ICML | 2020
Thirty-seventh International Conference on Machine Learning

Google AI

Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels

Lu Jiang, Di Huang, Mason Liu, Weilong Yang.


Lu Jiang


Di Huang


Mason Liu


Weilong Yang

Deep Learning on Noisy Labels

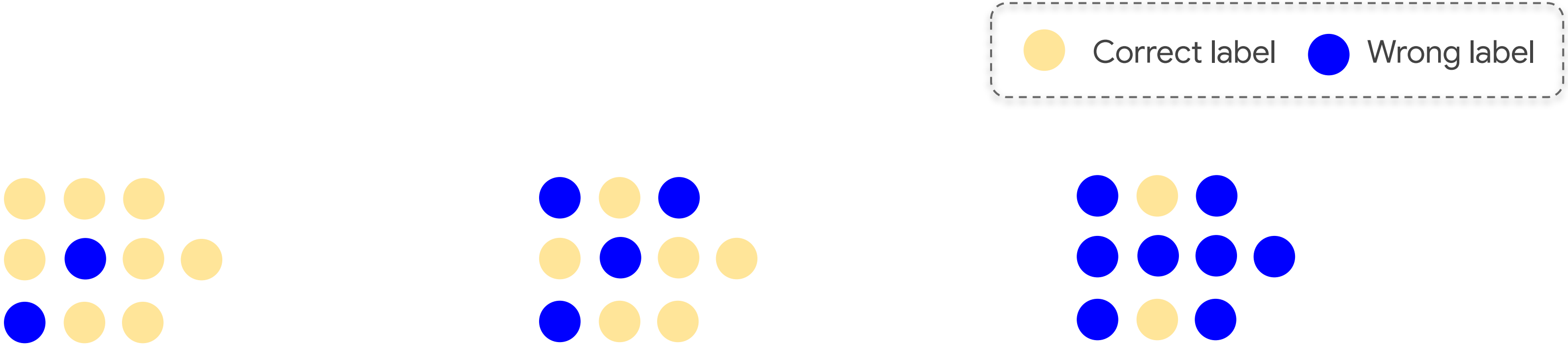
Deep networks are very good at memorizing the noisy labels (*Zhang et al. 2017*).

Memorization leads to a critical issue since noisy labels are inevitable in big data.

Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." *ICLR* (2017).

Controlled Noisy Labels

Performing controlled experiments on noisy labels is essential in existing works.



noise level=20%

40%

80%

Issues with Controlled Synthetic Labels

Issue: existing studies only perform controlled experiments on synthetic labels (or random labels).

Issues with Controlled Synthetic Labels

Issue: existing studies only perform controlled experiments on synthetic labels (or random labels).

1. Contradictory findings.
   For example, DNNs are robust to massive label noise?

UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Our central finding can be summarized as:

Deep neural networks easily fit random labels.

(Zhang et al. 2017)

OR

Deep Learning is Robust to Massive Label Noise

David Rolnick [*1] **Andreas Veit** [*2] **Serge Belongie** [2] **Nir Shavit** [3]

(Rolnick et al. 2017)

Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." *ICLR* (2017).
Rolnick, D., et al. Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694, 2017.

Issues with Controlled Synthetic Labels

Issue: existing studies only perform controlled experiments on synthetic labels (or random labels).

2. Inconsistent empirical results
   We found that methods that perform well on synthetic noise may not work as well on real-world noisy labels.

- Motivation of our research project.

Our Contributions:

1. We establish the first benchmark of controlled real-world label noise (from the web).

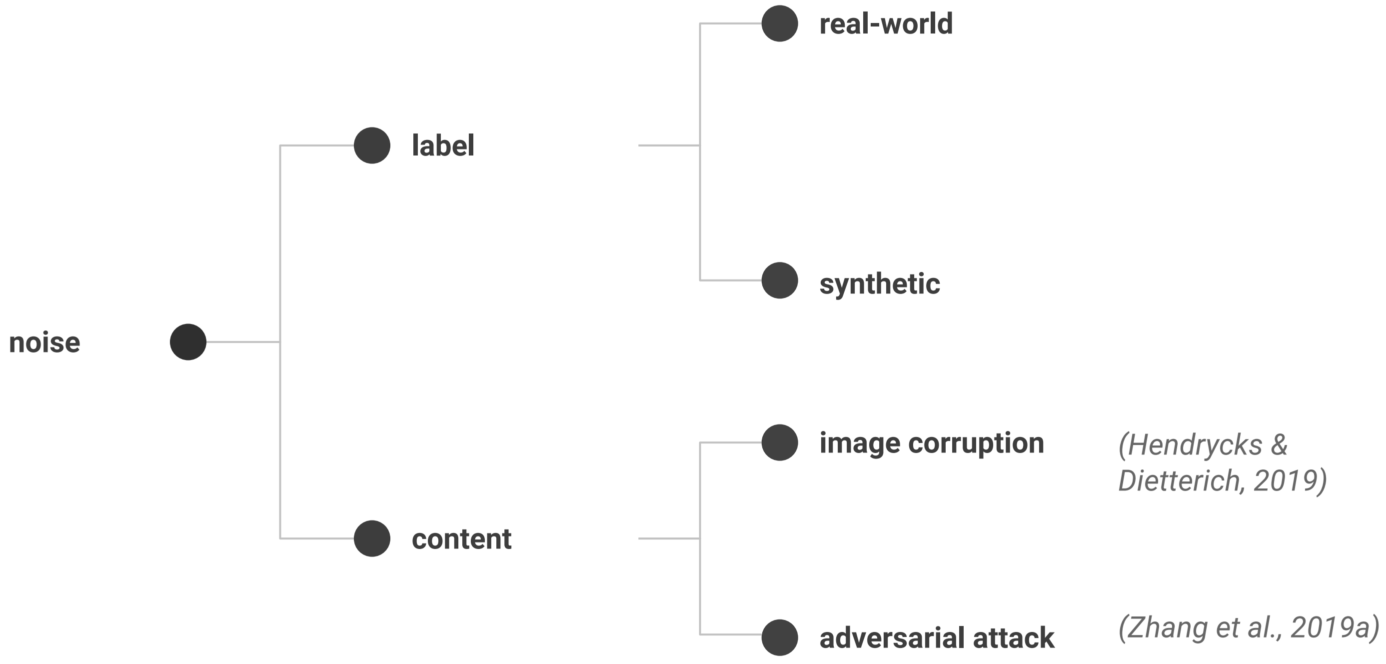
2. A simple but highly effective method to overcome both synthetic and real-world noisy labels (best results on the WebVision benchmark)

3. We conduct the largest study by far into understanding deep neural networks trained on noisy labels across different noise levels, noise types, network architectures, methods, and training settings.

