Learning from Training Dynamics: Identifying Mislabeled Data beyond Manually Designed Features

Qingrui Jia1,2*, Xuhong Li3,4, Lei Yu5,6,7, Jiang Bian3, Penghao Zhao2, Shupeng Li3, Haoyi Xiong3,4, Dejing Dou4

1 Sino-French Engineer School, Beihang University, Beijing, China
2 Baidu Inc. Beijing, China
3 Beihang Hangzhou Innovation Institute Yuhang, Hangzhou, China
4 BCG Greater China, Beijing, China

Abstract

While mislabeled or ambiguously-labeled samples in the training set could negatively affect the performance of deep models, diagnosing the dataset and identifying mislabeled samples helps to improve the generalization power. Training dynamics, i.e., the traces left by iterations of optimization algorithms, have recently been proved to be effective to localize mislabeled samples with hand-crafted features. In this paper, beyond manually designed features, we introduce a novel learning-based solution, leveraging a noise detector, instantiated by an LSTM network, which learns to predict whether a sample was mislabeled using the raw training dynamics as input. Specifically, the proposed method trains the noise detector in a supervised manner using the dataset with synthesized label noises and can adapt to various datasets (either naturally or synthesized label-noised) without retraining. We conduct extensive experiments to evaluate the proposed method. We train the noise detector based on the synthesized label-noised CIFAR dataset and test such noise detector on Tiny ImageNet, CUB-200, Caltech-256, WebVision and Clothing1M. Results show that the proposed method precisely detects mislabeled samples on various datasets without further adaptation, and outperforms state-of-the-art methods. Besides, more experiments demonstrate that the mislabel identification can guide a label correction, namely data debugging, providing orthogonal improvements of algorithm-centric state-of-the-art techniques from the data aspect.

Introduction

Deep learning models have achieved remarkable performances across various tasks (Salakhutdinov 2014) due to the rapid development of paralleled devices, advanced network architectures and learning algorithms, as well as the large dataset collection and high-quality annotations (Deng et al. 2009a; Li et al. 2017a; Zhou et al. 2018). However, the difficulty and expensiveness of acquiring high-quality labels inevitably leave an amount of label noises on the datasets. In many real-world settings, training samples and their labels are collected through proxy variables or web scraping (Chen 2009a; Li et al. 2017a; Zhou et al. 2018). However, the aforementioned distinguishable features are usually extracted manually from the training dynamics with prior knowledge and domain expertise, and it is not always fitting well and being efficient. Tracing back the history of machine learning, the deep learning methods are much more effective than traditional learning with handcraft features, by learning features from raw data and solving the learning tasks in end-to-end fashions. In light of the evolution of machine learning, we propose a novel solution through learning from training dynamics to identify mislabeled samples. Beyond manually designed features, our
proposed method trains a noise detector, instanced by an LSTM network (Hochreiter and Schmidhuber 1997), on the raw training dynamics, with the objective of learning distinguishable features to solve the binary classification task, i.e., separating mislabeled samples from clean ones.

The most efficient and effective way to train the noise detector is the supervised learning, where the binary annotations whether samples being mislabeled or not are required. However, this kind of information is usually not provided by public datasets. To cope with this issue, we purposely wrong-set the labels of certain samples from an (almost) clean dataset, and conduct the standard training process on it. In this way, the raw training dynamics can be trained with the supervision of the synthesized label noises.

We carry out three types of evaluation experiments to validate the effectiveness of our method in identifying mislabeled samples. The results show that (1) The noise detector trained on the synthesized noisy CIFAR-100 can precisely identify the mislabels on other synthesized and real-world noisy datasets without further adaptation or retraining, showing a good transfer ability; (2) Compared to the state-of-the-art approaches from the data aspect, our method outperforms with a clear margin; and (3) Combining with methods from the algorithm aspect, our method provides an orthogonal improvement over the current methods.

