EFFICIENT DISRUPTIONS OF BLACK-BOX IMAGE TRANSLATION DEEPFAKE GENERATION SYSTEMS

Nataniel Ruiz, Sarah Adel Bargal, Stan Sclaroff
Department of Computer Science, Boston University
{nruiz9,sbargal,sclaroff}@bu.edu

ABSTRACT

In this work, we develop efficient disruptions of black-box image translation deepfake generation systems. We are the first to demonstrate black-box deepfake generation disruption by presenting image translation formulations of attacks initially proposed for classification models. Nevertheless, a naive adaptation of classification black-box attacks results in a prohibitive number of queries for image translation systems in real-world applications. We present *Leaking Transferable Perturbations (LTP)*, an algorithm that significantly reduces the number of queries needed to attack an image. LTP consists of two phases: (1) a short *leaking phase* where we attack the network using traditional black-box attacks and gather information on successful attacks on a small dataset and (2) an *exploitation phase* where we leverage said information to subsequently attack the network with improved efficiency. Our attack reduces the total number of queries necessary to attack GANimation and StarGAN by more than half.

1 INTRODUCTION

The term “deepfake” has recently been adopted in a broader context and can be used to refer to any altered media of someone’s likeness. Recently there have been remarkable advances in face modification algorithms and controllable face synthesis Wiles et al. (2018); Ranjan et al. (2018); Geng et al. (2019); Nguyen-Phuoc et al. (2019); Ghosh et al. (2020). Some algorithms only need a single image and can create modified versions of that person under different poses, expressions, lighting and other attribute changes Choi et al. (2018); Pumarola et al. (2018); Choi et al. (2019). The most advanced algorithms can create puppeteering videos using as few as one image Zakharov et al. (2019); Tewari et al. (2020). This few-shot deepfake technology based on image translation networks has gained popularity in the mainstream with apps such as FaceApp fac that allow for transformation of images such as putting a smile on someone’s face and making them appear older or younger, among other interventions. These technologies can be used in malicious ways to produce undesirable content of someone without their consent.

Instead of detecting deepfakes after the fact, Ruiz et al. (2020) recently proposed using *white-box adversarial attacks* to protect an image from modification by image translation networks. While this work assumes that the adversary has access to the model’s structure, weights and gradients, in a real scenario, these might not be accessible. In this work, we focus on the *black-box scenario* where model parameters are unknown and show the vulnerability of several popular image translation networks. Specifically, we are the first to explore black-box adversarial attacks on image translation systems with an application of disrupting deepfake generation (Figure 1).

We present a simple, yet highly effective, algorithm that we call *Leaking Transferable Perturbations (LTP)* that sharply decreases the average number of queries required to generate attacks in this scenario. LTP is composed of two phases, a short *leaking phase* during which the network is attacked using a traditional black-box attack on a small dataset of images, and an *exploitation phase*, where the algorithm leverages the information obtained during the *leaking phase* to subsequently attack the network with improved efficiency.

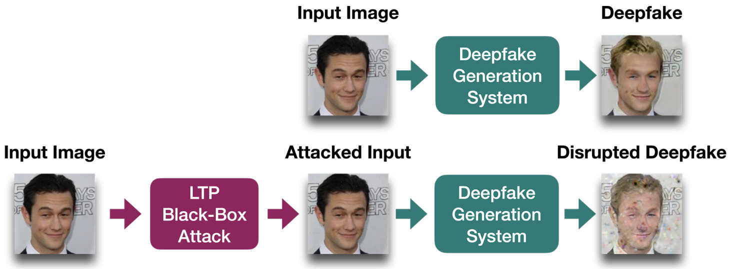

Figure 1: Illustration of our proposed black-box attack LTP disrupting a deepfake generation system. After applying an imperceptible filter on the input image (bottom), output of the deepfake generation system is successfully disrupted preventing malicious change of hair color.

