SAiDL Assignment 2025
Core ML

Soham Kalburgi

CS&IS, BITS Pilani, KK Birla Goa Campus
f20230460@goa.bits-pilani.ac.in

1 Introduction

Robustness in machine learning is essential for handling real-world datasets, where noisy labels can degrade model performance. Traditional loss functions like Cross Entropy (CE) are sensitive to label noise, leading to *overfitting*. Normalized loss functions have been proposed to enhance robustness. However, while they mitigate noise sensitivity, they often suffer from a problem of *underfitting*.

This report provides an analysis of the effectiveness of normalized loss functions, such as Normalized Cross Entropy (NCE) and Normalized Focal Loss (NFL), in managing noisy labels. Additionally, the *Active Passive Loss* (APL) framework is explored, which pairs active and passive loss functions to balance robustness and performance. Through comparative analysis on the CIFAR-10 dataset [2] under varying noise rates, the trade-offs between noise resilience and model accuracy are assessed. Most of this report aligns with the official paper [3].

2 Theoretical Understanding

A loss function in machine learning quantifies the difference between a model's predictions and the actual labels. In simple terms, it measures how well or poorly the model is performing on a given task. Given a K-class dataset (e.g., $K = 10$ for CIFAR-10) with noisy labels as $\mathcal{D} = \{(\mathbf{x}, y)^{(i)}\}_{i=1}^{n}$ with $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ denoting a sample and $y \in \mathcal{Y} = \{1, \ldots, K\}$ its annotated label (possibly incorrect), the distribution over different labels for sample $\mathbf{x}$ is denoted by $\mathbf{q}(k|\mathbf{x})$ and $\sum_{k=1}^{K} \mathbf{q}(k|\mathbf{x}) = 1$. The main focus is on the common case where there is only one single label y for $\mathbf{x}$: i.e. $\mathbf{q}(y|\mathbf{x}) = 1$ and $\mathbf{q}(k \neq y|\mathbf{x}) = 0$.

2.1 Types of Noise

Under the assumption that the noise is conditionally independent to the inputs, label noise can be either *symmetric* (or uniform), or *asymmetric* (or class-conditional). The overall noise rate is denoted by $\eta \in [0, 1]$, and the class-wise noise rate from class j to class k by η_{jk}.

For symmetric noise, $\eta_{jk} = \frac{\eta}{K-1}$ for $j \neq k$ and $\eta_{jk} = 1 - \eta$ for $j = k$. For asymmetric noise, η_{jk} is conditioned on both the true class j and mislabeled class k.

2.2 Loss Functions

For a sample $\mathbf{x}$, the probability output of a classifier $f(\mathbf{x})$ is denoted as $\mathbf{p}(k|\mathbf{x}) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}$ where z_k denotes logits output of the network with respect to class k. Training a classifier f is to find a set of optimal parameters $\boldsymbol{\theta}$ that minimize the empirical risk defined by the loss function: $\boldsymbol{\theta} := \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \mathcal{L}(f(\mathbf{x}_i, y_i))$ where $\mathcal{L}(f(\mathbf{x}), y)$ is the loss of f w.r.t. label y.

The commonly used Cross Entropy (CE) loss on sample $\mathbf{x}$ is defined as (proven not robust to label noise):

$$CE = -\sum_{k=1}^{K} \mathbf{q}(k|\mathbf{x}) \log \mathbf{p}(k|\mathbf{x}) \tag{1}$$

Mean Absolute Error (MAE) is defined as (proven robust to label noise):

$$MAE = \sum_{k=1}^{K} |\mathbf{p}(k|\mathbf{x}) - \mathbf{q}(k|\mathbf{x})| \tag{2}$$

Reverse Cross Entropy (RCE) loss is defined as (proven robust to label noise):

$$RCE = -\sum_{k=1}^{K} \mathbf{p}(k|\mathbf{x}) \log \mathbf{q}(k|\mathbf{x}) \tag{3}$$

with $\mathbf{q}(k \neq y|\mathbf{x}) = 0$ truncated to a small value such that $\log(\mathbf{q}(k \neq y|\mathbf{x})) = A$.

