PDAML: A PSEUDO DOMAIN ADAPTATION PARADIGM FOR SUBJECT-INDEPENDENT EEG-BASED EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain adaptation (DA) and domain generalization (DG) methods have been successfully adopted to alleviate the domain shift problem caused by the subject variability of EEG signals in subject-independent affective brain-computer interfaces (aBCIs). Usually, the DA methods give relatively promising results than the DG methods but require additional computation resources each time a new subject comes. In this paper, we first propose a new paradigm called *Pseudo Domain Adaptation* (PDA), which is more suitable for subject-independent aBCIs. Then we propose the pseudo domain adaptation via meta-learning (PDAML) based on PDA. The PDAML consists of a feature extractor, a classifier, and a sum-decomposable structure called domain shift governor. We prove that a network with a sum-decomposable structure can compute the divergence between different domains effectively in theory. By taking advantage of the adversarial learning and meta-learning, the governor helps PDAML quickly generalize to a new domain using the target data through a few self-adaptation steps in the test phase. Experimental results on the public aBICs dataset demonstrate that our proposed method not only avoids the additional computation resources of the DA methods but also reaches a similar generalization performance of the state-of-the-art DA methods.

1 INTRODUCTION

Affective brain-computer interfaces (aBCIs), which focus on developing machines to recognize human emotion automatically and provide more humanized interaction, have attracted widespread attention from academics and industries (Mühl et al. (2014); Shanechi (2019)). Various studies have demonstrated that the electroencephalography (EEG) signal is especially reliable to recognize human emotion in the subject-dependent emotion model, in which the training and test data are from one subject (Jenke et al. (2014); Alarcão & Fonseca (2019)). However, due to the non-stationary nature of EEG signal and structural variability between different subjects, subject-independent model based on the assumption of Independent Identity Distribution(i.i.d.) usually shows bad generalization performance in real aBCIs applications, which is called the problem of Domain Shift (Sugiyama et al. (2007); Samek et al. (2013); Sussillo et al. (2016); Zheng & Lu (2016); Li et al. (2018d)).

Domain adaptation (DA) is one of the promising ways to solve this problem. DA uses data from both source and target domain to promote the adaption performance. One of the most sufficient studied DA is mapping the two distributions to one common feature space where they have the same marginal distribution. Though DA has demonstrated significant success in subject-independent EEG-based emotion recognition (Zheng & Lu (2016); Li et al. (2018b;c); Luo et al. (2018)), the additional computation and time to apply the DA methods is an exasperating problem in real-world scenarios and causes poor user experience. As a consequence, the concept of domain generalization (DG) arises in situations where multiple source domains can be accessed but unlabeled target samples are not available. The DG methods have also been successfully adopted to build the subject-independent emotion model (Ma et al. (2019b;a)). However, since there is no prior information about the target domain during training, it's challenging for DG to perform as promising as DA.

A compromise solution is Fast Domain Adaptation (FDA), which trains the main model in advance and uses a small amount of test samples to adjust efficiently. Even though FDA methods can avoid consuming time cost by adaptation, most of them need to storage both source and target domains in the test phase (Chai et al. (2017); Zhao et al. (2021)), which needs extra storage space and causes poor portability. However, in real-world applications, whether a EEG-based affective model can adapt to different subjects quickly as well as keep its portability does matter.

In this paper, we propose a new paradigm called *Pseudo Domain Adaptation*(PDA) for subject-independent EEG-based emotion recognition. Compared to typical FDA, only target domain is required in the test phase of PDA, which makes PDA capable to output the prediction more quickly than DA and FDA. From the perspective of the real-world application, PDA is more suited to build subject-independent affective model.

To implement PDA, we first prove that with the sum-decomposable structure, a network is equivalent to domain discrepancy metrics in traditional DA methods like MMD or $\mathcal{H}$-divergence. Based on our theory, we propose a method named Pseudo Domain Adaptation via Meta Learning(PDAML). The PDAML consists of a feature extractor, a classifier, and a sum-decomposable structure called domain shift governor. By taking advantage of the adversarial learning and meta-learning, the governor helps PDAML quickly generalize to a new domain using the target data through a few self-adaptive steps in the test phase.

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

- We propose a more suitable paradigm to build subject-independent affective model.
- We prove that a sum-decomposable network is equivalent to domain discriminator for representing any type of domain discrepancy in theory.
- We propose PDAML that is portable and can fast adapt to different subjects in EEG-based emotion recognition.
- We conduct extensive experiments on the SEED dataset[1], which is a public available EEG-based affective dataset. The experimental results demonstrate that our method has better performance than DG methods. From the perspective of time and storage cost, the PDAML performs as well as DG methods.

