Robust Synthesis of Adversarial Visual Examples Using a Deep Image Prior

Thomas Gittings[1], Steve Schneider[2], and John Collomosse[1,3]

[1]Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, UK.
[2]Surrey Centre for Cyber Security, University of Surrey, UK.
[3]Adobe Research, Creative Intelligence Lab, San Jose, CA.
{t.gittings,s.schneider,j.collomosse}@surrey.ac.uk

Abstract

We present a novel method for generating robust adversarial image examples building upon the recent 'deep image prior' (DIP) that exploits convolutional network architectures to enforce plausible texture in image synthesis. Adversarial images are commonly generated by perturbing images to introduce high frequency noise that induces image misclassification, but that is fragile to subsequent digital manipulation of the image. We show that using DIP to reconstruct an image under adversarial constraint induces perturbations that are more robust to affine deformation, whilst remaining visually imperceptible. Furthermore we show that our DIP approach can also be adapted to produce local adversarial patches ('adversarial stickers'). We demonstrate robust adversarial examples over a broad gamut of images and object classes drawn from the ImageNet dataset.

1. Introduction

Convolutional neural networks (CNNs) are known to be vulnerable to adversarial examples; inputs that exhibit marginal visual difference from the distribution of correctly classified images but cause dramatically different classification decisions [20, 9]. These 'adversarial images' may be crafted to attack an *a priori* trained network, via back-propagation of gradients to induce pixel-level perturbations of a target image either globally [20, 9, 15, 2] or within a restricted local region ('adversarial stickers' [1, 6]) in order to nudge the classification outcome over the decision boundary. The enthusiastic adoption of neural networks *e.g.* for autonomous vehicles and robotics opens a new facet of cyber-security, motivated on one hand to train networks resilient to such attacks, and on the other to evaluate resilience by developing new attacks.

The contribution of this paper is a new algorithm for synthesizing adversarial image examples that exhibit equal or better robustness to image deformation than contemporary methods, whilst maintaining low perceptibility to a human observer. Currently adversarial image attacks are limited in their scope and real-world applicability due to either their fragility or the perceptibility of the perturbations introduced to the target image. The perturbations typically induced to generate an adversarial image manifest as high-frequency noise (see Fig. 1). The capacity of perturbations to induce misclassification is thus greatly attenuated by simple image deformations such as re-sampling under affine transformation, which are inherent in any eventual manifestation of an attack 'in the wild'. Furthermore, high-frequency artifacts are readily detectable by the human visual system so revealing the presence of attacks.

In this work we explore an alternative method for generating covert adversarial image examples, leveraging the recently proposed 'Deep Image Prior' (DIP) for image reconstruction [22]. The surprising result of DIP is that the statistics of natural images can be encoded through a CNN architecture; *i.e.* the CNN structure rather than actual weights of the filters. The translation equivariance of CNNs enables DIP to exploit the internal recurrence of visual texture in images [12], in a similar way as the classical non-parametric patch based approaches to texture synthesis [8]. Under DIP an image is reconstructed by training a deep encoder-decoder CNN from scratch (random weights) to overfit a reconstruction loss function for that single image. Our core technical contribution is to frame adversarial image synthesis as a reconstruction problem, leveraging the DIP to reconstruct an image from a randomly initialized (white noise) image under a dual reconstruction and adversarial loss. Reconstructing

Figure 1. Overview: We propose a reconstruction-based approach to synthesising adversarial image attacks using DIP (middle) that produces images more resilient to affine transformation without human perceptible artifacts vs. the original image (left) than state of the art adversarial image attacks *e.g.* Carlini and Wagner [2] (right).

the image under this constraint affords greater flexibility for perturbation across the whole image, in contrast to existing methods that rely upon backpropagation to update pixel values from a local minimum (the initial target image). The resulting perturbations are regularized by the DIP to resemble the appearance of natural images, further mitigating against sporadic high frequency noise patterns characteristic of adversarial images. Our reconstruction framework can be used to induce perturbations within the whole image, that are hard to perceive yet exhibit superior robustness to affine transformation than state of the art adversarial image algorithms. Further we show that our method can also be adapted to restrict perturbations to a local region of interest resulting in adversarial stickers that are competitive with the state of the art at inducing misclassification.

