O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks

Jinchi Huang
Alibaba Group
Hangzhou, China
jinchi.hjc@alibaba-inc.com

Lie Qu
Alibaba Group
Hangzhou, China
qulie.ql@alibaba-inc.com

Rongfei Jia
Alibaba Group
Hangzhou, China
rongfei.jrf@alibaba-inc.com

Binqiang Zhao
Alibaba Group
Hangzhou, China
binqiang.zhao@alibaba-inc.com

Abstract

*This paper proposes a novel noisy label detection approach, named O2U-net, for deep neural networks without human annotations. Different from prior work which requires specifically designed noise-robust loss functions or networks, O2U-net is easy to implement but effective. It only requires adjusting the hyper-parameters of the deep network to make its status transfer from **o**verfitting to **u**nderfitting (O2U) cyclically. The losses of each sample are recorded during iterations. The higher the normalized average loss of a sample, the higher the probability of being noisy labels. O2U-net is naturally compatible with active learning and other human annotation approaches. This introduces extra flexibility for learning with noisy labels. We conduct sufficient experiments on multiple datasets in various settings. The experimental results prove the state-of-the-art of O2S-net.*

1. Introduction

Although deep neural networks have already achieved tremendous success in computer vision, their performance suffers from noisy labels in training data. Noisy labels refer to labels which areassigned to wrong classes in supervised learning. In real-world situations, acquiring high-quality annotated data is costly and time-consuming. It needs massive human annotation and verification. As a result, most of the deep models applied in industry have to be trainedbased on data with a large amount of noise. As deep neural networks have the capability to memorize all training samples [20], noisy labels would be overfitted. That greatlydegenerates the performance of deep models.

Recent studies draw attention to learning with noisy labels. There are two types of solutions: 1) directly training noise-robust models on *unclean* data; 2) detecting and cleansing noisy labels before training. The noise-robust solutions [3, 14, 19, 13] typically focus on introducing regularization to reduce the effect of the overfitting on noisy labels. In the solutions of noise cleansing, potential noisy labels are first detected, and then removed from the training set [5] or fed to the model after clean samples [7, 4] to reduce their negative impact. Although these two types of solutions have their own advantages in different cases, the noise-cleansing-based approaches have value add for practical usage in industry because of the following reasons:

- **Clean Dataset**: Data is the most expensive and valuable asset for industries. Removing noisy labels naturally generates clean datasets, which can be reused for other tasks via transfer learning without considering the impact of noisy labels.

- **Human Annotations**: The combination of noisy label detection and active learning [16] can further benefit supervised learning. In industry, a raw dataset is typically allowed to be verified and annotated for multiple rounds to guarantee its cleanness. Active learning can be conducted after noisy label detection to further reduce human annotations.

- **Applicability**: Noisy label detection can also benefit noise-robust models. Recent studies [4, 7] leverage curriculum learning [2] to build noise-robust models. Estimating the probabilities of noisy labels can help develop such a curriculum to model the difficulty of samples. That extends the applicability of noise cleansing models.

In this paper, we address noisy label detection in supervised learning. We propose a simple but effective approach

to identify mislabeled samples. The details of our contributions are summarized as follows:

- We propose a novel noisy label detection approach, named *O2U-net*, without human annotation and verification. Different from prior work, O2U-net does not require specifically designed noise-robust loss functions or networks. It is quite easy to implement and can be embedded in any network. O2U-net only requires adjusting the hyper-parameters of the network to make it transfer from ***o****verfitting* to ***u****nderfitting* cyclically. By calculating and ranking the normalized average loss of every sample, the mislabeled samples can be identified. In general, the higher the loss of a sample, the higher the probability of being a noisy one. O2U-net is naturally compatible with active learning and other human annotation approaches. It would further reduce annotation cost.
- We conduct extensive experiments on multiple datasets including both synthetic label noise and real-world label noise and compare O2U-net to several recent baselines. The experimental results show that O2U-net achieves the state-of-the-art performance. In almost all the cases, O2U-net outperforms the baselines by a large margin on noisy label detection. After removing noisy labels, the performance of the neural network is further improved, compared to other baselines.

In the following sections, we briefly introduce the related work of learning with noisy labels in Section 2, and then present the details of O2U-net in Section 3. We illustrate the training process of O2U-net in Section 4 and present our experimental results in Section 5. We conclude our work in Section 6.

2. Related Work

In the literature, the solutions of learning with noisy labels can be classified into two types: 1) detecting noisy labels and then cleansing potential noisy labels or reduce their impacts in the following training; 2) directly training noise-robust models with noisy labels.

