Adaptive Sample Selection with Joint Loss for Medical Image Classification Under Label Noise

Tongqing Xue[1], Aiping Qu[1, b*], and Han Hong[1 1]

[1*]School of Computer, University of South China HengYang, China.

ABSTRACT.

Deep Neural Networks (DNNs) have been extensively employed in the classification of medical images and achieve impressive performance. But the success of DNNs is dependent on the large amounts of correctly labelled images. However, noisy labels are often unavoidable in the real-world clinical scenarios, which significantly impact the performance of the model. In this manuscript, we introduce a new sample selection method which could select the clean samples adaptively without knowing the prior knowledge, such as label noise rates. We also integrated semi-supervised learning during sample selection to fully utilize the noised dataset. Specifically, we calculate batch statistics in each mini-batch and divide the samples into clean and noisy based on the statistics, then they are served as labelled and unlablled in the semi-supervised manner. Furthermore, we use a joint loss to leverage useful information from unlabelled data along with a supervised loss, which strengthens the model's robustness. To evaluate the effectiveness of our method, we conduct sufficient experiments on a medical image dataset: Chaoyang. The results show that the proposed method could deal with the noisy labels in real-world scenarios.

Keywords: Noisy labels; medical image classification; sample selection; semi-supervised learning.

1. INTRODUCTION

Deep learning methods have found broad application in the classification of medical images [1]. As is commonly recognized, deep learning based methods exceedingly depend on large-scale correctly annotated training datasets [2]. However, it is challenging to collect such large and clean datasets in the real medical scenarios, because the annotation is time-consuming and the noisy labels (incorrect labels) are often unavoidable during the artificial annotation. Therefore, robust approaches that can copy with noisy labels are highly longed-for.

Previous works have achieved great success in learning with noisy labels (LNL), and they also demonstrated many strategies to address this problem. Among these methods, a potential way to address noisy labels is sample selection. Its main idea is to filter out noisy samples from the original noised training set. Then consider either dropping out them for obtaining robustness against noisy labels or utilizing specific strategies to take advantage of the noisy data, such as semi-supervised learning (SSL). Sample selection based approaches have yield unprecedented results in many noisy image classification scenarios, demonstrating a high tolerance for label-noise. For example, several works [3,4,5] train two networks concurrently to select the clean samples based on a strategy called “small loss” to combat noisy labels. However, the aforementioned methods require information of noise rate which would be limited in real clinical scenarios since the noise rate of the medical image datasets is typically unknown in their nature.

To address the issue mentioned above, in this manuscript, we present an adaptive sample selection method with a joint loss for medical image classification, which could differentiate between clean and noisy samples based on the statistics of samples’ prediction probability in each mini-batch. In addition, we take account of the class labels in the process of calculating the batch statistics, so the selection is also related to the given labels of the samples. Moreover, a dual-networks co-training strategy (DCS), derived from Mixmatch [6], is devised to make full use of the training data and to explore useful information within the noisy data. We perform experiments on a publicly available medical image dataset: Chaoyang. The experimental results demonstrate that our method achieves superior accuracy when compared to other approaches. Our contributions are as follows:

[1*b] qap@usc.edu.cn

The International Conference Optoelectronic Information and Optical Engineering (OIOE2024), edited by Yang Yue, Lu Leng, Proc. of SPIE Vol. 13513, 135131W · © 2025 SPIE · 0277-786X · © The Authors. Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3045588

• We propose a robust noisy label learning method for medical image classification, which does not need any prior knowledge, thus it is more suitable for real-world noisy datasets.

• Dual-networks co-training strategy (DCS) is applied to take full advantage of the noisy data, which can explore the useful information within the noisy samples.

2. METHODS

In this section, we showcase the primary contribution of this paper, a robust noisy-label learning algorithm for medical image classification based on the batch statistics. An overview of our proposed architecture is shown in Fig. 1. It mainly contains three modules: the dual-network framework, the mini-batch selection module, the mixmatch module.

