Uncertainty Quantification in End-to-End Implicit Neural Representations for Medical Imaging

Francisca Vasconcelos∗
francisca.vasconcelos@keble.ox.ac.uk

Bobby He∗
bobby.he@stats.ox.ac.uk

Yee Whye Teh
y.w.teh@stats.ox.ac.uk
Department of Statistics
University of Oxford
Oxford, UK OX13LB

Abstract

Implicit neural representations (INRs) have recently achieved impressive results in image representation. This work explores the uncertainty quantification quality of INRs for medical imaging. We propose the first uncertainty aware, end-to-end INR architecture for computed tomography (CT) image reconstruction. Four established neural network uncertainty quantification techniques – deep ensembles, Monte Carlo dropout, Bayes-by-backpropagation, and Hamiltonian Monte Carlo – are implemented and assessed according to both image reconstruction quality and model calibration. We find that these INRs outperform traditional medical image reconstruction algorithms according to predictive accuracy; deep ensembles of Monte Carlo dropout base-learners achieve the best image reconstruction and model calibration among the techniques tested; activation function and random Fourier feature embedding frequency have large effects on model performance; and Bayes-by-backpropagation is ill-suited for sampling from the INR posterior distributions. Preliminary results further indicate that, with adequate tuning, Hamiltonian Monte Carlo may outperform Monte Carlo dropout deep ensembles.

1 Introduction

In computed tomography (CT), improving reconstructed image quality via increased measurement also increases patient exposure to harmful radiation [1, 2, 3]. As a result, there is interest in reconstruction techniques which achieve high image quality from few measurements. Machine learning approaches based on deep learning have proved promising in this regard. However, they require large training data sets, which are difficult to collect in the medical setting. A significant recent advance was the introduction of implicit neural representations (INRs), which represent complex coordinate-based signals as functions encoded by small neural networks. For example, an image can be represented as a function mapping \((x, y)\) coordinates to \((r, g, b)\) pixel intensities. INRs have taken the field of computer graphics by storm, achieving impressive results in novel view synthesis [4–7], shape representation [8–10, 11–12, 13], and texture synthesis [14–15]. More recent work has also demonstrated the applicability of this technique to medical imaging [16, 17].

In all these applications, INRs were assessed on their predictive accuracy and reconstructed signal plausibility. However, in medical imaging, which affects doctor decisions and patient well-being, it is also important to understand model confidence in the reconstructed image. For example, a model

∗Denotes equal author contribution. Order decided by rock-paper-scissors (best of 3).
can quantify its uncertainty in each pixel by outputting a per-pixel variance. If this variance is large in critical image regions, such as the location of a potential tumor, a doctor should order additional measurements to ensure proper diagnosis. Uncertainty quantification can also be used to decrease healthcare cost via automated triage, e.g. by assigning images with varying degrees of uncertainty to healthcare providers of relevant expertise. Finally, understanding model uncertainty could inform more efficient measurement procedures, leveraging techniques such as active learning [18].

2 Methods
Encoding 2D CT image reconstruction in INRs
As shown in Figure 1, CT measurement data comes in the form of a sinogram, \( p_\theta(r) \), where \( r \) is the X-ray detector location and \( \theta \) is the measurement angle. However, the goal of reconstruction is to generate a photo of the 2D cross-section of attenuation coefficients, \( f(x,y) \). This is achieved using an end-to-end approach to image reconstruction, in which our model predicts the final cross-section attenuation coefficient function, as illustrated in Figure 1. The model input is pixel coordinate \( (x,y) \) and its output is the corresponding predicted attenuation coefficient value \( \hat{f}(x,y) \). The sinogram data is incorporated in the model via the training loss function. Given a ground truth sinogram \( p_\theta(r) \), the loss of the INR output \( \hat{f}(x,y) \) is defined as

\[
L(p_\theta(r), \hat{f}(x,y)) = \frac{1}{2|\Theta \times R|} \sum_{\theta \in \Theta} \sum_{r \in R} \left( p_\theta(r) - \int_\mathcal{Y} \int_\mathcal{X} \hat{f}(x,y) \delta(x \cos \theta + y \sin \theta - r) \, dx \, dy \right)^2,
\]

