Deep Learning Based Multi-Label Chest X-Ray Classification with Entropy Weighting Loss

Shaocong Mo[✉], Ming Cai*[✉]
College of Computer Science and Technology
Zhejiang University
Hangzhou 310027, China

***Abstract*—With large chest X-ray image datasets open-accessed, it is feasible to build diagnosis system based on deep learning methods. In this work, we formulated chest X-ray diagnosis as multi-label classification problem. However, multi-labels are treated independently in traditional methods and existing loss can easily lose important information for class with fewer cases. We proposed an entropy weighting loss to observe inter-label dependencies and make full use of class whose cases are fewer than others. From the global perspective of all pathological classes, digging relations between all classes, the proposed loss could improve model performance. We experimented with three basic deep learning models (VGG16, ResNet50, DenseNet121) at first, and decided to use DenseNet121 as a baseline model. We evaluated our approach and achieved better results, whose AUC score is 0.8430 on average, under the Chest X-ray14 and its patient-wise split. We demonstrate that our proposed loss can dig relations between different disease labels just as illness complications in reality, and we do not consider every label is independent.**

***Keywords*-Chest X-ray; Multi-label learning; DenseNet; Cross entropy**

I. Introduction

Medical imaging can provide non-invasive information for diseases, and it is an indispensable component for disease diagnosis and treatment[1]. With medical imaging playing a considerable role, doctors monitor and analyze the occurrence, development, and feedback of treatment. Chest X-ray has been used to find diseases such as tuberculosis, pneumonia, and lung cancer[2]. It is a low-cost method for diagnosing diseases and playing an essential role in clinical and epidemiological studies.

With machine learning and deep learning methods having enormous potential usage, precision medicine relies on more heterogeneous data[3]. However, a primary key issue is that, facing large-scale medical image data, how to make full use of them to process and analyze. Image-level medical image classification can seem as multi-label classification problem with huge amounts of labels. Different from natural images, predicting absence of each label in a medical image is as important as predicting its presence to reduce the likelihood of misdiagnosis.

Chest X-ray diagnosis can be taken as a particular multi-label classification challenge. Chest X-ray images contain dozens of patterns, corresponding to hundreds of underlying conditions. It leads to disagreements and unnecessary follow-up procedures among radiologists. Complex interactions between different anomalous patterns often have significant clinical implications. In the context of sophisticated pathological features, early diagnosis is of great help in the choice of treatment options for lung diseases, especially cancers [4]. From the perspective of medical characteristics or clinical cognition, there is no quantitative standard for judgment of intricate patterns in chest X-ray images. Doctors often only make preliminary judgments on the burr, lobulation, and edge of images, so it is very difficult to give precise diagnosis.

Deep learning technology is a data-driven method. Although there have been open-access chest X-ray image datasets that contain a large number of cases, these datasets are too small. Differences remain compared to traditional computer vision datasets. Typically, labels could be incomplete, and samples could be unrelated or wrong, leading to uncertain prior information.

To resolve this under weak prior, we trained the DenseNet architecture with weighted cross entropy loss as baseline firstly. Furthermore, we proposed an entropy weighting loss that takes the overall perspective of all pathological classes. We design that loss so that the model can inherently learn labels dependency between all classes.

We can summary the contributions of this paper as follows,

- We fine-tune DenseNet121 as baseline model.
- We propose an entropy weighting loss that is specifically tailored for multi label classification task.
- We experiment with two ways to calculate loss and then analysis results.

II. Related Work

With the advantage of large-scale dataset, Wang et al. [5] tested different deep CNNs architecture and got qualitative

*Corresponding author
[✉]E-mail: mosc@zju.edu.cn, cm@zju.edu.cn

978-1-7281-4653-9/19/$31.00 ©2019 IEEE
DOI 10.1109/ISCID.2019.10111

results. Yao et al. [6] proposed a methed to find interdependencies between labels by using RNNs. Li et al. [7] proposed a unified network system for pathological classification, but also used to locate diseases. They combined category information with image local information, implicitly encoding disease label, and location bounding-box. Kumar et al. [8] presented cascaded architecture model inspired by Adaboost with pair-wise error loss. Guan et al. [9] and Wang et al. [10] introduced attention guided solutions exploiting the mutual relationship between labels and the locations of diseases.

