CS182/282 Final Project Commentary

1. Introduction

For the final project, our group has created a homework assignment which allows students to engage with a state-of-the-art unsupervised learning method presented in the paper, *'Unsupervised representation learning by predicting image rotations'* [1]. The idea involves learning semantic image features by predicting rotations of unlabeled images, and using the learned features for further tasks of interests. Throught the assignment, students are expected to learn the core deep neural network concepts, such as self-supervised learning, feature extraction, and transfer learning. Except the dataloader, the entire assignment was created using `Jax` and `Flax` python packages.

2. Rotation Net Unsupervised Learning

2.1 Main Idea

The suggested self-supervised learning technique uses a ConvNet model trained to estimate rotations of unlabelled images as a feature extraction method. The core intuition behind the method is as follows: a model cannot accurately estimate image rotations unless it first learns salient visual features present in images. Once a model trained on image rotations, it can can further be transferred and used for other visual tasks of interests.

2.2 Method

The main objective of the self-supervised learning technique is to extract useful semantic features from a large amount of unlabeled training images. To achieve this goal, unlabeled training images are rotated by multiples of 90°, (90°, 180°, 270°and 360°) and a ConvNet classifier model $F(\cdot)$ ("RotNet" hereafter) is trained to estimate the rotations of images. Mathmatically, a rotation operator $g(\cdot|y)$ (where $y = 90°, 180°, 270°$ and $360°$), is applied to an image X, and the model $F(\cdot)$ outputs a probability distribution of image rotations $F(X^{y^*}|\theta) = \{F^y(X^{y^*}|\theta)\}_{y=1}^{K}$, where y^* and θ are a rotation label (unknown to the model $F(\cdot)$) and a set of learnable model parameters, respectively. The goal of the model is to solve the following objective function from a set of N training images $\{X_i\}_{i=0}^{N}$,

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} loss(X_i, \theta)$$

where *loss* is a multi-class cross-entropy loss,

$$loss(X_i, \theta) = -\frac{1}{4} \sum_{y} \log F^y(X^y|\theta)$$

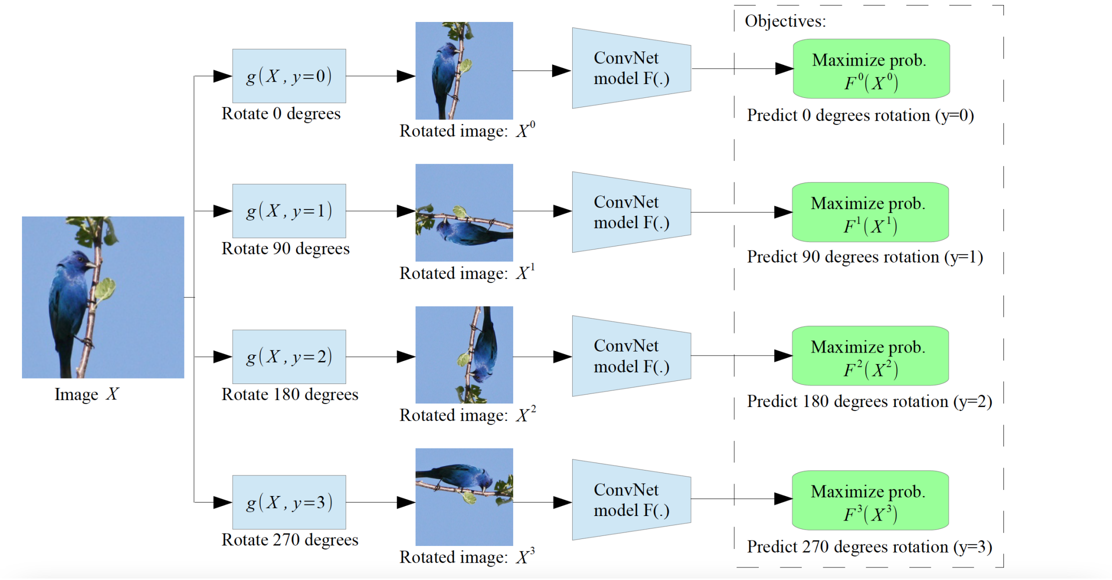

Figure 1: A schematic diagram for the suggested RotNet sel-supervision task. The RotNet outputs a full probability distribution of possible rotations (0, 90, 180 and 270 degrees). Image from [1]

Once the RotNet is trained, the RotNet classifier layer, along with the last few convolutional blocks, are removed from the RotNet and the remaining structure is directly connected to a new ConvNet (PredNet hereafter) designed for a new task of interest. In other words, once the RotNet is trained on rotated images, it is used for a visual feature generator. It is reported in the original paper that the suggested RotNet unsupervised method achieves test accuracy as high as 91.16% on CIFAR-10 dataset, which is comparable to the test accuracy of a fully supervised model, 92.80%.

2.3 Assignment

An assignment was created for students to engage with the core ideas of the suggested self-supervision method. The assignment includes the followings: a detailed instruction on how to setup an environment, a dataloader for CIFAR-10 training and test datasets, fully implemented RotNet and PredNet models for transfer learning, an assignment where students are asked to implement some core methods.

Students will accomplish an image classification task using both unsupervised representation learning together with transfer learning and fully supervised learning. They will compared the two method to understand that the former can give better performance than the latter with a faster training speed. In addition, they will conduct experiments on using the learned feature representation instead of the images for the classification task and find out that the performance is not much worse than the fully supervised method, which means the most important information of the images are captured in the learned representation.

Moreover, students will try extracting the feature representation from different layers to verify that there is indeed a difference between which layer to extract the features.

3. Reference

1. Gidaris, S., Singh, P. & Komodakis, N. *Unsupervised Representation Learning by Predicting Image Rotations* arXiv:1803.07728 [cs]. Mar. 2018. http://arxiv.org/abs/1803.07728 (2022).