Six Pixels to Kevin Bacon *
Robust Face Recognition in the Deep Learning Era

Leo Mehr
leomehr@stanford.edu

Luca Schroeder
lucschr@stanford.edu

Abstract

Modern facial recognition systems rely on deep neural networks which are known to be susceptible to adversarial attacks. We evaluate the robustness of a state-of-the-art facial recognition system, FaceNet, under four modern attacks and two defenses. We find several surprising results – for instance, that uninformed attackers are extremely ineffective against even the most basic defense – and conclude our work with several recommendations for practitioners. We publish all code for this project on Github. [2]

1 Introduction

Recent advances in facial recognition technology have revolutionized a plethora of applications, ranging from device authentication (e.g. Apple's Face ID) and e-commerce (e.g. Mastercard's 'selfie' payment technology) to public safety (e.g. police identifying criminals at large public events). And in the near future, facial recognition could be deployed in even higher-stake situations, such as targeting on military rifles [1] and management of pain medication for patients [2].

Thus, the risks of attacks on face identification technology and the consequences of misidentification grow ever larger—and so it is crucial to understand how brittle face identification currently is and how robust it can be made. These questions are particularly relevant in the wake of recent research on adversarial examples that has shown the vulnerability of DNNs to even small perturbations.

2 Related Work

Face recognition. Over the last ten years, error rates on face recognition systems have decreased by two orders of magnitude [3, 4] and surpassed human-level performance for the first time [5]. This improvement was fueled by: deep CNN architectures that eliminated the need for handcrafted feature extraction [6], rich labeled datasets of up to 100s of millions of images [7], and innovations in loss functions and training strategies. Perhaps the most common approach in modern face recognition systems, exemplified by the FaceNet model [7] we attack in this paper, is to use a ResNet [8] architecture in combination with a distance-based loss function, embedding face images in a Euclidean space, then reducing intra-variance and increasing inter-variance therein [6].

Adversarial attacks/defenses. Deep neural networks have been shown to be vulnerable to a variety of white-box [9, 10] and black-box attacks [11, 12] which can trigger misclassification even when modifying just a single pixel in a natural image [13]. A number of mitigating defenses have been proposed, from adversarial training [14] to attack detection [15] to defensive distillation [16], but recent research suggests most if not all of these in fact can be defeated [17]. Security through obscurity has also been ruled out as an option as adversarial examples have demonstrated strong transferability across different models and training sets [18, 19]. Our work transfers existing proposals for attacks and defenses to the context of face recognition, which has received comparatively little attention in this domain; this contribution is valuable not only because face recognition is a critical application in its own right but also because existing research has focused on datasets like MNIST which are known to have peculiarities that challenge generalization of results [17].

3 Dataset

We evaluate our attacks and defenses on the Labeled Faces in the Wild (LFW) dataset [20], widely recognized as the *de facto* face verification benchmark. The LFW dataset contains 13,233 images of 5,749 public figures, with 1,680 individuals

*https://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon

[2]https://github.com/leomehr/cs230project

CS230: Deep Learning, Winter 2019, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

represented by 2+ photos in the dataset. The LFW test set is a sequence of pairs of images which need to be classified as being photos of the same person or photos of different people. Figure 1 shows examples of both types of pairs. This face verification task is the basic building block of a face recognition system, which may compare a new face image with a number of images in its database to authenticate or identify users. LFW images are aligned with a Multi-task Cascaded Convolutional Network (MTCNN) [21] and scaled to 160×160. 0-1 RGB scale is used.

