Adaptive transfer learning to enhance Domain transfer in Brain Tumor Segmentation

Yuan, Liqiang[a], Marius, Erdt[a,b], and Wang, Lipo[a]

[a]Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
[b]Fraunhofer Singapore, 50 Nanyang Avenue, Singapore 639798

Abstract

Supervised deep learning has greatly catalyzed the development of medical image processing. However, reliable predictions require a large amount of labeled data, which is hard to attain due to the required expensive manual efforts. Transfer learning serves as a potential solution for mitigating the issue of data insufficiency. But up till now, most transfer learning strategies for medical image segmentation either fine-tune only the last few layers of a network or focus on the decoder or encoder parts as a whole. Thus, improving transfer learning strategies is of critical importance for developing supervised deep learning, further benefits medical image processing. In this work, we propose a novel strategy that adaptively fine-tunes the network based on *policy value*. Specifically, the encoder layers are fine-tuned to extract latent feature followed by a fully connected layer that generates policy value. The decoder is then adaptively fine-tuned according to these *policy value*. The proposed approach has been applied to segment human brain tumors in MRI. The evaluation has been performed using 769 volumes from public databases. Domain transfer from T2 to T1, T1ce, and Flair shows state-of-the-art segmentation accuracy.

Key words- Deep Learning, Transfer Learning, Human Brain Tumor Segmentation

1 Introduction

The development in the field of deep learning greatly benefits from the improvements in the network structure[1] and the availability of massive data sets for training[2]. However, in medical imaging, labeled data is often not available due to expensive manual cost. This challenge motivates researchers to explore various methods that can cope with the scarcity of data in the medical domain. TL (Transfer Learning) has demonstrated its potential to improve the training efficiency[3] by transferring knowledge from one machine learning classifier (source domain) to another machine learning classifier (target domain).

Previous works by He et.al[4], Bar et.al[5] and Chang et.al.[6] have demonstrated that TL knowledge from non-medical domain to 2D medical imaging is applicable. However, this TL strategy is less suitable for three-dimensional medical image problems. Although this issue is mitigated by expanding three-dimensional images to many two-dimensional slices, missing contextual information about neighboring slices in the 3D image stack may degrade the performance of the neural network, especially for tasks like brain tumor segmentation.

In this work, we focus on the human brain tumor segmentation problem in different Magnetic Resonance Imaging (MRI) modalities. A MRI T2 modality is chosen to be the source domain data and MRI T1, Flair and, T1ce modalities are selected to be our target domain data. An Auto-Encoder like architecture is preferred for segmentation task e.g. U-Net[7] and V-Net[8]. In medical segmentation TL literature, Ghafoorian et.al[9] fine-tune the last K layers and fixes the remain layers in a pre-trained neural network to segment Brain Lesion. Kaur et.al[10] compare three strategies which are fine-tuning only the encoder, fine-tuning only the decoder, and fine-tuning all layers to transfer from brain lesion segmentation to brain tumor segmentation. To transfer knowledge from natural domain to medical domain, Amiri et.al[11] put forward two strategies that training either the encoder part progressively or the decoder part progressively.

All the aforementioned TL strategies are based on human cognition that which layers in the neural network should be fine-tuned or fixed. However, such intervention is not able to generate a reliable TL strategy. A new strategy [12] is to utilize a policy network that based on target data to determine whether layers in the network should be fine-tuned or fixed. There are two drawbacks. The first is doubling the number of the parameters in the whole network that, dramatically increases the parameter size and training time while handling with three-dimension tasks. On the other hand, it cannot control every channel at a fine-grained level while the importance of different kernels in the pre-trained network is not the same for the target data. To improve this strategy , we propose a novel TL strategy for a three-dimension Auto-encoder like segmentation neural network. Instead of fine-tuning or fixing all the layers in the network, we choose to fine-tune the decoder part adaptively. The encoder part is fine-tuned to generate latent code and policy value. To step further, Squeeze and Excitation block[13] is utilized to acquire more information on a channel-wise level. Our contributions are as followed:

