A CROSS-MODAL ADAPTATION APPROACH FOR BRAIN DECODING

Pouya Ghaemmaghami, Moin Nabi, Yan Yan, Giuseppe Riccardi and Nicu Sebe

Department of Information Engineering and Computer Science, University of Trento, Italy

ABSTRACT

Brain decoding has become a hot topic in many recent brain studies. In a typical neuroimaging experiment, participants are presented with different categories of stimuli while their concurrent brain activity is recorded. Then a classifier is trained on the features extracted from the recorded brain data to discriminate different target stimuli classes. It is a common practice to hypothesize that the stimulus-related information exists in the brain data if the decoder can accurately predict the target stimulus category. However, most of the neuroimaging studies suffer from few and noisy samples. These constraints affects the performance of such decoding systems. In order to cope with this limitation, a dictionary learning approach is used in this paper to transfer knowledge from the multimedia domain to the brain domain. We show that such cross-modal domain adaptation yields better performance of the learning algorithm in the brain domain. This is the first study in the direction of cross-modal adaptation by joint dictionary learning on multimedia and brain modality.

Index Terms— brain decoding, signal processing, multimedia information retrieval, domain adaptation, genre classification

1. INTRODUCTION

Brain Computer Interface (BCI) has recently become a hot topic outside neuroscience and rehabilitation communities. Other disciplines such as artificial intelligence have already started to contribute to the field by bringing into play machine learning and signal processing algorithms to brain data. Recent progress in brain studies demonstrates the possibility of brain decoding, which is typically a classification of stimuli into a set of categories. In a typical brain decoding paradigm, experimental participants are presented by different categories of stimuli while their brain activity is simultaneously being recorded. Then a machine learning algorithm is employed to categorize the features extracted from the measured signal into the target stimuli classes [1, 2, 3, 4, 5, 6].

However, recording brain data is very costly resulting in very few samples. Additionally, the recordings are very noisy due to the low signal-to-noise ratio and the non-stationarity nature of the signals. These two constraints lead to a sudden drop in the performance of machine learning algorithms. For example, in [7, 8], authors show the possibility of music and movie genre classification using brain signals but they obtained better results on the same dataset when low-level multimedia features were used instead of brain features.

Most of the machine learning approaches work well when being trained and tested in one specific domain. They may however fail when applied to another domain, i.e., the distribution changes, which is the case in many real world applications. In such cases, transferring knowledge across domains and tasks would be desirable. This is the situation where the performance in the target domain depends on the performance in the source domain and the similarity between the source and the target domain. Transfer learning can be very beneficial in the cases where collecting data is immensely difficult and costly [9, 10]. Such situations emerge often in the neuroimaging studies.

Sparse dictionary learning is one of the most widely used approaches for domain adaptation where the goal is finding sparse representations to minimize domain divergence and model error. The strategies to find these representations are dataset dependent and can be supervised or unsupervised (if there is no label information) [9]. Prior works in neuroimaging studies have shown that low-level audio-visual features such as orientation, direction of motion and color of visual stimulus are encoded in the human brain[11, 12, 13]. Some of these low-level audio-visual features were used also in multimedia retrieval literature for specific tasks (e.g,. genre classification [14, 15]). Inspired by these facts, in this paper, we address the specific problem of cross-modal adaptation by learning jointly a sparse dictionary on the low-level audio-visual features and brain features. We are the first showing that such cross-modal adaptation is feasible between multimedia and brain features. The method has been tested on two neuroimaging datasets: DECAF dataset [16] and DEAP dataset [17]. These datasets contain the magnetoencephalography (MEG) and electroencephalography (EEG) data of subjects who watched music/movie clips and they have been used previously by [7, 8] regarding music/movie genre classification. Motivated by recent successes in domain adaptation in machine learning literature [18, 19, 20, 21], we hypothesize that brain/multimedia adaptation can be done successfully and our evaluation confirms that such adaptation shows a significant performance gain. Figure 1 illustrates the overview of the framework proposed in this study.

