Specialized Machine Learning Technologies Course Project

Affandy Fahrizain (J41332c)

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Introduction

Self-attention-based model like Transformers has successfully become a new trend and state-of-the-art in the most NLP tasks. While in the Computer Vision area, CNN-based model still widely used. Inspired by the successful application of attention in NLP, there are several evidence to apply attention mechanism in Computer Vision task too. Some of them combined CNN-based architecture with attention, while some others are trying to fully replace convolutions. On the selected paper, the authors try to replace convolutions with vanilla Transformers model(Dosovitskiy et al., 2021).

Application of self-attention in images requires each pixel attends to every other pixel, which of course will be very expensive in terms of computational. Hence, several techniques have been applied such as self-attention only in local neighborhoods instead globally, using local multi-head dot-product self-attention blocks to completely replace convolutions, postprocessing CNN outputs using self-attention, etc. All these techniques shown promising results on computer vision tasks, but quite hard to be scaled yet and requires complex engineering to be implemented efficiently on hardware accelerators. On the other hand, as the Transformers' model is based on MLP networks, it is more computational efficient and more scalable especially if one want to train big models with over than 100B parameters.

Code experiment available at: https://github.com/fhrzn/sml-tech/

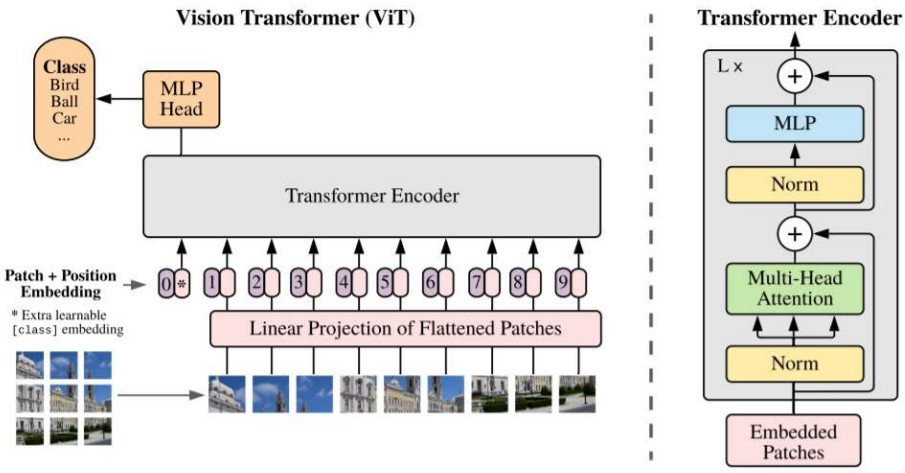


Figure 1. Vision Transformer Illustration

Another approach was very similar with the selected paper, that is extract patches of image with size 2x2 and apply self-attention on top of that later. Nevertheless, it only applicable to small-resolution image due to its patch size, while on this work medium-resolution images will be handled as well.

As Transformer requires sequences as the input which very different approach with CNN-based models, the authors decide to extract images into fixed-size patches, add position embedding to each patch, and finally prepend extra learnable "classification token" like [CLS] token in BERT to the sequence. Similar to BERT CLS token, the prepended classification token intended to represent image and would be fed into classification head later.

Model

The architecture of Vision Transformer (ViT) is like vanilla Transformer as depicted in Figure 1. The main difference is only at the input embedding. In this work the authors presented Patch Embedding, that is a way to convert image into sequences by splitting into patches. In the input part, the Position Embedding is added to retain positional information of patch. Like the original Transformer architecture, inside

Transformer Encoder layer there are exists Multiheaded Self-Attention (MSA) layers and MLP blocks with Layer Norm (LN) and Residual Connection applied before and after every block respectively. Lastly, an MLP classification head with one hidden layer.

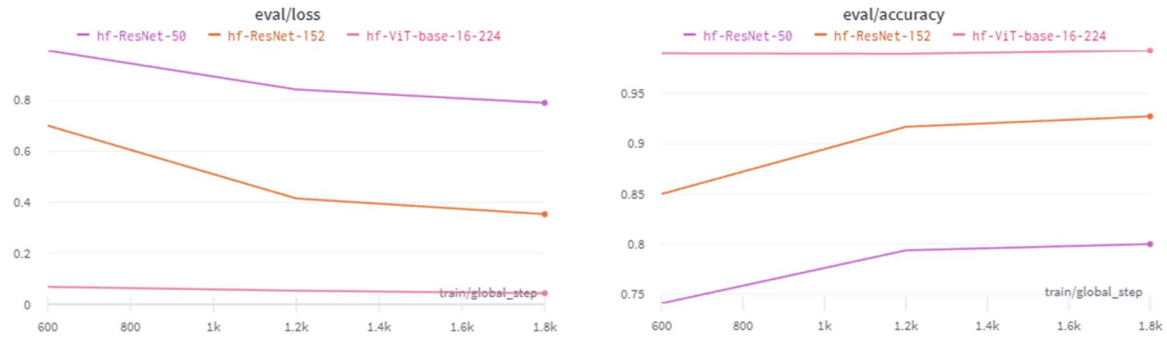
As the authors want to completely replace CNN with Transformer, there is a shortcoming within this approach. The Vision Transformer has much less image-specific inductive bias than CNNs which its locality, two-dimensional neighborhood, and translational equivariance are transferred into each layer. On the other hand, MLP layers are local and translationally equivariance while the self-attentions are global in ViT.

