# Deep Learning with Noisy Supervision

Ivor W. Tsang

Centre for Artificial Intelligence, University of Technology Sydney, Australia

Jun 16th, 2019

# Outline

# Machine Learning in last two decades

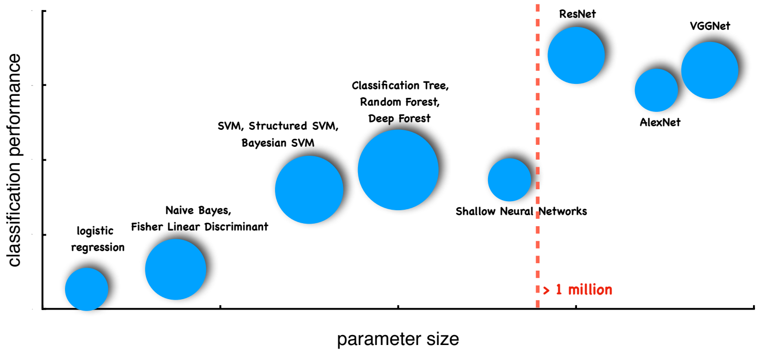



Image classification as one fundamental task in computer vision has been well investigated for a long time. Benefiting from the development of deep learning, a significant improvement have been achieved in many practical applications, e.g., clothing, food or car classification.

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

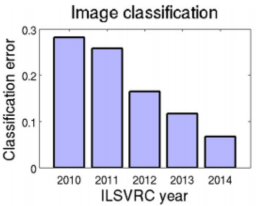
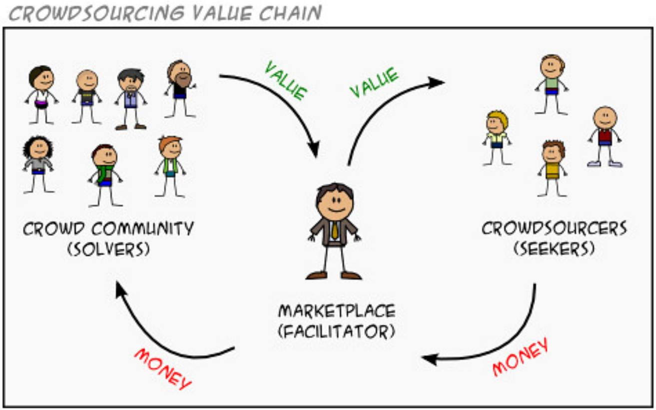



Figure: There is a steady reduction of error every year in object classification on large scale dataset (1000 object categories, 1.2 million training images) [Russakovsky et al., 2015].

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

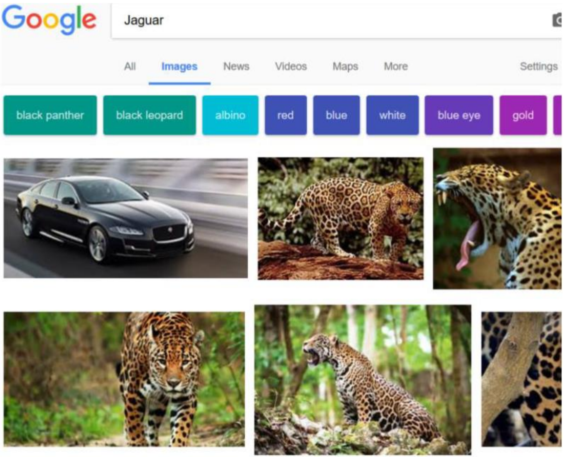
# Noisy labels from crowdsourcing platforms.




Credit: *Torbjørn Marø*

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

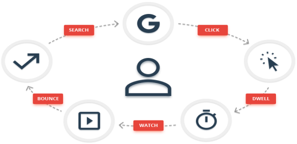
# Noisy labels from web search/crawler.

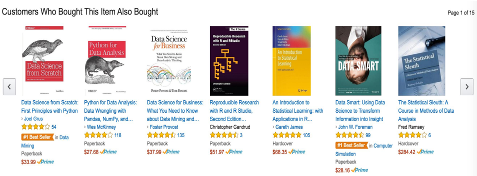



Screenshot of Google.com

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

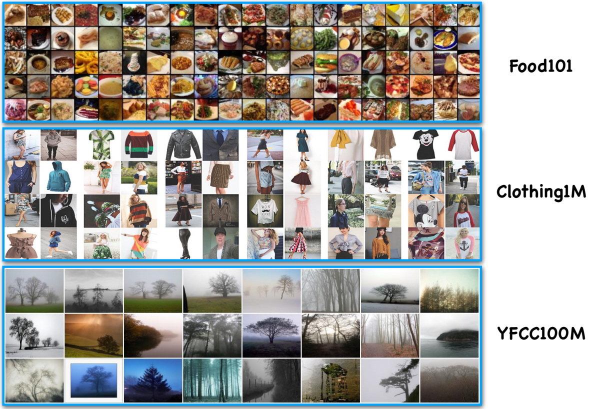
# Noisy labels from implicit feedback.




- Customers may accidentally miss some links in a quick browse.

# Real-world Noisy Databases



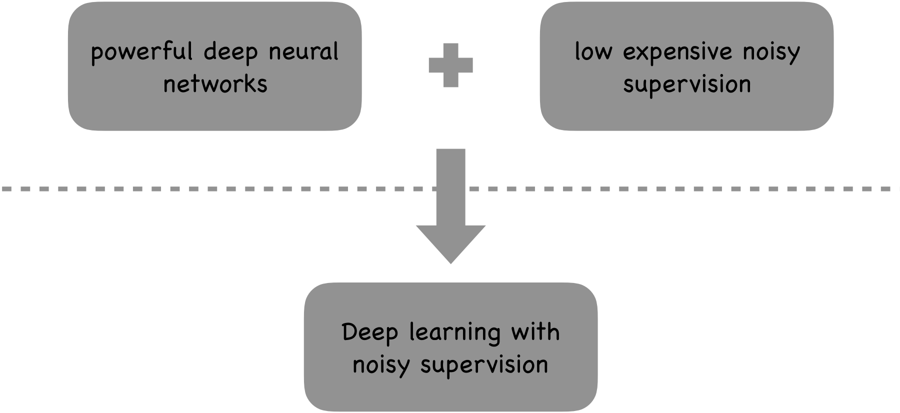

There are almost inexhaustible noisy annotated images available on the social and e-commerce websites at very low cost of human labor.

# Processing Noisy Data

## Bottleneck on Labor Annotation


Considering the expensive human labor in the complex and arbitrary applications e.g., medical diagnostic and fine-grained visualization, collecting a large-scale dataset with accurate annotations is usually impractical.

