Review-2

Momentum Contrast for Unsupervised Visual Representation Learning

Unsupervised learning has seen a tremendous growth in the development of visual tasks. Various unsupervised methods in deep learning consist of self-supervision, or contrastive learning, wherein the loss function is defined on a pretext task consisting of true labels as intrinsically generated entities. This aids in efficient recognition of images from large datasets and successfully transfer the learned features to downstream tasks such as object detection. The work presents a contrastive learning algorithm called Momentum Contrast (MoCo). MoCo models contrastive learning as a dictionary-lookup task with the query being the input to the model and the keys being augmentations of the input image. The encoded query is contrasted against the momentum-encoded keys from a queue which aid in learning rich representations that are transferrable to downstream tasks. MoCo is shown to outperform various supervised learning algorithms on the ImageNet dataset and 7 detection tasks from other large-scale bechmarks.

The MoCo algorithm comprises of a query encoder and a key encoder. The query encoder encodes the visual query and compares it against the encoded keys using the similarity dot product. Keys are generated from augmentations of the input image and encoded using a momentum encoder which is a slow moving average of the query encoder. Utilizing momentum-based updates leads to consistency in keys of the dictionary which has proven to be a hindrance in the memory-bank approach and previous works on contrastive learning. Moreover, MoCo makes use of a queue for comparing dynamically generated keys with the query. The queue leads to the construction of a larger dictionary in comparison to the simpler end-to-end method which can only accomodate smaller batch-sizes due to memory constraints. Additionally, MoCo makes use of shuffling batch-normalization which shuffles the sample order of the mini-batch before distributional execution. Shuffling batch-normalization prevents the leakage of information in comparison to its standard counterpart in the ResNet architecture.

MoCo generalizes on a number of benchmarks including ImageNet-1M (IM-1M) and Instagram-1B (IG-1B). Representations learned as a result of the momentum updates are readily transferrable to downstream object detection tasks on the PASCAL-VOC, COCO and various other datasets. Moreover, the ablation study presented on the various constituents of MoCo highlight the effectiveness of the proposed approach. However, the MoCo algorithm makes use of only a single pretext task (instance discrimination as dictionary lookup) and does not yield any insights into what kind of other pretext tasks may be used to improve the performance of the model. Additionally, MoCo is able to demonstrate only incremental gains from IM-1M to IG-1B given that the addition of data is significant. These gains highlight the need for proper exploitation of large-scale datasets.

MoCo presents the novel constituents of a momentum encoder combined with a queue for keys. To that end, MoCo throws light on two new areas of work. While the usage of self-supervised methods is increasing rapidly, what other pretext tasks could be adopted for the unsupervised setting? Addition of new pretext tasks from an energy-based perspective would aid in better generalization of models to large-scale benchmarks. Secondly, Moco demonstrates suitability of the momentum scheme on visual inputs. While the application of a slow-moving encoder is yet to be presented in other high dimensional settings such as natural language and video inputs, MoCo paves the way for necessary changes and directions for improvements of self-supervised encoding methods.

The growing applicaiton of unsupervised learning on visual tasks has motivated self-supervised methods to effectively outperform conventional supervised schemes. However, self-supervision is restricted by memory as a result of large batch-sizes and consistency of the pretext task. MoCo aims to address these issues by introducing a dictionary-lookup as a pretext task consisting of queries as the input image and keys as augmentations of the input image. While queries are encoded using a query encoder, key encodings are produced using a slow-moving average of the query encoder which is proven essential for maintaining consistency among dynamic keys. Keys are compared to the query upon utilizing a dynamic queue which expands the dictionary size and leads to higher batch-sizes. Utilizing the momentum encoder in combination with the dynamic queue, MoCo outperforms various supervised learning algorithms on IM-1M and IG-1B benchmarks and suitably transfers the learned representations to downstream object detection tasks.