ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Single-pass uncertainty estimation with layer ensembling for regression: application to proton therapy dose prediction for head and neck cancer

Robin Tilman[1], Margerie Huet-Dastarac[2],
Ana Maria Barragan-Montero[2], and John A. Lee[1,2] *

1- UCLouvain.be - ICTEAM
Place du Levant 3, 1348 Louvain-la-Neuve - Belgium

2- UCLouvain.be - IREC/MIRO
Avenue Hippocrate 55 B1.54.07, 1200 Brussels - Belgium

Abstract. We developed a new uncertainty quantification method for deep learning regression models, based on Layer Ensembles [1], which is competitive with state-of-the-art ensembling and Monte Carlo (MC) dropout techniques. The method was implemented in a UNet-like architecture and applied to predicting 3D dose maps for head and neck cancer patients who are treated with proton therapy. The new approach runs approximately 8 times faster than MC Dropout. Our statistical analysis showed no significant difference in prediction accuracy between the two different methods (p-value = 0.09). Moreover, the correlation uncertainty/error in the body is only -3%. These findings demonstrate the potential of the new method in enabling fast and accurate uncertainty quantification for regression problems and, in particular, for proton therapy dose prediction.

1 Introduction

U-Net architectures have showcased promising results for predicting 3D dose distributions in radiotherapy treatments [2]. However, quantifying prediction uncertainty remains a crucial aspect yet to be further explored to achieve safe and effective treatment planning.

MC Dropout [3] and Deep Ensembling [4] are two widely used methods to estimate a model's prediction uncertainty, without being too computationally intensive as Bayesian Neural Networks (BNNs) [5]. The former involves applying dropout at test time and making multiple (different) predictions for the same input, which results in an approximation of BNNs. The variance of the predictions - computed voxelwise - is considered as a proxy of the uncertainty of the model and gives a 3D uncertainty map. Deep Ensembling works in a similar manner, by training multiple versions of the same model with different subsets of data, and then computing their prediction variance as the uncertainty. Despite their popularity, MC Dropout and Ensembling have a big drawback: the need of performing several-passes at inference time to later compute the variance. This is time consuming and hampers the implementation of these methods for clinical

*M. H.-D. and A.M. B.-M. are funded by the Walloon region (PROTHERWAL/CHARP, grant 7289). J.A. L. is a Senior Research Associate with the F.R.S.-FNRS.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

situations with tight schedules, like adaptive protontherapy, where the patient is waiting in the treatment room.

In this work, we describe a new and fast method of quantifying the uncertainty of regression problems with a single inference pass, using a 3D U-Net. We developed a new method based on Layer Ensembles [1], introduced by Kushibar et al., to generate a 3D voxelwise uncertainty map for predicted proton therapy dose distributions in head and neck cancer patients. Regarding training time, the new method is significantly faster than Deep Ensembling as it involves only a single model. Regarding inference time, the 3D uncertainty map is directly computed in single pass, which makes it competitive against MC Dropout. Our approach is compared with these two commonly used methods.

2 Material and methods

Patient dataset. We used a data set of 60 patients with head and neck cancer to train 11 versions of a dose prediction model. Each patient anatomy is described by a CT scan as well as multiple masks indicating the location of organs at risk (OARs) and prescribed doses on clinical target volumes (CTV). The dataset was randomly divided into 11 groups of 47 training, 5 validation, and 5 test patients, with each test set exclusive to its corresponding model. Moreover, 3 patients were kept out of each set to assess the consistency of predictions across all 11 models.

Architecture of the prediction model. The prediction model used consists in a 3D U-Net with both long and short residual skip-connections. The model takes as input 16 channels which describe the anatomy of the patient and outputs a single-channel dose prediction. We evaluated variations of this architecture. Specifically, we assessed the effectiveness of incorporating attention gates in the decoder [6], as well as Project-and-Excite (PE) modules [7] after each convolution block. We trained all models for 200 epochs, with a dropout rate of 0.3, using a learning rate of 10^{-4}.

Layer Ensembles. Our implementation of Layer Ensembles is based on the architecture detailed in [1]: a head is attached after every decoder block in the 3D U-Net, described in the previous paragraph with the input of a head being the output of the corresponding decoder block (Figure 1). Each head consists of a 3D convolution module followed by an upsampling operation and a ReLU activation function. The output of each head has the same shape as the final dose prediction and can be consider as an approximation of it. The uncertainty is obtained by computing the variance of the outputs from the different heads as well as the final output of the model. The intuition behind this approach is as follows: hard to predict samples must have been through more decoder blocks in order for a head to predict an accurate dose [1]. As a result, the outputs of the heads will vary significantly from block to block, leading to increased uncertainty.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Prediction error. The quality of a radiotherapy dose is typically assessed with dose-volume metrics, as the mean (D_{mean}) and D_k (e.g., D_2, D_{95}, or D_{99}) for specific organs at risk (OAR) and target volumes (CTV), where D_k is dose received by the x % of the OAR/CTV volume. Thus, we evaluate the quality of a predicted dose by computing the compound absolute error in clinically relevant dose-volume metrics for the considered OARs and CTVs:

$$E_{\text{compound}} = \sum_{j=0}^{OARs,CTVs} |E_{j,k}|$$

being $E_{j,k}$ the error in D_k metric for a given, j, OAR or CTV (e.g. $E_{CTV,95} = D_{CTV,95} - \hat{D}_{\text{CTV},95}$, where $D_{\text{CTV},95}$ and $\hat{D}_{CTV,95}$ are the metric values for the ground truth and predicted doses, respectively. Note that for MC Dropout and Deep Ensembling, the prediction is the average of the different models, while for LayerEnsembles there is only a single predicted dose.

