Machine Learning (2022) 111:1103–1124
https://doi.org/10.1007/s10994-021-06081-9

Improving deep label noise learning with dual active label correction

Shao-Yuan Li[1] · Ye Shi[1] · Sheng-Jun Huang[1] · Songcan Chen[1]

Received: 15 May 2021 / Revised: 13 August 2021 / Accepted: 22 September 2021 /
Published online: 6 January 2022

Abstract

Label noise is now a common problem in many applications, which may lead to significant learning performance degeneration. To deal with the label noise, Active Label Correction (ALC) was proposed to query the true labels for a small subset of instances. As the true labels costs can be high, the focus of ALC is to maximally improve the learning performance with minimal query costs. Existing ALC methods mainly proceed by querying the most likely mislabeled instances, or using criteria derived from standard active learning. In this paper, we focus on deep neural network models and show that due to their intrinsic memorization effect, the true labels of a large proportion of mislabeled instances can be correctly predicted with early stopped training, even under severe noise. Inspired by this, we propose to train deep label noise learning models robustly with dual ALC (DALC): on one hand, we select the most useful instances for classifier improvement and query their true labels from external experts; on the other hand, due to the active data sampling bias, the label noise model estimation can be highly biased, which may in turn hurt the classifier learning. To alleviate this issue, we propose to identify the instances that are most likely predicted with true labels by the classifier, and take the predictions as their true labels. By integrating the two sources of true labels, we experiment on multiple benchmark datasets with various label noise rate and show the effectiveness of the proposed DALC on both the classification accuracy and the label noise model estimation. The code is available at https://github.com/lilylisy/mlj21DALC.

Editors: Yu-Feng Li, Mehmet Gönen, Kee-Eung Kim.

✉ Shao-Yuan Li
lisy@nuaa.edu.cn

Ye Shi
shiye1998@nuaa.edu.cn

Sheng-Jun Huang
huangsj@nuaa.edu.cn

Songcan Chen
s.chen@nuaa.edu.cn

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Keywords Label noise · Deep neural networks · Active learning · Noise model

1 Introduction

With the advancement of deep learning on many domains, datasets are becoming bigger and bigger. However, in many real-world applications, obtaining large scale accurate labels is often infeasible due to the high cost or labeling difficulty, thus extensive data come along with noisy labels. For example, web crawling images with automatic label extractions (Krause et al., 2016; Xiao et al., 2015), crowdsourcing annotated tasks (Horvitz, 2007; Li et al., 2021), and data poisoning (Steinhardt et al., 2017). Inspiring studies have shown that label noise can significantly degenerates the performance of learning models (Angluin & Laird, 1987; Zhang et al., 2016), thus label noise robust algorithms have attracted much attention in recent years.

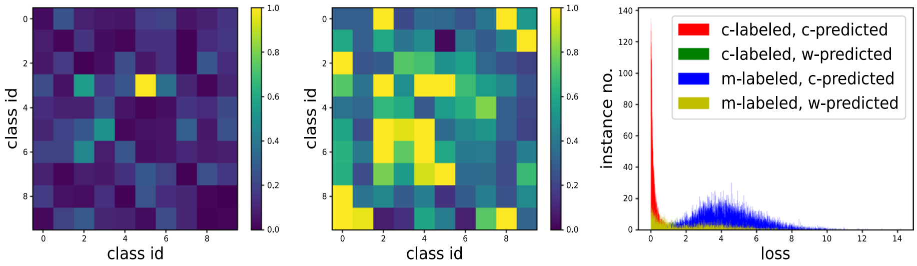
Recent researches have investigated acquiring true labels for some of the training instances, such that the generalization performance can be increased. As the cost of true labels can be high, e.g., through additional measurements or human experts, Active Label Correction (ALC) has been proposed from the active learning perspective (Kremer et al., 2018; Nallapati et al., 2009; Rebbapragada et al., 2012; Samel & Miao, 2018; Urner et al., 2012). Similar to standard active learning, ALC involves querying the true labels for a subset of training instances through human interaction. Unlike standard active learning, which acquires labels for unlabeled instances, for ALC, the training data are all labeled with a certain number of errors.

Early ALC works (Nallapati et al., 2009; Rebbapragada et al., 2012) focused themselves on detecting the mislabeled instances and re-labeling, without considering label noise modeling, which however is supposed to be easier to estimate with relatively small number of instances and can correct the noise. Following works (Kremer et al., 2018; Samel & Miao, 2018; Urner et al., 2012) remedied this by incorporating a label noise model, and adopting the active selection strategies derived from standard active learning, e.g., maximum expected model change (Kremer et al., 2018), uncertainty (Samel & Miao, 2018; Urner et al., 2012). A common limitation of these methods is that, as the true labels are not queried uniformly at random, the label noise model estimation can be seriously biased, which in turn would lead to inferior performance. Besides, these methods do not yet exploit the intrinsic fitting characteristics of deep learning models.

Taking the widely adopted transition matrix noise model as an example. Formally, let Y denote the clean label, $\bar{Y}$ denote the noisy label, X the instance, the transition matrix is defined as, $T_{ij}(x) = P(\bar{Y} = j|Y = i, X = x)$, which represents the transition probabilities that clean labels flip into noisy labels. As $T(x)$ is generally hard to learn, current state-of-the-art methods typically assume that T is class-dependent and instance independent, i.e., $P(\bar{Y} = j|Y = i, X = x) = P(\bar{Y} = j|Y = i)$. Transition matrix plays an essential role in building classifier-consistent algorithm in label noise learnig, i.e., by using the transition matrix T and noisy class posterior $P(\bar{Y} = j|X = x)$ (which can be estimated using noisy data), the clean class posterior $P(Y = j|X = x)$ can be inferred with strong theoretical ground (Han et al., 2018; Natarajan et al., 2013; Patrini et al., 2017).

Figure 1 shows an illustrative example of learning deep neural network classifiers with forward loss correction via leveraging transition matrix (Patrini et al., 2017). The experiment is conducted on CIFAR-10 with 40% symmetric label errors.[1] (a)–(b) show the

[1] The experimental details can be found in the experiments setting in Sect. 4.1.



(a) random **(b)** uncertainty **(c)** loss distribution

Fig. 1 Illustrative experimental results on CIFAR-10 with symmetric label noise ratio 40% as an example. **a**, **b** are respectively the estimation errors of the transition matrix with 5% training instances with true labels queried by random and uncertainty strategies, which are calculated as $|T - \hat{T}|/T$, with $\hat{T}$ denoting the estimated transition matrix, T the groundtruth transition matrix. **c** is the cross entropy loss distribution of the dnn classifier on all training instances without active true label query when trained early stopped

estimation errors of the transition matrix, when 5% training instances are queried with true labels and used for transition matrix estimation. The true labels are queried by random sampling and uncertainty sampling (largest entropy) respectively. The error is calculated by $|T - \hat{T}|/T$, with $\hat{T}$ and T as the estimated and groundtruth transition matrix. It can be seen that due to the data sampling bias, the transition matrix estimation error of uncertainty query is severely larger than that of the random strategy.

Figure 1c demonstrates the fitting characteristics of the deep neural network classifier on noisy data. Note that in this case, we have not use ALC, i.e., no extra true labels are used in training. It demonstrates the prediction loss distribution of the learned classifier on the training instances. For better illustration, we split the training instances into four groups based on their training labels and the classifier's predictions: c-labeled (m-labeled) means the training instances are correctly labeled (mislabeled), c-predicted (w-predicted) means the classifier's predictions are correct (wrong) with respect to their groundtruth labels. With early stopped training, it can be seen that, (1) the losses of the correctly labeled instances are much smaller than that of the mislabeled ones, (2) rather inspiring, for a large proportion of mislabeled instances, although their training labels are wrong, the classifier can correctly predict their true labels.

The results of Fig. 1c can be explained by the critical memorization effects of deep neural networks found by Arpit et al. (2017), i.e., deep neural networks tend to memorize and fit easy (clean) patterns first, and then gradually overfit hard (noisy) patterns. This phenomenon has inspired early stopping (Li et al., 2020) and small-loss tricks (Han et al., 2018; Jiang et al., 2018) to combat with noisy labels, which respectively avoid overfitting noisy labels to some degree by ending training early, and treat the small-loss instances as clean instances and only back propagates them to update the model parameters. However, such memorization effects are rarely leveraged in ALC to help select the most helpful instances with low query cost.

