Note: Masked Autoencoders Are Scalable Vision Learners

Sinkoo

April 17, 2022

1 Introduction

This paper proposed masked autoencoders(MAE): mask random patches of the input image and reconstruct the missing pixels using a encoder-and-decoder architecture. Encoder work only on unmasked data and decoder tries to reconstruct the data from the masked tokens and latent representations.

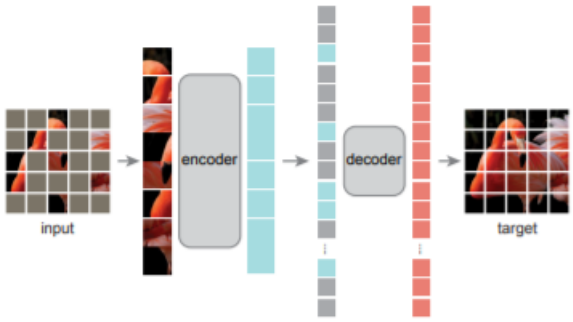

Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

2 Method

Masking Divide image into patches and randomly mask some of them with uniform distribution sampling(with high ratio mask 75% in this paper).

MAE encoder A ViT but only apply on unmasked patches.

MAE decoder The input to the MAE decoder is the full set of tokens consisting of (i) encoded visible patches, and (ii) mask tokens. The decoder has another series of Transformer blocks.
The MAE decoder is only used during pre-training to perform the image reconstruction task (only the encoder is used to produce image representations for recognition). So using a slight decoder can decrease the pre-training time.

Reconstruction target MAE predict the pixel values for masked patched to reconstruct the image. Calculate the MSE loss of the original image and the reconstructed image on pixel wise and only on masked patches.

Simple implementation

3 Summary

This is a really classic paper on applying autoencoder or transformers to CV. And it proposed a strong pretraining model to learning useful information for downstreaming works by force encoder to deal with mostly masked data.