Poster: A boostershot for transferable physically realizable adversarial examples

Verheyen Willem, Sander Joos, Tim Van hamme, Davy Preuveneers, Wouter Joosen
imec-DistriNet, KU Leuven

***Abstract*—Adversarial perturbations are claimed to enlarge the attack surface of machine learning models. However, as the most prominent attack methodologies require unrealistically strong adversaries, they are hardly used in attacks against real-world systems. In this paper, we alleviate the constraints on the threat model and attack a face recognition system with physically realizable perturbations in a black-box scenario, provided a single attack attempt. As such, we are forced to rely on more pragmatic, but less effective, attack methods that leverage transferability – adversarial perturbations successful on known models tend to also work on unknown ones. We overcome the poor attack success rate of transferability by using adversarially trained surrogate models.**

I. Introduction

A well-known problem of neural networks is their susceptibility to adversarial examples, e.g., images perturbed in such a way that changes are imperceptible to humans but impair the standard operation of neural networks. Despite a large body of work on methodologies to generate adversarial examples, the number of attacks on real-world models that take advantage of them is limited. This low prevalence in real-world attacks can be explained by the restrictiveness of the threat model that is present in practical ML-based systems. Attacking such systems requires a strong adversary with capabilities that are often unrealistic in practice. Moreover, the adversary's aim is to evade a face recognition system without knowledge about the network architecture but does possess information about a limited number of identities that can be recognized. Furthermore, the adversary's goal is to evade detection on the first try and is therefore limited to a single query without digital access to the target model.

Therefore, the adversary needs to rely on robust surrogate models to find a physical adversarial perturbation in the shape of glasses [1]. It is shown by previous studies [2]–[4] that robust models learn more universal representations of the training data as opposed to non-robust models. As such, the robust model learns better generalizing features, which serves as evidence that adversarial perturbations generated on robust surrogate models target features which are also present in other networks that fulfill a similar purpose.

II. Related work

In this section, we introduce related work that motivates the use of adversarially trained surrogate models to increase transfer-based adversarial examples.

Transferability allows an adversary to generate adversarial examples on a known model and use these to attack unknown models. Although transfer-based attacks have a relatively low success rate compared to other attack methods [5]–[8], their major benefit is the limited query amount to the target model.

One proposed method to increase the low success rate of transfer-based adversarial examples is to generate them on surrogate models that have been adversarially trained on attacks of similar nature [9]. The reason behind this is that models robust against adversarial examples learn more generalizing features that are shared with other DNNs. Therefore, adversarial examples that exploit these features transfer better than those generated on non-robust DNNs.

Increasing the success rate of transfer-based adversarial examples has been explored further in the scope of digital adversarial examples [3], [4], [10]–[12]. Yet, similar to the transferability of adversarial examples, current understanding of this topic remains incomplete, especially for physical adversarial examples.

III. Attack methodology and experimental setup

In this section, we first describe the attack methodology followed by the experimental setup; among others, we specify how the adversarial perturbations are generated, and describe the data and models used.

A. Attack methodology

1) Obtain a collection of samples to train the surrogate model. This contains samples of the attacker that ultimately need to be misclassified, but also samples from other identities assumed to be recognized by the model under attack.
2) Perform adversarial training on the surrogate model with the collected set of samples by finding optimal perturbations.
3) Use the newly robust surrogate model to craft adversarial examples and only utilize those that are successful on the surrogate while their benign counterpart is also classified correctly.
4) Use the attacks that successfully fooled the surrogate model to attack the target model.

B. Experimental setup

To perform our evaluation in light of face recognition we require: 1) a sufficiently large face dataset which can be split into two training and one attack portion, 2) different model architectures, 3) a methodology to generate physical realizable attacks, and 4) models with a varying degree of robustness to obtain the adversarial test sets.

In the following we describe how we obtain each of the aforementioned requirements.

Req. 1: In contrast to prior work that investigates transferability between DNNs that classify identical classes [9], [13], we make a distinction between the identities recognized by the face recognition surrogate- and target model with some degree of overlap. We therefore use a portion of the VGGFACE2-dataset [14] and split this into three distinct subsets, each consisting of 400 identities that represent one core dataset shared by both surrogate- and target model and one additional set of identities for both the surrogate and target model respectively.

Req. 2: We consider the following architectures for both the surrogate- and target models, for which we use pre-trained weights: VGGFace pretrained on the VGGFace dataset [14], Facenet pretrained on the VGGFACE2 dataset [15] and VGG19 pretrained on the VGGFACE2 dataset. We fine-tune these models with our data, such that they classify the adequate set of identities.

Req. 3: Our physically realizable adversarial examples are based on the work by Sharif et al. [1]. we consider physically realizable adversarial examples with localized perturbations in the shape of glass frames. The adversarial glasses are always cropped to match the size of a person's face and rotated accordingly.