2. Related Work

Adversarial attacks on deep networks for visual recognition have received significant attention in recent years, prompted by the step change in performance delivered by CNNs across diverse object classes [10, 19]. Szegedy *et al.* pioneered adversarial attacks for visual classifiers [20], demonstrating that minor perturbations of pixel values can induce significant CNN misclassification rates despite little human perceptible visual difference. Goodfellow *et al.* demonstrated linearity of this effect in input space, introducing the fast gradient sign method (FGSM) [9] to quickly compute adversarial perturbations via backpropagation without need for solving costly optimizations. An iterative form of this method for more robust attacks was later presented [11]. DeepFool [15] explictly optimises to minimise perceptibility and maximise robustness in order to form adversarial examples. Carlini and Wagner optimise for similar goals using alternative norms [2]. All of the above induce high frequency noise within the whole image with limited human

perceptibility. Whilst the attacks are covert they are also fragile to image resampling *e.g.* due to image transformation or printing. Adversarial patches take a complementary, overt approach via synthesis of vivid 'stickers' [1] that occupy only a small region yet induce misclassification [1, 6] or misdetection [3, 21]. Such approaches are reminiscent of attacks against classical facial detection algorithms [7, 18] in that they can be physically manifested for real-world deployment. Nevertheless, like whole image methods, such stickers are sensitive to affine transformation limiting their use in practical attacks [13] for the time being. Like prior work, we take an optimization approach but uniquely leverage an encoder-decoder network and Deep Image Prior (DIP) [22] to reconstruct the attack image from scratch incorporating an adversarial constraint. The latitude afforded by this reconstruction mitigates the introduction of sporadic noise which both improves robustness and limits perceptibly of attacks. Whilst our focus is on whole image covert attack our method can also be adapted to synthesise overt attacks via adversarial patches.

3. Methodology

Let $f : \mathbb{R}^m \to \{1, \ldots, l\}$ be a CNN classifier taking image pixel values and returning an object class label, and let $J_f : \mathbb{R}^m \times \{1, \ldots, l\} \to \mathbb{R}^+$ be its associated loss function. For an image $x \in \mathbb{R}^m$ we aim to find a small perturbation $r \in \mathbb{R}^m$ such that $f(x+r) \neq f(x)$. We say that $x+r$ is an *adversarial image*. If we pick a class $c \in \{1, \ldots, l\}$ and aim to get $f(x+r) = c$ then we say the adversarial example is *targeted*, and if not we say it is *untargeted*. We let $P_f(x)$ be the output probabilities of the classifier, and we also let $Z_f(x)$ be the output to the logits layer of the classifier. For consistency, subscripts to indicate the components of vectors, and bracketed superscripts to indicate iterations, i.e. x_i is the ith component of x and $x^{(i)}$ is the ith iteration of x.
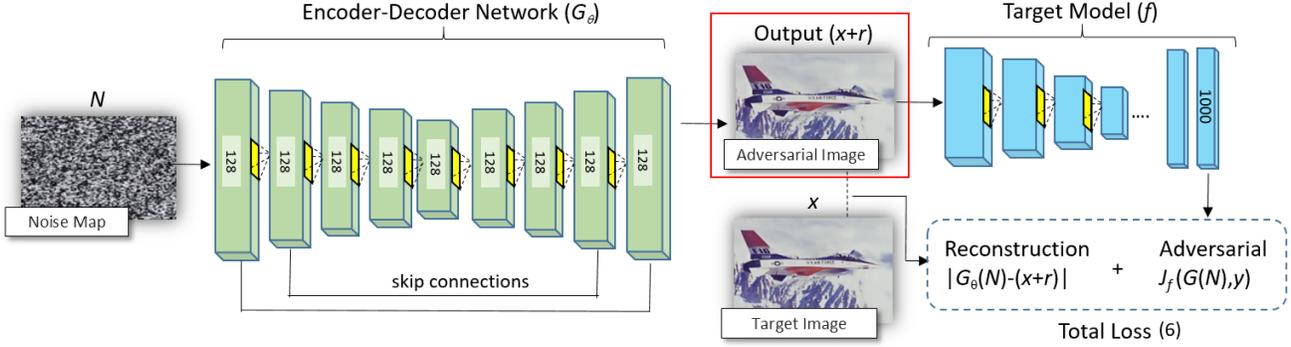
Figure 2. Proposed architecture for adversarial image generation via image reconstruction (DIP). Model $G_\theta(.)$ is trained on single target image x to map $N \mapsto (x+r)$ such that $f(x+r) \neq f(x)$ for pre-trained target CNN (the subject of the attack) under dual loss (6) comprising reconstruction and adversarial loss terms.