Noise-Cleansing-based Approaches Koh and Liang [8] propose an influence functions to measure which samples are "*harmful*" to model training. As the proposed approach requires intensive computation on the impact of every training sample on all the validation samples, it is hardly implemented in industry. In [21], Zhang et al. propose an approach to detect both outlier samples and hard training set bugs using a small group of trusted data. As this approach requires a strong convex assumption on the objective function, it cannot be applied to most of the deep models because such an assumption can hardly hold. In [10], Lee et al. propose a joint neural embedding network named CleanNet. This approach summarizes the knowledge of label noise from a fraction of manually verified classes. Transfer learning is then conducted to transfer the knowledge to other classes to handle label noise. The human verification lowers the applicability of this work. In [5], Han et al. propose a noisy label detection approach, named *Co-teaching*, in which two deep networks are trained simultaneously. Each network selects which samples the other network uses for training. Either of the networks teaches each other to identify noisy labels. Another similar work is proposed in [11]. In recent studies, curriculum learning [2] is applied to learning with noisy labels. In [4], Guo et al. propose *CurriculumNet*, in which training data are divided into several subsets by ranking their complexity via distribution density. The subsets are formed as a curriculum to teach the model in understanding label noise gradually. A similar idea is proposed in [7]. In this work, a *MentorNet* is trained to identify potential noisy labels. It then provides a data-driven curriculum for a *StudentNet* which is trained on the relatively clean data samples.

Noise-Robust Models In [3], label noise is modeled by additional softmax layers to estimate the transition between correct labels and noisy labels. In [19], Xiao et al. propose a probabilistic model to describe the relations among images, truth labels, noisy labels and noise types. The probabilistic model requires a small set of verified clean labels. In [14], Reed and Lee propose the notion *consistent* to model noisy labels. Sample reconstruction errors are applied as the consistency objective to estimate the noise distribution. All the above noise-transition-estimation-based approaches aim at discovering the pattern of noise in data.

Note that all the prior work of learning with noisy labels requires either particular assumptions (e.g., noise distribution estimation) or extra specifically designed loss functions or networks (e.g., Co-teaching and MentorNet). Those limit their applicability in practice. Different from the prior work, O2U-net only requires only appropriately adjusting the hyper-parameters of deep networks. It is straightforward but surprisingly effective in various situations.

3. The Proposed Model

We propose *O2U-net* which aims at detecting noisy labels without human annotations. In our setting, potential noisy labels are detected and removed from the original dataset. A final classifier is then re-trained based on the clean dataset. The final performance would be improved because of the cleansing of label noise.

3.1. Intuition

The intuition of O2U-net comes from the training process of common deep neural networks. In a typical training

process, the status of a network goes from underfitting to overfitting. At the early stage of training, the convergence speed of the network is fast. The network trends to first learn the knowledge from the samples which are "easy" to fit [1]. In the gradient-based optimization, such easy samples contribute more to the gradient computation at the early stage, and as a result, their losses decrease sharply. Conversely, the "hard" samples are usually learned at the late stage of training. If the training continues to the very late stage of training, the network would memorize every single training sample through its massive parameters and thus get overfitted. The negative impact of label noise is mainly caused by the overfitting of noisy labels.

By observing the whole training procedure on the dataset including label noise, it is found that noisy labels are usually memorized at the late stage of training as the "hard" samples. At the beginning of the training, the losses of noisy labels are larger than those of clean samples because clean samples quickly get fit at that beginning. At the late stage of training, the losses generated from noisy labels and clean labels are indistinguishable because both of them are memorized by the network. Therefore, by tracking the variation of loss of every sample at the different stages of training, it is possible to detect noisy labels to some extent. However, in an ordinary training process, the status of the network would change from underfitting to overfitting only once. Once the noisy labels are memorized, their losses would fast decrease. Moreover, when the noisy labels are overfitted is unknown. As a result, the loss tracking for every sample may not be reliable because of the lack of sufficient statistics. To overcome this issue, we introduce multiple rounds of status transfer in training. We try to keep the status of the network in changing between underfitting and overfitting cyclically. In O2U-net, we apply the cyclical learning rate (introduced in Section 3.2) to make the network transfer from overfitting to underfitting repeatedly. Fig. 1 illustrates this process. As a result, the noisy labels are identified through the statistics of their losses in the cyclical training. In general, the larger the average loss of a sample after the cyclical training, the higher the probability of being a mislabeled sample.

Recent studies of learning with noisy labels, which are based on *Curriculum Learning* (e.g., CurriculumNet and MentorNet), also share the same intuition. In these studies, a curriculum is designed to rank the difficulty of training samples. Easy samples are trained before hard samples to introduce robustness to the network. Although the ways in which these approaches model sample difficulty are different, the proposed difficulty method can be described as a function of sample losses. In their work, the potential noisy labels are not removed because they argue that noisy labels and real hard cases may not be correctly distinguished. However, in terms of our experiments presented in Section 5, removing the potential noisy labels achieves the best performance in most of the cases. It is worthy to note that both their work and O2U-net work are proposed based on the assumption that the gradient computation is dominated by the clean samples when the network is underfitting. Therefore, the proportion and the distribution of noisy labels have a huge impact on label noise detection.

