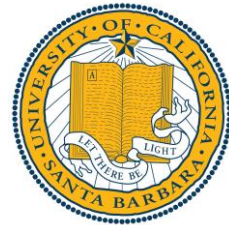


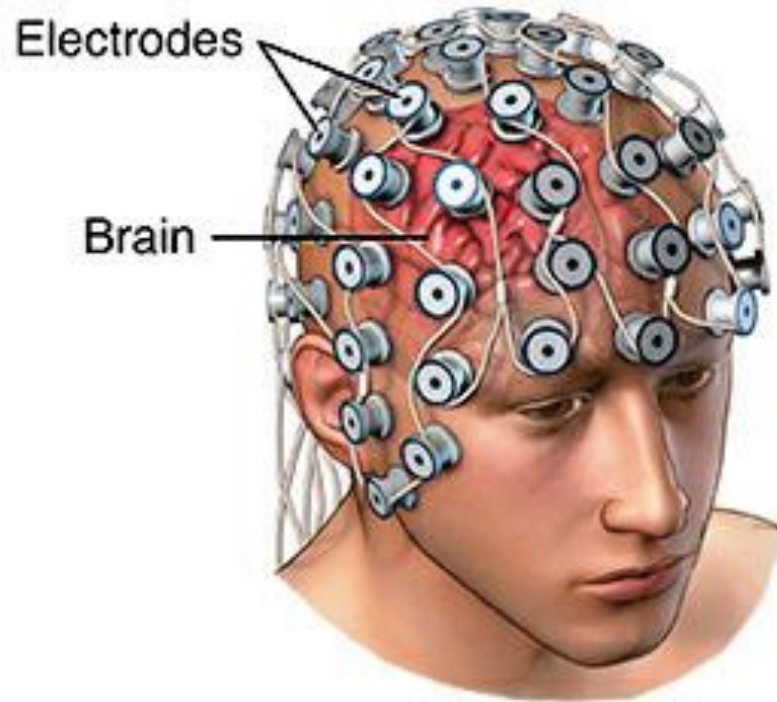
SAGA: Sparse Adversarial Attack on EEG-Based Brain Computer Interface

Boyuan Feng, Yuke Wang, Yufei Ding

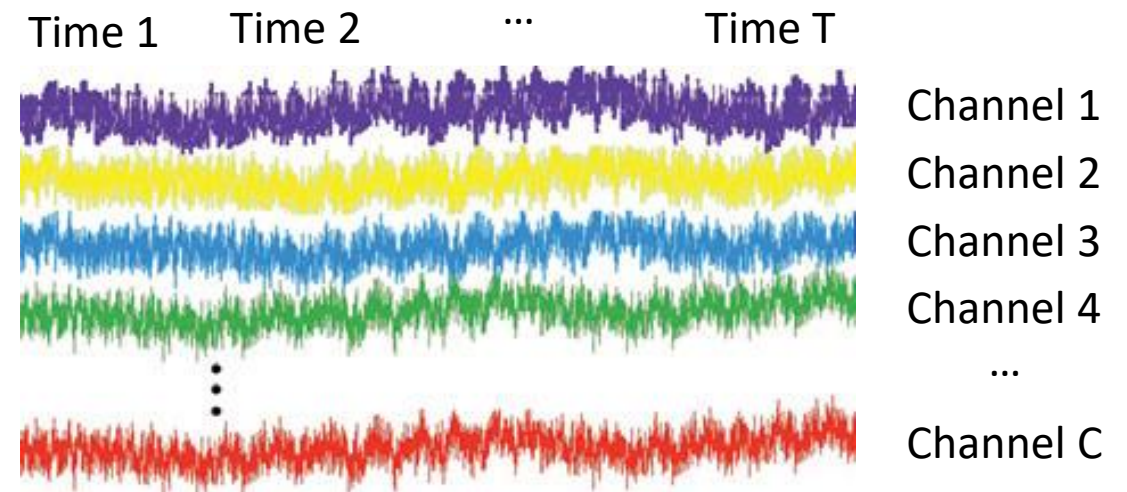


UC SANTA BARBARA

Background



Electroencephalography (EEG)



