Side Information Dependence as a Regularizer for Analyzing Human Brain Conditions across Cognitive Experiments

Shuo Zhou,[1] **Wenwen Li,**[2] **Christopher R. Cox,**[3] **Haiping Lu**[1,4]
[1]Department of Computer Science, University of Sheffield, United Kingdom
[2]Centre for Clinical Brain Sciences, University of Edinburgh, United Kingdom
[3]Department of Psychology, Louisiana State University, United States
[4]Sheffield Institute for Translational Neuroscience, United Kingdom
{szhou20, h.lu}@sheffield.ac.uk, wenwen.li@ed.ac.uk, chriscox@lsu.edu

Abstract

The increasing of public neuroimaging datasets opens a door to analyzing homogeneous human brain conditions across datasets by transfer learning (TL). However, neuroimaging data are high-dimensional, noisy, and with small sample sizes. It is challenging to learn a robust model for data across different cognitive experiments and subjects. A recent TL approach minimizes domain dependence to learn common cross-domain features, via the *Hilbert-Schmidt Independence Criterion* (HSIC). Inspired by this approach and the multi-source TL theory, we propose a *Side Information Dependence Regularization* (SIDeR) learning framework for TL in brain condition decoding. Specifically, SIDeR simultaneously minimizes the empirical risk and the statistical dependence on the domain side information, to reduce the theoretical generalization error bound. We construct 17 brain decoding TL tasks using public neuroimaging data for evaluation. Comprehensive experiments validate the superiority of SIDeR over ten competing methods, particularly an average improvement of 15.6% on the TL tasks with multi-source experiments.

Introduction

In cognitive neuroscience, neuroimaging can help relate different cognitive functions to patterns of neural activity using functional magnetic resonance imaging (fMRI) (Ogawa et al. 1990). This often takes the form of a classification problem (Cox and Savoy 2003), e.g., distinguishing between *brain conditions* associated with experimental stimuli. While fMRI produces volumes with the number of voxels in the order of 10^5, a typical experiment will have on the order of 100 discrete trials. This severely constrains the number of training examples available for the classifier. Moreover, neuroimaging data are noisy and contain a significant amount of physiological, respiratory, and mechanical artifacts, which requires robust modeling against noise (Aydore, Thirion, and Varoquaux 2019).

Transfer learning (TL) is an attractive machine learning scheme that can improve the classification performance on a learning task by leveraging the knowledge from related tasks. The task of interest is called the *target domain*, while the task(s) to be leveraged is called the *source domain* (Pan et al. 2011). A TL problem is *homogeneous* when the feature and label space of the source and target domains are the same, and *heterogeneous* if they are different.

Transfer learning techniques have been studied in some fMRI applications. Mensch et al. (2017) take a multi-task learning approach to use *resting-state data* from the Human Connectome Project (Van Essen et al. 2012) to learn a general representation via matrix decomposition and then jointly optimize multiple *heterogeneous* task-based fMRI classification tasks. Zhang, Chen, and Ramadge (2018) take a matrix factorization approach that relies on *shared subjects* between datasets to learn better subject factor matrices. Deep models such as autoencoder (Velioglu and Vural 2017; Li, Parikh, and He 2018) and AlexNet (Zhang et al. 2019) pre-trained on generic source data have also been used to represent the target fMRI data for classification.

However, the source data used by the existing fMRI TL studies are independent to the target classification task. While public neuroimaging data from multiple sites, e.g., the OpenNeuro (Gorgolewski et al. 2017), have many similar brain conditions across different cognitive experiments. This enables homogeneous TL studies to leverage the power of overlapping labels across domains. Furthermore, it may potentially offer interpretation/insights from the domain shift perspective for neuroscientists. Here, we make the first attempt, to the best of our knowledge, to investigate *homogeneous TL for brain condition decoding*.

Homogeneous TL methods are mainly studied in the fields of computer vision (CV) and natural language processing (NLP). They focus on minimizing data distribution mismatch, i.e., making features from the source and target to have as similar distributions as possible, via 1) learning a feature mapping (Pan et al. 2011; Long et al. 2013b); or 2) jointly optimizing the distribution mismatch and classifier parameters (Long et al. 2013a; Wang et al. 2018).

Brain condition decoding presents TL challenges different from those in CV/NLP. fMRI data are generated by brain signals, which are not natural images that human visual system has adapted to interpret. Consequently, fMRI analysis relies heavily on statistics. Furthermore, cognitive stimuli are implemented varying across experiments. Even

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

the same information can be encoded as different patterns of activity by different brains (Chen et al. 2015). Hence, each subject can be considered as a unique learning task to extract subject-specific features (Rao et al. 2013) in fMRI studies. Additionally, as mentioned before, fMRI data are noisy. Therefore, for TL in brain condition decoding, it can be beneficial to take more domain information, such as experiment designs and subjects, into account to learn a robust model.

Recently, the *maximum independence domain adaptation* (MIDA) (Yan, Kou, and Zhang 2018) introduces a new domain dependence minimization approach to TL. It learns common, cross-domain features by minimizing statistical dependence on auxiliary domain side information, as measured by the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. 2005). This inspired us to encode different experiment designs and subjects as auxiliary domain covariates for TL in brain condition decoding.

In this paper, we propose a *Side Information Dependence Regularization* (SIDeR) framework for homogeneous TL in brain condition decoding. The contributions are threefold: **(1)** We discover the relationship between HSIC and maximum mean discrepancy (MMD) and derive two HSIC-based generalization bound for and multi-source TL. The theoretical studies enable the formulation of the SIDeR framework that simultaneously minimizes the empirical prediction risk and the dependence on domain side information. **(2)** Under this framework, we construct a simplified HSIC and incorporate the hinge loss that can take unlabeled samples into account following the Manifold Regularization framework formulation (Belkin, Niyogi, and Sindhwani 2006). This gives us the SIDeR$_{\text{SVM}}$ algorithm. **(3)** We construct 17 new homogeneous brain decoding TL tasksby identifying datasets with homogeneous brain conditions from public repositories. Experiments on these tasks show the superior performance of SIDeR over ten competing methods.

Preliminaries

Hilbert-Schmidt Independence Criterion (HSIC) is a non-parametric criterion for measuring the statistical dependence between two sets $\mathbf{X} = \{\mathbf{x}_i\}$ and $\mathbf{Y} = \{\mathbf{y}_i\}$, both with size n. HSIC tests whether $\Pr(\mathbf{x}, \mathbf{y}) = \Pr(\mathbf{x}) \Pr(\mathbf{y})$. Denoting the empirical HSIC as $\rho_h(\mathbf{X}, \mathbf{Y})$, it can be computed via (Gretton et al. 2005)

$$\rho_h(\mathbf{X}, \mathbf{Y}) = \text{tr}(\mathbf{KHLH})/(n-1)^2, \tag{1}$$

where $\mathbf{K}, \mathbf{H}, \mathbf{L} \in \mathbb{R}^{n \times n}$, $\mathbf{K}_{i,j} := k_x(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{L}_{i,j} := k_y(\mathbf{y}_i, \mathbf{y}_j)$, $k_x(\cdot, \cdot)$ and $k_y(\cdot, \cdot)$ are two kernel functions, e.g., linear, polynomial, or radial basis function (RBF), $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix, and tr $(\cdot)$ is the trace function. HSIC is zero if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent. A larger HSIC value suggests stronger statistical dependence.