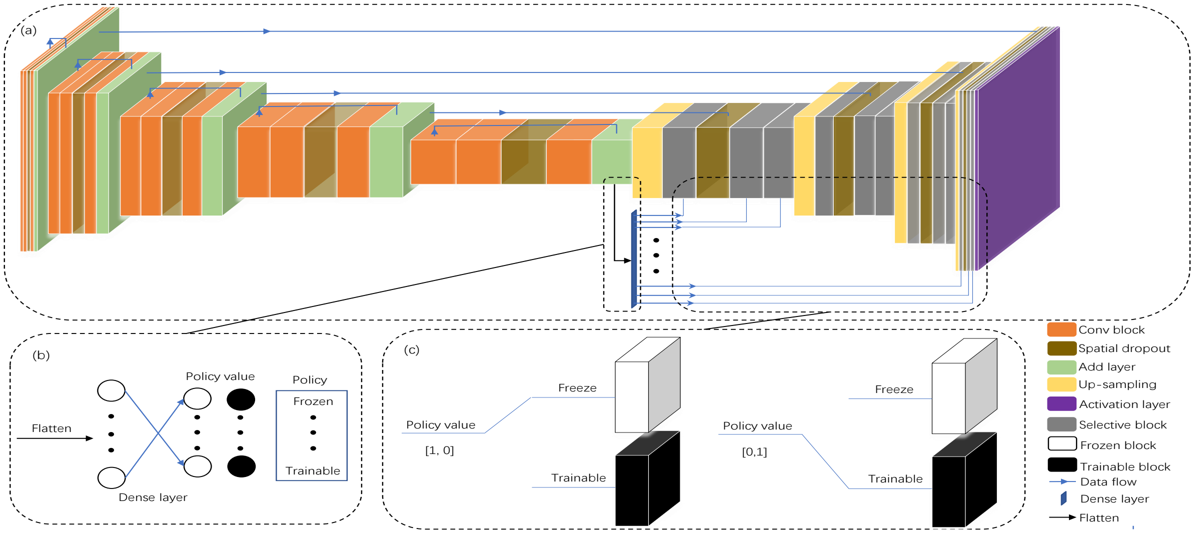

Figure 1: Whole network structure. (a) The backbone is 3D-U-Net. (b) The latent feature output by the encoder goes to a flatten operation and then a dense layer followed by a softmax acitvation to get policy value. Policy value is the one-hot encode of choosing either frozen or trainable block. (c) Every grey blocks has two branches but only one block allows data to flow. When policy value is [1,0], the data flows to the frozen block, when policy value is [0,1], the data flows to the trainable block.

- * We propose a novel strategy for TL in a segmentation task that fine-tunes the decoder part adaptively.
- * We leverage a Squeeze and Excitation (SE) block to obtain a fine-grained control for every channel in the decoder.
- * Through a series of experiments, we demonstrate that our method achieves state-of-the-are (SOTA) while transferring knowledge from MRI T2 modality to MRI T1, Flair, and T1ce modalities.

2 Methods

We already have a well trained network F on source data (X_S, Y_S), our goal is to transfer the pre-trained network F to adapt target data (X_T, Y_T). An overview of the method is introduced in section 2.1. Then, section 2.2 describes the policy value based TL strategy in detail. At last, a fine-grained control of kernels on the channel level is shown in section 2.3

2.1 Overview

A 3D U-Net F is first trained on the source domain. The output of l_{th} convolution block in the decoder part of F is:

$$O_l = O_{l-1} * F_l, \tag{1}$$

We copy a new trainable block $\widehat{F_l}$ and freeze the original block F_l. To determine which block will be used in the network, a weighted sum that obtains the new output of this block is preferred:

$$O_l = wO_{l-1} * F_l + (1 - w)O_{l-1} * \widehat{F_l}, \tag{2}$$

where w and 1-w are the policy value. As w is either 0 or 1, the output can be only affected by one block. Figure 1 (a) illustrates the overall structure of our method. The overall training process can be concluded as followed:

1. Given a target training sample, the latent code is extracted by the encoder. Afterwards, the latent code flows to the policy branch and the decoder.
2. In fig 1(b), the policy value, listed as w and 1-w, is generated through a process where latent feature proceeds to a flatten layer, a dense layer, and a softmax activation successively.
3. In fig 1(c), the policy value controls the switch button in the gray block. In case of policy value is [1,0], the button turns to the frozen block. While policy value equals [0,1], the button opts to the trainable block.
4. The weights in the encoder part are updated to extract latent features. The weights in the dense layer are updated to assign proper policy value to the layers in the decoder part. The weights in the decoder part are adaptively fine-tuned according to the policy value.

