Research on Diagnosis of Dermatology Based on Deep Residual Neural Network

Jiayuan Wang\textsuperscript{1, a}, Weiye Wang\textsuperscript{1, b}*, and Tian Tian\textsuperscript{2}

\textsuperscript{1}School of Automation, Beijing Information Science and Technology University, Beijing, China
\textsuperscript{2}Beijing Information Science and Technology University, Beijing, China

\textsuperscript{a}wjy@wjyai.com; \textsuperscript{b}afatfox@163.com;

Abstract. The methods of dermatological clinical examination are mainly skin images, including dermoscopy. Residual neural network (ResNet) can predict diseases according to dermoscopy images and provide effective proposals for doctors. Based on the ResNet model, this article migrated the pre-trained model on ImageNet to simulation experiment, and used the Focal Loss function to solve the problem of experimental sample imbalance, including but not limited to operations such as flip, rotation, scaling, and loss function replacement, thereby improving network performance. The experimental results show that the model trained by our method can reach completely correct when it classified a small number of samples. Our model can reach accuracy rate of 90.08\%, recall rate of 88.44\%, and F1 score of 85.25\%. Compared with the model with unmodified loss function at the same depth, our model has respectively improved by 1.3\%, 4.62\%, and 3.58\% in the above three aspects, which indicates that our method is effective in predicting rare diseases, and in predicting common diseases the accuracy rate also achieves good results.

1. Introduction
With the rapid development of computer, computer-aided diagnosis technology has made good progress in the medical field\cite{1-3}. In recent years, the incidence of skin diseases has continued to increase. There are many types of dermatoses that require an experienced dermatologist to diagnose\cite{4-5}. Dermatoscope is a non-invasive clinical diagnostic imaging technology with high diagnosis rate\cite{6}. At the grassroots level, there is a lack of experienced clinicians. Faced with this problem, image analysis algorithms of deep convolutional neural networks can be used to predict the symptoms of dermoscopy images, thereby assisting doctors to give diagnosis results and improve the accuracy of diagnosis results\cite{7-8}. Labeling dermoscopy images requires extremely high dermatological medical knowledge. The quality of training sample data determines the final predictive performance of a network. Therefore, expert-level specialists are required to label a large amount of sample data. The cost is extremely high. Nowadays there are only a handful of shaped skin disease sample datasets, so it is difficult to study the classification of dermoscopy images. Through its own training, traditional CNN (Convolutional Neural Networks) obtains features from the data set. As the depth increases, the model can better approximate the sample data, but the difficulty of the model increases dramatically. The effect of model training may even degrade. ResNet proposed by Kaiming He et al. \cite{9} can effectively solve the problem of network degradation as the number of layers increases.
This article based on ResNet model applies the Focal Loss function to solve the problem of experimental sample imbalance, and migrated the pre-trained model on ImageNet to simulation experiment in order to improve the network performance. The experimental simulation illustrates that our model is effective in predicting rare diseases, and in predicting common diseases the accuracy rate also achieves good results.

The model of deep ResNet is detailed in the second part. In the third part, we describe the experimental simulation and results. The paper is concluded in the fourth part.

2. The model

The rapid development of deep learning has made it a place in the field of medical image analysis, including classification, detection, and recognition. Compared with traditional algorithms, deep learning does not need to provide features to the network. It only needs to pass a large amount of sample data to the network. Through convolution-pooling operations, it is usually possible to train a good model. In this way, there is no need to manually extract features, which greatly reduces the threshold for use and saves manual costs, especially in domains with strong medical specialties. Medical image features require expert-level or even top-level physicians to multiply mark in order to obtain the approved labels and features which all experts unanimously agree with. The labor and time costs are greatly increased.

