Skull Stripping using Confidence Segmentation Convolution Neural Network

Kaiyuan Chen*, Jingyue Shen*, and Fabien Scalzo

Department of Computer Science and Neurology
University of California, Los Angeles (UCLA), USA
{chenkaiyuan,brianshen}@ucla.edu

Abstract. Skull stripping is an important preprocessing step on cerebral Magnetic Resonance (MR) images because unnecessary brain structures, like eye balls and muscles, greatly hinder the accuracy of further automatic diagnosis. To extract important brain tissue quickly, we developed a model named Confidence Segmentation Convolutional Neural Network (CSCNet). CSCNet takes the form of a Fully Convolutional Network (FCN) that adopts an encoder-decoder architecture which gives a reconstructed bitmask with pixel-wise confidence level. During our experiments, a crossvalidation was performed on 750 MRI slices of the brain and demonstrated the high accuracy of the model (dice score: 0.97 ± 0.005) with a prediction time of less than 0.5 seconds.

Keywords: MRI · Machine Learning · Skull Stripping · Semantic Segmentation

1 Introduction

Computer-aided diagnosis based on medical images from Magnetic Resonance Imaging (MRI) is used widely for its 'noninvasive, nondestructive, flexible' properties [3]. With the help of different MRI techniques like fluid-attenuated inversion recovery (FLAIR) and Diffusion-weighted (DW) MRI, it is possible to obtain the anatomical structure of human soft tissue with high resolution. For brain disease diagnosis, in order to check interior and exterior of brain structures, MRI can produce cross-sectional images from different angles. However, those slices produced from different angles pose great challenges in skull stripping. It is hard to strip those tissue of interest, from extracranial or non-brain tissue that has nothing to do with brain diseases such as Alzheimers disease, aneurysm in the brain, arteriovenous malformation-cerebral and Cushings disease [3].

As a prerequisite, skull stripping needs to produce fast prediction speed and accurate representation of original brain tissue. In addition, since the MR images can be taken from different angles, depth and light conditions, the algorithm needs to have great generalization power while maintaining high accuracy. Figure 1 illustrates the challenging nature of 2D skull stripping with some examples

* Equal Contribution

from our dataset. It can be seen that the non-brain structure appears very similar to the brain tissue. Sometimes brain tissue only occupies small part of the image and the boundaries between brain tissue and non-brain structures are not clear. The parameters of the MRI may lead to different intensity profiles. In addition, the image structure varies from one slice to another. Sometimes the slices do not hold any brain tissue.

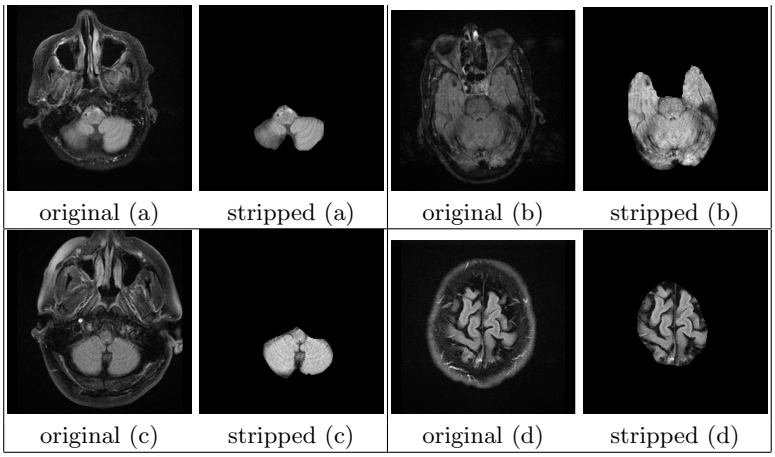

Fig. 1. Illustration of challenges posed during skull stripping.

Our contributions are as follows:

- We design CSCNet, a deep learning architecture that can be applied to skull stripping with confidence level.
- We use series of experiments to show our model is fast and accurate for reconstructing skull stripped images.

We organize this paper as following: we first introduce basic terms like CNN and related algorithms on skull stripping in Section 2; then we continue to strip skulls with confidence level in Section 3; we show our experimental results in Section 4; conclusion and future work are in Section 5.

