Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations

A PREPRINT

Jiaheng Wei* UC Santa Cruz

Zhaowei Zhu* UC Santa Cruz

Hao Cheng UC Santa Cruz

Tongliang Liu University of Sydney

Gang Niu RIKEN

Yang Liu † UC Santa Cruz

陈明猜

2021/12/15

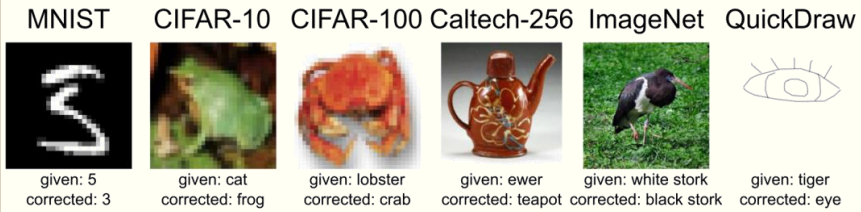
Standard settings of learning with noisy labels

Definition: The presence of noisy labels in these datasets introduces two problems. How can examples with label errors be identified, and how can learning be done well in spite of noisy labels.

Real-world applications: (1). Labelers may lack the necessary experience, (2). data can be too complex to be correctly classified, even for the experts, (3). adversarial poisoning purposes.

Common settings :

- A typical approach for modeling label noise assumes that the corruption process is conditionally independent of data features when the true label is given.
- Random label noise is class-conditional and the flip probability depends on the class.
- For more realistic noise modeling, the corruption probability is assumed to be dependent on both the data features and class labels.



label errors are pervasive across popular datasets used in most machine learning research[1].

Tasks: Symmetric and asymmetric noise, ranging from 20%-90%.

- Symmetric noise is generated by randomly replacing the labels for a percentage of the training data with all possible labels.
- Asymmetric noise is designed to mimic the structure of real-world label noise, where labels are only replaced by similar classes (e.g. deer-horse, dog-cat).

[1]. *https://l7.curtisnorthcutt.com/label-errors*

[2]. *Song, Hwanjun, et al. "Learning from noisy labels with deep neural networks: A survey." arXiv preprint arXiv:2007.08199 (2020).*

Existing efforts suffer from two caveats:

Motivation: the lack of ground-truth verification makes it hard to study the property and treatment of real-world label noise:

- Complex task (High-resolution): when learning with large-scale and relative high-resolution data, the complex data pattern, various augmentation strategies, the use of extra train or clean data, different computation power (for hyper-parameter tuning such as batch-size, learning rate, etc) jointly contribute to the model performance and then result in unfair comparison.
- Missing clean labels: the lack of clean labels for verification in most existed noisy-label datasets makes the evaluation of robust methods intractable.
- Interventions: human interventions in data generation and non-representative data collection process might disturb the original noisy-label pattern.

This work presents two new benchmark datasets, which we name as CIFAR-10N, CIFAR-100N, equipping the training datasets of CIFAR-10 and CIFAR-100 with human-annotated real-world noisy labels that we collect from Amazon Mechanical Turk.

Data collection

We randomly split the training dataset of CIFAR-10 without replacement into ten batches. In the Mturk interface, each batch contains 500 HITs with 10 images per HIT. The training images and test dataset remain unchanged. Each HIT is then randomly assigned to three independent workers. Workers gain base reward $0.03 after submitting the answers of each HIT. We reward workers with bonus salary if the worker contributes more HITs than the averaged number of submissions. We did not make use of any ground-truth clean labels to approve or reject submissions. We only block and reject workers who submit answers with fixed/regular distribution patterns.

