DISCRIMINATIVE FEATURE SELECTION GUIDED DEEP CANONICAL CORRELATION ANALYSIS

Nour El Din El Madany, Yifeng He, and Ling Guan

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

ABSTRACT

This paper proposes a novel Discriminative Feature Selection Guided Deep Canonical Correlation Analysis (D^2CCA) for multiview learning. The proposed (D^2CCA) enhances the discriminative power of the learned featured representation by imposing the selection of the most discriminative features. Moreover, it learns to maximize the correlations between two views. Also, an alternating iterative learning algorithm is presented to find the sub-optimal solution. The experimental results demonstrated that the proposed (D^2CCA) can achieve a higher average accuracy compared to several existing methods.

***Index Terms*—** Discriminative Feature Selection Guided Deep Canonical Correlation Analysis, Multiview Learning, Deep CCA

1. INTRODUCTION

In many computer vision applications, an object is observed from different complementary views or modalities or sets. Moreover, the presence of multiple views creates an opportunity of learning better representations by effectively utilizing the information obtained from different views.

Researchers invested substantial efforts to investigate several techniques for multiview learning. Hardon *et al.* [1] presented Canonical Correlation Analysis (CCA) which aims to maximize the correlation between two different views. Kernel CCA ($KCCA$) [1] was presented to reveal the potential nonlinear relations between the two views which can not be revealed by (CCA). As (CCA) and ($KCCA$) are limited to revealing relations between two views only, ($MCCA$) [2] was introduced to interpret the relations among more than two views. Rasiwasia *et al.* [3] proposed Cluster Canonical Correlation Analysis ($Cluster - CCA$) to consider the correlations between samples from the same classes only. Kan *et al.* [4] extended Linear Discriminant Analysis LDA to Multiview Discriminant Analysis ($MvDA$). It maximizes the trace ratio between the within-class and between-class matrices for all views or sets.

Due to the revolutionary success of deep learning, researchers proposed to learn unified representation from different views using deep neural networks. Ngiam *et al.* proposed to capture the middle level relationship between two views using Deep Bimodal Autoencoder [5].

Conversely, such models does not discover the correlations across the views explicitly which limits the performance of multiview learning. Andrew *et al.* introduced the Deep Canonical Correlation Analysis ($Deep - CCA$) which learns jointly complex nonlinear transformations of two views such that the resulting representations are highly correlated [6]. Wang *et al.* proposed Deep Canonically Correlated Autoencoders ($DCCAE$) which is basicly adding a reconstruction regularization term to the ($Deep - CCA$) objective function [7]. The extensive experiments showed the importance of the reconstruction regularization term. In [8], Chang *et al.* presented stochastic decorrelation loss to ($Deep - CCA$) objective function.

The architecture of ($Deep - CCA$) encourages the network to seek an effective representation of data. However, ($Deep - CCA$) dismisses supervision incorporation leading to limited performance. According to [9], the learned representation is composed of task-relevant and task-irrelevant units. Only task-relevant units, which are related to the objective, carries the discriminative features representation. To address this issue, a unified framework is proposed to integrate discriminative feature selection and ($Deep - CCA$). Intuitively, supervised feature selection is applied on the learned feature representation of ($Deep - CCA$) to select the most discriminative features. The selected units are used to optimize the deep neural network to improve the discriminability on the selected units and maximize the correlation between the two views.

The rest of this paper is organized as follows. Section 2 presents background of (CCA). The proposed Discriminative Feature Selection Guided ($Deep - CCA$) is described in Section 3. The experimental results are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. BACKGROUND ON CANONICAL CORRELATION ANALYSIS

