Automatic segmentation of stroke lesions in non-contrast computed tomography with convolutional neural networks

Author(s) names withheld
Address withheld

Editors: Under Review for MIDL 2020

Abstract
Manual lesion segmentation for non-contrast computed tomography (NCCT), a common modality for volumetric follow-up assessment of ischemic strokes, is time-consuming and subject to high inter-observer variability. Our approach uses a combination of a 3D convolutional neural network (CNN) combined with post-processing methods. A total of 272 multi-center clinical NCCT datasets were used: 204 for CNN training, 48 for validation and developing post-processing methods, and 20 for testing. The testing datasets were from centers that did not contribute to the training and validation sets, and were segmented by two neuroradiologists. We achieved a median Dice score of 0.63, which was significantly improved to 0.66 with post-processing. The automatically segmented lesion volumes were not significantly different from the lesion volumes determined by the two manual observers. As the model was trained on datasets from multiple centers, it is broadly applicable.

Keywords: stroke, computed tomography, segmentation, deep learning, CNN

1. Introduction
NCCT is the most common imaging modality for volumetric assessment of stroke lesions (Eswaradass et al., 2016). Manual lesion segmentation in NCCT images is time-consuming and associated with high inter-observer variability. Semi-automatic lesion segmentation tools have been developed (Kuang et al., 2019), but still require human interaction while previous work on automatic NCCT lesion segmentation is very limited.

CNNs show superior performance for various segmentation tasks in medical imaging because of their ability to learn complex patterns and relationships in the data (Zaharchuk et al., 2018). The use of multi-scale features and three-dimensional kernels (Kamnitsas et al., 2017) would allow an automated segmentation algorithm to take advantage of the spatial contiguity of stroke lesions while maintaining localized focus. However, for stroke lesion segmentation, these methods have only been applied to magnetic resonance imaging (MRI) (Chen et al., 2017; Wu et al., 2019; Liu et al., 2019) or computed tomography perfusion and angiography datasets (Öman et al., 2019; Kasasbeh et al., 2019). CNN-based lesion segmentation has not been evaluated in NCCT datasets, despite its common application in stroke imaging.

The aim of this work was to train and evaluate a CNN model for stroke lesion segmentation in NCCT datasets. To improve upon CNN segmentations we investigated post processing methods. We tested the models generalizability by evaluating it on a holdout test set acquired from entirely different studies from those used in training and validation.
2. Materials and Methods

A total of 272 clinical follow-up NCCT datasets acquired at 24 centers and corresponding manual segmentations were available. The in-slice resolution ranged from 0.355 to 0.637 mm, the slice thickness ranged from 1.00 to 10.0 mm, while the number of slices ranged from 10 to 141. 204 datasets were used for training of the 3D CNN-based lesion segmentation model, 48 datasets for validation, and 20 datasets for testing. The 20 hold-out datasets used for testing were acquired at two centers not contributing to the training or validation sets and were manually segmented by two neuroradiologists.

All datasets were preprocessed identically to ensure data consistency across the different scanners and acquisition protocol. A 3D multi-scale CNN was trained with 204 datasets and corresponding lesion segmentations using the previously described DeepMedic (Kamnitsas et al., 2017) framework (v0.7.2). The CNN-based lesion segmentations were post-processed to improve accuracy. In brief, post-processing consisted of a connected component analysis to exclude small lesion components, most likely caused by noise artifacts, and an automatic hole-filling approach. The minimum lesion size and hole-filling kernel size were systematically optimized using the validation datasets, resulting in final values of 1.5 mL and 3 voxels, respectively. Automatic segmentations were evaluated using the Dice similarity coefficient (DSC) and lesion volume. Further details on data pre-processing, model design, training parameters, post-processing and model evaluation are available in Appendix A.

3. Results and Discussion

The median lesion volumes for the training and validation sets were 40.4 [14.1–96.3] mL and 41.5 [20.0–107.1] mL, respectively. The test set segmented by two observers had lesion volumes of 49.0 [15.7–166.9] mL and 56.3 [22.4–192.1] mL. No significant differences in lesion volumes were found between the training, validation and test sets were found (P > 0.05).

