3D Supervised Learning for CT Hematoma Segmentation via Transfer Learning from a 2D Trained Network

By

Zihao Wu

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Electrical Engineering

May 31, 2020

Nashville, Tennessee

Approved:
Bennett A. Landman, Ph.D.
Yuankai Huo, Ph.D.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS ii

LIST OF FIGURES iii

LIST OF TABLES iv

Chapter

I. Introduction 1

1. Traumatic Brain Injury 1
2. Hematoma Segmentation 1

II. Related Works 3

1. 2D Trained Network 3

III. Materials and Methods 4

1. Data Collection 4
2. Prediction using 2D Model 5
3. Visual Quality Assurance and Evaluation 5
4. Preprocessing 7
5. Network 8
6. Loss Function for Hematoma Segmentation 9

IV. Results 12

1. Quantitative Evaluation 12
2. Qualitative Evaluation 14

V. Conclusion 15

REFERENCES 16

LIST OF FIGURES

Figure **Page**

1. Representative 5.0 mm thick transverse CT sections in 3 subjects with TBI 4

2. Montage image of CT brain volume 5

3. Histogram of the quality score of CT images 6

4. GUI Tool for manual QA 6

5. CT image (Left) and preprocessed image (Right) 7

6. Overview of data employed in the experiment.. 8

7. False positive in segmentation predictions 9

8. Traditional Dice loss (Left) and designed Dice loss for healthy data (Right) 10

9. Loss curve with (R) and without (L) designed Dice loss 11

10. Dice Similarity Coefficient (DSC) for 2D and 3D model 12

11. Sensitivity or true positive rate (TPR) for 2D and 3D model 13

12. Specificity or true negative rate (TNR) for 2D and 3D model 13

13. The resulting predictions of 2D model and 3D model 14

LIST OF TABLES

Table **Page**

1. Average Dice, Sensitivity and Specificity for the 2D model and 3D model 13

Chapter I. Introduction

1. Traumatic Brain Injury

Traumatic brain injury (TBI) is a major cause of death and disability (Cozza et al., 2014). In the United States, the total number of traumatic brain injuries cases approaches 3.5 million. According to reports, the incidence of nerve trauma is very high due to motor vehicle accidents, civilians and military personnel participating in contact movements and theaters. TBI may lead to hematoma and local tissue disruption and may lead to brain swelling (edema) and increased intracranial pressure (ICP), thereby further causing hypotension and hypoxemia to damage local tissues. TBI can be classified into two main categories; primary injury that occurs during impact and secondary injury that caused by primary injury (He et al., 2013).

More than 30% of TBI patients have hematoma. The prevention and proper treatment of these sequelae can reduce the incidence of injury, the risk of death and the cost of care. CT and magnetic resonance imaging (MRI) are commonly used image modes for TBI diagnosis. Although MRI has been shown to have higher sensitivity and specificity than CT, it is usually used in the later stages of treatment that require more detailed information. In addition, MRI is not routinely used in mild TBI (Kurča et al., 2006). On the other hand, CT imaging is faster, lower cost, and can show severe abnormalities, such as fractures and hematoma (W. Chen et al., 2009). In addition, CT imaging also plays an important role in allowing the emergency department to perform rapid TBI assessments. Therefore, it is used as the gold standard for emergency care for diagnosis of intracranial injury after TBI (Babcock et al., 2012). It is also used for the evaluation of epidural and subdural hematoma in patients with TBI. Visual inspection of these images can be time-consuming, and related to errors, especially when a significant number of images need to be evaluated. Computer-aided imaging analysis can improve diagnostic efficiency and help reduce mortality, long-term complications, and related costs.

2. Hematoma Segmentation

Before deep learning came about, there were a lot of statistical methods that used manual feature selection to accomplish this. Most of them are unsupervised clustering-based methods or atlas-based methods. And this is still done now, but nowadays, a lot of stuff is deep learning focused.

Fuzzy c means clustering (FCM) is an unsupervised technique that has been successfully used in medical image segmentation (Bezdek et al., 1984). An image can be represented in various feature spaces and FCM segments the image into objects by grouping similar pixels in the feature space. The clustering is achieved by iteratively minimizing the cost function, which depends on the distance from pixels in the feature domain to the clustering center. By clustering the pixels, FCM can divide an image into meaningful regions. More specifically, segmentation is the process of dividing the entire image into c largest connected regions so that each region is uniform relative to some criteria.

However, since the standard FCM does not consider any spatial information in the context of images, it has the disadvantage of being sensitive to noise and other imaging artifacts. Recently, many researchers have incorporated local spatial information into the standard FCM (Chuang et al., 2006). One of the important features of an image is that adjacent pixels are highly correlated, that is, they have a high probability of belonging to the

same cluster. In spatial FCM, spatial functions have been incorporated into membership functions.

Another widely used segmentation methods is atlas-based segmentation, especially in medical image segmentation. In multi atlas label propagation with origins (Heckemann et al., 2006; Rohlfing et al., 2004), each semi-automatically or completely manually annotated atlas is aligned with the unsegmented target image. Then the propagated segmentation is merged into the consensus label of each voxel in the target image. More detailed overview of atlas-based methods is provided by (Cabezas et al., 2011). In 2013, a more powerful atlas-based method called joint label fusion was proposed by (Wang et al., 2013). In this state-of-the-art approach, the segmentation bias can be reduced by estimating the joint segmentation error of different atlas pairs. Atlas-based methods rely on the accurate registration of atlases and unsegmented MR images to determine the spatial transformation of atlas labels to the target space. This may be difficult if the target image is different from the available atlas due to pathological reasons.

