MRI segmentation of the developing neonatal brain: Pipeline and training strategies for label scarcity

Leonie Richter
Department of Computing
Imperial College London
richter.leo94@gmail.com

Ahmed E. Fetit
UKRI CDT in AI for Healthcare & Dept. of Computing
Imperial College London
afetit@imperial.ac.uk

Abstract

We here summarise and discuss our published work on semantic segmentation of 3D neonatal brain MRI with deep networks. In addition to developing an accurate, end-to-end segmentation pipeline specifically designed for neonatal brain MRI, we investigated two approaches that can help alleviate the problem of label scarcity often faced in neonatal imaging. First, we examined different strategies of distributing a limited budget of annotated 2D slices over 3D whole-brain images. In the second approach, we compared the segmentation performance of pre-trained models with different strategies of fine-tuning on a small subset of preterm infants. We illustrated our findings using publicly available MRI scans obtained retrospectively from the *Developing Human Connectome Project* (ages at scan: 26-45 weeks).

1 Introduction

Initiatives such as the *Developing Human Connectome Project (dHCP)* [1] and the *Baby Connectome Project* [2] aim to develop a blueprint of the developing human brain with Magnetic Resonance Imaging (MRI) techniques. Mapping out neonatal brain development could accelerate the detection of cerebral palsy, autism, hypoxic ischemic encephalopathy and congenital deformations, which are thought to originate during the perinatal period of human development [1], [3]. Accurate structural processing of MRI data, particularly semantic segmentation, is an important step towards delivering a precise connectome of the developing brain. Yet, despite advances in deep learning, obtaining sufficient ground truth labels for data-driven methods can be expensive, cumbersome, and time-consuming. With neonatal brain MRI, there are also unique complications associated with this type of data (see [4, 5, 6]); one example is that the scans are oftentimes heterogeneous in morphology and texture, which is caused by the rapid brain development taking place over narrow time-scales [4].

In our recently published paper [7], we made three novel contributions to the specific area of neonatal brain MRI segmentation: *i)* We developed a fully automated, deep learning pipeline for segmenting highly complex 3D neonatal brain MRI that achieves high segmentation performance on subjects of a wide age range; *ii)* We studied different strategies for dealing with insufficient labels in neonatal brain MRI, by framing the problem as having a limited budget of labels that would be distributed over a training set, e.g., *'what is the best way to distribute a constrained budget of 2D labels over 3D data?' iii)* Finally, we examined the extent to which transfer learning can alleviate the problem of label scarcity in the context of neonatal imaging.

2 Materials and Methods

Dataset: We used publicly available data provided by the *Developing Human Connectome Project (dHCP)* consortium. The most recent, publicy available *Third Data Release* includes 783 neonatal MRI scans. For the purpose of this research, T1- and T2-weighted 3D scans, as well as minimal

Medical Imaging Meets NeurIPS (Med-NeurIPS 2022), New Orleans, USA.

meta data were used. We utilised labels that were generated by the *dHCP* consortium using an automated software pipeline (not deep learning-based, see [1]), and are publicly available as part of the aforementioned data release. The labels represented the following classes: Cerebrospinal Fluid (CSF), Cortical Grey Matter (cGM), White Matter (WM), Ventricles, Cerebellum, Deep Grey Matter (dGM), Brainstem (BS), Hippocampus, as well as inner and outer background classes [8]. When developing the deep learning pipeline, we selected an age-representative subset consisting of 20% of the *complete* data, resulting in 142 samples. We further divided the 142 samples into 114 training and 28 validation samples, and used an independent test set of 100 separate samples.

Pipeline Design: U-Net [9] was a natural starting point in terms of architecture, given the excellent performance demonstrated on variety of tasks [10, 11, 12]. We cropped training and validation images to non-zero value regions using a bounding box in order to increase information density and save computation power. We cropped the images into patches and we introduced a range of spatial and intensity-based transformations using random spatial cropping, 3D elastic deformation, flipping, Gaussian noising and smoothing, as well as intensity scaling and shifting.

We introduced three key modifications to the U-Net architecture: *i)* Kernels and strides were modified for 3D data; *ii)* Residual connections were used within the convolution blocks; *iii)* Deep supervision was utilised as per [13]. We computed the loss not only from the final output of the network and the ground truth label, but by taking outputs from deeper layers with a lower resolution into account as well. The model compared the outputs from deeper layers with downsampled versions of the ground truth segmentations. The number of deeper U-Net layers used is a hyperparameter of the network architecture; however, we did not add undersampled output near the bottleneck layer [12]. Following [11], we implemented a combined loss of Dice Loss and Cross Entropy (CE) Loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} \tag{1}$$

The Dice Loss, which is derived from Dice Similarity Coefficient (DSC), was computed as per [11]:

$$\mathcal{L}_{\text{Dice}} = -\frac{2}{|K|} \sum_{k \in K} \frac{\sum_{i \in I} u_i^k v_i^k}{\sum_{i \in I} u_i^k + \sum_{i \in I} v_i^k} \tag{2}$$

For each voxel i of the image, a softmax output vector $u_i \in \mathbb{R}^K$ was assumed, with K = # classes. During training, we used stochastic gradient descent with an initial learning rate of $1\mathrm{e}-2$, a Nesterov momentum of 0.99 and weight decay, and a scheduler which decays the learning rate via the scheme $(1 - \text{epoch} \,/\, \text{epoch}_{\text{max}})^{0.9}$.

