

Multi-View Representation Learning: A Survey from Shallow Methods to Deep Methods

Yingming Li, Ming Yang, Zhongfei (Mark) Zhang, *Senior Member, IEEE*

Abstract—Recently, multi-view representation learning has become a rapidly growing direction in machine learning and data mining areas. This paper first reviews the root methods and theories on multi-view representation learning, especially on canonical correlation analysis (CCA) and its several extensions. And then we investigate the advancement of multi-view representation learning that ranges from shallow methods including multi-modal topic learning, multi-view sparse coding, and multi-view latent space Markov networks, to deep methods including multi-modal restricted Boltzmann machines, multi-modal autoencoders, and multi-modal recurrent neural networks. Further, we also provide an important perspective from manifold alignment for multi-view representation learning. Overall, this survey aims to provide an insightful overview of theoretical basis and current developments in the field of multi-view representation learning and to help researchers find the most appropriate tools for particular applications.

Index Terms—Multi-view representation learning, canonical correlation analysis, multi-modal deep learning.

1 INTRODUCTION

Multi-view data have become increasingly available in real-world applications where examples are described by different feature sets or different “views”, such as image+text, audio+video, and webpage+click-through data. The different views usually contain complementary information, and multi-view learning can exploit this information to learn representation that is more expressive than that of single-view learning method. Therefore, multi-view representation learning has become a very promising topic with wide applicability.

Canonical correlation analysis (CCA) [1] and its kernel extensions [2–4] are representative techniques in early studies of multi-view representation learning. A variety of theories and approaches are later introduced to investigate their theoretical properties, explain their success, and extend them to improve the generalization performance in particular tasks. While CCA and its kernel versions show their abilities of effectively modeling the relationship between two or more sets of variables, they have limitations on capturing high level associations between multi-view data. Inspired by the success of deep neural networks [5–7], deep CCAs [8] have been proposed to this problem, with a common strategy to learn a joint representation that is coupled between multiple views at higher level after learning several layers of view-specific features in the lower layers.

However, how to learn a good association between multi-view data still remains an open problem. In 2016, a workshop on multi-view representation learning is held in conjunction with the 33rd international conference on machine learning to help promote a better understanding of various approaches and the challenges in specific applications. So far, there have been increasing research activities in this direction and a large number of multi-view representation learning algorithms have been presented based on

the fundamental theories of CCAs and progress of deep neural networks. For example, the advancement of multi-view representation learning ranges from shallow methods including multi-modal topic learning [9–11], multi-view sparse coding [12–14], and multi-view latent space Markov networks [15, 16], to deep methods including multi-modal restricted Boltzmann machines [17], multi-modal autoencoders [18–20], and multi-modal recurrent neural networks [21–23]. Further, manifold alignment also provides an important perspective for multi-view representation learning [24].

Therefore, we review the literature of multi-view representation learning from shallow methods to deep methods in accordance with its progress. In particular, each part is surveyed by following two broad, parallel lines of this learning mechanism: one rooted in probabilistic graphical models and the other rooted in non-linear embedding models including the kernel trick and neural networks. In fact, the fundamental difference between the two paradigms is whether the layered architecture of a learning model is to be interpreted as a probabilistic graphical model or as a computation network. The connection between these two paradigms becomes more obvious when we consider deeper multi-view representation learning models. The computational networks have become increasingly important for big data learning since the exact inference of probabilistic models usually becomes intractable with deep structures.

The goal of this survey is to review the theoretical basis and key advances in the area of multi-view representation learning and to provide a global picture of this active direction. We expect this survey to help researchers find the most appropriate approaches for their particular applications and deliver perspectives of what can be done in the future to promote the development of multi-view representation learning.

The remainder of this paper is organized as follows. In Section 2, we introduce theories and advances on CCA and its extensions including probabilistic CCA, sparse CCA, kernel CCA and deep CCA. Section 3 surveys representative shallow multi-view representation learning approaches according to the modeling mechanisms involved. Then, in Section 4, we investigate the popular deep multi-view representation learning methods which have

• Y. Li, M. Yang, Z. Zhang are with College of Information Science & Electronic Engineering, Zhejiang University, China.
E-mail: {yingming, cauchym, zhongfei}@zju.edu.cn

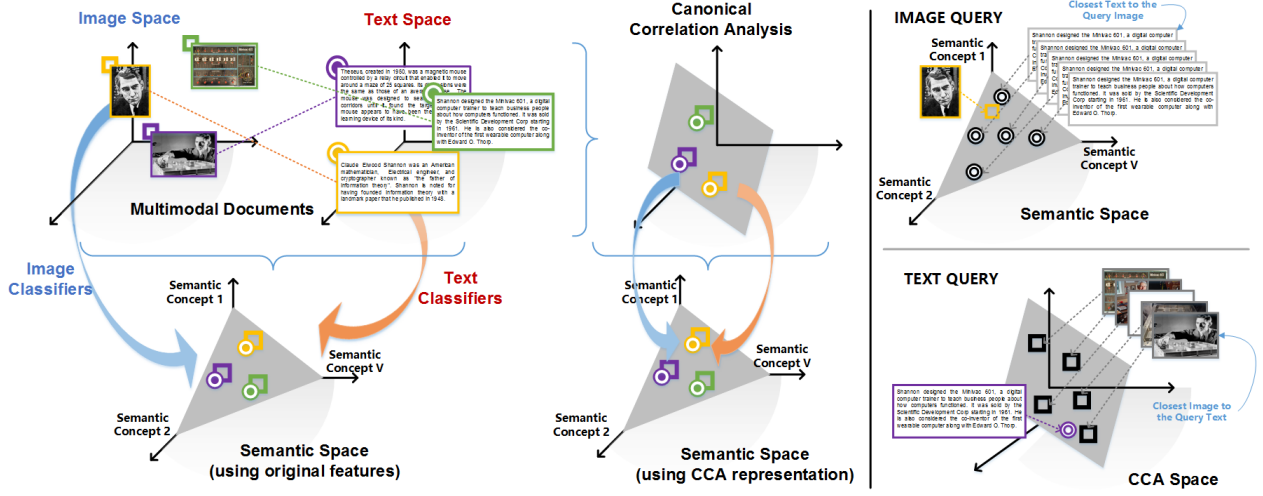


Fig. 1. An illustrative example application of CCA in cross-modal retrieval (adapted from [25]). Left: Embedding of the text and image from their source spaces to a CCA space, Semantic Space and a Semantic space learned using CCA representation. Right: examples of cross-modal retrieval where both text and images are mapped to a common space. At the top is shown an example of retrieving text in response to an image query with a common semantic space. At the bottom is shown an example of retrieving images in response to a text query with a common subspace using CCA.

attracted much attention and been widely applied in cross-media retrieval. Further, Section 5 surveys manifold alignment which provides an important perspective for multi-view representation learning. Finally, we provide a conclusion in Section 6.

2 CANONICAL CORRELATION ANALYSIS AND ITS EXTENSIONS

In this sections we will review the root of multi-view representation learning techniques, canonical correlation analysis (CCA) and its extensions that range from probabilistic graphical modeling to non-linear deep embedding. In particular, we will investigate the related work of probabilistic CCA, sparse CCA, kernel CCA and deep CCA to illustrate the probabilistic, sparse, and non-linear views of multi-view representation learning theories.

2.1 Canonical Correlation Analysis

Canonical Correlation Analysis [1] has become increasingly popular for its capability of effectively modeling the relationship between two or more sets of variables. From the perspective of multi-view representation learning, CCA computes a shared embedding of both sets of variables through maximizing the correlations among the variables between the two sets. More specifically, CCA has been widely used in multi-view learning tasks to generate low dimensional feature representations [25–27]. Improved generalization performance has been witnessed in areas including dimensionality reduction [28–30], clustering [31–33], regression [34, 35], word embeddings [36–38], discriminant learning [39–41], etc. For example, Figure 1 shows a fascinating cross-modality application of CCA in multi-media retrieval.

Given a pair of datasets $X = [x_1, \dots, x_n]$ and $Y = [y_1, \dots, y_n]$, CCA tends to find linear projections w_x and w_y , which make the corresponding examples in the two datasets maximally correlated in the projected space. The correlation coefficient between the two datasets in the projected space is given by

$$\rho = \text{corr}(w_x^T X, w_y^T Y) = \frac{w_x^T C_{xy} w_y}{\sqrt{(w_x^T C_{xx} w_x)(w_y^T C_{yy} w_y)}} \quad (1)$$

where the covariance matrix C_{xy} is defined as

$$C_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)^T \quad (2)$$

where $\mu_x = \frac{1}{n} \sum_{i=1}^n x_i$ and $\mu_y = \frac{1}{n} \sum_{i=1}^n y_i$ are the means of the two views, respectively. The definition of C_{xx} and C_{yy} can be obtained similarly.

Since the correlation ρ is invariant to the scaling of w_x and w_y , CCA can be posed equivalently as a constrained optimization problem.

$$\begin{aligned} \max_{w_x, w_y} \quad & w_x^T C_{xy} w_y \\ \text{s.t.} \quad & w_x^T C_{xx} w_x = 1, \quad w_y^T C_{yy} w_y = 1 \end{aligned} \quad (3)$$

By formulating the Lagrangian dual of Eq.(3), it can be shown that the solution to Eq.(3) is equivalent to solving a pair of generalized eigenvalue problems [3],

$$\begin{aligned} C_{xy} C_{yy}^{-1} C_{yx} w_x &= \lambda^2 C_{xx} w_x \\ C_{yx} C_{xx}^{-1} C_{xy} w_y &= \lambda^2 C_{yy} w_y \end{aligned} \quad (4)$$

Besides the above definition of CCA, there are also other different ways to define the canonical correlations of a pair of matrices, and all these ways are shown to be equivalent [42]. In particular, Kettenring [43] show that CCA is equivalent to a constrained least-square optimization problem. Further, Golub and Zha [42] also provide a classical algorithm for computing CCA by first QR decomposition of the data matrices which whitens the data and then an SVD of the whitened covariance matrix. However, with typically huge data matrices this procedure becomes extremely slow. Avron et al. [30, 44] propose a fast algorithm for CCA with a pair of tall-and-thin matrices using subsampled randomized Walsh-Hadamard transform [45], which only subsamples a small proportion of the training data points to approximate the matrix product. Further, Lu and Foster [46] consider sparse design matrices and introduce an efficient iterative regression algorithm for large scale CCA.

Another line of research for large scale CCA considers stochastic optimization algorithm for CCA [47]. Ma et al. [48] introduce an augmented approximate gradient scheme and further extend it to a stochastic optimization regime. Recent work [49, 50] attempts to transform the original problem of CCA into sequences of least squares problems and solve these problems with accelerated gradient descent (AGD).

Since local correlations usually exist for specific applications besides the global correlation, e.g. in earth science applications, Fern et al. [31] propose to use a mixture of local linear canonical correlation models to find correlation clusters from the data. The proposed method uses a two-step procedure: the compound data set is first partitioned into clusters such that each cluster contains instances whose x features and y features are linearly correlated; then CCA is independently applied to each cluster to form a mixture of correlations that are locally linear. This procedure is named as the correlation clustering to differentiate it from the traditional clustering. Further, Sun et al. [27] formulate CCA as a least squares problem in multi-label classification.

While CCA has the capability of conducting multi-view feature learning and has been widely applied in different fields, it still has some limitations in different applications. For example, it ignores the nonlinearities of multi-view data. Consequently, many algorithms based on CCA have been proposed to extend the original CCA in real-world applications. In the following sections, we review its several widely-used extensions including probabilistic CCA, sparse CCA, kernel CCA and Deep CCA.