Req. 4: We first construct a number of increasingly robust classifiers and then use these to construct test sets containing adversarial examples. Robust classifiers $\hat{f}_i$ are obtained with adversarial training on the initial classifier f, for a number of epochs i with adversarial examples generated from our dataset D. For each surrogate model, we then generate a test set $\hat{D}_i$ containing adversarial examples that are misclassified, and thus successful as an attack. Adversarial attacks are only added to the adversarial test set if their benign counterpart is still classified correctly by the surrogate model. This allows us to partially mitigate the effect of the well-known problem where adversarial training causes a drop in standard accuracy [16], [17]. Subsequently, we use these test sets to evaluate the transferability of our robustly generated adversarial examples to other classifiers. We only consider adversarial examples on the target model when their benign counterpart is also classified correctly. This is an important assumption, as it is very likely whenever a benign sample is misclassified, its resulting adversarial examples are also likely to be misclassified.

IV. Evaluation

We use the adversarial test sets generated on the surrogate models to attack the different target models and measure the transfer rate across them.

First, we demonstrate that adversarial examples generated on robust surrogate models have a higher probability of success when used to attack target models. In order to do so, we compare the transfer rates of attacks in adversarial test sets $\hat{D}_0$ and $\hat{D}_{max}$, where $\hat{D}_0$ contains adversarial examples generated on the non-adversarially trained surrogate and $\hat{D}_{max}$ contains adversarial examples generated on the most robust surrogate. Fig. 1 shows the distribution of transfer rates for both $\hat{D}_0$ (non-robust) and $\hat{D}_{max}$ (robust) when they are used to attack different target models. The target models considered are both not adversarially trained, and adversarially trained using the adversarial examples of similar nature. The transfer rate of adversarial examples generated on robust surrogate models increases from an average of 27% to 40% and from 4% to 15% for non-robust and robust target models respectively.

This shows that even when the target model is robust against the considered adversarial examples, the attack success rate can be increased when using robust surrogate models.

Next, Fig. 2 shows that as model robustness increases as a result of adversarial training, so does the transferability of physical realizable adversarial examples. This is in line with findings in previous work that claims that the classifier becomes more robust to adversarial examples, they rely more on robust features instead of non-robust features. As a result, features learned by robust classifiers benefit a higher degree of universality, whereas non-robust features are less universal and thus transfer worse [12].

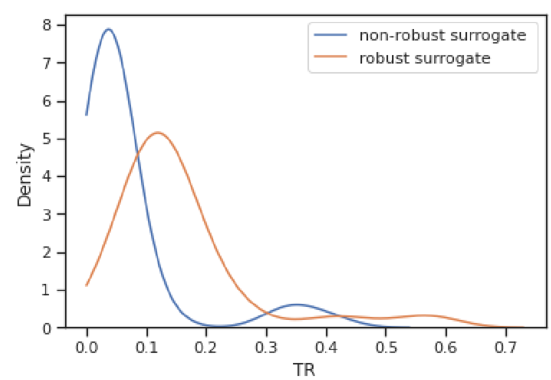


Fig. 1: KDE-plot of the transfer ratio of physical realizable adversarial examples generated on robust and non-robust surrogate models and transferred to non-robust and robust target models.

V. Discussion

This section discusses the main implications of our work, shortcomings, and possible directions for future work.

*a) **Security implications**:* In this work, we aimed to increase the transferability of physical adversarial perturbations to better accommodate threat models that require physically realizable perturbations against black-box models with a near zero-query budget. Our results extend the findings of recent works [3], [9]–[11] that already provided evidence that the transfer rate of adversarial examples with unlocalized perturbations increases when generated on robust classifiers. In this work, we leverage these findings and propose model robustness as a prior for the generation of physical realizable adversarial examples. Specifically, we demonstrate an impressive increase in attack success rate between 1.5x and 7x against

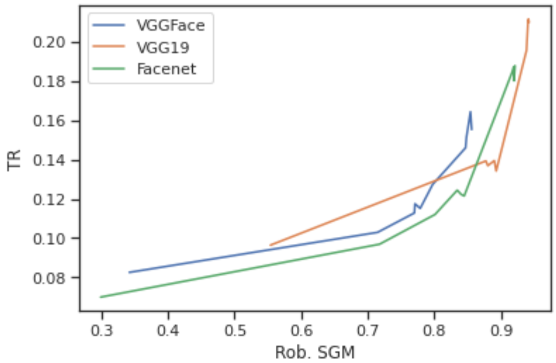


Fig. 2: Average transfer rates for increasingly robust surrogate models. The X-axis represents the accuracy on adversarial examples of the surrogate model, the y-axis represents the transfer rate to different black box target models.

face recognition systems when using robust surrogate models over their non-robust counterparts. On the one hand, with an absolute success rate which varies between 6 and 56% we are not consistently evading detection by face recognition systems. On the other hand, in an impersonation scenario against an authentication system, such an increase is significant, as often multiple authentication attempts are allowed [18], e.g., FaceID allows five authentication attempts before switching to PIN input.

