SkinDeep: Diagnosing Dermatological Images via Computer Vision

Liying Chen
University of Michigan
liyche@umich.edu

Farzad Siraj
University of Michigan
fsiraj@umich.edu

Jinhao Su
University of Michigan
sujinhao@umich.edu

Chris Wong
University of Michigan
wongch@umich.edu

Marisa Wong
University of Michigan
mariwong@umich.edu

Abstract

In this paper, we use the state-of-the-art VGG19 pretrained on the ImageNet dataset to classify common skin diseases and cancers. We train on the ISIC 2019 challenge dataset consisting of 25,331 images across 8 classes. Due to the large class imbalances in the dataset, by using careful preprocessing and weighted cross-entropy loss, we acheive a average weighted recall score of 0.72.

1. Introduction

Skin cancers are the most common human malignancy and are usually diagnosed in multiple steps. The very first step is a visual examination by a dermatologist who then determines the appropriate course of action and prescribes further testing. This initial step of visual examination is labor-intensive and has potential to be automated with recent advances in computer vision. Like most doctors, dermatologists face fatigue and overwork due to the rapidly increasing demand for their services. As training new dermatologists is an expensive endeavor, we hope an automated tool such as SkinDeep will reduce workload and enhance dermatologist performance.

1.1. Patient Outcomes

There are 5.4 million new cases of skin cancer in the United States every year. One in five Americans will be diagnosed with a cutaneous malignancy in their lifetime. Although melanomas represent fewer than 5% of all skin cancers in the United States, they account for approximately 75% of all skin-cancer-related deaths, and are responsible for over 10,000 deaths annually in the United States alone. Early detection is critical, as the estimated 5-year survival rate for melanoma drops from over 99% if detected in its earliest stages to about 14% if detected in its latest stages.

1.2. Role of Computer Vision

With recent advances in computer vision, deep convolutional networks have surpassed the golden standard in many domains - even beating out humans in many tasks such as classification on the ImageNet dataset. In many applications, these algorithms have proven to be much more efficient than humans with acceptable levels of accuracy - in segmentation tasks for example.

In the medical field however, an algorithm that works is not sufficient. Healthcare is a sensitive domain where people's lives are often at stake. An algorithm that performs well needs to be thoroughly analysed, gaining acceptance from various stakeholders including regulatory bodies, physicians, and hospitals. In order to do so, not only does it have to meet diagnostic standards, but also need to be explainable and predictable to a great extent. As such, stakeholders need to be informed about the strengths and weaknesses of the model - for example performance metrics on a per-class granularity.

1.3. Intended Use

Due to the sensitive nature of the medical field, SkinDeep is not intended to replace the diagnostic expertise of a trained dermatologist. It is intended to be used as a tool integrated into their workflow and serve as an additional piece of evidence in the overall diagnosis. This will require the dermatologist to be familiar with the specifications, strengths, and weaknesses of the model, but we believe this effort will lead to a smoother more informed dermatologist workflow.

1.4. SkinDeep

SkinDeep uses the VGG19 w/ Batch Normalization architecture pretrained on the ImageNet dataset as a feature extractor. The features extracted by its convolutional layers are fed into a 2-layer dense classifier. This classifier was trained using the 2019 International Skin Imaging Col-

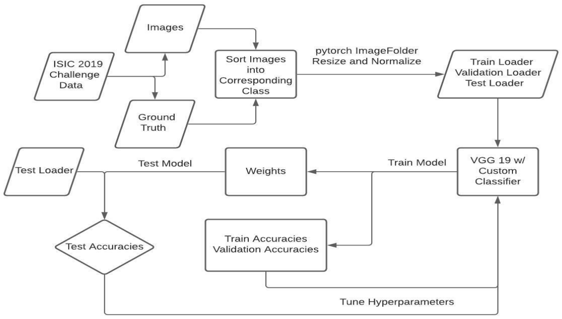
Images
ISIC 2019 Challenge Data
Sort Images into Corresponding Class
pytorch ImageFolder Resize and Normalize
Train Loader Validation Loader Test Loader
Ground Truth
Test Loader
Test Model
Weights
Train Model
VGG 19 w/ Custom Classifier
Test Accuracies
Train Accuracies Validation Accuracies
Tune Hyperparameters

Figure 1. Flow Chart of the Training Process

laboration (ISIC) dataset which consists of 25,331 images across 8 classes.

