Towards Deep and Discriminative Canonical Correlation Analysis

Matthias Dorfer MATTHIAS.DORFER@JKU.AT
Gerhard Widmer GERHARD.WIDMER@JKU.AT
Department of Computational Perception, Johannes Kepler University Linz, Altenberger Str. 69., Linz, 4040 Austria

Abstract

We introduce a discriminative extension of Deep Canonical Correlation Analysis (DCCA) for the purpose of multi-view representation learning. The objective of DCCA is to learn two groups of latent features which are highly correlated when projected into the common CCA-space. Representations learned with DCCA pre-training have proven to be beneficial when used in a subsequent classification tasks. In this work we tackle exactly the problem of multi-view classification by incorporating a discriminative regularizer on the hidden representations already at train time. Inspired by a deep learning interpretation of Linear Discriminant Analysis (DeepLDA) we design a joint optimization target that encourages the network to learn representations which are not only correlated but also highly discriminative. Preliminary results show that the joint optimization of correlation and separation is feasible and helps to enhance the classification power of the learned representations.

1. Introduction

Canonical Correlation Analysis (CCA) (Hotelling, 1936) is a method from multivariate statistics and measures the linear dependency between two groups of variables. Along with the quantification of correlation (canonical coefficients) it yields a linear projection into a subspace where the two groups of features exhibit maximum correlation. Linear Discriminant Analysis (LDA) (Fisher, 1936) on the other hand operates only on a single set of observations but also tries to find a linear projection into a lower dimensional space. When the preconditions of LDA are fulfilled the observations become linearly separable in the resulting space. The central idea of this paper is to exploit both concepts in a discriminative multi-view deep neural network setup to simultaneously learn representations which are both highly correlated and discriminative at the same time. In the following we review related work where such representations have proven to be useful.

Related Work[1]**:** Andrew et al. (2013) introduce in their work a non-linear neural network extension of classic CCA. The multi-view representations learned by Deep Canonical Correlation Analysis (DCCA) have proven to be useful when used in subsequent classification tasks. Wang et al. (2015) pick up their idea and combine DCCA with different types of multi-view auto-encoder architectures with the result of further improving the discriminative power of the learned feature representations. They simultaneously optimize correlation as well as reconstruction errors of both views by an architecture called the Deep Canonical Correlated Autoencoder (DCCAE). Wang & Livescu (2015) recently proposed another improvement which allows to apply Kernel CCA to large-scale problems. In their work on Deep Linear Discriminant Analysis (DeepLDA) Dorfer et al. (2015) propose an end-to-end deep neural network interpretation of LDA. They introduce an LDA inspired optimization target which allows to learn linearly separable latent spaces on top of the hidden representations of their networks.

The aim of the present work is to show that it is possible to take advantage of both correlation and separability. We propose a joint optimization target unifying DCCA and DeepLDA in a single end-to-end deep multi-view network.

2. Methods

We start by introducing a common notation which will be used throughout this paper. Based on this notation we review DCCA as well as DeepLDA and show how the concepts are used to jointly learn latent feature representations fulfilling the requirements of both methods.

Under review for the ICML 2016 Workshop on Multi-View Representation Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

[1] This is by no means a comprehensive overview of the present state of the art. However, as this is a workshop paper we only review work that is directly related to our approach.

2.1. Notation

Let $\mathbf{x}_1, ..., \mathbf{x}_N = \mathbf{X} \in \mathbb{R}^{N \times d_x}$ and $\mathbf{y}_1, ..., \mathbf{y}_N = \mathbf{Y} \in \mathbb{R}^{N \times d_y}$ denote a set of N multi-view observations belonging to C different classes $c \in \{1, ..., C\}$. According to Wang et al. (2015) we define $\mathbf{f}$ and $\mathbf{g}$ to be non-linear feature extractors (mappings) used for pre-processing the input data. In the present case $\mathbf{f}$ and $\mathbf{g}$ are two different deep neural networks producing hidden feature representations $\mathbf{f}(\mathbf{X}) \in \mathbb{R}^{N \times h}$ and $\mathbf{g}(\mathbf{Y}) \in \mathbb{R}^{N \times h}$ for their corresponding input views. The parameters of the two models are referred to as $\Theta_{\mathbf{f}}$ and $\Theta_{\mathbf{g}}$ and we fix the dimensionality h of the hidden representation to be the same for both views. For a briefer notation we will denote $\mathbf{f}(\mathbf{X})$ and $\mathbf{g}(\mathbf{Y})$ by $\mathbf{f}_X$ and $\mathbf{g}_Y$ respectively in the following.

