CS 230 Project Report: Automatic Fetal Brain Segmentation for Prenatal MRI Screening

Cheng Chen
Department of Electrical Engineering
Stanford University
cchen91@stanford.edu

Pablo G. Diaz-Hyland
Department of Chemical Engineering
Stanford University
pablo98@stanford.edu

Abstract

Fetal brain imaging is the cornerstone of prenatal screening and diagnosis of congenital abnormalities. While fetal brain MRI has unrivaled diagnostic efficacy in detecting neurodevelopmental abnormalities, interpreting fetal brain MRI remains a tremendous challenge for radiologists and clinicians due to the rapidly changing architecture of the normally maturing brain. To facilitate fetal brain analytics, we propose the development of an automated semantic segmentation model for detecting fetal brain. We apply the models U-Net and SegNet to a new and private dataset from Stanford School of Medicine, and observe that both models perform similarly in training, with an IoU score of > 0.9. However, the IoU score of the dev set reaches a maximum of 0.78 despite regularization and early stopping. This variance problem may be associated to the limited amount of training data as well as inaccurate masks. Overall, we have proved the effectiveness of applying deep learning to the novel task of fetal brain segmentation, which can serve as the foundation for the development of a classifier to detect abnormalities in fetal brain growth.

1 Introduction

Fetal brain imaging has played a significant role in detecting brain growth abnormalities that may lead to brain disorders such as Alzheimer's disease, autism and schizophrenia. Rigorous evaluation of fetal brain maturity and development is particularly critical due to the high prevalence of anomalies (an estimated 3 in 1000 pregnancies) [1, 2]. To quantify neural growth, semantic segmentation measurements are needed. Semantic segmentation describes the process of associating each pixel of an image with a class label and its need is highlighted by the fact that an increasingly number of applications benefit from inferring knowledge from images. Medical imaging segmentation is a special type of class activation mapping problem [3]. The difference is mainly on number of object classes and required accuracy.

To facilitate fetal brain analytics, we propose the development of an *automated segmentation (pixel classification) model for detecting fetal brain*. Such a method can isolate fetal brain that can later be used as input for studies that evaluate normal *in utero* brain development from congenital malformations and disease-states using fetal brain MRI.

2 Related work

As the popularity of deep learning has grown over recent years, many semantic segmentation problems are being tackled with architectures involving convolutional neural networks. In 2014,

CS230: Deep Learning, Fall 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

Fully Convolutional Networks (FCNs) popularized the use of trained end-to-end CNN architectures for semantic segmentation [4]. Instead of having fully connected layers at the end which can hurt spatial information of pixels, FCNs apply up-sampling to the low-resolution maps inside the network.

Semantic segmentation models generally consist of two parts: encoder and decoder. The encoder is a pre-trained classification network that contains several stages of two or three convolution layers followed by a pooling layer (e.g. AlexNet, VGG-16, GoogLeNet and ResNet). The decoder projects the discriminative features learned by the encoder into the pixel space to obtain a classification. The main difference between prior work is on the decoder part. U-Net [5] has recently emerged as an efficient CNN-based algorithm for biomedical imaging segmentation. By using a contracting path to capture context and a symmetric expanding path that enables precise localization, U-Net can be trained with very few images to outperform prior methods. SegNet [6] is another deep CNN that was designed because many pixel segmentation algorithms produce coarse segmentation maps. This is due to max-pooling and sub-sampling methods that reduce image resolution. SegNet contains the same encoder network as the VGG16 network. SegNet's decoder network uses the max-pooling indices received from encoders for unpooling and deconvolution to produce dense feature maps. This reduces the number of parameters enabling end-to-end training, making SegNet memory-efficient. U-Net, on the other hand, does not need the max-pooling indices from the encoder for unpooling. Instead, the corresponding encoder layer is concatenated with the upsampled layer before it is passed to the following convolution layers.

3 Dataset and Features

The dataset is provided by Prof. Kristen Yeom (Stanford University School of Medicine), and it was collected and labeled recently. We are the first team that works on it.

The dataset consists of 690 fetal MRI scans in `dicom` format and the corresponding masks in `nifti` format. Each MRI scan contains a varying number of 512×512 greyscale images (slices), and is either from *axial*, *coronal*, or *sagittal* view (three anatomy views). The masks are manually generated by an experienced radiologist, and are of the same dimension as the input. Figure 1 shows a typical image and the corresponding mask in the dataset.

