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

Zhanghan Ke[1,2] * Daoye Wang[2] Qiong Yan[2] Jimmy Ren[2] Rynson W.H. Lau[1]
[1] City University of Hong Kong [2] SenseTime Research

Appendix A: Convergence of the EMA

In our paper, we state that the EMA teacher is coupled with the student in the existing Teacher-Student methods. We provide below a formal proposition for this statement and a simple proof.

Proposition 1. *Given a sequence* $\{ s_t \}_{t\in\mathbb{N}} \subseteq \mathbb{R}^m$ *and let* $s'_t = \alpha\, s'_{t-1} + (1-\alpha)\, s_t$*, where* $0 < \alpha < 1$*,* $t \in \mathbb{N}$*,* $s'_0 \in \mathbb{R}^m$*. If* $\{ s_t \}_{t\in\mathbb{N}}$ *converges to* $S \in \mathbb{R}^m$*, then* $\{ s'_t \}_{t\in\mathbb{N}}$ *converges to* S *as well.*

Proof. By the definition of convergence, if $\{ s_t \}_{t\in\mathbb{N}}$ converges to S, we have: $\forall \epsilon > 0$, $\exists T \in \mathbb{N}$ such that $\forall t > T$, $|s_t - S| < \epsilon$. First, when $t > T$, by the formula of the sum of a finite geometric series, we rewrite S and s'_t as:

$$\begin{aligned} S &= (1-\alpha)\,\frac{1-\alpha^{t-T}}{1-\alpha}\,S + \alpha^{t-T}\,S \\ &= (1-\alpha)\sum_{i=T+1}^{t} \alpha^{t-i} S + \alpha^{t-T} S\,, \\ s'_t &= \alpha^t s'_0 + (1-\alpha)\sum_{i=1}^{t}\alpha^{t-i} s_i \\ &= \alpha^t s'_0 + (1-\alpha)\sum_{i=1}^{T}\alpha^{t-i} s_i + (1-\alpha)\sum_{i=T+1}^{t}\alpha^{t-i} s_i\,. \end{aligned} \tag{1}$$

Since T is finite, $\alpha^T s'_0$ and $\sum_{i=1}^{T} \alpha^{T-i} s_i$ are bounded. Thus, $\exists C \in \mathbb{R}^+$ such that:

$$|\alpha^T s'_0 + (1-\alpha)\sum_{i=1}^{T}\alpha^{T-i} s_i| < C\,.$$

Since $0 < \alpha < 1$, we have $\lim_{t\to\infty} \alpha^t = 0$. Thus, $\exists T' > 0$ such that $\forall t > T', \alpha^t < min\{ \frac{\epsilon}{C}, \frac{\epsilon}{|S|} \}$. Then, after substituting Eq. 1 into $|s'_t - S|$ and applying the Triangular Inequality, we have:

$$\begin{aligned} |s'_t - S| \le\; & |\alpha^t s'_0 + (1-\alpha)\sum_{i=1}^{T}\alpha^{t-i} s_i| \\ & + |(1-\alpha)\sum_{i=T+1}^{t}\alpha^{t-i}(s_i - S)| + |\alpha^{t-T} S|\,. \end{aligned} \tag{2}$$

Then $\forall t > (T + T')$, we have:

$$\begin{aligned} & |\alpha^t s'_0 + (1-\alpha)\sum_{i=1}^{T}\alpha^{t-i} s_i| \\ & = \alpha^{t-T}\,|\alpha^T s'_0 + (1-\alpha)\sum_{i=1}^{T}\alpha^{T-i} s_i| < \frac{\epsilon}{C}\,C < \epsilon\,, \end{aligned} \tag{3}$$

$$\begin{aligned} & |(1-\alpha)\sum_{i=T+1}^{t}\alpha^{t-i}(s_i - S)| \\ & \le (1-\alpha)\sum_{i=T+1}^{t}\alpha^{t-i}|s_i - S| = (1-\alpha^{t-T})\,\epsilon < \epsilon\,, \end{aligned} \tag{4}$$

$$|\alpha^{t-T} S| < \frac{\epsilon}{|S|}|S| < \epsilon\,. \tag{5}$$

Combining Eq. 2, 3, 4, 5, we have $|s'_t - S| < 3\epsilon$, $\forall t > (T + T')$, i.e., $\{s'_t\}_{y\in\mathbb{N}}$ converges to S. □

Appendix B: Model Architectures

The model architecture used in our CIFAR-10, CIFAR-100, and SVHN experiments is the 13-layer convolutional network (13-layer CNN), which is the same as previous works [6, 3, 1, 4, 5]. We implement it following FastSWA [1] for comparison. Table 1 describes its architecture in details. For ImageNet experiments, we use a 50-layer ResNeXt [7] architecture, which includes 3+4+6+3 residual blocks and uses the group convolution with 32 groups.

*kezhanghan@outlook.com

Table 1: The 13-layer CNN for our SSL experiments.