2 Basic Concepts and Related Works

2.1 Traditional Skull Stripping Algorithms

As a preliminary step for further diagnosis, skull stripping needs both speed and accuracy in practice, so these two factors should be considered in any algorithms proposed. By Kalavathi et al. [3], skull stripping algorithms can be classified into five categories: mathematical morphology-based methods, intensity-based methods, deformable surface-based methods, atlas-based methods, and hybrid

methods. In past few years, machine learning-based methods also show great results in skull stripping. For example, Yunjie et al.[6] developed a skull stripping method with an adaptive Gaussian Mixture Model (GMM) and a 3D mathematical morphology method. The GMM is used to classify brain tissue and to estimate the bias field in the brain tissue. Butman et al.[11] introduced a robust machine learning method that detects the brain boundary by random forest. Since random forest has high expressive power on voxels of brain boundary, this method can reach high accuracy. However, these methods of skull stripping usually take local patches and conduct prediction pixel by pixel. so if the training dataset is not complete or testing data contains too much noise, the resulting images will have high variance as we shown in our experiments.

2.2 Convolutional Neural Networks

Convolutional Neural Networks Convolutional Neural Networks (CNNs) have shown great success in vision-related tasks like image classification, as shown in the performance of AlexNet[14],VGGNet[15], GoogLeNet[16]. Compared to traditional pixel-wise prediction by feeding local patches to linear or nonlinear predictors, CNNs can give better results due to their ability of joint feature and classifier learning[17].

Semantic Segmentation In this paper, we model skull stripping as a special case of semantic segmentation. Semantic Segmentation is a process that associates each image pixel with a class label. Fully Convolutional Network (FCN), proposed by Long et al.[8] provides a hierarchical scheme for this problem. Many works on semantic segmentation like SegNet[1]and U-Net[2] are built on top of FCN[8]. FCN and most of its variations take an encoder-decoder architecture, where encoders try to learn features at different granularity while decoders try to upsample these feature maps from lower resolution to higher resolution for pixel-wise classification. The encoder part is usually modified version of pre-trained deep classification network like VGG[15] or ResNet[18], and the decoder part is where most works differ. In FCN, each decoder layer takes the combination of the corresponding encoder max-pooling layer's prediction result (by applying a 1×1 convolution on the max-pooling layer) and previous decoder layer's result as input to do $2\times$ upsampling, in an effort to make local predictions that respect global structure[8]. In U-Net, a major modification from FCN is that it keeps the feature channels in upsampling. Instead of combining prediction results from corresponding encoder layer and the previous decoder layer, it combines feature maps of the two layers as input to next decoder layer, allowing the network to propagate context information to higher resolution[2]. For SegNet, it is more efficient in terms of memory and computation time during inference, and produces similar or better performance than FCN on different metrics. SegNet discards the fully connected layers of VGG16 in encoder part, which makes the network much smaller and easier to train. It features a more memory-efficient decoding technique compared to FCN and U-Net. It uses unpooling[21] rather than

deconvolution to do upsampling[1]. Instead of remembering the whole feature maps in encoder part, SegNet only memorizes the indices selected during max-pooling, which only requires 2 bits to store per 2×2 pooling window, and uses them to perform upsampling[1]. This method eliminates the need for learning in upsampling step, and further reduces the number of parameters.

Architectures not based on FCN[8] are also proposed for semantic segmentation, like ENet[5] and PSPNet[9]. ENet also takes the general encoder-decoder approach, but designs its own encoder and decoder part from ground up. It has very fast inference time while having similar performance compared to SegNet, which shows great potential in applying to real-time low-power mobile devices[5]. PSPNet uses global scene context of an image as prior to help better predict pixels' labels. It utilizes a Pyramid Pooling Module to extract global contextual prior of input images and combines the feature map from its modified ResNet architecture with the prior to produce final prediction map[9].

Another work worth mentioning is Baysian SegNet by Kendall et al[19]. The authors tried to model predictive system's uncertainty and produce probablistic segmentation output by adding several dropout[20] layers in SegNet. They trained the model with dropout, and at test time they used dropout to do Monte Carlo dropout sampling over the network's weights to get the softmax class probabilities. This method shows better results than SegNet, but can lead to longer inference time due to sampling[19].

CNN-related works on skull stripping For 3D CT images, Kleesiek et al.[7] used non-parametric 3D Convolutional Neural Network (CNN) to learn important features from input images. However, 3D images usually contain more structural information than 2D images, so for models applying to 2D gray scale MRI should rely less on positional and spacial structures than 3D CNN. Some recent works also proposed FCN-based architecture and achieved better performance than traditional methods. For example, Zeynettin et al. [13] utilized fully convolutional network similar to U-Net to perform skull stripping. Raunak,D., and Yi H.proposed an architecture called CompNet[4] that learns features from both brain tissue and its complementary part outside the brain to do skull stripping.

