Class-Dependent Label-Noise Learning with Cycle-Consistency Regularization

De Cheng[1*], Yixiong Ning[1*], Nannan Wang[1†], Xinbo Gao[2], Heng Yang[3], Yuxuan Du[4], Bo Han[5], Tongliang Liu[6]

[1]Xidian University, [2]Chongqing University of Posts and Telecommunications, [3]Shenzhen Aimo Technology Co., LTD, [4]JD Explore Academy, [5]Hong Kong Baptist University, [6]TML Lab, The University of Sydney.

Motivation

Problem: In label-noise learning, estimating the transition matrix plays an important role in building statistically consistent classifier. Current state-of-the-art consistent estimator for the transition matrix has been developed under the newly proposed sufficiently scattered assumption, through incorporating the minimum volume constraint of the transition matrix T into label-noise learning. To compute the volume of T, it heavily relies on the estimated noisy class posterior. However, the estimation error of the noisy class posterior could usually be large as deep learning methods tend to easily overfit the noisy labels.

Solution: We propose to estimate the transition matrix under a forward-backward cycle-consistency regularization, of which we have greatly reduced the dependency of estimating the transition matrix T on the noisy class posterior.

Problem Settings:

Clean data distribution: (X, Y)

Noisy data distribution: $(X, \overline{Y})$ Training data: $\overline{D} := \{(\mathbf{x}_i, \overline{y}_i)\}_{i=1}^{N}$

Noisy class-posterior probability: $P(\overline{Y}|X)$

Clean class-posterior probability: $P(Y|X)$

Transition Matrix: $T_{ij}(\mathbf{x}) = P(\overline{Y} = \mathrm{j}|Y = i, X = x)$

Main Contributions

This paper proposes a class-dependent label-noise learning with cycle-consistency regularization, which creatively propose to estimate the transition matrix under a forward-backward cycle-consistency regularization, of which we have greatly reduced the dependency of estimating the transition matrix T on the noisy class posterior:

- We propose to estimate the transition matrix T under a forward-backward cycle-consistency regularization, of which we could greatly reduce the dependency of minimizing the volume of the transition matrix T on the estimated noisy class posterior.
- We show that such cycle-consistency regularization could help to minimize the volume of the transition matrix T without directly exploiting the estimated noisy class posterior, which encourages the estimated transition matrix T to converge to the optimal solution.
- Experimental results on four datasets (two synthetic and two real-world datasets) with different label-noise settings consistently justify the effectiveness of the proposed method, on reducing the estimation error of the transition matrix and greatly improving the classification performance.

Methodology

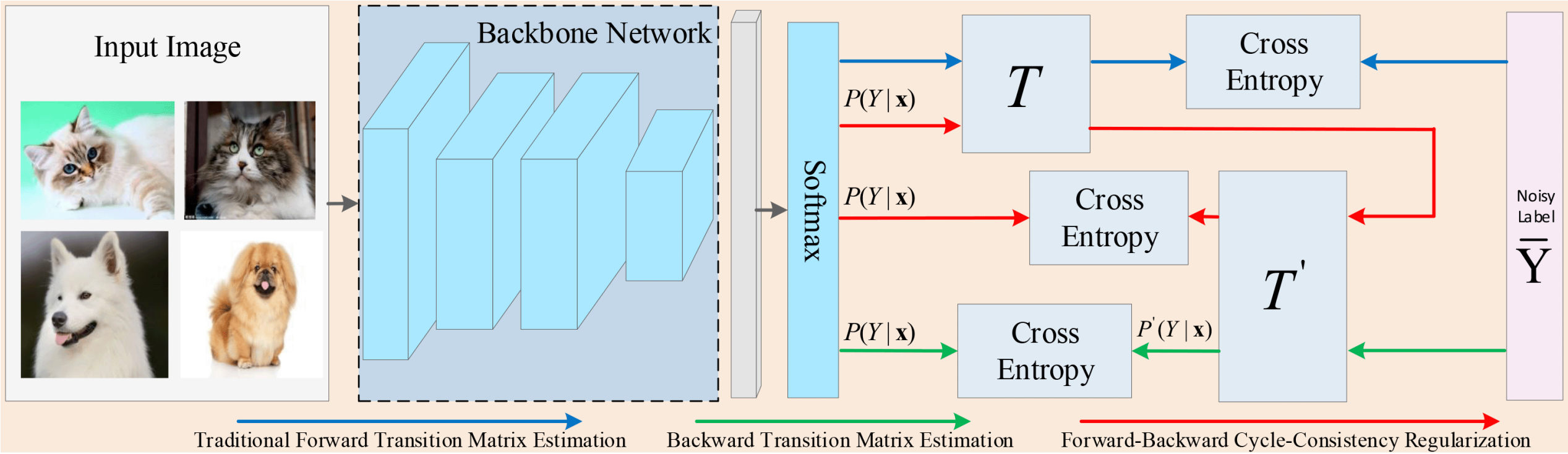


Figure 1. The proposed Cycle-Consistency regularization label-noise learning framework.

