MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Learning Invariant Representations from EEG via Adversarial Inference

Ozdenizci, Ozan; Wang, Ye; Koike-Akino, Toshiaki; Erdogmus, Deniz

TR2020-049 April 18, 2020

Abstract

Discovering and exploiting shared, invariant neural activity in electroencephalogram (EEG) based classification tasks is of significant interest for generalizability of decoding models across subjects or EEG recording sessions. While deep neural networks are recently emerging as generic EEG feature extractors, this transfer learning aspect usually relies on the prior assumption that deep networks naturally behave as subject- (or session-) invariant EEG feature extractors. We propose a further step towards invariance of EEG deep learning frameworks in a systemic way during model training. We introduce an adversarial inference approach to learn representations that are invariant to inter-subject variabilities within a discriminative setting. We perform experimental studies using a publicly available motor imagery EEG dataset, and state-of-the-art convolutional neural network based EEG decoding models within the proposed adversarial learning framework. We present our results in cross-subject model transfer scenarios, demonstrate neurophysiological interpretations of the learned networks, and discuss potential insights offered by adversarial inference to the growing field of deep learning for EEG.

IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.DOI

Learning Invariant Representations from EEG via Adversarial Inference

OZAN ÖZDENIZCI[1], (Student Member, IEEE), YE WANG[2], (Senior Member, IEEE), TOSHIAKI KOIKE-AKINO[2], (Senior Member, IEEE), and DENIZ ERDOĞMUŞ[1], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA
[2]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA

Corresponding author: Ozan Özdenizci (oozdenizci@ece.neu.edu).

Ozan Özdenizci and Deniz Erdoğmuş are partially supported by NSF (IIS-1149570, CNS-1544895, IIS-1715858), DHHS (90RE5017-02-01), and NIH (R01DC009834).

ABSTRACT Discovering and exploiting shared, invariant neural activity in electroencephalogram (EEG) based classification tasks is of significant interest for generalizability of decoding models across subjects or EEG recording sessions. While deep neural networks are recently emerging as generic EEG feature extractors, this transfer learning aspect usually relies on the prior assumption that deep networks naturally behave as subject- (or session-) invariant EEG feature extractors. We propose a further step towards invariance of EEG deep learning frameworks in a systemic way during model training. We introduce an adversarial inference approach to learn representations that are invariant to inter-subject variabilities within a discriminative setting. We perform experimental studies using a publicly available motor imagery EEG dataset, and state-of-the-art convolutional neural network based EEG decoding models within the proposed adversarial learning framework. We present our results in cross-subject model transfer scenarios, demonstrate neurophysiological interpretations of the learned networks, and discuss potential insights offered by adversarial inference to the growing field of deep learning for EEG.

INDEX TERMS adversarial learning, brain-computer interface, deep neural networks, electroencephalogram, invariant representation, motor imagery

I. INTRODUCTION

RAPID progress of deep learning in computer vision with the emergence of large image data sets and computational resources over the last decade motivated a variety of studies exploring deep neural networks in decoding information from electroencephalographic (EEG) data [1], [2]. This interest was particularly focused on EEG-based brain-computer interface (BCI) technology which is primarily motivated by an aim to provide a neural control channel for individuals with severe neuromuscular disorders [3], [4]. Developing BCI systems mainly rely on robust decoding of user (subject) intentions from EEG, under the prior belief that EEG encodes the information on such intent. To that end, convolutional neural network (CNN) based feature extractors became powerful generic EEG signal processing tools, alleviating the need for manual feature extraction [2], [5].

One of the main challenges in EEG classification is coping with the change in data distributions across different subjects or recording sessions, well known as the problem of *transfer learning* [5–7]. Particularly in cross-subject transfer, the aim is to discover and exploit shared, invariant neural structures across subjects towards the primary goal of eliminating or reducing system calibration times for people with neuromuscular disabilities. Conventional machine learning approaches in addressing cross-subject invariance mostly focus on regularizing classifiers [8] or feature extractors [9], [10] using other subjects' data, as well as learning population level common spatial bases dictionaries [11], [12]. Such methods are shown to yield promising results when learned representations are regularized not to overfit to the subject pool. However from a deep feature learning standpoint, current approaches rely on the hypothesis that the deep, capable network architectures will internally learn robust representations (features) during training, that are generalizable across subjects and/or sessions [13–16] (cf. Section II-A for a detailed look). Nevertheless this assumption can be naturally constrained given that most neuroimaging datasets are of smaller scale than those of images or videos, which further restrains the progress of deep learning in cognitive neuroscience.

In light of recent work on invariant representation learning

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2971600, IEEE Access

with neural networks [17], [18], we present in this paper an adversarial inference approach to learn nuisance-invariant representations from EEG. Particularly, we aim to learn representations that are invariant to cross-subject variabilities within a discriminative neural network setting. We recently explored a similar idea in EEG-based biometric identification systems for inter-recording invariance with promising results [19]. Here, we hypothesize that an adversarial regularization towards learning subject-invariant representations can shed light on current EEG deep learning studies to develop EEG decoding models that systematically consider the generalizability problem. We propose our adversarial training approach independent of the EEG deep learning architecture. In experimental evaluations, using a publicly available EEG dataset, we demonstrate the impact of adversarial discriminative training on three state-of-the-art neural network architectures for EEG decoding (i.e., EEGNet [15], DeepConvNet and ShallowConvNet [14]). We further employ the layer-wise relevance propagation method [20] for the trained neural networks to investigate the neurophysiological signatures that subject-invariant models exploit. We compare our results with regards to the non-adversarially trained counterpart of each architecture, and finally discuss the benefits offered by adversarial inference to the field of deep learning for EEG.

Contributions of this paper are three-fold: (1) an adversarial inference approach to learn invariant representations for deep learning based EEG decoding models is presented, (2) implementation and evaluations of this approach for subject-invariant discriminative EEG feature learning are performed in cross-subjects model transfer scenarios, and (3) visual demonstrations of the neurophysiological interpretability of invariant representation learning models are revealed.

II. RELATED WORK

A. DEEP LEARNING IN EEG

Over the last two decades, deep neural networks have been widely explored as generic feature extractors for EEG, particularly in the context of developing brain interfaces [2]. A significant collection of work uses convolutional architectures that are capable of exploiting temporal, spectral and spatial structures from input raw EEG. Applications of such models were thoroughly studied for decoding motor imagery [14], [21], [22], visually evoked potentials (VEP), which was first demonstrated with P300 detection on two users' data [23], steady-state visually evoked potentials [24], as well as for rhythm perception from EEG during auditory stimuli [25]. In other respects, EEG is translated into different network input forms, such as topographical images [26], combinations of different spectral EEG components [16], topology-preserving multi-spectral images (i.e., EEG movies) within recurrent-CNNs [13], or frequency domain representations [27], [28]. Nevertheless, a large portion of existing works were either limited by not being generalizable to different EEG decoding problems, or being offline studies lacking demonstrations of cross-subjects generalization [5].

Recent examples of CNNs in EEG decoding introduce non-task-specific architectures for discriminative feature extraction; specifically DeepConvNet, ShallowConvNet [14], and EEGNet [15]. Further progress on assessing neurophysiological features extracted within the deep learning black-boxes made these tools more interpretable [14], [29], [30]. Yet, most studies rely on the intuition that the deeper and more capable the architecture, learned features would be less sensitive to variations across a large dataset and potentially be transferable across-subjects [13], [16]. With a similar aim in [31], transfer capability of a convolutional autoencoder was assessed by training with cross-subject validation sets, which tends to introduce a model selection bias at early validation stopping and makes the learned models inapplicable for plug-in model transfer. Also recently, across-subjects transfer capabilities of motor imagery [21], as well as VEP [32] decoding CNN models were demonstrated only by fine-tuning global parameters to reduce calibration times. Notably in [16], leave-one-out cross-subject generalizability of the proposed architecture was demonstrated successfully, however relying on the deep capability of the network with no explicit approach towards imposing subject-invariance within the model. In [33] joint adversarial training methods were used to transfer knowledge from large, annotated image databases to learn generalizable EEG feature extractors, while restricting the EEG input representations to match with image dataset CNN architectures. Recently, an end-to-end CNN for cross-subject EEG decoding using a deep domain adaptation approach was presented [34], while assuming availability of target domain data during model training which makes it hardly applicable to real-time brain interface control problems. In this respect, we highlight that existing EEG deep learning methods do not explicitly ensure inferring subject-invariant representations during discriminative model training, which is left to be explored.