III. Method

Given images with disease labels, we design a model producing disease diagnosis simultaneously. Due to well-known performance, *DenseNet121*[11] is utilized as baseline neural network architecture. Then entropy weighting loss function is applied to improve model diagnosis performance. The deep CNNs model is used to do classification tasks by the type and quality of labels under the premise that the corresponding labels can not wholly determine images.

A. Problem Formulation

We firstly define multi-label chest X-ray image classification task.

Under the premise that labels cannot be completely correct, and labels can not completely determine samples, the deep CNNs are to classify chest X-ray images based on types and numbers of existing corresponding labels. In general, deep CNNs are used to optimize the correctness of diagnostic information with uncertain prior information.

B. Loss Analysis

Since the negative samples (healthy or no typical disease label) are much more compared to the positive samples (disease label), the positive and negative samples are incredibly unbalanced, weighted cross entropy loss can be used to optimize training phase. Positive and negative factors avoid extreme value problem, which is a limitation of the original cross entropy loss.

Though above loss can hack with the imbalanced distribution between healthy and unhealthy instances as well as possible. It is still unable to resolve the problem when prediction probability is half by half. Considering the sum of loss for each pathology, for every index in label sequence, sum up binary cross entropy each. Followed as (1),

$$Loss(I,l) = \sum_{c=1}^{class} \begin{aligned}[t] &[-l_c \log(p(L_c=1|I)) \\ &-(1-l_c)\log(p(L_c=0|I))] \end{aligned} \tag{1}$$

where $p(L_c = 1|I)$ is the probability of pathology label c, and the $p(L_c = 0|I)$ is the opposite. l_c is the label represented for each pathology in one label sequence, $l_c = 0$ or $l_c = 1$. For imbalanced problem, a positive factor can be multiplied with the first part in binary cross entropy.

However, the loss function above do not consider label relation problem when there are three or more classes. To improve this, we propose a modified one. From the overall perspective of pathological classes, multi-class entropy weighting loss was proposed. Following the fully connected layer, *softmax* produced probability of each class.

$$Loss(I,L) = -\sum^{nclass} l_i p_i \log(p_i) \tag{2}$$

p_i is the final probability of each class. We consider the confusion degree of each class represented classification performance. For extremely unbalanced data, the fewer ones can be observed more.

IV. Experiments

A. Dataset

To estimate the ability of our proposed loss, the Chest X-ray14 dataset is used. Chest X-ray14 [5] contains more than 110 thousands chest X-ray images from more 30 thousands individual patients. These images are initially rescaled into the size of 1024×1024. Each image has one label at least. The category of label consists 14 diseases, another label called 'No Finding' means patient of one image is normal or does not consist the listed 14 diseases. The dataset is officially splited into three subsets ($7:1:2$) by patient individually, training, validation and testing respectably.

B. Training Details

We downscaled the images to 224×224 and then normalized. When using DenseNet121 as the baseline model, we optimized the network using Adam [12] with 32 mini-batch size. We used learning rate initialed as 0.001, and we trained each model for 50 epochs. During other stages, we also downscaled the images to the same size as training and normalized with the same mean, standard deviation. Random horizontal flipping was performed as data augmentation.

We did all experiments on a PC with Intel Core i7 8700K, one NVIDIA GTX 1080Ti, 32GB of RAM, and Ubuntu Linux 16.04 in practice. And all experiments were implemented by Keras deep learning framework.

