UNIVERSITY OF CALIFORNIA

Los Angeles

An Application of Split Attention Networks: Melanoma Detection

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Andrew Mashhadi

2023

ABSTRACT OF THE THESIS

An Application of Split Attention Networks:
Melanoma Detection

by

Andrew Mashhadi
Master of Applied Statistics
University of California, Los Angeles, 2023
Professor Yingnian Wu, Chair

(Abstract temporarily omitted)

The thesis of Andrew Mashhadi is approved.

Michael Tsiang

Frederic Paik Schoenberg

Yingnian Wu, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Background

In the past decade, machine learning has exploded in popularity. Machine learning methods have demonstrated endless applciations to a variety of industries including but not limited to engineering, science, finance, medicine, and technology. A large branch of machine learning that has recently taken the world by storm is *deep learning*. Deep learning is made up of *artificial neural networks* (ANN) and is generally trained with a form of *feature learning*.

Neural network architures have evolved and expanded since the original perceptron and ANN models. Through different areas of application, variations in neural network architecture such as *deep neural networks*, *convolutional neural networks*, *recurrent neural networks*, and more recently *transformers*, have all been proposed and adopted. It is no secret that the transformer model has quickly become a front-runner for applications in computer vision due to its successes in natural language processing. However, deep convolutional neural networks are still state-of-the-art for tasks such as image classification, object detection, semantic segmentation, and instance segmentation

Different learning methods have also been adopted. However, the type of learning method used generally depends on the model's particular objective. *Supervised Learning* is one of the most common methods for classification or regression models. It is the process of using labeled datasets to train machine learning models to classify or predict outcomes appropriately. *Unsupervised Learning* generally involves the analysis or clustering of unlabeled datsets. And *Reinforcement Learning* is based on rewarding desired behaviors and/or punishing undesired ones.

More recently, the medical community has been opening it's doors to modern deep learning techniques. In particular, convolutional neural networks (CNNs) have often been used to develop more efficient, and accurate, diagnostic tools to analyze medical images. Due to the growth of effective image recognition models, collection of medical images for specific applciations in the healthcare community have been been growing.

1.2 Problem Statement

Skin cancer is one of the most common types of cancer. Although melanoma only accounts for about 1% of skin cancer, the death rate was still about 2.1 per 100,000 men and women per year based on 2016-2020 deaths [Ins23]. In 2023, the American Cancer Society estimates that bout 7,990 people are expected to die from a total of about 97,610 new melanoma cases in the United States alone [Soc23]. It is well known that early detection of melanoma will provide the best chance for successful treatment and greater chance of survival.

Image analysis tools that automate the diagnosis of melanoma will improve dermatologist's diagnostic accuracy, and better detection of melanoma has the opportunity to positively impact millions of people. Providing an accurate machine learning model to aid dermatologist's in their evaluations of patients moles may lead to an earlier diagnoses, and could therefore provide the best chance for appropriate intervention. We want to use a new deep convolutional-neural-network architecture that uses patients skin-lesion images along with any other patient-level contextual information to determine which patients are likely to have melanoma skin cancer. In particular, the goal of this paper is to identify the presence of melanoma from images of skin lesions using the novel *ResNeSt* architecture that was designed to apply channel-wise attention on different network branches.

CHAPTER 2

Dataset

2.1 ISIC 2020 Challenge Dataset

For this paper, we used the "ISIC 2020 Challenge Dataset". This was the official dataset of the SIIM-ISIC Melanoma Classification Challenge hosted as a Kaggle sponsored competition in the Summer of 2020. It contains over 30,000 dermoscopic images of distinct skin lesions from approximately 2,000 patients. Each image has an associated record of metadata. The metadata includes the corresponding "beniegn" or "malignant" status, as well as the following patient-level features:

- *patient_id*: unique patient idenitfier
- *sex*: sex of the patient
- *age_approx*: approximate age of the of the patient
- *anatom_site_general_challenge*: the general location of imaged lesion
- *diagnosis*: additional details regarding the diagnosis

We should note that all associated malignant and benign diagnoses have been confimed using histopathology, expert agreement, or longitudinal follow-up [ISI20].

The International Skin Imaging Collaboration (ISIC) was responsible for compiling the images from the Hospital Clínic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, University of Queensland, and the University of Athens Medical School to form this official dataset. The ISIC Archive contains

the largest collection of quality-controlled dermoscopic images of skin lesions available to the public.

