ROBUST CURRICULUM LEARNING: FROM CLEAN LABEL DETECTION TO NOISY LABEL SELF-CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural nets training can easily overfit to noisy labels and end with poor generalization performance. Existing methods address this problem by (1) filtering out the noisy data and only using the clean data for training or (2) relabeling the noisy data by the model in-training or another model trained on a clean dataset. However, the former strategy ignores the wrongly-labeled data's useful information, while the latter may introduce extra noise if the relabeling quality is poor. In this paper, we propose a smooth transition and interplay between these two strategies as a curriculum that selects training samples by a dynamic criterion. In particular, we start with learning from clean data and then gradually move to learn noisy-labeled data with pseudo labels produced by a time-ensemble of the model and data augmentations. Unlike the instantaneous loss widely used for noise detection, our data selection is based on the dynamics of both the loss and output consistency for each sample across training steps and different data augmentations, resulting in more precise detection of both clean labels and correct pseudo labels. On multiple benchmarks of noisy labels, we show that our curriculum learning strategy can significantly improve the test accuracy without any auxiliary model or extra clean data.

1 INTRODUCTION

The expressive power and high capacity of deep neural nets (DNNs) result in accurate modeling and promising generalization if provided with sufficient data and clean(correct) labels. However, recent studies show that the training process is fragile and can easily overfit to noisy labels (Zhang et al., 2017), which commonly appear in real-world data since precise annotation is not always available or affordable. Hence, it is important to study the training dynamics affected by imperfect labels and develop robust learning strategies that ideally eliminate the negative impact of noisy labels while fully exploiting the information from all the available data.

Numerous approaches have been developed to address this challenge from various perspectives, e.g., loss correction (Xiao et al., 2015; Vahdat, 2017; Lee et al., 2018; Veit et al., 2017; Li et al., 2017b), robust loss functions (Ghosh et al., 2017; Zhang & Sabuncu, 2018; Wang et al., 2019; Ma et al., 2020) with provable noise tolerance, sample re-weighting (Patrini et al., 2017), curriculum learning(Kumar et al., 2010; Jiang et al., 2018; Guo et al., 2018), model co-teaching (Han et al., 2018), etc. A principal methodology behind a vast majority of works is to detect clean labels while discard/down-weigh the wrong ones, so the model mainly learns from correct labels. A broadly applied criterion is to select the samples with small losses and treat them as clean data. It is inspired by empirical observations that DNNs tend to learn simple patterns first before memorizing the noisy labels (Zhang et al., 2017; Arpit et al., 2017). Several curriculum learning methods utilize this criterion (Kumar et al., 2010; Jiang et al., 2014), and in each step, select/up-weigh samples with small losses. Robust loss functions also suppress the large losses associated with the possibly wrong labels. More recent approaches use mixture models (Arazo et al., 2019) to estimate the distribution of losses for clean and noisy data.

However, the instantaneous loss of an individual sample (i.e., the loss of the sample at one training step) is an unstable signal that can rapidly fluctuate due to DNN training's randomness. The error generated by such an unstable metric accumulates when the selected samples are used to train the model producing the losses. Co-teaching methods alleviate this problem by training two DNNs and using the loss computed on one model to guild the other. Also, as the model changes during training, each sample's loss needs to be re-evaluated even when it is not selected, which requires extra inference cost. MentorNet (Jiang et al., 2018) and Data Parameters (Saxena et al., 2019)

train an extra model to produce the sample weights or selection results without computing the loss. Furthermore, it may not be efficient to repeatedly train the model only on clean data that consistently have small losses, since the model have already learned, well memorized or overfitted to them.

A primary drawback of training only on clean data detected is that discarding the data with wrong labels removes potentially useful information about the data distribution. Hence, there have been growing interests in leveraging noisy data. Loss correction methods aim to correct the predicted class probabilities based on an estimated noise transition matrix of mislabeling probability between classes. Some other methods seek to relabel them by using the model itself (e.g., bootstrapping loss (Reed et al., 2014)) or another model/mechanism (e.g., directed graphical models, conditional random fields, or CNNs) trained on an additional set of clean data, which, however, is not always available. Self-training and unsupervised learning techniques have also been employed to generate pseudo labels to replace noisy labels. The pseudo labels are optimized together with the model or generated by the model with data augmentations to encourage the output consistency on the same sample's augmentations. Unfortunately, the pseudo labels' quality may vary across different samples and significantly degenerate when the noise ratio is high, or the model fails to produce stable and correct predictions. In such a case, relabeling all the noisy data for training may introduce extra noises or errors.

In this paper, we address the aforementioned problems of noise-label learning by developing a curriculum learning strategy called *Robust Curriculum Learning* (**RoCL**) that smoothly transits between two phases: (1) detection and supervised training on clean data; and (2) relabeling and self-supervision on noisy data. Specifically, we apply multiple training episodes, each starting from phase(1) and gradually moving to phase(2). Contrary to existing approaches, we only select samples with accurate given/pseudo labels that are most informative to the current model-in-training. Our data selection criterion takes both the dynamics of per-sample loss and output consistency (across multiple data augmentations) into account. Using an exponential moving average of the loss and consistency over training history, it overcomes the instability of instantaneous losses and does not incur any additional inference cost. In addition, by adjusting a temperature parameter, the criterion can interpolate between the two phases and keep the training focus on the data that the model mostly needs to improve on, e.g., clean data with unsatisfying output consistency or wrongly-labeled data with accurate pseudo labels. Thus, we can fully exploit both clean and noisy data more efficiently with less risk of introducing extra noise or error accumulation. We further show that our data selection can be derived from a novel optimization formulation for robust curriculum learning. We evaluate our method on multiple noisy learning benchmarks and show that our method outperforms a diverse set of recent noisy-label learning approaches.

1.1 Related Works

Early curriculum learning (CL) (Khan et al., 2011; Basu & Christensen, 2013; Spitkovsky et al., 2009) proposes the idea of generating an optimized sequence of training samples (i.e., a curriculum) to improve the models' performance, where the curriculum can be designed by human experts. Self-paced learning (SPL)(Kumar et al., 2010; Tang et al., 2012a; Supancic III & Ramanan, 2013; Tang et al., 2012b) selects samples with preference to smaller losses. It starts with selecting a small subset of samples with small losses, and gradually increases the selection size to cover all the training data. Self-paced curriculum learning (Jiang et al., 2015) combines the human expert in CL and loss-adaptation in SPL. SPL with diversity (SPLD) (Jiang et al., 2014) applies a negative group sparse regularization term to SPL to promote the diversity of selected samples. Minimax curriculum learning Zhou & Bilmes (2018) promotes the diversity of samples during early learning to encourage exploration and focus on hard samples in later stages.