Overall Objective Function:

$$L = L_1(\mathbf{w}; T) + L_2(\mathbf{w}; T') + \lambda L_3(\mathbf{w}; T, T')$$

Methodology

- **Forward Transition Matrix**:

  Given the training dataset $\overline{D} = \{(\mathrm{x}_i, \overline{y}_i)\}_{i=1}^{N}$, we optimize the empirical risk by jointly optimizing the transition matrix T and the consistent classifier $f(\mathrm{x};\ \mathrm{w})$ for label-noise learning. Specifically, we minimize the approximation error between the inferred noisy class-posterior probability $\mathrm{T}f(\mathrm{x};\ \mathrm{w})$

$$\min_{\mathbf{w},T} L_1(\mathbf{w}; T) = -\frac{1}{N}\sum_{i=1}^{N} \overline{y}_i \log(T \cdot f(x_i; \mathbf{w}))$$

- **Backward Transition Matrix**:

  Minimizing the volume of the transition matrix is equivalent to maximizing the volume of the clean class posterior. We directly use the noisy labels to maximize the volume of the clean class posterior. Intuitively, we maximize the volume of the learned clean class posteriors by matching it with the projected C-dimensional simplex in the noisy class posterior which also needs to estimate.

$$\min_{\mathbf{w},T'} L_2(\mathbf{w}; T') = -\frac{1}{N}\sum_{i=1}^{N} f(x_i; \mathbf{w}) \log(T' \cdot SoftMax(\overline{y}_i))$$

Methodology

- **Cycle-Consistency Regularization**:

  we build an "indirect" cycle-consistency through minimizing the difference between $P(Y|X = \mathrm{x})$ and $T'\big(\mathrm{T}P(Y|X = \mathrm{x})\big)$, where "indirect" means that we would make use of the invertible relationship between these two matrices T and T ′ indirectly through the original and circularly computed clean class-posterior.

$$\min_{\mathbf{w}} L_3(\mathbf{w}; T, T') = -\frac{1}{N}\sum_{i=1}^{N} f(x_i; \mathbf{w}) \log(T'(T \cdot f(x_i; \mathbf{w})))$$

Experiments

- Comparison with SOTAs On Two Synthetic Datasets With Symmetry Noise.

Method	Cifar-10			CIFAR-100		
	Sym-20%	Sym-40%	Sym-60%	Sym-20%	Sym-40%	Sym-60%
CE (baseline)	84.58± 0.18	80.78± 0.38	68.31± 0.33	51.93 ±0.39	40.11± 0.86	25.81± 0.74
GCE [40]	89.31 ±0.07	86.61± 0.23	79.40± 0.41	66.07± 0.24	59.03± 0.21	45.68± 0.39
PeerLoss [17]	88.78 ±0.18	84.87± 0.15	75.28± 0.31	57.34± 0.34	43.39± 0.33	28.66± 0.67
Co-teaching [6]	85.76 ±0.26	83.12± 0.31	70.89± 1.06	56.83± 0.28	43.38± 0.51	28.04± 0.69
Co-teaching++ [38]	86.39±0.33	83.80±0.30	72.51±0.46	57.64±0.34	44.28±0.83	29.60±1.16
T-Revision [33]	88.01±0.16	84.52±0.11	71.53±0.82	62.66±0.53	55.25±0.36	39.94±1.28
VolMinNet [13]	89.69±0.19	85.46±0.19	73.55±0.78	64.70±0.60	56.25±0.45	41.06±0.45
DualT [37]	89.88±0.13	86.23±0.64	72.21±1.67	65.75±0.38	56.80±0.18	42.56±0.55
T-For(T)	89.53±0.11	85.38±0.13	73.01±0.54	64.23±0.64	56.02±0.39	40.89±0.37
T-Back(T')	88.40±0.12	84.97±0.16	73.12±0.79	63.39±0.62	54.96±0.43	41.15±0.82
$T + T'$	89.64±0.16	85.47±0.32	73.39±0.40	64.95±0.91	56.36±0.51	41.94±0.43
Ours	**90.44**±0.19	**87.30**±0.25	**81.01**±0.25	**67.74**±0.17	**61.71**±0.20	**49.30**±0.82

- Comparison with SOTAs On Two Synthetic Datasets With Asymmetry Noise.

