How does Disagreement Help Generalization against Label Corruption?

Center for Advanced Intelligence Project, RIKEN, Japan
Centre for Artificial Intelligence, University of Technology Sydney, Australia

Jun 12th, 2019

Outline

Big and high quality data drives the success of deep models.

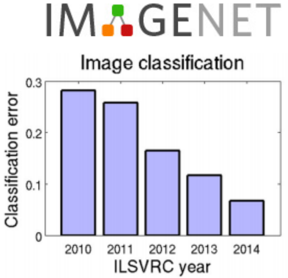
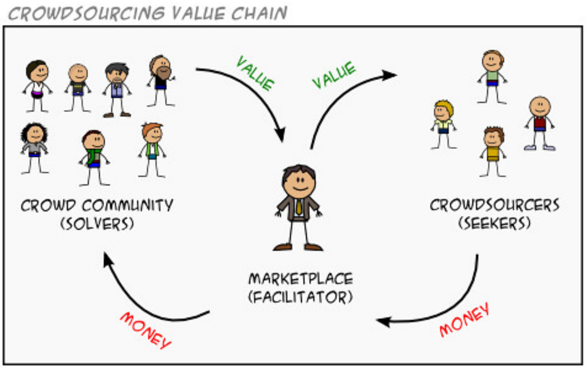


Figure: There is a steady reduction of error every year in object classification on large scale dataset (1000 object categories, 1.2 million training images) [Russakovsky et al., 2015].

- However, what we usually have in practice is big data with noisy labels.

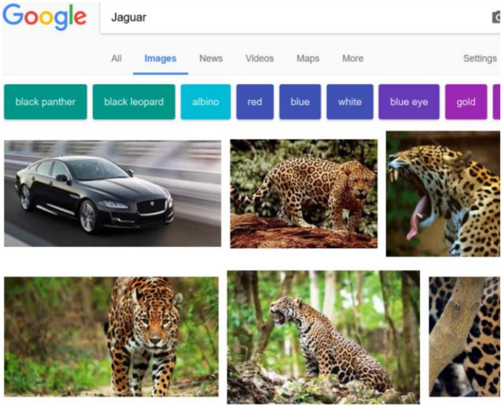
Noisy labels from crowdsourcing platforms.



Credit: *Torbjørn Marø*

- Unreliable labels may occur when the workers have limited domain knowledge.

Noisy labels from web search/crawler.



Screenshot of Google.com

- The keywords may not be relevant to the image contents.

How to model noisy labels?

- Class-conditional noise (CCN):
  Each label y in the training set (with c classes) is flipped into $\tilde{y}$ with probability $p(\tilde{y}|y)$. Denote by $T \in [0,1]^{(c\times c)}$ the noise transition matrix specifying the probability of flipping one label to another, so that $\forall_{i,j} T_{ij} = p(\tilde{y}=j|y=i)$.

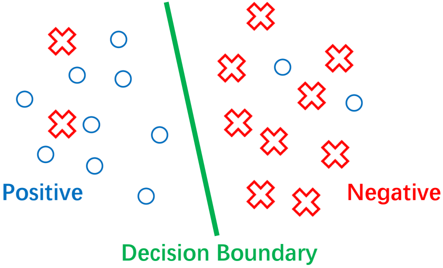


Figure: Illustration of noisy labels.

What happens when learning with noisy labels?

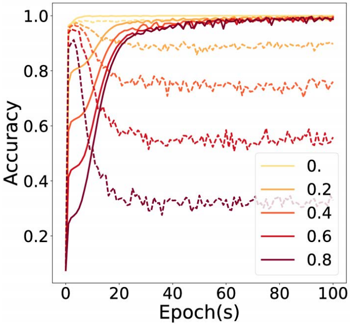


Figure: Accuracy of neural networks on noisy MNIST with different noise rate (0., 0.2, 0.4, 0.6, 0.8). (Solid is train, dotted is validation.) [Arpit et al., 2017]

Memorization: Learning easy patterns first, then (totally) over-fit noisy training data.

Effect: Training deep neural networks directly on noisy labels results in accuracy degradation.

How can wen robustly learn from noisy labels?