2 Related Work

There is a large amount of work on white-box attacks against discriminative models Szegedy et al. (2014); Goodfellow et al. (2015); Moosavi-Dezfooli et al. (2016); Papernot et al. (2016); Carlini & Wagner (2017); Nguyen et al. (2015); Moosavi-Dezfooli et al. (2017); Kurakin et al. (2017); Madry et al. (2018) as well as black-box attacks against discriminative models Liu et al. (2016); Narodytska & Kasiviswanathan (2016); Chen et al. (2017); Papernot et al. (2017); Ilyas et al. (2018; 2019); Bhagoji et al. (2018); Cheng et al. (2019); Tu et al. (2019); Guo et al. (2019); Andriushchenko et al. (2020). There exists a limited amount of work demonstrating adversarial attacks on generative models Tabacof et al. (2016); Kos et al. (2018); Ruiz et al. (2020) and our work is the first to tackle adversarial attacks on image translation networks.

3 Method

In the black-box adversarial attack setting, we are given a budget of black-box queries for each image we would like to attack. In this setting, we have the same number of maximum allowed queries for all images in the dataset. That is, for each image $\boldsymbol{x}$ and target $\boldsymbol{r}$ we want to solve the optimization problem

$$\min_{\boldsymbol{\eta}} L(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}), \boldsymbol{r}), \quad \text{subject to } p(\boldsymbol{\eta}) \leq \boldsymbol{\epsilon}, \mathbf{Q} \leq \boldsymbol{B}, \tag{1}$$

where $\boldsymbol{G}$ is the generator of the image translation system, $\boldsymbol{\eta}$ is the perturbation, $\mathbf{Q}$ is the number of queries used, $\boldsymbol{B}$ is the maximum number of queries allowed for a single image and L is an image-level regression loss.

Our proposed algorithm seeks to reduce $\mathbf{Q}$ by, first, leaking elements of transferable perturbations from a small auxiliary dataset and then exploiting these transferable components on the images in the larger test dataset. It has two phases (1) the *leaking phase*, where we attack the model using a traditional attack and gather information on successful attacks on a small auxiliary dataset (2) the *exploitation phase* where we attack the model using the leaked information from the first phase on the larger test set. This allows us to sharply reduce the number of amortized queries needed. The two phases of LTP are demonstrated in Figure 2.

Algorithm Our algorithm has two phases, the *leaking phase* and the *exploitation phase*. During the *leaking phase*, it performs a traditional black-box attack on a separate dataset $\mathcal{D}_s$ consisting of N_s images, drawn from the same distribution as our test dataset $\mathcal{D}$. We extract principal components from these perturbations using principal component analysis (PCA). During the *exploitation phase* we use the principal components to improve the efficiency of our black-box attacks on the test dataset $\mathcal{D}$. We accomplish this by querying the black-box using the leaked principal components using a modified IT-*alg*. We implement IT-NES, IT-Bandits-TD, IT-SimBA, and IT-Square as variants for IT-*alg*. We achieve strong attacks using fewer queries.

Leaking Phase During the *leaking phase* we apply a black-box attack on a leaking dataset $\mathcal{D}_s$. We attack all images $\boldsymbol{x} \in \mathcal{D}_s$ until we achieve successful attacks ($L(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}), \boldsymbol{r}) < \tau$, where τ is the success threshold) or until we use a maximum number of $\mathbf{Q}$ queries. We create a set $\mathcal{P}$ of generated

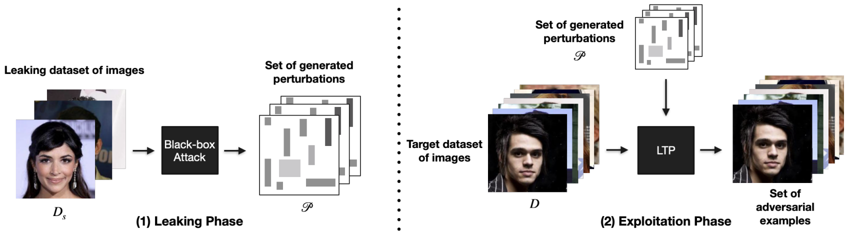

Figure 2: Illustration of our LTP pipeline. During the leaking phase, black-box image translation attacks are performed on the small leaking dataset and a set of generated perturbations $\mathcal{P}$ are collected. Using this set of generated perturbations, the algorithm finds strong attacks efficiently during the exploitation phase by exploring the perturbation directions given by the principal components of $\mathcal{P}$.

perturbations $\boldsymbol{\eta}$. Our framework is general and any attack or combination of attacks can be used for the leaking phase. We use PCA on perturbations $\boldsymbol{\eta} \in \mathcal{P}$ and extract principal components $\boldsymbol{q} \in \mathcal{Q}$.