The approach presented in this work is based on CCA and LDA, two methods from classic multivariate statistics, which both rely on the covariance structures of the respective input feature distributions. Equation (1) introduces the covariance matrices for the learned feature representations of both views.

$$\mathbf{\Sigma}_X = \frac{1}{N-1} \bar{\mathbf{f}}_X^T \bar{\mathbf{f}}_X \text{ and } \mathbf{\Sigma}_Y = \frac{1}{N-1} \bar{\mathbf{g}}_Y^T \bar{\mathbf{g}}_Y \tag{1}$$

$\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ are also referred to as *total scatter* in terms of LDA (see Subsection 2.3) (Fisher, 1936). In addition to the individual covariance matrices we define the cross-covariance $\mathbf{\Sigma}_{XY}$ between the feature representations of the two different views:

$$\mathbf{\Sigma}_{XY} = \frac{1}{N-1} \bar{\mathbf{f}}_X^T \bar{\mathbf{g}}_Y \tag{2}$$

For the formulation of LDA we also require the C per-class covariance matrices $\mathbf{\Sigma}_{X_c}$ as well as their average $\mathbf{\Sigma}_{X_w} = (1/C) \sum_c \mathbf{\Sigma}_{X_c}$ over all individual class covariances (*within scatter* in terms of LDA). Finally we introduce the *between scatter* matrix $\mathbf{\Sigma}_{X_b} = \mathbf{\Sigma}_X - \mathbf{\Sigma}_{X_w}$ as the difference between *total* and *within scatter*.

2.2. Deep Canonical Correlation Analysis (DCCA)

In this section we review a deep neural network extension to classical CCA introduced by (Andrew et al., 2013). In their work CCA is used to combine the topmost feature representations of two different neural network $\mathbf{f}$ and $\mathbf{g}$ as shown in Figure 1a. The DCCA optimization target pushes the networks to learn highly correlated feature representations. Based on the covariances introduced above CCA defines a matrix $\mathbf{T} = \mathbf{\Sigma}_X^{-1/2} \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_Y^{-1/2}$. The total correlation between $\mathbf{f}_X$ and $\mathbf{g}_Y$ is then computed as the sum over the singular values $\mathbf{d}$ with corresponding singular value problem $\mathbf{T} = \mathbf{UDV}$ and $\mathbf{D} = diag(\mathbf{d})$. $\mathbf{U}$ and $\mathbf{V}$ are the projection matrices which transform the two views into the linear CCA sub-space. The correlation itself is optimized by maximizing the sum over the singular values $\mathbf{d}$ with respect to the network parameters $\Theta_{\mathbf{f}}$ and $\Theta_{\mathbf{g}}$:

$$\arg\max_{\Theta_{\mathbf{f}}, \Theta_{\mathbf{g}}} \sum_{i=1}^{h} d_i \tag{3}$$

If $\mathbf{f}$ and $\mathbf{g}$ have the same feature dimensionality h it is also possible to optimize the canonical correlation by maximizing the matrix trace norm $||\mathbf{T}||_{tr} = tr((\mathbf{T}^T \mathbf{T})^{1/2})$. For a detailed derivation of the DCCA optimization target we refer to the work of Andrew et al. (2013).