2 Related Work

aBCIs have received considerable attention very recent. Mühl *et al.* presented the definition of aBCIs (Mühl et al. (2014)) by introducing the affective factors into traditional brain-computer interfaces (BCIs) (Zander & Jatzev (2011)). More applied works focus on studying EEG-based emotion recognition. Zheng and Lu recruited 15 subjects to watch 15 selected Chinese movie clips to elicit three emotions: happy, neutral and sad. They developed a public emotion dataset called SEED by recording the EEG signals of the subjects (Zheng & Lu (2015)). Based on the SEED dataset, researchers have made great progress in developing EEG-based emotion recognition model, especially for subject-dependent model.

Due to the non-stationary nature of EEG signal and structural variability between different subjects, it is hard to develop subject-independent EEG-based emotion recognition model by directly using typical machine learning approaches. Researchers have focused on applying DA and DG methods to subject-independent EEG-based emotion recognition. Typical DA methods are discrepancy-based, and they alleviate the domain shift problem by minimizing traditional metrics, such as Maximum Mean Discrepancy (MMD) (Pan et al. (2011); Long et al. (2017); Wang et al. (2018)), Kullback-Leibler (*KL*) divergence (Zhuang et al. (2015)), and $\mathcal{H}$-divergence (Bendavid et al. (2010)), between the different domains. Zheng and Lu adopted transfer component analysis (TCA) (Pan et al. (2011)), which minimizes MMD (Gretton et al. (2007)) between two domains by constructing kernel matrix, and successfully built personalized EEG-based emotion models (Zheng & Lu (2016)).

Recently, adversarial DA methods have made great successes in different fields (Ganin & Lempitsky (2015); Tzeng et al. (2017); Shen et al. (2018)). The basic idea of the adversarial training is similar to generative adversarial networks (GANs) (Goodfellow et al. (2014)), which play an adversarial

[1]http://bcmi.sjtu.edu.cn/∼seed/index.html

game to make the generated distribution approximate to the real distribution. After the adversarial training, the data distribution of the target domain is similar to the source domain, and the domain shift is diminished. Researchers have successfully adopted adversarial DA methods to aBCIs. Li *et al.* (Li et al. (2018b)) adopted domain-adversarial neural networks (DANN) (Ganin & Lempitsky (2015)) to EEG-based emotion recognition and improved the recognition accuracies of subject-independent models. And they (Li et al. (2018c)) also achieved significant performance in subject-independent vigilance estimation by implementing DANN and adversarial discriminative domain adaptation (ADDA) (Tzeng et al. (2017)). Luo *et al.* proposed Wasserstein GAN (Arjovsky et al. (2017)) adversarial domain adaptation (WGANDA) and successfully adopted it to build subject-independent emotion recognition models (Luo et al. (2018)).

In real-world aBCIs applications, each subject can be viewed as an individual domain. DA methods, which require high additional computation resources for each new domain, hinder the development of aBCIs from lab to real scenarios. DG methods, which can be utilized by data manipulation, representation learning, or meta-learning (Wang et al. (2021)), aim to generalize to unseeing target domains without additional data collection from target domains. Researchers have focused on adopting DG methods to aBCIs very recently. Ma *et al.* generalized the structure of DANN into DG and proposed an adversarial structure called Domain Residual Network (DResNet). They adopted DResNet to subject-independent EEG-based vigilance estimation and emotion recognition (Ma et al. (2019b;a)). The experimental result demonstrated that the DG method could improve the generalization ability without data collection from the target domains. However, the DA methods usually give relatively promising results than DG methods in aBCIs applications. As a compromise way, FDA was adopted in aBCIs. Zhao *et al.* proposed a Plug-and-Play Domain Adaptation (PPDA) method to fast adapt the model in EEG-based emotion recognition (Zhao et al. (2021)).

Meta-learning, also known as learning to learn, has received a resurgence in interest recently with applications, one of which is domain generalization. Meta-learning aims to learn episodes sampled from the related tasks (Finn et al. (2017)). Meta-Learning for Domain Generalization (MLDG) first introduces meta-learning strategy to DG (Li et al. (2018a)) , then MetaReg (Balaji et al. (2018)) and Feature-Critic (Li et al. (2019)) are subsequently proposed to enhance the model's generalizing ability by introducing an auxiliary loss in the training. Compared to most previous DG work that designs a specific model, meta-learning-based DG methods focus on model agnostic training strategy by exposing the model to domain shift in the training phase. To the best of our knowledge, we are the first to introduce meta-learning based methods to aBCIs tasks.

3 THEORY

The motivation of PDAML is using a simple network to compute domain shift, taking only the target domain as input. Traditional DA methods usually taking two domains as input when comparing the target domain with a specific source domain. Thus they need extra storage space for source data and sophisticated methods (e.g. GAN) to represent domain shift in the test phase, which is time-consuming and storage-consuming. Either or both of these problems obstruct the practical application of DA method in EEG-based diagnosis. However, in multi-source setting, we will show that minimizing the discrepancy between all pairwise domains is equivalent to minimizing the discrepancy between each domain and an implicit domain. Additionally, we will prove that a network with what is termed **sum-decomposition** form can represent any domain shift metrics in theory.

