Deep Probabilistic Canonical Correlation Analysis

Mahdi Karami, Dale Schuurmans
Department of Computer Science
University of Alberta
Edmonton, Alberta, Canada
{karami1, daes}@ualberta.ca

Abstract

We propose a deep generative framework for multi-view learning based on a probabilistic interpretation of canonical correlation analysis (CCA). The model combines a linear multi-view layer in the latent space with deep generative networks as observation models, to decompose the variability in multiple views into a shared latent representation that describes the common underlying sources of variation and a set of view-specific components. To approximate the posterior distribution of the latent multi-view layer, an efficient variational inference procedure is developed based on the solution of probabilistic CCA. The model is then generalized to an arbitrary number of views. An empirical analysis confirms that the proposed deep multi-view model can discover subtle relationships between multiple views and recover rich representations.

Introduction

When a dataset consists of multiple co-occurring groups of observations, views or modalities from the same underlying source of variation, a learning algorithm should leverage the complementary information to alleviate learning difficulty (Chaudhuri et al. 2009) and improve accuracy. A well-established method for two-view analysis is given by canonical correlation analysis (CCA) (Hotelling 1992), a classical subspace learning technique that extracts the common information between two multivariate random variables by projecting them onto a subspace. Due to its ease of interpretation and closed-form solution, CCA has become a standard model for unsupervised two-view learning (Klami, Virtanen, and Kaski 2013) which has been used in a broad range of tasks such as dimensionality reduction, visualization and time series analysis (Xia et al. 2014).

The goal of representation learning is to capture the underlying essence of data and extract natural features; for example, to reveal implicit categories or cluster memberships. In multi-view data the relationship between different views should also be leveraged to enhance feature extraction. Additionally, given that representation learning in real-world applications poses significant challenges, where data is typically high-dimensional with complex structure, it is necessary to exploit expressive yet scalable models such as deep generative neural networks.

It has been shown in (Chaudhuri et al. 2009) that projecting multi-view data onto low-dimensional subspaces using CCA allows cluster memberships to be more easily recovered under a weak separation condition. Nevertheless, CCA exhibits poor generalization when trained on small training sets, hence (Klami and Kaski 2007; Klami, Virtanen, and Kaski 2013; Mukuta et al. 2014) adopt a Bayesian approach to solve a probabilistic interpretation of CCA. However, real applications involve nonlinear subspaces where more than two views are available. Recently, deep learning has received renewed interest as an effective approach for recovering expressive models of complex datasets. For multi-view learning, several deep learning based approaches have been successfully extended (Ngiam et al. 2011; Andrew et al. 2013; Wang, Livescu, and Bilmes 2015). Subsequently (Wang et al. 2016; Tang, Wang, and Livescu 2017) have proposed VCCA and VCCA-private, which are deep two-view autoencoders that integrate a generative two-view model over a shared representation with a generative model that combines shared plus view-specific factors. However, these methods adopt a black box variational inference approach that does not exploit the probabilistic CCA formulation established in (Bach and Jordan 2005). These methods are also primarily targeted to the two-view setting, and do not provide a simple extension to the generic multi-modal setting (with an arbitrary number of views) where all modalities are available at test time.

In this work, we first develop a modified formulation of probabilistic CCA, then show how such a linear probabilistic layer can be extended to a deep generative multi-view network. The proposed model captures data variation by a *shared latent representation* that isolates the common underlying sources of variation (i.e. the essence of multi-view data) while combining this with a set of *view-specific latent factors* to obtain an interpretable latent representation. Importantly, the model can be naturally extended to an arbitrary number of views. We design the learning algorithm using a variational inference method, which is known to be a powerful tool for scaling probabilistic models to complex problems and large datasets (Rezende, Mohamed, and Wierstra 2014). In contrast to VCCA and VCCA-private (Wang et al. 2016; Tang, Wang, and Livescu 2017), the proposed variational inference is grounded in the probabilistic CCA formulation, yielding a more principled and expressive multi-view approach. Furthermore, the learning algorithm offers a flexible data fusion

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

method in the latent space, which makes it appropriate for modeling general multi-modal datasets. Empirical studies confirm that the proposed deep generative multi-view model can efficiently integrate multiple views to alleviate learning difficulty in different downstream tasks.

Probabilistic CCA

Canonical correlation analysis (CCA) (Hotelling 1992) is a classical subspace learning method that extracts information from the cross-correlation between two variables.[1] Let $\mathbf{z}_1 \in \mathbb{R}^{d_1}$ and $\mathbf{z}_2 \in \mathbb{R}^{d_2}$ be a pair of random vectors corresponding to two different views with their mean and covariance matrices denoted as $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, respectively, and their cross-covariance denoted $\boldsymbol{\Sigma}_{12}$. CCA linearly projects these onto the subspace $\mathbb{R}^{d_0}$ as $\mathbf{r}_1 = \boldsymbol{U}_1^\top \mathbf{z}_1$ and $\mathbf{r}_2 = \boldsymbol{U}_2^\top \mathbf{z}_2$, where matrices $\boldsymbol{U}_1 \in \mathbb{R}^{d_1 \times d_0}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{d_2 \times d_0}$ are composed of first d_0 *canonical pairs of directions* vectors, $(\boldsymbol{u}_{1i}, \boldsymbol{u}_{2i})$, and $0 < d_0 \leq \min\{d_1, d_2\}$. This projection is such that each pair of components $(\mathbf{r}_1(i), \mathbf{r}_2(j))$ are maximally correlated if $i = j$, with correlation coefficient p_i, and uncorrelated otherwise, hence forming a diagonal *matrix of canonical correlations* $\boldsymbol{P}_{d_0} = \text{diag}([p_0, ..., p_{d_0}])$. The optimal solution for $\{\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{P}_{d_0}\}$ can be computed using singular value decomposition of the correlation matrix $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Refer to Appendix A for a detailed formulation of the CCA problem and its solution.

Bach and Jordan (2005) and Browne (1979) proposed a probabilistic generative interpretation of the classical CCA problem that reveals the shared latent representation explicitly. An extension of their results to a more flexible model can be expressed as follows.

