DOMAIN-INVARIANT REPRESENTATION LEARNING FROM EEG WITH PRIVATE ENCODERS

David Bethge[1,2]*, Philipp Hallgarten*[1,3]*, Tobias Grosse-Puppendahl*[1]*, Mohamed Kari*[1]*,*
Ralf Mikut[3]*, Albrecht Schmidt*[2]*, Ozan Özdenizci*[4,5]

[1] Dr. Ing. h.c. F. Porsche AG, Stuttgart, Germany
[2] Ludwig-Maximilians University Munich, Germany
[3] Karlsruhe Institute of Technology, Germany
[4] Institute of Theoretical Computer Science, Graz University of Technology, Austria
[5] TU Graz - SAL Dependable Embedded Systems Lab, Silicon Austria Labs, Austria

ABSTRACT

Deep learning based electroencephalography (EEG) signal processing methods are known to suffer from poor test-time generalization due to the changes in data distribution. This becomes a more challenging problem when privacy-preserving representation learning is of interest such as in clinical settings. To that end, we propose a multi-source learning architecture where we extract domain-invariant representations from dataset-specific private encoders. Our model utilizes a maximum-mean-discrepancy (MMD) based domain alignment approach to impose domain-invariance for encoded representations, which outperforms state-of-the-art approaches in EEG-based emotion classification. Furthermore, representations learned in our pipeline preserve domain privacy as dataset-specific private encoding alleviates the need for conventional, centralized EEG-based deep neural network training approaches with shared parameters.

1. INTRODUCTION

Over the past decades, electroencephalography (EEG) based emotion recognition gained significant interest towards developing affective neural-machine interfaces [1]. However, due to the costly data collection processes, large EEG datasets recorded under emotion eliciting experimental paradigms are not ubiquitous. Several small scale affective EEG datasets were previously introduced, which however highly differ in their experimental setups (*e.g.*, stimuli, emotion labels). This introduces the well-known *domain adaptation* problem, i.e., models trained to recognize emotions on one of these datasets commonly fails to solve the same task for another dataset. In order to make EEG-based emotion recognition algorithms suitable for a variety of experiments and real-life scenarios, it is crucial to obtain a model that can tackle this task across multiple datasets. One approach to achieve this from a deep learning perspective is to extract and exploit domain-invariant representations from multi-channel EEG data.

Recent work have shown significant promise in using invariant representation learning from time-series EEG data for cross-subject generalization [2–4], while learning across multiple dataset sources remains an open question. Notable methods successfully introduced adversarial censoring for domain invariance, widely known as DANN [5], in various EEG decoding tasks [2, 3], including emotion recognition [4]. Adversarial domain regularization with DANN [5] considers a cross-domain shared encoder that extracts features from data of all subjects, and two classifiers: a task and a domain classifier. During training the encoder is adversarially penalized on the domain classification loss, which enforces the model towards learning domain-invariant representations. In this work we take a different approach to invariant EEG representation learning by further considering to preserve domain privacy that is of critical importance in clinical settings [6, 7].

We propose a multi-source learning framework for domain invariant representation learning from time-series signals such as multi-channel EEG recordings. From a different perspective than adversarial learning methods, our framework consists of a private feature encoder per domain and a cross-domain shared classifier, where we utilize a maximum-mean-discrepancy (MMD) [8] based domain alignment loss across private feature encoders to minimize domain-specific leakage within the learned representations. Our contributions in this work are three-fold: (1) We introduce a deep neural signal processing architecture where EEG time-series are privately encoded at the source end and only the learned representations are shared to a global classifier that enables decentralized cross-dataset learning by preserving domain privacy. (2) We reveal large dataset domain specific variances in conventionally trained centralized pipelines, and demonstrate that regularizing latent representations via an MMD-based domain alignment loss enables dataset source independent representation learning. (3) We show that the use of adaptive batch normalization layers in such multi-source settings prevent performance decrease at inference time.

2. MATERIALS AND METHODS

2.1. Experimental Data

Benchmark Affective EEG Datasets: We used four publicly available EEG-based emotion recognition datasets: *DEAP* [9], *DREAMER* [10], *SEED* [11, 12], *SEED-IV* [13]. During EEG recordings subjects were presented audio-visual stimuli that are expected to elicit distinct emotions. Differences in the experimental setups lead to a variability in the structure of data samples and their labels. *SEED* and *SEED-IV* contain 62-channel EEG recordings sampled at $200\,\text{Hz}$, whereas *DEAP* contains recordings from 32 electrodes and *DREAMER* from 14, both sampled at $128\,\text{Hz}$. Emotion labels of *SEED* and *SEED-IV* represents one out of three and four discrete emotions respectively, labeled by the stimuli that the subjects were presented. For *DEAP* and *DREAMER* post-recording subject self-assessment ratings on the *valence* and *arousal* continuous scales were included as the labels.

