南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Learning with Noisy Labels From A Causal Perspective

Reporter: Tong Jin

SSL-based or Model-based Method?

Which is Better for Learning with Noisy Labels: The Semi-supervised Method or Modeling Label Noise?

Yu Yao [1 2] Mingming Gong [3 1] Yuxuan Du [4] Jun Yu [5] Bo Han [6] Kun Zhang [1 2] Tongliang Liu [1 7]

ICML 2023

Noisy Label

DNN can achieve remarkable performance when accurately annotated large-scale training datasets are available.

However, annotating a large number of examples accurately is **expensive** and **infeasible** in real life.

Cheap datasets which contain label errors are easy to obtain.

Label Noise

Poor Test Performance

Two Major Streams of Methods

- **SSL-based Methods**

Training Data → Labeled Set (labels are likely to be correct)

Training Data → Unlabeled Set (labels are likely to be incorrect)

e.g., by exploiting the memorization effect of DNN.

SOTA performance on many benchmark datasets.

- **Model-based Methods**

Designing statistically consistent methods by employing the noise transition matrix $\boldsymbol{T}(\mathrm{x})$.

$$\begin{aligned}&\boldsymbol{T}(\boldsymbol{x})[P(Y=1|\boldsymbol{x}),\ldots,P(Y=L|\boldsymbol{x})]^{\top}\\&=[P(\tilde{Y}=1|\boldsymbol{x}),\ldots,P(\tilde{Y}=L|\boldsymbol{x})]^{\top}.\end{aligned}$$

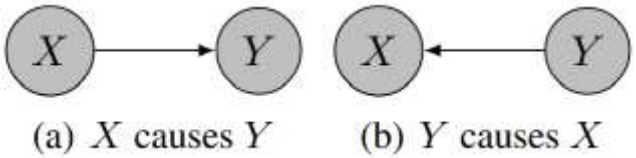
Providing statistical guarantees.

Which stream of methods should be exploited when given a real-world dataset?

It is closely dependent on **the generative process of the dataset**, and none of the two streams of methods are dominating.

Learning with Noisy Labels From A Causal Perspective

- **Data Generation Process**

P (Y |X) 

P (X |Y)

(a) X causes Y (b) Y causes X

Fig.1 An illustration of different data generative processes without label noise.

Modularity property： the conditional distribution of each variable given its causes does not inform or influence the other conditional distributions.

- When X causes Y, P (X) can not inform P (Y|X),

i.e., **P (X) does not contain the relevant information of P (Y |X).**

- When Y causes X, P (Y) can not inform P (X |Y).

P (X) and P (Y |X) do not follow the underlying causal direction. They do not satisfy the modularity property. Bayes Formula: P (X) can inform P (Y |X).

i.e., **P (X) generally contains the relevant information of P (Y |X).**

Different data generative processes can influence the performance of SSL-based method

To make use of the unlabeled data to help learn classifiers, SSL relies on the condition that **P (X) has to contain the information of P (Y |X).**

- When Y causes X, because P (X) contains the information of P (Y |X). It is possible to help learn P (Y |X) by exploiting P (X) by SSL-based method.

  SSL can improve the generalization ability of a classifier.

- When X causes Y , because P (X) generally does not contain the information of P (Y |X).

  Exploiting unlabeled data by using SSL then generally is not helpful.

Different noisy data generative influence SSL-based methods

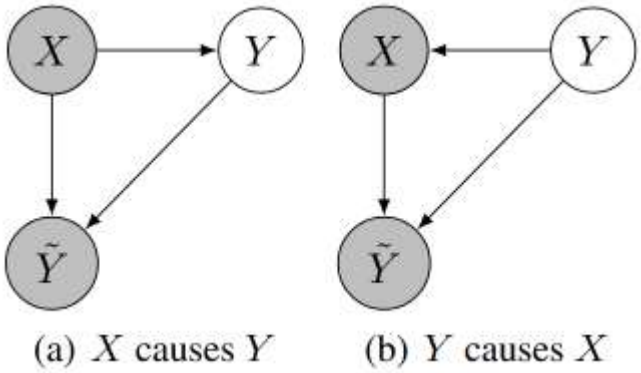


Fig. 2: An illustration of different noisy data generative processes. Both the instance X and the noisy label $\tilde{Y}$ are observable, and the clean label Y is latent.

When X causes Y, X is a cause of both Y and $\tilde{Y}$.

Causal modularity suggests that P (X) does not contain the information on either P ($\tilde{Y}$ |X) or P (Y |X), the unlabeled set can not help learn a classifier in general.

When Y causes X, P (X) contains the information of P (Y |X) because P (X) and P (Y |X) do not satisfy the modularity property, the unlabeled set can help learn a classifier.

How about Model-based Methods?

Estimating the noise transition matrix $T(x)$, which is learned in a **supervised manner** on the whole noisy training set and **does not require exploiting P (X).**
Then, P (Y |X) can be learned by using the estimated $T(x)$ to correct the loss on the whole noisy training set, which is also learned in a supervised manner.

Therefore, the performance of the model-based methods is **not influenced by the different data generative processes**.

However, these methods **usually require a large number of training examples** to accurately estimate the transition matrix. If the transition matrix is poorly estimated, the estimation error of P (Y |X) will be large.

Which is Better for Learning with Noisy Labels?

- **SSL-based methods**

  Strength: easily incorporate heuristics (e.g., prior knowledge) to make use of the finite training sample.

  Weakness: do not work when X causesY in the data generative process.

- **Model-based methods**

  Strength: do not influenced by the data generative process.

