Dual Student: Breaking the Limits of the Teacher in Semi-supervised Learning

(2019)

Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, Rynson W.H. Lau

Summary

Contributions.

The authors propose an improvement over the Teacher-Student setting in semi-supervised learning. With Mean Teachers, a consistency of predictions is imposed between the student and the teacher, where the teacher is an exponential moving average (EWA) of the weights of the student model. The objective is to force the unlabeled data to meet the smoothness assumption (i.e, the learned decision boundary will lie in low density regions). However, the authors state that having a coupled teacher is not sufficient for the student, and propose two independent students. Instead of enforcing a consistency between the two models they introduce a stability constraint, and show better performance over the previous methods.

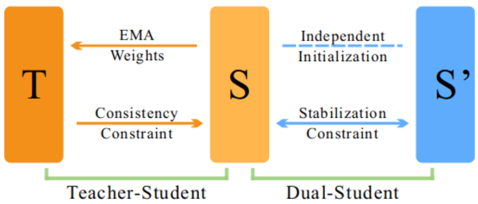

Figure 1: Teacher-Student versus Dual Student. The teacher (T) in Teacher-Student is an EMA of the student (S), imposing a consistency constraint on the student. Their weights are tightly coupled. In contrast, a bidirectional stabilization constraint is applied between the two students (S and S') in Dual Student. Their weights are loosely coupled.

Drawback of Teacher-Student setting.

In Mean Teacher, the cluster assumption - *"If two data points in a high-density region are close, then so should be the corresponding outputs"* is enforced between the predictions of a student model (S), and its EWA version (a teacher model T). The weights θ' of the teacher model are computed at each iteration using the student's θ weights and a weighting factor α: $\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t$. However, given a large number of training iterations, the teacher model will converge to the same weights as that of the student, and no additional and meaningful knowledge can be extracted compared to the student.

This can be see in the two figures bellow. In Figure 2, the distance between the S and the T weights decreases as the training progress, and so does their predictions. Additionally, in Figure 3, we see that any biased prediction from the teacher is carried over to student.

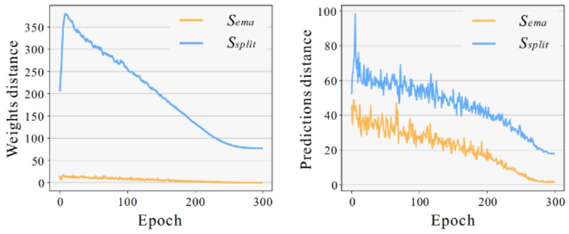

Figure 2: Left: S_{ema} contains two models with similar weights, while the weights of the two models in S_{split} keep a certain distance. Right: The predictions of the two models in S_{split} keep a larger distance than those of S_{ema}.

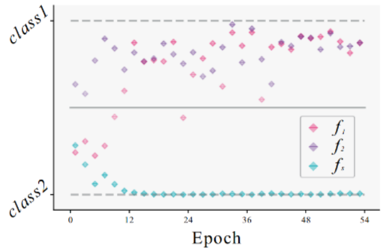

Figure 3: Our method can alleviate the confirmation bias. f_1 and f_2 are the independent students from our Dual Student, while f_s is the student guided by the Mean Teacher. For a misclassified sample (belonging to *class1*), f_1 can correct it quickly with the knowledge from f_2. However, f_s is unable to correct its prediction due to the wrong guidance from the EMA teacher.

Dual-Student

Stability As stated above, with two student models with different initialization, enforcing a consistency between their prediction might result is the models collapsing into each other. To this end, the authors propose a stability constrain to train the models. A stable sample must satisfy these two conditions:

- The perditions using two versions, a clean x and a perturbed version $\bar{x}$ give the same results: $f(x) == f(\bar{x})$.
- Both predictions are confident, ie, are far from the decision boundary. This can be tested by seeing if $f(x)$ (resp. $f(\bar{x})$) is greater than a confidence threshold ϵ (such as 0.1) (if the output is probability density over many classes / values, the max p value needs to be greater than ϵ).

Training To train the two model, we can use the stability constraint to train one of the models for a given unlabeled data point x. First we create a perturbed version of the input, giving us $\bar{x}$. With the two students f_1 and f_2, we compute the four predictions $f_1(x); f_1(\bar{x}); f_2(x); f_2(\bar{x})$. And see if the predictions of each model are stable, if one of them is stable and the other is not, we only train the one with the stable predictions using an unsupervised loss: $\mathcal{L}_{mse}(x) = \|f_1(x) - f_2(x)\|^2$. If both models gave stable predictions. We train the one with the smallest variation between its two predictions: $\mathcal{E}_x^i = \|f_i(x) - f_i(\bar{x})\|^2$. The following Figure summarizes the process.

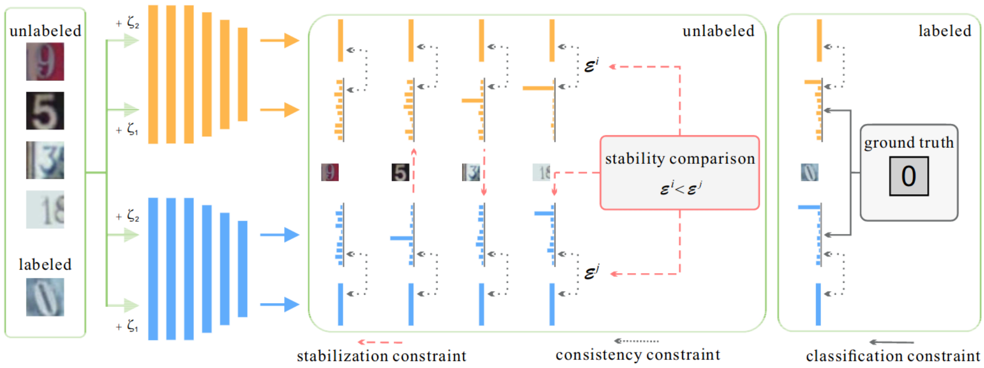

Figure 4: Dual Student structure overview. We train two student models separately. Each batch includes labeled and unlabeled data and is forwarded twice. The stabilization constraint based on the *stable samples* is enforced between the students. Each student also learns labeled data by the classification constraint and meets the *smooth assumption* by the consistency constraint.

