Note: ConvMAE: Masked Convolution Meets Masked Autoencoders

sgc

June 5, 2022

1 Introduction

This paper introduced ConvMAE using hybrid convolution-transformer architectures and masked convolution into the masked auto-encoders. To prevent information leakage, the convolution blocks at early stages are equipped with masked convolutions.

2 Method

3 ConvMAE

This method is a simple and effective derivative of the popular MAE with minimal but effective modifications on the encoder design and the masking strategy. Conventional MAE makes transformer layers keeping all tokens during the pretraining resulting low efficiency. So authors introduce a hierarchical masking strategy coupled with masked convolution for the convolution stages to ensure only a small number of visible tokens are input into the transformer layers.

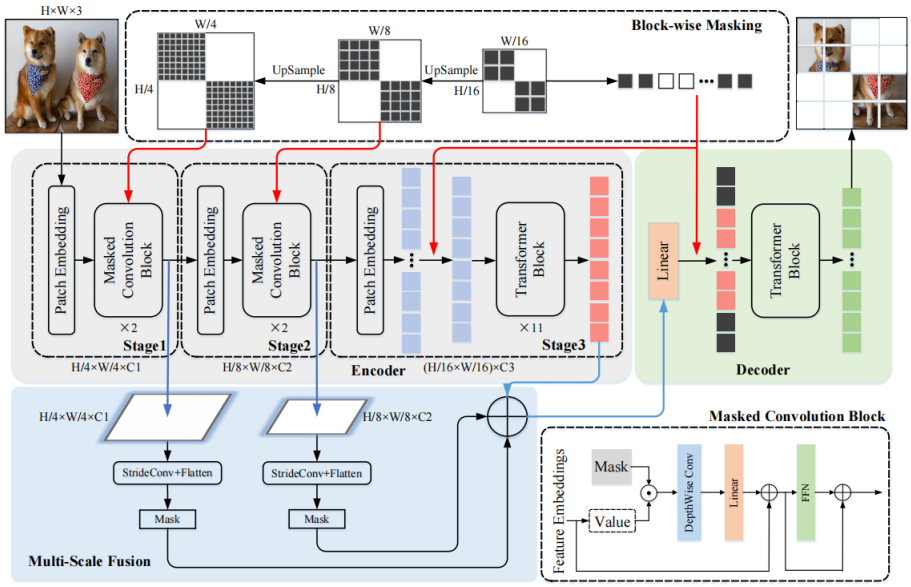

Figure 1: The pipeline of our proposed ConvMAE which consists of a hybrid convolution-transformer encoder, block-wise masking strategy with masked convolution and multi-scale decoder.

The Hybrid Convolution-transformer Encoder The encoder consists of 4 stages. The convolution blocks follow the design principle of the transformer block by only replacing the self-attention operation with the 5×5 depth-wise convolution, The third transformer stage uses commonly used self-attention blocks to obtain token embeddings

Block-wise Masking with Masked Convolutions The masked convolution in the first 2 stages to avoid the information leakage and ensure the quality of pre-training. And the masking strategy is upsample the masks in the stage-3.

The Multi-scale Decoder and Loss To get the coarse- and fine- information, the decoder combines the feature of stage 1, 2, 3.

$$E_d = \text{Linear}(\text{StrideConv}(E_1, 4) + \text{StrideConv}(E_2, 2) + E_3),$$

ConvMAE for Object Detection and Semantic Segmentation The main idea is replace some replace all but 1st, 4th, 7th, 11th global self-attention layers in stage-3 to shifted-window local self-attention layers. And shared global and local relative position bias are introduced to mitigate the high computational loss in stage-3.

4 Results

ConvMAE can achieve higher performance(mIoU) than the benchmark with less parameters. Setting the SOTA.

5 Summary

This work improved the MAE through building a pyramid structure and implement convolution + transformer structure to the encoder. Experiments show that this method sets new SOTA.

References