Auxiliary Decoder and Classifier for Imbalanced Skin Disease Diagnosis

Zili Xu\textsuperscript{1}, Jiaxin Zhuang\textsuperscript{1}, Rong Zhang\textsuperscript{1}, Ruixuan Wang\textsuperscript{1*}, Xuemei Guo\textsuperscript{1} and Wei-Shi Zheng\textsuperscript{1,2,3}

\textsuperscript{1} School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong, China
\textsuperscript{2} Department of Computer Science and Engineering, Southern University of Science and Technology, Guangzhou, Guangdong, China
\textsuperscript{3} Pazhou Lab, Guangzhou, Guangdong, China
E-mail: wangruix5@mail.sysu.edu.cn

Abstract. To date, deep learning has been widely adopted in medical diagnosis systems and made great achievements in real-world applications. However, in medical image-based intelligent diagnosis, the phenomenon of class imbalance often appears due to the substantially smaller available training data for rare diseases compared to common diseases, which usually degrade the performance of classification dramatically. In this paper, we propose a novel learning framework to effectively alleviate the impact of class imbalance, adding an auxiliary decoder to reconstruct original images and reusing the original CNN classifier to help the classifier more likely extract disease-relevant features for both rare and common diseases. Throughout experiments on two skin disease datasets support that the proposed framework outperforms strong baselines with a very explicit margin. The proposed method is also independent of network architecture so that it can be flexibly combined with different model structures. Besides, applying our proposed method with existing training strategies designed for class imbalance can still improve the classification performance.

1. Introduction
Nowadays, significant progress has been achieved in deep learning and relevant techniques have been applied in many medical diagnosis systems [1-8]. However, accurate diagnosis often relies on a large amount of training data, while in many real applications, such as skin disease diagnosis [9-11], 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 the class imbalance in deep learning. One widely adopted approach is to augment small classes by simply over-sampling the data from these classes [12]. Oversampling can be indirectly implemented by various transformations of original data, such as flipping horizontally, and random rotation within a certain range of degrees. Besides generating more data, cost-sensitive methods, such as class weighting [13] and focal loss [14], 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 [15]. 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 that 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]. 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 the 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 a 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 $L$:

$$L = L_c + \alpha L_r + \beta L_t$$

(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. $\alpha$ and $\beta$ are coefficients to balance the three loss terms. $L_r$ can further be decomposed to

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

(2)

where $L_k$ is the reconstruction loss for the $k$th class of training images, and $K$ is the total number of classes. $\omega_k$ is the class weight, with a 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.
3. Experiments

3.1. Experimental Settings
Dataset. Two medical image datasets were used to evaluate the proposed approach. The first one is the Skin-7 dataset provided by ISIC2018 Challenge with 7 disease categories [17, 18], 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 [19]. 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 a training set and the other one as the test set.

Implementation and protocol. In the experiments, each encoder backbone was pretrained on ImageNet, while the decoder was initialized by Kaiming Normal Initialization [20]. The decoder consists of 2 blocks, and each block contains 1 deconvolutional layer. α and β in equation (1) were set to 10.0 and 0.2 respectively. SGD optimizer was used throughout, with an 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 overall 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), and 2) focal loss [14] 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. For a fair comparison with the 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 1-2, 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 all training is converged and the proposed framework consistently outperforms the baselines. Besides, 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.

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

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

**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 | 83.26 | 69.22 | | 64.16 | 65.63 | 68.58 | |
| | (6.90) | (5.50) | (11.24) | | (2.59) | (4.36) | (3.84) | |
| WCE | 75.40 | 83.09 | 70.58 | | 65.78 | 69.13 | 69.87 | |
| | (8.12) | (5.09) | (10.95) | | (3.26) | (4.62) | (4.76) | |
| CD (ours) | 78.02 | 88.20 | 70.61 | | 69.32 | 71.53 | 73.37 | |
| | (6.46) | (5.84) | (9.89) | | (4.66) | (6.03) | (6.56) | |
| CDC (ours) | 80.28 | 90.25 | 72.78 | | 69.05 | 75.09 | 75.33 | |
| | (7.64) | (3.99) | (10.06) | | (4.71) | (3.94) | (4.94) | |
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 CNN classifier more likely to extract disease-relevant visual features. Experiments on two skin image datasets showed that the proposed learning framework can improve not only the overall average classification performance overall 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] Hesamian M H, Jia W J, He X J and Kennedy P 2019 Deep learning techniques for medical image segmentation: Achievements and challenges Journal of Digital Imaging 32 (4) 582-596.