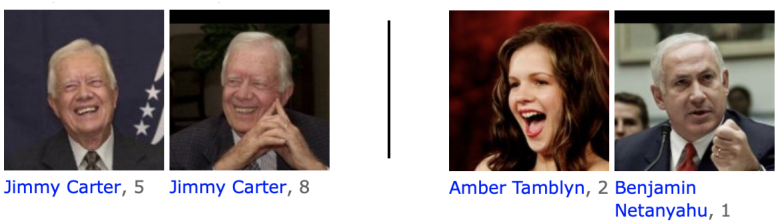


Figure 1: Sample LFW test pairs, to be classified as "same" (left) and "not same" (right)

4 Methods

The model we target with our attacks is a TensorFlow implementation of FaceNet [7], which gives a mapping from face images to a 128-D embedding in a Euclidean space. This model uses the Inception-ResNet-v1 architecture [22] and is pre-trained on the VGGFace2 dataset [23], which contains more than 3 million face images of 9,000+ individuals. The model achieves 99.65% accuracy on LFW.

To stage our attacks we use the Python package Foolbox [24], which contains implementations for many adversarial attack techniques. As Foolbox expects a classifier which takes one input image we wrap our FaceNet model as shown in Figure 2. For each LFW test pair (f_1, f_2), we fix the second face image f_2 and compute its embedding x_2. The adversary then feeds perturbed versions of f_1 into our model and our "classifier" outputs probabilities that the pairs of images are of the "same" class or of the "different" class. If the images were photos of the same person to begin with, the adversary's goal is to find perturbation Δ such that $d(\texttt{embedding}(f_1 + \Delta), x_2) > \texttt{threshold}$, i.e. such that FaceNet thinks the two faces are of different people. This set-up follows [25].

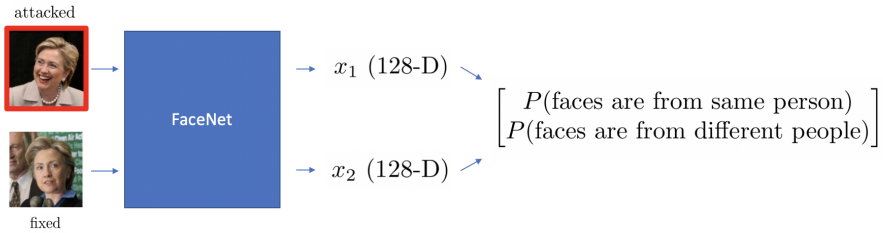


Figure 2: Turning FaceNet into a classifier for Foolbox

In particular following [25] we set $P(\text{same}) = 1 - 0.5 \cdot \frac{d}{\texttt{threshold}}$ if the distance d in the embedding space between two face images is $< \texttt{threshold}$ and set it to $0.5 - \frac{0.5 * (d - \texttt{threshold})}{\texttt{max\_distance} - \texttt{threshold}}$ otherwise. This is an arbitrarily chosen mapping from distances to probabilities and it would be possible to instead learn this by appending and training further layers to FaceNet.

We focus on four popular adversarial attack techniques:

- Additive Uniform Noise Attack (Uniform): i.i.d. Uniform noise is added to the image; the standard deviation of the noise is increased until misclassification is achieved;
- Additive Gaussian Noise Attack (Gaussian): same as Uniform but with i.i.d. Gaussian noise;
- Fast Gradient Sign Method (FGSM) [9]: find smallest ϵ such that $f_1 + \Delta(\epsilon)$ is misclassified, where $\Delta = \epsilon \cdot \text{sign}(\nabla_f \mathcal{L}(f_1, \ell_0))$ and ℓ_0 is the true classification label. i.e. FGSM perturbs the image in the gradient direction, increasing loss and triggering misclassification;
- Deep Fool [10]: iteratively perturb f_1 in the direction of the gradient of the loss function, generating a sequence of perturbations $\epsilon_0, \epsilon_1, \ldots$ that terminates once the decision boundary is crossed, and yields the final perturbation $\epsilon = \sum_i \epsilon_i$. Deep Fool is roughly an iterative extension of FGSM and while it is more computationally expensive, it produces more effective adversarial examples and with much smaller perturbations.

