CS 230 – Final Project Report
Daniel Jun

Pruned Neural Networks for Skin Cancer Detection

https://github.com/djdude36/Pruned-Neural-Networks-for-Skin-Cancer-Detection

Introduction

Convolution neural networks (CNN's) can provide more accessible healthcare to underserved regions. One possibility is a model that can detect skin cancer from pictures taken by a smartphone. Skin cancer is the most common type of cancer in the United States with about 5.4 million new cases every year [1, 2]. The most dangerous are melanomas that account for approximately 75% of all skin-cancer-related deaths with over 10,000 deaths annually in the United States. Early detection and treatment are critical as the estimated 5-year survival rate for melanoma drops from over 99% if detected in its earliest stages to about 14% in its latest stages [3]. However, according to IMS Health, there are only 9,600 dermatologists and 7,800 dermatology practices to serve 323 million people in the United States [4]. Therefore, a skin cancer detection system would be valuable, especially to those who do not have easy access to dermatologists.

Related Work

In 2017, Esteva *et al.* [5] presented a CNN classifier that performed as well as dermatologists in identifying malignancies from skin lesion images. The model used a pre-trained Inception v3 CNN that was fine-tuned on 129,450 skin lesion images with 757 training classes. The model was then tested on two binary classification tasks: keratinocyte carcinomas (most common cancer) versus benign seborrheic keratoses; and malignant melanomas (deadliest skin cancer) versus benign nevi. The area under the curve (AUC) for both tasks was around 0.95 and the model outperformed the average of 21 dermatologists, demonstrating the effectiveness of deep learning in healthcare.

Other studies showed similar results in comparing CNNs to dermatologists in classifying skin lesions. Brinker *et al.* [6] used ResNet50 and trained on 12,378 open-source dermatoscopic images to outperform 136 dermatologists. Han *et al.* [7] used ResNet152 to classify 12 skin diseases from datasets with different patient demographics, Asian and Caucasian. The model was found to be sensitive to demographics (skin color), but the performance was still comparable to that of 16 dermatologists.

Model Pruning

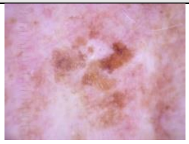
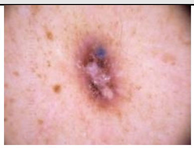
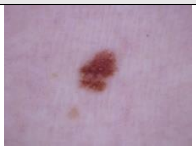
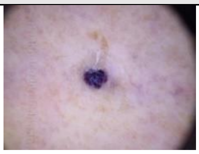
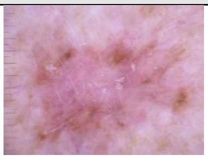
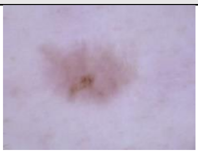
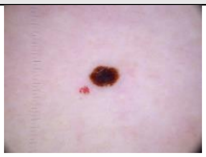
CNN's, however, have non-trivial requirements that may not be available in resource limited regions. The model Inception v3 contains 21 million (M), ResNet50 23M, and ResNet152 58M parameters. These models usually require large memory and GPU to train and deploy. Because of their size and computation requirements, they may not be effective on mobile devices, such as smart phones.

To make models more efficient, they can be pruned to remove redundant model parameters that are not sensitive to the model's performance. In Han *et al.* [8] pre-trained networks were pruned of low-weight connections to create more sparse models. The models were then retrained to learn the final weights on the remaining sparse connections. AlexNet was pruned from 61M to 6.7M parameters and VGG-16 from 138M to 10.3M parameters while maintaining predictive performance.

This project explored using the CNN's Inception v3, ResNet50 v2, ResNet152 v2, and the smaller Mobile Net v2 on the open-source HAM10000 dataset [9] to classify skin lesions as benign or malignant. The models were pruned using Keras [10] to create more sparse models that may be easier to deploy on mobile devices for use in resource-limited regions.

Dataset and Features

The data used in this project was from the HAM10000 dataset [9], which consists of 10,015 dermatoscopic images of common pigmented skin lesions. There were seven classes that include a representative collection of all important diagnostic categories in pigmented lesions: actinic keratoses and intraepithelial carcinoma / Bowen's disease, basal cell carcinoma, benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses), dermatofibroma, melanoma, melanocytic nevi, and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage). The seven classes can be grouped as benign or malignant as shown below with sample images.

