Face-Off: Improving Imperceptibility of Adversarial Examples

Brian Tang
University of Wisconsin - Madison

1 INTRODUCTION

As face recognition becomes more prevalent in contexts such as social media, photo storage, and law enforcement, it becomes increasingly important to consider the privacy of users' data. Automated face recognition systems can exploit uploaded photos to associate users with locations and activities. Recent work allows privacy-conscious users to obfuscate their faces from face recognition without any loss of usability. These approaches rely on adversarial attacks[3] or data poisoning attacks[11] to preserve user privacy while retaining full usability of the social media platform. Other approaches completely blur faces or obfuscate faces from detection systems, but doing so neglects the privacy vs. utility trade-off inherent on social media and online photo platforms. The system proposed in Face-Off[3] uses adversarial attacks to induce errors in the face recognition model and result in misclassifications without affecting face detection performance. However, as these privacy systems operate in a black-box threat model with no access to proprietary model parameters, Face-Off relies on the transferability property of adversarial examples to ensure usability. Amplification of the perturbation mask is one such approach to enhancing this property, however, it negatively impacts usability (see Figure 1). Image scaling attacks, a recent discovery which exploits the scaling aspect of preprocessing in deep neural network (DNN) pipelines, camouflage downsized images within high resolution images[13]. This effectively produces a completely different result after a downsampling algorithm is applied. By combining this step with Face-Off in the postprocessing of adversarial example generation, we can greatly decrease the perceptibility of the adversarial perturbations while increasing the strength of the attack[1]. We produce the following contributions from this work:

(1) The addition of image scaling attacks to the pipeline of Face-Off.
(2) Preliminary results of the effectiveness of adversarial image scaling attacks.
(3) Model parameters from the AWS Rekognition and Azure Face services.

[1]Concurrent research is being done to evaluate the effectiveness of camouflaging powerful adversarial perturbations using image scaling

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Figure 1: An example amplified output from Face-Off.

2 BACKGROUND

2.1 Adversarial Machine Learning

In the traditional deep learning classification setting, adversarial examples are images with minor imperceptible perturbations which result in an incorrect classification output from a model[4, 12]. Adversarial examples are defined with the following threat model: Given white-box access to a model's parameters, weights, and architecture, and where $f(X) = Y$, find an X' close to X such that $f(X') \neq Y$. For the context of Face-Off, we assume the user (adversary) has black-box access to the model in which they cannot query the model. Querying the face recognition model would defeat the purpose of Face-Off and leak user data. Thus, Face-Off must rely on the transferability property of adversarial examples, where adversarial examples generated for one deep learning also result in misclassifications in other models[7, 8]. Face-Off uses adversarial attack algorithms such as Projected Gradient Descent (PGD)[6] and Carlini-Wagner (CW)[2] to apply an imperceptible layer of strategic noise to the original image. Adversarial examples can also be amplified[1, 5] to increase the likelihood the attack will transfer to other models as well as decrease matching confidence.

2.2 Face Recognition and Face-Off

Typically, face recognition determines matches between faces by detecting a face within a photo and matching it with another face or centroid of face embeddings. A distance metric (l_2 norm or cosine similarity) is used in tandem with a threshold to calculate the *closeness* of faces. Face-Off applies a layer of strategic adversarial perturbations onto an uploaded face which allows a user to mask their data from proprietary datasets and malicious third-parties[3]. Through these mismatches, the user's face is no longer able to be automatically recognized, thus giving privacy conscious users a practical option for online face obfuscation. One of the main drawbacks of Face-Off is the privacy vs. utility trade-off inherent from amplifying adversarial perturbations. The more a noise mask is amplified, the worse the image quality appears.

2.3 Image Scaling Attacks

Image scaling attacks take advantage of scaling algorithms by injecting a camouflaged image within a larger resolution image so that downscaling the image using an algorithm, such as linear interpolation, results in a completely different image. An attack image $S' \sim S$ is created by a source image S and target image T such that the downscaled image $D \simeq T$. See Figure 2 for a visual representation. This attack can be used to induce failures within deep learning models via data poisoning attacks[10] and can likewise be used to hide adversarial perturbations. Image scaling attacks can be detected and defended against[9], and can produce perceptible artifacts in lower resolution images, especially if the target image is very different from the source image. Since adversarial examples are already quite similar to the original source image, combining the two approaches would greatly reduce perceptibility of any image tampering. An image scaling attack with source image X and target image X' would produce an attack image $X'' \simeq X$ and an downscaled image $D \simeq X'$.

