University of Salamanca

DOCTORAL DISSERTATION

Deep Learning for Computer Vision in Smart Cities

Departament of Computer Science and Automation

Faculty of Science

Author:

David García Retuerta

Supervisors:

Sara Rodríguez González

Pablo Chamoso Santos

A dissertation submitted in fulfillment of the requirements for the degree of Doctor of Philosophy (Ph.D.) in Computer Science at the University of Salamanca

Salamanca, 2022

Statement of Authorship

This thesis project is presented by David García Retuerta, under the title "Deep Learning for Computer Vision in Smart Cities" in fulfillment of the requirements for the Doctorate Degree in Computer Engineering, University of Salamanca. This thesis has been carried out under the supervision of Dr. Sara Rodríguez González and Dr. Pablo Chamoso Santos; Professors at the Department of Informatics and Automation Control, University of Salamanca.

Salamanca, May 28, 2022

Author:

Firmado por DAVID GARCIA con un certificado emitido por AC FNMT Usuarios

David García Retuerta

Supervisors:

RODRIGUEZ GONZALEZ SARA - 70864126E

Firmado digitalmente por RODRIGUEZ GONZALEZ SARA - 70864126E
Fecha: 2022.06.22 18:51:03 +02'00'

Dr. Sara Rodríguez González

Firmado digitalmente por CHAMOSO SANTOS PABLO - 70888952P
Fecha: 2022.06.21 13:55:43 +02'00'

Dr. Pablo Chamoso Santos

Abstract

The Digital Age has caused a rapid shift from traditional industry to an economy mainly based upon information technology. According to recent studies, 74 zettabytes (ZB) of data have been generated, captured and replicated in the world in 2021, with video accounting for 82% of internet traffic. This figure has been amplified due to the coronavirus pandemic, and it is expected to keep increasing, reaching 149 ZB by 2024. Processing this impressive amount of information is one of the main scientific challenges of our time. Against this backdrop, Machine Learning (ML) and two related paradigms have emerged: big data and deep learning. These disciplines take advantage of mathematical optimization methods, bioinspiration and modern Graphics Processing Units (GPUs) to manage large datasets efficiently and effectively.

Cities from around the world have adapted the previous methods to make use of the newly available data, promoting themselves as “smart”. Apart from aiming to integrate innovative technologies in their daily operation, Smart Cities (SCs) aim to attract new residents and external investors.

Some of the key motivations of the Horizon projects and NextGenerationEU funds are precisely to make cities more digital, greener, healthier and robust. Artificial Intelligence (AI) can greatly contribute to the achievement of those objectives. Several lines of action have been identified in SCs, such as: smart mobility, smart environment, smart people, smart living and smart economy.

This dissertation focuses on vision applications of deep learning within the scope of SCs. Theoretical and practical research gaps are identified and suitable solutions are proposed. As a result, the state of the art has been pushed forward and new use cases have been successfully implemented. A novel solution is proposed for each of the identified lines of action.

Two models have been designed and evaluated with special attention to efficiency and scalability, and a third model has been created and tested focusing on accuracy within a high-resource environment. Moreover, two novel methods have been developed: a method for automatising crucial healthcare challenges, making early diagnosis an option; and another method for automatic unbiased cadastral categorization.

Acknowledgements

I would like to express my gratitude to all those who, with their support, encouragement and kindness; have made this great milestone of my career possible — this dissertation is dedicated to you.

First of all, I must thank my supervisors, Sara Rodríguez and Pablo Chamoso, for their dedication and guidance. This thesis wouldn't have been possible without their valuable support.

Secondly, I want to express my profound gratitude to John A. Lee and the rest of the MIRO Machine Learning Group from the Université catholique de Louvain for their overwhelming hospitality during my research stay in their lab. The months I spent working with them greatly enriched my scientific capabilities and the research which lead to this dissertation.

I also want to thank Besik Dundua for his earnest hospitality, valuable pieces of advice and for going out of his way to welcome me. His interest in quality research, not quantity, taught me to deal with the state of the art as an old friend with whom to have an enjoyable conversation, instead of as a intimidating professor I must learn from. The months I spent under your supervision truly touched me.

Moreover, I also want to thank my colleagues at BISITE research group for their company and friendship over the years. May the future take us apart, may the destiny bring us together.

Lastly, I am grateful to my encouraging relatives, my fellow professors of "Programación II" and my hard-working students.

Contents

Lists of Figures

Lists of Tables

Abbreviations

AI	**A**rtificial **I**ntelligence
AUROC	**A**rea **U**nder the **R**eceiver **O**perating **C**haracteristic
AutoML	**Auto**matic **M**achine **L**earning
CT	**C**omputed **T**omography
CTV	**C**linical **T**arget **V**olume
HAR	**H**uman **A**ctivity **R**ecognition
IMRT	**I**ntensity-**M**odulated **R**adiation **T**herapy
IoT	**I**nternet **o**f **T**hings
gMLP	**g**ated **M**ulti-**L**ayer **P**erceptron
GPU	**G**raphics **P**rocessing **U**nit
ML	**M**achine **L**earning
MLP	**M**ulti **L**ayer **P**ercepton
MSE	**M**ean **S**quare **E**rror
NLP	**N**atural **L**anguage **P**rocessing
OAR	**O**rgan **A**t **R**isk
PET	**P**ositron **E**mission **T**omography
RNN	**R**ecurrent **N**eural **N**etwork
SC	**S**mart **C**ity
SVM	**S**upport **V**ector **M**achine
ViT	**Vi**sion **T**ransformer

Practice makes the master.

— *Patrick Rothfuss, The Name of the Wind.*

Chapter 1

Introduction

Chapter 1

Introduction

1.1 Introduction

According to recent reports, over the course of 2021, 74 ZB of data have been generated, captured and replicated in the world (Aundhkar and Guja [2021]). Moreover, estimates for 2021 stated video would account for 82% of internet traffic (Cisco [2016]). In addition, the COVID-19 pandemic has amplified the amount of data generated globally and it is expected to continue increasing, reaching 149 ZB by 2024 (Turi and Li [2021]). The efficient processing of such a vast amount of information is one of the main research challenges the scientific community is facing at the moment.

Additionally to the data generated directly through human interaction with electronic devices, a notable share of today's information flow is generated by the estimated 18 billion connected Internet Of Things (IoT) devices (Simiscuka and Muntean [2021]). A great share of this data is being generated in the metropolis all around the world. SCs comprehend that new data sources offer opportunities to attract new citizens, awaken the interest of external investors, integrate technology in their daily operation and generate economic growth.

Precisely, some of the key motivations of the Horizon projects and NextGenerationEU funds are to make cities more digital, greener, healthier and more robust; ML has the potential to help achieve those objectives. Over the last few years, SC developers have been working on those fields. In order to attain greater focus, five vertical markets have been defined: Smart Mobility, Smart Environment, Smart People, Smart Living and Smart Economy. To achieve a balanced development of the metropolis, none of those vertical markets should be left out. Keeping this fact in mind, the current research work proposes a novel method for each of the mentioned vertical markets.

Novel deep learning methods are presented in this dissertation, focusing on vision applications within the scope of SCs. Theoretical and practical research gaps are identified, considered, and suitable solutions are presented. The proposed methods make extensive usage of the following deep learning methods: convolutional layers, transformers, U-Net-shaped networks and transfer learning.

The hypothesis put forward in this dissertation is that SCs can greatly benefit from innovative computer vision algorithms, which are based on deep learning techniques. There are two distinct dimensions of the hypothesis: applying known methods to existing problems (practical dimension) and optimising the current cutting-edge algorithms (theoretical dimension). Results indicate that the proposed methods are cost-effective and efficient, achieving an appropriate accuracy in new real-life applications and improving the performance of known architectures.

The present chapter is organized as follows: problem description and motivation are introduced in Sec. 1.2. Sec. 1.3 shows research hypothesis and objectives of this dissertation, and the methodology is presented in Sec. 1.4. Finally, Sec. 1.5 provides the structure of the dissertation.

1.2 Problem Description and Motivation

Innovative approaches in urban areas are currently triggering the “Smart City Revolution”. The ever-increasing urbanization of the world is straining the existing infrastructure of metropolises, in which the quality of life and socio-economic development are facing numerous issues. Cities can effectively tackle challenges by ensuring a high level of citizen involvement, ensuring a wide-spread usage of internet-based applications, boosting collaboration among their institutions, and a great many other initiatives. When clustered into vertical markets, some of the most important opportunities are (Kumar [2020]):

- **Smart Economy**. Embracing open data and smart systems can boost innovation, development of businesses, and reduce administration costs. A start-up which developed a solution for a local city can export that technology all over the world.
- **Smart People**. Committed citizens can inform of infrastructural break downs, ground objects in urgent need of maintenance and any kind of security risk.
- **Smart Environment**. Better usage of natural resources and environment protection can positively affect the health of citizens and boost the city sustainability.
- **Smart Mobility**. A higher efficiency in all transport can improve the air quality, decrease commuting times and reduce road fatality rates. Smart parking lots can greatly reduce the search traffic on the streets.

- **Smart Living**. IoT devices can improve people's healthcare, analysing the vital signs and rising accurate early warnings.

Considering all the previous information, and the fact that cameras have become ubiquitous in every corner of the developed world; computer vision emerges as an outstanding source of knowledge for a wide variety of applications. Indeed, the combination of computer vision and data science tools has many potential applications which could be used to promote efficient and sustainable development in SCs.

Currently, the scientific community is spending great amounts of time and funding on finding new applications of ML and improving the existing methods. So much so that a) it is considered a key part of the fourth industrial revolution, b) it is a buzzword year after year since the 2010s, c) important actors have asserted that the so-called "AI Revolution" has already started (Harari [2017]). ML brings many advantages to society, the greatest of which are:

1. **A wide range of applications in all fields**. It represents a major leap forward in the way computers can learn from data. It is becoming ubiquitous in social media, recommendation systems and virtual personal assistants, among other applications.

2. **Increased automation of tasks**. ML is a technology which can make some jobs easier by mimicking human behaviour. For instance, surveillance companies no longer need a human operator to check the video captured by each CCTV camera, as there are ML algorithms capable of understanding video.

3. **Easy and prompt identification of trends and patterns**. Shifts in the distribution of a considered variable can be detected, first as an anomaly and then as an upcoming trend. This feature contributed to the adoption of ML in investment companies.

4. **Constant improvement**. New architectures and approaches are being constantly proposed, with new revolutionary proposals appearing every 5-10 years. Since 2010, convolutional networks, long short-term memory blocks and transformers have been responsible for breakthroughs in the field.

5. **Capacity to deal with multi-dimensional and multi-variety data**. Data science projects typically include data from different sensors and different sources, which an artificial neural network has the ability to use as input.

Some sectors which can greatly benefit from ML include:

- **Financial markets**. Within the past 10 years, the financial industry has spent a lot of resources on using complex models in stock forecast. For example, in the last few years, it has become common to use Natural Language Processing (NLP) to predict the stock market on the basis of the news (Kim et al. [2014]).

- **Automotive industry**. In 2018, 78% of automotive companies invested in skills and training for ML, heavily using ML in their marketing campaigns (Schrage and Kiron [2018]). For example, Tesla released its *Autopilot* in 2015, announcing their intent for a yet-to-come update to offer SAE Level 5 (full autonomous driving).

- **Healthcare industry**. ML has the potential to speed up the analysis of the great amounts of data gathered for each patient. In particular, clinical decision support tools are being used to process large datasets in order to identify a new diseases.

- **Agriculture**. Rich recommendations and insights about crops can be obtained using data science and ML models. Algorithms can be used to make the most of pre-harvesting, harvesting and post-harvesting periods.

- **Military industry**. Life-and-death decisions on the battlefield require brief consideration before taking action, a task with ML could take responsibility for. Simulating battles and maximizing the likelihood of the most advantageous outcome is an option most armies would be interested in.

- **Social networks and search engines**. Big tech companies heavily rely on ML for providing friend recommendations or search results. For example, the CLIP network (Radford et al. [2021]) can predict the existence of any label within a given image, a perfect suit for reverse image search.

- **Most engineering disciplines**. With ML becoming an ongoing trend for the last decades, most technical fields are starting to integrate it in their research.

- **Social sciences**. Their current era of data abundance perfectly suits the ML paradigm, where loose approaches can closely model societies behaviour.

- **Art and advertising industries**. New art styles are being created by machines paintings, compositions and writings. Moreover, advertising content can now be delivered to a more targeted audience than ever using ML.

Interestingly, despite all the recent and revolutionary advances in this field, there is still much room for improvement in ML. In particular, its applications for computer vision have only gained momentum since the popularization of deep learning, less than 20 years ago, and its usage in SCs their still limited.

Thus, there is a great opportunity to design and test new methods which will optimise this field. In addition, there is a vast unexplored potential for applications which could satisfy a great portion of citizens, and which could optimise the use of resources in their administrations.

The following questions motivated the research:

- What time-consuming tasks could be automatised?
- Are modern vision algorithms optimised regarding their computational cost?
- Will the public sector become more data-centred in the future?
- What time-sensitive tasks are currently burdened by manual execution?
- Could the reliability of automatic video understanding be improved?
- Could the current data of the administrations be successfully used in ML?

This dissertation tries to find answers to the previous questions by: a) using ML methods to solve certain challenges of SCs, b) improving the state of the art of the related algorithms.

The selected ML methods focus on computer vision, due to the recent explosion in video content and the increased popularity of filming hardware. Moreover, recent advances in object detection, face, action and activity recognition and human pose estimation have overcome many of their past limitations, creating a solid scientific foundation which can be used to solve several real-life problems (Voulodimos et al. [2018]).

1.3 Hypothesis and Purpose

This research work provides different solutions to the identified research gaps in the current state of the art regarding deep learning applications for computer vision within the context of SCs. In particular:

The initial hypothesis of this research work is that it is possible to enhance, optimise or accelerate the current algorithms and techniques used for computer vision in SCs, by applying novel deep learning methods.

Several research gaps have been identified in the design of computer vision algorithms and in their applications, which could increase the survival rate of patients, minimize the cost of state-of-the-art solutions and open new research lines. We modify existing deep learning models to make them more efficient, accurate and effective. Moreover, we develop new solutions to automatise tasks which are performed manually at the moment. Therefore, the final result of this work is to push forward the field of computer vision and, in particular, its applications related to SCs.

The main objective of this dissertation is to enhance the accuracy and to lower the computational requirements of existing computer vision models, as well as to use them for achieving effective solutions in SCs.

This dissertation touches various research areas within the computer vision field: medical image segmentation, camera-based human activity recognition (HAR), and image classification. They were chosen due to their relevance and improvement potential. Before developing new methods, the state of the art of the computer vision field in the mentioned research areas have been studied, research gaps have been identified and the found problems resolved.

1.4 Methodology

A formal and well-structured methodology is necessary to perform quality research which produces valid results, and to ensure that a valid outcome is obtained at each research step. The *action-research* method (Reason and Bradbury [2001]) was chosen and applied in the current work. It stands out as it is action and change-oriented, enabling the researcher to focus on well-defined problems to produce new knowledge based on other works over a period of time. It has become a common approach to empirical research. The process is as follows: (1) identification of the real problem, (2) study of the possible hypotheses, selection of one and development of a proposal, (3) verification of the selected hypothesis, and (4) to draw conclusions after the evaluation of the obtained results. In our case:

1. Identification and description of the characteristics of the faced problem. The characteristics of the SC infrastructure and its used algorithms are defined and all of the possible hypotheses proposed.
2. Study of the possible hypotheses, selection of one and development of a proposal. An incremental study of the state of the art is performed and the previous hypotheses analysed. A theoretical framework is thus obtained and the most promising hypotheses selected. A proposal with strong scientific foundations is then put forward.

3. Verification of the selected hypotheses. An iterative and progressive design of a solution is then carried out. The relevant available visual data is then gathered, several components which take care of the different aspects of the problem implemented, and they are combined into a comprehensive model.

4. To draw conclusions after the evaluation of the obtained results. The new model is tested to analyse its behaviour: its components, functionality and each of its iterations are evaluated and a large number of raw results obtained. Such a results are analysed to perform the formulation of conclusions.

In parallel to the mentioned stages, a continuous dissemination of knowledge, results and experiences with the scientific community was carried out. This process materialised in the form of publications in scientific journals, attendance to international conferences and presentation of works.

1.5 Structure of the Thesis Dissertation

This doctoral thesis is divided into eight chapters and two appendices. Their structure is described below.

Chapter 1 provides an introduction to the carried out research and to the dissertation. Specifically, it describes the motivation of the research and the problems which were intended to be solved. The hypothesis, objectives and methodology which have led to this dissertation are also presented.

Chapter 2 provides a review of the state of the art as well as describes the concepts which are important for the readers' understanding of the chapters that follow. ML, computer vision, SCs and their respective sub-fields are described in detail; as well as their relation.

Chapter 3 is related to the vertical market of "Smart Mobility". An introduction to its related concepts and importance is presented first, and a particular use case afterwards. The most promising natural language processing and computer vision network of the last few years is modified to lower its computational cost and resource requirements, all while slightly increasing its accuracy. It has the potential to increase the spread of self-driving vehicles and bring computer vision closer to medical techniques. This chapter is longer than others due to the extensive testing performed in order to validate the improvements of the proposal.

Chapter 4 is related to the vertical market of "Smart People". An introduction to its related concepts and importance is presented first, and a particular use case afterwards.

An attention network is developed to improve camera-based HAR, a core task of a wide variety of applications. The performance of the new network and its architecture allow developers to run it reliably even when on modest hardware.

Chapter 5 is related to the vertical market of “Smart Living”. An introduction to its related concepts and importance is presented first, and a particular use case afterwards. An automated ML framework is used for performing clinical target volume and organ segmentation in an application related to proton therapy for patients with esophageal cancer. This use case presents an improvement over previous techniques found on the state of the art, getting the treatment planning one step closer to automation in real-life clinical cases.

Chapter 6 is related to the vertical market of “Smart Economy”. An introduction to its related concepts and importance is presented first, and a particular use case afterwards. Transfer learning is used to automatically categorise the constructive typology of residential buildings according to their facade. A method for automatising this administrative task is put forward, and recommendations related on the data intake process formulated. This use case has the potential to reduce the economical and human requirements to run a cadastre.

Chapter 7 is related to the vertical market of “Smart Environment”. An introduction to its related concepts and importance is presented first, and a particular use case afterwards. A method is developed to obtain a species recognition and possible disease infection after a tick bite. It has the potential to save lives thanks to enabling a fast diagnostic and to help spread this medical feature to sparsely populated regions.

Chapter 8 draws conclusions from the research work and enumerates the contributions of the developments to the state of the art. Future research lines which have been opened during the course of the Ph.D. studies are also presented.

Appendix A contains a list of the tangible outcomes of this Ph.D. program. First, the list of scientific publications which took place during the Ph.D. program by the Ph.D. candidate. The list includes book chapters and international journals listed as part of the Journal Citation Reports (JCR). Secondly, the research projects in which the Ph.D. candidate participated are also mentioned as, to some extent, they have contributed to the scientific development of the research. Lastly, his research stays in renowned research institutions abroad are mentioned.

Appendix B provides a translation into Spanish of the dissertation.

Finally, a list with all the bibliographical references used and referenced during this dissertation is presented.

Chapter 2

Background

VNiVERSiDAD D SALAMANCA

Chapter 2

Background

2.1 Introduction

This chapter introduces the fundamental concepts and techniques that will be relevant in the chapters that follow. First, we introduce machine learning, describing its evolution, importance, cornerstones and one of its most robust models (Support Vector Machines). Secondly, two of ML's main subfields are presented — deep learning and transfer learning. Thirdly, computer vision is considered, listing and describing in detail two of its most important architectures at the moment — U-Net shaped networks and Transformers. Lastly, the concept of smart city is put forward, its current importance analysed, and its tendencies identified. Open data is examined as it is part of its foundation, and the different aspects of SCs are categorized in five vertical markets.

2.2 Machine Learning

The last decade has seen an impressive rise of ML based methods, affecting numerous industries such as autonomous driving, healthcare, finances, manufacturing, energy production, etc. ML is considered to be a turning point for humanity in our age, the same way that computers revolutionised the world in the 80's and 90's. At a basic level, the objective of ML is to identify patterns in data, extracting knowledge which can then be used for a wide variety of purposes.

ML is part of the broader concept of *Artificial Intelligence* — a recently established field based on mathematical optimization, statistical learning, data mining and computer science. It is a subfield that provides systems the ability to automatically learn and improve from experience without being explicitly programmed, i.e., to do what people do naturally: learn by using examples.

Some of its most important milestones have been: IMB's Deep Blue beats Gary Kasparov at chess (Campbell et al. [2002]), personal assistants become mainstream in smartphones (Strayer et al. [2017]), the ResNet network achieves a better classification accuracy than humans (He et al. [2016]), Google's Deepmind Alphago defeats Go's champion (Holcomb et al. [2018]), and Mercedes-Benz develops a level 3 autonomous car (van der Aalst [2022]).

One of the most important models in ML is Artificial Neural Networks (ANNs), a biologically-inspired connectionist system. Their "perceptron" is an algorithm which loosely mimics the neurons in a biological brain, and its connections act like the synapses in a biological brain transmitting information.

The Universal Approximation Theorem (Gelenbe et al. [1999]) proved mathematically that there is always a feed-forward neural network which can approximate any given continuous function $f(x)$ between two Euclidean spaces, with respect to the compact convergence topology, for any number of inputs and outputs. This promising property of ANNs, together with the improved computational resources at the time and the increased data availability, generated an interest explosion in the field which left behind the so-called "AI winter" of the late 80's and early 90's. This boom eventually led to deep learning and the data-hungry models we currently coexist with.

There are three main tasks in ML (Russell Stuart [2010]):

- **Supervised Learning**. It relies on labelled data and, using a loss function, obtains feedback from its predictions. It is used to forecast an output, and its goal is to generalise using the input, so that it works properly with new unseen data. Classification and regression problems fall within this category. Some notable algorithms are: Artificial Neural Networks, Support Vector Machines, and Random Forest.

- **Unsupervised Learning**. It relies on unlabelled data and, without receiving any feedback, finds the hidden patterns. Clustering and associations problems fall within this category. Some notable algorithms are: K-means, k-Nearest Neighbours (k-NN), Principal Component Analysis (PCA).

- **Semi-supervised Learning**. It relies on a small amount of labelled data and a large amount of unlabelled data, a special case of weak supervision. It is a less important category than the former two.

2.2.1 Support Vector Machine

A Support Vector Machine (SVM) is a supervised learning model commonly used in classification and regression tasks. It is considered to be one of the most robust classifiers for the time being, and it is one of the most common choices to retrain the last layer of pretrained networks (Lu et al. [2015]).

Let $(x_1, y_1), ..., (x_n, y_n)$ be a training set of n-points, where $y_i = 1$ if $x_i \in C_1$, and $y_i = -1$ if $x_i \in C_2$. Consequently, C_1 and C_2 are the two classes to model, and $x_i \in \mathbb{R}^n$ $\forall i \in \{1, ..., n\}$.

The goal is to find the maximum-margin hyperplane which divides the $x_i/y_i = 1$ from the $x_i/y_i = -1$. That is, the goal is to maximize the distance from the hyperplane to the nearest x_i from either group.

Any hyperplane can be written as the $x_j \in \mathbb{R}^k/w^T x - b = 0$, where w is the normal vector to the hyperplane.

We can face two different scenarios:

- **Linearly separable training data**. In this case a *hard margin* can be used, and the optimization problem can be defined as:

  Minimizing $||w||_2$ subject to the constraint $y_i \cdot (w^T x - b) \geq 1$ for $i \in \{1, ..., n\}$.

  $||w||_2$ refers to the Euclidean distance from a point to a plane, which in the case of the distance from hyperplane H to an arbitrary point y is:

$$d(H, y) = \frac{|(\sum\limits_{i=1}^{n} a_i y_i) - d|}{\sqrt{\sum\limits_{i=1}^{n} a_i^2}} \tag{2.1}$$

  with $y = (y_1, ..., y_n)$, H defined by a point $p = (p_1, ..., p_n)$ and its normal vector $a = (a_1, ..., a_n)$, and $d = p \cdot a$.

  Note that y_i is the i-th target and $(w^T x - b)$ is the i-th output.

- **Not linearly separable training data**. In this case a *soft margin* must be used, making use of the squared hinge loss function:

$$l(x_i, y_i) = \max(0, 1 - y_i \cdot (w^T x_i - b)) \tag{2.2}$$

  Therefore, the goal is to minimize:

$$\lambda||w||_2^2 + \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i \cdot (w^T x_i - b))^2 \tag{2.3}$$

where $\lambda \in \mathbb{R}^+$ is a new parameter to optimize.

2.2.2 Deep Learning

Deep learning is a family of ML methods characterised by the use of a high number of layers in the network. The main advantages of *deeper* networks are: “distributed representations” and “the power of depth” (Eldan and Shamir [2016], LeCun et al. [2015]). The former refers to the ability of deep learning models to divide the space of features efficiently by increasing the number of input examples linearly, as opposed to traditional methods which need to increase the number of examples exponentially (due to the so-called *curse of dimensionality*). The latter refers to the need of traditional methods to take exponential number of nodes in a single hidden layer to represent an arbitrary function, when deep learning takes a linear number of nodes by increase the depth of the network.

Deep learning-based research has demonstrated excellent performance in numerous study domains, including computer vision (Tan et al. [2018]), object detection and recognition (Ijjina and Mohan [2017]) and NLP (Angeleas and Bourbakis [2016], Young et al. [2018]). Early studies, such as (Ha et al. [2015], Lane and Georgiev [2015], Yang et al. [2015]), that investigated the applicability of deep learning in HAR, have inspired researchers to work in this field actively. The main advantage of deep learning over traditional ML algorithms is the reduced effort needed while picking up the right features, by automatically extracting abstract features through several hidden layers. Other advantages include:

- Layer-by-Layer architecture that enables deep models to learn descriptive features from complex and multimodal sensory data and get high accuracy rates using powerful GPUs.
- Neural network structure diversification, which provides flexibility to choose an appropriate model based on the learning environment. For instance, Convolutional Neural Networks (CNNs) are preferred to analyse multimodal sensory data by exploring local connections (Hammerla et al. [2016]). While Recurrent Reural Retworks (RNNs) are suitable for streaming sensory data in HAR because they extract temporal connections and incrementally learn information through time intervals.

- Optimization function role in unified network composition, providing that deep neural networks are detachable. This feature allows effective solutions in a variety of deep learning techniques such as deep transfer learning (Akbari and Jafari [2019]), deep active learning (Gudur et al. [2019]), and deep attention mechanisms (Murahari and Plötz [2018]).

Regarding its programming languages, Python is the absolute leader with more than 60% of ML developers are using and prioritizing it. In particular, three of its libraries stand out: PyTorch, TensorFlow and Keras (a high-level interface of TensorFlow).

TensorFlow gained a great popularity in all deep learning sectors initially, but there is currently a tendency towards PyTorch. State-of-the-art models and new architecture are most commonly implemented in Pytorch over the last two years, showing that it became the preferred programming language for top-researchers. Overall, PyTorch surpassed TensorFlow for the first time in April 2021, and has continued to gain popularity ever since (Fig. 2.1).

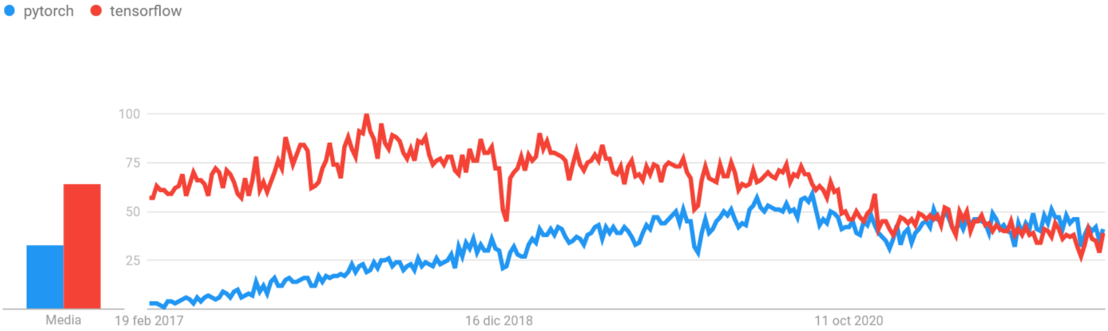

Fig. 2.1: Popularity of PyTorch and TensorFlow over the period 2017-2021. Source: Google Trends

2.2.2.1 Transfer Learning

Transfer learning is a typical ML approach that allows the classification ability of the learning model to be transferred from the predefined environment to a dynamic setting. Transfer learning is especially useful for resolving problems with distribution discrepancies (Chen et al. [2021]). It prevents learning model performance from degrading when the distributions of the training and test data differ. This problem emerges in the activity recognition context when activity recognition models are applied to new configurations such as recognition of new activities, involvement of new sensors and users, etc.

The source domain in transfer learning corresponds to domains incorporating vast amounts of annotated data and the main objective is to use the information from the source domain to annotate the samples in the target domain. In the activity recognition field, the source domain relates to the initial configuration, while the target domain denotes a new deployment that the system has never seen before. A diagram summarising of the basic functioning of transfer learning can be found in Fig. 2.2.

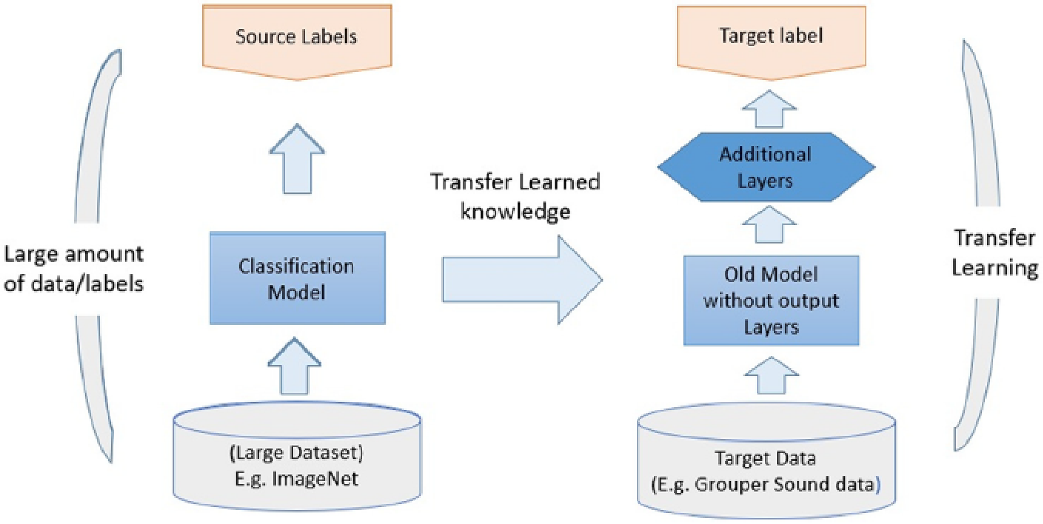
Source Labels
Target label
Additional Layers
Transfer Learned knowledge
Large amount of data/labels
Classification Model
Old Model without output Layers
Transfer Learning
(Large Dataset) E.g. ImageNet
Target Data (E.g. Grouper Sound data)

FIG. 2.2: Transfer learning diagram

2.3 Computer Vision

Computer vision is an interdisciplinary scientific field which studies how computers can gain high-level knowledge from digital images or videos. It is also bio-inspired, as it aims to reproduce the functioning of the human visual system.

Deep learning has fuelled research in a great variety of computer vision applications, as it allows for high amounts of data to be processed efficiently. It has allowed for many stages of common vision-related tasks to be automatised. The process of computer vision tasks is commonly divided in: (1) acquiring the data, (2) processing, (3) analysing, and (4) understanding.

The image data can take many forms, such as: video sequences, medical images (PET, CT, MRI, . . .), 360º images, input from multiple cameras, 3D images, etc.

Some of the most frequent applications of computer vision are:

- **Medicine**. Cancer detection, cell classification and disease progression modelling have already benefited from computer vision methods.

- **Machine Vision**. Animal and crop monitoring, customer tracking and multi-player pose tracking are among the latest successful applications of computer vision.
- **Military**. Unmanned combat aerial vehicles, long-range missiles and surveillance satellite make use of computer vision to achieve superiority in the battlefield.
- **Autonomous vehicles**. Traffic sign detection, collision avoidance systems, road condition monitoring are some of the new computer vision-based systems which have allowed the current popularisation of self-driving cars.

An important milestone of the deep learning related to computer vision took place in 2015, when ResNet achieved a better image classification than humans in the ImageNet competition He et al. [2016]. Moreover, deep learning-based methods won the competition year after year since 2011, when XRCE won with a “traditional” computer vision approach (Sánchez and Perronnin [2011]). Fig. 2.3 shows the ImageNet results from 2011 to 2016.

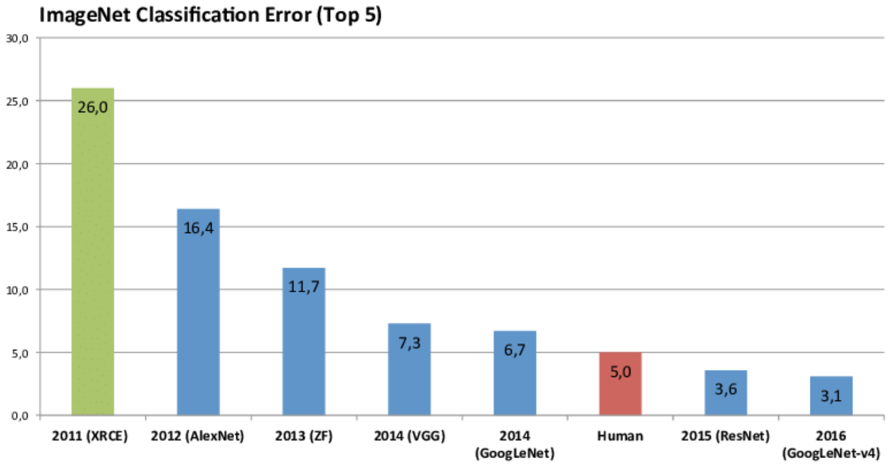

Fig. 2.3: ImageNet top-5 classification winners. Traditional approaches (green), deep learning approaches (blue) and humans (red).

2.3.1 U-Net shaped networks

Since the introduction of U-net in 2015, it has revolutionised the field of biomedical image segmentation, with only Transformer-based architectures achieving better performance in some datasets (Tang et al. [2021]).

The original U-net network is based on a contracting path and an expansive path, as it can be observed in Fig. 2.4. It makes use extensively of convolutions, max pooling and the ReLU activation function (Ronneberger et al. [2015]).

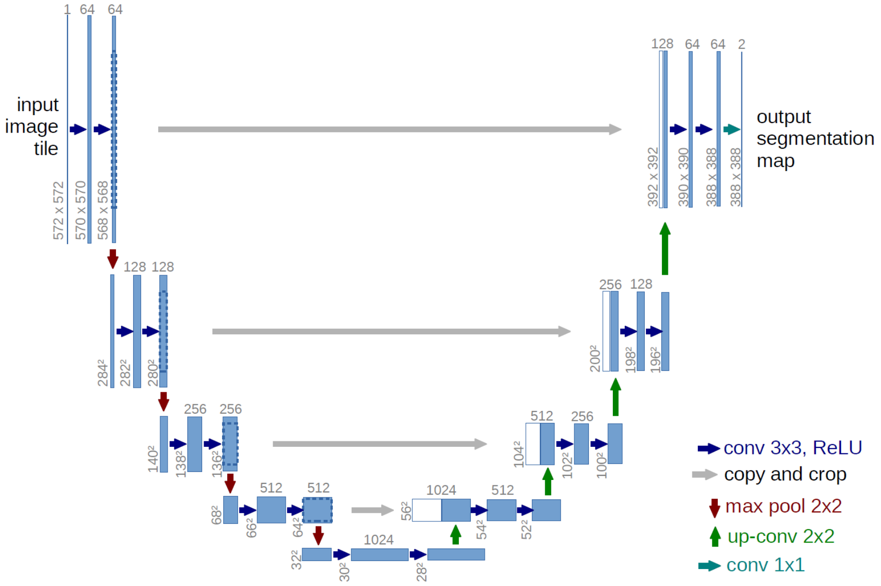

FIG. 2.4: U-net architecture summary. Example for 32x32 pixels in the lowest resolution. Extracted from Ronneberger et al. [2015]

The contraction phase reduces the spatial information and increases the feature information, learning an abstract representation of the input image. Meanwhile, the expansive phase combines spatial information and the obtained features through a sequence of up-convolutions and concatenations with high-resolution features of previous layers (skipped connections), using the abstract representation to produce a semantic segmentation mask.

U-net defines a cross-entropy energy function which it tries to minimise, therefore understanding "learning" as minimizing a loss functional. The used energy function is thus defined to penalise at each position the derivation of $p_{l_{(x)}}(x)$ from 1:

$$E = \sum_{x \in \Omega} \omega(x) \log(p_{l_{(x)}}(x)) \tag{2.4}$$

where $p_k(x)$ is the approximated maximum-function:

$$p_k(x) = \exp(a_k(x))/(\sum_{k'=1}^{K} \exp(a_{k'}(x))) \tag{2.5}$$

with $a_k(x)$ being the activation in feature channel k at the pixel position $x \in \Omega$, where $\Omega \in \mathbb{Z}^2$; K being the number of input classes.

This definition results in $p_k(x)$ being close to 1 for the k which has the maximum activation $a_k(x)$, and $p_k(x)$ being close to 0 in all other cases.

Moreover, data augmentation is also extensively used in combination with the network, as medical datasets usually have a reduced amount of instances, and expanding them is a very costly and time-consuming process (Tajbakhsh et al. [2020]).

2.3.2 Transformers

Transformer models have revolutionised the fields of NLP, text data processing and computer vision since their introduction in 2017 (Vaswani et al. [2017]), completely replacing RNNs and CNNs in most advanced applications. Their increased parallelization allows training on larger datasets than was once possible, leading to the development of very successful models such as BERT (Devlin et al. [2018]) and the three iterations of GPT (Brown et al. [2020]).