# Deep Learning + Noisy Labels



# 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 $\forall_{i,j} T_{ij} = p(\tilde{y} = j | y = i)$.

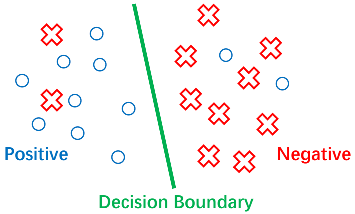



Figure: Illustration of noisy labels.

# What happens when learning with noisy labels?

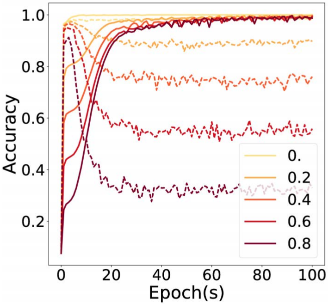



Figure: Accuracy of neural networks on noisy MNIST with different noise rate (0., 0.2, 0.4, 0.6, 0.8).
(Solid is train, dotted is validation.) [Arpit et al., 2017]

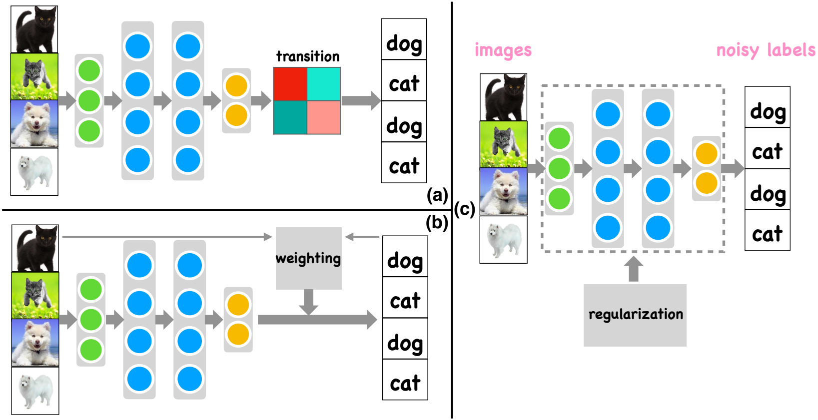
Memorization: Learning easy patterns first, then (totally) over-fit noisy training data.

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

# Deep Learning with Noisy Supervision

## How to do in this area?

Three popular methodologies currently applied in this area.



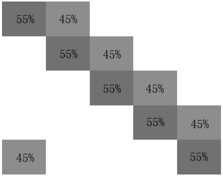
# Current Works

Current progress in three orthogonal directions:

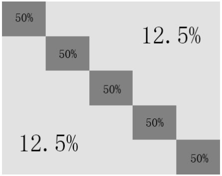
- Learning with noise transition:
  Forward Correction (Australian National University, CVPR'17)
  S-adaptation (Bar Ilan University, ICLR'17)
  Masking (UTS, NeurIPS'18)
- Learning with selected samples:
  MentorNet (Google AI, ICML'18)
  Learning to Reweight Examples (University of Toronto, ICML'18)
  Co-teaching (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)

# Estimating Noise Transition Matrix

- Main idea: estimate the matrix and learn the classifier
- Benefit: with theoretical guarantees
- Drawback: hard to estimate the matrix for large-class cases



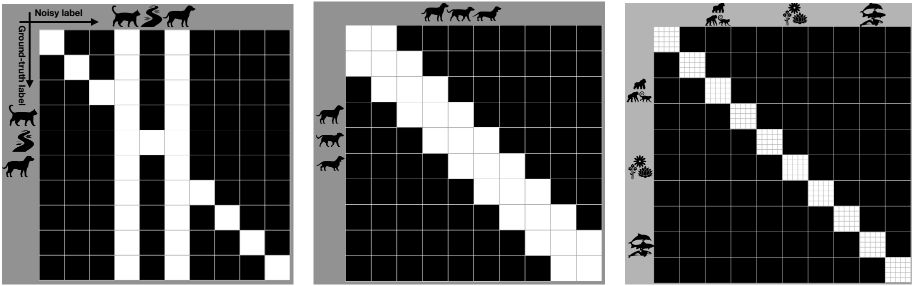

(a) pair ($\epsilon = 45\%$).




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

Figure: The noise transition matrix $T$, where $T_{ij} = \Pr(\tilde{y} = e^j | y = e^i)$.

# Data Perspective




(a) Column-diagonal (b) Tri-diagonal (c) Block-diagonal

Figure: Three types of noise structure.

- (a) beach $\not\leftrightarrow$ mountain; beach $\leftrightarrow$ dog.
- (b1) Australian terrier $\not\leftrightarrow$ Norwich terrier;
- (b2) Norfolk terrier $\leftrightarrow$ Norwich terrier $\leftrightarrow$ Irish terrier.
- (c) aquatic mammals $\not\leftrightarrow$ flowers; beaver $\leftrightarrow$ dolphin.

# Deficiency of Benchmarks

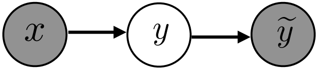



Figure: Benchmark models. $(x, \tilde{y})$ denotes the instance with the noisy label.

- Independent framework: the estimation is not for agnostic noisy data.
- Unified framework: the brute-force estimation suffer local minimums.

# Our Solution: Structure-aware probabilistic model

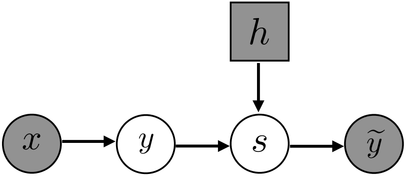



Figure: MASKING models the matrix $T$, where $T_{ij} = \Pr(\tilde{y} = e^j | y = e^i)$, by an explicit variable $s$. Thus, we embed a structure constraint ($h$) on the variable $s$.

- Human cognition masks the invalid class transitions.
- The model focuses on estimating the noise transition probability.
- The estimation burden will be largely reduced.

# Straightforward Dilemma

- In deep learning, hard to choose a distance measure (e.g., L2).
- Clean validation: repeat the training procedure to tune parameters.

# When Structure Meets Generative Model

- The latent ground-truth label $y \sim P(y|x)$ (Categorical).
- The transition $s \sim P(s)$ and its structure $s_o \sim P(s_o)$, where $P(s)$ is an implicit distribution modeled by DNN, $P(s_o) = P(s)\frac{ds}{ds_o}\Big|_{s_o=f(s)}$. $f(\cdot)$ is the mapping function from $s$ to $s_o$.
- The noisy label $\tilde{y} \sim P(\tilde{y}|y, s)$, where $P(\tilde{y}|y, s)$ models the transition from $y$ to $\tilde{y}$ given $s$.