The resolution of each image varied drastically throughout the dataset, with some images reaching as high as 4000×6000 pixels. The set of images consisted of over 110GB, so we hosted the data on UCLA's Hoffman2 Linux Cluster. To support the size of each image and the large number of individual images within the dataset, we used computing resources from the Hoffman2 cluster and Google Colab to tune, train, and test our models with powerful *Graphical Processing Units* (GPUs). Although the GPU would occasionally change, we mostly used a single NVIDIA A100 GPU with approximately 40GB VRAM to train our larger models.

2.2 Data Preparation

Throughout training, we found it to be difficult to train our models without batch-sizes greater than or equal to 16 images. However, computational constraints limited our ability to utilize larger batch-sizes with such large lesion images. We ultimately found that the only way to train efficiently with suffcient batch-size ($B \geq 16$) was to center crop, or resize (using bilinear interpolation), each original lesion image to a fixed-sized 512×512 image. Further transformations and resizing are discussed in the methodology section.

The patient-level features within the metadata also required minor preprocessing. We created dummy variables with treatment coding (also known as *one-hot encoding*) for the *sex* and *anatom_site_general_challenge* features, while keeping *age_approx* as a continuous variable.

2.3 Exploratory Data Analysis

In this section, we provide an exploratory data analysis of both the images and the patient-level features within the metadata provided in our dataset.

2.3.1 Skin Lesion Images

The images of skin lesions are the most important part of data for our model. They will be used to train the image classification network, so we decided to investigate the data a bit more, before beginning to model. First, we investigated the proportion of "positives" (malignant) and "negative" (beniegn) images within our dataset. Table 2.1 presents the balance of malignant and beniegn skin lesions within our dataset. As we can see from the table, there is significantly large class imbalance. Only about 1.8% of the skin lesion image are classified as malignant.

	Beniegn (−)	Malignant (+)
Count	32542	584
Proportion	0.9824	0.0176

Table 2.1: Proportion of Beniegn & Malignant Skin Lesions

As mentioned above, we fixed the resolution of the images in our dataset to 512x512 pixels. We visually inspected the quality of several randomly selected skin lesions images after cropping and resizing to make sure that the *object of interest* (the actually skin lesion itself) was still present and easily distinguishable. In Figures 2.1-2.2 , we present 6 images from our dataset. The top 3 skin lesions were randomly selected from the images flagged as malignant, while the bottom 3 skin lesisons were randomly selected from the images marked as beniegn. Although the relative and absolute size of the skin lesions differ, they are all clearly distinguishable in the images.

We also visually compared a random collection of images flagged as malignant with another random collection of images flagged as benign. The objective here was to see if our "untrained" eye could easily classfiy the skin lessions as malignant or beniegn. We can see from Figures 2.1-2.2 that without proper training and eductaion, it is very difficult to determine whether or not a skin lesion is considered malignant. Even with the proper training, it can still be difficult to accurately determine if melanoma is present. According to the Skin Cancer Foundation, features such as assymmetry shapes, irregular borders, uneven

Figure 2.1: Examples of Malignant Skin Lesions

Figure 2.2: Examples of Beniegn Skin Lesions

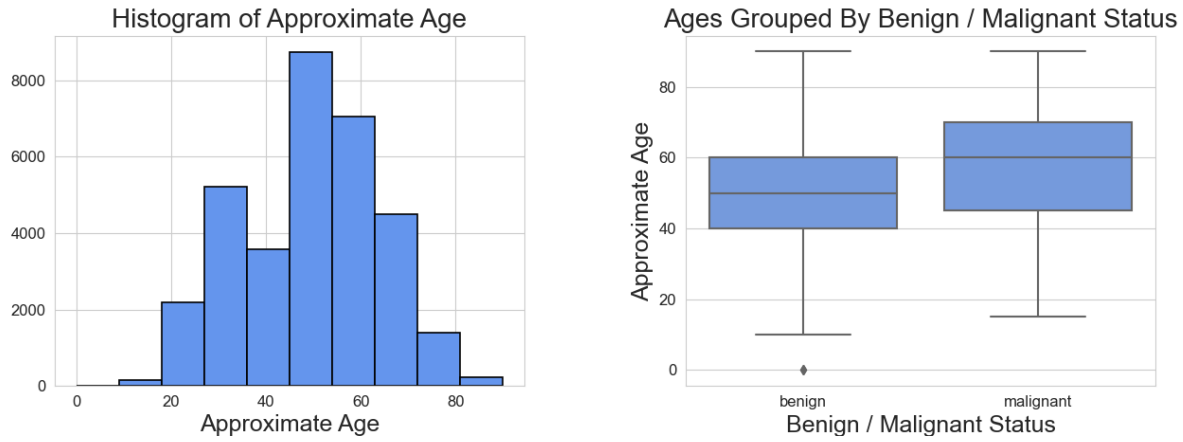
distribution of color, and large relative size may indicate the presence of melanoma (or other skin cancers) [Fou23]. Ideally, our network will pick up on these characteristics while training.

