Final Report: Skin Cancer Recognition via Computer Vision/ Healthcare

Yinuo Yao (yaoyinuo@stanford.edu)
Chen Chen (chenc2@stanford.edu)
Tao Jia (taoj@stanford.edu)

Github: https://github.com/yaoyn/CS230Proj.git

1 Introduction

In this project, we plan to use computer vision techniques to recognize different types of skin lesions and provide a supporting tool for the diagnosis of skin cancer. In current medical diagnosis, identifying skin cancer has always been challenging because of its close assemblance to other types of skin diseases. Based on the test sets[1], two dermatologists can only achieve an accuracy of about 66% in identifying skin cancer. As a result, this false detection outcome is detrimental. In this project, by applying deep learning techniques, we aim to assist doctor in identifying and classifying different types of skin cancer based on images of skin diseases on different locations of the patient's body.

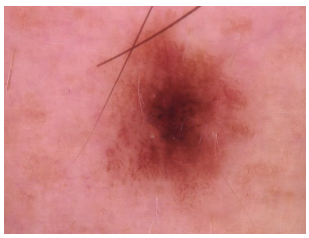
1.1 Research Background

A detailed background of the dataset has been published by Philipp, Cliff and Harald[2]. It has been proven that using deep neural network to identify the skin images[1][3] and dermatoscopy[4] has a high accuracy. An early study by Binder et al applied neural network to identify pigmented skin lesions[5]. Esteva et al presented an influential work on computer vision diagnosis of skin cancers.[6]

2 Data Description

The dataset we use is HAM10000 (“Human Against Machine with 10000 training images”) that consists of 10015 dermatoscopic images, seven diagnostic categories and a metadata file with information of sex, age, and the location of the skin feature. All photos have sizes of 600 * 450 pixels. The data can be found online at https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000.

A typical data entry with images and metadata is:



lesion_id	image_id	dx	dx_type	age	sex	localization
HAM_0000550	ISIC_0024306	nv	follow_up	45.0	male	trunk

Figure 1, a typical data entry

We can read from the example that the diagnostic category “dx” is “nv” (standing for melanocytic nevi, a type of benign neoplasms), the method for diagnosis “dx_type” is “follow_up”.

A summary of data is shown below:

Diagnosis category	Full name of disease	Description	Counts in dataset	Counts in dataset after augmentation
nv	Melanocytic nevi	Benign	6705, 66.95%	6705, 34.1%
mel	Melanoma	Malignant cancer, invasive or non-invasive	1113, 11.11%	2226, 11.3%
bkl	Benign keratosis-like lesions	Benign	1099, 10.97%	2198, 11.2%
bcc	Basal cell carcinoma	Malignant cancer, non-invasive	514, 5.13%	2056, 10.5%
akiec	Actinic keratoses	Benign, may turn to malignant cancer	327, 3.27%	2289, 11.6%
vasc	Vascular lesions	Mostly Benign	142, 1.42%	2130, 10.8%
df	'Dermatofibroma'	Benign skin lesion	115, 1.15%	2070, 10.5%

Table 1. Summary of data

3 Network Architect

3.1 Mission

Although there are 7 types of diagnostic categories, only two (“mel” and “bcc”) are malignant, and they only take ~16% of all data. The distribution is highly uneven, and the most important focus is to identify malignant cancers (bcc, mel) and potentially malignant features (akiec) correctly. Therefore, instead of accuracy, we take the recall of cancers (probability for cancer cases to be correctly predicted as cancer) as our pivotal metric.

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

3.2 Tools

We use keras 2.3.1 with tensorflow 2.0.0 backend. This way we can focus on the high-level architecture of our neural network instead of fighting against the details of backpropagating, and can adapt simple neural networks and residual networks from our homework. We set up an instance of p2.xlarge in AWS EC2 with 1 GPU as an economic computation method.

3.3 Baseline

We adopted and modified the framework of the code in reference [7] as our baseline model.

