EEG-Based Brain-Computer Interfaces are Vulnerable to Adversarial Attacks

Dongrui Wu
Ministry of Education Key Laboratory of Image Processing and Intelligent Control
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology, China
Email: drwu@hust.edu.cn.

Original paper:
X. Zhang, D. Wu, L. Ding, H. Luo, C-T Lin, T-P Jung and R. Chavarriaga, "Tiny noise, big mistakes: Adversarial perturbations induce errors in Brain-Computer Interface spellers," *National Science Review*, 2020.
https://doi.org/10.1093/nsr/nwaa233

Machine learning has been extensively used in EEG-based brain-computer interfaces (BCIs) for brain signal classification. Most studies so far focused on making the BCI classifiers faster and more reliable; however, few have considered their security. It has been found in other application domains that adversarial examples [1], which are normal examples contaminated by deliberately designed tiny perturbations, can easily fool machine learning models. These perturbations are usually so small that they are indistinguishable to human eyes.

EEG-based BCI systems may also be attacked by adversarial examples. The consequence could range from merely user frustration to severe misdiagnosis in clinical applications [2]. We believe a new and more detailed understanding of how adversarial EEG perturbations affect BCI classification can inform the design of BCIs to defend against such attacks.

State-of-the-Art

Adversarial attacks of EEG-based BCIs were first investigated by Zhang and Wu in 2019 [2]. We considered three attack types, whose characteristics are summarized in Table I. White-box attacks assume that the attacker has access to all information of the target model, including its architecture and parameters. Gray-box attacks assume the attacker knows some but not all information about the target model, e.g., the training data that the target model is tuned on, but not its architecture and parameters. Black-box attacks assume the attacker knows neither the architecture nor the parameters of the target model, but can observe its responses to inputs. White-box attacks are the easiest, and black-box attacks are the hardest.

TABLE I
SUMMARY OF THE THREE ATTACK TYPES [2].

Target model information	White-Box	Gray-Box	Black-Box
Know its architecture	✓	×	×
Know its parameters $\boldsymbol{\theta}$	✓	×	×
Know its training data	−	✓	×
Can observe its response	−	−	✓

We [2] proposed an adversarial attack framework, shown in Fig. 1, for EEG-based BCIs. A jamming module is injected between signal preprocessing and feature engineering to generate adversarial examples. We proposed an unsupervised fast gradient sign method to construct the jamming module and attack three popular convolutional neural network classifiers in BCIs. Its effectiveness was verified in three different BCI paradigms (P300 evoked potentials, feedback error-related negativity, and motor imagery). To reduce the number of training samples in black-box attacks, we further proposed query synthesis based active learning to improve the query efficiency in training the substitute classification model [3]. Meanwhile, we also performed white-box target attacks for two

BCI regression problems (EEG-based driver drowsiness estimation, and EEG-based user reaction time estimation in psychomotor vigilance tasks) [4].

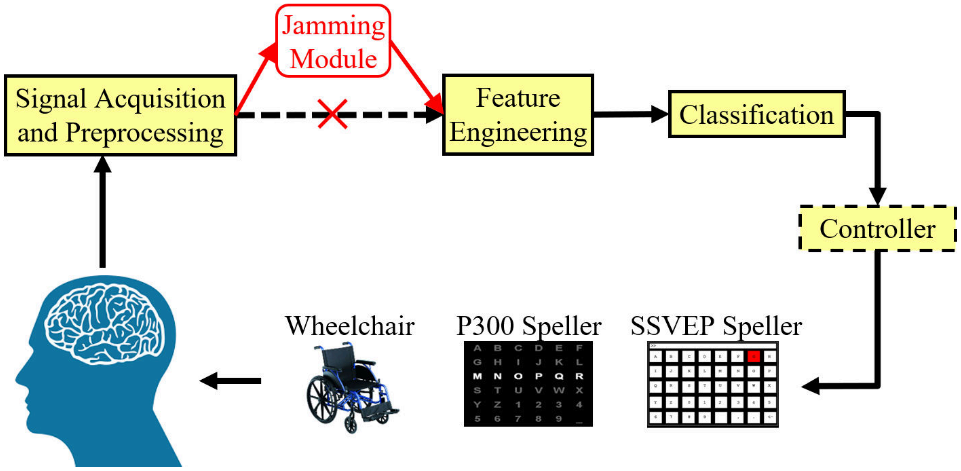

Fig. 1. The proposed attack framework [2] to a closed-loop BCI system, which injects a jamming module between signal preprocessing and feature engineering to generate adversarial examples. A controller may not be needed in certain applications, e.g., BCI spellers.

