Machine Learning Engineer Nanodegree

Capstone Proposal: Skin Cancer Detection using Transfer Learning of Deep Convolution Networks

Daniel C. Chan
August 2018

Domain Background

Skin cancer is the uncontrolled growth of abnormal skin cells. It occurs when damaged skin cells (most often caused by ultraviolet radiation) triggers mutations, or genetic defects, that lead the skin cells to multiply rapidly and form malignant tumors. Melanoma, is the most dangerous form of skin cancer, kills an estimated 10,130 people in the US annually. Skin cancer is also the most common type of cancer and the easiest to cure, if diagnosed and treated early. Each year in the U.S. over 5.4 million cases of non-melanoma skin cancer are treated in more than 3.3 million people [1].

To spot potential Melanoma skin cancer, the American Cancer Society recommend conducting a monthly self-visual inspection using the "ABCDE rule" to look for Asymmetry, Border, Color, Diameter and Evolving, respectively, of a suspicious skin lesion [2].

To aid visual inspection one can purchase an affordable dermatoscope for less than $300 from an ecommerce site such as Amazon [3]. It is equipped with its own bright light source and a magnification lens that can illuminate deep structures in the skin not visible to naked eyes due to the glare from skin surface [4]. High resolution pictures can be taken with a smart phone or traditional camera for applying the "ABCDE rule". Using these images to train a machine-learning model could form the basis of an affordable early detection system and potentially save more human lives.

Problem Statement

If basic structural differences exist that can distinguish different types of skin cancer from each other, a Deep Convolution Neural Network (DCNN) should be able to learn from these dermoscopic images and classify them consistently. However, publicly available images are limited, and in most cases less than a thousand for each skin cancer type and will not lend themselves for training DCNN from scratch. Therefore, transfer learning/feature extraction or fine tuning would be a more appropriate approach. In addition, benign cancer data are more abundant than malignant ones, thus creating a highly imbalance data set for training, if it is not treated appropriately, the resulting model could have a bias towards benign predictions and defeat the purpose of an early detection system. Computational efficiency is also a concern, since fine tuning can take hours or days to perform for each case and can limit the number of parameters that can be explored.

The goodness of the prediction will be measured by the balance between sensitivity and specificity as well as the AUC (to be described) for binary classification. Sensitivity measure the proportion of actual positives that are correctly classified, whereas, specificity is the proportion of actual negatives that are correctly classified. The average specificity by dermatologists at a sensitivity of 0.8 is about 0.4 [5]. To assess prediction repeatability, I will be using the 5-fold cross-validation method.

Datasets and Inputs

The International Skin Imaging Collaboration (ISIC) is an international effort to improve Melanoma diagnosis, sponsored by the International Society for Digital Imaging of the Skin (ISDIS). The ISIC Archive [6] contains the largest publicly available collection of quality controlled dermoscopic images of skin lesions. ISIC started hosting open challenges for skin lesion analysis towards Melanoma detection in 2016. Training and test data are readily available through a simple registration process. For the 2017 challenge, ISIC provided 2000 images and allowed the use of external data for training purpose only. Models that have been trained with these data were then put to the test by classifying the disease captured in 600 different images. There were 23 final test set submissions. All top submissions implemented various ensembles of deep learning networks. All used additional data sources to train, either from ISIC, in-house annotations or external sources [7]. I believe the key motivation by this approach is to prevent overfitting caused by small training dataset and avoid the bias inducted by an imbalanced dataset. Liberal use of external data resources can achieve higher score; however, it can also make the results difficult to replicate and not easy to understand key driving factors for building a robust early detection system.

For this study I chose to focus on the 2017 datasets and opted not to use data from other sources that may contain varying degrees of quality. The objective is to provide a systematic way to evaluate the capability and limitations of transfer learning.