Contribution I: New Dataset

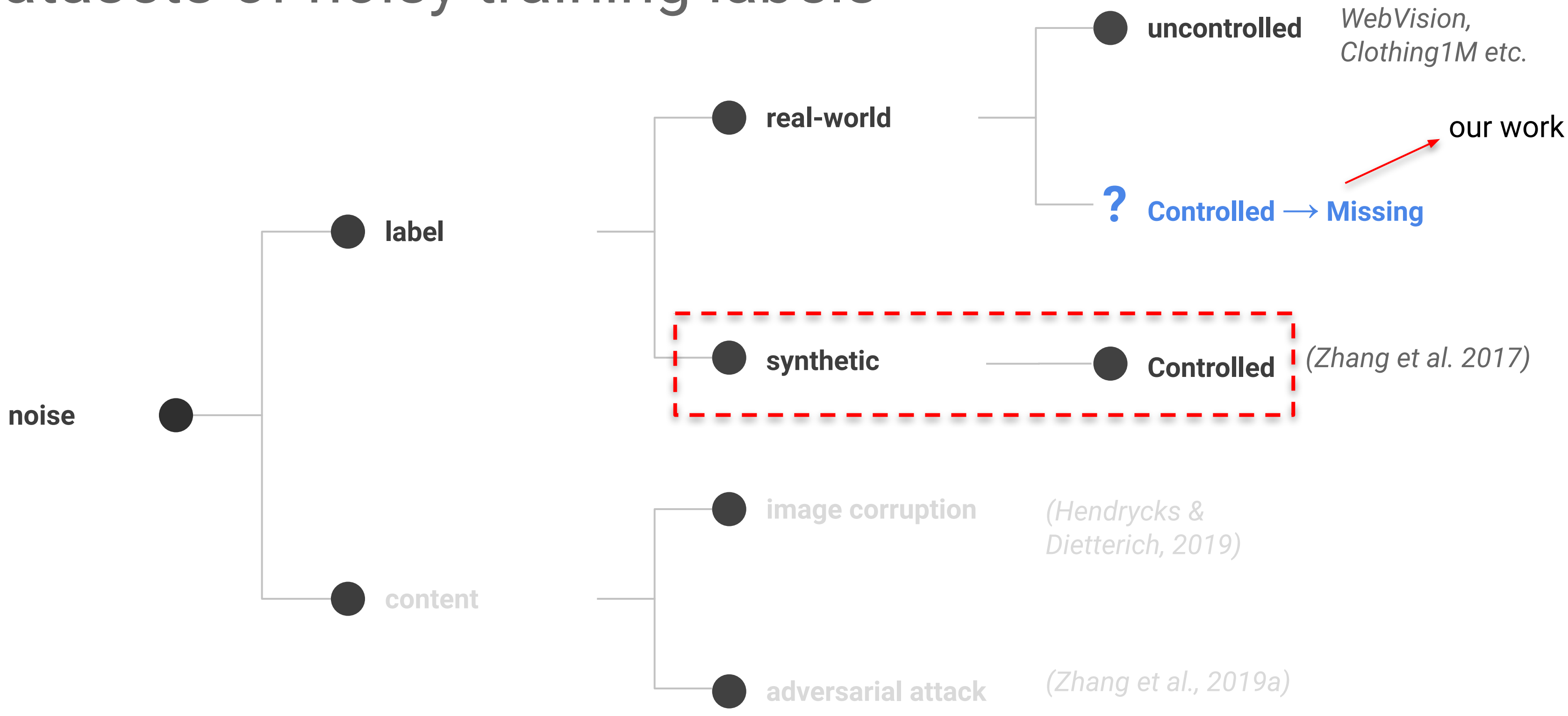
First benchmark of controlled real-world label noise

Datasets of noisy training labels

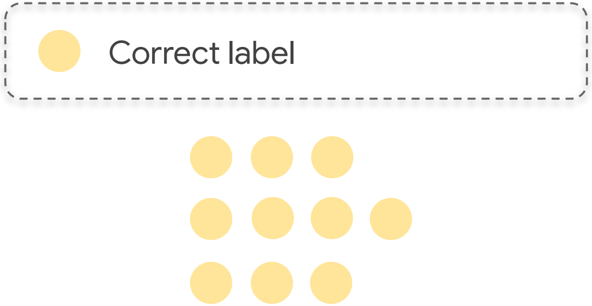


Datasets of noisy training labels

Datasets of noisy training labels



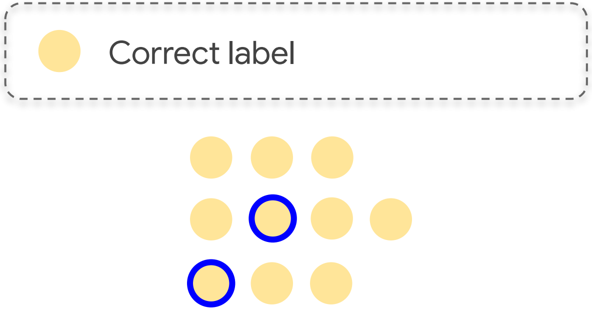
Construction of controlled synthetic label noise



Mini-ImageNet

1. **Starts with a well-labeled dataset.**
2. Randomly selects p% examples.
3. Independently flips each label to a random incorrect class (symmetric or asymmetric).
4. Repeats Step 1-3 with a different p (noise level)

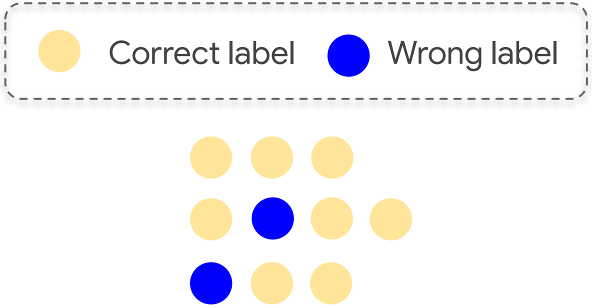
Construction of controlled synthetic label noise



noise level p = 20%

1. Starts with a well-labeled dataset.
2. **Randomly selects p% examples.**
3. Independently flips each label to a random incorrect class (symmetric or asymmetric).
4. Repeats Step 1-3 with a different p (noise level)

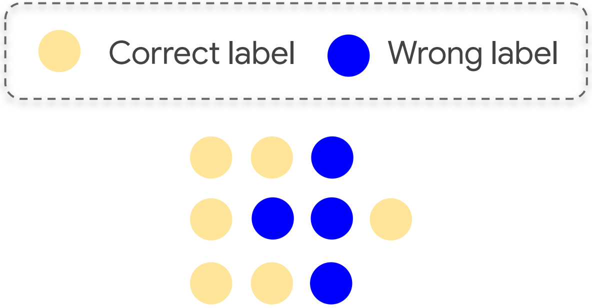
Construction of controlled synthetic label noise



noise level p = 20%

1. Starts with a well-labeled dataset.
2. Randomly selects p% examples.
3. **Independently flips each label to a random incorrect class (symmetric or asymmetric).**
4. Repeats Step 1-3 with a different p (noise level)

