Localized adversarial artifacts for compressed sensing MRI

R. Alaifari and G. Alberti and T. Gauksson

Localized adversarial artifacts for compressed sensing MRI

Rima Alaifari*   Giovanni S. Alberti†   Tandri Gauksson*

Abstract

As interest in deep neural networks (DNNs) for image reconstruction tasks grows, their reliability has been called into question (Antun et al., 2020; Gottschling et al., 2020). However, recent work has shown that compared to total variation (TV) minimization, they show similar robustness to adversarial noise in terms of ℓ^2-reconstruction error (Genzel et al., 2022). We consider a different notion of robustness, using the ℓ^∞-norm, and argue that localized reconstruction artifacts are a more relevant defect than the ℓ^2-error. We create adversarial perturbations to undersampled MRI measurements which induce severe localized artifacts in the TV-regularized reconstruction. The same attack method is not as effective against DNN based reconstruction. Finally, we show that this phenomenon is inherent to reconstruction methods for which exact recovery can be guaranteed, as with compressed sensing reconstructions with ℓ^1- or TV-minimization.

1   Introduction

Following the success of deep learning in computer vision, deep neural networks (DNNs) have now found their way to a wide range of imaging inverse problems [3, 18, 19]. In some applications, learning the distribution of images from data is the only option. In others, existing methods based on hand-crafted priors are well established. Magnetic resonance image (MRI) reconstruction, for which sparsity based methods have been highly successful, is an example of the latter [16]. However, recent work suggests that image quality can be improved and computation times shortened significantly by the use of DNNs in MRI reconstruction [9].

At the same time, it is well known that DNNs trained for image classification admit so called adversarial examples – images that have been altered in minor but very specific ways to change the label predicted by the network [4, 21]. In [2], it was discovered that DNNs used in inverse problems (MRI and CT) exhibit similar behaviour. Namely, the authors show that perturbing the measurements slightly can lead to undesirable artifacts in the image reconstructed by the network, and that the same perturbations do not cause problems for state-of-the-art compressed sensing methods. On the other hand, [12] shows quantitatively that DNNs can be made robust, to a comparable level with total variation (TV) minimization, by injecting statistical noise to the measurement data during training. Here, robustness is measured by the mean relative reconstruction error as a function

---

*Seminar for Applied Mathematics, Department of Mathematics, ETH Zürich (rima.alaifari@math.ethz.ch, tandri.gauksson@math.ethz.ch).

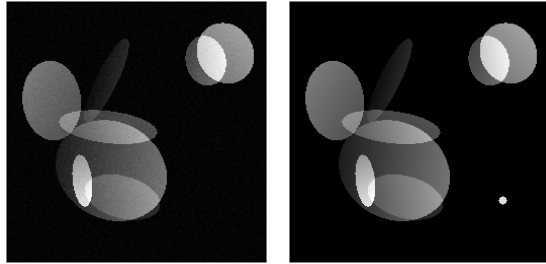†MaLGa Center, Department of Mathematics, University of Genoa (giovanni.alberti@unige.it).

Figure 1: The same image corrupted by two different perturbations of equal ℓ^2-norm (10% of the norm of the original image): Gaussian noise (left) and a potentially meaningful feature (right). The goal of this work is to find small adversarial perturbations to MRI measurements, such that artifacts in the spirit of the latter appear in the reconstruction process.

of relative (adversarial or statistical) noise level, where both are defined by the ℓ^2-norm. Although DNNs and TV-regularized reconstruction behave similarly to adversarial perturbations from this quantitative perspective, the reconstruction artifacts are qualitatively very different. TV minimization suffers from global degradation of image quality (due to staircasing effects), while DNNs tend to introduce new meaningful features to the image. The latter type of artifacts is arguably worse, but it is only severe at relatively high noise levels and the introduced features can already be seen embedded in the adversarial noise and are therefore not created by the network.

In the current work, we focus on the two-dimensional compressed sensing MRI problem [16] and aim to create adversarial perturbations for TV-regularization that result in more localized reconstruction artifacts as displayed in Figure 1. To this end, we simply modify the attack method used in [12] by replacing the loss function with a weighted seminorm, with a weight vector supported on a targeted location. Several locations are targeted, and the one on which the largest artifact appears is selected. Our experimental results show that such localized adversarial perturbations for TV-regularization do exist, and that the effects on DNNs are milder. The resulting artifacts are often manifested as isolated spikes in the image, and in the penultimate section we provide a mathematical justification for the appearance of these artifacts, based on the theory of compressed sensing. Curiously enough, the same positive compressed sensing results on undersampled MRI guaranteeing exact recovery of sparse signals have a negative counterpart regarding the appearance of sparse artifacts. While the analysis we provide is in the context of MRI, the motivations behind this phenomenon are more general and are mainly due to the fact that the forward operator in undersampled MRI has a nontrivial kernel (see also [14]), which happens in many inverse imaging problems.

We note that in regularized reconstruction for imaging applications, the standard benchmark for stability is, and always has been, a quantitative one described by a norm estimate in some chosen (appropriate) norm. At the same time, it is interesting to see the influence of the adversarial attacks research on DNN based image reconstruction. It seems a valid argument to point out that suitable attacks can lead to artifacts that are qualitatively relevant: after all, the objects of interest are images. So to us, this also opens the discussion about benchmarks for classical regularization algorithms. One might argue that also in that case, qualitative benchmarks (which are of course harder to define) should play a relevant

role as well.

