Seth Thompson

Machine Learning

Final Project

Dermatologist-level classification of skin cancer with deep neural networks

Book Report/Analysis/Explanation

Abstract
Skin cancer rates have been or the rise for the last 30 years in America as well as other parts of the world. Skin cancer is the most common cancer in the U.S.[1]. Utilizing Convolutional Neural Networks (CNN's) for image classification of dermatological images shows potential to advance the medical treatment of certain types of skin cancer by reducing the time before a patient receives treatment. Utilizing GoogLeNet Inception v3 architecture, it is possible to achieve on par performance with board certified Dermatologists on biopsy-proven images. The images used for comparison consist of two cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi.

Introduction
Human skin is made up of basal cells, squamous cells, and melanocytes. Skin cancer is the most common type of cancer in the U.S. According to cancer.org, about 4.3 million people in the U.S. will be diagnosed with a Basal or Squamous cell carcinoma each year resulting in around 3,000 deaths from these diagnoses. There are approximately 91,270 diagnosed with melanoma each year resulting in approximately 9,320 deaths[2]. While melanoma accounts for less than 5% of all skin cancer diagnoses, it accounts for over 75% of all skin cancer deaths[1][2]. Melanoma can become deadly within as little as 6 weeks of development. With the average wait time at a dermatologist of more than 30 days[3], this time spent waiting for a diagnosis can prove invaluable to a patient that potentially has skin cancer. The ABCDE's of skin cancer are one of the most common guidelines for skin cancer classification. A-Asymmetry, B-Border, C-Color, D-Diameter, E-Evolution. It is also notable that the A,B,C, & D also work well as identifiable characteristics of an image using CNN's. Utilizing CNN's to identify Skin cancer shows great potential for advancing the treatment of skin cancer patients in terms of classification accuracy and by lessening the amount of initial patients for dermatologists.

Methods
CNN Description
CNN's have shown excellent accuracy in image classification. A CNN utilizes a series of convolutional layers with filters to analyze and extract data from an image. The CNN then produces tensors as outputs. These tensors, or outputs, are then processed by another convolutional layer or a pooling layer. Before entering the pooling layer, but after convolution, the image units are processed by an ReLU layer. This ReLU layer rectifies the linear units, a process that removes all negative values from the extraction process and converts them to zero.[4] The pooling layer is used to compress the size of the image, in this case, taking the average or maximum value, for the size of the pooling filter. The pooling filter then moves along based on the step size, or stride, until the entire image has been pooled and compressed into a new output. This combination of convolution and pooling can be done repeatedly, creating deep stacks, with different sequences, different filter sizes, and different strides for pooling before reaching a fully connected layer before output. Once the process has reached the fully connected layer, a series of values is used to create a prediction for the image. The CNN is trained using 757 disease classes from the taxonomy consisting of 2,032 diseases. Each of these values will have different weights depending on the images that are input into the process. This process is done by Backpropagation. Backpropagation utilizes gradient descent to minimize error. This will determine the features and weights applied to each value for the prediction. In finding this

minimum, the network is also optimizing performance. The last step before output is the softmax function. In this function, Softmax takes in the vector that represents every value from the last pooling layer and normalizes it to a probability distribution.

GoogLeNet Inception v3 and training via transfer learning

The GoogLeNet Inception v3 architecture utilizes the same structure of a CNN, but with an important variation. In the Inception v3 architecture, instead of specifying the CNN to utilize a 1x1 convolution layer, 3x3 convolution layer, 5x5 convolution layer, or a pooling layer, GoogleLeNet's v3 architecture utilizes all of the choices[5][6]. This approach results in a more dynamic network with better results and performance. GoogLeNet's Inception v3 comes pre-trained on 1.28 million images and achieves 93.33% accuracy among 1000 object classes in the 2014 ImageNet Large Scale Visual Recognition Challenge. This neural network was trained for more than 2 weeks using 8 tesla GPU's and is therefore not practically trainable on a standard desktop PC. However, through transfer learning using 129,450 images, 3,374 of which are dermoscopy images, this network was trained for use on dermatological images. 127,463 images were used to train the network and 1,942 images consisting of biopsy labeled test images were used to test the networks accuracy. This transfer learning is done by removing the last layer of the v3 Inception of the architecture, which is the fully connected layer, and training it on the aforementioned set of images. Backpropagation is then used to train this network for use on the dataset.