Solution Statement

In summary, I am proposing a solution that consists of the following steps:

1. Using transfer learning, I will explore different DCNN models that have been trained with ImageNet data and evaluate their capability in predicting skin cancer types, they may include VGG16, ResNet50 and InceptionV2. Their performance will be evaluated using a 5-fold cross validation and comparing with the ISIC 2017 test data
2. Imbalanced data will be treated by adjusting their weights in the loss function calculation
3. Computational efficiency will be achieved algorithmically by adjusting the learning rate cyclically [8] and by using Nvidia GPU working in concert with multiprocessing at the CPU level. Cross Entropy loss and computational time will be monitored at each epoch to make sure the training loss is close to machine error (10^{-6}) and that the validation loss exhibits no sign of overfitting.

Benchmark Model

In essence, the transfer-learning solution I am assembling together is for multi-class classification. To quantify its capability, I have applied it to the well-known CIFAR-10 dataset [9] which contains 60,000 32x32 color images in 10 different classes. The 10 different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. There are 6,000 images of each class. I used 50,000 of them as training data and 10,000 for validation to train ResNet50 and VGG16 models. Weights were initialized to those obtained from ImageNet. The quality of the prediction is measured by its accuracy in classifying the test data correctly. The figure below shows the convergence rate of Stochastic Gradient Decent Method (SGD), with momentum=0.8, batch size=32, cyclic learning rates

that vary between 0.01 and 10^{-4}, for 50 epochs. It took 6 hours of computational time on a Nvidia 1080Ti GPU or 7.2 minutes/epoch. Validation accuracy of the current approach is 93.7% and 95.51% for VGG16 and ResNet50, respectively, which are very competitive to the top result of 96.53% reported by Rodrigob [10]. For the ResNet50 model, it reaches a validation accuracy of 95.51% after 2 epochs, the computation could have stopped there using an early stopping strategy and resulted in a computational time of 14.4 mins which would make it a top-5 training time in the DAWNBench competition [11].

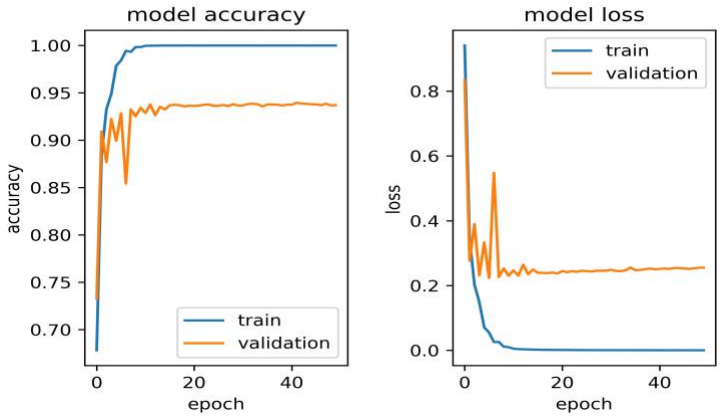


Figure 1. Convergence characteristics of the VGG16 model. Training accuracy reaches 100% and training loss is nearly zero. No overfitting is detected. Cyclic learning rates with the upper bound set to 10^{-3} to prevent numerical divergence.

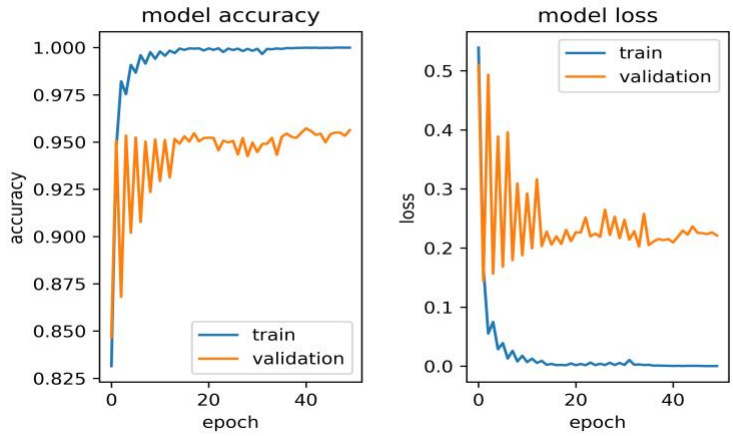


Figure 2. Convergence characteristics of the ResNet50 model. Training accuracy reaches 100% and training loss is at 2.84×10^{-4}. No overfitting is detected. Cyclic learning rates with the upper bound set to 10^{-2}. Model achieves the steady-state validation accuracy and loss after 2 epochs.