3.1 PROBLEM SETUP

Now we give a formal definition of the problem. Let $\mathcal{X}$ denote input EEG data space and $\mathcal{Y}$ denote output space. We define a domain $\mathcal{D}$ to be the joint distribution $\mathbb{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$. This distribution changes for many reasons and we assume it follows distribution $\mathscr{P}$. Domains cannot be observed directly. What we observe are samples $\{D_i\}$ of domains where each D_i denotes a set of $\{X_i, Y_i\}$.

The inconsistency of domains may cause poor generalization ability. One way to handle this is to use a functional $\mathcal{T}$ mapping each domain into another while reducing the discrepancy between new domains. Typically, a divergence loss function $d(\cdot, \cdot)$ is selected, which takes the marginal or joint distribution into consideration. And the final $\mathcal{T}$ is chosen by minimizing Equation (1),

$$\mathcal{T}_{da} = \arg\min_{\mathcal{T}} d(\mathcal{T}(D_i), \mathcal{T}(D_j)). \tag{1}$$

Since such functional $\mathcal{T}_{da}$ can transfer domains into a common feature space, the model trained on $\mathcal{T}(D)$ will not suffer domain shift problem. This method is called alignment.

3.2 SHIFT-FREE DOMAIN

In the multi-source DA or DG setting, an easy way to introduce DA methods is simultaneously minimizing divergence between every pairwise source domains Equation (2),

$$\mathcal{T}_{mda} = \arg\min_{\mathcal{T}} \sum_{D_i, D_j \in D_S} d(\mathcal{T}(D_i), \mathcal{T}(D_j)). \tag{2}$$

In the ideal situation, the final discrepancy $\sum_{D_i, D_j \in D_S} d(\mathcal{T}(D_i), \mathcal{T}(D_j))$ limits to zero. All domains will be the same. We define that final domain as **shift-free domain**.

Definition 1. *A shift-free domain u_{sf} is any $\mathcal{T}(D_i)$ in the limit that $\sum_{i \neq j} d(\mathcal{T}(D_i), \mathcal{T}(D_j)) \to 0$.*

Theorem 1. *in the limit, optimizing total discrepancy $\sum_{D_i, D_j \in D_S} d(\mathcal{T}(D_i), \mathcal{T}(D_j))$ is equivalent to optimizing a loss function $\sum_{D_i \in D_S} \mathcal{L}_{\mathcal{T}(D_i)}$, where $\mathcal{L}_{\mathcal{T}(D_i)} := d(\mathcal{T}(D_i), u_{sf})$.*

As shown in figure 1, aligning all the pairwise domains simultaneously is equivalent to aligning all single domains with the shift-free domain. As long as we derive the shift-free domain, we can construct a network to compute the shift of the target domain and use it to govern the pseudo domain adaptation. Next, we will show how to construct the network.

3.3 SUM-DECOMPOSITION

A key requirement for a function to represent domain discrepancy is the **permutation-invariant** constraints. That is to say, the order of the source domain data should be irrelevant to the output. This property has been well studied in previous works (Zaheer et al. (2017); Qi et al. (2017)). Usually, the summation is introduced to enforce permutation-invariance. This form is termed *sum-decomposition*. And the definition of sum-decomposable is shown in Definition 2.

Definition 2. *A function g is **sum-decomposable** via $\mathbb{R}^N$ if there exist function $\psi : \mathbb{R} \to \mathbb{R}^N$ and $\rho : \mathbb{R}^N \to \mathbb{R}$, such that $g(X) = \rho\left(\sum_{x \in X} \psi(x)\right)$.*

Theorem 2. *An continues map $F : \mathbb{R}^M \to \mathbb{R}$ is permutation invariant if and only if it is continuously sum-decomposable via $\mathbb{R}^M$.*

Proposition 1. *Traditional domain discrepancy like MMD or $\mathcal{H}$-divergence can be computed equivalently using a sum-decomposable function.*

The proof of Theorem 2 can be found in previous study (Wagstaff et al. (2019)). By adding a summation layer or averaging layer , one can easily enforce permutation-invariant property. Theorem 2 suggests that sum-decomposable network via a latent space with sufficient dimension should suffice to model any permutation-invariant function, including $d(\cdot, u_{sf})$ occurs in Theorem 1.

4 METHODOLOGY

(a) (b)

Figure 1: The sketch map of the total discrepancy between pairwise domains and discrepancy between each domain and the shift-free domain.