Experimental Results

There are 3 ViT model variants released by authors, namely ViT-Base, ViT-Large, and ViT-Huge with different number of layers, hidden layers, MLP size, attention heads, and params. All these models are pretrained on large dataset such as ImageNet, ImageNet-21k, and JFT which then make it possible to perform well on the downstream tasks.

Due to limited resources, we performed image classification task using ViT-Base variant. We used the huggingface implementation of pretrained model ViT. We performed zero-shot and fine-tuning on Shoe vs Sandal vs Boot image dataset contains 15k images that publicly available in Kaggle. We also compare ViT model performances with ResNet50 and ResNet152. We use Kaggle environment with Tesla P100 GPU. We fine tune all models in 3 epochs, with learning rate 1e-5, and batch size 16.

Table 1 and table 2 show model performances zero-shot and fine-tuned scenario on test set. In zero-shot scenario, ResNet-152 outperform ViT and ResNet-50.



(a) (b)

Figure 2. (a) Validation Loss and (b) Validation Accuracy Plot

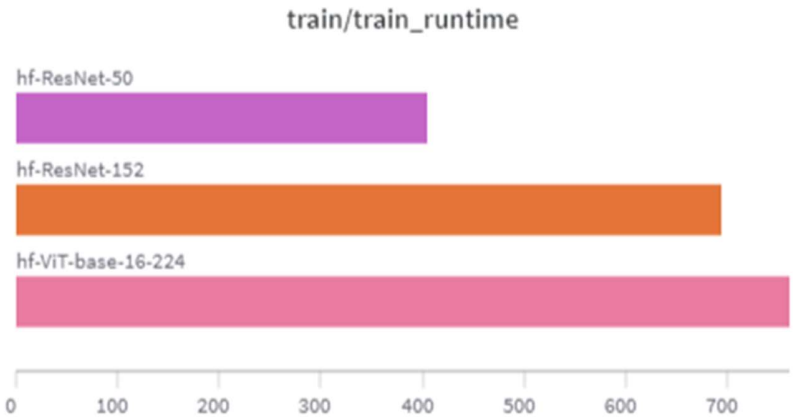


Figure 3. Training runtime

Table 1. Zero shot performances

	ViT	**ResNet-152**	**ResNet-50**
Accuracy	32.90	**42.30**	36.60
F1 Score	30.70	**36.20**	26.90

Table 2. Fine-tuned performances

	ViT	**ResNet-152**	**ResNet-50**
Accuracy	**99.33**	92.76	80.03
F1 Score	**99.33**	92.77	79.62

Surprisingly, ResNet-50 also outperform ViT in Accuracy but still behind in F1 Score. On the other hand, after 3 epochs of fine-tuning ViT successfully outperform the rest models, followed by ResNet-152 and ResNet-50 in the last position.

In the fine-tuning scenario, as depicted in Figure 2 (a) and (b), ViT model already reach very low loss and high accuracy in validation set. While the ResNet models performance improving by several points on each epoch. Another interesting plot is runtime plot, that as we can see ViT model took longest time to train compared to ResNet models. We assume this is happened because the complexity of models are different, for example ResNet-50 indeed having shallower network and simpler complexity compared to ViT and ResNet-152.

Conclusion

Vision Transformer (ViT) is a Transformer based model that intended to apply self-attention and completely replace CNNs in computer vision tasks. The authors introducing Patch Embedding to convert image into sequences which is necessary as Transformer input.

In this task, we also performed zero-shot and fine-tuned experiments with our selected dataset. As the competitor of ViT, we employed ResNet-152 and ResNet-50 for the experiments. The result shows in the zero-shot scenario ViT have lowest performance in accuracy compared to the others, while ResNet-152 having the best performance. On the other hand, in fine-tuned scenario ViT outperforms the ResNet models. We also showed that ViT already outperforms the other since first epoch of fine-tuning, while the ResNets needs several epochs to improve its performance.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.