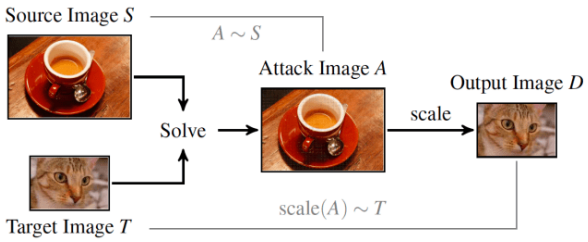

Figure 2: The image scaling attack pipeline.

3 METHODOLOGY

Given the scope of this project, we identify the following as the key features to focus our efforts on:

(1) The integration of image scaling attacks into the offline Face-Off evaluation pipeline and the website backend functionality.
(2) An analysis of online APIs and the model parameters required to be known for the image scaling attack algorithm.

The code for the image scaling attacks has been reproduced and combined with the existing Face-Off functionality. In addition to this, the website has been updated with full adversarial scaling capabilities. The updated pipeline of Face-Off is as follows:

(1) Detect and crop face via MTCNN
(2) Generate adversarial example
(source = person 1, target = person 2)
(3) Apply image scaling attack
(source = cropped, target = downscaled adversarial example)
(4) Stitch double-attacked face back onto original image

Image scaling attacks rely on knowledge of model input size, image scaling algorithm, and, for our use case, face detection scheme. In order to query this information, we develop a dataset of scaled faces of **Meryl Streep** with **Matt Damon** hidden within. Figure 3 shows an example of one such dataset. Each attacked face is compared with a regular image of Matt Damon, and APIs returning confidence scores of $\geq 50\%$ are deemed matches. Matches indicate a successful set of transferable parameters. Faces were generated with a combination of these attributes as detailed in Table 1.

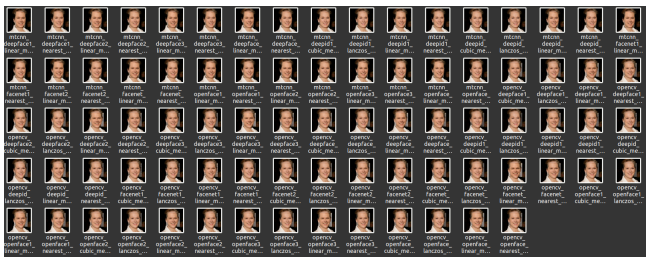

Figure 3: Dataset of Meryl Streep images used to query APIs for model knowledge.

Input	Input	Input	Detector	Scaling
55x47	144x144	158x158	OpenCV	Nearest
64x64	152x152	160x160	MTCNN	Linear
96x96	154x154	162x162	-	Cubic
100x100	156x156	164x164	-	Lanczos
120x120	-	-	-	-

Table 1: Combinations of parameters for online API queries

Figure 4: Left: Unperturbed image; **Middle**: Image scaling attack (output is a 112x96 Matt Damon); **Right**: Adversarially perturbed image

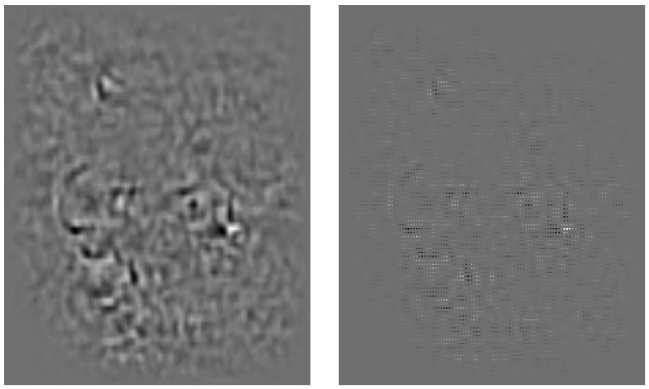

Figure 5: Left: Adversarial perturbation mask; Right: Perturbation mask + Image scaling attack

