NC STATE UNIVERSITY

Local Clustering with Mean Teacher for Semi-supervised learning

Authors: Zexi Chen, Benjamin Button, Bharathkumar Ramachandra, Tianfu Wu, Ranga Raju Vatsavai

Introduction to Consistency-based approaches

Consistency-based approaches for semi-supervised learning (SSL)

- Encourage consistent probability predictions between a teacher-student pair for the same data under perturbations

Loss term: $\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{cons}$

$$\mathcal{L}_{cons} = \mathop{\mathbb{E}}_{\{x_i\}_{i=1}^{n_l+n_u}} D[f(x_i, \xi'; \theta'), f(x_i, \xi''; \theta'')]$$

Random perturbation

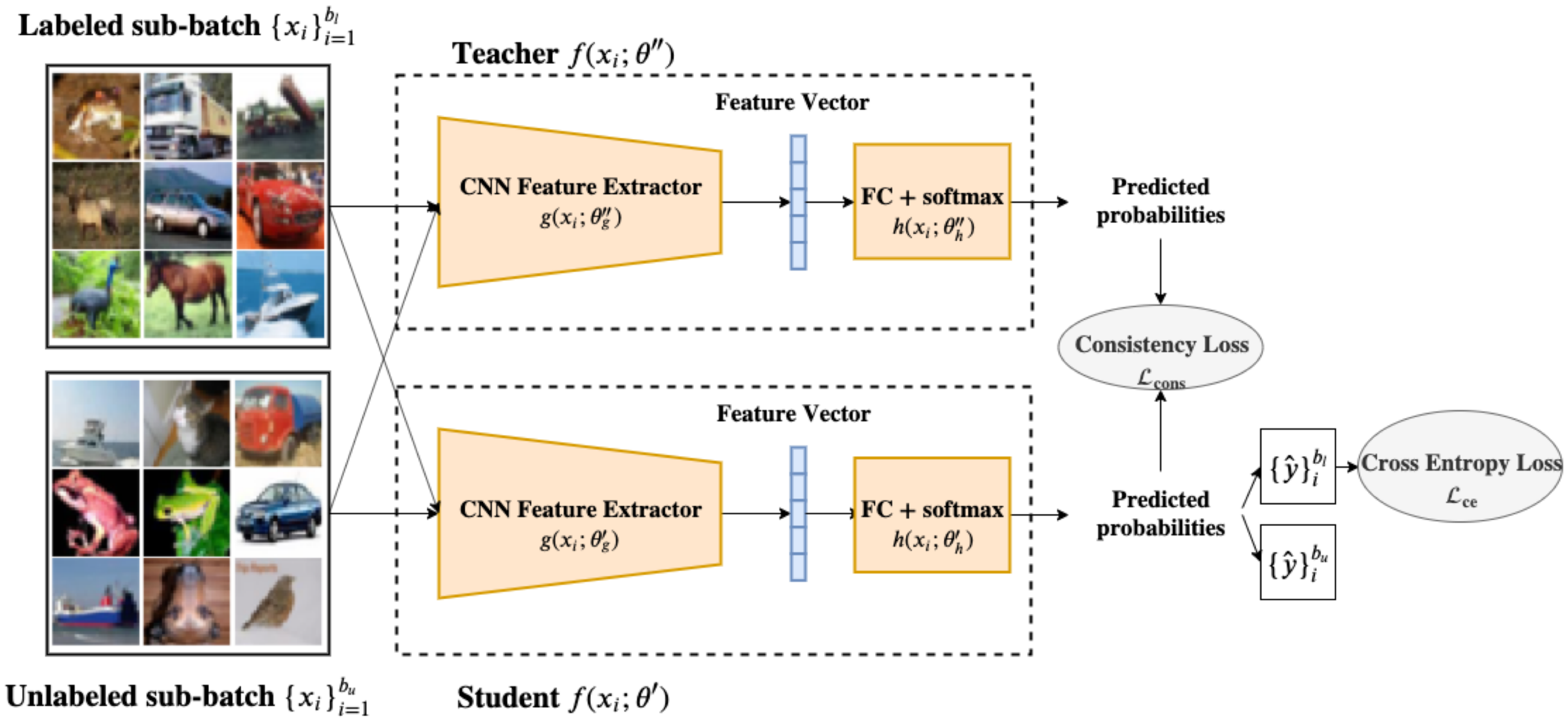


Figure 1. General framework for consistency-based approaches

Research Motivation

Issue with consistency-based methods: Confirmation bias

- Caused by inaccurate learning targets generated by the teacher model
- It would trap some unlabeled data samples in low-density regions or enforce them into high-density regions of incorrect class in feature space.