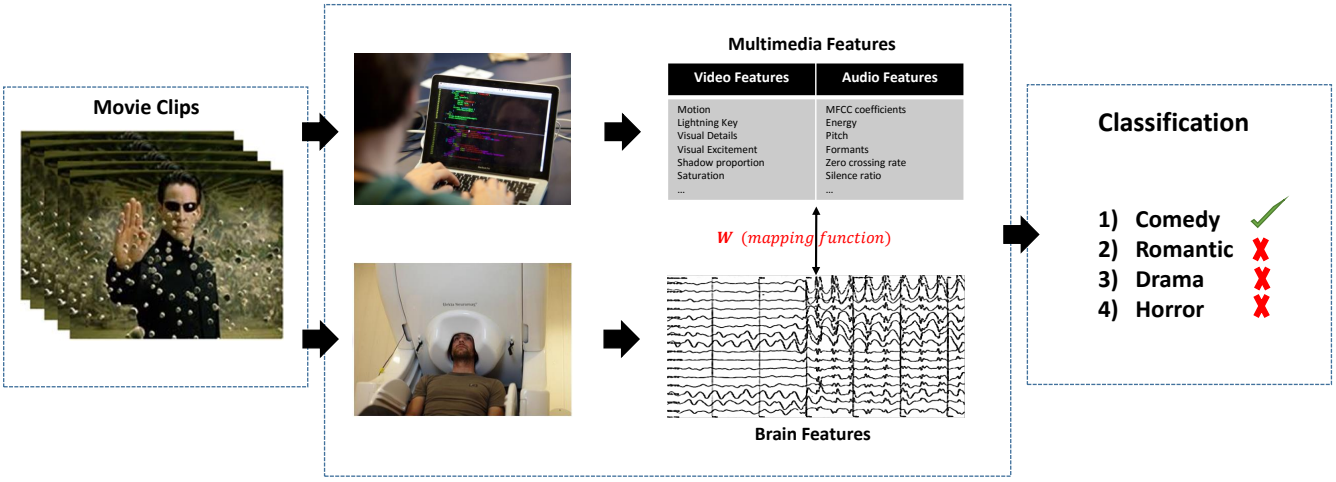

Fig. 1. Overview of our proposed framework: During training, a dictionary learning approach is used to learn a mapping function for brain/multimedia adaptation. Once the mapping function is learned, the genre of a test movie clip is predicted using the adapted brain features.

The remainder of this paper is structured as follows: In section 2 we present our experimental setup including the dataset, feature extraction steps and our adaptation procedure. Section 3 elaborates our experimental results with a brief discussion. And finally section 4 concludes the paper and highlights some future directions.

2. MATERIALS AND METHOD

In this section, we describe the used datasets, feature extraction scheme, and the adaptation procedure.

2.1. Datasets

In this study, we employed two publicly available neuoroimaging cross-modal datasets.

DEAP dataset[1]: This dataset contains the electroencephalographic (EEG) data of 32 participants as each watched 40 music video clips. These music video clips were projected onto a screen at a screen refresh rate of 60 Hz that was located about a meter in front of the subject. The electroencephalographic data were recorded using a 32 channel Biosemi ActiveTwo system at a sampling rate of 512 Hz.

DECAF dataset[2]: This dataset contains the magnetoencephalographic (MEG) data of 30 participants as each watched 36 movie clips and 40 music video clips. These clips were projected (20 frames/second) onto a screen located about a meter in front of the subject inside the acquisition room. The MEG data were recorded using a 306 channel Electa Neuromag device (102 magnetometers and 204 gradiometers) with 1KHz sampling rate in a magnetically shielded room with controlled illumination.

We specifically selected these two datasets in this study because, for each music/movie clip, the corresponding brain features (i.e. MEG features and EEG features) and multimedia features can be extracted.

2.2. Annotation

Music genres: Each music clip (in both datasets) is labeled with one of the following two broad genres: Pop or Rock (see [8] for the details of genre annotation). Note that the music video clips used in the DECAF dataset are the same clips used in the DEAP dataset.

Movie genres: Each movie clip is assigned with a label out of the following four genres: Comedy, Romantic, Drama and Horror (see [7] for more details on genre annotation).

2.3. Source Domain: Multimedia Features

Following [16, 7], for each music/movie clip, the low-level audio-visual features are extracted. These low-level Multimedia Content Analysis (MCA) features include 49 video features and 56 audio features (see table 1). These MCA features are extracted for each second of the movie clips and then they were averaged by the length of the clip.