# ELBO of MASKING

$$\ln P(\tilde{y}|x) \geq \mathbb{E}_{Q(s)}\left[\underbrace{\ln \sum_{y} P(\tilde{y}|y,s)P(y|x)}_{\text{previous model}} - \ln\left(Q(s_o)/\underbrace{P(s_o)}_{\text{structure prior}}\right)\Bigg|_{s_o=f(s)}\right],$$

where $Q(s)$ is the variational distribution to approximate the posterior of the noise transition matrix $s$, and $Q(s_o) = Q(s)\frac{ds}{ds_o}\big|_{s_o=f(s)}$ is the corresponding variational distribution of the structure $s_o$.

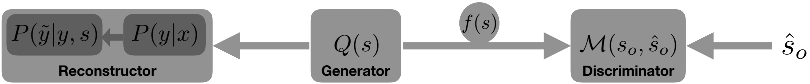
## Remark

*MASKING benefits from the human guidance (the second term) in the procedure of learning with noisy supervision (the first term).*

# Principled Realization

Q: Challenge from structure alignment.

A: GAN-like structure to model the structure instillation.



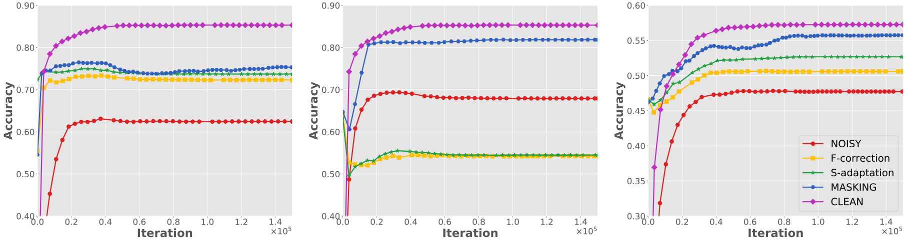
# Datasets

Table: Benchmark CIFAR10 and CIFAR100; Industrial-level Clothing1M.

| | # of training | # of testing | # of class | size |
|---|---|---|---|---|
| *CIFAR10* | 50,000 | 10,000 | 10 | 32×32 |
| *CIFAR100* | 50,000 | 10,000 | 1000 | 32×32 |
| *Clothing1M* | 1,000,000(N) + 5,000(C) | 1,000 | 14 | 256×256 |


Figure: Mislabeled images often share similar visual patterns in Clothing1M.

# CIFAR10 and CIFAR100




(a) Column-diagnoal (b) Tri-diagnoal (c) Block-diagnoal

Figure: The test accuracy vs iterations on benchmark datasets.

# Clothing1M

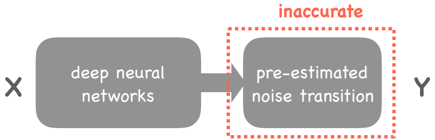
Table: Test accuracy on Clothing1M.

| Models | Performance(%) |
|---|---|
| NOISY | 68.9 |
| F-correction | 69.8 |
| S-adaptation | 70.3 |
| MASKING | 71.1 |
| CLEAN | 75.2 |

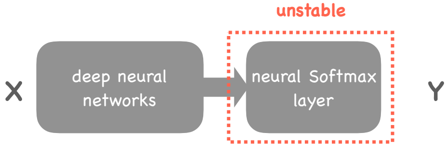
# LCCN

## Motivation

One-step pre-estimation of noise transition.




Adapt the noise transition via the neural layer.

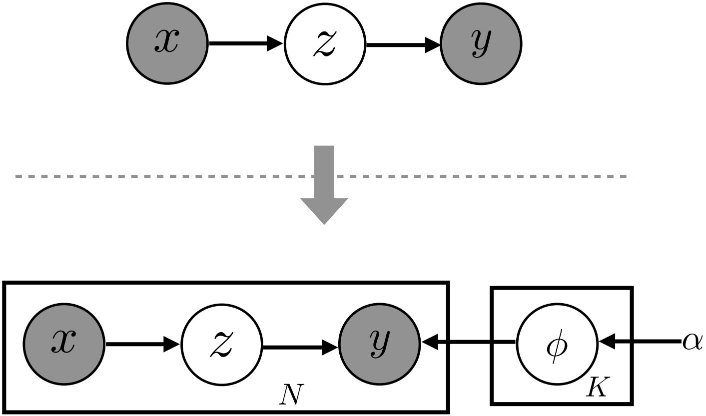



Issues: ignore the global dependency of noise transition.

# LCCN

## Latent Class-Conditional Noise Model

Reformulate original model.



# LCCN

## Dynamic Label Regression

Inference: Autoencoded Gibbs Sampling

$$P(z_n|Z^{\neg n}, X, Y; \alpha) \propto \underbrace{P(z_n|x_n)}_{\text{Classifier encoder}} \underbrace{\frac{\alpha_{y_n} + N^{\neg n}_{z_n y_n}}{\sum_{k'}^{K}(\alpha_{k'} + N^{\neg n}_{z_n k'})}}_{\text{Conditional transition}}. \quad (1)$$

Learning: Independent Optimization

$$\begin{cases} \min -\frac{1}{n}\sum_{n=1}^{N} \ell_1(z_n, P(z_n|x_n)) \\ \min -\frac{1}{n}\sum_{n=1}^{N} \ell_2(y_n, P(y_n|z_n)). \end{cases} \quad (2)$$

# LCCN

## Guarantee

### Theorem

*Suppose $\alpha_i$ is a positive smoothing scalar, $N_i$ is the current sample number of the ith category (i=1,...,K), $M_i$ is the sum of the sample numbers newly allocated into (positive) and removed from (negative) the ith category after a batch of training samples, and $\widehat{M}_i$ is its absolute sum of such two cases. Then, for the transition vector $\phi_i$ of the ith category, its variation via a training batch is characterized by the below inequality,*

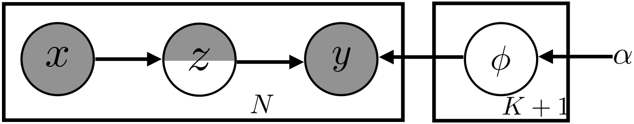
$$\left|\phi_i^{new} - \phi_i^{old}\right| \le \frac{|r_i| + \widehat{r}_i}{1 + r_i} \tag{3}$$

*where $r_i = \frac{M_i}{N_i + \sum_{j=1}^{K} \alpha_j}$ and $\widehat{r}_i = \frac{\widehat{M}_i}{N_i + \sum_{j=1}^{K} \alpha_j}$. According to the definition, we have $r_i > -1$, $\widehat{r}_i \ge 0$ and $\widehat{r}_i \ge |r_i|$.*

