DLMI Challenge Report: Group 9

Xingjian Zhang XINGJIAN.ZHANG@POLYTECHNIQUE.EDU
École polytechnique, IP Paris, Palaiseau, France

Haiwei Fu HAIWEI.FU@STUDENT-CS.FR
CentraleSupélec, Gif-sur-Yvette, France

Editors: Under Review for MIDL 2023

Abstract

Radiation therapy is a common treatment for cancer, but determining the appropriate radiation dose can be challenging. Currently, radiation oncologists rely on guidelines and their own clinical experiences to plan the dose, resulting in variability in treatment. Deep learning models can provide a faster, potentially more accurate, and personalized approach to radiation dose planning.

In this report, we explore the use of convolutional neural networks (CNNs), generative adversarial networks (GANs) and U-Net architecture to predict radiation dose based on CT images, organ contours and a possible irradiation mask. CT images provide detailed anatomical information while organ contours help to define the shape and location of the organs that are being targeted. We use a 2D dataset derived from the OpenKBP challenge (Babier et al., 2021), which consists of 7800 training, 1200 validation and 1200 test samples. Our results show that deep learning models can accurately predict radiation dose compared to a ground truth simulation with a mean absolute error of less than 0.32.

The results highlight the potential of deep learning models in predicting radiation dose. Further validation and refinement of these models could aid in reducing the variability of radiation therapy and ultimately lead to improved patient outcomes.

Codalab username: **xingjian.zhang**.

Keywords: radiation dose prediction, deep learning, medical imaging.

1. Introduction

Radiation therapy is a crucial treatment option for cancer patients, and it involves delivering a specific amount of radiation dose to the tumor site while minimizing exposure to healthy tissues. However, determining the appropriate radiation dose can be challenging and requires the expertise of radiation oncologists who rely on guidelines and clinical experiences to plan the treatment. This variability in treatment planning can result in inconsistent outcomes for patients.

Recent advancements in deep learning offer the potential to provide a personalized approach to radiation dose planning, leading to faster and more accurate predictions. In this report, we extensively explore the use of CNNs and GANs to predict radiation dose. We study the performances of UNet architectures and conditional GANs, which have strong capabilities in semantic segmentation and image-to-image translation tasks.

© 2023 X. Zhang & H. Fu.

2. Architecture and methodological components

2.1. Model architecture

Our best performing model is a Pix2pix GAN (Isola et al., 2016) with a pretrained custom UNet++ (Zhou et al., 2018) architecture as the generator. We use cGAN because the problem can be formulated as a style transfer task: we are given images (CT, contours) and we want to learn to map them to images of another style (radiation dose maps). For our generator, we chose to experiment with UNet architectures. UNets are state-of-the-arts for semantic segmentation tasks, with the encoder-decoder design capable of capturing abstract feature representations.

Initially, we relied solely on UNet models and did not consider using cGANs. We experimented with several UNet architectures including UNet (Ronneberger et al., 2015), UNet++, and UNet3+ (Huang et al., 2020). We varied their depths, feature sizes, and convolution blocks to include residual, attention, and dilation. After thorough evaluation, we found that a custom UNet++ model with dilated convolutions performed the best. Increasing dilation rate in deeper layers could allow our model to capture more global representations with larger receptive fields. Additionally, we implemented separable convolutions to allow larger feature sizes under memory constraints. Based on our experiments, we found that feature sizes of [80, 160, 320, 640, 1280] performed better than conventional feature sizes up to 1024.

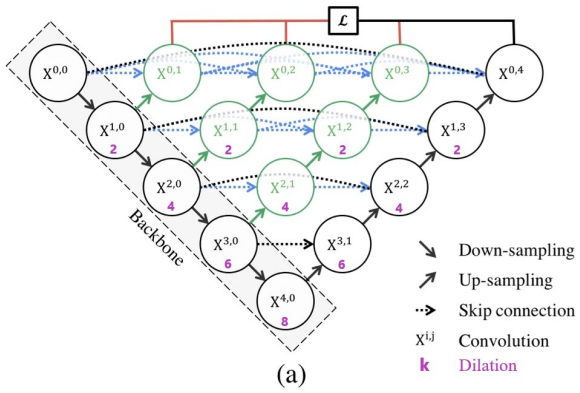

Figure 1: Custom UNet++ architecture with dilated convolution blocks.

During the competition, we observed a significant difference between the ground truth dose maps and our predicted dose maps - the absence of beam shapes. While the ground truth simulations displayed clear beam shapes, our UNet models alone could not learn this shape information. To tackle this issue, we explored generative networks, which use adversarial networks to distinguish between real and fake dose maps. We used our best performing custom UNet++ model as a pretrained generator and experimented with multiple discriminators including NLayerDiscriminators, PixelDiscriminators, and PatchDiscriminators, ultimately settling on a 3-Layer NLayerDiscriminator. With this approach, we achieved our best performance with a score of 0.31191 on the test set.

