Measuring Robustness in Deep Learning Based Compressive Sensing

Mohammad Zalbagi Darestani 1 Akshay S. Chaudhari 2 Reinhard Heckel 1 3

Abstract
Deep neural networks give state-of-the-art accuracy for reconstructing images from few and noisy measurements, a problem arising for example in accelerated magnetic resonance imaging (MRI). However, recent works have raised concerns that deep-learning-based image reconstruction methods are sensitive to perturbations and are less robust than traditional methods: Neural networks (i) may be sensitive to small, yet adversarially-selected perturbations, (ii) may perform poorly under distribution shifts, and (iii) may fail to recover small but important features in an image. In order to understand the sensitivity to such perturbations, in this work, we measure the robustness of different approaches for image reconstruction including trained and un-trained neural networks as well as traditional sparsity-based methods. We find, contrary to prior works, that both trained and un-trained methods are vulnerable to adversarial perturbations. Moreover, both trained and un-trained methods tuned for a particular dataset suffer very similarly from distribution shifts. Finally, we demonstrate that an image reconstruction method that achieves higher reconstruction quality, also performs better in terms of accurately recovering fine details. Our results indicate that the state-of-the-art deep-learning-based image reconstruction methods provide improved performance than traditional methods without compromising robustness.

1. Introduction
Neural networks outperform traditional (e.g., sparsity-based) methods in a variety of image reconstruction tasks across common metrics of image quality. For instance, consider the compressive sensing problem arising in magnetic resonance imaging (MRI), where the goal is to reconstruct a diagnostic quality image using linear under-sampled measurements for accelerating the MRI scans. A large body of literature shows that neural networks can enable higher reconstruction quality and faster reconstruction computations for MRI compared to clinically utilized traditional sparsity-based reconstruction methods (Hammernik et al., 2018; Zbontar et al., 2018; Knoll et al., 2020; Sriram et al., 2020a; Wang et al., 2019; Schlemper et al., 2017; Putzky & Welling, 2019; Arora et al., 2020; Zalbagi Darestani & Heckel, 2020).

However, recent works have raised robustness concerns about neural-network-based image recovery (Cohen et al., 2018; Huang et al., 2018; Antun et al., 2020; Gottschling et al., 2020). Specifically, Antun et al. (2020) demonstrated that small, adversarially-selected perturbations in the under-sampled measurements may result in significant reconstruction artifacts. Based on those findings, Antun et al. (2020) and Gottschling et al. (2020) concluded that despite providing worse reconstruction quality than neural networks, traditional sparsity-based compressive sensing methods tend to be more robust to adversarial perturbations. Moreover, the results from the fastMRI challenge, a competition for improving the performance of image reconstruction systems for accelerated MRI, have shown that while all top performing deep networks yield high scores according to various image quality metrics, there is also a potential to miss small, clinically relevant pathologies (Knoll et al., 2020, Fig. 3), which may lead to false-negative diagnoses.

These findings, however, do not address the question whether the identified robustness concerns are specific to training a network for image reconstruction, or whether un-trained neural networks and traditional un-trained sparsity-based reconstruction methods are similarly sensitive to perturbations.

In addition, it is unknown whether neural networks for image reconstruction are sensitive to various distribution shifts, which can commonly arise in clinical practice. Natural distribution shifts have shown to degrade performance in learning-based systems, for example for classification problems (Recht et al., 2019; Taori et al., 2020; Miller et al., 2021; Yadav & Bottou, 2019) and for question answering models in natural language processing (Miller et al., 2020).
In this work, we study the robustness of three popular families of compressive sensing reconstruction methods for MRI:

i) **Neural networks trained end to end.** All leading methods in terms of image reconstruction quality of the 2019 and 2020 fastMRI competitions belong to this class of methods. We consider U-net-based recovery (Ronneberger et al., 2015), as it is a simple baseline, and the end-to-end variational network (Sriram et al., 2020a), which is a state-of-the-art method on the fastMRI dataset (Zbontar et al., 2018).

ii) **Traditional CS methods.** We consider imposing a sparsity prior and image reconstruction with $\ell_1$-norm minimization, which is the leading classical compressive sensing method (Lustig et al., 2007).

iii) **Un-trained neural networks.** We consider a variation of the Deep Image Prior (Ulyanov et al., 2018) and Deep Decoder (Heckel & Hand, 2019) for MRI introduced in (Zalbagi Darestani & Heckel, 2020), which is a neural network based method, but works without training data.

