Skin Lesion Analysis Towards Melanoma Detection Using Deep Neural Network Ensemble

Jiaxin Zhuang[†], Weipeng Li[†], Siyamalan[‡] [*]
Roy Wang[†], JianGuo Zhang, Jihan Liu[†], Jiahui Pan[†], Ziyu Yin[†]
[†]MIA Group, School of Data and Computer Science, Sun Yet-sen University, China
[‡]Computing School of Science and Engineering, University of Dundee.

Email: zhuangjx5@mail2.sysu.edu.cn

Abstract

This paper reports the method of MIA group from Sun Yat-Sen University for ISIC 2018 Challenge Task 3: Disease Classification. Our approach aims to combine deep neural network ensemble with several state-of-the-art techniques to get better performance in this task. Our method finally received a validation score of 92.3% online.

1 Introduction

The task 3 of ISIC 2018 *Skin Lesion Analysis Towards Melanoma Detection* is a common but challenging image classification task. It requires the proposed algorithm to automatically classify dermoscopic lesion images [1] into seven disease categories as shown in **Tab.** 1. Participants were ranked by the mean precision across seven categories as follows:

$$MCA = \sum^{n} P_i/n \tag{1}$$

where MCA denotes *Multi-Class Accuracy* and P_i denotes the precision of the ith class.

To address this task, we proposed a deep learning based classification architecture in this paper. In Section 2, we describe the major difficulty encountered in the classification task. Section 3 introduces our methodology and shows our proposed classification system. Section 4 reports our evaluation results and Section 5 is the conclusion of our work.

2 Major Difficulty for Classification

With the rapid development of computer vision techniques, especially in the field of deep learning, a large number of Convolutional Neural Networks (CNNs) have been widely used in image classification and detection. It is well acknowledged that CNNs show excellent performance in classifying images [2] and achieve higher accuracy than many traditional methods. Therefore, we applied CNN-based models in Lesion Diagnosis Task. However, the data provided for the Task 3 indicate two characters that inhibit the performance of deep neural networks.

1. Data is imbalanced between seven categories. For example, images of Melanocytic Nevus is approximately 58 times as many as Dermatofibroma.

2. All the categories except Melanocytic Nevus have too small training datasets to meet the need of a deep neural network.

[*] The first three author contributed equally to this work

3 Proposed Classification System

After comparing the performance of different CNN-based models in this task, we adopted the ensemble of Squeeze-and-Excitation Network (SENet) [3] and Progressive Neural Architecture Search Net (PNASNet-5-Large) [4] as our final classification model. These two models exhibit powerful abilities in learning features for classification task in the case of insufficient training data. We used the ensemble of these two models for final prediction and thus received better results. All our codes are put in github for reference[1].

3.1 Framework of Classification System

At first, we separately trained the SENet and PNASNet-5-Large with the original data and the preprocessed data. Before the models behave a bit overfitted in training dataset, we stopped training and used the networks to predict the results on test dataset. Then, the results are combined together to generate the final classification output. **Fig.** 1 shows the framework of our classification system.

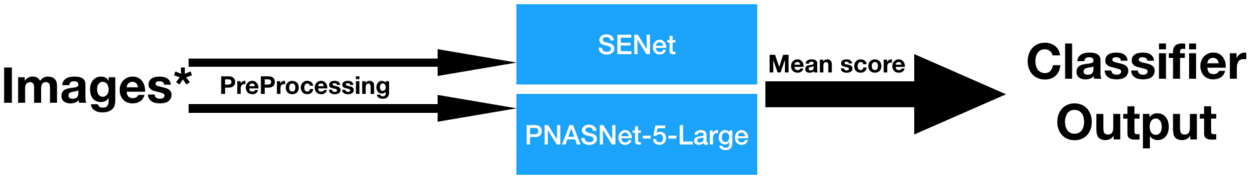

Figure 1: Framework of Classification System

3.2 Preprocessing and Preparation of Ground Truth

Image preprocessing is one of the most important parts in image classification tasks. We adopted two preprocessing methods for our data, which are color constancy algorithms and data augmentation.

3.2.1 Color Constancy Algorithms

Firstly, in order to eliminate the variance of luminance and color, we used color constancy [5] algorithms to normalize original images. The difference between original image and processed image by using color constancy algorithm are shown in **Fig.** 2 and **Fig.** 3.

