Interval Neural Networks as Instability Detectors for Image Reconstructions

Jan Macdonald[1], Maximilian März[1], Luis Oala[2], Wojciech Samek[2]

[1]Institut für Mathematik, Technische Universität Berlin
[2]Machine Learning Group, Fraunhofer HHI
macdonald@math.tu-berlin.de

Abstract. This work investigates the detection of instabilities that may occur when utilizing deep learning models for image reconstruction tasks. Although neural networks often empirically outperform traditional reconstruction methods, their usage for sensitive medical applications remains controversial. Indeed, in a recent series of works, it has been demonstrated that deep learning approaches are susceptible to various types of instabilities, caused for instance by adversarial noise or out-of-distribution features. It is argued that this phenomenon can be observed regardless of the underlying architecture and that there is no easy remedy. Based on this insight, the present work demonstrates, how uncertainty quantification methods can be employed as instability detectors. In particular, it is shown that the recently proposed *Interval Neural Networks* are highly effective in revealing instabilities of reconstructions. Such an ability is crucial to ensure a safe use of deep learning-based methods for medical image reconstruction.

1 Introduction

Deep learning has shown the potential to outperform traditional schemes for solving various signal recovery problems in medical imaging applications [1,2]. Typically, such tasks are modelled as finite-dimensional linear inverse problems

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \eta \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the unknown signal of interest, $\mathbf{A} \in \mathbb{R}^{m\times n}$ denotes the forward operator representing a physical measurement process, and $\eta \in \mathbb{R}^m$ is modelling noise in the measurements. Important examples include choosing $\mathbf{A}$ as a subsampled Fourier matrix (magnetic resonance imaging) or a discrete Radon transform (computed tomography). Solving the inverse problem (1) amounts to computing an approximate reconstruction of $\mathbf{x}$ from its observed measurements $\mathbf{y}$. The difficulty of this task is mainly determined by the strength of the noise and the degree of ill-posedness of (1), which is typically governed by the amount of undersampling in the measurement domain [3].

In many cases, sparse regularization provides state-of-the-art solvers for (1), which are additionally backed up by theoretical guarantees, e.g. by compressed sensing [3]. However, it has been demonstrated that data-based deep learning methods are able to outperform their traditional counterparts in terms of empirical reconstruction quality and speed; see [2] for a recent overview.

In image classification, the susceptibility of deep neural networks to adversarial exploitation is well documented [4]. Recent works have reported similar instabilities for image reconstruction tasks [5,6], which can be caused by visually imperceptible adversarial noise or features that have not been seen during training. The former can be found by solving a problem of the form

$$\underset{\mathbf{e}\in\mathbb{R}^m}{\text{maximize}}\ \|\mathrm{Rec}(\mathbf{y}+\mathbf{e})-\mathbf{x}\|_2 \quad \text{subject to} \quad \|\mathbf{e}\|_2 \leq \delta, \tag{2}$$

where $\mathrm{Rec}\colon \mathbb{R}^m \to \mathbb{R}^n$ is a solution method for (1) and $\delta > 0$ is small. In other words, given measurements $\mathbf{y}$, the goal is to find a perturbation $\mathbf{e}$ that maximizes the error of a reconstruction algorithm.

Although there has been a first attempt to alleviate these shortcomings, [6] argues that such instabilities are in fact an unavoidable price for improvements in performance over classical methods. Hence, this work is motivated by the following premise: *if instabilities occur, we want to be able to detect them.* To that end, we demonstrate the potential of the recently proposed Interval Neural Network framework [7] as an instability detector. Its superiority over two other uncertainty quantification (UQ) methods [8,9] is shown.