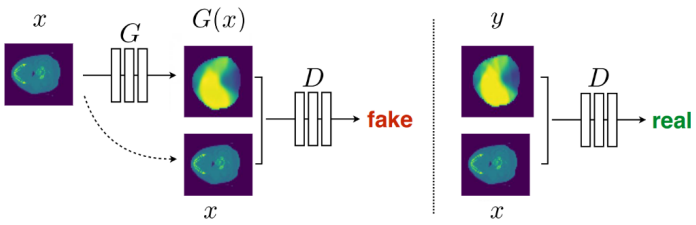

Figure 2: Pix2pix cGAN architecture: The generator (UNet++) generates dose predictions from inputs. The discriminator takes dose mappings and the generator inputs then try to distinguish between real and predicted dose maps.

2.2. Methodological components

2.2.1. Data treatment and augmentation

The data provided for this study comprises of gray-scale CT images and binary masks that represent organs, as well as a possible mask indicating the irradiation area. The CT images are standard clinical images with a color depth of 12 bits. The structural masks denote organs at risks (OARs) and radiation dosage plannings need to ideally avoid these organs. We created distance maps (DMs) based on the organ masks. Within an organ contour, pixels are assigned negative values that represent their Euclidean distance to the closest contour edge. Pixels outside the organ mask are assigned a value of zero. This approach is taken to allow the model to learn meaningful relationships between the dose quantity and the proximity of OARs to the radiation target.

Normalization is applied to the CT images and DMs. To introduce more noise and variability, we applied augmentations to the training data using Albumentations (Buslaev et al., 2020). Spatial transformations were applied in order: horizontal/vertical flip, rotation, scaling, and translation with 0.2 probabilities each. We do not apply shear or other spatial distortions as we expect all medical images to have realistic shapes.

2.2.2. Loss functions

As the challenge is evaluated using mean absolute error (MAE), we initially trained our models using just the L1 loss function. The UNet++/3+ architectures allow for intermediate predictions using deep supervision (DS) to obtain a more generalized prediction. The MAE is computed for multiple output predictions averaged

$$\mathcal{L}_{\text{DS}} = \frac{1}{N}\sum_{i=1}^{N}(\frac{1}{M}\sum_{j=1}^{M}|P_{i,j} - GT_i|) \tag{1}$$

where $P_{i,j}$ denotes the j'th prediction for sample i, GT_i is its corresponding ground truth. To help our model focus more on the regions where the irradiation is possible, we add an additional L1 loss where only the possible dose mask's area is considered. We define the weighted-L1 loss (WL) as follows

$$\mathcal{L}_{\text{weighted}} = \mathcal{L}_{\text{DS}} + 0.5 \times \mathcal{L}_{\text{DS possible dose mask}} \tag{2}$$

For the Pix2pix GAN, we follow the same objectives as the paper (Isola et al., 2016). We replaced L1 distance with our weighted-L1 loss. The final objective is thus

$$G^* = arg \min_G \max_D \mathcal{L}_{\text{cGan}}(G, D) + 100 \times \mathcal{L}_{\text{weighted}}(G) \tag{3}$$

2.2.3. Hyperparameters

AdamW optimizers and CyclicLR schedulers were used in the training process. The initial learning rate was set to 0.0001 and gradually increased to a maximum learning rate of 0.02 using the `triangle2` scheme which reduces the maximum learning rate after each cycle. A batch size of 16 was used with mixed precision during training. The Pix2pix model used 0.01 max learning rate for the generator, and 0.02 max learning rate for the discriminator. The models were trained for 60 epochs. We trained using Standard Google Colab GPUs, the best validation epochs were chosen as the results. The training duration varied from one to five hours depending on the size of the model and GPU allocation.

3. Model tuning and comparison

3.1. Result discussion

In Table 2, we detail the performances of some models that we tested. From the results, we see that using distance maps and weighted-L1 loss both improved our model predictions. Using dilated convolutions improved our test score from 0.320 to 0.316 for UNet++[1024], and increasing the size of features further lowered the test score to 0.314. We arrived at our lowest test score 0.312 using the Pix2pix GAN approach. However, we note that the validation score is the lowest in our UNet++[1280] model. This may suggest that the difference in test score between Pix2pix and UNet++[1280] is simply due to random variations.

During the competition, we tried many approaches that did not work. For example, we applied VOI LUT to the images during preprocessing, windowing the full range of pixels down to 8-bit color depth within the range of $[0, 255]$. We thought windowing may allow models to better generalize the organs as they have less variation in intensity, however they prove to be a detriment due to the loss of information. We tried to include Gabor filtered CT images as additional channels, or using residual connections and attention gates in our networks. We also tried to modify our weighted-L1 loss function to use Smooth-L1 loss, as well as adding another loss only for the pixels inside the organ contours. These approaches did not improve our model, and their test scores are omitted from the table.

