Mini-Project 2-3 Proposal

Shunyi Zhu, Zhen Wang, Haoze He
Department of Eletrical and Computer Engineering
Tandon School of Engineering, New York Univeristy
sz3719, zw2655, hh2537@nyu.edu

1 Key Idea and Datasets

We will take several datasets for training and testing. The expected dataset will be Image-Net-1K image classification with 1.28M training images and 50K validation images from 1000 classes.

1.1 Combiner of Transformer and Convolution Neural Network

Both convolution neural network(CNN) and vision transformer(ViT) are success in past few years. They are two dominant frameworks in the field of computer vision. However, these two frameworks have their own drawbacks. As for CNN, the convolution layer has limited respective field, which makes it hard to capture the associated information between global and local area. As for ViT, it can capture the associated information using self-attention. But without similarity comparisons among all tokens also leads to high redundancy. In addition, the low-efficiency flaw of ViT transformer caused by high computational/space complexity in Self-Attention. To combine the advantage of both ViT and CNN, we plan to reduce the number of tokens using submodular optimization, merge tokens with high similarity into one, and use it as input for CNN architecture.

1.2 Vision Transformer

Transformer have shown extraordinary success in the field of computer vision in recent years. However, the redundancy of token and high computation/space complexity of self-Attention remain to be problem. To solve these problems, we will try to reduce the number of tokens using submodular optimization and propose new self-attention mechanism to reduce overall complexity.

2 Deliverable

In the final demo, We will test our framework on the given datasets. The performance of our algorithm will be evaluated by comparing with the baseline provided by Transformer and Convolution Neural Network in different perspectives. Our expected goal is to combine the strengths given by Transformers and CNN to increase the learning accuracy. We will also try to reduce the complexity of self-attention. Finally, all the code will be open source on github.

References

[3] Lu, Jiachen (2021) SOFT: Softmax-free Transformer with Linear Complexity. *NeurIPS*.

[2] Kunchang Li, Yali Wang (2022) UniFormer: Unifying Convolution and Sqelf-attention for Visual Recognition *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[3] Lu, Jiachen (2021) SOFT: Softmax-free Transformer with Linear Complexity. *NeurIPS*.

[4] Ma, Teli (2021) Oriented Object Detection with Transformer. *Computer Vision and Pattern Recognition (cs.CV), FOS: Computer and information sciences, FOS: Computer and information sciences*, arXiv.

[5] Naseer, Muzammal (2021) Intriguing Properties of Vision Transformers. *Computer Vision and Pattern Recognition (cs.CV), Artificial Intelligence (cs.AI), Machine Learning (cs.LG), FOS: Computer and information sciences, FOS: Computer and information sciences*. arXiv.