Team 37: Skin Lesion Classification

Zhiyin Liu
Electrical and Computer Engineering
University of California San Diego
San Diego, CA 92161
zhl114@ucsd.edu

Xingtong Yang
Electrical and Computer Engineering
University of California San Diego
San Diego, CA 92161
xiy035@ucsd.edu

Abstract

Skin lesion is a terrifying disease for human beings. The sooner the type of skin lesion can be diagnosed, the quicker the appropriate treatment can be applied. So, machine learning is a powerful tool to help detect and classify the type of skin lesion. In this report, we would describe how we used VGG16, and ResNet50 from transfer learning and scratch to classify the seven types of skin lesion images from the HAM10000 dataset and also compare the results from four models. After fine-tuning the hyperparameters, we got that the validation accuracy for VGG16 transfer learning is 74%; for ResNet50 transfer learning is 81%; for VGG16 from scratch is 96%; for ResNet50 from scratch is also 96%.

1 Introduction

There are many types of the skin lesion. Some of the them are not fatal and can be cured with appropriate treatment. However, some of them are actually cancer and they are fatal. Therefore, how to correctly and quickly find out the type of the skin lesion is very important. From this inspiration, we are interested in exploring the use of machine learning algorithms to aid medical professionals in the diagnosis skin lesions. To help explore the usage of the machine learning models, we used HAM10000 dataset from Harvard which it contains seven types of the skin lesions [1]. We used VGG16 and ResNet50 with transfer learning and from scratch to classify the lesions. The main contributions for this project are:

- The paper that we referenced with doesn't use transfer learning but we think it may be useful and tried it.
- We can classify the lesions with an maximum validation accuracy: 96%.
- The best model for this project we found is: VGG16 from scratch.

2 Related Works

The paper that we are reading and referencing is: Soft-Attention Improves Skin Cancer Classification Performance[2]. In this paper the authors presented how they used VGG, ResNet, Inception ResNet V2, DenseNet to classifiy those skin lesion images into the seven disease class with or without soft attention [2]. Due to time limit of this quarter, we looked into VGG16 and ResNet model applied on the skin lesion dataset. This paper really list out all key parts on how to evaluate performance that trained on medical image dataset such as pre-process dataset, details of the models usage and different metrics evaluation.

NeurIPS 2021 AI for Science Workshop.

3 Problem Formulation

As we discussed in the previous section, we would like to use HAM10000 dataset[1]. This dataset contains seven skin lesions which are: 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 nevus (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc).[1], where melanocytic nevus (nv) is the skin cancer and we would focus on classifying this more than others. Since the original dataset doesn't separate the images by the type of lesion, the first problem we need to solve is to process the images into categories and establish the training and validation set. The second problem that we needs to solve is to classify the seven skin lesion and compare the results between the models we used. The models that we used in this project are the transfer learning for VGG16 and ResNet50 and then the VGG16 and ResNet50 from scratch.

4 Dataset

HAM10000 dataset contains approximately 10015 color images (224x224x3) from 7 different skin Lesions (see figures below).