We motivate our method by first briefly recapping fast gradient methods which create minor image perturbations via a short (ε) linear jump in the input domain determined via backpropagation through f:

$$r = \varepsilon \frac{\nabla_x J_f(x,y)}{\|\nabla_x J_f(x,y)\|}. \qquad (1)$$

We try to choose ε as small as possible such that the attack is still effective. A popular and fast approximation (FGSM) due to Goodfellow *et al.* [9] is to set $r = \varepsilon \, \mathrm{sign}(\nabla_x J_f(x,y))$. A piece-wise linear extension of FGSM [11] is to perform FGSM iteratively in the hope of obtaining a successful adversarial example without having to make ε large. More concretely, we set $r^{(0)} = 0$, and then for $t > 0$ we have

$$r^{(t+1)} = r^{(t)} + \alpha \, \mathrm{sign}\left(\nabla_x J_f\left(x + r^{(t)}, y\right)\right) \qquad (2)$$

where the $\alpha = \varepsilon/T$ for T the number of iterations. The *momentum iterative method* [5] adds a momentum term to this basic iterative method in order to escape from local minima. As before we set $r^{(0)} = 0$ but also $g^{(0)} = 0$. Now for $t > 0$ we iterate using:

$$g^{(t+1)} = \mu g^{(t)} + \frac{\nabla_x J_f(x + r^{(t)}, y)}{\|\nabla_x J_f(x + r^{(t)}, y)\|_1} \qquad (3)$$

$$r^{(t+1)} = r^{(t)} + \alpha \, \mathrm{sign}(g^{(t+1)}) \qquad (4)$$

where μ is a decay factor that degenerates to the basic iterative method as t increases.

3.1. Adversarial Images via DIP Reconstruction

The disadvantage of FGSM and derivative methods is their dependence upon backpropagation to perturb the image from a local minima (x). Despite mitigation strategies (*e.g.* momentum) all tend to converge to an r composed of high frequency speckle noise (c.f. Fig. 3) which presents a trade-off between visibility (high ε) and fragility (low ε).

We therefore propose an alternative to these local methods, leveraging the recently proposed Deep Image Prior (DIP) [22] to synthesise $x+r$ via global image reconstruction, such that $f(x+r) \neq f(x)$.

The core idea is to learn a generative CNN G_θ (where θ are the network parameters) to reconstruct x from a noise map N of identical height and width to x, with pixels drawn from a *uniform* random distribution. We use a symmetric encoder-decoder network architecture with skip connections (Fig. 2) for G_θ, comprising five (up-)convolutional layers with filter size 3×3 and 128 en/decoder channels per layer, skip connections between all layer pairings. Ulyanov *et al.* originally proposed a reconstruction loss to learn G_θ for image restoration applications such as denoising, in-painting and super-resolution:

$$\theta^* = \arg\min_\theta \|G_\theta(N) - x\|_2^2 \qquad (5)$$

where $\|.\|_2$ is the L_2 norm. In practice the generator G_θ provides implicit regularisation via the structure of the CNN in lieu of an explicit regularisation in the loss (*e.g.* total variation loss [14]). Whilst G_θ is under-constrained, its output converges faster to natural images than to unnatural ones, thus DIP provides a useful reconstruction prior.

We propose an alternative, dual term loss function to learn θ to synthesise $G_\theta(N) = x + r$ that both approximates and misclassifies x:

$$\theta^* = \arg\min_{\theta^{(t)}} J_f(G_{\theta^{(t)}}(N), y) + \lambda \|G_{\theta^{(t)}}(N) - x\|_2^2 \qquad (6)$$

where in our work $\lambda = 10^{-5}$ balances the order of magnitude of the adversarial (first) and reconstruction (second) terms, and $\theta^{(t)}$ indicates network weights at iteration t of training. We train to overfit θ to x via the Adam optimizer yielding the adversarial perturbation $r = G_{\theta^*}(N) - x$. We explore popular CNN configurations for $\{f, J_f\}$ to attack in subsec. 4.2.

