Boosting Self-Supervised Learning via Knowledge Transfer

Seungmin Lee (profile2697@gmail.com; 2013-11420), Dept. of Computer Science and Engineering, Seoul National University

1. Motivation

Self-supervised learning (SSL) aims to learn meaningful representations from pretext tasks for boosting the performance of target tasks. The pretext tasks do not require hand-made labels. Otherwise, the target tasks are typically assumed that the tasks do not have enough labels. Traditional SSL methods transfer the knowledge learned from the pretext task using fine-tuning. Despite using fine-tuning is a simple and effective way, the transfer methodology limits the design choices of network architectures since a pretext network and target network should have the same structure for the fine-tuning. In this setting, the architecture is limited to a simple structure because the target task does not have enough labels. For the same reason, the difficulty of the pretext task is also limited to an easy task because it should be solvable using the simple model. The proposed method tries to decouple the architectures of pretext tasks and target tasks using distillation.

2. Method

2.1. Knowledge Transfer using Distillation

The proposed method uses distillation instead of fine-tuning for knowledge transfer. The distillation removes the coupling of the target task network architecture and pretext task network architecture. Additionally, the distillation also helps not to transfer specific knowledge of the pretext task like the weights of the fully-connected layers. The proposed knowledge transfer procedure consists of four steps: **(a)** train a pretext model on a pretext task. **(b)** extract cluster centroids of features from the pre-trained pretext model using the K-means algorithm. In this step, the features are generated using the images of the pretext tasks. **(c)** assign pseudo-labels to the target task images by finding the closest centroid of each image. Note a different dataset(like target dataset) than the one used in step (a) can be utilized in this step.**(d)** pre-train a task model using a classification problem of the pseudo-labels.

2.2. More Difficult Pretext Task: Jigsaw++

By decoupling the architectures of the pretext task and the target task, the proposed method can use a deeper pretext task model and thus more difficult but representative pretext tasks. This paper proposed a Jigsaw++ pretext task. This task is similar to the ordinary jigsaw puzzle, except it can contain some random tiles from other images. In this paper, the pretext task consists of 9 tiles, and at most, 2 tiles are from other images. The proposed task is harder than other pretext tasks, but it gives more characteristic features.

3. Experiments and Analysis

There are many interesting experiments and analysis. In this section, I will briefly summarize the results.

Impact of Cluster Centroids The authors measured the performance by varying the number of centroids from 500 to 2000. They used the ImageNet dataset, so the optimal amount of the centroids is maybe about 1000. In these experiments, the performance increases as the number of centroids increases. However, the difference between the lowest and largest performance is marginal, which indicates the proposed method is pretty robust to the centroids number if the technique does not choose extremely a small amount.

Comparison with Other SSL Methods The authors compared the proposed method to other SSL methods. According to the results, the pretext task model size seems pretty essential. The difference between distillation and fine-tuning using the same pretext task and the same model is marginal. However, using vgg as the pretext model instead of AlexNet boosts the performance, which indicates the decoupling the network structures is crucial.

Classification with a Linear Classifier In this experiment, the authors compared the performances of linear classifiers trained on features from an arbitrary layer. Compared to other SSL methods, the proposed method seems to encode more characteristic features. The difference between the performances of the conv1 and conv4 is quite significant. This difference shows the choice of the intermediate layer used for transferring knowledge is crucial. Therefore, if we can find the optimal layer, it would be helpful to transfer adequate knowledge.

Visualization The visualization also shows interesting results (Figure. 1). The images that lie around the same cluster centroid share similar characteristics.

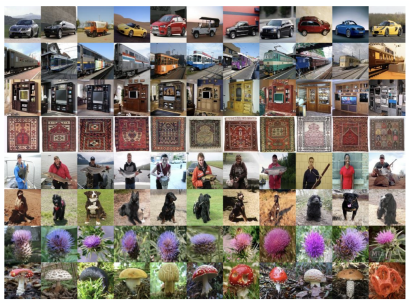

Figure 1. Each rows are the images that lie around the same centroid.