We study the aforementioned methods with respect to the following three notions of robustness (see Figure 1):

i) **Small adversarial perturbations.** Small perturbations are selected adversarially to cause a large reconstruction error for a given image and an image reconstruction method. Antun et al. (2020) studied the effect of such perturbations, but only selected them for neural networks trained end-to-end. However, for a comparison among methods, it is essential to find these perturbations separately for each method. While there is no explicit adversary in an MRI scanner (as an MRI scanner is a closed system), studying adversarial perturbations is important since measurement perturbations can be due to scanner-induced artifacts (coil failure, B0 magnet drift, bad gradient shimming, etc.) or patient-induced artifacts (motion, off-resonance, etc.) (Bellon et al., 1986).

ii) **Robustness to distribution shifts.** Neural networks are typically evaluated by first collecting a set of images and measurements, second partitioning the data into training and test sets, and third training and evaluating the network on the train and test data. Thus, both train and test data are drawn from the same distribution. However, in practice train and test distributions may vary: for example, a network can be trained on data from one set of patients, anatomies, and contrasts, but may be used for varying datasets at test time. For classification, it is well known that distribution shifts are common in practice and have a large impact on performance (Recht et al., 2019), but for image reconstruction methods the effect of distribution shifts is unknown.

iii) **Robustness in recovering fine details.** Small features in an MRI image (e.g., 10-pixel sized features or even smaller) are often important for accurate diagnostics. Therefore, it is crucial for a reconstruction method to be able to recover such features in the image.

Understanding the robustness of algorithms with respect to these perturbations is important, especially in medical imaging where errors may result in a faulty diagnosis. Towards improving our understanding in this realm, the main contributions of our work and findings of our study are:

- All studied methods—trained and un-trained—are sensitive to small, adversarially-selected perturbations, and the performance loss is not unique to neural networks trained end-to-end. This result is in contrast to that of Antun et al. (2020), which found that “deep learning typically yields unstable methods,” whereas $\ell_1$-minimization methods “are not affected by the perturbation.” Those two results can co-exist because
Antun et al. (2020) found that an adversarial perturbation selected for a particular neural network does not affect $\ell_1$-minimization much. In this work (in contrast to Antun et al. (2020)), we also compute adversarial perturbations for $\ell_1$-minimization and show that while those significantly impair the performance of $\ell_1$-minimization, they affect neural networks less.

- We provide the first study of distribution shifts in the context of image reconstruction. Even un-trained methods such as $\ell_1$-minimization are affected by distribution shifts, because their hyper-parameters are tuned on a given distribution. We study three notions of natural distribution shifts for accelerated MRI reconstruction: (i) slight shift to a close domain (i.e., same anatomy but a dataset with a different acquisition technique), (ii) anatomy shift (i.e., training on brains and testing on knees), and (iii) adversarially-filtered shifts. Perhaps surprisingly, we find that both un-trained and trained methods are similarly affected by all three types of distribution shifts (see Figure 2 as an example for case (i)), and typically the best performing method is also the best performing method under a distribution shift. For adversarially-filtered shifts, we find that challenging-to-reconstruct-images are difficult to reconstruct for all methods, from which we conclude that some images are naturally difficult to reconstruct.

- Finally, we quantify the recovery of small features in the images i) through introducing artificial small features, and ii) through studying small, clinically relevant features in real data. From studying artificial small feature recovery, we find that each reconstruction method is sensitive to specific regions in an image and faces difficulty in recovering small features in those regions. By studying small pathologic features in real data, we find that traditional CS methods are less robust in recovering small details compared to neural networks. This confirms the intuition that small feature recovery ability should correlate with overall reconstruction performance.

