Improving Robustness of Deep-Learning-Based Image Reconstruction

Ankit Raj [1] **Yoram Bresler** [1] **Bo Li** [2]

Abstract

Deep-learning-based methods for various applications have been shown vulnerable to adversarial examples. Here we address the use of deep-learning networks as inverse problem solvers, which has generated much excitement and even adoption efforts by the main equipment vendors for medical imaging including computed tomography (CT) and MRI. However, the recent demonstration that such networks suffer from a similar vulnerability to adversarial attacks potentially undermines their future. We propose to modify the training strategy of end-to-end deep-learning-based inverse problem solvers to improve robustness. To this end, we introduce an auxiliary network to generate adversarial examples, which is used in a min-max formulation to build robust image reconstruction networks. Theoretically, we argue that for such inverse problem solvers, one should analyze and study the effect of adversaries in the measurement-space, instead of in the signal-space used in previous work. We show for a linear reconstruction scheme that our min-max formulation results in a singular-value filter regularized solution, which suppresses the effect of adversarial examples. Numerical experiments using the proposed min-max scheme confirm convergence to this solution. We complement the theory by experiments on non-linear Compressive Sensing (CS) reconstruction by a deep neural network on two standard datasets, and, using anonymized clinical data, on a state-of-the-art published algorithm for low-dose x-ray CT reconstruction. We show a significant improvement in robustness over other methods for deep network-based reconstruction, by using the proposed approach.

1. Introduction

Adversarial examples for deep learning based methods have been demonstrated for various problems (Szegedy et al., 2013; Kurakin et al., 2016; Cisse et al., 2017a; Eykholt et al., 2017; Xiao et al., 2018), showing that easily obtained minute perturbations can make deep networks produce unexpected results. There has been plethora of work to defend against these attacks as well (Madry et al., 2017; Tramèr et al., 2017; Athalye et al., 2018; Wong et al., 2018; Jang et al., 2019a; Jiang et al., 2018; Xu et al., 2017; Schmidt et al., 2018). Recently, (Antun et al., 2020; Choi et al., 2019) introduced adversarial attacks on image reconstruction networks, but no defences have been proposed. In this work, we address this gap by proposing an adversarial training scheme for image reconstruction deep networks.

Image reconstruction involving the recovery of an image from indirect measurements is used in many applications, including critical applications such as medical imaging, e.g., Magnetic Resonance Imaging (MRI), Computerised Tomography (CT), etc. Such applications demand the reconstruction to be stable and reliable. On the other hand, in order to speed up the acquisition, reduce sensor cost, or reduce radiation dose, it is highly desirable to subsample the measurement data, while still recovering the original image. This is enabled by the compressive sensing (CS) paradigm (Candes et al., 2006; Donoho, 2006). CS involves projecting a high dimensional, signal $x \in \mathbb{R}^n$ to a lower dimensional measurement $y \in \mathbb{R}^m, m \ll n$, using a small set of linear, non-adaptive frames. The noisy measurement model is:

$$y = Ax + v, A \in \mathbb{R}^{m \times n}, \tag{1}$$

where A is the measurement matrix and v is the measurement noise. The goal is to recover the unobserved natural image x, from the compressive measurement y. Although the problem with $m \ll n$ is severely ill-posed and does not have a unique solution, CS achieves nice, stable solutions for a special class of signals x - those that are sparse or sparsifiable, by using sparse regularization techniques (Candes et al., 2006; Donoho, 2006; Elad & Aharon, 2006; Dong et al., 2011; Wen et al., 2015; Liu et al., 2017; Dabov et al., 2009; Yang et al., 2010; Elad, 2010; Li et al., 2009; Ravishankar & Bresler, 2012).

Recently, deep learning-based methods have also been proposed as an alternative method for performing image recon-

[1] Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign (UIUC) [2] Department of Computer Science, UIUC. Correspondence to: Ankit Raj <ankitr3@illinois.edu>, Yoram Bresler <ybresler@illinois.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

struction (Zhu et al., 2018; Jin et al., 2017; Schlemper et al., 2017; Yang et al., 2017; Hammernik et al., 2018). While these methods have achieved state-of-the-art (SOTA) performance, the networks have been found to be very unstable (Antun et al., 2020), as compared to the traditional methods. Adversarial perturbations have been shown to exist for such networks, which can degrade the quality of image reconstruction significantly. (Antun et al., 2020) studies three types of instabilities: ***(i)*** Tiny (small norm) perturbations applied to images that are almost invisible in the original images, but cause a significant distortion in the reconstructed images. ***(ii)*** Small structural changes in the original images, that get removed from the reconstructed images. ***(iii)*** Stability with increasing the number of measurement samples. In this work, we try to address instability (i) above.

We argue that studying the instability for image reconstruction networks in the x-space (Antun et al., 2020) is sub-optimal, and instead, we consider perturbations in the measurement, y-space. For robustness, we modify the training strategy: we introduce an auxiliary network to generate adversarial examples on the fly, which are used in a min-max formulation. This results in an adversarial game between two networks while training, similar to the Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017). However, since the goal here is a robust reconstruction network, the training strategy is modified.

Our theoretical analysis for a special case of a linear reconstruction scheme shows that the min-max formulation results in a singular-value filter regularized solution, suppressing the effect of adversarial examples. For a linear reconstruction network, our experiment using the min-max formulation with a learned adversarial example generator confirms convergence to the theoretically-obtained solution. For a complex non-linear deep network, our experiments show that compared to other methods, the proposed training scheme results in a more robust network, both qualitatively and quantitatively. Further, experiments and analysis show qualitatively different behavior as a function of the conditioning of the measurement matrix. Finally, the practical significance of the proposed formulation is demonstrated in experiments using clinical CT data with the FBPConvNet (Jin et al., 2017) for low dose CT reconstruction, where we again achieve significant improvement in robustness.

2. Proposed Method

2.1. Adversarial Training

One of the most powerful methods for training an adversarially robust network is adversarial training (Madry et al., 2017; Tramèr et al., 2017; Sinha et al., 2017; Arnab et al., 2018). It involves training the network using adversarial examples, enhancing its robustness to attacks during inference. This strategy has been effective in classification settings. Standard adversarial training involves solving the following min-max optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\in\mathbb{D}}[\max_{\delta:\|\delta\|_p\leq\epsilon} \mathcal{L}(f(x+\delta;\theta),y)] \qquad (2)$$

where $\mathcal{L}(\cdot)$ represents the applicable loss function, e.g., cross-entropy for classification, and δ is the perturbation added to each sample, within an ℓ_p-norm ball of radius ϵ. This min-max formulation encompasses possible variants of adversarial training. It consists of solving two optimization problems: an inner maximization and an outer minimization problem. This corresponds to an adversarial game between the attacker and robust network f. The inner problem tries to find the optimal $\delta : \|\delta\|_p \leq \epsilon$ for a given data point (x, y) maximizing the loss, which essentially is the adversarial attack, whereas the outer problem aims to find a θ minimizing the same loss. For an optimal θ^* solving (2), $f(;\theta^*)$ will be robust (in expected value) to all x_{adv} lying in the ℓ_p ball of radius ϵ-around the true x.