B. ADVERSARIAL REPRESENTATION LEARNING

Adversarial representation learning can be viewed as simultaneously learning to predict a dependent variable from a representation, while exploiting an adaptive dependence measure between these two to also learn the representation itself such that this dependence is minimized. The history of adversarial learning methods in data science goes back as far as Schmidhuber's *principle of predictability minimization* introduced for unsupervised learning of distributed non-redundant representational units from data [35]. The principle suggests learning an adaptive predictor of each unit that uses the remaining units, while each individual unit trying to minimize its predictability. This eventually enables learning statistically independent representational units, which still combine to become descriptive. More recently with generative adversarial networks (GAN), a generative model can be learned to synthesize realistic data samples from random noise, while an adversarial classifier has the antagonistic objective of identifying real and generated data samples [36].

Progressive work of our interest focuses on using adversarial training for the latent space, instead of the output space

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2971600, IEEE Access

as in GANs, particularly to learn invariant latent representations by disentangling specific attributes (e.g., nuisance variables) from the representation. A significant amount of work tackles this problem from a generative perspective, where variational autoencoders (VAEs) are censored by constraining the encoded latent space to be invariant to specific attributes either with an invariance-enforcing kernel-based penalty term [37], or with adversarial training objectives that exploits a simultaneously learned attribute classifier loss [38]. Similar approaches to invariant representation learning are also proposed with discriminative perspectives tailored to specific prediction tasks, such as learning attribute-invariant *fair clusters* [39], or jointly training a classifier with an adversarially censored VAE to learn *fair classifiers* [40].

Considering fully-discriminative approaches that do not require learning a generative decoder counterpart, adversarial discriminative representation learning can be observed as a domain adaptation problem when the nuisance variable is binary. Assuming that the two domains (source and target) are the nuisance variables, domain-invariant predictive representations can be adversarially learned to minimize a measure of domain discrepancy. Most of the existing approaches to this problem in image processing assume target data to be available at training time [41–43], which makes them inapplicable to the problem of transferring EEG representations across subjects. From another perspective, recent work studies this as an adversarial training game that aims to maximize task-specific prediction certainty from learned representations, while minimizing the certainty of inferring the nuisance variables causing such domain shift from these representations [17], or classifier outputs [18]. Importantly, these advancements in deep, invariant, and discriminative feature learning were not particularly explored for EEG.

It is important to note that in this work we are only focusing on partially supervised cases where nuisance variables that represent variability across data are specified, while at the other extreme are fully unsupervised methods where deep networks are trained to disentangle factors of variation in data without explicitly specifying the source of variability.

III. METHODS

A. NOTATION AND PROBLEM DESCRIPTION

Let $\{(\boldsymbol{X}_i, y_i, s_i)\}_{i=1}^{n}$ denote the data set consisting of n observations coming from a data generation process with $\boldsymbol{X} \sim p(\boldsymbol{X}|y,s)$, $y \sim p(y)$, and $s \sim p(s)$, where $\boldsymbol{X}_i \in \mathbb{R}^{C\times T}$ is the raw EEG data at trial i recorded from C channels for T discretized time samples, $y_i \in \{0,1,\ldots,L-1\}$ is the corresponding condition (i.e., class) label, and $s_i \in \{1,2,\ldots,S\}$ denotes the subject identification (ID) number for the person that the trial EEG data is collected from across S subjects that our data set is constituted with. Note that for our problem of interest, the underlying but reasonable assumption here is s and y being marginally independent.

Given training data, the aim is to learn a discriminative EEG decoder model that predicts y from observations $\boldsymbol{X}$. For such a model to be generalizable across subjects, ideally the predictions should be invariant to s, which will be unknown at test time. We regard s as some nuisance parameter that is involved in the EEG data generation process, and aim to learn a parametric model which can be generalized across subjects and learns features (representations) that are invariant to s. A similar methodology was recently utilized in our previous work for session-to-session feature invariance [19].

B. ADVERSARIAL DISCRIMINATIVE MODEL TRAINING

Given the training data set, we train a deterministic *encoder* network with parameters θ_e to learn representations $h = f(\boldsymbol{X};\theta_e)$. Specifications of the encoder network are further discussed in Section IV-B. Obtained representations are used as input separately to both a *classifier* with parameters θ_c to estimate y, as well as an *adversary* network with parameters θ_a, which aims to recover the nuisance variable s. Respectively, the classifier and adversary networks are modeling the likelihoods $q_{\theta_c}(y|h)$ and $q_{\theta_a}(s|h)$. In order to filter factors of variation caused by s within h, we propose an adversarial game. The adversary is trained to predict s by maximizing the likelihood $q_{\theta_a}(s|h)$, while at the same time, the encoder is trying to conceal information regarding s that is embedded in h by minimizing that likelihood, as well as retaining sufficient discriminative information for the classifier to estimate y by maximizing $q_{\theta_c}(y|h)$. Overall, we train these networks simultaneously towards the objective:

$$\hat{\theta}_e, \hat{\theta}_c, \hat{\theta}_a = \arg\min_{\theta_e,\theta_c} \max_{\theta_a} \mathcal{L}(\theta_e, \theta_c, \theta_a), \tag{1}$$

where the loss function is denoted by:

$$\mathcal{L} = \mathbb{E}_h \mathbb{E}_y[-\log q_{\theta_c}(y|h)] + \lambda \mathbb{E}_h \mathbb{E}_s[\log q_{\theta_a}(s|h)], \tag{2}$$

with θ_e represented through $h = f(\boldsymbol{X};\theta_e)$, and a higher *adversarial regularization weight* $\lambda > 0$ enforcing stronger invariance trading-off with discriminative performance. The optimization algorithm uses stochastic gradient descent (or ascent) alternatingly for the adversary and the encoder-classifier networks to optimize Eq. (1) (see Algorithm 1). This approach is motivated by the work on adversarially learned invariant representations in discriminative model training [17], [18]. Accordingly, the theoretical foundations on the convergence of such an adversarial game was previously studied in various settings [17], [18], [43]. Note that in Algorithm 1, setting $\lambda = 0$ would indicate training a regular CNN, whereas $\lambda < 0$ would correspond to forcing the encoder to exploit subject-variant task-discriminative features, which is not expected to be favorable for transfer learning. An overview of the network is illustrated in Figure 1.

IV. EXPERIMENTAL STUDIES

We perform experiments on a publicly available EEG dataset for motor imagery decoding [44]. Particularly, motor imagery based BCI systems rely on detection of evident contralateral desynchronization of oscillatory EEG rhythms over sensorimotor areas following imagination of a movement [4].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2971600, IEEE Access

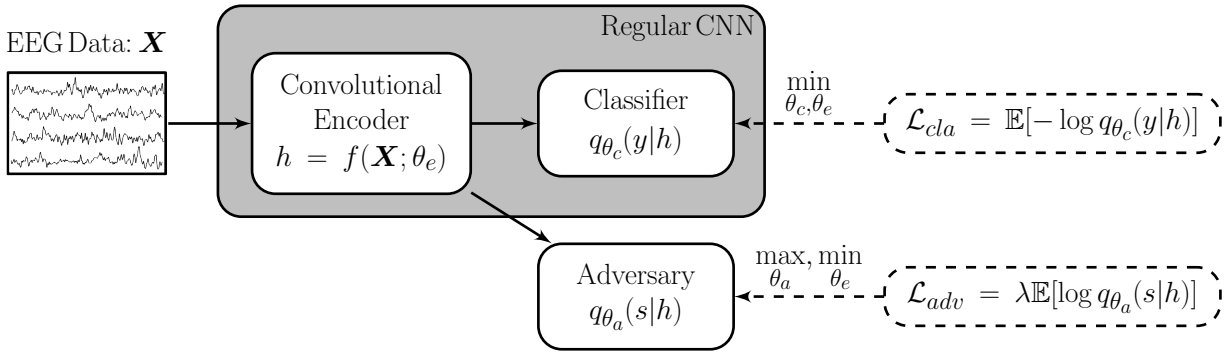


FIGURE 1. Overall adversarial model training architecture for invariant discriminative representation learning, illustrating the convolutional EEG feature encoder, adversary, and classifier networks. Networks are simultaneously trained towards the objective in Eq. (1), as illustrated by the loss functions of the classifier ($\mathcal{L}_{cla}$) and the adversary ($\mathcal{L}_{adv}$) in the dashed boxes. The gray shaded region indicates the regular CNN components with a cascaded encoder and classifier.

