Masking: A New Perspective of Noisy Supervision

Bo Han[*1,2] Jiangchao Yao[*3,1] Gang Niu[2] Mingyuan Zhou[4]
Ivor W. Tsang[1] Ya Zhang[3] Masashi Sugiyama[2,5]

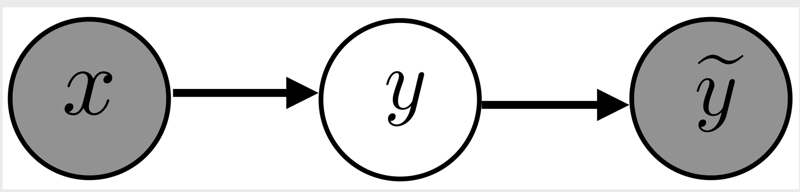
[1]CAI, University of Technology Sydney [2]AIP, RIKEN
[3]Shanghai JiaoTong University [4]The University of Texas at Austin [5]The University of Tokyo

Overview

TL;DR: The masking conveys **human cognition** of invalid class transitions, and speculates the **structure** of the noise transition matrix. Therefore, we only learn the noise **transition probability** to reduce the estimation burden.

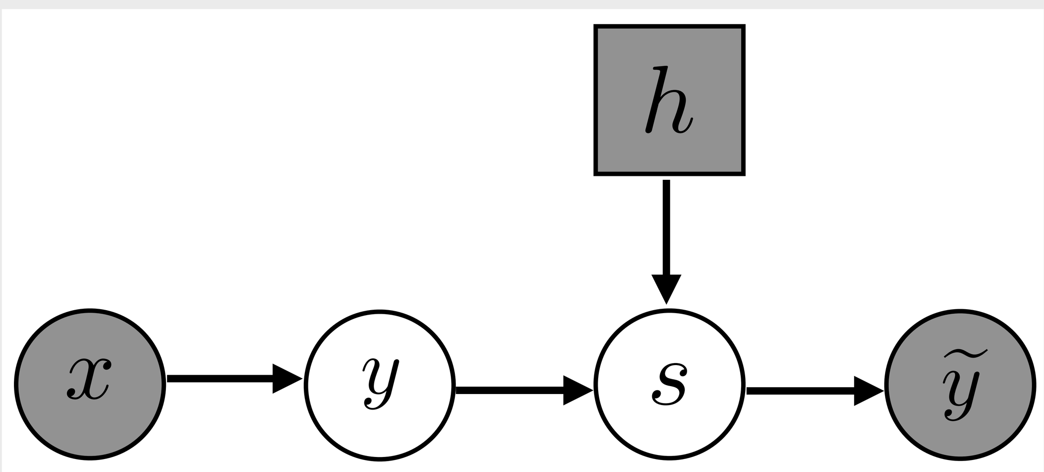
- **Noisy labels** are corrupted from ground-truth labels, which degenerates the robustness of learning models.
- **Deep neural networks** have the high capacity to fit any noisy labels. The solutions are as follows.
  - ⋄ Noise **transition** matrix estimation. E.g., F-correction.
  - ⋄ **Regularization**. E.g., VAT and Mean teacher.
  - ⋄ Training on **selected** samples. E.g., MentorNet.
- We present a human-assisted approach called **Masking** combating with noisy labels.
  - ⋄ Conveying human cognition of invalid class transitions.
  - ⋄ Speculating the structure of the noise transition matrix.
  - ⋄ Deriving a **structure-aware probabilistic model**, which incorporates a **structure prior**.
- Empirical results on **CIFAR10, CIFAR100** with three noise structures and **Clothing1M** demonstrate that, our approach can improve the robustness of classifiers.

Deficiency of Benchmarks



- Independent framework: not for **agnostic** noisy data.
- Unified framework: suffer from **local** minimums.

Structure-aware Probabilistic Model (MASKING)