2.2. Problem Formulation

(Antun et al., 2020) identify instabilities of a deep learning based image reconstruction network by maximizing the following cost function:

$$Q_y(r) = \frac{1}{2}\|f(y+Ar)-x\|_2^2 - \frac{\lambda}{2}\|r\|^2 \qquad (3)$$

As evident from this framework, the perturbation r is added in the x-space for each y, resulting in perturbation Ar in the y-space. We argue that this formulation can miss important aspects in image reconstruction, especially in ill-posed problems, for the following three main reasons:

1. It may not be able to model all possible perturbations to y. The perturbations $A\delta$ to y modeled in this formulation are all constrained to the range-space of A. When A does not have full row rank, there exist perturbations to y that cannot be represented as $A\delta$.

2. It misses instabilities created by the ill-conditioning of the reconstruction problem. Consider a simple ill-conditioned reconstruction problem:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} \text{ and } f = \begin{bmatrix} 1 & 0 \\ 0 & 1/a \end{bmatrix} \qquad (4)$$

where A and f define the forward and reconstruction operator respectively, and $|a| \ll 1$. For $\delta = [0, \epsilon]^T$ perturbation in x, the reconstruction is $f(A(x+\delta)) = x+\delta$, and the reconstruction error is $\|f(A(x+\delta)) - x\|_2 = \epsilon$, that is, for small ϵ, the perturbation has negligible effect. In contrast, for the same perturbation δ in y, the reconstruction is $f(Ax+\delta) = x+[0, \epsilon/a]^T$, with reconstruction error $\|f(A(x+\delta)) - x\|_2 = \epsilon/a$, which can be arbitrarily large if $a \to 0$. This aspect is completely missed by the formulation based on (3).

3. For inverse problems, one also wants robustness to perturbations in the measurement matrix A. Suppose A used in training is slightly different from the actual $A' = A + \tilde{A}$ that generates the measurements. This results in perturbation $\tilde{A}x$ in y-space, which may be outside the range space of A, and therefore, as in 1 above, may not be possible to capture by the formulation based on (3).

The above points indicate that studying the problem of robustness to perturbations for image reconstruction problems in x-space misses possible perturbations in y-space that can have a huge adversarial effect on reconstruction. Since many of the image reconstruction problems are ill-posed or ill-conditioned, we formulate and study the issue of adversaries in the y-space, which is more generic and able to handle perturbations in the measurement operator A as well.

2.3. Image Reconstruction

Image Reconstruction deals with recovering the clean image x from noisy and possibly incomplete measurements $y = Ax + v$. Recently, deep-learning-based approaches have outperformed the traditional techniques. Many deep learning architectures are inspired by iterative reconstruction schemes (Rick Chang et al., 2017; Raj et al., 2019; Bora et al., 2017; Wen et al., 2019). Another popular way is to use an end-to-end deep network to solve the image reconstruction problem directly (Jin et al., 2017; Zhu et al., 2018; Schlemper et al., 2017; Yang et al., 2017; Hammernik et al., 2018; Sajjadi et al., 2017; Yao et al., 2019). In this work, we propose modification in the training scheme for the end-to-end networks.

Consider the standard MSE loss in x-space with the popular ℓ_2-regularization on the weights (aka weight decay), which mitigates overfitting and helps in generalization (Krogh & Hertz, 1992)

$$\min_{\theta} \mathbb{E}_x \|f(Ax;\theta) - x\|^2 + \mu\|\theta\|^2 \tag{5}$$

In this paper, we experiment both with $\mu > 0$ (regularization present) and $\mu = 0$ (no regularization). No regularization is used in the sequel, unless stated otherwise.

2.3.1. Adversarial Training for Image Reconstruction

Motivated by the adversarial training strategy (2), several frameworks have been proposed recently to make classification by deep networks more robust (Jang et al., 2019b; Kurakin et al., 2016; Wang & Yu, 2019). For image reconstruction, we propose to modify the training loss to the general form:

$$\min_{\theta} \mathbb{E}_x \max_{\delta:\|\delta\|_p \leq \epsilon} \|f(Ax;\theta) - x\|^2 + \lambda\|f(Ax+\delta;\theta) - x\|^2$$

The role of the first term is to ensure that the network f maps the non-adversarial measurement to the true x, while the role of the second term is to train f on worst-case adversarial examples within the ℓ_p-norm ball around the nominal measurement Ax. We want δ to be the worst case perturbation for a given f. However, during the initial training epochs, f is mostly random (assuming random initialization of the weights) resulting in random perturbation, which makes f diverge. Instead, we use only the first term during initial epochs to get an f that provides reasonable reconstruction. Then, reasonable perturbations are obtained by activating the second term, which results in robust f.

Now, solving the min-max problem above is intractable for a large dataset as it involves finding the adversarial example by solving the inner maximization for each training sample $y = Ax$. This may be done using projected gradient descent (PGD), but is very costly. A possible sub-optimal approximation (with $p = 2$) for this formulation is:

$$\min_{\theta} \max_{\delta:\|\delta\|_2 \leq \epsilon} \mathbb{E}_x \|f(Ax;\theta) - x\|_2^2 + \lambda\|f(Ax+\delta;\theta) - x\|_2^2$$

This formulation finds a common δ that is adversarial to all measurements y on the average, and tries to minimize the reconstruction loss for the adversarial examples together with that for clean examples. Clearly this is sub-optimal as using a perturbation δ common to all y's need not be the worst-case perturbation for any of the y's, and optimizing for the common δ won't result in a highly robust network.