The paper is structured as follows: In Section 2 we introduce the compressed sensing MRI model, in Section 3 we formulate our adversarial attack method, in Section 4 we present the results of our numerical experiments, in Section 5 we give a mathematical explanation for our observations, and in Section 6 we make concluding remarks.

The code for the experimental part of the paper builds on the code from [12], and is available on `https://gitlab.math.ethz.ch/tandrig/localadvmri`.

2 Compressed sensing MRI

The goal of magnetic resonance imaging (MRI) is to recover an object's density from its Fourier coefficients. In the fully sampled case, this problem is readily solved by applying the inverse Fourier transform. However, acquisition times can be reduced significantly by undersampling along non-Cartesian trajectories in the frequency domain. This leads to an underdetermined linear system, and it is clear that some additional assumptions need to be made on the object density in order to get a good reconstruction [16].

Let us model the cross section of an object by an image of resolution $n \times n$ for some $n \in \mathbb{N}$. Let $\mathcal{F} \colon \mathbb{C}^{n \times n} \to \mathbb{C}^{n \times n}$ be the two-dimensional discrete Fourier transform, and let $\mathrm{P}_\Omega \colon \mathbb{C}^{n \times n} \to \mathbb{C}^m$ be a projection onto a set of $m \geq 1$ indices, $\Omega \subseteq \{(i,j) \colon i,j = 1, \ldots, n\}$. The forward operator of the subsampled MRI problem is $\mathcal{A} = \mathrm{P}_\Omega \circ \mathcal{F}$, and the measured values corresponding to an image $x \in \mathbb{C}^{n \times n}$ are

$$y = \mathcal{A}x + \epsilon, \tag{1}$$

where ϵ is zero-mean random noise which we assume is bounded in ℓ^2-norm by some constant $\eta > 0$. In order to recover x from the measurement vector, we must search for an image z with

$$\|\mathcal{A}z - y\|_2 \leq \eta.$$

Since $m < n^2$, the equation $y = \mathcal{A}z$ does not have a unique solution. The pseudoinverse $\mathcal{A}^\dagger = \mathcal{F}^{-1} \circ \mathrm{P}_\Omega^\dagger$ certainly provides a solution, but the resulting image may be of low quality or exhibit severe aliasing artifacts (depending on Ω), and not at all represent the desired object's density (see Figure 2). Instead, we must impose some conditions on z based on our a priori knowledge of x.

Compressed sensing (CS) refers to the approach of favoring solutions that are sparse under some given transform Ψ. Under certain conditions, the sparsest solution is also the one with the smallest ℓ^1-norm [11], and hence we are left with the following convex optimization problem:

$$\min_{z \in \mathbb{C}^{n \times n}} \|\Psi z\|_1 \text{ subject to } \|\mathcal{A}z - y\|_2 \leq \eta. \tag{2}$$

Several choices of the transform Ψ can be found in the CS-MRI literature, including the identity, the wavelet transform, and the image gradient [16]. In this work, we focus on the last one of these choices. More precisely, we choose $\Psi = \nabla$, where $\nabla \colon \mathbb{C}^{n \times n} \to \mathbb{C}^{n \times n} \times \mathbb{C}^{n \times n}$ is the two-dimensional finite difference operator on $\mathbb{C}^{n \times n}$ with periodic boundary conditions. The quantity $\|\nabla z\|_1$ is known as the *total variation* (TV) of the image z. Minimization of the total variation promotes sparse gradients, and hence piecewise constant solutions. For computational efficiency, we solve the unconstrained formulation of (2):

$$\mathrm{rec}_{\mathrm{TV}}(y; \eta) = \operatorname*{arg\,min}_{z \in \mathbb{C}^{n \times n}} \|\mathcal{A}z - y\|_2^2 + \lambda_\eta \|\nabla z\|_1, \tag{3}$$
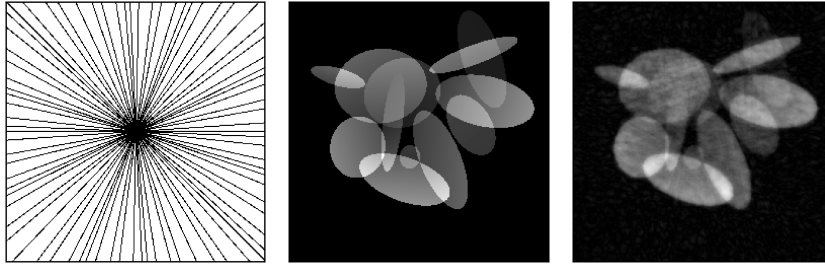
3

Figure 2: Left: The sampling mask Ω used in the experiments, comprising 40 lines through the origin. The retained frequencies are shown in black. Middle: An example image x. Right: A low quality linear reconstruction, using the pseudoinverse of the measurement matrix $\mathcal{A}^\dagger \mathcal{A} x$.

where $\lambda_\eta > 0$ is a regularization parameter. An appropriate choice of λ_η ensures that $\mathrm{rec}_{\mathrm{TV}}(y;\eta)$ is a solution to (2) [11].

2.1 Deep neural networks as an alternative to CS

Another approach to the MRI problem is to learn the inverse mapping $\mathbb{C}^m \to \mathbb{C}^{n\times n}$ from fully sampled data, i.e. to replace $\mathrm{rec}_{\mathrm{TV}}$ by a neural network. Several different strategies appear in the literature. Fully learned networks learn the entire inversion without any knowledge of the forward model at all [24], while others use the linear operator $\mathcal{A}^\dagger \colon \mathbb{C}^m \to \mathbb{C}^{n\times n}$ as a first layer so that only the post-processing step $\mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is learned [22]. The forward model can also be incorporated at several stages in the network, as in networks that are based on unrolled iterative optimization algorithms [23].