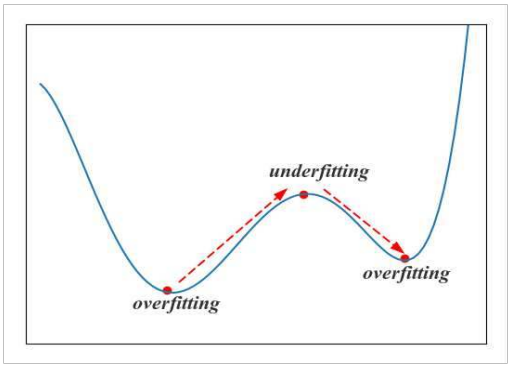

Figure 1. Cyclical Training

3.2. O2U-Net

We adjust the hyper-parameters of a deep network to make its status transferring from overfitting to underfitting cyclically. A straightforward way is to apply the cyclical learning rate. At the beginning of training, a large learning rate is set. The learning rate linearly decreases to some extent during training and is then reset to the original learning rate. This whole process repeats for multiple rounds until enough loss statistics are gathered. The idea behind is that, when the network almost converges to some minimum (nearly overfitting), a large learning rate makes the network jump out of the minimum. As a result, the network would abruptly become underfitting. We repeat this process and track the loss of every sample. We find that noisy labels generate larger losses than clean ones during the cyclical training. It should be clarified that we apply the same network to detect noisy labels and train the final classifier. The network can be any common network for image classification, e.g., ResNet, ImageNet or other customized CNNs.

The whole training process of O2U-net comprises three steps, which are introduced as follows:

1. **Pre-training:** Firstly, we follow the common setting of hyper-parameters to train the network directly on the original dataset including noisy labels. At this step, a common constant learning rate is applied. A large batch size is applied to reduce the impact of label noise [15]. We use a validation set to monitor the performance of training. The network is trained until the accuracy in the validation set stays stable.

2. **Cyclical Training:** Secondly, the cyclical learning rate is applied to continue training the network. A smaller batch size is chosen to make the network more

easily transfer from overfitting to underfitting. The network is then trained for multiple rounds based on the cyclical learning rate. The loss of every sample is recorded during the cyclical training. For a training epoch, we subtract the average loss of all the samples in this epoch from the loss of every sample to normalize the losses in different epochs.

In the cyclical train, suppose the maximum cyclical learning rate is r_1, and the minimum learning rate is r_2, where $r_1 > r_2$. We adopt a linear decrease function to cyclically adjust the learning rate. The equation for learning rate adjustment during the cyclically training is as follows:

$$\begin{aligned} s(t) &= \frac{(1 + ((t-1) \mod c))}{c}; \\ r(t) &= (1 - s(t)) \times r_1 + s(t) \times r_2, \end{aligned} \tag{1}$$

where t refers to the tth epoch in the cyclical training, c is the total number of epochs in each cyclical round and $r(t)$ is the learning rate applied at t. An example of cyclical learning rate is illustrated in Fig. 3.

After the whole cyclical training, the average of the normalized losses of every sample is computed. All the average losses are then ranked in descending order. The top $k\%$ of samples are removed from the original dataset as noisy labels, where k depends on the prior knowledge on the dataset. Such prior knowledge can be obtained by manually verifying a small group of randomly selected samples.

3. **Training on Clean Data:** Lastly, we re-initialize the parameters of the network, and re-train it on the cleansing dataset ordinarily until achieving stable accuracy and loss in the validation set. Algorithm 1 presents the whole training process (Step 1 to Step 3) of O2U-net.

4. Illustration

In the section, we illustrate the process of cyclical training (Step 2) to help explain the effectiveness of O2U-net.

In this illustration, we use *ResNet-101* [6] and the dataset *CIFAR-100* [9] to train an image classifier. As CIFAR-100 is a clean dataset, we follow the setting in [20], in which each sample is independently assigned to a uniform random label other than its true label with the probability $p = 0.2$, i.e., there are nearly 20% noisy labels. After the pre-training (Step 1), we compare the variation of sample losses in the cases of a constant learning rate and cyclical learning rate. Fig. 2 and Fig. 3 show the loss variation of the constant rate and cyclical rate respectively. In Fig. 3, it is observed that the training losses fluctuate periodically with the cyclical adjustment of learning rate. With the decrease of the learning rate, the network converges back to some minimum.

Algorithm 1 Training of O2U-Net

Input: the dataset D including a fraction of noisy labels.
Output: : the ranking R of the probabilities of being noisy labels for every samples; a classifier CLS for image classification.

Step 1: Pre-training
Initialization: the network parameters W; constant learning rate η; a large batch size b_l.
repeat
  $t = 1 \dots$ max epoch num:
    fetch mini-batch D_m from D;
    compute loss l_m on D_m;
    update $W^t = W^{t-1} - \eta \nabla l_m$.
until stable accuracy and loss in the validation set.

Step 2: Cyclical Training
Initialization: a small batch size b_s, where $b_l > b_s$; cyclical learning rate bounds r_1 and r_2; the length of a cyclical round c; the training loss for each sample $l_n = 0$.

repeat
  $t = 1 \dots$ max epoch num:
    $\eta \leftarrow r(t)$ via Eq. 1;
    fetch mini-batch D_m from D;
    compute loss l_m on D_m;
    update $W^t = W^{t-1} - \eta \nabla l_m$;
    record the loss l_n of evey sample;
    normalize l_n.
until max epoch num.
Compute the normalized average loss $\overline{l_n}$ of every sample in all the epochs;
Obtain R by ranking all the samples in descending order according to $\overline{l_n}$;
Remove top-$k\%$ samples from D to obtain a dataset D'.

