Domain Knowledge Driven Multi-modal Segmentation of Anatomical Brain Barriers to Cancer Spread

Xiaoyang Zou$^1$ and Qi Dou$^2$

$^1$ School of Engineering Sciences, Huazhong University of Science and Technology, Wuhan, China
$^2$ Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China

Abstract. It is important to accurately segment anatomical brain barriers to cancer spread with multi-modal images, in order to assist definition of the clinical target volume (CTV). In this work, we explore a multi-modal segmentation method largely driven by domain knowledge. We apply 3D U-Net as the backbone model. In order to reduce the learning difficulty of deep convolutional neural networks, we employ a label merging strategy for the symmetrical structures which have both left and right labels, to highlight the structural information regardless of the locations. Moreover, considering the existence of visual preference for certain modality and mismatches in co-registration, we adopt a multi-modality ensemble strategy for multi-modal learning to enable the models better driven by domain knowledge of this task, which is different from fully data-driven methods, like early fusion strategy for multi-modal images. By contrast, multi-modality ensemble strategy yields better segmentation results. Our method achieved an average score of 0.895 on MICCAI 2020 Anatomical Brain Barriers to Cancer Spread Challenge’s final test dataset (https://abcs.mgh.harvard.edu/). Detailed methodologies and results are described in this technical report (This work was done when X. Zou did remote internship with CUHK.).

1 Introduction

Glioma is the most common type of malignant brain tumors with an incidence of 6.9 per 100,000 population [8], comprising about 25.5% of all brain and other central nervous system tumors and 80.8% of all malignant brain tumors [10]. An effective treatment for gliomas is radiotherapy [7]. In the process of tumor’s radiotherapy planning, accurate definition of the clinical target volume (CTV) is a highly critical step. Inadequate definition of CTV may lead to geometric miss of the tumors, resulting in underdosage of certain areas in radiotherapy, which may lead to higher risk of recurrence for the patients [15]. Actually, brain anatomical structures can serve as a natural barrier to the spreading of brain tumors. Thus,
their boundaries can effectively assist the definition of CTV [14]. In addition, in order to prevent the sensitive and important healthy organs from being affected by radiation during radiotherapy, some healthy organs also need to be accurately delineated (like eyes, cochleae, optic nerves, chiasm, etc.) [9]. However, manual delineation of brain anatomical structures and healthy organs is time-consuming and error-prone. Thus, developing a fully automatic segmentation algorithm will effectively improve the efficiency and consistency of radiotherapy planning [3], which is exactly what the Anatomical Brain Barriers to Cancer Spread (ABCs) Challenge is aimed at.

To conduct this task, multi-modal images are often required in clinical practice. Specifically, Computed Tomography (CT) and multi-sequence Magnetic Resonance Imaging (MRI) can visualize different features of brain anatomical structures [13]. There exist both shared information and complementary information between different modalities. Looking for a multi-modal learning strategy to exploit valuable modality information may improve the segmentation performance effectively.

In recent years, deep convolutional neural networks (DCNNs) have become the de facto method for automated medical image segmentation. Ronneberger et al. proposed U-Net, a landmark network architecture for medical image segmentation [12]. Çiçek et al. proposed 3D U-Net for 3D medical image segmentation on this basis [1]. Isensee et al. completed the nnU-Net training framework, making U-Net achieve a relatively high level [4–6]. On this basis, two strategies are often used for the fusion of co-registered multi-modal images, which are early fusion and late fusion [2]. The early fusion is to concatenate multi-modal images as different channels of the network’s input, which has already been widely used for brain tissue’s segmentation in multi-sequence MR images [17]. And the late fusion is to fuse multi-modal images at a semantic level in the middle of the network. However, most existing methods are merely data-driven, without carefully consideration to incorporate valuable domain knowledge into learning process.

