Project Report

Machine Learning for Computer Vision

by Hanying Zhang

University of Bologna

June 11, 2022

1 Executive Summary

In this project the Vision Transformer(ViT)[1] paper is studied and implemented. An experiment is also carried out to test the performance of the model on a new dataset. The visual attention and position embedding are also studied.

2 Vision Transformer

Transformer[2] was firstly introduced in Natural Language Processing field and has dominated the field since then. Inspired by the success of Transformer in NLP, the Vision Transformer(ViT) paper applied a standard transformer directly to images, using as few as possible modifications. The paper shows that when trained on large($14M - 300M$) images, ViT could get excellent results when pre-trained at sufficient scale and transferred to tasks with fewer datapoints.

2.1 Model Overview

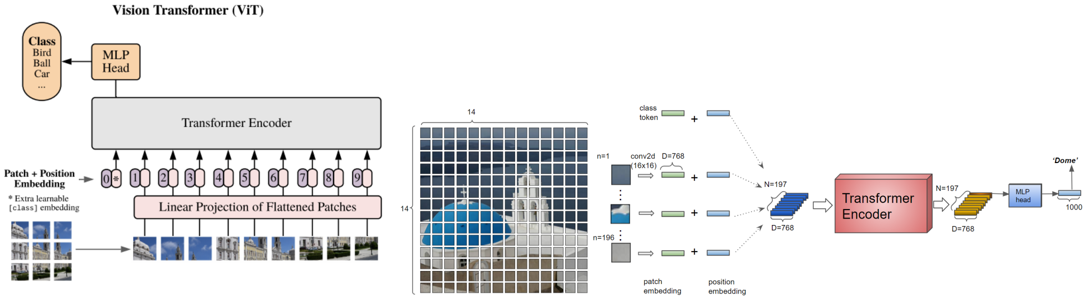

Figure 1. The main structure of ViT

The figures above show the main structure of ViT. The left figure is from the original paper, and the right one is more detailed. It can be seen from the figures that the model design is followed the original Transformer as closely as possible.

Generally speaking, the input image is reshaped into a sequence of flattened 2D patches. These patches are then flattened and mapped to constant D dimensions with a trainable linear projection(or a 2D convolution), which are named patch embeddings.

A learnable class token which represents the image class is prepended to the sequence of the patch embeddings. Position embeddings are also added to the patch embeddings to retain position information, forming the input of Transformer encoders.

[1] https://arxiv.org/abs/2010.11929

[2] https://arxiv.org/abs/1706.03762

The Transformer encoder is slightly different from the one in the original Transformer paper. In this model the encoder consists of alternating layers of multiheaded self-attention and MLP blocks. Layernorm is applied before every block, and residual connections after every block, as shown in the figures below.

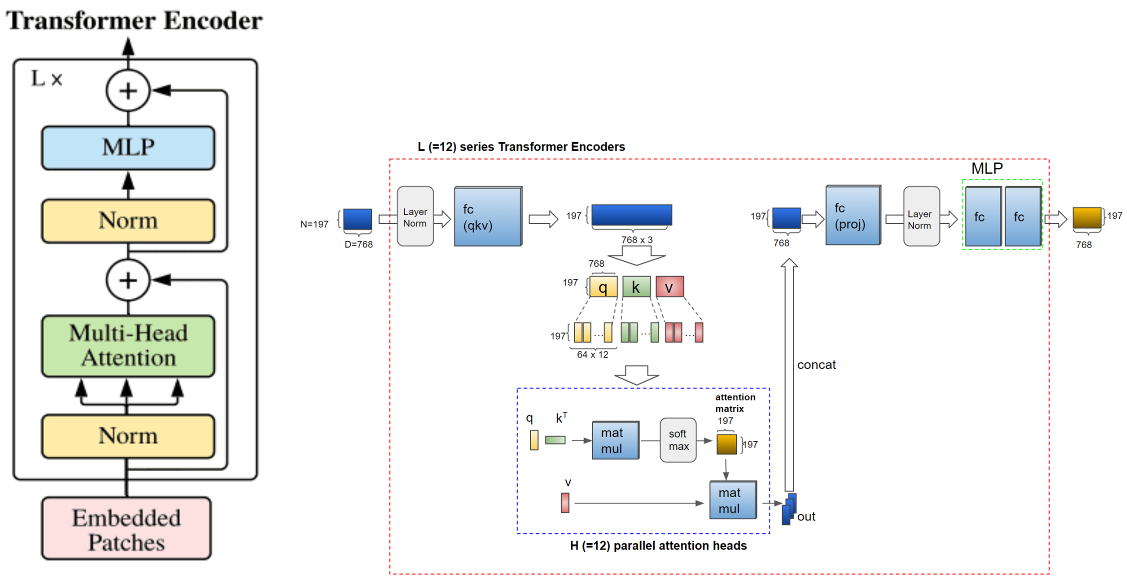

Figure 2. Transformer Encoder

When fine-tuning for downstream tasks, the pre-trained prediction head is replaced by a $D \times K$ feedforward layer, where K is the number of downstream classes.