Step 3: training on clean data
repeat
  conduct ordinary classifier training on D'.
until stable accuracy and loss in the validation set.
Obtain the image classifier CLS.

After the cyclical training, a sample rank is obtained according to their losses. We plot the samples in terms of four groups, which are clean samples, top $0\% - 20\%$ ranked noisy samples, top $20\% - 40\%$ noisy samples and top $40\% - 60\%$ noisy samples. These top $k\%$ samples for the constant learning rate setting and cyclical learning rate setting are selected according to their corresponding loss ranks. Every point plotted in Fig. 2 and Fig. 3 is the average loss of each group in that epoch.

It is observed that the losses of the top $0\% - 20\%$ noisy

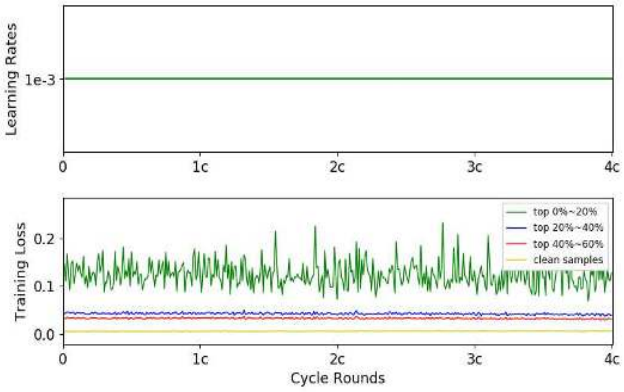

Figure 2. Loss Variation for Constant Learning Rate

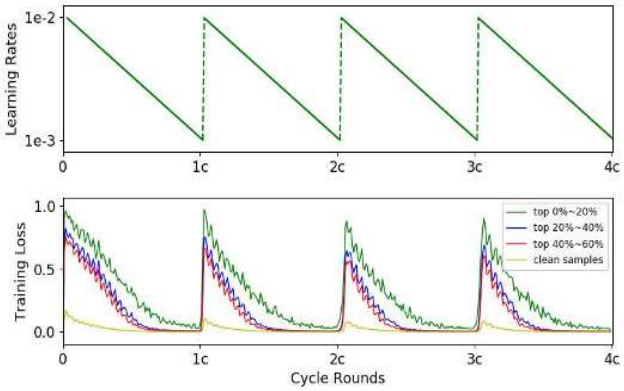

Figure 3. Loss Variation for Cyclical Learning Rate

samples fluctuate drastically for both the constant learning rate setting and the cyclical learning rate setting. The loss value gaps between the top $0\% - 20\%$ group and the clean sample group in the constant learning rate setting are much smaller than those in the cyclical learning rate setting. In this case, both the constant learning and the cyclical learning have the ability to distinguish clean samples and the most remarkable noisy samples. However, for the groups of top $20\% - 40\%$ and top $40\% - 60\%$ noisy samples, their loss variations in the cyclical learning rate setting are more notable than those in the constant learning rate setting. The loss gaps between these two groups and the group of clean samples in the cyclical learning rate setting are much larger than those in the constant learning rate setting. A larger gap implies stronger distinguishability between clean samples and noisy samples. During cyclical training, noisy samples tend to produce much larger losses than clean samples. The multiple-round cyclical training reduces the statistical bias of sample losses. Therefore, training the network from overfitting to underfitting repeatedly can not only identify remarkable label noise but produce a more accurate rank of the probabilities of being noisy samples.

The same conclusion can be seen from the precision-recall curves on noisy label detection. In Fig. 4, the precision-recall (PR) curve shows that, when recalling the same proportion of noisy samples from the corresponding loss ranks in both of the settings, cyclical training always produces higher precision on detecting noisy samples than constant learning because cyclical learning can more effectively rank noisy samples at the top.

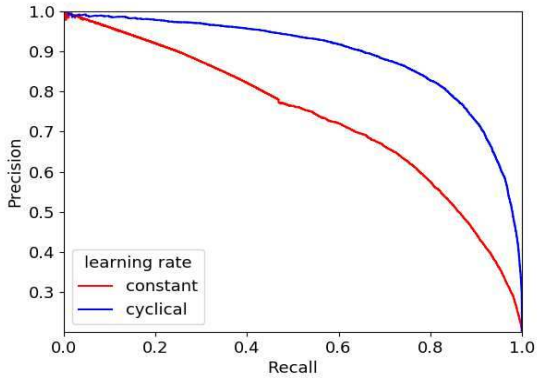

Figure 4. Precision-Recall Curves for Different Types of Learning Rate

	Training #	Test #	Class #	Image Size
CIFAR-10	$60K$	$10K$	10	28×28
CIFAR-100	$50K$	$10K$	100	28×28
Mini-ImageNet	$50K$	$10K$	100	84×84
Clothing1M	$1M + 48K$	$10K$	14	256×256

Table 1. Datasets

5. Experiments

We conduct experiments in various settings and compare O2U-net to recent outstanding baselines.

