Interval Neural Networks as Instability Detectors for Image Reconstructions

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

[1] Dept. of Mathematics, Technical University of Berlin, 10623 Berlin, Germany
{maerz,macdonald}@math.tu-berlin.de
[2] Machine Learning Group, Fraunhofer HHI, 10587 Berlin, Germany
{luis.oala,wojciech.samek}@hhi.fraunhofer.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 on two use cases 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.

Keywords: Inverse Problems · Deep Learning · Adversarial Attacks.

1 Introduction

Deep learning has shown the potential to outperform traditional schemes for solving various signal recovery problems in medical imaging applications [19,18,14,1]. Typically, such tasks are modelled as finite-dimensional linear inverse problems,

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

where $\boldsymbol{x} \in \mathbb{R}^n$ is the unknown signal of interest, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ denotes the forward operator representing a physical measurement process, and $\boldsymbol{\eta} \in \mathbb{R}^m$ is modelling noise in the measurements. Important examples include choosing $\boldsymbol{A}$ as the identity (denoising), 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 $\boldsymbol{x}$ from its observed measurements $\boldsymbol{y}$. The difficulty of this task is mainly determined by the strength

* Equal contribution in alphabetical order

of the noise and the degree of ill-posedness of (1), which is typically governed by the amount of undersampling in the measurement domain; cf. [15,9].

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 [9]. Recently, 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, however, the field is still in an early stage. Focusing primarily on recovery performance, aspects such as the reliability of reconstructions have not yet been extensively explored; see [2,4] for exceptions.

In image classification, the susceptibility of deep neural networks to adversarial exploitation is well documented [34,25,8]. Recent works have reported similar instabilities for image reconstruction tasks [17,3,12], which can be caused by visually imperceptible adversarial noise or features that have not been seen during training. Although there have been first attempts to alleviate these shortcomings [30,6], [12] 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 uncertainty quantification (UQ) as an instability detector. Out of the three compared UQ methods, the recently proposed Interval Neural Network framework of [28] is shown to be particularly well suited for this task.

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 [37,19,18]. Thus, the reconstruction is given by

$$\boldsymbol{x}_{\mathrm{rec}} = \boldsymbol{\Phi}(\boldsymbol{A}^{\dagger}\boldsymbol{y}), \tag{2}$$

where $\boldsymbol{\Phi}\colon \mathbb{R}^n \to \mathbb{R}^n$ denotes the prediction network (trained to minimize the loss $\|\boldsymbol{x} - \boldsymbol{\Phi}(\boldsymbol{A}^{\dagger}\boldsymbol{y})\|_2^2$) and $\boldsymbol{A}^{\dagger}$ symbolizes the non-learned model-based inversion.[3] Based on this reconstruction method, we then focus on two use cases. First, the standard imaging task of removing white Gaussian noise (i.e., $\boldsymbol{A}$ is the identity), which can be seen as a well-conditioned inverse problem, is examined.[4] Second, we consider the severely ill-posed problem of limited angle computed tomography ($\boldsymbol{A}$ is a subsampled Radon transform), which has applications in dental tomography, breast tomosynthesis or electron tomography. While $\boldsymbol{\Phi}$ is only used for a plain removal of Gaussian noise in the first case, the latter application requires a removal of structured artifacts as well as an "inpainting" of missing

[3] There is a variety of other possibilities to utilize neural networks to solve (1); see [5] for a comprehensive overview. However, [3] suggests, that the issue of instabilities occurs independently of the considered reconstruction scheme. Thus, we restrict our study to the simple "image-to-image" post-processing setting described above.

[4] Note that there is an intimate connection between denoising and solving general ill-posed inverse problems, that can for instance be exploited by "plug-and-play" schemes [36]; see also [29]. Thus, this application is chosen as a prototypical example for image-to-image regression by neural networks with a broad scope of implications.

edge information. On each of the two use cases, we investigate the capacity of three UQ schemes (see Section 2) to localize possible instabilities in the output of the prediction network $\boldsymbol{\Phi}$. 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 Section 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 described by [12,3,17].
c) We demonstrate that one UQ approach in particular, the so called Interval Neural Network, performs best as an instability detector.