Theorem 1 *Assume the probabilistic generative model for the graphical model in Figure 1 as:*

$$\begin{aligned} \boldsymbol{\phi} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{I}_{d_0}), \quad 0 < d_0 \leq \min\{d_1, d_2\} \\ \mathbf{z}_m | \boldsymbol{\phi} &\sim \mathcal{N}(\boldsymbol{W}_m \boldsymbol{\phi} + \boldsymbol{\mu}_{\epsilon_m}, \boldsymbol{\Psi}_m), \\ &\boldsymbol{W}_m \in \mathbb{R}^{d_m \times d_0}, \boldsymbol{\Psi}_m \succcurlyeq 0 \quad \forall m \in \{1, 2\} \end{aligned} \tag{1}$$

where $\boldsymbol{\phi}$ is the shared latent representation. The maximum likelihood estimate of the parameters of this model for view $m \in \{1, 2\}$ can be expressed in terms of the canonical correlation directions as

$$\begin{aligned} \hat{\boldsymbol{W}}_m &= \boldsymbol{\Sigma}_{mm} \boldsymbol{U}_m \boldsymbol{P}_{d_0}^{1/2} \boldsymbol{R} \\ \hat{\boldsymbol{\Psi}}_m &= \boldsymbol{\Sigma}_{mm} - \hat{\boldsymbol{W}}_m \hat{\boldsymbol{W}}_m^\top \\ \hat{\boldsymbol{\mu}}_{\epsilon_m} &= \boldsymbol{\mu}_m - \hat{\boldsymbol{W}}_m \boldsymbol{\mu}_0 \end{aligned} \tag{2}$$

where $\boldsymbol{R}$ is an arbitrary rotation matrix and the residual errors terms can be defined as $\boldsymbol{\epsilon}_m := \mathbf{z}_m - \boldsymbol{W}_m \boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}_{\epsilon_m}, \boldsymbol{\Psi}_m)$, $m \in \{1, 2\}$. This probabilistic graphical model induces conditional independence between $\mathbf{z}_1$ and $\mathbf{z}_2$ given $\boldsymbol{\phi}$. The parameter $\boldsymbol{\mu}_0$ is not identifiable by maximum likelihood.

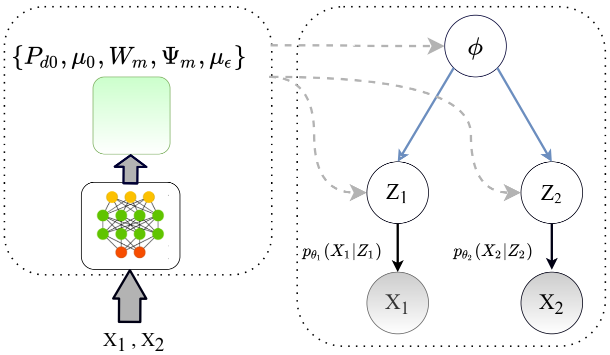

Figure 1: Graphical representation of the deep probabilistic CCA model, where the blue edges belong to latent linear probabilistic CCA model and the black edges represent the deep nonlinear observation networks (decoders) $p_{\theta_m}(\mathbf{x}_m|\mathbf{z}_m) = g_m(\mathbf{z}_m; \theta_m)$. Shaded nodes denotes observed views and dashed line represent the stochastic samples drawn from the approximate posteriors.

Proof: See Appendix A for the proof. ■

In contrast to the results in (Bach and Jordan 2005) where $\boldsymbol{\mu}_0 = \mathbf{0}$, here we assume $\boldsymbol{\mu}_0$ as an extra model parameter. Besides adding an extra degree of freedom in optimizing the objective function of the deep generative model, $\boldsymbol{\mu}_0$ is used as an estimate of the shared representation for the downstream learning tasks in our experiments. We will also derive an analytical form to identify it based on the other parameters of the probabilistic multi-view layer.

Generalization to Arbitrary Number of Views

As an extension to an arbitrary number of views for probabilistic CCA, (Archambeau and Bach 2009) proposed a general probabilistic model as follows:

$$\begin{aligned} \mathbf{z}_m &= \boldsymbol{W}_m \boldsymbol{\phi}_0 + \boldsymbol{T}_m \boldsymbol{\phi}_m + \boldsymbol{\mu}_m + \boldsymbol{\nu}_m, \\ \boldsymbol{\nu}_m &\sim \mathcal{N}(\mathbf{0}, \tau_m^{-1} \mathbf{I}_{d_m}), \\ \boldsymbol{W}_m &\in \mathbb{R}^{d_m \times d_0}, \boldsymbol{T}_m \in \mathbb{R}^{d_m \times q_m}, \forall m \in \{1, ..., M\} \end{aligned} \tag{3}$$

where $\{\boldsymbol{\mu}_m\}_{m=1}^M$ and $\{\boldsymbol{\nu}_m\}_{m=1}^M$ are the view specific offsets and residual errors, respectively. This model can also be viewed as a multibattery factor analysis (MBFA) (Klami et al. 2014; Browne 1980) in the statistics literature, which describes the statistical dependence between all the views by a single shared latent vector, $\boldsymbol{\phi}_0$, and the factor loading matrices $\boldsymbol{W}_m$, and also explains away the view-specific variations by factors that are private to each view, $\boldsymbol{\phi}_m$ with factor loading $\boldsymbol{T}_m$. Restricting to a single view, this model includes the probabilistic factor analysis as a special case if the prior on the view-specific factor is multivariate independent Gaussian, and reduces to probabilistic PCA if the prior is also

[1] **Notation and Definitions:** Throughout the paper, bold lowercase variables denote vectors (*e.g.* $\boldsymbol{x}$) or vector-valued random variables (*e.g.* $\mathbf{x}$), bold uppercase are used for matrices (*e.g.* $\boldsymbol{X}$) or matrix-valued random variables (*e.g.* $\mathbf{X}$) and unbold lowercase are scalars (*e.g.* x) or random variables (*e.g.* x). There are M views in total and subscripts are intended to identify the view-specific variable, (*e.g.* $\mathbf{x}_m, \boldsymbol{\Sigma}_{mm}$), which is different from an element of a vector that is specified by subscript (*e.g.* x_{mi}). The difference should be clear from context.

isotropic. Archambeau and Bach (2009) followed a Bayesian approach to the linear generative model (3) and proposed a variational Expectation-Maximization algorithm to estimate the model parameters. A reformulation for the parameters of this general model inspired by the maximum likelihood solutions of probabilistic CCA in Theorem 1 is presented in Appendix B.

