Adversarial Robustness of Deep Inpainting Models

Gowthami Somepalli[1], Phillip Pope[1], Soheil Feizi[1]
[1]University of Maryland, College Park
{gowthami,ppope,soheil}@cs.umd.edu

Abstract

The image inpainting is an important task in computer vision with applications ranging from restoration of damaged photographs to removing artifacts from medical images. Several inpainting methods based on deep generative models such as GANs and flow-based models or discriminative Image-to-Image Translation models have been proposed. However, the robustness of inpainting models against adversarial attacks has not been studied so far; i.e. the possibility of adding imperceptible perturbations to input images to the inpainter to degrade its performance. Despite differences in the inherent nature of various inpainting models, through extensive experiments, we show that they are prone to adversarial attacks. In particular, we consider an image inpainting methods based on Flow-based generative models (e.g. Conditional-GLOW) and develop adaptive adversarial attacks to break the model. We observe that our adversarial attacks on Conditional-GLOW is quite successful. Finally, we robustify inpainting model using an adversarial training approach. We observe that adversarially trained inpainter is robust against adversarial attacks while obtaining on par performance with original models on clean examples.

1. Introduction

The robustness is an intrinsically important desiderata of machine learning models, and inpainting models are no exception. The need for robustness arises from the failure of modeling assumptions, which often occurs in practice. Much recent work in the machine learning community has focused on *adversarial* robustness, the amelioration of the well known sensitivity of deep models towards adversarial perturbations. In this work we extend this line of research to inpainting models, asking the two-fold questions (1) *do adversarial attacks exist for inpainting models*? and (2) *can they be made robust towards such attacks*?

Given an image with missing pixels, inpainting is the process of restoring the image with plausible content. It has several applications in photo restoration, image retouching, in object elimination in photos etc. Most of the early techniques are single-image methods where they use the properties of the image to fill its holes. For example Total Variation based approaches used the smoothness property of images to fill the gaps [1]. Patch based methods search for relevant patches in the image or other images in dataset in recursive fashion to fill the holes [3, 9, 2]

Deep inpainting models leverage learned priors to fill the missing information, and significantly improved state of the art. Most work in deep inpainting has focused an encoder-decoder architectures with specialized connections, convolutions and losses [14, 6, 18, 19, 10, 20]. Recent inpainting models are based on structured predictions techniques like using conditional normalizing flows [12].

In this paper, to the best of our knowledge, for the first time, we present the rigorous evaluation of robustness of several deep inpainting models to adversarial attacks.

Our contributions are as follows:

- We show that c-Glow is quite susceptible to attacks even with small perturbations. And we show how the performance drops in relation to the attack strength.
- We robustify the c-Glow model using *adversarial training*, and show improved resistance towards adversarial attacks.

2. Related work

The following are the three main topics discussed in the paper.

2.1. Inpainting

Image inpainting is a task of filling up the missing regions in a given input image. The earlier models are primarily diffusion based methods or patch based. These techniques relied on neighboring pixels to fill the holes using methods like distance field [15] or they iteratively search for a patches with similar neighborhood features to fill the holes. [2] [3] [4].

The advent of Deep neural networks especially CNNs and GANs changed the landscape of inpainting methods.

Deep inpainting techniques are significantly better at predicting non-texture holes like faces or objects where the traditional methods fail. These methods learn the semantics of the objects by training over large-scale datasets and are able to synthesise the images well in prediction setting. Earlier methods include Context Encoder [14], Context Encoder with Local and Global discriminators [6], DCGANi [18], Deep Image prior [17], Inpainting with Contextual Attention [19] etc. Recent methods include Conditional GLOW, Partial covolution, Gated Convolution etc which utilize latest Deep learning architectures to achieve SOTA inpainting results.

