Noisy Label Learning in Deep Learning

Xuefeng Liang, Longshan Yao, Xingyu Liu

Noisy Label Learning in Deep Learning *

Xuefeng Liang[1], Longshan Yao[1], and XingYu Liu[1]

1.School of Artificial Intelligence, Xidian University, China

Abstract. Currently, the construction of a large-scale manual annotation databases is still a prerequisite for the success of DNN. Although there is no shortage of data, there is a lack of clean label data in many fields, because it takes a lot of time and huge labor costs to build such a database. As many studies have shown that noisy label will seriously affect the stability and performance of the DNN. Learning from noisy labels has become more and more important, and many methods have been proposed by scholars. The purpose of this paper is to systematically summarize the different ideas for solving the noisy label learning problem, analyze the problems with existing methods, and try to analyze how to solve these problems. First, we will describe the problem of learning with label noise from the perspective of supervised learning. And then we will summarize the existing methods from the perspective of dataset usage. Subsequently, we will analyze the problems with the data and existing methods. Finally we will give some possible solution ideas.

Keywords: DNN · Noisy label · Large-scale manual annotation dataset.

1 Introduction

In recent years, the good performance of deep learning and its successful application in many fields greatly depend on the establishment of large-scale manual label databases, such as ImageNet [5], etc. However, the construction of large-scale databases is time consuming and labor-intensive. Therefore, there are still many fields that lack large-scale and reliable databases, which hinders the development and application of deep learning. In order to overcome the difficulty of large-scale database construction, scholars have proposed many low-cost alternatives in recent years. For example, we can collect a large amount of data from search engines or social media and then label the data based on surrounding text and tags, or using Amazon's Mechanical Turk and so on. The datasets obtained by these methods inevitably contain a large number of noisy labels. Recent studies have shown that deep neural networks will overfit with noisy label data [34, 1], thereby affecting the generalization performance and stability of the network. In addition, the data in some fields has strong ambiguity and is extremely difficult to be labeled. Thus, it's also inevitable to generate noisy labels, such as speech emotion, lip language, and facial aesthetics.

Noisy label learning has been one of the core areas of deep learning and scholars have proposed a large number of solutions to address this problem.

* This work was supported by the Guangdong Provincial Key Research and Development Programme under Grant 2021B0101410002.

Each of these methods has its own advantages and disadvantages as well as specific constraints. This paper hopes to systematically expatiate the different approaches to solve the problem of noisy label learning, discuss the problems that may be neglected, and try to give some ideas about solving the problems.

Firstly, we will explain the preliminary knowledge in Section 2; in Section 3, we will try to expatiate the various ideas and methods for solving the noisy label learning problem; in Section 4, the potential problems in the existing methods and possible solutions will be discussed; Section 5 for summary.

2 Preliminary Knowledge

In order to facilitate the reader's understanding, this section will introduce the relevant terms used in this paper and briefly clarify related issues.

2.1 Noisy labels in deep learning

The goal of deep learning tasks under supervised learning is to learn a mapping function f from dataset $D = \{(x_i, y_i)|i = 1, 2, 3 \ldots n\}$, where the parameters of the mapping $f : x_i - y_i$ are expressed as θ, which can also be called the parameters of the neural network, f called the classifier. Look for the best classifier, by calculating the loss $L(f(x_i, \theta), y_i)$ of each sample data, and find the best mapping θ^* by optimizing the cost function.

$$R_D(f(\theta)) = \sum_{i=0}^{n} L(f(x_i, \theta), y_i) \tag{1}$$

When the dataset D contains noisy labels $(x_i, \widetilde{y}_i)$, $\widetilde{y}_i \neq y_i$, the sample loss $L(f(x_i, \theta), \widetilde{y}_i)$ cannot truly represent the loss of the sample. So the parameters $\widetilde{\theta}_i$ obtained by optimizing the cost function is not the best parameters, where $\widetilde{\theta}^* \neq \theta^*$.

2.2 Noisy label dataset and noisy label types

Synthetic dataset The types of simulated noise label: Pair noise [7], Symmetric noise [27], Asymmetric noise [29], as Fig.1. Applying the simulated noise type to the basic dataset, such as CIFAR-10, CIFAR-100, MNIST, etc., generates new dataset containing noisy label which called Synthetic dataset.

