

Thoracic Organ Segmentation using U-Net Deep Learning Architecture

Will Argus
wargus@eng.ucsd.edu

Harmeet Gill
hsgill@eng.ucsd.edu

Karndeeep Singh Rai-Bhatti
kraibhat@eng.ucsd.edu

Abstract— In this paper, we will discuss our implementation of deep learning to segment organs in CT scans of the human thorax. The dataset, and methods are provided by a competition hosted during the Annual Meeting of American Association of Physicists in Medicine in 2017.

Keywords—AAPM, Dice, U-Net, DICOM, Thoracic, Binary Cross entropy

I. INTRODUCTION

In 2017, the Annual Meeting of American Association of Physicists in Medicine (AAPM) hosted a challenge intended to encourage exploration and evaluation of machine learning approaches for auto-segmenting Computed tomography (CT) scans. In medicine, CT scans are used to map the internal anatomy of the body based on x-ray attenuation measurements. These measurements are made at different angles around the body and used to reconstruct cross sectional images.

Even if a CT scan can be used to visualize the organs within a body, more work is required to make the CT scan useful. In the case of radiation therapy, a common use case for CT scans, the location of the entire organ of interest must be identified and annotated on a CT scan in a process called segmentation. Typically, this is done manually, a tedious process where a radiologist combs through the entire volume of images marks each pixel that shows the organ of interest. The idea behind auto-segmentation is that this process can be automated such that it is faster, cheaper, and gives more consistent results.

Returning to the topic of the 2017 annual meeting of the AAPM, all groups were given the same dataset and graded using the same criteria. The designated organs of interest were the heart, spinal cord, esophagus, and the right and left lungs. This paper will detail the authors' implementation of one of the approaches involving a U-Net using the same dataset used in the competition

II. RELATED WORK

This problem was approached by a group in the University of South Carolina, which trained a deep CNN to segment the CT scans using specific orthogonal 2D slices of the scans. The CNN used coarse feature recognition and then fine extraction to segment the pixels of the 3D scans. This solution focused on real time segmentation. They segmented 2

organs, the left kidney and the pancreases, with an intersection over union score of 88 and 65 percent accuracy, respectively.

The general CNN architecture used by the competition team whose approach the authors are implementing here, is known as a U-Net and was developed by a group at the University of Freiburg, Germany. The U-Net architecture is common in biomedical image processing due to its ability to assign a label to each pixel in an image, meaning that it can identify specific desired regions in an image.

III. DATASET AND FEATURES

Our data set was put together by the AAMP to serve as a benchmark for comparing Auto-Segmentation methods. The Dataset consists of 60 thoracic CT scans. All cross sectional images are 512 by 512 pixels. The height of a given volume varies, but is on average 130 images. Images were taken using 3 different scanners at different institutes (20 scans each). This leads to significant variation between volumes. These variations are both due to differences in patients (body fat, bone structure), as well as scanner settings (slice thickness, scanner voltage, scanner quality). CT data was presented in the native dicom format.

Each CT volume was accompanied by a segmentation file. Volumes were manually and independently annotated by 3 radiologists following guidelines. The provided segmentations were generated via pairwise comparison of the radiologists' annotations. Every included the following labeled and non-overlapping regions: Heart, Esophagus, Spinal Cord, and Lungs.

For this task, the features that we hope to detect are well defined. We wanted to identify organs, each with its own set of defining features. For example, the spinal cord can be recognized by its shape as well as the sharp change in attenuation coefficients from the bone to tissue. The Lungs similarly benefit, mainly being filled with air and being surrounded by tissue.

IV. METHODS AND MODEL

A. U-Net Architecture

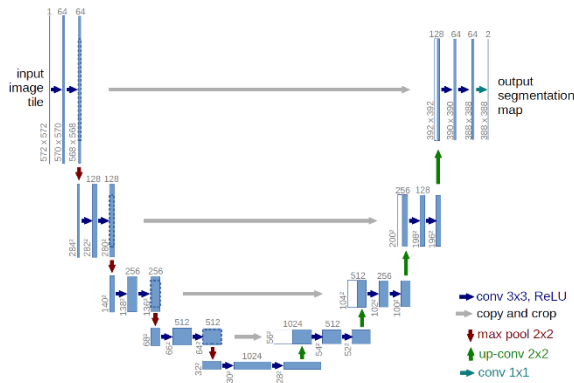


Figure 1 - Diagram of U-net architecture published by the Brox research group

The U-Net is a Convolutional Network designed specifically for biomedical image segmentation. An image is input into the network. The network outputs a binary mask corresponding to the region of interest.

