Radiation dose prediction for Cancer Patients using UNet-based deep learning approach

Paul FOTSO KAPTUE ,Codalab Name: E.Ch PAUL.FOTSO-KAPTUE@POLYTECHNIQUE.EDU

Emna CHOURIA, Codalab Name: E.Ch EMNA.CHOURIA@PLOYTECHNIQUE.EDU

Editors: Under Review for MIDL 2023

Keywords: Radiation Dose Prediction, UNet, Encoder-bottleneck-decoder architecture.

1. Introduction

Radiation therapy is a widely used cancer treatment that uses high-energy Gamma rays to target and destroys cancer cells. However, radiation therapy can also damage healthy tissues surrounding the cancerous region. Therefore, accurately predicting the radiation dose needed to effectively treat cancer while minimizing harm to healthy tissues is critical. To achieve this goal, medical professionals use computer-based treatment planning systems to calculate the optimal radiation dose. These systems rely on complex mathematical models and physics-based simulations to predict the radiation dose distribution. However, they are very slow and approximative since the dosimetrist goes through many iterations to adjust and tune treatment planning parameters. This is where machine learning, and specifically deep learning, can play a vital role by allowing for more accurate and efficient dose prediction. Many methods have been proposed for predicting radiotherapy doses. In this work, we trained a U-Net using a 2D dataset to predict the dose of radiation that is delivered to a patient's tumor and surrounding healthy tissue.

2. Dataset

For network training, a dataset coming from the Open Knowledge-Based Planning (OpenKBP) Grand Challenge. It provides several real clinical data from many institutions available on The Cancer Imaging Archive TCIA (Clark et al., 2013). The dataset is originally composed of 3D images downsampled to 128 × 128 × 128 voxel tensors. But in this work it is modified into a 2D (128 x 128) dataset divided into train (7800 samples), validation (1200samples) and test (1200samples). The folder of each sample contains a CT scan of the patient, 10 structural masks which are binary masks of the 10 organs involved in the treatment, possible dose mask that is binary mask of where the irradiation is allowed and ground-truth dose which is the desired output.

3. Data preparation

Data preparation is a fundamental step in the workflow of network training. It may be more important than choosing the model itself for accurate predictions. We perform multistep preprocessing. First, we performed data normalization to CT scan input data for each patient. And finally, we stacked the data of each patient to create a 12-channel 2D positional information tensor of 128 × 128 as inputs to the network. The input data for each patient were arranged in a tensor of 12 channels as follows:

CT scan: 128×128, Structural masks: $10 \times 128 \times 128$, Possible dose masks: 128×128.

This unique representation of the contours in distinct channels helps avoid the possibility of structures overlapping if they are all represented in a single channel due to their coarse size.

4. Architecture and methodological components

4.1. Deep Learning Terminology

In this section, we will define some of the deep learning terminology used in this paper.

- **Convolutional neural networks** (CNN) were first proposed by (Lecun et al., 1998), and quickly found their use in deep learning for computer vision and imaging tasks. By using kernels and convolution, CNN can easily extract image features, such as edges. Additionally, CNN feature extraction is shift-invariant and uses overall fewer weights than fully connected networks. Each convolution layer of a CNN calculates a set of feature maps from the input or a set of feature maps from the previous layer.

- **Pooling Layer** Pooling in deep learning refers to dividing a feature map into rectangular patches, and then aggregating the pixel information in each patch to create a new, lower-resolution layer that retains essential features of the high-resolution map. Typically, the patches are 2 x 2 for 2D with a stride of 2, which effectively halves each dimension on the output feature map. This greatly reduces the computational expense and helps the network see the image more globally using a standard-size convolution kernel. Max pooling (Nagi et al., 2011), where the single largest pixel value is carried over from each rectangular patch, became one of the most popular pooling methods to use with CNNs. There is no direct inverse operation to max pooling, but some common techniques to increase the resolution include upsampling and deconvolution.

- **U-shaped Network** are a type of end-to-end convolutional neural network that is popular for medical image segmentation. The original U-net was proposed by (Ronneberger et al., 2015). The U-net architecture has since then outperformed the prior best method in 2D segmentation in several different biomedical applications. The application of U-shaped networks has during the last years extended to the area of predicting spatial dose as well (Nguyen et al., 2017)(Nguyen et al., 2019). A Typical U-shaped network consists of a contracting part and an expanding part to generate output with the correct resolution. Each level in the network consists of convolutional layers. In-between levels in the contracting and the expanding path it is standard to use skip connections to preserve fine feature information that otherwise is lost during down-sampling. The information that is passed through is concatenated with the data in the expanding path.

