MULTIVIEW LEARNING VIA DEEP DISCRIMINATIVE CANONICAL CORRELATION ANALYSIS

Nour El Din Elmadany, Yifeng He, Ling Guan

Department of Electrical & Computer Engineering, Ryerson University, Toronto, Ontario, Canada

ABSTRACT

In this paper, we propose Deep Discriminative Canonical Correlation Analysis (DDCCA), a method to learn the nonlinear transformation of two data sets such that the within-class correlation is maximized and the inter-class correlation is minimized. Parameters of the two deep transformations are jointly learned. Unlike CCA and Discriminative CCA, the proposed DDCCA does not need inner product. The proposed DDCCA was evaluated in two applications, handwritten digit recognition and speech-based emotion recognition. The experimental results demonstrated that the proposed DDCCA can get a higher recognition accuracy compared to the existing Deep CCA method.

Index Terms— DDCCA, Deep CCA, Discriminative CCA, CCA.

1. INTRODUCTION

In many applications, data arrives in multiple sets or styles or views or modalities. Usually multi-view feature learning refers to learning from different views or sets of data that present different characteristics of data. The views can take different shapes. It may be different multimodalities like a simultaneously recorded audio and video [1], or different information or features extracted from the same source [2]. The presence of multiple sources of information gives us the chance for better representation by incorporating and analyzing different sources of data. Recently, multi-view learning gained the interest of researchers and became a vital research direction.

Canonical Correlation Analysis (CCA) targets at maximizing correlation between the pairwise samples. The idea of using CCA in feature learning was explored by Sun *et al.* [3]. The CCA demonstrated its effectiveness in many applications like medical imagining [4], and audio visual synchronization [5]. However, simple linear correlation cannot describe the dependencies between the input sets. In other words, CCA may not correctly correlate the samples in case of the non-linear correlation between samples. To tackle this problem, Fyfe *et al.* [6] proposed Kernel Canonical Correlation Analysis (KCCA). KCCA has been applied in many applications like the fusion of ear and profile face for multimodal biometric recognition [7], and learning acoustic features for speech recognition [8]. However, CCA can neither perfectly reveal the similarity among samples nor dissimilarity among samples from different classes. So, Sun *et al.* introduced Discriminative CCA to solve this problem [9]. Not only Discriminative CCA is able to maximize the within-class similarity and minimize the between-class similarity. So, it is a more suitable feature representation. Recently, the success of deep learning grabbed the attention of the researchers. Andrew *et al.* [10] proposed to combine deep networks and CCA to tackle the problems of KCCA. KCCA faces mainly two problems. First, the presentation is limited by a certain fixed kernel. Second, KCCA needs a huge amount of time to compute the kernel for new data. Andrew *et al.* proposed to combine deep networks and CCA [10]. In Deep CCA, two deep mapping networks learn jointly a representation that maximizes the correlation between the data sets [10]. Deep CCA has been applied in different applications. Wang *et al.* [11] applied Deep CCA in speech recognition. Lu *et al.* [12] used Deep CCA in word embedding. Yan *et al.* [13] addressed the problem of matching images and texts using Deep CCA.

In this paper, we propose Deep Discriminative CCA (DDCCA) which simultaneously learns two deep mapping networks of the two sets to maximize the within-class correlation and minimize the inter-class correlation. In principle, the proposed DDCCA can be applied to any task that has been used by CCA or Deep CCA. We evaluated the proposed DDCCA in two applications, which are handwritten digits recognition and emotion recognition from speech.

The rest of the paper is organized as follows. We give a brief background on CCA, Discriminative CCA, and Deep CCA in Section 2. The proposed DDCCA is presented in Section 3. Experimental results are provided in Section 4. Finally, the conclusion is drawn in Section 5.

2. BACKGROUD ON CCA, DISCRIMINATIVE CCA, AND DEEP CCA

In this section, a review on CCA, Discriminative CCA, and Deep CCA is presented.

