Classifying Skin Lesions with Simulated Lens Blur

Christina Le
Department of Biomedical Engineering
Duke University
Durham, NC 27708

Allison Li
Department of Biomedical Engineering
Duke University
Durham, NC 27708

Abstract

In 2020, skin cancer is the most prevalent form of cancer in the United States. However, the survival rate is an astonishing 99% when detected early [1]. The following experiment explores the possibility of using a convolutional neural network (CNN) to aid physicians in the diagnosis and recognition of melanomas using solely image recognition. Using the ISIC 2017 Skin Lesion dataset, a model combining Visual Geometry Group 16 (VGG16) and a physical layer to simulate the effect of blurring was trained on multiple augmented variations of an original set of 2000 images, validated with 200 images, and tested for performance on 600 images. The addition of the physical layer evaluates the relationship between the quality of an image to the accuracy of diagnosis. The final results for this model show that blurring does not have a negative effect with an average area under the curve (AUC) of 82.5% that measures the degree of separability between the two classes of melanoma and seborrheic keratosis or benign nevi and a maximum average loss of 0.0324.

1 Introduction

Image recognition tasks have become very relevant and important. These tasks involve training a computer model to classify images given a set of labels. Convolution neural networks have been instrumental in building accurate models to execute image classification. These deep learning neural networks take images with corresponding labels as inputs and generally contain millions of weights to be trained on these inputs. Due to the high performance of convolutional neural networks, image classification has begun to play a larger role in the medical field, specifically for disease diagnosis. Medical image classification places a large burden on medical professionals to acquire many patient image data and annotate each image with appropriate labels. At times, annotation requires a consensus among doctors or further medical tests, so the manual cost is high. In general, the cost of acquiring accurate medical image data for image classification can be significant.

In this work, we are creating a melanoma binary classifier with images of skin lesions from the ISIC 2017 challenge. Melanoma is a form of skin cancer that may be dangerous if not treated early. Skin lesions cannot accurately be classified as melanoma just by looking and generally require a biopsy for a concrete diagnosis. It is possible that certain features of a skin lesion that may not be immediately visible to the human eye may be captured and differentiated by a convolutional neural network. To evaluate a deep learning approach to classifying melanoma, we are utilizing a convolutional neural network. Before input into the neural network, we are first passing our images through a physical layer that may simulate images that originate from a lower resolution camera and are blurred by the camera lens in hopes of studying the effect of blurring on the network's ability to classify skin lesions. The simulated physical layer can have implications on image resolution and quality requirements for accurate diagnosis of melanoma.

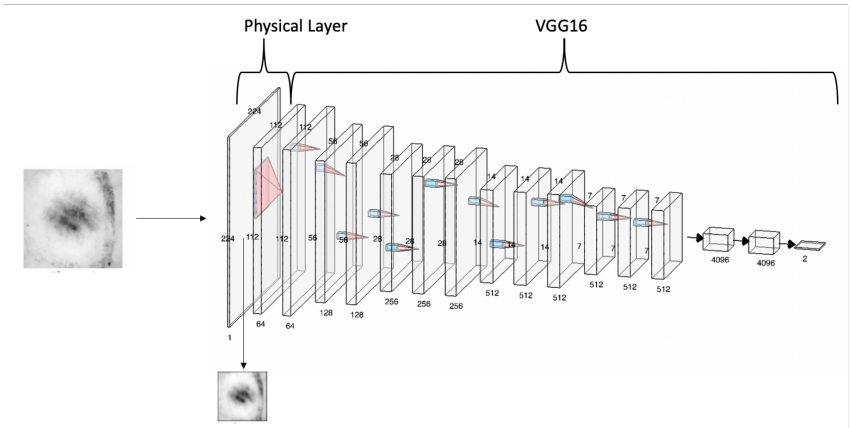

Figure 1: Model used for training, including the physical layer and VGG16 architecture. An input skin lesion image is shown on the left. The image at the bottom is an example of the input image after being fed through the physical layer. After the physical layer, the image has been blurred and downsampled.

2 Related Work

Melanoma identification and diagnosis through machine learning is problem that has been tackled before. The context of many published works pertaining to melanoma diagnosis highlights the pertinence of these models healthcare, alluding to future mobile applications that will help detect and diagnose skin cancer. Many of the first published models for melanoma classification focused on segmentation masks, clustering, and supervised learning [2]. The commonality between these models is the importance of border, diameter, and asymmetry for feature extraction [2, 3]. The complexity of the models used in these papers is directly proportional to the scale of the classification task. When only performing binary classification on melanomas versus all other types of skin lesions, Visual Geometry Group 16 (VGG16) or AlexNet are commonly used as the base architecture for the model. However with larger classification tasks that branch out to classify subsets of benign, malignant, and non-neoplastic lesions, more complex networks like DenseNet or Google's Inception v3 CNN are used for their expressive power [4].