In a nutshell, the take-away of our study is that deep learning methods that perform best based on reconstruction quality are also best under realistic distribution shifts and for small feature recovery, and we could not find them to be more sensitive to adversarial perturbations than classical methods.

2. Problem setup: Accelerated multi-coil MRI

We study robustness in the context of accelerated multi-coil MRI reconstruction, because this is one of the most popular applications of compressive sensing and an important medical imaging technology.

In multi-coil MRI, $n_c$ multiple radiofrequency coils each record a measurement that is sensitive to a spatially local anatomical region. We consider multi-coil MRI over simpler single-coil MRI experiments since using multiple coils is common in clinical practice.

The goal of accelerated multi-coil MRI reconstruction is to reconstruct an image $x^* \in \mathbb{C}^N$ from a set of measurements (often called $k$-space measurements) obtained as

$$y_i = \text{MFS}_i x^* + \text{noise} \in \mathbb{C}^M, \quad i = 1, \ldots, n_c.$$

In the above equation, $S_i$ is a complex-valued position-dependent coil sensitivity map, that is applied through element-wise multiplication to the image $x^*$, $F$ implements the 2D discrete Fourier transform, and $M$ is a mask that implements under-sampling of $k$-space data.

For all experiments, we use the fastMRI dataset (Zbontar et al., 2018), designed for training and evaluating deep-learning-based MRI reconstruction methods. The fastMRI dataset consists of fully-sampled measurements (i.e., taken with an identity mask $M = 1$) of knees taken with $n_c = 15$ coils, and of brains taken with a varying number of coils. The dataset also contains reference images that are obtained by reconstructing the coil images from each full coil measurement as $\hat{x}_i = F^{-1} y_i$ and then combining them via the root-sum-of-squares (RSS) algorithm to a final single image: $\hat{x} = \sqrt{\sum_{i=1}^{n_c} |\hat{x}_i|^2}$. Here, $|\cdot|$ and $\sqrt{\cdot}$ denote element-wise absolute value and squared root operations. Alternatively, one could use SENSE (Roemer et al., 1990), a signal-to-noise ratio optimal method, to combine coil images.

To under-sample $k$-space measurements (for acceleration), we employ a standard 1D mask (random or equi-spaced vertical or horizontal lines in the frequency domain), which is the default in the fastMRI challenge. We consider 4x acceleration throughout the paper as this is one of the acceleration factors considered in the fastMRI challenge, although 2x
3. Image reconstruction methods

We consider three families of image reconstruction methods: i) trained neural networks, ii) traditional sparsity-based CS methods, and iii) un-trained neural networks. In this section, we provide a brief overview of each method.

**Trained neural networks.** Convolutional neural networks are trained either to map the measurement directly to an artifact-free image, or to map from a coarse least-squares reconstruction from the under-sampled measurement to an artifact-free image.

Let \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) be a training set consisting of pairs of target image \( x_j \) and a under-sampled measurement \( y_j \). The measurement in our setup consists of the physical k-space measurements from multiple receiver coils. A network \( f_\theta \) with parameters (weights) \( \theta \) takes as input the measurement and generates an image. The network is typically trained by minimizing the loss

\[
\mathcal{L}(\theta) = \frac{1}{n} \sum_{j=1}^{n} \| x_j - f(y_j) \|_2^2,
\]

which yields the trained method \( f_\theta \). At test time, the network generates an image as \( f_\theta(y) \) based on the measurement \( y \).

The best performing methods in the fastMRI competition are all trained networks, and yield significant improvements over classical methods (Putzky & Welling, 2019; Sriram et al., 2020b,a). We consider two methods of this type: U-net (Ronneberger et al., 2015) based reconstruction (Jin et al., 2017), as this is a simple baseline, and the end-to-end variational network (VarNet) (Sriram et al., 2020a), as this is the current state-of-the-art network.