Related Work In addition to the work cited above there exist strands of research in deep learning occupied with the detection of adversarial and OoD inputs. Maximum Mean Discrepancy, Kernel Density Estimation and other tools, see [8] for an overview, have been successfully employed for adversarial input detection. Popular methods for OoD input detection include Minimum Covariance Determinant [32], Support Vector Data Description [35], as well as methods geared particularly towards the deep model setting such as ODIN [22], Outlier Exposure [16], or detection in latent space [13].

The detection of adversarial and OoD inputs in these works is typically done in the classification setting. We emphasize that image-to-image regression by $\boldsymbol{\Phi}$ is a fundamentally different task: While classification is inherently discontinuous, $\boldsymbol{\Phi}$ addresses a problem that allows for stable reconstruction methods in many cases, e.g. by sparse regularization. Furthermore, we are not interested in a crude, outright rejection of data points in the *input space* but rather seek to obtain fine-grained information about erroneous artifacts in the *output space*. More closely related to our goal is the work of [20,11] where uncertainty quantification was considered for segmentation and depth-estimation tasks. Hence, we include their approaches as detection methods which are described next.

2 Detection Methods

We consider three methods for uncertainty quantification of neural network predictions and compare their capacity to detect reconstruction instabilities caused by adversarial noise and OoD features.

Interval Neural Network 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)

$$\boldsymbol{\Phi}_{\mathrm{INN}}\colon \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n, \quad \widetilde{\boldsymbol{x}} \mapsto \big(\boldsymbol{\Phi}(\widetilde{\boldsymbol{x}}), \underline{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{x}}), \overline{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{x}})\big) \tag{3}$$

where $\underline{\boldsymbol{\Phi}}$ and $\overline{\boldsymbol{\Phi}}$ are mappings to lower and upper interval bounds for the prediction of the INN, cf. supplementary material. Given labeled samples $(\widetilde{\boldsymbol{x}}_i, \boldsymbol{x}_i) = (\boldsymbol{A}^\dagger \boldsymbol{y}_i, \boldsymbol{x}_i)$ it is suggested in [28] to train the INN by minimizing the empirical loss

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

subject to constraints that guarantee $\underline{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{x}}) \leq \boldsymbol{\Phi}(\widetilde{\boldsymbol{x}}) \leq \overline{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{x}})$ for all $\widetilde{\boldsymbol{x}}$. 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., $\boldsymbol{u}_{\text{INN}}(\widetilde{\boldsymbol{x}}) = \overline{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{x}}) - \underline{\boldsymbol{\Phi}}(\widetilde{\boldsymbol{x}})$. We refer to [28] for further details on INNs and their evaluation in the context of uncertainty quantification.

Monte Carlo Dropout In MCDrop proposed by [10,20], uncertainty scores are obtained through the sample variance of multiple stochastic forward passes on the same input data point. In other words, if $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_T$ are realizations of independent draws of random dropout masks for the same prediction network $\boldsymbol{\Phi}$, then the pixel-wise uncertainty estimate is given by

$$\boldsymbol{u}_{\text{MCDrop}}(\widetilde{\boldsymbol{x}}) = \tfrac{1}{T-1} \left(\sum_{t=1}^T \boldsymbol{\Phi}_t(\widetilde{\boldsymbol{x}})^2 - \tfrac{1}{T} \left(\sum_{t=1}^T \boldsymbol{\Phi}_t(\widetilde{\boldsymbol{x}}) \right)^2 \right).$$