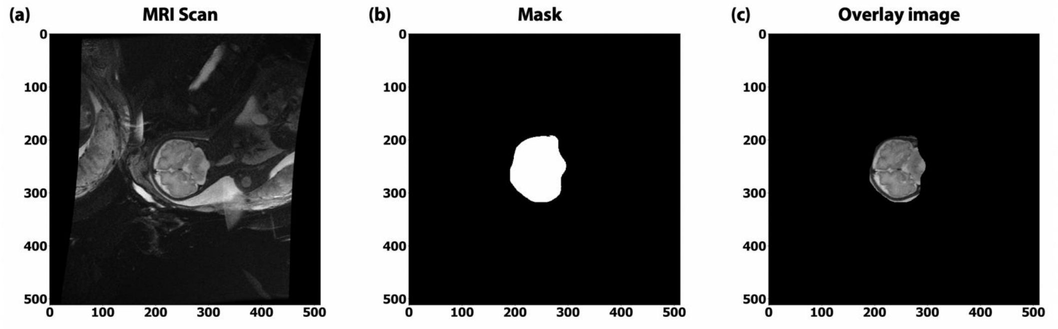

Figure 1: An example in the dataset. (a) One image (slice) in the MRI scan; (b) corresponding mask; (c) visualization for sanity check with the overlay image.

The varying and limited number ($15 \sim 35$) of slices for each data point hinders the implementation of a 3D model. As a result, we flatten the 3D scan into a series of 2D images, on which the model will perform the segmentation task. With this method, we obtained approximately 10,000 grayscale 2D images (input) and the corresponding masks (output).

The major challenge of the dataset is that fetal head is only a small part of the image, whereas most tissues in view are from the mother side. At pixel level, the positive in the masks consists of 2.7% of all pixels.

There are two additional challenges of the dataset: (1) the brain from different anatomy views have different features and boundaries from the surrounding tissues, and (2) in the same anatomy view at different slicing depth, the brain features as well as boundaries from surrounding tissues, are different. As a result, the model needs to learn a large number of different brain structures.

4 Methods

We applied U-Net and SegNet to the fetal brain segmentation task. A U-Net model from Github [7] was used as a starting point. We also referred to [8] when developing our SegNet model.

4.1 Models

Figure 2 shows the architectures of both models, which can be divided into ten stages. The first four stages ("encoder") perform convolution and max-pooling, followed by the fifth stage consisting of only convolution layers. The sixth to ninth stage perform unpooling/upsampling and (de)convolution ("decoder"). The last/output layer is a 1×1 convolution layer with the sigmoid activation function.

The main difference between the two models is that U-Net copies the layers from the "encoder" and concatenate them to corresponding layers in the "decoder" after upsampling. In comparison, SegNet passes only the maxpooling indices from "encoder" to "decoder" for unpoolng, which greatly reduces the amount of memory.

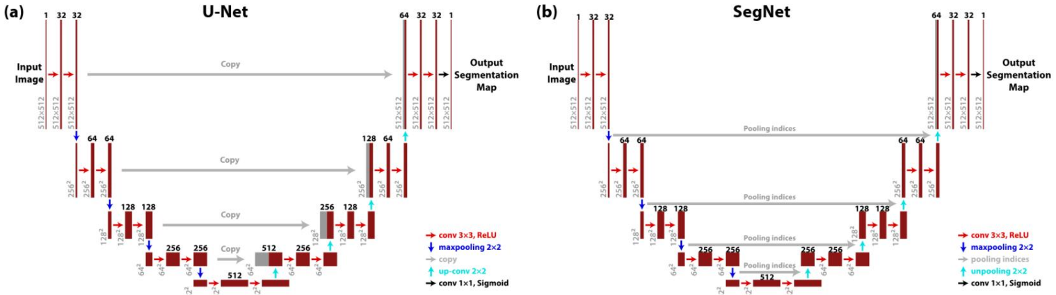

Figure 2: Comparison of (a) U-Net [5] model and (b) SegNet [6] model. The main difference between the two models rests on what information is transferred between the "encoder" and the "decoder".