  Weakness: need a large training sample to perform well.

Casual Structure Detection Method (CDNL estimator)

Y': pseudo labels estimated by an unsupervised classification method (clustering).

Flip rate P($\tilde{Y}$ |Y'): pseudo labels Y' be flipped into noise labels $\tilde{Y}$.

$Y^* = \arg\max_i P(Y = i|\boldsymbol{x})$: the Bayes label on the clean class-posterior distribution.

Flip rate P($\tilde{Y}$ |Y^*): clean label Y^* be flipped into noise labels $\tilde{Y}$.

If X causes Y, P(X) does not contain labeling information, then ***Y′* should be very different from clean label Y**. Therefore, the estimation error of the flip rate (the difference between P($\tilde{Y}$ |Y') and P($\tilde{Y}$ | Y^*)) is usually large.

If Y causes X, P(X) contains information of P (Y |X), the ***Y′* should be close to clean label Y**. Therefore the estimation error of P ($\tilde{Y}$ |Y') is usually small.

$$d(P(\tilde{Y}|Y^*), P(\tilde{Y}|Y')) = \sum_i^L \sum_j^L \frac{|P(\tilde{Y} = j|Y^* = i) - P(\tilde{Y} = j|Y' = i)|}{L^2}.$$

Experimental Results

Baselines:

Model-based methods: Forward, Reweighting, T-Revision

SSL-based methods: JoCoR, MoPro, Dividemix, Mixup

Datasets:

Synthetic datasets: XYgaussian and YXguassian;

Real-world datasets:

X causes Y: KrKp, Balancescale, Splice;

Y causes X: Waveform, MNIST, and CIFAR10.

Experimental Results

Estimation error of $P(\widetilde{Y}|Y^*)$

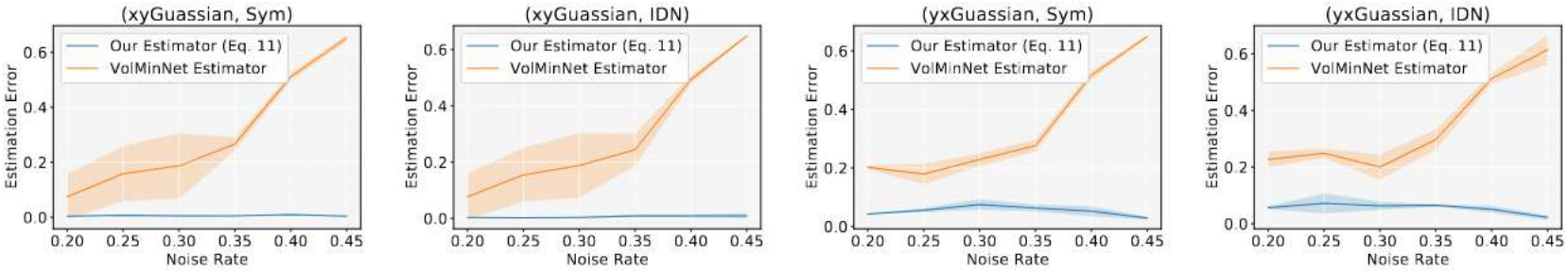


Figure 3: Estimation error of $P(\tilde{Y}|Y^*)$ on synthetic datasets with instance-independent and instance-dependent label noise. Our estimator outperforms the state-of-the-art method by a large margin.

Experimental Results

Classification Accuracies (Synthetic datasets)

Table 1: Test accuracies (%) of different methods on XYgaussian (causal) and YXgaussian (anticausal) datasets with different types of label noise. Estimation errors obtained by CDNL estimator are shown in the parentheses after noise rates.

XYguassian		Sym			Instance	
(causal)	20% (0.196)	30% (0.131)	40% (0.142)	20% (0.101)	30 (0.127)	40% (0.191)
Forward	98.98±0.15	98.28±0.48	96.46±1.07	98.98±0.23	98.60±0.19	97.29±0.51
Reweighting	99.26±0.23	**98.57**±0.34	**96.85**±0.71	99.42±0.30	98.42±0.35	97.14±1.07
T-Revision	**99.32**±0.24	98.55±0.37	96.82±0.72	**99.45**±0.28	**98.55**±0.65	**97.22**±0.91
JoCoR (SSL)	98.19±0.26	89.66±5.31	89.12±19.43	99.03±0.08	89.51±9.62	73.41±13.44
MoPro (SSL)	96.41±0.45	95.70±0.93	77.32±6.99	95.98±0.87	94.63±0.64	77.79±8.95
Dividemix (SSL)	97.20±0.25	96.98±0.12	95.39±0.83	97.15±0.56	97.13±0.17	90.72±0.47
Mixup (SSL)	97.15±0.14	96.88±0.35	94.12±0.76	96.93±0.44	96.15±0.65	87.68±9.06
YXguassian		Sym			Instance	
(antiausal)	20% (0.026)	30% (0.031)	40% (0.028)	20% (0.027)	30 (0.031)	40% (0.043)
Forward	86.26±0.13	85.97±0.19	84.85±0.93	86.10±0.11	85.56±0.47	83.94±2.14
Reweighting	86.31±0.18	85.85±0.27	84.68±0.55	86.22±0.25	86.03±0.26	84.19±0.84
T-Revision	**86.32**±0.17	85.81±0.32	84.42±0.56	86.25±0.23	86.02±0.23	84.18±0.83
JoCoR (SSL)	86.26±0.10	85.99±0.09	85.86±0.21	86.16±0.14	86.13±0.14	85.43±0.34
MoPro (SSL)	84.79±0.72	84.17±0.61	83.67±1.32	85.36±0.63	84.43±1.27	81.07±3.03
Dividemix (SSL)	**86.32**±0.20	**86.28**±0.11	**86.23**±0.19	**86.37**±0.09	**86.37**±0.12	**86.06**±0.15
Mixup (SSL)	86.15±0.19	85.64±0.63	82.48±2.56	85.74±0.43	85.01±0.92	81.47±5.76

Experimental Results

Classification Accuracies (Real-world datasets)

Table 2: Comparing test accuracies (%) of different methods on causal and anticausal datasets with different levels and types of label noise. Estimation errors obtained by employing CDNL estimator are shown in the parentheses after noise rates.

