RADEMACHER COMPLEXITY REGULARIZATION FOR CORRELATION-BASED MULTIVIEW REPRESENTATION LEARNING

Maurice Kuschel[⋆] *Tanuj Hasija*[⋆] *Timothy Marrinan*[†]

[⋆]Signal and System Theory Group, Paderborn University, Germany
[†] Pacific Northwest National Lab, Seattle, United States

ABSTRACT

Deep correlation-based multiview representation learning techniques have become increasingly popular methods for extracting highly correlated representations from multiview data. However, their ability to find highly complex mappings between the views can also lead to overfitting and overly correlated representations. In this work, we propose a regularizer for this specific problem, based on the Rademacher complexity of the DNNs, tailored for multiview correlation maximization. We demonstrate that the proposed regularization leads to less noisy representations in synthetic data and improved performance of downstream tasks in real-world multiview datasets.

Index Terms— Neural network complexity, deep canonical correlation analysis, overfitting, unsupervised learning

1. INTRODUCTION

Canonical correlation analysis (CCA) and its extensions [1, 2] are widely used for multiview representation learning (MRL) in many applications [3, 4, 5] for distilling meaningful low-dimensional representations from high-dimensional data by exploiting the joint information contained in multiple views of an object or phenomenon [6]. This popularity is due to the fact that CCA provides a simple yet effective solution for learning shared or correlated representations from two views while discarding the information present in only one of the views. However, the simplicity comes with a cost. Since CCA mappings are linear, it is not effective in applications where the views are connected by nonlinear relationships [7].

CCA can be generalized using kernel methods [8, 9] or deep neural networks (DNNs) [10]. DNN-based methods like deep CCA (DCCA) have the advantage of approximating nonlinear functions in a data-driven manner and are not restricted to a particular class of nonlinearities, compared to kernel methods. However, relaxing the linear model introduces new challenges, such as the possibility of identifying non-existent relationships between the views. The more complex a mapping is allowed to be, the more prone it is to overfitting [11, 12]. Some further extensions of DCCA address this issue indirectly by extending the networks to autoencoders [11] and restricting the class of allowable mappings to make representations identifiable [13]. However, neither of these methods explicitly prevents learning overly complex mappings to maximize correlated representations, especially when the two views are heterogeneous [12].

In this work, we propose a regularization term for DNNs used in deep correlation maximization using an upper bound on the Rademacher complexity (RC) of those networks. By definition, RC measures the ability of a network to fit random noise, and it is the fitting of this noise that allows the networks to learn subspaces that are overly correlated [14]. Inspired by the upper bound for the RC of a DNN for a supervised task [15], we propose a regularization term for the given unsupervised multiview problem which combines the RC upper bounds of multiple networks and penalizes the K largest weights of each layer in a neural network. We demonstrate its advantages with both synthetic and real-world multiview data.

2. BACKGROUND AND RELATED WORK

Let $\mathbf{x}(m) \in \mathbb{R}^{N_x}$ and $\mathbf{y}(m) \in \mathbb{R}^{N_y}$ be an observation from View 1 and View 2, respectively, where N_x and N_y denote the dimensions of the observations. The M paired observations from each view form the columns of data matrices, $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(M)]$ and $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(M)]$. Suppose that a nonlinear mapping of a fully connected neural network, $\mathbf{f}$, consists of L fully connected layers. If the l-th layer has $J^{(l)}$ neurons, then each layer will have an associated set of weights, $\mathbf{W}^{(l)} \in \mathbb{R}^{J^{(l-1)} \times J^{(l)}}$. We denote the full set of weights by $\mathcal{W} = \{\mathbf{W}^{(1)}, ..., \mathbf{W}^{(L)}\}$.

2.1. Correlation analysis-based MRL techniques

2.1.1. Deep canonical correlation analysis (DCCA)

Given data matrices $\mathbf{X}$ and $\mathbf{Y}$, the goal of DCCA is to learn associated linear mappings, $\mathbf{A}_x \in \mathbb{R}^{D_x \times D_x}$ and $\mathbf{A}_y \in \mathbb{R}^{D_y \times D_y}$, with D_x and D_y denoting the dimensions of the latent space, as well as nonlinear mappings, $\mathbf{f}_x : \mathbb{R}^{N_x} \to \mathbb{R}^{D_x}$ and $\mathbf{f}_y : \mathbb{R}^{N_y} \to \mathbb{R}^{D_y}$, without supervision such that the correlation