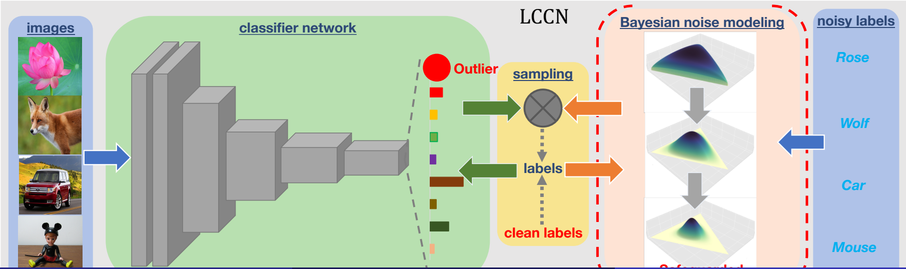
# LCCN

## Illustration

LCCN can be easily extended to the open-set noise setting and the semi-supervised learning with the similar optimization.




The illustration of the extended training procedure.



# LCCN

## Experiments on toy datasets

Table: The average accuracy (%) over 5 trials on CIFAR-10 and CIFAR-100 with different noise.

| Dataset | | CIFAR-10 | | | | | CIFAR-100 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # | Method \ Noise Ratio | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| 1 | CE | 90.10 | 88.12 | 76.93 | 59.01 | 56.85 | 66.15 | 64.31 | 60.11 | 51.68 | 33.37 |
| 2 | Bootstrapping | 90.73 | 88.12 | 76.29 | 57.04 | 56.79 | 66.48 | 64.61 | 63.01 | 55.27 | 34.52 |
| 3 | Forward | 90.86 | 89.03 | 82.47 | 67.11 | 57.29 | 65.43 | 62.72 | 61.28 | 52.64 | 33.82 |
| 4 | S-adaptation | 91.02 | 88.83 | 86.79 | 72.74 | 60.92 | 65.52 | 64.11 | 62.39 | 52.74 | 30.07 |
| 5 | LCCN | 91.35 | 89.33 | 88.41 | 79.48 | 64.82 | 67.83 | 67.63 | 66.86 | 65.52 | 33.71 |
| 6 | CE with the clean data | 91.63 | | | | | 69.41 | | | | |

Table: The average accuracy (%) over 5 trials on CIFAR-10 and CIFAR-100 with different noise under the extended settings.

| Dataset | | CIFAR-10 | | | | | CIFAR-100 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # | Method \ Noise Ratio | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| 1 | CE | 89.13 | 87.06 | 74.63 | 62.29 | 57.07 | 62.94 | 59.73 | 54.71 | 45.57 | 31.74 |
| 2 | Bootstrapping | 90.13 | 84.58 | 74.76 | 54.87 | 55.56 | 63.73 | 60.88 | 59.77 | 40.23 | 31.86 |
| 3 | Forward | 88.63 | 84.97 | 78.47 | 58.23 | 56.52 | 63.69 | 62.63 | 61.86 | 51.47 | 35.71 |
| 4 | S-adaptation | 88.58 | 87.28 | 61.17 | 57.12 | 56.73 | 63.51 | 61.50 | 60.59 | 53.22 | 32.19 |
| 5 | LCCN | 88.63 | 88.06 | 82.15 | 69.48 | 55.12 | 63.97 | 62.84 | 61.79 | 60.34 | 33.52 |
| 6 | LCCN* | 89.59 | 88.43 | 84.34 | 72.33 | 56.28 | 64.71 | 63.05 | 62.48 | 62.02 | 32.37 |
| 7 | LCCN+ | 90.30 | 88.93 | 88.21 | 87.42 | 86.33 | 65.67 | 64.24 | 63.52 | 63.19 | 62.39 |

# LCCN

## Experiments

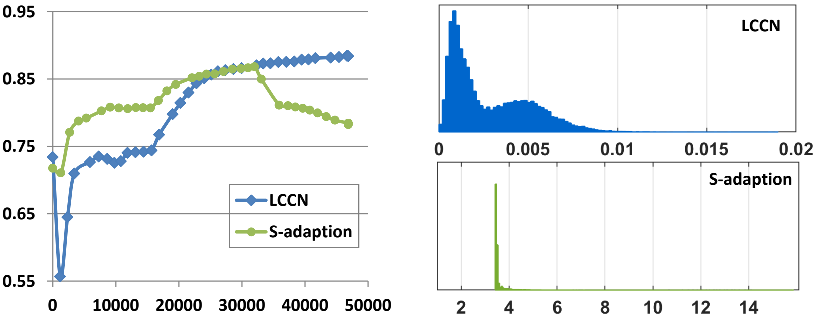
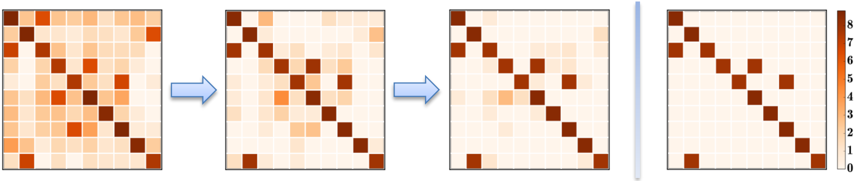



Figure: The test accuracy of LCCN and S-adaptation in the training on CIFAR-10 with $r$=0.5 and the corresponding histograms for the change of noise transition $\phi$ via a mini-batch of samples.

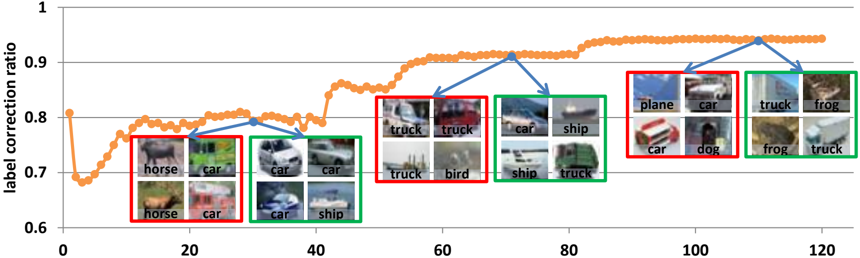
# LCCN

## Experiments

The transition learning in dynamic label regression.




The label inference in dynamic label regression.