Another type of hematoma segmentation algorithm is deep learning-based methods. The latest progress of Convolutional Neural Network (CNN) have demonstrated that deep learning methods perform excellent in automatically image classification and segmentation tasks (Litjens et al., 2017a). Some recent examples include brain tumor segmentation (Havaei et al., 2017), ischemic lesion segmentation (L. Chen et al., 2017), lung tumor segmentation (Hwang & Park, 2017). Specifically, in hematoma segmentation diagnostic procedures, deep learning algorithms have excellent ability to perform complex tasks with high speed and high precision similar to human experts. Recurrent Attention DenseNet (Grewal et al., 2018) uses recurrent attention DenseNet (Jegou et al., 2017) with LSTM to segment and classify brain hemorrhage from CT scans, RADnet utilizes 3D context from neighboring slices to improve predictions at each slice and subsequently, aggregates the slice-level predictions to provide diagnosis at CT level. ICHNet (Islam et al., 2019) develops by integrating dilated convolutional neural networks (CNN) with hypercolumn features, sampling a suitable number of pixels and concatenating corresponding features from multiple layers. Because pixels can be freely sampled instead of image patches, the model trains in the brain area and ignored CT background to boosting the convergence time. (Kuo et al., 2019) reports a deep learning algorithm with accuracy comparable to that of radiologists for the evaluation of acute intracranial hemorrhage on head CT by using a strong pixel-level supervision approach and a relatively small training dataset. It demonstrates the highest classification accuracy to date with a receiver operating characteristic (ROC) area under the curve (AUC) of 0.991 ± 0.006 for identification of examinations positive for acute intracranial hemorrhage.

In these primary types of hematoma segmentation algorithms, especially the deep learning-based methods, a significant number of CT brain data are always needed for more accurate training. However, there is only one publicly available dataset called CQ500 for the detection of ICH sub-types (Chilamkurthy et al., 2018) that consists of 491 head CT scans. There is no publicly available dataset for the hematoma segmentation. Therefore, there is a need for a benchmark dataset that could help to extend the work in ICH segmentation. Our work proposed an efficient way to reduce the manually effort for collecting a well labeled dataset for hematoma segmentation.

Chapter II. Related Works

1. 2D Trained Network

In deep learning tasks, more data means more accuracy and more robustness, especially in the domain of medical imaging (Halevy et al., 2009). However, to protect the anonymity of patient, usually it is not allowed (or even illegal) to transfer the protected health information (PHI) between institutions. So far, there is not a large public dataset available for deep learning tasks in many application areas including medical imaging. Therefore, a contradiction lies here: machine learning models benefit greatly from large amounts of data, but healthcare-related data sets cannot be easily shared between sites.

In order to address this problem, researchers proposed to transfer the models themselves between sites instead of a dataset (Sheller et al., 2019; Yang et al., 2019). (Remedios et al., 2019) proposed an extensible framework through which multiple sites can train the same model using private data and the validation of the efficacy of two different training schema on the segmentation of hematoma in traumatic brain injury (TBI) CT scans and demonstrate that the multi-site model have a better generalization ability.

However, 2D network is mostly used in natural images, such as photographic images, to extract the spatial features only in two dimension (Krizhevsky et al., 2017). Though 2D CNN can be applied to the volumetric data set (such as cross-sectional CT images), some researchers (Litjens et al., 2017) have reported the benefit of 3D CNN to incorporate the volumetric information in medical images to get a higher accuracy. In CT brain images, the spatial information of the brain hemorrhage in 2D CT image extends to three dimension. As the main motivation, in our experiment, we extent a previous published 2D hematoma segmentation model to a robust 3D whole brain hematoma segmentation pipeline that integrate human quality assurance if a significant number of unlabeled CT brain images for better performance and generalization. Specifically, we choose one of the three 2D models in this previous published work to do transfer learning. The original 2D patch-based model is trained on 33 images with a 0.748 average Dice coefficient on 29 testing images. We selected good resulting segmentations generated from the predictions of 2D segmentation model and used these machine-predicted inlier data to improve the segmentation performance.

The rest of the paper is organized as follows. Details of the proposed method are presented in Chapter 3. We show the performance of the proposed method in Chapter 4 and finally we conclude the paper in Chapter 5.

Chapter III. Materials and Methods

1. Data Collection

11979 CT image volumes from 4033 patients were retrospectively acquired in deidentified from consecutive trauma patients at Vanderbilt University Medical Center. All CT image volumes were in Digital Imaging and Communications in Medicine (DICOM) format, a QA Processing tool (Gao, 2019) was used to do the Quality Assurance for DICOM files. This tool involves 5 steps:

1. Step 1: Check the instance number of DICOMs (need to read DICOM header) if the instance number can match the number of DICOMs for the session.
2. Step 2: Check the slice distance if DICOMs (need to read DICOM header) to avoid those sessions lose slices.
3. Step 3: filter some sessions with very limited slices (here we set 15) but still can pass the instance number check.
4. Step 4 Use the dcm2niix tool (Li et al., 2016)to convert DICOMs to NIFTI by:

dcm2niix -m y -o output_folder DICOM_folder

Here, -m means Merge 2D slices: ("-m y" or "-m n"). If selected, images from the same series are stacked into a single NIfTI image regardless of study time, echo, coil, orientation, etc.