Focal Loss (FL) is a generalization of CE and is defined as (proven not robust to label noise):

$$FL = -\sum_{k=1}^{K} \mathbf{q}(k|\mathbf{x})(1 - \mathbf{p}(k|\mathbf{x}))^{\gamma} \log \mathbf{p}(k|\mathbf{x}) \tag{4}$$

where $\gamma \geq 0$ is a tunable parameter. FL reduces to CE loss when $\gamma = 0$.

2.3 Normalized Loss Functions

A loss function is normalized as:

$$\mathcal{L}_{\text{norm}} = \frac{\mathcal{L}(f(\mathbf{x}), y)}{\sum_{j=1}^{K} \mathcal{L}(f(\mathbf{x}), j)} \tag{5}$$

A normalized loss has the property: $\mathcal{L}_{\text{norm}} \in [0, 1]$. Normalizing the loss functions in Section 2.2 gives:

1. The Normalized Cross Entropy (NCE) loss is defined as:

$$NCE = \frac{-\sum_{k=1}^{K} \mathbf{q}(k|\mathbf{x}) \log \mathbf{p}(k|\mathbf{x})}{-\sum_{j=1}^{K} \sum_{k=1}^{K} \mathbf{q}(y = j|\mathbf{x}) \log \mathbf{p}(k|\mathbf{x})} = \log_{\prod_{k}^{K} \mathbf{p}(k|\mathbf{x})} \mathbf{p}(y|\mathbf{x}) \tag{6}$$

2. The Normalized Mean Absolute Error (NMAE) is defined as:

$$NMAE = \frac{\sum_{k=1}^{K} |\mathbf{p}(k|\mathbf{x}) - \mathbf{q}(k|\mathbf{x})|}{\sum_{j=1}^{K} \sum_{k=1}^{K} |\mathbf{p}(k|\mathbf{x}) - \mathbf{q}(y = j|\mathbf{x})|} = \frac{1}{K-1}(1 - \mathbf{p}(y|\mathbf{x})) = \frac{1}{2(K-1)} \cdot MAE \tag{7}$$

   NMAE is simply a scaled version of MAE by a factor of $\frac{1}{2(K-1)}$.

3. The Normalized Reverse Cross Entropy (NRCE) loss is defined as:

$$NRCE = \frac{-\sum_{k=1}^{K} \mathbf{p}(k|\mathbf{x}) \log \mathbf{q}(k|\mathbf{x})}{-\sum_{j=1}^{K} \sum_{k=1}^{K} \mathbf{p}(k|\mathbf{x}) \log \mathbf{q}(y = j|\mathbf{x})} = \frac{1}{K-1}(1 - \mathbf{p}(y|\mathbf{x})) = \frac{1}{A(K-1)} \cdot RCE \tag{8}$$

   NRCE is a scaled version of RCE by a factor of $\frac{1}{A(K-1)}$.

4. The Normalized Focal Loss (NFL) is defined as:

$$\begin{aligned} NFL &= \frac{-\sum_{k=1}^{K} \mathbf{q}(k|\mathbf{x})(1 - \mathbf{p}(k|\mathbf{x}))^{\gamma} \log \mathbf{p}(k|\mathbf{x})}{-\sum_{j=1}^{K} \sum_{k=1}^{K} \mathbf{q}(y = j|\mathbf{x})(1 - \mathbf{p}(k|\mathbf{x}))^{\gamma} \log \mathbf{p}(k|\mathbf{x})} \\ &= \log_{\prod_{k}^{K} (1 - \mathbf{p}(k|\mathbf{x}))^{\gamma} \mathbf{p}(k|\mathbf{x})} (1 - \mathbf{p}(y|\mathbf{x}))^{\gamma} \mathbf{p}(y|\mathbf{x}) \end{aligned} \tag{9}$$

As NMAE and NRCE are scaled versions of MAE and RCE respectively, their robustness property is preserved. The normalization of CE and FL, however, derives new loss functions that are robust. To show noise tolerance, two lemmas (for symmetric and asymmetric noise) are stated in the paper. Stating them without their proofs:

Lemma 1. *In a multi-class classification problem, any normalized loss function $\mathcal{L}_{\text{norm}}$ is noise tolerant under symmetric (or uniform) label noise, if noise rate $\eta < \frac{K-1}{K}$.*

