Transfer Learning on fMRI Datasets

Hejia Zhang
Princeton University

Po-Hsuan Chen
Princeton University

Peter J. Ramadge
Princeton University

Abstract

We explore transferring learning between fMRI datasets. A method is introduced to improve prediction accuracy on a primary fMRI dataset by jointly learning a model using other secondary fMRI datasets. We assume the secondary datasets are directly or indirectly linked to the primary dataset through sets of partially shared subjects. This method is particularly useful when the primary dataset is small. Using six fMRI datasets linked by various subsets of shared subjects, we show that the method yields improved performance in various predictive tasks. Our tests are performed on a variety of regions of interest in the brain and across various stimuli.

1 INTRODUCTION

Functional magnetic resonance imaging (fMRI) is a noninvasive neuroimaging technique that measures brain activity through the proxy of blood-oxygen-level-dependent (BOLD) contrast imaging [1]. One typically collects an fMRI dataset for a group of subjects using a fixed experimental design. The design could be block based, a naturalistic stimulus, or a combination of forms. The resulting multi-subject dataset is then analyzed to address a particular hypothesis about brain function, or used in an exploratory way to develop new hypotheses. Various data analysis methods exist for achieving these objectives. Here we investigate a distinct approach that seeks to exploit other secondary datasets (collected for different purposes, at different times, by other researchers, etc.) to improve the analysis of the primary dataset. There has been prior work related to this idea using canonical correlation analysis [2, 3]. We take a fresh look at this problem from a transfer learning perspective [4]. By training a model jointly on the primary and secondary datasets, we are transferring information from the secondary datasets to the primary dataset. Alternatively, by training on secondary datasets alone, we can then generalize to the primary dataset. Prior work [5–9] on transfer learning in fMRI, MRI, and EEG has used techniques such as domain adaptation and common-pattern identification. Here we employ a distinctive technique to exploit secondary datasets.

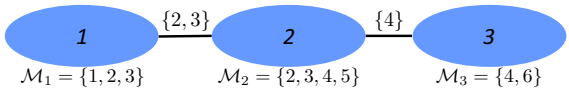

Figure 1: A simple dataset graph. Nodes represent datasets, edges indicate the presence of shared subjects, and the edge labels indicate the set of indices of the shared subjects. $\mathcal{M}_d$ is the set of subject indices in dataset d.

The key challenge in directly analyzing the primary dataset is accounting for structural and functional variation across subjects. Of the many approaches to this issue, functional methods [3, 10–24] have proven most successful. By modeling the functional variability across subjects, these methods yield better accuracy on predictive tasks, e.g., [20, 25, 26], but work best when there are many subjects and/or many time samples (TRs) per subject. This creates the opportunity for secondary datasets to assist the analysis of a small primary dataset. One can visualize the primary and secondary datasets as a labeled graph (Fig. 1) with datasets as nodes and an edge between nodes p and q labeled with the set of indices of the corresponding shared subjects. Not all edges need to be present, and the labels can be distinct and of different sizes. Under reasonable assumptions, the greater connectivity of the dataset graph, and the larger the sets of shared subjects on its edges, the more improvement we expect to see in the analysis of the primary dataset. An alternative way to think about the problem is to artificially assume that all subjects are included in all datasets, but that some of the data is missing (e.g., the data for a subject that didn't participate in an experiment is missing). We can then view the problem of leveraging secondary datasets as an instance of the more general problem of handling missing data. The methods we discuss cover both situations, but to simplify the presentation, we focus on the first.

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

The existence of datasets with shared subjects is important for this research. Some datasets of this form are currently publicly available (see Fig. 3), and it is likely that many more will be available in the future. The existence of shared subjects across fMRI datasets is already common in university research laboratories where each principal investigator (PI) draws subjects from the same slowly varying pool of participants. Thus, it is common for some subjects in each experiment to also participate in several other experiments. Besides, longitudinal fMRI studies (e.g. [27, 28]) can also provide multi-datasets with shared subjects. The National Institute of Mental Health (NIMH) now requires, that PIs "Obtain Informed Consent that allows for broad sharing of the research subject's de-identified data" and that the PIs "Collect Personally Identifiable Information" from subjects that permits the creation of a de-identified subject ID [29, 30]. Shared subjects between datasets can then be located in the NIMH data archive using the subject ID. This adds tremendous potential value to the fMRI research community and will increase the number of publicly available datasets containing ID-identified shared subjects.

Our contribution is to show that utilizing a model with fixed subject-specific basis, and a common stimulus within each dataset, can account for the differences between subjects and datasets, and successfully transfer information between datasets.

2 PROBLEM SETUP

Let m subjects, indexed by $i = 1:m$, participate in n different experiments, indexed by $d = 1:n$. Experiment d is conducted over a subset $\mathcal{M}_d$ of the subjects with $|\mathcal{M}_d| = m_d$. We make two assumptions. First, that in experiment d all subjects are recorded for the same number of time samples (TRs). Each time sample (TR) is one 3D brain scan. This is not a strong assumption. Second, that across all experiments in which subject i participates, data is available for the same set of (anatomically aligned) voxels of our choice. This, slightly stronger, assumption requires that a common subset of brain regions are imaged across the datasets. Under these assumptions, the fMRI data from subject $i \in \mathcal{M}_d$ can be represented as a matrix $X_{di} \in \mathbb{R}^{v_i \times t_d}$, where v_i is the number of voxels for subject i and t_d is the number of TRs for experiment d. Each column of X_{di} contains the vectorized brain scan from a specific time sample. Dataset d, denoted by $\{X_{di}\}_{i \in \mathcal{M}_d}$, is the fMRI data from subjects in experiment d.