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

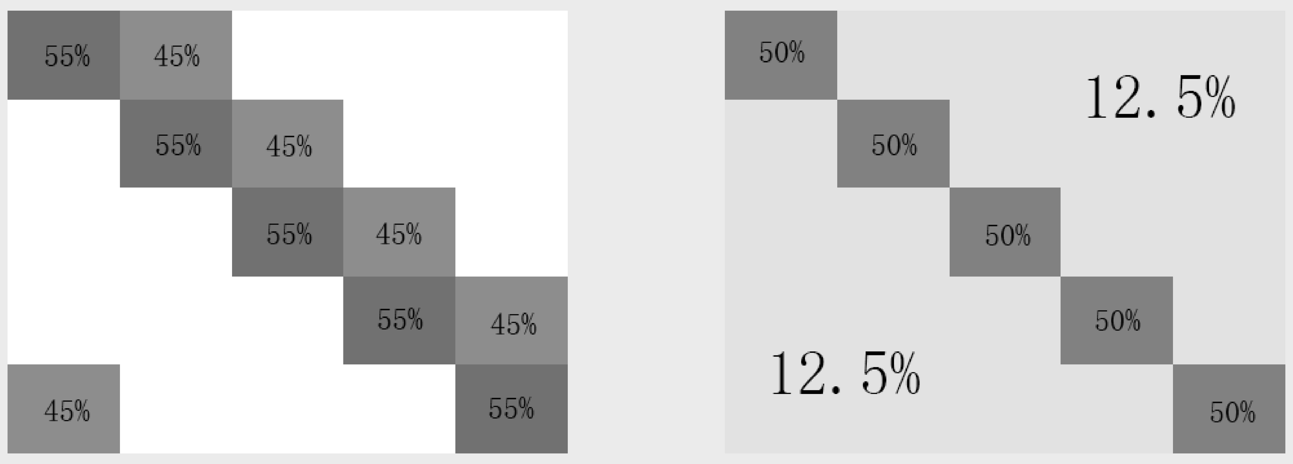
ELBO of MASKING

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

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.

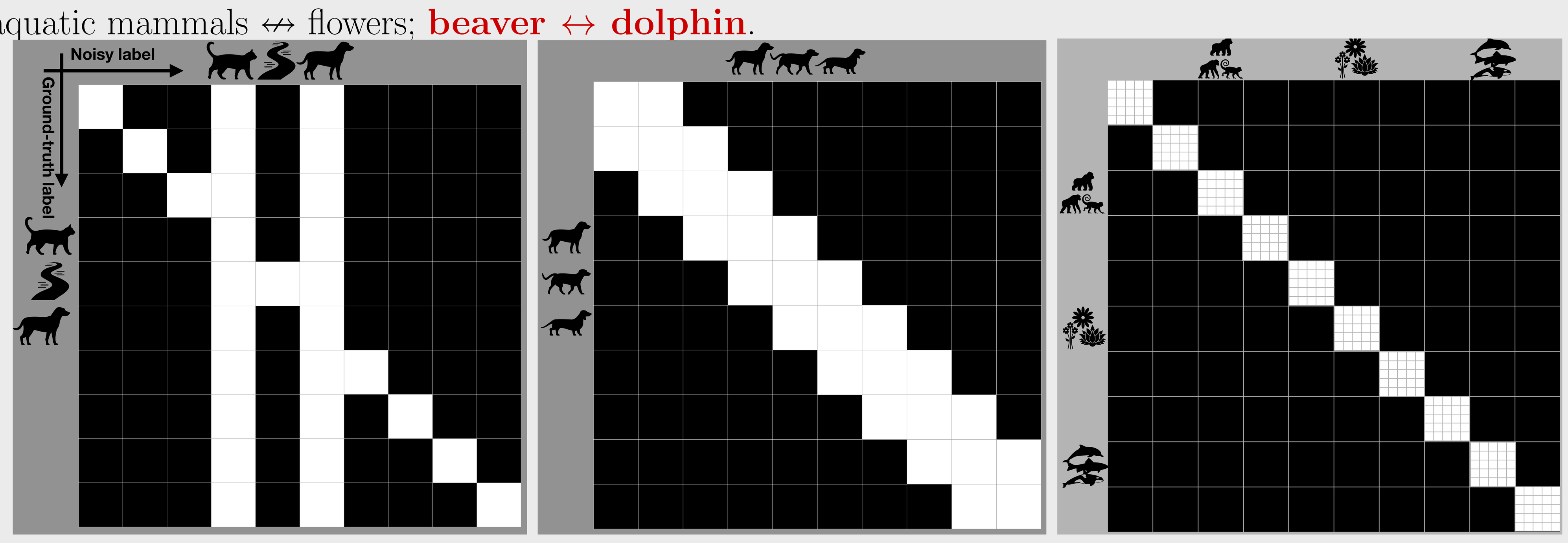
QR Code

Estimating Noise Transition Matrix



Motivation: Data Perspective

- (a) beach ↮ mountain; **beach ↔ dog**.
- (b1) Australian terrier ↮ Norwich terrier; (b2) **Norfolk terrier ↔ Norwich terrier ↔ Irish terrier**.
- (c) aquatic mammals ↮ flowers; **beaver ↔ dolphin**.

