Integrating Multiple-Study Multiple-Subject fMRI Datasets Using Canonical Correlation Analysis

Indrayana Rustandi[1,2], Marcel Adam Just[3] and Tom M. Mitchell[1,4]

[1] Computer Science Department, Carnegie Mellon University, USA,
[2] Center for the Neural Basis of Cognition, Carnegie Mellon University, USA,
[3] Department of Psychology, Carnegie Mellon University, USA,
[4] Machine Learning Department, Carnegie Mellon University, USA.

Abstract. We present an approach to integrate multiple fMRI datasets in the context of predictive fMRI data analysis. The approach utilizes canonical correlation analysis (CCA) to find common dimensions among the different datasets, and it does not require that the multiple fMRI datasets be spatially normalized. We apply the approach to the task of predicting brain activations for unseen concrete-noun words using multiple-subject datasets from two related fMRI studies. The proposed approach yields better prediction accuracies than those of an approach where each subject's data is analyzed separately.

1 Introduction

The predictive style of fMRI data analysis, in which we try to predict some quantity of interest based on some fMRI data, has recently become more widespread (for an overview, see [1]). Nonetheless, most of the predictive approaches for fMRI data analysis have been limited in the sense that they can be applied only individually to a particular subject's data from a particular fMRI study. A few approaches have been proposed to get around this limitation. The most naïve approach is to first register the different subjects' brains to a canonical spatial coordinate frame, then pool all the data together and treat them as coming from one subject in one study. However, this approach ignores the variability that is likely to exist across subjects both within a particular study and also across studies. This can be corrected using, for instance, hierarchical Bayes techniques, as was proposed in [2]. Even then, these approaches assume that the same voxel after spatial normalization behaves similarly across subjects and studies of interest, even though the spatial normalization process is imperfect and variability in activations across subjects can still exist even after spatial normalization.

We present a new approach to integrate multiple-subject multiple-study fMRI data in the context of predictive fMRI data analysis. Unlike the approaches mentioned above, our approach treats the data for each subject as a distinct set, and it does not require that the disparate datasets be in a common normalized space. We apply our approach to the task of predicting fMRI data for new stimuli, similar to the task described in [3].

2 Methods

2.1 Datasets

We use previously analyzed datasets based on two fMRI studies: the **WP (Word-Picture)** study [3] and the **WO (Word-Only)** study [4]. In both studies, each participant was presented with stimuli corresponding to sixty concrete-noun words, which can be grouped into twelve semantic categories. In the WP study, each stimulus consisted of a line-drawing picture and the associated word label, e.g. "house". In the WO study, each stimulus consisted only of the word label, without the line drawing. In each trial, the stimulus was presented for three seconds followed by a seven-second period of fixation before the next trial started. The participants were instructed to actively think about the properties of the object described by the stimulus. Each participant went through six runs of the experiment during a single session, where each of the sixty words was presented once in each run. Data from nine (WP) and eleven (WO) right-handed adult participants are available for the analysis, including data from three participants who participated in both studies.

Acquisition and Preprocessing Parameters In both studies, fMRI images were acquired on a Siemens Allegra 3.0T scanner using a gradient echo EPI pulse sequence with TR = 1000ms. The data were processed using the SPM2 software to correct for slice timing, motion, and linear trend, and then temporally filtered using a 190s cutoff and spatially normalized into MNI space resampled to $3\times3\times6$ mm^3 voxels. The percent signal change relative to the fixation condition was computed at each voxel for each stimulus presentation and then a single fMRI mean image was created for each of the 360 trials (60 words $\times$ 6 runs) by taking the mean of the images collected 4s, 5s, 6s, and 7s after stimulus onset, to account for the delay in the hemodynamic response. To reduce the effect of noise in our analysis, for each participant in each study, we analyze the canonical image for each of the sixty words, obtained by averaging the six images associated with the corresponding word across the six presentations/runs.