Algorithm 1 Adversarial discriminative model training

Input: training data $\{(\boldsymbol{X}_i, y_i, s_i)\}_{i=1}^{n}$, adv. weight $\lambda > 0$
Output: $\hat{\theta}_e, \hat{\theta}_c, \hat{\theta}_a$

Randomly initialize $\theta_e, \theta_c, \theta_a$
for $t = 1$ to $\#epochs$ **do**
  for $b = 1$ to $\#batches$ **do**
    Sample batch of M trials $\{(\boldsymbol{X}_m, y_m, s_m)\}_{m=1}^{M}$
    Update θ_a with stochastic gradient ascent by:
    $\nabla_{\theta_a} \sum_{m=1}^{M} \lambda \log q_{\theta_a}(s_m|h_m = f(\boldsymbol{X}_m;\theta_e))$
    Update θ_e, θ_c with stochastic gradient descent by:
    $\nabla_{\theta_e,\theta_c} \sum_{m=1}^{M} [-\log q_{\theta_c}(y_m|h_m) + \lambda \log q_{\theta_a}(s_m|h_m)]$
  end for
end for

A. DATASET DESCRIPTION

The original dataset [44] consisted of single-session data from 52 healthy subjects, however we discarded 4 subjects' data due to irregular timestamp alignments and unequal number of trials per class. This resulted in a set of 48 subjects' EEG data for our empirical assessments. During the experiments, subjects were sitting in front of a computer screen and were instructed to perform cue-based tasks while 64-channel EEG [45] were recorded at a sampling rate of 512 Hz. These tasks included movement imagination of the left or right hand during three second trials, for 100 trials per hand in randomized order. This resulted in a total of 200 trials per subject, with an associated binary class label (0 for left, 1 for right hand). The original dataset also included other preliminary cue-based recordings as well, which were however not part of our experimental analyses. Further specifications of the dataset can be accessed from [44].

B. NEURAL NETWORK ARCHITECTURE

Beyond design specifications of the network architecture, naturally, any discriminative representation learning network can be adversarially trained with a same approach. We demonstrate our empirical results using three state-of-the-art CNN models proposed for EEG decoding, namely the EEGNet [15], DeepConvNet and ShallowConvNet [14] architectures. Within the convolutional layers, temporal, spatial, and spatio-temporal convolutions for aggregation of neural features in h are performed. Subsequently, all three architectures have a final dense linear classification layer which we separated from the preceding encoder layers, as the classifier block. This resulted in the complete convolutional architectures except the final dense layer constructing the encoder, whereas the final dense layer constructing the classifier network. Further specifications on how the encoder architectures were implemented can be accessed in Appendix A. Parameter choices were based on the original descriptions in the manuscripts, as well as their provided software implementations online [14], [15].

Regarding the classifier and adversary blocks, we simply used the linear classification approach of the networks we inherited. The classifier utilizes h as an input to a fully-connected layer with L softmax units for task discrimination. Similarly for the adversary, h is used as input to a fully-connected layer with S softmax units for subject ID discrimination, to obtain normalized log-probabilities that will be used to calculate the cross-entropy losses in Eq. (2).

C. MODEL TRAINING AND EVALUATION

All raw EEG data was initially resampled to 128 Hz. This was performed both to save computational time, as well as to construct a common network input basis for all three architectures [14], [15]. As the EEG pre-processing steps, we common average referenced each subject's EEG data, and bandpass filtered the signals between 4 and 40 Hz with a causal third order Butterworth filter. We epoched each trial in the [0.5-2.5] seconds of post-cue time interval. No offline channel selection or artifact correction was performed. This resulted in EEG trials with dimensions of 64-channels by 256 time samples as inputs to the networks.

We evaluated adversarial and non-adversarial (regular CNN) training of each encoder network in simulated online decoding studies (i.e., in cross-subjects decoding scenarios

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2971600, IEEE Access

with direct transfer of learned models to novel subjects without subject-specific calibration or fine-tuning). We generate the *transfer set* in 6 folds (i.e., 8 of the 48 subjects were held out in turns), yielding cross-subject predictions for each subject with models that are learned from a separate group of 40 subjects. This 6-fold process was also repeated 10 times by randomly changing the 8-subject transfer set and the remaining 40-subject group folds. In total, this resulted in 10 transfer learning prediction accuracies for every subject in the dataset, with models that are learned from a different (but intersecting) group of 40 subjects. During model learning from the 40 subjects, we generate a *training set* and a *validation set* by randomly assigning 20% of the trials (i.e., 40 out of 200) from each of the 40 subjects for the validation set, and the remaining 80% of the trials for the training set. This resulted in 6400 model training set trials, and 1600 trials for the validation sets that are used by the neural network models to monitor losses of the classifier and/or adversary.

D. IMPLEMENTATION

We implemented all models in Tensorflow [46] using the Keras API [47]. Networks were trained with 40 training trials per batch for at most 500 epochs with early stopping based on the classifier loss on the validation set. Specifically, if the validation loss for class prediction did not improve (i.e., reach a new lowest value) for 10 epochs, training was stopped and the model which resulted in the lowest validation loss was saved. Parameter updates were performed once per batch with Adam [48]. For the models described in Section IV-B, and the training approach with the classifier output $L = 2$ and adversary output $S = 40$ (see Section IV-C), the number of parameters to be learned during training for EEGNet are: 1,872 for encoder, 258 for classifier, 5,160 for adversary, for DeepConvNet are: 172,150 for encoder, 4,802 for classifier, 96,040 for adversary, and for ShallowConvNet are: 103,040 for encoder, 2,402 for classifier, 48,040 for adversary. Our implementations are available at: https://github.com/oozdenizci/AdversarialEEGDecoding.

E. INTERPRETATION OF LEARNED NETWORKS

To explore the neurophysiological signatures that the networks exploit, we employ layer-wise relevance propagation (LRP) [20] as a feature interpretation method which was recently shown as a powerful approach to study interpretability of EEG deep learning models [29]. Specifically, LRP decomposes the network output score into *relevances* of each unit of the network input (i.e., pixels of the EEG data matrix $\boldsymbol{X}$), according to its contribution to the classification decision. These relevance scores for each pixel of $\boldsymbol{X}$ are then visualized as what we denote as a *feature relevance* map.

Let $R_i^{(l)}$ denote the *relevance* of neuron i in layer l. To investigate classification decisions, firstly, the neuron with the highest score at the network output layer prior to softmax activation is assigned a relevance value that is equal to its score, while all the other output layer neurons are assigned a relevance value of zero. Subsequently, layer by layer, relevances of each neuron at an upper layer $l+1$ are redistributed to the neurons at the adjacent lower layer l through a backward pass until the input layer $l = 1$ is reached, according to the following rule:

$$R_i^{(l)} = \sum_j \frac{z_{ji}}{\sum_{i'}(z_{ji'})} R_j^{(l+1)}, \tag{3}$$

where z_{ji} is the weighted activation of a neuron i at layer l onto neuron j at layer $l+1$ during the forward pass after training. In our implementations, we utilize a slight variant of the LRP framework called ϵ-LRP from the original work [20], which only differs with an additional term in the denominator to preserve numerical stability.

To investigate the feature relevances for classifier decisions of cross-subjects transferred models, the backward pass was initiated from the classifier output neuron with the highest score out of the L neurons, prior to softmax activation. Similarly, to demonstrate how the networks can exploit user-specific EEG patterns into the highest score out of the S output neurons of the adversary for user identification, we also generated feature relevance maps for the adversary decisions on the validation set after completion of model training.