3 Proposed Method

In this section we introduce our proposed Confidence Segmentation convolutional Neural Network (CSCNet). We model this problem as semantic segmentation, and adopt an architecture similar to SegNet[1]. By making use of properties of gray scale MR images, we generate a confidence level matrix as a bitmask and apply the bitmask to the original MR Image.

3.1 Semantic Segmentation with Confidence level

We model 2D Skull Stripping problem as a semantic segmentation problem, i.e. given an image as matrix $\mathbf{X}$, we can view it as a sum of skull matrix $\mathbf{S}$ and

stripped matrix $\mathbf{X}'$ with dimension w and h, i.e.

$$X = S + X'$$

Given $\mathbf{X}$, our algorithm will produce a confidence matrix $\mathbf{\Theta}$ that, with a hyper-parameter confidence level α, it will give a skull matrix

$$\mathbf{S}_{ij} = \begin{cases} \mathbf{X}_{ij} & \mathbf{\Theta}_{ij} < \alpha \\ 0 & \mathbf{\Theta}_{ij} \geq \alpha \end{cases}$$

and reconstructed image

$$\mathbf{X}'_{ij} = \begin{cases} \mathbf{X}_{ij} & \mathbf{\Theta}_{ij} \geq \alpha \\ 0 & \mathbf{\Theta}_{ij} < \alpha \end{cases}$$

3.2 Architecture

Encoder-decoder Architecture To handle pixel-wise labeling, most Fully Convolutional Network (FCN)s adopt an encoder-decoder scheme. FCN is translation invariant since it leverages operations like convolution, pooling and activation functions on relative distance space.

Under this encoder-decoder scheme, input images first go through convolutions and pooling layers to reduce to a more compact low dimensional embedding, and then the target images are reconstructed by deconvolutions and upsampling. Because of the characteristics of FCN, models are usually trained in an end-to-end fashion by back-propagation and weights are updated by Stochastic Gradient Descent (SGD) efficiently.

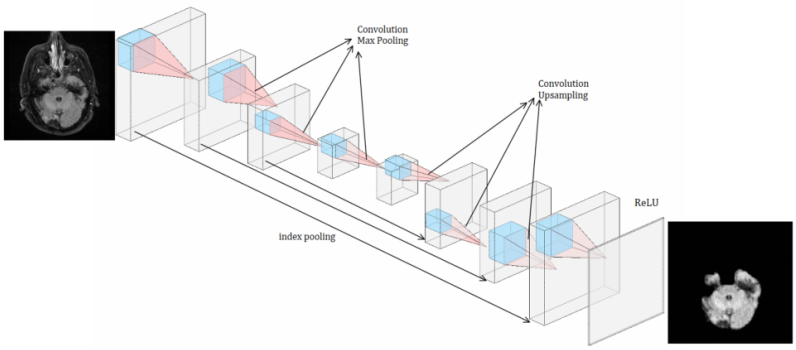

Fig. 2. Illustrattion of the CSCNet architecture for skull stripping.

Encoder network for our model has 8 convolutional layers. For each encoder layer, it uses a filter bank to do convolutional with bias, followed by batch normalization and element-wise rectified linear non-linearity (ReLU), i.e. $\max(0, x)$.

After going through encoder's convolution, batch normalization and ReLU, we perform a max-pooling operation. Our pooling window is 2×2 and to prevent overlapped pooling, our stride is also 2.

In decoder network starting from low dimensional encoded embedding, we adopted a deconvolutional scheme similar to Semantic Segmentation Fully Convolutional Networks [8]. We have 8 deconvolutional layers with upsampling and deconvolutions. These decoder layers upsample from input using transferred pool indices to generate feature maps. However, instead of strictly applying softmax in original SegNet, we apply a ReLU activation function and compare the results with gray scale images. The Relu is applied to construct a confidence level for each pixel in the image.

From Confidence level to output Image Empirically, our model cannot reproduce exact brain tissue while generalizing a large training dataset. Thus, we need to use the output of activation function trained based on target image, which will have higher confidence to brain tissue if it has higher activation state. Thus, we perform bitmask by equation 3. Empirically we choose 0.5 in our own training dataset described in 4.1 and the procedure is on figure 3. Tuning the parameter α is not required, but it does represent a true negative and false positive trade-off in terms of classification results.

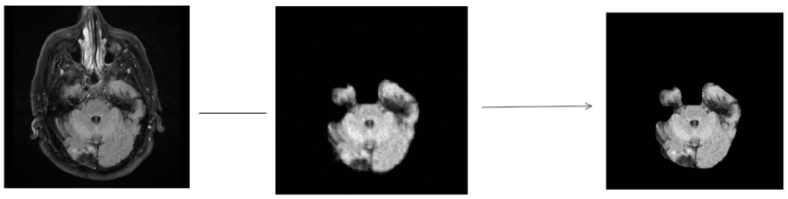

Fig. 3. Constructing brain-tissue images from confidence level matrix.