2.1.1 Conditional GLOW

Conditional Glow (C-Glow) is structured prediction model, derived from Glow [7] model with additional neural networks capturing the conditional relationship between input data and the output. A normalizing flow is a sequence of invertible functions $f = f_1 o f_2 o .....f_n$ which takes target y as an input and generate the mapping to the latent variable. In C-GLOW, y is the inpainted image, x is the corrupted image, and each function in flow is modified from f_{ϕ_i} to $f_{\phi_i,x}$. The conditional likelihood of the C-Glow can be written as -

$$\log p(y|x,\theta) = \log p_z(z) + \sum_{i=1}^{M} \log |det(\frac{\partial f_{\phi_i,x}}{\partial r_{i-1}})|$$

where $r_i = f_{\phi_i}(r_{i-1}), r_0 = x, r_M = z$

2.2. Adversarial attacks

To perform an adversarial attack on a model, we seek the smallest perturbation(bound by a limit and sometimes barely noticeable to humans) to cause large errors in the final task the model is trained for. If the attacker have the access to the model architecture, parameters, input data, we call it a white-box attack.

In generative inpainting setting, we can set frame this problem as - x is the input to the inpainter, θ be the learned parameters and y be the ground truth, $L(x, y; \theta)$ be the loss function on which the model is optimized on. We want to find the perturbation δ such that $L(x + \delta, y; \theta)$ is really high even though δ is small. Depending on the inpainting techniques, the loss function changes, for Conditional glow, it is NLL of the normalizing flow.

Fast Gradient Sign Method(FGSM) FGSM [5] is the fastest adversarial attack if not the best attack method. It is a single step If $L(x, y; \theta)$ is the loss function defined for the model, this attack sets the perturbation δ as - $\delta = \epsilon\text{sign}(\nabla L)$ where ϵ is the acceptable perturbation level in each pixel.

This attack has been proposed in the context of image classification but we have adapted it for semantic structured prediction problem of inpainting in this paper.

2.3. Robustness training

The defense strategy to overcome the earlier white-box attacks has been explored by [8, 13]. The adversarial training as proposed by [13] on a model with x, y, θ as input, output and model parameters respectively is modeled on minmax objective -

$$\min_{\theta} E_{(x,y)}[\max_{\delta \in S} L(x + \delta, y; \theta)]$$

where δ is the perturbation, S is the acceptable set of perturbation. This is same as iteratively training the model on the in-situ generated adversarial examples. Even though [16] found that adversarially trained models are still susceptible to black-box attacks generated from other networks, it still remains to be one of the most successful/ easy-to-train defense mechanisms which can withstand many white-box attacks.

3. Approach

In this paper we are performing a white box attack: we assume the attacker has access to a copy of the model with which to devise the attack.

Even though the end goal is the same, the models we considered here differ in philosophy, architecture and implementation. The relative ease of attack on the models vary too.

Let x be the input to the model, image that needs to be inpainted, with square occlusion in the center in this case. Let y be the ground truth and $L(x, y; \theta)$ is the loss function defined for optimizing the model. If we have access to the ground truth, for FGSM attack, we can perturbate the x in the direction $\nabla_x L(x, y; \theta)$.

Since the attacker does not always have access to ground truth, we have performed our attacks on the prediction of the model. Let's y' be the prediction on x for a given model. We can calculate the perturbed x using the following equation -

$$x_{pert} = x + \epsilon.\text{sign}(\nabla_x L(x, y'; \theta))$$

Now depending on the model, loss function L changes. In case of C-Glow, the loss function is negative log-likelihood

4. Experiments and Results

Evaluation metrics As noted in [18, 20], inpainting models lack good quantitative metrics. Commonly used is the *peak signal to noise ratio* (PSNR) and l_1 norm based metrics. These metrics are flawed since there may be many qualitatively good restorations of an image which may

