Deep Learning with Noisy Supervision

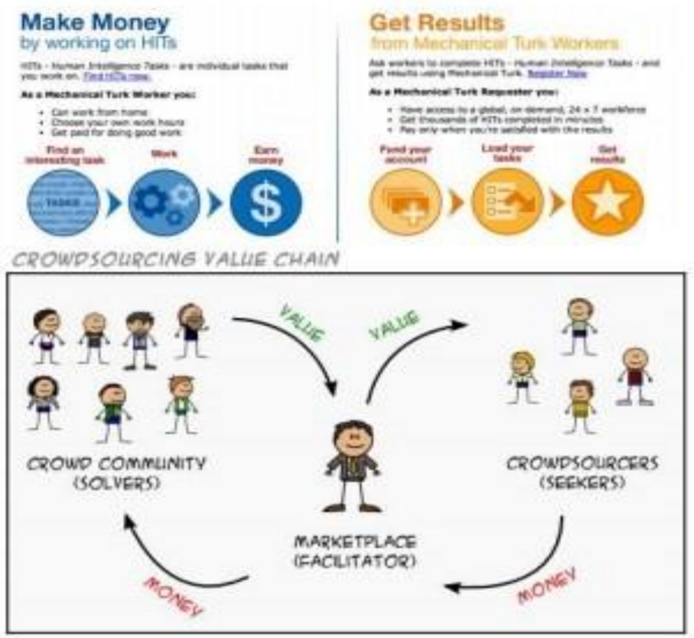
Dr. Bo Han
Trustworthy Machine Learning Group
Computer Science Department, Hong Kong Baptist University
https://bhanml.github.io/

Overview of This Tutorial

- Part I: Why and What Noisy Labels
- Part II: Current Progress and Tutorial Perspectives
- Part III: Training Perspective
- Part IV: Data Perspective
- Part V: Regularization Perspective
- Part VI: Future Directions

Part I: Why Noisy Labels

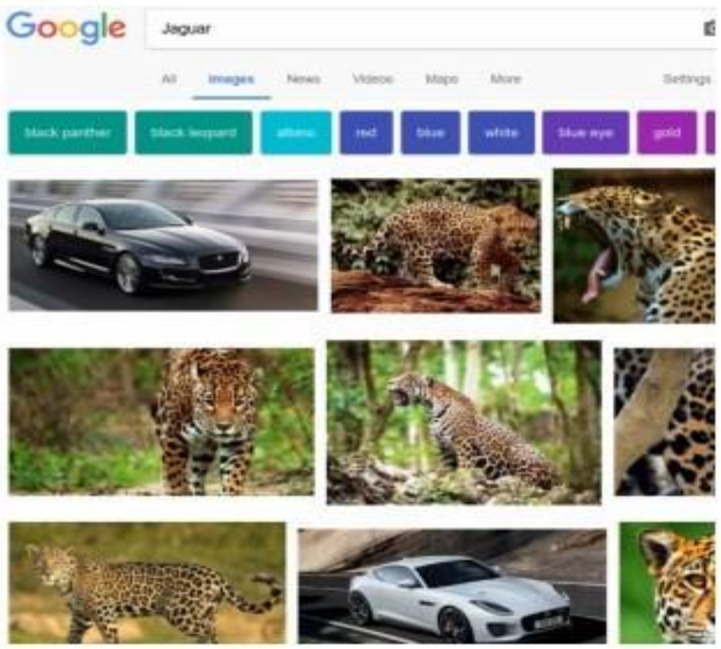
Active label collection



In crowdsourcing,
labels are from non-experts

(Credit to Amazon)

Passive label collection



In web search,
labels are from users' clicks

(Credit to Google)

Why Noisy Labels

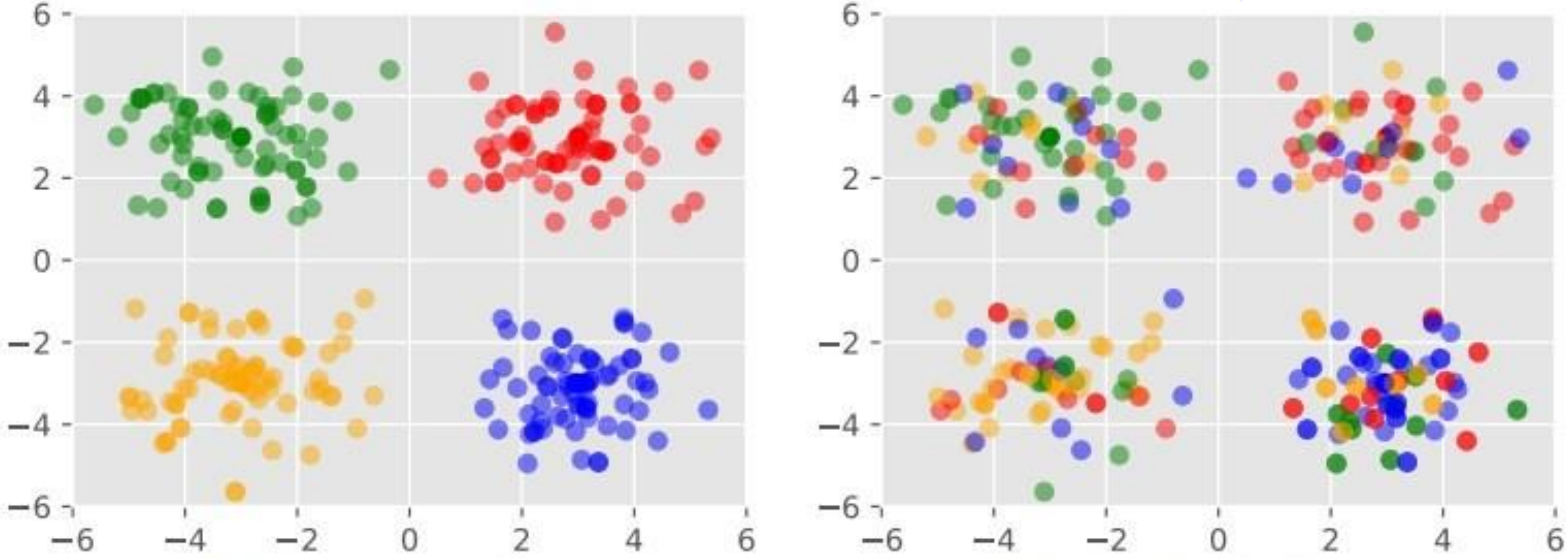
(Credit to Clothing1M)

(Credit to Outlook)

What are Noisy Labels

$R_{\ell,D}(f_\theta) := \mathbb{E}_{(x,y)\sim D}[\ell(f_\theta(x), y)]$

$\hat{R}_{\tilde{\ell},\bar{D}}(f_\theta) := 1/N \sum_{i=1}^{N} \tilde{\ell}(f_\theta(x_i), \bar{y}_i)$



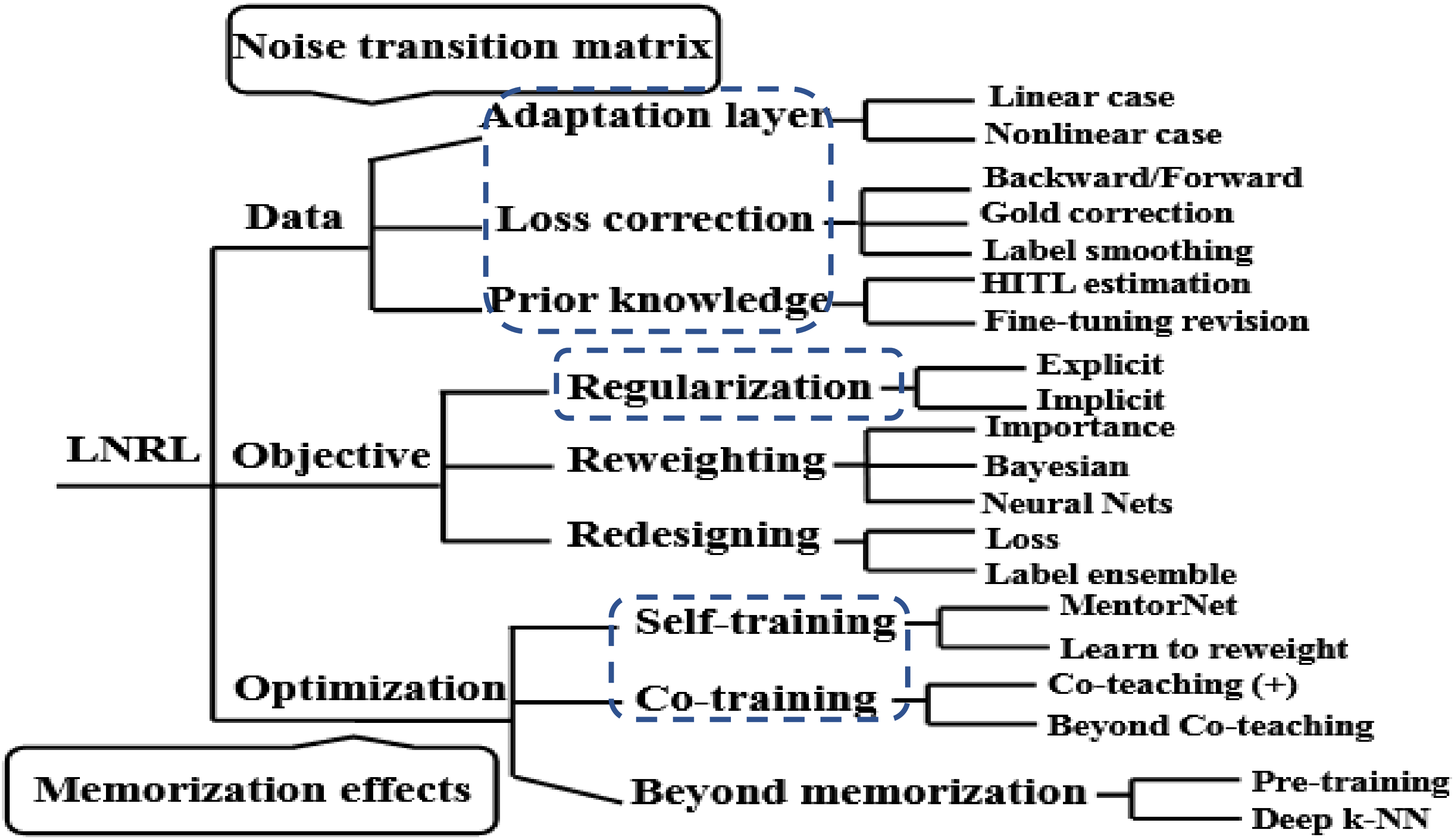
Clean training data
$(x_i, y_i)_{i=1}^{n} \sim p(x, y)$

Noisy training data
$(x_i, \tilde{y}_i)_{i=1}^{n} \sim p(x, \tilde{y})$

?

