Robust Training Deep Neural Networks with Noisy Labels: A Survey

Sixu Li

1 Background Introduction

Deep neural networks have been widely used for many complex supervised learning tasks because of their strong learning abilities. However, this powerful learning capacities also brings an obvious drawback, i.e., DNNs can memorize random noise in the training process. And meanwhile, in practice, it is expensive to obtain large clean datasets because of the highly cost of manually annotation. On the contrary, data with noisy labels are easy to get. Therefore, how to train Deep Neural Networks with noisy labels is a challenging but worthy research direction.

2 Noise Generation

Most commonly, there are two kinds of classical label corruptions generated by noise transition matrix, where the entry p_{ij} represents the probability of flipping label of instance from class i to class j:

(1) Symmetry Flipping: ground-truth label y has the same probability corrupted into any other class labels and the noise transition matrix can be written as:

$$Q = \begin{bmatrix} 1-\epsilon & \frac{\epsilon}{n-1} & \cdots & \frac{\epsilon}{n-1} & \frac{\epsilon}{n-1} \\ \frac{\epsilon}{n-1} & 1-\epsilon & \frac{\epsilon}{n-1} & \cdots & \frac{\epsilon}{n-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \frac{\epsilon}{n-1} & \cdots & \frac{\epsilon}{n-1} & 1-\epsilon & \frac{\epsilon}{n-1} \\ \frac{\epsilon}{n-1} & \frac{\epsilon}{n-1} & \cdots & \frac{\epsilon}{n-1} & 1-\epsilon \end{bmatrix} \tag{1}$$

(2) Pair Flipping: ground-truth label y may only be corrupted into one kind of noisy label $\widetilde{y}$, which may quite similar to the origin class label y. For instance, human annotators

may easily regard "dog" as "cat" when the images are not clear. And in this circumstance, noise transition matrix is given by:

$$Q = \begin{bmatrix} 1-\epsilon & \epsilon & 0 & \cdots & 0 \\ 0 & 1-\epsilon & \epsilon & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1-\epsilon & \epsilon \\ \epsilon & 0 & \cdots & 0 & 1-\epsilon \end{bmatrix} \tag{2}$$

3 Summary of Current Papers

3.1 Noise Model Based Methods

Methods in this section try to model the behavior of noise and utilize this information during training for better performance.

3.1.1 Noisy Channel

By modeling the noise transition matrix, true class probabilities of data can be extracted and fed into classifier for robust noise training. Along this direction, different methods used to find the transition matrix were proposed.

- **Add an extra layer:** [27] add a linear fully connected layer on the top of the base classifier and use this layer to approximate the true noise transition matrix. In [12], additional softmax layer, which simplifies the optimization process comparing to full connected layer, is added and meanwhile, use dropout regularization to prevent the base model from learning the noisy labels directly.

- **Using another network:** [34] proposed a new parameter, called quality embedding, which tells the trustworthiness of data and can be estimated by a neural network.(**It is interesting to figure out that how they define the trustworthiness of data**).

- **Explicit Calculation:** In [5], EM algorithm is used to iteratively train network to match given distribution and estimate noise transition matrix.In [1], they argued that when there are a mixture of correct and mislabelled targets, networks tend to fit the former before the latter. Based on this observation, a beta mixture that estimates the probability of a sample being mislabelled and corrects the loss by relying on the network

prediction(namely, boostrapping loss) is proposed. And this methods also uses EM algorithm to train the model iteratively. **However, because EM algorithm need to estimate in iteration, it is hard to implement this method with DNNs with large data sets.** [10] assume they have some clean data and first train networks on the noisy data and then make preditions on trusted data. And the softmax probabilities were used to construct noise transition matrix. (I think detailed reading is needed in order to find the drawback of this method.)

In this Noisy Channel direction, I think methods proposed by[34, 5, 1, 10] are intriguing and deserve more detailly reading.

3.1.2 Label Noise Cleansing

An intuitive method to deal with noisy labels is to remove samples with suspicious labels or correct their their corrupted label to corresponding true class. In this direction, following three papers seem interesting and can read in detail.