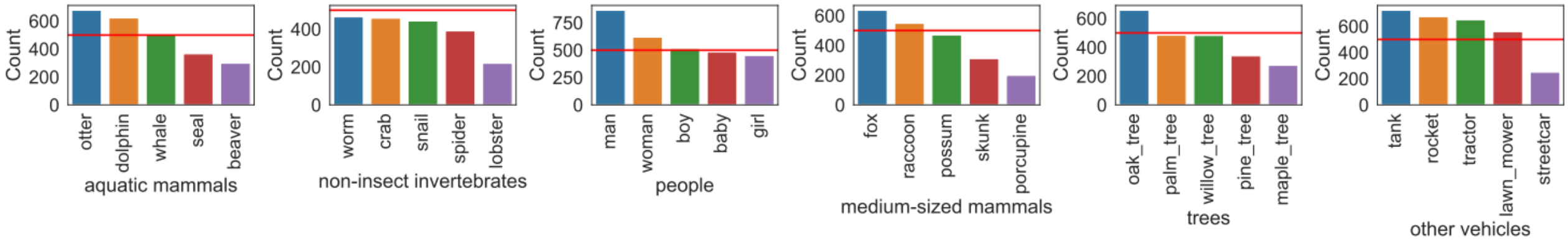
Dataset statistics For CIFAR-10N dataset, each training image contains one clean label and three human annotated labels. We provide five noisy-label sets as follows.

- **Aggregate:** aggregation of three noisy labels by majority voting. If the submitted three labels are different for an image, the aggregated label will be randomly selected among the three labels.
- **Random** i $(i \in \{1, 2, 3\})$**:** the i-th submitted label for each image. Note our collection procedure ensures that one image cannot be repeatedly labeled by the same worker.
- **Worst:** dataset with the highest noise rate. For each image, if there exist any wrongly annotated labels in three noisy labels, the worst label is randomly selected from wrong labels. Otherwise, the worst label is equal to the clean label.

60.27% of the training images have received unanimous label from three independent labelers. The noise rates of prepared five noisy label sets are **9.03% (Aggregate), 17.23% (Random 1), 18.12% (Random 2), 17.64% (Random 3)** and **40.21% (Worst)**. A complete dataset comparison among existing benchmarks and ours are given in Table 1. We

Preliminary observations on CIFAR-10N, CIFAR-100N

Observation 1: Imbalanced annotations Our first observation is the imbalanced contribution of labels. Note that while the number of images are the same for each clean label, across all the five noisy label sets of CIFAR-10N, we observe that human annotators have different preferences for similar classes. For instance, they are more likely to annotate an image to be an automobile rather than the truck, to be the horse rather than the deer (see Figure 10 in the Appendix). The aggregated labels appear more frequently in automobile and ship, and less frequently in deer and cat. This gap of frequency becomes more clear in the worst label set. In CIFAR-100N, human annotators annotate frequently on classes which are outside of the clean-coarse, i.e., 25% noisy labels fall outside of the super-class and 15% inside the super-class. And the phenomenon of imbalanced annotations also appears substantially as shown in Figure 1, which presents the distribution of noisy labels for each selected fine class. "Man" appears $\geq$ 750 times, while "Streetcar" only has $\approx$ 200 annotations.



Preliminary observations on CIFAR-10N, CIFAR-100N

Observation 2: Noisy label flips to similar features In CIFAR-100N, most fine classes are more likely to be mislabeled into less than four fine classes. In Figure 2, we show top three wrongly annotated fine labels for several fine classes that have a relative large noise rate. Due to the low-resolution of images, a number of noisy labels are annotated in pair of classes, i.e, $\approx 20\%$ of "snake" and "worm" images are mislabeled between each other, similarly for "cockroack"-"beetle", "fox"-"wolve", etc. While some other noisy labels are more frequently annotated within more classes, such as "boy"-"baby"-"girl"-"man", "shark"-"whale"-"dolphin"-"trout", etc, which share similar features.