where \( \Theta = \{\theta_1, ..., \theta_n\} \) are the view angles, \( R = \{r_1, ..., r_n\} \) the X-ray detector locations, and \( \mathcal{X} \times \mathcal{Y} \) the image pixels \((x,y)\). The integral surrounding \( \hat{f} \) is the Radon transform. Since the loss is calculated directly on the desired output, end-to-end training minimizes the propagation of error, but has a training complexity cost. Each training iteration requires sampling the model \(|\mathcal{X} \times \mathcal{Y}| \) times, once per image pixel, and a Radon transform must be calculated for all \(|\Theta \times R| \) points in the sinogram. However, given the relatively small nature of INRs by deep learning standards, we did not find this computationally barring, with networks taking no more than a few minutes to train.

Uncertainty quantification of INRs
We implemented and compared multiple methods for uncertainty quantification of INR parameters and predictions. Experimental details are reported in Appendix A. Bayes-by-backpropagation (BBB) [19], Monte Carlo dropout (MCD) [20], and Hamiltonian Monte Carlo (HMC) [21, 22, 23] were used for approximate Bayesian neural network (BNN) inference, while deep ensembles of size \( N \) (DE-\( N \)) [24] were used to aggregate the results of \( N \) MCD base learners. Besides HMC, which is regarded as the gold standard inference scheme for BNNs [25], all these approximate inference schemes are computationally efficient and common choices within the uncertainty deep learning community [26]. Model performance was assessed according to peak signal-to-noise ratio (PSNR), negative log likelihood (NLL), and expected calibration error (ECE).

Baseline
As a baseline, we compared our INRs to standard medical image reconstruction algorithms: filtered back-projection (FBP), conjugate-gradient least squares (CGLS), expectation max-
while too high a frequency results in high-frequency image artifacts. Consistent with previous work, the RFF frequency must be consistent with the amount of data used in training the INR. Too low a frequency prohibits higher frequency learning, while too high a frequency results in high-frequency image artifacts. Consistent with previous work, we further found that ensembling MCD architectures can improve image reconstruction quality and model calibration, achieving significant improvements even for small numbers of base learners. In all, DEs of the top 5 or 10 performing MCD models achieved the best results in terms of image reconstruction and calibration. In future work, to better separate the influence of inference method from INR prior choice, we will further explore the performance of tuned HMC for INRs.

While this work is the first use of uncertainty quantification for INRs, it is not the first proposal of INRs for CT image reconstruction. The CoIL architecture [16] was recently demonstrated to improve

### Table 1: INR accuracy and calibration results of all four uncertainty quantification approaches are presented. Classical approaches do not produce uncertainty estimates. For BBB, MCD, and DEs results are averaged over all validation and all test set images.

<table>
<thead>
<tr>
<th># View</th>
<th>Recon</th>
<th>Validation Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PSNR</td>
<td>NLL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FBP</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGLS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SART</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>HMC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BBB</td>
<td>-1.090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCD</td>
<td>-1.701</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-2</td>
<td>-1.737</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-5</td>
<td>-2.226</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-10</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>HMC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BBB</td>
<td>1.050</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCD</td>
<td>0.701</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-2</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-5</td>
<td>-0.364</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-10</td>
<td></td>
</tr>
</tbody>
</table>

Data  The Shepp-Logan phantom [28] approach was used to generate artificial brain images, with corresponding sinograms generated via the Radon transform. In this preliminary work, no noise was added to the images. In all, 10 ground truth images and 20 corresponding sinograms were generated: 5 validation and 5 test set sinograms each for the 5- and 20-view (θ) cases.

