Stroke Lesion Segmentation with 3D Convolutional Neural Networks

Haren Bhavnani, Margaret Zheng, Yang Xiao, Zhangyi Fan
University of Pennsylvania, USA

I. Introduction

Approximately 800,000 people in the United States suffer from a stroke each year, from which approximately 133,000 deaths occur. Furthermore, up to two thirds of stroke survivors experience long-term disabilities that can impair their daily life going forward. It is due to this possibility of long-term impairment that the procedures taken in both the acute stage, where early recovery can be promoted, and at the subacute/chronic stages, at which point long term recovery can be promoted. While there is evidence that different patients going through this subacute/chronic phase may be best suited to different rehabilitation measures, a barrier preventing research in the area is accurate legion segmentation.

The current gold standard for lesion segmentation is manual tracing, however this poses a large problem. Manual segmentation requires specialised tracers and can be extremely time consuming and subjective. While techniques that automate this process exist, the results are often not precise enough for use in research as a substitute for manual tracing. In this project, we design a deep learning model that has the ability to segment the voxels of the image into lesion and non-lesion classes, using 3D T1 weighted MRI images.

The uneven shape and location of lesions, coupled with lesion size variance between stroke patients pose as major obstacles to the segmentation process, as displayed in Figure 1. Unlike most image processing tasks which involve primarily with 2D images, our project handles 3D MRI scans. Thus we implement a 3D Convolutional Neural Network with 5 convolutional layers and 2 fully connected layers. Finally, when processing each voxel, we also feed in the symmetric voxel in order to utilize the quasi-symmetry property of the brain for our segmentation.

II. Related Works

Current research in the are of lesion segmentation is plentiful, with challenges such as the Ischemic Stroke Lesion Segmentation (ISLES) challenge that occur yearly. All current research in the area, however, use multimodal clinical MRI data. These such algorithms are not scalable to high-quality T1-weighted MRIs, which are conventionally used for subacute/chronic stroke rehabilitation research. To address this issue our model will use the recently released open source Anatomical Tracings of Lesions after Stroke (ATLAS) Dataset, which consists of 304 T1-weighted MRIs with manually segmented lesions. The closest papers to our work include [1], [4] and [3]. [4] and [3] both use 2D convolutional nets, while our method proposes the use of a 3D Neural Network for lesion segmentation. [1] is the first paper to use a 3D Neural Network for classification, once again however this is on the multimodal MRI data.

Our novel process, while using different data to [1], will also involve image contrast adjustment (since lesions are visible as darker areas on T1-weighted MRI images), and also will investigate the unique symmetry property proposed in [3] and its application on 3D Convolutional Nets.

III. Methods

A. Dataset Overview

ATLAS (Anatomical Tracings of Lesions After Stroke) is an open-source dataset of 304 T1-weighted MRI scans with manually segmented lesions and metadata. This provides a large, standardized dataset for comparing the performance of different segmentation methods.

The dataset includes:

- **304 T1-weighted MRI scans with lesion segmentation**
  For each MRI, brain lesions were identified and masks were manually drawn on each individual brain in native space using MRIcron, an open-source tool for brain imaging visualization and defining volumes of interest. To identify lesions, each T1-weighted MRI image was displayed using the multiple view option in MRIcron20, which displays the brain in the coronal, sagittal, and axial view.

- **229 MNI152 standard-space T1-weighted average structural template images**
  Lesion segmentation algorithms vary in whether the input should be in native space or a standardized space. The dataset above is provided in native subject space and a subset of the dataset was also defaced, intensity normalized, and provided in standard space (normalized to the MNI-152 template).

- **Lesion metadata file**
  For each lesion, the dataset also provides metadata on the lesion properties to give additional qualitative information, beyond the binary lesion mask. This information includes specific lesion characteristics as well as scanner strength, brand/model, and image resolution.

