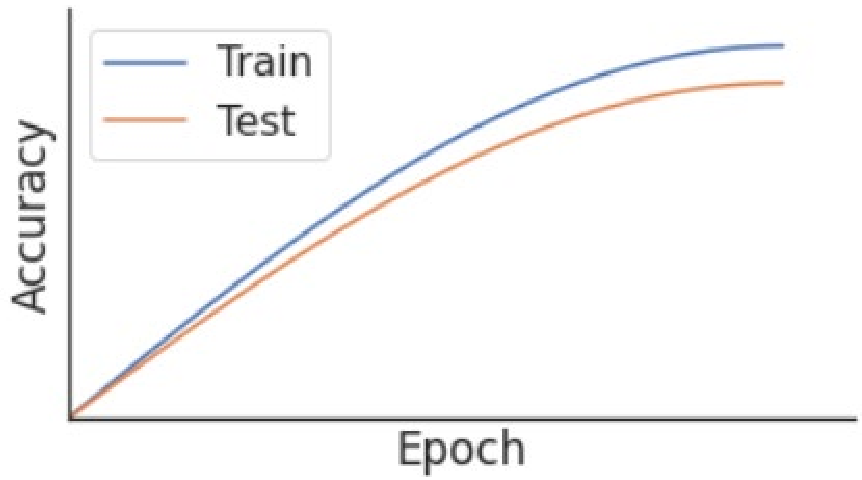
Robust and On-the-fly Data Denoising For Image Classification

Jiaming Song, Yann Dauphin, Michael Auli, Tengyu Ma

Automatically finds “leopards” in CIFAR100 training set!

Supervised learning in deep learning

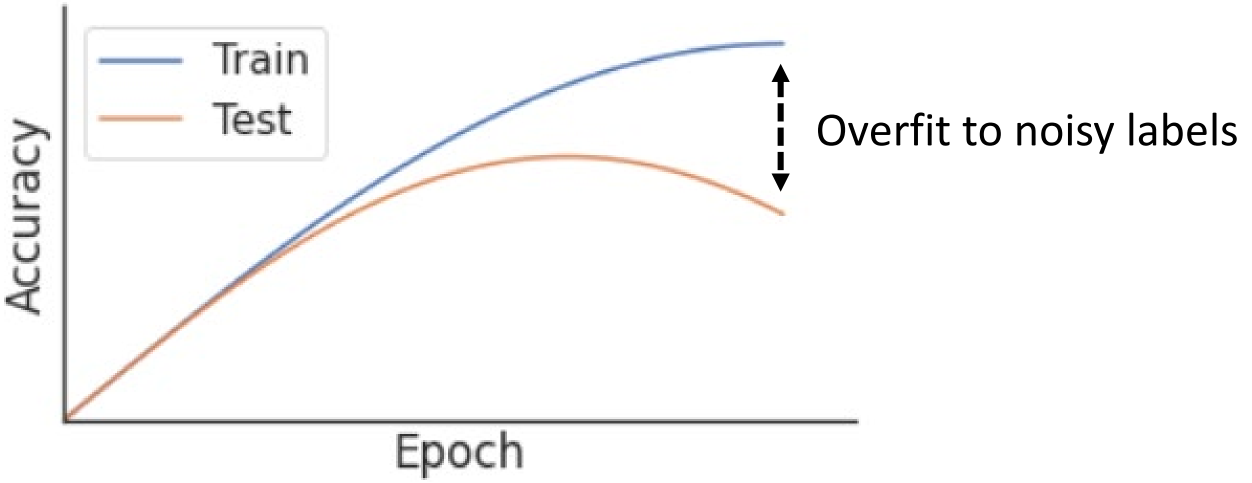


Train and test set from same distribution

- Low generalization error
- High train accuracy -> high test accuracy

Noisy labels negative impact performance!

- What if the train distribution has noisy labels?



- High generalization error
- High train accuracy -> low test accuracy

- Noisy labels arise from web supervision, mechanical turk...

Challenges for Image Classification

- Deep neural networks can overfit noisy labels easily
- Noisy labels are common in practice
  - web supervision, mechanical turk...
- Lack of domain-specific knowledge about noisy labels
  - e.g. % of labels are noisy, or noise transition matrix

Can we identify noisy labels under these restrictions?

