

DEEP GENERALIZED CANONICAL CORRELATION ANALYSIS

**Adrian Benton, Huda Khayrallah, Biman Gujral,
Drew Reisinger, Sheng Zhang, Raman Arora**

Center for Language and Speech Processing

Johns Hopkins University

Baltimore, MD 21218, USA

adrian[†], huda^{*}, bgujral1^{*}, reisinger[◊], zsheng2^{*}, arora[†]

*@jhu.edu, ◊@cogsci.jhu.edu, †@cs.jhu.edu

ABSTRACT

We present Deep Generalized Canonical Correlation Analysis (DGCCA) – a method for learning nonlinear transformations of arbitrarily many views of data, such that the resulting transformations are maximally informative of each other. While methods for nonlinear two-view representation learning (Deep CCA, (Andrew et al., 2013)) and linear many-view representation learning (Generalized CCA (Horst, 1961)) exist, DGCCA is the first CCA-style multiview representation learning technique that combines the flexibility of nonlinear (deep) representation learning with the statistical power of incorporating information from many independent sources, or views. We present the DGCCA formulation as well as an efficient stochastic optimization algorithm for solving it. We learn DGCCA representations on two distinct datasets for three downstream tasks: phonetic transcription from acoustic and articulatory measurements, and recommending hashtags and friends on a dataset of Twitter users. We find that DGCCA representations soundly beat existing methods at phonetic transcription and hashtag recommendation, and in general perform no worse than standard linear many-view techniques.

1 INTRODUCTION

Multiview representation learning refers to settings where one has access to many “views” of data, at train time. Views often correspond to different modalities or independent information about examples: a scene represented as a series of audio and image frames, a social media user characterized by the messages they post and who they friend, or a speech utterance and the configuration of the speaker’s tongue. Multiview techniques learn a representation of data that captures the sources of variation common to all views.

Multiview representation techniques are attractive for intuitive reasons. A representation that is able to explain many views of the data is more likely to capture meaningful variation than a representation that is a good fit for only one of the views. They are also attractive for the theoretical reasons. For example, Anandkumar et al. (2014) show that certain classes of latent variable models, such as Hidden Markov Models, Gaussian Mixture Models, and Latent Dirichlet Allocation models, can be optimally learned with multiview spectral techniques. Representations learned from many views will generalize better than one, since the learned representations are forced to accurately capture variation in all views at the same time (Sridharan & Kakade, 2008) – each view acts as a regularizer, constraining the possible representations that can be learned. These methods are often based on canonical correlation analysis (CCA), a classical statistical technique proposed by Hotelling (1936).

In spite of encouraging theoretical guarantees, multiview learning techniques cannot freely model nonlinear relationships between arbitrarily many views. Either they are able to model variation across many views, but can only learn linear mappings to the shared space (Horst, 1961), or they simply cannot be applied to data with more than two views using existing techniques based on kernel CCA (Hardoon et al., 2004) and deep CCA (Andrew et al., 2013).

Here we present Deep Generalized Canonical Correlation Analysis (DGCCA). Unlike previous correlation-based multiview techniques, DGCCA learns a shared representation from data with arbitrarily many views and simultaneously learns nonlinear mappings from each view to this shared space. The only (mild) constraint is that these nonlinear mappings from views to shared space must be differentiable. Our main methodological contribution is the derivation of the gradient update for the Generalized Canonical Correlation Analysis (GCCA) objective (Horst, 1961). As a practical contribution, we have also released an implementation of DGCCA¹.

We also evaluate DGCCA-learned representations on two distinct datasets and three downstream tasks: phonetic transcription from aligned speech and articulatory data, and Twitter hashtag and friend recommendation from six text and network feature views. We find that downstream performance of DGCCA representations is ultimately task-dependent. However, we find clear gains in performance from DGCCA for tasks previously shown to benefit from representation learning on more than two views, with up to 4% improvement in heldout accuracy for phonetic transcription.

The paper is organized as follows. We review prior work in Section 2. In Section 3 we describe DGCCA. Empirical results on a synthetic dataset, and three downstream tasks are presented in Section 4. In Section 5, we describe the differences between DGCCA and other non-CCA-based multiview learning work and conclude with future directions in Section 6.

2 PRIOR WORK

Some of most successful techniques for multiview representation learning are based on canonical correlation analysis (Wang et al., 2015a;b) and its extension to the nonlinear and many view settings, which we describe in this section. For other related multiview learning techniques, see Section 5.

2.1 CANONICAL CORRELATION ANALYSIS (CCA)

Canonical correlation analysis (CCA) (Hotelling, 1936) is a statistical method that finds maximally correlated linear projections of two random vectors and is a fundamental multiview learning technique. Given two input views, $X_1 \in \mathbb{R}^{d_1}$ and $X_2 \in \mathbb{R}^{d_2}$, with covariance matrices, Σ_{11} and Σ_{22} , respectively, and cross-covariance matrix, Σ_{12} , CCA finds directions that maximize the correlation between them:

$$(u_1^*, u_2^*) = \operatorname{argmax}_{u_1 \in \mathbb{R}^{d_1}, u_2 \in \mathbb{R}^{d_2}} \operatorname{corr}(u_1^\top X_1, u_2^\top X_2) = \operatorname{argmax}_{u_1 \in \mathbb{R}^{d_1}, u_2 \in \mathbb{R}^{d_2}} \frac{u_1^\top \Sigma_{12} u_2}{\sqrt{u_1^\top \Sigma_{11} u_1 u_2^\top \Sigma_{22} u_2}}$$

Since this formulation is invariant to affine transformations of u_1 and u_2 , we can write it as the following constrained optimization formulation:

$$(u_1^*, u_2^*) = \operatorname{argmax}_{u_1^\top \Sigma_{11} u_1 = u_2^\top \Sigma_{22} u_2 = 1} u_1^\top \Sigma_{12} u_2 \quad (1)$$

This technique has two limitations that have led to significant extensions: First, it is limited to learning representations that are *linear* transformations of the data in each view, and second, it can only leverage two input views.