KrKp	Sym			Instance		
(causal)	20% (0.297)	30% (0.196)	40% (0.070)	20% (0.262)	30% (0.166)	40% (0.072)
Forward	93.31±1.0	89.31±1.96	77.78±7.4	94.0±0.8	87.25±3.1	80.75±2.31
Reweighting	93.88±1.43	91.16±1.09	77.31±5.26	93.5±2.63	89.25±1.53	78.22±6.61
T-Revision	**94.72**±0.62	**91.81**±1.93	**77.97**±5.0	**94.5**±1.63	**90.78**±2.35	**79.06**±4.89
JoCoR (SSL)	93.69±0.23	89.53±0.84	67.81±2.07	93.44±0.71	87.44±2.95	67.75±6.51
MoPro (SSL)	89.47±1.13	79.47±7.03	65.94±2.06	89.31±3.82	79.59±6.2	62.62±4.78
Dividemix (SSL)	93.75±0.32	88.31±0.65	74.31±1.44	93.47±0.15	93.34±0.72	63.94±1.45
Mixup (SSL)	93.31±1.1	88.81±1.03	73.84±1.18	93.19±1.31	87.25±1.49	74.31±3.42
Splice	Sym		Pair		Instance	
(causal)	20% (0.136)	40% (0.146)	20% (0.140)	40% (0.148)	20% (0.151)	40% (0.153)
Forward	71.25±3.07	66.18±3.61	73.73±1.03	65.8±3.67	65.8±4.08	61.6±5.67
Reweighting	76.96±1.69	71.91±2.68	**75.55**±1.88	**66.68**±1.54	75.64±1.95	**63.54**±7.21
T-Revision	**76.99**±1.73	**71.94**±2.68	75.49±2.05	66.61±1.5	**75.67**±1.89	63.45±7.17
JoCoR (SSL)	69.81±4.61	63.2±1.89	59.37±1.44	57.71±3.7	59.66±2.44	55.3±5.87
MoPro (SSL)	53.6±0.19	53.51±0.0	53.51±0.0	53.25±0.43	53.79±0.38	52.17±3.27
Dividemix (SSL)	75.11±1.66	53.45±0.0	53.45±0.0	56.14±2.1	59.97±0.55	51.41±1.79
Mixup (SSL)	67.43±3.2	62.16±2.52	68.15±2.63	63.67±6.63	65.52±2.22	49.03±9.86
MNIST	Sym		Pair		Instance	
(anticausal)	20% (0.034)	40% (0.038)	20% (0.041)	40% (0.20)	20% (0.025)	40% (0.026)
Forward	98.75±0.08	97.86±0.22	98.84±0.10	94.92±0.89	96.87±0.15	90.30±0.61
Reweighting	98.71±0.11	98.13±0.19	98.54±.63	91.50±1.27	97.99±0.13	90.30±0.61
T-Revision	98.91±0.04	98.34±0.21	98.89±0.08	91.83±1.08	98.39±0.09	96.50±0.31
JoCoR (SSL)	98.06±0.13	96.64±0.19	98.01±0.19	96.85±0.43	98.62±0.06	96.07±0.31
MoPro (SSL)	98.51±0.92	95.14±1.23	96.79±1.04	94.96±1.32	98.53±0.52	96.45±1.20
Dividemix (SSL)	**99.24**±0.03	**99.21**±0.05	**99.25**±0.03	**98.50**±0.08	**99.31**±0.02	**97.75**±0.1
Mixup (SSL)	97.45±0.21	95.75±0.43	97.57±1.08	92.46±1.43	96.54±1.20	90.38±1.30
CIFAR10	Sym		Pair		Instance	
(anticausal)	20% (0.010)	40% (0.009)	20% (0.010)	40% (0.026)	20% (0.037)	40% (0.042)
Forward	88.21±0.48	78.44±0.89	88.21±0.48	77.44±6.89	85.29±0.38	74.72±3.24
Reweighting	86.77±0.40	83.16±0.46	89.60±1.01	77.06±6.47	88.72±0.41	84.52±2.65
T-Revision	90.33±0.52	84.94±2.58	89.75±0.41	80.94±2.58	90.46±0.13	85.37±3.36
JoCoR (SSL)	85.96±0.25	79.65±0.43	80.33±0.20	71.62±1.05	89.80±0.28	73.78±1.39
MoPro (SSL)	78.15±0.15	67.70±0.56	77.92±0.81	69.89±1.02	78.75±0.15	67.61±0.24
Dividemix (SSL)	**95.60**±0.10	**94.80**±1.10	**95.72**±0.04	**87.02** ±0.41	**95.50**±1.17	**94.50**±0.23
Mixup (SSL)	93.20±0.31	86.20±0.30	92.23±0.71	82.43±1.02	93.32±0.25	87.61±0.56

Table 4: Test accuracies (%) of different methods on Balancescale (causal) with different types of label noise. Estimation errors obtained by CDNL estimator are shown in the parentheses after noise rates.