Datasets. We evaluate O2U-Net on four benchmark datasets: CIFAR-10, CIFAR-100 [9], Mini-ImageNet [18] and Clothing1M [19]. CIFAR-10 and CIFAR-100 are the most popular datasets used in the literature of learning with noisy labels [5, 7, 13, 14]. Mini-ImageNet is a popular dataset frequently used in the area of few-shot learning [18, 17, 12]. As these three datasets are clean without noisy labels, we follow the common setting in the literature [4, 5, 7] to add synthetic noise into the training sets. No noisy labels are added in the test sets. The noisy labels are added in two ways:

- *Random Noise:* Each sample in the training set is independently assigned to a uniform random label other than its true label with the probability p, where p = 10%, 20%, 40% and 80% in our experiments.

- *Pair Noise:* The samples in a class can only be misla-

	ResNet-101					9-Layer CNN				
CIFAR-10										
	10%	20%	40%	80%	Pair 10%	10%	20%	40%	80%	Pair 10%
Training with Constant Learning Rate	10.23%	19.81%	39.96%	80.06%	9.96%	9.98%	20.71%	39.41%	79.89%	10.02%
Co-Teaching	58.46%	72.32%	84.75%	83.22%	54.16%	56.80%	69.58%	80.10%	82.51%	47.59%
Co-Teaching (top 10%)	58.46%	73.43%	74.86%	94.32%	54.16%	56.80%	70.37%	82.15%	84.19%	47.59%
Curriculum	68.13%	68.51%	59.35%	80.01%	63.24%	29.51%	24.24%	42.99%	80.03%	20.02%
Curriculum (top 10%)	68.13%	75.58%	62.23%	80.23%	63.24%	29.51%	24.72%	43.19%	80.06%	20.02%
O2U-net	**94.34%**	**95.47%**	**95.67%**	**89.02%**	**91.56%**	**84.68%**	**86.56%**	**86.98%**	**84.30%**	**74.84%**
O2U-net (top 10%)	**94.34%**	**97.96%**	**98.88%**	**97.38%**	**91.56%**	**84.68%**	**95.00%**	**95.72%**	**90.94%**	**74.84%**
CIFAR-100										
	10%	20%	40%	80%	Pair 10%	10%	20%	40%	80%	Pair 10%
Training with Constant Learning Rate	9.63%	20.56%	40.41%	79.61%	10.11%	10.21%	20.18%	40.07%	79.87%	9.93%
Co-Teaching	49.60%	65.35%	78.60%	84.72%	44.94%	51.45%	65.77%	78.12%	85.08%	44.95%
Co-Teaching (top 10%)	49.60%	66.62%	79.60%	87.80%	44.94%	51.45%	70.45%	80.05%	87.98%	44.95%
Curriculum	73.03%	86.01%	76.15%	82.31%	62.19%	59.21%	78.19%	60.08%	81.20%	**63.02%**
Curriculum (top 10%)	73.03%	92.24%	91.31%	88.18%	62.19%	59.21%	87.18%	76.63%	82.14%	**63.02%**
O2U-net	**90.76%**	**92.28%**	**92.64%**	**91.69%**	**64.68%**	**80.62%**	**83.71%**	**86.34%**	**87.06%**	60.08%
O2U-net (top 10%)	**90.76%**	**96.64%**	**96.60%**	**96.02%**	**64.68%**	**80.62%**	**95.96%**	**97.40%**	**95.94%**	60.08%
Mini-ImageNet										
	10%	20%	40%	80%	Pair 10%	10%	20%	40%	80%	Pair 10%
Training with Constant Learning Rate	10.02%	19.91%	39.93%	80.05%	9.97%	10.02%	20.12%	39.98%	80.04%	9.92%
Co-Teaching	47.10%	62.16%	75.22%	81.60%	37.02%	47.39%	62.06%	73.85%	81.82%	37.14%
Co-Teaching (top 10%)	47.10%	63.78%	76.35%	86.11%	37.02%	47.39%	64.80%	75.73%	87.94%	37.14%
Curriculum	62.77%	71.19%	67.61%	80.79%	55.74%	56.95%	62.43%	63.89%	80.05%	**58.38%**
Curriculum (top 10%)	62.77%	79.78%	80.11%	83.59%	55.74%	56.95%	72.79%	73.57%	80.06%	**58.38%**
O2U-net	**81.35%**	**84.94%**	**87.23%**	**90.21%**	**59.23%**	**71.45%**	**75.63%**	**81.05%**	**85.52%**	56.55%
O2U-net (top 10%)	**81.35%**	**96.26%**	**98.71%**	**98.90%**	**59.23%**	**71.45%**	**90.28%**	**95.73%**	**93.66%**	56.55%

Table 2. Comparison on Noisy Label Detection

beled to the same one of the other classes. We follow the same noise transition matrix described in [5]. The probability of sample mislabelling in a class is 0.1.

We further evaluate O2U-Net on a large real-world dataset - Clothing1M, which is composed of clothing data crawled from online shopping websites. Clothing1M comprises 1M images with real noisy labels with additional 48K verified clean data for training. Its overall noise proportion is approximately 38%. The summary of all the datasets in our experiments is introduced in Table 1.