2.2 Probabilistic CCA

CCA can be interpreted as the maximum likelihood solution to a probabilistic latent variable model for two Gaussian random vectors. This formulation of CCA as a probabilistic model is proposed by Bach and Jordan [51].

Probabilistic CCA can be formulated by first introducing an explicit latent variable z corresponding to the principal-component subspace. And then a Gaussian prior distribution $p(z)$ over the latent variable, together with two Gaussian conditional distributions $p(x|z)$ and $p(y|z)$ for the observed variables x and y conditioned on the value of the latent variable. In particular, the prior distribution over z is given by a zero-mean unit-covariance Gaussian

$$p(z) = \mathcal{N}(z|0, I_M) \quad (5)$$

Similarly, the conditional distributions of the observed variable x and y , conditioned on the value of the latent variable z , is also Gaussian and given as follows,

$$\begin{aligned} p(x|z) &= \mathcal{N}(x|W_x z + \mu_x, \Psi_x) \\ p(y|z) &= \mathcal{N}(y|W_y z + \mu_y, \Psi_y) \end{aligned} \quad (6)$$

where the mean of x and y is in general linear functions of z governed by the $d_x \times d_z$ matrix W_x and the d_x -dimensional vector μ_x , and by the $d_y \times d_z$ matrix W_y and the d_y -dimensional vector μ_y , respectively. The crucial idea lies in that the latent variables z is shared by the two data sets, while other variables or parameters are independent.

Consequently, Bach and Jordan [51] provide a detailed connection between CCA and probabilistic CCA based on the result of the maximum likelihood estimation of Eq.(6) that $W_x = \Sigma_{xx} U_x P^{1/2} R$ and $W_y = \Sigma_{yy} U_y P^{1/2} R$, where U is composed of the canonical directions, P is a diagonal matrix containing

the canonical correlations, and R is an arbitrary rotation matrix. Further, the posterior expectations $E(z|x)$ and $E(z|y)$ lie in the same subspace that the traditional CCA searches since the obtained canonical weights are the same as the ones for CCA. EM algorithm is applied to find the maximum likelihood solution and is shown to converge to the global optimum. In addition, Browne [52] also proves that the maximum likelihood solution of Eq.(6) is identical to the solution of classical CCA.

However, since probabilistic CCA is based on a Gaussian density model, outliers may lead to serious biases in the parameter estimates when using a maximum likelihood estimation method. To handle the influence of outliers, Archambeau et al. [53] introduce robust canonical correlation analysis by replacing Gaussian distributions with Student- t distributions, which constructs mixtures of robust CCA and can deal with missing data quite easily. Fyfe and Leen [54] consider a Dirichlet process method for performing a non-Bayesian probabilistic mixture CCA.

By treating the mixture case in length, Klami and Kaski [55] present a full Bayesian treatment of probabilistic canonical analyzer. Further, Wang [56] applies a hierarchical Bayesian model to probabilistic CCA and learn this model by variational approximation. By performing the Bayesian model algorithm for probabilistic CCA, the number of canonical correlations can be determined automatically and one would not encounter the issue of selecting the number of canonical correlations, which has been a well known downside for using CCA. Both [55] and [56] exploit the inverse Wishart matrix distribution as a prior for the covariance matrices $\Psi_{x(y)}$ in Eq.(6) and use the automatic relevance determination prior for the linear transformation matrices $W_{x(y)}$.

Later Viinikanoja et al. [57] introduce a mixture of robust canonical correlation analyzers and provide a variational Bayesian inference for learning from noisy data. Such kinds of mixture models can be considered as locally linear models that find several correlation clusters and fit a separate CCA model for each cluster. The clustering process is incorporated into the solution procedure and is also closely related to the CCA models themselves.

Based on exponential family extensions of principal component analysis [58, 59], Klami et al. [60] extend Bayesian CCA to the exponential family by generalizing the Gaussian noise assumption to noise with any distribution in the exponential family. Using the natural parameter formulation of exponential family distributions, certain common choices can be incorporated as special cases, which leading to an efficient way of exploiting the robustness of exponential family distribution in practical models [61, 62]. Recently, Mukuta and Harada [63] follow Wang's approach [56] and introduce a Bayesian extension of partial CCA [64]. Kamada et al. [65] propose a probabilistic semi-CCA by incorporating the mechanism of data missing.

Despite the apparent superiority of Bayesian CCA for model selection, it is worth pointing out that the inference of the Bayesian CCA is difficult for high-dimensional data. Most of the practical applications of Bayesian CCA in the earlier efforts focus on relatively low-dimensional data [55, 56]. To make Bayesian CCA feasible for high-dimensional data, Huopaniemi et al. [66, 67] propose to use multi-way analysis setup by exploiting dimensionality reduction techniques.

2.3 Sparse CCA

Recent years have witnessed a growing interest in learning sparse representations of data. Correspondingly, the problem of sparse

CCA also has received much attention in the multi-view representation learning. The quest for sparsity can be motivated from several aspects. First is the ability to account for the predicted results. The big picture usually relies on a small number of crucial variables, with details to be allowed for variation. The second motivation for sparsity is regularization and stability. Reasonable regularization plays an important role in eliminating the influence of noisy data and reducing the sensitivity of CCA to a small number of observations. Further, sparse CCA can be formulated as a subset selection scheme which reduces the dimensionality of the vectors and makes possible a stable solution.

The problem of sparse CCA can be considered as finding a pair of linear combinations of w_x and w_y with prescribed cardinality which maximize the correlation. In particular, sparse CCA can be defined as the solution to

$$\begin{aligned} \rho = \max_{w_x, w_y} & \frac{w_x^\top C_{xy} w_y}{\sqrt{w_x^\top C_{xx} w_x w_y^\top C_{yy} w_y}} \\ \text{s.t.} \quad & \|w_x\|_0 \leq s_x, \quad \|w_y\|_0 \leq s_y. \end{aligned} \quad (7)$$

Most of the approaches to sparse CCA are based on the well known LASSO trick [68] which is a shrinkage and selection method for linear regression. By formulating CCA as two constrained simultaneous regression problems, Haroon and Shawe-Taylor [69] propose to approximate the non-convex constraints with ∞ -norm. This is achieved by fixing each index of the optimized vector to 1 in turn and constraining the 1-norm of the remaining coefficients. Similarly, Waaijenborg et al. [70] propose to use the elastic net type regression.

In addition, Sun et al. [27] introduce a sparse CCA by formulating CCA as a least squares problem in multi-label classification and directly computing it with the Least Angle Regression algorithm (LARS) [71]. Further, this least squares formulation facilitates the incorporation of the unlabeled data into the CCA framework to capture the local geometry of the data. For example, graph laplacian [72] can be used in this framework to tackle with the unlabeled data.

In fact, the development of sparse CCA is intimately related to the advance of sparse PCA [73–75]. The classical solutions to generalized eigenvalue problem with sparse PCA can be easily extended to that of sparse CCA [76, 77]. Torres et al. [76] derive a sparse CCA algorithm by extending an approach for solving sparse eigenvalue problems using D.C. programming. Based on the sparse PCA algorithm in [74], Wiesel et al. [77] propose a backward greedy approach to sparse CCA by bounding the correlation at each stage. Witten et al. [78] propose to apply a penalized matrix decomposition to the covariance matrix C_{xy} , which results in a method for penalized sparse CCA. Consequently, structured sparse CCA has been proposed by extending the penalized CCA with structured sparsity inducing penalty [79].

Another line of research for sparse CCA is based on sparse Bayesian learning [80–82]. In particular, they are built on the probabilistic interpretation of CCA outlined by [51]. Fyfe and Leen [80] investigate two methods for sparsifying probabilistic CCA. Fujiwara et al. [83] present a variant of sparse Bayesian CCA, using an element-wise automatic relevance determination (ARD) prior configuration, where non-effective parameters are automatically driven to zero. Rai and Daumé III [82] propose a nonparametric, fully Bayesian framework that can automatically select the number of correlation components and capture the sparsity underlying the projections. This framework exploits the

Indian Buffet Process [84] to discover latent feature representation of a set of observations and to control the overall complexity of the model.

Recently, some authors have used probabilistic inter-battery factor analysis (IBFA) [52, 85], which can be considered as an extended CCA model which complements CCA by providing a certain variation not captured by the correlation components, to build more complex hierarchical sparse Bayesian CCA models. By extending IBFA with group-wise sparsity, Klami et al. [86] introduce a Bayesian treatment of the IBFA model, which describes not only the correlations between data sets but also provides components explaining the linear structure within each of the data sets.

Built on multi-battery factor analysis (MBFA) by McDonald [87] and Browne [88] as a generalization of IBFA, Klami et al. [89] introduce a problem formulation of group factor analysis, which extends CCA to more than two sets with structural sparsity, resulting in that it is more flexible than the previous extensions [86]. The solution to this framework can be formulated as variational inference of a latent variable model with a structural sparsity prior.

2.4 Kernel CCA

Canonical Correlation Analysis is a linear multi-view representation learning algorithm, but for many scenarios of real-world multi-view data revealing nonlinearities, it is impossible for a linear embedding to capture all the properties of the multi-view data [90]. Since kernelization is a principled trick for introducing non-linearity into linear methods, kernel CCA [91–93] provides an alternative solution. As a non-linear extension of CCA, kernel CCA has been successfully applied in many situations, including independent component analysis [2], cross-media information retrieval [3, 94, 95], computational biology [3, 96, 97], multi-view clustering [32, 98], acoustic feature learning [99, 100], and statistics [2, 101, 102].

The key idea of kernel CCA lies in embedding the data into a higher dimensional feature space $\phi_x : \mathcal{X} \rightarrow \mathcal{H}$, where \mathcal{H}_x is the reproducing kernel Hilbert space (RKHS) associated with the real numbers, $k_x : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and $k_x(x_i, x_j) = \langle \phi_x(x_i), \phi_x(x_j) \rangle$. k_y, \mathcal{H}_y , and ϕ_y can be defined analogously.

By adapting the representer theorem [103] to the case of multi-view data to state that the following minimization problem,

$$\begin{aligned} \min_{f_1, \dots, f_k} & L((x_1, y_1, f_x(x_1), f_y(y_1)), \dots, (x_n, y_n, f_x(x_n), f_y(y_n))) \\ & + \Omega_x(\|f\|_K^2, \|f\|_K^2) \end{aligned} \quad (8)$$

where L is an arbitrary loss function and Ω is a strictly monotonically increasing function, admits representation of the form

$$f_x(x) = \sum_i \alpha_i k_x(x_i, x), \quad f_y(y) = \sum_i \beta_i k_y(y_i, y) \quad (9)$$

Correspondingly, we replace vectors w_x and w_y in our previous CCA formulation Eq.(1) with $f_x = \sum_i \alpha_i \phi_x(x_i)$ and $f_y = \sum_i \beta_i \phi_y(y_i)$, respectively, and replace the covariance matrices accordingly. The kernel CCA objective can be written as follows:

$$\rho = \frac{f_x^\top \hat{C}_{xy} f_y}{\sqrt{f_x^\top \hat{C}_{xx} f_x f_y^\top \hat{C}_{yy} f_y}} \quad (10)$$

In particular, the kernel covariance matrix \hat{C}_{xy} is defined as

$$\hat{C}_{xy} = \frac{1}{n} \sum_{i=1}^n (\phi_x(x_i) - \mu_{\phi_x(x)}) (\phi_y(y_i) - \mu_{\phi_y(y)})^\top, \quad (11)$$

where $\mu_{\phi_x(x)} = \frac{1}{n} \sum_{i=1}^n \phi_x(x_i)$ and $\mu_{\phi_y(y)} = \frac{1}{n} \sum_{i=1}^n \phi_y(y_i)$ are the means of the two views' kernel mappings, respectively. The form of \hat{C}_{xx} and \hat{C}_{yy} can be obtained analogously.