2.1. Deep ResNet

When the deep neural network reaches a certain number of layers, the performance will approach saturation. Increasing the number of layers further will lead to the degradation of network performance, namely the disappearance of gradients and the explosion of gradients. This is not caused by overfitting. The accuracy of both training and verification is decreasing, so deep neural networks become difficult to train. ResNet makes up for this flaw. Its basic structure is a residual block, and multiple residuals form a residual network. Residual network is based on CNN, adding jump layer connections outside the weight layer to form one residual block as a part of the network. Let the input of the residual block be \( x \) and the output be \( H(x) = F(x) + x \). If \( F(x) \) is 0, \( H(x) = x \), that is, the input is equal to the output, which is called identity mapping. The residual is \( F(x) = H(x) - x \). The goal of training is to make \( F(x) \) approaching 0. Only identity mappings left. In this way, the network is always in an optimal state in theory, thereby avoiding the situation where the network's effect is degraded due to the increase in depth. Experiments have shown that \( F(x) \) is closer to 0 than \( F(x) = x \) is much easier [10].

The networks involved in this article are ResNet34, ResNet50, ResNet101, and ResNet152. The detailed structure of the model is shown in Table 1.

| layer name | Output size | 34-layer | 50-layer | 101-layer | 152-layer |
|------------|-------------|----------|----------|-----------|-----------|
|            |             | 3×3 max pool, stride 2 |          |           |           |
| Conv1      | 112×112     | 7×7, 64, stride 2 |
| Conv2_x    | 56×56       | \([3×3, 64] ×2\) | \(1×1, 64\) | \(1×1, 64\) | \(1×1, 64\) |
|            |             | \([3×3, 64] ×3\) | \(3×3, 256\) | \(3×3, 256\) | \(3×3, 256\) |
| Conv3_x    | 28×28       | \([3×3, 128] ×4\) | \(1×1, 128\) | \(1×1, 128\) | \(1×1, 128\) |
|            |             | \([3×3, 128] ×3\) | \(3×3, 512\) | \(3×3, 512\) | \(3×3, 512\) |
| Conv4_x    | 14×14       | \([3×3, 256] ×6\) | \(1×1, 256\) | \(1×1, 256\) | \(1×1, 256\) |
|            |             | \([3×3, 256] ×3\) | \(3×3, 1024\) | \(3×3, 1024\) | \(3×3, 1024\) |
| Conv5_x    | 7×7         | \([3×3, 512] ×4\) | \(1×1, 512\) | \(1×1, 512\) | \(1×1, 512\) |
|            |             | \([3×3, 512] ×3\) | \(3×3, 2048\) | \(3×3, 2048\) | \(3×3, 2048\) |
|            |             | \([1×1, 2048] \) | \(1×1, 2048\) | \(1×1, 2048\) | \(1×1, 2048\) |

| 1×1       | Average pool, 1000-d fc, softmax |
2.2. Loss function

For classification problems, the cross entropy loss function is generally used. If a task is divided into \( n \) categories, the cross-entropy loss function is defined as:

\[
CE = - \sum_{i=1}^{n} y_i \log(\hat{y}_i)
\]

where \( y_i \) is the actual category. In the one-hot method, for the sample of the \( m \) class, \( y_m = 1 \) and another \( y \) are 0. \( \hat{y}_i \) is the predicted output. All comparative experiments in this paper use cross-entropy loss functions.

Aiming at the problem of uneven sample size, this paper proposes to use Focal Loss function for correction [11]. It is defined as:

\[
FL(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t)
\]

where \( \alpha \) is the balance factor, used to balance the imbalanced categories. Introducing \( \gamma \) can reduce the loss of easy-to-classify samples and make the model pay more attention to difficult and misclassified samples.

In a study by Kaiming He et al., It was found that \( \alpha = 0.25 \) and \( \gamma = 2 \) work best. The paper introduces that \( \alpha = 0.5 \) and \( \gamma = 0.25 \) have similar results. But after the experimental verification of our data set, it is found that \( \alpha = 0.25 \) is better than \( \alpha = 0.5 \). Therefore, the experiments in this paper use \( \alpha = 0.25 \) and \( \gamma = 2 \) to compare with the results of the original training model.

3. Experiments and results

The experiments in this paper are run on Google server Colab. The experimental platform uses Ubuntu 18.04.3 system, Intel (R) Xeon (R) CPU @ 2.20GHz processor, 12GB memory, graphics cards are NVIDIA Tesla K80 11GB and Tesla P100 16GB. Two graphics cards are randomly assigned. Because the computing power of different graphics cards is different, the time required for training is also different under the same model. Therefore, the experimental results in this article do not include time comparison. The experiments in this paper are based on the PyTorch framework combined with the fast.ai library, using Python 3.6.8 language.