The network architecture resembles that of an autoencoder, consisting of a contracting path, a bottleneck, and an expanding path. The contracting path consists of a single unit repeated 4 times. The unit consists of a convolutional layer followed by a ReLU activation layer repeated twice, followed by a max pooling layer. The bottleneck consists of two more convolutional layers with ReLU activations functions. The expansive layer consists of an analogous unit to the contracting path being repeated 4 times. This consists of upsampling followed by a 2x2 convolutional layer and the same convolutional layers with ReLU activations.

Prior to every maxpool layer in the contracting path, a skip connection connects the layer to the corresponding layer in the expanding path. These skip connections serve to ensure the robust maintenance of spatial information within the network. The final layer consisted of a sigmoid function. This is coupled to a binary cross entropy loss. In training, the Adam optimizer was used.

$$\text{Binary CrossEntropy} = \sum_{i=0} y_{i,pred} \log(y_{i,truth})$$

Due to the kernel size used for each of the 23 convolutional layers, the model is only able to accurately predict information in the central two thirds of the image. This is due to a loss of information at the edges of the matrix with every convolution.

B. Our Model Specifics

Dropout layers were included within the model to reduce over fitting with the goal of improving the generalization of the model. One dropout layer was included at the end of the contracting path and another following the bottleneck. This specific configuration was chosen to optimize the compact representation at the end of the bottleneck. No earlier dropouts were included to maintain the integrity of the high level spatial information used in reconstruction. Similarly, no dropouts were included in the expansive (generative) path.

Within the dataset, positively labeled voxels for several of the target labels suffered from class imbalance. For example, the heart made up approximately 1% of the pixels in a CT volume. The approach taken in our study was to segment every organ individually, further exacerbating the problem of class imbalance.

To solve this, the dataset was resampled. All slices containing any positive label were separated out. Along with this, a set number of slices were randomly selected from the rest of the volume to represent the remainder of the images. The model was then trained on this subset.

This solution was only applicable to the heart. Organs such as the spine and esophagus with very small segmented areas in a majority of slices. Accounting for this would require implementing alternative loss functions.

V. EXPERIMENTS, RESULTS AND DISCUSSION

Results of the model varied between the four organs. A separate but identical neural network was used for each of the organ segmentations, with 36 CT volumes for training, 9 for validation and 15 for testing. During the training process, CT volumes were broken down by individual slices, and passed to the training model via a generator function by a batch size of 5 stacks (5x512x512x1 array density). Each epoch stepped through all the training images for all volumes, thus a large number of epochs were not necessary before the model started to over fit.

The Dice coefficient was used to judge performance of the model. This metric divides the union between truth and prediction with the total sum of both objects, and gives a better indication of whether our prediction remained in the same space as the ground truth.

$$\text{Dice} = \frac{2|y_{truth} \cup y_{pred}|}{|y_{truth}| + |y_{pred}|}$$

A summary of the Dice scores for each organ can be found in Table 1.

A. Heart

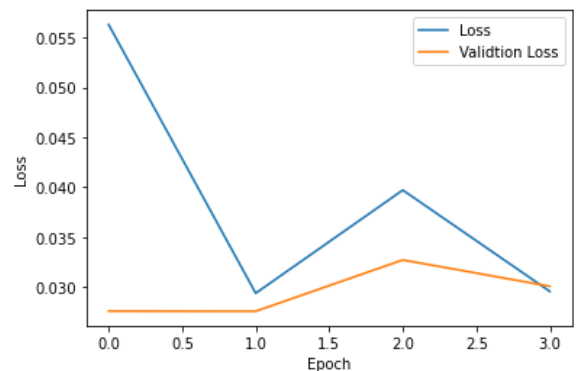


Figure 2 - Binary Cross-Entropy loss of heart training for 4 epochs

Our heart region was well predicted, with a loss of less than 3.5 percent. The loss function in any instance was calculated using the images in a single batch, therefore does

not necessarily indicate a fully stacked volumetric loss value. Because of this, low loss models were still observed to have volumetric inconsistency as seen in the stacked dimension. This can be seen in Figure 3, where the top of the heart prediction in the right object has discontinuous slices, while the left ground truth is consistent.