Balancescale	Sym		Pair		Instance	
(Causal)	20% (0.099)	40% (0.071)	20% (0.113)	40% (0.109)	20% (0.110)	40% (0.090)
Forward	74.24±8.74	78.8±10.53	83.36±2.23	72.48±9.12	75.36±5.53	69.6±9.71
Reweighting	89.76±3.37	89.28±1.87	**94.08**±2.41	79.36±15.02	**90.72**±2.8	**86.24**±1.38
T-Revision	**92.64**±0.93	**89.76**±3.14	92.32±3.97	**81.12**±13.91	89.12±3.45	85.28±2.06
JoCoR (SSL)	76.96±3.87	58.08±13.43	72.32±10.43	60.16±12.88	73.28±4.34	51.2±6.13
MoPro (SSL)	84.29±2.38	84.13±1.81	84.73±3.16	80.79±7.93	86.19±2.59	78.1±7.28
Dividemix (SSL)	88.16±0.32	86.56±0.93	81.12±0.39	62.96±1.47	87.52±0.64	79.04±1.18
Mixup (SSL)	86.08±2.51	83.68±3.49	86.72±1.3	67.68±17.1	84.96±2.17	75.36±5.46

Table 5: Test accuracies (%) of different methods on Waveform (anticausal) datasets with different types of label noise. Estimation errors obtained by CDNL estimator are shown in the parentheses after noise rates.

Waveform	Sym		Pair		Instance	
(Anticausal)	20% (0.138)	40% (0.257)	20% (0.257)	40% (0.12)	20% (0.099)	40% (0.089)
Forward	74.66±7.68	74.76±3.3	70.02±10.79	66.46±3.84	59.78±12.14	56.62±12.87
Reweighting	**84.58**±1.89	83.92±1.38	**83.30**±2.28	**73.22**±4.51	**85.02**±0.93	**83.3**±3.02
T-Revision	84.24±1.3	**85.70**±0.66	82.72±6.03	68.86±8.56	84.04±2.38	83.5±1.87
JoCoR (SSL)	83.44±0.83	60.28±1.46	80.64±1.29	57.14±4.17	63.84±8.8	54.56±4.44
MoPro (SSL)	76.62±7.16	76.37±7.0	79.55±2.32	58.44±7.11	77.36±4.04	65.14±5.61
Dividemix (SSL)	83.36±0.63	82.06±1.25	69.74±1.9	58.48±0.98	73.00±2.30	66.86±1.26
Mixup (SSL)	81.38±1.67	79.48±1.05	80.54±2.51	72.34±4.58	78.88±1.05	71.26±5.44

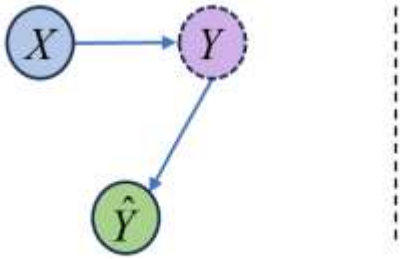
Learning Causal Transition Matrix for Instance-dependent Label Noise

Jiahui Li[1,2*], Tai-Wei Chang[2*], Kun Kuang[1†], Ximing Li[2], Long Chen[3], Jun Zhou[2]

[1]Zhejiang University
[2]Ant Group
[3]The Hong Kong University of Science and Technology
jiahuil@zju.edu.cn, taiwei.twc@antgroup.com, kunkuang@zju.edu.cn
xili.lxm@antgroup.com, longchen@ust.hk, jun.zhoujun@antgroup.com