Construction of controlled synthetic label noise

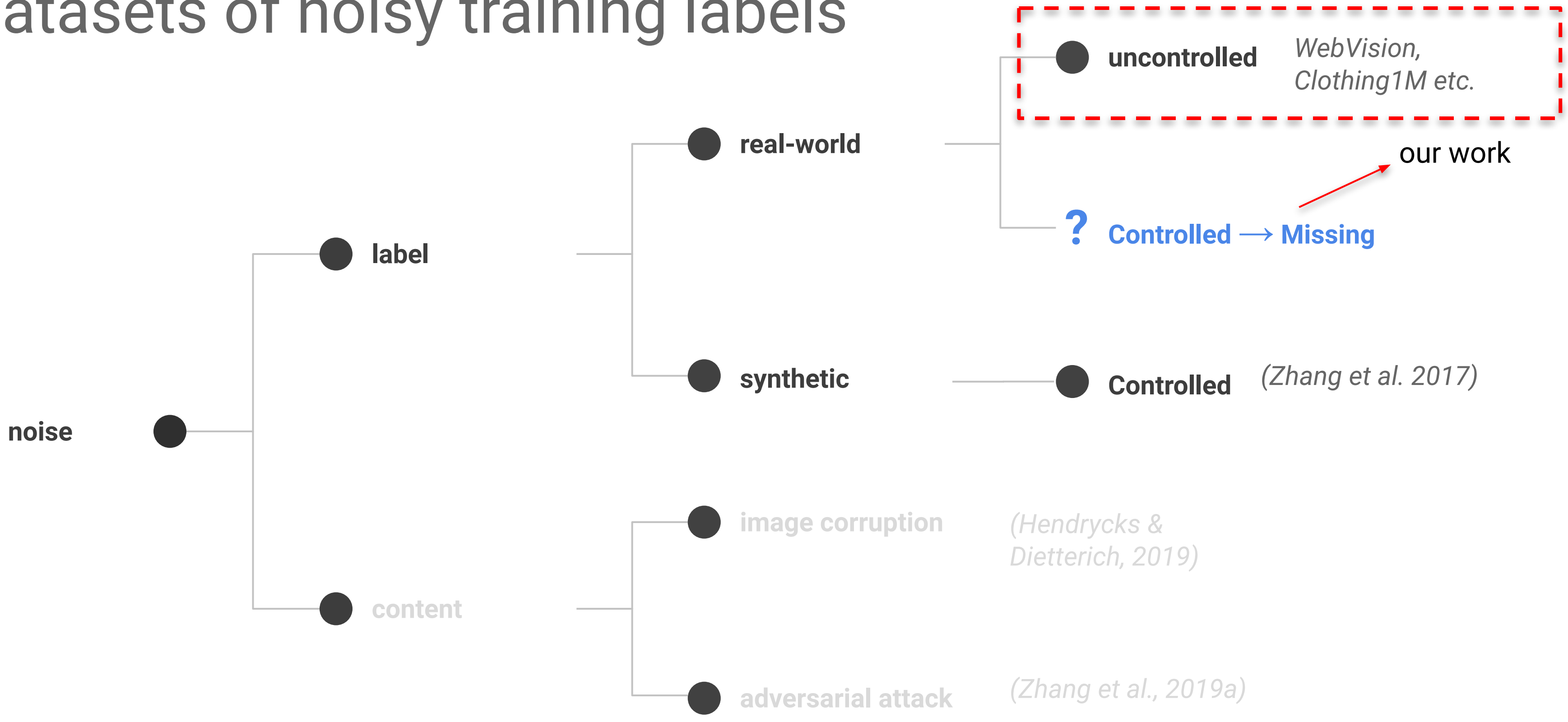


noise level **p = 40%**

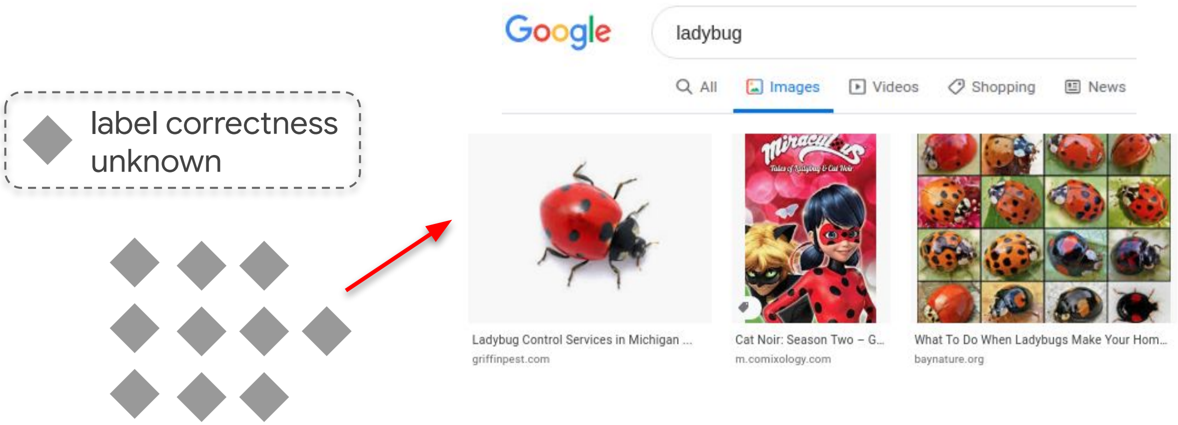
1. Starts with a well-labeled dataset.
2. Randomly selects p% examples.
3. Independently flips each label to a random incorrect class (symmetric or asymmetric).
4. **Repeats Step 1-3 with a different p (noise level)**

This process generates controlled synthetic label noise.

Datasets of noisy training labels



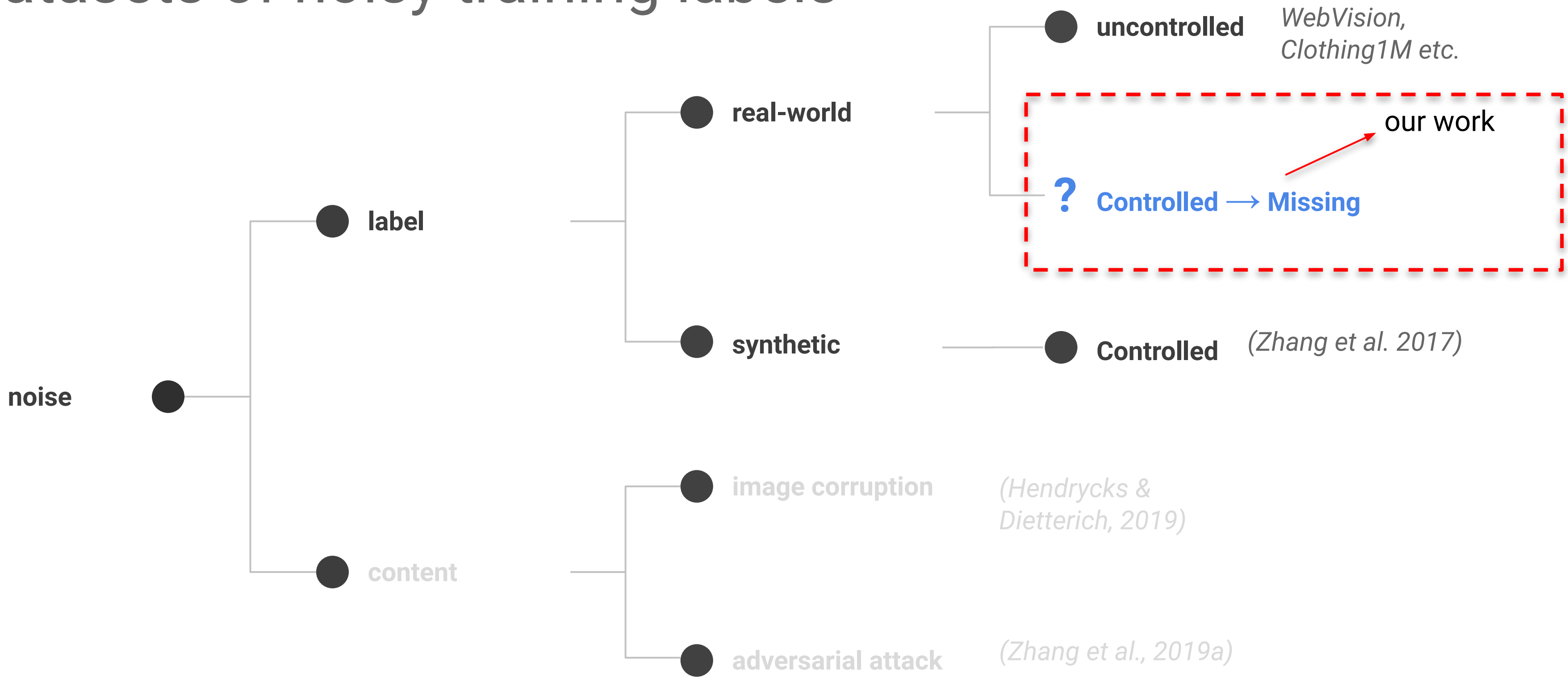
Construction of uncontrolled web label noise