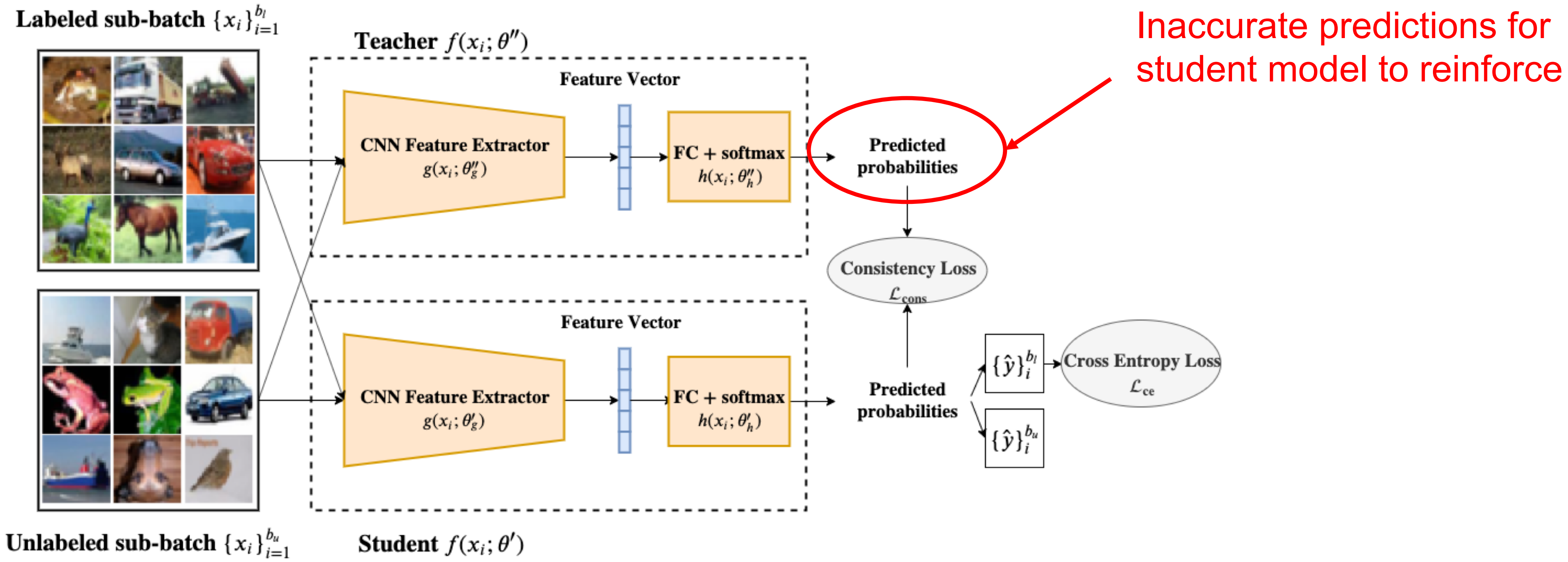


Figure 2. General framework for consistency-based approaches

Method-Idea illustration

Clustering assumption: Samples are likely to have the same class label if there is a path connecting them passing through regions of high density only.
Local consistency assumption: Nearby samples are likely to have the same label. Samples on the same structure (typically, a manifold) are likely to have the same label.

Propose a local clustering method that clusters data points locally by minimizing the pairwise distance between neighboring points in feature space

