**Author**: Su Kara

**Subject**: Project Log Book

Please note that I’ve always taken notes on my computer rather than handwriting in a log book.

**8/15/19 - First meeting with Dr. Chang**

* Discussed several project ideas and decided on the following:  
  Project description: In this study, we will design a deep learning based image analysis algorithm for assessment of endotracheal tube (ETT) position from chest radiographs. Specifically, the algorithm will be comprised of two serial convolutional neural networks---a first algorithm to detect presence or absence of ETT; and a second algorithm for determination of the distal tube tip relative to the carina.
* He sent me the following links to get familiar with chest x-ray datasets and download the latest MIMIC-CXR (Medical Information Mart for Intensive Care - Chest X-Rays) dataset from PhysioNet:  
  NIH CXR: <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview>MIMIC CXR: <https://mimic.physionet.org/gettingstarted/access/>
* He also asked me to complete his lessons and assignments:
  + Intro to Python I: <https://bit.ly/2JCaELu>
  + Intro to Python II: <https://bit.ly/2XKkkw6>
  + Intro to Python III: <https://bit.ly/2GhhIuL>
  + Assignment 1: <https://bit.ly/2XR3neF>
  + Assignment 2: <https://bit.ly/32EwBAU>
  + Assignment 3: <https://bit.ly/32QFJ5t>
  + Assignment 4a: <https://bit.ly/2Y5ZEi9>
  + Assignment 5: <https://bit.ly/2OuViNj>

**8/19/19 - Collaborative Institutional Training Initiative (CITI Program)**

* Completed this PhysioNet certification program to be able to download the MIMIC-CXR dataset from MIT
* Required the completion of the Data or Specimens Only Research course in the Human Research curriculum

**9/2/19 - Status Update for Dr. Chang**

* Downloaded the NIH CXR dataset and installed OsiriX Lite to review the images.
* Completed the PhysioNet program to get certified.
* Applied for the MIMIC CXR dataset at MIT and got approved. However, I couldn’t download the data because of an issue on their website. I sent them an email and they said it would be fixed in a week but I’m still waiting. As soon as it is fixed, I will download the data and review it.
* Completed the first week of Convolutional Neural Network Coursera class by Andrew Ng.
* Completed Intro to Python I and II.
* Finished Assignment 1.

**9/23/19 - Status Update for Dr. Chang**

* Completed the second week of Coursera classes.
* Completed Intro to Python III and all the assignments.
* Attended the SCAS meeting at Loyola Marymount University on September 7

**9/30/19 - MIMIC-CXR Dataset Available at MIT**

* <https://physionet.org/content/mimic-cxr/2.0.0/>

MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0

377,110 images corresponding to 227,835 radiographic studies

- A set of 10 folders (p10 - p19), each with ~6,500 sub-folders. Sub-folders are named according to the patient identifier, and contain free-text reports and DICOM files for all studies for that patient

- cxr-record-list.csv.gz - a compressed file providing the link between an image, its corresponding study identifier, and its corresponding patient identifier

- cxr-study-list.csv.gz - a compressed file providing a link between anonymous study and patient identifiers

Folder structure

Free-text reports and images are provided in individual folders. An example of the folder structure for a single patient's images is as follows:

files

└── p10

└── p10000032

├── s50414267

│ ├── 02aa804e-bde0afdd-112c0b34-7bc16630-4e384014.dcm

│ └── 174413ec-4ec4c1f7-34ea26b7-c5f994f8-79ef1962.dcm

├── s50414267.txt

├── s53189527

│ ├── 2a2277a9-b0ded155-c0de8eb9-c124d10e-82c5caab.dcm

│ └── e084de3b-be89b11e-20fe3f9f-9c8d8dfe-4cfd202c.dcm

├── s53189527.txt

├── s53911762

│ ├── 68b5c4b1-227d0485-9cc38c3f-7b84ab51-4b472714.dcm

│ └── fffabebf-74fd3a1f-673b6b41-96ec0ac9-2ab69818.dcm

├── s53911762.txt

├── s56699142

│ └── ea030e7a-2e3b1346-bc518786-7a8fd698-f673b44c.dcm

└── s56699142.txt

* Above, we have a single patient, p10000032. Since the first three characters of the folder name are p10, the patient folder is in the p10/ folder. This patient has four radiographic studies: s50414267, s53189527, s53911762, and s56699142. These study identifiers are completely random, and their order has no implications for the chronological order of the actual studies. Each study has two chest x-rays associated with it, except s56699142, which only has one study.

