Assessing robustness to adversarial attacks in attention-based networks: Case of EEG-based motor imagery classification

Nour El Houda Sayah Ben Aissa a, Ahmed Korichi a, Abderrahmane Lakas b, Chaker Abdelaziz Kerrache c,*, Carlos T. Calafate d

a LINATI Laboratory, Department of Computer Science and Information Technology, Faculty of New Technologies of Information and Communication, University of Kordi Merbah, Ouargla 30000, Algeria
b College of Information Technology (CIT), Emirates Centre for Mobility Research, United Arab Emirates University, United Arab Emirates
c Laboratoire d’Informatique et de Mathématiques, Université Amar Telidji de Laghouat, Algeria
d Computer Engineering Department (DISCA), Universitat Politècnica de València, 46022 Valencia, Spain

A R T I C L E I N F O

Keywords:
Brain–computer interfaces (BCI) EEG
Electroencephalography (EEG)
Classification
Attention based networks
Adversarial attacks

A B S T R A C T

The classification of motor imagery (MI) using Electroencephalography (EEG) plays a pivotal role in facilitating communication for individuals with physical limitations through Brain-Computer Interface (BCI) systems. Recent strides in Attention-Based Networks (ATN) have showcased remarkable performance in EEG signal classification, presenting a promising alternative to conventional Convolutional Neural Networks (CNNs). However, while CNNs have been extensively analyzed for their resilience against adversarial attacks, the susceptibility of ATNs in comparable scenarios remains largely unexplored. This paper aims to fill this gap by investigating the robustness of ATNs in adversarial contexts. We propose a high-performing attention-based deep learning model specifically designed for classifying Motor Imagery (MI) brain signals extracted from EEG data. Subsequently, we conduct a thorough series of experiments to assess various attack strategies targeting ATNs employed in EEG-based BCI tasks. Our analysis utilizes the widely recognized BCI Competition 2a dataset to demonstrate the effectiveness of attention mechanisms in BCI endeavors. Despite achieving commendable classification results in terms of accuracy (87.15%) and kappa score (0.8287), our findings reveal the vulnerability of attention-based models to adversarial manipulations (accuracy: 9.07%, kappa score: -0.21), highlighting the imperative for bolstering the robustness of attention architectures for EEG classification tasks.

1. Introduction

The emergence of electroencephalogram (EEG) technology has facilitated direct interaction with brain activity. Human brains manifest diverse responses to visual stimuli, as recorded by EEG, and this has led to the development of applications in various fields like Neuroscience [1], Biometrics [2], neuromarketing [3], etc. Among the different paradigms of EEG, motor imagery (MI) signals have garnered interest and represent the most frequently applied paradigms. Since MI emerges as a reaction to cognitive tasks, EEG-based MI holds varied applications, including medical rehabilitation, entertainment, and communication avenues for individuals facing motor or speech limitations [4]. In fact, they prove to be beneficial for patients with paralysis or limb amputation, aiding in the control of prosthetic limbs, wheelchairs, or even speech restoration. They also have the potential to improve gaming accessibility, enabling individuals with physical disabilities to engage in gaming experiences previously beyond their reach. However, Brain Computer interface (BCI) technology faces challenges like a poor signal-to-noise ratio, a long user training period, and hardware limitations. Decoding brain signals accurately is challenging due to noise and substantial variability in data, especially in non-invasive BCIs. Many BCIs require user training, and ongoing challenges include miniaturization and improvements in electrode technology.

BCIs have undergone significant evolution through the integration of advanced computational techniques with deep learning, playing a crucial role. Inspired by the human brain’s structure, Deep learning (DL) utilizes multi-layered artificial neural networks to automatically extract features from raw data. This approach has surpassed traditional machine learning methods, which relied on manual feature extraction. DL’s feature extraction from raw neural data minimizes the need for manual intervention, showcasing enhanced performance in tasks such as...
as motor imagery classification. Despite its advantages, DL models face challenges like the need for substantial data, as well as potential overfitting due to the high dimensionality of BCI data. Moreover, these models are vulnerable to adversarial attacks, where subtle manipulations of the input data can mislead the model’s predictions [5].