In more recent works, the performance of these models is compared directly to the diagnoses made by dermatologists when looking at the images with the raw eye [4]. Models have been shown to have higher accuracy while dermatologists' accuracy is reduced by false positive rates. While these models generally have high accuracy and low loss, the most accurate diagnosis for skin cancer is by performing a biopsy. Many of the models discussed and cited here were built to illuminate the possibility of cancer as a push for early detection.

Inspiration for the physical layer used in this experiment came from a variety of experiments, specially those simulating a phase-coded aperture physical layer before their model [5]. These experiments show that allowing a model to determine the most optimal state for the feature being studied (blur in this specific case) will generally result in better performance.

3 Method

3.1 Training

At the start of designing this experiment, a variety of established architectures including AlexNet, DenseNet, and VGG16/19 as well as a simple convolutional model built from four convolutional layers and two dense layers were in consideration. When evaluating the performance of the models against each other, the main concern was the direct relationship between overfitting and extensive training due to number of trainable parameters. We also wanted to select a model that would not be severely affected by the class imbalance in our datasets. While universally accepted to perform well for many classification tasks, DenseNet's network architecture was deeper than necessary for

our classification task and prone to overfitting because of the large amount of parameters. After comparing our implementations with related works, we ultimately chose to go with VGG16 as our model of choice as a medium between excessive parameters and the ability to learn and establish more features that could not be captured in our simple model.

All data used throughout this model comes from the ISIC 2017 Skin Lesion dataset. To split up our dataset into a training set and validation set, we used K-fold cross-validation with K=10 to randomly partition the training set of 2000 images, resulting in 1800 images used for training and 200 images used for validation. 600 images were used to evaluate the model after 20 epochs of training. The most interesting, reported statistics in our model include the area under the curve (AUC) that ties together sensitivity and specificity for classification and focal loss, which will be explained in the following paragraphs.

To reduce overfitting and increase the generalizability of our model, we utilized data augmentation and generated a "new" set of images for each batch. The data augmentation techniques we explored include rotation through an angle of 90° and horizontal/vertical flipping. These augmentations make sense in context of skin lesions, which generally do not have a specific orientation. Before augmenting our data, we summed the images along their color channels, resulting in a modified dataset of grayscale images. We compared the model's performance with grayscale images to colored images and saw no significant difference, which led to the decision of using grayscale images for our training for better compatibility with our physical layer.

The optimizer we chose was Adam with a learning rate of 0.0001 because of its consistent performance compared to Stochastic Gradient Descent (SGD) without specifying additional parameters and momentum. However, our results indicate that a learning rate schedule could have improved performance.

Our proposed model utilizes focal loss to train and optimize its predictions. Focal loss is cross-entropy loss with a modulation term to account for class imbalance. Focal loss was originally applied to the task of object detection, where there may be large foreground/background class imbalance [7]. The focal loss function in our model was built to calculate loss based on one-hot predictions and labels [6]. The ISIC 2017 Skin Lesion dataset contains skewed image data with respect to labelling, where there are many more non-melanoma examples than melanoma examples (Table 1). Focal loss allows us to down-weight the "easy" non-melanoma examples, which are "easy" due to the class imbalance, since the model is training with many more non-melanoma examples compared to melanoma examples. Compared to standard cross-entropy loss, we consistently saw better performance using focal loss for this classification task.

3.2 Physical Layer

The physical layer is composed of a convolutional and an average pooling layer. In our model, this physical layer is applied to the images prior to the VGG16 model. In effect, we are attempting to simulate a much lower resolution incoherent imaging system with lens blur and to train our model as if these skin lesions images originated from such a system.

We are additionally attempting to optimize the lens blur by making the weights of the convolutional kernel trainable. The weights of the convolutional kernel are initialized randomly. A kernel constraint of non-negative values is placed on the layer because we are modeling incoherent imaging, in which negative values have no meaning. The network must learn a blur that is optimal to accurately classify the skin lesions. The average pooling layer is fixed and pools two-by-two pixel regions, downsampling the image so that an input 224x224-sized image becomes a 112x112-sized image. The resulting blur kernel could be implemented in hardware to develop a lower-resolution image system, with the lens blur as specified by the learned kernel, that creates images which are optimized for melanoma classification by our VGG16 model.