For our experiments, we use the networks Tira and TiraFL from [12], which are based on the fully convolutional Tiramisu architecture [15]. The difference between the two networks lies in that TiraFL is a fully learned network $\mathbb{C}^m \to \mathbb{C}^{n\times n}$, while Tira applies the pseudoinverse $\mathcal{A}^\dagger$ to the measurements first. Both networks are trained using the mean squared error as a loss function, and Gaussian noise is added to the input as a means of regularization.

3 Adversarial perturbations

The study of adversarial examples originates in image classification. In that context, given an image and a classifier, an *adversarial perturbation* is one that is imperceptible when added to the image but changes the output of the classifier. The perturbed image is called an *adversarial example*. The imperceptibility of the perturbation is difficult to define, so this requirement is commonly replaced by a bound on the norm of the perturbation. A popular choice is the ℓ^∞-norm [13, 17], since if each pixel in the image is changed only by a small value, then one can be sure that the semantic meaning of the image stays the same. In contrast, the meaning of an image may be changed (say a handwritten 1 to a 7) by introducing a localized perturbation with a small ℓ^2-norm. It should be noted that there exist other adversarial image transformations that are not small in the ℓ^∞-norm, such as rotations, translations, or smooth deformations [1].

Two difficulties arise when adapting the notion of adversarial examples to MRI. Firstly,

since measurements are in the frequency domain, the imperceptibility of perturbations is not very meaningful. Secondly, since the output of the reconstruction method is a continuous variable (as opposed to discrete labels in classification), a notion of severity is needed to quantify the effect of an adversarial perturbation.

It is natural to tackle the first problem by referring to the noise model, i.e. a perturbation is "imperceptible" if it is small in ℓ^2-norm, as is done in both [2] and [12]. For the second problem, [2] inspects the visual quality of the reconstructed images, while [12] uses the ℓ^2-error for quantitative analysis of stability to perturbations. Thus, for given measurements y and noise level η, [12] defines an adversarial perturbation e by

$$e = \arg\max_{e \in \mathbb{C}^m} \|\text{rec}_{\text{TV}}(y + e; \eta) - \text{rec}_{\text{TV}}(y; \eta)\|_2 \quad \text{such that} \quad \|e\|_2 \leq \eta \quad (4)$$

(and in fact, [2] solves a similar unconstrained optimization problem). We refer to

$$\rho = \text{rec}_{\text{TV}}(y + e; \eta) - \text{rec}_{\text{TV}}(y; \eta) \quad (5)$$

as the *reconstruction artifact* induced by the perturbation e.

In this work, we argue that the reconstruction ℓ^2-error does not sufficiently capture the most harmful reconstruction artifacts. In the medical setting, a misdiagnosis might be based on a localized anomaly in the image rather than an overall poor quality of reconstruction (see Figure 1). Thus we aim to create perturbations that cause localized reconstruction artifacts. We replace the Euclidean norm in the objective with a weighted seminorm:

$$e = \arg\max_{e \in \mathbb{C}^m} \|\phi \odot (\text{rec}_{\text{TV}}(y + e; \eta) - \text{rec}_{\text{TV}}(y; \eta))\|_2 \quad \text{such that} \quad \|e\|_2 \leq \eta, \quad (6)$$

where $\odot$ denotes entrywise product and $\phi \in \mathbb{C}^{n \times n}$ is a weight vector. To promote localized artifacts, we select a weight vector that is the (discrete) indicator function of a disk of radius $\sigma > 0$, centered at $\mu = (\mu_1, \mu_2) \in [1, n]^2$. In other words, we let $\phi = \phi^{\mu, \sigma} \in \mathbb{C}^{n \times n}$ with

$$\phi_{ij}^{\mu, \sigma} = \begin{cases} 1 & \text{if } (i - \mu_1)^2 + (j - \mu_2)^2 \leq \sigma^2, \\ 0 & \text{otherwise,} \end{cases}$$

for $i, j = 1, \ldots, n$.

Although we target a specific location μ in the image, we are interested in any large artifacts that may appear in reconstruction. Therefore, we judge the severity of the reconstruction artifact ρ by its ℓ^∞-norm. We solve (6) for all locations μ on a regular grid in $[1, n]^2$, and select the perturbation e_μ that maximizes

$$\|\text{rec}_{\text{TV}}(y + e_\mu; \eta) - \text{rec}_{\text{TV}}(y; \eta)\|_\infty$$

as our adversarial perturbation. The radius, σ, is fixed at a small value relative to n. Note that for a large enough radius, (6) is equivalent to (4).