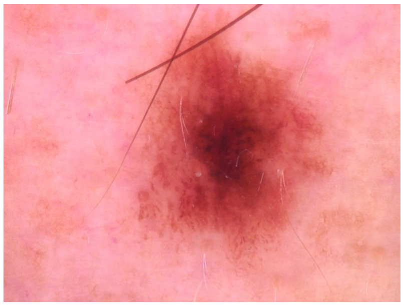

Figure 2: Training image without using color constancy

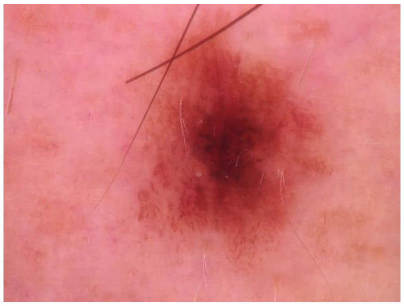

Figure 3: Training image with using color constancy

[1] https://github.com/orgs/SYSU-MIA/

3.2.2 Data Augmentation

The original images are resized to 300×300 for SENet and 441×441 for PNASNet to reduce computational costs and improve the performance. In addition, we randomly flipped the images horizontally and vertically. We also randomly changed the brightness, contrast, saturation of the images, rotated an image from $-180°$ to $+180°$ and applied affine transformation. The images are randomly cropped to 224×224 for SENet and 331×331 for PNASNet.

3.3 Overcome the Major Difficulty

Tab. 1 shows that the major difficulty in this task is imbalanced data. The largest number of images for one class is up to 6705 whereas the smallest one is only 115.

Disease	Amount
Melanoma (MEL)	1113
Melanocytic nevus (NV)	6705
Basal cell carcinoma (BBC)	514
Actinic keratosis / Bowens disease (intraepithelial carcinoma) (AKIEC)	327
Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis) (BKL)	1099
Dermatofibroma (DF)	115
Vascular lesion (VASC)	143

Table 1: Numbers of images for each class in ISIC 2018 Challenge Task 3.

To address the imbalanced classes, we tried the following techniques that could improve the MCA [2] by at least 10%.

1. Class Weighted Loss Function.
2. Focal Loss Function [6].

We used 5-fold cross-validation on the provided training data to evaluate the performance of our methods. Class-weighted cross-entropy loss are selected as the loss function so as to punish harder on the false predictions on those classes with smaller datasets. As a result, we got a big promotion in MCA metric. The confusion matrix 4 on one of our validation datasets further proves that the model could pay more attention to categories with smaller dataset than normal model after using class weighted loss function.

Although we decided to use class-weighted cross-entropy as our loss function finally, focal loss is still worth trying. According to the experimental results, we found the model using focal loss has more stable performance (i.e., smaller variance on prediction accuracy) on validation dataset than using class weighted loss.

3.4 Other Trials for the Imbalanced Data

Although the ensemble model of SENet and PNASNet achieved the best classification results so far, we have also been trying several other methods to address the imbalanced problem.

3.4.1 Triplet Loss and Contrastive Loss

Triplet loss is the loss function implemented in the state-of-the-art face recognition FaceNet [7] which achieved accuracy of 99.63% on the LFW dataset. Contrastive loss is the loss function adopted in the Siamese Network [8] which also proved to have powerful ability in face verification. We tried using triplet loss and contrastive loss as our loss function in place of cross entropy loss to reduce the negative effect of imbalanced data. However, the results are not so good as we expected since we couldn't select as many hard pairs or triplets easily for the skin lesion as we do in face recognition task. We are considering to use the K nearest neighbor clustering and support vector machines to classify the learned embeddings in the future.