Though theoretically important, our previous approaches (actually also most adversarial attack approaches to time-series signals in the literature [2], [5]–[7]) have a serious limitation: they need to know the full EEG trial before computing the adversarial perturbations, i.e., they are not causal and hence cannot be implemented in online applications. This limitation may be more easily explained by taking adversarial attacks to speech signals as an example. To attack a voice command, most current approaches need to record the entire voice command first, and then design the perturbation. However, once the perturbation is obtained, the voice command has already been sent out (e.g., to a smartphone or Amazon Echo), so there is no chance to add the perturbation to the voice command to perform the attack online.

Our Approach

Two new adversarial attack approaches to EEG-based BCIs were proposed in our latest publication [8]. What distinguishes them most from previous ones is that they explicitly considered the causality in designing the perturbations. The adversarial perturbation template is constructed directly from the training set and then fixed. So, there is no need to know the test EEG trial and compute the perturbation specifically for it. The perturbation can be directly added to a test EEG trial as soon as it starts, hence satisfies causality and can be implemented online.

We show for the first time that one can generate tiny adversarial EEG perturbation templates for target attacks for both P300 and SSVEP spellers, i.e., mislead the classification to any character the attacker wants, regardless of what the user intended character is. For the P300 speller, the attacker can construct a perturbation template, which makes the P300 speller classify any perturbed EEG epoch into a target one, whether the benign EEG epoch is target or non-target. For the SSVEP speller, the attacker can generate many perturbation templates, each corresponding to a different character. By adding a specific perturbation template to a benign EEG trial, the attacker can mislead the SSVEP speller to output the corresponding character, no matter what the user intended output is.

Fig. 2 illustrate the attack procedure for P300 spellers.

Experimental Validations

We considered adversarial attacks to both P300 and SSVEP based BCI spellers. Their effectiveness was demonstrated on three publicly available datasets: P300 Dataset II in BCI Competition III[1], P300 speller with ALS patients[2]

[1] http://www.bbci.de/competition/iii/#data_set_ii

[2] http://bnci-horizon-2020.eu/database/data-sets

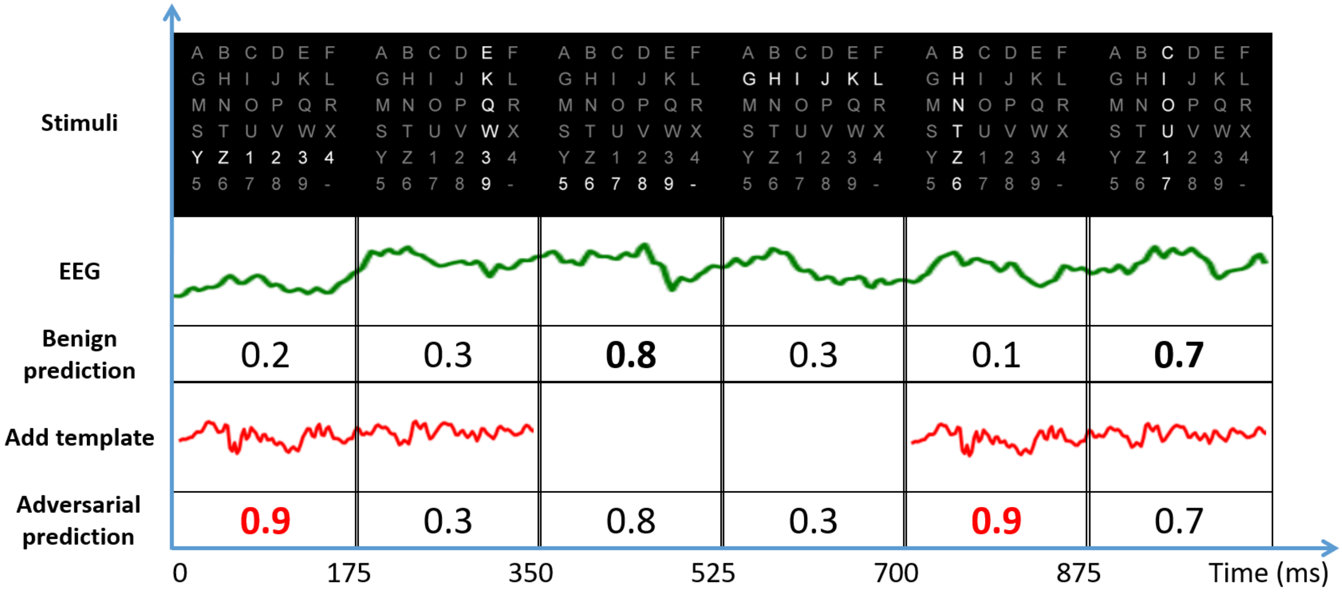

Fig. 2. Illustration of the attack procedure in the P300 protocol [8]. The attacker character is *Z*, whereas the user character is *7*. For the benign EEG trial, the P300 speller can correctly identify that P300 is elicited by the intensifications of the last row and the third column. To mislead the P300 speller, adversarial perturbation template is added during the periods of 0-350ms and 700-1050ms, so that the fifth row and the second column are believed to elicit P300 with the highest probability. The added adversarial perturbation templates do not influence the results of the second and the last stimuli, because their corresponding periods are out of synchronization with the templates. As a result, the P300 speller misclassifies the perturbed trial to attacker character *Z*.