Lemma 2. *In a multi-class classification problem, given $R(f^*) = 0$ (risk) and $0 \leq \mathcal{L}_{norm}(f(\mathbf{x}), k) \leq \frac{1}{K-1}, \forall k$, any normalized loss function $\mathcal{L}_{norm}$ is noise tolerant under asymmetric (or class-conditional) label noise, if noise rate $\eta_{jk} < 1 - \eta_y$.*

2.4 Active Passive Loss Framework

The CE and FL losses become robust after normalization, however, this robustness leads to decreased model accuracy (supported empirically in Fig. 1). To address this *underfitting* problem, the paper proposes the APL framework. If $\mathcal{L}(f(x), y) = \sum_{k=1}^{K} \ell(f(x), k)$ then:

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

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

Intuitively, an *active loss* function (CE, NCE, FL, NFL) only optimizes for the correct class y, meaning it encourages the model to increase the probability $\mathbf{q}(k = y|x)$ while ignoring incorrect classes $k \neq y$, whereas, a *passive loss* function (MAE, NMAE, RCE, NRCE) explicitly minimizes the probabilities of alteast one incorrect class $k \neq y$ in addition to maximizing the probability of the correct class.

Definition of APL: To combine a robust active loss and a robust passive loss for both robust and sufficient learning.

$$\mathcal{L}_{\text{APL}} = \alpha \cdot \mathcal{L}_{\text{active}} + \beta \cdot \mathcal{L}_{\text{passive}} \tag{10}$$

where $\alpha, \beta > 0$ are parameters to balance the two terms. A nonrobust loss function should be normalized to be used in the APL scheme. To show noise tolerance, the paper states the following lemma:

Lemma 3. *$\forall\alpha, \forall\beta$, if $\mathcal{L}_{\text{active}}$ and $\mathcal{L}_{\text{passive}}$ are noise tolerant then $\mathcal{L}_{\text{APL}} = \alpha \cdot \mathcal{L}_{\text{active}} + \beta \cdot \mathcal{L}_{\text{passive}}$ is noise tolerant.*

There are four possible combinations satisfying the APL principle: (1) αNCE + βMAE (2) αNCE + βRCE (3) αNFL + βMAE (4) αNFL + βRCE.

3 Empirical Study

All experiments were conducted on CIFAR-10 dataset using the ResNet-18 architecture [1]. The networks were trained for 50 epochs with a batch size of 128 and an initial learning rate of 0.01 with cosine learning rate annealing. Stochastic Gradient Descent (SGD) optimizer with momentum 0.9 and weight decay 10^{-4} was used. Data augmentations including random crop and horizontal flip were applied. The final test accuracy is reported based on a single-run evaluation. For FL/NFL loss functions, $\gamma = 0.5$ and for RCE/NRCE loss functions, $A = -4$.

Symmetric noise was generated by flipping labels in each class randomly to incorrect labels of other classes. For asymmetric noise, TRUCK → AUTOMOBILE, BIRD → AIRPLANE, DEER → HORSE and CAT ↔ DOG were flipped.

The test accuracies for Sections 3.1 and 3.2 are reported in Table 1 (symmetric noise) and Table 2 (asymmetric noise).