2.3. Deep Linear Discriminant Analysis (DeepLDA)

The central idea of DeepLDA is to put LDA on top of a deep neural network to learn latent representations which maximize the separation between the C individual classes (Dorfer et al., 2015). We illustrate this in Figure 1b. LDA in general finds a projection matrix $\mathbf{A}$ that maximizes the ratio of *between scatter* $\mathbf{\Sigma}_{X_b}$ and *within scatter* $\mathbf{\Sigma}_{X_w}$:

$$\arg\max_{\mathbf{A}} \frac{|\mathbf{A} \mathbf{\Sigma}_{X_b} \mathbf{A}^T|}{|\mathbf{A} \mathbf{\Sigma}_{X_w} \mathbf{A}^T|} \tag{4}$$

Projection matrix $\mathbf{A}$ transforms the data into a $C-1$ dimensional space where the observations become linearly separable. The linear combinations $\mathbf{A}$ which maximize the class separation are determined by solving the generalized LDA eigenvalue problem $\mathbf{\Sigma}_{X_b} \mathbf{e} = \mathbf{v} \mathbf{\Sigma}_{X_w} \mathbf{e}$. The resulting eigenvalues $\mathbf{v}$ quantify the separation in direction of the eigenvectors and the projection matrix $\mathbf{A}$ is exactly this set of corresponding eigenvectors $\mathbf{e}$.

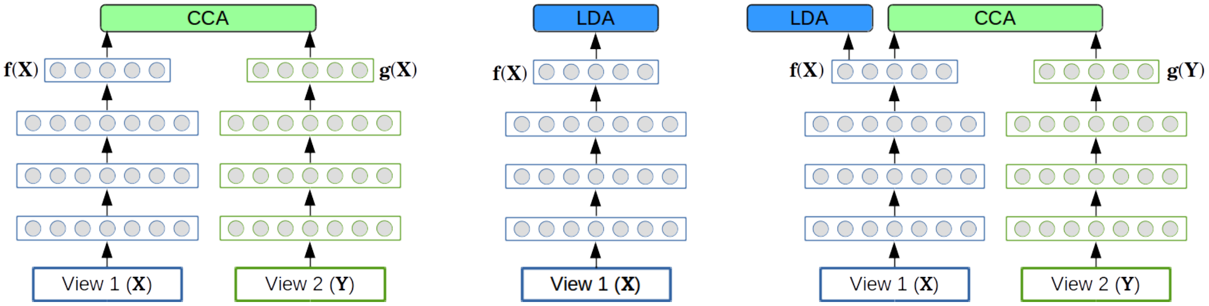
DeepLDA makes use of the beneficial properties of LDA by casting it as a deep learning optimization target. The related optimization target focuses on maximizing the k smallest eigenvalues $\{v_1, ..., v_k\}$ as follows:

$$\arg\max_{\Theta} \frac{1}{k} \sum_{i=1}^{k} v_i \tag{5}$$

where $\{v_1, ..., v_k\} = \{v_j | v_j < \min\{v_1, ..., v_{C-1}\} + \epsilon\}$. The design of this objective encourages the network to push separation (discriminative power) into all available dimensions of the eigenspace (Dorfer et al., 2015). The feature space resulting from DeepLDA optimization has also proven to be suitable for subsequent classification tasks.

2.4. Discriminative Canonical Correlation Analysis

The two methods introduced above (DCCA and DeepLDA) are based on the optimization of the eigenvalue structure of their corresponding eigenvalue problems. DCCA optimization tackles the singular values of matrix $\mathbf{T}$ with the goal of maximizing the correlation of the hidden representations learned by two different neural networks. DeepLDA on the other hand maximizes the separation of classes, which is



(a) DCCA (b) DeepLDA (c) Joint DCCA / DeepLDA.

Figure 1. Schematic sketches of all methods involved in the joint optimization of correlation and separation. All methods have in common that they are built on top of a neural network. DeepLDA operates only on a single view whereas DCCA and the joint model operate simultaneously on both views. $\mathbf{f}(\mathbf{X})$ and $\mathbf{g}(\mathbf{X})$ are the hidden representations learned by the receptive networks.

quantified by the magnitude of the eigenvalues of the corresponding generalized eigenvalue problem. However, both methods have in common that they back-propagate an error resulting from an eigenvalue problem to tune the parameters of a deep neural network.