T.M. was partially supported by Mathematics for Artificial Reasoning in Science (MARS) initiative via the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL).

between the mapped representations $\mathbf{Z}_x = \mathbf{A}_x \mathbf{f}_x(\mathbf{X})$, $\mathbf{Z}_y = \mathbf{A}_y \mathbf{f}_y(\mathbf{Y})$ is maximized. DCCA then solves the problem

$$\underset{\mathbf{A}_x, \mathbf{A}_y, \mathbf{f}_x, \mathbf{f}_y}{\arg\max} \quad \text{trace}\left(\frac{1}{M} \mathbf{A}_x \mathbf{f}_x(\mathbf{X}) \mathbf{f}_y(\mathbf{Y})^T \mathbf{A}_y^T \right) \tag{1}$$

$$\text{such that} \quad \frac{1}{M} \mathbf{A}_x \mathbf{f}_x(\mathbf{X}) \mathbf{f}_x(\mathbf{X})^T \mathbf{A}_x^T = \mathbf{I}, \tag{2}$$

$$\frac{1}{M} \mathbf{A}_y \mathbf{f}_y(\mathbf{Y}) \mathbf{f}_y(\mathbf{Y})^T \mathbf{A}_y^T = \mathbf{I}, \tag{3}$$

$$\frac{1}{M} \mathbf{A}_x \mathbf{f}_x(\mathbf{X}) \mathbf{f}_y(\mathbf{Y})^T \mathbf{A}_y^T \text{ is diagonal.} \tag{4}$$

In DCCA, $\mathbf{f}_x$ and $\mathbf{f}_y$ are designed using two separate neural networks [10], and Problem (1) is typically solved iteratively using either full-batch optimization algorithms [10] or stochastic gradient descent with mini-batches [16].

2.1.2. Canonical correlation analysis (CCA)

If $\mathbf{f}_x$ and $\mathbf{f}_y$ are both restricted to be identity functions, (1) to (4) describe the problem of CCA, which finds linear mappings, $\mathbf{A}_x$ and $\mathbf{A}_y$, such that the correlation between the representations $\mathbf{Z}_x = \mathbf{A}_x \mathbf{X}$ and $\mathbf{Z}_y = \mathbf{A}_y \mathbf{Y}$ is maximized. It is well-known that this can be solved algebraically via the singular value decomposition (SVD) [1].

2.1.3. Deep canonical correlated autoencoders (DCCAE)

What distinguishes DCCAE from prior works is the inclusion of an autoencoder regularization [11]. Let $\mathbf{g}_x : \mathbb{R}^{D_x} \to \mathbb{R}^{N_x}$ and $\mathbf{g}_y : \mathbb{R}^{D_y} \to \mathbb{R}^{N_y}$ denote decoder functions to be learned by the network. The autoencoder regularization term for each view is the mean-squared reconstruction error between the input and the decoded outputs. The term for View 1 is

$$\text{Reg}_{AE}(\mathbf{X}, \mathbf{f}_x, \mathbf{g}_x) = \lambda \tfrac{1}{M} \textstyle\sum_{m=1}^{M} \left(\mathbf{x}(m) - \mathbf{g}_x\Big(\mathbf{f}_x\big(\mathbf{x}(m)\big)\Big) \right)^2, \tag{5}$$

where λ is a hyper-parameter and the term for View 2 is defined analogously. The DCCAE objective is the DCCA objective in (1) minus the autoencoder regularization for each view.

2.1.4. Identifiable DCCAE (IDCCAE)

DCCA and DCCAE are data-driven frameworks that do not provide theoretical guarantees for recovering the shared components in general. However, [13] was able to show that the solution of DCCAE is identifiable, i.e. we can provably recover the range space of the shared data sources up to an affine shift, if some weak assumptions about the data-generating model are satisfied. The proof requires the data to be generated by a Post-Nonlinear Mixing (PNL) model, the non-linearities to be applied channel-wise, and the shared data sources to be exactly common. Further, [13] also assumes that the learned functions $\mathbf{f}_x$, $\mathbf{f}_y$, $\mathbf{g}_x$, and $\mathbf{g}_y$ must be channelwise nonlinear mappings.

