Deep Learning Approaches for Image Segmentation with Lung CT Scans

Dimitrios Gotsis (gotsis2), Francis Roxas (roxas2), Harsh Agrawal (hagrawa3)

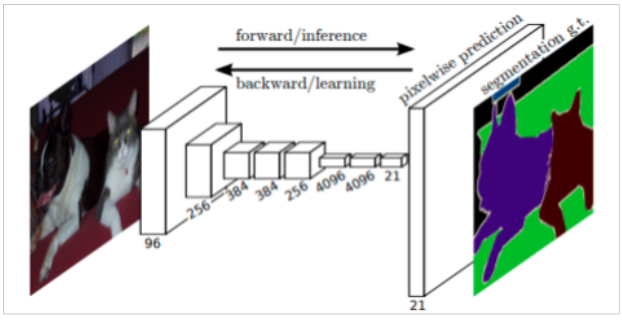
1 Introduction

The last few decades have seen huge advances in biomedical imaging and biomedical devices. This has allowed for doctors and medical practitioners to gain further insight into patients' medical conditions saving countless lives. Additionally, recent advancements in computational capacity, data collection and machine learning have seen the biomedical imaging field take a much more computationally and machine learning driven approach to its problems. Specifically, major advancements in machine learning such as deep neural networks (DNNs) have become ubiquitous in a variety of fields, biomedical imaging included. These DNNs have improved results for classification problems, image enhancement, image segmentation, etc. For our problem, which we will discuss in detail below, we wish to leverage such new approaches for image segmentation for biomedical images.

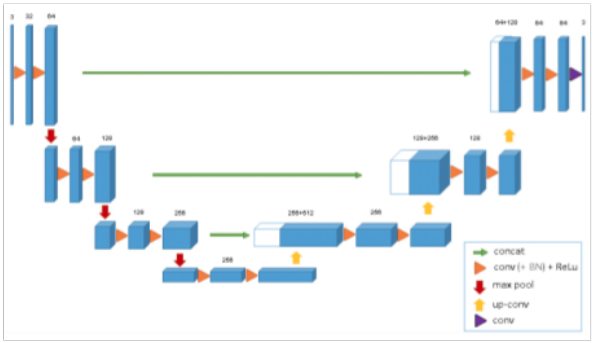
2 Background

2.1 Image Segmentation

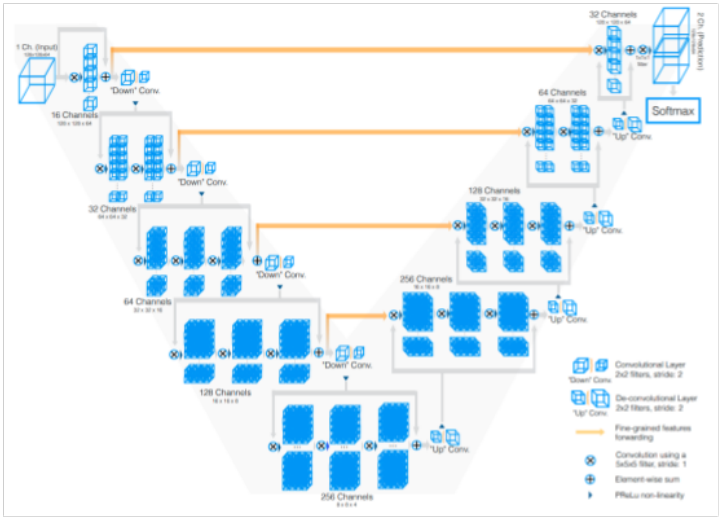
Image segmentation is a process of partitioning an image into regions or segments that look similar. Semantic image segmentation is the process of partitioning an image into meaningful segments, for example grouping all pixels belonging to the same object into a segment. This consists of predicting a class label for each pixel in an image. Image segmentation is typically used by medical professionals for work such as computer-aided diagnosis and patient treatment plans. Common approaches for image segmentation include clustering (KNN and WPCA) and architectures such as FCN [1], 3D U-Net [2] and V-Net [3]. FCN has an encoder which applies convolution layers followed by pooling repeatedly and finally upsamples to original image resolution to output a mask. 3D U-Net is similar to FCN except it has a decoder path which upsamples step-by-step and has skip connections which help recover any spatial information lost during downsampling. V-Net is similar to 3D U-Net except it has residual connections which helps solve the gradient vanishing problem.



