Dose Prediction Challenge

Hugo Malard HUGO.MALARD@ENS-PARIS-SACLAY.FR
Denys Sikorskyi DENNIS.SIKORSKYI@ENS-PARIS-SACLAY.FR
ENS Paris-Saclay

Editors: Under Review for MIDL 2023

Abstract

Medical image segmentation is a crucial step in medical image analysis, as it plays a critical role in diagnosis, treatment planning, and patient monitoring. Deep learning techniques, especially convolutional neural networks (CNNs), have shown remarkable results in medical image segmentation. In this work, we study the use of U-net for brain segmentation using multi modal inputs.

Keywords: Brain Segmentation, U-Net, Depthwise Convolutions, Multi-modal.

1. Introduction

Medical imaging is a significant diagnostic and treatment-planning tool in a variety of medical applications. Radiation dose delivery accuracy and precision are crucial for optimal treatment results in radiation therapy. Medical imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI), give helpful information for radiation treatment planning, but the precise delineation of target structures and healthy tissues are also essential.

The technique of finding and isolating certain structures or regions of interest from medical imaging is known as segmentation. It is critical in radiation therapy treatment planning because it correctly identifies and delineates target structures and healthy tissues in CT and MRI images. Medical picture segmentation is a difficult process because of the complexity of anatomical features, diversity in patient anatomy, and image noise and abnormalities. For various reasons, accurate segmentation is critical. For starters, it allows for precise target delineation, which is essential for accurately delivering radiation doses to the tumor while reducing damage to surrounding healthy tissues. Second, proper segmentation can increase radiation therapy planning efficiency by minimizing the time necessary for manual demarcation and inter-observer variability.

As a result, establishing precise and effective segmentation algorithms is critical for radiation therapy treatment planning. During the challenge, we use deep learning-based algorithms to address the problem of segmentation in radiation therapy treatment planning. We present a method for reliably segmenting target structures and healthy tissues in CT images by using the capabilities of deep neural networks.

2. Model Architecture

2.1. U-net architecture

The U-Net(Olaf Ronneberger, 2015) architecture was first introduced in a 2015 paper by Ronneberger et al. The U-Net is a convolutional neural network (CNN) architecture that

is designed for segmentation tasks. The U-Net architecture consists of an encoder and a decoder. The encoder is a series of convolutional layers that are used to extract features from the input image. The decoder is a series of convolutional layers that are used to reconstruct the output image from the features extracted by the encoder. The architecture of the U-Net is shown in Figure 1.
The U-Net architecture is unique in that it has skip connections that connect the encoder and decoder layers. These skip connections allow information to flow directly from the encoder to the decoder, bypassing the intermediate layers. The skip connections are shown in Figure 1 as arrows that connect corresponding layers in the encoder and decoder.
U-Net has shown impressive performance in segmentation tasks, especially in the medical field. The U-Net has been used for a wide range of segmentation tasks, including the segmentation of organs, tumors, and lesions. The U-Net has shown state-of-the-art performance on several benchmarks.
The effectiveness of the U-Net can be attributed to several factors. First, the skip connections in the U-Net allow information to flow directly from the encoder to the decoder, which helps to preserve spatial information and reduce the loss of fine-grained details. Moreover, the U-Net is a relatively small and efficient architecture that can be trained on small datasets, making it ideal for medical imaging applications where large datasets are often not available.

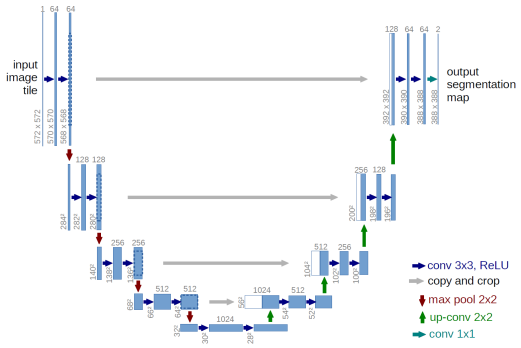

Figure 1: U-net architecture

2.2. Depthwise convolution

U-net architecture seems appropriate for our task, but we are not only using a single image as input but a concatenation of different inputs (CT, structure mask, and possible dose mask) on the channel axis. Therefore, using standard convolution would treat the channels together while they contain each different information.
Therefore using depthwise convolution would allow computing a different kernel for each input. The depthwise convolution applies a 2D convolution to each input channel independently, using a separate kernel for each channel. Depthwise convolution allows to the extraction of different features from each of the inputs but has the effect of increasing a lot the number of parameters. Therefore using only depthwise convolutions in the encoder

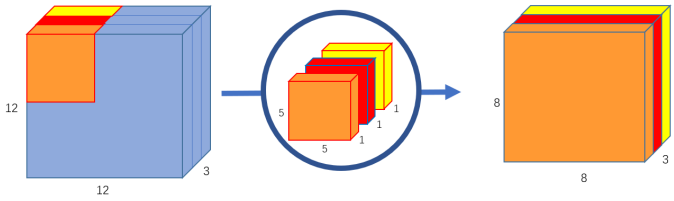

Figure 2: Depthwise convolution illustration

make the model overfit. We found out that only using it in the first block (features are then combined with standard convolutions) allows good results without big overfitting.