**10/7/19 - Examine report files**

* Review several hundred report files and noticed several different kinds of tubes mentioned as below:  
    
   Here is a small subset of my findings:

p10/p10001401/s50225296.txt: Nasogastric (NG) tube

p10/p10001401/s56534136.txt: Enteric tube

p10/p10001884/s56722923.txt: Esophageal (ET) tube

p10/p10001884/s59305618.txt: Intubated

p10/p10001884/s50376803.txt: Endotracheal (ETT) tube

p10/p10002013/s55969846.txt: Chest tube

p10/p10002428/s57961288.txt: Dobbhoff tube

p10/p10002428/s59659695.txt: Endotracheal tube, Dobbhoff feeding tube

* Found a general pattern to identify an ETT in a report.
* Wrote a Python program to search for "endotracheal or ett \* ? cm \* carina" in 227,835 report files by using a proper regular expression, and found 8,900 files that had a match. Wrote the distance to carina in centimeters (? position above) into a file. Here is a list of the first 5 lines from the output file:

subject\_id,study\_id,dist\_carina

10001884,50376803,6

10002428,50292543,2.3

10002428,56597576,5.6

10002428,56836542,4.6

**10/21/19 - Second meeting with Dr. Chang**

* Reviewed the MIMIC-CXR dataset:  
   65,389 patients  
   227,835 studies/reports  
   377,110 images  
   15,344 reports (studies) contain Endotracheal or ETT  
   8,712 reports (studies) contain Endotracheal or ETT with a distance to carina
* Dr. Chang requested two files (one positive and one negative), each containing 8,000 studies

**10/23/19 - Label files ready**

* Sent the following two label files to Dr. Chang:  
   ett.csv - 8,000 studies with positive labels with distance values  
   no-ett.csv - 8,000 studies with no labels

**11/11/19 - Image preprocessing completed**

* Dr. Chang completed the preprocessing of images from their original size of 3056x2544 15 MB 0.139 pixel spacing (some images 1760x2140 7 MB 0.2 pixel spacing) to 512 x 512.
* Sent me 24,010 image files that correspond to 16,000 studies (8,000 positive and 8,000 negative)
* Downloaded and opened the zip file.
* Saved the images and labels into x.pkl and y.pkl files.
* Loaded the images and labels from the pickle files.
* Checked the 24,010 labels and found out that there are 9036 positive and 14974 negative labels even though the studies were evenly distributed as 8000 and 8000.
* Used the same CNN algorithm from assignment 4 and the two times I tried, it crashed in a couple of minutes. This is most likely due to the big size of x.pkl as it contains 24010\*512\*512 int16. In assignment 4, we had only 60000\*28\*28 int8.

**11/19/19 - Third meeting with Dr. Chang**

* Dr. Chang sent a new zip file with cropped images in the size of 256x128.
* Some of the cropped images were not in the given size of 256x128.
* 23,444 images: Removed the ones that are not 256x128, which dropped the count to 20,558
* 20,558 images: 7,685 positive and 12,873 negative

CNN Architecture:

- BLOCK 1 (filter depth of 16)

- BLOCK 2 (filter depth of 32)

- BLOCK 3 (filter depth of 64)

- BLOCK 4 (filter depth of 128)

- BLOCK 5 (filter depth of 256)

- RESHAPE (8 \* 4 \* 256, 128)

- HIDDEN LAYER (128 nodes)

- LOGIT SCORES (2 nodes)

- iterations=2000, batch\_size=256

- training took 3.5 hours

Accuracy:

- train accuracy: 0.98

- test accuracy: 0.94

**11/21/19 - Test different models**

* Dr. Chang recommended testing the following changes in the model
  + Change the total number of iterations between 1000 and 8000
  + Change the filter depth between 32 and 256
  + Change the batch size between 32 and 256
  + Best combination was 4000 iterations, 64 filter depth, and 64 batch size
  + It was the fastest and highest accuracy

**11/27/19 - Separate training and test images**

* Dr. Chang pointed out that if a patient has more than one DICOM file, which is very common, then make sure that all images for that patient are either used for training OR validation, but *not* both. This is because the two images are likely to be similar, and we shouldn't mix. This should not increase accuracy; in fact if anything it may decrease. The reason is that when two images from the same patient are used in training and validation, the algorithm can "cheat" and our estimate of accuracy will be erroneously high. So this is a measure just to ensure we can trust our results. Best case scenario, there is no change in performance indicating that our algorithm was able to generalize well.
* Since patient images were kept in 10 main directories (p10-p19), I used p10-p17 (8 directories) for training and p18-p19 (2 directories) for testing.