[2] Lundervold A S and Lundervold A 2019 An overview of deep learning in medical imaging focusing on MRI Zeitschrift für Medizinische Physik 29 (2) 102-127.

[3] Sahiner B, Pezeshk A, Hadjiiski L M, Wang X S, Drukker K, Cha K H, Summers R M and Giger M L 2019 Deep learning in medical imaging and radiation therapy Medical Physics 46 (1) 1-36.

[4] Wang X S, Peng Y F, Lu L, Lu Z Y, Bagher M and Summers R 2017 Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases CVPR.

[5] Shen D G, Wu G R and Suk H 2017 Deep learning in medical image Analysis Biomedical Engineering 19 (1) 221-248.

[6] Yi X, Walia E and Babyn P 2019 Generative adversarial network in medical imaging: A review Medical Image Analysis 58 101552.

[7] Jo S, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B and Rueckert D 2019 Attention gated networks: Learning to leverage salient regions in medical images Medical Image Analysis 53 197-207.

[8] Guha B, Zhao A, Sabuncu M, Guttag J and Dalca A 2019 VoxelMorph: A learning framework for deformable medical image registration IEEE Transactions on Medical Imaging 38 (8) 1788-1800.

[9] Andre E, Kuprel B, Novoa R, Ko J, Swetter S, Blau H and Thrun S 2017 Dermatologist-level
classification of skin cancer with deep neural networks *Nature* 542 (7639) 115-18.

[10] Zhang J P, Xie Y T, Wu Q and Xia Y 2019 Medical image classification using synergic deep learning *Medical Image Analysis* 54 10-19.

[11] Titus B, Hekler A, Enk A, Berking C, Haferkamp S, Hauschild A, Weichenthal M, Klode J, Schadendorf D and Holland-Letz T 2019 Deep neural networks are superior to dermatologists in melanoma image classification *European Journal of Cancer* 119 11-17.

[12] Mateusz B, Maki A and Mazurowski M 2018 A systematic study of the class imbalance problem in convolutional neural networks *Neural Networks* 106 249-259.

[13] Salman K, Munawar H, Mohammed B, Ferdous S and Roberto T 2017 Cost-sensitive boosting for classification of imbalanced data *IEEE Transactions on Neural Networks and Learning Systems* 29 (8) 3573-3587.

[14] Lin T, Goyal P, Girshick R, He K M and Dollar P 2017 Focal loss for dense object detection *ICCV*.

[15] Huang C, Li Y N, Loy C and Tang X O 2016 Learning deep representation for imbalanced classification *CVPR*.

[16] Zhuang J X, Cai J B, Wang R X, Zhang J G and Zheng W S 2019 Care: Class attention to regions of lesion for classification on imbalanced data *MIDL*.

[17] Codella N, Gutman D, Celebi E, Helba B, Marchetti M, Dusza S, Kalloo A, Liopyris K, Mishra N and Kittler H 2018 Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging *ISBI*.

[18] Tschandl P, Rosendahl C and Kittler H 2018 The ham10000 dataset, a large collection of multisource dermatoscopic images of common pigmented skin lesions *Scientific Data* 5 180161.

[19] Sun X X, Yang J F, Sun M and Wang K 2016 A benchmark for automatic visual classification of clinical skin disease images *ECCV*.

[20] He K M, Zhang X Y, Ren S Q and Sun J 2015 Delving deep into rectifiers: Surpassing human level performance on imagenet classification *ICCV*.