Evaluation Metrics

Class prediction accuracy is used for the CIFAR-10 benchmark. It measures the proportion of data with the correct class predicted. It is defined as:

$$\text{Accuracy}=\frac{(TP+TN)}{(TP+TN+FP+FN)}$$

where TP=True Positive, TN=True Negative, FP=False Positive and FN=False Negative.

Under certain circumstances, however, it could be a misleading indicator. In the case where both TP and FP are zero, the accuracy can reach 80% if the ratio between TN and FN is at 4. This is an example which the model could be making all negative predictions and still achieved a respectful accuracy. The reason it works for CIFAR-10 is because the number of test samples for each class is the same making this is an unlikely scenario.
For the ISIC 2017 data, the ratio between benign and malignant data is 4.3, so accuracy would not be an appropriate metric. Instead, specificity and sensitivity will be used as the evaluation metrics [12]. Sensitivity (also known as Recall or True Positive Rate) is defined as:

$$\text{Sensitivity}=\frac{TP}{TP+FN}$$

and Specificity (also known as True Negative Rate or 1 minus False Positive Rate) is defined as:

$$\text{Specificity}=\frac{TN}{TN+FP}$$

One can also plot True Positive Rate against False Positive Rate to measure Receiver Operating Characteristic (ROC) and the area under the ROC curve (AUC) is an effective metric, even for imbalance data set, to gauge the performance of a binary classification model [13].

Project Design

According to the "ABCDE" rules, asymmetry and diameter of a skin lesion are key attributes to look for, therefore it is important not to elongate or squish any of the images. As all the DCNN models available in Keras [14] only work with square images in a size of either 224x224 or 299X299, I need to run these images through several preprocessing steps:

1. Locate the center of each image
2. Determine the shorter dimension of 2 sides
3. Use the above dimension, crop a square image and output it in jpeg format to a folder with name corresponds to the skin cancer type
4. These images will be resized to the appropriate dimensions during training

As shown in Table 1, these images have very large sizes with aspect ratio ranging from 0.75 to 1.55. Not only will cropping maintain the original image aspect ratio, it will reduce the image size for computational efficiency, especially in memory requirement, as the GPU in use has a limit of 11GB.

Table 2 shows there is an imbalance in different skin cancer types. Nevus has the highest number of training images with Melanoma being the least. The ratio is about 3.7 to 1. I plan to compensate this

imbalance by either adjusting the class weights in the entropy loss calculation or oversampling the training data with various geometric transformations.

Skin Cancer Type	Minimum Image Size (width, length)	Maximum Image Size (width length)	Average image width and length	Min, max Aspect Ratio
Melanoma	(679, 566)	(6708, 4459)	2715, 1876	0.83, 1.55
Nevus	(576, 768)	(6748, 4499)	2545, 1752	0.75, 1.52
Seborrheic Keratosis	(1007, 704)	(6708, 4439)	3780, 2554	1.34, 1.53

Table 1. ISIC 2017 training data set image sizes and aspect ratios

Skin Cancer Type	Number of training images	Number of validation images	Number of test images
Melanoma	374	30	117
Nevus	1372	78	393
Seborrheic Keratosis	254	42	90

Table 2. Number of images that are available for the ISIC 2017 competition

Once training data are prepared, I will conduct the following steps using the Keras (version 2.2) API and Tensorflow (version 1.8) backend:

1. feed images into computer memory using the Generator API in Keras
2. Load pre-trained models (e.g. VGG16, VGG19, ResNet50 or InceptionV2) without the top layer, since it is problem specific, refer to this model as base_model
3. Use the penultimate layer to predict a set of image features and write them to disk
4. Input these features to a model constructed with one dense and one softmax layer, refer to this as the top_model
5. Train the top_model with SGD method and cyclic learning rates to numerical convergence
6. Save the top_model weights to a file
7. From here, I will explore 2 possible options:
    a. Concatenate base_model and top_model into one model, initialize the top 2 layers with the weights from step 6 and then train the entire model (also known as fine tuning) by unfreezing a few layers at a time from top to bottom
    b. Work with the image features in step 4 and input them to various classification methods such as dense neural network and Gradient Boost Tree method (e.g. Lightgbm), then aggregate the results using an ensemble method (details will be provided in the project report)

Results will be output as probabilities for each cancer type. I will be making 2 sets of binary classification comparison. First one is for Melanoma (malignant) versus the rest of benign cancer types. The second one is for Seborrheic Keratosis versus Melanoma and Nevus. True and False Positive Rates will be calculated and plotted for comparison. From them, we can also determine the AUC using the Scikit-Learn Metrics API [15].

Additionally, for classification of melanoma, Specificity will be measured on the ROC curve where Sensitivity is equal to 82%, 89%, and 95%, respectively. They correspond to dermatologist classification and management performance levels [16].

Figures 3 and 4 show some of the preliminary results. In Figure 3, Melanoma features are less distinguishable than the other skin cancer types supporting the argument that more sophisticated classification methods are needed. Figures 4 show a significant variation in AUC during a 5-fold cross validation. This result hints the need for an ensemble method.

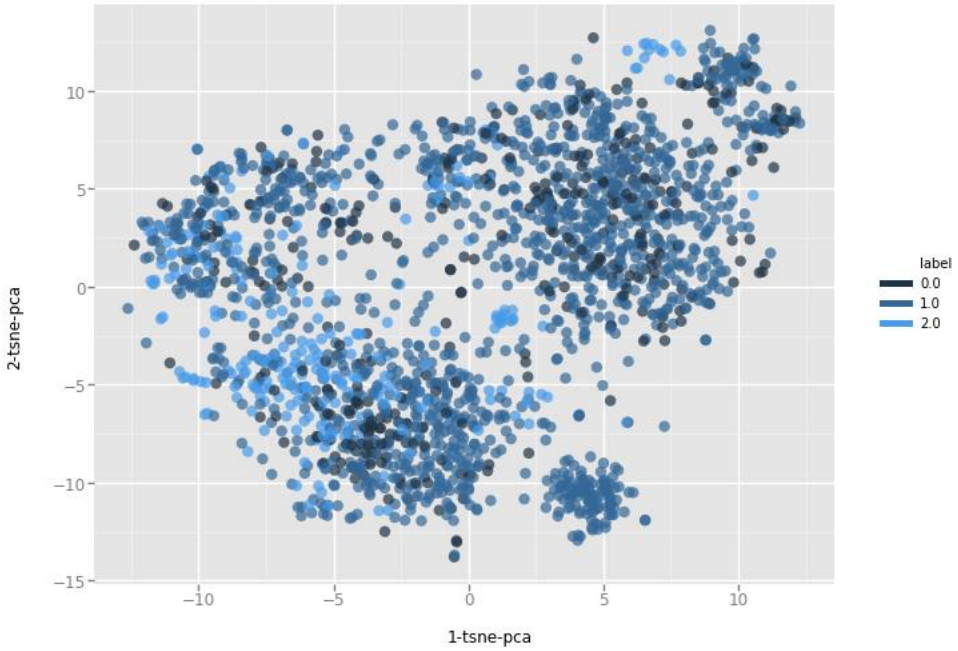
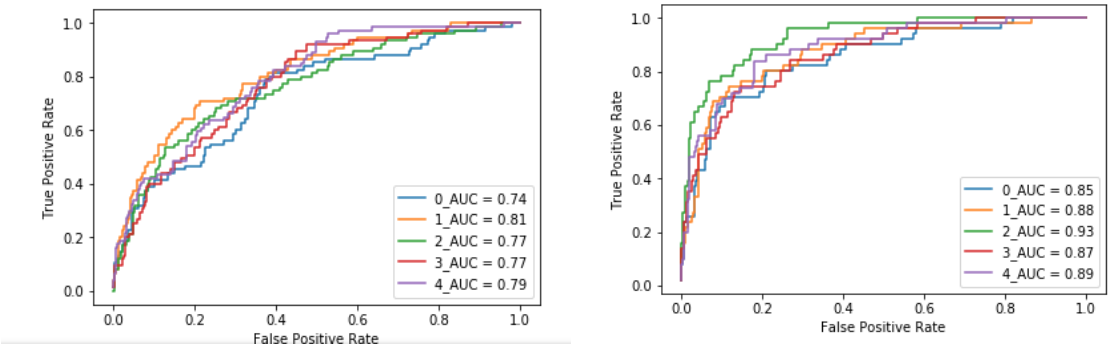