2.2 DEEP CANONICAL CORRELATION ANALYSIS (DCCA)

Deep CCA (DCCA) (Andrew et al., 2013) is an extension of CCA that addresses the first limitation by finding maximally linearly correlated *non-linear transformations* of two vectors. It does this by passing each of the input views through stacked non-linear representations and performing CCA on the outputs.

Let us use $f_1(X_1)$ and $f_2(X_2)$ to represent the network outputs. The weights, W_1 and W_2 , of these networks are trained through standard backpropagation to maximize the CCA objective:

$$(u_1^*, u_2^*, W_1^*, W_2^*) = \operatorname{argmax}_{u_1, u_2} \operatorname{corr}(u_1^\top f_1(X_1), u_2^\top f_2(X_2))$$

DCCA is still limited to only 2 input views.

¹See <https://bitbucket.org/adrianbenton/dgcca-py3> for implementation of DGCCA along with data from the synthetic experiments.

2.3 GENERALIZED CANONICAL CORRELATION ANALYSIS (GCCA)

Another extension of CCA, which addresses the limitation on the number of views, is Generalized CCA (GCCA) (Horst, 1961). It corresponds to solving the optimization problem in Equation (2), of finding a shared representation G of J different views, where N is the number of data points, d_j is the dimensionality of the j th view, r is the dimensionality of the learned representation, and $X_j \in \mathbb{R}^{d_j \times N}$ is the data matrix for the j th view.²

$$\begin{aligned} & \underset{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}}{\text{minimize}} && \sum_{j=1}^J \|G - U_j^\top X_j\|_F^2 \\ & \text{subject to} && GG^\top = I_r \end{aligned} \quad (2)$$

Solving GCCA requires finding an eigendecomposition of an $N \times N$ matrix, which scales quadratically with sample size and leads to memory constraints. Unlike CCA and DCCA, which only learn projections or transformations on each of the views, GCCA also learns a view-independent representation G that best reconstructs all of the view-specific representations simultaneously. The key limitation of GCCA is that it can only learn *linear* transformations of each view.

3 DEEP GENERALIZED CANONICAL CORRELATION ANALYSIS (DGCCA)

In this section, we present deep GCCA (DGCCA): a multiview representation learning technique that benefits from the expressive power of deep neural networks and can also leverage statistical strength from more than two views in data, unlike Deep CCA which is limited to only two views. More fundamentally, deep CCA and deep GCCA have very different objectives and optimization problems, and it is not immediately clear how to extend deep CCA to more than two views.

DGCCA learns a nonlinear map for each view in order to maximize the correlation between the learnt representations across views. In training, DGCCA passes the input vectors in each view through multiple layers of nonlinear transformations and backpropagates the gradient of the GCCA objective with respect to network parameters to tune each view’s network, as illustrated in Figure 1. The objective is to train networks that reduce the GCCA reconstruction error among their outputs. At test time, new data can be projected by feeding them through the learned network for each view.

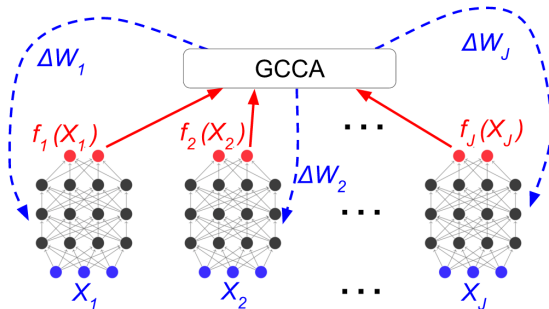


Figure 1: A schematic of DGCCA with deep networks for J views.

We now formally define the DGCCA problem. We consider J views in our data, and let $X_j \in \mathbb{R}^{d_j \times N}$ denote the j th input matrix.³ The network for the j th view consists of K_j layers. Assume, for simplicity, that each layer in the j th view network has c_j units with a final (output) layer of size o_j . The output of the k th layer for the j th view is $h_k^j = s(W_k^j h_{k-1}^j)$, where $s : \mathbb{R} \rightarrow \mathbb{R}$ is a nonlinear activation function and $W_k^j \in \mathbb{R}^{c_k \times c_{k-1}}$ is the weight matrix for the k th layer of the j th view network. We denote the output of the final layer as $f_j(X_j)$.

²Horst’s original presentation of GCCA is in a very different form from the one used here but has been shown to be equivalent (Kettenring, 1971).

³Our notation for this section closely follows that of Andrew et al. (2013)

DGCCA can be expressed as the following optimization problem: find weight matrices $W^j = \{W_1^j, \dots, W_{K_j}^j\}$ defining the functions f_j , and linear transformations U_j (of the output of the j^{th} network), for $j = 1, \dots, J$, that

$$\begin{aligned} & \underset{U_j \in \mathbb{R}^{o_j \times r}, G \in \mathbb{R}^{r \times N}}{\text{minimize}} && \sum_{j=1}^J \|G - U_j^\top f_j(X_j)\|_F^2, \\ & \text{subject to} && GG^\top = I_r \end{aligned} \quad (3)$$

where $G \in \mathbb{R}^{r \times N}$ is the shared representation we are interested in learning.

