Normalized Loss Functions for Deep Learning with Noisy Labels

Xingjun Ma [*1] **Hanxun Huang** [*1] **Yisen Wang** [2] **Simone Romano** **Sarah Erfani** [1] **James Bailey** [1]

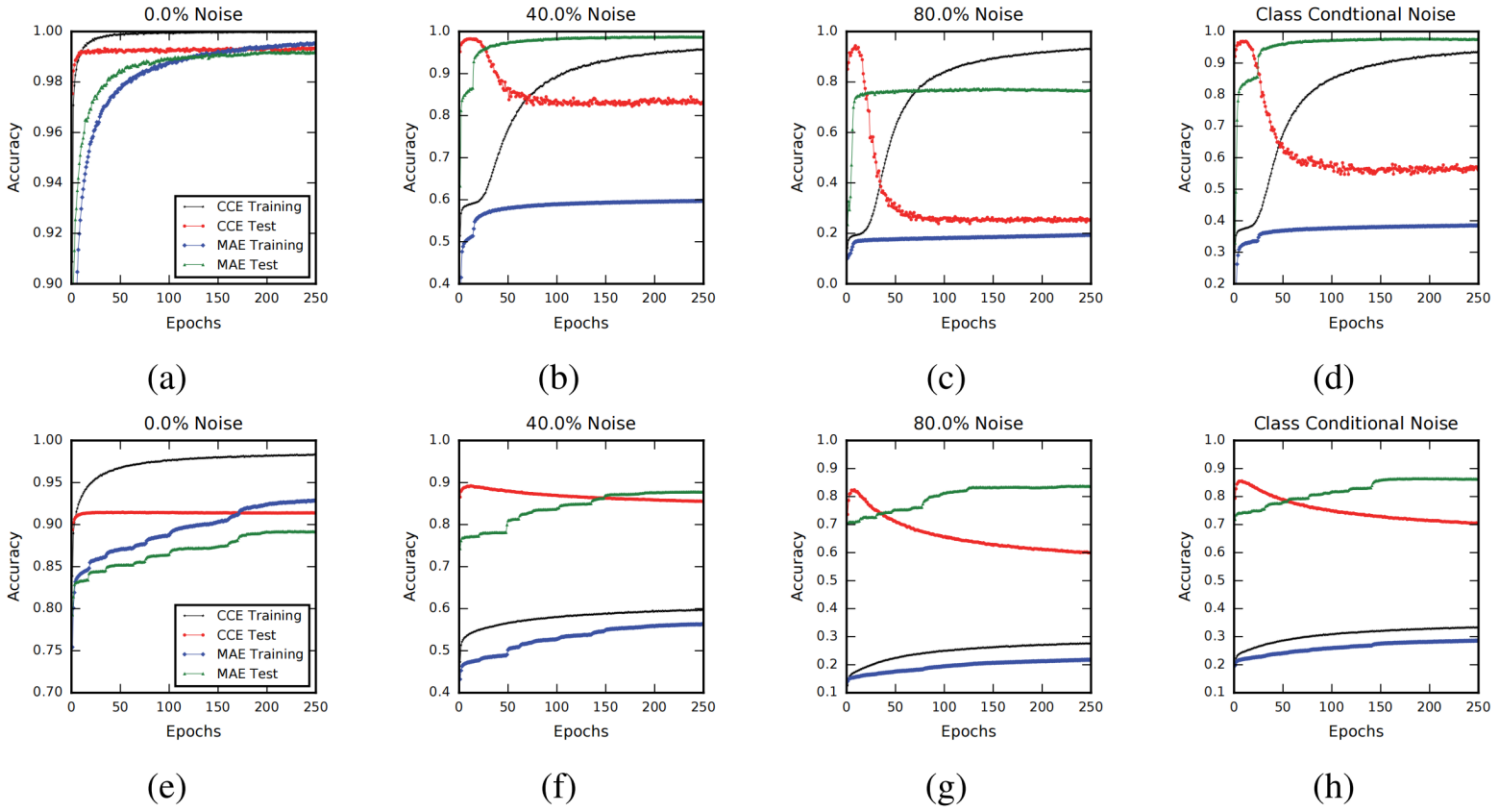
[1]**The University of Melbourne, Australia** [2]**Shanghai Jiao Tong University, China.**

ICML2020

Robust loss functions under label noise for deep neural networks

We call a loss function L symmetric if it satisfies, for some constant C

$$\sum_{i=1}^{k} L(f(\mathbf{x}), i) = C, \forall \mathbf{x} \in \mathcal{X}, \forall f$$



Proof 1 under symmetric noise

Theorem 1 *In a multi-class classification problem, let loss function L satisfy Eq 2. Then L is noise tolerant under symmetric or uniform label noise if $\eta < \frac{k-1}{k}$.*

Proof 1 *Recall that for any f,*

risk

$$R_L(f) = \mathbb{E}_{\mathbf{x},y_\mathbf{x}} L(f(\mathbf{x}), y_\mathbf{x}) \tag{3}$$

For uniform noise, we have, for any f,[1]

$$\begin{aligned} R_L^\eta(f) &= \mathbb{E}_{\mathbf{x},\hat{y}_\mathbf{x}} L(f(\mathbf{x}), \hat{y}_\mathbf{x}) \\ &= \mathbb{E}_\mathbf{x}\mathbb{E}_{y_\mathbf{x}|\mathbf{x}}\mathbb{E}_{\hat{y}_\mathbf{x}|\mathbf{x},y_\mathbf{x}} L(f(\mathbf{x}), \hat{y}_\mathbf{x}) \\ &= \mathbb{E}_\mathbf{x}\mathbb{E}_{y_\mathbf{x}|\mathbf{x}}\left[(1-\eta)L(f(\mathbf{x}), y_\mathbf{x}) + \frac{\eta}{k-1}\sum_{i\neq y_\mathbf{x}} L(f(\mathbf{x}), i)\right] \\ &= (1-\eta)R_L(f) + \frac{\eta}{k-1}(C - R_L(f)) \\ &= \frac{C\eta}{k-1} + \left(1 - \frac{\eta k}{k-1}\right) R_L(f). \end{aligned}$$

Thus, for any f,

$$R_L^\eta(f^*) - R_L^\eta(f) = (1 - \frac{\eta k}{k-1})(R_L(f^*) - R_L(f)) \leq 0$$

because $\eta < \frac{k-1}{k}$ and f^* is a minimizer of R_L. This proves f^* is also minimizer of risk under uniform noise.