Motivated by the above phenomena, in this paper, we propose our Dual Active Label Correction (DALC) approach for deep label noise learning, which try to learn an accurate dnn classifier as well as low biased noise model with less labeling cost. In a high level, we introduce true labels iteratively for some selected instances, and then conduct learning. For the truly labeled instances and noisy labeled instances, we respectively use normal cross entropy loss, and forward corrected cross entropy loss via leveraging the transition matrix noise model. Specifically, different from existing ALC methods, our DALC framework proposes to select instances to get true labels from external human experts as well as from the internal classifier prediction. For the external experts labeling, we propose an uncertainty query strategy to select instances with most uncertain label prediction to query their true labels. The most uncertain instances are supposed to contribute most to improving the classifier. For the internal classifier labeling, we propose one confidence query strategy to identify instances that are most likely correctly predicted, and use the classifier's prediction as the true label. The confidence query strategy leverages the memorization effects of deep neural networks with early stopped training. Combing the two sources of instances, the transition matrix estimation bias induced by the uncertainty strategy can be alleviated. In comparison with the small-loss tricks which tend to select correctly labeled instances, we empirically show that the confidence strategy can also detect and re-correct noisy labels.

The rest of the paper is organized as follows. In Sect. 2, we give a brief review of related work. Then, we formulate our problem and propose the DALC method. Section 4 reports several empirical results on a number of label-noise learning benchmark data. Finally, we give a conclusion in Sect. 5.

2 Related work

Label noise is not new in the machine learning field, whose study can be dated back to the work of Angluin and Laird (1987), which discussed the possibility of learning algorithms to cope with incorrect training data. Since then, with the ubiquitous noisy and imperfect labels or annotations in real world environment, designing noise robust learning models has become urgent. According to whether the learned classifier is statistically consistent, i.e., whether the classifier guarantees to converge to the optimal classifier trained on the clean labels, existing label noise learning algorithms can be roughly categorized into two groups: statistically inconsistent methods and statistically consistent ones. Many state-of-the-art approaches with statistically inconsistent classifiers are specifically designed through some heuristic, e.g., reliable example filtering (Han et al., 2018; Jiang et al., 2018; Li et al., 2020; Ren et al., 2018), robust losses (Liu & Guo, 2020; Ma et al., 2020; Zhang & Sabuncu, 2018), label correction (Reed et al., 2014; Tanaka et al., 2018), and regularization (Li et al., 2017, 2020; Liu et al., 2020).

The classifier consistent algorithms basically introduce a noise transition matrix which models the probabilities of clean labels flipping into noisy labels, and build up the relationship between the latent true labels and observed noisy labels (Goldberger & Ben-Reuven, 2017; Natarajan et al., 2013; Patrini et al., 2017). For such line of work, the major concern is how to effectively estimate the transition matrix and leverage the estimated matrix to combat with the label noise. Liu and Tao (2016) proposed to estimate the transition matrix through cross-validation, and then build an importance weighted loss function using the estimated matrix. As the computational complexity for the transition matrix estimation is prohibited for multi-class problems, this method only applies to binary classification. Goldberger and Ben-Reuven (2017), Sukhbaatar et al. (2015) proposed to add an adaptation layer with different constraints after the softmax output layer of the classifier. The adaptation layer can be regarded as a transition matrix function. Treating the classifier and the transition matrix as components of the same network, the transition matrix and the classifier are estimated simultaneously in an end to end manner by back propagating the cross-entropy loss. Patrini et al. (2017) proposed to leverage the noise transition matrix to conduct loss correction, such that training on the noisy labels via the corrected loss should be approximately equal to training on clean labels via the original loss. In Patrini et al. (2017), two loss correction methods have been proposed, i.e., the forward correction and the backward correction, which respectively corrects the classifier's predictions and the cross entropy loss by transition matrix. Empirically the backward correction method has been reported to perform worse than the forward method. As one state-of-the-art noisy label learning method, forward loss correction has been widely used as the base model of various following methods (Hendrycks et al., 2018; Xia et al., 2019; Yao et al., 2020). For a comprehensive understanding of the label noise learning field, we recommend the recent reviews (Frenay & Verleysen, 2014; Han et al., 2020; Song et al., 2020).

In this paper, we also adopt the forward loss correction method (Patrini et al., 2017) as our base model, and propose to actively select a small set of training instances to obtain their true labels to ensure robustness gains. Rebbapragada et al. (2012), Urner et al. (2012) seemed to be the earliest work exploring this idea, respectively formulated as active label correction and learning from weak teachers. Nallapati et al. (2009) shared the similar idea with Rebbapragada et al. (2012), which try to detect and re-label the mislabeled instances, and then conduct learning for classic statistical models such as logistic regression, svm on the cleaned data, without modeling and combating with the label noise during model training. Urner et al. (2012) assumed that the instances labels are likely to be correct in label-homogeneous regions and deteriorate near classification boundaries, and analyzed the sample complexity of the setting. The most closely related work to us are Kremer et al. (2018), Samel and Miao (2018), which considered using deep neural network as learning model and incorporated a label noise model. Kremer et al. (2018) adopted the active selection strategies derived from maximum expected model change, Samel and Miao (2018) used one margin measure defined as the discrepancy of the predicted label and the observed label. One common drawback of Kremer et al. (2018), Samel and Miao (2018) is that, they ignore the label noise model estimation bias which could be serious due to the active data sampling bias. They also ignored exploiting the intrinsic characteristics of deep learning models to help reduce this bias and further save labeling cost.

Note that there are also some other label noise learning methods which leverage clean labels for neural network pretrain or knowledge distillation or transition matrix estimation, e.g., (Hendrycks et al., 2018; Li et al., 2017; Ren et al., 2018; Veit et al., 2017; Xiao et al., 2015). Whereas these work assume readily available clean labels, which is different from our active label correction concern.

3 The proposed approach

We use $\mathcal{X} \subset \mathbb{R}^d$ to denote the instance space, $\mathcal{Y} = \{1, 2, \dots, C\}$ to denote the label space, $x \in \mathcal{X}$ to denote one specific instance and $y \in \mathcal{Y}$ to denote one specific label value. With X, Y denoting the random variables, we use $P(X, Y)$ to represent the ground-truth joint probability distribution over the instance label pairs (x, y). In label noise learning, rather than ground-truth labels, a training set $D = \{(x_i, \bar{y}_i)_{i=1}^N\}$ with noisy labels $\bar{y}$ drawn from a corrupted label probability distribution $P(\bar{Y}|X)$ is given. Here we use $\bar{Y}$ to denote the corrupted label. The learning target is to learn the mapping from X to Y for the ground-truth joint distribution $P(X, Y)$, i.e., predicting the true label y for any given instance $x \in \mathcal{X}$.

In the following, we will first introduce the key concept of noise transition matrix, then the forward loss correction base model we used as our noisy label learning model, and then propose our dual active label correction (DALC) approach.

3.1 Transition matrix

Definition 1 (*Noise transition matrix* van Rooyen & Williamson, 2017) Suppose that the observed label $\bar{y}$ is noisy i.i.d. drawn from a corrupted distribution $P(Y|x)$, where features are intact. Meanwhile, there exists a corruption process, transitioning from the latent clean label y to the observed noisy label $\bar{y}$. Such corruption process can be approximately modeled via a label transition matrix T, where $T_{ij} = P(\bar{Y} = j|Y = i)$.

From the definition, the (i, j)-th entry $T_{ij} = P(\bar{Y} = j|Y = i)$ represents the probability that instance x belonging to class i having a noisy class label j. Thus the following import equation between the noisy label posterior $P(\bar{Y}|X)$ and the clean label posterior $P(Y|X)$ can be induced as:

$$P(\bar{Y} = j|X = x) = \sum_{i=1}^{C} T_{ij} P(Y = i|X = x;\theta). \tag{1}$$

Based on Eq. (1), it can be seen that the clean label posterior $P(Y|X)$ can be inferred by using the T and the noisy label posterior $P(\bar{Y}|X)$. This equation has been widely used in label noise domain to learn *statistically consistent classifiers*. That is to say, the classifier learned by using the noisy labels $\bar{y}$ will asymptotically converges to the optimal classifier defined on the clean label y (Natarajan et al., 2013; Patrini et al., 2017; Xia et al., 2019; Yao et al., 2020).

