A GAN based Framework for Multi-Modal Medical Image Segmentation

Nurun Nahar[1]
MSc Student
South China University of Technology
Guangzhou, China

Sarwan Soomro[2]
MSc Student
South China University of Technology
Guangzhou, China

Ahmed Afif Monrat[3]
PhD Student
Lulea˚ University of Technology
Skellefteå, Sweden

---***---

***Abstract*—**Image segmentation is the procedure of dividing a digital image into a multiple set of pixels. The prior goal of the segmentation is to make things simpler and transform the representation of medical images into a meaningful subject. Many variant modalities, such as CT, X-ray, MRI, microscopy, positron emission tomography, single photon emission computer tomography, among others, make segmentation difficult. The challenging problem is for segmenting the regions with missing edges, absence of texture contrast, region of interest (ROI), and background. Most of the current researches only focuses on single-mode or paired multimodal images, and there are few researches on single-mode processing of unpaired multimodal images (Unified multimodal), the latter is more flexible and has good generalization ability in processing medical images. Based on the analysis of the existing medical image problems, this paper focuses on the unified multimodal segmentation of medical images by leveraging General Adversarial Network (GAN). GAN provides a simple and effective paradigm for image generation, which has been increasingly used in different fields such as migration learning and data enhancement. In this research, we have proposed a GAN based framework for addressing the issues concerning multi-modal medical image segmentation.

***Index Terms*—Multi-modal image segmentation; Generative Adversarial Network (GAN); Generator; Discriminator; Image classification.**

1. INTRODUCTION

In recent years, with the rapid development of deep learning theory and hardware, deep learning is more and more widely used in the field of medical image processing, and has greatly improved the performance of traditional methods. However, the following problems still need to be solved in the field of medical image:

a) Due to the particularity of medical field, the data needed for training is scarce, and it is difficult to obtain.

b) The new model structure is mostly for a single training set, which is inevitable to over fit, and the generalization ability of the model structure is poor. Both researchers and doctors are trapped in how to choose the best model structure.

c) Most of the current researches only focus on single-mode or paired multimodal images, and there are few researches on single-mode processing of unpaired multimodal images (Unified multimodal), the latter is more flexible and has good generalization ability in processing medical images.

Generation antagonism network (GAN) provides a simple and effective paradigm for image generation, which has been increasingly used in different fields such as migration learning and data enhancement. On the one hand, the generated training data can alleviate the problem of data scarcity to a certain extent, on the other hand, the idea of generating confrontation is used to map the data of different modes to the unified feature space, which is conducive to solving the unified multimodal problem.

Based on the analysis of the existing medical image problems and the development of Gan, this paper focuses on the unified multimodal segmentation of medical images. This research is focused on multi-modal medical image segmentation by leveraging GAN.

2. LITERATURE REVIEW & RELATED WORKS

Recently, the problem of multimodal segmentation has been extensively studied. Nie et al. [1] trained a network for three modalities and fused their high-layer features for the final segmentation. Tseng et al. [2] proposed cross-modality convolution layers to better leverage multimodal information. However, these methods were limited because they required paired registered images. An alternative approach is to train different models for different modalities in a shared latent space by extracting a common representation from different modalities. Kuga et al. [3] trained multimodal encoder-decoder networks with a shared representation. Valindria et al. [4] extracted modality-independent features by sharing the last layers of the encoder. However, these methods were not unified and required more parameters because of the requirement of specific encoder-decoder architectures for each modality. To extract modality-invariant features efficiently, Xu et al.

[5] represented a multimodal distillation module. Hu et al. [6] performed feature recalibration using SE (squeeze-and-excitation) blocks.

Recently, adversarial learning has been regarded as an effective way to transfer knowledge across different image domains. Huo et al. [7] presented an end-to-end synthesis and segmentation network with unpaired MRI and CT images.

To address limited scalability and robustness in translating among more than two domains, Choi et al. [8] developed a scalable approach (StarGAN) that can perform image-to-image translation for multiple domains using an unified model. Yuan et al. [9], proposed a Unified Attentional Generative Adversar- ial Network for Brain Tumor Segmentation From Multimodal Unpaired Images, which is a two-stream translation and seg- mentation unified attentional generative adversarial network (UAGAN), which can perform any-to-any image modality translation and segment the target objects simultaneously in the case where two or more modalities are available. Our research is going to be heavily influenced by this novel approach for analyzing and segmenting modality images.