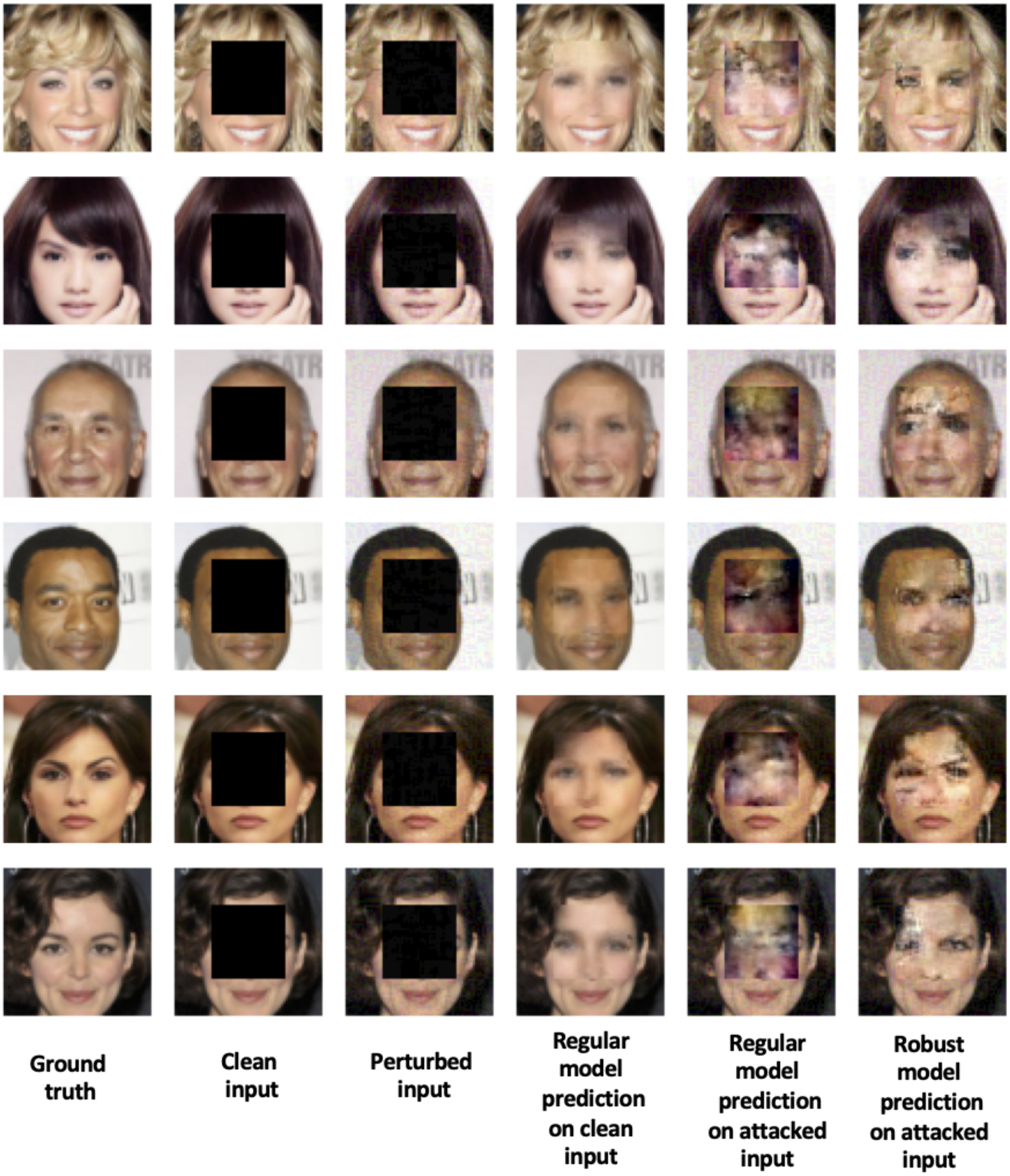
Ground truth
Clean input
Perturbed input
Regular model prediction on clean input
Regular model prediction on attacked input
Robust model prediction on attacked input

Figure 1. Select results of c-GLOW attacks. From left to right the panels are (1) Ground truth (2) Clean input (3) Perturbed input (4) Prediction of clean input on clean model (5) Prediction of attacked input on clean model (6) Prediction of attacked input on robust model

nonetheless have large discrepancies from the ground truth. Quantitative metrics are essential for experimental validation. We note these flaws and proceed in spite of them.

Datasets We have trained the C-Glow model on CelebA dataset [11] since the training details and hyperparameters shared are relevant to this dataset. It is a largescale dataset with 202,599 celebrity face images.

Table 1. Comparision of various metrics of regular and robust trained inpainting models. The models are trained on adversarial examples at $\epsilon = 8$, all the attack results presented here are at $\epsilon = 8$.

