Auxiliary Decoder and Classifier for Imbalanced Skin Disease Diagnosis

Zili Xu[1], Jiaxin Zhuang[1], Rong Zhang[1], Ruixuan Wang[1], Xuemei Guo[1] and Wei-Shi Zheng[1,2,3]

[1]School of Data and Computer Science, Sun Yat-sen University, China
[2]Department of Computer Science and Engineering, Southern University of Science and Technology, China
[3]Pazhou Lab, Guangzhou, China

E-mail: wangruix5@mail.sysu.edu.cn

Abstract. In medical image based intelligent diagnosis, class imbalance issue often appears due to the substantially smaller available training data for rare diseases compared to common diseases. Here we propose a novel learning framework to effectively solve this issue, adding an auxiliary decoder and reusing the original CNN classifier to help the classifier more likely extract disease-relevant features for both rare and common diseases. Experiments on two skin disease datasets supports that the proposed framework outperforms strong baselines and can be flexibly combined different model structures and existing training strategies.

1. Introduction

Nowadays, significant progress has been achieved in deep learning and relevant techniques have been applied in many medical diagnosis systems [4, 7, 9, 13, 14]. However, accurate diagnosis often relies on large amount of training data, while in many real applications, such as skin disease diagnosis [2, 15], the collected data may be very limited especially for rare diseases. The imbalanced training data between common and rare diseases could make deep learning models largely ignore discriminative features of rare diseases during model training resulting in biased predictions towards the common (large-sample) diseases. Since the class-imbalance issue has been extensively investigated over the last few decades, some traditional but effective methods have been adopted to deal with class imbalance in deep learning. One widely adopted approach is to augment small classes by simply over-sampling the data from these classes [8]. Oversampling can be indirectly implemented by various transformations of original data, such as flipping horizontally, and random rotation within certain range of degrees. Besides generating more data, cost sensitive methods, such as class weighting [11] and focal loss [6], are also proved to effective to alleviate the class-imbalance issue. The class-weighting method can help deep models pay more attention to small-class data during training, and focal loss can help models automatically select and focus on hard training data large of which are often from small classes In addition, transfer learning via pretrained model can be also useful to alleviate the class-imbalance issue [5]. While the traditional approaches have been widely adopted to handle the class-imbalance issue in training deep learning models, deep learning technique itself has seldomly been explored to help train models. One exception is the recent work which uses the

Grad-CAM attention map to help models focus on the lesion region in images of rare diseases during model training, resulting in improved diagnosis performance on both common and rare diseases [16]. Different from the attention method, in this paper, a simple but novel deep learning based framework is proposed to alleviate the class-imbalance mainly with the help of a decoder network. This is inspired by the idea that better image reconstruction from the higher layer of the CNN classifier would help the classifier likely extract more visual content, especially when the reconstruction faithfulness is enforced for images of rare diseases. Extensive experiments on two skin image datasets proved the effectiveness of the proposed framework.

2. Method

The objective of interest is to effectively handle the class imbalance problem such that small class(es) such as rare diseases can be well learned by the classifier. To help the classifier learn small classes, larger weights are often assigned to smaller classes such that the importance of each training data from small classes is emphasized and therefore can be correctly recognized by the classifier. However, due to limited training data for each small class, the classifier might be trained to overfit the data of small classes, i.e., learn to recognize each small class of data not entirely by the class-specific characteristics but by certain superficial features. With this consideration, it would be desirable if the feature vector for class prediction can contain more information of the original data, such that the class-specific information can be more likely encoded in the feature vector. Here we propose a simple way to achieve this goal, i.e., by adding a subsidiary decoder and an auxiliary CNN classifier to the (original) CNN classifier (Figure 1). Intuitively, if any input image can be well reconstructed from the feature output of the original CNN classifier, such feature vector should contain all essential visual information of the input image, including the discriminative information for accurate class prediction. Therefore, during training the CNN classifier, a decoder can be attached to the end of the feature extractor part of the original classifier to help the feature extractor output contain as much visual information of the input as possible, where the feature extractor would also work as an encoder. On the other hand, since the output of the decoder is often over-smoothed compared to the input image (e.g., discarding detailed information like high-frequency edges and textures), the class-specific information in the input image could be partly or mostly removed during the decoding process. To help the output of the decoder contain the essential discriminative information for each input image, we propose applying the original classifier again, as an auxiliary CNN classifier sharing the mode parameters with the original CNN classifier, to the classification of the reconstructed image from the decoder ('twin classifier network' in Figure 1). Overall, the twin CNN classifiers and the decoder can be jointly trained by minimizing the loss L,