Related works. We summarize the state-of-the-art homogeneous TL methods as the following three approaches.

Distribution mismatch minimization mapping is a popular approach to homogeneous TL, which learns a mapping via minimizing the marginal or joint distribution mismatch between a source and a target, e.g., as in *Transfer Component Analysis* (TCA) (Pan et al. 2011) and *Joint Distribution Adaptation* (JDA) (Long et al. 2013b). The distribution mismatch is typically measured by the *maximum mean discrepancy* (MMD) criterion (Borgwardt et al. 2006). TCA also has a semi-supervised version, semi-supervised TCA (SSTCA), that introduces an additional label dependence objective $\rho_h(\phi(\mathbf{X}^l), \mathbf{Y})$ to maximize, where $\mathbf{X}^l$ denotes labeled data, and $\mathbf{Y}$ is a label matrix.

Domain-invariant classifier is another approach that learns a classifier by optimizing the prediction loss and distribution mismatch jointly. Long et al. (2013a) proposed Adaptation Regularization based Transfer Learning (ARTL) framework by incorporating joint distribution mismatch (as in JDA) into the manifold regularization framework (Belkin, Niyogi, and Sindhwani 2006). Based on ARTL, Wang et al. (2018) proposed *Manifold Embedded Distribution Alignment* (MEDA) by introducing a trade-off between marginal and conditional distribution mismatch for dynamic transfer.

Domain dependence minimization mapping. Yan, Kou, and Zhang (2018) proposed the MIDA method using a new approach that extracts cross-domain features by learning a mapping to minimize the dependence on domain information, e.g., device and time, which is not directly modeled in the previous two approaches. There is also a semi-supervised version of MIDA, i.e., SMIDA, which maximizes the label dependence (as in SSTCA).

Proposed Method

This section first defines the transfer learning problem. Then we perform theoretical studies on HSIC to reveal the relationships between HSIC and MMD. This enables us to derive a generalization bound for the multi-source TL setting. Subsequently, the bound motivate the formulation of the Side Information Dependence Regularization (SIDeR) learning framework for homogeneous TL.

Problem Definition

In a cognitive experiment, each subject is presented a set of stimuli (conditions) designed by neuroscientists. An experiment typically features one or a few (if repeated) samples per condition per subject. We consider a target dataset (experiment) have both labeled and unlabeled samples, and there are labeled samples with the homogeneous brain conditions that acquired from one or more source experiments, where the experiment designs are different. The objective is to predict the human brain conditions of *unlabeled target samples*.

The target cognitive experiment has n_t fMRI data samples $\mathbf{X}_t = [\mathbf{X}_t^l, \mathbf{X}_t^u] \in \mathbb{R}^{d \times n_t}$ of m brain conditions for classification. $\mathbf{X}_t^l \in \mathbb{R}^{d \times \tilde{n}_t}$ and $\mathbf{X}_t^u \in \mathbb{R}^{d \times (n_t - \tilde{n}_t)}$ are labeled and unlabeled target data, respectively, d is the number of fMRI features, e.g., voxels.

The source consists of data from one or more cognitive experiments with n_s labeled samples $\mathbf{X}_s \in \mathbb{R}^{d \times n_s}$ in total, with the same m brain conditions as the target data.

Domain covariate encoding. Denote the target and source data jointly as $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_t] \in \mathbb{R}^{d \times n}$, $n = n_s + n_t$. Each fMRI sample $\mathbf{x}_i$ ($i = 1, \cdots, n$) is collected with a particular experiment implementation j from a particular subject k, where $j = 1, \cdots, p$ and $k = 1, \cdots, q$, i.e., there are p unique experiment implementations and q unique

subjects. These are the *domain covariates* to be utilized in our TL method. We use a simple *one-hot-encoding* strategy to encode such domain covariates. Specifically, we construct a one-hot *experiment implementation* covariate matrix $\mathbf{E} \in \mathbb{R}^{n\times p}$, where its (i,j)th element $e_{i,j} = 1$ if $\mathbf{x}_i$ is collected from experiment j and $e_{i,j} = 0$ otherwise. Similarly, we construct a one-hot *subject* covariate matrix $\mathbf{S} \in \mathbb{R}^{n\times q}$, where $s_{i,k} = 1$ if $\mathbf{x}_i$ is from subject k and $s_{i,k} = 0$ otherwise. We then obtain the auxiliary domain covariate matrix $\mathbf{D} \in \mathbb{R}^{\hat{d}\times n}$ by concatenating $\mathbf{E}^\top$ and $\mathbf{S}^\top$, where $\hat{d} = p + q$.

Multi-source view of brain decoding. As mentioned in Sec. Introduction, each cognitive experiment can be designed differently, and each subject can encode the stimuli differently, i.e., $P(\mathbf{X}|E_i) \neq P(\mathbf{X}|E_j)$, and $P(\mathbf{X}|S_i) \neq P(\mathbf{X}|S_j)$, E and S denote an experiment and a subject, respectively. Traditional TL methods consider two different datasets (experiments in brain decoding) as different domains. If we also consider each subject as a domain as in (Rao et al. 2013), then each unique experiment-subject combination is a domain, i.e., brain decoding TL tasks is essentially a multi-source transfer problem. Therefore, in the following, we study HSIC in the multi-source TL setting.

Theoretical Studies on HSIC

The domain dependence of the data can be computed via $\rho_h(\mathbf{X}, \mathbf{D})$, using the domain covariate matrix $\mathbf{D}$ defined above. Now we show when using *one-hot encoding* and a *linear kernel* for $\mathbf{D}$ in HSIC, we can derive an equivalence between HSIC and MMD, as shown in the following lemma. Based on this lemma, we can derive generalization bounds for HSIC-based TL and formulate our new framework.