Ideally, we would want the best of both worlds: i.e., to generate δ for each y independently, together with tractable training. To this end, we propose to parameterize the worst-case perturbation $\delta = \arg\max_{\delta:\|\delta\|_2 \leq \epsilon} \|f(y+\delta;\theta) - x\|_2^2$ by a deep neural network $G(y;\phi)$. This also eliminates the need to solve the inner-maximization to find δ using hand-crafted methods. Since $G(\cdot)$ is parameterized by ϕ and takes y as input, a well-trained G will result in optimal perturbation for the given $y = Ax$. The modified loss function becomes:

$$\begin{aligned}\min_{\theta} \max_{\phi:\|G(\cdot,\phi)\|_2 \leq \epsilon} & \mathbb{E}_x \|f(Ax;\theta) - x\|^2 \\ & + \lambda\|f(Ax + G(Ax;\phi);\theta) - x\|^2\end{aligned}$$

This results in an adversarial game between the two networks: G and f, where G's goal is to generate strong adversarial examples that maximize the reconstruction loss for the given f, while f tries to make itself robust to the adversarial examples generated by the G. This framework is illustrated in the Fig. 1. This min-max setting is quite similar to the Generative adversarial network (GAN), with the difference in the objective function. Also, here, the main goal is to build an adversarially robust f, which requires some empirical changes compared to standard GANs to make it work. Another change is to reformulate the constraint $\|G(\cdot,\phi)\|_2 \leq \epsilon$ into a penalty form using the hinge loss, which makes the training more tractable:

$$\min_{\theta}\max_{\phi} \quad \mathbb{E}_x\|f(Ax;\theta)-x\|^2 + \lambda_1\|f(Ax+G(Ax;\phi);\theta)-x\|^2 + \lambda_2\max\{0,\|G(Ax;\phi)\|_2^2-\epsilon\} \tag{6}$$

Note that λ_2 must be negative to satisfy the required constraint $\|G(\cdot,\phi)\|_2 \leq \epsilon$.

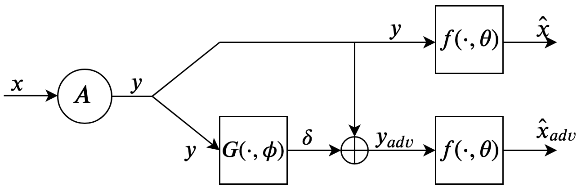


Figure 1. Adversarial training framework of image reconstruction network f with another network G generating the perturbations

2.3.2. Training Strategy

We apply some modifications and heuristic changes to train a robust f jointly with training G in a mini-batch set-up. In each iteration, we update G to generate adversarial examples and train f using those adversarial examples along with the non-adversarial or clean samples to make it robust. Along with the training of robust f, G is being trained to generate worst-case adversarial examples. To generate strong adversarial examples by G in the mini-batch update, we divide each mini-batch into K sets. Now, G is trained over each set independently and we use adversarial examples after the update of G for each set. This fine-tunes G for the small set to generate stronger perturbations for every image belonging to the set. Then, f is trained using the entire mini-batch at once but with the adversarial examples generated set-wise. G obtained after the update corresponding to the K^{th} set is passed for the next iteration or mini-batch update. This is described in Algorithm 1.

2.4. Robustness Metric

To define a metric for the robustness of a network, we measure the following quantity for network f:

$$\Delta_{\text{max}}(x_0,\epsilon) = \max_{\|\delta\|_2\leq\epsilon}\|f(Ax_0+\delta)-x_0\|^2 \tag{7}$$

This determines the reconstruction error due to the worst-case additive perturbation over an ϵ-ball around the nominal measurement $y = Ax_0$ for each image x_0. The final robustness metric for f is $\rho(\epsilon) = \mathbb{E}_{x_0}[\Delta_{\text{max}}(x_0,\epsilon)]$, which we estimate by the sample average of $\Delta_{\text{max}}(x_0,\epsilon)$ over a test dataset,

$$\hat{\rho}(\epsilon) = \frac{1}{N}\sum_{i=1}^{N}\Delta_{\text{max}}(x_i,\epsilon) \tag{8}$$

Algorithm 1 Algorithm for training at iteration T

```
Input: Mini-batch samples $(x_T, y_T)$, $G_{T-1}$, $f_{T-1}$
Output: $G_T$ and $f_T$
$G_{T,0} = G_{T-1}$, $f = f_{T-1}$ Divide mini-batch into $K$ parts.
while $k \leq K$ do
    $x = x_{T,k}, G = G_{T,k-1}$
    $G_{T,k} = \arg\max_G \lambda_1\|f_{T-1}(Ax+G(Ax;\phi);\theta)-x\|^2 + \lambda_2\max\{0,\|G(Ax;\phi)\|_2^2-\epsilon\}$
    $\delta_{T,k} = G_{T,k}(x)$
end while
$\delta_T = [\delta_{T,1}, \delta_{T,2}, ..., \delta_{T,K}]$
$f_T = \arg\min_f \|f(Ax_T)-x_T\|^2 + \lambda_1\|f(Ax_T+\delta_T)-x_T\|^2$
$G_T = G_{T,K}$
return $G_T, f_T$
```

The smaller $\hat{\rho}$, the more robust the network.

We solve the optimization problem in (7) using projected gradient ascent (PGA) with momentum (with parameters selected empirically). Importantly, unlike training, where computation of $\Delta_{\text{max}}(x_0)$ is required at every epoch, we need to solve (7) only once for every sample x_i in the test set, making this computation feasible during testing.

3. Theoretical Analysis

We theoretically obtained the optimal solution for the min-max formulation in (2.3.1) for a simple linear reconstruction. Although this analysis doesn't extend easily to the non-linear deep learning based reconstruction, it gives some insights for the behavior of the proposed formulation, and how it depends on the conditioning of the measurement matrices.

Theorem 1. *Suppose that the reconstruction network f is a one-layer feed-forward network with no non-linearity i.e., $f = B$, and assume that the data is normalized, i.e., $E(x) = 0$ and $\text{COV}(x) = I$. Denote the SVD of the measurement matrix A by $A = USV^T$, where S is a diagonal matrix with singular values ordered in (increasing) order, and the SVD of matrix B by $B = MQP^T$. Then the optimal B obtained by solving (2.3.1) is a modified pseudo-inverse of A, with $M = V$, $P = U$ and Q a filtered inverse of S, given by*

$$Q = \text{diag}\left(q_m, \ldots, q_m, 1/S_{m+1}, \ldots, 1/S_n\right), \qquad q_m = \frac{\sum_{i=1}^{m} S_i}{\sum_{i=1}^{m} S_i^2 + \frac{\lambda}{1+\lambda}\epsilon^2} \tag{9}$$

with largest entry q_m of multiplicity m that depends on ϵ, λ and $\{S_i\}_{i=1}^{n}$.