$$L = L_c + \alpha L_r + \beta L_t \,, \tag{1}$$

where L_c can be the general cross-entropy loss or its variants like class-weighted cross-entropy for the original CNN classifier, L_t represents the same type of loss as that of the original CNN classifier for the auxiliary CNN classifier, and L_r is the (L_2 or L_1) reconstruction loss for the decoder. α and β are coefficients to balance the three loss terms. L_r can further decomposed to

$$L_r = \sum_{k=0}^{K} \omega_k L_k \,, \tag{2}$$

where L_k is the reconstruction loss for the k-th class of training images, and K is the total number of classes. ω_k is the class weight, with higher value for smaller classes and thus emphasizing that images from smaller classes should be reconstructed more faithfully. It is expected that the class weight would further help the output of feature extractor in the model keep all important (including the class-specific) information especially for small classes.

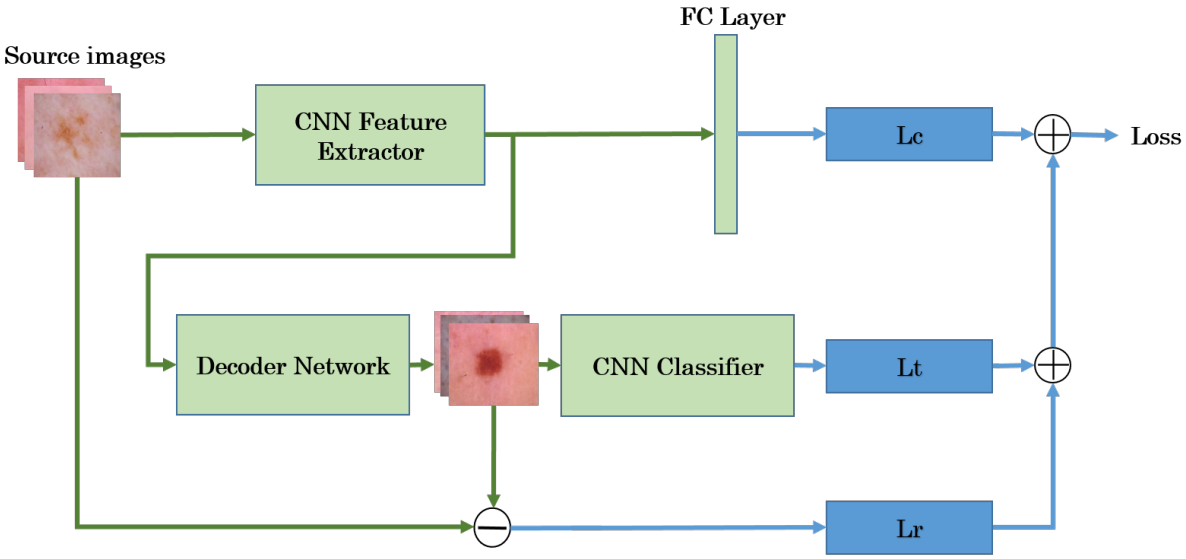

Figure 1. The proposed model framework. The green part represents the model structure, and the blue part represents the loss terms for model training.

3. Experiments

3.1. Experimental settings

Dataset. Two medical image datasets were used to evaluate the proposed approach. The first one is *Skin-7 dataset* provided by ISIC2018 Challenge with 7 disease catogories [1, 12], in which 6705 images are for Melanocytic nevus and only 115 images for Dermatofibroma, clearly having serious data imbalance between classes. The other is the *Skin-198 dataset* with 198 categories [10]. The smallest class contains only 10 samples and more than 70 classes contain less than 20 samples. All images were resized to 300×300 pixels and then randomly cropped to 224×224 pixels. For each dataset, images are randomly split into five folds with stratification for five-cross validation. Each time, we gather four folds as training set and the other one as the test set.