Although constraining the observation models to a classical linear model (1) offers closed form inference for the latent variable(s), as well as efficient training algorithms, the resulting expressiveness is very limited for modeling complex data distributions. On the other hand, the generative descriptions of the probabilistic multi-view models (1) and (3) can be extended naturally as the building blocks of more complex hierarchical models (Klami, Virtanen, and Kaski 2013).

Deep Probabilistic CCA

Deep generative networks are known to be powerful techniques for increasing modeling capacity and improving its expressiveness. Therefore, we append deep generative networks as *observation models* on top of the linear probabilistic model to obtain a combined model, which we denote as *deep probabilistic CCA* or a *deep probabilistic multi-view* network. A graphical representation of this model is depicted in Figure 1. Let $\mathbf{x} := \{\mathbf{x}_m \in \mathbb{R}^{d'_m}\}_{m=1}^M$ denote the collection of observations of all views and $\mathbf{z} := \{\boldsymbol{\phi} \in \mathbb{R}^{d_0}\} \cup \{\mathbf{z}_m \in \mathbb{R}^{d_m}\}_{m=1}^M$ be the collection of the shared latent representation and latent variables corresponding to each view. The nonlinear observation models, also called the *decoders* in the context of variational auto-encoders, are described by deep neural networks $p_{\theta_m}(\mathbf{x}_m|\mathbf{z}_m) = g_m(\mathbf{z}_m; \theta_m)$ with the set of model parameters $\theta = \{\theta_m\}_{m=1}^M$.

In this deep probabilistic model, the *latent linear probabilistic CCA layer* of the form presented in (1) models the linear cross-correlation between all variables $\{\mathbf{z}_m\}_{m=1}^M$ in the latent space, while the nonlinear generative observation networks are responsible for expressing the complex variations of each view. In the following, an approximate variational inference approach is presented for training of this deep generative multi-view model.

Variational Inference

To obtain the maximum likelihood estimate of the model parameters, it is desirable to maximize the marginal data log-likelihood averaged on the dataset $\mathcal{D} = \{x^{(i)}\},\ i = 1, .., N$, which can be expressed as $\log p_\theta(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^N \log p_\theta(\boldsymbol{x}^{(i)}) \simeq \mathbb{E}_{\mathbf{x}\sim\hat{P}_{data}}[\log p_\theta(\mathbf{x})]$.

This objective requires marginalization over all latent variables which entails computing the expectation of the likelihood function $p_\theta(\mathbf{x}|\mathbf{z})$ over the prior distribution on the set of latent variables, $p(\mathbf{z})$. The marginalization is typically intractable for complex models. One work around is to follow the variational inference principle (Jordan et al. 1999), by introducing an approximate posterior distribution $q_\eta(\mathbf{z}|\mathbf{x})$ — also known as *variational inference network* in the context of *amortized variational inference* and is often modeled by deep NNs with model parameters η — then maximize the resulting variational lower bound on the marginal log-likelihood

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\eta}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}[q_\eta(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})] \quad (4)$$

This approach has recently attained renewed interest and, due to its success in training deep generative models, is considered a default, flexible statistical inference method (Rezende, Mohamed, and Wierstra 2014; Kingma and Welling 2013). This bound, also known as the *evidence lower bound (ELBO)*, can be decomposed into two main terms: the first, the expectation of the log-likelihood function $\log p_\theta(\mathbf{x}|\mathbf{z})$, is known as the *negative reconstruction error*. The conditional independence structure of the deep generative multimodal model implies that the likelihood function can be factored, allowing the negative reconstruction error to be expressed as

$$\mathbb{E}_{q_\eta}[\log p_\theta(\mathbf{x}|\mathbf{z})] = \sum_{m=1}^M \mathbb{E}_{q_\eta}[\log p_{\theta_m}(\mathbf{x}_m|\mathbf{z}_m)].$$

Although the expectations above do not typically provide a closed analytical form, they can be approximated using Monte Carlo estimation by drawing L random samples from the approximate posterior $q_\eta(\mathbf{z}|\mathbf{x})$ for each data point $\mathbf{x} = \boldsymbol{x}^{(i)}$.[2]

The second term in the ELBO is the *KL divergence* between the approximate posterior and the prior distribution of the latent variables, which acts as a regularizer that injects prior knowledge about the latent variable into the learning algorithm. Considering the conditional independence of the latent variables $\{\mathbf{z}_m|\boldsymbol{\phi}\}$ induced by the probabilistic graphical model of latent linear layer (1), the approximate posterior of the set of latent variables can be factorized as $q_\eta(\mathbf{z}|\mathbf{x}) = q_\eta(\boldsymbol{\phi}|\mathbf{x})\prod_{m=1}^M q_\eta(\mathbf{z}_m|\boldsymbol{\phi}, \mathbf{x})$ therefore, the KL divergence term can be decomposed into (refer to Appendix C for the derivation)

$$\begin{aligned} D_{\mathrm{KL}}[q_\eta(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})] =& D_{\mathrm{KL}}[q_\eta(\boldsymbol{\phi}|\mathbf{x})\|p(\boldsymbol{\phi})] + \\ & \sum_{m=1}^M D_{\mathrm{KL}}[q_\eta(\boldsymbol{\epsilon}_m|\mathbf{x})\|p(\boldsymbol{\epsilon}_m)] \quad (5) \end{aligned}$$