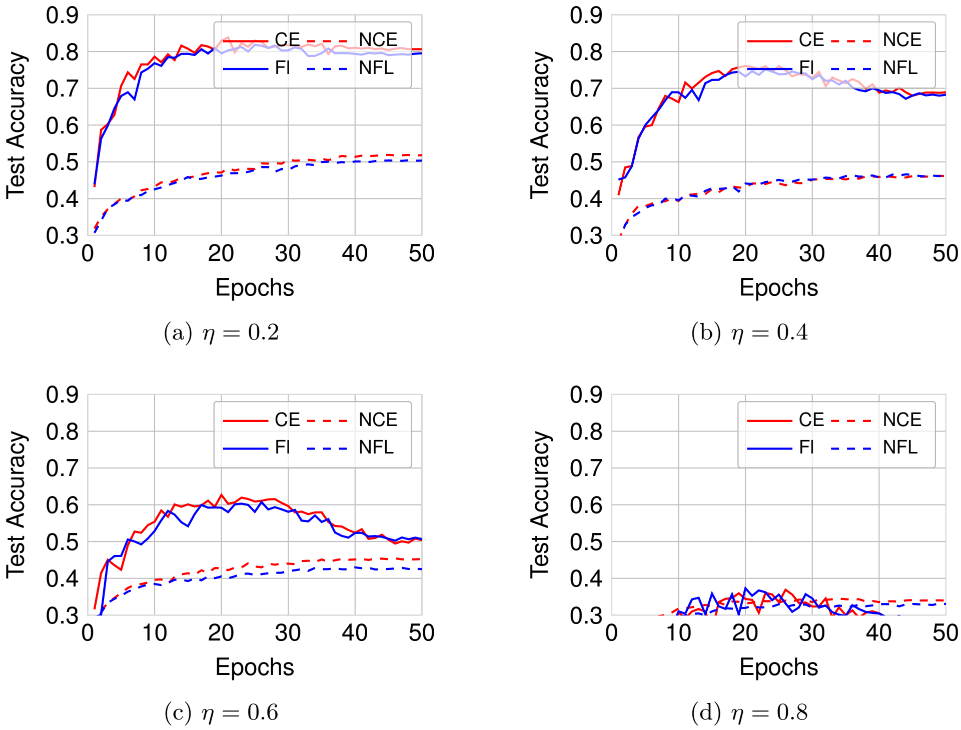
3.1 Normalized Loss Functions

As seen in Fig. 1, both CE and FL loss functions begin to overfit around epoch 25, where their test accuracies decline. By contrast, the normalized counterparts (NCE and NFL) continue to improve or remain stable, showing no signs of overfitting. This trend holds consistently across different levels of symmetric noise ($\eta \in [0.2, 0.8]$). As the noise rate increases, overfitting becomes more pronounced in the nonrobust losses (CE and FL), leading to a sharper drop in test accuracy. The normalized loss functions, though robust, face the problem of *underfitting*, leading to lower accuracies. This remains true for asymmetric noise as well, as seen in Fig. 3.

3.2 Active Passive Loss Framework

For all APL losses, α and β are set as 1.0. As seen in Fig. 2, at lower noise rates ($\eta = 0.2$) all four methods converge to high accuracies. As the noise rate increases, the methods NFL+RCE and NCE+RCE degrade more noticeably, while the MAE-based methods remain relatively stable, consistently outperforming the traditional CE loss function and its normalized form.

In the case of asymmetric noise, as seen in Fig. 4, all four methods converge to similar accuracies except $\eta = 0.4$ where RCE-based losses degrade.