2.2 Policy value

The policy value is the one-hot encoding for choosing the trainable block or frozen block. More specifically, [1,0] represents choosing a frozen block, and [0,1] means choosing a trainable block.

Guo et.al [12] utilize a side network as a feature extractor that consists of 3 Residual blocks to extract the features from the input target data. In this work, the pre-trained encoder is applied as a feature extractor, since it is an already well-trained extractor in an Auto-encoder like network[14]. Thus, without involving any side networks, a flatten layer-a fully connected layer-a conditional softmax is added after the encoder. The conditional softmax equation is:

$$S_i = \frac{e^{i/T}}{\sum_j e^{j/T}} \tag{3}$$

Here i represents the ith element in a vector V and S_i is the softmax value for this element. In this work, we have two elements which are choosing a frozen block, and the other one is choosing a trainable block. T is the conditional number that equals 0.1 in this work. Original softmax can not promise a dichotomous form. Introducing a conditional number pushes a larger value to 1 and a smaller value to 0.

In [12], a Gumbel softmax[15] is proposed to improve the robustness of the whole network. By involving Gumbel softmax Equation (3) becomes:

$$S_i = \frac{e^{(\log i + G_i)/T}}{\sum_j e^{(\log j + G_j)/T}} \tag{4}$$

where G is a standard Gumbel distribution when $G = -\log(-log(U))$ with $U \sim uniform[0,1]$. G is actually a regularizer, since it adds randomness to the training to improve its robustness. The Gumbel softmax has the probability to swap the policy value from [1,0] to [0,1]. However, through our ablation study, we find that using a Gumbel softmax degrades the performance of the whole network when cooperating with fine-grained control.

2.3 Fine-grained control

In TL, not all kernels in Conv layers are equally important to the target task. The attention mechanism [16] is a powerful operation in deep learning area, which assigns larger weights to the key feature and smaller weights to the less important feature. Squeeze and Excitation (SE) block[13] is a kind of attention mechanism on the channel level. In addition, the SE block also explores the channel-wise relationship. Figure (2) illustrates the process of the SE block in this work. A 3D feature map $L * W * H * C$ goes to a Squeeze and Excitation branch. Max-pooling layer catches the significant feature in every channel. This process is termed as squeeze operation. Two dense layers followed by a sigmoid activation explore the channel-wise relationship. This process is termed as the excitation operation. The output of the SE block is generated by dot producting the result of Squeeze and Excitation and original feature map.

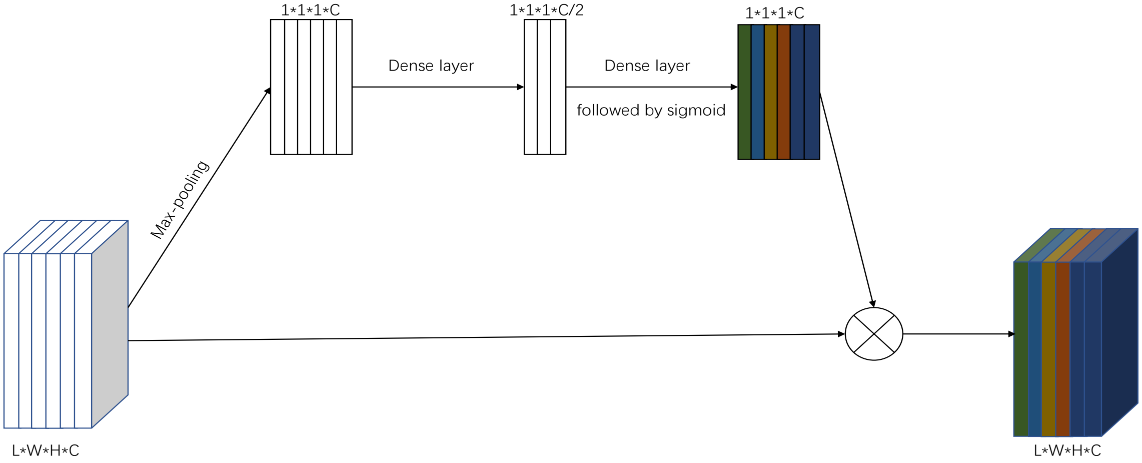

Figure 2: SE block. L, W, H represent length, width and height of the 3D-feature map respectively, C represents channel. Different color represents different weights, $\otimes$ means dot product.