Note that here a simplified assumption is made that the noise transition matrix T is only *class-dependent*, i.e., $T(x)_{ij} = T_{ij}$ for any x. While in practice the noise pattern can be more complex, this simplification makes T to be identifiable under mild conditions and has been shown to be rather effective in capturing aspects of the label noise and guiding learning by vast works. Next, we introduce the forward loss correction model proposed by Patrini et al. (2017) as our base learning model, which is one state of the art noisy label learning method and base model of a variety of works (Hendrycks et al., 2018; Kremer et al., 2018; Xia et al., 2019; Yao et al., 2020).

3.2 Forward loss correction base model

Loss correction is an important branch in label noise learning via leveraging the transition matrix. The aim of loss correction is that, training on noisy labels via the corrected loss should be approximately equal to training on clean labels via the original loss.

Using the deep neural network as learning model, Patrini et al. (2017) introduced the forward loss correction technique, which corrects the network predictions by the transition matrix T. Using $g(x;\theta)$ to denote the softmax output of the deep neural network classifier $g(\cdot)$ for some instance x with parameter θ:

$$g(x;\theta) = \hat{P}(Y|X = x) \approx P(Y|X = x). \tag{2}$$

The noisy label prediction $f(x)$ is derived with the transition matrix as:

$$f(x) = T^\top g(x;\theta) = \hat{P}(\bar{Y}|X = x) \approx P(\bar{Y}|X = x). \tag{3}$$

Given the noisy data $(x_i, \bar{y}_i)$, a specific loss function ℓ, e.g., the cross entropy loss we used in our paper, the network parameter θ and the transition matrix T, the forward correction loss $\ell^{\rightarrow}$ is defined as:

$$\ell^{\rightarrow}(x_i, \bar{y}_i;\theta) := \ell(T^\top g(x_i;\theta), \bar{y}_i). \tag{4}$$

For the forward loss correction defined in Eq. 4, Patrini et al. (2017) gives a formal theoretical guarantee *w.r.t* the clean data distribution as following:

Theorem 1 (Forward Correction, Theorem 1 in Patrini et al. 2017) *Suppose that the label transition matrix T with $T_{ij} = P(\bar{Y} = j|Y = i)$ is non-singular. Given loss ℓ and network parameter θ, the minimizer of the corrected loss $\ell^{\rightarrow}$ under the noisy distribution is the same as the minimizer of the original loss ℓ under the clean distribution*:

$$\arg\min \mathbb{E}_{x_i,\bar{y}_i} \ell^{\rightarrow}(x_i, \bar{y}_i;\theta) = \arg\min \mathbb{E}_{x_i,y_i} \ell(x_i, y_i;\theta). \tag{5}$$

Theorem 1 shows that the noise transition matrix plays an essential role in the loss correction model, whose estimation quality will signficantly impact the learning performance. Patrini et al. (2017) originally proposed one heuritic approach to estimate the transition matrix, and then minimize the corrected loss in Eq. (4) over all instances to learn the neural network classifier with transition matrix fixed. They first train a neural network on the noisy data $D = \{(x_i, \bar{y}_i)_{i=1}^N\}$, then use this network to get the noisy class posterior prediction $\hat{P}(\bar{Y} = i|X = x)$ for each instance x. For each class i, they identify the set of instances A_i with largest $\hat{P}(\bar{Y} = i|X = x)$ as truely belonging to class i, i.e., $\hat{P}(Y = i|X = x) \approx 1$, and then estimate the transition matrix using the following equation, which is derived from Eq. (1):

$$\hat{T}_{ij} = \frac{1}{A_i} \sum_{x \in A_i} \hat{P}(\bar{Y} = j|X = x). \tag{6}$$

Algorithm 1 summarizes the process of the forward loss correction algorithm. From Eq. (6), it can be seen that the transition matrix estimation depends on the subset of data A_i with clean labels. Actually, a number of statistically consistent methods (Goldberger & Ben-Reuven, 2017; Hendrycks et al., 2018; Patrini et al., 2017; Sukhbaatar et al., 2015; Xia et al., 2019; Yao et al., 2020) have been proposed mainly on how to better estimate the subset A_i and the transition matrix.

In this paper, we focus on actively selecting instances with clean labels. We emphasize that the data sampling bias of standard active strategies can lead to highly biased transition matrix estimation. E.g., when using the common uncertain strategy (largest entropy) to query clean labels, i.e., instances with most uncertain predictions $g(x;\theta) = \hat{P}(Y|X = x)$ are selected, their noisy class posterior predictions $f(x) = T^\top g(x;\theta) = \hat{P}(\bar{Y}|X = x)$ are inevitably to be more uncertain with larger entropy. By Eq. (6), each row $\hat{T}_{i\cdot}$ is computed as the average of $\hat{P}(\bar{Y} = j|X = x)$ for $x \in A_i$, meaning that the uncertain strategy would prefer $\hat{T}_{i\cdot}$ with large entropy.

To alleviate this effect, we propose our dual active label correction method by incoporating a subset of instances with most confident predictions $g(x;\theta)$, whose noisy class posterior predictions $f(x)$ tend to be confident with small entropy. Combing the two sources of instances, the transition matrix estimation bias is expected to be reduced. We will explain the details in the next subsection.

Algorithm 1 The Forward Loss Correction algorithm

```
Input:
  $D = \{(x_i, \bar{y}_i)_{i=1}^{N}\}$: training data with label noise
  $\ell$: loss function
Phase1: transition matrix estimation
  train a neural network $f(x;\gamma)$ with parameter $\gamma$ using loss $\ell(f(x_i;\gamma), \bar{y}_i)$ on $D$
  for each $x_i$, take the softmax output of $f(x_i;\gamma)$ as its noisy label posterior $\hat{P}(\bar{Y}|X = x_i)$
  for $i = 1, \cdots, C$
    identify the set of instances $A_i$ with largest $\hat{P}(\bar{y} = i|x)$
    for $j = 1, \cdots, C$
      estimate the $(i, j)$-th entry of $\hat{T}_{ij}$ using Eq.( 6)
    end for
  end for
Phase2:classifier training
  with $T$ fixed as $\hat{T}$, train $g(x;\theta)$ using the corrected loss $\ell^{\rightarrow}$ defined in Eq.( 4) on $D$
Output:
  $g(x;\theta)$: the deep neural network classifier
  $\hat{T}$: the transition matrix estimation
  $\hat{P}(\bar{Y}|X = x_i), i = 1, \cdots, N$: the noisy label posterior estimation of each $x_i$
```

3.3 The dual active label correction (DALC) framework

We use the forward loss correction model as our base learning model. During the training, DALC progressively queries the true labels of a subset of the training data, and feeds them into the model to update the transition matrix T and the neural network $g(x;\theta)$. The true labels come from two sources, i.e., from both external human experts and the internal classifier's predictions. The instances for which the classifier is most uncertain are selected for querying true labels from human experts, which are supposed to be diffcult instances for the classifier. Considering that the active sampling bias may lead to serious transition matrix estimation error, which in turn would hurt the classifier learning, we take advantage of the memorization effects of deep neural networks. Specifically, we propose one confidence measure to select instances that are most likely correctly predicted by the classifier, and use the predicions as their true labels.

In the following we will first explain the uncertainty and confidence measures, and then summarize the algorithm.

3.3.1 The selection measures

Uncertainty Uncertainty is a commonly used measure in traditional active learning, which measures how uncertain the prediction of the current classifier is for some instance. With $\hat{P}(Y|X = x)$ denoting the label probability prediction of the classifier $g(x;\theta)$ on some instance x, we propose two uncertainty criteria for the training instances.