2.4. Target Domain: Brain Features

MEG Features: Following [7], the MEG data is processed as follows:

[1] http://www.eecs.qmul.ac.uk/mmv/datasets/deap/

[2] http://mhug.disi.unitn.it/wp-content/DECAF/DECAF.html

Table 1. Extracted MCA features for each music/movie clip (the number of features is listed in the parenthesis).

Audio features	Video features
MFCC features (13)	Motion (1)
Derivative of MFCC features (13)	Visual Excitement (1)
AMFCC features (13)	Visual Details (1)
Energy (1)	Color Variance (1)
Pitch (1)	Lighting Key (1)
Zero crossing rate (1)	Shadow Proportion (1)
Silence ratio (2)	Grayness (1)
Formants (4)	Saturation for frames (1)
MSpectrum flux (2)	Lightness for frames (1)
Spectral centroid (2)	Lightness (20)
Delta spectrum magnitud (2)	Hue (20)
Band energy ratio (2)	
56 Audio features	**49 Video features**

1. Down-sampling the MEG signal to 300 Hz.
2. Bandpass frequency filtering (1 - 95 Hz).
3. Estimating the spectral power of the 102 combined-gradiometer sensors of each trial with a window size of 300 samples.
4. Calculating MEG features by averaging the signal power over time and over four frequency bands: theta (3:7 Hz), alpha (8:15 Hz), beta (16:31 Hz) and gamma (32:45 Hz). The output of this procedure for each trial is a matrix with the following dimensions: 102 (number of the MEG sensors) $\times$ 4 (frequency bands).

EEG Features: We used the same pre-processed EEG data as in [17]. These pre-processing steps are as follows:

1. Down-sampling the EEG signal to 128 Hz.
2. EOG artifacts removal.
3. Bandpass frequency filtering (1 - 45 Hz).
4. Estimating the spectral power of each channel of the EEG trials with a window size of 128 samples.
5. Calculating EEG features by averaging the signal power over time and over four frequency bands: theta (3:7 Hz), alpha (8:15 Hz), beta (16:31 Hz) and gamma (32:45 Hz). The output of this procedure for each trial is a matrix with the following dimensions: 32 (number of the EEG sensors) $\times$ 4 (frequency bands).

2.5. Domain Adaptation

To benefit sparsity-inducing properties, we first sparsify the features in both modalities. Once sparse representations are obtained, we adopted the Semi-Coupled Dictionary Learning (SCDL) approach [22] in order to adapt the sparse MCA features to the sparse brain features. This was done for each subject separately. We refer to these features as **Adapted-Brain** features.

The intuition behind such cross-modal adaptation is that a mapping function can be found to associate the given sample in the brain domain to the corresponding sample in the multimedia domain. Since each pair of samples in two modalities refer to the same video clip, it is reasonable to assume that there exists a hidden space where the knowledge can be transferred across the two modalities. Therefore, we employ a coupled dictionary learning method with the assumption that there exists a dictionary pair over which the representations of two modalities have a stable mapping. Once the dictionary pair and mapping are learned, cross-modal domain adaptation can be performed.

We denote $\mathbf{X}$ and $\mathbf{Y}$ as source and target domain feature matrix, respectively. $\mathbf{D}_x$ and $\mathbf{D}_y$ are the dictionaries learned in the source and the target domain. $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ are the codes learned in the source and the target domain. We propose to optimize the following objective function below:

$$\begin{aligned} &\min_{(\mathbf{D}_x,\mathbf{D}_y,\mathbf{W})} \|\mathbf{X}-\mathbf{D}_x\mathbf{\Lambda}_x\|_F^2+\|\mathbf{Y}-\mathbf{D}_y\Lambda_y\|_F^2 \\ &+\gamma\|\mathbf{\Lambda}_y-\mathbf{W}\mathbf{\Lambda}_x\|_F^2+\lambda_x\|\mathbf{\Lambda}_x\|_1+\lambda_y\|\mathbf{\Lambda}_y\|_1+\lambda_w\|\mathbf{W}\|_F^2 \\ &s.t. \quad \|d_{x,i}\|_{l_2}\leq 1, \|d_{y,i}\|_{l_2}\leq 1, \quad \forall i \end{aligned} \tag{1}$$

where γ, λ_x, λ_y, λ_W are regularization parameters to balance the terms in the objective function. The objective function in (1) is not jointly convex to $\boldsymbol{D}_x$, $\boldsymbol{D}_y$, $\boldsymbol{W}$. However, it is convex w.r.t. each of them if others are fixed. An iterative algorithm is designed to alternatively optimize the variables.