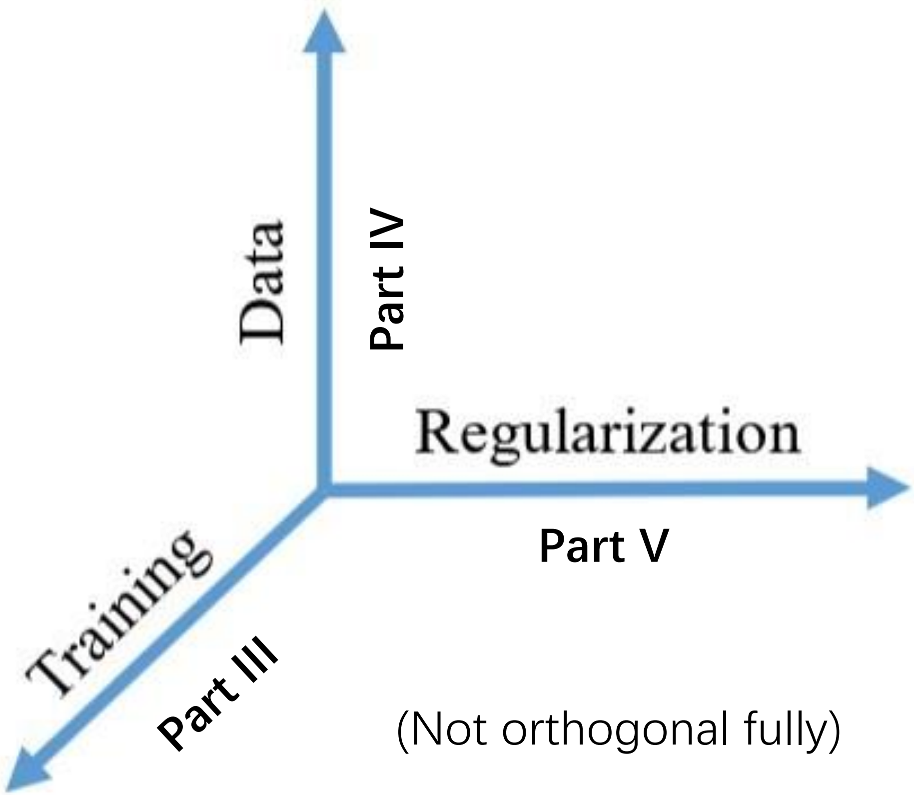
(Credit to Dr. Gang Niu)

Part II: Current Progress



B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama.
A Survey of Label-noise Representation Learning: Past, Present and Future. *arXiv preprint: 2011.04406*, 2020.

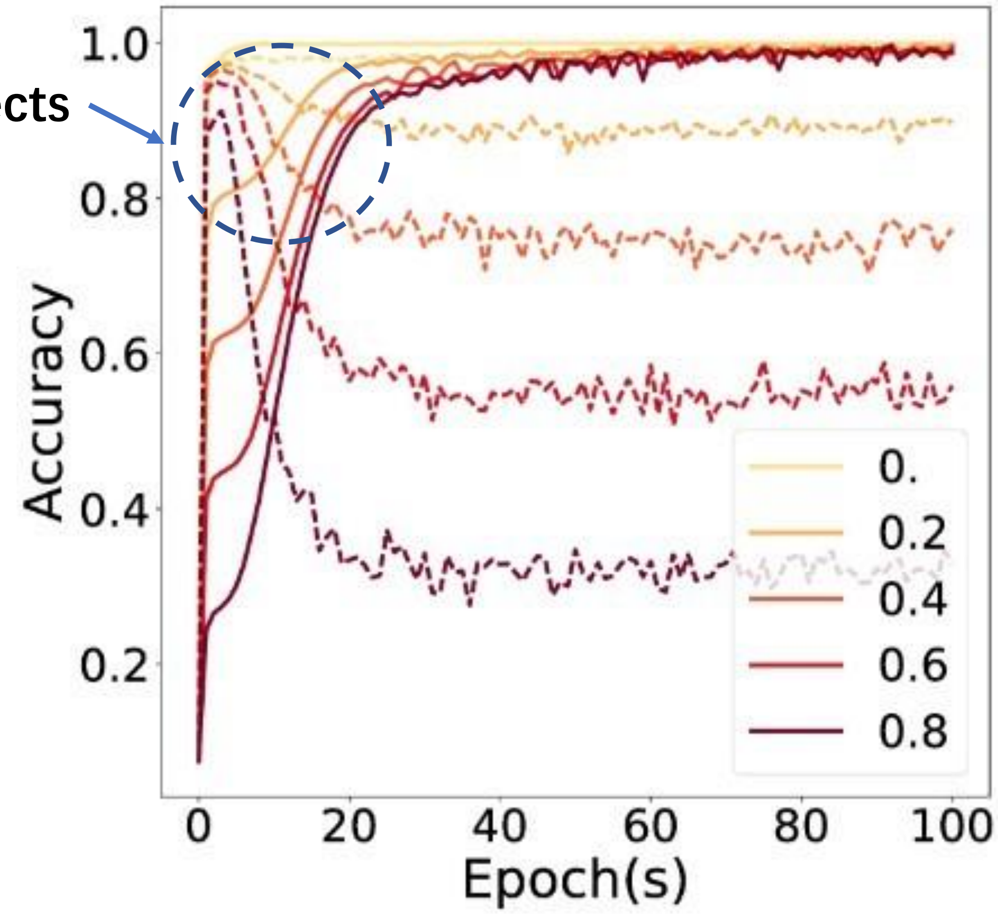
Tutorial Perspectives



(Not orthogonal fully)

Part III: Training Perspective

Memorization Effects



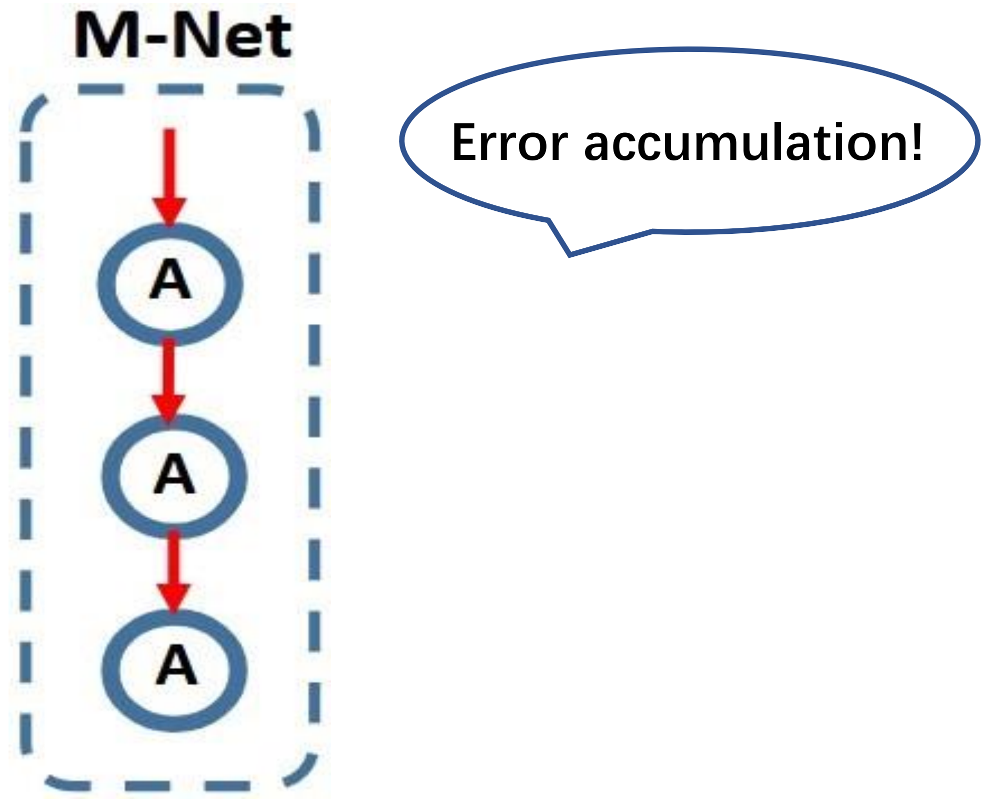
D. Arpit et al. A Closer Look at Memorization in Deep Networks. In *ICML*, 2017.

Training on Selected Samples

Algorithm 1 General procedure on using sample selection to combat noisy labels.

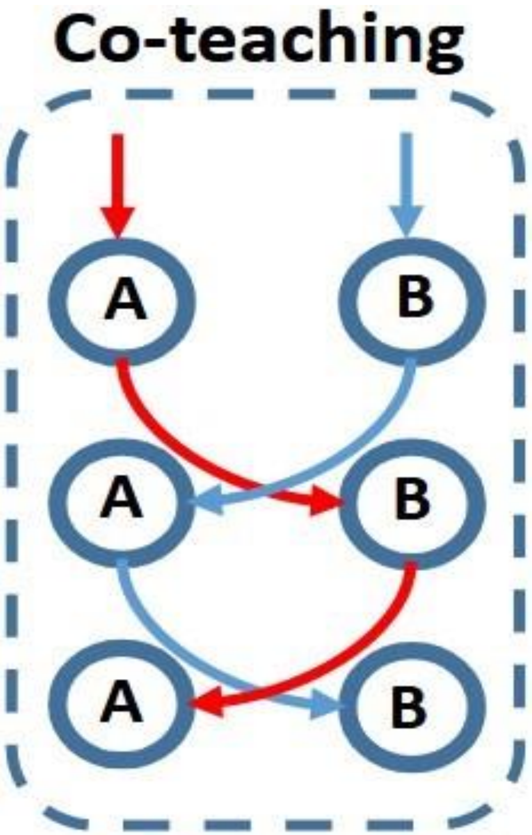
for $t = 0, \dots, T-1$ **do**
  draw a mini-batch $\bar{\mathcal{D}}$ from $\mathcal{D}$;
  select $R(t)$ small-loss samples $\bar{\mathcal{D}}_f$ from $\bar{\mathcal{D}}$ based on network's predictions;
  update network parameter using $\bar{\mathcal{D}}_f$;
end for

Self-teaching (MentorNet)