**11/27/19 - Cross-validation to catch failed images**

* Dr. Chang asked me to find all the exams the algorithm has made a mistake on. The only way we can tell a mistake was made is by assessing that exam as part of the validation group (not training, since it is likely that the algorithm will overfit and get most of them correct even if the label is wrong). So we need to do what's known as cross-validation. Split your data into five groups. Then run a total of five experiments (using your top model), where each experiment uses a different 1/5th of the data for validation. After you are done, each data has had an opportunity to be "validation" data once. From the above, find a list of all patients where the algorithm made a "mistake". Then convert those over to PNG.
* Identified 1,308 mistakes, which is 6% of 20,558 images since the overall test accuracy is 94%.
* Visually inspected those images and moved them into two folders as TruePositive and TrueNegative for proper labelling.
* This cleanup of labels increased the accuracy to 97%.
* Successful completion of the CNN to classify the image as having ETT or not

**12/19/2019 - Fourth meeting with Dr. Chang**

* Dr. Chang put me in touch with Scott Refugio and Charles Lin to use their viewer/coordinate annotation tool written in Python.
* Used the tool to locate the tip of ETT by clicking the mouse on it.
* The tool generates a CSV file with the list of image files and the X and Y coordinates of the point clicked on.

**1/10/2020 - Annotate the tip of ETT**

* Generated an annotation.csv file for the 25 images I examined.
* Dr. Chang asked me to convert the images to grayscale and superimpose a 10x10 pixel square on the estimated coordinate.
* He reviewed my work and confirmed that they were all correct.

**1/21/2020 - Fifth meeting with Dr. Chang**

* Dr. Chang asked me to convert coordinates into floating points between 0 and 1, normalize them: y/255, x/127
* Use sigmoid function, so prediction is between 0 and 1 as shown in the following template:

logits = tf.matmul(h1, w7)

logits = tf.nn.sigmoid(logits)

# — L1 loss

loss = tf.reduce\_sum(logits - labels)

# — L2 loss

loss = tf.reduce\_sum((logits - labels) \*\* 2)

**1/30/2020 - Regression CNN to find the coordinates of the tip of ETT**

* Replaced prediction and accuracy with RMSE (root mean square error) based on the loss.
* Tested on the 25 annotations successfully.
* L2 seems to be performing better than L1.
* Converted the true positive images into grayscale and resized as 512x256 to annotate easily.

**2/4/2020 - Annotate the position of the tip of ETT on the true positive images**

* Used the tool to annotate the ETT tip on 3,555 images.
* Tested the CNN regression algorithm with 5-fold cross validation.
* Found ETT tip within 0.78cm of annotations.

**2/7/2020 - Fix 5-fold cross validation**

* Noticed that the accuracy was improving in every fold of 5-fold cross validation and sent an email to check Dr. Chang to confirm.
* Dr. Chang confirmed that I had to reset the model in every fold to start fresh.
* Updated the 5-fold cross validation logic in both regression and classification algorithms.
* The accuracy decreased in both algorithms.

**2/10/2020 - Issue with the position of carina and the cropping algorithm**

* Noticed that the position of carina was not correct.
* The cropping algorithm seemed to have cropped wrong parts of the original images rather than (0.75, 0.5) around the carina.
* Asked Dr. Chang whether I should redo those parts by adding a new regression algorithm similar to the one that found the ETT tip, and he confirmed that I could.

**2/11/2020 - Annotate the position of carina on 4,000 images**

* Used the tool to annotate the position of carina on 4,000 images.
* Regression algorithm found the carina within 0.63cm of annotations.

**2/14/2020 - A new cropping algorithm**

* Used the model to find the carina on the 17,050 frontal images that were 512x512.
* Cropped 256x128 around the carina to have it located at (0.75, 0.5), i.e. (191, 63).

**2/16/2020 - Use the new cropped images**

* Visually inspected around 1,000 more images to identify as true positive and true negative.
* Reran the classification algorithm with the new cropped images to have an accuracy of 96%.

**2/17/2020 - Annotate the ETT tip on the new cropped images**

* Used the tool to annotate the ETT tip on 3,000 images.
* Found ETT tip within 0.66cm of annotations.

**2/24/2020 - Scientific paper**

* Wrote a scientific paper for the project based on the standards provided in the Research Training Program (RTP) of Southern California Academy of Sciences (SCAS).

**3/5/2020 - Poster board**

* Prepared a poster board for OCSEF 2020.