2.3.2 Patient-Level Features

In this section we explore the one-way frequency tables for the approximate ages, patient sexes, and the locations of imaged lesions. Then we explored the two-way frequency tables for each of these variables with the response variable (malignant or beniegn).

Patient Age

We can see from the histogram of approximate ages in Figure 2.3(a) that most of the patients in this data are in their 40's, 50's, and 60's. The box-plot in Figure 2.3(b) clearly indicates that older patients are generally associated with a larger number of malignant skin lesions.



(a) (b)

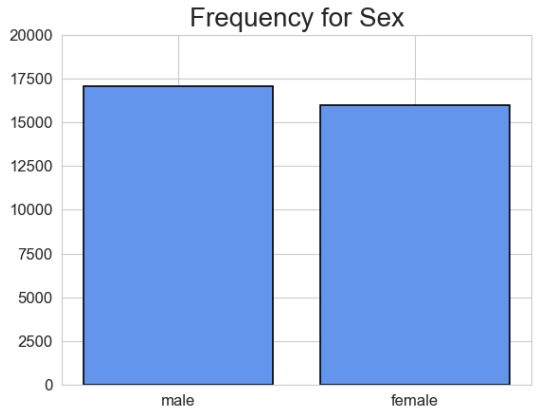
Figure 2.3: Histogram and Box-Plot For Appoximate Age

Patient Sex

We can see from the frequencies shown in Figure 2.4(a) that there are a similar number of males and females in this dataset, with just one or two thousand more males. The two-way contingency table between the response variable and patient sex is also shown in Figure 2.4(b). We can see that the malignant cases have a significantly larger proportion of males than the beniegn cases do. In fact, the χ^2-test of independence reported a p-value very close to 0. Therefore, we reject the hypothesis that patient sex and the presence of melanoma are independent, and conclude that there is a relationship between the two variables.

Skin-Lesion Location

There were 6 categories for the general skin-lesion location: oral/genital, palms/soles, head/neck, upper extremity, lower extremity, and torso. Figure 2.5(a) presents the associated frequen-



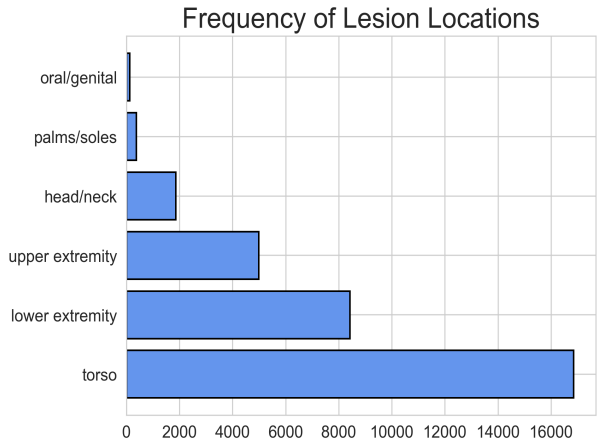
(a)

	Beniegn	Malignant
Female	15761	220
Male	16716	364

(b)

Figure 2.4: Frequency Plot and Two-way Contingency Table for Patient Sex

cies within our dataset. None of the general locations have a similar number of counts, and almost half of the skin-lesions come from the torso region. We also present the two-way contingency table between the response variable and lesion location in Figure 2.5(b). While it is somewhat difficult to tell, the does appear to be some differences in the balance between beniegn and malignant cases when conditioned on the skin-lesion location. In this case, the χ^2-test of independence also reported a p-value very close to 0, indicating that there is a relationship between the location of the skin-lesion and the presence of melanoma.