We model the variational approximate posteriors by joint multivariate Gaussian distributions with marginal densities $q_\eta(\mathbf{z}_m|\mathbf{x}_m) = \mathcal{N}(\mathbf{z}_m; \boldsymbol{\mu}_m(\mathbf{x}_m), \boldsymbol{\Sigma}_{mm}(\mathbf{x}_m))$, which are assumed for simplicity to be elementwise independent per each view, so having diagonal covariance matrices $\boldsymbol{\Sigma}_{mm} = \mathrm{diag}(\boldsymbol{\sigma}_m^2(\mathbf{x}_m))$, $\boldsymbol{\sigma}_m \in \mathbb{R}^{d_m}$. The cross correlation specified by canonical correlation matrix $\boldsymbol{P}_{d_0} = \mathrm{diag}(\boldsymbol{p}(\mathbf{x})), \boldsymbol{p} \in \mathbb{R}^{d_0}$. The parameters of these variational posteriors are specified by separate deep neural networks, also called *encoders*. In this model, a set of encoders are used to output the view-specific moments $\{(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2) = f_m(\mathbf{x}_m; \eta_m)\}_{m=1}^M$, and an encoder network describes the cross correlation $\boldsymbol{p} = f_0(\mathbf{x}^*; \eta_0)$. Depending on the application, $\mathbf{x}^*$ can be either one (or a subset) of the views, when only one (or a subset) of the views are available at the test time (*e.g.* in the multi-view setting where

[2] This, indeed, leads to the Monte Carlo approximation of the gradient of the expected log-likelihood, required for stochastic gradient descent training (Rezende, Mohamed, and Wierstra 2014)

$\mathbf{x}^* = \mathbf{x}_1$), or a concatenation of all the views (*e.g.* in the multi-modal setting).

Combined, the inference model is parameterized by $\eta = \{\eta_0\} \cup \{\eta_m\}_{m=1}^M$. Having obtained the moments of approximate posteriors, we can obtain the canonical directions and subsequently the parameters of the probabilistic CCA model, according to the results presented in Theorem 1.

It is worth noting that the diagonal choices for covariance matrices $\{\mathbf{\Sigma}_{mm}\}_{m=1}^M$ simplify the algebraic operations significantly, resulting in a trivial SVD computation and matrix inversion required for CCA solution and its probabilistic form in Theorem 1.[3] Consequently, we can verify that the canonical pairs of directions will be $(\boldsymbol{u}_{1i}, \boldsymbol{u}_{2i}) = (\sigma_{1i}^{-1}\boldsymbol{e}^{(i)}, \sigma_{2i}^{-1}\boldsymbol{e}^{(i)})$

where $\boldsymbol{e}^{(i)}$ is the standard basis vector $[0,\dots,0,1,0,\dots,0]^\top$ with a 1 at ith position (refer to Appendix A).

Assuming isotropic multivariate Gaussian priors on the latent variables, $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \lambda_0^{-1}\mathbf{I})$ and $\boldsymbol{\epsilon}_m \sim \mathcal{N}(\mathbf{0}, \lambda_m^{-1}\mathbf{I})$, results in closed form solutions for the KL divergence terms (Kingma and Welling 2013). In the following, we provide an analytical approach to optimally identify the mean of the shared latent variable, $\boldsymbol{\mu}_0$, that is not identifiable by likelihood maximization in Theorem 1, from the parameters of the model.

Lemma 1 *Rewriting the KL divergences with respect to the terms depending on the mean of latent factors gives rise to the following optimization problem*

$$\min_{\boldsymbol{\mu}_0, \{\boldsymbol{\mu}_{\epsilon_m}\}_{m=1}^M} \frac{1}{2}\lambda_0\|\boldsymbol{\mu}_0\|^2 + \frac{1}{2}\sum_{m=1}^M \lambda_m\|\boldsymbol{\mu}_{\epsilon_m}\|^2 + \mathcal{K} \quad (6)$$
$$\textit{s.t. } \boldsymbol{\mu}_{\epsilon_m} = \boldsymbol{\mu}_m - \boldsymbol{W}_m\boldsymbol{\mu}_0, \forall m \in \{1,...,M\}$$

where $\mathcal{K}$ is sum of the terms not depending on the means.

Solving this optimization problem results in the unique minimizer

$$\boldsymbol{\mu}_0^* = (\lambda_0\mathbf{I} + \sum_{m=1}^M \lambda_m \boldsymbol{W}_m^\top \boldsymbol{W}_m)^{-1}(\sum_{m=1}^M \lambda_m \boldsymbol{W}_m^\top \boldsymbol{\mu}_m). \quad (7)$$

Having obtained the optimal $\boldsymbol{\mu}_0^*$, one can subsequently compute the means of the view-specific factors, $\{\boldsymbol{\mu}_{\epsilon_m}\}_{m=1}^M$.

Proof: See Appendix C for the proof. ■

According to the inference network, the optimal $\boldsymbol{\mu}_0$ obtained via (7) is a function of all the views, which can be viewed as a type of data fusion in the latent space customized for the variational inference learning of our model. This makes it an appropriate choice for the multi-modal setting. On the other hand, in the multi-view setting we are interested in a solution that depends only on the primary view available at test time. To deal with this, we can solve a revised version of the optimization problem (6) by ignoring the terms that depend on the non-primary views, leading to the minimizer

$$\hat{\boldsymbol{\mu}}_0 = (\lambda_0\mathbf{I} + \lambda_1\boldsymbol{W}_1^\top\boldsymbol{W}_1)^{-1}\lambda_1\boldsymbol{W}_1^\top\boldsymbol{\mu}_1. \quad (8)$$

We further assume that the rotation matrix $\boldsymbol{R}$ is identity in the solution to the probabilistic linear models (2), while leaving it to the deep generative network to approximate the rotation. Specifically, in our neural network architecture, we select a fully connected first layer of the decoder to exactly mimic the rotation matrix.

In summary, the encoders, together with the parameterization of the latent probabilistic CCA layer in (2), provide a variational inference network to estimate the parameters of latent linear model, $\{\boldsymbol{P}_{d_0}(\mathbf{x}_1), \boldsymbol{\mu}_0, \boldsymbol{W}_m(\mathbf{x}_m), \boldsymbol{\Psi}_m(\mathbf{x}_m), \boldsymbol{\mu}_{\epsilon_m}\}_{m=1}^M$ as non-linear functions of the observations.