5. Step 5: use MIPAV to check one by one (time consuming, optional).

After preprocessing QA, 11477 quality insured volumes were acquired in nifty file format from total 11979 DICOM scans. 502 DICOM scans are not exit or cannot be loaded or fail in the quality assurance.

Figure 1 Representative 5.0 mm thick transverse CT sections through the head in 3 subjects with TBI. In-plane resolution is approximately 0.5 × 0.5 mm. In each case, the hemorrhagic lesion appears intermediate density between normal brain tissue and bone. Note the heterogeneity of size, location, density and configuration (Remedios et al., 2019)

2. Prediction using 2D Model

The baseline 2D model was trained using 33 CT brain volumes with manual label and test on 29 CT brain images. The best performance 2D model achieved 0.748 Dice score on the testing set. This model was used to predict the predicted masks on 11477 CT volumes. The 2D model preform prediction for every slice of a given volume and then concatenate the prediction mask of every slice into one mask volume. It takes 8 days to compute 11477 prediction masks with Nvidia GTX Titian X.

3. Visual Quality Assurance and Evaluation

For easier visualize the whole volume, 11477 3D images were sliced and each of them was aligned as one single 2D montage image (Figure 2). Converting was performed via MATLAB (Jimmy Shen, 2020). For every brain slice in a 2D montage image, the prediction masks were overlaid on the corresponding CT brain slice for manual inspection.

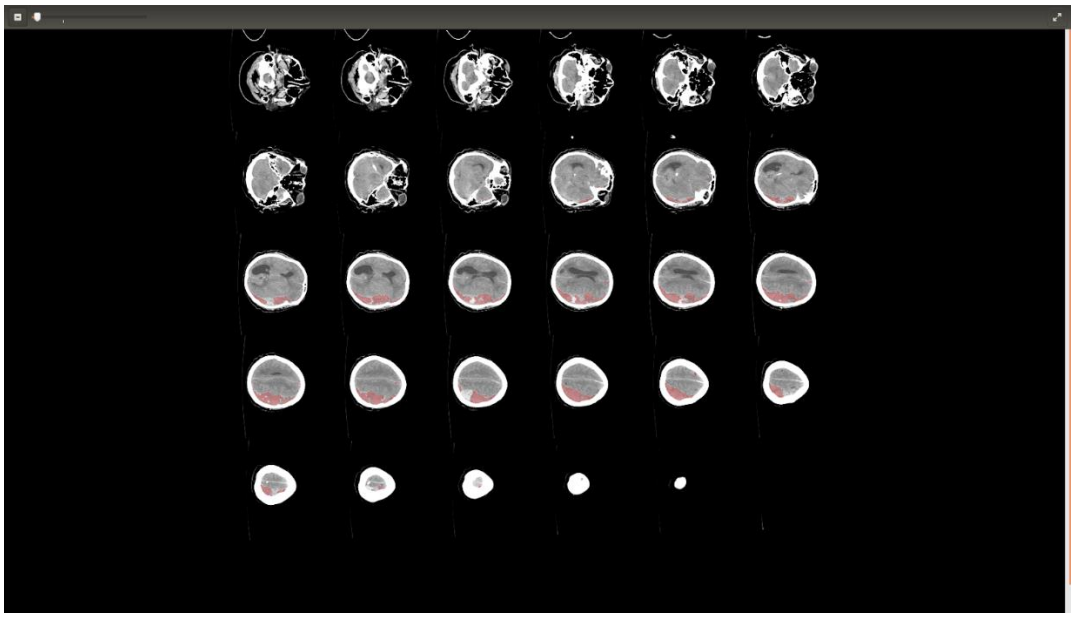


Figure 2 Montage image of CT brain volume

A quality score for every resulting segmentation was manually rated by a trained evaluator for evaluating the resulting prediction masks. The quality score for each montage image was determined by the accuracy of the segmentation of hematoma. If the segmentation of hematoma were well located in an unhealthy brain with hematoma, the montage image was coded with score 0. Score 1 was coded for a reasonable quality segmentation with some errors in an unhealthy brain with hematoma. Score 2 was coded for a failed segmentation in an unhealthy brain with hematoma. Score 3 was coded for a healthy brain image without hematoma with blank prediction mask. Score 4 was coded for invalid data. In total 11477 images, 1199 of them are labeled with score 0, 2475 of them are labeled with score 1, 2340 of them are labeled with score 2 and 4995 of them are labeled with Score 3.