Canonical correlation analysis (CCA) aims at maximizing the correlation between two different views [1]. Consider two views of N zero mean variables $X \in \Re^{p \times N}$and $Y \in \Re^{q \times N}$, p and q are the dimensions of feature samples in X and Y, re-

spectively. (CCA) aims at learning pairs of linear projection of the two views that are maximally correlated as follows:

$$\underset{w_x, w_y}{\text{maximize}} \quad \frac{w_x^T \Sigma_{xy} w_y}{\sqrt{w_x^T \Sigma_{xx} w_x}\sqrt{w_y^T \Sigma_{yy} w_y}} \tag{1}$$

where Σ_{xy} is the cross-covariance between X and Y. Σ_{xx} and Σ_{yy} are the covariances of X and Y, respectively. When finding the pairs (w_x^i, w_y^i), subsequent constraints are constrained to be uncorrelated $(w_x^i \Sigma_{xx} w_x^j = w_y^i \Sigma_{yy} w_y^j = 0)$ and $i \neq j$. The problem is reformulated as follows:

$$\begin{aligned} \underset{W_x, W_y}{\text{maximize}} \quad & Trace(W_x^T \Sigma_{xy} W_y) \\ \text{subject to} \quad & W_x^T \Sigma_{xx} W_x = I; W_y^T \Sigma_{yy} W_y = I \end{aligned} \tag{2}$$

The optimization problem Eq. (2) can be solved through different ways. The authors in [6] defined $T = \Sigma_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}}$ and used singular value decomposition UDV, where U and V are the left and right singular vectors and D is the a diagonal matrix of singular values. The optimum projections are computed by finding $W_x = \Sigma_{xx}^{-\frac{1}{2}} U$ and $W_y = \Sigma_{yy}^{-\frac{1}{2}} V$.

3. THE PROPOSED FRAMEWORK

3.1. Discriminative Feature Selection Guided Deep CCA

In this section, we propose the learning framework of Discriminative Feature Selection Guided Deep CCA (D^2CCA) as in Fig. 1. We integrated the feature selection in the final layer units and reformulated the problem as follows:

$$\begin{aligned} \underset{\theta_x, \theta_y, P_x, P_y}{\text{maximize}} \quad & corr(f_x(X;\theta_x), f_y(Y;\theta_y)) \\ & +\lambda_1 C(P_x; F_x) + \lambda_2 C(P_y; F_y) \end{aligned} \tag{3}$$

where $F_x = f_x(X;\theta_x)$ and $F_y = f_y(Y;\theta_y)$ and represent the output of the two deep neural network of each view, θ_x and θ_y are the network parameters of X and Y, respectively. λ_1 and λ_2 are balancing parameters. $corr(f_x(X;\theta_x), f_y(Y;\theta_y)) = Tr(TT^T)$ and $T = \Sigma_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}}$. $C(P_x; f_x)$ and $C(P_y; f_y)$ are the feature selecting regularized term, with a learned feature selection matrix P_x and P_y performed on the output feature representation F_x and F_y, respectively. Specifically, $i-th$ row vector in P_x denoted by p_x^i is all zero vector except the $i-th$ element corresponding to the $i-th$ feature dimension as follows:

$$p_x^i = [0, ..., 0, 1, ..., 0, 0] \tag{4}$$

where P_x and $P_y \in \Re^{m \times d}$, d is the size of output feature representation and m is the size of the selected feature representation.

Feature selection can be mainly split into three categories: unsupervised, supervised, and semi-supervised. In order to improve the discriminative information, supervised feature selection is adopted. Here, $C(P_x, F_x)$ is chosen as follows:

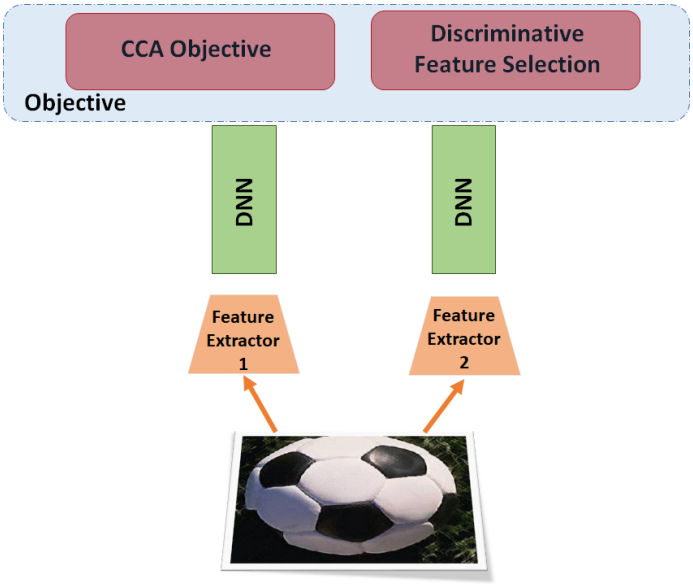

Fig. 1. The proposed D^2CCA.