Layer	**Details**
input	$32 \times 32 \times 3$ RGB image
augmentation	random translation, horizontal flip
convolution	128, 3×3, pad = *same*, LReLU $\alpha = 0.1$
convolution	128, 3×3, pad = *same*, LReLU $\alpha = 0.1$
convolution	128, 3×3, pad = *same*, LReLU $\alpha = 0.1$
pooling	2×2, type = *maxpool*
dropout	$p = 0.5$
convolution	256, 3×3, pad = *same*, LReLU $\alpha = 0.1$
convolution	256, 3×3, pad = *same*, LReLU $\alpha = 0.1$
convolution	256, 3×3, pad = *same*, LReLU $\alpha = 0.1$
pooling	2×2, type = *maxpool*
dropout	$p = 0.5$
convolution	512, 3×3, pad = *valid*, LReLU $\alpha = 0.1$
convolution	256, 1×1, LReLU $\alpha = 0.1$
convolution	128, 1×1, LReLU $\alpha = 0.1$
pooling	$6 \times 6 \Rightarrow 1 \times 1$, type = *avgpool*
dense	$128 \Rightarrow 10$, softmax

Appendix C: Semi-supervised Learning Setups

In our work, all experiments use the SGD optimizer with the nesterov momentum set to 0.9. The learning rate is adjusted by the function $\gamma = \gamma_0 * (0.5 + \cos((t-1) * \pi / N))$, where t is the current training step, N is the total number of steps, and γ_0 is the initial learning rate. We present the settings of the experiments on each dataset as follows.

CIFAR-10: On CIFAR-10, we set the batch size to 100 and half of the samples in each batch are labeled. The initial learning rate is 0.1. The weight decay is $1e^{-4}$. For the stabilization constraint, we set its coefficient $\lambda_2 = 100$ and ramp it up in the first 5 epochs. We set $\lambda_1 = 10$. The confidence threshold for the *stable samples* is 0.8.

CIFAR-100: On CIFAR-100, each minibatch contains 128 samples, including 31 labeled samples. We set the initial learning rate to 0.2 and the weight decay to $2e^{-4}$. The confidence threshold is $\xi = 0.4$. Other hyperparameters are the same as CIFAR-10.

SVHN: The batch size on SVHN is 100, and each minibatch contains only 10 labeled samples. The initial learning rate is 0.1, and the weight decay is $1e^{-4}$. The stabilization constraint is scaled by 10 (ramp up in 5 epochs). We use the confidence threshold $\xi = 0.8$.

ImageNet: We validate our method on ImageNet by the ResNeXt-50 architecture on 8 GPUs with batch size 320 and half of the batch are labeled samples. Each sample is augmented following [2] and is resized to 224×224. We warm-up the learning rate from 0.08 to 0.2 in the first 2 epochs. The model is trained for 60 epochs with the weight decay set to $5e^{-5}$, the stabilization constraint coefficient set to 1000, and a small confidence threshold of 0.01.

Table 2: The small CNN for domain adaptation.

Layer	**Details**
input	$28 \times 28 \times 1$ Gray image
augmentation	gaussian noise $\zeta = 0.15$
convolution	16, 3×3, pad = *same*, LReLU $\alpha = 0.1$
pooling	2×2, type = *maxpool*
convolution	32, 3×3, pad = *same*, LReLU $\alpha = 0.1$
pooling	2×2, type = *maxpool*
dropout	$p = 0.5$
convolution	32, 3×3, pad = *same*, LReLU $\alpha = 0.1$
pooling	$6 \times 6 \Rightarrow 1 \times 1$, type = *avgpool*
dense	$32 \Rightarrow 10$, softmax

Appendix D: Domain Adaptation Setups

We design a small convolutional network for the domain adaptation from USPS (source domain) to MNIST (target domain). The structure is shown in Table 2. We train all experiments for 100 epochs by the SGD optimizer with the nesterov momentum set to 0.9 and the weight decay set to $1e^{-4}$. The learning rate declines from 0.1 to 0 by a cosine adjustment. Each batch includes 256 samples while 32 of them are labeled. We randomly extract 7000 balanced samples from MNIST for target-supervised experiments, and other experiments are done by using the training set of USPS. The coefficient of the stabilization constraint is $\lambda_2 = 1.0$. We also ramp it up in the first 5 epochs. The confidence threshold is $\xi = 0.6$. We discover that the input noise with $\zeta = 0.15$ is vital for the Mean Teacher but not for our method in this experiment.

References

[1] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *Proc ICLR*, 2019. 1

[2] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv:1709.01507*, 2017. 2

[3] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proc. ICLR*. 2017. 1

[4] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proc. CVPR*. 2018. 1

[5] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan L. Yuille. Deep co-training for semi-supervised image recognition. In *Proc. ECCV*. 2018. 1

[6] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NIPS*. 2017. 1

[7] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. CVPR*. 2017. 1