3.2. Ablation study

We perform an ablation study on our Pix2pix/UNet++ model. In Figure 3 and Table 1, we show the results when dilated convolutions, distance maps, and deep supervision modules are subsequently removed. We also show the effects when we decrease our feature size to 1024, or use L1 loss instead of weighted-L1 loss.

Table 1: Ablation study for Pix2pix model components.

Models	**Loss**	**Validation score**	**Test score**
Pix2pix[UNet++[1280]], DS, DM, DL	WL, BCE	0.36768	0.31191
UNet++[1280], DS, DM, DL	WL	0.36553	0.31357
UNet++[1280], DS, DM, DL	L1	0.36945	0.31588
UNet++[1280], DS, DM	WL	0.36693	0.31599
UNet++[1280], DS	WL	0.36928	0.31942
UNet++[1280]	WL	0.37530	0.35149
UNet++[1024], DS, DM, DL	WL	0.36754	0.31644

We observe that when each module is removed, the MAE score on the test set increases. A visual ablation study shows that the dose map from the Pix2pix model had the most beam-like shapes on the left side of the image. When examine closely, we also note that it is able to simulate a dose with variations near the center, while worse models had constant intensity in the middle. Overall, the ablation results suggest that the added modules are beneficial to our model.

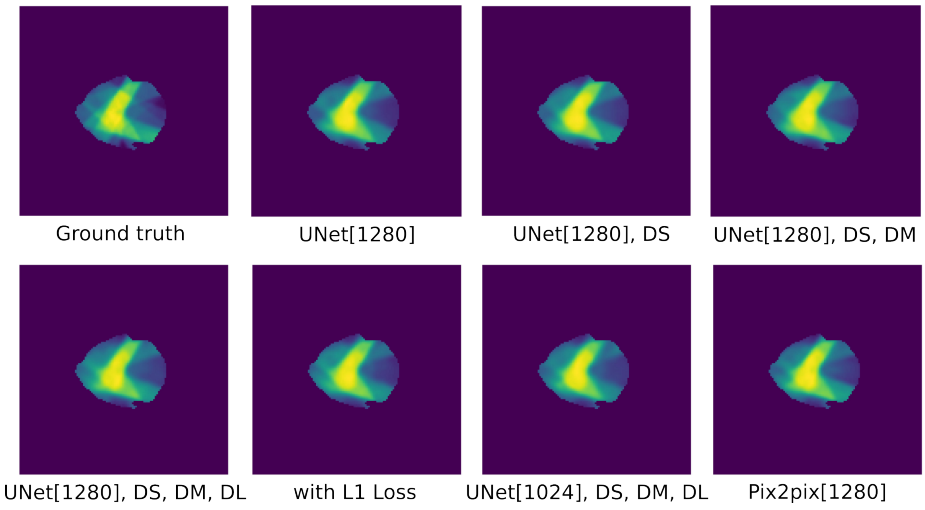

Figure 3: Visual ablation study on a validation sample.

4. Conclusion

In this report, we have shown that deep learning can be a powerful tool for radiation dose prediction. Through careful feature engineering and hyperparameter tunings, we were able to create a cGAN model that is robust and accurate at simulating radiation dosage given limited information. We demonstrated and validated our design choices through quantitative and qualitative ablation studies. To improve, we could investigate other architectures such as Vision Transformers and test on additional unseen datasets.

References

Aaron Babier, Binghao Zhang, Rafid Mahmood, Kevin L. Moore, Thomas G. Purdie, Andrea L. McNiven, and Timothy C. Y. Chan. OpenKBP: The open-access knowledge-based planning grand challenge and dataset. *Medical Physics*, 48(9):5549–5561, June 2021. doi: 10.1002/mp.14845. URL https://doi.org/10.1002/mp.14845.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. ISSN 2078-2489. doi: 10.3390/info11020125. URL https://www.mdpi.com/2078-2489/11/2/125.

Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation, 2020. URL https://arxiv.org/abs/2004.08790.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2016. URL https://arxiv.org/abs/1611.07004.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018. URL https://arxiv.org/abs/1807.10165.

Appendix A. Model components and results.

Table 2: MAE scores for different models and components: DS - deep supervision, DM - distance maps, WL - weighted loss, DL - dilated convolutions. The best scores are in **bold**.

Models	**Loss**	**Validation score**	**Test score**
UNet[1024]	L1	0.37589	0.32951
UNet++[512], DS	L1	0.37574	0.32851
UNet++[512], DS, DM	L1	0.37459	0.32383
UNet++[512], DS, DM	WL	0.37046	0.32360
UNet++[1024], DS, DM	WL	0.37070	0.32021
UNet++[1024], DS, DM, DL	WL	0.36754	0.31644
UNet++[1280], DS, DM, DL	WL	**0.36553**	0.31357
UNet3+[512], DS, DM, DL	WL	0.37891	0.32161
UNet3+[1024], DS, DM, DL	WL	0.37102	0.31738
Pix2pix[UNet++[1280]], DS, DM, DL	WL, BCE	0.36768	**0.31191**