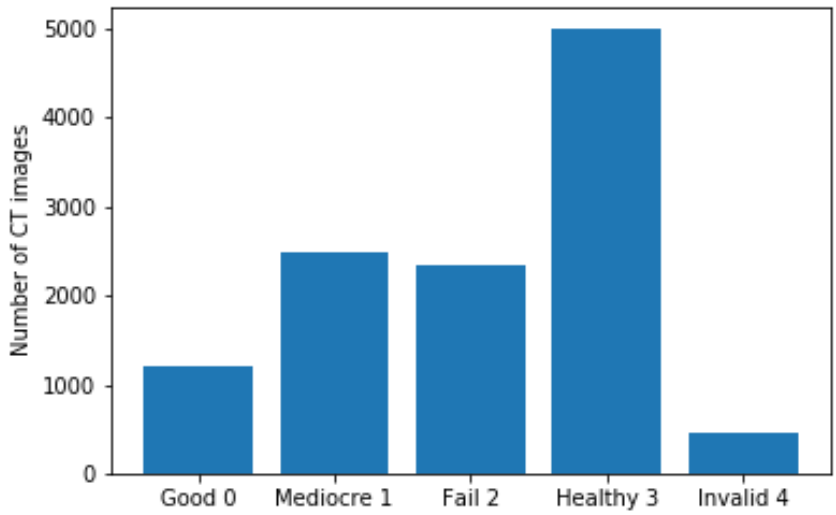


Figure 3 Histogram of the quality score of CT images

In order to speed up the manual evaluation process, a GUI application was developed using Tkinter in python3 (Figure 4). With this GUI application, montage images can be shown one by one and it is easily to zoom in or zoom out for more detail information. After the coded score is chosen by evaluator, it is automatically saved in a text list for further use.

Figure 4 GUI Tool for manual QA

4. Preprocessing

All manual inspected CT brain volumes are saved on hardware with the predicted masks. The voxel resolutions of all volumes were approximately 0.5 × 0.5 × 2.0 - 5.0 mm^3 with approximately dimensions of (500, 600, 30 - 70). Voxel intensities preserved in Hounsfield units, which places hematoma density at a value of about +50 to +70HU with air and bone at opposite extreme values of -1000HU and +1000HU, grey matter and white matter at values of +20 to +50HU.

Subsequently, each volume underwent the same preprocessing steps:

1. Skull-stripped with CT BET (Muschelli et al., 2015).
2. Rigidly transformed to the "RAI" orientation (the x axis is running from Left to Right, the y axis is running from Posterior to Anterior, the z axis is running from Superior to Inferior).
3. Resampled to 0.5 × 0.5 × 5 mm^3 via 3dresample, a part of the Analysis of Functional Neuroimages (AFNI) software package, using nearest interpolation method for masks resampling and bicubic interpolation method for CT images resampling to avoid artifacts.
4. For every image, cropped the image to the smallest image that still contains the non-zero-pixel (foreground brain) locations and save the cropping values.
5. Cropped the corresponding mask with the saved cropping values.
6. Padded the images with masks to 208 × 208 × 48 pixels.
7. Cut off the intensity to the soft tissue window [0,1000].

After preprocessing, all 3D CT brain images have a same voxel resolution with 0.5 × 0.5 × 5 mm^3 and a same dimension with (208 x 208 x 48). The preprocessing part significantly removed a lot of background data with no information.

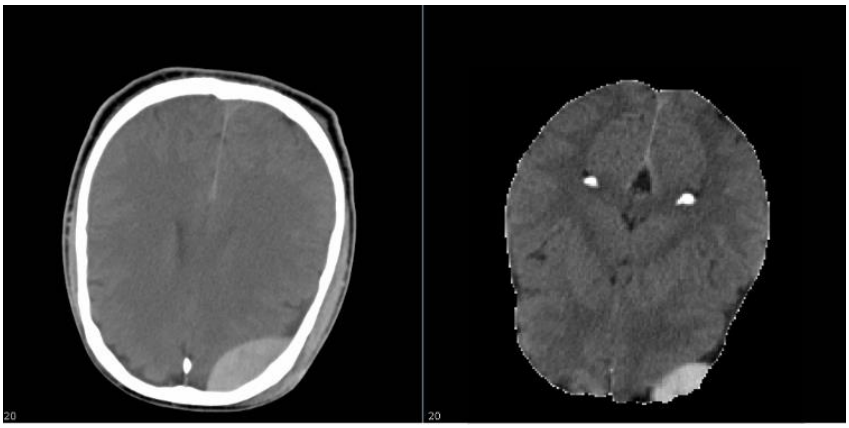


Figure 5 CT image (Left) and preprocessed image (Right)

5. Network

The network architecture used for hematoma segmentation is 3D-Unet (Çiçek et al., 2016). In our 3D Unet network in hematoma segmentation, the input CT brain images only consist of two classes: background (brain) and foreground (hematoma). The network predictions are volumes with the same shape as the input CT brain volumes and are processed through a sigmoid layer which outputs the probability of each voxel to belong to foreground. Then, the output probability map was classified by a hard threshold (0.5 in this pipeline). If the prediction for a given voxel is greater than the hard threshold, it is assigned as foreground class to that voxel. If the prediction is less than the hard threshold, it does not get assigned a label (i.e. is labelled as background). Therefore, the input of this network is a 3D CT whole brain volume with corresponding ground truth volume, both in the shape of [1x208x208x48], and the output is a binary predicted mask volume with shape of [1x208x208x48].

We trained with a batch size of 1 using whole 3D volume. The initial learning rate was set to 1×10^{-4} with the Adam optimizer. Convergence was defined as no improvement of validation loss in 50 epochs. The selected deep learning framework was TensorFlow v.1.14 and an NVIDIA 1080 TI was used for hardware acceleration.

Image volumes with scores of 1, 2 and 4 were omitted from training set selection due to the low quality of segmentation results. 2398 total 3D CT brain volumes with scores 0 (n = 1199) along with the same number of scores 3 were selected to be used in the 3D hematoma segmentation pipeline. With total 2398 images as the dataset, 20% were randomly selected to be the validation set, the testing set was the original testing set of previous published 2D network. Patients were not mixed in the data split, and the samples with scores 0 and scores 3 were randomly undersampled such that data were balanced for training and validation.