We focus on factor models that acknowledge functional variability across subjects. So we want to represent the j-th column of X_{di} in the factored form $W_i s_{dj}$ where the columns of W_i form a subject-specific basis of dimension k. Each column of X_{di} is a "brain map" indicating the level of activations across the voxels. So W_i is a subject-specific basis of brain maps. With this in mind, we select three factor models for consideration: Group-ICA (GICA) [20], Dictionary Learning (DL) [13], and Hyperalignment (HA)/Shared Response Model (SRM) [12, 16]. These methods differ in how the data dimensions (space, time, subjects) are treated, how regularization is imposed to estimate the factors, and what assumptions are made on the dataset. We will use these methods applied to a single primary dataset as comparison benchmarks for any method making use of secondary datasets. We will also explore how secondary datasets can be incorporated into these methods. Below we briefly review each method. Additional single-dataset multi-subject models are discussed in the supplementary material.

GICA is a deterministic algorithm that applies two PCAs and one ICA along the temporal dimension. It finds an independent basis arranged as the columns of a matrix W for all subjects in a dataset [19]. Since GICA estimates only one brain basis W, it does not directly estimate subject-specific functional variability. However, applying GICA in the spatial dimension (see, e.g., [31]) allows estimation of a brain basis W_i for each subject i. We use GICA in this way.

DL learns subject-specific bases W_i and loadings U_i while requiring all W_i to be similar to a group template W. The objective is to minimize

$$\sum_i \tfrac{1}{2}(\|X_{di}^T - U_i W_i^T\|_F^2 + \mu\|W_i - W\|_F^2) + \mu\alpha\Omega(W),$$

where Ω is a spatial regularizing function and μ, α are parameters. DL has shown good performance in finding brain maps using resting-state data [13].

HA/SRM methods learn a shared response S for the dataset and a subject-specific brain basis W_i by minimizing $\sum_i \|X_i - W_i S\|_F^2$ under the constraint $W_i^T W_i = I$. HA is a deterministic algorithm in which the number of factors is the number of voxels. SRM is a probabilistic latent variable model with a user selectable number of factors. To be consistent with GICA and DL, we only consider SRM.

Note that GICA, DL, and SRM account for subject variability by finding subject-specific bases W_i. GICA (applied in the spatial dimension) and SRM both assume an identical stimulus for each subject in an experiment and identify a time domain shared loading matrix S for each experiment. DL does not require this assumption, and the shared component is identified through the spatial group template W.

We make two modeling assumptions to help transfer information from the secondary datasets. First, we assume that the subject-specific spatial bases W_i are invariant across datasets. This seems to be a reasonable assumption when the stimuli in the various experiments

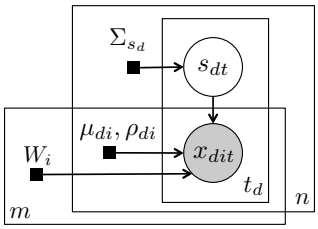

Figure 2: Graphical model for MDMS. Brain activation pattern $x_{dit} \in \mathbb{R}^{v_i}$ (v_i voxels) is observed from subject i at time t, $t = 1{:}t_d$, $i = 1{:}m$, $d = 1{:}n$. Each observation x_{dit} is a linear combination of subject-specific orthogonal basis (columns of W_i) using the weights specified by s_{dt}. Shaded nodes: observations, unshaded nodes: latent variables, and black squares: parameters and hyperparameters.

have similar characteristics and the subjects remain in a healthy state. Note that the bases can be whole-brain or for voxels within a region of interest (ROI). Second, to exploit multiple subjects within a dataset, we need some form of connection across subjects. This can be achieved through a group spatial basis shared across subjects, as in DL, or through an identical stimulus, as in GICA and SRM. We take the latter approach assuming an identical stimulus within a dataset. This will be modeled by a common set of temporal features representing the stimulus.

Henceforth, we assume each subject within a dataset receives the same stimulus. Both DL and SRM find subject-specific spatial bases W_i. We assume these are invariant across datasets. The SRM method also learns temporal loadings S that are specific to the experiment. The DL method does not require an identical stimulus for each subject in an experiment, so its temporal loadings can be subject-specific. It hence has no explicit dataset-specific shared component in the time domain. To address this, we tried modifying the model to enforce shared loadings across subjects within a dataset during training. However, this did not perform well in testing. Alternatively, we can average the subject-specific loadings U_i across subjects in each dataset and treat the averaged loadings as the shared temporal component within the dataset. With this modification, DL can also be used to leverage secondary datasets. Applying GICA in the spatial domain estimates subject-specific bases W_i. But W_i is dataset-specific and is not invariant across datasets. So GICA needs more extensive reformulation to meet our multi-dataset assumptions.

3 SECONDARY DATASETS

We now use DL and SRM to transfer information from the secondary datasets to the primary dataset. Since DL does not require an identical stimulus across subjects, we concatenate each subject's data from all datasets and treat this combination as a single dataset. Then apply DL to the aggregated dataset. The dataset-specific time domain loadings can be computed after training as discussed above. We do not change the DL model. We simply explore how well the DL method can leverage the secondary datasets. However, to better differentiate DL applied on multiple datasets, we label it multi-dataset DL (MDDL). Since DL does not require an identical stimulus, it is potentially at a slight disadvantage in our experimental paradigm. This needs to be kept in mind when interpreting test results.