- *Largest Entropy* The first is the commonly used unsupervised entropy measure. Instances with largest prediction entropy are supposed to be difficult for the classifier. We select them to query the true labels from the human experts. The entropy of the classifier prediction $g(x;\theta) = \hat{P}(Y = i|X = x)$ for instance x is defined as:

$$entropy(x) = -\sum_{i=1}^{C} g(x;\theta) \cdot \log g(x;\theta). \tag{7}$$

- *Largest Loss* As we have labels for the training data, the classifier's prediction loss on each instance can be used as one uncertainty measure. Given one instance $(x, \bar{y})$, the loss is computed between the classifier's prediction $g(x;\theta)$ and the given training label $\bar{y}$:

  $$loss(x) = \ell(g(x;\theta), \bar{y}). \tag{8}$$

  This measure coincides with the small-loss tricks and stems from the memorization effects of deep neural networks (Arpit et al., 2017), i.e., during the optimization process, the deep neural networks tend to memorize and fit easy (clean) patterns first, and then gradually overfit hard (noisy) patterns. With early stopped training, the small loss and large loss instances can be respectively regarded as correctly labeled and mislabeled instances. The large loss measure tends to detect the mislabeled instances and query their true labels.

Confidence To combat with the transition matrix estimation bias induced by the active sampling bias, and further save query cost, we propose to identify the most likely correctly predicted instances and use their predicions as the true labels. In parallel with the uncertainty measure, we also make use of the entropy and loss defined in Eqs. (7) and (8) to get the confidence measure. Different from the external human query which select the most *uncertain* instances with *largest entropy* and *largest loss*, for the internal clasifier's prediction selection, the most *confident* instance predictions with *least entropy* and *least loss* are selected.

After define the uncertainty and confidence meansures, next we will introduce the selection and utilization details of queried true labels.

3.3.2 The DALC algorithm

The main steps for the proposed DALC (Dual Active Label Correction) algorithm are summarized in Algorithm 2. Before conducting the active data sampling, one initial classifier $g(x;\theta)$ and the noisy label posterior estimation $\hat{P}(\bar{Y}|X)$ are obtained by using the forward loss correction method on D described in Algorithm 1. Then based on the predictions of $g(x;\theta)$, two subsets of most uncertain instances A_h (with largest entropy) and most confident instances A_g (with smallest entropy or loss score) are selected according to their entropy or loss scores defined in Eqs. (7, 8). For each instance $x \in A_h$, its true label is queried from human experts; for each instances $x \in A_g$, the classifier's prediction $g(x;\theta)$ is used as its true label. Then we remove $A_h \cup A_g$ from D, and get two groups of instances $A_h \cup A_g$ and D respectively with true labels and noisy labels. After that, each entry of the transition matrix $\hat{T}_{ij}$ is estimated using Eq. (6). The classifier $g(x;\theta)$ is trained using loss $\ell(g(x_i;\theta), \bar{y}_i)$ for instances belonging to $A_h \cup A_g$ and $\ell^{\rightarrow}(x_i, \bar{y}_i;\theta)$ for instances belonging to $D - A_h - A_g$, i.e., the mixed loss defined as following:

$$\ell^m := \sum_{x_i \in A_h \cup A_g} \ell(g(x_i;\theta), y_i) + \sum_{x_i \in D} \ell(\hat{T}^\top g(x_i;\theta), \bar{y}_i). \tag{9}$$

The active process repeats until the the query budget is reached or the classifier obtains a specified performance.

Algorithm 2 The Dual Active Label Correction algorithm

```
Input:
  $D = \{(x_i, \bar{y}_i)_{i=1}^N\}$: training data with label noise
  $\ell$: loss function
Initialization:
  apply the forward loss correction method in Algorithm  1 on $D$
  obtain $g(x;\theta)$, $\hat{P}(\bar{Y}|X = x_i), i = 1, \cdots, N$: the output of Algorithm  1
Repeat:
  Phase1: active instance query
  for each instance $x \in D$
    calculate the its entropy score and loss score according to Eq.( 7),( 8)
  end for
  select a batch of instances $A_h$ from $D$ with largest entropy score
  query the true labels of $A_h$ from human experts, and remove $A_h$ from $D$
  select a batch of instances $A_g$ from $D$ with smallest entropy score or loss score
  take the output of $g(x;\theta)$ for $x \in B$ as their true labels, and remove $A_g$ from $D$
  Phase2: transition matrix update
  for $i = 1, \cdots, C$
    identify the set of instances $A_i \subset (A_h \cup A_g)$ with label $i$
    for $j = 1, \cdots, C$
      estimate the $(i, j)$-th entry of $\hat{T}_{ij}$ using Eq.( 6)
    end for
  end for
  Phase3: classifier update
  with $T$ fixed as $\hat{T}$, train $g(x;\theta)$ using the mixed loss defined in Eq.( 10) on $D \cup A_h \cup A_g$
Until query budget or expected performance reached
Output:
  $g(x;\theta)$: the deep neural network classifier
  $\hat{T}$: the transition matrix estimation
```

Note that in Algorithm 2, our DALC algorithm employs the entropy score rather than the loss to measure the uncertainty. The loss score based uncertainty defined in Eq. (8) is actually the main movitation of Nallapati et al. (2009), Rebbapragada et al. (2012), whose target is detecting and relabeling the mislabeled instances, and then training the classic statistical learning models with cleaned data. However, as shown in Fig. 1c, or the deep neural networks, they are characterized by correctly predicting the true labels for noisy instances even with large loss. In the experiment, we will show more details about the effects of the loss score based uncertainty measures.

4 Experiments

4.1 Settings

Dataset We perform experiments on CIFAR-10 (Krizhevsky, 2009), CIFAR-100 (Krizhevsky, 2009) and MNIST (LeCun et al., 1998), which are commonly used benchmark dataset in label noise learning tasks. CIFAR-10 has 10 classes of images including 50, 000 training images and 10, 000 test images. CIFAR- 100 also has 50, 000 training images and 10, 000 test images, but 100 classes. MNIST has 10 classes of images including 60, 000 training images and 10, 000 test images.

The original datasets contain clean data. To generate noisy labels, we corrupt the training data of each dataset according to a transition matrix T. We employ the symmetry flipping setting (van Rooyen et al., 2015), which models the most common label noise classification task by uniformly flipping the class of clean label into the other classes. For each data with true label i, its corrupted label is sampled from the categorical distribution parameterized by the ith row of T. We generates label noise from light to heavy at 40, 60, 80 corruptions fractions, which leads aroung 40%, 60%, 80% of instances to have noisy labels.

Network and optimization For fair comparison, we implement the methods following the settings in Hendrycks et al. (2018). Specifically, for CIFAR-10 and CIFAR-100, we train a Wide Residual Network (Zagoruyko & Komodakis, 2016) of depth 40 and a widening factor of 2. We train for 75 epochs using SGD with Nesterov momentum and a cosine learning rate schedule (Loshchilov & Hutter, 2017). For MNIST, we use a 2-layer fully connected network with 256 hidden dimensions. The training is conducted by using Adam for 10 epochs with batch size of 32 and a learning rate of 0.001. l_2 weight decay regularization is used on all layers with $\lambda = 1 * 10^{-6}$.

Baselines To assess the performance of the proposed DALC, we conduct comparisons for the following implementations:

- Random which randomly selects instances and query their true labels from human experts.
- Entropy which selects the most uncertain instances with the maximum entropy score according to Eq. (7), and query their true labels from human experts.
- Loss which selects the most uncertain instances with the maximum loss score according to Eq. (8), and query their true labels from external human experts.
- DALC-*l* which selects the most uncertain instances with the maximum entropy score according to Eq. (7), and query their true labels from human experts; also selects the most confident instances with the minimum loss score according to Eq. (8), and use the classifier's prediction as their true labels.
- DALC-*e* which selects the most uncertain instances with the maximum entropy score according to Eq. (7), and query their true labels from human experts; also selects the most confident instances with the minimum entropy score according to Eq. (7), and use the classifier's prediction as their true labels.

