How does Disagreement Help Generalization against Label Corruption?

Xingrui Yu[1] Bo Han[2] Jiangchao Yao[3] Gang Niu[2]
Ivor W. Tsang[1] Masashi Sugiyama[2,4]

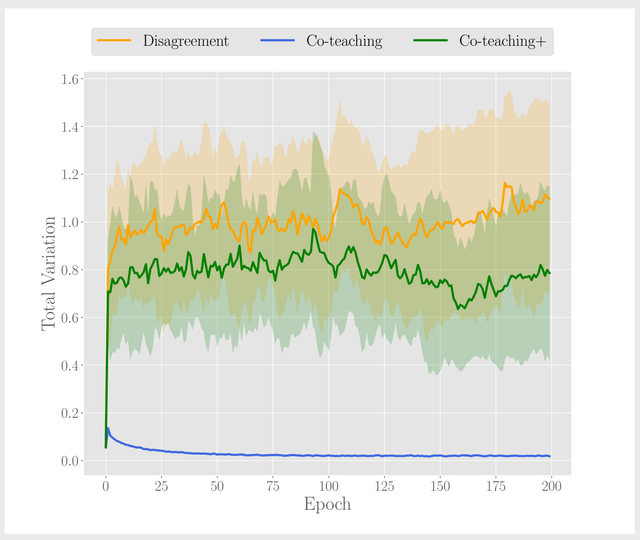
[1]CAI, University of Technology Sydney [2]AIP, RIKEN [3]Alibaba Damo Academy [4]The University of Tokyo

Overview

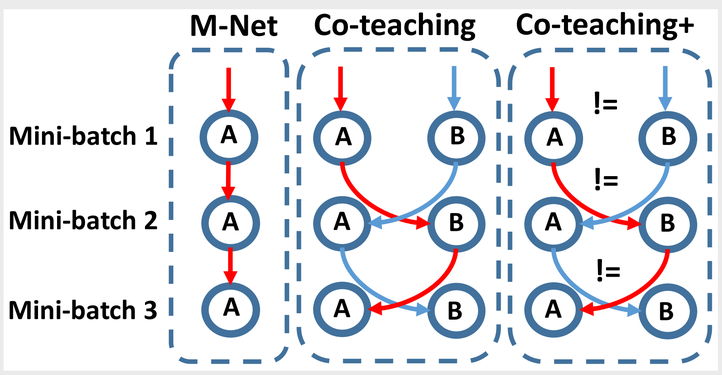
TL;DR: Two networks feed forward and predict all data, but keep prediction **disagreement data** only. Among such disagreement data, each network selects its **small-loss data**, but back propagates the small-loss data from its **peer network** and updates its own parameters.

- **Noisy labels** are corrupted from ground-truth labels, which degenerates the robustness of learning models.
- **Deep neural networks** have the high capacity to fit any noisy labels. The solutions are as follows.
  - ◇ Noise **transition** matrix estimation. E.g., F-correction.
  - ◇ **Regularization**. E.g., VAT and Mean teacher.
  - ◇ Training on **selected** samples. E.g., MentorNet.
- We present a new paradigm called **Co-teaching+** combating with noisy labels.
  - ◇ We train **two** networks simultaneously, where they feed forward and predict all data, but keep prediction **disagreement data** only.
  - ◇ In each mini-batch **disagreement** data, each network **filters** noisy instances based on memorization effects.
  - ◇ It teaches the **remaining** instances to its **peer** network for updating the parameters.

Motivation



Beyond Co-teaching



Illustrative Example

- Peer learning will be better than solo learning.
- The optimal peer should be complementary: Student good at math should review with another good at literature.