The core of the present work is to reuse both concepts in a multi-view representation learning setting. We formulate a joint optimization target that simultaneously optimizes the correlation between the hidden representations of two different views as well as the discriminative power of the learned representations (see Figure 1c):

$$\underset{\Theta_{\mathbf{f}},\Theta_{\mathbf{g}}}{\arg\max}\ \lambda_{CCA}\frac{1}{h}\sum_{i=1}^{h} d_i + \lambda_{LDA}\frac{1}{k}\sum_{i=1}^{k} v_i \qquad (6)$$

λ_{CCA} and λ_{LDA} are weighting factors that balance the relative importance of the two tasks at hand. In addition, we normalize the canonical correlation by the dimensionality h of the hidden spaces $\mathbf{f}_X$ and $\mathbf{g}_Y$. The maximum correlation achievable is therefore 1.0 compared to h as for example described in (Wang & Livescu, 2015). We emphasize that we apply the discriminative regularizer only to the first of the two views (compare also Figure 1c). This will also be the view which is used at test time for a classification task in our experiments. The design choice is made to be in line with the experiments described in (Andrew et al., 2013). However, it is straightforward to extend the application of the discriminative regularizer to view 2 or oven to both of the hidden representations.

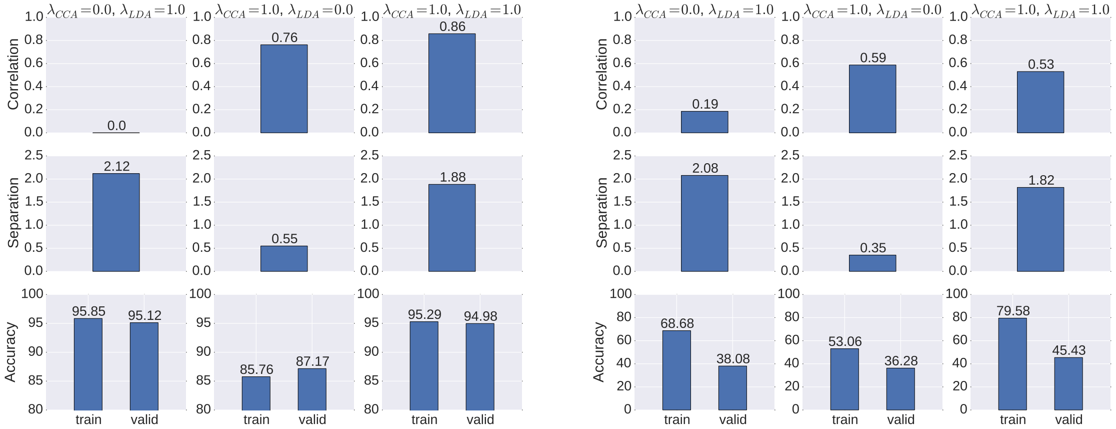
3. Experiments

In the following we evaluate our joint optimization approach on two different benchmark datasets as for example used in (Wang & Livescu, 2015). Besides focusing our evaluation on the magnitude of canonical correlations between the two different input views we also investigate the discriminative power of the learned representations. As evaluation measures we report the canonical correlation between the views, the classification accuracy as well as the magnitude of the eigenvalues (separation) of an LDA on hidden representation $\mathbf{f}_X$. In particular, we evaluate the discriminative power (accuracy, separation) of the learned representations on the CCA projected data $\mathbf{U}\mathbf{f}_X$ of view 1 ($\mathbf{U}$ is the linear projection yielded by the singularvalue problem introduced in Section 2.2). This is an important fact and means that the second view is not required at test time at all (Andrew et al., 2013; Wang & Livescu, 2015).

3.1. Network Architecture and Optimization

In this subsection we briefly outline the network architectures as well as the optimization strategies used to train our models. For the split MNIST task (Section 3.2) we select a network with three densely connected layers each having 1024 units. The dimensionality of the topmost hidden representations is 50 for both views and the batch size is set to 2500 examples. For the Wisconsin X-ray Microbeam (XRMB) Speech Database (Section 3.3) we use the same network architecture as described in (Wang & Livescu, 2015). For view 1 we use three dense layers with 1500 units and one dense layer with linear activation for view 2. The topmost hidden representations have again 50 output neurons and the batch size is set to 5000 examples. One thing we did differently in our architectures is the activation function of the three dense layers. Instead of rectified linear units (ReLUs) (Nair & Hinton, 2010) we apply exponential linear units (ELUs) (Clevert et al., 2015) [2]. Finally, we train both models with the *adam* optimizer (Kingma & Ba, 2014), an initial learning rate of 0.0005 and a learning rate decay of 0.5.