Lemma 1. *HSIC is proportional to MMD when there are only two discrete domains, i.e., with a degenerated one-hot domain covariate vector, e.g.,* $\mathbf{d}_0 = [\overbrace{0\cdots0}^{n_s}\overbrace{1\cdots1}^{n_t}] \in \mathbb{R}^n$, *and linear kernel is used for* $\mathbf{d}_0$ *in HSIC.*

Proof. The MMD between the two domains is

$$\text{MMD}(\mathbf{X}_s, \mathbf{X}_t) = \left\| \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbf{x}_{s_i} - \frac{1}{n_t}\sum_{i=1}^{n_t} \mathbf{x}_{t_i} \right\|_{\mathcal{H}_k}^2, \quad (2)$$

where $\mathcal{H}_k$ denotes a reproducing kernel Hilbert space (RKHS). The empirical MMD can be computed via $\text{tr}(\mathbf{KL}')$ (Pan et al. 2011), where $\mathbf{K} = k(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n\times n}$, $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_t] \in \mathbb{R}^{d\times n}$, and $\mathbf{L}' \in \mathbb{R}^{n\times n}$ is defined as

$$\mathbf{L}'_{ij} = \begin{cases} \frac{1}{{n_s}^2} & \text{if} \quad \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_s, \\ \frac{1}{{n_t}^2} & \text{if} \quad \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_t, \\ -\frac{1}{n_s n_t} & \text{otherwise.} \end{cases} \quad (3)$$

By Eq. (1), $\rho_h(\mathbf{X}, \mathbf{d}_0) = \text{tr}(\mathbf{KHLH})/(n-1)^2$, where $\mathbf{K} = k(\mathbf{X}, \mathbf{X})$ is exactly the same kernel matrix as in the MMD, and $\mathbf{L} = \mathbf{d}_0^\top \mathbf{d}_0$, i.e., $\mathbf{L}_{i,j} = 1$, if $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_t$, and otherwise $\mathbf{L}_{i,j} = 0$. Let $\hat{\mathbf{L}} = \mathbf{HLH}$, resulting

$$\hat{\mathbf{L}}_{ij} = \begin{cases} \frac{{n_t}^2}{n^2} & \text{if} \quad \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_s, \\ \frac{{n_s}^2}{n^2} & \text{if} \quad \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_t, \\ -\frac{n_s n_t}{n^2} & \text{otherwise.} \end{cases} \quad (4)$$

By comparing Eq. (3) and Eq. (4), we have

$$\text{MMD}(\mathbf{X}_s, \mathbf{X}_t) = u\rho_h(\mathbf{X}, \mathbf{d}_0), \quad (5)$$

where $u = \frac{n^2(n-1)^2}{(n_s n_t)^2}$. For a learning task, u is a constant. This completes the proof. □

Generalization bound. For the multi-source setting, we define a domain j as the samples $\mathbf{X}_j$ drawn from a distribution D_j on the inputs $\mathcal{X}$ and a labeling function $f_j : \mathcal{X} \to \{0,1\}$. A hypothesis $f \in \mathcal{H}$ is a function $f : \mathcal{X} \to \{0,1\}$. We consider a classifier f trained on a total of n_s samples $\mathbf{X}_s$ that drawing from J ($J \geq 1$) distinct source domains with a domain weight vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_J]$, where $\sum_{j=1}^J \alpha_j = 1$, and derive the bounds on its generalization performance on a target domain, i.e., $\epsilon_t(f)$ or $\epsilon_t(f, f_t)$.

By the proof of Theorems 4 and 5 in (Ben-David et al. 2010), $\epsilon_t(f) \leq \hat{\epsilon}_s(f) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{X}_s, \mathbf{X}_t) + \Lambda^* + \mathcal{O}$, where $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{X}_s, \mathbf{X}_t) = 2\sup_{f,f'\in\mathcal{H}} |\hat{\epsilon}_j(f, f') - \hat{\epsilon}_t(f, f')|$ is the empirical symmetric $\mathcal{H}$-divergence, Λ^* is the risk of an ideal joint hypothesis, and $\mathcal{O}$ denotes the complexity of hypothesis space. Here we derive a bound for Λ^* as a lemma first and then an HSIC-based bound for the multi-source setting.

Lemma 2. *Let* $\mathcal{H}$ *be a hypothesis space,* $\alpha_{J+1} = 1$, *and* $D_{J+1} = D_t$, *for* $j \in \{1, \ldots, J+1\}$, *let* $\mathbf{X}_j$ *be samples drawing from* D_j *with domain weight* α_j *and labeled by function* f_j, *let* $f^* = \arg\min_{f\in\mathcal{H}} \sum_{j=1}^{J+1} \alpha_j \epsilon_j(f)$ *to be the ideal joint hypothesis,* $\Lambda^* = \sum_{j=1}^{J+1} \alpha_j \epsilon_j(f^*)$, *then*

$$\Lambda^* \leq \frac{1}{2}\sum_{j=1}^{J+1} \alpha_j \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{X}_j, \mathbf{X}_{\alpha\setminus j}) + \Omega, \quad (6)$$

where $\mathbf{X}_\alpha$ *is the mixture of* $\mathbf{X}_s$ *and* $\mathbf{X}_t$, $\mathbf{X}_{\alpha\setminus j}$ *denotes exclude* $\mathbf{X}_j$ *from* $\mathbf{X}_\alpha$, *and* $\Omega = \sum_{j=1}^{J+1} \alpha_j \epsilon_{\alpha\setminus j}(f^*, f_j) + \mathcal{O}$.

Proof. For any $j \in \{1, \ldots, J+1\}$

$$\begin{aligned} \epsilon_j(f^*) &= \epsilon_j(f^*) + \epsilon_{\alpha\setminus j}(f^*, f_j) - \epsilon_{\alpha\setminus j}(f^*, f_j) \\ &\leq |\epsilon_j(f^*, f_j) - \epsilon_{\alpha\setminus j}(f^*, f_j)| + \epsilon_{\alpha\setminus j}(f^*, f_j) \\ &\leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_{\alpha\setminus j}) + \epsilon_{\alpha\setminus j}(f^*, f_j) \\ &\leq \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{X}_j, \mathbf{X}_{\alpha\setminus j}) + \mathcal{O}_j + \epsilon_{\alpha\setminus j}(f^*, f_j), \end{aligned} \quad (7)$$

then $\Lambda^* = \sum_{j=1}^{J+1} \alpha_j \epsilon_j(f^*) \leq \sum_{j=1}^{J+1} \alpha_j \epsilon_{\alpha\setminus j}(f^*, f_j) + \frac{1}{2}\sum_{j=1}^{J+1} \alpha_j \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{X}_j, \mathbf{X}_{\alpha\setminus j}) + \mathcal{O}$. □

Theorem 1 (Multi-source)**.** *Let* $\mathcal{H}$ *be a hypothesis space,* $\alpha_{J+1} = 1$, *and* $D_{J+1} = D_t$, *for* $j \in \{1, \ldots, J\}$, *let* $\mathbf{X}_j$ *be labeled samples of size* n_j *drawn from* D_j *with domain weight* α_j *and labeled by function* f_j. *Let* $\mathbf{D}$ *be a one-hot domain covariate matrix, then for* $f \in \mathcal{H}$:

$$\epsilon_t(f) \leq \hat{\epsilon}_s(f) + \frac{3}{2}\rho_h(\mathbf{X}, \mathbf{DU}) + \Omega, \quad (8)$$

where $\hat{\epsilon}_s(f)$ *is the empirical risk of* f *on the source data,* $\mathbf{U} = \text{diag}(\mathbf{u})$, $\text{diag}(\cdot)$ *is the diagonal function,* $\mathbf{u} \in \mathbb{R}^n$ *is a vector,* $u_i = \alpha_i n^2(n-1)^2/(n_j^2(n-n_j)^2)$, *if* $\mathbf{x}_i \in \mathbf{X}_j$, $i = 1, \ldots, n$, *and* $\Omega = \sum_{j=1}^{J+1} \alpha_j \epsilon_{\alpha\setminus j}(f^*, f_j) + \mathcal{O}$.