$$C(P_x, F_x) = \frac{Tr(P_x F_x L_b^x F_x^T P_x^T)}{Tr(P_x F_x L_w^x F_x^T P_x^T)} \tag{5}$$

where L_b^x is the between class laplacian matrix, and L_w^x is the within class laplacian matrix for view X. The between-class laplacian matrix $L_b^x = D_b^x - S_b^x$, where D_b^x is the diagonal matrix and its entries are column sum of S_b^x for view X. The within-class laplacian matrix $L_w^x = D_w^x - S_w^x$, where D_w^x is the diagonal matrix and its entries are column sum of S_w^x for view X. S_{wx}^{ij} is the $(i,j) - th$ element in matrix S_w^{ij} and is computed as follows:

$$S_{wx}^{ij} = \begin{cases} exp(-\|x_i - x_j\|^2 / t_S) & i,j \in c, c \in C, i \neq j \\ 0 & otherwise \end{cases} \tag{6}$$

where C is the set of classes and t_S is chosen to be 1. S_{bx}^{ij} is the $(i,j) - th$ element in matrix S_b^{ij} and is computed as follows:

$$S_{bx}^{ij} = exp(-\|\bar{x}_i - \bar{x}_j\|^2 / t_S) \tag{7}$$

where $\bar{x}_i$ is the mean of data samples belonging to the same class and view Y has a similar expression. Generally, the trace ratio does not have a closed form solution, therefore the problem is changed to trace difference problem to achieve a globally optimum solution [10]. Following [10], Eq. (3) is reformulated as follows:

$$\begin{aligned} \underset{\theta_x, \theta_y, P_x, P_y, \gamma_x, \gamma_y}{\text{maximize}} \quad & corr(f_x(X;\theta_x), f_y(Y;\theta_y)) \\ & +\lambda_1 (P_x F_x (L_b^x - \gamma_x L_w^x) F_x^T P_x^T) \\ & +\lambda_2 (P_y F_y (L_b^y - \gamma_y L_w^y) F_y^T P_y^T) \end{aligned} \tag{8}$$

where γ_x and γ_y are the trace ratio score for view X and Y, respectively.

3.2. Optimization

Solving the final objective function is hard due to the non linearity introduced by the deep neural network. Therefore, alternating optimization approach is employed to iteratively learn the network parameters (θ_x, θ_y), the feature selection matrices (P_x, P_y), and the trace ratio score (γ_x, γ_y). In other words, we optimize two sub problems, the feature selection problem and canonical correlation maximization problem.

3.2.1. Feature Selection Problem

First, the parameters of the two deep neural networks are fixed and the parameters of feature selection are learned. Specifically, for view X, we compute both the feature selection matrix P_x and the trace ratio score γ_x which are optimized in an alternating manner. Suppose P_x is computed from the previous iteration, γ_x is computed as follows:

$$\gamma_x = \frac{Tr(P_x F_x L_b^x F_x^T P_x^T)}{Tr(P_x F_x L_w^x F_x^T P_x^T)} \tag{9}$$

where L_w and L_b are computed as in Eq. (6) and Eq. (7). Using the obtained γ_x, each feature in view X is ranked according to its discriminative power score $r(\gamma_x)$ according to:

$$r(\gamma_x) = Tr(p_x F_x (L_b^x - \gamma_x L_w^x) F_x^T p_x^T) \tag{10}$$

where p_x is a one hot vector corresponding to each feature dimension as in Eq. (4). This procedure will continue in iterative manner till convergence. Similarly for the second view Y, we follow the same steps. Algorithm 1 summarizes the optimization procedure for both X and Y.

