Addressing The False Negative Problem of Deep Learning MRI Reconstruction Models by Adversarial Attacks and Robust Training

Abstract
Deep learning models have been shown to have success in accelerating MRI reconstruction, over traditional methods. However, it has been observed that these methods tend to miss the small features that are rare, such as meniscal tears, subchondral osteophyte, etc. This is a concerning finding as these small and rare features are the most relevant in clinical diagnostic settings. In this work, we propose a framework to find the worst-case false negatives by adversarially attacking the trained models and improve the models’ ability to reconstruct the small features by robust training.

Keywords: MRI Reconstruction, Adversarial Attack, Robust Training.

1. Introduction
High data quality is a priority in medical image analysis. Magnetic resonance imaging (MRI) has the capability of satisfying such requirement when it comes to screening soft tissues. However, MRI has a limitation of requiring long scanning time. As a consequence, over the past few years, acceleration of MRI has received an increasing level of attention, which did not restrict only to medical physicist but extended also to the deep learning community. The book chapter from Hammernik and Knoll (2020), the Fast-MRI challenge that was held at NeurIPS 2019 (Zbontar et al., 2018), the upcoming AccelMR 2020 challenge at ISBI and dedicated sessions at prestigious conferences such as ISMRM, MICCAI are all examples that accelerated MRI powered by deep learning is an active topic of research.

1.1. Hypotheses for false negative
The false negative phenomenon of MRI reconstruction models refers to qualitative observations provided during the announcement of NeurIPS 2019 FastMRI challenge results\(^1\). The best and worst performing models in terms of structural similarity metric (SSIM) and radiologists’ image quality assessment, were shown to have failed in reconstructing some relatively small abnormalities such as meniscal tear and subchondral osteophyte. In this paper, we investigate two intuitive orthogonal hypotheses, with the aim to better explain this phenomenon:

\(^1\) https://slideslive.com/38922093/medical-imaging-meets-neurips-4
1 The information of small abnormality features was completely lost through the undersampling process.

2 The information of small abnormality features was not completely lost, but it is attenuated and lays in the tail-end of the distribution, hence is rare.

Were the first hypothesis true, it would be impossible for any method to reconstruct a small abnormality feature, unless the presence of the abnormality is confounded with other structural changes. We are unable to formally verify whether the first hypothesis is always false. Nonetheless, we show that the condition stated in hypothesis 1 is unlikely to occur. Were the second hypothesis true, it would be possible for a reconstruction model to reconstruct it. This is especially true, if the method is data-driven and learning-based. In this work, we show that the second hypothesis is true in many cases and it is possible for a deep learning reconstruction model to reconstruct the small abnormality, using the limited information at disposal.

To investigate these two hypotheses, we define ‘false-negative adversarial feature’ (FNAF), a perceptible small feature that is present in the ground truth MRI but has disappeared upon MRI reconstruction via a learning model.

2. Related works

2.1. MRI Reconstruction with Deep Learning

MRI reconstruction from undersampled k-space has a key-role in fast MRI (Liang et al., 2019; Hammernik and Knoll, 2020). Liang et al. (2019) explains that deep learning-powered MRI reconstruction can be accomplished following either data-driven, model-driven or integrated approaches. Data-driven approaches are generally data hungry and do not require prior knowledge, mainly because they take advantage of huge amount of data to learn the mapping between the raw data and the reconstructed MRI. In model-based approaches, the solution space is restricted by injecting task prior knowledge. This can be obtained for instance by reproducing the iterative approach of compressed sensing. Integrated approaches combine positive aspects of both previous solutions.

2.2. Adversarial attack by small perturbation

Many instances of adversarial attacks, where small imperceptible perturbations are added to images with the aim to attack machine learning models, can be found in the published scientific literature (Biggio et al., 2013; Szegedy et al., 2014; Goodfellow et al., 2015). Goodfellow et al. (2015); Bubeck et al. (2018); Gilmer et al. (2018); Mahloujifar et al. (2019); Shafahi et al. (2018) attempt to explain these adversarial examples by means of a variety of theories. One notable theory is that adversarial examples are a consequence of data scarcity (Schmidt et al., 2018), as the true data distribution is not being captured by non-sufficiently large dataset. Another profound explanation is provided by Ilyas et al. (2019), which shows that adversarial successes are mainly supported by model’s ability to generalize on standard test set by using non-robust features. In other words, adversarial examples are more likely a product of datasets rather than that of machine learning models. To make a model resistant to adversarial attacks without more data, one could employ adversarial training
and provide the model with a prior that remarks the fact that non-robust features are not useful (Goodfellow et al., 2015; Madry et al., 2018). These findings are orthogonal to the second investigated hypothesis. If we interpret the distribution of FNAF as the distribution of robust features, we can attribute FNAF reconstruction failure, to the dataset’s inability to capture FNAF’s distribution.