In the context of robust learning with noisy labels, label correction methods aim to identify the wrong labels and possibly correct them to get more consistent labels for training. Previous works often apply an extra noise model (directed graphical model (Xiao et al., 2015), conditional random fields (Vahdat, 2017), neural networks (Lee et al., 2018; Veit et al., 2017) and knowledge graphs (Li et al., 2017b)) to correct the noisy labels, which often require extra clean data and as well as training/inference of the extra noise model. Another line of research is loss correction, which modifies the loss or prediction probabilities during training to correct the misinformation from the noisy labels. Patrini et al. (2017) uses two noise transition matrix (backward and forward) to correct the prediction probabilities. Label Smoothing Regularization (Szegedy et al., 2016; Pereyra et al., 2017) alleviates overfitting to noisy labels by using soft labels instead of one-hot labels. Reed et al. (2014) augments the loss with a notion of perceptual consistency. Jiang et al. (2018) trains a mentor network to re-weight samples during the training of a student network. Guo et al. (2018) designs a curriculum by measuring the

complexity of data using its distribution density in a feature space, and rank the complexity in an unsupervised manner. Ren et al. (2018) proposes a meta-learning algorithm that learns to assign weights to samples based on their gradient directions in training compared to those on validation data, which assumes the availability of clean data. Co-teaching (Han et al., 2018) feeds in the network with the most confident samples of another network for reducing the confirmation bias.

Some approaches focus on designing loss functions that have robust behaviors and provable tolerance to label noise. Ghosh et al. (2017) theoretically proves that the Mean Absolute Error(MAE) is a robust loss. The Generalized Cross Entropy (Zhang & Sabuncu, 2018) uses a negative Box-Cox transformation to obtain a loss function that generalizes MAE and Cross Entropy loss. Wang et al. (2019) proposes a Symmetric Cross Entropy that combines Cross Entropy loss and Reverse Cross Entropy loss. Ma et al. (2020) proposes a loss normalization method and shows that any loss can be made robust to noisy labels.

RoCL shares similar spirits with some CL methods that RoCL starts with learning easy and clean samples and gradually moves to hard and noisy ones. RoCL is more related to the loss correction approach in noisy-label learning literature as RoCL generates a curriculum dynamically assigning weight(probability) to each sample. RoCL differs from existing methods in: (1) it only selects a subset of informative and reliable labels for training in each step; (2) it creates a smooth transition not only from clean data to noisy data but also from supervised learning to self-supervision; (3) it runs multiple episodes of the curriculum to avoid getting in a local minimum dominated by a small set of clean/noisy data or a specific type of loss; (4) it does not assume the availability of an extra set of clean data; (5) it does not require extra computation or any modification to the model.

2 Dynamic Patterns of Clean/Noisy Labels in Training

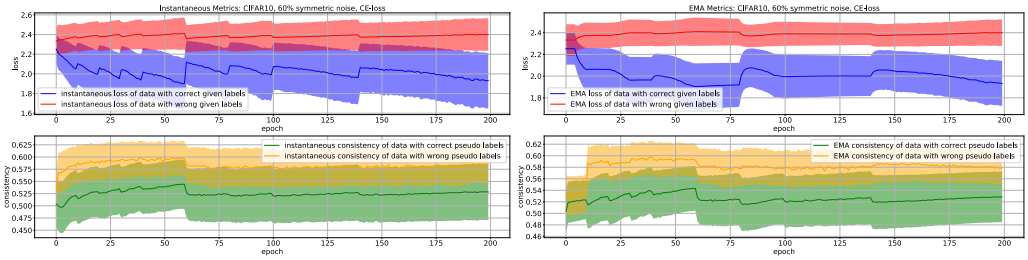

Figure 1: Dynamic patterns of instantaneous metrics (LEFT) and exponential moving average (EMA) metrics (RIGHT) when applying Baseline (Algorithm 2 in Appendix) that alternates for multiple episodes between supervised learning using given labels and self-supervised learning using pseudo labels. We use cross entropy for the loss (TOP) and 0-1 loss for the consistency (BOTTOM) averaged over $m = 7$ random augmentations.

2.1 Loss Dynamics and Clean Label Detection

A key challenge for most noise-label learning methods is to design a reliable criterion to select/re-weigh clean data and distinguish them from the noisy ones, so all the clean data can be fully exploited while most noisy labels are filtered out of the training process. Instantaneous losses have been widely used for this purpose according to the observation that clean data usually incur smaller losses than the noisy ones. One important reason is that the clean labels are consistent with each other in producing gradient updates, and therefore, the model can fit them better and faster. On the contrary, the noisy labels may contain inconsistent information with others, so the tug of war among them takes longer. For example, it can be hard for the model to find consistent visual patterns from images with noisy labels to make the desired predictions. However, this criterion of selecting samples with small instantaneous losses may lose its accuracy under high noise ratios and the randomness of DNN training, e.g., random initialization, random data augmentation, etc. In such scenarios, the instantaneous loss for a sample can be unstable across training steps, as shown in the top-left plot of Figure 1. Moreover, evaluating the instantaneous losses in each step requires extra inference cost on samples not selected for training.

The dynamic patterns of losses over the course of training give us a new insight for better clean data detection even when the noise ratio is high. In particular, we posit that a sample's label is more likely to be correct if its losses persistently retain low values over training steps. Given a sample (x_i, y_i) with x_i being the features and y_i being the label, we describe its loss dynamics using a simple exponential moving average (EMA) of the instantaneous loss $\ell(f(x_i; \theta_t), y_i)$ (where $f(x_i; \theta_t)$ denotes the model output and θ_t is the model parameters at step t) along the training history, which

is defined and computed recursively as

$$l_{t+1}(i) = \begin{cases} \gamma \times \ell(f(x_i;\theta_t), y_i) + (1-\gamma) \times l_t(i) & \text{if } i \in S_t \\ l_t(i) & \text{else ,} \end{cases} \tag{1}$$

Where $\gamma \in [0, 1]$ is a discounting factor, V is the set of all n training samples, and $S_t \subseteq V$ is the set of samples used for training at step t, e.g., a subset selected by certain curriculum learning method. Not that we only update the EMA losses for selected samples, so it only depends on a by-product of training and does not require extra inference.

In the top row of Figure 1, we show how the losses and EMA losses associated with clean/noisy data change throughout the training process. Specifically, we train a ResNet34 (He et al., 2016) model on CIFAR10 with 60% of the original labels randomly changed to a wrong class. To avoid quick overfitting to the noise, we train the model for multiple episodes (each composed of several training epochs of all data with a cosine annealing learning rate) and alternate between the supervised learning episode that minimizes the cross-entropy loss against the given noisy labels and the self-supervision episode that minimizes the consistency loss (ref. Eq. (2)). Comparing the dynamics of instantaneous loss and EMA loss in the figure, we see that the variance for each group of samples decreases as expected with EMA. Also, the confusing region of clean and noisy labels is reduced, demonstrating the effectiveness of using dynamic patterns of losses for clean data detection.