In summary, our main contributions are the following:

- To identify the mislabeled samples, we propose a novel supervised-learning-based solution to train an LSTM-instanced label noise detector based on the synthesized noisy dataset and its training dynamics.
- The noise detector trained on a noised CIFAR-100 is an off-the-shelf detector for identifying the mislabeled samples, i.e., it can accurately find label noises without any further adaptation or retraining on other noised datasets, including Tiny ImageNet, CUB-200, Caltech-256, WebVision and Clothing 100K, where the latter two are real-world noisy datasets.
- On three types of evaluation experiments, our proposed method outperforms all the existing methods of identifying the mislabeled samples. Analyses and discussions are also provided to validate the effectiveness of our method.
- Combining with the learning-with-noisy-labels algorithms, our proposed method further boosts test accuracies of the models trained on noisy datasets. This also proves that the improvements from the data aspect can be orthogonal to those from the algorithm aspect.

### Related Work

Learning with label noise has achieved remarkable success. We review a good amount of prior works most relevant to ours and categorize them into two aspects, algorithms or data. For a complete review, we refer to Song et al. (2020); Algan and Ulusoy (2021); Mazumder et al. (2022).

**From Algorithms Aspect.** Tremendous approaches provide algorithms-centric perspectives and stay in the conventional mainstream for robust learning in presence of mislabeled samples. Some approaches proposed to estimate the noise transition matrix via pretrained model (Patrini et al. 2017), clean validation set (Hendrycks et al. 2018) or expectation-maximization (EM) algorithm (Goldberger and Ben-Reuven 2017). A number of methods propose to improve the noise immunity by designing robust loss functions (Ghosh, Kumar, and Sastry 2017; Wang et al. 2019), or modify loss or probabilities (Arazo et al. 2019; Yao et al. 2019; Reed et al. 2014) to compensate for negative impact owing to noisy labels. Some alternative methodologies aim to refurbish the corrupted labels, by characterizing the distribution of noisy labels (Xiao et al. 2015; Vahdat 2017; Li et al. 2019; Lee et al. 2018) or in a meta-learning manner (Li et al. 2017c; Zheng, Awadallah, and Dumas 2021).

Remarkably, the state-of-the-art methods are usually hybrid, integrated of above components. For examples, DivideMix (Li, Socher, and Hoi 2020) leverages Gaussian mixture model to distinguish mislabel ones, then regards them as unlabeled samples, and trains the model in a semi-supervised learning manner. Nishi et al. (2021) improves the DivideMix framework by applying different data augmentation strategies for finding noisy labels and learning models. Through curriculum learning (Bengio et al. 2009), RoCL (Zhou, Wang, and Bilmes 2021) first learns from clean data and then gradually involves noisy-labeled data corrected by the predictions from an ensemble based model.

In brief, algorithm-centric training with label noises requires a modification of the optimization algorithm. The proposed approach in this paper, from a data-centric aspect, does not modify the optimization procedure but filters the label noises in the first step. By all means, algorithm-centric methods are used to improve the optimization procedures, orthogonal to data-centric methods introduced as follows.

**From Data Aspect.** Fundamentally different from the...
above methods in learning with label noise, a lot of approaches focus on data-centric amelioration and construct higher quality datasets, decoupled from training procedures. A straightforward thought of harnessing mislabeled samples is excluding them from the training data and then retraining on the remaining samples (Brody and Friedl 1999). Therefore, various studies endeavor to detect mislabeled samples or sort clean samples. One of the prevalent methods is unsupervised outliers removal (Liu, Hua, and Smith 2014; Xia et al. 2015), which is based on the assumption that outliers are incorrectly labeled.

Recently, instead of simply checking the final predictions, more works analyze the training dynamics, i.e., the traces left by the steps during SGD. For instance, Toneva et al. (2019) defines a forgetting event using training dynamics, and samples prone to be forgotten with high frequency probably correspond to be mislabeled. Swayamdipta et al. (2020) builds a data map with training dynamics, then marks off a region composed of instances that the model finds hard to learn in dataset mapping, which might be labeling errors. Pleiss et al. (2020) proffers a statistic called area under margin (AUM) and introduces threshold samples to separate clean and mislabeled samples. In addition, other solutions filter noisy data from perspective of influence functions (Koh and Liang 2017) for identifying training points that exert the most influence on the model. Ghorbani and Zou (2019) introduced data Shapley to quantify the value of the mislabeled samples. As shown in Figure 1, the mean of the mislabeled samples is more complicated to fit than the clean samples. However, with these visually visible differences, it is still challenging to identify the mislabeled samples precisely. Designing more expert features will surely be helpful, but instead, we propose to benefit from the advantage of deep learning models to learn good features for classifying them.