2.2. Empirical Rademacher Complexity (RC)

RC reports the complexity of a class of functions by measuring its ability to fit random noise [14]. The better a class can match random noise, the more complex it is. Let $\mathbf{\Gamma}$ be a sample space and $\mathcal{M} = \{\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_S\}$ a set of drawn samples. Further, let $\mathcal{H}$ be a set of real-valued functions, such that $\mathbf{h} \in \mathcal{H} : \mathbf{\Gamma} \mapsto \mathbb{R}$ and let $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_S)$ be i.i.d. random variables, uniformly chosen from $\{-1, 1\}^S$. Then, the empirical RC of $\mathcal{H}$ with respect to $\mathcal{M}$ is defined as [17]

$$R_{\mathcal{M}}(\mathcal{H}) := \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\mathbf{h} \in \mathcal{H}} \tfrac{1}{S} \textstyle\sum_{s=1}^{S} \sigma_s \mathbf{h}(\boldsymbol{\gamma}_s) \right]. \tag{6}$$

For a given set of samples $\mathcal{M}$ and Rademacher vector $\boldsymbol{\sigma}$, the supremum measures the maximum correlation between σ_s and $\mathbf{h}(\boldsymbol{\gamma}_s)$ over all $\mathbf{h} \in \mathcal{H}$. When $\mathcal{H}$ is rich enough, it contains functions that can properly resemble all combinations of -1 and 1. By applying the expectation operator to $\boldsymbol{\sigma}$ we average the ability of the function class $\mathcal{H}$ to fit random samples over all possible sample combinations.

2.2.1. Rademacher Complexity Bound

The empirical RC in (6) is defined as the supremum across an abstract class of functions, making it difficult to compute for a specific neural network instance. However, a useful upper bound on (6) is derived in [15], which can be computed for a given neural network with respect to the input. Based on that upper bound, the authors propose a regularization term for supervised classification with c classes, where the neural network employs Bernoulli dropout with a retain-rate of θ. The regularization term proposed by [15] is

$$\text{Reg}_{\text{RCDrop}}(\mathbf{X}, \mathcal{W}_{\mathbf{f}_x}) = \left(\textstyle\prod_{l=1}^{L} \|\theta^{(l-1)}\|_1^{1/q} \max_j \|\mathbf{W}_{\mathbf{f}_x}^{(l)}(:, j)\|_p\right) \cdot T_{\mathbf{X}}, \tag{7}$$

where $\mathbf{W}_{\mathbf{f}_x}^{(l)}(:, j)$ denotes the jth column of $\mathbf{W}_{\mathbf{f}_x}^{(l)}$, $p \geq 1$ meets $1/p + 1/q = 1$ and $T_{\mathbf{X}} = c 2^L \sqrt{\frac{\log N_x}{M}} \|\mathbf{X}\|_{\max}$ is constant with $\|\mathbf{X}\|_{\max} = \max_{i,j} |\mathbf{X}(i, j)|$. If we choose $p = \infty$ and $q = 1$, as in [15], we have $\max_j \|\mathbf{W}_{\mathbf{f}_x}^{(l)}(:, j)\|_p = \|\mathbf{W}_{\mathbf{f}_x}^{(l)}\|_{\max}$.

3. RADEMACHER COMPLEXITY FOR DCCA

Training neural networks to maximize correlation without any restriction on the class of allowable nonlinear mappings can be a blessing and a curse in MRL. On one hand, unconstrained networks have greater expressivity, but they may also learn overly complex mappings to extract perfectly correlated representations from the data, even when such representations are spurious and do not match the true correlations of the underlying signals [12, 13]. One natural strategy to discourage DCCA-based methods from learning representations that are overly correlated is to promote models that are as simple as possible, but as complex as necessary. In this vein, we introduce a multiview-specific regularization term that penalizes

the complexity of the mapping during training. Recall that RC measures the ability of a network to fit random noise – the proposed regularization penalizes this complexity explicitly. The inherent assumption is that if a minor gain in correlation leads to a disproportionate increase in network complexity, it may be an indication of overfitting, which is undesirable.