2.1. CCA

In multi-view learning, different data features are extracted from the same underlying signal. Let $X \in R^{m \times N}$, $Y \in R^{d \times N}$ denote two random matrices with zero mean representing the available features with N samples for

training, and m and d represent the dimension…s of X and Y, respectively. The objective of CCA is to find pairs of projection vectors that maximize the correlation between the projections of the two views. It can be formulated as follows:

$$\max_{\{W_x, W_y\}} \frac{{W_x}^T C_{xy} W_y}{\sqrt{{W_x}^T C_{xx} W_x}\sqrt{{W_y}^T C_{yy} W_y}} \quad (1)$$

where C_{xy} is the cross-covariance matrix of X and Y, C_{xx} and C_{yy} are the covariance of X and Y, respectively. The above objective can be rewritten in different ways, one of which is

$$\max_{\{W_x, W_y\}} Tr({W_x}^T C_{xy} W_y) \quad (2)$$
$$s.t.\ {W_x}^T C_{xx} W_x = {W_y}^T C_{yy} W_y = \mathrm{I}$$

The optimal solution to the optimization problem (2) can be determined by computing the sum of the top singular values of the matrix $T = {C_{xx}}^{-1} C_{xy} {C_{yy}}^{-1}$. Assume U_x and U_y are the first K left and right singular vectors of T. The optimal projection matrices for the optimization problem (2) are given by

$$W_x = {C_{xx}}^{-1} U_x,\ W_y = {C_{yy}}^{-1} U_y\,, \quad (3)$$

The projections of input features with CCA are given by $P_x = {W_x}^T X$ and $P_y = {W_y}^T Y$.

2.2. Discriminative CCA

Discriminative CCA aims to find pairs of projection matrices W_x and W_y such that within-class correlation is maximized and inter-class correlation is minimized. The Discriminative CCA can be presented as follows:

$$\max_{\{W_x, W_y\}} \frac{{W_x}^T \check{C}_{xy} W_y}{\sqrt{{W_x}^T \check{C}_{xx} W_x}\sqrt{{W_y}^T \check{C}_{yy} W_y}} \quad (4)$$

where $\check{C}_{xy} = C_w - \xi C_b$, C_w and C_b denote the within-class correlation and between-class correlation, respectively, and ξ is a tuning parameter that represents the relative contribution of the within-class correlation and between-class correlation. $\check{C}_{xy}$ is the difference between the within-class correlation and between class correlation. Moreover, the within-class correlation C_w can be calculated as follows:

$$C_w = XAY^T \quad (5)$$

$$A = \begin{bmatrix} \mathbf{1}_{n_1 \times n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{1}_{n_c \times n_c} \end{bmatrix} \quad (6)$$

where n_i is the number of samples in the i^{th} class.
The between-class correlation C_b is given by

$$C_b = -XAY^T \quad (7)$$

Thus, $\check{C}_{xy}$ can be rewritten as $(1 + \xi)C_w$. This means that maximizing C_w minimizes C_b automatically. So, Discriminative CCA is rewritten as follows:

$$\max_{\{W_x, W_y\}} ({W_x}^T (1 + \xi) C_w W_y) \quad (8)$$
$$s.t.\ {W_x}^T C_{xx} W_x = {W_y}^T C_{yy} W_y = \mathrm{I}$$

The optimization problem (9) can be solved by finding the eigenvector associated with the largest eigenvalue, which can be solved as the generalized eigenvalue problem. In [9], Sun *et al.* proved that the highest recognition accuracy can be obtained at the dimension $d \leq C$, where C is the number of classes.

2.2. Deep CCA

Deep CCA is a method to learn nonlinear transformation of two views data such that the resulting representations are highly correlated. One of the advantages of Deep CCA is that it does not require inner product performed in CCA. Deep CCA tries to find representations of the two data sets. Both sets pass through multi-layer of non-linear transformation. The output of the first hidden layer h_1^x can be computed as follows:

$$h_1^x = s(W_1^x X + b_1^x) \quad (9)$$

where X represents the first view, W_1^x is a matrix of weights, b_1^x represents a vector of biases, and s is a nonlinear function. h_1^x can be applied as an input to the next hidden layer. The final output $f_x(X)$ can be computed as follows:

$$f_x(X) = s(W_d^x h_{d-1}^x + b_d^x) \quad (10)$$

where d is the number of hidden layers. Similarly, the output of the second set $f_y(Y)$ is computed as follows:

$$f_y(Y) = s(W_d^y h_{d-1}^y + b_d^y) \quad (11)$$

The objective of Deep CCA is to jointly learn weights and biases of the two sets such that the correlation between $f_x(X)$ and $f_y(y)$ is maximized. The objective can be formulated as follows:

$$\max_{\{\theta_x, \theta_y\}} corr(f_x(X, \theta_x), f_y(Y, \theta_y)) \quad (12)$$

where θ_x is a vector of the parameters W_l^x and b_l^x for $l = 1,2,..d$, and similarly for θ_y.