Arch.	Method	None	Rot-S	Rot-L	Scale-S	Scale-L	JPEG
VGG	DIP *(Ours)*	**1.00**	**0.97**	**0.45**	**0.99**	**0.65**	**0.63**
	FGSM-IterL [11]	1.00	0.99	0.76	1.00	0.84	0.99
	FGSM-IterS [11]	**1.00**	0.53	0.07	0.79	0.11	0.11
	L-BFGS [20]	**1.00**	0.45	0.05	0.78	0.08	0.06
	SMM [16]	0.59	0.26	0.04	0.37	0.07	0.10
	C&W [2]	**1.00**	0.37	0.05	0.65	0.07	0.07
	DeepFool [15]	0.98	0.87	0.37	0.94	0.48	0.57
GNet	DIP *(Ours)*	0.93	**0.91**	**0.59**	0.92	**0.79**	**0.71**
	FGSM-IterL [11]	0.88	0.87	0.72	0.88	0.79	0.86
	FGSM-IterS [11]	0.75	0.39	0.22	0.49	0.25	0.23
	L-BFGS [20]	0.98	0.40	0.07	0.71	0.11	0.07
	SMM [16]	0.25	0.10	0.04	0.11	0.05	0.05
	C&W [2]	**0.99**	0.34	0.07	0.55	0.10	0.07
	DeepFool [15]	0.98	0.77	0.40	0.86	0.53	0.45

Table 1. Misclassification rate induced by attack for all methods and architectures trained over ImageNet, for several affine transformations. Note FGSM-IterL is an overt attack (FGSM with large ε) for comparison.

Baseline	Ours Less Visible	No Difference	Ours More Visible
FGSM-IterL [11]	**0.76**	0.16	0.08
FGSM-IterS [11]	0.14	**0.50**	0.36
L-BFGS [20]	0.14	**0.52**	0.34
SMM [16]	0.24	**0.52**	0.24
C&W [2]	0.17	0.05	**0.78**
DeepFool [15]	0.04	0.37	**0.59**

Table 2. Perceptibility study. Quantifying visibility of perturbations due to our DIP method vs. baselines; values indicate fraction of response in which our method has less, equally, or more visible artifacts vs. a baseline method (over 100 classes).

Method	FGSM-IterS	FGSM-IterL	L-BFGS	SMM	C&W	DeepFool	Ours
Time(m)	0.10	0.02	2.88	0.72	1.87	0.02	4.49

Table 3. Time taken to generate adversarial image examples for all methods.

3.1.1 Local patch based attack

Our method can optionally be adapted to an overt attack, in which an adversarial patch is synthesised and composited into a region of an image in order to induce misclassification. We define a region of interest (ROI) via binary mask $M \in [0, 1]$. In this case we seek perturbation r to create a composite image $\hat{x} = x + r^{(t)}$:

$$\hat{x} = M \odot G_{\theta^{(t)}}(N) + (1 - M) \odot x \qquad (7)$$

where $\odot$ is element-wise multiplication. We optimize as before, but without any reconstruction constraint in the loss:

$$\theta^* = \underset{\theta^{(t)}}{\arg\min} J_f(\hat{x}, y). \qquad (8)$$

A single adversarial patch capable of attacking multiple images (similar to the adversarial stickers of Brown *et al.* [1]) can be created by sampling x in mini-batches from a set of training images (versus learning θ over a single image, as

in the whole image case). We evaluate the performance of such stickers in subsec. 4.3.

4. Evaluation and Discussion

We evaluate the performance of our method in terms of its efficacy and robustness against affine image transformation, and the perceptibility of artifacts introduced into the covert adversarial image examples created.