- **Joint Optimization Framework:** [28, 35] utilize joint optimization framework for both training classifer and updating noisy labels. Their work consists of two progressive steps(like EM algorithm) 1) Fix labels and update classifier's parameters with SGD. 2) Fix network's parameters and update labels.

- **Cluster Based Method:** In [8], by clustering the deep feature, of data with same class label, extracted by neural network, they construct prototypes in each class and regard them as trusted data. Then, corrected label is found by checking similarity among the data sample and prototypes. **One drawback I guess: In this framework, they should cluster the deep feature in every training iterations, which has heavy workloads and leads to slow training speed.**

3.1.3 Sample Choosing

Another natural idea to combat corrupted labels is that fitering out the noisy labels and training DNNs only on the selected clean samples. **Hence, In my opinion,in this direction, the main problem to think about is that what are the different characteristics between clean data and corrupted data. In other word, how can we utilize the output of DNNs to distinguish this two kinds of data.** Several approaches were proposed:

- **Consistency of Label and Model Prediction:** [25, 20] argued that if both label and model prediction of the given sample are consistent, then this data can be regard as trust data.

- **Complexity of Sample:** In [6], they used the distribution density of data in the feature space to represent the complexity of it and samples with lower complexities are more likely to be clean.

- **Sample Loss:** In, [7, 11], they consider the sample with small loss as clean samples. **we can think about what's disadvantage of using loss and under what circumstances this measure may fail.**

In these above methods, they both utilize an empirical property of DNNs, i.e., DNNs will first learn from the easy pattern(clean data) and then gradually fit the noise. **I think this is an interesting property related to the generalization process of DNNs and deserve to do some researches on it.** Besides, claimed by [2], that the test accuracy can be quantitatively characterized in terms of the noise ratio in datasets, which is also interesting to read it in detail.
In summary, I think [20, 6, 7, 11, 2] deserve to read thoroughly.

3.1.4 Sample weighting

Sample weighting can be regard as an extention method of sample choosing. In stead of just dropping the suspicious samples, lower weights were assigned on this kinds of data and higher attention were paid to trusted one. Three benchmark papers deserved to read are as follows:

- In [14], feature extractor network is used to extract features from both reference set and queried samples. Based on their features, similarity loss is defined to measure relevance of the image to its label, which is then used to weight importance of the particular sample to learn.

- In [26], they proposed a meta paradiam to determine the weight on each sample. In each iteration, gradient descent step on given mini-batch for weighting factor is performed, so that if minimizes the loss on given small noise-free data sets. **By theoretical deriation, they provided an impressive result that if a pair of training and clean validation examples are very similar, and they also provide similar gradient directions, then this training example is helpful and should be up-weighted,**

and conversely, if they provide opposite gradient directions, this training example is harmful and should be downweighed. Besides, weighting factor being validated in each iteration is argued to have regularizer effect on the model, which is interesting and deserve to do some researches on it.

- [29] extends cross-entropy loss and introduces abstention mechanism, which gives option to abstain samples, depending o their cross-entropy error, with an abstention penalty. Therefore, network learns to abstain confusing samples during learning which helps to learn underlying structure of noise.

3.2 Noise Model Free Methods

These methods aims to achieve label noise robustness without explicitly modeling it, but rather designing robustness in the proposed algorithm. Various methodologies are presented in the following subsections.

3.2.1 Robust Loss Functions

A loss function is said to be noise robust if the classifier learned with noisy and noise-free data, both achieve the same classification accuracy[17].
(1) MAE and it's Variants: It has shown by [3], mean absolute error(MAE) is much more robust than commonly used loss categorical cross-entropy (CCE). Yet, training a network under MAE loss would be slow because the gradient can quickly saturate while training. In order to solve this problem, some variants of MAE were proposed.

- [31] provided an improved version of MAE, called IMAE and meanwhile, gave two findings: (1) MAE does not treat examples equally and (2) The variance of gradient magnitude matters.

- [36] proposed a generalization of MAE and CCE, which converges in a desirable speed and meanwhile is robust to the noise.