Proof. Please refer to the supplementary material. □

The modified inverse B reduces the effect of ill-conditioning in A for adversarial cases in the reconstruction. This can be easily understood, using the simple example from (4). As explained previously, for the A in (4) with $|a| < 1$, an exact inverse, $f = \begin{bmatrix}1 & 0\\ 0 & \frac{1}{a}\end{bmatrix}$, amplifies the perturbation. Instead the min-max formulation (2.3.1) (with $\lambda = 1$) results in a modified pseudo inverse $\hat{f} = \begin{bmatrix}1 & 0\\ 0 & \frac{a}{a^2+0.5\epsilon^2}\end{bmatrix}$, suppressing the effect of an adversarial perturbation $\delta = [0, \epsilon]^T$ in y as $\|\hat{f}\delta\| \ll \|f\delta\|$ for $a \to 0$ and $\epsilon \nrightarrow 0$. It can also be seen that $\hat{f}$ won't be optimal the for the unperturbed y as it's not actual an inverse and the reconstruction loss using f for the unperturbed case would be smaller than that for $\hat{f}$. However, for even very small adversaries, f would be much more sensitive than $\hat{f}$. This shows the trade-off between the perturbed and unperturbed case for the reconstruction in the case of an ill-conditioned A.

This trade-off behavior will not manifest for a well-conditioned A, as an ideal linear inverse f for this case won't amplify the small perturbations, and a reconstruction obtained using (2.3.1) with linear $\hat{f}$ will be very close to f (depending on ϵ): for well-conditioned A, $a \nrightarrow 0$. In that case $a^2 \gg 0.5\epsilon^2$, which reduces $\hat{f}$ to f.

Our experiments with deep-learning-based non-linear image reconstruction methods for CS using as sensing matrices random rows of a Gaussian matrix (well-conditioned) vs. random rows of a DCT sub-matrix (relatively ill-conditioned) indeed show the qualitatively different behavior with increasing strength of the perturbations.

4. Experiments

Network Architecture: For the reconstruction network f, we follow the architecture of deep convolutional networks for image reconstruction. They use multiple convolution, deconvolution and ReLU layers, and use batch normalization and dropout for better generalization. As a pre-processing step, which has been found to be effective for reconstruction, we apply the transpose (adjoint) of A to the measurement y, feeding $A^T y$ to the network. This transforms the measurement into the image-space, allowing the network to operate purely in image space.

For the adversarial perturbation generator G we use a standard feed-forward network, which takes input y as input. The network consists of multiple fully-connected and ReLU layers. We trained the architecture shown in fig. 1 using the objective defined in (6).

We designed networks of similar structure but different number of layers for the two datasets, MNIST and CelebA used in the experiments.

We used the Adam Optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, learning rate of 10^{-4} and mini-batch size of 128, but divided into $K = 4$ parts during the update of G, described in the algorithm 1. During training, the size ϵ of the perturbation has to be neither too big (affects performance on clean samples) nor too small (results in less robustness). We empirically picked $\epsilon = 2$ for MNIST and $\epsilon = 3$ for the CelebA datasets. However, during testing, we evaluated $\hat{\rho}$, defined in (8) for different ϵ's (including those not used while training), to obtain a fair assessment of robustness.

We compare the adversarially trained model using the min-max formulation defined in the objective (6), with three models trained using different training schemes:

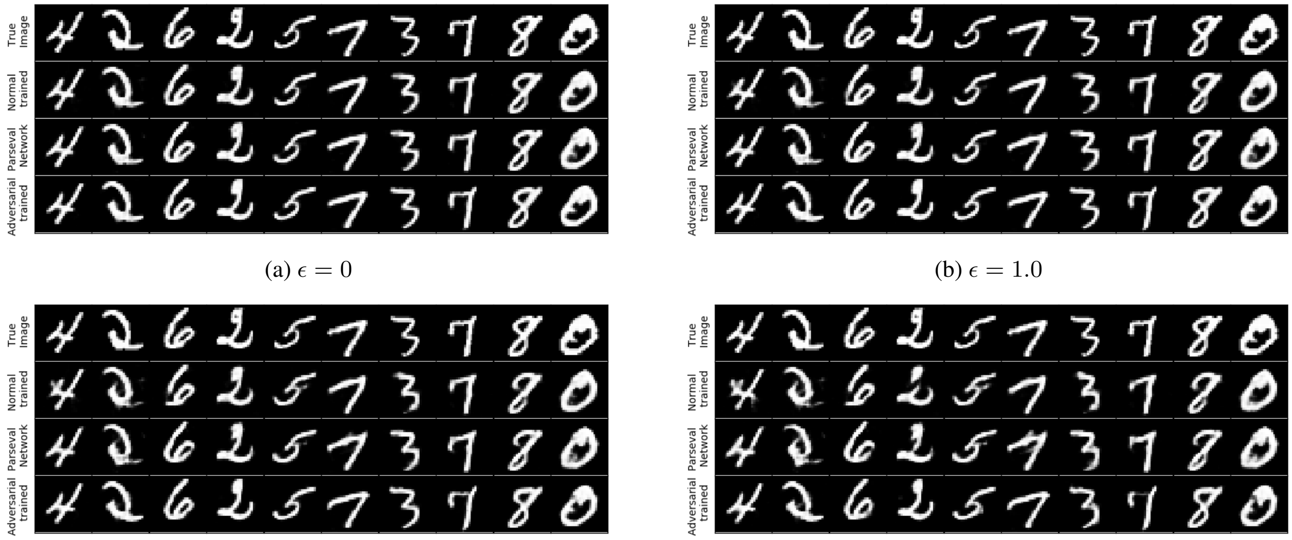
1. Normally trained model with no regularization, i.e., $\mu = 0$ in (6).
2. ℓ_2-norm weight regularized model, using (5) with $\mu > 10^{-6}$ (aka weight decay), chosen empirically to avoid over-fitting and improve robustness and generalization of the network.
3. Lipschitz constant ($\mathcal{L}$)-constrained Parseval network (Cisse et al., 2017b). The idea is to constrain the overall Lipschitz constant $\mathcal{L}$ of the network to be ≤ 1, by making $\mathcal{L}$ of every layer, ≤ 1. Motivated by the idea that regularizing the spectral norm of weight matrices could help robustness, this approach proposes to constrain the weight matrices to also be orthonormal, making them *Parseval tight frames*. Let S_{fc} and S_c define the set of indices for fully-connected and convolutional layers respectively. The regularization term to penalize the deviation from the constraint is

$$\frac{\beta}{2}\Big(\sum_{i\in S_{fc}} \|W_i^T W_i - I_i\|_2^2 + \sum_{j\in S_c} \|\mathbf{W_j}^T \mathbf{W_j} - \frac{I_j}{k_j}\|_2^2\Big) \tag{10}$$

   where W_i is the weight matrix for ith fully connected layer and $\mathbf{W_j}$ is the transformed or unfolded weight matrix of jth convolution layer having kernel size k_j. Transformation requires input to the convolution to shift and repeat k_j^2 times. Hence, to maintain the *Parseval tight frames* constraint on the convolution operator, we need to make $\mathbf{W_j}^T\mathbf{W_j} \approx \frac{I_j}{k_j}$. I_i and I_j are identity matrices whose sizes depend on the size of W_i and $\mathbf{W_j}$ respectively. β controls the weight given to the regularization compared to the standard reconstruction loss. Empirically, we picked β to be 10^{-5}.