Optimization: We solve the DGCCA optimization problem using stochastic gradient descent (SGD) with mini-batches. In particular, we estimate the gradient of the DGCCA objective in Problem 3 on a mini-batch of samples that is mapped through the network and use back-propagation to update the weight matrices, W^j 's. However, note that the DGCCA optimization problem is a constrained optimization problem. It is not immediately clear how to perform projected gradient descent with back-propagation. Instead, we characterize the objective function of the GCCA problem at an optimum, and compute its gradient with respect to the inputs to GCCA, i.e. with respect to the network outputs. These gradients are then back-propagated through the network to update W^j 's.

Although the relationship between DGCCA and GCCA is analogous to the relationship between DCCA and CCA, derivation of the GCCA objective gradient with respect to the network output layers is non-trivial. The main difficulty stems from the fact that there is no natural extension of the correlation objective to more than two random variables. Instead, we consider correlations between every pair of views, stack them in a $J \times J$ matrix and maximize a certain matrix norm for that matrix. For GCCA, this suggests an optimization problem that maximizes the sum of correlations between a shared representation and each view. Since the objective as well as the constraints of the generalized CCA problem are very different from that of the CCA problem, it is not immediately obvious how to extend Deep CCA to Deep GCCA.

Next, we show a sketch of the gradient derivation, the full derivation is given in appendix A. It is straightforward to show that the solution to the GCCA problem is given by solving an eigenvalue problem. In particular, define $C_{jj} = f(X_j)f(X_j)^\top \in \mathbb{R}^{o_j \times o_j}$, to be the scaled empirical covariance matrix of the j^{th} network output, and $P_j = f(X_j)^\top C_{jj}^{-1} f(X_j) \in \mathbb{R}^{N \times N}$ be the corresponding projection matrix that whitens the data; note that P_j is symmetric and idempotent. We define $M = \sum_{j=1}^J P_j$. Since each P_j is positive semi-definite, so is M . Then, it is easy to check that the rows of G are the top r (orthonormal) eigenvectors of M , and $U_j = C_{jj}^{-1} f(X_j) G^\top$. Thus, at the minimum of the objective, we can rewrite the reconstruction error as follows:

$$\sum_{j=1}^J \|G - U_j^\top f_j(X_j)\|_F^2 = \sum_{j=1}^J \|G - G f_j(X_j)^\top C_{jj}^{-1} f_j(X_j)\|_F^2 = rJ - \text{Tr}(GMG^\top)$$

Minimizing the GCCA objective (w.r.t. the weights of the neural networks) means maximizing $\text{Tr}(GMG^\top)$, which is the sum of eigenvalues $L = \sum_{i=1}^r \lambda_i(M)$. Taking the derivative of L with respect to each output layer $f_j(X_j)$ we have:

$$\frac{\partial L}{\partial f_j(X_j)} = 2U_j G - 2U_j U_j^\top f_j(X_j)$$

Thus, the gradient is the difference between the r -dimensional auxiliary representation G embedded into the subspace spanned by the columns of U_j (the first term) and the projection of the actual data in $f_j(X_j)$ onto the said subspace (the second term). Intuitively, if the auxiliary representation G is far away from the view-specific representation $U_j^\top f_j(X_j)$, then the network weights should receive a large update. Computing the gradient descent update has time complexity $O(JNrd)$, where $d = \max(d_1, d_2, \dots, d_J)$ is the largest dimensionality of the input views.

4 EXPERIMENTS

4.1 SYNTHETIC MULTIVIEW MIXTURE MODEL

In this section, we apply DGCCA to a small synthetic data set to show how it preserves the generative structure of data sampled from a multiview mixture model. The data we use for this experiment are

plotted in Figure 2. Points that share the same color across different views are sampled from the same mixture component.

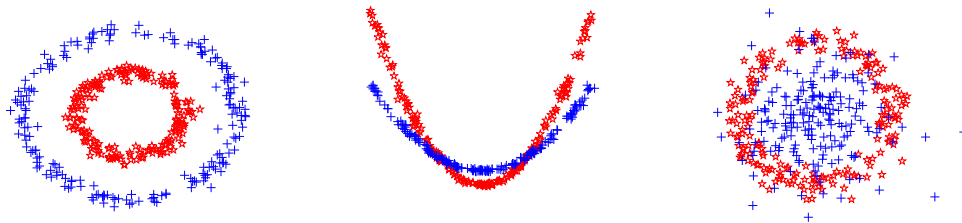
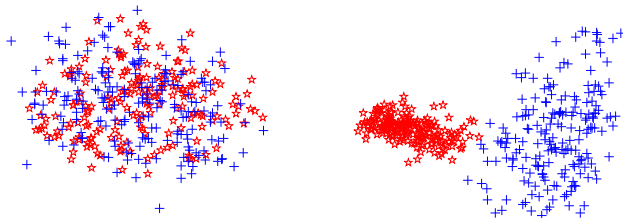


Figure 2: Synthetic data used in in Section 4.1 experiments.

Importantly, in each view, there is no linear transformation of the data that separates the two mixture components, in the sense that the generative structure of the data could not be exploited by a linear model. This point is reinforced by Figure 3(a), which shows the two-dimensional representation G learned by applying (linear) GCCA to the data in Figure 2. The learned representation completely loses the structure of the data.



(a) GCCA

(b) DGCCA

Figure 3: The matrix G learned from applying (linear) GCCA or DGCCA to the data in Figure 2.

We can contrast the failure of GCCA to preserve structure with the result of applying DGCCA; in this case, the input neural networks had three hidden layers with ten units each with weights randomly initialized. We plot the representation G learned by DGCCA in Figure 3 (b). In this representation, the mixture components are easily separated by a linear classifier; in fact, the structure is largely preserved even after projection onto the first coordinate of G .