1.1 Overview and Contributions

We consider a straight-forward approach to solving (1), which is based on post-processing a standard model-based inversion by a neural network [1]. Thus, the reconstruction is given by

$$\mathbf{x}_{\mathrm{rec}} = \mathbf{\Phi}(\mathbf{A}^\dagger\mathbf{y}) \tag{3}$$

where $\mathbf{\Phi}\colon \mathbb{R}^n \to \mathbb{R}^n$ denotes the prediction network (trained to minimize the loss $\|\mathbf{x}-\mathbf{\Phi}(\mathbf{A}^\dagger\mathbf{y})\|_2^2$) and $\mathbf{A}^\dagger$ symbolizes the non-learned model-based inversion. This scheme is studied for solving the severely ill-posed problem of limited angle computed tomography ($\mathbf{A}$ is a subsampled Radon transform), which has applications in dental tomography, breast tomosynthesis or electron tomography. We investigate the capacity of three UQ schemes (see Sec. 2) to localize possible instabilities in the output of the prediction network $\mathbf{\Phi}$. As possible causes for such instabilities we consider: (i) adversarial noise on the input and (ii) imposed structural characteristics that have not been seen during training, i.e., out-of-distribution (OoD) features (see Sec. 3). We believe that detecting OoD-instabilities is of particular importance in the context of medical imaging, since pathological changes are typically rare events in the training data. In summary, the contributions of this work are as follows:

a) We show that UQ can be utilized to detect the lack of robustness of deep learning-based image reconstruction methods.

b) Three UQ schemes for artificial neural networks are compared with respect to their capacity of revealing reconstruction instabilities.

c) We demonstrate that one UQ approach in particular, the so called Interval Neural Network, performs best as an instability detector.

2 Materials and Methods

We briefly present three methods for UQ of neural network predictions and discuss the considered limited angle CT task.

2.1 Uncertainty Quantification Methods

In this work we consider only UQ methods that rely on the training of a single neural network and exclude computationally more costly approaches like ensemble learning or cross-validation.

Interval Neural Network The recent work [7] has shown that by using interval arithmetic a baseline network $\boldsymbol{\Phi} \colon \mathbb{R}^n \to \mathbb{R}^n$ can be extended to an Interval Neural Network (INN) $\boldsymbol{\Phi}_{\mathrm{INN}} \colon \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n, \quad \widetilde{\mathbf{x}} \mapsto \big(\boldsymbol{\Phi}(\widetilde{\mathbf{x}}), \underline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}}), \overline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}})\big)$, where $\underline{\boldsymbol{\Phi}}$ and $\overline{\boldsymbol{\Phi}}$ are mappings to lower and upper interval bounds for the prediction of the INN. Given training samples $(\widetilde{\mathbf{x}}_i, \mathbf{x}_i) = (\mathbf{A}^{\dagger}\mathbf{y}_i, \mathbf{x}_i)$, the INN is trained by minimizing

$$\sum_i \| \max\{\mathbf{x}_i - \overline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}}_i), 0\}\|_2^2 + \| \max\{\underline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}}_i) - \mathbf{x}_i, 0\}\|_2^2 + \beta \|\overline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}}_i) - \underline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}}_i)\|_1,$$

subject to constraints that guarantee $\underline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}}) \leq \boldsymbol{\Phi}(\widetilde{\mathbf{x}}) \leq \overline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}})$ for all $\widetilde{\mathbf{x}}$. Hence, the idea of INNs is to produce output intervals that contain the true labels with high probability, while remaining as tight as possible. The pixel-wise uncertainty estimate of an INN is then given by the width of the prediction interval, i.e., $\mathbf{u}_{\mathrm{INN}}(\widetilde{\mathbf{x}}) = \overline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}}) - \underline{\boldsymbol{\Phi}}(\widetilde{\mathbf{x}})$.

Monte Carlo Dropout In MCDrop proposed by [8], uncertainty scores are obtained through the sample variance of multiple stochastic forward passes on the same input data point. If $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_T$ are realizations of independent draws of random dropout masks of the prediction network $\boldsymbol{\Phi}$, then the pixel-wise uncertainty estimate is given by $\mathbf{u}_{\mathrm{MCD\textsc{rop}}}(\widetilde{\mathbf{x}}) = \frac{1}{T-1}(\sum_{t=1}^{T} \boldsymbol{\Phi}_t(\widetilde{\mathbf{x}})^2 - \frac{1}{T}(\sum_{t=1}^{T} \boldsymbol{\Phi}_t(\widetilde{\mathbf{x}}))^2)$.