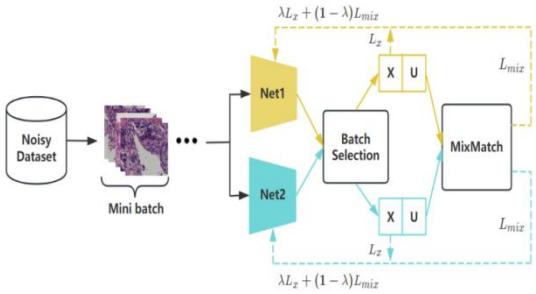

Figure. 1 The overall architecture of our method

2.1 Adaptive selection through batch statistics

Under the noise label over K classes, the labels supplied in the training set could be incorrect. We aim to train a classifier $f(\cdot;\theta)$ on this noisy-labeled training set so that it can exhibit good generalization on an unknown test set. As mentioned above, most algorithms need to know the prior knowledge which is impractical in real scenarios. Inspired by work BARE [7], in order to make our approach more applicable to real-world datasets, we introduce an adaptive curriculum which can be described as a weighted loss [8, 9], disregarding the regularization term and taking $l_i = L(f(x_i;\theta), y_i)$, the optimization becomes

$$\min_{\theta,\mathbb{w}} L(\theta,\mathbb{w}) = \sum_{i=1}^{m} \omega_i l_i - \varsigma\|\mathbb{w}\|_1 = \sum_{i=1}^{m} (\omega_i l_i + (1-\omega_i)\varsigma - m\varsigma) \quad (1)$$

where $-\varsigma\|\mathbb{w}\|_1, (\varsigma > 0)$ represents the curriculum, m is the size of a mini-batch. The optimal $\mathbb{w}$ for any fixed θ is: $\omega_i = 1$ if $l_i < \varsigma$ and $\omega_i = 0$ otherwise. When we put ς lie on the class label (ς_j is a function of θ and of all x_i whose labels corresponding to class j), we could maintain an adaptive curriculum where the ς_j is derived from all x_i of that class within a mini-batch and the present θ. Since the inverse relationship between the loss and the posterior probability, our selection criterion for a sample involves the requirement that the assigned posterior probability surpasses a threshold derived from a statistic of the observed posterior probabilities within the mini-batch. That is to say, we assign weights to samples in every mini-batch as

$$\omega_i = \begin{cases} 1, & f(x_i;\theta) \geq \varsigma_j = \mu_j + \sigma_j \\ 0, & \text{else} \end{cases} \quad (2)$$

where μ_j and σ_j represent the mean and variance of the class posterior probabilities for samples that are labeled as j.

2.2 Semi-supervised learning with MixMatch

Based on the adaptive selection, we have effectively distinguished between clean and noisy samples. In order to make full use of the dataset, inspired by DivideMix [10], we further employ the dual-network co-training strategy (DCS) with MixMatch, which helps alleviate the confirmation bias inherent in self-training where a single model may accumulate its errors. Given a mini-batch, after the mini-batch selection module, the samples are classified into clean and noisy subsets. We view the clean and noisy as labelled and unlabelled in a semi-supervised setting. Then, we utilize the combined predictions from the two networks to generate the labels for unlabeled samples. Specifically, MixMatch mixes the data, in which each sample is combined with another sample randomly selected from the merged mini-batch of X and U. After MixMatch, we get X^* and U^*, the loss on X^* is the cross-entropy loss and the loss on U^* is the mean squared error. Furthermore, to avoid allocating all samples to a single category, we use a regularization term L_{reg} in the mini-batch which is applied by [11] and [12], to regularize the mean output of the model throughout all instances in the mini-batch.

$$L_{reg} = \sum_{k=1}^{K} \frac{1}{K} \log\left(\frac{1}{K} \Big/ \frac{1}{|X^*|+|U^*|} \sum_{x \in X^*+U^*} f^k(x;\theta)\right) \quad (3)$$

2.3 Loss Functions

In our experiments, the total loss is:

$$L = \lambda L_x + (1-\lambda) L_u + L_{reg} \quad (4)$$

L_x is calculated on labelled samples with clean labels, L_u is calculated on unlabelled samples with noisy labels, L_{reg} is the regularization term. λ is a parameter, we set it to be 0.5 in our experiment.

3. EXPERIMENTS

3.1 Datasets and Experimental details

To validate the effectiveness of our method, we conducted comparative experiments on a challenging medical image dataset: Chaoyang [13].

