Residual H-Net: A high-level feature extraction approach for brain lesion segmentation

Kefan Song[1, 5,†], Yi Sun[2, †], Lincong Wang[3, †] and Yuqiao Xue[4, †]

[1]Emory University, Atlanta, USA
[2]University of Washington, Seattle
[3]University of California Irvine, Irvine
[4]University of Florida, Florida, USA

[5]ksong23@emory.edu
[†]These authors contributed equally.

Abstract. In the case of acute stroke patients, mere diagnosis falls short, and segmentation is needed. Recent development in deep learning and image processing has provided us with the potential to automatically perform brain lesion segmentation. However, many of the approaches ended up failing to generalize to new data by overfitting the ATLAS R1.2 dataset and ignoring information extraction of the high-level features. We propose a novel Residual H-Net that addresses these two issues by adding a special residual block in the middle of the U-Net and increasing dilation size to better extract the high-level features. The presented model is less susceptible to overfitting and much easier to train. The Residual H-Net is tested on a subset of ATLAS R2.0 data and shows promising performance against the previous state-of-the-art model.

Keywords: Brain lesion segmentation, deep learning, residual network, deep convolutional neural networks, high-level feature abstraction

1. Introduction

A stroke is a medical condition in which a blood clot blocks the flow of blood and prevents the brain from getting oxygen. As a matter of fact, stroke is one the most common leading cause of death around the world. Each year, fifteen million people worldwide suffer strokes, of which five million die and five million become permanently disabled [1]. There are two main types of strokes: ischemic and hemorrhagic. Both prevent the brain from functioning properly and eventually lead to cell death. This report will be focused on Ischemic Stroke, and we will be using computer vision techniques and convolutional neuro network models to diagnose the infarct lesions caused by Ischemic Stroke on patients' brains.

Ischemic Stroke is very common among people who are over forty years old. It happens when a blood clot blocks or narrows the blood vessel in the brain, which stops the blood from flowing into the brain. The decreased blood supply would cause dysfunction of brain tissues, and the brain cells are likely to die very soon. The area of the brain where the blood supply is decreased and cells are damaged or dead is called a lesion. Simply put, a lesion is any damaged region of the brain caused by an Ischemic Stroke. When treating ischemic Stroke, it is helpful to first identify the damaged areas, or the lesions, in the patient's brain. There are various methods that are being used to diagnose stroke lesions, including CT imaging, MRI scans, and lab blood tests. Besides the clinician's judgment that is based on human expertise, it is of great value to have an automatic brain lesion segmentation to provide a secondary decision. With the advance of the deep convolutional neural network, automatically segmenting the brain lesion area on a large clinical dataset becomes a possibility.

In this paper, we will be using images produced by MRI scans. More specifically, we will first construct a convolutional neural network model and then train this model to identify and locate the lesions of patients' brains in these MRI images. In the task of image segmentation, U-Net, a specific form of the convolutional neural network, is often used. However, many approaches fail to generalize to new clinical data since they tend to overfit the training dataset, most commonly used ATLAS r1.2, as mentioned on the dataset's website [2]. Two factors come into play of such a phenomenon. Firstly, even though the ATLAS r1.2 dataset contains tens of thousands of images, it is considered a very small dataset by the number of patients: it only contains 229 patients' 3D MRI brain scans. The adjacent brain lesion images within a patient's scan are considered correlated with each other, lacking the variability necessary for the model to generalize well. Secondly, since all of the training and testing data are open to the public, most methods proposed in the literature lack a true test set: the reported scores are the validation scores. To address this overfitting issue, we trained our Residual H-Net on a larger dataset (ATLAS r2.0), and utilized the normalization properties of residual blocks.