(a)



(b)



(c)

Figure 1: Common architectures for image segmentation. (a) FCN (b) 3D U-Net (c) V-Net.

2.2 Generative Adversarial Networks (GANs)

One modern architecture model for DNNs is a generative adversarial network (GAN). These models are known for creating hyper-realistic images for a variety of tasks, from image enhancement to data augmentation. GANs are made up of two networks during training, a generator and a discriminator. The generator takes in as an input a random noise image and tries to output an image from the data distribution. On the other hand, the discriminator takes in an image from either the dataset or the output of the generator and determines whether the image is from the dataset or not. Thus, the two networks are trained to compete against each other. When training is complete, the discriminator is discarded and only the generator is kept. [4] Additionally, one commonly used variant is the conditional GAN (cGAN). The only main difference with a traditional GAN is that the cGAN takes an input which pairs with an appropriate output from the dataset. The cGAN may also take a random noise image as an additional input if necessary. [5]

3 Problem Statement

1. We will compare several deep-learning approaches to segment CT imaging volumes of the chest for the COVID-19 Challenge [1].
2. We will use GANs to improve the performance of the best architecture for this problem and show that it captures finer details and enforces spatial contiguity.

3.1 Method

We began with experimenting with simple approaches learned in this course such as K-Means and K-Nearest Neighbors to familiarize ourselves with the dataset. We improved this result using deep learning methods, specifically we experimented with popular image segmentation architectures such as FCN [1], 3D U-Net [2], V-Net [3]. We converted the 2D FCN to handle 3D volumes by replacing the 2D convolutional layers by 3D layers.

These techniques have been known to occasionally produce coarse or non-contiguous segments which we confirmed by our experiments. This has been improved by applying post processing techniques such as Conditional Random Fields (as used in Deeplab). However, these techniques are slow and cannot be easily trained end-to-end. Newer methods have used GANs to improve the model, where they train a segmentor as the generator and add a discriminator network to predict whether a given segmentation result is ground truth or synthetic (generated by the segmentor) as shown in Figure 2. This approach has shown to enforce spatial contiguity and can be trained end-to-end. We used the training set up and discriminator from Cirillo et al. [6]. This approach has a PatchGAN discriminator which produces a class label (1=ground truth, or 0=predicted) for each patch. This helps penalize the mask at the scale of local patches.