Pair noisy label dataset [7] Pair noise is the transfer of labels between two adjacent classes. As shown in the left of Fig.1, Pair noise with a noise ratio of 45%, the first class of data retains 55% of the total, and the remaining 45% of the data is labeled as the second class, and so on. (The row represents the real category, and the column represents the label category.)

Symmetric noisy label dataset [27] Symmetric noise is to keep a certain proportion of the main class label unchanged, and the rest is uniform distributed to other classes. As shown in the mid of Fig.1. The first class of data retains

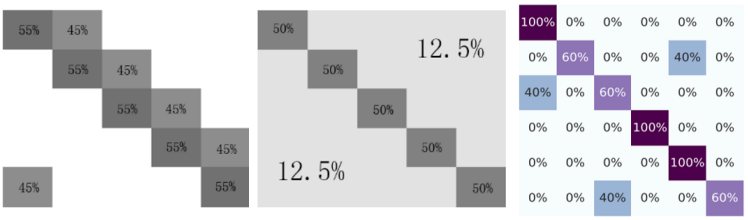

Fig. 1. Example of noise transfer matrix (From left to right are Pair noise, Symmetric noise, and Asymmetric noise.)

50% of the whole, and the remaining 50% of data are equally divided into the remaining 4 classes, and so on.

Asymmetric noisy label dataset [29] Asymmetric noise transfers the label according to the provided noise transfer matrix. As shown in the right of Fig.1, the first class of data is not transferred to other classes; the sixth class of data is retained at 60% of the whole, and the remaining 50% of data is changed to the third class, and so on.

Actual dataset The datasets containing noisy labels collected in the real world are called Actual datasets: Clothing1M [30], Webvision [16], Food101 [4], etc.

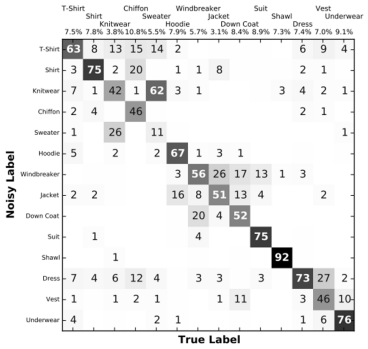

Fig. 2. Noise transfer matrix of Clothing1M.

Fig. 3. Example from Clothing1M.

Clothing1M [30] The database builder used crawler technology to collect 14 categories of pictures on several shopping websites, gave them labels through the description text around the pictures. More than 1 million data is collected and the correct ratio of labels is about 61.54%. The noise transition matrix is shown in Fig. 2.

As shown in Fig. 2, the noisy label composition of Clothing1M is very complicated, and there are many categories similar to Pair noise type, such as Wind-

breaker, Chiffon and Shirt, etc. Noise similar to Symmetric noise type are Hoodie, T-Shirt, Dress And Vest. There are also three types of analogue noise that do not fall into the three categories at all, for example, almost 1% of the samples in each category are assigned to the other category. In addition, there are noise forms that deviate from normal values, such as Knitwear and Sweater.

Webvision [16] There are two versions of Webvision, both of which are collected on the Flickr website and Google Image Search through crawler technology. The first version has 1000 categories and uses the same 1000 concepts as ILSVRC2012 for querying images, and the label is the keyword used in the query. The total number is 2.4 million. In the second version, the number of visual concepts was expanded from 1,000 to 5,000, and the total number of images in the training set reached 16 million. The first version is mostly used in existing papers.

The Webvision database is a typical long-tail dataset, where the amount of data for each category ranges from 300 to 10,000, as shown in Fig.4. To explore the noise form of the data, the builders took 200,000 photos, 200 of each category, and posted the task on Amazon's Mechanical Turk (AMT). The user is required to tell the label of each image is correct or not. Each image is judged by three users. If more than two users find it is correct, it is the correct label data. Finally, statistics (Fig.5) show that 20% of the data is real noise (0 votes), and 66% have the correct label (2 votes or 3 votes).