Mean and Variance Estimation Another possibility is to double the number of outputs of the prediction network and train it to approximate the mean and variance of a Gaussian distribution. In [9], this is referred to as lightweight probabilistic networks (ProbOut) $\boldsymbol{\Phi}_{\mathrm{P\textsc{rob}O\textsc{ut}}} \colon \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n, \quad \widetilde{\mathbf{x}} \mapsto (\boldsymbol{\Phi}_{\mathrm{mean}}(\widetilde{\mathbf{x}}), \boldsymbol{\Phi}_{\mathrm{var}}(\widetilde{\mathbf{x}}))$, trained by minimizing $\sum_i \|(\mathbf{x}_i - \boldsymbol{\Phi}_{\mathrm{mean}}(\widetilde{\mathbf{x}}_i))/\sqrt{\boldsymbol{\Phi}_{\mathrm{var}}(\widetilde{\mathbf{x}}_i)}\|_2^2 + \| \log \boldsymbol{\Phi}_{\mathrm{var}}(\widetilde{\mathbf{x}}_i)\|_1$. The pixel-wise uncertainty score is given by $\mathbf{u}_{\mathrm{P\textsc{rob}O\textsc{ut}}}(\widetilde{\mathbf{x}}) = \boldsymbol{\Phi}_{\mathrm{var}}(\widetilde{\mathbf{x}})$.

2.2 Inverse Problem, Neural Network and Data

We consider a simulation of the noiseless Radon transform with a moderate missing wedge of 30° for the forward model (1). The non-learned inversion $\mathbf{A}^\dagger$ in (3) is based on the filtered backprojection algorithm (FBP). The underlying prediction network is a U-Net variant. Our experiments are based on a data set consisting of 512×512 human CT scans from the AAPM Low Dose CT Grand Challenge data [10].[1] In total, it contains 2580 images of 10 patients. Eight of these ten patients were used for training (2036 samples), one for validation (214 samples) and one for testing (330 samples).

3 Results

We perform two experiments on detecting instabilities in the context of limited angle CT; code can be found at `https://github.com/luisoala/inn`.

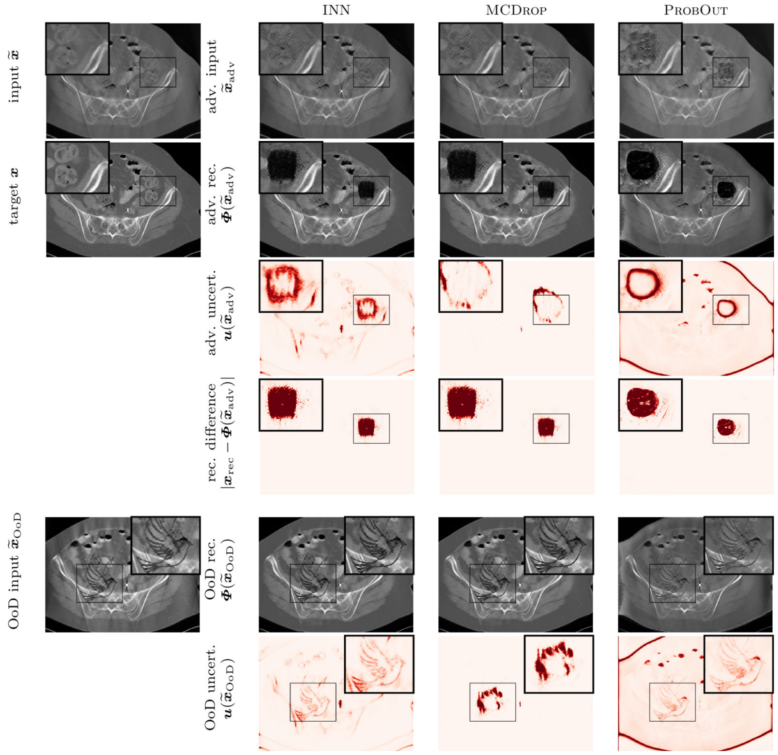
3.1 Adversarial Artifact Detection (AdvDetect)