(a)

	Beniegn	Malignant
Head/Neck	1781	74
Lower Extremity	8293	124
Oral/Genital	120	4
Palms/Soles	370	5
Torso	16588	257
Upper Extremity	4872	111

(b)

Figure 2.5: Frequency Plot and Two-way Contingency Table for Lesion Location

CHAPTER 3

Methodology

In this section, we describe the methods and models used to best predict the presence of melanoma. As mentioned above, we used deep convolutional neural networks to generate classifications. More specifically, the latest variant of the residual network, the *ResNeSt*, was used to assess whether or not its channel-wise attention architecture would find further success in melanoma detection. Additionally, we trained a smaller, more traditional, convolutional neural network to compare the performance with the split-attention network.

To make use of the additional metadata features within our dataset, we simultaneously trained a standard *multi-layer perceptron* (MLP) with inputs from our patient-level features. Therefore, the deep convolutional networks is used to effectively extract any contextualized features from the skin-lesion images and the multi-layer perceptron is used to extract any important patient-level information. The two networks are then connected with eachother before generating a final probability. Since the outputted probabilities are a result from two seperate networks communicating with eachother, this model may be described as a *Multi-Network Ensemble*.

3.1 Training, Validation, and Testing Sets

Before discussing the actual models used, we breifly discuss the how the dataset was split and sampled for training, validating, and testing.

We first randomly split 80% of the data into a training set and used the other random 20% as the test set. The test set was not used by any network during the training process. Additionally, we partitioned another 20% of the training data into a validation set. Recall

that our dataset is largely imbalanced. As shown in the Exploratory Data Analysis section, only about 2% of the skin lesions in the dataset are flagged as malignant, while the remaining skin lesions are marked as benign. We ensured that each split of the data had a similar sample proportional of about 2% malignant skin lesions.

Since the response variable (malignant or beniegn) in the training data is heavily imbalanced, we used randomized oversampling of the malignant observations to combat any potential bias toward a "beniegn" classification. By oversampling, we hoped that the networks would better detect when skin lesions were truly malignant. It's important to note that since oversampling was used, we considered a single epoch as when the entire *oversampled* dataset was passed forward and backward through the network exactly once. Therefore, a single epoch will contain multiple repeated malignant skin lesion images.

Note that using standard k-fold cross validation would require a model to be trained k times for each set of hyper-parameters. Given the size of our data and our models, the time it would take to perform this procedure was simply infeasible. Therefore, we used the independent validation set to tune each network's hyper-parameters effectively. This method minimized overall training time, and produced optimal hyper-parameters. Once the hyper-parameters were obtained, we retrained the model on a new oversampled dataset made up from combining the training and validation sets.

We tuned each model with the area under the *Receiver Operating Characteristic* (ROC-AUC) as our performance metric. The ROC-AUC was chosen over other metrics because it summarizes how well our network separates the two response classes over *all* possible thresholds (as opposed to a single predetermined threshold).

3.2 Image Augmentations

Image augmentations were used as a preprocessing step for each image before training. Some image augmention methods are often used to alter the original images in the dataset to effectively create more "unseen" examples for the network to use while training. This technique artificially extends the dataset by randomly providing alerations to existing data. Given

the limited number of malignant examples, this technique is very important to artificially extend our malignant skin-lesion training images. In our networks, we randomly employed the following image augmentations (in this order) on the images in our training data:

1. **HueSaturationValue**: With probability $p = 0.5$, we randomly changed the hue, saturation, and value of the input image. The shift in hue was between $(-5, 5)$, the shift in saturaion was between $(-10, 10)$, and the shift in value was between $(-5, 5)$.

2. **VerticalFlip**: With probability $p = 0.5$, we veritially flipped the imput image. Note that vertically flipping the image should not change the true class of the skin lesion.

3. **HorizontalFlip**: With probability $p = 0.5$, we veritially flipped the imput image. Note that vertically flipping the image should not change the true class of the skin lesion.

4. **GaussNoise**: we applied gaussian noise to the input image. We did not want to corrupt the image too much and to ooften, so we used a mean of 0 and a variance between $(10, 50)$ with a low probability $p = 0.2$.