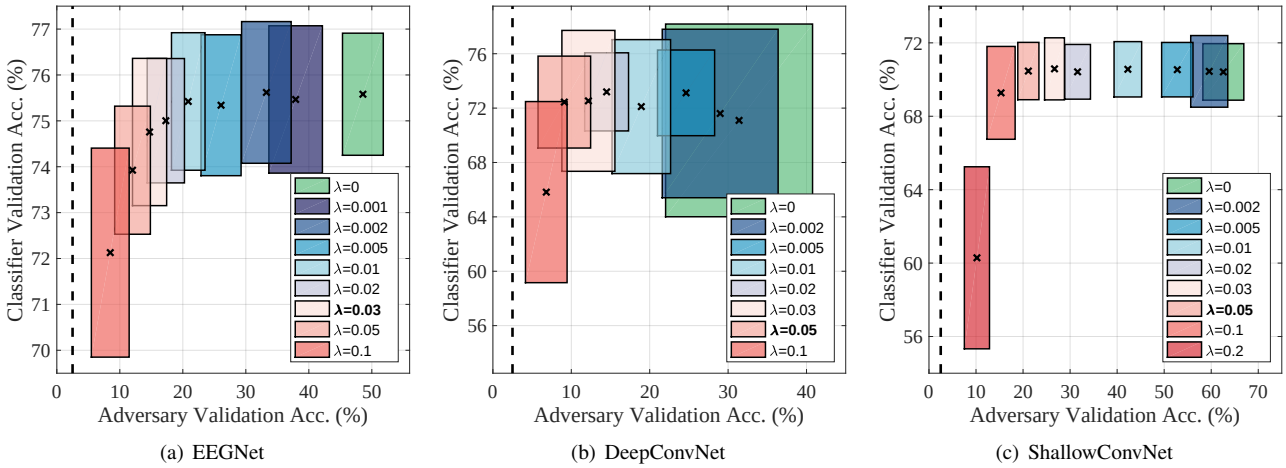
V. RESULTS

A. CHOOSING THE ADVERSARIAL REGULARIZATION WEIGHT PARAMETER

An intuitive way to choose the adversarial regularization weight λ is by cross-validation (parameter sweep). We train models with various choices of $\lambda > 0$, and favor decreases in adversary accuracy with increasing λ, while maintaining a similar classifier accuracy on the validation sets with respect to not using an adversary ($\lambda = 0$). Figure 2 demonstrates these changes by varying λ for each architecture. In this context, we define the adversary accuracy as the percentage of correctly predicted trials in subject identification by the adversary network (i.e., it is favored if this value is small), whereas the classifier accuracy is defined as the percentage of correctly predicted trials in class label discrimination by the classifier network (i.e., it is favored if this value is high).

For the non-adversarial models ($\lambda = 0$) we trained the adversary network alongside the encoder-classifier with no adversarial loss feedback, and assessed the amount of subject-discriminative information (i.e., leakage) in the encoded representations. We observe that regular CNNs can indeed learn features that exploit subject-specific information, leading to a 48.5% average adversary accuracy to discriminate 40 subjects with EEGNet, 31.4% with DeepConvNet and 62.6% with ShallowConvNet. Increasing λ censors the encoder as expected and suppresses adversary accuracies. However a very strong λ can force the encoder to lose task-discriminative information, leading to decreasing classifier accuracies on the within-subject validation sets as observed in Figure 2. Hence we determine an operating λ range where the classifier does not start to perform very poorly (i.e., similar performance as $\lambda = 0$) and adversary accuracy is low. Specifically, we proceed by choosing $\lambda = 0.03$ for EEGNet,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2971600, IEEE Access



(a) EEGNet (b) DeepConvNet (c) ShallowConvNet

FIGURE 2. Classifier versus adversary accuracies on the validation set after model training for (a) EEGNet, (b) DeepConvNet, and (c) ShallowConvNet. Vertical dashed black lines denote the chance level for adversary accuracy (i.e., 40-class subject identification). For the colored box patches, center marks denote the means across training folds and repetitions, and widths denote ± 1 standard deviation intervals in both dimensions.

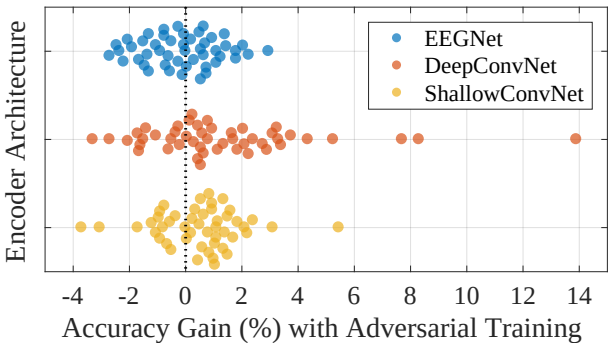


FIGURE 3. Distribution of the differences between adversarial and regular training accuracies averaged across repetitions. Each row/color indicates one encoder architecture, and each dot indicates a single subject.

and $\lambda = 0.05$ for DeepConvNet as well as ShallowConvNet, which are indicated with bold legend texts in Figure 2.

B. CROSS-SUBJECT DECODING MODEL TRANSFER

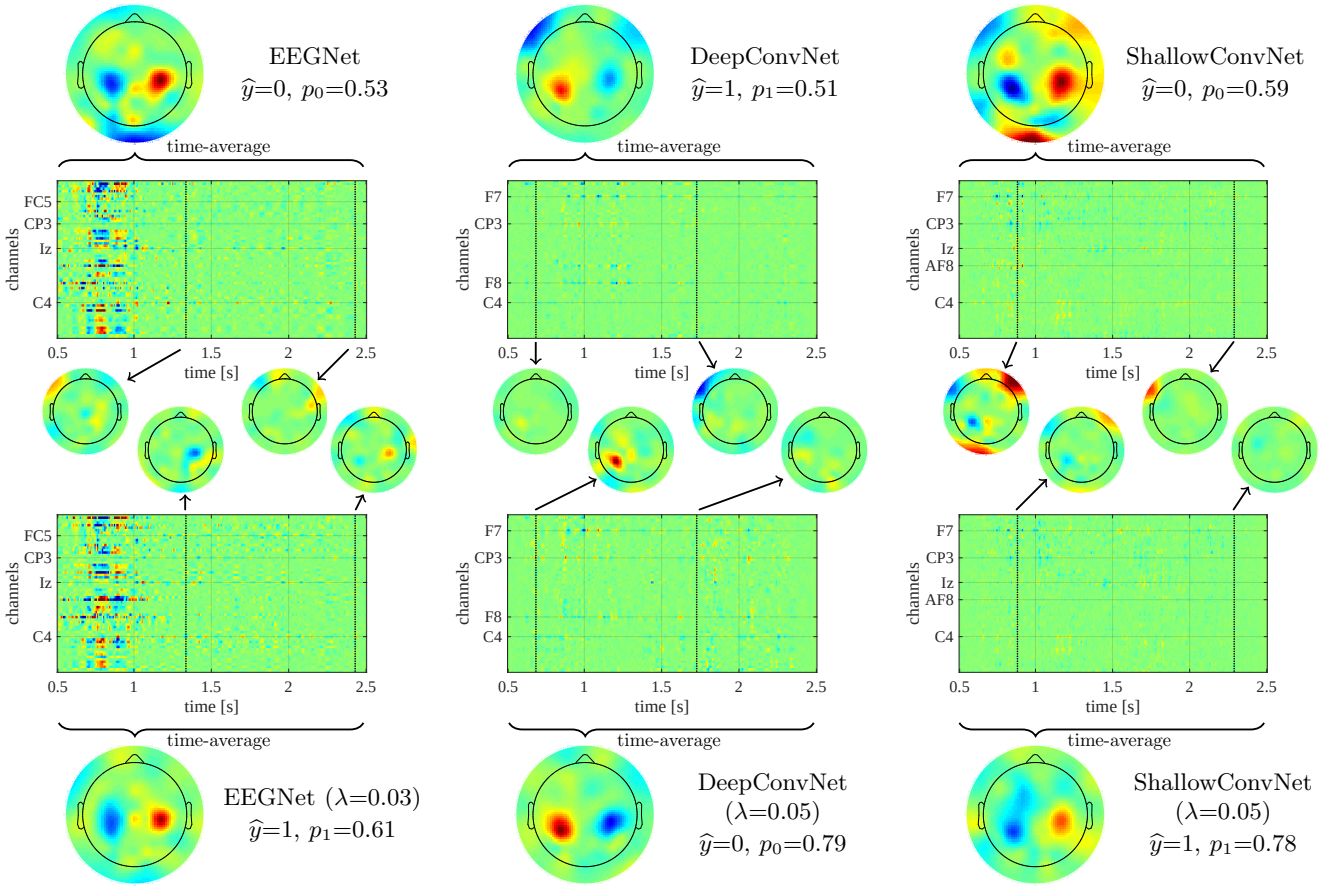
We investigate cross-subjects generalization of adversarially learned invariant representations ($\lambda > 0$), in comparison to non-adversarial CNN classifiers ($\lambda = 0$). Comparisons between the performances of adversarial versus non-adversarial learning methods were evaluated by repeated-measures Analysis of Variance (ANOVA) statistical tests for each architecture. We model the classification accuracies as the dependent variables obtained from the same subject group, and the training approach (i.e., adversarial or non-adversarial) as the categorical independent variable. We use a repeated-measures test design since we make 10 different cross-subject model based predictions for each subject, hence accommodate for within-subject performance variabilities with the same method by considering the repetitions.