[2] In our initial experiments we encountered problems with numerical stability resulting in *Not a Number* (NaN) errors during optimization. We already talked to the authors but we do not know yet why this problem is reduced when using ELU activations.



(a) Results on the split MNIST dataset.

(b) Results on XRMB dataset.

Figure 2. Comparative evaluation of accuracy, separation and correlation on two experimental datasets. We report results for three different weightings for the discriminative (λ_{LDA}) and the canonical correlation (λ_{CCA}) part of the joint loss.

3.2. Split MNIST

The setup for this experiment has already been used in (Andrew et al., 2013; Wang & Livescu, 2015) as a proof of concept for their methods. MNIST contains small 28×28 pixel images showing digits ranging from 0 to 9. The dataset holds 70000 images where 50000 are used for training and 10000 for evaluation and testing, respectively. The experimental setup is as follows. Each of the images is split along the vertical center line resulting in two sub-images having size 28×14. The two pieces are interpreted as two different views on the same digit and used to train the networks $\mathbf{f}$ and $\mathbf{g}$ to maximizes the canonical correlation between the representations of the pieces. In contrast to the methods mentioned above we also use the class labels at train time as they are required to compute the individual class covariance matrices for DeepLDA (see Section 2.3). Figure 2a shows our results on the split MNIST task. The rows in the plots report the three evaluation measures correlation, separation and accuracy. For separation and correlation we report only results on the train set as this is the data where the eigenvalue problems are solved and the projection matrices are estimated. The columns show three different settings for weighting the correlation and the separation part of the proposed joint loss. One observation attracting attention is that in the case $\lambda_{CCA} = 0.0$, $\lambda_{LDA} = 1.0$ there is no correlation at all between the hidden representations. This is due to the fact that the LDA loss applies only to sub-network $\mathbf{f}$ processing the data originating from view 1 (this is the view which is used at test time). Sub-network $\mathbf{g}$ remains completely untouched in this case. The main observation we would like to underline becomes clear when investigating column two and three in the plot. The correlation is high with values of 0.76 and 0.86 for both settings. However, the separation as well as the classification accuracy on both train and validation set are highly increased for the multi-task case $\lambda_{CCA} = 1.0$, $\lambda_{LDA} = 1.0$. This suggests that it is feasible to jointly optimize correlation and separation at the same time. Our results on the XRMB dataset reported in the following section will further emphasize this observation.

3.3. Wisconsin X-ray Microbeam Speech Database

The XRMB database introduced by Westbury (1994) is a multi-view speech production dataset. It contains articulatory as well as acoustic features recorded simultaneously from 47 different English speakers. Along with the multimodal data there are phonetic labels available for classification. Andrew et al. (2013) show in their work on DCCA that optimizing the correlation between the two sets of different features also has a positive effect on the recognition accuracy of the individual phonemes. In terms of input features we follow the works of Andrew et al. (2013); Wang et al. (2015); Wang & Livescu (2015). The acoustic features (13 mel frequency cepstral coefficients (MFCCs)) are presented to the network along with their first and second derivative as 7 frame windows around each target frame ($\mathbf{X} \in \mathbb{R}^{N \times 273}$). The articulatory data ($\mathbf{Y} \in \mathbb{R}^{N \times 112}$) is a recording of the displacements of eight pellets placed on the speaker's lips, tongue and jaws. The features are further sub-sampled to be in line with their acoustic counterpart. We would like to emphasize that the results reported in this section are preliminary and we only report results for three different speakers of a three fold cross validation (two train speakers, one validation speaker)[3]. The experimental