- Data are divided into train (90%), validation (5%) and test set (5%). Data is first normalized by division of 255, as normalization by mean and standard deviation will cause major distortion due to our very unbalanced dataset.
- Baseline model structure: [CONV2D -> RELU] x 2 -> MAXPOOL -> Dropout -> [CONV2D -> RELU] x 2 -> MAXPOOL -> Dropout -> FLATTEN -> FULLY-CONNECTED -> DROPOUT -> FULLY-CONNECTED

After 50 epochs, the baseline model achieves accuracy of 79%, 76% and 78% for training, validation and test set, respectively. The recall for cancer (probability for cancer cases to be correctly predicted as cancer) is 0.44, which is not satisfactory. Besides, the accuracies show little overfitting, so it may be challenging to significantly improve from this simple CNN model. Therefore, we switch gear to the more complex and flexible ResNet.

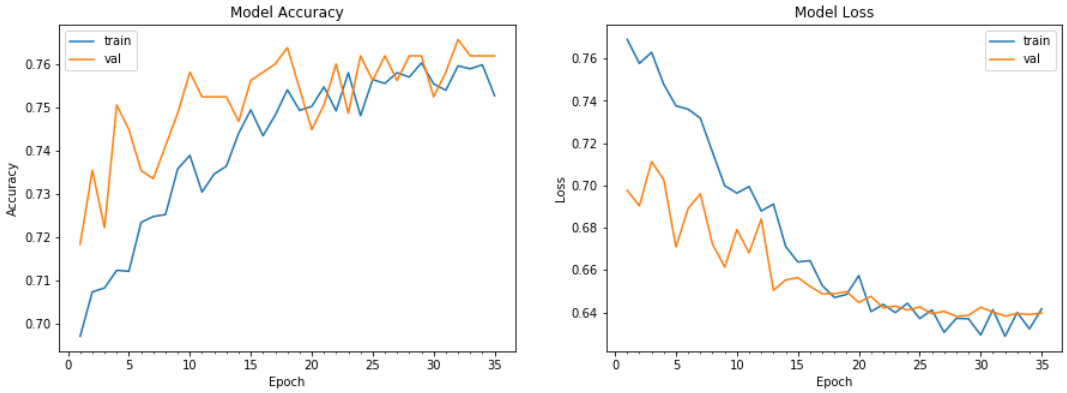


Figure 2. the training of baseline model (from 16th epochs to 50 epochs)

3.4 Our Approach: Residual Networks

We applied ResNet 50 with initial weights imported from ImageNet, and added GlobalAveragePooling, Dropout and Softmax for classification. After constructing a vanilla model and conducting initial data analysis, we concluded that the main challenge with our dataset is due to an imbalance where entire dataset only consists of approximately 20% as cancerous disease (bcc, mel and akiec). This would make the training of model to identify cancerous disease more difficult. Therefore, aiming to resolve dataset imbalance, we have decided to focus our tuning on the **weight of each category on the loss function**. The weight is calculated as the following,

$$W_i = \lambda \frac{N_i}{\sum_{i=1}^{C} N_i},$$

where W_i is the weight of category i, N_i is the number of samples in category i, λ is the multiplier and C is the number of categories. By applying λ, we could penalize the respective category more than its sample weight. In addition, we also tuned **dropout rate** and applied **data augmentation** to prevent overfitting. Table 2 summarizes the parameters of various model we tried. The naming follows “Model-X-Y” where X is the number of the model and Y is the dropout rate.

Model #	**Weights**	**Multiplier (λ) for cancerous**			**Dropout rate**	**Data Aug**
		akiec	bcc	mel		
Baseline	Sample	No multiplier			0.3	No
Model 1-3	Sample	No multiplier				No
Model 2-3	Computed	1	1	1	0.3	No
Model 3-3	Computed	1.5	1.5	1.5	0.3	No
Model 4-3	Computed	2	2	2	0.3	No
Model 5-3	Computed	3	3	3	0.3	No
Model 6-3	Computed	4.5	1.5	1.5	0.3	No
Model 7-3	Computed	6	2	2	0.3	No
Model 8-3	Computed	9	3	3	0.3	No
Model 8-1	Computed	9	3	3	0.1	No
Model 8-2	Computed	9	3	3	0.2	No