U-net based reconstruction simply trains a U-net end-to-end on the training set to map the under-sampled measurement to the original image. VarNet-based reconstruction (Sriram et al., 2020a) is more intricate, with a more complicated network including coil sensitivity estimation, image-domain refinement, and data consistency steps, but conceptually similar as it is also trained end-to-end.

**Traditional compressive sensing methods.** Traditional compressive sensing methods are sparsity based and either impose a sparse representation by minimizing the \( \ell_1 \)-norm, or perform total-variation norm minimization. Traditional CS methods are popular for MRI reconstruction (Chen & Huang, 2012; Block et al., 2007; Lustig et al., 2007), and are used commonly in clinical practice. \( \ell_1 \)-norm minimization relies on assuming sparsity in a transform domain. We impose Wavelet sparsity following common practice in MRI (Chen & Huang, 2012; Lustig et al., 2007). Sparsity-based reconstruction recovers an image by minimizing the (convex) loss

\[
\mathcal{L}_1(x) = \sum_{i=1}^{n_c} \left\| y_i - A\hat{S}x \right\|_2^2 + \lambda \| Hx \|_1.
\]

Here, \( H \) denotes the 2D Wavelet transform and \( \hat{S} \) are coil sensitivity maps estimated from the under-sampled measurement using the ESPIRiT method (Uecker et al., 2014). Sparsity-based approaches provably succeed provided that the signal is sparse (Cardés et al., 2006; Lustig et al., 2007).

**Un-trained neural networks.** Perhaps surprisingly, convolutional neural networks can regularize inverse problems without training, as has first been demonstrated by the Deep Image Prior (Ulyanov et al., 2018) for denoising, super-resolution, and inpainting problems.

Such un-trained networks are also powerful for compressive sensing (Veen et al., 2018), and simple convolutional architectures such as the Deep Decoder (Heckel & Hand, 2019) work well in practice. Arora et al. (2020) and Zalbagi Darestani & Heckel (2020) used variants of the deep decoder for multi-coil MRI and achieved noticeable improvements over traditional CS methods. Zalbagi Darestani & Heckel (2020) further demonstrated that un-trained networks even perform similar to the U-net—the trained approach mentioned earlier. Un-trained neural networks provably recover smooth signals from few measurements (Heckel & Soltanolkotabi, 2020a).

In a nutshell, an un-trained network recovers an image by first fitting a randomly initialized, image generating network to a measurement, and then taking the network output as the recovered image. The network is not trained, and uses the structure of the network alone as a prior for the images.

We finally note that there is a fourth class of neural network-based reconstruction methods, pioneered by (Bora et al., 2017), which impose a learned prior (Bora et al., 2017; Hand & Voroninski, 2018; Asim et al., 2020; Daras et al., 2021). At the time of writing, this class of neural network-based reconstruction methods has not been applied to MRI, and was therefore not included in our study. In the meantime, Kelkar et al. (2021) applied this method to MRI images, and it would be interesting to include this class of methods into further robustness studies.

4. Small, adversarially-selected perturbations

Adversarial perturbations are often studied to evaluate the robustness of machine learning systems. For classification problems, there exists a large body of literature on the sensitivity of neural networks to adversarial perturbations (Goodfellow et al., 2015; Szegedy et al., 2014; Moosavi-Dezfooli et al., 2017). These works show that the predicted label for
The study of adversarial robustness in classification is motivated by the concept of an adversary that may alter the input of a machine learning system with imperceptible perturbations. In image reconstruction problems, there is typically no such adversary, but studying adversarial robustness in image reconstruction methods is important as it enables certifying worst-case robustness.

The study of small perturbations for image reconstruction problems was initiated by Cohen et al. (2018); Huang et al. (2018); Antun et al. (2020). Antun et al. (2020) generated these perturbations for a given network and image through optimization and concluded that “deep learning for inverse problems is typically unstable” because of the impact of adversarial perturbations. In a recent subsequent work, Genzel et al. (2020) unrolled a classical method, total variation minimization (TV), and generated small perturbations for this unrolled network in the same way as Antun et al. (2020) did, and found “superior robustness of the learned reconstruction schemes [a U-net] over TV.” Thus there is a disagreement on whether neural network or classical methods are more sensitive to adversarial perturbations.