The total correlation can be written as follows:

$$corr\left(f_x(X), f_y(Y)\right) = Tr(T^T T)^{1/2} \quad (13)$$

where T is computed for the outputs from the two multilayer networks. The weights and biases are trained using back

propagations from the trace norm layer. The gradient of $corr\left(f_x(X), f_y(Y)\right)$ can be computed as follows:

$$\frac{\partial corr\left(f_x(X), f_y(Y)\right)}{\partial f_x(X)} = \frac{1}{N}\left(2\nabla_{xx} f_x(X) + \nabla_{xy} f_y(Y)\right) \quad (14)$$

where $\nabla_{xy} = {C_{xx}}^{-1} U V^T {C_{yy}}^{-1}$ and $\nabla_{xx} = -\frac{1}{2} {C_{xx}}^{-1} U D U^T {C_{yy}}^{-1}$, and $T = UDV^T$. In the same way, $\frac{\partial corr\left(f_x(X), f_y(Y)\right)}{\partial f_y(y)}$ can be computed.

Deep denoising autoencoder is adopted to initialize the values of weights in the two deep networks [10] [14].

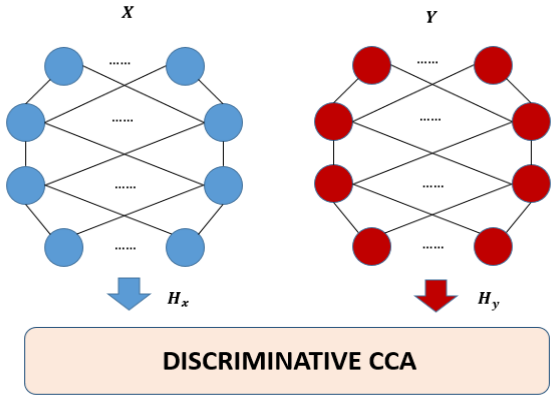

Figure 1. A schematic diagram of the proposed DDCCA.

3. THE PROPOSED DDCCA

The proposed DDCCA aims to find complex nonlinear representations such that within-class correlation is maximized and between-class correlation is minimized. Similarly to Deep CCA, the proposed DDCCA does not need any inner product. Figure 1 shows a schematic diagram of DDCCA. Let H^x and H^y are the top-level representation of the two matrices. $\bar{H}^x$ and $\bar{H}^y$ are the zero mean data matrix. Similar to Deep CCA, $\frac{\partial corr(H^x, H^y)}{\partial H^x}$ should be computed as follows:

$$\frac{\partial corr(H^x, H^y)}{\partial H^x_{ij}} = \sum_{ab} \nabla^{xx}_{ab} \frac{\partial C^{xx}_{ab}}{\partial H^x_{ij}} + \sum_{ab} \nabla^{xy}_{ab} \frac{\partial C^{xy}_{ab}}{\partial H^x_{ij}} \quad (15)$$

where H^x_{ij} represents the top level representation of the set X, and i, j are the subscripts that index into matrices. ∇^{xx}_{ab} and ∇^{xy}_{ab} can computed as in (14), and $\frac{\partial C^{xx}_{ab}}{\partial H^x_{ij}}$ can be calculated as follows:

$$\frac{\partial C^{xx}_{ab}}{\partial H^x_{ij}} = \frac{1}{N-1}\left(\mathbf{1}_{\{a=i\}} H^x_{ij} + \mathbf{1}_{\{b=i\}} H^x_{ij}\right) \quad (16)$$

where N is the number of samples. However, $\frac{\partial C^{xy}_{ab}}{\partial H^x_{ij}}$ can be computed as follows:

$$\frac{\partial C^{xy}_{ab}}{\partial H^x_{ij}} = \frac{1}{N-1}\left(\mathbf{1}_{\{a=i\}}(A + A^T) H^x_{ij}\right) \quad (17)$$

where A can be calculated similarly as in (6). However, A is considered as a symmetric positive semi-definite matrix. Therefore, Equation (17) can be rewritten as follows:

$$\frac{\partial C^{xy}_{ab}}{\partial H^x_{ij}} = \frac{2}{N-1}\left(\mathbf{1}_{\{a=i\}} A H^x_{ij}\right) \quad (18)$$

Finally, Equation (15) can be rewritten as follows:

$$\frac{\partial corr(H^x, H^y)}{\partial H^x} = \frac{1}{N-1}\left(\nabla^{xx} \bar{H}^x + \nabla^{xy} \bar{H}^y\right) \quad (19)$$

Calculating $\frac{\partial corr(H^x, H^y)}{\partial H^y}$ is similar to $\frac{\partial corr(H^x, H^y)}{\partial H^x}$.