Model 8-4	Computed	9	3	3	0.4	No
Model 8-5	Computed	9	3	3	0.5	No
Model 9-3	Computed	9	3	3	0.3	Yes

Table 2. Summary of hyperparameters in different model explored

4 Presentation & Analysis of Results

4.1 Loss function weight

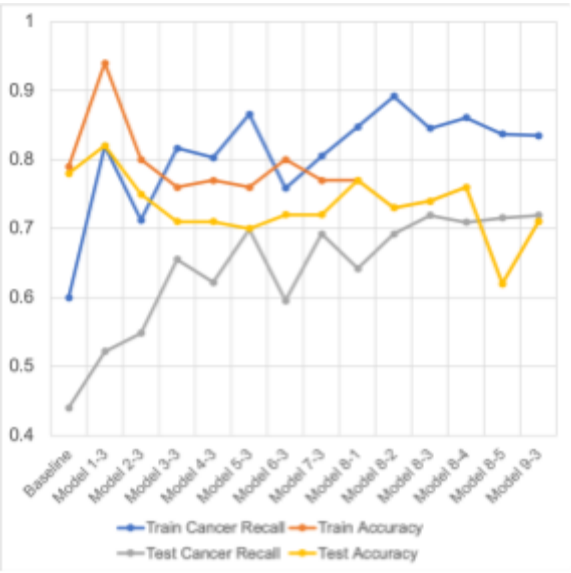


Figure 3. Cancer recall and accuracy for train and test

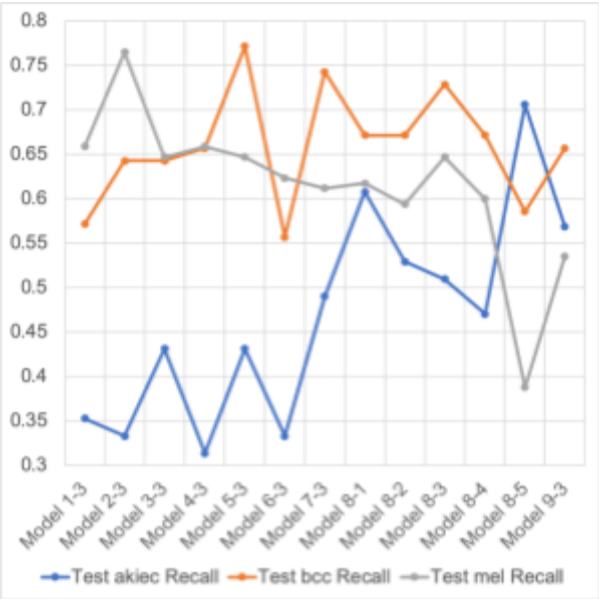


Figure 4. Test recall rate for akiec, bcc and mel

In order to improve the recall rate for cancers, our key performance indicator, we tried different ways to penalize errors in predicting cancer. Model 1 used the same weight on the loss function. Model 2 computed weights based on the fraction of each category in the dataset. In Models 3-5, a constant multiplier is added to penalize cancerous disease more than non-cancerous disease. In Models 6-8, we tripled the constant multiplier on disease “akiec” because we found out the recall rate for “akiec” was unsatisfactory.

From Model 1 to Model 8, all of them achieved test accuracy greater than 70% which proved the capability of these models. Since test accuracy was not the main concern in our model, we shifted the focus to the recall rate of cancerous disease. As shown in Figure 3, the test accuracy decreased by approximately 10% while the recall rate for cancerous diseases increased by 20%. The decrease in the accuracy was contributed by the misclassification of non-cancerous disease “nv” (Figure 5). In addition, as shown in Figure 4, by using a larger multiplier on “akiec”, the recall rate for “akiec” increased monotonically (Model 6-3, Model 7–3 and Model 8-3). At the same time, we did observe the recall rate for “mel” and “bcc” (the other two types of cancerous diseases) following a concave trend with increasing followed by decreasing (Model 6-3, Model 7–3 and Model 8-3). The concave trend in the recall rate suggested that there is an optimal weight for trade off among each cancerous disease. Based on our results, we found that model 8-3 performed the best.