- In [32], it shows that Cross Entropy(CE) exhibits overfitting to noisy labels on some classes("easy" classes), but also suffers from significant under learning on some other classes("hard" classes). And then, they proposed an approach of Symmetric cross entropy Learning, boosting CE symmetrically with a noise robust counterpart Reverse Cross Entropy(RCE).

(2) 0-1 Loss and it's Variants: Shown by [17], 0-1 loss has noise tolerance much more than commonly used convex losses. However, 0-1 loss is non-convex and non-differentiable, which is difficult to be optimized by gradient based methods. Inspired by this problem, [15] proposed Curriculum loss(CL), which can be efficiently optimized while keeping 0-1 loss's robust properties.
(3) Loss Correction Approach: [23] estimate the noise transition matrix T and use it to correct loss function. And the corrected loss are called "backward" loss and "forward" loss. Additionally, they prove that for ReLU networks the Hessian of loss is independent from label noise.
(4) Information-theoretic Loss: A novel information theoretic loss function $\mathcal{L}_{DMI}$ is proposed by [33]. $\mathcal{L}_{DMI}$ is the first loss function that is provably robust to instance-independent label noise, regardless of noise pattern.
Because these seven papers [36, 15, 32, 31, 3, 23, 33] both have theoretical guarantees and great performances, I think they are worth to read in detail.
And there are some other Robust loss functions always used in traditional machine learning methods such as SVM and ridge regression proposed by[18, 19, 22, 30, 18, 4] can be used as reference.

3.2.2 Regularizers

Regularizations are wide used methods to prevent DNNs from overfitting noisy labels. Along this direction, there are two interesting papers as follows:

- In [9], it is shown that the adversarial pre-training contributes to label noise robustness of model.

- In [16], they propose a new perspective for understanding DNN generalization by investigating the dimensionality of the deep representation subspace of training samples. It is shown that, from a dimensionality perspective, DNNs exhibit different learning styles on clean and noisy labels. (**It deserve to read thoroughly because, in my opinion, robust training DNNs with noisy labels, in some sense, is trying to understanding the generalization of DNNs.**)

3.2.3 Others

- [21] uses semi-supervised paradiam and iteratively filters label noise by checking the consistency between moving average of model prediction and given labels. It seems achieving outperformed results.

- In [13], they start with an indirect learning method called Negative Learning(NL) in which the DNNs are trained using a complementary label as in "input image does not belong to this complementary label". I think this is a interesting considering direction.

- In [24], they proposes a statistic named Area Under the Margin(AUM), which exploits differences in the training dynamics of clean and mislabeled samples. A simple procedure—adding an extra class populated with purposefully mislabled indicator samples—learns a threshold that isolates mislabeled data based on this metric.

References

[1] Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. Unsupervised label noise modeling and loss correction. https://arxiv.org/pdf/1904.11238.pdf, 2019.

[2] Chen, P., Liao, B., Chen, G., and Zhang, S. Understanding and utilizing deep neural networks trained with noisy labels. https://arxiv.org/pdf/1905.05040.pdf, 2019.

[3] Ghosh, A., Kumar, H., and Sastry, P. Robust loss functions under label noise for deep neural networks. https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14759/14355, 2017.

[4] Ghosh, A., Manwani, N., and Sastry, P. Making risk minimization tolerant to label noise. https://arxiv.org/pdf/1403.3610.pdf, 2015.

[5] Goldberger, J., and Ben-Reuven, E. Training deep neural-networks using a noise adaptation layer, 2016.

[6] Guo, S., Huang, W., Zhang, H., Zhuang, C., Dong, D., Scott, M. R., and Huang, D. Curriculumnet: Weakly supervised learning from large-scale web images. https://arxiv.org/pdf/1808.01097.pdf, 2018.

[7] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. https://arxiv.org/pdf/1804.06872.pdf, 2018.

[8] Han, J., Luo, P., and Wang, X. Deep self-learning from noisy labels. https://arxiv.org/pdf/1908.02160.pdf, 2019.

[9] Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. https://arxiv.org/pdf/1901.09960.pdf, 2019.

[10] Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. https://arxiv.org/pdf/1802.05300.pdf, 2018.

