Capstone Project Proposal

by David Soto - dasoto@gmail.com

@Galvanize Data Science Immersive Program

CNN to identify malign moles on skin

1. Project Summary and motivation

The purpose of this project is to create a tool that considering the image of a mole, can calculate the probability that a mole can be malign.

Skin cancer is a common disease that affect a big amount of peoples. Some facts about skin cancer:

- Every year there are more new cases of skin cancer than the combined incidence of cancers of the breast, prostate, lung and colon.
- An estimated 87,110 new cases of invasive melanoma will be diagnosed in the U.S. in 2017.
- The estimated 5-year survival rate for patients whose melanoma is detected early is about 98 percent in the U.S. The survival rate falls to 62 percent when the disease reaches the lymph nodes, and 18 percent when the disease metastasizes to distant organs.

2. Development process and Data

The idea of this project is to construct a CNN model that can predict the probability that a specific mole can be malign.

2.1 Data:

To train this model I'm planning to use a set of images from the International Skin Imaging Collaboration: Mellanoma Project ISIC https://isic-archive.com.

The specific datasets to use are:

- ISIC*UDA-2*1: Moles and melanomas. Biopsy-confirmed melanocytic lesions. Both malignant and benign lesions are included.
    - Benign: 23
    - Malign: 37
- ISIC*UDA-1*1 Moles and melanomas. Biopsy-confirmed melanocytic lesions. Both malignant and

benign lesions are included.

  - Benign: 398
  - Malign: 159

- ISIC*MSK-2*1: Benign and malignant skin lesions. Biopsy-confirmed melanocytic and non-melanocytic lesions.

  - Benign: 1167 (Not used)
  - Malign: 352

- ISIC*MSK-1*2: Both malignant and benign melanocytic and non-melanocytic lesions. Almost all images confirmed by histopathology. Images not taken with modern digital cameras.

  - Benign: 339
  - Malign: 77

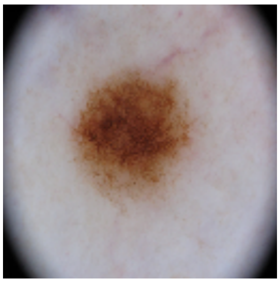
- ISIC*MSK-1*1: Moles and melanomas. Biopsy-confirmed melanocytic lesions, both malignant and benign.

  - Benign: 448
  - Malign: 224

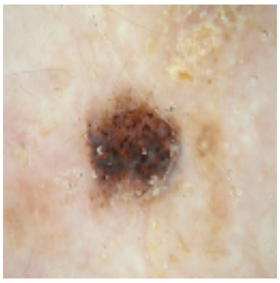
As summary the total images to use are:

Benign Images	**Malign Images**
1208	849

Some sample images are shown below: 1. Sample images of benign moles:



- Sample images of malign moles:



2.2 Preprocessing:

The following preprocessing tasks are going to be developed for each image: 1. Visual inspection to detect images with low quality or not representative 2. Image resizing: Transform images to 128x128x3 3. Crop images: Automatic or manual Crop 4. Other to define later in order to improve model quality

2.3 CNN Model:

The idea is to develop a simple CNN model from scratch, and evaluate the performance to set a baseline. The following steps to improve the model are: 1. Data augmentation: Rotations, noising, scaling to avoid overfitting 2. Transferred Learning: Using a pre-trained network construct some additional layer at the end to fine tuning our model. (VGG-16, or other) 3. Others to define.

2.4 Model Evaluation:

To evaluate the different models we will use ROC Curves and AUC score. To choose the correct model we will evaluate the precision and accuracy to set the threshold level that represent a good tradeoff between TPR and FPR.

3. Results presentation

As mention before the idea is to generate a tool to predict the probability of a malign mole. To do it, I'm planning to provide the following resources:

1. Web App: The web app will have the possibility that a user upload a high quality image of an specific mole. The results will be a prediction about the probability that the given mole be malign in terms of percentage. The backend that contain the web app and model loaded will be located in Amazon Web Services.

2. Iphone App: Our CNN model will be loaded into the iPhone to make local predictions. Advantages: The image data don't need to be uploaded to any server, because the model predictions can be done through the pre-trained model loaded into the iPhone.