Recent developments in deep learning have shown the effectiveness of training a deep convolution neural network in improving the performance of various computer vision tasks, for example image classification [3]. Residual connection is proposed by He et al., in order to make deeper neural network easy to train with performance gain. Residual connection is widely used in medical image segmentation [4, 5, 6]. However, previous models merely utilize residual connections in a way to focus on extracting the low-level features, as most of the convolution layers and residual connections are designed during the downsampling and the upsampling procedure, while leaving the entire high-level features only to an attention module [7], a non-local module [6], or, in many cases, skipped [4, 5]. In order to better exploit the high-level features, we proposed Residual H-Net, with 15 and potentially more residual blocks at the center of a typical U-Net architecture that solely focus on high-level feature extractions.

2. Related works

Previous research has implemented fully convolutional layers to biomedical image segmentation. It has a fundamental but solid method that can largely help convolutional neural networks have a better performance. Badrinarayanan et al [4]. propose SegNet, an encoder-decoder convolutional neural network for image segmentation. The encoder network is basically a VGG-16 model, whereas the decoder network functions to transform the low-resolution encoded features back into an image. The novelty here is that SegNet makes sure the decoding upsampling procedure uses the corresponding pooling indexes of that of the encoding layers. Such design makes sure that the model can focus on upscaling features into images without learning the upsampling per se, which gives the model a superior performance. Ronneberger et al [4]. add additional layers to a typical contracting network, replacing the pooling operators with upsampling operators. As a result, the output's resolution is increased by these layers. High-resolution features from the contracting route are merged with the upsampled for localization. Based on this knowledge, a subsequent convolution layer may subsequently learn to put together a more exact result. He et al. [8] proposed a spatial pyramid pooling layer in a convolutional neural network to accept arbitrary input sizes of the image. The SPP layer (spatial pyramid pooling layer) is added at the end of the convolution layers before densely connected layers, which would accept the arbitrary size of convolution output and produces a fixed-sized output. However, some [6] argues that such a design, when implemented in image segmentation, would result in too many redundant parameters. Chen et al. [9] propose an ensemble of DeconvNets CNN and a critic convolutional neural network that takes the output of the ensemble models as input, then identifies and removes potential false positives, achieving a good result. Qi et al. [6] proposed a revision to the traditional U-Net model, namely X-Net, by decreasing the number of trainable parameters and changing the convolutional layers during MaxPooling and UpSampling into residual blocks, which is said to improve overfitting and reduce dimensions. X-Net also has a non-local block in the middle of the high-level features, which the authors claim to improve Long-Range Dependencies Extraction. Zhou et al. [10] propose a novel D-UNet that performs as fast as a 2D convolution neural network but performs close to a 3D convolution neural network. The proposed method cleverly passed the 3D convolution during the downsampling phase before squeezing it into a 2D shape and adding it back to the original 2D convolution. This method currently outperforms all previous models.

3. Approach

To classify each pixel into lesion pixels or non-lesion pixels, we apply the U-Net architecture and residual blocks to output an image showing the brain lesion image segmentation. The U-Net architecture generally contains MaxPooling layers in the first half of the model to turn the high-resolution data into low-resolution features; and Upsampling layers in the second half of the model, as shown in figure 1. The convolution layers in the U-net are contained in residual blocks. These two design choices are also adopted in other segmentation implementations, such as X-Net [6]. In this paper, we propose a special design of U-Net, with residual block placing at the center, named Residual H-Net, to better harness the information from the low-resolution features.