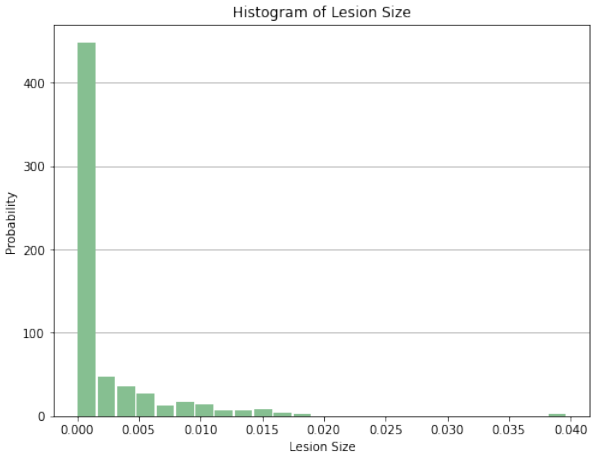

Fig. 1. Relative size of lesion across all samples

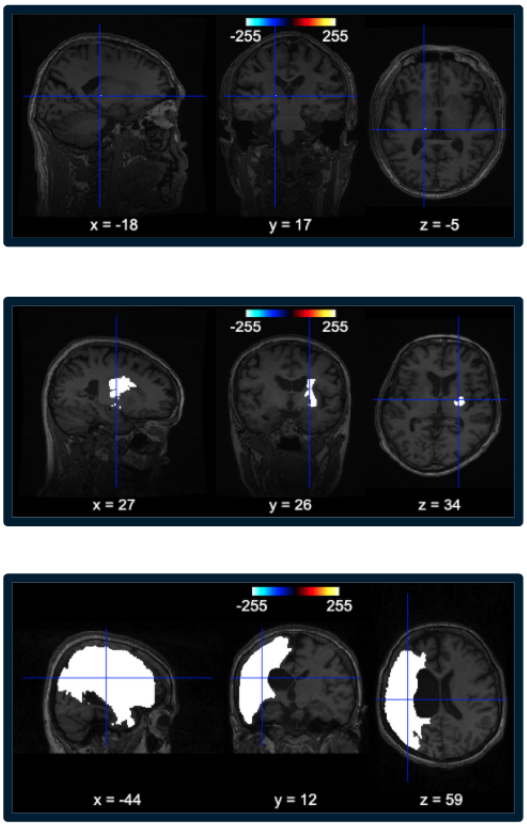

Fig. 2. Minimum, Median, and Maximum size Lesion

1) Exploratory Data Analysis: We begin with EDA to understand our dataset and objective. From the manually identified lesions in each scan, we can calculate the size as a proportion of the image itself. The distribution of lesion sizes is shown in Figure 1.

Figure 2 displays the smallest and largest lesions in the dataset, as well as the median lesion size. There is high inconsistency in lesion size and We anticipate that identifying smaller lesions may be a more challenging task than identifying larger ones.

We note that given the small size of many lesions, the majority of voxels are not part of the lesion. Specifically, only $\sim$ 2% of the voxels are in lesions. This dataset imbalance motivates our decision to upsample our training data.

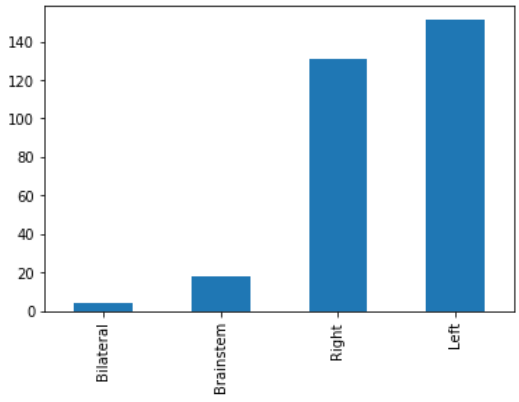

Fig. 3. Proportion of Lesions by Location

Lastly, we find that our dataset contains roughly equal representation of right and left hemisphere lesions, as well as some bilateral and brainstem lesions.

B. Preprocessing

In the dataset, we encountered some minor inconsistencies that required addressing. First, there were some images were the lesions mask and the brain scan mismatched in size. Since these images were incompatible for direct comparison, they were excluded.

Furthermore, the dataset contained images of varying sizes. To use these images as inputs to our models, we require that they are of the same size. So, during our Exploratory Data Analysis, we found the largest size for each dimension and padded all images to this largest size, 256 x 256 x 256, to ensure no loss of data. We padded with 0's, consistent with the background of each scan, along all three dimensions. This ensured that the brains still remained center on each image after padding.

We split our data into training, testing, and validation sets of sizes 100, 25, and 25 respectively.