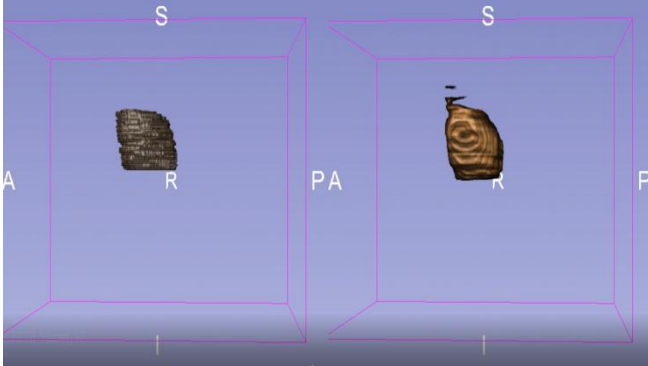


Figure 3 - 3-D view of heart with ground truth on left, and prediction on right

One problem that was prevalent in heart training was class imbalance. Initially, most of the training images were void of any true pixel labels. As a result, our model did not predict the heart region well after training. To mitigate this, the training files were split into sections that contained true labels and sections that did not. All sections that contained true labels were fed into the model, and an equally volumetric stack containing randomized selection of sparse images were also fed in.

B. Esophagus

The esophagus, of all the organ predictions, was the worst predicted. Figure 4 shows the binary-cross entropy loss of the esophagus training for 6 epochs. Although the training and validation loss was low (approximately 0), this is not indicative of the accuracy of the model with respect to the segmented masks. Because the esophagus occupies a small region at any given slice (as can be seen in Figure 5), the majority of weight in the loss calculation is being done on pixels with a value of zero, which is skewing the value down. A more sensitive measurement of volumetric similarity can be achieved using the dice coefficient. Our mean Dice coefficient against 15 test volumes was 8.8 percent, which is indicative of a bad fit.

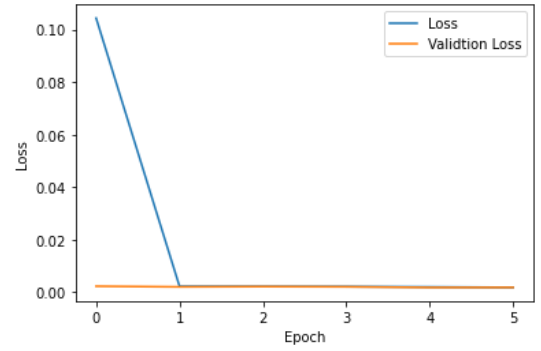


Figure 4 - Binary Cross-Entropy loss of esophagus training for 6 epochs

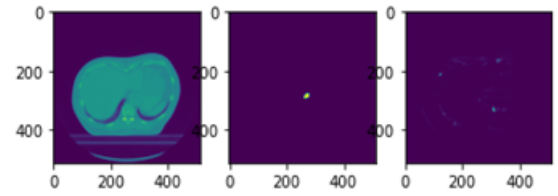


Figure 5 - Slice comparison of esophagus, left image is raw CT scan, middle image indicated binary mask, and right image indicated the neural net prediction

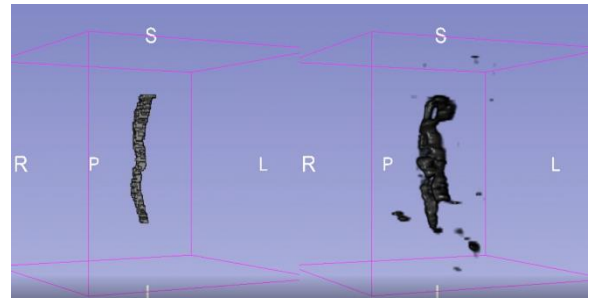


Figure 6 - 3-D view of esophagus with ground truth on left, and prediction on right