Incorrect segmentations from the low-resolution outputs of deeper layers were assigned lower weights than final segmentation maps [12], i.e.:

$$\mathcal{L}_{total} = \omega_{final} \cdot \mathcal{L}_{final} + \omega_{-1} \cdot \mathcal{L}_{-1} + \omega_{-2} \cdot \mathcal{L}_{-2} + \dots \tag{3}$$

where ω_{-1} is the weight given to the loss based on the penultimate layer of the U-Net, ω_{-2} is the weight given to the loss based on the layer before that, etc.

Moreover, we extended the pipeline aiming to perform age prediction and semantic segmentation simultaneously (AgeU-Net) as subject age has been shown to be an important source of variance in neonatal brain MRI and a reason for the particular difficulty of neonatal brain segmentation [14].

Label Budgeting Experiments: The lack of sufficient ground truth labels is frequently mentioned in the medical imaging literature, including in relation to neonatal data, e.g., [15, 16, 8, 14, 3], amongst others. Since manual labeling of scans by experts is directly constrained by time and financial costs, one can reframe the problem as having a certain limited budget of labels that ought to be efficiently allocated. This label budget may contain, for example, a maximum number of 2D slices that ought to be annotated on a certain modality and along a certain axis. Hence, in addition to developing the pipeline, we investigated empirically how this budget can be allocated as efficiently as possible, i.e., *'which allocation strategy would lead to the best performance?'*. We defined the baseline ideal condition as the labeling of all slices in the 3D brain scans.

Transfer Learning Experiments: Transfer learning can overcome the absence of sufficient labels. Inspired by [16, 17], we investigated six strategies to adapt a baseline model pre-trained on a source

dataset of older infant brains to the task of segmenting preterm infant brain MRI. These strategies differed in the degree to which they incorporated preterm training data; full details can be found in our published paper [7].

3 Results

The results obtained with the segmentation pipeline are summarised in Table 1. The table shows mean DSCs averaged over 4 runs for the 10 classes (including the two background classes), as well as mean DSCs that only take into account the 8 physiological classes. The highest DSC obtained on the test set was 0.913 averaged over all classes, and 0.920 averaged over the physiologically relevant tissue classes. Table 2 further breaks down these results per specific tissue classes.

	Validation Set		Test Set	
	Mean DSC	Mean DSC (no BG)	Mean DSC	Mean DSC (no BG)
Mean DSC	0.934 $\pm$4e$-$4	0.943 $\pm$4e$-$4	0.913 $\pm$3e$-$3	0.917 $\pm$3e$-$3

Table 1: Mean DSC values averaged over 4 runs, computed on both the validation and held-out test sets. Note that the two background classes were not taken into account in *Mean Dice (no BG)* [7].

	CSF	cGM	WM	Ventr.	Cereb.	dGM	BS	Hippoc.
Validation DSC	0.925	0.943	0.961	0.907	0.950	0.960	0.961	0.903
Test DSC	0.919	0.926	0.932	0.903	0.895	0.940	0.940	0.877

Table 2: DSC values for the 8 physiologically relevant tissue classes, averaged over 4 runs [7].

With regards to the label budgeting experiments, we investigated four strategies of distributing a limited number of labelled 2D slices over 3D training data; sagittal, coronal, axial, and random. In each condition, the proportion of selected 2D slices per 3D brain was 33%, i.e., two thirds of the slices along the respective axis were hidden during training (60 epochs). To enable a fair comparison, we made sure that the number of data units in the training set was tripled compared to the first set of experiments. The results of the 4 runs, as well as of the baseline model after 20 epochs, are summarised in Table 3.

	Validation Set		Test Set	
	Mean DSC	Mean DSC (no BG)	Mean DSC	Mean DSC (no BG)
Random	0.913	0.913	0.910	0.914
Sagittal	0.912	0.912	0.900	0.907
Coronal	0.911	0.910	0.913	0.917
Axial	0.913	0.912	0.905	0.908
Full (20 epochs)	0.908	0.909	0.852	0.849

Table 3: Mean DSC values obtained with the 4 budgeting conditions. The values obtained with the baseline model, which was trained on fully annotated brain scans, are summarised in the last row [7].

Lastly, when comparing different transfer learning strategies, we observed that including more parameters during training (i.e., deeper fine-tuning) ultimately yielded better performance. Fine-tuning only the bottleneck and adjacent layer was clearly sub-optimal, but the performance did not reflect linear dependence on the number of parameters included during training. To illustrate, the difference between the shallow and medium fine-tuning strategies was primarily the addition of two blocks, while with the deep fine-tuning strategy, four new blocks were included during training. In spite of this, the improvement observed between the medium and shallow conditions was *higher* than that between deep and medium conditions.