*b) **Validity threats**:* The experimental setup inherits several less realistic assumptions on the proposed threat model which we attempt to solve in our current ongoing work.

We assume both target- and surrogate models use partially overlapping datasets in their training procedure. This is an exaggerated simplification of a real-world scenario where an attacker has limited to no knowledge of the data used to train the target model. However, we assume that further minimizing this overlap will have a limited impact on our results because the intuition behind deep neural networks is that they can distribute feature space evenly across different classes, separating each class with similar distances to each other. Therefore, we propose the use of feature extractors to overcome relying on overlapping datasets which, in turn, also relates better to face recognition/authentication as an open-world problem. Furthermore, we plan on considering impersonation attacks as well to bridge the transition to authentication.

VI. Conclusion

In this paper, we propose a method to increase the attack success rate of adversarial examples to face recognition systems in a highly restrictive, yet realistic black-box setting. We do so by leveraging and enhancing the transferability property of adversarial examples that are realizable in the physical world by generating attacks on adversarially trained surrogate models. Specifically, we found that using a robust surrogate model over its non-robust counterpart drastically increases transferability with a factor of 1.5 up to 7 for single attempt attacks compared to the state-of-the-art. Moreover, even in the case of low absolute attack success rates such increases are significant for attacks against applications that allow for more than one attempt but implement rate limiting, e.g., face authentication systems. In conclusion, we believe that this work provides a compelling contribution to the creation of adversarial examples that impose a significant threat to practical machine learning applications.

References

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

[2] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," 2019.

[3] M. Terzi, A. Achille, M. Maggipinto, and G. A. Susto, "Adversarial Training Reduces Information and Improves Transferability," *arXiv:2007.11259 [cs, stat]*, Dec. 2020. arXiv: 2007.11259.

[4] J. M. Springer, M. Mitchell, and G. T. Kenyon, "A little robustness goes a long way: Leveraging robust features for targeted transfer attacks," 2021.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv:1412.6572 [cs, stat]*, Mar. 2015. arXiv: 1412.6572.

[6] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into Transferable Adversarial Examples and Black-box Attacks," *arXiv:1611.02770 [cs]*, Feb. 2017. arXiv: 1611.02770.

[7] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," 2017.

[8] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples," *arXiv:1605.07277 [cs]*, May 2016. arXiv: 1605.07277.

[9] J. M. Springer, M. Mitchell, and G. T. Kenyon, "Adversarial perturbations are not so weird: Entanglement of robust and non-robust features in neural network classifiers," 2021.

[10] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry, "Do Adversarially Robust ImageNet Models Transfer Better?," *arXiv:2007.08489 [cs, stat]*, Dec. 2020. arXiv: 2007.08489.

[11] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," 2020.

[12] J. M. Springer, M. Mitchell, and G. T. Kenyon, "Adversarial Perturbations Are Not So Weird: Entanglement of Robust and Non-Robust Features in Neural Network Classifiers," *arXiv:2102.05110 [cs]*, Feb. 2021. arXiv: 2102.05110.

[13] F. Utrera, E. Kravitz, N. B. Erichson, R. Khanna, and M. W. Mahoney, "Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification," *arXiv:2007.05869 [cs, stat]*, Apr. 2021. arXiv: 2007.05869.

[14] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," 2018.

[16] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," 2019.

[17] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," 2019.

[18] P. Markert, D. V. Bailey, M. Golla, M. Dürmuth, and A. J. Aviv, "This pin can be easily guessed: Analyzing the security of smartphone unlock pins," 2021.

KU LEUVEN

DEPARTMENT OF COMPUTER SCIENCE

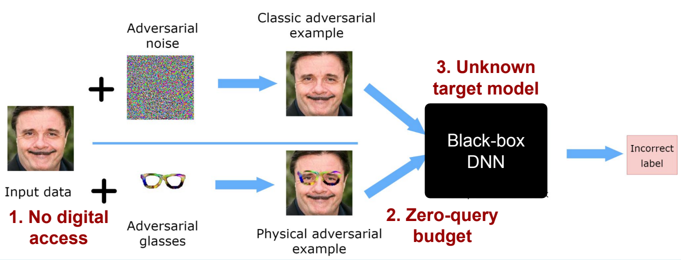
A boostershot for transferable physically realizable adversarial examples

Willem Verheyen, Sander Joos, Tim Van hamme, Davy Preuveneers, Wouter Joosen

imec-DistriNet, KU Leuven

Problem

Adversarial examples: Seemingly benign images that cause misclassification in a machine learning model.



