Summary - Realistic Evaluation of Deep Semi-Supervised Learning Algorithms

Jaechan Park

k11850713@students.jku.at

1 Introduction

It is a great deal of manually selecting reliable labels for an unlabeled dataset requiring a lot of human effort and capital. In recent, a new way of machine learning method, Semi-Supervised Learning (SSL), is treated as to address those as mentioned cumbersome and improve prediction performance. The main idea of SSL behinds learning on the structure of unlabeled data samples for the sake of appointing proper labels for the unlabeled dataset. The aim of this paper [1] provides more direct and pragmatic evaluation rules of the SSL technique, which apply to a dataset of real-world problems.

2 Semi-Supervised Learning Techniques

The purpose of the Supervised learning method is to formulate a parameterized prediction function. In the supervised learning model's classification task, the devised prediction function could provide the correct class or unseen dataset label. Hence, a high-quality labeled dataset, $(x, y) \in D$, is a fundamental resource to train a supervised model. Also, an optimal parameter has to be set in the prediction function $f_\theta(x)$ to achieve successful prediction outcomes for unseen input samples.

Semi-Supervised learning (SSL) can leverage unlabeled data samples, $x \in D_{UL}$, contributing to enhancing the performance of prediction function, $f_\theta(x)$, by learning the structure of the unlabeled data samples. Therefore, the different structure of the unlabeled data samples provides a satisfying label of which is functioning as an actual label in the training of supervised learning fashion.

In this paper [1], two major classes - Consistency regularization and Entropy Minimization - are used in the image classification task. Consistency regularization addresses input data samples' interference for obtaining stable predicted outputs by prediction function $f_\theta(x)$. And, Entropy Minimization is to enhance

the degree of confidence for predictions on unlabeled data.

To be more specific, the Π-Model is a simple stochastic method applied to consistency regularization. In the process, adding a loss term is to minimize the variance of the outputs when the same input sample passes through the neural network. The Mean-Teacher model addresses the issue mentioned above of Π-Model, which is a potential risk of different outputs for the same input. Thus, the outcome is predicted by using an exponential moving average of parameters derived from previous training steps in the Mean-Teacher method. Virtual Adversarial Training applies a little noise to input data after directly estimated the tiny noise. The applied noise to the input dataset would play an essential role in predicting outputs by the prediction function. Entropy Minimization involves a loss term to achieve more confident expected outcomes for unlabeled data samples. *Pseudo-Label* is the method to assign proper labels of an unlabeled dataset by a pre-trained model. The assigned labels, *Pseudo-Labels*, participate in the training of a standard supervised learning as the "true" labels.

3 Experimental Criteria and Results

As this study's fundamental purpose is to produce a better accurate and practical assessment scheme, the accompanying suggestive criteria for assessment promote evaluations more realistic and applicable to a dataset acquired from real-world problems.

3.1 A Shared Implementation for Reproduction

To approach a more direct and realistic comparison of evaluations, it requires to minimize the discrepancy of assessment across the underlying SSL models in the study. Thus, the introduction of a shared implementation across the underlying SSL architectures could alleviate further possible variability in results of which could cause potential danger of reproduction with the underlying SSL models.

Wide ResNet is a typical neural network architecture on the image classification task. The model, WRN-28-2, consists of ResNet with depth 20 and width 2. Also, batch normalization and leaky ReLU nonlinearities use in the model (WRN-28-2). As well, the ubiquitous Adam optimizer applies to train the model. To the fairness and equality for evaluation comparisons, the best performance outcomes use for all of each SSL model after 1000 trials of Gaussian Process-based black-box optimization in Google Cloud ML Engine.

There are two common datasets for image classification task - SVHN and CIFAR-10 - used to conduct this work. For supervised learning fashion, 1,000 and 4,000 labeled data samples perform for each SVHN and CIFAR-10, respectively. To be ready for both the training and validation set, the standard split uses the two datasets. Hence, the number of training dataset for SVHN and CIFAR-10

are 65,932 and 45,000, respectively. Whereas, the number of validation dataset for SVHN and CIFAR-10 are 7,325 and 5,000, respectively. Table 1 provides more detailed results of error rates from the various SSL models and baseline for each of CIFAR-10 and SVHN.