Figure 3 presents the differences in accuracies obtained with adversarial versus non-adversarial methods per subject, averaged across repetitions. In some cases, adversarial training yields more than 4% increases in cross-subject model transfer accuracies (e.g., 14% with one subject for DeepConvNet), indicating potential benefits of invariant representations for some subjects. Repeated-measures ANOVA tests indicated a significant performance increase with adversarial training for DeepConvNet ($p = 0.003$) and ShallowConvNet ($p = 0.02$), rejecting the null hypothesis that average accuracies across repetitions and subjects are equal. However we did not observe significant differences across the population for EEGNet ($p = 0.59$). We consider this to be potentially due to EEGNet being a more optimized architecture than DeepConvNet or ShallowConvNet in terms of the number of parameters to be learned and manipulated. Most importantly, generalization performances do not degrade significantly by adversarial regularization of deep EEG feature extractors.

C. SINGLE-TRIAL FEATURE INTERPRETATIONS

Figure 4 illustrates feature relevance maps for classifier decisions in three arbitrary single-trial cases, when learned models are transferred for cross-subject prediction. For all relevance maps, green color indicates a zero relevance score whereas an intensity of red indicates a positive, and an intensity of blue indicates a negative score. To exemplify from Figure 4(a) for a trial of subject 8 where $y_{\text{true}} = 1$ (right hand), the regular EEGNet architecture performs a wrong prediction of $\widehat{y} = 0$ with a confidence of $p_0 = 0.53$. However, the adversarially regularized EEGNet ($\lambda = 0.03$) performs a correct prediction of $\widehat{y} = 1$ with probability $p_1 = 0.61$, through the demonstrated feature relevance maps. An artifact-free classification of motor imagery is ideally expected to be performed via EEG evidences from motor cortical regions (i.e., electrodes C3, C4). However the regular EEGNet is entangling classifier predictions with class-irrelevant EEG artifacts from occipital electrodes (i.e.,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2971600, IEEE Access



(a) Subject: 8, Trial: 27, $y_{\text{true}} = 1$ (b) Subject: 30, Trial: 177, $y_{\text{true}} = 0$ (c) Subject: 43, Trial: 77, $y_{\text{true}} = 1$

FIGURE 4. Feature relevance map illustrations of non-adversarial versus adversarial models for class prediction in cross-subject model transfer scenarios. Different encoder architectures are demonstrated in (a), (b) or (c) during three arbitrary trials from different subjects. Specifications of the subject ID, trial number, and the true class label of the trials are provided in the subcaptions. Image matrices demonstrate the relevance maps for each EEG channel and time sample. Top image matrices relate to the non-adversarial models (λ=0), whereas the below image matrices relate to their adversarially trained counterparts. Black dotted vertical lines on these image matrices indicate arbitrary time points for when the associated relevance scalp maps are also demonstrated where indicated with an arrow. Scalp maps above the top image matrix, and under the below image matrix depict the time-averaged trial relevance maps. Predicted class labels in each case are indicated with $\widehat{y}$ being equal to 0 (class left) or 1 (class right), together with the confidence of predictions (i.e., probability value of p_0 for left or p_1 for right).

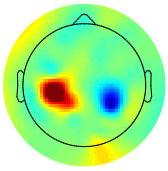
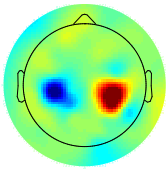
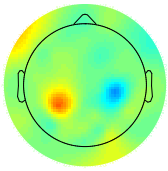
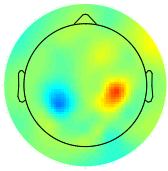
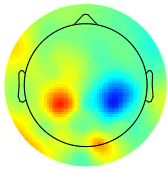
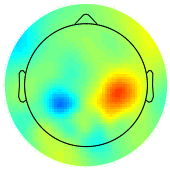
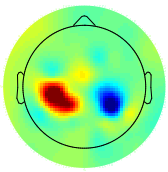
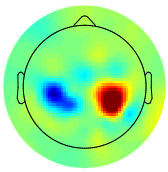
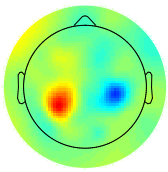
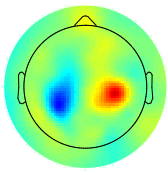
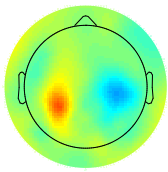
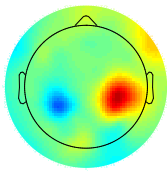
Iz) as observed with strong relevance scores on the time-averaged relevance topography. In this example, the adversarially learned model was able to censor this information from decision making as observed with the EEGNet ($\lambda = 0.03$) time-averaged relevance topography. Same differences can be also tracked at different time-points of the raw feature relevance maps as shown in Figure 4(a). Particular training set subjects who demonstrate artifactual activities across trials can influence deep learned models for decision making in this manner. This example illustrates how adversarial regularization can overcome these cases to perform robust decisions. Similar behaviors with adversarial regularization are further presented in Figure 4(b) for DeepConvNet, and in Figure 4(c) for ShallowConvNet, where eye blinks and jaw/muscle movement related artifacts (e.g., electrodes F7, AF8) are influencing incorrect decisions by regular CNNs.

Table 1 illustrates feature relevance scalp maps for cross-subject classifier decisions, averaged across time for each trial and across correctly predicted trials per class. For each architecture, a different subject's average relevance scalp maps for left and right class predictions are presented. For example, in the non-adversarial ShallowConvNet class left topography for subject 13, artifactual occipital patterns are observed. These were discarded in the adversarial counterpart below, leading to an ideal correct decision making with the invariant model. Similar behaviors are shown for class left in EEGNet (for subject 4), as well as in DeepConvNet (for subject 3) with jaw/muscle movement related artifacts that were unattended by adversarial training. Note that in the DeepConvNet example, relevance scores over the motor cortical areas are also strengthened with the adversarial models.

Taking a step back from cross-subject model transfer learning, Figure 5 illustrates feature relevance maps for adversary decisions in three arbitrary validation set trials after completion of model trainings. To exemplify from Figure 5(b) for a trial of subject 6 where $y_{\text{true}} = 1$ (right hand imagery), the regular DeepConvNet architecture is able to discriminate the subject for this trial ($\widehat{s} = 6$) with a very high confidence

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2971600, IEEE Access

TABLE 1. Average feature relevance scalp map illustrations for cross-subject model transfer. Topographies are obtained by first averaging raw relevance maps across time for each trial, and then averaging across correctly predicted trials per class. Each architecture is demonstrated with a different subject. Non-adversarial models indicate λ=0. Adversarially trained models utilize their optimal λ choices (i.e., EEGNet λ=0.03, DeepConvNet λ=0.05, ShallowConvNet λ=0.05).

	EEGNet (for subject 4)		**DeepConvNet (for subject 3)**		**ShallowConvNet (for subject 13)**	
	Left	Right	Left	Right	Left	Right
Non-Adversarial						
Adversarial						

across 40 subjects ($\widetilde{p}_6 = 0.88$), mainly relying on the eye blink patterns of this subject as observed by the time-averaged relevance scalp map. A further look into relevance scalp maps for the classifier decisions (illustrated in the dashed boxes alongside) also reveals an incorrect prediction of the class label as $\widehat{y} = 0$ even though the model weakly exploits motor cortical patterns. Nevertheless, the adversarial counterpart successfully misclassifies the subject of the trial ($\widehat{s} = 34$) with a close to chance level probability ($\widetilde{p}_{34} = 0.11$). As the encoder was trained to censor user-specific information, the adversary decision relies on any arbitrary EEG pattern (e.g., motor cortical rhythms in this case) rather than the eye blink artifacts. Accordingly, classifier prediction is also successfully performed with high confidence for the same validation set trial ($p_1 = 0.78$). In Figure 5(a) and (c), similar behaviors are presented when the adversarial models perform fooled, incorrect subject classifications with arbitrary EEG patterns and low confidences, whereas the regular CNNs were successfully learning user-discriminative EEG patterns that are encoded in the deep learned representations. These illustrations further demonstrate how classifier predictions were corrected with the invariant models.