Except for the active study with label noise modeling consideration, we aslo compare with the *Noise-agnostic* estimator, i.e, which learns on the training data without modeling the label noise and using the normal loss ℓ:

$$\ell^{agnostic} := \sum_{x_i \in A_h \cup A_g} \ell(g(x_i;\theta), y_i) + \sum_{x_i \in D} \ell(g(x_i;\theta), \bar{y}_i). \tag{10}$$

With this comparison, we show the importance of label noise modeling. We examine the classification accuracy on the test set, and the discrepancy between the learned transition matrix $\hat{T}$ and the true one T. To avoid the influence of randomness, we repeat the experiments for 5 times and report the average results. We omit the standard deviations which

Table 1 The classification accuracy on CIFAR-10, the best performance are bold

Noise (%)	Model	Strategy	Percent of queried instances					
			5%	10%	15%	20%	25%	30%
40	Noise-agnostic	Random	0.875	0.884	0.892	0.889	0.878	0.900
		Entropy	0.881	0.896	0.910	0.922	0.925	0.924
		Loss	0.860	0.862	0.866	0.870	0.884	0.901
		DALC-*l*	0.895	0.910	0.917	0.925	0.927	0.933
		DALC-*e*	0.890	0.908	0.922	0.929	0.931	0.934
	Forward loss correction	Random	0.889	0.898	0.904	0.910	0.910	0.916
		Entropy	0.901	0.918	0.927	0.932	0.936	0.938
		Loss	0.887	0.902	0.902	0.914	0.920	0.930
		DALC-*l*	0.904	0.918	0.927	0.930	0.938	**0.942**
		DALC-*e*	**0.907**	**0.920**	**0.930**	**0.934**	**0.939**	0.941
60	Noise-agnostic	Random	0.827	0.835	0.800	0.783	0.796	0.852
		Entropy	0.829	0.852	0.861	0.898	0.913	0.913
		Loss	0.717	0.643	0.702	0.730	0.786	0.787
		DALC-*l*	0.818	0.797	0.869	0.863	0.869	0.854
		DALC-*e*	0.534	0.596	0.760	0.822	0.846	0.860
	Forward loss correction	Random	0.849	0.864	0.878	0.891	0.899	0.906
		Entropy	**0.881**	0.895	0.906	0.920	0.927	0.932
		Loss	0.840	0.847	0.854	0.867	0.878	0.891
		DALC-*l*	0.864	0.892	0.910	0.923	0.932	0.934
		DALC-*e*	0.872	**0.900**	**0.914**	**0.924**	**0.934**	**0.939**
80	Noise-agnostic	Random	0.659	0.623	0.629	0.668	0.663	0.723
		Entropy	0.683	0.638	0.777	0.845	0.860	0.871
		Loss	0.717	0.643	0.702	0.730	0.786	0.787
		DALC-*l*	0.624	0.429	0.624	0.730	0.757	0.776
		DALC-*e*	0.534	0.596	0.760	0.822	0.846	0.860
	Forward loss correction	Random	0.806	0.839	0.860	0.878	0.888	0.902
		Entropy	0.799	0.861	0.887	0.899	0.918	0.922
		Loss	0.784	0.811	0.838	0.859	0.875	0.888
		DALC-*l*	0.802	0.864	0.890	**0.915**	0.923	0.923
		DALC-*e*	**0.816**	**0.873**	**0.897**	0.910	**0.925**	**0.929**

mainly vary in range $[0.1\%, 0.6\%]$ for the test accuracy and [0.01, 0.1] for the transition matrix discrepancy due to the overflowed table format.

4.2 Comparison for classification accuracy

Tables 1, 2 and 3 respectively show the classification accuracy of compared methods as the number of queried instances increases on CIFAR-10, CIFAR-100 and MNIST. Here the number denotes the number of queried true labels from external human experts, which makes a fair comparison between DALC and the random, entroy, loss baselines. As for the internal classifier prediction aquisition for DALC, a equal number of most confident instance are incoporated in the experiment for implementation simplicity.

Table 2 The classification accuracy on CIFAR-100, the best performance are bold

Noise (%)	Model	Strategy	Percent of queried instances					
			5%	10%	15%	20%	25%	30%
40	Noise-agnostic	Random	0.606	0.628	0.643	0.657	0.657	0.665
		Entropy	0.625	0.644	0.668	0.678	0.694	0.703
		Loss	0.609	0.632	0.652	0.666	0.683	0.704
		DALC-*l*	0.633	0.667	0.686	0.689	0.702	0.716
		DALC-*e*	0.632	0.663	0.674	0.697	0.709	0.713
	Forward loss correction	Random	0.640	0.670	0.679	0.691	0.701	0.708
		Entropy	0.654	0.693	0.707	0.717	0.722	0.734
		Loss	0.663	0.687	0.706	0.717	0.728	0.737
		DALC-*l*	0.667	0.697	0.709	0.719	0.732	0.737
		DALC-*e*	**0.674**	**0.701**	**0.714**	**0.727**	**0.736**	**0.737**
60	Noise-agnostic	Random	0.511	0.543	0.587	0.580	0.610	0.624
		Entropy	0.522	**0.565**	0.594	0.632	0.648	0.670
		Loss	0.517	0.553	0.565	0.590	0.612	0.625
		DALC-*l*	0.561	0.510	0.583	0.618	0.636	0.637
		DALC-*e*	0.570	0.558	0.615	0.647	0.665	0.686
	Forward loss correction	Random	0.582	0.610	0.637	0.650	0.668	0.676
		Entropy	0.556	0.629	0.651	0.673	0.690	0.695
		Loss	0.578	0.619	0.647	0.670	0.685	0.698
		DALC-*l*	**0.585**	0.633	0.656	0.682	0.696	0.713
		DALC-*e*	0.584	**0.639**	**0.670**	**0.691**	**0.704**	**0.716**
80	Noise-agnostic	Random	0.292	0.226	0.242	0.359	0.434	0.514
		Entropy	0.282	0.309	0.340	0.396	0.481	0.582
		Loss	0.300	0.334	0.380	0.434	0.496	0.545
		DALC-*l*	0.081	0.131	0.226	0.354	0.377	0.424
		DALC-*e*	0.144	0.190	0.350	0.460	0.513	0.541
	Forward loss correction	Random	**0.305**	**0.475**	**0.554**	**0.596**	**0.621**	**0.650**
		Entropy	0.235	0.414	0.512	0.558	0.609	0.641
		Loss	0.302	0.430	0.517	0.564	0.605	0.642
		DALC-*l*	0.284	0.438	0.526	0.556	0.577	0.597
		DALC-*e*	0.262	0.360	0.431	0.481	0.517	0.552

First, we make a high level comparison between the noise-agnostic model and the forward loss correction model. Except for the case of MNIST dataset, we can see a clear margin between the two models for almost all noise rates and active scenarioes i.e., the performance of forward loss correction model easily dominate the noise-agnostic model. The MNIST dataset is kind of easy and relatively more robust to label noise. Even learning the neural network directly on the noisy training data with normal cross entropy loss can achieve satisfactory performance. Whereas for common complex datasets for which the effect of label noise can be severe, modeling the label noise is rather helpful.

Second, we compare between the entropy sampling strategy and the loss sampling strategy. On CIFAR-10, both for the noise-agnostic model and the forward loss correction model, the entropy baseline achieves much higher classification accuracy than the loss baseline. The underlying reason is due to the phenomenon that we have shown and

Table 3 The classification accuracy on MNIST, the best performance are bold

Noise (%)	Model	Strategy	Number of queried instances					
			18	54	90	126	162	180
40	Noise-agnostic	Random	0.964	0.963	0.962	0.961	0.962	0.964
		Entropy	0.962	0.962	0.962	0.962	0.964	0.964
		Loss	0.959	0.963	0.963	0.963	0.967	0.961
		DALC-*l*	0.962	0.960	0.962	0.964	0.962	0.964
		DALC-*e*	0.961	0.966	0.963	0.962	0.963	0.960
	Forward loss correction	Random	0.961	0.964	0.966	0.971	0.966	0.969
		Entropy	0.965	0.966	0.967	0.969	0.969	**0.970**
		Loss	0.964	0.965	0.964	0.965	0.966	0.966
		DALC-*l*	0.958	0.866	0.864	0.966	0.964	0.966
		DALC-*e*	**0.966**	**0.967**	**0.969**	**0.971**	**0.970**	0.968
60	Noise-agnostic	Random	0.949	0.949	0.951	0.948	0.946	0.950
		Entropy	0.949	0.951	0.952	0.952	0.952	0.951
		Loss	0.949	0.952	0.951	0.951	0.952	0.949
		DALC-*l*	0.948	0.953	0.949	0.953	0.951	0.952
		DALC-*e*	0.951	0.953	0.950	0.946	0.951	0.951
	Forward loss correction	Random	0.949	0.944	0.954	0.950	0.952	0.950
		Entropy	0.906	0.920	0.952	0.954	0.957	0.955
		Loss	0.916	0.950	0.951	0.954	**0.958**	0.950
		DALC-*l*	0.925	0.935	0.854	0.939	0.949	0.955
		DALC-*e*	**0.951**	**0.956**	**0.956**	**0.956**	**0.958**	**0.959**
80	Noise- agnostic	Random	**0.923**	0.913	0.915	0.912	0.913	0.920
		Entropy	0.906	0.912	0.907	0.922	0.917	0.929
		Loss	0.913	0.913	0.912	0.912	0.910	0.915
		DALC-*l*	0.914	0.919	0.917	0.917	0.918	0.910
		DALC-*e*	0.917	0.915	0.905	0.899	0.908	0.905
	Forward loss correction	Random	0.906	0.909	0.919	0.922	0.925	0.927
		Entropy	0.713	0.617	0.853	0.856	0.901	0.919
		Loss	0.791	0.628	0.733	0.829	0.895	0.913
		DALC-*l*	0.721	0.918	0.905	0.924	0.913	0.929
		DALC-*e*	0.796	**0.926**	**0.921**	**0.934**	**0.929**	**0.934**