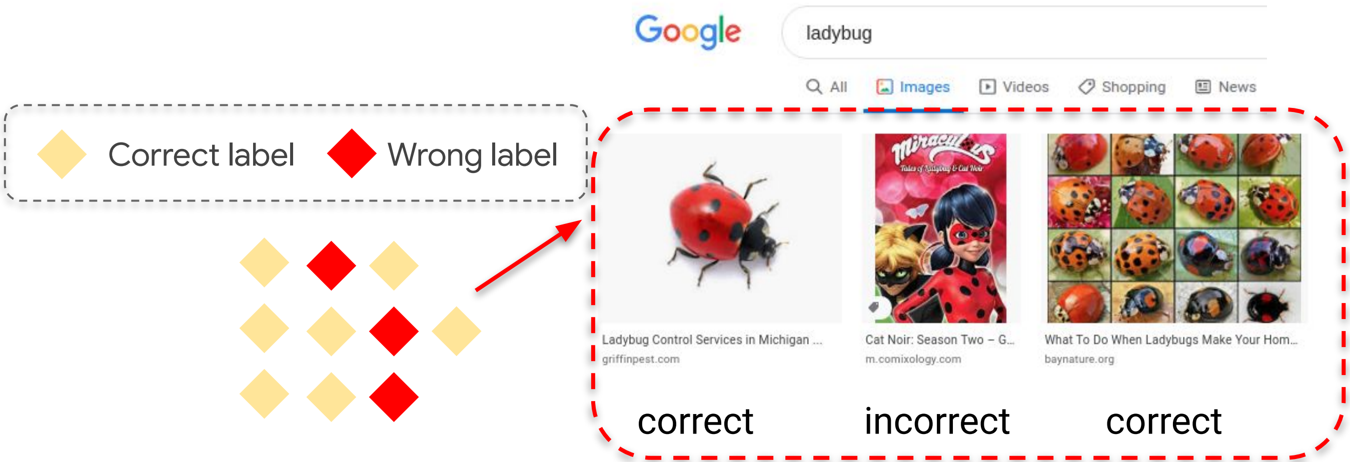
noise level **p = ??%**

This process can automatically collect noisy labeled images from the web.
But the noise level is fixed and unknown (unsuitable for controlled studies).

Datasets of noisy training labels



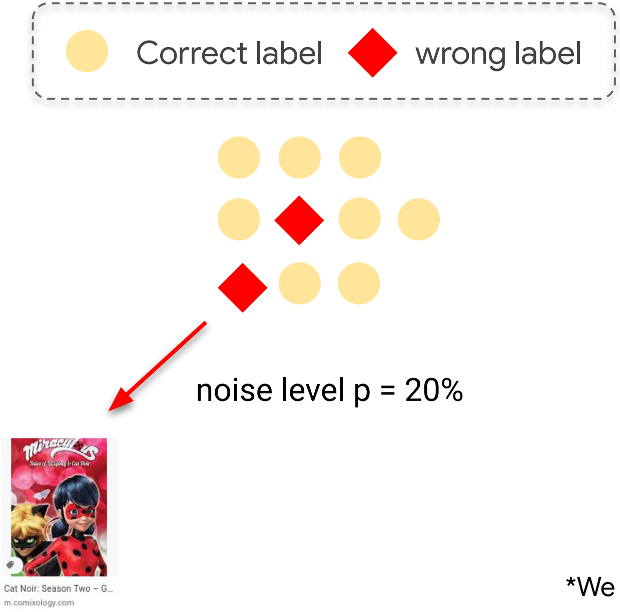
From uncontrolled to controlled noise



noise level **p is known**

We have each retrieved image annotated by 3-5 works using Google Cloud Labeling Service
https://cloud.google.com/ai-platform/data-labeling/docs

Construction of our dataset



1. Starts with a well-labeled dataset.

2. Randomly selects p% examples.

3. **Replaces the clean images with the incorrectly labeled web images while leaving the label unchanged*.**

4. Repeats Step 1-3 with a different p (noise level)

*We show that an alternative way to construct the dataset by removing all image-to-image results leads to consistent results in the Appendix

Our Dataset: Controlled Noisy Labels from the Web

Manually annotate 212K images through 800K annotations.
We establish the first benchmark of controlled web label noise for two classification tasks: coarse (Mini-ImageNet) and fine-grained (Stanford Cars)

Table 1. Overview of our datasets of controlled red (web) label noise. Blue (synthetic) label noise is also included for comparison.

Dataset	#Class	Noise Source	Train Size	Val Size	Controlled Noise Levels (%)
Red Mini-ImageNet	100	image search label	50,000	5,000	0, 5, 10, 15, 20, 30, 40, 50, 60, 80
Blue Mini-ImageNet		symmetric label flipping	60,000		0, 5, 10, 15, 20, 30, 40, 50, 60, 80
Red Stanford Cars	196	image search label	8,144	8,041	0, 5, 10, 15, 20, 30, 40, 50, 60, 80
Blue Stanford Cars		symmetric label flipping	8,144		0, 5, 10, 15, 20, 30, 40, 50, 60, 80

Our Dataset: Controlled Noisy Labels from the Web

Manually annotate 212K images through 800K annotations.
We establish the first benchmark of controlled web label noise for two classification tasks: coarse (Mini-ImageNet) and fine-grained (Stanford Cars)