The Attention mechanism, introduced in the groundbreaking paper ‘Attention Is All You Need’ by Vaswani et al. in 2017 [6], has revolutionized natural language processing (NLP), and has extended its impact to various machine learning tasks beyond NLP, including vision and BCI. Although initially designed for NLP, attention-based networks (ATNs) have demonstrated great potential in other domains, including BCI. The success of the initial model has led to the development of numerous variants tailored for different tasks, such as BERT [7], GPT [8], and Vision Transformers [9].

While the transformative impact of the ATN architecture on machine learning and BCI applications is evident, it becomes crucial to assess its robustness against adversarial attacks. To address such issue, this paper adopts the recent attention-based model, ATCNet [10], to test its robustness in decoding MI-EEG brain signals. The evaluation focuses on examining the model’s performance and reliability, both in the absence and the presence of adversarial attacks, thereby contributing with significant insights into its applicability and effectiveness in practical scenarios.

The main contributions of this study can be summarized as follows:

• We introduce a high-performance attention-based deep learning model to classify EEG-based MI brain signals, which outperforms state-of-the-art models.
• We design a series of experiments aimed at providing a comprehensive evaluation of attack strategies on EEG-based attention mechanism.

The remainder of this paper is organized as follows. In Section 2 we provide an overview of the latest studies on the application of DL models in EEG based classification tasks, along with the impact of adversarial attacks on these models. Section 3 presents the foundational aspects of our research, while Sections 4 and 5 provide the details of the experimental setup, and the performance results obtained from our investigations. Finally, we provide concluding remarks in Section 6.

2. Related works

This section provides an overview of recent developments in DL architectures for EEG task classification 2.1, and sheds light on the emerging challenge of adversarial attacks 2.2. Through an exploration of DL models and vulnerability assessments, this review underscores the critical need for robust and secure BCI systems in practical applications.

2.1. DL in EEG based BCI system

In recent years, the field of EEG task classification has witnessed a significant surge in studies employing DL methods. A multitude of DL architectures, including CNNs [11], Recurrent Neural Networks (RNNs) [12], Deep Belief Networks (DBNs) [13], and autoencoders (AEs) [14], have been proposed for this purpose. Among these architectures, CNNs have emerged as the most prevalent choice for MI classification, with various models such as inception-based [15], residual-based [16], and attention-based [10] CNNs being developed. A recent innovation in DL architecture is the Temporal Convolutional Network (TCN), designed specifically for time series modeling and classification. TCN has demonstrated superior performance compared to traditional RNNs, like LSTM and GRU, in sequence-related tasks, and has been applied to classify MI tasks [17]. Efforts have been made to enhance TCN’s performance through techniques such as feature fusion [18]. Moreover, recent research explored the integration of scientific machine learning (SciML) and attention mechanisms with the TCN architecture. SciML, which combines machine learning and scientific computing, aims to develop domain-aware ML models tailored for scientific data analysis [19].

Several studies have employed attention-based DL models for MI-EEG signal classification, yielding promising results. However, despite these advancements, a critical gap in the literature still remains, as existing models have yet to address the issue of robustness against adversarial attacks, which is crucial for real-world applications. Adversarial attacks pose a significant threat to the reliability and security of MI-EEG signal classification systems, highlighting the need for further research in this area to ensure the effectiveness and integrity of these models in practical settings.