It is also illustrative to consider the view-specific representations learned by DGCCA, that is, to consider the outputs of the neural networks that were trained to maximize the GCCA objective. We plot the representations in Figure 4. For each view, we have learned a nonlinear mapping that does remarkably well at making the mixture components linearly separable. Recall that absolutely no direct supervision was given about which mixture component each point was generated from. The only training signals available to the networks were the reconstruction errors between the network outputs and the learned representation G .

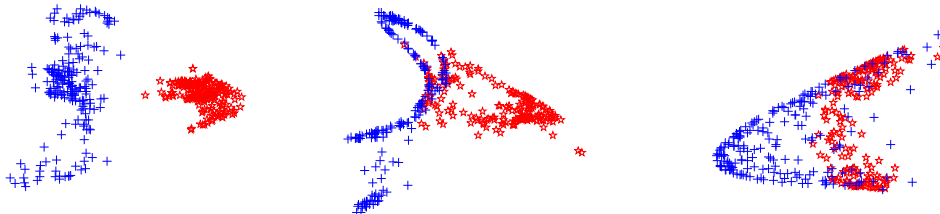


Figure 4: Outputs of the trained input neural networks in Section 4.1 applied to the data in Figure 2.

4.2 PHONEME CLASSIFICATION

In this section, we discuss experiments on the University of Wisconsin X-ray Microbeam Database (XRMB) (Westbury, 1994). XRMB contains acoustic and articulatory recordings as well as phonemic labels. We present phoneme classification results on the acoustic vectors projected using DCCA,

GCCA, and DGCCA. We set acoustic and articulatory data as the two views and phoneme labels as the third view for GCCA and DGCCA. For classification, we run K-nearest neighbor classification (Cover & Hart, 1967) on the projected result.

4.2.1 DATA

We use the same train/tune/test split of the data as Arora & Livescu (2014). To limit experiment runtime, we use a subset of speakers for our experiments. We run a set of *cross-speaker* experiments using the male speaker JW11 for training and two splits of JW24 for tuning and testing. We also perform parameter tuning for the third view with 5-fold cross validation using a *single speaker*, JW11. For both experiments, we use acoustic and articulatory measurements as the two views in DCCA. Following the pre-processing in Andrew et al. (2013), we get 273 and 112 dimensional feature vectors for the first and second view respectively. Each speaker has $\sim 50,000$ frames. For the third view in GCCA and DGCCA, we use 39-dimensional one-hot vectors corresponding to the labels for each frame, following Arora & Livescu (2014).

4.2.2 PARAMETERS

We use a fixed network size and regularization for the first two views, each containing three hidden layers with sigmoid activation functions. Hidden layers for the acoustic view were all width 1024, and layers in the articulatory view all had width 512 units. L2 penalty constants of 0.0001 and 0.01 were used to train the acoustic and articulatory view networks, respectively. The output layer dimension of each network is set to 30 for DCCA and DGCCA. For the 5-fold speaker-dependent experiments, we performed a grid search for the network sizes in $\{128, 256, 512, 1024\}$ and covariance matrix regularization in $\{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}\}$ for the third view in each fold. We fix the hyperparameters for these experiments optimizing the networks with minibatch stochastic gradient descent with a step size of 0.005, batch size of 2000, and no learning decay or momentum. The third view neural network had an L2 penalty of 0.0005.

4.2.3 RESULTS

As we show in Table 1, DGCCA improves upon both the linear multiview GCCA and the non-linear 2-view DCCA for both the cross-speaker and speaker-dependent cross-validated tasks.

In addition to accuracy, we examine the reconstruction error, i.e. the objective in Equation 3, obtained from the objective in GCCA and DGCCA.⁴ This sharp improvement in reconstruction error shows that a non-linear algorithm can better model the data.

In this experimental setup, DCCA under-performs the baseline of simply running KNN on the original acoustic view. Prior work considered the output of DCCA stacked on to the central frame of the original acoustic view (39 dimensions). This poor performance, in the absence of original features, indicates that it was not able to find a more informative projection than original acoustic features based on correlation with the articulatory view within the first 30 dimensions.

Table 1: KNN phoneme classification performance

METHOD	CROSS-SPEAKER			SPEAKER-DEPENDENT		
	DEV ACC	TEST ACC	REC ERROR	DEV ACC	TEST ACC	REC ERROR
MFCC	48.89	49.28		66.27	66.22	
DCCA	45.40	46.06		65.88	65.81	
GCCA	49.59	50.18	40.67	69.52	69.78	40.39
DGCCA	53.78	54.22	35.89	72.62	72.33	20.52

To highlight the improvements of DGCCA over GCCA, Figure 5 presents a subset of the the confusion matrices on speaker-dependent test data. In particular, we observe large improvements in the classification of *D*, *F*, *K*, *SH*, *V* and *Y*. GCCA outperforms DGCCA for *UH* and *DH*. These matrices also highlight the common misclassifications that DGCCA improves upon. For instance,

⁴For 2-view experiments, correlation is a common metric to compare performance. Since that metric is unavailable in a multiview setting, reconstruction error is the analogue.