(008-2014) [9], and the Tsinghua University SSVEP dataset[3]. All source code is available on GitHub[4].

For the two subjects in Dataset II, the attacker can manipulate the P300 speller to spell whatever character he/she wants, regardless of what the user intended character is, with a higher than 90% average success rate. For four of the eight ALS patients, the attack made the P300 speller almost completely useless. For six of the eight subjects in the SSVEP dataset, their output character can be manipulated to any character the attacker wanted, at 70%-100% success rate.

Importance

We show, for the first time, that tiny noise can significantly manipulate the outputs of P300 and SSVEP spellers, exposing a critical security concern in BCIs. Our generated adversarial perturbation templates satisfy the causality of time-series signals, which rarely drew much attention before.

We need to emphasize that the goal of this study is not to damage EEG-based BCIs. Instead, we aim to demonstrate that serious adversarial attacks to EEG-based BCIs are possible, and hence expose a critical security concern, which has received little attention before. Interestingly, a very recent Nature Medicine study also revealed that deep learning models for electrocardiograms are susceptible to adversarial attack [10].

Our future research will develop strategies to defend against such attacks. Meanwhile, we hope our study can attract more researchers' attention to the security of EEG-based BCIs, and more broadly, wearable devices.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int'l Conf. on Learning Representations*, Banff, Canada, Apr. 2014.

[2] X. Zhang and D. Wu, "On the vulnerability of CNN classifiers in EEG-based BCIs," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 5, pp. 814–825, 2019.

[3] X. Jiang, X. Zhang, and D. Wu, "Active learning for black-box adversarial attacks in EEG-based brain-computer interfaces," in *Proc. IEEE Symposium Series on Computational Intelligence*, Xiamen, China, Dec. 2019.

[3] http://bci.med.tsinghua.edu.cn/download.html

[4] https://github.com/ZhangXiao96/Speller-Attacks

[4] L. Meng, C.-T. Lin, T.-P. Jung, and D. Wu, "White-box target attack for EEG-based BCI regression problems," in *Proc. Int'l Conf. on Neural Information Processing*, Sydney, Australia, Dec. 2019.

[5] N. Carlini and D. A. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Symposium on Security and Privacy*, San Francisco, CA, May 2018, pp. 1–7.

[6] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Adversarial attacks on deep neural networks for time series classification," in *Int'l Joint Conf. on Neural Networks*, Budapest, Hungary, Jul. 2019, pp. 1–8.

[7] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proc. 36th Int'l Conf. on Machine Learning*, Long Beach, CA, May 2019, pp. 5231–5240.

[8] X. Zhang, D. Wu, L. Ding, H. Luo, C.-T. Lin, T.-P. Jung, and R. Chavarriaga, "Tiny noise, big mistakes: Adversarial perturbations induce errors in brain-computer interface spellers," *National Science Review*, 2020, in press.

[9] A. Riccio, L. Simione, F. Schettini, A. Pizzimenti, M. Inghilleri, M. Olivetti Belardinelli, D. Mattia, and F. Cincotti, "Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis," *Frontiers in Human Neuroscience*, vol. 7, p. 732, Nov. 2013.

[10] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath, "Deep learning models for electrocardiograms are susceptible to adversarial attack," *Nature Medicine*, vol. 3, pp. 360–363, 2020.

Dongrui Wu received the BE degree in automatic control from the University of Science and Technology of China in 2003, the ME degree in electrical engineering from the National University of Singapore in 2005, and the PhD degree in electrical engineering from the University of Southern California in 2009. He is now Professor in the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China, and Deputy Director of the Key Laboratory of Image Processing and Intelligent Control, Ministry of Education. His research interests include affective computing, brain computer interfaces, computational intelligence, and machine learning. He has more than 160 publications, including a book entitled *Perceptual Computing* (Wiley-IEEE Press, 2010). He received the IEEE SMC Society Early Career Award in 2017, First Prize of the China Brain-Computer Interface Competition in 2019 and 2020, and USERN Prize in Formal Sciences in 2020.