3. EXPERIMENTS AND RESULTS

For the sake of compatibility with [7, 8], we employed the same classifiers under the leave-one-clip-out cross-validation schema to classify brain features (*Brain* features and *Adapted-Brain* features) into the target genre classes. Such evaluation, provides us with comparing brain features before and after adaptation. The above-mentioned pipeline was performed in the two following scenarios: Subject-level analysis and Population-level analysis.

3.1. Subject-level analysis

3.1.1. movie genre classification

Following [7], at subject level, the Naive Bayes classifier was employed on the brain data of each subject separately. Then, the average accuracy of all subjects using MEG features (Brain) and Adapted-MEG features (Adapted-Brain) are obtained. The results of such analysis is demonstrated in Figure 2 (DECAF-MOVIE). The significant difference ($p-value = 0.0038$) between the average accuracy of Adapted-MEG features (0.42 ± 0.05) and MEG features (0.36 ± 0.11) suggests the efficacy of adapting the brain domain to the multimedia domain.

Besides genre, we also adopt our classifier to classify movie clips using their affective labels provided in [16]. Table2 shows the average accuracy of all subjects using MEG features and Adapted-MEG features. The result of the Adapted-MEG features is significantly ($p - value = 0.0224$) better than the MEG features.

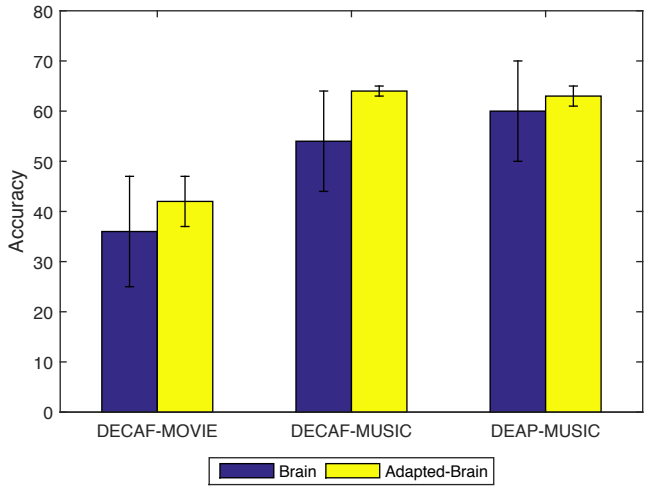

Fig. 2. Comparison between the accuracy of Brain and Adapted-Brain features in classifying the genre of the music/movie clip in the single-subject level scenario on different datasets.

3.1.2. music genre classification

Following [8], at the subject level, A Linear SVM classifier was employed on the brain data for each subject separately in both DECAF and DEAP datasets. Figure 2 (DECAF-MUSIC and DEAP-MUSIC) compares the results of music genre classification using brain features (MEG and EEG) before and after adaptation. In both DECAF and DEAP datasets, the distribution of the obtained classification accuracies using Adapted-Brain features is far superior to Brain features. This difference implies the effectiveness of adapting brain domain to multimedia domain. In the case of DECAF-MUSIC dataset, this difference is significant ($p - value < 0.001$).

Table 2. Comparison between the accuracy of MEG and Adapted-MEG features in classifying the affective content in the single-subject level scenario.

Feature-Space	Affective Content (valence) Accuracy
MEG	0.55 ± 0.10
Adapted-MEG	$\mathbf{0.61} \pm 0.08$

3.2. Population-level analysis

To evaluate the efficacy of the Adapted-Brain features at the population level, the genre of each music/movie clip is computed by majority voting over the predicted labels of single-subject predictions across all subjects.

The results are summarized in Table 3. In case of movie genre classification (DECAF-MOVIE), the population level accuracy for Adapted-MEG features is 63.9% which is higher than the accuracy of MEG features (55.6%). In case of music genre classification using MEG signals (DECAF-MUSIC), the population level accuracy for Adapted-MEG features is 65% and this is also higher than the accuracy of MEG features (57.5%). However, in the case of music genre classification using EEG signals (DECAF-MUSIC), the population level accuracy for Adapted-EEG features is 62.5% which is below the accuracy of EEG features (75%). Considering the higher accuracy of the Adapted-EEG features in the Subject-Level analysis, this phenomena is probably due to the low agreement between the predictions of all subjects.