Sensory Data

Motivation

- Important workload with wide applications
 - Emotion recognition, intention recognition, etc.
- Important to understand robustness of EEG-based BCI
 - Impact of small perturbation on the EEG data on the outputs

Related Work

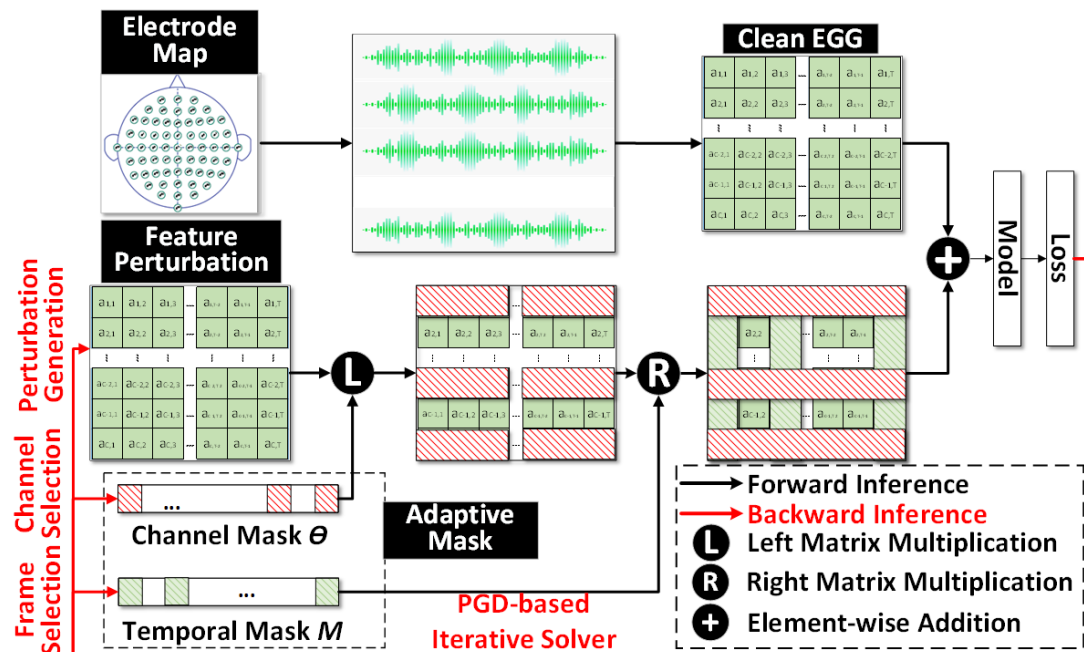
- Initial study [10]
 - Results
 - EEG-based BCI is vulnerable to FGSM and iFGSM
 - **Assumption**
 - Strong attack capability that perturbs all channels and all time steps
- We study sparse adversarial attack on EEG analytics
 - **Our assumption**
 - Weak attack capability that only sparsely attacks a small portion of channels and time steps
 - E.g., 3 out of 64 channels (electrodes) fail at a small portion of time steps

Challenges

- Need a uniform optimization framework
 - To formulate the attacking capability under various sparsity constraints
- Need a sophisticated method
 - To solve the optimization problem under given sparsity constraints

Overview of SAGA

- **Adaptive mask** on the time dimension and the channel dimension
 - Uniformly encodes diverse sparsity constraints
- **PGD-based iterative solver**
 - Effectively generate adversarial examples while satisfying sparsity constraints