2.2 Consistency Dynamics and Pseudo Label Selection

Simply removing noisy data from training discards important information about data distribution. Including all the noisy data for self-supervision, bootstrapping or relabeling can also harm the training process since pseudo labels' quality is not equally high for different samples. The pseudo labels also much depend on the model generating them, where the model's predictions may contain noises or biases that can accumulate if adopted for training. Hence, a careful selection of noisy data is necessary. However, without access to a known set of clean data or a reliable pre-trained model, it is nontrivial to evaluate pseudo labels' correctness. By empirically analyzing the training dynamics of model outputs observed in the above experiment, we discover that the model output for a sample tends to be an accurate pseudo label if the pseudo labels are consistent over steps and across augmentations.

We first define the instantaneous consistency of a sample x_i at step t as the discrepancy of the model output $f(x_i;\theta)$ between step t and $t-1$ on x_i and its m data augmentations $\{x_i^{(j)}\}_{j=1}^m$, where the discrepancy can be measured by any loss function $\ell(\cdot,\cdot)$. However, the discrepancy can be trivially small if the models of step t and $t-1$ are too similar and make similar errors. Therefore, we follow the mean teacher (Tarvainen & Valpola, 2017) method and instead use an exponential moving average of the model parameters to compute the model prediction at step $t-1$: $\overline{f}_t(x_i) \triangleq f(x_i;\overline{\theta}_t),\ \ \overline{\theta}_t \triangleq \gamma\theta_{t-1} + (1-\gamma)\overline{\theta}_{t-1}$.

The pseudo label is usually constructed by averaging $\overline{f}_t(x_i)$ over multiple data augmentations, i.e., $1/m \sum_{j=1}^m \overline{f}_t(x_i^{(j)})$. The instantaneous consistency $\zeta_t(i)$ of sample x_i at step t is then defined as

$$\zeta_t(i) \triangleq \frac{1}{m}\sum_{j=1}^{m} \ell(f(x_i;\theta_t), \overline{f}_t(x_i^{(j)})), \tag{2}$$

Which can also be minimized as a consistency loss for x_i in self-supervised learning when no label is given. Similar to the EMA loss in Eq. (1), we define the EMA consistency over training history as

$$c_{t+1}(i) = \begin{cases} \gamma \times \zeta_t(i) + (1-\gamma) \times c_t(i) & \text{if } i \in S_t \\ c_t(i) & \text{else} \end{cases} \tag{3}$$

The EMA consistency $c_t(i)$ measures both the time-consistency over multiple steps and spatial consistency over different augmentations for the model output of sample x_i. If the output prediction is wrong and contradicts other samples' labels, these two types of consistency can hardly hold since the output can easily change or get flipped by training on different random mini-batches. In the bottom-right plot of Figure 1, we report the mean and standard deviation of the EMA consistency for two groups of data at each epoch, i.e., the ones with correct pseudo labels and the ones with incorrect pseudo labels. Comparing to the instantaneous consistency in the bottom-left plot, EMA consistency provides a more reliable criterion for allocating the correct pseudo labels. Thus, we can safely exploit the noisy data by learning using their pseudo labels with a small risk of introducing harmful noises or biases.

3 ROBUST CURRICULUM LEARNING

In this section, we present the RoCL method. We first introduce the selection criterion covering both the clean label detection and pseudo label selection. By adjusting two temperature parameters τ_1 and τ_2, it can smoothly interpolate between the two criteria and control their trade-offs. We then show that the criterion is derived from a novel optimization formulation for robust curriculum learning. We show the RoCL algorithm using our data selection criterion and propose several practical modifications.

3.1 DATA SELECTION CRITERION

To obtain the capability of both clean label detection and pseudo label selection, we combine the two criteria from the previous sections and use a temperature parameter to control the preference of small/large losses or consistency and their trade-offs. We sample x_i at step t with probability:

$$\mathcal{P}_t(i) = \lambda \times p_t(i) + (1 - \lambda) \times q_t(i), \tag{4}$$

where $p_t(i)$ and $q_t(i)$ are defined as softmax probabilities, i.e.,

$$p_t(i) \triangleq \frac{\exp[\tau_1 l_t(i)]}{\sum_{j=1}^n \exp[\tau_1 l_t(j)]}, \;\; q_t(i) \triangleq \frac{\exp[\tau_2 c_t(i)]}{\sum_{j=1}^n \exp[\tau_2 c_t(j)]}, \tag{5}$$

and $\lambda \in [0, 1]$ controls trade-off between them. Here, $p_t(i)$ is a softmax probability computed from EMA losses and serves for clean data detection: samples with smaller(larger) EMA losses and thus clean(noisy) labels tend to have high probability $p_t(i)$ when τ_1 is negative(positive). Similarly, $q_t(i)$ is computed from EMA consistency and serves for pseudo label selection: samples with smaller(larger) EMA consistency and thus correct(wrong) pseudo labels tend to have high probability $q_t(i)$ when τ_2 is negative(positive). The probabilities can be seen to approximate the max(min)-operators when $\tau_1, \tau_2 \to +\infty$ ($\tau_1, \tau_2 \to -\infty$) and reduce to uniform sampling when $\tau_1, \tau_2 = 0$.

Thereby, either $p_t(i)$ or $q_t(i)$ can be independently employed to select/re-weigh samples for noise-label learning. However, selecting data with high probabilities $p_t(i)$ when $\tau_1 \ll 0$ tends to make the training focus on clean data the model has already learned and memorized, which carry limited new information and prevents exploration of data near the margin, thus deterring the learning progress. Similarly, selecting data with high probabilities $q_t(i)$ when $\tau_2 \ll 0$ is not informative since the model outputs are consistently correct for those data, and little progress can be made.

We can encourage exploration by manipulating the temperature parameters. When moving τ_1 and τ_2 towards zero, we obtain a more uniform exploration of all data. Nonetheless, a more effective strategy is coupling the options of τ_1 and τ_2. For example, a negative τ_1 associated with a positive τ_2 lift the preference over clean data that have not been fully exploited and memorized by the model. Alternatively, a positive τ_1 with a negative τ_2 emphasizes those noisy data whose pseudo labels tend to be correct, so relabeling them provides new information in addition to the clean data.