In this work, instead of creating some novel DCNN architectures, we primarily focus on exploring how to facilitate the model learning with domain knowledge of this task. We employ the label merging strategy for symmetrical structures, which have both left and right labels, to reduce the learning difficulty of DCNNs. Thus, structural information can be highlighted for the network to learn. Besides, we adopt multi-modality ensemble strategy to help the network learn the domain knowledge embedded in multi-modal images, thus to realize a domain knowledge driven model instead of a fully data-driven model. We find that, with the existence of visual preference for certain modality and mismatches in co-registration, the segmentation performance of using multi-modality ensemble strategy goes beyond that of using early fusion strategy for multi-modal images.
2 Methods

In this section, we present our proposed methods including label merging strategy for symmetric structures and multi-modality ensemble strategy. In addition, we show details about the dataset, network architecture and training protocol.

2.1 Data Description

In this work, we use the 45 cases training data which are officially provided by ABCs challenge for training, and do not use any external data. In addition to this, 15 cases are used for validation leaderboard and another 15 cases are used for final test scoring. Each case contains images of three modalities, including CT, T1-weighted MR and T2-weighted FLAIR MR images. And the multi-modal images are aligned in same size and resolution by co-registration and resampling. What is noteworthy is that, in the final submission, we only use CT and T1-weighted MR images for training, without any T2-weighted FLAIR MR images.

2.2 Network Architecture

All models are trained with plain 3D U-Net implemented on nnU-Net training framework\(^1\), as shown in Fig. 1. The network consists of two paths, which are down-sampling path and up-sampling path. In each path, there are 4 stages of down-sampling or up-sampling. And each stage consists of two blocks. In down-sampling path, the first block includes a $3 \times 3 \times 3$ convolutional layer strided $1 \times 2 \times 2$ or $2 \times 2 \times 2$ for images down-sampling, an instance normalization layer and then a leaky ReLU layer. The second block includes a $3 \times 3 \times 3$ convolutional layer strided $1 \times 1 \times 1$ with padding, an instance normalization layer and then a leaky ReLU layer. In up-sampling path, we just replace the convolutional layer in first block with transposed convolutional layer. The final block consists of a $1 \times 1 \times 1$ convolutional layer and a softmax layer. Skip connections are used to forward the information from down-sampling path to up-sampling path.

2.3 Label Merging for Symmetric Structures

The brain anatomical structures often have typical sagittal plane’s symmetry. And some of them can be directly divided into left and right two parts, such as eyes, optic nerves, cochleae, lacrimal glands, etc. In Task 2 of ABCs Challenge, the left and right parts of these structures are distinguished with two labels, which are scored independently. Therefore, it’s of great significance to identify the structures’ left and right parts correctly.

In a sense, the difference in location between left and right can be regarded as a kind of location information, which is relatively independent of structural information. In fact, this kind of location information does not change case by

---

\(^1\) https://github.com/MIC-DKFZ/nnUNet.
case, especially when the images are scanned in the same direction. For DCNNs, such kind of redundant location information means a great learning difficulty, and may drown out the structural information. Actually, it is not necessary for DCNNs to learn such fixed location information. Instead, judgments can be made directly from left to right based on the predictions during post-processing.

In order to help DCNNs get rid of the difficulty of learning location information, we firstly merge the left and right labels of structures in Task 2 (see Fig. 2). In this way, the left and right structures have the same label. The DCNNs no longer need to learn the difference between left and right structures, but only focus on structural information.

Our model is trained with the new merged ground truth. However, predictions based on this model do not contain any location information. Thus, during post-processing, the left and right parts need to be separated again according to their actual locations in the predictions. In this challenge, images are pre-processed to same direction and have good symmetry. Based on this, we find that the y-z plane can be regarded as the sagittal plane, so as to determine whether the structure is located on the left or right side. For example, we can simply consider the minimum to half of the x-axis as the right part and the maximum to half of the x-axis as the left part. It has been verified that such a judgment is completely right for this data set.

In this way, the evaluation metrics will not decline due to the wrong judgment of left and right of DCNNs, and the segmentation task can be focused on segmenting the structures themselves.
2.4 Multi-modality Ensemble

By observing the characteristics of the data set, we find that visual preference for certain modal images always exists when segmenting different structures or dealing with different tasks. For example, Fig. 3 shows the brainstem in different modal images. We can see that the brainstem in T1-weighted MR images has the highest contrast and the best visual preference, compared with that in CT images or T2-weighted FLAIR MR images. In other words, for the brainstem, visual preference is for T1-weighted MR images.