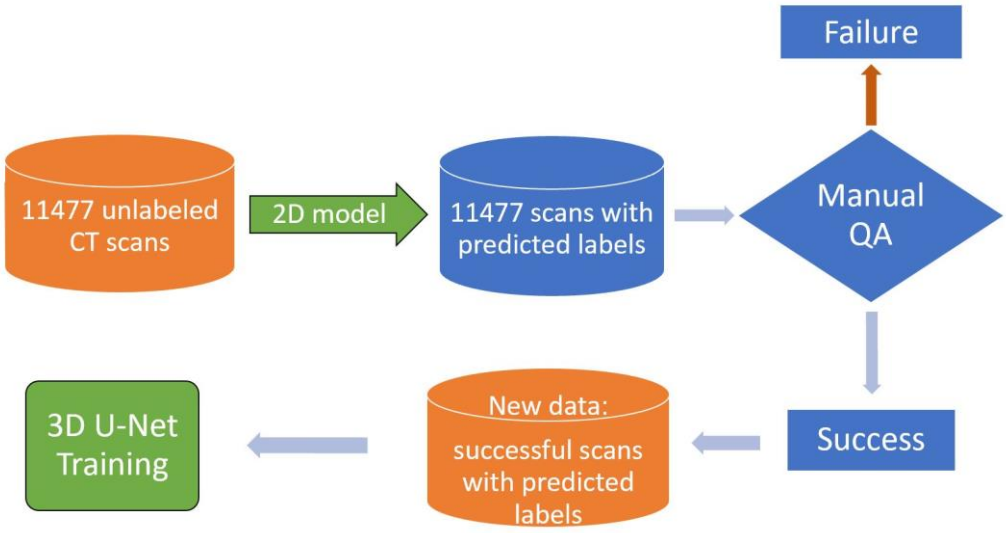


Figure 6 Overview of data employed in the experiment. The 2D hematoma segmentation model trained on 33 training data was applied on 11477 unlabeled data, and the resulting predictions were evaluated using manual quality assurance. Specifically, the failed segmentations were ignored and the successful scans (n = 2398) with predicted labels were used for 3D U-Net training.

6. Loss Function for Hematoma Segmentation

In most segmentation tasks, the Dice score coefficient (DSC) is a measure of overlap widely used to assess segmentation performance when a gold standard or ground truth is available. The use of the Dice loss in CNN was proposed in (Milletari et al., 2016).

However, since Dice coefficient is not stable for small structures (Wong et al., 2018), which are very common in medical images such as the CT brain images with traumatic brain injury we are processing in this work, the misclassification of several pixels can cause the coefficients to drop significantly. In CT brain images with traumatic brain injury, several hematomas without connection can occur in different brain area, also some small tissues or lesions are very similar to the hematoma. Therefore, in hematoma segmentation, it is very likely that the network may generate more false positive areas in the output predictions.

As a result, when dealing with the false positive predictions especially in a healthy brain volume input data which the corresponding ground truth is empty, the traditional Dice loss which only compute the foreground overlaps cannot well represent the performance of network predictions and may causes the learning process to get trapped in local minima. The loss value will be infinitely close to zero when the network prediction has false positives for a healthy brain with empty ground truth.

CT scan Ground Truth Prediction

Figure 7 False positive in segmentation predictions. In CT brain with traumatic brain injury, it is very common that several hematomas may show in several different places without connection in one brain volume. In addition, the air bubbles shown in the top area of the brain volume is very similar to the hematoma. Therefore, in hematoma segmentation, it is very likely that the network generates more false positive areas in the output predictions.

In order to alleviate such problem, in this work, in terms of the two kinds of training data we used, unhealthy brain with hematoma (score 0) and healthy brain without hematoma (score 1), two 3D Dice loss functions have been designed for finding the best parameters of 3D Unet model. These two loss functions have been applied for different kinds of input data, the unhealthy CT brain images and healthy CT brain images, respectively:

1. Traditional Dice loss (Sudre et al., 2017) for unhealthy brain volumes:

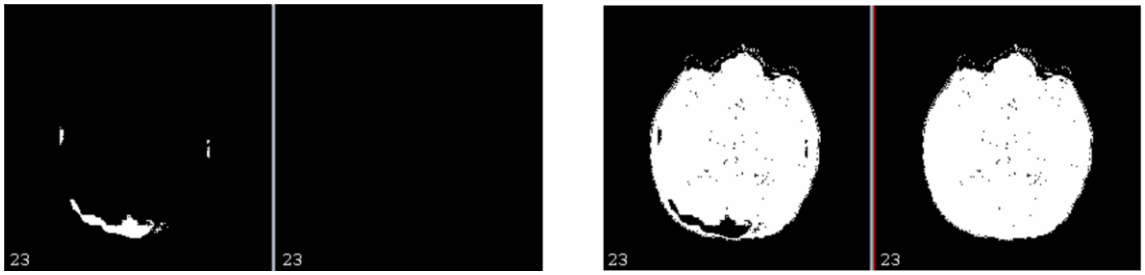
$$DL_u = 1 - \frac{2\Sigma_i^N p_i g_i + \epsilon}{\Sigma_i^N p_i + \Sigma_i^N g_i + \epsilon}$$

where G is the reference binary volume (ground truth) with voxel values g_i, and P is the predicted binary segmentation volume for the foreground label over N image elements p_i, the ϵ term is used here to ensure the loss function stability by avoiding the numerical issue of dividing by 0, i.e. p and g empty.

