**Meeting Note**

**Author**: Y.K. Lin
**Date**: 2024/09/30 – 2024/10/04

**Weekly Summary**
Masked Autoencoders Are Scalable Vision Learners [1].

**Plan for Next Week**
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv: 1810.04805)

**Details**
In the realm of natural language processing (NLP), the leak of data has been successfully addressed by self-supervised learning (e.g. using non-masked data to predict the masked part). For instance, Generative Pre-trained Transformer (GPT) [2] takes a sequence as input (i.e. non-masked data) and predict the next token (i.e. masked data) as output; the "Cloze task [3]" in BERT [4], the authors randomly mask some tokens of the input sequence and train the language model to predict the original sequence as the objective.

On the other hands, many of the dataset, especially large image datasets, are publicly unavailable in the field of computer vision. In order to address this issue and draw on the solutions and successful experiences from the field of NLP, He et al. [1] proposed the masked autoencoder (MAE). Firstly, they analysis the difference of masked autoencoder between language and vision, and list three major points: (1) The architecture is different. The vision and language models typically use convolution neural networks (CNNs) [6] and Transformer [7] as the backbone network, respectively. In 2020, Google proposed the Vision Transform (ViT) [5], which filled this gap by using Transformer model on vision tasks; (2) The density of information is different. Language is generated by human and contains a highly density of information, such that we can use a few words to describe a complex image. However, vision contains much less information, and has heavy redundancy [Re. 1]. To address this issue, the authors masked a high portion of pixels randomly, to minimize the information redundancy in the images; (3) The decoder plays different role between reconstructing text and images. As we mentioned at point (2), the decoder for text reconstructing tasks which predicts the missing tokens with rich sematic information can be very simple, such as a multiplayer perceptron (MLP), and the decoder for image reconstruction tasks should be well-designed [Re. 2].

With the analysis, the authors designed a simple and asymmetric masked autoencoder, MAE,

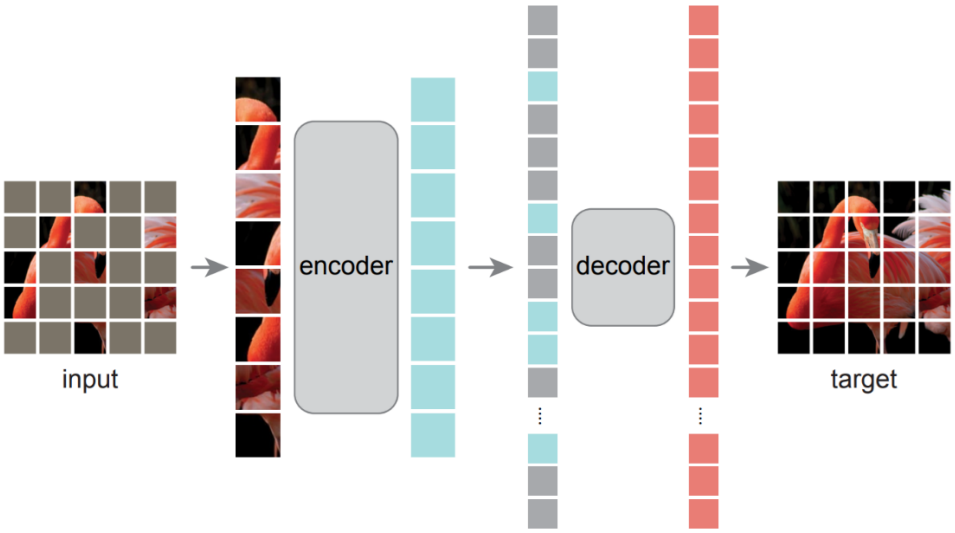


Figure 1. The overall architecture of MAE.

which is effective for vision representation learning. As the Figure 1 shows, the encoder [Re. 3] only processes on non-masked patches as inputs, this allows MAE to require less computational power while utilizing a very large encoder (because a high portion (e.g. 75%) of patches are masked); and the decoder processes on all masked and non-masked patches since the decoder would have no position information if only the non-masked patches were used. Despite the decoder should be well-designed, it can be very small, flexible, and independent of the encoder, unlike traditional autoencoder, because it is used only during the training phase.

For the pre-training tasks, the authors use image reconstruction as the objective. They first mask the image randomly, and then use the mean square error (MSE) as the loss function to train the model [Re. 4], where the loss is computed only on the masked patches.