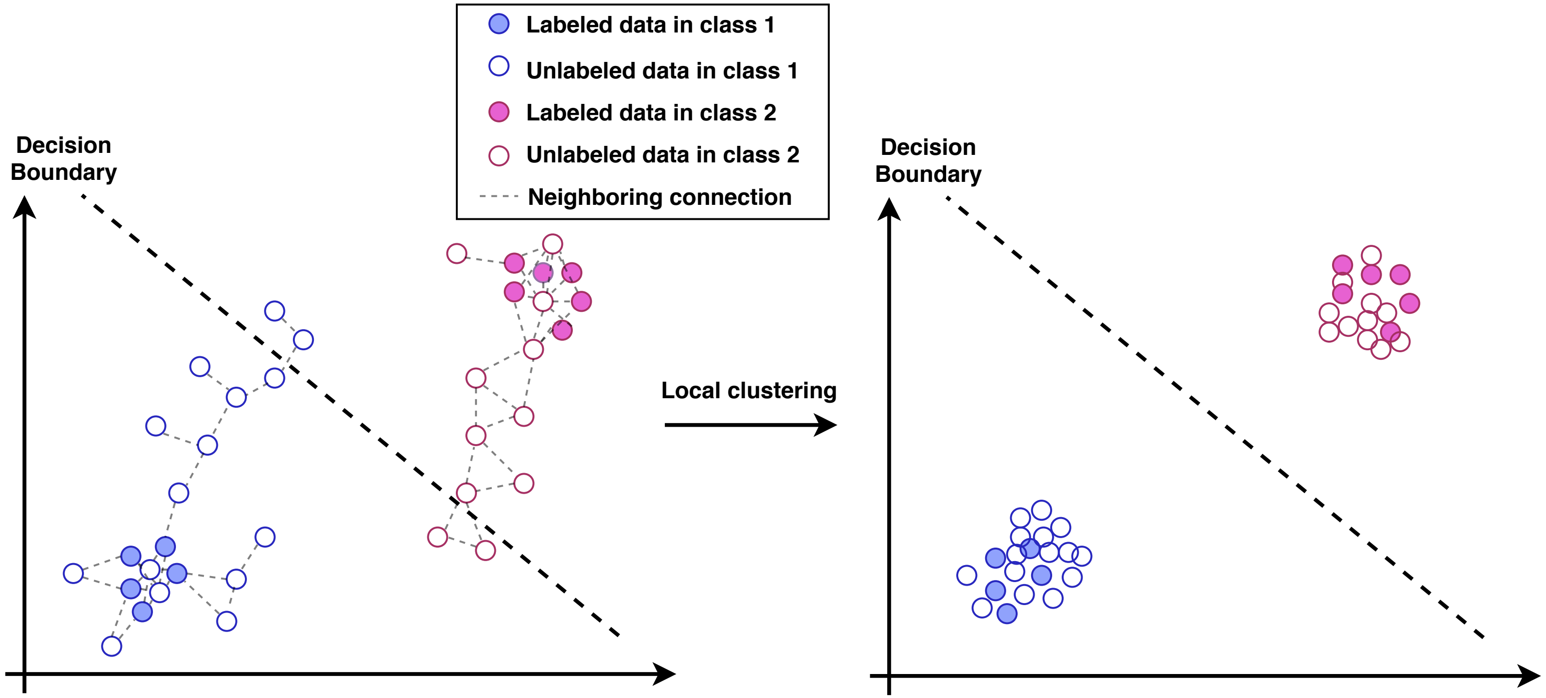


Figure 3. An illustration of the intuition behind Local clustering in feature space. Each point represents the intermediate learned representation of one data sample

Related Work

Several types of consistency-based methods:

Π model[2]

$$\mathcal{L}_{cons} = \mathop{\mathbb{E}}_{\{x_i\}_{i=1}^{n_l+n_u}} D[f(x_i, \xi'; \theta'), f(x_i, \xi''; \theta'')]$$

$$\theta' = \theta''$$

Temporal Ensembling[2]

- The EMA of probability predictions of the student model as the teacher model's predictions

Virtual Adversarial Training (VAT)[3]

- Impose adversarial perturbations to either the inputs or intermediate feature vectors that would maximize the difference in predictions between the student model and teacher model.

Mean Teacher[1]