3.1. Data set

The data set used for the experiments in this paper is from the International Skin Imaging Collaboration (ISIC)[12-13]. The dataset is input as a dermatoscopic lesion image of 600*450 size three-channel RGB in JPEG format with a total of 10015 images. We will randomly sample 10% of the data in each disease as the validation set. Because random extraction is performed by taking integers from small ones, in the end, there are 998 items as the validation set, and other items as the training set. Labels are stored in a CSV file. The dataset contains a total of 7 diseases, and each image is given a consistent label by at least three dermatologists. Therefore, the reliability of this data set is relatively high. The disease distribution of this data set is shown in Table 2.

| Disease | The number of images |
|---------|----------------------|
| MEL     | 1113                 |
| NV      | 6705                 |
| BCC     | 514                  |
| AKIEC   | 327                  |
| BKL     | 1099                 |
| DF      | 115                  |
| VASC    | 142                  |

Table 2. Distribution of data set.
It can be seen from Table 2 that there is an imbalance in the category of the data set. Therefore, we performed an oversampling operation on this data set, so that the number of training dataset images for each type of disease is 6035. This can prevent the prediction result from biasing to the category with many samples, thereby avoiding the phenomenon of higher accuracy but poor generalization ability.

In order to improve the robustness and prevent over-fitting, the experiments in this paper have performed image enhancement to improve the quality of the data set. Enhancement operations include random flip, random rotation, random scaling, changes in brightness and contrast, random symmetrical transformation, vertical or horizontal flip 90°. Crop or pad the image to a size of 224 x 224 for subsequent batch processing. The comparison between the original image and the enhanced image is shown in Figure 1.

![Original image and enhanced pictures](image)

**Figure 1.** Original and enhanced pictures.

3.2. Network performance evaluation standard
The evaluation methods used in the experiments in this paper are Recall, Accuracy, and F1 Score. The Recall is the probability of being predicted correctly in the actual positive sample. Accuracy is the proportion of correctly predicted results in the total number of samples. The F1 score considers both precision and recall and considers both to be equally important. Note: Precision is not accuracy. Precision is the proportion of correctly predicted positive samples among all predicted positive samples. They are defined as:

\[
Re = \frac{TP}{TP+FN} \tag{3}
\]

\[
Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}
\]

\[
F1 = \frac{2 \cdot Re \cdot Pre}{Re + Pre} \tag{5}
\]

TP is true positive, that is, the number of correct predictions for the classification. FN is false positive, which is predicted as this type of disease, but it is not actually this type of disease. TN is true negative, that is, it is predicted to be not a disease of this type and correct. FN is false negative, i.e. underreporting. It means that the type of disease cannot be predicted correctly. Pre is accuracy.

3.3. Training and results
The experiments in this paper use pre-trained ResNet network models with different depths on the ImageNet dataset for migration training. Process the data set using the data enhancement methods described above. Change data to databunch format after data enhancement. And use imagenet_stats standard for data normalization. Training is divided into two phases. The first stage: First, only 5 cycles of training are performed on the parameters of the fully connected layer. The second stage: 5
more cycles of training under the parameters of the thawed convolution layer. The last 5 cycles set the learning rate between $10^{-6}$ and $10^{-4}$.

The experiments in this paper compare the different depths of the ResNet network model to find the best model depth. The comparison results are shown in Table 3.

**Table 3.** Comparison of different model results.

| The model   | Recall   | Accuracy | F1 Score |
|-------------|----------|----------|----------|
| ResNet 34   | 83.31%   | 80.86%   | 76.30%   |
| ResNet 50   | 83.93%   | 86.47%   | 79.91%   |
| ResNet 101  | 83.82%   | 88.78%   | 81.67%   |
| ResNet 152  | 84.69%   | 87.07%   | 81.60%   |
| ResNet 101 + FL | 86.95%   | 90.08%   | 85.22%   |

Through comprehensive comparison of different depth network models, it is found that ResNet 101 has the best network performance. For this model, we adopted a strategy to modify the loss function and found that the effect was improved in all aspects compared with the original model. Recall increased by 3.13%, accuracy increased by 1.3%, and F1 score increased by 3.55%. This shows that the modified loss function method proposed in this paper can improve the effect of training on uneven samples.