3. CHALLENGES WITH MULTI-MODALITY IMAGE

With the rapid development of deep medical image segmentation, the following problems still need to be solved in the field of medical image segmentation:

a) Medical images are complex and lack of simple linear features [10]. The segmentation algorithm is affected by artifacts and other factors. On the one hand, medical image involves patient privacy. Different hospitals have different image collection standards and inconsistent data, which makes it difficult to collect large-scale medical data, but the success of deep learning often depends on a large number of annotation data. On the other hand, the training data with good quality needs a lot of time annotation by professional doctors. Pixel level annotation is time-consuming, laborious and tedious. How to solve the problem of medical data scarcity has become an ongoing topic in the field of medical image research.

b) Most of the new model structures aim at a single training set, which inevitably leads to over fitting and poor generalization ability of the model structure. For example, the shape and size of liver and brain tumors are different, and the segmentation model designed for liver segmentation is not necessarily applicable to the brain. Both researchers and doctors are trapped in how to choose the best model structure, and the complex super parameters are also a major factor limiting the landing of segmentation model. Whether there is a general model or how to design a general model has become another hot topic in medical image research, which is related to the actual landing of computer-aided diagnosis. Isensee et al. proposed a general NNU net automation framework [11], which can automatically adjust the structure of the model according to the problem domain. This work has won the champion of medical segmentation triathlon, which provides a feasible idea for the general design of the model.

c) At present, most researches only focus on single-mode or paired multimodal images, and few researches deal with non-paired multimodal images with a single general model [12]. Due to the complexity of medical image itself, the imaging principles of different modal images are different, and the resulting images are inconsistent. At present, there is no fully automatic segmentation method applicable to all modal medical images. Taking liver segmentation as an example, CT or MRI can be used to map the liver to different feature spaces. Although they have different image features, they have potential sharing features of liver. How to use a general model to process multi-modal medical image data at the same time is still a blank in related research fields.

Combined with the development of medical image model and the above mentioned problems in the field of medical im- age, this study focuses on the unified multimodal segmentation problem, that is, using a single unified segmentation model to process different modes.

4. PROBLEM STATEMENT & RESEARCH OBJECTIVES

In medical applications, the same anatomical structures may be observed in multiple modalities despite the different image characteristics [13]. Currently, most deep models for multimodal segmentation rely on paired registered images. However, multimodal paired registered images are difficult to obtain in many cases. Therefore, developing a model that can segment the target objects from different modalities with unpaired images is significant for many clinical applications. Deep learning algorithms produces state-of-the-art results for different machine learning and computer vision tasks. To perform well on a given task, these algorithms require large dataset for training. However, deep learning algorithms lack generalization and suffer from over-fitting whenever trained on small dataset, especially when one is dealing with medical images. For supervised image analysis in medical imaging, having image data along with their corresponding annotated ground-truths is costly as well as time consuming since annotations of the data is done by medical experts manually. In this paper, we propose a new Generative Adversarial Network for Medical Imaging. This approach generates synthetic medical images and their segmented masks, which can then be used for the application of supervised analysis of medical images.

Inspired by classic generative adversarial networks (GAN), we propose an end-to-end adversarial neural network, for the task of medical image segmentation. Since image segmentation requires dense, pixellevel labeling, the single scalar real/fake output of a classic GAN's discriminator may be ineffective in producing stable and sufficient gradient feedback to the networks. Instead, we use a fully convolutional neural network as the segmentor to generate segmentation label maps, and propose an adversarial critic network with a multi-scale L1 loss function to force the critic and segmentor to learn both global and local features that capture long- and short- range spatial relationships between pixels. In our proposed framework, the segmentor and critic networks are trained in an alternating fashion in a min-max game: The critic takes as input a pair of images, (original image predicted label map, original image ground truth label map), and then is trained by maximizing a multi-scale loss function; The segmentor is trained with only gradients passed along by the critic, with the aim to minimize the multi-scale loss function. We show that such a framework is more effective and stable for the segmentation task, and it leads to better performance than the state-of-the-art U-net segmentation method.

5. METHODOLOGY & PROPOSED ARCHITECTURE

One neural network, called the generator, generates new data instances, while the other, the discriminator, evaluates them for authenticity; i.e. the discriminator decides whether each instance of data that it reviews belongs to the actual training dataset or not. The generator is creates new, synthetic images that it passes to the discriminator. It does so in the hopes that they, too, will be deemed authentic, even though they are fake. The goal of the generator is to generate passable hand-written digits: to lie without being caught. The goal of the discriminator is to identify images coming from the generator as fake.