Algorithm 1 Feature Selection Problem

Require: Two learned feature representation F_x and F_y, number of selected features m, laplacian matrices L_b^x, L_w^x, L_b^y, and L_w^y

Ensure: P_x, P_y, γ_x, and γ_y

1: Initialize P_x=P_y= $I \in \Re^d$
2: **Repeat**
3: Obtain γ_x using Eq. (9) and γ_y similarly.
4: Compute the score of each j feature with Eq. (10) for each view.
5: Rank the features according to their discriminative power score descendingly for each view.
6: Update the feature selection matrix P_x and P_y.
7: **Until** Convergence

3.2.2. Discriminative Feature Selection Guided Deep CCA Learning

When P_x, P_y, γ_x, and γ_y are fixed, the parameters of the two deep neural networks are optimized using Stochastic Gradient descent or any of its variants. The gradient of the objective function in Eq. (8) with respect to the learned feature representation F_x is computed as follows:

$$\begin{aligned} \frac{\partial L}{\partial F_x} &= \frac{1}{N}(2\nabla_{xx}\bar{F}_x + \nabla_{xy}\bar{F}_y) \\ &+ \lambda_1 P_x^T P_x F_x^T (L_b^x - \gamma_x L_w^x) \end{aligned} \tag{11}$$

where

$$\nabla_{xy} = \Sigma_{xx}^{-\frac{1}{2}} U V^T \Sigma_{yy}^{-\frac{1}{2}} \tag{12}$$

and

$$\nabla_{xx} = -\frac{1}{2}\Sigma_{xx}^{-\frac{1}{2}} U D U^T \Sigma_{xx}^{-\frac{1}{2}} \tag{13}$$

N is the number of data samples and $\frac{\partial L}{F_y}$ has a symmetric expression. The details of optimization is summarized in Algorithm 2.

Algorithm 2 Discriminative Feature Selection Guided Deep CCA

Require: Two views X and Y, balancing parameters λ_1 and λ_2, number of selected features m

Ensure: θ_x, and θ_y

1: Initialize θ_x and θ_y
2: **Repeat**
3: Fix θ_x and θ_y, and update P_x, P_y, γ_x, and γ_y using Algorithm 1
4: Fix P_x, P_y, γ_x, and γ_y and update θ_x and θ_y.
for each view using Eq. (12-14).
5: **Until** Convergence

4. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed D^2CCA for object recognition and handwritten digits recognition using two datasets, Caltech101 object database [11] and MNIST handwritten digit database [12].

4.1. Experiments on Caltech101 Dataset

Caltech101 is a widely used database for object recognition containing a 9144 images from 102 classes of 101 object classes (animals, vehicles, tree, etc) and one background class. Following the experimantal setting [11], we chose 10 random objects per class for training and the rest for testing.

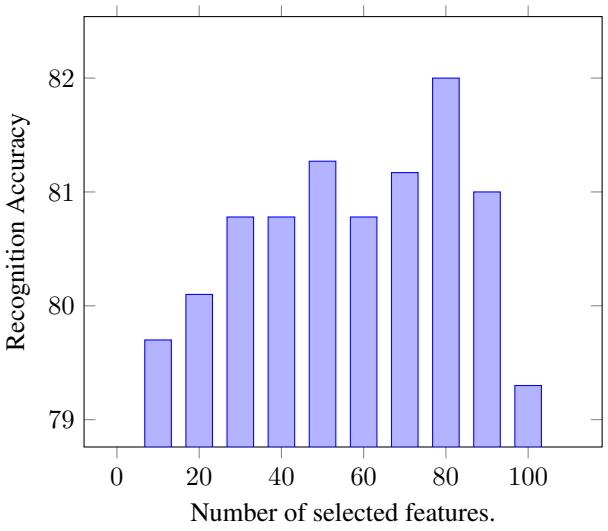

Fig. 2. Performance analysis of the influence of the number of discriminative selected features m.