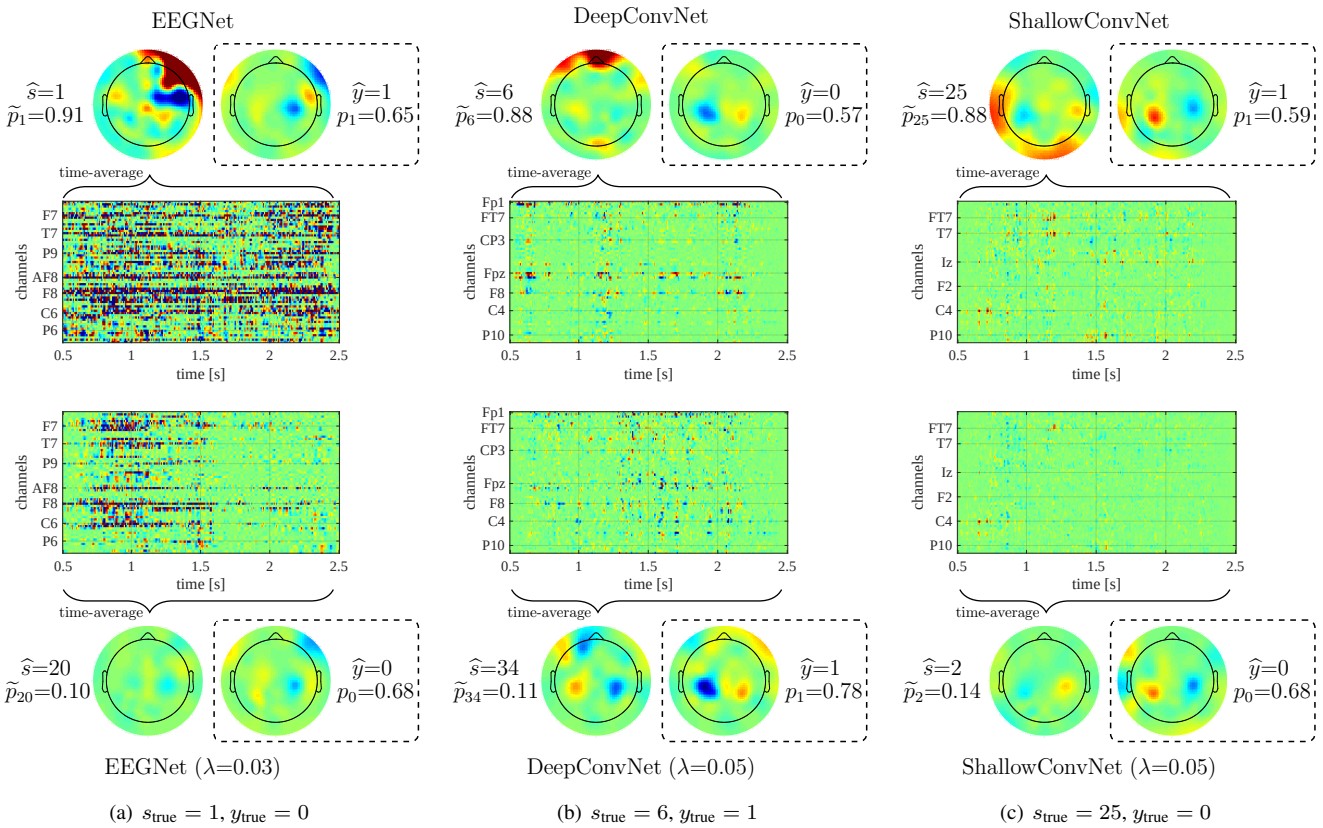
VI. DISCUSSION

In this work we propose a step towards invariance of deep EEG feature extractors in a systemic way using adversarial training methods within a discriminative framework. To the contrary of the widely relied on assumption that deep EEG neural network architectures internally generalize across-subjects, we argue that an adversarial regularization approach towards learning subject-invariant representations is likely to extend EEG deep learning approaches. Empirical results show that explicitly learning invariant EEG representations from a particular subject group can indeed be useful to generalize predictive models to novel subjects. Neurophysiological interpretations of the exploited EEG patterns further demonstrate the usefulness of our approach in cases where artifactual training data can affect model performances.

As one concerning observation, cross-subject decoding accuracies did not consistently show very high increases with all networks. We highlight this to be affected by various factors such as the network architecture, as well as the size and recording quality of the dataset to be used for training the model. As demonstrated, potentially due to being a more optimized architecture in terms of the number of parameters to be learned and manipulated, EEGNet did not significantly benefit in transfer accuracies. However the performances did not degrade by adversarial regularization, further showing benefits when deeper architectures were considered. Hence we argue that our approach provides a robust basis on invariant feature learning, and can particularly thrive when little or artifactual training data is under consideration. More importantly, feature relevance interpretations consistently demonstrated significant advantages in certain single-trial cases, which strongly supports our hypothesis on the need to systematically impose invariance constraints during conventional EEG deep learning model training.

One other limitation of our approach is related to the selection of the model learning subject group, which can lead to variations in accuracies for cross-subject model transfer. Although adversarial learning addresses potential confounders regarding subject-specific variations with respect to the rest of the model learning subject group, variability in transfer accuracies may still be caused due to a specific set of model learning subjects yielding good or bad discriminative performance for the decoding problem. Hence, one important aspect that still remains to be addressed is active selection of subjects from a pool with better discriminative task performance for transferable model learning. On another note, even though we make online decoding evaluations, our current approach requires temporal segmenting (e.g., trials), which does not extend to asynchronous EEG decoding yet.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2971600, IEEE Access



(a) $s_{\text{true}} = 1, y_{\text{true}} = 0$ (b) $s_{\text{true}} = 6, y_{\text{true}} = 1$ (c) $s_{\text{true}} = 25, y_{\text{true}} = 0$

FIGURE 5. Feature relevance map illustrations of non-adversarial versus adversarial models for user identification (i.e., adversary decisions) on samples from the validation sets, after model training. Different encoder architectures are demonstrated in (a), (b) or (c) during three arbitrary validation set trials from different subjects. Specifications of the true subject label (s_{true}) and true class label (y_{true}) of the trials are provided in the subcaptions. Image matrices demonstrate the relevance maps for each EEG channel and time sample. Top image matrices relate to the non-adversarial models, whereas the below image matrices relate to their adversarially trained counterparts. Scalp maps above the top image matrix, and under the below image matrix on the left sides depict the time-averaged trial relevance maps for the adversary. The scalp maps on the right sides inside the dashed boxes indicate the time-averaged relevance scalp maps for the classifier decision of the same trial. Predicted labels in each case are indicated with $\widehat{s} \in \{1, 2, \ldots, 40\}$ or $\widehat{y} \in \{0, 1\}$, together with the confidence of predictions (i.e., probability value of p_0 or p_1 for classifier predictions, and $\widetilde{p}_{\widehat{s}}$ for adversary predictions).

Many research studies have investigated calibration-less EEG classification models to develop simple BCI systems for communication [49–52]. At one end of calibration-free EEG classification, without considering an attempt for invariant representation learning, there exists several work on on-the-fly calibration of adaptive BCI classifiers [53]. Most common approaches include adjusting classifier parameters throughout BCI system use [54], [55], where models are initialized either by pre-trained classifiers on a subject pool [50], or simply initialized randomly [51], [52]. In terms of initializing such classifier models, our approach has the capability of constructing a subject-invariant baseline as well. To further extend this idea, besides a discriminative approach, ongoing recent work explores EEG data augmentation using GANs [56–61]. Such data augmentation would provide significant insights for model training with subject-invariant augmented EEG data, which is basically a generative approach to our problem of interest. Recently, we approached this invariant generative model aspect in our preliminary work for transfer learning [62], which is further currently being explored in an invariant EEG data augmentation context.

It is important to highlight that rather than proposing a new, alternative deep learning architecture for EEG feature extraction, we present a framework that can naturally be used to regularize any existing discriminative architecture to learn nuisance-invariant representations. Generally, regularization of neural networks is performed with dropout layers during model training [63]. In the context of this paper, we also exploit our knowledge on the source of intended invariance by adversarial censoring, and empirically demonstrate its benefits in learning invariant EEG representations. Since we were not interested in comparison of different deep learning models, or comparison of deep learning methods with respect to conventional EEG feature extraction protocols (e.g., common spatial patterns [64], [65]), we restricted our analyses to the comparison of adversarially trained versus regularly trained CNNs, importantly with neurophysiological interpretations of these models. In the light of the presented empirical results, we believe our approach would provide a more robust feature-invariance basis for existing deep learning models that are proposed for EEG-based decoding tasks.

.

