Summary - Pseudo-Label : The Simple and Efficient Semi-Supervised Learning method for Deep Neural Network

Jaechan Park

k11850713@students.jku.at

1 Introduction

This paper [1] suggests a new handy and efficient way of determining proper labels for an unlabeled dataset used for supervised learning to predict performance better. Most training work by deep neural network models has two essential parts - *unsupervised pre-training* and *fine-tuning* of a supervised learning method. The most difficulty of using a supervised learning method is acquiring a high-quality labeled dataset, challenging human effort, and capital. Thus, Semi-Supervised learning could be a remedy to address those mentioned above cumbersome of achieving a high-quality labeled dataset by reduction of human resource and cost.

Some researchers proposed notable Semi-Supervised learning methods of being applied both supervised and unsupervised learning methods at the same time. In this paper [1], a new approach of Semi-Supervised learning proposes to assign proper "$fake$" labels for each unlabeled data samples, and the newly assigned "$fake$" labels are *Pseudo-Label*. The assigning criterion of *Pseudo-Label* is on the maximum forecast probability after weights update from understanding the distribution and formation of unlabeled data samples.

2 Theoretical backgrounds for Deep Neural Network

As Semi-Supervised learning relies on the model architecture of deep neural network (DNN), the basic concept of DNN architecture is essential to make a better understanding of further explanation in this study.

2.1 Basic Knowledge for Deep Neural Networks

The beauty of DNN architecture is multiple-layer neural networks able to outperform conventional machine learning methods. The equation (1) describes the M layers of hidden units in the mathematical statement.

$$h_i^k = s^k \left(\sum_{j=1}^{d^k} W_{ij}^k h_j^{k-1} + b_i^k \right), \ k = 1, ..., M+1 \tag{1}$$

where s^k represents a non-linear activation function of the k-th layer and h_i^{M+1} is output units for predicting target class.

In the DNN model, an appropriate activation function must be used for units to train a model accordingly and have advanced prediction performance. There are two popular functions for binary classification task can be thought being applied for activation unit function in this research.

The first activation unit function is the Sigmoid function. This function yields a binary probability after activation, which enables to produce a decision of labeling for unlabeled data samples. Another activation function is the Rectified Linear activation function. This activation function returns sparse representations of a considerable volume of real zeros after activating its inputs in the units. With the sparse representations, deep neural network models could expect less demanding resources to be developed and better prediction performance.

Minimized supervised loss function uses to train the entire network. Also, Cross-Entropy can be another choice for supervised loss function in case of using the sigmoid function for output units. The mathematical expression for loss function and *Cross Entropy* are defined to the equation (2) and equation (3), respectively.

$$\sum_{i=1}^{C} L(y_i, f_i(x)) \tag{2}$$

$$L(y_i, f_i) = -y_i log f_i - (1 - y_i) log(1 - f_i) \tag{3}$$

where, $y_i =$ is the 1-of-K code of the label, $f_i =$ is the network output for i'th label and $x =$ represents input vector.

2.2 Denoising Auto-Encoder and Dropout

During the *unsupervised pre-training* stage, Denoising Auto-Encoder (DAE) algorithm is to develop neural network layers with initializing the weights through training autoencoders. In this paper [1], it adopts a probability 0.5, adding to the DAE algorithm as noise for the corruption of input data structure.

Over-fitting is a big deal issue in supervised learning fashion that has to address to obtain optimized prediction performance. Dropout is a well-known technique enabling reducing over-fitting matter by alleviating complex co-adaptations on hidden representations. Stochastic Gradient Descent (SGD) with dropout is for training the neural network in this study. Momentum with a high initial value of exponentially decaying learning rate is to accelerate training speed. All parameter settings remain to the original dropout paper except for the weight regularization.

3 Pseudo-Label

3.1 Classification in Low-Density and Entropy Regularization

Newly given labels of unlabeled datasets serve as true labels and use supervised learning tasks with updating every weight during the *fine-tuning* phase. Equation (4) represents the overall loss function of balancing the number of data samples between labeled and unlabeled datasets. The value of $\alpha(t)$ represents a coefficient of adjusting between labeled and unlabeled datasets. Fiding a proper value of the coefficient is a vital task in this work.

$$L = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{C} L(y_i^m, f_i^m) + \alpha(t)\frac{1}{n'}\sum_{m=1}^{n'}\sum_{i=1}^{C} L(y_i^{'m}, f_i^{'m}) \tag{4}$$

where, n = represents the number of mini-batch in labeled data for SGD, n' = represents the number of mini-batch in unlabeled data for SGD, f_i^m = represents the output units of m's sample in labeled data, y_i^m = represents the label for m's sample, $f_i^{'m}$ = represents activation function for unlabeled, $y_i^{'m}$ = represents the pseudo-label of m's unlabeled sample, $\alpha(t)$ = represents coefficient for balancing terms.