Note that during our model building process, we downsample the size of the images and use much smaller training, test, and validation sets. This is to allow our models to be run locally without overloading RAM or taking an unreasonable amount of time. However, our model results are all based on the split mentioned above and our 3D model is constructed entirely using the full dataset.

As mentioned in the Exploratory Data Analysis, in the dataset, the target of interest (the lesions) only occupies a small portion of the entire scan, approximately $\sim$ 2% overall. Thus, most of the extracted features come from the healthy brain tissues and the background. This high class imbalance would make training more difficult and tempt our models to simply predict all zero's (non-lesion) for high performance. In order to avoid bias toward class 0 (healthy tissues and background), we deployed an epsilon method to avoid taking in too many non-target voxels but still have them represented

in the training data. With a probability of epsilon (0.01) we extract a class 0 point for training. We were able to upsample and create a balanced training set where class 1 makes up 43.7% of the dataset. However, having a model that can only classify well with balanced input would not be a useful model. With true unlabeled scans, the location and size of the lesion would be unknown, as would the proportion of lesion to non-lesion pixels. So, we validate our models on data that is not upsampled.

We also normalized the images to standard Gaussian distribution to improve the network convergence speed. For our advanced 3D CNN (final model), we replace this with histogram equalization.

C. Evaluation Metric

We assess the performance of our models using the Sorenson-Dice coefficient as the primary outcome measure (for training and testing). The Dice coefficient is a statistic to gauge the similarity between two samples by measuring the spatial overlap of the ground truth human expert-identified lesion segmentation and the model identified lesion, making 0 for no overlap and 1 for complete overlap.

The Dice coefficient is a principal evaluation measure for comparing the quality of lesion segmentation from a model to reference ground-truth and is a standard in stroke lesion research.

For two sets X and Y, which here would be the true and predicted lesion labels:

$$DC = \frac{2 \cdot |A \cap B|}{|A| \cap |B|}$$

Although we could have created a Dice loss function, research papers in this field tend to prefer Cross Entropy loss for its desirable properties. So, we opt to use Cross Entropy Loss.

D. Feature Engineering

Apart from the aforementioned pre-processing to standardized the images, we also employed various different feature configuration to supplement and improve our algorithmic performance.

1) *Adjusting Contrast*:

In order to enhance the properties and features of the images,

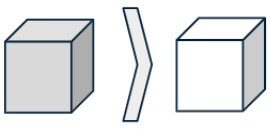

Fig. 4. Contrast Adjustment

we utilized histogram equalization on the images. histogram equalization is a way to augment images, whereby, it assigns the intensity values of pixels in the input image such that the output image contains a uniform distribution of intensities. Through such, the equalization improves the contrast and enhances the image overall.

2) *Spatial Information*:

As opposed to feeding in the model with the whole image

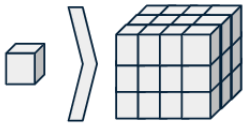

Fig. 5. Spatial Relevance

and expecting an output of the same size, we are instead treating this segmentation on a pixel by pixel (voxel by voxel) basis. Specifically, for every pixel/voxel of the image, we are obtaining all the relevant pixels/voxels that are within 12 pixel/voxel of the pixel/voxel of interest along with the true label of this pixel/voxel (In the 2D Case, this would yield a 25 x 25 square, while for the 3D case, this would yield a 25 x 25 x 25 cube). In doing so, we preserve the

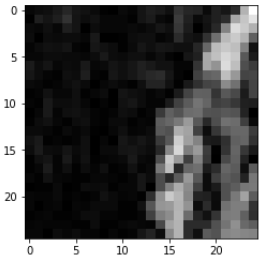

Fig. 6. Sample 25 x 25 Square

benefit of feeding in the model with the whole image, which is to account for the surroundings of every pixel and making contextualized prediction, but we also allow for the model to concentrate on a local region to perform localized prediction. Combining these two potential benefit, we would expect such reconfiguration to yield better result. It is also worth mentioning that for our non deep learning baseline, such a localized approach is still used, whereby we simply took the 25 x 25 square for the 2D case and flattened it to a 625 sized 1D vector.