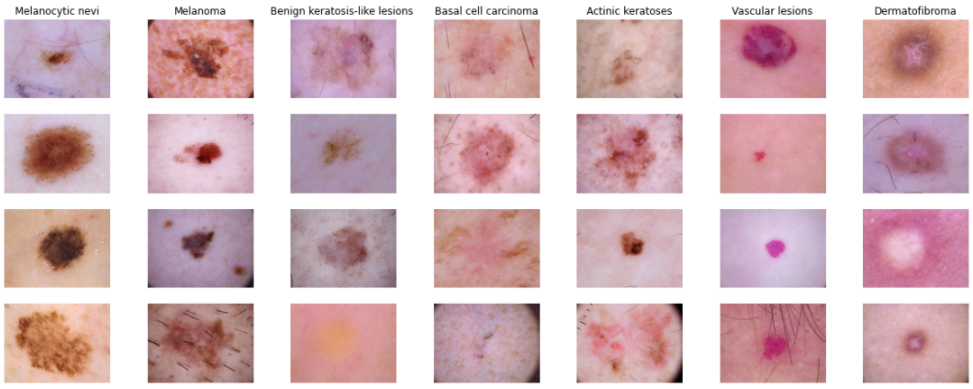

Figure 1: Skin Lesions

The original dataset contains three files: HAM10000 part 1, HAM1000 part 2, and a metadata file. The raw data contains all skin lesion images with a specific ID. We first filter out repeated IDs by extracting the IDs from metadata.Then based on the lesion type of each image, we separate all images into 7 classes.
The number of images per class is as follows: 327 akiek (pre-cancerous), 514 for bcc (benign), 1099 for bkl(benign), 115 for df (benign), 1113 for mel(cancerous), 6705 for nv (benign), 142 for vasc (benign). The images come from different body parts, but by simple inspection is not clear the body part it comes from. The earliest patient had an age of around 8 years and the oldest around an age of 85 with most patients being between 35 and 50 years old. 54 percent being identified as male and 45 percent as female. We are not however explicitly using this information on our experiments at the moment. One more important thing to note is that the dataset does not contain images from people of color or healthy skin images; however, for our initial experiments is not a problem. To conduct the initial experiments we randomly divide 85% of the original dataset into training data and rest to be testing data regardless of ages, gender, and skin color. Then we save them to google drive in a form that TensorFlow can directly train models on.
The training set has the following distribution 3% akiec, 5% bcc, 11% bkl, 1% df, 11% mel, 66% nv, and 2% vasc. The testing set has the distribution 2% akiec, 3% bcc, 8% bkl, 1% df, 4% mel, 80% nv, and 1% vasc.

5 Technical Approach

As we discussed in the dataset section, the first step of the project is to pre-process the images in the dataset. After separating the skin lesion images into seven categories and established the training and validation dataset, we started to pick the model and compare the performance of the models. Since the base objective of this project is to do image classification, we choose the convolution neural network (CNN), especially VGG16 and ResNet50.

5.1 Transfer Learning

We tried the transfer learning for VGG16 and ResNet50 first. Since transfer learning used online pre-trained weight for the convolution layers, so it should help us save time in training. As we know that the convolution layers help to find the useful features of the images to do the classification in fully connected linear layers, so we choose the weight in ImageNet dataset to be our weight because ImageNet contains a huge number of the images[3]. For the architecture of the VGG16 and ResNet50 we used build-in models from tenserflow keras. To be able to update the weight better, we used Adam optimizer with learning rate 0.001. The activation function that we used in hidden layers of the fully connected network is ReLU:

$$a = max(0, input)$$

The activation function that we used for the output layer is softmax:

$$\sigma(z) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Since we have seven skin lesion, so the loss function we used is categorical cross-entropy function[5]:

$$Loss = -\sum_{i=1}^{outputsize} y_i * log(\hat{y}_i)$$

To keep the consistency, VGG16 and ResNet50 use the same activiation function and the loss function. We ran those models several times and we tried the learning rate 0.01 first, but turns out learning rate 0.01 overfit our modle so we changed to the 0.001 in the end. Unfortunately, we didn't get a pretty good result by using the transfer learning. So we decided to build VGG16 and ResNet50 from scratch to try to improve the results.

5.2 VGG16 and ResNet50 from Scratch

5.2.1 VGG16

To build the models from scratch, we first tried the VGG16 because the architecture of the VGG16 is relatively simpler than ResNet50. The VGG16 architecture that we used for VGG16 is the figure below:

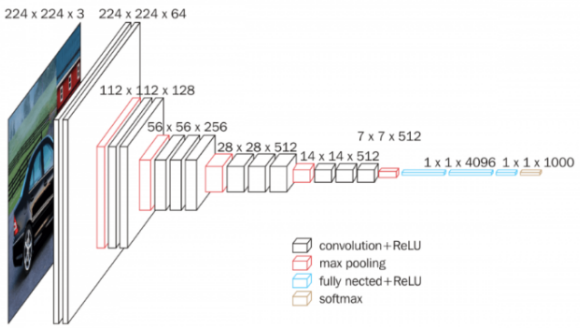

Figure 2: Architecture of VGG16 [4]

VGG16 is a powerful tool to classify RGB images because it combined the feature of convolution and fully connected network to first capture the useful feature of the images then to classify the images.

