Deep Learning for Pigmented Skin Lesions Recognition

Abstract

Research has showed, skin cancer accounts for 1/3 of all diagnosed cancers worldwide. Therefore, scientists and Doctors have been researched on pigmented skin lesion for many years. Still, the average experts' diagnosis has only about 60% accuracy. To improve diagnosis accuracy, we plan to implement advanced neural network model on diagnosis of pigmented skin lesions using 12-layer CNN, Resnet150v2, Inceptionv3, Vgg19. By comparison, Inceptionv3 model has best performance, with 89% accuracy on validation, 86% accuracy on test dataset. Overall, transfer learning models are always show better performance than training a model directly (CNN).

Key words: CNN, Resnet150v2, Resnet150v2, Vgg19, Transfer Learning

Introduction

Skin cancer accounts for 1/3 of all diagnosed cancers worldwide. The dermatologists have an average accuracy of 62% to 80% in skin cancer diagnosis. To improve accuracy of diagnosing skin cancer, we use "The HAM10000 dataset" [5] to train our deep learning models and make predictions to evaluate them. The HAM10000 dataset consists of 10015 images about seven types of skin lesions and 193 images for test dataset. We split 10015 images into training and validation datasets and use 193 images for test. The seven skin lesions are as follows: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (**akiec**), basal cell carcinoma (**bcc**), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, **bkl**), dermatofibroma (**df**), melanoma (**mel**), melanocytic nevi (**nv**) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, **vasc**).
We first preprocess images to deal with imbalanced categories, and then normalize it. We mainly use 4 types of models that are CNN, Resnet150v2, Inception v3, Vgg19.

Background

Thanks to the development of science and technology, people can apply the deep learning model to leverage diagnosis. In the past decade, neural networks have developed rapidly, and many creative and breakthrough new inventions have been discovered. Top segmentation algorithms

still have problems in classifying over 10% of images on average, and algorithms with equal performance on test data can have different abilities to generalize. However, machine-learning classifiers still outperform human experts in the diagnosis of pigmented skin lesions. By comparison, in 2020 Chaturvedi, S.S. used ResNeXt101 model gains maximum accuracy of 93.20% [3]. In 2022, A. K. Sharma et al., used Cascaded ensembled deep learning model gains 98.3% accuracy [4].

Approach

Deep Learning techniques are widely used in the medical image field. We are going to build an automated diagnostic system using ANN, support human expertise in diagnosis of benign and malignant pigmented skin lesions. We expect our model to successfully diagnose 7 kinds of pigmented skin lesions with high accuracy. Initially, we will build Convolutional Neural networks, build deep models to gain high accuracy. As far as we learned, it is possible that if there are too many layers, the model's performance will decrease and even over-fitting. So, we will also use ResNet150v2, Inception v3, Vgg19.

VGG

The improvement of vgg16 over Alexnet is to replace the larger convolution kernel (11x11, 7x7, 5x5) in Alexnet with several consecutive 3x3 convolution kernels [6]. For a given receptive field (the local size of the input picture related to the output), using a stacked small convolution kernel is better than using a large convolution kernel, because the multi-layer nonlinear layer can increase the network depth to ensure the learning of more complex patterns, and the cost is relatively small (fewer parameters).

ResNet

ResNet solve the degradation issue. ResNet provides two ways to solve the degradation problem: identity mapping and residual mapping. Identity mapping refers to the "curved line" part in the figure, and residual mapping refers to the rest of the non-curved line. $F(x)$ is the network map before summation and $H(x)$ is the network map input after summation [7].