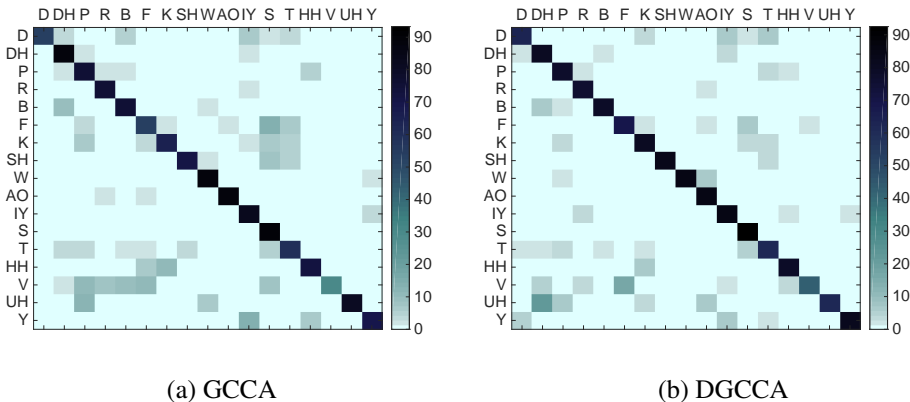


Figure 5: The confusion matrix for speaker-dependent GCCA and DGCCA

DGCCA rectifies the frequent misclassification of V as P , R and B by GCCA. In addition, commonly incorrect classification of phonemes such as S and T is corrected by DGCCA, which enables better performance on other voiceless consonants such as like F , K and SH . Vowels are classified with almost equal accuracy by both the methods.

4.3 TWITTER USER HASHTAG & FRIEND RECOMMENDATION

Linear multiview techniques are effective at recommending hashtag and friends for Twitter users (Benton et al., 2016). In this experiment, six views of a Twitter user were constructed by applying principal component analysis (PCA) to the bag-of-words representations of (1) tweets posted by the ego user, (2) other mentioned users, (3) their friends, and (4) their followers, as well as one-hot encodings of the local (5) friend and (6) follower networks. We learn and evaluate DGCCA models on identical training, development, and test sets as Benton et al. (2016), and evaluate the DGCCA representations on macro precision at 1000 (P@1000) and recall at 1000 (R@1000) for the hashtag and friend recommendation tasks described there.

We trained 40 different DGCCA model architectures, each with identical architectures across views, where the width of the hidden and output layers, c_1 and c_2 , for each view are drawn uniformly from $[10, 1000]$, and the auxiliary representation width r is drawn uniformly from $[10, c_2]^5$. All networks used ReLUs as activation functions, and were optimized with Adam (Kingma & Ba, 2014) for 200 epochs⁶. Networks were trained on 90% of 102,328 Twitter users, with 10% of users used as a tuning set to estimate heldout reconstruction error for model selection. We report development and test results for the best performing model on the downstream task development set. Learning rate was set to 10^{-4} with an L1 and L2 regularization constants of 0.01 and 0.001 for all weights⁷.

Table 2: Dev/test performance at Twitter friend and hashtag recommendation tasks.

ALGORITHM	FRIEND		HASHTAG	
	P@1000	R@1000	P@1000	R@1000
PCA[TEXT+NET]	0.445/0.439	0.149/0.147	0.011/0.008	0.312/0.290
GCCA[TEXT]	0.244/0.249	0.080/0.081	0.012/0.009	0.351/0.326
GCCA[TEXT+NET]	0.271/0.276	0.088/0.089	0.012/0.010	0.359/0.334
DGCCA[TEXT+NET]	0.297/0.268	0.099/0.090	0.013/0.010	0.385/0.373
WGCCA[TEXT]	0.269/0.279	0.089/0.091	0.012/0.009	0.357/0.325
WGCCA[TEXT+NET]	0.376/0.364	0.123/0.120	0.013/0.009	0.360/0.346

Table 2 displays the performance of DGCCA compared to PCA[*text+net*] (PCA applied to concatenation of view feature vectors), linear GCCA applied to the four text views, [*text*], and all

⁵We chose to restrict ourselves to a single hidden layer with non-linear activation and identical architectures for each view, so as to avoid a fishing expedition. If DGCCA is appropriate for learning Twitter user representations, then a good architecture should require little exploration.

⁶From preliminary experiments, we found that Adam pushed down reconstruction error more quickly than SGD with momentum, and that ReLUs were easier to optimize than sigmoid activations.

⁷This setting of regularization constants led to low reconstruction error in preliminary experiments.

views, $[text+net]$, along with a weighted GCCA variant (WGCCA). We learned PCA, GCCA, and WGCCA representations of width $r \in \{10, 20, 50, 100, 200, 300, 400, 500, 750, 1000\}$, and report the best performing representations on the development set.

There are several points to note: First is that DGCCA outperforms linear methods at hashtag recommendation by a wide margin in terms of recall. This is exciting because this task was shown to benefit from incorporating more than just two views from Twitter users. These results suggest that a nonlinear transformation of the input views can yield additional gains in performance. In addition, WGCCA models sweep over every possible weighting of views with weights in $\{0, 0.25, 1.0\}$. WGCCA has a distinct advantage in that the model is allowed to discriminatively weight views to maximize downstream performance. The fact that DGCCA is able to outperform WGCCA at hashtag recommendation is encouraging, since WGCCA has much more freedom to discard uninformative views, whereas the DGCCA objective forces networks to minimize reconstruction error equally across all views. As noted in Benton et al. (2016), only the friend network view was useful for learning representations for friend recommendation (corroborated by performance of PCA applied to friend network view), so it is unsurprising that DGCCA when applied to all views cannot compete with WGCCA representations learned on the single useful friend network view⁸.