Chaoyang dataset comprises 4021 training samples and 2139 testing samples. There are 664 adenoma, 1404 adenocarcinoma, 842 serrated, and 1111 normal in the training data. There are 273 adenoma, 840 adenocarcinoma, 321 serrated, and 705 normal in the testing data. It consists of 4 classes of colon slides, sourced from Chaoyang hospital. The patch dimensions are 512 × 512, and each patch is labeled by three experts. The testing set reflects a consensus opinion from all three pathologists. Approximately 40% of the training samples exhibit inconsistency and are labelled by one expert. For preprocessing, we perform random horizontal flip and resize the image to 256 × 256. The label noise in Chaoyang arises from a real-world scenarios, and the noise denotes the labeled samples that are genuinely incorrect, rather than artificially adding.

We utilized PyTorch to implement the codes and executed them on a workstation equipped with 32 GB NVIDIA Tesla V100 GPU. For the two networks, we use Resnet34 and Adam as optimizer with initial learning rate of 3e-3. The maximum epoch is set to 40. At the 10th, 20th, 30th epochs, the learning rate is reduced by half. We set the batch size to 16 for the dataset.

3.2 Comparison experiments and results

3.2.1 Comparison methods

We conduct experiments and compare our method to BARE [7], DM [10], SPR-LNL [14], LongReMix [15]. For these methods all do not need know the noise rate.

BARE [7] (Deep Patel et al. 2018). It proposed an adaptive sample selection method for learning with noisy labels.

DM [10] (Junnan Li et al. 2020). It used semi-supervised learning (SSL) for learning with noisy labels, which adopted Gaussian Mixture Model (GMM) to model loss distribution of each sample. Then, the clean samples and noisy samples were used as labelled and unlabelled for SSL.

SPR_LNL [14] (Yikai Wang et al. 2022). It was based on statistical sample selection, which proved to be guaranted

LongReMix [15] (Filipe R. Cordeiro et al. 2023). It adopted unsupervised learning to separate the training data into clean and noisy, then used semi-supervised learning to minimize the empirical vicinal risk (EVR).

3.2.2 Evaluation criteria

We conducted all experiments using the open-source codes provided by the authors. Additionally, we trained the model three times for each experiment and reported the mean value along with the standard deviation. Our evaluation metrics include Accuracy (ACC), Precision, Recall, F1 Score (F1), AUC, and Specificity.

3.2.3 Experiment results

Table 1 shows the comparative results on Chaoyang dataset. From the table, we can see that our method demonstrates superior performance, achieving an ACC, AUC, F1 score, Precision, Recall, Specificity of 83.47%, 94.22%, 76.72%, 79.37%, 75.19%, 94.26%, respectively. SPR-LNL gets comparable performance to ours and has a slightly higher recall with 75.38%, it can be explained that they adopt a penalized regression to help identify the noisy samples which is a theoretically guaranteed framework. Furthermore, they also combine it with semi-supervised algorithm to utilize the information with the noisy samples. In contrast, the results of BARE are the poorest with 61.46% ACC. This indicates that methods designed for natural images may not necessarily perform well on medical images. Moreover, this once

again confirms that simply discarding noisy samples can lead to the loss of essential information, especially in medical imaging scenarios.

Table 1. Comparison results with other methods.

Method	ACC	AUC	F1 score	Precision	Recall	Specificity
BARE	61.46±2.31	75.30±1.61	54.12±1.55	53.75±1.53	54.83±1.65	86.71±0.76
DM	75.67±1.62	88.46±2.53	66.31±2.27	69.06±2.67	65.92±1.69	91.18±0.61
SPR-LNL	83.06±0.23	94.18±0.14	76.45±0.12	79.02±0.68	75.38±0.17	94.00±0.04
LongReMix	76.41±0.88	89.28±0.72	65.96±1.80	70.05±2.06	64.81±1.38	91.48±0.33
Ours	83.47±0.20	94.22±0.21	76.72±0.14	79.37±0.62	75.19±0.14	94.26±0.06

3.3 Ablation study

In this subsection, we analyze the performance benefits from each component in our method. We conduct ablation experiment on Chaoyang dataset, the results are presented in Table 2. The batch selection (abbreviated as BS) means to train the network without MixMatch and just use one network, which degenerates to the work BARE. To validate the gain of dual-networks, we train two networks with batch selection and use cross-entropy loss (abbreviated as DCS + BS). To validate the proposed joint loss, we train two networks with batch selection and MixMatch (abbreviated as DCS + BS + Mix). The results indicate that each component contributes to a certain improvement in the overall performance. Specifically, dual-networks co-training strategy improves a lot compared to the single-net BS. It gets a further modest improvement by adding the joint loss. In this study, the method with DCS and Mix achieves a higher ACC, AUC, F1 and Precision.