Adaptive Mask

- Channel Mask: $\theta \in [0,1]^C$
- Temporal Mask: $M \in [0,1]^T$
- Sparsity cardinality constraints:
 - $card(\theta) \leq \epsilon_\theta$
 - $card(M) \leq \epsilon_M$
- Attack Formulation
 - $\min \quad -\ell(X + diag(\theta) \cdot X_\delta \cdot diag(M), Y)$
s.t. $card(\theta) \leq \epsilon_\theta, card(M) \leq \epsilon_M$

Iterative Attack

Algorithm 1: Iterative Attack to solve Problem 6.

```
1 Input: Given  $X$ , fixed weight  $W$ , learning rate  $\eta_k$ , and  
   iteration number  $K$   
2 Randomly initialize  $\theta^{(1)}$ ,  $M^{(1)}$ , and  $X_\delta^{(1)}$ .  
3 for  $k = 1, 2, \dots, K$  do  
4   | Channel Selection on  $\theta$ :  
5   |  $\theta^{(k+1)} = \Pi_{S_1}[\theta^{(k)} + \eta_k g_\theta^{(k)}]$  with Eq 8.  
6   | Frame Selection on  $M$ :  
7   |  $M^{(k+1)} = \Pi_{S_2}[M^{(k)} + \eta_k g_M^{(k)}]$  with Eq 9.  
8   | Perturbation Generation on  $X_\delta$ :  
9   |  $X_\delta^{(k+1)} = X_\delta^{(k)} + \eta_k g_X^{(k)}$  with Eq 10.  
10 end  
11 Return channel mask  $\theta^{(K)}$ , frame mask  $M^{(K)}$ , and feature  
   perturbation  $X_\delta^{(K)}$ .
```

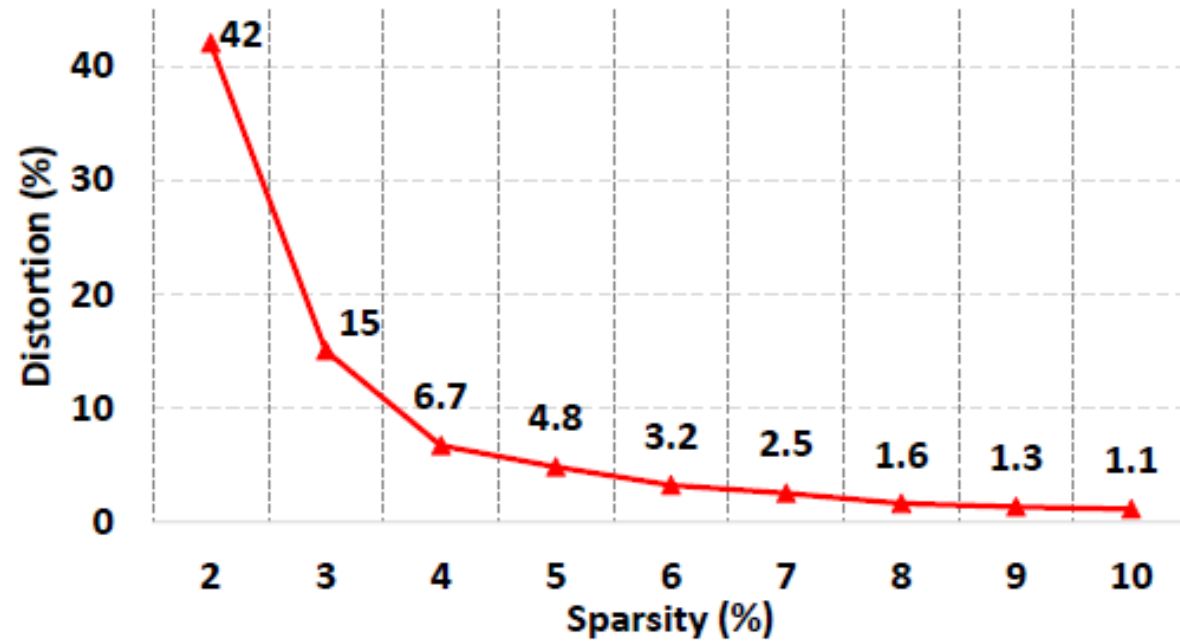
$$S_1 = \{\theta \in [0,1]^C \mid \text{card}(\theta) \leq \epsilon_\theta\}$$
$$S_2 = \{M \in [0,1]^T \mid \text{card}(M) \leq \epsilon_M\}$$