3. Android App: (Optional if time allow it)

4. Project Schedule

Activity	Days	Current Status	Progress
1. Data Acquisition	1	Done	++++
2. Initial Preprocessing and visualizations	1	Done	++++
3. First Model Construction and tuning	2	Done	++++
4. Model Optimization I (Data augmentation)	1	Pending	----
5. Model Optimization II (Transferred learning)	2	Pending	----
6. Model Optimization III (Fine Tuning)	2	Pending	----
7. Web App Development + Backend Service	2	Pending	----
8. Ios App Development	2	In Progress	++--
9. Android App Development	2	Pending	----
10. Presentation preparation	1	Pending	----

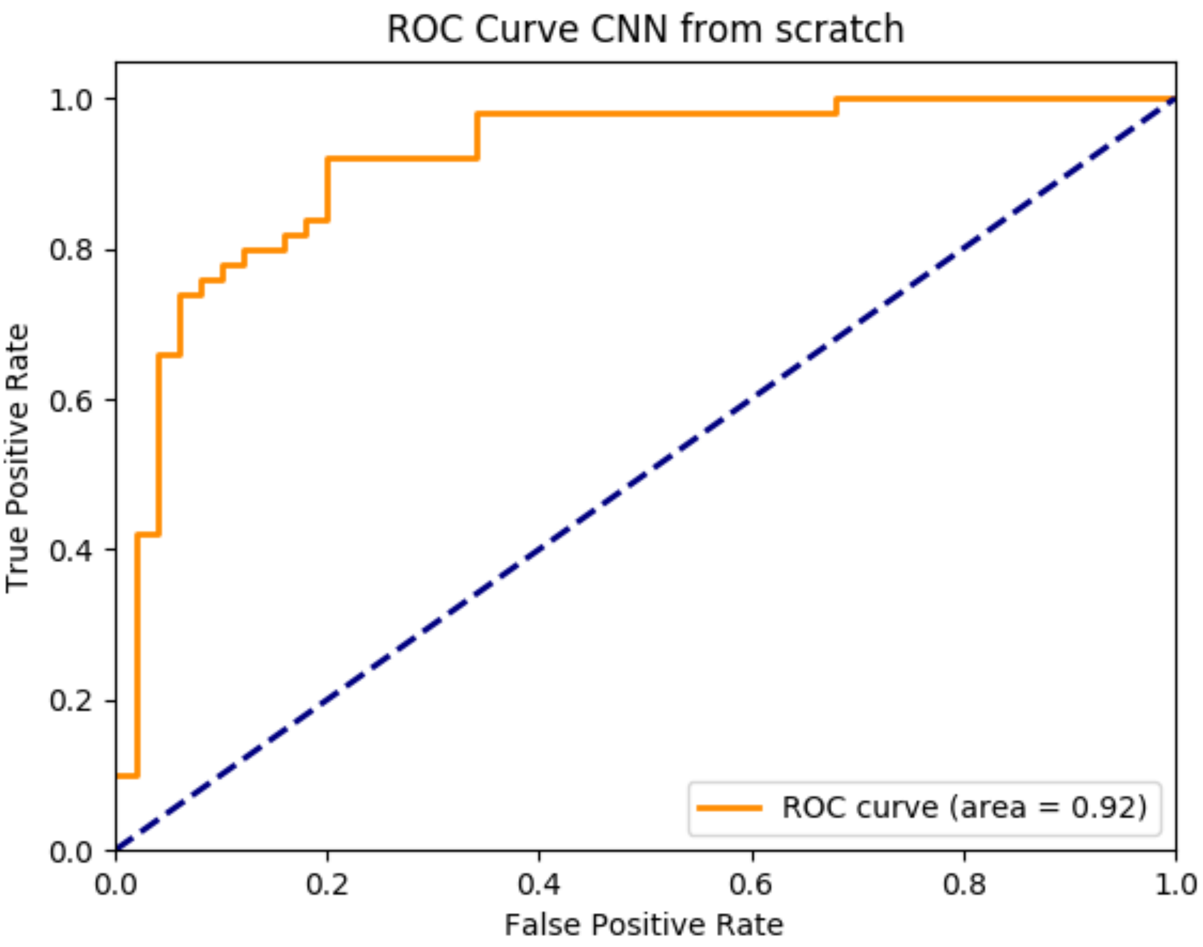
5. Tools to Use

- Tensorflow (GPU High performance computing - NVIDIA)
- keras
- Python
- matplotlib
- scikit-learn
- AWS (EC2 - S3)
- IoS swift + core ML
- Flask

6. Current Progress

First Model: CNN from scratch, no data augmentation

Simple Convolutional Neural Network with 3 layers. The results obtained until now can be shown on the ROC curve presented below:



7. Next Steps

TBD