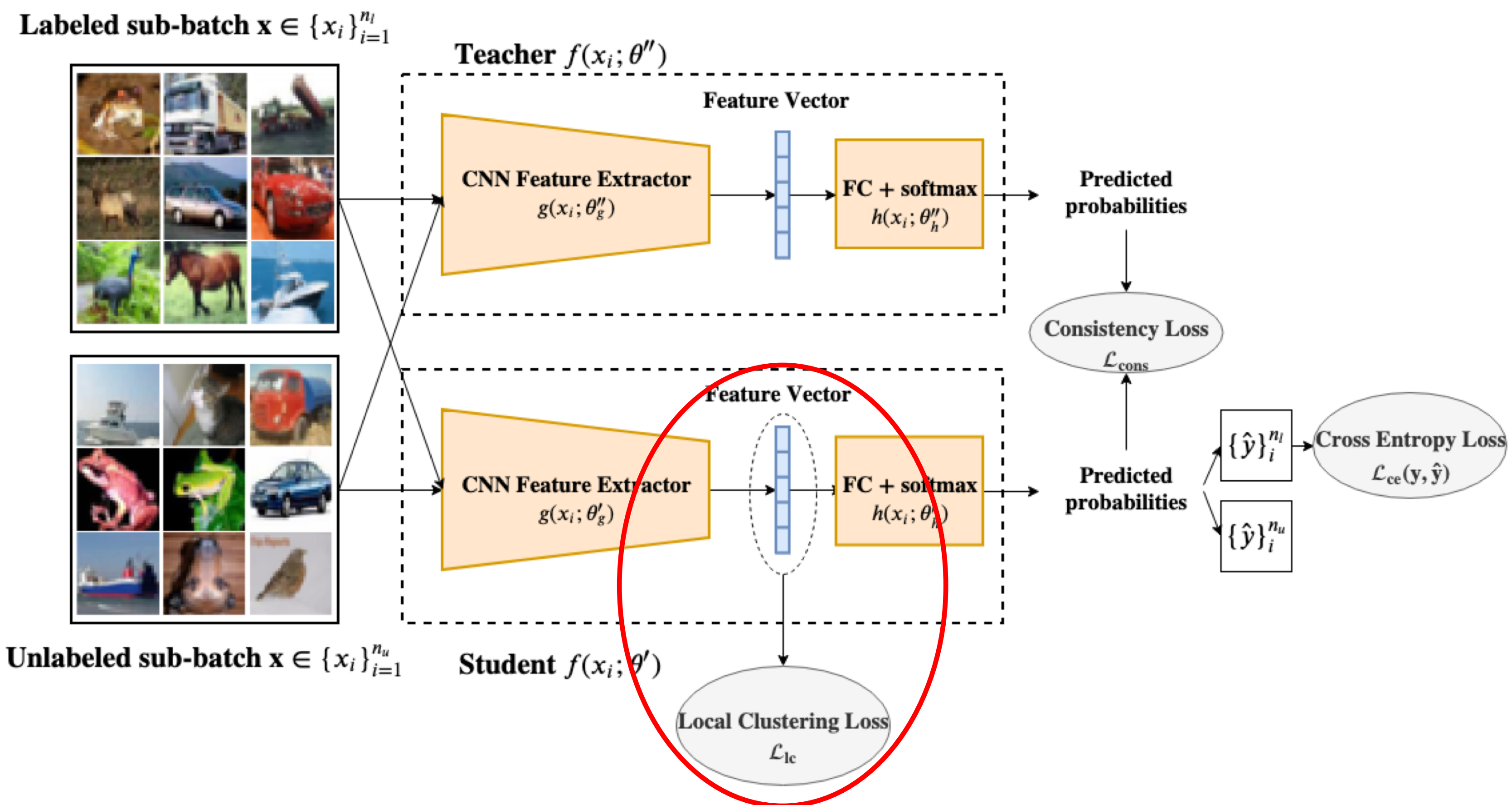
$$\theta''_t = \alpha\theta'_{t-1} + (1-\alpha)\theta'_t$$

Note: the citation-style references are presented in the end of the slides

Method

Local clustering:

- Build on top of Mean Teacher
- Idea is to pull those misclassified unlabeled data to the high-density regions of their correct class in feature space
- Can be treated as a new regularizer to Mean Teacher



Method

Local clustering:

- At each training iteration, build a weighted graph in feature space from a sub-batch of labeled data and a sub-batch of unlabeled data

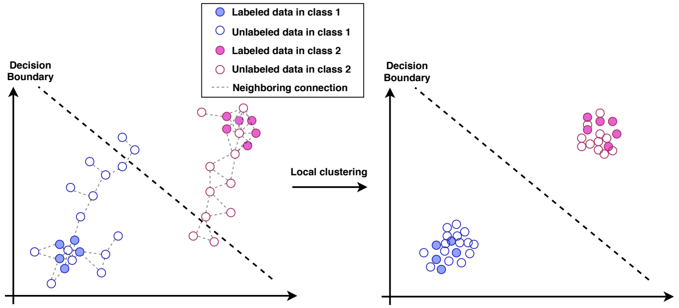
The local clustering loss:

$$\mathcal{L}_{lc} = \mathop{\mathbb{E}}_{\{x_i\}_{i=1}^{n_l}, \{x_j\}_{j=1}^{n_u}} [w_{ij}||g(x_i;\theta'_g) - g(x_j;\theta'_g)||^2] + \mathop{\mathbb{E}}_{\{x_m\}_{m=1}^{n_u}, \{x_n\}_{n=1}^{n_u}} [w_{mn}||g(x_m;\theta'_g) - g(x_n;\theta'_g)||^2]$$

The edge weight function

$$w_{ij} = \begin{cases} exp\left(-\frac{||z_i - z_j||^2}{\epsilon}\right) & \text{if } ||z_i - z_j||^2 \leq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

Total loss: $$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{cons} + \lambda_2 \mathcal{L}_{lc}$$



Results – Datasets

We conduct experiments on two widely used semi-supervised image classification benchmark datasets: SVHN and CIFAR-10