To summarise, there is a need within the neonatal neuroimaging community for accurate and automated computational tools for the segmentation of MRI scans, but the complexity of data impedes the use of deep learning and data-driven techniques. Despite this, we developed a highly accurate, end-to-end deep learning pipeline for segmenting 3D neonatal brain MRI of a wide age range. Additionally, we studied the efficacy of different label budgeting strategies for dealing with the ground truth bottleneck in the context of neonatal MRI. Finally, we investigated the extent to which different transfer learning strategies can alleviate the label scarcity bottleneck in neonatal imaging.

4 Broader Impact

The data used in this research is fully anonymised and publicly available from the *dHCP* consortium as part of the *Third Data Release*, there were therefore no risks of leaking patients' identifiable information. The problem of label scarcity that is often encountered in medical imaging is intensified with neonatal neuroimaging data, mainly due to low signal-to-noise ratio, low contrast-to-noise ratio, high occurrence of motion artifacts, as well as inverted signals; addressing the ground truth bottleneck is therefore essential for deep learning to find its way in neonatal neuroimage segmentation tasks. Developing a robust pipeline for neonatal neuroimaging data based on deep networks would usually require large amounts of labels to be annotated manually by experienced annotators. We avoided this by making use of structural segmentation maps that were generated by the *dHCP* initiative using an automated structural pipeline (not deep learning-based), and are publicly available as part of the aforementioned data release. We also used these labels in the budgeting and transfer learning experiments. Note that the *dHCP* structural data and segmentations had undergone a quality assurance process detailed in their release notes, which reported small regions of common inaccuracies; however, we did not manually refine the labels used in this research. The work discussed here presents a proof-of-concept and is based on retrospective data; translating this research for clincial adoption would require ethical and regulatory approvals, as well as large-scale prospective trials. To aid reproducibility, all the software we developed was made publicly available on Github [1].

5 Acknowledgements

The work of Ahmed E. Fetit was supported by the UK Research and Innovation Centre for Doctoral Training in Artificial Intelligence for Healthcare in his role as Senior Teaching Fellow (Grant Number: EP/S023283/1). The authors would like to thank Dr Benjamin Hou for technical feedback on the work. Data were provided by the developing Human Connectome Project, KCL-Imperial-Oxford Consortium funded by the European Research Council under the European Union Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. (319456). We are grateful to the families who generously supported this trial.

References

[1] Antonios Makropoulos, Emma C Robinson, Andreas Schuh, Robert Wright, Sean Fitzgibbon, Jelena Bozek, Serena J Counsell, Johannes Steinweg, Katy Vecchiato, Jonathan Passerat-Palmbach, et al. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *Neuroimage*, 173:88–112, 2018.

[2] Brittany R Howell, Martin A Styner, Wei Gao, Pew-Thian Yap, Li Wang, Kristine Baluyot, Essa Yacoub, Geng Chen, Taylor Potts, Andrew Salzwedel, et al. The UNC/UMN baby connectome project (BCP): An overview of the study design and protocol development. *NeuroImage*, 185:891–905, 2019.

[3] Chelli N Devi, Anupama Chandrasekharan, VK Sundararaman, and Zachariah C Alex. Neonatal brain MRI segmentation: A review. *Computers in Biology and Medicine*, 64:163–178, 2015.

[4] Daniela Prayer. Fetal MRI. *Topics in Magnetic Resonance Imaging*, 22(3):89, 2011.

[5] Ricardo C Sampaio and Charles L Truwit. Myelination in the developing human brain. *Handbook of Developmental Cognitive Neuroscience*, pages 35–44, 2001.

[6] Helen M Branson. Normal myelination: a practical pictorial review. *Neuroimaging Clinics*, 23(2):183–195, 2013.

[7] Leonie Richter and Ahmed E Fetit. Accurate segmentation of neonatal brain MRI with deep learning. *Frontiers in Neuroinformatics*, 16, 2022.

[8] Antonios Makropoulos, Ioannis S Gousias, Christian Ledig, Paul Aljabar, Ahmed Serag, Joseph V Hajnal, A David Edwards, Serena J Counsell, and Daniel Rueckert. Automatic whole brain MRI segmentation of the developing neonatal brain. *IEEE Transactions on Medical Imaging*, 33(9):1818–1831, 2014.

[1] https://github.com/richterleo/neonatal_brain_segmentation/

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[10] Lei Cai, Jingyang Gao, and Di Zhao. A review of the application of deep learning in medical image classification and segmentation. *Annals of Translational Medicine*.

[11] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[12] Fabian Isensee, Paul F Jäger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019.

[13] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision support*, pages 3–11. Springer, 2018.

[14] Antonios Makropoulos, Serena J Counsell, and Daniel Rueckert. A review on automatic fetal and neonatal brain MRI segmentation. *NeuroImage*, 170:231–248, 2018.

[15] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

[16] Mina Amiri, Rupert Brooks, and Hassan Rivaz. Fine tuning u-net for ultrasound image segmentation: Which layers? In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 235–242. Springer, 2019.

[17] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.