4.2 Loss function

Our segmentation task is in essence a binary classification problem at pixel level. *Binary cross-entropy (BCE) loss* is thus a good starting point. However, our dataset has an imbalance of positives and negatives. As a result, we also use two other loss functions that are insensitive to such imbalance, i.e. *DICE loss* and *IOU loss*,

$$\mathcal{L}_{DICE} = 1 - \frac{2\sum_{i=0}^{n} y_i \hat{y}_i + \epsilon_0}{\sum_{i=0}^{n}(y_i + \hat{y}_i) + \epsilon_0},$$

$$\mathcal{L}_{IOU} = 1 - \frac{\sum_{i=0}^{n} y_i \hat{y}_i + \epsilon_0}{\sum_{i=0}^{n}(y_i + \hat{y}_i - y_i \hat{y}_i) + \epsilon_0},$$

where ϵ_0 is a small value to prevent dividing by 0.

DICE loss and IOU loss are based on two common metrics: DICE score and IOU score. Training on minimizing the inverse of the metric is equivalent to increasing the metric. As a result, either DICE loss and IOU loss are insensitive to imbalance between positives and negatives.

5 Experiments/Results/Discussion

The novelty of the project involves applying state-of-the-art segmentation models in the new application of fetal brain segmentation. Our experiments focus on training the best U-Net and SegNet models for fetal brain segmentation.

5.1 Training

The dataset of approximately 10,000 images is split into training/dev/test sets with a ratio of 80%/10%/10%. We evaluated the U-Net and SegNet models with different sizes (number of filters), regularization and early stopping. For the early stopping procedure, we wrote a script that stops the training if the validation loss does not decrease after 5 consecutive epochs. During our hyperparameter search, we experimented with different values of learning rate, dropout rate and batch normalization. We kept the Adam optimizer parameters constant as the default except for the learning rate, which was tuned for optimal performance on the training set in 100 epochs. In the end, the learning rate ranges between $\alpha = 10^{-5}$ and $\alpha = 10^{-4}$ depending on the model size (bigger model has a smaller α). For each convolutional layer, ReLU is the activation function and He initialization is used.

Due to the privacy requirement of medical data, the model training was done on Prof. Yeom's private server in the Stanford School of Medicine. The GPU memory available for us was 12GB, and we always chose the mini-batch size as the maximum that fits to the GPU memory. For a large U-Net, the largest mini-batch size contained 4 images, while for SegNet it contained 8 images since this model is smaller.

5.2 Results

The metric we use to evaluate model performance is the IoU score, defined as

$$IoU = \frac{TP}{TP + FP + FN},$$

which is a commonly used metric for segmentation (TP: true positives; FP: false positives; FN: false negatives). Another popular metric is the DICE score. Since the two metrics are basically interchangeable, we only stick to IoU score throughout the project.

Table 1 lists some typical results for good combinations of model size, loss function and hyper-parameters. We found that both U-Net and SegNet perform similarly in training and dev sets. The IoU scores for both models on the training set are very high (>0.9) and the scores on the dev set are between 0.7 and 0.8. This shows a typical variance problem since the dev IoU is lower than the training IoU by approximately 0.15, despite regularization and early stopping.

Model	Loss	Dropout	IoU (Training Set)	IoU (Dev Set)
U-Net (Small)	BCE	0.5	0.91	0.77
U-Net (Medium)	BCE	0.5	0.94	0.78
U-Net (Medium)	IoU	0.5	0.90	0.71
U-Net (Large)	BCE	0.5	0.91	0.74
SegNet	BCE	N/A	0.96	0.73

Table 1: Typical results of good combinations of model size, loss function and hyperparameter tuning. Results are for 100 epochs, with the learning rate tuned for optimal performance on the training set. Early stopping is omitted here as it does not have a significant impact.

5.3 Discussion

To analyze the variance problem, we generated the overlay images using output masks of the model. Figure 3 shows two typical examples when the model makes good or bad predictions.

One possible reason for the large variance is limited training data. Generally speaking, 10,000 is not a small number for a segmentation task. However, as can be seen in Figure 3, the structure of the brain varies with the slicing depth. We might need much more training data for the model to learn the brain structure at different depths and in different anatomy views (axial, sagittal and coronal).

Another possible reason is the inaccurate "ground truth". Illustrated in Figure 1 and 3, the overlay images with "ground truth" show that the boundary area has non-brain tissues (the mask of the whole dataset was generated by only one radiologist in a short amount of time by drawing a boundary line manually). During the model training process, if the mask provided as the ground truth is not accurate enough, the model will over-fit to this inaccurate mask pattern, preventing model generalization. As a result, variance problem is unavoidable.