2.2 Model

We analyze the datasets in the context of the predictive computational model proposed in [3], shown in figure 1. This model assumes that there are intermediate semantic features—denoted as *base features* in figure 1—that represent the meaning of each stimulus word, and that underlie the brain activations associated with thinking about that stimulus word. For example, one base feature to describe semantics of arbitrary stimulus nouns might be the frequency with which that noun co-occurs with the verb "eat" in a large collection of text. Given a particular word and its associated base features, the model assumes that each subject's brain activations can be modeled as linear combinations of the base features. Finding the mapping from the base features to the fMRI brain activations amounts to learning the coefficients (the β coefficients in figure 1) for the

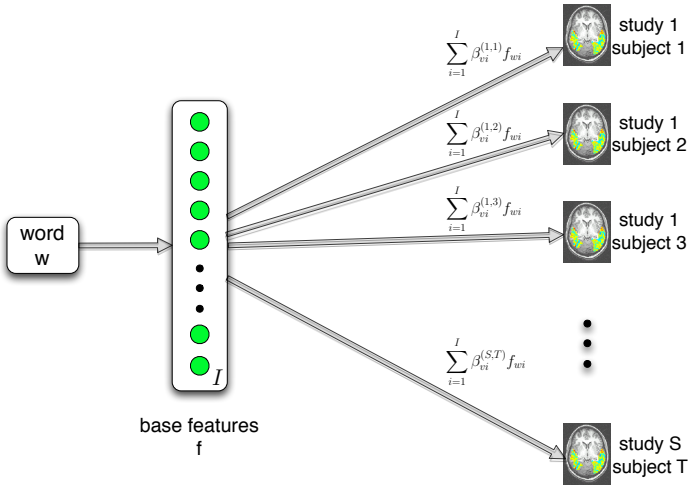

Fig. 1. The baseline predictive computational model from [3] (I base features)

multivariate linear regression problem with the base features (the f variables in figure 1) as covariates and the brain activations in each voxel as responses.

In the baseline model, there exist different mappings from the base features to the brain activations for the different participants. On the other hand, we might expect some similarity among the mappings for the various subjects in one or more related studies. Incorporating this knowledge into the model can potentially give us better predictive ability by leveraging the similar information available across subjects and studies. In addition, it potentially allows us to better quantify the similarities and differences among the various subjects and studies. Yet, we also would like to avoid the restrictions of the existing methods for integrating multiple fMRI datasets mentioned in section 1. With these in mind, we propose an enhancement to the baseline model, which we call the *CCA-mult* approach.

In the CCA-mult model, shown in figure 2, we introduce a common abstraction for brain activations for the various subjects, denoted as *learned common features* in figure 2. The learned common features are essentially the shared low-dimensional representation for the various subjects' brain activations data, and they are learned based on the regularities present in the various subjects' brain activations data. In particular, as figure 2 shows, we focus on a *linear* low-dimensional representation of the brain activations data, i.e. a low-dimensional representation such that the brain activations for each subject can be reconstructed as linear combination of the features in this representation. Now, instead of having a subject-specific direct mapping from the base features to the brain activations, we have a (linear) subject-independent mapping from the base features to the subject-independent learned features, and then subject-specific mappings from the learned common features to the brain activations. The model parameters are obtained using a two-step process. First, the subject-specific β's

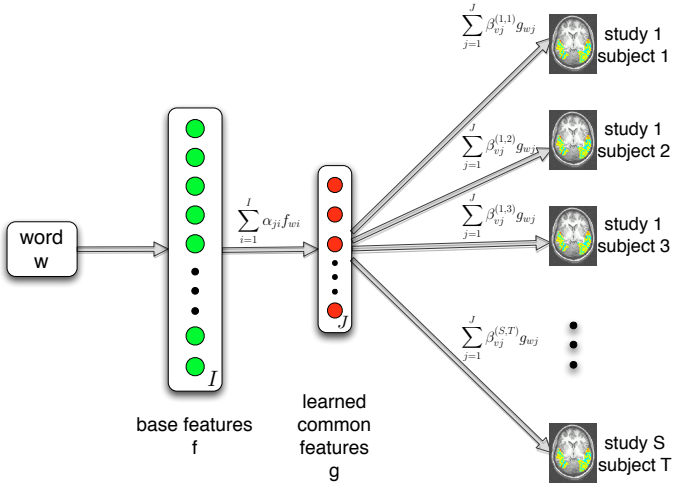

Fig. 2. The model for the CCA-mult approach (I base and J learned common features)

are estimated using canonical correlation analysis, described next. Second, the subject-independent α's are estimated using multivariate linear regression.