2. Approach

Our approach to the classification problem is to apply the idea of transfer learning from pretrained CNN models. There are many classification tasks that adopt this technique and it has proven to perform well for many medical tasks [3] [1].

2.1. Data Preprocessing

To process the input data, we resize each image to 224 × 224 × 3 and normalize the images to have mean of [0.485, 0.456, 0.406] and standard deviation of [0.229, 0.224, 0.225]. The train-validation-test split ratio is 8:1:1, where we carefully divide the dataset to make sure the split within each class also reflects this ratio.

2.2. Model and Hyperparameters

We use the VGG-19[4] model with pretrained weights from ImageNet and freeze all layers except the custom classifier we build. This custom classifier includes 3 fully connected layers that go from original input dimension of the classifier of VGG19 to 1024 and then from 1024 to 128 and finally from 128 to 8 which is the number of classes we have. The first two fully connected layers are activated with ReLU and have a dropout of 0.2. The loss function is cross-entropy loss and the optimizer is Adam with initial learning rate of 7e-4 and weight decay of 5e-6. We experiment with different learning rate and weight decay to arrive at these values which give the best test accuracy. All of the train, validation and test data are loaded with batch size of 128. For each session, we train the model for 20 epochs and save the state dict after each epoch. The model seems to eventually overfit for various learning rate and weight decays.

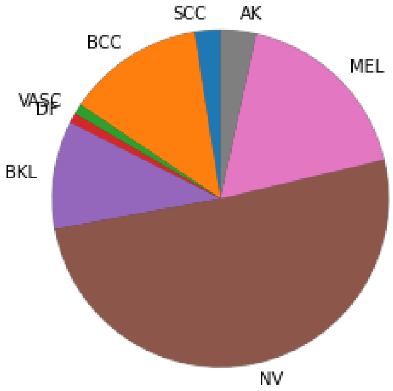

Figure 2. Distribution of diagnoses within 2019 ISIC dataset

3. Experiments

We used the 2019 ISIC challenge dataset for fine-tuning and evaluation. ISIC, the International Skin Imaging Collaboration, holds an annual competition for the diagnosis of various skin conditions using machine learning. The 2019 dataset contains approximately 23,000 dermoscopic JPEG images, each with an associated ground-truth diagnosis, as well as a spreadsheet of metadata containing patient age, sex, and where on the body the image came from. The dataset contains 8 categories of diagnoses, with each image only belonging to a single category. Unfortunately, the sizes of the categories are not evenly distributed as seen in Figure 2, so this discrepancy was accounted for in our training loop and evaluation. We used this dataset for testing because it is our goal to automate the diagnostic process by classifying skin cancer into their respective diagnoses using a single dermoscopic image. Dermoscopic images increase the model performance, as they show skin lesions unobstructed by skin surface reflection.

3.1. Metrics

Success was measured quantitatively with F1 scores, recall, precision, and accuracy. We strived to achieve the highest metrics we possibly could. Our model achieved an F1 score of 0.7291, recall of 0.7288, precision of 0.7326, and accuracy of 0.7288. This was reasonable because of the size of the dataset. From these scores, we know the proportion of skin lesions that are correctly classified, positive cases that are actually correct, as well as the proportion of true positive cases that are detected out of each category. A potential downside to this would be not analyzing the results for trends in how the model incorrectly labels images. For example, if there are two categories that are very similar to each other and those two categories make up a majority of the incorrect labels, then the metric scores of the other 6 categories are higher than our results and the metric scores for the two similar categories should be lower.