We apply the data selection criterion of Eq. (4) in each step and gradually change its parameters $(\tau_1, \tau_2, \lambda)$ over the course of training according to the properties discussed above. We start from a negative τ_1 associated with a positive τ_2 and a large λ ($p_t(i)$ dominates), then gradually increase τ_1 while decrease τ_2 and λ, and end with a positive τ_1, a negative τ_2 and a small λ ($q_t(i)$ dominates). In this way, we get a curriculum with a smooth transition between supervised learning of clean data (using correct given labels) and self-supervision of noisy data (using reliable pseudo labels). Moreover, the coupling strategy on τ_1 and τ_2 encourages to select informative samples that the model mostly needs to improve on, i.e., clean data with inconsistent model outputs or noisy data with correct pseudo labels. In practice, we can further reduce the parameters: (1) given the sequence for τ_1 in the curriculum as $\tau_{1:T}$, we can reverse it as the sequence for τ_2, i.e., $\tau_{T:1}$; (2) we can make λ monotone increase with τ_1, e.g., setting the initial value $\lambda_1 = a\tau_1 + b$ and ending value $\lambda_T = a\tau_T + b$, solving this linear system for a and b, which generate $\lambda_{1:T} = a\tau_{1:T} + b$.

3.2 ROBUST CURRICULUM LEARNING AS AN OPTIMIZATION

In fact, the data selection criterion is derived from an optimization formulation of our robust curriculum learning (RoCL), in which we aim to minimize a combination of supervised loss and consistency(self-supervised) loss in the following form.

$$\min_\theta F(\theta) \triangleq \frac{\lambda}{\tau_1} \log \left(\frac{1}{n} \sum_{i=1}^n \exp[\tau_1 \ell(f(x_i; \theta), y_i)] \right) + \frac{1-\lambda}{\tau_2} \log \left(\frac{1}{n} \sum_{i=1}^n \exp[\tau_2 \zeta(i)] \right), \tag{6}$$

where the consistency loss $\zeta(i)$ is defined in Eq. 2 (with t removed here). Unlike the conventional empirical risk minimization (ERM) with arithmetic average loss, e.g., $1/n \sum_{i=1}^{n} \ell(f(x_i;\theta), y_i)$ and $1/n \sum_{i=1}^{n} \zeta(i)$, we use LogSumExp loss with an additional temperature parameter, so it can approximate both the min-loss (when $\tau \to -\infty$), max-loss (when $\tau \to +\infty$), and any interpolation between them, where the min-loss focuses on the the easiest samples with the smallest losses and the max-loss focuses on the hardest ones. Note it reduces to the arithmetic average loss when $\tau \to 0$. It is called "tilted loss" in a recent work (Li et al., 2020), which shows several intriguing properties of it in different learning settings. We mainly use the properties of its gradient to derive the data selection criterion given in Eq. (4). By simple algebra, the gradient of $F(\theta)$ at step t is

$$\begin{aligned}\nabla_\theta F(\theta_t) &= \lambda \sum_{i=1}^{n} p_t(i) \nabla_\theta \ell(f(x_i;\theta_t), y_i) + (1-\lambda) \sum_{i=1}^{n} q_t(i) \nabla_\theta \zeta_t(i) \\ &= \sum_{i=1}^{n} \mathcal{P}_t(i) \left[\frac{\lambda p_t(i)}{\mathcal{P}_t(i)} \nabla_\theta \ell(f(x_i;\theta_t), y_i) + \frac{(1-\lambda) q_t(i)}{\mathcal{P}_t(i)} \nabla_\theta \zeta_t(i) \right] = \mathbb{E}_{i \sim \mathcal{P}_t(i)} G_t(i), \quad (7)\end{aligned}$$

where

$$G_t(i) \triangleq \frac{\lambda p_t(i)}{\mathcal{P}_t(i)} \nabla_\theta \ell(f(x_i;\theta_t), y_i) + \frac{(1-\lambda) q_t(i)}{\mathcal{P}_t(i)} \nabla_\theta \zeta_t(i). \quad (8)$$

Here we have an abuse of notation for $p_t(i)$ and $q_t(i)$, in which the EMA loss $l_t(i)$ and EMA consistency $c_t(i)$ in their definition of Eq. (5) are replaced by their instantaneous counterparts $\ell(f(x_i;\theta_t), y_i)$ and $\zeta_t(i)$, respectively. Note in Section 2, we already discussed that the EMA metrics are better alternatives to the instantaneous metrics when used for data selection in noisy-label learning. An unbiased estimator of the gradient in Eq. (7) can be achieved by drawing a subset of samples according to $\mathcal{P}_t(i)$ and averaging their gradients $G_t(i)$ in Eq. (8).

3.3 Algorithm of Robust Curriculum Learning

We now elaborate on our curriculum learning algorithm, RoCL, based on the data selection criterion in Section 3.1 and the optimization formulation in Section 3.2. We apply a warm starting episode of a few epochs (e.g., 5-10) over all the data and given labels with label smoothing to obtain a stable EMA metrics. After that, we apply multiple episodes of curriculum learning, each including a sequence of steps following the curriculum at the end of Section 3.1 for data selection per step. We repeat the transition between clean data learning to noisy data learning for multiple episodes to avoid getting trapped in a local minimum dominated by a small set of clean/noisy data or a specific type of loss. It also reinforces the memorization of clean labels and correct pseudo labels learned in previous episodes. Moreover, under the coupling strategy of τ_1 and τ_2, each episode is encouraged to explore the clean/noisy data that the previous episode fails to learn. Considering the undertrained model (producing inaccurate pseudo labels) and the relatively high variance of the EMA metrics at the earlier episodes, we further propose to starts from a small budget for the selected subset size and gradually increase in later episodes.

Algorithm 1 Robust Curriculum Learning (RoCL)

input: $\{(x_i, y_i)\}_{i=1}^{n}$, $\pi(\cdot;\eta)$, $\ell(\cdot,\cdot)$, $f(\cdot;\theta)$, $T_{0:K}$; $\tau_1 < 0, \tau_T > 0$; $\lambda_1, \lambda_T > 0 \in [0,1]$; $\gamma, \gamma_b \in [0,1]$
initialize: θ_0, $b_0 \in (0,n)$, $l_0(i) = c_0(i) = 0 \; \forall i \in [n]$
for $k \in \{0, \cdots, K\}$ **do**
  Schedule $\tau_{1:T_k}$ and $\lambda_{1:T_k}$ by Eq. (9)-(10);
  for $t' \in \{1, \cdots, T_k\}$ **do**
    $t \leftarrow t' + T_{k-1}$;
    if $k = 0$ **then**
      $S_t \leftarrow [n]$;
      $\theta_t \leftarrow \theta_{t-1} + \pi\left(\nabla_\theta \frac{1}{n} \sum_{i=0}^{n} \ell_t(i);\eta\right)$;
    else
      Draw a subset $S_t \subseteq [n]$ of b_k samples according to probability $\mathcal{P}_t$ in Eq. (4) with $\tau_1 = \tau_{t'}, \tau_2 = \tau_{T_k - t'}, \lambda = \lambda_{t'}$;
      $\theta_t \leftarrow \theta_{t-1} + \pi\left(\frac{1}{b_k} \sum_{i \in S_t} G_t(i);\eta\right)$ (Eq. (8));
    end if
    Update $l_{t+1}(i)$ and $c_{t+1}(i)$ by Eq. (1) and Eq. (3);
  end for
  $b_{k+1} \leftarrow (1+\gamma_b) \times b_k$;
end for