Proof 2 under class-conditional noise

Theorem 3 *Suppose L satisfies Eq 2 and $0 \leq L(f(\mathbf{x}), i) \leq C/(k-1), \forall i \in [k]$. If $R_L(f^*) = 0$, then, L is noise tolerant under class conditional noise when $\bar{\eta}_{ij} < (1 - \eta_i), \forall j \neq i$, $\forall i, j \in [k]$.*

Proof 3 *For class-conditional noise, we have*

$$
\begin{aligned}
R_L^{\eta}(f) &= \mathbb{E}_{\mathcal{D}}(1 - \eta_{y_{\mathbf{x}}})L(f(\mathbf{x}), y_{\mathbf{x}}) + \mathbb{E}_{\mathcal{D}} \sum_{i \neq y_{\mathbf{x}}} \bar{\eta}_{y_{\mathbf{x}} i} L(f(\mathbf{x}), i) \\
&= \mathbb{E}_{\mathcal{D}}(1 - \eta_{y_{\mathbf{x}}})(C - \sum_{i \neq y_x} L(f(\mathbf{x}), i)) \\
&\quad + \mathbb{E}_{\mathcal{D}} \sum_{i \neq y_x} \bar{\eta}_{y_{\mathbf{x}} i} L(f(x), i) \\
&= C\mathbb{E}_{\mathcal{D}}(1 - \eta_{y_{\mathbf{x}}}) - \mathbb{E}_{\mathcal{D}} \sum_{i \neq y_{\mathbf{x}}} (1 - \eta_{y_{\mathbf{x}}} - \bar{\eta}_{y_{\mathbf{x}} i}) L(f(\mathbf{x}), i)
\end{aligned}
\tag{6}
$$

Since f_η^* is the minimizer of R_L^η, we have $R_L^\eta(f_\eta^*) - R_L^\eta(f^*) \leq 0$ and hence from Eq.(6) we have

$$
\mathbb{E}_{\mathcal{D}} \sum_{i \neq y_{\mathbf{x}}} (1 - \eta_{y_{\mathbf{x}}} - \bar{\eta}_{y_{\mathbf{x}} i}) (L(f^*(\mathbf{x}), i) - L(f_\eta^*(\mathbf{x}), i)) \leq 0 \tag{7}
$$

> 0 $\quad$ ≥ 0

$$
R_L(f^*) = L(f^*(\mathbf{x}), y_{\mathbf{x}}) = 0 \Longrightarrow L(f^*(\mathbf{x}), i) = C/(k-1), \quad i \neq y_{\mathbf{x}}
$$

$$
R_L(f_\eta^*) = L(f_\eta^*(x), y_{\mathbf{x}}) \geq 0 \Longrightarrow L(f_\eta^*(\mathbf{x}), i) \leq C/(k-1), \forall i
$$

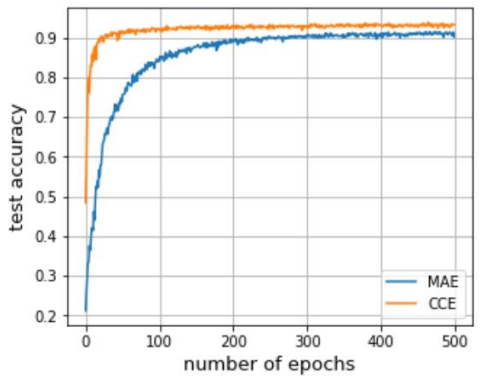
$$
L(f_\eta^*(\mathbf{x}), i) = C/(k-1) \Longrightarrow L(f_\eta^*(\mathbf{x}), y_{\mathbf{x}}) = 0
$$

$$
\Downarrow
$$

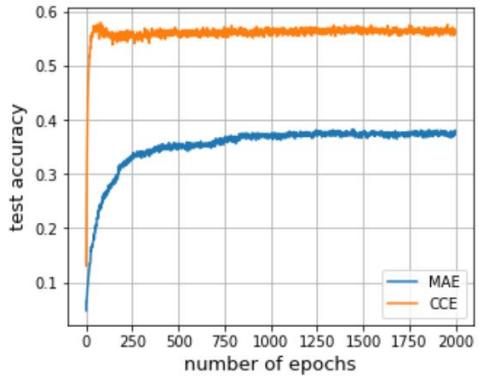
$$
f_\eta^*(\mathbf{x}) = f^*(\mathbf{x})
$$

GCE-Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels

MAE is robust but underperform Cross Entropy



CIFAR-10(clean)



CIFAR-100(clean)

$$\sum_{i=1}^{n} \frac{\partial \mathcal{L}(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i)}{\partial \boldsymbol{\theta}} = \begin{cases} \sum_{i=1}^{n} -\frac{1}{f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta}) & \text{for CCE} \\ \sum_{i=1}^{n} -\nabla_{\boldsymbol{\theta}} f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta}) & \text{for MAE/unhinged loss.} \end{cases}$$

$$\mathcal{L}_q(f(\boldsymbol{x}), \boldsymbol{e}_j) = \frac{(1 - f_j(\boldsymbol{x})^q)}{q} \quad \text{where } q \in (0, 1]$$

q→0

$$\lim_{q \to 0} \mathcal{L}_q(f(\boldsymbol{x}), \boldsymbol{e}_j) = -\ln f_j(x)$$

q=1

$$\mathcal{L}_q(f(\boldsymbol{x}), \boldsymbol{e}_j) = 1 - f_j(x)$$

$$\mathcal{L}_{MAE}(f(\boldsymbol{x}), \boldsymbol{e}_j) = \|\boldsymbol{e}_j - f(\boldsymbol{x})\|_1 = 2 - 2f_j(\boldsymbol{x})$$