Let K_x denote the kernel matrix such that $K_x = H \tilde{K}_x H$, where $[\tilde{K}_x]_{ij} = k_x(x_i, x_j)$ and $H = I - \frac{1}{n} \mathbf{1} \mathbf{1}^\top$ is a centering matrix, $\mathbf{1} \in \mathbb{R}^n$ being a vector of all ones. And K_y is defined similarly. Further, we substitute them into Eq.(10) and formulate the objective of kernel CCA as the following optimization problem:

$$\max_{\alpha, \beta} \frac{\alpha^\top K_x K_y \beta}{\sqrt{\alpha^\top K_x^2 \alpha \beta^\top K_y^2 \beta}} \quad (12)$$

As discussed in [3], the above optimization leads to degenerate solutions when either K_x or K_y is invertible. Thus, we introduce regularization terms and maximize the following regularized expression

$$\max_{\alpha, \beta} \frac{\alpha^\top K_x K_y \beta}{\sqrt{\alpha^\top (K_x^2 + \epsilon_x K_x) \alpha \beta^\top (K_y^2 + \epsilon_y K_y) \beta}} \quad (13)$$

Since this new regularized objective function is not affected by re-scaling of α or β , we assume that the optimization problem is subject to

$$\begin{aligned} \alpha^\top K_x^2 \alpha + \epsilon_x \alpha^\top K_x \alpha &= 1 \\ \beta^\top K_y^2 \beta + \epsilon_y \beta^\top K_y \beta &= 1 \end{aligned} \quad (14)$$

Similar to the optimized case of CCA, by formulating the Lagrangian dual of Eq.(13) with the constraints in Eq.(14), it can be shown that the solution to Eq.(13) is also equivalent to solving a pair of generalized eigenproblems [3],

$$\begin{aligned} (K_x + \epsilon_x I)^{-1} K_y (K_y + \epsilon_y I)^{-1} K_x \alpha &= \lambda^2 \alpha \\ (K_y + \epsilon_y I)^{-1} K_x (K_x + \epsilon_x I)^{-1} K_y \beta &= \lambda^2 \beta \end{aligned} \quad (15)$$

Consequently, the statistical properties of KCCA have been investigated from several aspects [2, 104]. Fukumizu et al. [105] introduce a mathematical proof of the statistical convergence of kernel CCA by providing rates for the regularization parameters. Later Hardoon and Shawe-Taylor [106] provide a detailed theoretical analysis of KCCA and propose a finite sample statistical analysis of KCCA by using a regression formulation. Cai and Sun [107] also provide a convergence rate analysis of kernel CCA under an approximation assumption. However, the problems of choosing the regularization parameter in practice remain largely unsolved.

In addition, KCCA has a closed-form solution via the eigenvalue system in Eq.(15), but this solution does not scale up to the large size of the training set, due to the problem of time complexity and memory cost. Thus, various approximation methods have been proposed by constructing low-rank approximations of the kernel matrices, including incomplete Cholesky decomposition [2], partial Gram-Schmidt orthogonalisation [3], and block incremental SVD [99, 108]. In addition, the Nyström method [109] is widely used to speed up the kernel machines [110–113]. This approach is achieved by carrying out an eigen-decomposition on a lower-dimensional system, and then expanding the results back up to the original dimensions.

In early applications, KCCA has shown its ability as a feature preprocessing step that improves subsequent analysis in, for example, classification using a support vector machine [114]. While the combination of KCCA and SVM seems effective, there appears no guarantee that the multi-view representation learned by KCCA will be best suited to the classification task. Consequently, Farquhar et al. [115] investigate the possibility of combining the two distinct stages of KCCA and SVM into a single optimization named as SVM-2K, which has been widely used in classification tasks where two views of the same phenomenon are available.

With the development of image understanding, kernel CCA has already been successfully used to associate images or image regions with different kinds of captions, including individual words, sets of tag, and even sentences [3, 94, 95, 116, 117]. Hardoon et al. [3] first apply KCCA to cross-modality retrieval task, in which images are retrieved by a given multiple text query and without using any label information around the retrieved images. Consequently, KCCA is exploited by Socher and Li [94] to learn a mapping between textual words and visual words so that both modalities are connected by a shared, low dimensional feature space. Further, Hodosh et al. [117] make use of KCCA in a stringent task of associating images with natural language sentences that describe what is depicted.

Moreover, kernel CCA also has attracted much attention in multi-view clustering [32, 98]. Blaschko and Lampert [32] propose a correlational spectral clustering method, which exploits KCCA to do unsupervised clustering of images and text in latent meaning space. For webpage clustering task, Trivedi et al. [98] use a regularized variant of the KCCA algorithm to learn a lower dimensional subspace from heterogeneous data sources. This approach leverages tag information as a complementary part of webpage contents to extract highly discriminative features.

While KCCA can learn representation that is closely related to the underlying generative process of the multi-view data, it may suffer from the small sample effect when the data acquisition in one or more modalities is expensive or otherwise limited. In order to robustly learn the relevant directions maximizing the correlation, Blaschko et al. [97] propose to modify the objective of KCCA with semi-supervised Laplacian regularization [118] to favor directions that lie along the data manifold [118].

2.5 Deep CCA

The CCA-like objectives can be naturally applied to neural networks to capture high-level associations between data from multiple views. In the early work, by assuming that different parts of perceptual input have common causes in the external world, Becker and Hinton [119] present a multilayer nonlinear extension of canonical correlation by maximizing the normalized covariation between the outputs from two neural network modules. Further, Becker [120] explores the idea of maximizing the mutual information between the outputs of different network modules to extract higher order features from coherence inputs.

Later Lai and Fyfe [121, 122] investigate a neural network implementation of CCA and maximize the correlation (rather than canonical correlation) between the outputs of the networks for each view. Hsieh [123] formulates a nonlinear canonical correlation analysis (NLCCA) method using three feedforward neural networks. The first network maximizes the correlation between the canonical variates (the two output neurons), while the remaining two networks map the canonical variates back to the original two sets of variables.

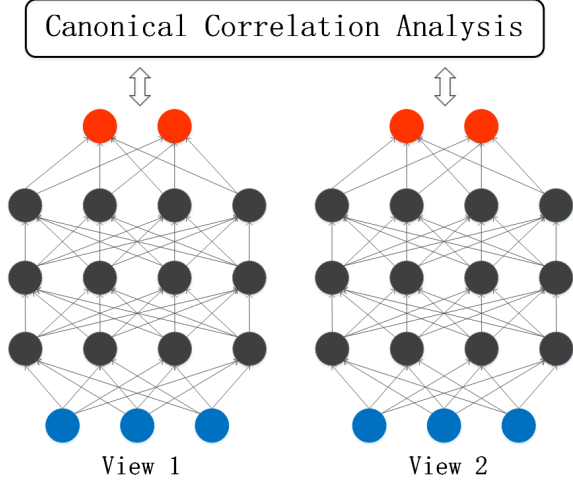


Fig. 2. The framework of deep CCA (adapted from [8]), in which the output layers of two deep networks are maximally correlated.

Although multiple CCA-based neural network models have been proposed for decades, the full deep neural network extension of CCA, referred as deep CCA, has recently been developed by Andrew et al. [8]. Inspired by the recent success of deep neural networks [6, 124], Andrew et al. [8] introduce deep CCA to learn deep nonlinear mappings between two views $\{X, Y\}$ which are maximally correlated. The deep CCA learns representations of the two views by using multiple stacked layers of nonlinear mappings. In particular, assume for simplicity that the network has d intermediate layers, deep CCA first learns deep representation from $f_x(x) = h_{W_x, b_x}(x)$ with parameters $(W_x, b_x) = (W_x^1, b_x^1, W_x^2, b_x^2, \dots, W_x^d, b_x^d)$, where $W_x^l(i, j)$ denotes the parameters associated with the connection between unit i in layer l , and unit j in layer $l + 1$. Also, $b_x^l(j)$ denotes the bias associated with unit j in layer $l + 1$. Given a sample of the second view, the representation $f_y(y)$ is computed in the same way, with different parameters (W_y, b_y) . The goal of deep CCA is to jointly learn parameters for both views such that $\text{corr}(f_x(X), f_y(Y))$ is as high as possible. Let θ_x be the vector of all the parameters (W_x, b_x) of the first view and similarly for θ_y , then

$$(\theta_x^*, \theta_y^*) = \arg \max_{(\theta_x, \theta_y)} \text{corr}(f_x(X; \theta_x), f_y(Y; \theta_y)). \quad (16)$$

For training deep neural network models, parameters are typically estimated with gradient-based optimization methods. Thus, the parameters (θ_x^*, θ_y^*) are also estimated on the training data by following the gradient of the correlation objective, with batch-based algorithms like L-BFGS as in [8] or stochastic optimization with mini-batches [125–127].

Deep CCA and its extensions have been widely applied in learning representation tasks in which multiple views of data are provided. Yan and Mikolajczyk [128] learn a joint latent space for matching images and captions with a deep CCA framework, which adopts a GPU implementation and could deal with overfitting. To exploit multilingual context when learning word embeddings, Lu et al. [126] learn deep non-linear embeddings of two languages using the deep CCA.

Recently, a deep canonically correlated autoencoder (DCCAE) [20] is proposed by combining the advantages of the deep CCA and autoencoder-based approaches. In particular, DCCAE consists

of two autoencoders and optimizes the combination of canonical correlation between the learned bottleneck representations and the reconstruction errors of the autoencoders. This optimization offers a trade-off between information captured in the embedding within each view on one aspect, and the information in the relationship across views on the other.

3 SHALLOW METHODS ON MULTI-VIEW REPRESENTATION LEARNING

In this section, we will first review the shallow multi-view representation learning methods from probabilistic modeling perspective, and then survey the related methods from directly parameterized representation learning perspective.

3.1 Probabilistic Models

From the probabilistic modeling perspective, the problem of multi-view feature learning can be interpreted as an attempt to recover a compact set of latent random variables that describe a distribution over the observed multi-view data. We can express $p(x, y, h)$ as a probabilistic model over the joint space of the latent variables h , and observed two-view data x, y . Feature values are determined by the posterior probability $p(h|x, y)$. Parameters are usually estimated by maximizing the regularized likelihood of the training data.

3.1.1 Multi-View Probabilistic Latent Semantic Analysis

Probabilistic latent semantic analysis (PLSA) [129] is a statistical variant of Latent Semantic Analysis (LSA) [130] and a significant step toward probabilistic modeling of text. It performs a probabilistic decomposition by modeling each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of “topics” $z \in \mathcal{Z} = \{z_1, \dots, z_K\}$. Each topic z_k is associated with relative estimates $P(t_i|z_k)$ for each term in the corpus. A document d_j is then represented as a convex combination of factors with mixing weights $P(z_k|d_j)$, i.e., the predictive probabilities for terms in a particular document have the form $P(t_i|d_j) = \sum_k P(t_i|z_k)P(z_k|d_j)$, with non-negative probabilities and constraints $\sum_i P(t_i|z_k) = 1$ for all k and $\sum_k P(z_k|d_j) = 1$ for all j .