The Sequence-to-Sequence (Seq2Seq) architecture is their cornerstone, as the transformer's design mimics the encoder-decoder attention mechanism of the former. It works transforming a given sequence of elements into another sequence. The process consists of an encoder module which maps the input into a higher dimensional space (n-dimensional vector), and then a decoder module which turns the n-dimensional vector into a m-dimensional output sequence; with $n, m \in \mathbb{R}$ (Sutskever et al. [2014]).

The leap forward of Transformers is their *attention mechanism* which does not imply any kind of recurrent network. It decides at each step what are the important parts of the input sequence based on the Scaled Dot-Product Attention and the Multi-Head Attention, represented in Fig. 2.5.

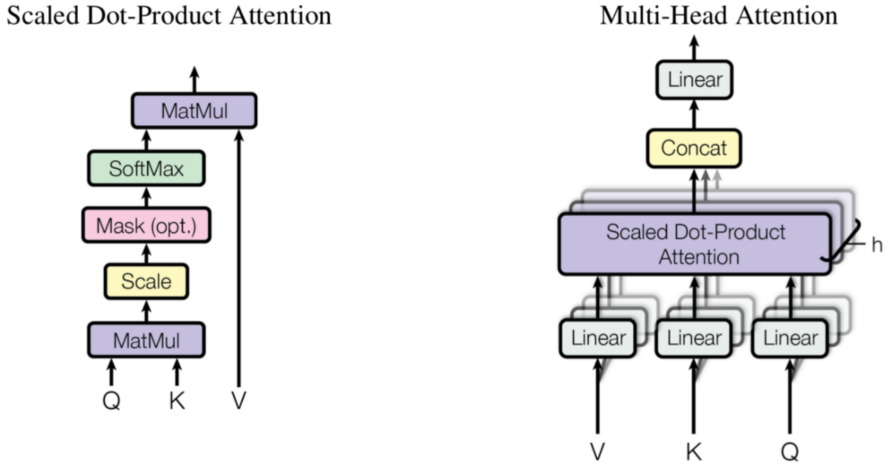

Fig. 2.5: Scaled Dot-Product Attention (left) and Multi-Head Attention (right). Extracted from Vaswani et al. [2017]

In particular, the matrix of outputs of the Scaled Dot-Product is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{2.6}$$

where d_k is the dimension of K, and Q, K, V are the so-called *query, key, value.*

This block is used in a bigger structure called *Multi-Head attention.* A single Multi-Head attention head, averaging outputs, is defined as:

$$\text{MultiHead}(Q, K, V) = \text{concat}(\text{head}_1, ..., \text{head}_h) \cdot W^O \tag{2.7}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, and W_x^y represents the corresponding parameter matrices. In particular: $W_i^Q \in \mathbb{R}^{d_{model} \times d_Q}, W_i^K \in \mathbb{R}^{d_{model} \times d_K}, W_i^V \in \mathbb{R}^{d_{model} \times d_V}$ and $W^O \in \mathbb{R}^{d_{model} \times h \cdot d_V}$.

These blocks are used on top of other modules to create an encode-decoder architecture. Transformers are mainly formed by Multi-Head attention and feed-forward layers. Fig. 2.6 shows the full transformer model architecture and all its modules.

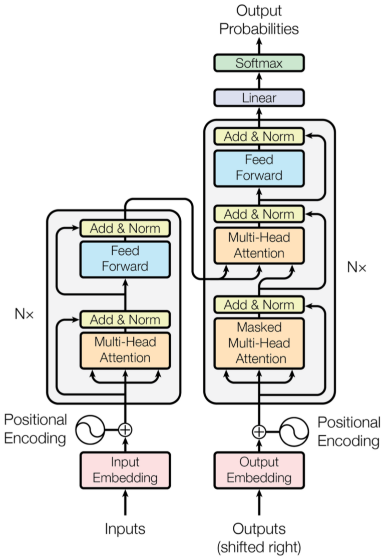

FIG. 2.6: Transformer model architecture. Encoder (left) and decoder (right). Extracted from Vaswani et al. [2017]

This complex architecture leads to a complexity per layer of $O(r \cdot n \cdot d)$, significantly lower than in traditional self-attention architectures, $O(n^2 \cdot d)$, recurrent networks, $O(n \cdot d^2)$, and convolutional networks, $O(k \cdot n \cdot d^2)$. n being the sequence length, d being the representation dimension, k being the kernel size of convolutions and r being the size of the neighbourhood.

2.4 Smart Cities

Data is the primary source of progress in almost all sectors of human activity. It can be a driver of change that improves people's lives (Yeh [2017]), generates value (Dameri and Rosenthal-Sabroux [2014]) and optimizes decision-making processes. In recent years, new data-related paradigms have emerged: IoT devices which help gather a wide range of data (Marjani et al. [2017]) and ML algorithms. As a result, the technology for SCs to thrive is already easily available and widespread, foretelling their future importance. In this context, SCs aim to provide sustainable urban development worldwide (Vidiasova et al. [2017]).

Smart city is an emerging concept in constant revision which currently refers to a type of urban development based on sustainability that is able to respond adequately to the basic needs of institutions, businesses and the inhabitants themselves, both economically, operationally, socially and environmentally.

While AI and SCs have been trending for a long time, their adoption has been slow (Allam and Dhunny [2019]). As popularity analyses show (Fig. 2.7), the interest in SCs and ML has been equivalent for a long time; where both were booming in 2013. However, SCs reached a plateau in 2015, while ML continued to grow. This time is considered to be the beginning of the "AI revolution" (Walsh [2017]). During these years, the SC concept has undergone some changes and offered many advances, commonly including air quality monitoring, easy-to-use public transport, improved commuting and collaborative governance.

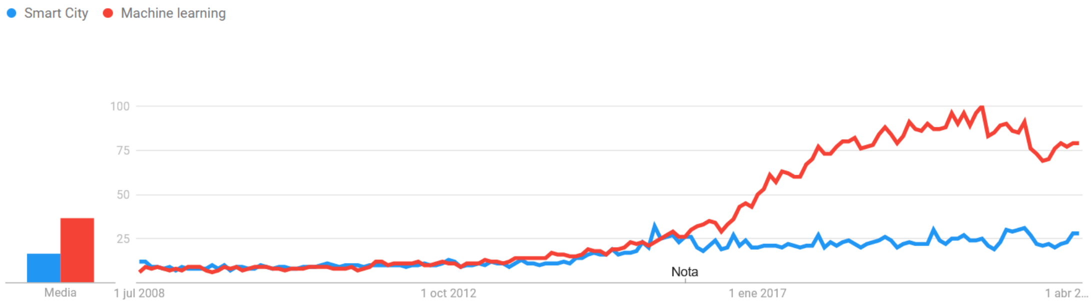

Fig. 2.7: Popularity of SCs and ML in the period 2008-2021. Source: Google Trends

IoT provides solutions for the integration of different devices and technologies, and enables them to interact with each other. It can boost productivity and performance of different SC domains. They are a basis for the successful establishment of SC infrastructure and related services (Park et al. [2018]). Nowadays, there are already proven, real-world solutions which integrate IoT devices in Smart Territories, for example, Dynamic Street Light Control systems (Ouerhani et al. [2016]).

Another important technology to consider is Big Data. It is of strategic value to Smart Territories, given the ever-growing number of devices for data capture, the improved processing capabilities and the good results of their models. The large amounts of data gathered by all kinds of sensors give SCs the potential to obtain valuable information on their cities and develop real-world solutions (Hashem et al. [2016]).

Many authors are already preparing us for the forthcoming “AI revolution”. ML models are already better than humans at some classification tasks and they keep improving continuously, which has a direct effect on the economy due to an increase in robot productivity (Alonso et al. [2020]). Thanks to advanced computer vision techniques, ophthalmology will be revolutionised (Hallak and Azar [2020]) and post-school education will be improved by adapting its educational programming (Butler-Adam [2018]).

Vertical markets provide opportunities for the achievement of efficient and sustainable development (Garcia-Retuerta et al. [2021]). Categorising the different domains of a SC can help metropolises develop well-focused and balanced lines of action with clearly set priorities. In combination with ML, IoT and other technologies, this approach can make a city bloom.

2.4.1 Open Data

The concept of open data is a philosophy based on increasing the availability of data, without any restriction of copyright, patents or other control mechanisms. As a result, any person can make use of them and it can boost cooperation between institutions; and enterprises with a specific expertise can search for customers who could benefit from their solutions. Moreover, certain fields, such as SCs and cybersecurity, are known to greatly benefit from the open data/open-source philosophy (Weiss and Bailetti [2015]), with outstandingly successful projects, such as the UNIX operating systems (AlMarzouq et al. [2005]).

Open data has a capital importance to SCs all around the world, as it boosts innovation, eases the knowledge extraction and use in many sectors, and provides a positive change to their complex ecosystems (Neves et al. [2020]). Paris is an outstanding example of its effect, as their “Welcome City Lab” has developed several applications for tourism which work within the SC infrastructure based on its available open data (Courmont [2016]).

Many high-income countries have National Open Data Portals, which provide a unified source that developers may turn to. An exceptional example of this is the official portal for European data (https://data.europa.eu), which includes 1,282,950 open datasets from 36 EU and non-EU countries, grouped in 81 catalogues and labelled within thirteen categories.

Some low-income countries have also developed National Open Data Portals, although their number is much lower in comparison to those of high-income countries. Countries such as India have launched successful yet relatively small portals. The (Indian) Open Government Data Platform (https://data.gov.in) was launched in 2012 and currently includes almost half a million datasets clustered into around 10,000 catalogues. It accepts new dataset suggestions and shows some examples of successful solutions.

Open data has the potential of improving economic efficiency. With a predicted economic impact of €199.51 - 334.20 billion in EU+ countries by 2025 (Spichtinger and Blumesberger [2020]), there is no doubt of the growing importance of its market size. Efficiency gains include: improving environment protection (e.g., by improving the energy efficiency of a household), optimising time efficiency (e.g., improving the public transport timetables and routes), encouraging public sector savings (e.g., lowering translation expenditure) and even saving lives (e.g., improving the response to health emergencies). Furthermore, up to €26 billion could be saved by reducing traffic congestion (Goodwin [2004]), up to 25% of the final energy consumption in the developed world could be saved by improving energy efficiency in buildings (Pérez-Lombard et al. [2008]), and deaths could be prevented by up to 4.5% through the provision of first aid by bystanders (Ashour et al. [2007]).

Some countries with notable open data portals are:

- United States (https://www.data.gov).
- Czech Republic (https://data.gov.cz/english).
- Australia (https://data.gov.au).
- Canada (http://open.canada.ca/en).
- United Kingdom (https://data.gov.uk).
- Taiwan(https://data.gov.tw).
- Indonesia (https://data.go.id).
- Mexico (http://datos.gob.mx).
- France (https://www.data.gouv.fr/en).

All these initiatives have shown that there may be many possible benefits of open government data for public sector bodies (Kucera and Chlapek [2014]), such as:

- Increased transparency.
- Improved public relations and attitudes towards government.
- Increased reputation of a public sector body.
- Transparent way of informing the general public about the infringement of legislation.
- Improved government services.
- Improved government data and processes.
- Better understanding and management of data within public sector bodies.
- Supporting reuse.
- Increasing value of the data.
- Stimulating economic growth.
- Minimizing errors when working with government data.
- Easier translations.
- Lower number of requests for data.

Another advantage of open data is that a large amount of data may be made available rapidly. During the COVID-19 pandemic, lockdowns were put into effect by different organisations all around the world, due to the lack of knowledge regarding the new virus. Different variables (such as seasonal behaviour, regional mortality rate and the effectiveness of government measures) were then identified and analysed by researchers, easing the path towards a better understanding of the situation (Alamo et al. [2020]). Open data on COVID-19 variables helped scientists find a solution much more rapidly than they would have if not such a large amount of data were available.

2.4.2 Vertical Markets

Vertical Markets are a powerful approach for the classification of the different aspects of SCs and Smart Territories, providing insights on the strong and weak points of the infrastructure. This and other forms of clustering the available information enable a city to prioritise certain services that can be directly perceived by its citizens; those that improve the living standards in the city (Tay et al. [2018]).

This segmentation could ease the development of projects in several sectors: The public sector could use it as a positive incentive for innovative projects, and the private sector could either participate in the results of open bidding processes from the previous case, or ask for funding for their own projects with tangible proposals and predicted revenues. Table 2.1 shows the proposed clustering of SC components and their classification into vertical markets.

Tab. 2.1: Vertical markets of Smart Cities and their domains

Vertical markets	Domain
Smart Economy	• Innovation • Productivity • Entrepreneurship • Flexible Labour Market
Smart Mobility	• Connected Public transport • Multimodality • Logistics • Accessibility
Smart Environment	• Environmental Protection • Resource Management • Energy Efficiency
Smart People	• Digital Education • Creativity • Inclusive Society
Smart Living	• Tourism • Security • Healthcare • Culture

Vertical markets can contribute to the creation of the cities of tomorrow and to enhance all aspects of everyday life in a balanced manner. An overview of their principles, use cases and potential is shown in Fig. 2.8.

2.5 Conclusions

Despite several successful applications in industrial and commercial applications, ML remains a young field with many research opportunities available. New data capture possibilities have boosted big data methods, and therefore deep learning have become a predominant field with great potential in the last few years. In particular, camera-based applications have enabled new medical diagnosis and treatment practises, made self-driving cars a viable commercial product and have lowered the human resources need for several classification problems.

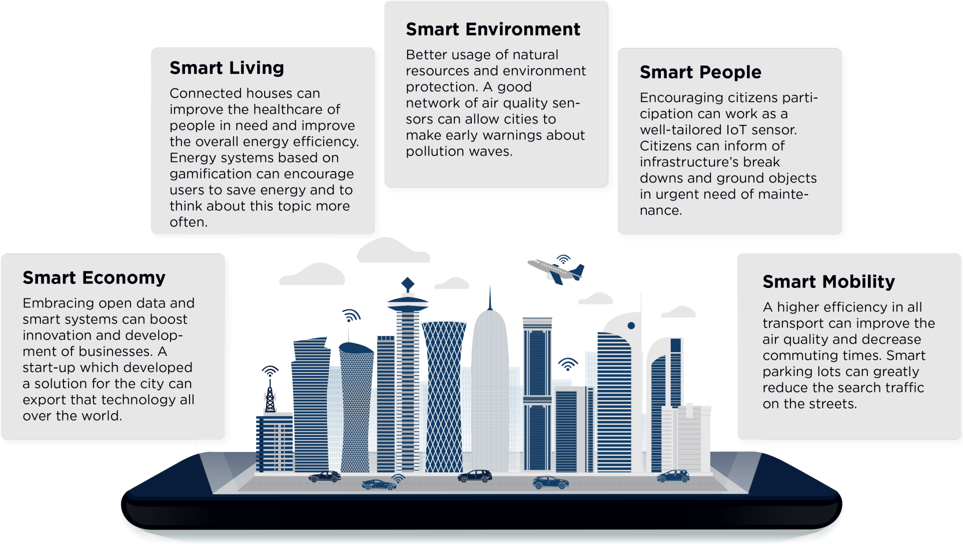
Smart Living
Connected houses can improve the healthcare of people in need and improve the overall energy efficiency. Energy systems based on gamification can encourage users to save energy and to think about this topic more often.
Smart Environment
Better usage of natural resources and environment protection. A good network of air quality sensors can allow cities to make early warnings about pollution waves.
Smart People
Encouraging citizens participation can work as a well-tailored IoT sensor. Citizens can inform of infrastructure's break downs and ground objects in urgent need of maintenance.
Smart Economy
Embracing open data and smart systems can boost innovation and development of businesses. A start-up which developed a solution for the city can export that technology all over the world.
Smart Mobility
A higher efficiency in all transport can improve the air quality and decrease commuting times. Smart parking lots can greatly reduce the search traffic on the streets.

FIG. 2.8: SC vertical markets. Their principles, use cases and potential

Moreover, new trends in technologically modern urban areas have greatly capture the former advances of the last decade. SCs have embraced IoT sensors to gather data from different sources, brought their information in real time to users through mobile applications, automated several services using ML and favoured economic development through information technology and open data.

As a result, this Ph.D. dissertation focuses in identified research gap in ML within the SC scope. Particularly, vision-based applications are our aim since hardware advances have recently made their processing possible, and many new data sources have become available recently.

The vertical markets of SCs are used to maximise the impact of the present research, optimising the gains to the citizens and improving their life quality. One research has been carried out per vertical market, although it is important to notice that some research opportunities lay in the intersection of two or more vertical markets.

Chapter 3

Smart Mobility

Chapter 3

Smart Mobility

3.1 Introduction

Smart Mobility strives towards a future that is defined by cleanliness, safety, efficiency and connectedness. Multimodality is one of its key components, which plays a crucial role in reshaping the current urban mobility patterns (Pop and Proștean [2018]).

The integration and improvement of traffic management, public transport, logistics and ICT (Information & Communications Technology) infrastructure will result in fresher air, more mobility alternatives and lower traffic in cities. Traffic jams will become less common in metropolitan areas, and missing a train due to unexpected situations will become a thing of the past.

Some key principles followed by successful smart mobility project are (Moscholidou and Pangbourne [2020]): reducing the number of journeys taken by car; economic, social or digital inclusion; favouring cleaner technologies; working towards safer and quieter streets; sharing mobility information with the city to enable the monitoring and planning of the transport network.

Regarding traffic management, Smart Territories could use modern technologies to produce real-world changes. Several examples can be found below, as well as a link to one or more data sources related to each case:

- Cars circling for parking are considered to account for approximately 30% of total traffic in cities (Shoup [2018]). Detecting congested areas and providing this information to drivers could reduce noise and air pollution in cities.
  (of Seattle [2020]) dataset contains the 2020 paid parking occupancy in Seattle (USA). It is of a large size (21.5GB). It contains GPS coordinates. Smaller datasets and older datasets are also available on the same webpage.

- Strategic roads could be managed more optimally by predicting near-future traffic levels on the basis of historical and real-time data. Travellers could receive this information in real-time to optimise their route.
  The collection of webpages gathered in (GraphHopper [2022]) contain traffic information with data which can be used free of charge.

- Analysing traffic accidents to reduce future casualties.
  (Moosavi [2021]) is a dataset of countrywide traffic accidents in the United States. It has around 3 million records. From February 2016 to January 2021.

- Safety could be boosted by modeling past accidents and subsequently extracting information about their causes.
  (Council [2020]) dataset contains information on the accidents that occurred in Leeds (UK) from 2009 to 2019).

- Traffic restrictions, especially for private vehicles, are becoming a new trend in many European cities. To analyse how they would affect the city traffic, past temporary restrictions could be analysed and checked against the traffic volume at different times.
  Several datasets related to traffic in Brisbane, Australia (Council [2017]). The first dataset contains the temporal closures information, the second dataset has data on traffic volume at intersections, while the third one provides information on the traffic volume on some roads. There are no GPS coordinates, however, they can be inferred from the webpage. All data is not simultaneously available and must be gathered. The first dataset contains only future data, the second one only data from the last day, and the third one monthly data since July 2019.

- Walking became an activity people longed for during the COVID-19 pandemic. To ensure the pedestrians' safety, a crowd detection system could be implemented and recommendations could be made to citizens regarding the less crowded areas (Garcia-Retuerta et al. [2021]).
  (of Melbourne [2022]) dataset contains information on 75 pedestrian counting points in Melbourne (Australia). It counts the number of pedestrians on an hourly basis. It also provides the hourly average of the last 4 and 52 weeks for each sensor. Data are available from July 2009 onwards.

- Accurate location data can improve emergency response times, as well as allow for a better spatial analyses and development of under-utilized areas.
  (Française [2022]) dataset contains an accurate list of addresses of the whole country of France, including the GPS coordinates of streets, ZIP code and land registry code. It is regularly updated. Data are divided in over 200 csv files of

different sizes (85MB maximum). The webpage is only available in French. It can be used to improve current maps, as the inputs also include manually checked registers.

3.2 Proposed Method: Low Computational Cost Transformers

We live in the age of data. As IoT and AI become more and more ubiquitous technologies, ever-larger amounts of data are being generated. Information can be stored in the form of audio, text, images, or a mixture of the previous. Images are responsible for a big part of global data creation, as photos and videos are becoming omnipresent in our society.

The computer vision field has several open problems of great importance, namely: image classification, image segmentation, object detection, anomaly detection and activity recognition. In particular, image classification stands out as the most important challenge, as it is the foundation for solving other vision problems. It is used in many real-world applications, such as medical imaging, machine vision, object identification in satellite images, autonomous driving, etc.

The method proposed in this chapter considers one of the most important deep learning architectures, Transformers, and its behaviour in NLP and computer vision. An alteration to the well-establish architecture is put forward. The main goal of the modification was to lower the computational cost of the model while preserving its accuracy, however, results indicate that it actually results in a slight accuracy increase on average.

An obvious application of the findings would be to boost driverless car technologies, as less powerful (cheaper) hardware would be required for their manufacturing, while reaction times are lowered (due to a more lightweight processing). Road safely would therefore be improved.

3.2.1 Related Works

This section provides a presentation of the techniques used in the proposed method, namely: use of transformers in computer vision, with particular attention to their self-attention mechanism and complexity; and the Vision Transformer architecture.

Transformers (Vaswani et al. [2017]) appeared in 2017 as a new, scalable architecture capable of state-of-the-art results in NLP (Lan et al. [2019], Liu et al. [2019], Yang et al. [2019]). What makes this model unique is that its performance, even with billions of parameters and billions of training examples, does not saturate. Hence, a common result in Transformer-based methods is that the deeper/larger the model, the more accurate the results are. Transformers were already introduced in Section 2.3.2, and they are further described in the current section.

Due to their promising results, researchers have also applied them to other fields, such as computer vision, where they are beginning to replace CNNs in some state-of-the-art applications (Carion et al. [2020], Doersch et al. [2020], Girdhar et al. [2019], Liu et al. [2021c], Su et al. [2019], Ye et al. [2019]).

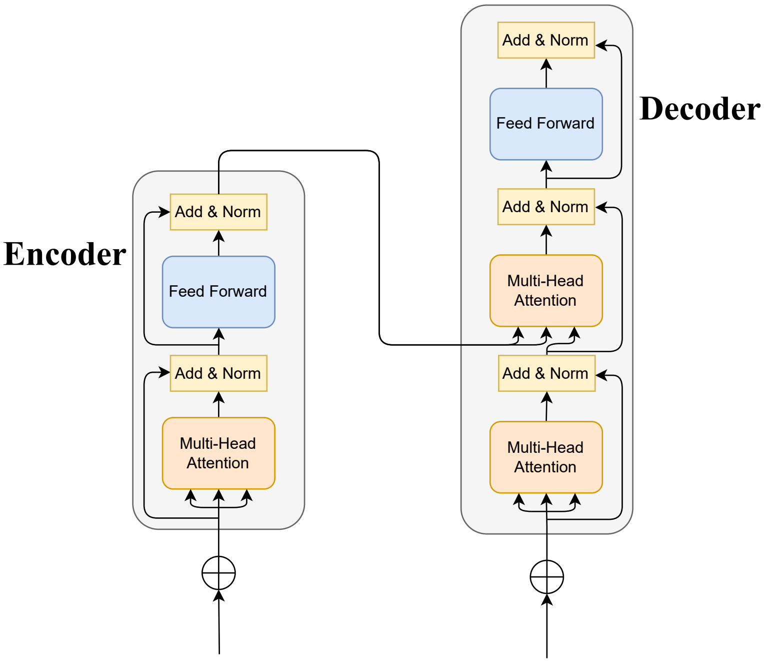

FIG. 3.1: The transformer — illustration of its overall structure

Transformers differentiate from previous architectures by: (1) not necessarily processing data in order, which allows them to compute the representations for each individual token in parallel, and (2) identifying the context of each token in the multi-head attention blocks, which aggregates spatial information across tokens. Moreover, their attention mechanism introduces the inductive bias that the spatial interactions should be dynamically parameterized based on the input representations (Liu et al. [2021a]). They are divided in *encoder* block and *decoder* block, which basic components can be found in Fig. 3.1.

Self-attention is one of its most critical qualities: in the original NLP approach, it allows the model to associate words from another part of the text to the currently processing word (e.g., in the sentence “the ball was blue and it was hot”, it associates the word “ball” to the word “it”). Therefore, the model is capable of understanding the most relevant words related to the one considered.

Self-attention linearly projects each word embedding in the input text into three different feature spaces, commonly called: *query, key* and *value.* First, query and key are processed together, their relation is modeled and then, the resulting features are multiplied by the value’s features. Each of the three spaces is preceded by a linear projection, which can therefore adapt its weights to better capture their relations. Fig. 3.2 shows the parts of the Transformer responsible for the self-attention.

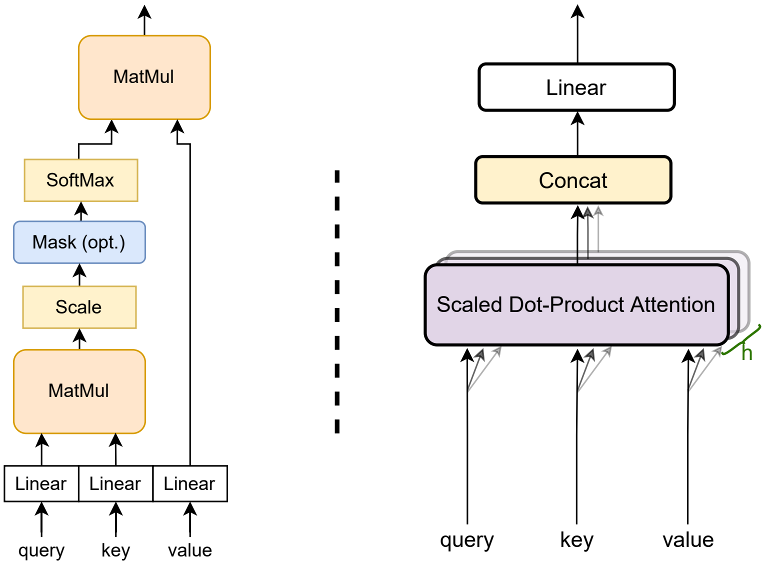

FIG. 3.2: Illustration for the self-attention-related blocks of the Transformer: Scaled Dot-Product Attention (left) and Multi-Head Attention (right)

Transformers make use of positional encoding to provide information about the relative position of the words in the sentence to the model. The positional encoding vector is added to the embedding vector.

The positional encoding vector is added to the embedding vector. Embeddings represent a token in a d-dimensional space in which tokens with similar meaning are closer together. However, embeddings do not encode the relative position of words in a sentence. Therefore, after adding positional encoding, words are closer to each other, in accordance with the similarity of their meaning and their position in the sentence in d-dimensional space. Fig. 3.3 represents the similarity of a word to another in a position x-units away, depending on the depth of the model.

The formulas for calculating the positional encoding are:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{3.1}$$

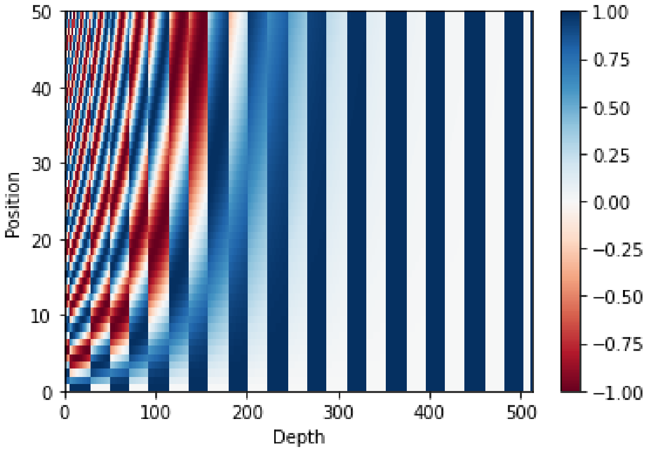

FIG. 3.3: Similarity of a word to another in a position x-units away, depending on the depth of the model

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{3.2}$$

Regarding the **Transformer complexity**, it is important to note that its structure allows for good parallelization while employing self-attention. The other common self-attention method, recurrent neural networks, requires $O(n)$ sequential operations per layer, while Transformer's layers connect all positions with a constant number of sequentially executed operations, $O(1)$. In particular, Transformer's layers are faster than recurrent layers when the sequence length n is less than the representation dimensionality d, a very common scenario in machine translations.

Compared to CNNs, a stack of O(n/k) convolutional layers is required to connect all pairs of inputs and outputs, when kernel width $k < n$ and contiguous kernels are used. As a result, its layers are generally k-times computationally more expensive. Separable convolutions and dilated convolutions are know to decrease the complexity to $O(k \cdot n \cdot d + n \cdot d^2)$ and $O(log_k(n))$, respectively.

A comparison of the complexity Transformer's self-attention layers against recurrent and convolutional layers is provided in Table 3.1.

TAB. 3.1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention

Layer type	Complexity per layer	Sequential operations	Max. path length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$

Moreover, the self-attention blocks of Transformers increase the interpretability of the model, as individual attention heads exhibit behaviours related to the grammatical, syntactic and semantic structure of the sentences.

There are some real-life applications of Transformers already, such as the Tesla Autopilot which uses a Transformer-based architecture on the multi-camera system of the electric cars of the brand to achieve its self-driving capabilities.

Vision Transformer (ViT) is a newer iteration of Transformers which modifies the architecture to optimise it in vision-related tasks (Dosovitskiy et al. [2020]). The model is designed for image classification based on a Transformer-like architecture which is applied over patches of the image (Fig. 3.4). The process for feeding an image into a Transformer is as follows: An image is split into patches of a previously fixed size, then each patch is linearly embedded, position embeddings are added, and the resulting sequence of vectors is used as the input for a standard Transformer encoder.

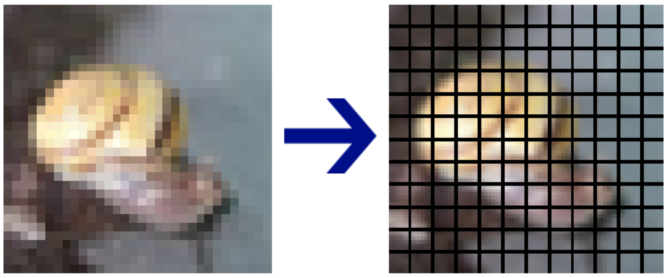

FIG. 3.4: Patches obtained from a CIFAR-100 image

Its most important results are obtained with datasets of more than 14M images, where it surpasses state-of-the-art CNNs (Han et al. [2020]). Subsequently, it has become commonplace to pre-train the ViT on a large dataset and then fine-tune it with a small dataset.

Swim transformer (Liu et al. [2021c]) is the most promising ViT-based architecture at the moment, as it has obtained state-of-the-art results on some demanding object detection datasets such as COCO. It modified the attention mechanism and developed a new multi-stage approach.

3.2.2 Materials and Methods

This section presents the used datasets and the proposed modification of the transformer architecture.

CIFAR-10

The CIFAR-10 (Canadian Institute For Advanced Research-10) dataset is a collection of images commonly used for training ML and computer vision algorithms. It is one of the most commonly used datasets for ML research. It contains 60,000 color images, 32x32 pixels, in 10 different classes: airplanes, cars, birds, cats, deer, dogs, frogs, toads, horses, ships, and trucks. It is split with 50,000 training images and 10,000 test images, where there are 6,000 images in each class in total. Classes are completely mutually exclusive, as there is no overlapping between them.

CIFAR-10 is a labeled subset of the “80 million tiny images” dataset. It was created by remunerated students who labelled the images, and it is protected under the MIT License. Fig. 3.5 shows an example of some of its images.

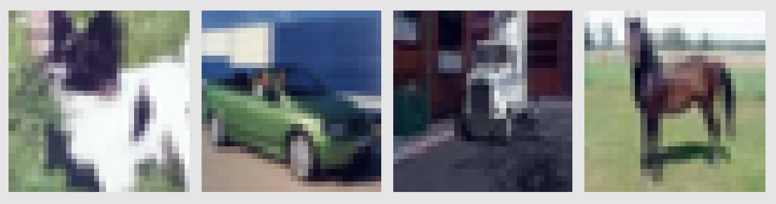

Fig. 3.5: Four images from the CIFAR-10 dataset belonging to the classes: cat, car, truck and horse (respectively)

CIFAR-100

The CIFAR-100 (Canadian Institute For Advanced Research-100) dataset is a collection of images commonly used for training ML and computer vision algorithms. It is similar to CIFAR-10, but it contains 100 classes with 600 images each. It contains 60,000 32x32 color images in total too, and their split consists of 500 training images and 100 testing images per class. The 100 classes in the CIFAR-100 are grouped into 20 superclasses, therefore each image has a “fine” label (the class to which it belongs) and a “coarse” label (the superclass to which it belongs).

CIFAR-100 is also a subset of the “80 million tiny images” dataset, created by the same students who created CIFAR-10, and it is also protected under the MIT License. Fig. 3.6 shows an example of its images, grouped for an anomaly detection task.

NLP

The Portuguese-English translation dataset from the TED Talks Open Translation Project was used for the NLP tests. It is derived from TED talk transcripts, where

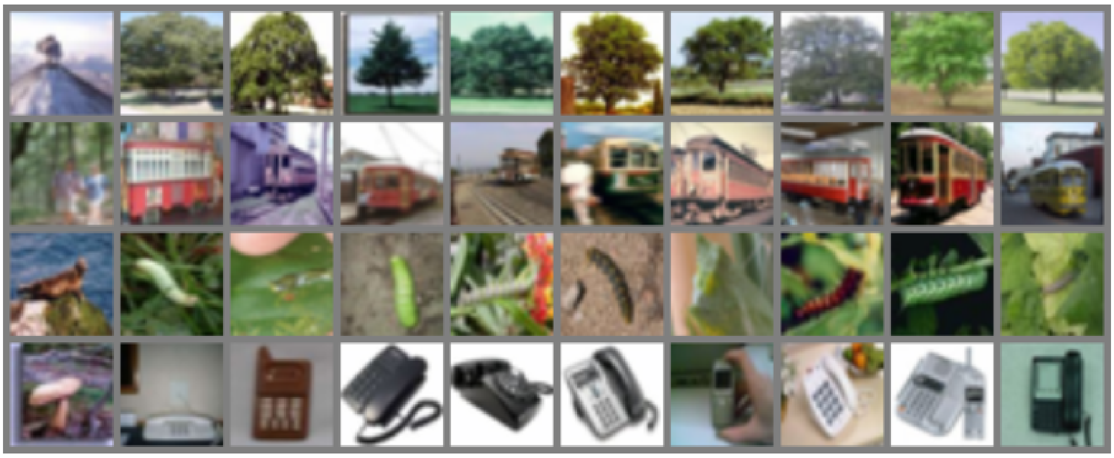

FIG. 3.6: Images extracted from the CIFAR-100 dataset. Each row shows a different input set, where the first 9 are from the same class and the last image belongs to a different class (the anomaly)

one language is high resource and the other language is low resource. It is used for comparing similar language pairs. It contains around 50,000 training examples, 1,100 validation examples, and 2,000 test examples.

Do (not) Value Attention

Narang et al. [2021] state in their paper: "most of the Transformer variants we found beneficial were either developed in the same codebase that we used or are relatively minor changes". The Transformer modifications they analysed consisted mainly of altering the activation function, embedding, normalization, depth, parameter sharing, the softmax function, or the Transformer architecture itself. They propose a series of tasks meant to ensure the effectiveness of new modifications. We performed our research to comply with them. We propose a relatively minor change to the Transformer architecture but with major implications. This modification is capable of lowering the computational cost of training in all our experiments related to computer vision. The alteration consists of removing one of the key components of the self-attention block: the third dimension in which the input is split, commonly called *value*.

We define the matrix of outputs of the Scaled Dot-Product of layer l is defined as:

$$\text{Attention}_l(Q_l, K_l) = \text{softmax}(\frac{Q_l \cdot K_l^T}{\sqrt{d_k}}) \cdot K_l \tag{3.3}$$

where d_k is the dimension of K, and Q, K are the so-called *query* and *key*, respectively.

Then, the previous block is used in the multi-Head attention. The output of a multi-Head attention head in layer l is defined as:

$$s_l(Q_l, K_l) = \text{LayerNorm}([a_1(Q_l, K_l), ...a_h(Q_l, K_l)] \cdot W^O + p_{l-1}(Q_{l-1}, K_{l-1})) \tag{3.4}$$

where $a_i = \text{Attention}_l(Q_l W_{i,l}^Q, K_l W_{i,l}^K)\ \forall i$, and $W_{i,l}^y$ represents the projection matrix of head i, in layer l respect to the dimensionality y_l in layer l. Also, $W^O \in \mathbb{R}^{d_{model} \times h \cdot d_{K_l}}$ and $p_{l-1}(Q_{l-1}, K_{l-1})$ is the output of the previous multi-Head attention block. h is the total number of heads.

Afterwards, a projection is applied to the output, then a RelU nonlinearity and another projection:

$$f_l(Q_l, K_l) = \max(0,\ s_l(Q_l, K_l) W_{l,1}^O + b_{l,1}) \cdot W_{l,2}^O + b_{l,2} \tag{3.5}$$

where $W_{l,1}^O \in \mathbb{R}^{d_{model} \times d_f}$ and $W_{l,2}^O \in \mathbb{R}^{d_f \times d_{model}}$ are the projection matrices and $b_{l,1} \in \mathbb{R}^{d_f}$, $b_{l,2} \in \mathbb{R}^{d_{model}}$ are the biases. d_f is the hidden dimension.

Lastly, a residual connection is applied and a layer normalization, resulting in:

$$p_l(Q_l, K_l) = \text{LayerNorm}(s_l(Q_l, K_l) + f_l(Q_l, K_l)) \tag{3.6}$$

Fig. 3.7 shows a high level representation of the proposed attention structure.