In VGG16 convolution layer block, there are two to three convolution layers followed by a max pooling layer to both extract the features and down size the input images. Additionally, we added a batch normalization after every max pooling layer to smooth the learning. During the process of training the model and compared the results after each run, we found out that the usable feature of the skin lesion is very limited. Also, the skin lesions in our dataset are similar to each other. So we adjusted the architecture of the VGG16 model a little bit by referencing to the paper:Soft Attention Improves Skin Cancer Classification Performance [2]. For each convolution blocks of the VGG16, the last convolution layer (the layer before max pooling) we changed the kernel size to be 1. The reason that we changed the filter size of the last convolution layer is that we want to maintain the benefit of using the convolution and max pooling layers but we also want to capture more detailed feature of the skin lesions, so we only changed the last convolution layer before max pooling of each convolution blocks. Other than the last convolution layer of each convolution blocks and added batch normalization layers, the other layers are the same following VGG16 architecture.
For the fully connected layer of the VGG16 architecture, the activation function for hidden layer we used ReLU function as we descripted in the transfer learning section. Similar to the activiation function for the output layer that we used softmax. The loss function for this model is also the same as the transfer learning: categorical cross-entropy. To help update the weights better in fully connected layer, we used Adam optimizer with learning rate of 0.001.

5.2.2 ResNet50

After building the VGG16, we also tried to build ResNet50 from scratch. The main architecture of the ResNet50 is the figure below.

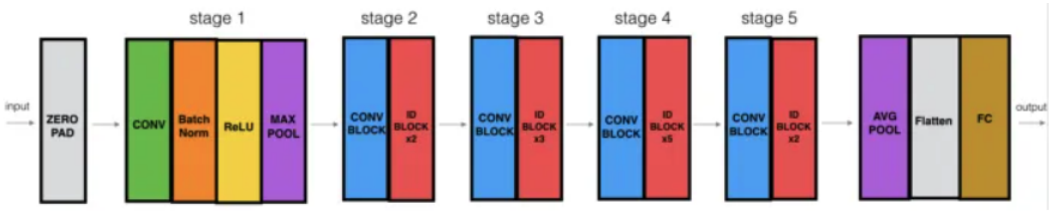

Figure 3: Architecture of ResNet50 [6]

ResNet50 is a deep learning network and it is also powerful tool for object classification. The benefit in using the ResNet50 is that the gradients could skip some layers in backward propagation and since it is very deep network so it can learn the features by many aspects [6]. However, ResNet50 is a complex architecture so we didn't modify any layer, we built our model exactly follow the ResNet50 architecture from scratch.
Similar to the VGG16, the activation function for hidden layers is ReLU. The activation function for the output layer is softmax and the loss function here is the categorical cross-entropy. We still used the Adam optimizer with a learning rate of 0.001 to update weights.

5.3 Metrics

We use "Accuracy" as our metric evaluation during the training. After training, we also used 'Precision', 'Recall', and 'F1-score' to evaluate the model performance. The formula on how to compute these metrics are listed below:

$$Accuracy = \frac{TruePositive(TP) + TrueNegative(TN)}{Total\ Number\ of\ TestSet}$$

$$Precision = \frac{TruePositive(TP)}{TruePositive(TP) + FalsePositive(FP)}$$

$$Recall = \frac{TruePositive(TP) + TrueNegative(TN)}{TruePositive(TP) + FalseNegative(FN)}$$

$$F1\_Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

True Positive means when we have a skin lesion image with a specific, say Type A, the model correct predict as Type A. False positive means that the model incorrectly predict it to be some other types. True Negative means when we have a skin lesion image that is not type A, the model correct predict it as not Type A. False negative means that the model incorrectly predict it as Type A.
According to the equations, Accuracy is basically the percentage of correct predictions on validation dataset while training.

6 Results

Below we have shown a metrics evaluation score tale of four models we have trained.

Model name/Metrics	Best Accuracy	Precision After Training (NV/Averaged other 6 classes)	Recall after Training (NV/Averaged other 6 classes)	F1_score after Training (NV / Averaged other 6 classes)
VGG16 Transfer Learning	0.74	NV: 0.92 Avg 6: 0.44	NV: 0.91 Avg 6: 0.36	NV: 0.91 Avg 6: 0.366
VGG16 From Scratch	0.96	NV: 1.00 Avg 6: 0.86	NV: 0.95 Avg 6: 0.98	NV: 0.98 Avg 6: 0.91
ResNet50 Transfer Learning	0.81	NV: 0.91 Avg 6: 0.15	NV: 0.93 Avg 6: 0.18	NV: 0.92 Avg 6: 0.16
ResNet50 From Scratch	0.96	NV: 1.00 Avg 6: 0.85	NV: 0.95 Avg 6: 0.98	NV: 0.98 Avg 6: 0.91