Multi-modal PLSA was originally proposed for the problem of jointly modeling content and connectivity. In the original Link-PLSA model [9], both term-based PLSA [129] and citation-based PHITS [131] are merged into a joint probabilistic model, explaining terms and citations in terms of a common set of underlying factors. In particular, Link-PLSA defines the following joint model for predicting citations/links and terms in documents:

$$\begin{aligned} P(t_i|d_j) &= \sum_k P(t_i|z_k)P(z_k|d_j), \\ P(c_l|d_j) &= \sum_k P(c_l|z_k)P(z_k|d_j). \end{aligned} \quad (17)$$

where $P(z_k|d_j)$ denotes the shared document-specific mixing proportions. It couples the conditional probabilities for terms and citations: each topic has some probability $P(c_l|z_k)$ of linking to document d_l as well as some probability $P(t_i|z_k)$ of containing an occurrence of term t_i . This coupling integrates content and link information in a principled manner.

Further, Link-PLSA aims at maximizing the following (normalized) log-likelihood function,

$$\mathcal{L} = \sum_j \left[\alpha \sum_i \frac{C_{ij}}{\sum_{i'} C_{i'j}} \log \sum_k P(t_i|z_k)P(z_k|d_j) + (1 - \alpha) \sum_i \frac{A_{lj}}{\sum_{l'} A_{l'j}} \log \sum_k P(c_l|z_k)P(z_k|d_j) \right] \quad (18)$$

where C is the term-document matrix of word counts and A is a matrix of document-citation pair. C_{ij} denotes how often a term t_i occurs in document d_j and A_{ij} is nonzero if d_j contains a hyperlink d_i . Following the EM approach it is straightforward to derive the parameter estimation.

Further, Monay et al. [132] present a multi-view PLSA-based approach called PLSA-words. The principal idea of PLSA-words is to constrain the latent space by focusing on textual features and then learning visual variations conditioned on the space learned from text. If the conditional dependence assumption holds, then the parameter learning for visual features is effective, subject to the generated informative latent space from textual features. If the dependence assumption is violated, then on average the parameter learning is less effective and multi-modal PLSA learning may not be successful. Instead of conditional single directed dependence from textual modality to visual modality, Zhang et al. [133] propose a PLSA based probabilistic multi-modal semantic model, in which the visual features and the textual words are fully connected via a hidden layer to allow much more mutual dependence between each other than PLSA-words.

Considering that multi-view PLSA methods have good performance on image annotations, extensions are designed to conduct image annotation or clustering tasks [134–141]. For example, inspired by the knowledge from neuroscience that our neocortex consists of multiple layers, Lienhart et al. [136] propose a multilayer multimodal PLSA model, which can handle different modalities as well as different features within a mode.

3.1.2 Multi-Modal Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) [142] is a generative probabilistic model for collections of a corpus. It proceeds beyond PLSA through providing a generative model at words and document level simultaneously. In particular, LDA is a three-level hierarchical Bayesian network that models each document as a finite mixture over an underlying set of topics.

As a generative model, LDA is extendable to multi-view data. Barnard et al. [10] present a mixture of multi-modal LDA model (MoM-LDA), which describes the following generative process for the multi-modal data: each document (consisting of visual and textual information) has a distribution for a fixed number of topics (mixture components), and given a specific topic the visual features and textual words are generated. Further, the probability distribution of the topics is different for each image-caption pair, which is achieved by imposing a Dirichlet prior for the distribution of topics.

Consequently, Blei and Jordan [11] propose a correspondence LDA (Corr-LDA) model, which not only allows simultaneous dimensionality reduction in the representation of region descriptions and words, but also models the conditional correspondence between their respectively reduced representations. The graphical model of Corr-LDA is depicted in Figure 3. This model can be viewed in terms of a generative process that first generates the region descriptions and subsequently generates the caption words.

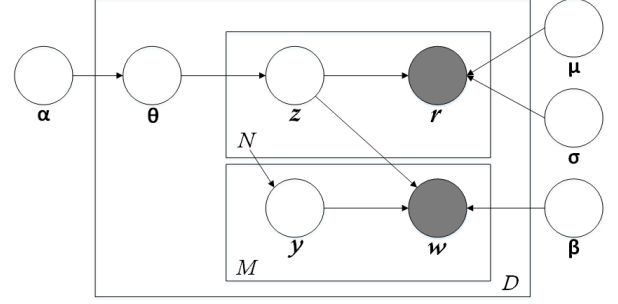


Fig. 3. The graphical model representation of the Corr-LDA model (adapted from [11]).

In particular, let $\mathbf{z} = \{z_1, z_2, \dots, z_N\}$ be the latent variables that generate the image, and let $\mathbf{y} = \{y_1, y_2, \dots, y_M\}$ be discrete indexing variables that take values from 1 to N with equal probability. Each image and its corresponding caption are represented as a pair (\mathbf{r}, \mathbf{w}) . The first element $\mathbf{r} = \{r_1, \dots, r_N\}$ is a collection of N feature vectors associated with the regions of the image. The second element $\mathbf{w} = \{w_1, \dots, w_M\}$ is the collection of M words of the caption. Given N and M , a K -factor Corr-LDA model assumes the following generative process for an image/caption (\mathbf{r}, \mathbf{w}) :

- 1) Sample $\theta \sim \text{Dir}(\theta|\alpha)$.
- 2) For each image region $r_n, n \in \{1, \dots, N\}$:
 - a) Sample $z_n \sim \text{Mult}(\theta)$
 - b) Sample $r_n \sim p(r|z_n, \mu, \sigma)$ from a multivariate Gaussian distribution conditioned on z_n .
- 3) For each caption word $w_m, m \in \{1, \dots, M\}$:
 - a) Sample $y_m \sim \text{Unif}(1, \dots, N)$
 - b) Sample $w_m \sim p(w|y_m, \mathbf{z}, \beta)$ from a multinomial distribution conditioned on the z_{y_m} factor.

Consequently, Corr-LDA specifies the following joint distribution on image regions, caption words, and latent variables:

$$p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}) = p(\theta|\alpha) \left(\prod_{n=1}^N p(z_n|\theta) p(r_n|z_n, \mu, \sigma) \right) \cdot \left(\prod_{m=1}^M p(y_m|N) p(w_m|y_m, \mathbf{z}, \beta) \right) \quad (19)$$

Note that exact probabilistic inference for Corr-LDA is intractable and we employ variational inference methods to approximate the posterior distribution over the latent variables given a particular image/caption.

Both MoM-LDA and Corr-LDA achieve great success in image annotation. However, there are still limitations with these models. First, selecting the number of mixture components to be used in a MoM-LDA or Corr-LDA model is difficult. In particular, different choices of the number of mixture components correspond to different MoM-LDA or Corr-LDA models. Second, MoM-LDA and Corr-LDA models neglect the discriminative information which may be helpful for multi-view learning. For example, image captions provide evidence for the class label, and the class label provides evidence for image captions. Third, both models discard substantial visually auxiliary information existing in the images,

e.g., the position information of the visual features. For multi-modal topic modeling, we are ultimately interested in defining a latent space that is consistent in semantic terms, while discarding as little useful information about the multi-modal data as possible.

Amount of work have been presented to solve the above limitations. Based on hierarchical Dirichlet process (HDP) [143], Yakhnenko and Honavar [144] introduce a multi-modal hierarchical Dirichlet process model (MoM-HDP), which is an extension of MoM-LDA with an infinite number of mixture components. Thus, MoM-HDP is capable of removing the need for a prior choice of the number of mixture components or the computational expense of model selection. Note that in practice, the Dirichlet process is approximated by truncating it [145]. Qi et al. [146] also present a correspondence hierarchical Dirichlet process model (Corr-HDP), which is an extension of the Corr-LDA model using a hierarchical Dirichlet process instead.

Following the supervised LDA algorithms [147], supervised multi-modal LDA models are subsequently proposed to make effective use of the discriminative information. Since supervised topic models can be classified into two categories: *downstream models* and *upstream models* [147], supervised multi-modal LDA models can also be categorized accordingly. For a downstream supervised multimodal model, the supervised response variables are generated from topic assignment variables. For instance, Wang et al. [148] develop a multi-modal probabilistic model for jointly modeling the image, its class label, and its annotations, called multi-class supervised LDA with annotations, which treats the class label as a global description of the image, and treats annotation terms as local descriptions of parts of the image. Its underlying assumptions naturally integrate the multi-modal data with their discriminative information so that it takes advantage of the merits of Corr-LDA and supervised LDA [149] simultaneously.

For an upstream supervised multi-modal model, the response variables directly or indirectly generate latent topic variables [150–152]. For instance, Cao et al. [151] propose a spatially coherent latent topic model (Spatial-LTM), which represents an image containing objects in two different modalities: appearance features and salient image patches. Further, a supervised spatial-LTM model is presented by incorporating label information into distribution of topics. Nguyen et al. [153] propose a Multi-modal Multi-instance Multi-Label LDA model (M3LDA), in which the model consists of a visual-label part, a textual-label part, and a label-topic part. The underlying idea is to make the topic decided by the visual information to be consistent with the topic decided by the textual information, leading to the correct label assignment.

3.1.3 Shared Gaussian Process Latent Variable Model

Gaussian Processes (GPs) [154, 155] are generalizations of Gaussian distributions defined over infinite index sets. GPs have become powerful models for classification and regression that subsume numerous classes of function approximators, such as single hidden-layer neural networks and RBF networks. Lawrence [156] proposes the Gaussian Process Latent Variable Model (GPLVM) as a non-linear dimensionality reduction technique. Consequently, Shon et al. [157] present the shared GPLVM (SGPLVM) model as a generalization of GPLVM that can handle multiple observation spaces, where each set of observations is parameterized by a different set of kernel parameters. This shared GPLVM model can be considered as a nonlinear extension to CCA.

Let X, Y be matrices of observations drawn from spaces of dimensionality d_X, d_Y respectively, and Z be a latent space of dimensionality $d_Z \leq d_X, d_Y$. Assume that each latent variable z_i generates a pair of observations x_i and y_i through GPs parameterized as non-linear functions $f_X : Z \rightarrow X$ and $f_Y : Z \rightarrow Y$. By employing an RBF kernel, the similarity between two data points z, z' can be defined as:

$$k(z, z') = \alpha_X \exp\left(-\frac{\gamma_X}{2} \|z - z'\|^2\right) + \delta_{z, z'} \beta_X^{-1} \quad (20)$$

where $\theta_X = \{\alpha_X, \beta_X, \gamma_X\}$ are the given hyperparameters for the X space and δ represents the delta function.

The priors $P(\theta_X), P(\theta_Y), P(X)$, the likelihoods $P(X), P(Y)$ for the X, Y observation spaces, and the joint likelihood $P_{GP}(X, Y, Z, \theta_X, \theta_Y)$ are given by:

$$\begin{aligned} P(X|\theta_X, Z) &= \frac{|W|^N}{\sqrt{(2\pi)^{ND_X}} |K|^{D_X}} \exp\left(-\frac{1}{2} \sum_{k=1}^{D_X} w_k^2 X_k^\top K_X X_k\right) \\ P(Y|\theta_Y, Z) &= \frac{|V|^N}{\sqrt{(2\pi)^{ND_Y}} |K|^{D_Y}} \exp\left(-\frac{1}{2} \sum_{m=1}^{D_Y} v_m^2 Y_m^\top K_Y Y_m\right) \\ P(\theta_X) &\propto \frac{1}{\alpha_X \beta_X \gamma_X} \quad P(\theta_Y) \propto \frac{1}{\alpha_Y \beta_Y \gamma_Y} \\ P(Z) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \sum_i \|z_i\|^2\right) \end{aligned}$$

$$\begin{aligned} P_{GP}(X, Y, Z, \theta_X, \theta_Y) \\ = P(X|\theta_X, Z) P(Y|\theta_Y, Z) P(\theta_X) P(\theta_Y) P(Z) \end{aligned} \quad (21)$$

where $\alpha_X, \beta_X, \gamma_X$ are hyperparameters for the X space, and w_k, v_m respectively denote the diagonal entries for matrices W, V .

