Mean teachers are better role models
(2017)

Antti Tarvainen, Harri Valpola
Notes

1 Introduction

In this work the authors explore semi supervised learning setting, where they use an exponentially weighted average of the models weight to obtain more stable prediction of the unlabeled data, that are then used as pseudo labels to calculate the unsupervised loss term and train the model using unlabeled data.

Given that the model must predict the same predictions even if the inputs did change slightly by adding some noise, by comparing the prediction with two different views (with regularization, noise or data augmentation) of the input we can enable the model to learn more abstract invariances, the noise may be added to intermediate representations, by some regularization technic such as Dropout. So in semi supervised learning, rather than minimizing the classification cost at the zero-dimensional data points of the input space like the supervised setting, the regularized model minimizes the cost on a manifold around each data point, thus pushing decision boundaries away from the labeled data points, this is illustrated in the following figure:

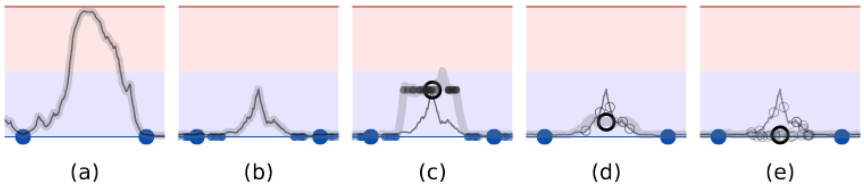

Figure 1: A sketch of a binary classification task with two labeled examples (large blue dots) and one unlabeled example, demonstrating how the choice of the unlabeled target (black circle) affects the fitted function (gray curve). **(a)** A model with no regularization is free to fit any function that predicts the labeled training examples well. **(b)** A model trained with noisy labeled data (small dots) learns to give consistent predictions around labeled data points. **(c)** Consistency to noise around unlabeled examples provides additional smoothing. For the clarity of illustration, the teacher model (gray curve) is first fitted to the labeled examples, and then left unchanged during the training of the student model. Also for clarity, we will omit the small dots in figures d and e. **(d)** Noise on the teacher model reduces the bias of the targets without additional training. The expected direction of stochastic gradient descent is towards the mean (large blue circle) of individual noisy targets (small blue circles). **(e)** An ensemble of models gives an even better expected target. Both Temporal Ensembling and the Mean Teacher method use this approach.

2 Mean Teacher

The authors argue that to improve the target quality, we can either choose the perturbation very carefully or use a better teacher model instead of using the same as the student model (like in temporal labeling), so with a weighted average of the models weights we can form a better teacher model without any additionnal training, and even without storing any eponentially weighted average of the prediction like temporal labeling that can be very expensive memory wise with large datasets, so this approch reduces the training time and scales to larger dataset linearly.

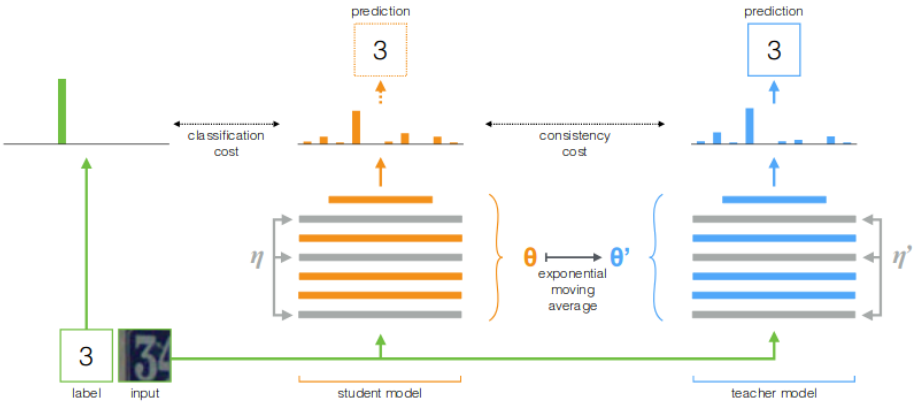

Figure 2: The Mean Teacher method. The figure depicts a training batch with a single labeled example. Both the student and the teacher model evaluate the input applying noise (η, η') within their computation. The softmax output of the student model is compared with the one-hot label using classification cost and with the teacher output using consistency cost. After the weights of the student model have been updated with gradient descent, the teacher model weights are updated as an exponential moving average of the student weights. Both model outputs can be used for prediction, but at the end of the training the teacher prediction is more likely to be correct. A training step with an unlabeled example would be similar, except no classification cost would be applied.

