FID3018 Review of Papers: Self Review

Self supervised learning in computer vision

Abubakrelsedik Karali

Karali@kth.se

Introduction

Due to the powerful ability to learn different levels of general visual features, deep neural networks have been used as the basic structure to many computer vision applications. Large-scale labelled data are generally required to train deep neural networks to obtain better performance in visual feature learning from images or videos for computer vision applications. The models trained from large-scale image datasets are widely used as the pre-trained models and fine-tuned for other tasks for two main reasons: (1) the parameters learned from large-scale diverse datasets provide a good starting point, therefore, networks training on other tasks can converge faster, (2) the network trained on large-scale datasets already learned the hierarchy features which can help to reduce over-fitting problem during the training of other tasks, especially when datasets of other tasks are small or training labels are scarce. Therefore, the performance of deep convolutional neural networks greatly depends on their capability and the amount of training data

However, collection and annotation of large-scale datasets are time-consuming and expensive. To avoid extensive cost of collecting and annotating large-scale datasets, self-supervised learning methods are proposed to learn general image and video features from large-scale unlabelled data without using any human-annotated labels.

The general, a pipeline of self-supervised learning consists of pretext task and downstream task. During the self-supervised training phase, a predefined pretext task is designed for ConvNets to solve, and the pseudo labels for the pretext task are automatically generated based on some attributes of data. Then the ConvNet is trained to learn object functions of the pretext task. After the self-supervised training finished, the learned visual features can be further transferred to downstream tasks (especially when only relatively small data available) as pretrained models to improve performance and overcome overfitting. Generally, shallow layers capture general low-level features like edges, corners, and textures while deeper layers capture task related high-level features. Therefore, visual features from only the first several layers are transferred during the supervised downstream task training phase.

Since no human annotations are needed to generate pseudo labels during self-supervised training, very largescale datasets can be used for self-supervised training. Trained with these pseudo labels, self-supervised methods achieved promising results and the gap with supervised methods in performance on downstream tasks becomes smaller.

This review will focus on three papers published in very reputable conferences ICCV, ICLR and CVPR: the first one is the first attempt towards self-supervision in computer vision. The second focuses on a very simple yet very successful approach for self-supervision. Lastly an empirical evaluation of different approaches showing different behavior and patterns.

1. Unsupervised Visual Representation Learning by Context Prediction. Carl Doersch, Abhinav Gupta, Alexei A. Efros

The method proposed by Doersch et al. is one of the pioneers works of using spatial context cues for self-supervised visual feature learning. This work explores the use of spatial context as a source of free and plentiful supervisory signal for training a rich visual representation. Given only a large, unlabelled image collection, Random pairs of image patches are extracted from each image, then a ConvNet is

trained to recognize the relative positions of the two image patches. To solve this puzzle, ConvNets need to recognize objects in images and learn the relationships among different parts of objects. To avoid the network learns trivial solutions such as simply using edges in patches to accomplish the task, heavy data augmentation is applied during the training phase. They demonstrated that the feature representation learned using spatial context captures visual similarity across images. Moreover, the representation allows to perform unsupervised visual discovery of objects like cats, people, and even birds from the Pascal VOC dataset.

Notes,The main principle of designing puzzle tasks is to find a suitable task which is not too difficult and not too easy for a network to solve. If it is too difficult, the network may not converge due to the ambiguity of the task or can easily learn trivial solutions if it is too easy. Therefore, a reduction in the search space is usually employed to reduce the difficulty of the task.

2. UNSUPERVISED REPRESENTATION LEARNING BY PREDICTING IMAGE ROTATIONS. Spyros Gidaris, Praveer Singh, Nikos Komodakis

Gidaris et al proposed to learn image features by training ConvNets to recognize the 2D image rotation that is applied to the image that it gets as input. They demonstrated both qualitatively and quantitatively that this simple task can actually provide a very powerful supervisory task for semantic feature learning. They evaluated the approach in different unsupervised feature learning benchmarks and we exhibit in all of them state-of-the-art performance. The results on those benchmarks demonstrated huge improvements compared to prior state-of-the-art approaches in unsupervised representation learning which significantly closed the gap with supervised feature learning. For example, in PASCAL VOC 2007 detection task the unsupervised pre-trained AlexNet model achieves the state-of-the-art for unsupervised methods mAP of 54.4% that is only 2.4 points lower from the fully supervised case. The results are the same when they transfer the unsupervised learned features on various other tasks, such as ImageNet classification, PASCAL classification, PASCAL segmentation, and CIFAR-10 classification.

Notes, Images contain rich spatial context information such as the relative positions among different patches from an image which can be used to design the pretext task for self supervised learning. The pretext task can be to predict the relative positions of two patches such as the first paper. The context of full images can also be used as a supervision signal to design pretext tasks such as to recognize the rotating angles of the whole images which is the case of this study. To accomplish these pretext tasks, ConvNets need to learn spatial context information such as the shape of the objects and the relative positions of different parts of an object. Also in this study ffeature maps are visualized to show the attention of networks. Larger activation represents the neural network pays more attention to the corresponding region in the image. Feature maps are usually qualitatively visualized and compared with that of supervised models.

3. Revisiting Self-Supervised Visual Representation Learning. Alexander Kolesnikov, Xiaohua Zhai, Lucas Beyer

In this study the compared different self-supervised approaches including the first the second papers discussed in this report. They have investigated self-supervised visual representation learning from empirical and unexplored angles. They experiments was systematically test different pretext tasks- different architectures- different dataset with some variations like widening and changing the features level. Doing so, they concluded multiple important insights, namely that lessons from architecture design in the fully supervised setting do not necessarily translate to the self supervised setting, In contrary to popular architectures like AlexNet, in residual architectures, the final feature layer

consistently results in the best performance, also the widening factor of CNNs has a drastic effect on performance of self-supervised techniques, stochastic gradient descent training of linear logistic regression takes long time to converge. In this study they also demonstrated that performance of existing self-supervision techniques can be consistently boosted and that this leads to halving the gap between self supervision and fully supervision. Most importantly they reveal that neither is the ranking of architectures consistent across different methods, nor is the ranking of methods consistent across architectures. This implies that pretext tasks for self-supervised learning should not be considered in isolation, but in conjunction with underlying architectures.

Notes. They revisited numerous previously proposed self-supervised models, conduct a thorough large-scale study and, as a result, uncover multiple crucial insights. The experiments was empirical and lacks any theory in comparing the approaches. Which made some results looks quite abstract and in some cases meaningless for example "*Most importantly they reveal that neither is the ranking of architectures consistent across different methods, nor is the ranking of methods consistent across architectures*". But also it was very nice that they challenge a number of common practices in self supervised visual representation learning and observe that standard recipes for CNN design do not always translate to self-supervised representation learning. As part of the study, they boosted the performance of previously proposed techniques and outperform previously published state-of-the-art results by a large margin, which is not very common contribution of a survey paper.

1. C. Doersch, A. Gupta and A. A. Efros, "Unsupervised Visual Representation Learning by Context Prediction," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1422-1430, doi: 10.1109/ICCV.2015.167.

2. Nikos Komodakis, Spyros Gidaris. "Unsupervised representation learning by predicting image rotations." *International Conference on Learning Representations (ICLR)*, Apr 2018, Vancouver, Canada.

3.

4. Alexander Kolesnikov, Xiaohua Zhai, Lucas Beyer; "Revisiting Self-Supervised Visual Representation Learning" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern* Recognition (CVPR), 2019, pp. 1920-1929