3 Experiments

The main experiment is a classification task of a new dataset using ViT model as backbone, before and after fine-tuning. After finishing the task, the position embeddings and Transformer attentions are to be visualized.

3.1 Dataset

The dataset used here is Dogs Vs. Cats[3] on Kaggle. But the competition is already finished and the test is not available. However, another website providing similar compitioin which still accepts test submissions is found. The dataset of this new competition is slightly different from the original one. The training data contains $20k$ images and the validation and test sets both contain $2k$ images.

The *Dataset* and *DataLoader* classes from *Pytorch* are used to customize a new *Dataset* class to import all this dataset.

3.2 Fine-tune

A pre-trained model weights is used in this project. Specifically, the model used is '*jx_vit_base_p16_224-80ecf9dd.pth*'. The performance before fine-tune is also important as a benchmark. After loading the weights into the model, the head needs to be modified according to the number of classes in the new dataset. As the new dataset has only two classes, the new head is a FC layer whose weight is $[2, 768]$, where $D = 768$ is the output dim of the transformer. During this experiment, all parameters are fixed except for the ones in the new head.

Then, during the fine-tune process, all parameters in the model including the new head are made trainable.

[3] https://www.kaggle.com/c/dogs-vs-cats

In both experiments, the training process run for 5 epochs, the batch size are both 16 and the optimizer are both *Adam*. The learning rate is set to 0.0001 and a CosineAnnealingLR decay is used.

3.3 Results and Analysis

The test accuracy before and after fine-tune are 96.5% and 98.65%. It's obvious that after fine-tune, the accuracy would become better.

Training with more epochs and bigger learning rates are also explored, but these parameters failed to give better results.

3.4 Position Embedding Visualization

Several variations of position embedding methods were tried in the original paper, i.e., no positional embedding, 1D positional embeddings, 2D positional embeddings and relative positional embeddings. The results showed that only the case of no positional embeddings got worse results. The rest 3 cases showed little difference, which means that 1D positional embedding is enough for this model.

Thus I would like to visualize the 1D positional embedding for explanation. The method here is computing the *cosine similarity* between the $i-$th embedding and **all** the embeddings(except for the *class_ token* one).

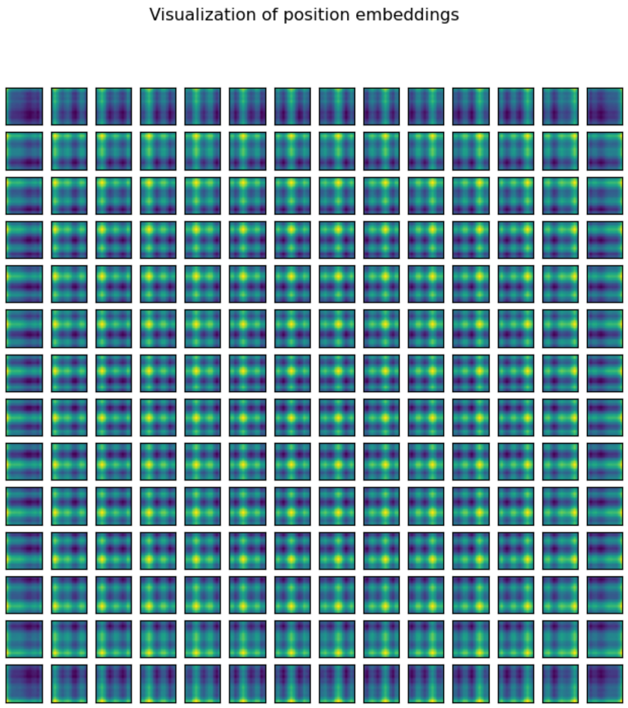

Figure 3. Similarity of Positional Embeddings

From the above figure we can see that each patch of the image has the embedding correctly representing the positions in the image. Thus we can say that 1D positional embedding is enough for this model.

3.5 Visual Attention

The visual attentions, i.e., matmul$(Q, K^T)/_{\sqrt{D_K}}$, is also visualized in order to see how the attention mechanism works. One image from the test set which contains two objects is selected to show the attentions.