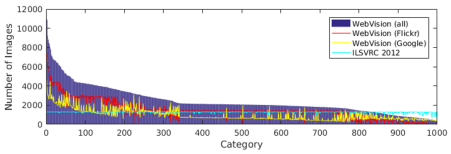

Fig. 4. Number of each class in Webvision

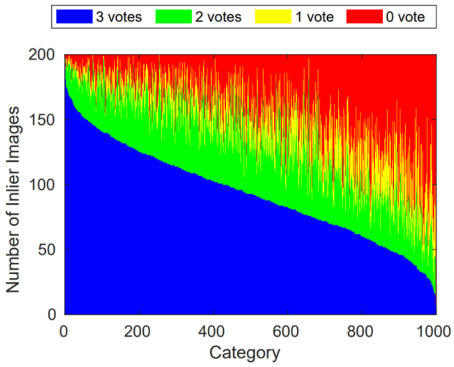

Fig. 5. The result of manual voting in Webvision. Samples with more than 2 votes are considered clean labels, with 0 votes are noise labels.

2.3 Analysis the problems in noisy label learning

Mathematical representation: The theoretical performance of the neural network $f(\theta)$ with the parameter θ on the dataset D can achieve : $P_D = f(\theta, D)$.

The theoretical performance on different sub-dataset:

$$P_{D_{all}} \leq P_{D_{clean}} < P_{D_{clean}+D_{unlabel}} < P_{D_{all\_clean}} \tag{2}$$

Where, $D_{all} = Shuffle\{D_{clean}, D_{noise}\}$, D_{clean} is the clean label data subset, D_{noise} is the noisy label data subset, $D_{unlabel}$ is the unlabeled data subset after D_{noise} deletes the label, and D_{all_clean} is the union of D_{all_clean} and D_{clean} after D_{noise} is given the true label.

3 Existing Methods of Noisy Label Learning

In this section, combined with the analysis of the problems with noisy label learning in Section 2.3, we summarize the existing methods into three categories from the perspective of how to use dataset: Full-equal-using method, Clean-based method and Full-differ-using method.

3.1 Full-equal-using method

This type of methods do not divide the noisy label data and the clean label data during training, and treat all the data equally. Typical methods include estimating noise transfer matrix, designing noise robust loss function, etc.

Estimating the noise transfer matrix [18, 23] is to correct the model prediction and adjust the loss function so that its gradient can descend in the right direction. A further method, Dual T [32], converts the estimating noise transfer matrix into two steps. First, it will estimate the transfer matrix from clean label to intermediate category label (network prediction label). Then, estimating the transfer matrix from intermediate category label to noisy label. It uses two consecutive matrices to simplify the problem. This kind of method generally requires a clean dataset to estimate the noise transfer matrix. In addition, the SIGUA [6] methods also attempt to solve the noise label learning problem from the perspective of correcting the gradient back propagation, which can adjust the gradient of the clean label data in each batch to reduce the learning rate of the noisy label data gradient.

The goal of designing robust loss function method is that the loss can be effectively calculated for the clean label data, and the influence of the loss of noisy label data can be controlled within a certain range. In the paper [36], the author finds that Mean Absolute Error (MAE) deals with equal weight to each category, while Cross Entropy (CE) only focuses on the current category, leading to preference for hard samples. Therefore, for noisy labels, the MAE is more robust than the CE, but the CE loss has higher accuracy and fitting speed than MAE. Based on the above analyses, the author proposes GCE loss by combining the advantages of MAE and CE. Different from the analysis of the adaptive characteristics of various loss functions, Symmetric loss [28] is proposed by focusing on the prediction results of different classes. The author found that DNN learning with CE can be **class-biased** on clean label data and **under**

learning on noise label data. To balance this problem, the author puts forward the idea of symmetric Cross Entropy loss. 'Symmetric' means weighted sum of the Cross Entropy loss and inverse Cross Entropy loss (label and predicted value are interchanged). APL [19] divides the existing loss functions into 'Active' and 'Passive', and proposes the 'activate passive learning' method to use the advantages of different types loss functions at different training stages. Different from the distance-based loss function, L_DMI [31] based on information entropy is not only theoretically prove robust to instance-independent label noise, but also easily apply to any classification network.