SE block is added directly after every convolution block in the decoder section, thereby achieving adaptive control on the layer level and a fine-grained control on the channel level.

3 Experiment

All our experiments are carried out on a server with an AMD Ryzen Threadripper 1950X 16-Core Processor, and an NVIDIA Corporation GV100 TITAN V. We implement our code by Python 3.6, Tensorflow 2.1. We use the same training hyper-parameter along with the whole experiment. More specifically, our batch size is 1 due to GPU capacity limit. Our initial learning rate is 0.001, which will be reduced to half if there is no improvement in the validation process, and we train for 150 epochs. We utilize Dice loss as our loss function:

$$d = 1 - \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{5}$$

where d is the dice loss, X and Y are the predicted and ground truth respectively.

3.1 Data and preprocessing

In this work, Medical Segmentation Decathlon Brain Tumor MRI T2 modality is selected to be our source data, which consists of BraTS 2016 and BraTS 2017 [17], with 484 training data. BraTS 2018

	Flair	T1ce	T1
Scratch	0.866 ± 0.079	0.714 ± 0.124	0.719 ± 0.149
Ft-encoder	0.873 ± 0.081	0.722± 0.156	0.756 ± 0.106
Ft-decoder	0.865 ± 0.085	0.710 ± 0.148	0.734 ± 0.089
Ft-all	0.872 ± 0.089	0.727 ± 0.135	0.761 ± 0.089
Spottune [12]	0.876 ± 0.079	0.735 ± 0.135	0.750 ± 0.091
Our method	**0.884 ± 0.079**	**0.738 ± 0.134**	**0.768 ± 0.085**

Table 1: The Dice score of whole tumor segmentation result for the test data under different modality from Brats 2018 by different methods. And 'ft' means fine-tune.

MRI Flair, T1, T1ce are chosen to be our target data[18] with 285 training data. The ratio of training, validation, and testing is 0.6:0.2:0.2.

All data are pre-processed by N4 Bias Field correction that enhances the information in the MRI and resized to 128*128*128 to fit 3D U-net. We focus on the whole tumor segmentation, so we set all the labels in the ground truth to 1 and other places to 0.

3.2 Segmentation result

For a segmentation TL task, common strategies are fine-tuning encoder, fine-tuning decoder, and fine-tuning all layers. Except for the aforementioned baselines, we also compare our method with [12]method and training from scratch. The standard deviation and mean dice score within the data set are recorded. The detail of the segmentation result is shown in table (1).

Our method excels all the baselines under three different MRI modalities. Apparently, TL assists the segmentation task. Besides, we find that [12]work performs relatively poor than our work. Doubling the parameter size tends to over-fit in TL. The parameter size for our work is 17 m and the parameter size for [12] is 30 m.Instead, fine-tuning the encoder to obtain policy value and fine-tuning the decoder adaptively is more suitable for segmentation task. Moreover, we have 20% time-saving and 40% parameter-saving in training compared to the [12].

3.3 Ablation study

In this section, an ablation study has been carried out to explore the impact of the Gumbel softmax and SE block to the segmentation result. Comparative Test has been held among the situation (1) Without both Gumbel softmax and SE block; (2) With Gumbel softmax; (3) With SE block; (4) With both Gumbel softmax and SE block.

	Flair	T1ce	T1
Without G and SE	0.872	0.723	0.757
With G without SE	0.876	0.724	0.750
Without G with SE	**0.884**	**0.738**	**0.768**
With both G and SE	0.877	0.727	0.758

Table 2: Ablation study on Gumbel softmax and SE block, G is the abbreviation of Gumbel softmax.

It can be concluded from table (2), compared with the scenario without G and SE, dice score is improved by incorporating with either Gumbel softmax or SE block. However, adding only SE block outperforms adding both of them. We believe that is because Gumbel softmax is a too strong regularizer that damages the channel-wise attention mechanism. Therefore, only the SE block is exploited in our work.