The Proposed Approach
Training dynamics are distinguishable features to identify the mislabeled samples. As shown in Figure 1, the mean of the given-label probability over noisy samples is lower than the one over clean samples; but the variance is higher, especially at the latter half of the training, indicating that the mislabeled samples are more complicated to fit than the clean samples. However, with these visually visible differences, it is still challenging to identify the mislabeled samples precisely. Designing more expert features will surely be helpful, but instead, we propose to benefit from the advantage of deep learning models to learn good features for classifying the mislabeled samples from clean ones.

Definition of Training Dynamics
We give notations and definitions within the context of classification learning problems without loss of generality.

In the context of SGD or its variant, we define the training dynamics as the probability values of all training samples obtained by the model at each training epoch, denoted as \( L \). Specifically, \( L(i, t) \) is the probabilities of the \( i \)-th sample at the \( t \)-th training epoch. At each epoch, the model would pass through forward and backward propagation for each training sample, with the probability values as intermediate results that can be recorded during the training process without additional computations\(^2\). Note that \( L(i, t) \) is a vector of \( C \) elements, where \( C \) is the number of classes, and we note \( L_c(i, t) \) as the probability of the \( c \)-th class. Noting the deep model as \( f \), then we have
\[
L(i, t) = f^{(t)}(x_i),
\]
where \( x_i \) is the input data. For computing gradients and updating the model’s parameters, the label \( y_i \) is needed.

Supervised Learning from Training Dynamics
We present the supervised learning from training dynamics in three steps as following.

**Step 1.** Given an (almost) clean dataset \( D \), we first generate the label noises and obtain the noised dataset \( D^N \). Specifically, \( D \) contains the data instances \((x_i, y_i)\) while \( D^N \) contains \((x_i, y_i^N, k_i)\). The additional term \( k_i \) indicates that the sample is intentionally mislabeled \((k_i = 1)\) or not \((k_i = 0)\). The design of the noise generator follows the previous works (Pleiss et al. 2020; Northcutt, Jiang, and Chuang 2021), where the symmetric and asymmetric noises are used in our experiments. Following the noise generator, the new values of \( y_i^N \) will overwrite the original information in \( D \), and \( k_i \) can be simply obtained by
\[
k_i = \begin{cases} 1, & \text{if } y_i^N \neq y_i; \\ 0, & \text{if } y_i^N = y_i. \end{cases}
\]

The noised datasets \( D^N \) are only used to train the LSTM-instanced noise detector, denoted as \( g \). The trained noise detector \( g \) (as introduced in Eq (3)) is directly applicable to other real-world datasets.

As for the (almost) clean dataset \( D \), though some famous datasets, such as CIFAR-10 and ImageNet, contain a small ratio (\(< 3\%)\) of mislabeled samples (Northcutt, Athalye, and Mueller 2021), our method is well-adapted to a mild quantification of underlying noises in the dataset. This is shown in Subsection .

**Step 2.** We train the deep model \( f \) on \( D^N \) using SGD or one of its variants, and record the training dynamics \( L^N \).

\(^2\)For cases when the value of a certain sample at some epoch is missing, interpolations across training epochs can be used.

![Figure 2: Supervised Learning from Training Dynamics.](image-url)
supervise the noise detector $g$ taking the training dynamics $L^N$ as feature inputs and the label noise information $K = \{k_i\}$ as supervision labels. Specifically, the noise detector is optimized following

$$\hat{g} = \arg \min_{g} \sum_i L(g(L^N(i), k_i)),$$

where $L(\cdot, \cdot)$ is the binary cross entropy loss function, and the optimization process is guided by an AdamW optimizer (Kingma and Ba 2015; Loshchilov and Hutter 2019). Moreover, to deal with the epoch-wise training dynamics, $g$ is instanced by an LSTM model which is a popular choice for time-series sequences. In practice, for efficiency and generality, we only take the given-label probability as input, i.e., $L^N(i)$. Involving other classes would increase the parameters and computations in the LSTM model and the efficiency would be encumbered for datasets with over hundreds of classes.