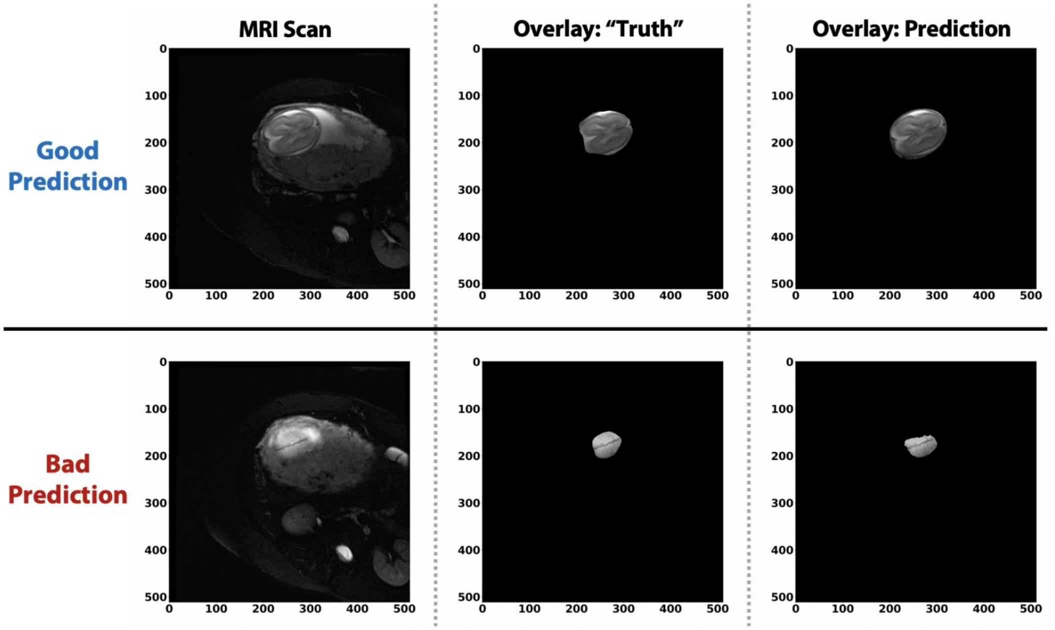

Figure 3: Two typical examples of model output. The top row shows an example when the model makes a good prediction. The bottom row shows an example when the model makes a bad prediction.

6 Conclusion/Future Work

In this project, we applied state-of-the-art segmentation models to the novel task of fetal brain segmentation. Our current models have very good performance on the training set and relatively good performance on the dev set. We believe our project can serve as a first step for the development of a classifier to detect abnormalities in fetal brain growth.

For future work, we would like to address the persistent variance problem with the following steps:

- Apply data augmentation (rotation, flipping, change of contrast, etc.);
- Obtain better masks from the collaborator at the Stanford University School of Medicine;
- Apply other regularization techniques, including L2;
- Experiment with other semantic segmentation algorithms such as DeepLab.

Contributions

The code to the project is at `https://github.com/cchen91/cs230_submission`. Cheng adapted existing U-Net and SegNet models to the problem, as well as wrote the data pre-processing code. Pablo performed literature review on different methods of image segmentation and worked on hyperparameter tuning.

Acknowledgements

We would like to thank Aarti Bagul for her mentoring and feedback during project meetings. We also extend our gratitude to Prof. Kristen Yeom from the Stanford School of Medicine for providing us the dataset and computational resources.

References

[1] "European anomaly registers." `http://www.eurocat-network.eu/prenatalscreeeninganddiagnosis`. Accessed: Oct. 13, 2018.

[2] "Congenital anomaly statistics 2015." https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/716574/Congenital_anomaly_statistics_2015_v2.pdf. Accessed: Oct. 13, 2018.

[3] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

[4] 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*, pp. 3431–3440, 2015.

[5] 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*, pp. 234–241, Springer, 2015.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 2481–2495, 2017.

[7] https://github.com/zhixuhao/unet. Accessed: Oct. 20, 2018.

[8] https://github.com/ykamikawa/SegNet. Accessed: Nov. 21, 2018.