	Succ%, $\|\|\Delta\|\| < \infty$		Succ%, $\|\|\Delta\|\| < 5$		Succ%, $\|\|\Delta\|\| < 1$		Average $\|\|\Delta\|\|$	
	Anon.	Impers.	Anon.	Impers.	Anon.	Impers.	Anon.	Impers.
FGSM	**100%**	79.17%	**100%**	75%	50%	16.67%	1.23	2.41
DeepFool	**100%**	**100%**	**100%**	**100%**	**92.31%**	**70.83 %**	**0.50**	**0.90**
Uniform	**100%**	0%	7.69%	0%	3.84%	0%	33.29	-
Gaussian	**100%**	4.17%	7.69%	0%	3.84%	0%	32.79	12.91

Table 1: Success rates of attacks with constrained perturbation size $||\Delta||$, and average perturbation size $||\Delta||$ of successful attacks. Performance segmented into anonymization/impersonation tasks.

In addition to evaluating these attacks on an undefended face recognition system, we also analyze their success against two defense mechanisms:

- Low-pass filter: input images to the system are first blurred with a 3x3 mean filter, which averages each pixel value with itself and the immediate surrounding pixels.
- Deep image restoration: input images are first passed through a Noise2Clean [26] network pretrained for Gaussian noise removal on 50K perturbed ImageNet examples; Noise2Clean is a convolutional auto-encoder network which uses the Red30 [27] architecture with symmetric skip connections.

(a) (b) (c) (d)

Figure 3: (a) attacked with Gaussian noise (b), corrected with a mean filter (c) and Noise2Clean (d).

Figure 3 shows both defenses applied to an attacked image for a visual comparison of their corrective power. Following [28, 17] we analyze the interaction of these attacks and defenses under two threat models:

- Zero Knowledge: the attacker has no knowledge that a defense mechanism is being used. Adversarial examples are constructed against the base facial recognition system and then tested against the protected system.
- Perfect Knowledge: the attacker has complete knowledge of the complete network architecture/parameters, including those of any defense layers. Adversarial examples are constructed & run against the protected system.

5 Results

Since adversarial attacks can be time-consuming to run it is impractical to generate an entire adversarial dataset from LFW test pairs. Instead, we randomly chose 50 LFW test pairs and ran each attack on these. Similar to [25] we report the performance of the attacks on two conceptually distinct tasks:

1. *Anonymization*. If (f_1, f_2) are photos of the same person, the attacker tries to perturb f_1 such that $f_1' = f_1 + \Delta$ and f_2 are regarded as being faces of different people. An example of a real-world anonymization attack would be a person of interest trying to mask their presence from law enforcement face recognition software.
2. *Impersonation*. If (f_1, f_2) are photos of different people, the attacker tries to perturb f_1 such that $f_1' = f_1 + \Delta$ and f_2 are regarded as being faces of the same person. An example of a real-world impersonation attack would be an individual trying to gain access to someone else's device which uses face authentication.

Table 1 gives the success rate (percentage of test pairs for which a misclassification was achieved) for the 4 different attacks against the undefended FaceNet system. It also shows how the success rate changes as the max perturbation size is constrained, and the average perturbation size (L2-norm of Δ) for successful misclassification attacks with no such size constraint. Table 2 gives the success rate for different attacks in the presence of the simple mean filter defense, in both zero- and perfect-knowledge settings. Zero knowledge results for Noise2Clean defense are similar and are omitted for space, although interestingly, Noise2Clean works better for defending against the black-box Additive Gaussian and Uniform attacks which have substantially more noise but is not as competitive at eliminating the white-box FGSM and DeepFool attacks. Unresolvable compatibility issues prevented perfect knowledge experimentation for Noise2Clean.