2. Designed Dice loss for healthy brain volumes:

$$DL_h = 1 - \frac{2\Sigma_i^N (b_i - p_i) b_i + \in}{\Sigma_i^N (b_i - p_i) + \Sigma_i^N b_i + \in}$$

where B is the whole brain volume in the input image volume with voxel values b_i, and P is the predicted binary segmentation volume for the foreground label over N image elements p_i, the ϵ term is used here to ensure the loss function stability by avoiding the numerical issue of dividing by 0, i.e. p and g empty.



Prediction Ground Truth Brain - Prediction Brain

Figure 8 Traditional Dice loss (Left) and designed Dice loss for healthy data (Right). In this loss function for healthy brain volumes, instead of simply computing the background overlap between the prediction and ground truth, first we use the input volume to get the whole brain volume, then the Dice loss is defined by computing the overlap of the whole brain volume and the background prediction in the brain volume area. This aims to avoid errors caused by empty areas outside the brain in the images. Here the Dice loss of this segmentation prediction is -0.773 instead of $-1.63e^{-10}$ computed by the traditional Dice loss function.

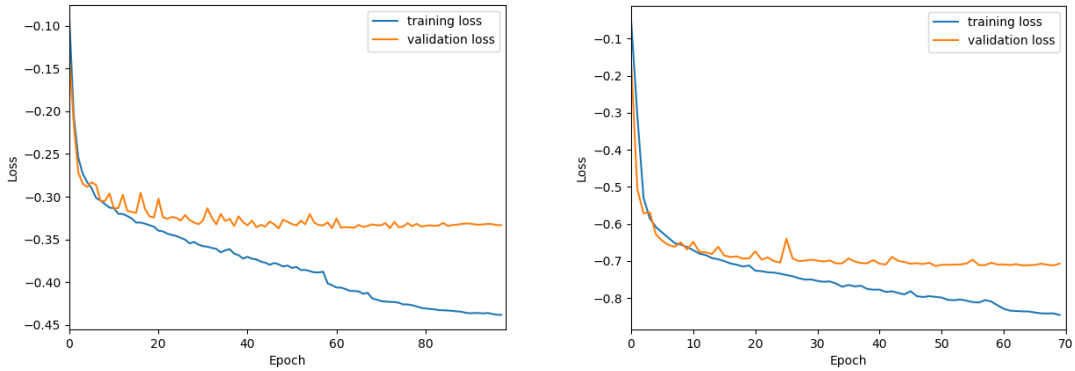


Figure 9 Loss curve with (R) and without (L) designed Dice loss. By using two different Dice loss functions during the training, the loss curve can properly represent the performance of the model and make the model can handle with healthy brain images.

Chapter IV. Results

The training loss and validation loss were recorded during the training and shown in Figure 9 as a function of epoch. The loss converges after 147 epoch training. Based on the validation loss curve, the epoch 97 model was selected as the best performance model with the lowest validation loss value.

1. Quantitative Evaluation

After training, we evaluated the selected model on the original testing set, which is the same testing set as the 2D patch-based model. The box plot of Dice similarity coefficient of the model was shown in Figure 10 as quantitative evaluation. As shown below, the mean Dice Similarity Coefficient on 2D model predictions and 3D model predictions is 0.748 and 0.729. the median Dice score of 2D model and 3D model is 0.792 and 0.763. The best Dice similarity coefficient performed on 2D and 3D model predictions are 0.920 and 0.967. However, there are two outliers with Dice similarity coefficient lower than 0.4, as a result the mean Dice similarity coefficient is reduced to 0.729. The outlier segmentations in the testing results may be caused by several reasons. In one hand, the ground truth of the testing set is manually labeled by humans, which is not 100% precise, especially in a low-dose CT brain image which the resolution is too low to find the edge of hematoma accurately. Also, as the hematoma ages, it becomes hypo dense so that the hematoma will show different HU value in patients with different ages (Rao et al., 2016). However, the age of the patients did not include as the network input.

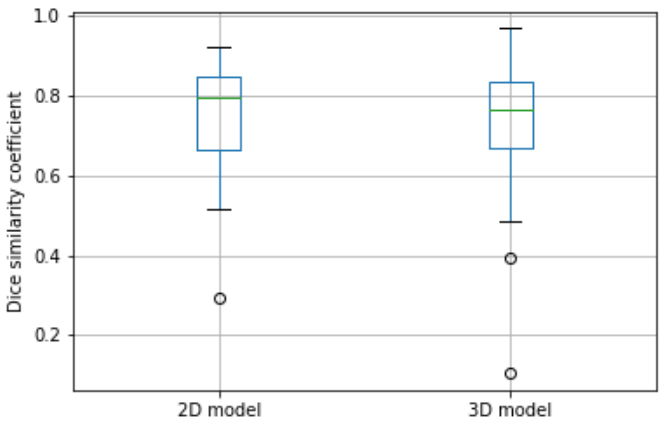


Figure 10 Dice Similarity Coefficient (DSC) for 2D and 3D model (p= 0.2471 >= 0.05 using paired t-test). The mean Dice Similarity Coefficient on 2D model predictions and 3D model predictions is 0.748 and 0.729. the median Dice score of 2D model and 3D model is 0.792 and 0.763. The best Dice similarity coefficient performed on 2D and 3D model predictions are 0.920 and 0.967.