SL/SCE-Symmetric Cross Entropy for Robust Learning with Noisy Labels

$$\ell_{ce} = -\sum_{k=1}^{K} q(k \mid \mathbf{x}) \log p(k \mid \mathbf{x})$$

Clean targets

dist of true labels

If given labels are noisy ? (q is not clean)

$$\ell_{rce} = -\sum_{k=1}^{K} p(k \mid \mathbf{x}) \log q(k \mid \mathbf{x})$$

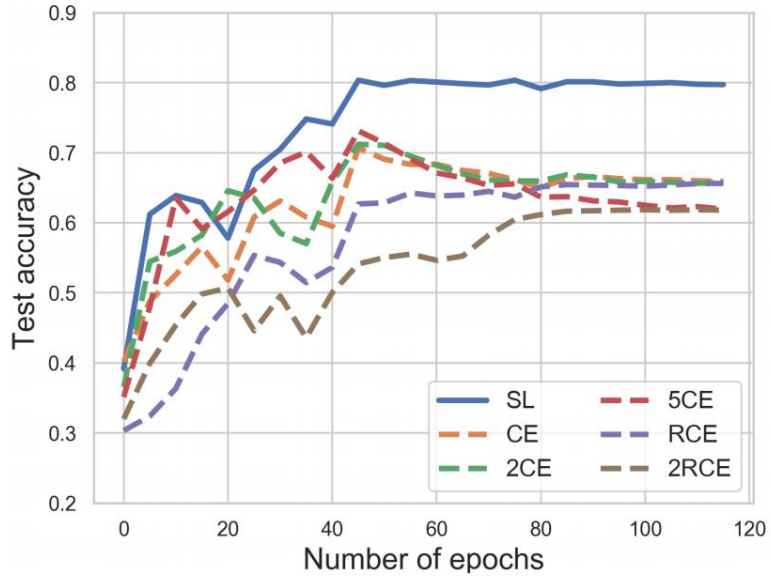
$\log(\boldsymbol{q}(k \neq y|\boldsymbol{x})) = A$ when $\boldsymbol{q}(k \neq y|\boldsymbol{x}) = 0$

$$\ell_{rce} = -\sum_{k=1}^{K} p(k \mid \mathbf{x}) \log q(k \mid \mathbf{x})$$

$$= -p(y \mid \mathbf{x}) \log 1 - \sum_{k \neq y} p(k \mid \mathbf{x}) A = -A \sum_{k \neq y} p(k \mid \mathbf{x})$$

$$= -A(1 - p(y \mid \mathbf{x}))$$

$$\sum_{k=1}^{K} \ell_{rce}(f(x), j) = A(1 - K)$$



CIFAR-10 with
60% symmetric label noise.

$$\ell = \alpha \ell_{ce} + \beta \ell_{rce}$$

Motivations

1. If a loss function satisfies symmetric condition, it is noise-robust under symmetric and class-conditional condition
2. CE is non-robust but better for convergence, robust loss functions (at least MAE and RCE) suffer from underfitting
3. Combine CE and a robust loss function can make them benefit from each other

$$\mathcal{L}_{\text{norm}} = \frac{\mathcal{L}(f(\boldsymbol{x}), y)}{\sum_{j=1}^{K} \mathcal{L}(f(\boldsymbol{x}), j)}.$$

Is normalized loss term still robust?

Robust loss functions suffer from underfitting

$$NCE = \frac{-\sum_{k=1}^{K} \boldsymbol{q}(k|\boldsymbol{x}) \log \boldsymbol{p}(k|\boldsymbol{x})}{-\sum_{j=1}^{K}\sum_{k=1}^{K} \boldsymbol{q}(y=j|\boldsymbol{x}) \log \boldsymbol{p}(k|\boldsymbol{x})} = \log_{\prod_k^K \boldsymbol{p}(k|\boldsymbol{x})} \boldsymbol{p}(y|\boldsymbol{x}),$$

$$NMAE = \frac{\sum_{k=1}^{K} |\boldsymbol{p}(k|\boldsymbol{x}) - \boldsymbol{q}(k|\boldsymbol{x})|}{\sum_{j=1}^{K}\sum_{k=1}^{K} |\boldsymbol{p}(k|\boldsymbol{x}) - \boldsymbol{q}(y=j|\boldsymbol{x})|} = \frac{1}{K-1}(1 - \boldsymbol{p}(y|\boldsymbol{x})) = \frac{1}{2(K-1)} \cdot MAE.$$

$$NRCE = \frac{-\sum_{k=1}^{K} \boldsymbol{p}(k|\boldsymbol{x}) \log \boldsymbol{q}(k|\boldsymbol{x})}{-\sum_{j=1}^{K}\sum_{k=1}^{K} \boldsymbol{p}(k|\boldsymbol{x}) \log \boldsymbol{q}(y=j|\boldsymbol{x})} = \frac{1}{K-1}(1 - \boldsymbol{p}(y|\boldsymbol{x})) = \frac{1}{A(K-1)} \cdot RCE$$

$$NFL = \frac{-\sum_{k=1}^{K} \boldsymbol{q}(k|\boldsymbol{x})(1-\boldsymbol{p}(k|\boldsymbol{x}))^{\gamma} \log \boldsymbol{p}(k|\boldsymbol{x})}{-\sum_{j=1}^{K}\sum_{k=1}^{K} \boldsymbol{q}(y=j|\boldsymbol{x})(1-\boldsymbol{p}(k|\boldsymbol{x}))^{\gamma} \log \boldsymbol{p}(k|\boldsymbol{x})} = \log_{\prod_k^K (1-\boldsymbol{p}(k|\boldsymbol{x}))^{\gamma}\boldsymbol{p}(k|\boldsymbol{x})} (1-\boldsymbol{p}(y|\boldsymbol{x}))^{\gamma}\boldsymbol{p}(y|\boldsymbol{x}).$$

