Capstone Proposal

Domain Background

Skin cancer is the most prevalent type of cancer. Melanoma, specifically, is responsible for 75% of skin cancer deaths, despite being the least common skin cancer. The American Cancer Society estimates over 100,000 new melanoma cases will be diagnosed in 2020. It's also expected that almost 7,000 people will die from the disease. As with other cancers, early and accurate detection—potentially aided by data science—can make treatment more effective.

Currently, dermatologists evaluate every one of a patient's moles to identify outlier lesions or “ugly ducklings” that are most likely to be melanoma. Existing AI approaches have not adequately considered this clinical frame of reference. Dermatologists could enhance their diagnostic accuracy if detection algorithms take into account “contextual” images within the same patient to determine which images represent a melanoma. If successful, classifiers would be more accurate and could better support dermatological clinic work.

As the leading healthcare organization for informatics in medical imaging, the Society for Imaging Informatics in Medicine (SIIM)'s mission is to advance medical imaging informatics through education, research, and innovation in a multi-disciplinary community. SIIM is joined by the International Skin Imaging Collaboration (ISIC), an international effort to improve melanoma diagnosis. The ISIC Archive contains the largest publicly available collection of quality-controlled dermoscopic images of skin lesions.

Melanoma is a deadly disease, but if caught early, most melanomas can be cured with minor surgery. Image analysis tools that automate the diagnosis of melanoma will improve dermatologists' diagnostic accuracy. Better detection of melanoma has the opportunity to positively impact millions of people.

Source

Problem Statement

In this competition, participants will identify melanoma in images of skin lesions. In particular, they will use images within the same patient and determine which are likely to represent a melanoma. Specifically, participants need to predict a binary target for each image, the probability (floating point) between 0.0 and 1.0 that the lesion in the image is malignant (the target).

Datasets and Inputs

Images

The images are provided in DICOM & JPEG formats. DICOM images can be accessed using commonly-available libraries like pydicom, and contains both image and metadata. It is a commonly used medical imaging data format.

Total image files in training set ~ 33.1k (DICOM & JPEG each) Total image files in test set ~ 11.0k (DICOM * JPEG each)

Images are also provided in TFRecord format (in the tfrecords directory). Images in TFRecord format have been resized to a uniform 1024x1024 and there are overall 32 files in the directory.

Image Metadata

Metadata is also provided outside of the DICOM format, in CSV files.

Files

- train.csv - the training set (1.96 MB)
- test.csv - the test set (479 KB)
- sample_submission.csv - a sample submission file in the correct format (161 KB)

Total size of the dataset (Images + Files) - 108.19 GB .

Columns

- image_name - unique identifier, points to filename of related DICOM image
- patient_id - unique patient identifier
- sex - the sex of the patient (when unknown, will be blank)
- age_approx - approximate patient age at time of imaging
- anatom_site_general_challenge - location of imaged site
- diagnosis - detailed diagnosis information (train only)
- benign_malignant - indicator of malignancy of imaged lesion
- target - binarized version of the target variable, the value 0 denotes `benign`, and 1 indicates `malignant`

Source

All images and files can be downloaded from Kaggle on this link.

Evaluation Metrics

For this Kaggle competition, submissions are evaluated on area under the ROC curve between the predicted probability and the observed target. AUROC is a suitable metric for this image classification problem.

Solution Statement

For this competition, we are going to build an image classifier using deep learning. Such image analysis tools will also support clinical dermatologists in the future. The proposed image classifier model should be able to improve upon the benchmark model, described below, in terms of the evaluation metric (Area Under ROC).

Benchmark Model

While dermatologists are currently identifying lesions manually, and no existing automated solution exists, we can build a simple baseline model against which we can compare the accuracy of our image classification model.

A good baseline model can be created by using 3 features, namely, age, sex and the location of the images site. We can calculate the grouped mean value for each combination of these features in the train set and use that to make predictions on the test set.

Predictions using this simple mean value of the target variable gives an Area under ROC value of `0.699` on the public leaderboard !!

Our image classifier should be able to beat atleast this benchmark to be deemed useful.

The code for the baseline model is added in the Github repo. Here is the Kaggle Kernel link.

Project Design

Image processing steps might differ for images with different file sizes, for eg., 1024x1024x3 vs 512x512x3 etc. We can combine the given tfrecords data for images with the image metadata to create new tfrecords, and then use them for modeling.

Since we are working with tfrecords, we can use Tensorflow library to build our neural network. We can use stratified k-folds for model validation before making predictions on the test set.

Since training a deep learning model on a large image dataset (~120 GB) is going to be a compute heavy task, Kaggle notebooks which offer free GPUs (and TPUs) can serve as the ideal solution for training this model.

Additionally, pretrained models such as ImageNet might be explored to get a good score.

We have already seen that the benchmark model using patient info described above has given good results (AUROC 69.9%), so the final model can be some kind of ensemble. For eg, bagging the results of the image classifier with the benchmark model might be worth trying out.