The AdvDetect experiment assesses the capacity of the considered UQ methods to capture artifacts in the output that were caused by adversarial noise. To that end, we create perturbed inputs for each measurement sample $\mathbf{y}$ in the test set by employing the box-constrained L-BFGS algorithm to minimize the function $\|\mathbf{\Phi}(\widetilde{\mathbf{x}}_{\mathrm{adv}}) - \mathbf{x}_{\mathrm{adv.\ tar.}}\|_2^2$ over the domain $\widetilde{\mathbf{x}}_{\mathrm{adv}} \in [0,1]^n$. Here, $\mathbf{x}_{\mathrm{adv.\ tar.}}$ represents a corresponding adversarial target, which is created by subtracting 1.5 times its mean value from $\mathbf{x}_{\mathrm{rec}}$ within a random 50×50 square, leading to clearly visible artifacts in the corresponding reconstructions; see Fig. 1. It is arguable, whether the technical aspects of such an adversarial perturbation (i.e., attacking subsequently to a model-based inversion) is a realistic scenario in the context of inverse problems. However, for our purposes, such a simple setup (see also [5]) is sufficient.

In order to assess the adversarial artifact detection capacity, the different UQ schemes are then used to produce uncertainty heatmaps for the generated adversarial inputs. A quantitative evaluation is carried out by computing the mean Pearson correlation coefficient between the pixel-wise change in the uncertainty heatmaps $|\mathbf{u}(\widetilde{\mathbf{x}}) - \mathbf{u}(\widetilde{\mathbf{x}}_{\mathrm{adv}})|$ and the change of reconstructions $|\mathbf{x}_{\mathrm{rec}} - \mathbf{\Phi}(\widetilde{\mathbf{x}}_{\mathrm{adv}})|$. The results are summarized in Tab. 1 and illustrated in Fig. 1. We observe that both INN and ProbOut are able to detect the image region of adversarial perturbations. In particular INN highlights the effect of almost imperceptible input perturbations on the reconstructions. Overall, the uncertainty predictions of all three methods mostly emphasize boundary features in the image. While MCDrop shows fewer “False Positives”, it also exhibits more “False Negatives” compared to INN and ProbOut.

[1] See: https://www.aapm.org/GrandChallenge/LowDoseCT/; We would like to thank Dr. Cynthia McCollough, the Mayo Clinic, and the American Association of Physicists in Medicine as well as the grants EB017095 and EB017185 from the National Institute of Biomedical Imaging and Bioengineering for providing the AAPM data.

Fig. 1. Results of the three UQ methods for the AdvDetect and ArtDetect experiments for one exemplary slice. The plotting windows are slightly adjusted for better contrast.


INN
MCDrop
ProbOut
input $\widetilde{\boldsymbol{x}}$
target $\boldsymbol{x}$
adv. input $\widetilde{\boldsymbol{x}}_{\mathrm{adv}}$
adv. rec. $\boldsymbol{\Phi}(\widetilde{\boldsymbol{x}}_{\mathrm{adv}})$
adv. uncert. $\boldsymbol{u}(\widetilde{\boldsymbol{x}}_{\mathrm{adv}})$
rec. difference $|\boldsymbol{x}_{\mathrm{rec}} - \boldsymbol{\Phi}(\widetilde{\boldsymbol{x}}_{\mathrm{adv}})|$
OoD input $\widetilde{\boldsymbol{x}}_{\mathrm{OoD}}$
OoD rec. $\boldsymbol{\Phi}(\widetilde{\boldsymbol{x}}_{\mathrm{OoD}})$
OoD uncert. $\boldsymbol{u}(\widetilde{\boldsymbol{x}}_{\mathrm{OoD}})$

3.2 Atypical Artifact Detection (ArtDetect)

The ArtDetect experiment is designed analogously to the setup described by [6], i.e., an atypical artifact, which was not present in the training data, is randomly placed in the input. We insert the silhouette of a peace dove in each image of the test set; see Fig. 1. The simulation of the measurements and model-based inversions is carried out on the new test set as before.

