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

2020.10.30

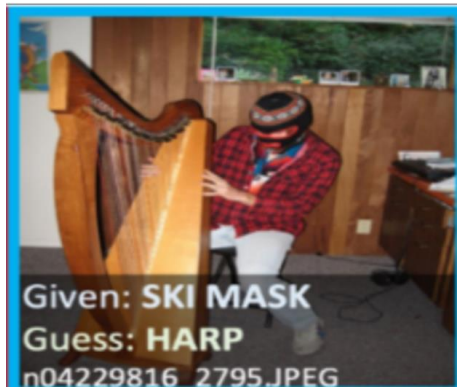
Jinchi Huang Lie Qu Rongfei Jia

ICCV 2019

motivation

There is no completely clean data set

- Multi-label image
- Ontological problem
- label error



High-quality annotated data is expensive and time consuming

- As a result, most of the deep models applied in industry have to be trained based on data with a large amount of noise

Ideas

O2U was inspired by the training process of the model: the model gradually transitioned from underfitting to overfitting.

- **In the early stage of training**, the model tends to **learn simple samples** first. After such simple samples are learned, loss will have a rapid decline and the network will converge quickly.

- **In the later stages of training**, the model **learns difficult samples** slowly.

According to loss, we can **roughly distinguish** between clean samples and nois samples

- There is a **large loss gap** between clean samples and noise samples **in the early stage** of training, and the gap between clean samples and noise samples **in the late stage of training** will **not be so large again**.

- If we can count the loss value of each sample in each period, and count its mean value and variance, we can find out the noise label to some extent according to the loss value.

Ideas

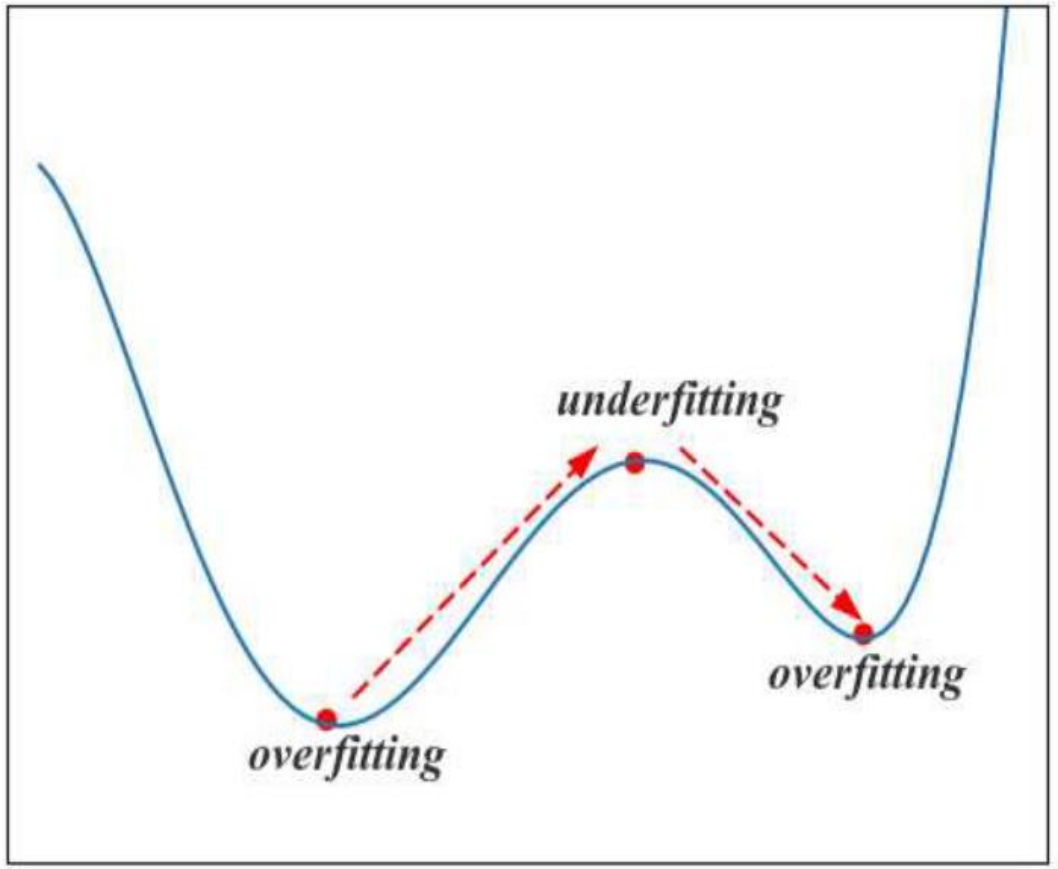
Problem: When is the noisy sample fitted by the model ?

