Email: nickdh98@gmail.com; cg801@cam.ac.uk
GitHub: github.com/SciCodePhy

Vision Transformers for End-to-End Particle Reconstruction for the CMS Experiment

Chenguang Guan
Department of Applied Mathematics and Theoretical Physics, University of Cambridge
(Dated: April 4, 2023)

Compact Muon Solenoid (CMS) detector is one of the most important components of the Large Hadron Collider (LHC) experiments, detecting the energies, momenta, and trajectories of the particles produced in the LHC collisions. However, traditional approaches potentially lose information, which limits the possibility of discovering new physics such as Beyond Standard Model (BSM). Therefore, machine learning techniques such as Vision Transformer can be used for end-to-end physics event classification. In this proposal, we propose various potential improvements of the current ViT-type Transformer, including new position embedding, hierarchical attention, an-isotropic attention, Fourier Transform linear mixer, and etc.

Contents

I. BACKGROUND AND MOTIVATION

"The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator" [1], where experiments are conducted to investigate the fundamental properties of matter. The Compact Muon Solenoid (CMS) detector is one of the most important components of the experiments, which measures the energies, momenta, and trajectories of the particles produced in the LHC collisions. [2]. In the CMS experiment, we need to first reconstruct the low-level detector data into progressively more physically motivated quantities until obtaining tabular-like particle-level data [5, 6]. "Traditional analysis approaches use these condensed inputs to construct an event classifier that capitalizes on the decay structure or topology of the processes involved" [5, 6]. However, traditional approaches potentially lose information, which limits the possibility of discovering new physics such Beyond Standard Model (BSM).

Therefore, a good alternative is end-to-end event classification method with machine learning techniques, which is the motivation of this project. Developers in the ml4sci community have developed CNN-based architectures such as ResNet-type architecture [4] for end-to-end classification [5, 6].

Furthermore, Large Language Model (LLM) becomes one of the most popular topics in the NLP and Machine Learning community. Motivated by recent advancement in LLM (such as GPT-4, ChatGPT and previous GPT-3 and Dall-E-2), this project aims to study the applicability of a large-scale transformer-based model for end-to-end physics event classification [3].

A. A Brief Review of Transformer and Vision Trasnformer

In the original Transformer for Natural Language Processing (NLP) [7], there are both attention-based encoders and decoders.

However, there is only attention-based encoder in ViT. The ViT introduces patch embedding to transform 2D image-patches to flatten embedding vectors ("Images to Words"), while they also develop some other positional embedding including 2-D/relative/learnable embedding beyond the 1-D positional embedding in NLP Transformer. Finally, they borrows the idea of class token from BERT [9] to classify the images.

- 1. Patch Embedding: $\mathbf{x} \in R^{H\times W\times C} \rightarrow \mathbf{x}_p \in R^{N\times\left(P^2\cdot C\right)}$
- 2. Class Token: Adding an extra token to gather information.
- 3. Position Embeidding: learnable 1D postion embeddings, 2D postion embeddings, etc.
- 4. Incorporating position information in the model: Added before feeding into the transformer encoder; Add before each encoder block; Added before each encoder block (shared weights).
- 5. Transformer Encoder Block: same as NLP transformer.
- 6. Multilayer perceptron (MLP) with the class token as inputs for classification task.

B. Bottleneck of Current Models

In the literature [5, 6] and my evaluation test (github.com/SciCodePhy/E2E_CMS_ml4sci_GSoC), we can find that the highest accuracy is about 75% and highest ROC-AUC score is a bit over 0.8, which is not so satisfying. The reason might be the differences between CMS data and image data.

Without considering the batch dimension (N_{events} or N_{samples}), the dimension of CMS data and image data are both 3-D. However, their physical meanings are considerably different. Three dimensions of CMS data are all spatial dimensions (r, ϕ, z) in cylindrical coordinates, which encodes positional information/locality (inductive bias).

Therefore, we might need some efficient mixer across all three dimension rather than 2-D convolutional layer.

II. PROPOSED DIRECTIONS BEYOND VIT

Because CMS data are considerably different from images in terms of sparsity, range of values, dimensions, etc, new models can be potentially developed in various directions. We can not only apply ViT-type Transformer to the CMS data, but can also explore models beyond ViT.