3.1 Visualizing perturbations

While we search for reconstruction artifacts by perturbing the measurement vector $y = \mathcal{A}x$, it is important to consider whether the artifact is in some sense already encoded in the perturbed measurements $y + e$, rather than "invented" by the reconstruction method. Indeed, in some cases shown in [12] (see for example Fig. 7 therein) the same perturbation induces
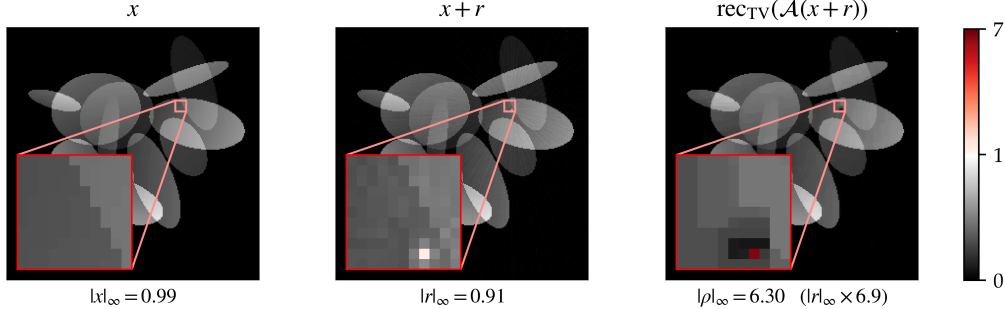
Figure 3: Left: The original image. Middle: The perturbed image. A close up of the selected location, μ, shows that the image perturbation, $r = \mathcal{A}^\dagger e$, introduces a spike. Right: The TV-reconstruction of the perturbed measurements. The spike is greatly amplified. Note that pixel values from 0 to 1 are shown in grayscale, while values beyond 1 are shown in red (the colorbar on the right is not linear).

similar artifacts for both TV-regularized reconstruction and DNN-based reconstruction, indicating that these artifacts are present in the perturbation itself and are not created in the process of reconstruction.

To see if the perturbation $e \in \mathbb{C}^m$ encodes an artifact in this way, we visualize it by an image perturbation $r \in \mathbb{C}^{n \times n}$ such that $\mathcal{A}r = e$. We select $r = \mathcal{A}^\dagger e$, as this perturbation is orthogonal to $\ker \mathcal{A}$, which means that $\|r\|_2 = \|e\|_2$. Then we can compare the perturbed image $x + r$ with the reconstruction $\text{rec}_{\text{TV}}(y + e; \eta) = \text{rec}_{\text{TV}}(\mathcal{A}(x + r); \eta)$, both visually and in terms of the ℓ^∞-norm.

4 Results

We demonstrate the strategy described above by creating adversarial perturbations for synthetically generated 256×256 images of phantom ellipses (as in [12]) with pixel values ranging from 0 to 1. To generate the measurement vectors, we apply the MRI forward operator $\mathcal{A} = \text{P}_\Omega \circ \mathcal{F}$ with a sampling mask Ω defined by 25, 40, or 80 lines through the origin (see Figure 2), which corresponds to 11%, 17%, or 32% of the coefficients. We train the neural networks, Tira and TiraFL, only on the data sampled on 40 lines. Moreover, all the figures in this paper are based on measurements using that same mask.

The reconstruction map in (3), rec_{TV}, is realized by the alternating direction method of multipliers (ADMM) [5], and the implementation is taken from [12]. For each noise level η, the parameter λ_η is selected by grid search (along with the one free parameter of ADMM), averaged over 50 samples of the ellipse data.

We create adversarial perturbations for 50 ellipse images according to (6) at relative noise levels $(\eta / \|y\|_2)$ ranging from 0.5% to 10%. For each noise level, and each image, we search for the best position of localization weight vector, μ, on an 8×8 grid. The radius of the localization disk is fixed at $\sigma = 5$. Experimentation showed that similar reconstruction artifacts appear for other values of σ, as long as it is small. For larger σ, the artifacts are not well localized and resemble those of [12].

Figure 3 shows an example image x, a perturbed image $x + r$ (where $r = \mathcal{A}^\dagger e$ is a visual representation of the perturbation), and the reconstruction from perturbed measurements
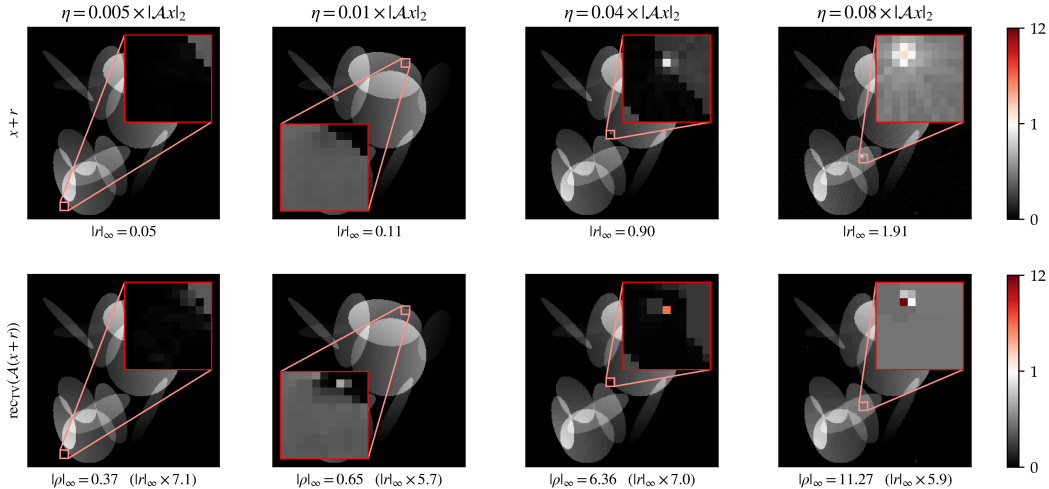
6

Figure 4: Adversarial attacks at 0.5%, 1%, 4%, and 8% relative noise. Upper row: the original image with the image perturbation added. Lower row: TV-regularized reconstruction of the perturbed measurements. The reconstruction artifact ρ is consistently much larger than the perturbation r in the ℓ^∞-norm.