The SRM method models the stimulus of a dataset using a Gaussian latent variable and models the subject-specific basis as a hyperparameter. To extend SRM into a multi-dataset setting, we need to modify the model while respecting the two modeling assumptions. We extend the single Gaussian latent variable in SRM into a set of Gaussian latent variables modeling all n stimuli. The latent variable for dataset d can be represented as $s_{dt} \sim \mathcal{N}(\mathbf{0}, \Sigma_{s_d})$ taking values in $\mathbb{R}^k$. In accordance with modeling assumptions, we keep W_i fixed while introducing multiple datasets. The observation in dataset d for subject i at time t is denoted by $x_{dit} \in \mathbb{R}^{v_i}$. x_{dit} is modeled by a multivariate Gaussian distribution $x_{dit} \sim \mathcal{N}(W_i s_{dt} + \mu_{di}, \rho_{di}^2 I)$, where $W_i \in \mathbb{R}^{v_i \times k}$ is the basis for subject i, $i \in \mathcal{M}_d$, μ_{di} and ρ_{di}^2 are dataset-subject-specific mean and noise, respectively. Note that the subject-specific basis W_i is used across the datasets.

Combining the dataset-specific latent variables s_{dt}, the subject-specific hyperparameter W_i, and the observations x_{dit} yields the joint model (see Fig. 2),

$$\begin{aligned} s_{dt} &\sim \mathcal{N}(\mathbf{0}, \Sigma_{s_d}), \\ x_{dit}|s_{dt} &\sim \mathcal{N}(W_i s_{dt} + \mu_{di}, \rho_{di}^2 I), \text{ s.t. } W_i^T W_i = I_k. \end{aligned}$$

When processing dataset d, we mask out subjects that are not in the dataset. Let dataset d contain subjects $\mathcal{M}_d = \{1 \dots m_d\}$. We use following notation to indicate subject masking:

$$\begin{aligned} W_{(d)}^T &= [W_1^T \dots W_{m_d}^T] \\ x_{dt}^T &= [{x_{d1t}}^T \dots {x_{dm_d t_d}}^T] \\ \Psi_d &= \text{diag}(\rho_{d1}^2 I, \dots, \rho_{dm_d}^2 I) \\ \mu_d^T &= [\mu_{d1}^T \dots \mu_{dm_d}^T]. \end{aligned}$$

Recall that $x_{dit} \in \mathbb{R}^{v_i}$ is the observation in dataset d from subject i at time t.

We then derive a constrained EM algorithm to estimate the posterior distributions of the latent variables and the hyperparameters (details in Sup. Mat.). The EM coordinate descent update equations are:

$$\begin{aligned} \mathbb{E}_{s_d|x_d}[s_{dt}] &= (W_{(d)}\Sigma_{s_d})^T (W_{(d)}\Sigma_{s_d} W_{(d)}^T + \Psi_d)^{-1} \\ &\qquad (x_{dt} - \mu_d), \qquad (1) \\ \mathbb{E}_{s_d|x_d}[s_{dt} s_{dt}^T] &= \Sigma_{s_d} - \Sigma_{s_d}^T W_{(d)}^T (W_{(d)}\Sigma_{s_d} W_{(d)}^T + \Psi_d)^{-1} \end{aligned}$$

$$W_{(d)}\Sigma_{s_d} + \mathbb{E}_{s_d|x_d}[s_{dt}]\mathbb{E}_{s_d|x_d}[s_{dt}]^T,$$

$$\mu_{di}^{\text{new}} = \frac{1}{t_d}\sum_{t=1}^{t_d} x_{dit},$$

$$W_i^{\text{new}} = A_i(A_i^T A_i)^{-1/2},$$

$$A_i = \frac{1}{2}\Big(\sum_{d:i\in\mathcal{M}_d}\sum_{t=1}^{t_d}(x_{dit} - \mu_{di}^{\text{new}})\mathbb{E}_{s_d|x_d}[s_{dt}]^T\Big),$$

$$\rho_{di}^{2\ \text{new}} = \frac{1}{t_d v_i}\sum_{t=1}^{t_d}\Big(\|x_{dit} - \mu_{di}^{\text{new}}\|^2 - 2(x_{dit} - \mu_{di}^{\text{new}})^T W_i^{\text{new}}\mathbb{E}_{s_d|x_d}[s_{dt}] + \text{tr}(\mathbb{E}_{s_d|x_d}[s_{dt}s_{dt}^T])\Big),$$

$$\Sigma_{s_d}^{\text{new}} = \frac{1}{t_d}\sum_{t=1}^{t_d}(\mathbb{E}_{s_d|x_d}[s_{dt}s_{dt}^T]).$$

We adopt techniques from [32] to speed up inversion of the matrix in (1) (details in Sup. Mat.). For ease of reference, we call this multi-dataset multi-subject (MDMS) SRM analysis (for brevity just MDMS).

Next, we investigate how to introduce new data into MDDL and MDMS. New data and new subjects can be introduced to MDDL in the same way as DL [13]. So we will not discuss this in detail. Let us focus instead on MDMS. To simplify the discussion, we use the deterministic version:

$$\min_{W_i,S_d}\sum_{d=1}^{n}\sum_{i\in\mathcal{M}_d}\|X_{di} - W_i S_d\|_F^2,$$

(see the Supp. Mat.). First, consider introducing a new subject i into dataset d, with the data X_{di} gathered under the stimulus of dataset d. We assume dataset d has already been modeled by MDMS. The basis for subject i can be estimated by solving $\min_{W_i,W_i^T W_i=I}\|X_{di} - W_i S_d\|_F^2$ while keeping S_d fixed. Second, consider introducing a new dataset d. Features modeling the stimulus of dataset d can be estimated by solving $\min_{S_d}\sum_{i\in\mathcal{M}_d}\|X_{di} - W_i S_d\|_F^2$ while keeping W_i fixed. This is possible when there are shared subjects between dataset d and the existing datasets. When we are only interested in estimating the features S_{di} of a new data point X_{di} from an existing subject i, we simply project the new data through W_i, giving us $S_{di} = W_i^T X_{di}$. We have made additional connections between GICA, DL, SRM, MDDL, and MDMS. These are included in the Supp. Mat..