To compare different training schemes, we follow the same scheme (described below) for each dataset. Also, we extensively compare the performance for the two datasets for Compressive Sensing (CS) task using two matrices: one well-conditioned and another, relatively ill-conditioned. This comparison complements the theoretical analysis in the previous section.

The **MNIST** dataset (LeCun et al., 1998) consists of 28×28 gray-scale images of digits with $50,000$ training and $10,000$ test samples. The image reconstruction network consists of



(a) $\epsilon = 0$

(b) $\epsilon = 1.0$

(c) $\epsilon = 2.0$

(d) $\epsilon = 3.0$

Figure 2. Qualitative Comparison for the MNIST dataset for different perturbations. *First row* of each sub-figure corresponds to the true image, *Second row* to the reconstruction using normally trained model, *Third row* to the reconstruction using Parseval Network, *Fourth row* to the reconstruction using the adversarially trained model (***proposed scheme***).

4 convolution layers and 3 transposed convolution layers using re-scaled images between $[-1, 1]$. For the generator G, we used 5 fully-connected layers network. Empirically, we found $\lambda_1 = 1$ and $\lambda_2 = -0.1$ in (6), gave the best performance in terms of robustness (lower $\hat{\rho}$) for different perturbations.

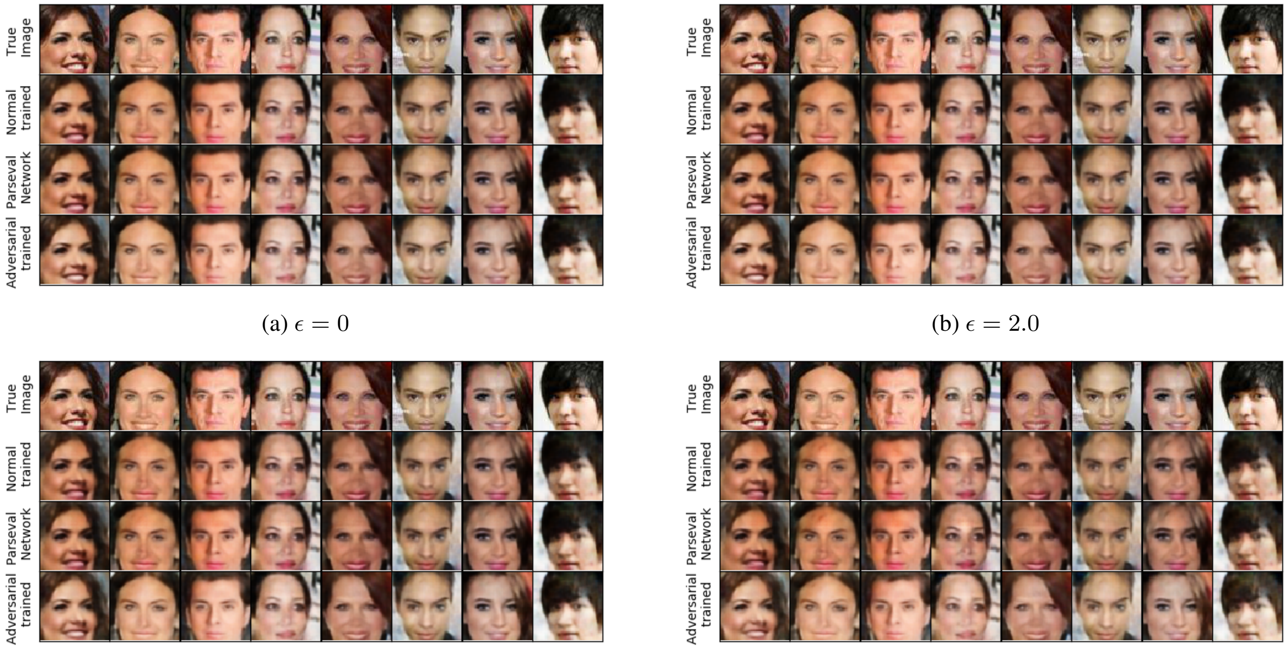
The **CelebA** dataset (Liu et al., 2015) consists of more than $200,000$ celebrity images. We use the aligned and cropped version, which pre-processes each image to a size of $64 \times 64 \times 3$ and scaled between $[-1, 1]$. We randomly pick $160,000$ images for the training. Images from the $40,000$ held-out set are used for evaluation. The image reconstruction network consists of 6 convolution layers and 4 transposed convolution layers. For the generator G, we used a 6 fully-connected layers network. We found $\lambda_1 = 3$ and $\lambda_2 = -1$ in (6) gave the best robustness performance (lower $\hat{\rho}$) for different perturbations.

4.1. Gaussian Measurement matrix

In this set-up, we use the same measurement matrix A as (Bora et al., 2017; Raj et al., 2019), i.e. $A_{i,j} \sim N(0, 1/m)$ where m is the number of measurements. For MNIST, the measurement matrix $A \in R^{m \times 784}$, with $m = 100$, whereas for CelebA, $A \in R^{m \times 12288}$, with $m = 1000$. Figures 2 and 3 show the qualitative comparisons for the MNIST and CelebA reconstructions respectively, by solving the optimization described in Section 2.4. It can be seen clearly in both the cases that for different ϵ the adversarially trained models outperform the normally trained and Parseval networks. For higher ϵ's, the normally trained and Parseval models generate significant artifacts, which are much smaller for the adversarially trained models. Figures 4a and 4b show this improvement in performance in terms of the quantitative metric $\hat{\rho}$, defined in (8) for the MNIST and CelebA datasets respectively. It can be seen that $\hat{\rho}$ is lower for the adversarially-trained models compared to other training methods: no regularization, ℓ_2-norm regularization on weights, and Parseval networks (Lipschitz-constant-regularized) for different ϵ's, showing that adversarial training using the proposed min-max formulation indeed outperforms other approaches in terms of robustness. It is noteworthy that even for $\epsilon = 0$ (unperturbed case), adversarial training reduces the reconstruction loss, indicating that it acts like an excellent regularizer in general.