$\text{rec}_{\text{TV}}(y + e; \eta)$. The output of the reconstruction method is a vector in $\mathbb{C}^{256 \times 256}$, which we visualize by taking the entrywise modulus. We now list some observations based on this quantitative experiment and visual inspection of examples.

TV-regularized reconstruction amplifies artifacts already present in the perturbation: We observe that the adversarial perturbations indeed induce localized reconstruction artifacts in the form of a spike. A spike is also noticeably present in the image perturbation, r, itself. The image perturbation is therefore *not* imperceptible in the ℓ^∞-sense, except at very low noise levels. However, in ℓ^∞-norm, the reconstruction artifact ρ from (5) is much larger than the perturbation. We consider the *amplification factor*

$$\alpha = \|\rho\|_\infty \, / \, \|r\|_\infty$$

to quantify this effect. In Figure 4, we observe a large α at different noise levels, but visually, this phenomenon is especially conspicuous at higher noise levels.

The amplification factor is roughly equal to the subsampling factor: Table 1 shows the average amplification factor of the 50 ellipse samples for each noise level. Although the amplification factor varies more at low noise levels, for TV-regularization its average does not seem to depend on the noise level. There is however a clear dependence on the number of coefficients in the measurements. In fact, for reasons that will become clear in the next section, the average amplification factor is approximately the subsampling factor n^2/m, which for the 25, 40, and 80 line sampling masks is approximately 9.4, 6.0, and 3.1.

The amplification factor is smaller for DNNs: Applying the attack strategy (6) to the neural networks Tira and TiraFL does produce reconstruction artifacts. However, at $> 1\%$ noise, the attack is far less effective than for rec_{TV}. Figure 5 shows a typical artifact created by Tira, which has a much lower ℓ^∞-norm than those created by rec_{TV}.

Perturbations made for DNNs transfer to rec_{TV}, but not vice versa: In Figure 6, we see that a perturbation created for a DNN can transfer to all three reconstruction

$\eta/\,\|y\|_2$	$\mathrm{rec_{TV}}$ (25 lines)	$\mathrm{rec_{TV}}$ (40 lines)	$\mathrm{rec_{TV}}$ (80 lines)	Tira	TiraFL
0.5%	8.12 ± 2.83	6.23 ± 2.16	3.15 ± 0.15	7.29 ± 5.00	7.36 ± 3.83
1.0%	9.05 ± 2.56	5.64 ± 2.24	3.55 ± 0.26	5.91 ± 4.76	5.67 ± 4.40
2.0%	9.05 ± 2.40	5.88 ± 0.40	3.22 ± 0.19	3.20 ± 1.53	3.59 ± 2.61
4.0%	9.06 ± 0.43	6.88 ± 0.56	3.19 ± 0.10	2.38 ± 0.81	2.34 ± 0.69
8.0%	9.75 ± 0.52	5.91 ± 0.28	3.07 ± 0.07	1.89 ± 0.44	1.96 ± 0.58
10.0%	8.80 ± 0.71	5.65 ± 0.42	2.99 ± 0.11	1.80 ± 0.48	1.77 ± 0.60

Table 1: The amplification factor $\alpha = \|\rho\|_\infty / \|r\|_\infty$ across different relative noise levels, averaged over 50 samples from the ellipse data set. For $\mathrm{rec_{TV}}$, we reconstruct from measurements based on 25, 40, and 80 sampling lines in the frequency domain. The DNNs are tested on the 40 lines sampling mask only.
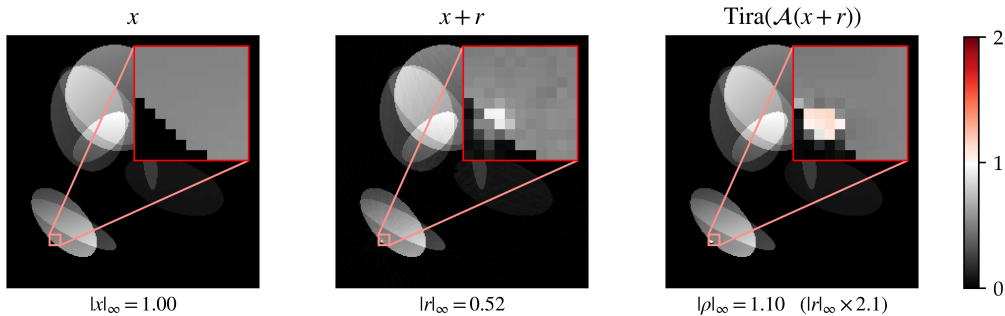

Figure 5: Analogous to Figure 3, with the reconstruction method $\mathrm{rec_{TV}}$ replaced by the DNN Tira. Here, the amplification factor is $\alpha = 2.13$.

methods, with $\mathrm{rec_{TV}}$ showing the most severe artifact. On the other hand, using the DNNs to reconstruct from $y + e$, where e is a perturbation created for TV-regularized reconstruction, produces images of good quality and no visible artifacts (Figure 7). In fact, the peak of the image perturbation r is dampened by the DNN reconstruction.

5   Explaining localized artifacts

Many results exist that guarantee exact recovery of sparse vectors from partial Fourier measurements (1), depending on properties of the underlying signal, the sampling mask Ω, and the reconstruction method. The reconstruction artifacts seen in Figures 3 and 4 consist mainly of a single pixel spike. We now show how such exactness guarantees can imply the existence of low ℓ^∞-norm perturbations which give rise to precisely that type of artifacts.