Table 2. Module ablation on Chaoyang dataset.

Method	ACC	AUC	F1 score	Precision	Recall	Specificity
BS	61.46±2.31	75.30±1.61	54.12±1.55	53.75±1.53	54.83±1.65	86.71±0.76
DCS+BS	83.09±0.32	93.90±0.75	76.54±0.67	77.68±0.57	75.82±0.57	94.27±0.14
DCS+BS+Mix	83.47±0.20	94.22±0.21	76.72±0.14	79.37±0.62	75.19±0.14	94.26±0.06

4. CONCLUSION

In medical imaging tasks, it is challenging to collect a large amount of samples with correct labels due to the complex situations in medical imaging field. There have been many algorithms to address the issue of noisy labels, but many of them require some prior knowledge, such as noise ratio or a small clean dataset. However, the prior knowledge is unavailable in real medical datasets. In this manuscript, we present a robust noisy label learning method for medical image classification to solve the above problem. Specifically, we adopt adaptive sample selection based on batch statistics to distinguish between clean and noisy samples. Unlike other methods, we do not discard the noisy samples, we propose a joint loss to utilize them in a semi-supervised manner. We demonstrate the effectiveness of our method on a real-world medical image dataset, which proves its reliability in dealing with real medical image classification with noisy labels.

5. REFERENCES

[1] Shen D, Wu G, Suk H I. Deep learning in medical image analysis[J]. Annual review of biomedical engineering, 2017, 19: 221-248.

[2] Zhou S K, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises[J]. Proceedings of the IEEE, 2021, 109(5): 820-838.

[3] Han B, Yao Q, Yu X, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels[J]. Advances in neural information processing systems, 2018, 31.

[4] Yu X, Han B, Yao J, et al. How does disagreement help generalization against label corruption?[C]//International Conference on Machine Learning. PMLR, 2019: 7164-7173.
[5] Wei H, Feng L, Chen X, et al. Combating noisy labels by agreement: A joint training method with co-regularization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 13726-13735.
[6] Berthelot D, Carlini N, Goodfellow I, et al. Mixmatch: A holistic approach to semi-supervised learning[J]. Advances in neural information processing systems, 2019, 32.
[7] Patel D, Sastry P S. Adaptive sample selection for robust learning under label noise[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 3932-3942.
[8] Jiang L, Zhou Z, Leung T, et al. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels[C]//International conference on machine learning. PMLR, 2018: 2304-2313.
[9] Kumar M, Packer B, Koller D. Self-paced learning for latent variable models[J]. Advances in neural information processing systems, 2010, 23.
[10] Li J, Socher R, Hoi S C H. Dividemix: Learning with noisy labels as semi-supervised learning[J]. arXiv preprint arXiv:2002.07394, 2020.
[11] Tanaka D, Ikami D, Yamasaki T, et al. Joint optimization framework for learning with noisy labels[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5552-5560.
[12] Arazo E, Ortego D, Albert P, et al. Unsupervised label noise modeling and loss correction[C]//International conference on machine learning. PMLR, 2019: 312-321.
[13] Zhu C, Chen W, Peng T, et al. Hard sample aware noise robust learning for histopathology image classification[J]. IEEE transactions on medical imaging, 2021, 41(4): 881-894.
[14] Wang Y, Sun X, Fu Y. Scalable penalized regression for noise detection in learning with noisy labels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 346-355.
[15] Cordeiro F R, Sachdeva R, Belagiannis V, et al. Longremix: Robust learning with high confidence samples in a noisy label environment[J]. Pattern Recognition, 2023, 133: 109013.