Dataset	Labels	Supervised	$\Pi - Model$	Mean Teacher	VAT	VAT + EntMin	Pseudo-Label
CIFAR-10	4000	20.26 ±0.38%	16.37 ±0.63%	15.87 ±0.28%	13.86 ±0.27%	13.13 ±0.39%	17.78 ±0.57%
SVHN	1000	12.83 ±0.47%	7.19 ±0.27%	5.65 ±0.47%	5.63 ±0.20%	5.35 ±0.19%	7.62 ±0.29%

Table 1: Test Error rates of SSL models and baselines

3.2 High-quality Fully-Supervised Baselines

A high-quality baseline is an essential prior condition to make the evaluation comparisons across the models more reliable. In order to obtain a high-quality baseline in fully-supervised learning, 1000 trials use to set optimized hyperparameters. As well, the same trial procedure applies to all the underlying SSL models in this study. According to table 2 in the paper [1], it represents the error rate change between fully-supervised baseline and reported Semi-supervised learning models. Based on the outcomes, the general tendency for the difference of error rates between the fully-supervised learning and each experimented SSL models show smaller than the reported results in the literature.

3.3 Transfer Learning

"Transfer" learning technique is a powerful and practical technique when the size of available labeled data is limited and costly to obtain. This technique's idea is that a new model's initialized parameters can be ready through training with a separate labeled dataset. To apply "Transfer" learning method, the independent labeled dataset's domain properties have to be related to a new dataset to be used for *fine-tuning* in the same network.

The standard WRN-28-2 model with ImageNet downsampled to 32 x 32, the original size of CIFAR-10, is used for this study. And, same hyperparameters of the supervised baseline use for training of the network. After the previous training procedures, *fine-tuning* for the network is conducted using 4,000 labeled data samples from CIFAR-10. The test set's error rate with the trained network model is 12.09%, which is lower than the error rate from any SSL models. Therefore, the lower error rate is an indication that the two used datasets are related, and the network is favorable for "Transfer" learning.

Another experiment is performed along the same settings and procedures to investigate the effect of overlapping of related classes between 252 ImageNet and CIFAR-10 after removing the two datasets' comparable classes. The result of the error rate, 12.91%, shows no significant effect. For the SVHN dataset, there are no results reported in the paper [1] because no conclusive results reached using "Transfer" learning from ImageNet to the dataset of SVHN.

3.4 Class Distribution Mismatch

After excluding 252 ImageNet classes that are expected comparable, An ideal case of obtaining the best result from Semi-Supervised learning is that class distributions of labeled and unlabeled data sets are identical. However, the possibility only occurs in the theoretical concept, not in real studies. Hence, measurement of the degree of class distribution matching is needed to reach a better performance when a SSL model runs for training. In this study, the labeled dataset has to separate into two individual data sets with a proper ratio. A separated dataset is required to dispose of corresponding labels as acting to the unlabeled dataset.

In order to test class distribution matching, six animal classes from CIFAR-10 are produced for classification tasks. On the other hand, four classes are set in the unlabeled dataset. As well, a fully supervised training without an unlabeled dataset runs for the comparison. 400 data samples are given to each class; thus, the entire amount of manually labeled data samples is 2,400.

According to figure 2 in the paper [1], it gives a hint that improper predicted labels would cause a severe negative impact on forecast performance compared to the results without using all of the unlabeled data samples.

3.5 Varying the Amount of Labeled *and* Unlabeled Data

Two feasible methods are to obtain different sizes of the labeled dataset. A typical way of acquiring different sizes for labeled data sets is that withdrawing a particular portion of labels in the labeled dataset. Another method is less common practice and divided into two realistic scenarios. The first scenario is that the unlabeled dataset size is vast, whereas the second one is that the unlabeled size is relatively tiny.