To generate the whole schedule of $\tau_{1:T}$ in each episode, we can apply any monotone interpolation between τ_1 and τ_T whose values are predefined. In our curriculum for each episode, we need to keep a high quality of the selected clean(pseudo) labels in earlier(later) stages and make the exploration stages in between shorter since their selected labels contain more noise. Therefore, we choose "s"-shaped functions such as $\tanh$ or the logistic function for the interpolation. When using $\tanh$, for

example, we can pick a value σ close to 1 (e.g., $\sigma = 0.95$) and define

$$\forall t \in [T], \tau_t = \frac{\tau_T - \tau_m}{\sigma} \times \tanh\left(\tanh^{-1}(-\sigma) + \frac{2t}{T}\tanh^{-1}(\sigma)\right) + \tau_m, \tau_m = \frac{\tau_1 + \tau_T}{2}. \quad (9)$$

The corresponding schedule for λ can then be defined as an affine transformation of $\tau_{1:T}$:

$$\forall t \in [T], \lambda_t = a_\lambda(\tau_t - \tau_1) + \lambda_1, a_\lambda = \frac{\lambda_T - \lambda_1}{\tau_T - \tau_1}. \quad (10)$$

We list the detailed algorithmic procedures of RoCL in Algorithm 1. We denote the update of θ_t produced by the adopted optimizer as $\pi(\nabla_\theta F(\theta_t), \eta)$, where η contains all hyperparameters of the optimizer at step t, e.g., the learning rate. We denote $\ell(f(x_i; \theta_{t-1}), y_i)$ by a brief notation $\ell_t(i)$.

3.4 Practical Modifications

We apply the following techniques broadly studied and adopted in recent robust learning methods.

- We apply the **class-balance regularization** used in (Tanaka et al., 2018), which prevents the model from predicting the same class for all the samples within a mini-batch B. We add $(1/B)\sum_{i \in B} \ell(f(x_i; \theta), 1/C \cdot \mathbf{1})$ (where C is the number of classes) to the objective for a mini-batch B with regularization weight of 1.
- We apply **label smoothing** whose effectiveness in noisy-label learning has been studied in (Lukasik et al., 2020). We modify each one-hot label y_i to be $\bar{y}_i \leftarrow (1-\alpha)y_i + \alpha/C$ (e.g., we use $\alpha = 0.5$).
- We apply **Mix-Up** (Zhang et al., 2018) to all the selected data. Mix-Up of (1) two clean labels and (2) one clean label + one pseudo label are both helpful: the former encourages smooth model outputs between the two samples; the latter can improve the pseudo label's effectivenss by learning its interpolation with a clean label and also reduces the risk when the pseudo label is incorrect. However, Mix-Up of two (soft) pseudo labels can significantly increase the entropy of the mixed label if both pseudo labels are under-confident (or even incorrect). We apply a curriculum to the beta distribution's parameter of Mix-Up and gradually reduce it from large to small (e.g., from 8.0 to 0.2 in our experiments) within each episode.

4 Experiments

We evaluate RoCL with other state-of-the-art approaches for noisy-label learning on three widely used benchmarks, i.e., CIFAR10/100 with two types of synthetic noises (i.e., symmetric and asymmetric), and WebVision (Li et al., 2017a) (the first 50 classes) containing unknown noises from web labels. Symmetric noise is generated by flipping each label randomly to an incorrect class with probability ρ (i.e., noise rate), and our experiments cover $\rho = \{0.4, 0.6, 0.8\}$. For asymmetric noise, we flip the labels within a specific set of classes. For CIFAR10, flipping TRUCK→ AUTOMOBILE, BIRD→AIRPLANE, DEER→HORSE, CAT→DOG. In CIFAR100, the 100 classes are grouped into 20 super-classes with each has 5 sub-classes, we then flip each class within the same super-class into the next in a circular fashion with probability ρ. Our experiments cover $\rho = \{0.2, 0.3, 0.4\}$.

Table 1: Accuracy (%) evaluated on WebVision and ILSVRC2012 validation sets for DNNs trained by noisy-label learning methods on WebVision training set (first 50 classes), which contains **real-world web-label noises**.

Val. set	WebVision		ILSVRC2012	
Accuracy	Top-1	Top-5	Top-1	Top-5
F-correct	61.12	82.68	57.36	82.36
Decoupling	62.54	84.74	58.26	82.26
Co-teaching	63.58	85.20	61.48	84.70
MentorNet	63.00	81.40	57.80	79.92
MentorMix	76.00	90.20	72.90	91.10
D2L	62.68	84.00	57.80	81.36
INCV	65.24	85.34	61.60	84.98
RoCL (ours)	**78.80**	**92.52**	**75.72**	**92.20**

We apply RoCL to train ResNet34 on CIFAR10/100 and ResNet50 on WebVision, which are the most widely used models in other baseline papers. We apply SGD with momentum of 0.9, weight decay of 10^{-4} and cosine annealing learning rate in each training episode. The initial learning rate is set to 0.1 for CIFAR10/100 and 1.0 for WebVision. In all RoCL experiments, we apply $T_0 = 10$ warm starting epochs followed by $K = 10$ episodes of curriculum learning, whose lengths start from $T_1 = 10$ and increase by 10 for every episode afterwards. We initialize the subset size $b_0 = 0.2n$ and set factors $\gamma = \gamma_b = 0.1$. We set $\tau_1 = -4$ and $\tau_T = 1$ in order to rely more on the correct given/pseudo labels (when $\tau < 0$) and suffer less from the risk of exploring more uncertain/noisy labels (when $\tau > 0$). We set $\lambda_1 = 0.9$ and $\lambda_T = 0.3$ so each episode starts from supervised learning and gradually move to self-supervision. We use Cubuk et al. (2020) for data augmentations. We compare RoCL to the following baselines: F-correct (Patrini et al., 2017), Decoupling (Malach & Shalev-Shwartz, 2017),

Co-teaching (Han et al., 2018), D2L (Ma et al., 2018), INCV (Chen et al., 2019), MD-DYR-SH (Arazo et al., 2019), MentorNet (Jiang et al., 2018), MentorMix (Jiang et al., 2020), O2U-net (Huang et al., 2019), RoG+D2L (Lee et al., 2019), PENCIL (Yi & Wu, 2019), GCE (Zhang & Sabuncu, 2018), SCE (Wang et al., 2019), NFL/NCE variants (Ma et al., 2020), and Bootstrap (Reed et al., 2014).