Given a trained SGPLVM, the parameters in one observation space can be inferred from the given parameters in the other observation space (e.g., infer the corresponding y given x). This problem can be solved in two steps. First, the most likely latent coordinate z given the observation x can be determined by maximizing the log likelihood of (z, x) . Once the correct latent coordinate z has been inferred for a given x , the corresponding observation y can be predicted with the trained SGPLVM.

Following the idea of discriminative Gaussian process latent variable model (DGPLVM) [158], Eleftheriadis et al. [159] propose a discriminative shared Gaussian process latent variable model (DS-GPLVM) for multi-view analysis by exploiting a shared discriminative manifold by multi-view data. Further, Xu et al. [160] propose a large margin multi-view Gaussian process via an integration of Gaussian process and large-margin principle. Recently, by imposing cross-modal similarity/dissimilarity constraint as a prior to the latent space, Song et al. [161] propose a multi-modal regularized similarity Gaussian Process variable model to enforce that the semantically similar/dissimilar cross-modal observations are also similar/dissimilar in the latent space, which maximizes the cross-modal semantic consistency.

3.1.4 Collective Matrix Factorization

Collective Matrix factorization (CMF) have become an important tool for boosting the overall factorization quality in the case of multiple view related matrices [162–169]. The learned factor matrices can be used for estimating missing values in collaborative filtering, or as multi-view representation for classification and clustering.

In general, a simple form of CMF considers multiple low rank data matrices $\{X_i \in \mathbb{R}^{n \times d_i}\}_{i \in \mathcal{I}}$ of the same row dimension, and simultaneously factorizes them based on the following form

$$X_i = UV_i^\top \text{ for all } i \in \mathcal{I} \quad (22)$$

where it is assumed that these matrices share the same factor matrix $U \in \mathbb{R}^{n \times k}$ and each matrix has its own loading matrix $V_i \in d_i \times k$. Bias for different factors and loadings can also be incorporated [170].

To use CMF, we must compute the factors and loadings given multiple observed data matrices. The common approach is to minimize the regularized squared error loss with respect to U and $\{V_i\}_{i \in \mathcal{I}}$,

$$\min_{U, V_i} \sum_{i \in \mathcal{I}} \lambda_{X_i} \|X_i - UV_i^\top\|_F^2 + \lambda_U \|U\|_F^2 + \sum_{i \in \mathcal{I}} \lambda_{V_i} \|V_i\|_F^2 \quad (23)$$

where λ_U and λ_{V_i} are regularization parameters.

The above regularized form of CMF can be generalized as a probabilistic model. In probabilistic collective matrix factorization (PCMF), we assume the following generative distributions,

$$\begin{aligned} p(U) &= \mathcal{N}_{n,k}(0, \lambda_U^{-1} I_n \otimes I_k) \\ p(V_i) &= \mathcal{N}_{d_i,k}(0, \lambda_{V_i}^{-1} I_{d_i} \otimes I_k) \\ p(X_i) &= \mathcal{N}_{n,d_i}(UV_i^\top, \lambda_{X_i}^{-1} I_n \otimes I_{d_i}) \end{aligned} \quad (24)$$

where I_k is a k -dimensional identity matrix. This is the probabilistic interpretation of CMF that other complex models build on [171]. CMF model is usually fit by finding a locally optimal solution of factor matrix U and loading matrices V_i , with an iterative algorithm. Further, from the above generative distributions of model components, we can see that the main difference between probabilistic CMF and probabilistic CCA in Eq.(6) depends on the choice of the noise distribution on top of the shared components.

In early studies of joint matrix factorization, Yu et al. [163] propose a multi-label informed latent semantic indexing algorithm which exploits the idea of joint matrix factorization to learn a shared latent semantic factor. Consequently, Zhu et al. [172] introduce a joint matrix factorization model to combine the information of content and links for web page analysis. By sharing factors between text content and link structure, this model could learn a consistent and compact feature representation. Further, supervised information is incorporated into the joint matrix factorization process to obtain a better representation for the classification task.

Later Singh and Gordon [162] propose collective matrix factorization in which several matrices are factorized simultaneously, sharing parameters among factors. This CMF approach allows nonlinear relationships between the parameters and outputs, using Bregman divergences to measure the error. Further, a hierarchical Bayesian variant [173] is proposed to provide an estimate of uncertainty during the relational modeling process. In particular, a block Metropolis-Hastings sampler has been presented to speed up the parameter estimation.

Further, CMF has been applied for multi-view hashing [174, 175]. A multimodal latent binary embedding (MLBE) method is proposed by Zhen and Yang [174] to learn hash functions from a probabilistic generative perspective. MLBE encodes the intra-modality and inter-modality similarities through a latent factor model in which hash codes are binary latent factors in a common Hamming space. Further, Ding et al. [175] put forward collective matrix factorization hashing to learn unified hash codes

by CMF with a latent factor model from different modalities of one instance.

In addition, joint nonnegative matrix factorization (NMF) also has been proposed for multi-view analysis [176–179]. For example, Liu et al. [177] introduce a joint NMF algorithm to learn a factorization that gives compatible clustering solutions across multiple views. The key idea lies in formulating a joint matrix factorization process with the constraint that pushes clustering solution of each view towards a common consensus instead of fixing it directly.

3.1.5 Multi-View Sparse Coding

Like CMF, multi-view sparse coding also has both a probabilistic and non-probabilistic interpretation. Multi-view sparse coding [12–14] relates a latent representation (either a vector of random variables or a feature vector, depending on the interpretation) to the multi-view data through a set of linear mappings, which we refer to as the *dictionaries*. It has the property of finding shared representation h^* which picks out the most appropriate bases and zeros others, given a high degree of correlation with the input. This property is owing to the explaining away effect which arises naturally in directed graphical models [180].

The main difference between multi-view sparse coding and CMF is that multi-view sparse coding adopts a penalty to ensure a sparse activation of h . Given a pair of datasets $\{X \in \mathbb{R}^{n \times d_x}, Y \in \mathbb{R}^{n \times d_y}\}$, a non-probabilistic multi-view sparse coding scheme can be seen as recovering the code or feature vector associated with a new multi-view input via:

$$h^* = \arg \min_h \|x - W_x h\|_2^2 + \|y - W_y h\|_2^2 + \lambda \|h\|_1 \quad (25)$$

Learning the pair of dictionaries $\{W_x, W_y\}$ can be accomplished by optimizing the following training criterion with respect to W_x and W_y :

$$\mathcal{J}_{W_x, W_y} = \sum_i (\|x_i - W_x h_i^*\|_2^2 + \|y_i - W_y h_i^*\|_2^2) \quad (26)$$

where x_i and y_i are the two modal inputs and h^* is the corresponding sparse code determined by Eq.(25). In particular, W_x and W_y are usually constrained to have unit-norm columns.

The probabilistic interpretation of multi-modal sparse coding differs from that of CMF, in that instead of a Gaussian prior on the latent variable h , we use a sparsity inducing Laplace prior (corresponding to an L1 penalty):

$$\begin{aligned} p(h) &= \prod_j \frac{2}{\lambda} \exp(-\lambda |h_j|) \\ \forall_{i=1}^n : \quad & p(x_i|h) = \mathcal{N}(x_i; W_x h + \mu_{x_i}, \sigma_{x_i}^2 \mathbf{I}) \\ & p(y_i|h) = \mathcal{N}(y_i; W_y h + \mu_{y_i}, \sigma_{y_i}^2 \mathbf{I}) \end{aligned} \quad (27)$$

In the case of two-view sparse coding, because we seek a sparse multi-view representation (i.e., one with many features set to zero), we are interested in recovering the MAP (maximum *a posteriori*) value of h : i.e., $h^* = \arg \max_h p(h|x, y)$ rather than its expected value $E[h|x, y]$. Under this interpretation, learning dictionaries W_x and W_y proceeds as maximizing the likelihood of the data given these MAP values of h^* : $\arg \max_{W_x, W_y} \prod_i p(x_i|h^*)p(y_i|h^*)$ subject to the norm constraint on W_x and W_y . Note that this parameter learning scheme, subject to the MAP values of the latent h , is not a standard practice in the probabilistic graphical model literature.

Typically, the likelihood of the multi-view data is maximized directly. In the presence of latent variables, expectation maximization is employed where the parameters are optimized with respect to the marginal likelihood, i.e., summing or integrating the joint log-likelihood over the all values of the latent variables under their posterior, rather than considering only the single MAP value of h . The theoretical properties of this form of parameter learning are not yet well understood but seem to work in practice (e.g., Gaussian mixture models).

One might expect that multi-view sparse representation would significantly leverage the performance especially when features for different views are complementary to one another and indeed it seems to be the case. There are numerous examples of its successful applications as a multi-view feature learning scheme, including human pose estimation [12], image classification [181], web data mining [182], as well as cross-media retrieval [183, 184].

3.1.6 Multi-View Latent Space Markov Networks

Undirected graphical models, also called Markov random fields that have many special cases, including the exponential family Harmonium [185] and restricted Boltzmann machine [186]. Within the context of unsupervised multi-view feature learning, Xing et al. [15] first introduce a particular form of multi-view latent space Markov network model called multi-wing harmonium model. This model can be viewed as an undirected counterpart of the aforementioned directed aspect models such as multi-modal LDA [11], with the advantages that inference is fast due to the conditional independence of the hidden units and that topic mixing can be achieved by document- and feature-specific combination of aspects.

For simplicity, we begin with dual-wing harmonium model, which consists of two modalities of input units $X = \{x_i\}_{i=1}^n$, $Y = \{y_j\}_{j=1}^n$, and a set of hidden units $H = \{h_k\}_{k=1}^n$. In this dual-wing harmonium, each modality of input units and the hidden units constructs a complete bipartite graph where units in the same set have no connections but are fully connected to units in the other set. In addition, there are no connections between two input modalities. In particular, consider all the case where all observed and hidden variables are from exponential family; we have

$$\begin{aligned} p(x_i) &= \exp\{\theta_i^T \phi(x_i) - A(\theta_i)\} \\ p(y_j) &= \exp\{\eta_j^T \psi(y_j) - B(\eta_j)\} \\ p(h_k) &= \exp\{\lambda_k^T \varphi(h_k) - C(\lambda_k)\} \end{aligned} \quad (28)$$

where $\phi(\cdot)$, $\psi(\cdot)$, and $\varphi(\cdot)$ are potentials over cliques formed by individual nodes, θ_i , η_j , and λ_k are the associated weights of potential functions, and $A(\cdot)$, $B(\cdot)$, and $C(\cdot)$ are log partition functions.

Through coupling the random variables in the log-domain and introducing other additional terms, we obtain the joint distribution $p(X, Y, H)$ as follows:

$$\begin{aligned} p(X, Y, H) &\propto \exp\left\{\sum_i \theta_i^T \phi(x_i) + \sum_j \eta_j^T \psi(y_j) + \sum_k \lambda_k^T \varphi(h_k) \right. \\ &\quad \left. + \sum_{ik} \phi(x_i)^T W_{ik} \varphi(h_k) + \sum_{jk} \psi(y_j)^T U_{jk} \varphi(h_k)\right\} \end{aligned} \quad (29)$$

where $\phi(x_i)\varphi(h_k)$, $\psi(y_j)\varphi(h_k)$ are potentials over cliques consisting of pairwise linked nodes, and W_{ik} , U_{jk} are the associated

weights of potential functions. From the joint distribution, we can derive the conditional distributions

$$\begin{aligned} p(x_i|H) &\propto \exp\{\tilde{\theta}_i^T \phi(x_i) - A(\tilde{\theta}_i)\} \\ p(y_j|H) &\propto \exp\{\tilde{\eta}_j^T \psi(y_j) - B(\tilde{\eta}_j)\} \\ p(h_k|X, Y) &\propto \exp\{\tilde{\lambda}_k^T \varphi(h_k) - C(\tilde{\lambda}_k)\} \end{aligned} \quad (30)$$

where the shifted parameters $\tilde{\theta}_i = \theta_i + \sum_k W_{ik} \varphi(h_k)$, $\tilde{\eta}_j = \eta_j + \sum_k U_{jk} \varphi(h_k)$, and $\tilde{\lambda}_k = \lambda_k + \sum_i W_{ik} \phi(x_i) + \sum_j U_{jk} \psi(y_j)$.