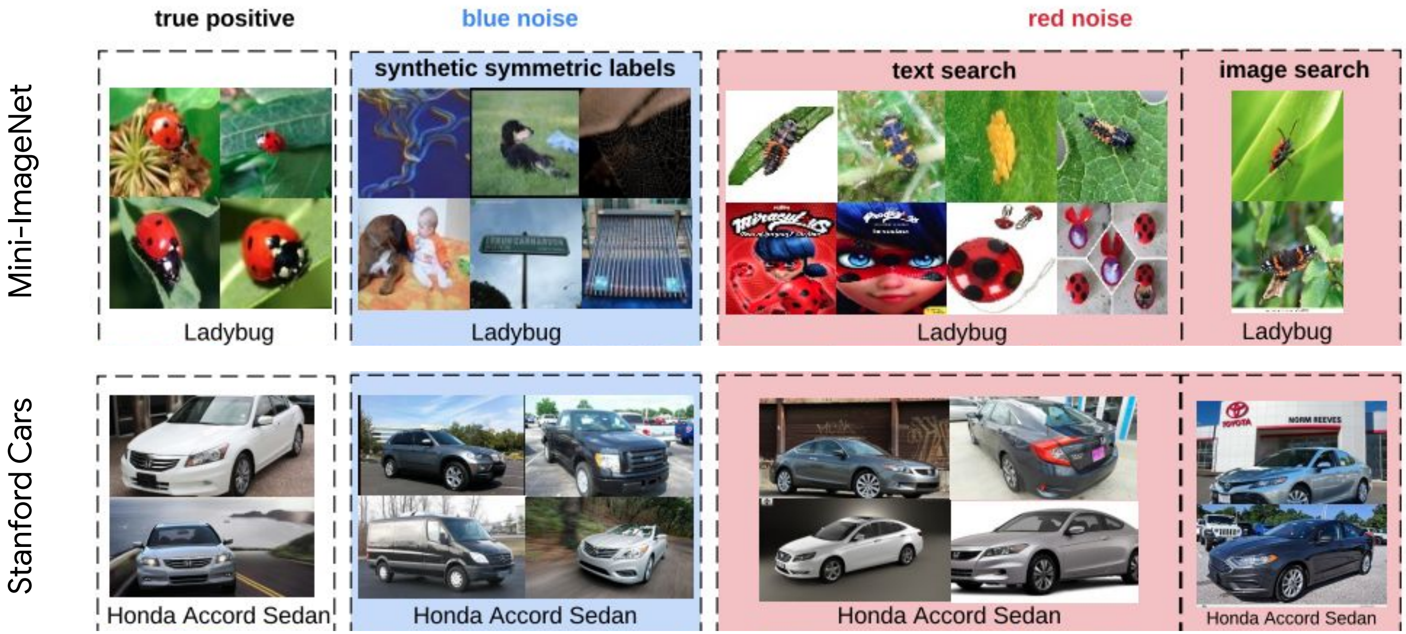
Table 1. Overview of our datasets of controlled red (web) label noise. Blue (synthetic) label noise is also included for comparison.

Dataset	#Class	Noise Source	Train Size	Val Size	Controlled Noise Levels (%)
Red Mini-ImageNet	100	image search label	50,000	5,000	0, 5, 10, 15, 20, 30, 40, 50, 60, 80
Blue Mini-ImageNet		symmetric label flipping	60,000		0, 5, 10, 15, 20, 30, 40, 50, 60, 80
Red Stanford Cars	196	image search label	8,144	8,041	0, 5, 10, 15, 20, 30, 40, 50, 60, 80
Blue Stanford Cars		symmetric label flipping	8,144		0, 5, 10, 15, 20, 30, 40, 50, 60, 80

Red noise: label noise from the web

Blue noise: synthetic label noise

Difference	Blue Noise	Red Noise
Visual & semantic similarity to true positive images	Low	High
Instance-level noise	No	Yes
Latent class vocabulary from which images are sampled	Fixed vocabulary	Open vocabulary



Contribution II: New Method

to overcome synthetic and real-world label noise

Overview

Problem: Given a noisy dataset of some unknown noise level, find a robust learning method that generalizes well on the clean test data.

Prior works: Many techniques tackle it from multiple directions, among others,

- Regularization (Azadi et al., 2016; Noh et al., 2017; etc.)
- Label cleaning (Reed et al., 2014; Goldberger, 2017; Li et al., 2017b; Veit et al., 2017; Song et al., 2019; etc.)
- Example weighting (Jiang et al., 2018; Ren et al., 2018; Shu et al., 2019; Jiang et al., 2015; Liang et al., 2016; etc.)
- Data augmentation (Zhang et al., 2018; Cheng et al., 2019)
- … …

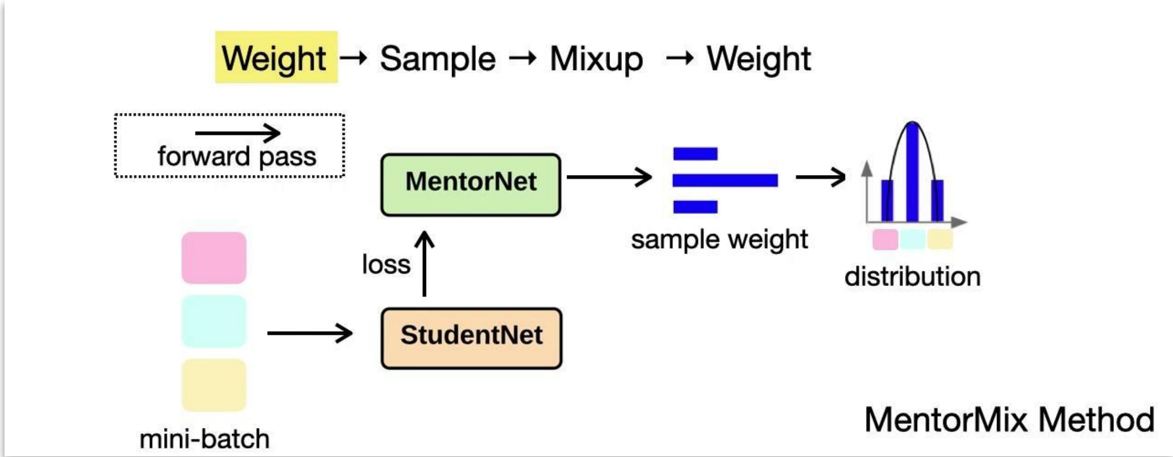
Our Method: a simple and effective method called MentorMix.

Why need yet another method? We show our method overcomes both synthetic and real-world noisy labels.

Google

Method

MentorMix is inspired by **MentorNet** (for curriculum learning) and **Mixup** (for vicinal risk minimization). It comprise four steps: weight[1], sample, mixup, and weight again[2].



1. The simplest MentorNet form is a loss thresholding function: $v_i^* = \mathbf{1}(\ell(x_i, y_i) < \gamma)$
2. We found second weighting is useful for high noise levels.

Jiang, Lu, et al. "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels." *ICML 2018*.
Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." *ICLR 2017*.

Experimental Results

MentorMix: A simple but highly effective method to overcome both synthetic and real-world noisy labels.

On our dataset

each cell is the mean of 10 different noise levels from 0% to 80%

Table 2. Peak accuracy (%) of the best trial of each method averaged across 10 noise levels. – denotes the method is failed to train.