**Step 3.** Finally, we conduct the standard training process on the clean dataset $D'$, and record the training dynamics $L$. The previously trained $\hat{g}$ is then used to estimate the probability of the label $y_i$ being mislabeled according to the training dynamics $L_y(i)$, i.e., computing $\hat{g}(L_y(i))$.

The procedure of the proposed approach is summarized in Algorithm 1. We introduce three extensions of this proposed approach in the following subsection to show the applicability and robustness of our approach for practical usages.

**Characteristic Analysis**

The proposed approach effectively identifies the mislabeled samples, demonstrated by the experiments in Section 4. Here we provide the characteristic analysis to show the transferability, robustness and applicability of our method.

**Off-The-Shelf Noise Detector and Transferability.** A trained $\hat{g}$ on one dataset is capable of identifying the mislabeled samples on other datasets. Once the noise detector $\hat{g}$ is trained, Step 3 can process standalone for any dataset $D'$, which proves that $\hat{g}$ is generalizable and has learned some common features for identifying mislabeled samples.

In Section 4, we show that $\hat{g}$ trained from the noised CIFAR-100 works quite well on other (supposedly-) clean datasets and noisy datasets. Note that further fine-tuning of $\hat{g}$ on $D'$, following the exact procedure described in Section 4, can always provide further improvements.

**Robustness against Underlying Noises.** In real-world scenarios, the dataset $D$ may contain a small or mild ratio of noises. This would introduce noises in the training process of $g$. We simulate this scenario by manually changing the labels of the (supposedly-) clean dataset and setting it to $D$. Then we process the three steps described previously and use $\hat{g}$ to detect the mislabeled samples. According to our analytical experiments, the noise detector $\hat{g}$ is still effective in this setting, up to a large ratio of noises. This relaxes the constraint on the label quality of $D$.

**Usability of Super-Classes Noises.** During the dataset annotation, human experts usually mislabel the samples into similar categories but rarely to those with very different categories. This pattern of mislabeling is easy to mimic, especially in CIFAR-100, which has 20 super-classes, followed by 5 finer categories. We follow this setting in Section 4 to generate the noisy labels within the super-class, instead of across all 100 classes. Experiments show that our method is still capable of accurately identifying the mislabeled samples under this super-classes noise setting.

**Experiments**

On synthesized and real-world noisy datasets, three types of evaluations are conducted and presented in Section 4, and , respectively: (1) On synthesized datasets, where the mislabeled samples are known, the precision and recall can be directly calculated. (2) Following the procedure of previous works (Brodley and Friedl 1999; Pleiss et al. 2020), the evaluations are performed by checking the performance of the supervised learning with/without (by excluding) mislabeled samples in the datasets. A higher test accuracy trained on the dataset without mislabeled samples indicates a better identification of mislabeled samples. (3) Similar to the data debugging evaluation scheme (Mazumder et al. 2022), the evaluation procedure first identifies the label errors and then corrects them with the true labels or pseudo labels generated by the model in (2), i.e., trained excluding label errors. Afterward, the corrected dataset is incorporated with the state-of-the-art learning algorithms, e.g., DivideMix (Li, Socher, and Hoi 2020) and DM-AugDesc (Nishi et al. 2021), to enjoy the orthogonal improvements from the data aspect.

**Identify Mislabeled Samples**

This section validates our approach on the datasets with synthesized label noises. We follow the commonly used noise distributions (Northcutt, Jiang, and Chuang 2021; Pleiss et al. 2020) to generate label noises at different ratios and change the labels of the training set with symmetry (or asymmetry) noises but leave the testing set unchanged.

Here we mainly compare our method with AUM (Pleiss et al. 2020), one of the most effective approaches that also use the training dynamics to identify mislabeled samples.