Current progress in three orthogonal directions:

- Learning with noise transition:
  Forward Correction (Australian National University, CVPR'17)
  S-adaptation (Bar Ilan University, ICLR'17)
  Masking (RIKEN-AIP/UTS, NeurIPS'18)
- Learning with selected samples:
  MentorNet (Google AI, ICML'18)
  Learning to Reweight Examples (University of Toronto, ICML'18)
  Co-teaching (RIKEN-AIP/UTS, NeurIPS'18)
- Learning with implicit regularization:
  Virtual Adversarial Training (Preferred Networks, ICLR'16)
  Mean Teachers (Curious AI, NIPS'17)
  Temporal Ensembling (NVIDIA, ICLR'17)

A promising research line: Learning with small-loss instances

- Main idea: regard small-loss instances as “correct” instances.

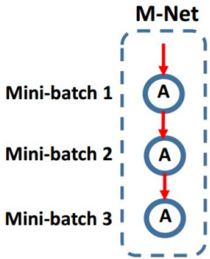


Figure: Self-training MentorNet[Jiang et al., 2018].

- Benefit: easy to implement & free of assumptions.
- Drawback: accumulated error caused by sample-selection bias.

A promising research line: Learning with small-loss instances

Consider the standard class-conditional noise (CCN) model.

- We can learn a reliable classifier if a set of clean data is available.
- Then, we can use the reliable classifier to filter out the noisy data, where "small loss" serves as a gold standard.
- However, we usually only have access to noisy training data. The selected small-loss instances are only likely to be correct, instead of totally correct.
- (Problem) There exists accumulated error caused by sample-selection bias.
- (Solution 1) In order to select more correct samples, can we design a "small-loss" rule by utilizing the memorization of deep neural networks?

Related work: Decoupling

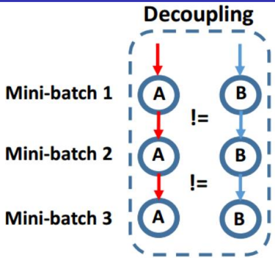


Figure: Decoupling[Malach and Shalev-Shwartz, 2017].

- Easy samples can be quickly learnt and classified (memorization effect).
- Decoupling focus on hard samples, which can be more informative.
- Decoupling use samples in each mini-batch that two classifiers have disagreement in predictions to update networks.
- (Solution 2) Can we further attenuate the error from noisy data by utilizing two networks?

Co-teaching: Cross-update meets small-loss

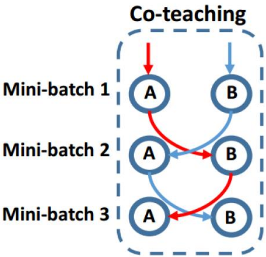


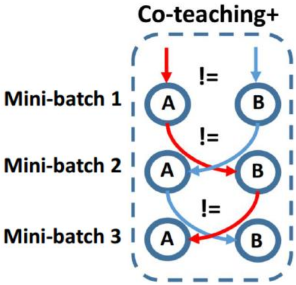
Figure: Co-teaching[Han et al., 2018].

- Co-teaching maintains two networks (A & B) simultaneously.
- Each network samples its small-loss instances based on memorization of neural networks.
- Each network teaches such useful instances to its peer network. (Cross-update)

Divergence

- Two networks in Co-teaching will converge to a consensus gradually.
- However, two networks in Disagreement will keep diverged.
- We bridge the "Disagreement" strategy with Co-teaching to achieve Co-teaching+.

How does Disagreement Benefit Co-teaching?



- Disagreement-update step: Two networks feed forward and predict all data first, and only keep prediction disagreement data.
- Cross-update step: Based on disagreement data, each network selects its small-loss data, but back propagates such data from its peer network.

Co-teaching+ Paradigm

1: Input $w^{(1)}$ and $w^{(2)}$, training set $\mathcal{D}$, batch size B, learning rate η, estimated noise rate τ, epoch E_k and $E_{\max}$;

for $e = 1, 2, \ldots, E_{\max}$ do

  2: Shuffle $\mathcal{D}$ into $\frac{|\mathcal{D}|}{B}$ mini-batches; //noisy dataset

  for $n = 1, \ldots, \frac{|\mathcal{D}|}{B}$ do

    3: Fetch n-th mini-batch $\bar{\mathcal{D}}$ from $\mathcal{D}$;

    4: Select prediction disagreement $\bar{\mathcal{D}}' = \{(x_i, y_i) : \bar{y}_i^{(1)} \neq \bar{y}_i^{(2)}\}$;