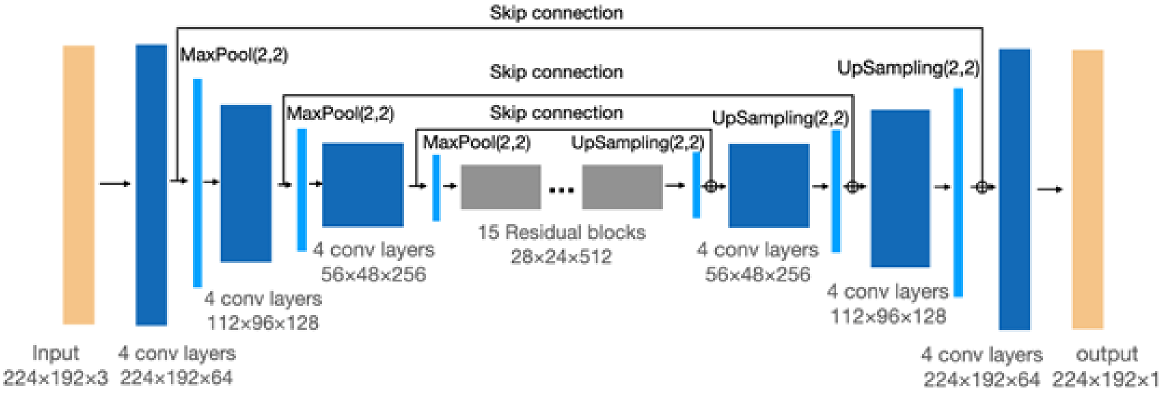


Figure 1. The model architecture of our Residual U-Net, where additional 15 residual blocks are added after MaxPooling and before UpSampling, transforming the conventional U-Net into an "H" shape architecture.

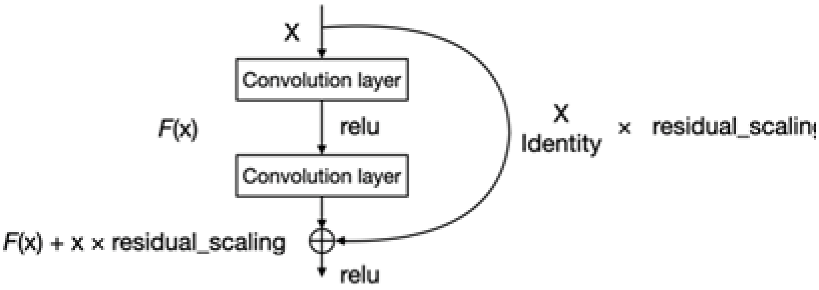


Figure 2. The residual block with residual rescaling used in our model.

3.1. Residual Blocks for Low-Resolution features

A residual block is shown in Figure 2. It is also named skip connection, because besides the traditional forward pass, the residual block also sends the output of previous layers to the current layer by performing a piecewise addition operation. Therefore, by adding the identity of the previous layer times a residual scaling ratio (in our case is set to 0.1), a residual block preserves the original information of the input, thus allowing the model to go as deep as 2000 layers while being easy to train and superior in performance[11].

The brain lesion image segmentation is often evaluated using a dice score, which is the intersection over the union of the truth lesion image and the predicted lesion image. Previous methods focus on capturing high-resolution features by adding more layers both at the start and the end of the model [2, 4, 8]. However, the best models among them are only getting a dice score of roughly 0.5 [8, 9], about half the size of the target area. This shows the models couldn't mostly circle out the lesion yet. This may indicate that pushing for higher resolution would only provide marginal benefit. A pure pixel-level oriented design is limited by its capacity to identify features with large receptive fields, which mostly lie in the middle of a U-Net model. Therefore, here we propose a special design of residual block to better extract information from these high-level features.

For each residual block, we start with a convolution layer with 512 channels, 3 kernels, and stride equals 1, which is followed by a ReLU layer. The same convolution layer and ReLU layer are repeated once before the output of the first convolution layer is added to the output of the last ReLU layer. The result feature mapping is then the final output of the residual block. In our special residual block, we intentionally remove the batch normalization layer, as such design improves performance in other U-Net [12].