3) *Symmetry*:

While recent research suggests that human brains are not

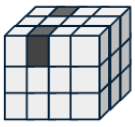

Fig. 7. Symmetry

entirely symmetrical, the quasi-symmetry properties of brain can nonetheless be exploited in creating more informative feature space for our model. By including a supposedly symmetrical and adjacent healthy side of a brain, cube B, in the categorization of its counter part that is damaged by

lesion, cube A, the algorithms will likely be more capable of spotting anomalies that exists in cube A in comparison to the healthier parts of the brain. To accomplish this, we, in particular, extracted adjacent 25 x 25 x 25 cubes in two sides of the brain sliced from the top angle and appended them together, where by the cube of interest is positioned before the adjacent cube, to form a final cube with dimension of 50 x 25 x 25.

E. Non-Deep Learning Baseline

We create a baseline model to compare our advanced approaches against, without using any deep learning techniques. Deep learning is the preferred and more traditional approach for class based pixel-wise segmentation of images. Researchers have found reasonable success with pixel-wise segmentation using Random Forest [2]. So, we use Random Forest as our baseline model, classifying each pixel as being either part of a lesion (1) or not part of a lesion (0). We flatten the 25 x 25 2D square representation into a 625 1D vector which we feed in as input into our model.

We tune the Random Forest Classifier with a Grid Search of for number of estimators, max depth, and the minimum number of samples required to split an internal node. We settle on a model with 50 trees that have max depth of 6 and minimum samples required to split at 2. Despite extensive tuning, the Random Forest Classifier performs poorly on this task, predicting 0's for most pixels. On the test set, the Dice coefficient is 0.015.

F. 2D Deep Learning Baseline

We choose to use Transfer Learning with a Residual Network or ResNet for our baseline deep learning model. Since we have a limited amount of data, we choose to use a strong, pretrained model as we expect it will have better performance. We use the Pytorch's ResNet-18, which is pretrained on ImageNet. We swap out the last fully connected layer of the network with our own classification layer and train only the last layer. For this model, we take the 25 x 25 2D representation for every pixel and its surrounding as input. After tuning the learning rate, we select 1e-5 learning rate and Adam optimizer. The ResNet significantly outperforms the baseline, with Dice coefficient 0.146 on test set.

G. 3D Deep Learning Approach

For our next approach we chose to implement a 3D Convolutional Neural Network. The structure of our network involved 5 Convolutional layers, the first two of which were followed by a max pooling layer, and then 2 fully connected layers. Our first 6 layers were followed by the ReLU activation function, with a Softmax being used as the non-linearity after the final fully connected layer. Our approach to constructing this network involved tuning both the number of layers as well as the number of channels in each of the convolutions layers. We tried combinations of 3,4, or 5 Convolutional layers with 2 or 3 fully connected layers and found that the choice of 5 Convolutional layers and 2 fully connected layers performed best on our validation set. We chose the Adam optimizer with a learning rate of 1e-5 (we experimented with 1e-4, 1e-5 and 1e-6). The Ultimately on the test set, the 3D CNN outperforms both previous models, with Dice coefficient 0.389 on the test set.

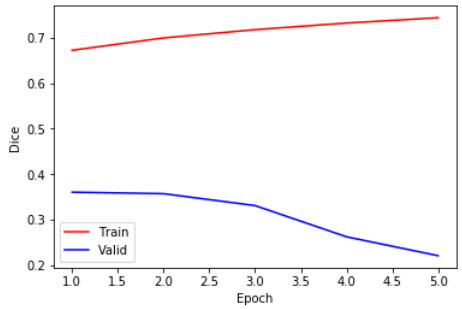

Fig. 8. 3D CNN Dice Coefficient Curve

Furthermore we chose to train for a single epoch, as we noticed that the model seemed to overfit with multiple epochs, which can be seen in the 3D CNN Dice Coefficient Curve.

H. Advanced 3D Deep Learning Approach

For our advanced featurization approach, we chose to use the 3D CNN from the previous subsection as it had the best performance on our test dataset. For this approach we implement our novel method based on the inherent quasi-symmetry property of brains to construct more advanced features that would fully represent the important properties in the data to help with classification. After experimentation we chose the same learning rate as we did with the base 3D model (1e-5), and once again we found that the model overfit after 1 epoch, as can be seen in the Advanced 3D CNN Dice Coefficient Curve.

