Application of Deep Symbolic Learning in NGS

Ángel Canal-Alonso1, Pedro Jiménez1 and Noelia Egido1, Javier Prieto[1], Juan Manuel Corchado1

[1]Department of Bioinformatics and Computational Biology, AIR Institute, Carbajosa de la Sagrada, Spain

Email: acanal@air-institute.com

Summary

The application of Deep Symbolic Learning in genomic analysis has begun to gain traction as a promising approach to interpret and understand vast data sets derived from DNA sequencing. Next-generation sequencing (NGS) techniques have revolutionized the field of clinical genetics and human biology, generating massive volumes of data that require advanced tools for analysis. However, traditional methods are often too abstract or complicated for clinical staff. This work focuses on exploring how Deep Symbolic Learning, a subfield of explainable artificial intelligence (XAI), can be effectively applied to NGS data. A detailed evaluation of the suitability of different architectures will be carried out,

Keywords: Next-Generation sequencing, Explainable Artificial Intelligence, Deep Symbolic Learning

Introduction

Deep Learning

Deep Learning, also known as deep learning, represents a subcategory of techniques within machine learning that has revolutionized multiple fields of study and application, from computer vision to natural language processing, including bioinformatics and computational genomics. . Specifically, this branch of machine learning is based on artificial neural network architectures with numerous hidden layers, which are called "deep." These layers allow the model to learn hierarchies of features from the data, from the most basic to the most complex.

The foundation of Deep Learning is the simulation of neural structures analogous to those present in the human brain, although in a considerably more simplified manner. Each neuron in these networks is connected to others and can transmit information between them. As the data passes through each layer, a nonlinear transformation is performed that allows for gradual feature detection and construction. For example, in the context of computer vision, early layers might identify edges, while deeper layers might identify complex structures such as faces or specific patterns.

A key aspect that has driven the success of Deep Learning is the ability of these deep neural networks to perform machine learning of features. Unlike traditional approaches, where features are extracted manually, in deep learning the network is capable of autonomously learning the most relevant features directly from the raw data.

The efficiency and accuracy of these techniques, however, come at the cost of the need for large volumes of data and significant computational power. The constant feedback and adjustment of the weights of the connections in these networks during training requires powerful processing units, with Graphics Processing Units (GPU) being the most used due to their ability to handle matrix operations in parallel, essential in the neural network training.

In the context of genomics and next generation sequencing (NGS), Deep Learning has shown potential to improve accuracy and speed in tasks such as genomic annotation, protein structure prediction and, of relevance to our study, in the Variant Calling phase, where the aim is to identify genetic variants from sequencing data. Deep Learning's ability to handle large data sets and learn complex features makes it a valuable tool for addressing the challenges inherent in high-resolution genomics.

Symbolic Learning

Symbolic learning, or Symbolic Learning, is a classic approach in the field of artificial intelligence (AI) that focuses on the representation and manipulation of knowledge in the form of symbols and rules. Instead of relying exclusively on numerical or statistical calculations, as in other machine learning approaches, symbolic learning is based on the construction of symbolic representations of information, allowing logical reasoning and deductions based on these representations.

The essence of symbolic learning lies in its ability to model complex and structured relationships in data. These models are generally interpretable, as they are made up of sets of rules, facts or logical structures, which can be easily understood and examined by humans. For example, a system based on symbolic learning could express knowledge in the form of rules of the type "If A, then B", allowing us to reason about these rules and reach specific conclusions.

Symbolic learning has been fundamental in the evolution of AI, especially in the early years of the discipline, and has given rise to expert systems, inference engines, and knowledge bases. These systems are especially effective in domains where prior knowledge is essential and can be clearly defined and structured, such as in medicine, law or engineering.

In the field of genomics and bioinformatics, symbolic learning offers a unique perspective by providing tools to represent and manipulate biological knowledge in a structured way. Genetic relationships, metabolic pathways or protein-protein interactions, for example, can be encoded in symbolic structures that facilitate their analysis and understanding.

When combined with approaches such as Deep Learning, symbolic learning allows systems to benefit from both the feature generalization and machine learning capabilities of neural networks, as well as the precision, transparency, and interpretability of symbolic reasoning. This combination, called "Deep Symbolic Learning", integrates the best of both worlds, and in the context of Variant Calling in NGS pipelines, can offer robust and highly interpretable solutions for the identification and classification of genetic variants.

Deep Symbolic Learning

Deep Symbolic Learning (DSL) is an emerging approach in the field of artificial intelligence that seeks to combine the strengths of Deep Learning and symbolic learning. This integration aims to resolve one of the main criticisms of Deep Learning, which is the lack of interpretability and transparency of its models, offering solutions that are not only powerful in terms of performance, but also understandable and justifiable.