In addition to allowing the weights of the kernel to be trainable, we manually varied the kernel size of the convolutional layer to test how the network may behave with different numbers of weights to train. The final optimized convolutional kernel was extracted from the network after each instance of training and applied to sample images from our training set to understand visually how the learned convolutional kernel was blurring skin lesion images.

Table 1: Distribution of Datasets

Set	Melanoma	Seborrheic Keratosis or Benign Nevi
Training	337	1468
Validation	37	163
Testing	117	483
		*reports the number of images found per class

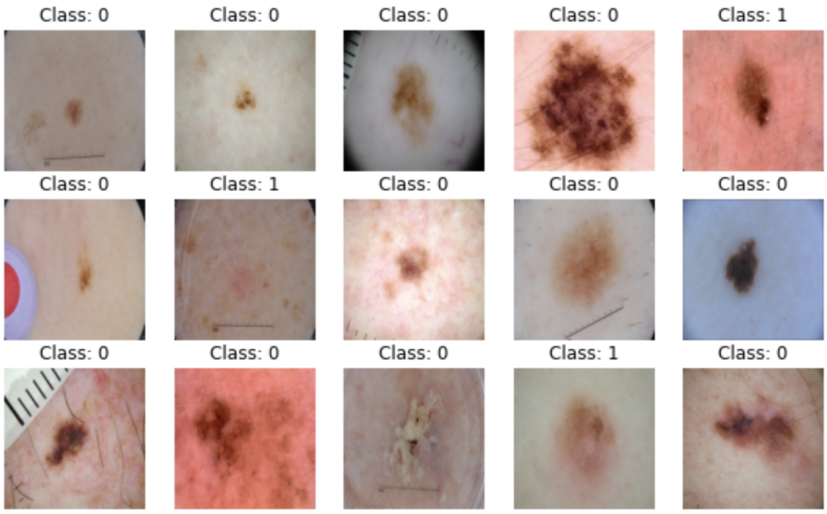

Figure 2: Skin lesion images with class labels from the 2017 ISIC Challenge Part 3. A class label of 1 indicates that the skin lesion is melanoma.

4 Results

The ISIC Skin Lesion data contains 2000 training images and 600 test images. We split the training images into 1800 for the training set and 200 for the validation set. Example images are shown in Figure 2. As mentioned prior, the images were summed along their color channel to convert them into one-channel, grayscale images.

Experiments were conducted to understand the effect of the physical layer on classification performance. In each experiment, the model was trained using the same optimizer and loss function for 15-20 epochs. Data augmentation, including horizontal and vertical flipping and rotation, was also applied to the training set. Table 1 shows the distribution of class labels among each dataset. Each experiment was conducted 5 times. After training, the test loss and test AUC were computed using the test dataset. The average results of each experiment are listed in Table 2. Training with solely the VGG16 model and no physical layer was conducted to establish a baseline loss and AUC; this is the No blur instance in the table. Physical Layer denotes an experiment in which the convolutional layer weights were trainable. Gaussian Blur denotes an experiment in which the convolutional layer weights were fixed to a Gaussian function and not learned by the model.

5 Discussion

Throughout our experiments, the loss remained relatively the same. We saw greater differences in the AUC results, which indicate that allowing the model to learn weights for the convolutional kernel resulted in better performance than fixing the kernel to be a Gaussian. In all kernel size experiments, the learned convolutional layer outperformed the fixed Gaussian layer.

We also notably saw that the AUC for learned convolutional layers was higher than the AUC for the original experiment with no blur or downsampling applied. This result is very interesting given that larger kernel sizes relate to a greater degree of blurring. We saw increased AUC and slightly

Table 2: Test Loss and Test AUC after training

Experiment	Kernel Size	Loss	AUC
No blur	–	0.0321 ± 0.000187	0.815 ± 0.00939
Physical Layer	10	0.0321 ± 0.000186	0.813 ± 0.00862
Physical Layer	25	0.0322 ± 0.000346	0.825 ± 0.00470
Physical Layer	50	0.0322 ± 0.000626	0.823 ± 0.00358
Physical Layer	100	0.0320 ± 0.000165	0.827 ± 0.00207
Physical Layer	200	0.0320 ± 0.000104	0.827 ± 0.00215
Gaussian Blur	10	0.0321 ± 0.000217	0.813 ± 0.01064
Gaussian Blur	25	0.0321 ± 0.000268	0.812 ± 0.00920
Gaussian Blur	50	0.0321 ± 0.000219	0.807 ± 0.00766
Gaussian Blur	100	0.0322 ± 0.000394	0.807 ± 0.00428
Gaussian Blur	200	0.0324 ± 0.001076	0.810 ± 0.00126