Figure 4: Models Metric Evaluation Score Summary

In this form, the "NV" means the cancer: melanocytic nevus. Avg 6 means the average result of other six skin lesions. As we mentioned in the beginning of the report, we would pay more attention on classifying the cancer, so we show the result of the cancer more detailed than others. To be able to pay more attention on this cancer, the weight when we trained the dataset for NV is five and one for the rest lesions. Also as we discussed before, the learning rate that we used in this project is 0.001. From the metrics evaluation form, even though the transfer learning for VGG16 and ResNet50 both has a good performance on detecting the cancer “NV’ with good precision, recall, and F1 scores, it has very pool result over the other 6 classes. None of the scores are above 0.5. Since the transfer learning is using the pre-trained weights, each convolution layer is not trainable, but only train on forward layers after flatten from last convolution layer. Since we have 80% dataset on NV, The forward layers tend out to be over-fitting on this skin lesion dataset. Comparing with VGG16 and ResNet50 that are both built from scratch and set each convolution layer to be trainable, the performance of each model is actually learning from the lesion images. Since we trained each model under the same number of epochs with early stop criteria based on the validation lost, VGG16 has a better precision score than ResNet50 does. Therefore, we decided that VGG16 is the best model to be considered when classifying 7 skin lesions.

Below we have accuracy and loss plot on both on training and validation.

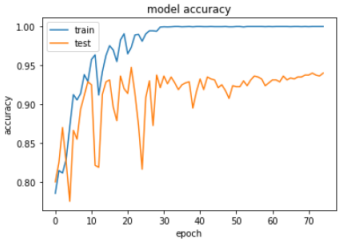

Figure 5: VGG16 from scratch Accuracy Plot

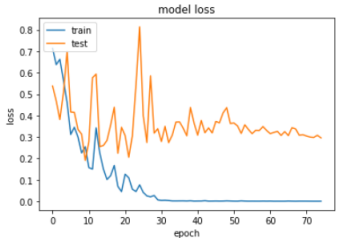

Figure 6: VGG16 from scratch Lost Plot

The baseline for VGG16 that set each convolution layer to be trainable is 47% on training data and 67% on testing dataset. By continuously updating weights in each layer with Adam optimizer, it reached highest 96% on testing dataset and almost 100% on training dataset. This tell us that our skin lesion dataset is representative to be trained on VGG16. The loss is decreasing , and accuracy is increasing on both test and train dataset during training.

Below we have accuracy and loss plot on both on training and validation for VGG16 Transfer Learning.

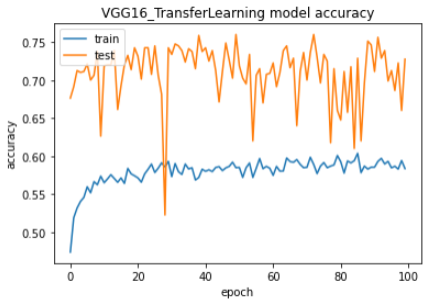

Figure 7: VGG16 Transfer Learning Accuracy Plot

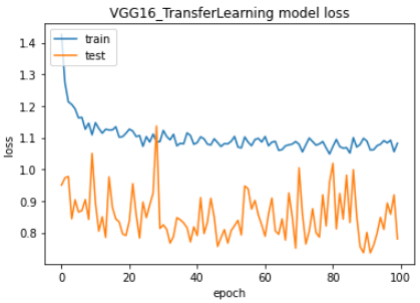

Figure 8: VGG16 Transfer Learning Lost Plot

We can tell from the figures above that the VGG16 transfer learning has a really large fluctuations accuracy and loss on the testing, therefore the model is actually over-fitting on this dataset.
Below are results from both ResNet50 build from scratch and transfer learning.