L. Jiang et al. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Data. In *ICML*, 2018.

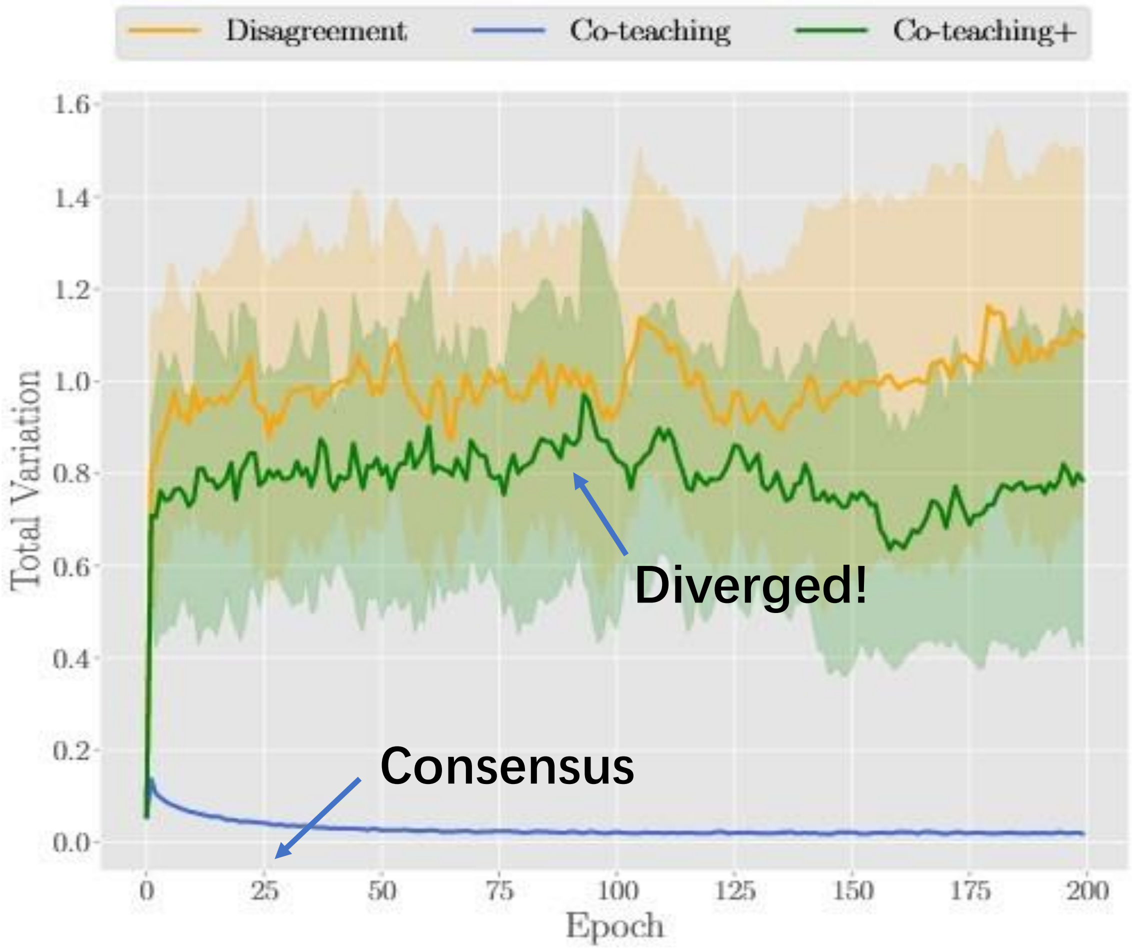
Co-teaching



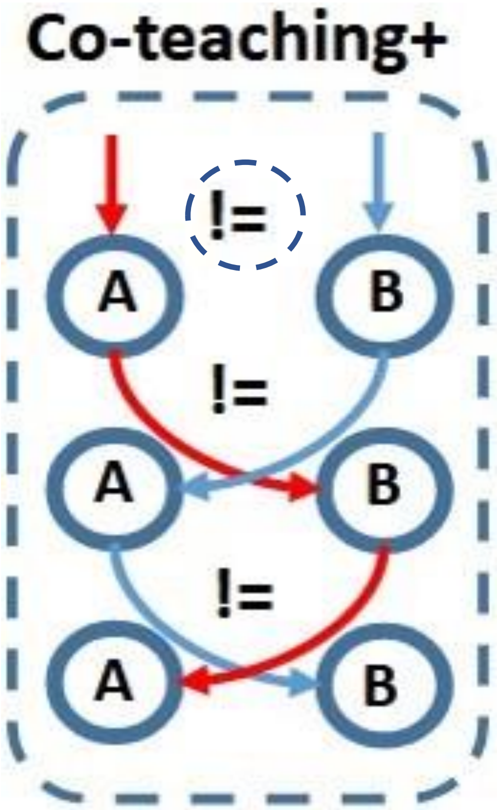
Find “bugs” by peers

B. Han et al. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *NeurIPS*, 2018.

Divergence Matters



Co-teaching+



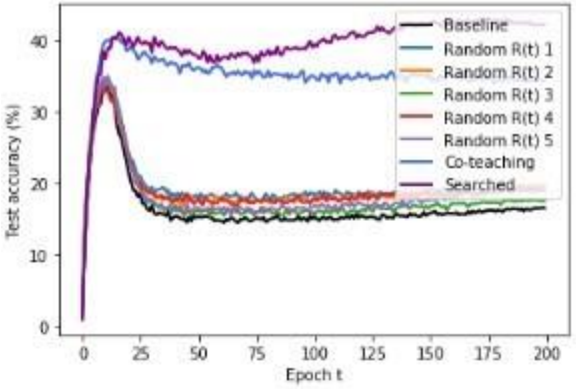
Divergence meeting Co-teaching

X. Yu et al. How does Disagreement Help Generalization against Label Corruption? In *ICML*, 2019.

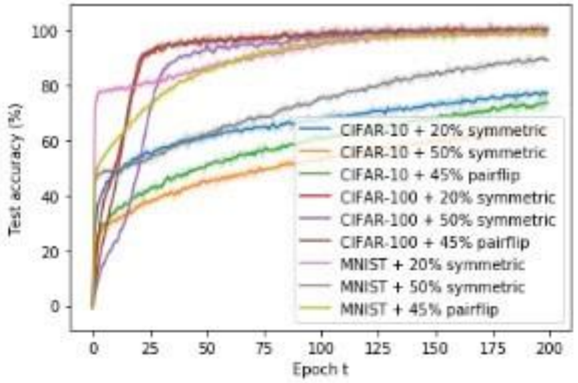
IJCAI 2021 MONTREAL

Rethinking R(t)

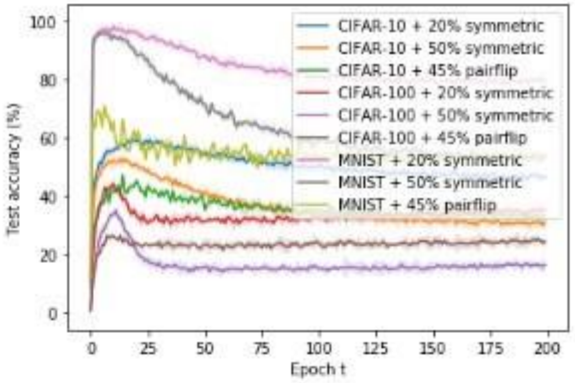
$$R(t) = 1 - \tau \cdot \min((t/t_k)^c, 1)$$



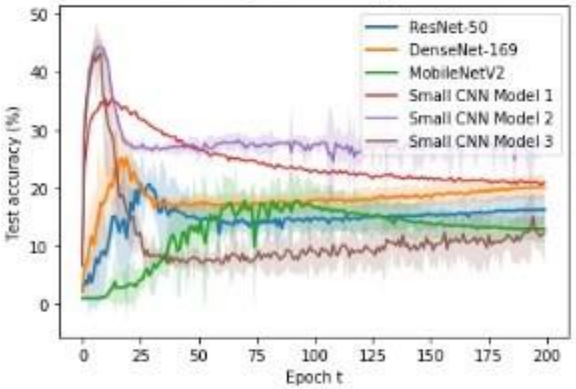
(a) Impact of $R(t)$.



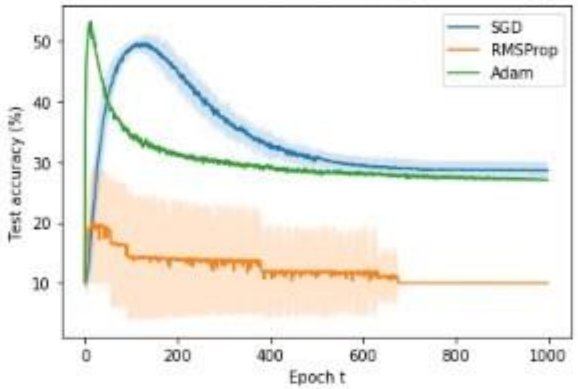
(b) Different data sets (training accuracy).



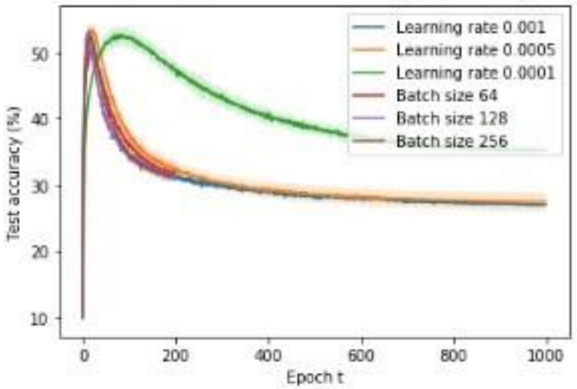
(c) Different data sets (testing accuracy).



(d) Different architectures.



(e) Different optimizers.