explained in the introduction section, i.e., a number of mislabeled instances with large loss can be correctly predicted by the classifier. In this case, querying their true labels from the human experts are redundant and waste of labeling cost. This result can also be observed for CIFAR-100 with noise ratio 40% and 60%. When the noise ratio is as high as 80% for CIFAR-100, the accuracy of both the forward loss correction model and noise-agnostic model collapse, i.e., 0.136 and 0.206. In such case, the classifier is too weak to give good predictions, and the active sampling strategies are less effective than the random strategy. Due to the same reason, the loss sampling strategy performs better than the entropy strategy. As the number of queried instances increase and the performance increases, e.g., 30% queried instances, the performance of entropy strategy becomes comparable. For the relatively easy MNIST dataset, the two strategies are comparable.

Thirdly, we compare the proposed DALC approach with other baselines. It can be seen that combined with the forward loss correction model, except for the difficult case of CIFAR-100 with 80% noise ratio, for almost all the other cases, built on the entropy based uncertainty strategy, the DALC-*e* strategy enhanced with entropy based confidence sampling achieves the best performance. The DALC-*l* with loss based confidence sampling also achieves comparable performance. This validate the positive role of the internal active query strategy.

In the next subsection, we will demonstrate the specific contribution of the dual active label correction strategy to estimating high quality transition matrix.

4.3 Comparison for estimating transition matrix

Figure 2 show the comparison between the estimation error of the three active sampling strategies: random, entropy and DALC. Note that the forward loss correction model is used for this experiment. The DALC is implemented using the entropy based confidence score, i.e., DALC-*e*. The estimation error is calculated using the relative L1 ratio, i.e., $|T - \hat{T}|/T$, with $\hat{T}$ denoting the estimated transition matrix, and T the groundtruth transition matrix.

It can be seen that in all cases, although the entropy sampling stratey achieves good performance for classification accuracy as shown in the last subsection, its estimation error is quite large compared with the random sampling strategy, almost two times the CIFAR-10 and MNIST dataset. With our dual active sampling strategy, when the high confident predictions from the classifier are incoporated, the estimation error for transition matrix reduces significantly, almost the same as that of random sampling, and even lower in some cases, e.g., CIFAR-100 and MNIST with noise rate 40%, 60%.

This result tells us two things: (1) in the noisy label learning field, simply adopting the standard active sampling strategies can lead to highly biased noise model estimation; (2) we can take advantage of the high confidence instances to combat with bias. However, such aspects are rarely considered in previous active label correction work. We believe this point would be worthy of more attention.

In Fig. 3, we show the training and testing loss of DALC-*e* for noise ratio 40%, 60% and 80% on CIFAR data during the training procedure. With early stopped 75 epochs training, DALC-*e* effectively converges without overfitting even for high noise ratio.

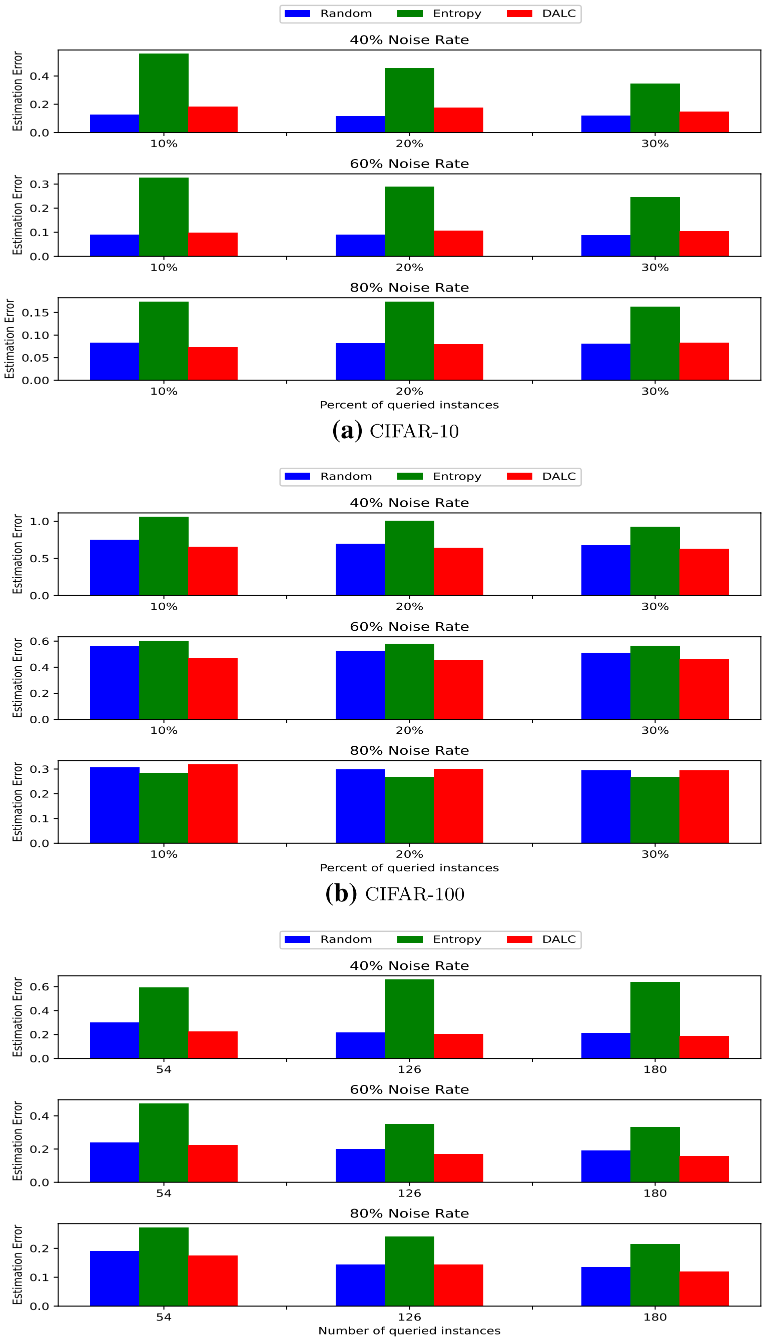
4.4 Additional experiment results

This subsection introduces more experimental results to better help understand the proposed DALC approach.

4.4.1 Comparison to state-of-the-art

Table 4 shows the test accuracy comparison with two state-of-the-art robust label noise learning methods DivideMix (Li et al., 2020), ELR (Liu et al., 2020) and their variants on CIFAR-10 and CIFAR-100. We report results for the symmetry flipping noise 40%, 60%, 80%.