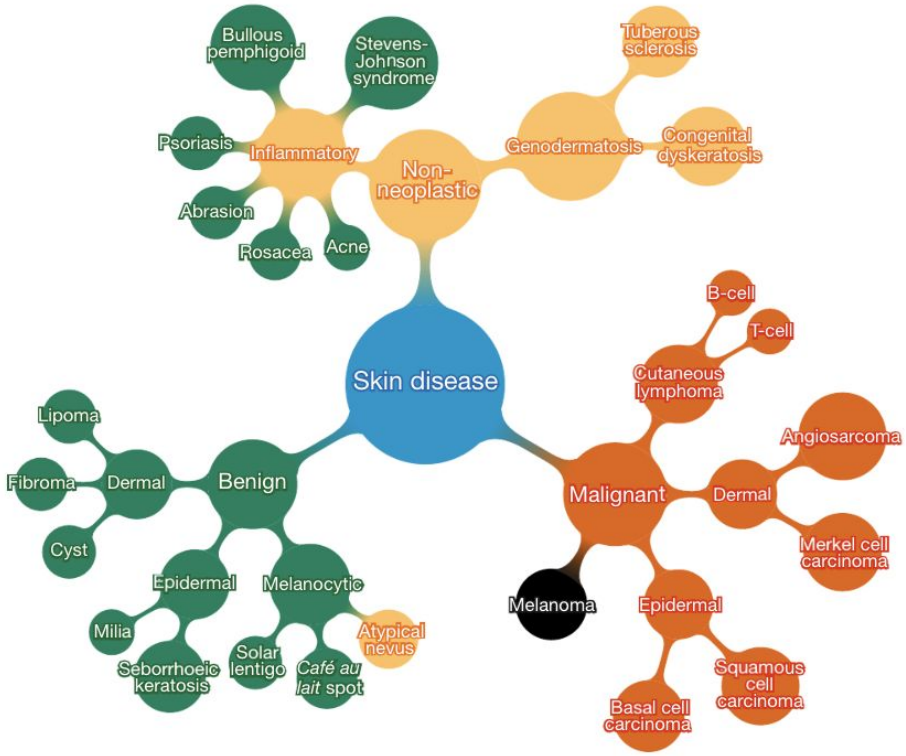
Taxonomy

The taxonomy represents 2,032 diseases arranged in a tree structure with first level nodes representing benign lesions, malignant lesions, and non-neoplastic lesions. This taxonomy was built from the bottom up by dermatologists with individual disease representing leaf nodes. These nodes were merged according to clinical and visual similarity. This taxonomy is both useful in classification and is medically relevant as the leaf nodes represent diseases that have similar clinical treatment plans.

Disease Partitioning Algorithm

The partitioning algorithm used to partition diseases into training classes is a recursive algorithm designed specifically with the taxonomy in mind. The algorithm leverages the taxonomy to create training classes with diseases that are visually similar and also have similar clinical treatment processes. The algorithm prevents training classes that are overly fine grained and lack the information to be trained properly. It also prevents generating classes that are too coarse and create a bias. The algorithms only hyperparameter, maxClassSize, initialized to be 1000 creates a disease partition of 757 classes.