**Experiment Setups.** Evaluations are conducted on
four datasets with synthesized noisy labels: CIFAR-10/100 (Krizhevsky and Hinton 2009), CUB-200-2011 (Wah et al. 2011) and Caltech256 (Griffin, Holub, and Perona 2007). Evaluation metrics include mAP, ROC AUC scores and Precision@95, where Precision@95 represents the precision when the recall reach 95%, which can be directly measured because of the available ground-truth.

Training dynamics are insensitive to the deep model architectures (Toneva et al. 2019; Pleiss et al. 2020). We therefore choose ResNets as the deep model \( f \) for the higher testing accuracy in original datasets, while we have also tested VGGs (Simonyan and Zisserman 2015), which show similar results in preliminary experiments. Concerning the noise detector, for efficiency, an LSTM of two hidden layers with 64 dimensions for each is used. All the noise detectors share the same architecture, allowing the possible transfer between datasets.

**Results.** Following the steps in Section 3, we supervise the noise detector instantiated by an LSTM model on the noised CIFAR-10/100 datasets with the corrupted labels. Our method is competitive with AUM without threshold strategy. Figure 3 depicts the mAP, AUC and Precision@95 scores for identifying mislabeled samples from the noised CIFAR-10/100 datasets. Our procedure excels AUM in both symmetric and asymmetric noise settings, the later is confirmed in Figure 4.

Additional experiments have been conducted on the noised CUB-200-2011 and Caltech-256. As shown in Figure 5, our approach benefits from the well-trained noise detector \( \hat{g} \) on CIFAR-100 with 30% noise ratio. Then we fine-tune \( \hat{g} \) on the noised CUB-200-2011 and Caltech-256, respectively, to further improve the performance of the noise detection with a clear margin.

**Retrain after Excluding Mislabeled Samples.**

Mislabeled samples are harmful to the model (Arora et al. 2019; Zhang et al. 2021), and removing them should be beneficial for its accuracy and generalizability. To validate the effectiveness of identification, we remove samples predicted to be mislabeled using the proposed approach and then retrain the deep model. Based on the testing accuracy after excluding samples, we are able to measure the effectiveness of mislabeled sample identification. Within this setting, we compare our method against several existing ones, introduced in the following paragraph.

**Baselines.** We compare our approach with several methods from the existing literature. **DY-Bootstrap** (Arazo et al. 2019) proposes to fit a beta mixture model to estimate the probability that a sample is mislabeled. **Data Param** (Saxena, Tuzel, and DeCoste 2019) allows a learnable parameter to each sample or class, which controls their importance during the training stage. **INCV** (Chen et al. 2019) presents a strategy to select a clean subset successively harnessing iterative cross-validation. **AUM** (Pleiss et al. 2020) has been described above. **Standard** represents training with the entire dataset, without removing any samples. **Oracle** means training without any synthesized noisy samples.

**Experiment Setups.** We test our approach on 7 datasets,
Table 2: Testing accuracy on real-world datasets (ResNet-32 for CIFAR/Tiny I.N., ResNet-50 for others).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CUB-200-2011</th>
<th>20% Sym.</th>
<th>mini WebVision 20%</th>
<th>Top1</th>
<th>Top5</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM-AugDesc</td>
<td>79.65</td>
<td>79.01</td>
<td>78.64</td>
<td>93.20</td>
<td></td>
</tr>
<tr>
<td>+ 5% D.D.[Ours]</td>
<td>79.55</td>
<td>79.15</td>
<td>79.32</td>
<td>92.56</td>
<td></td>
</tr>
<tr>
<td>+ 10% D.D.[Ours]</td>
<td>80.14</td>
<td>79.65</td>
<td>78.84</td>
<td>92.84</td>
<td></td>
</tr>
<tr>
<td>DivideMix</td>
<td>75.96</td>
<td>73.89</td>
<td>77.32</td>
<td>91.64</td>
<td></td>
</tr>
<tr>
<td>+ 5% D.D.[Ours]</td>
<td>76.06</td>
<td>74.44</td>
<td>77.84</td>
<td>92.16</td>
<td></td>
</tr>
<tr>
<td>+ 10% D.D.[Ours]</td>
<td>76.77</td>
<td>74.99</td>
<td>78.24</td>
<td>92.72</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Testing accuracy on noisy CUB-200-2011 and original mini WebVision. D.D. denotes data debugging for short. Competitive results (better than baselines) are bolded.