Yes!

Our Approach

Step 1: identify noisy labels under these restrictions

Step 2: remove identified examples

Step 3: train with remaining examples

Result: simple approach that with SOTA performance!

Our Approach

Step 1: identify noisy labels under these restrictions

Step 2: remove identified examples

Step 3: train with remaining examples

Result: simple approach that with SOTA performance!

Step 1: entropy-based assumption

Assumption: noisy labels have higher conditional entropy

"entropy of clean labels" < "entropy of noisy labels"

Intuition: labeling sources have different opinions

→ chair

chair

chair

clean labels

→ leopard

panther

bear

noisy labels

Step 1: noisy labels -> higher loss

Assumption: noisy labels have higher conditional entropy

“entropy of clean labels” < “entropy of noisy labels”

Intuition: labeling sources have different opinions

Cross entropy loss = KL divergence + Entropy

When KL = 0, noisy labels will have higher loss!

Step 1: uniform noisy labels

But we know almost nothing about noisy labels!

What if the dataset contains uniform noisy labels?

X -> Uniform(Y)

leopard

chair

tree

Uniform noisy labels -> high entropy -> high loss!

Step 1: a simplified case

Let us consider an easier, *counterfactual* situation:

- Only source of noisy labels in dataset is Uniform(Y).
- Can we identify these labels (regardless of %)?

Yes!

The loss values of **uniform noisy labels**

- *(when trained on ResNets with large learning rates)*
- almost does not decrease / depend on the amount
- and can be estimated **with the model parameters**!

Step 1: simulate loss distribution

The loss values of **uniform noisy labels**

- almost does not decrease / depend on the amount
- and can be estimated **with the model parameters**!

How to simulate?

$$\boldsymbol{y} = \mathrm{fc}(\max(\boldsymbol{x}, 0))$$

fc = last fully connected layer

$$\text{Cross Entropy}(\boldsymbol{y}, k)$$

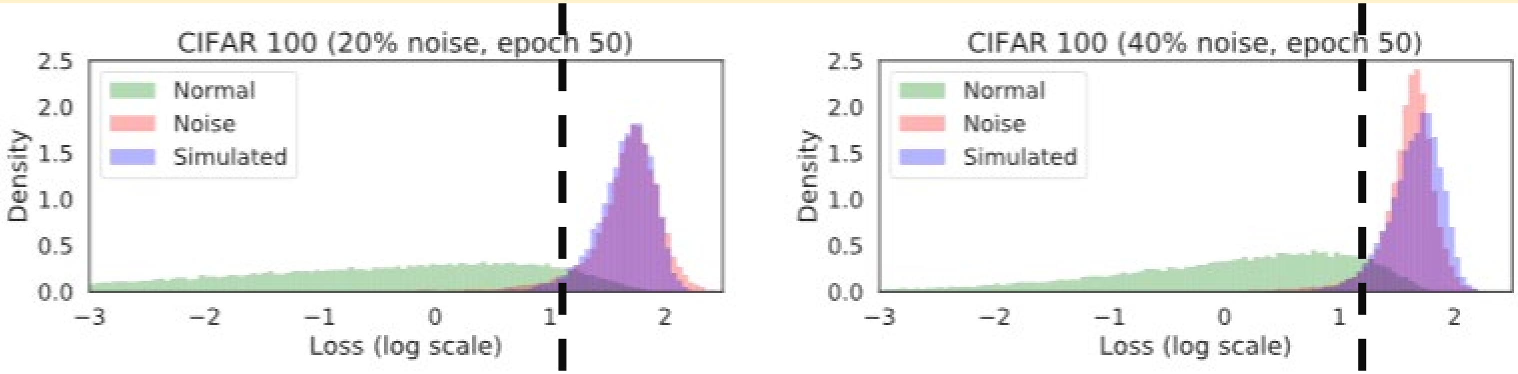
$$\boldsymbol{x} \sim \mathcal{N}(0, I)$$

$$k \sim \text{Uniform}(Y)$$

Step 1: validate our claims

Setup: CIFAR-100, 20% / 40% of noise, lr = 0.1

- Only source of noisy labels in dataset is Uniform(Y).