By having an average of different (student) models instead of averaging the predictions, we can produce more accurate model and more stable labels to train the student model, so at the end of each epoch, we update the teacher models weights using the current weights of the student model (there is no parameter sharing between the two):

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t$$

And at each training iteration, we reduce the cost funtion between the predictions of the teacher model and the student model:

$$J(\theta) = \mathbb{E}_{x,\eta',\eta}\left[\|f\left(x,\theta',\eta'\right) - f(x,\theta,\eta)\|^2\right]$$

3 Experiments

Table 1: Error rate percentage on SVHN over 10 runs (4 runs when using all labels). We use exponential moving average weights in the evaluation of all our models. All the methods use a similar 13-layer ConvNet architecture. See Table 5 in the Appendix for results without input augmentation.

	250 labels 73257 images	500 labels 73257 images	1000 labels 73257 images	73257 labels 73257 images
GAN [25]		18.44 ± 4.8	8.11 ± 1.3	
Π model [13]		6.65 ± 0.53	4.82 ± 0.17	2.54 ± 0.04
Temporal Ensembling [13]		5.12 ± 0.13	4.42 ± 0.16	2.74 ± 0.06
VAT+EntMin [16]			**3.86**	
Supervised-only	27.77 ± 3.18	16.88 ± 1.30	12.32 ± 0.95	2.75 ± 0.10
Π model	9.69 ± 0.92	6.83 ± 0.66	4.95 ± 0.26	2.50 ± 0.07
Mean Teacher	$\mathbf{4.35 \pm 0.50}$	$\mathbf{4.18 \pm 0.27}$	3.95 ± 0.19	$\mathbf{2.50 \pm 0.05}$

Table 2: Error rate percentage on CIFAR-10 over 10 runs (4 runs when using all labels).

	1000 labels 50000 images	2000 labels 50000 images	4000 labels 50000 images	50000 labels 50000 images
GAN [25]			18.63 ± 2.32	
Π model [13]			12.36 ± 0.31	5.56 ± 0.10
Temporal Ensembling [13]			12.16 ± 0.31	$\mathbf{5.60 \pm 0.10}$
VAT+EntMin [16]			**10.55**	
Supervised-only	46.43 ± 1.21	33.94 ± 0.73	20.66 ± 0.57	5.82 ± 0.15
Π model	27.36 ± 1.20	18.02 ± 0.60	13.20 ± 0.27	6.06 ± 0.11
Mean Teacher	$\mathbf{21.55 \pm 1.48}$	$\mathbf{15.73 \pm 0.31}$	12.31 ± 0.28	5.94 ± 0.15

Table 3: Error percentage over 10 runs on SVHN with extra unlabeled training data.

	500 labels 73257 images	500 labels 173257 images	500 labels 573257 images
Π model (ours)	6.83 ± 0.66	4.49 ± 0.27	3.26 ± 0.14
Mean Teacher	$\mathbf{4.18 \pm 0.27}$	$\mathbf{3.02 \pm 0.16}$	$\mathbf{2.46 \pm 0.06}$

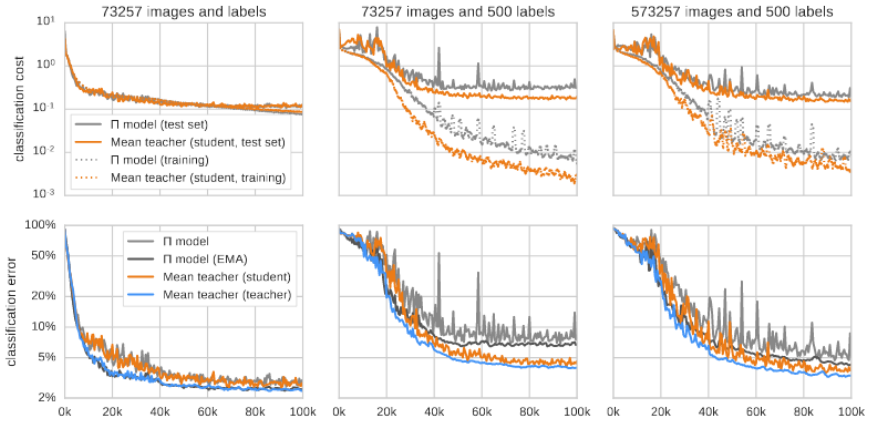

Figure 3: Smoothened classification cost (top) and classification error (bottom) of Mean Teacher and our baseline Π model on SVHN over the first 100000 training steps. In the upper row, the training classification costs are measured using only labeled data.