HHS Public Access

Author manuscript

Domain Adapt Represent Transf Med Image Learn Less Labels Imperfect Data (2019).

Author manuscript; available in PMC 2020 February 21.

Published in final edited form as:

Domain Adapt Represent Transf Med Image Learn Less Labels Imperfect Data (2019). 2019 October ; 11795: 243–251. doi:10.1007/978-3-030-33391-1_28.

Multi-task Learning for Neonatal Brain Segmentation Using 3D Dense-Unet with Dense Attention Guided by Geodesic Distance

Toan Duc Bui¹, Li Wang¹, Jian Chen², Weili Lin¹, Gang Li¹, Dinggang Shen^{1,3}

¹Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

²School of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, China

³Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea

Abstract

The deep convolutional neural network has achieved outstanding performance on neonatal brain MRI tissue segmentation. However, it may fail to produce reasonable results on unseen datasets that have different imaging appearance distributions with the training data. The main reason is that deep learning models tend to have a good fitting to the training dataset, but do not lead to a good generalization on the unseen datasets. To address this problem, we propose a multi-task learning method, which simultaneously learns both tissue segmentation and geodesic distance regression to regularize a shared encoder network. Furthermore, a dense attention gate is explored to force the network to learn rich contextual information. By using three neonatal brain datasets with different imaging protocols from different scanners, our experimental results demonstrate superior performance of our proposed method over the existing deep learning-based methods on the unseen datasets.

Keywords

Neonatal brain segmentation; Multi-task learning; Attention; Geodesic distance

1 Introduction

Brain tissue segmentation is a fundamental step in the baby brain MRI analysis. However, due to the low tissue contrast in the early infantile phase, accurate and automatic segmentation of the neonatal brain is still a challenge. Many efforts [2,10,14] have been proposed to improve segmentation accuracy. Over the past few years, deep convolutional neural network (DCNN) is considered as a potentially promising approach for infant brain segmentation. The DCNN aims to learn a supervised model from relevant features in the training images. A benchmark for infant brain segmentation based on DCNN can be found

in [14], where the DCNN methods demonstrated great success, in comparison with non-deep learning based methods. However, these methods achieve a good segmentation accuracy only on the testing datasets that have the similar imaging appearance as the training dataset, but may fail to produce reasonable results, when the testing data have different imaging appearance (i.e., from different acquisition conditions). The main reason is that these deep learning models tend to have a good fitting to the training dataset, but they generally do not lead to good generalization on unseen data [5].

Multi-task learning [1] has been proposed to improve the generalization of the DCNN by forcing a single model to learn several related tasks at once. For instance, Myronenko [9] added a variational autoencoder branch for image reconstruction and jointly trained it with segmentation branch to regularize the shared encoder for brain tumor segmentation. Dangi et al. [4] proposed multitask learning for cardiac MR image segmentation by jointly learning a segmentation network and a Euclidean distance map regression network. Wang et al. [13] employed the Euclidean distance to refine the segmentation result for infant brains. However, the Euclidean distance uses ground-truth label information to compute the distance to the tissue boundary, thus it does not leverage the rich image contextual information. Contrary to the Euclidean distance, the geodesic distance [3] allows encoding both spatial regularization and contrast-sensitivity. Wang et al. [12] demonstrated the effectiveness of the geodesic distance over the Euclidean distance for the interactive segmentation. However, it is not designed for multi-task learning.

In this paper, we propose multi-task learning for neonatal brain segmentation. Instead of learning a single segmentation task, the proposed network simultaneously trains both the tissue segmentation task and a geodesic distance regression task together to regularize the shared encoder network. Meanwhile, we observe that the regression features can be used to further refine the segmentation result. Hence, we further concatenate the segmentation features and the regression features together followed by three $3 \times 3 \times 3$ convolutions to produce a refined segmentation result. Furthermore, we propose a dense attention gate (DAG) to guide the network to focus on learning the contextual information. The DAG utilizes contextual information from high-level features to provide unambiguous information for the correct category as the guidance of low-level features.