Observations: loss distribution for uniform labels

- is very different from that of normal labels
- are similar, regardless of percentage (20%, 40%)
- and can be estimated **with the model parameters**!

Step 1: uniform case -> practical cases

How about non uniform noise?

1. Uniform noisy labels -> high entropy -> high loss!
2. Uniform loss distribution does not depend on %

In practice

- 0% percent uniform noise
- Estimate “high loss” regions based on model parameters
- If an example has “high loss”, then it is probably noisy!

Step 1: validate the proposed method

Example: identify CIFAR-100 “noisy” labels in train set

Automatically find clearly mislabeled examples in CIFAR-100!

Mislabeled “leopards” (most are tigers and panthers)

Our Approach

Step 1: identify noisy labels under these restrictions

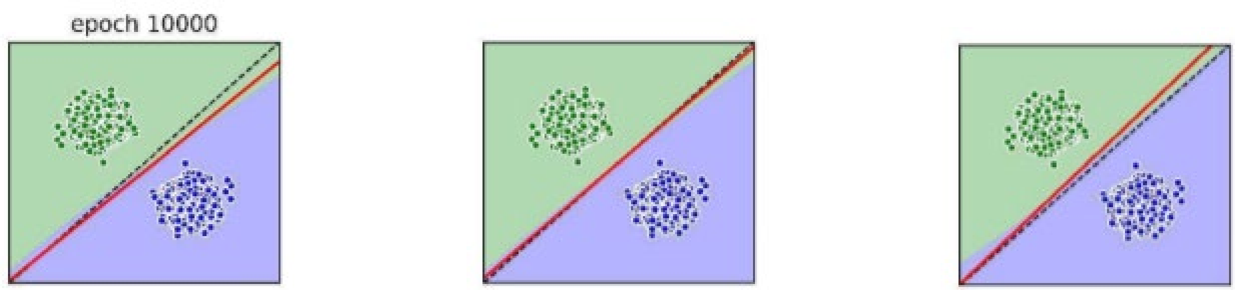
Step 2: remove identified examples

Step 3: train with remaining examples

Result: simple approach that with SOTA performance!

Step 2: remove identified examples (why)

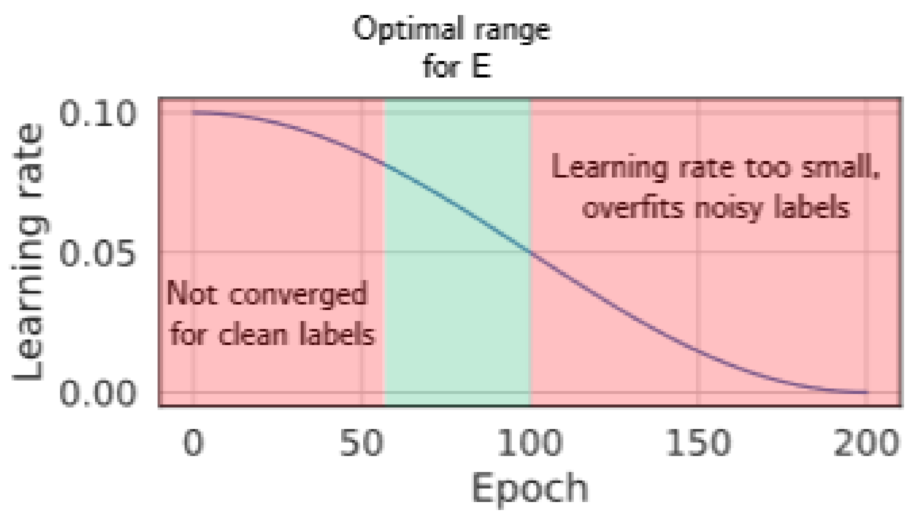
Why? Reweighting does not entirely prevent overfitting .



- Weighted by 10:1, 1:1, 1:10 (figure from Byrd and Lipton, 2019)
- Decision boundary does not change much from weighting!