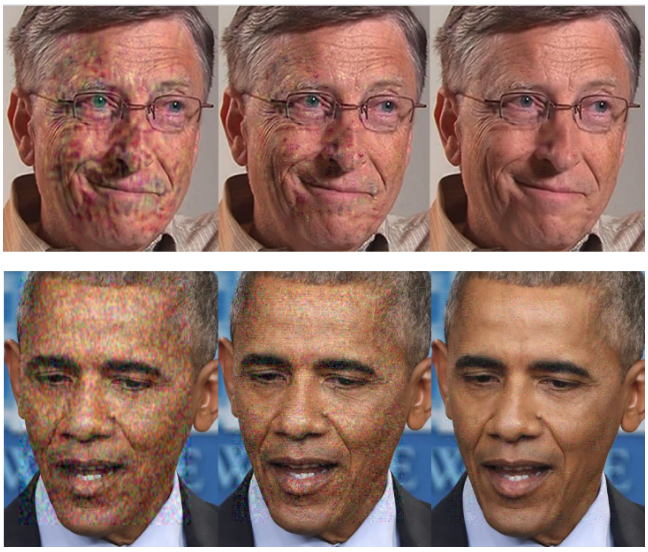

Figure 6: Top: Carlini-Wagner L2; **Bottom**: Projected Gradient Descent L2; **Left**: Adversarial example; **Middle**: Robust linear scaled adversarial example; **Right**: Weak linear scaled adversarial example;

Figure 7: Left: Adversarial example (Carlini-Wagner L2); **Right**: Image scaling attack and adversarial example;

API	Scaling	Input	Detector
Azure	Linear	100x100 to 164x164	MTCNN
AWS	Linear	120x120 to 164x164	MTCNN or OpenCV

Table 2: Online API configurations

4 EVALUATION

5 DISCUSSION

5.1 Challenges

After additional literature review and initial experimentation, some challenges have arisen with the direction of the project. Image scaling attacks rely on the interpolation algorithm being performed on the exact same image, since the perturbations are required to be embedded within the image at exact pixel positions, so that the scaling algorithm is exploited. Several implementation issues specific to the face recognition domain develop as a result.

(1) Face detection strategies (MTCNN, dlib, SSD, OpenCV) may not universally detect and segment images the same, so replacing a face with an image scaling attack may not necessarily result in the desired output since pixel positions could be slightly offset.
(2) Generating image scaling attacks is computationally expensive which could decrease the usability aspect of Face-Off.

These issues came up when the initial query dataset was run on face recognition APIs. The extensive online evaluation failed to transfer to any APIs. Shifting the approach to a robust image scaling algorithm which perturbs a percentage of neighboring pixels was a viable solution. Unfortunately during these experiments, I realized the costs incurred by uploading images to these APIs would accumulate quickly. In order to get any meaningful results, I will likely need more funding and time to generate and evaluate a much larger dataset of images.

5.2 Future Directions

(1) **December 20**: Perform online evaluation to ensure faces are successfully obfuscated.
(2) **December 25**: Generate adversarial examples using other face recogntion architectures.
(3) **January 10**: Evaluate perceptibility score (LPIPS metric) of adversarial examples vs. image scaling + adversarial examples on several face. recognition datasets (Labeled Faces in the Wild, Famous Celebrities).
(4) **January 25**: Reduce computational overhead and implement runtime enhancements.
(5) **February 15**: Create a smarter robust scaling attack algorithm specific to this application.
(6) **March 15**: Submit work to a workshop.

REFERENCES

[1] Xiaoyu Cao and Neil Zhenqiang Gong. Mitigating evasion attacks to deep neural networks via region-based classification. In *Proceedings of the 33rd Annual Computer Security Applications Conference. ACM* (2017).
[2] Nicholas Carlini and David Wagner. 2016. Towards Evaluating the Robustness of Neural Networks. *arXiv:1608.04644 [cs.CR]* (2016).
[3] Varun Chandrasekaran, Chuhan Gao, Brian Tang, Kassem Fawaz, and Somesh Jha. Face-Off: Adversarial Face Obfuscation. In *The 21st Privacy Enhancing Technologies Symposium* (2021).
[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *arXiv:1412.6572 [stat.ML]* (2014).
[5] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv:1611.02770* (2016).
[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083 [stat.ML]* (2017).
[7] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv:1605.07277* (2016).
[8] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2016. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv:1602.02697* (2016).
[9] Erwin Quiring, David Klein, Daniel Arp, Martin Johns, and Konrad Rieck. Adversarial Preprocessing: Understanding and Preventing Image-Scaling Attacks in Machine Learning. In *Proceedings of the 29th USENIX Security Symposium* (2020).
[10] Erwin Quiring and Konrad Rieck. 2020. Backdooring and Poisoning Neural Networks with Image-Scaling Attacks. *arXiv:2003.08633* (2020).
[11] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. In *Proceedings of the 29th USENIX Security Symposium* (2020).
[12] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intruiging properties of neural networks. *arXiv:1312.6199 [cs.CV]* (2014).

[13] Qixue Xiao, Yufei Chen, Chao Shen, Yu Chen, and Kang Li. Seeing is Not Believing: Camouflage Attacks on Image Scaling Algorithms. In *Proceedings of the 28th USENIX Security Symposium* (2019).