Baselines. We compare O2U-net to the recent outstanding approaches for learning with noisy labels:

- **Direct Training:** Direct training is the most fundamental baseline in which the image classifier is directly trained on the original dataset with noisy labels.
- **Training with Bootstrapping [14]:** This work proposes a consistency objective in which the current prediction of the model is used to resist the impact of noisy labels. We compare O2U-net to both hard-bootstrapping and soft-bootstrapping.
- **Co-teaching [5]:** This work proposes a noise-robust model that comprises two simultaneously trained networks. Each network guides the other one to select the clean samples in training.
- **MentorNet [7]:** This work leverages curriculum learning to model the difficulty of training samples. We compare O2U-Net to the proposed data-driven curriculum design method (MentoNet DD).
- **CurriculumNet [4]:** This work proposes a density-based clustering algorithm to model sample difficulty in curriculum learning.

All the baselines are re-implemented based on their open-source codes with minor modifications to fit our setting.

Networks. We evaluate O2U-Net on two networks: ResNet-101 [6] and 9-Layer CNN [5]. ResNet-101 is a proven network applied to diverse image-related tasks. The 9-Layer CNN is the network applied in the baseline Co-teaching. We slightly modify its structure to fit it to different image sizes.

Experiment Settings. We compare O2U-net to the baselines on two aspects:

- *Noisy Label Detection:* we compare the *precision* of noisy label detection of O2U-net and the other baselines. The precision is computed through the number of truly detected noisy labels over the total number of detected noisy labels. As the noise levels are set differently in our experiments, *precision* is a better metric

	ResNet-101					9-Layer CNN				
CIFAR-10										
	10%	20%	40%	80%	Pair 10%	10%	20%	40%	80%	Pair 10%
Direct Traing	88.31%	83.00%	65.66%	15.91%	88.17%	82.67%	76.42%	56.08%	17.67%	83.83%
Soft Bootstrapping	88.87%	83.20%	69.91%	18.12%	90.08%	82.68%	75.21%	54.55%	17.65%	83.55%
Hard Bootstrapping	89.69%	84.88%	68.90%	15.59%	89.17%	82.96%	75.00%	58.08%	18.18%	84.21%
MentorNet DD	92.80%	91.23%	88.64%	**46.31%**	91.02%	84.78%	80.71%	72.96%	28.19%	85.94%
CurriculumNet	90.59%	84.65%	69.45%	17.95%	90.45%	81.71%	74.02%	57.55%	16.23%	83.62%
Co-Teaching	90.36%	87.26%	82.80%	26.23%	90.77%	85.69%	82.66%	77.42%	22.60%	85.83%
O2U-net (Cycle Length 10)	93.58%	**92.57%**	**90.33%**	37.76%	**94.14%**	87.35%	84.85%	73.34%	33.18%	88.07%
O2U-net (Cycle Length 50)	**93.67%**	91.60%	89.59%	43.41%	93.99%	**87.64%**	**85.24%**	**79.64%**	**34.93%**	**88.22%**
CIFAR-100										
Direct Traing	68.89%	62.73%	48.87%	9.21%	69.10%	58.29%	49.32%	34.74%	7.25%	59.75%
Soft Bootstrapping	69.87%	62.71%	48.01%	9.05%	71.30%	58.29%	49.32%	34.74%	7.25%	60.17%
Hard Bootstrapping	70.31%	63.36%	48.55%	8.88%	70.77%	59.18%	48.97%	37.05%	7.53%	60.01%
MentorNet DD	73.14%	72.64%	67.51%	30.12%	71.96%	59.02%	52.12%	44.15%	11.21%	61.02%
CurriculumNet	73.23%	67.09%	51.68%	9.63%	73.30%	55.34%	46.31%	29.91%	4.39%	57.79%
Co-Teaching	68.81%	64.40%	57.42%	15.16%	70.02%	57.1%	53.79%	46.47%	12.23%	57.53%
O2U-net (Cycle Length 10)	75.39%	**74.12%**	**69.21%**	**39.39%**	**75.51%**	61.92%	59.32%	50.30%	15.18%	63.71%
O2U-net (Cycle Length 50)	**75.43%**	73.28%	67.00%	26.96%	75.35%	**62.32%**	**60.53%**	**52.47%**	**20.44%**	**64.50%**
Mini-ImageNet										
	10%	20%	40%	80%	Pair 10%	10%	20%	40%	80%	Pair 10%
Direct Traing	58.44%	51.27%	38.49%	7.98%	57.13%	42.64%	37.52%	25.09%	4.67%	45.08%
Soft Bootstrapping	57.42%	51.00%	38.54%	8.16%	59.11%	43.14%	37.51%	26.08%	4.63%	45.90%
Hard Bootstrapping	57.63%	50.97%	37.95%	7.66%	58.69%	43.76%	38.69%	26.58%	4.48%	45.98%
MentorNet DD	59.87%	57.66%	40.83%	15.11%	59.26%	44.98%	42.12%	33.12%	10.18%	46.12%
CurriculumNet	62.70%	55.82%	41.13%	8.75%	62.60%	41.69%	34.02%	21.02%	3.20%	44.16%
Co-Teaching	58.10%	53.41%	46.31%	6.13%	58.40%	44.85%	41.47%	34.81%	6.65%	45.38%
O2U-net (Cycle Length 10)	**63.90%**	**60.93%**	**54.77%**	**23.39%**	**63.13%**	47.63%	45.04%	38.20%	8.10%	49.45%
O2U-net (Cycle Length 50)	63.48%	60.09%	53.59%	23.15%	62.75%	**48.57%**	**45.32%**	**38.39%**	**8.47%**	**50.32%**