B. Spinal Cord

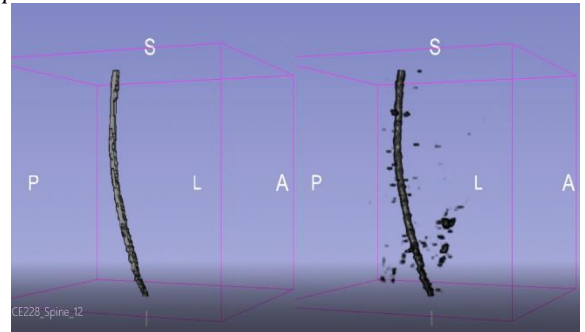


Figure 7 - 3-D view of spine with ground truth on left, and prediction on right

The spine prediction, although noisy in the sparse regions, predicted the volume and shape of the binary CT stack

very well. As seen in Figure 7, the curvature and volumetric consistency was tracked well over the stacks. Figure 8 shows the loss function over 4 epochs, with an end loss of less than 1 percent for both the training and validation set. The average spine Dice coefficient was 75.5 percent, which indicates the model intelligently predicting the spine, but with some false positives in each stack. It's worth noting that with some thresholding of the predicted stack, much of this noise can potentially be taken out. The currently produced results have no post-processing applied.

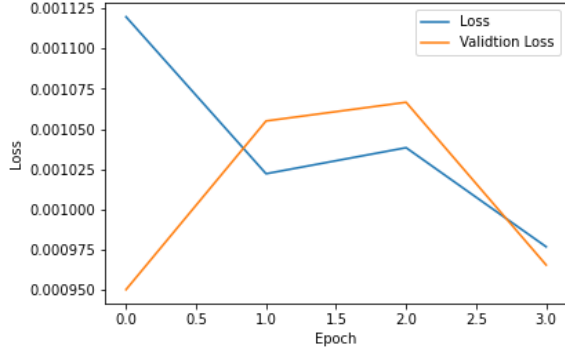


Figure 8 - Binary Cross-Entropy loss of spine training for 4 epochs

One interesting results produced by the spine can be seen in Figure 9. Because the 2017 competition only cared about segmenting the thoracic region, the spine segmentation stopped after reaching the neck area. Our model, however, was able to understand that the spine continued and kept predicting the correct spinal region into the neck. This obviously brought down the Dice score, but was a good indication of intelligent predicting.

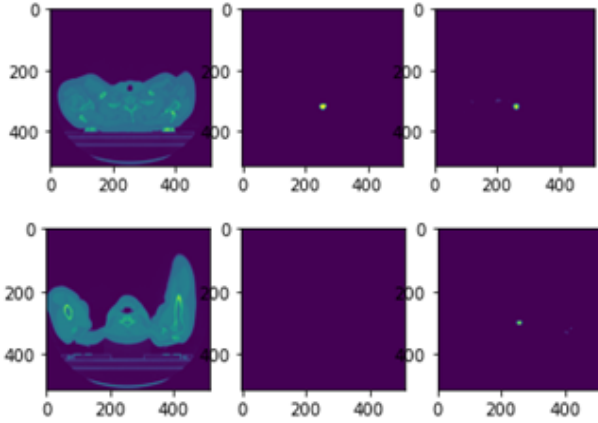


Figure 9 - Slice comparison of spine. left image is raw CT scan, middle image indicated binary mask, and right image indicated the neural net prediction. Above three images are showing the model correctly predicting spine, and below 3 images show our model prediction

C. Lungs

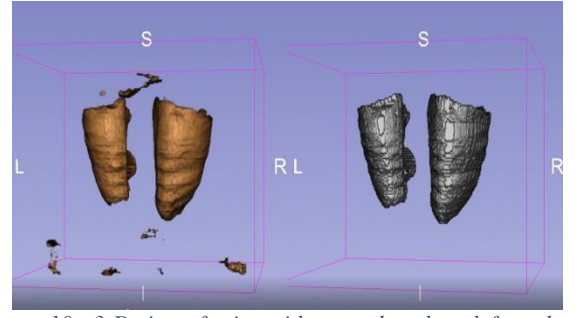


Figure 10 - 3-D view of spine with ground truth on left, and prediction on right

The lungs were well predicted with a loss of less than 5 percent. Similar issues with regard to volumetric noise can be seen in Figure 10, but otherwise the general shape and space was predicted well.

