AIN SHAMS UNIVERSITY

FACULTY OF ENGINEERING

Computers and Systems Engineering Department

Paper Summary Report

CSE616 Neural Networks and Their Applications

Submitted by

Mohammad Ahmad Khattab Mousa

Student Details

Name	ID
Mohammad Ahmad Khattab Mousa	2002639

Contents

1. Report Introduction

This report provides a summary for the paper **“An Image is Worth 16x16 Words: Transformer for image recognition at scale**” **[1]** by a group of researchers of Google Brain Team. The report has a section to summarize the contents of each section of the paper. The report is available over the github repository **[2]**.

2. Abstract

While transformer architecture has became the standard model for natural language processing (NLP) tasks, it is showing a slower progress for computer vision tasks. In computer vision, attention mechanism is used as an additional layer in conjunction with normal convolutional network layers. In this paper, the authors propose to use the attention mechanism without convolutional layers for image classification tasks. The authors claims that such model outperforms the well-know convolutional architectures.

3. Introduction

Transformer models are showing great progress over natural language processing tasks. Due to their computational efficiency, researchers were able to train models with more than 100 billion parameters.

In the opposite side, convolutional models are still the dominant models in computer vision tasks. Some researchers tried to add attention as a neural layer within a CNN architecture. In this work, the authors try to use attention mechanism exclusively without making use of any convolutional layers. The idea is to split an image into a sequence of sub-images (patches) which are processed through different linear embedding layers. The output of these layers is a sequence of vectors which are used as an input sequence to the well-known transformer model commonly used in NLP tasks. The model is used to solve supervised image classification problem and hence a many-to-one architecture is chosen.

When training the model over ImageNet dataset, the model achieves accuracy lower than ResNet architectures of comparable size. This is due to the fact transformers lacks the inductive biases inherited in CNN architectures like transformation equivariance, transformation invariance and input locality.

Due to missing such inductive biases, transformer model needed to have pre-training over extremely large datasets in order to achieve accuracy which is better than benchmark values recorded for the state-of-the art convolutional networks.

4. Related Work

Transformer models were proposed in the paper **“Attention is all you need” [3]** where attention layer was used to replace recurrent units in models dealing with data presented as a sequence. Later on, transformer model dominated the NLP translation domain within few years**.** Many trials were made to utilize attention layers in Computer Vision tasks. Basic application of attention layer requires to apply attention between each pair of pixels of an image which is not scalable and computationally inefficient. In addition, such way does not make use of the concept of locality within images. Researchers worked in different direction in order to apply self-attention in a more scalable and efficient ways. Some researchers applied attention between each pixel and the neighboring pixels. Other researchers applied attention over image blocks of a varying size. Other researchers applied attention over image blocks of 2x2 pixels. In this paper, the authors applied attention over image blocks of 16x16 pixels.

5. Method

5.1 Vision Transformer

The authors used the same architecture proposed in **[3]** while proposing a new layer to embed an image into sequences of input vectors. The model starts by splitting an image into a sequence of flattened 2D 16x16 patches. If we assume the resolution of the image is (H,W), number of channels (C), the resolution of the patch is (P,P), this leads to a patch size of P*P and number of patches $N = HW/P^2$. Flattened patches are mapped to D dimension with a trainable projection. A new learnable patch CLS is prepended to the sequence for classification purpose. The position of the patches is embedded and added to the patches in order to encode the spatial information of the patches. This sequence of embedded patches is fed as the input to the transformer model. The transformer block consists of interleaving layers of layer normalization, multi-head self-attention and Multi-Layer Perceptron. Skip connections are added as of the diagram shown in Figure1. Finally, a classification head of multiple fully-connected layers are added to CLS embedding (first vector in the sequence). GELU non-linearity is used in all MLP layers as the activation function.