Exploitation Phase Our *exploitation phase* consists of using a modified IT-SimBA using $\boldsymbol{q} \in \mathcal{Q}$ as candidate vectors. Since $\mathcal{Q}$ is not necessarily a basis of the image space (because $N_s < d^2$), and even though the initial iterations of the attack very rapidly decrease the loss, the attack might saturate. We switch to a full basis in image space Γ after a number of iterations n_{sat} of saturating loss. The resulting attacks achieve strong results using substantially fewer queries $\mathbf{Q}$.

4 Experiments

In this section we attack GANimation and StarGAN using IT-NES, IT-Bandits-TD, IT-SimBA, IT-Square and our proposed method LTP.

4.1 Experimental Setup

Architectures and Datasets We attack the GANimation Pumarola et al. (2018) and StarGAN Choi et al. (2018) architectures. For GANimation we attack three expressions coded with distinct facial action units (AUs) and present average results. The expressions correspond to “smile with closed eyes”, “smile with open eyes” and “surprised eyebrow raise”. For StarGAN we present average results over 5 different attribute classes. The classes are “black hair”, “blond hair”, “brown hair”, “female” and “old”. The dataset used for both architectures is the CelebA dataset Liu et al. (2015). For GANimation we attack 1,000 images using each expression, yielding 3,000 individual attacks. For StarGAN we attack 200 images using 5 different classes, yielding 1,000 individual attacks.

Implementation Details We use adapt versions of the official NES, Bandits-TD, SimBA and the state-of-the-art Square Attack implementations ban; sim; squ for the image translation scenario. We prepend the image translation versions with “IT”.

For GANimation we build our leaked PCA components using 100 images, for each of the three expressions evaluated. We attack them using IT-NES with a 0.005 success threshold and 1,000 max iterations. We perform 351.9 queries on average per image. For StarGAN we build our PCA components using 10 images and 5 classes. We attack them using IT-NES with a 0.05 success threshold and 1,000 max iterations. We perform 928.4 queries on average per image.

4.2 Experimental Results

GANimation We attack GANimation using an *identity attack*, where we select the target image $\boldsymbol{r}$ to be the input image, such that using our attack we push the network output to be the same as the input. We select a success threshold of $\tau = 0.005$, meaning that we halt the attack when $L(\boldsymbol{G}(\boldsymbol{x} + \boldsymbol{\eta}), \boldsymbol{x}) \leq \tau$. After a successful attack at this threshold the transformations by GANimation are not noticeable. We use a maximum number of queries $\boldsymbol{B} = 10,000$ for all methods. In Table 1 (left) we show comparisons between LTP, IT-NES, IT-Bandits-TD, IT-SimBA and IT-Square. We can see that LTP is much more efficient than other methods achieving a ˜56% reduction in average

GANimation

Attack	Avg. Queries ↓	Avg. Norm ↓	FID ↓	Success Rate ↑
IT-NES	598	1.82	6.30	98.8%
IT-Bandits-TD	855	4.38	8.49	96.3%
IT-SimBA	551	4.87	7.68	97.9%
IT-Square	531	5.00	8.96	98.8%
LTP	**231**	2.42	6.30	98.8%

StarGAN

Attack	Avg. Queries ↓	Avg. Norm ↓	FID ↓	Success Rate ↑
IT-NES	1,001	2.90	25.22	99.8%
IT-Bandits-TD	4,901	4.99	12.34	52.2%
IT-SimBA	444	5.93	50.39	100%
IT-Square	3,856	5.00	20.77	98.7%
LTP	**155**	5.28	45.01	100%

Table 1: Attack comparison on GANimation (left) and on StarGAN (right). We show the mean number of queries per image, the average norm of the perturbation, the FID score of the attacked outputs and the success rate percentage.