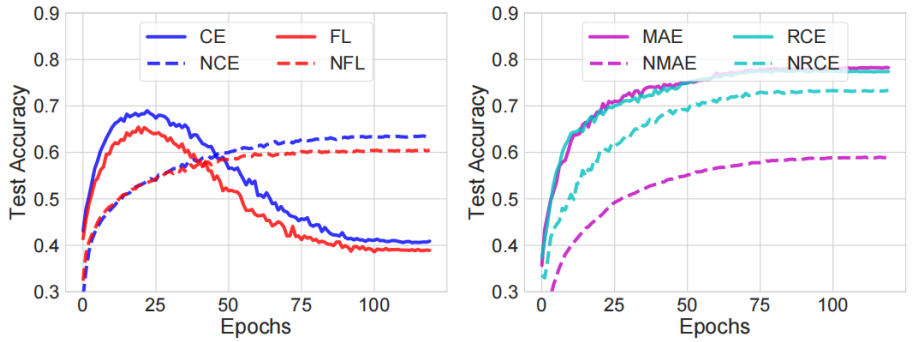


Figure 2. Test accuracies of unnormalized versus normalized loss functions on CIFAR-10 under 0.6 symmetric noise.

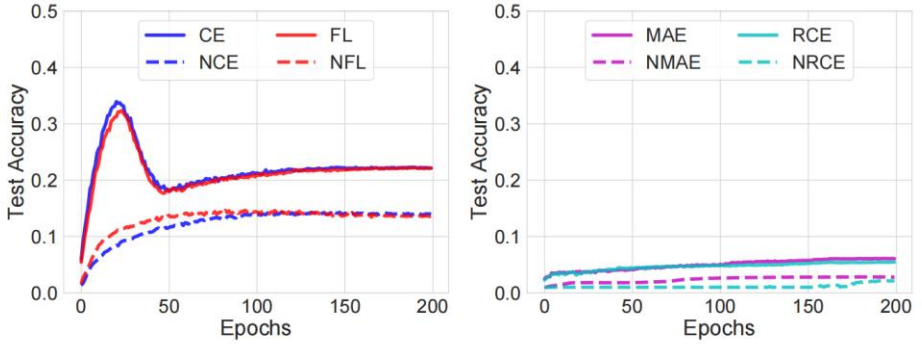


Figure 1. Test accuracies of unnormalized versus normalized loss functions on CIFAR-100 under 0.6 symmetric noise.

Active Loss and Passive Loss

Definition 1. *(Active loss function)* $\mathcal{L}_{Active}$ *is an active loss function if* $\forall(\boldsymbol{x}, y) \in \mathcal{D}\ \forall k \neq y\ \ell(f(\boldsymbol{x}), k) = 0.$

Definition 2. *(Passive loss function)* $\mathcal{L}_{Passive}$ *is a passive loss function if* $\forall(\boldsymbol{x}, y) \in \mathcal{D}\ \exists k \neq y\ \ell(f(\boldsymbol{x}), k) \neq 0.$

Only explictly optimize the target class

Besides explictly optimize other classes

Table 1. Examples of active and passive loss functions.

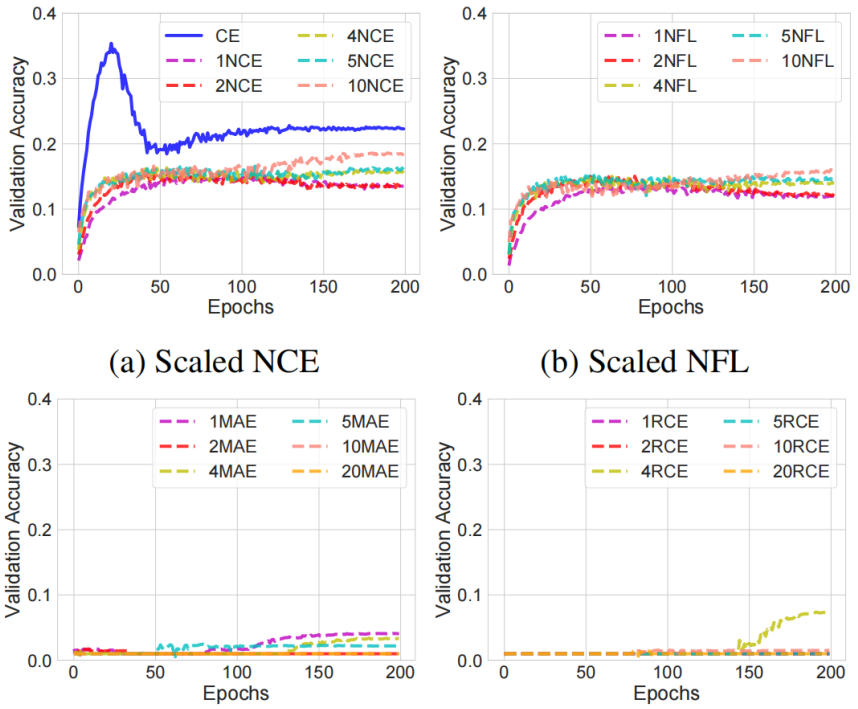
Loss Type	Active	Passive
Examples	CE, NCE, FL, NFL	MAE, NMAE, RCE, NRCE

$$\mathcal{L}_{\mathrm{APL}} = \alpha \cdot \mathcal{L}_{\mathrm{Active}} + \beta \cdot \mathcal{L}_{\mathrm{Passive}}$$

A robust loss function with two complementary loss

Q1: complementary is necessary?

Q2: robust L_{Active} is necessary?

