IL NUOVO CIMENTO **48 C** (2025) 116
DOI 10.1393/ncc/i2025-25116-0

Colloquia: IFAE 2024

Advancing High-Energy Physics through Machine Learning: Current state and future prospects(*)

F. Vaselli[1][2]

[1] *INFN, Sezione di Pisa - Pisa, Italy*
[2] *Scuola Normale Superiore - Pisa, Italy*

received 2 December 2024

Summary. — Recent advancements in AI and Machine Learning are having remarkable results in the field of High Energy Physics. The growing adoption of *Transformer* architectures make it easy to build and innovate from industry developments, while the community is increasingly looking to *foundation models* as the next promising approach. The following work is a brief review of current innovations and an outlook of possible future developments.

1. – The Transformer architecture for Science

In the past, Machine Learning architectures have often been developed by taking into account specific symmetries or domain-specific properties of the data. Thus, *Convolutional* Neural Networks were introduced to work on image-like data, while *Recursive* Networks were used for sequences and *Graph* Networks for sparse, unordered collections. This paradigm has produced remarkable results, but at the cost of having several different solutions, each specialized to its specific scientific domain or problem.

This changed in 2017 with the introduction of the *Transformer* architecture [2]. The key idea behind the Transformer is the use of *attention*, a simple but highly effective mechanism through which the network is capable of attending to the different parts of the input sequence and uncover the most important correlations between its parts.

The attention mechanism, as initially introduced in the context of language learning, is presented in fig. 1. Attention method simulates how human attention works by assigning varying levels of importance to different words in a sentence. It assigns importance to each word by calculating weights for the word's numerical representation, known as its embedding, to determine its importance.

Transformers networks, building on the attention mechanism, can thus be applied to any type of data which can be seen as a sequence of multiple elements. This can include electronic readouts, particle tracks, jet constituents listings, and in general almost any type of scientific data, also outside of HEP. Results are generally much better than

(*) IFAE 2024 - "New Technologies" session

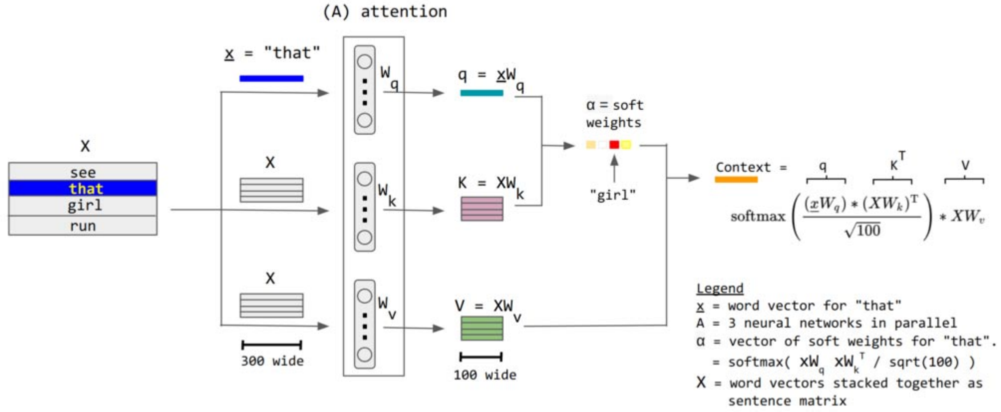
(A) attention
x = "that"
X
see
that
girl
run
X
X
300 wide
W_q
W_k
W_v
q = xW_q
K = XW_k
V = XW_v
100 wide
α = soft weights
"girl"
Context =
q
K^T
V
softmax ((xW_q) * (XW_k)^T / √100) * XW_v
Legend
x = word vector for "that"
A = 3 neural networks in parallel
α = vector of soft weights for "that".
= softmax(xW_q xW_k^T / sqrt(100))
X = word vectors stacked together as sentence matrix

Fig. 1. – The attention mechanism in action on an input sequence. Note that this approach can be applied to any data which can be seen as a sequence of tokens, *i.e.*, detector data, particle tracks, etc. From Numiri, CC BY-SA 4.0, via Wikimedia Commons.

previously proposed approaches, assuming the network is given enough training data. Additionally, any advancement in Transformer networks has a simultaneous impact on different data domains, since the architecture can be used in different data domains.