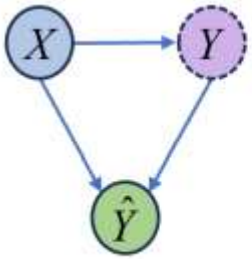
AAAI 2025

Model-based Method



(a) Instance-independent noise

$P(\hat{Y}|Y,X) = P(\hat{Y}|Y)$



(b) Instance-dependent noise

$P(\hat{Y}|Y,X)$

X : Instance

Y : Clean label

$\hat{Y}$: Observed label

Transition matrix: $T_{ij} = P(\hat{Y} = j \mid Y = i, X = x)$

Causal Graphs of the Data Generation Procedure

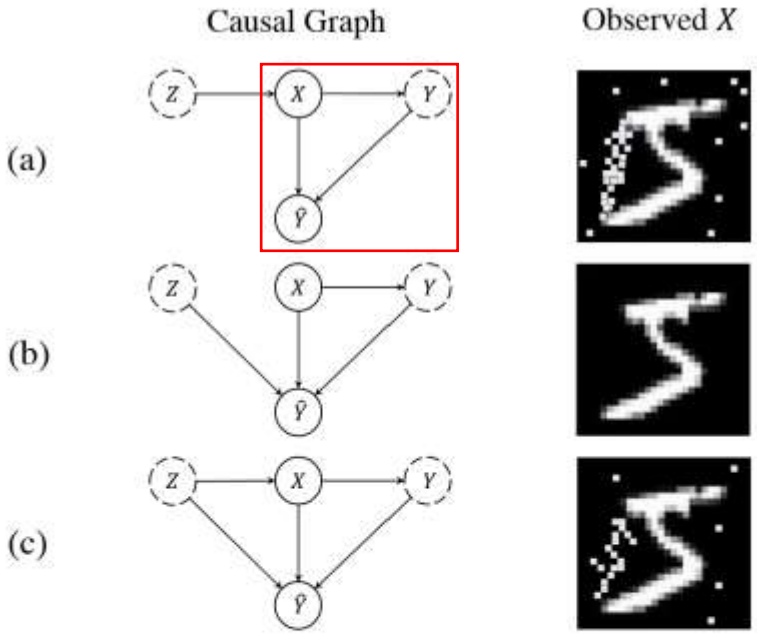


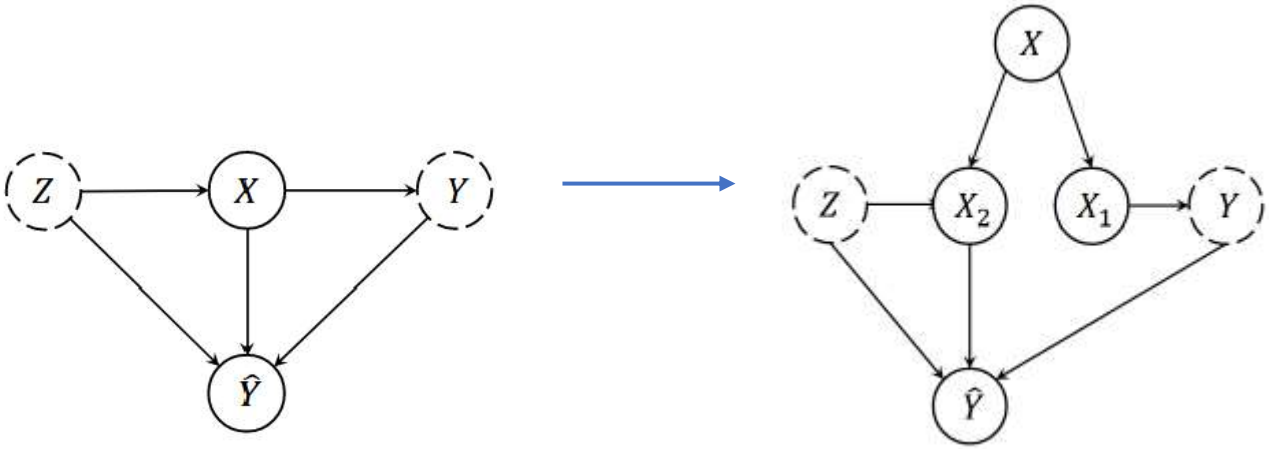
Figure 1: Examples of three causal graphs illustrating the mislabeling of "5" as "6" in MNIST, where X denotes instance(image), Y denotes the ground truth label, $\hat{Y}$ denotes the noisy label, and Z denotes the latent variable. The dashed circles represent the unobservable variable. (a) The instance is perturbed by noise, making "5" looks like "6". (b) The instance is clean, but it is mislabeled by an annotator. (c) The instance exhibits a mixture situation of (a) and (b).

Scenario (a), certain environmental factors Z, e.g. lighting, noise, shadow, impact the instance X. This, in turn, influences the annotators, leading to the production of a noisy label $\hat{Y}$.

Scenario (b), some factors Z do not influence X but directly cause the generation of the noisy label $\hat{Y}$, such as the annotator's negligence.

Scenario (c), both scenarios (a) and (b) can occur simultaneously, resulting in a noisy label $\hat{Y}$.

Causal Graph for Learning with Noisy Labels



X → Noise-resistant component X_1

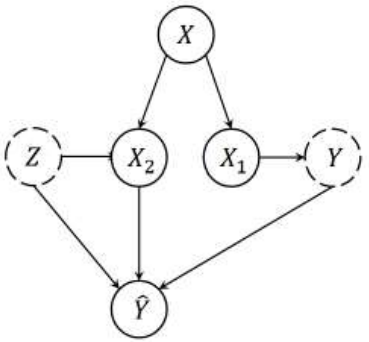
Noise-sensitive component X_2

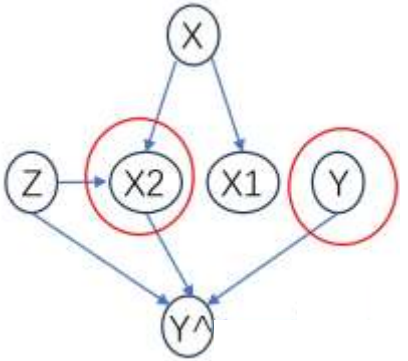
Causal Viewpoint for Denoising

Causal transition matrix: $T_{cau} = P(\hat{Y}|do(Y), X)$

Theorem 1 *The instance-dependent causal transition matrix $P(\hat{Y}|do(Y), X)$ is identifiable if we recover the noise predictive factor X_2.*

Proof: Let $G_{\bar{Y}}$ be the graph induced by removing the incoming edges of Y. Since $\hat{Y} \perp\!\!\!\perp X_1 | Y, X_2$ in $G_{\bar{Y}}$, we have $P(\hat{Y}|do(Y), X) = P(\hat{Y}|do(Y), X_1, X_2) = P(\hat{Y}|do(Y), X_2)$. Let $G_{\underline{Y}}$ be the graph induced by removing the outgoing edges of Y. Since $\hat{Y} \perp\!\!\!\perp Y | X_2$ in $G_{\underline{Y}}$, we have $P(\hat{Y}|do(Y), X_2) = P(\hat{Y}|Y, X_2)$. However, note that Y is an unobservable latent variable that we are interested in modeling, and as such we need causal estimand that gives us an unbiased estimation of Y as well.