Method	Mini-ImageNet				Stanford Cars			
	Fine-tuned		Trained from scratch		Fine-tuned		Trained from scratch	
	Blue	Red	Blue	Red	Blue	Red	Blue	Red
Vanilla	82.3±1.9	81.6±1.9	58.3±10.3	64.9±5.2	70.0±16.8	82.4±6.9	53.8±24.4	77.7±10.4
WeightDecay	81.9±1.8	81.5±1.8	—	—	72.2±17.5	84.3±6.6	—	—
Dropout	82.8±1.3	81.8±1.8	59.3±9.5	65.7±5.0	71.7±16.9	83.8±6.6	62.8±23.5	84.1±6.7
S-Model	82.3±1.8	82.0±1.9	58.7±10.2	64.6±5.1	69.7±16.8	82.4±7.1	53.9±23.5	77.6±10.2
Boostrap	83.1±1.6	82.7±1.8	60.1±9.7	65.5±4.9	71.7±16.9	82.8±6.7	55.6±23.9	78.9±9.6
Mixup	81.7±1.8	82.4±1.7	60.7±9.8	66.0±4.9	73.1±16.6	85.0±6.2	64.2±21.6	82.5±8.0
MentorNet	82.9±1.7	82.4±1.7	61.8±10.3	65.1±5.0	75.9±16.8	82.6±6.6	56.8±23.1	78.9±8.9
Ours (MentorMix)	**84.2±0.7**	**83.3±1.9**	**70.9±3.4**	**67.0±5.0**	**78.2±16.2**	**86.9±5.5**	**67.7±23.0**	**83.6±7.5**

Methods which perform well on synthetic noise may not work as well on real-world noisy labels, and vice versa. MentorMix is able to overcome both synthetic and real-world noisy labels

Experimental Results

MentorMix: A simple but highly effective method to overcome both synthetic and real-world noisy labels.

On public CIFAR (synthetic noise)

Table 3. Comparison with the state-of-the-art in terms of the validation accuracy on CIFAR-100 (top) and CIFAR-10 (bottom).

Data	Method	Noise level (%) 20	40	60	80
CIFAR100	Arazo et al. (2019)	73.7	70.1	59.5	39.5
	Zhang & Sabuncu (2018)	67.6	62.6	54.0	29.6
	MentorNet (2018)	73.5	68.5	61.2	35.5
	Mixup (2018)	73.9	66.8	58.8	40.1
	Huang et al. (2019)	74.1	69.2	39.4	-
	Ours (MentorMix)	**78.6**	**71.3**	**64.6**	**48.8**
CIFAR10	Arazo et al. (2019)	94.0	92.8	90.3	74.1
	Zhang & Sabuncu (2018)	89.7	87.6	82.7	67.9
	Lee et al. (2019)	87.1	81.8	75.4	-
	Chen et al. (2019)	89.7	-	-	52.3
	Huang et al. (2019)	92.6	90.3	46.3	-
	MentorNet (2018)	92.0	91.2	74.2	60.0
	Mixup (2018)	94.0	91.5	86.8	76.9
	Ours (MentorMix)†	**95.6**	**94.2**	**91.3**	**81.0**

On public WebVision (real-world noise)

Table 4. Comparison with the state-of-the-art on the clean validation set of ILSVRC12 and WebVision. The number outside (inside) the parentheses denotes the top-1 (top-5) classification accuracy (%). † marks the method trained using extra verification labels.

Data	Method	ILSVRC12	WebVision
Full	Lee et al. (2018)†	60.2(81.1)	68.5(86.5)
Full	Vanilla	61.7(82.4)	70.9(88.0)
Full	MentorNet (2018)†	64.2(84.8)	72.6(88.9)
Full	Guo et al. (2018)†	64.8(84.9)	72.1(89.2)
Full	Saxena et al. (2019)	—	65.7(—)
Full	Ours (MentorMix)	**67.5(87.2)**	**74.3(90.5)**
Mini	MentorNet (2018)	63.8(85.8)	—
Mini	Chen et al. (2019)	61.6(85.0)	65.2(85.3)
Mini	Ours (MentorMix)	**72.9(91.1)**	**76.0(90.2)**

Experimental Results

MentorMix: A simple but highly effective method to overcome both synthetic and real-world noisy labels.

On public CIFAR (synthetic noise)

Table 3. Comparison with the state-of-the-art in terms of the validation accuracy on CIFAR-100 (top) and CIFAR-10 (bottom).

Data	Method	Noise level (%) 20	40	60	80
CIFAR100	Arazo et al. (2019)	73.7	70.1	59.5	39.5
	Zhang & Sabuncu (2018)	67.6	62.6	54.0	29.6
	MentorNet (2018)	73.5	68.5	61.2	35.5
	Mixup (2018)	73.9	66.8	58.8	40.1
	Huang et al. (2019)	74.1	69.2	39.4	-
	Ours (MentorMix)	**78.6**	**71.3**	**64.6**	**48.8**
CIFAR10	Arazo et al. (2019)	94.0	92.8	90.3	74.1
	Zhang & Sabuncu (2018)	89.7	87.6	82.7	67.9
	Lee et al. (2019)	87.1	81.8	75.4	-
	Chen et al. (2019)	89.7	-	-	52.3
	Huang et al. (2019)	92.6	90.3	46.3	-
	MentorNet (2018)	92.0	91.2	74.2	60.0
	Mixup (2018)	94.0	91.5	86.8	76.9
	Ours (MentorMix)†	**95.6**	**94.2**	**91.3**	**81.0**

On public WebVision (real-world noise)

Table 4. Comparison with the state-of-the-art on the clean validation set of ILSVRC12 and WebVision. The number outside (inside) the parentheses denotes the top-1 (top-5) classification accuracy (%). † marks the method trained using extra verification labels.

Data	Method	ILSVRC12	WebVision
Full	Lee et al. (2018)†	60.2(81.1)	68.5(86.5)
Full	Vanilla	61.7(82.4)	70.9(88.0)
Full	MentorNet (2018)†	64.2(84.8)	72.6(88.9)
Full	Guo et al. (2018)†	64.8(84.9)	72.1(89.2)
Full	Saxena et al. (2019)	—	65.7(—)
Full	Ours (MentorMix)	**67.5(87.2)**	**74.3(90.5)**
Mini	MentorNet (2018)	63.8(85.8)	—
Mini	Chen et al. (2019)	61.6(85.0)	65.2(85.3)
Mini	Ours (MentorMix)	**72.9(91.1)**	**76.0(90.2)**

The best-published result on the WebVision benchmark!

Contribution III: New findings

on real-world label noise

Contribution III

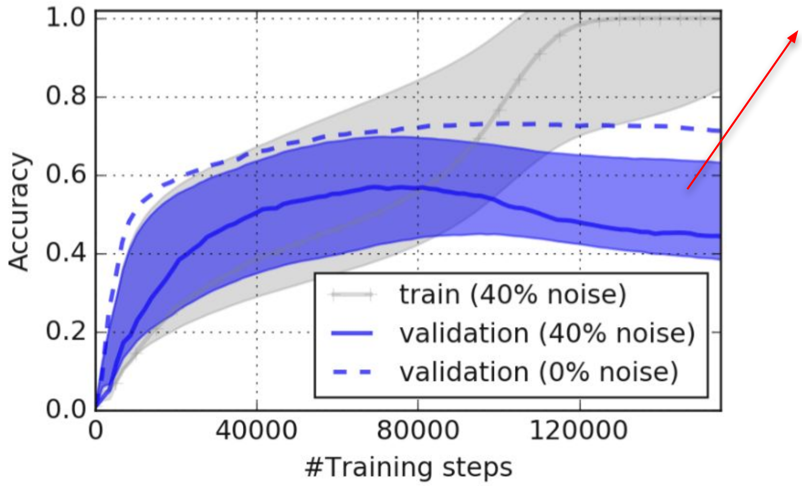
We conduct the largest study by far into understanding deep neural networks trained on noisy labels.