    5: Get $\bar{\mathcal{D}}'^{(1)} = \arg\min_{\mathcal{D}':|\mathcal{D}'| \geq \lambda(e)|\bar{\mathcal{D}}'|} \ell(\mathcal{D}'; w^{(1)})$; //sample $\lambda(e)$% small-loss instances

    6: Get $\bar{\mathcal{D}}'^{(2)} = \arg\min_{\mathcal{D}':|\mathcal{D}'| \geq \lambda(e)|\bar{\mathcal{D}}'|} \ell(\mathcal{D}'; w^{(2)})$; //sample $\lambda(e)$% small-loss instances

    7: Update $w^{(1)} = w^{(1)} - \eta \nabla \ell(\bar{\mathcal{D}}'^{(2)}; w^{(1)})$;//update $w^{(1)}$ by $\bar{\mathcal{D}}'^{(2)}$;

    8: Update $w^{(2)} = w^{(2)} - \eta \nabla \ell(\bar{\mathcal{D}}'^{(1)}; w^{(2)})$;//update $w^{(2)}$ by $\bar{\mathcal{D}}'^{(1)}$;

  end

  9: Update $\lambda(e) = 1 - \min\{\frac{e}{E_k}\tau, \tau\}$ or $1 - \min\{\frac{e}{E_k}\tau, (1 + \frac{e - E_k}{E_{\max} - E_k})\tau\}$; (memorization helps)

end

10: Output $w^{(1)}$ and $w^{(2)}$.

Co-teaching+: Step 4: disagreement-update; Step 5-8: cross-update.

Relations to other approaches

Table: Comparison of state-of-the-art and related techniques with our Co-teaching+ approach.
"small loss": regarding small-loss samples as "clean" samples;
"double classifiers": training two classifiers simultaneously;
"cross update": updating parameters in a cross manner;
"divergence": keeping two classifiers diverged during training.

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

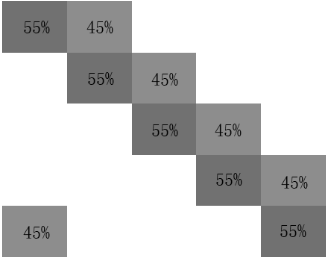
Datasets for CCN model

Table: Summary of data sets used in the experiments.

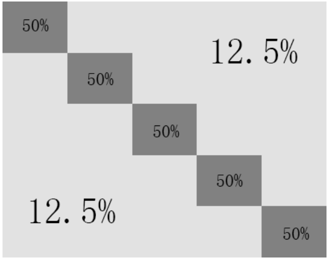
	# of train	# of test	# of class	size
MNIST	60,000	10,000	10	28×28
CIFAR-10	50,000	10,000	10	32×32
CIFAR-100	50,000	10,000	100	32×32
NEWS	11,314	7,532	7	1000-D
T-ImageNet	100,000	10,000	200	64×64

Noise Transitions for CCN model

We manually generate class-conditional noisy labels using two types of noise transitions:



(a) Pair ($\epsilon = 45\%$).



(b) Symmetry ($\epsilon = 50\%$).

Figure: Different noise transitions (using 5 classes as an example) [Han et al., 2018].

Baselines

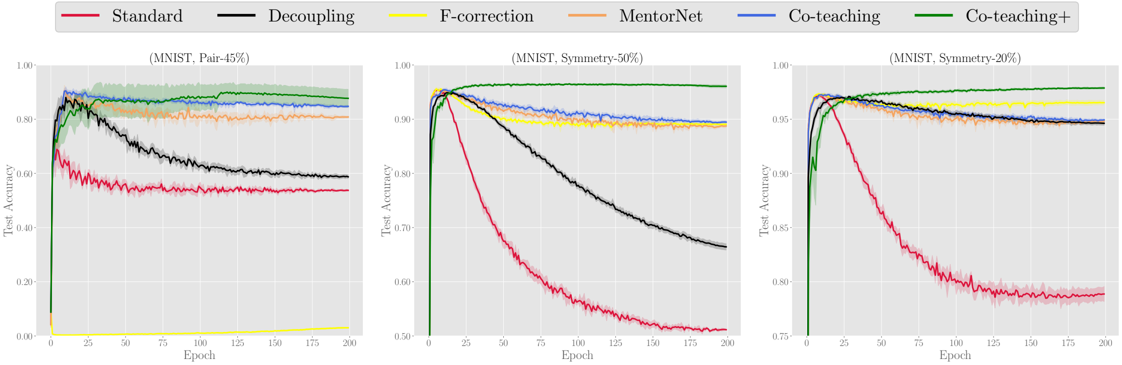
- MentorNet: small-loss trick;
- Co-teaching: small-loss and cross-update trick.
- Decoupling: instances that have different predictions;
- F-correction: loss correction on transition matrix;
- Standard: directly training on noisy datasets.

