Adversarial Attacks on Classifiers for Eye-based User Modelling

Inken Hagestedt
CISPA Helmholtz Center for
Information Security,
Saarland Informatics Campus
inken.hagestedt@cispa.saarland

Michael Backes
CISPA Helmholtz Center for
Information Security,
Saarland Informatics Campus
backes@cispa.saarland

Andreas Bulling
University of Stuttgart ,
Institute for Visualisation and
Interactive Systems
andreas.bulling@vis.uni-stuttgart.de

ABSTRACT

An ever-growing body of work has demonstrated the rich information content available in eye movements for user modelling, e.g. for predicting users' activities, cognitive processes, or even personality traits. We show that state-of-the-art classifiers for eye-based user modelling are highly vulnerable to *adversarial examples*: small artificial perturbations in gaze input that can dramatically change a classifier's predictions. On the sample task of eye-based document type recognition we study the success of adversarial attacks with and without targeting the attack to a specific class.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**.

KEYWORDS

Privacy-aware eye tracking, Gaze Data, Eye Movements, Gaze Behaviour, Eye-based user modelling

ACM Reference Format:
Inken Hagestedt, Michael Backes, and Andreas Bulling. 2020. Adversarial Attacks on Classifiers for Eye-based User Modelling. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Adjunct), June 2–5, 2020, Stuttgart, Germany*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3379157.3390511

1 INTRODUCTION

Recent advances in mobile eye tracking [Kassner et al. 2014; Tonsen et al. 2017] and gaze estimation using off-the-shelf cameras [Zhang et al. 2017, 2019] have spurred research on eye-based user modelling. That is, the prediction of various user characteristics from eye movements, such as users' activities [Bulling et al. 2010], cognitive processes and states [Bulling and Zander 2014; Sattar et al. 2015], or personality traits [Hoppe et al. 2018]. Combined with the continuing integration of eye tracking into head-mounted virtual and augmented reality headsets, this promises a range of exciting new applications, such as mental health monitoring [Vidal et al. 2012] or life logging and quantified self [Kunze et al. 2013a, 2015].

However, with widespread availability of gaze data comes an ever-increasing risk of misuse and privacy attacks. Examples for adversaries are headset manufacturers trying to obtain personal information about consumer interests or preferences, malicious applications running on the computer to which the headset is connected that spy on users' activities, or third parties launching targeted attacks on gaze data. Despite these diverse threats, research has mainly focused on ocular biometrics [Nigam et al. 2015] or secure user authentication using eye movements [Holland and Komogortsev 2011]. Researchers have only recently started to study these threats and proposed first solutions to making eye tracking privacy-aware – both at the hardware [Steil et al. 2019b] and software [John et al. 2019; Liu et al. 2019; Steil et al. 2019a] level.

This work contributes another building block to this emerging research field of *privacy-aware eye tracking* by studying, for the first time, adversarial attacks. Such attacks create small perturbations in gaze data input, better known as *adversarial examples*, that dramatically change the classifier's predictions. Adversarial examples have only recently started to being studied in the intersecting research field of computer vision and security [Papernot et al. 2018] but have not yet received any attention in the eye tracking community. We study adversarial attacks on classifiers for eye-based user modelling for the sample task of document type recognition from eye movements during reading [Kunze et al. 2013b]. We picked this task given that reading is a truly pervasive activity, widely studied in different fields including eye tracking research, and has been the subject of a recent study on using differential privacy for privacy-aware eye tracking [Steil et al. 2019a]. We study the feasibility and performance of this attack in different scenarios that we carefully chose to represent real-world use cases: Attacks with and without knowledge about the internal classifier gradients as well as with and without targeting the attack to a specific class. Our results demonstrate that classifiers can not be trusted blindly since they can be mislead by adversarial examples.

2 CREATING ADVERSARIAL EXAMPLES

Following previous work [Bulling et al. 2012, 2010, 2013; Steil et al. 2019a], we directly targeted Support Vector Machines (SVM) with radial basis function (RBF) kernels. Additionally, we indirectly targeted Random Forests (RF) [Kunze et al. 2013b] by generating adversarial examples against an SVM fitted on the same training data to study transferability. We compute adversarial examples with the Fast Gradient Sign Method [Goodfellow et al. 2014] (FGSM), a state-of-the-art method that is independent of neural network structures but only uses gradients which are well-defined for SVM. FGSM computed the gradient and perturbed the sample in this direction, we used its minimal mode which repeatedly computes a growing

ETRA '20 Adjunct, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7135-3/20/06…$15.00
https://doi.org/10.1145/3379157.3390511

adversarial perturbation until the sample is misclassified. We opted for the minimal attack because it lead to smaller perturbations.