Method	Cifar-10			CIFAR-100		
	Asym-20%	Asym-40%	Asym-60%	Asym-20%	Asym-40%	Asym-60%
CE (baseline)	84.71±0.21	81.26±0.04	68.40±1.16	52.16±0.37	40.21±0.23	26.56±0.64
GCE [40]	89.54±0.21	85.95±0.40	79.55±0.51	65.66±0.73	57.34±0.35	45.46±0.16
PeerLoss [17]	88.98±0.15	85.61±0.59	77.03±0.49	57.51±0.05	43.95±0.35	30.02±0.39
Co-teaching [6]	85.90±0.38	83.09±0.44	71.69±0.50	57.21±0.37	43.76±0.46	30.18±0.71
Co-teaching++ [38]	87.13±0.07	84.86±0.36	73.50±0.47	58.79±0.35	45.26±0.41	32.02±1.22
T-Revision [33]	87.99±0.32	85.17±0.07	72.93±0.27	63.94±0.19	57.19±1.28	42.36±1.09
VolMinNet [13]	89.62±0.15	86.12±0.16	74.80±0.15	65.91±0.25	58.35±0.35	42.16±0.94
DualT [37]	89.36±0.44	86.59±0.30	78.89±0.99	65.76±0.56	56.90±0.39	44.61±1.20
T-For (T)	89.46±0.21	85.74±0.11	74.54±0.12	65.30±0.01	56.31±0.42	42.21±0.58
T-Back (T')	89.97±0.14	85.81±0.31	73.40±0.81	64.56±0.34	55.09±0.55	41.73±0.73
$T + T'$	89.62±0.24	86.25±0.03	74.80±0.21	65.52±0.28	57.10±0.20	42.72±0.29
Ours	**90.55**±0.03	**87.29**±0.05	**82.58**±0.24	**68.34**±0.24	**62.64**±0.49	**50.29**±0.24

Experiments

Comparison with SOTAs On Real-world Noisy Datasets.

Classification accuracy (%) on the Clothing1M dataset.

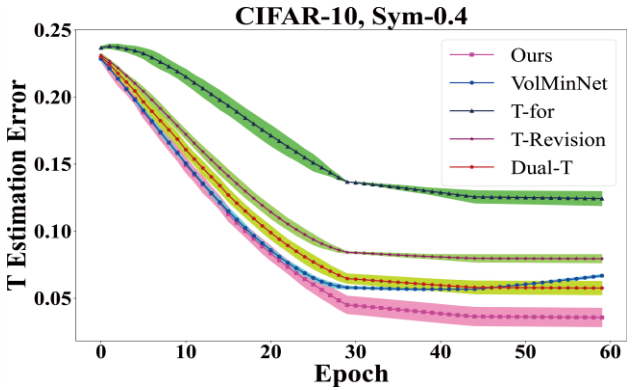
Methods	CE (Baseline)	GCE [40]	SL [26]	Co-teaching [6]	JointOpt [23]	L_{DMI} [35]
Accuracy	68.94	69.75	71.02	69.21	72.16	72.46
Methods	PTD-R-V [32]	ERL [15]	ForwardT [20]	JoCor [27]	CORES [5]	CAL [41]
Accuracy	71.67	72.87	69.84	70.30	73.24	74.17
Methods	MEIDTM [4]	VolMinNet* [13]	Ours	DivideMix* [11]	DivideMix+VolMinNet	DivideMix+Ours
Accuracy	73.05	69.82	70.73	74.67	74.83	**75.12**

Classification accuracy (%) on the Food101N dataset.

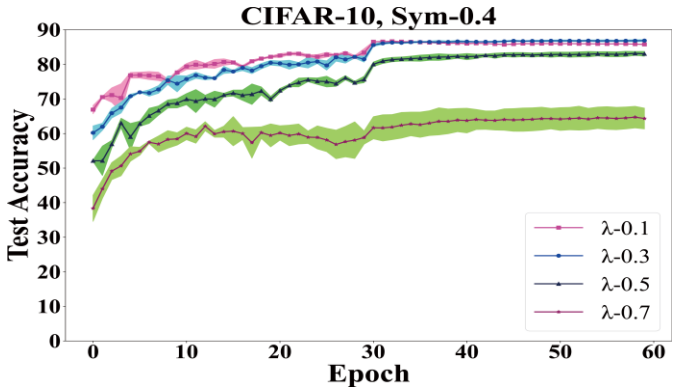
Methods	CE (Baseline)	CleanNet_{WH} [10]	CleanNet_{WS} [10]	DeepSelf [7]	NoiseResist [14]	VolMinNet* [13]
Accuracy	81.44	83.47	83.95	85.11	84.70	83.04
Methods	DivideMix* [11]	Ours	DivideMix+T-For (T)	DivideMix+T-Back (T')	DivideMix+VolMinNet	DivideMix+Ours
Accuracy	84.39	83.71	85.07	84.83	85.07	**86.11**

Ablation Study

Compare the estimation error of T between our method and other transition matrix based methods with symmetry noise on CIFAR-10.



Shows classification accuracy with various values of λ on CIFAR-10




XIDIAN UNIVERSITY
1931


NEURAL INFORMATION
PROCESSING SYSTEMS

THANK YOU