Cascaded Robust Learning at Imperfect Labels for Chest X-ray Segmentation

Cheng Xue[1(✉)], Qiao Deng[1], Xiaomeng Li[1,2], Qi Dou[1], and Pheng-Ann Heng[1,3]

[1] Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Sha Tin, Hong Kong
cxue@cse.cuhk.edu.hk
[2] Department of Radiation Oncology, Stanford University, Stanford, USA
[3] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology,
Chinese Academy of Sciences, Shenzhen, China

Abstract. The superior performance of CNN on medical image analysis heavily depends on the annotation quality, such as the number of labeled images, the source of images, and the expert experience. The annotation requires great expertise and labor. To deal with the high inter-rater variability, the study of the imperfect label has great significance in medical image segmentation tasks. In this paper, we present a novel cascaded robust learning framework for chest X-ray segmentation with imperfect annotation at the boundary. Our model consists of three independent networks, which can effectively learn useful information from peer networks. The framework includes two stages. In the first stage, we select the clean annotated samples via a model committee setting, the networks are trained by minimizing a segmentation loss using the selected clean samples. In the second stage, we design a joint optimization framework with label correction to gradually correct the wrong annotation and improve the network performance. We conduct experiments on the public chest X-ray image datasets collected by Shenzhen Hospital. The results show that our methods could achieve a significant improvement on the accuracy in segmentation tasks compared to the previous methods.

Keywords: Robust learning · Imperfect label · Lung segmentation

1 Introduction

Deep neural networks (DNNs) have achieved human-level performance on many medical image analysis tasks, such as melanoma diagnosis [5], pulmonary nodules detection [14], retinal disease [4], and lumpy node metastases detection [1]. These outstanding performances heavily rely on massive training data with high-quality annotations. Annotation of medical images, especially for pixel-level annotation for segmentation tasks, is costly and time-consuming. The process is experience-prone, while the annotations from different clinical experts may have disagreements that are usually inevitable for the blurred boundary of lesions and organs.

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12266, pp. 579–588, 2020.
https://doi.org/10.1007/978-3-030-59725-2_56

Previous studies show that the DNNs trained by noisy labeled datasets can cause performance degradation. That is because the huge memory capacity and strong learning ability of DNNs can remember the noisy labels and easily overfit to them [15,18,19]. Tackling the issue of annotation noises is a complicated and challenging topic. Manually reducing the presence of incorrect labels, for example by requiring a stronger committee of expert clinicians to do labelling, has to be expensive, time-consuming and impractical. In this paper, we address this problem in the insight of robust learning with the noisy labelled data inherent in the training procedure. Many studies have addressed the issue of the noisy label in medical analysis community. Goldberger et al. [6] added an additional softmax layer to estimate the correct labels. Xue et al. [17] proposed to consider the noisy sample and hard sample by an on-line sample selection module and re-weighting module. Zhu et al. [19] proposed the automatic quality evaluation module and overfitting control module to update the network parameters. Shu et al. [15] presented an LVC-Net losses function by combining noisy labels with image local visual cues to generate better semantic segmentation. Most of the approaches adopted the strategy of selecting samples for training [17,19], exhibited their feasibility in robust learning. However, these methods exist a strong accumulated error caused by sample selection bias. The wrongly selected samples will influence the network performance and further decrease the quality of selected samples. Le et al. [10] addressed the sample selection bias issue by utilized a small set of clean training samples to assign weights to training samples. The main drawback of this approach was the extra clean labels were usually unavailable in the real-world scenarios.

To tackle the challenging problem of noisy labeled segmentation masks, we present a cascaded learning framework for lung segmentation using the X-ray images with imperfectly annotated ground truth. In the first stage, our framework selects clean annotated samples according to the prediction confidence and uncertainty of samples, which is inspired by the ideas of Co-teaching [7]. Specifically, our model consists of three independent networks being trained simultaneously, each network is real-time updated according to the prediction results of the other two networks. For a clean annotated sample, the three networks tend to produce high confidence prediction with smaller inter-rater variance. Thus, the samples with close prediction and high confidence are selected as the high-quality sample, which will be used to contribute to the weight backpropagation process. Since the selection stage leads to a low utilization efficiency of the valuable training data, we propose a label correction module in the second stage, which can correct the imperfect label. Furthermore, a joint optimization scheme is designed to cooperatively supervise the three networks with the original label and the corrected one. Our method was extensively evaluated on the Shenzhen chest x-ray dataset [3,8,16]. The results demonstrate a good capability of our method to the issue of the noisy labeled boundary, that the cascaded robust learning framework can more accurately perform the lung segmentation comparing to other methods.