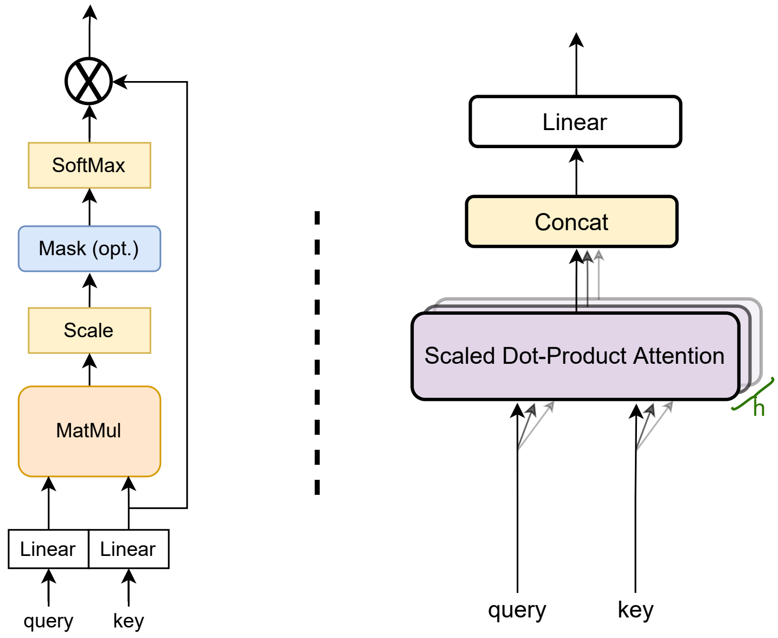

FIG. 3.7: Proposed scaled dot product (left) and multi-head attention (right)

3.2.3 Experimental Results and Discussion

Experiments were performed following the strict recommendations of Narang et al. [2021] for analysing the effects of modifications to Transformers. Robust alterations are wanted, capable of improving the models across different implementations and use cases, and which could set the base for a future iteration of the Transformers. All the code and results are available at: https://github.com/dvidgar/transformer_light

Two machines were available for the experiments of the current chapter. They have the following specifications:

1. GPU NVIDIA Tesla K80 with 24 GB GDDR5, 13GB RAM and two cores of a Intel Xeon CPU of 2.20GHz.
2. GPU NVIDIA GeForce RTX 3090 with 24 GB GDDR6X, 131GB RAM and a Intel i9-10940X CPU.

PyTorch: Vision Transformer

CIFAR-100 The CIFAR-100 dataset is used, in combination with the PyTorch library, for image classification. The model is based upon the Vision Transformer (described in Section 3.2.1).

The classification head of the ViT is implemented to minimise the Cross-Entropy loss, and GELU is used as the activation function. The feed-forward blocks are implemented using the following structure: Multi Layer Perceptron (MLP)⟶GELU⟶Dropout⟶MLP⟶Output, and 12 Transformer blocks are used in total.

Our modified architecture is found to improve all aspects of the model: **test accuracy is increased, train accuracy decreased, training times reduced and the number of trainable parameters shrinks**. Table 3.2 shows all the performance values. Machine 2 was used for these experiments.

Tab. 3.2: Results for the Vision Transformer on the CIFAR-100 dataset. Both training periods had the same duration

CIFAR-100

	original	modified	diff. (%)
test accuracy (%)	**41.7**	**42.6**	**+2.15**
train accuracy (%)	84.6	82.5	-2.09
parameters (M)	**16.0**	**12.9**	**-19.7**
epochs	90	100	+11.1

Similar tests were also carried out in a less-powerful environment (Machine 1). In this case, the modified architecture shows a 33.03% faster training time per epoch, considerably better than in the previous case. Thus, the proposed method especially benefits low-resource machines thanks to its lower computational cost.

CIFAR-10 The CIFAR-10 dataset is used, in combination with the lower level features from the pretrained ResNet-34 (He et al. [2016]), for image classification. The PyTorch library was used for coding. The model is based upon the Vision Transformer.

Vision Transformer technical details are described in Section 3.2.1 and the specific configuration used for this experiment is equivalent to the one used for the CIFAR-100 dataset (in the previous section).

Our modified architecture is found to improve all aspects of the model: **accuracy is increased, training times reduced and the number of trainable parameters shrinks**. Table 3.3 shows all the performance values. Machine 2 was used for these experiments.

Tab. 3.3: Results for the Vision Transformer on the CIFAR-10 dataset with the lower level features of ResNet-34. Both training periods had the same duration

CIFAR-10 pretrained

	original	modified	diff. (%)
test accuracy (%)	**63.7**	**65.3**	**+2.62**
train accuracy (%)	75.3	76.3	+1.38
epochs	**92**	**100**	**+8.69**
parameters (M)	16.5	13.4	-19.1

PyTorch: Transformaly

Transformaly Cohen and Avidan [2021] is a ViT-based architecture which currently achieves the best AUROC (Area Under the Receiver Operating Characteristic) results of the state of the art in anomaly detection on CIFAR-10 and CIFAR-100, in both the common unimodal setting and the multimodal setting. It works by using two independent feature spaces: a teacher-student network where the student is only trained on normal examples, and a pre-trained feature extractor. Both cases use a pre-trained ViT network as their backbone.

In our experiments, we use the CIFAR-10 dataset in the common unimodal setting. Since no pre-trained ViT with the proposed alteration can be found, our results refer only to networks trained from scratch. Therefore, they considerably differ from the paper outstanding results.

Our proposed alteration result has a **much higher average AUROC performance of the model, a faster training time per epoch and a reduced number of trainable parameters**. Table 3.4 shows all the performance values. Machine 2 was used for these experiments.

Tab. 3.4: Results for the Transformaly model trained from scratch on the CIFAR-10 dataset. Both trainings lasted 50 epochs

CIFAR-10 unimodal

	original	modified	diff. (%)
AUROC class 0 (%)	**68.2**	**73.2**	**+7.4**
class 1 (%)	62.7	78.4	+25
class 2 (%)	**64.4**	**67.1**	**+4.1**
class 3 (%)	62.6	65.2	+4.2
class 4 (%)	**70.3**	**71.6**	**+1.9**
class 5 (%)	59.1	62.9	+6.5
class 6 (%)	**79**	**77.9**	**-1.4**
class 7 (%)	56.6	64.4	+13.7
class 8 (%)	**64.4**	**75.6**	**+17.4**
parameters (M)	257.2	236	-8.3
train time (s/epoch)	**216**	**211**	**-2.3**

When both variants had the same training time, it resulted in 49 vs 50 epochs (original vs proposed architecture, respectively). In that case there is an even higher average AUROC: 11.23%.

PyTorch Lightning: Vision Transformer

The CIFAR-10 dataset is used, in combination with the lower level features from the pretrained ResNet-34 (He et al. [2016]), for image classification. The PyTorch Lightning library was used for coding. The model is based upon the Vision Transformer.

Vision transformer technical details are described in Section 3.2.1. In this case, an image of size $N \times N$ is split into $(N/M)^2$ patches of size $M \times M$. The patches represent the input words to the Transformer.

The pre-layer normalization variant of the Transformer blocks from Xiong et al. [2020] is used. It consists in applying a Layer Normalization as a first layer in the residual blocks, instead of applying it in between residual blocks as the original Transformer does. This results in a better gradient flow and removes the necessity of a warm-up stage.

Our modified architecture is found to improve all aspects of the model: **accuracy is increased, training times reduced and the number of trainable parameters shrinks**. Table 3.5 shows all the performance values.

Tab. 3.5: Results for the Vision Transformer on the CIFAR-10 dataset. Both trainings lasted 180 epochs

CIFAR-10

	original	modified	diff. (%)
test accuracy (%)	**76.0**	**76.3**	**+0.36**
train accuracy (%)	76.9	76.9	+0.08
train time (s/epoch)	**82**	**78**	**-4.88**
parameters (M)	3.2	2.8	-12.5

Note that both networks trained for the same number of epochs, so the original Transformer had extra training time compared to the modified version. Having the same training epochs is a much more challenging task than having the same training time. Namely, in this case the modified Transformer was capable of completing 180 epochs while the original Transformer could only have done 169 epochs.

Machine 1 was used for these experiments. Preliminary tests on more powerful hardware (Machine 2) show a lower training time gain, probably due to the ML-optimised capabilities of new GPUs (notably: mixed precision).

PyTorch Lightning: Transformer Encoder for Anomaly Detection

The CIFAR-100 dataset is used, in combination with the lower level features from the pretrained ResNet-34 (He et al. [2016]). The PyTorch Lightning library was used for coding. The model is based upon a Transformer encoder as shown in Fig. 3.1, with some modifications so that it performs visual anomaly detection. An example of the task is shown in Fig. 3.8.

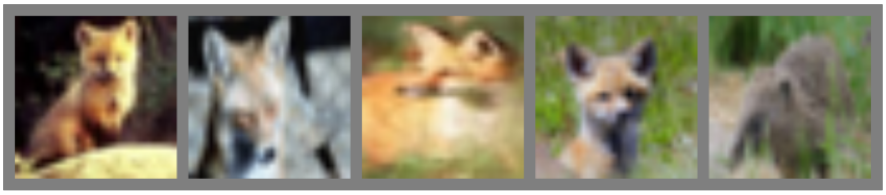

Fig. 3.8: Anomaly detection task based on CIFAR-100. First four images of the batch showing foxes, and the last representing a different animal

The used model is formed by an input network which maps the input into the model dimension (MLP + dropout), an input network which applies the positional encoding of Equation 3.7, the transformer architecture and an output network which maps the output encodings to the predictions dimensions (equal to the number of classes, formed by a MLP + layer normalization + ReLU activation + dropout + MLP). The positional encoding over several different hidden dimensions is shown in Fig. 3.9.

$$PE_{(pos,i)} = \begin{cases} \sin\left(\frac{pos}{10000^{i/d_{\text{model}}}}\right) & \text{if } \ i \bmod 2 = 0 \\ \cos\left(\frac{pos}{10000^{(i-1)/d_{\text{model}}}}\right) & \text{otherwise} \end{cases} \tag{3.7}$$

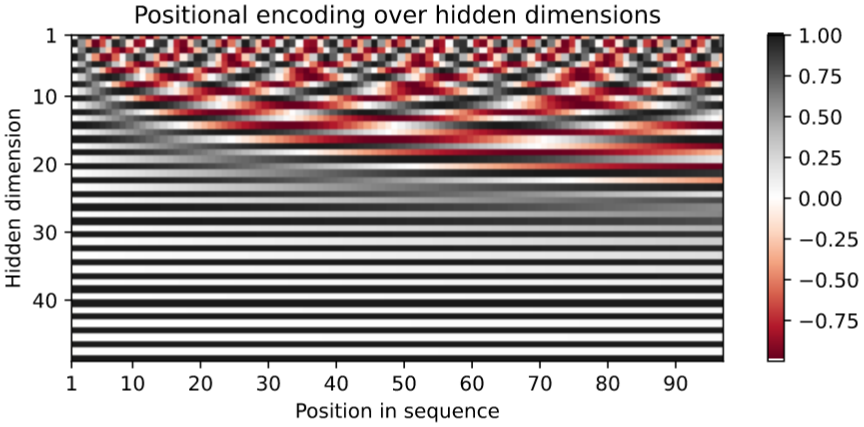

Fig. 3.9: Each pixel, therefore, represents the change of the input feature we perform to encode the specific position

The modified architecture is found to improve all aspects of the model: **accuracy was increased, training times reduced and the number of trainable parameters shrank**. Table 3.6 shows all the performance values.

Tab. 3.6: Results for the Vision Transformer on the CIFAR-100 dataset with the lower level features of ResNet-34. Both trainings lasted 180 epochs

CIFAR-100 pretrained

	original	modified	diff. (%)
train accuracy (%)	**98.1**	**98.7**	**+0.61**
test accuracy (%)	94.8	95.4	+0.63
val. accuracy (%)	**94.3**	**94.6**	**+0.32**
train time (s/epoch)	36	26	-27.8
parameters (M)	**2.3**	**2**	**-13.0**

Note that both networks trained for the same number of epochs, so the original Transformer had extra training time compared to the modified version. Having the same training epochs is a much more challenging task than having the same training time. Theoretically, if the same training time was allowed, the original Transformer could only reach 130 epochs, and the modified Transformer 180 epochs.

Machine 1 was used for these experiments. Preliminary tests on more powerful hardware show a lower training time gain, probably due to the ML-optimised capabilities of new GPUs (notably: mixed precision).

Regarding interpretability, the output of the different attention heads can be analysed and plotted similarly to the case in NLP tasks. However, in this case the results do not seem to provide much information (Fig. 3.10).

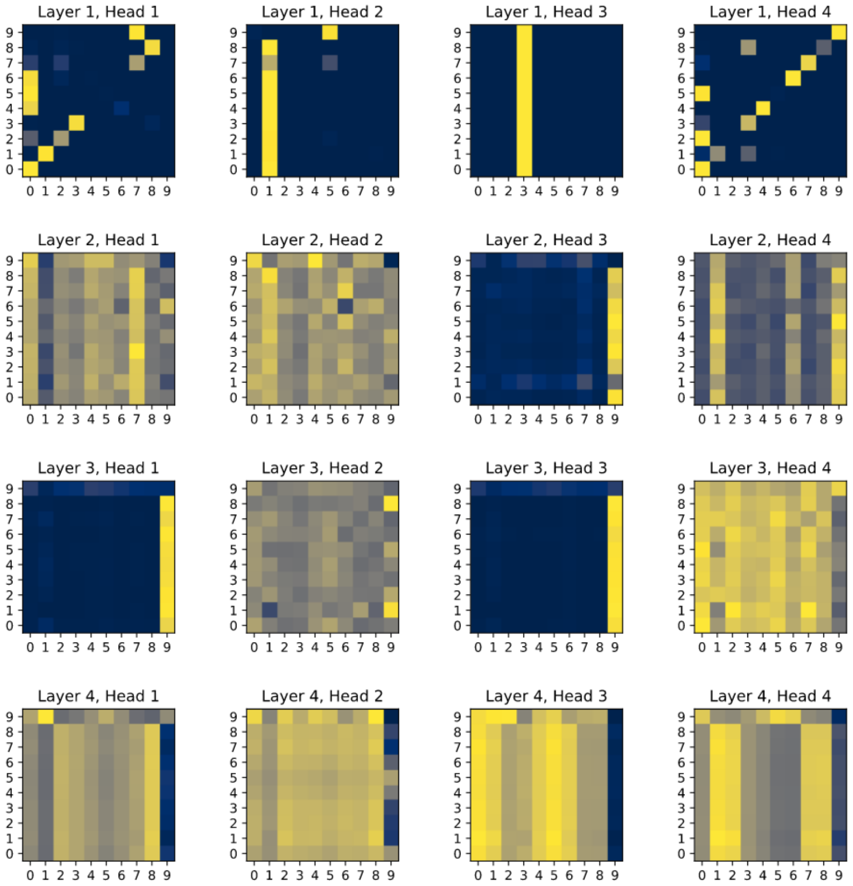

FIG. 3.10: Output of the different attention heads of the anomaly detection model

TensorFlow: Vision Transformer

The CIFAR-100 dataset was used used in these experiments, in combination with the TensorFlow library for coding. The model is based upon the Vision Transformer and used for image classification.

Vision transformer technical details are described in Section 3.2.1. In this case, data augmentation is also implemented in order to train the model from scratch successfully. The used data augmentation techniques are: horizontal random flips, random rotations and random zooming.

The modified architecture is found to improve all aspects of the model: **accuracy was increased, training times per epoch reduced and the number of trainable parameters slightly shrank**. Both trainings were carried out for the same amount of time. Table 3.7 shows all the performance values. Machine 2 was used for these experiments.

Tab. 3.7: Results for the Vision Transformer on the CIFAR-100 dataset. Both trainings had the same duration

CIFAR-100 pretrained

	original	modified	diff. (%)
train accuracy (%)	**66.25**	**66.56**	**+0.46**
test accuracy (%)	52.74	53.83	+2.07
test top 5 accuracy (%)	**80.9**	**80.96**	**+0.07**
epochs	70	82	+17.1
parameters (M)	**21.8**	**21.6**	**-0.92**

Note that the number of trainable parameters is slightly reduced due to the last 3 MLP layers. They contain 18.88, 2.1 and 0.1 million parameters respectively. If only the multihead attention block is considered, the original architecture has 66.368 parameters and the proposed version only 49.728 — a 25.1% decrease in the number of parameters.

NLP

The dataset described in Section 3.2.2 was used as input, and TensorFlow was the library in which the models were programmed. Machine 1 was used for these experiments.

The accuracy of the original Transformer was 0.4010, needing 41 seconds for each training epoch and having 4,981,913 parameters.

The accuracy of the proposed modified Transformer was 0.3916, needing 39 seconds for each training epoch and having 4,783,769 parameters.

Therefore, the new architecture results in a **2.34% accuracy drop**, while also decreasing the training time a 4.88% and the number of parameters by 3.98%. This result seems to indicate why the original architecture included the *values* space: it is a necessary component for determining the words relations in natural language processing tasks.

A few examples of the performed translations are:

- este é um problema que temos que resolver. (input)

  this is a problem we have to solve. (real translation)

  this is a problem that we have to solve it. (output original transformer)

  this is a problem that we have to fix. (output modified transformer)

- os meus vizinhos ouviram sobre esta ideia. (input)

  and my neighboring homes heard about this idea. (real translation)

  my neighbors heard about this idea of this idea. (output original transformer)

  my neighbors heard about this idea. (output modified transformer)

- vou então muito rapidamente partilhar convosco algumas histórias de algumas coisas mágicas que aconteceram. (input)

  so i ’ll just share with you some stories very quickly of some magical things that have happened. (real translation)

  so i ’m very quickly going to share with you some magic stories that had happening. (output original transformer)

  so i ’m going to share with you a few stories of some magic things that had happened. (output modified transformer)

Regarding the interpretability, Fig. 3.11 shows the output of different layers and attention blocks of the decoder of the modified Transformer. As it can be seen, it captures semantic relations among words, while paying most of its attention to the central words of the sentence (the most important ones). A similar behaviour to the original Transformer. Therefore, we conclude that it maintains this positive quality of the original architecture.

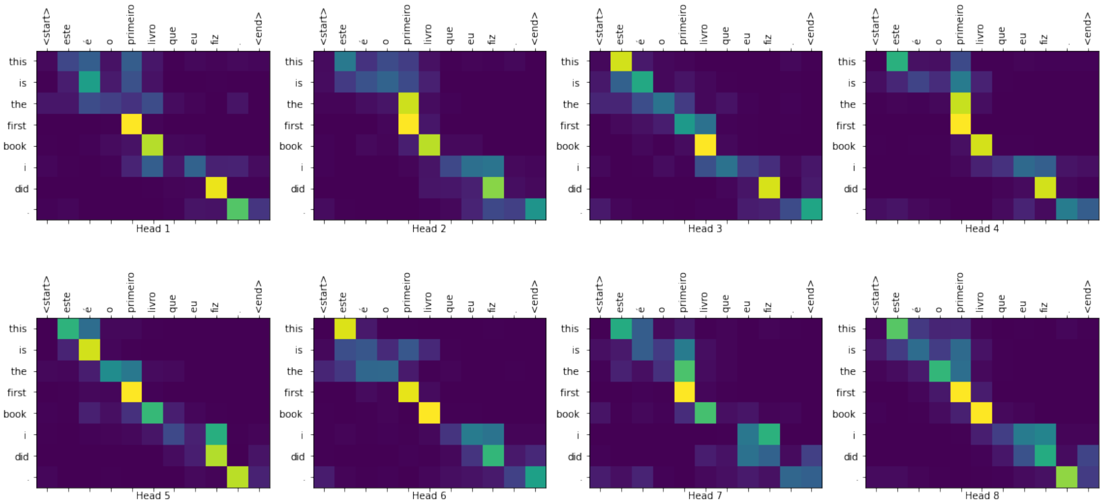

Fig. 3.11: Output of different layers and attention blocks of the decoder of the modified Transformer

3.2.4 Conclusions and Future Work

We have explored the consequences of modifying a key part of the Transformer architecture, obtaining promising results for equal training times. Accuracy gain ranged between 0.36 – 8.76% in the experiments, training was accelerated by 2.32 – 27.8% and the number of trainable parameters decreased by 10.81% on average. Less powerful machines seem to specially benefit from the proposed alteration due to its lower computational cost. However, the modification seems not to be suitable for NLP tasks, which may explain why the original wide-spread Transformer architecture does not already include it.

Challenging tests were performed to verify the robustness of the proposed modification: several implementations on different frameworks were used, hyper-parameters were not fine-tuned, the best performing model was not cherry-picked and two different machines (with rather different specifications) were used for the trainings. The proposed alteration showed a superior behaviour all the tests. Results show that the proposed modification is effective, efficient and therefore relevant for the computer vision field.

The Shapiro-Wilk test of the results (excluding Transformaly, as it is clearly an outlier) does not show any significance departure from the normality, with p values of 0.231 and 0.2314 in the case of same time-length training and same epochs-length training, respectively. The considered variable is $Y = \hat{X}_0 - \hat{X}_m$, where $\hat{X}_0$ is the test accuracy of the original network and $\hat{X}_m$ is the test accuracy of the proposed network. Therefore, it can be concluded that the variable Y presents a normal distribution centered in $\overline{x} = 1.566$ with a standard deviation $S = 1.0049$ and a skewness $\mu_3 = -0.4452$, in the case of same time-length training. Asserting that the proposed method increases the accuracy in the general case by $\sim 1.5\%$, however, there have been some variations in different cases.

It also follows that, in the case of same epochs-length training, the variable Y presents a normal distribution centered in $\overline{x} = -0.5948$ with a standard deviation $S = 1.3991$ and a skewness $\mu_3 = -1.5066$. Thus, it is found that the proposed method decreases the accuracy in the general case if the original network has extra training time.

While these initial results are encouraging, many challenges remain. More experiments shall be performed to validate the observed positive behaviour in visual tasks, and modern networks shall be modified accordingly in an attempt to push the state of the art forward. Researchers, companies and institutions with restricted computational resources could greatly benefit from this method, as well as applications which require a frequent network re-training.

Chapter 4

Smart People

Chapter 4

Smart People

4.1 Introduction

Smart People is a vertical market which implies many advances for SC residents. It seeks to evolve the manner in which citizens interact with the public and private sector, as individuals or as businesses. This will result in an increased overall efficiency, as more individuals become aware of the services available in their Smart Territory.

Moreover, an important dimension of ”Smart People” is related to education; using modern technology and learning methods in the classroom, developing students’ technological skills and creating local facilities for lifelong learning. This term also encompasses the residents’ professional life. Smart territories should facilitate career choices, sharing labour market opportunities with the broader public and expanding work flexibility (teleworking, flexible timetable, etc) (Benešová and Tupa [2017]).

Talent deployment is also a crucial aspect of this vertical market. Creative networks should be boosted, creative artists and individuals should be supported and partnerships with creative organizations developed.

As this vertical market fits within the social sciences scope, only scarce and short open data sources are available. The most trustworthy websites for obtaining good datasets in this topic are:

- (Bank [2019]) is a database of the World Bank. It includes complete, worldwide information about education, gender, poverty, labour and social protection, social development, etc.
- (OECD [2021]) is a database of the Organisation for Economic Co-operation and Development. It contains trustworthy, worldwide information regarding demography, population, development, education, training, globalisation, information and communication technology, labour, etc.

4.2 Proposed Method: a Light Pyramid Dilated Attention Network and modern Multi-Layer Perceptron-based Architectures for Human Activity Recognition

Recent advances in HAR have enabled a wide range of applications, including the IoT (Akbari et al. [2018], Alani et al. [2020], Jiang and Yin [2015]), healthcare (Lyu et al. [2017]) and enhanced manufacturing (Gumaei et al. [2019]). Activity recognition is critical to humanity because it records people's behaviours with data that computing systems can use to monitor, analyse, and assist them in their daily lives using input data sources such as sensing devices, including vision sensors and embedded sensors.

The improvement of video surveillance or CCTV technology (Dang et al. [2019]) has resulted in improved video quality, easier setup, lower costs, and secure communication. Hence, an increasing number of applications utilizing CCTV systems for security and monitoring purposes have successfully been applied.

The main domains of HAR include surveillance systems (Jalal et al. [2017], Ji et al. [2018]), gesture recognition (Oyedotun and Khashman [2016], Pigou et al. [2016], Xu et al. [2017]), behaviour analysis (Batchuluun et al. [2017]), patient monitoring systems (Prati et al. [2019]) and a range of healthcare systems (Avilés-Cruz et al. [2019], Qi et al. [2019]). As a result, tracking daily activities is required to provide clinicians with up-to-date reports and inform patients with real-time feedback on their progress. For instance, patients with cognitive decline or mental disorders must be continuously monitored in order to detect unusual behaviour on time and thus avoid unintended negative consequences (Dedabrishvili et al. [2021], Varatharajan et al. [2017]).

However, due to the fact that there is no standard procedure for associating the massive volume of collected data to a specific action, HAR is regarded as a difficult research problem. Other considerable challenges of the study area include: feature extraction, class imbalance, data segmentation, computational cost, and privacy (Chen et al. [2021]).

Video action recognition is the first step of video understanding, a critical component of vision-based HAR and, therefore, it is an active research area in recent years

4.2.1 Related Works

The **two-Stream Inflated 3D ConvNet** (I3D) has become as the *de facto* foundation of most modern HAR algorithms. It is based on the inflated Inception-V1 architecture, which is formed by Inception modules as well as convolutions and max-pooling (Fig. 4.1).

Several HAR proposals load a pre-trained I3D model, extract the features of the original data and, then, use them as inputs for their customised model. This process is believed to capture spatio-temporal information of the video. Piergiovanni and Ryoo [2018] makes use of I3D features and their "super-event" to obtain a 36.4% per-frame-mAP on the Multi-THUMOS dataset, and Mavroudi et al. [2020] make a clever usage of I3D, Bidirectional Gated Recurrent Units (BiGRU) and VS-ST-MPNN to obtain a 23.7% per-frame-mAP on the Charades dataset.

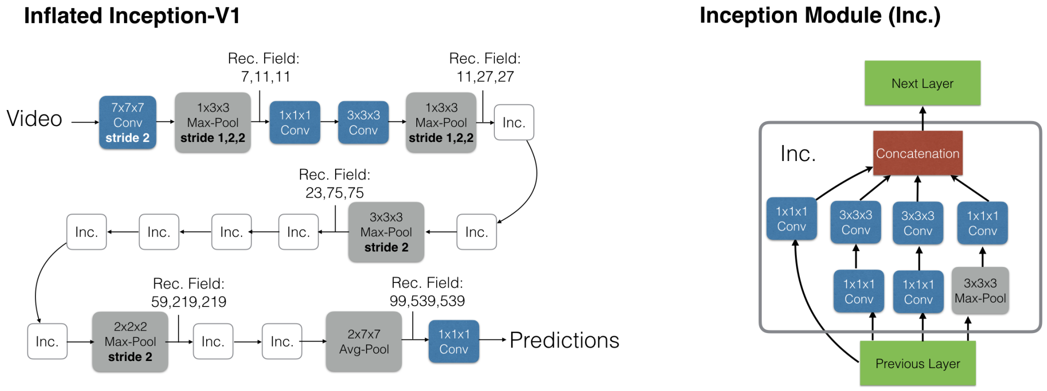

Fig. 4.1: Inflated Inception-V1 architecture (left) and the inception module (right). Extracted from Carreira and Zisserman [2017]

Transformers have gained a considerable popularity recently, being commonly used in state-of-the-art proposal for NLP and computer vision. Action Transformers have been proposed by Mazzia et al. [2021] to address the challenges of HAR, outperforming common models significantly (0.8-10.68% improvement). Their self-attention capabilities are a great leap forward with come with certain drawbacks, namely: increased complexity and larger number of parameters.

PDAN (Dai et al. [2021]) is an alternative architecture which includes an attention mechanism for HAR. Its core is the *Dilated Attention Layer (DAL)*, a block capable of processing the information with different dilatations. Its authors propose to apply one convolution, then 5 DAL blocks (dilatation equal to 2^i in the i-th block) and then a convolution. A summary of the PDAN architecture can be found in Fig. 4.2. PDAN uses local I3D features as input, therefore making indirect use of an Action Transformer trained with the ImageNet dataset and then fine-tuned for the Charades dataset (Carreira and Zisserman [2017]). This iterative process to develop a new model is a current trend in deep learning — instead of developing a model from scratch, re-use parts of previously trained deep models. It is called *transfer learning*.

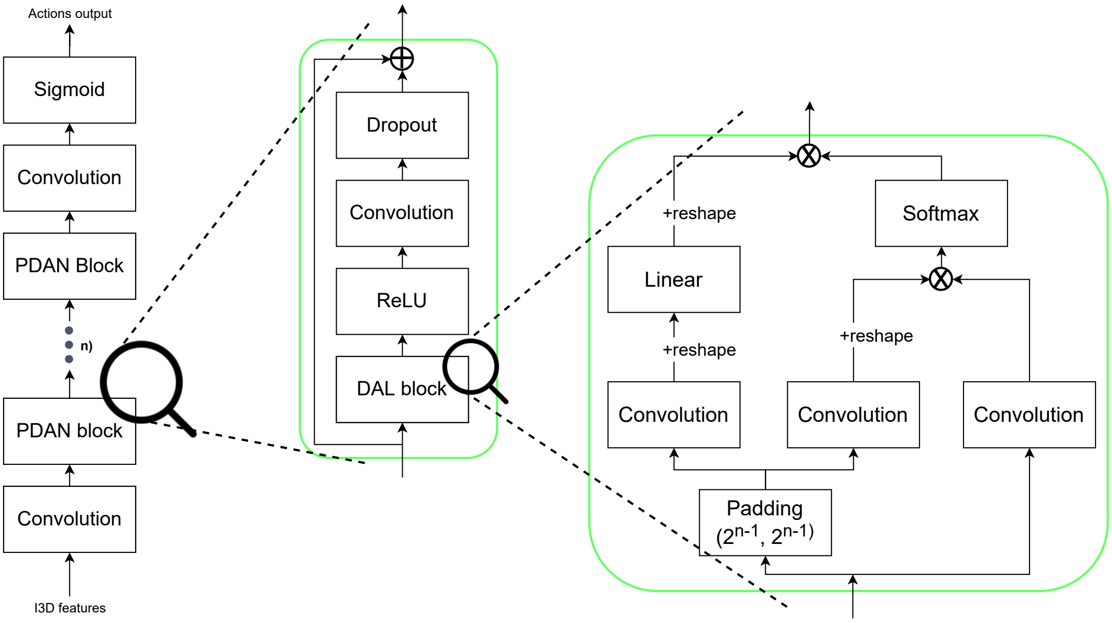

FIG. 4.2: PDAN deep learning architecture including n DAL blocks

PDAN outperforms all other methods for action detection in the Charades dataset challenge, both on the RGB modality and on the RGB+flow modality. Details can be found in Table 4.1, where PDAN and other architectures are compared against a basic classifier trained on top of the extracted I3D features.

	Modality	per frame-mAP difference (%)
WSGN (supervised)	RGB	+19.87
Stacked-STGCN	RGB	+22.43
PDAN	**RGB**	**+51.92**
Super event	RGB + Flow	+12.79
3 TGMs	RGB + Flow	+25.00
3 TGMs + Super event	RGB + Flow	+29.65
Dilated-TCN	RGB + Flow	+36.62
MS-TCN	RGB + Flow	+40.69
PDAN	**RGB + Flow**	**+54.06**

TAB. 4.1: Accuracy comparison of the different action recognition architectures. They are all based on the extracted I3D features, and compared against a basic classifier trained on top of the segment-level I3D features

Just recently, the Google Brain Team proposed a new deep learning method which has a great potential and has already surprised the researcher community — **Gated Multi-Layer Perceptron** (most commonly referred to as *gMLP*). The paper of Liu et al. [2021a] introduces a paradigm shift in ML: instead of developing more complex architectures with new characteristics, improving classic methods. gMLP achieves an accuracy comparable to Vision Transformers while using considerably less parameters and FLOPs, and it also achieved state-of-the-art results in NLP with fewer parameters.

gMLP consists of a stack of n blocks of the same size and structure, where the main novelty is the use of the *Spatial Gating Unit* (SPU) and the use of GELU (Gaussian Error Linear Unit) as the activation function. The SGU is used to capture the interactions across a sequence of elements, therefore carrying out the same role as attention in Transformer architectures, but it does not require encoding for element positions. Rather, it applies element-wise multiplication of part of its input with a linear projection of part of the same input. The SGU layer $s(\cdot)$ to contain a contraction operation over the spatial dimension as:

$$s(Z) = Z \odot f_{W,b}(Z) \tag{4.1}$$

where $\odot$ represents the element-wise multiplication, Z is the input matrix and $f_{W,b}$ is a simple linear projection of weights W and bias b:

$$f_{W,b}(Z) = W \cdot Z + b \tag{4.2}$$

This layer is meant to enable cross-token interactions within the network. Fig. 4.3 shows a representation of one of the n-blocks.

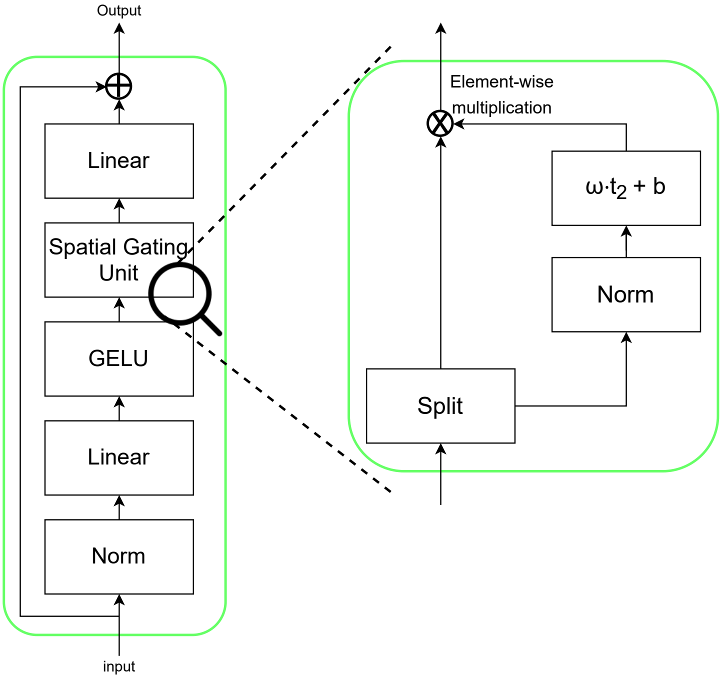

FIG. 4.3: gMLP deep learning architecture summary. One gMLP block is represented

The attention mechanism of transformers is believed to be one of their main ingredients which boost the remarkable effectiveness of Transformers. It introduces the inductive bias that the spatial interactions should be dynamically parameterized based on the input representations (Bahdanau et al. [2014]). However, it is known that MLPs with static parameterization can represent arbitrary functions as multilayer feed-forward networks are universal approximators (Hornik et al. [1989]). Thus, is a dynamic parametization required to accurately model complex visual data? gMLP results show that self-attention is not critical for Vision Transformers (Liu et al. [2021a]).

Another important deep learning architecture which have gained popularity recently is **Mixer-MLP** (Tolstikhin et al. [2021]), or simply Mixer for simplicity. The developers aim was to propose an alternative to Transformers and convolutions in computer vision, with a similar accuracy. Their proposal is based solely in MLPs which are repeatedly applied across the feature channels and the spatial locations. It relies only in matrix multiplications, scalar nonlinearities, reshapes and transpositions. Fig. 4.4 shows a summary of the architecture.

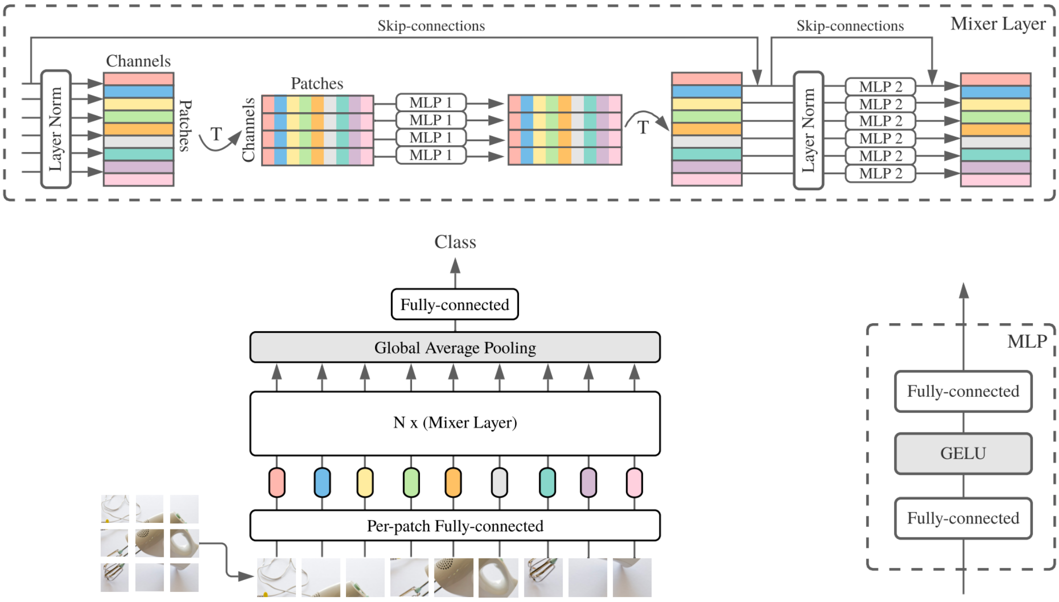

FIG. 4.4: Mixer-MLP deep learning architecture summary. Extracted from Moscholidou and Pangbourne [2020]

Let $\mathbf{X} \in \mathbb{R}^{S \times C}$ be a real-valuated two-dimensional table of S non-overlapping patches of the input image, and let C be its *hidden dimension* (non-learnable parameter). Mixer consists of multiple layers of identical size which input is geometrically altered, and each layer consists of two MLP blocks:

- *Token-mixing* layer is applied to the columns of $\mathbf{X}$ and maps $\mathbb{R}^S \rightarrow \mathbb{R}^S$.
- *Channel-mixing* layer is applied to the rows of $\mathbf{X}$ and maps $\mathbb{R}^C \rightarrow \mathbb{R}^C$.