C. Evaluation

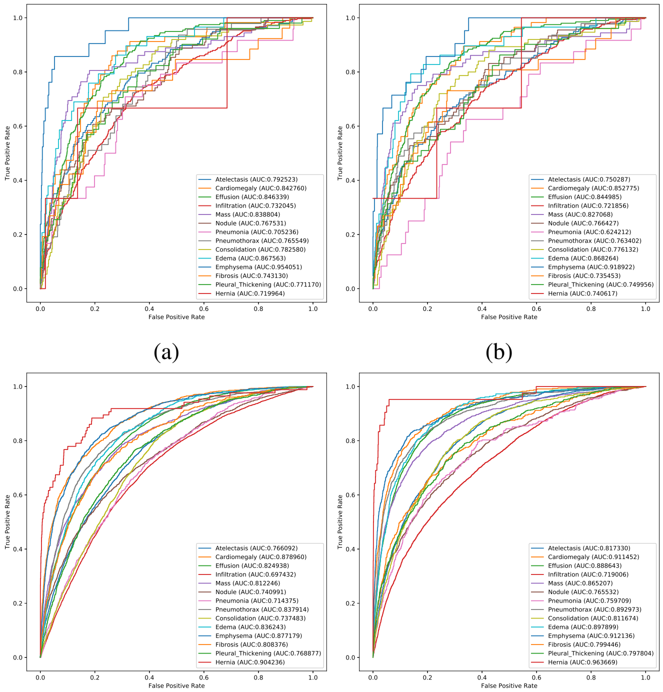
AUC score is used for model performance measurement, which is the area under ROC curve. A better classifier always has a higher AUC score.

D. Results

We tested three pre-trained baseline models, VGG16[13], ResNet50[14] and DenseNet121, all three models use weighted cross entropy loss. The corresponding AUC values of each pathology are given in **Table I** of 2nd to 4th columns from right to left. Since some state of arts split dataset

Table I: Comparison results of different baseline models (the 2nd to 4th columns from right to left, italic with blue) and classification results (the last two columns from left to right, bold with black) compared on different loss. For the convenience of results display, 'CE' means cross entropy.

Model	Wang et al.[5]	CNNs+RNNs[6]	CNNs + Patch Slicing[7]	Boosted Cascaded[8]	ChestNet[10]	VGGNet-16	ResNet-50	DenseNet121	DenseNet121
Loss	weighted CE loss		customized loss	pair-wise error loss	customized loss		weighted CE loss		proposed loss
Atelectasis	0.7200	0.8100	0.8000	0.7600	0.7433	*0.7925*	0.7503	0.7661	**0.8173**
Cardiomegaly	0.8100	0.9040	0.8100	0.9100	0.8748	0.8428	0.8528	*0.8790*	**0.9115**
Consolidation	0.7100	0.7880	0.8000	0.7800	0.7256	*0.7826*	0.7761	0.7375	**0.8117**
Edema	0.8300	0.8820	0.8800	0.8900	0.8327	0.8676	*0.8683*	0.8362	**0.8979**
Effusion	0.7800	0.8590	0.8700	0.8600	0.8114	*0.8463*	0.8450	0.8249	**0.8886**
Emphysema	0.8100	0.8290	0.9100	0.9000	0.8222	*0.9541*	0.9189	0.8772	**0.9121**
Fibrosis	0.7700	0.7670	0.7800	0.7600	0.8041	0.7431	0.7355	***0.8084***	0.7994
Hernia	0.7700	0.9140	0.7700	0.9000	0.8996	0.7200	0.7406	*0.9042*	**0.9637**
Infiltration	0.6100	0.6950	0.7000	0.6900	0.6772	*0.7320*	0.7219	0.6974	**0.7190**
Mass	0.7100	0.7920	0.8300	0.7800	0.7833	*0.8388*	0.8271	0.8122	**0.8652**
Nodule	0.6700	0.7170	0.7500	0.7000	0.6975	*0.7675*	0.7664	0.7410	**0.7655**
Pleural Thickening	0.7100	0.7650	0.7900	0.7700	0.7513	*0.7712*	0.7500	0.7689	**0.7978**
Pneumonia	0.6300	0.7130	0.6700	0.7100	0.6959	0.7052	0.6242	*0.7144*	**0.7597**
Pneumothorax	0.8100	0.8410	0.8700	0.8600	0.8098	0.7655	0.7634	*0.8379*	**0.8930**
AVE	0.7386	0.8054	0.8021	0.8050	0.7810	0.7949	0.7815	*0.8004*	**0.8430**