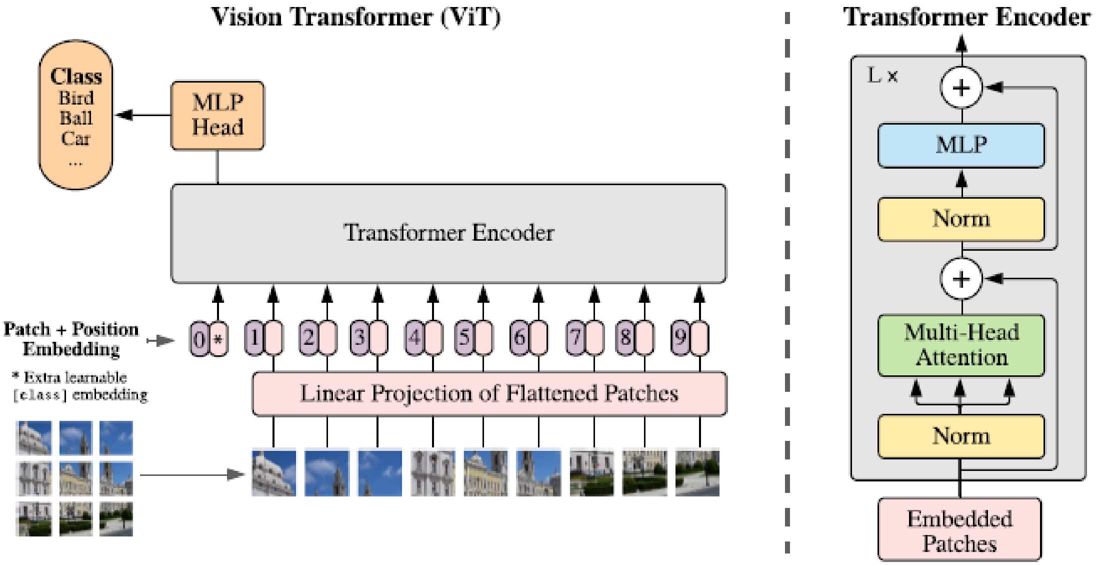


Figure 1 Architecture of Vision Transformer

5.2 Hybrid Model

In this approach, a CNN is used to process the images and resulting feature maps are used as the input to the embedding layer instead of dividing the image into patches.

5.3 Fine Tuning over higher resolution

The authors propose to pre-train the model on large dataset and then fine-tune to smaller downstream tasks instead of training from scratch over the downstream datasets. The classification head used during pre-training is removed and a new fully connected layer is used for fine-tuning. It is recommended by the authors to fine-tune over higher resolution images. When fine-tuning over higher resolution images, the authors propose to keep the same patch size which will lead to longer input sequence. The position embedding values are interpolated to avail the position embeddings of the new sequence length.

6. Experiment

The authors compare the performance of the Vision Transformer (ViT) model, ResNEt model and the hybrid model.

6.1 Setup

Pre-Training Datasets The authors pretrained the models using the below datasets:

1. ImageNet dataset with 1k classes and 1.3M images.
2. ImageNet-21k dataset with 21k classes and 14M images.
3. JFT-300 dataset with 18k classes and 303M high-resolution images.

Transfer Learning Tasks The trained models were transferred to several benchmark tasks which are ImageNet with the original validation labels, cleaned-up Real Labels, CIFAR-10/100, Oxford-IIIT Pets and Oxford Flowers-102. The training datasets were de-duplicated with respect to the test sets of the downstream tasks. Finally, the models were evaluated using 19-task VTAB classification suite.

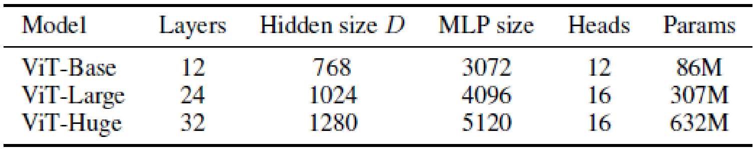
Model Variants The authors used three variants of Vision Transformers (ViT) which are ViT-Base, ViT-Large and ViT-Huge variants. These models were tested in combination with different image patch sizes. For example, ViT-L/16 means the "Large Variant Architecture" with 16x16 input patch size. For the ResNet model, the authors replaced the "Batch Normalization" with "Group Normalization" and used standardized convolution. With such changes, the ResNet Model is labeled as "ResNet (BiT)".

Training and Fine-Tuning Adam optimizer was used for all models with β_1 = 0:9, β_2 = 0:999,batch size of 4096 and weight decay of 0.1. For fine-tuning, SGD with momentum was used with a batch size of 512 for all models.

Metrics The authors mainly focused over the fine-tuning accuracy to access the model performance. When fine-tuning accuracy is too expensive, linear few-shot accuracies were used for fast on-the-fly evaluation.

6.2 Comparison to the State of the Art

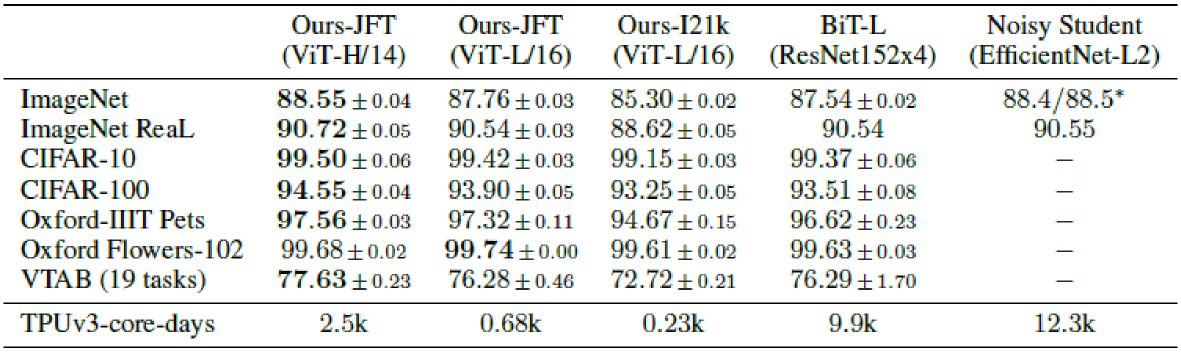
The performance of different variants of ViT models were compared to modified ResNet (BiT model) and modified EfficientNet (Noisy Student model). Table1 shows the architecture of different variants of the ViT models.



Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1 Architecture of different Vision Transformer model variants

Table2 shows the accuracy of different models over different tasks. First row is the pre-training dataset used for each model. First column is the target transfer learning task. Last row is the number of training days used to pre-train each model using the same hardware resources. BiT was pre-trained over JFT-300M dataset which is not mentioned in the table. Noisy Student model was pre-trained using semi-supervised learning on ImageNet and JFT-300M with the labels removed.



	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Table 2 Performance Comparison of ViT models and state-of-art CNN models

6.3 Pre-training Dataset size requirement

In this section, the authors show how training dataset size impacts the model performance. It is observed that increasing the training dataset size, enhances the model accuracy with the models achieving best results when pre-trained over JFT dataset (303 Million images). The results are summarized in Figure2. This is consistent with the intuition that CNN models could easily learn from small datasets due to the inductive biases inherited from the model architecture. For large datasets, transformer models have better accuracy due to learning the relevant patterns directly from the data.

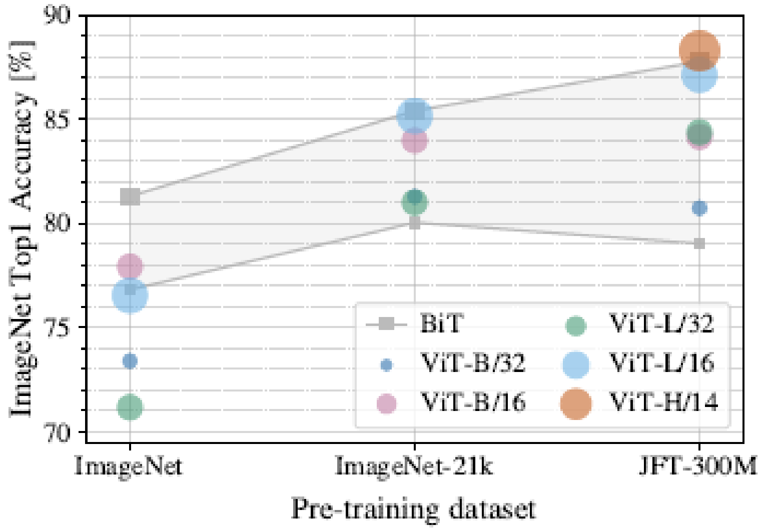


Figure 2 Models Accuracy when pre-trained over different datasets with different sizes

6.4 Scaling Study

Checking the pre-training computational cost, it is found that in general ViT models perform better that ResNet models with the same pre-training computational cost. Hybrid models (having CNN for patch generation) show better accuracy for small and medium model size but the difference disappears for large model size.

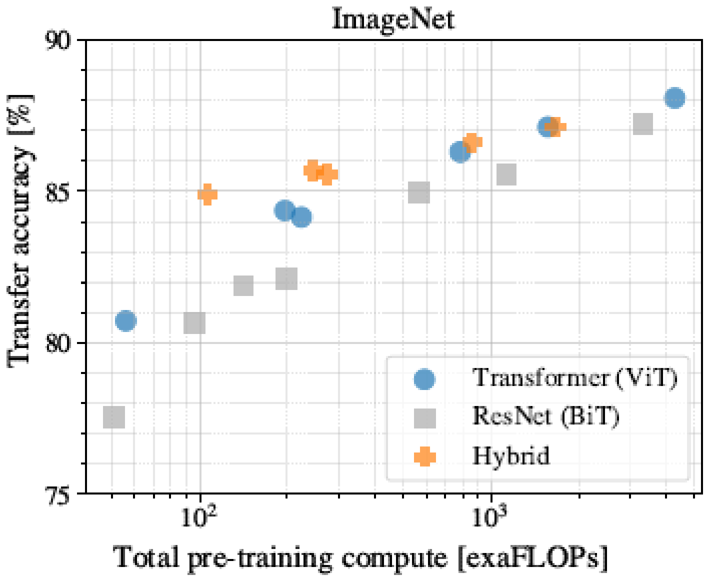


Figure 3 Comparing the accuracy of different models with the same pre-training computational cost

6.5 Self-Supervision

Vision transformers requires a huge pre-training dataset and building of such huge dataset is not a simple task. As a result, the authors tested self-supervision for pre-training vision transformer. Masked patch prediction for self-supervision was used for pre-training ViT-B/16 model and model achieved 79% accuracy over ImageNet lagging by 4% behind supervised pre-training. Contrastive pre-training was left for future work.

7. Conclusion

Vision transformer showed accuracy enhancement compared to CNN models when pre-trained over huge datasets and transferred to target downstream tasks. Training directly over the target dataset showed low accuracy as vision transformers lack many of the inductive biases found in CNN networks.

Authors recommended Future work could be summarized as follows:

1. Use Vision Transformer for computer vision tasks other than image classification
2. Explore better options to pre-train vision transformer using self-supervised techniques

8. References

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[2] https://github.com/Gr8Job/CSE616-Vision-Transformer

[3] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).