3.1. MV Regularization With Rademacher Complexity

The regularization term introduced in [15] is not directly applicable to MRL, so we propose a modification. First, MRL is an unsupervised task so the number of classes in the constant term $T_{\mathbf{X}}$ can be neglected because $c = 1$. Second, dropout is a common way to prevent overfitting [18], but it acts quite differently from norm-based regularizers by randomly deactivating neurons of a layer during each training step. This type of architecture-affecting regularization is complementary to all norm-based techniques. Therefore, in this work, the retain-rate is set to $\theta = 1$ to remove the impact of dropout and focus on the effect of regularizing the Rademacher complexity directly. As a result, (7) can be rewritten as

$$\text{Reg}_{\text{RC}}(\mathbf{X}, \mathcal{W}_{\mathbf{f}_x}) = \left(\prod_{l=1}^{L} \|\mathbf{W}_{\mathbf{f}_x}^{(l)}\|_{\max}\right) \cdot T_{\mathbf{X}}, \tag{8}$$

with $T_{\mathbf{X}} = 2^L \sqrt{\frac{\log N_x}{M}} \|\mathbf{X}\|_{\max}$, and can be interpreted as the product of the largest weight of each layer.

Regularizing just one weight from each layer of a network can be ineffective for layers with hundreds of neurons because very few neurons will be affected during training. We extend (8) by regularizing K elements from each layer, with K as a hyperparameter. To choose the K elements to regularize in layer l, we first find the largest magnitude element in each column of $\mathbf{W}_{\mathbf{f}_x}^{(l)}$ and choose the K largest from that set. To ensure that the regularizer is still an upper bound for the RC of the neural network, we take the sum of those K largest elements. Let $\|\mathbf{W}_{\mathbf{f}_x}^{(l)}\|_{\text{top-}K}$ denote the sum of the K largest magnitude elements from the columns of $\mathbf{W}_{\mathbf{f}_x}^{(l)}$. The regularization term for a single view of the multiview network is

$$\text{Reg}_{\text{RCTK}}(\mathbf{X}, \mathcal{W}_{\mathbf{f}_x}) = \left(\prod_{l=1}^{L} \|\mathbf{W}_{\mathbf{f}_x}^{(l)}\|_{\text{top-}K}\right) \cdot T_{\mathbf{X}} \tag{9}$$

where $T_{\mathbf{X}}$ is defined as before. In MRL, a network is trained for each view and requires its own regularization term. The proposed multiview RC regularization term is written as the product of all the terms for each view,

$$\text{Reg}_{\text{RCMV}}(\mathbf{X}, \mathbf{Y}, \mathcal{W}_{\mathbf{f}_x}, \mathcal{W}_{\mathbf{f}_y}) = \text{Reg}_{\text{RCTK}}(\mathbf{X}, \mathcal{W}_{\mathbf{f}_x}) \cdot \text{Reg}_{\text{RCTK}}(\mathbf{Y}, \mathcal{W}_{\mathbf{f}_y}). \tag{10}$$

4. EXPERIMENTAL RESULTS

In this section, we show that including the proposed regularization term in (10) improves the performance of existing DCCA-based techniques on synthetic and real-world datasets. The prefix "Core" will be used to denote a method that uses the 'complexity-regularization' from (10) in addition to the baseline regularization.

ρ	CCA	DCCA	DCCAE	CoreDCCAE	IDCCAE	CoreIDCCAE
0.9	0.342	0.243	0.319	0.287	0.009	0.012
0.7	0.353	0.676	0.443	0.321	0.021	0.029
0.5	0.410	0.749	0.456	0.386	0.065	0.044
0.3	0.579	0.865	0.655	0.489	0.311	0.251

Table 1: Average distance between ground truth subspaces and estimated representations for all techniques along with the proposed CoreDCCAE and CoreIDCCAE.

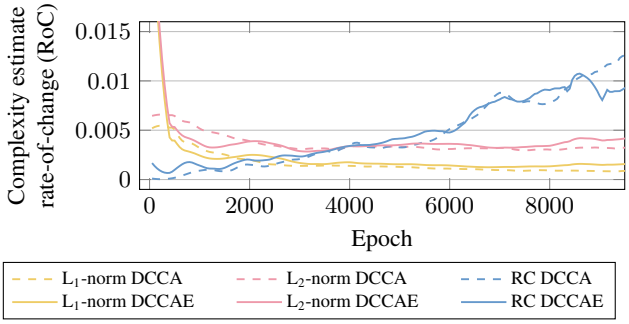

Fig. 1: Comparison of complexity estimates during training of unregularized DCCA and DCCAE. Measurements were normalized to be in $[0, 1]$ and the gradient w.r.t the course of training is shown.