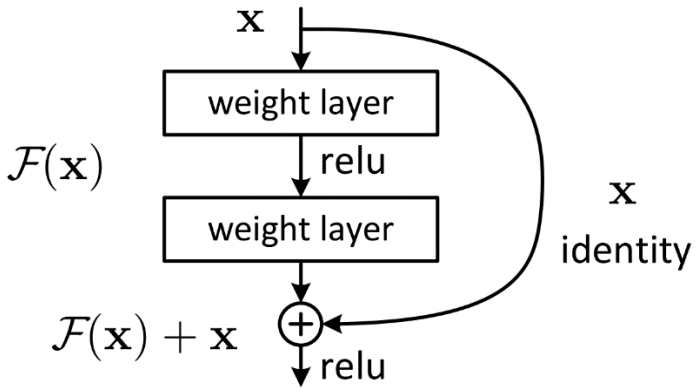


Figure 1 ResNet150V2 Structure

Inception

The InceptionV3 model is the third-generation model in Google's Inception series. Its model structure is placed in the same paper as the InceptionV2 model. In fact, the difference between the two models is not large. Compared with other neural network models, the biggest feature of the Inception network is to expand the convolution operation between neural network layers. The Inception network uses convolution kernels of different sizes, so that there are receptive fields of different sizes, and finally realizes the fusion of different scale features.

Preprocess

We make stratified train-validation-split using the official training data with ratio of 0.8. As seen from **Figure 1**, Ham10000 is a highly imbalanced dataset which affect model accuracy greatly. To balanced it, we use image generator function from TensorFlow to conduct training image oversample by zooming, shifting, rotating, flipping images. We keep original images composition for validation data and test data.

To deal with overfitting problem, we use early stopping function to monitor how validation loss changes. If there is no validation loss lower than the current lowest validation with difference larger than 0.001 for 50 epochs, the model will stop.

CNN, ResNet150v2, Inceptionv3, Vgg19 are used to train models. Weights are initialized using ImageNet for ResNet150, Inceptionv3, Vgg19. Last 3 outputs are transformed into 7 labels.

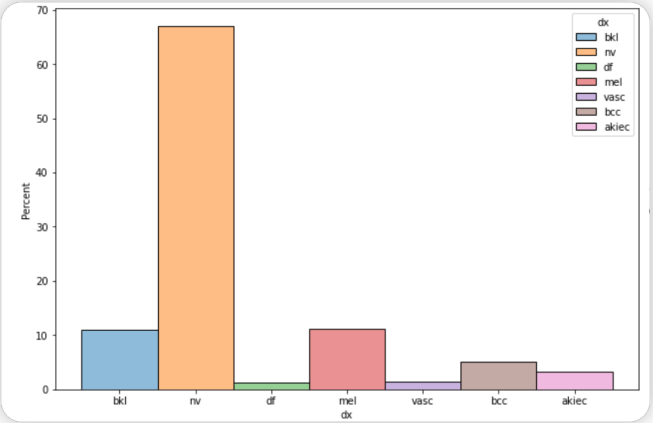


Figure 2 Classes Distribution

Results

We used the preprocessed dataset as described in the approach part.

In **Figure 2**, the average accuracy for CNN, ResNet, Inception, Vgg are as follows: 0.75, 0.81,0.86,0.79. ResNet150v2 works best for the test dataset, while CNN is the worst. However, in **Figure 3**, average AUC for all models is ranked from high to low: InceptionV3, Vgg19, ResNet150v2, CNN. InceptionV3, CNN keeps the same order, while ResNet150v2 becomes better than Vgg19. Considering our test dataset do not conduct image oversample, this may be due to ResNet150v2 predicts majority classes better in test dataset.

We use four models to train seven categories. InceptionV3 is the best model for ‘akiec’, InceptionV3 for ‘bcc’, InceptionV3 for ‘bkl’, InceptionV3 for ‘df’, ResNet150V2 for ‘mel’, InceptionV3 for ‘nv’, all model seems perform well for ‘vasc’. The rank is measured through the overall rate of precision, recall and F1-score. Since we only have one image with label of ‘df’ in test dataset, and only InceptionV3 predict it right.

As for parameters of all four models, Vgg19 has the greatest number of parameters, then comes the ResNet150v2, InceptionV3, CNN. Considering prediction evaluations concluded before, InceptionV3 uses few parameters to generate prediction with highest accuracy.

The type ‘vasc’ only has 3 images in the test dataset, but nearly all four models predict this type right. This may be due to ‘vasc’ is the most recognizable type in skin lesions. Prediction evaluations for ‘nv’ is also high. Considering ‘nv’ has the majority images in original train dataset and test dataset, it may be due to that image augmentation could make the number of all categories equal but still have lower effect for training compared to the real image data.

