Quality control in radiotherapy-treatment planning using multi-task learning and uncertainty estimation

Felix J.S. Bragman[1*] **Ryutaro Tanno**[1] **Zach Eaton-Rosen**[1] **Wenqi Li**[1] **David J. Hawkes**[1]

Sébastien Ourselin[2] **Daniel C. Alexander**[1,3] **Jamie R. McClelland**[1] **M. Jorge Cardoso**[2,1]

1. Centre for Medical Image Computing, University College London
2. Biomedical Engineering and Imaging Sciences, King's College London
3. Clinical Imaging Research Centre, National University of Singapore
*f.bragman@ucl.ac.uk

Abstract

Multi-task learning is ideally suited for MR-only radiotherapy planning as it can jointly simulate a synthetic CT (synCT) scan - a regression task - and an automated contour of organs-at-risk - a segmentation task - from MRI data. We propose to use a probabilistic deep-learning model to estimate respectively the *intrinsic* and *parameter* uncertainty. *Intrinsic* uncertainty is estimated through a heteroscedastic noise model whilst *parameter* uncertainty is modelled using approximate Bayesian inference. This provides a mechanism for data-driven adaptation of task losses on a voxel-wise basis and importantly, a measure of uncertainty over the prediction of both tasks. We achieve state-of-the-art performance in the regression and segmentation of prostate cancer scans. We show that automated estimates of uncertainty correlate strongly in areas prone to errors across both tasks, which can be used as mechanism for quality control in radiotherapy treatment planning.

1 Introduction

MR-only radiotherapy treatment (RT) planning has been recently proposed to mitigate issues with conventional treatment planning, that requires both the acquisition of an MR (magnetic resonance) scan and CT (computed tomography) scan. MR-only RT planning involves the simulation of a synthetic CT (synCT) scan from an MR scan and automated contouring of OARs. Convolutional neural networks (CNNs) have been applied to CT synthesis [6, 8]. They have employed conditional generative adversarial networks to capture fine texture details [6] with extensions using CycleGAN to leverage the abundance of unpaired training data [8]. These methods commit to a single prediction with no measure of confidence in the synCT and do not jointly learn the OAR segmentation, leading to anatomically inconsistent predictions between the synCT and the segmentation organs. If the predictive uncertainty of the model were known, this information could be leveraged to: 1) stochastically sample synCTs for probabilistic dose delivery estimation, 2) provide a basis for automated quality control and 3) assess when more training data is needed to decrease parameter uncertainty.

2 Methods

Dual-task architecture. We propose a probabilistic dual-task CNN-based algorithm which operates on a MR image and simultaneously provides four valuable outputs necessary for probabilistic radiotherapy planning: (1) the synCT; (2) the OAR segmentations and (3) quantification of predictive uncertainty in (1) and (2). Our work makes use of [7] to enrich the probabilistic multi-task learning method proposed in [3], enabling modelling of the spatial variation of intrinsic uncertainty via

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

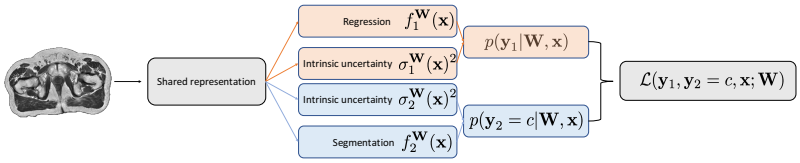


Figure 1: Multi-task learning architecture.

heteroscedastic noise, and integrating parameter uncertainty via dropout. The proposed architecture integrates the methods of uncertainty modeling in CNNs [4, 7] into a multi-task learning framework with hard-parameter sharing, in which the initial few layers of the network are shared across the two tasks (synthesis and segmentation) and branch out into task-specific layers (Fig.1). For each input patch $\mathbf{x}$, our dual-task model estimates the conditional distributions $p(\mathbf{y}_i|\mathbf{x})$ for tasks $i = 1, 2$ where $\mathbf{y}_1$ and $\mathbf{y}_2$ denote the Hounsfield Unit (HU) and class probabilities of OARs at the center of the input patch. At inference, the probability maps over the synCT and OARs are obtained by stitching together outputs from appropriately shifted versions of the input patches.

