Paper review for " Chen et al.: Generative pretraining from pixels"

EECS 598 Paper Review - Week 3 - Changyuan Qiu

Inspired by the enormous progress of unsupervised representation learning models for natural language processing like GPT-2 and BERT, this paper aims at utilizing generative models for pre-training in image classification tasks. The core approach of the paper can be divided into 3 stages: the pre-processing stage, the pre-training stage, and the fine-tuning stage. In the pre-processing stage, raw images are resized to a low resolution and reshaped into a 1D sequence without incorporating knowledge of the 2D input structure; in the pre-training stage, one of the 2 objectives (Autoregressive next pixel prediction and BERT masked pixel prediction) are optimized using an architecture built from GPT-2; in the fine-tuning stage, the performance of the model are evaluated with linear probing or fine-tuning.

Compared with prior work, this paper uses a dense connectivity pattern which does not encode the 2D spatial structure of images, yet is still able to match and even outperform approaches which do. For linear probing, the model achieves 96.3% on CIFAR-10, 82.8% on CIFAR- and 95.5% on STL-10, outperforming SOTA unsupervised pre-training methods like SimCLR and ResNet-152, it also achieves comparable performance as SOTA on ImageNet (72.0% vs. 76.5%); for fine-tuning (pre-trained on ImageNet), it achieves 99.0% (SOTA) on CIFAR-10, 88.5% on CIFAR-100 (worse than SOTA – 91.7%) and 66.3% on ImageNet (worse than SOTA, 70.2%).

On point worth mentioning is that while for linear probing the performance gap between objective of BERT and AR is approximately 1% on CIFAR-10 and 6% on ImageNet, after fine-tuning the prior gap diminishes to 0.4% on CIFAR-10 and on ImageNet BERT even outperforms AR slightly. I believe this might come from the fact that the model is pre-trained on ImageNet during fine-tuning. Besides, the approach performs worse on high resolution image (as shown in experiments on ImageNet) probably due to down-sampling. Finally, the approach in this paper ignores the spatial information of image, which do lead to loss of information.