The CNN achieved a median DSC of 0.63 [0.37–0.74] on the test set, compared to the two manual segmentations (Figure 1A). Post-processing of CNN-based segmentations significantly improved the DSC to 0.66 [0.41–0.75] (P < 0.05). Both were lower than the inter-observer DSC (P < 0.0001). Bland-Altman analysis showed no systematic bias of the automatic segmentations toward either observer in terms of DSC (Additional Figure 2) or lesion volume (Additional Figure 3). Importantly, no significant difference was found between the lesion volume estimates of the first observer (49.0 [15.7–166.9] mL), second observer (56.3 [22.4–191.1] mL), and the CNN-based lesion segmentation before (66.9 [27.2–114.3] mL) and after post-processing (66.4 [22.1–118.8] mL) (P > 0.05) (Figure 1B). Additionally, there was a strong correlation between lesion volumes from manual segmentations and CNN-based segmentations after post-processing (ρ = 0.77, P < 0.0001) (Figure 1C).

As NCCT is a standard imaging procedure available in most stroke centers for follow-up assessment, an automatic lesion segmentation pipeline for this modality is of high demand. To date, CNN models for follow-up lesion segmentation have primarily been investigated for MRI (Kamnitsas et al., 2017; Chen et al., 2017; Wu et al., 2019; Liu et al., 2019),
achieving DSCs (0.67–0.79) similar to the proposed method’s segmentation performance in the NCCT test set (median DSC=0.66). It bears mentioning that these results are not directly comparable because different datasets were used. Nevertheless, the proposed method is very promising given that lesion segmentation in NCCT is more challenging compared to typical MRI follow-up sequences such as diffusion weighted imaging (Fiebach et al., 2002) as the ischemic changes are subtler.

This study demonstrated the successful use of a CNN-based lesion segmentation in clinical NCCT datasets. Though the voxel-wise agreement of the CNN-based segmentations was inferior to the inter-observer agreement, the corresponding lesion volumes were not different from manual segmentations and strongly correlated with them. This suggests a potential application of the model for volumetric assessment of follow-up lesions. Within this context, it is reassuring that the model results were not biased toward either rater. Providing consistent results is an advantage of automatic segmentation algorithms, thereby reducing variability between sites or studies. This lays the foundation for developing automatic lesion analysis tools for NCCT images and can contribute toward consistent and high-throughput analysis of large multi-center studies. The trained model will be freely available, in the final non-anonymous manuscript, for NCCT lesion segmentation.
References


Acknowledgments
Acknowledgments withheld.
Appendix A. Supplemental Methods

A.1. Segmentation evaluation

The Dice similarity coefficient (DSC) was used as the primary outcome measurement for evaluation of the automatic lesion segmentations. The DSC measures the overlap between two segmentations and is defined between 0 and 1, whereas 1 indicates perfect consensus. The DSC is calculated as \( \text{DSC} = \frac{2 \cdot |A \cap B|}{|A| + |B|} \), where A and B are segmentations from manual and automated segmentations, respectively.

For inter-observer DSC, A and B were the segmentations from observer A and observer B, respectively.

DSC scores for training data were obtained by 10-fold cross-validation. Samples were randomly assigned to test folds. The training samples were evaluated when they were a part of a fold’s test data.

DSC scores and lesion volumes from automatic segmentations, for the validation and holdout test set were obtained using a single CNN model that was trained on the entire training data.

A.2. NCCT scan pre-processing

As the NCCT images were acquired from multiple centers, with differing scanners and imaging protocols, the datasets had to be pre-processed to ensure consistency.