4.1. Synthetic Data

For the synthetic experiment, we generate the data from an identifiable model[1] [13]. Since these subspaces are known, we can assess the accuracy of the learned representations by measuring the distance between the ground truth and estimated subspaces as 1 minus the cosine of the largest principal angle separating the subspaces [19]. The reported number for each method is the distance to the ground truth averaged across both views. Table 1 shows the performance of the techniques for different values of correlation coefficient ρ. We expect all methods to perform best (lowest average distance) when ρ is high because the two views share more information. When the ρ is low we expect to see improvements from including the proposed complexity regularization because the unregularized methods will tend to overfit. Of the unregularized methods, IDCCAE produces the lowest average distances to the ground truth. This is expected because the model assumptions of IDCCAE are satisfied by our synthetic data. Both of the 'Core' methods show significant improvements over their unregularized counterparts when $\rho = 0.3$ and $\rho = 0.5$, and CoreDCCAE outperforms DCCAE for all ρ values [2]. For smaller values of ρ, traditional techniques overfit more and the benefit of including the regularization is more pronounced.

4.2. Overfitting-Sensitivity of Complexity Estimates

To motivate the use of the RC in the proposed regularization term, we compare the sensitivity to overfitting of (9) with the L_1-norm and L_2-norm. We train (unregularized) networks for

[1] Code at: github.com/SSTGroup/RademacherRegCorrelationMRL

[2] We omit regularization results with DCCA since its autoencoder extensions have demonstrated comparable or superior performance, consistent with our observations.

DCCA and DCCAE using the before-mentioned synthetic data with known ground truth. For each network, we compute the estimates of complexity from the network parameters at each epoch of training. The estimates are independently normalized to the range $[0, 1]$. Fig. 1 plots the rate of change (RoC) of these complexity estimates as a function of the training epoch.

As we know the true subspace for synthetic data, we can estimate where overfitting starts, as the epoch when the distance to the ground truth is minimized and then begins to increase. All methods start overfitting at around the 500th epoch. Thus, we expect to see the complexity estimates increase more rapidly as the networks begin to fit the noise. The RoC for the L_1 and L_2-norms converge to small values, indicating that they do not increase at a greater rate due to overfitting. In contrast, the RoC for the RC bound increases throughout training, consistent with the process of learning complex representations. This indicates that the proposed RC regularization is more sensitive in this context.

For the remainder, L_2-norm regularization is used as the baseline for comparison because the L_1-norm is less sensitive to the investigated overfitting. We also omit IDCCAE in real data experiments due to the substantial computational burden of creating channel-wise networks for each input dimension.

4.3. Occluded Multiview MNIST

In this investigation, we use the well-known MNIST dataset [20] consisting of 28×28 pixel gray-scale images of handwritten digits. We construct a synthetic two-view version of the MNIST dataset by building image pairs of the same class [11]. We expect the CCA-based techniques to discover learned representations that represent the digit present in both images. To design views that are not perfectly correlated, we modify the two MNIST views by adding spatter augmentation [21] and adding a variable number of white boxes on top. We follow the preprocessing steps proposed in [12].

To evaluate the learned representations we cluster the projected views into ten clusters each using spectral clustering [22]. We use the tuning set to select the hyperparameters of all competitors[3]. For early stopping, we monitor the smoothed clustering accuracy of the first view and select the configuration with the best performance. We report the performance on test data, averaged over five runs. Table 2 shows the clustering accuracy for raw data, the single-view approach using principal component analysis (PCA), and the multiview techniques.

We expect the boxes to impair the spectral clustering accuracy, as the noise can be falsely taken as relevant for correlation by the encoder. This expectation is confirmed by the results, as the accuracy degrades when more boxes are present in the images. When comparing the results we can see that DCCA and DCCAE perform similarly, with a slight edge for DCCAE. Depending on the number of boxes, CoreDCCAE performs on par or better than the other methods. It is striking that the gap between CoreDCCAE and DCCAE gets larger as more noise, i.e., the number of boxes, is present in the images. This supports our hypothesis that the other methods overfit significantly despite being regularized by L_2-norm, and that penalizing the RC bound benefits the clustering accuracy.

[3]For details on the fully connected neural networks and hyperparameters, see the provided code.

#	Raw	PCA	CCA	DCCA	DCCAE	CoreDCCAE
0	70.2%	70.7%	84.8%	95.6%	**96.7%**	96.6%
1	50.3%	51.4%	71.2%	86.9%	87.0%	**88.7%**
2	34.5%	38.1%	66.0%	76.7%	77.0%	**81.1%**

Table 2: Spectral clustering accuracy comparison of the different methods for the occluded multiview MNIST datasets.