Learning the Common Features To learn the common features across the various subjects' brain activations data, we use *canonical correlation analysis* (CCA) [5]. The classical CCA is a multivariate statistical technique to discover correlated components across two datasets. More formally, given two datasets represented as matrices $\mathbf{X}$ ($D_X \times N$) and $\mathbf{Y}$ ($D_Y \times N$), CCA tries to find the vectors $\mathbf{w_X}$ ($D_X \times 1$) and $\mathbf{w_Y}$ ($D_Y \times 1$) such that the quantities $\mathbf{a_X} = \mathbf{w_X}^T\mathbf{X}$ and $\mathbf{a_Y} = \mathbf{w_Y}^T\mathbf{Y}$ are maximally correlated. Given this formulation, $\mathbf{w_X}$ and $\mathbf{w_Y}$ can be found as a solution to a generalized eigenvalue problem. The pair $\mathbf{a_X}$ and $\mathbf{a_Y}$ are called the first *canonical variate*, while we call $\mathbf{w_X}$ and $\mathbf{w_Y}$ the pair of *loadings* for the first canonical component. By deflating the data matrices with respect to canonical variates already found and reapplying the process, we can find subsequent canonical variates. The classical CCA can be extended to handle more than two datasets [6] [7], and to avoid overfitting, we can also regularize the loadings $\mathbf{w}_\cdot$'s similar to what is done in ridge regression [8] [7].

Tying back to the CCA-mult model, we apply CCA to the fMRI datasets, each dataset being a matrix with as many rows as voxels and as many columns as instances/trials. We then take the sample mean of the canonical variates over the different datasets as the learned common features. The loadings $\mathbf{w}_\cdot$'s define the inverse mappings from brain activations to common features. Notice that there are no restrictions that all the fMRI datasets have to have the same number of voxels, as long as we can match the instances/trials in those datasets, since CCA accepts data matrices with different numbers of rows (corresponding to voxels),

the constraint being that the matrices have to have the same number of columns (corresponding to instances). As a result, spatial normalization is not necessary.

Prediction In the baseline model, after learning the mapping from the base features to the brain activations, we can generate predicted brain activations associated with a new word by using the known base features for that word along with the learned multivariate linear regression coefficients. In the CCA-mult model, given the base features for a new word, we first generate a prediction for the common features for that word by linear regression. Given the predicted common features, we then generate the predicted brain activations for each subject by multiplying the predicted brain activations with the Moore-Penrose pseudoinverse of the loading matrix for that particular subject, obtained by aggregating that subject's loadings $\mathbf{w}_{\text{subj}}$ across all the canonical components. In essence, the pseudoinverses of the loading matrices correspond to the β's in figure 2.

Parameters We set the number of learned common features to ten ($J = 10$), using the first ten canonical variates; the optimal number to use still needs to be investigated. We use the regularized multiple-dataset version of CCA similar to that presented in [7], using 0.5 as the regularization coefficient, where the regularization coefficient ranges from 0 to 1.

3 Experiments

3.1 Setup

To compare the performance of the CCA-mult approach with that of the baseline approach, we ran experiments using the following methods:

1. **LR** The baseline method shown in figure 1.
2. **CCA-mult-subj** The CCA-mult method applied to all the subjects in a particular study, separately for the WP and the WO studies.
3. **CCA-mult-subj-study** The CCA-mult method applied to all the subjects from the two studies combined.

Besides integrating multiple datasets, the CCA-mult approach also performs dimensionality reduction. In order to contrast the contribution of dataset integration and the dimensionality reduction aspects, we also consider a fourth method (**PCA**) in which we individually run principal component analysis (PCA) on the fMRI data for each subject, and then perform a linear regression from the base features to each subject's first ten PCA components, to match the number of dimensions of the CCA-mult variations.