APPENDIX A ENCODER ARCHITECTURES

Tables 2, 3 and 4 demonstrate the parameter specifications of the EEGNet [15], DeepConvNet and ShallowConvNet [14] architectures based on the original descriptions in the manuscripts, as well as their provided software implementations online. Encoder network inputs were defined as 64 channel EEG recordings with a sampling rate of 128 Hz for two seconds (i.e., 256 time samples). Only for the DeepConvNet and ShallowConvNet architectures, since the original parameter choices were developed for input EEG signals with a sampling rate of 250 Hz, all temporal convolution and pooling kernel sizes were taken as the half of the values used in [14], in consistency with the re-implementations by [15].

REFERENCES

[1] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, E. A. Allen, H. J. Bockholt, J. D. Long, H. J. Johnson, J. S. Paulsen, J. A. Turner, and V. D. Calhoun, "Deep learning for neuroimaging: a validation study," Frontiers in Neuroscience, vol. 8, p. 229, 2014.

[2] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," Journal of Neural Engineering, vol. 16, no. 3, p. 031001, 2019.

[3] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," Clinical Neurophysiology, vol. 113, no. 6, pp. 767–791, 2002.

[4] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," Proceedings of the IEEE, vol. 89, no. 7, pp. 1123–1134, 2001.

[5] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," Journal of Neural Engineering, vol. 15, no. 3, p. 031005, 2018.

[6] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," Journal of Neural Engineering, vol. 4, no. 2, p. R1, 2007.

[7] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," IEEE Computational Intelligence Magazine, vol. 11, no. 1, pp. 20–31, 2016.

[8] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, "Subject-independent mental state classification in single trials," Neural Networks, vol. 22, no. 9, pp. 1305–1312, 2009.

[9] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," IEEE Signal Processing Letters, vol. 16, no. 8, pp. 683–686, 2009.

[10] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms," IEEE Transactions on Biomedical Engineering, vol. 58, no. 2, pp. 355–362, 2011.

[11] W. Tu and S. Sun, "A subject transfer framework for EEG classification," Neurocomputing, vol. 82, pp. 109–116, 2012.

[12] H. Morioka, A. Kanemura, J.-i. Hirayama, M. Shikauchi, T. Ogawa, S. Ikeda, M. Kawanabe, and S. Ishii, "Learning a common dictionary for subject-transfer decoding with resting calibration," NeuroImage, vol. 111, pp. 167–178, 2015.

[13] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in International Conference on Learning Representations, 2016.

[14] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," Human Brain Mapping, vol. 38, no. 11, pp. 5391–5420, 2017.

[15] V. Lawhern, A. Solon, N. Waytowich, S. M. Gordon, C. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," Journal of Neural Engineering, vol. 15, no. 5, p. 056013, 2018.

[16] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, "Inter-subject transfer learning with end-to-end deep convolutional neural network for EEG-based BCI," Journal of Neural Engineering, 2018.

[17] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig, "Controllable invariance through adversarial feature learning," in Advances in Neural Information Processing Systems, 2017, pp. 585–596.

[18] G. Louppe, M. Kagan, and K. Cranmer, "Learning to pivot with adversarial networks," in Advances in Neural Information Processing Systems, 2017, pp. 981–990.

[19] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Adversarial deep learning in EEG biometrics," IEEE Signal Processing Letters, vol. 26, no. 5, pp. 710–714, 2019.

[20] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PloS one, vol. 10, no. 7, p. e0130140, 2015.

[21] S. Sakhavi and C. Guan, "Convolutional neural network-based transfer learning and knowledge distillation using multi-subject data in motor imagery BCI," in 8th International IEEE/EMBS Conference on Neural Engineering. IEEE, 2017, pp. 588–591.

[22] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," IEEE Transactions on Neural Networks and Learning Systems, no. 99, pp. 1–11, 2018.

[23] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 3, pp. 433–445, 2011.

[24] N.-S. Kwak, K.-R. Müller, and S.-W. Lee, "A convolutional neural network for steady state visual evoked potential classification under ambulatory environment," PloS one, vol. 12, no. 2, p. e0172578, 2017.

[25] S. Stober, D. J. Cameron, and J. A. Grahn, "Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings," in Advances in Neural Information Processing Systems, 2014, pp. 1449–1457.

[26] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," Journal of Neural Engineering, vol. 14, no. 1, p. 016003, 2016.

[27] N. Lu, T. Li, X. Ren, and H. Miao, "A deep learning scheme for motor imagery classification based on restricted boltzmann machines," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 6, pp. 566–576, 2017.

[28] G. Ruffini, D. Ibañez, M. Castellano, L. Dubreuil-Vall, A. Soria-Frisch, R. Postuma, J.-F. Gagnon, and J. Montplaisir, "Deep learning with EEG spectrograms in rapid eye movement behavior disorder," Frontiers in Neurology, vol. 10, 2019.

[29] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial EEG classification," Journal of Neuroscience Methods, vol. 274, pp. 141–145, 2016.

[30] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," arXiv preprint arXiv:1711.06104, 2017.

[31] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Deep feature learning for EEG recordings," in International Conference on Learning Representations, 2016.

[32] M. Hajinoroozi, Z. Mao, Y.-P. Lin, and Y. Huang, "Deep transfer learning for cross-subject and cross-experiment prediction of image rapid serial visual presentation events from EEG data," in International Conference on Augmented Cognition. Springer, 2017, pp. 45–55.

[33] C. Tan, F. Sun, W. Zhang, T. Kong, C. Yang, and X. Zhang, "Adaptive adversarial transfer learning for electroencephalography classification," in International Joint Conference on Neural Networks, 2018, pp. 1–8.

[34] W. Hang, W. Feng, R. Du, S. Liang, Y. Chen, Q. Wang, and X. Liu, "Cross-subject EEG signal recognition using deep domain adaptation network," IEEE Access, vol. 7, pp. 128 273–128 282, 2019.

[35] J. Schmidhuber, "Learning factorial codes by predictability minimization," Neural Computation, vol. 4, no. 6, pp. 863–879, 1992.

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[37] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," in International Conference on Learning Representations, 2016.

[38] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer et al., "Fader networks: Manipulating images by sliding attributes," in Advances in Neural Information Processing Systems, 2017, pp. 5967–5976.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2971600, IEEE Access

TABLE 2. EEGNet feature encoder implementation

Layer	Filters × (Kernel Size)	Output Dim.	Options	Number of Parameters
Input EEG		$(1, 64, 256)$		
Conv2D	$8 \times (1, 32)$	$(8, 64, 256)$	no bias, padding = same	256
BatchNorm		$(8, 64, 256)$		16
Depthwise Conv2D	$2 \times (64, 1)$	$(16, 1, 256)$	no bias, max norm = 1	1024
BatchNorm + ELU		$(16, 1, 256)$		32
Mean Pooling	$(1, 4)$	$(16, 1, 64)$		
Dropout		$(16, 1, 64)$	p = 0.25	
Separable Conv2D	$16 \times (1, 16)$	$(16, 1, 64)$	no bias, padding = same	512
BatchNorm + ELU		$(16, 1, 64)$		32
Mean Pooling	$(1, 8)$	$(16, 1, 8)$		
Dropout		$(16, 1, 8)$	p = 0.25	
Features (h)		$(1, 128)$	flatten	

TABLE 3. DeepConvNet feature encoder implementation

Layer	Filters × (Kernel Size)	Output Dim.	Options	Number of Parameters
Input EEG		$(1, 64, 256)$		
Conv2D	$25 \times (1, 5)$	$(25, 64, 252)$		150
Conv2D	$25 \times (64, 1)$	$(25, 1, 252)$	no bias	40,000
BatchNorm + ELU		$(25, 1, 252)$	epsilon = 10^{-5}, momentum = 0.1	50
Max Pooling	$(1, 2)$	$(25, 1, 126)$	strides = $(1, 2)$	
Dropout		$(25, 1, 126)$	p = 0.5	
Conv2D	$50 \times (1, 5)$	$(50, 1, 122)$	no bias	6,250
BatchNorm + ELU		$(50, 1, 122)$	epsilon = 10^{-5}, momentum = 0.1	100
Max Pooling	$(1, 2)$	$(50, 1, 61)$	strides = $(1, 2)$	
Dropout		$(50, 1, 61)$	p = 0.5	
Conv2D	$100 \times (1, 5)$	$(100, 1, 57)$	no bias	25,000
BatchNorm + ELU		$(100, 1, 57)$	epsilon = 10^{-5}, momentum = 0.1	200
Max Pooling	$(1, 2)$	$(100, 1, 28)$	strides = $(1, 2)$	
Dropout		$(100, 1, 28)$	p = 0.5	
Conv2D	$200 \times (1, 5)$	$(200, 1, 24)$	no bias	100,000
BatchNorm + ELU		$(200, 1, 24)$	epsilon = 10^{-5}, momentum = 0.1	400
Max Pooling	$(1, 2)$	$(200, 1, 12)$	strides = $(1, 2)$	
Dropout		$(200, 1, 12)$	p = 0.5	
Features (h)		$(1, 2400)$	flatten	