# LCCN

## Experiments on Clothing1M

Table: The average accuracy over 5 trials on Clothing1M.

| # | Method | Accuracy |
|---|---|---|
| 1 | CE | 68.94 |
| 2 | Bootstrapping | 69.12 |
| 3 | Forward | 69.84 |
| 4 | S-adaptation | 70.36 |
| 5 | Joint Optimization | 72.16 |
| 6 | LCCN | 71.63 |
|  | LCCN warmed-up by $\phi$ in ? | 73.07 |
|  | LCCN* | 72.80 |
| 7 | CE on the clean data | 75.28 |
| 8 | Forward+ | 80.38 |
| 9 | LCCN+ | 81.25 |

# LCCN

## Experiments on Clothing1M

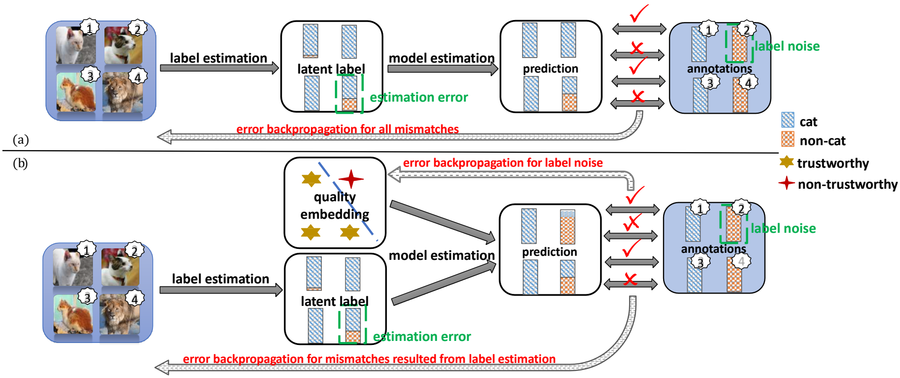
The learned noise transition on Clothing1M by LCCN.

| class | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 77 | | 6 | | | 4 | | | | | | | | | T-shirt |
| 2 | 2 | 88 | | 3 | | | | | | | | | | | shirt |
| 3 | 12 | 4 | 48 | 4 | 7 | 7 | | | | | | 5 | | | knitwear |
| 4 | 9 | 15 | 5 | 51 | | | | | | | | 9 | | | chiffon |
| 5 | 11 | 4 | 43 | | 16 | 9 | | | | | | 4 | | | sweater |
| 6 | 3 | | 3 | | | 86 | | | | | | | | | hoodie |
| 7 | | | | | | | 87 | 3 | | 5 | | | | | windbreaker |
| 8 | | | | | | | | 92 | | | | | | | jacket |
| 9 | | | | | | | 2 | | 96 | | | | | | downcoat |
| 10 | | | | | | | 4 | 2 | | 91 | | | | | suit |
| 11 | | | 3 | | | | | | | | 81 | 3 | | | shawl |
| 12 | | | | 3 | | | | | | | | 84 | 3 | | dress |
| 13 | | | | 3 | | | | | | | | 6 | 76 | 3 | vest |
| 14 | 1 | | | | | | | | | | | | | 95 | underwear |
| 15 | 12 | 10 | | 7 | | 9 | 6 | 4 | | 7 | | 14 | 6 | 3 | outlier |

# Quality Embedding

## Composite reasoning way

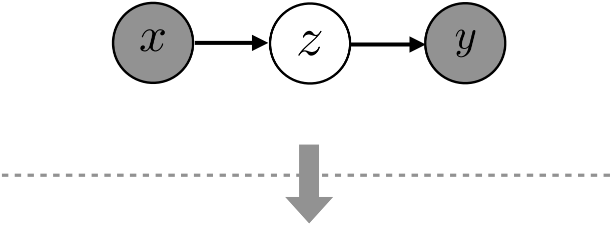
The intuitive idea to deal with the residual noise effect.

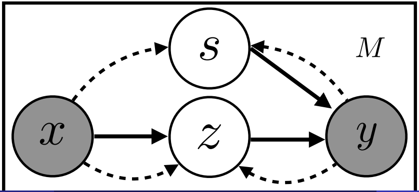


# Quality Embedding

## Quality Augmented Probabilistic Model

Reformulate the original model.



# Quality Embedding

## Objective

The regularized objective

$$\min \ \hat{L} = -\mathrm{E}_{q(z|x,y),q(s|x,y)}\left[\ln P(y|z,s)\right]$$

$$+ \mathrm{D}_{\mathrm{KL}}\left[q(z|x,y)||\underbrace{P(z|x)}_{\text{classifier}}\right] + \mathrm{D}_{\mathrm{KL}}\left[q(s|x,y)||P(s)\right]$$

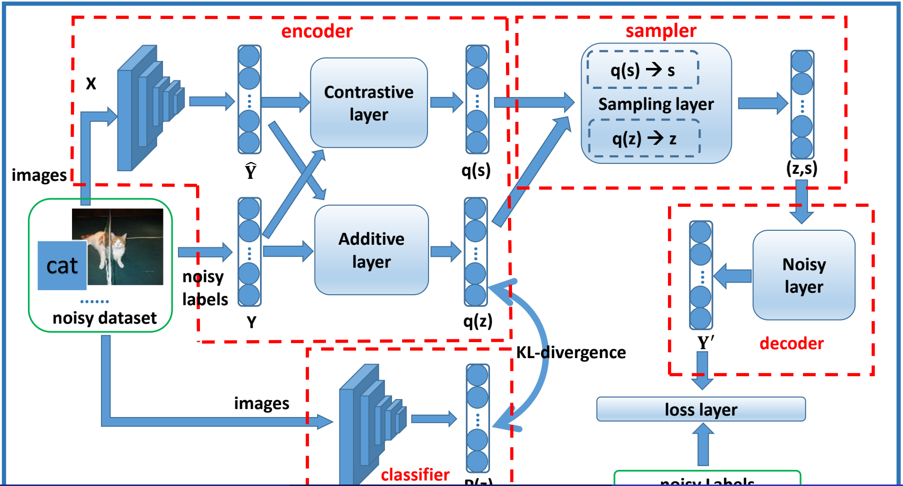
$$- \lambda \underbrace{\left(\mathrm{E}_{q(z|x,y)}\left[\ln q(z|x,y)\right] + \mathrm{E}_{q(s|x,y)}\left[\ln q(s|x,y)\right]\right)}_{\text{variational mutual regularizer}}.$$

This can be optimized by reparameterization tricks as VAE ?.

# Quality Embedding

## Contrastive-Additive Neural network (CAN)

The network implementation of the proposed model.