Mean & Variance Estimation The work by [27] proposed another simple recipe for uncertainty scores: the number of output components of the prediction network is doubled and trained to approximate the mean and variance of a Gaussian distribution. This approach has been recast by [11] as so-called lightweight probabilistic networks (ProbOut)

$$\boldsymbol{\Phi}_{\text{ProbOut}} \colon \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n, \quad \widetilde{\boldsymbol{x}} \mapsto (\boldsymbol{\Phi}_{\text{mean}}(\widetilde{\boldsymbol{x}}), \boldsymbol{\Phi}_{\text{var}}(\widetilde{\boldsymbol{x}}))$$

which are trained by minimizing the empirical loss

$$\sum_i \left\| \frac{\boldsymbol{x}_i - \boldsymbol{\Phi}_{\text{mean}}(\widetilde{\boldsymbol{x}}_i)}{\sqrt{\boldsymbol{\Phi}_{\text{var}}(\widetilde{\boldsymbol{x}}_i)}} \right\|_2^2 + \| \log \boldsymbol{\Phi}_{\text{var}}(\widetilde{\boldsymbol{x}}_i) \|_1.$$

The pixel-wise uncertainty score of ProbOut is then simply given by the variance estimate, i.e., $\boldsymbol{u}_{\text{ProbOut}}(\widetilde{\boldsymbol{x}}) = \boldsymbol{\Phi}_{\text{var}}(\widetilde{\boldsymbol{x}})$.

3 Experiments and Results

In this section, we first briefly report on the general deep learning setup of the experiments. Detailed technicalities are listed in the supplement for the sake of reproducability. Finally, we describe the actual experiments for the detection of instabilities and their results.

3.1 Inverse Problems, Neural Networks and Data

Image Denoising This task consists of removing additive Gaussian noise with standard deviation 25/255 from greyscale images (rescaled to the intensity range $[0, 1]$) from the Berkeley Segmentation Dataset [23]. The prediction network underlying all uncertainty methods is a fully-convolutional residual network with 17 convolution layers, inspired by [38].