For simplicity, we consider the noiseless one-dimensional setting. Let $n \geq 2$ be an even integer and let $\mathcal{F}_n \colon \mathbb{C}^n \to \mathbb{C}^n$ denote the one-dimensional discrete Fourier transform,

$$\mathcal{F}_n(z) = \left(\frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} z_j \mathrm{e}^{-2\pi \mathrm{i} k j / n} \right)_{k=-n/2+1}^{n/2}.$$

Let $\Omega_n \subseteq \{-n/2+1, \ldots, n/2\}$ be a set of indices, and define the operator $\mathcal{A}_n = \mathrm{P}_{\Omega_n} \circ \mathcal{F}_n$, where P_{Ω_n} is the projection onto the index set Ω_n. Consider a signal $x_n \in \mathbb{C}^n$. We
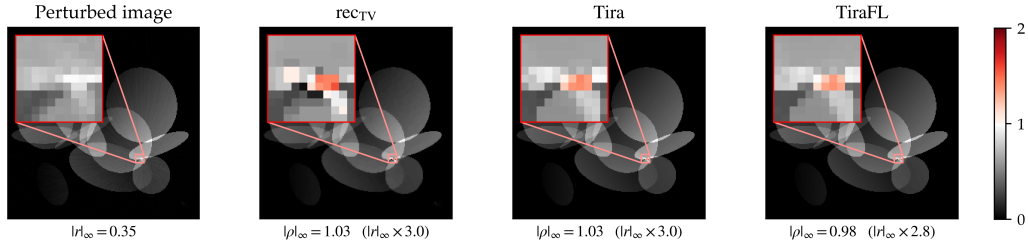
Figure 6: An adversarial perturbation is created with respect to Tira, and then rec$_{\text{TV}}$, Tira, and TiraFL are used to reconstruct from the perturbed measurements. All methods show similar reconstruction artifacts.
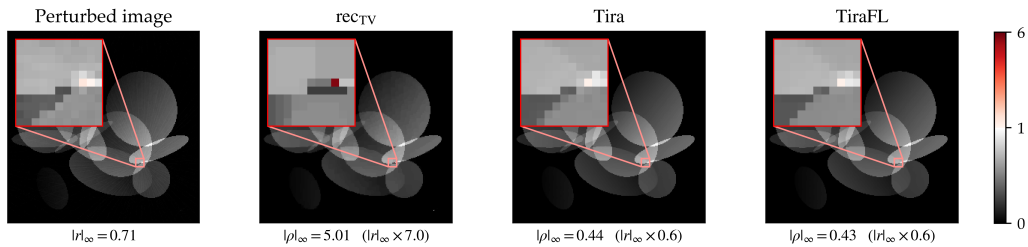

Figure 7: Analogous to Figure 6, with the perturbation constructed with respect to rec$_{\text{TV}}$. The reconstruction artifact in the second image does not appear in the second and third images, and hence the perturbation is not transferable.

wish to perturb the measurement vector $y_n = \mathcal{A}_n(x_n)$ such that a spike appears in the reconstruction. Without loss of generality, let $\delta_n = (1, 0, \ldots, 0) \in \mathbb{C}^n$ be our desired single spike artifact, and simply define the perturbation

$$e_n = \mathcal{A}_n(\delta_n) = \left(1/\sqrt{n}\right)_{k \in \Omega_n}$$

(which, incidentally, is an ℓ^∞-minimal perturbation given an ℓ^2-budget). The corresponding "image perturbation" is

$$r_n = \mathcal{A}_n^\dagger(e_n) = \left(\frac{1}{n} \sum_{k \in \Omega_n} e^{2\pi \mathrm{i} k j / n}\right)_{j=0}^{n-1}.$$

If we can guarantee the exact recovery of x_n from $\mathcal{A}_n x_n$ and of $x_n + \delta_n$ from its measurements (see Theorems 5.1 and 5.2 below),

$$\mathcal{A}_n(x_n + \delta_n) = y_n + e_n = \mathcal{A}_n(x_n + r_n),$$

then a reconstruction artifact is created with an amplification factor $\alpha_n = \|\delta_n\|_\infty / \|r_n\|_\infty = 1 / \|r_n\|_\infty$. The ℓ^∞-norm of r_n can easily be bounded from above:

$$\|r_n\|_\infty = \max_{j=0,\ldots,n-1} \left| \frac{1}{n} \sum_{k \in \Omega_n} e^{2\pi \mathrm{i} k j / n} \right| \leq \max_{j=0,\ldots,n-1} \frac{1}{n} \sum_{k \in \Omega_n} \left| e^{2\pi \mathrm{i} k j / n} \right| = \frac{m_n}{n},$$

9

where $m_n = |\Omega_n|$, so that

$$\alpha_n \geq \frac{n}{m_n}.$$

In other words, the amplification factor is at least as big as the subsampling factor, which is in accordance with Table 1. If the size of the sampling mask Ω_n grows slowly with the dimension n, then the amplification factor α_n can be made arbitrarily large simply by increasing the resolution. In fact, since

$$\|r_n\|_2 = \|e_n\|_2 = \sqrt{m_n/n}, \quad \|\delta_n\|_2 = 1,$$

the same is true for amplification in the ℓ^2-norm, although the growth is at a slower rate.