In this section we study small adversarial perturbations, but in contrast to those previous works, (i) we include un-trained neural networks in our study, (ii) we include the state-of-the-art method VarNet beyond the baseline trained method U-net, and (iii) we introduce a method to generate small adversarial perturbations for un-trained methods without unrolling and thus without treating un-trained methods as a neural network.

Our experimental setup is as follows. We consider 10 randomly-chosen proton-density-weighted knee MRI images from the fastMRI validation set. For each image, we generate a small perturbation added to the measurement \( (k\text{-space}) \) of a given \( \ell_2 \) norm. For \( \ell_1 \)-minimization and the ConvDecoder, we generate adversarial perturbations through a new optimization-based method detailed in Appendix A.1. For U-net- and VarNet-based reconstruction, we generate perturbations with Projected Gradient Descent (PGD) as described in Appendix A.1. We then apply each perturbation to each image, and reconstruct with all four methods. Figure 3 shows the results and the supplement contains reconstructions examples.

Our experiment shows that both trained and un-trained methods are sensitive to small adversarial perturbations. Recall that (Antun et al., 2020; Gottschling et al., 2020) found that perturbations adversarially selected for neural networks only mildly affect sparsity-based methods, and concluded that traditional CS methods are more robust than trained methods to such small perturbations. In contrast, our results show that while adversarially-selected perturbation for a trained network (e.g., U-net) has a relativelly mild impact on \( \ell_1 \)-norm minimization, the converse is also true: perturbations found for \( \ell_1 \)-norm minimization have a significant effect for \( \ell_1 \)-minimization, but only a mild effect on U-net’s performance. Thus, both methods are vulnerable to perturbations specifically tailored to them.

We note that at first sight, the findings in Figure 3 suggest that ConvDecoder is slightly more robust than both U-net and \( \ell_1 \)-based reconstruction, which in turn is slightly more robust than VarNet-based reconstruction. However, it is not possible to draw such absolute comparisons, simply because in order to find adversarial perturbations, we solve a non-convex optimization problem with a numerical method (PGD), and because of the non-convexity, we are not guaranteed to find a worst-case perturbation. It could be that for \( \ell_1 \)-minimization we find a worst-case perturbation but for U-net we do not. This issue is inherent to the problem setup and applies to all current methods for finding adversarial perturbations, including those from (Antun et al., 2020; Genzel et al., 2020).
5. Distribution shifts

While robustness to distribution shifts has gained lots of attention over the past few years for image classification tasks (Recht et al., 2019; Ovadia et al., 2019; Taori et al., 2020; Hendrycks et al., 2020), there is little understanding on the effect of distribution shifts in image reconstruction problems. In particular, we are not aware of a systematic study on distribution shifts in accelerated MRI reconstruction. Understanding the robustness to distribution shifts in image reconstruction problems is important, since given the paucity of training datasets, it may be common to train a method on one patient population, but use the trained model on another; or train on the machine of one manufacturer and test on the machine of another, etc.

In this section, we study three variants of distribution shifts: (i) A dataset shift from the fastMRI-knee dataset to a knee dataset from mridata.org that we call the Stanford set, (ii) an anatomy shift from brain images to knee images and vice versa, and finally (iii) an adversarially-filtered shift to evaluate different models on a set of difficult-to-reconstruct images, inspired by adversarially-filtered shifts for classification introduced by Hendrycks et al. (2021).

We note all methods, even un-trained ones are affected by distribution shifts, because un-trained methods have hyper-parameters (such as the penalty $\lambda$ in $\ell_1$-minimization), that are tuned on a given distribution. Moreover, we consider multiple variants of each method (e.g., VarNet with different hyper-parameters) in order to obtain a larger variety of models. The details of each reconstruction method and its variants are provided in the supplement.