4.1. Experimental Setup

We evaluate our approach against 6 baselines (subsec.4.1.1) over 2 popular architectures: VGG-19 and GoogLeNet Inception v3 trained using ImageNet [4] and evaluate on a test set of 1000 images (hereafter, ImageNet-TS1K) comprising a random image sampled from each category in the ImageNet test partition. We evaluate the *success rate*, defined as the fraction of target images in which the method induces an incorrect classification decision. For covert attacks we also determine
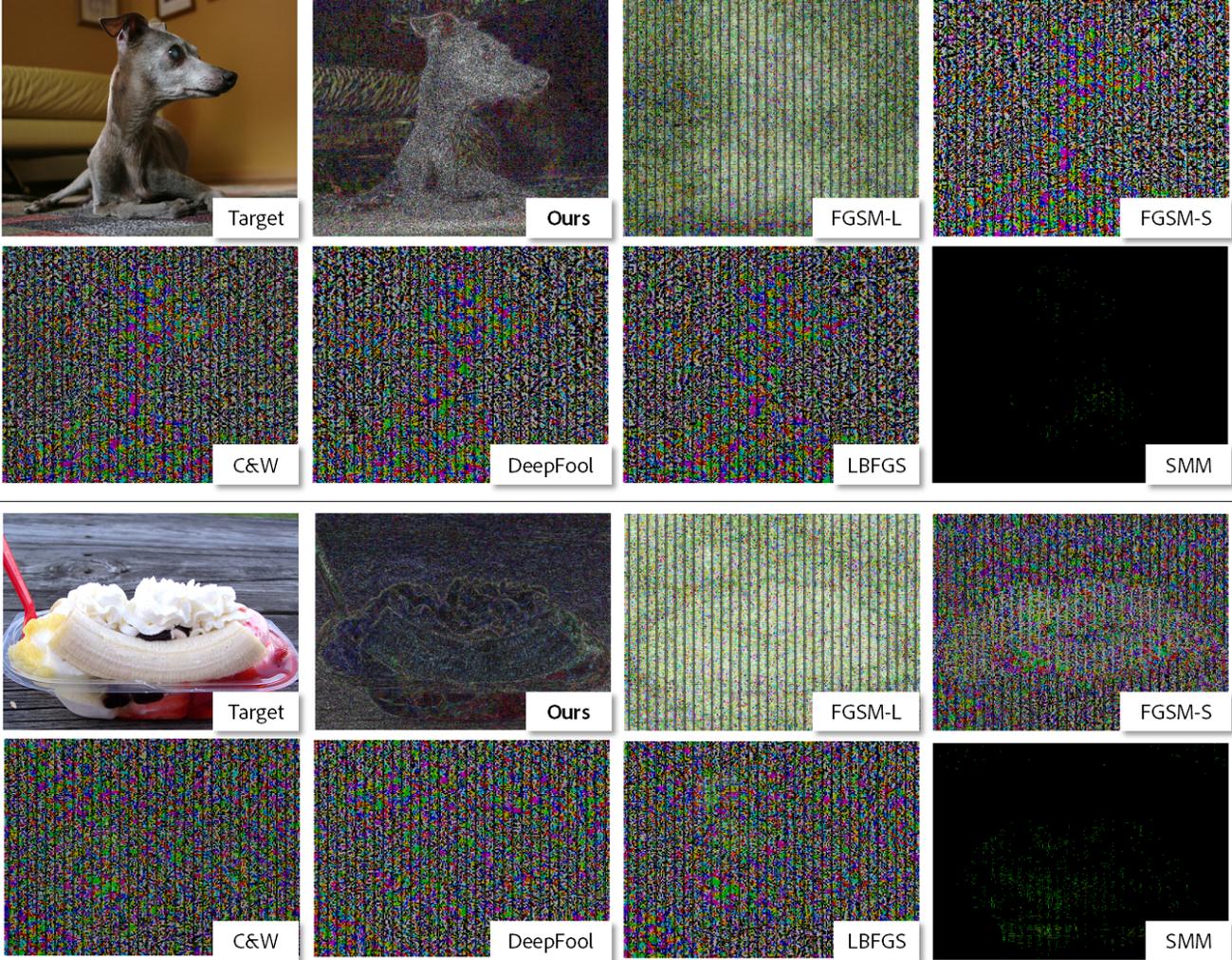
Figure 3. Representative perturbations (r) our proposed method vs. baselines running on two target images x; normalised and gamma corrected for visibility. In practice all the perturbations except FGSMIter-L are non- or barely perceptible.

the perceptibility of induced image artifacts via user study on Amazon Mechanical Turk (MTurk) using a randomly sampled 10% of the test set (hereafter, ImageNet-TS100) for practicality. We present the original image alongside the perturbed images from our technique and a baseline (in random order) and ask which is closer to the original. Each triplet is presented five times (each to independent MTurkers); in total 500 annotations are collected for the perceptual user study.