Figure 1. Taxonomy Illustration



Datasets

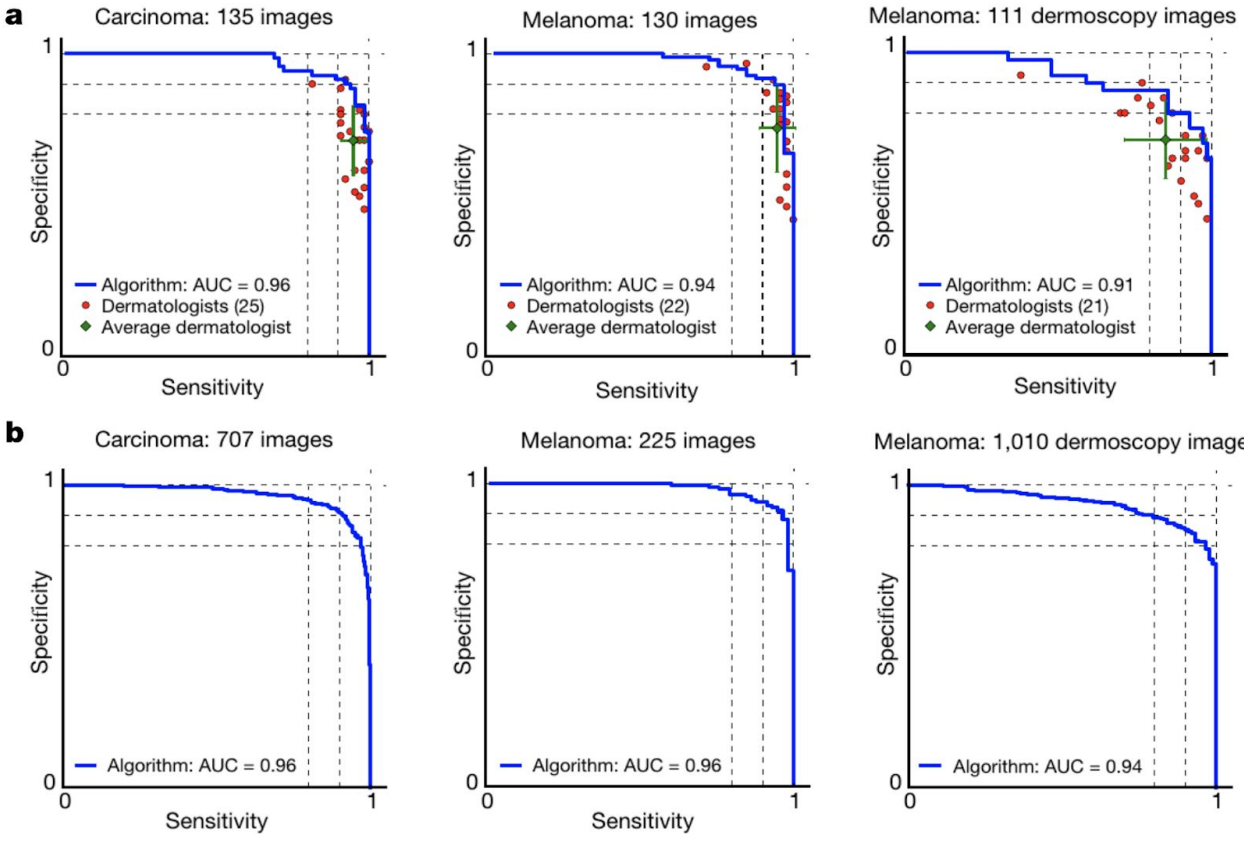
Datasets from Stanford hospital, the University of Edinburgh Dermofit Image Library and the ISIC Dermoscopic Archive were used in training the CNN. The dataset of images is more than two orders of magnitude larger than any previous dataset used to train a model for skin cancer classification. Blurred images and long range images were excluded for validation sets, but were used in training. All images were of 299x299 pixels for compatibility with the GoogLeNet CNN, although larger pictures were used for dermatologist classification.