- Once the noise sample starts fitting then the loss will go down very quickly
- It is not easy for us to determine when the noise sample starts to fit .
- If we directly count the loss of each sample, the statistical result will be not so reliable.

Ideas

Solution: Make model switch between under fitting and over fitting .

- At the **beginning** of training, **a large learning rate** is set.
- The learning rate linearly **decreases** during training .
- The loss of clean samples decreases rapidly, and the noise samples gradually decrease in the later stage. The model is slowly overfitting.
- Then reset to the original learning rate.
- Jump the model from over fitting to under fitting.
- The **loss** of Both clean samples and noise samples will **increase**, but loss of clean samples will rapidly decrease, while loss of noisy samples will slowly decrease.



Algorithm

Step1 Pre-training:

Constant learning rate and large Batch size are used to train on the full data set.

Initialization: the network parameters W; constant learning rate η; a large batch size b_l.

repeat

$t = 1 \ldots$ 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.

Generator

Step 2: Cyclical Training

r1: maximum learning rate
r2: minimum learning rate
C: The total number of epochs in Cyclical Training.
t: The tth epoch in the cyclical training.
r(t) : The learning rate at t

$$s(t) = \frac{(1 + ((t-1) \mod c))}{c};$$
$$r(t) = (1 - s(t)) \times r_1 + s(t) \times r_2, \quad (1)$$

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
  $t = 1 \ldots$ 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'.

illustration

To illustrate the cyclical training is more conducive to the separation of noise and clean samples in the dataset.