5 OTHER MULTIVIEW LEARNING WORK

There has been strong work outside of CCA-related methods to combine nonlinear representation and learning from multiple views. Kumar et al. (2011) elegantly outlines two main approaches these methods take to learn a joint representation from many views: either by 1) explicitly maximizing pairwise similarity/correlation between views or by 2) alternately optimizing a shared, “consensus” representation and view-specific transformations to maximize similarity. Models such as the siamese network proposed by Masci et al. (2014), fall in the former camp, minimizing the squared error between embeddings learned from each view, leading to a quadratic increase in the terms of the loss function size as the number of views increase. Rajendran et al. (2015) extend Correlational Neural Networks (Chandar et al., 2015) to many views and avoid this quadratic explosion in the loss function by only computing correlation between each view embedding and the embedding of a “pivot” view. Although this model may be appropriate for tasks such as multilingual image captioning, there are many datasets where there is no clear method of choosing a pivot view. The DGCCA objective does not suffer from this quadratic increase w.r.t. the number of views, nor does it require a privileged pivot view, since the shared representation is learned from the per-view representations.

Approaches that estimate a “consensus” representation, such as the multiview spectral clustering approach in Kumar et al. (2011), typically do so by an alternating optimization scheme which depends on a strong initialization to avoid bad local optima. The GCCA objective our work builds on is particularly attractive, since it admits a globally optimal solution for both the view-specific projections $U_1 \dots U_J$, and the shared representation G by singular value decomposition of a single matrix: a sum of the per-view projection matrices. Local optima arise in the DGCCA objective only because we are also learning nonlinear transformations of the input views. Nonlinear multiview methods often avoid learning these nonlinear transformations by assuming that a kernel or graph Laplacian (e.g. in multiview clustering) is given (Kumar et al., 2011; Xiaowen, 2014; Sharma et al., 2012).

6 CONCLUSION

We present DGCCA, a method for non-linear multiview representation learning from an arbitrary number of views. We show that DGCCA clearly outperforms prior work when using labels as a third view (Andrew et al., 2013; Arora & Livescu, 2014; Wang et al., 2015c), and can successfully exploit multiple views to learn user representations useful for downstream tasks such as hashtag recommendation for Twitter users. To date, CCA-style multiview learning techniques were either restricted to learning representations from no more than two views, or strictly linear transformations of the input views. This work overcomes these limitations.

⁸The performance of WGCCA suffers compared to PCA because whitening the friend network data ignores the fact that the spectrum of the decays quickly with a long tail – the first few principal components made up a large portion of the variance in the data, but it was also important to compare users based on other components.

REFERENCES

- Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.
- Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1247–1255, 2013.
- Raman Arora and Karen Livescu. Multi-view learning with supervision for transformed bottleneck features. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 2499–2503. IEEE, 2014.
- Adrian Benton, Raman Arora, and Mark Dredze. Learning multiview embeddings of twitter users. In *The 54th Annual Meeting of the Association for Computational Linguistics*, pp. 14, 2016.
- Sarath Chandar, Mitesh M. Khapra, Hugo Larochelle, and Balaraman Ravindran. Correlational neural networks. *CoRR*, abs/1504.07225, 2015. URL <http://arxiv.org/abs/1504.07225>.
- Thomas M Cover and Peter E Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.
- David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
- Paul Horst. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17(4), 1961.
- Harold Hotelling. Relations between two sets of variates. *Biometrika*, pp. 321–377, 1936.
- Jon R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.
- Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in neural information processing systems*, pp. 1413–1421, 2011.
- Jonathan Masci, Michael M Bronstein, Alexander M Bronstein, and Jürgen Schmidhuber. Multimodal similarity-preserving hashing. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):824–830, 2014.
- Kaare Petersen and Michael Pedersen. The matrix cookbook, Nov 2012. URL <http://www2.imm.dtu.dk/pubdb/p.php?3274>. Version 20121115.
- Janarthanan Rajendran, Mitesh M Khapra, Sarath Chandar, and Balaraman Ravindran. Bridge correlational neural networks for multilingual multimodal representation learning. *arXiv preprint arXiv:1510.03519*, 2015.
- Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2160–2167. IEEE, 2012.
- Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. In *Proceedings of COLT*, 2008.
- Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. Unsupervised learning of acoustic features via deep canonical correlation analysis. In *Proc. of the IEEE Int. Conf. Acoustics, Speech and Sig. Proc. (ICASSP’15)*, 2015a.
- Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proc. of the 32nd Int. Conf. Machine Learning (ICML 2015)*, 2015b.
- Weiran Wang, Raman Arora, Karen Livescu, and Nathan Srebro. Stochastic optimization for deep cca via nonlinear orthogonal iterations. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control and Computing (ALLERTON)*, 2015c.
- John R. Westbury. X-ray microbeam speech production database users handbook. In *Waisman Center on Mental Retardation & Human Development University of Wisconsin Madison, WI 53705-2280*, 1994.
- Dong Xiaowen. *Multi-View Signal Processing and Learning on Graphs*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2014.

APPENDIX A DERIVING THE GCCA OBJECTIVE GRADIENT

In order to train the neural networks in DGCCA, we need to compute the gradient of the GCCA objective with respect to any one of its input views. This gradient can then be backpropagated through the input networks to derive updates for the network weights.

Let N be the number of data points and J the number of views. Let $Y_j \in \mathbb{R}^{c_K^j \times N}$ be the data matrix representing the output of the j th neural network, i.e. $Y_j = f_j(X_j)$, where c_K^j is the number of neurons in the output layer of the j th network. Then, GCCA can be written as the following optimization problem, where r is the dimensionality of the learned auxiliary representation:

$$\begin{aligned} & \underset{U_j \in \mathbb{R}^{c_K^j \times r}, G \in \mathbb{R}^{r \times N}}{\text{minimize}} && \sum_{j=1}^J \|G - U_j^\top Y_j\|_F^2 \\ & \text{subject to} && GG^\top = I_r \end{aligned}$$