The L_2 norm was used to measure perturbations, it is wildly used for the generation of adversarial examples and remains small under many small changes to the features [Carlini and Wagner 2017]. We kept the perturbation per step ε_s at 0.1 during the experiments and evaluated two different methods to find the ε_{max} hyperparameter for FGSM: The general ε_{max} was chosen such that the average accuracy over multiple participants was lowest, this corresponds to an attack without prior knowledge about the target's data. We chose the person-specific ε_{max} by determining the maximal perturbation for the lowest accuracy for each person individually. The goal of the untargeted attack is to misclassify the sample into any of the other classes, while targeted attacks perturbed samples of one class only such that they were misclassified as one specific other class.

3 EVALUATION

We used the public dataset by [Steil et al. 2019a] that contains recordings of 20 participants reading three different document types (comic, newspaper, textbook) in virtual reality. We first detected fixations using a dispersion-based algorithm, and then extracted 52 high-level features from the basic eye movements following [Bulling et al. 2010], using a sliding window. Two more features from [Kunze et al. 2013b] were added, to give a general estimate of the reading direction and distance covered during the time window.

SVM Training. We studied the task of recognising different document types from eye movements during reading. Using sklearn's SVM implementation [Buitinck et al. 2013], we trained with leave-one-person-out cross-validation. The optimal window size of 45 seconds was selected based on the validation accuracy on 200 samples per participant and document type. We kept the penalty parameter C on the error term at its default of 1.0 and the RBF-kernel hyperparameter γ that controls the locality of the kernel at $\frac{1}{54}$.

Random Forest Training. Similarly to before, we selected the best window size (45 seconds again) using leave-one-person-out cross-validation on the same data split into training, validation, and test data. We evaluated for the key hyperparameters, namely the number of trees (100, 50, 10, 200) and the number of samples per leaf (50, 10, 100, 5) to avoid overfitting.

3.1 Results

Attack Success: The accuracy after attack is well below chance level of 1/3 with only one exception, thus the SVM is vulnerable to our attack. The accuracy of general and person-specific choice of ε_{max} are very close, thus we show the general choice of ε_{max} in Figure 1. That means, it is not necessary to know data of the target for mounting a successful attack.

Distance Evaluation: In oder to evaluate the consequences of our adversarial perturbation, we measured the euclidean distance between all test samples before the attack. This naturally occurring distance between samples was compared to the distances between test samples and their corresponding perturbed version. We studied three different ways to select the FGSM hyperparameter ε_{max}: person-specific, general or Additionally, we selected the smallest ε_{max} such that on average over all participants, an accuracy of 0.3 is reached. If the goal is only guessing accuracy, smaller perturbations often suffice. We observe that in most cases, on average the distance between original and perturbed point is smaller than the average distance between benign test points. The p-values of Welch's t-test between the two distributions of distances are below 0.01 except for the targeted attacks misclassifying comic as textbook and textbook as comic, respectively. This demonstrates that the perturbations we computed are indeed mostly "small".

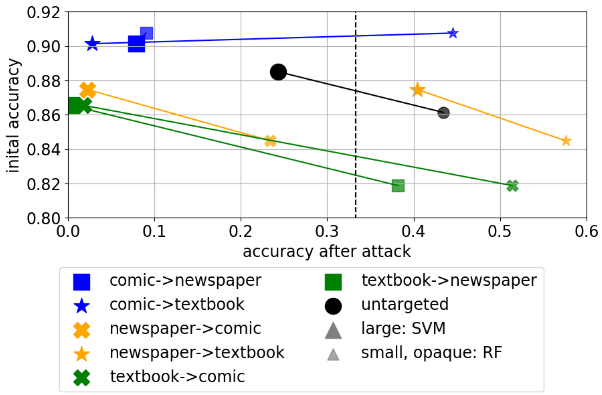

Figure 1: Accuracies after attacks on SVM with FGSM (larger markers) and transfer to RF (smaller markers). Different colors and markers show different document types under attack. The dashed black line visualizes the chance level.

Transferability: Finally, we study whether the perturbations carry over to a different family of classifiers, namely, RF classifiers, who had a similar accuracy on test data. So we used RF to classify the adversarial examples against the SVM. We observe that the accuracy drops after the attack, but only rarely below guessing accuracy for RF. That means, the decision boundaries between SVMs and RF are similar enough for some samples to transfer, however, not so similar that all of them carry over. The distances are similar to those observed for the SVM, but in some cases higher. We conclude that knowledge on the type of classifier does increase attack accuracy, however, hiding the type of classifier does not mitigate all attacks.