Proof. By the theoretical results in (Ben-David et al. 2010) mentioned above and Lemma 2, we have

$$\begin{aligned}\epsilon_t(f) &\leq \hat{\epsilon}_s(f) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{X}_{\alpha\backslash(J+1)}, \mathbf{X}_{(J+1)}) \\ &\quad + \frac{1}{2}\sum_{j=1}^{J+1}\alpha_j \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{X}_j, \mathbf{X}_{\alpha\backslash j}) + \Omega \\ &\leq \hat{\epsilon}_s(f) + \frac{3}{2}\sum_{j=1}^{J+1}\alpha_j \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{X}_j, \mathbf{X}_{\alpha\backslash j}) + \Omega.\end{aligned} \tag{9}$$

Empirical $\mathcal{H}$-divergence can be estimated by MMD as in the existing TL studies, e.g., (Long et al. 2013a). Hence, by Lemma 1, we have

$$\begin{aligned}\epsilon_t(f) &\leq \hat{\epsilon}_s(f) + \frac{3}{2}\sum_{j=1}^{J+1} u_j \mathrm{tr}(\mathbf{K}\mathbf{H}\mathbf{L}_j\mathbf{H}) + \Omega \\ &= \hat{\epsilon}_s(f) + \frac{3}{2}\rho_h(\mathbf{X}, \mathbf{D}\mathbf{U}) + \Omega,\end{aligned} \tag{10}$$

where $\mathbf{L}_j = \mathbf{d}_j^\top \mathbf{d}_j$, $\mathbf{d}_j \in \mathbb{R}^n$ is the jth row of $\mathbf{D}$, e.g., $\mathbf{d}_1 = [\overbrace{1\cdots1}^{n_1}\overbrace{0\cdots0}^{n-n_1}]$. This completes the proof. □

The Framework

Our ultimate goal is to learn a classifier for the unlabeled target data. From Theorem 1, the bound of $\epsilon_t(f)$ can be decreased by simultaneously minimizing **1)** the empirical error on labeled data, and **2)** the dependence on domain covariates. This observation enables us to propose a new Side Information Dependence Regularization (SIDeR) learning framework that optimizes these two objectives. Here, we follow the Manifold Regularization framework that can take unlabeled samples into account and formulate SIDeR as

$$\min_f \ \mathcal{L}(f(\mathbf{X}^l), \mathbf{Y}) + \sigma\|f\|_K^2 + \lambda\rho_h(f(\mathbf{X}), \mathbf{D}), \tag{11}$$

where $\sigma, \lambda \geq 0$ are hyper-parameters, $\mathbf{X}^l \in \mathbb{R}^{d\times\tilde{n}}$ denotes all labeled samples, $f(\cdot)$ is the decision function of a classifier, $\|f\|_K^2$ is the Tikhonov regularization term, and $\mathbf{Y}$ denotes training labels. For each term in SIDeR framework, $\mathcal{L}(f(\mathbf{X}^l), \mathbf{Y})$ minimizes the empirical risk, $\|f\|_K^2$ minimizes the model complexity, and $\rho_h(f(\mathbf{X}), \mathbf{D})$ minimizes the domain dependence in the label decision space.

Connection to existing methods. SIDeR minimizes prediction error and domain dependence simultaneously, and therefore it can also be viewed as combining the virtues from both domain-invariant classifier methods and domain dependence minimization mapping. We summarize the relationship between SIDeR and related methods as follows.

SIDeR vs. ARTL. By Lemma 1, ARTL without manifold regularization and conditional distribution mismatch is equivalent to SIDeR with the degenerated domain covariate matrix $\mathbf{D}_0$. However, SIDeR can model multiple sources and domain covariates, making it more flexible than ARTL. Moreover, it is easier to extend SIDeR to leverage the rich continuous side information in public neuroimaging dataset, such as subjects' age, IQ, and handiness score. For the same reason, TCA is equivalent to MIDA with $\mathbf{d}_0$.

Algorithm 1 Side Information Dependence Regularization (SIDeR) with SVM Loss

Input: Input data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ (first $\tilde{n}$ samples are labeled), label vector $\mathbf{y} \in \mathbb{R}^{\tilde{n}}$, and domain covariates.
Hyper-parameters: Penalty C, trade-off parameter λ, kernel function $k_x(\cdot,\cdot)$ and corresponding hyper-parameters.
Output: Coefficient vector $\mathbf{w}$.

1: Encode domain covariates into a matrix $\mathbf{D} \in \mathbb{R}^{\hat{d}\times n}$ with one-hot encoding;
2: Construct matrix $\tilde{\mathbf{Y}} \in \mathbb{R}^{\tilde{n}\times n}$, where $\tilde{\mathbf{Y}}_{i,i} = \mathbf{y}_i$, and the rest are zeros, identity matrix $\mathbf{I}$, and centering matrix $\mathbf{H}$;
3: Construct kernel matrices $\mathbf{K} = k_x(\mathbf{X}, \mathbf{X})$, $\mathbf{L} = \mathbf{D}^\top\mathbf{D}$;
4: Learn the optimal Lagrange multipliers $\mathbf{a}^*$ by solving the QP problem of Eq. (14);
5: Compute $\mathbf{w} = (\mathbf{I} + \lambda\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{K})^{-1}\tilde{\mathbf{Y}}^\top\mathbf{a}^*$.
6: **return** Coefficient vector $\mathbf{w}$.

SIDeR vs. SMIDA. We can also view SIDeR as 1) replacing the label dependence term $\rho_h(\phi(\mathbf{X}), \mathbf{Y})$ in SMIDA with the prediction loss, and 2) learning a mapping to a *one-dimensional* classification space (i.e., a line) rather than a low-dimensional subspace.

SIDeR vs. Multi-task learning (MTL). MTL learns N different hypotheses for predicting unlabeled samples from N tasks. SIDeR learns only one hypothesis (homogeneous task) for predicting unlabeled target samples.

