Seminar Reivew: Efficient Learning in Single- and Multi-Modal Vision Transformers

Mingji Han

Introduction

In this project, we review a paper "**SupMAE: Supervised Masked Autoencoders Are Efficient Vision Learners**" [2]. This paper comes from BU ECE Distinguished Lecture "Efficient Learning in Single- and Multi-Modal Vision Transformers" on Sept. 26th. by Prof. Diana Marculescu. (https://www.bu.edu/eng/news-events/events-calendar/?eid=285332)

Topic

Transformer [8] has been the fundamental building block of large vision models (ViT [3] and CLIP [7]). It demonstrates outstanding performance on various computer vision tasks. Masked Autoencoder (MAE) [1] is a self-supervised learning method for transformer-based vision models, which achieve high accuracies on ImageNet datasets and other downstream tasks. The key idea of MAE is that masking the original images and use encoder-decoder models to reconstruct the original images. However, training transformer-based models needs large scale of data and multiple training epochs, taking significant computational resources. The computational costs make model deployment costly. To address this challenge, the author purpose Supervised Masked Autoencoders (SupMAE), reducing the computational costs while maintaining learning effectiveness.

Main Contributions

The paper makes the following major contributions:

1. SupMAE is the first research work that proves that using supervised learning can improve the learning process of self-supervised learning masked autoencoder.
2. SupMAE reduces the computational costs significantly. It only takes MAE's 30% computational cost but achieve the similar performance on ImageNet-1K and other learning tasks.
3. SupMAE makes MAE learned the robust and transferable features and concepts, making the models perform well on downstreak learning tasks like transfer learning.

To be more specific, the SupMAE extends the MAE with a supervised learning branch (model) with classification losses. The learning objective is not simply reconstruction loss but a sum of reconstruction loss and classification loss. With the branch of supervised learning and classification losses, the SupMAE model can learn more robust

and transferable features from classification labels. The classification branch also enables to learn the features and concepts faster during the training, making SupMAE use 75% less epochs than MAE to complete learning. Therefore reducing the computational costs significantly.

Results

1. On ImageNet-1K datasets, compared with other supervised or self-supervised ViT models including MoCov3 [4], BEiT [6] , ViT [3], DeiT [5], and MAE [2], SupMAE achieve comparable training accuracies.
2. Compared with supervised and self-supervised ViT models, it takes only about 30% trainincg costs and 25% training epochs to achieve the comparable training accurancies.
3. On robustness evaluation, SupMAE achieves better performance on “In-Rendition” and “In-Sketch” than MAE and DeiT. The overall robustness of SupMAE is better than MAE and DeiT.
4. On downstream tasks like few-shot transfer learning and semantic segmentation, SupMAE outperforms MAE.

My View of Work

This research work use a simple but effective method, i.e. adding a supervised classification branch and model to improve the performance of MAE. The highlight points of this paper is that it reduce the training computational time and cost significantly while achieving comparable performance on various vision tasks. As the large pre-trained models has become standard of computer vision and natural language processing. Reducing the training costs has become an important issue in model pre-traning. This paper gives us an insight on how to reducing the training costs in an algorithm view.

What Should I Do as Next Steps

There are several directions can be explored based on this research work. The first is that using different tasks for supervised learning branch on SupMAE. In SupMAE, the supervised learning branch uses classification task. It is possible to use other learning tasks like object detection and visual recognition. Moreover, we can explore the supervised learning approach in multi-modal vision-language models. We can test

whether supervised learning method with a few high-quality data can make improvements on multi-model models.

Reference:

[1] Masked Autoencoders Are Scalable Vision Learners https://arxiv.org/abs/2111.06377

[2] SupMAE: Supervised Masked Autoencoders Are Efficient Vision Learners https://arxiv.org/abs/2205.14540

[3] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale https://arxiv.org/abs/2010.11929

[4] An Empirical Study of Training Self-Supervised Vision Transformers https://arxiv.org/abs/2104.02057

[5] Training data-efficient image transformers & distillation through attention https://arxiv.org/abs/2012.12877

[6] BEiT: BERT Pre-Training of Image Transformers https://arxiv.org/abs/2106.08254

[7] Learning Transferable Visual Models From Natural Language Supervision https://arxiv.org/abs/2103.00020

[8] Attention Is All You Need https://arxiv.org/abs/1706.03762