Implementation and Protocol. In the experiments, each encoder backbone was pretrained on ImageNet, while decoder was initialized by Kaiming Normal Initialization [3]. The decoder consists of 2 blocks, and each block contains 1 deconvolutional layers. α and β in Equation(1) were set to 10.0 and 0.2 respectively. SGD optimizer was used throughout, with initial learning rate set as 0.001 and momentum set as 0.9. The learning rate was divided by 10 at the $100th$ epoch. Each model was trained for up to 200 epochs, with the consistent observation of training convergence within 160 epochs. Considering the imbalance distribution across classes, *mean class f1-score* (MF1, i.e., average f1-score over all classes), *Precision* (i.e., average precision over all classes) and *Recall* (i.e., average recall over all classes) at the last training epoch were calculated on each validation set, and the mean and standard deviation of the measurements over all the five cross-validation sets were reported.

3.2. Results

In order to test the effectiveness of the proposed approach, we compared our method to two widely-used training strategies for handling data imbalance, namely, 1) *cost sensitive learning* (i.e., class-weighted cross-entropy loss, denoted by WCE) [11], and 2) *focal loss* [6] denoted by FL, a representative method of hard negative mining. The traditional cross-entropy loss (BCE) and the class-weighted focal loss (WFL) were also used as baseline training strategies.

Table 1. Comparison between the proposed approach and baseline methods based on cross-entropy loss on Skin-7 and Skin-198 datasets.

Approaches	Skin-7			Skin-198		
	MF1	Precision	Recall	MF1	Precision	Recall
BCE	84.44 (0.55)	86.78 (1.48)	82.92 (1.35)	65.75 (1.21)	65.72 (1.20)	66.85 (1.01)
WCE	84.53 (1.01)	86.12 (1.24)	83.52 (1.22)	65.45 (1.46)	65.45 (1.45)	66.31 (1.50)
CD (ours)	85.93 (1.30)	88.58 (1.37)	83.92 (1.35)	67.69 (1.56)	67.67 (1.55)	68.62 (1.54)
CDC (ours)	**86.75** (0.82)	**89.26** (0.37)	**84.81** (0.98)	**67.97** (1.00)	**67.96** (1.01)	**69.16** (0.89)

Table 2. Comparison between the proposed approach and baseline methods based on focal loss on Skin-7 and Skin-198 datasets.

Approaches	Skin-7			Skin-198		
	MF1	Precision	Recall	MF1	Precision	Recall
FL	85.34 (0.62)	87.94 (0.74)	82.97 (0.89)	65.33 (1.35)	65.33 (1.33)	66.11 (1.32)
WFL	86.48 (0.58)	89.37 (0.36)	84.22 (1.12)	65.46 (1.97)	65.41 (1.98)	66.37 (1.93)
CD (ours)+FL	86.35 (0.95)	88.56 (1.21)	84.59 (0.88)	67.33 (1.41)	67.35 (1.41)	68.21 (1.16)
CDC (ours)+FL	**86.92** (1.08)	**89.62** (0.84)	**84.84** (1.70)	**67.52** (2.32)	**67.52** (2.36)	**68.61** (1.85)

For fair comparison with each baseline, the proposed model (denoted by CDC) was trained by the same baseline training strategy each time. The performance of the ablation version without the auxiliary CNN classifier (denoted by CD) was also reported. From Tables 1 and 2, it can be observed that the proposed framework outperforms all the baselines on both Skin-7 and Skin-198 datasets. In particular, the improvement is also clear on the small-sample classes (Table 3, average performance over the smallest classes on Skin-7 and the 40 smallest classes on Skin-198) compared to the baselines BCE and WCE. Similar improvement was also observed when compared to the FC and WFC baselines on the small-sample classes (not shown due to limited space). Figure 2 demonstrates the performance of various approaches on one validation set during the training process, confirming that the all trainings are converged and the proposed framework consistently outperform the baselines. In addition, while all the reported results were based on the ResNet50 backbone, similar performance was also observed when using the VGG and DenseNet backbones, supporting that the proposed framework is generalizable and not limited to specific CNN backbone.

4. Conclusion

In conclusion, this paper proposed a novel and effective way to help handle the class-imbalance issue, mainly by using a subsidiary decoder to help the CNN classifier more likely extract disease-

Table 3. Performance of methods based on cross-entropy loss on small-samples classes of Skin-7 and Skin-198 datasets.