Prediction uncertainty. Let U_{map}, the 3D dose uncertainty map obtained by taking the voxelwise standard deviation of the different dose predictions in the body. Let E_{map}, the error map representing the voxelwise MAE between the predicted dose and the actual dose in the body. The quality of U_{map} is determined by the correlation ρ_{map} between U_{map} and E_{map}, which should be as close as possible to 1.

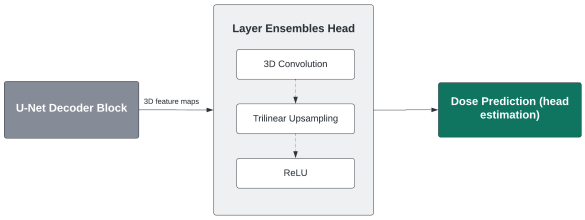

Fig. 1: Layer Ensembles process

3 Results and discussion

For each uncertainty estimation method described in Section 2 we shall evaluate the quality of the uncertainty quantification, the accuracy of the predictions and the inference time. The objective is to achieve swift and accurate predictions across all test patients, while simultaneously maximizing the correlation between error and uncertainty.

3.1 Layer Ensembles:

We trained the Layer Ensembles model using a different loss function from the one used to train a standalone U-Net. We differentiate between two distinct

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

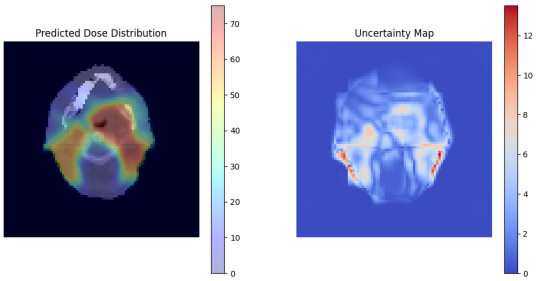

Fig. 2: Slice of 3D dose prediction and uncertainty map obtained with the best Layer Ensembling model

losses: $loss_{\text{base}}$, which computes the average loss of the predictions from each individual head and the final prediction, and $loss_{\text{modified}}$ which gives more weight (50% of the total loss) to the final model prediction. The usage of the two losses are analyzed for each version of Layer Ensembles trained. The heads and the model are jointly trained for 200 epochs, then the heads alone are fine-tuned for 200 more epochs.

We analyzed multiple U-Net architectures trained with 3 or 4 heads attached to 3 or 4 decoder blocks[1]: a base version of the 3D U-Net, as described in Section 2, another version with PE modules added after each convolution block of the U-Net, and a 3D Attention U-Net with PE modules.

The model with the lowest mean E_{compound}, 20.77 Gy ($\pm$ 19.78), uses the architecture with 3 heads, PE and attention modules, and is trained with $loss_{modified}$. The baseline Layer Ensembles model with 3 heads, trained with $loss_{modified}$, has an average error of 21.87 Gy ($\pm$ 16.62).

3.2 Comparison with Ensembling and MC Dropout

We compared the results obtained with Layer Ensembles to the performance of Deep Ensembling and MC Dropout. For MC Dropout, we predicted 20 different doses for each patient. We tried different test dropout rates between 0.1 and 0.5 for MC Dropout. A rate of 0.3 gave a good tradeoff between the quality of the uncertainty quantification and the final prediction accuracy. For Deep Ensembling, we trained 10 different 3D U-Net with PE blocks to predict 10 different doses for each test patient. Previous literature supports the use of a smaller number of ensembling models, where they found that 5 ensembling models gave similar results that scaling up to 50 models [8].

Figure 3 compares the average E_{compound} for the different uncertainty quantification methods. We conducted 5 Wilcoxon tests, with Benjamini-Hochberg correction [9], to compare Deep Ensembling and MC Dropout (rate = 0.3) with

[1]We also tried other versions of Layer Ensembles with heads attached to the encoder or the bottleneck, but observed a poor uncertainty estimation.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