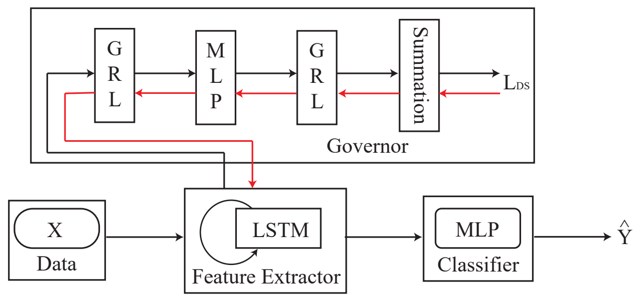


Figure 2: The working flow of PDAML. When a new domain comes, the feature extractor first changes the data to feature. Then the governor judges its discrepancy and computes loss L_{DS}. Under the governing of G, the feature extractor makes quick self-adaptation. After several adaptation steps, the data X will be passed to a new feature extractor and then forward propagate to the classifier.

4.1 Overview

As an implementation of the theory, our method PDAML couples a sum-decomposable domain shift governor with the task network. As illustrated in Figure 2, the task network is decomposed into a feature extractor $F(\theta)$ and a classifier $C(\phi)$. Domain shift governor $G(\omega)$ takes the feature from $F(\theta)$ to analyze the discrepancy between the current domain and the shift-free domain. When we use this network to classify some data of the target domain, G will first propagate forward to compute the domain shift, and then automatically propagate backward to fine-tune the feature extractor F. It seems like a smart governor who knows how to modify the network according to its performance on producing shift-free features. Under the governing of G, the entire network can generalize to any unseen domains via pseudo domain adaptation. In the rest of this section, we will introduce how to design a domain shift governor, and then introduce our training strategy.

4.2 Designing the governor

Since the governor needs to be permutation-invariant, an easy way to satisfy this requirement is to insert a summation layer into a neural network, as what Feature-Critic did. The Feature-Critic network simply adds a summation layer at the end of a multi-layer perceptron (MLP), omitting the outside map ρ in Definition 2. Hence it may not represent a domain-shift governor perfectly. Besides, we found in the experiment that introducing adversarial in the network can promote its capability. Specifically, we insert two **Gradient Reversal Layers** (GRL) before and after a two-layer MLP and add another layer after it.

The motivation for adding GRL into the network comes from traditional methods representing domains discrepancy. between two domains, like MMD or adversarial-based methods. Those methods all have one thing in common, that we should use the 'biggest' difference between two domains to represent the divergence.

To enforce our network the same capability of simulating domain shift like these traditional methods, we define a new divergence similar to MMD and h-divergence, which we term Maximum Mean Norm Discrepancy(MMND).

Definition 3. *The MMND between two domains D_i, D_j is defined as,*

$$MMND(D_i, D_j) := \max_{f \in \mathcal{T}} \left|\left|\mathbb{E}_{x \in D_i} f(x) - \mathbb{E}_{x \in D_j} f(x)\right|\right|_2, \tag{3}$$

where f is a function mapping x into a vector space.

According to Theorem(1), we can choose an implicit domain as the shift-free domain to avoid direct domain comparison. Intuitively, the implicit domain should be chosen to make the total divergence as small as possible, such that the optimization will be fast. So our proposed objective function contains both minimization and maximization as equation (4) shows,

$$\mathcal{L}_{ds} = \sum_{D_i \in D_S} \max_{f \in \mathcal{T}} \min_{\mu \in V_{latent}} ||\mathbb{E}_{x \in D_i} f(x) - \mu||_2 . \tag{4}$$

In domain shift governor, f is the inner map ψ in Equation (2), which is represented by the first several layers of the network. The summation layer computes the mean value of $f(x)$. Finally, the difference's norm will be computed by the last layer of the domain shift governor. As for maximization and minimization, we adopt Goodfellow et al. (2014) method that adding Gradient Reversal Layers in the governor network. The GRL works as an identity map during forwarding propagation but reverses the gradient direction in backward. Compared with Domain Adaptation Regularizer in DANN, our proposed domain shift governor has one more minimization task. Thus we need two GRL in total. As Figure 2 shows, the left GRL fools the governor to find out the biggest difference of domains, and the right GRL corrects the feature extractor to generate uniform features of different domains. This two-GRL structure could be unstable in experiments. To handle this, we freeze part of the network between two GRL when the iteration reaches a threshold.

4.3 TRAINING THE NETWORK

The next step is to learn all the parameters. To guarantee the generalization ability of the governor, we introduce the meta-learning strategy. The algorithm can be divided into training the governor and training the network. They are represented in Algorithm 1 and Algorithm 2, respectively.