Our overarching finding is that the performance drop under each of the distribution shifts is similar for all considered trained and un-trained methods. Thus, out-of-distribution performance is strongly correlated with in-distribution performance. We found that surprising because un-trained methods depend only very mildly on the distribution through hyper-parameter tuning. As a consequence, the advantage achieved by a given method over another typically retains this advantage even under distribution shifts.

This complements an emerging line of works starting with (Recht et al., 2019) that finds a strong correlation between out-of-distribution and in-distribution generalization for a large variety of datasets and models for image classification problems (Miller et al., 2021; Taori et al., 2020; Yadav & Bottou, 2019) and even for question-answering models (Miller et al., 2020). Our results, presented next, indicate that this relation persists even in the context of image recovery, and even when including un-trained methods.

5.1. Dataset shift

We start with studying the performance of models trained or tuned on the fastMRI knee dataset, but tested on a different knee dataset. Specifically, we test on the Stanford dataset retrieved by collecting all available 18 knee volumes from mridata.org (Epperson et al., 2013). The Stanford set contains knees of the same size as the fastMRI images ($320 \times 320$), but the dataset is different in that (i) the frequency-domain representation of the Stanford set has a $320 \times 320$ resolution as opposed to $640 \times 360$ on average for fastMRI, (ii) the slice thickness is lower for the Stanford dataset (0.6mm vs 3mm), resulting in lower SNR, and (iii) the Stanford set is acquired using 3D MRI (single volumetric MRI measurement) vs. 2D fastMRI (multiple slice-wise measurements), resulting in varying blurring and SNR. All those slight differences induce a clinically relevant distribution shift.

Since all of the Stanford set samples are fat-suppressed images, when considering the shift from fastMRI to the Stanford set, we only consider fat-suppressed images from the fastMRI dataset as well. Our main finding is that all reconstruction methods perform worse on the new MRI samples, but the absolute performance drop is similar. In-distribution and out-distribution performances are linearly correlated. Figure 2 shows average SSIM values when training on the fastMRI dataset and evaluating on both fastMRI and Stanford datasets. Reconstruction examples are provided in the supplement.

We finally remark that when naively applied, the state-of-the-art VarNet is particularly sensitive to this dataset shift, in that the frequency resolution changes in the Stanford set and this affects VarNet as it is based on estimating the unknown $k$-space. This is not reflected in Figure 2, since we manually increased the resolution of Stanford set data points for VarNet to have a fair comparison among all methods with respect to only the dataset shift (and not resolution shift). For an example of VarNet reconstruction without such resolution fix, we refer to the supplement.

5.2. Anatomy shift

We next consider an anatomy shift where we move from a certain image type (knees) to another (brains), and vice versa. To understand the robustness to anatomy shifts, we perform the following experiment: We train U-net and VarNet on the whole knee training set and also optimize the hyper-parameters of the ConvDecoder and $\ell_1$-norm minimization on that set. Then we test all methods on the brain validation set. We conversely train on brain MRIs and test on knee MRIs.

Our main finding from the corresponding results in Figure 4 is that, again, all reconstruction methods perform worse on the new anatomy, but the absolute performance drop is similar. Note that there is more variance in the scores for knee images since there are two considerably different knee image types. Those different types arise from different contrasts and different MRI sequences for obtaining the measurements. For the brain images, there are 5 categories...
We finally study the performance on images that are partic-
ularly difficult to reconstruct as measured by their recon-
struction error, in order to understand whether any of the
considered methods degrade or shine on such naturally dif-
ficult examples. This experiment is inspired by the study
of “adversarially-filtered” data in image classification. Such
points refer to a set of challenging samples that cause a
significant performance loss to most of the classifiers, and
was introduced by Hendrycks et al. (2021) as ImageNet-A.

ImageNet-A consists of all ImageNet (Deng et al., 2009)
images that ResNet-50 misclassifies.

Here, we create fastMRI-A, a subset of the fastMRI dataset
that contains the most challenging to reconstruct samples,
and test all methods on this set of difficult images.