Dataset	Number of training images	Number of test images	Number of action classes
SVHN	73,257	26,032	10
CIFAR-10	50,000	10,000	10

Table 8: Video action recognition datasets used for this task

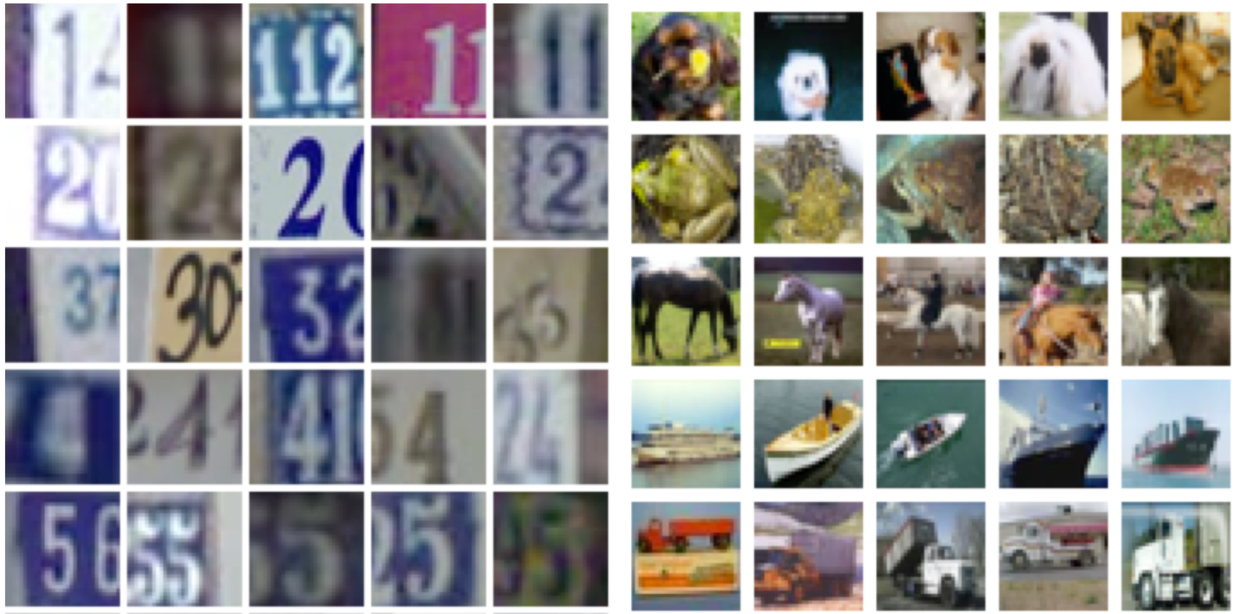


Figure 4. Sample images in SVHN (left) and CIFAR-10(right)

Results

Compares with state-of-the-art methods on SVHN and CIFAR-10

- Train the models on SVHN training images with 500 and 1,000 randomly labeled samples
- Train the models on CIFAR-10 training images with 2,000 and 4,000 randomly labeled samples
- Test error rate percentage is reported as evaluation metric

Method	SVHN		CIFAR-10	
	$n_l = 500$	$n_l = 1000$	$n_l = 2,000$	$n_l = 4,000$
FM-GAN (Salimans et al. 2016)	18.44 ± 4.80	8.11 ± 1.30	19.61 ± 2.09	18.63 ± 2.32
Bad GAN (Dai et al. 2017)	-	7.42 ± 0.65	-	14.41 ± 0.30
Local GAN (Qi et al. 2018)	5.48 ± 0.29	4.73 ± 0.29	-	14.23 ± 0.27
Π model (Laine and Aila 2016)	6.65 ± 0.53	4.82 ± 0.17	-	12.36 ± 0.31
TempEns (Laine and Aila 2016)	5.12 ± 0.13	4.42 ± 0.16	-	12.16 ± 0.31
Mean Teacher (Tarvainen and Valpola 2017)	4.18 ± 0.27	3.95 ± 0.19	15.73 ± 0.31	12.31 ± 0.28
VAdD (Park et al. 2018)	-	4.16 ± 0.08	-	11.32 ± 0.11
VAT + EntMin (Miyato et al. 2018)	-	3.86 ± 0.11	-	10.55 ± 0.05
TempEns + SNTG (Luo et al. 2018)	4.46 ± 0.26	3.98 ± 0.21	13.64 ± 0.32	10.93 ± 0.14
MT + SNTG (Luo et al. 2018)	3.99 ± 0.24	3.86 ± 0.27	-	-
MT*	3.91 ± 0.11	3.80 ± 0.09	12.37 ± 0.29	9.93 ± 0.16
MT + LC **(ours)**	$\mathbf{3.54} \pm \mathbf{0.17}$	$\mathbf{3.35} \pm \mathbf{0.09}$	$\mathbf{11.56} \pm \mathbf{0.31}$	$\mathbf{9.26} \pm \mathbf{0.16}$