DivideMix (Li et al., 2020) addresses noisy label learning in an iterative semi-supervised manner. It uses two networks to perform sample selection via a two-component mixture model (Arazo et al., 2019), and applies the semi-supervised learning technique



(a) CIFAR-10

(b) CIFAR-100

(c) MNIST

Fig. 2 The estimation error of the transition matrix on CIFAR-10, CIFAR-100 and MNIST. Results for three different numbers of actively queried instances in noise rate 40%, 60% and 80% by random, entropy, and DALC are shown. The estimation error is calculated as $|T - \hat{T}|/T$, with $\hat{T}$ denoting the estimated transition matrix, T the groundtruth transition matrix

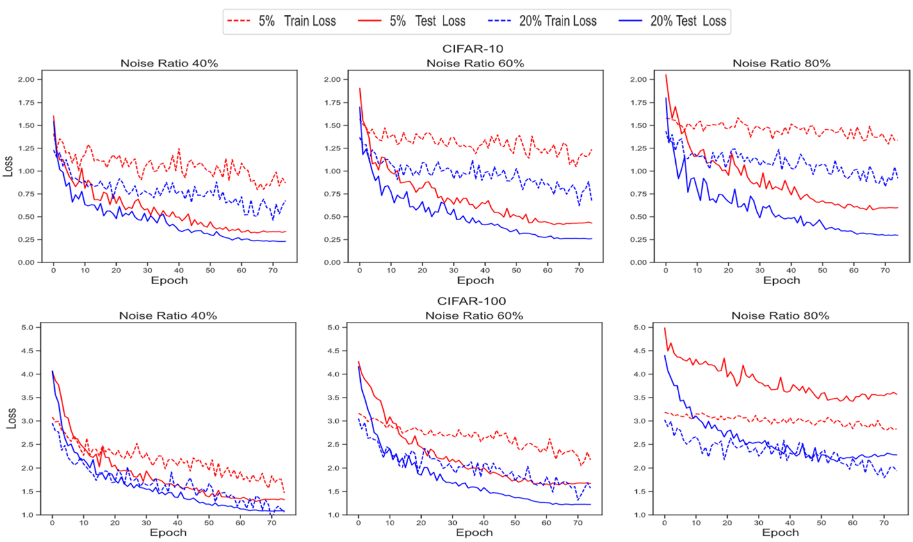


Fig. 3 The training and testing loss of DALC-*e* on CIFAR-10 and CIFAR-100 during the training epochs. Results for two different numbers of actively queried instances 5% and 20% in noise rate 40%, 60% and 80% are shown

Table 4 Test accuracy comparison on CIFAR-10 and CIFAR-100 with symmetry noises

Dataset	CIFAR-10			CIFAR-100		
Methods/noise ratio	40%	60%	80%	40%	60%	80%
DivideMix	0.942	0.945	0.932	0.757	0.704	0.551
DivideMix w/o co-tr	0.921	0.929	0.896	0.715	0.670	0.504
DivideMix w/o mixup	0.921	0.880	0.604	0.662	0.558	0.326
ELR	0.892	0.861	0.739	0.683	0.593	0.298
ELR*	0.914	0.889	0.807	0.684	0.601	0.303
DALC-e 10%	0.920	0.900	0.873	0.701	0.639	0.360
DALC-e 20%	0.934	0.924	0.910	0.727	0.691	0.481

MixMatch (Berthelot et al., 2019) with mixup data augmentation (Zhang et al., 2018). It was shown by numerous works that using two networks and mixup data augmentation significantly help learning (Han et al., 2018; Liu et al., 2020; Ortego et al., 2021). As our DALC approach falls in the iterative sample selection and learning process, but not enhanced with two networks and mixup data augmentation, we also compare with two variants of DivideMix: DivideMix w/o co-tr which uses one network and DivideMix w/o mixup which doesn't use data augmentation.

We have run DivideMix and its variants using its official implementations, as the original paper didn't report results for noise level $40\%, 60\%$. To get the best test performance during training, DivideMix adopts specific configurations for different noise ratios/noise types/datasets, which however are not publicly revealed. In this paper, we adopt the parametrization reported in Li et al. (2020) with default $\lambda_u = 25$ in the code. Results of ELR and ELR* (which uses cosine annealing learning rate for better performance) are take

Table 5 Test accuracy comparison on CIFAR-10 and CIFAR-100 with pair flipping noises

Dataset	CIFAR-10			CIFAR-100		
Methods/noise ratio	40%	60%	80%	40%	60%	80%
DivideMix	0.934	0.596	0.509	0.721	0.295	0.005
ELR	0.901	0.578	0.576	0.733	0.009	0.003
ELR*	0.904	0.606	0.587	0.737	0.006	0.002
Random 10	0.934	0.934	0.922	0.733	0.731	0.719
DALC-*e* 10	0.933	0.929	0.924	0.728	0.722	0.717

from Liu et al. (2020). Motivated by the early-learning and memorization phenomena, ELR proposed a regularization term to implicitly prevent memorization of the false labels. The ELR+ variant in Liu et al. (2020) enhanced by using two networks and mixup is not used for fair comparison. Note that on CIFAR benchmarks, ELR/ELR* used 10% of clean training data as validation set to choose hyperparameters, and DivideMix and its variants report the best test performance, thus we regard DALC with 10% queried true labels as a fair comparison setting. We append results of DALC with 10%, 20% true labels in Table 4 for clearer demonstration. It can be seen that while DivideMix performs the best, results of DivideMix w/o mixup data augmentation drop significantly, outperformed by DALC-*e* 10% at a large margin. When 20% true labels are queried, DALC performs better than or is comparable with DivideMix w/o co-tr in most cases except on CIFAR-100 with 80% noise rate. ELR/ELR* are consistently inferior to DALC. It is believed that when combining the proposed DALC with two networks and data augmentation, we can achieve much more performance improvement. We leave this for future work.

4.4.2 Pair flipping noise

Pair flipping noise models the fine-grained classification task where the class of clean label can flip into its adjunct class instead of far-away class. Existing label noise learning approaches normally studied this setting with relatively low noise rate (Li et al., 2020; Liu et al., 2020), e.g., 40%, because theoretically the flipped classes cannot be distinguished without additional assumptions for noise larger than 50%. We show in this paper that, with properly modeled label noise process, e.g., the noise transition matrix, with a small number of instances with queried clean labels, the flipped classes can be identified. Thus the label noise model can be accurately estimated with rather good learning performance, even in high noise rate cases.

Table 5 shows the test accuracy results for pair flipping noise 40%, 60%, 80% on CIFAR-10 and CIFAR-100. We have run DivideMix and ELR/ELR* using their official implementations for noise level 60%, 80%, and take the reported results for 40% from their papers. We have observed that for pair flipping noise, very few randomly selected clean instances are enough to identify the flipped classes and re-correct the estimated transition matrix, e.g., 10 instances for each class. Whereas different active sampling strategies make no much difference. Thus here we report results of selecting 10 instances for each class for Random and DALC-*e* strategies, and omit results of other active strategies.

It can be seen that DivideMix and ELR/ELR* fail to identify the flipped classes for high noise rate 60% and 80%, resulting in poor performance. However, with only 10 queried

Table 6 Test accuracy comparison on Clothing1M dataset

DivideMix	ELR	ELR*	Random 10%	Random 20%	DALC-*e* 10%	DALC-*e* 20%
0.644	0.589	0.62	0.602	0.652	0.662	0.699

correctly labeled instances for each class, even with 80% noise ratio, on CIFAR-10, Random achieves 0.922 test classification accuracy, which is only 0.019 absolute accuracy lower than the neural network classifier trained without label noise (0.941); on CIFAR-100, Random achieves 0.719 test accuracy, 0.023 lower than the neural network classifier trained without label noise (0.742). The proposed active DALC-*e* performs quite close to Random.

4.4.3 Realistic noise

We further corroborate the proposed method by considering the real-world dataset Clothing1M (Xiao et al., 2015), which contains 47, 570 clean labeled training data and 10^6 noisy labeled training data, with clean validation and test set respectively having 14, 313 and 10, 526 images. For the training data, a subset of 24, 637 images are tagged with both clean and noisy labels. In this experiments, we use this subset for training, and report accuracy on the clean test data. A pre-trained ResNeXt-50 on ImageNet Xie et al. (2017) is used as our backbone network. We train for 10 epochs using SGD according to the 1cycle learning rate policy (Smith & Topin, 2018) with initial learning rate 0.1 and maximum learning rate 0.01. For DivideMix, ELR and ELR*, their official implementations are used. Results for Random strategy and DALC-*e* with 10%, 20% true labels are reported. Table 6 shows the comparison results. Although DivideMix and ELR/ELR* use the additional 14, 313 clean validation data to get best performance, DALC-*e* consistently outperforms them and Random strategy.