We report the results and comparisons to baselines in three tables: real-world noise in Table. 1, symmetric noise in Table. 2 and asymmetric noise in Table. 3. RoCL achieves the best performance in every setting, and for most of the cases, improves upon the existing methods by large margins. The closest rival to RoCL is MentorMix, which utilizes MentorNet and Mix-Up to assign weights to each sample. We note that MentorMix requires training of an extra mentor network to generate the sample weights, while RoCL is more flexible and only makes changes to the training process without modifying the model.

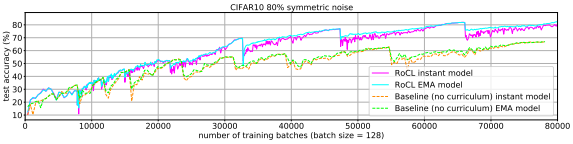

Figure 2: RoCL (Algorithm 1) vs. Baseline without any curriculum (Algorithm 2 in Appendix) during the training of ResNet34 on CIFAR10 with 80% symmetric noises.

Table 2: Test accuracy (%) of noisy-label learning methods on CIFAR10/100 corrupted by **symmetric(uniform)** label noises of different levels. All the baselines' results are from the original papers or the following-up works. There are two formats of these reported results: "mean±variance" of 5 trials and single-trial accuracy.

Dataset	CIFAR10			CIFAR100		
Noise Rate	40%	60%	80%	40%	60%	80%
MD-DYR-SH	92.3	86.1	74.1	70.1	59.5	39.5
MentorNet	91.2	74.2	60.0	68.5	61.2	35.5
MentorMix	94.2	91.3	81.0	71.3	64.6	41.2
O2U-net	90.3	-	43.4	69.2	-	39.4
RoG+D2L	87.0	78.0	-	64.9	40.6	-
PENCIL	-	-	-	69.12 ± 0.62	57.79 ± 3.86	fail
GCE	87.62 ± 0.26	82.70 ± 0.23	67.92 ± 0.60	62.64 ± 0.33	54.04 ± 0.56	29.60 ± 0.51
SCE	85.34 ± 0.07	80.07 ± 0.02	53.81 ± 0.27	53.69 ± 0.07	41.47 ± 0.04	15.00 ± 0.04
NFL+MAE	83.81 ± 0.06	76.36 ± 0.31	45.23 ± 0.52	58.18 ± 0.08	46.10 ± 0.50	24.78 ± 0.82
NCE+RCE	86.02 ± 0.09	79.78 ± 0.50	52.71 ± 1.90	59.48 ± 0.56	47.12 ± 0.62	25.80 ± 1.12
RoCL (ours)	$\mathbf{94.55 \pm 0.12}$	$\mathbf{92.06 \pm 0.23}$	$\mathbf{85.76 \pm 0.26}$	$\mathbf{74.64 \pm 0.43}$	$\mathbf{66.79 \pm 0.58}$	$\mathbf{47.24 \pm 0.75}$

Table 3: Test accuracy (%) of noisy-label learning methods on CIFAR10/100 corrupted by **asymmetric(class-dependent)** noises of 3 levels. All the baselines' results are from the original papers or the following-up works.

Dataset	CIFAR10			CIFAR100		
Noise Rate	20%	30%	40%	20%	30%	40%
PENCIL	92.43	91.84	91.01	74.70 ± 0.56	72.52 ± 0.38	63.61 ± 0.23
Bootstrap	86.57 ± 0.08	84.86 ± 0.05	79.76 ± 0.07	63.44 ± 0.35	63.18 ± 0.35	62.08 ± 0.22
F-correct	89.09 ± 0.47	86.79 ± 0.36	83.55 ± 0.58	42.46 ± 2.16	38.13 ± 2.97	34.44 ± 1.93
GCE	86.07 ± 0.31	80.78 ± 0.21	74.98 ± 0.32	59.99 ± 0.83	53.99 ± 0.29	41.49 ± 0.79
SCE	83.92 ± 0.07	79.70 ± 0.27	78.20 ± 0.03	58.22 ± 0.47	49.85 ± 0.91	42.19 ± 0.19
NFL+MAE	86.81 ± 0.32	83.91 ± 0.34	77.16 ± 0.10	63.10 ± 0.22	56.19 ± 0.61	43.51 ± 0.42
NCE+RCE	88.56 ± 0.17	85.58 ± 0.44	79.59 ± 0.40	62.68 ± 0.79	57.82 ± 0.41	46.79 ± 0.96
RoCL (ours)	$\mathbf{95.38 \pm 0.21}$	$\mathbf{94.19 \pm 0.28}$	$\mathbf{92.31 \pm 0.35}$	$\mathbf{80.03 \pm 0.34}$	$\mathbf{77.59 \pm 0.45}$	$\mathbf{73.28 \pm 0.83}$

Table. 4 reports RoCL's performance when applied with different loss functions on CIFAR10 under high noise rates, i.e., 60% and 80%. We observe significant improvements over their performance without using RoCL in Table. 2. It indicates that RoCL is compatible with any loss function and can further enhance their performance. In Figure 2, we compare RoCL with the Baseline algorithm used in Section 2, which also takes the advantage of switching between supervised learning with the given labels and self-supervision with the pseudo labels but does not have any curriculum to dynamically select data for each step. RoCL learns much faster than Baseline for the same amount of training batches, indicating that our curriculum is effective in selecting more informative data and labels. Note Baseline's final accuracy 67.03% already outperforms most methods in Table 2, which verifies the effectiveness of multi-episode training that alternates between the two phases.

Table 4: Test accuracy (%) of **RoCL** applied with different loss functions on CIFAR10 corrupted by $\{60\%, 80\%\}$ **symmetric(uniform)** noises (CE-cross entropy).

Noise Rate	60%	80%
CE	90.22 ± 0.24	77.47 ± 0.67
GCE	89.30 ± 0.68	79.84 ± 1.12
SCE	$\mathbf{92.06 \pm 0.23}$	74.25 ± 0.86
NFL+MAE	88.73 ± 0.47	$\mathbf{85.76 \pm 0.26}$
NFL+RCE	87.68 ± 0.35	80.09 ± 0.41
NCE+MAE	90.37 ± 0.43	82.16 ± 0.93
NCE+RCE	88.03 ± 0.39	80.33 ± 0.80

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. *arXiv preprint arXiv:1904.11238*, 2019.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242, 2017.