Table 1: Error rate percentage comparison with the sota methods on SVHN and CIFAR-10

Results

Visualize the test error as a function of training epoch on SVHN and CIFAR-10

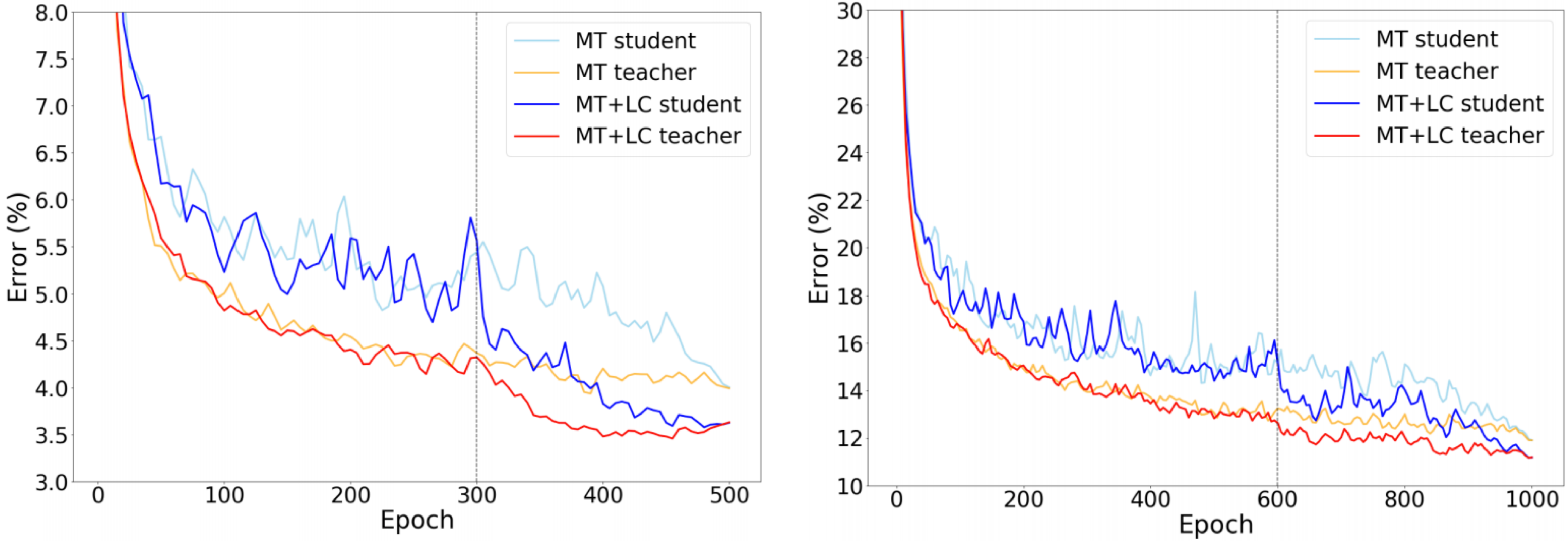
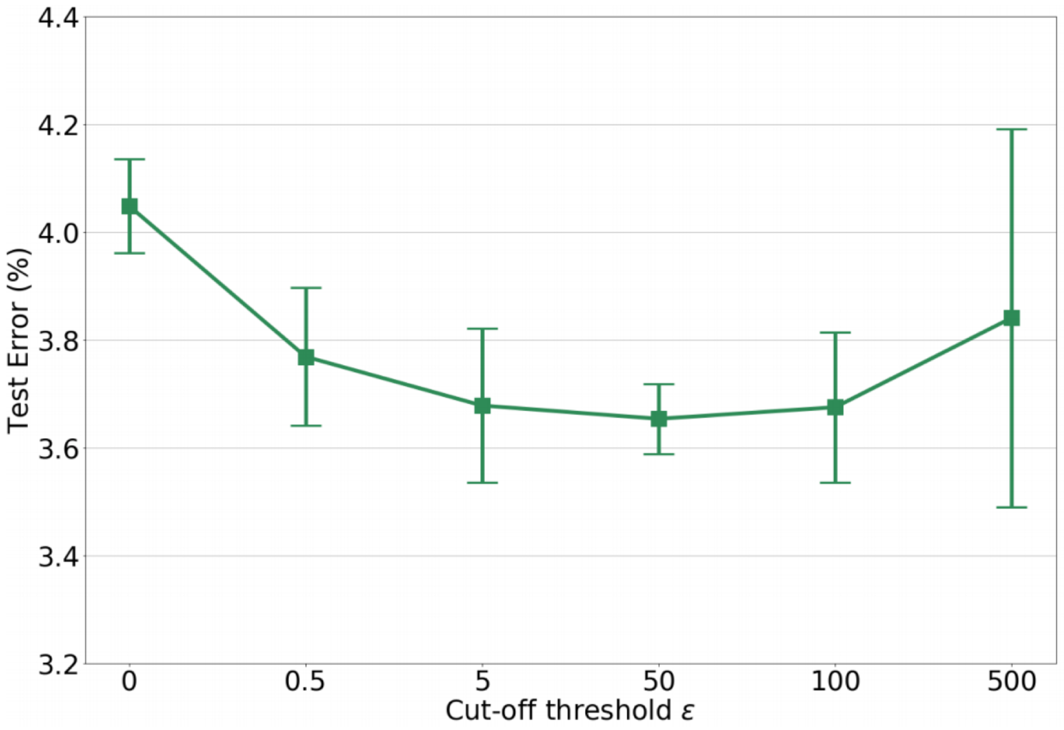