Designing noise-robust network structure, adding a specific layer or branch to the network to deal with noisy labels has the advantage of very targeted, but the network is not extended well. CleanNet [13], in the training process, abstracts the paradigm of each category by clustering method according to clean data, which will be used as the similarity measurement calculating loss during training. MetaCleaner [35] divides the training process into two steps. The first step is to estimate the confidence of each data. The second step is to generate a clean abstract training sample by summarizing the confidence score of the data. By classifying noise types, ClothingNet [30] can calculate the posterior probability of the noise type after the noise type is learned. Self-learning [8] adds additional clustering modules to the network, maintains the abstract category features generated by the clustering module in the training process, and gives corresponding pseudo-labels through the similarity between the training data and the abstract category features.

3.2 Clean-based method

This type of method only uses clean label subset during training, and the key problem is to improve the classification accuracy between clean label data and noisy label data. Once the subset of cleanly labeled data is divided, the model can be trained according to the normal dataset, so that the influence of noisy labeled data can be avoided. The main directions of this type of method are: screening clean label data and adjusting the weight of samples. Currently, The methods used to identify noisy label and clean label data include: Small-Loss criterion, Gaussian mixture model(GMM), Bayesian mixture model(BMM), etc.

Small-Loss criterion: Data with a small training loss is considered clean data.

Decouple [20] first propose the idea of decoupling the problem of 'how to update' from 'when to update' in the training process, and give an update strategy based on network inconsistent information. This method first initializes two networks randomly, and in the subsequent training process, only when the two networks have a disagreement, the reverse update is performed. Co-teaching [7] is a representative method of dealing with noisy labels based on Co-training [2] ideas. During the training, the two networks calculate the loss of the same batch data and make a ranking. Then, according to the Small-Loss Criterion, the samples with Small loss value are selected and regarded as clean label data, and then they are passed to each other for gradient update. Mandal et al. [21] adds the idea of self-supervision on the basis of Co-teaching to improve the

accuracy of clean label recognition, and then increases the proportion of clean labels in each iteration and improves the model performance. Also on the basis of Co-teaching, Jo-CoR [29] introduces the idea of 'agreement' into two networks (under the same database, different models will reach agreement on most sample labels. In addition, different models trained on the same view are unlikely to reach agreement on incorrect labels [24, 12]), which improves the accuracy of data screening. The author adds contrast loss (JS divergence) between the two networks to realize the 'agreement' of the two networks, and filters the clean label data according to the Small-Loss criterion. MentorNet [10] uses the idea of curriculum learning (referencing from human learning mode) to implement the sequential learning of the model from easy to difficult in data that contain noisy label through two networks (a teacher network and a student network).

3.3 Full-differ-using method

This type of method can be regarded as a further extension of Clean-based method. There is another reason that when the noise ratio is higher than a certain threshold, the clean label data in the training data is not enough for the network to solve the problem, the noise label data should be considered.

The idea of correcting noisy labels comes from the the theory that network has the possibility of self-correcting. Based on this idea, PENCIL [33] treats the pseudo-label (corrected label of the network) as an independent parameter and updates it during the training process as the network parameter, so as to obtain a correct pseudo-label. However, this method sets all the training data as correctable, leading to the fact that in the training process, the network often 'corrects' clean labels into false labels by mistake. With the similar idea, Joint-optimization [25] use a single network to realize the function of training and correcting noisy labels, and regularize the noise sample loss after correction to reduce the impact of errors.

As for the method of designing robust network structure, there are also some methods divide the clean label and noisy label data and then use them separately during training. For example, based on the idea of weakly supervised learning [9], a residual network branch is specially designed to process noisy label data and maintain noisy label and clean label subsets continuously to train the network. Mean Teacher [26] based on knowledge distillation method and Meta-learning based on meta learning method [15]. These methods are similar to Mentornet [10], which requires a teacher network to provide soft label which is the label for student network to learn. Differently, it will consider to divide clean label and noisy label data in the process and only correct the noisy label. Other methods, such as NLNL [11], enhance the use of noisy label data through 'negative learning'. Its core idea is that if the label of the sample is wrong, it can be sure that the sample does not belong to the labeled category. This information is also useful, which can assist network training and improve network performance.