TABLE 4. ShallowConvNet feature encoder implementation

Layer	Filters × (Kernel Size)	Output Dim.	Options	Number of Parameters
Input EEG		$(1, 64, 256)$		
Conv2D	$40 \times (1, 13)$	$(40, 64, 244)$		560
Conv2D	$40 \times (64, 1)$	$(40, 1, 244)$	no bias	102,400
BatchNorm		$(40, 1, 244)$	epsilon = 10^{-5}, momentum = 0.1	80
Square: $f(x) = x^2$		$(40, 1, 244)$		
Mean Pooling	$(1, 35)$	$(40, 1, 30)$	strides = $(1, 7)$	
Log: $f(x) = \log(x)$		$(40, 1, 30)$		
Dropout		$(40, 1, 30)$	p = 0.5	
Features (h)		$(1, 1200)$	flatten	

[39] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in International Conference on Machine Learning, 2013, pp. 325–333.
[40] H. Edwards and A. Storkey, "Censoring representations with an adversary," in International Conference on Learning Representations, 2016.
[41] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in Computer Vision and Pattern Recognition, vol. 1, no. 2, 2017, p. 4.
[42] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in The Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[43] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 2096–2030, 2016.
[44] H. Cho, M. Ahn, S. Ahn, M. Kwon, and S. C. Jun, "EEG datasets for motor imagery brain–computer interface," GigaScience, vol. 6, no. 7, p. gix034, 2017.
[45] G. H. Klem, H. O. Lüders, H. Jasper, and C. Elger, "The ten-twenty electrode system of the international federation," Electroencephalogr Clin Neurophysiol, vol. 52, no. 3, pp. 3–6, 1999.
[46] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., "Tensorflow: A system for large-scale machine learning," in 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.
[47] F. Chollet, "Keras," 2015.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in International Conference on Learning Representations, 2015.
[49] H. Cecotti, "A self-paced and calibration-less SSVEP-based brain–computer interface speller," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 18, no. 2, pp. 127–133, 2010.
[50] S. Lu, C. Guan, and H. Zhang, "Unsupervised brain computer interface based on intersubject information and online adaptation," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 17, no. 2, pp. 135–145, 2009.
[51] M. Spüler, W. Rosenstiel, and M. Bogdan, "Online adaptation of a c-VEP brain-computer interface (BCI) based on error-related potentials and unsupervised learning," PloS one, vol. 7, no. 12, p. e51077, 2012.
[52] P.-J. Kindermans, M. Schreuder, B. Schrauwen, K.-R. Müller, and M. Tangermann, "True zero-training brain-computer interfacing–an online study," PloS one, vol. 9, no. 7, p. e102504, 2014.
[53] J. R. Millan, "On the need for on-line learning in brain-computer interfaces," in IEEE International Joint Conference on Neural Networks, vol. 4, 2004, pp. 2877–2882.
[54] F. Gembler, P. Stawicki, and I. Volosyak, "Autonomous parameter adjustment for SSVEP-based BCIs with a novel BCI wizard," Frontiers in Neuroscience, vol. 9, p. 474, 2015.
[55] X. Song and S.-C. Yoon, "Improving brain–computer interface classification using adaptive common spatial patterns," Computers in Biology and Medicine, vol. 61, pp. 150–160, 2015.
[56] I. A. Corley and Y. Huang, "Deep EEG super-resolution: Upsampling EEG spatial resolution with generative adversarial networks," in IEEE EMBS International Conference on Biomedical & Health Informatics, 2018, pp. 100–103.
[57] Y. Luo and B.-L. Lu, "EEG data augmentation for emotion recognition using a conditional wasserstein GAN," in International Conference of the IEEE Engineering in Medicine and Biology Society, 2018, pp. 2535–2538.
[58] S. M. Abdelfattah, G. M. Abdelrahman, and M. Wang, "Augmenting the size of EEG datasets using generative adversarial networks," in International Joint Conference on Neural Networks, 2018, pp. 1–6.
[59] N. K. N. Aznan, A. Atapour-Abarghouei, S. Bonner, J. Connolly, N. A. Moubayed, and T. Breckon, "Simulating brain signals: Creating synthetic EEG data via neural-based generative models for improved SSVEP classification," in International Joint Conference on Neural Networks, 2019.
[60] W. Ko, E. Jeon, J. Lee, and H.-I. Suk, "Semi-supervised deep adversarial learning for brain-computer interface," in 2019 7th International Winter Conference on Brain-Computer Interface (BCI). IEEE, 2019, pp. 1–4.
[61] S. Hwang, K. Hong, G. Son, and H. Byun, "EZSL-GAN: EEG-based zero-shot learning approach using a generative adversarial network," in 2019 7th International Winter Conference on Brain-Computer Interface (BCI). IEEE, 2019, pp. 1–4.
[62] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Transfer learning in brain-computer interfaces with adversarial variational autoencoders," in 9th International IEEE/EMBS Conference on Neural Engineering (NER), 2019, pp. 207–210.
[63] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.
[64] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in IEEE International Joint Conference on Neural Networks, 2008, pp. 2390–2397.
[65] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," IEEE Signal Processing Magazine, vol. 25, no. 1, pp. 41–56, 2008.

OZAN ÖZDENIZCI (S'18) received the B.S. degree in electronics engineering with a minor in mathematics and the M.S. degree in electronics engineering from Sabancı University, Istanbul, Turkey, in 2014 and 2016 respectively. He is currently pursuing his Ph.D. degree in electrical engineering at Northeastern University, Boston, MA, USA. He was a Research Intern at the Max Planck Institute for Intelligent Systems, Tübingen, Germany in 2013 and 2014, and at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA in 2018. His research interests include statistical signal processing and machine learning, with applications to biomedical information processing.

YE WANG (SM'19) received the B.S. degree in electrical and computer engineering from Worcester Polytechnic Institute, Worcester, MA, USA in 2005, and the M.S. and Ph.D. degrees in electrical and computer engineering from Boston University, Boston, MA, USA in 2009 and 2011 respectively. In 2012, he joined Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA, where he had also previously completed an internship in 2010. His broad research interests are information theory, machine learning, signal processing, communications, and data privacy/security.

TOSHIAKI KOIKE-AKINO (SM'11) received the B.S. degree in electrical and electronics engineering, M.S. and Ph.D. degrees in communications and computer engineering from Kyoto University, Kyoto, Japan, in 2002, 2003, and 2005, respectively. During 2006–2010 he was a Postdoctoral Researcher at Harvard University, Cambridge, MA, USA, and joined Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA, in 2010. His research interests include signal processing for data communications and sensing. He received the YRP Encouragement Award 2005, the 21st TELECOM System Technology Award, the 2008 Ericsson Young Scientist Award, the IEEE GLOBECOM'08 Best Paper Award in Wireless Communications Symposium, the 24th TELECOM System Technology Encouragement Award, and the IEEE GLOBECOM'09 Best Paper Award in Wireless Communications Symposium.

DENIZ ERDOĞMUŞ (SM'07) received the B.S. degrees in electrical engineering and mathematics in 1997, and the M.S. degree in electrical engineering in 1999 from Middle East Technical University, Ankara, Turkey. He received his Ph.D. in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2002. He held a Postdoctoral position with the University of Florida until 2004. He was with the computer science and electrical engineering, and the biomedical engineering departments at the Oregon Health and Science University, Portland, OR, USA until 2008. Since 2008, he has been with the electrical and computer engineering department at Northeastern University, Boston, MA, USA. His research focuses on statistical signal processing and machine learning with applications to contextual signal/image/data analysis with applications in cyber-human systems including brain computer interfaces and technologies that collaboratively improve human performance. He has served as an associate editor and program committee member for a number of journals and conferences in these areas, including IEEE Signal Processing Letters, and the following IEEE Transactions: Signal Processing, Biomedical Engineering, and Neural Networks.