3. Training procedure and model tuning

3.1. Data description and preprocessing

The medical imaging dataset contains four picture types: CT, dose, structure masks, and possible dose. The task's purpose is to estimate the dosage picture from the other three types of photographs. The dataset is divided into 7800 training samples, 1200 validation samples, and 1200 test samples. The data is preprocessed prior to the deployment of various models to guarantee that the input to the models is uniform. All four types of images have a size of 128x128 pixels. The CT pictures are normalized and then concatenated with the other two types of images to generate the final input tensor for the model. This uniform input guarantees that the models can properly learn from the data. The model will be supplied with batches of the input tensor and matching dosage pictures during the training phase. Using multiple neural network topologies and training methodologies, the model will learn to transfer the input tensor to the appropriate dosage picture. The validation set is used to monitor the model's performance during the training phase and to prevent overfitting. Finally, the test set is used to evaluate the model's performance on new, previously unknown data.

3.2. Baseline and Transfer learning

Firstly, in order to find out the base MAE loss for different possible architectures we tried to implement different possible image segmentation models using transfer learning and see the final results of them. We have implemented different architectures such as MAnet(Yaxin Zhao, 2020), Linknet(Abhishek Chaurasia, 2017), FPN(Tsung-Yi Lin, 2016), PSPNet(Hengshuang Zhao, 2016), PAN(Can Zhang, 2020), DeepLabV3(Liang-Chieh Chen, 2017), and DeepLabV3+(Liang-Chieh Chen, 2018) with pre-trained backbones on ImageNet dataset like ResNet50(Ross Wightman, 2021) and ResNet101(Kaiming He, 2015).
During the training of these models, we found that the best possible result of these models is 0.61114 MAE loss with DeepLabV3+. Therefore, we based our training performance on these simple transfer learning models and used the loss of 0.61.
We also tried to start from a U-net trained for (uni-model) brain segmentation, by just

Model	Best MAE loss
Manet	0.82124
Linknet	0.78416
FPN	0.76247
PSPNet	0.72081
PAN	0.75528
DeepLabV3	0.63323
DeepLabV3+	0.61114

Table 1: ImageNet pre trained models performances

changing the first block to a depthwise convolution layer. We started by only training the first and last blocks and after some epochs training the full model. The initialization turned out to be not necessarily better than the random.

Model	Best MAE loss
Transfer learning	0.59
Random Init	0.57

Table 2: Brain segmentation pre-trained models performances

3.3. Training procedure and fine tuning

The final model that was used in this challenge was UNet which was trained from scratch on brain images. Its architecture is a U-net with depthwise convolution in the first block and transposed convolutions for upsampling. After different trials, we found that the optimal size of the model was a 4 blocks encoder and a 4 blocks decoder where each block contains 2 convolutional layers each followed by batch normalization and ReLU activation function. The optimal performances were achieved using an initial number of channels of 32 in the first block, multiplying it by 2 at each block in the encoder and dividing it by 2 at each decoder's block.
The final model is trained using the concatenation of CT, structure masks, and possible dose. After the predictions, the possible dose is used to mask everything outside of the zone of interest. This logically greatly helps the performances. We also provided the possible dose mask as input to the model since it can contain some additional information that could be useful to extract, even though it is also used after the prediction to constrain them.
The model was optimized using AdamW optimizer which we found a bit better than the standard Adam. We also found weight decay to play a crucial role in the final performances. Since our U-net makes use of a lot of batch normalization layers, we used the biggest batch size that we could afford: 32.

4. Results

The final model performs nicely on both the validation and the test set (0.4 and 0.36 respectively). We noticed that adding some augmentations to the input images (random horizontal flip and center crop) does not improve the performances, it even decreases them. We hypothesize that it may be due to the fact that medical images are not standard images and therefore changing the input image in a standard way impact the semantics of the image and just fool the model.

5. Conclusion and going further

To conclude, our study showed that depthwise convolutions can be nicely integrated into the U-net architecture without increasing a lot the number of parameters, allowing to the extraction of different features from the input modalities.
We observed that using the mask of the possible dose as a hard constraint (by multiplying the prediction by the mask) after the output of the model gives a real boost to the performances of the model.
Fusing the information of the different modalities is a tough task and to go further one may use a deep fusion strategy (having a different network that extracts features from the different modalities before merging them together in the decoder). But those kinds of approaches usually require a bigger number of parameters and are therefore more suited to a bigger dataset. (**?**)

References

Eugenio Culurciello Abhishek Chaurasia. *LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation*. 2017.

Guang Chen Lei Gan Can Zhang, Yuexian Zou. *PAN: Towards Fast Action Recognition via Learning Persistence of Appearance*. 2020.

Xiaojuan Qi Xiaogang Wang Jiaya Jia Hengshuang Zhao, Jianping Shi. *Pyramid Scene Parsing Network*. 2016.

Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. *Deep Residual Learning for Image Recognition*. 2015.

Florian Schroff Hartwig Adam Liang-Chieh Chen, George Papandreou. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017.

George Papandreou Florian Schroff Hartwig Adam Liang-Chieh Chen, Yukun Zhu. *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*. 2018.

Thomas Brox Olaf Ronneberger, Philipp Fischer. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015.

Hervé Jégou Ross Wightman, Hugo Touvron. *ResNet strikes back: An improved training procedure in timm*. 2021.

Ross Girshick Kaiming He Bharath Hariharan Serge Belongie Tsung-Yi Lin, Piotr Dollár. *Feature Pyramid Networks for Object Detection.* 2016.

Tangkun Zhang Yaxin Zhao, Jichao Jiao. *MANet: Multimodal Attention Network based Point- View fusion for 3D Shape Recognition.* 2020.