2.3. Adversarial attack on generative networks

While most of adversarial attacks focus on discriminative models, Kos et al. (2017) propose a framework for attacking variational autoencoders (VAE) and the VAE-GAN. Specifically, input images are imperceptibly perturbed so that the generative models generate target images belonging to a different class. Although reconstruction models can be seen as generative, we differ from this work, as we focus on generating perceptible features that perform un-targeted attacks.

2.4. Adversarial attacks via bi- and three-dimensional transformations or physical attacks

Going beyond small perturbations, a set of more realistic attacks produced by 2D and 3D transformation has been proposed in Xiao et al. (2018); Athalye et al. (2018). Similarly to our work, these studies perform perceptible attacks. Arguably, the most realistic attacks are physical attacks, which are achieved by altering the physical space before an image is captured digitally (Kurakin et al., 2017). Kügler et al. (2018) propose a physical attack on Dermoscopy images by drawing on the skin, around areas of interest. Although these attacks could more easily translate to real world scenarios, it would be nearly impossible to perform physical attacks with imaging modalities such as MRI.

Previously described adversarial attacks do not alter the image semantics. This work builds upon the fact that MRI reconstruction models should reconstruct all features of an under-sampled image. Furthermore, the reconstructed image quality must match that of a fully sampled image or at least guarantee that the complete information is present.

3. Methods

3.1. False negative adversarial attack on reconstruction networks

Madry et al. (2018) formalize that the adversarial attack maximizes the loss of a machine learning model, parameterized by $\theta$. This is feasible by changing $\delta$ in the set of allowed distribution of $S \subseteq R^d$. This can be formalized as:

$$\max_{\delta \in S} L(\theta, x + \delta, y)$$

(1)

We conduct a similar strategy with the exception that in our case, the perturbation $\delta$ is also present in the ground truth whereas the input perturbation is modified. The objective function becomes:

$$\max_{\delta \in S} L(\theta, x + \delta', y + \delta)$$

(2)
with:
\[
\delta' = U(\delta)
\]  
(3)

where \( U \) is an under-sampling function with a k-space mask function \( M \):

\[
U(y) = \mathcal{F}^{-1}(M(\mathcal{F}(y)))
\]  
(4)

We restrict the \( S \) to be a set of visible small bright features in all the locations of an image.

Since we synthetically constructed the small features, we can measure the loss within the area of each features to know exactly if the feature has been reconstructed. In practice, we place a mask on the reconstructed image and the perturbed target image, so that only the area of the small feature is highlighted. The area is relaxed so that a small region at a distance \( d \) from the feature boundaries is also included. The loss was adapted from Kervadec et al. (2018) and Calivá et al. (2019): we only keep the foreground and a small portion of the background with respect to the border of the synthetic feature. This encourages the model to reconstruct a feature and the distinctiveness of that feature compared to the background.

### 3.2. Under-sampling information preservation verification

A benefit of having a synthetic feature generator is that one can quantify the amount of preserved information after k-space under-sampling. To make sure the information of \( \delta \) is preserved through under-sampling in the k-space, we make sure the following condition is fulfilled:

\[
D(x + \delta', x) > \epsilon
\]  
(5)

where \( D \) is a distance function, and \( \epsilon \) is a noise error tolerance threshold. We obtain \( x + \delta' \) and \( x \) through the following:

\[
U(y + \delta) = U(y) + U(\delta) = x + \delta'
\]  
(6)

as \( U \) is linear and closed under addition.

### 3.3. Random search

Random search (Bergstra and Bengio, 2012) has been shown to be an effective optimization technique, especially in a low-dimensional search space. We generate random shapes of feature \( \delta \) at random locations in the image and find the \( \delta \) that maximizes the loss from Equation (2).