[2] Multi-Class Accuracy

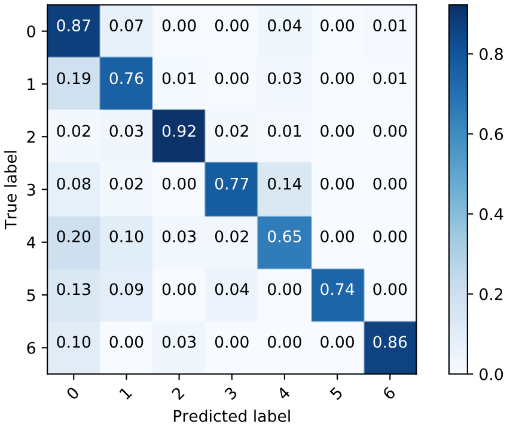

Figure 4: Confusion Matrix of Classification System
on one of our own validation dataset

3.4.2 Clustering over More Categories

Just as mentioned above, the major difficulty of this task is imbalanced data. We tried to train a VGG19 [9] to extract the feature maps of original images, and then used clustering method to divide large groups' images into small group. Therefore, we could get many small classes to classify.

3.4.3 Undersampling and Oversampling

For imbalanced class, we also tried undersampling over more images categories and oversampling over less images categories, but neither of them could bring obvious improvement and even sometimes got worse.

3.4.4 Teacher-Student Network

We trained a teacher network [10] and used all training data and obtained its corresponding probabilities for each categories. Then we used training data with probabilities of each categories to train the student model. Nevertheless, the accuracy didn't achieve as high as what we desired.

3.4.5 Traditional Machine Learning

We also tried to obtain the feature maps produced by CNN at first and then used SVM model to be the classifier. The MCA of it could achieve a bit more than 70%.

3.5 Candidate Models and Ensemble Models

After comparing the results with all the other methods above, we finally adopted the ensemble model of SENet-154 and PNASNet-5-Large, used the class weighted loss function and fine-tuned the network for skin lesion classification. Both the SENet-154 and PNASNet-5-Large were pre-trained on ImageNet and it turns out that the models which show higher accuracy in ImageNet also prove to be of higher accuracy in classification.

For one particular test sample x, we have N sub-models for ensemble model. For those models, we have N 7-dimensional vectors named $s_1, s_2, s_3, \ldots, s_N$. We have tried three ways to do model ensemble as follows.

Direct Average: Direct average is the simplest and most effective way to do model ensemble.

$$Finalscore = \frac{\sum_{i=1}^{N} s_i}{N} \tag{2}$$

Weighted Average: The weighted average method is based on direct average, and add weight to adjust the importance of the output of different models. The equation 3 shows the algorithm, where ω_i is the weight of the i-th model, $w_i >= 0$ and $\sum_{i=1}^{N} \omega_i = 1$

$$Finalscore = \frac{\sum_{i=1}^{N} \omega_i s_i}{N} \tag{3}$$

Voting: Firstly, we need to convert the output of each model into predicted label. And then, get the final result by using majority voting system.

The ensemble systems above showed a slightly different performance. So we took *Weighted Average* as the final model ensemble.

4 Evaluation Results

By comparing different models on online validation dataset and our own validation dataset, we finally decided to ensemble PNASNet and SENet to generate our final predictions. The evaluation results are shown in the **Tab.** 2.

Model	MCA on Official Validation Data	MCA on 5-fold Validation Data
PNASNet	88.7 ± 1.0	82.6 ± 2.0
SENet	89.8 ± 1.2	81.0 ± 1.0
Ensemble SENet	91.7	—
Ensemble SENet with PNASNet	92.3	—

Table 2: Evaluation Result

5 Conclusions

To summarize, we provided an ensemble of SENet and PNASNet as our final deep neural network architecture. To make full use of the imbalanced data in Task 3, we tried different preprocessing techniques and implemented a lot of trials in the loss functions. Although some inspiring ideas may not work very well in the actual experiments due to the lack of effective training data, we have made great progress and obtain a high accuracy of 92.3% on the validation data. Our future emphasis may lie on more effective pre-processing or sample techniques as well as network designs to solve the bottleneck of the imbalanced data in skin lesion analysis.

References

[1] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *CoRR*, abs/1803.10417, 2018.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017.

[4] Chenxi Liu, Barret Zoph, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. *arXiv preprint arXiv:1712.00559*, 2017.

[5] Catarina Barata, M Emre Celebi, and Jorge S Marques. Improving dermoscopy image classification using color constancy. *IEEE journal of biomedical and health informatics*, 19(3):1146–1152, 2015.

[6] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[7] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[8] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[10] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *ArXiv e-prints*, March 2015.