We also perform histogram equalization in order to emphasise the contrast between lesion and non-lesion voxels. A further description of these two methods can be found in section III.D.3 and III.D.1. Use of these methods improved our dice score on the test dataset by 0.11, from 0.389 to 0.499.

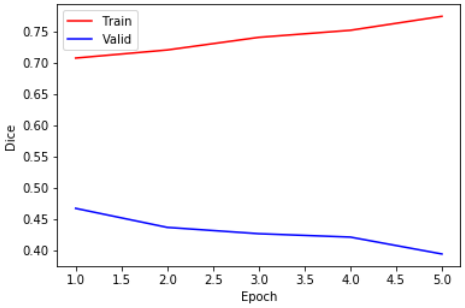

Fig. 9. Advanced 3D CNN Dice Coefficient Curve

Models	Testing Dice Coefficient
Non Deep Learning Baseline *(Random Forest)*	0.015
2D Deep Learning Baseline *(Transfer Learning)*	0.146
3D Deep Learning Models *(3D CNN)*	0.389
Advanced 3D Deep Learning Models *(3D CNN w/ Advanced Features)*	0.499

Fig. 10. Models Performance Metrics Comparison

IV. Analysis

A. Non Deep Learning Baseline

We chose the Random Forest as it is known to excel in accuracy among current machine learning algorithms. So it serves as a good baseline for comparison with deep learning models.

It was expected that Random Forest would not be able to capture the complexity of the data as it does not take spatial information into consideration. Also due to the sheer volume of data (billions of data points) we have, Random Forest cannot offer enough granularity that the deep Neural Nets can provide with much more parameters. Furthermore, Random Forest or for that matter, any non deep learning models is incapable of dynamically processing and transforming the features to represent and uncover latent pattern in visual images. This largely constrains the algorithm to only the static features at hand, limiting its accuracy.

B. 2D Deep Learning Baseline

For deep learning baselines we experimented with applying transfer learning using a pre-trained 2D ResNet (Residual Network) model.

The 2D network takes in an input of 25 x 25 square and is trained to predict the label of a central pixel according to the content of surrounding patch. 2D network provides richer spatial information of neighboring pixels in one slice. The richer spatial information that the 2D network provides is crucial to lesion segmentation.

Transfer learning often proves to be successful because it is capable of borrowing the knowledge learned from previous sources (ImageNet) and transferring the network parameters from a pre-trained network performs better than random initialization. ResNet has deep structures and we wanted to reduce the convergence time and computational load so we froze the model except the last layer during training to stop any updates on the pre-trained weights.

Although ResNet exceeded the performance of non deep learning baseline, the result was still not satisfactory. The reason why it did not do well lies in the fact that the sources ResNet trained on are far away from our target. Biomedical images have very different appearance and size so transfer learning from models trained on ImageNet did not offer enough representational ability.

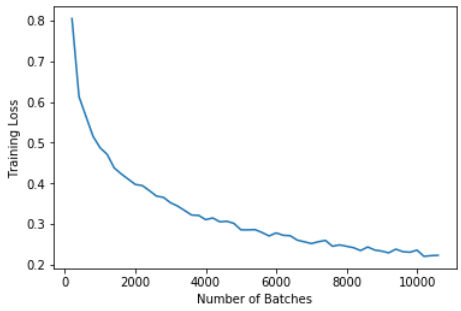

Fig. 11. 2D ResNet Loss Curve

C. 3D Deep Learning Approach

The 3D CNN is trained to predict the label of a central voxel according to its surrounding volumes. There is important spatial information in the third dimension (the one that would be sliced over for 3D Convolutions), and that this information will be useful for classification. 3D offers the most complete spatial information of neighboring voxels. It leverages comprehensive information across all three dimensions rather than just having one in the 2D approaches.

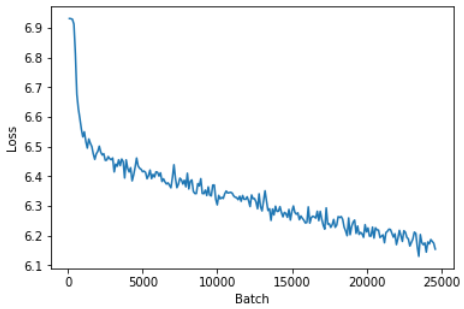

Fig. 12. 3D CNN Loss Curve

Note: We scaled up the magnitude of our loss to increase interpretability.