### 3.4. Finite-difference approximated gradient ascent

We notice that the location of the \( \delta \) feature is an important factor in finding FNAF. To optimize for the low-dimensional non-differentiable parameter (i.e. set of the \((x, y)\) coordinates of \( \delta \)), we approximate the partial derivatives for each parameter \( p \) with the finite central difference:

\[
\frac{\partial L}{\partial p} = \frac{L(p + \frac{h}{2}) - L(p - \frac{h}{2})}{h}
\]  
(7)

where \( h \) is the step size. Gradient ascent is used to update the location parameter \( p \) to maximize Equation (2).
3.5. FNAF-robust training

Small perturbations-based adversarial training requires models to be trained only on the robust features (Madry et al., 2018). Our attack formulation allows the reconstruction models to undergo standard training and adversarial training at the same time. This allows us to do FNAF-robust training on a pre-trained model and speed up convergence. We further adopt ideas from Shafahi et al. (2019) to accelerate training. In essence, to do FNAF-robust training, our model uses a training set which includes the original examples, the adversarial examples and also all the examples that were generated during the search for the worst case adversarial examples.

4. Experiments and results

4.1. Experimental setup

We conduct our experiments on the FastMRI knee dataset with single-coil setting, including 4x and 8x acceleration factors (Zbontar et al., 2018).

We evaluate our methods with two 2-D deep learning based methods, U-Net (Ronneberger et al., 2015), and invertible Recurrent Inference Machines (I-RIM) (Putzky and Welling, 2019) – the winner of the single-coil FastMRI challenge. For U-Net, we followed the training procedures described in Zbontar et al. (2018). For I-RIM we followed the training procedures described in Putzky et al. (2019).

4.2. Implementation details

We perform the FNAF attack on the models with a mean-square error (MSE) loss. Considering that MRI image pixel values were clamped within the range (-6, +6), we constrained the FNAF to comprise 10 connected pixels, with each pixel values ranging between 3 and 6. Finally, the attack mask was placed within the center 120-by-120 crop of the image. The constraint is to ensure the feature is small, bright and placed in a reasonable location. For random search, 11 randomly shaped FNAFs were generated at random locations for each sample in the validation set and the highest adversarial loss is recorded. For finite-difference gradient ascent (FD), we perform the optimizations for the location x and y in 2 iterations. The number of iteration are chosen to have a reasonable computation time and keep the number of forward passes for one sample constant for both methods. We choose the $h$ to be 10 and the learning rate to be $10^{-5}$.

The attack is rejected when the information-preservation (IP) loss is lower than 0.0001. This is especially important for FNAF-robust training, as we do not want the FNAF-robust model to go to the other extreme and produce hallucination of non-existing features.

For FNAF-robust training, we follow the training procedure described in Section 4.2. The adversarial loss is weighted by 100 to force the model to focus on the small features. To ensure we do not overfit to the validation in terms of the FNAF attack success, we pick the best model in terms of the standard reconstruction loss on the validation set, ignoring the adversarial loss.
Table 1: Standard validation set evaluation with SSIM and normalized mean-square error (NMSE)

<table>
<thead>
<tr>
<th></th>
<th>SSIM</th>
<th>NMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>0.6851 ± 0.2869</td>
<td>0.04212 ± 0.05527</td>
</tr>
<tr>
<td>I-RIM</td>
<td>0.689 ± 0.2896</td>
<td>0.04092 ± 0.05566</td>
</tr>
<tr>
<td>FNAF-robust U-Net</td>
<td>0.6846 ± 0.2873</td>
<td>0.04215 ± 0.05538</td>
</tr>
</tbody>
</table>

Table 2: FNAF attack evaluations.

<table>
<thead>
<tr>
<th></th>
<th>RS (Attack Rate %)</th>
<th>FD (Attack Rate %)</th>
<th>RS (MSE)</th>
<th>FD (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>85.28</td>
<td>77.07</td>
<td>0.001561</td>
<td>0.001423</td>
</tr>
<tr>
<td>I-RIM</td>
<td>81.52</td>
<td>69.57</td>
<td>0.001516</td>
<td>0.001366</td>
</tr>
<tr>
<td>FNAF-robust U-Net</td>
<td>14.98</td>
<td>12.09</td>
<td>0.000516</td>
<td>0.000444</td>
</tr>
</tbody>
</table>

4.3. Attack evaluation metrics

We calculate the average attack loss for the validation set and the hit rate with the attack. We consider the attack to be a hit, when the loss is higher than a threshold value \( k \). From observation, we choose 0.001 for \( k \) as the feature is mostly lost if the loss is higher than that. To keep the hit rate conservative, \( k \) is set to be high so there might be cases where the feature is lost when the loss is lower than \( k \). The actual hit rates are likely higher than the values reported in this work.