To additionally investigate the transferability and generalizability of the attacks produced on our FaceNet model we fed the attack images to Amazon Rekognition, a commercial state-of-the-art deep learning system. Given a pair (f_1, f_2) of images

	Zero Knowledge Attacks				Perfect Knowledge Attacks			
	Succ%, $\|\|\Delta\|\| < \infty$		Average $\|\|\Delta\|\|_{\text{def}}$		Succ%, $\|\|\Delta\|\| < \infty$		Average $\|\|\Delta\|\|$	
	Anon.	Impers.	Anon.	Impers.	Anon.	Impers.	Anon.	Impers.
FGSM	7.69%	0%	2.20	2.38	**100%**	79.17%	1.40	2.48
DeepFool	7.69%	0%	**2.05**	**1.88**	**100%**	**100%**	**0.58**	**1.05**
Uniform	7.69%	0%	11.50	-	**100%**	0%	41.45	-
Gaussian	**11.53%**	0%	11.34	4.38	**100%**	4.17%	41.58	52.52

Table 2: Evaluation of mean filter defense against zero- and perfect-knowledge attacks with unconstrained perturbation size $||\Delta||$. $||\Delta||_{\text{def}}$ refers to the norm of the difference between the original image and the attacked image after applying the defense correction to the latter.

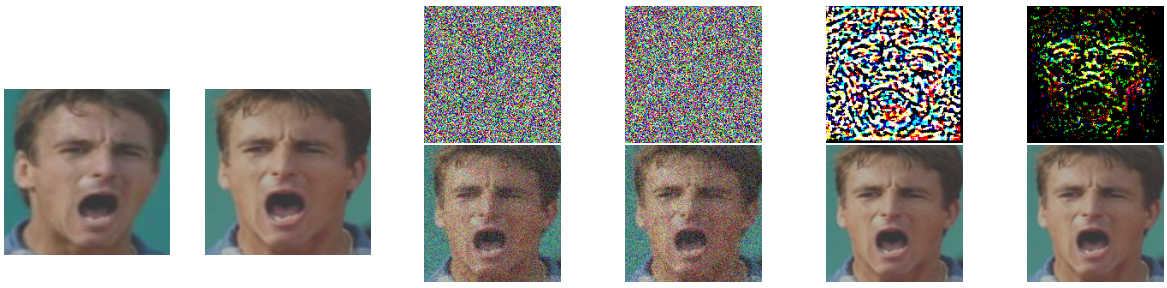
	No Preprocessing		Mean Filter		Noise2Clean	
	Anon.	Impers.	Anon.	Impers.	Anon.	Impers.
No Attack	98.60	87.74	98.40	87.71	98.62	87.59
FGSM	97.46	**78.51**	97.06	79.74	97.38	78.86
DeepFool	96.68	78.65	96.65	**79.11**	96.71	**78.77**
Uniform	**58.12**	-	92.16	-	94.64	-
Gaussian	61.97	90.75	**92.14**	92.25	**94.57**	91.88

Table 3: Transferability of attacks/defenses to Amazon Rekognition. The numbers are a proxy for Rekognition's confidence in the correct classification (lower scores $\implies$ more successful attacks).

this web service outputs a similarity score s between faces in the two images. If f_1 and f_2 are the same person we use s as a measurement for Rekognition's confidence in the correct prediction ('same'); if they are of different people we use $1 - s$ instead. Table 3 gives the average confidence score of Rekognition on the unperturbed test set, the attacked test set, and the attacked test set corrected with the mean filter and Noise2Clean defenses before being fed to Rekognition.

6 Discussion

Impersonation is harder than anonymization. As evidenced in Table 1, anonymization attacks are always able to succeed for unconstrained perturbation sizes; by contrast impersonation has lower success rates and requires higher perturbations (~2x greater $||\Delta||$ for FGSM/DeepFool). This disparity makes sense: for anonymization, the attacker simply has to make the system classify an image as *anyone* other than its original identity, whereas for impersonation the attacker has to convince the system that an image is of a *particular* person. The specificity of the impersonation target accounts for the difficulty.