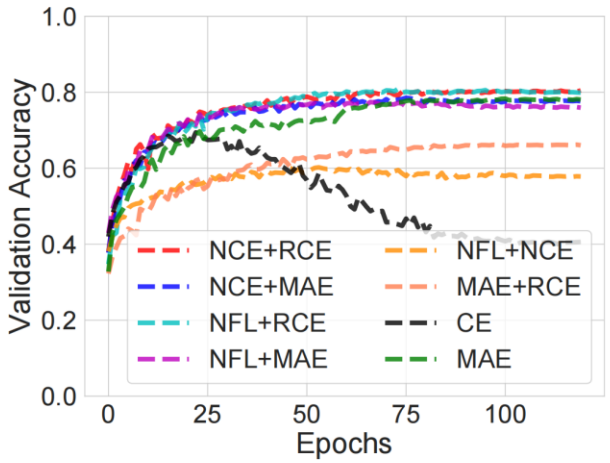


(a) Scaled NCE (b) Scaled NFL

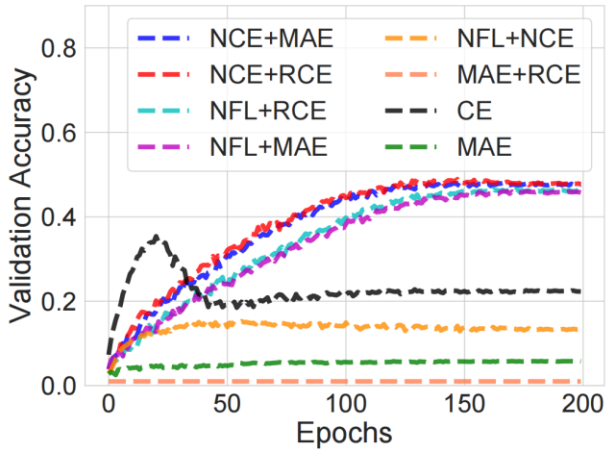
(c) Scaled MAE (d) Scaled RCE

Figure 3. Test accuracies of scaled loss functions on CIFAR-100 with 0.6 symmetric noise.

Experiments



(a) CIFAR-10



(b) CIFAR-100

Under 60% symmetric noise

Active + Passive:
NCE + RCE, NCE + MAE, NFL + RCE, NFL + MAE

Active + Active :
NFL + NCE

Passive + Passive :
MAE + RCE

Baselines:
CE, MAE

Experiments-Symmetric Noise

Table 2. Test accuracies (%) of different methods on benchmark datasets with clean or symmetric label noise ($\eta \in [0.2, 0.8]$). The results (mean±std) are reported over 3 random runs and the top 2 best results are **boldfaced**.

Datasets	Methods	Clean (η=0.0)	Symmetric Noise Rate (η) 0.2	0.4	0.6	0.8
MNIST	CE	99.25 ± 0.08	97.42 ± 0.06	94.21 ± 0.54	86.00 ± 1.48	47.08 ± 1.15
	FL	99.30 ± 0.02	97.45 ± 0.19	94.71 ± 0.25	85.76 ± 1.85	49.77 ± 2.26
	GCE	99.27 ± 0.01	99.18 ± 0.06	98.72 ± 0.05	97.43 ± 0.23	12.77 ± 2.00
	NLNL	99.27 ± 0.02	97.49 ± 0.30	96.64 ± 0.52	97.22 ± 0.06	10.32 ± 0.73
	SCE	99.24 ± 0.08	99.15 ± 0.04	98.78 ± 0.09	97.45 ± 0.29	73.70 ± 0.84
	NFL+MAE	99.39 ± 0.04	99.12 ± 0.06	98.74 ± 0.14	96.91 ± 0.09	**74.98 ± 1.99**
	NFL+RCE	99.38 ± 0.02	**99.19 ± 0.06**	**98.79 ± 0.10**	**97.46 ± 0.03**	74.59 ± 2.23
	NCE+MAE	99.37 ± 0.02	99.14 ± 0.05	98.78 ± 0.00	96.76 ± 0.34	74.66 ± 1.11
	NCE+RCE	99.37 ± 0.02	**99.20 ± 0.04**	**98.79 ± 0.12**	**97.48 ± 0.13**	**75.18 ± 1.19**
CIFAR-10	CE	90.36 ± 0.03	75.90 ± 0.28	60.28 ± 0.27	40.90 ± 0.35	19.65 ± 0.46
	FL	89.63 ± 0.25	74.59 ± 0.49	57.55 ± 0.39	38.91 ± 0.62	19.43 ± 0.27
	GCE	89.38 ± 0.23	87.27 ± 0.21	83.33 ± 0.39	72.00 ± 0.37	29.08 ± 0.80
	NLNL	91.93 ± 0.20	83.98 ± 0.18	76.58 ± 0.44	72.85 ± 0.39	51.41 ± 0.85
	SCE	91.30 ± 0.22	88.05 ± 0.26	82.06 ± 0.24	66.08 ± 0.25	30.69 ± 0.63
	NFL+MAE	89.25 ± 0.19	87.33 ± 0.14	83.81 ± 0.06	76.36 ± 0.31	45.23 ± 0.52
	NFL+RCE	90.91 ± 0.02	**89.14 ± 0.13**	**86.05 ± 0.12**	**79.78 ± 0.13**	**55.06 ± 1.08**
	NCE+MAE	88.83 ± 0.34	87.12 ± 0.21	84.19 ± 0.43	77.61 ± 0.05	49.62 ± 0.72
	NCE+RCE	90.76 ± 0.22	**89.22 ± 0.27**	**86.02 ± 0.09**	**79.78 ± 0.50**	**52.71 ± 1.90**
CIFAR-100	CE	70.89 ± 0.22	56.99 ± 0.41	41.40 ± 0.36	22.15 ± 0.40	7.58 ± 0.44
	FL	70.61 ± 0.44	56.10 ± 0.48	40.77 ± 0.62	22.14 ± 1.00	7.21 ± 0.25
	GCE	69.00 ± 0.56	65.24 ± 0.56	58.94 ± 0.50	45.18 ± 0.93	16.18 ± 0.46
	NLNL	68.72 ± 0.60	46.99 ± 0.91	30.29 ± 1.64	16.60 ± 0.90	11.01 ± 2.48
	SCE	70.38 ± 0.45	55.39 ± 0.18	39.99 ± 0.59	22.35 ± 0.65	7.57 ± 0.28
	NFL+MAE	67.98 ± 0.52	63.58 ± 0.09	58.18 ± 0.08	46.10 ± 0.50	24.78 ± 0.82
	NFL+RCE	68.23 ± 0.62	64.52 ± 0.35	58.20 ± 0.31	46.30 ± 0.45	25.16 ± 0.55
	NCE+MAE	68.75 ± 0.54	**65.25 ± 0.62**	**59.22 ± 0.36**	**48.06 ± 0.34**	**25.50 ± 0.76**
	NCE+RCE	69.02 ± 0.11	**65.31 ± 0.07**	**59.48 ± 0.56**	**47.12 ± 0.62**	**25.80 ± 1.12**