4.4. Attack results

The result of the attack shown in Table 2 confirms that hypothesis 2 is true in many cases. The attack with FD is weaker than that with random search (RS), which is counter-intuitive. This might be due to various reasons, such as the tuning of the optimizer hyper-parameters, the small number of iterations, etc. Nonetheless, the high success rate of the random search method for both models shows that it is fairly easy to find a FNAF in the search space that we have heuristically defined. Although I-RIM is more resilient to the attacks than U-Net, the attack rate is still fairly high. This is concerning but also understandable given that deep learning methods are not explicitly optimized for such objective, so these FNAFs are at the tail-end of the distribution or even out-of-distribution with respect to the model. Fortunately, we can modify the objective as specified in Section 3.5 to produce a FNAF-robust model which is empirically fairly resilient to the attacks.

4.5. Under-sampling information preservation verification

To investigate hypothesis 2, we measure the acceptance rate of the adversarial examples based on the information-preservation loss. A very high acceptance rate is observed across all settings, showing that in most cases the small feature’s information is not completely
Figure 1: The top row shows a ”failed” FNAF attack. The bottom row shows a ”successful” FNAF attack. Row 1 contains the under-sampled zero-filled images. Row 2 contains the fully-sampled ground truth images. Row 3 contains U-Net reconstructed images. Row 4 contains FNAF-robust U-Net reconstructed images. (C-G-D-H) FNAF reconstruction: (C) adversarial loss of 0.000231. (G) adversarial loss of 0.00206. (D) adversarial loss of \(8.25 \cdot 10^{-5}\). (H) adversarial loss of 0.000305.

Table 3: Information preservation

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>U-Net FNAF</th>
<th>I-RIM FNAF</th>
<th>Robust U-Net FNAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptance Rate (%)</td>
<td>99.82</td>
<td>99.72</td>
<td>99.78</td>
<td>99.34</td>
</tr>
<tr>
<td>IP Loss (MSE)</td>
<td>0.00064</td>
<td>0.00050</td>
<td>0.00050</td>
<td>0.00052</td>
</tr>
</tbody>
</table>

lost through under-sampling, at least for the way we constructed the features. We suspect that the same could hold true for real-life abnormalities.

Figure 2 shows a small negative correlation between IP loss and FNAF loss. In fact, we expect that more information would weaken the attack. However, such negative correlation is weak, indicating that there is no simple linear correlation. Therefore the preservation of information alone cannot predict the FNAF-robustness of the model. So the information loss due to under-sampling is a valid but insufficient explanation for the existence of FNAFs.
4.6. Location distribution of adversarial features

We visualize the location distributions of the worst case FNAF on the image. There seems to be no apparent pattern to the location of the FNAF. However, the location distributions seems to be similar across non-FNAF-robust models. We investigate this in the next section.

4.7. Transferability of adversarial features across reconstruction networks

We take FNAF examples from U-Net and apply them to I-RIM, and observed a 92.13% attack rate. The high transferability is similar to what is observed in Goodfellow et al. (2015) and Alcorn et al. (2019). This is indicating that the training data does not capture the distribution of FNAF.
Table 4: Abnormality Reconstructions

<table>
<thead>
<tr>
<th></th>
<th>Cartilage Lesion Rate</th>
<th>Meniscus Lesion Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>1/8</td>
<td>8/9</td>
</tr>
<tr>
<td>FNAF-robust U-Net</td>
<td>3/8</td>
<td>9/9</td>
</tr>
</tbody>
</table>

4.8. Generalization to real-world abnormalities

A musculoskeletal (MSK) imaging trained M.D. inspected and identified abnormalities of clinical relevance in 51 volumes from the validation set. The abnormalities include cartilage lesions, meniscal tears, and meniscal degenerations.

The results in Table 4 show that the FNAF-robust U-Net is marginally better out of the small number of abnormalities found. Although further extensive evaluation is needed, this is an encouraging result, considering that there is no guarantee that the synthetic feature would look like real-world abnormalities. The detailed comments of the abnormality findings are included in Appendix A.

5. Conclusions

In this work, we investigate two hypothesis for the false negative problem in deep-learning-based MRI reconstruction. By developing the FNAF adversarial robustness framework, we show that this problem is difficult, but not impossible. Within this framework, there is much potential to bring the extensive theoretical and empirical ideas from the adversarial robustness community, especially in the area of provable defenses (Wong and Kolter, 2018; Mirman et al., 2018; Raghunathan et al., 2018; Balunovic and Vechev, 2020) to tackle the problem. We also hope to inspire future work in the direction of defining a better (realistic) search space for the FNAF, towards generalization to real-world abnormalities. The findings from Appendix A could be a toy validation set for future works.