In training probabilistic models parameters are typically updated in order to maximize the *likelihood of the training data*. The updating rules can be obtained by taking derivative of the log-likelihood of the sample defined in Eq.(29) with respect to the model parameters. The multi-wing model can be directly obtained by extending the dual-wing model when the multi-modal input data are observed.

Further, Chen et al. [16] present a multi-view latent space Markov network and its large-margin extension that satisfies a weak conditional independence assumption that data from different views and the response variables are conditionally independent given a set of latent variables. In addition, Xie and Xing [187] propose a multi-modal distance metric learning (MMDML) framework based on the multi-wing harmonium model and metric learning method by [188]. This MMDML provides a principled way to embed data of arbitrary modalities into a single latent space where distance supervision is leveraged.

3.2 Directly Learning A Parametric Embedding from Multi-View Input to Representation

From the framework of multi-view probabilistic models discussed in the previous section, we can see that the learned representation is usually associated with shared latent variables, specifically with their posterior distribution given the multi-view observed input. Further, this posterior distribution often becomes complicated and intractable if the designed models have hierarchical structures. It has to resort to sampling or approximate inference techniques, which need to endure the associated computational and approximation error cost. In addition, a posterior distribution over shared latent variables is not yet a reasonable feature vector that can be directly fed to the classifier. Therefore, if we intend to obtain stable deterministic feature values, an alternative non-probabilistic multi-view embedding learning paradigm is the directly parameterized feature or representation functions. The common perspective between these methods is that they learn a direct encoding for multi-view input. Consequently, in this section we review the related work from this perspective.

3.2.1 Partial Least Squares

Partial Least Squares (PLS) [189–191] is a wide class of methods for modeling relations between sets of observed variables. It has been a popular tool for regression and classification as well as dimensionality reduction, especially in the field of chemometrics [192, 193]. The underlying assumption of all PLS methods is that the observed data are generated by a process which is driven by a small number of latent variables. In particular, PLS creates orthogonal latent vectors by maximizing the covariance between different sets of variables.

Given a pair of datasets $X = [x_1, \dots, x_n] \in \mathbb{R}^{d_x \times n}$ and $Y = [y_1, \dots, y_n] \in \mathbb{R}^{d_y \times n}$, a k -dimensional PLS solution can

be parameterized by a pair of matrices $W_x \in \mathbb{R}^{d_x \times k}$ and $W_y \in \mathbb{R}^{d_y \times k}$ [47]. The PLS problem can now be expressed as:

$$\begin{aligned} \max_{W_x, W_y} & \text{tr}(W_x^\top C_{xy} W_y) \\ \text{s.t.} & W_x^\top W_x = I, W_y^\top W_y = I. \end{aligned} \quad (31)$$

It can be shown that the columns of the optimal W_x and W_y correspond to the singular vectors of covariance matrix $C_{xy} = \mathbb{E}[xy^\top]$. Like the CCA objective, PLS is also an optimization of an expectation subject to fixed constraints.

In essence, CCA finds the directions of maximum correlation while PLS finds the directions of maximum covariance. Covariance and correlation are two different statistical measures for describing how variables covary. It has been shown that there is close connections between PLS and CCA in several aspects [190, 193]. Guo and Mu [194] investigate the CCA based methods, including linear CCA, regularized CCA, and kernel CCA, and compare to the PLS models in solving the joint estimation problem. In particular, they provide a consistent ranking of the above methods in estimating age, gender, and ethnicity.

Further, Li et al. [195] introduce a least square form of PLS, called cross-modal factor analysis (CFA). CFA aims at finding orthogonal transformation matrices W_x and W_y that minimize the following expression:

$$\begin{aligned} & \|X^\top W_x - Y^\top W_y\|_F^2 \\ \text{subject to: } & W_x^\top W_x = I, W_y^\top W_y = I \end{aligned} \quad (32)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. It can be easily verified that the above optimization problem in Eq.(32) has the same solution as that of PLS. Several extensions of CFA are presented by incorporating non-linearity and supervised information [196–198].

Among the variants of PLS, Orthonormalized PLS (OPLS) [199–201] is a popular representative. OPLS does not consider the variance of one of the two views and has been shown to be competitive with the other PLS variants. In particular, OPLS computes the orthogonal transform vectors for X by solving the following optimization problem:

$$\begin{aligned} \max_{W_x} & \text{tr}(W_x^\top XY^\top YX^\top W_x) \\ \text{s.t.} & W_x^\top XX^\top W_x = I. \end{aligned} \quad (33)$$

It can be shown that the columns of the optimal W are given by the eigenvectors of the following generalized eigenvalue problem:

$$XY^\top YX^\top w_x = \eta XX^\top w_x \quad (34)$$

From the above computation process of the projection of X in OPLS, we can see that the obtained transformation matrix W_x is directed by the information encoded in Y . This is especially useful in the supervised learning context in which OPLS is applied for supervised dimensionality reduction.

Connections between CCA and OPLS also have been studied in the past decade. Based on the unified framework for CCA and PLS in [190], CCA and OPLS also can be considered as special cases of this framework by selecting different values of regularization parameters. Further, Sun et al. [202] investigate the equivalence relationship between CCA and OPLS and show that the difference between the CCA solution and the OPLS solution is a mere orthogonal transformation.

Another commonly used variant of PLS is PLS Regression (PLSR), which has been proven to be an efficient robust method to do quantity analysis for signal processing [192, 203]. PLSR uses the two-block predictive PLS model to capture the relationship between X and Y [192]. It is especially true for the case that one set of variables are the target variables of a regression problem as it comprises regression tasks as well as dimensionality reduction techniques. By following the way of probabilistic CCA, Li et al. [204] present a probabilistic PLS to explain PLSR from a probabilistic viewpoint and describes the physical meaning of PLSR model.

3.2.2 Multi-View Discriminant Analysis

CCA and its kernel extensions are used to match sets of features by maximizing within-set correlation and minimizing between-set correlation, which are unsupervised scenarios. A number of multi-view analysis approaches [205–207] have also been proposed by considering the scenarios in which the data have multiple different views, along with supervised information.

Given a pair of datasets $X = [x_1, \dots, x_n]$ and $Y = [y_1, \dots, y_n]$ with the corresponding labels $l = (l_1, \dots, l_n)$, the two-view regularized Fisher discriminant analysis (FDA) [205] chooses two sets of weights w_x and w_y to solve the following optimization problem,

$$\rho = \frac{w_x^\top X l l^\top Y^\top w_y}{\sqrt{(w_x^\top X B X^\top w_x + \mu \|w_x\|^2) \cdot (w_y^\top Y B Y^\top w_y + \mu \|w_y\|^2)}} \quad (35)$$

where w_x and w_y are the weight vectors for each view, and B is a matrix incorporating the label information and the balance of the datasets. Since the equation is not affected by rescaling of w_x or w_y , the optimization is subjected to the following constraints

$$\begin{aligned} w_x^\top X B X^\top w_x + \mu \|w_x\|^2 &= 1 \\ w_y^\top Y B Y^\top w_y + \mu \|w_y\|^2 &= 1 \end{aligned} \quad (36)$$

By introducing the Lagrange multipliers λ_x and λ_y , the corresponding Lagrangian for this optimization is:

$$\begin{aligned} \mathcal{L} = & w_x^\top X l l^\top Y^\top w_y - \frac{\lambda_x}{2} (w_x^\top X B X^\top w_x + \mu \|w_x\|^2 - 1) \\ & - \frac{\lambda_y}{2} (w_y^\top Y B Y^\top w_y + \mu \|w_y\|^2 - 1), \end{aligned} \quad (37)$$

which can be solved by differentiation with respect to the weight vectors w_x and w_y .

By substituting the two weight vectors with $w_x = X\alpha$ and $w_y = Y\beta$, and a shared regularization parameter κ for both sets of weights, we obtain

$$\begin{aligned} \rho = & \frac{\alpha^\top X^\top X l l^\top Y^\top Y \beta}{\sqrt{(\alpha^\top X^\top X B X^\top X \alpha + \kappa \|w_x\|^2)}} \\ & \cdot \frac{1}{\sqrt{(\beta^\top Y^\top Y B Y^\top Y \beta + \kappa \|w_y\|^2)}} \end{aligned} \quad (38)$$

Then the kernel form of it is:

$$\rho = \frac{\alpha^\top K_x l l^\top K_y \beta}{\sqrt{(\alpha^\top K_x B K_x \alpha + \kappa \alpha^\top K_x \alpha) \cdot (\beta^\top K_y B K_y \beta + \kappa \beta^\top K_y \beta)}} \quad (39)$$

As in the primal form, the equation is not affected by rescaling w_x or w_y ; the optimization is subjected to the following constraints

$$\begin{aligned} \alpha^\top K_x B K_x \alpha + \kappa \alpha^\top K_x \alpha &= 1 \\ \beta^\top K_y B K_y \beta + \kappa \beta^\top K_y \beta &= 1 \end{aligned} \quad (40)$$

where K_x and K_y are the corresponding kernels for the two views. The corresponding Lagrangian for this optimization can be written as

$$\begin{aligned} \mathcal{L} = & \alpha^\top K_x l l^\top K_y \beta - \frac{\lambda_x}{2} \left(\alpha^\top K_x B K_x \alpha + \kappa \alpha^\top K_x \alpha - 1 \right) \\ & - \frac{\lambda_y}{2} \left(\beta^\top K_y B K_y \beta + \kappa \beta^\top K_y \beta - 1 \right) \end{aligned} \quad (41)$$

By differentiation with respect to the weight vectors α and β , we can solve the above problem. In addition, based on the approach outlined in [208], both regularized two-view FDA and its kernel extension can be casted as equivalent disciplined convex optimization problems. Consequently, Diethe et al. [206] introduce Multi-view Fisher Discriminant Analysis (MFDA) that learns classifiers in multiple views, by minimizing the variance of the data along the projection while maximizing the distance between the average outputs for classes over all of the views.

However, MFDA can only be used for binary classification problems. In [207], Kan et al. propose a Multi-view Discriminant Analysis (MvDA) method, which seeks for a discriminant common space by maximizing the between-class and minimizing the within-class variations, across all the views. Later based on bilinear models [209] and general graph embedding framework [210], Sharma et al. [211] introduce Generalized Multi-view Analysis (GMA). As an instance of GMA, Generalized Multi-view Linear Discriminant Analysis (GMLDA) finds a set of projection directions in each view that tries to separate different contents' class means and unify different views of the same class in the common subspace.

3.2.3 Cross-Modal Ranking

Motivated by incorporating ranking information into multi-modal embedding learning, cross-modal ranking has attracted much attention in the literature [212–214]. Bai et al. [212] present a supervised semantic indexing (SSI) model which defines a class of non-linear models that are discriminatively trained to map multi-modal input pairs into ranking scores. Unlike CCA and CFA, SSI is trained from a supervised way on the ranking task of interest so that it retains more discriminative information.