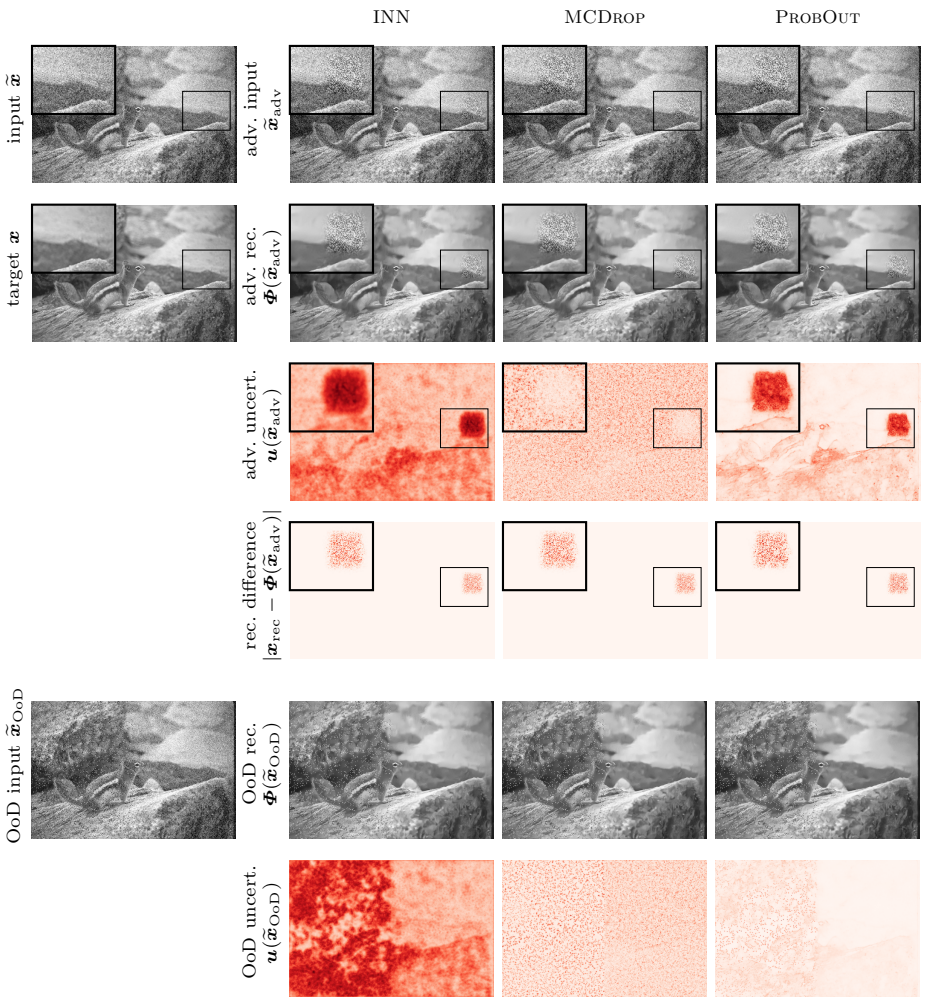

Fig. 1. Results of three UQ methods for the AdvDetect and ArtDetect experiments for one exemplary data sample of the Image Denoising task.

Limited Angle Computed Tomography (CT) For this task, 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 $\boldsymbol{A}^\dagger$ in (2) is based on the filtered backprojection algorithm (FBP) [26]. The underlying prediction network is a U-Net [31] 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 [24].[5] 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).

3.2 Instability Detection

Two experiments are performed on the two tasks described above. The first one, Adversarial Artifact Detection, examines the capacity of uncertainty quantification methods to detect adversarial inputs. The second experiment, Atypical Artifact Detection, exposes the prediction network to a novel structure that was not present during training, analogous to the out-of-distribution test in [3,12]. Both experiments are explained in detail below.

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 $\boldsymbol{y}$ in the test set by employing the box-constrained L-BFGS algorithm [7] to solve

$$\text{minimize}_{\widetilde{\boldsymbol{x}}_{\text{adv}} \in [0,1]^n} \; \|\boldsymbol{\Phi}(\widetilde{\boldsymbol{x}}_{\text{adv}}) - \boldsymbol{x}_{\text{adv. tar.}}\|_2^2 + \lambda \|\widetilde{\boldsymbol{x}}_{\text{adv}} - \widetilde{\boldsymbol{x}}\|_2^2, \tag{4}$$

where $\widetilde{\boldsymbol{x}} = \boldsymbol{A}^\dagger \boldsymbol{y}$ denotes the model based inversion, $\boldsymbol{x}_{\text{adv. tar.}}$ represents a corresponding adversarial target, and $\lambda \geq 0$ is a parameter for balancing the two terms in (4). It is arguable, whether the technical aspects of such an adversarial pertubation (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 [17]) is sufficient. We refer to [3,12], where adversarial noise is mapped to the measurement domain. For the Image Denoising data we use $\lambda = 0.5$, and the adversarial targets are created by adding noise to a random 50×50 patch in the reconstruction $\boldsymbol{x}_{\text{rec}} = \boldsymbol{\Phi}(\widetilde{\boldsymbol{x}})$ obtained via (2). Thus, the denoising network is forced to fail its task in that region; see Figure 1. For the Limited Angle CT task we found that the second term in (4) is not required, i.e., we use $\lambda = 0$. Adversarial targets are created by subtracting 1.5 times its mean value from $\boldsymbol{x}_{\text{rec}}$ within a random 50×50 square, leading to clearly visible artifacts in the corresponding reconstructions; see Figure 2. In order to assess the

[5] 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.

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 $|\boldsymbol{u}(\widetilde{\boldsymbol{x}}) - \boldsymbol{u}(\widetilde{\boldsymbol{x}}_{\text{adv}})|$ and the change of reconstructions $|\boldsymbol{x}_{\text{rec}} - \boldsymbol{\Phi}(\widetilde{\boldsymbol{x}}_{\text{adv}})|$. The results are summarized in Table 1 and illustrated in Figures 1 and 2. We observe that both INN and ProbOut are able to detect the image region of adversarial perturbations, with ProbOut achieving slightly higher correlations in the denoising task and INN having the highest correlation in the CT task. This shows that both methods are able to visually highlight the effect that almost imperceptible input perturbations have on the reconstructions.