First, the bone structures were removed from each dataset, retaining only the brain tissue in the images. To remove the bone structures, which have high Hounsfield values, a six-step procedure following the approach described by Muschelli et al (Muschelli et al., 2015) was performed in a slice-wise manner. The approach was implemented using the Insight Segmentation and Registration Toolkit (ITK) (Yoo et al., 2002). A Gaussian filter with a variance of 4 pixels was used to smooth each slice. Next, the intensities were thresholded between 0 and 100 Hounsfield units and a circular structural element with a radius of 1 pixel was used to erode the resulting segmentation. Afterwards, the largest connected component in each slice is extracted and a circular structural element with a radius of 1 pixel was used to dilate this component in order to create a brain mask for the slice. Finally, after performing the first three steps in each slice, the masks from each slice are combined into a final mask for the entire image and any holes in this final mask are filled using the VotingBinaryHoleFillingImageFilter in ITK.

Second, the images were thresholded between 0 and 100 Hounsfield units to remove noise and hypo- or hyper density artifacts. Finally, the images were normalized to unit variance to account for potential differences in scanner tube potential and different reconstruction algorithms. All images in the training, validation, and test datasets underwent the same pre-processing procedure.

A.3. 3D Convolutional Neural Network Architecture

The DeepMedic (Kamnitsas et al., 2017) framework consists of two convolutional pathways for processing images at two different scales. This is done by using high- and low-resolution versions of the image as the input to the pathways. Both pathways use convolutional kernels of size 3x3x3 and have 8 convolutional layers consisting of 30, 30, 40, 40, 40, 40, 50 and...
50 feature maps. After these 8 parallel convolutional layers, the pathways are combined in two convolutional layers. These two convolutional layers each had 150 neurons and used 1x1x1 convolutional kernels. A final classification softmax classification layer produces lesion probability maps. A threshold of > 0.5 was used to binarize the probability map to a final lesion segmentation.

**A.4. 3D Convolutional Neural Network Training**

All model training was performed in Python 2.7 on Compute Canada and Calcul Quebec computing clusters. The network was trained for 35 epochs, each with 20 sub epochs. 1000 segments were loaded for training per sub epoch, with the batch size set at 10. Data augmentation consisted of reflection along the sagittal axis. An initial learning rate of 0.001 was used and decreased through training by polynomial decay. Root mean square propagation was used as the optimizer. L1 and L2 regularizations of 0.000001 and 0.0001 were used, respectively.

**A.5. Post-processing**

Post-processing was done exclusively using the ITK toolkit via the SimpleITK implementation. Post-processing was conducted on the binary lesion segmentations produced by the trained DeepMedic model. These segmentations were passed through the CastImageFilter with OutputPixelType set at 4 (32-bit signed integer) in order for the images to be compatible with the necessary filters.

Using a connected components analysis, components below 1.5 mL in volume were removed. The exception was in segmentations where the largest connected component was smaller than the cutoff, in which case no cutoff was used. Finally, hole-filling with a radius of 3 voxels was used to fill gaps within the segmentation.

The validation dataset was used to optimize the minimum object size threshold and the hole-filling kernel radius. The minimum object size threshold was optimized first, by varying the threshold range from 0.3 mL to 2.5 mL. The value that maximized the DSC was 1.5 mL. Using this threshold, the hole-filling radius was optimized next using values of 2, 3, 5, 7, and 10 voxels. As hole-filling causes the segmented lesion volumes to grow, and subsequently increased the error in lesion volume estimates, both the DSC and lesion volume error were considered when choosing the optimal value. More precisely, the DSC was maximized while the lesion volume error was minimized. The optimal radius was found to be 3 voxels.
Appendix B. Additional Figures

Figure 2: Agreement analysis of DSC score between individual human observers and automated segmentations in the test dataset. Bland-Altman plots show very little bias toward either observer A or observer B in segmentations by (A) CNN (−0.012 ± 0.044) or (B) CNN + post processing (−0.017 ± 0.054).
Figure 3: Agreement analysis of lesion volume estimates from individual human raters and automated segmentations in the test dataset. Bland-Altman plots showed minimal bias in lesion volume estimates when comparing (A) rater A to CNN (14.0 ± 53.7 mL), (B) rater A to CNN + post-processing (11.9 ± 50.5 mL), (C) rater B to CNN (19.5 ± 53.7 mL), (D) rater B to CNN + post-processing (17.4 ± 50.6 mL).