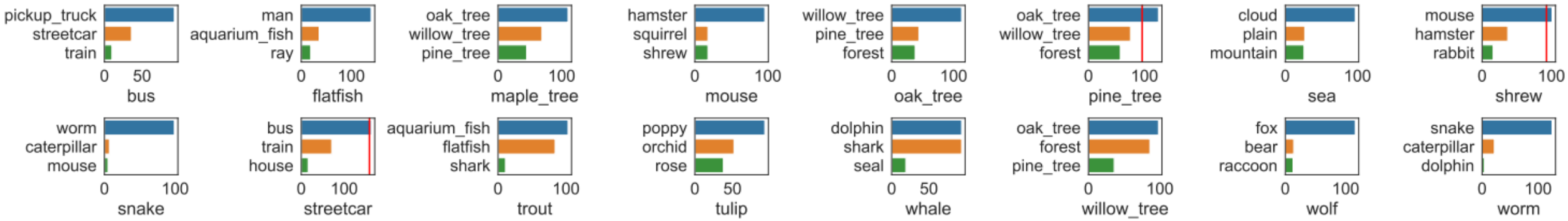
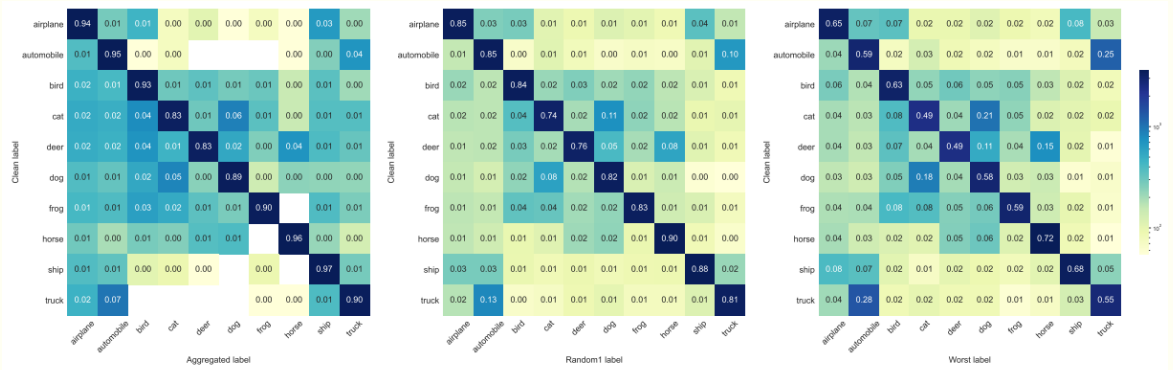


Figure 2: Top 3 wrongly annotated fine labels in selected fine classes. For "pine tree", "shrew", "streetcar", the dominant class is the **wrong** class. The corresponding number of correct annotations are highlighted with red lines.

Preliminary observations on CIFAR-10N, CIFAR-100N

Observation 3: The pattern of noise transition matrices In the class-dependent label noise setting, suppose the label noise is conditional independent of the feature, the noise transition matrices of CIFAR-10N and CIFAR-100N are best described by a mixture of symmetric and asymmetric T. For CIFAR-10N, we heatmap the aggregated noisy labels, random1 noisy labels and worst noise labels w.r.t. the clean labels. In Figure 3, the three noisy label sets share a common pattern: the clean label flips into one or more similar classes more often. The remaining classes largely follow the symmetric noise model with a low noise rate. For example, "truck" and "automobile" flip between each other more often ($\approx 25\% - 30\%$ percentage), which is much larger than that of all other classes. Besides, in the central area of each transition matrix, it is quite obvious that the clean label of animal classes flips more often to other animals. Similar observations hold in CIFAR-100N, where each class flips to a few misleading classes with much higher probability than that of remaining ones (see Figure 13 in the Appendix). Apparently, current synthetic class-dependent noisy settings are not as complex as the real-world human annotated label noise.



(a) Aggregated noisy labels (b) Random noisy labels (c) Worst noisy labels

Figure 3: Transition matrix of CIFAR-10N noisy labels (color bar is log-norm transformed).

Preliminary observations on CIFAR-10N, CIFAR-100N

Observation 4: Label noise: bad news or good news? During the label collection, there exist a non-negligible amount of the wrongly annotated classes that indeed co-exist in the corresponding images. In other words, training images of CIFAR-100 may contain **multiple labels** rather than a single one. We select several exemplary training images of CIFAR-100 where multiple labels appear (in Figure 4). The annotated class also appears in the corresponding image while is deemed as a wrong annotation by referring to the officially provided clean label. The most frequent case is best described by the scenario where a man holding a flatfish in hands. The clean label usually comes to "flatfish", while human annotators are more likely to categorize these images into "man". We conjecture that with the increasing label dimension, the phenomenon of multiple clean labels might be a more common issue. We leave more explorations for the future work.

Figure 4: Exemplary CIFAR-100 training images with multiple labels. The text below each picture denotes the CIFAR-100 clean label (first row) and the human annotated noisy label (second row).

Preliminary observations on CIFAR-10N, CIFAR-100N

Table 4: Performance gap between human noise and class-dependent noise: test accuracy (trained on synthetic noise) - test accuracy (trained on human noise). Negative gaps are highlighted in red.

