Review-4

Data-Efficient Image Recognition With Contrastive Predictive Coding

Growing advancements in unsupervised learning have given rise to data-efficient algorithms which commonly fall in the regime of self-supervised learning. These methods provide improved performance in comparison to supervised learning with their representations readily transferrable to downstream tasks. The work presents an improved version of the Contrastive Predictive Coding method (CPCv2) which makes use of latent representations for image recognition in a data-efficient manner. In comparison to CPCv1, CPCv2 demonstrates improved performance as a result of 5 improvements carried out to the training setting. These improvements aid in data-efficient learning when trained on 1% of ImageNet dataset. Furthermore, representations learnt by the model are finetuned to for efficient object-detection on the PASCAL-VOC 2007 dataset.

CPCv2 is a significant improvement in comparison to the previous counterpart. While the model makes use of pre-existing InfoNCE Loss, the training setup introduces 7 improvements which lead to data-efficient learning of images. (1) CPCv2 consists of increased model capacity with 46 residual blocks resulting in ResNet-161. (2) Batch normalization is replaced by Layer normalization as it does not harm downstream performance by eliminzating dependency between image patches. (3) Instead of carrying out bottom-up predictions of patches, CPCv2 increases the prediction directions to bottom, up left and right. (4) CPCv2 consists of patch-based augmentations such as color-dropping and color-transforms in order augment low-level variability. (5) Lastly, CPCv2 makes use of larger patch sizes in conjunction with other randomized patch-based augmentations.

CPCv2 leads to better learning of representations and data-efficient transfer to downstream tasks such as object detection. The primary component of CPCv2 which allows it to scale on the ImageNet dataset is the model size. Larger model sizes have demonstrated performance gains compared to previous methods. CPCv2 increases the size of ResNet architecture to accomplish this objective. Additionally, the method makes use of patch-level augmentations which allow the model to learn fine-grained variability in images. While CPCv2 improves classification accuracies with reduced data requirements, it presents 2 major drawbacks. Firstly, While other methods such as MoCo make use of lesser augmentations to yield performance improvements, CPCv2 does not fully exploit the data and depends on increased number of augmentation techniques to learn the semantics of images. Secondly, improved performance and data-efficiency of CPCv2 are a result of increased model size and patch-based operations which do not account for performance gains in CPC loss.

While the novel contributions of CPCv2 directly depend on improvements in training setup and model architecture, the method opens two new directions for future work. Firstly, CPCv2 paves the way for data-efficient image recognition by making use of effective augmentations and their variations for learning fine-grained semantics. Secondly, the method demonstrates transfer of fine-tuned representations on PASCAL-VOC 2007 dataset. The work can be further extended to other complex datasets such as NORB dataset which consists of images in different frames of reference.

Image recognition in supervised learning is greatly hindered by data-efficiency. The work aims to address the open problem of data-efficiency by making use of the self-supervised CPCv2. In comparison to CPCv1, the method demonstrates data-efficiency and performance gains on the ImageNet dataset by using as less as 1% samples. This is a direct consequence of increased model capacity, patch-level augmentations and prediction directions which improve feature extraction and predictive power of the model. Additionally, the learnt representations can be finetuned and transferred to downstream object detection tasks with state-of-the-art-performance.