Our study confirms existing findings on synthetic noisy labels, and brings forward new findings that may challenge our preconception.

Blue Noise (symmetric)

(1) DNNs generalize poorly on synthetic label noise (Zhang et al., 2017).

Colored belt plots the 95% confidence interval across 10 noise levels.
Wider belt ⟶ poorer generalization



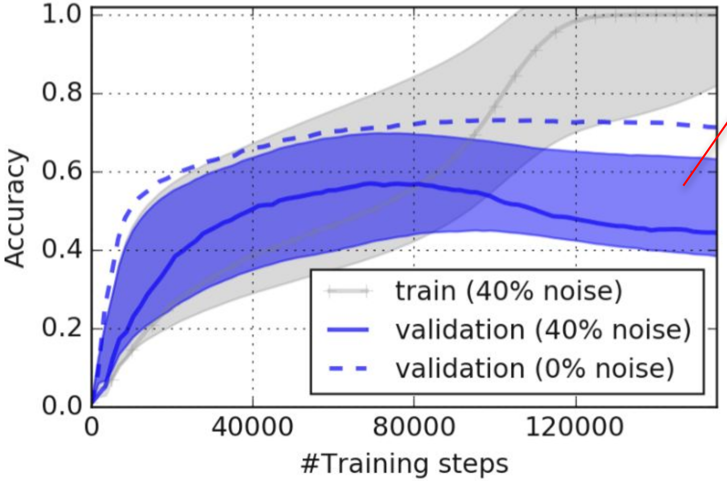
Blue Noise (symmetric)

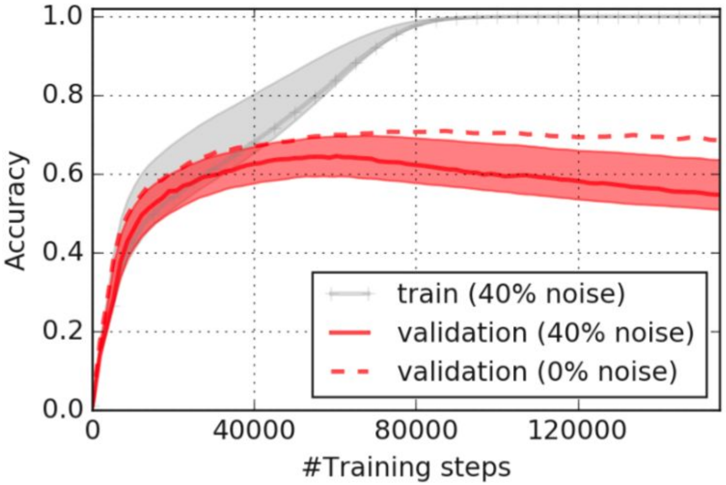
(1) DNNs generalize poorly on synthetic label noise (Zhang et al., 2017).

Red Noise (web)

DNNs generalize much better on the web label noise.

Colored belt plots the 95% confidence interval across 10 noise levels.
Wider belt ⟶ poorer generalization



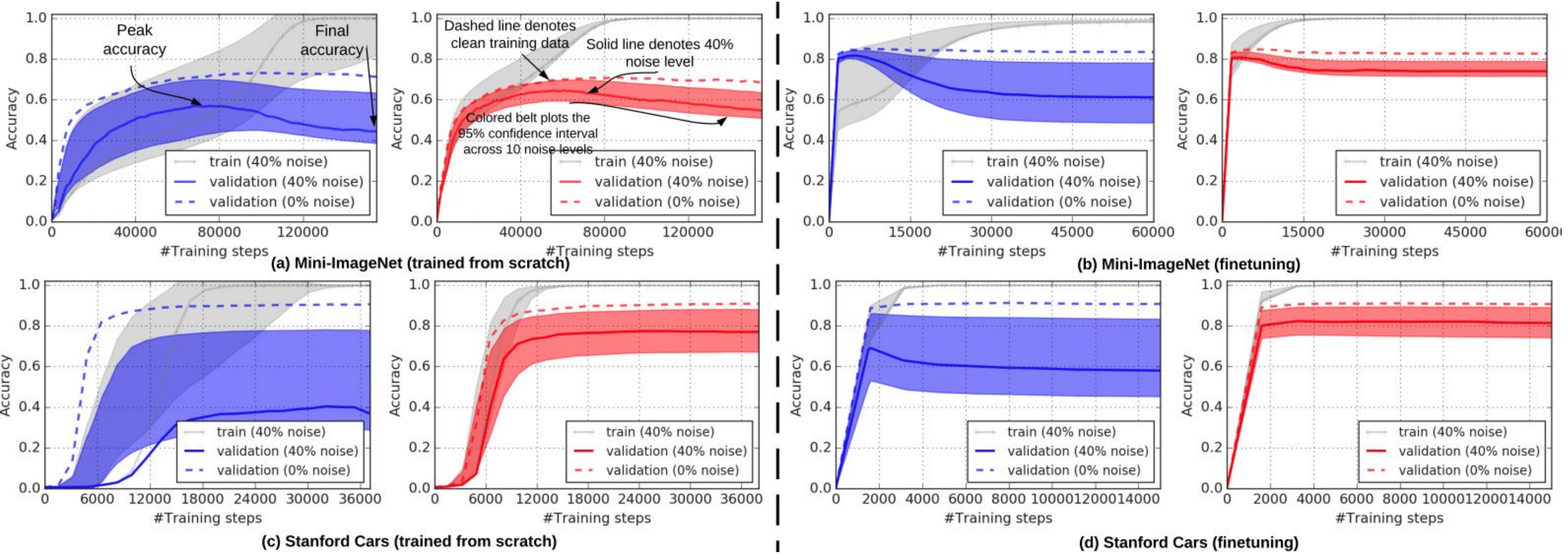


Blue Noise (symmetric)

(1) DNNs generalize poorly on synthetic label noise (Zhang et al., 2017).

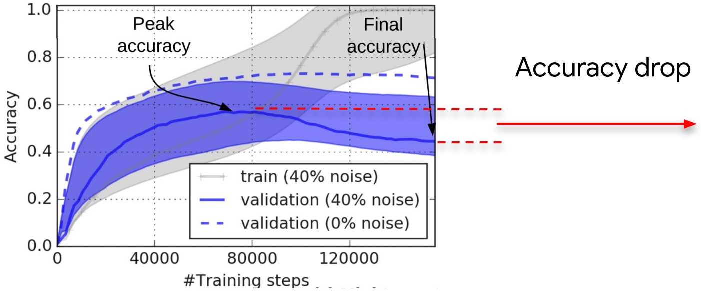
Red Noise (web)

DNNs generalize much better on the web label noise.