Method	CIFAR-10 Gap					CIFAR-100 Gap
	Aggregate	*Random 1*	*Random 2*	*Random 3*	*Worst*	*Noisy*
CE (Standard)	4.35	6.01	4.50	5.82	9.00	1.20
Forward T [Patrini et al., 2017]	4.30	4.82	4.86	4.27	7.08	-0.14
Co-teaching+ [Yu et al., 2019]	0.89	0.92	0.86	1.05	2.63	-0.61
Peer Loss [Liu and Guo, 2020]	1.90	2.42	2.74	1.95	4.67	-0.85
ELR [Liu et al., 2020]	-0.78	-0.81	-0.97	-0.51	-1.64	1.05
F-Div [Wei and Liu, 2020]	0.72	1.62	1.33	1.65	4.14	1.31
Divide-Mix [Li et al., 2020]	1.39	0.46	-0.38	0.08	0.99	0.65
Negative-LS [Wei et al., 2021]	0.77	1.31	1.08	1.36	4.00	1.26
JoCoR [Wei et al., 2020]	0.35	0.78	0.68	1.01	2.43	-0.48
CORES^2 [Cheng et al., 2021]	1.49	1.69	1.52	1.66	1.67	-0.72
CAL [Zhu et al., 2021b]	0.25	0.04	0.04	0.09	0.44	0.47

Preliminary observations on CIFAR-10N, CIFAR-100N

In Figure 8, we train CE loss with a ResNet-34 [He et al., 2016] neural network on three noisy label sets of CIFAR-10N: aggre-label (left column), random-label1 (middle column) and worst-label (right column). While visualizing the memorization ($\eta = 0.95$) on training samples, we split the train data into two parts: images with clean labels (the annotation matches the clean label) and wrong labels (the rest). Denote f_H and f_S as models trained on human noisy labels and synthetic noisy labels (with the same T), respectively. f_S gradually memorizes both correct and wrong predictions, while f_H only memorizes correct predictions on clean labels and wrong predictions on wrong labels.