Results

Table 1: Test error rate on CIFAR-10 averaged over 5 runs. Parentheses show numbers of training epochs (default 300).

Model	1k labels	2k labels	4k labels	all labels
Π [17]	$31.65 \pm 1.20^\dagger$	$17.57 \pm 0.44^\dagger$	12.36 ± 0.31	5.56 ± 0.10
Π + SN [19]	21.23 ± 1.27	14.65 ± 0.31	11.00 ± 0.13	5.19 ± 0.14
Temp [17]	$23.31 \pm 1.01^\dagger$	$15.64 \pm 0.39^\dagger$	12.16 ± 0.24	5.60 ± 0.10
Temp + SN [19]	18.41 ± 0.52	13.64 ± 0.32	10.93 ± 0.34	5.20 ± 0.14
MT [33]	$18.78 \pm 0.31^\dagger$	$14.43 \pm 0.20^\dagger$	$11.41 \pm 0.27^\dagger$	$5.98 \pm 0.21^\dagger$
MT + FSWA [1]	16.84 ± 0.62	12.24 ± 0.31	9.86 ± 0.27	$\mathbf{5.14 \pm 0.07}$
CS	17.38 ± 0.52	13.76 ± 0.27	10.24 ± 0.20	5.18 ± 0.11
DS	$\mathbf{15.74 \pm 0.45}$	$\mathbf{11.47 \pm 0.14}$	$\mathbf{9.65 \pm 0.12}$	5.20 ± 0.03
MT + FSWA (1200) [1]	15.58 ± 0.12	11.02 ± 0.23	9.05 ± 0.21	4.73 ± 0.18
Deep CT (600) [27]	-	-	9.03 ± 0.18	-
DS (600)	$\mathbf{14.17 \pm 0.38}$	$\mathbf{10.72 \pm 0.19}$	$\mathbf{8.89 \pm 0.09}$	$\mathbf{4.66 \pm 0.07}$

Table 2: Test error rate on CIFAR-100 averaged over 5 runs.

Model	10k labels	all labels
Temp [17]	38.65 ± 0.51	26.30 ± 0.15
Π [17]	39.19 ± 0.36	26.32 ± 0.04
Π + FSWA [1]	35.14 ± 0.71	22.00 ± 0.21
MT [33]	$35.96 \pm 0.77^\dagger$	$23.37 \pm 0.16^\dagger$
MT + FSWA [1]	34.10 ± 0.31	$\mathbf{21.84 \pm 0.12}$
DS	$\mathbf{33.08 \pm 0.27}$	21.90 ± 0.14
MT + FSWA (1200) [1]	33.62 ± 0.54	$\mathbf{21.52 \pm 0.12}$
Deep CT (600) [27]	34.63 ± 0.14	-
DS (480)	$\mathbf{32.77 \pm 0.24}$	21.79 ± 0.11

Table 3: Test error rate on SVHN averaged over 5 runs.

Model	250 labels	500 labels
Supervised [33]	27.77 ± 3.18	16.88 ± 1.30
MT [33]	4.35 ± 0.50	4.18 ± 0.27
DS	$\mathbf{4.24 \pm 0.10}$	$\mathbf{3.96 \pm 0.15}$

Table 4: Test error rate on ImageNet averaged over 2 runs

Model	10% labels-top1	10% labels-top5
Supervised	42.15 ± 0.09	19.76 ± 0.11
MT [33]	37.83 ± 0.12	16.65 ± 0.08
DS	$\mathbf{36.48 \pm 0.05}$	$\mathbf{16.42 \pm 0.07}$

Table 5: Test error rate of two variants of Dual Student (all using the 13-layer CNN) on the CIFAR benchmark averaged over 3 runs. Parentheses of Multiple Student (MS) indicate the numbers of students. Parentheses of Imbalanced Student (IS) indicate the numbers of parameters for the strong student.

Model	CIFAR-10 1k labels	CIFAR-100 10k labels
DS	15.74 ± 0.45	33.08 ± 0.27
MS (4 models)	14.97 ± 0.36	32.89 ± 0.32
MS (8 models)	14.77 ± 0.33	32.83 ± 0.28
IS (3.53M params)	13.43 ± 0.24	32.59 ± 0.27
IS (11.6M params)	$\mathbf{12.39 \pm 0.26}$	$\mathbf{31.56 \pm 0.22}$

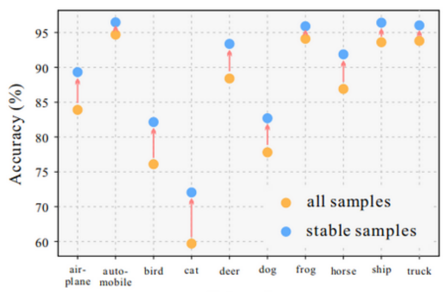

Figure 6: Test accuracy of each category on the *stable samples* and on all samples of CIFAR-10. The performance gap indicates that the *stable samples* represent relatively more reliable knowledge of a model. The average ratio of the *stable samples* on the test set is about 85% w.r.t. the model.