4.2. Discrete Cosine Transform (DCT) sub-matrix

To empirically study the effect of conditioning of the matrix, we did experiment by choosing A as random m rows and n columns of a $p \times p$ DCT matrix, where $p \gg n$. This makes A relatively ill-conditioned compared to the random Gaussian A, i.e. the condition number for the random DCT sub-matrix is higher than that of random Gaussian one. The number of measurements has been kept same as the previous case, i.e. $(m = 100,\ n = 784)$ for MNIST and $(m = 1000,\ n = 12288)$ for CelebA. We trained networks having the same configuration as the Gaussian ones. Figure 4 shows the comparison for the two measurement matrices. Based on the figure, we can see that $\hat{\rho}$ for the



(a) $\epsilon = 0$

(b) $\epsilon = 2.0$

(c) $\epsilon = 5.0$

(d) $\epsilon = 10.0$

Figure 3. Qualitative Comparison for the CelebA dataset for different perturbations. *First row* of each sub-figure corresponds to the true image, *Second row* to the reconstruction using normally trained model, *Third row* to the reconstruction using Parseval Network, *Fourth row* to the reconstruction using the adversarially trained model (***proposed scheme***).

DCT sub-matrix, MNIST (Fig. 4d) and CelebA (Fig. 4e), are very close for models trained adversarially and using other schemes for the unperturbed case ($\epsilon = 0$), but the gap between them increases with increasing ϵ's, with adversarially trained models outperforming the other methods consistently. This behavior is qualitatively different from that for the Gaussian case (Fig. 4a and Fig. 4b), where the gap between adversarially trained networks and models trained using other (or no) regularizers is roughly constant for different ϵ.

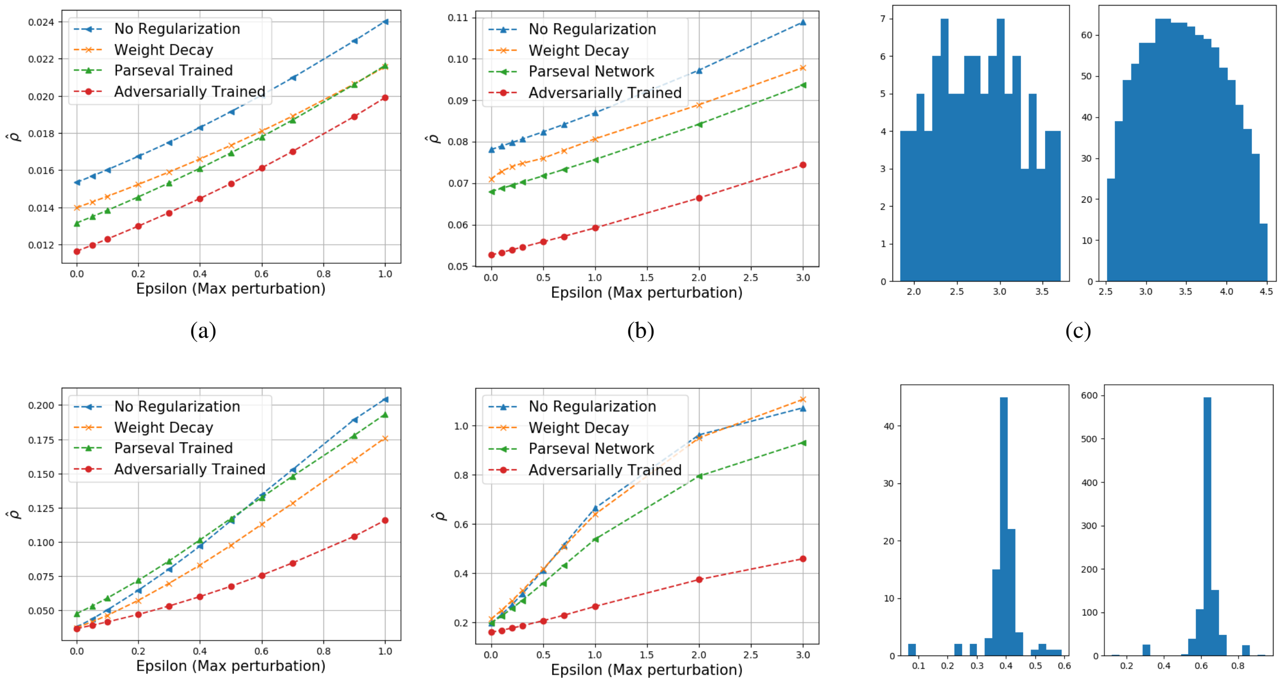
4.3. Analysis with respect to Conditioning

To check the conditioning, Fig.4c shows the histogram for the singular values of the random Gaussian matrices. It can be seen that the condition number (ratio of max. and min. singular value) is close to 2 which is very well conditioned for both data sets. On the other hand, the histogram of the same for the random DCT sub-matrices (Fig.4f) shows higher condition numbers – 8.9 for the 100×784 and 7.9 for the 1000×12288 dimension matrices, which is ill-conditioned relative to the Gaussian ones.

Referring to the above analysis of conditioning and plots of the robustness measure $\hat{\rho}$ for the two types of matrices: random Gaussian vs. random DCT indicate that the behavior of the proposed min-max formulation depends on how well (or relatively ill)-conditioned the matrices are. This corroborates with the theoretical analysis for a simple reconstruction scheme (linear network) described in Sec. 3.

4.4. CT Reconstruction

In this experiment, we implement the proposed adversarial training method on the state-of-the-art deep-learning-based network, *FBPConvNet* (Jin et al., 2017) for low-dose x-ray CT reconstruction. As in (Jin et al., 2017), the measurements y are obtained by computing projections of the CT images at 143 views uniformly spaced between $[0, 180°]$. The input to the reconstruction network in this case is different from that in our previous examples, where the input was $A^T y$. Instead, the reconstruction network FBPConvNet is fed with an initial reconstruction estimate $\hat{x} = Ry$ where R is the FBP reconstruction operator, and y is the set of projections at 143 views. Note that $R = A^* H$, where H is a so-called ramp filter, and the backprojection operator A^* is the adjoint of A. Because in the matrix case, $A^* = A^T$, it follows that in the case of the FBPConvNet, there is an extra step of applying the filter H on the input. For ground truth, we used FBP reconstructions from projections at 1000 views, also called full-view FBP. For fast computation of forward projection (Radon transform) and filtered back-projection (FBP - numerical inverse Radon transform) on GPUs, we used the Astra toolbox (Van Aarle et al., 2016).