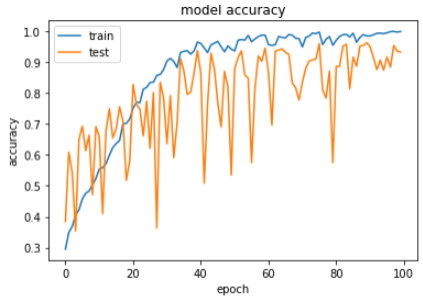

Figure 9: ResNet50 Accuracy Plot

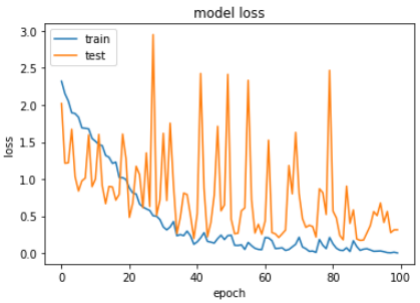

Figure 10: ResNet50 Lost Plot

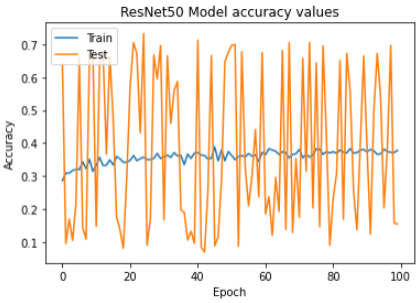

Figure 11: ResNet50 Transfer Learning Accuracy Plot

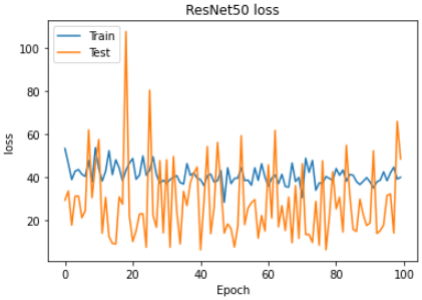

Figure 12: ResNet50 Transfer Learning Lost Plot

As we can tell from the above graph, the RestNet50 has similar situation on model with/out transfer learning that are trained under the same criteria. So based on all those results that we showed above

we can say that transfer learning is not good method for us to use in this project and VGG16 from scratch has the best performance.

7 Conclusion and Future Work

After training the four models by fine-turning all hyper-parameters and trained under same criteria. We can conclude that transfer learning is really not a good model for classifying the skin lesions. The reason that transfer learning didn't perform well may be because the online weight didn't work well with skin lesion images.
As we talked about previously that skin lesions are similar to each other, so we changed the convolution layer kernel size a little bit in VGG16. However, we didn't do the same thing for ResNet50 because it is a relatively large and complex network. Based on the metric evaluation score table, VGG16 has a generally better performance than ResNet50 does. As a result, we reached 96% validation accuracy for VGG16 and ResNet50. VGG16 is the best model so far that can be used to make a good prediction on 7 different skin lesions.
We mentioned that we would try ResNet34 as well to compare the performance with the other two models that we tried. However, since one of our team member dropped from the course at the end of the semester and that he was in charged of that model so we didn't try at the end. We would also spend some time on trying ResNet34 to compare the performance with what we did in this project.
Soft-attention mechanism is introduced in this paper[2], which will helps to better focus on the key features of the lesion images. In the future, we can implement the soft-attention mechanism to improve the performance of any ML models applied on the lesion images dataset. Also, we could change the architecture of ResNet50 based on what we have done in VGG16 to improve the performance.

8 Acknowledge

Group 37 Github link: https://github.com/leoisthebesta/ECE228-Group37
Link to part of codes in data pre-processing and metric evaluation generation are referenced below: https://github.com/skrantidatta/Attention-based-Skin-Cancer-Classification/blob/main/Ham10000%20models/Model%20without%20Soft%20Attention/Vgg16.ipynb Link to part of codes from ResNet Architecture in tensorflow: https://machinelearningknowledge.ai/keras-implementation-of-resnet-50-architecture-from-scratch/

9 Team Member Contribution

The idea of this project was proposed by the member who dropped from this course.
All the coding was completed by Xingtong and Zhiyin together. Zhiyin in charged of running the models and building VGG16 and Resnet50 from scratch. Xingtong in charged of running the models and building the models for transfer learning.
All the proposal, milestone and final report writing is separated equally to the two members. We also create poster together.
We spent every hour working on everything together!

References

[1] 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

[2] Datta, S.K., Shaikh, M.A., Srihari, S.N., Gao, M. (2021). Soft Attention Improves Skin Cancer Classification Performance. In: , et al. Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data. IMIMIC TDA4MedicalData 2021 2021. Lecture Notes in Computer Science(), vol 12929. Springer, Cham. https://doi.org/10.1007/978-3-030-87444-5_2

[3] Cassimiro, Gabriel. “Transfer Learning with VGG16 and Keras.” Medium, Towards Data Science, 17 June 2021, https://towardsdatascience.com/transfer-learning-with-vgg16-and-keras-50ea161580b4.

[4] Simonyan, Karen and Zisserman, Andrew, 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition, https://arxiv.org/abs/1409.1556

[5] Categorical crossentropy https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/categorical-crossentropy

[6] Sachin Mohan "Keras Implementation of ResNet-50 (Residual Networks) Architecture from Scratch" 16 Decemeber, 2020, https://machinelearningknowledge.ai/keras-implementation-of-resnet-50-architecture-from-scratch/