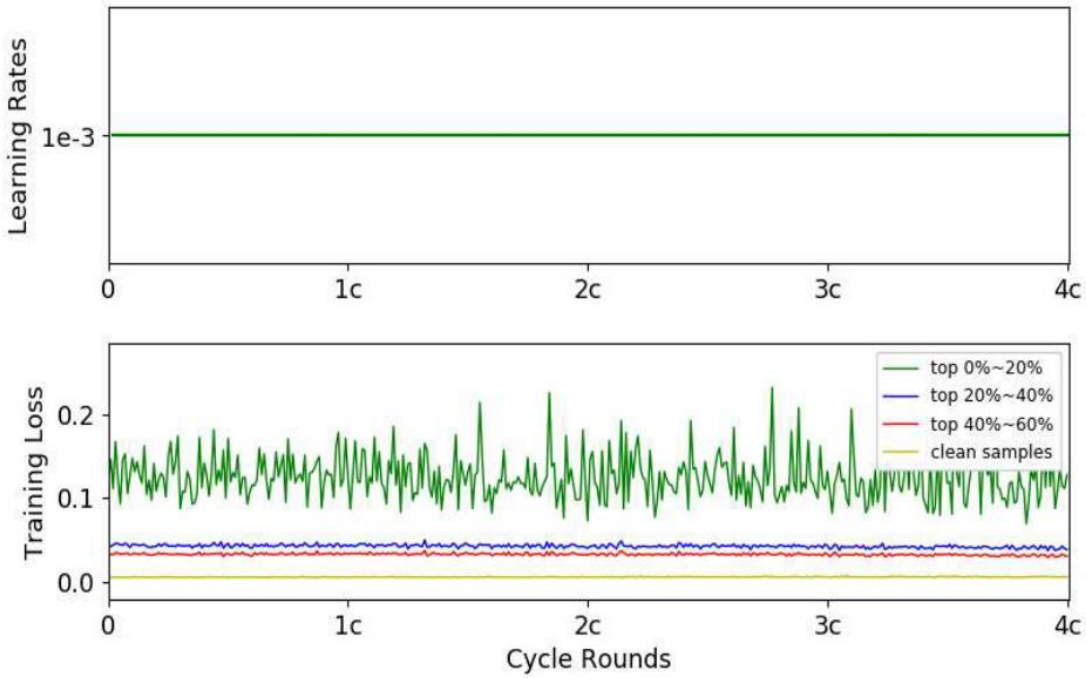


Figure 2. Loss Variation for Constant Learning Rate

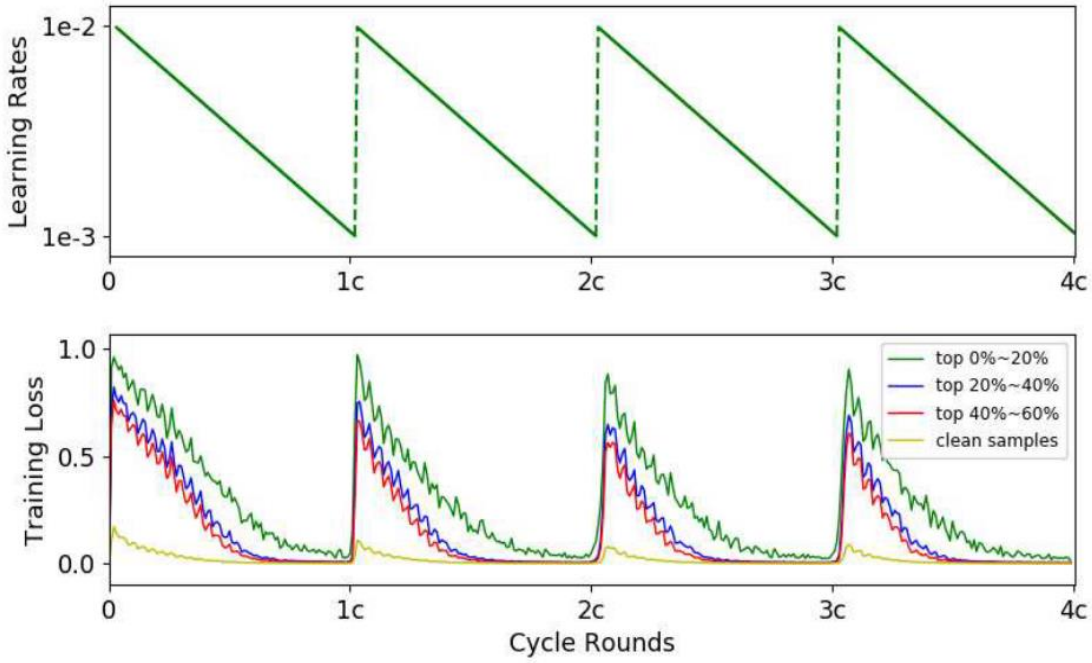


Figure 3. Loss Variation for Cyclical Learning Rate

Discriminator

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.

Experiment

Datasets

- **CIFAR-10**
- **CIFAR-100**
- **Mini-ImageNet**
- **Clothing1M**

Noisy label

- **Random Noise:**

Each sample in the training set is independently assigned to a uniform random label other than its true.

- **Pair Noise:**

The samples in a class can only be mislabeled to the same one of the other classes.

Experiment

Baselines

- **Direct Training**

The image classifier is directly trained on the original dataset with noisy labels.

- **Training with Bootstrapping**

This work **proposes a consistency objective** in which the current prediction of the model is used to resist the impact of noisy labels.

- **Co-teaching**

This work proposes that simultaneously train networks. Each network guides the other one to select the clean samples in training.

- **CurriculumNet**

This work **proposes a density based clustering algorithm** to model sample difficulty in curriculum learning.

- **MentorNet**

This work leverages curriculum learning to model the difficulty of training samples.

Experiment

Experiment Settings

This work compares O2U-net to the baselines on two aspects:

- **Noisy Label Detection:**

The work compares the precision that computed through the number of truly detected noisy labels over the total number of detected noisy labels.

- **Image Classification：**

The work compares the accuracy of the final image classifier.

Experiment

	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

Experiment

	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

Thanks