Acknowledgments

Acknowledgments withheld.

References


Figure 4: (A) Ground truth: small cartilage lesion in femur. (B) U-Net: Area of cartilage lesion not defined and resembles increased signal intensity. (C) FNAF-robust U-Net: Cartilage lesion preserved but less clear.


Appendix A. Detailed Comments of Real-World Abnormalities
Table 5: Comments of the MSK radiologist involved in the study. Cases where FNAR-robust U-Net improved compared to U-Net are bolded.

<table>
<thead>
<tr>
<th>File</th>
<th>Slice number</th>
<th>Comments on ground truth</th>
<th>Comments on U-Net reconstruction</th>
<th>Comments on FNAR-robust U-Net reconstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>27</td>
<td>Signal change</td>
<td>Original lesion preserved but less clear</td>
<td>Original lesion preserved</td>
</tr>
<tr>
<td>26</td>
<td>16</td>
<td>Cartilage lesion</td>
<td>Cartilage lesion in original now looks like signal change</td>
<td>Cartilage lesion in original now looks like signal change</td>
</tr>
<tr>
<td>52</td>
<td></td>
<td>Metal artifacts in tibia</td>
<td>No change in metal artifacts</td>
<td>Metal artifacts preserved</td>
</tr>
<tr>
<td>71</td>
<td>23</td>
<td>Cartilage lesion in tibia</td>
<td>Original cartilage lesion not seen</td>
<td>Original cartilage lesion preserved but less clear</td>
</tr>
<tr>
<td>73</td>
<td>23</td>
<td>Intrasubstance degeneration</td>
<td>Intrasubstance degeneration preserved</td>
<td>Intrasubstance degeneration preserved</td>
</tr>
<tr>
<td>107</td>
<td>16</td>
<td>Cartilage lesion in femur</td>
<td>Original cartilage lesion not seen</td>
<td>Original cartilage lesion not seen</td>
</tr>
<tr>
<td>114</td>
<td>26</td>
<td>Vertical tear in meniscus</td>
<td>Original tear preserved but less clear</td>
<td>Original tear preserved but less clear</td>
</tr>
<tr>
<td>178</td>
<td>14-21</td>
<td>Meniscectomy</td>
<td>Menisectomy preserved</td>
<td>Menisectomy preserved</td>
</tr>
<tr>
<td>196</td>
<td>26</td>
<td>Horizontal meniscal tear</td>
<td>Meniscal tear preserved</td>
<td>Meniscal tear preserved</td>
</tr>
<tr>
<td>201</td>
<td>25</td>
<td>Signal change in femoral cartilage</td>
<td>Cartilage lesion not preserved</td>
<td>Cartilage lesion preserved but less clear</td>
</tr>
<tr>
<td>267</td>
<td>24</td>
<td>Meniscal tear</td>
<td>Original tear preserved but less clear</td>
<td>Original tear preserved but less clear</td>
</tr>
<tr>
<td>280</td>
<td>22</td>
<td>Cartilage lesion in tibia</td>
<td>Original lesion not preserved</td>
<td>Original lesion not preserved</td>
</tr>
<tr>
<td>314</td>
<td>14-20</td>
<td>Meniscal degeneration/menisectomy</td>
<td>Meniscal degeneration preserved</td>
<td>Meniscal degeneration preserved</td>
</tr>
<tr>
<td>325</td>
<td>24</td>
<td>Signal change in cartilage</td>
<td>Original cartilage lesion not preserved</td>
<td>Signal change in cartilage partially preserved</td>
</tr>
<tr>
<td>356</td>
<td>21</td>
<td>Cartilage lesion</td>
<td>Original lesion preserved but not clear</td>
<td>Original cartilage lesion preserved</td>
</tr>
<tr>
<td>464</td>
<td>26</td>
<td>Intrasubstance degeneration</td>
<td>Intrasubstance degeneration preserved but not clear</td>
<td>Intrasubstance degeneration preserved</td>
</tr>
<tr>
<td>480</td>
<td>21</td>
<td>Cartilage lesion</td>
<td>Cartilage lesion not preserved</td>
<td>Cartilage lesion not preserved</td>
</tr>
<tr>
<td>528</td>
<td>28</td>
<td>Intrasubstance degeneration</td>
<td>Intrasubstance degeneration not preserved</td>
<td>Intrasubstance degeneration preserved</td>
</tr>
</tbody>
</table>