Proposed Algorithm

Simplified HSIC. In SIDeR framework, we aim to optimize the domain dependence in the decision space. If we view the coefficient vector $\mathbf{w}$ as a *classifier-based feature mapping*, this mapping projects input features to a one-dimensional space (i.e., a line), where the projected values represent the decision scores. Following the principle of dependence minimization, we aim to learn a domain-independent classifier by minimizing the dependence of the decision scores (projected values) on domain side information, i.e., experiment implementations and subjects. By the Representer Theorem (Schölkopf, Herbrich, and Smola 2001), we can simplify the HSIC $\rho_h(f(\mathbf{X}), \mathbf{D})$ to the following version

$$\begin{aligned}\rho_{sh}(f(\mathbf{X}), \mathbf{D}) &= \mathrm{tr}((\mathbf{w}^\top\mathbf{K})^\top(\mathbf{w}^\top\mathbf{K})\mathbf{H}\mathbf{L}\mathbf{H}) \\ &= \mathbf{w}^\top\mathbf{K}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{K}\mathbf{w},\end{aligned} \tag{12}$$

where $\mathbf{L} = \mathbf{D}^\top\mathbf{D}$ (linear kernel) according to Lemma 1.

SIDeR with SVM loss. We can plug in any loss function for the first term in SIDeR of Eq. (11), such as the square loss, logistic loss, or hinge loss. In this paper, we consider only binary classification with $\mathbf{y} \in \mathbb{R}^{\tilde{n}}$, $y_i \in \{-1, 1\}$, $i = 1, \cdots \tilde{n}$, i.e., decoding $m = 2$ brain conditions. Here we choose hinge loss, which is robust to binary classification problems, for empirical risk minimization as in support vector machines (SVMs). We define $f(\mathbf{X}) = \mathbf{w}^\top\phi(\mathbf{X})$, ϕ is a linear or non-linear kernel mapping, $\mathbf{w}$ is a coefficient vector, $y = \mathrm{sgn}(f(\mathbf{x}))$, where $\mathrm{sgn}(\cdot)$ is the sign function that extracts the sign of a real number, i.e., (1 or -1). Using the Representer Theorem

Table 1: Information on the OpenfMRI data used. 'Exp' indexes the six cognitive experiments A–F. #AC is the accession number of an OpenfMRI project, where the same group of subjects are used in each project and there is no overlapping subject between projects. #Sub indicate the number of unique subjects for each dataset. Each of the six experiments has two brain conditions to classify. Each subject in each experiment contributed two positive and two negative brain condition samples, respectively.

Exp	#AC	Exp Description	#Sub
A	ds007	Stop signal with spoken pseudo word naming (Xue, Aron, and Poldrack 2008)	20
B	ds007	Stop signal with spoken letter naming (Xue, Aron, and Poldrack 2008)	20
C	ds007	Stop signal with manual response (Xue, Aron, and Poldrack 2008)	20
D	ds008	Conditional stop signal (Aron et al. 2007)	13
E	ds101	Simon task [Unpublished]	21
F	ds102	Flanker task (Kelly et al. 2008)	26

again, we have $f(\cdot) = \sum_{i=1}^{n} w_i k_x(\cdot, \mathbf{x}_i)$, and therefore $f(\mathbf{x}_j) = \sum_{i=1}^{n} w_i k_x(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{w}^\top k_x(\mathbf{X}, \mathbf{x}_j)$. By incorporating SVM loss into Eq. (12), we formulate the primal objective function of SIDeR$_{\text{SVM}}$ as

$$\min_{\mathbf{w},\xi,b} \frac{1}{2}\mathbf{w}^\top \mathbf{K}\mathbf{w} + C\sum_{i}^{\tilde{n}} \xi_i + \frac{\tilde{\lambda}}{2}\mathbf{w}^\top \mathbf{K}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{K}\mathbf{w}, \quad \text{s.t. } y_i(\mathbf{w}^\top \mathbf{k}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \ldots, \tilde{n}, \tag{13}$$

where ξ_i is the "slack variable" for the ith sample, b is a bias term, $C > 0$ controls the trade-off between penalty and the margin, $\tilde{\lambda} = \lambda/(n-1)^2$, and λ controls the significance of simplified HSIC regularizer.

The role of unlabeled samples. The unlabeled target samples have no labels, but they have domain side information (covariates) available. Thus, they should affect (regularize) $\mathbf{w}$ directly via the simplified HSIC term.

Optimization. To solve Eq. (13) effectively, we follow the steps in (Belkin, Niyogi, and Sindhwani 2006) and reformulate Eq. (13) via Lagrange dual. We consider the first $\tilde{n}$ samples are labeled and construct a matrix and denote $\mathbf{a}$ as the Lagrange multipliers, resulting the dual problem

$$\min_{\mathbf{a}} \frac{1}{2}\mathbf{a}^\top \mathbf{Q}\mathbf{a} - \sum_{i=1}^{\tilde{n}} a_i, \quad \text{s.t. } \mathbf{a}^\top \mathbf{y} = 0,\ 0 \leq a_i \leq C,\ i = 1, \ldots, \tilde{n}, \tag{14}$$

where $\mathbf{Q} = \tilde{\mathbf{Y}}\mathbf{K}(\mathbf{I} + \tilde{\lambda}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{K})^{-1}\tilde{\mathbf{Y}}^\top$, $\tilde{\mathbf{Y}} \in \mathbb{R}^{\tilde{n}\times n}$ is a matrix where $\tilde{\mathbf{Y}}_{i,i} = \mathbf{y}_i$, $i = 1, \ldots \tilde{n}$, and the rest are zeros. Denoting $\mathbf{a}^*$ as the optimal solution to Eq. (14), $\mathbf{w} = (\mathbf{I} + \tilde{\lambda}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{K})^{-1}\tilde{\mathbf{Y}}^\top \mathbf{a}^*$. Equation (14) is a quadratic programming (QP) problem that can be solved by standard QP tools. Algorithm 1 is the pseudocode for SIDeR$_{\text{SVM}}$.

Computational complexity. The complexity of computing HSIC is $O(n(d^2 + \tilde{d}^2))$ when linear kernel is used, i.e., $\mathbf{K} = \mathbf{X}^\top\mathbf{X}$, $\mathbf{L} = \mathbf{D}^\top\mathbf{D}$ (Gretton et al. 2005). In brain decoding problems, $d \gg n$ and $d \gg \tilde{d}$, so the overall computational complexity of HSIC is $O(nd^2)$. However, HSIC only needs to be computed once. The complexity of solving the quadratic programming problem for Eq. (14) is $O(n^3)$.

Table 2: Domain differences from the psychological perspective. A⇌B means when A is used as target experiment, B will be used as source experiment, and vice versa.

Tasks	Paradigm	Subjects	Control	Response
A⇌B	Same	Same	Same	Different
A⇌C	Same	Same	Different	Different
B⇌C	Same	Same	Different	Different
C⇌D	Same	Different	Similar	Different
A⇌D	Same	Different	Different	Different
B⇌D	Same	Different	Different	Different
E⇌F	Different	Different	Similar	Different

Experiments

This section evaluates SIDeR$_{\text{SVM}}$ against ten competing methods on 17 TL tasks in brain decoding. [1]

Experimental Setup

Dataset selection. We selected six datasets (A to F) that are most meaningful from psychological perspective from the public OpenfMRI repository,[2] as summarized in Table 1. Each dataset is from an experiment. Subjects from the same accession number (ds$\times\times\times$) are the same and there is no overlapping subject between accession numbers. There are two brain conditions selected from each dataset, with each as a *class* and having the same number of samples. Thus, we have binary classification problems that discriminate between brain conditions in an experiment.