Deep Symbolic Learning operates through the conjunction of neural networks, capable of learning rich and hierarchical representations of data, with symbolic structures that allow the construction of logical and semantically coherent models. Instead of considering only numerical or statistical patterns, as pure Deep Learning does, DSL incorporates symbols, rules, and logical relationships into its learning process, providing greater context and structure to the knowledge acquired.

A notable aspect of DSL is its ability to leverage previously established knowledge, encoded in symbolic representations, to inform and guide the learning of neural networks. This is especially useful in domains where there is a rich knowledge base, such as biology and genomics.

In the context of Variant Calling in NGS pipelines, DSL offers a promising approach. Neural networks can learn complex patterns and subtleties in sequencing data, while the symbolic component can incorporate rules and known facts about genetic variants, mutations, and their biological relevance. This could not only improve the accuracy of variant identification, but also provide logical and knowledge-based explanations for why a certain sequence is considered a variant.

The hybrid nature of Deep Symbolic Learning also allows for greater flexibility in modeling. While neural networks can adjust to quirks and noise in the data, the symbolic component can act as a regulator, ensuring that predictions and conclusions are consistent with established biological knowledge. In this way, the DSL is positioned as a robust and cutting-edge tool to face the intrinsic challenges of the Variant Calling phase and of genomics in general.

Application of DSL in NGS

Next-generation sequencing (NGS) has revolutionized the field of genomics, allowing large volumes of genomic data to be obtained in significantly reduced times and costs compared to traditional techniques. These advances, while providing an unprecedented wealth of information, also present significant challenges in terms of data processing, analysis and interpretation. It is in this context where Deep Symbolic Learning (DSL) emerges as a potential solution to address and overcome such challenges.

Advantages of your application

The union of symbolic learning and deep learning in the context of next generation sequencing (NGS) offers a range of advantages that capitalize on the strengths of both approaches:

- Interpretability and Transparency: One of the main challenges of Deep Learning is its “black box” nature, which means that although the model may have high performance, it can be difficult to understand how it arrives at a particular decision. By integrating symbolic learning, a layer of transparency and explainability is introduced to the model. Decisions based on symbolic rules can be inspected, tracked and justified, facilitating the understanding and validation of results in the context of NGS.

- Incorporation of Prior Knowledge: In genomics, there is a vast body of accumulated knowledge about genetics, mutations, and genomic relationships. Symbolic learning allows the explicit incorporation of this knowledge in the form of rules and relationships. This not only informs and guides the model, but can also increase accuracy and robustness by ensuring that the system does not contradict well-established genomic principles.
- Generalization and Adaptability: While Deep Learning is excellent for detecting and learning patterns in large data sets, symbolic learning gives the system the ability to generalize from specific examples and adapt to new data or contexts. This is essential in NGS, where data can vary depending on the sequencing technique, the organism studied, or the experimental conditions.
- Robustness to Noise: NGS data can be noisy due to sequencing errors or biological variations. While deep neural networks may be susceptible to overfitting to this noise, the structured and logical nature of symbolic learning can act as a moderator, preventing hasty conclusions based on noisy or atypical information.
- Computational Optimization: The integration of symbolic knowledge can direct and focus the learning process, potentially reducing the need for computationally expensive iterations. By having a guiding structure based on known rules and relationships, the system can converge more quickly to optimal solutions, saving time and computational resources.

Multi-Modality Integration: In genomics, different types of data are often combined, such as genomic sequences, gene expression, and proteomics data. While deep learning can efficiently handle the integration of multiple data modalities, symbolic learning can provide a coherent and structured framework for understanding and reasoning about how these different types of data relate to each other.

Applicability in the phases of a pipeline

The interpretation of genetic sequences is a critical task in genomics and bioinformatics, as it involves the identification and understanding of variants and mutations that may have clinical, evolutionary or functional implications. Deep Symbolic Learning (DSL), by combining deep learning and symbolic learning, has significant potential to improve and enrich this interpretation. Some specific applications of DSL in the area of genomic interpretation are explored here:

- Variant Identification: One of the main tasks in genomic interpretation is to identify variants, such as SNPs and structural mutations, from NGS sequences. The DSL can be particularly useful here, as neural networks can identify complex patterns in the data, while the symbolic component can validate these identifications against previously established rules and knowledge. This combination can significantly reduce false positives and negatives.
- Functional Analysis: Not all identified variants have a functional impact. DSL can help predict the impact of a variant, combining machine learning based on data from gene expression, protein structure, and other modalities, with symbolic rules that encode prior knowledge about functional sites, protein domains, and biological pathways.
- Clinical Interpretation: For variants with potential clinical importance, it is essential to interpret their meaning in terms of diseases, phenotypes or response to treatments. Here, the symbolic component of DSL can leverage databases of clinical variants and scientific literature, while deep learning can identify subtle patterns in the data that correlate specific variants with clinical outcomes.
- Evolutionary Understanding: DSL can also be applied to understand the evolutionary implications of variants, combining the ability of deep learning to analyze large genomic data sets from different species, with symbolically encoded evolutionary rules and theories.
- Multi-Omics Data Integration: Modern genomics goes beyond just DNA sequences, also incorporating transcriptomic, proteomic and metabolomic data. The DSL is especially well suited for this integrative task, as it can learn unified representations of different types of data while reasoning about them in a coherent symbolic framework.
- Automation and Scalability: As the amount of genomic data grows exponentially, it is essential that interpretation systems are automatic and scalable. DSL, by combining the computational efficiency of deep learning with the structure and coherence of symbolic learning, offers a solution that can process large volumes of data efficiently and accurately.