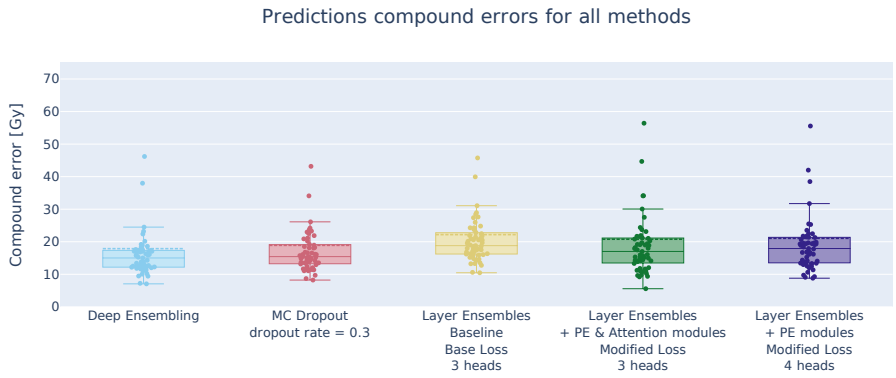

Fig. 3: Predictions compound errors for different methods

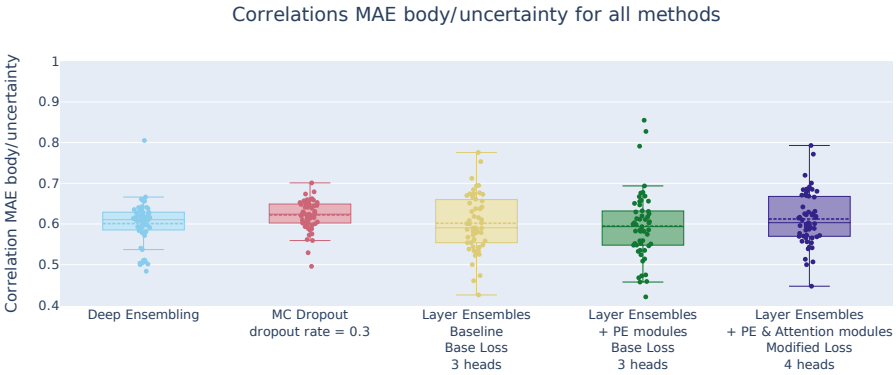

Fig. 4: Correlation uncertainty/MAE in body for different methods

the Layer Ensembles model that has the lowest average E_{compound} (in green on Figure 3), as well as with the baseline Layer Ensembles (in yellow). The 5 tests gave the following results: Deep Ensembling performs significantly better than MC Dropout in CTVs and OARs ($p = 0.048$). The best Layer Ensembles model exhibits no significant error difference compared to MC Dropout (p-value $= 0.09$), but underperforms compared to Deep Ensembling (p-value $= 0.0014$). Finally, the baseline Layer Ensembles model is significantly less accurate than MC Dropout (p-value $< 10^{-4}$) and Deep Ensembling (p-value $< 10^{-6}$).

Figure 4 reports the values of ρ_{map} the correlation between U_{map} and E_{map}. Deep Ensembling does not perform significantly better than the Layer Ensembles model with attention modules depicted on the graph ($p = 0.9$). The average ρ_{map} are 0.6 for both methods. Nonetheless, with a 0.62 for ρ_{map}, MC Dropout outperforms Layer Ensembles ($p = 0.03$) and Deep Ensembling ($p = 0.002$).

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

We used an A100-PCIE-40GB GPU to measure the time required to predict dose and quantify uncertainty for 55 test patients. Our experiments revealed a significant difference in inference time between the three methods, with Layer Ensembles being the fastest (00:17:26), followed by Deep Ensembling (01:08:56) and MC Dropout being the slowest (02:25:50). Notice that Deep Ensembling inference time is shorter than for MC Dropout, but the method requires multiple models to be trained at first.

4 Conclusion and perspectives

We used a new uncertainty quantification method based on Layer Ensembles [1] to quantify uncertainty of proton therapy dose predictions. Because it only requires a single pass of the data and a unique model, our approach has the advantage of being significantly faster than state-of-the-art methods used in practical applications to predict an accurate dose and its uncertainty. Our model can provide a 3D map of the uncertainty, which might help make decisions and interpret the prediction of the algorithm, as well as its reliability. The new method can also be used to perform out-of-distribution detection, e.g., regions with large uncertainty might indicate the presence of noise or an anomaly in the input CT scan of the patient. Finally, Layer Ensembles could be used for Active Learning tasks for unannotated new samples of proton therapy dose.

References

[1] Kaisar Kushibar, Victor Campello, Lidia Garrucho, Akis Linardos, Petia Radeva, and Karim Lekadir. Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation. *Lecture Notes in Computer Science*, pages 514–524, 2022.

[2] Dan Nguyen et al. 3d radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture. *Physics in Medicine & Biology*, 64(6):065020, 2019.

[3] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, Oct 2016.

[4] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, Nov 2017.

[5] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, and et al. A survey of uncertainty in deep neural networks, Jan 2022.

[6] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, and et al. Attention u-net: Learning where to look for the pancreas, May 2018.

[7] Anne-Marie Rickmann, Abhijit Guha Roy, Ignacio Sarasua, Nassir Navab, and Christian Wachinger. 'project & excite' modules for segmentation of volumetric medical scans, Jun 2019.

[8] Yaniv Ovadia et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

[9] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.