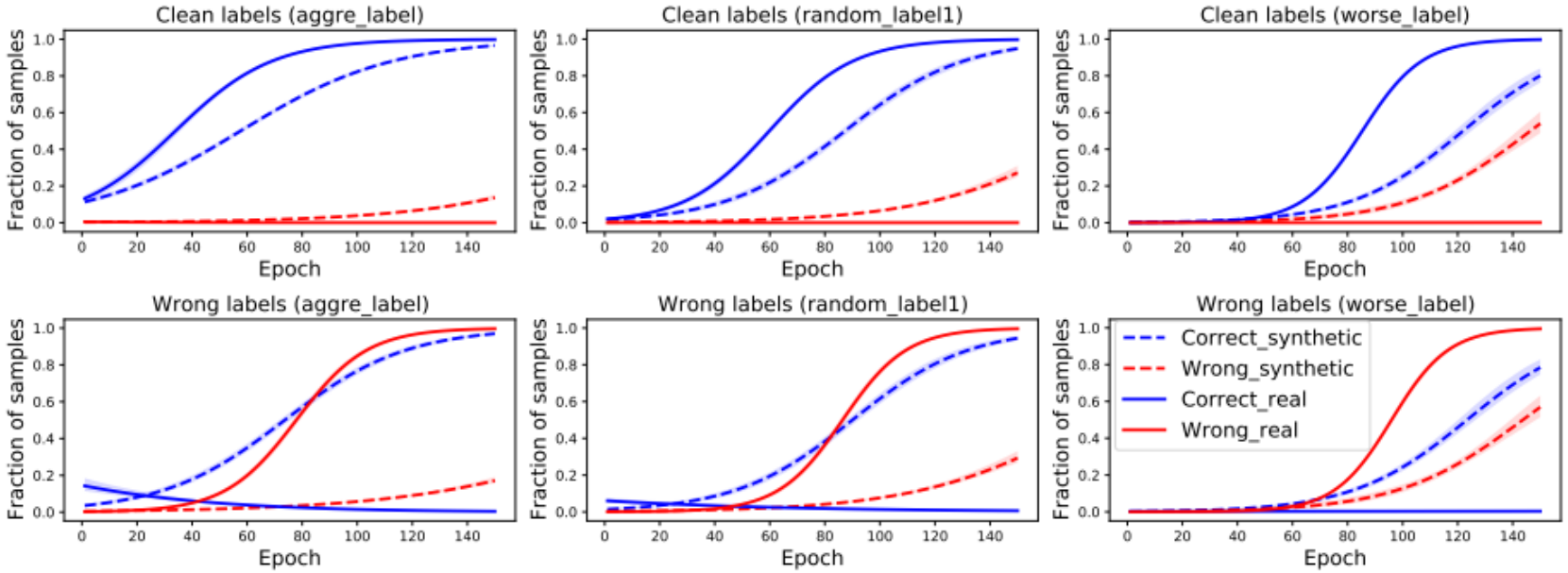
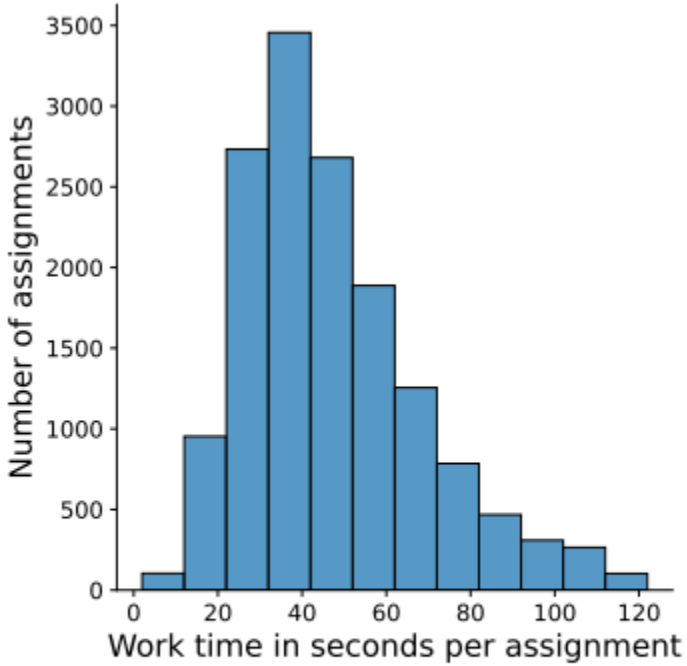
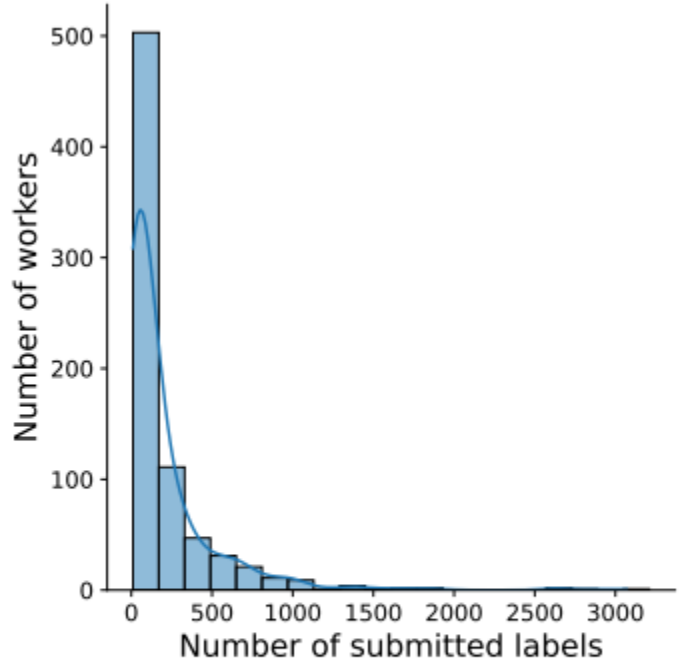


Figure 8: Memorization of clean and wrongs labels on CIFAR-10N and synthetic noise with same T: red line denotes the percentage of memorized (wrongly predicted) samples, blue line denotes that of correctly predicted ones. f_S: dashed line (- -), f_M: solid line (—).

Preliminary observations on CIFAR-10N, CIFAR-100N



(a) Distribution of work time in seconds per HIT.



(b) Distribution of the amount of submitted labels.

Figure 9: The behaviors of workers in the collection of CIFAR-10N.

Thanks