Step 2: remove identified examples (when)

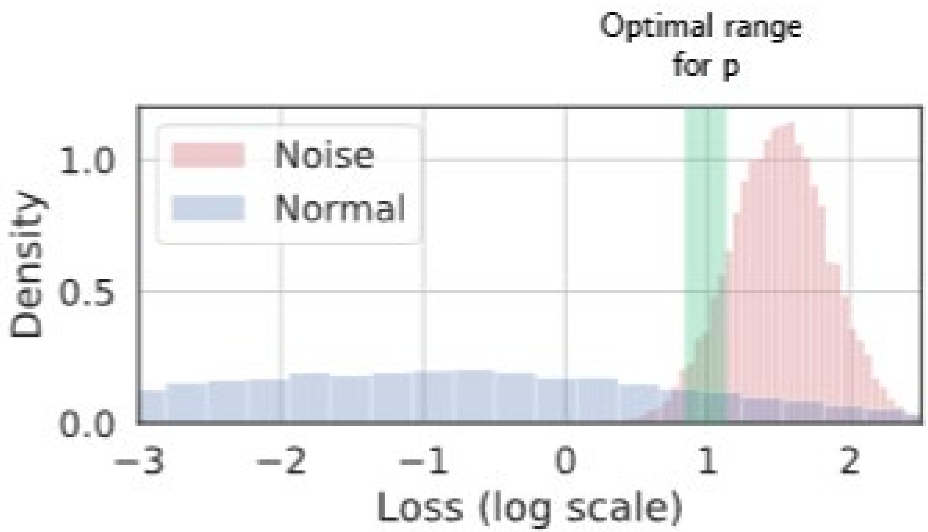
When? Remove samples when learning rate is still high.



- **Too early**: clean labels are not properly learned
- **Too late**: small learning rate, overfits noisy labels

Step 2: remove identified examples (what)

What? Remove samples with loss larger than p-th quantile



- **Aggressive threshold:** risk removing more clean examples
- **Weak threshold:** risk keeping more noisy examples

Our Approach

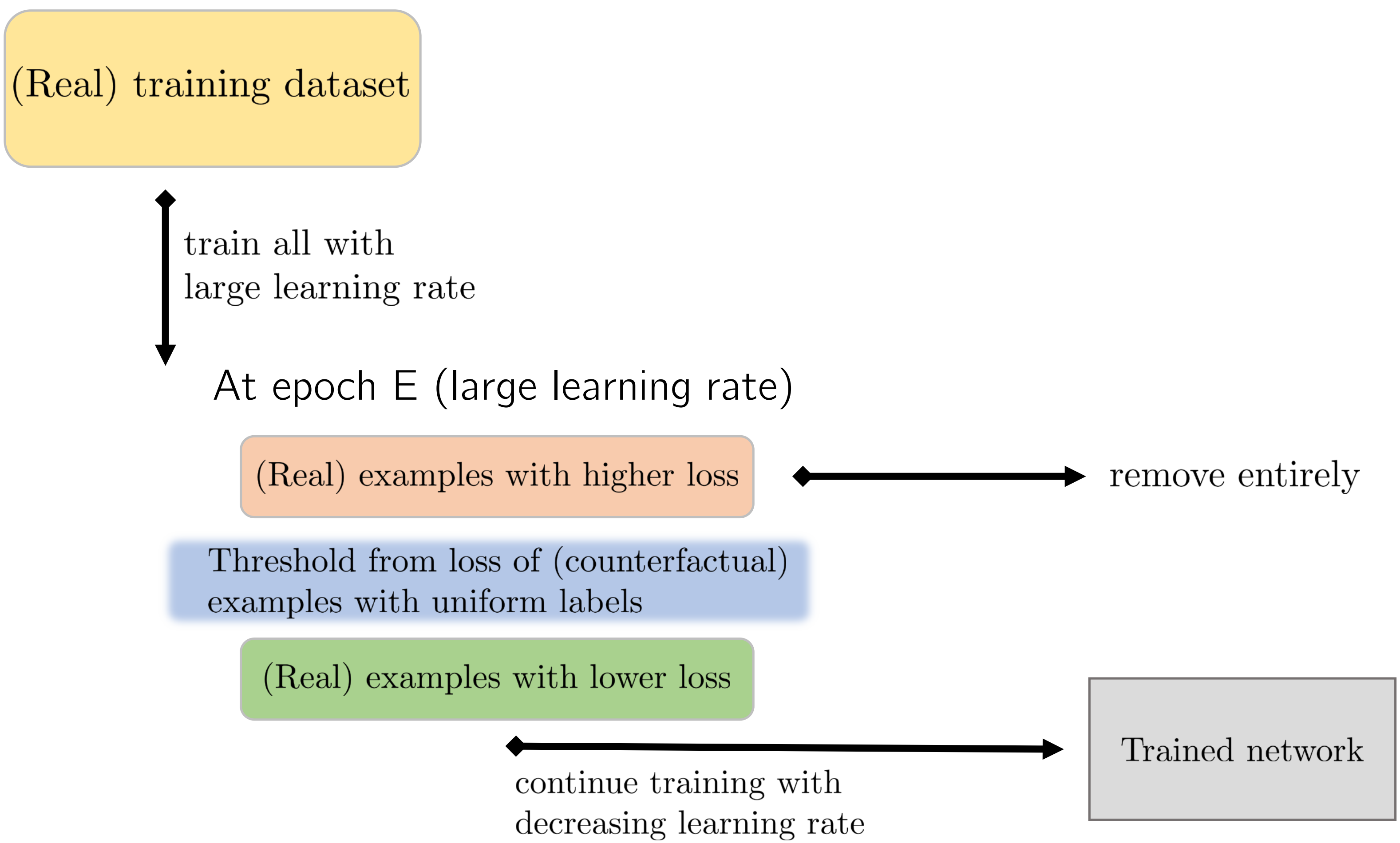
Step 1: identify noisy labels under these restrictions