Training the governor. To enhance the model's generalization ability, we not only optimize the governor's output $\mathcal{L}_{DS}$ to reduce domain shift but also couple it with the generalization process via meta-learning. First, all accessible domains are randomly split into meta-train domains D_{tr} and meta-test domains (also known as meta valid domains D_{val}). We use the governor to optimize the feature extractor $F(\theta)$ on meta-train domains, then evaluate the promotion of $F(\theta')$ on meta-test domains.

Let ℓ denote the classification loss function. After θ updates to θ', the loss function will change from $\ell(x^{(i)}, y^{(i)}; \theta)$ to $\ell(x^{(i)}, y^{(i)}; \theta')$. Following Li *et al.*'s work Li et al. (2019), we construct the meta loss functionin Equation (5),

$$\mathcal{L}_{meta} = \sum_{(x_i, y_i) \in D_{val}} \tanh\Big(\ell(x^{(i)}, y^{(i)}; \theta') - \ell(x^{(i)}, y^{(i)}; \theta)\Big). \tag{5}$$

Finally, the total loss for updating ω is shown in Equation (6),

$$\mathcal{L}_{DS}(\theta, \phi, \omega; D_{tr}) + \lambda \mathcal{L}_{meta}(\theta', \phi, \omega; D_{val}). \tag{6}$$

Here λ is a hyper-parameter, θ, ϕ, ω are parameters of F, C, G. By optimizing Equation(6), $G(\omega)$ finally comes to its real update.

Training the entire network. Unlike MetaReg or Feature-Critic that first trains the auxiliary network and then trains the task network, we propose to train our governor and task network alternately. Because we need G still working even in the test phase, which means we must keep updating G during the training of other parts of the network. Besides, to fully implement the principle of meta-learning, we use the MAML framework to train the network, instead of directly optimizing it. We view the governor and classification as two tasks, then use episode training to update their parameters.

Still, the domains are split into D_{tr} and D_{val}. G uses data from D_{tr} to compute the domain shift loss $\mathcal{L}_{DS}(\theta, \omega; D_{tr})$ and classification loss $\mathcal{L}_{ce}^{(tr)}$, which is further used to update the feature extractor $F(\theta)$. With the updated parameters, $F(\theta')$ take data from D_{val} and the task network computes a corresponding $\mathcal{L}_{ce}^{(val)}$. The total loss for optimizing $F(\theta)$ and $C(\phi)$ is as follows,

$$\lambda \mathcal{L}_{DS}(\theta, \omega; D_{tr}) + \mathcal{L}_{ce}^{(tr)}(\theta, \phi) + \mathcal{L}_{ce}^{(val)}(\theta', \phi). \tag{7}$$

Algorithm 1 Training the governor

Input: $\mathcal{D}, \theta, \phi, \omega$
Output: ω
Random Split $\mathcal{D}$:
$(\mathcal{D}_{tra}, \mathcal{D}_{val}) \leftarrow \mathcal{D}$
Meta train:
$\mathcal{L}_{DS}(\theta, \omega; D_{tr}) \leftarrow G(F(X))$
$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{DS}(\theta, \omega; D_{tr})$
Meta validate:
$\mathcal{L}_{meta}(\theta, \theta', \phi; D_{val}) \leftarrow$ Equation (5)
Meta optimization:
update ω using $\mathcal{L}_{DS} + \lambda \mathcal{L}_{meta}$

Algorithm 2 Training the entire network

Input: $\mathcal{D}, \theta, \phi, \omega$
Output: θ, ϕ, ω
Random Split $\mathcal{D}$
Meta train:
$\mathcal{L}_{DS}(\theta, \omega; D_{tr}) \leftarrow G(F(X))$
$\mathcal{L}_{ce}^{(tr)}(\theta, \phi) \leftarrow \ell(C(F(X_{tr})), Y_{tr})$
$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{DS}(\theta, \omega; D_{tr})$
Meta validate:
$\mathcal{L}_{ce}^{(val)}(\theta', \phi) \leftarrow \ell\left(C(F(X_{val})), Y_{val}\right)$
Meta optimization:
update θ, ϕ, ω using $\lambda \mathcal{L}_{DS} + \mathcal{L}_{ce}^{(val)} + \mathcal{L}_{ce}^{(tr)}$

Analysis in the view of meta-learning. The way we introduce PDAML is in the view of alignment-based domain adaptation. However, this method can also be explained in the view of meta-learning. According to Li et al. (2018a), meta-learning can be seen as coupling different tasks by making their gradients in the same direction. In our setting, the governor is a task artificially set to be coupled with the ”domain generalization” process’. When the model accesses a new domain, the optimization for the model to adapt to this new domain is in the same direction as the optimization promoted by the governor.