With a further analysis, we noticed that there was an approximately 10% difference between the recall rate on training data and test data which suggested potential overfitting of the data. Therefore, in the next step, we tried to tune our dropout rate (keep probability) to regularize the model.

4.2 Dropout rate

In this part, we used Model 8-3 as the baseline and tuned its dropout rate from 0.1 to 0.5 (Model 8-1 to Model 8-5). The accuracy difference between train and test decreased from 7% to 3% suggesting reduced overfitting. In addition, as we increased the dropout rate from 0.1 to 0.3, the difference between train and test recall rate for cancerous diseases decreased from 20% to 10%. No improvement was observed as we increased the dropout rate from 0.3 to 0.5. By examining the individual recall rate for each cancerous disease, we observed test recall rate for “akiec” increased monotonically while both the test recall rate for “bcc” and “mel” increased then decreased (Model 8-1 to Model 8-5). Again, by finding an optimal among the tradeoff, we concluded that dropout rate of 0.3 performed the best.

4.3 Data augmentation

To further reduce overfitting, we adopted data augmentation by adding more samples into cancerous diseases through random flipping and cropping. Table 1 summarizes the new images distribution. In this model, we used the same architecture and parameters as model 8-3 except for augmenting the data.

Results have shown that the accuracy difference between train and test decreased to 1%. However, recall rate difference for cancerous diseases did not improve.

5 Project Insights and Discussion

In the previous analysis, we focused more on recall rate for all cancerous and potentially cancerous diseases. And our models is capable of improving the recall rate to 70%. This demonstrated the capability of our model. To analyze the results further, we constructed the confusion matrix of our best results produced by Model 8-3. By examining the confusion matrix, we found out that in certain circumstances, false classification might not be detrimental. There are actually two levels of classification: 1. Cancerous or non-cancerous 2. Disease type. For example, since “akiec” is cancerous, classifying “akiec” as “bcc” and “mel” will not be as detrimental as classifying it as the other non-cancerous diseases. This is because doctor will still treat it as cancer. However, we admit that this does have an impact on the treatment procedures. On the other hand, classifying cancer as non-cancer will be detrimental. Therefore, this observation provides a path for future improvement on the model by constructing two separate neural networks in classification. The first layer will focus on identifying whether it is cancerous or non-cancerous. The second layer will then focus on identifying the type of the disease given it is cancerous or non-cancerous. In our opinion, this framework will probably improve the overall accuracy and recall rate.

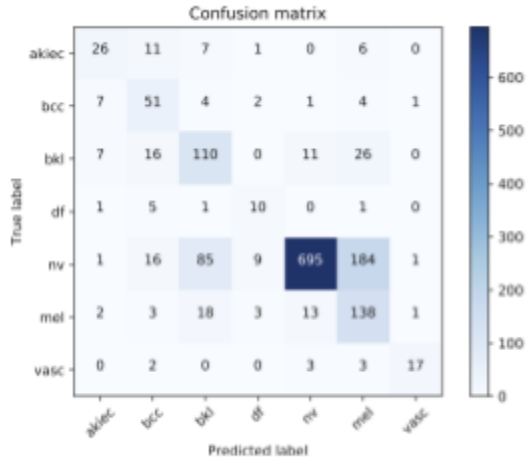


Figure 5. Test confusion matrix of our best model (Model 8-3)

References

1. Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542:115–18.

2. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset: a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018; 5:180161.
3. Han SS, Kim MS, Lim W et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermat 2018; 138:1529–38.
4. Haenssle HA, Fink C, Schneiderbauer R et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018; 29:1836–42.
5. Binder, M., Steiner, A., Schwarz, M., Knollmayer, S., Wolff, K., & Pehamberger, H. Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study. British Journal of Dermatology 1994; 130(4): 460–465. doi: 10.1111/j.1365-2133.1994.tb03378.x
6. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639): 115–118. doi: 10.1038/nature21056
7. Siddhartha, M. Step wise Approach : CNN Model, https://www.kaggle.com/sid321axn/step-wise-approach-cnn-model-77-0344-accuracy