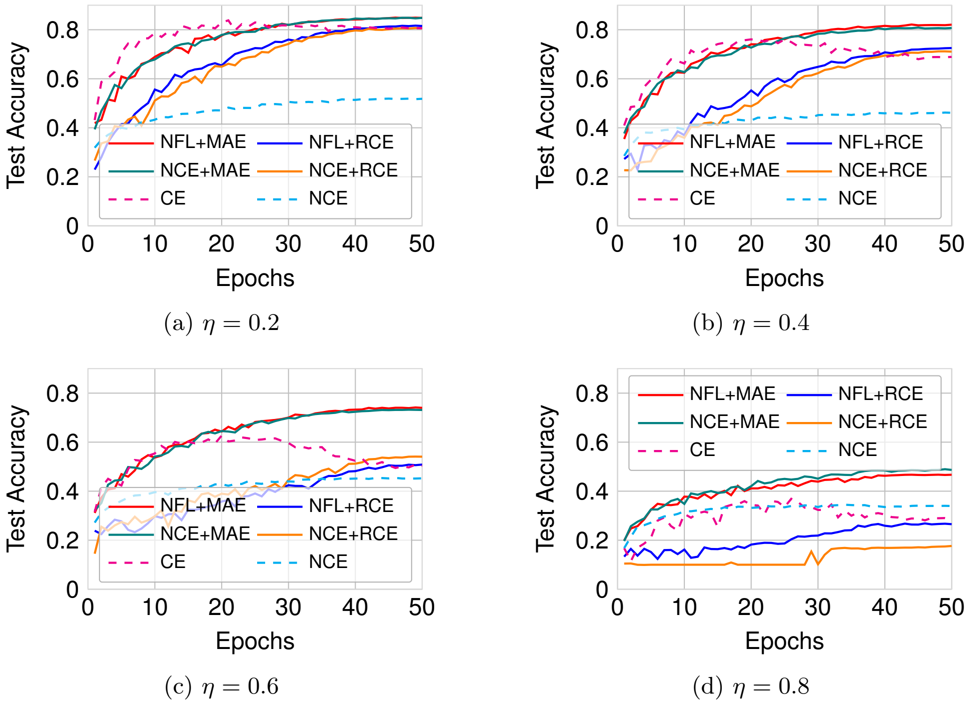
(a) $\eta = 0.2$

(b) $\eta = 0.4$

(c) $\eta = 0.6$

(d) $\eta = 0.8$

Fig. 1. Test accuracies of unnormalized (CE and FL) vs. normalized (NCE and NFL) loss functions on CIFAR-10 under symmetric noise rates $\eta \in [0.2, 0.8]$.



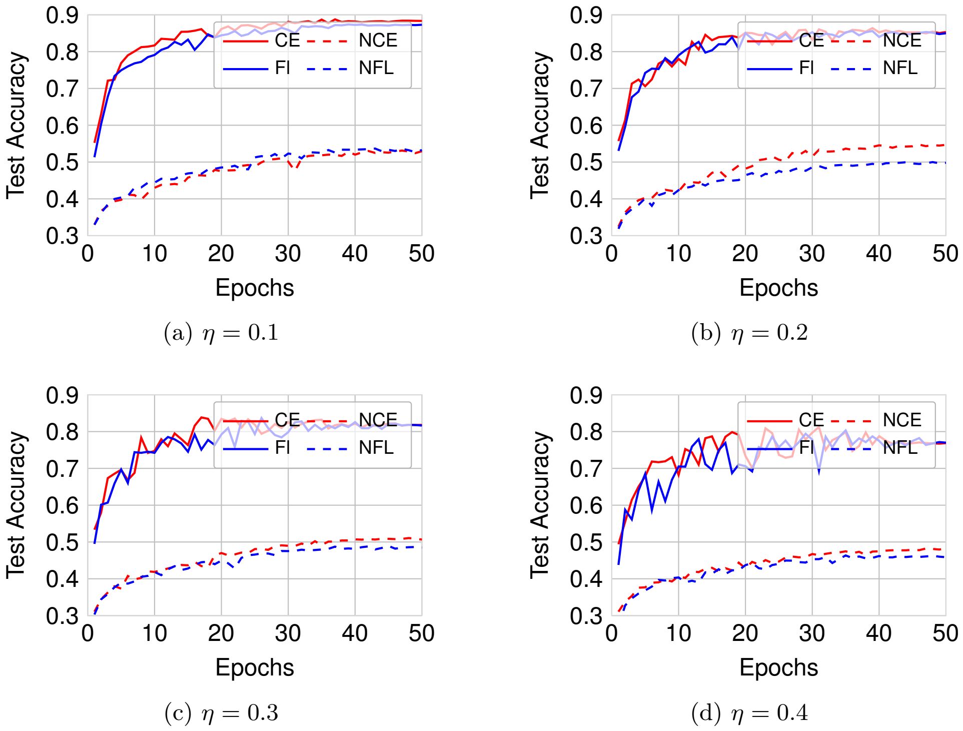
(a) $\eta = 0.2$

(b) $\eta = 0.4$

(c) $\eta = 0.6$

(d) $\eta = 0.8$

Fig. 2. Test accuracies of APL loss functions (NFL+MAE, NFL+RCE, NCE+MAE, NCE+RCE) on CIFAR-10 dataset with symmetric noise rates $\eta \in [0.2, 0.8]$.