Raw	PCA	CCA	DCCA	DCCAE	CoreDCCAE
43.0%	42.9%	42.9%	57.9%	58.0%	**59.8%**

Table 3: SVM accuracy comparison for the XRMB dataset.

4.4. X-Ray Microbeam (XRMB) Database

Now we investigate the Wisconsin X-Ray Microbeam (XRMB) database [23]. It consists of acoustic speech recordings and concurrently recorded articulatory measurements. For further details of the dataset see [23]. Previous work has shown that this second view improves phonetic recognition performance compared to audio-only approaches [24]. In contrast to the previous datasets, these two views are more complementary and heterogeneous, as we combine audio and image data. Thus, the shared information between the two views may not be identical and, therefore, is more prone to overfitting.

We again use the tuning set to select the hyperparameters of all competitors. We evaluate the learned representations via linear support vector classification [25]. For Rademacher regularization, we considered the $K = 20$ largest values. Results on the test data are shown in Table 3. The effect seen in the previous experiment persists for this dataset, and CoreDCCAE outperforms both DCCA and DCCAE, reinforcing the idea that complex, overly correlated representations are less meaningful for real, heterogeneous data, and the proposed regularizer improves the classification performance.

5. CONCLUSION

State-of-the-art correlation-based MRL methods are designed to discover complex nonlinear mappings, which can lead to overfitting, especially when dealing with heterogeneous views. We proposed a multiview regularization term, based on Rademacher complexity, that penalizes overly complex mappings. Synthetic experiments show that comparable norm-based regularizers are relatively insensitive to this type of overfitting. Further, experiments on real-world multiview data demonstrate the occurrence of overfitting when data is truly heterogeneous and that the methods produce more meaningful representations when using the proposed regularizer. An in-depth analysis regarding the effect of K and the performance in combination with dropout is left for future work.

6. REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.

[2] J. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[3] W. Shao, S. Xiang, Z. Zhang, K. Huang, and J. Zhang, "Hyper-graph based sparse canonical correlation analysis for the diagnosis of alzheimer's disease from multi-dimensional genomic data," *Methods*, vol. 189, pp. 86–94, 2021.

[4] S. Vieluf, T. Hasija, M. Kuschel, C. Reinsberger, and T. Loddenkemper, "Developing a deep canonical correlation-based technique for seizure prediction," *Expert Systems with Applications*, vol. 234, pp. 120986, 2023.

[5] T. Hasija, M. Gölz, M. Muma, P. J. Schreier, and A. M. Zoubir, "Source enumeration and robust voice activity detection in wireless acoustic sensor networks," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 1257–1261.

[6] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[7] K. Friston, J. Phillips, D. Chawla, and C. Buchel, "Nonlinear PCA: characterizing interactions between modes of brain activity," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 355, no. 1393, pp. 135–146, 2000.

[8] D. R. 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.

[9] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 05, pp. 365–377, 2000.

[10] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[11] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*. PMLR, 2015, pp. 1083–1092.

[12] M. Kuschel, T. Marrinan, and T. Hasija, "Geodesic-based relaxation for deep canonical correlation analysis," in *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, 2023, pp. 1–6.

[13] Q. Lyu and X. Fu, "Nonlinear multiview analysis: Identifiability and neural network-assisted implementation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2697–2712, 2020.

[14] M. M. Wolf, "Mathematical foundations of supervised learning," *Lecture notes from Technical University of Munich*, 2018.

[15] K. Zhai and H. Wang, "Adaptive dropout with Rademacher complexity regularization," in *ICLR*, 2018.

[16] W. Wang, R. Arora, K. Livescu, and N. Srebro, "Stochastic optimization for deep CCA via nonlinear orthogonal iterations," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2015, pp. 688–695.

[17] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*, Cambridge university press, 2014.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[19] Å. Björck and G. H. Golub, "Numerical methods for computing angles between linear subspaces," *Mathematics of computation*, vol. 27, no. 123, pp. 579–594, 1973.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[21] N. Mu and J. Gilmer, "MNIST-C: A robustness benchmark for computer vision," *arXiv preprint arXiv:1906.02337*, 2019.

[22] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 14, 2001.

[23] J. R. Westbury, G. Turner, and J. Dembowski, "X-ray microbeam speech production database user's handbook," *University of Wisconsin*, 1994.

[24] R. Arora and K. Livescu, "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7135–7139.

[25] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.