By observing the confusion matrix of the ResNet 101 network model, it can be seen that the oversampling method used in this paper is effective for processing data sets, as shown in Figure 2. On the validation set, VASC diseases with a small sample size are all correctly classified. DF class with the smallest sample size has 3 errors. Observe the model confusion matrix of ResNet 101 + FL after modifying the loss function. As shown in Figure 3, it can be seen that there are only 2 classification errors in the project with the smallest sample size, and the accuracy rate is 9% higher than the former. By comparing the two confusion matrices, it can be found that the network model after modifying the loss function has improved the classification performance of fewer samples in multi-classification tasks. This shows that the Focal Loss function is more effective than the cross-entropy loss function in multi-classification tasks with uneven sample sizes.

![Figure 2. ResNet 101 confusion matrix.](image1)

![Figure 3. ResNet 101 + FL confusion matrix.](image2)

The 0-1 in the confusion matrix represents seven different diseases, namely, 'MEL', 'NV', 'BCC', 'AKIEC', 'BKL', 'DF', and 'VASC', totaling 998 validation set samples.
4. Concluding remarks
This article mainly studies the application of deep learning in the field of assistant diagnosis of dermatology. Deep neural networks with different depths trained. And by comparing the results of the validation set, finding a depth with better performance, and achieving multi-class recognition of skin diseases based on deep learning. The deep learning model proposed in this article improves accuracy rate, recall rate, and F1 score. Compared with traditional machine learning, the methods of deep learning and transfer learning avoid the labor and time costs of feature extraction by experts. Through the oversampling data enhancement method and the operation of modifying the loss function, the performance of the model on multi-classification tasks on an imbalanced sample dataset is improved. Experiments have proved that the method proposed in this paper has better performance than conventional methods, and has certain auxiliary effects for clinical diagnosis of skin diseases.

Acknowledgments
This work is supported by General Project of Science and Technology Plan of Beijing Municipal Education Commission (Grant No. KM202011232005).

References
[1] Miller D D, Brown E W. Artificial intelligence in medical practice: the question to the answer?[J]. Am J Med, 2018, 131:129-133.
[2] Smith, L N, Smith, M L, Farooq, A R, Sun, J, Ding, Y, Warr, R. Machine vision 3D skin texture analysis for detection of melanoma[J]. Sensor Review, 2011, 31(2).
[3] Mahmoud, Mohamed Khalad Abu, Al-Jumaily, Adel, Maali, Yashar, Anam, Khairul. Classification of Malignant Melanoma and Benign Nevi from Skin Lesions Based on Support Vector Machine[P]. 2013.
[4] Esteva A, Kuprel B, Novoa R A et al. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. Nature, 2017, 542(7639):115-118.
[5] Lu Wang. Machine Learning Based Skin Disease Image Classification[D]. Zhejiang University, 2018.
[6] Guan Qiu, Li Jiang, Hu Haigen, Gong Mingjie, Chen Feng. Melanoma skin lesion recognition based on deep residual network[J]. Journal of Zhejiang University of Technology, 2019, 47(04):430-435.
[7] Magalhaes Carolina, Mendes Joaquim, Vardasca Ricardo. The role of AI classifiers in skin cancer images.[J]. Skin research and technology: official journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI), 2019, 25(5).
[8] Fu Yu, Bao Ying-qi, Yie Yi-bo. Application Research of Artificial Intelligence in the Skin Diseases Assisted Diagnosis[J]. China Digital Medicine, 2018, 13(10):29-31+38.
[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016, pp.770-778.
[10] He K, Zhang X, Ren S et al. Deep residual learning for image recognition[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016:770-778.
[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollar. Focal Loss for Dense Object Detection. The IEEE International Conference on Computer Vision (ICCV), 2017, pp.2980-2988.
[12] Tschandl P., Rosendahl C., Kittler H. The HAM10000 dataset. A large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, sdata.2018.161.
[13] Noel Codella, Veronica Rotemberg, Philipp Tschandl, Emre Celebi, Stephen Dusza et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC), 2018. https://arxiv.org/abs/1902.03368.