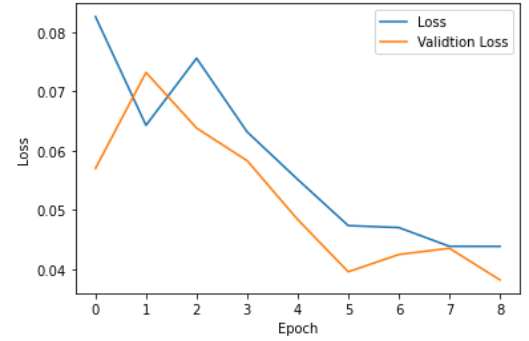


Figure 11 - Binary Cross-Entropy loss of lung training for 9 epochs

Table 1 - Dice Comparison of Results versus Paper [1]

	Dice (%) - Results	Dice (%) - Paper
Lung	95.2 +/- 5.6	95.6 +/- 1.9
Heart	89.3 +/- 9.5	93.1 +/- 1.5
Spine	75.5 +/- 7.3	86.2 +/- 3.8
Esophagus	8.8 +/- 11.4	81.8 +/- 3.9

VI. CONSLUTION

In conclusion, we were able to produce similar results to those in the AAPM competition for 3 of the 4 organs. It should be noted, that due to limited computational resources, our code had to be adjusted accordingly.

For future experiments, we would like to try consolidating all the binary masks and using one neural network model that outputs a one-hot-encoded results with a channel for each organ. This would work to address the problem of class imbalance and would also attribute the sparse space that is otherwise treated homogeneously to other organs. Additionally, we would like to test with other loss models as well as number of epochs to improve our results.

REFERENCES

- [1] Yang, J., Veeraraghavan, H., Armato, S. G., Farahani, K., Kirby, J. S., Kalpathy-Kramer, J., ... Sharp, G. C. (2018). *Auto-segmentation for Thoracic Radiation Treatment Planning: A Grand Challenge at AAPM 2017. Medical Physics*. doi:10.1002/mp.13141
- [2] Ronneberger, Olaf, et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241., doi:10.1007/978-3-319-24574-4_28.
- [3] Çiçek, Özgün, et al. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation." *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 Lecture Notes in Computer Science*, 2016, pp. 424–432., doi:10.1007/978-3-319-46723-8_49.
- [4] Zhou, Xiangrong, et al. "Automated segmentation of 3D anatomical structures on CT images by using a deep convolutional network based on end-to-end learning approach." *Department of Computer Science and Engineering, University of South Carolina – Song Wang*, 2016

ACKNOWLEDGMENT

A. Will Argus

Contributed to the literature search as well as to the 3D data visualization, and helped define some of the methods used in the final model.

B. Harmeet Gill

Helped with code prototyping and testing different functions and data processing. Contributed in label generating code and final result generation.

C. Karandeep Singh

Created most of the final code after consolidating the coding and model experiments, and contributed to all other aspects of the project as well.

Responses to Critical Reviews

The authors of this paper would like to thank Groups 52, 58, and 80 for their constructive criticism and the help it provided in improving this work.

Critical Review from Team 52:

“The motivation and background behind the project are made clear immediately. The presentation use images well in conveying specific learning points. Arguments regarding methodology behind the design of the neural network are broken down to smaller chunks to be easily understood. The arguments are well reasoned.

There are a lot of information regarding loss and how it affects segmentation results. It would do well to put the segmentation images to be in their own slides. It will give more room for the presenters to explain the learning points regarding the loss and have a good transition to the resulting images from changing training methods.

The formatting of images is all over the place. Images from various result sections sometimes are all at bottom right and sometimes they are from top right to bottom. Again, it would be better for both audience and presenter if few more slides are dedicated to just images and graph without text. As of now, there can be too much information in one slide.

Overall, there are plenty of good information from the beginning background to the results. The methodologies on preventing overfitting and improving the training process are well thought. The message and concepts are conveyed well and clearly, but some slides towards the end that are not as well formatted.”

The authors' response: The positive comments about the well-reasoned arguments are appreciated. With regards to the critiques, placing the segmentation images on their own slides is a good suggestion. These authors felt that by placing the segmentation images on the same slide as the loss explanation, it would help justify the loss used. However given that it detracted from the explanation of the loss, they will likely be moved to their own slides in future presentations. The formatting of images was done with the goal of placing relevant images next to the information so the images could be viewed while the information was discussed in the presentation. In future presentations, the images will be placed in the same spots for ease of viewing for the reader. As per the final suggestion, the slides at the end can be reformatted in future presentations in order to keep consistency and make it easier on the reader.