Atypical Artifact Detection (ArtDetect) The ArtDetect experiment is designed analogous to the setup described by [12], i.e., an atypical artifact, which was not present in the training data, is randomly placed in the input. For the Image Denoising task this is achieved by locally changing the noise distribution, i.e., we replace the Gaussian noise by Salt & Pepper noise in one half of each image in the test set; see Figure 1. For the Limited Angle CT task the silhouette of a peace dove is inserted in each image of the test set; see Figure 2. 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 $|\boldsymbol{u}(\widetilde{\boldsymbol{x}}) - \boldsymbol{u}(\widetilde{\boldsymbol{x}}_{\text{OoD}})|$ and a binary mask marking the region of change in the inputs. The results are summarized in Table 1 and illustrated in Figures 1 and 2. All three UQ methods are correlated with the input change, however INN achieves the highest correlation in both the Image Denoising and CT task. 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.

4 Conclusion

We demonstrated qualitatively and quantitatively on two use-cases, image denoising and limited angle computed tomography, that uncertainty quantification, in particular INN and ProbOut, bears great potential as a fine-grained instability detector. Furthermore, Interval Neural Networks performed best overall in three out of four experiments. 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.

Table 1. Mean Pearson correlation coefficients, averaged (± standard deviation) over three experimental runs, for the Adversarial Artifact Detection and Atypical Artifact Detection experiments.

	Image Denoising		CT	
UQ Method	AdvDetect	ArtDetect	AdvDetect	ArtDetect
INN	0.77 ± 0.008	$\mathbf{0.69 \pm 0.006}$	$\mathbf{0.56 \pm 0.05}$	$\mathbf{0.52 \pm 0.03}$
MCDrop	0.20 ± 0.001	0.44 ± 0.02	0.28 ± 0.02	0.26 ± 0.01
ProbOut	$\mathbf{0.81 \pm 0.002}$	0.44 ± 0.01	0.48 ± 0.12	0.34 ± 0.04

