BEiT: BERT Pre-Training of Image Transformers Reading Report

Tianyu Shi

June 30, 2022

1 Introduction

Transformer has already achieved promising performance in computer vision. However, it shows that vision Transformer requires far more data to train than CNNs. To address this problem of requiring large amounts of data for training, self-supervised pre-training is a promising approach that can make use of large-scale image data. This paper addresses the problem of self-supervised learning of visual models with generative pre-training. They propose a BEiT model, similar to BERT in natural language processing, that reduces the reliance on annotated data and thus enables efficient visual pre-training models. The challenge of the problem is that there is no previous research on the use of generative self-supervised pre-training in the visual domain, so the authors and theirs have done a lot of exploration.

2 Method

The core approach of this paper is to pre-train a visual model using Masked Image Modeling. Given an input image x, BEiT encodes it as a vector representation of the context. The main goal of Masked Image Modeling is to recover the masked image patches from the encoded vector.

2.1 Image Representations

2.1.1 Image Patch

An image will be split into a sequence of patches, so a standard Transformer can take this image input directly. The image patches will be expanded into vectors and projected linearly, similar to the way words are processed in BERT. The image patch retains the original pixels and will be used as a feature of the input in BEiT.

2.1.2 Visual Token

There are two modules in the learning process of Visual token, tokenizer and encoder. The encoder learns to reconstruct the input image x based on the visual tokenizer. Because the potential visual token is discrete, the training of the model is non-differentiable.

2.2 Pre-Training BEIT ： Masked Image Modeling

The authors randomly masked about 40% of the image patches so that the corrupted image patches would be fed to the Transformer. Pre-training aims to maximize the log-likelihood of a correct visual marker z_i given a corrupted image:

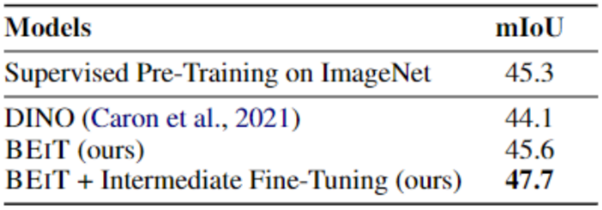
$$max \sum_{x \in D} E_M[\sum_{i \in M} logp_{MIN}(z_i|x^M)]$$

where D is the training corpus, M represents randomly masked positions, and x^M is the corrupted image that is masked according to M.

3 Experiment

3.1 Semantic Segmentation

The purpose of semantic segmentation is to predict a corresponding category for each pixel of the input image. The authors evaluated BEiT on the ADE20K benchmark, which has 25,000 images and 150 semantic categories. As shown in Table 3, the authors compared BEiT with supervised pre-training that relies on ImageNet labeled data, and their proposed method achieved better performance than supervised pre-training.

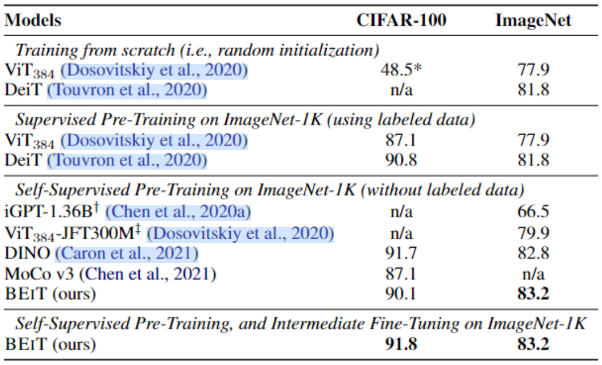


Models	mIoU
Supervised Pre-Training on ImageNet	45.3
DINO (Caron et al., 2021)	44.1
BEIT (ours)	45.6
BEIT + Intermediate Fine-Tuning (ours)	**47.7**

Figure 1: Results of semantic segmentation on ADE20K.

3.2 Image Classification

After unsupervised pre-training, BEiT-L was able to achieve the same performance on ImageNet-1k as Google's JFT-3B data (3B labeled images, 224 times more than ImageNet-22k) with supervised pre-training using only ImageNet-22k (14M labeled images) of labeled data. Furthermore, BEiT-L was able to consistently improve performance on ImageNet-1k to 89.5% after fine-tuning on a dataset of 70M size. This shows that the self-supervised pre-training of the BEiT drives the SOTA of the large model visual Transformer while significantly reducing the amount of supervised labeled data required.



Models	CIFAR-100	ImageNet
Training from scratch (i.e., random initialization)		
ViT$_{384}$ (Dosovitskiy et al., 2020)	48.5*	77.9
DeiT (Touvron et al., 2020)	n/a	81.8
Supervised Pre-Training on ImageNet-1K (using labeled data)		
ViT$_{384}$ (Dosovitskiy et al., 2020)	87.1	77.9
DeiT (Touvron et al., 2020)	90.8	81.8
Self-Supervised Pre-Training on ImageNet-1K (without labeled data)		
iGPT-1.36B† (Chen et al., 2020a)	n/a	66.5
ViT$_{384}$-JFT300M‡ (Dosovitskiy et al., 2020)	n/a	79.9
DINO (Caron et al., 2021)	91.7	82.8
MoCo v3 (Chen et al., 2021)	87.1	n/a
BEIT (ours)	90.1	**83.2**
Self-Supervised Pre-Training, and Intermediate Fine-Tuning on ImageNet-1K		
BEIT (ours)	**91.8**	**83.2**

Figure 2: Results of image classification on ImageNet-1k.

4 Conclusion

For the first time, BEiT demonstrates that the generative pre-training model can achieve better fine-tuning results than contrastive learning. It achieves superior results in image classification as well as semantic segmentation. However, BEiT has not been tested enough on downstream tasks, or perhaps the results were not good enough to include in the article. In the future, it may be possible to continue to scale up the BEiT and extend it to multimodal learning in text pre-training.

References

[1] Ding, Xiaohan and Zhang, Xiangyu and Zhou, Yizhuang and Han, Jungong and Ding, Guiguang and Sun, Jian. *Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs*, CVPR, 2022.