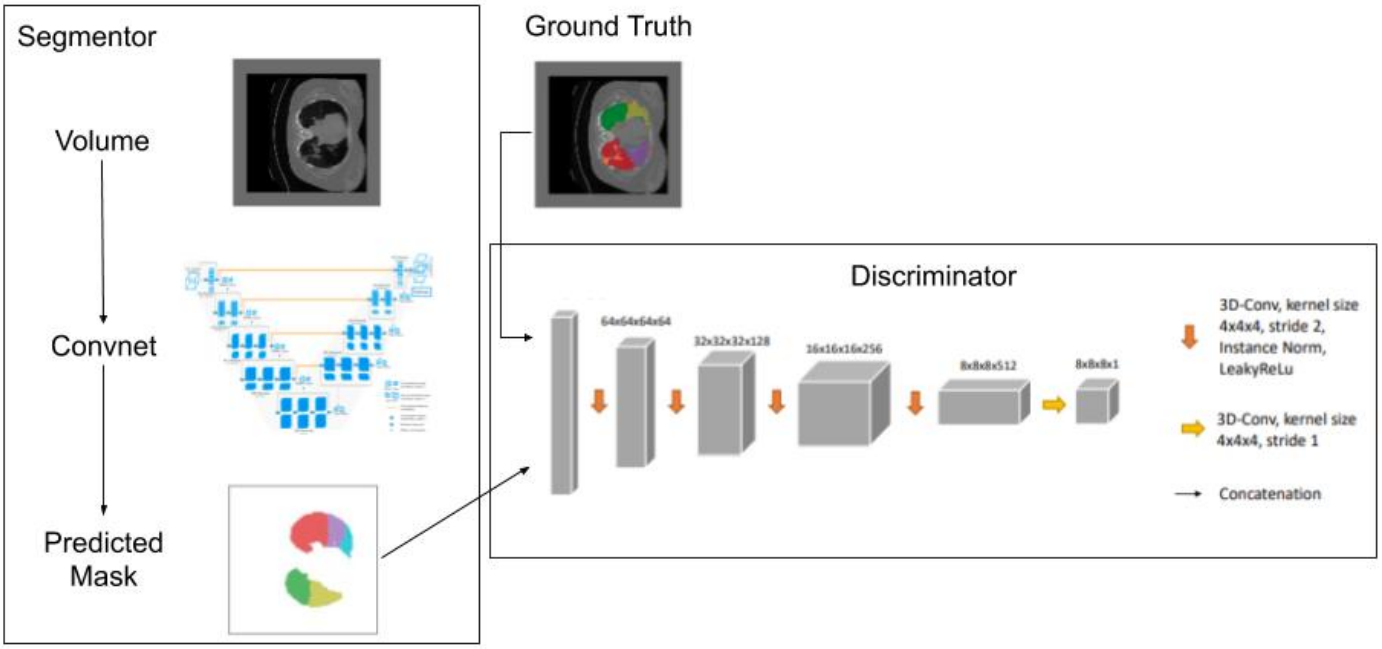


Figure 2: Overview of our approach. Left: segmentation net takes RGB image as input, and produces per-pixel class predictions. Right: Discriminator takes ground truth and predicted label map as input and produces a patch-wise class label (1=ground truth, or 0=predicted) [6].

[1] covid19challenge.eu

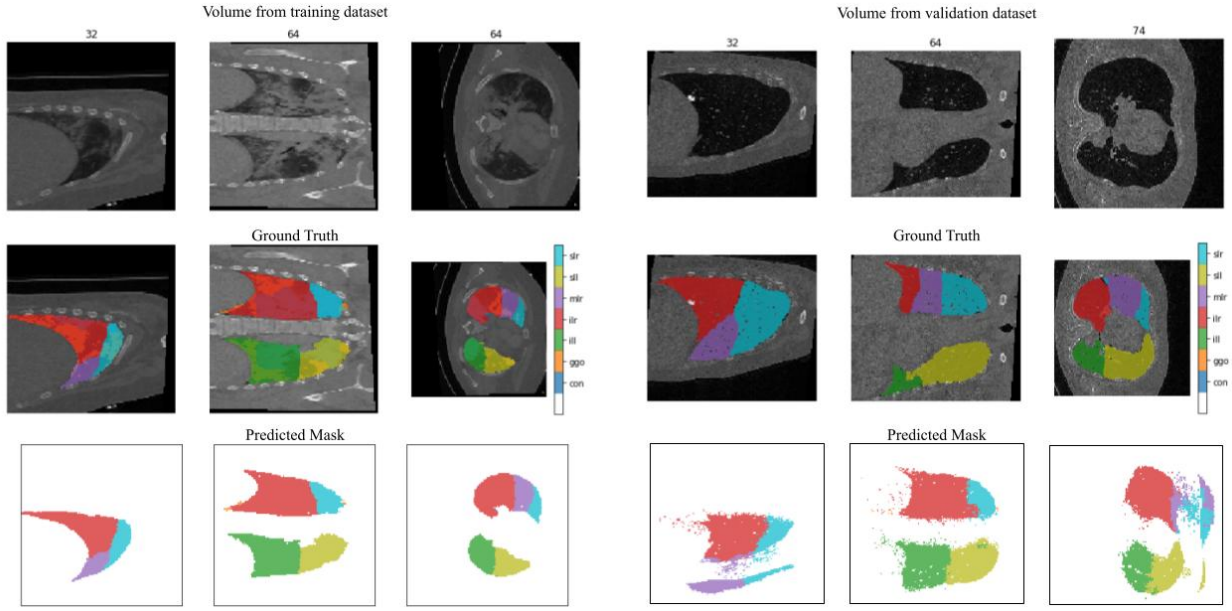


Figure 3: Results from KNN with 10 nearest neighbors and pixels with their coordinates as the feature vector.