[3] Unfortunately we did not have the entire XRMB dataset available when preparing this manuscript. To pro-

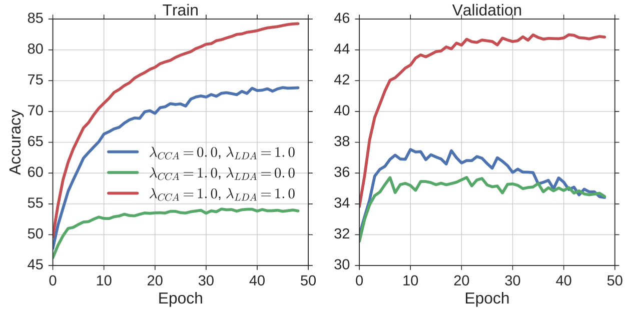


Figure 3. Evolution of train and validation error for the three different weighting schemes averaged over three folds. The joint optimization has a clear advantage compared to an optimization on the individual tasks alone.

setup is analogous to the one reported above. We evaluate three different joint weight parametrizations and report the phoneme recognition accuracy on the CCA projected hidden representation of the acoustic output $\mathbf{f}_X$ of the network. Figure 3 shows the evolution of train and validation error over the training epochs of the networks averaged over the tree folds. The main finding is that when jointly optimizing on both objectives ($\lambda_{CCA} = 1.0$, $\lambda_{LDA} = 1.0$) the learned representation is capable of achieving higher validation accuracies compared to an optimization on the correlation and discriminative task alone. In Figure 2b we again outline our results in detail and select for each parametrization the results of the epoch where the validation accuracy is highest. We already observed that the classification accuracy on the validation set is highest when optimizing on the joint optimization target. When taking a closer look at the correlation and separation structure of the representations we see that in the case of $\lambda_{CCA} = 1.0$, $\lambda_{LDA} = 1.0$ the measures are not at the maximum of the three configurations. However, when investigating both at the same time they are highly increased compared to the single objective cases. We expect that the optimization of both targets at the same time forces the network to learn more general representations which are less prone to over-fitting. Although our results are preliminary we think they are very promising and suggest further investigations.

4. Conclusion

We have presented a discriminative extension to DCCA multi-view representation learning. DCCA optimizes the correlation between the hidden representations of two neural networks which process the data of two different input views. In our work we not only maximize the canonical correlation but propose to add an additional discriminative regularizer on the hidden representation of the view intended to be used at test time. We do this by reusing the eigenvalue objective function of DeepLDA. Preliminary experiments on two multi-view datasets indicate that joint correlation separation optimization has the potential to improve the classification performance of the learned representations. We think that the results are promising and conclude that joint multi-view optimization is a promising future direction which deserves further investigation.

References

Andrew, Galen, Arora, Raman, Bilmes, Jeff, and Livescu, Karen. Deep canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning*, pp. 1247–1255, 2013.

Clevert, Djork-Arné, Unterthiner, Thomas, and Hochreiter, Sepp. Fast and accurate deep network learning by exponential linear units (elus). *International Conference on Learning Representations (ICLR) (arXiv:1511.07289)*, 2015.

Dorfer, Matthias, Kelz, Rainer, and Widmer, Gerhard. Deep linear discriminant analysis. *International Conference on Learning Representations (ICLR) (arXiv:1511.04707)*, 2015.

Fisher, Ronald A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Hotelling, Harold. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nair, Vinod and Hinton, Geoffrey E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.

Wang, Weiran and Livescu, Karen. Large-scale approximate kernel canonical correlation analysis. *International Conference on Learning Representations (ICLR) (arXiv:1511.04773)*, 2015.

Wang, Weiran, Arora, Raman, Livescu, Karen, and Bilmes, Jeff. On deep multi-view representation learning. In *Proceedings of the International Conference on Machine Learning*, 2015.

Westbury, J. R. X-ray microbeam speech production database user's handbook. *Waisman Center, University of Wisconsin: Madison, USA*, pp. 1–100, 1994.

vide initial results we take the data of three speakers available at https://github.com/corbyrosset/Correlation-Analysis-and-Friends.