We extracted two types of features; hand crafted (Dense SIFT + Bag of visual words $BOVW$) and Learning based features $VGG-f$ [13] each representing one of the two views. Each of the two features was applied to PCA and the top 100 dimensions were selected as preprocessing. For classification, Linear SVM is chosen [14]. Table I shows the recognition accuracy of each feature. From figure, we can notice the superiority of $VGG-f$ over the hand crafted feature.

Table 1. Recognition Accuracy for each View on Caltech101 Database.

View	Recognition Accuracy %
Dense SIFT+ BOVW	48
VGG-f feature	69

We choose the same layer size for each network [100, 1200, 1200, 1200, 100] with ReLU as a non linear activation function. Our code was implemented using Pytorch. First, we conducted several experiments to study the effect of the number of selected feature on the recognition accuracy. Fig.2 shows that choosing 80 features out of 100 has the highest recogntion accuracy. In other words, this phenomena proves that some of output units are irrelevant to the recognition task as the may not represent the object itself. Moreover, we compared D^2CCS with CCA [1], $ClusterCCA$ [3], $DeepCCA$ [15], $DCCA$ [16] and $MvDA$ [4]. The results are summarized in the table 2 which clearly shows that D^2CCA outperforms other existing methods.

Table 2. Recognition accuracy comparison between the proposed methods D^2CCA and the other multi-view techniques on Caltech101 Dataset.

Multi-view Technique	Recognition Accuracy %
MvDA	80
CCA	70.69
DCCA	71.86
Cluster CCA	70.69
Deep CCA	79.3
D^2CCA	**82**

4.2. Experiments on MNIST database

MNIST handwritten digits dataset [12] consists of 60000 training images and 10000 for testing image. It contains 10 classes for numbers between 0 to 9. Each image is 28 $\times$ 28 . Each image is divided into two halves containing 14 columns to produce two views as in [6]. 50000 images were used in training phase. 10000 images were used in validation and the rest is for testing. We followed the same architecture deployed in [7]. Table 3 shows the comparison between Deep CCA and the proposed D^2CCA. As shown in the table, the proposed D^2CCA has a recognition accuracy higher than Deep CCA.

Table 3. Recognition accuracy comparison between the proposed method D^2CCA and the other multi-view techniques on MNIST Dataset.

Multi-view Technique	Recognition Accuracy %
deep CCA [7]	97.2
D^2CCA	**98.1**

5. CONCLUSIONS

We proposed a novel approach called D^2CCA for multiview learning. D^2CCA has the ability to satisfy maximizing the correlations between the two views and selecting the discriminative features to dismiss the irrelevant feature representation for the recognition task. The proposed D^2CCA is applied to fuse multiview features. The experimental results demonstrated that D^2CCA outperformed the other methods compared in terms of average recognition accuracy.

Acknowledgement

We acknowledge NVIDIA corporation for the donation of GPU used for this research.

6. REFERENCES

[1] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis, an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[2] M.A. Hasan, "On multi-set canonical correlation analysis," in *International Joint Conference on Neural Networks*, 2009, pp. 1128–1133.

[3] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

[4] M. Kan, S. Shan, H. Zhang, and S. Lao gand X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, 2016.

[5] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[6] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[7] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 1083–1092.

[8] Xiaobin Chang, Tao Xiang, and Timothy M Hospedales, "Deep multi-view learning with stochastic decorrelation loss," *arXiv preprint arXiv:1707.09669*, 2017.

[9] Shuyang Wang, Zhengming Ding, and Yun Fu, "Feature selection guided auto-encoder.," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, (AAAI-17)*.

[10] Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan, "Trace ratio criterion for feature selection.," in *AAAI*, 2008, vol. 2, pp. 671–676.

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer vision and Image understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[12] Y LeCun and C Cortes, "The mnist database of handwritten digits," 1998.

[13] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: delving deep into convoltional networks," in *The British Machine Vision Conference (BMVC)*, 2014.

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

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

[16] Ting-Kai Sun, Song-Can Chen, Zhong Jin, and Jing-Yu Yang, "Kernelized discriminative canonical correlation analysis," in *Wavelet Analysis and Pattern Recognition, 2007. ICWAPR'07. International Conference on*. IEEE, 2007, vol. 3, pp. 1283–1287.