# Experiments

## On PASCAL VOC From Web Sources

Table: classification results on the 20 categories of *VOC 07*.

| Model | Resnet-N | LearnQ | ICNM | Bootstrap | CAN |
|---|---|---|---|---|---|
| aeroplane | 98.4 | 98.4 | 98.1 | 98.6 | 98.8 |
| bicycle | 81.1 | 83.8 | 82.9 | 84.1 | 84.1 |
| bird | 92.9 | 93.8 | 93.6 | 93.6 | 95.3 |
| boat | 88.7 | 88.5 | 88.9 | 90.9 | 93.2 |
| bottle | 57.0 | 53.5 | 53.4 | 56.3 | 62.1 |
| bus | 87.4 | 87.8 | 87.7 | 89.8 | 90.8 |
| car | 73.2 | 73.7 | 72.3 | 75.5 | 77.0 |
| cat | 96.6 | 96.5 | 96.2 | 96.3 | 97.9 |
| chair | 63.3 | 64.3 | 64.7 | 69.8 | 72.6 |
| cow | 90.0 | 90.6 | 91.2 | 91.6 | 94.4 |
| table | 63.9 | 62.6 | 66.3 | 69.9 | 73.5 |
| dog | 94.3 | 94.6 | 94.2 | 94.4 | 96.1 |
| horse | 95.0 | 96.1 | 96.2 | 95.8 | 97.7 |
| motorbike | 92.9 | 91.6 | 91.4 | 93.2 | 94.3 |
| person | 76.8 | 78.4 | 78.0 | 82.2 | 82.4 |
| plant | 43.8 | 46.8 | 44.0 | 43.2 | 45.5 |
| sheep | 92.9 | 92.8 | 93.5 | 92.8 | 95.8 |
| sofa | 67.2 | 69.0 | 69.3 | 70.9 | 71.4 |
| train | 93.1 | 94.0 | 94.4 | 95.4 | 95.8 |
| tv | 65.1 | 65.4 | 66.9 | 67.4 | 68.6 |
| mAP | 80.7 | 81.1 | 81.2 | 82.6 | 84.4 |

# Experiments

## On PASCAL VOC From Web Sources

Table: classification results on the 20 categories of *VOC 12*.

| Model | Resnet-N | LearnQ | ICNM | Bootstrap | CAN |
|---|---|---|---|---|---|
| aeroplane | 98.4 | 98.4 | 98.1 | 98.6 | 98.8 |
| bicycle | 81.1 | 83.8 | 82.9 | 84.1 | 84.1 |
| bird | 92.9 | 93.8 | 93.6 | 93.6 | 95.3 |
| boat | 88.7 | 88.5 | 88.9 | 90.9 | 93.2 |
| bottle | 57.0 | 53.5 | 53.4 | 56.3 | 62.1 |
| bus | 87.4 | 87.8 | 87.7 | 89.8 | 90.8 |
| car | 73.2 | 73.7 | 72.3 | 75.5 | 77.0 |
| cat | 96.6 | 96.5 | 96.2 | 96.3 | 97.9 |
| chair | 63.3 | 64.3 | 64.7 | 69.8 | 72.6 |
| cow | 90.0 | 90.6 | 91.2 | 91.6 | 94.4 |
| table | 63.9 | 62.6 | 66.3 | 69.9 | 73.5 |
| dog | 94.3 | 94.6 | 94.2 | 94.4 | 96.1 |
| horse | 95.0 | 96.1 | 96.2 | 95.8 | 97.7 |
| motorbike | 92.9 | 91.6 | 91.4 | 93.2 | 94.3 |
| person | 76.8 | 78.4 | 78.0 | 82.2 | 82.4 |
| plant | 43.8 | 46.8 | 44.0 | 43.2 | 45.5 |
| sheep | 92.9 | 92.8 | 93.5 | 92.8 | 95.8 |
| sofa | 67.2 | 69.0 | 69.3 | 70.9 | 71.4 |
| train | 93.1 | 94.0 | 94.4 | 95.4 | 95.8 |
| tv | 65.1 | 65.4 | 66.9 | 67.4 | 68.6 |
| mAP | 80.7 | 81.1 | 81.2 | 82.6 | 84.4 |

# Experiments

## On Standford Dog Datasets From Crowdsourcing

Table: Classification results on 4 categories of Standford Dog.

| Model | nft | nwt | iwh | swh | mAP |
|---|---|---|---|---|---|
| MLP-N | 78.1 | 73.2 | 80.9 | 76.5 | 77.2 |
| LearnQ | 80.5 | 73.7 | 83.0 | 77.7 | 78.7 |
| ICNM | 80.5 | 72.8 | 83.9 | 78.3 | 78.9 |
| Bootstrap | 80.7 | 72.5 | 83.7 | 78.1 | 78.8 |
| CAN | 82.0 | 79.0 | 81.8 | 83.8 | 81.7 |

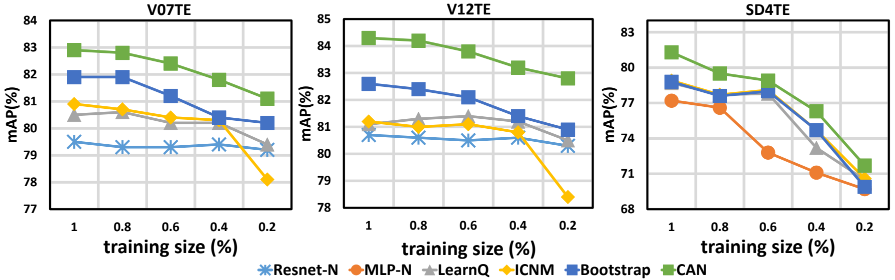
# Experiments

## On Lambda

Results with different regularization coefficient $\lambda$ in CAN.

| $\lambda$ | 0 | 0.2 | 0.5 | 1 | 2 | 5 | 10 |
|---|---|---|---|---|---|---|---|
| *V07TE* | 82.9 | 83.5 | 84.8 | 83.6 | 80.7 | 78.8 | 77.0 |
| *V12TE* | 84.3 | 85.2 | 84.1 | 83.0 | 80.8 | 78.3 | 76.6 |
| *SD4TE* | 78.6 | 80.7 | 80.4 | 79.9 | 76.4 | 73.9 | 71.3 |

Classification results with different training sizes.