Table 3. Comparison between the accuracy of Brain features and Adapted-Brain features in the population-level analysis.

Dataset	Feature-Space	Accuracy
DECAF-MOVIE	MEG [7]	55.6%
	Adapted-MEG	**63.9%**
DECAF-MUSIC	MEG [8]	57.5%
	Adapted-MEG	**65%**
DEAP-MUSIC	EEG [8]	**75%**
	Adapted-EEG	62.5%

4. CONCLUSIONS

In this paper, we applied a dictionary learning approach for the brain/multimedia adaptation. Despite the difference between these two modalities, our adaptation procedure outperformed the previous state of the art of the movie/music genre classification task (using brain signals). We evaluated our approach on two different neuroimaging modalities (MEG and EEG) and our cross-modal domain adaptation approach led to improved results in both of them. We believe that such approaches can overcome the limitations of the neuroimaging studies (namely, few and noisy samples) and consequently boost the performance of the decoding algorithms. As future work, we will explore such adaptation procedure by employing other neuroimaging modalities (e.g. fMRI) on other tasks (e.g. Action Recognition) .

5. REFERENCES

[1] Francisco Pereira, Tom Mitchell, and Matthew Botvinick, "Machine learning classifiers and fmri: a

tutorial overview," *Neuroimage*, vol. 45, no. 1, pp. S199–S209, 2009.

[2] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller, "Introduction to machine learning for brain imaging," *Neuroimage*, vol. 56, no. 2, pp. 387–399, 2011.

[3] Mo Chen, Junwei Han, Xintao Hu, Xi Jiang, Lei Guo, and Tianming Liu, "Survey of encoding and decoding of visual stimulus via fmri: an image analysis perspective," *Brain imaging and behavior*, vol. 8, no. 1, pp. 7–23, 2014.

[4] Raheel Zafar, Aamir Saeed Malik, Nidal Kamel, Sarat C Dass, Jafri M Abdullah, Faruque Reza, and Ahmad Helmy Abdul Karim, "Decoding of visual information from human brain activity: A review of fmri and eeg studies," *Journal of integrative neuroscience*, vol. 14, no. 02, pp. 155–168, 2015.

[5] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva, "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence," *Scientific reports*, vol. 6, 2016.

[6] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Oliva, "Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks," *NeuroImage*, 2016.

[7] Pouya Ghaemmaghami, Mojtaba Khomami Abadi, Seyed Mostafa Kia, Paolo Avesani, and Nicu Sebe, "Movie genre classification by exploiting MEG brain signals," in *ICIAP*. 2015.

[8] Pouya Ghaemmaghami and Nicu Sebe, "Brain and music: Music genre classification using brain signals," in *EUSIPCO*, 2016.

[9] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[10] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding," in *ICML*, 2009.

[11] Kalanit Grill-Spector and Kevin S Weiner, "The functional architecture of the ventral temporal cortex and its role in categorization," *Nature Reviews Neuroscience*, vol. 15, no. 8, pp. 536–548, 2014.

[12] Yukiyasu Kamitani and Frank Tong, "Decoding motion direction from activity in human visual cortex," *Journal of Vision*, vol. 5, no. 8, pp. 152–152, 2005.

[13] Keiji Tanaka, "Mechanisms of visual object recognition: monkey and human studies," *Current opinion in neurobiology*, vol. 7, no. 4, pp. 523–529, 1997.

[14] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah, "On the use of computable features for film classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 52–64, 2005.

[15] Howard Zhou, Tucker Hermans, Asmita V Karandikar, and James M Rehg, "Movie genre classification via scene categorization," in *ACM Multimedia*, 2010.

[16] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.

[17] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[18] Nitish Srivastava and Ruslan Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Journal of Machine Learning Research*, vol. 15, pp. 2949–2980, 2014.

[19] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa, "Visual domain adaptation: A survey of recent advances," *Signal Processing Magazine, IEEE*, vol. 32, no. 3, pp. 53–69, 2015.

[20] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko, "Simultaneous deep transfer across domains and tasks," in *ICCV*, 2015.

[21] Saurabh Gupta, Judy Hoffman, and Jitendra Malik, "Cross modal distillation for supervision transfer," *CVPR*, 2016.

[22] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *CVPR*, 2012.