DEEP MULTIWAY CANONICAL CORRELATION ANALYSIS FOR DECODING THE AUDITORY BRAIN

Jaswanth Reddy Katthi, Rohit Kumar

ABSTRACT

Decoding of the auditory brain for an acoustic stimulus involves finding the relationship between the audio input and the brain activity measured in terms of Electroencephalography (EEG) recordings. Prior methods in this domain focus on analysing a subjects' activity separately using linear analysis methods like Canonical Correlation Analysis (CCA) and non-linear methods like Deep CCA. A recent attempt was made, called multiway CCA, to combine the brain activity readings from a bunch of subjects and extract useful information from each subject which is irrespective of the subject to obtain a large dataset of stimulus and response to work with. In this project, we tried to introduce a deep learning framework to perform correlation analysis in this setup. We try to replace the block of multiway CCA, which is one linear formulation of a Generalized Canonical Correlation Analysis with a deep version of Generalized CCA. The corresponding results obtained in performing the existing multiway CCA method onto the data and the comparison of the correlations obtained for each subject with and without the influence of other subjects' data are presented.

1. CANONICAL CORRELATION ANALYSIS

1.1. Linear CCA

For a pair of multi-variate datasets, Canonical Correlation Analysis (CCA) [1] solves the problem of finding the optimal linear transforms that maximize the Pearson correlation between the transformed vectors.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote $\mathcal{D}_1$ and $\mathcal{D}_2$ dimensional vectors respectively. Let n denote the dimension of the canonical sub-space where the correlation between transformed vectors is maximal. For example, if $n = 1$, let $\boldsymbol{u_1}$, $\boldsymbol{v_1}$ denote the pair of vectors which project $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively into 1-dimensional space. Now, the problem is to find $\boldsymbol{u_1}$ and $\boldsymbol{v_1}$ such that the correlation, ρ, between $x' = \boldsymbol{u_1}^T\boldsymbol{x}$ and $y' = \boldsymbol{v_1}^T\boldsymbol{y}$ is maximized. The problem can be equivalently given as to maximize,

$$\rho = \frac{\boldsymbol{u_1}^T\mathbf{C}_{xy}\boldsymbol{v_1}}{\sqrt{\boldsymbol{u_1}^T\mathbf{C}_{xx}\boldsymbol{u_1}\boldsymbol{v_1}^T\mathbf{C}_{yy}\boldsymbol{v_1}}} \tag{1}$$

where, $\mathbf{C}_{xy} = E[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{y} - \boldsymbol{\mu_y})^T]$ and $\mathbf{C}_{xx}$, $\mathbf{C}_{yy}$ are the auto-correlation matrices of $\boldsymbol{x}$, $\boldsymbol{y}$ respectively.

Let $\boldsymbol{T} \triangleq \mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}$. Then, the solution to the CCA problem ($\mathbf{u}_1^*$ and $\mathbf{v}_1^*$) are given as the first left and right singular vectors of the $\boldsymbol{T}$ matrix and the maximum correlation is the top singular value [2]. This can be extended for $n > 1$ also by finding the subsequent singular vectors.

1.2. Linear Multiway Correlation Analysis

The goal of MCCA [3],will be to find dimensions in multi variant data that maximize the correlation between the multiple data sets.

Consider the data in the l^{th} set with $\mathrm{x}_i^l \in \mathbb{R}^{d_l}$,where i=1,...,T enumerates exemplars,l=1,...,N enumerated the dataset,and d_l are the dimension of each data set.Where the dimensions D=$\sum_{i=1}^{N} d_l$,will be for total of N dataset.The goal is to identify on each dataset a projection vector $v^l \in \mathbb{R}_l^d$ such that inter-set correlation(ISC), is maximized.The problem can be equivalently given as to maximize .

$$\rho = \frac{\sum_{l=1}^{N}\sum_{k=1,k\neq l}^{N}\boldsymbol{v}^{lT}\boldsymbol{R}^{lk}\boldsymbol{v}^{k}}{(N-1)\sum_{l=1}^{N}\boldsymbol{v}^{lT}\boldsymbol{R}^{ll}\boldsymbol{v}^{l}} \tag{2}$$

where R^{lk} are the cross covariance matrices between x_i^l and x_i^k:

$$R^{lk} = \sum_{i=1}^{T}(x_i^l - \overline{x}_*^l)(x_i^l - \overline{x}_*^l)^T \tag{3}$$

where $\overline{x}_*^l = T^{-1}\sum_{i=1}^{T} x_i^l$,is the sample mean for the data set l.