(a) f_2 (b) original, f_1 (c) Uniform (d) Gaussian (e) FGSM (f) DeepFool

Figure 4: (a) is f_2, the fixed image and (b) is f_1, the image to attack in a LFW test pair; the rest are the perturbed images f_1' produced by the four attacks, all of which successfully cause the classifier into outputting 'not same'. The top row images show the delta between the perturbed and original images (scaled to be visible).

White-box attacks are more powerful than black-box. As expected, the gradient-based white-box attacks are much more effective than their additive noise black-box counterparts, with order-of-magnitude smaller perturbation sizes required for a misclassification on anonymization tasks. This can be seen in Figure 4; changes resulting from successful white-box attacks are imperceptible. For impersonation tasks black-box attacks are essentially useless; intuitively it is plausible to add enough uniform noise that a face no longer belongs to its original identity but it is unrealistic to expect to be able to add enough uniform random noise to suddenly match a precise target identity.

Even simple defenses eliminate attacks in zero knowledge settings. As Table 2 demonstrates, even a mean filter cripples attacks if the adversary is unaware it is being applied; impersonation attacks have 0% success across the board and even

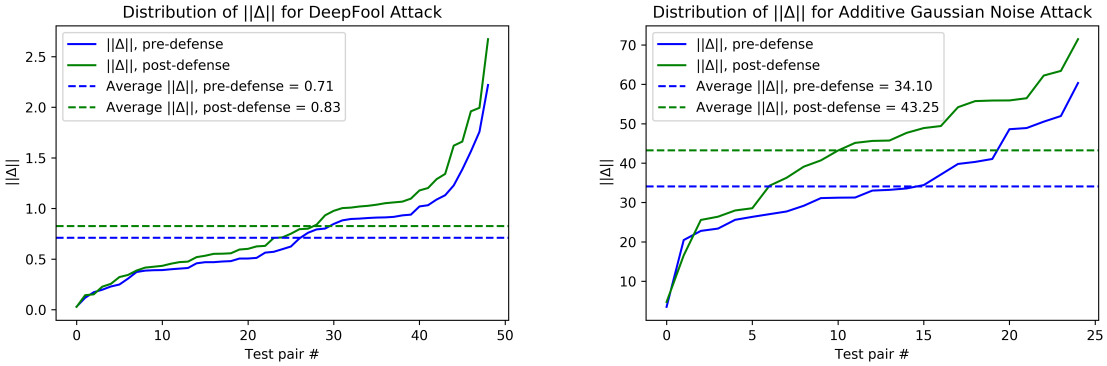


Figure 5: Distribution of the L2-norm of the attack perturbation Δ, before and after applying the mean filter defense. Attacker has full knowledge.

anonymization attacks are essentially useless. Comparing $||\Delta||_{\text{def}}$ in Table 2 to $||\Delta||$ in Table 1, we see that the mean filter greatly reduces the distance between the perturbed and original images for black-box attacks.

Perfect knowledge of defenses renders them ineffective. As Table 2 shows, when the attacker can generate adversarial examples against the defended network, the success rates are almost the same as against the undefended network in Table 1. However, as Figure 5 illustrates, attacks against the defended network are understandably larger (17% higher $||\Delta||$ for white-box and 27% higher $||\Delta||$ for black-box attacks) and therefore also take iterative attacks longer to generate.

Adversarial examples are moderately transferable. As Table 3 demonstrates, *all* attacks lower the confidence of correct classification in Amazon Rekognition. White-box impersonation attacks and black-box anonymization attacks seem to transfer well while white-box anonymization attacks do not, although it is unknown if this is due to any hidden preprocessing by Rekognition. In general, we expect white-box attacks to transfer less well because the perturbation sizes are small and tuned to gradients resulting from our specific network architecture and parameters. Applying defenses significantly improves performance against black-box noise attacks, suggesting universal defenses can be developed against such attacks; it is interesting to note that neither defense mitigates white-box impersonation attacks in contrast to our results in Table 2.