Task weighting with heteroscedastic uncertainty. We adapt the *heteroscedastic* (data-dependent) noise model to multi-task learning. For the CT synthesis (task $i = 1$), we define our likelihood as a normal distribution $p\left(\mathbf{y}_1|\mathbf{W},\mathbf{x}\right) = \mathcal{N}(f_1^{\mathbf{W}}(\mathbf{x}), \sigma_1^{\mathbf{W}}(\mathbf{x})^2)$. For the segmentation (task $i = 2$), we define the classification likelihood as softmax function of scaled logits i.e. $p\left(\mathbf{y}_2|\mathbf{W},\mathbf{x}\right) = \text{Softmax}\left(f_2^{\mathbf{W}}(\mathbf{x})/2\sigma_2^{\mathbf{W}}(\mathbf{x})^2\right)$. The NLL loss for the dual-task network is consequently derived as

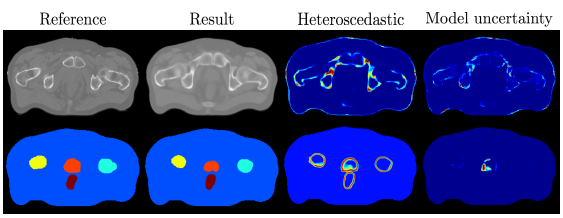
$$\mathcal{L}(\mathbf{y}_1, \mathbf{y}_2 = c, \mathbf{x}; \mathbf{W}) = \frac{||\mathbf{y}_1 - f_1^{\mathbf{W}}(\mathbf{x})||^2}{2\sigma_1^{\mathbf{W}}(\mathbf{x})^2} + \frac{\text{CE}(f_2^{\mathbf{W}}(\mathbf{x}), \mathbf{y}_2 = c)}{2\sigma_2^{\mathbf{W}}(\mathbf{x})^2} + \log\left(\sigma_1^{\mathbf{W}}(\mathbf{x})^2\sigma_2^{\mathbf{W}}(\mathbf{x})^2\right)$$

where the MSE and CE terms are weighted by the inverse of heteroscedastic intrinsic uncertainty terms $\sigma_i^{\mathbf{W}}(\mathbf{x})^2$, that enables automatic weighting of task losses on a per-sample basis.

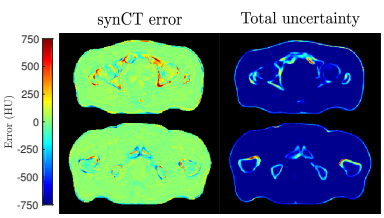
Parameter uncertainty with approximate Bayesian inference. In this work, we use dropout in our model to assess the benefit of modelling parameter uncertainty in the context of our multitask learning problem. During training, for each input, network weights are drawn from the approximate posterior $w' \sim q(\mathbf{W})$ to obtain the multi-task output (predictive mean and predictive variance), $\mathbf{f}^{w'}(\mathbf{x}) := [f_1^{w'}(\mathbf{x}), f_2^{w'}(\mathbf{x}), \sigma_1^{w'}(\mathbf{x})^2, \sigma_2^{w'}(\mathbf{x})^2]$. At test time, for the input $\mathbf{x}$, we collect output samples $\{\mathbf{f}^{w^{(t)}}(\mathbf{x})\}_{t=1}^{\text{T}}$ by performing T stochastic forward-passes with $\{w^{(t)}\}_{t=1}^{\text{T}} \sim q(\mathbf{W})$. For the regression, we calculate the expectation over the T samples in addition to the variance, which is the *parameter* uncertainty. For the segmentation, we compute the expectation of class probabilities to obtain the final labels. The *parameter* uncertainty in the segmentation is obtained by considering variance of the stochastic class probabilities. The final predictive uncertainty is the sum of the *intrinsic* and *parameter* uncertainties. In this paper, we study the utility of the predictive uncertainty as a mechanism for automated quality control of RT treatment planning.

3 Results and discussion

Training. We trained our network on 15 prostate cancer patients using 3-fold cross-validation. Prior to training, the CT scans were spatially aligned with the T2 scans [1]. We trained our model on randomly selected 2D axial slices, reconstructing the 3D volume at test time. The representation network was an adapted version of HighResNet [5] with the following features per layer



(a) Model output.



(b) synCT error and regression uncertainty.

Figure 2: Correlation between uncertainty and model output.