### Results and discussion

Although restricted to noiseless phantom data, this work presents the first large-scale study of model parameterization for INRs with uncertainty quantification. Experimental results are presented in Table 1 sample reconstructed images, variances, coverage plots, and reliability curves in the 20-view case are presented in Figure 2 and boxplots of MCD and BBB model performance by hyperparameter are presented in Appendix A. MCD significantly outperformed BBB, with activation function substantially affecting reconstruction quality. We found the Sine activation produced top-performing MCD models, as expected from the recent SIREN work [29]. Silu, Tanh, and Relu achieved slightly lower performance, but greater consistency. We also confirmed previous findings that random Fourier feature (RFF) embeddings enable neural networks to learn high-frequency image components better [30]. However, we found that the RFF frequency must be consistent with the amount of data used in training the INR. Too low a frequency prohibits higher frequency learning, while too high a frequency results in high-frequency image artifacts. Consistent with previous work [31], we further found that ensembling MCD architectures can improve image reconstruction quality and model calibration, achieving significant improvements even for small numbers of base learners. In all, DEs of the top 5 or 10 performing MCD models achieved the best results in terms of image reconstruction and calibration. In future work, to better separate the influence of inference method from INR prior choice, we will further explore the performance of tuned HMC for INRs.

### Figure 2: The pixel-wise mean, pixel-wise variance, pixel-wise mean squared error (MSE), pixel-wise coverage, and image reliability curves are shown for BBB, MCD, and DE image reconstructions of a test set 20-view sinogram. The same is shown for HMC on a 20-view sinogram, outside the test set.
image reconstruction pipelines by learning a functional form of the measurement sinogram. However, this necessitates use of classical image reconstruction to generate the final, desired image cross-section. It also lacks support for uncertainty estimation, since the relation between the uncertainty of sinogram values and image pixels is not immediately evident. Instead, we propose an end-to-end image reconstruction pipeline where the network output is the desired image cross-section, which, as shown in this work, provides seamless support for uncertainty estimation. In future work, we aim to compare the performance of our end-to-end approach to that of the CoIL architecture & consider approaches to increase the training efficiency of the proposed end-to-end solution, as well as extending our approach to other medical imaging settings, such as 3D CT & MRI.

**Potential Negative Societal Impact**

We do not foresee many potential negative societal impacts to our work. Given the importance of robust and reliable predictions in the medical imaging domain, we believe calibrated uncertainty quantification is an important quality for any model that is to be deployed in practice, including potentially Implicit Neural Representations.

**Acknowledgements**

We would like to acknowledge Nalini Singh for her useful discussions in this work.

**References**


A Appendix - Uncertainty Experiments

For MCD and BBB, large-scale hyperparameter sweeps were performed to find optimal model parameters for: activation function (Relu, SiLU, Sine, Softplus, Tanh), model depth (3, 6, 9), model width (16, 64, 256, 1024), and random fourier feature (RFF) frequency (1, 5, 10, 15). For MCD probability of dropout (0.2, 0.5, 0.8) and for BBB prior standard deviation (10, 100, 1000) and KL factor (ξ) were additionally swept over. Hyperparameter sweep models were assessed according to average PSNR, negative log likelihood (NLL), and expected calibration error (ECE) on the validation set. These experiments were run in both 5- and 20-view cases. Boxplots of the model performance for MCD and BBB according to each hyperparameter are presented in Figures 3 and 4. The best performing model in each case was further assessed on the test set.

Since MCD outperformed BBB, shown in Table 1, MCD networks were used as base learners for the DEs. DEs of size \( N \) (DE-\( N \)) were created by ensembling the \( N \) top-performing unique MCD model architectures from the hyper-parameter sweeps, in both the 5- and 20-view case. The DEs were tested on the same validation and test sets as MCD and BBB.

HMC was implemented via Hamiltonorch, sampling from an INR of width 256, depth 3, ReLU activation, and RFF frequency 10. In this preliminary work, no hyper parameter tuning was performed, so no validation set was used. Further, given the computationally intensive nature of HMC, the network was tested only on a single Shepp-Logan phantom, outside the previously described validation and test sets. In future work, we will perform a hyperparameter search, similar to those of MCD and BBB, and report scores for the full validation and test sets.
