CSCI 1470 Final Project Report:
Effect of Bone Suppression on X-Ray Classification

Bone Vision: Hossam Zaki, Nasheath Ahmed, Andrew Aoun, Mohamad Abouelafia.
Brown University
10th December 2020

1. Introduction

Modern technological advances have fundamentally changed the intersection between medicine and computer science. Through the fields of artificial intelligence (machine/deep learning), computational models have become valuable tools for both data analysis and classification. In this project, we focus on the latter and aim to compare classification performance across images with and without feature suppression.

We chose to focus on medical imaging data sets and, specifically, X-ray images where bone shadows have been shown to impede visual classification by medical professionals. We therefore aim to investigate whether the suppression of such shadows can also improve disease recognition in computational models. Here we use a Convolutional Neural Network (CNN) for classification and an Auto-encoder for bone suppression. This is a challenging task as images vary highly in quality, contrast and bone shadow intensity. A significant performance increase for bone shadow suppressed images would provide evidence that bone shadow suppression is an effective de-noising process that should be applied before X-ray classification models.

2. Methodology

2.1. Bone Suppression

The data set for the bone suppression methodology consists of normal x-ray images that contain the bone shadow of the rib cages and the clavicle. The target images consist of images with these shadows removed from the images. The data was organized into a train, validation, and test split of .70/.15/.15 respectively. The total data set consisted of 4080 images in total that were trained and tested with. Images were pre-processed with values between 0-1 to be passed into the model architecture and the full size image of 1024*1024*3 were used. Furthermore, images contained random horizontal flips for data augmentation purposes in order to decrease overfitting of the images.

Following the pre-processing, we adopted an auto encoder methodology for training the bone suppression algorithm. The model architecture was composed of three encoding layers to compress the information of the images and three decoding that upsample and increase the dimensions of the newly produced images. The encoding layers consisted of the the convolution layers with output filter sizes of 16, 32, 64 respectively with strides of 2 in order to downsize the images. The decoding layer consisted of three Upsampling2D layers with a size of 2 to double the size of the image after each upsample followed by a Conv2D layer with strides of 1x1 to maintain the size of the images. The output filter sizes for the Conv2D layers were 32, 16, and 3 respectively. In all runs of the model, the ADAM optimizer was used with an initial learning rate of 0.001 with default parameters 1 = 0.9 and 2 = 0.999. After 25 epochs, the learning rate was decreased to 0.0001 with default parameters of 1 = 0.9 and 2 = 0.999. We used a batch-size of 4 and the loss was calculated using mean-squared-error and ssim[1]. MSE was found to work well as a loss function for image reconstruction, but fails to take into account the intricacy of the human body. Another useful loss function is the structural similarity index, SSIM for short. SSIM accounts for the overall "perceptual" difference between the two images, rather than pixel-for-pixel difference in MSE[2]. We also defined α to be .80[2]. Our loss function is defined as (Eq. 1):

$$L = \alpha * SSIM + (1 - \alpha) * MSE \tag{1}$$

The model was implemented using Tensorflow, and training was done on an NVIDIA P100 GPU for a total of 75 epochs.

2.2. Classification

This model will contain two instances. Since we hope to draw conclusions about accuracy changes for x-ray images with and without bone shadow removal, the first instance of the second model (CNN) will take in transformed x-ray images (Bone shadow suppressed) using the best weights from the auto encoder bone suppression model, and the second instance of the second model will take in untransformed

x-ray images. Therefore these instances will have the same architecture but the inputs and, more importantly, the learned weights, will vary.

The data set for the classification model comes from the NIH clinical center that provides over one hundred and twenty thousand chest x-rays, from more than thirty thousand patients. The data set has a csv file with the labels and various other metadata concerning the patients. The csv was used to read in the x-rays and their associated labels. The following criteria was used to pre-process the data: removing regular x-rays (x-rays with a 'no finding' label), removing diseases with less than 1000 x-rays available, and x-rays that do not have multiple labels. Furthermore, we augmented the data by doing horizontal flips. This resulted in 9 classifications and 28284 unique x-rays. This provided us with a well distributed number of x-rays per classification. The x-ray images used in the model were down scaled to (512, 512,3) from (1024, 1024, 3), and were standardized to values between 0-1.

The classification model can be seen below in Figure 1 below. It consisted of the built in Keras application RESNET50 (with pre-loaded Imagenet weights), a dense layer (Relu activation), a dropout layer (dropout rate of .5), and a dense layer with the size of the number of classifications (softmax activation). Additionally, we used Adam as our optimizer with a learning rate of .00001, and our metric for loss is 'categorical crossentropy'. We used several metrics accuracy that we deemed relevant: AUC, categorical accuracy, and binary accuracy. These metrics provided us with different ways to interpret how our model varied with and without bone suppressed images, which is discussed further under results. The models were run using a Tensorflow framework and training was done using a single NVIDIA P100 GPU for a total of 50 epochs.