(a) (b)

(c) (d)

Figure 1: From (a) to (c): Multi-label classification performance (ROC curve) in different model with weighted cross entropy loss. VGG16, ResNet50, DenseNet121. (d) Multi-label classification performance (ROC curve) with our proposed loss, using pre-trained DenseNet121.

depending on images not patients, and Wang et al.[5] have given baseline results, we choose it to compare. According to the blue italic results, we can conclude that VGG16 and DenseNet121 have better performance as a baseline model. However, DenseNet121 achieves good result on average, especially for Hernia pathology(AUC = 0.9042). From Chest X-ray14, we can find that there are fewer samples to learn for Hernia. **Figure 1 (a)-(c)** shows results on 14 classes with three different basic pre-trained models: VGG16, ResNet50 and DenseNet121. We can find that DenseNet121 also performs well on these 4 diseases: Cardiomegaly(AUC = 0.8790), Pneumothorax(AUC = 0.8379), Fibrosis(AUC = 0.8040).

From left to right, the 1st to 5th, and last 2 columns in **Table I** shows performance comparison of weighted cross entropy loss, our proposed loss and other approaches, the first two of which use pre-trained DenseNet121 as a baseline model. From the black bold results, we can conclude that there are 13 pathology classification results better than baseline using our proposed loss. Compared to the state of arts, our proposed loss improves classification results for these diseases: cardiomegaly (AUC=0.9915), Mass (AUC=0.8652), Consolidation (AUC = 0.8117), edema (AUC=0.8979), hernia (AUC=0.9637). Due to the particularity of the hernia patient group, the number of this pathology is far less than other major diseases. Weight cross entropy loss considers the loss of each class, instead of considering overall. In general, the proposed loss is more sensitive to pathology with fewer samples, which leads to better performance. **Figure 1 (c)-(d)** shows the multi-label classification performance on 14 classes with two different loss using pre-trained DenseNet121 as a baseline model.

V. Conclusion

In this paper, we formulated chest X-ray diagnosis as multi-label classification problem. Taking advantage of enhancing feature propagation and reusing feature, DenseNet121 was utilized as a baseline model with its ability to avoid gradient disappearance. With implicitly coding the pathological labels, we fine-tuned the pre-trained model. Then we proposed an entropy weighting loss. From the global perspective of all pathological classes, digging relations between all classes, the proposed entropy weighting loss could improve model performance. The classification result on testing sub-dataset is much improved, especially for the pathology of "hernia".

Acknowledgment

This work was supported by National Natural Science Foundation of China (Grant Nos. 51775496, 51875503).

References

[1] A. Michienzi, T. Kron, J. Callahan, N. Plumridge, D. Ball, and S. Everitt, "Cone-beam computed tomography for lung cancer–validation with ct and monitoring tumour response during chemo-radiation therapy," *Journal of medical imaging and radiation oncology*, vol. 61, no. 2, pp. 263–270, 2017.

[2] R. H. Daffner and M. Hartman, *Clinical radiology: the essentials*. Lippincott Williams & Wilkins, 2013.

[3] G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005

[4] C. Parmar, R. T. Leijenaar, P. Grossmann, E. R. Velazquez, J. Bussink, D. Rietveld, M. M. Rietbergen, B. Haibe-Kains, P. Lambin, and H. J. Aerts, "Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer," *Scientific reports*, vol. 5, p. 11044, 2015.

[5] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 3462–3471. [Online]. Available: https://doi.org/10.1109/CVPR.2017.369

[6] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv preprint arXiv:1710.10501*, 2017.

[7] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 8290–8299.

[8] P. Kumar, M. Grewal, and M. M. Srivastava, "Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 546–552.

[9] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *arXiv preprint arXiv:1801.09927*, 2018.

[10] H. Wang and Y. Xia, "Chestnet: A deep neural network for classification of thoracic diseases on chest radiography," *arXiv preprint arXiv:1807.03058*, 2018.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90