In order to assess the atypical artifact detection capacity, the different UQ schemes are then used to produce uncertainty heatmaps on the resulting OoD inputs. A quantitative evaluation is carried out by computing the mean Pearson correlation coefficient between the change in the uncertainty heatmaps $|\mathbf{u}(\widetilde{\mathbf{x}}) - \mathbf{u}(\widetilde{\mathbf{x}}_{\mathrm{OoD}})|$ and a binary mask marking the region of change in the inputs. The results are summarized in Tab. 1 and illustrated in Fig. 1. All three UQ methods are correlated with the input change, however INN achieves the highest correlation. This shows that UQ in general, and INNs in particular, can serve as a warning system for inputs containing atypical features that might otherwise lead to unnoticed and possibly erroneous reconstruction artifacts.

Table 1. Mean Pearson correlation coefficients, averaged (± standard deviation) over three experimental runs, for both instability detection experiments.

UQ Method	AdvDetect	ArtDetect
INN	$\mathbf{0.56 \pm 0.05}$	$\mathbf{0.52 \pm 0.03}$
MCDrop	0.28 ± 0.02	0.26 ± 0.01
ProbOut	0.48 ± 0.12	0.34 ± 0.04

4 Discussion

We demonstrated qualitatively and quantitatively that uncertainty quantification, in particular by INNs, bears great potential as a fine-grained instability detector. This was shown for limited angle CT as a prototypical example of a severely ill-posed inverse problem. The presented UQ methods are versatile and can be employed for various types of neural networks and other clinical applications.

The implication and goal of this work is to ultimately move deep learning technology closer to a level of reliability that makes it a serious contender for integration in medical imaging workflows. If we want to harness the prowess of deep learning we will need to find strategies for accounting for its instabilities. Uncertainty quantification can be an important tool to that end.

References

1. Jin KH, McCann MT, Froustey E, et al. Deep Convolutional Neural Network for Inverse Problems in Imaging. IEEE Trans Image Process. 2017;26:4509–4522.
2. Arridge S, Maass P, Öktem O, et al. Solving inverse problems using data-driven models. Acta Numerica. 2019;28:1–174.
3. Foucart S, Rauhut H. A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis. Birkhäuser; 2013.
4. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. In: International Conference on Learning Representations; 2014. .
5. Huang Y, Würfl T, Breininger K, et al. Some Investigations on Robustness of Deep Learning in Limited Angle Tomography. In: Frangi AF, Schnabel JA, Davatzikos C, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018; 2018. p. 45–153.
6. Gottschling NM, Antun V, Adcock B, et al. The troublesome kernel: why deep learning for inverse problems is typically unstable?; 2020. ArXiv:2001.01258.
7. Oala L, Heiß C, Macdonald J, et al. Interval Neural Networks: Uncertainty Scores; 2020. ArXiv preprint arXiv:2003.11566.
8. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Balcan MF, Weinberger KQ, editors. Proceedings of The 33rd International Conference on Machine Learning. vol. 48 of Proceedings of Machine Learning Research. New York, New York, USA; 2016. p. 1050–1059.
9. Gast J, Roth S. Lightweight Probabilistic Deep Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018; p. 3369–3378.
10. McCollough CH. TU-FG-207A-04: Overview of the Low Dose CT Grand Challenge. Med Phys. 2016;43(6 Part 35):3759–3760.