2.2. Adversarial attacks in the EEG-based BCI system

Recent research has shown that EEG-based BCIs are vulnerable to adversarial attacks. These attacks include intentionally introducing small disturbances into EEG data, resulting in undesired outcomes such as misclassification, along with a decrease in the accuracy of the model. Zang et al. [20] introduced the first work which showcased the efficacy of an unsupervised fast gradient sign method (UFGSM) by implementing it on three widely used CNN classifiers for BCIs. Authors in [21] have shown that adversarial perturbations may easily deceive machine learning models, including those that depend on P300 and steady-state visual-evoked potential BCI spellers. The study demonstrated that adversarial perturbations may lead to significant misdiagnosis in medical applications, posing a significant security threat. Meng et al. [22] introduced a backdoor attack approach employing narrow period pulses, achieving high success rates with minimal perturbations. Simultaneously, the study [23] proposed a sparse adversarial attack method tailored for EEG-based BCIs, producing adversarial examples characterized by low-amplitude perturbations and high success rates. Lastly, Yu et al. [24] investigated how adversarial attacks may compromise deep learning models used to diagnose epilepsy, underscoring how subtle perturbations in EEG or BEAMs could lead to misdiagnoses, thereby highlighting safety concerns associated with deploying such models in epilepsy diagnosis (see Table 1).

3. Methodology

Our research addresses the crucial question of identifying the best architecture that ensures both security and accuracy for BCIs. By delving into the robustness of attention-based networks against adversarial attacks, and exploring their performance in EEG-based motor imagery decoding, we contribute to the understanding of how to design BCI systems that are not only accurate in their decoding capabilities, but also resilient against potential adversarial threats. Hence, in this section, we focus on the foundational aspects of our research, beginning with an overview of the ATCNet architecture, a pivotal component in our exploration. We then transition to discussing key adversarial attack models, which pose challenges to the robustness and reliability of neural networks. Understanding these components lays the groundwork for our investigation into enhancing the resilience of MI-EEG decoding systems against adversarial perturbations.

3.1. ATCNet architecture

The choice of the Attention-Based Temporal Convolutional Network (ATCNet) [10] as the central point of our research stems from its unique and powerful features tailored to address the challenges inherent in decoding brain signals. First, ATCNet stands out for its ability to effectively capture and process temporal dependencies within MI-EEG data. The model’s temporal convolutional layers are effective at extracting high-level temporal features, making it well-suited for dealing with the time-sensitive nature of EEG signals associated with motor imagery. This capacity is crucial in ensuring that the intricate
patterns and nuances in the MI-EEG data are adequately represented and leveraged for accurate classification. The incorporation of an attention layer in ATCNet adds another layer of power to the model. By focusing on the most relevant information within the temporal sequence, the attention mechanism enables the network to prioritize critical features and discard noise. This not only contributes to the robustness of the model, but also enhances its interpretability, allowing for a more transparent understanding of the decision-making process. Moreover, the utilization of multi-head self-attention in ATCNet is a strategic choice aimed at capturing the diverse aspects of the temporal sequence simultaneously. This enables the model to consider various perspectives when analyzing MI-EEG data, leading to a more comprehensive and nuanced representation. In the context of MI classification, where different EEG patterns may hold varying degrees of significance, this multi-head self-attention mechanism proves invaluable.

The convolutional-based sliding window technique further reinforces the model’s capability to capture intricate temporal patterns, especially in scenarios where the timing of neural events is crucial. This approach not only contributes to the model’s accuracy, but also ensures that it remains adaptive to the dynamic nature of MI-EEG data. The main architecture of the EEG ATCNet Model can be seen in Fig. 1. For more details about the ATCNET model, we refer the reader to [10].

### 3.2. Generating adversarial examples

Adversarial attacks on BCI represent a growing concern as BCIs become more prevalent in various applications including healthcare, gaming, and communication. Adversarial attacks typically involve the manipulation of input data to deceive machine learning algorithms, leading to incorrect classifications or behaviors (see Fig. 2). In the context of BCIs, these attacks can have serious implications, including compromising the integrity and security of neural data, and potentially causing harm to users (see Fig. 3). Algorithm 1 serves as a systematic procedure for crafting adversarial examples, a crucial task in assessing the robustness of machine learning models. At its core, the algorithm takes as input the original data $X$, the true label $Y_{true}$, the model $M$ to be attacked, and parameters governing the attack process, such as the maximum allowable perturbation magnitude ($\epsilon$), and the maximum number of iterations.