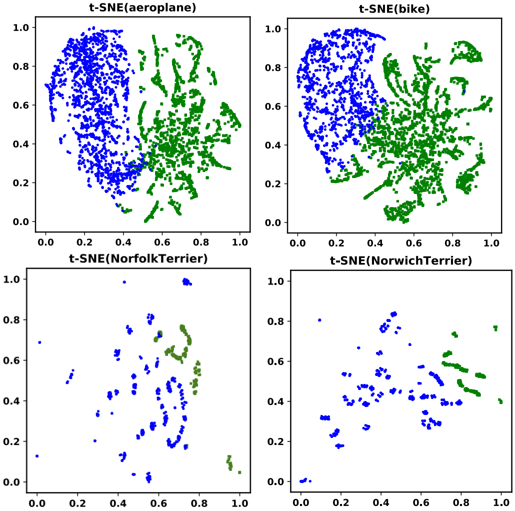
# Experiments

## Classification with Controlable Noise

| Dataset | $P_{noise}$ | 1.0 | 0.8 | 0.6 | 0.4 | 0.2 | 0.0 |
|---|---|---|---|---|---|---|---|
| *V07TE* | Resnet-N | 6.4 | 33.4 | 53.0 | 70.2 | 78.2 | 86.8 |
| | LearnQ | 9.1 | 28.0 | 56.4 | 72.0 | 80.1 | 85.4 |
| | ICNM | 9.2 | 28.5 | 57 | 71.6 | 79.6 | 85.4 |
| | Bootstrap | 8.9 | 30.1 | 59.3 | 73.3 | 81.0 | 85.5 |
| | CAN | 8.6 | 36.1 | 63.2 | 79.4 | 83.6 | 85.3 |
| *V12TE* | Resnet-N | 5.2 | 26.6 | 49.2 | 69.0 | 80.0 | 89.7 |
| | LearnQ | 8.4 | 23.7 | 49.7 | 70.3 | 81.3 | 88.3 |
| | ICNM | 8.4 | 23.8 | 49.6 | 70.5 | 81.4 | 88.3 |
| | Bootstrap | 8.2 | 25.1 | 51.8 | 72.6 | 82.2 | 88.5 |
| | CAN | 10.5 | 28.0 | 55.3 | 78.4 | 84.5 | 87.3 |
| *SD4TE* | MLP-N | 29.6 | 41.6 | 51.5 | 73.4 | 86.1 | 96.4 |
| | LearnQ | 26.9 | 39.6 | 60.4 | 72.7 | 89.0 | 95.9 |
| | ICNM | 27.0 | 39.7 | 60.8 | 73.1 | 89.2 | 95.8 |
| | Bootstrap | 27.8 | 38.6 | 58.7 | 73.5 | 89.3 | 96.2 |
| | CAN | 30.1 | 49.7 | 63.9 | 77.1 | 91.1 | 94.3 |

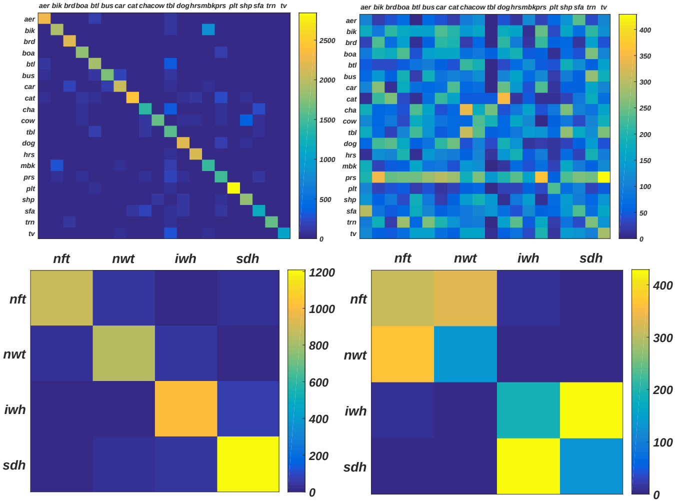
# Experiments

## On Quality Embedding Visualization



# Experiments

## On Conditional Transition



# Experiments

## Analysis of latent labels.


Exemplars on latent label estimation of *WEB* dataset (the first two rows) and *AMT* dataset (the third row) as well as some failures (the fourth row). We forward the noisy label (black word)

# A promising research line: Learning with small-loss instances

- Main idea: regard small-loss instances as "correct" instances.

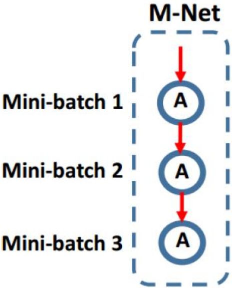



Figure: Self-training MentorNet[Jiang et al., 2018].

- 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?

# Co-teaching: Cross-update meets small-loss

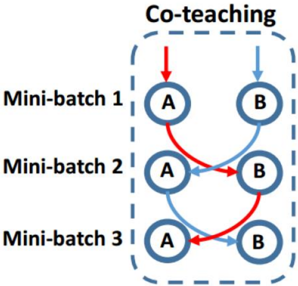



Figure: Co-teaching[Han et al., 2018].

- 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)

# Co-teaching Paradigm

Input: $w_f$ and $w_g$, learning rate $\eta$, fixed $\tau$, epoch $T_k$ and $T_{\max}$, iter $N_{\max}$;
for $T = 1, 2, \ldots, T_{max}$ do
  Shuffle: training set $\mathcal{D}$; //noisy dataset;
  for $N = 1, \ldots, N_{\max}$ do
    Draw: mini-batch $\bar{\mathcal{D}}$ from $\mathcal{D}$;
    Sample: $\bar{\mathcal{D}}_f = \arg\min_{\bar{\mathcal{D}}} \ell(f, \bar{\mathcal{D}}, R(T))$; //$R(T)$% small-loss;
    Sample: $\bar{\mathcal{D}}_g = \arg\min_{\bar{\mathcal{D}}} \ell(g, \bar{\mathcal{D}}, R(T))$; //$R(T)$% small-loss;
    Update: $w_f = w_f - \eta\nabla f(\bar{\mathcal{D}}_g)$; //update $w_f$ by $\bar{\mathcal{D}}_g$;
    Update: $w_g = w_g - \eta\nabla g(\bar{\mathcal{D}}_f)$; //update $w_g$ by $\bar{\mathcal{D}}_f$;
  end
  Update: $R(T) = 1 - \min\left\{\frac{T}{T_k}\tau, \tau\right\}$;
end
Output: $w_f$ and $w_g$

Algorithm 1: Co-teaching Paradigm.

# 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+.

# Decoupling

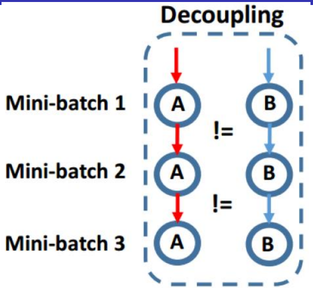
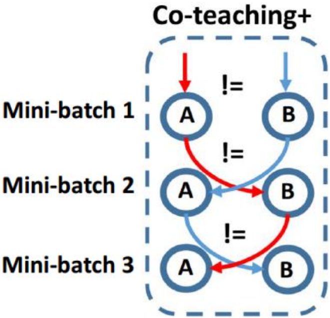



Figure: Decoupling[Malach and Shalev-Shwartz, 2017].