To examine performance results over different unlabeled dataset sizes, the first method for different sizes of each CIFAR-10 and SVHN is set several different labeled sizes for each of the two datasets. After experiments with various labeled datasets, a significant outcome found that all of the SSL methods converge as increasing the number of labels. Although there are some notable outcomes have been found in the VAT and Π-model methods, it is hard to lead a deterministic insight across all of the SSL methods. There is another way to set the different sizes of the unlabeled dataset. An extended SVHN dataset uses various unlabeled data sizes for the SVHN dataset with additional 531,131 digit images.

Eight million additional unlabeled images augment with the "Tiny Images" dataset to set different unlabeled CIFAR-10 data sizes. However, the expanded supplementary dataset with the CIFAR-10 dataset is not favorable for this study since the mismatching distribution between the labeled dataset, and the unlabeled dataset is significantly huge.

In general, the results of the error rate of SSL methods on SVHN with 1,000 labels and the different sizes of SVHN-extra unlabeled data show improvement of the performance of SSL methods as growing the size of the unlabeled dataset. However, the SSL methods of both Pseudo-Labeling and Π-Model tend to diminish lightly along with increasing the unlabeled dataset's size.

To sum up, the analysis of the effect of different amounts of data samples shows that the measurement of sensitivity across the SSL models is highly diverse.

3.6 Realistically Small Validation Sets

In real-world applications, the amount of validation dataset is relatively small compared to the training dataset size. However, an unusual augmented SSL dataset has an abnormal amount of validation data samples. The odd validation data size would cause a severe issue for hyperparameters settings while training in a neural network. Hoeffding's inequality estimates a reliable theoretical amount of validation data samples.

$$P(|\bar{V} - E[V]| < p) > 1 - 2exp(-2np^2)$$

where, $\bar{V}$, $E[V]$, p, and n represent the empirical estimate of the validation error, its hypothetical true value, the desired maximum deviation between our estimate and the true value, and the number of examples in validation set, respectively.

To analyze validation error theoretically, it requires computing the average value of independent indicator function variables. However, the theoretical analysis can not be applied to realistic problems because it uses the validation accuracy derived from the average independent binary indicator. Thus, in approaching a more practical way for this study, each SSL model is applied to be trained with the baseline models on the SVHN dataset. The 1,000 labeled samples and several different validation data sizes use for evaluation comparison. Figure 6 in the paper [1] is the line graph providing the mean and standard deviation for validation error across the percent of 10 randomly-sampled non-overlapping validation dataset size to training dataset size. With the 10% validation data size, which is a realistic amount, the performance with the validation data size does not significantly differ from other validation sizes. It means that a massive validation data size requires setting optimal hyperparameters, not applicable to real-world problems. A possible remedy of the validation data size is cross-validation, but it has to deal with variance.

According to figure 6 in the paper [1], it shows the measurement of correlation across the SSL models when the experiment coducts with the same validation dataset. It require to compute the mean and standard deviation values of the difference in validation error between each SSL technique and Π-model. As the

overlapping regions of the difference of validation error across the SSL models are increasing along the decreasing the percentage of validation data compared to training data size, the validation data size does not influence distinction for model selection.

4 Conclusion

This study provides three useful tips to use SSL model in machine learning tasks. The first considerable case to apply SSL model is that when it is not possible to gain a high-quality labeled dataset from similar domains for fine-tuning. The second case is that the labeled data samples are independent and identically distributed and withdrawn from the unlabeled dataset pool. The last favorable case for SSL model is that when the amount of the labeled dataset size is sufficiently large to provide a correct assessment of validation accuracy, it contributes to making a model selection and hyperparameters tuning.

References

[1] Colin Raffel Ekin D. Cubuk Ian J. Goodfellow Avital Oliver, Augustus Odena. *Realistic Evaluation of Deep Semi-Supervised Learning Algorithms*. NeurIPS 2018 Proceedings, 2018.