1.3. Deep MCCA

The extension of the linear transformation based CCA analysis to deep transformation learning based CCA was first proposed by Andrew et.al [2]. Two input sets of vectors are passed through a pair of feed-forward connections to undergo a set of non-linear transformations. The outputs of each network are the final representations on which the cross correlation is computed. The neural network is trained to maximize the correlation cost.

Let the non-linear transform performed by the first neural network on $\boldsymbol{x}$ be denoted as $f_1(\cdot)$. Similarly, let the second network transformation on $\boldsymbol{y}$ be denoted as $f_2(\cdot)$. Let $\boldsymbol{\theta_1}$ be the set of all trainable parameters of the first neural network and $\boldsymbol{\theta_2}$ be that of the second network. The deep CCA optimization can be given as,

$$(\boldsymbol{\theta_1^*}, \boldsymbol{\theta_2^*}) = \underset{(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})}{\operatorname{argmax}} \operatorname{corr}\left(f_1\left(\boldsymbol{x}; \boldsymbol{\theta_1}\right), f_2\left(\boldsymbol{y}; \boldsymbol{\theta_2}\right)\right) \tag{4}$$

Now, we tried to push the deep CCA method in the traits of linear MCCA [3]. The deep version formulation of the generalized CCA is adopted from [4]. In the paper, the linear Generalized CCA formulation of finding the shared representation (say G) of J different views X_j and U_j is the transform for j^{th} view is as follows :

$$\min_{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}} \left\| G - U_j^T X_j \right\|_F^2 \text{ subject to } GG^T = I_r \tag{5}$$

And the deep version of the Generalized Canonical Correlation Analysis i.e., DGCCA formulation is as follows :

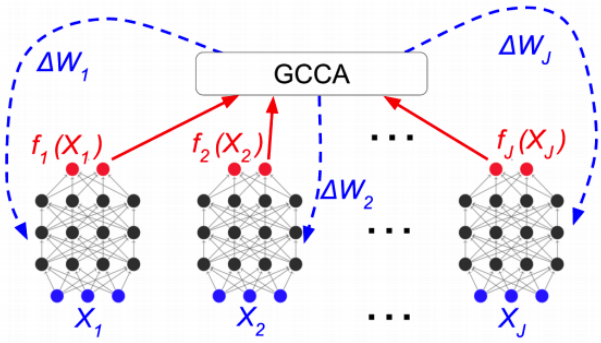

Fig. 1. The J different views provided to the DGCCA block which finds the final representations are processed EEG, collected from the $J-1$ subjects and 1 view which represents the processed audio stimuli which is common to all the subjects.

$$\min_{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}} \left\| G - U_j^T f_j(X_j) \right\|_F^2 \text{ subject to } GG^T = I_r \tag{6}$$

If we denoted the $f_j(\cdot)$ as the nonlinear transformation of the j^{th} neural network, then we can see that the representations X_j are replaced by the final representations obtained by passing j^{th} view through the j^{th} neural network. It is same as GCCA equation where $G \in \mathbb{R}^{r \times N}$ is the shared representation we are interested in learning.

The work in [4] showed that the DGCCA learned a nonlinear mapping that does remarkably well at making a mixture previously linearly non-separable mixture components into a linearly separable mixture. Results also show improvements in phoneme classification when acoustic and articulatory data as the two views and phoneme labels as the third view for GCCA and DGCCA.

In our work, we have followed the architecture of [2] used for finding the representations of the left and right halves of MNIST hand-written digit images such that the correlation between them is maximized. [2] has shown that the correlation can be increased significantly by using the deep CCA model over the linear CCA model.