<table>
<tr><th colspan="12">Benign</th></tr>
<tr>
<td colspan="3">
Benign keratosis-like lesions</td>
<td colspan="3">
Dermatofibroma</td>
<td colspan="3">
Melanocytic nevi</td>
<td colspan="3">
Vascular lesions</td>
</tr>
<tr><th colspan="12">Malignant</th></tr>
<tr>
<td colspan="4">
Actinic keratoses</td>
<td colspan="4">
Basal cell carcinoma</td>
<td colspan="4">
Melanoma</td>
</tr>
</table>

The data was imbalanced with only 20% (1,954) being malignant and the rest benign. The data was manually balanced by selecting all 1,954 malignant images and 1,954 random benign images for a total dataset of 3,908 images. The data was then split 60%, 20%, 20% into train, validation, and test sets with batch size 16 and resized to two resolutions: 224 x 224 pixels to be compatible with Mobile Net v2, ResNet50 v2, and ResNet152 v2 and 299 x 299 pixels to be compatible with Inception v3.

Methods

Since our skin lesion dataset was small, models that have been pre-trained on ImageNet with 1.28 million images for 1,000 object classes were used [11]. Specifically, Inception v3, ResNet50 v2, and ResNet152 v2 were tested as they were used in other studies to classify skin lesions. Mobile Net v2 [12] was also included to evaluate the performance of a much smaller model in comparison to the other models that have at least 9 times more parameters. For the pre-trained models, the final classification layer was replaced with a binary classification layer with a sigmoid activation function to output probabilities for our binary classes, benign and malignant. The Adam optimizer was used for gradient descent to update model parameters during training.

To maximize the performance of the models on classifying skin lesions, two factors were experimented with: the number of layers to freeze with preset ImageNet parameters and the learning rate. With small datasets, it is helpful to use transfer learning and extract parameters that have been pre-trained on a larger dataset [13]. However, if the dataset is very different, the parameters would also need to be trained on the target dataset to reduce bias. This concept was explored by freezing different layers and changing the number of trainable parameters for Mobile Net v2 and Inception v3 models. Different learning rates were also tested for all four models to obtain the best performance.

Performance was measured by evaluating the AUC on the test set. The AUC is the area under the receiver operating characteristic (ROC) curve and it indicates how good the model is in distinguishing between classes. Generally, the larger the AUC is for a model, the better the model's performance is across different threshold levels (which trades off sensitivity and specificity) along the ROC curve. This also makes the AUC a more reliable metric than accuracy, as the latter can be deceiving with imbalanced data.

After high prediction performance was achieved, the models were pruned at sparsity 25%, 50%, 75% and 90%. The models were then retrained on the training set so the remaining parameters can fine-tune their weights to the new sparse network. The models' performances were then evaluated again against the test set.

Google's Tensorflow v1.1.5 deep learning framework was used to train, validate, and test the models with Google Cloud and NVIDIA Tesla P1000 GPU.

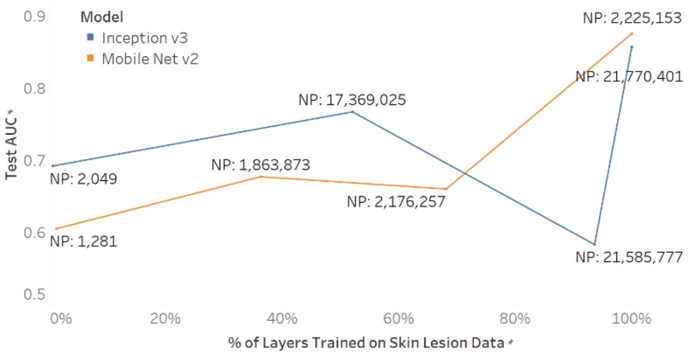
Experiments/Results/Discussion

Freezing Layers

Details of the four models used to classify skin lesions are shown in the table below.