For demonstration purposes, we now cite a known recovery guarantee for ℓ^1-minimization, which can then be translated to results on TV-minimization by a simple argument. The number of measurements required depends on the sparsity of the signal, which we shall assume does not vary with the dimension. This is true if x_n is a finite spike train, and thus suitable for ℓ^1-minimization. We employ TV-minimization if x_n is a piecewise constant function, in which case the sparsity of ∇x_n is constant with respect to n. Moreover, if x_n is s-sparse, then $x_n + \delta_n$ is $(s+1)$-sparse. Similarly, if ∇x_n is s-sparse, then $\nabla(x_n + \delta_n)$ is $(s+2)$-sparse (where ∇ is understood as the periodic finite difference operator). Hence, exactness results hold for $x_n + \delta_n$ as well as x_n.

Theorem 5.1 ([7, Theorem 1.1] as stated in [11, Theorem 12.20]). *Let $x \in \mathbb{C}^n$ be s-sparse, $\varepsilon > 0$, and suppose that the indices of Ω_n are chosen uniformly at random from $\{-n/2+1, \ldots, n/2\}$. Then, there exists a constant $C > 0$ such that if*

$$m_n \geq Cs \log(n) \log(\varepsilon^{-1}),$$

then x is the unique solution to

$$\min_{z \in \mathbb{C}^n} \|z\|_1 \;\; subject\;to\;\; \mathcal{A}_n z = \mathcal{A}_n x,$$

with probability at least $1 - \varepsilon$.

Keeping ε at a small but constant value, and using the minimal number of measurements means that $m_n \lesssim \log n$. Applying Theorem 5.1 to $x = x_n + \delta_n$, we see that ℓ^1-minimization combined with this uniformly random sampling scheme leads to amplification of

$$\alpha_n \gtrsim \frac{n}{\log n} \xrightarrow[n \to \infty]{} \infty,$$

with high probability. In the same way, the corresponding TV-minimization result implies that TV-minimization for gradient sparse signals also leads to amplification of $\alpha_n \gtrsim n/\log n$.

Theorem 5.2. *Let $x \in \mathbb{C}^n$ be such that ∇x is s-sparse, $\varepsilon > 0$, and suppose that $\Omega_n = \{0\} \cup \Omega'$ where Ω' contains m_n indices chosen uniformly at random from $\{-n/2+1, \ldots, n/2\}$. Then, there exists a constant $C > 0$ such that if*

$$m_n \geq Cs \log(n) \log(\varepsilon^{-1}),$$

then x is the unique solution to

$$\min_{z \in \mathbb{C}^n} \|\nabla z\|_1 \;\; subject\;to\;\; \mathcal{A}_n z = \mathcal{A}_n x,$$

with probability at least $1 - \varepsilon$.

Proof. (The following argument can be used to translate results on recovery from Fourier measurements by ℓ^1-minimization to TV-minimization, and appears, for example, in [8].) First, note that for any $z \in \mathbb{C}^n$, and $k = -n/2 + 1, \ldots, n/2$, we have

$$
\begin{aligned}
(\mathcal{F}_n (\nabla z))_k &= \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \left(z_j - z_{(j-1) \bmod n} \right) \mathrm{e}^{-2\pi \mathrm{i} k j / n} \\
&= \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} (z_j) \, \mathrm{e}^{-2\pi \mathrm{i} k j / n} - \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} (z_j) \, \mathrm{e}^{-2\pi \mathrm{i} k (j+1) / n} \\
&= \left(1 - \mathrm{e}^{-2\pi \mathrm{i} k / n} \right) (\mathcal{F}_n z)_k \, ,
\end{aligned}
$$

and therefore

$$
\mathcal{A}_n z = \mathcal{A}_n x \quad \Longleftrightarrow \quad \mathcal{A}_n (\nabla z) = \mathcal{A}_n (\nabla x) \text{ and } \mathbf{1} \cdot z = \mathbf{1} \cdot x
$$

(where $\mathbf{1} = (1, \ldots, 1)$ and the condition $\mathbf{1} \cdot z = \mathbf{1} \cdot x$ comes from the assumption that $0 \in \Omega_n$).

By successively loosening the constraints, we see that

$$
\begin{aligned}
& \min_{z \in \mathbb{C}^n} \| \nabla z \|_1 \text{ subject to } \mathcal{A}_n z = \mathcal{A}_n x \\
= & \min_{z \in \mathbb{C}^n} \| \nabla z \|_1 \text{ subject to } \mathcal{A}_n (\nabla z) = \mathcal{A}_n (\nabla x) \text{ and } \mathbf{1} \cdot z = \mathbf{1} \cdot x \\
\geq & \min_{z \in \mathbb{C}^n} \| \nabla z \|_1 \text{ subject to } \mathcal{A}_n (\nabla z) = \mathcal{A}_n (\nabla x) \\
\geq & \min_{z \in \mathbb{C}^n} \| z \|_1 \text{ subject to } \mathcal{A}_n z = \mathcal{A}_n (\nabla x) \, .
\end{aligned}
$$

By Theorem 5.1, ∇x is the unique solution to the last of these minimization problems, and therefore x also solves the first one. Uniqueness follows from the fact that the only z which satisfies both $\nabla z = \nabla x$ and $\mathbf{1} \cdot z = \mathbf{1} \cdot x$ is $z = x$. $\qquad \square$

Of course, other sampling schemes exist, and if additional information on the signal is utilized, this result can be improved upon. For example, if the minimal distance between two non-zero entries of $x_n + \delta_n$ is greater than $2n/c$ for some positive integer c, then ℓ^1-minimization recovers $x_n + \delta_n$ exactly as long as the frequencies $\{-c, c+1, \ldots, c\}$ are included in the sampling mask Ω_n [6]. Likewise, if $x_n + \delta_n$ is non-negative and s-sparse, then it suffices that $\{-s, s+1, \ldots, s\} \subseteq \Omega_n$ [10]. In these cases m_n is constant, and therefore $\alpha_n \gtrsim n$. However, we cannot apply these specialized results to our example with TV-minimization, since $\nabla(x_n + \delta_n)$ is neither non-negative nor well separated. It is worth mentioning the main result of [20] on TV-minimization, which uses a random sampling mask that concentrates on the lower frequencies with $m_n \gtrsim \log(n)$.