Network structures

Table: MLP and CNN models used in our experiments on *MNIST*, *CIFAR-10*, *CIFAR-100*/*Open-sets*, and *NEWS*.

MLP on *MNIST*	CNN on *CIFAR-10*	CNN on *CIFAR-100*/*Open-sets*	MLP on *NEWS*
28×28 Gray Image	32×32 RGB Image	32×32 RGB Image	1000-D Text
	5×5 Conv, 6 ReLU 2×2 Max-pool	3×3 Conv, 64 BN, ReLU 3×3 Conv, 64 BN, ReLU 2×2 Max-pool	300-D Embedding Flatten → 1000×300 Adaptive avg-pool → 16×300
Dense 28×28 → 256, ReLU	5×5 Conv, 16 ReLU 2×2 Max-pool	3×3 Conv, 128 BN, ReLU 3×3 Conv, 128 BN, ReLU 2×2 Max-pool	Dense 16×300 → 4×300 BN, Softsign
	Dense 16×5×5 → 120, ReLU Dense 120 → 84, ReLU	3×3 Conv, 196 BN, ReLU 3×3 Conv, 196 BN, ReLU 2×2 Max-pool	Dense 4×300 → 300 BN, Softsign
Dense 256 → 10	Dense 84 → 10	Dense 256 → 100/10	Dense 300 → 7

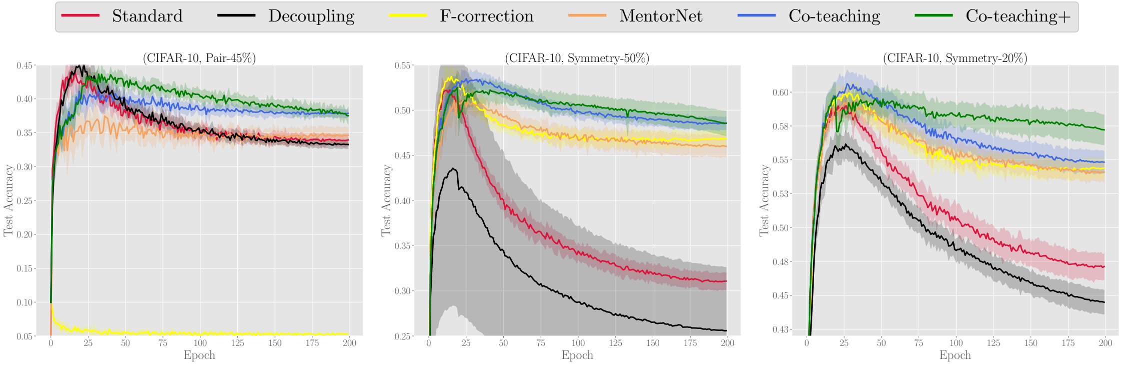
MNIST



(a) Pair-45%. (b) Symmetry-50%. (c) Symmetry-20%.

Figure: Test accuracy vs number of epochs on *MNIST* dataset.

CIFAR-10



(a) Pair-45%. (b) Symmetry-50%. (c) Symmetry-20%.