A detailed discussion of alternative methods for recovering $\boldsymbol{\mu}_0$, processes for drawing samples from the latent variables, and techniques for obtaining more expressive approximate posteriors using normalizing flow (Rezende and Mohamed 2015) are presented in Appendix D.

Related Work

To capture nonlinearity in multi-view data, several kernel-based methods have been proposed (Hardoon, Szedmak, and Shawe-taylor 2004; Bach and Jordan 2003). Such methods, in general, require a large memory to store a massive amount of training data for the test phase. Kernel-CCA in particular requires an $N \times N$ eigenvalue decomposition which is computationally expensive for large datasets. To overcome this issue, some kernel approximation techniques based on random sampling of training data are proposed in (Williams and Seeger 2001) and (Lopez-Paz et al. 2014).

Probabilistic non-linear multi-view learning has been considered in (Shon et al. 2006; Damianou et al. 2012). As an alternative, deep neural networks (DNNs) offer powerful parametric models that can be trained for large pools of data using the recent advances in stochastic optimization algorithms. In the multi-view setting, a deep auto-encoder model, called (SplitAE), was proposed in (Ngiam et al. 2011) in which an encoder maps the primary view to a latent representation and two decoders are trained so that the reconstruction error of both views is minimized.

The classical CCA was extended to deep CCA (DCCA) in (Andrew et al. 2013) by replacing the linear transformations of both views with two deep nonlinear NNs, then learning the model parameters by maximizing the cross correlation between the nonlinear projections. DCCA is then extended to deep CCA autoencoder (DCCAE) in (Wang, Livescu, and Bilmes 2015) where autoencoders are leveraged to additionally reconstruct the inputs, hence introducing extra reconstruction error terms to the objective function. While DCCAE can improve representation learning over DCCA, empirical studies have shown that it tends to ignore the added reconstruction error terms, which results in poorly reconstructed views (Wang, Livescu, and Bilmes 2015). Training algorithms for these classical CCA-based methods require

[3] These types of simplifying assumption on the approximate posteriors have also been used in various deep variational inference models (Rezende, Mohamed, and Wierstra 2014; Kingma and Welling 2013). Although the representation power of such linear latent model is limited but using flexible enough deep generative models, that can explain away the complex nonlinear structures among the data, can justify these choices.

sufficiently large training batches to approximate the covariance matrices and the gradients. Moreover, they do not naturally provide an inference model to estimate the shared latent factor, nor do they enable generative sampling from the model in the input space, while also being restricted to the two-view setting. In contrast, the reconstruction error terms appear naturally in the objective for variational inference, the ELBO, and therefore play a fundamental role in training of the decoder. Furthermore, the stochastic backpropagation method with small mini-batches has proven to be a standard and scalable technique for training deep variational autoencoders (Rezende, Mohamed, and Wierstra 2014). Finally, the probabilistic multi-view model enables enforcing desired structures such as sparsity (Archambeau and Bach 2009) by adopting a broader range of exponential family distributions for priors and approximate posteriors on the latent factors to capture while this property is not immediately apparent in the classical CCA-based variants.

It is worth noting that, although our proposed deep generative model is built upon a single shared latent factor (and also a single correlation matrix to specify the relationship between all the views), it can be seen that the contribution of the shared factor in *m*th view is controlled by the factor loading $\boldsymbol{W}_m$ that is, in turn, a function of $\boldsymbol{P}_{d_0}$ and the view-specific parameter $\boldsymbol{\Sigma}_{mm}$. Thus, the shared factor does not equally influence the views but instead its effect on each view varies by the strength of its projection, $\boldsymbol{W}_m\boldsymbol{\phi}$, which results in dissimilar cross-covariances $\boldsymbol{\Sigma}_{ml}$ for each pair $m \neq l$. This property, in fact, offers flexibility to model uneven dependencies between different subsets of views which is crucial for expressive multi-view modeling when $M > 2$. On the other hand, in the variational two-view autoencoders in (Tang, Wang, and Livescu 2017; Wang et al. 2016), the shared latent representation equally contributes in both views,

so these variational two-view methods can be viewed as special cases of the more generic model proposed here when the posterior factor loading $\{\boldsymbol{W}_m\}_{m=1}^2$ are substituted with identity matrix, hence, they are expected to offer lower flexibility. This can explain why they offer less expressive representation than DCCAE in some experimental studies.

More recently, different VAE based multi-modal deep generative models has been proposed that model the variational posterior of the shared latent variable given all modalities as the product of unimodal posteriors, namely product of experts (PoE) (Wu and Goodman 2018) or as a weighted summation of unimodal posteriors, namely mixture of experts (MoE) (Shi et al. 2019). On the other hand, the linear probabilistic CCA layer describes set of cross-correlated multivariate Gaussian posteriors for random variables $\{\mathbf{z}_m\}_{m=1}^M$. As a potential future direction one can use this form of multivariate Gaussian distribution to specify the base unimodal approximate posteriors, *i.e.* $\{q(\mathbf{z}_1...\mathbf{z}_M|\mathbf{x}_m)\}_{m=1}^M$, for such combination of expert methods (PoE or MoE) to obtain a more expressive multi-view modeling.

Experiments

We empirically evaluate the representation learning performance of the proposed method and compare against well established baselines in two scenarios: I) when several views are available at training time but only a single view (the primary view) is available at test time, namely the multi-view setting, and II) all views are available at training and testing time, namely the multi-modal setting.

Multi-View Experiments

For the experimental study, we used the two-view noisy MNIST datasets of (Wang, Livescu, and Bilmes 2015) and (Wang et al. 2016), where the first view of the dataset was synthesized by randomly rotating each image while the image of the second view was randomly sampled from the same class as the first view, but not necessary the same image, then was corrupted by random uniform noise. As a result of this procedure, both views share only the same digit identity (label) but not the handwriting style. Details of data generation process and samples are available in appendix F.