In this work, we have tried to generalize the goal of obtaining better representations of a subject's EEG and the auditory stimulus which was previously tried to capture using DCCA accessing each subject's data separately. We combine the EEG data of all the subjects and the common stimulus to find representations for all subjects such that the common signals from each subject corresponding to the stimulus is obtained, taking the advantage of having more data from different subjects.

The main goal is to solve the problem of not having a lot of EEG data from a single subject by tapping the EEG available from many subjects, by finding the subject-independent/signals-common-across-subjects representations using the DGCCA formulation.

2. EXPERIMENTS AND RESULTS

The linear MCCA analysis performed in [3] forms the baseline for this work. We use the same stimuli response data collected by Liberto et. al. [5]. Specifically, the EEG recordings from 128 channels are recorded when subjects are listening to a male speaker reading snippets of a novel. A Biosemi system was used for EEG data collection which was sampled at 512 Hz. We use 20 speech excerpts, each of duration approximately 3 minutes presented diotically via headphones. The EEG data were down-sampled to 64 Hz. It was further processed using de-trending and de-noising using noise tools software [6]. The data were processed with band-pass filtering between $0.1 - 12$ Hz. The stimulus data was obtained from audio sampled at 44100 Hz. The audio envelope was obtained by a squaring and smoothing operation by convolution with a square window and downsampled to 64 Hz. The envelope was further compressed to the power $1/3$. In all our experiments, we perform DCCA projection to one dimension and compare with the linear CCA projection to one dimension. More details about the EEG pre-processing and the audio envelope extraction are described in De Chevigne [7].

Table 1. Linear MCCA used for De-noising matrix

Users	Linear CCA3	MCCA(no stimulus)	MCCA(stimulus view)
User 1	0.22	0.23	**0.23**
User 2	0.25	0.25	**0.26**
User 3	0.16	0.10	0.15
User 4	0.29	0.3	**0.31**
User 5	0.32	0.28	**0.32**
User 6	0.31	0.29	**0.32**

We have tested the linear MCCA and the Deep GCCA on the prepocessed EEG collected from 6 subjects.

The linear part is done as : In linear MCCA part,MCCA has been used to obtain the denoising matrix and than that denoising matrix has been applied to CCA. The main idea was that each data matrix X_n may be denoised by projecting it to the overcomplete basis of Canonical correlates,selecting the first m ¡ D components,and than projecting back.This is refer as "denoising",as it can be used to attenuate the components that are least shared across subjects,likely to be the noise.This can be summarized by a denoising matrix D_n product of the first 'm' columns of subject specific projection matrix V_n,by the first 'm' rows of its pseudo inverse. The thing we tried to do is in calculation of denoising matrix,considered the stimulus response as one of the view and try to estimate the denoising matrix than.The parameters are important while doing this,i.e 30 PCs are kept in the first PCA,resulting in 128*280 analysis matrix(6 for users and last 40 are stimulus matrix providing 40 delay to it).Than the first 50 columns of this matrix were multiplied by the first 50 rows of its pseudo inverse to yield a 128*128 subject specific denoising matrix. Results can been seen in table 1.

The deep GCCA is tested as follows : All the subjects' EEG data is sent through a a filter-bank of 21 FIR band-pass filters whose characteristics like centre frequency, bandwidth and duration of impulse response are uniformly distributed on a logarithmic scale, separately for each subject.The 21 D output of the audio envelope after passing through the filter-bank is used as the representation from the stimulus side, which is common for all the subjects. The 60 D PCA output of the original 128 channel EEG through the 21 filter filter-bank gives 1260 dimensional response for each subject. This high dimensional vector is processed with another PCA transformation to 139 dimensional output for each subject separately. Now, the six EEG response views of the 6 subjects plus 1 stimulus view, combined together as 7 different views of the same data is provided to the corresponding number of different neural networks such that a new representation for each view is obtained using the DGCCA.

The configuration of the DGCCA block that is tested in this work is shown in Fig. 1.

