Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey

Seri Lee
Computer Science and Engineering
Seoul National University
Seoul, Republic of Korea
sally20921@snu.ac.kr

***Abstract*—Large-scale labeled data are generally required to train deep neural networks in order to obtain better performance in visual feature learning from images or videos for computer vision applications. To avoid extensive cost of collecting and annotating large-scale datasets, as a subset of unsupervised learning methods, self-supervised learning methods are proposed to learn general image and video features from large-scale unlabeled data without using any human-annotated labels. This paper provides an extensive review of deep learning-based self-supervised general visual feature learning methods from images or videos. First, the motivation, general pipeline, and terminologies of this field are described. Then the common deep neural network architectures that is used for self-supervised learning are summarized. Next, the schema and evaluation metrics of self-supervised learning methods are reviewed followed by the commonly used image and video datasets and the existing self-supervised visual feature learning methods. Finally, quantitative performance comparisons of the reviewed methods on benchmark datasets are summarized and discussed for both image and video feature learning. At last, this paper is concluded and lists a set of promising future directions for self-supervised visual feature learning.**

***Index Terms*—Self-supervised Learning, Convolutional Neural Network, Transfer Learning, Deep Learning**

I. Introduction

A. Motivation

Due to the powerful ability to learn different levels of general visual features, deep neural networks [1] have been used as the basic structure to many computer vision applications such as object detection, semantic segmentation, image captioning, etc. The models trained from large-scale image datasets like ImageNet 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.

The performance of deep convolutional neural networks (ConvNets) greatly depends on their capability and the amount of training data. Different kinds of network architectures were developed to increase the capacity of network models, and larger and larger datasets were collected. Various networks including AlexNet, VGG, GoogLeNet, ResNet, and DenseNet, and large scale datasets such as ImageNet, OpenImage have been proposed to train very deep ConvNets. With the sophisticated architectures and large-scale datasets, the performance of ConvNets keeps breaking the state-of-the-arts for many computer vision tasks.

However, collection and annotation of large-scale datasets are time-consuming and expensive. As one of the most widely used datasets for pre-training very deep 2D convolutional neural networks (2DConvNets), ImageNet contains about 1.3 million labeled images covering 1,000 classes while each image is labeled by human workers with one class label. Compared to image datasets, collection and annotation of video datasets are more expensive due to the temporal dimension. The Kinetics dataset, which is mainly used to train ConvNets for video human action recognition, consists of 500,000 videos belonging to 600 categories and each video lasts around 10 seconds. It took many Amazon Turk workers a lot of time to collect and annotate a dataset at such a large scale.

To avoid time-consuming and expensive data annotations, many self-supervised methods were proposed to learn visual features from large-scale unlabeled images or videos without using any human annotations. To learn visual features from unlabeled data, a popular solution is to propose various pretext tasks for networks to solve, while the networks can be trained by learning objective functions of the pretext tasks and the features are learned through this process. Various pretext tasks have been proposed for self-supervised learning including colorizing grayscale images, image inpainting, image jigsaw puzzle, etc. The pretext tasks share two common properties: (1) visual features of images or videos need to be captured by ConvNets to solve the pretext tasks, (2) pseudo labels for the pretext task can be automatically generated based on the attributes of images or videos.

The general pipeline of self-supervised learning is shown in Fig 1. 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 learning is finished, the learned visual features can be further transferred to downstream tasks (especially when only relatively small data available) as pre-trained models to improve performance and overcome over-fitting. Generally,

shallow layers capture general low-level features like edges, corners, and textures while deeper layers capture task related to high-level features. Therefore, visual features from only the first several layers are transferred during the supervised downstream task training phase.

B. Term Definition

To make this survey easy to read, we first define the terms used in the remaining sections.

- **Human-annotated label**: Human-annotated labels refer to labels of data that are manually annotated by human workers.
- **Pseudo label**: Pseudo labels are automatically generated labels based on data attributes for pretext tasks.
- **Pretext Task**: Pretext tasks are pre-designed tasks for networks to solve, and visual features are learned by learning objective functions for pretext tasks.
- **Downstream Task**: Downstream tasks are computer vision applications that are used to evaluate the quality of features learned by self-supervised learning. These applications can greatly benefit from the pre-trained models when training data are scarce. In general, human-annotated labels are needed to solve the downstream tasks. However, in some applications, the downstream task can be the same as the pretext task without using any human-annotated labels.
- **Semi-supervised Learning**: Semi-supervised learning refers to learning methods using a small amount of labeled data in conjunction with a large amount of unlabeled data.
- **Self-supervised Learning**: Self-supervised learning refers to learning methods in which ConvNets are explicitly trained with automatically generated labels.