7 Conclusion & Future Work

While some of the results we found were expected, such as white-box attacks being more effective than black-box attacks, and impersonation being more difficult than anonymization, other results were truly surprising to us and illustrate some of the peculiar properties of deep neural network facial recognition systems.

Because even basic defenses eliminate zero-knowledge attacks and raise the computational cost of attacks in a perfect knowledge setting (with no degradation in accuracy on images from honest users), we recommend that facial recognition systems use at least basic defenses such as a low-pass filter. This means, however, that a rational attacker should assume that defenses are in play and attacks will therefore never be zero knowledge. More work is needed, then, to analyze attacks in a partial knowledge setting, where attackers are aware that defenses exist but do not know what those defenses are; in particular the question is whether attacks against one defense can transfer successfully to a model using another defense.

While perfect knowledge attacks are extremely *effective* against the low-pass filter, one major area for future work is to evaluate whether more sophisticated defenses such as Noise2Clean are effective against perfect knowledge attackers. We hypothesize that at least white-box attacks can still break through such defenses. Another direction is to investigate an ensemble of defenses from which one or more is selected randomly at runtime; this may provide probabilistic protection.

For a defender, the success of perfect knowledge attacks clearly suggests a need to reduce the attacker's knowledge as much as possible, e.g. by carefully securing the network architecture & weights or the defense implementations. We therefore suggest defenders avoid using popular pre-trained models simply out-of-the-box. Although our work demonstrates there is indeed some transferability of adversarial examples in facial recognition, transferred examples (particularly white-box attacks) still have significantly lower success rates on the new facial recognition system.

There are many further ideas for future work, some of which we highlight here. Can adversarial attacks be prevented while maintaining verification accuracy by making the embedding distance threshold for a 'same' classification stricter? Rather than correcting adversarial images, is it easier and more reliable to instead detect attacks, e.g. by a SafetyNet? Finally, it may be interesting to study a "group impersonation" task that falls between the "anonymization" and "impersonation" tasks studied in this paper; here, the attacker tries fooling the network into classifying the input as any one of N identities. Which defenses work best for different group impersonation attacks?

8 Contributions

Luca did the initial work setting up the FaceNet model and integrating with Foolbox, as well as the zero knowledge attack and transferability experiments. Leo did the initial work setting up experiments on AWS, as well as the perfect knowledge attack experiments and integration with Noise2Clean. Both authors contributed to the writing of the report and presentation.

The authors thank Hao Sheng for mentoring this project and Amazon Web Services for sponsoring CS230 by providing GPU credits.

References

[1] Matthew Cox. Army's next infantry weapon could have facial-recognition technology. *Military.com*, 2019. [Online; accessed 10-October-2019].

[2] Clarice Smith. Facial recognition enters into healthcare. *Journal of AHIMA*, 2018. [Online; accessed 10-October-2019].

[3] Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*, pages 189–248. Springer, 2016.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[5] Chaochao Lu and Xiaoou Tang. Surpassing human-level face verification performance on lfw with gaussianface. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[6] Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.

[7] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CVPR*, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

[10] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal 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.

[11] Chuan Guo, Jacob R Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q Weinberger. Simple black-box adversarial attacks. *arXiv preprint arXiv:1905.07121*, 2019.

[12] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017.

[13] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019.

[14] 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.

[15] Jiajun Lu, Theerasit Issaranon, and David A. Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[16] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *IEEE Symposium on Security and Privacy*, 2016.

[17] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017.

[18] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[20] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[21] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016.

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[23] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[24] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox v0.8.0: A python toolbox to benchmark the robustness of machine learning models. *CoRR*, 2017.

[25] Bruno López Garcia. Crafting adversarial faces. *brunolopezgarcia.github.io*, 2018. [Online; accessed 7-November-2019].

[26] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.

[27] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in neural information processing systems*, pages 2802–2810, 2016.

[28] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.