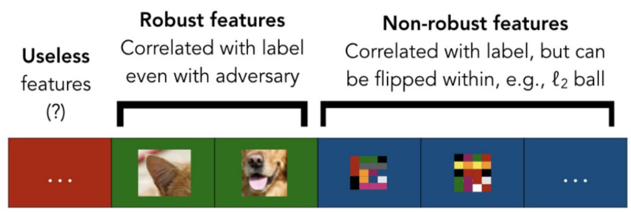
Motivation

Classic adversarial examples
- ➔ Requires digital access to target
- ➔ Limit on perturbation magnitude
- ➔ Global perturbations are not always possible

Physical adversarial examples
- ➔ Only requires physical access to target
- ➔ No limited perturbation magnitude
- ➔ Localized perturbations

Transfer-based attacks against black box model
- ➔ Relates to real-world scenario
- ➔ No query access required
- ➔ On-the-spot attack generation

Background

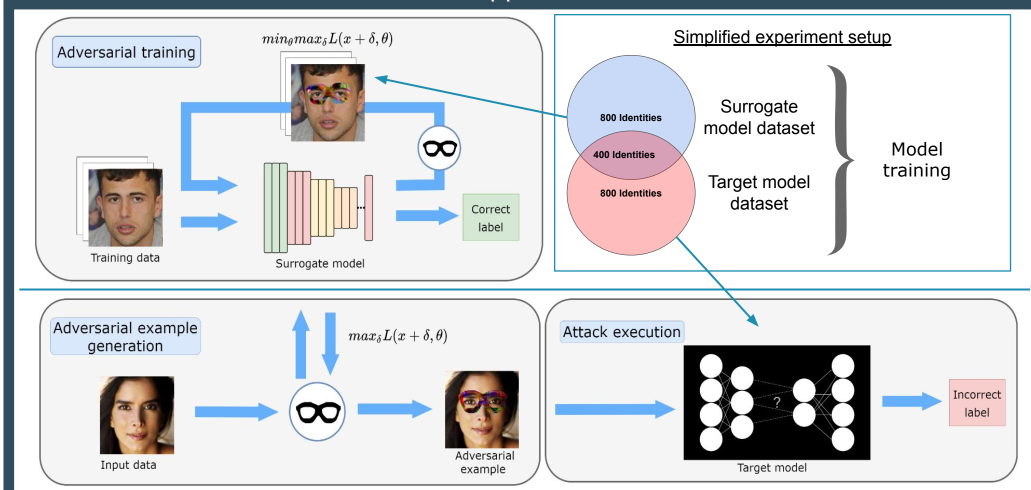


Adversarial training reduces influence of non-robust features[1]

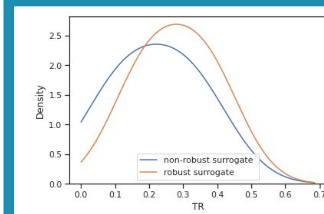
Robust features are shared between different models = universal[2]

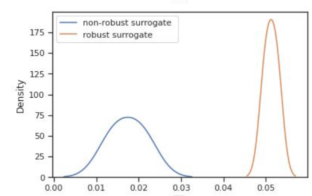
Adversarially examples generated on adversarially trained DNN exploit universal features.[2]

Approach



Results





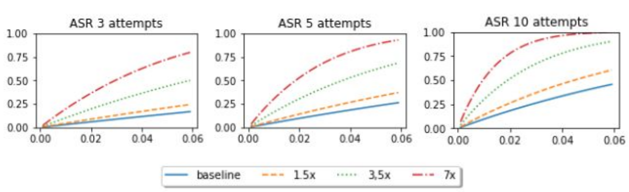
Top-1 ASR against standard models (top) & robust models (bottom).

Contributions

1. Improved single-attempt success rate against black box models.

Increased top1-ASR with a factor up to 7.

2. Security implications against black-box models with low query budget.



Future

(1) closed-world → open-world assumption:
- ➔ Current simplified experiment setup: classifier that recognizes predetermined set of identities used during model training.
- ➔ Extend to rely on feature extraction: recognize any identity independent of training based on distance metrics.

(2) (Evasion +) impersonation attack goal:
- ➔ Current experiments only involve untargeted attacks = evasion
- ➔ Extend with targeted attacks = impersonate some chosen target identity

1. *Adversarial Examples Are Not Bugs, They Are Features.* Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry
2. *Adversarial Perturbations Are Not So Weird: Entanglement of Robust and Non-Robust Features in Neural Network Classifiers.* Jacob M. Springer, Melanie Mitchell, Garrett T. Kenyon

KU Leuven
Department of Computer Science
Celestijnenlaan 200A box 2402
3001 Leuven, Belgium

Contact
Verheyen Willem
willem.verheyen@kuleuven.be

DistriNet