4 EXPERIMENTS

Datasets: To test the transfer of information from secondary datasets to a primary dataset, we use six publicly available fMRI datasets collected from a total of 85 subjects (see Fig. 3). The datasets were collected while subjects were provided with distinct stimuli, different preprocessing steps, and are explored using different regions of interest (ROI) (details in Tab. 1). All datasets are aligned to the Montreal Neuroscience Institute (MNI) anatomical template [44]. Each dataset uses a distinct stimulus, but each subject in a dataset is presented with the same stimulus. Datasets *greeneyes*, *milky*, and *vodka* are collected while subjects were listening to a narrated story. Dataset *sherlock* contains two components for the same group of subjects. The first part (movie) was collected when the subjects were watching a movie and the second part (recall) was collected when the same subjects recalled the scenes in the movie without any external prompts. Dataset *schema* is a concatenation of 8 small datasets where each small dataset was collected while subjects listened to a 3-minute narrated story. References to the source of each dataset are given in the Table in Fig. 3. By treating a dataset as a node and shared subjects between two datasets as an edge, the network structure of these datasets is shown in Fig. 3. Three subjects are shared by *greeneyes*, *milky*, and *sherlock*, and 1 subject by *vodka*, *schema*, and *greeneyes*. No subject is shared by more than 3 datasets. (See Supp. Mat. for details).

Dataset	Type	Samples	Num. Subjs
greeneyes [33, 34]	Audio	450 TRs	40
milky [35, 36]	Audio	297 TRs	18
vodka [35, 36]	Audio	297 TRs	18
schema [37]	Audio	937 TRs	31
sherlock [38, 39]	Movie	1973 TRs	16
sherlock-recall [38, 39]	Recall	34 scenes	16

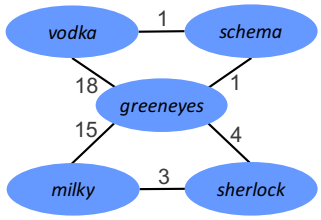

Figure 3: The Datasets. **Top:** Information on the fMRI datasets. Each TR is 1.5 secs. In scene recall, each scene is the average response over recall period of that scene. **Bottom:** Structure of datasets as a graph. The number of shared subjects is labeled on the graph edges.

Evaluation Method: We use test accuracy on various prediction tasks as a performance metric. All the accuracies are computed based on data from one or more left-out subjects. For each model, the number of factors k is selected from the set $[25, 50, 75, 100, 125, 150]$ using 4-fold cross-validation on the training data.

Experiment 1: Do Secondary Datasets Help? We first test if the prediction accuracy of a time segment matching experiment and a scene recall matching experiment on the primary dataset can be improved by leveraging secondary datasets. The time segment matching experiment tries to predict the time point of a given segment of fMRI response from a testing subject after training the model on data from other subjects. The scene recall matching experiment tries to predict the scene from recall data after training the model on data from other subjects. The experiments treat one dataset as the primary dataset and the other datasets as secondary datasets. Prediction accuracy

Region of Interest (ROI)	Num. Voxels	Description
default mode network (DMN) [40]	2329	language processing, etc.
early auditory cortex (EAC) [41]	1189	low level auditory processing
planum temporale (PT) [42]	1318	high level auditory processing
posterior medial cortex (PMC) [43]	813	mental state, etc.

Table 1: Information of ROIs used in the experiments.

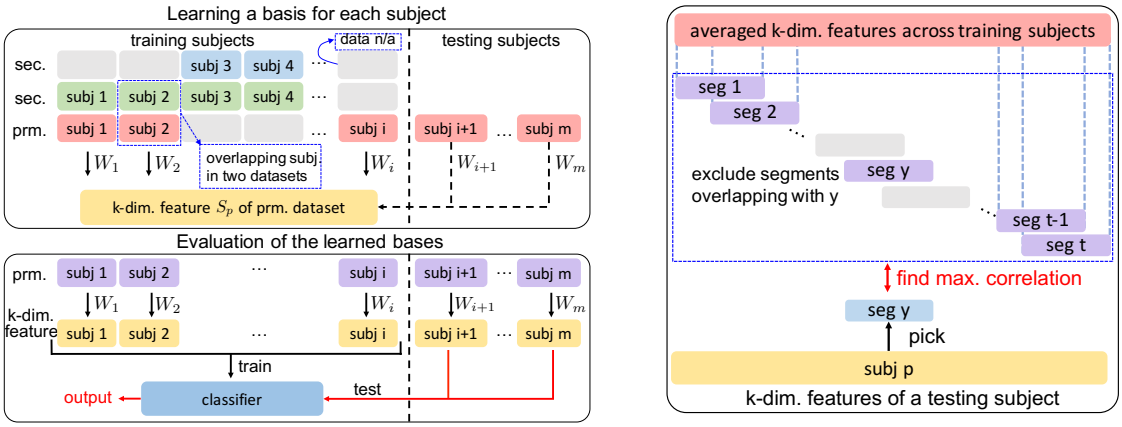

Figure 4: Experiment 1. See Exp. 1 paragraph 2 for full details. **Left**: Use sec. datasets to help prm. dataset learn a better set of subject bases. k-dimensional feature of secondary datasets are learned but not used in this experiment. Test on left-out testing subjects. **Right**: Time segment matching correlation classifier. Repeat for all possible segments.

is calculated on left-out subjects and left-out data in the primary dataset. We randomly partition all subjects from all datasets into 65 training subjects and 20 testing subjects. We ensure all datasets are linked through training subjects, and each dataset has at least three testing subjects. Testing subjects are completely left-out from all datasets during training. The splitting of training and testing subject is illustrated in Fig. 6. This random partition is repeated five times and averaged results are reported with standard error.