Combine with Algorithm-Centric Approaches
From an orthogonal aspect, we evaluate the enhancement to algorithm-centric state-of-the-art solution attributed to de-contamination of training data. Following the debugging algorithm described in Mazumder et al. (2022), we first select a number of samples with the most suspicion as label noise. Labels of these samples are then replaced by error-free ones, namely data debugging. We conduct experiments on both synthesized and real-world noisy datasets, i.e., a noised CUB-200 with 20% symmetric noises and mini WebVision. Note that the noise detector used here is trained in advance on a noised CIFAR-100. We follow the same experiment setups as DivideMix and DM-AugDesc, where ResNet-50 (He et al. 2016) and Inception-ResNet v2 (Szegedy et al. 2017) are used respectively.

Data Debugging Boosts. Table 3 shows the performance after debugging some training instances detected by our approach. By simply applying our data-centric solution to algorithm-centric states of the art, i.e., DivideMix and DM-AugDesc, testing accuracies on both synthesized and real-world noisy datasets are further improved without any changes in model architectures or optimization processes.

Analyses
We provide three analyses on the proposed noise detector.

Tolerance to Underlying Noise. Underlying noises may exist in real-world datasets, even in CIFAR-10. The super-

3mini WebVision denotes the Google-resourcing mini WebVision as previous works (Li, Socher, and Hoi 2020; Nishi et al. 2021).
The performance of the noise detector is satisfied if underlying noise ratios are no larger than 20%, showing good robustness against underlying noise. The mAP scores are measured on an independent synthesized noised Caltech256 with 40% symmetric noises, using an LSTM network as the noise detector trained on the twice-contaminated CIFAR-10 datasets. Two facts are observed: 1) Under the same synthesized noise ratios, more underlying noises would harm the detector. 2) The performance of the noise detector is satisfied if underlying noise ratio is no larger than 20%, showing good robustness against underlying noise.

Noises within Super-Classes of CIFAR-100. The 100 classes in the CIFAR-100 are grouped into 20 super-classes. To simulate the mislabeling in the annotation process, we contaminate the dataset twice, where the first contamination is representative of underlying noises and the second one for synthesized noises. Empirically, we demonstrate that the noise detector can tolerate large underlying noises ratios.

Feature Importance Analysis through LIME. Though the LSTM network does not necessarily have a set of interpretable rules to identify the mislabeled samples, we could analyze the importance of features through local explanations. While the exact formula of identifying the noise is hard to deduce, the local explanations indicate that such LSTM has significant variance than clean samples, as shown in Figure 1. The trained LSTM network also takes another half into consideration, where the given-label probabilities of mislabeled samples have a more significant variance than clean samples, as shown in Figure 1. While the exact formula of identifying the noise is hard to deduce, the local explanations indicate that such LSTM has learned relevant features to find mislabeled samples.

Conclusion

In this paper, we introduced a learning-based approach to identify mislabeled samples from training dynamics, which trains label noise detectors instanced by an LSTM network using synthesized noises. Beyond handcraft features, detectors can automatically learn relevant features from the time-series data to identify the mislabeled samples in the training set. This trained noise detector is then applicable to other noisy datasets without further adaptation or retraining. To validate the effectiveness of the proposed approach in detecting label noises, we conducted experiments on synthesized and real-world noisy datasets, where our approach excels the existing data-centric methods in both settings. We further combined our approach with the mainstream algorithms of learning with noisy labels, and provided orthogonal improvements from the data aspect. More analyses on the robustness and applicability of our approach were also provided in details.

For both researchers and practitioners, identifying mislabeled data can prelude the model development pipeline, at the sacrifice of only one additional step to obtain the corresponding training dynamics. Samples predicted to be mislabeled can be further re-used in other manners, such as advanced data debugging and semi-supervised learning.
Acknowledgements
This work is supported in part by National Key R&D Programs of China under the grant No. 2021ZD0110303.

References