To train the deep neural network, the procedures of Deep CCA are followed [10]. The initialization of the parameters for each layer is obtained by a denoising autoencoder [14].Then, Limited memory-Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) [15] is adopted to find a local minimum of reconstruction error plus a quadratic penalty [10] [16]. The obtained weights and biases are used as the initial values to maximize the objective of DDCCA. The s adopted rectified linear unit (ReLU) function as our proposed as a nonlinear function in our proposed technique DDCCA [17].

4. EXPERIMENTAL RESULTS

In this section, the experimental evaluation for the proposed DDCCA is presented. The evaluation of the DDCCA is conducted on two publicly available datasets for two different applications. The DDCCA is applied first to handwriting recognition using MNIST dataset [18] and emotion recognition from speech using RML dataset [19].

4.1. Handwritten Digit Recognition

MNIST handwritten image dataset [18] consists of 60,000 training images and 10,000 testing images. The dataset contains numbers from 0 to 9. Each image is 28×28 pixels. The image is divided into two halves: right half and left half, each containing 14 columns of handwritten digits. We used 50,000 images for training, 10,000 images for tuning, and 10,000 images for testing. The number of hidden layers was chosen to be two hidden layers with 800 nodes in each layer. For the output layer, the number of nodes is 50. Linear SVM [20] was adopted as a classifier.

Table 1. Results of handwritten digit recognition

Method	Average recognition accuracy
DCCA [10]	96.87%
The proposed DDCCA	97.24%

Table 1 shows the comparison of the recognition results between the proposed DDCCA and the existing DCCA. As shown in Table 1, the proposed DDCCA has a higher average recognition accuracy than DCCA due to the better discriminancy in DDCCA.

4.1. Speech based Emotion Recognition

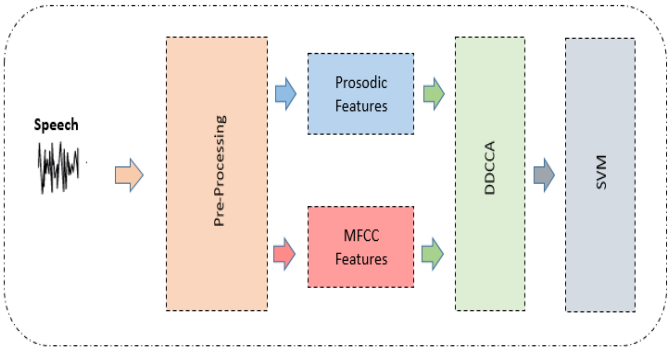

Figure 2 Emotion recognition framework from speech based on DDCCA.

To evaluate the performance of the proposed method in emotion recognition, experiments were conducted on RML dataset [19]. The dataset contains six principal emotions (angry, disgust, fear, surprise sadness and happiness) which are performed by eight different subjects speaking different languages (English, Mandarin, Urdu, Punjabi, Persian, and Italian. A total of 288 samples from dataset are selected. Among the samples, 144 samples were used for training, 48 samples for tuning, and 96 were used for testing.

It is a challenge to choose features that can represent the characteristics of the emotion. Prosodic features were adopted in emotion recognition systems [19], as it showed good performance in emotion recognition. MFCC was adopted in speech emotion recognition [19]. Then, the prosodic and MFCC features are fused together using the proposed DDCCA. The details of prosodic and MFCC features can be found in [19]. Linear SVM [20] was adopted as a classifier in our framework. The framework of emotion recognition is shown in figure 2

First, the performance of emotion recognition using a single feature (e.g., prosodic feature or MFCC feature) is provided in Table 2. The results show that prosodic feature is better than MFCC feature in emotion recognition.

Table 2. Results for emotion recognition with single feature.

Feature	Average recognition accuracy [%]
Prosodic	56.25
MFCC	52.08

Table 3. Comparison between DDCCA and DCCA.