Preprocessing. Each sample was preprocessed using FSL (Jenkinson et al. 2012) with the protocol in (Poldrack et al. 2013) to obtain the Z-score statistical parametric map (SPM) (Friston et al. 1994; 1998) of size $91 \times 109 \times 91$, which is then reduced to a vector of size $228,546$ by masking the voxels outside of the brain.

Seventeen TL tasks. We constructed 17 TL problems with increasing psychological difficulty, as determined by discrepancy across experimental paradigms, subjects involved, cognitive *control* demands (e.g., inhibiting a planned response or ignoring a distracting or misleading stimulus), and complexity or modality of response. Table 2 summarizes how each pair of tasks relate on these dimensions. We denote source and target experiments as S and T and classify a positive condition against a negative condition. We define 17 TL tasks ($S \to T$) as listed in the first columns of Tables 3 and 4 with the three classification problems below:

- "*Successful stop*" vs "*Unsuccessful stop*": Twelve TL tasks with single-source experiment.

[1] The code for reproducing the experiments (including data preprocessing) is available at https://github.com/sz144/sider

[2] https://legacy.openfmri.org or https://openneuro.org.

Table 3: Classification accuracy in percentage for 14 single-source experiment TL tasks (mean ± standard deviation over ten repetitions of cross-validation). 'Avg' is the average over the 14 tasks. Each task is denoted as *Source→Target* experiment, e.g., A→B means A is the source and B is the target. Subscript r denotes the RBF kernel gives better results, and for those without subscript r, the linear kernel gives better results. The best result for each task is in **bold**, and the second best is underlined.

	SVM^t	**PCA**^{t}	**PCA**$^{s+t}$	**TCA**	**SSTCA**$_r$	**JDA**	**ARSVM**	**MEDA**	**MIDA**	**SMIDA**	**SIDeR**$_{\text{SVM}}$
A→B	63.4±2.6	62.5±4.4	59.7±4.5	55.0±4.3	48.0±3.9	63.7±2.8	65.7±2.2	57.8±5.3	64.5±5.0	54.9±2.8	**78.6±2.9**
B→A	59.2±4.0	60.4±5.7	60.8±3.1	58.9±3.2	50.4±4.6	54.4±1.8	64.0±2.8	62.5±4.5	71.2±3.5	52.4±1.5	**79.6±2.7**
A→C	68.4±2.6	66.9±6.3	74.0±5.9	70.6±6.3	48.0±4.6	58.4±3.1	70.9±2.7	70.1±2.1	78.4±2.9	57.5±2.1	**87.4±2.1**
C→A	59.2±4.0	60.4±5.7	67.6±5.2	63.6±4.0	51.8±3.4	53.7±3.6	59.5±3.4	59.7±3.1	**70.6±4.0**	52.4±1.3	68.8±3.8
B→C	68.4±2.6	66.9±6.3	78.9±3.4	86.4±5.3	49.9±3.6	63.1±5.2	73.5±2.1	73.6±3.6	80.1±4.4	58.5±2.5	**90.5±1.5**
C→B	63.4±2.6	62.5±4.4	66.6±4.0	75.3±4.0	52.6±4.4	54.1±3.5	62.5±2.1	57.0±2.1	73.2±3.3	58.4±2.4	**77.4±2.8**
C→D	74.6±3.2	79.6±8.2	68.9±4.9	74.4±5.3	44.8±2.6	71.7±2.3	81.3±4.0	64.6±3.4	74.2±6.6	66.5±2.8	**87.1±2.3**
D→C	68.4±2.6	66.9±6.3	75.0±1.4	71.9±4.6	56.0±5.3	58.0±2.4	70.9±2.6	62.8±4.1	74.4±4.2	61.4±3.6	**87.4±2.6**
A→D	74.6±3.2	78.7±2.6	51.9±4.3	54.0±3.2	53.7±3.2	64.2±9.9	78.5±1.9	52.9±4.5	63.3±6.0	72.7±4.6	**86.2±1.9**
D→A	59.2±4.0	60.4±5.7	63.7±5.3	58.0±4.8	50.0±3.1	55.5±1.9	59.2±3.4	58.4±3.6	50.6±2.4	48.1±1.7	**67.1±3.8**
B→D	74.6±3.2	79.6±8.2	68.5±2.1	67.9±2.6	60.2±4.0	62.5±4.7	85.6±2.3	55.4±4.7	66.9±4.2	78.7±3.2	**94.2±1.9**
D→B	63.4±2.6	65.2±4.4	60.5±5.6	50.1±5.4	54.0±4.0	54.7±3.1	61.4±2.4	60.4±5.7	64.7±3.6	55.5±2.3	**73.7±3.1**
E→F	66.9±1.7	67.5±6.2	51.4±1.8	52.3±3.2	56.0±2.3	50.5±2.2	62.5±2.0	60.2±2.2	66.4±3.0	61.5±3.8	**74.0±4.0**
F→E	53.3±2.8	51.8±3.9	51.3±2.4	49.3±1.9	52.1±2.1	50.2±1.7	49.9±2.2	52.7±1.5	**53.7±4.5**	49.2±1.7	51.0±4.4
Avg	65.5±3.0	66.5±5.7	64.2±3.8	63.4±4.1	51.9±3.6	58.2±3.4	67.5±2.6	60.6±3.7	68.0±4.2	59.1±2.6	**78.8±2.8**

- "*Congruent correct*" vs "*Incongruent correct*": Two TL tasks with singe-source experiment.

- "*Successful stop*" vs "*Unsuccessful stop*": Three TL tasks with multiple (two) source experiments.

Ten methods compared. We evaluate SIDeR$_{\text{SVM}}$ against ten methods: three simple baselines 1) **SVM**^{t}, 2)**PCA**^{t}, and 3) **PCA**$^{s+t}$; and seven state-of-the-art TL methods discussed in *Preliminary*: 4) **TCA** (Pan et al. 2011), 5) **SSTCA** (Pan et al. 2011), 6) **JDA** (Long et al. 2013b), 7) **ARSVM** (Long et al. 2013a) of ARTL, 8) **MEDA** (Wang et al. 2018), 9) **MIDA** (Yan, Kou, and Zhang 2018), and 10) **SMIDA** (Yan, Kou, and Zhang 2018). SVM^t and PCA^t use only the target data while PCA^{s+t} uses both source and target data. For multi-source experiments TL, the multiple source experiments are used as one single source domain by MMD based methods, i.e., TCA, SSTCA, JDA, and ARSVM. PCA, TCA, SSTCA, JDA, MIDA, and SMIDA only learn a feature mapping so they use SVM as the classifier. Both linear and RBF kernels were studied for such SVM classifiers, and also for ARSVM, MEDA and SIDeR$_{\text{SVM}}$. We will report the result from the best performing kernel.