In particular, SSI attempts to learn a similarity function $f(q, d)$ between a text query q and an image d according to a pre-defined ranking loss. The learned function f directly maps each text-image pair to a ranking score based on their semantic relevance. Given a text query $q \in \mathbb{R}^m$ and an image $d \in \mathbb{R}^n$, SSI intends to find a linear scoring function to measure the relevance of d given q :

$$f(q, d) = q^\top W d = \sum_{i=1}^m \sum_{j=1}^n q_i W_{ij} d_j \quad (42)$$

where $f(q, d)$ is defined as the score between the query q and the image d , and the parameter $W \in \mathbb{R}^{m \times n}$ captures the correspondence between the two different modalities of the data: W_{ij} represents the correlation between the i -th dimension of the text space and the j -th dimension of the image space. Note that this way of embedding allows both positive and negative correlations between different modalities since both positive and negative values are allowed in W_{ij} .

Given the similarity function in Eq.(42) and a set of tuples, where each tuple contains a query q , a relevant image d^+ and an irrelevant image d^- , SSI attempts to choose the scoring function $f(q, d)$ such that $f(q, d^+) > f(q, d^-)$, expressing that d^+ should be ranked higher than d^- . For this purpose, SSI exploits the

margin ranking loss [215] which has already been widely used in information retrieval, and minimizes:

$$\sum_{(q, d^+, d^-)} \max(0, 1 - q^\top W d^+ + q^\top W d^-) \quad (43)$$

This optimization problem can be solved through stochastic gradient descent [216],

$$\begin{aligned} W \leftarrow W + \lambda \left(q(d^+)^\top - q(d^-)^\top \right), \\ \text{if } 1 - q^\top W d^+ + q^\top W d^- > 0 \end{aligned} \quad (44)$$

In fact, this method is a special margin ranking perceptron [217], which has been shown to be equivalent to SVM [218]. In contrast to classical SVM, stochastic training is highly scalable and is easy to implement for millions of training examples. However, dealing with the models on all the pairs of multi-modalities input features is still computationally challenging. Thus, SSI also proposes several improvements to the above basic model for addressing this issue, including low-rank representation, sparsification, and correlated feature hashing. In the following, we take the low-rank constrained SSI model for example.

Motivated by dealing with the model memory, speed, and capacity control issues, a low-rank configuration for W is set up with $W = U^\top V$ in Eq.(42), which results in the following low-rank scoring function:

$$f(q, d) = q^\top U^\top V d = (Uq)^\top V d \quad (45)$$

where $U \in \mathbb{R}^{k \times m}$ is used to map the query text q from the m -dimensional text space to the k -dimensional subspace and $V \in \mathbb{R}^{k \times n}$ is used to map the retrieved image d from the n -dimensional image space to the k -dimensional subspace. Consequently, the text query and the retrieved image are mapped to a common k -dimensional latent semantic space, and then their similarity is measured with a cosine similarity score in the k -dimensional space.

When the W matrix is constrained in the low-rank way, training is performed in a similar way as before. Gradient steps are still applied to optimize the parameters U and V :

$$\begin{aligned} U \leftarrow U + \lambda V(d^+ - d^-)q^\top, & \text{ if } 1 - f(q, d^+) + f(q, d^-) > 0 \\ V \leftarrow V + \lambda Uq(d^+ - d^-)^\top, & \text{ if } 1 - f(q, d^+) + f(q, d^-) > 0 \end{aligned} \quad (46)$$

Inspired by the success of the SSI models for cross-media retrieval tasks, a variety of ‘‘cross-modal ranking’’ methods have been proposed to achieve better generalization performance. To exploit the advantage of online learning of kernel-based classifiers, Grangier and Bengio [214] propose a discriminative cross-modal ranking model called Passive-Aggressive Model for Image Retrieval (PAMIR), which not only adopts a learning criterion related to the final retrieval performance, but also considers different image kernels. In contrast to SSI, the parameterization of PAMIR can benefit from effective kernels for image comparison. Further, several recent efforts [219–221] incorporate latent semantic representation learning into cross-modal ranking. Lu et al. [219] propose a Latent Semantic Cross-Modal Ranking (LSCMR) algorithm, which regards the cross-modal retrieval as a list-wise ranking problem and optimizes the list-wise ranking loss with a low-rank embedding setup. In particular, the low-rank embedding space is discriminatively learned by the structural large margin method. Moreover, Wu et al. [220] introduce a Bi-directional Cross-media Semantic Representation Model, which incorporates

the bi-directional list-wise ranking loss into the learning of cross-modal embeddings to take advantage of bi-directional ranking examples.

3.2.4 Cross-Modal Hashing

CCA and its extensions have been widely applied to conduct cross-view similarity search [25, 38, 95]. A promising way to speed up the cross-view similarity search is the hashing technique which makes a tradeoff between accuracy and efficiency. Hence, several multi-modal hashing methods have been proposed for fast similarity search in multi-modal data [222–231]. The principle of the multi-modal hashing methods is to map the high dimensional multi-modal data into a common hash codes so that similar cross-modal data objects have the same or similar hash codes.

Bronstein et al. [222] propose a hashing-based model, called cross-modal similarity sensitive hashing (CMSSH), which approaches the cross-modality similarity learning problem by embedding the multi-modal data into a common metric space. The similarity is parameterized by the embedding itself. The goal of cross-modality similarity learning is to construct the similarity function between points from different spaces, $X \in \mathbb{R}^{d_1}$ and $Y \in \mathbb{R}^{d_2}$. Assume that the unknown binary similarity function is $s : X \times Y \rightarrow \{\pm 1\}$; the classical cross-modality similarity learning aims at finding a binary similarity function \hat{s} on $X \times Y$ approximating s . Recent work attempts to solve the problem of cross-modality similarity learning as an multi-view representation learning problem.

In particular, CMSSH proposes to construct two maps: $\xi : X \rightarrow \mathbb{H}^n$ and $\eta : Y \rightarrow \mathbb{H}^n$, where \mathbb{H}^n denotes the n -dimensional Hamming space. Such mappings encode the multi-modal data into two n -bit binary strings so that $d_{\mathbb{H}^n}(\xi(x), \eta(y))$ is small for $s(x, y) = +1$ and large for $s(x, y) = -1$ with a high probability. Consequently, this hamming embedding can be interpreted as cross-modal similarity-sensitive hashing, under which positive pairs have a high collision probability, while negative pairs are unlikely to collide. Such hashing also acts as a way of multi-modal dimensionality reduction when $d_1, d_2 \gg n$.

The n -dimensional Hamming embedding for X can be thought of as a vector $\xi(x) = (\xi_1(x), \dots, \xi_n(x))$ of binary embeddings of the form

$$\xi_i(x) = \begin{cases} 0 & \text{if } f_i(x) \leq 0, \\ 1 & \text{if } f_i(x) > 0, \end{cases} \quad (47)$$

parameterized by a projection $f_i : X \rightarrow \mathbb{R}$. Similarly, η_i is a binary map parameterized by projection $g_i : Y \rightarrow \mathbb{R}$.

Following the greedy approach [232], the Hamming metric can be constructed sequentially as a superposition of weak binary classifiers on pairs of data points,

$$\begin{aligned} h_i(x, y) &= \begin{cases} +1 & \text{if } \xi_i(x) = \eta_i(y), \\ -1 & \text{otherwise,} \end{cases} \\ &= (2\xi_i(x) - 1)(2\eta_i(y) - 1), \end{aligned} \quad (48)$$

Here, a simple strategy for the maps is affine projection, such as $f_i(x) = p_i^\top x + a_i$ and $g_i(y) = q_i^\top y + b_i$. It can be extended to complex projections easily.

Observing the resemblance with sequentially binary classifiers, boosted cross-modality similarity learning algorithms are introduced based on the standard AdaBoost procedure [233]. CMSSH has shown its utility and efficiency in several multi-view learning applications including cross-representation shape retrieval and alignment of multi-modal medical images.

However, CMSSH only considers the inter-view correlation but ignores the intra-view similarity [234]. Kumar and Udupa [223] extend Spectral Hashing [235] from the single view setting to the multi-view scenario and present cross view hashing (CVH), which attempts to find hash functions that map similar objects to similar codewords over all the views so that inter-view and intra-view similarities are both preserved. Gong and Lazebnik [224] combine iterative quantization with CCA to exploit cross-modal embeddings for learning similarity preserving binary codes. Consequently, Zhen and Yang [225] present co-regularized hashing (CRH) for multi-modal data based on a boosted co-regularization framework. The hash functions for each bit of the hash codes are learned by solving DC (difference of convex functions) programs, while the learning for multiple bits is performed via a boosting procedure. Later Zhu et al. [226] introduce linear cross-modal hashing (LCMH), which is adopts a “two-stage” strategy to learn cross-view hashing functions. The data within each modality are first encoded into a low-rank representation using the idea of anchor graph [236] and then hash functions for each modality are learned to map each modality’s low-rank space into a shared Hamming space. Song et al. [237] introduce an inter-media hashing (IMH) model by jointly capturing inter-media and intra-media consistency.

Further, recent work has incorporated sparse coding into multi-modal hashing. Wu et al. [184] present sparse multi-modal hashing by introducing dictionary learning into multi-view hashing, where the intra-modality and inter-modality similarities are modeled with a hypergraph. In particular, the learned dictionary for each modality is considered as the hash functions and the sparse code of each data point over its corresponding dictionary equals to the acquired hash code. Inspired by this effectiveness of incorporating sparse learning into multi-view hashing, Yu et al. [234] additionally exploit discriminative information when learning coupled dictionaries to make the shared dictionary space interpretable. Zhou et al. [238] present a latent semantic sparse hashing model by unifying sparse coding and matrix factorization.

Based on taking both hashing function learning and quantization of hash codes into consideration, Wu et al. [239] present quantized correlation hashing (QCH). QCH adopts a joint learning scheme that consolidates the minimization of quantization loss and maximization of domain correlation into a single objective function. In this way, the quantizer and binary code learning are simultaneously optimized so that the quantizer is more fit for cross-modality data search.

4 DEEP METHODS ON MULTI-VIEW REPRESENTATION LEARNING

Inspired by the success of deep neural networks [5, 6, 124], a variety of deep multi-view feature learning methods have been proposed to capture the high-level correlation between multi-view data. In this section, we continue to review the deep multi-view representation models from probabilistic and directly embedding perspectives.

4.1 Probabilistic Models

Restricted Boltzmann Machines (RBM) [240] is an undirected graphical model that can learn the distribution of training data. The model consists of stochastic visible units $\mathbf{v} \in \{0, 1\}^{d_v}$ and

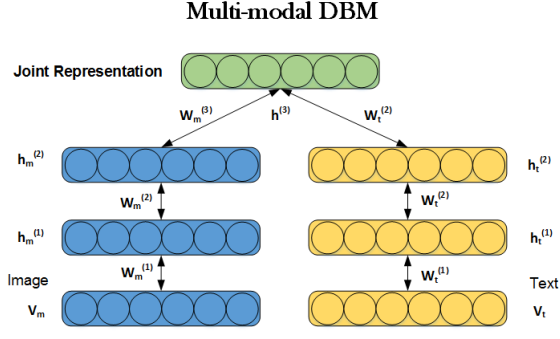


Fig. 4. The graphical model of deep multi-modal RBM (adapted from [17]), which models the joint distribution over image and text inputs.

stochastic hidden units $\mathbf{h} \in \{0, 1\}^{d_h}$, which seeks to minimize the following energy function $E : \{0, 1\}^{d_v+d_h} \rightarrow \mathbb{R}$:

$$E(\mathbf{v}, \mathbf{h}; \theta) = - \sum_{i=1}^{d_v} \sum_{j=1}^{d_h} v_i W_{ij} h_j - \sum_{i=1}^{d_v} b_i v_i - \sum_{j=1}^{d_h} a_j h_j \quad (49)$$

where $\theta = \{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$ are the model parameters. Consequently, the joint distribution over the visible and hidden units is defined by:

$$P(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{Z(\theta)} \exp(-E(\mathbf{v}, \mathbf{h}; \theta)). \quad (50)$$

When considering modeling visible real-valued or sparse count data, this RBM can be easily extended to corresponding variants, e.g., Gaussian RBM [185] and replicated softmax RBM [241].