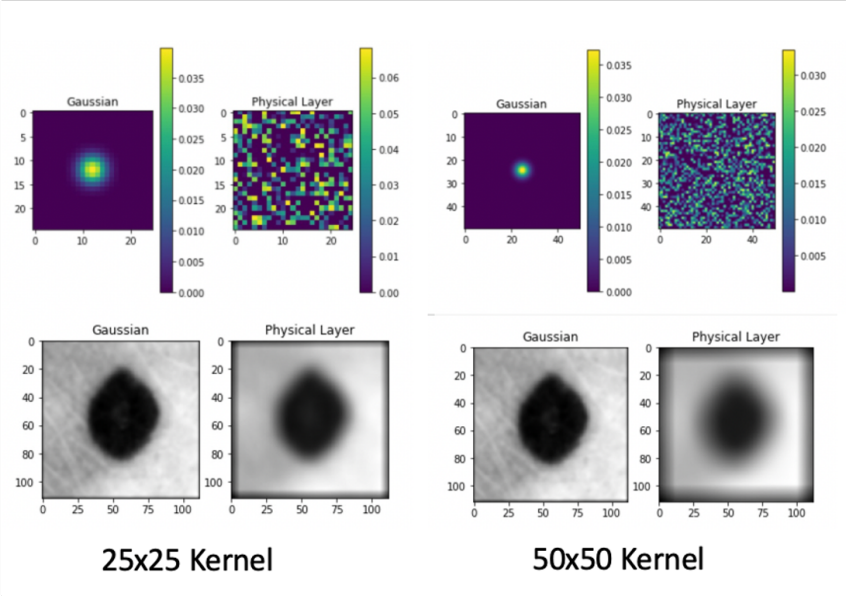

Figure 3: Resulting skin lesion image given a convolutional kernel. The kernel from gaussian and physical layer experiments are shown above. The physical layer weights were learned by the model after 15 epochs of training. Each kernel is applied to the same skin lesion to illustrate differences in the fixed applied blur and the learned applied blur.

decreased loss with larger blurs in the physical layer experiments. This conflicts with our general assumption that blurring might make it more difficult to classify images. It is possible that the learned blur is somehow highlighting important features in the skin lesion or blurring out less important features.

Our work is limited by the skin lesion dataset we utilized. Due to the class imbalance, our network may not have enough examples of melanoma to accurately classify skin lesions. With more examples of melanomas, we would expect to have additional important features to help the classification. It is possible that lack of melanoma images currently is preventing our model from effectively training and assigning meaningful values to all of the parameters, resulting in a higher chance for overfitting when analyzing the statistics. Additionally, it is possible that our results involving the physical layer and seemingly positive effect of lens blur and resolution on performance would not hold given a more balanced dataset. Nonetheless, we have established a framework for simulating lens blur in an incoherent, lower-resolution imaging system, so we could further expand our work by acquiring more skin lesion data, and specifically more melanoma examples, and performing additional training with the same physical layer.

6 Conclusion

The prevalence of skin cancer in the United States sheds a light on the importance of the development of additional methods for diagnosis and identification, specifically through the aid of a CNN image recognition model. A physical layer imposed on a traditional CNN can also specify hardware equipment that might best aid the CNN in classification performance. While this experiment touches on the differentiation between melanoma versus benign skin lesions, future work could tackle a larger classification problem to correctly identify the differences within subsets of malignant and benign lesions. The accuracy of those models could be furthered compared to classification performed by dermatologists to gauge the practicality of such a model in the modern world.

References

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

[2] Yuexiang Li and Linlin Shen. Skin Lesion Analysis Towards Melanoma Detection Using Deep Learning Network, 2017; arXiv:1703.00577

[3] Aminur Rab Ratul, M. Hamed Mozaffari, Won-Sook Lee, and Enea Parimbelli. Skin Lesions Classification Using Deep Learning Based on Dilated Convolution, 2019;

[4] Amirreza Rezvantalab and Habib Safigholi and Somayeh Karimijeshni. Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms, 2018; arXiv:1810.10348.

[5] Shay Elmalem, Raja Giryes, and Emanuel Marom. Learned phase coded aperture for the benefit of depth of field extension, 2018.

[6] Multi-class classification with focal loss for imbalanced datasets. DLology. https://www.dlology.com/blog/multi-class-classification-with-focal-loss-for-imbalanced-datasets/

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He and Piotr Dollár. Focal Loss for Dense Object Detection, 2017; arXiv:1708.02002.