We performed 10 × 5-fold cross-validation on the target domain. For each split, the target training samples and all source samples (except for SVM^t and PCA^t) were used for training, with the target test samples for testing. On each training set, the optimal hyper-parameters for all methods are determined using the search strategy in (Pan et al. 2011) with 10 further random splits (20% for validation, 80% for training). To compare the difference between MMD and HSIC as a regularizer, manifold regularization and conditional distribution mismatch of ARSVM are not considered. Sensitivity studies for SIDeR are provided later in this section, which will validate SIDeR can offer stable performance for a wide range of hyper-parameter settings.

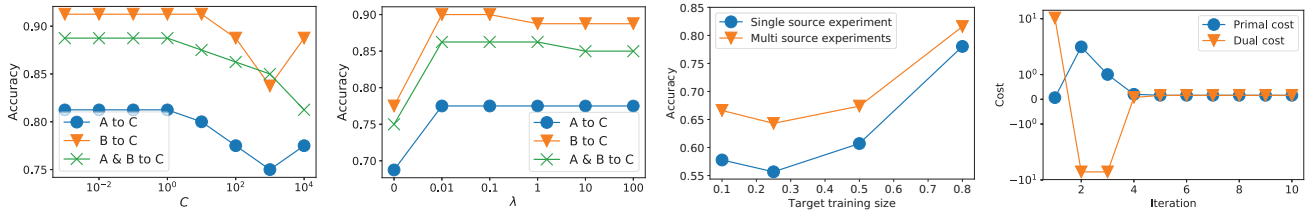
Results and Discussions

Tables 3 and 4 summarize the decoding accuracy of the 17 TL tasks with both the mean and standard deviation. The best result for each task is in **bold** and the second best is underlined. We have five key observations:

- On the whole, SIDeR$_{\text{SVM}}$ outperformed all the comparing methods. From results of using experiment A or B as target domain, we can observe the performance gain decreases from easy to difficult tasks. This indicates a plausible correlation between the TL improvements and the transfer difficulties from psychological perspective.
- SIDeR$_{\text{SVM}}$ outperformed the best existing method (ARSVM) by **15.6%** (83.2% vs. 67.6%) in TL tasks with multi-source experiments (Table 4). On the other hand, SIDeR$_{\text{SVM}}$ obtained lower accuracy on A&B→C compared to B→C, and the rest results show using multiple source experiments is better. Thus, source selections can influence the transfer performance. If there is no clear preference of a particular source dataset, transfer with multiple sources is preferred in our opinion.
- MIDA and SIDeR$_{\text{SVM}}$ outperformed the corresponding MMD-counterparts TCA and ARSVM, respectively. Based on Theorem 1, this confirmed that making use of multiple domain side information (experiments and subjects) is beneficial in brain decoding.
- SIDeR$_{\text{SVM}}$ outperformed SMIDA, and ARSVM outperformed SSTCA, which indicates prediction loss (hinge loss) is more robust than variance preserving and label dependence maximization in brain decoding problems.
- In Talbe 3, TCA and MIDA outperformed the corresponding semi-supervised version SSTCA and SMIDA. However, their performance were close when more source samples were used (Table 4). This observation indicates that label dependence is more susceptible to overfitting. Additionally, the performance of methods with conditional distribution alignment, i.e., JDA and MEDA, were

Table 4: Classification accuracy in percentage for three multi-source experiment TL tasks.

	SVM^t	PCA^t	PCA_r^{s+t}	TCA_r	SSTCA	JDA	ARSVM	MEDA	MIDA	SMIDA	$\text{SIDeR}_{\text{SVM}}$
B&C→A	59.2±4.0	60.4±5.7	51.5±1.5	52.1±1.1	52.1±3.4	51.7±2.8	63.2±2.5	52.6±1.4	51.4±1.8	55.4±2.4	**80.3±3.1**
A&C→B	63.4±2.6	62.5±4.4	53.7±2.1	54.5±3.1	52.2±3.2	53.8±3.3	67.6±1.6	52.6±1.3	61.7±2.2	60.4±2.2	**79.5±2.0**
A&B→C	68.4±2.6	66.9±6.3	50.6±2.5	52.1±1.1	56.5±3.5	52.0±3.1	71.9±2.6	57.0±0.8	65.4±2.2	60.5±3.2	**89.9±1.7**
Avg	63.7±3.1	62.6±5.5	52.0±2.0	52.9±1.8	53.6±3.4	52.5±3.0	67.6±2.2	54.1±2.6	59.5±2.0	58.8±2.6	**83.2±2.3**



(a) C. (b) λ. (c) Target training size. (d) Convergence.

Figure 1: The sensitivity of the classification accuracy with respect to hyper-parameters C, λ, and labeled target training data size, and the convergence study for $\text{SIDeR}_{\text{SVM}}$ with linear kernel.

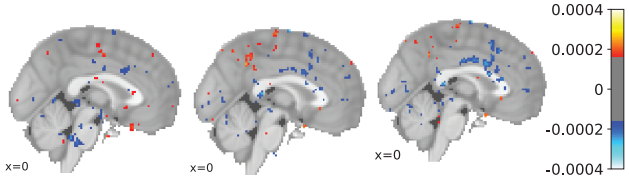
inferior to marginal distribution alignment methods, i.e., TCA and ARSVM. In summary, using label dependence or conditional distribution alignment may lead to inferior performance in small sample brain decoding TL tasks.

Hyper-parameter sensitivity. We evaluated the sensitivity of $\text{SIDeR}_{\text{SVM}}$ with linear kernel against hyper-parameters C and λ under five-fold cross validation. Figure 1a shows the sensitivity against $C \in [10^{-3}, 10^4]$ when fixing $\lambda = 1$. We can observe that the accuracy stays stable when $C \leq 1$, and shows a trend of decreasing when $C \in [10^0, 10^4]$. Since a smaller value of C can lead to a larger SVM classification margin, we expect a classifier with a larger margin to generalize better and have higher prediction accuracy. Figure 1b shows the sensitivity against $\lambda \in \{0, 0.01, 0.1, 1, 10, 100\}$ when fixing $C = 1$. We can observe that the prediction accuracy of $\text{SIDeR}_{\text{SVM}}$ kept almost constant when $\lambda > 0$ and it was not sensitive to λ. When $\lambda = 0$, i.e., without minimizing domain dependence, $\text{SIDeR}_{\text{SVM}}$ becomes a standard kernel SVM and the performance dropped significantly.

Sensitivity to sample size. Figure 1c shows the sensitivity of $\text{SIDeR}_{\text{SVM}}$ with respect to labeled target training sample size for single/multi-source experiment transfer, averaged over fourteen/three single/multi-source experiment transfer tasks over 20 random splits, respectively. Multi-source experiment transfer obtained better performance, especially when the labeled target training sample size is smaller.