To begin, the algorithm initializes an adversarial example $X_{adv}$, initially identical to the original input $X$. It then computes the gradient of the model’s loss function with respect to $X_{adv}$, employing backpropagation to determine the direction in which the input should be perturbed to maximize loss and cause misclassification. Subsequently, based on the calculated gradient, the algorithm generates adversarial perturbations using methods like the Fast Gradient Sign Method (FGSM), the Basic Iterative Method (BIM), the Projected Gradient Descent (PGD) method, or the Carlini & Wagner (C & W) attack. These methods iteratively adjust the input data to induce misclassification while adhering to constraints such as the maximum allowable perturbation magnitude. The process repeats for a specified number of iterations, or until a predefined stop criterion is met. This criterion could include achieving a target label confidence, or reaching a maximum perturbation threshold. After termination, the generated adversarial example $X_{adv}$ is used to evaluate the model’s robustness by testing its susceptibility to adversarial attacks.

### 3.3. Adversarial attacks

We now provide a concise overview of key techniques used in our work for generating adversarial examples.

#### 3.3.1. Fast gradient sign method (FGSM)

FGSM [26] is a widely used non-targeted adversarial attack model. It involves perturbing input data by a small amount in the direction of the gradient of the loss function with respect to the input. Mathematically, the adversarial example $X_{adv}$ is generated as follows:

$$X_{adv} = X + \epsilon \cdot \text{sign}(\nabla_{X} J(X, Y_{true}))$$

(1)

where $X$ is the original input, $\epsilon$ is a small constant controlling the perturbation magnitude, $J$ is the loss function, and $Y_{true}$ is the true label.
3.3.2. Basic iterative method (BIM)

BIM [27] is a non-targeted attack, being an iterative extension of FGSM. It performs multiple small perturbations in the direction of the gradient over several iterations. The adversarial example at each iteration is given by:

$$X_{\text{adv}}^{(i+1)} = \text{clip}(X_{\text{adv}}^{(i)} + \alpha \cdot \text{sign}(\nabla_X J(X_{\text{adv}}^{(i)}, Y_{\text{true}})), X - \epsilon, X + \epsilon)$$

where $X_{\text{adv}}^{(i)}$ is the adversarial example at the $i$th iteration, $\alpha$ is the step size, and clip($\cdot$) ensures that the perturbed data stays within the $\epsilon$-ball around the original input.

3.3.3. Projected gradient descent (PGD)

PGD [28] is a non-targeted attack, being an iterative adversarial attack method designed to find adversarial examples within a specified $\epsilon$-ball. It combines gradient descent with a projection step to ensure that the perturbed data remains within the allowable perturbation range. The optimization problem is formulated as:

$$\arg \min_x J(x + \delta, y_{\text{true}}) \text{ subject to } \|\delta\|_\infty \leq \epsilon$$

3.3.4. Carlini & Wagner (C & W)

The C&W attack [29] is a non-targeted, representing an optimization-based method which aims to find the minimum perturbation for misclassification. This approach considers various attack objectives, making it a powerful and versatile model. The algorithm introduces an auxiliary variable $w$, defines $x_0 = \frac{2}{\tanh(w) + 1}$, and formulates the optimization problem:

$$\min_w \left\| \frac{1}{2} \left( \tanh(w) + 1 \right) - x_0 \right\|_2 + \epsilon \cdot f \left( \frac{1}{2} \left( \tanh(w) + 1 \right) \right)$$

where $\epsilon$ is a parameter found via dichotomy. The objective function $f(x_0)$ maximizes the gap between non-target class scores and the target class score, controlled by $k$ for confidence. Higher values of $k$ increase the error probability, but also the computation involved. Hence, the choice of $k$ balances accuracy and computational complexity.
4. Experimental setup

In this section, we detail the experimental framework, encompassing the dataset, training process with hyperparameters, and evaluation metrics. Additionally, we provide information about the platform and hardware accelerator used for our experiments.

4.1. Dataset

The BCI Competition IV 2a dataset [30], established in 2008, consists of motor imagery samples across four classes: left hand, right hand, feet, and tongue movements, collected from nine subjects. Electroencephalographic data, recorded with 22 Ag/AgCl electrodes at 250 Hz, underwent bandpass filtering (0.5 to 100 Hz), and 50 Hz notch filtering for electrical interference removal. Each trial has a duration of four seconds. Dataset characteristics are: $T = 1125$ time points, $C = 22$ EEG channels, $n = 4$ motor imagery classes, and $m = 5184$ motor imagery trials.