Evaluation To evaluate the predictive ability of these four methods, we use the cross-validation (CV) scheme described in [3]. In this scheme, for each CV fold, we hold out two words out of the sixty words used and train the model

using the data associated with the 58 remaining words ($\binom{60}{2}$ or 1770 total folds). The trained model is then used to generate the predicted fMRI activations for the two held-out words. We compare the predicted activations with the true observed fMRI activations using the cosine similarity metric as described in [3], obtaining a binary accuracy score for each fold indicating whether the model predicts the fMRI images for the two held-out words well enough to distinguish which held-out word is associated with which of the two held-out images. These results are aggregated across all the folds to obtain an accuracy figure.

Base features In [3], a set of base features derived from the statistics of a large text corpus data was used. In particular, they used co-occurrence counts of the stimulus words with a set of 25 verbs as base features, the counts derived from a collection of English Web pages collected by Google. In this paper, we consider the co-occurrence counts with the following sets of words as base features:

- 25 verbs used in [3] ($I = 25$)
- 1000 most familiar nouns from the MRC psycholinguistic database[1] ($I = 1000$)
- 1000 most familiar nouns, 1000 most familiar verbs, and 814 most adjectives, also from the MRC psycholinguistic database ($I = 2814$)

Voxel selection and data processing To decrease the effect of noise, we perform our analysis on a subset of the voxels considered relevant: in each CV fold and for each participant in each study, we rank the voxels based on the stability [3] of its activations for the 58 words used to train the model across the six presentations/runs, and choose 500 voxels with the highest stability.

3.2 Results

Figure 3 shows the average accuracies across all the subjects in each study for all the approaches, across the three sets of base features. As the figure shows, for all three sets of base features, both CCA-mult variations give better accuracies compared to the LR and PCA approaches. On the other hand, the differences in accuracies between the two CCA-mult variations are relatively small.

When using the CCA-mult method, we can also look into what kind of semantic information each of the learned common features represents. We focus on the common feature corresponding to the dominant CCA component for the CCA-mult-subj-study variation. In particular, we can see how this common feature is mapped to the entire brain by regressing it to the full-brain activations. Figure 4 shows the results for one of the participants that took part in both studies. In the WP case, we see significant loading magnitude in the fusiform gyri, highlighted by the pink ellipses, while the WO loadings exhibit significant magnitude around the superior parietal cortex, highlighted by the purple ellipses.

[1] `http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm`

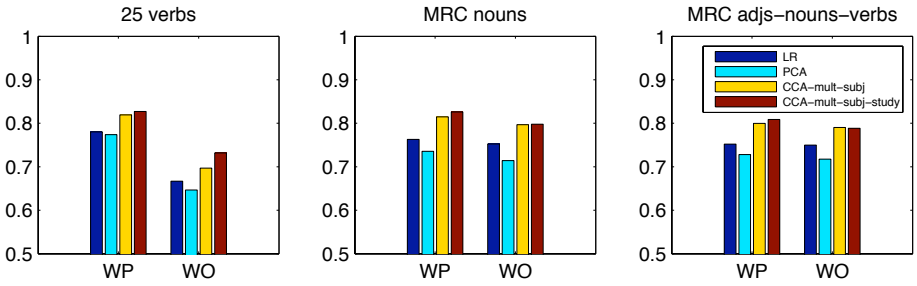

Fig. 3. Accuracies of the three sets of base features for the WP and WO datasets.

On the other hand, white ellipses denote some of the areas exhibiting similar loading values in both cases, located among others in the left extrastriate, left pars opercularis, and left precentral gyrus.

Next, we can check whether this common feature represents some semantic dimensions by looking at the distributions of its value across all sixty stimulus words. The top five stimulus words with the most positive value for the first common feature—*knife, cat, spoon, key, pliers*—roughly represent the "manipulability" concept, while the top five stimulus with the most negative value for the same feature—*apartment, church, closet, house, barn*—roughly represent the "shelter" concept, mirroring some of the findings of [4].