Omitting layer indices, mixer layers can be described as:

$$\begin{aligned} U_{*,i} &= \mathbf{X}_{*,i} + W_2\ \sigma(W_1 \mathrm{LayerNorm}(\mathbf{X})_{*,i}), \quad \text{for } i \in \{1, ..., C\} \\ Y_{j,*} &= \mathbf{X}_{j,*} + W_4\ \sigma(W_3 \mathrm{LayerNorm}(\mathbf{X})_{j,*}), \quad \text{for } j \in \{1, ..., S\} \end{aligned} \tag{4.3}$$

Compared to gMLP, Mixer-MLP contains three times more parameters and is roughly 3% less accurate. However, its clear distinction of per-location (channel-mixing) operations and cross-location (token-mixing) operation, along with its skipped connections, make it a very interesting architecture to integrate within other networks. We theorise that either channel-mixing or token-mixing can capture information which the previous network missed.

One last deep learning architecture which must be considered is the **Vision Permutator**. Its distinctive feature is that it separately encodes the feature representations along the height and width dimensions with linear projections, unlike most MLP-like methods which encode the spatial information along the flattened spatial dimensions. This design feature allows the architecture to capture long-range dependencies along one spatial direction and meanwhile preserve precise positional information along the other direction. Fig. 4.5 provides a summary of the Vision Permutator architecture. Technical details are described in the end of Section 4.2.2, as the original network has been extensively modified for the integration. A summary of the modified block will be provided in Fig. 4.9.

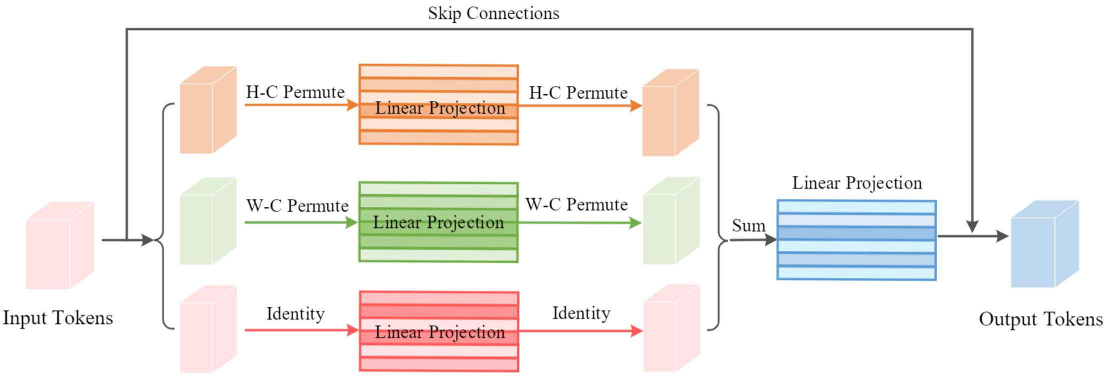

Fig. 4.5: Vision Permutator deep learning architecture summary. Extracted from Hou et al. [2021]

4.2.2 Materials and Methods

The **Charades** dataset, introduced by Sigurdsson et al. [2016], is a densely annotated dataset composed of 9,848 videos of daily indoors activities with an average length of 30 seconds, involving interactions with 46 objects classes in 15 types of indoor scenes and containing a vocabulary of 30 verbs leading to 157 action classes. Each video in the dataset is annotated by multiple free-text descriptions, action labels, action intervals and classes of interacting objects. 267 different users were presented with a sentence, which includes objects and actions from a fixed vocabulary, and they recorded a video acting out the sentence. In total, the dataset contains 66,500 temporal annotations for 157 action classes, 41,104 labels for 46 object classes, and 27,847 textual descriptions of the videos. In the standard split there are 7,986 training videos and 1,863 validation videos. Only the actions annotations are used.

Another considered dataset is **Toyota Smarthome Untrimmed (TSU)**. It is a dataset for activity detection in long untrimmed videos. It contains 536 videos with an average duration of 21 minutes, densely annotated with 51 activities. The dataset presents a unique combination of challenges, namely: high intra-class variation, high-class imbalance, and activities with similar motion and high duration variance. Activities are annotated with both coarse and fine-grained labels.

The dataset has been recorded in an apartment equipped with 7 Kinect v1 cameras. It contains common daily living activities of 18 subjects, with a resolution of 640×480. Due to privacy-preserving reasons, the face of the subjects is blurred. An example of the dataset videos, as well of Charades videos, can be found in Fig. 4.6.

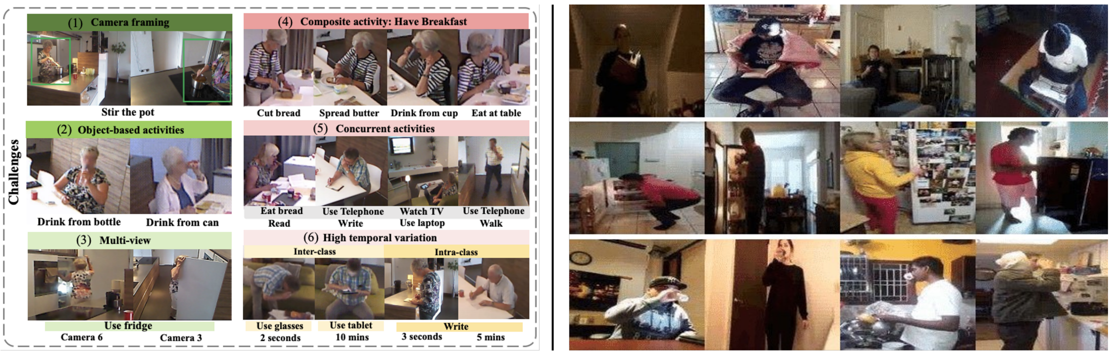

Fig. 4.6: Examples of actions from the TSU (left) and charades (right) datasets. Extracted from Dai et al. [2022], Sigurdsson et al. [2016].

These datasets were selected for the experimental phase, and the PDAN architecture was used to set a baseline.

Besides, several authors (like Mavroudi et al. [2020], Piergiovanni and Ryoo [2019, 2018]) have shown that developing a deep learning network based on I3D features provides a state-of-the-art accuracy as well as speed up the training phase. In particular, the combination of I3D features + PDAN obtains the best per frame-mAP accuracy at the moment. As a result, this promising network is the cornerstone of the developed models, which main goal is to adapt it and to successfully integrate it with other modern deep learning blocks.

The comparative study of Liu et al. [2021b] provides a good starting point for discovering state-of-the-art deep learning architectures derived from the classical MLP. In particular, the Google's proposals (i.e., gMLP and Mixer-MLP) and the Vision Permutator were the most interesting for integrating with the PDAN architecture.

During this research work, the original PDAN architecture was implemented and tested following the authors indications, and it was also adapted to run with the available computer resources. The I3D features were extracted with an I3D network trained for the Charades challenge. an evaluation accuracy of **23.1639% per-frame-mAP** was obtained for the Charades dataset, which is set to be the **baseline** of this research work.

PDAN was modified by removing one of the convolutional layers in the DAL block which was found to be redundant, and adapting the structure of the new block. Fig. 4.7 shows the proposed DAL block. The resulting architecture is henceforth called **PDAN light**.

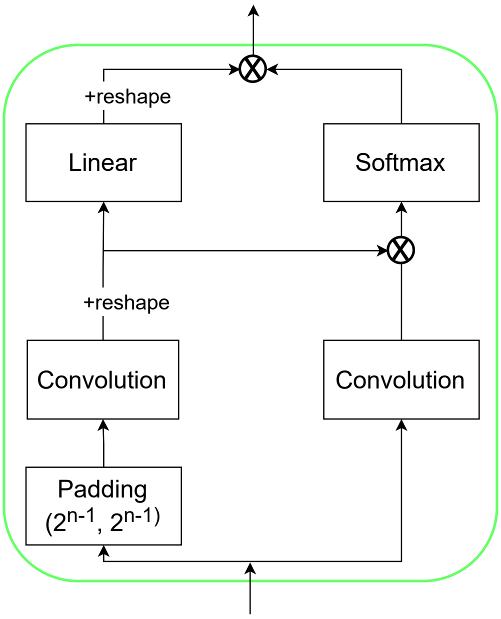

Fig. 4.7: DAL block in the proposed PDAN light architecture

The new DAL-light block maintains the multiple dilation rates, which inherently makes it learn the attention weights at different temporal scales. The information is processed across the temporal domain to preserve the spatial information. It only contains two learnable $1x1$ kernels, resulting in **22.4% less parameters** than the original architecture (4,547,741 and 5,858,461 parameters respectively).

Let's consider the n-th DAL-light block. Then, the attentional operation (output) $a_n(\cdot)$ of input features vector f_{nt} at time $t \in \{1, ..., T\}$ is:

$$a_n(f_{nt}) = Q_n(f_{nt})[\text{softmax}(Q_n(f_{nt})K_n(f'_{nt}))] \tag{4.4}$$

where $f_{nt} \in \mathbb{R}^{1\times C_2}$ and $f'_{nt} \in \mathbb{R}^{KS\times C_2}$ with KS being the kernel size; $K_n(f'_{nt}) = W_{K_n} f'_{nt}$ and $Q_n(f_{nt}) = W_{Q_n} f_{nt}$ are two independent bottleneck convolutions; and $W_{Q_n}, W_{K_n} \in \mathbb{R}^{C_2\times C_2}\ \forall n$.

Therefore, the output of the n-th DAL block for the whole video is:

$$output_n = [a_n(f_{n1}), ..., a_n(f_{nT})] \tag{4.5}$$

Note that so far only the DAL block was modified, the overall structure of the network remains unchanged for now.

In order to further enhance the network, integration of PDAN-light was performed with gMLP, Mixer-MLP and Vision Permutator blocks. In all the cases, the best performing architecture consisted having 4 PDAN blocks with dilatations starting in $(i+1)$ and adding the new block as represented in Fig. 4.8. The network state in the epoch with the higher accuracy is used as the trained network.

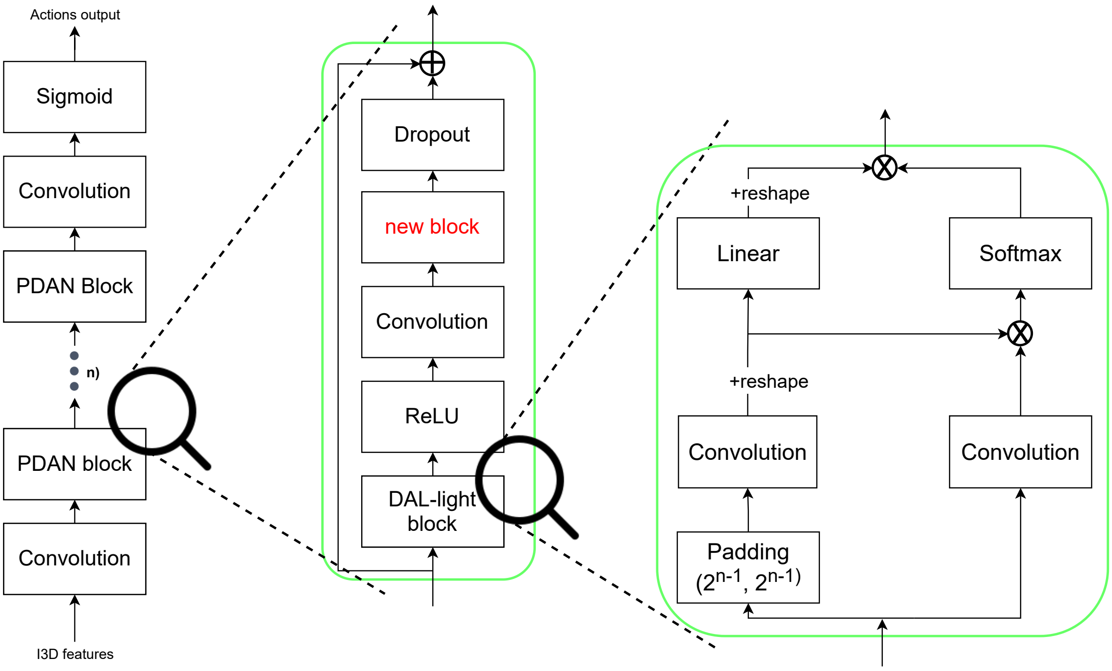

FIG. 4.8: Integration architecture for new block into PDAN light

Vision Permutator needed to be specially altered to integrate well and provide a good performance. Alterations are described in detail below, with Fig. 4.9 providing a graphical description of the proposal.

Let $X \in \mathbb{R}^{H\times W\times C}$ be the block input, with $H \times W \times C$ being the shape of the input. We then apply three independent fully connected layers with weights $W_1, W_2, W_3 \in$

$\mathbb{R}^{C \times C}$. Let X_1, X_2, X_3 be their respective output. Then, a fully connected layer (linear projection) is applied to the addition of all X_i, resulting in:

$$\hat{X} = \text{proj}(\sum_i X_i) \tag{4.6}$$

In this case, the projection has a similar structure to the feed-forward layer of Transformers: two fully connected layers with a GELU activation in the middle.

Afterwards, a softmax function is applied, and the components of $\hat{X}$ are multiplied by $X_i \ \forall i$:

$$Y = \text{softmax}(\hat{X})_x \cdot X_1 + \text{softmax}(\hat{X})_y \cdot X_2 + \text{softmax}(\hat{X})_z \cdot X_3 \tag{4.7}$$

Lastly:

$$output = X + \text{proj}(Y) \tag{4.8}$$

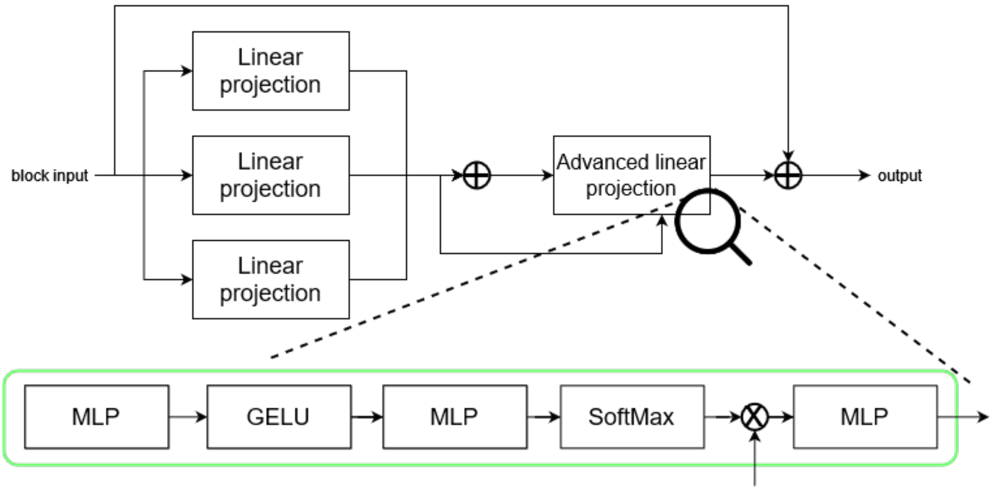

FIG. 4.9: Proposed weighted vision permutator deep learning block

The number of training epochs for evaluation is selected using the regularization method *early stopping*.

4.2.3 Experimental Results and Discussion

The hardware used in this research was a PC with a i7-8700K processor, 16GB RAM memory and a Nvidia GeForce GTX 1070 GPU. Note that the available hardware caused limitations in the study as the I3D features could not be extracted as other authors did. The I3D network had to be modified to shrink its GPU memory requirements. Only up to 1,000 frames could be inputted into the network at once, as opposed to the original configuration of up to 16,000 frames. A new baseline with the extracted features was used for the experiments, as described in the previous section.

Let's first consider the Charades dataset. The best performing architecture in this case is the integration of the original PDAN and Vision permutator, obtaining a **23.965 per-frame mAP**. This represents a **3.34% improvement** compared to the baseline. However, the most balanced architecture is the one including MLP-mixer and the DAL-light: it has a reduced number of parameters, which results in fast training; and achieves a 23.595 per-frame mAP accuracy. Results indicate that PDAN light does not adapt well to the modern MLP blocks, as the performance increase is lower than in the integration with the original network. However, the PDAN light architecture has a positive behaviour alone, with a 23.455 per-frame mAP, a 22.4% parameter reduction and a 30.96% decrease in training times. A table summarising the accuracy obtained by the different architectures on the Charades dataset can be found in Tab. 4.2.

Networks	Parameters (M)	Train size	Test size	Eval. accuracy
PDAN	5.85	7985	1863	23.164
+ MLP-Mixer	4.89	7985	1863	23.664
+ Vision Permutator	10.32	7985	1863	23.965
+ gMLP	4.86	7985	1863	23.549
PDAN light	4.54	7985	1863	23.455
+ MLP-Mixer	3.84	7985	1863	23.595
+ Vision Permutator	9.27	7985	1863	23.617
+ gMLP	3.82	7985	1863	23.520

Tab. 4.2: Accuracy comparison of the different architectures on the Charades dataset, with the baseline on in red.

It is also interesting to analyse the accuracy in each epoch of the different architectures during the training. As it can be seen in Fig. 4.10, PDAN light behaves very similarly to the original PDAN architectures, despite having significantly less parameters. The architecture including Vision Permutator achieves a higher accuracy in the early stages, which then decreases fast; indicating an overfitting behaviour due to the high number of parameters. MLP-Mixer provides a balanced improvement: it reduces the number of parameters, achieves a higher peak accuracy and overfits less over time. Lastly, gMLP makes the training slower but more consistent: it starts with a lower accuracy than the baseline, but then surpasses it and overfits less over time.

Moreover, experiments on the TSU dataset were carried out. Tab. 4.3 provides a summary of the accuracy obtained with the different integrated architectures.

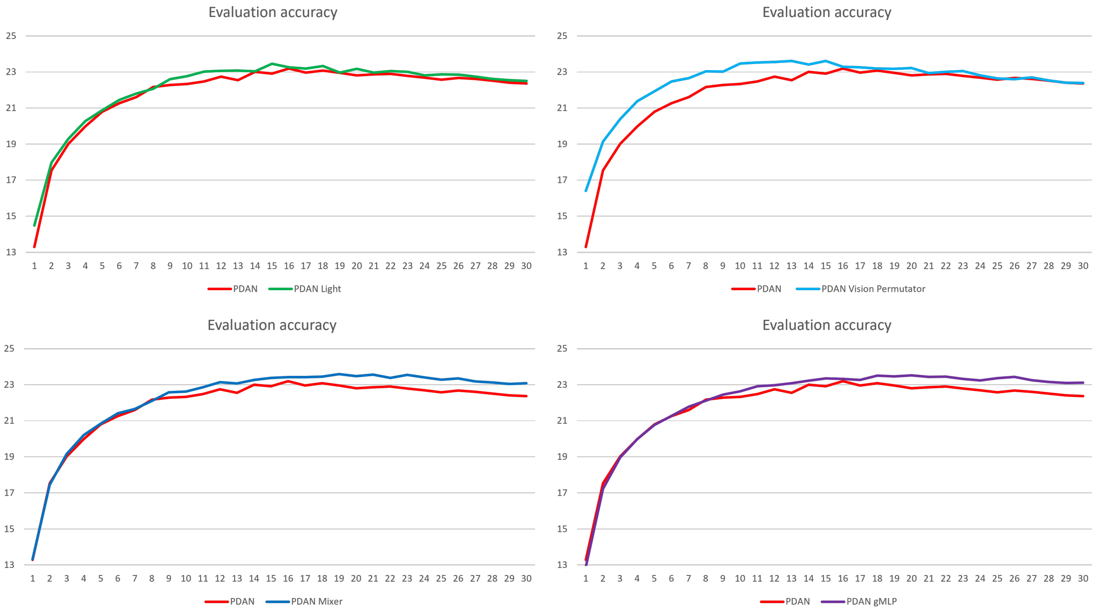

Fig. 4.10: Evaluation accuracy of the proposed architectures. Baseline is always shown in red in all charts. The x-axis shows the epoch number and the y-axis shows the per-frame mAP accuracy

Networks	Parameters (M)	Train size	Test size	Eval. accuracy
PDAN	5.80	351	185	31.714
+ MLP-Mixer	4.83	351	185	31.929
+ Vision Permutator	10.27	351	185	30.870
+ gMLP	4.81	351	185	31.345
PDAN light	4.49	351	185	32.554
+ MLP-Mixer	3.78	351	185	31.715
+ Vision Permutator	9.22	351	185	31.612
+ gMLP	3.76	351	185	31.406

Tab. 4.3: Accuracy comparison of the different architectures on the TSU dataset, with the baseline on in red.

In the TSU dataset, PDAN light obtained the best results with a **+2.7% accuracy increase** and a **22.6% reduction of trainable parameters**, compared to the baseline PDAN. This increase accounts for a 32.554 per-frame mAP. It is also noteworthy that the integration on PDAN light and MLP-Mixer resulted in the greatest reduction in the number of parameters, a −34.8%, while maintaining an accuracy statistically similar to the baseline. The architecture including Vision Permutator performs badly in this dataset, below the baseline; and again PDAN light seems not to integrate well with the modern MLP blocks, as the accuracy of PDAN light alone is the highest.

All in all, results show that PDAN light is an improvement of the original architecture regarding both the computational cost and the model performance. Furthermore, the integration on PDAN light and MLP-Mixer provides an outstanding reduction of the computational cost of the architecture, while providing an comparatively similar accuracy to the original architecture.

However, the integration of PDAN architectures with Vision Permutator is not successful as a rule. Results seem to indicate that the pronounced increase of parameters makes the model prone to overfitting, as there is a sharp increase in training accuracy during the first epochs; and also because the good performance does not transfer across different datasets. The integration of PDAN architectures with gMLP is also not successful as a rule, with good results in one dataset and bad results in another.

4.2.4 Conclusions and Future Works

In this chapter we have explored how to model complex temporal relations in densely annotated video streams. A modified version of the PDAN architecture has been proposed, improving the accuracy, the performance and reducing its computational complexity. Moreover, state-of-the-art deep learning blocks have been integrated with the proposed architecture, resulting in an even higher accuracy and the capability to better learn the features representation across time. The method has been evaluated on two densely annotated multi-label datasets: Charades and Toyota Smarthome Untrimmed. The former is a common benchmark for Action Detection in the state of the art, and the latter poses a unique combination of challenges with very long videos. Results indicate that our PDAN light architecture outperforms all other state-of-the-art methods.

A decrease of up to 34.87% in the number trainable parameters was achieved on the Charades dataset, as well as an evaluation accuracy gain of up to 3.34%. In the TSU dataset, a decrease of up to 34.8% in the number trainable parameters was achieved and an evaluation accuracy gain of up to 2.7%

Chapter 5

Smart Living

Chapter 5

Smart Living

5.1 Introduction

Smart Living seeks to increase the quality of life and safety across all age groups and demographics in the SC. Optimizing the available services and easing citizen's access to them are two key goals of Smart Living. This results in greater social and digital inclusion, enhanced security, improved healthcare and better smart buildings.

With security at the centre of the Smart Living concept, both digital and urban security must be taken into account, and it must be possible to access security services online. Public safety is becoming a growing concern, especially in developing countries that are undergoing an intense urbanisation process (Ismagilova et al. [2019]).

Tourism digitization is also a major factor: tourist information should be available online, landmark tickets sold online, a tourist card developed, cultural visits improved through the usage of modern technology, comprehensive cultural heritage management should be implemented and cultural programs should be broadcast online (Su et al. [2011]).

Healthcare is one of the most important indications of the quality of life of residents. Innovative solutions regarding health information and education can boost efficiency, as well as help prevent infectious diseases and guide the general public towards a healthier lifestyle (Casino et al. [2017]). In addition, it can improve access to healthcare and the life of people with disabilities.

Several research gaps have been identified for Smart Living, as well as useful datasets which can be used to develop a novel solution:

- According to estimations, 20–50 billion devices will be connected to the internet in a few years (Goel et al. [2021]). Such a large number confirms that IoT is already affecting our day to day and will continue to do so in the near future,

raising important security concerns among SC residents. An automatic anomaly detection system for the identification of security breaches in IoT networks, would be a great step forward in increasing the cities' digital security.
The dataset (Naveed [2020]) contains real traffic data from IoT sensors. Data is presented in the form of statistics packages such as the weight of the stream, the covariance between two streams, etc. There are 8 GB of data in several csv files. Attacks have been carried out by two botnets, infecting 9 commercial IoT devices by Mirai and BASHLITE (Meidan et al. [2018]).

- Police response to street shootings is currently delayed as bystanders must call an emergency number and explain the incident. A gunshot detection system could analyse street noise and send an alarm to the police department instantly when a shooting takes place, with a precise GPS location. A system could be trained to distinguish gunshots from other street noises, enabling the city's microphones to monitor incidents in their area.
  The dataset (Lilien et al. [2017]) contains approximately 10,000 individual gunshot recordings (format "wav"). It is necessary to register to get access (free) and files of different weapons and recording devices must be downloaded separately. It includes 18 different gun models, recorded with two different smart phones and an advanced mobile microphone.

- Preventing public safety problems is a new possibility thanks to the power of Big Data and Analytics, based on historical public safety data. Public services can use predictive models to identify areas in the city where crime is most likely to take place, anticipating and then dispatching officers to deter potential criminals. In particular, the average crime rate of each region can be modelled and abnormal numbers can then be detected and corrected.
  The dataset (Denmark [2021]) contains 543,780 data cells of reported criminal offences by region and type of offence in Denmark. Currently, there are quadrimestral data from 2017 to 2021. A maximum of 10,000 data cells can be downloaded per query.

- 75% of the buildings in the EU are not designed according to any efficiency code, whereas around 45% of the world energy is being used in the residential sector. Therefore, this is one of the main challenges any major city will face, and which it can stand up to via smart home development, energy modelling for buildings and reward system implementation for those who use energy efficiently.
  The dataset (Batra et al. [2014a]) contains the energy-related measurements of a commercial building (IIIT Delhi, a university campus in India). The refresh rate of data is 30 seconds, and it was captured over 30 days in June 2014. There are

independent datasets from 8 different sources (lifts, lights, sockets, floor totals...) whose frequency is synchronised but not their start time. Several aspects of data are vaguely explained in the article Batra et al. [2014b] or not explained at all. Following the information provided in the article, it has been possible to deduce that “power.csv” is watts, “current.csv” is hertz and “energy.csv” is kWh.

5.2 Proposed Method: nnU-Net for CTV & Organ Segmentation (Proton Therapy)

Recent clinical trials have shown the enormous potential of proton therapy in cancer treatment, due to its unique physical dose deposition properties (Elhammali et al. [2019], Lesueur et al. [2019], Wang et al. [2021]). Its main benefit is the steep dose fall-off at the proton’s end-of-range, the so-called “Bragg Peak”. This property allows for a standard target volume coverage while sparing the healthy tissue and organs located behind it. Unfortunately, Intensity Modulated Proton Therapy (IMPT) planning is a very demanding procedure: first, Computed Tomography (CT) images of the patient must be obtained; then, Organs At Risk (OARs) and the Clinical Target Volume (CTV) must be identified and carefully delimited following strict guidelines; afterwards, dose objectives are prescribed to each OAR and to the CTV; next, the overall dose prescription for the patient is generated; then, inverse planning is used to simulate optimal beamlet configuration which would result in the prescribed dose using robust optimization methods; and eventually, the proton therapy is delivered to the patient typically using Pencil Beam Scanning (PBS).

As a result, treatment planning is a time-consuming process which currently requires several specialised physicians, sending their results back and forth until a successful outcome is agreed upon, and which can last up to a week. Delaying cancer treatment is known to cause a lowering in survival rate, with recent studies asserting that every month delayed in cancer treatment can raise the risk of death by 6 - 13% (Hanna et al. [2020]).

The proposed method uses a deep learning approach to conduct the CTV and OARs segmentation in seconds, and it is being adapted to also carry out the overall dose prescription.

5.2.1 Related Works

Automated machine learning (AutoML) is a new paradigm for automating the task of applying ML to real-world problems. Traditional ML methods have the need of an expert to fine-tune a developed model to a particular use case, by carefully adapting the hyper-parameters. However, autoML allows for non-experts to adapt and deploy a model to their use case without any need for modifications. This high degree of automation allows developers to produce solutions faster and more easily, while outperforming many hand-designed models. In particular, it is well-suited to computer vision problems because of of its generalised data preparation and feature engineering methods (He et al. [2021]).

In the medical field, U-Net shaped approaches to computer vision have gained a great popularity. The architecture of the original network can be found in Fig. 5.1. Barragán-Montero et al. [2019] used a hierarchically densely connected U-Net (HD U-Net), which combines the original U-Net and DenseNet, to perform 3D dose prediction for lung intensity modulated radiotherapy (IMRT) patients; and Wu et al. [2021] used the same architecture to boost the accuracy of pencil beam dose calculation. Other authors have adapted this architecture to perform medical image segmentation, for example creating the "UNet 3+" which makes use of full-scale skip connections and deep supervisions (Huang et al. [2020]).

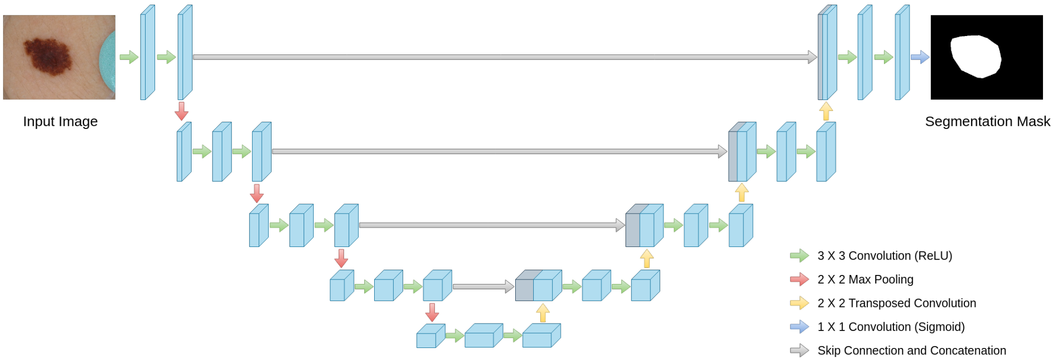
Input Image
Segmentation Mask
3 X 3 Convolution (ReLU)
2 X 2 Max Pooling
2 X 2 Transposed Convolution
1 X 1 Convolution (Sigmoid)
Skip Connection and Concatenation

FIG. 5.1: U-Net deep learning architectural framework

Some of the most common applications of the U-Net architecture are: pixel-wise regression, dense volumetric segmentations, image-to-image translation and, our use case, image segmentation.

In order to evaluate the results of medical image segmentation, and in particular OAR segmentation, several previous studies have focused on geometric evaluations between manual and automatic organ segmentation (Zhu et al. [2020]). Geometric evaluations are most commonly based on the previously described DICE coefficient, the

Jaccard similarity coefficient and the Mean Distance to Agreement (MDA). The DICE and Jaccard coefficients focus on evaluating the similarity based on the overlapping area between the two delineations, while the Mean Distance to Agreement focuses on evaluating the distance of outline points. Henceforth, the DICE coefficient is used.

One of the most common loss (error) functions in ML organ segmentation is the Sørensen–Dice coefficient, also known as *dice similarity coefficient (DSC)* and simply *DICE coefficient.* This is a statistical measurement which gauges the similarity of two examples. In this particular case it measures the overlapping of ground truth contours and the contours generated with the developed network. It is defined as:

$$DICE(X,Y) = \frac{2 \cdot |X \cap Y|}{|X| + |Y|} \tag{5.1}$$

where X is the ground truth and Y is the generated contours.

In order to establish a baseline for our experiments, we consider the model of Dong et al. [2019] as the current state of the art for OARs segmentation, because of its high accuracy in several delicate organs. They applied their method to five OARs: left and right lungs, spinal cord, esophagus, and heart; using a database consisting of 35 patients' chest CTs. The mean DICE score of the left and right lungs was identical (0.97 and 0.97), the spinal cord's score was very high (0.9), as well as the heart's score (0.87), while the esophagus's score was significantly lower (0.75).

Concerning the CTV segmentation, the new *DeepTarget* model (Jin et al. [2021]) is considered as the baseline, due to its outstanding performance compared to other state-of-the-art works. The authors obtained a DICE coefficient of 0.826 using as input Positron Emission Tomography (PET) / CT images + information of the Gross Tumor Volume/CTV/OARs, encoded using a novel method they call "SDM" (Signed Distance Transform Map).

5.2.2 Materials and Methods

This section presents the used datasets and the proposed AutoML architecture.

Patients Database

Two different databases were considered in this study. They were proposed in the article of Barragán-Montero et al. [2021]. They will be referred to as "variable database" (VarDB) and "homogenized database" (HomDB) from now on.

The variable database was extracted retrospectively from the patient directory, and contained 60 patients with esophageal cancer treated with IMRT in the University Hospitals Leuven, from 2016 to 2020. As the treatment protocol has evolved over time, this database contained three different treatment machines (Clinac 2100C/D, TrueBeam, and Halcyon, Varian Medical Systems, Palo Alto CA) and different beam configurations (from 5 to 9 coplanar beams) and beam energies (6 or 10 MV). In addition, it involved different physicians and medical physicists for contouring and planning, respectively. Dose calculation and optimization was done using the AAA algorithm (version 10.0.28 or 15.6.03), from the treatment planning system (Eclipse, Varian Medical Systems, Palo Alto CA). The version of the treatment planning system evolved during the time frame of the study from Eclipse 13, Eclipse 15.1 and Eclipse 15.6.

The homogenized database was intentionally built to reduce the variabilities in the variable database, by re-contouring and re-planning the same 60 patients from the variable database, carefully following organ at risk delineation guidelines (Jabbour et al. [2014]) and a fixed planning protocol. The 60 patients were uniformly planned using a seven beam IMRT class-solution on Halcyon (beam configuration at 0°, 30°, 60°, 155°, 220°, 300°, 330° angles) and an updated list of dose constraints (Ajani et al. [2019]). Optimisation was performed by the same observer (medical physicist with 10 years of experience in inverse planning optimization with Eclipse), starting from a fixed set of objectives and weights and clinical judgement by one radiation oncologist, for all plans. Higher priorities were set for the dose constraints in the lungs, since recent studies (Beukema et al. [2020]) have found more evidence for the correlation between lung dose volume parameters and pulmonary toxicity and survival (Wang et al. [2006]). Hence, all treatment plans in this database were guided towards the lowest possible lung V5, V10, V20, V40 and Dmean. The dose calculation algorithm was AAA version 15.6.03, from the treatment planning system (Eclipse, Varian Medical Systems, Palo Alto CA). The version of Eclipse was the same (15.6) for all HomDB plans. Detailed information about the contour delineation and the replanning objectives to generate the HomDB can be found in the Supplementary material. For the VarDB plans, the constraints, objectives and priorities evolved over time, but the exact information could not be retrieved due to the retrospective nature of the database.

For both the variable and homogenized databases, the prescribed total radiation dose to the planning target volume (PTV) was 45.0 Gy in fractions of 1.8 Gy.“ Gy” is the symbol for a standard unit for ionizing radiation called “Gray”, which equals to $m^2 \cdot s^{-2}$. The dose map grid size was the same for both databases and equal to 2.5 mm × 2.5 mm (in plane) x 3 mm (slice thickness).

The manual delimitation of the homogenize database is henceforth considered as the ground truth, but both databases were described as they are closely related.

NOTE: The use of patient data for the study was approved by the Institutional Ethical Review Board of the University Hospitals Leuven (S59667).

AutoML architecture

Forty eight patients diagnosed with esophageal cancer before 2020 were used for training the OAR-segmentation models. The OARs selected for segmentation were: heart, lungs, liver, left kidney, right kidney, spinal canal plus 3 millimetres, left ventricle. Twelve patients were selected to validate the deep learning model. Training and validation patients were selected randomly from the database described in section 5.2.2. CTV segmentation was also carried out in the same sets of patients.

The nnU-Net architecture (Isensee et al. [2021]) was selected for developing the deep learning models used for OARs and CTV segmentation. This method is based upon the well-establish U-Net network, but it is capable of adapting itself to the considered datasets, removing the need to fine-tune the hyper-parameters of the network when applying it to a new use case. This was achieved buy generalizing the manual parameter-tuning carried out by researchers for the Medical Segmentation Decathlon (Antonelli et al. [2021]). Fig. 5.2 shows a summary of the method.

The autoML component of the network is based on obtaining the so-called *data fingerprint* and applying a set of rules. It consists of the following procedures:

- Cropping the CT-images of the training cases to the non-zero region. This reduces the training dataset size and lower the training computational complexity.
- Calculating the number of voxels per spatial dimension before and after cropping.
- Adapting the input based on the modality found in the metadata.
- Mean, std, 0.5 and 0.995 percentiles, of the intensity values in the foreground regions.
- Normalizing CT images.
- Detecting and correcting outliers, as it is possible to find anomalous labels.
- The global dataset percentile clipping and the z score are calculated based on the global foreground mean and its standard deviation for the CT images modality, and on the per-image mean and standard deviation otherwise.

- If the input is anisotropic, the distribution of spacings is calculated in-plane with third-order spline, and out-plane with the nearest neighbour. Also, it is equal to the tenth percentile in the lower resolution axis and to the median in the other axes.

  If it is not anisotropic, third-order spline is used in every data point and the median spacing for every axis.

- Adapting the network topology, batch size and patch size according to the available GPU.