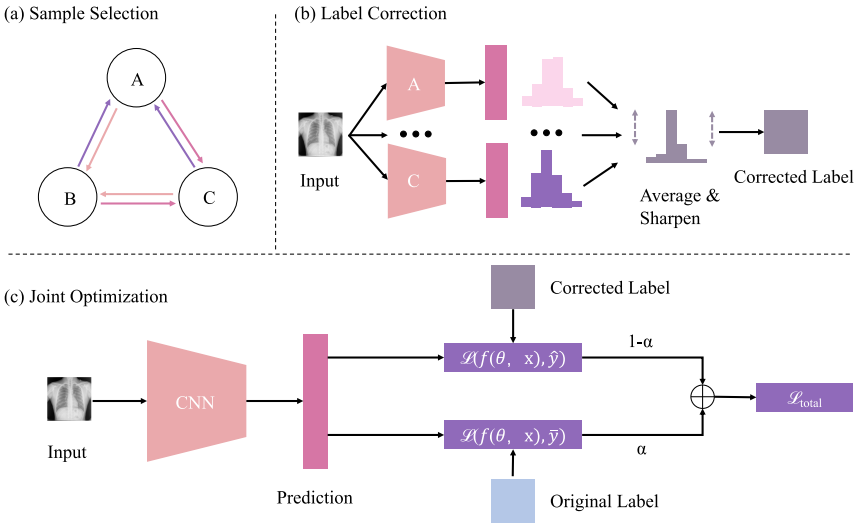


Fig. 1. Illustration of the pipeline of our cascaded robust learning framework. (a) shows the first sample selection stage, where three networks trained independently, but updated according to the prediction of the other two peer networks. (b) and (c) are the second stage. (b) shows our proposed label correction module, using the average prediction of three networks followed by a sharpening function to produce the corrected label $\bar{y}$. (c) shows the joint optimization scheme, the network is supervised by the original label $\hat{y}$ and the corrected label $\bar{y}$. The final output is given by the average of the three networks.

2 Method

Figure 1 illustrates the framework of our cascaded robust learning method. In the first stage, we design a three-networks sample selection module. The module filters the clean samples and updates the three networks with the selected clean samples. In the second stage, our method starts to correct the imperfect labels, then use the corrected label and original label to jointly optimize the three networks.

2.1 Sample Selection Stage

We study the task of chest x-ray segmentation, where the training set contains images x and noisy labeled ground truth $\hat{y}$, while the clean ground truth y is unknown. The goal for this fully supervised segmentation task is to minimize the following object function:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(f(x_i;\theta), \hat{y}_i) \tag{1}$$

where $\mathcal{L}$ denotes the loss function (e.g., cross-entropy loss) to evaluate the quality of the network output on inputs. $f(\theta)$ denotes the segmentation neural network with weights θ.

Recent studies show that by updating the network with high confidence samples can improve the robustness to noisy labels [7,9,12]. Therefore, we propose a novel sample selection framework (SS) to select high confidence samples as the useful training instances. Our framework consisted of three independent networks, where they have identical architecture. We adopt the vanilla U-Net [13] as the classifier in our experiment. In the training process, we select the high confidence samples with small uncertainty to update each network, because those samples are more likely to be clean labeled instances. In our experiment, we empirically select half batch data as useful information. Concretely, the three networks feed forward and predict the same mini-batch of data. Then for each network, the useful samples for weight updating are obtained by the other two networks as shown in Fig. 1(a). Taking network A as an example, the useful sample for network A is obtained from network B and C, where we first filter out the high uncertainty (μ) samples by excluding the ones showing disagreed prediction, then among the low uncertainty samples, the small loss samples was further selected as useful samples for network A. We employ the agreement between two models as uncertainty of each samples and calculate the uncertainty according to Eq. 2.