4 Conclusion

We have presented the CCA-mult approach to integrate data from multiple subjects and multiple fMRI studies. The CCA-mult approach does not require that the datasets be spatially normalized. Our results show that by using the lower-dimensional feature space discovered by the CCA-mult method, we obtain better accuracies compared to using the baseline approach from [3] and to using the dataset-specific lower-dimensional feature space discovered through PCA.

The experiments reported in this paper support our thesis that it is possible to train more accurate computational models by integrating training data from multiple subjects participating in multiple related fMRI studies, by incorporating latent variables that capture commonalities across subjects, yet still allow the model to estimate parameters that are specific to each participant and study. Given that many fMRI analyses are limited by the sparsity of training data relative to the complexity of the phenomena to be modeled, it is important to develop models like ours that integrate data from multiple subjects and studies. To that end, one specific direction for future research on our model is to remove its current restriction that the different studies have matched trials (e.g., in our case study, each data set must include the same 60 semantic stimuli). We are currently exploring methods that relax this assumption, to enable training a model from data in which different subjects are presented with different stimuli.

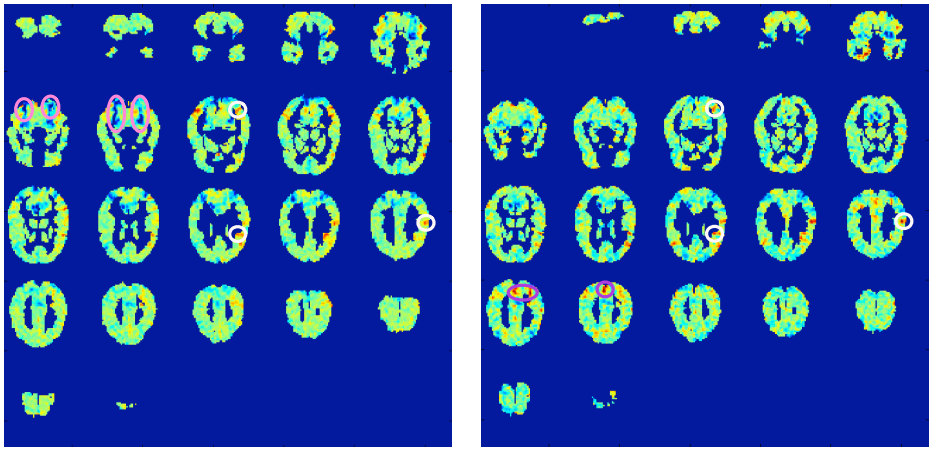

Fig. 4. The full-brain loadings for one participant in both the WP study (left) and the WO study (right). White ellipses denote significant loadings present in both cases, while pink ellipses denote significant loadings present in only the WP case, and purple ellipses denote significant loadings present in only the WO case.

5 Acknowledgments

We thank Vlad Cherkassky for preprocessing the fMRI data and Kai-min Chang for referring us to the MRC database. This research was supported by a grant from the Keck Foundation, an NSF CCBI research grant, and also NSF through TeraGrid resources provided by the Pittsburgh Supercomputing Center.

References

1. Haynes, J.D., Rees, G.: Decoding mental states from brain activity in humans. Nature Reviews Neuroscience **7** (2006) 523–534
2. Rustandi, I.: Classifying Multiple-Subject fMRI Data Using the Hierarchical Gaussian Naïve Bayes Classifier. In: 13th Conference on Human Brain Mapping. (2007)
3. Mitchell, T.M., et al.: Predicting human brain activity associated with the meanings of nouns. Science **320** (2008) 1191–1195
4. Just, M.A., et al.: A neurosemantic theory of concrete noun representation based on the underlying brain codes. Under review
5. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4) (1936) 321–377
6. Kettenring, J.R.: Canonical analysis of several sets of variables. Biometrika **58**(3) (1971) 433–451
7. Hardoon, D.R., et al.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation **16** (2004) 2639–2664
8. Vinod, H.D.: Canonical ridge and econometrics of joint production. Journal of Econometrics **4** (1976) 147–166