Critical Review from Team 58:

Great job with the presentation! It thoroughly explained the dataset, model, ground-truth reference. The results are great too.

Improvements -

1. In medical imaging a 512x512 resolution image can be thought to have great quality. Reducing the resolution and dimensions of input may help in reducing the total trainable parameters allowing more samples to be handled in a batch and may also reduce the training time without much effect on performance.
2. To attribute low performance on esophagus segmentation to either model or dataset, you could train a separate model to demarcate just esophagus. If the performance remains poor, it is probably a dataset issue and little preprocessing that exaggerates features of esophagus (texture/contrast) may be used.

Unclear -

3. Does the architecture of U-net use pretrained weights or does transfer learning?
4. Choice of batch size seems to be low, Is there a specific reason?
5. Data shuffling - The data set is said to have 60 scans with each having multiple slices. From the model parameters during rundown it appeared that each slice of 3D scan is taken as a separate input. Is the shuffling done by considering sets of images as a single unit?
6. Based on the application the CT image would be used for, the accuracy requirement of demarcation may defer. Say, if it was to understand perfusion in heart an underestimate of the volume of heart may call in for unnecessary clinical intervention. Does the dice coefficient differentiate overestimation and under-estimation of organ volume?

The authors' response:

The positive comments regarding the explanation of the dataset, model, and results are appreciated.

1. With regards to reducing the quality of the image to reduce the training time, it is a good suggest and the authors plan to look into doing that. This would allow for larger batch sizes and potentially faster training with more data.
2. The esophagus problems may be partially solved with a separate model, but the authors also postulated that the esophagus is the most complex and difficult of the organs to segment. Compared to the others, there is very little clear denomination between the esophagus and the surrounding tissue to enable the model to accurately predict it. For this reason, its segmentation will likely be inferior to the others no matter what method is used.
3. The architecture does not use pretrained weights nor transfer learning. They are randomly initialized by TensorFlow.

4. The small batch size is used due to the size of the images. 5 images was the maximum we could consistently use to run the model without running out of GPU memory.
5. To elaborate on how the data is parsed:
The first 15 complete volumes are maintained as complete volumes for testing
Remaining 45 are used for training and validation.

These are randomly shuffled on a slice basis. Slices for training and slices for validation both may come from the same volume

Final testing is however done on the complete volumes and a dice score is calculated using an entire volume.

6. No, the Dice coefficient cannot differentiate between overestimation and underestimation since it only compares the intersection of the prediction and the true to the sum of their areas. This is a good thing to point out and this weakness was one of the reasons the Intersection over Union loss function was considered and is talked about in the presentation as well.

It is also worth noting that myocardial perfusion imaging requires a contrast agent to image vasculature. Our model is not trained in any capacity to see that kind of data. Similarly, they probably would not scan a patient's entire torso to examine the heart.

Critical Review from Team 80:

Group 43 successfully achieved the image segmentation of thoracic organs via the U-Net deep learning model and also did a great presentation about how they analyzed the challenge, how they got the dataset and how they did their machine learning experiment. The logic flow of their presentation is clear and well-organized. They first gave a detailed background of TC technology and a sufficient survey of related research. Then an informative introduction of the dataset and U-Net model was provided, which makes a lot of sense to reviewers. Finally, they presented the results of their U-Net model and did a clear comparison with the literature.

We think there are some improvements which can be done to make this work better:

- Some texts on the figures are too small to recognize, it can be better if authors can organize the axis, labels and titles of the figures in a more convenient way.
- For the results part, we think the authors can show loss curves with more epochs so that the readers can learn more information about the entire training process.
- It would be better for the audience if the video voice could be louder and clearer. Overall, group 43 did an excellent job on the image segmentation problem and provided a great presentation.

The authors' response: The positive comments about the research survey, U-Net explanation, and results presentation are appreciated.

- With regards to the text on the figures, the authors plan to make it larger and more uniformly organized for future presentations.
- With regards to showing more epochs on the loss curves, we terminated training at the end of what was shown in the figures. From prior testing, we knew continued training would cause over fitting.
- Additionally in the future when recording a presentation, the authors will use a better audio recoding set-up to make the voices clearer.