3.2 Dataset

We will use the dataset provided by the COVID-19 Challenge. This dataset consists of 113 segmented CT imaging volumes of the chest, 79 from COVID-19 cases and 34 from non-COVID-19 cases. The labels for each pixel are provided in multi-hot encoding for five lung lobes and two lesions types (consolidation and ground-glass opacities). The original images were 512×512 times a variable dimension. For simplicity all images were resized to $128 \times 128 \times 128$. Figure 3 shows the volumes and their annotated labels.

4 Experiments and Results

We trained a K-Nearest Neighbors classifier with 10 neighbors using the pixel values and their coordinates as the features. Figure 3 shows the results from this approach. This approach was computationally expensive and did not generalize well to unseen data. Therefore, we decided to explore learning based methods.

As mentioned above we have partitioned our experiments for a deep learning based approach into two parts. First we trained our neural network models (FCN, U-net and V-net) in an end-to-end fashion, i.e. the output from the networks is directly compared to the ground truth. For training we used Adam Optimizer with a learning rate of 0.0002 and trained for a total of 50 epochs with batch size 1.

Since metrics like mean squared error (MSE) are not ideal for segmentation, for our loss function we used used the dice loss since we want an overlap metric between the guess and the ground truth. The dice loss can be described as follows:

$$DL(a, b) = 1 - 2\frac{\sum_i a_i b_i}{\sum_i a_i^2 + b_i^2} \tag{1}$$

Where a is the output from our model, b is the ground truth and i is the coordinate index. Note that we want to minimize the dice loss, thus the smaller the loss the better our result. As we can see in Table 1, V-net performed the best, followed by U-net and then FCN.

End-to-End Training Models			
	FCN	U-net	V-net
Dice Loss	0.34689	0.33250	0.32914

Table 1: End-to-End Training; Results on Validation Set

Adversarial Training Models		
	Adversarial U-net	Adversarial V-net
Dice Loss	0.33185	0.31393