The type of method based on semi-supervised learning treats the noisy label data as unlabeled data and uses the semi-supervised learning method to deal with it. Because there are many effective algorithms for semi-supervised learning,

such methods usually only need to consider how to optimize the partitioning of training data. DivideMix [14] also refers to the ideas of the two models in Co-training, using Gaussian mixture model to select clean samples and noisy label samples, and treat the noisy label samples as unlabeled data, combined with the excellent algorithm MixMatch [3] in semi-supervised learning for training; ERL [17] found the characteristics of early learning of the network, and added regular means to prevent the network from remembering noisy labels.

4 Problems in Existing Methods

4.1 Difference between Synthetic dataset and the Actual dataset

Difference in the size: The basic dataset used in the Synthetic dataset only contains 60,000 on average, and the pixel size is within 32*32; in the contrast, the Actual dataset is generally more than 1 million, and the pixel size is more than 224*224.
Difference in noise type: Only have three types of synthetic noise, Pair noise [7], Symmetric noise [27], Asymmetric noise [29], and there are some regularities in them. AS for Actual datasets such as Clothing1M, it's noise transfer matrix(Fig.2) is very complicated than synthetic noise dataset.
Differences in the data itself: The basic databases of Synthetic datasets such as CIFAR-10, CIFAR-100, MNIST always have single and prominent objects in the pictures, while the complexity of the data in the Actual dataset is far greater than the above three, such as shown in Fig.3.
Difference in frequency of use: Through the statistics of the databsets used in the experiment of 47 related papers, the frequency of each databset is obtained, as shown in Table 2. We find that CIFAR-10 is used as the baseset to generate Synthetic datasets with a frequency of 75%, CIFAR-100 is used at 62%, and MNIST is used at 36%. According to statistics, the usage ratios of the three Actual datasets are 55% of Clothing1M, 12.7% for both Webvision and Food101.

Table 1. Dataset usage frequency statistics (Only shown used more than 2 times).

Dataset	CIFAR-10	CIFAR-100	minist	ILSVRC	News	F-minist	Clothing1M	Food101	Webvision
Time	35	29	17	6	3	6	26	6	6
Frequency	75%	62%	36%	13%	6%	13%	55%	13%	13%

4.2 Problems with existing methods

Due to space limitations, we only show some results of representative methods, as shown in Table 2. CE (Cross Entropy Loss) is a benchmark for comparison.

It can be seen from Table 2 that the improvement of the current optimal method comparing to the benchmark method is as follows: on CIFAR-10 and CIFAR-100 with noise ratio of 20%, the improvement of 10% and 15% is achieved respectively, and the optimal results are 96.1% and 77.3%, respectively. With noise ratio of 50%, the improvement of 15% and 28% is achieved respectively, and the optimal results are 94.6% and 74.6%. With noise ratio of 80%, the improvement of 31% and 40% were obtained, respectively, and the optimal results

Table 2. Experimental results (Symmetrical noise sym., Asymmetric noise asym., %).

Noise ratio-Data set-Noise type	CE	F-correction [22]	Co-teaching [7]	PENCIL [33]	Meta-Learning [15]	DivideMix [14]	ELR [17]
20%- ciar10-sym.	86.80	86.80	89.50	92.40	92.90	96.10	94.60
50%- ciar10-sym.	79.40	79.80	85.70	89.10	89.30	94.60	93.80
80%- ciar10-sym.	62.90	63.30	67.40	77.50	77.40	93.20	91.10
20%- ciar100-sym.	62.00	61.50	65.60	69.40	68.50	77.30	77.50
50%- ciar100-sym.	46.70	46.60	51.80	57.50	59.20	74.60	72.40
80%- ciar100-sym.	19.90	19.90	27.90	31.10	42.40	60.20	58.20
40%-ciar10-asym.	83.20			88.50		93.40	92.70
40%-ciar100-asym.						72.10	76.50
Clothing1M	69.21	69.84			73.47	74.76	74.81
Webvision		61.12	63.58			77.32	77.78

were 93.6% and 60.2%, respectively. There is a 5% improvement on the Clothing1M comparing to the benchmark method, and the best result is 74.81%. The current optimal result on the Webvision dataset is 77.78%, and the noise label in the Webvision accounts for about 34%. Due to the lack of experimental results of the benchmark method on this dataset, we use the F-correction [22] method as the evaluation benchmark. Further statistics are made based on Table 2, as shown Table 3. Combining Table 2 and Table 3, it is easy to find the problems existing in the current methods:

The performance improvement of existing methods on Actual datasets is much lower than on Synthetic datasets

Table 3. The improvement of the best results on each dataset compare to the benchmark.