Here are the steps a GAN takes:

a) The generator takes in random numbers and returns an image.

b) This generated image is fed into the discriminator along- side a stream of images taken from the actual, ground- truth dataset.

c) The discriminator takes in both real and fake images and returns probabilities, a number between 0 and 1, with 1 representing a prediction of authenticity and 0 representing fake. Double feedback loop:

d) The discriminator is in a feedback loop with the ground truth of the images, which we know.

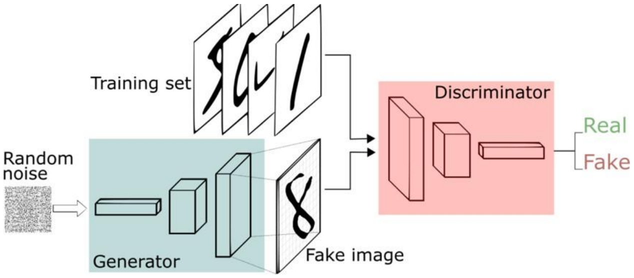

Fig. 1. Functional diagram of Generative Adversarial Network (GAN)

e) The generator is in a feedback loop with the discriminator.

From Fig 1. We can observe that, GAN, consisting of two modules, a discriminator D and a generator G. D learns to distinguish between real and transformed images and classify the real images to its corresponding domain. G takes in as input both the image and target domain label and generates a transformed image. The target domain label is spatially replicated and concatenated with the input image. G tries to reconstruct the original image from the transformed image given the original domain label. G tries to generate images indistinguishable from real images and classifiable as target domain by D.

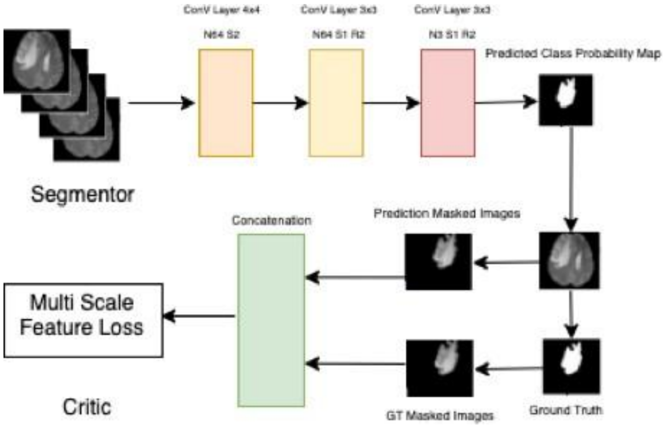

Fig. 2. A Network architecture with segmentor and critic networks. Marketplace Application's State Transition [14]

Fig 2. Illustrates another GAN network architecture with segmentor and critic networks. 4 × 4 convolutional layers with stride 2 (S2) and the corresponding number of feature maps (e.g., N64) are used for encoding, while image resize layers with a factor of 2 (R2) and 3 × 3 convolutional layers with stride 1 are used for decoding. Masked images are calculated by pixel-wise multiplication of a label map and (the multiple channels of) an input image. Note that, although only

one label map (for whole tumor segmentation) is illustrated here, multiple label maps (e.g. also for tumor core and Gd-enhanced tumor core) can be generated by the segmentor in one path.

Fig 3. is an overview of StarGAN, consisting of two modules, a discriminator D and a generator G. D learns to distinguish between real and fake images and classify the real images to its corresponding domain. G takes in as input both the image and target domain label and generates an fake image. The target domain label is spatially replicated and concatenated with the input image. G tries to reconstruct the original image from the fake image given the original domain label. G tries to generate images indistinguishable from real images and classifiable as target domain by D.

Both of these models are going to be the main research motivation for our research. The thesis work aims to study and incorporate the concept of StarGAN and SeGAN model for developing a robust framework for higher visual quality comparing to existing model and minimize the multi scale loss function.

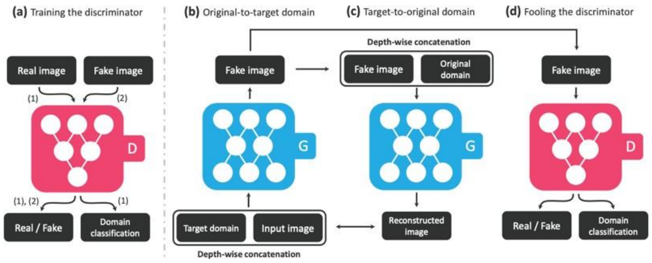

Fig. 3. StarGAN model

A. Proposed Methodology

The proposed model consists of two parts: the segmentor network S and the critic network C. The segmentor is a fully convolutional encoder-decoder network that generates a probability label map from input images. The critic network is fed with two inputs: original images masked by ground truth label maps, and original images masked by predicted label maps from S. The S and C networks are alternately trained in an adversarial fashion: the training of S aims to minimize our proposed multi-scale L1 loss, while the training of C aims to maximize the same loss function.