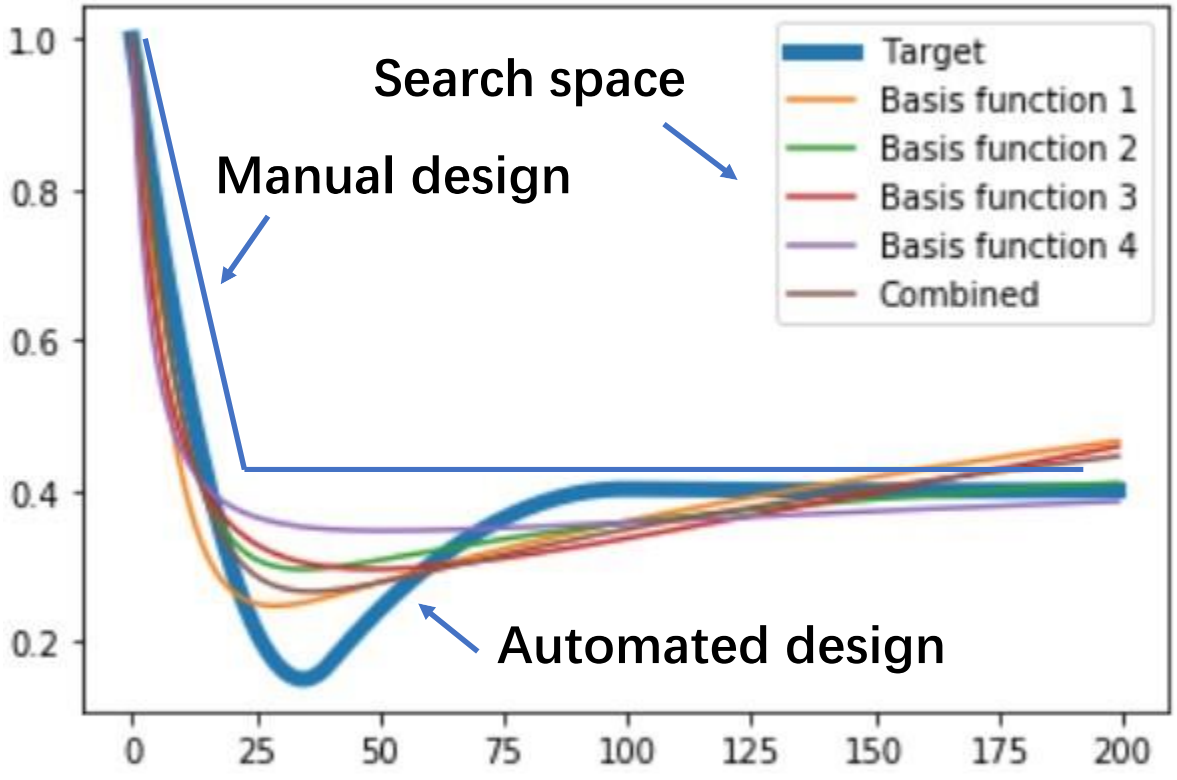


(f) Different optimizer settings.

S2E: Searching to Exploit

$$R^* = \arg\min_{R(\cdot)\in\mathcal{F}} \mathcal{L}_{\text{val}}(f(\boldsymbol{w}^*; R), \mathcal{D}_{\text{val}}),$$
$$\text{s.t. } \boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \mathcal{L}_{\text{tr}}(f(\boldsymbol{w}; R), \mathcal{D}_{\text{tr}}).$$

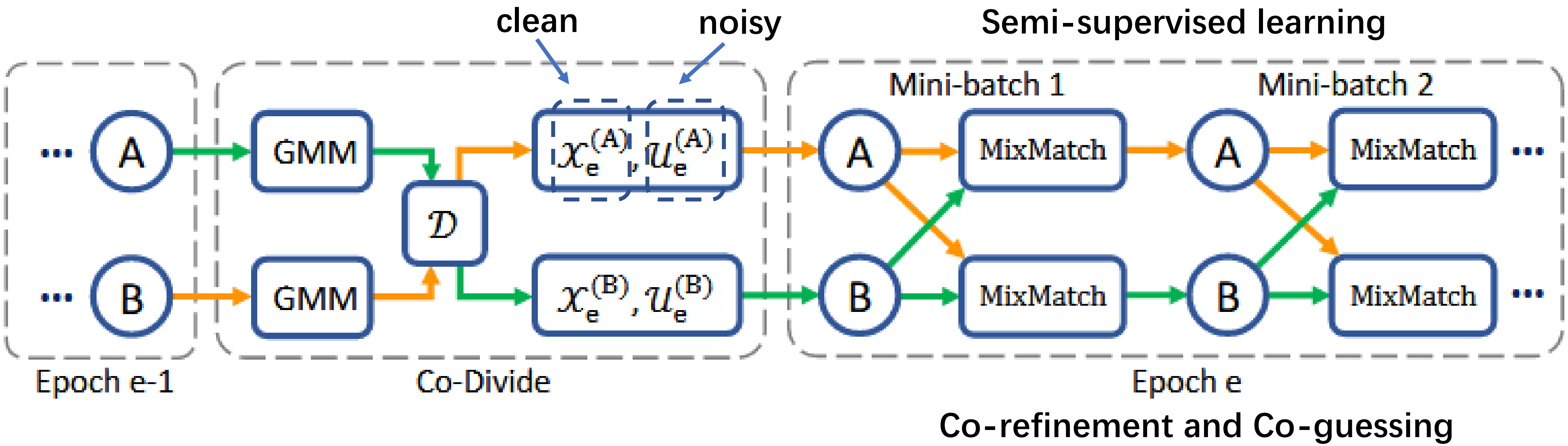
Bi-level Optimization



Q. Yao et al. Searching to Exploit Memorization Effect in Learning from Noisy Labels. In *ICML*, 2020.

DivideMix

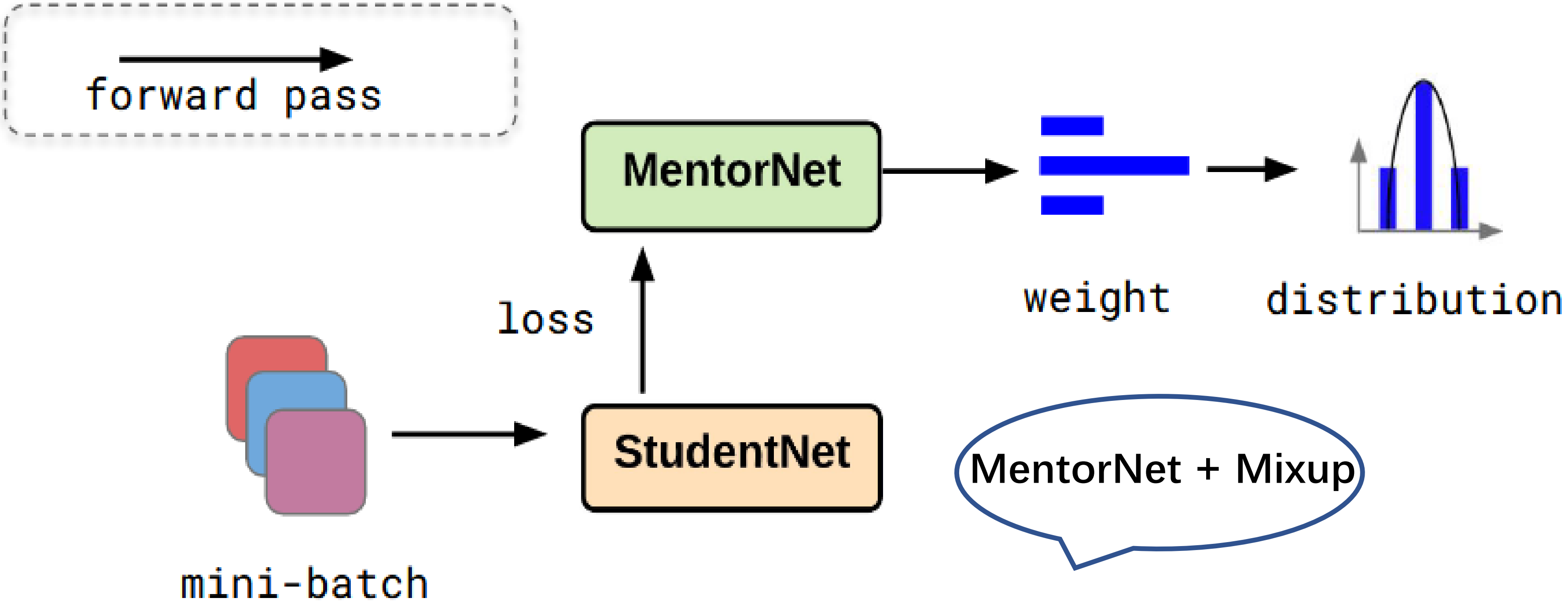
IJCAI 2021 MONTREAL



J. Li et al. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *ICLR*, 2020.

MentorMix

Weight → Sample → Mixup → Weight



L. Jiang et al. Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels. In *ICML*, 2020.

Summary

- **Memorization effect** in deep learning is new and important.
- MentorNet and Co-teaching series are developed.
- Many **applications** have leveraged Co-teaching series.

B. Han et al. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *NeurIPS*, 2018.

Code: https://github.com/bhanML/Co-teaching

Part IV: Data Perspective

$$
\begin{array}{cc}
 & \begin{array}{cccc} & \text{Cat} & \text{Wolf} & \tilde{\boldsymbol{y}} \end{array} \\
\begin{array}{c} \text{Dog} \\ \\ \boldsymbol{y} \\ \\ \end{array} &
\begin{bmatrix}
1-\tau & \frac{\tau}{n-1} & \cdots & \frac{\tau}{n-1} \\
\frac{\tau}{n-1} & 1-\tau & & \frac{\tau}{n-1} \\
\vdots & & \ddots & \vdots \\
\frac{\tau}{n-1} & \frac{\tau}{n-1} & \cdots & 1-\tau
\end{bmatrix}
\end{array}
$$

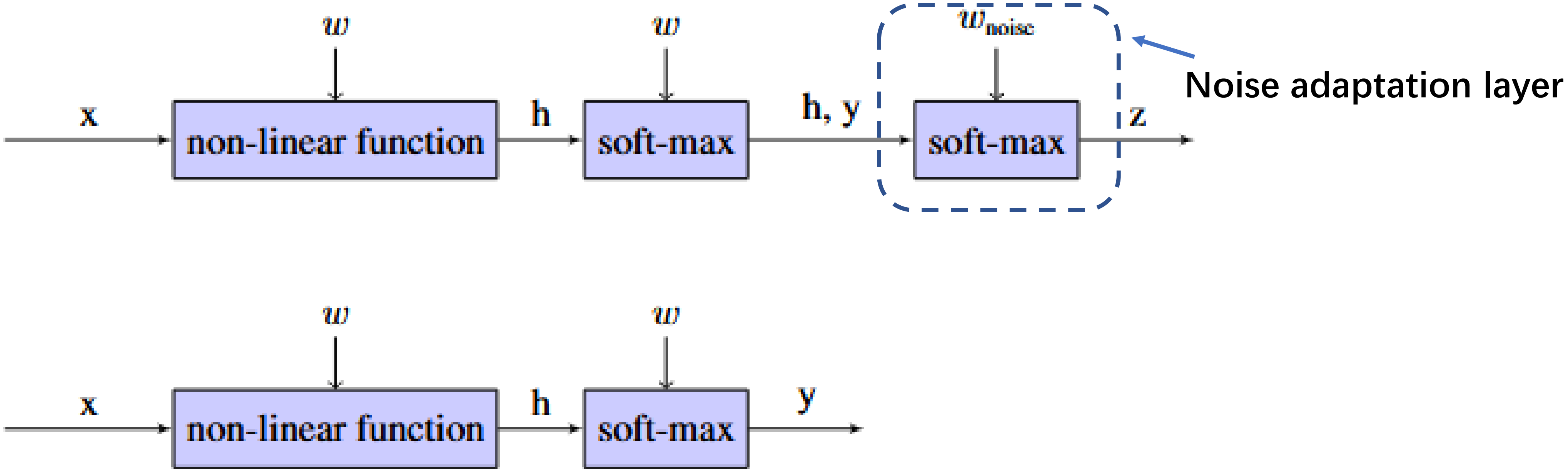
(a) Sym-flipping.

$$
\begin{bmatrix}
1-\tau & \tau & 0 & 0 \\
0 & 1-\tau & \tau & 0 \\
\vdots & & \ddots & \vdots \\
0 & & & \tau \\
\tau & 0 & \cdots & 1-\tau
\end{bmatrix}
$$

(b) Pair-flipping.