2 Method

2.1 Dense Attention Gate (DAG)

Let $\mathbf{U}^l = [u_1^l, u_2^l, \dots, u_{C_l}^l]$ be a feature map at the l^{th} layer, where $u_i^l \in \mathbb{R}^{D \times H \times W}$ and C_l is the number of channel. A statistic vector $\mathbf{z}^l = [z_1^l, z_2^l, \dots, z_{C_l}^l]$ is achieved by the global average pooling [8], and each element z_i^l of vector $\mathbf{z}^l$ is computed as

$z_i^l = \frac{1}{D \times H \times W} \sum_{d=1}^D \sum_{h=1}^H \sum_{w=1}^W u_i^l(d, h, w)$. Since high-level features often have a larger number of channels, we apply $1 \times 1 \times 1$ convolution to reduce the number of channels to C_l . It provides a balance of important features among different levels. An attention vector $\mathbf{a}^l$ at the layer l^{th} is formulated via a sigmoid activation [6]:

$$\mathbf{a}^l = \sigma(\mathbf{F}(\mathbf{z}^l, \mathbf{W}_0^l)) \quad (1)$$

where $\mathbf{W}_0^l = [w_1, w_2, \dots, w_{C_l}]$ is learnable weight for each element of the vector $\mathbf{z}^l$, and $\mathbf{F}(\cdot)$ is a transformation function.

To integrate the contextual information from high-level features, we define our proposed dense attention gate (DAG) as follows:

$$\mathbf{a}_{DAG}^l = \sum_{i=l}^n w_i \mathbf{a}^i \quad (2)$$

where the dense attention vector $\mathbf{a}_{DAG}^{n-1}$ at the layer $(n-1)^{th}$ computes as follows:

$\mathbf{a}_{DAG}^{n-1} = w_n \mathbf{a}^n + w_{n-1} \mathbf{a}^{n-1}$. In which, $\mathbf{a}^n = \sigma(\mathbf{F}(\mathbf{z}^n, \mathbf{W}_0^n))$ is the attention vector at the layer n^{th} . In the same way, the dense attention vector $\mathbf{a}_{DAG}^{n-2}$ at the layer $(n-2)^{th}$ calculates as follows: $\mathbf{a}_{DAG}^{n-2} = w_n \mathbf{a}^n + w_{n-1} \mathbf{a}^{n-1} + w_{n-2} \mathbf{a}^{n-2} = \mathbf{a}_{DAG}^{n-1} + w_{n-2} \mathbf{a}^{n-2}$. Thus, the Eq. 2 can be rewritten as follows:

$$\mathbf{a}_{DAG}^l = \mathbf{a}_{DAG}^{l-1} + w_l \mathbf{a}^l \quad (3)$$

The Eq. 3 indicates that the dense attention vector $\mathbf{a}_{DAG}^l$ at layer l^{th} can be computed via linear combination of the dense attention vector $\mathbf{a}_{DAG}^{l-1}$ at the higher level feature $l-1^{th}$ and the attention vector at the layer l^{th} . Hence, the dense attention vector $\mathbf{a}_{DAG}^l$ not only contains the information of the layer l^{th} , but also includes the contextual information of higher level features. We employ the weight $\mathbf{W}_1^l = [w_l, w_{l+1}, \dots, w_n]$ to learn the contribution of each high-level feature to the attention vector $\mathbf{a}_{DAG}^l$. Figure 1 shows our proposed dense attention gate, which aims to force the network to learn the contextual information of features. By using the contextual information from high-level features, our proposed DAG enables the network to emphasize useful information and ignore ambiguous information in the low-level features.

2.2 Geodesic Distance

Given an image $I: \Omega \rightarrow \mathbb{R}^{D \times H \times W}$ and a corresponding ground-truth label $Y: \Omega \rightarrow \mathbb{R}^{D \times H \times W}$. We use one-hot encoding scheme to convert the ground-truth label Y to M_i binary mask images where $M_i(d, h, w) = 1$ if $M_i(d, h, w)$ belongs to the class i^{th} ; otherwise it takes 0 value, $i = 1, \dots, C$, where C is the number of tissue classes. For the binary mask M_i , the geodesic distance [12] from a voxel x to the boundary of the mask M_i is defined:

$$GD(x, M_i, \nabla I) = \min_{x' \in M_i} d(x, x', \nabla I) \quad (4)$$

where $d(x, x', \nabla I) = \min_{\Gamma \in P(x, x')} \int_0^1 \sqrt{\Gamma'(s)^2 + \gamma^2 (\nabla I(s) \cdot u(s))^2} ds$, and $P(x, x')$ denotes a set of all possible paths between voxels x and x' , and $\Gamma(s)$ is a path in the set, parameterized by $s \in [0, 1]$. The unit vector $u(s) = \frac{\Gamma'(s)}{\|\Gamma'(s)\|}$ presents the direction of the path. The factor γ controls the contribution of the image gradient $\nabla I(s)$ and the spatial distances, $\Gamma'(s)$. Note that the geodesic distance reduces to the Euclidean distance when setting $\gamma = 0$.