We construct the fastMRI-A (A for adversarial) dataset as
follows. We take 5 mid-slice images from each of the 199
knee validation volumes of the fastMRI dataset. This results
in a set of 995 images. From this set, we select the images
that result in the 100 lowest SSIM scores (bottom 10%) when reconstructing them via the i-RIM architecture (Putzky
& Welling, 2019). We use the i-RIM architecture, a fifth
reconstruction method, because this network is one of the
best-performing methods in the fastMRI competition, and
the winner of the single-coil challenge track. Note that for
our experiment, it is important not to choose the difficult
examples with any of the methods we study (i.e., VarNet, U-
net, ConvDecoder, and $\ell_1$-norm minimization), because the
goal is to understand whether challenging samples for one
method are also challenging for other methods. We include
a few examples from fastMRI-A and their corresponding 4x-
accelerated reconstructions in the supplementary materials.

Figure 6 shows the performance of the four methods we con-
sider on the fastMRI-A dataset. We plot the performance
on those images as a function of the performance on the
fastMRI dataset, in order to compare the relative perfor-
ance change caused by the distribution shift. For testing
these methods on the fastMRI-A dataset, we used the same
set of validation images that we chose in Section 5.2.

There is a constant gap between the linear interpolation
shown in Figure 6 and $y = x$ which demonstrates that all
four reconstruction methods are equally sensitive to a shift to
naturally challenging samples. Thus, we again find (see Figure
6) that the performance of all methods drops by a
similar amount under this distribution shift—meaning
that neither un-trained nor trained methods degrade or
shine on the difficult-to-reconstruct images. This estab-
ishes that there are difficult images to reconstruct, on which
all methods perform worse than on an average image.

Moreover we find that challenging images are naturally
difficult to reconstruct, since both trained and un-trained
methods are equally prone to this shift. Our hypothesis for
this natural difficulty is that fastMRI-A samples contain
more high-frequency information compared to an average
fastMRI example, and thus are harder to reconstruct
because less high-frequency information is available in the
measurements. The experiments in the supplement confirm
this hypothesis.

6. Recovering small features in an image

For medical applications, it is important to recover small
details of an image, because such details can be critical for
Measuring Robustness in Deep Learning Based Compressive Sensing

$\ell_1$ minimization  U-net  VarNet  ConvDecoder  ground truth

Figure 5: Anatomy shift from knee to brain. End-to-end variational network (VarNet) and U-net cannot remove under-sampling aliasing artifacts for some images when being trained on knee and tested on brain. $\ell_1$-norm minimization (an un-trained method), also induces aliasing artifacts (irrespective of the distribution shift) similar to most of the traditional CS methods. The ConvDecoder (an un-trained network) is more stable in this setup.

Figure 6: Challenging data points are naturally challenging and lower scores for these data points are not due to learning. Trained and un-trained methods perform equally poorly on these samples as there is a constant gap between $y = x$ and the best linear fit.

In this section, we study the ability of the four methods to recover such small details. We first propose a simple framework based on artificial features in order to determine whether there is any location dependency for a given reconstruction method when recovering the small feature. Furthermore, we evaluate the ability to recover small features on a set of 22 annotated fastMRI knee images which contain real-world pathologies.

Artificial feature recovery. Our proposed framework for small feature recovery is as follows. We take a small window consisting of $3 \times 3$ pixels, fill it with the maximum value of a pixel in the image and slide it through the image to have this spot in different locations. We then generate a measurement from this perturbed image and perform the reconstruction using the four mentioned methods. We then measure the error in reconstructing this feature only (i.e., we compute the Mean-Squared Error (MSE) for the $3 \times 3$ region). We performed this experiment for every location in a test image to understand whether there is a location dependency for different methods.

The results, displayed in Figure 7, show that different reconstruction methods are sensitive to errors at different regions in the image. In Figure 7, we see that $\ell_1$-minimization and VarNet perform worse in dark regions of the image (note that the feature itself is bright). Note that this location dependency is not due to learning, in that both trained and un-trained methods have location dependent recovery performance.