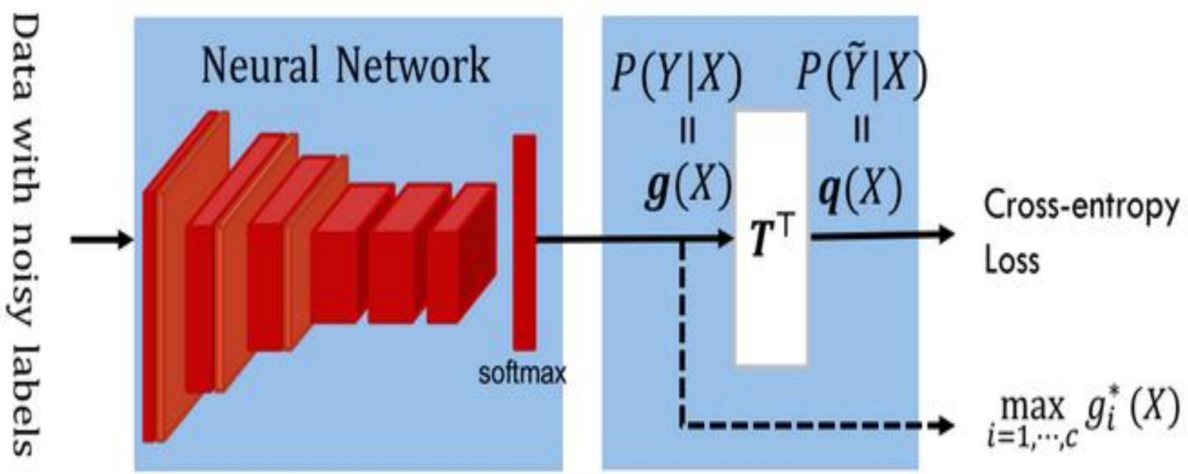
Noise Transition Matrix

Adaptation layer



J. Goldberger et al. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

Forward Correction



(Credit to Dr. Tongliang Liu)

Theorem 2. *(Forward Correction, Theorem 1 in [22]) Suppose that the label transition matrix T is non-singular, where $T_{ij} = p(\bar{y} = j | y = i)$ given that corrupted label $\bar{y} = j$ is flipped from clean label $y = i$. Given loss ℓ and network function f, Forward Correction is defined as*

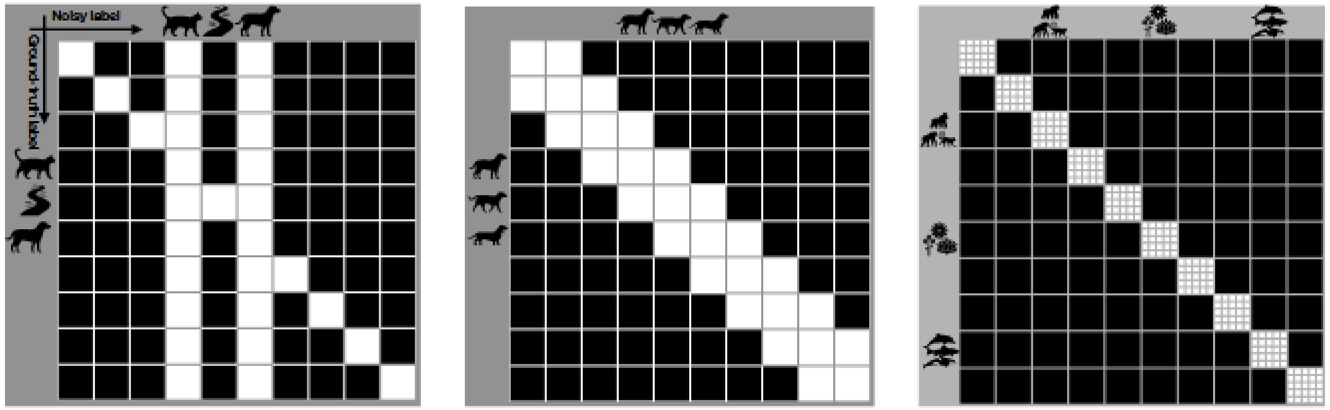
$$\ell^{\rightarrow}(f(x), \bar{y}) = [\ell_{y|T^{\top} f(x)}]_{\bar{y}}, \tag{6}$$

where $\ell_{y|T^{\top} f(x)} = (\ell(T^{\top} f(x), 1), \dots, \ell(T^{\top} f(x), k))$. Then, the minimizer of the corrected loss under the noisy distribution is the same as the minimizer of the orginal loss under the clean distribution, namely,

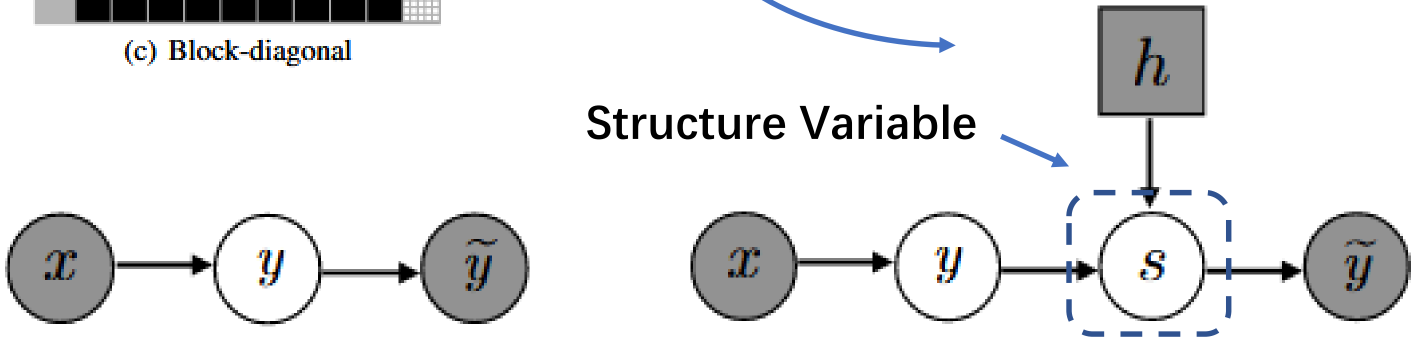
$$\arg\min_{f} \mathbb{E}_{x,\bar{y}} \ell^{\rightarrow}(f(x), \bar{y}) = \arg\min_{f} \mathbb{E}_{x,y} \ell(f(x), y). \tag{7}$$

G. Patrini et al. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *CVPR*, 2017.

Masking



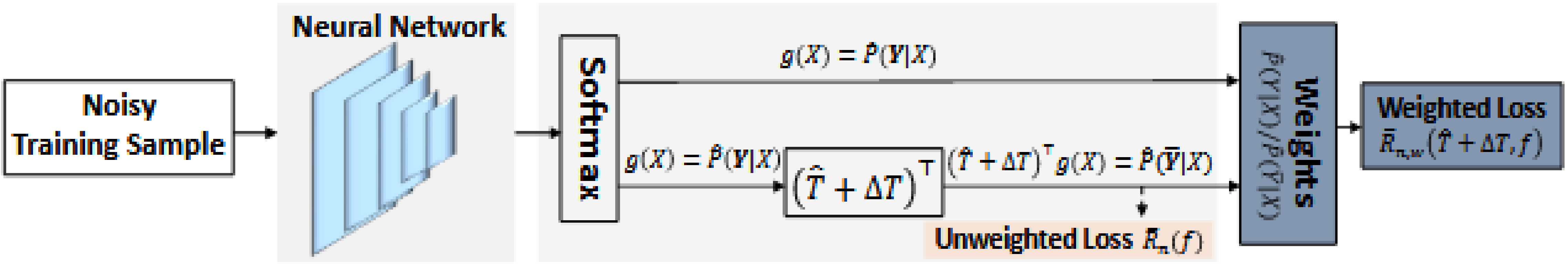
(a) Column-diagonal (b) Tri-diagonal (c) Block-diagonal



(a) Benchmark model. (b) MASKING model.