[11] Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. https://arxiv.org/pdf/1712.05055.pdf, 2017.

[12] Jindal, I., Nokleby, M., and Chen, X. Learning deep networks from noisy labels with dropout regularization. https://arxiv.org/pdf/1705.03419.pdf, 2016.

[13] Kim, Y., Yim, J., Yun, J., and Kim, J. Nlnl: Negative learning for noisy labels. https://arxiv.org/pdf/1908.07387.pdf, 2019.

[14] Lee, K.-H., He, X., Zhang, L., and Yang, L. Cleannet: Transfer learning for scalable image classifier training with label noise. https://arxiv.org/pdf/1711.07131.pdf, 2018.

[15] Lyu, Y., and Tsang, I. W. Curriculum loss: Robust learning and generalization against label corruption. https://arxiv.org/pdf/1905.10045.pdf, 2019.

[16] Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S. M., Xia, S.-T., Wijewickrema, S., and Bailey, J. Dimensionality-driven learning with noisy labels. https://arxiv.org/pdf/1806.02612.pdf, 2018.

[17] Manwani, N., and Sastry, P. Noise tolerance under risk minimization. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6342929, 2013.

[18] Mnih, V., and Hinton, G. E. Learning to label aerial images from noisy data, 2012.

[19] Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. https://papers.nips.cc/paper/5073-learning-with-noisy-labels.pdf, 2013.

[20] Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., and Brox, T. Self: Learning to filter noisy labels with self-ensembling. https://arxiv.org/pdf/1910.01842.pdf, 2019.

[21] Nguyen, D. T., Ngo, T.-P.-N., Lou, Z., Klar, M., Beggel, L., and Brox, T. Robust learning under label noise with iterative noise-filtering. https://arxiv.org/pdf/1906.00216.pdf, 2019.

[22] Patrini, G., Nielsen, F., Nock, R., and Carioni, M. Loss factorization, weakly supervised learning and label noise robustness. https://arxiv.org/pdf/1602.02450.pdf, 2016.

[23] Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. https://arxiv.org/pdf/1609.03683.pdf, 2017.

[24] Pleiss, G., Zhang, T., Elenberg, E. R., and Weinberger, K. Q. Identifying mislabeled data using the area under the margin ranking. https://arxiv.org/pdf/2001.10528.pdf, 2020.

[25] Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. https://arxiv.org/pdf/1412.6596.pdf, 2014.

[26] Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. https://arxiv.org/pdf/1803.09050.pdf, 2018.

[27] Sukhbaatar, S., and Fergus, R. Learning from noisy labels with deep neural networks. https://arxiv.org/pdf/1406.2080.pdf, 2014.

[28] Tanaka, D., Ikami, D., Yamasaki, T., and Aizawa, K. Joint optimization framework for learning with noisy labels. https://arxiv.org/pdf/1803.11364.pdf, 2018.

[29] Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., and Mohd-Yusof, J. Combating label noise in deep learning using abstention. https://arxiv.org/pdf/1905.10964.pdf, 2019.

[30] Van Rooyen, B., Menon, A., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. https://arxiv.org/pdf/1505.07634.pdf, 2015.

[31] Wang, X., Kodirov, E., Hua, Y., and Robertson, N. M. Imae for noise-robust learning: Mean absolute error does not treat examples equally and gradient magnitude's variance matters. https://arxiv.org/pdf/1903.12141.pdf.

[32] Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. https://arxiv.org/pdf/1908.06112.pdf, 2019.

[33] Xu, Y., Cao, P., Kong, Y., and Wang, Y. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. https://arxiv.org/pdf/1909.03388.pdf, 2019.

[34] Yao, J., Wang, J., Tsang, I. W., Zhang, Y., Sun, J., Zhang, C., and Zhang, R. Deep learning from noisy image labels with quality embedding. https://arxiv.org/pdf/1711.00583.pdf, 2018.

[35] Yi, K., and Wu, J. Probabilistic end-to-end noise correction for learning with noisy labels. https://arxiv.org/pdf/1903.07788.pdf, 2019.

[36] Zhang, Z., and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. https://arxiv.org/pdf/1805.07836.pdf, 2018.