5 EXPERIMENTS

Datasets. We evaluate our proposed PDAML method on the SEED dataset (Zheng & Lu (2015)). The SEED dataset contains the EEG signals of 15 subjects. They were recruited to watch 15 well-prepared video clips that can elicit exactly one of the three kinds of emotion: happy, neutral, and sad. The criteria of film clip selection ensure that each clip is well-edited to create coherent emotion eliciting and maximize emotional meanings. The signals were sampled at the rate of 1000 Hz with ESI NeuroScan System from a 62-electrode headset.

Feature Extraction. We extract the same feature following the existing studies (Zheng & Lu (2015)). Since the SEED dataset has been preprocessed, we could directly extract the feature. Differential entropy (DE) feature (Duan et al. (2013)) has been extracted. Previous works have shown that the DE feature of EEG signals is efficient for EEG-based emotion recognition (Zheng & Lu (2015); Zheng et al. (2019); Yang et al. (2018)). Shi *et al.* have demonstrated that the value of DE is equal to the logarithmic spectral energy for a fixed-length EEG sequence in a certain band (Shi et al. (2013)). So we firstly use Short Time Fourier Transform (STFT) with a 1-s-long non-overlapping Hanning window to extract the spectral energy of EEG signal from five frequency bands: δ: 1-3 Hz, θ: 4-7 Hz, α: 8-13 Hz, β: 14-30 Hz, and γ: 31-50 Hz. Then we can calculate the DE feature. Taking account into the dynamic characteristics of EEG-based emotion recognition tasks, we employ the linear dynamic system (LDS) approach to filter the DE feature. The dimension of each sample is 310 (62 channels $\times$ 5 frequency bands). Since the EEG data are time series, we re-sampling the feature with a time-step of 15 and a 1-s-long overlapping, and each subject has 3184 samples.

Implementation Details. Following the previous work of Plug-and-play, we adopt a leave-one-subject-out strategy to study the generalization ability of PDAML. In each iteration, we select one subject as the target and use the other 14 subjects to train our model. In the test phase, we choose the prediction results after 10 steps of self-adaptation. The feature extractor of PDAML is an LSTM with 2 layers, output dimension 256, and time step 15. The classifier is an MLP with 2 layers, hidden size 100. We use Adam optimizer with learning rate 2e-4, weight decay 1e-4 for both the task network and governor. λ is set to be 0.1. First, we pre-train the task network to reach more than 85% accuracy in the training set. Then we use PDAML training the domain shift governor as well as the task network at the same time. The threshold for freezing the part of the governor inside GRL is 40, the max iteration is 200.

6 RESULT AND DISCUSSION

To evaluate our method, we adopt the leave-one-subject-out evaluation scheme and compare PDAML with various DA and DG methods on the SEED dataset. The results, including mean accuracy (avg.) and standard deviations (std.), are reported in Table 1. In comparison with the baseline of aggregating the data from all source domains and directly using a Support Vector Machine (SVM) to train a single model, all these methods show great improvement of the accuracy by at least 13%. Among them, PDAML outperforms all DG methods. When compared with the DA methods, PDAML still achieves the considerable results. Only WGANDA and PPDA have slightly higher accuracy than PDAML. But WGANDA needs all of the source domains and PPDA needs part of them to apply adaptation, which cannot compare PPDA's fast generalization ability.

Methods	TYPE	Avg.	Std.
SVM (Zheng & Lu (2016))	Baseline	0.567	0.163
DICA (Ma et al. (2019b))	DG	0.694	0.078
DResNet (Ma et al. (2019b))		0.853	0.080
PPDA_NC (Zhao et al. (2021))		0.854	0.071
MLDG (Li et al. (2018a))		0.795	0.120
Feature-Critic (Li et al. (2019))		0.806	0.120
TCA (Zheng & Lu (2016))	DA	0.640	0.146
TPT (Zheng & Lu (2016))		0.752	0.128
DANN (Li et al. (2018b))		0.792	0.131
DAN (Li et al. (2018b))		0.838	0.086
WGANDA (Luo et al. (2018))		0.871	0.071
PPDA ((Zhao et al. (2021))	Fast DA	0.867	0.071
PDAML (Ours)	Pseudo DA	0.864	0.094

Table 1: Results on the SEED dataset.