Label Transformation: To realize our cross-dataset experimental analyses, we transformed the label spaces across all four datasets into a common set. We determined the three discrete emotions *Negative*, *Neutral*, and *Positive* (as also used in the *SEED*-dataset) as the common label-space. Using *k*-means* clustering in the two-dimensional label space of *Valence* and *Arousal* for *DEAP*, and *DREAMER* datasets, we determined four clusters corresponding to the four discrete emotions used as labels in the *SEED-IV* dataset: *Fear*, *Sad*, *Neutral*, *Happy*. We merged the former two into being the *Negative* class, and considered *Happy* to be the *Positive* class. Hence we obtained three class labels in a unified manner.

2.2. Multi-Source EEG Processing for Private Encoding

We will denote multiple datasets $\mathcal{D}_k$ consisting of time-series EEG signal epochs and corresponding emotion label of the experiment paradigm as pairs (s_j, l_j). We pre-process the signals to serve as input to our architecture as follows. First, we perform baseline correction for signals s_j using a three-second time-window and apply a $4-40\,\text{Hz}$ Butterworth band-pass filter. Since the length of the available s_j differ within each dataset, we used only the last T seconds of each available signal with T being the length of the shortest time-series in the dataset, i.e., 60, 64, 185, 50 for *DEAP*, *DREAMER*, *SEED* and *SEED-IV* respectively. Then, we segment the signals into non-overlapping $2\,\text{s}$ windows that will be used as inputs x_i to our model. We assign the label l_{j_i} to each window x_i. All pairs of x_i and corresponding labels y_i obtained from a dataset are considered as a processed data-source $\mathcal{X}_k$.

To balance the number of samples across data-sources and with respect to the label for each data-source individually, we applied stochastic undersampling while constructing our training set. We split the data into the training (60%), validation (20%) and test (20%) sets, while we assure that the stratification constraints hold for each subset individually.

2.3. Multi-Source Learning Framework

Domain Aligned Private Encoders (DAPE): Our model consist of one private encoder per data-source and a shared classifier. While the encoder can be designed arbitrarily for each data-source, for proof of concept we used the *DeepConvNet* architecture [14] for our all data-sources. We replaced the final pooling layer in *DeepConvNet* by an adaptive average pooling layer, that introduces the flexibility to control the number of features n_z in the latent representation z. In our models we chose $n_z = 50$ as the best results in terms of the balance between emotion classification and domain invariance were achieved. One extension of our model towards achieving better domain adaptation, which we denote by adaptive DAPE (aDAPE), considers the use of adaptive batch normalization layers [15]. Such layers were also recently used in deep EEG analysis by [16, 17], instead of conventional batch normalization. By doing so, we ensure each layer of any given encoder to receive data from a similar distribution at test time as well, as samples are normalized by the statistics estimated from the utilized mini-batch.

During each forward pass, a batch of B samples from $x_i \sim \mathcal{X}_i$ are drawn and used as input to each of the M encoders, represented through the parameters $\theta_i, i \in \{1, \ldots, M\}$, with M also being the number of data-sources. We consider the output representations $z_i = f(\theta_i, x_i) \in \mathbb{R}^{B \times n_z}$ of each encoder as samples drawn from independent data-source distributions. To overcome the problem of domain shift across encoders and regularize the encoders towards learning domain-invariant representations, we use an MMD [8] based domain-alignment loss $\mathcal{L}_{\text{DA}}$ using Gaussian kernels with $\sigma_j \in \{10, 15, 20, 50\}$. Rather than computing the MMD for all combinations of data-sources, we randomly sample M pairs of distributions $\{(z_p^{(i)}, z_q^{(i)})\}_{i=1}^{M}$, $p, q \in \{1, \ldots, M\}$ such that $p \neq q$, calculate the MMD loss [8] between them and accumulate in the domain alignment regularizer:

$$\mathcal{L}_{DA} = \sum_{j=1}^{|\sigma|} \sum_{i=1}^{M} \text{MMD}^2(z_p^{(i)}, z_q^{(i)}, \sigma_j), \quad (1)$$

where $|\sigma|$ denotes the cardinality for the pre-defined set σ_j is chosen from and $\text{MMD}^2(z_p^{(i)}, z_q^{(i)}, \sigma_j)$ is defined as follows:

$$\begin{aligned}
\text{MMD}^2 = & \frac{1}{B(B-1)} \sum_{m=1}^{B} \sum_{\substack{n=1 \\ n \neq m}}^{B} \varphi(z_{p_m}^{(i)}, z_{p_n}^{(i)}, \sigma_j) \\
& + \frac{1}{B(B-1)} \sum_{m=1}^{B} \sum_{\substack{n=1 \\ n \neq m}}^{B} \varphi(z_{q_m}^{(i)}, z_{q_n}^{(i)}, \sigma_j) \\
& - \frac{2}{B^2} \sum_{m=1}^{B} \sum_{n=1}^{B} \varphi(z_{p_m}^{(i)}, z_{q_n}^{(i)}, \sigma_j), \quad (2)
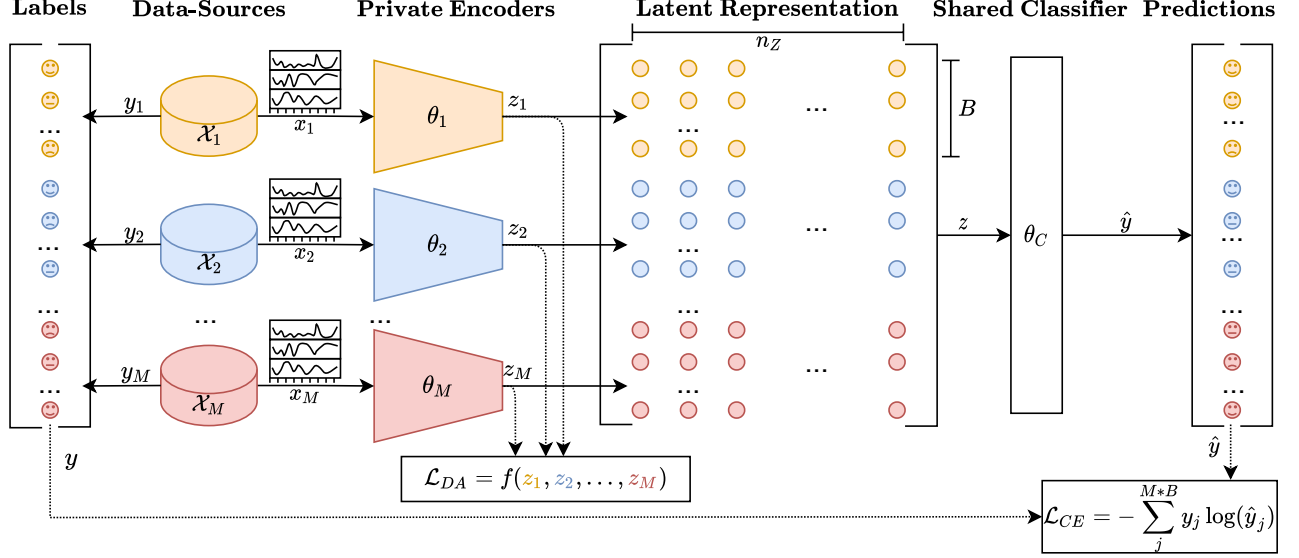\end{aligned}$$

Fig. 1. DAPE Architecture. Each color represents one data-source. Signals x_1 from each data-source $\mathcal{X}_i$ are processed by a private encoder. Output representations z_i are then used to calculate the domain-alignment loss $\mathcal{L}_{DA}$ via Eq. (1). In addition, output representations are concatenated to a representation $z \in \mathbb{R}^{(B*M) \times n_z}$ and used as input for a shared classifier. Classifier outputs a batch of predictions $\hat{y}_i$ for each batch of signals x_i. Corresponding batches of labels y_i are concatenated in the same way as the latent representations z_i and used together with the predictions for the calculation of the cross-entropy loss $\mathcal{L}_{CE}$.

with $\varphi(z_{p_m}^{(i)}, z_{q_n}^{(i)}, \sigma_j)$ being the Gaussian kernel function:

$$\varphi(z_{p_m}^{(i)}, z_{q_n}^{(i)}, \sigma_j) = \exp\left(\frac{-\|z_{p_m}^{(i)} - z_{q_n}^{(i)}\|^2}{2\sigma_j^2}\right). \quad (3)$$

All batches of representations z_i are then stacked to one vector of representations $z \in \mathbb{R}^{(B*M) \times n_z}$, and used as input for the classifier with parameters θ_C to compute predicted emotions $\hat{y}$. We train the network with a categorical cross-entropy loss function $\mathcal{L}_{CE} = -\sum_j^{M*B} y_j \log(\hat{y}_j)$. Note that only the parameters θ_C and θ_i (parameters of the encoder that processed a sample x_j) contribute to the prediction of an individual sample $\hat{y}_j$, which ensures that only the classifier and the encoder i are optimized based on the gradient of the loss.