Figure 5. Smoothed test error curves of MT and MT+LC on SVHN (left) and CIFAR-10 (right)

Results

Ablation Study:

- Study the effect of **cut-off threshold ϵ**
- Study the effect of **LC loss weight λ_2**



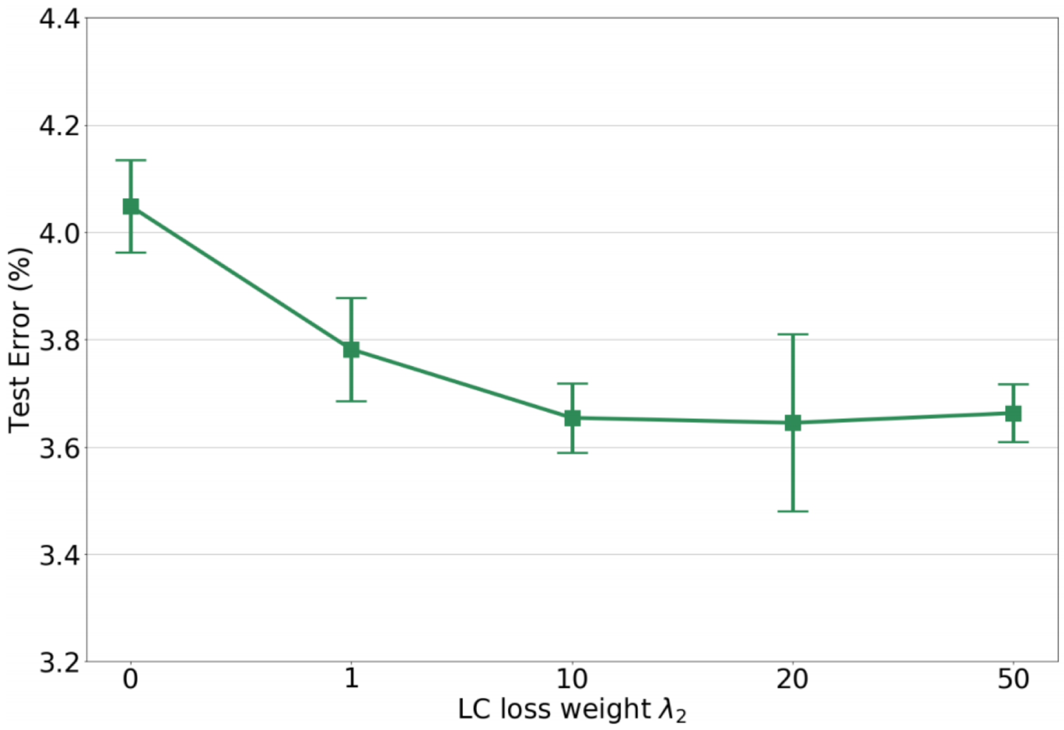


Figure 6. The test errors of MT + LC with different cut-off thresholds (left) and loss weights(right)

Visualization

Visual comparison of MT and MT + LC on intermediate feature representations on test data