Noisy label dataset	Noise ratio	Increase
CIFAR-10 sys.	20%	10%
	50%	15%
	80%	31%
CIFAR-100 sys.	20%	15%
	50%	28%
	80%	40%
CIFAR-10 asys.	40%	10%
Clothing1M	38%	5%
Webvision	34%	16%

Detailed analysis: On the Synthetic dataset, when the noise rate is 20%, the results of the three type methods are similar. At this time, the clean label data still has a large proportion of the training data, and the three methods have no significant difference in the amount of available data. When the noise ratio continues to rise to 50%, the gap between the F-correction method and other methods gradually widens. This is because as the proportion of noisy label data increases, this method does not distinguish between clean and noisy labels, resulting in the network being gradually affected by noise. Although the performance of Co-teaching method is not much different from PENCIL and other methods, the gap between it and DivideMix and ELR methods is gradually in-

creasing. This is because DivideMix and ELR consider the correction and use of noisy label data by MixMatch and other methods. Effective use of noisy label data, increased training data. This phenomenon also exists in the experimental results with the noise ratio of 80%.

Clothing1M is the most frequently used Actual dataset in paper, and its noise label accounts for about 38%. Through analysis, we can find that the performance of Meta-Learning, which has a quite gap with the optimal method on Synthetic dataset, is similar with the optimal method on Clothing1M. Combined with further analysis, we still found that the performance of Full-equal-using method(such as F-correction) on the Actual dataset is far worse than Clean-based method and Full-differ-using method, which is also in line with the findings on the Synthetic dataset. Compared with the benchmark results on the Clothing1M and Synthetic datasets, we found that the benchmark method result on the Webvision is 61.12%, which does not reach 66% (100% -34%). The reason for this phenomenon is probably because Webvision is a long-tailed dataset. Although the benchmark method has the ability to deal with noisy labels, it does not have the ability to deal with the imbalance of the data itself.

4.3 Possible solutions

This section will put forward some feasible solutions for the series of method problems summarized above.

More analysis of the characteristics of noise in Actual datasets: Through the data analysis in Section 4.1, we can clearly know that the complexity of the noisy label in the Actual dataset is much higher than the Synthetic dataset in all dimensions. For example, Clothing1M include both Symmetric noise type and Pair noise type, while Synthetic datasets only have one type of noise. Webvision, the vote count of 0 is considered as noisy label, and the vote of 2 and 3 is clean label. For data with one vote, we should do more analysis instead of doing nothing. Therefore, to solve the problem of noisy labels, it is necessary to increase the analysis of the characteristics of different noisy labels in the dataset.

Adjust accordingly with the dataset: Because the actual data collected in the real world has a wide variety and inconsistent complexity. Even without considering the noisy label, the sample complexity in the Actual dataset is much higher than that in the Synthetic dataset. As shown in Fig.3, in Actual datasets, even among samples of the same category, there are usually large differences. Webvision is a typical long-tail dataset. When dealing with this kind of noisy dataset, increasing the processing capacity for long-tail data will make the algorithm more effective.

Improve the accuracy of dividing between clean labels and noisy labels: Comparing the results from perspective of data usage, we found that: Full-equal-using methods are very sensitive to the noise in training set. When the noise ratio is larger, it is difficult to exclude the influence of the noisy label on the model. The performance of the Clean-based methods greatly depend on the purity and quantity of the clean label data subset. It can be seen from the experimental results that the better of subset division, the better results can get.

Full-differ-using method is the more stable and more effective method among the three methods. How to effectively use noisy label data greatly affects the final result.

At present, the commonly used distinguishing methods are: Small-loss, GMM, BMM, etc. These methods all rely on the calculated loss value, but the loss value is deviated from the actual loss value, so that the above methods cannot completely eliminate the influence of noise labels. Therefore, more attention should be paid to how to divide clean label and noisy label data.

Pay attention to how to use noisy labels data: Comparing the experimental results in Section 4.2, we can see that in the case of using the same basic method, due to the different ways of using noisy label data, the performance of each method on the Actual datasets appears to be quite different, such as PENCIL and DivideMix.