Experimental design: To provide a fair comparison, we used neural network architectures with the same capacity as those used in (Wang, Livescu, and Bilmes 2015) and (Wang et al. 2016). Accordingly, for the deep network models, all inference and decoding networks were composed of 3 fully connected nonlinear hidden layers of 1024 units, with `ReLU` gates used as the nonlinearity for all hidden units. The first and the second encoder specify $(\boldsymbol{\mu}_1, \text{diag}(\sigma_1^2)) = f_1(\mathbf{x}_1;\theta_1)$, $(\boldsymbol{\mu}_2, \text{diag}(\sigma_2^2)) = f_2(\mathbf{x}_2;\theta_2)$ with the variances specified by a `softplus` function, and an extra encoder modeling the canonical correlations $\text{diag}(p_i)$ using the `sigmoid` function as the output gate. Independent Bernoulli distributions and independent Gaussian distributions were selected to specify the likelihood functions of the first and the second view, respectively, with the parameters of each view being specified by its own decoder network; `sigmoid` functions were applied on outputs used to estimate the means of both views while the variances of the Gaussian variables were specified by `softplus` functions. To prevent over-fitting, stochastic drop-out (Srivastava et al. 2014) was applied to all the layers as a regularization technique. The details of the experimental setup and training procedure can be found in Appendix F.

To evaluate the learned representation, the discriminative and clustering tasks were examined on the shared latent variable. For the discriminative goal, a one-versus-one linear SVM classification algorithm was applied on the shared representation $\boldsymbol{\phi}$. The parameters of the SVM algorithm were tuned using the validation set and the classification error was measured on the test set. We also performed spectral clustering (Von Luxburg 2007) on the *k*-nearest-neighbor graph constructed from the shared representation. To comply with the experiments in (Wang, Livescu, and Bilmes 2015) the degree (number of neighbors) of the nodes was tuned in the set $\{5, 10, 20, 30, 50\}$ using the validation set, and *k*-means was used as the last step to construct a final partitioning into 10 clusters in the embedding space. The proposed deep probabilistic CCA is compared against available multi-view methods in terms of performance at the downstream tasks, reported in Table 1, where the results highlight that the proposed variational model significantly improves representation learning from multi-view datasets.

Method	Error (%)	NMI (%)	ACC (%)
Linear CCA	19.6	56.0	72.9
SpliAE	11.9	69.0	64.0
KCCA	5.1	87.3	94.7
DCCA	2.9	92.0	97.0
DCCAE	2.2	93.4	97.5
VCCA	3.0	-	-
VCCA-private	2.4	-	-
VPCCA	**1.9**	**94.8**	**98.1**

Table 1: Performance of the downstream tasks for different multi-view learning algorithms on the noisy two-view MNIST digit images. Performance measures are classification error rate (the lower the better), normalized mutual information (NMI) and accuracy (ACC) of clustering (the higher the better) (Cai, He, and Han 2005). *VPCCA*: multiview setting, *i.e.* only primary view is available at the test time so $\boldsymbol{\mu}_0$ of equation (8) is used. The results of variational PCCA method are averaged over 3 trials where the results of the baseline methods are from (Wang, Livescu, and Bilmes 2015; Wang et al. 2016). The baseline methods are *Linear CCA*: linear single layer CCA, *DCCA*: deep CCA (Andrew et al. 2013), *Randomized KCCA*: randomized kernel CCA approximation with Gaussian RBF kernels and random Fourier features (Lopez-Paz et al. 2014), *DCCAE*: deep CCA-Auto encoder (Wang, Livescu, and Bilmes 2015), *VCCA(-private)*: (shared-private) multi-view variational auto-encoder (Wang et al. 2016)

Repeating the experiments in the multi-modal setting (*i.e.* both views available at test time, namely VPCCA-2) and using (7) to recover the mean of the shared latent variable significantly improves downstream task performance, resulting in classification error=0.4% and clustering NMI=98.3% or ACC=99.4%. These findings support the merit of the proposed algorithm for successfully integrating information from different modalities.

Figures 2 depict the 2D t-SNE embeddings of the shared latent representations and private factor of the 1st view for multi-view setting (VPCCA), multi-modal setting (VPCCA-2v when both views are available at test time) and VCCA-private (Wang et al. 2016). They verify that the representation of the images of different classes are well separated in the shared latent space while VPCCA can separate the classes better; among them, VPCCA-2v results in the cleanest 2D embedding. For more qualitative experiments, refer to Appendix E.

Multi-Modal Clustering

An important and interesting application of the proposed deep generative model is in clustering multi-modal datasets. Recently, a deep multi-modal subspace clustering method (Abavisani and Patel 2018b) has successfully extended the idea of deep subspace clustering (DSC) (Ji et al. 2017) to multiple modalities. A key component of such approaches is applying a self-expressive layer on a non-linear mapping of the data obtained by deep auto-encoders, which represents the projection of data points as a linear combination of other data point projections. Although offering significant improvement in clustering performance for data lying in nonlinear subspaces, such methods require a self-representation coefficient matrix of size $N \times N$ where N is the number of data points, making this approach prohibitively expensive for large datasets. The clustering performance of the proposed method is evaluated on the following standard datasets.

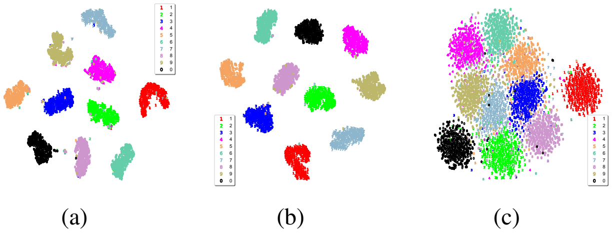
(a) (b) (c)

Figure 2: 2D t-SNE embedding of samples of the shared representation ϕ are from models (a) VPCCA when only 1st view is available at test time, *i.e.* $\phi \sim q_\eta(\phi|\mathbf{x}_1)$, (b) VPCCA-2v when both views are available at test time *i.e.* $\phi \sim q_\eta(\phi|\mathbf{x}_1, \mathbf{x}_2)$, (c) VCCA-private when only 1st view is available at test time, *i.e.* $\phi \sim q_\eta(\phi|\mathbf{x}_1)$.