In the time segment matching experiment, we test on scenarios when *greeneyes*, *milky*, *vodka*, *sherlock* each becomes the primary dataset. *schema* is not used as a primary dataset because it has large temporal discontinuities due to the concatenation of smaller datasets. On each primary dataset, we perform a version of the time segment matching experiment adapted from [16]. The major difference is that we test on left-out subjects instead of left-out data of training subjects. There are two phases in the experiment. The first phase has two steps. In the first step, we partition the primary dataset into two halves in time. Then for all training subjects, we use one half of the primary dataset, and all TRs of the secondary datasets to jointly learn the subject-specific bases W_i. We also learn the k-dimensional temporal feature S_p for the selected half of the primary dataset. In the second step, for all testing subjects, the subject-specific bases W_i are computed using the other half of the primary dataset and the temporal feature S_p learned in the first step (details in §3).

In the second phase, we evaluate the usefulness of the estimated bases for the testing subjects in the primary dataset. The left-out data of all subjects in the primary dataset is transformed to the k-dimensional feature space using the bases learned in the first phase. Features are averaged across training subjects. For each testing subject, a 9 TR time segment y in the feature is selected and correlated with all 9 TR segments in the averaged feature across training subjects. Segments overlapping with y are excluded. If the highest correlation appears at segment y in the averaged feature across training subjects, then we say segment y is correctly matched. This procedure is repeated for both halves of the primary dataset, and all possible 9 TR segments and results are averaged. This time segment matching experiment tests two properties of the learned subject-specific bases. First, that the bases are stimulus-invariant. Second, that the bases transform subjects' data into a common feature space such that features of similar stimuli are highly correlated. In this case, a high matching accuracy suggests a high quality of the learned bases. Results on default mode network (DMN) (Tab. 1) are summarized in Fig. 5 (other ROIs in Sup. Mat.). For each model, k values are selected using cross-validation. MDMS has the highest accuracy in this matching experiment among all methods we have tested. This suggests MDMS can effectively leverage information in secondary datasets to aid the analysis in the primary dataset. We also observe more improvement when the primary dataset is small (*milky* and *vodka*). This is expected as it is hard for a primary dataset with a small number of training samples to learn bases with good generalizability, so adding

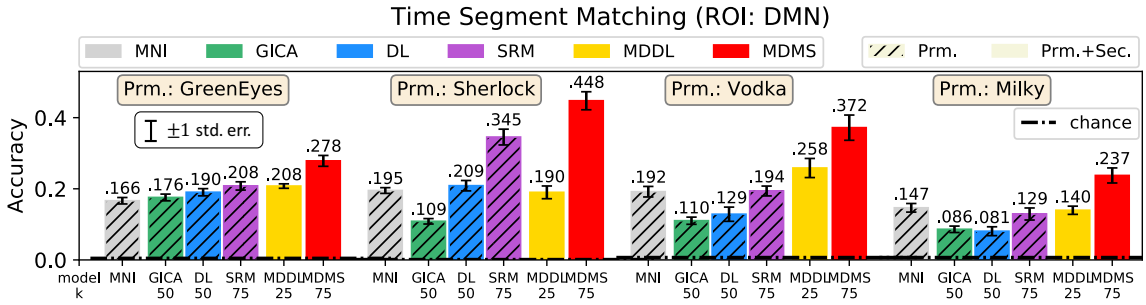

Figure 5: Experiment 1. Results of time segment matching experiment on ROI DMN (other ROIs in Supp. Mat.). Chance accuracy: *greeneyes*: 0.005; *sherlock*: 0.001; *vodka*: 0.008; *milky*: 0.008. k is selected based on 4-fold cross-validation. MNI is anatomical alignment. Mean and standard error computed using 5 random partitions of training and testing subjects.

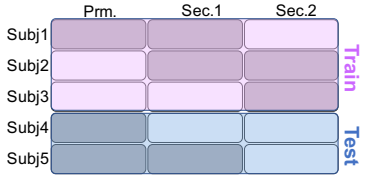

Figure 6: An example of a random partition of the training and testing subjects. Available observations are grey blocks, missing observations are white blocks. For all datasets, testing subjects are completely left-out.

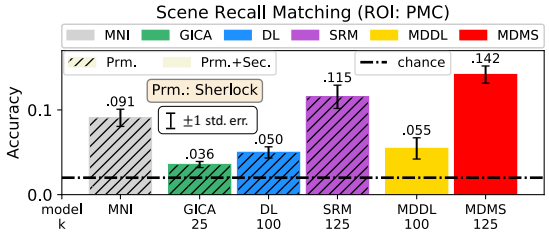

Figure 7: Experiment 1. Scene recall matching on PMC ROI. On average each subject has data for 34 recalled scenes; there are 50 possible scenes (classes) in total. So the chance accuracy is 0.02. For each method, k is selected using 4-fold cross-validation on the training data. MNI is anatomical alignment only. The mean and standard error are computed using 5 random partitions of the training and testing subjects.

information from secondary datasets yields a better generalization to left-out data. A primary dataset with many samples is less prone to overfitting; hence we expect less improvement.

In the scene recall matching experiment, the testing data in the primary dataset has a more distinct form from the secondary datasets. *sherlock* is used as the primary dataset and all the others as secondary datasets. We are interested in classifying the scenes of the *sherlock-recall* data from testing subjects while using a classifier learned from the training subjects. This experiment also has a two-phase procedure. In the first phase, we fit the model to the movie part of *sherlock* and all secondary datasets as in the time segment matching experiment. In the second phase, we test the effectiveness of the learned bases. An SVM classifier is trained with the scene label and training subjects' recall data transformed into the common feature space using the learned bases. The SVM is then used to classify testing subjects' transformed recall data.