4.2. Signal preprocessing and data preparation

In this study, we processed raw EEG signals related to MI without any preprocessing, utilizing the complete frequency band and all EEG channels. Each MI trial, denoted as $X_i \in \mathbb{R}^{C \times T}$, was associated with a class label $y_i$ using a dataset $S = \{(X_i, y_i)\}_{i=1}^{m}$, where $y_i$ belongs to $\{1, \ldots, n\}$, and $n$ represents the total classes. The BCI-2a dataset [30] used in this context has the following characteristics: $T = 1125$ time points, $C = 22$ EEG channels, $n = 4$ motor imagery classes, and $m = 5184$ motor imagery trials.

4.3. Training procedure

The training procedure outlined in this paper adopts a subject-dependent evaluation approach, consistent with the original competition’s setup, and all models are trained and tested on the Kaggle platform using a GPU accelerator (GPU X 4) to enhance computational efficiency, Python3 as the programming language of choice, and Tensorflow with Keras API as the main library. The configurations entail initializing weights with a Glorot uniform initializer, employing the Adam optimizer with a fixed learning rate of 0.0009 for optimization, utilizing a batch size of 64 for mini-batch training, applying the categorical cross-entropy loss function, and conducting training for 1000 epochs with early stopping patience set at 300.

4.4. Performance metrics

In this study, the performance of all models is assessed using two key metrics: accuracy (ACC) and the Kappa score ($\kappa$ score). Accuracy is computed as the ratio of the sum of true positives ($TP_i$) for all classes to the total number of samples ($n$), where $TP_i$ represents the number of correctly predicted samples in class $i$. Mathematically, it is expressed as:

$$ACC = \frac{\sum_{i=1}^{n} TP_i}{n}$$  \hspace{1cm} (5)

Meanwhile, the Kappa score evaluates the agreement between observed and expected classifications, considering chance agreement. It is calculated as the average of the difference between the actual percentage of agreement ($P_i$) and the expected percentage chance of agreement ($P_e$), normalized by the total number of classes ($n$). The equation for the Kappa score is:

$$\kappa_{score} = \frac{1}{n} \sum_{i=1}^{n} \frac{P_i - P_e}{1 - P_e}$$  \hspace{1cm} (6)

Here, $n$ denotes the number of classes, $P_i$ represents the actual percentage of agreement, and $P_e$ refers to the expected percentage chance.
of agreement. These metrics collectively provide a comprehensive assessment of the models’ performance by considering both accuracy and the agreement beyond chance, thereby facilitating a robust evaluation framework.

5. Experimental results

In this section, we present classification results for four reproduced models: ATCNET, EEGNet, EEG TCNet, and TCNet Fusion. Then, we present the evaluation of the ATCNET baseline model’s robustness against adversarial attacks. Using FGSM, C&W, PGD, and BIM methods with an epsilon value of 0.01, we analyze the model’s classification performance in terms of accuracy and kappa coefficients. Confusion matrices illustrate the impact of these attacks on the model’s reliability.

5.1. Contrast models

In order to validate the effectiveness of ATCNET model, we conducted experiments on 3 alternative baseline models:

- **EEGNet** is a compact CNN designed specifically for EEG-based BCI tasks. It comprises three one-dimensional convolutional layers optimized for processing EEG signals. Despite its simplicity, EEGNet demonstrates robustness in learning various properties relevant to a wide range of BCI tasks, making it a popular choice in the field of EEG-based machine learning and neuroscience research.

- **EEG-TCNet** is a model structured to learn frequency filters from a two-dimensional temporal convolutional layer, and subsequently employs deep convolutional layers to capture spatial filters specific to certain frequencies within EEG signals. EEG-TCNet also utilizes separable convolutional layers to extract temporal summaries of each feature map, and incorporates a Temporal Convolutional Network (TCN) to further leverage temporal information in EEG data processing.