Current proposals for DSL architectures

Recently, researchers from IBM Research Zürich and ETH Zürich designed an architecture that combines deep neural networks and vector-symbolic models, known as neuro-vector-symbolic architecture (NVSA). This architecture overcomes previous limitations, providing a unified framework for solving tasks involving high-level perception and reasoning. NVSA has proven effective in solving Raven's progressive matrices, an abstract reasoning task, with remarkable efficiency and accuracy compared to other architectures.

The neuro-vector-symbolic architecture (NVSA) proposed by researchers at IBM Research Zürich and ETH Zürich represents an innovative step in the evolution of artificial intelligence systems. Below is a more detailed description of how it works and what makes it special.

Deep neural networks (DNNs) are a subclass of neural networks that have multiple hidden layers between the input and output. These layers allow DNNs to model and learn complex, non-linear patterns. They have been used successfully in a wide variety of tasks, especially those related to perception, such as image recognition and speech processing.

On the other hand, vector-symbolic models are based on symbolic representations, which means they work with abstract concepts and relationships between them rather than direct patterns of data. These models are especially useful for tasks that require reasoning and manipulation of symbols, since they can represent and work with logical and semantic structures.

The NVSA architecture combines the power of DNNs and vector-symbolic models. While DNNs deal with perception and feature extraction from input data, vector-symbolic models deal with high-level reasoning and symbolic manipulation.

This hybrid design allows NVSA to overcome previous limitations by providing a unified framework. Instead of relying solely on DNNs for all tasks or relying only on symbolic systems, this architecture uses the strengths of both approaches where they are most relevant.

A good example of its effectiveness is Raven's progressive matrix solver. These matrices are psychometric tests designed to evaluate an individual's abstract reasoning. They require both perception (identifying visual patterns) and logical reasoning (deducing the relationship between different elements and predicting the next in the sequence). The NVSA has proven to be remarkably efficient and accurate in this task, outperforming other architectures that only use one of the two approaches.

The neuro-vector-symbolic architecture represents a promising integration of DNN-based perception and reasoning based on vector-symbolic models. Its ability to address tasks that combine both needs shows its potential to take artificial intelligence to new horizons in terms of versatility and efficiency.

References

Garcia-Retuerta D, Canal-Alonso A, Casado-Vara R, Rey AM, Panuccio G, Corchado JM. Bidirectional-Pass Algorithm for Interictal Event Detection. In Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020). PACBB 2020. Advances in Intelligent Systems and Computing, vol 1240. Springer, Cham. https://doi.org/10.1007/978-3-030-54568-0_20

Castillo Ossa LF, Chamoso P, Arango-López J, Pinto-Santos F, Isaza GA, Santa-Cruz-González C, Ceballos-Marquez A, Hernández G, Corchado JM. A Hybrid Model for COVID-19 Monitoring and Prediction. Electronics. 2021; 10(7):799.
https://doi.org/10.3390/electronics10070799

Intelligent Platform Based on Smart PPE for Safety in Workplaces. Márquez-Sánchez S, Campero-Jurado I, Herrera-Santos J, Rodríguez S, Corchado JM. Sensors (Basel). 2021 Jul 7;21(14):4652
https://doi.org/10.3390/s21144652

A. Canal-Alonso, R. Casado-Vara and J. Manuel Corchado, "An affordable implantable VNS for use in animal research," 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2020, pp. 1-4,
doi: 10.1109/ICECS49266.2020.9294958

An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray. Ramos J, Castellanos-Garzón JA, González-Briones A, de Paz JF, Corchado JM. Interdiscip Sci. 2017 Mar;9(1):1-13
DOI 10.1007/s12539-017-0219-6

Acknowledgments

This study has been funded by the AIR Genomics project (with file number CCTT3/20/SA/0003), through the call 2020 R&D PROJECTS ORIENTED TO THE EXCELLENCE AND COMPETITIVE IMPROVEMENT OF THE CCTT by the Institute of Business Competitiveness of Castilla y León and FEDER funds