3.2. The Modified U-Net

As suggested by the X-Net paper, the vanilla U-Net may lead to overfitting since it doesn't contain a residual block. In addition, The architecture of X-Net contains 4 blocks of convolution and pooling layers during each upsampling and downsampling process, requiring layers with the high-level features to have 1024 channels. In this paper, we follow the same design of the residual block, and uses only three blocks of convolution and pooling layers at each sampling stage to avoid making our model parameters unnecessary large. Each of our residual blocks now only needs to have 512 channels. The kernel size is (3,3). They are followed by a 1 x 1 convolution layer before the residual neural network output. The first block is 224 by 192 by 64, where (224, 192) is the input image size and 64 is the channel size. Then a MaxPooling layer of (2, 2) is added. This transforms the output feature into the shape of (112, 96, 64). Then the same convolution block with MaxPooling is repeated again, changing the channel from 64 to 128, making the output feature into the shape of (56,48,128), after which the previously mentioned special residual blocks are added. In the second half of the U-Net, the model has the shape of convolution blocks in reverse. Starting with input shape of (56,48,128), outputting the shape of (112, 96, 64). Each is preceded by an upsampling layer of (2, 2), thus reversing the low-resolution feature back into a full-fledged image. To speed up training, we only use 3 convolution blocks for each sampling stage in U-Net, instead of 4 proposed in X-Net with acceptable performance compromise.

To make our model focuses on large features instead of pixel-level details, the receptive field of convolution layers in the U-Net is implemented. The dilation size of the first two and the last two convolutional layers are to 6 and 8 respectively.

4. Experiments

4.1. Dataset

Our model is built around brain lesion datasets such as Anatomical Tracings of Lesions After Stroke (ATLAS) R2.0, which is an open sourced dataset. This dataset is an improvement over the ATKAS R1.2 dataset from 2018, which has the same format as the data. The dataset is increased from 229 T1-weighted normalized 3D MRI images with manually-segmented lesion masks to 955 T1-weighted MRI scans. 655 of them are training dataset, while the rest 300 is used for test dataset, which is not yet released. Each 3D MRI image represents one patient's brain scan reading, which consists of 189 slices of images. Since we only formulate this problem as 2D convolutional neural network mask image generation, the preprocessed dataset will contain 123795 images, which is too large to run on the Google Colab environment. Therefore, we take a subset of the training data to speed up experimentation and comparison between methods. The positive samples of 255 patients were chosen as the training set, and 30 patients were arbitrarily chosen as the validation dataset.

Implementation Our model is implemented in Keras. Adam is used to automatically adjust the learning rate. The learning rate is set to 0.001 to begin with. We use the sum of cross entropy loss and the dice loss as the loss function. The training batch size is set to 8, and the evaluation batch size is set to 1. The experiment is run on NVIDIA T4 GPU on Google Colab.

4.2. Evaluation Metrics

The performance of the models is measured by Dice, which is widely used in medical image segmentation tasks[13]. It is slightly different than accuracy but aimed at the same goal: describe the overall accuracy of the model. The new metric dice is calculated by 2 * the Area of Overlap divided by the total number of pixels in both images. Briefly meaning is the 2 * the sum of the intersection of the predicted response variable and the true response divided by the union of the two responses [1]. The reason why accuracy is abandoned is that since every brain slice has a size (224,192) after cutting the edges, (edges are cut because they have a paucity of brain pixels and will add bias and noises to the model), there is a relatively very small region that could have lesions on the brain. It means that if we use accuracy, the main part of the image will remain the same. That output will make accuracy extremely high, which makes the real performance overrated. Besides, it is very hard to tell the difference between each model since their accuracy will all be very high. By contrast, dice is a suitable metric that has a much small value due to its calculation. Since the goal is to learn and predict every pixel on the slices instead of making predictions of the final result (whether it has lesions or not), dice can precisely generate the overall performance based on each pixel.

