Efficient Self-supervised Vision Pretraining with Local Masked Reconstruction [2]

Minji Kim (minji@snu.ac.kr; 2020-28702), Dept. of Electrical and Computer Engineering, Seoul National University

1. Background and Motivation

Generative self-supervised learning approaches such as MAE [3] and BEiT [1], which reconstruct the original image from a small fraction of image patches, have shown promising performance in the domain of image classification. However, their global masked reconstruction strategy is computationally heavy and time-consuming. For example, pretraining an MAE-Huge on ImageNet with 224×224 resolution takes 34.5 hours on 128 TPU-v3 GPUs [3]. This problem becomes much more severe with high-resolution images which are essential in many vision tasks such as object detection. Thus, improving the *efficiency* in self-supervised vision pretraining is demanding. Inspired by the observation that the Transformer model mostly attends to patches close to the target patch, this paper proposes to restrict the range of attention used in the reconstruction so that the redundant computation in global self-attention could be resolved.

2. Local Masked Reconstruction (LoMaR)

MAE. MAE (shown on the left side of Fig. 1) is based on an asymmetric encoder-decoder architecture. First, a large portion (*e.g.*, 75%) of image patches are masked out while the remaining patches are projected into the latent space by the encoder. Then, the representations are fed into the decoder together with the placeholders for the masked patches, whereas those patches should be reconstructed as same with the original input image before masking. MSE loss is employed for the reconstruction loss.

Local masked reconstruction. Unlike MAE which globally reconstruct the patches, LoMaR randomly samples n windows where each contains $m \times m$ patches and reconstructs each window. This will reduce the complexity from $O(h^2w^2)$ to $O(hw+nm^4)$ given $h \times w$ of total patches. Instead of the asymmetric encoder-decoder of MAE, LoMaR is composed of a simple Transformer encoder followed by a simple MLP head. First, all patches in a sampled local window including the masked ones are fed into the encoder. The latent representations from the encoder output are converted to the original feature dimension with the MLP head. The MSE loss are computed between the final outputs and the normalized ground-truth image.

Finetuning. LoMaR is pretrained on ImageNet-1K in a self-supervised manner and then finetuned on ImageNet-1K with supervision from the labels. During finetuning, all the image patches are fed into the model and the average of their features are used as the final representation for classification.

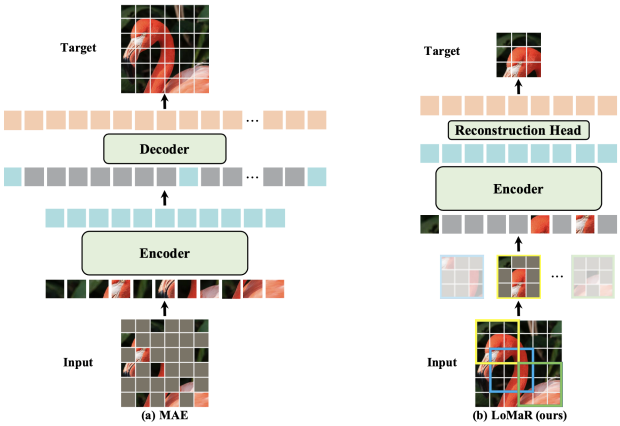

Figure 1. The illustration of LoMaR.

Experimental results. LoMaR reaches the same accuracy with MAE within 66 hours of pretraining, which is $3.5\times$ faster. Application of LoMaR in object detection and instance segmentation has shown an $0.3 \sim 0.4$%p improvement of AP over MAE.

3. Personal Note

The idea of limiting the effective region of self-attention is simple and effective, although not novel nor surprising. Also, I think the computation efficiency should also be compared other than the training time, *e.g.*, the number of parameters and throughput.

References

[1] H. Bao, et al. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[2] J. Chen, et al. Efficient self-supervised vision pretraining with local masked reconstruction. *arXiv preprint arXiv:2206.00790*, 2022. 1

[3] K. He, et al. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1