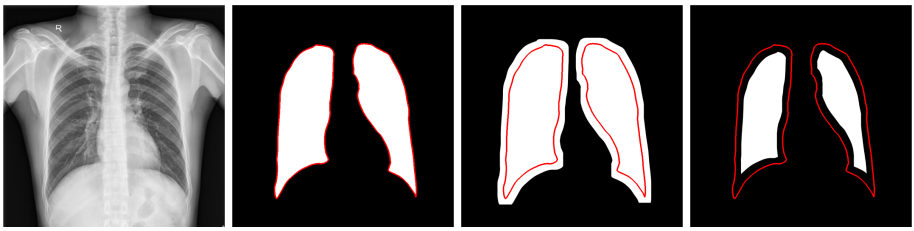
$$\mu = |\mathcal{L}(f_B(x_i;\theta_B),\hat{y}_i) - \mathcal{L}(f_C(x_i;\theta_C),\hat{y}_i)| \tag{2}$$

where $\mathcal{L}$ denotes the cross-entropy loss. f_B and f_C denote the network B and network C. θ_B and θ_C represent the weight of network B and C. Note that the three networks has different training parameters as they are updated by different selected samples in each mini-batch, they did not learn the bias in the noisy labels at the same speed, and μ is not close to 0.

2.2 Joint Optimization with Label Correction

In the stage of sample selection, only partial samples can be used for training, where it does not take full advantage of the imperfect training data. Therefore, we design a joint optimization (JO) framework to train the network with the original label and corrected label, so that the utilization efficiency of training data can be maintained. In order to correct the noisy label, we design a label correction module to work together with the joint optimization scheme.

Label Correction. The sample selection stage first trains an initial network by using image x with noisy label $\hat{y}$. Then we proceed to the label correction phase, as shown in Fig. 1 (b). We compute the average of three model' prediction in each iteration, that is followed by an entropy minimization step widely adopted in semi-supervised learning [2,11]. Specifically, for the average prediction of the three models, we apply a sharpening function to reduce the entropy of the per pixel label distribution through adjusting the temperature:



(a) chest X-ray (b) ground truth (c)dilation (d)erosion

Fig. 2. Examples of the noisy annotations. Red line indicates the clean ground truth. (Color figure online)

$$q = \frac{1}{3}(f_A((x_i;\theta_A),\hat{y}_i) + f_B((x_i;\theta_B),\hat{y}_i) + f_C((x_i;\theta_C),\hat{y}_i))$$
$$\bar{y} = sharpen(q,T)_i = q_i^{\frac{1}{T}} / \sum_{j=1}^{L} q_j^{\frac{1}{T}} \tag{3}$$

where q is the average prediction feature map over three models, T is a hyperparameter that adjusts the temperature. As T closes to zero, the output of $Sharpen(q,T)$ will approach a one-hot distribution. Since we will use $\bar{y} = Sharpen(q,T)$ as a corrected target for the model's prediction later, following the setting of [2], $T = 0.5$ is chosen to encourage the model to produce lower-entropy prediction.

Joint Optimization. We start the joint optimization stage after k epochs of sample selection. For each uncertain sample, we produce a corrected label for the imperfect input by the label correction module. The corrected label is used in the training process together with the original label as a complementary supervision to jointly supervise the network:

$$\mathcal{L}_{total} = \alpha \times \mathcal{L}(f(x_i;\theta),\hat{y}_i) + (1-\alpha) \times \mathcal{L}(f(x_i;\theta),\bar{y}_i) \tag{4}$$

where $\mathcal{L}$ is the cross entropy loss, $\hat{y}$ is the original noisy label, and $\bar{y}$ is the corrected label produced by the label correction phase. The weight factor α controls the weights of the two terms. In our study, we set $\alpha = 0.5$ that gives the best results.

3 Experiments

3.1 Dataset and Pre-processing

We evaluated our method on the public Shenzhen chest x-ray dataset [3,8,16], the segmentation masks were prepared manually by Computer Engineering Department, Faculty of Informatics and Computer Engineering, National Technical University of Ukraine. The dataset contains 566 chest x-ray images and

(a) Chest X-ray (b) ground truth (c) vanilla U-Net (d)ours

Fig. 3. Two examples of the segmentation results in the test data by different methods. (a) is the input image, (b) is the ground truth. (c) and (d) shows the results of U-Net and our method under 75% noise ratio.

each image has the left and the right lungs. We split the 566 chest x-ray images into 396 images for training and 170 for evaluation. All the images were resized to 256×256, and normalized as zero mean and unit variance.