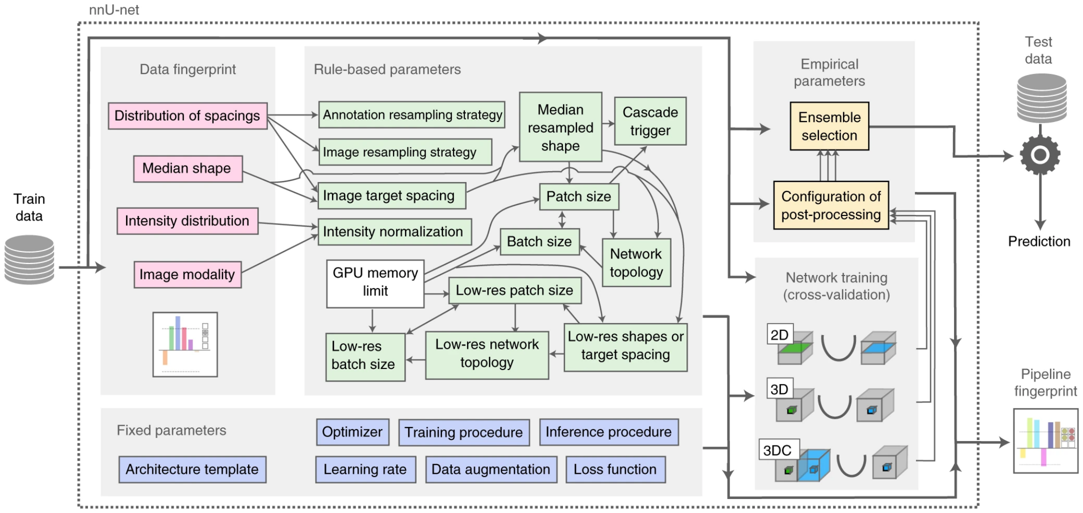

FIG. 5.2: nnU-Net deep learning architectural framework. Isensee et al. [2021]

Next, a standard U-Net network is applied. There are four possible variants: 1) 2D U-Net, 2) 3D U-Net, 3) 3D cascade U-Net, 4) an ensemble of two of the previous architectures. The best performing model (or combination of two) is selected according to the 5-fold cross validation performance. Therefore, training of all architectures must be performed, which has a great computational cost.

Interestingly, strong data augmentation is carried out during the training to avoid overfitting, and dropout is subsequently not needed. The used data augmentation methods are: elastic deformations, mirroring, flipping, scaling, rotating, random cropping, varying the brightness in each channel.

In the post-processing phase, is it calculate whether the all-but-largest-component-suppression increases the cross-validation performance for each target class (OARs and CTV in our case). This technique is used only on the classes in which it results in an improvement.

An important method used to improve the overall network performance was **mixed precision**. Mixed precision (Micikevicius et al. [2017]) consists of using both 16-bit and 32-bit floating-point types (FP16 and FP32 respectively) during model training, obtaining equal accuracy than traditional 32-bit networks but lowering the memory consumption, and the step time for each epoch. Only modern GPUs can make use of this optimization.

Mixed precision uses FP16 to store the weights, gradients and activations during training iterations, and FP32 to store a master copy of the weights. During a training iteration, first the FP32 master eight is converted to FP16, then the forward pass and the backpropagation are computed, then the weights are updated and, at the end of the iteration, the weight gradients are used to update the master weights.

5.2.3 Experimental Results and Discussion

The performance of the model regarding the OARs, with several difference configurations, is described in Table 5.1. 5-fold cross validation has been used to assess how the obtained results would generalize to an independent dataset. The DICE coefficient of the heart, lungs, liver, left kidney, right kidney, spinal canal plus 3 millimetres and left ventricle in the best performing model were: 0.9128018, 0.9805966, 0.960512, 0.9296532, 0.9298872, 0.880695, 0.9014234; respectively. Therefore, it can be concluded that the model can reliably delimit the OARs in the CT images of patients with esophageal cancer. Moreover, it achieved a better performance that the current state-of-the-art approaches in several OARs (+4.28% in heart delimitation and +1.05% in lungs delimitation against the previously described baseline), although falling behind in the spinal cord delimitation (-1.93%).

Note that we delimited the spinal cord with a three millimetres isotropic expansion. The performance in this regard is directly compared against other authors because it is possible to retrieve the original spinal cord segmentation by shrinking our delimitation uniformity in all orientations. Thus, the regions used for calculating the DICE coefficient could be shrank correspondingly, obtaining the same result than if the original spinal cord segmentation were computed.

				MT_Heart		MT_Lungs		MT_Liver		MT_Kidney_L	MT_Kidney_R		MT_SpinalCan_03	MT_LeftVentricle	
resolution	**epochs**	**image**		raw	posprocess	raw	posprocess	raw	posprocess	raw	raw	posprocess	raw	raw	posprocess
3mm	**100**	**cropped**	fold 0	0.8977753	0.8977753	0.9819038	0.9818916	0.9554385	0.9554385	0.9407314	0.9357668	0.5459083	0.8345155	0.9010424	0.9010424
			fold 1	0.8926716	0.8926716	0.9812017	0.981207	0.9556981	0.9556981	0.8509012	0.8588125	0.7811218	0.8522994	0.9025545	0.9025545
			fold 2	0.8920798	0.8920798	0.9779444	0.9779561	0.9568113	0.9568113	0.9360619	0.9416874	0.6319471	0.8281737	0.896413	0.896413
			fold 3	0.8960825	0.8960825	0.983388	0.983399	0.9599064	0.9599064	0.9399134	0.9315057	0.7765992	0.865514	0.8815938	0.8815938
			fold 4	0.874644	0.874644	0.9763848	0.9763701	0.9596455	0.9596455	0.9454952	0.9419605	0.6252012	0.8405918	0.8170492	0.8170492
			MEAN	0.8906506	0.8906506	0.9801645	0.9801648	0.9575	0.9575	0.9226206	0.9219466	0.6721555	0.8442189	0.8797306	0.8797306
		original	fold 0	0.9155024	0.9155024	0.9807017	0.9807087	0.9610952	0.9610952	0.9491651	0.946783	0.6328655	0.8958114	0.9200651	0.9200651
			fold 1	0.9169476	0.9169476	0.9823967	0.9824204	0.9597017	0.9597017	0.9051068	0.9043206	0.6674315	0.8829147	0.9155271	0.9155271
			fold 2	0.9129121	0.9129121	0.9797382	0.9797475	0.9623471	0.9623495	0.9435687	0.9439884	0.5858306	0.8702765	0.8728912	0.8728912
			fold 3	0.9114912	0.9114912	0.9808411	0.9808605	0.9600675	0.9600675	0.9049719	0.909453	0.5938079	0.875907	0.9188593	0.9188593
			fold 4	0.9071558	0.9071558	0.9793052	0.9793186	0.9593476	0.9602269	0.9454537	0.944891	0.5905645	0.8785636	0.8797746	0.8797746
			MEAN	0.9128018	0.9128018	0.9805966	0.9806112	0.9605118	0.9606882	0.9296532	0.9298872	0.6141	0.8806947	0.9014234	0.9014234
original	**1000**	**original**	fold 0	0.895598	0.89555	0.9842871	0.8140165	0.9567442	0.9567442	0.9446204	0.9405317	0.9405317	0.8591702	0.8955085	0.8955085
			fold 1	0.8960039	0.8961105	0.7115638	0.7115638	0.9589397	0.9589569	0.8577113	0.8660246	0.8660251	0.8553654	0.9094451	0.9094451
			fold 2	0.8972901	0.8972839	0.980284	0.7494418	0.9581713	0.9581713	0.9385045	0.9444338	0.9444338	0.8651018	0.9052236	0.9052236
			fold 3	0.899917	0.8998864	0.9849167	0.8235456	0.9651122	0.9655959	0.93837	0.9345211	0.9353185	0.8832659	0.9008072	0.9008072
			fold 4	0.8770196	0.8769766	0.9789694	0.7395799	0.9620913	0.962092	0.9483026	0.9448126	0.9448126	0.8501049	0.8152452	0.8152452
			MEAN	0.8931657	0.8931615	0.9280042	0.7676295	0.9602117	0.9603121	0.9255018	0.9260648	0.9262244	0.8626016	0.8852459	0.8852459

Tab. 5.1: Results of the model regarding the OARs. Different configurations were tested and 5-fold cross validation was always used

The segmentation accuracy of the models regarding the CTV can be found in Table 5.2 and Table 5.3. 5-fold cross validation was used. The common clinical procedure for manual CTV segmentation also makes use of the PET/CT and (echo-)endoscopy, a technique which is not standardized and which data consists of a text file with information of where the lymph nodes are located and how long the tumour is. Thus, the automatic deep learning segmentation which only makes use of the CT images is not expected to be very reliable in this case. Results show that the best performance has a DICE coefficient of 77.13%, making use of CT images and OARs segmentation, both with the original voxel size of 1 mm x 1 mm.

resolution	organs	epochs	raw	pp_all
original	included	100	0.746116	0.745864
		500	0.771254	0.771268
3mm	not included	100	0.708519	0.708723
		500	0.740884	0.740897

TAB. 5.2: DICE coefficient for clinical target volume's segmentation, fold 0

	raw	**pp_all**
fold 0	0.771254	0.771268
fold 1	0.714262	0.714009
fold 2	0.775222	0.775135
fold 3	0.712402	0.712062
fold 4	0.725337	0.725345
MEAN	**0.739695**	**0.739564**

TAB. 5.3: 5-fold cross-validation for CTV's segmentation, 500 epochs, organs included, original resolution

Results indicate that the obtained segmentations are in close agreement with the ground truth. A qualitative example of OAR segmentation example can be found in Fig. 5.3. From these experiments, nnU-Net shows a great potential for image segmentation in CT images of esophageal cancer patients. In the used dataset, the model successfully delimited all the OARs (>90% DICE) besides the spinal cord. The spinal cord and CTV delineation were partly successful but did no achieve a high enough performance for medical usage. Furthermore, as the model was optimised for the Medical Segmentation Decathlon, it can seamlessly adapt to other datasets, without any hyper-parameters tuning nor CT image preprocessing.

5.2.4 Conclusions and Future Work

The findings of this study show that a reliable automatic delineation of OARs can be used in clinical applications, and that the CTV segmentation, although improved, is not accurate enough to replace manual delineation. Automatic delineation has the potential

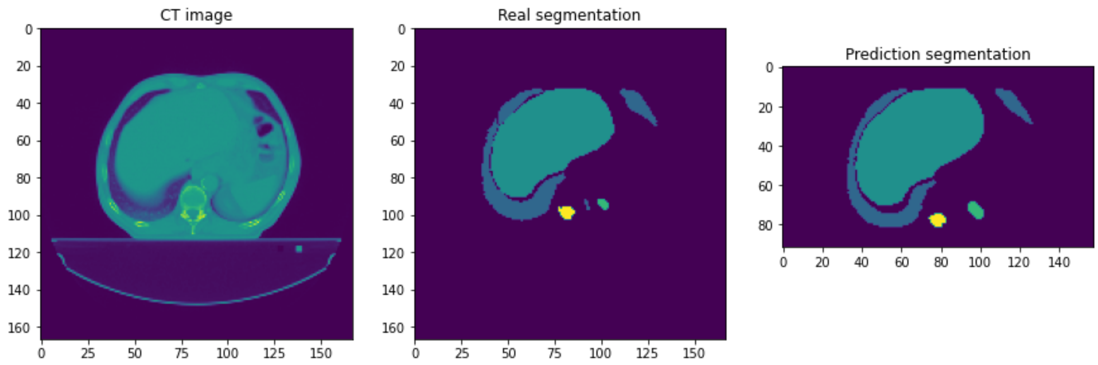

FIG. 5.3: Qualitative example of OARs segmentation results. CT image (left), manual segmentation (center) and obtained segmentation (right)

to reduce the planning phase of a proton therapy from up to a week, to mere minutes; a desirable step forward as the survival rate of esophageal cancer patients decreases over time. In addition, it can reduce the number of physicians needed to plan a proton therapy, consequently lowering its cost and releasing valuable human resources back to the hospital.

The reliability of this model is supported by the >90% DICE coefficient of all organ's segmentation (besides the spinal cord). A possible use of this deep learning approach would be to make the initial CT images segmentation based on it. Then, an experienced physician could adjust and validate the segmentation of the OARs, and carry out the CTV delimitation. This would result in faster planning without accuracy drop. However, it is important that the physician carries out a segmentation from scratch occasionally, so that they don't lose this crucial skill.

Future research lines are based on adapting the nnU-Net architecture to produce a continuous output, and subsequently using it for automatically carrying out the dose prediction for a proton therapy of the considered patient (Fig. 5.4). A modified version of a U-Net network has already been developed for this task, and some of the core ideas of the nnU-Net architectures have been added to it, improving its performance.

In particular, the following improvements were implemented in the modified U-Net network:

- Mixed precision.
- An optimization algorithm which reduces the learning rate when the training accuracy has stopped improving
- A tailored loss function for the training phase: random locations of each class were selected during training and their Mean Squared Error (MSE) computed. This is

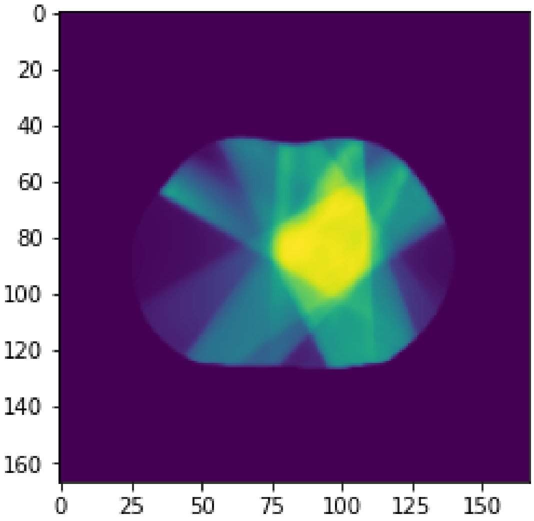

FIG. 5.4: Proton therapy prescribed dose for an esophageal cancer patient, based on their CT images

believed to lower the computational cost of training while preventing overfitting in the network.

Preliminary results already show a lower training time per epoch (up to 6.1% faster) and an outstanding boost in validation accuracy, with a 27.63% loss decrease. In particular, the validation loss lowered from 4.4010 to 3.1847.

Acknowledgements

Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

Chapter 6

Smart Economy

VNiVERSiDAD D SALAMANCA

Chapter 6

Smart Economy

6.1 Introduction

Smart economy is a vertical market based on innovation and entrepreneurship. Both action strategies can help fight unemployment, boost productivity and improve the overall labour market.

Since the 18th century, globalization has been an unceasing process due to advances in transportation and communication technology, resulting in great but challenging opportunities for local businesses and companies (Bretos and Marcuello [2017]). Adaptations to changes must be reasonably fast as there is always the risk that another actor on the global market will take advantage of it and dominate the market.

City internationalization can be easily achieved via the internet, and technological information can be obtained online effortlessly. A comprehensive list of the research gaps identified for Smart Economy, as well as some data to develop a solution, can be found below:

- Business and commerce networks are a key part of the economy of any Smart Territory. Clustering similar economic activities has the potential to boost their success and productivity, and finding information on the best performing areas can help detect promising businesses early.
  The dataset (Hall [2019]) contains information on the economic activities at street level in the city of Barcelona (Spain) from 2014, 2016 and 2019. 105 fields are available and their description can be found on the same website (although it is only available in Catalonian language).
- Entrepreneurial education and training emerge naturally in Smart Territories, as a consequence of user-centred designs and practices. New entities, which develop into SCs, can take advantage of the available data resources and propose novel

solutions to the challenges associated with entrepreneurial education and training. The dataset (Zillow [2022]) contains values of different states, countries, cities, ZIP code areas and neighbourhoods in the USA, for the period 2000-2022. This is a large collection of data with many missing values. It can be used by any new economic actor to predict the future market behaviour. There is information about several types of dwellings. All the provided values refer to the housing value over time.

- Startups play a major role in today's economic growth as they revitalise the local economy, favour innovation and generate employment. Predicting and encouraging their growth is an important factor for the success of any new SC.
  The dataset (KC [2020]) contains 48 economic descriptors of 924 startups in the USA. It contains businesses which have already been closed and their closure data. The columns include information on "industry trends, investment insights and individual company information".

6.2 Proposed Method: Transfer Learning for Automatic House Categorization in Cadastres Registries

Land typology categorization is one of the biggest challenges facing land registries around the world. Historically, this task has been performed manually, requiring a large workforce, creating a big workload and resulting in a low process. In recent years, satellite imaging has revolutionised the cadastral activities in rural areas, as the new information retrieval method has allowed for less physical expeditions and various automatic classification methods have been proposed (Matikainen et al. [2004], Müller and Zaum [2005]).

Nevertheless, land registries cannot be considered as *smart* at the moment. Automatization has not yet arrived nor has been proposed to bring it to cadastral urban applications. This chapter proposes a method to fulfil this research gap. A deep learning solution is proposed to perform automatic building categorization based on a dataset of the cadastre of Salamanca city (Spain).

6.2.1 Related Works

Convolutional Neural Networks have gained a great popularity since 2011, when they won their first computer vision competition, obtaining superhuman performance in the German Traffic Sign Recognition Benchmark of IJCNN 2011. They are exceptionally useful as they can learn to extract the most important features from an image automatically, a process which in the past people had to carry out and design features (feature engineering) manually, being very time-consuming and challenging.

The core building block of CNNs is the convolutional layer. Mathematically, it is a discrete convolution defined as:

$$(f * h)[m, n] = \sum_j \sum_k f[j, k] \cdot h[m - j, n - k] \tag{6.1}$$

where f refers to the input image, h is the chosen kernel (also called *filter*) to be applied and m, n refer to the indexes of rows and columns of the resulting matrix, respectively.

h is a trainable parameter which has a small receptive field. However, the receptive field is expended over the full depth of the input, resulting in a 2-dimensional activation map of the kernel during the forward pass. The network learns by choosing kernels which activate when a specific feature is found in a particular spatial location of the input, thus achieving classification capabilities.

Based on these basic blocks and ideas, several successful neural networks have been developed for computer vision.

Inception V3 (Szegedy et al. [2016]) is a CNN based on the Inception network, but which adds several enhancements to the original architecture. In particular, these enhancements are: label smoothing, factorized 7 × 7 convolutions and batch normalization in some layers. Moreover, it makes usage of cross entropy as a loss function and adds an auxiliary classifier in order to propagate label information along the network. It defines the probability of each label $k \in \{1, \ldots, K\}$ for a training example x as:

$$p(k|x) = \frac{\exp(z_k)}{\sum_i z_i} \tag{6.2}$$

where z_i, with $i \in \{1, \ldots, K\}$, are the original (unnormalized) log-probabilities. Fig. 6.1 shows a summary of the network.

Xception (Chollet [2017]) is a CNN which is based on depthwise separable convolution layers exclusively. The network is formed by 36 convolutional layers, organised in 14 modules. Fig. 6.2 shows a detailed representation of it.

Depthwise Separable Convolutions split the input processing into two steps: (1) a single convolutional filter is applied to each input channel, and (2) a linear combination of the output is created using a pointwise convolution.

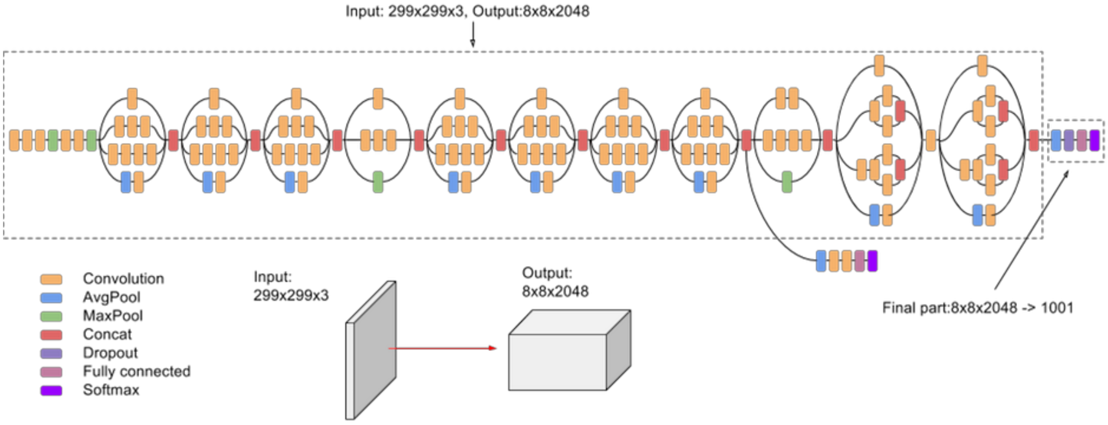

FIG. 6.1: High level representation of the Inception V3 network.

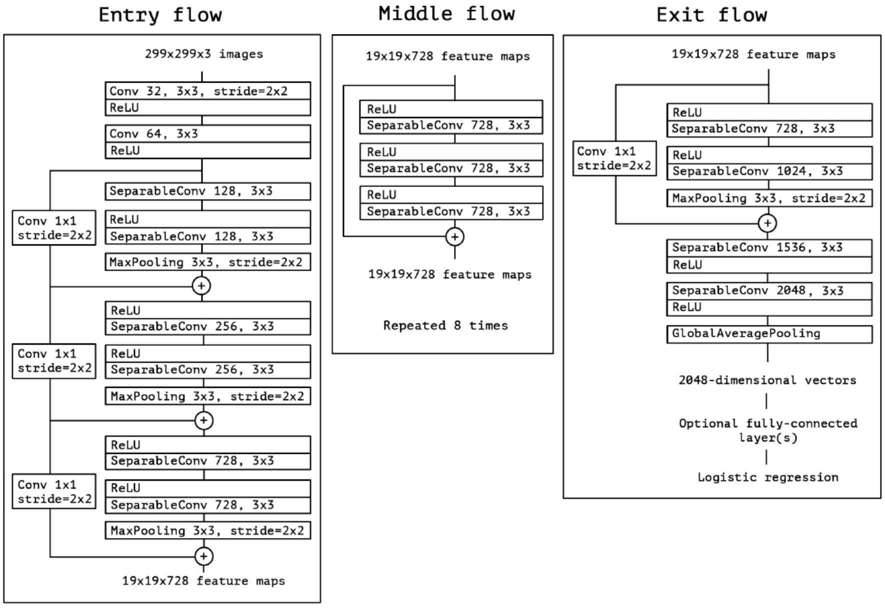

FIG. 6.2: High level representation of the Xception network.

Compared to Inception V3, it has more parameters ($\sim 3\%$) but its training is around 9% faster.

VGG16 (Simonyan and Zisserman [2014]) is a CNN based on the classical VGG architecture used for ImageNet. It is characterised by its simplicity: it only uses pooling, convolutional and fully connected layers, and it accepts only images of a fixed size (224×224 RGB). Fig. 6.3 shows a summary of the network.

Support vector machines are a common choice for dividing the solution space generated after extracting the low-level features of the previous ML architectures. It is a well-established method, from 1963, which first applications were already related to image classification and obtained good results (Cortes and Vapnik [1995]). Over the

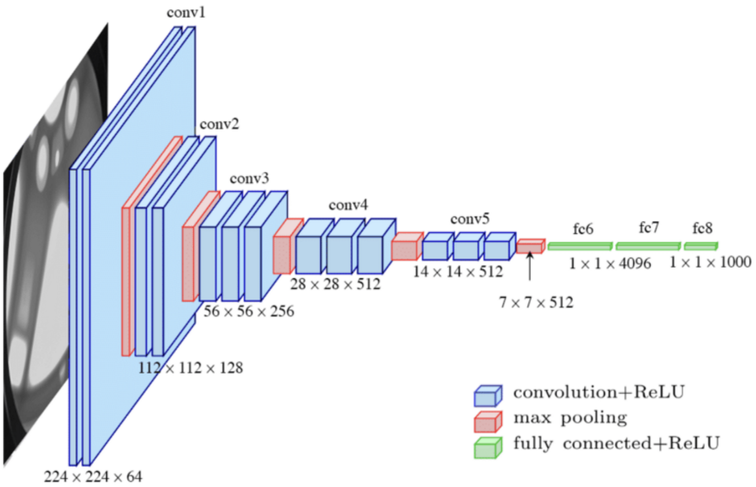

Fig. 6.3: High level representation of the VGG16 network.

years, it has become a popular technique in ML due to its good balance of performance and simplicity. Sec. 2.2.1 further develops this subject.

6.2.2 Materials and Methods

This section presents the used datasets and the proposed transfer learning solution.

Dataset

The main goal of this research work is to automatically predict the typology and category of urban dwellings making usage of images of their facade. Images which do not contain metadata related to the building nor any kind of non-visual information.

Thus, a dataset from the land registry of Salamanca (Spain) was selected. It contains images of 5802 dwellings as well as their corresponding building category, taken with a high-resolution camera. The Spanish cadastre classifies the urban dwellings into main three building categories, although one of them is extremely rare to be found in the city (1-34 ratio against other categories) and therefore was dismissed. The building categories are: “1.1.2 collective houses in a closed bloc”, “1.2.1 single or semi-detached building” (dismissed) and “1.2.2 single-family houses in line or closed bloc”.

Note that other building categories can be found in the registries of the Spanish cadastre. However, some of them are too outdated to be found in cities at the moment, and the rest of them do not refer to urban dwellings.

The input data has been subdivided into two sets: training and testing. 80% of the images were randomly assigned to the training set, and the remaining 20% of images were assigned to the testing set. Figure 6.5 shows a visual example of the dataset and the performed classification.

Deep Learning Model

Several works in the literature make use of an integration of CNNs and a SVM for image classification tasks (Elleuch et al. [2016], Niu and Suen [2012], Sampaio et al. [2011], Xue et al. [2016]). We also take advantage of this idea in the developed model.

We make use of a deep CNN which is already trained to extract the lower level features of the images. Then, a SVM is trained on the features to obtain predictions related to the building category, and the output is compared against the real label.

For the training phase, a squared L2 penalty and a radial basis function kernel are used:

$$Kernel(x, x') = \exp(-\gamma \cdot \|x - x'\|^2) \tag{6.3}$$

where the parameter γ is sometimes replaced by the parameter σ, which relation is $\gamma = \frac{1}{2\sigma^2}$. They must be set *a priori.*

The squared L2 penalty is a regulation term based on the l_2 norm:

$$\|x\|_2 = \sum_i \mid x_i \mid^2 \tag{6.4}$$

6.2.3 Experimental Results and Discussion

Several pretrained network architectures have been tested as the backbone of the model, where a SVM use their lower level features to perform a classification. The performance of the different deep networks regarding their accuracy and their mean squared error can be found in Table 6.1. The best performing architecture was obtained by Xception.

Tab. 6.1: Performance of the different trained networks+SVM

	Inception V2	**Inception V3**	**VGG16**	**Xception**
Accuracy	**0.6442**	**0.6809**	**0.6354**	**0.8557**
MSE	**0.3557**	**0.319**	**0.3645**	**0.1442**

A 85.6% accuracy was achieved by the best performing architecture, Xception backbone plus SVM. The training phase lasted only 19 minutes 52 seconds, and it took only 373 milliseconds to classify a new image. Tests were carried out in a PC with a Intel i7-8700K processor and 16 GB RAM. The confusion matrix of the selected model can be found in Fig. 6.4.

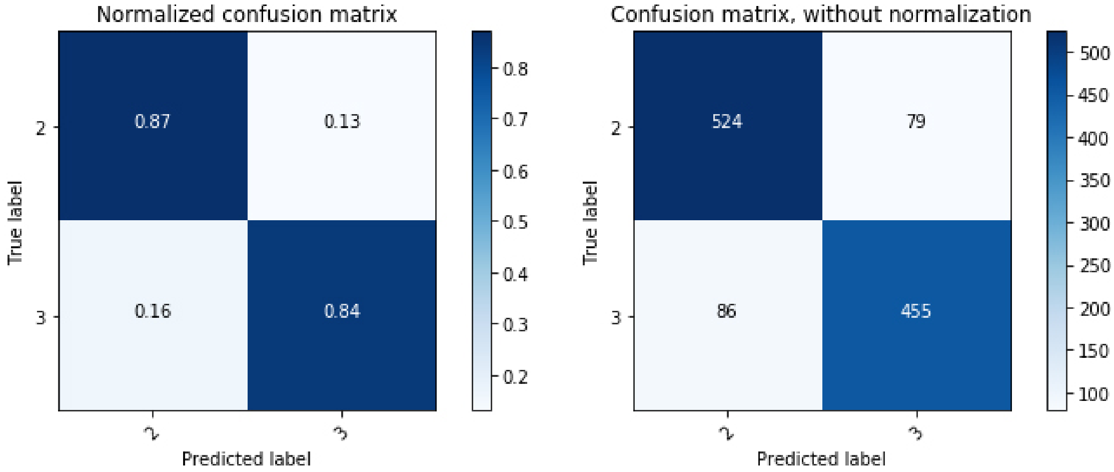

Fig. 6.4: Confusion Matrix of the best the Xception and SVM-based model, absolute and relative results.

Moreover, it is interesting to observe which images where successfully classified and which led to a classification mistake. An example regarding this idea, as well as representing the proportion of success-failure, can be found in Fig. 6.5.

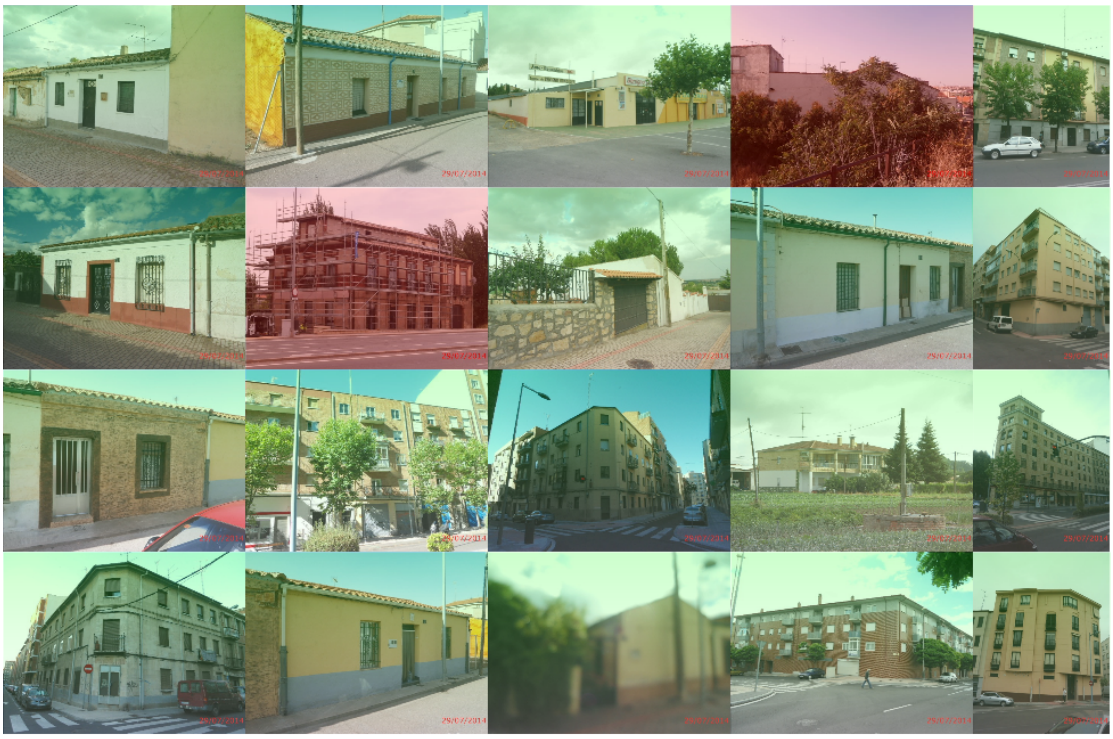

Fig. 6.5: Obtained results for building classification. Images in red show errors in classification, images in green show successful cases.

The developed algorithm could several purposes, such as:

1. Verify human classification. A system could be implemented to compare the classification performed by cadastre workers against the machine prediction. In case of disagreement, a different worker should re-classify the image, untying the situation. It would detect and solve biased decisions, such as, an operator facing a classification of the house of a close relative. In general, it would be a useful tool for quality control and internal fraud prevention.

2. Improve worker's productivity. The automatic prediction could be used so that a worker only needs to accept or reject classifications, and then manually select the building category of the rejected set, examining the corresponding images in detail. This alternative would diminish the workload while maintaining the classification quality.

Furthermore, the only computationally expensive task of the method is the network training. This task could be performed only once a year in a big server, and then used in basically any PC of the local sections of the cadastre. Funding costs would be low in this case.

6.2.4 Conclusions and Future Work

A method capable of automatically classifying the typology and category of dwellings has been obtained. Results indicate that the classification is made with a notable degree of confidence, but human intervention is still necessary to ensure trust-worthy results. Low funding is needed to implement the method in a real-world scenario.

The proposal has the potential to improve the productivity of cadastres with a visual database of their assigned buildings. We live in a historical moment were machine-based image classification equals or surpasses human classification in many applications. It makes us optimist that in the future more and more land registry tasks can be automatised and the number of mistakes lowered.

In order to achieve above-human classification accuracy, a bigger database would be required, and the network could be re-trained. Moreover, this method could also include other typological categories or to be applied to rural dwellings using orthophotographs.

Chapter 7

Smart Environment

Chapter 7

Smart Environment

7.1 Introduction

Smart Environment is an innovative concept that refers to the way in which natural and built resources are managed, its aim is to improve the habitability of the Smart Territory and the well-being of its residents. Recently it became a major concern of the younger generations and the Paris Agreement secured the political commitment of all countries around the world (Schleussner et al. [2016]).

The implementation of regulatory and cultural changes in big metropolises is key if air and noise pollution is to reduce. Moreover, the long-term goal of achieving a sustainable city must be based on efficient waste management, high-quality food and water management, rational consumption patterns, thorough urban planning and increased awareness.

Furthermore, environmental control involves assessing disaster risks and enabling the development of systems for the early detection of natural disasters (Moe and Pathranarakul [2006]). It would therefore be easier to predict power outages and their outreach could be mitigated with a smart grid and the use of local renewable energy. These and other possible use cases, as well as open data which could be used to develop a solution, can be found in the following list:

- Modelling the bodies of water surrounding a SC is of critical importance when dealing with natural disasters. A system for early anomaly detection can maximise the time available to issue a warning and perform a fast evacuation. Moreover, the effects of climate change can be studied and their implications for the coastal areas of the city may be predicted before they actually appear.
  The dataset (Institute [2007]) contains complete weather and ocean data in Ireland, for the period 2001-present (currently 2022). There are several missing values but even if the related data is omitted, there are still over 200,000 data points.

GPS coordinates are included, as well as waves and wind information, sea and air information, relative humidity and the dew point.

- Climate change is considered to be one of the biggest threats of our age, and world data is the central focus of debate. Smart territories can use models and visualizations to raise awareness among their citizens and to test their carbon emission reduction initiatives. In particular, the ambitious goals of the Paris Agreement can be constantly analysed and each country's commitments checked. The dataset (Boden et al. [2013]) contains the total carbon emissions from fossil fuel consumption and cement production for every country, as well as from solid/liquid/gas fuel consumption, cement production, gas flaring, bunker fuels, and per capital missions. Data dates back to 1751 in the case of some countries, and end in 2014.

- Air pollution is a serious issue all over the world due to the negative consequences it has on health. It results in millions of deaths every year (Organization et al. [2005]). Pollution waves can be shortened using an early detection system and citizens can be warned against spending long periods in areas where air pollution levels are high.
  The dataset (Government [2019]) contains information on air quality in Seoul (South Korea). The measured elements are: $SO_2, NO_2,\ O_3, CO, PM_{10}, PM_{2.5}$. There are 24 stations and their GPS locations are included. There are hourly data from 01/01/2017 to 31/12/2019.

- People are increasingly worried about the effect of new residential and industrial sectors on underground/surface water quality, and the environment in general. Accessible, real-time monitoring of waters can provide valuable information to citizens and boost their trust in the sanitation system.
  The dataset (government [2019]) contains a chemical analysis of water in several locations all around India, ranging from 2003 to 2014. It has 2,000 data points. No GPS data are available, only state and location names.

- For many centuries human beings have produced all types of waste. In modern times, the urbanization of the world has led to an increase in waste production, which affects human health. Smart Territories could make use of advanced, versatile vehicles such as drones to take images of the streets and then, street cleaners could be directed to the dirties areas.
  The dataset (Cchangcs [2019]) contains 2,527 garbage images annotated with 6 categories (cardboard, glass, metal, paper, plastic and trash). The background is mostly white.

7.2 Proposed Method: Transfer Learning for Tick Identification and Transmitted Disease Diagnostic

As urbanization becomes more widespread, city residents are more exposed to pollution, toxic air and sicknesses. Outdoor activities and occasional nature excursions have become more prevalent in the last few years. Warm and humid climates, without extreme temperatures, favor the spread of insects and small organisms that can cause serious health problems. As Solanas et al. [2014] state in their paper, SCs can efficiently tackle this problem by using the context-aware network and sensing infrastructure.

The combination of big data and smart systems has the potential to improve the healthcare sector. Several different concepts have been proposed within this scope: (1) *e-Health*, consists of using electronic health records, (2) *m-Health*, consists of using mobile devices to access medical data, (3) *s-Health*, consists of interactive health-related information which is made available to citizens, and (4) *m-Health augmented with s-Health*, which uses mobile devices to provide health authorities with patient's emergencies.

This chapter focuses on tick-borne diseases. They have a diverse incidence in different geographical regions and contexts. However, early and accurate diagnostics are know to notably reduce mortality and morbidity rates (Bratton and Corey [2005]).

In particular, the Crimean–Congo hemorrhagic fever (CCHF) has become an important health concern in a wide geographic range including Africa, Asia, the Middle East, Eastern Europe and, soon, Western Europe. It consists of a zoonotic disease with no commercially available vaccine, difficult prevention, and a case fatality rate as high as 40% in some outbreaks (Organization [2013]).

We propose a method for automatic tick recognition based on image capture, where the possible transmitted diseases are identified.

7.2.1 Related Works

Convolutional Neural Networks, already introduced in Sec. 6.2.1, are commonly used to replicate the receptive fields of the human eyes and brain, in order to understand visual stimuli. Mathematically, they are a regularised version of the multilayer perceptron. Some of their top applications include: facial recognition, climate understanding, document analysis, advertisement, and image enhancement.

MobileNet (Howard et al. [2017]) is a type of CNN designed for mobile and embedded vision applications. It considerably reduces the number of trainable parameters when compared to a traditional CNN network. It makes use of depthwise separable convolutions, a concept based on two operations: depthwise convolution and pointwise convolution.