The averaged accuracy over five random partitions with standard error is reported in Fig. 7. This experiment is only conducted on posterior medial cortex (PMC) (Tab. 1), the ROI used in [38]. We observe improved prediction performance for MDMS on the distinct scene recall task. This suggests that by leveraging secondary datasets, the primary dataset can learn spatial bases that can generalize well to distinct (but related) tasks while using several distinct stimuli. Note that in our case no secondary dataset has the same type of stimulus as the left-out data, but we still observe improvement in prediction accuracy. The ability to learn basis that works for various types of stimulus is essential when exploring new types of stimulus.

Experiment 2: Semantic Embedding We now explore if secondary datasets can help in an fMRI to text semantic embedding experiment adapted from [45]. This is a harder task than experiment 1. A major difference is that we test on completely left-out subjects instead of left-out data of training subjects, and we are testing on multiple datasets. We use *greeneyes*, *vodka*, *milky*, and *sherlock*, datasets (augmented with text annotation for each TR of the stimuli) to learn a linear mapping between the common feature space and a text semantic embedding space.

The experiment has three phases. In the first phase, we preprocess the text annotation of each TR to a 300-dimensional semantic vector using methods in [46]. We also partitioned all subjects into training and testing subjects as in experiment 1. In the second phase, we partition each dataset into two halves in time. Bases for all training subjects and k-dimensional features for each dataset are learned from one time half of all datasets. For MDDL and MDMS, bases and features are learned jointly from the multiple datasets and for other methods, bases and features are learned separately for each dataset. The features of all datasets are then concatenated along the temporal axis to form a

matrix S_{all}, and their corresponding text embeddings are concatenated in the same way to form a matrix T_{all}. The linear map Ω from the common feature space to the semantic embedding space is learned by solving

$$\min_{\Omega} \quad \|T_{all} - \Omega S_{all}\|_F^2 \qquad \text{s.t.} \quad \Omega^T \Omega = I.$$

In the last phase, Ω and our learned subject basis are tested with a scene classification task conducted on datasets *greeneyes* and *sherlock* separately (*vodka* and *milky* have too few TRs to form meaningful scenes for testing). Left-out half of *greeneyes* and *sherlock* are evenly divided into 7 and 25 non-overlapping scenes, respectively. Testing subjects' left-out half of data in dataset d is transformed to k-dimensional space with learned bases and averaged across subjects. Then, we predict the semantic embedding vectors of all scenes, as columns of T_{pred}, by mapping the averaged k-dimensional feature S_p through Ω to the semantic embedding space, $T_{pred} = \Omega S_p$. The predicted semantic vector of a scene y is then selected and correlated with the true semantic vectors of all scenes. If the highest correlation appears at scene y, then it is correctly classified. This procedure is repeated for all scenes in T_{pred}. The experiment pipeline and scene matching procedure is shown in Fig. 8. The extra requirement of a learning a linear mapping makes this experiment harder than experiment 1. The secondary datasets are leveraged when learning this linear mapping.

The accuracies averaged across five random partitions of training and testing subjects on planum temporale (PT) are reported in Fig. 8 (other ROIs in Sup. Mat.). k values are selected with cross-validations. We observe that MDMS outperforms related methods, but MDDL also performs comparably on *greeneyes*. The results show that the secondary datasets in this harder task help the analysis of the primary dataset and the potential of using secondary datasets in bridging fMRI feature space and semantic embedding space.

Experiment 3: Complete Transfer Learning We now explore transfer learning from secondary datasets to a completely left-out primary dataset. Only secondary datasets are used during the learning phase. We then test generalization of the learned subject-specific bases to the primary dataset. This is done by conducting a time segment matching experiment on the shared subjects in the primary dataset. We focus on MDMS from this point forward since it consistently performs well in the previous experiments.

We compare the generalizability of the subject bases when these are learned using one versus two secondary datasets. We also test generalizability when a small secondary dataset is used as a bridge between the primary and another larger secondary dataset. We do not expect the small secondary dataset alone to

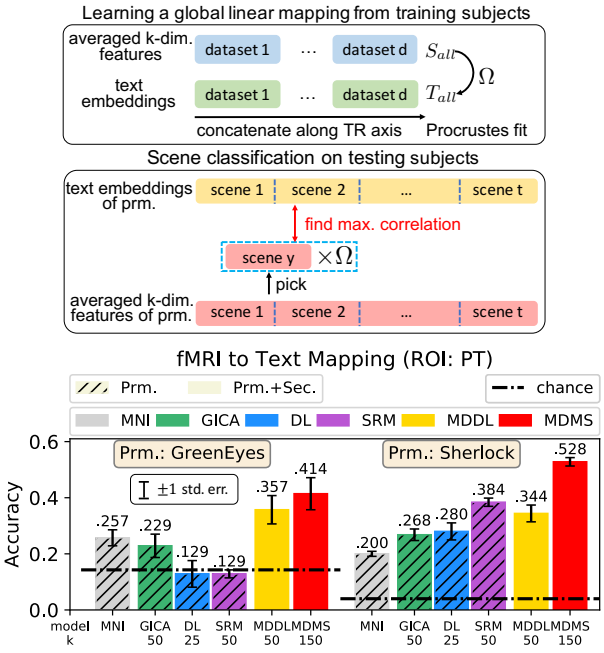

Figure 8: Experiment 2. See Exp. 2 for full details. **Top (First):** Learn a linear map Ω from shared response to text embedding space. **Top (Second):** Perform scene classification by using Ω to map a scene in shared space to semantic space, then find the best matching of scene in semantic space. **Bottom:** fMRI-data to text embedding classification accuracy. Results on other ROIs in Supp. Mat. Chance accuracy: *greeneyes*: 0.14; *sherlock*: 0.04. k selected based on 4-fold cross-validation. MNI is anatomical alignment only. Mean and standard error computed using 5 random partitions of training and testing subjects.