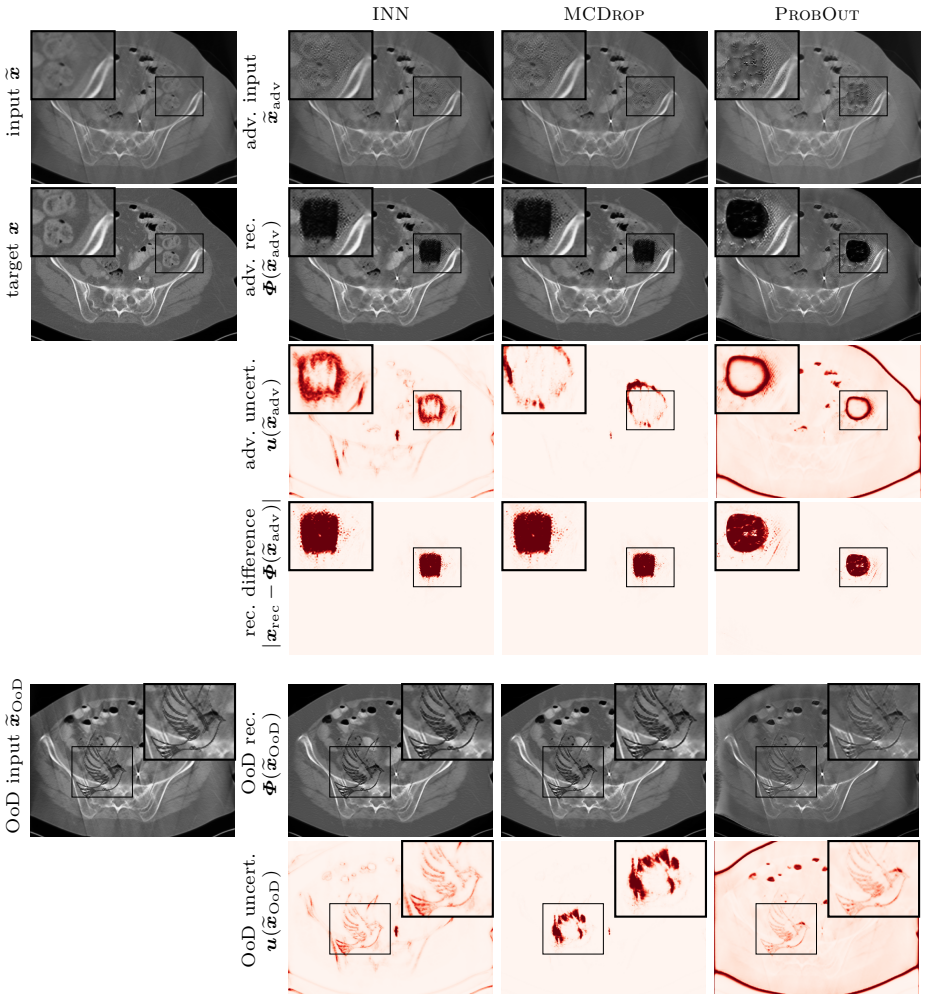

Fig. 2. Results of three UQ methods for the AdvDetect and ArtDetect experiments for one exemplary data sample of the Computed Tomography Reconstruction task. The plotting windows are slightly adjusted for better contrast.

A Supplementary Material

A.1 Details of Experimental Setup

Table 2. Summary of the technical details regarding the neural network architecures, training, and data sets for the two use cases of Image Denoising and Computed Tomography Reconstruction. Image Denoising data is available at `https://github.com/husqin/DnCNN-keras`(not affiliated with authors of this paper).

	Image Denoising	**Limited Angle CT**
Base Network	based on [38] dropout (0.05) after every other conv. trained with Adam[21], 50 epochs learning rate: 10^{-4} mini-batch size: 128 no batch normalization as in [38] 128 instead of 64 conv. channels, cf. [38]	U-Net of [31] dropout (0.7) after down-/up-sampling trained with Adam, 400 epochs learning rate: $7.5 \cdot 10^{-5}$ mini-batch size: 12
INN	10 epochs with Adam learning rate: 10^{-6} $\beta = 10^{-3}$ mini batch size: 96 interval arithmetic in last 8 layers	15 epochs with Adam learning rate: 10^{-6} $\beta = 10^{-4}$ mini batch size: 6 interval arithmetic in last 12 layers
MCDrop	$T = 128$ forward passes	$T = 16$ forward passes
ProbOut	additional output channel otherwise same setup as base network	additional output channel 400 more epochs with Adam learning rate: 10^{-7} mini-batch size: 12
Data	Berkeley Segmentation Dataset [23] 400 128×128-images; see [33,38] overlapping 40×40-patches, stride 10 rescaled to intesity range $[0, 1]$ Gaussian noise, standard dev. 25/255 testing: 68 images of varying size; cf. [38]	AAPM Low Dose CT Grand Challenge 10 patients: 2580 512×512-images (8/1/1 for training/validation/testing) noiseless Radon transform 30° missing wedge Ramp-filter for FBP

A.2 Interval Arithmetic in Neural Networks

We give a derivation of the lower and upper interval bounds $\underline{\boldsymbol{\Phi}}$ and $\overline{\boldsymbol{\Phi}}$ in equation (3) of the main paper. Interval Neural Networks (INNs) make use of interval arithmetic that deviates from customary arithmetic. The forward pass through a ReLU neural network layer $\boldsymbol{x} \mapsto \varrho(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$ in interval arithmetic is as follows: Given a component-wise interval valued input $[\underline{\boldsymbol{x}}, \overline{\boldsymbol{x}}]$ and interval valued