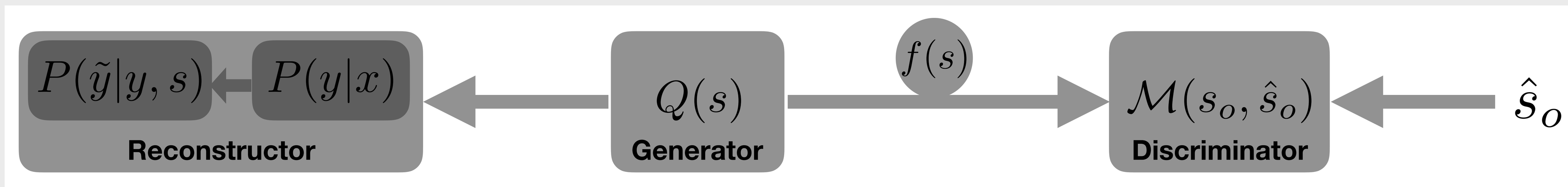


Principled Realization

Q1: Challenge from **structure extraction**.
A1: the **tempered sigmoid func** as $f(\cdot)$ to map from s to s_o,

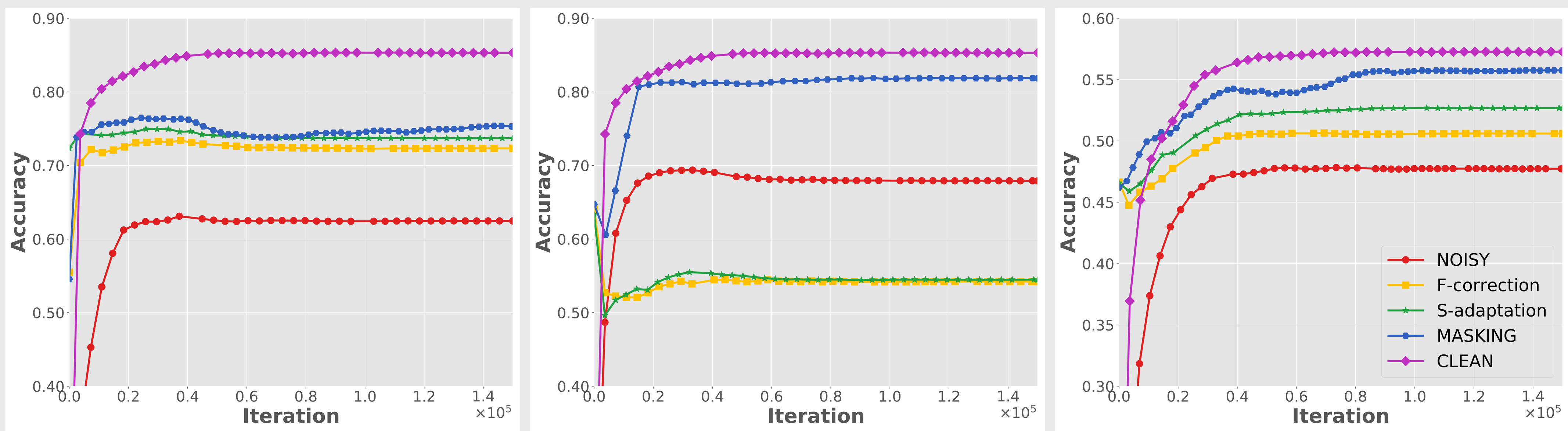
$$f(s) = \frac{1}{1+\exp(-\frac{s-\alpha}{\beta})}, \quad \text{where } \alpha \in (0,1), \beta \ll 1.$$

Q2: Challenge from **structure alignment**.
A2: **GAN-like structure** to model the structure instillation.



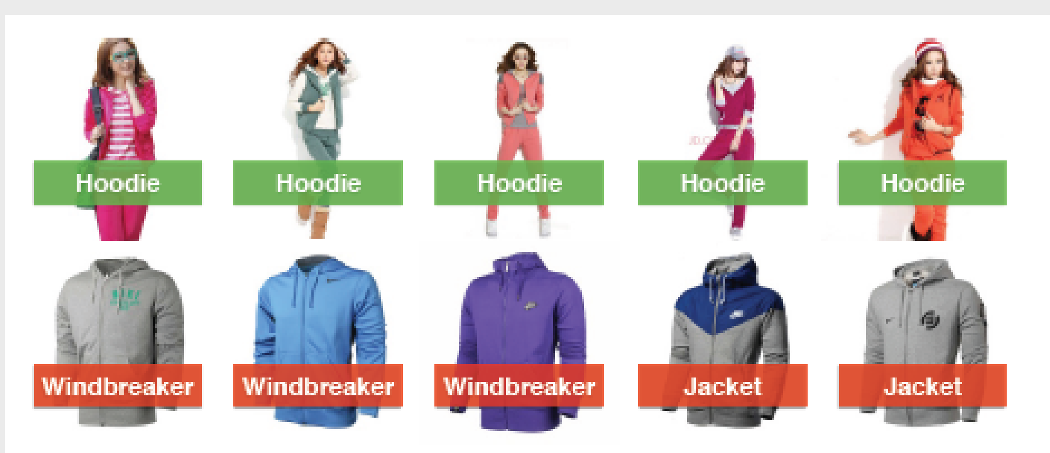
CIFAR10 and CIFAR100

- The test accuracy vs iterations on **benchmark** datasets.



Clothing1M

- The test accuracy on **Clothing1M** from Taobao.com.



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