B. Han et al. Masking: A New Perspective of Noisy Supervision. In *NeurIPS*, 2018.

Fine-tuning

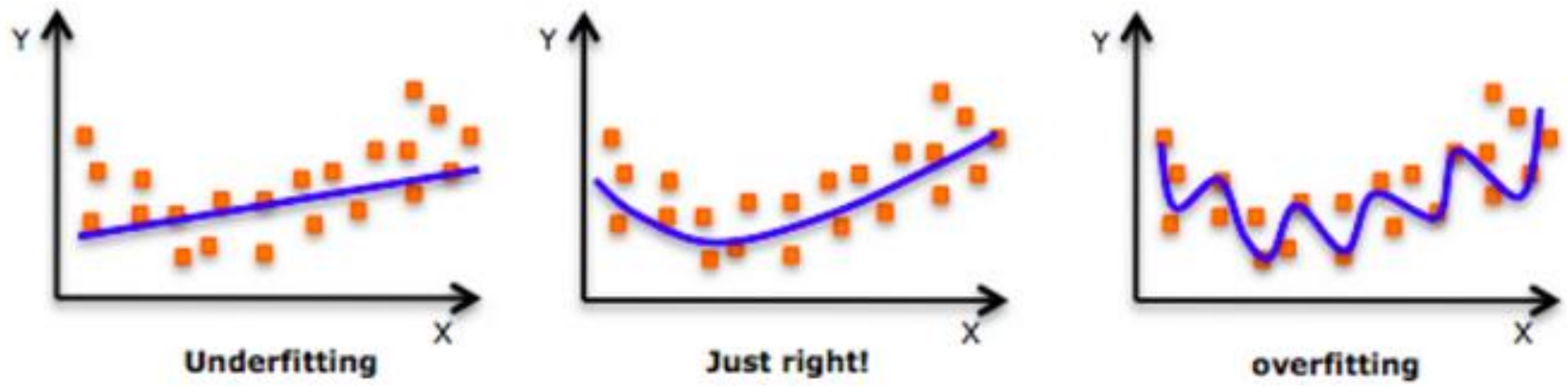


X. Xiao et al. Are Anchor Points Really Indispensable in Label-noise Learning? In *NeurIPS*, 2019.

Summary

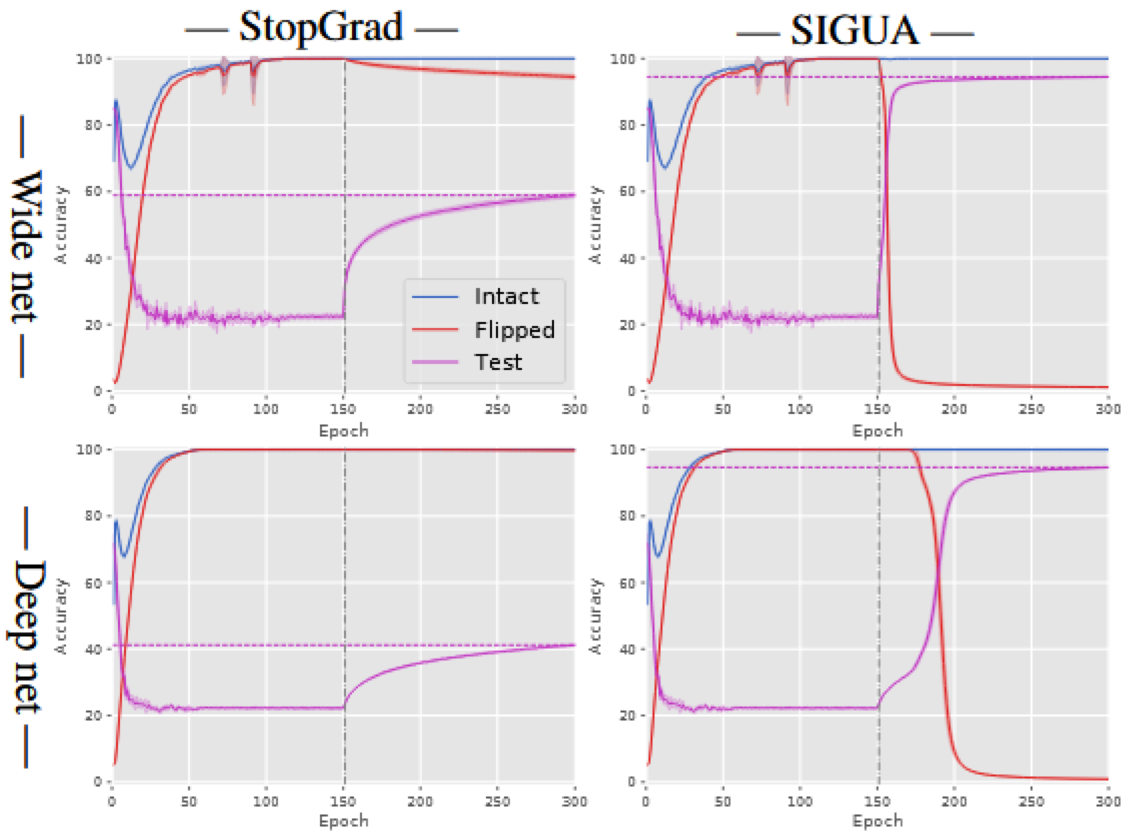
- **Noise transition matrix** is the key in data perspective.
- A potential direction is how to estimate this matrix **easily**.
- Another potential direction is how to leverage this matrix **effectively**.

Part V: Regularization Perspective



(Credit to Analytics Vidhya)

SIGUA



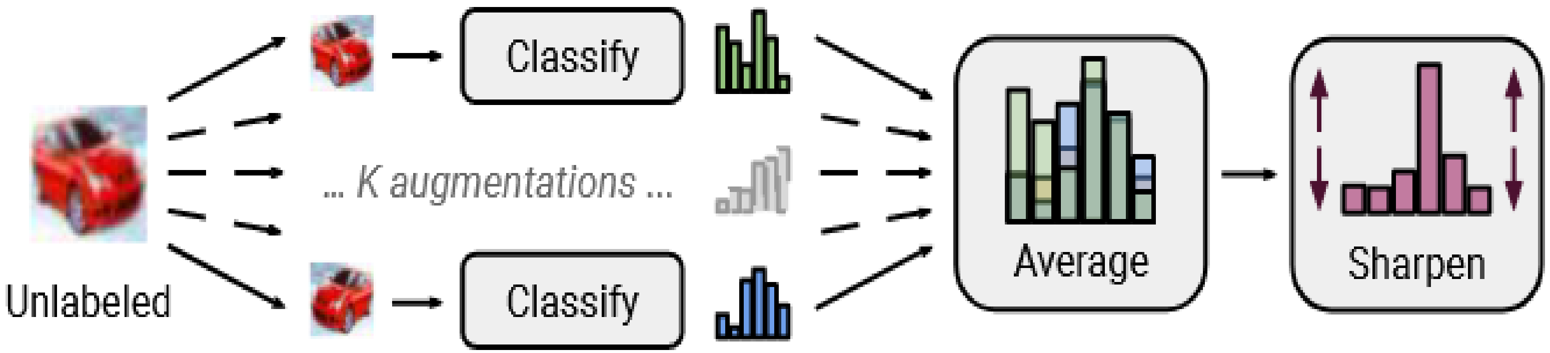
Algorithm 1 SIGUA-prototype (in a mini-batch).

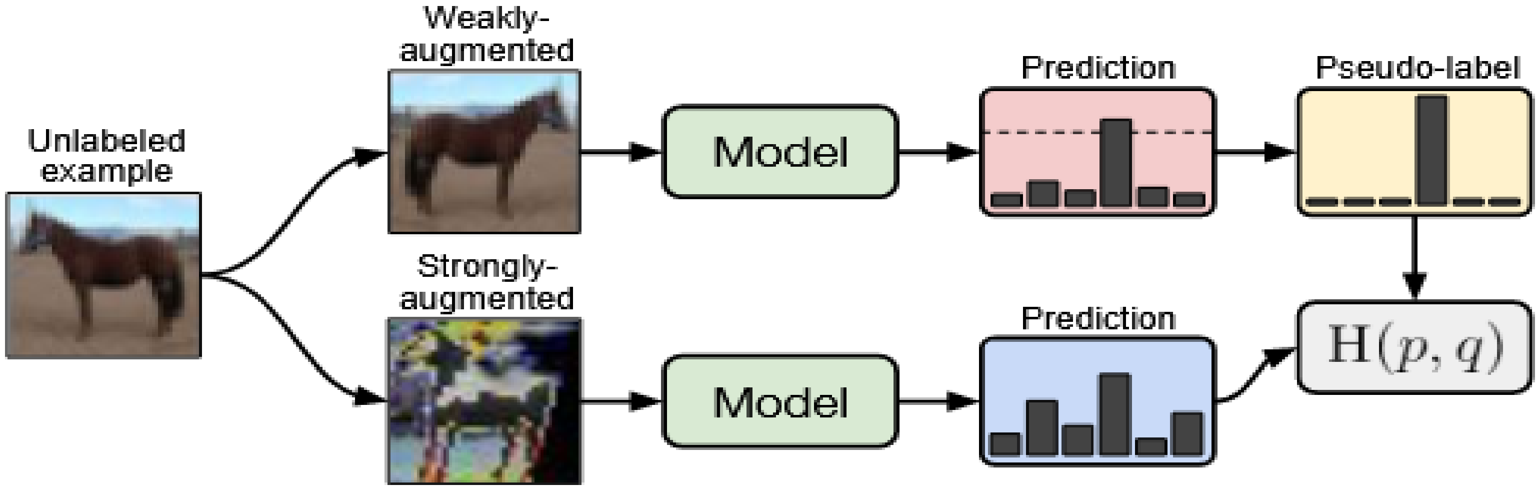
Require: base learning algorithm $\mathfrak{B}$, optimizer $\mathfrak{O}$, mini-batch $\mathcal{S}_\mathrm{b} = \{(x_i, \tilde{y}_i)\}_{i=1}^{n_\mathrm{b}}$ of batch size n_b, current model f_θ where θ holds the parameters of f, good- and bad-data conditions $\mathfrak{C}_\mathrm{good}$ and $\mathfrak{C}_\mathrm{bad}$ for $\mathfrak{B}$, underweight parameter γ such that $0 \le \gamma \le 1$

$\{\ell_i\}_{i=1}^{n_\mathrm{b}} \leftarrow \mathfrak{B}$.forward($f_\theta, \mathcal{S}_\mathrm{b}$) # forward pass
$\ell_\mathrm{b} \leftarrow 0$ # initialize loss accumulator
for $i = 1, \dots, n_\mathrm{b}$ **do**
  if $\mathfrak{C}_\mathrm{good}(x_i, \tilde{y}_i)$ **then**
    $\ell_\mathrm{b} \leftarrow \ell_\mathrm{b} + \ell_i$ # accumulate loss positively
  else if $\mathfrak{C}_\mathrm{bad}(x_i, \tilde{y}_i)$ **then** ← Gradient Ascent
    $\ell_\mathrm{b} \leftarrow \ell_\mathrm{b} - \gamma\ell_i$ # accumulate loss negatively
  end if # ignore any uncertain data
end for
$\ell_\mathrm{b} \leftarrow \ell_\mathrm{b} / n_\mathrm{b}$ # average accumulated loss
$\nabla_\theta \leftarrow \mathfrak{B}$.backward($f_\theta, \ell_\mathrm{b}$) # backward pass
$\mathfrak{O}$.step(∇_θ) # update model

B. Han et al. SIGUA: Forgetting May Make Learning with Noisy Labels More Robust. In *ICML*, 2020.

MixMatch & FixMatch





D. Berthelot et al. MixMatch: A Holistic Approach to Semi-supervised Learning. In *NeurIPS*, 2019.

K. Sohn et al. FixMatch: Simplifying Semi-supervised Learning with Consistency and Confidence. In *NeurIPS*, 2020.

Bootstrapping

target prediction

$$\ell_{\text{soft}}(q,t) = \sum_{k=1}^{L} [\beta t_k + (1-\beta) q_k] \log(q_k)$$

$$\ell_{\text{hard}}(q,t) = \sum_{k=1}^{L} [\beta t_k + (1-\beta) z_k] \log(q_k)$$

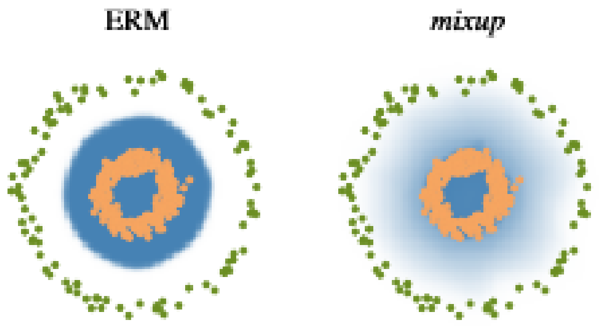
S. Reed et al. Training Deep Neural Networks on Noisy Labels with Bootstrapping. In *ICLR Workshop*, 2015.