Table 3. Comparison on Robust Image Classifier

than *accuracy*. In our experiments, we compute two types of precisions. The first is to compute the overall precision among all the noisy labels. For example, if the proportion of noisy labels is set to 20%, then we select top 20% samples according to their final loss rank, and compute the precision based on these 20% samples. In the second type, for different noise levels, we always select top 10% samples as the detected noisy labels and compute the precision. We compute these two types of precisions because noise levels are usually unknown in real-world datasets. We always select the top 10% noisy labels for a fair comparison.

Note that, training with bootstrapping and MentorNet are not compared in this experiment because both of them conduct end-to-end training for an image classifier without an explicit process of noisy label detection.

- *Image Classification:* we compare the accuracy of the final image classifier. In O2U-net, we remove the noisy labels detected from cyclical training, and use the rest of the samples for the classifier training. O2U-net and all the other baselines are evaluated on the same clean testing set. In a cycle round of cyclical training, we adopt two different cycle lengths for further comparison, i.e., we set 10 or 50 epochs per cycle length.

	MentorNet DD	Co-Teaching	CurriculumNet	O2U-Net (Cycle Length 10)
ResNet-101	79.30%	78.52%	80.46%	**82.38%**
9-Layer CNN	70.33%	68.74%	73.33%	**75.61%**

Table 4. Comparison on Clothing1M

Hyper-parameters. We follow the original settings of ResNet-101 and 9-Layer CNN. The batch sizes in Step 1&2&3 of O2U-Net are set to 128, 16 and 128 respectively. In Step 1, the constant learning rate is 0.001. In Step 2, the cyclical learning rate is linearly adjusted from 0.01 to 0.001 in a cycle round. In a cycle round, we adopt two different cycle lengths, 10 or 50. The maximum number of epochs in Step 2 is 200. We apply the SGD optimizer with the momentum factor 0.9 and L2 penalty factor 5e-4.

5.1. Comparison Results

Noisy Label Detection. Tables 2 demonstrates the comparison results of noisy label detection between O2U-net and the other baselines. O2U-net significantly improves the precision of noise detection in most cases. With 10% pair noisy labels, CurriculumNet performs slightly better on CIFAR-100 and Mini-ImageNet. In addition, ResNet-101 can produce better performance than the 9-Layer CNN in most of the cases. It should be noted that Co-teaching and CurriculumNet are not originally proposed for noisy label

detection, but both of them involve mechanisms to compute the "difficulty" of samples. In Co-teaching, a proportion of difficult samples are removed according to their losses in every training iteration. In CurriculumNet, all the samples are classified into several groups (curriculum design) according to their density which describes the difficulty of training. In our experiments, we follow their definitions of "difficulty" to detect noisy labels and make comparisons. However, such comparisons may not be fair. To this end, the comparisons of the performance of the final image classifier are conducted in our experiments.

Image Classification. Table 3&4 shows the comparison results on image classification. In O2U-net, we remove all the detected noisy labels and train the classifier on the cleansing dataset. The other baselines are implemented in terms of their original settings. The results show that O2U-net exceeds all the other baselines in the majority of cases with both synthetic noise and real-world noise. More specifically, ResNet-101 with the cycle length 10 produces the best results in most cases. Although some clean but "hard" samples may be mistakenly removed as label noise in O2U-net. Much more clean samples are correctly kept as a tradeoff. The overall performance is thus improved.

5.2. Batch Size and Circle Length

We explore the impact of the hyper-parameters *batch size* and *circle length* in O2U-net. This experiment is conducted in the setting of 20% random noise using ResNet-101. Figs. 5&6 show the PR curves of noisy label detection using different batch sizes in cyclical training on CIFAR-10 and CIFAR-100, respectively. If the batch size is small, the gradient computation tends to be inaccurate. That, however, helps the trained network more easily deviate from the current local minimum, i.e., jumping out of overfitting. As a result, the performance of noisy label detection gets slightly better when the batch size decreases from 128 to 16. When the batch size becomes too small, e.g., 4, the performance gets worse. Such a small batch size leads to a very slow convergence speed. As a result, the network cannot be well-trained in a cycle round. We have tested 10%, 20%, 30% and 40% noise levels on every dataset in our experiments. All the results follow the similar trend. Overall, different batch sizes lead to a minor effect on the performance of O2U-Net, but a major effect on the efficiency of training.

Fig. 7 shows the precisions using different cycle lengths during cyclical training on CIFAR-100. It is observed that the impact of cycle length is weak on noisy label detection. The only requirement is that the cycle length should be long enough to ensure the status of the network transferring from underfitting to overfitting in a cycle round. The same results can also be observed on the other datasets.