- 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?

# 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 $\eta$, estimated noise rate $\tau$, 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 | × | ✓ | × | ✓ | ✓ |

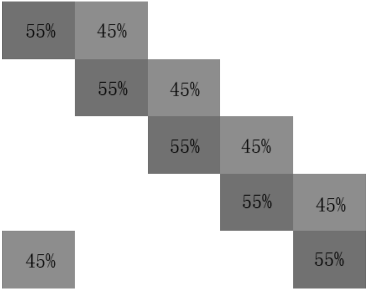
# Datasets for CCN model

Table: Summary of data sets used in the experiments.

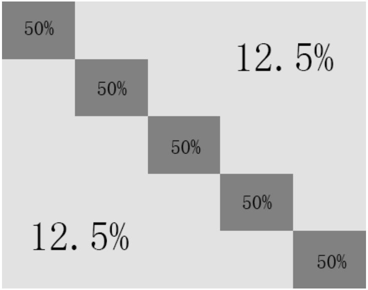
| | # of train | # of test | # of class | size |
|---|---|---|---|---|
| *MNIST* | 60,000 | 10,000 | 10 | 28×28 |
| *CIFAR-10* | 50,000 | 10,000 | 10 | 32×32 |
| *CIFAR-100* | 50,000 | 10,000 | 100 | 32×32 |
| *NEWS* | 11,314 | 7,532 | 7 | 1000-D |
| *T-ImageNet* | 100,000 | 10,000 | 200 | 64×64 |

# 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

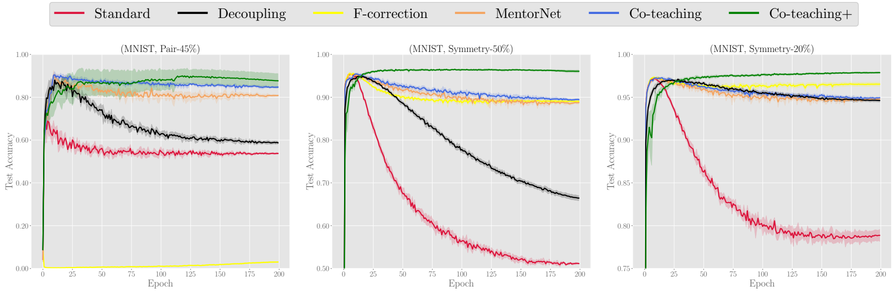
- MentorNet: small-loss trick;
- Co-teaching: small-loss and cross-update trick.
- Decoupling: instances that have different predictions;
- F-correction: loss correction on transition matrix;
- Standard: directly training on noisy datasets.

# 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 |
| Dense 28×28 → 256, ReLU | 5×5 Conv, 6 ReLU<br>2×2 Max-pool | 3×3 Conv, 64 BN, ReLU<br>3×3 Conv, 64 BN, ReLU<br>2×2 Max-pool | 300-D Embedding<br>Flatten → 1000×300<br>Adaptive avg-pool → 16×300 |
| | 5×5 Conv, 16 ReLU<br>2×2 Max-pool | 3×3 Conv, 128 BN, ReLU<br>3×3 Conv, 128 BN, ReLU<br>2×2 Max-pool | Dense 16×300 → 4×300<br>BN, Softsign |
| | Dense 16×5×5 → 120, ReLU<br>Dense 120 → 84, ReLU | 3×3 Conv, 196 BN, ReLU<br>3×3 Conv, 196 BN, ReLU<br>2×2 Max-pool | Dense 4×300 → 300<br>BN, Softsign |
| Dense 256 → 10 | Dense 84 → 10 | Dense 256 → 100/10 | Dense 300 → 7 |

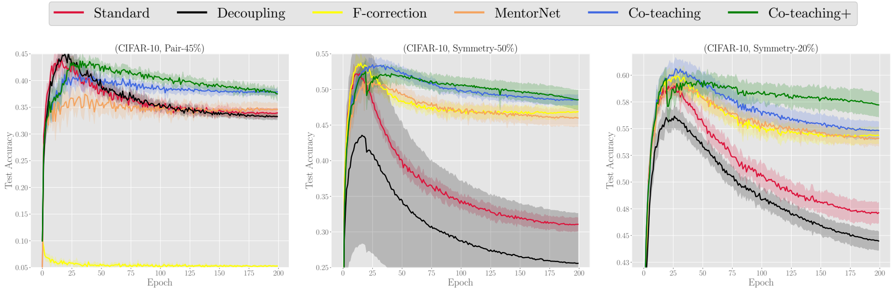
# *MNIST*




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

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

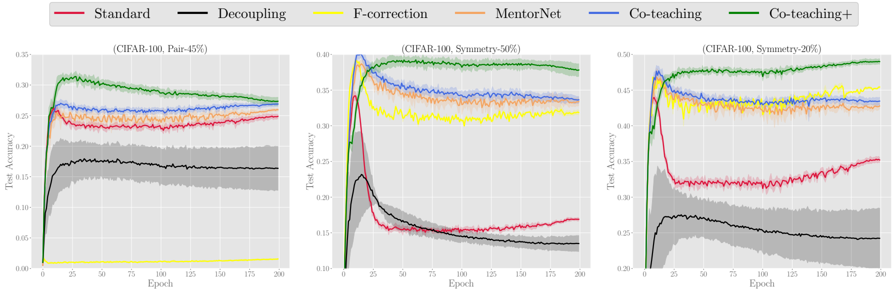
# *CIFAR-10*




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

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

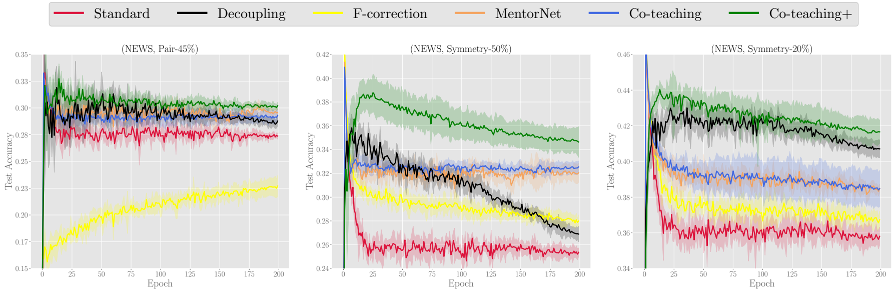
# CIFAR-100




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

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].

# Papers and Codes

- Masking: A New Perspective of Noisy Supervision. *NIPS*, 2018.
- Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. *NIPS*, 2018.
- How does Disagreement Help Generalization against Label Corruption? *ICML*, 2019.

# 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.