1˙1. *Example application.* – More specifically, in HEP this type of architecture is being used in very different domains, from calorimetry to pile-up mitigation. An interesting use case is the tagging of particle jets, where the CMS experiment has deployed a *particle Transformer* [1], which obtains state-of-the-art results by having the single constituents of the jets interact through the attention mechanism. This architecture is illustrated in fig. 2.

2. – Foundation Models for HEP

The Transformer architecture is increasingly being adopted in the HEP field. The emergence of such a powerful, general-purpose network has lead to the idea of creating a so-called *foundation model*. The main idea is that we can avoid training similar networks multiple times on different datasets for different tasks. Instead, we train a single network on a very large dataset, representative of our domain or experiment, and fine-tune it for each application scenario. The idea is illustrated in fig. 3.

This paradigm requires a loss function for the unsupervised learning of the foundation model on the initial dataset. *Contrastive learning* [3] is a possible solution, where the network is trained by comparing two different examples from the dataset, learning a space clustering together similar examples and separating different ones.

Once the pre-training step has been performed, the network can be fine-tuned to perform multiple sub-tasks, showcasing better results with less training data when compared with similar networks trained from scratch on the same sub-task.

2˙1. *Example application.* – The authors of [4] pre-trained the network on a dataset consisting of particle jets from different sources, and then fine-tuned it on the task of Top tagging, *i.e.*, identifying those jets coming from a Top quark decay. Figure 4 shows

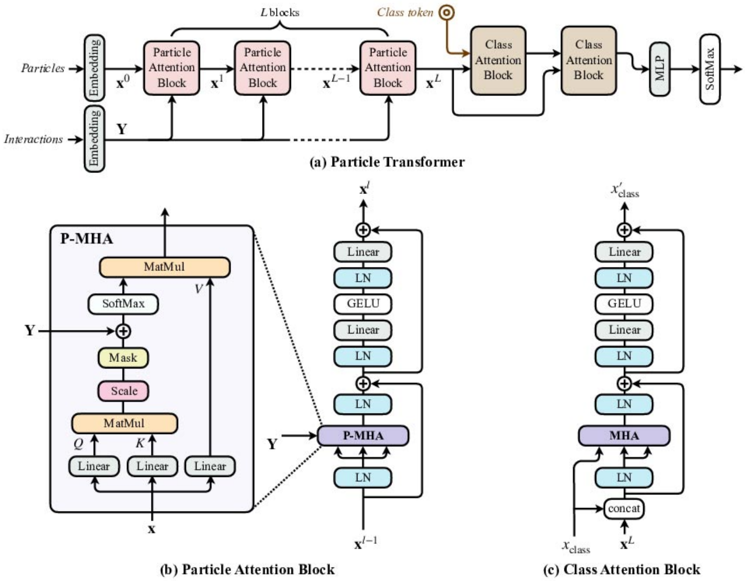

Fig. 2. – How the particle Transformer allows the constituents of the jets to "interact" through the attention mechanism for state-of-the-art results in jet tagging. From [1].

that the accuracy of the pre-trained tagger is consistently better than the tagger trained from scratch for each different number of labelled training samples.