4.3. Results

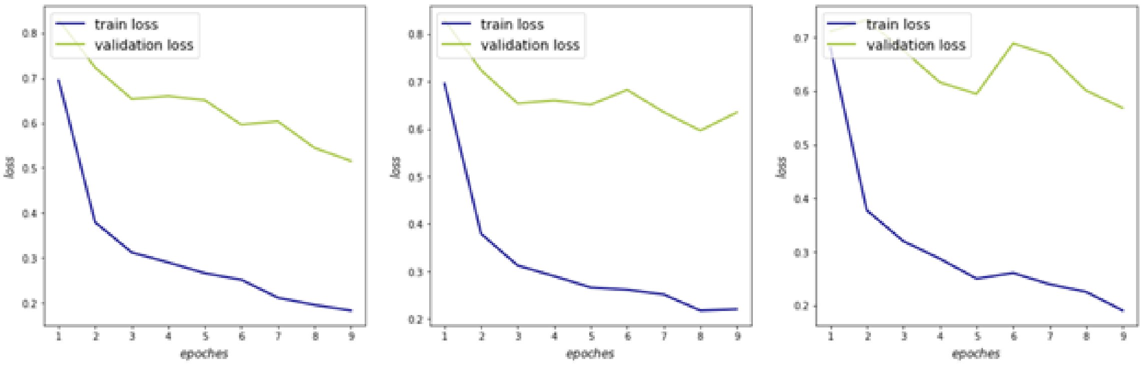
After several trials and improvements, the new model built for predicting Ischemic stroke in patients' images dataset has an obvious improvement compared to the known X-net model. The results of all models were analyzed based on the same subset of the original data and the same data generator pipeline settings.

We conducted experiments on 3 different models: the original X-net model, the U-Net model with 3 X-net conv blocks, and our model. Since our model only has 3 X-net conv blocks, we not only compared it with the original X-net model but also with the U-Net model with 3 conv blocks to observe the real performance of our model.

Here are the results of these 3 models below:

Table 1. Results of the models.

	our model	U-Net (X-net with 3 conv blocks)	original X-net
train dice	0.8696	0.8535	0.8751
validation dice	**0.5777**	0.4850	0.5329



(a) Our Model
Model

(b)U-Net (X-Net model with 3 conv blocks)

(c) Original X-Net

Figure 3. Loss Figure.

The results table labels that our model has a higher validation dice compared to the original X-net model. That indicates that with fewer layers and parameters, our model still has a better performance. In other words, with similarly high accuracy performance, our model has better efficiency and saves more memory. When comparing it to the U-net model with 3 X-Net conv blocks, the dice are higher than the model with only 3 conv blocks. That means our model has far exceeded the U-Net model.

Some of the actual segmentation results are presented below in figure 3. We can see that our purposed model is able to output superior results against the state of the art X-net in many cases. In addition, we can verify that the 15 added residual blocks are critical in terms of improving the model performance. When the residual blocks are ablated, our model becomes identical to U-Net with 3 conv blocks, which has a significantly worse segmentation results.