4.2. Network architecture

In our work, we extended the standard 2D U-net architecture (Ronneberger et al., 2015) to a version for dose distribution prediction to account for the dependence on the anatomical geometry in the adjacent regions of the given anatomy. For our baseline model we adapt the architecture described in (Ahn et al., 2021) to fit the shape of our input data.

Our network architecture consists of four multiscale hierarchical levels made of a series of 2D convolutional layers.

The encoder part contains two convolutional layers(3x3 kernel size) at each hierarchy level to extract a set of low to high-level features. In addition to feature extraction, these convolutional layers also find hierarchical representations over a wide receptive field of the CT image and contour structures input data. Each convolutional layer was followed by a rectified linear unit (ReLu) and a maximum pooling layer.

On the decoding path, we used a 2×2 transposed convolution and two 3×3 convolution layers followed by a ReLu activation function. Concatenation was performed with the corresponding feature map from the skip connection path and two convolution layers with 3×3 filters. To avoid overfitting during training, batch normalization 3 was added to the layers. In the final layer, we used a 1×1 convolution network with a relu activation function.

5. Model tuning and comparison

To optimize the performance of the model, our pipeline included some crucial steps including the preprocessing of the data, the adjustment of the architecture and the tuning of the hyperparameters. An Ablation study was also performed Ablation studies can be used to determine which features or components are essential.

5.1. Additional models tested

We first implemented the basic UNet architecture described in (Nguyen et al., 2017), where the encoder and decoder networks have a symmetric structure, and the feature maps are passed directly from the encoder to the corresponding decoder layer through skip connections. However its performance was not satisfactory. It may indeed lose some information during the encoding process, as the feature maps are passed from the encoder to the corresponding decoder layer without any compression. This can affect the model's ability to capture fine-grained details in the input image. It also turned out that it has a reduced ability to capture complex features due to its larger number of parameters. It besides requires more computation and memory resources to process the input images. Thus because of the low accuracy of this model, its slow convergence and its reduced efficiency, we opted for bottleneck architecture where the encoder network is compressed into a bottleneck layer reducing the dimensionality of the feature maps before they are passed to the decoder network. It effectively offers improved generalization, faster training and it helps the model to extract more informative and compact feature representations, which can lead to better image generation quality.

5.2. Preprocessing

In order to improve the robustness of the model, we normalized the input CT images. Normalization is a standard step that refers to scaling pixel values so that they fall within a certain range. Here, the CT data is normalized to [0, 1], and all the other inputs are 0 or 1 because they are binary masks. Normalization enhances the performance as it can reduce the impact of variations in lighting, color balance, and other factors that can affect the appearance of images. And indeed there is a significant difference between results with and without normalization : the MAE on validation set is around 0.5 before normalization and around 0.4 after.

5.3. Hyperparameters tnuning

By searching over a range of hyperparameters values, one can find a set of hyperparameters that achieve good performance without overfitting the data or requiring excessive computational resources. We tested different values for the learning rate, number of epochs and the batch size and we opted for a learning rate of 2x1e-4, a batch size of 8 and 60 epochs.

5.4. Ablation study

In order to understand the contribution of specific components of the model to its overall performance, we used an ablation study consisting in removing or modifying components to evaluate how these changes affect the model. First, results from ablation study on dropout layers have shown that higher dropout rates are more effective in preventing overfitting with a high number of epochs, but it can also harm the model's performance, as it may cause the network to lose important information. Ablating the dropout layer in our model allows to have slightly better performance when we use 60 epochs which is the minimum number of epochs at which the validation loss is the lowest. We also tested different depths of the UNet by modifiying the number of donwsampling and upsampling layers. We found that increasing the depth of the U-Net by adding two layers to the encoder and and two to the decoder improves significantly its ability to capture complex features and patterns in the input images and hence results in better performance. And as we explained before, adding the bottleneck layer is an important modification that substantially changes the results. We used scheduler which is a technique used to adjust the learning rate during training and we noticed that it has no effect on the final performance. Finally we studied the impact of the activation function. To do this we tested GELU (Gaussian Error Linear Units) and ReLU (Rectified Linear Units) and GELU has been shown to perform better than ReLU in our case.