Method	Average recognition accuracy [%]
The proposed DDCCA	65.63
DCCA [10]	60.42

Then, the proposed DDCCA is compared with DCCA in Figure 3. In Figure 3, the size of output layer is varied from 3 to 25 nodes. The number of hidden layers was chosen to be two hidden layers with 200 nodes in each layer. The results show that the proposed DDCCA has a better recognition accuracy in average than DCCA. It can be observed that the proposed DDCCA has an average recognition accuracy greater than 65%, which is 5% higher than DCCA. The maximum recognition accuracy is found to be at six output nodes, where 6 is the number of classes. This matches the property of Discriminative CCA that the maximum accuracy should be at a dimension $d \leq C$. Table 3 shows a comparison between DDCCA and DCCA.

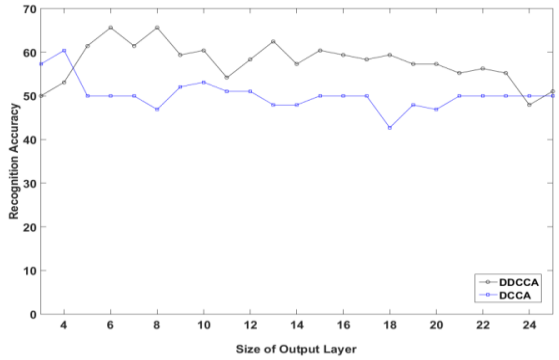

Figure 3. The recognition accuracy with different number of output nodes.

5. CONLUSION

We have proposed an effective multi-view learning method, called DDCCA, which jointly maximizes the within-class correlation and minimizes the inter-class correlation. The proposed DDCCA has a better discriminative capability, leading to an improvement of recognition accuracy. The proposed method was evaluated in handwritten digit recognition and speech-based emotion recognition. The experimental results demonstrated an improved recognition accuracy over the existing deep CCA.

6. REFERENCES

[1] E. Kidron, Y. Schechner and M. Elad, "Pixels that sound*," IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 88-96, 2005.

[2] J. Pennington, R. Soccher, and C. Manning, "Glove: Global vectors for word representation," *Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543, 2014.

[3] Q. Sun, S. Zeng, Y. Liu, P. Heng, and D. Xia, " A new method of feature fusion and its application in image recognition," Pattern Recognition, vol.36, no. 1, pp.2437-2448,2005.

[4] N. Correa, T. Adali, Y. Li, and V. Calhoun, "Canonical correlation analysis for data fusion and group ingerences: examining applications of medical imaging data," *IEEE Signal Processing Magazine*, vol.27, no.4, pp.39-50, 2010.

[5] M. Sarin, Y. Yemez, E. Erzin, and A. Teklap, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transaction of Multimedia*, vol. 9, no.7, pp. 1396-1403, 2007.

[6] C. Fyfe and P. Lai, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol.10, pp.365-374, 2001.

[7] X. Xu and Z. Mu, "Feature fusion method based on KCCA for ear and profile face based multimodal recognition," *IEEE International Conference on Automation and Logistics*, pp.620-623, 2007.

[8] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," *Symposium on Machine Learning in Speech and Language (MLSLP)*, 2012.

[9] T. Sun, S. Chen, Z. Jin, and J. Yang. "Kernalized discriminative canonical correlation analysis," *International Conference on Wavelet Analysis and Pattern Recognition*, pp.1283-1287, 2007.

[10] G. Andrew, R. Arora, J. Bilmes and K. Livescu, "Deep canonical correlation analysis," *International Conference on Machine Learning (ICML)*, 2013.

[11] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2015.

[12] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu, "Deep Multilingual Correlation for Improved Word Embeddings," *Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT),* 2015.

[13] F. Yan and K. Milkolajczyk, "Deep correlation for matching images and text," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3441-3451, 2015.

[14] P.Vincent, H. Larochelle, I. lajoie, Y. Bengio, P. Manzagol, "Stacked denoising autoencoder:learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, pp. 3371-3408, 2010.

[15] J. Nocedal and S. Wright, "Numerical Optimization," *Springer*, 2nd editions, 2006.

[16] Q. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, A. Ng, " On Optimization Methods for Deep Learning," *International Conference on Machine Learning (ICML)*, 2010.

[17] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," *International Conference on Machine Learning (ICML)*, 2010.

[18] Y. LeCun, L. Buttou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of IEEE*, pp.755-824, 1998.

[19] Y. Wang and L. Guan, "Recognizing human emotional state from audio visual signals," IEEE Transactions on Multimedia, vol.10, no.5, pp. 936-946, 2008.

[20] C. Chang and C. Lin, " LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol.2, no.3, 2011.