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

CIFAR-100

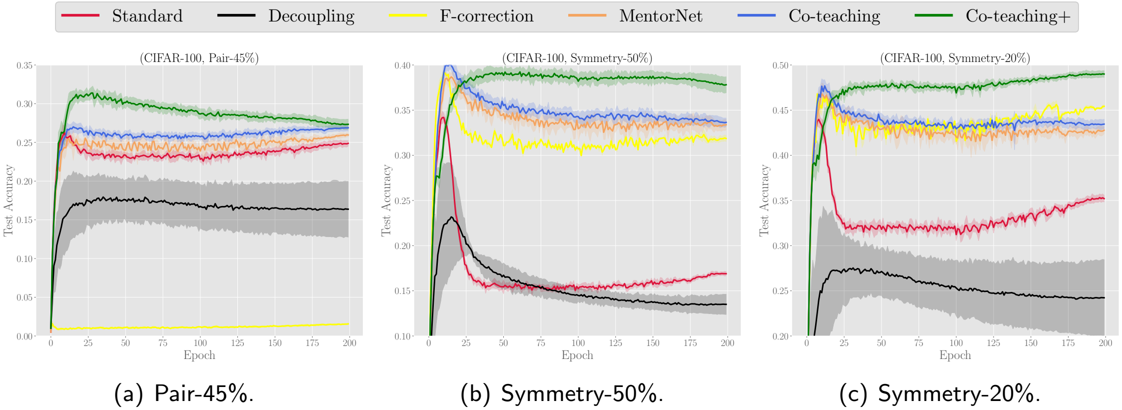
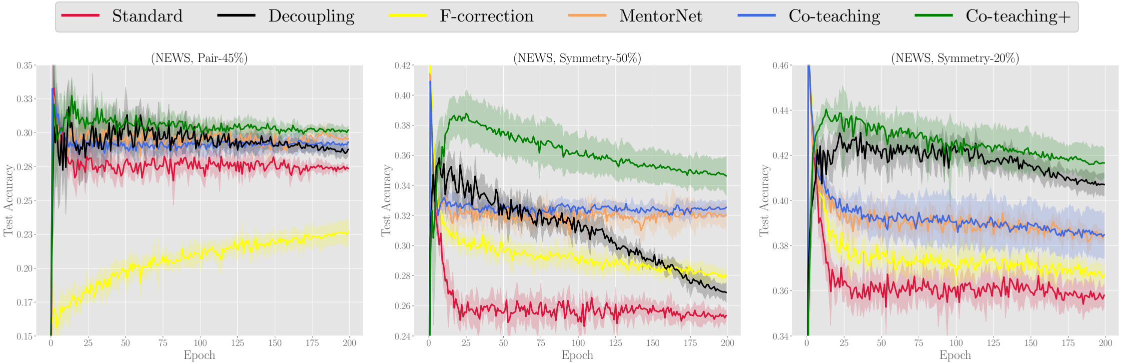


Figure: Test accuracy vs number of epochs on *CIFAR-100* dataset.

NEWS



(a) Pair-45%. (b) Symmetry-50%. (c) Symmetry-20%.

Figure: Test accuracy vs number of epochs on *NEWS* dataset.

T-ImageNet

Table: Averaged/maximal test accuracy (%) of different approaches on *T-ImageNet* over last 10 epochs. The best results are in blue.

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

Open-sets

Open-set noise:
An open-set noisy label occurs when a noisy sample possesses a true class that is not contained within the set of known classes in the training data.

Open-sets: CIFAR-10 noisy dataset with 40% open-set noise from CIFAR-100, ImageNet32, and SVHN.

Figure: Examples of open-set noise for "airplane" in CIFAR-10 [Wang et al., 2018].

Open-sets

Table: Averaged/maximal test accuracy (%) of different approaches on *Open-sets* over last 10 epochs. The best results are in blue.

Open-set noise	Standard	MentorNet	Iterative[Wang et al., 2018]	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

Summary

Conclusion:

- This paper presents Co-teaching+, a robust model for learning on noisy labels.
- Three key points towards robust training on noisy labels:
  1) use small-loss trick based on memorization effects of deep networks;
  2) cross-update parameters of two networks;
  3) keep two networks diverged during training.

Future work:

- Investigate the theory of Co-teaching+ from the view of disagreement-based algorithms [Wang and Zhou, 2017].

Link to our paper:

Our poster will be:

Wed Jun 12th 06:30 – 09:00 PM@Pacific Ballroom #21

Thank you very much for your attention!

References

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. (2017). A closer look at memorization in deep networks. In *International Conference on Machine Learnin*, pages 233–242.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, pages 8527–8537.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. (2018). Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*.

Malach, E. and Shalev-Shwartz, S. (2017). Decoupling" when to update" from" how to update". In *Advances in Neural Information Processing Systems*, pages 960–970.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Wang, W. and Zhou, Z.-H. (2017). Theoretical foundation of co-training and disagreement-based algorithms. *arXiv preprint arXiv:1708.04403*.

Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., and Xia, S.-T. (2018). Iterative learning with open-set noisy labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8688–8696.