Note: BEIT: BERT Pre-Training of Image Transformers

sgc

June 5, 2022

1 Introduction

BERT can not directly be applied to image tasks since there is no pre-exist vocabulary like in NLP.
This paper introduced a self-supervised vision representation model BEIT(stands for Bidirectional Encoder representation from Image Transformers). The main procedure is that first tokenize the original image into visual tokens. Then randomly mask some image patches and fed them into the backbone Transformer. Authors proposed to split the image into a grid of patches that are the input representation of backbone Transformer. And tokenize the image to discrete visual tokens, which is obtained by the latent codes of discrete VAE. The objective is predict the visual tokens of the masked patches.

2 Method

The image has two views in this method: image patch and visual tokens. The number of patches and tokens are the same. In this work authors directly use the image tokenizer described in [1].

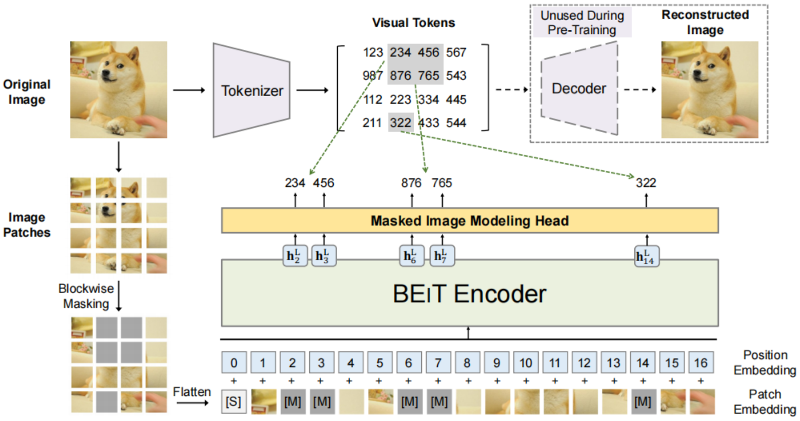

Figure 1: Overview of BEiT pre-training. Before pre-training, we learn an "image tokenizer" via autoencoding-style reconstruction, where an image is tokenized into discrete visual tokens according to the learned vocabulary. During pre-training, each image has two views, i.e., image patches, and visual tokens. We randomly mask some proportion of image patches (gray patches in the figure) and replace them with a special mask embedding [M]. Then the patches are fed to a backbone vision Transformer. The pre-training task aims at predicting the visual tokens of the *original* image based on the encoding vectors of the *corrupted* image.

2.1 Backbone Network: Image Transformer

Like conventional transformer, special token [S] and learnable position embeddings are applied.

2.2 Pre-Training BEIT: Masked Image Modeling

First randomly mask some percentage of image patches, and then predict the visual tokens that are corresponding to the masked patches. This model replace the masked patch with a learnable embedding $e_{[M]}$. The objective is to maximize the log-likelihood of the correct visual tokes z_i given the corrupted image.

$$\max \sum_{x\in\mathcal{D}} \mathbb{E}_{\mathcal{M}} \left[\sum_{i\in\mathcal{M}} \log p_{\text{MIM}}(z_i|x^{\mathcal{M}}) \right]$$

blockwise masking is employed.

3 Results

For segmentation tasks, the proposed method achieves better performance than supervised pretraining. With intermediate fine tuning it works even better. Extra ablation study shows that the proposed masked image modeling task significantly outperforms naive pixel-level auto-encoding.

4 Summary

This paper proposed a self-supervised pre-training framework. The proposed method is critical to make BERT-like pre-training work well for image Transformers.

References

[1] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021. 2