Blue Noise (symmetric)

(2) DNNs learn pattern first on noisy training labels (Arpit et al., 2017)

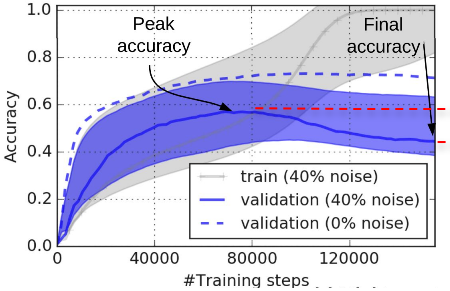


Blue Noise (symmetric)

Red Noise (web)

(2) DNNs learn pattern first on noisy training labels (Arpit et al., 2017)

DNNs may NOT learn pattern first on the web label noise



Accuracy drop

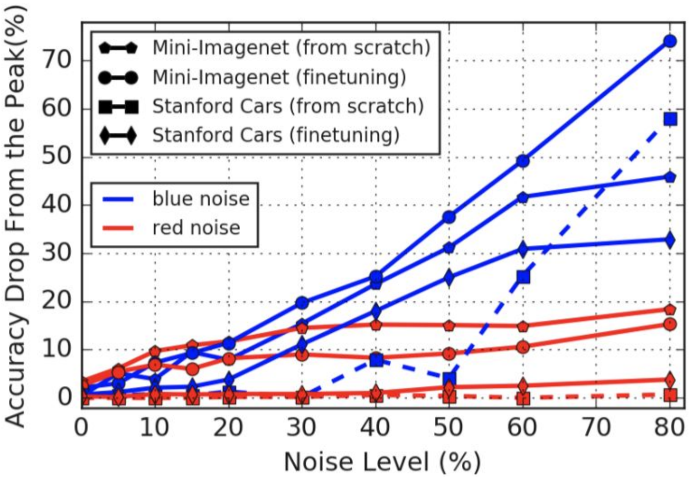


Figure 3. Performance drop from the peak accuracy at different noise levels. Colors are used to differentiate noise types.

Google

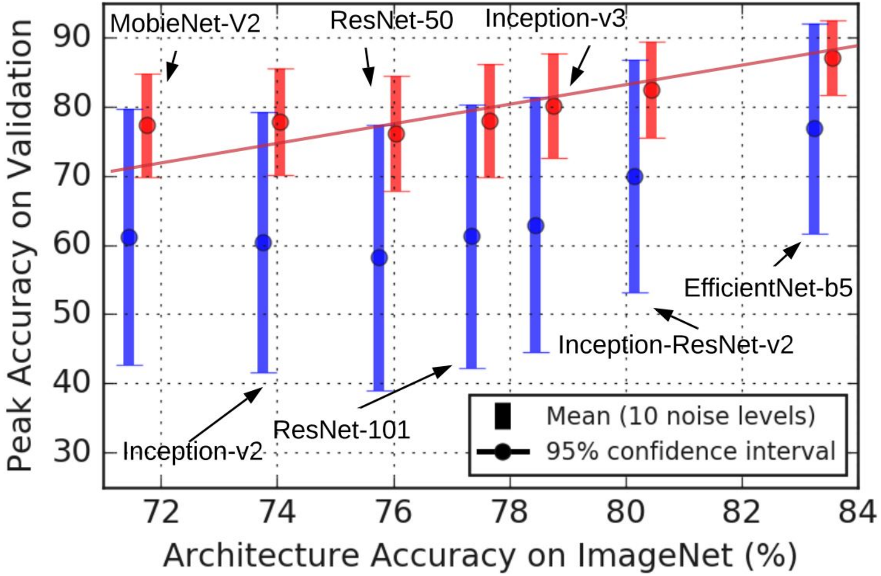
Conclusions

Clean Data

ImageNet architectures generalize on clean training labels when the networks are fine-tuned (Kornblith et al., 2019).

Blue Noise and Red Noise

It also holds on noisy labels.



ImageNet architectures generalize on noisy labels when the networks are fine-tuned.

Key takeaways:

1. We proposed:
   a. the first benchmark of real-world controlled label noise (from the web),
   b. a simple method (MentorMix) to overcome both synthetic and real-world noisy labels.

Key takeaways:

1. We proposed:
    a. the first benchmark of real-world controlled label noise (from the web),
    b. a simple method (MentorMix) to overcome both synthetic and real-world noisy labels.

2. We found:
    a. Deep networks may NOT learn patterns first but generalize much better on the real-world label noise from the web.
    b. Methods which perform well on synthetic noise may not work as well on the real-world noisy labels from the web.
    c. Advanced pretrained architectures are better at overcoming noisy labels.
    d. Further using MentorMix yields the best results.

Thanks for watching. Please find our data and code at:
http://www.lujiang.info/cnlw

Appendix

Contribution II

MentorMix consists of two key operations:
MentorNet (for curriculum learning) and **Mixup** (for vicinal risk minimization).

Algorithm 1 The proposed MentorMix method.

Input : mini-batch $\mathcal{D}_m$; two hyperparameters γ_p and α
Output : the loss of the mini-batch

For every $(\mathbf{x}_i, y_i)$ in $\mathcal{D}_m$ compute $\ell(\mathbf{x}_i, y_i)$
Set $\ell_p(\mathcal{D}_m)$ to be the γ_p-th percentile of the loss $\{\ell(\mathbf{x}_i, y_i)\}$.
$\gamma \leftarrow \mathrm{EMA}(\ell_p(\mathcal{D}_m))$ `// update the moving average`
$v_i^* \leftarrow \mathrm{MentorNet}(\ell(\mathbf{x}_i, y_i), \gamma)$ `// MentorNet weight`
Compute $P_{\mathbf{v}} = \mathrm{softmax}(\mathbf{v}^*)$, where $\mathbf{v}^* = [v_1^*, \cdots, v_{|\mathcal{D}_m|}^*]$
Stop gradient
foreach $(\mathbf{x}_i, \mathbf{y}_i)$ **do**
  Draw a sample $(\mathbf{x}_j, \mathbf{y}_j)$ with replacement from $P_{\mathbf{v}}$
  $\lambda \leftarrow Beta(\alpha, \alpha)$
  $\lambda \leftarrow v_i^* \max(\lambda, 1-\lambda) + (1 - v_i^*) \min(\lambda, 1-\lambda)$
  $\tilde{\mathbf{x}}_{ij} \leftarrow \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j$
  $\tilde{\mathbf{y}}_{ij} \leftarrow \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j$
  Compute $\ell_i = \ell(\tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{y}}_{ij})$
end
return $(1/|\mathcal{D}_m|) \sum_{i=1}^{|\mathcal{D}_m|} \ell_i$

MentorNet as importance sampling

Mixup for minimizing the vicinal risk

We use the simplest MentorNet here which is a thresholding function:

$$v_i^* = \mathbf{1}(\ell(x_i, y_i) < \gamma)$$

Jiang, Lu, et al. "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels." *ICML 2018*
Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." *ICLR 2017.*

Weight → Sample → Mixup → Weight