D. Advanced 3D Deep Learning Approach

We Believe the Advanced 3D Deep Learning Approach proved most favorable for the following reasons.

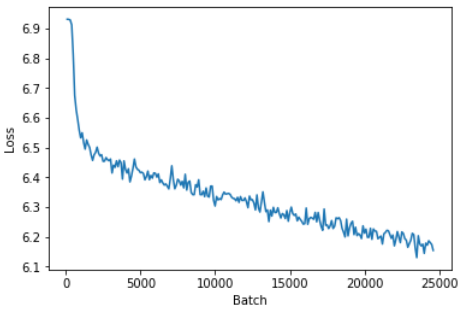

Fig. 13. Advanced 3D CNN Loss Curve

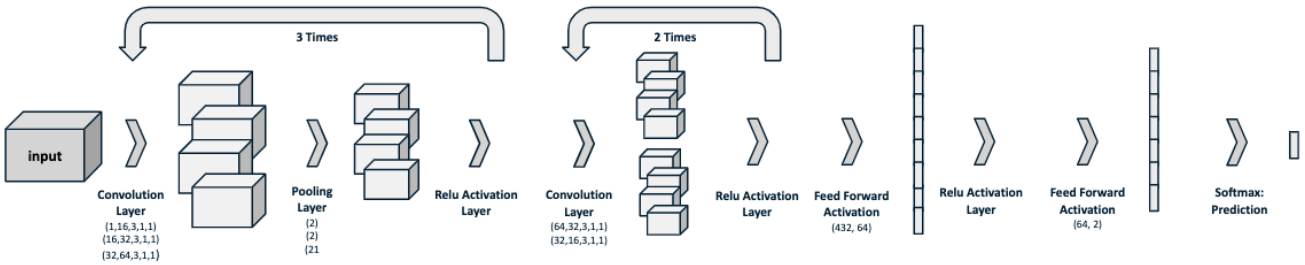

Fig. 14. 3D Convolutional Neural Network Architecture

- **Symmetry**
  We leveraged the quasi-symmetry property of the brain to aid the classification task. This inherent symmetry property breaks down when a lesion is present in a hemisphere of the brain. We appended the mirrored voxel representation to the end of each feature so this additional information would lead to a more accurate segmentation by the model.
- **Histogram Equalization**
  The 3D model also built on traditional methods such as histogram equalization. The ambiguous boundary with a small contrast between lesions and the neighboring brain tissues proposes a big challenge. Performing histogram equalization on the images increased the variance between the target and neighboring voxels thus improving the result of classification.

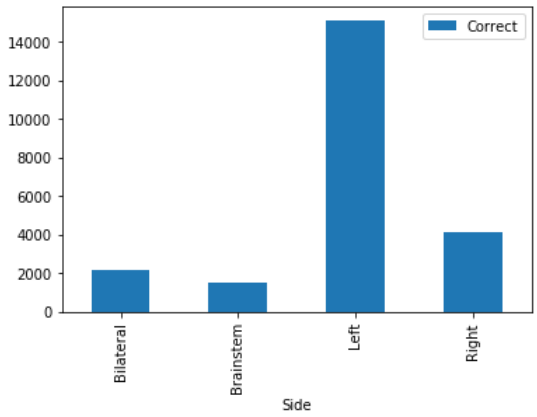

Fig. 15. Proportion of Correct Classifications by Lesion Location

We notice that our model labels disproportionately more of the left hemisphere lesion pixels than the right hemisphere lesions correctly. This is not due to class imbalance, recall that Figure 3 showed that our dataset actually contains equal proportions of right and left hemisphere lesions. There is no clear explanation for this phenomenon. However, this does help us identify where we can improve our model further.

E. Advanced 3D Network Sample Prediction

Figure 16 displays a sample of a prediction of lesion compared to the actual true lesion. The model identifies the general area of the lesion mostly correctly, although it seems to mislabel some pixels within the lesion.