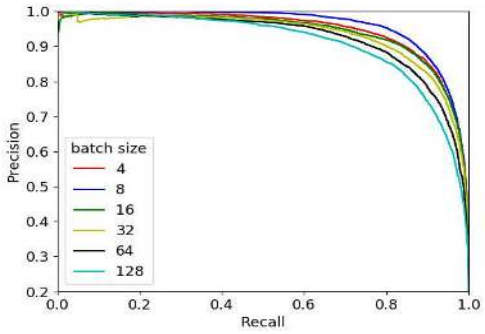

Figure 5. Batch Size Comparison on CIFAR-10

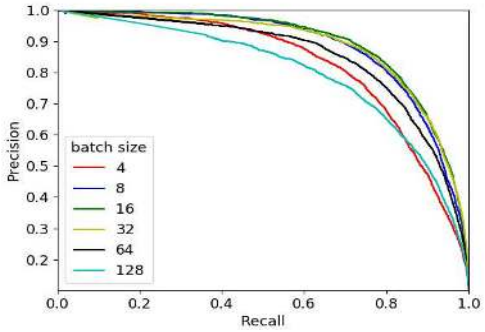

Figure 6. Batch Size Comparison on CIFAR-100

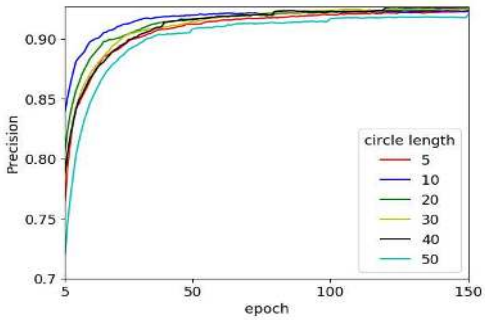

Figure 7. Cycle Length Comparison on CIFAR-100

6. Conclusions

In this paper, we have proposed O2U-net, a novel noisy label detection approach for deep networks. Different from prior studies which require extra specifically designed loss functions or networks, O2U-net is straightforward but achieve the state-of-the-art performance. It only requires adjusting the network hyper-parameters to keep the status of the trained network in transferring from overfitting to underfitting cyclically. During cyclical training, the sample losses are recorded as the indicator of the probability of label noise. We have conducted sufficient experiments on both synthetic datasets and real-world dataset. The results prove the superiority of O2U-net in various cases. O2U-net achieves high applicability because of its ease of use. It can thus apply to diverse practical demands in industry.

References

[1] Devansh Arpit, Stanislaw K. Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A Closer Look at Memorization in Deep Networks. In *the 34th International Conference on Machine Learning (ICML2017)*, pages 233–242, 2017.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. In *the 26th annual international conference on machine learning*, pages 41–48, 2009.

[3] Jacob Goldberger and Ehud Ben-Reuven. Training Deep Neural-networks Using a Noise Adaptation Layer. In *the 5th International Conference on Learning Representations (ICLR2017)*, 2017.

[4] 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 *the 15th European Conference on Computer Vision (ECCV2018)*, pages 139–154, 2018.

[5] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy labels. In *the 32nd Conference on Neural Information Processing Systems (NeurIPS2018)*, pages 8536–8546, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *the IEEE conference on computer vision and pattern recognition (CVPR2016)*, pages 770–778, 2016.

[7] 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 *the 35th International Conference on Machine Learning (ICML2018)*, pages 2309–2318, 2018.

[8] Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *the 34th International Conference on Machine Learning (ICML2017)*, pages 1885–1894, 2017.

[9] Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. Technical report, 2009.

[10] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. CleanNet: Transfer Learning for Scalable Image Classifier Training With Label Noise. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018)*, pages 5447–5456, 2018.

[11] Eran Malach and Shai Shalev-Shwartz. Decoupling "When to Update" from "How to Update". In *the 31st Conference on Neural Information Processing Systems (NeurIPS2017)*, pages 961–971, 2017.

[12] Tsendsuren Munkhdalai and Hong Yu. Meta Networks. In *the 34th International Conference on Machine Learning (ICML2017)*, pages 2554–2563, 2017.

[13] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)*, pages 2233–2241, 2017.

[14] 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.

[15] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep Learning is Robust to Massive Label Noise. *arXiv preprint arXiv:1705.10694*, 2017.

[16] Burr Settles. Active Learning Literature Survey. Technical report, 2010.

[17] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning. In *the 31st Conference on Neural Information Processing Systems*, pages 4080–4090, 2017.

[18] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. In *the 30th Conference on Neural Information Processing Systems (NeurIPS)*, pages 3630–3638, 2016.

[19] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from Massive Noisy Labeled Data for Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015)*, pages 2691–2699, 2015.

[20] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding Deep Learning Requires Rethinking Generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[21] Xuezhou Zhang, Xiaojin Zhu, and Stephen J. Wright. Training Set Debugging Using Trusted Items. In *the 32th Conference on Artificial Intelligence (AAAI2018)*, pages 4482–4489, 2018.