3.2 Implementation

The framework was implemented in PyTorch, using a TITAN Xp GPU. We used the Stochastic Gradient Descent optimizer to update the network parameters with weight decay of 0.001 and a momentum of 0.9. We adopt an exponential learning rate with an initial learning rate set as 0.001. We totally trained 100 epochs, the batch size was 32. We adopted the data augmentation including random rotation and random horizontal flipping. In order to produce noisy labels for the training data, we simulate imperfect annotation with noisy boundary in real word scenarios. We randomly selected (noise ratio) 25%, 50%, 75% samples from the training set to erode or dilate with the number of iterations (noise level) n between 5–15 ($5 \leq n \leq 15$). We adopted the dice coefficient as evaluation criteria for segmentation accuracy evaluation. Figure 2 shows the example of some noisy annotation of the segmentation mask.

3.3 Quantitative Evaluation

The experiments were conducted on the Chest X-ray dataset. We trained the network on the samples with different ratio of noisy labels and tested it by the clean labels. Table 1 presents the segmentation performance of vanilla U-Net (baseline) and our cascaded robust learning framework that were all trained by

Table 1. Comparison between our method and various methods.

Noise ratio	Noise level	Strategy	Dice (%)	k		
				20	50	80
No noise	–	Vanilla U-Net	89.89	–	–	–
No noise	–	Co-teaching [7]	91.46	–	–	–
No noise	–	Ours	–	**92.52**	92.50	92.36
25%	$5 \leq n \leq 15$	Vanilla U-Net	87.58	–	–	–
25%	$5 \leq n \leq 15$	Co-teaching [7]	89.06	–	–	–
25%	$5 \leq n \leq 15$	Ours-SS	91.42	–	–	–
25%	$5 \leq n \leq 15$	Ours	–	92.11	92.81	**93.06**
50%	$5 \leq n \leq 15$	Vanilla U-Net	86.65	–	–	–
50%	$5 \leq n \leq 15$	Co-teaching [7]	88.56	–	–	–
50%	$5 \leq n \leq 15$	Ours-SS	88.87	–	–	–
50%	$5 \leq n \leq 15$	Ours	–	**90.14**	90.05	89.56
75%	$5 \leq n \leq 15$	Vanilla U-Net	84.96	–	–	–
75%	$5 \leq n \leq 15$	Co-teaching [7]	90.23	–	–	–
75%	$5 \leq n \leq 15$	Ours-SS	90.41	–	–	–
75%	$5 \leq n \leq 15$	Ours	–	91.07	90.19	**91.17**

noisy labels. We first trained the fully supervised vanilla U-Net with the noisy ratio set to zero, which can be regarded as the upper-line performance. Compared with the vanilla U-Net, our framework improves the segmentation performance and achieves an average Dice of 0.925 on the clean annotated dataset, indicating that the sample selection stage and joint-optimization stage can encourage the model to learn more distinguishing features.

For the training dataset with different ratio of noisy labels, we observed that as the noise ratio increases, the segmentation performance of the vanilla U-Net decreases dramatically. Compared with vanilla U-Net, the sample selection stage (SS) can consistently improve the performance by encouraging the model to be trained by the selected data. Through the joint optimization (JO) stage supervised by the corrected label and original ones, the segmentation accuracy is further improved, suggesting that our method can effectively eliminate the effect of the noise and gain performance by producing the correct label. In Fig. 3, we show some segmentation results under 75% noise, in which our results have higher Dice score than the vanilla U-Net. At all the noise ratio, we compared our method with the state-of-the-art noise robust method [7], which select the small loss samples according to the prediction of peer network. The results show that our method outperforms the state-of-the-art method in all the noise ratio setting.

In our experiment, we also investigated the impact of the starting epoch k on the performance of our method. As shown in Table 1, the joint optimization

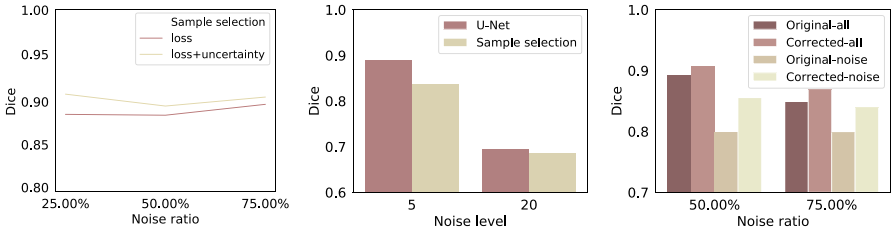


Fig. 4. (a) The segmentation accuracy of different sample selection criteria. (b) The segmentation accuracy of the U-Net and U-Net with only sample selection stage on 100% noise ratio setting. (c) The label accuracy of labels in the original dataset, and labels corrected by the model at the end of training.