Multi-Objective Optimization: Training of *DAPE* can be considered as a multi-objective optimization. On one hand, we want to minimize the cross-entropy loss for the classification task and one the other hand we want to minimize the domain-alignment loss $\mathcal{L}_{DA}$ for learning domain-invariant representations. We assume that these goals are not contradictory and that the classification task gets more robust using domain-invariant representations of the samples. To control the domain-invariance of representations we use a hyperparameter κ as the tunable weight in the optimization problem:

$$\theta_1^*, \ldots, \theta_M^*, \theta_C^* = \operatorname*{arg\,min}_{\theta_1, \ldots, \theta_M, \theta_C} \mathcal{L}_{CE} + \kappa \mathcal{L}_{DA}. \quad (4)$$

We use an annealing update scheme for κ as proposed by [18,

19]. To ensure that the domain alignment does not lead the encoders to learn task-irrelevant representations, we start the training with $\kappa = 0$ and increase it with the beginning of the 5th epoch by a rate of 0.25 per epoch. To prevent the domain-alignment loss dominating the cross-entropy loss in the later epochs, we stop updating κ by the 70th epoch and retain the value of 16.25 for all subsequent epochs.

2.4. Quantifying Domain-Invariance of Representations

One of the motivating goals in our architecture is to preserve domain privacy at the data-source end, as well as ensuring that an attacker not being able to deduce the information concerning the data-source from latent representations. To assess invariance of learned representations via DAPE, we firstly compute the latent representations z_i of all samples in the test set. We use 80% of the samples to train a *domain classifier* using the data-source ID as label. We subsequently evaluate the domain-invariance of the representations by calculating the achieved accuracies of these domain classifiers on the remaining 20% of the test set representations. Note that a lower 4-class domain-classifier accuracy corresponds to a higher domain-invariance of the learned representations, which is favorable for multi-source learning. We considered a multitude of learning machines as domain classifiers (e.g., support vector machines (SVM), quadratic discriminant analysis). Since all domain classifiers showed similar results we report our results with linear SVMs in Section 3.2.

Table 1. Accuracies of 3-class emotion classification evaluated with *DAPE*, *aDAPE* and baseline methods: local and global baseline models, and *DANN* architecture. Bottom row shows the results of the linear SVM used to assess the 4-class domain-invariance. Local baseline corresponds to using four independent networks (i.e., dataset-specific encoder and classifiers) without any regularization or domain-alignment.

		Local Models	Global Model	*DANN* [5]	*DAPE* (**Ours**)	*aDAPE* (**Ours**)
Emotion Classification ($\nearrow$)	*DEAP*	50.01% (±0.58)	40.02% (±1.09)	39.81% (±0.87)	47.81%(±0.24)	47.99%(±1.30)
	DREAMER	59.03% (±2.41)	43.78% (±1.28)	42.30% (±1.60)	49.13%(±1.54)	48.70%(±1.99)
	SEED	56.19% (±1.97)	42.93% (±0.95)	42.63% (±0.26)	51.72%(±0.67)	53.45%(±2.05)
	SEED-IV	42.42% (±3.64)	34.55% (±0.78)	34.57% (±0.56)	41.94%(±0.64)	43.25%(±1.07)
	Mean	51.91% (±2.15)	40.32% (±0.42)	39.82% (±0.33)	47.65%(±0.13)	**49.09**%(±1.49)
Domain Classification ($\searrow$)	**Mean**	99.87% (±0.15)	60.11% (±3.47)	66.17%(±4.45)	82.83%(±3.18)	**52.76**%(±2.92)

3. EXPERIMENTAL RESULTS

3.1. Emotion Classification

Table 1 summarizes our results on the emotion classification task. Here *local models* indicate four independent networks processed on each data-source individually (i.e., *DAPE* with $M = 1$), which would be a baseline approach if one is not interested in domain-privacy or representation invariance. To the contrary, the *global model* indicates a baseline where one trains a single, unified model by pooling all data-source inputs into one unified training set, however constraining both the amount of training data and the encoder specifications to be uniform and shared across all data-sources. DANN [5] depicts an invariant representation learning approach that was previously considered for EEG decoding in [2, 4], where one can train a model adversarially by censoring the data-source ID relevant information from latent representations to impose invariance, without considering data-source privacy.

While the global model and DANN perform only slightly above chance level (33%) in emotion classification, we observe that *DAPE* and *aDAPE* clearly outperform the baseline approaches that learns a unified model for emotion classification, while simultaneously ensuring data-source privacy. Furthermore, we observe that *aDAPE* is also only slightly below the average of the extensive local models (49.09% vs. 51.91%), although solving this task across multiple domains poses a more challenging problem than learning locally.

