Adversary on Multimodal BCI-based Classification*

Chetan Kumar[1], James Patrick Donohue[1], Rohan Gonjari[1], Neela Rahimi[1], John McLinden[2], Yalda Shahriari[2] and Ming Shao[1]
[1]University of Massachusetts Dartmouth, MA, USA, [2]University of Rhode Island, RI, USA

I. Introduction

Neural networks (NN) has been adopted by brain-computer interfaces (BCI) to encode brain signals acquired using electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS). However, it has been found that NN models are vulnerable to adversarial examples, i.e., corrupted samples with imperceptible noise. Once attacked, it could impact medical diagnosis and patients' quality of life. While early work focuses on interference using external devices at the time of signal acquisition, recent research shifts to collected signals, features, and learning models under various attack modes (e.g., white-, grey-, and black-box) [1]. However, existing work only considers single-modality attacks and ignores the topological relationships among different observations, e.g., samples having strong similarities. Different from previous approaches, we introduce graph neural networks (GNN) to multimodal BCI-based classification and explore its performance and robustness against adversarial attacks. This study will evaluate the robustness of NN models with and without graph knowledge on both single and multimodal data.

II. Dataset and Methodology

Dataset: EEG and fNIRS data of nine amyotrophic lateral sclerosis (ALS) subjects and nine healthy controls (HC) were used in experiments and written consent was obtained from all the subjects [2]. In total, there are 252 observations for each group. Signals of each observation were converted to feature vectors with a fixed size. The training/test split is set to 70:30 in all experiments.

Classification: The classification model is implemented through GNN [3] and a 3-layer NN, to account for learning models with and without graph. Graph has been playing increasingly important roles in patient networks to identify distinct relationships within/between patient groups to assist representation learning. Note we construct a k-nearest-neighbor (KNN) graph between observations (nodes) as there is no built-in graph for the ALS dataset. Given the dataset $X \in \mathbb{R}^{N\times d}$ and KNN graph $G \in \mathbb{R}^{N\times N}$, where N is the number of observations and d is the features size, we learn a GNN model f_G and a plain NN model f_{NN} for classification.

Adversarial Attack: We consider different attack strategies for f_{NN} and f_G. For f_{NN}, we generate the adversarial features X' by fast gradient sign algorithm (FGSM) [4] where $x' = x + \epsilon sign\nabla_x$ and ϵ controls the magnitude of the attack. For f_G, we apply Meta-attack [5] to generate perturbed graph G', and ensure changes on graph are bounded by $\|G - G'\|_0 \leq \epsilon$ where $\|\cdot\|_0$ is ℓ_0 norm. As a comparison, we also include a Gaussian noise baseline with varied standard deviations.

* This work is supported in part by the NSF under Grant No. 2050972.

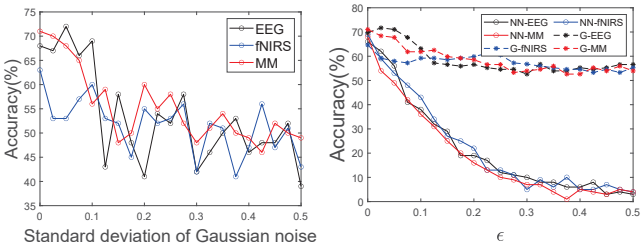

Fig. 1: Left: Gaussian noise. Right: FGSM on f_{NN} and Meta-attack on f_G. MM for "multimodality" and G for "GNN."

III. Results and Discussion

Experiments show that the NN model and its feature are most vulnerable to FGSM attack, for both single and multimodal data. This can be identified from comparisons with baseline Gaussian noise in the left figure and GNN model in the right figure. Attacks on GNN model, however, are less significant given the same attack magnitude ϵ. Note when $\epsilon = 0.5$, 400 edges in G have been flipped after attacks. In addition, multimodal BCI data did not strengthen the robustness of the model, while they empirically provided better performance in BCI-based classification, compared to single modal data. We believe one of the reasons is the preference for high-dimensional data by the adversarial attack. To develop defense strategies, researchers may focus on effective adversarial training, feature squeezing, and defensive distillation and most importantly, interpretable models should be learned to identify and understand the cause of adversarial examples. Overall, the research community must work towards developing robust and secure BCI systems for accurate diagnoses and safe treatments.

References

[1] X. Zhang and D. Wu, "On the vulnerability of cnn classifiers in eeg-based bcis," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 27, no. 5, pp. 814–825, 2019.
[2] R. J. Deligani, S. B. Borgheai, J. McLinden, and Y. Shahriari, "Multimodal fusion of eeg-fnirs: a mutual information-based hybrid classification framework," *Biomedical optics express*, vol. 12, no. 3, pp. 1635–1650, 2021.
[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
[5] D. Zügner and S. Günnemann, "Adversarial attacks on graph neural networks via meta learning," in *International Conference on Learning Representations (ICLR)*, 2019.