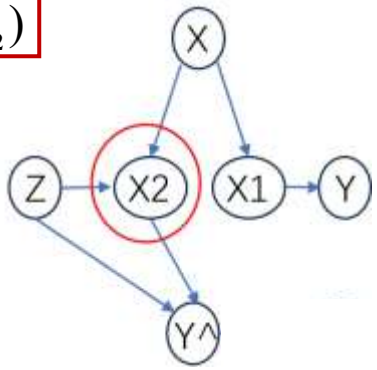
$G_{\bar{Y}}$

$$P(\hat{Y} \mid do(Y), X) = P(\hat{Y} \mid Y, X_2)$$

Chain: X → Y → Z

Fork: X ← Y → Z

Collider: X → Y ← Z



$G_{\underline{Y}}$

Causal Viewpoint for Denoising

Theorem 2 *The effect of X_1 on Y can be identified if we recover X_1.*

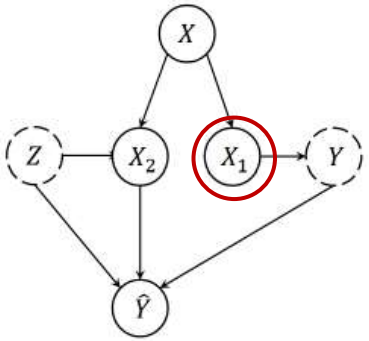
Proof: Since we have $P(Y|do(X_1)) = \int_{X_2} P(Y|X_1, X_2)P(X_2)dX_2$ by using the backdoor criterion and that $Y \perp\!\!\!\perp X_2|X_1$ by d-separation, we have

$$\int_{X_2} P(Y|X_1, X_2)P(X_2)dX_2 = \int_{X_2} P(Y|X_1)P(X_2)dX_2$$

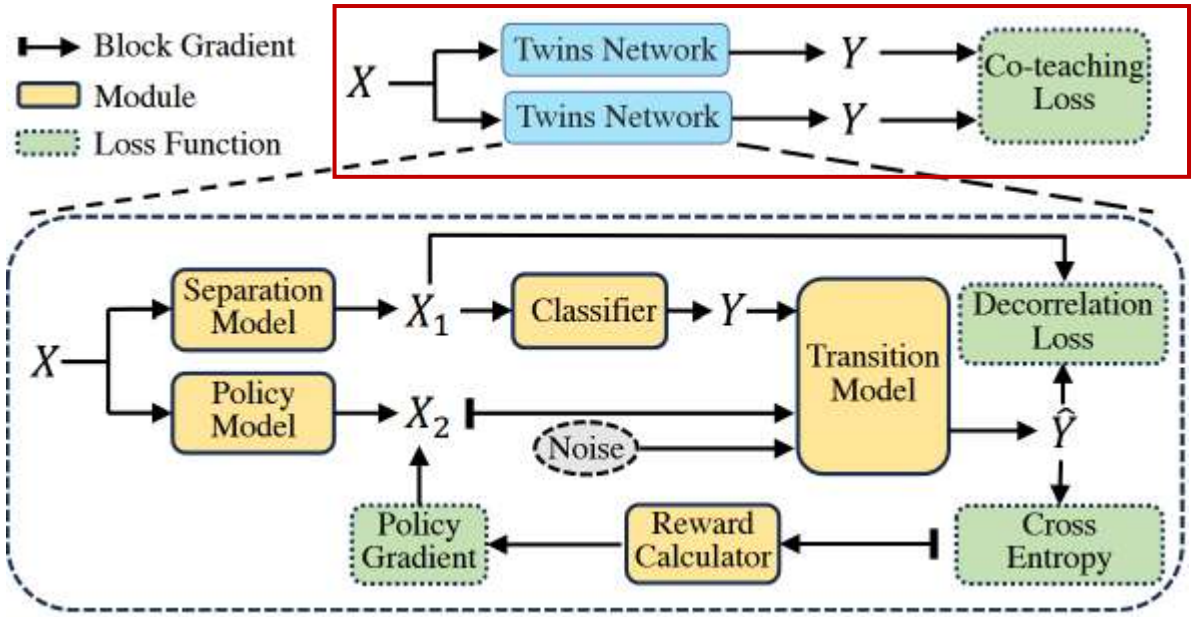
$$= P(Y|X_1)\int_{X_2} P(X_2)dX_2 = P(Y|X_1).$$

Based on Theorem 2, it is possible to obtain an unbiased classifier based solely on X_1. Consequently, we can recover X_1 by decorrelating it from Z. According to Theorem 1, the causal transition matrix can be identified if we can identify the contributing factor $X_2 \subseteq X$ to $\hat{Y}$. Intuitively, X_1 can be omitted, since it is the parent of Y and the *do* operation effectively eliminates the incoming edge of Y.