Figure 3: t-SNE visualization of the image features output by the penultimate layer of ResNet50 model showing the concentration of different cancer types. Label 0, 1 and 2 is for Melanoma, Nevus and Seborrheic Keratosis, respectively.



Figures 4: ROC predicted by ResNet50 generated features and a Dense layer. 5 different curves represent the result of a 5-fold cross validation. They are on par with other predicted results using large amount of data sourced externally to the competition.

References

[1] Skin Cancer Facts and Statistics, https://www.skincancer.org/skin-cancer-information/skin-cancer-facts

[2] How to spot cancer? https://www.cancer.org/latest-news/how-to-spot-skin-cancer.html

[3] The cost of a pocket dermascope in Amazon as of April 28, 2018, https://www.amazon.com/3Gen-DermLite-DL100-Dermatology-Dermascope/dp/B000X2INPY/ref=sr_1_1?ie=UTF8&qid=1524953402&sr=8-1&keywords=dermatoscope

[4] How does a dermatoscope work? http://www.dermatoscope.info/how-does-a-dermatoscope-work

[5] A second set of eyes - Using computers to aid melanoma detection, by Michael Marchetti, Oct 4, 2017, IBM Blog Research https://www.ibm.com/blogs/research/2017/10/computers-to-aid-melanoma-detection/

[6] International Skin Imaging Collaboration (ISIC) Challenges https://challenge.kitware.com/ - challenges

[7] Skin Lesion Analysis toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC) https://www.arxiv-vanity.com/papers/1710.05006/

[8] A Disciplined Approach to Neural Network Hyper-Parameters, by Leslie N. Smith US Naval Research Laboratory, https://arxiv.org/pdf/1803.09820.pdf

[9] CIFAR-10, https://www.cs.toronto.edu/~kriz/cifar.html

[10] CIFAR-10 Benchmark Results, http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html#43494641522d3130

[11] DAWNBench: An End-to-end Deep Learning Benchmark and Competition, https://dawn.cs.stanford.edu/benchmark/index.html#cifar10-train-time

[12] Sensitivity and specificity, Wikipedia, https://en.wikipedia.org/wiki/Sensitivity_and_specificity

[13] What you wanted to know about AUC, Sept 19, 2013, http://fastml.com/what-you-wanted-to-know-about-auc/

{14] Keras Documentation, https://keras.io/

[15] Scikit-learn, Machine Learning in Python, http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

[16] Dermatologist-level classification of skin cancer with deep neutral networks, by Andre Esteva, Brett Kuprel, Roberto Novoa, Justin Ko, Susan Swetter, Helen Blau and Sebastian Thrun, doi:10.1038/nature21056, Nature, https://www.nature.com/articles/nature21056.epdf?shared_access_token=tPhBdW2oX8dei9v85_BGR9RgN0jAjWel9jnR3ZoTv0NXpMHRAJy8Qn10ys2O4tuPtovyxVYI1vj6wDkw5KrpdIP0sH9ajTy-n5Cd09NMTEXy6JYE9xBvS-qGgozoAxMpYuSM-K7S7WQj_lmdQ-oPw8KLoGEsyCynW-FDUGJ7M2k%3D