The conventional GANs have an objective loss function defined as:

$$\min_{\theta_G}\max_{\theta_D} L(\theta_G, \theta_D)$$

$$= E_{x\sim P_{data}}[\log D(x)] + E_{z\sim P_z}[\log\left(1 - D\left(G(z)\right)\right)]\ (1)$$

In this objective function, x is the real image from an unknown distribution P_{data}, and z is a random input for the generator, drawn from a probability distribution (such as Gaussion) P_z. G and D represent the parameters for the generator and discriminator in GAN, respectively.

In our proposed framework, given a dataset with N training images x_n and corresponding ground truth label maps y_n, the multi-scale objective loss function L is defined as:

$$\min_{\theta_S}\max_{\theta_C} L(\theta_S, \theta_C)$$

$$= \frac{1}{N}\sum^{N} l_{mae}(f_C(x_n \circ S(x_n)), f_C(x_n \circ y_n))\ (2)$$

where l_{mae} is the Mean Absolute Error (MAE) or $L1$ distance; x_n *o* $S(x_n)$ is the input image masked by segment or predicted label map (i.e., pixel-wise multiplication of predicted label map and original image); x_n *o* y_n is the input image masked by its ground truth label map (i.e., pixel-wise multiplication of ground truth label map and original image); and $f_C(x)$ represents the hierarchical features extracted from image x by the critic network. More specifically, the l_{mae} function is defined as:

$$l_{mae}\left(f_C(x), f_C(x')\right) = \tfrac{1}{L}\textstyle\sum_{i=1}^{L} ||f_C^i(x) - f_C^i(x')||_{1.}\ (3)$$

Where L is the total number of layers/scales in the critic network, and $f_C(x)$ is the extracted feature map of image x at the ith layer of C.

6. CONCLUSION

Based on the analysis of the existing medical image problems, this paper focuses on the unified multimodal segmentation of medical images by leveraging General Adversarial Network (GAN). In this research, we have proposed a GAN based framework for addressing the issues concerning multi-modal medical image segmentation. Currently we are working on finding proper dataset and conduct experiments based on the proposed methodology for developing a proof of concept. The next research article will contain the experimental observations and the efficiency of the proposed method.

REFERENCES

[1] D. Nie, L. Wang, Y. Gao, and D. Shen, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," in 2016 IEEE 13Th international symposium on biomedical imaging (ISBI). IEEE, 2016, pp. 1342–1345.

[2] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learn- ing and cross-modality convolution for 3d biomedical segmentation," in Proceedings of the IEEE

conference on Computer Vision and Pattern Recognition, 2017, pp. 6393–6400.

[3] R. Kuga, A. Kanezaki, M. Samejima, Y. Sugano, and Y. Matsushita, “Multi-task learning using multi-modal encoder-decoder networks with shared skip connections,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 403–411.

[4] V. V. Valindria, N. Pawlowski, M. Rajchl, I. Lavdas, E. O. Aboagye,

A. G. Rockall, D. Rueckert, and B. Glocker, “Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri,” in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 547–556.

[5] D. Xu, W. Ouyang, X. Wang, and N. Sebe, “Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 675–684.

[6] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[7] Y. Huo, Z. Xu, S. Bao, A. Assad, R. G. Abramson, and B. A. Landman, “Adversarial synthesis learning enables segmentation without target modality ground truth,” in 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, 2018, pp. 1217–1220.

[8] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan:Unified generative adversarial networks for multi-domain image-to-image translation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.

[9] W. Yuan, J. Wei, J. Wang, Q. Ma, and T. Tasdizen, “Unified attentional generative adversarial network for brain tumor segmentation from multi- modal unpaired images,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 229–237.

[10] 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. Springer, 2015, pp. 234–241.

[11] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” Nature Methods, pp. 1–9, 2020.

[12] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest et al., “The multimodal brain tumor image segmentation benchmark (brats),” IEEE transactions on medical imaging, vol. 34, no. 10, pp. 1993–2024, 2014.

[13] Z. Zhang, L. Yang, and Y. Zheng, “Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9242–9251.

[14] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, “Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation,” Neuroinformatics, vol. 16, no. 3-4, pp. 383–392, 2018.