4.1.1 Baseline Methods

Our baselines are momentum iterative FGSM [5] with small (-S) and large (-L) parameter choices for ε. For the former we pick the smallest ε that the attack is effective (*i.e.* inducing minimal perturbation). For the latter we pick a constant large value (0.1). Four contemporary methods (below) are also evaluated. Open implementations from

Rauber *et al.*'s Foolbox [17] are used for all baselines.

L-BFGS. Szegedy *et al.*'s method [20] uses L-BFGS search to generate adversarial examples by finding minimum $(d > 0)$ such that the r minimising $\|r\|_2 + dJ_f(x,c)$ subject to $x + r \in [0,1]^m$ satisfies $f(x + r) = c$.

Carlini & Wagner (C&W) use a binary search to choose the smallest $d > 0$ for which the solution w to the problem

$$\text{minimise } \|h(w) - x\|_2^2 + dg(h(w)) \qquad (9)$$

satisfies $f(h(w)) = c$; where g is defined as

$$g(x') = \max\big\{ \max\{Z_f(x')_i : i \neq c\}$$
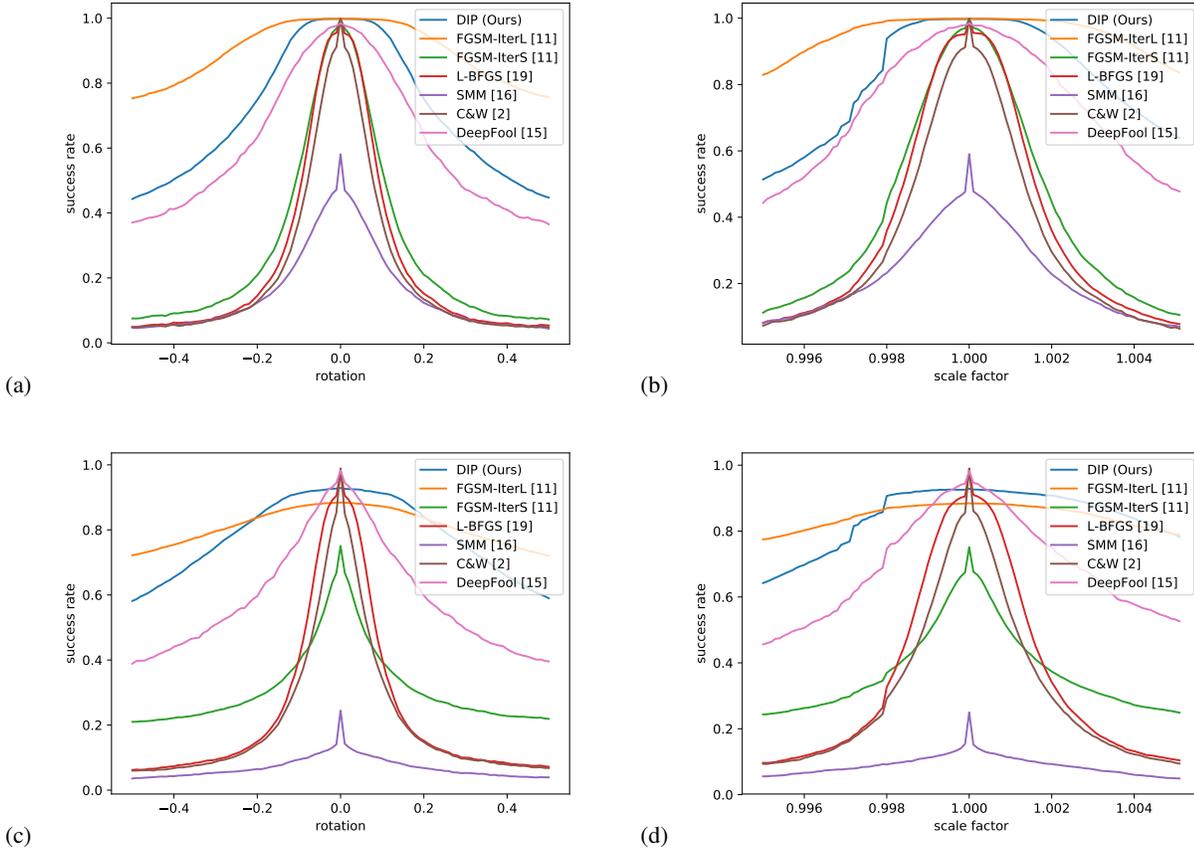$$- Z_f(x')_c, -\kappa\big\}, \quad (10)$$

Figure 4. Characterising the robustness of the proposed method for whole image attack to affine image transformation; rotation (a,c), scaling (b,d) over ImageNet trained VGG-19 (a,b) and GoogLeNetv3 (c,d) models. Our method degrades more gracefully than contemporary baselines.