Model	# of Layers	Trainable Parameters	Non-trainable Parameters	Total Parameters
Mobile Net v2	157	2,225,153	34,112	2,259,265
Inception v3	313	21,770,401	34,432	21,804,833
ResNet50 v2	192	23,521,409	45,440	23,566,849
ResNet152 v2	566	58,189,953	143,744	58,333,697

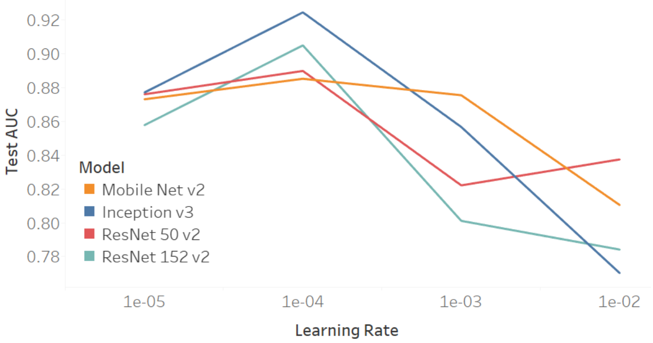
These models were initially pre-trained on the ImageNet dataset and contained preset, learned parameters. However, as our skin lesion images were very different from those of ImageNet, various degrees of freezing layers (i.e., keeping the preset, ImageNet-learned parameters) were tested to determine the number of layers that should be trained on our skin lesion dataset. Different layers of Mobile Net v2 and Inception v3 were frozen from the bottom-up and the figure on the right shows the % of layers and number of parameters (NP) trained on the skin lesion data and the resulting test AUC scores.



When only the top classification layer was trained (~1% of layers), the test AUC's were relatively low for both models. As more layers were unfrozen and made trainable, the AUC's increased, then decreased, and then increased again to maximum AUC's of 0.876 and 0.857 for Mobile Net v2 and Inception v3, respectively, when all layers were trained on the skin lesion data.

This trend in AUC highlights the large difference between our skin lesion images and those of ImageNet as well as the complexity of the Mobile Net v2 and Inception v3 models. To reduce bias error on our task, it is more beneficial to not use the preset ImageNet weights, but to train all the parameters on the skin lesion data. In addition, the dip in AUC's at 94% and 68% of layers trained for Inception and Mobile Net v2, respectively, may be due to the models' architecture and the interdependencies of the lower layers. Since lower layers learn low-level features like colors and edges, "interrupting" lower layers by having some trained on ImageNet and others trained on skin lesions may disrupt the learning process and result in lower performance.

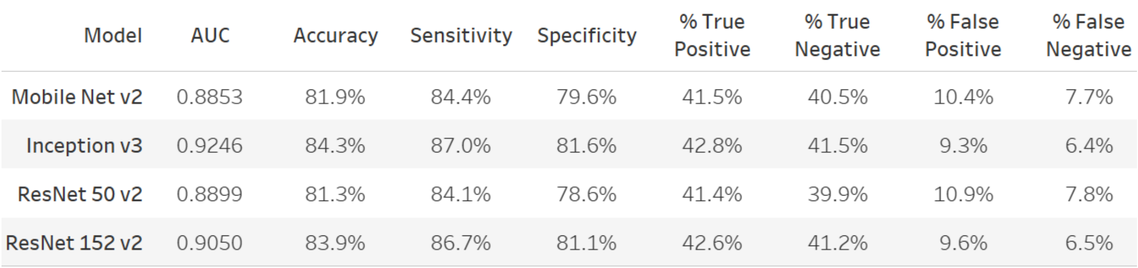
Learning Rates



The default learning rate of 1e-03 was used to test various degrees of freezing layers, but other learning rates were also tested for all four models (with all layers trained) as shown in the right figure. The figure highlights the importance of learning rate on performance. At the default learning rate of 1e-03, Mobile Net v2 had the best test AUC at 0.876 and ResNet152 v2 the worst at 0.801. Decreasing the learning rate to 1e-04, however, resulted with Mobile Net v2 being the worst performer with an AUC of 0.883 and ResNet152 v2 being the second best with 0.905, a 13% increase. Learning rate 1e-04 provided the maximum AUC for all models with Inception v3 having the highest at 0.925. Smaller learning rate 1e-05 may have taken too long to converge on an optima while larger learning rates 1e-03 and 1e-02 may have been too big and skipped over optimas.