Model type:	Metric	Regular model		Robust model	
Data:		Clean	Attack	Clean	Attack
c-Glow	PSNR	25.20 ± 0.32	18.42 ± 0.10	25.26 ± 0.09	22.03 ± 0.07
c-Glow	$l1$-norm	0.02 ± 0.00	0.07 ± 0.00	0.02 ± 0.00	0.05 ± 0.00
c-Glow	SSIM	0.89 ± 0.01	0.68 ± 0.01	0.89 ± 0.00	0.76 ± 0.00

Adversarial Attacks For this paper, c-Glow is trained only on data with a square occluded region at the center.The mask is of size 64 x 64 on CelebA images, and 128 x 128 on CelebA-HQ images. With this setting, we are able to replicate the PSNR values claimed in paper for C-Glow on the clean models.

We performed white-box attacks on c-Glow with perturbation limit set to $\epsilon = 8$. For the attack, we have calculated the perturbation on the input masked image using FGSM attack. We have calculated the perturbation with respect to negative log-likelihood (NLL) function.

The results of the clean model on clean input and attacked input may be seen in Figure 1. We observe c-Glow is quite sensitive to the perturbations in the input. For c-Glow, the prediction on the attacked input is just random pixels with no semblance to facial features.

Robustness results We have performed the adversarial training using the attacked images at perturbation level $\epsilon = 8$. In every iteration, for a given batch, we find the adversarial examples insitu and use them for training. You can see in the Table 1 quantitative comparison of various metrics in regular vs robust Conditional GLOW models. The values are mean and standard deviation values across 10 batches of size 200 each. The adversarially trained model seems to be outperforming the regular trained model in both attack and no-attack scenarios.

You can see in the Figure 2, qualitative results of attacks at different epsilons for each of the models. These are white-box attacks based on FGSM.

5. Conclusion

We propose an adversarial attack and robustness training formulation for deep inpainting models. In particular we study c-GLOW and show that it is quite susceptible to adversarial attacks. We further train the model on adversarial examples and show such models are more robust to adversarial attacks. While we explored DCGANi[18], it is a fixed generator model, hence adversarial attack is essentially not possible. In future, we are planning to explore different adversarial attacks like single-pixel attack and so on.

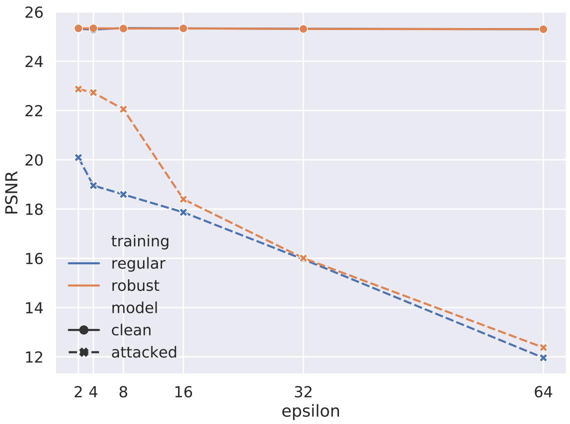

Figure 2. Regular vs Robust model performance at different levels of attack perturbations on c-Glow model.

References

[1] Manya V Afonso, José M Bioucas-Dias, and Mário AT Figueiredo. An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE Transactions on Image Processing*, 20(3):681–695, 2010. 1

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24. ACM, 2009. 1

[3] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001. 1

[4] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999. 1

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[6] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 1, 2

[7] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018. 2

[8] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2

[9] Vivek Kwatra, Irfan Essa, Aaron Bobick, and Nipun Kwatra. Texture optimization for example-based synthesis. In *ACM SIGGRAPH 2005 Papers*, pages 795–802. 2005. 1

[10] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 1

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 3

[12] You Lu and Bert Huang. Structured output learning with conditional generative flows. *arXiv preprint arXiv:1905.13288*, 2019. 1

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[14] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1, 2

[15] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004. 1

[16] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 2

[17] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 2

[18] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017. 1, 2, 4

[19] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 1, 2

[20] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019. 1, 2