Step 2: remove identified examples

Step 3: train with remaining examples

Result: simple approach that with SOTA performance!

Overview of On-the-fly Data Denoising



Experiments

Datasets

- CIFAR-10, CIFAR-100, ImageNet (clean)
- WebVision, Clothing1M (noisy)

Noise

- Artificial (uniform, non-homogenous)
- Natural (inherent in dataset)

Our method (ODD)

- achieves SOTA-level performance
- has virtually no computational overhead

CIFAR-10 and CIFAR-100

Uniform label noise (0%, 20%, 40%)

Table 1. Validation accuracy (in percentage) with uniform label noise. $\star$ denotes methods trained with knowledge of 1000 additional clean labels

	CIFAR-10			CIFAR-100		
% mislabeled	0	20	40	0	20	40
ERM	96.3	88.5	84.4	81.6	69.6	55.7
mixup	97.0	93.9	91.7	81.4	71.2	59.4
GCE	-	89.9	87.1	-	66.8	62.7
LUO	96.2	**96.2**	94.9	81.4	80.6	74.2
REN*	-	-	86.9	-	-	61.4
MENTORNET*	-	92.0	89.0	-	73.0	68.0
ODD	96.2	94.7	92.8	81.8	77.2	72.4
ODD + *mixup*	**97.2**	**95.6**	**95.5**	**82.5**	**79.1**	**76.5**

WebVision / ImageNet

- 1000 classes, 2M images labeled with web supervision

Table 4. Top-1 (top-5) accuracy on WebVision and ImageNet validation sets when trained on WebVision.

Method	WebVision	ImageNet
LASS [1]	66.6 (85.6)	59.0 (80.8)
CleanNet [20]	68.5 (86.5)	60.2 (81.1)
ERM	69.7 (87.0)	62.9 (83.6)
MentorNet [16]	70.8 (88.0)	62.5 (83.0)
CurriculumNet [9]	73.1 (89.2)	64.7 (84.9)
ODD	72.6 (89.3)	64.8 (85.5)

Clothing1M

- 14 classes, containing 50k clean and 1M noisy images

Table 5. Validation accuracy on Clothing1M.

Method	Setting	Accuracy
ERM	noisy	68.9
GCE	noisy	69.1
Loss Correction [30]	noisy	69.2
LCCN [43]	noisy	71.6
Joint Opt. [39]	noisy	72.2
DMI [42]	noisy	72.5
ODD	noisy	**73.5**
ERM	clean	75.2
Loss Correction	noisy + clean	**80.4**
ODD	noisy + clean	**80.3**

Summary

Problem: dataset contains labels that are incorrect / noisy

Solution: implicit regularization helps find noisy examples!

Advantages:

- Virtually no computational overhead
- Does not require prior knowledge of noise
- State-of-the-art performance

Automatically finds “leopards” in CIFAR100 training set!