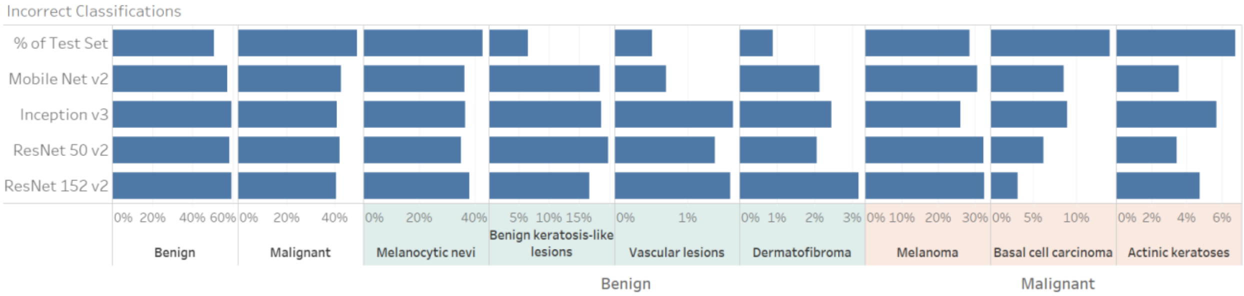
Test Evaluation

The four models had the best performance when all layers were trained on the skin lesion data with learning rate 1e-04. A summary of the models' performance on the test set and the percent of test data that were classified as true positive, true negative, false positive, and false negative is shown in the table below. Since this task involves classifying skin lesions as benign or malignant with high risks in missing positive cases (false negative), we are interested in high sensitivity and few false positives. All four models reported sensitivity to be higher than both accuracy and specificity, with Inception v3 having the highest sensitivity at 87.0%. In addition, of the incorrect classifications made, a majority of them were false positive than false negative; most errors were in classifying skin lesions as malignant when in actuality they were benign, which is the lower risk error.

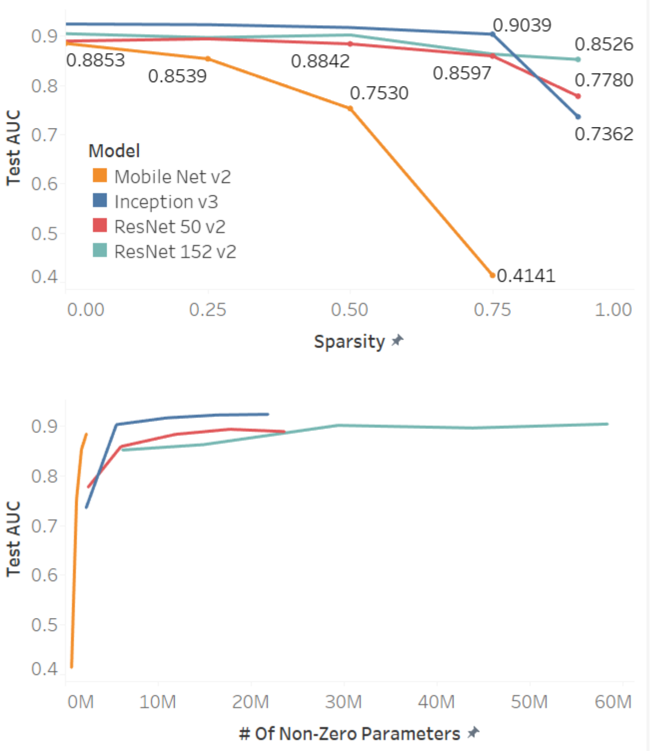


Model	AUC	Accuracy	Sensitivity	Specificity	% True Positive	% True Negative	% False Positive	% False Negative
Mobile Net v2	0.8853	81.9%	84.4%	79.6%	41.5%	40.5%	10.4%	7.7%
Inception v3	0.9246	84.3%	87.0%	81.6%	42.8%	41.5%	9.3%	6.4%
ResNet 50 v2	0.8899	81.3%	84.1%	78.6%	41.4%	39.9%	10.9%	7.8%
ResNet 152 v2	0.9050	83.9%	86.7%	81.1%	42.6%	41.2%	9.6%	6.5%