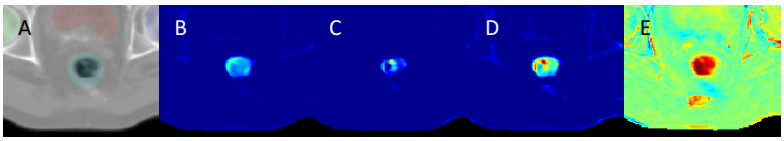


Figure 3: Quality control of patient with variations in anatomy. a) synCT with overlaid reference segmentation, b) heteroscedastic noise, c) parameter uncertainty, d) total uncertainty and e) z-score.

$f_R = [64, 64, 128, 256, 2048]$. Each task-specific branch was a set of 5 convolutional layers of size $[256_{l=1,2,3,4}, n_{i,l=5}]$ where $n_{i,l=5}$ is 1 for regression and $\boldsymbol{\sigma}$ whilst equal to the number of segmentation classes. The first two layers were 3×3 kernels and the final convolutional layers were fully connected. The model was implemented in NiftyNet [2]. We minimised the loss using ADAM with a learning rate of 10^{-3}, converging at 17,500 iterations and training until 19,000. At test time, we sampled 10 stochastic samples at 18,000 and 19,000 iterations leading to $T = 20$ stochastic samples.

Model performance. We illustrate the output of the framework for a patient in Fig.2a, displaying its ability to generate a synCT, OAR segmentations and measures of uncertainty across both tasks. Our method performed better than the state-of-the-art in pelvic CT synthesis, tested on the same dataset [1] and measured by the mean absolute error (HU) 43.3 ± 2.9 compared to 45.7 ± 4.6. We also achieved equivalent segmentation performance compared to [1] achieving fuzzy DICE scores of 0.91 ± 0.02, 0.92 ± 0.02, 0.70 ± 0.06, 0.74 ± 0.12 and 0.93 ± 0.04 versus 0.89 ± 0.02, 0.90 ± 0.01, 0.73 ± 0.06, 0.77 ± 0.06 and 0.90 ± 0.03 at the left and right femur, prostate, rectum and bladder[1].

Uncertainty for quality control. We observe that uncertainty correlates in regions prone to errors, notably close to organ boundaries and near the bone Fig. 2a and 2b. By modelling heteroscedastic and parameter uncertainty, we are able to disentangle the origin of uncertainty in the predictions with a view towards assessing the clinical utility of the model. Heteroscedastic uncertainty is data-dependent, representing regions in a scan that are intrinsically likely to be uncertain. However, parameter uncertainty relates to model ambiguity, resulting from a lack of variability in training data necessary to generalise to unseen examples. In safety critical applications such as RT planning, this facilitates analysing cases where uncertainty stems from sparsity in the training data, which may lead towards the decision to acquire more training data before clinical deployment. This issue is seen in Fig.3, which demonstrates a case where variations in rectum filling across the CT and T2 scan lead to isolated regions of high uncertainty and error in the regression. In these situations, the uncertainty can potentially be used as mechanism for further intervention by the clinical team. Uncertainty can thus be potentially exploited for automated quality control in the absence of ground-truth with an uncertainty-based traffic-light system in RT planning. This will be investigated in future work.

References

[1] N. Burgos et al. Iterative framework for the joint segmentation and CT synthesis of MR images: application to MRI-only radiotherapy treatment planning. *Phys. Med. Biol.*, 62, 2017.

[2] E. Gibson et al. Niftynet: a deep-learning platform for medical imaging. *Comput. Methods Programs Biomed.*, 158, 2018.

[3] A. Kendall et al. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018.

[4] A Kendall and Y Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, pages 5580–5590, 2017.

[5] W. Li et al. On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task. In *IPMI*, pages 348–360, 2017.

[6] D. Nie et al. Medical image synthesis with context-aware generative adversarial networks. In *MICCAI Part III*, pages 417–425, 2017.

[7] R. Tanno et al. Bayesian image quality transfer with cnns: Exploring uncertainty in dmri super-resolution. In *MICCAI Part I*, pages 611–619, 2017.

[8] J. Wolterink et al. Deep mr to ct synthesis using unpaired data. In *SASHIMI*, pages 14–22, 2017.

[1] T1 and T2-weighted scans were used in [1] in contrast to our work, which only used T2-weighted scans.