Approaches	Skin-7			Skin-198		
	MF1	Precision	Recall	MF1	Precision	Recall
BCE	75.13 (6.90)	83.26 (5.50)	69.22 (11.24)	64.16 (2.59)	65.63 (4.36)	68.58 (3.84)
WCE	75.40 (8.12)	83.09 (5.09)	70.58 (10.95)	65.78 (3.26)	69.13 (4.62)	69.87 (4.76)
CD (ours)	78.02 (6.46)	88.20 (5.84)	70.61 (9.89)	**69.32** (4.66)	71.53 (6.03)	73.37 (6.56)
CDC (ours)	**80.28** (7.64)	**90.25** (3.99)	**72.78** (10.06)	69.05 (4.71)	**75.09** (3.94)	**75.33** (4.94)

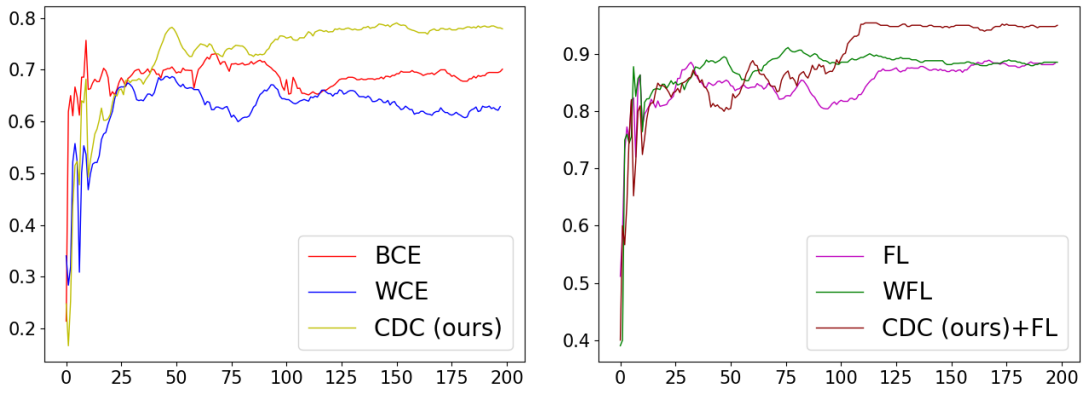

Figure 2. MF1 curves of different methods over the smallest classes on one Skin-7 validation set with respect to training epochs. The training was converged around 160 epochs and the performance of the proposed framework is consistently better than corresponding baselines.

relevant visual features. Experiments on two skin image datasets showed that the proposed learning framework can improve not only the overall average classification performance over all diseases, but more importantly on the small-class (often corresponding to rare) diseases. The proposed framework is independent of existing training strategies and model backbones, and therefore can be easily combined with existing strategies and various CNN models.

References

[1] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, and Harald Kittler. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging. In *ISBI*, 2018.

[2] Yasuhiro Fujisawa, Sae Inoue, and Yoshiyuki Nakamura. The possibility of deep learning-based, computer-aided skin tumor classifiers. *Frontiers in Medicine*, 6:191, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[4] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

[5] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016.

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.
[7] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.
[8] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.
[9] Berkman Sahiner, Aria Pezeshk, Lubomir M Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H Cha, Ronald M Summers, and Maryellen L Giger. Deep learning in medical imaging and radiation therapy. *Medical physics*, 46(1):e1–e36, 2019.
[10] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *ECCV*, 2016.
[11] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007.
[12] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.
[13] X Wang, Y Peng, L Lu, Z Lu, M Bagheri, and RM Summers. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.
[14] Lihteh Wu, Priscilla Fernandez-Loaiza, Johanna Sauma, Erick Hernandez-Bogantes, and Marissé Masis. Classification of diabetic retinopathy and diabetic macular edema. *World journal of diabetes*, 4(6):290, 2013.
[15] Jianpeng Zhang, Yutong Xie, Qi Wu, and Yong Xia. Medical image classification using synergic deep learning. *Medical image analysis*, 54:10–19, 2019.
[16] Jiaxin Zhuang, Jiabin Cai, Ruixuan Wang, Jianguo Zhang, and Weishi Zheng. Care: Class attention to regions of lesion for classification on imbalanced data. In *MIDL*, 2019.