We define a signed geodesic distance map for the tissue i^{th} , $D_i(x)$, by taking negative value $-GD(\cdot)$ if the voxel inside the tissue M_i , and positive value $GD(\cdot)$ if the voxel outside the tissue. Figure 2 indicates an example of the signed geodesic distance for three classes. For a given voxel, the red color shows the voxels far away from the tissue boundary, while the blue color indicates the voxels close to the tissue boundary. By considering the geodesic distance, the network enables learning the spatial relationship for each voxel in each class.

2.3 Network Architecture of Multi-task Learning

Figure 3 illustrates our proposed multi-task network architecture, which has two main branches: segmentation branch and regression branch. The segmentation branch is a 3D-Unet [2] with dense connection [7] to improve information flow between blocks. The segmentation branch includes two paths: an encoder and a decoder. The encoder path comprises four dense blocks, followed by max-pooling layers to enlarge the receptive field. To preserve the spatial resolution of feature maps, we used dilated convolution with the rate of 2 for two last blocks. The proposed DAG is added on the skip connection to force the network to learn the contextual information. In the decoder path, 3D convolutions with stride of 2 are used to recover the input resolution. We apply a $1 \times 1 \times 1$ convolution to map the segmentation features to one of the four tissue classes, i.e., GM, WM, CSF and BG (background).

The regression branch shares the same encoder with the segmentation branch and has a similar decoding structure. Instead of pixel-wise classification, the distance regression learns pixel-wise regression. We use the mean square error loss to measure the difference between the estimated geodesic distance and the ground-truth distance. A discrimination network is added to encourage the regression network to produce similar output as the ground-truth distance. The features from two last layers of segmentation and regression branch are concatenated together, and then fed to a refinement network (RN) that consists of three convolutional layers $3 \times 3 \times 3$ to fuse the two features and provide a refinement segmentation result. By using the shared encoder, the proposed network enables learning a joint representation between segmentation and regression task. Hence, the regression task helps regularize the shared features and improves the generalization of the network, thus yielding a better segmentation on the unseen datasets. The proposed network not only provides the pixel-wise prediction, but also estimates the spatial distance of each pixel to the tissue boundary.

The proposed multi-task loss is thus defined as follows:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{seg} + \alpha_2 \mathcal{L}_{reg} + \alpha_3 \mathcal{L}_D + \mathcal{L}_{seg_reg} \quad (5)$$

where $\mathcal{L}_{seg}$ and $\mathcal{L}_{seg_reg}$ are the cross-entropy losses, and $\mathcal{L}_{reg}$ and $\mathcal{L}_D$ denote the mean square error losses for distance regression and the discrimination network, respectively. The parameters $\alpha_i (i = 1, \dots, 3)$ control the balance among losses.

3 Experiments and Results

3.1 Datasets and Training

We compare the performance of our proposed network with several state-of-the-art baselines on three datasets of neonatal brains, as shown in Table 1. Herein, we mainly focus on T1-weighted neonatal images, which typically have higher spatial resolutions than T2-weighted images in most neonatal imaging studies. However, our method is generic and can also be applied in T2-weighted images. All T1-weighted images used in our experiments were acquired on 3T scanners. The performance of each method was compared with the manual segmentation using the following two metrics: Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD). A higher DSC value and a lower HD value indicates a superior performance.

All networks were implemented and trained using Pytorch framework on the 12GB Titan X GPU. We normalized the input image to zero mean and unit variance, and randomly cropped a sub-region with a size of $64 \times 64 \times 64$ before inputting them into the network. The Adam optimizer with a batch size of four were used to train the network. The initialization learning rate was set as 0.0002 and was decreased ten times every 4,000 epochs. The total number of iteration was 20,000 epochs. The balance weights were set as $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = 0.5$. We set $\gamma = 0.2$ for the geodesic distance map. We selected the dataset A as the training data and the two remaining datasets for the validation datasets.