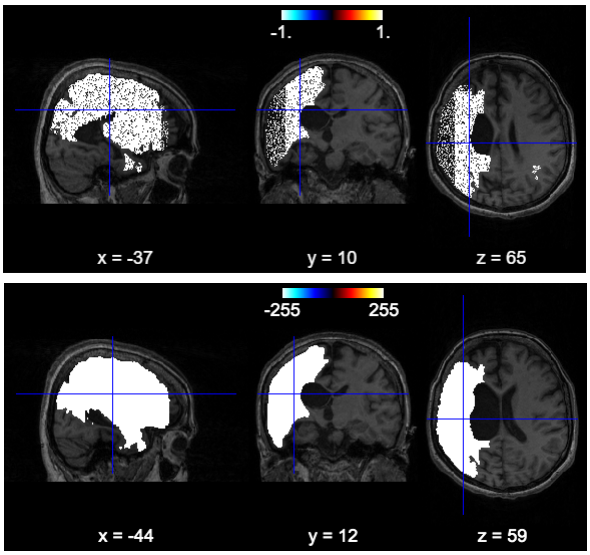

Fig. 16. Top: Predicted Lesion, Bottom: Actual Lesion

V. Discussion

Although our advance model does outperform all the baselines, there is still room for improvement. We expect that with some adjustments, we could be able to improve the dice coefficient. If we had more time, we would try the following changes:

A. Using Standardized Dataset

Provided alongside the dataset of 304 T1 weighted MRI images is an additional set of 229 Standardised MRI images. These images came from the initial dataset, however they were processed using defacing, intensity normalization, and normalization to a standard space (MNI-152 template). Initially we had chosen to use the non-standardized images due to the fact that it gave us access to a larger quantity of data, (it is mentioned in the documentation that ”Due to technical difficulties and differences in scanner image quality, a subset of brains is not included in the standard space conversion”), however since computational limitations led to us only being able to use a subset of the data anyhow, an interesting experiment would be to attempt our process on these standardized images and observe the results.

B. Other Datasets

Ideally, we would have tested our model on datasets other than ATLAS. In research papers, the performance of the same model varies significantly across different datasets, (in the case of [1], their results ranged from 0.35 to 0.85 on different datasets while using the same model). However, we were unable to obtain access to another data source, since most available datasets for Lesion Segmentation are only released to the public during active competitions and access is revoked once the competition ends.

C. Mask R-CNNs / Taking a Mask Approach

Our 2D Deep Learning Baseline involved the use of a pre-trained ResNet in order to make the pixel-wise classification. Our approach included fine-tuning the ResNet by only training the last layer of the network to act as a 'feature extractor' for our task. Another possible approach that could have been taken is the use of a Mask R-CNN for this task instead of a ResNet. Mask RCNNs are regarded as the state of the art in image segmentation, and therefore could have produced better results than the ResNet. Similarly to how we used a pretrained ResNet we could have also chosen to use a pretrained Mask R-CNN, specifically one trained on the MS COCO dataset, and fine-tuned that model for our task. The Mask R-CNN takes in an image as input and returns a class label, bounding box coordinates for each object, and the object mask.

So, using this model would be a different approach then the one we take. We have chosen to treat this segmentation process as a pixel-by-pixel classification of lesion or non-lesion. However, we could generally have approached the problem differently, aiming to generate a mask and bounding box for each lesion, rather than classifying pixel by pixel. Given that Figure 16 demonstrates some difficulty on the part of our model for identifying pixels clearly within a lesion, a mask type model may have had more success.

D. Improve Performance on Right Hemisphere Lesions

As discussed in analysis, our model performs much better with left hemisphere lesions than right hemisphere lesions. To combat this shortcoming, we could upsample right hemisphere lesions in training.

Overall we believe our model and analysis for this project proved generally successful in attaining reasonable result that may help automate and more accurately segment lesions. Through this and echoing our overall goal, we hope to advance further discovery in the space of stroke related research.

References

[1] M. Havaei. "Multiple Sclerosis, Stroke and Traumatic Brain Injuries". In: *Lecture Notes in Computer Science* 9556 (2016).

[2] F. Schroff, A. Criminisi, and A. Zisserman. "Object Class Segmentation using Random Forests". In: (2016).

[3] K. Raina, U. Yahorau, and T. Schmah. "Exploiting bilateral symmetry in brain lesion segmentation". In: (2019).

[4] K. Kamnitsas. "Multi-Scale 3D Convolutional Neural Networks for Lesion Segmentation in Brain MRI". In: *Medical Image Analysis* 46 ().