In This table, all the experiments were performed with the same preprocessing, learning rate, batch size and number of epochs. They are all done with ReLU activation except the last one, and they are all without dropout except the first one.

5.5. Internal validation procedure

In order to quantify the quality of the generated doses in addition to the loss value obtained for the validation set and the visual inspection, we used specific metrics used to compare images: Mean Absolute Error (MAE), the Peak Signal-to-Noise Ratio (PSNR) and the

Ablated components	Validation MAE score	Test MAE score given by Codalab
with dropout(p=0.15)	0.41	0.35
without dropout	0.39	0.34
3 donwsampling 3 upsampling layers	0.61	–
5 donwsampling 5 upsampling layers	0.41	0.35
with/witout scheduler	0.41	–
ReLU activation	0.41	0.35
GELU activation	0.39	0.34

Table 1: Summary of ablation study results

Structural Similarity index (SSIM). We have therefore defined functions that return the values of these different metrics for each trained model. The value obtained by MAE is very close to the score given by the Codalab on the test set.

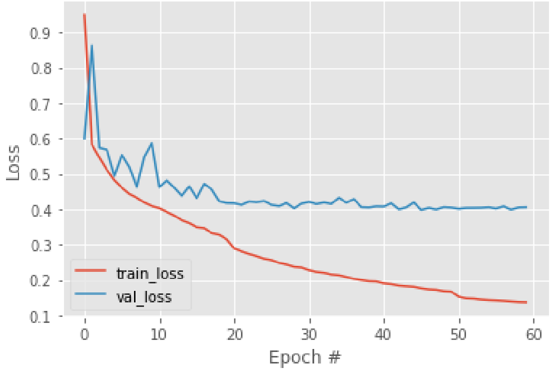

Figure 1: Loss function for train and validation sets using GELU activation, without dropout, with lr=2x1e-4

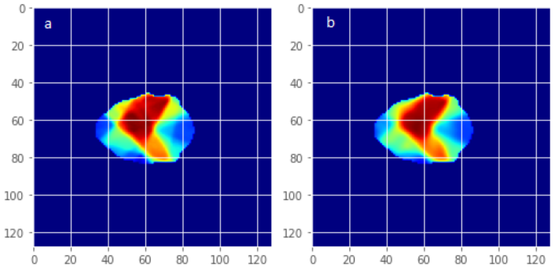

Figure 2: a. example of real dose b. corresponding generated dose

References

Sang Hee Ahn, EunSook Kim, Chankyu Kim, Wonjoong Cheon, Myeongsoo Kim, Se Lee, Young Lim, Hong Sup Kim, Dongho Shin, Dae Kim, and Hwi Jeong. Deep learning method for prediction of patient-specific dose distribution in breast cancer. *Radiation Oncology*, 16, 08 2021. doi: 10.1186/s13014-021-01864-9.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and F. Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of digital imaging*, 26, 07 2013. doi: 10.1007/s10278-013-9622-7.

Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 12 1998. doi: 10.1109/5.726791.

Jawad Nagi, Frederick Ducatelle, Gianni A. Di Caro, Dan Cireşan, Ueli Meier, Alessandro Giusti, Farrukh Nagi, Jürgen Schmidhuber, and Luca Maria Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 342–347, 2011. doi: 10.1109/ICSIPA.2011.6144164.

Dan Nguyen, Troy Long, Xun Jia, Weiguo Lu, Xuejun Gu, Zohaib Iqbal, and Steve B. Jiang. Dose prediction with u-net: A feasibility study for predicting dose distributions from contours using deep learning on prostate imrt patients. *ArXiv*, abs/1709.09233, 2017.

Dan Nguyen, Rafe McBeth, Azar Sadeghnejad Barkousaraie, Gyanendra Bohara, Chenyang Shen, Xun Jia, and Steve Jiang. Incorporating human and learned domain knowledge into training deep neural networks: A differentiable dose-volume histogram and adversarial inspired framework for generating pareto optimal dose distributions in radiation therapy. *Medical Physics*, 47(3):837–849, dec 2019. doi: 10.1002/mp.13955. URL https://doi.org/10.1002%2Fmp.13955.

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