Paper review for "Dosovitskiy et al: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"

EECS 598 Paper Review – Week 8 - Changyuan Qiu

Self-attention-based architectures, in particular Transformers, have been dominant in natural language processing (NLP) recently, while its applications to computer vision remain limited, where attention is either applied in conjunction with CNN or used to replace certain components of CNN while keeping the overall structure in place. This paper explores the performance of the pure Transformer architecture in computer vision tasks and the authors find that Transformers applied directly to image patches and pre-trained on large datasets work well on image classification.

Here the authors propose Vision Transformer (ViT), which take linear embeddings of the image patches concatenated with position embeddings and learnable classification tokens as input (like tokens or words in an NLP model) and feed the input to a standard Transformer encoder which consists of alternating layers of multiheaded self-attention and MLP blocks with layernorm (LN) and residual connections as proposed in Vaswani et al. (2017).

Compared with prior works which mainly focus on combining CNN with self-attention, this paper does not rely on CNN and uses a pure Transformer architecture. Regarding experiments, they pre-train ViT model of different scale (ViT-Base, ViT-Large and ViT-Huge) on ILSVRC-2012, ImageNet-21k and JFT and evaluate the representation learning capabilities through fine-tuning to several benchmark tasks: ImageNet on the original validation labels and the cleaned-up ReaL labels, CIFAR-10/100, Ixfird-IIIT Pets, Oxford Flowers-102 and the 19-task VTAB classification suite. And they find that there largest model (ViT-Huge with 14x14 input patch size pre-trained on JFT) achieve SOTA on all benchmark tasks except Oxford Flowers-102, for which ViT-Large with 16x16 input patch size pre-trained on JFT achieve SOTA.

Moreover, they also found ViT to be much more efficient to train than previous SOTA models. For example, the ViT-H/14 model takes 2.5k TPUv3-core-days to train, while the previous SOTA on ImageNet, EfficientNet-L2 takes 12.3k TPUv3-core-days to train. And that the ViT-L/16 model achieves comparable performance as ResNet152x4 on all tasks while the former only takes 0.68k TPUv3-core-days to train and the later takes 9.9k TPUv3-core-days to train.

Regarding limitations of this paper, I discovered that in the introduction the authors contrast Transformer with CNN trained on mid-sized datasets such as ImageNet and find that the performance of ViT is slightly worse than CNN. And they stated that "Transformers lack some inductive biases inherent to CNNs, such as translation equivariance and locality, and therefore do not generalize well when trained on insufficient amounts of data." However, since there exist different inductive biases in standard convolution (like static size of kernels, locality and translation equivalence), it is not clear which inductive bias of CNN prevents better generalization and further analysis is needed.

Reference

Dosovitskiy, Alexey & Beyer, Lucas & Kolesnikov, Alexander & Weissenborn, Dirk & Zhai, Xiaohua & Unterthiner, Thomas & Dehghani, Mostafa & Minderer, Matthias & Heigold, Georg & Gelly, Sylvain & Uszkoreit, Jakob & Houlsby, Neil. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS,* 2017.