Experiments-Class-conditional Noise

Table 3. Test accuracy (%) of different methods on benchmark datasets with clean or asymmetric label noise ($\eta \in [0.1, 0.4]$). The results (mean±std) are reported over 3 random runs and the top 2 best results are **boldfaced**.

Datasets	Methods	Asymmetric Noise Rate (η) 0.1	0.2	0.3	0.4
MNIST	CE	98.53 ± 0.11	96.75 ± 0.31	92.98 ± 1.41	85.74 ± 2.70
	FL	98.97 ± 0.10	98.35 ± 0.17	96.57 ± 0.36	91.18 ± 2.02
	GCE	99.25 ± 0.03	99.11 ± 0.04	96.99 ± 0.53	88.56 ± 2.40
	NLNL	98.38 ± 0.17	95.98 ± 0.58	91.52 ± 1.14	86.36 ± 0.40
	SCE	99.15 ± 0.07	99.05 ± 0.05	97.96 ± 0.40	91.89 ± 3.32
	NFL+MAE	99.31 ± 0.05	99.09 ± 0.12	97.88 ± 0.16	$\mathbf{93.52 \pm 0.19}$
	NFL+RCE	$\mathbf{99.33 \pm 0.06}$	99.13 ± 0.01	$\mathbf{97.99 \pm 0.05}$	$\mathbf{93.59 \pm 0.82}$
	NCE+MAE	99.26 ± 0.02	$\mathbf{99.21 \pm 0.04}$	$\mathbf{98.99 \pm 0.03}$	93.40 ± 1.28
	NCE+RCE	$\mathbf{99.34 \pm 0.06}$	$\mathbf{99.17 \pm 0.02}$	97.94 ± 0.21	93.12 ± 1.17
CIFAR-10	CE	87.38 ± 0.16	83.62 ± 0.15	79.38 ± 0.28	75.00 ± 0.50
	FL	86.35 ± 0.30	82.97 ± 0.14	79.48 ± 0.21	74.60 ± 0.15
	GCE	88.42 ± 0.07	86.07 ± 0.31	80.78 ± 0.21	74.98 ± 0.32
	NLNL	88.54 ± 0.25	84.74 ± 0.08	81.26 ± 0.43	76.97 ± 0.52
	SCE	88.13 ± 0.21	83.92 ± 0.07	79.70 ± 0.27	78.20 ± 0.03
	NFL+MAE	88.46 ± 0.20	86.81 ± 0.32	83.91 ± 0.34	77.16 ± 0.10
	NFL+RCE	$\mathbf{90.20 \pm 0.15}$	$\mathbf{88.73 \pm 0.29}$	$\mathbf{85.74 \pm 0.22}$	$\mathbf{79.27 \pm 0.43}$
	NCE+MAE	88.25 ± 0.09	86.44 ± 0.23	83.98 ± 0.52	78.23 ± 0.42
	NCE+RCE	$\mathbf{89.95 \pm 0.20}$	$\mathbf{88.56 \pm 0.17}$	$\mathbf{85.58 \pm 0.44}$	$\mathbf{79.59 \pm 0.40}$
CIFAR-100	CE	65.42 ± 0.22	58.45 ± 0.45	51.09 ± 0.29	41.68 ± 0.45
	FL	64.79 ± 0.18	58.59 ± 0.81	51.26 ± 0.18	42.15 ± 0.44
	GCE	61.98 ± 0.81	59.99 ± 0.83	53.99 ± 0.29	41.49 ± 0.79
	NLNL	59.55 ± 1.22	50.19 ± 0.56	42.81 ± 1.13	35.10 ± 0.20
	SCE	64.15 ± 0.61	58.22 ± 0.47	49.85 ± 0.91	42.19 ± 0.19
	NFL+MAE	$\mathbf{66.06 \pm 0.23}$	$\mathbf{63.10 \pm 0.22}$	56.19 ± 0.61	43.51 ± 0.42
	NFL+RCE	$\mathbf{66.13 \pm 0.31}$	$\mathbf{63.12 \pm 0.41}$	54.72 ± 0.38	42.97 ± 1.03
	NCE+MAE	65.71 ± 0.34	62.38 ± 0.60	$\mathbf{58.02 \pm 0.48}$	$\mathbf{47.22 \pm 0.30}$
	NCE+RCE	65.68 ± 0.25	62.68 ± 0.79	$\mathbf{57.82 \pm 0.41}$	$\mathbf{46.79 \pm 0.96}$

Experiments-NGCE & real-world Noise

Table 4. Test accuracy (%) of APL losses NGCE+MAE and NGCE+RCE on CIFAR-10 under both symmetric and asymmetric noise. The top-2 best results are in **bold**.

