Paper Review and Notes For
BEIT: BERT Pre-Training of Image Transformers

Jin Hyoung Joo
hyoungjoo.j@gmail.com

Abstract

This paper [1] introduces BEIT *(Bidirectional Encoder representation from Image Transformers), which is a self-supervised vision model that is trained by the proposed masked image modeling task.*

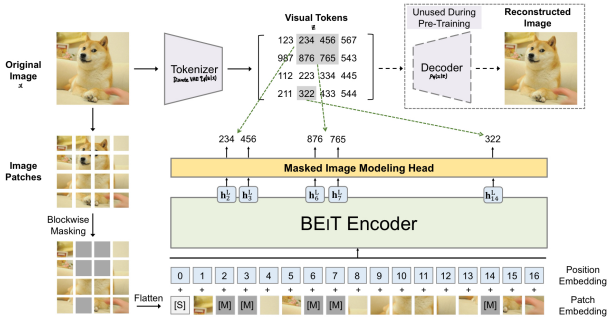

Figure 1. Model architecture of BEIT.

1. Key Points

1.1. Background of BEIT

Even though Vision Transformers are performant, they require large amounts of training data. To solve this problem, self-supervised pre-training is neccessary.

BERT achieved great success in NLP using the masked language modeling task, which is to predict the masked tokens of a given text based on the Transformer's encoding results. Naively applying this method to Vision Transformers have the following problems.

- There is no pre-existing vocabulary for image patches.
- Pixel-level recovery tasks waste modeling capabilities on pre-training short-range dependencies and high-frequency details.

1.2. Proposed Method

The pre-training of BEIT operates in the following steps.

1. The image is split into a grid of patches.
2. Using a pre-trained discrete VAE, the image is tokenized into discrete visual tokens.
3. Some proportion of the image patches are masked, and the corrupted input is used as input to the Transformer.
4. The model learns to recover the visual tokens of the original image (output of Step 2).

1.3. Advantages of BEIT

BEIT has improved convergence speed, high stability, and lower training costs on end tasks. A self-supervised BEIT can also learn reasonable semantic regions via pre-training even with no human supervision.

2. Technical Details

2.1. Blockwise Masking

2.2. Training Objective

3. Further Research

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1