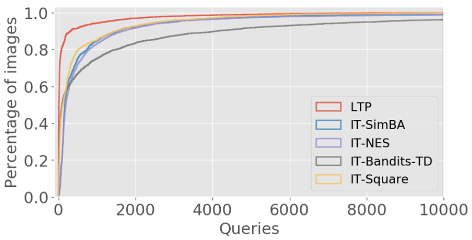

Figure 3: Success rate by number of queries on 3,000 attacks on GANimation (left) and 1,000 attacks on StarGAN (right).

queries (231 *vs.* 531) compared to the next best method (IT-Square). Our method also achieves a lower average perturbation norm than the comparable IT-SimBA attack as well as an improved success rate. We also compute FID scores Heusel et al. (2017) for the disrupted network outputs, comparing the feature distributions of attacked outputs with the feature distribution of the original images. Thus, FID measures how similar the attacked output images are to the intact inputs. LTP achieves the lowest FID score, reflecting that the images have been preserved to a greater extent. Figure 3 (left) shows the cumulative histogram of images successfully attacked for the number of queries represented by the x-axis. We observe that LTP achieves superior results compared to other algorithms.

StarGAN We attack 200 images on the StarGAN architecture using 5 different attribute classes. We use a *maximum distortion attack* (called *optimal attack* in Ruiz et al. (2020)), where the target image $\boldsymbol{r}$ is the non-attacked output image $\boldsymbol{G}(\boldsymbol{x})$ and we maximize the loss instead of minimizing it to achieve the maximum amount of distortion in the output image. We present results for a threshold $\tau = 0.05$, where the output image is visibly distorted. We use a maximum number of queries $\boldsymbol{B} = 10,000$ for all methods. In Table 1 (right) we show comparisons between IT-NES, IT-Bandits-TD, IT-SimBA, IT-Square and LTP. We can see that LTP is much more efficient than other methods achieving a reduction in mean queries of ˜65% compared to the next best attack and achieving a 100% success rate. We compute FID scores for the disrupted network outputs, comparing the feature distributions of attacked outputs with the feature distribution of the ground-truth network outputs. Thus, FID measures how different the attacked output images are to the deepfake generated images. LTP achieves a very high FID score, reflecting that the images have been corrupted to a large extent using the *maximum distortion* attack. In this case the average norm is higher than some competing methods. Qualitative analysis of images shows that the attack remains imperceptible. Figure 3 (right) shows the cumulative histogram of images successfully attacked for a specific number of queries. We observe that LTP achieves superior results compared to IT-SimBA, IT-NES, IT-Bandits-TD and IT-Square.

Conclusion We presented results for the strongest attack types for StarGAN (*maximum distortion*) and GANimation (*identity*). These are the first successful black-box attacks on image translation systems. Our results demonstrate that LTP is more efficient than competing methods. This is a consequence of the transferability of the leaked PCA components that are subsequently used as candidate vectors during the exploitation phase. We find that image translation architectures have specific vulnerabilities and that there exist correlations between attacks constructed for different images. This is the nugget of intuition that motivates our approach.

REFERENCES

https://github.com/MadryLab/blackbox-bandits.

https://www.faceapp.com.

https://github.com/cg563/simple-blackbox-attack.

https://github.com/max-andr/square-attack.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020.

Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 154–169, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

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*, pp. 15–26, 2017.

Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJlk6iRqKX.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. *arXiv preprint arXiv:1912.01865*, 2019.

Z. Geng, C. Cao, and S. Tulyakov. 3D guided fine-grained face manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9821–9830, 2019.

Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael Black, and Timo Bolkart. Gif: Generative interpretable faces. *arXiv preprint arXiv:2009.00149*, 2020.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. ICLR*, 2015.

Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2484–2493. PMLR, 2019. URL http://proceedings.mlr.press/v97/guo19a.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pp. 6626–6637, 2017.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 2142–2151, 2018.

Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=BkMiWhR5K7`.

Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 36–42. IEEE, 2018.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=HJGU3Rodl`.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *CoRR*, abs/1611.02770, 2016.

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.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

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*, pp. 2574–2582, 2016.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *arXiv preprint arXiv:1612.06299*, 2016.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7588–7597, 2019.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519. ACM, 2017.

Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 818–833, 2018.

A. Ranjan, T. Bolkart, S. Sanyal, and M. Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 704–720, 2018.

Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. *CoRR*, abs/2003.01279, 2020. URL `https://arxiv.org/abs/2003.01279`.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *In Proc. ICLR*, 2014.

Pedro Tabacof, Julia Tavares, and Eduardo Valle. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*, 2016.

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. *arXiv preprint arXiv:2004.00121*, 2020.

Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 742–749, 2019.

O. Wiles, A Sophia K., and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 670–686, 2018.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9459–9468, 2019.