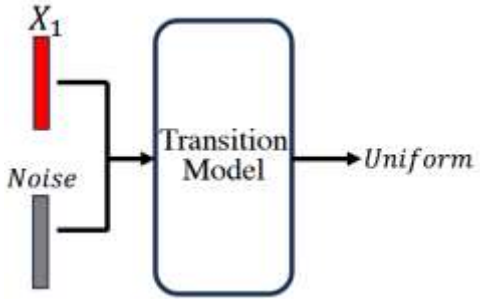
$$P\left(Y \mid do\left(X_1\right)\right) = P\left(Y \mid X_1\right)$$



Training Framework for Denoising



Decorrelating X_1 and $\hat{Y}$



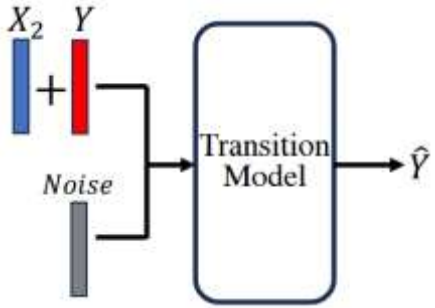
Transition model take X_1 and a Gaussian noise as inputs and output a tensor with dimension of k:

$$\hat{Y}_{X_1} = f_{tran}(X_1, Z_1),\ Z_1 \sim \mathcal{N}(0,1),\ \hat{Y}_{X_1} \in \mathbb{R}^k.$$

We aim to ensure that the probability of $\hat{Y}$ given X_1 is uniform for each class candidate. To achieve this, we constrain $\hat{Y}_{X_1}$ so that X_1 of each instance predicts an all-one vector.

$$\mathbf{Reg}(X_1) = -\sum_{i=1}^{N} \tilde{y}_{x_1} \text{log\_softmax}(\mathbf{1}_k),$$

Modeling the relationship from X_2, Y, Z to $\hat{Y}$



For modeling the causal transition matrix, we take Y and another noise as input, affiliated with the noise-sensitive component X_2:

$$\hat{Y} = f_{tran}(m(Y, X_2), Z_2),\ Z_2 \sim \mathcal{N}(0,1),\ \hat{Y} \in \mathbb{R}^k,$$

$$m(Y, X_2) = gs(Y + \beta * X_2),$$

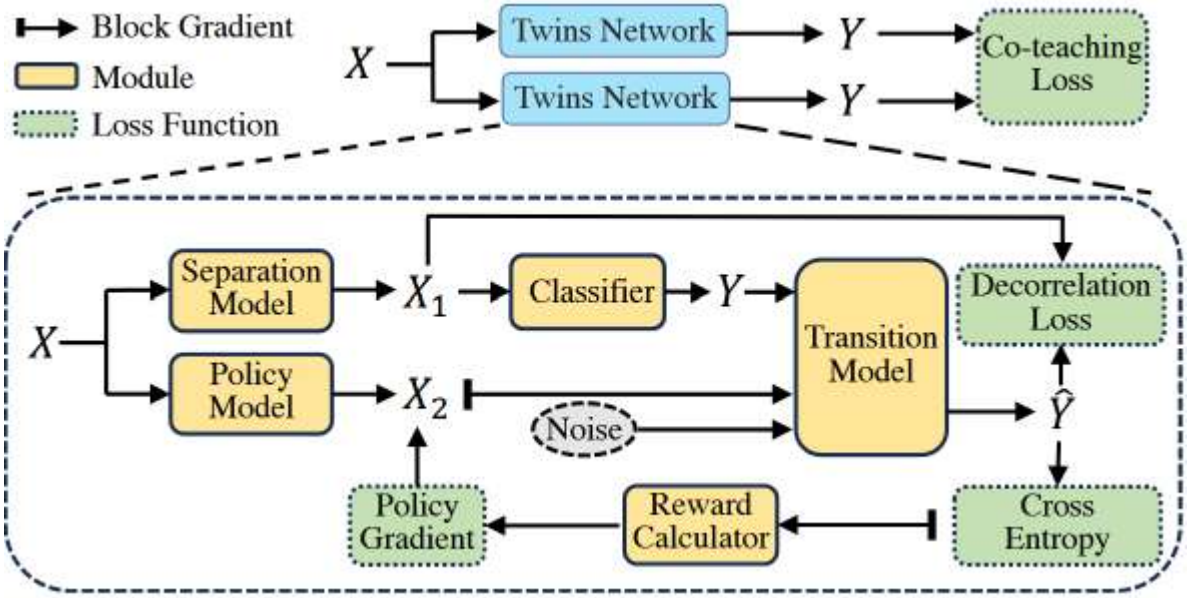
Since our main focus is on the causal transition, we block the gradient of X_2 to make the relationship from Y to $\hat{Y}$ more precise. Therefore, X_2 can be considered as the compensation tensor that bridges the gap $P(\hat{Y}|do(Y), X)$ and $P(\hat{Y}|do(Y))$

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} \tilde{y}_i \log(\hat{y}_i),$$

where $\tilde{y}$ denotes the predicted noisy label, and $\hat{y}$ denotes the noisy label.

Training Framework for Denoising

$$\mathcal{L} = \mathcal{L}_{\text{co-teaching}} + \alpha_1 * \mathcal{L}_{CE} + \alpha_2 * \mathcal{L}_{PG} + \alpha_3 * \mathbf{Reg}(X_1),$$



Reward calculator: $R = \frac{1}{1-\tilde{y}_i \log(\hat{y}_i)}$

Loss function: $\mathcal{L}_{pg} = -\sum_{i=1}^{N} R \log \pi(x_i).$

Experimental Results

Performance Results

Method	SYM		ASYM		IDN	
	20%	80%	20%	40%	20%	40%
CE	74.0	27.0	81.0	77.3	68.4	52.1
Early Stop	83.6	49.5	84.1	76.6	79.5	55.4
Co-teaching	82.5	64.2	88.2	73.6	81.8	75.4
Joint	82.0	6.0	82.1	82.3	82.7	82.4
JoCoR	86.0	27.6	88.9	79.4	86.3	83.2
CORES2	74.6	8.9	77.6	74.3	80.0	58.1
SCE	74.0	27.0	82.0	77.4	68.3	52.0
LS	73.9	27.8	81.5	77.0	69.0	52.5
REL	84.6	70.1	82.8	76.2	84.6	75.5
Forward	77.4	24.3	88.3	79.2	75.2	56.9
DualT	84.5	10.0	86.9	83.1	85.1	68.5
TVR	72.6	24.9	80.6	76.4	66.3	51.7
CausalNL	84.0	51.5	88.8	87.4	90.8	90.0
Ours w.o/ pg	**92.1**	**75.8**	89.6	82.4	**91.3**	90.3
Ours	**92.1**	71.5	**91.4**	**88.7**	91.0	**90.4**

Table 1: Results on FashionMNIST with symmetric, asymmetric, and instance-dependent label noise.