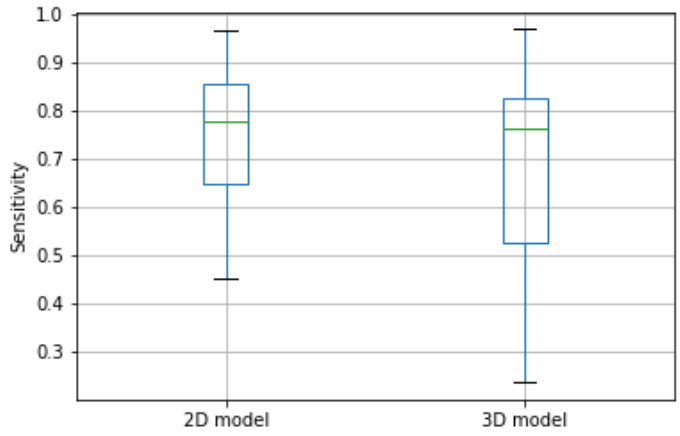


Figure 11 Sensitivity or true positive rate (TPR) for 2D and 3D model (p= 0.0025 < 0.05 using paired t-test)

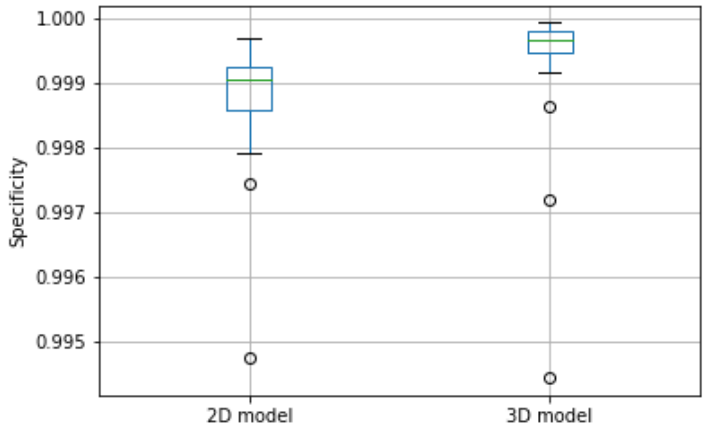


Figure 12 Specificity or true negative rate (TNR) for 2D and 3D model (p= 0.0000 < 0.05 using paired t-test)

Table 1 Average Dice Similarity Coefficient, Sensitivity and Specificity for the 2D model and 3D model. The average result over both models is shown to illustrate each model's general ability. The p value indicates significant improvements in Specificity (p < 0.05), and a not significant (p > 0.05) reduction in Dice Similarity Coefficient between the 2D and 3D models as evaluated by using paired t-test. The improved specificity or true negative rate (TNR) of 3D model shown that it is better for handling the false positive predictions.

	Dice (DSC) p= 0.2471 >= 0.05	**Sensitivity (TPR)** p= 0.0025 < 0.05	**Specificity (TNR)** p= 0.0000 < 0.05
2D model	0.7478	0.7443	0.9988
3D model	0.7286	0.6918	0.9994

2. Qualitative Evaluation

The predicted segmentation of test CT volumes are shown in Figure 13 for qualitative evaluation. According to the Figure, the model has lower false positive predictions. Compare with the 2D patch-based model, the predicted segmentations from our pipeline become more detailed and more comprehensive. However, in some scans, it is not very sensitive to find the hematoma as shown in lower case in Figure 13.

CT scan Ground Truth 2D model prediction 3D model prediction

Figure 13 The resulting predictions of 2D model and 3D model. The 3D model has a more comprehensive prediction on large scale hematoma but ignored the tiny hematoma which is not connected with the main body. 3D network allows the model to learn the global location and context at the same time so that it has a better performance on a connected target.

Chapter V. Conclusion

In this work, we proposed a 3D supervised learning method for CT hematoma segmentation with the integration of manual quality assurance of the predicted annotations from a 2D trained network. Briefly, a previously published patched-based segmentation model was applied to 11477 scans from 4033 patients that were retrospectively acquired in deidentified from consecutive trauma patients. The resulting predictions were manually rated for quality assurance and classified into four levels based on the segmentation accuracy. From all the resulting segmentations, 2400 good segmentation scans with predicted masks were selected as training data to train a 3D segmentation model. Two different Dice loss functions were designed and applied on two kinds of input data, respectively. The mean Dice Similarity Coefficient (DSC) obtained from the 3D model was 0.729 on the same testing set as 2D Network.

To conclude, feedback of successful automated results presents an efficient manner to transfer a 2D model to 3D and increase robustness of supervised deep learning algorithms at substantially reduced manual effort (here in 35 hours for quality assurance versus an estimated 11477 hours for full tracing of 11477 scans). The designed Dice loss function can be used in the further segmentation tasks in order to handle the false positives generated during the training.

