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

Transformers for Dark Matter Morphology with Strong Gravitational Lensing

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

Strong gravitational lensing can provide valuable information about the distribution of matter in the universe, including the distribution of dark matter, which cannot be directly observed. This project aims to study the applicability of a large-scale transformer-based model for lensing image classification and regression. In the proposal, we also propose various potential improvements of the current ViT-type Transformer, including new position embedding, hierarchical attention, anisotropic attention, Fourier Transform linear mixer, and etc.

Contents

I. BACKGROUND AND MOTIVATION

Strong gravitational lensing can provide valuable information about the distribution of matter in the universe, including the distribution of dark matter, which cannot be directly observed. "Deep learning methods have the potential to accurately identify images containing substructure, and differentiate WIMP dark matter from other well motivated models, including vortex substructure of dark matter condensates and superfluids" [5], which is called DeepLense [2, 5].

Developers in the ml4sci community have applied various pre-trained models (such as ResNet, VGG, DenseNet and AlexNet) and other methods including unsupervised learning to lensing images [3, 5].

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 lensing image classification and regression [2].

A. A Brief Review of Transformer and Vision Trasnformer

In the original Transformer for Natural Language Processing (NLP) [6], 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 [8] 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. Failure of Toy Models

In my evaluation task (github.com/SciCodePhy/DeepLense_ml4sci_GSoC), we can find that shallow models such as MLP, LeNet, two-block ResNet's performances are poor on lensing image classification task, while ResNet has much better performance than VGG and Alexnet.

The reason might be the differences between lensing images and images in real life. There are high similarities between lensing images with different kinds of substructure (FIG. 1).

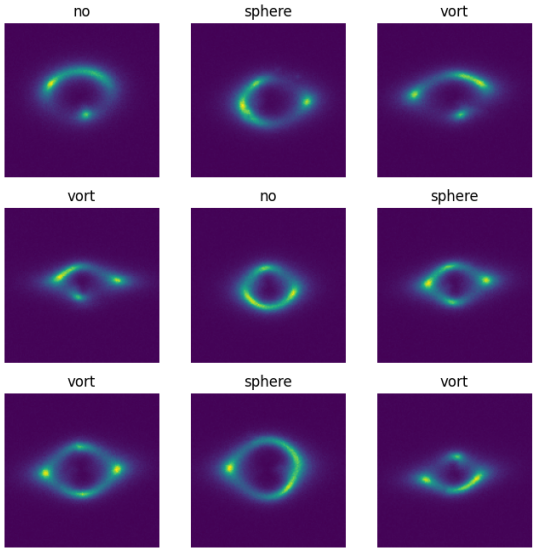

FIG. 1: Lensing images with three kinds of substructure

II. PROPOSED DIRECTIONS BEYOND VIT

We can not only apply ViT-type Transformer to the gravitational lensing, but can also explore models beyond ViT. New models can be potentially developed in various directions.

A. Position Embedding

The original ViT paper shows that there are no significant performance differences between different positional embedding in ViT. However, this might be different in lensing image data:

- 1-D fixed position embedding
- 2-D fixed position embedding
- learnable position embedding

B. Hierarchical Attention – Swin Transformer

Furthermore, we can also change the ViT-type architecture to Swin Transformer [9]. The lensing images are coarse-grained along height and width dimensions. Denoting the size of each patch as (H, W), we can have the following hierarchical attention structure:

$$(H, w) \to ((2H, 2W) \to (4H, 4W) \to ...$$

C. Fourier Transform: FNet

Google research proposes a Transformer architecture "by replacing the self-attention sublayers with simple linear transformations that "mix" input tokens" [10]. This kind of linear mixer can also be transferred to vision and lensing 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).

D. MLP-based mixer

There are some work such as MLP-mixer and ResMLP [13, 14] 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 [15].

E. Hybrid Architecture

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

- Convolutional Layer (2D) + 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 lensing and then feed the resulting feature maps into a transformer-based layer to capture global dependencies.

F. Data Augementation

In our numeric experiment (github.com/SciCodePhy/DeepLense_ml4sci_GSoC), we take the random rotations up to 0 degrees as the method of data augmentation. However, in some cases, we find that the data augmentation will make the training more difficult. Therefore, systematic numeric experiments are needed to determine the effect of data augmentation on lensing image dataset.

III. TIMELINE

A. Tasks, Objectives and Expected Deliverable

Task-1: Pushing the performance limit of ViT and other existing models (such as pre-trained models) [3, 5]:

- Hype-parameter tuning
- Numeric Experiment for Data Augmentation

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 other existing models (such as pre-trained models) [3, 5]:

- Week1 (0 - 20 hours): Running experiments of exisisting models [3, 5]
- Week2 (20 - 40 hours): hyper-parameter tuning for ViT and numeric experiment of data augmentation

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] GSoC 2023 Projects Related To DEEPLENSE https://ml4sci.org/gsoc/projects/2023/project_DEEPLENSE.html
[2] Transformers for Dark Matter Morphology with Strong Gravitational Lensing https://ml4sci.org/gsoc/2023/proposal_DEEPLENSE4.html
[3] https://arxiv.org/abs/2008.12731
[4] https://arxiv.org/abs/1512.03385
[5] https://arxiv.org/abs/1909.07346
[6] https://arxiv.org/abs/1706.03762
[7] https://arxiv.org/abs/2010.11929
[8] https://arxiv.org/abs/1810.04805
[9] https://arxiv.org/abs/2103.14030
[10] https://arxiv.org/abs/2105.03824
[11] https://arxiv.org/abs/2012.09841
[12] https://arxiv.org/abs/2212.09748
[13] https://arxiv.org/abs/2105.01601
[14] https://arxiv.org/abs/2105.03404
[15] https://arxiv.org/abs/1803.03635
[16] https://arxiv.org/abs/1710.10903