Methods	Symmetric noise 0.4	Symmetric noise 0.8	Asymmetric noise 0.4
GCE	83.33 ± 0.39	29.08 ± 0.80	$\mathbf{74.98 \pm 0.32}$
NGCE+MAE	$\mathbf{84.14 \pm 0.15}$	$\mathbf{50.55 \pm 1.08}$	$\mathbf{76.55 \pm 0.48}$
NGCE+RCE	$\mathbf{85.76 \pm 0.26}$	$\mathbf{44.69 \pm 4.93}$	71.65 ± 0.68

Table 5. Top-1 validation accuracies (%) on clean ILSVRC12 validation set of ResNet-50 models trained on WebVision using different loss functions, under the Mini setting in (Jiang et al., 2018). The top-2 best results are in **bold**.

Loss	CE	GCE	SCE	**NCE+MAE**	**NCE+RCE**
Acc	58.88	53.68	61.76	**62.36**	**62.64**

Learning from Noisy Labels with Complementary Loss Functions

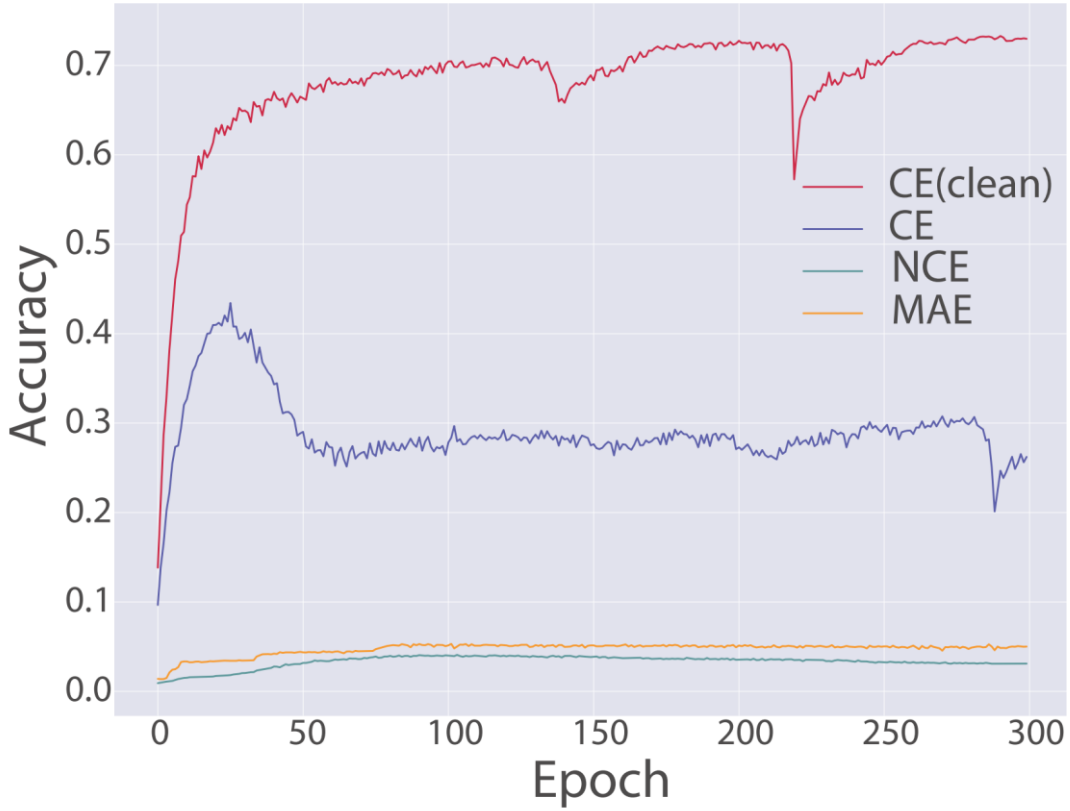
Deng-Bao Wang,[1,2] **Yong Wen,**[3] **Lujia Pan,**[4,3] **Min-Ling Zhang**[1,2,5*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
[3]Noah's Ark Lab, Huawei Technologies
[4]NSKEYLAB, Xi'an Jiaotong University
[5]Collaborative Innovation Center of Wireless Communications Technology, China
wangdb@seu.edu.cn, {wenyong4, panlujia}@huawei.com, zhangml@seu.edu.cn

AAAI 2021

Motivations

1. Cross entropy achieving good convergence but not robust (clean instances / easy instances)
2. Robust loss function suffers from underfitting (noisy instances / hard instances)



CIFAR-100 under 0.5 symmetric noise.

Learning with Complementary Losses

Algorithm 1: Learning with Complementary Losses.

Input: Training set $\mathcal{D}$, robust loss ℓ_{ROB}, T, T_{warm}, α, γ_1, γ_2, batch size B and Optimizer $\mathcal{O}$.

Output: Model parameters $\boldsymbol{\theta}$.

for $t = 1; t \leq T$ **do**
  for $b = 1; b \leq \frac{|\mathcal{D}|}{B}$ **do**
    Fetch mini-batch $\mathcal{B}$ from $\mathcal{D}$;
    if $t < T_{warm}$ **then** ▷ Warm up
      $\mathcal{L} = \sum_{(\boldsymbol{x},\bar{y})\in\mathcal{B}} \ell_{CE}(f(\boldsymbol{x}), \bar{y})$;
      Update $\boldsymbol{\theta} = \mathcal{O}(\mathcal{L}, \boldsymbol{\theta})$;
    end
    else ▷ Main train
      Calculate the average prediction for each sample using Eq. (7);
      Obtain $\widetilde{\mathcal{B}}$ using Eq. (8);
      Obtain $\widetilde{\mathcal{B}}'$ using Eq. (9);
      Obtain $\bar{\mathcal{B}}$ using Eq. (10);
      Calculate $\mathcal{L}_{CE} = \sum_{(\boldsymbol{x},\widetilde{y})\in\widetilde{\mathcal{B}}'} \ell_{CE}(f(\boldsymbol{x}), \widetilde{y})$;
      Calculate $\mathcal{L}_{ROB} = \sum_{(\boldsymbol{x},\bar{y})\in\bar{\mathcal{B}}} \ell_{ROB}(f(\boldsymbol{x}), \bar{y})$;
      $\mathcal{L}_{CL} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{ROB}$;
      Update $\boldsymbol{\theta} = \mathcal{O}(\mathcal{L}_{CL}, \boldsymbol{\theta})$
    end
  end
  Preserve the model outputs of current epoch;
end
return θ.