6.1 COMPARED WITH META-LEARNING BASED DG

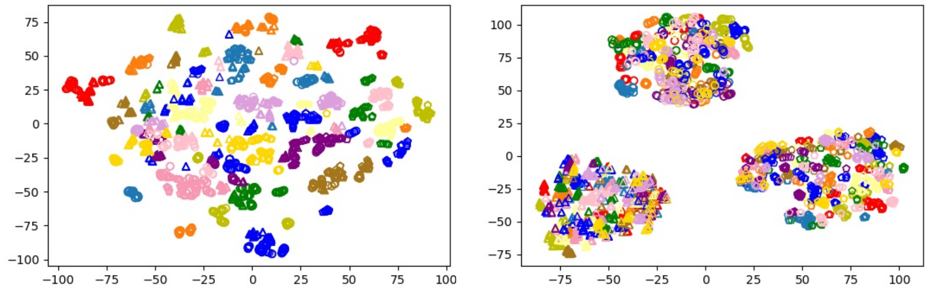
It should be noticed that as our method is an implementation of meta-learning strategy, its accuracy outperforms MLDG's and Feature-Critic's by 6.9% and 5.8%, respectively. Results demonstrate that a method like MLDG that directly adopts episode training to generalize the model is not competent to overcome the subject variability in EEG-based emotion recognition. Nevertheless, Feature Critic uses a sum-decomposable MLP to simulate domain shift in the training phase, but still does not improve the result significantly. It implies that applying domain shift governor in the test phase is useful. As a PDA method, our PDAML predicts any target set with only several steps of self-adaptation, combining the advantage of both the DA and DG.

6.2 VISUALIZATION

The domain shift governor is designed to perform a similar way to a conventional domain discriminator that helps feature extractor generate domain-invariant features. As shown in Figure 3, with the trained governor, the feature extractor can reduce the domain shift between data from each subject. Data with the same emotion label from different domains share a similar distribution in the common space. And this distribution is shift free domain.

6.3 FAST ADAPTATION

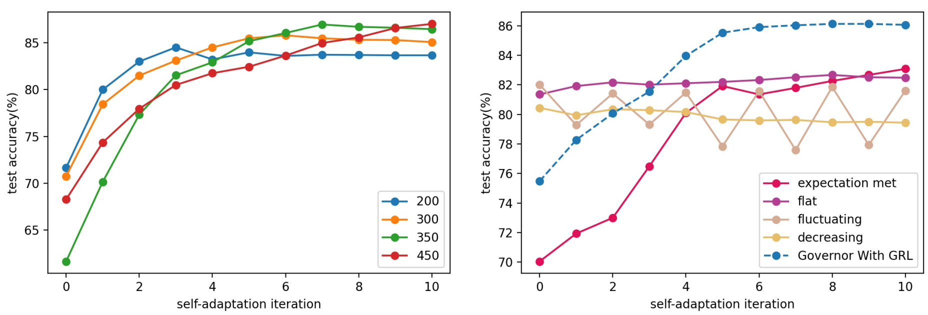
To evaluate the domain shift governor, we record the accuracy of the self-adaptation in the test phase. As shown in Figure 4, we see that PDAML can adapt to the a new domain stably, achieving good performance in just a few self-adapt steps. The left one is the adaptation performance on the target domain during the training phase. Each self-adaptation requires no extra information except the input data for prediction. Good performance is achieved within only a few self-adaptation steps.



(a) original domain (b) shift-free domain

Figure 3: Visualization of the data using t-SNE. Different colors stand for different subjects. Different shapes stand for different emotion labels.

The right one is the comparison between the governor with and without GRL. We can see that the introduce of the adversarial strategy makes the domain shift governor more stable and efficient, otherwise there may be fluctuating or decrease.



(a) (b)

Figure 4: The test accuracy of self-adaptation. (a) Results of standard PDAML; (b) Results when GRLs are removed.

7 Conclusion

In this work, we propose a PDA paradigm, in which no source domain is required in the test phase, for building subject-independent EEG-based affective model. As an implementation of PDA, we propose PDAML method with a sum-decomposable domain shift governor to make PDA. By taking advantage of adversarial learning and meta-learning, our PDAML generalizes to a new domain within only a few self-adaptive steps. Experimental results on the SEED dataset demonstrate that PDAML outperforms the DG methods and converges quickly, which is more suit for building subject-independent affective model than typical DA, DG and FDA methods.

REFERENCES

S. M. Alarcão and M. J. Fonseca. Emotions recognition using EEG signals: A survey. *IEEE Trans. Affective Computing*, 10(3):374–393, July 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. MetaReg: towards domain generalization using meta-regularization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 1006–1016, Red Hook, NY, USA, December 2018. Curran Associates Inc.

Shai Bendavid, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Xin Chai, Qisong Wang, Yongping Zhao, Yongqiang Li, Dan Liu, Xin Liu, and Ou Bai. A Fast, Efficient Domain Adaptation Technique for Cross-Domain Electroencephalography(EEG)-Based Emotion Recognition. *Sensors*, 17(5):1014, May 2017. doi: 10.3390/s17051014. URL `https://www.mdpi.com/1424-8220/17/5/1014`. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for EEG-based emotion classification. In *International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84. IEEE, 2013.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, July 2017. URL `http://proceedings.mlr.press/v70/finn17a.html`. ISSN: 2640-3498.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML'15*, volume 37, pp. 1180–1189, 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS'14*, pp. 2672–2680, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *NIPS'07*, pp. 513–520, 2007.