yield a good transfer learning. We are interested to see if the small dataset can be used as a bridge to incorporate the information in the larger secondary dataset. For this purpose, we use three datasets of different sizes, *milky*, *greeneyes*, and *sherlock*. For each primary dataset, the first secondary dataset is small but has many shared subjects with the primary dataset. The second secondary dataset is larger but may have few or no shared subjects with the primary dataset. Prediction accuracy is computed with and without the larger dataset. We observe increased accuracies after adding the larger dataset (see Fig. 9). This shows

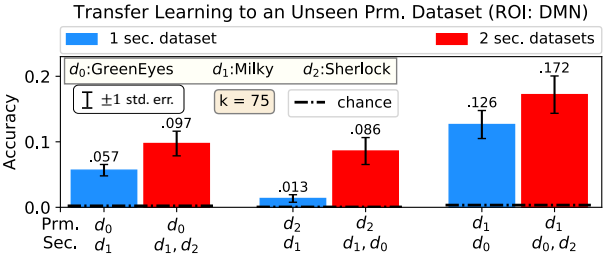

Figure 9: Experiment 3. Complete transfer learning. Time segment matching accuracy on prm. dataset using subject bases learned only from sec. datasets. Results on other ROIs in Supp. Mat. Chance accuracy: *greeneyes*: 0.0025; *milky*: 0.004; *sherlock*: 0.0005. k same as in exp. 1. Mean and st. error computed using shared subjects.

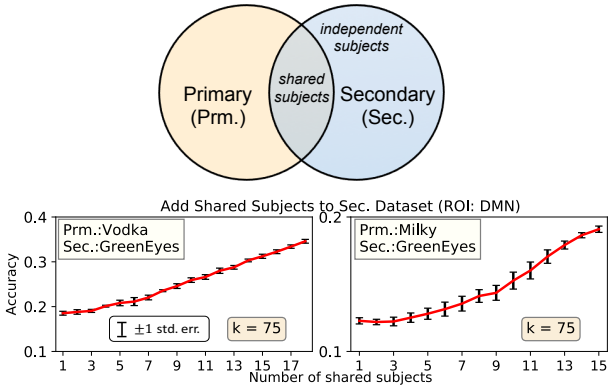

Figure 10: Experiment 4. **Top:** Shared and independent subjects. **Bottom:** Time segment matching accuracy on primary dataset as a function of the number of shared subjects. Chance accuracy and k are as in Experiment 1. Mean and standard error computed using all subjects in the primary dataset.

that we can effectively transfer the information from a secondary dataset, which has few or no shared subjects with the primary dataset, by using another secondary dataset as a bridge.

Experiment 4: Number of Shared Subjects Finally, we investigate how the number of shared subjects in a secondary dataset impacts fitting the model to the primary dataset. To evaluate this, we only use one secondary dataset. Let the independent subjects be those in the secondary, but not the primary dataset and the shared subjects be those in both datasets. This is illustrated in Fig. 10. We use the accuracy of the time segment matching experiment on the primary dataset to measure the effectiveness of integrating information from the secondary dataset. Experiment 1 indicates that small datasets have more headroom for improvement. Hence we focus on this situation. All subjects in the primary dataset are included in the basis-learning phase. The primary dataset is partitioned into two halves in time, and test accuracy is computed using the left-out half of the primary dataset.

We start with a training set containing all independent subjects in the secondary dataset, then add the shared subjects one by one. Results from two pairs of primary and secondary datasets are summarized in Fig. 10. Accuracy is averaged over both folds of the dataset. We observe significant improvement in accuracy as the number of shared subjects increases. This verifies that more shared subjects in the secondary dataset helps to learn better bases for time segment matching in the primary dataset. We also explored the role of independent subjects. More independent subjects in a single secondary dataset do not necessarily help the learning in the primary dataset (see Supp. Mat.).

5 DISCUSSION AND CONCLUSION

By assuming a fixed basis for each subject and a common feature representing the stimulus for each dataset, we pose an approach that enables multi-dataset multi-subject modeling. Following this approach, we have extended two multi-subject models into a multi-dataset setting. We observed improved performance on the primary dataset in various setups. MDMS has consistently outperformed other comparison methods in experiments. We also demonstrated that secondary datasets could be more helpful with more shared subjects with the primary dataset.

These are particularly useful in various aspects. First, the improved accuracy in predictive tasks facilitates analysis to identify important signals from the primary dataset. Second, if we are conducting a series of experiments and a subject is missing from an experiment, we provide a way to still leverage the subject's existing data in the primary analysis. Third, it reduces the cost of doing science. Collecting a large dataset is expensive. Even though there have been many efforts in collecting large datasets [47–49], they might not be directly aligned with our scientific question. We envision a scenario that there are many large datasets with stimuli provided online. For our research, we will have some subjects receiving stimuli from both the online datasets and our own designed experiment. Using these subjects to bridge the primary dataset and the online datasets as secondary datasets has the potential to improve statistical sensitivity in the analysis on the primary dataset, hence reduce the amount of data that we need to collect and result in reduced cost on data collection. Lastly, with the spirit of ensuring our research is publicly available and easily reproducible, we will release code for our implementation and experiments.

Acknowledgements

This research was performed at Princeton, and supported by NSF Grant IIS-1607801, and by a Google PhD Fellowship (P.-H. Chen). We thank U. Hasson, J. Chen, Y. Yeshurun, and C. Baldassano for sharing fMRI data, and K. Norman for fruitful discussions.

References

[1] S. A. Huettel et al. *Functional magnetic resonance imaging*. Vol. 1, Sinauer Associates Sunderland, 2004.

[2] M. J. Rosa et al. Estimating multivariate similarity between neuroimaging datasets with sparse canonical correlation analysis: an application to perfusion imaging. *Frontiers in neuroscience*, 9, 2015.