3. – Shortcoming and pitfalls

When working with Transformers and foundation models, the speed and the advantages of these new approaches make it possible to delegate more and more of our decisions in analyses and detector studies. While in a certain way this is desirable, insofar as it

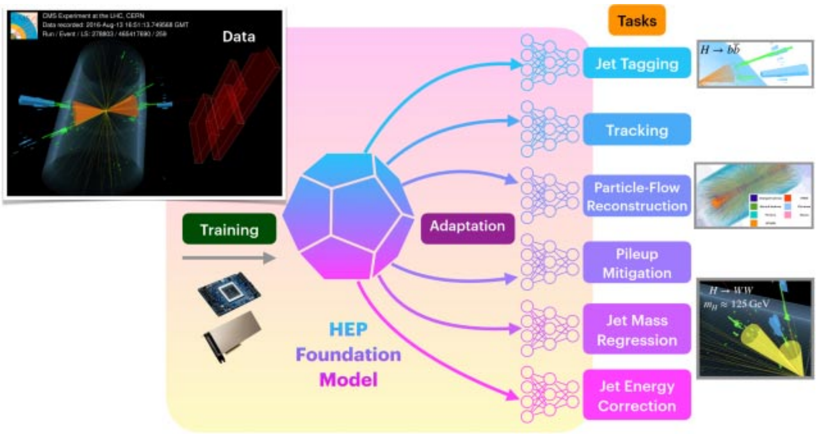

Fig. 3. – The main idea behind the design of a foundation model: train on general domain data and fine-tune on specific application modes. From [4].

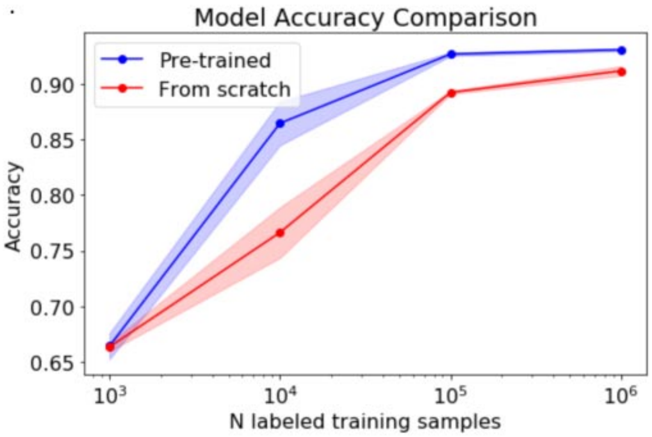

Fig. 4. – The foundation model consistently performs better than networks trained from scratch. From [4].

reduces human mistakes and the amount of time spent optimizing certain steps, on the other hand there are a number of scientists advocating for care in how much decision power we want to give to these networks.

The different steps in our current simulation or analysis chains give us the opportunity to make informed decisions and enrich the results through our physical understanding. It would therefore be desirable that these types of algorithms are applied to improve this existing approach, instead of replacing it completely in an end-to-end approach.

4. – Conclusions

We discussed how the Transformer architecture is innovating the landscape of Machine Learning application in Physics, by allowing the deployment of general-purpose, efficient solutions across a wide range of data domains.

Building on them, foundation models promise to reduce the training times of these algorithms while increasing their data efficiency.

Care must be taken to ensure that these applications remain a tool in the hand of the physicists, to empower our understanding of Nature instead of increasingly replacing our decisions in the discovery process.

We are confident that, thanks to these novel architectures as well as all the other exceptional efforts of the community, the future will see an increase of remarkable results across the field of High Energy Physics.

* * *

The author would like to acknowledge and thank the participants to the 2024 edition of the ACAT workshop, at Stony Brook, NY, for the helpful discussion and insights regarding ML applications to the field of HEP.

REFERENCES

[1] Qu Huilin, Li Congqiao and Qian Sitian, *Particle transformer for jet tagging* (2024).

[2] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Łukasz and Polosukhin Illia, *Attention is all you need*, in *Advances in Neural Information Processing Systems*, Vol. **30**, edited by Guyon I., Von Luxburg U., Bengio S., Wallach H., Fergus R., Vishwanathan S. and Garnett R. (Curran Associates Inc.) 2017.

[3] Weng Lilian, *Contrastive representation learning*, `lilianweng.github.io` (May 2021).

[4] Zhao Zihan, Mokhtar Farouk, Kansal Raghav, Li Billy and Duarte Javier, *Self-supervised learning (ssl) for jet tagging* (2024).