Results

The algorithm for the CNN was validated using two methods. The first is a 3 class disease partition utilizing the first level nodes of the taxonomy. These nodes consist of benign lesions, malignant lesions, and non-neoplastic lesions. In this validation method, the CNN achieves 72.1% overall accuracy with an margin of error ± 0.9%. Two dermatologists who performed the same task on a subset of the validation achieved 65.56% and 66.0% overall accuracy. The second validation method is a nine class disease partition. These nine classes are the children of the first level nodes and represent diseases that are visually similar and that also have similar treatment plans. In this task the CNN achieves 55.4% overall accuracy with an margin of error ± 1.7%. The two dermatologist achieved 53.5% and 55.0% overall accuracy respectively. These images are only labeled by dermatologists and not confirmed by biopsy, so the results are not conclusive. However, it does show that the CNN is learning from the training dataset. In order to

conclusively validate the the algorithm, further testing was conducted to distinguish between: keratinocyte carcinoma or benign seborrheic keratoses; malignant melanoma or benign nevus using standard images; and malignant melanoma or benign nevus using dermoscopic images. This testing was done using only biopsy-proven images. The metrics used for comparison are sensitivity and specificity. Sensitivity is defined and the amount of correctly identified positive predictions divided by the total number of positive cases of cancer. Specificity is defined as the number of correctly identified negative predictions divided by the total number of negative cases of cancer. By defining the prediction as $\hat{y} = P \geq t$ where t is a threshold probability it is possible to compute the sensitivity and specificity. Varying t from 0-1 generates a curve of sensitivities and specificities that are achievable from this CNN. The area under the curve (AUC) is used to measure the performance. With a maximum value of 1 (specificity maximum value of 1 times sensitivity maximum value of 1), the CNN achieves an AUC score of 0.96 classifying keratinocyte carcinoma or benign seborrheic keratoses. In classifying malignant melanoma or benign nevus using standard images, the CNN achieves a 0.94 AUC. In the last case, classifying malignant melanoma or benign nevus using dermoscopic images the AUC is 0.91. These results are compared against 21 board certified dermatologists. Fig. 2a below shows the blue line as a AUC, the red points as the dermatologists performance, and the green points as the average of the dermatologists. The CNN outperformed dermatologists in any case where the red point falls below the blue curve. Fig 2b show the sames test on the entire testing dataset. The changes seen are less than 0.03 when compared to the previous test results in Fig. 2a. This validates the reliability of the results on larger datasets.

Figure 2 AUC curve vs dermatologist performance for 3 test cases



Conclusion/Discussion

The use of CNN's for skin cancer classification is emerging as a very powerful tool in modern scientific medicine. The results achieved by this CNN have proven to be on par or better than 21 board certified dermatologists. A variety of important factors contributed to the successful results. The development of the taxonomy and disease partitioning algorithm for the process was a crucial step, as they classify images both visually and clinically, similar to the manner in which it would occur in a dermatologist's office. The ability to implement transfer learning and modify only a small section of the Inception v3 CNN allows for time and cost efficient training of the model. The size of the dataset used to train shows the robustness of the model. With the technology of CNN's and the ability of transfer learning, this is the first step in improving access to time sensitive, potentially life saving skin cancer diagnoses. The use of CNN's in the scientific community can also be adapted to any medical specialty utilizing images in disease classification such as radiology and pathology.

References

1. Skin cancer. (n.d.). Retrieved December 10, 2018, from https://www.aad.org/media/stats/conditions/skin-cancer
2. Key Statistics for Melanoma Skin Cancer. (n.d.). Retrieved December 10, 2018, from https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html
3. Kimball, A. B., & Resneck, J. S. (2008). The US dermatology workforce: A specialty remains in shortage. *Journal of the American Academy of Dermatology,59*(5), 741-745. doi:10.1016/j.jaad.2008.06.037
4. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2015.7298594
5. Advanced Guide to Inception v3 on Cloud TPU | Cloud TPU | Google Cloud. (n.d.). Retrieved December 10, 2018, from https://cloud.google.com/tpu/docs/inception-v3-advanced
6. Vanhoucke, Vincent, Sergey, Jonathon, & Zbigniew. (2015, December 11). Rethinking the Inception Architecture for Computer Vision. Retrieved from https://arxiv.org/abs/1512.00567
7. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, International Journal of Science,542*, 115-118. Retrieved December 10, 2018.