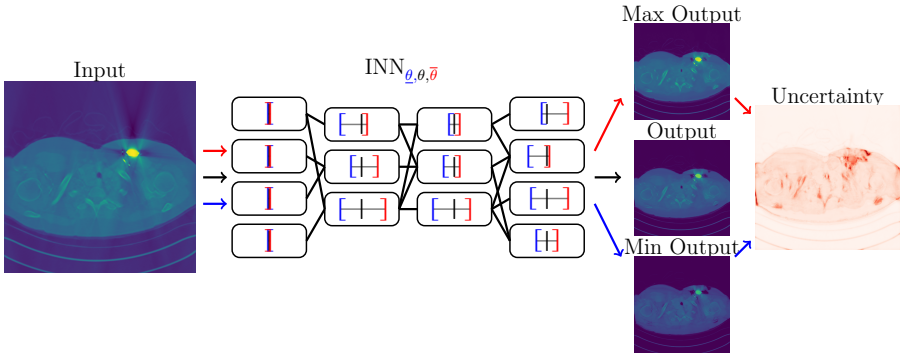

Fig. 3. INN Schematic Overview. The structure of an Interval Neural Network figure reproduced from [28] with permission from the authors.

weight matrices $\left[\underline{\boldsymbol{W}}, \overline{\boldsymbol{W}}\right]$ and bias vectors $\left[\underline{\boldsymbol{b}}, \overline{\boldsymbol{b}}\right]$ the output interval $[\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}]$ after propagation through the layer is formally expressed as

$$[\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}] = \varrho\left(\left[\underline{\boldsymbol{W}}, \overline{\boldsymbol{W}}\right][\underline{\boldsymbol{x}}, \overline{\boldsymbol{x}}] + \left[\underline{\boldsymbol{b}}, \overline{\boldsymbol{b}}\right]\right).$$

In the special case where $[\underline{\boldsymbol{x}}, \overline{\boldsymbol{x}}]$ is non-negative—for example image inputs scaled to the intensity range $[0, 1]$ or outputs of a previous ReLU layer—this can be explicitly calculated via

$$\begin{aligned}
\underline{\boldsymbol{z}} &= \varrho\left(\max\left\{\underline{\boldsymbol{W}}, 0\right\}\underline{\boldsymbol{x}} + \min\left\{\underline{\boldsymbol{W}}, 0\right\}\overline{\boldsymbol{x}} + \underline{\boldsymbol{b}}\right), \\
\overline{\boldsymbol{z}} &= \varrho\left(\min\left\{\overline{\boldsymbol{W}}, 0\right\}\underline{\boldsymbol{x}} + \max\left\{\overline{\boldsymbol{W}}, 0\right\}\overline{\boldsymbol{x}} + \overline{\boldsymbol{b}}\right),
\end{aligned}$$

where the maximum and minimum functions are applied component-wise. Applying this for all network layers finally yields $\underline{\boldsymbol{\Phi}}$ and $\overline{\boldsymbol{\Phi}}$.

Acknowledgements The authors would like to thank Sören Becker for feedback on the final draft. M.M. acknowledges support by the DFG Priority Programme DFG-SPP 1798 Grants KU 1446/21 and KU 1446/23.