R. Jenke, A. Peer, and M. Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Trans. Affective Computing*, 5(3):327–339, July 2014.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to Generalize: Meta-Learning for Domain Generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018a. ISSN 2374-3468. URL `https://ojs.aaai.org/index.php/AAAI/article/view/11596`. Number: 1.

He Li, Yi-Ming Jin, Wei-Long Zheng, and Bao-Liang Lu. Cross-subject emotion recognition using deep adaptation networks. In *Neural Information Processing*, pp. 403–413, 2018b.

He Li, Wei-Long Zheng, and Bao-Liang Lu. Multimodal vigilance estimation with adversarial domain adaptation networks. In *IJCNN*, pp. 1–6, 07 2018c.

Xiang Li, Dawei Song, Peng Zhang, Yazhou Zhang, Yuexian Hou, and Bin Hu. Exploring EEG features in cross-subject emotion recognition. *Frontiers in Neuroscience*, 12:162, 03 2018d. doi: 10.3389/fnins.2018.00162.

Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3915–3924. PMLR, 09–15 Jun 2019.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML'17*, pp. 2208–2217, 2017.

Yun Luo, Si-Yang Zhang, Wei-Long Zheng, and Bao-Liang Lu. WGAN domain adaptation for EEG-based emotion recognition. In *Neural Information Processing*, pp. 275–286, 2018.

Bo-Qun Ma, He Li, Yun Luo, and Bao-Liang Lu. Depersonalized cross-subject vigilance estimation with adversarial domain generalization. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019a. doi: 10.1109/IJCNN.2019.8852347.

Bo-Qun Ma, He Li, Wei-Long Zheng, and Bao-Liang Lu. Reducing the subject variability of eeg signals with adversarial domain generalization. In *Neural Information Processing*, pp. 30–42, Cham, 2019b. Springer International Publishing.

Christian Mühl, Brendan Allison, Anton Nijholt, and Guillaume Chanel. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 1 (2):66–84, 2014.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Networks*, 22(2):199–210, 2011.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://papers.nips.cc/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html`.

Wojciech Samek, Frank C. Meinecke, and Klaus-Robert Muller. Transferring subspaces between subjects in brain–computer interfacing. *IEEE Trans. Biomedical Engineering*, 60(8):2289–2298, 2013.

Maryam Shanechi. Brain–machine interfaces from motor to mood. *Nature Neuroscience*, 22:1554–1564, 10 2019. doi: 10.1038/s41593-019-0488-y.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *AAAI'18*, pp. 4058–4065, 2018.

Li-Chen Shi, Ying-Ying Jiao, and Bao-Liang Lu. Differential entropy feature for EEG-based vigilance estimation. *International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013 (2013):6627–6630, 2013.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *JMLR*, 8:985–1005, 2007.

David Sussillo, Sergey D Stavisky, Jonathan C Kao, Stephen I Ryu, and Krishna V Shenoy. Making brain–machine interfaces robust to future neural variability. *Nature Communications*, 7(1):13749, 2016.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR'17*, pp. 7167–7176, 2017.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A. Osborne. On the Limitations of Representing Functions on Sets. In *International Conference on Machine Learning*, pp. 6487–6494. PMLR, May 2019. URL `http://proceedings.mlr.press/v97/wagstaff19a.html`. ISSN: 2640-3498.

Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *ACMMM'18*, pp. 402–410, 2018.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4627–4635, 8 2021. doi: 10.24963/ijcai.2021/628.

Yimin Yang, Q M Jonathan Wu, Wei-long Zheng, and Bao-liang Lu. EEG-based emotion recognition using hierarchical network with subnetwork nodes. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2):408–419, 2018.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan R. Salakhutdinov, and Alexander J. Smola. Deep Sets. January 2017. URL https://openreview.net/forum?id=HJbcbv-_WH.

T. O. Zander and S Jatzev. Context-aware brain-computer interfaces: exploring the information space of user, technical system and environment. *Journal of Neural Engineering*, 9(1):016003, 2011.

Li-Ming Zhao, Xu Yan, and Bao-Liang Lu. Plug-and-Play Domain Adaptation for Cross-Subject EEG-based Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):863–870, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16169.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. AMD*, 7(3):162–175, 2015.

Wei-Long Zheng and Bao-Liang Lu. Personalizing EEG-based affective models with transfer learning. In *IJCAI'16*, pp. 2732–2738, 2016.

Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. Identifying stable patterns over time for emotion recognition from eeg. *IEEE Trans. Affective Computing*, 10(3):417–429, July 2019. ISSN 2371-9850. doi: 10.1109/TAFFC.2017.2712143.

Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: transfer learning with deep autoencoders. In *IJCAI'15*, pp. 4119–4125, 2015.

A Appendix

You may include other additional sections here.