4 Experiment

In this section, we describe our experimental protocol and present the results.

4.1 Dataset Description

Our dataset consists of around 750 manually labeled skull stripped fluid-attenuated inversion recovery (FLAIR) brain slices from 35 patients provided by University of California, Los Angeles. The use of this dataset was approved by the Institutional Review Board (IRB). Each patient's brain MRI was composed of between 15 to 30 MR slices. The difficulty of skull stripping lies in the fact that brain structure varies due to different diseases patients have, and the challenges mentioned in Figure 1 in Section 1.

4.2 Metrics

We use Dice Score as our major evaluation metric, which is defined as:

$$DiceScore = \frac{2|X \cap Y|}{|X|+|Y|},$$

where X represents the skull stripping results of our model and Y represents the ground truth stripped images. $|X \cap Y|$ means the intersection between ground truth and model's result. Here, for each image in our dataset, we consider $|X \cap Y|$ as the number of correctly classified pixels in an image. Beside this metric, which is commonly used in skull stripping, we also evaluate our model on other traditional classification metrics like Precision and Recall. In the context of classification problem, they are defined as follows:

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive}$$

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative}$$

4.3 Experiments Setup

We performed five-fold crossvalidation with 600 images as training set and remaining 150 images as validation set. The images are of size 256 $\times$ 256 pixels. We treated each patient as a learning unit and put all images of one patient either in training set or in test set. By this way all testing models are able to see different structures of brain and skull in both training and validation set. This approach increases the validity of our experiments' results. As a trade-off between accuracy and prediction time, we choose filter size 16 and 32.

4.4 Descriptions of competing Models

We choose traditional pixel-by-pixel prediction random forest classifier as our baseline model. There are many existing classifiers like SVM, logistic regression, random forest that take in a local patch surrounding a single pixel and predict as a bitmask. Taking into account the training efficiency and prediction time, we choose random forest [11] as our baseline model and use 3×3 local patches along with brightness and position information as features.

Other competing state-of-the-art models include FCN [13] and CompNet [4]. For FCN [13], the architecture consists of two parts. The first part has three convolutional layers with three max pooling layers. The second part consists of three convolutional layers with three deconvolutional layers and one final reconstruction layer. For all models, we treat the outputs as bitmasks and then applied them to unstripped images to generate clearer and better results.

The accuracy of the competing models are shown in Table 1:

Model	Accuracy	Dice Score	Precision	Recall
Random Forest [11]	0.8538	0.6439	0.8816	0.50621
FCN[8]	0.9647	0.9066	0.9005	0.9213
CompNet [4]	0.9804	0.9468	0.9276	0.9711
CSCNet-16 (our model)	0.9787	0.9510	0.97664	0.9344
CSCNet-32 (our model)	0.9859	0.9701	0.96157	0.9830

Table 1. Accuracy of models in different metrics

Model	Prediction Time
Random Forest	35.32
FCN	0.88
CompNet	75.46
CSCNet-16 (our model)	0.31
CSCNet-32 (our model)	0.43

Table 2. Prediction time of the models

4.5 Prediction Time Experiment

We ran all predictions on a personal laptop with single-core Skylake processor and 8G RAM without GPU to simulate actual scenario.

As shown in Table 1 and Table 2, our models (both CSCNet-16 and CSCNet-32) have high accuracy and dice score. They outperform the baseline Random Forest model and other state-of-art FCN-based models. Although CompNet has high prediction accuracy, its reliance on dense convolutional blocks makes its prediction time not applicable to be used in practice, since skull stripping is only a preprocessing step which requires both accuracy and fast speed.

4.6 Examples of Skull Stripped Results

Figure 4 illustrates some of our skull stripping results. Our model is able to strip skulls from brain tissue given slices of a brain taken from different levels. During our experiments, we found that brain tissue in some images cannot be stripped cleanly from skulls. Such challenging case is illustrated in the upper right corner of figure 4. This could be caused by the blurry boundary condition in images, perhaps due to a motion artifact during the acquisition.