4 DISCUSSION AND CONCLUSION

Our results demonstrated that it is easy to fool classifiers for eye tracking data, thus we can not rely on their outputs. For desirable classifications, our findings raise the question whether the current techniques are robust enough for naturally occurring noise. Phrased differently, knowing that there exist small shifts in the data that change the classifiers' outcomes, how can we ensure these shifts do not occur naturally and lead to failure of the eye tracking system? On the other hand, for undesirable classifications such as extraction of privacy-sensitive information or activity surveillance, adversarial examples may be a way to circumvent these classifications. We leave it to future work to explore the use of adversarial examples as a protection mechanism.

REFERENCES

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. PKDD, Prague, Czech Republic, 108–122.

Andreas Bulling, Jamie A. Ward, and Hans Gellersen. 2012. Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors. *ACM Transactions on Applied Perception* 9, 1 (2012), 2:1–2:21. https://doi.org/10.1145/2134203.2134205

Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. 2010. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence* 33, 4 (2010), 741–753.

Andreas Bulling, Christian Weichel, and Hans Gellersen. 2013. EyeContext: recognition of high-level contextual cues from human visual behaviour. In *Proceedings of the sigchi conference on human factors in computing systems*. ACM, ACM, Paris, 305–308.

Andreas Bulling and Thorsten O Zander. 2014. Cognition-aware computing. *IEEE Pervasive Computing* 13, 3 (2014), 80–83.

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

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* 1412.6572, 14126572 (2014), 0.

Corey Holland and Oleg V Komogortsev. 2011. Biometric identification via eye movement scanpaths in reading. In *2011 International joint conference on biometrics (IJCB)*. IEEE, IEEE, United States, 1–8.

Sabrina Hoppe, Tobias Loetscher, Stephanie A Morey, and Andreas Bulling. 2018. Eye movements during everyday behavior predict personality traits. *Frontiers in human neuroscience* 12 (2018), 105.

Brendan John, Sanjeev Koppal, and Eakta Jain. 2019. EyeVEIL: degrading iris authentication in eye tracking headsets. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. ACM, ACM, New York, NY, 37.

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adj. Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, Seattle, US, 1151–1160. https://doi.org/10.1145/2638728.2641695

Kai Kunze, Masakazu Iwamura, Koichi Kise, Seiichi Uchida, and Shinichiro Omachi. 2013a. Activity recognition for the mind: Toward a cognitive" Quantified Self". *Computer* 46, 10 (2013), 105–108.

Kai Kunze, Katsutoshi Masai, Masahiko Inami, Ömer Sacakli, Marcus Liwicki, Andreas Dengel, Shoya Ishimaru, and Koichi Kise. 2015. Quantifying reading habits: counting how many words you read. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, ACM, Osaka, 87–96.

Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. 2013b. I know what you are reading: recognition of document types using mobile eye tracking. In *Proceedings of the 2013 International Symposium on Wearable Computers*. ACM, ACM, Zurich, Swizerland, 113–116.

Ao Liu, Lirong Xia, Andrew Duchowski, Reynold Bailey, Kenneth Holmqvist, and Eakta Jain. 2019. Differential privacy for eye-tracking data. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ETRA 19*. ACM, ACM, Denver, Colorado, 1–10.

Ishan Nigam, Mayank Vatsa, and Richa Singh. 2015. Ocular biometrics: A survey of modalities and fusion approaches. *Information Fusion* 26 (2015), 1–35.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. 2018. SoK: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, IEEE, London, United Kingdom, 399–414.

Hosnieh Sattar, Sabine Müller, Mario Fritz, and Andreas Bulling. 2015. Prediction of Search Targets From Fixations in Open-world Settings. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, 981–990. https://doi.org/10.1109/CVPR.2015.7298700

Julian Steil, Inken Hagestedt, Michael Xuelin Huang, and Andreas Bulling. 2019a. Privacy-aware eye tracking using differential privacy. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ETRA 19*. ACM, ACM, Denver, 27.

Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. 2019b. Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. ACM, ACM, Denver, 26.

Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 3 (2017), 106:1–106:21. https://doi.org/10.1145/3130971

Mélodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. 2012. Wearable Eye Tracking for Mental Health Monitoring. *Computer Communications* 35, 11 (2012), 1306–1311. https://doi.org/10.1016/j.comcom.2011.11.002

Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2017. Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery. In *Proc. ACM Symposium on User Interface Software and Technology (UIST)*. ACM, Quebec City, Canada, 193–203. https://doi.org/10.1145/3126594.3126614

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2019. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41, 1 (2019), 162–175. https://doi.org/10.1109/TPAMI.2017.2778103