QR Code

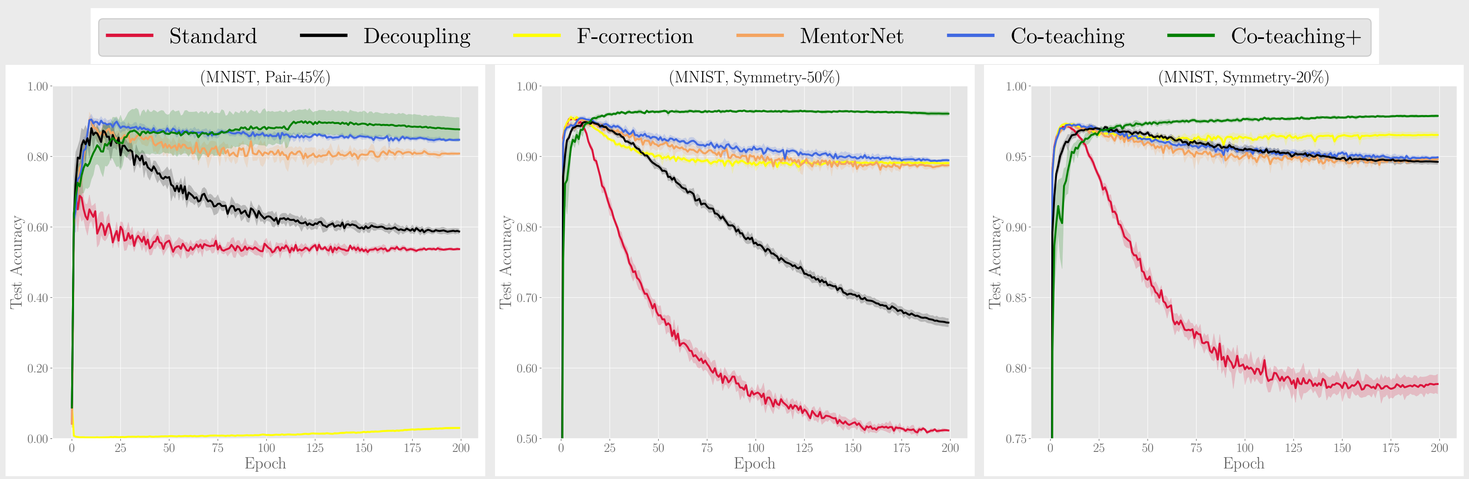
Key Factors of Co-teaching+

- **"small loss":** regarding small-loss samples as "clean" samples;
- **"double classifiers":** training two classifiers simultaneously;
- **"cross update":** updating parameters in a cross manner instead of a parallel manner;
- **"divergence":** keeping two classifiers diverged during the whole training epochs.

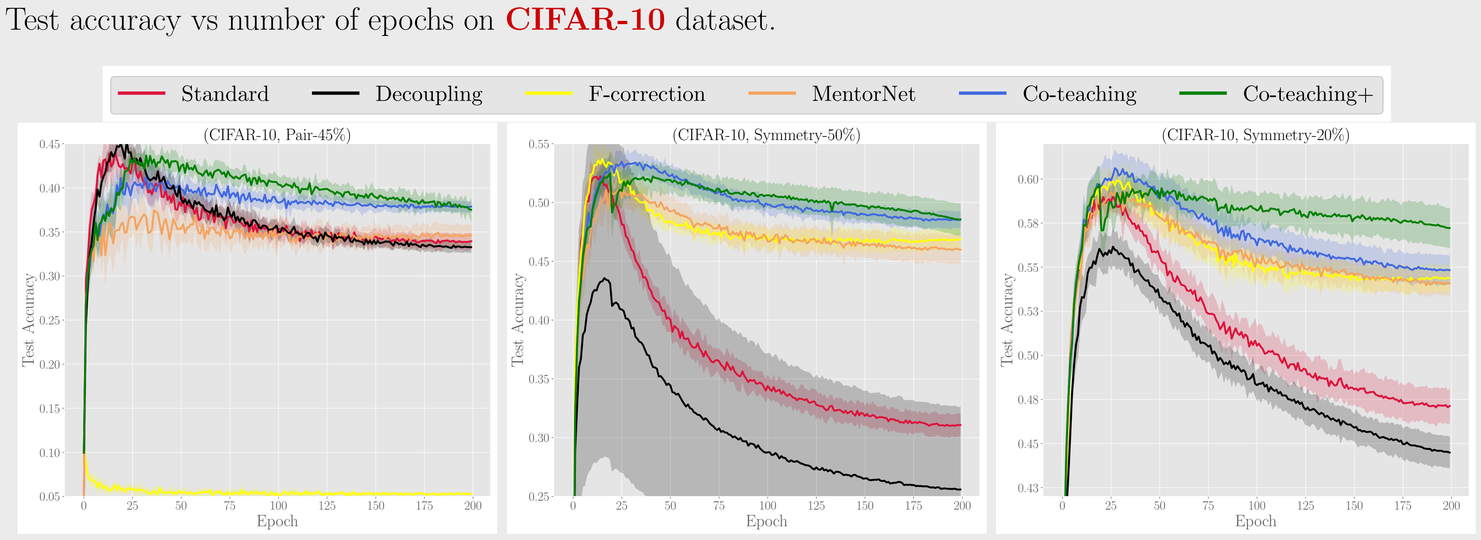
	MentorNet	Co-training	Co-teaching	Decoupling	Co-teaching+
small loss	✓	×	✓	×	✓
double classifiers	×	✓	✓	✓	✓
cross update	×	✓	✓	×	✓
divergence	×	✓	×	✓	✓