[3] I. Rustandi et al. Integrating multiple-study multiple-subject fmri datasets using canonical correlation analysis. In *Proceedings of the MICCAI Workshop*, 2009.

[4] S. J. Pan et al. A survey on transfer learning. *IEEE Trans. on knowledge and data engineering*, 2010.

[5] Y. Schwartz et al. Improving accuracy and power with transfer learning using a meta-analytic database. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 248–255, 2012.

[6] Y. Schwartz et al. On spatial selectivity and prediction

across conditions with fmri. In *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, pages 53–56. IEEE, 2012.

[7] A. van Engelen et al. Multi-center mri carotid plaque component segmentation using feature normalization and transfer learning. *IEEE transactions on medical imaging*, 34(6):1294–1305, 2015.

[8] C. Wachinger et al. Domain adaptation for alzheimer's disease diagnostics. *Neuroimage*, 139:470–479, 2016.

[9] W. Zheng et al. Personalizing eeg-based affective models with transfer learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2732–2738. AAAI Press, 2016.

[10] J. D. Cohen et al. Computational approaches to fMRI analysis. *Nature Neuroscience*, 2017.

[11] E. Dohmatob et al. Learning brain regions via large-scale online structured sparse dictionary learning. In *NIPS*, 2016.

[12] P.-H. Chen et al. A reduced-dimension fMRI shared response model. In *NIPS*, 2015.

[13] G. Varoquaux et al. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Int. Conference on Information Processing in Medical Imaging*. Springer, 2011.

[14] Y.-O. Li et al. Joint blind source separation by multiset canonical correlation analysis. *IEEE Transactions on Signal Processing*, 57(10):3918–3929, 2009.

[15] N. M. Correa et al. Canonical correlation analysis for data fusion and group inferences. *IEEE signal processing magazine*, 27(4):39–50, 2010.

[16] J. V. Haxby et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

[17] J. S. Guntupalli et al. A model of representational spaces in human cortex. *Cerebral Cortex*, 2016.

[18] J. S. Guntupalli et al. A computational model of shared fine-scale structure in the human connectome. *bioRxiv*, page 108738, 2017.

[19] V. D. Calhoun et al. A method for making group inferences from functional MRI data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

[20] V. D. Calhoun et al. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*, 45(1):S163–S172, 2009.

[21] J. R. Manning et al. Topographic factor analysis: a bayesian model for inferring brain networks from neural data. *PloS one*, 2014.

[22] J. R. Manning et al. Hierarchical topographic factor analysis. In *IEEE Pattern Rec. in Neuroimaging*, 2014.

[23] D. Bzdok et al. Semi-supervised factored logistic regression for high-dimensional neuroimaging data. In *NIPS*, 2015.

[24] P.-H. Chen et al. A convolutional autoencoder for multi-subject fMRI data aggregation. *arXiv*, 2016.

[25] J. V. Haxby et al. Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37:435–456, 2014.

[26] T. Naselaris et al. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, 2011.

[27] N. S. Ward et al. Neural correlates of motor recovery after stroke: a longitudinal fMRI study. *Brain*, 126(11):2476–2496, 2003.

[28] D. J. Simmonds et al. Protracted development of executive and mnemonic brain systems underlying working memory in adolescence: a longitudinal fMRI study. *NeuroImage*, 2017.

[29] NIMH data archive. `https://data-archive.nimh.nih.gov`.

[30] NIMH notice: Data archive data sharing terms and conditions.

[31] H. Zhang et al. A searchlight factor model approach for locating shared information in multi-subject fMRI analysis. *arXiv:1609.09432*, 2016.

[32] M. J. Anderson et al. Enabling factor analysis on thousand-subject neuroimaging datasets. *arXiv*, 2016.

[33] Y. Yeshurun et al. Same story, different story: the neural representation of interpretive frameworks. *Psychological science*, 28(3):307–319, 2017.

[34] Y. Yeshurun et al. Same story, different story: the neural representation of interpretive frameworks, 2016. `http://arks.princeton.edu/ark:/88435/dsp0141687k93v`.

[35] Y. Yeshurun et al. The butterfly effect: amplification of local changes along the temporal processing hierarchy. *bioRxiv*, page 102590, 2017.

[36] Y. Yeshurun et al. Amplification of local changes along the timescale processing hierarchy, 2017. `http://arks.princeton.edu/ark:/88435/dsp01ks65hf84n`.

[37] C. Baldassano et al. Representation of real-world event schemas during narrative perception. SfN, 2016.

[38] J. Chen et al. Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1):115–125, 2017.

[39] J. Chen. Sherlock movie watching dataset, 2016. `http://arks.princeton.edu/ark:/88435/dsp01nz8062179`.

[40] E. Simony et al. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature communications*, 7, 2016.

[41] R. J. Zatorre et al. Structure and function of auditory cortex: music and speech. *Trends in cog. sci.*, 2002.

[42] T. D. Griffiths et al. The planum temporale as a computational hub. *Trends in neurosciences*, 2002.

[43] D. Margulies et al. Precuneus shares intrinsic functional architecture in humans and monkeys. *Proceedings of the National Academy of Sciences*, 2009.

[44] J. Mazziotta et al. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Phil. Trans. of the Royal Society of London B: Biological Sciences*, 2001.

[45] K. Vodrahalli et al. Mapping between natural movie fMRI responses and word-sequence representations. *arXiv preprint arXiv:1610.03914*, 2016.

[46] S. Arora et al. A latent variable model approach to PMI-based word embeddings. *Trans. of the ACL*, 2016.

[47] R. A. Poldrack et al. Making big data open: data sharing in neuroimaging. *Nature neuroscience*, 2014.

[48] D. C. Van Essen et al. The WU-Minn human connectome project: an overview. *Neuroimage*, 2013.

[49] M. Hanke et al. A high-resolution 7-tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific data*, 1, 2014.