It can be shown that the solution is found by solving a certain eigenvalue problem. In particular, define $C_{jj} = Y_j Y_j^\top \in \mathbb{R}^{c_K^j \times c_K^j}$, $P_j = Y_j^\top C_{jj}^{-1} Y_j$ (note that P_j is symmetric and idempotent), and $M = \sum_{j=1}^J P_j$ (since each P_j is psd, so is M). Then the rows of G are the top r (orthonormal) eigenvectors of M , and $U_j = C_{jj}^{-1} Y_j G^\top$. Thus, at the minima of the objective, we can rewrite the reconstruction error as follows:

$$\begin{aligned} \sum_{j=1}^J \|G - U_j^\top Y_j\|_F^2 &= \sum_{j=1}^J \|G - G Y_j^\top C_{jj}^{-1} Y_j\|_F^2 \\ &= \sum_{j=1}^J \|G(I_N - P_j)\|_F^2 \\ &= \sum_{j=1}^J \text{Tr}[G(I_N - P_j)G^\top] \\ &= \sum_{j=1}^J \text{Tr}(I_r) - \text{Tr}(GMG^\top) \\ &= Jr - \text{Tr}(GMG^\top) \end{aligned}$$

Note that we can write the rank-1 decomposition of M as $\sum_{k=1}^N \lambda_k g_k g_k^\top$. Furthermore, since the k th row of G is g_k , and since the matrix product $Gg_k = \hat{e}_k$,

$$GMG^\top = \sum_{k=1}^N \lambda_k Gg_k (Gg_k)^\top = \sum_{k=1}^r \lambda_k \hat{e}_k \hat{e}_k^\top$$

But this is just an $N \times N$ diagonal matrix containing the top r eigenvalues of M , so we can write the GCCA objective as

$$Jr - \sum_{i=1}^r \lambda_i(M)$$

Thus, minimizing the GCCA objective (w.r.t. the weights of the neural nets) means maximizing the sum of eigenvalues $\sum_{i=1}^r \lambda_i(M)$, which we will henceforth denote by L .

Now, we will derive an expression for $\frac{\partial L}{\partial Y_j}$ for any view Y_j . First, by the chain rule, and using the fact that $\frac{\partial L}{\partial M} = G^\top G$ (Petersen & Pedersen, 2012),

$$\begin{aligned} \frac{\partial L}{\partial (Y_j)_{ab}} &= \sum_{c,d=1}^N \frac{\partial L}{\partial M_{cd}} \frac{\partial M_{cd}}{\partial (Y_j)_{ab}} \\ &= \sum_{c,d=1}^N (G^\top G)_{cd} \frac{\partial M_{cd}}{\partial (Y_j)_{ab}} \end{aligned}$$

Since $M = \sum_{j'=1}^J P_{j'}$, and since the only projection matrix that depends on Y_j is P_j , $\frac{\partial M}{\partial Y_j} = \frac{\partial P_j}{\partial Y_j}$.
 Since $P_j = Y_j^\top C_{jj}^{-1} Y_j$,

$$(P_j)_{cd} = \sum_{k,\ell=1}^{c_K^j} (Y_j)_{kc} (C_{jj}^{-1})_{k\ell} (Y_j)_{\ell d}$$

Thus, by the product rule,

$$\begin{aligned} \frac{\partial (P_j)_{cd}}{\partial (Y_j)_{ab}} &= \delta_{cb} \sum_{\ell=1}^{c_K^j} (Y_j)_{\ell d} (C_{jj}^{-1})_{a\ell} + \\ &\quad \delta_{db} \sum_{k=1}^{c_K^j} (Y_j)_{kc} (C_{jj}^{-1})_{ka} + \\ &\quad \sum_{k,\ell=1}^{c_K^j} (Y_j)_{kc} (Y_j)_{\ell d} \frac{\partial (C_{jj}^{-1})_{k\ell}}{\partial (Y_j)_{ab}} \\ &= \delta_{cb} (C_{jj}^{-1} Y_j)_{ad} + \delta_{db} (C_{jj}^{-1} Y_j)_{ac} \\ &\quad + \sum_{k,\ell=1}^{c_K^j} (Y_j)_{kc} (Y_j)_{\ell d} \frac{\partial (C_{jj}^{-1})_{k\ell}}{\partial (Y_j)_{ab}} \end{aligned}$$

The derivative in the last term can also be computed using the chain rule:

$$\begin{aligned} \frac{\partial (C_{jj}^{-1})_{k\ell}}{\partial (Y_j)_{ab}} &= \sum_{m,n=1}^N \frac{\partial (C_{jj}^{-1})_{k\ell}}{\partial (C_{jj})_{mn}} \frac{\partial (C_{jj})_{mn}}{\partial (Y_j)_{ab}} \\ &= - \sum_{m,n=1}^N \left\{ (C_{jj}^{-1})_{km} (C_{jj}^{-1})_{n\ell} \right. \\ &\quad \left. [\delta_{am} (Y_j)_{nb} + \delta_{an} (Y_j)_{mb}] \right\} \\ &= - \sum_{n=1}^N (C_{jj}^{-1})_{ka} (C_{jj}^{-1})_{n\ell} (Y_j)_{nb} \\ &\quad - \sum_{m=1}^N (C_{jj}^{-1})_{km} (C_{jj}^{-1})_{a\ell} (Y_j)_{mb} \\ &= -(C_{jj}^{-1})_{ka} (C_{jj}^{-1} Y_j)_{\ell b} \\ &\quad - (C_{jj}^{-1})_{a\ell} (C_{jj}^{-1} Y_j)_{kb} \end{aligned}$$