$$\boldsymbol{p}(\boldsymbol{x}) = \frac{1}{w}\sum_{i=k-w}^{k-1} f^{(i)}(\boldsymbol{x})$$ Model ensembling

$\widetilde{\mathcal{B}} = \{(\boldsymbol{x}, \widetilde{y}) | (\boldsymbol{x}, \bar{y}) \in \mathcal{B}\}$ where $\widetilde{y} = \arg\max_{j\in[c]} p_j$ Pseudo labels

$\widetilde{\mathcal{B}}' = \arg\min_{\mathcal{B}':|\mathcal{B}'|\geq\gamma_1|\mathcal{B}|} \sum_{(\boldsymbol{x},y)\in\widetilde{\mathcal{B}}'} H(\boldsymbol{p}(\boldsymbol{x}))$ Easy instances

$\bar{\mathcal{B}} = \arg\max_{\mathcal{B}':|\mathcal{B}'|\geq\gamma_2|\mathcal{B}|} \sum_{(\boldsymbol{x},y)\in\mathcal{B}'} E(\boldsymbol{p}(\boldsymbol{x}))$ Hard instances

Experiments(CIFAR-10)

Noise type	Uniform			Class-conditional		
Noise ratio	0.2	0.4	0.6	0.2	0.3	0.4
Standard CE	83.78±0.32	67.73±0.72	47.79±0.82	86.54±0.52	81.86±0.70	76.02±0.93
GCE (2018)	90.44±0.08	88.08±0.13	81.13±0.21	90.30±0.16	88.68±0.10	84.77±0.87
SCE (2019)	91.68±0.17	87.54±0.47	78.88±0.69	89.91±0.58	84.86±0.77	76.52±1.33
APL (2020)	87.79±0.48	79.13±0.75	66.51±1.38	90.14±0.34	83.70±1.08	76.02±1.20
Ours (CE+MAE)	**93.49±0.14**	**92.04±0.20**	**89.37±0.14**	**94.10±0.17**	93.01±0.20	91.52±0.21
Ours (CE+APL)	92.20±0.54	90.47±0.87	88.01±0.55	94.08±0.14	**93.39±0.17**	**91.89±0.19**
M-correction (2019)	93.69±0.32	93.18±0.10	90.59±0.33	——	——	——
DivideMix (2020)	95.23±0.22	93.62±0.15	92.84±0.19	93.20±0.28	92.38±0.24	91.15±0.69
Ours* (CE)	93.62±0.21	93.21±0.17	92.33±0.24	93.83±0.18	92.82±0.22	91.39±0.26
Ours* (CE+MAE)	**95.37±0.09•**	**94.79±0.11•**	**93.59±0.19•**	**94.98±0.14•**	**94.33±0.24•**	**92.18±0.42•**
Ours* (CE+APL)	94.70±0.13	93.80±0.14	92.68±0.27	94.34±0.14	93.38±0.28	91.10±0.40

Experiments(CIFAR-100)

Noise type	Uniform			Class-conditional		
Noise ratio	0.2	0.4	0.6	0.2	0.3	0.4
Standard CE	62.82±0.27	49.10±0.37	31.11±0.36	63.43±0.20	54.37±0.30	44.89±0.40
GCE (2018)	69.43±0.20	64.44±0.14	55.96±0.47	68.96±0.20	64.68±0.22	50.67±0.44
SCE (2019)	61.60±0.31	46.21±0.40	30.05±0.73	62.18±0.21	53.67±0.30	44.29±0.39
APL (2020)	70.74±0.33	61.92±0.52	48.64±0.83	63.11±0.45	52.91±0.46	42.48±0.68
Ours (CE+MAE)	**71.63±0.26**	**68.61±0.35**	**62.52±0.22**	**69.99±0.23**	**68.32±0.20**	**65.72±0.28**
Ours (CE+APL)	71.26±0.36	68.50±0.28	62.30±0.23	69.54±0.17	67.71±0.23	65.39±0.52
M-correction (2019)	68.95±0.53	65.43±0.30	59.43±0.36	——	——	——
DivideMix (2020)	74.80±0.28	72.92±0.20	69.38±0.26	74.90±0.41	72.14±0.45	50.79±0.62
Ours* (CE)	71.04±0.40	69.20±0.27	65.09±0.33	67.80±0.37	66.12±0.36	65.19±0.29
Ours* (CE+MAE)	**77.61±0.22•**	**75.81±0.37•**	**72.17±0.30•**	**76.74±0.53•**	**75.32±0.39•**	68.07±0.79
Ours* (CE+APL)	77.21±0.39	75.62±0.27	71.34±0.60	76.50±0.41	74.47±0.50	**68.18±1.16•**

Experiments

Noise ratio	Uniform (0.5)	Class-cond. (0.3)
Standard CE	23.16±0.53	43.67±0.22
SCE	23.89±0.32	38.92±0.44
APL	5.43±0.14	Fail
Ours (CE+MAE)	50.59±0.33	51.99±0.29
Ours (CE+APL)	50.67±0.29	51.31±0.19
Ours* (CE)	48.98±0.32	45.88±0.44
Ours* (CE+MAE)	**56.73±0.35**	**57.27±0.40**
Ours* (CE+APL)	55.24±0.39	56.64±0.29

TinyImageNet

	CE	GCE	SCE	M-corr.	D-mix.	Ours
Acc.	68.80	69.75	71.02	71.00	74.76	73.59

clothing1M