References

1. Adler, J., Öktem, O.: Learned Primal-dual Reconstruction. IEEE T. Med. Imaging **37**(6), 1322–1332 (2018)
2. Adler, J., Öktem, O.: Deep Bayesian Inversion (Nov 2018), arXiv: 1811.05910
3. Antun, V., Renna, F., Poon, C., Adcock, B., Hansen, A.C.: On instabilities of deep learning in image reconstruction - does AI come at a cost? (2019), arXiv:1902.05300
4. Ardizzone, L., Kruse, J., Wirkert, S.J., Rahner, D., Pellegrini, E.W., Klessen, R.S., Maier-Hein, L., Rother, C., Köthe, U.: Analyzing inverse problems with invertible neural networks. In: International Conference on Learning Representations (2018)
5. Arridge, S., Maass, P., Öktem, O., Schönlieb, C.B.: Solving inverse problems using data-driven models. Acta Numerica **28**, 1–174 (2019)
6. Bubba, T.A., Kutyniok, G., Lassas, M., März, M., Samek, W., Siltanen, S., Srinivasan, V.: Learning the invisible: a hybrid deep learning-shearlet framework for limited angle computed tomography. Inverse Problems **35**(6), 064002 (2019)
7. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995)
8. Carlini, N., Wagner, D.: Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. p. 3–14. AISec '17 (2017)
9. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis, Birkhäuser (2013)
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. New York, New York, USA (2016)
11. Gast, J., Roth, S.: Lightweight Probabilistic Deep Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 3369–3378 (2018)
12. Gottschling, N.M., Antun, V., Adcock, B., Hansen, A.C.: The troublesome kernel: why deep learning for inverse problems is typically unstable? (2020), arXiv:2001.01258
13. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., Aspuru-Guzik, A.: Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science **4**(2), 268–276 (2018)
14. Hammernik, K., Klatzer, T., Kobler, E., Recht, M.P., Sodickson, D.K., Pock, T., Knoll, F.: Learning a variational network for reconstruction of accelerated mri data. Magnetic Resonance in Medicine **79**(6), 3055–3071 (2018)
15. Hansen, P.C.: Discrete Inverse Problems. Society for Industrial and Applied Mathematics (2010)
16. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. Proc. of the International Conference on Learning Representations (2019)
17. Huang, Y., Würfl, T., Breininger, K., Liu, L., Lauritsch, G., Maier, A.: Some investigations on robustness of deep learning in limited angle tomography. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G.

(eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 145–153 (2018)
18. Jin, K.H., McCann, M.T., Froustey, E., Unser, M.: Deep Convolutional Neural Network for Inverse Problems in Imaging. IEEE Trans. Imag. Proc. **26**, 4509–4522 (2017)
19. Kang, E., Min, J., Ye, J.C.: A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. Med. Phys. **44**(10), 360–375 (2017)
20. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 5580–5590. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
21. Kingma, D.P., Ba, J.A.: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Liang, S., Li, Y., Srikant, R.: Principled detection of out-of-distribution examples in neural networks. Proceedings of the International Conference on Learning Representations (2019)
23. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 8th Int'l Conf. Computer Vision. vol. 2, pp. 416–423 (July 2001)
24. McCollough, C.: Tu-fg-207a-04: Overview of the low dose ct grand challenge. Med. Physs **43**(6 Part 35), 3759–3760 (2016)
25. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. pp. 2574–2582 (06 2016)
26. Natterer, F.: The Mathematics of Computerized Tomography. SIAM (2001)
27. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94). vol. 1 (Jun 1994)
28. Oala, L., Heiß, C., Macdonald, J., März, M., Samek, W., Kutyniok, G.: Interval neural networks: Uncertainty scores (2020), arXiv:2003.11566
29. Oymak, S., Hassibi, B.: Sharp mse bounds for proximal denoising. Found. Comput. Math. **16** (2016)
30. Raj, A., Bresler, Y., Li, B.: Improving Robustness of Deep-Learning-Based Image Reconstruction (2020), arXiv:2002.11821
31. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241 (2015)
32. Rousseeuw, P.J.: Least median of squares regression. Journal of the American Statistical Association **79**(388), 871–880 (1984)
33. Schmidt, U., Roth, S.: Shrinkage fields for effective image restoration. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2774–2781 (2014)
34. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014)
35. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Mach. Learn. **54**(1), 45–66 (2004)
36. Venkatakrishnan, S.V., Bouman, C.A., Wohlberg, B.: Plug-and-play priors for model based reconstruction. In: 2013 IEEE Global Conference on Signal and Information Processing. pp. 945–948 (2013)

37. Zhang, H., Li, L., Qiao, K., Wang, L., et al.: Image Prediction for Limited-angle Tomography via Deep Learning with Convolutional Neural Network. arXiv:1607.08707 (2016)
38. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Trans. Imag. Proc. **26**, 3142–3155 (2017)