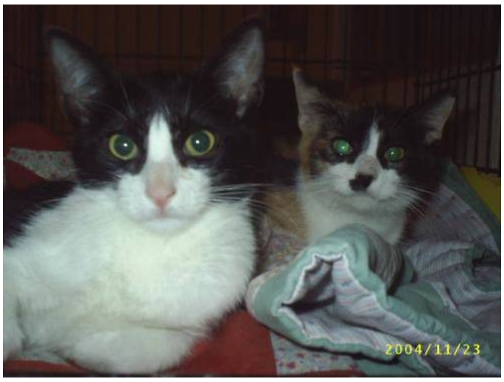

Figure 4. Selected Image to show visual attention

The following two figures show that the areas of the two objects(left) and the patch positional information(right) are being considered.

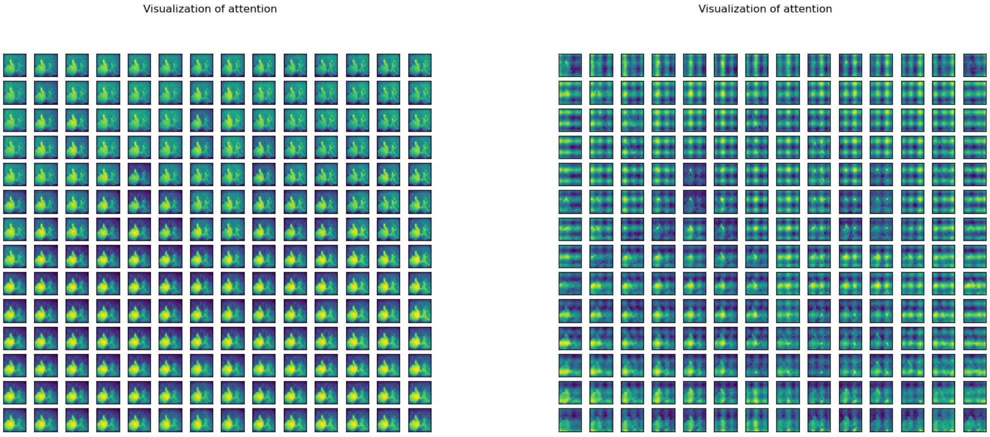

Figure 5. Attention Visualization I

The information about horizontal(left) and vertical(right) positions where the objects are could also been shown in the attentions.

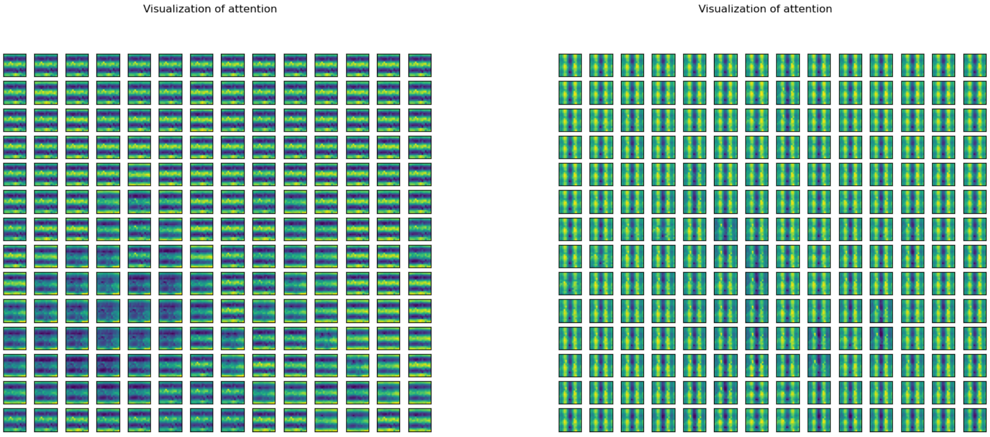

Figure 6. Attention Visualization II

4 Conclusion

In this project an experiment was carried out to test the performance of ViT as the backbone in a classification task. Specifically, a new dataset was used to test the model before and after fine-tune. This experiment proved that ViT matches or exceeds the state of the art on many image classification datasets.

After fine-tune, the positional embedding and attention were also studied. According to the figures observed, it is convincible that the 1D positional embedding is good enough for this model and visual attentions are also functioning very well.

More experiments, such as self supervised learning, could also be carried out on this model to unveil more information of ViT model. Moreover, there have been many improvements and variants of ViT model, such as Simple ViT[4], Distillation[5] and Deep ViT[6], a study on these models could also be carried out. More efficient attention mechanisms[7] have also been proposed, studying on these papers would also be valuable.

[4] https://arxiv.org/abs/2205.01580

[5] https://arxiv.org/abs/2012.12877

[6] https://arxiv.org/abs/2103.11886

[7] https://arxiv.org/abs/2009.14794