Increase training epoch: Through the analysis of all the papers using Actual datasets, we found that the average epoch of network training is not less than 200 in experiment of Synthetic datasets. And in the experiment of the Actual dataset, the training time is much longer than the Synthetic dataset due to the huge amount of data, the average epoch of training is only about 20. Due to the greatly reduced training epochs, the test results obtained may be incorrect.

5 Conclusion

This paper summarizes the current methods form the perspective of how to use data, analyzing the existing problems, and giving some potential solutions. At present, due to the efforts of many scholars, the noise label learning methods have been able to achieve satisfactory results on three types of Synthetic datasets. However, these methods still have some deficiencies in Actual datasets, which need to be further studied.

References

1. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International Conference on Machine Learning. pp. 233–242. PMLR (2017)
2. Balcan, M.F., Blum, A., Yang, K.: Co-training and expansion: Towards bridging theory and practice. Advances in neural information processing systems **17**, 89–96 (2005)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019)
4. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Han, B., Niu, G., Yu, X., Yao, Q., Xu, M., Tsang, I., Sugiyama, M.: Sigua: Forgetting may make learning with noisy labels more robust. In: International Conference on Machine Learning. pp. 4006–4016. PMLR (2020)
7. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. arXiv preprint arXiv:1804.06872 (2018)
8. Han, J., Luo, P., Wang, X.: Deep self-learning from noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5138–5147 (2019)
9. Hu, M., Han, H., Shan, S., Chen, X.: Weakly supervised image classification through noise regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11517–11525 (2019)
10. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference on Machine Learning. pp. 2304–2313. PMLR (2018)
11. Kim, Y., Yim, J., Yun, J., Kim, J.: Nlnl: Negative learning for noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 101–110 (2019)
12. Kumar, A., Saha, A., Daume, H.: Co-regularization based semi-supervised domain adaptation. Advances in neural information processing systems **23**, 478–486 (2010)
13. Lee, K.H., He, X., Zhang, L., Yang, L.: Cleannet: Transfer learning for scalable image classifier training with label noise. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5447–5456 (2018)
14. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)
15. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Learning to learn from noisy labeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5051–5059 (2019)
16. Li, W., Wang, L., Li, W., Agustsson, E., Van Gool, L.: Webvision database: Visual learning and understanding from web data. arXiv preprint arXiv:1708.02862 (2017)
17. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Advances in Neural Information Processing Systems **33** (2020)

18. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence **38**(3), 447–461 (2015)
19. Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., Bailey, J.: Normalized loss functions for deep learning with noisy labels. In: International Conference on Machine Learning. pp. 6543–6553. PMLR (2020)
20. Malach, E., Shalev-Shwartz, S.: Decoupling" when to update" from" how to update". arXiv preprint arXiv:1706.02613 (2017)
21. Mandal, D., Bharadwaj, S., Biswas, S.: A novel self-supervised re-labeling approach for training with noisy labels. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1381–1390 (2020)
22. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1944–1952 (2017)
23. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: International Conference on Machine Learning. pp. 4334–4343. PMLR (2018)
24. Sindhwani, V., Niyogi, P., Belkin, M.: A co-regularization approach to semi-supervised learning with multiple views. In: Proceedings of ICML workshop on learning with multiple views. vol. 2005, pp. 74–79. Citeseer (2005)
25. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5552–5560 (2018)
26. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)
27. Van Rooyen, B., Menon, A.K., Williamson, R.C.: Learning with symmetric label noise: The importance of being unhinged. arXiv preprint arXiv:1505.07634 (2015)
28. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 322–330 (2019)
29. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13726–13735 (2020)
30. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2691–2699 (2015)
31. Xu, Y., Cao, P., Kong, Y., Wang, Y.: L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In: NeurIPS. pp. 6222–6233 (2019)
32. Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., Sugiyama, M.: Dual t: Reducing estimation error for transition matrix in label-noise learning. arXiv preprint arXiv:2006.07805 (2020)
33. Yi, K., Wu, J.: Probabilistic end-to-end noise correction for learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7017–7025 (2019)
34. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 (2016)
35. Zhang, W., Wang, Y., Qiao, Y.: Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7373–7382 (2019)

36. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. arXiv preprint arXiv:1805.07836 (2018)