5 Conclusion

The challenging part of skull stripping on MRI is that, given the limited structural information, the model needs to identify brain tissue and distinguish it from skull which usually has similar intensity and appearance. Thus, modeling MRI brain skull stripping in terms of semantic segmentation shows a way to think about labeling these images with pixel confidence level. We propose

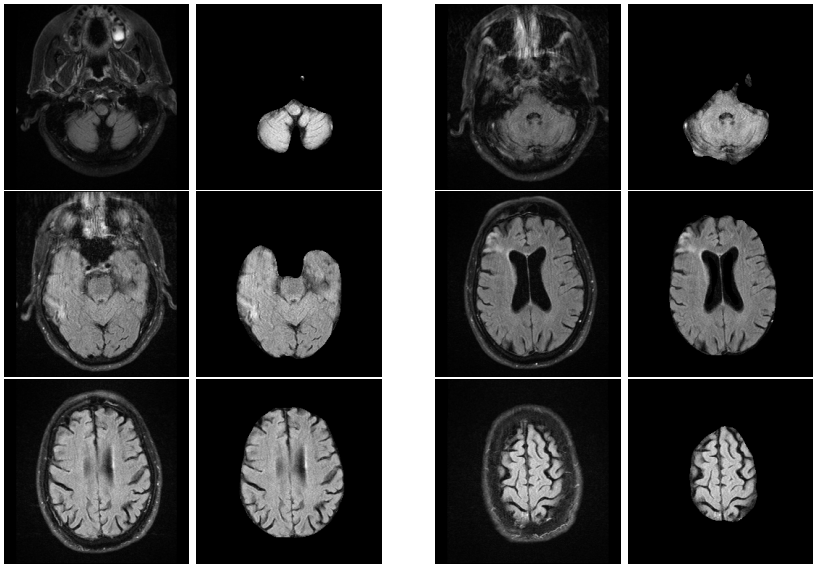

Fig. 4. Some test results of the skull stripping model on different brain slices.

Confidence Segmentation CNN, a deep learning neural network that augments semantic segmentation models and makes use of the knowledge that higher pixel intensity implies higher confidence in skull classification. Based on a series of experiments, our model has high dice score, precision and recall and is faster that state-of-the-art models.

References

1. Vijay Badrinarayanan, Alex Kendall and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, 2015; arXiv:1511.00561.
2. O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp. 234-241.
3. Kalavathi,P., and V.B. Surya Prasath. 2016. Methods on Skull Stripping of MRI Head Scan Images – a Review. Advances in Pediatries., U.S. National Library of Medicine. www.ncbi.nlm.nih.gov/pmc/articles/PMC4879034.
4. Raunak,D., and Yi H. 2018. CompNet: Complementary Segmentation Network for Brain MRI Extraction. *arXiv preprint arXiv:1804.00521v2*
5. Adam Paszke, Abhishek Chaurasia, Sangpil Kim and Eugenio Culurciello. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation, 2016; arXiv:1606.02147v1
6. Yunjie C, Jianwei Z, Shunfeng W. A new fast brain skull stripping method, biomedical engineering and informatics. Tianjin: Proc. 2nd International Conference on Biomedical Engineering and Informatics, BMEI09; 2009.

7. Kleesiek J, Urban G, Hubert A, Schwarz D, Maier -Hein K, Bendszus M, Biller A. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. NeuroImage . 2016;129:460-469.
8. J. Long, E. Shelhamer, and T. Darrell, Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 34313440.
9. Hengshuang Z., Jianping S., Xiaojuan Q., Xiaogang W., and Jiaya J. Pyramid scene parsing network. *CoRR*, abs/1612.01105, 2016.
10. Two-dimensional fast magnetic resonance brain segmentation. Suri JS IEEE Eng Med Biol Mag. 2001 Jul-Aug; 20 (4):84-95.
11. Butman J, Roy S, Pham D. Robust skull stripping using multiple MR image contrasts insensitive to pathology. NeuroImage . 2017;146:132-147.
12. Kleesiek J, Urban G, Hubert A, Schwarz D, Maier-Hein K, Bendszus M, Biller A. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. NeuroImage . 2016;129:460-469.
13. Zeynettin Akkus, Petro M. Kostandy, Kenneth A. Philbrick, Bradley J. Erickson.: Extraction of brain tissue from CT head images using fully convolutional neural networks. Proc. SPIE 10574, Medical Imaging 2018: Image Processing, 1057420 (2 March 2018). https://doi.org/10.1117/12.2293423
14. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
15. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
16. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.
17. Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, and Gang Wang. Recent advances in convolutional neural networks. *CoRR*, abs/1512.07108, 2015.
18. Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385, 2015.
19. Alex Kendall and Vijay Badrinarayanan and Roberto Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *CoRR*, abs/1511.02680, 2015.
20. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural net- works from overfitting. *The Journal of Machine Learning Research*, 15(1):19291958, 2014.
21. Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *CoRR*, abs/1311.2901, 2013.