REFERENCES

Babcock, L., Byczkowski, T., Mookerjee, S., & Bazarian, J. J. (2012). Ability of S100B to predict severity and cranial CT results in children with TBI. In *Brain Injury*. https://doi.org/10.3109/02699052.2012.694565

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences*. https://doi.org/10.1016/0098-3004(84)90020-7

Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., & Bach Cuadra, M. (2011). A review of atlas-based segmentation for magnetic resonance brain images. *Computer Methods and Programs in Biomedicine*. https://doi.org/10.1016/j.cmpb.2011.07.015

Chen, L., Bentley, P., & Rueckert, D. (2017). Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *NeuroImage: Clinical*. https://doi.org/10.1016/j.nicl.2017.06.016

Chen, W., Smith, R., Ji, S. Y., Ward, K. R., & Najarian, K. (2009). Automated ventricular systems segmentation in brain CT images by combining low-level segmentation and high-level template matching. *BMC Medical Informatics and Decision Making*. https://doi.org/10.1186/1472-6947-9-S1-S4

Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., Mahajan, V., Rao, P., & Warier, P. (2018). Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet*. https://doi.org/10.1016/S0140-6736(18)31645-3

Chuang, K. S., Tzeng, H. L., Chen, S., Wu, J., & Chen, T. J. (2006). Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*. https://doi.org/10.1016/j.compmedimag.2005.10.001

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-net: Learning dense volumetric segmentation from sparse annotation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-46723-8_49

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*. https://doi.org/10.1006/cbmr.1996.0014

Cozza, S. J., N. Goldenberg, M., & Ursano, R. J. (2014). Care of Military Service Members, Veterans, and Their Families. In *Care of Military Service Members, Veterans, and Their Families*. https://doi.org/10.1176/appi.books.9781585625161

Gao, R. (2019). *GitHub-RiqiangGao/QA_tool*. Https://Github.Com/RiqiangGao/QA_tool.

Grewal, M., Srivastava, M. M., Kumar, P., & Varadarajan, S. (2018). RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans. *Proceedings - International Symposium on Biomedical Imaging*. https://doi.org/10.1109/ISBI.2018.8363574

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*. https://doi.org/10.1109/MIS.2009.36

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P. M., & Larochelle, H. (2017). Brain tumor segmentation with Deep Neural Networks. *Medical Image Analysis*. https://doi.org/10.1016/j.media.2016.05.004

Heckemann, R. A., Hajnal, J. v., Aljabar, P., Rueckert, D., & Hammers, A. (2006). Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2006.05.061

He, W., Wang, L. S., Li, H. Z., Cheng, L. G., Zhang, M., & Wladyka, C. G. (2013). Intraoperative contrast-enhanced ultrasound in traumatic brain surgery. *Clinical Imaging*. https://doi.org/10.1016/j.clinimag.2013.08.001

Hwang, S., & Park, S. (2017). Accurate lung segmentation via network-wise training of convolutional networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-67558-9_11

Islam, M., Sanghani, P., See, A. A. Q., James, M. L., King, N. K. K., & Ren, H. (2019). ICHNet: Intracerebral hemorrhage (ICH) segmentation using deep learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-11723-8_46

Jegou, S., Drozdzal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017). The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. https://doi.org/10.1109/CVPRW.2017.156

Jimmy Shen. (2020). Tools for NIfTI and ANALYZE image. *MATLAB Central File Exchange*.

Kuo, W., Häne, C., Mukherjee, P., Malik, J., & Yuh, E. L. (2019). Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1908021116

Kurča, E., Sivák, Š., & Kučera, P. (2006). Impaired cognitive functions in mild traumatic brain injury patients with normal and pathologic magnetic resonance imaging. *Neuroradiology*. https://doi.org/10.1007/s00234-006-0109-9

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017a). A survey on deep learning in medical image analysis. In *Medical Image Analysis*. https://doi.org/10.1016/j.media.2017.07.005

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017b). A survey on deep learning in medical image analysis. In *Medical Image Analysis*. https://doi.org/10.1016/j.media.2017.07.005

Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*. https://doi.org/10.1109/3DV.2016.79

Rao, M. G., Singh, D., Khandelwal, N., & Sharma, S. K. (2016). Dating of early subdural haematoma: A correlative clinico-radiological study. *Journal of Clinical and Diagnostic Research*. https://doi.org/10.7860/JCDR/2016/17207.7644

Remedios, S., Roy, S., Blaber, J., Bermudez, C., Nath, V., Patel, M. B., Butman, J. A., Landman, B. A., & Pham, D. L. (2019). *Distributed deep learning for robust multi-site segmentation of CT imaging after traumatic brain injury*. 9. https://doi.org/10.1117/12.2511997

RiqiangGao/QA_tool. (n.d.). Retrieved April 19, 2020, from https://github.com/RiqiangGao/QA_tool

Rohlfing, T., Russakoff, D. B., & Maurer, C. R. (2004). Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/TMI.2004.830803

Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2019). Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-11723-8_9

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). *Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations*. https://doi.org/10.1007/978-3-319-67558-9_28

Tools for NIfTI and ANALYZE image - File Exchange - MATLAB Central. (n.d.). Retrieved April 19, 2020, from https://www.mathworks.com/matlabcentral/fileexchange/8797-tools-for-nifti-and-analyze-image?s_tid=gn_loc_drop

Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., & Yushkevich, P. A. (2013). Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2012.143

Wong, K. C. L., Moradi, M., Tang, H., & Syeda-Mahmood, T. (2018). 3D segmentation with exponential logarithmic loss for highly unbalanced object sizes. *Lecture Notes in Computer Science*

(Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *11072 LNCS*, 612–619. https://doi.org/10.1007/978-3-030-00931-1_70

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*. https://doi.org/10.1145/3298981