The deep architecture we used [2] contains two hidden layer network, for each view, with the first hidden layer having 2038 units and the second layer having 1608 units followed by a one dimen-

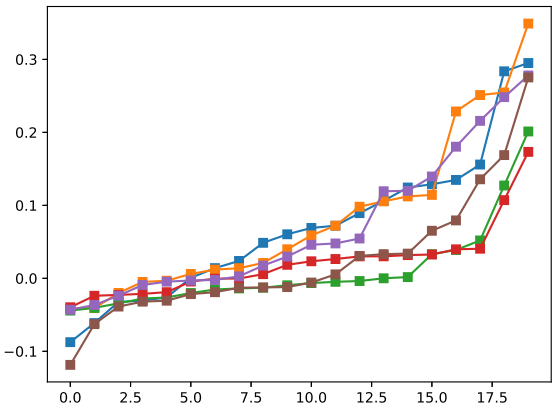

Fig. 2. Results of the DGCCA for different subjects.

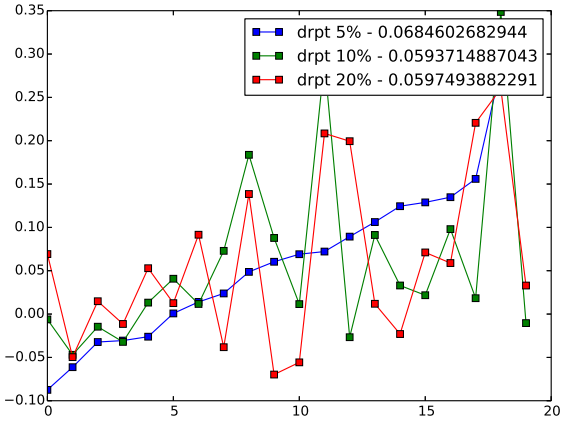

Fig. 3. Results of the DGCCA performed for one subject for different dropouts value of 5%, 10%, 20%.

sional output layer. We use the leaky ReLU activation function at the output [8]. We also incorporate dropout regularization [9, 10] in the deep CCA model training to avoid over-fitting in the noisy conditions. With varying amounts of dropout regularization.

Each subject had 20 sessions with approximately 160 seconds of audio recording in each session. All the results are obtained for 20 fold validation experiments in which one of the sessions is held out as the test data while the 19 other sessions are used in training the model (both the linear models as well as the DCCA models). For the models training, this set of training instances were further split randomly into training and validation with a $90 - 10$ split for each subject and the common stimulus.

The results for each subject for each of the 20 cross validation folds when performed DGCCA is compared to the correlation values obtained for the final representations obtained for each subject separately without the influence of the other subjects' data. The plot which shows the corresponding comparision is shown in the 2

3. SUMMARY

In this project, we tried to marry the DGCCA and multiway CCA models for decoding the auditory EEG activity which were introduced in two different literature works separately. It is not evident that the marriage is successful but we feel that there is still scope for a lot of hyperparameters tuning, to make it work out. The DGCCA model performs a non-linear mapping of the responses of all subjects and the common stimulus where the correlation is maximized. We still need to experiment several configurations of the deep CCA model. In summary, this work shows that there is still a lot scope for experimenting the replacement of the linear MCCA module with a deep version of it to constitute as a useful method for analyzing complex relationships between stimulus and EEG recordings.

4. REFERENCES

[1] Harold Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*, pp. 162–190. Springer, 1992.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, 2013, pp. 1247–1255.

[3] Alain de Cheveigne, Giovanni M Di Liberto, Dorothee Arzounian, Daniel DE Wong, Jens Hjortkjaer, Søren Fuglsang, and Lucas C Parra, "Multiway canonical correlation analysis of brain data," *NeuroImage*, vol. 186, pp. 728–740, 2019.

[4] Adrian Benton, Huda Khayrallah, Biman Gujral, Drew Reisinger, Sheng Zhang, and Raman Arora, "Deep generalized canonical correlation analysis," *CoRR*, vol. abs/1702.02519, 2017.

[5] Giovanni M Di Liberto, James A O'Sullivan, and Edmund C Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

[6] Alain de Cheveigné and Dorothée Arzounian, "Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data," *NeuroImage*, vol. 172, pp. 903–912, 2018.

[7] Alain de Cheveigné, Daniel DE Wong, Giovanni M Di Liberto, Jens Hjortkjaer, Malcolm Slaney, and Edmund Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[8] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[9] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[10] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.