5. **ShiftScaleRotate**: With probability $p = 0.5$, we randomly applied the three affine transforms: translate, scale and rotate the input. The shift factor range for both height and width was set to $(-0.25, 0.25)$, the scaling factor range was set to $(-0.25, 0.25)$, and the rotation range was set from -30 degrees to 30 degrees.

6. **RandomBrightnessContrast**: With probability $p = 0.5$, we randomly modified the brightness and contrast of the input image. Since many of the images were already very dark, the factor range for changing brightness was set to $(0.9, 1.1)$. The factor range for changing contrast was also set to $(0.9, 1.1)$.

Some image augmentations techniques are also used to make sure the images in the dataset work with our models, and may promote faster convergence. For instance, in addition to the above augmentations, we also randomly cropped each 512×512 image to a 416x416 image and normalized the pixel values to have a similar distribution. Although the

randomness of the cropping does somewhat "artificially" extend our training set, the main point of cropping the images to 416×416 is because the ResNeSt network we used requires an input size of 416×416 pixels. The normalization of the pixel values will reduce the risk of exploding gradients, which has been shown to increase training time and generally slows down convergence. We normalized each channel using the associated sample means and sample standard deviations calculcated from the *ImageNet* database. The ImageNet database consists of millions of images and was designed for use in visual object recognition software research. Although we could have used the sample means and sample standard deviations from our own dataset, using the sample statistics from the ImageNet database is common practice and was recommended by the authors of the *ResNeSt* network.

We should note that most of the above image augmentations were only used on the training set. The validation and testing sets only center cropped the images to 416×416 and normalized the pixels to work with the trained network.

3.3 Network for Skin-Lesion Images

3.3.1 Convolutional Neural Network

Write here. Discuss final implementation of Batch Normalization, Adam Optimizer, L2 Regularization, and Dropout layers.

3.3.2 ResNeSt

Write about evolution from CNN to ResNet to ResNeSt. Also use the following description of ResNeSt somewher (from pytorch website):

While image classification models have recently continued to advance, most downstream applications such as object detection and semantic segmentation still employ ResNet variants as the backbone network due to their simple and modular structure. We present a simple and modular Split-Attention block that enables attention across feature-map groups. By stacking these Split-Attention blocks ResNet-style, we obtain a new ResNet variant which we

call ResNeSt. Our network preserves the overall ResNet structure to be used in downstream tasks straightforwardly without introducing additional computational costs. ResNeSt models outperform other networks with similar model complexities, and also help downstream tasks including object detection, instance segmentation and semantic segmentation.

Discuss final implementation of Batch Normalization, Adam Optimizer (different rates), L2 Regularization, and Dropout layers.

A New ResNet Variant, is modularized architecture is designed to apply the channel-wise attention on different network branchesresidual-network variant known as *ResNeSt*.

Disuss the use of the mentioned Image Augmentations and then discuss the idea of Batch Normalization, Adam Optimizer, L2 Regularization, and Dropout layers. Discuss the difference between starting the CNN weights at random values, vs starting the ResNeSt weights at the IMAGENET trained weights and finetuned for our set of data (effectively leveraging their training).

3.4 Network for Patient-Level Metadata

Give back-ground on Multi-Layer Perceptron for metadata.

CHAPTER 4

Results

4.1 CNN + MLP Ensemble Results

Write here. Give all plots from analysis notebook. Give table of statistics and confusion matrix.

4.2 ResNeSt + MLP Ensemble Results

Write here. Write here. Give all plots from analysis notebook. Give table of statistics and confusion matrix. Compare to regular CNN + MLP model.

CHAPTER 5

Conclusion and Future Work

Write here.

REFERENCES

[Fou23] Skin Cancer Foundation. "Melanoma Warning Signs.", 2023. Last Accessed: 2023-03-20.

[Ins23] National Cancer Institute. "Cancer Stat Facts: Melanoma of the Skin.", 2023. Last Accessed: 2023-03-20.

[ISI20] "The ISIC 2020 Challenge Dataset.", 2020. Last Accessed: 2023-02-13.

[Soc23] American Cancer Society. "Key Statistics for Melanoma Skin Cancer.", 2023. Last Accessed: 2023-03-20.