Residual Neural Networks (ResNet) was introduced by He et al. [2016]. The paper proved that, against popular believe at the time, that extra layers do not progressively learn more complex features. In fact, the authors showed empirically that there is a maximum threshold for depth after which traditional CNN models have a higher training and test error (Fig. 7.1). It is a clear signal of the network overfitting. The problem of training extremely deep networks should be solved, they proposed, with a new neural network layer — the residual block.

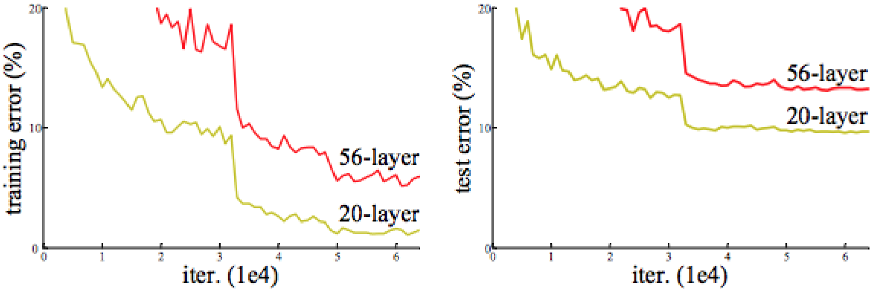

FIG. 7.1: Training error (left) and test error (right) on the CIFAR-10 dataset of a 20-layer and 56-layer CNN. Extracted from He et al. [2016]

The residual block introduced *skipped connections* for the first time, which is basically an identity mapping from a previous layer to the output of some stacked layers:

$$y = f(x, \{W_i\}) + x \tag{7.1}$$

where $f(\cdot)$ is the mapping of the chosen stacked layers, with their corresponding weights $\{W_i\}$; and x is the layer input. Fig. 7.2 represent the residual block concept.

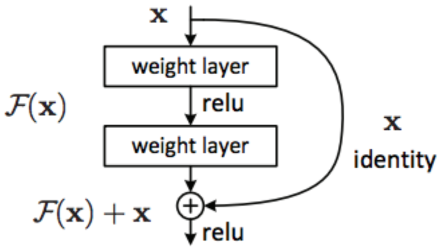
x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

FIG. 7.2: Basic building block for residual networks. Extracted from He et al. [2016]

ResNet gained a great important on 2015 when the ResNet50 won the ImageNet challenge, becoming the first computer vision algorithm to obtain a lower classification error than humans. A milestone where machines displayed a suprahuman results.

There is also a particularly interesting case of ResNet where the models have several parallel skips called **DenseNets** (Huang et al. [2017]). They are characterised by their *Dense Blocks*, where all the layers are directly connected with each other via dense connections. In order to preserve the feed-forward nature of the network, layers can only obtain inputs from previous layers and pass its output (feature maps) to subsequent layers.

DenseNets solve the overfitting problem of very deep CNNs by ensuring a maximum information flow, by easing the path for gradients passage. They exploit the potential of feature re-usage, instead of obtaining representational power from ever-deeper architectures. Fig. 7.3 shows an example of a DenseNet network used for a computer vision task, namely image classification.

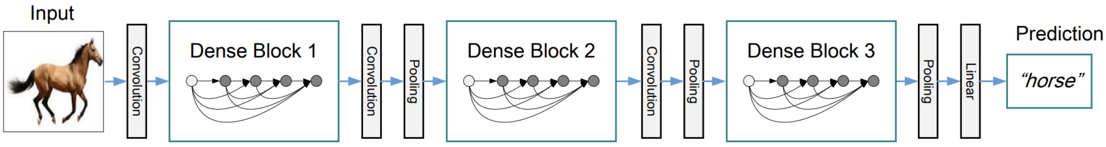

FIG. 7.3: DenseNet including of three Dense Blocks. Extracted from Huang et al. [2017]

EfficientNet (Tan and Le [2019]) is an architecture which cleverly combines all the previous concepts: it is a CNN with a scaling method that uniformly scales all dimensions via a *compound coefficient*, a scaling method which it also uses to improve ResNets. It scales all the dimensions of the network (width, depth, and resolution) uniformly, making use of fixed scaling coefficients.

The EfficientNet-B0 network is a model based on a modern iteration of MobileNet (MobileNetV2) and its inverted bottleneck residual blocks. It stands out due to its "compound coefficient", which is the base to perform a uniform scaling of all depth, width and resolution dimensions of the network. This method stands on the idea that, if the input image is bigger, then the networks need to be deeper so that the receptive fields are able of capturing more detailed patterns on the bigger image. Figure 7.4 shows a comparison of the EfficientNet scaling method against other solutions.

The previous architectures can be used as the backbone of a new network, in a fashion which is called **Transfer Learning**. Transfer learning consists of gaining knowledge from a (deep) pre-trained network for an initial purpose (i.e., feature extraction), and then adapting it in order to solve a new use case. It allows ML engineers to use to

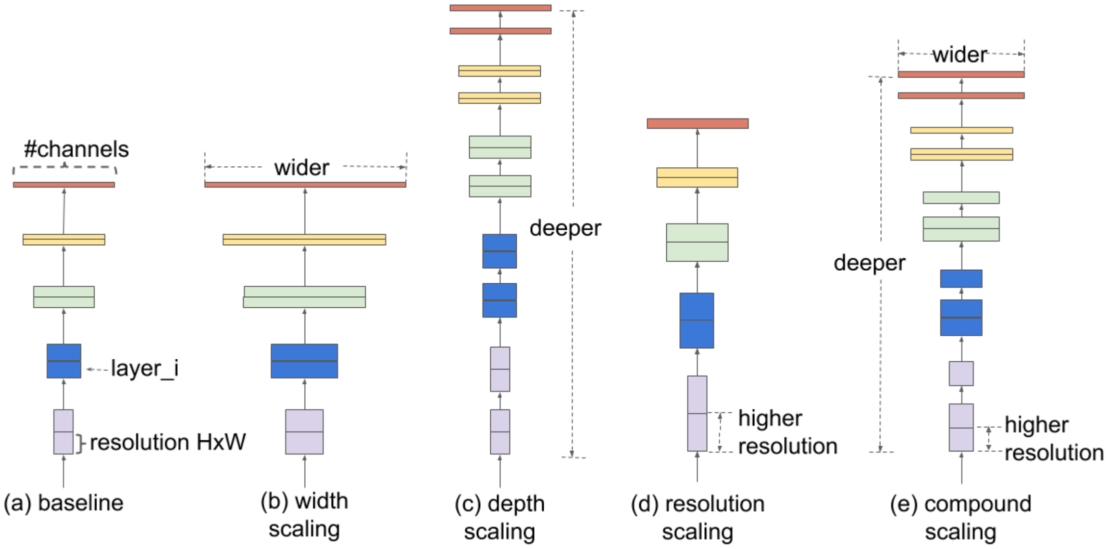

FIG. 7.4: Comparison of different network scaling methods. (a) is the baseline, (b)-(d) are traditional solutions which only increase one dimension, (e) uniformly scales the three dimensions based on a set ratio. Extracted from Tan and Le [2019]

powerful networks without the computational resources their training would require, and to use them for problem for which they were not originally designed. It has become a great leap forward in computer vision projects, as it brings the results of supercomputers to common applications. Section 2.2.2.1 further introduces this topic.

The **Otsu's method** for threshold calculation must also be described as it is a powerful tool for image processing. The general algorithm's pipeline can be described as:

1. Calculate histogram and intensity level probabilities of the input image.
2. Let $\omega_i(t), i \in \{1, 2\}$ be the probabilities of two classes divided by a threshold t, $\mu_i(t)$ be the mean of class i, and $max_{intensity} = 255$. Initialize $\omega_i(0)$, and $\mu_i(0)$.
3. Iterate over all possible intensity thresholds $t \in \{0, ..., max_{intensity}\}$. In each step:
    - Update the values of ω_i and μ_i.
    - Calculate the between-class variance $\sigma^2(t) = \omega_1(t)\omega_2(t)[\mu_1(t) - \mu_2(t)]^2$.
4. Obtain $\max(\sigma^2(t))$, which is the desired threshold.

7.2.2 Materials and Methods

This section presents the used datasets, the procedure to photograph a tick, the image preprocessing and the proposed ML architecture.

Data

One image dataset was used for this study. It was extracted from the ticks received by the CIETUS (Centre for Tropical Diseases Research of the University of Salamanca, in Spanish): in January-February 2021, all the analysed ticks were photographed, classified and sorted in a NoSQL database.

The considered dataset contains colour images of ticks extracted from the patients, either by a professional or a nonprofessional, which were manually examined by an expert to determine the subject species. For the most part, each specimen was photographed twice to capture all of its characteristic features (up to five photos were required in some cases).

Five tick species can be found in the region in which the study took place (Castile and León):

- Ixodes. They are characterized by a dorsal shield without grooves. They have an elongated face without eyes. They are medium-sized ticks (3 to 4 mm including mouthparts).
- Hyalomma. They are characterized by an elongated face with eyes and a dotted back. It has either ringed legs or marbled and long legs. They are large-sized ticks (5 to 6 mm including mouthparts).
- Rhipicephalus. They are characterized by a hexagonal-shaped basis capituli, with eyes and with triangular adanal plates. Their palpi are wider than long and they are medium-sized ticks (3 to 5 mm including mouthparts).
- Dermacentor. They are characterized by its ornamented shield, and its basic morphology is of ixodid tick family. They are medium-sized ticks (3 to 5 mm including mouthparts).
- Haemaphysalis. They are characterized by its scalloped, no eyes, rectangular capitulum base, no grooves.

The original dataset contains approximately 2200 images in total, of which 460 are labelled. The subset of the labelled images photographed with an HD camera is selected for the tests, as the subset of labelled imaged photographed with a smartphone presented a very low quality.

Haemaphysalis images was also dismissed as their number is not statistically significant. In the study region, only around five specimens are found per year, and the considered database only contains images of one haemaphysalis tick. Moreover, the Rhipicephalus images also presented the following issues and are not considered: low number of images (17), little diversity of specimens (11), and relatively bad pictures (most of the specimens seem to be very small).

Subsequently, 174 images are used in the experiments, split as: 141 training images and 33 validation images. Data leakage was a main concern as the dataset contains several different images taken from the same specimen. This issue was addressed by holding back a validation dataset which was selected manually, where two different images of the same specimen must always belong to the same subset. The train-validation ratio is close to 80-20.

The resulting dataset contains images taken with a HD camera, a device which rural doctors would have no problem in using (most ticks are extracted in rural areas and then analysed in cities). The four more common ticks found in the considered region are shown in Fig. 7.5, with real-life examples with show their most characteristic features.

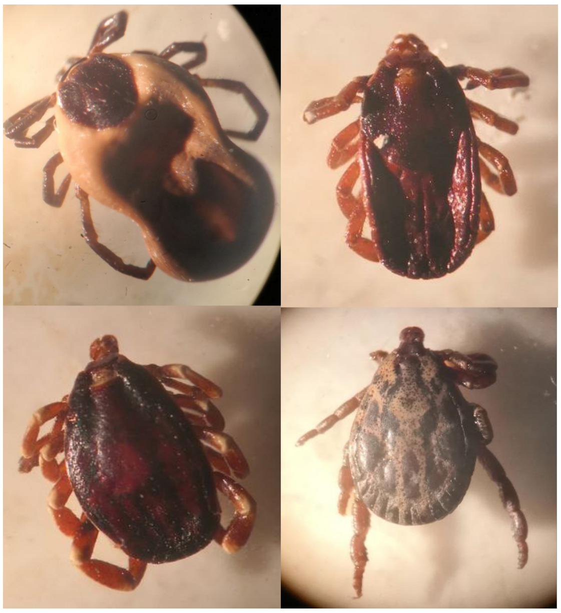

Fig. 7.5: Ticks found in Castile and León every year. Ixodes (top left), Rhipicephalus (top right), Hyalomma (bottom left) and Dermacentor (bottom right)

The diseases which can be transmitted by each tick species are:

- Ixodes: Lyme disease, anaplasmosis and the central European encephalitis.
- Hyalomma: Crimean–Congo hemorrhagic fever.
- Rhipicephalus: Boutonneuse fever, also called Mediter-ranean spotted fever.
- Dermacentor: rickettsiosis, tick-borne lymphadenopathy (TIBOLA) – also called Dermacentor-Borne Necrosis Erythema Lymphade-nopathy (DEBONEL).

- Haemaphysalis: rickettsiosis.

User Process

In order to ease the spread of the proposal and improve its effectiveness, a method has been developed so that any individual can successfully carry out the analysis. The process is as follows:

1. The tick is extracted from the patient's skin using fine-tipped tweezers.
2. The tick specimen is placed on a flat surface, on top of a white piece of paper. Most ticks die when they are extracted from the skin, so is very unlikely that for the specimen to run away.
3. The camera is positioned 20 centimetres above the specimen and the picture taken. Several pictures can be used to confirm the classification.
4. The photo is opened by an app or webpage, which uploads it to a server responsible for applying the proposed method and send back the classification.
5. The app or webpage shows the species prediction to the user, and its associated possible diseases.

Data Preparation

Each image needed to be processed before feeding it to the ML module to obtain the best results. The goal is to have a consistency so that all the input is as homogeneous as possible. The main issue is to crop the area of the tick within the image, maintaining a constant image width and length.

The applied method is as follows:

1. Conversion of the original image to grayscale.
2. Low-pass filter application to remove image noise, losing some image details. In particular, a Gaussian filter is applied.
3. Calculation of an image threshold t using the Otsu's method (Bangare et al. [2015]).

4. Using the threshold to convert the image into black and white. Pixel values higher than t become black and values lower than t become white. As all images have a white paper background, only the tick region stays white after applying the threshold t.

   NOTE: In graysacale images, a pixel value of 255 represents the colour white, and 0 represents the colour black.

5. Creation of a bounding rectangle s which includes all the white pixels. The smallest rectangle containing all the white pixels is selected.

6. Cropping the original image according to the coordinates of s.

7. Outputting the new image, resized to some constant dimensions.

A real-life example of the results of the previously described method can be found in Fig. 7.6.

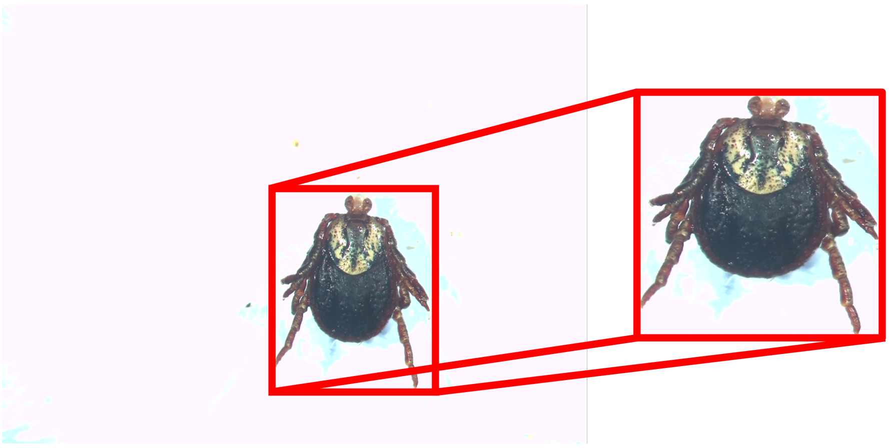

Fig. 7.6: Image processing method for automatically cropping the image to the region of interest

Found issues: some images contain ticks with broken legs, resulting in an abnormally large bounding rectangle.

Machine Learning Module

Transfer learning was used to obtain a successful classification with the very limited of images available. Several deep networks, pretrained on the ImageNet-1k dataset, are used as the backbone. They are described in the next section.

A support vector machine is trained from scratch to adapt the precious networks to the considered classes. In particular, it replaces their last layer.

A summary of the whole process can be found in Fig. 7.7.

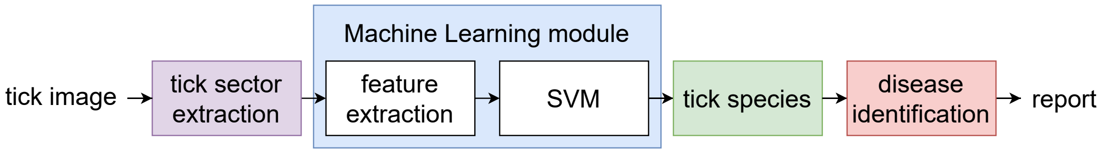

Fig. 7.7: Process for camera-based tick disease identification

7.2.3 Experimental Results and Discussion

The performance of the different deep networks regarding their accuracy and their mean squared error can be found in Tab. 7.1.

Tab. 7.1: Results of the methods with different backbones

	accuracy score	mean squared error
DenseNet201	**0.47**	**0.53**
ResNet50	0.53	0.63
ResNet50V2	**0.32**	**0.68**
MobileNet	0.50	1.21
EfficientNetB1	**0.76**	**0.32**
EfficientNetB2	0.82	0.18
EfficientNetB3	**0.89**	**0.26**
EfficientNetB4	0.76	0.32
EfficientNetB5	**0.79**	**0.21**
EfficientNetV2S	0.29	1.50

It is clear that EfficientNets provide the best performance in our application. In particular, EfficientNetB3 stands out over every other network — 89% accuracy.

The confusion matrix of the EfficientNetB3-based model can be found in Fig. 7.8.

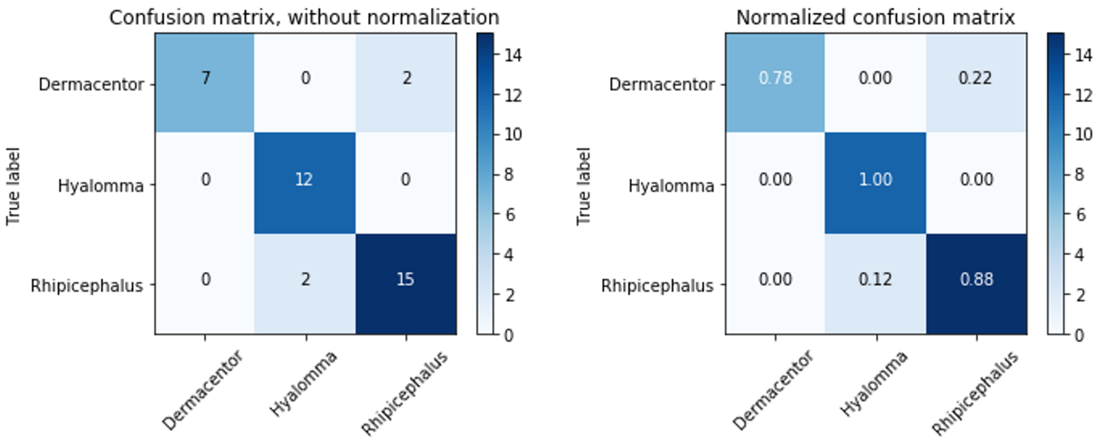

Fig. 7.8: Confusion matrix without normalization (left) and normalized confusion matrix (right) of the best developed model

As it can be seen, most of the results lay within the main diagonal of the confusion matrix, a signal of a positive behaviour. Only 10.5% of the images are mistakenly classified, a good outcome of a freshly proposed application based on real-world data.

All Hyalomma ticks were successfully classified, indicating that this method could be a useful tool in the fight against the Crimean–Congo hemorrhagic fever, which could reduce its current expansion phase, lower its high fatality rate and increase the public health levels of regions with constant outbreaks (such as Africa, Asia, the Balkans, the Middle East, and Russia). The recall of the Rhipicephalus was 88.2% and the recall of the Dermacentor 77.8%. Therefore, a low type II error was obtained in the most critical case, and a low type II error in other cases.

7.2.4 Conclusions and Future Work

A novel method for identifying the possible diseases transmitting by a tick bite has been proposed. Several deep learning alternatives were considered and compared, resulting in the election of the EfficientNetB3 + SVM architecture. This model obtains strong classification accuracy (89%) and opens the door to real-world applications, with the potential of achieving an early-diagnostic which can save many lives.

The method only requires a flat surface, a piece of paper and a HD camera, which can be easily obtained by many doctors in most of the developed world. Then automatic image preprocessing and classification is performed without the need of human intervention, bringing this solution within reach of non-technical users.

Tick-borne diseases are a high health risk for society, as ticks are abundant in woodlands for a long season every year in many parts of the world. An early and accurate diagnosis can ease this health challenge, which is now possible thanks to the current technology and research efforts of the scientific community.

A promising future research line would be to recreate the performed experiments with images taken with a smartphone, and verify whether robust results are obtained. This could be a big leap forward to bring this technology to the masses. Moreover, a larger dataset could greatly benefit the model and its reliability.

Chapter 8

Conclusions

VNiVERSiDAD
D SALAMANCA

Chapter 8

Conclusions

This research work has presented several algorithms and new applications for improvements across SCs, mostly related to healthcare, automation and autonomous driving. A growing need for efficient ML solutions, capable of running in environments with low computational resources has been identified, as well as a demand for novel early-diagnostic methods and for lowering the cost of repetitive activities.

The world is experiencing the ever-growing amount of data continuously generated by humans, much of which is stored waiting to be processed. ML, and in particular deep learning, have become integral methods to solve this situation and to improve life quality and economic performance. Moreover, modern hardware has allowed for a widespread use of AI applications, which are already present in the smart phones of most of the world population. New, efficient algorithms should be developed to further expand its use. The outcome of this research work are such algorithms.

The main contributions achieved in this Ph.D. dissertation are:

- We have performed an analysis of the particular problems of computer vision within SCs revising the state of the art, tackling the main problems which deep learning could solve. This has been used to establish the initial requirements for the design and development of the solution.
- We have studied the techniques and technologies being used in SCs, which has allowed to improve existing solutions and to propose new applications. This has led to the development of new deep learning structures capable of efficient and effective training, as well as provided useful results.
- We have used formal mechanisms to solve the problem of optimising network behaviour in vision tasks. In particular, the proposals have increased the efficiency of the methods used for automatic decision-making in SCs.
- We have demonstrated the remarkable performance of the proposals in realistic and challenging scenarios. Real-life data has been evaluated and widely-used

datasets have been considered. The obtained results provide strong evidence of improvements regarding the networks performance and computational efficiency.

The research presented in this doctoral dissertation validates the vision-based models proposed for SCs. They have a high potential, as an evident expansion of SCs is currently taking place, and EVs and smartphones are democratizing the use of deep learning-based applications. Moreover, the initial hypothesis is confirmed: it has been proven that the current techniques used for computer vision in SCs can be enhanced (Sec. 4.2), optimised (Sec. 7.2) and accelerated (Sec. 3.2) by applying original deep learning techniques. Novel methods regarding these aspects have been presented.

Furthermore, all the research questions which motivated the present research have been answered:

- *What time-consuming tasks could be automatised?*

  This dissertation proposes a ML-based method for tick identification and the prediction of potential transmitted diseases Sec. 7.2. It removes the need for an expert to manually classify the specimen and it provides a prompt diagnosis. Also, Sec. 6.2 proposes a ML-based method for automatic house categorization which could be used by cadastres, and Sec. 5.2 proposes a ML-based method for automatic medical image segmentation.

- *Are modern vision algorithms optimised to lower their computational cost?*

  Results seem to indicate that their accuracy was heavily optimised, but their computational cost still has much room for improvement. In particular, Sec. 3.2 proposes alterations to the well-known Transformer architecture which, on average, lower their number of trainable parameters by 10.81% while maintaining or improving their accuracy. Also, Sec. 4.2 proposes modifications to the PDAN architecture (related to HAR), achieving a reduction of up to 34.87% in the number of trainable parameters while slightly improving the accuracy of the model.

- *Will the public sector become more data-centred in the future?*

  New data-driven techniques have a clear potential to improve the functioning of administrations in SCs, and their managers seem to have noticed it in the last few years (Retuerta et al. [2018]). Sections 6.2, 7.2, 5.2 provide solutions to real-life problems which the public sector is currently facing, and which could be solved using more data-centred methods. Therefore, the public sector currently has the tools and the opportunity to make extensive usage of data in its operations, an occasion it is likely to take advantage of.

- *What time-sensitive tasks are currently burdened by manual execution?*

  Tick species recognition and organ delineation for esophageal cancer patients are the most prominent cases identified in this dissertation. In both cases, an early diagnosis is known to lower the risk of mortality in patients suffering from this illness, however, manual techniques currently in use can take up to one week, due to high-demand periods and organization issues.

- *Could the reliability of automatic video understanding be improved?*

  Yes. Section 4.2 describes how to improve the accuracy of a state-of-the-art network for HAR, better locating the temporal component of actions within a video and identifying the correct activities.

- *Could the data of the administrations already be successfully used in machine learning?*

  Yes. Sections 7.2 and 6.2 use a real dataset of the local authorities, achieving successful results and validating the potential of applying known ML methods.

The main line of research for the future, regarding the presented work, is to include new mathematical techniques and modern deep learning structures to optimize the behaviour of algorithms, just as studying the internal working of their different layers to better understand the usefulness of each of their components. Crucial components should be maintained, and less important structures should either be altered or completely removed, ensuring a balanced performance. Finally, the last line of research which has been opened is related to Transformers. State of the art architectures will be modified with the proposed alterations, in an effort to challenge the current state of the art.

On the basis of the obtained insights and results, possible lines of research could explore the following aspects:

- Testing and validation. Comprehensive tests are needed for a detailed evaluation of the proposed algorithms in several dimensions, such as real-life performance, empirical validation, usefulness of the applications, balanced functioning, real-time application, quality output, etc. The obtained results would allow to further optimise the methods and to achieve robust algorithms and systems.

- Solving new practical problems. The proposed methods could prove themselves useful in a variety of applications. The use of new variables and new use cases could verify their correct behaviour. An algorithm which adapts itself well to new problems is well needed, as it would be a step from *narrow AI* towards the much-wanted *broad AI*.

- Incorporation of new techniques. New mathematical techniques and novel deep learning structures have the potential to improve and optimize the current algorithms used in SCs. They should be explored to increase effectiveness and efficiency in SCs.

In closing, deep learning methods can actively improve the computer vision field and its applications to SCs. These three emerging fields enjoy great popularity at the moment, and new models are constantly coming and going. This dissertation systematically studies the field, and proposes methods to improve several key points which, statistically, could amount to little more than one drop in a vast ocean. Yet what is any ocean, but a multitude of drops?

Appendix A

Publications and related works

Appendix A

Publications and related works

A.1 Introduction

This doctoral dissertation comprises research in the field of deep learning, which has been conducted over a period of three academic years, within the BISITE research group of the University of Salamanca. The knowledge acquired during this period has made it possible to develop the current doctoral dissertation.

During the course of the candidate's Ph.D. studies, an FPI grant has been awarded to him as part of the project "Towards sustainable intelligent mobility supported by multi-agent systems and edge computing (InEDGEMobility)", Reference: RTI2018-095390-B-C32, financed by the Ministry of Science and Innovation (MICINN), the State Research Agency (AEI) and the European Regional Development Fund (FEDER).

A list of some of the most relevant publications related to this work, which have been published since enrolment in the first year of doctoral studies, can be found below. The list encompasses publications in international journals and in book chapters, sorted in chronological order. Afterwards, the R&D projects in which the author has participated as part of his Doctoral studies are described, as well as his research stays and educational activities.

A.1.1 Papers in International Journals

- Chamoso, P., Bartolomé, A., **García-Retuerta, D.**, Prieto, J., & De La Prieta, F. (2020). Profile generation system using artificial intelligence for information recovery and analysis. Journal of Ambient Intelligence and Humanized Computing. (JCR, Q1)
- Corchado, J., Chamoso, P., Hernández, G., Gutierrez, A., Camacho, A., González-Briones, A., Pinto-Santos, F., Goyenechea, E., **Garcia-Retuerta, D.**,

Alonso-Miguel, M., & others (2021). Deepint.net: A rapid deployment platform for smart territories. Sensors, 21(1), 236. (JCR, Q1)

- **García-Retuerta, D.**, Chamoso, P., Hernández, G., Guzmán, A., Yigitcanlar, T., & Corchado, J. (2021). An Efficient Management Platform for Developing Smart Cities: Solution for Real-Time and Future Crowd Detection. Electronics, 10(7), 765. (JCR, Q2)

- **García-Retuerta, D.**, Rivas, A., Guisado-Gámez, J., Antoniou, E., & Chamoso, P. (2021). Reputation System for Increased Engagement in Public Transport Oriented-Applications. Electronics, 10(9), 1070. (JCR, Q2)

A.1.2 Book Chapters

- **García-Retuerta, D.** (2020). AI-Based Proposal for Epileptic Seizure Prediction in Real-Time. In International Symposium on Ambient Intelligence (pp. 289–292).

- **García-Retuerta, D.**, Casado-Vara, R., Martin-del Rey, A., Prieta, F., Prieto, J., & Corchado, J. (2020). Quaternion neural networks: state-of-the-art and research challenges. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 456–467).

- **García-Retuerta, D.**, Casado-Vara, R., Calvo-Rolle, J., Quintián, H., & Prieto, J. (2020). Deep Learning for House Categorisation, a Proposal Towards Automation in Land Registry. In International Conference on Hybrid Artificial Intelligence Systems (pp. 698–705).

- **García-Retuerta, D.**, Casado-Vara, R., & Rodríguez, S. (2021). Transfer Learning for Arthropodous Identification and its Use in the Transmitted Disease Diagnostic. In Practical Applications of Agents and Multi-Agent Systems (pp. 253–260).

A.1.3 Project Participation

- **Title**: InEDGEMobility: Movilidad inteligente y sostenible soportada por Sistemas Multi-agentes y Edge Computing.

  Founder: Spanish ministry of Science, Innovation and Universities. 01/2019 - 12/2021. (RTI2018-095390-B-C32)

  PI: Juan Manuel Corchado Rodríguez, Sara Rodríguez González.

- **Title**: HERMES: Hybrid Enhanced Regenerative Medicine Systems.

  Founder: European Commission. 02/2019 - 01/2024. (H2020-FETPROACT-2018-2020)

  PI: Gabriella Panuccio.

- **Title**: CHROMOSOME: Change and Analysis of Consumer Behavior at Smart Homes via Social Machine.

  Founder: Salamanca City of Culture and Knowledge Foundation. 09/2018 - 08/2021.

  PI: Javier Prieto Tejedor.

A.1.4 Research Stays

Below is a list of the stays the candidate has carried out abroad in renowned higher education institutions, at which research work related to this doctoral thesis has been conducted by the Ph.D. candidate.

1. Catholic University of Louvain (UCLouvain).

   - Host Research Group: Molecular Imaging, Radiotherapy and Oncology (MIRO).
   - Location: Belgium.
   - Period: July 2021 - October 2021.
   - Activities undertaken: Deployment of an AutoML framework for U-Net (an image-to-image convolutional neural network) which was then used to predict treatment doses in radiation oncology on the basis of CT images and organ contours. Results in Chapter 5.2.

2. Tbilisi State University (TSU).

   - Host Research Group: Ilia Vekua Institute of Applied Mathematics (VIAM).
   - Location: Georgia.
   - Period: November 2021 - February 2022.
   - Activities undertaken: Deployment of new deep learning architectures for vision-based Human Activity Recognition. The architectures modelled complex temporal relations in densely annotated video streams, which contributed to the advancement of the state of the art. Results in Chapter 4.2.

A.1.5 Performed Educational Activities

1. Research seminars: Attendance to three research seminars, 9 hours in total.
2. Follow-up meetings regarding projects, work or research results: eight meetings with the Ph.D. supervisors and other fellow Ph.D. candidates to share the progress of the research, 12 hours in total.
3. Methodological, specialised or practical training courses: Completion of three courses, 16 hours in total.
4. Attendance to national or international congresses: Attendance to eight international conferences, 160 hours in total.
5. Scientific publications: 14 works as described in the previous sections.
6. Research stays in other research centres: Two internships in renowned foreign research groups, over six months in total.
7. Mobility actions and criteria: Two internships in renowned foreign research groups and attendance to eight international conferences.

Appendix B

Dissertation Summary in Spanish

VNiVERSiDAD
Đ SALAMANCA

Appendix B

Dissertation Summary in Spanish

Este anexo recoge la traducción al español del **título**, **resumen**, **índice**, **introducción** y **conclusiones** de la tesis doctoral “*Deep Learning for Computer Vision in Smart Cities*”, de David García Retuerta, así como un **resumen significativo** de los capítulos centrales. Todos los detalles pueden ser encontrados en el trabajo original, escrito en inglés.

Autor:

Firmado por DAVID GARCIA con un
certificado emitido por AC FNMT
Usuarios

David García Retuerta

Directores:

RODRIGUEZ
GONZALEZ
SARA -
70864126E

Firmado digitalmente
por RODRIGUEZ
GONZALEZ SARA -
70864126E
Fecha: 2022.06.22
19:12:01 +02'00'

Dr. Sara Rodríguez González

Firmado digitalmente por
CHAMOSO SANTOS PABLO -
70888952P
Fecha: 2022.06.21 13:59:37 +02'00'

Dr. Pablo Chamoso Santos

Contents

Resumen

**Aprendizaje profundo para la visión
por ordenador en ciudades inteligentes**

por David García Retuerta

La era de la información ha provocado un rápido cambio de la industria hacia una economía basada principalmente en la tecnología de la información. Según estudios recientes, 74 zettabytes (ZB) de datos han sido generados, capturados y reproducidos en el mundo en 2021, con el contenido visual representando el 82% del tráfico de Internet. Esta cifra ha crecido debido a la pandemia de COVID-19, y se espera que siga aumentando hasta alcanzar los 149 ZB en 2024. El procesamiento de esta ingente cantidad de información es uno de los principales retos científicos de nuestro tiempo. Este contexto ha contribuido a la aparición del aprendizaje automático y de dos nuevos paradigmas relacionados: el big data y el aprendizaje profundo. Estas disciplinas aprovechan los métodos de optimización matemática, la bioinspiración y las tarjetas gráficas más modernas para gestionar grandes conjuntos de datos.

Ciudades en todo el mundo se han adaptado para hacer uso de los nuevos datos disponibles, promocionándose a sí mismas como inteligentes. Las *Smart Cities* además pretenden atraer a nuevos ciudadanos, interesar a inversores externos e integrar la tecnología en su funcionamiento diario. Algunas de las motivaciones clave de los proyectos Horizon y de los fondos NextGenerationEU son precisamente hacer que las ciudades sean más digitales, más verdes, más sanas y más resilientes; objetivos que pueden beneficiarse en gran medida de la inteligencia artificial. Se han identificado cinco líneas de actuación: Smart Mobility, Smart Environment, Smart People, Smart Living y Smart Economy.

Esta tesis se centra en las aplicaciones de visión del aprendizaje profundo en el ámbito de las SC, identificando áreas de actuación y proponiendo soluciones. Como resultado, se ha contribuido al estado del arte y se ha propuesto una nueva solución para cada una de las líneas de actuación identificadas.

Se han diseñado y evaluado dos modelos con especial atención a la eficiencia y la escalabilidad, y se desarrolló y probó un tercero con el objetivo de conseguir precisión

en un entorno de altos recursos. Además, se elaboró un método para automatizar retos sanitarios cruciales, haciendo posible el diagnóstico precoz; y se creó un método para automatizar la categorización catastral urbana.

Capítulo 1

Introducción

1.1 Introducción

De acuerdo a informes recientes, a lo largo de 2021 se han generado, capturado y replicado 74 ZB de datos en el mundo (Aundhkar and Guja [2021]). Entre ellos, se espera que el vídeo represente el 82% del tráfico de Internet en 2021 (Cisco [2016]). Además, la pandemia de COVID-19 ha ampliado la cantidad de datos generados a nivel mundial, que se espera que siga aumentando, hasta alcanzar los 149 ZB en 2024 (Turi and Li [2021]). El procesamiento eficiente de esta gran cantidad de información es uno de los principales retos de investigación a los que se enfrenta la comunidad científica en estos momentos.

Además de los datos generados directamente por la interacción humana con los dispositivos electrónicos, una parte notable del flujo de información actual es generada por los 18.000 millones de dispositivos IoT conectados que se calcula (Simiscuka and Muntean [2021]). Las SC han aprovechado la oportunidad que suponen estas nuevas fuentes de datos para atraer a nuevos ciudadanos, interesar a inversores externos, integrar la tecnología en su funcionamiento diario y generar crecimiento económico.

Precisamente, algunas de las motivaciones clave de los proyectos Horizon y los fondos NextGenerationEU son hacer que las ciudades sean más digitales, más verdes, más sanas y más fuertes; objetivos que el ML tiene el potencial de facilitar. Los desarrolladores de SC han estado trabajando en estos campos recientemente. Para centrar su atención, se han propuesto cinco mercados verticales: Smart Mobility, Smart Environment, Smart People, Smart Living y Smart Economy. Un desarrollo equilibrado de la metrópolis debería incluirlos todos y no dejar ninguno atrás. En consecuencia, este trabajo de investigación propone un método novedoso para cada uno de los mercados verticales anteriores.

En esta tesis se presentan métodos novedosos de aprendizaje profundo, centrados en aplicaciones de visión en el ámbito de las SC. Se identifican lagunas de investigación teóricas y prácticas, y se presentan soluciones adecuadas. Los métodos propuestos utilizan ampliamente los siguientes métodos de aprendizaje profundo: redes convolucionales, transformers, U-shaped networks y transfer learning.

La hipótesis planteada en esta tesis doctoral es que las SC pueden beneficiarse en gran medida de los algoritmos innovadores de visión por ordenador, que se basan en técnicas de aprendizaje profundo. La hipótesis tiene dos dimensiones distintas: la aplicación de métodos conocidos a problemas existentes (dimensión práctica) y la optimización de los algoritmos actuales de vanguardia (dimensión teórica). Los resultados indican que los métodos propuestos son rentables y eficientes, logrando una precisión justa en nuevas aplicaciones reales y mejorando el rendimiento de las arquitecturas conocidas.

El presente capítulo está organizado como sigue: la descripción del problema y la motivación se introducen en la Sec. 1.2. En la Sec. 1.3 se muestran las hipótesis de investigación y los objetivos de esta tesis, y en la Sec. 1.4 se presenta la metodología. Por último, la Sec. 1.5 presenta la estructura de la tesis.