3.2. Domain-Invariance of the Learned Representations

Domain-invariance of the learned representations estimated through the linear SVM are shown at the bottom row of Table 1. We observe that *aDAPE* shows the highest domain-invariance in the learned representations, where in all other models the linear SVM was able to deduce the data-source ID from the representations with $> 60\%$ accuracy. Achieved domain-invariance is $\sim 14\%$ higher for *aDAPE* than for *DANN*, and $\sim 47\%$ higher than for the local models. Fur-

thermore, it is important to note that the performance of the domain classifier is much higher ($\sim 30\%$) for *aDAPE* than for *DAPE*, from an ablation study perspective. This significant gap results in the use of adaptive batch normalization layers. Overall emotion classification accuracy of *aDAPE* is higher than of *DAPE*, and achieved domain-invariance of representations is also stronger (while still being slightly above the chance-level), confirming our hypothesis that using domain-invariant representations in multi-source settings helps by making the classification task more robust.

4. CONCLUSION

We present *Domain Aligned Private Encoders* as a multi-source learning framework for domain-invariant representation learning, and demonstrate its feasibility on an EEG-based emotion classification task using data from four publicly available datasets. Different than state-of-the-art adversarial training approaches to learn invariant EEG representations [2–4], we utilize an MMD based domain alignment loss [16, 20] across dataset-specific private feature encoders. Proposed deep neural signal processing architecture encodes multi-channel EEG time-series signals privately at the source end and only the learned representations are shared to a global classifier. Our decentralized invariant representation learning approach and use of adaptive batch normalization layers improved performance in our experimental analyses.

Accessible EEG data is generally present in small datasets, which are also distributed across various countries and/or laboratories. Due to naturally occurring data privacy concerns, we believe that such architectures as proposed in this work make common use of private data representations more convenient and real-life applicable (e.g., clinical time-series monitoring [21], personalized BCI-based stroke rehabilitation protocols [6, 22]), hence our work advances the area of cross-domain transfer learning for EEG signals in that sense.

5. REFERENCES

[1] Christian Mühl et al., "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.

[2] Ozan Özdenizci et al., "Learning invariant representations from EEG via adversarial inference," *IEEE Access*, vol. 8, pp. 27074–27085, 2020.

[3] Eunjin Jeon et al., "Mutual information-driven subject-invariant and class-relevant deep representation learning in BCI," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[4] Soheil Rayatdoost et al., "Subject-invariant EEG representation learning for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 3955–3959.

[5] Yaroslav Ganin et al., "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[6] Anisha Agarwal et al., "Protecting privacy of users in brain-computer interface applications," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 8, pp. 1546–1555, 2019.

[7] Ce Ju et al., "Federated transfer learning for EEG signal classification," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, 2020, pp. 3040–3045.

[8] Arthur Gretton et al., "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[9] Sander Koelstra et al., "DEAP: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[10] Stamos Katsigiannis and Naeem Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[11] Ruo-Nan Duan et al., "Differential entropy feature for EEG-based emotion classification," in *6th International IEEE/EMBS Conference on Neural Engineering*, 2013, pp. 81–84.

[12] Wei-Long Zheng and Bao-Liang Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[13] Wei-Long Zheng et al., "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.

[14] Robin Tibor Schirrmeister et al., "Deep learning with Convolutional Neural Networks for EEG Decoding and Visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[15] Yanghao Li et al., "Revisiting batch normalization for practical domain adaptation," *ICLR Workshops*, 2017.

[16] Magdiel Jiménez-Guarneros and Pilar Gómez-Gil, "Custom domain adaptation: A new method for cross-subject, EEG-based cognitive load recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 750–754, 2020.

[17] Magdiel Jiménez-Guarneros and Pilar Gómez-Gil, "Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition," *Pattern Recognition Letters*, vol. 141, pp. 54–60, 2021.

[18] Edgar Schonfeld et al., "Generalized zero-and few-shot learning via aligned variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8247–8255.

[19] Samuel R Bowman et al., "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[20] Hao Chen et al., "MS-MDA: Multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition," *arXiv preprint arXiv:2107.07740*, 2021.

[21] Ozan Özdenizci et al., "Time-series prediction of proximal aggression onset in minimally-verbal youth with autism spectrum disorder using physiological biosignals," in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2018, pp. 5745–5748.

[22] Anastasia-Atalanti Mastakouri et al., "Personalized brain-computer interface models for motor rehabilitation," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2017, pp. 3024–3029.