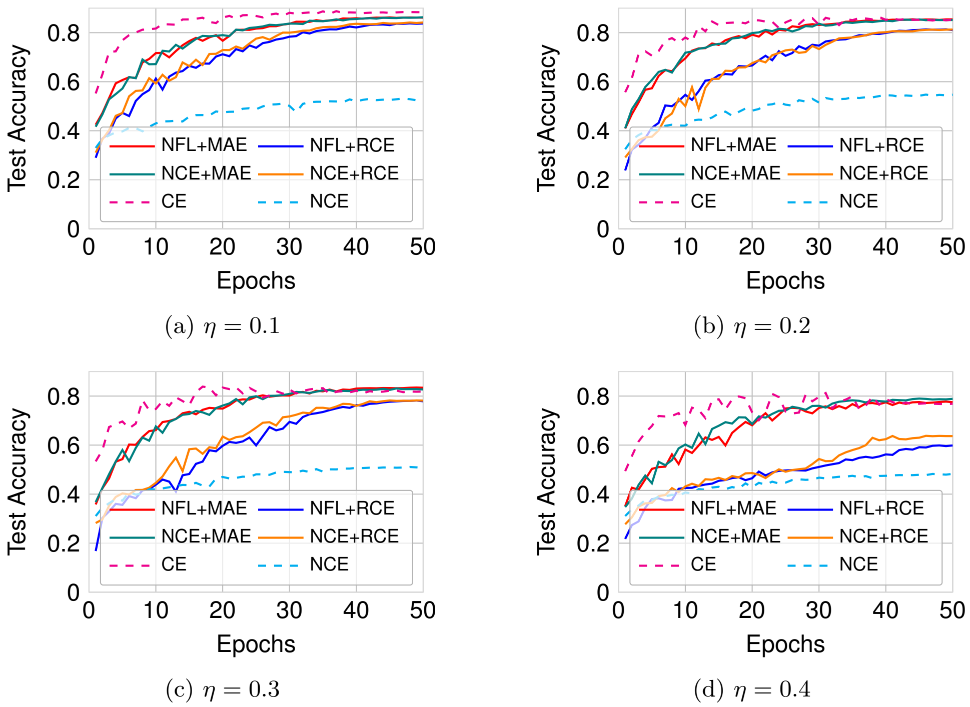
(a) $\eta = 0.1$

(b) $\eta = 0.2$

(c) $\eta = 0.3$

(d) $\eta = 0.4$

Fig. 3. Test accuracies of unnormalized (CE and FL) vs. normalized (NCE and NFL) loss functions on CIFAR-10 under asymmetric noise rates $\eta \in [0.1, 0.4]$.



(a) $\eta = 0.1$

(b) $\eta = 0.2$

(c) $\eta = 0.3$

(d) $\eta = 0.4$

Fig. 4. Test accuracies of APL loss functions (NFL+MAE, NFL+RCE, NCE+MAE, NCE+RCE) on CIFAR-10 dataset with asymmetric noise rates $\eta \in [0.1, 0.4]$.

Table 1. Test accuracies (%) of different loss functions on CIFAR-10 dataset with symmetric label noise $\eta \in [0.2, 0.8]$. Final accuracy is reported based on single-run evaluation. **Boldfaced** denotes the top-2 accuracies for a noise rate.

Dataset	Loss Functions	Symmetric Noise Rate (η)			
		0.2	0.4	0.6	0.8
CIFAR-10	CE	83.83	76.07	62.69	36.90
	FL	81.83	75.18	60.74	37.30
	NCE	51.90	46.21	45.43	34.48
	NFL	50.44	46.62	42.98	33.17
	NFL+MAE	**84.95**	**82.13**	**74.16**	**46.73**
	NFL+RCE	81.67	72.53	50.85	26.85
	NCE+MAE	**84.88**	**80.88**	**73.23**	**48.97**
	NCE+RCE	80.69	71.18	54.11	17.65

Table 2. Test accuracies (%) of different loss functions on CIFAR-10 dataset with asymmetric label noise $\eta \in [0.1, 0.4]$. Final accuracy is reported based on single-run evaluation. **Boldfaced** denotes the top-2 accuracies for a noise rate.

Dataset	Loss Functions	Asymmetric Noise Rate (η)			
		0.1	0.2	0.3	0.4
CIFAR-10	**CE**	**88.77**	**86.40**	**83.85**	**81.24**
	FL	**87.39**	**85.60**	**83.68**	**80.32**
	NCE	53.12	54.74	51.07	48.27
	NFL	53.66	49.99	48.93	46.35
	NFL+MAE	86.24	85.46	83.49	77.74
	NFL+RCE	83.79	81.37	78.17	59.87
	NCE+MAE	86.29	85.41	83.11	78.98
	NCE+RCE	84.16	81.29	78.22	63.78

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385
2. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto (2009)
3. Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., Bailey, J.: Normalized loss functions for deep learning with noisy labels. In: ICML (2020)