Convergence. Figure 1d illustrates the primal and dual cost of $\text{SIDeR}_{\text{SVM}}$ on the TL task A→B. We can observe that the costs converged within ten iterations.

Model visualization. We visualize the three models with the best performance on the learning task A&B→C, which are SVM^t, ARSVM, and $\text{SIDeR}_{\text{SVM}}$, using the Python package Nilearn (Abraham et al. 2014). Figure 2 depicts the learned weights averaged over 5-fold cross validation in the voxel space. We can observe that without training with source data, SVM^t (Fig. 2a) highlight different areas compare to ARSVM (Fig. 2b) and $\text{SIDeR}_{\text{SVM}}$ (Fig. 2c), which



(a) SVM^t (b) ARSVM (c) $\text{SIDeR}_{\text{SVM}}$

Figure 2: Visualization of the top 1% learned voxel weights in magnitude for three classifiers averaged over 5-fold cross validation on A&B→C.

identified some common clusters around cingulate gyrus (shaded in blue). These clusters in Fig. 2c are clearer and less noisy, suggesting that $\text{SIDeR}_{\text{SVM}}$ has identified more coherent brain functional areas.

Conclusion

In this paper, we proposed Side Information Dependence Regularization (SIDeR) learning framework for TL in analyzing human brain conditions across subjects and experiments. We incorporated the SVM loss into SIDeR to simultaneously minimize the empirical prediction risk and the dependence on domain side information measured by a simplified HSIC. We evaluated SIDeR against SVM, PCA and seven state-of-the-art TL methods on seventeen TL tasks. Experimental results showed the superior overall performance of SIDeR over other methods, particularly on multi-source experiment transfer, with a 15.6% improvement. This confirmed the benefits of leveraging domain side information and HSIC in TL for brain condition decoding.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant

EP/R014507/1 and the European Research Council (ERC) grant GAP: 670428 - BRAIN2MIND_NEUROCOMP.

References

Abraham, A.; Pedregosa, F.; Eickenberg, M.; Gervais, P.; Mueller, A.; Kossaifi, J.; Gramfort, A.; Thirion, B.; and Varoquaux, G. 2014. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics* 8:14.

Aron, A. R.; Behrens, T. E.; Smith, S.; Frank, M. J.; and Poldrack, R. A. 2007. Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *Journal of Neuroscience* 27(14):3743–3752.

Aydore, S.; Thirion, B.; and Varoquaux, G. 2019. Feature grouping as a stochastic regularizer for high-dimensional structured data. In *ICML*, 385–394.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* 7(Nov):2399–2434.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine Learning* 79(1-2):151–175.

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22(14):e49–e57.

Chen, P.-H. C.; Chen, J.; Yeshurun, Y.; Hasson, U.; Haxby, J.; and Ramadge, P. J. 2015. A reduced-dimension fmri shared response model. In *NeurIPS*, 460–468.

Cox, D. D., and Savoy, R. L. 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19(2):261–270.

Friston, K. J.; Holmes, A. P.; Worsley, K. J.; Poline, J.-P.; Frith, C. D.; and Frackowiak, R. S. 1994. Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping* 2(4):189–210.

Friston, K. J.; Fletcher, P.; Josephs, O.; Holmes, A.; Rugg, M.; and Turner, R. 1998. Event-related fMRI: characterizing differential responses. *NeuroImage* 7(1):30–40.

Gorgolewski, K.; Esteban, O.; Schaefer, G.; Wandell, B.; and Poldrack, R. 2017. Openneuro - a free online platform for sharing and analysis of neuroimaging data. In *OHBM*, 1677.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 63–77. Springer.

Jenkinson, M.; Beckmann, C. F.; Behrens, T. E.; Woolrich, M. W.; and Smith, S. M. 2012. FSL. *NeuroImage* 62(2):782–790.

Kelly, A. C.; Uddin, L. Q.; Biswal, B. B.; Castellanos, F. X.; and Milham, M. P. 2008. Competition between functional brain networks mediates behavioral variability. *NeuroImage* 39(1):527–537.

Li, H.; Parikh, N. A.; and He, L. 2018. A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. *Frontiers in Neuroscience* 12:491.

Long, M.; Wang, J.; Ding, G.; Pan, S. J.; and Philip, S. Y. 2013a. Adaptation regularization: A general framework for transfer learning. *TKDE* 26(5):1076–1089.

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2013b. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2200–2207.

Mensch, A.; Mairal, J.; Bzdok, D.; Thirion, B.; and Varoquaux, G. 2017. Learning neural representations of human cognition across many fMRI studies. In *NeurIPS*, 5883–5893.

Ogawa, S.; Lee, T.-M.; Kay, A. R.; and Tank, D. W. 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *PNAS* 87(24):9868–9872.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2):199–210.

Poldrack, R. A.; Barch, D. M.; Mitchell, J.; Wager, T.; Wagner, A. D.; Devlin, J. T.; Cumba, C.; Koyejo, O.; and Milham, M. 2013. Toward open sharing of task-based fMRI data: the OpenfMRI project. *Frontiers in Neuroinformatics* 7:12.

Rao, N.; Cox, C.; Nowak, R.; and Rogers, T. T. 2013. Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. In *NeurIPS*, 2202–2210.

Schölkopf, B.; Herbrich, R.; and Smola, A. J. 2001. A generalized representer theorem. In *COLT*, 416–426. Springer.

Van Essen, D. C.; Ugurbil, K.; Auerbach, E.; Barch, D.; Behrens, T.; Bucholz, R.; Chang, A.; Chen, L.; Corbetta, M.; Curtiss, S. W.; et al. 2012. The human connectome project: a data acquisition perspective. *NeuroImage* 62(4):2222–2231.

Velioglu, B., and Vural, F. T. Y. 2017. Transfer learning for brain decoding using deep architectures. In *ICCI*CC*, 65–70. IEEE.

Wang, J.; Feng, W.; Chen, Y.; Yu, H.; Huang, M.; and Yu, P. S. 2018. Visual domain adaptation with manifold embedded distribution alignment. In *ACMMM*, 402–410. ACM.

Xue, G.; Aron, A. R.; and Poldrack, R. A. 2008. Common neural substrates for inhibition of spoken and manual responses. *Cerebral Cortex* 18(8):1923–1932.

Yan, K.; Kou, L.; and Zhang, D. 2018. Learning domain-invariant subspace using domain features and independence maximization. *IEEE Transactions on Cybernetics* 48(1):288–299.

Zhang, C.; Qiao, K.; Wang, L.; Tong, L.; Hu, G.; Zhang, R.-Y.; and Yan, B. 2019. A visual encoding model based on deep neural networks and transfer learning for brain activity measured by functional magnetic resonance imaging. *Journal of Neuroscience Methods* 108318.

Zhang, H.; Chen, P.-H.; and Ramadge, P. 2018. Transfer learning on fMRI datasets. In *AISTATS*, 595–603. PMLR.