Our models predict well in types of ‘nv’, ‘vasc’, while poor in types of ‘bkl’ and ‘akiec’. To improve it, we should use more advanced models to train it. We could also train a model to recognize the outline of the main area which could be the attention of images.

	precision	recall	f1-score	support
akiec	0.60	0.38	0.46	8
bcc	0.92	0.73	0.81	15
bkl	0.47	0.32	0.38	22
df	0.00	0.00	0.00	1
mel	0.50	0.24	0.32	21
nv	0.80	0.93	0.86	123
vasc	1.00	1.00	1.00	3
accuracy			0.75	193
macro avg	0.61	0.51	0.55	193
weighted avg	0.73	0.75	0.73	193

Figure 2.1 CNN

	precision	recall	f1-score	support
akiec	0.71	0.62	0.67	8
bcc	1.00	0.80	0.89	15
bkl	0.73	0.73	0.73	22
df	1.00	1.00	1.00	1
mel	0.68	0.62	0.65	21
nv	0.90	0.94	0.92	123
vasc	1.00	1.00	1.00	3
accuracy			0.86	193
macro avg	0.86	0.82	0.84	193
weighted avg	0.86	0.86	0.86	193

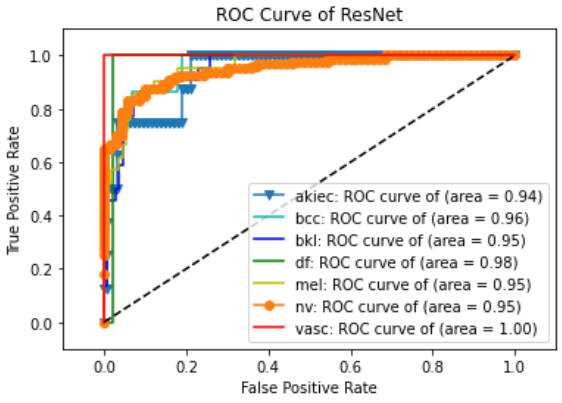
Figure 2.2 Inception

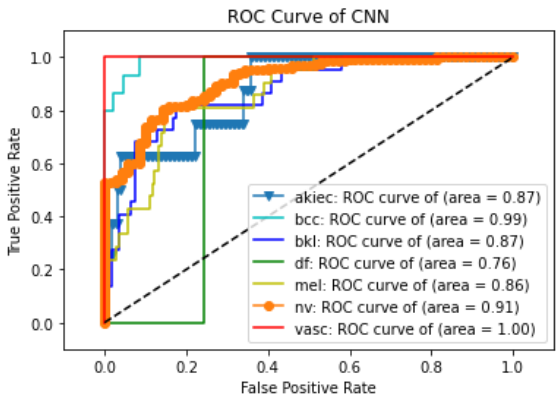
	precision	recall	f1-score	support
akiec	0.67	0.50	0.57	8
bcc	0.60	0.80	0.69	15
bkl	0.65	0.68	0.67	22
df	0.00	0.00	0.00	1
mel	0.71	0.57	0.63	21
nv	0.90	0.92	0.91	123
vasc	1.00	0.33	0.50	3
accuracy			0.81	193
macro avg	0.65	0.54	0.57	193
weighted avg	0.81	0.81	0.81	193

Figure 2.3 ResNet150V2

	precision	recall	f1-score	support
akiec	0.38	0.38	0.38	8
bcc	0.62	0.67	0.65	15
bkl	0.57	0.36	0.44	22
df	0.00	0.00	0.00	1
mel	0.62	0.76	0.68	21
nv	0.90	0.92	0.91	123
vasc	1.00	1.00	1.00	3
accuracy			0.79	193
macro avg	0.58	0.58	0.58	193
weighted avg	0.78	0.79	0.78	193