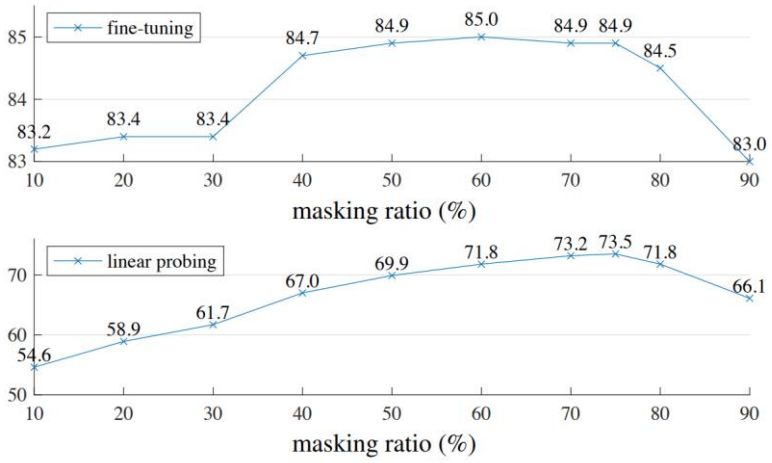


Figure 2. The experimental results of different masking ratio for (top) fine-tuning and (bottom) linear probing. The model has better performance with higher masking ratio (75%).

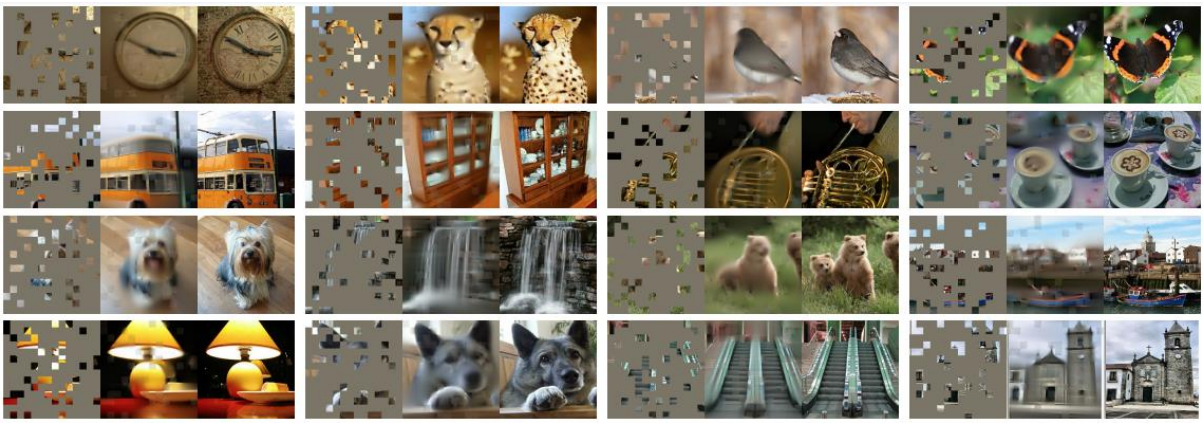

Figure 3. Example results from ImageNet dataset [8].


Figure 4. Different masking strategies. (left) Random mask. (middle) Large block mask. (right) Grid mask which remove blocks at equal intervals [Re. 5].

**Remarks**

[Re. 1]

The authors give an example, that is, to predict a missing word in a sentence requires complete understanding of language, but is much easier to predict a few missing pixels on an image because the model only need to reconstruct the missing parts from the neighborhood pixels.

[Re. 2]

The authors experiment on the depth and width of decoder. The results shows that sufficiently depth is important for decoder on reconstruction tasks, also, because it shows that decoder with 512 width performs well, the authors use 512 dimensions as default width.

Table 1. The experimental results on (a) the depth of decoder and (b) the width of decoder.

| blocks | ft | lin |
|---|---|---|
| 1 | 84.8 | 65.5 |
| 2 | **84.9** | 70.0 |
| 4 | **84.9** | 71.9 |
| 8 | **84.9** | **73.5** |
| 12 | 84.4 | 73.3 |

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

| dim | ft | lin |
|---|---|---|
| 128 | **84.9** | 69.1 |
| 256 | 84.8 | 71.3 |
| 512 | **84.9** | **73.5** |
| 768 | 84.4 | 73.1 |
| 1024 | 84.3 | 73.1 |

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

[Re. 3]
The authors mentioned that “the encoder is a ViT”.

[Re. 4]
One of the key points they emphasized is “simplicity”. They use a straightforward random masking strategy and apply MSE as the loss function.

[Re. 5]
Maybe we can use the weighted loss map to force the model to better learn the boundary between masked and non-masked areas to address the issue that the images appear blocky, although it is not the purpose of this paper.

**References**

[1] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16000-16009).
[2] Radford, A. (2018). Improving language understanding by generative pre-training.
[3] Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 30, 415-433.
[4] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
[5] Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
[6] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, *36*(4), 193-202.
[7] Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
[8] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee, 2009.

**Comments**

N/A