A. Position Embedding

The CMS data is in cylindrical coordinates (r, ϕ, z) rather than Cartesian coordinates (x, y, z). The symmetry is an important factor, which can be considered not only in position embedding, but also in attention mechanism, network architecture etc.

- We can use 3-D learnable embeddings rather than fixed 3-D embeddings.
- We can use $x = r\cos(\phi)$ and $y = r\sin(\phi)$ to re-design a fixed positional embedding, which integrates our priori knowledge about coordinate transformation into the position embeddings.

However, the original ViT paper shows that there are no significant performance differences between different positional embedding in ViT. Therefore, systematic numeric experiments are needed to determine whether other kinds of position embeddings are necessary.

B. Hierarchical Attention – Swin Transformer

Furthermore, we can also change the ViT-type architecture to Swin Transformer [10]. In our case, we can use attention layers that operate at the level of individual particles, clusters of particles, and the entire detector. We can coarse-grain all three dimensions of data simultaneously, rather than only coarse-grain on Height and Width dimensions in original Swin Transformer.

Denoting the size of each patch as (R, Φ, Z), we can have the following hierarchical attention structure:

$$(R*2, \Phi*2, Z*2) \to (R*4, \Phi*4, Z*4) \to (R*8, \Phi*8, Z*8) \to ...$$

Another point we should note is that the size of each dimension is not same. Therefore, we can stop coarse-graining one certain direction when achieving the limit.

$$(R*8, \Phi*8, Z*8 \text{ (max)}) \to (R*16, \Phi*16, Z*8 \text{ (max)}) \to ... \to (R*2^{n_R}, \Phi*2^{n_\Phi}, Z*2^{n_Z}),$$

where $(R*2^{n_R}, \Phi*2^{n_\Phi}, Z*2^{n_Z})$ is the max size of a patch (in the top layer).

C. An-isotropic Attention

We also propose a new attention mechanisms that operate along the radial, azimuthal and vertical directions separately. The reception field of a patch is along each direction:

- Radial direction: (R, Φ, Z): $(R \pm \Delta\Phi, \Phi, Z)$, $(R \pm 2\Delta\Phi, \Phi, Z)$, $(R \pm 3\Delta\Phi, \Phi, Z)$, ...
- Azimuthal direction: (R, Φ, Z): $(R, \Phi \pm \Delta\Phi, Z)$, $(R, \Phi \pm 2\Delta\Phi, Z)$, $(R, \Phi \pm 3\Delta\Phi, Z)$, ...
- Vertical direction: (R, Φ, Z): $(R, \Phi, Z \pm \Delta Z)$, $(R, \Phi, Z \pm 2\Delta Z)$, $(R, \Phi, Z \pm 3\Delta Z)$, ...

This can be as seen as a "three-head" masked attention. Three kinds of masks allow only the attention along three directions.

After operating attention along three directions, we can:

- sum three "head"s by brute-force;
- or concatenate three "head"s and use a feed-forward neural network (MLP) to transform three heads into one head.

D. Fourier Transform: FNet

Google research proposes a Transformer architecture "by replacing the self-attention sublayers with simple linear transformations that "mix" input tokens" [11]. This kind of linear mixer can also be transferred to vision and CMS data.

The 1-D Discrete Fourier Transform is defined as

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} nk}, \quad 0 \le k \le N-1.$$

Therefore, in the NLP case (which is also 1-D), the fixed weight matrix is:

$$W_{nk} = \left(e^{-\frac{2\pi i}{N} nk} / \sqrt{N} \right),$$

where $n, k = 0, \ldots, N-1$.

We can easily promote the 1-D Fourier Transformer mixer to 2-D case (vision) and 3-D case (CMS data).

E. MLP-based mixer

There are some work such as MLP-mixer and ResMLP [14, 15] showing that the MLP architecture based on patch embedding can have comparable performance. There are generally two kinds of mixer, one of which mixes the information of each feature dimension of the patch embedding, another of which mixes the information across the patches.

These mixers can also be incorporated into our architecture.

Furthermore, we can apply pruning to the MLP-mixer layer [16].

F. Graph Neural Network (optional)

CMS data can be seen as a variant of point cloud data in cylindrical coordinates. We can transform the CMS data to graphs based on different criterion. A naive criterion is to transform the CMS data to fully connected graph, which is simple but computational expensive. Another criterion is to use the local information and $k - NN$ to construct graph.