Map	IDN				
	20%	30%	40%	45%	50%
CE	91.51±0.45	91.21±0.43	87.87±1.12	67.15±1.65	51.01±3.62
Co-teaching	93.93±0.31	92.06±0.31	91.93±0.81	89.33±0.71	67.62±1.99
Decoupling	90.02±0.25	91.59±0.25	88.27±0.42	84.57±0.89	65.14±2.79
MentorNet	94.08±0.12	92.73±0.37	90.41±0.49	87.45±0.75	61.23±2.82
Mixup	89.73±0.37	90.02±0.35	85.47±0.63	82.41±0.62	68.95±2.58
Forward	91.89±0.31	91.59±0.23	89.33±0.53	80.15±1.91	62.53±3.35
Reweight	92.44±0.34	92.32±0.51	91.31±0.67	85.93±0.84	64.13±3.75
T-Revision	93.14±0.53	93.51±0.74	92.65±0.76	88.54±1.58	64.51±3.42
BLTM-V	**95.12±0.40**	**94.69±0.24**	88.13±3.23	80.43±4.12	78.71±4.37
CausalNL	94.06±0.23	93.86±0.65	93.82±0.64	93.19±0.93	85.41±2.95
Ours w.o/ pg	93.86±0.17	93.82±0.18	93.50±0.25	93.19±1.25	**92.91±1.54**
Ours	94.13±0.08	93.97±0.11	**93.94±0.16**	**93.33±1.12**	92.57±1.56

Table 2: Results on SVNH with instance-dependent noise.

Experimental Results

Map	CIFAR10-IDN					CIFAR100-IDN				
	20%	30%	40%	45%	50%	20%	30%	40%	45%	50%
CE	75.81±0.26	69.15±0.65	62.45±0.86	51.72±1.34	39.42±2.52	30.42±0.44	24.15±0.78	21.45±0.70	15.23±1.32	14.42±2.21
Co-teaching	80.96±0.31	78.56±0.61	73.41±0.78	71.60±0.79	45.92±2.21	37.96±0.53	33.43±0.74	28.04±1.43	25.60±0.93	23.97±1.91
Decoupling	78.71±0.15	75.17±0.58	61.73±0.34	58.61±1.73	50.43±2.19	36.53±0.49	30.93±0.88	27.85±0.91	23.81±1.31	19.59±2.12
MentorNet	81.03±0.24	77.22±0.47	71.83±0.49	66.18±0.64	47.89±2.03	38.91±0.54	34.23±0.73	31.89±1.19	27.53±1.23	24.15±2.31
Mixup	73.17±0.34	70.02±0.31	61.56±0.71	56.45±0.67	48.95±2.58	32.92±0.76	29.76±0.87	25.92±1.26	23.13±2.15	21.31±1.32
Forward	76.64±0.26	69.75±0.56	60.21±0.75	48.81±2.59	46.27±1.30	36.38±0.92	33.17±0.73	26.75±0.93	21.93±1.29	19.27±2.11
Reweight	76.23±0.25	70.12±0.72	62.58±0.46	51.54±0.92	45.46±2.56	36.73±0.72	31.91±0.91	28.39±1.46	24.12±1.41	20.23±1.23
T-Revision	76.15±0.37	70.36 ±0.54	64.09±0.37	54.42±1.01	49.02±2.13	37.24±0.85	36.54±0.79	27.23±1.13	25.53±1.94	22.54±1.95
BLTM-V[1]	80.37±1.98	78.82±1.07	72.93±4.00	64.83±4.65	60.33±5.29	-	-	-	-	-
CausalNL	81.47±0.32	80.38±0.44	77.53±0.45	78.60±1.06	77.39±1.24	41.47±0.43	40.98±0.62	34.02±0.95	33.34±1.13	32.13±2.23
Ours w.o/ pg	82.57±0.33	81.24±0.36	79.36±0.81	78.43±0.51	75.59±2.07	45.50±0.99	**44.67±0.60**	38.44±1.40	34.88±2.53	33.05±1.47
Ours	**82.94±0.29**	**82.15±0.25**	**81.04±0.23**	**80.24±0.39**	**78.37±0.93**	**46.37±0.46**	43.34±0.39	**39.61±1.04**	**37.04±1.83**	**34.44±1.86**

Table 3: Results on CIFAR dataset.

Method	Dataset	
	Food101	Clothing1M
CE	78.37	68.88
Early Stop	73.22	67.07
Co-teaching	78.35	60.15
SCE	75.23	67.77
REL	78.96	62.53
Forward	83.76	69.91
DualT	57.46	70.18
TVR	77.37	69.44
CausalNL	85.64	68.90
Ours w.o/ pg	85.52	70.48
Ours	**85.86**	**72.25**

Table 4: Results on real-world dataset.

THANKS