3.2 Training and Testing Within the Same Dataset

We first evaluate the performance of those methods on the dataset A. We randomly select sixteen samples for training and four samples for validation and nine samples for testing. Table 2 shows the segmentation accuracy in term of DSC values for 3-fold cross validation and testing on the dataset A of different methods. We employ 3D-Unet [2] with dense connection [7] as a baseline model. The proposed architecture is built upon the baseline model and extended by including the regression branch. Meanwhile, the concurrent spatial and channel (ScSE) [11] is included to the baseline model for comparing with the proposed DAG block.

Table 2 reports the performance of the proposed method and the competing methods on validation and testing sets when training on the same dataset. Both proposed DAG and ScSE blocks provide better results in comparison with the baseline, demonstrating the advantage of the attention mechanism for neonatal brain segmentation. By adding the distance regression branch, the proposed network overcomes the existing methods in term of average

DSC. From Table 2, we can conclude that these methods provide good generalization when the training and testing data are from the same scanning center.

3.3 Training and Testing in Different Datasets

To demonstrate the effectiveness of the proposed multi-task learning, we further evaluate the proposed network on the unseen datasets: B and C. All testing images are resampled to the training resolution of $0.9375 \times 0.9375 \times 1 \text{ mm}^3$.

Figure 4 illustrates the segmentation results obtained by different methods in all three datasets. The baseline method, 3D-Unet [2], provides a good result when testing and training data have the similar distribution, but has misclassification on the unseen datasets indicated by white rectangle as shown in Fig. 4 (b). This indicates that the existing methods have a poor generalization on the unseen datasets. In contrast, the proposed method provides better stability and generalization capability among different datasets as shown in Fig. 4(d).

Table 3 compares the mean and standard deviation of segmentation accuracy of those methods on the unseen datasets in term of DSC, and HD metrics. It is observed that the proposed multi-task learning provides a better generalization than the existing networks on unseen datasets, i.e., achieving more accurate segmentation. The proposed architecture not only improved DSC from 89.02% to 90.29%, but also reduced HD distance from 1.32mm to 1.11mm, compared with the existing methods. Since the all networks are trained on low resolution images (dataset A), thus they led to a low performance on the testing images with a high resolution (dataset B).

4 Conclusion

We presented an effective of multi-task learning architecture that allows training segmentation and distance regression tasks simultaneously to address the generalization issue on unseen datasets in neonatal brain segmentation. We also proposed a dense attention gate to force learning the important features in the low level-features by using the information from high-level features. We show that the proposed multi-task learning network can provide a good generalization, yielding a better segmentation accuracy on unseen datasets.

Acknowledgment.

This work was supported in part by NIH Grants MH107815, MH109773, MH116225, and MH117943.

References

1. Caruana R: Multitask learning. *Mach. Learn.* 28(1), 41–75 (1997)
2. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). 10.1007/978-3-31946723-8_49

3. Criminisi A, Sharp T, Blake A.: GeoS: geodesic image segmentation In: Forsyth D, Torr P, Zisserman A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 99–112. Springer, Heidelberg (2008). 10.1007/978-3-540-88682_29
4. Dangi S, Yaniv Z, Linte C.: A distance map regularized CNN for cardiac cine MR image segmentation. arXiv preprint arXiv:1901.01238 (2019)
5. Davatzikos C.: Machine learning in neuroimaging: progress and challenges. *NeuroImage* 197, 652 (2018) [PubMed: 30296563]
6. Hu J, Shen L, Sun G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE CVPR*, pp. 7132–7141 (2018)
7. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ: Densely connected convolutional networks. In: *Proceedings of the IEEE CVPR*, pp. 4700–4708 (2017)
8. Lin M, Chen Q, Yan S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
9. Myronenko A.: 3D MRI brain tumor segmentation using autoencoder regularization In: Crimi A, Bakas S, Kuijff H, Keyvan F, Reyes M, van Walsum T. (eds.) *BrainLes 2018*. LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019). 10.1007/978-3-030-11726-9_28
10. Ronneberger O, Fischer P, Brox T.: U-Net: convolutional networks for biomedical image segmentation In: Navab N, Hornegger J, Wells WM, Frangi AF (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). 10.1007/978-3-319-24574-4_28
11. Roy AG, Navab N, Wachinger C.: Concurrent spatial and channel squeeze & excitation in fully convolutional networks. arXiv preprint arXiv:1803.02579 (2018)
12. Wang G, et al.: DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. *IEEE Trans. PAMI* 41(7), 1559–1572 (2018)
13. Wang L, et al.: Volume-based analysis of 6-month-old infant brain MRI for autism biomarker identification and early diagnosis In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G. (eds.) *MICCAI 2018*. LNCS, vol. 11072, pp. 411–419. Springer, Cham (2018). 10.1007/978-3030-00931-1_47
14. Wang L, et al.: Benchmark on automatic 6-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. *IEEE Trans. Med. Imaging* 38(9), 2219–2230 (2019)