A further breakdown of the incorrect classifications is shown in the figure below. The top row indicates the percent of the test set that belongs to a specific class, such as 51% of the test set is benign and 28% is melanoma. The other rows show that of the incorrect classifications made by a model, the percent that belongs to a specific class. This figure allows us to see if there are certain classes that are unproportionally misclassified by the models. As previously mentioned, more benign than malignant examples are misclassified with unproportionally high errors with benign keratosis-like lesions. Of the malignant examples, the models appear to have unproportionally low errors with basal cell carcinoma, indicating that the models may be better at detecting this type of malignant skin lesion than melanoma or actinic keratoses.



Pruning



After high performance was achieved, the models were pruned at sparsity levels 0.25, 0.50, 0.75, and 0.90. The right figure shows the test AUC performance against sparsity (top) and number of non-zero parameters (bottom).

As sparsity increased (and number of non-zero parameters decreased), the test AUC remained relatively stable for the larger models. For ResNet50 v2 and ResNet152 v2, the performance actually peaked at sparsity 0.25 and 0.50, respectively. This highlights the redundancies of parameters in large networks (as shown in Han *et al.* [8]) and their tendency to overfit on the training and/or validation data.

Increasing sparsity on the smaller Mobile Net v2, however, significantly decreased its performance. This shows that Mobile Net v2 is an already compact and efficient model with small redundancy in parameters. In addition, when the larger models are made sparse to have similar number of non-zero parameters as Mobile Net v2 (around 2M), Mobile Net v2's performance is much better than those of Inception v3 and ResNet50 v2. Therefore, in terms of making the most efficient CNN's, it may be more useful to create an efficient CNN from scratch as opposed to starting with a large network and then pruning it to a smaller network.

Conclusion/Future Work

Four pre-trained CNN's were explored to classify skin lesions as benign or malignant from the open-source HAM10000 dataset. The models Mobile Net v2, Inception v3, ResNet50 v2, and ResNet152 v2 were evaluated at different degrees of freezing layers and learning rates to find the best performance. The models were then pruned at different sparsity levels and their performance were reevaluated. For the large models Inception v3, ResNet50 v2, and ResNet152 v2, the performance were retained even at 0.75 sparsity. For the already small model Mobile Net v2, performance decreased quickly with sparsity. However, Mobile Net v2 at 0 sparsity still had fewer number of parameters and better performance than the most pruned versions of the larger models. For future work, it would be interesting to apply quantization on the pruned models to compress their size even more.

Contributions

Daniel Jun is the sole contributor to this project.

References

1. G. Guy, et. al., "Vital signs: Melanoma incidence and mortality trends and projections—United States, 1982–2030," MMWR Morb Mortal Wkly Rep, vol. 64, no. 21, pp. 591-596, 2015.
2. Rogers, H. W. et al. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012. JAMA Dermatology 151.10, 1081–1086 (2015).
3. American Cancer Society. *Cancer Facts & Figures 2020*. Atlanta, Ga: American Cancer Society; 2020.
4. M.P. Seraly, "A new patient management service for dermatologists," [Online]. Available: http://iagnosis.com/sites/default/files/DermatologistOnCallWhitepaper.pdf.
5. Esteva, Andre, et al. "Dermatologist-level classification of skin cancer." *Nature* (2016).
6. Brinker, Titus J., et al. "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task." *European Journal of Cancer* 113 (2019): 47-54.
7. Han, Seung Seog, et al. "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm." *Journal of Investigative Dermatology* 138.7 (2018): 1529-1538.
8. Han, Song, et al. "Learning both weights and connections for efficient neural network." *Advances in neural information processing systems*. 2015.
9. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi:10.1038/sdata.2018.161 (2018).
10. "Pruning in Keras Example." Pruning in Keras Example, Tensorflow, www.tensorflow.org/model_optimization/guide/pruning/pruning_with_keras.
11. Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115.3 (2015): 211-252.
12. Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).
13. "Transfer Learning with a Pretrained ConvNet: TensorFlow Core." TensorFlow, www.tensorflow.org/tutorials/images/transfer_learning#create_the_base_model_from_the_pre-trained_convnets.