Model: "chest_class_model"

Layer (type)	Output Shape	Param #
resnet50 (Functional)	(None, 16, 16, 2048)	23587712
flatten (Flatten)	(None, 524288)	0
dense (Dense)	(None, 256)	134217984
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 9)	2313

Total params: 157,808,009
Trainable params: 157,754,889
Non-trainable params: 53,120

Figure 1. Architecture for the classification model

3. Results

Following the auto-encoder architecture described, the evidence of bone shadow suppression was clear. This was witnessed both visually by inspecting the images against the labels (figures 4 and 5) as well as numerically by decreasing loss values over the training period with an average loss of 0.03. However, it is important to acknowledge the quality of these images as well as the level of bone suppression. As will be covered in the discussion section, the auto-encoder does not produce images with high enough quality and essentially 'blurs' the region around bone shadows (figure 4) for the classification images. This does in fact suppress bone shadows but leaves the question as to the impact on the classification results. This result is not as visible when testing images from the bone-shadow data set. Additionally, bone suppression appears to be minimal and focused towards the inner part of the lungs only on lung classification images (figure 4)

For both classification tasks without the bone suppression algorithm auto encoder applied and with the bone suppression algorithm applied to the images, we calculate the the area under the ROC curve (AUC). The ROC calculates the true positive rate over the false positive rate of the classifications, so we decided this would be a valid metric for this task we are trying to examine with different diseases. Figures 2 and 3 below represent the AUC curves for both classification tasks. The AUC was calculated over the total span of all the test images. The average AUC value for the curve without bone suppression was 0.78, and the average AUC value for the images with bone suppression was 0.74. The AUC curve value was greater than 0.80 for over 4 of the disease types which were Atelectasis, Cardiomegaly, Effusion, and Pneumothorax. This was the same for the classification results with bone suppression, but the values for the classification with bone suppression contained lower values. There didn't appear to be a significant difference between the AUC curve values for the classification task with bone suppressed images and without bone suppression images. For the AUC curve with Nodule and Plueral Thickening, the AUC values were 0.75 and 0.71 respectively for images without bone suppression applied to it and 0.68 and 0.66 for the images with bone suppression applied to it. The lower values for the AUC curve in the images with the bone suppressed can be attributed to the difference in the images and the slight blur that is generated from the autoencoder generated images.

4. Challenges

The issues we encountered related almost entirely to the processing power, memory and the google cloud computing (GCP) instance. We found that our original bone-suppression model would take nearly a week to complete. We explored the NVIDIA V100 GPU as an alternative, however, issues arose here related to both the lack of credits as well as starting the V100 GPU. Currently we have managed to reduce the length of time for the Auto-encoder model to train to 20-24 hours by reducing the size of our training data set as well as various manipulations of the base model. These include: filter sizes, number of filters, batch sizes and other general structural changes to reduce this time.

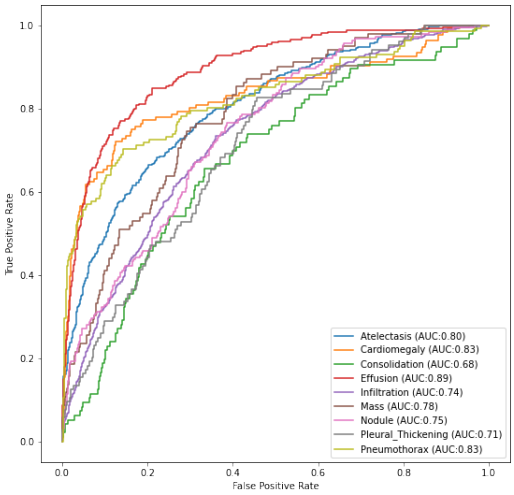

Figure 2. AUC curves of Classification task without Bone Suppression Algorithm Applied to the images

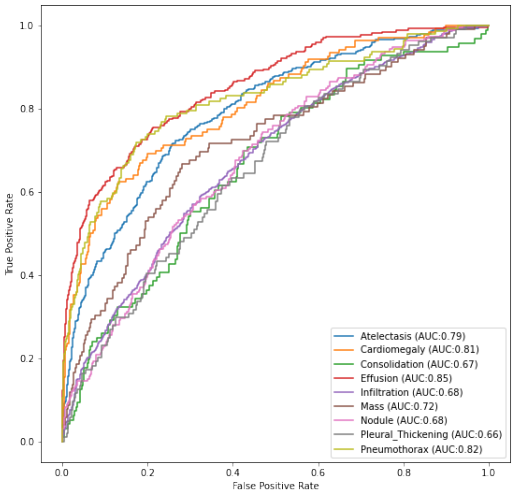

Figure 3. AUC curves of Classification task with Bone Suppression Algorithm Applied to the images