What’s more, there exists obvious mismatches in the co-registration of multi-modal images. Considering that each modality shares the ground truth, when we compare each single modal image with the ground truth, it is not difficult to discover that some structures are mismatched with the ground truth. As Fig. 4 shows, we compare the same slice from all modal images and the ground truth in both Task 1 and Task 2. The mismatches between the location of actual structures and the ground truth are highlighted, which mainly exist at ventricles, sinuses, eyes, etc. And we can see that T1-weighted MR images almost have no mismatch in Task 1, while CT images almost have no mismatch in Task 2.

Although early fusion is a commonly used multi-modal fusion strategy, DCNNs are hard to distinguish the domain difference mentioned above with data-driven method only. It is necessary to try a domain knowledge driven method of this task. Thus, we further explore the segmentation performance when training with each single modality, hoping to train a better domain knowledge driven model with multi-modality ensemble strategy. Finally, we use T1-weighted MR images for training to segment all the structures in Task 1 and the brainstem in Task 2, while use CT images for training to segment all the structures in Task 2 except the brainstem.
2.5 Training Protocol

Pre-processing methods include non-zero region cropping and normalization. For CT images, we clip the CT value from 0.5% to 99.5% for training to reduce less informative pixels. All CT cases are normalized together by z score, considering that CT value is quantitative. For MR images, each case is individually normalized by z score due to the distribution difference. Patches are extracted randomly with a size of [112, 160, 128]. And the batch size is 2. Images are augmented through random rotation, scaling, mirroring, gamma, brightness and contrast transformation. Stochastic gradient descent optimizer is used for training, with an initial learning rate of 1e-2 and a weighted decay of 3e-5. The total loss is the sum of dice loss and cross-entropy loss. All models are trained using only one NVIDIA TITAN Xp GPU and PyTorch library [11].

In the final submission, we adopt the model ensemble strategy to improve the model’s robustness. Models are test on the validation leaderboard every 100 epochs trained. We finally select the best two models to employ model ensemble strategy. The final prediction is obtained by averaging the softmax output of the two models and taking 0.5 as the threshold for binarization.

3 Results

Both of dice similarity coefficient (DSC) and surface dice similarity coefficient (SDSC) at the tolerance of 2 mm are used to evaluate the accuracy of our automatic delineation. Specific calculations are based on an open source library from
Fig. 4. An example of mismatches in both Task 1 and Task 2. In each task, same slice is taken from all modal images and the ground truth. The contrast is auto-adjusted by ITK-SNAP [16].
DeepMind\textsuperscript{2}, which is recommended by the challenge organizers. And the average score over all test data of both DSC and SDSC is taken as the final evaluation. All experiments were completed during the ABCs 2020 Challenge, and all models were evaluated by 5-fold cross-validation on the training set or with data for the validation leaderboard.

Table 1 shows the effectiveness of label merging strategy for symmetric structures. For Task2, we compare the average DSC with and without label merging strategy and calculate their difference, under the condition of training with the same plain 3D U-Net and T1-weighted MR images. It can be seen that label merging strategy realizes a huge improvement of DSC in Task 2, about 33.3\%, especially to the structures that can be divided into left and right two parts.