Handwritten Digits: A two-modal dataset is built by pairing each image in the MNIST dataset with an arbitrary sample of the same class from USPS dataset (Hull 1994) so that the images of both modalities share only the same digit identity but not the style of the handwriting

Multi-modal Facial Components: We also evaluated the proposed method on the multi-modal facial dataset used in (Abavisani and Patel 2018a), based on the Extended Yale-B dataset (Lee, Ho, and Kriegman 2005), where 4 facial components (eyes, nose and mouth) and the whole face image formed 5 different modalities. For this multi-modal data, we train the general deep probabilistic multi-view model (eqn. (11) in apdx. B) that extends the deep probabilistic CCA to an arbitrary number of views.

Please refer to Appendix F for more details and depicted samples of the above multi-modal datasets.

Experimental design: To provide a fair comparison, the encoders and decoders in this set of experiments were defined by neural networks with similar architectures as those used in (Abavisani and Patel 2018a), except that our model does not require the self-expressive layer (*i.e.* a linear fully connected layer with parameter matrix of size $N \times N$ coefficients where N is the number data points). This is a key advantage of the proposed model that significantly reduces the total number of parameters, especially for large input sizes. Thus, the proposed architecture is sufficiently scalable to take advantage of all the training samples.

Accordingly, the encoders (inference networks) of all modalities were composed of convolutional NN (CNN) layers while the decoders (observation networks) were built of transposed convolution layers. `ReLU` gate was used as the nonlinearity for all the hidden units of the deep networks.

The encoders specified $(\boldsymbol{\mu}_m, \mathrm{diag}(\sigma_m^2)) = f_m(\mathbf{x}_m; \theta_m)$, where the variances were modeled by a `softplus` function. An extra encoder network modeled the canonical correlations, diag(p_i), using the `sigmoid` function as the output gate. The observation likelihood functions of all the views, $p_{\theta_m}(\mathbf{x}_m|\mathbf{z}_m)$, were modeled by independent Bernoulli distributions with the mean parameter being specified by decoder networks, $g_m(\mathbf{z}_m; \theta_m)$; with `sigmoid` functions applied to estimate valid means for the distributions. We observed that an optimal choice of the ratio of prior noise precision, λ_0/λ_i, can significantly improve the learned representation for the proposed model. This may be explained by the fact that adjusting the priors of the latent linear probabilistic layer can control the view specific factors to be flexible enough to capture the variations private to each view but restricted enough so as not to describe the relationships between the views. A related idea was elaborated in the formulation of group factor analysis (Klami et al. 2014). In contrast, VCCA-private (Wang et al. 2016) did not exhibit such behavior in our experiments and required less parameter tuning. Details of the model architecture and experimental setup, together with more empirical results are presented in Appendix F.

To estimate shared latent features, VPCCA used the optimal data fusion of (6) in the latent space while, in VCCA-private, we applied a dense linear layer on the outputs of the encoders, $\{f_m(\mathbf{x}_m; \theta_m)\}_{m=1}^{M}$, to estimate $\boldsymbol{\mu}_0$. Clustering is, then, performed on the shared latent factor $\boldsymbol{\phi}$ using spectral clustering (Von Luxburg 2007) on the k-nearest-neighbor graph, with the number of neighbors set to $k = 5$. As the last step, spectral clustering is used to discretize the real-valued representation in the embedding space to extract the final partitioning. The results summarized in Table 2 show that the proposed deep generative model achieves state-of-the-art results, which highlights the fact that the proposed method can efficiently leverage the extra modalities and extract the common underlying information, the cluster memberships, among the modalities.

Extra experiments with multi-modal facial datasets when subset of modalities are missing at test time are presented in appendix E.

Conclusion

In this work, a deep generative multi-view problem was formulated based on the probabilistic interpretation of CCA. A variational inference is customized based on the linear probabilistic CCA model, which resulted in a flexible data fusion method in the latent space. The proposed model is flexible enough to describe arbitrary number of views. Experimental results have shown that the proposed model is able to efficiently integrate the relationship between multiple views to learn a rich common representation, achieving state-of-the-art performance on several downstream tasks, including multi-modal clustering, where the extra modalities were leveraged to uncover the cluster memberships. These indeed suggest that the proposed method is a proper way of extending variational inference to deep probabilistic canonical correlation analysis.

	Digits			**Extended Yale-B**		
	ACC	NMI	ARI	ACC	NMI	ARI
CMVFC	47.6	73.56	38.12	66.84	72.03	40
TM-MSC	80.65	83.44	75.67	63.12	67.06	38.37
MSSC	81.65	85.33	77.36	80.3	82.78	50.18
MLRR	80.6	84.13	76.53	67.62	73.36	40.85
KMSSC	84.4	89.45	79.61	87.65	81.5	63.83
KMLRR	86.85	80.34	82.76	82.45	85.43	59.71
DMSC	95.15	92.09	90.22	99.22	98.89	98.38
VCCA-private	90.02	92.43	85.09	97.52	98.09	96.07
VPCCA	**98.78**	**96.72**	**97.35**	**99.72**	**99.56**	**99.22**

Table 2: Performance for different multi-modal clustering algorithms on two-modal handwritten digits made from MNIST and USPS and multi-modal facial components extracted from Yale-B dataset. Performance metrics are clustering Accuracy rate (ACC), Normalized Mutual Information (NMI) (Cai, He, and Han 2005) and Adjusted Rand Index (ARI) (Rand 1971); all measures are in percent and the higher means the better. Here, we assume that all modalities are available at test time so *VPCCA* uses $\boldsymbol{\mu}_0$ of equation (7). The results of the variational PCCA method are averaged over 3 trials. The clustering performance is compared against the well established subspace clustering methods TM-MSC (Zhang et al. 2015), CMVFC (Cao et al. 2015), MSSC, MLRR , KMSSC, KMLRR (Abavisani and Patel 2018b) and DMSC (Abavisani and Patel 2018a). The results of the above baseline methods are from (Abavisani and Patel 2018a).