(JO) with label correction stage is started at 20, 50, and 80 epochs, respectively. The experimental results show that the joint optimization stage can consistently produce good results with different starting epoch k.

3.4 Analysis of Our Method

Sample Selection. Compared with the vanilla U-Net, our sample selection stage (SS) shows higher segmentation accuracy under different noisy ratio, as shown in Table 1. To validate the criteria of our sample selection, we conducted another experiment that only selected the small loss sample. Figure 4(a) shows the test accuracy with different sample selection criteria. It reveals that the test accuracy significantly improved when considering the uncertainty in the selection stage. To further validate the effectiveness of our method at the sample selection stage, we applied our method on training dataset with 100% noise and noise level $n = 5, 20$. As shown in Fig. 4(b), under this setting, the sample selection stage shows worse segmentation accuracy than vanilla U-Net, because no clean sample can be selected. The results decreased due to the low sample utilization efficiency.

Joint Optimization. To analyze the contribution of the joint optimization stage, we explore the label accuracy with and without the stage of joint optimization and label correction. We calculated the Dice coefficient of the initial noisy label ($\hat{y}$) and the corrected label ($\bar{y}$) of the final model at the end of the training. Figure 4(c) shows the overall accuracy for severe noise situation (50%, 75%), where the Dice coefficient for all the original (Original-all) and corrected label (Corrected-all), and the Dice coefficient only for the original noise label (Original-noise) and corrected noise label (Corrected-noise) are presented. We see that the label quality is improved by the scheme of joint optimization and label correction, especially for those original noise labels.

4 Conclusion

In this paper, we present a novel Cascaded Robust Learning framework for the segmentation of noisy labeled chest x-ray images. Our method consists of two stages: sample selection stage, and the stage of joint optimization with label correction. In the first stage, the clean annotated samples are selected for network updating, so that the influence of noisy sample can be interactively eliminated in the three networks. In the second stage, the label correction module works together with the joint optimization scheme to revise the imperfect labels. Thus the training of the whole network is supervised by the corrected labels and the original ones. Compared with other state-of-the-art models, our cascaded robust learning framework keeps high robustness when the training data contains imperfect annotate boundaries. Experimental results on the benchmark dataset demonstrate that our network outperforms other methods on segmentation tasks and achieves very competitive results on the noisy labeled dataset.

Acknowledgments. This work is supported by Hong Kong Innovation and Technology Commission (Project No. ITS/311/18FP), Shenzhen Science and Technology Program (JCYJ20170413162256793) and a CUHK Direct Grant for Research.

References

1. Bejnordi, B.E., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: MixMatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems, pp. 5050–5060 (2019)
3. Candemir, S., et al.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans. Med. Imaging **33**(2), 577–590 (2013)
4. De Fauw, J., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. **24**(9), 1342–1350 (2018)
5. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)
6. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer. In: ICLR (2017)
7. Han, B.,et al.: Co-teaching: robust training of deep neural networks with extremely noisy labels. In: Advances in Neural Information Processing Systems, pp. 8527–8537 (2018)
8. Jaeger, S., et al.: Automatic tuberculosis screening using chest radiographs. IEEE Trans. Med. Imaging **33**(2), 233–245 (2013)
9. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: MentorNet: regularizing very deep neural networks on corrupted labels. In: ICML (2018)
10. Le, H., Samaras, D., Kurc, T., Gupta, R., Shroyer, K., Saltz, J.: Pancreatic cancer detection in whole slide images using noisy label annotations. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 541–549. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_60

11. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst. (2020)
12. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: ICML (2018)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Setio, A.A.A., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med. Image Anal. **42**, 1–13 (2017)
15. Shu, Y., Wu, X., Li, W.: LVC-Net: medical image segmentation with noisy label based on local visual cues. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 558–566. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_62
16. Stirenko, S., et al.: Chest X-ray analysis of tuberculosis by deep learning with segmentation and augmentation. In: 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO), pp. 422–428. IEEE (2018)
17. Xue, C., Dou, Q., Shi, X., Chen, H., Heng, P.A.: Robust learning at noisy labeled medical images: applied to skin lesion classification. In: ISBI (2019)
18. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: ICLR (2017)
19. Zhu, H., Shi, J., Wu, J.: Pick-and-learn: automatic quality evaluation for noisy-labeled image segmentation. arXiv preprint arXiv:1907.11835 (2019)