\textbf{Table 1.} Models are only trained on T1-weighted MR images. The DSC scores trained based on the original ground truth are shown in column T1, and the DSC scores trained based on the ground truth after label merging are shown in column T1+Label merging. The column Difference calculates the difference value between those two DSC scores. The results are evaluated based on 5-fold cross-validation.

| Task 2 structure | T1   | T1 + label merging | Difference |
|------------------|------|--------------------|------------|
| Brainstem        | 0.885| 0.899              | +0.014     |
| Chiasm           | 0.450| 0.510              | +0.060     |
| Left Cochlea     | 0.340| 0.715              | +0.375     |
| Right Cochlea    | 0.464| 0.694              | +0.230     |
| Left Eye         | 0.671| 0.896              | +0.225     |
| Right Eye        | 0.692| 0.910              | +0.218     |
| Left Lacrimal    | 0.504| 0.672              | +0.168     |
| Right Lacrimal   | 0.527| 0.676              | +0.149     |
| Left OpticNerve  | 0.474| 0.673              | +0.199     |
| Right OpticNerve | 0.493| 0.683              | +0.190     |
| \textbf{Average} | \textbf{0.550} | \textbf{0.733} | \textbf{+0.183} |

On the basis of adopting label merging strategy, we compare the results of the early fusion strategy of multi-modal images with the multi-modality ensemble strategy, as shown in Table 2. The scores listed here are based on the validation set for leaderboard. It can be seen that the results based on our multi-modality ensemble strategy completely exceed the results based on early fusion strategy. Besides, segmenting brainstem with T1-weighted MR images can help achieve better results indeed.

Through model ensemble, we further improve our performance on the validation leaderboard. Table 3 shows the final results of top 5 teams on the ABCs 2020 Challenge’s validation leaderboard. Our method surpassed all the other 20

\textsuperscript{2} https://github.com/deepmind/surface-distance.
Table 2. Comparison of early fusion strategy and multi-modality ensemble strategy. The early fusion adopted here is to concatenate all the three modal images as the networks input. The contents in parentheses indicate the scope of different modal images.

| Method | Task 1 DSC | Task 1 SDSC | Task 2 DSC | Task 2 SDSC |
|--------|------------|-------------|------------|-------------|
| Task 1: early fusion Task 2: early fusion | 0.867 | 0.973 | 0.755 | 0.921 |
| Task 1: T1 (all) Task 2: CT (all) | 0.877 | 0.975 | 0.777 | 0.924 |
| Task 1: T1 (all) Task 2: T1 (brainstem), CT (others) | **0.877** | **0.975** | **0.779** | **0.929** |

competing teams and ranked the 1st place on the validation leaderboard by a narrow margin. And we still won the 3rd place for the final test scoring with an average score of 0.895, which is 0.002 behind the 1st place.

Table 3. Top 5 results on ABCs 2020 Challenge’s validation leaderboard. Submissions are ranked by the overall average score of DSC and SDSC of each structure in Task1 and Task2. Note that the data used for validation leaderboard and the data used for final scoring are not same batch data. So this result does not represent the final scoring of the challenge.

| Team name      | Task 1 DSC | Task 1 SDSC | Task 2 DSC | Task 2 SDSC | Average score |
|----------------|------------|-------------|------------|-------------|--------------|
| MedAIR (ours)  | **0.877**  | **0.975**   | 0.781      | 0.930       | **0.891**    |
| Holo           | 0.876      | 0.974       | 0.782      | 0.929       | 0.890        |
| Tree           | 0.875      | 0.974       | 0.782      | 0.929       | 0.890        |
| AIViewSjtu     | 0.869      | 0.974       | 0.772      | 0.939       | 0.888        |
| HILab          | 0.869      | 0.972       | 0.771      | 0.933       | 0.886        |
| Sen            | 0.863      | 0.972       | 0.758      | 0.930       | 0.881        |

4 Discussion

It should be noted that the method we use to separate the merged labels in predictions has some limitations. We just simply make the judgment between left and right according to the x-coordinate, considering that all images provided by ABCs 2020 Challenge have been pre-processed to same direction and have good symmetry. However, for images with different direction and poor symmetry, a more essential approach to solve this problem is required. For example, we may need to convert the images to the same direction, or even find and combine the middle surface place of the symmetric structures for judgment.

For the segmentation of multi-modal images, we believe that different modal images contribute differently to the segmentation of different structures or different tasks. Although we have discovered that multi-modality ensemble strategy
shows better segmentation performance compared with early fusion of multi-modal images in ABCs challenge, the selection of best modality needs further study. Our final selection suggests two possibilities. The best modality might be the one has the best visual contrast or the one that is used for the data set’s labeling. In addition, considering the mismatches of co-registration, we need to think about whether it is worthwhile to adopt a multi-modality fusion strategy if obvious co-registration noise exists.

