Stochastic Consensus: Enhancing Semi-Supervised Learning with Consistency of Stochastic Classifiers

Hui Tang[1], Lin Sun[2], and Kui Jia[1]

[1] South China University of Technology, Guangzhou, China
eehuitang@mail.scut.edu.cn, kuijia@scut.edu.cn
[2] Magic Leap, Sunnyvale, CA, USA
lsun@magicleap.com

A Contribution Summary

Our contributions are summarized below.

- We propose Stochastic Consensus (STOCO), a novel sample filter for better semi-supervised learning, where the consistency criterion among multiple, stochastic classifiers is introduced. Our proposed STOCO carries forward the advantages of self-training and tri-training to select unlabeled samples and generate their pseudo labels more reliably while circumventing the disadvantage of model size expansion.
- Specifically, we sample the classifiers from a Gaussian distribution whose parameters are jointly optimized in training, which can dynamically capture the pattern of sensible decision boundaries; learning a Gaussian classifier also acts as an implicit regularization to alleviate overfitting to small amounts of labeled data. Moreover, the pseudo labels are generated by deep discriminative clustering in order to consider the intrinsic data structures and encourage cluster size balance.
- We also provide theoretical analysis by connecting with the probably approximately correct (PAC) theory on learning from noisy data, which confirms the rationality of our method. Our theoretical analysis connects noisy label learning to SSL, and can serve as a *general* analytical method for pseudo-labeling based SSL frameworks including our STOCO. The result of such an analysis is an important indicator for effective pseudo label generation.
- Experiments on four typical benchmark datasets have demonstrated that our proposed STOCO outperforms existing methods and achieves the state of the art, especially in label-scarce cases.

B Classifier Variance Visualization

To intuitively inspect the behaviour of the learned Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ during model training, we visualize the average variance over all dimensions by $1/d^{\boldsymbol{\sigma}} \sum_i |\sigma_i|$, where $d^{\boldsymbol{\sigma}}$ is the number of elements in the learned $\boldsymbol{\sigma}$, namely the

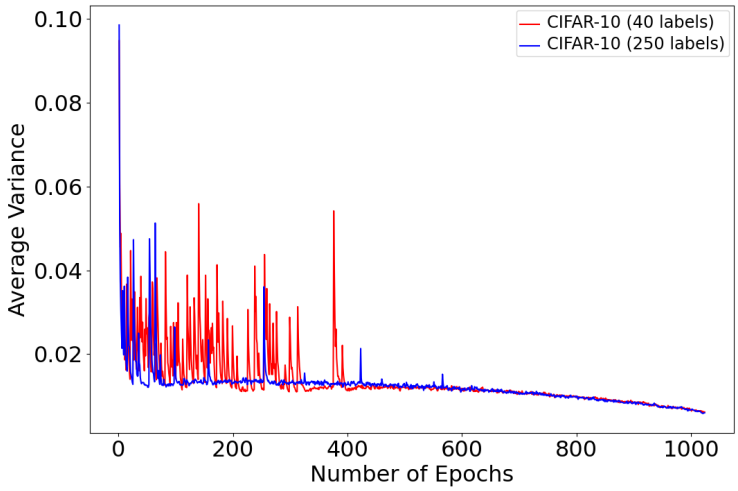

Fig. A1: Average variance of the learned Gaussian distribution during model training. Please refer to the main text for how it is computed.

dimension of the flattened classifier weight matrix, and σ_i is the i-th element of $\boldsymbol{\sigma}$. The results on individual 40- and 250-label splits from CIFAR-10 [3] are shown in Fig. A1. We observe that as the learning process proceeds, the average variance gradually decreases, meaning that the discrepancy between stochastic classifiers reduces. This observation also suggests the convergence of the learned Gaussian distribution, which guarantees the stability of model training and performance improvement.

Specifically, STOCO learns different classifiers sampled from a Gaussian distribution to minimize the same loss function during training; namely, the Gaussian is learned to yield a set of classifiers with reduced loss (cf. Fig. 2) and *decreased variance* (cf. Fig. A1, where it *gradually converges*). Our STOCO achieves higher accuracy than FixMatch (cf. Table 2), verifying that more sensible decision boundaries are learned by STOCO. What makes STOCO special is a consistency criterion among multiple stochastic classifiers, which contributes to the advantages (cf. Table 1). These analyses can support that the Gaussian in STOCO learning does dynamically capture the pattern of sensible decision boundaries.

C Design Justification for the Ensemble of Stochastic Classifiers

We have justified the ensemble design on a single 40-label split from CIFAR-10 and find that compared to the element-wise product, an average of category predictions degenerates.

D Validation on ImageNet

We also verify the effectiveness of STOCO on ImageNet-1K [1], which is a much more realistic and complicated dataset. It includes 1.28M training images and 50K validation ones, which are distributed in 1K object categories. During training, we sample 100 labeled instances per class at random from the training set and the remaining ones are unlabeled. At inference, we evaluate the classification result on the validation set. We follow FixMatch [4] and use a pre-activation ResNet-50 [2] as the backbone. The batch size and weight decay are set to 32 and 3e-4 respectively. The rest of hyperparameter settings are the same as those for other datasets in Sec. 4.1. On a single 100K-label split from ImageNet-1K, our method outperforms FixMatch by a large margin (3.69%), though more studies on different splits are certainly necessary to be conducted.

E Comparison of Inference Time

On ImageNet-1K with the same baseline condition, the training speeds of STOCO ($m = 5$) and FixMatch are 1.16 s/iter and 0.97 s/iter respectively; at inference, they have the same time cost since they both forward through one classifier only. As mentioned in Sec. 4.1, for inference, we use a fixed classifier determined by the learned mean $\boldsymbol{\mu}$. We emphasize that our method as a plug-in adds very little computation.

F Future Work

The underlying mechanism behind STOCO is that the model would be less likely to be misled if there are fewer and fewer mislabeled samples in the selected pseudo-labeled set. On the other hand, the model's generalization performance would be continuously improved if there are more and more representative samples, i.e., high-quality ones. In this work, we push forward an important step along this line and more explorations are expected. For example, more careful studies in different self-training algorithmic frameworks are to be conducted; the noise rate can be iteratively reduced by increasing the number of stochastic classifiers based on a pre-defined schedule during training. A new perspective of uniting SSL and active learning may be also worth studying to reduce the labeling cost by annotating more semantically representative samples.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
2. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV. pp. 630–645 (2016)
3. Krizhevsky, A.: Learning multiple layers of features from tiny images. In: Technical report (2009)
4. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: NeurIPS. vol. 33, pp. 596–608 (2020)