3.4 Methods ranking

In the machine learning field, ranking methods are utilized to check if there exists a significant difference among different methods. Ranking in statistics generates a more reliable result. Friedman test [19] is used in this work. For each MRI modality (T1, Flair and T1ce), five data sets are constructed of random numbers of training data. Totally, 15 data sets are selected to carry out the Friedman test. Figure (3) illustrates the ranking result by the Friedman test. A smaller overlap between the

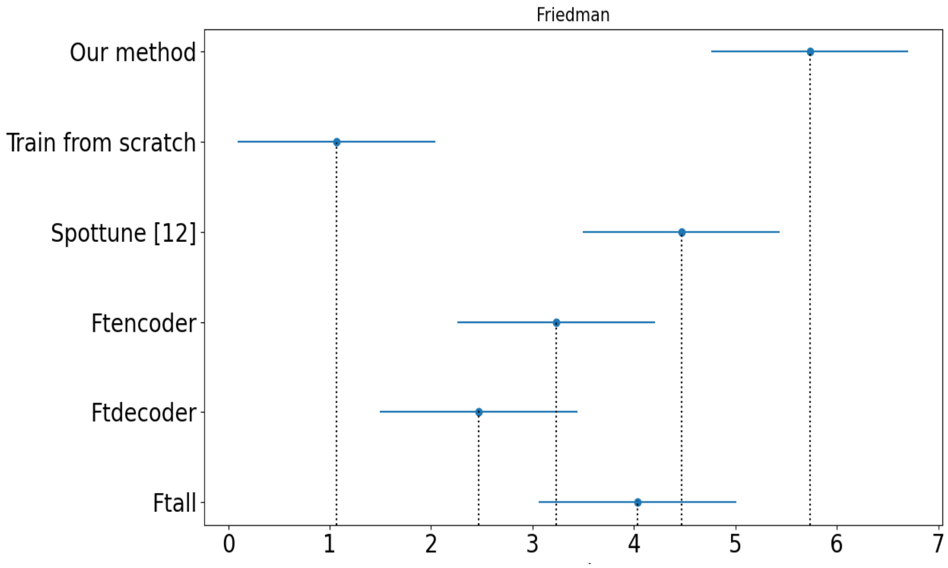

Figure 3: Friedman test. The horizontal axis is the ranking result, higher ranking result means better.

two methods indicates a more significant difference between them. Statistically, TL assists in the segmentation task, and our method outperforms the other methods.

4 Conclusion

In this work, a novel TL strategy is proposed for the segmentation task. In previous TL methods, human interference is required to determine which layer in the network to be fine-tuned or fixed. On the contrary, we put forward an auto policy strategy that fine-tunes the segmentation network adaptively. Through a series of experiments, we demonstrate the feasibility of our method in the medical area. In addition to that, it is practical in other segmentation tasks.

The link to our paper is https://ieeexplore.ieee.org/document/9434100.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[4] Xuehai He, Xingyi Yang, Shanghang Zhang, Jinyu Zhao, Yichen Zhang, Eric Xing, and Pengtao Xie. Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medRxiv*, 2020.

[5] Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan. Chest pathology detection using deep learning with non-medical training. In *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pages 294–297. IEEE, 2015.

[6] Jongwon Chang, Jisang Yu, Taehwa Han, Hyuk-jae Chang, and Eunjeong Park. A method for classifying medical images using transfer learning: A pilot study on histopathology of breast

cancer. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–4. IEEE, 2017.

[7] 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.

[8] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[9] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram Van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 516–524. Springer, 2017.

[10] Barleen Kaur, Paul Lemaître, Raghav Mehta, Nazanin Mohammadi Sepahvand, Doina Precup, Douglas Arnold, and Tal Arbel. Improving pathological structure segmentation via transfer learning across diseases. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 90–98. Springer, 2019.

[11] 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.

[12] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4805–4814, 2019.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[14] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[15] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[17] Christopher T Lloyd, Alessandro Sorichetta, and Andrew J Tatem. High resolution global gridded data for use in population studies. *Scientific data*, 4(1):1–17, 2017.

[18] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[19] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.