1.2 Descripción del problema y motivación

Los planteamientos innovadores en las zonas urbanas están desencadenando actualmente la " Smart City Revolution ". La creciente urbanización del mundo está ejerciendo presión sobre las infraestructuras existentes de las metrópolis, cuya calidad de vida y desarrollo socioeconómico está sufriendo múltiples y nuevas cuestiones por parte de diversos actores. Las ciudades pueden hacer frente a estos retos garantizando un alto nivel de participación de los ciudadanos, asegurando un uso generalizado de las aplicaciones basadas en Internet, impulsando la colaboración entre sus instituciones y muchas otras iniciativas. Si se agrupan en mercados verticales, algunas de las oportunidades más importantes son (Kumar [2020]):

- **Economía inteligente**. Adoptar datos abiertos y sistemas inteligentes puede impulsar la innovación, el desarrollo de las empresas y reducir los costes de administración. Una start-up que desarrolle una solución para la ciudad puede exportar esa tecnología a todo el mundo.
- **Personas inteligentes**. Los ciudadanos comprometidos pueden informar de las averías de las infraestructuras, de los objetos del suelo que necesitan un mantenimiento urgente y de cualquier tipo de riesgo para la seguridad.
- **Medio ambiente inteligente**. Un mejor uso de los recursos naturales y la protección del medio ambiente pueden afectar positivamente a la salud de los ciudadanos e impulsar la sostenibilidad de la ciudad.
- **Movilidad inteligente**. Una mayor eficiencia en todos los transportes puede mejorar la calidad del aire, disminuir los tiempos de desplazamiento y reducir los índices de mortalidad en carretera. Los aparcamientos inteligentes pueden reducir en gran medida el tráfico de búsqueda en las calles.
- **Vida inteligente**. Los dispositivos IoT pueden mejorar la atención sanitaria de las personas, analizando las constantes vitales y emitiendo alertas tempranas precisas.

Teniendo en cuenta toda la información anterior, más el hecho de que las cámaras se han vuelto omnipresentes en todos los rincones del mundo desarrollado; la visión por ordenador emerge como una fuente de conocimiento excepcional para una amplia variedad de aplicaciones. De hecho, la combinación de herramientas de visión por computador y de ciencia de datos tiene muchas aplicaciones potenciales que podrían utilizarse para promover un desarrollo eficiente y sostenible en las SC.

En la actualidad, la comunidad científica está dedicando gran cantidad de tiempo y fondos a encontrar nuevas aplicaciones del ML y a mejorar los métodos existentes. Tanto es así que se considera una pieza clave de la cuarta revolución industrial, es una palabra de moda año tras año desde la década de 2010, e importantes actores han afirmado que la llamada “Revolución de la IA” ya ha comenzado (Harari [2017]). El ML aporta muchas ventajas a la sociedad, las mayores de las cuales son:

1. **Una amplia aplicación en todos los campos**. Representa un gran avance en la forma en que los ordenadores pueden aprender de los datos. Se está haciendo omnipresente en las redes sociales, los sistemas de recomendación y los asistentes personales, entre otras aplicaciones.

2. **Mayor automatización de tareas**. El ML es una tecnología que puede facilitar algunos trabajos imitando el comportamiento humano. Por ejemplo, las empresas de vigilancia ya no necesitan un operador humano para comprobar los vídeos captados por sus cámaras, ya que existen algoritmos de ML capaces de comprender lo que ocurre en ellos.

3. **Fácil y rápida identificación de tendencias y patrones**. Los cambios en la distribución de una variable considerada pueden detectarse, primero como una anomalía y luego como una tendencia próxima. Esta característica ha contribuido a la adopción del ML en las empresas de inversión financiera.

4. **Mejora constante**. Constantemente se presentan nuevas arquitecturas y enfoques, y cada 5-10 años aparecen nuevas propuestas revolucionarias. Desde 2010, las redes convolucionales, los bloques LSTM y los transformers han sido los responsables de los avances en este campo.

5. **Capacidad para tratar datos multidimensionales y de múltiples variedades**. Los proyectos de data science suelen incluir datos procedentes de diferentes sensores y fuentes, que una red neuronal artificial tiene la capacidad de utilizar como entrada.

Algunos sectores que pueden beneficiarse enormemente del ML son:

- **Mercados financieros**. En los últimos 10 años, el sector financiero ha dedicado muchos recursos a utilizar modelos complejos en la previsión de valores. Por ejemplo, en los últimos años, se ha hecho común el uso de NLP para predecir el mercado de valores sobre la base de las noticias (Kim et al. [2014]).

- **Industria automovilística**. En 2018, el 78% las empresas de automoción invirtieron en habilidades y formación relacionados el ML, utilizando en gran medida el ML en sus campañas de marketing (Schrage and Kiron [2018]). Por ejemplo, Tesla lanzó su *Autopilot* en 2015, anunciando su intención de una futura actualización para ofrecer el nivel 5 de SAE (conducción autónoma completa).

- **Sector sanitario**. El ML tiene el potencial de acelerar el análisis de las grandes cantidades de datos recogidos para cada paciente. En particular, se están utilizando herramientas de apoyo a la decisión clínica para procesar grandes conjuntos de datos con el fin de identificar una nueva enfermedad.

- **Agricultura**. Mediante la ciencia de los datos y los modelos de inteligencia artificial se pueden obtener recomendaciones y conocimientos sobre los cultivos. Se pueden utilizar algoritmos para aprovechar al máximo los periodos de precosecha, cosecha y poscosecha.

- **Industria militar**. Las decisiones a vida o muerte en el campo de batalla requieren de un breve tiempo de consideración antes de actuar, una tarea de la que podría encargarse el ML. Simular batallas y maximizar la probabilidad victoria es una opción que interesaría a la mayoría de los ejércitos.

- **Redes sociales y motores de búsqueda**. Las grandes empresas tecnológicas se basan en gran medida en el ML para ofrecer recomendaciones de amigos o resultados de búsqueda. Por ejemplo, la red CLIP (Radford et al. [2021]) puede predecir la existencia de cualquier etiqueta dentro de una imagen dada, un algoritmo perfecto para la búsqueda inversa de imágenes.

- **Gran parte de las disciplinas de ingeniería**. Con el ML siendo en una tendencia durante las últimas décadas, la mayoría de los campos técnicos están empezando a integrarlo en su investigación.

- **Ciencias sociales**. Su actual era de abundancia de datos se adapta perfectamente al paradigma de ML, donde los enfoques difusos pueden modelar de cerca el comportamiento de las sociedades.

- **Industrias del arte y la publicidad**. Se están creando nuevos estilos artísticos mediante pinturas, composiciones y escritos realizados por máquinas. Además, los contenidos publicitarios pueden llegar a un público más específico que nunca gracias al ML.

Curiosamente, a pesar de todos los recientes y revolucionarios avances en este campo, todavía hay un gran margen de mejora en el ML. En particular, sus aplicaciones para la visión por ordenador han cobrado impulso desde la popularización del aprendizaje profundo, hace menos de 20 años, y su uso en las SC es todavía limitado.

Por lo tanto, existe una gran oportunidad para diseñar y probar nuevos métodos que optimicen este campo. Además, existen vastas aplicaciones potenciales inexploradas que podrían afectar y complacer a una gran parte de los ciudadanos, y que podrían optimizar el gasto de recursos de sus administraciones.

Las siguientes preguntas motivaron la investigación:

- ¿Qué tareas que consumen mucho tiempo podrían automatizarse?
- ¿Están optimizados los algoritmos de visión modernos para reducir su coste computacional?
- ¿El sector público se centrará más en los datos en el futuro?
- ¿Qué tareas sensibles al tiempo se ven actualmente sobrecargadas por la ejecución manual?
- ¿Podría mejorarse la fiabilidad de la comprensión automática de vídeos?
- ¿Podrían utilizarse con éxito los datos de las administraciones en ML?

Esta tesis trata de encontrar respuestas a las preguntas anteriores, tanto utilizando métodos de ML para resolver ciertos retos de las SC, como mejorando el estado del arte de los algoritmos relacionados.

Los métodos de ML seleccionados se centran en la visión por ordenador, debido a la reciente explosión del contenido de vídeo y a la creciente popularidad del hardware de filmación. Además, los recientes avances en la detección de objetos, el reconocimiento de caras, acciones y actividades; y la estimación de la pose humana han superado muchas de sus limitaciones pasadas, creando una sólida base científica que puede utilizarse para resolver varios problemas de la vida real (Voulodimos et al. [2018]).

1.3 Hipótesis y objetivo

Este trabajo de investigación aporta diferentes soluciones a las lagunas de investigación identificadas en el estado del arte existente respecto a las aplicaciones de aprendizaje profundo para la visión por ordenador en el contexto de las SC. En particular:

La hipótesis inicial de este trabajo de investigación es que es posible mejorar, optimizar o acelerar los actuales algoritmos y técnicas utilizadas para la visión por computador en SCs, aplicando novedosos métodos de aprendizaje profundo.

Se han identificado varias lagunas de investigación en el diseño de algoritmos de visión por computador y en sus aplicaciones, que podrían aumentar la tasa de supervivencia de los pacientes y minimizar el coste de las soluciones de última generación. Modificamos los modelos de aprendizaje profundo existentes para hacerlos más eficientes, precisos y eficaces. Además, desarrollamos nuevas soluciones para automatizar tareas que actualmente se realizan de forma manual. Por tanto, el resultado final de este trabajo es impulsar el campo de la visión por computador y, en particular, sus aplicaciones relacionadas con las SC.

El objetivo principal de esta tesis es mejorar la precisión y disminuir los requerimientos computacionales del modelo de visión por computador existente, así como aplicarlos para lograr soluciones efectivas en SCs.

Esta tesis incluye varias áreas de investigación dentro del campo de la visión por computador: segmentación de imágenes médicas, reconocimiento de actividades humanas basado en cámaras, clasificación de imágenes y detección de anomalías. Se han elegido por su relevancia y potencial de mejora. Antes de desarrollar nuevos métodos, se ha estudiado el estado del arte del campo de la visión por ordenador, las áreas de investigación mencionadas, se han identificado las lagunas de investigación y se han resuelto los problemas encontrados.

1.4 Metodología

Una metodología formal y bien estructurada es necesaria para realizar una investigación de calidad que produzca resultados válidos, y para garantizar que se obtiene un resultado válido en cada paso de la investigación. El método *investigación-acción* (Reason and Bradbury [2001]) fue elegido y aplicado en el presente trabajo. Destaca por estar orientado a la acción y al cambio, lo que permite al investigador centrarse en problemas bien definidos para producir nuevos conocimientos basados en otros trabajos durante un periodo de tiempo. Se ha convertido en un enfoque habitual de la investigación empírica. El proceso es el siguiente (1) identificación del problema real, (2) estudio de las posibles hipótesis, selección de una y desarrollo de una propuesta, (3) verificación de la hipótesis seleccionada, y (4) sacar conclusiones tras la evaluación de los resultados obtenidos. En nuestro caso:

1. Identificación y descripción de las características del problema planteado. Se definen las características de la infraestructura de SC y sus algoritmos utilizados y se proponen todas las hipótesis posibles.

2. Estudio de las posibles hipótesis, selección de una y desarrollo de una propuesta. Se realiza un estudio incremental del estado del arte y se analizan las hipótesis anteriores. Se obtiene así un marco teórico y se seleccionan las hipótesis más prometedoras. A continuación, se presenta una propuesta con sólidos fundamentos científicos.

3. Verificación de las hipótesis seleccionadas. A continuación se lleva a cabo un diseño iterativo y progresivo de una solución. A continuación se reúnen los datos visuales relevantes disponibles, se implementan varios componentes que se ocupan de los diferentes aspectos del problema y se combinan en un modelo integral.

4. Sacar conclusiones tras la evaluación de los resultados obtenidos. El nuevo modelo se pone a prueba para analizar su comportamiento: se evalúan sus componentes, su funcionalidad y cada una de sus iteraciones y se obtiene un gran número de resultados brutos. Dichos resultados se analizan para realizar la formulación de conclusiones.

Paralelamente a las etapas mencionadas, se llevó a cabo una continua difusión de conocimientos, resultados y experiencias con la comunidad científica. Este proceso se materializó en forma de publicaciones en revistas científicas, asistencia a congresos internacionales y presentación de trabajos.

1.5 Estructura de la tesis doctoral

Esta tesis doctoral se divide en ocho capítulos y un apéndice. Su estructura se describe a continuación.

El capítulo 1 proporciona una introducción a la investigación realizada y a la tesis. En concreto, describe la motivación de la investigación y los problemas que se pretenden resolver. También se presentan las hipótesis, los objetivos y la metodología que han dado lugar a esta tesis.

En el capítulo 2 se hace una revisión del estado del arte y se describen algunos conceptos previos que el lector necesitará en los siguientes capítulos. Se describen en detalle el ML, la visión por ordenador, los SC y sus respectivos subcampos; así como la relación de los tres conceptos anteriores.

El capítulo 3 está relacionado con el mercado vertical de la ”Movilidad Inteligente”. Primero se presenta una introducción a los conceptos relacionados con ella y su importancia, y después un caso de uso particular. La red de procesamiento de lenguaje natural y visión por ordenador más prometedora de los últimos años se modifica para reducir su coste computacional y sus necesidades de recursos, al tiempo que aumenta ligeramente su precisión. Tiene el potencial de aumentar la difusión de los vehículos de autoconducción y acercar la visión por ordenador a las técnicas médicas. Este capítulo es más largo que otros debido a las extensas pruebas realizadas para validar las mejoras de la propuesta.

El capítulo 4 está relacionado con el mercado vertical de las ”personas inteligentes”. Se presenta primero una introducción a los conceptos relacionados y su importancia, y después un caso de uso particular. Se desarrolla una red de atención para mejorar el reconocimiento de la actividad humana basado en cámaras, una tarea central con una amplia variedad de aplicaciones. El rendimiento de la nueva red permite a los desarrolladores ejecutarla de forma fiable incluso con un hardware modesto.

El capítulo 5 está relacionado con el mercado vertical de la ”vida inteligente”. Primero se presenta una introducción a los conceptos relacionados con ella y su importancia, y después un caso de uso particular. Se utiliza un marco de trabajo automatizado de ML para realizar la segmentación clínica de volúmenes y órganos en una aplicación relacionada con la terapia de protones para pacientes con cáncer de esófago. Este caso de uso presenta una mejora sobre las técnicas anteriores encontradas en el estado del arte, consiguiendo que la planificación del tratamiento esté un paso más cerca de la automatización en casos clínicos reales.

El capítulo 6 está relacionado con el mercado vertical de la ”economía inteligente”. Primero se presenta una introducción a los conceptos relacionados con ella y su importancia, y después un caso de uso particular. Se utiliza el aprendizaje de transferencia para categorizar automáticamente la tipología constructiva de los edificios de viviendas en función de su fachada. Se presenta un método para automatizar este aspecto administrativo y se formulan recomendaciones relacionadas con el proceso de toma de datos. Este caso de uso tiene el potencial de reducir los requisitos económicos y humanos para gestionar un catastro.

El capítulo 7 está relacionado con el mercado vertical del ”entorno inteligente”. Primero se presenta una introducción a los conceptos relacionados con él y su importancia, y después un caso de uso particular. Se desarrolla un método para obtener un reconocimiento de la especie y la posible infección de la enfermedad tras la picadura de una garrapata. Tiene el potencial de salvar vidas gracias a un diagnóstico rápido y de ayudar a difundir esta característica médica en regiones poco pobladas.

En el capítulo 8 se extraen las conclusiones del trabajo de investigación y se enumeran las aportaciones de los desarrollos al estado de la técnica. También se presentan las futuras líneas de investigación que se han abierto en el transcurso de los estudios de doctorado.

En el Apéndice A se recoge una lista de los resultados tangibles de este programa de doctorado. En primer lugar, la lista de publicaciones científicas que tuvieron lugar durante el programa de doctorado por el candidato al doctorado. La lista incluye capítulos de libros y revistas internacionales que figuran en el Journal Citation Reports (JCR). En segundo lugar, también se mencionan los proyectos de investigación en los que ha participado el doctorando y que, en cierta medida, han contribuido al desarrollo científico de la investigación. Por último, se mencionan sus estancias de investigación en instituciones de investigación de renombre en el extranjero.

Por último, se presenta una lista con todas las referencias bibliográficas utilizadas y referenciadas durante esta tesis.

Capítulo 2

Fundamentos

Este capítulo introduce los conceptos y técnicas fundamentales que serán utilizados en los siguientes capítulos. En primer lugar, se realiza una presentación del campo del Machine Learning y sus algoritmos relacionados, incidiendo particularmente en las máquinas de soporte vectorial, el aprendizaje profundo y el transfer learning. Tras ello, el campo de la visión por ordenador es presentado y dos de sus aplicaciones que más potencial han mostrado en los últimos años son descritas: los transfomers y las redes de tipo U-Net. En tercer lugar, se aborda el tema de las Smart Cities, sus verticales y las posibilidades que sus datos abiertos conllevan. Por último, se extraer unas conclusiones de todo lo anterior para la tesis actual.

2.1 Machine Learning

En la última década se ha producido un gran aumento de los métodos basados en ML, que ha afectado a numerosos sectores tales como la conducción autónoma, la sanidad, las finanzas, la fabricación, la producción de energía, etc. El ML se considera un punto de inflexión para la humanidad en la actualidad, del mismo modo que la informática revolucionó el mundo en los años 80 y 90. A un nivel básico, el objetivo del ML es identificar patrones en los datos, extrayendo conocimientos que luego pueden ser utilizados para una amplia variedad de propósitos.

El ML forma parte del concepto más amplio de Inteligencia Artificial — un campo recientemente establecido basado en la optimización matemática, el aprendizaje estadístico, la minería de datos y la informática. Es una aplicación que proporciona a los sistemas la capacidad de aprender y mejorar automáticamente a partir de la experiencia sin ser programados explícitamente, es decir, de hacer lo que las personas hacen de forma natural: aprender utilizando ejemplos.

Uno de sus primeros y más robustos algoritmo fue la máquina de vectores de soporte (SVM). Se trata de un modelo de aprendizaje supervisado comúnmente utilizado en tareas de clasificación y regresión. Se considera uno de los clasificadores más robustos por el momento, y es una de las opciones preferidas para reentrenar la última capa de las redes preentrenadas (Lu et al. [2015]).

En los últimos años se ha popularizado una de sus más variantes más avanzadas: el aprendizaje profundo. Se trata de una familia de métodos de ML caracterizados por el uso de un elevado número de capas en la red. Las principales ventajas de las redes *profundas* son: las representaciones distribuidas y el poder de la profundidad (Eldan and Shamir [2016], LeCun et al. [2015]).

El transfer learning ha sido el encargado de traer todas las ventajas del aprendizaje profunto a los científicos con poderes de computación más limitados. El transfer learning es un enfoque típico de ML que permite transferir la capacidad de clasificación del modelo de aprendizaje del entorno predefinido a un entorno dinámico.

2.2 Visión por ordenador

La visión por ordenador es un campo científico interdisciplinar que estudia cómo los ordenadores pueden obtener conocimientos de alto nivel a partir de imágenes digitales o vídeos. También se inspira en la biología, ya que pretende reproducir el funcionamiento del sistema visual humano.

El aprendizaje profundo ha impulsado la investigación en una gran variedad de aplicaciones de visión por ordenador, ya que permite procesar grandes cantidades de datos de forma eficiente. Ha permitido automatizar etapas de tareas comunes relacionadas con la visión. El proceso completo de las tareas de visión por ordenador suele dividirse en: (1) adquisición de los datos, (2) procesamiento, (3) análisis y (4) comprensión.

Algunas de las aplicaciones más frecuentes de la visión por ordenador son: la medicina, actividad militar, los vehíclos autónomos, la visión artificial.

Desde la introducción de U-net en 2015, esta red ha revolucionado el campo de la segmentación de imágenes biomédicas, siendo sólo las arquitecturas basadas en Transformer las que consiguen el mejor rendimiento en algunos conjuntos de datos (Tang et al. [2021]). La red U-net se basa en un contracting path y un expansive path. Utiliza ampliamente las convoluciones, el max pooling y la función de activación ReLU (Ronneberger et al. [2015]).

El siguiente gran avance en el campo se debió a los modelos basados en *transformers*. Estos han revolucionado los campos del NLP, el procesamiento de datos de texto y la visión por ordenador desde su introducción en 2017 (Vaswani et al. [2017]), sustituyendo completamente a las RNNs y CNNs en la mayoría de las aplicaciones punteras. Su mayor grado de paralelización permite entrenar en conjuntos de datos más grandes de lo que antes era posible, lo que ha llevado al desarrollo de modelos muy exitosos como BERT (Devlin et al. [2018]) y las tres iteraciones de GPT (Brown et al. [2020]).

2.3 Smart Cities

Smart City es un concepto emergente en constante revisión que en la actualidad hace referencia a un tipo de desarrollo urbano basado en la sostenibilidad que es capaz de responder adecuadamente a las necesidades básicas de las instituciones, las empresas y los propios habitantes, tanto desde el punto de vista económico como operativo, social y medioambiental.

Además, el concepto de datos abiertos es una filosofía basada en aumentar la disponibilidad de los datos, sin ninguna restricción de derechos de autor, patentes u otros mecanismos de control. De este modo, cualquier persona puede hacer uso de ellos y puede impulsar la cooperación entre instituciones. Los datos abiertos tienen una importancia fundamental para las SC de todo el mundo, ya que impulsan la innovación, facilitan la extracción y el uso del conocimiento en muchos sectores y proporcionan un cambio positivo a sus complejos ecosistemas (Neves et al. [2020]).

La correcta utilización de estos y otros datos presenta un reto para las ciudades inteligentes. Los verticales son un enfoque prometedor para la clasificación de los diferentes aspectos de las SC y los territorios inteligentes, proporcionando información sobre los puntos fuertes y débiles de la infraestructura. Esta y otras formas de agrupar la información disponible permiten a una ciudad priorizar determinados servicios que pueden ser percibidos directamente por sus ciudadanos; aquellos que mejoran el nivel de vida en la ciudad (Tay et al. [2018]).

2.4 Conclusión

A pesar su gran cantidad de aplicaciones exitosas, el ML sigue siendo un campo joven con muchas oportunidades de investigación disponibles. Las nuevas posibilidades de captura de datos han impulsado los métodos de big data, lo que ha provocado que el aprendizaje profundo se convierta en un campo en auge en los últimos años.

Además, las nuevas tendencias de las zonas urbanas tecnológicamente modernas han captado en gran medida los anteriores avances de la última década. Las ciudades han adoptado los sensores IoT para recopilar datos de diferentes fuentes, han llevado su información en tiempo real a los usuarios a través de aplicaciones móviles, han automatizado varios servicios utilizando ML y han favorecido el desarrollo económico a través de las tecnologías de la información y los datos abiertos.

Por todo ello, esta tesis doctoral se centra en el vacío de investigación identificado en el ámbito de las comunicaciones electrónicas. En particular, nuestro objetivo son las aplicaciones basadas en la visión, ya que los avances en el hardware han hecho posible su procesamiento, y muchas nuevas fuentes de datos han estado disponibles recientemente.

Se utilizarán los verticales de las SC para maximizar el impacto de la presente investigación, optimizando los beneficios para los ciudadanos y mejorando su calidad de vida. Se realizará un trabajo de investigación por cada vertical, aunque es importante señalar que algunas oportunidades de investigación se encuentran en la intersección de dos o más mercados verticales.

Capítulo 3

Smart Mobility

Smart Mobility es un vertical que aspira a un futuro definido por la limpieza, la seguridad, la eficiencia y la conectividad. La multimodalidad es uno de sus componentes clave, que desempeña un papel crucial en la remodelación de los actuales patrones de movilidad urbana. La integración y la mejora de la gestión del tráfico, el transporte público, la logística y la infraestructura de las TIC (Tecnologías de la Información y las Comunicaciones) darán lugar a un aire más fresco, más alternativas de movilidad y menos tráfico en las ciudades. Los atascos serán menos habituales en las áreas metropolitanas y perder un tren debido a contratiempos será cosa del pasado. En cuanto a la gestión del tráfico, los territorios inteligentes podrían utilizar las tecnologías modernas para producir cambios en el mundo real.

Este capítulo se centra en uno de los componentes de ML clave para las ayudas a la conducción y los coches autónomos: los transformers. A continuación se presenta una modificación estructurar a su multiattention head capaz de disminuir su coste computacional y aumentar su precisión en aplicaciones visuales.

3.1 Resumen del capítulo

En este capítulo se presenta un una nueva variante de un algoritmo ampliamente utilizado en diversos sub-campos de visión por ordenador. En esencia, el objetivo es recudir la complejidad computacional del estado del arte, manteniendo la precisión y rendimiento. Esto conlleva una serie de desafíos técnicos estamos tratando con unas de las *attention networks* más optimizadas y con mejores resultados de la actualidad.

Elegir la variante adecuada nos permitirá obtener una clasificación más veloz, reducir los recursos computacionales necesarios y mantener un rendimiento equiparable al de la bibliografía actual.

En primer lugar analizamos los fundamentos de la arquitectura transformers y su versión adaptada para el campo de la visión por ordenador. Esto nos permitirá formar unas bases teóricas sólidas desde las que desarrollar nuestra propuesta.

Esto nos permite alterar el todo-conocido *self attention block* y reducir su número de canales de tres (querry, key, value) a solo dos (querry, key). Una capa de tipo MLP es como resultado eliminada, implicando que ya no es necesario tener en cuenta todos sus parámetros asociados para el entrenamiento.

En cuanto a la eliminación del canal *value* que proponemos, teorizamos que su importancia era elevada para tareas de procesamiento de lenguaje natural, y se trata de un componente que fue heredado pero que no parece aportar ningún beneficio a la red. Los resultados obtenidos en unas pruebas de traducción automática corroboran esta teoría.

Los resultados experimentales relacionados con las aplicaciones de visión por ordenador sugieren que el canal *value* no es necesario en la red, pues sin él la precisión aumenta ligeramente, los tiempos de entrenamiento se reducen y el número de parámetros entrenables disminuye considerablemente.

Esto conlleva grandes implicaciones, ya que la clasificación/segmentación de imágenes, detección de objetos, detección de anomalías visuales y el reconocimiento de actividades se convertirían en tareas accesibles para investigadores con menor financiación disponible. Además los recursos que utilizan estos métodos en su día a día reducirían su coste y mejorarían su adopción masiva en la sociedad. Por ejemplo, la seguridad vial aumentaría y se contribuiría al aumento del número de vehículos autónomos en nuestras carreteras.

En definitiva, los resultados confirman el potencial de la propuesta para diversos campos relacionados con la visión por ordenador, y ayudan a comprender mejor el funcionamiento del *self attention block*.

Capítulo 4

Smart People

Smart People es un vertical que entraña diversos beneficios para los residentes de la SC. Trata de hacer evolucionar la forma en que los ciudadanos interactúan con el sector público y privado, tanto como individuos como como empresas. Esto conllevará un aumento de la eficiencia general, a medida que un mayor número de individuos conozca los servicios que tiene disponibles en su territorio inteligente. El despliegue del talento es también un aspecto crucial de este mercado vertical. Hay que impulsar las redes creativas, apoyar a los artistas e individuos creativos y desarrollar asociaciones con organizaciones creativas.

Este capítulo propone una nueva iteración de un método conocido para realizar reconocimiento de actividades humanas en vídeos. Esto conlleva una mejor comprensión por parte de las máquinas del comportamiento humano, lo cual tiene grandes posibilidades de cara a facilitar la interacción persona-tecnología. Bloques modernos basados en el perceptrón multicapa son utilizados, y cambios estructurales en la red PDAN. Los resultados indican que se coste computacional del nuevo modelo es considerablemente inferir que el original encontrado en el estado del arte, aun cuando la accuracy ha sido levemente mejorada.

4.1 Resumen del capítulo

En este capítulo se presenta un nuevo método para el reconocimiento de actividades humanas. En esencia, el objetivo es que el programa sea capaz de etiquetar las acciones que tienen lugar en los vídeos, pudiéndose dar acciones simultáneas o individuales. Esto conlleva una serie de desafíos técnicos debido a la elevada cantidad de datos a considerar (gran cantidad de fotogramas en cada vídeo), incluyendo sus relaciones temporales y causales. La arquitectura desarrollada se basa en el algoritmo PDAN (Dai et al. [2021]) pues es el que mejores resultados ha obtenido para el dataset Charades dentro la bibliografía.

En primer lugar, debido a las limitaciones de hardware con las que contamos, extraemos las *lower level features* de la red I3D para el dataset considerado. Como se trata de una red preentranada, podemos descargar los pesos y preprocesar el dataset, reduciento varias órdenes de magnitud la complejidad del problema.

Esto nos permite diseñar una arquitectura aplicable al mundo real con los recursos disponibles. En nuestra propuesta, modificamos la estructura básica de la red PDAN, dentro del bloque DAL, para reducir el número de parámetros y mejorar su precisión. Además, analizamos el comportamiento de bloques avanzados basados en el perceptrón multicapa y su posible entegración con la red.

Los resultados experimentales sugieren que gMLP, MLP-Mixer y Vision Permutator se integran de forma satisfactoria con la red desarrollada basada en la arquitectura PDAN. En concreto, gMLP y MLP-Mixer permiten reducir aun más el número de parámetros entrenables en la red mientras aumentan la precisión de la misma. Sin embargo, Vision Permutator aumenta el número de parámetros totales aunque consigue la precisión más elevada de todas las alternativas.

En general, los resultados sugieren que es posible reducir el coste computacional de las redes que requieren un procesado intenso de datos, en el caso del reconocimiento de actividades humanas en vídeos.

En definitiva, los resultados confirman que los nuevos bloques basados en el perceptrón multicapa pueden integrarse de manera exitosa con arquitecturas existentes, y que aun hay margen de mejora dentro del campo del reconocimiento de la actividad humana, tanto respecto del coste computacional como de la precisión de los algoritmos.

Capítulo 5

Smart Living

Smart Living es un vertical que busca aumentar la calidad de vida y la seguridad en todos los grupos de edad y demográficos del SC. Optimizar los servicios disponibles y facilitar el acceso de los ciudadanos a ellos son dos de sus objetivos clave. Esto puede dar lugar a una mayor inclusión social y digital, una mayor seguridad, una mejor asistencia sanitaria y mejores edificios inteligentes. La seguridad y la digitalización del turismo también son objetivos clave, y la sanidad se considera uno de los indicadores más importantes de la calidad de vida de los residentes.

Este capítulo presenta los avances obtenidos en el campo de la terapia de protones contra el cáncer de esófago. Una segmentación automática, veloz y fiable de los órganos y el CTV es el resultado del método propuesto, lo cual posibilita un tratamiento temprano de la enfermedad y puede contribuir a aumentar la esperanza de vida de la población. La red nnU-Net es utilizada para procesar imágenes de TAC (Tomografía Axial Computerizada), y una variante de U-Net utilizada para predicción de dosis es mejorada.

5.1 Resumen del capítulo

En este capítulo se presenta una solución para agilizar el tratamiento de pacientes de cáncer de esófago, en el caso de una terapia de protones. En esencia, el objetivo es realizar una segmentación automática tanto de órganos como del CTV sobre imágenes de TAC. Esto conlleva una serie de desafíos técnicos como procesar gran cantidad de información tridimensional, la gran variabilidad entre las distintas bases de datos según el/los médico(s) que la etiquetó/etiquetaron, y todos los desafíos relacionados con el paradigma *Small Data*.

En primer lugar analizamos los modelos que pueden ser utilizados para el caso de uso considerado, destacando la red nnU-Net. Esta red forma parte de un framework de Auto-ML que presentan un gran poder de adaptación de cara a segmentar órganos en distintas configuraciones, ubicaciones y con distintos datos de entrada.

Esto nos permite diseñar una arquitectura con posibilidades de aplicarla al mundo real. En nuestra propuesta, los médicos podrían obtener una segmentación de calidad en cuestión de segundos, lo cual adelantaría el comienzo del tratamiento y incrementaría la tasa de supervivencia de los pacientes. La única intervención humana sería para verificar la segmentación, un proceso que tras varios ciclos de utilización generaría suficientes datos de calidad como para poder re-entrenar la red y prescindir de la supervisión humana.

Aunque la metodología utilizada era robusta, nos encontramos con el problema de tener un número muy reducido de muestras etiquetadas con las que entrenar los modelos. Esto fue solucionado mediante el uso intensivo de *data augmentation*.

En cuanto al desafío generado por el gran tamaño de los datos a procesar, esto fue solucionado mediante la implementación de técnicas como "Mixed precision" para optimizar el funcionamiento de las GPU de última generación.

Los resultados experimentales sugieren que el método es superior a otras alternativas presentes en la bibliografía, la mayoría de ellas basadas en la red U-Net.

En general, el índice de Sørensen-Dice indica que el método es efectivo segmentando los OARs y el CTV, lo cual ha sido validado utilizando cross-validation. La segmentación obtenida en todos los órganos considerados muestra un índice de Sørensen-Dice superior al 90% (excepto para la columna vertebral).

Además, se proponen algunas mejoras para las redes de tipo U-Net utilizadas en "dose prediction" y ciertas lineas de trabajo futuro.

En definitiva, los resultados evidencian el potencial de las redes de tipo nnU-Net para la segmentación automática de órganos y del CTV, con las implicaciones médicas que ello conlleva.

Capítulo 6

Smart Economy

Smart economy es un vertical basado en la innovación y el espíritu empresarial. Ambas acciones estratégicas pueden ayudar a combatir el desempleo, impulsar la productividad y mejorar el mercado laboral. Desde el siglo XVIII, la globalización ha sido un proceso incesante debido a los avances en la tecnología de los transportes y las comunicaciones, lo que ha dado lugar a grandes pero desafiantes oportunidades para los negocios y las empresas locales (Bretos and Marcuello [2017]). La adaptación a los cambios debe ser razonablemente rápida, ya que siempre existe el riesgo de que otro actor del mercado global se aproveche de ello y se haga con el control de la mayor parte del mercado.

Este capítulo presenta un nuevo método para lograr una clasificación tipológica automática de edificios, utilizando fotos de sus fachadas como única información disponible. Se trata de una actuación decidida a adaptar las capacidades de las administraciones regionales a las nuevas posibilidades tecnológicas. Transfer Learning es utilizado para desarrollar el modelo, cuyos resultados muestran que tiene un gran potencial para ser utilizado en casos reales por los distintos catastros regionales.

6.1 Resumen del capítulo

En este capítulo se presenta un nuevo método para la asignación automática de categorías constructivas a edificios. En esencia, el objetivo es asignar una categoría constructiva a distintos edificios basada solamente en una fotografía de su fachada, de acuerdo a las tipología constructivas del catastro local. Esto conlleva una serie de desafíos técnicos como desarrollar un clasificador fiable, procesar imágenes de las fachadas tomadas desde diversos ángulos, y no verse afectado por posibles objetos externos que entorpezcan la visión de la fachada. Elegir la arquitectura correcta nos permitirá obtener una clasificación en la cual las instituciones puedan confiar en el futuro.

En primer lugar analizamos los modelos que pueden contribuir a desarrollar la arquitectura de ML, como por ejemplo redes convolucionales, redes del tipo Inception, Xception, VGG y máquinas de soporte vectorial.

Esto nos permite diseñar una solución aplicable al mundo real. En nuestra propuesta, los operarios del catastro solo tendrían que verificar una categorización ya realizara relativa a las tipologías constructivas del catastro urbano. Esto mejoraría su productividad y la calidad de las asignaciones finales.

En cuanto a las categorías constructivas presentes en el dataset, una de ellas contenía significativamente menos ejemplos que las demás, por lo que tuvo que ser descartada. Sin embargo, el método desarrollado es capaz de integrarla si fuera posible.

Los resultados experimentales sugieren que la red Xception obtiene una precisión satisfactoria para esta primera propuesta, y que las otras redes se comportan significativamente peor.

En general, los resultados experimentales muestran el potential de la inteligencia artificial al ser puesta al servicio de las unidades catastrales locales. Cuanto más datos sean utilizados para entrenar los algoritmos, mejor se comportarán y más fiables serán sus predicciones.

En definitiva, la asignación automática de categorías constructivas es actualmente un proceso manual en que el han de participar dos operarios del catastro anualmente — uno para recolectar los datos y otro para etiquetarlos. Este sistema puede adaptarse fácilmente a los métodos de aprendizaje profundo actuales, lo cual contribuirá a reducir costes y a reducir sesgos en la clasificación.

Capítulo 7

Smart Environment

Smart Environment es un concepto innovador referido a la forma de gestionar los recursos naturales. El objetivo de este vertical es mejorar la habitabilidad del territorio inteligente y el bienestar de sus residentes. Recientemente, el medio ambiente se ha convertido en una de las principales preocupaciones de las nuevas generaciones, con el Acuerdo de París reflejando el compromiso político de todos los países del mundo en este aspecto. La aplicación de cambios normativos y culturales en las grandes metrópolis es clave si se quiere reducir la contaminación atmosférica y acústica. Además, el objetivo a largo plazo de conseguir una ciudad sostenible debe basarse en una gestión eficiente de los residuos, una gestión de alta calidad de los alimentos y el agua, unos patrones de consumo racionales, una planificación urbana exhaustiva y una mayor concienciación.

Este capítulo propone un nuevo método para adaptarse mejor a desafíos generados por la naturaleza, en concreto las enfermedades infecciosas transmitidas por insectos. Se ha desarrollado un clasificador capaz de identificar de manera automática las posibles enfermedades transmitidas por una garrapata en base a una foto del espécimen en cuestión. Transfer Learning es utilizado para desarrollar el modelo, así como técnicas de computer vision para preprocesar las imágenes. Los resultados indican que el método es efectivo y podría ser considerado para su utilización en casos reales.