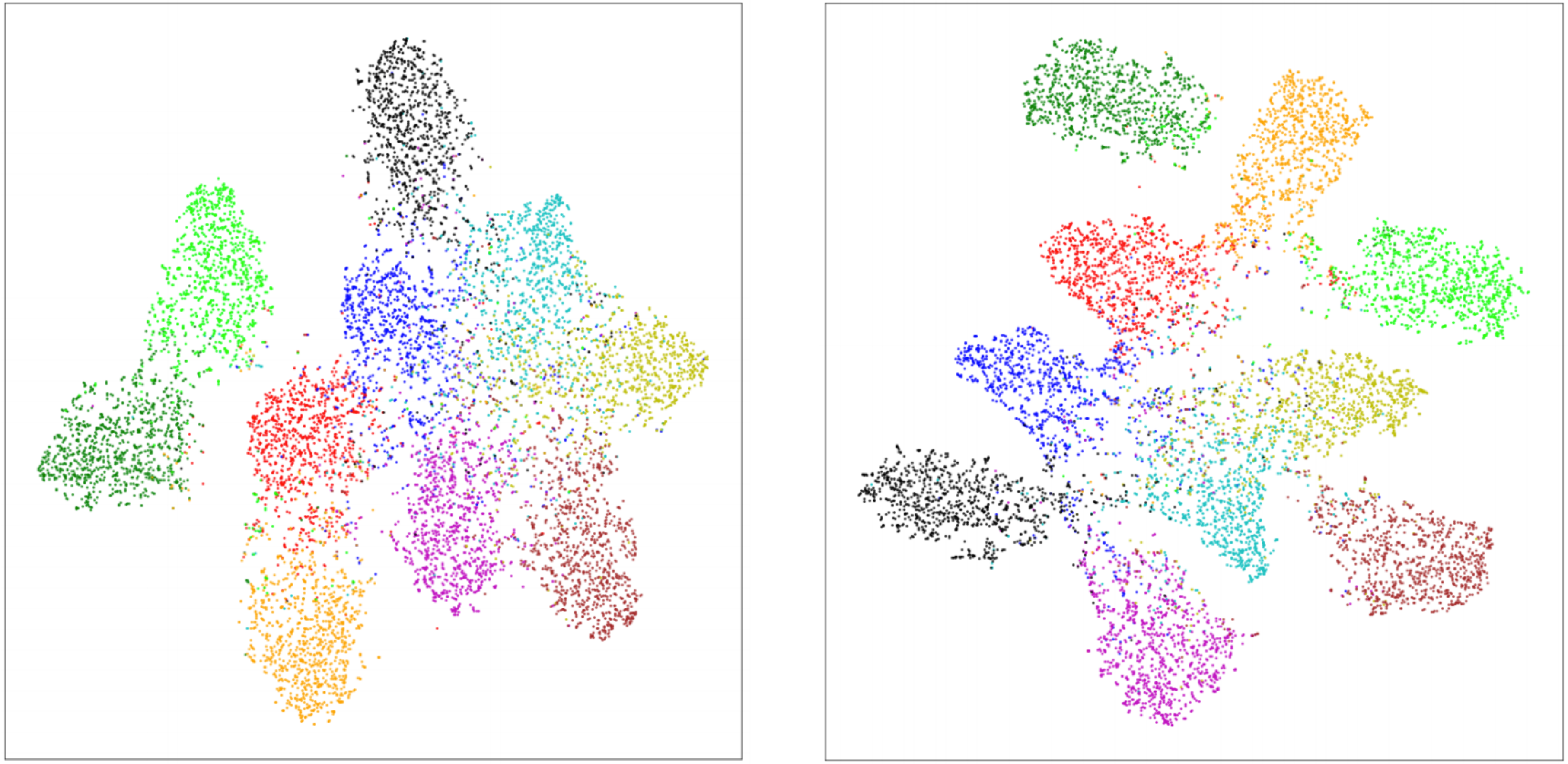


Figure 7. t-SNE visualization of CIFAR-10 test data features obtained by MT (left) and MT + LC (right)

NC STATE UNIVERSITY

Summary & future work

- We developed a novel local clustering method to address the confirmation bias issues existing in Mean Teacher method
- Future work:
  - Design adaptive cut-off threshold
  - Use a pair-wise metric learning method to determine the pairwise similarity

NC STATE UNIVERSITY

References for slide 5

[1] Tarvainen, Antti, and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results." *Advances in neural information processing systems*. 2017

[2] Laine, Samuli, and Timo Aila. "Temporal ensembling for semi-supervised learning." *arXiv preprint arXiv:1610.02242* (2016).

[3] Miyato, Takeru, et al. "Virtual adversarial training: a regularization method for supervised and semi-supervised learning." *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018): 1979-1993.

Thank You!