Real-world feature recovery. Next, we study natural features by performing an evaluation on a set of 22 annotated images (Cheng et al., 2020) from the fastMRI knee dataset which contain real-world pathologies. We study the performance of those methods in recovering the small features by measuring the MSE only for the region in which the feature is located. The results, depicted in Figure 8, show that the ranking of the methods in terms of small-feature-recovery performance is VarNet > U-net > ConvDecoder > $\ell_1$-norm minimization, which coincides with the ranking in terms of overall reconstruction quality. The figure also shows that the reconstruction quality is perfectly linearly correlated with accurate recovery of fine details, which is intuitively expected.

With regards to prior works on small feature recovery, Cheng et al. (2020) made the first step toward this direction by proposing an optimization framework to synthesize small features for trained networks, by generating different random perturbations of small spatial extent and selecting the worst one for a given method. However, we found it difficult to build a comparison among different methods based on such framework, mainly because of the reasons we discussed for optimization-based perturbations in Section 4 (i.e., (i) worst-case features are not guaranteed, and (ii) hyper-parameters are different across methods).

7. Conclusion

Deep-learning-based image reconstruction methods yield high-quality reconstructions from under-sampled data. However, recent works raised the concern that the improvements in reconstruction quality come at a loss in robustness. In
Measuring Robustness in Deep Learning Based Compressive Sensing

This paper, we have studied the robustness (i) against small adversarial perturbations, (ii) to distribution shifts, and (iii) in recovering fine details, for three families of MRI reconstruction methods: trained deep networks, un-trained deep networks, and classical sparsity-based approaches.

We find that both deep-learning-based as well as classical sparsity-based image reconstruction methods are sensitive to small, adversarially-selected perturbations.

Moreover, the reconstruction quality is correlated with small feature recovery, and hence improving overall reconstruction performance also improves performance for recovering fine details of an image.

Finally, we find that the performance drop under each of three different realistic distribution shifts is similar for all considered trained and un-trained methods. The out-of-distribution accuracy is linearly correlated with the in-distribution accuracy, and the performance ranking of the methods typically remains accurate even under distribution shifts. This is perhaps surprising, because un-trained methods only depend on the training distribution through hyper-parameter tuning.

To improve performance in practice, it is important overcome the performance drop due to distribution shifts: Distribution shifts occur in practice and incur a significant loss in performance. For un-trained methods, very little data for hyper-parameter tuning is needed and thus the performance gap may be closed via access to only few images of the new domain, and it might even be possible to do hyper-parameter tuning on a single under-sampled measurement (Zalbagi Darestani & Heckel, 2020).

For trained methods, this problem might be addressed through larger and more diverse datasets, or through data-augmentation (Fabian et al., 2021). However, distribution shifts are difficult to overcome: Taori et al. (2020) finds for image classification that robustness enhancing methods—apart from training on large and more diverse datasets—help little for natural distribution shifts, and conclude that “distribution shifts arising in real data are currently an open research problem.”

In a nutshell, the take-away of our study is that the deep learning methods that perform best based on reconstruction quality are also best under realistic distribution shifts and for small feature recovery, and we could not find them to be more sensitive to adversarial perturbations.

Reproducibility

The code to reproduce all results in this paper is available at https://github.com/MLI-lab/Robustness-CS.

Acknowledgment

M. Zalbagi Darestani and R. Heckel are (partially) supported by NSF award IIS-1816986, and R. Heckel acknowledges support of the IAS at TUM and the DFG. R. Heckel also thanks Ludwig Schmidt for discussions on distribution shifts, and particular on linear fit phenomena. A. Chaudhari is supported by National Institutes of Health (NIH) grants NIH R01 AR063643, R01 EB002524, K24 AR062068, P41 EB015891, received research support from GE Healthcare and Philips, provided consulting services to Skope MR Inc, Subtle Medical, Image Analysis Group, Edge Analytics, Culvert Engineering, and Chondrometrics GmBH, and is a share-holder of Subtle Medical, LVIS Corporation, and Brain Key.
References