Table 2: Adversarial Training; Results on Validation Set

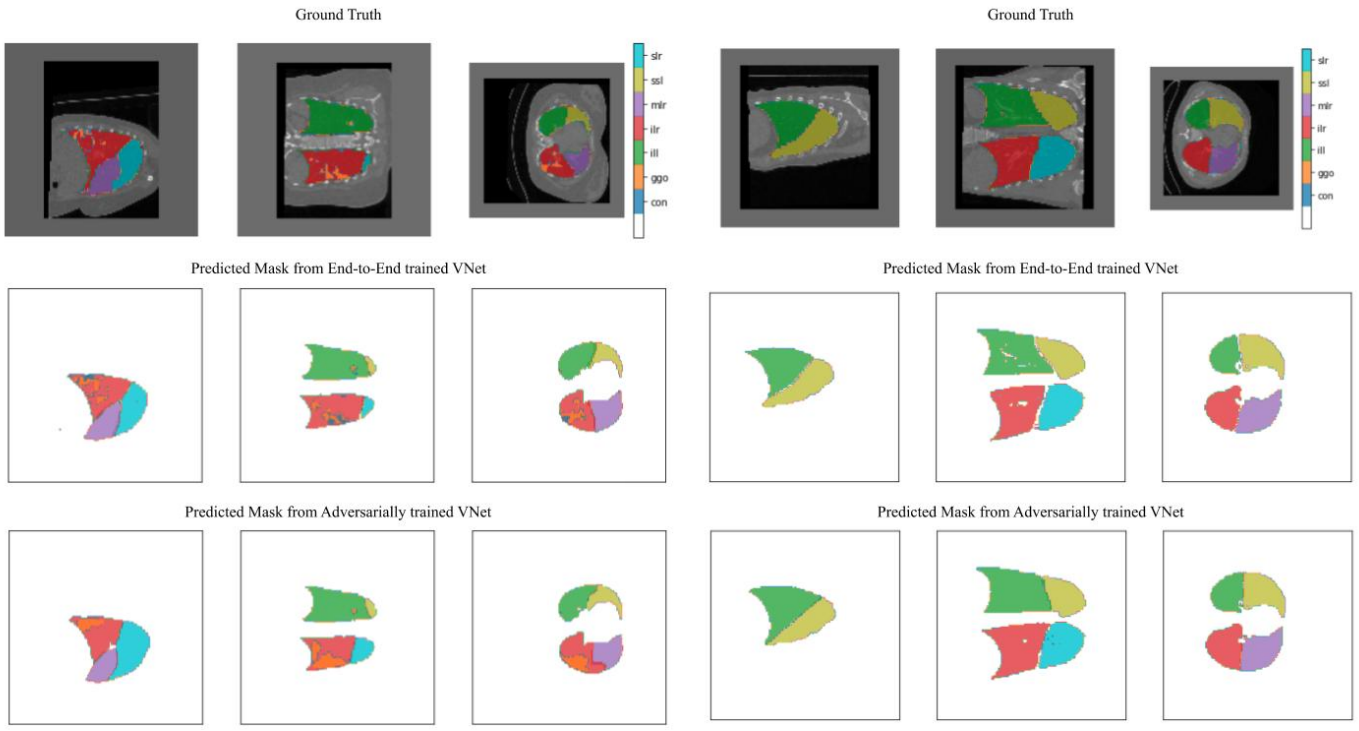


Figure 4: Comparison of results from VNet trained End-to-End and Adversarially. Left - The orange region is much smoother in the output from the adversarially trained model. Right - The output from the adversarially trained model has little to no holes compared to the end-to-end trained model.

Second we ran each of these models as generators in adversarial training. As we can see in Figure 4, the predicted mask from the adversarially trained model is smoother and has fewer holes or discontinuities. Table 2 also shows that the 3D U-Net and V-Net trained adversarially perform slightly better than the end-to-end trained models. It is worth mentioning that the adversarially trained FCN did not provide meaningful results and thus the results are not in the table above. However, for the 3D U-Net and the V-Net it seems that adversarial training is beneficial as it yielded results that looked closer to the ground truth.

5 Conclusion

In this project, we learned about state-of-the-art deep learning methods for image segmentation such as FCN, 3D U-Net and V-Net. We implemented these architectures and found that they often produce a mask with coarse or non-contiguous segments. To capture finer details and make the predicted mask resemble the ground truth, we trained these models using a GAN where we used this architecture as the generator and a PatchGAN as the discriminator. This approach allowed us to achieve a slightly lower dice loss and predictions that were smoother and had fewer holes. We found that it was difficult to tune the GAN to show performance improvements. Further experiments need to be conducted to verify the robustness of this approach. We look forward to applying this approach to other medical imaging modalities and hope to develop efficient, reliable, and robust methods that would be able to analyze scans for applications including diagnosis and robot-assisted surgery.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. arXiv: 1411.4038. [Online]. Available: http://arxiv.org/abs/1411.4038.

[2] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," *CoRR*, vol. abs/1606.06650, 2016. arXiv: 1606.06650. [Online]. Available: http://arxiv.org/abs/1606.06650.

[3] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *CoRR*, vol. abs/1606.04797, 2016. arXiv: 1606.04797. [Online]. Available: http://arxiv.org/abs/1606.04797.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML].

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, *Image-to-image translation with conditional adversarial networks*, 2018. arXiv: 1611.07004 [cs.CV].

[6] M. D. Cirillo, D. Abramian, and A. Eklund, *Vox2vox: 3d-gan for brain tumour segmentation*, 2020. arXiv: 2003.13653 [cs.CV].