- **TCNet fusion** is a fixed hyperparameter-based CNN model that utilizes multiple techniques, such as temporal convolutional networks (TCNs), separable convolution, depth-wise convolution, and the fusion of layers.

5.2. Classification in the absence of an attack

Before delving into the examination of how adversarial attacks affect the EEG ATCNET classification model, we first assess the classification performance of ATCNET model in the absence of any attack with well known EEG based models. Table 2 and confusion matrix 4 outline
5.3. Classification in the presence of the adversarial attacks

We investigate the performance of the ATCNet baseline model under various adversarial attacks, where epsilon (\(\epsilon\)) is set to 0.01. Fig. 7 summarizes the classification results obtained through different attacks, including FGSM, C&W, PGD, and BIM. The accuracy and kappa metrics for each attack scenario are provided, highlighting the model’s response to adversarial perturbations. It is evident from the table that, under adversarial conditions, the model’s classification performance experiences notable variations compared to its baseline accuracy. For instance, the FGSM attack notably decreases its accuracy, with values ranging from 0.2448 to 0.8490, and kappa coefficients fluctuating between -0.0069 and 0.7986. The C&W attack showcases the most severe impact, with accuracy as low as 0.0174, and consistently negative kappa coefficients, indicating poor agreement with ground truth labels. PGD and BIM attacks exhibit varying impacts, both reducing accuracy and kappa coefficients across subsets. These findings emphasize the need to assess the robustness of the ATCNet model in real-world scenarios, where adversarial attacks may compromise its reliability. Analyzing the confusion matrices for each attack given in Fig. 5 allows us to identify specific patterns of misclassifications induced by adversarial attacks, providing valuable information on the model’s vulnerabilities.

6. Conclusion

This study highlights the effectiveness of Attention-Based Networks in decoding EEG signals, particularly in the context of motor imagery. These models, specifically referred to as ATCNet in the study, showcase notable capabilities in capturing temporal dependencies, and in extracting high-level features from MI-EEG data, thus positioning themselves as valuable tools for real-time applications in various contexts. However, the study also sheds light on an important concern regarding the susceptibility of ATCNet to adversarial attacks. Adversarial attacks refer
to intentional manipulations of input data to mislead machine learning models, potentially leading to erroneous outputs. In the context of MI-EEG decoding, such vulnerabilities could have serious implications for applications relying on the accurate interpretation of brain signals.

The identification of ATCNet’s susceptibility to adversarial attacks underscores the need for developing robust defense mechanisms to protect MI-EEG decoding systems. These defense mechanisms could involve strategies such as adversarial training, which entails training models with adversarially perturbed data to enhance their resilience against such attacks.

7. Future directions

Our future research will address the following aspects:

- Integrate ATCNet with advanced adversarial defense algorithms to further improve its performance and resilience against adversarial attacks, thereby enhancing the reliability and accuracy of MI-EEG decoding systems in real-world scenarios.
- The susceptibility analysis of other machine learning models based BCIs during both training and testing phases, revealed...
significant vulnerabilities. These vulnerabilities underscore the necessity for the development of novel attack strategies tailored specifically for BCIs.

- Exploring methods to defend BCI against adversarial attacks is our primary focus. Various defense techniques, such as adversarial training, defensive distillation, and ensemble adversarial training, have been proposed for different application domains. However, the absence of a universal defense strategy persists due to the ongoing theoretical ambiguity surrounding the main causes of adversarial examples in deep learning. Goodfellow [26] attributed the existence of adversarial examples to the linearity of deep learning models, while Gilmer [31] contended that these examples arise from the high-dimensional geometry of the data manifold. Our investigation aims to uncover the fundamental reasons behind adversarial examples in EEG classification/regression, thus enabling the development of robust defense strategies for enhancing the security of BCIs.

**CRediT authorship contribution statement**


**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to bias their work. The authors have no financial relationships that might be perceived as posing a conflict of interest.

**Acknowledgments**

This paper has been funded by Emirates Centre for Mobility Research at United Arab Emirates University, Grant #:31R269.

**References**