What’s more, with the help of the powerful nnU-Net training framework, it seems using some novel DCNN architectures or training methods cannot make significant improvements. We have tried to use 3D U-Net with residual blocks, and to modify the loss function as the sum of dice loss and focal loss or the sum of dice loss and weighted cross entropy loss. However, we did not obtain any improvement on the validation leaderboard. Compared with this, it seems more important to make the model become more robust. Thus, we adopt the model ensemble strategy here, which is to average the softmax outputs of the top two best performing models that we submitted on the leaderboard and then predict with a 0.5 threshold value. This strategy helps to improve the robustness of the model, thus enabling the model to achieve better performance.

5 Conclusion

In this work, we use plain 3D U-Net to segment anatomical brain barriers to cancer spread. For Task 2, we adopt a label merging strategy for symmetrical structures to reduce the network’s difficulty of learning location information. That is, the DCNNs only need to focus on structural information at the semantic level, then the left and right parts will be automatically separated according to their real location in the predictions. This strategy can significantly improve the DCNNs’ segmentation performance on those structures that have both left and right labels. In addition, we adopt multi-modality ensemble strategy to facilitate the model better driven by domain knowledge. The domain knowledge is important for the networks to learn especially when visual preference for certain modality and mismatches in co-registration exist. And the segmentation performance surpasses that of employing only data-driven method like multi-modal early fusion strategy.

References

1. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
2. Dou, Q., Liu, Q., Heng, P., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. IEEE Trans. Med. Imaging 39, 2415–2425 (2020)
3. Duanmu, H., et al.: Automatic brain organ segmentation with 3D fully convolutional neural network for radiation therapy treatment planning. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 758–762. IEEE (2020)

4. Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 (2019)

5. Isensee, F., et al.: nnU-Net: self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)

6. Isensee, F., Petersen, J., Kohl, S.A., Jäger, P.F., Maier-Hein, K.H.: nnU-Net: breaking the spell on successful medical image segmentation, vol. 1, pp. 1–8. arXiv preprint arXiv:1904.08128 (2019)

7. Laperriere, N., Zuraw, L., Cairncross, G., Cancer Care Ontario Practice Guidelines Initiative Neuro-Oncology Disease Site Group, et al.: Radiotherapy for newly diagnosed malignant glioma in adults: a systematic review. Radiother. Oncol. 64(3), 259–273 (2002)

8. Li, K., et al.: Trends and patterns of incidence of diffuse glioma in adults in the united states, 1973–2014. Cancer Med. 7(10), 5281–5290 (2018)

9. Mlynarski, P., Delingette, H., Alghamdi, H., Bondiau, P.Y., Ayache, N.: Anatomically consistent segmentation of organs at risk in MRI with convolutional neural networks. arXiv preprint arXiv:1907.02003 (2019)

10. Ostrom, Q.T., et al.: CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2012–2016. Neuro-Oncol. 21(Suppl. 5), v1–v100 (2019)

11. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8026–8037 (2019)

12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

13. Rumboldt, Z., Castillo, M., Huang, B., Rossi, A.: Brain Imaging with MRI and CT: An Image Pattern Approach. Cambridge University Press, Cambridge (2012)

14. Shusharina, N., Söderberg, J., Edmunds, D., Löfman, F., Shihi, H., Bortfeld, T.: Automated delineation of the clinical target volume using anatomically constrained 3D expansion of the gross tumor volume. Radiother. Oncol. 146, 37–43 (2020)

15. Weiss, E., Hess, C.F.: The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy. Strahlentherapie und Onkologie 179(1), 21–30 (2003)

16. Yushkevich, P.A., Gao, Y., Gerig, G.: ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3342–3345. IEEE (2016)

17. Zhang, W., et al.: Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage 108, 214–224 (2015)