Semi-supervised learning reinforces the usage of the unlabeled dataset to increase generalization performance by the *cluster assumption*. With the assumption, it empowers to provide the classification boundary in low-density areas. There are two methods used to support the *cluster assumption* in this study - *Semi-Supervised Embedding* and *Manifold Tangent Classifier*. Applying embedding-based regularization in the neural network is to predict the class of an unlabeled data sample developed from its neighbors' in a high-density region. Another method of the *cluster assumption*, *Manifold Tangent Classifier*, is to make less sensitive to variations in the directions of the low-dimensional manifold.

Entropy Regularization enables an unlabeled dataset to advantage its availability obtained from maximum a posterior estimation (MAP). This approach is to

reduce the conditional entropy of class probability for unlabeled data. Then, the minimized entropy makes a less dense population of data instances in low-density regions. The equation (5) is defined as the Cross-Entropy in mathematical terms in the paper[1].

$$H(y|x') = -\frac{1}{n'}\sum_{m=1}^{n'}\sum_{i=1}^{C} P(y_i^m = 1|x'^m) logP(y_i^m = 1|x'^m) \tag{5}$$

where, n' represents the number of unlabeled data, C represents the number of classes, y_i^m represents the unknown label of the m-th unlabeled sample, x'^m represents the input vector of m-th unlabeled sample.

Equation (6) is a description of the mathematical expression on how to assess the posterior distribution maximum. The first term of equation (5) is the conditional log-likelihood of labeled data, which has to maximize for optimized performance. Whereas, the second term, the entropy of unlabeled data, has to be reduced for the same purpose as the first term in the equation. The regions of overlapping classes can be defined as entropy. Reducing the overlapping area makes less dense data points at the classification or decision boundary.

$$C(\theta, \lambda) = \sum_{m=1}^{n} logP(y^m|x^m;\theta) - \lambda H(y|x';\theta) \tag{6}$$

where, n is the number of labeled data, x^m is the m-th labeled sample, and λ is balance coefficient of two terms.

3.2 Training with Pseudo-Label

According to figure 1 in the paper [1] of t-SNE 2-D embedding on MNIST, using 60,000 unlabeled samples and Pseudo-Labels shows the superior achievement of clustering than using 600 labeled samples solely. Hence, the cluster assumption in this work seems to work to increase generalization performance.

4 Experimental Methods and Results

Each reduced 100, 600, 1000, and 3000 of MNIST labeled data is prepared for training trials. Also, a separated 1000 labeled dataset is prepared for the validation set to optimize hyperparameters. Labels of the remaining labeled data samples after the preparation for training sets and validation sets are treated as unlabeled samples after throwing away corresponding labels. Except for the experiment with 100 reduced labeled data set, 10 different trials for each reduced set is conducted with the same hyper-parameters and network. Because the result of the 100 labeled data samples set heavily depends on the data split, 30 experiment trials with identical network and hyper-parameter settings. The confidence interval for each 100, 600 and 1000 labeled data is $95\% \pm 1 \sim 1.5\%$,

$\pm 0.1 \sim 0.15\%$ and less than $\pm 0.1\%$, respectively.

This study adopts a hidden layer with 5000 hidden units in the neural network. The Recitified Linear function is used for the hidden activation function, whereas the sigmoid activation function uses for the output unit. The rate of initial learning uses at 1.5, and the number of mini-batch for labeled data and unlabeled data is 32, 256, respectively. Table 1 in the paper [1] reports a specific error rate comparison between the proposed method and the conclusions from (Weston et.al., 2008; Rifai er al., 2011b).

In conclusion, the new approach of Semi-Supervised learning, *Pseudo-Label* with Denoising Auto-Encoder, outperforms other compared techniques for most experimental trials with the different number of the labeled data set.

References

[1] Dong-Hyun Lee. *Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks.* ICML 2013 Workshop in Challenges in Representation Learning: The Black Box Learning Challenge., 2013.