BEIT: BERT Pre-Training of Image Transformers

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/38b9d095-39d5-4b6a-8f75-2c520e758403/BEIT_BERT_pretraining_of_image_transformers.pdf

Abstract

Bidirectional Encoder representation from Image Tansformers (BEIT)

: masked image modeling task to pretrain vision Transformers

1. tokenize : original image→visual token
2. randomly mask some image patch and fed them into backbone Transformer

   (pre-training_recover the original visual tokens based on corrupted image patches)

   →(fine-tune)

introduction

[contribution]

- proposed masked image modeling task to pretrain vision Transformers, self-supervised
  - VAE

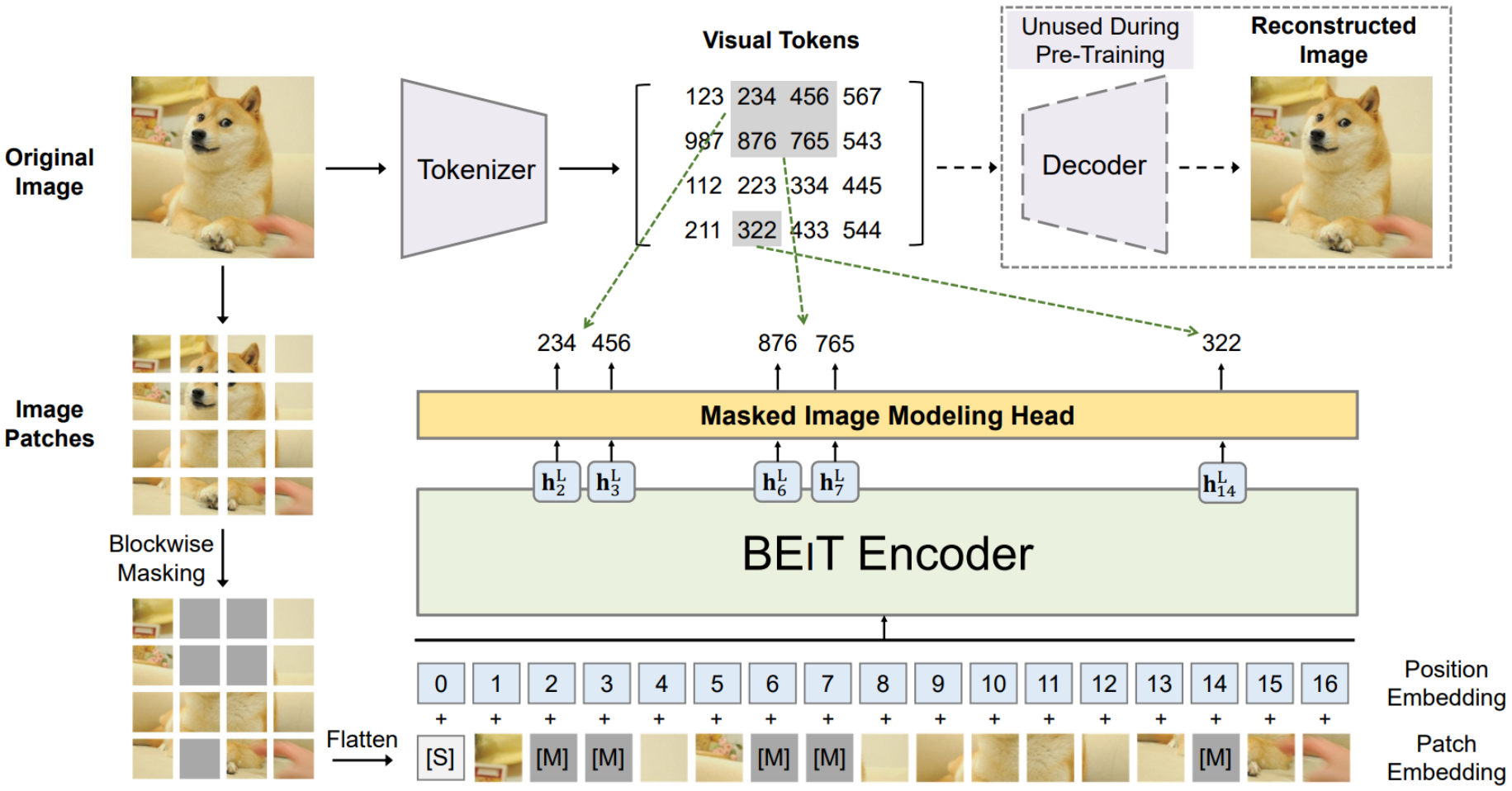

Figure 1: Overview of BEIT pre-training. Before pre-training, we learn an "image tokenizer" via autoencoding-style reconstruction, where an image is tokenized into discrete visual tokens according to the learned vocabulary. During pre-training, each image has two views, i.e., image patches, and visual 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.

Methods

input image : x

pretrain task : masked image modeling (MIM)

→ goal) recover masked image patches based on encoding vectors

image representations

image patch

image : $x \in \mathbb{R}^{H \times W \times C}$

patch : $x^p \in \mathbb{R}^{N \times (P^2 C)}$ _ $N = HW/P^2$개

C : #channels (H, W) : input image resolution (P, P) : resolution of each patch

$\{x_i^p\}_{i=1}^N$→ flatten into vectors → linearly projected (BERT' word embedding)

(experiment) 224×224 → 14×14개 (16×16)

visual token

-image tokenizer > raw pixe|

$x \in \mathbb{R}^{H \times W \times C}$ → $z = [z_1, \cdots, z_N] \in \mathcal{V}^{h \times w}$ $(\mathcal{V} = \{1, \cdots, |\mathcal{V}|\}$: contrain discrete token)

*learn image tokenizer via discrete variational autoencoder (dVAE)

tokenizer + decoder

backbone network: image transformer

:standard Transformer

pre-training BEIT : masked image modeling (MIM)

(image representation) tokenize

(backbone network:image transformer) L-layer Transformer

final hidden = encoded representations of the input patches

goal) maximize the log-likelihood of the correc visual token fiven the corrupted image

from the perspective of variational autoencoder