Mixup

```python
# y1, y2 should be one-hot vectors
for (x1, y1), (x2, y2) in zip(loader1, loader2):
    lam = numpy.random.beta(alpha, alpha)
    x = Variable(lam * x1 + (1. - lam) * x2)
    y = Variable(lam * y1 + (1. - lam) * y2)
    optimizer.zero_grad()
    loss(net(x), y).backward()
    optimizer.step()
```

interpolation

(a) One epoch of *mixup* training in PyTorch.



(b) Effect of *mixup* ($\alpha = 1$) on a toy problem. Green: Class 0. Orange: Class 1. Blue shading indicates $p(y = 1|x)$.

H. Zhang et al. Mixup: Beyond Empirical Risk Minimization. In *ICLR*, 2018.

Summary

- Regularization is very popular for **semi-supervised learning**.
- Explicit regularization is in the level of **objective function**.
- Implicit regularization is in the level of **algorithm** and **data**.

Part VI: Future Directions

IJCAI 2021 MONTREAL

A Survey of Label-noise Representation Learning: Past, Present and Future

Bo Han, Quanming Yao, Tongliang Liu, Gang Niu,
Ivor W. Tsang, James T. Kwok, *Fellow, IEEE* and Masashi Sugiyama

Abstract—Classical machine learning implicitly assumes that labels of the training data are sampled from a clean distribution, which can be too restrictive for real-world scenarios. However, statistical-learning-based methods may not train deep learning models robustly with these noisy labels. Therefore, it is urgent to design Label-Noise Representation Learning (LNRL) methods for robustly training deep models with noisy labels. To fully understand LNRL, we conduct a survey study. We first clarify a formal definition for LNRL from the perspective of machine learning. Then, via the lens of learning theory and empirical study, we figure out why noisy labels affect deep models' performance. Based on the theoretical guidance, we categorize different LNRL methods into three directions. Under this unified taxonomy, we provide a thorough discussion of the pros and cons of different categories. More importantly, we summarize the essential components of robust LNRL, which can spark new directions. Lastly, we propose possible research directions within LNRL, such as new datasets, instance-dependent LNRL, and adversarial LNRL. We also envision potential directions beyond LNRL, such as learning with feature-noise, preference-noise, domain-noise, similarity-noise, graph-noise and demonstration-noise.

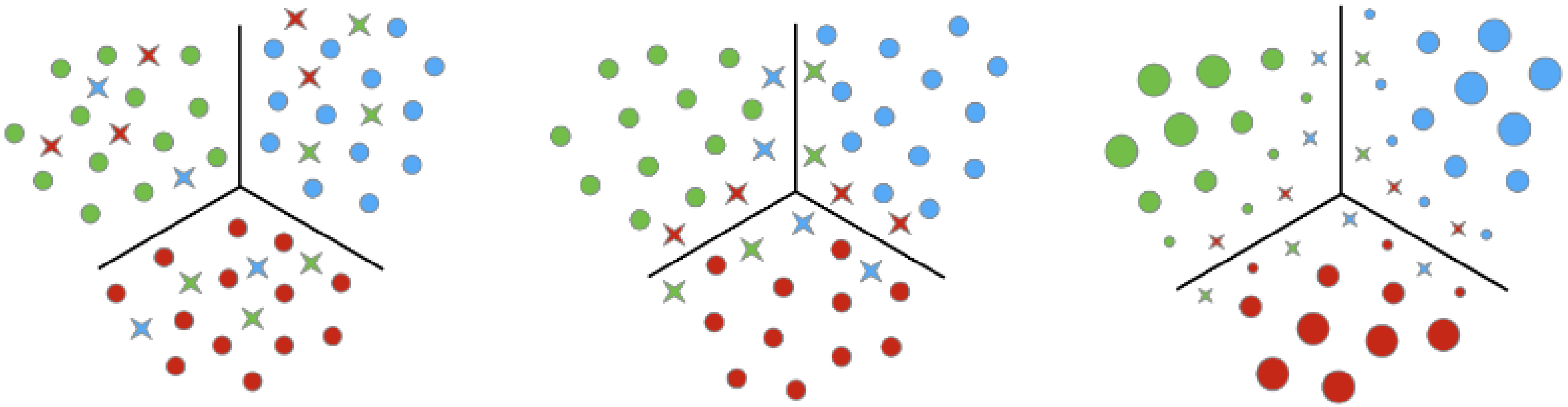
Index Terms—Machine Learning, Representation Learning, Weakly Supervised Learning, Label-noise Learning, Noisy Labels.

B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama.
A Survey of Label-noise Representation Learning: Past, Present and Future. *arXiv preprint: 2011.04406*, 2020.

New Datasets

L. Jiang et al. Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels. In *ICML*, 2020.

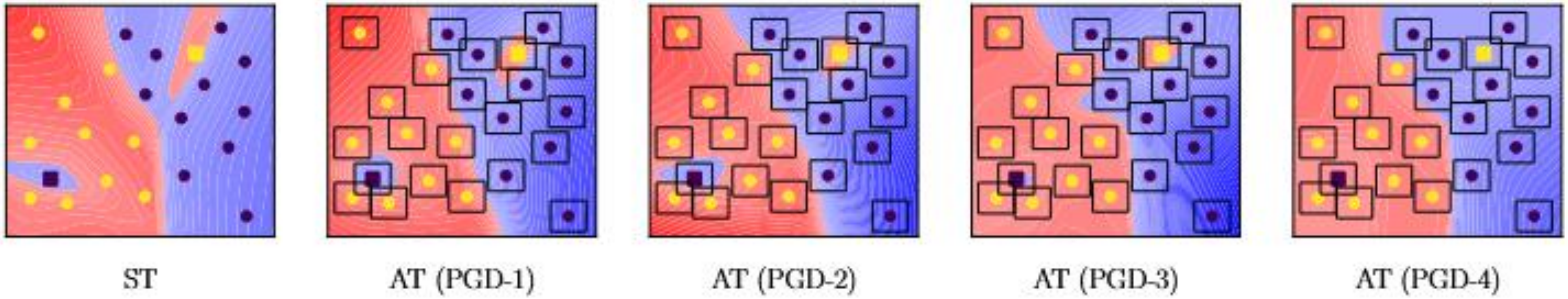
Instance-dependent LNRL



(a) Class-conditional noise. (b) Instance-dependent noise (boundary-consistent noise). (c) Confidence-scored instance-dependent noise.

A. Berthon et al. Confidence Scores Make Instance-dependent Label-noise Learning Possible. In *ICML*, 2021.

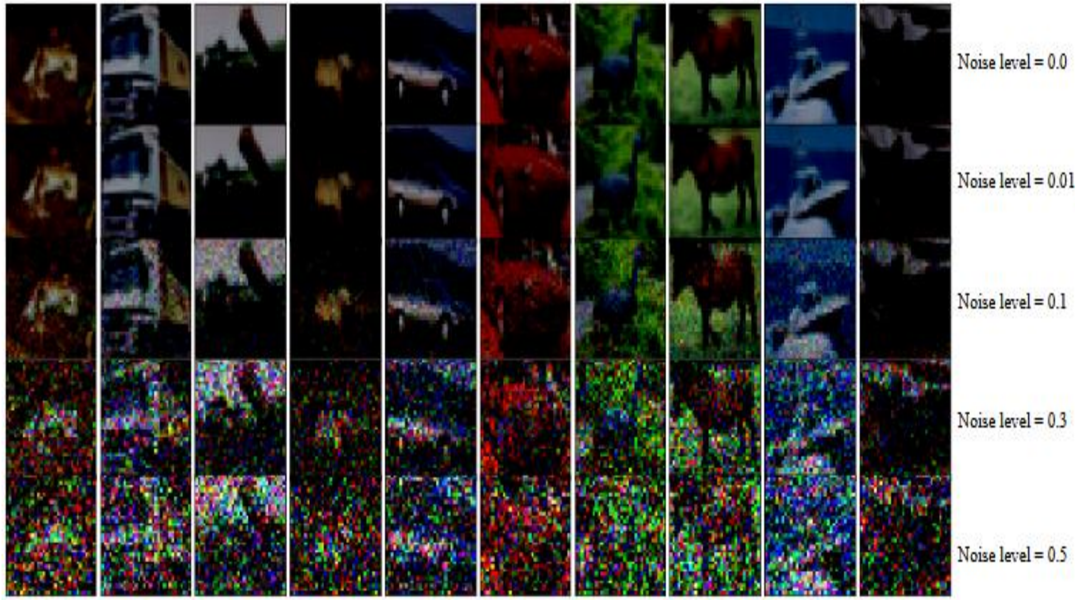
Adversarial LNRL



weak → **strong**

J. Zhu et al. Understanding the Interaction of Adversarial Training with Noisy Labels. *arXiv preprint*, 2021.

Noisy Feature



Image

video games good for children computer games can promote problem-solving and team-building in children, say games industry experts. (Noise level = 0.0)

vedeo games good for dhildlenzcospxter games can iromote problem-sorvtng and teai-building in children, sby games industry experts. (Noise level = 0.1)

video nawvs zgood foryxhilqretngomvumer games cahcprocotubpnoblex-szbvina and tqlmmbuaddiagjin whipdren, saywgsmes ildustry exmrrts. (Noise level = 0.3)

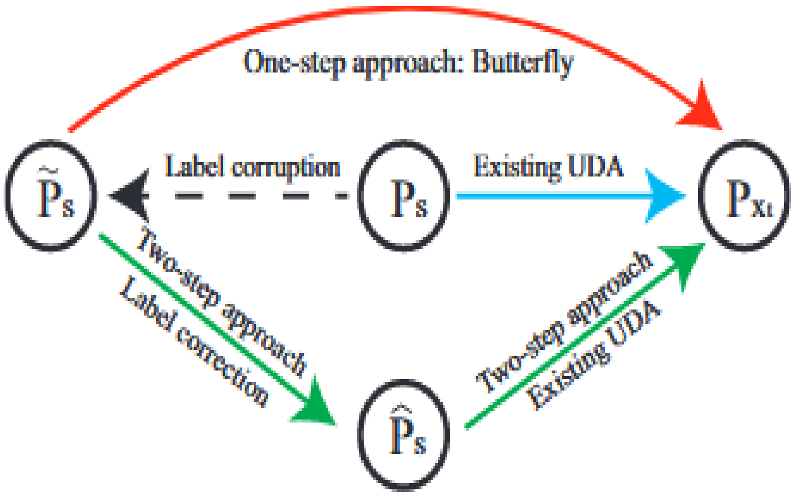
tmdeo gakec jgopd brr cgildrenjcoogwdeh lxdeu vanspromote xrobkeh-svlkieo and termwwuojvinguinfcojbdses, sacosamlt cndgstoyaagpbrus. (Noise level = 0.5)