Sumit Basu and Janara Christensen. Teaching classification boundaries to humans. In *AAAI*, pp. 109–115, 2013.

Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. *arXiv preprint arXiv:1905.05040*, 2019.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. *arXiv preprint arXiv:1712.09482*, 2017.

Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150, 2018.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pp. 8527–8537, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3326–3334, 2019.

Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander G. Hauptmann. Self-paced learning with diversity. In *NeurIPS*, pp. 2078–2086, 2014.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. Self-paced curriculum learning. In *AAAI*, pp. 2694–2700, 2015.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313, 2018.

Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. ICML, 2020.

Faisal Khan, Xiaojin (Jerry) Zhu, and Bilge Mutlu. How do humans teach: On curriculum learning and teaching dimension. In *NeurIPS*, pp. 1449–1457, 2011.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *NeurIPS*, pp. 1189–1197, 2010.

Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. *arXiv preprint arXiv:1901.11300*, 2019.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456, 2018.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017a.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1910–1918, 2017b.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? *arXiv preprint arXiv:2003.02819*, 2020.

Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah M Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. *arXiv preprint arXiv:1806.02612*, 2018.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. *arXiv preprint arXiv:2006.13554*, 2020.

Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". In *Advances in Neural Information Processing Systems*, pp. 960–970, 2017.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018.

Shreyas Saxena, Oncel Tuzel, and Dennis DeCoste. Data parameters: A new family of parameters for learning a differentiable curriculum. In *Advances in Neural Information Processing Systems*, pp. 11095–11105, 2019.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. Baby Steps: How "Less is More" in unsupervised dependency parsing. In *NeurIPS 2009 Workshop on Grammar Induction, Representation of Language and Language Learning*, 2009.

James Steven Supancic III and Deva Ramanan. Self-paced learning for long-term tracking. In *CVPR*, pp. 2379–2386, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5552–5560, 2018.

Kevin Tang, Vignesh Ramanathan, Li Fei-fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In *NeurIPS*, pp. 638–646, 2012a.

Ye Tang, Yu-Bin Yang, and Yang Gao. Self-paced dictionary learning for image classification. In *MM*, pp. 833–836, 2012b.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 1195–1204. 2017.

Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Advances in Neural Information Processing Systems*, pp. 5596–5605, 2017.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 839–847, 2017.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 322–330, 2019.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.

Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7017–7025, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pp. 8778–8788, 2018.

Tianyi Zhou and Jeff Bilmes. Minimax curriculum learning: Machine teaching with desirable difficulties and scheduled diversity. In *ICLR*, 2018.

A APPENDIX

A.1 BASELINE (NO CURRICULUM) IN SECTION 2 AND FIGURE 2

Algorithm 2 Baseline (no curriculum)

input: $\{(x_i, y_i)\}_{i=1}^n$, $\pi(\cdot;\eta)$, $\ell(\cdot,\cdot)$, $f(\cdot;\theta)$, $T_{0:K}$; $\gamma \in [0,1]$
initialize: θ_0, $l_0(i) = c_0(i) = 0\ \forall i \in [n]$
for $k \in \{0, \cdots, K\}$ **do**
  for $t' \in \{1, \cdots, T_k\}$ **do**
    $t \leftarrow t' + T_{k-1}$;
    $S_t \leftarrow [n]$;
    if $k\%2 = 0$ **then**
      $\theta_t \leftarrow \theta_{t-1} + \pi\left(\nabla_\theta \frac{1}{n} \sum_{i=0}^n \ell_t(i);\eta\right)$; {supervised learning using given labels}
      Update $l_{t+1}(i)$ by Eq. (1); {update EMA loss}
    else
      $\theta_t \leftarrow \theta_{t-1} + \pi\left(\nabla_\theta \frac{1}{n} \sum_{i=0}^n \zeta_t(i);\eta\right)$; {self-supervised learning using pseudo labels}
      Update $c_{t+1}(i)$ by Eq. (3); {update EMA consistency}
    end if
  end for
end for

A.2 ADDITIONAL EXPERIMENTS

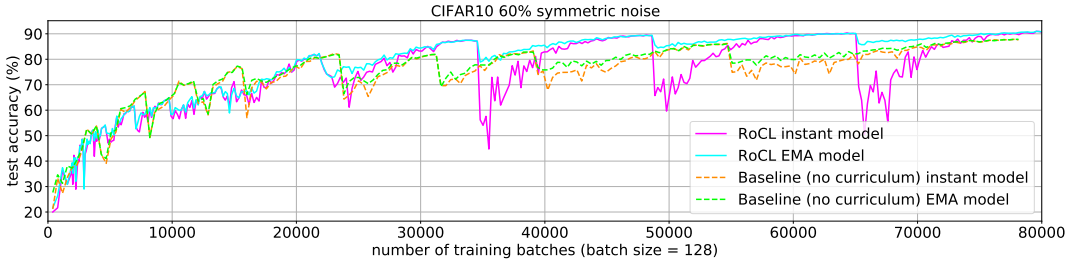

Figure 3: RoCL (Algorithm 1) vs. Baseline without any curriculum (Algorithm 2 in Appendix) during the training of ResNet34 on CIFAR10 containing **60% symmetric noises** on labels.

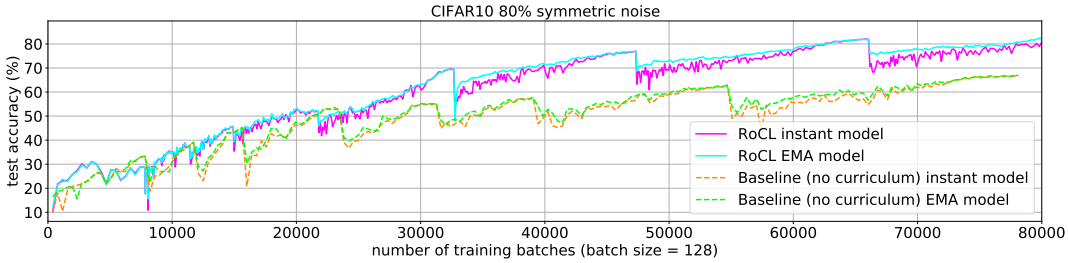

Figure 4: RoCL (Algorithm 1) vs. Baseline without any curriculum (Algorithm 2 in Appendix) during the training of ResNet34 on CIFAR10 containing **80% symmetric noises** on labels.