$h(w) = \frac{1}{2}(\tanh(w)+1)$, and $\kappa \geq 0$ is a parameter that can guarantee a desired confidence. Then our adversarial example is $x + r = \frac{1}{2}(\tanh(w)+1)$.

Saliency Map Method (SMM) Papernot *et al.* differentiate the softmax output P_f (or in another variant, the logits Z_f) and apply a *saliency map* to this derivative, to target features to perturb. A typical example of a saliency map would be to choose two pixels (p_1, p_2):

$$(p_1, p_2) = \operatorname*{arg\,min}_{(p_1,p_2)} - \left(\sum_{i=p_1,p_2} \frac{\partial (P_f)_t}{\partial x_i} \right) \times \left(\sum_{i=p_1,p_2} \sum_{j \neq t} \frac{\partial (P_f)_j}{\partial x_i} \right) \quad (11)$$

where t is the target class and we only consider (p_1, p_2) for which the first term in (11) is positive and the second negative.

DeepFool [15] finds optimal adversarial examples for the

case of an affine classifier. To apply this to a general classifier we iteratively linearise it. Begin by setting $x^{(0)} = 0$. We continue the following procedure until we achieve a change of class, *i.e.* it is an *untargeted* attack. For every $i \in \{\}, t > 0$ we define

$$w_{i,t} = \nabla P_f \left(x^{(t)} \right)_i - \nabla P_f \left(x^{(t)} \right)_{f(x)} \quad (12)$$

$$v_{i,t} = P_f \left(x^{(t)} \right)_i - P_f \left(x^{(t)} \right)_{f(x)} \quad (13)$$

$$k = \operatorname*{arg\,min}_{i \neq f(x)} \frac{|v_{i,t}|}{\|w_{i,t}\|_2} \quad (14)$$

$$x^{(t+1)} = x^{(t)} + \frac{|v_{i,t}|}{\|w_{i,t}\|_2^2} w_{k,t} \quad (15)$$

and stop when $f\left(x^{(t)}\right) \neq f(x)$.

4.2. Evaluating Adversarial Images

We quantify success rate of our proposed DIP method against baselines in Tbl. 1. All methods run as targetted attacks; for each image in ImageNet-TS1K we pick a random
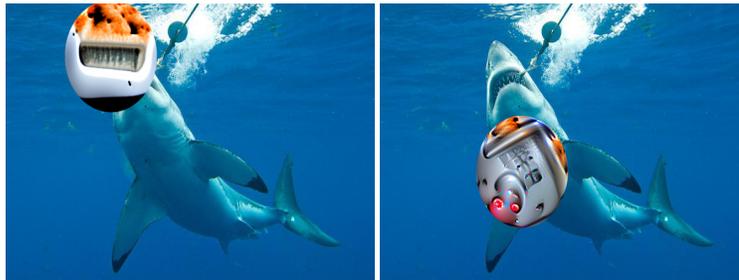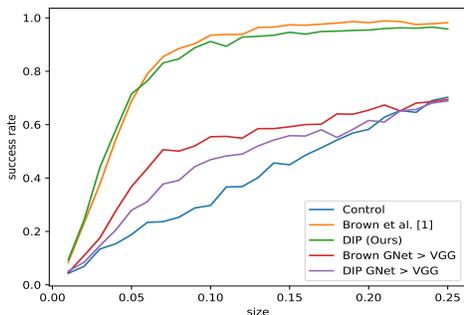
Figure 5. Adversarial patch attack: Ours (subsec. 3.1.1) vs. Brown *et al*. [1]. Left: Plotting misclassification rate vs. size of patch for VGG-19 and GoogLeNetv3; Control line (red) indicates success rate of super-imposing a photographic patch of target object. Successful patch attacks by us (middle) and Brown *et al*. [1] (right).