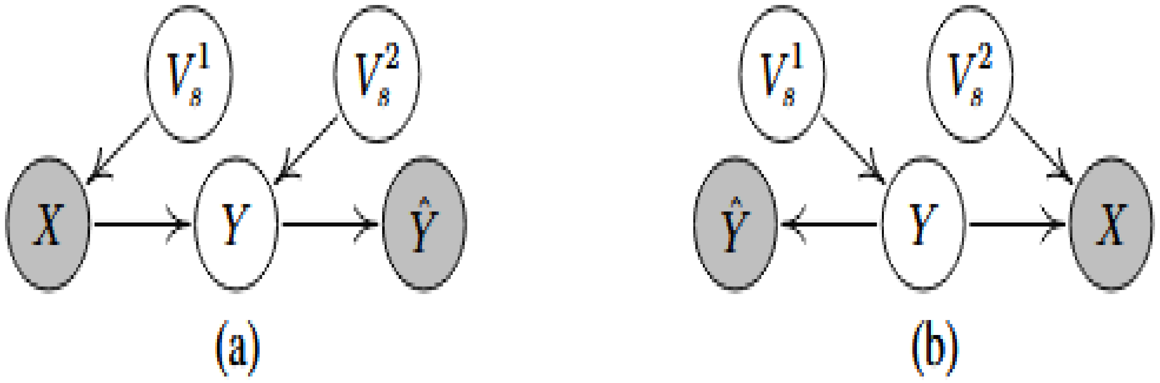
vizwszgbrwjtguihcxfoatbhivrrwvq cxmpgugflziwls clfnzrommtohprtblef-solvynx mjnyiaf-gjwlcergwklskqibdtjn,aoty gameshinzustrm oxpertsdm (Noise level = 0.8)

Text

J. Zhang et al. Towards Robust ResNet: A Small Step but a Giant Leap. In *IJCAI*, 2019.

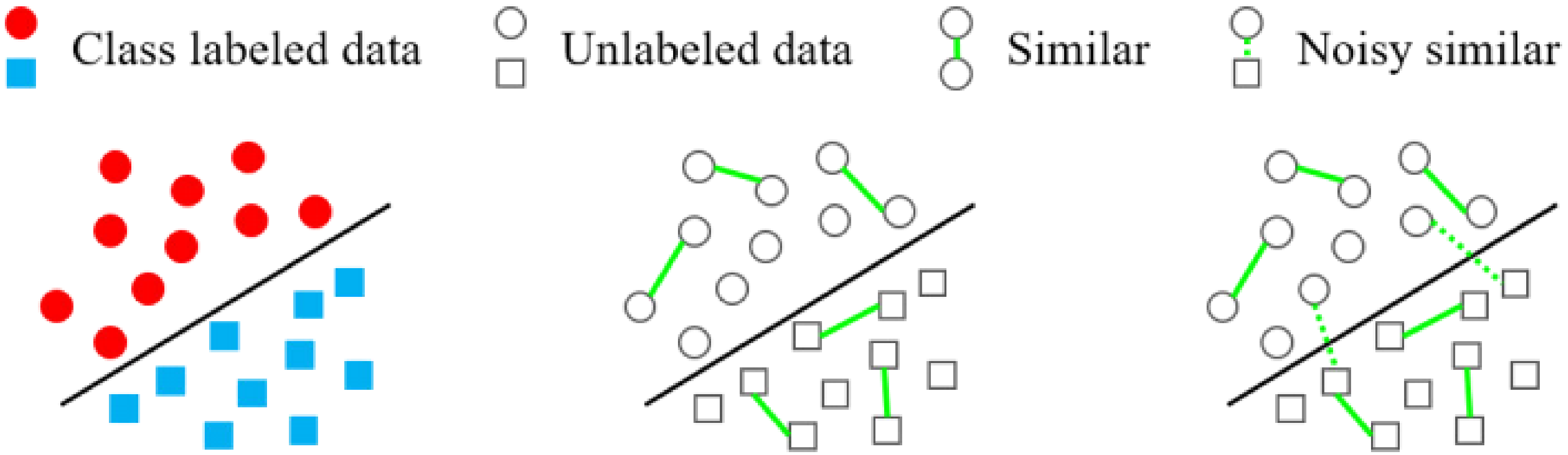
Noisy Domain





F. Liu et al. Butterfly: One-step Approach towards Wildly Unsupervised Domain Adaptation. *arXiv preprint*, 2019.
X. Yu et al. Label-noise Robust Domain Adaptation. In *ICML*, 2020.

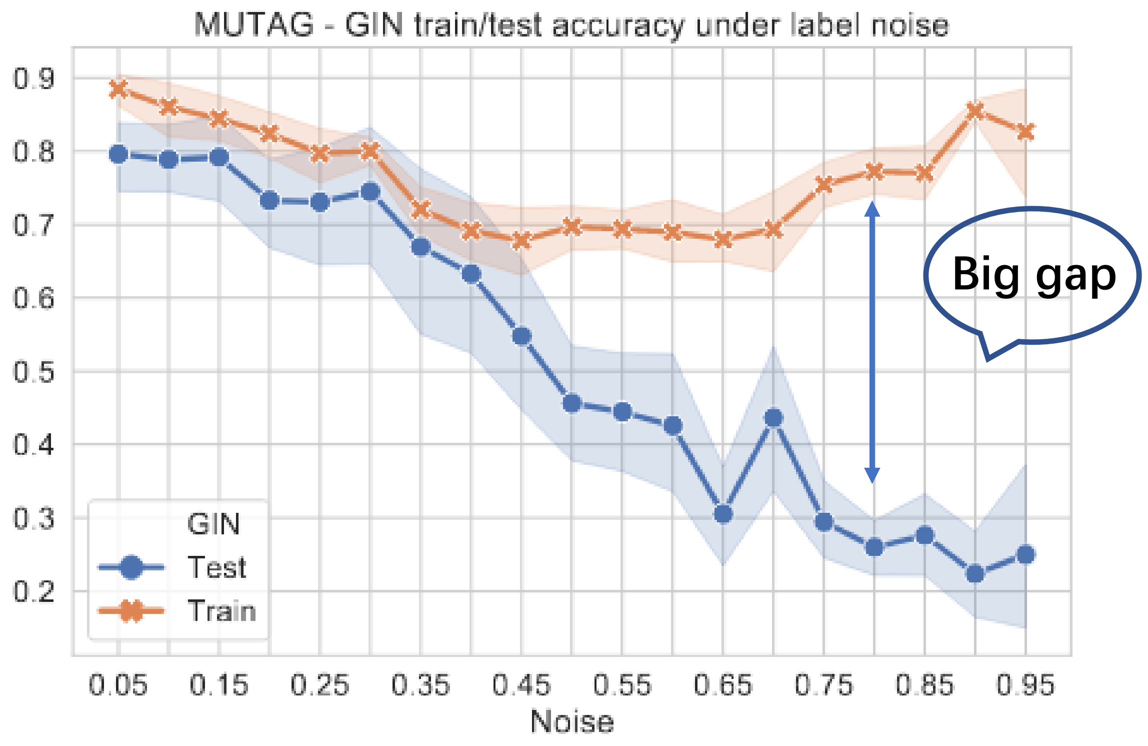
Noisy Similarity



(a) Supervised Classification (b) SU Classification (c) NSU Classification

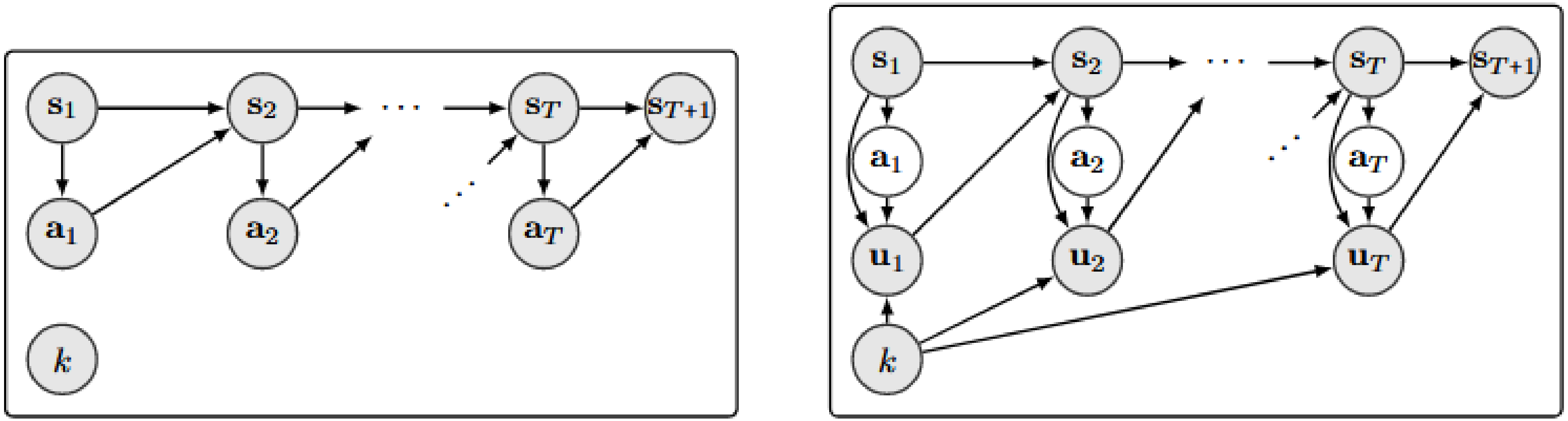
S. Wu et al. Multi-class Classification from Noisy-similarity-labeled Data. arXiv preprint, 2020.

Noisy Graph



Hoang NT et al. Learning Graph Neural Networks with Noisy Labels. In *ICLR Workshop*, 2019.

Noisy Demonstration



(a) Expert demonstrations

(b) Diverse-quality demonstrations

V. Tangkaratt et al. Variational Imitation Learning from Diverse-quality Demonstrations. In *ICML*, 2020.

Noisy Machine Translation

IJCAI 2021 MONTREAL

Chinese-English (ISI bitext)	
Src:	美国提出的报复清单是中国政府绝对不能接受的。
Trg:	And the Chinese side would certainly not accept the unreasonable demands put forward by the Americans concerning the protection of intellectual property rights.
Human:	The revenge list proposed by America will definitely not be accepted by Chinese government.

P. Dakwale et al. Improving Neural Machine Translation Using Noisy Parallel Data through Distillation. In *MT Summit*, 2019.

Conclusions

- Current progress mainly focuses on **class-conditional noise**.
- The new trend focuses on **instance-dependent noise**.
- Besides noisy labels, we should pay more efforts on **noisy data**.