5 Conclusion

In this paper, we consider the problem of deep label noise learning from the Active Label Correction (ALC) perspective, i.e., querying the true labels for some instances with minimal query costs and maximally improving the learning performance. A common limitation of existing ALC work is that, they ignore the fact that due to the active data sampling bias, the label noise model estimation can be seriously biased. Besides, they do not yet exploit the intrinsic fitting characteristics of deep learning models. We found that due to the memorization effect of deep neural networks, a large proportion of mislabeled instances with rather large loss can be correctly predicted, even under severe noise. Thus we propose one dual ALC (DALC) approach to select the most useful instances for classifier improvement and identify the most likely correctly predicted instances. The true labels of the two sources of instances are respectively queried from external human experts and the classifier predictions. Experiments on multiple datasets show that the proposed dual active query strategy is effective for both classification accuracy and combating the noise model estimation bias.

Author contributions S-YL, S-JH and SC contributed to the study conception and design. Material preparation, data collection and analysis were performed by S-YL and YS. The first draft of the manuscript was written by S-YL and all authors commented on the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by National Natural Science Foundation of China (Grant No. 61906089), National Key R&D Program of China (Grant No. 2020AAA0107000), Jiangsu Province Basic Research Program (Grant No. BK20190408) and China Postdoc Science Foundation (Grant No. 2019TQ0152).

Availability of data and material The datasets used in the study are publically available from their corresponding authors.

Code availability The code for this study are not publicly available until the paper is published, but are available from the corresponding author Shao-Yuan Li on reasonable request.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

Angluin, D., & Laird, P. D. (1987). Learning from noisy examples. *Machine Learning, 2*(4), 343–370.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2019). Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th international conference on machine learning* (pp. 312–321).

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. C., Bengio, Y., & Lacoste-Julien, S. (2017). A closer look at memorization in deep networks. In *Proceedings of the 34th international conference on machine learning* (pp. 233–242).

Berthelot, D., Carlini, N., Ian J. Goodfellow, N. P., Oliver, A., & Raffel, C. (2019). Mixmatch: A holistic approach to semi-supervised learning. In *Advances in neural information processing systems* (vol. 32).

Frenay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 25*(5), 845–869.

Goldberger, J., & Ben-Reuven, E. (2017). Training deep neural-networks using a noise adaptation layer. In *Proceedings of the 5th international conference on learning representations* (pp. 8778–8788).

Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., & Sugiyama, M. (2018). Masking: A new perspective of noisy supervision. In *Advances in neural information processing systems* (vol. 31, pp. 5841–5851).

Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I. W., Kwok, J. T., & Sugiyama, M. (2020). A survey of label-noise representation learning: Past, present and future. arXiv:abs/2011.04406

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., & Sugiyama., M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems* (vol. 31, pp. 8536–8546).

Hendrycks, D., Mazeika, M., Wilson, D., & Gimpel, K. (2018). Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in neural information processing systems* (vol. 31, pp. 10477–10486).

Horvitz, E. (2007). Reflections on challenges and promises of mixed-initiative interaction. *AI Magazine, 28*(2), 13–22.

Jiang, L., Zhou, Z., Leung, T., Li, L. J., & Fei-Fei, L. (2018). Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th international conference on machine learning* (pp. 2309–2318).

Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., & Fei-Fei, L. (2016). The unreasonable effectiveness of noisy data for fine-grained recognition. In *Proceedings of the 14th European conference on computer vision* (pp. 301–320).

Kremer, J., Sha, F., & Igel, C. (2018). Robust active label correction. In *Proceedings of the 21th international conference on artificial intelligence and statistics* (pp. 308–316).

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report.

LeCun, Y., Cortes, C., & Burges, C. J. (1998). The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/

Li, J., Socher, R., & Hoi, S. C. (2020). Dividemix: Learning with noisy labels as semi-supervised learning. In *Proceedings of the 8th international conference on learning representations*.

Li, M., Soltanolkotabi, M., & Oymak, S. (2020). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *The 23rd international conference on artificial intelligence and statistics* (vol. 108, pp. 4313–4324).

Li, S. Y., Huang, S. J., & Chen, S. (2021). Crowdsourcing aggregation with deep bayesian learning. *Science China Information Science, 64*(3)

Li, Y., Yang, J., Song, Y., Cao, L., & Li, L. (2017). Learning from noisy labels with distillation. In *Proceedings of 2017 IEEE international conference on computer vision* (pp. 1928–1936).

Liu, S., Niles-Weed, J., Razavian, N., & Fernandez-Granda, C. (2020) Early-learning regularization prevents memorization of noisy labels. In *Advances in neural information processing systems* (vol. 33).

Liu, T., & Tao, D. (2016). Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(3), 447–461.

Liu, Y., & Guo, H. (2020). Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th international conference on machine learning* (pp. 6226–6236).

Loshchilov, I., & Hutter, F. (2017). Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the 5th international conference on learning representations*.

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S. M., & Bailey, J. (2020). Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th international conference on machine learning* (pp. 6543–6553).

Nallapati, R., Surdeanu, M., & Manning, C. (2009). Corractive learning: Learning from noisy data through human interaction. In *IJCAI workshop on intelligence and interaction*.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., & Tewari, A. (2013). Learning with noisy labels. In *Advances in neural information processing systems* (vol. 26, pp. 1196–1204).

Ortego, D., Arazo, E., Albert, P., O'Connor, N. E., & McGuinness, K. (2021). Multi-objective interpolation training for robustness to label noise. In *Proceedings of 2021 IEEE/CVF conference on computer vision and pattern recognition* (pp. 6606–6615).

Patrini, G., Rozza, A., Menon, A. K., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of 2017 IEEE conference on computer vision and pattern recognition* (pp. 2233–2241).

Rebbapragada, U., Brodley, C. E., Sulla-Menashe, D., & Friedl, M. A. (2012). Active label correction. In *Proceedings of the 12th IEEE international conference on data mining* (pp. 1080–1085).

Reed, S. E., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., & Rabinovich, A. (2014). Training deep neural networks on noisy labels with bootstrapping. In *Proceedings of the 3rd international conference on learning representations*.

Ren, M., Zeng, W., Yang, B., Urtasun., & R. (2018). Learning to reweight examples for robust deep learning. In *Proceedings of the 35th international conference on machine learning* (pp. 4331–4340).

Samel, K., & Miao, X. (2018). Active deep learning to tune down the noise in labels. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 685–694).

Smith, L. N., & Topin, N. (2018). Super-convergence: Very fast training of neural networks using large learning rates. arXiv:1708.07120

Song, H., Kim, M., Park, D., & Lee, J. G. (2020). Learning from noisy labels with deep neural networks: A survey. arXiv:abs/2007.08199

Steinhardt, J., Koh, P. W., & Liang., P. (2017). Certified defenses for data poisoning attacks. In *Advances in neural information processing systems* (vol. 30, pp. 3517–3529).

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., & Fergus, R. (2015). Training convolutional networks with noisy labels. In *ICLR Workshop*.

Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition, pp. 5552–5560 (2018)

Urner, R., Ben-David, S., & Shamir, O. (2020). Learning from weak teachers. In *Proceedings of the 15th international conference on artificial intelligence and statistics* (pp. 1252–1260).

van Rooyen, B., Menon, A. K., & Williamson, R. C. (2015). Learning with symmetric label noise: The importance of being unhinged. In *Advances in neural information processing systems* (vol. 28, pp. 10–18).

van Rooyen, B., & Williamson, R. C. (2017). A theory of learning with corrupted labels. *Journal of Machine Learning Research, 18*(1), 8501–8550.

Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., & Belongi, S. J. (2017). Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of 2017 IEEE conference on computer vision and pattern recognition* (pp. 6575–6583).

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., & Sugiyama, M. (2019). Are anchor points really indispensable in label-noise learning. In *Advances in neural information processing systems* (vol. 32, pp. 6835–6846).

Xiao, T., Xia, T., Yang, Y., Huang, C., & Wang, X. (2015). Learning from massive noisy labeled data for image classification. In *Proceedings of 2015 IEEE conference on computer vision and pattern recognition* (pp. 2691–2699).

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of 2017 IEEE conference on computer vision and pattern recognition* (p. 5987).

Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., & Sugiyama, M. (2020). Dual t: Reducing estimation error for transition matrix in label-noise learning. In *Advances in neural information processing systems* (vol. 33).

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. In *Proceedings of the british machine vision conference 2016*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. In *Proceedings of the 5th international conference on learning representations*. arXiv:1611.03530

Zhang, H., Ciss'e, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In *Proceedings of the 6th international conference on learning representations*.

Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems* (vol. 31, pp. 8792–8802).