(a) (b) (c)

(d) (e) (f)

Figure 4. **Row 1** corresponds to the random rows of Gaussian measurement matrix: (a) MNIST, (b) CelebA, (c) Distribution of the singular values for MNIST (left, $m = 100$) and CelebA (right, $m = 1000$) cases. **Row 2** corresponds to random rows and columns of the DCT measurement matrix: (a) MNIST, (b) CelebA, (c) Distribution of the singular values for MNIST (left, $m = 100$) and CelebA (right, $m = 1000$) cases.

For normal training of the FBPConvNet, the hyperparameters are set exactly the same as in (Jin et al., 2017). Next, we train the FBPConvNet using our proposed adversarial scheme with attack generator G consisting of 6 fully-connected layers. We used anonymized clinical CT images (Vannier, 2007) of size 512×512 – 884 for training & validation and 221 for evaluation. To determine the attack and evaluate the trained reconstruction network for robustness, we employed the set-up described in Sec. 2.4, using (7) and (8), where the operator A is the Radon Transform, and $f(y) = \mathrm{N}(Ry)$, with R equal to the FBP and N denoting the neural network. To solve the optimization problem given by (7), we use projected gradient ascent. The gradient with respect to δ involves back-propagating through R, and therefore uses its adjoint, $R^* = H^*A = HA$, where the second equality follows because H is self-adjoint. This is different from the work in (Antun et al., 2020), which used R in the computation of gradient during the attack, because their attack was in the image domain, whereas ours is in the measurement domain.

Fig. 5 shows a qualitative comparison of reconstruction results in the presence of an attack (a slight perturbation in the data). It can be seen that the streaks and artifacts appearing due to the attack for a normally trained network, are not typically present in the proposed adversarial training scheme.

4.5. Linear Network for Reconstruction

We perform an experiment using a linear reconstruction network in a simulated set-up to compare the theoretically obtained optimal robust reconstruction network with the one learned by our scheme by optimizing the objective (2.3.1). We take $50,000$ samples of a signal $x \in \mathbb{R}^{20}$ drawn from $\mathcal{N}(0, I)$, hence, $\mathbb{E}(x) = 0$ and $\mathrm{COV}(x) = I$. For the measurement matrix $A \in \mathbb{R}^{10\times 20}$, we follow the same strategy as in Sec. 4.1, i.e. $\tilde{A}_{ij} \sim \mathcal{N}(0, 1/10)$. Since such matrices are well-conditioned, we replace 2 singular values of $\tilde{A}$ by small values (10^{-3} and 10^{-4}) keeping the other singular values and singular matrices fixed. This makes the modified matrix A ill-conditioned. We obtain the measurements $y = Ax \in \mathbb{R}^{10}$. For reconstruction, we train a linear network f having 1 fully-connected layer with no non-linearity i.e. $f = B \in \mathbb{R}^{20\times 10}$. The reconstruction is given by $\hat{x} = \hat{B}y$, where $\hat{B}$ is obtained by solving

$$\arg\min_{B} \max_{\delta:\|\delta\|_2\le\epsilon} \mathbb{E}_x\|BAx - x\|^2 + \lambda\|B(Ax+\delta) - x\|^2 \tag{11}$$

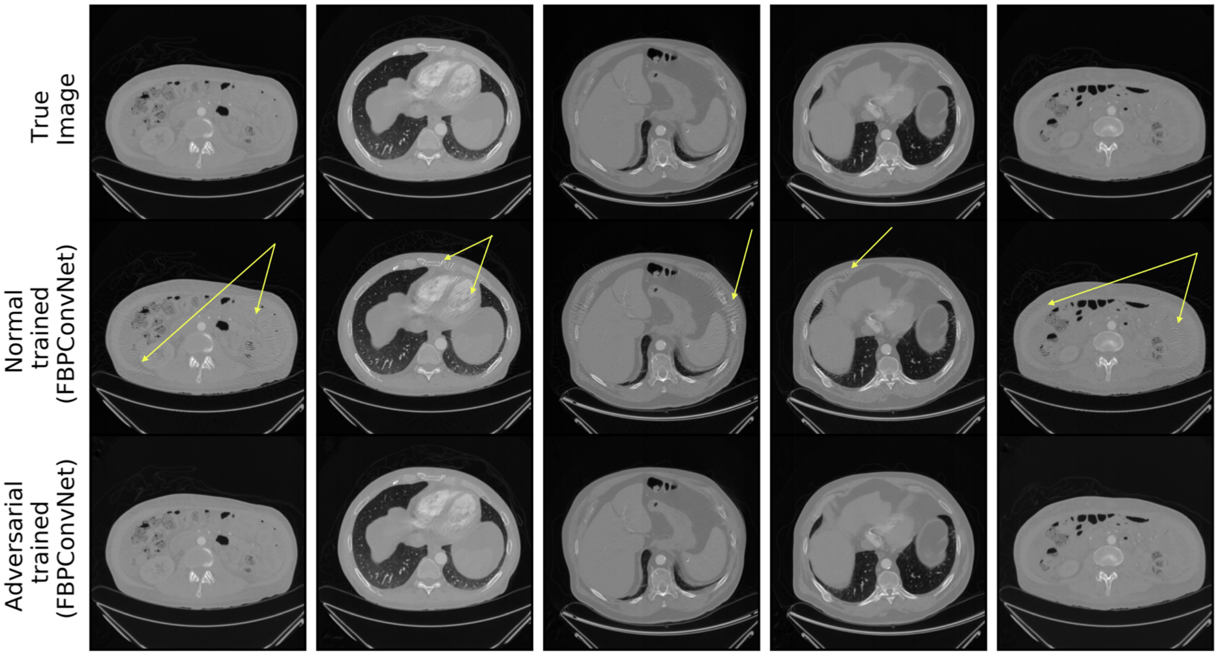


Figure 5. Qualitative Comparison for the reconstruction after attack for the X-ray dataset using FBPConvNet. *First row* corresponds to the true image. *Second row* to the reconstruction using normally trained model. *Yellow* arrows indicate the streaks and artifacts appearing after attack. *Third row* to the reconstruction using the adversarially trained model (***proposed scheme***)

.