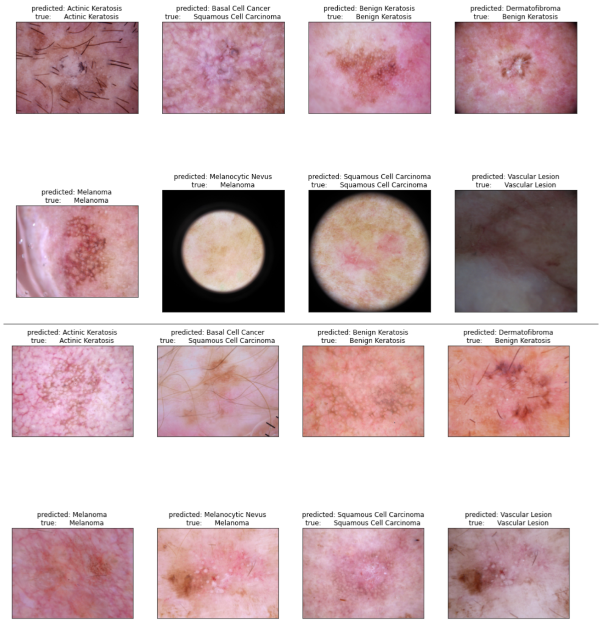
predicted: Actinic Keratosis
true: Actinic Keratosis
predicted: Basal Cell Cancer
true: Squamous Cell Carcinoma
predicted: Benign Keratosis
true: Benign Keratosis
predicted: Dermatofibroma
true: Benign Keratosis
predicted: Melanoma
true: Melanoma
predicted: Melanocytic Nevus
true: Melanoma
predicted: Squamous Cell Carcinoma
true: Squamous Cell Carcinoma
predicted: Vascular Lesion
true: Vascular Lesion
predicted: Actinic Keratosis
true: Actinic Keratosis
predicted: Basal Cell Cancer
true: Squamous Cell Carcinoma
predicted: Benign Keratosis
true: Benign Keratosis
predicted: Dermatofibroma
true: Benign Keratosis
predicted: Melanoma
true: Melanoma
predicted: Melanocytic Nevus
true: Melanoma
predicted: Squamous Cell Carcinoma
true: Squamous Cell Carcinoma
predicted: Vascular Lesion
true: Vascular Lesion

Figure 3. Sample from test set with predicted labels and true labels

3.2. Results

We can compare our accuracy to random chance and our initial proposition of 60% of the accuracy achieved in the Stanford paper [2]. Since we classified images into 8 categories, random chance would have a 12.5% chance of correctly classifying the images. Our model's accuracy of 0.7288 exceeded this baseline. In the Stanford paper, they achieved a 9-way accuracy of about 50% [2], so 60% of this would give us a goal of 30% accuracy. Our model's accuracy also exceeded this goal. See Figure 3 for examples of how our model labeled images against the true labels.

4. Implementation

The implementation that we used was based initially on the results of a 2017 paper [2] which was also attempting to diagnose dermatological conditions using machine learning. The model uses an Inception V3 CNN pretrained using the ImageNet dataset. However, our model architecture deviated significantly after initial experimentation.

Our model also makes use of transfer learning via the pretrained VGG19bn model imported from PyTorch, the results of which are then fed through a customized classifier. However, as seen in Figure 2, the distribution of the samples used was heavily skewed, and required correction in the training loop. Therefore, we took advantage of weighted loss functions to account for these discrepancies, with moderate success. All of our code was run on Google Colab notebooks, with notebook formatting heavily influenced by that of homework 5. Each iteration took approximately 3 hours to train, validate, and test. Then, for the testing partition of the dataset, random images were selected for visualization, and each of the selected images are plotted, with labels indicating the predicted diagnosis as well as the ground truth diagnosis.

References

[1] M. S. Elmahdy, S. S. Abdeldayem, and I. A. Yassine. Low quality dermal image classification using transfer learning. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 373–376, 2017.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature (London)*, 542(7639):115–118, 2017.

[3] B. Q. Huynh, H. Li, and M. L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):1 – 5, 2016.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.