Figure 2.4 VGG19





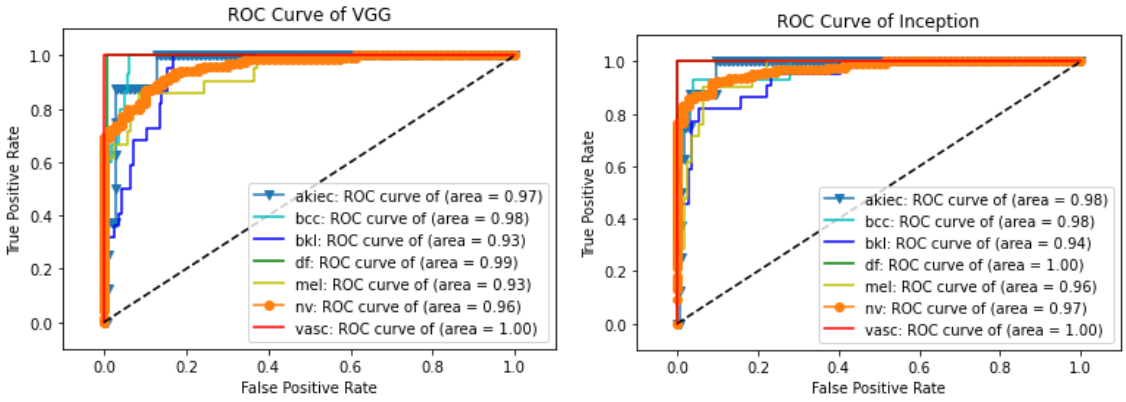


Figure 3 ROC Curve

Conclusion

We trained models that could automate diagnosis of skin lesions. Inception V3 is the best model for our dataset. Inception V3 is also the most efficient model using relative low parameters. Predictions for 'nv', 'vasc' is relatively high in our model while 'bkl', 'akiec' is relatively low. To improve, we could train a model for distinguishing the attention of images to improve performance.

The reason why InceptionV3 is the best model can be from 3 aspects:

Factorized Convolutions:

In InceptionV3, it replaces bigger convolutions with smaller convolutions leads to faster training. Say a 5 × 5 filter has 25 parameters; two 3 × 3 filters replacing a 5 × 5 convolution has only 18 (3*3 + 3*3) parameters instead. This implementation reduces computation cost as it reduces the number of parameters in network. Therefore, it can include more layers in model.

Asymmetric convolutions：

A 3 × 3 convolution could be replaced by a 1 × 3 convolution followed by a 3 × 1 convolution. If a 3 × 3 convolution is replaced by a 2 × 2 convolution, the number of parameters would be slightly higher than the asymmetric convolution proposed.

Auxiliary classifier：

The auxiliary classifier is a small CNN inserted between layers during training, and the loss incurred is added to the main network loss. In GoogLeNet auxiliary classifiers were used for a deeper network, whereas in Inception v3 an auxiliary classifier acts as a regularizer.

References

[1]N Codella, V Rotemberg, P Tschandl, ME Celebi, S Dusza, D Gutman, B Helba, A Kalloo, K Liopyris, M Marchetti, H Kittler, A Halpern. "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)."; 2019

[2]Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, Gutman D, Halpern A, Helba B, Hofmann-Wellenhof R, Lallas A, Lapins J, Longo C, Malvehy J, Marchetti MA, Marghoob A, Menzies S, Oakley A, Paoli J, Puig S, Rinner C, Rosendahl C, Scope A, Sinz C, Soyer HP, Thomas L, Zalaudek I, Kittler H. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol. 2019

[3] Chaturvedi, S.S., Tembhurne, J.V. & Diwan, T. A multi-class skin Cancer classification using deep convolutional neural networks. *Multimed Tools Appl* **79,** 28477–28498 (2020). https://doi.org/10.1007/s11042-020-09388-2

[4] A. K. Sharma et al., "Dermatologist-Level Classification of Skin Cancer Using Cascaded Ensembling of Convolutional Neural Network and Handcrafted Features Based Deep Neural Network," in IEEE Access, vol. 10, pp. 17920-17932, 2022, doi: 10.1109/ACCESS.2022.3149824.

[5] Tschandl, Philipp, 2018, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", https://doi.org/10.7910/DVN/DBW86T, Harvard Dataverse, V3, UNF:6:/APKSsDGVDhwPBWzsStU5A== [fileUNF]

[6] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[7] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." *Thirty-first AAAI conference on artificial intelligence*. 2017.