Results on Simulated Noisy Datasets

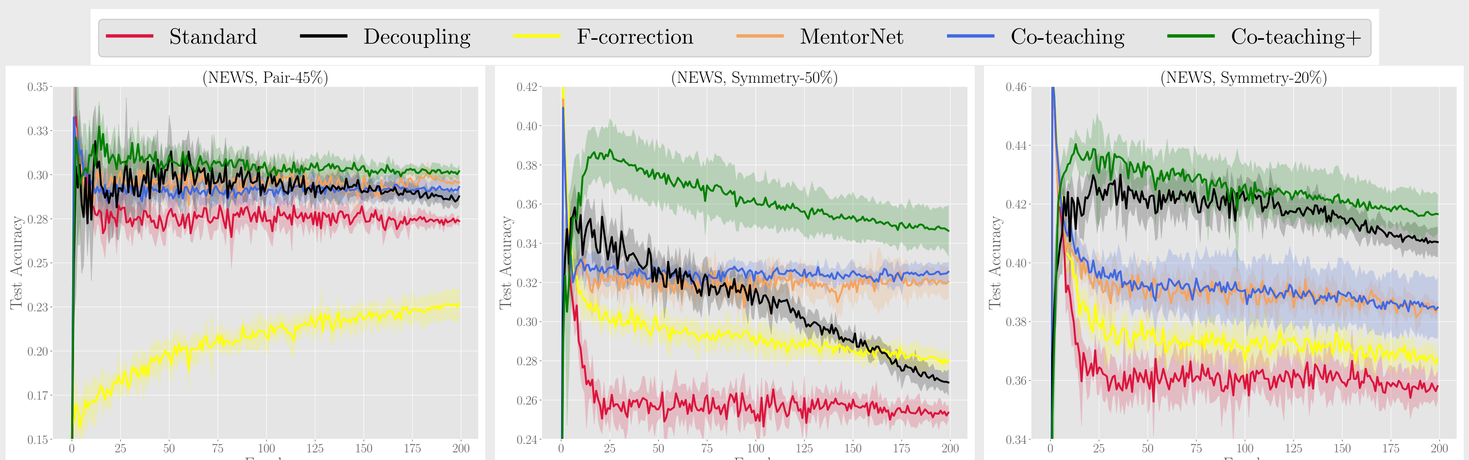
- Test accuracy vs number of epochs on **MNIST** dataset.



- Test accuracy vs number of epochs on **CIFAR-10** dataset.



- Test accuracy vs number of epochs on **NEWS** dataset.



Results on Real-world Noisy Datasets

- Averaged/maximal test accuracy (%) of different approaches on **T-ImageNet** over last 10 epochs.

Flipping-Rate(%)	Standard	Decoupling	F-correction	MentorNet	Co-teaching	Co-teaching+
Pair-45%	26.14/26.32	26.10/26.61	0.63/0.67	26.22/26.61	27.41/**27.82**	26.54/26.87
Symmetry-50%	19.58/19.77	22.61/22.81	32.84/33.12	35.47/35.76	37.09/37.60	41.19/**41.77**
Symmetry-20%	35.56/35.80	36.28/36.97	44.37/44.50	45.49/45.74	45.60/46.36	47.73/**48.20**

- Averaged/maximal test accuracy (%) of different approaches on **Open-sets** over last 10 epochs.

Open-set noise	Standard	MentorNet	Iterative	Co-teaching	Co-teaching+
CIFAR-10+CIFAR-100	62.92	79.27/79.33	79.28	79.43/79.58	79.28/**79.74**
CIFAR-10+ImageNet-32	58.63	79.27/79.40	79.38	79.42/79.60	79.89/**80.52**
CIFAR-10+SVHN	56.44	79.72/79.81	77.73	80.12/80.33	80.62/**80.95**