7.1 Resumen del capítulo

En este capítulo se presenta un novedoso algoritmo para la detección automática de las posibles enfermedades transmitidas por la picadura de garrapatas. En esencia, el objetivo final es que en el futuro un ciudadano cualquiera, sin formación médica ni tecnológica, sea capaz de identificar las enfermedades que le han podido ser transmitidas tras extraerse una garrapata. Esto una serie de desafíos técnicos como identificar el insecto dentro de la imagen tomada, clasificarlo con la suficiente precisión, hallar la correlación especie-posibles enfermedades transmitidas y comunicar esta información al usuario. Elegir la arquitectura de ML adecuada nos permitirá obtener una clasificación efectiva, eficiente y eficaz.

En primer lugar analizamos los modelos de aprendizaje profundo que pueden tener un impacto positivo en la investigación. Transfer learning fue reconocido como el paradigma que mejor se ajusta al caso de uso en consideración, pues se dedica a almacenar conocimientos adquiridos durante la resolución de un problema muy complejo y su aplicación a un problema diferente pero relacionado. Las redes de tipo DenseNet, ResNet, MobileNet y EfficientNet fueron seleccionadas como la estructura principal del modelo, y los pesos de versiones pre-entrenadas de ellas con el dataset ImageNet-1k fueron obtenidas. Tras ello, una máquina de soporte vectorial fue entrenada localmente para obtener un clasificador con las clases deseadas.

Esto nos permite diseñar una arquitectura aplicable en el mundo real. En nuestra propuesta, el usuario ha de tomar una fotografía del insecto sobre un fondo blanco (un folio), con la cámara a 20 centímetros de distancia. Tras ello, la imagen es subida a un servidor que se encarga de: (1) preprocesar la imagen para extraer el área donde se encuentra el insecto y (2) utilizar un clasificador para determinar la especie del espécimen.

Aunque la metodología propuesta es robusta, los datos disponibles para los experimentos presentaban diversos desafíos técnicos tales como: propensión a desarrollar data leakage, un número reducido de ejemplos para determinadas clases, gran cantidad de ejemplos no etiquetados, etc.

En cuanto a la correspondencia enfermedades-especies, se identificó en la bibliografía el relaciones encontradas hasta la fecha y los posibles síntomas a presentar. Por lo tanto, sabida la especie a la que pertenece la garrapata es inmediato identificar las posibles enfermedades transmitidas, así como sus posibles síntomas.

Los resultados experimentales sugieren que el pre-procesado de imágenes desarrollado es exitoso para extraer el insecto de la fotografía. Además, las redes de tipo EfficientNet mostraron rendimientos y precisiones sobresalientes, considerablemente superiores a los de todos los otros tipos de redes. Como consecuencia, el modelo propuesto se basa en la combinación de EfficientNetB3 con una máquina de soporte vectorial.

En general, el error de tipo II es reducido en el caso de las enfermedades más preocupantes. La exhaustividad de la clasificación fue elevada para la clasificación de todas las clases, lo cual indica que el método tiene un gran potencial desde un punto de vista médico.

Además, el modelo presenta la posibilidad de ser re-entrenado cuando más datos sean recolectados, lo cual sin duda incrementará su precisión y demás métricas. Ya que anualmente nos enfrentamos a temporadas de garrapatas, entre marzo y octubre, la obtención de nuevos datos no presenta un desafío.

En definitiva, los resultados verifican el potencial médico del método desarrollado. Médicamente, puede salvar vidas así como aumentar la esperanza de vida de la gente que pasa gran parte de su tiempo en el campo, contribuyendo al desarrollo de una sociedad más inteligente.

Capítulo 8

Conclusiones

Este trabajo de investigación ha presentado varios algoritmos y nuevas aplicaciones para mejorar varios aspectos de las SC, principalmente relacionados con la asistencia sanitaria, la automatización y la conducción autónoma. Se ha identificado una necesidad creciente de soluciones de ML eficientes, capaces de funcionar en entornos de bajos recursos computacionales, así como una demanda de métodos novedosos de diagnóstico temprano y de reducción del coste de las actividades repetitivas.

El mundo experimenta una cantidad cada vez mayor de datos generados por el ser humano, muchos de los cuales se almacenan a la espera de ser procesados. El ML, y en particular el aprendizaje profundo, han demostrado ser métodos integrales para resolver esta situación y mejorar la calidad de vida y el rendimiento económico. El hardware moderno ha permitido un uso generalizado de las aplicaciones de IA, que ya están presentes en los teléfonos inteligentes de la mayor parte de la población mundial. Es necesario desarrollar nuevos y eficientes algoritmos para ampliar aún más su uso. El resultado de este trabajo de investigación son dichos algoritmos.

Las principales aportaciones realizadas en esta tesis doctoral son:

- Se ha realizado un análisis de los problemas concretos de la visión por computador dentro de los SC revisando el estado del arte, abordando los principales problemas que el aprendizaje profundo podría resolver. Se ha aprovechado para establecer los requisitos iniciales para el diseño y desarrollo de la solución.
- Se han estudiado las técnicas y tecnologías en uso en las SCs lo que ha permitido mejorar las soluciones existentes y proponer nuevas aplicaciones. Esto ha

conducido al desarrollo de nuevas estructuras de aprendizaje profundo capaces de un entrenamiento eficiente y eficaz, así como de resultados útiles.

- Hemos utilizado mecanismos formales para resolver el problema de la optimización del comportamiento de las redes en tareas de visión. En particular, las propuestas han aumentado la eficiencia de los métodos utilizados para la toma de decisiones automáticas en SC.
- Hemos demostrado el notable rendimiento de las propuestas en escenarios realistas y desafiantes. Se han evaluado datos reales y se han utilizado conjuntos de datos ampliamente utilizados. Los resultados obtenidos proporcionan una fuerte evidencia de las mejoras en cuanto al rendimiento de las redes y la eficiencia computacional.

La investigación presentada en esta tesis doctoral valida los modelos basados en la visión propuestos para las SC. Tienen un alto potencial, ya que se está produciendo una clara expansión es SCs, y los VE y los smartphones están democratizando el uso de aplicaciones basadas en deep learning. Además, se confirma la hipótesis de partida: se ha comprobado que las técnicas actuales utilizadas para la visión por computador en SCs pueden ser mejoradas (Sec. 4.1), optimizadas (Sec. 7.1) y aceleradas (Sec. 3.1) aplicando técnicas originales de aprendizaje profundo, y se han presentado métodos novedosos en estos aspectos.

Además, se ha dado respuesta a todas las preguntas de investigación que motivaron la presente investigación:

- *¿Qué tareas que consumen tiempo podrían automatizarse?*

  Esta tesis propone un método basado en ML para la identificación de garrapatas y la predicción de sus posibles enfermedades transmitidas en la Sec. 7.1, eliminando la necesidad de que un experto clasifique manualmente el espécimen y proporciona un diagnóstico rápido. Además, la Sec. 6.1 propone un método basado en ML para la categorización automática de casas que podría ser utilizado por los catastros, y la Sec. 5.1 propone un método basado en ML para la segmentación automática de imágenes médicas.

- *¿Se han optimizado los algoritmos de visión modernos para reducir su coste computacional?*

  Los resultados parecen indicar que su precisión ha sido fuertemente optimizada, pero su coste computacional todavía tiene un gran margen de mejora. En particular, la Sec. 3.1 propone modificaciones en la conocida arquitectura de los transformadores que, en promedio, reducen el número de parámetros entrenables en 10,81%, manteniendo o mejorando su precisión. Asimismo, la Sec. 4.1 propone modificaciones a la arquitectura PDAN (relacionada con HAR), consiguiendo una reducción de hasta el $34,87\%$ en el número de parámetros entrenables mientras se mejora ligeramente la precisión del modelo.

- *¿El sector público se centrará más en los datos en el futuro?*

  Las nuevas técnicas centradas en los datos tienen un claro potencial para mejorar el funcionamiento de las administraciones en los TS, y sus gestores parecen haberlo notado en los últimos años (Retuerta et al. [2018]). Las secciones 6.1, 7.1, 5.1 aportan soluciones a problemas reales a los que se enfrenta el sector público en la actualidad, y que podrían resolverse con métodos más centrados en los datos. Por lo tanto, el sector público dispone actualmente de las herramientas y la oportunidad de hacer un amplio uso de los datos en su funcionamiento, ocasión que probablemente aprovechará.

- *¿Qué tareas sensibles al tiempo se ven actualmente sobrecargadas por la ejecución manual?*

  El reconocimiento de las especies de garrapatas y la delimitación de los órganos de los pacientes con cáncer de esófago son los casos más destacados identificados en esta disertación. En ambos casos, se sabe que un diagnóstico precoz reduce la tasa de mortalidad del paciente por la enfermedad sufrida, pero las técnicas manuales que se utilizan actualmente pueden tardar hasta una semana, debido a los periodos de alta demanda y a problemas de organización.

- *¿Podría mejorarse la fiabilidad de la comprensión automática del vídeo?*

  Sí. La sección 4.1 describe cómo mejorar la precisión de una red de última generación para HAR, localizando mejor el componente temporal de las acciones dentro de un vídeo e identificando las actividades correctas.

- *¿Podrían utilizarse ya con éxito los datos de las administraciones en el aprendizaje automático?*

  Sí. Las secciones 7.1 y 6.1 utilizan un conjunto de datos reales de las administraciones locales, logrando resultados exitosos y validando el potencial de aplicar métodos de ML conocidos.

La principal línea de investigación futura respecto al trabajo presentado, será incluir nuevas técnicas matemáticas y estructuras modernas de aprendizaje profundo para optimizar el comportamiento de los algoritmos, así como estudiar el funcionamiento interno de sus diferentes capas para entender mejor la utilidad de cada uno de sus componentes. Hay que mantener los componentes cruciales y alterar o eliminar las estructuras menos importantes, asegurando un rendimiento equilibrado. Finalmente, la última línea de investigación que se ha abierto está relacionada con los transformadores. Las arquitecturas más modernas se modificarán con las alteraciones propuestas, en un esfuerzo por desafiar el estado actual de la técnica.

A partir de los conocimientos y resultados obtenidos, algunas posibles líneas de investigación futuras pueden explorar los siguientes aspectos:

- Pruebas y validación. Se necesitan pruebas exhaustivas para evaluar en detalle los algoritmos propuestos en varias dimensiones, como el rendimiento en la vida real, la validación empírica, la utilidad de las aplicaciones, el funcionamiento equilibrado, la aplicación en tiempo real, la calidad de los resultados, etc. Los resultados obtenidos permitirían seguir optimizando los métodos y conseguir algoritmos y sistemas robustos.

- Resolución de nuevos problemas prácticos. Los métodos propuestos podrían demostrar su utilidad en diversas aplicaciones. El uso de nuevas variables y nuevos casos de uso podría verificar su correcto comportamiento. Un algoritmo que se adapte bien a nuevos problemas es muy necesario, ya que sería un paso desde la *narrow AI* hacia la tan deseada *broad AI*.

- Incorporación de nuevas técnicas. Las nuevas técnicas matemáticas y las novedosas estructuras de aprendizaje profundo tienen el potencial de mejorar y optimizar los actuales algoritmos utilizados en las SC. Deben ser exploradas para aumentar la eficacia y la eficiencia en las SC.

Para terminar, los métodos de aprendizaje profundo pueden mejorar activamente el campo de la visión por ordenador y sus aplicaciones a las SC. Estos tres campos emergentes gozan de gran popularidad en la actualidad, y constantemente aparecen y desaparecen nuevos modelos. Esta tesis estudia sistemáticamente el campo, y propone métodos para mejorar varios puntos clave que, estadísticamente, podrían suponer poco más que una gota en un vasto océano. Sin embargo, ¿qué es un océano, sino una multitud de gotas?

Bibliography

Ajani, J. A., D'Amico, T. A., Bentrem, D. J., Chao, J., Corvera, C., Das, P., Denlinger, C. S., Enzinger, P. C., Fanta, P., Farjah, F., et al. (2019). Esophageal and esophagogastric junction cancers, version 2.2019, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 17(7):855–883.

Akbari, A. and Jafari, R. (2019). Transferring activity recognition models for new wearable sensors with deep generative domain adaptation. *2019 18th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 85–96.

Akbari, A., Wu, J., Grimsley, R., and Jafari, R. (2018). Hierarchical signal segmentation and classification for accurate activity recognition. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers.*

Alamo, T., Reina, D. G., Mammarella, M., and Abella, A. (2020). Open data resources for fighting covid-19. *arXiv preprint arXiv:2004.06111.*

Alani, A. A., Cosma, G., and Taherkhani, A. (2020). Classifying imbalanced multi-modal sensor data for human activity recognition in a smart home using deep learning. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Allam, Z. and Dhunny, Z. A. (2019). On big data, artificial intelligence and smart cities. *Cities*, 89:80–91.

AlMarzouq, M., Zheng, L., Rong, G., and Grover, V. (2005). Open source: Concepts, benefits, and challenges. *Communications of the Association for Information Systems*, 16(1):37.

Alonso, C., Berg, A., Kothari, S., Papageorgiou, C., Rehman, S., AFR, P. N., and RES, C. P. (2020). Will the ai revolution cause a great divergence? *IMF Working Papers*, 2020(184).

Angeleas, A. and Bourbakis, N. G. (2016). A two formal languages based model for representing human activities. *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 779–783.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., van Ginneken, B., et al. (2021). The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*.

Ashour, A., Cameron, P., Bernard, S., Fitzgerald, M., Smith, K., and Walker, T. (2007). Could bystander first-aid prevent trauma deaths at the scene of injury? *Emergency Medicine Australasia*, 19(2):163–168.

Aundhkar, A. and Guja, S. (2021). A review on enterprise data lake solutions.

Avilés-Cruz, C., Rodríguez-Martínez, E., Villegas-Cortéz, J., and Ferreyra, A. (2019). Granger-causality: An efficient single user movement recognition using a smartphone accelerometer sensor. *Pattern Recognit. Lett.*, 125:576–583.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bangare, S. L., Dubal, A., Bangare, P. S., and Patil, S. (2015). Reviewing otsu's method for image thresholding. *International Journal of Applied Engineering Research*, 10(9):21777–21783.

Bank, T. W. (2019). Education statistics - all indicators. [Online] Available: https://databank.worldbank.org/source/education-statistics-%5e-all-indicators and https://www.kaggle.com/datasets/theworldbank/education-statistics. Accessed: 04/04/2022.

Barragán-Montero, A. M., Nguyen, D., Lu, W., Lin, M.-H., Norouzi-Kandalan, R., Geets, X., Sterpin, E., and Jiang, S. (2019). Three-dimensional dose prediction for lung imrt patients with deep neural networks: robust learning from heterogeneous beam configurations. *Medical physics*, 46(8):3679–3691.

Barragán-Montero, A. M., Thomas, M., Defraene, G., Michiels, S., Haustermans, K., Lee, J. A., and Sterpin, E. (2021). Deep learning dose prediction for imrt of esophageal cancer: The effect of data quality and quantity on model performance. *Physica Medica*, 83:52–63.

Batchuluun, G., Kim, J. H., Hong, H. G., Kang, J. K., and Park, K. R. (2017). Fuzzy system based human behavior recognition by combining behavior prediction and recognition. *Expert Syst. Appl.*, 81:108–133.

Batra, N., Parson, O., Berges, M., Singh, A., and Rogers, A. (2014a). Commercial building energy dataset. [Online] Available: https://combed.github.io. Accessed: 04/04/2022.

Batra, N., Parson, O., Berges, M., Singh, A., and Rogers, A. (2014b). A comparison of non-intrusive load monitoring methods for commercial and residential buildings. *arXiv:1408.6595*.

Benešová, A. and Tupa, J. (2017). Requirements for education and qualification of people in industry 4.0. *Procedia Manufacturing*, 11:2195–2202.

Beukema, J. C., Kawaguchi, Y., Sijtsema, N. M., Zhai, T.-T., Langendijk, J. A., van Dijk, L. V., van Luijk, P., Teshima, T., and Muijs, C. T. (2020). Can we safely reduce the radiation dose to the heart while compromising the dose to the lungs in oesophageal cancer patients? *Radiotherapy and Oncology*, 149:222–227.

Boden, T. A., Marland, G., and Andres, R. J. (2013). Co2 emissions from fossil fuels since 1751, by nation. [Online] Available: https://datahub.io/core/co2-fossil-by-nation. Accessed: 04/04/2022.

Bratton, R. L. and Corey, G. R. (2005). Tick-borne disease. *American family physician*, 71(12):2323–2330.

Bretos, I. and Marcuello, C. (2017). Revisiting globalization challenges and opportunities in the development of cooperatives. *Annals of Public and Cooperative Economics*, 88(1):47–73.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Butler-Adam, J. (2018). The fourth industrial revolution and education. *South African Journal of Science*, 114(5-6):1–1.

Campbell, M., Hoane Jr, A. J., and Hsu, F.-h. (2002). Deep blue. *Artificial intelligence*, 134(1-2):57–83.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Casino, F., Patsakis, C., Batista, E., Borràs, F., and Martínez-Ballesté, A. (2017). Healthy routes in the smart city: A context-aware mobile recommender. *IEEE Software*, 34(6):42–47.

Cchangcs (2019). Garbage classification. [Online] Available: https://www.kaggle.com/datasets/asdasdasasdas/garbage-classification. Accessed: 04/04/2022.

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. (2021). Deep learning for sensor-based human activity recognition. *ACM Computing Surveys (CSUR)*, 54:1 – 40.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Cisco (2016). Cisco visual networking index forecast and methodology. [Online] Available: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf. Accessed: 01/04/2022.

Cohen, M. J. and Avidan, S. (2021). Transformaly–two (feature spaces) are better than one. *arXiv preprint arXiv:2112.04185*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Council, B. C. (2017). Planned temporary road occupancies. [Online] Available: https://www.data.brisbane.qld.gov.au/data/dataset/planned-temporary-road-occupancies and https://www.data.brisbane.qld.gov.au/data/dataset/traffic-data-at-intersection-api and https://www.data.brisbane.qld.gov.au/data/dataset/traffic-management-key-corridor-monthly-performance-report. Accessed: 04/04/2022.

Council, L. C. (2020). Road traffic accidents. [Online] Available: https://data.europa.eu/data/datasets/road-traffic-accidents. Accessed: 04/04/2022.

Courmont, A. (2016). *Politiques des données urbaines: ce que l'open data fait au gouvernement urbain*. PhD thesis, Institut d'études politiques de paris-Sciences Po.

Dai, R., Das, S., Minciullo, L., Garattoni, L., Francesca, G., and Bremond, F. (2021). Pdan: Pyramid dilated attention network for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2970–2979.

Dai, R., Das, S., Sharma, S., Minciullo, L., Garattoni, L., Bremond, F., and Francesca, G. (2022). Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dameri, R. P. and Rosenthal-Sabroux, C. (2014). Smart city and value creation. In *Smart city*, pages 1–12. Springer.

Dang, L. M., Hassan, I., Im, S., and Moon, H. (2019). Face image manipulation detection based on a convolutional neural network. *Expert Syst. Appl.*, 129:156–168.

Dedabrishvili, M., Dundua, B., and Mamaiashvili, N. (2021). Smartphone sensor-based fall detection using machine learning algorithms. In *IEA/AIE*.

Denmark, S. (2021). Straf11: Reported criminal offences by region and type of offence. [Online] Available: https://www.statbank.dk/STRAF11. Accessed: 04/04/2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doersch, C., Gupta, A., and Zisserman, A. (2020). Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33:21981–21993.

Dong, X., Lei, Y., Wang, T., Thomas, M., Tang, L., Curran, W. J., Liu, T., and Yang, X. (2019). Automatic multiorgan segmentation in thorax ct images using u-net-gan. *Medical physics*, 46(5):2157–2168.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eldan, R. and Shamir, O. (2016). The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR.

Elhammali, A., Blanchard, P., Yoder, A., Liao, Z., Zhang, X., Zhu, X. R., Allen, P. K., Jeter, M., Welsh, J., and Nguyen, Q.-N. (2019). Clinical outcomes after intensity-modulated proton therapy with concurrent chemotherapy for inoperable non-small cell lung cancer. *Radiotherapy and Oncology*, 136:136–142.

Elleuch, M., Maalej, R., and Kherallah, M. (2016). A new design based-svm of the cnn classifier architecture with dropout for offline arabic handwritten recognition. *Procedia Computer Science*, 80:1712–1723.

Française, R. (2022). Le site national des adresses. [Online] Available: https://adresse.data.gouv.fr/. Accessed: 04/04/2022.

Garcia-Retuerta, D., Chamoso, P., Hernández, G., Guzmán, A. S. R., Yigitcanlar, T., and Corchado, J. M. (2021). An efficient management platform for developing smart cities: Solution for real-time and future crowd detection. *Electronics*, 10(7):765.

Gelenbe, E., Mao, Z.-H., and Li, Y.-D. (1999). Function approximation with spiked random networks. *IEEE Transactions on Neural Networks*, 10(1):3–9.

Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A. (2019). Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253.

Goel, S. S., Goel, A., Kumar, M., and Moltó, G. (2021). A review of internet of things: qualifying technologies and boundless horizon. *Journal of Reliable Intelligent Environments*, pages 1–11.

Goodwin, P. (2004). The economic costs of road traffic congestion. *UCL (University College London) eprints*.

government, I. (2019). Indian water quality data. [Online] Available: https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data. Accessed: 04/04/2022.

Government, S. M. (2019). Air pollution in seoul. [Online] Available: https://www.kaggle.com/datasets/bappekim/air-pollution-in-seoul. Accessed: 04/04/2022.

GraphHopper (2022). Graphhopper open traffic collection. [Online] Available: https://github.com/graphhopper/open-traffic-collection. Accessed: 04/04/2022.

Gudur, G. K., Sundaramoorthy, P., and Umaashankar, V. (2019). Activeharnet: Towards on-device deep bayesian active learning for human activity recognition. *ArXiv*, abs/1906.00108.

Gumaei, A. H., Hassan, M. M., Alelaiwi, A., and Alsalman, H. (2019). A hybrid deep learning model for human activity recognition using multimodal body sensing data. *IEEE Access*, 7:99152–99160.

Ha, S., Yun, J.-M., and Choi, S. (2015). Multi-modal convolutional neural networks for activity recognition. *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3017–3022.

Hall, B. C. (2019). Economic activities census on the ground floor of the city of barcelona. [Online] Available: https://opendata-ajuntament.barcelona.cat/data/en/dataset/cens-activitats-comercials. Accessed: 04/04/2022.

Hallak, J. A. and Azar, D. T. (2020). The ai revolution and how to prepare for it. *Translational Vision Science & Technology*, 9(2):16–16.

Hammerla, N. Y., Halloran, S., and Plötz, T. (2016). Deep, convolutional, and recurrent models for human activity recognition using wearables. *ArXiv*, abs/1604.08880.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al. (2020). A survey on vision transformer. *arXiv preprint arXiv:2012.12556*.

Hanna, T. P., King, W. D., Thibodeau, S., Jalink, M., Paulin, G. A., Harvey-Jones, E., O'Sullivan, D. E., Booth, C. M., Sullivan, R., and Aggarwal, A. (2020). Mortality due to cancer treatment delay: systematic review and meta-analysis. *bmj*, 371.

Harari, Y. N. (2017). Reboot for the ai revolution. *Nature*, 550(7676):324–327.

Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E., and Chiroma, H. (2016). The role of big data in smart city. *International Journal of information management*, 36(5):748–758.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, X., Zhao, K., and Chu, X. (2021). Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622.

Holcomb, S. D., Porter, W. K., Ault, S. V., Mao, G., and Wang, J. (2018). Overview on deepmind and its alphago zero ai. In *Proceedings of the 2018 international conference on big data and education*, pages 67–71.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Hou, Q., Jiang, Z., Yuan, L., Cheng, M.-M., Yan, S., and Feng, J. (2021). Vision permutator: A permutable mlp-like architecture for visual recognition. *arXiv preprint arXiv:2106.12368*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J. (2020). Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE.

Ijjina, E. P. and Mohan, C. K. (2017). Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recognit.*, 72:504–516.

Institute, I. M. (2007). Weather buoy network. [Online] Available: https://data.gov.ie/dataset/weather-buoy-network. Accessed: 04/04/2022.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.

Ismagilova, E., Hughes, L., Dwivedi, Y. K., and Raman, K. R. (2019). Smart cities: Advances in research—an information systems perspective. *International Journal of Information Management*, 47:88–100.

Jabbour, S. K., Hashem, S. A., Bosch, W., Kim, T. K., Finkelstein, S. E., Anderson, B. M., Ben-Josef, E., Crane, C. H., Goodman, K. A., Haddock, M. G., et al. (2014). Upper abdominal normal organ contouring guidelines and atlas: a radiation therapy oncology group consensus. *Practical radiation oncology*, 4(2):82–89.

Jalal, A., Kim, Y., Kim, Y.-J., Kamal, S., and Kim, D. (2017). Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.*, 61:295–308.

Ji, X., Cheng, J., Feng, W., and Tao, D. (2018). Skeleton embedded motion body partition for human action recognition using depth sequences. *Signal Process.*, 143:56–68.

Jiang, W. and Yin, Z. (2015). Human activity recognition using wearable sensors by deep convolutional neural networks. *Proceedings of the 23rd ACM international conference on Multimedia.*

Jin, D., Guo, D., Ho, T.-Y., Harrison, A. P., Xiao, J., Tseng, C.-K., and Lu, L. (2021). Deeptarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. *Medical Image Analysis*, 68:101909.

KC, M. (2020). Startup success prediction. [Online] Available: https://www.kaggle.com/datasets/manishkc06/startup-success-prediction. Accessed: 04/04/2022.

Kim, Y., Jeong, S. R., and Ghani, I. (2014). Text opinion mining to analyze news for stock market prediction. *Int. J. Advance. Soft Comput. Appl*, 6(1):2074–8523.

Kucera, J. and Chlapek, D. (2014). Benefits and risks of open government data. *Journal of Systems Integration*, 5(1):30–41.

Kumar, V. (2020). Smart environment for smart cities. In *Smart Environment for Smart Cities*, pages 1–53. Springer.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lane, N. D. and Georgiev, P. (2015). Can deep learning revolutionize mobile sensing? *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lesueur, P., Calugaru, V., Nauraye, C., Stefan, D., Cao, K., Emery, E., Reznik, Y., Habrand, J. L., Tessonnier, T., Chaikh, A., et al. (2019). Proton therapy for treatment of intracranial benign tumors in adults: a systematic review. *Cancer treatment reviews*, 72:56–64.

Lilien, R., Housman, J., Mettu, R., Weller, T., Haag, L., and Haag, M. (2017). Gunshot audio forensics dataset. [Online] Available: http://cadreforensics.com/audio. Accessed: 04/04/2022.

Liu, H., Dai, Z., So, D. R., and Le, Q. V. (2021a). Pay attention to mlps. *arXiv preprint arXiv:2105.08050*.

Liu, R., Li, Y., Liang, D., Tao, L., Hu, S., and Zheng, H.-T. (2021b). Are we ready for a new paradigm shift? a survey on visual deep mlp. *arXiv preprint arXiv:2111.04060*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021c). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., and Zhang, G. (2015). Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14–23.

Lyu, L., He, X., Law, Y. W., and Palaniswami, M. S. (2017). Privacy-preserving collaborative deep learning with application to human activity recognition. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., and Yaqoob, I. (2017). Big iot data analytics: architecture, opportunities, and open research challenges. *ieee access*, 5:5247–5261.

Matikainen, L., Hyyppä, J., and Kaartinen, H. (2004). Automatic detection of changes from laser scanner and aerial image data for updating building maps. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, 35:434–439.

Mavroudi, E., Haro, B. B., and Vidal, R. (2020). Representation learning on visual-symbolic graphs for video understanding. In *European Conference on Computer Vision*, pages 71–90. Springer.

Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., and Chiaberge, M. (2021). Action transformer: A self-attention model for short-time human action recognition. *arXiv preprint arXiv:2107.00606*.

Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Shabtai, A., Breitenbacher, D., and Elovici, Y. (2018). N-baiot: Network-based detection of iot botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, 17(3):12–22.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. (2017). Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Moe, T. L. and Pathranarakul, P. (2006). An integrated approach to natural disaster management. *Disaster Prevention and Management: An International Journal*.

Moosavi, S. (2021). Us-accidents: A countrywide traffic accident dataset. [Online] Available: https://smoosavi.org/datasets/us_accidents. Accessed: 04/04/2022.

Moscholidou, I. and Pangbourne, K. (2020). A preliminary assessment of regulatory efforts to steer smart mobility in london and seattle. *Transport Policy*, 98:170–177.

Müller, S. and Zaum, D. W. (2005). Robust building detection in aerial images. *International Archives of Photogrammetry and Remote Sensing*, 36(B2/W24):143–148.

Murahari, V. S. and Plötz, T. (2018). On attention models for human activity recognition. *Proceedings of the 2018 ACM International Symposium on Wearable Computers*.

Narang, S., Chung, H. W., Tay, Y., Fedus, W., Fevry, T., Matena, M., Malkan, K., Fiedel, N., Shazeer, N., Lan, Z., et al. (2021). Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2102.11972*.

Naveed, K. (2020). N-baiot dataset to detect iot botnet attacks. [Online] Available: https://www.kaggle.com/datasets/mkashifn/nbaiot-dataset. Accessed: 04/04/2022.

Neves, F. T., de Castro Neto, M., and Aparicio, M. (2020). The impacts of open data initiatives on smart cities: A framework for evaluation and monitoring. *Cities*, 106:102860.

Niu, X.-X. and Suen, C. Y. (2012). A novel hybrid cnn–svm classifier for recognizing handwritten digits. *Pattern Recognition*, 45(4):1318–1325.

OECD (2021). Oecd statistics. [Online] Available: https://stats.oecd.org. Accessed: 04/04/2022.

of Melbourne, C. (2022). Pedestrian counting system. [Online] Available: http://www.pedestrian.melbourne.vic.gov.au/#date=05-04-2022&time=1. Accessed: 04/04/2022.

of Seattle, C. (2020). 2020 paid parking occupancy (year-to-date). [Online] Available: https://data.seattle.gov/Transportation/2020-Paid-Parking-Occupancy-Year-to-date-/wtpb-jp8d. Accessed: 04/04/2022.

Organization, W. H. (2013). Crimean-congo haemorrhagic fever. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/crimean-congo-haemorrhagic-fever. Accessed: 01/04/2022.

Organization, W. H. et al. (2005). Effects of air pollution on children's health and development: a review of the evidence.

Ouerhani, N., Pazos, N., Aeberli, M., and Muller, M. (2016). Iot-based dynamic street light control for smart cities use cases. In *2016 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–5. IEEE.

Oyedotun, O. K. and Khashman, A. (2016). Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications*, 28:3941–3951.

Park, E., Del Pobil, A. P., and Kwon, S. J. (2018). The role of internet of things (iot) in smart cities: Technology roadmap-oriented approaches. *Sustainability*, 10(5):1388.

Pérez-Lombard, L., Ortiz, J., and Pout, C. (2008). A review on buildings energy consumption information. *Energy and buildings*, 40(3):394–398.

Piergiovanni, A. and Ryoo, M. (2019). Temporal gaussian mixture layer for videos. In *International Conference on Machine learning*, pages 5152–5161. PMLR.

Piergiovanni, A. and Ryoo, M. S. (2018). Learning latent super-events to detect multiple activities in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5304–5313.

Pigou, L., van den Oord, A., Dieleman, S., Herreweghe, M. V., and Dambre, J. (2016). Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, 126:430–439.

Pop, M.-D. and Proștean, O. (2018). A comparison between smart city approaches in road traffic management. *Procedia-social and behavioral sciences*, 238:29–36.

Prati, A., Shan, C., and Wang, K. I.-K. (2019). Sensors, vision and networks: From video surveillance to activity recognition and health monitoring. *J. Ambient Intell. Smart Environ.*, 11:5–22.

Qi, J., Yang, P., Hanneghan, M., Tang, S., and Zhou, B. (2019). A hybrid hierarchical framework for gym physical activity recognition and measurement using wearable sensors. *IEEE Internet of Things Journal*, 6:1384–1393.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Reason, P. and Bradbury, H. (2001). *Handbook of action research: Participative inquiry and practice*. sage.

Retuerta, D. G., Bondía, R. A., Tejedor, J. P., and Rodríguez, J. M. C. (2018). Inteligencia artificial para la asignación automática de categorías constructivas. *CT: Catastro*, 94:111–122.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Russell Stuart, J. (2010). Artificial intelligence/stuart j. russell, peter norvig. *A Modern Approach (Third ed.). Prentice Hall*, page 649.

Sampaio, W. B., Diniz, E. M., Silva, A. C., De Paiva, A. C., and Gattass, M. (2011). Detection of masses in mammogram images using cnn, geostatistic functions and svm. *Computers in Biology and Medicine*, 41(8):653–664.

Sánchez, J. and Perronnin, F. (2011). High-dimensional signature compression for large-scale image classification. In *CVPR 2011*, pages 1665–1672. IEEE.

Schleussner, C.-F., Rogelj, J., Schaeffer, M., Lissner, T., Licker, R., Fischer, E. M., Knutti, R., Levermann, A., Frieler, K., and Hare, W. (2016). Science and policy characteristics of the paris agreement temperature goal. *Nature Climate Change*, 6(9):827–835.

Schrage, M. and Kiron, D. (2018). Leading with next-generation key performance indicators. *MIT Sloan Management Review*, 16(June):1–2.

Shoup, D. (2018). Cruising for parking. In *Parking and the City*, pages 261–269. Routledge.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Simiscuka, A. A. and Muntean, G.-M. (2021). Remos-iot-a relay and mobility scheme for improved iot communication performance. *IEEE Access*, 9:73000–73011.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Solanas, A., Patsakis, C., Conti, M., Vlachos, I. S., Ramos, V., Falcone, F., Postolache, O., Pérez-Martínez, P. A., Di Pietro, R., Perrea, D. N., et al. (2014). Smart health: A context-aware health paradigm within smart cities. *IEEE Communications Magazine*, 52(8):74–81.

Spichtinger, D. and Blumesberger, S. (2020). Fair data and data management requirements in a comparative perspective: Horizon 2020 and fwf policies. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 73(2):207–216.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., and Hopman, R. J. (2017). The smartphone and the driver's cognitive workload: A comparison of apple, google, and microsoft's intelligent personal assistants. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 71(2):93.

Su, K., Li, J., and Fu, H. (2011). Smart city and the applications. In *2011 international conference on electronics, communications and control (ICECC)*, pages 1028–1031. IEEE.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z., and Ding, X. (2020). Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Tan, T.-H., Gochoo, M., Huang, S.-C., Liu, Y.-H., Liu, S.-H., and Huang, Y.-F. (2018). Multi-resident activity recognition in a smart home using rgb activity image and dcnn. *IEEE Sensors Journal*, 18:9718–9727.

Tang, Y., Yang, D., Li, W., Roth, H., Landman, B., Xu, D., Nath, V., and Hatamizadeh, A. (2021). Self-supervised pre-training of swin transformers for 3d medical image analysis. *arXiv preprint arXiv:2111.14791*.

Tay, K.-C., Supangkat, S. H., Cornelius, G., and Arman, A. A. (2018). The smart initiative and the garuda smart city framework for the development of smart cities. In *2018 International Conference on ICT for Smart Society (ICISS)*, pages 1–10. IEEE.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., et al. (2021). Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*.

Turi, A. N. and Li, X. S. (2021). Insight into unlocking entrepreneurial business potentials through data-driven decision making.

van der Aalst, W. (2022). Six levels of autonomous process execution management (apem). *arXiv preprint arXiv:2204.11328*.

Varatharajan, R., Manogaran, G., Kumar, P. M., and Sundarasekar, R. (2017). Wearable sensor devices for early detection of alzheimer disease using dynamic time warping algorithm. *Cluster Computing*, pages 1–10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vidiasova, L., Kachurina, P., and Cronemberger, F. (2017). Smart cities prospects from the results of the world practice expert benchmarking. *Procedia computer science*, 119:269–277.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018.

Walsh, T. (2017). The ai revolution. *NSW Department of Education Education: Future Frontiers*.

Wang, S.-l., Liao, Z., Vaporciyan, A. A., Tucker, S. L., Liu, H., Wei, X., Swisher, S., Ajani, J. A., Cox, J. D., and Komaki, R. (2006). Investigation of clinical and dosimetric factors associated with postoperative pulmonary complications in esophageal cancer patients treated with concurrent chemoradiotherapy followed by surgery. *International Journal of Radiation Oncology* Biology* Physics*, 64(3):692–699.

Wang, X., Hobbs, B., Gandhi, S. J., Muijs, C. T., Langendijk, J. A., and Lin, S. H. (2021). Current status and application of proton therapy for esophageal cancer. *Radiotherapy and Oncology*, 164:27–36.

Weiss, M. and Bailetti, T. (2015). Value of open source projects: A case for open source cybersecurity. In *2015 IEEE International Conference on Engineering, Technology and Innovation/International Technology Management Conference (ICE/ITMC)*, pages 1–8. IEEE.

Wu, C., Nguyen, D., Xing, Y., Montero, A. B., Schuemann, J., Shang, H., Pu, Y., and Jiang, S. (2021). Improving proton dose calculation accuracy by using deep learning. *Machine Learning: Science and Technology*, 2(1):015017.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. (2020). On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR.

Xu, C., Govindarajan, L. N., and Cheng, L. (2017). Hand action detection from ego-centric depth sequences with error-correcting hough transform. *Pattern Recognit.*, 72:494–503.

Xue, D.-X., Zhang, R., Feng, H., and Wang, Y.-L. (2016). Cnn-svm for microvascular morphological type recognition with data augmentation. *Journal of medical and biological engineering*, 36(6):755–764.

Yang, J., Nguyen, M. N., San, P. P., Li, X., and Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *IJCAI*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Ye, L., Rochan, M., Liu, Z., and Wang, Y. (2019). Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10502–10511.

Yeh, H. (2017). The effects of successful ict-based smart city services: From citizens' perspectives. *Government Information Quarterly*, 34(3):556–565.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13:55–75.

Zhu, J., Chen, X., Yang, B., Bi, N., Zhang, T., Men, K., and Dai, J. (2020). Evaluation of automatic segmentation model with dosimetric metrics for radiotherapy of esophageal cancer. *Frontiers in Oncology*, 10:1843.

Zillow (2022). Zillow housing data. [Online] Available: https://www.zillow.com/research/data. Accessed: 04/04/2022.