When we began the project, our bone suppression model was a Generative Adversarial Network (GAN), and provided good results on our data set. However, when we applied the GAN to the X-Ray Classification data set, we encountered a major issue. We found that the X-Rays in the NIH data set for classification were highly different to the X-Rays in the bone suppression data set. Therefore, when the GAN was applied to the NIH data set, we encountered large black blotches, obscuring the lung. We concluded that this was due

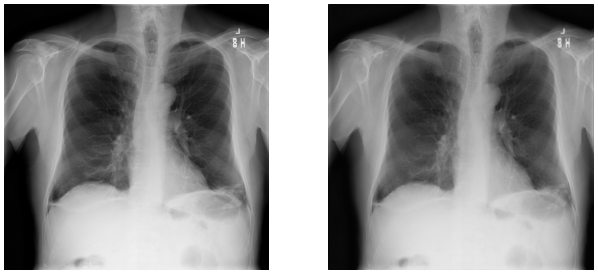

Figure 4. Image on the left:Represents the x-ray image without bone suppression algorithm. Image on right: Represents thee x-ray image with the bone suppression algorithm applied to it

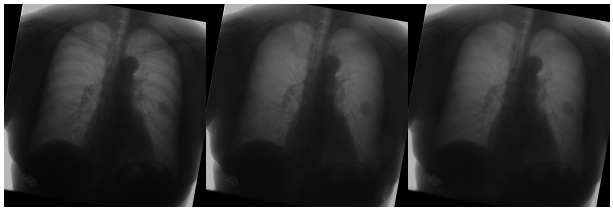

Figure 5. Left image:Source image; Middle image: Target Image; Right image:Outputted image by the auto encoder bone suppression algorithm

to the specificity of the GAN on the bone suppression data set. As a result, we transitioned to an auto-encoder, which solved this issue.

We also encountered issues implementing the classification model on the NIH Lung data set. The model seemed to not be learning with accuracy and loss metrics staying consistent over several epochs. In response, we lowered our learning rate to .0001 and initialized our RESNET-50 model with weights from Imagenet. This provided the baseline for the model to begin learning. Following this, we ran into issues with over fitting. We observed that our model had an average AUC of .95 on the training data compared with .70 on our testing data. To counteract this, we used Keras' data generator to build more images, lowered the number of epochs to train, as well as adding a dropout layer. This allowed our model to train more epochs before over fitting.

5. Reflection

This project was a stretch project for us. It consisted of two different problems, models, data sets, and implementations. We wanted to figure out if removing the bones from an x-ray would improve the overall classification using an neural network. As referenced in the introduction, radiologists are able to classify x-rays quicker and more accurately when the bones are moved from the x-rays, so we wanted to see if our neural network would perform the same. We are happy with how our project turned out, however, we think we could improve our results. First, we had to make a major pivot in our bone-suppression model. First, we used a GAN to remove the bones, and it ended up working very well.

However, it was not generalizable. When we used this model on the classification data, we saw massive blotches of black, which extremely hindered the classification process. Upon inspection, we found that the x-ray images in the suppression dataset and in the classification dataset, looked significantly different. It appeared that the dataset in the suppression x-rays were inverted, and a little more clear. This is a major problem for us, and moving forward, we plan on finding a new dataset that looks more like standard X-Ray images. We pivoted from using a GAN to a pure autoencoder so that it would be more generalizable. However, the autoencoder was less successful. If given more time, we would like to run our autoencoder for longer periods of time, and perhaps improve the architecture of our autoencoder as well.

For our classification model, we read several papers, and concluded that RESNET-50 worked well in a decent amount of time. We used an NIH Lung X-Ray dataset for this model as well, however we noticed that the quality of the x-rays aren't great, with large white blotches obscuring the lung completely. While our model still performs well, we would like to use another dataset. We found a dataset from Stanford which much cleaner X-Ray images, and as we continue to work on this project, we hope to use this dataset. Furthermore, we had hoped to have some time to add a heatmap to the lung images so as to allow a person to determine what features caused the model to predict a certain diagnosis. As we continue this project, this is something we would like to add.

Overall, we feel that we learned a lot throughout this project. We learned more about how GANs work, and how they differ from pure autoencoders. We also learned about the importance of data, and the hyper-parameters. Additionally, we learned more about reading papers and extracting important details from the paper and implementing them in our code.

We thoroughly enjoyed this project, and hope to continue this in the future.

References

[1] M. Gusarev et al. "Deep learning models for bone suppression in chest radiographs". In: *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2017, pp. 1–7. DOI: 10.1109/CIBCB.2017.8058543.

[2] Hang Zhao et al. "Loss Functions for Neural Networks for Image Processing". In: *CoRR* abs/1511.08861 (2015). arXiv: 1511.08861. URL: http://arxiv.org/abs/1511.08861.