SAGA Performance

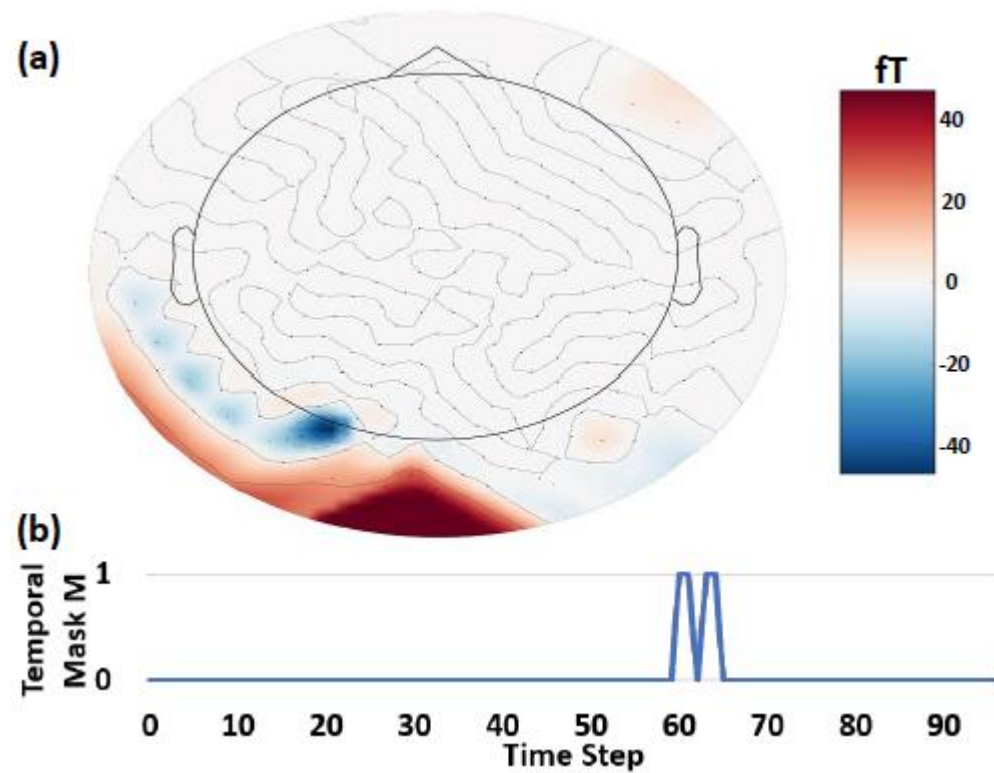
Dataset	Method	EEGNet		DeepConv	
		Acc. (%)	Norm	Acc. (%)	Norm
SPM	Clean	94.12	-	85.29	-
	FGSM	70.59	112.9	61.76	112.9
	iFGSM	70.59	112.8	58.82	112.7
	SAGA	0	112.1	0	111.2
MI	Clean	99.4	-	66.67	-
	FGSM	88.89	164.3	66.67	164.3
	iFGSM	88.89	161.3	55.56	133.3
	SAGA	11.11	155.2	11.11	132.3
ERP	Clean	91.67	-	72.22	-
	FGSM	54.17	1.94	63.89	1.94
	iFGSM	54.17	1.89	63.89	1.89
	SAGA	16.67	1.86	8.33	1.84

Overall performance under 5% sparsity constraints

Minimal Distortion



Visualization



(a) Channel Mask; (b) Temporal Mask