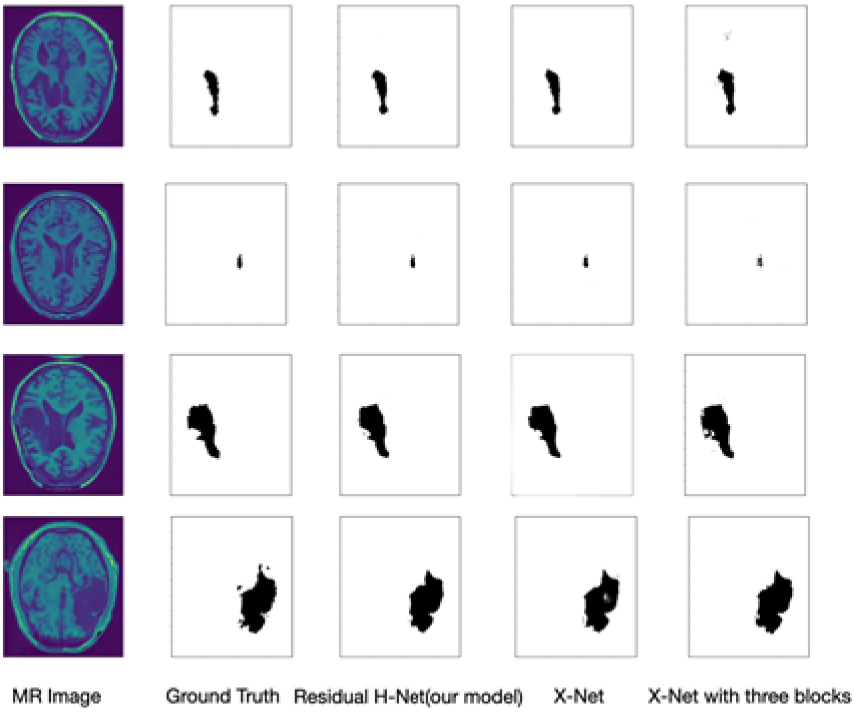


Figure 4. Segmentation results of our model and SOTA.

5. Conclusion

We present a modified model based on the vanilla U-Net Model for brain stroke lesion segmentation analysis and prediction. Due to the lack of feature extraction for high-level features, applying the U-net architecture will cause some underfitting during training. Therefore, our new modified model successfully addresses this issue by applying 15 residual blocks at the middle part of the U-net model, transforming into an “H” shape network. Such application of residual connections facilitates the information propagation, allowing the network to go deeper at the high-level features, and ultimately provides better performance.

Given the performance difference between training on positive samples and training on both samples, one future direction would be to improve performance on both negative and positive datasets by decomposing the lesion segmentation task into two tasks: a classification task and an image segmentation task. An additional classifier can be trained to specifically identify negative samples, helping the segmentation model to reduce false positives. With abundant computing resource, data augmentation and an even higher number of residual connections are also promising to further improve the performance.

References

[1] World Health Organization . Atlas of heart disease and stroke. Geneva, Switzerland: World Health Organization, 2004.

[2] Anatomical Tracings of Lesions After Stroke (ATLAS) R2.0 http://fcon_1000.projects.nitrc.org/indi/retro/atlas.html

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.90

[4] Badrinarayanan V., Kendall A., Cipolla R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12), 2481–2495 (2017)

[5] Ronneberger O., Fischer P., Brox T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer, Cham (2015)

[6] Qi, K., Yang, H., Li, C., Liu, Z., Wang, M., Liu, Q., &amp; Wang, S. (2019, December 30). X-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. arXiv.org Retrieved September 18, 2022, from https://arxiv.org/abs/1907.07000

[7] Hui, H., Zhang, X., Li, F., Mei, X., & Guo, Y. (2020). A partitioning-stacking prediction fusion network based on an improved attention U-Net for stroke lesion segmentation. IEEE Access, 8, 47419–47432. https://doi.org/10.1109/access.2020.2977946

[8] He K., Zhang X., Ren S., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(9), 1904-1916 (2015)

[9] Chen L, Bentley P, Rueckert D. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks[J]. NeuroImage: Clinical, 2017, 15: 633-643.

[10] Zhou, Yongjin, et al. “D-Unet: A Dimension-Fusion U Shape Network for Chronic Stroke Lesion Segmentation.” ArXiv.org, 14 Aug. 2019, https://arxiv.org/abs/1908.05104v1.

[11] Deep residual learning for image recognition - arxiv. (n.d.). Retrieved September 18, 2022, from https://arxiv.org/pdf/1512.03385.pdf

[12] Lim, B., Son,S., et al. ArXiv:1707.02921v1 [CS.CV] 10 jul 2017. (n.d.). Retrieved September 18, 2022, from https://arxiv.org/pdf/1707.02921.pdf

[13] Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., & Blaschko, M. B. (2019). Optimizing the DICE score and Jaccard Index for Medical Image Segmentation: Theory and Practice. Lecture Notes in Computer Science, 92–100.

https://doi.org/10.1007/978-3-030-32245-8_11