Since no human annotations are needed to generate pseudo labels during self-supervised training, very large-scale 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 paper provides a comprehensive survey of deep ConvNets-based self-supervised visual feature learning methods.

II. Commonly Used Pretext and Downstream Tasks

Generally, a pretext task is defined for ConvNets to solve and visual features can be learned through the process of accomplishing this pretext task. The pseudo labels P for pretext task can be automatically generated without human annotations. ConvNet is optimized by minimizing the error between the prediction of ConvNet O and the pseudo labels P. After the training on the pretext task is finished, ConvNet models that can capture visual features for images or videos are obtained.

A. Learning Visual Features from Pretext Tasks

To relieve the burden of large-scale dataset annotation, a pretext task is generally designed for networks to solve while pseudo labels for the pretext task are automatically generated based on data attributes. Many pretext tasks have been designed and applied for self-supervised learning such as foreground object segmentation, image inpainting, clustering, image colorization, temporal order verification, visual audio correspondence verification, and so on. Effective pretext tasks ensure that semantic features are learned through the process of accomplishing the pretext tasks.

Take the image colorization as an example; image colorization is a task to colorize gray-scale images into colorful images. To generate realistic colorful images, networks are required to learn the structure and context information of images. In this pretext task, the data X is the gray-scale image which can be generated by performing a linear transformation in RGB images, while the pseudo label P is the RGB image itself. The training pair X_i and P_i can be generated in real time with negligible cost. Self-supervised learning with other pretext tasks follow a similar pipeline.

B. Commonly Used Pretext Tasks

According to the data attributes used to design pretext tasks, we summarize the pretext tasks into four categories: generation-based, context-based, free semantic label-based, and cross modal-based.

Generation-based Methods: This type of method learn visual features by solving pretext tasks that involve image or video generation.

- **Image Generation**: Visual features are learned through the process of image generation tasks. This type of methods include image colorization, image super resolution, image inpainting, image generation with Generative Adversarial Networks (GANs).
- **Video Generation**: Visual features are learned thorugh the process of video generation tasks. This type of methods include video generation with GANs, and video prediction.

C. Commonly Used Downstream Tasks for Evaluation

To evaluate the quality of the learned image or video features by self-supervised methods, the learned parameters by self-supervised learning are employed as pre-trained models and then fine-tuned on downstream tasks such as image classification, semantic segmentation, object detection, and action recognition etc. The performance of the transfer learning on these high-level vision tasks demonstrates the generalization ability of the learned features. If ConvNets of self-supervised learning can learn general features, then the pre-trained models can be used as a good starting point for other vision tasks that require capturing similar features from images or videos.

Image classification, semantic segmentation, and object detection usually are used as the tasks to evaluate the generalization ability of the learned image features by self-supervised learning methods, while human action recognition in videos

is used to evaluate the quality of video features obtained from self-supervised learning methods.

III. Image Feature Learning

In this section, three groups of self-supervised image feature learning methods are reviewed including generation-based methods, context-based methods, and free semantic label-based methods.

A. Generation-based Image Feature Learning

Generation-based self-supervised methods for learning image features involve the process of generating images including image generation with GAN (to generate fake images), super-resolution (to generate high-resolution images), image inpainting (to predict missing image regions), and image colorization (to colorize gray-scale images into colorful images). For these tasks, pseudo training labels P usually are the images themselves and no human-annotated labels are needed during training, therefore, these methods belong to self-supervised learning methods.

The pioneer work about the image generation-based methods is the Autoencoder which learns to compress an image into a low-dimensional vector which then is un-compressed into the image that is closest to the original image with a bundle of layers. With an auto-encoder, networks can reduce the dimension of an image into a lower dimension vector that contains the main information of the original image. The current image generation-based methods follow a similar idea but with different pipelines to learn visual features through the process of image generation.

References

[1] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 1(2), 2020.