The single pixel reconstruction artifact may not be considered a meaningful feature, and can be disregarded by a practitioner. However, the artifacts seen in the experiments are more diverse and act as a proof of concept for localized adversarial artifacts for TV-regularization. Other criteria for artifact severity may lead to different results. The spike δ_n is only an idealization of the observed artifacts which maximizes the ℓ^∞-norm within a fixed ℓ^2-budget for the perturbation. A similar argument can be made for the existence of different types of sparse artifacts, but with a lower ℓ^∞-amplification factor. For any artifact $\rho_n \in \mathbb{C}^n$, the corresponding image perturbation can be bounded by the triangle inequality:

$$
\| r_n \|_\infty = \left\| \mathcal{A}_n^\dagger \mathcal{A}_n (\rho_n) \right\|_\infty \leq \| \rho_n \|_1 \frac{m_n}{n} ,
$$

so that an amplification of

$$\alpha_n \geq \frac{\|\rho_n\|_\infty}{\|\rho_n\|_1} \frac{n}{m_n}$$

can be attained. Depending on the sparsity properties of ρ_n, exact recovery of $x_n + \rho_n$ can be guaranteed by the theorems above. For example if $\rho_n = (1, \ldots, 1, 0, \ldots, 0)$ is a discrete rectangular function, then $\nabla \rho_n$ is 2-sparse and Theorem 5.2 can be applied.

6   Conclusion

In this work, we have studied localized adversarial perturbations in the context of MRI, both with TV-based and DNN-based reconstruction methods. In the case of TV-regularization, we have been successful in creating such localized reconstruction artifacts according to our criterion of large amplification of ℓ^∞-norms. Only at low noise levels does the same hold for the DNNs considered. We analysed the artifacts arising with TV, and showed that they are inherent to the compressed sensing MRI problem. While exact recovery of sparse signals from incomplete Fourier data is generally considered a positive result, we offer the perspective that it also guarantees the existence of adversarial perturbations. Keeping the ℓ^∞-norm of the image perturbation constant, the ℓ^∞-norm of the resulting reconstruction artifact is proportional to the subsampling factor. Thus the effect becomes more pronounced in high dimensions where proportionally fewer measurements may be used.

The results presented in this paper suggest that the ℓ^2-norm of the reconstruction error may not be a significant measure of the perturbation created, since localized artifacts have a very small ℓ^2 norm but a large ℓ^∞ norm, thereby yielding perceptible perturbations. It is natural to wonder whether it is possible to design a measure of the perturbations that could be both quantitatively computable and qualitatively meaningful. Furthermore, it would be interesting to generalize the experimental results of Section 4 and the mathematical insights of Section 5 to different kinds of artifacts and to more general inverse problems.

Acknowledgments

References

[1] R. Alaifari, G. S. Alberti, and T. Gauksson. ADef: an Iterative Algorithm to Construct Adversarial Deformations. In *International Conference on Learning Representations (ICLR 2019)*, 2019.

[2] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

[3] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[4] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[6] E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

[7] E. J. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE transactions on information theory*, 57(11):7235–7254, 2011.

[8] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[9] Y. Chen, C.-B. Schönlieb, P. Liò, T. Leiner, P. L. Dragotti, G. Wang, D. Rueckert, D. Firmin, and G. Yang. AI-based reconstruction for fast MRI—a systematic review and meta-analysis. *Proceedings of the IEEE*, 110(2):224–245, 2022.

[10] Y. De Castro and F. Gamboa. Exact reconstruction using Beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

[11] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer New York, 2013.

[12] M. Genzel, J. Macdonald, and M. März. Solving Inverse Problems With Deep Neural Networks - Robustness Included. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. doi: 10.1109/TPAMI.2022.3148324.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[14] N. M. Gottschling, V. Antun, B. Adcock, and A. C. Hansen. The troublesome kernel: why deep learning for inverse problems is typically unstable. *arXiv preprint arXiv:2001.01258*, 2020.

[15] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 11–19, 2017.

[16] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compressed Sensing MRI. *IEEE Signal Processing Magazine*, 25(2):72–82, 2008. doi: 10.1109/MSP.2007.914728.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[18] M. T. McCann, K. H. Jin, and M. Unser. Convolutional Neural Networks for Inverse Problems in Imaging: A Review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017. doi: 10.1109/msp.2017.2739299. URL `https://doi.org/10.1109%2Fmsp.2017.2739299`.

[19] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020.

[20] C. Poon. On the role of total variation in compressed sensing. *SIAM Journal on Imaging Sciences*, 8(1):682–720, 2015.

[21] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[22] S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, and D. Liang. Accelerating magnetic resonance imaging via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 514–517. IEEE, 2016.

[23] Y. Yang, J. Sun, H. Li, and Z. Xu. Deep ADMM-Net for compressive sensing MRI. *Advances in neural information processing systems*, 29, 2016.

[24] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, 2018.