Fig.1.
Our proposed dense attention gate (DAG).

Fig.2.

An example of computing geodesic distance: (a) original T1-weighted image (I), (b) label image (Y), geodesic distance maps of (c) cerebrospinal fluid (CSF) class M_1 , (d) gray matter (GM) class M_2 , and (e) white matter (WM) class M_3 .

Fig.3. Our proposed multi-task learning architecture for neonatal brain segmentation.

Fig.4.

An example of neonatal brain image: (a) axial slice of T1-weighted MRI from the dataset A (first row), the dataset B (second row) and the dataset C (third row), (b) results by 3D-Unet [2] (baseline), (c) results by baseline+SCSE [11], (d) results by our proposed method, and (e) the ground-truth results.

Table 1.

Datasets and imaging settings.

Datasets (# subject)	Scanners	Image resolutions	Imaging parameters (TR/TE)
Dataset A ($N=25$)	GE	$0.9375 \times 0.9375 \times 1\text{mm}^3$	10.47/4.76ms
Dataset B ($N=10$)	Siemens	$0.8 \times 0.8 \times 0.8\text{mm}^3$	2400/2.24ms
Dataset C ($N=10$)	Siemens	$1 \times 1 \times 1\text{mm}^3$	1900/4.38ms

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Performance comparison on 3-fold cross validation and testing using DSC.

	Method	CSF	GM	WM	Avg. DSC
Validation	3D-Unet [2] (Baseline)	97.59(0.39)	96.77(0.30)	98.43(0.28)	97.59
	Baseline+ScSE [11]	97.63(0.40)	96.83(0.29)	98.45(0.27)	97.64
	Baseline+DAG (Ours)	97.68(0.39)	96.97(0.28)	98.51(0.27)	97.72
	Baseline+DAG+DR (Ours)	97.80(0.41)	97.09(0.32)	98.55(0.28)	97.81
Testing	3D-Unet [2] (Baseline)	97.65(0.44)	96.85(0.33)	98.51(0.26)	97.67
	Baseline+ScSE [11]	97.71(0.44)	96.93(0.32)	98.54(0.26)	97.73
	Baseline+DAG (Ours)	97.75(0.49)	97.06(0.33)	98.59(0.26)	97.80
	Baseline+DAG+DR (Ours)	97.88(0.48)	97.18(0.35)	98.65(0.26)	97.90

ScSE: Spatial Channel Squeeze Excitation [11]; DAG: Dense Attention Gate, DR: Distance Regularization. Bold indicates a significant better performance with p -value < 0.05 using paired t-test

Table 3.

Segmentation accuracy on different datasets.

	Method	DSC (%)	HD (mm)
Dataset B	3D-Unet [2] (Baseline)	84.55 (0.93)	2.28 (0.20)
	Baseline+ScSE [11]	84.98 (0.98)	2.07 (0.17)
	Baseline+DAG+DR (Ours)	85.85 (1.19)	1.61 (0.12)
Dataset C	3D-Unet [2] (Baseline)	89.02 (3.15)	1.32 (0.11)
	Baseline+ScSE [11]	89.15 (3.01)	1.29 (0.09)
	Baseline+DAG+DR (Ours)	90.29 (2.43)	1.11 (0.06)