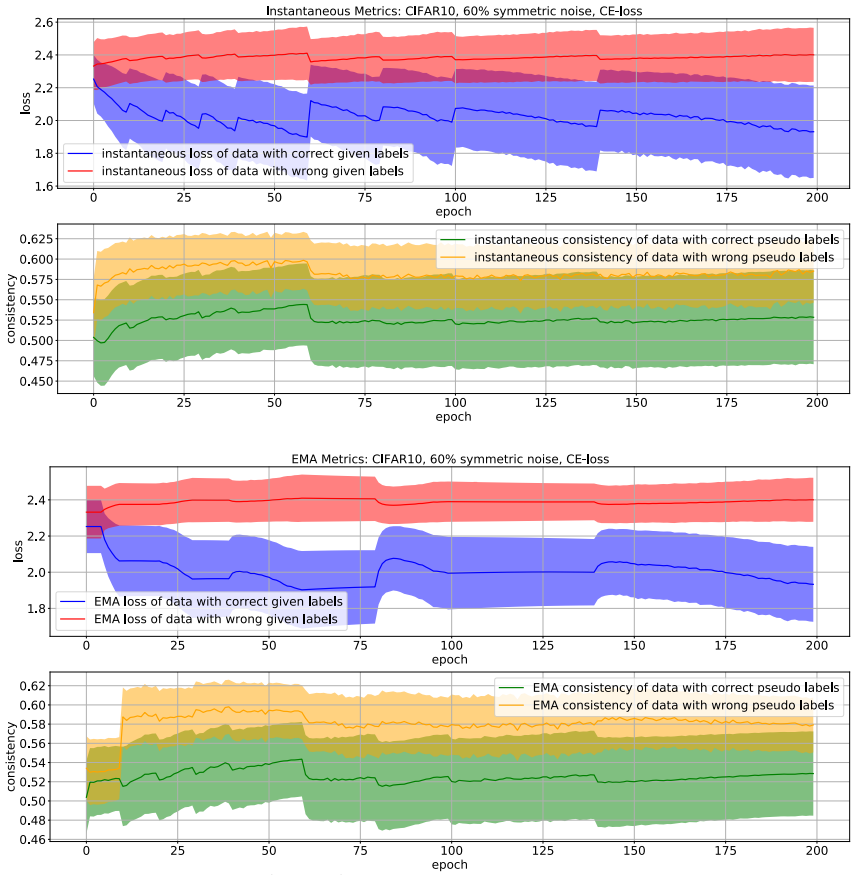

Figure 5: Large version of Figure 1.

Table 5: Extended version of Table 2 with two more baselines: NFL+RCE and NCE+MAE.

Dataset	CIFAR10			CIFAR100		
Noise Rate	40%	60%	80%	40%	60%	80%
MD-DYR-SH	92.3	86.1	74.1	70.1	59.5	39.5
MentorNet	91.2	74.2	60.0	68.5	61.2	35.5
MentorMix	94.2	91.3	81.0	71.3	64.6	41.2
O2U-net	90.3	-	43.4	69.2	-	39.4
RoG+D2L	87.0	78.0	-	64.9	40.6	-
PENCIL	-	-	-	69.12 ± 0.62	57.79 ± 3.86	fail
GCE	87.62 ± 0.26	82.70 ± 0.23	67.92 ± 0.60	62.64 ± 0.33	54.04 ± 0.56	29.60 ± 0.51
SCE	85.34 ± 0.07	80.07 ± 0.02	53.81 ± 0.27	53.69 ± 0.07	41.47 ± 0.04	15.00 ± 0.04
NFL+MAE	83.81 ± 0.06	76.36 ± 0.31	45.23 ± 0.52	58.18 ± 0.08	46.10 ± 0.50	24.78 ± 0.82
NFL+RCE	86.05 ± 0.12	79.78 ± 0.13	55.06 ± 1.08	58.20 ± 0.31	46.30 ± 0.45	25.16 ± 0.55
NCE+MAE	84.19 ± 0.43	77.61 ± 0.05	49.62 ± 0.72	59.22 ± 0.36	48.06 ± 0.34	25.50 ± 0.76
NCE+RCE	86.02 ± 0.09	79.78 ± 0.50	52.71 ± 1.90	59.48 ± 0.56	47.12 ± 0.62	25.80 ± 1.12
RoCL (ours)	$\mathbf{94.55 \pm 0.12}$	$\mathbf{92.06 \pm 0.23}$	$\mathbf{85.76 \pm 0.26}$	$\mathbf{74.64 \pm 0.43}$	$\mathbf{66.79 \pm 0.58}$	$\mathbf{47.24 \pm 0.75}$

Table 6: Extended version of Table 3 with two more baselines: NFL+RCE and NCE+MAE.

Dataset	CIFAR10			CIFAR100		
Noise Rate	20%	30%	40%	20%	30%	40%
PENCIL	92.43	91.84	91.01	74.70 ± 0.56	72.52 ± 0.38	63.61 ± 0.23
Bootstrap	86.57 ± 0.08	84.86 ± 0.05	79.76 ± 0.07	63.44 ± 0.35	63.18 ± 0.35	62.08 ± 0.22
F-correct	89.09 ± 0.47	86.79 ± 0.36	83.55 ± 0.58	42.46 ± 2.16	38.13 ± 2.97	34.44 ± 1.93
GCE	86.07 ± 0.31	80.78 ± 0.21	74.98 ± 0.32	59.99 ± 0.83	53.99 ± 0.29	41.49 ± 0.79
SCE	83.92 ± 0.07	79.70 ± 0.27	78.20 ± 0.03	58.22 ± 0.47	49.85 ± 0.91	42.19 ± 0.19
NFL+MAE	86.81 ± 0.32	83.91 ± 0.34	77.16 ± 0.10	63.10 ± 0.22	56.19 ± 0.61	43.51 ± 0.42
NFL+RCE	88.73 ± 0.29	85.74 ± 0.22	79.27 ± 0.43	63.12 ± 0.41	54.72 ± 0.38	42.97 ± 1.03
NCE+MAE	86.44 ± 0.23	83.98 ± 0.52	78.23 ± 0.42	62.38 ± 0.60	58.02 ± 0.48	47.22 ± 0.30
NCE+RCE	88.56 ± 0.17	85.58 ± 0.44	79.59 ± 0.40	62.68 ± 0.79	57.82 ± 0.41	46.79 ± 0.96
RoCL (ours)	$\mathbf{95.38 \pm 0.21}$	$\mathbf{94.19 \pm 0.28}$	$\mathbf{92.31 \pm 0.35}$	$\mathbf{80.03 \pm 0.34}$	$\mathbf{77.59 \pm 0.45}$	$\mathbf{73.28 \pm 0.83}$