Broader Impact

This work targets basic research at the heart of multi-view and multi-modal learning. The goal of unsupervised learning in general is to extract representations from data that reveal hidden regularity. If successful such a capability can provide a powerful tool for enhancing the understanding of complex sensory data, and support downstream tasks like supervised classification or clustering. The goal of this research is to improve the basic underlying learning principles rather than advance the technology in any specific application directions. However, because the techniques are fundamental data analysis capabilities, including visual data analysis, it is possible that abuse could perhaps occur, possibly in the form of making it easier to de-identify data sources from multiple measurements when the sources were not intended to be or did not want to be identified.

References

Abavisani, M.; and Patel, V. M. 2018a. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing* 12(6): 1601–1614.

Abavisani, M.; and Patel, V. M. 2018b. Multimodal sparse and low-rank subspace clustering. *Information Fusion* 39: 168–177.

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*, 1247–1255.

Antelmi, L.; Ayache, N.; Robert, P.; and Lorenzi, M. 2019. Sparse Multi-Channel Variational Autoencoder for the Joint

Analysis of Heterogeneous Data. In *Proceedings of the 36th International Conference on Machine Learning*, 302–311.

Archambeau, C.; and Bach, F. R. 2009. Sparse probabilistic projections. In *Advances in neural information processing systems*, 73–80.

Bach, F. R.; and Jordan, M. I. 2003. Kernel Independent Component Analysis. *J. Mach. Learn. Res.* 3: 1–48. ISSN 1532-4435.

Bach, F. R.; and Jordan, M. I. 2005. A probabilistic interpretation of canonical correlation analysis. *Technical Report 688, Department of Statistics, University of California, Berkeley* .

Browne, M. W. 1979. The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology* 32(1): 75–86.

Browne, M. W. 1980. Factor analysis of multiple batteries by maximum likelihood. *British Journal of Mathematical and Statistical Psychology* 33(2): 184–199.

Cai, D.; He, X.; and Han, J. 2005. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering* 17(12): 1624–1637.

Cao, X.; Zhang, C.; Zhou, C.; Fu, H.; and Foroosh, H. 2015. Constrained multi-view video face clustering. *IEEE Transactions on Image Processing* 24(11): 4381–4393.

Chaudhuri, K.; Kakade, S. M.; Livescu, K.; and Sridharan, K. 2009. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, 129–136. ACM.

Damianou, A.; Ek, C.; Titsias, M.; and Lawrence, N. 2012. Manifold relevance determination. *arXiv preprint arXiv:1206.4610* .

Hardoon, D. R.; Szedmak, S. R.; and Shawe-taylor, J. R. 2004. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Comput.* 16(12): 2639–2664. ISSN 0899-7667.

Hotelling, H. 1992. Relations between two sets of variates. In *Breakthroughs in statistics*, 162–190. Springer.

Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence* 16(5): 550–554.

Ji, P.; Zhang, T.; Li, H.; Salzmann, M.; and Reid, I. 2017. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems*, 24–33.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine learning* 37(2): 183–233.

Karami, M.; Schuurmans, D.; Sohl-Dickstein, J.; Dinh, L.; and Duckworth, D. 2019. Invertible Convolutional Flow. In *Advances in Neural Information Processing Systems*, 5636–5646.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Kingma, D. P.; Welling, M.; et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* 12(4): 307–392.

Klami, A.; and Kaski, S. 2007. Local dependent components. In *Proceedings of the 24th international conference on Machine learning*, 425–432. ACM.

Klami, A.; Virtanen, S.; and Kaski, S. 2013. Bayesian canonical correlation analysis. *Journal of Machine Learning Research* 14(Apr): 965–1003.

Klami, A.; Virtanen, S.; Leppäaho, E.; and Kaski, S. 2014. Group factor analysis. *IEEE transactions on neural networks and learning systems* 26(9): 2136–2147.

Lee, K.-C.; Ho, J.; and Kriegman, D. J. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (5): 684–698.

Lopez-Paz, D.; Sra, S.; Smola, A.; Ghahramani, Z.; and Schölkopf, B. 2014. Randomized nonlinear component analysis. In *International Conference on Machine Learning*, 1359–1367.

Mukuta, Y.; et al. 2014. Probabilistic partial canonical correlation analysis. In *International Conference on Machine Learning*, 1449–1457.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66(336): 846–850.

Rezende, D.; and Mohamed, S. 2015. Variational Inference with Normalizing Flows. In *Proceedings of The 32nd International Conference on Machine Learning*, 1530–1538.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082* .

Shi, Y.; Siddharth, N.; Paige, B.; and Torr, P. 2019. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In *Advances in Neural Information Processing Systems*, 15692–15703.

Shon, A.; Grochow, K.; Hertzmann, A.; and Rao, R. P. 2006. Learning shared latent structure for image synthesis and robotic imitation. In *Advances in neural information processing systems*, 1233–1240.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1): 1929–1958.

Tang, Q.; Wang, W.; and Livescu, K. 2017. Acoustic feature learning via deep variational canonical correlation analysis. *arXiv preprint arXiv:1708.04673* .

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing* 17(4): 395–416.

Wang, W.; Livescu, K.; and Bilmes, J. 2015. On Deep Multi-View Representation Learning. *Icml* 37.

Wang, W.; Yan, X.; Lee, H.; and Livescu, K. 2016. Deep Variational Canonical Correlation Analysis. *arXiv preprint arXiv:1610.03454* .

Williams, C. K.; and Seeger, M. 2001. Using the Nyström method to speed up kernel machines. In *Advances in neural information processing systems*, 682–688.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, 5575–5585.

Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Y. LeCun, C. C. 1998. The MNIST Database of Handwritten Digit.

Zhang, C.; Fu, H.; Liu, S.; Liu, G.; and Cao, X. 2015. Low-rank tensor constrained multiview subspace clustering. In *Proceedings of the IEEE international conference on computer vision*, 1582–1590.