Based on the graph data, some graph-based models such as graph attention network [17] can be used.

G. Hybrid Architecture

We can definitely use hybrid architectures of the above proposed mechanism:

- Convolutional Layer (3D) + Self-attention Layer + Hierarchical-attention Layer + An-isotropic attention layer + Fourier Transformer linear mixer + MLP mixer.

For example, we can use convolutional layer to extract local features from the detector data and then feed the resulting feature maps into a transformer-based layer to capture global dependencies.

III. TIMELINE

A. Tasks, Objectives and Expected Deliverable

Task-1: Pushing the performance limit of ViT and existing models such as ResNet on CMS data:

- Hype-parameter tuning

Expected deliverable:

- Jupyter notebook with numeric experiment results.

Task-2: Incorporating various proposed techniques and tricks into the Vision Transformer:

- New position embedding
- Hierarchical attention – Swin Transformer (modified encoder)
- An-isotropic attention (modified encoder)
- Fourier transform-based linear mixer –FNet (modified encoder)
- MLP-based mixer: MLP mixer and ResMLP (modified encoder)
- Hybrid Architecture

Expected deliverable:

- Jupyter notebook with numeric experiment results to compare different techniques.
- Integrated project code as a package.
- Note and other documentations for the project.

B. Detailed Plan

Warm-up/Community Bonding Period: Interacting with the community, implementing the existing models and running numeric experiment on existing models.

Task-1: Pushing the performance limit of ViT and existing models such as ResNet on CMS data:

- Week1 (0 - 20 hours): Running experiments of exisisting models [5, 6]
- Week2 (20 - 40 hours): hyper-parameter tuning for ViT

Task-2: Incorporating various proposed techniques and tricks into the Vision Transformer:

- Week 3 (40 - 60 hours): Position Embedding
- Week 4 - Week 5 (60 - 100 hours): Hierarchical Attention (Swin Transformer)
- Week 6 - Week 7 (100 - 140 hours): An-isotropic Attention
- Week 8 (140 - 160 hours): Fourier Transform
- Week 9 - Week 10 (160 - 200 hours): MLP-based mixer
- Week 11 - Week 12 (200 - 240 hours): Hybrid Architecture based on the above structures.
- Week 13 - Week 14 (240 - 280 hours): Writing final reports

Remaining Time: reserved for flexibility, such as literature review, exploring other models proposed in the proposal.

C. Availability

My personal final examination period will be June 1 - June 15, during the time I am only available for lightweight coding.

IV. PERSONAL BACKGROUND

I am now a master student of Mathematics Part III program at University of Cambridge.

I obtained my undergraduate degree from Nanjing University in physics and deferred the entry to Brown Physics PhD program by one year to attend the Cambridge Math Part III program (specializing in theoretical physics and mathematical statistics). My undergraduate research background before Part III is in Machine Learning for Physics (AI for Science) and theoretical physics (condensed matter theory, theory and numeric). I have experience in different kinds of machine learning methods especially generative models, graph neural network-based methods and various representation learning methods, while I am familiar with various classical datasets in vision and NLP.

I have two projects in Machine Learning for Physics (AI for Science), one of which utilized Transformer and graph neural network in representation learning for atom (atom2vec), and another of which was about generative flow models informed by physics (renormalization group).

---

[1] The Large Hadron Collider https://home.cern/science/accelerators/large-hadron-collider
[2] https://home.cern/science/experiments/cms
[3] Vision Transformers for End-to-End Particle Reconstruction for the CMS Experiment https://ml4sci.org/gsoc/2023/proposal_E2E4.html
[4] https://arxiv.org/abs/1512.03385
[5] https://arxiv.org/abs/1807.11916
[6] https://arxiv.org/abs/1902.08276
[7] https://arxiv.org/abs/1706.03762
[8] https://arxiv.org/abs/2010.11929
[9] https://arxiv.org/abs/1810.04805

[10] https://arxiv.org/abs/2103.14030
[11] https://arxiv.org/abs/2105.03824
[12] https://arxiv.org/abs/2012.09841
[13] https://arxiv.org/abs/2212.09748
[14] https://arxiv.org/abs/2105.01601
[15] https://arxiv.org/abs/2105.03404
[16] https://arxiv.org/abs/1803.03635
[17] https://arxiv.org/abs/1710.10903