Substituting this into the expression for $\frac{\partial (P_j)_{cd}}{\partial (Y_j)_{ab}}$ and simplifying matrix products, we find that

$$\begin{aligned} \frac{\partial (P_j)_{cd}}{\partial (Y_j)_{ab}} &= \delta_{cb} (C_{jj}^{-1} Y_j)_{ad} + \delta_{db} (C_{jj}^{-1} Y_j)_{ac} \\ &\quad - (C_{jj}^{-1} Y_j)_{ac} (Y_j^\top C_{jj}^{-1} Y_j)_{bd} \\ &\quad - (C_{jj}^{-1} Y_j)_{ad} (Y_j^\top C_{jj}^{-1} Y_j)_{bc} \\ &= (I_N - P_j)_{cb} (C_{jj}^{-1} Y_j)_{ad} + \\ &\quad (I_N - P_j)_{db} (C_{jj}^{-1} Y_j)_{ac} \end{aligned}$$

Finally, substituting this into our expression for $\frac{\partial L}{\partial(Y_j)_{ab}}$, we find that

$$\begin{aligned}\frac{\partial L}{\partial(Y_j)_{ab}} &= \sum_{c,d=1}^N (G^\top G)_{cd}(I_N - P_j)_{cb}(C_{jj}^{-1}Y_j)_{ad} \\ &\quad + \sum_{c,d=1}^N (G^\top G)_{cd}(I_N - P_j)_{db}(C_{jj}^{-1}Y_j)_{ac} \\ &= 2[C_{jj}^{-1}Y_j G^\top G(I_N - P_j)]_{ab}\end{aligned}$$

Therefore,

$$\frac{\partial L}{\partial Y_j} = 2C_{jj}^{-1}Y_j G^\top G(I_N - P_j)$$

But recall that $U_j = C_{jj}^{-1}Y_j G^\top$. Using this, the gradient simplifies as follows:

$$\frac{\partial L}{\partial Y_j} = 2U_j G - 2U_j U_j^\top Y_j$$

Thus, the gradient is the difference between the r -dimensional auxiliary representation G embedded into the subspace spanned by the columns of U_j (the first term) and the projection of the network outputs in $Y_j = f_j(X_j)$ onto said subspace (the second term). Intuitively, if the auxiliary representation G is far away from the view-specific representation $U_j^\top f_j(X_j)$, then the network weights should receive a large update.

APPENDIX B DGCCA OPTIMIZATION PSEUDOCODE

Algorithm 1 contains the pseudocode for the DGCCA optimization algorithm. In practice we use stochastic optimization with minibatches, following Wang et al. (2015c).

Algorithm 1 Deep Generalized CCA

Input: multiview data: X_1, X_2, \dots, X_J ,
number of iterations T , learning rate η

Output: O_1, O_2, \dots, O_J

Initialize weights W_1, W_2, \dots, W_J

for iteration $t = 1, 2, \dots, T$ **do**

for each view $j = 1, 2, \dots, J$ **do**

$O_j \leftarrow$ forward pass of X_j with weights W_j

 mean-center O_j

end for

$U_1, \dots, U_J, G \leftarrow$ gccca(O_1, \dots, O_J)

for each view $j = 1, 2, \dots, J$ **do**

$\partial F / \partial O_j \leftarrow U_j U_j^\top O_j - U_j G$

$\nabla W_j \leftarrow$ backprop($\partial F / \partial O_j, W_j$)

$W_j \leftarrow W_j - \eta \nabla W_j$

end for

end for

for each view $j = 1, 2, \dots, J$ **do**

$O_j \leftarrow$ forward pass of X_j with weights W_j

 mean-center O_j

end for

$U_1, \dots, U_J, G \leftarrow$ gccca(O_1, \dots, O_J)

for each view $j = 1 \dots J$ **do**

$O_j \leftarrow U_j^\top O_j$

end for

APPENDIX C RECONSTRUCTION ERROR AND DOWNSTREAM PERFORMANCE

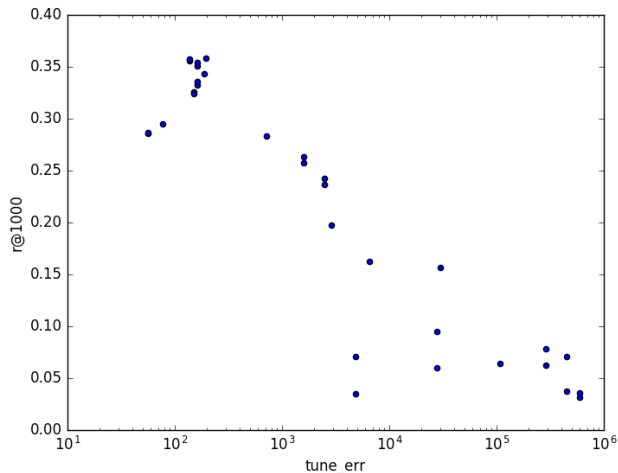


Figure 6: Tuning reconstruction error against Recall at 1000 for the hashtag prediction task. Each point corresponds to a different setting of hyperparameters.

CCA methods are typically evaluated intrinsically by the amount of correlation captured, or reconstruction error. These measures are dependent on the width of the shared embeddings and view-specific output layers, and do not necessarily predict downstream performance. Although reconstruction error cannot solely be relied on for model selection for a downstream task, we found that it was a useful as a signal to weed out very poor models. Figure 6 shows the reconstruction error against hashtag prediction Recall at 1000 for an initial grid search of DGCCA hyperparameters. Models with tuning reconstruction error greater than 10^3 can safely be ignored, while there is some variability in the performance of models with achieving lower error.

Since a DGCCA model with high reconstruction error suggests that the views do not agree with each other at all, it makes sense that the shared embedding will likely be noisy, whereas a relatively lowly reconstruction error suggests that the transformed views have converged to a stable solution.