incorrect class to define $J_f(.)$ over softmax loss in CNN f and penalise deviation from a one-hot vector for that incorrect class under MSE loss. The exception is DeepFool which as proposed in [15] defined $J_f(.)$ as an untargetted attack, using MSE loss from a negated one-hot vector for the correct class. We test robustness to five image transformations (plus no transformation). Specifically we test recompressing the image as JPEG with 80% quality, or scale the image by up to $1 \pm .005$ (-L) and $1 \pm .001$ (-S), or rotate it by up to $\pm.5$ (-L) and $\pm.1$ (-S) degrees. For all transformations and networks our method is significantly more robust than other covert baseline, observing by more graceful decay in success rate (Fig. 4). The exception is FGSM-IterL which we include an example of an overt attack, forcing highly visible image perturbations by setting ε high. This represents an indicative level of overtness that FGSM needs to perform at to match the barely visible perturbations of our method. We provide visual examples of r (enhanced via normalization) in Fig. 3, and include results from our MTurk evaluation of attack visibility over ImageNet-TS100 (Tbl. 2). In all but two cases the perturbations via our method are less visible, or there is no visible preference, to baselines. The exceptions are DeepFool and C&W. Minor loss of high frequency detail inherent in DIP reconstruction may influences responses. Finally we compare run-time speed of our method to baselines in Tbl.3. Our approach takes a few minutes to run per image, relatively slow but comparable to state of the art optimization approaches [20, 2] running on a NVIDIA 1080Ti GPU. The non-deterministic nature of DIP yields slightly different adversarial examples each run, however we found no significant performance difference between runs.

4.3. Evaluating Adversarial Patches

We evaluate the adaptation of our method (subsec. 3.1.1) to create adversarial patches for an overt targeted attack, comparing to Brown *et al*.'s Adversarial Stickers [1] (Fig. 5). We pick a random 'attack' class and generate a sticker us-

ing 999 training images (one per ImageNet class, holding out the attack class) sampled randomly from the ImageNet training partition. We then apply that single sticker to all ImageNet-TS1K excluding the attack class, and consider the misclassification rate by applying to sticker to a random location. The process is averaged over 10 random attack classes. The sticker is scaled to a proportion of image area (see plot, Fig. 5). We find our method performs similarly to Brown *et al*. [1] outperforming the control case of pasting a large image of the attack class into the image. We train on VGG and GNet to enable comparison to [1] but also compare application of the GNet trained patches to attack a VGG network; our method performs similarly beyond 0.25.

5. Conclusion

We proposed a novel algorithm for the generation of covert adversarial image examples. Leveraging DIP to reconstruct the image from scratch under dual reconstruction and adversarial loss avoids the introduction of fragile high frequency artifacts. The resulting adversarial image exhibits greater robustness to affine image warping than the state of the art methods [20, 9, 2, 15, 16] whilst exhibiting low human visual perceptibility. We showed the same framework can be adapted to synthesise adversarial patches with similar performance to the state of the art sticker attack [1]. We demonstrated successful attacks against popular CNN visual classification networks (VGG-19, GoogLeNet Inception v3) using diverse categories from ImageNet. Future work could further characterise other networks and datasets, however we do not feel such enhancements necessary to demonstrate the promise of our adversarial image synthesis using image reconstruction under DIP.

References

[1] T. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[3] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. ShapeShifter: Robust physical adversarial attack on faster r-CNN object detector. In *Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer International Publishing, 2019.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[6] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Dawn Song, Tadayoshi Kohno, Amir Rahmati, Atul Prakash, and Florian Tramer. Note on attacking object detectors with adversarial stickers. *arXiv preprint arXiv:1712.08062*, 2017.

[7] R. Feng and B. Prabhakaran. Facilitating fashion camouflage art. In *Proc. ACM Multimedia*, pages 793–802, 2013.

[8] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *Intl. Conference on Computer Vision (ICCV)*, 2009.

[9] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572v3*, 2014.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[11] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[12] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 105–114, 2017.

[13] J. Lu, H. Sibai, E. Fabry, and D. Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*, 2017.

[14] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[16] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[17] J. Rauber, W. Brendel, and M. Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

[18] M. Sharif, S. Bhagavatula, L. Bauer, and M. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540, 2016.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[20] 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.

[21] S. Thys, W. Van Ranst, and T. Goedeme. Fooling automated surveillance cameras: adversarial patches to attack person detection. *arXiv preprint arXiv:1904.08653*, 2019.

[22] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.