We used $\lambda = 1$, $\epsilon = 0.1$, learning rate $= 0.001$ and momentum term of 0.9 in our experiments. We obtain the theoretically derived reconstruction matrix B using the result given in (9) (from theorem 1). To compare B and $\hat{B}$, we examined the following three metrics:

- $\|\hat{B}-B\|_F/\|B\|_F = 0.024$, $\|\hat{B}-B\|_2/\|B\|_2 = 0.034$
- $\|I - BA\|_F/\|I - \hat{B}A\|_F = 0.99936$, where I is the identity matrix of size 20×20
- $\kappa(B) = 19.231$, $\kappa(\hat{B}) = 19.311$, κ: condition number

The above four metrics indicate that $\hat{B}$ indeed converges to the theoretically obtained solution B.

5. Conclusions

In this work, we proposed a min-max formulation to build robust deep-learning-based image reconstruction models. To make this more tractable, we reformulate this using an auxiliary network to generate adversarial examples for which the image reconstruction network tries to minimize the reconstruction loss. We theoretically analyzed a simple linear network and found that using the min-max formulation produces a singular-value filter regularized solution, which reduces the effect of adversarial examples for ill-conditioned matrices. Empirically, we found an adversarially-trained linear network to converge to the same solution. Additionally, extensive experiments with non-linear deep networks for Compressive Sensing (CS) using random Gaussian and DCT measurement matrices on the MNIST and CelebA datasets show that the proposed scheme outperforms other methods for different perturbations $\epsilon \geq 0$, however the behavior depends on the conditioning of the measurement matrices, as indicated by theory for the linear reconstruction scheme. Further, our experiment on real CT image reconstruction indicates that the state-of-the-art network trained using the proposed method outperforms the normal training methods in terms of robustness.

Acknowledgements

The authors would like to thank Prof. Michael Vannier, Department of Radiology, University of Chicago, for providing us the anonymized clinical CT images. This work was supported in part by US Army MURI Award W911NF-15-1-0479.

References

Antun, V., Renna, F., Poon, C., Adcock, B., and Hansen, A. C. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 2020.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Arnab, A., Miksik, O., and Torr, P. H. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 888–897, 2018.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. *arXiv preprint arXiv:1703.03208*, 2017.

Candes, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Choi, J.-H., Zhang, H., Kim, J.-H., Hsieh, C.-J., and Lee, J.-S. Evaluating robustness of deep image super-resolution against adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 303–311, 2019.

Cisse, M., Adi, Y., Neverova, N., and Keshet, J. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017a.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 854–863. JMLR. org, 2017b.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. Bm3d image denoising with shape-adaptive principal component analysis. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

Dong, W., Zhang, L., Shi, G., and Wu, X. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Elad, M. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

Elad, M. and Aharon, M. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hammernik, K., Klatzer, T., Kobler, E., Recht, M. P., Sodickson, D. K., Pock, T., and Knoll, F. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

Jang, Y., Zhao, T., Hong, S., and Lee, H. Adversarial defense via learning to generate diverse attacks. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019a.

Jang, Y., Zhao, T., Hong, S., and Lee, H. Adversarial defense via learning to generate diverse attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2740–2749, 2019b.

Jiang, H., Chen, Z., Shi, Y., Dai, B., and Zhao, T. Learning to defense by learning to attack. *arXiv preprint arXiv:1811.01213*, 2018.

Jin, K. H., McCann, M. T., Froustey, E., and Unser, M. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26 (9):4509–4522, 2017.

Krogh, A. and Hertz, J. A. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pp. 950–957, 1992.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, C., Yin, W., and Zhang, Y. User's guide for tval3: Tv minimization by augmented lagrangian and alternating direction algorithms. *CAAM report*, 20(46-47):4, 2009.

Liu, D., Wen, B., Liu, X., Wang, Z., and Huang, T. S. When image denoising meets high-level vision tasks: A deep learning approach. *arXiv preprint arXiv:1706.04284*, 2017.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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

Raj, A., Li, Y., and Bresler, Y. GAN-based projector for faster recovery with convergence guarantees in linear inverse problems. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5602–5611, 2019.

Ravishankar, S. and Bresler, Y. Learning sparsifying transforms. *IEEE Transactions on Signal Processing*, 61(5): 1072–1086, 2012.

Rick Chang, J., Li, C.-L., Poczos, B., Vijaya Kumar, B., and Sankaranarayanan, A. C. One network to solve them all–solving linear inverse problems using deep projection models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5888–5897, 2017.

Sajjadi, M. S., Scholkopf, B., and Hirsch, M. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4491–4500, 2017.

Schlemper, J., Caballero, J., Hajnal, J. V., Price, A., and Rueckert, D. A deep cascade of convolutional neural networks for mr image reconstruction. In *International Conference on Information Processing in Medical Imaging*, pp. 647–658. Springer, 2017.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

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

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Van Aarle, W., Palenstijn, W. J., Cant, J., Janssens, E., Bleichrodt, F., Dabravolski, A., De Beenhouwer, J., Batenburg, K. J., and Sijbers, J. Fast and flexible x-ray tomography using the astra toolbox. *Optics express*, 24(22): 25129–25147, 2016.

Vannier, M. *Personal communication*, 2007.

Wang, H. and Yu, C.-N. A direct approach to robust deep learning using adversarial networks. *arXiv preprint arXiv:1905.09591*, 2019.

Wen, B., Ravishankar, S., and Bresler, Y. Structured overcomplete sparsifying transform learning with convergence guarantees and applications. *International Journal of Computer Vision*, 114(2-3):137–167, 2015.

Wen, B., Ravishankar, S., Pfister, L., and Bresler, Y. Transform learning for magnetic resonance image reconstruction: From model-based learning to building neural networks. *arXiv preprint arXiv:1903.11431*, 2019.

Wong, E., Schmidt, F., Metzen, J. H., and Kolter, J. Z. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, pp. 8400–8409, 2018.

Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M., and Song, D. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P. L., Ye, X., Liu, F., Arridge, S., Keegan, J., Guo, Y., et al. Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2017.

Yang, J., Wright, J., Huang, T. S., and Ma, Y. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

Yao, H., Dai, F., Zhang, S., Zhang, Y., Tian, Q., and Xu, C. Dr2-net: Deep residual reconstruction network for image compressive sensing. *Neurocomputing*, 359:483–493, 2019.

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