A deep Boltzmann machine (DBM) is a generative network of stochastic binary units. It consists of a set of visible units $\mathbf{v} \in \{0, 1\}^{d_v}$, and a sequence of layers of hidden units $\mathbf{h}^{(1)} \in \{0, 1\}^{d_{h1}}, \mathbf{h}^{(2)} \in \{0, 1\}^{d_{h2}}, \dots, \mathbf{h}^{(L)} \in \{0, 1\}^{d_{hL}}$. Here only connections between hidden units are allowed in adjacent layers. Let us take a DBM with two hidden layers for example. By ignoring bias terms, the energy of the joint configuration $\{\mathbf{v}, \mathbf{h}\}$ is defined as

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\mathbf{v}^\top \mathbf{W}^{(1)} \mathbf{h}^{(1)} - \mathbf{h}^{(1)\top} \mathbf{W}^{(2)} \mathbf{h}^{(2)} \quad (51)$$

where $\mathbf{h} = \{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}\}$ represents the set of hidden units, and $\theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}\}$ are the model parameters that denote visible-to-hidden and hidden-to-hidden symmetric interaction terms. Further, this binary-to-binary DBM can also be easily extended to modeling dense real-valued or sparse count data.

By extending the setup of the DBM, Srivastava and Salakhutdinov [17] propose a deep multi-modal RBM to model the relationship between image and text. In particular, each data modality is modeled using a separate two-layer DBM and then an additional layer of binary hidden units on top of them is added to learn the shared representation.

Let $\mathbf{v}_m \in \mathbb{R}^{d_{v_m}}$ denote an image input and $\mathbf{v}_t \in \mathbb{R}^{d_{v_t}}$ denote a text input. By ignoring bias terms on the hidden units for clarity, the distribution of \mathbf{v}_m in the image-specific two-layer DBM is

given as follows:

$$\begin{aligned} P(\mathbf{v}_m; \theta) &= \sum_{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}} P(\mathbf{v}_m, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}; \theta) \\ &= \frac{1}{Z(\theta)} \sum_{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}} \exp\left(- \sum_{i=1}^{d_{v_m}} \frac{(v_{mi} - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^{d_{v_m}} \sum_{j=1}^{d_{h1}} \frac{v_{mi}}{\sigma_i} W_{ij}^{(1)} \right. \\ &\quad \left. + \sum_{j=1}^{d_{h1}} \sum_{l=1}^{d_{h2}} h_j^{(1)} W_{jl}^{(2)} h_l^{(2)}\right). \end{aligned} \quad (52)$$

Similarly, the text-specific two-layer DBM can also be defined by combining a replicated softmax model with a binary RBM.

Consequently, the deep multi-modal DBM has been presented by combining the image-specific and text specific two-layer DBM with an additional layer of binary hidden units on top of them. The particular graphical model is shown in Figure 4. The joint distribution over the multi-modal input can be written as:

$$\begin{aligned} P(\mathbf{v}^m, \mathbf{v}^t; \theta) &= \sum_{\mathbf{h}_m^{(2)}, \mathbf{h}_t^{(2)}, \mathbf{h}^{(3)}} P(\mathbf{h}_m^{(2)}, \mathbf{h}_t^{(2)}, \mathbf{h}^{(3)}) \left(\sum_{\mathbf{h}_m^{(1)}} P(\mathbf{v}_m, \right. \\ &\quad \left. \mathbf{h}_m^{(1)}, \mathbf{h}_m^{(2)}) \right) \left(\sum_{\mathbf{h}_t^{(1)}} P(\mathbf{v}_t, \mathbf{h}_t^{(1)}, \mathbf{h}_t^{(2)}) \right) \end{aligned} \quad (53)$$

Like RBM, exact maximum likelihood learning in this model is also intractable, while efficient approximate learning can be implemented by using mean-field inference to estimate data-dependent expectations, and an MCMC based stochastic approximation procedure to approximate the model's expected sufficient statistics [6].

Multi-modal DBM has been widely used for multi-view representation learning [242–244]. Hu et al. [243] employ the multi-modal DBM to learn joint representation for predicting answers in cQA portal. Ge et al. [244] apply the multi-modal RBM to determining information trustworthiness, in which the learned joint representation denotes the consistent latent reasons that underline users' ratings from multiple sources. Pang and Ngo [245] propose to learn a joint density model for emotion prediction in user-generated videos with a deep multi-modal Boltzmann machine. This multi-modal DBM is exploited to model the joint distribution over visual, auditory, and textual features. Here Gaussian RBM is used to model the distributions over the visual and auditory features, and replicated softmax topic model is applied for mining the textual features.

Further, Sohn et al. [246] investigate an improved multi-modal RBM framework via minimizing the variation of information between data modalities through the shared latent representation. Recently, Ehrlich et al. [247] present a multi-task multi-modal RBM (MTM-RBM) approach for facial attribute classification. In particular, a multi-task RBM is proposed by extending the formulation of discriminative RBM [248] to account for multiple tasks. And then multi-task RBM is naturally extended to MTM-RBM by combining a collection of unimodal MT-RBM, one for each visible modality. Unlike the typical multi-modal RBM, the learned joint feature representation of MTM-RBM enables interaction between different tasks so that it is suited for multiple attribute classification.

4.2 Directly Deep Parametric Embedding from Multi-View Inputs to Representation

With the development of deep neural networks, many shallow multi-view embedding methods have been extended to deep ones,

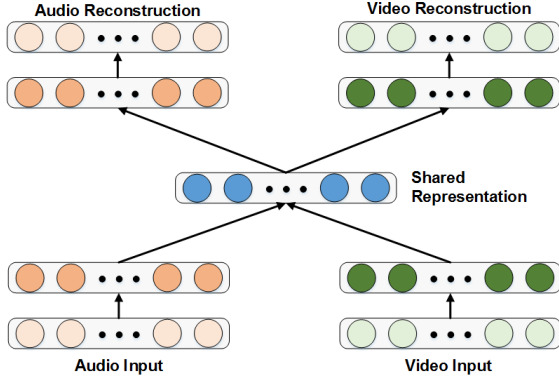


Fig. 5. The bimodal deep autoencoder (adapted from [18]).

respectively. In this section, we review the deep multi-view feature learning methods from the directly parametric embedding perspective.

4.2.1 Multi-Modal Deep Autoencoders

Although Multi-modal RBMs achieve great success in learning a shared representation, there are still limitations with them, e.g., there is no explicit objective for the models to discover correlations across the modalities such that some hidden units are tuned only for one modality while others are tuned only for the other. Multi-modal deep autoencoders [18, 249–251] gradually become good alternatives for learning a shared representation between modalities due to the sufficient flexibility of their objectives.

Inspired by denoising autoencoders [5], Ngiam et al. [18] propose to extract shared representations via training a bimodal deep autoencoder (Figure 5) using an augmented but noisy dataset with additional examples that have only a single modality as input. The key idea is to use greedy layer-wise training with an extension to RBMs with sparsity [252] followed by fine-tuning.

Further, Feng et al. [19] propose a correspondence autoencoder (Corr-AE) via constructing correlations between hidden representations of two uni-modal deep autoencoders. The details of the architecture of the basic Corr-AE is shown in Figure 6. As illustrated in Figure 6, this Corr-AE architecture consists of two subnetworks (each a basic deep autoencoder) that are connected by a predefined similarity measure on a specific coder layer. Each subnetwork in the Corr-AE serves for each modality.

Let $f(x; W_f)$ and $g(y; W_g)$ denote the mapping from the inputs $\{X, Y\}$ to the code layers, respectively. $\theta = \{W_f, W_g\}$ denotes the weight parameters in these two networks. Here the similarity measure between the i -th pair of image feature x_i and the given text feature y_i is defined as follows:

$$C(x_i, y_i; \theta) = \|f(x_i; W_f) - g(y_i; W_g)\|_2^2 \quad (54)$$

where f and g are logistic activation functions.

Consequently, the loss function on any pair of inputs is then defined as follows:

$$L(x_i, y_i; \theta) = (1 - \alpha) (L_I(x_i, y_i; \theta) + L_T(x_i, y_i; \theta)) + \alpha L_C(x_i, y_i; \theta) \quad (55)$$

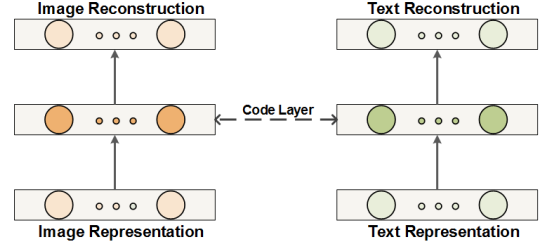


Fig. 6. The correspondence deep autoencoder (adapted from [19]).

where

$$\begin{aligned} L_I(x_i, y_i; \theta) &= \|x_i - \hat{x}_i\|_2^2 \\ L_T(x_i, y_i; \theta) &= \|y_i - \hat{y}_i\|_2^2 \\ L_C(x_i, y_i; \theta) &= C(x_i, y_i; \theta) \end{aligned}$$

L_I and L_T are the losses caused by data reconstruction errors for the given inputs of the two subnetworks, specifically image and text modality. L_C is the correlation loss and α is a parameter for trade-off between two groups of objectives. \hat{x}_i and \hat{y}_i are the reconstruction data from x_i and y_i respectively.

Overall, optimizing the objective in Eq.(55) enables Corr-AE to learn similar representations from bimodal features. Besides, based on two other multi-modal autoencoders [18], Corr-AE is extended to two other correspondence deep models, called Corr-Cross-AE and Corr-Full-AE.

Consequently, Silberer and Lapata train stacked multi-modal autoencoder with semi-supervised objective to learn grounded meaning representations. Wang et al. [253] propose an effective mapping mechanism based on stacked autoencoder for multi-modal retrieval. In particular, the cross-modal mapping functions are learned by optimizing an objective function which captures both intra-modal and inter-modal semantic relationships of data from heterogeneous sources.

Further, based on CCA and deep autoencoders, Wang et al. [20] propose deep canonically correlated autoencoders that also consist of two autoencoders and optimize the combination of canonical correlations between the learned bottleneck representations and the reconstruction errors of the autoencoders. Intuitively, this is the same principle as that of Corr-AE. The difference between them relies on the different similarity measures.

Recently, Rastegar et al. [254] suggest to exploit the cross weights between representations of modalities for gradually learning interactions of the modalities in a multi-modal deep autoencoder network. Theoretical analysis shows that considering these interactions in deep network manner (from low to high level) provides more intra-modality information. As opposed to the existing deep multi-modal autoencoders, this approach attempts to reconstruct the representation of each modality at a given level, with the representation of the other modalities in the previous layer.

4.2.2 Deep Cross-View Embedding Models

Deep cross-view embedding models have become increasingly popular in the applications including cross-media retrieval [255, 256] and multi-modal distributional semantic learning [257, 258]. Frome et al. [255] propose a deep visual-semantic embedding model (DeViSE), which connects two deep neural networks by a cross-modal mapping. As shown in Figure 7, DeVISE is first

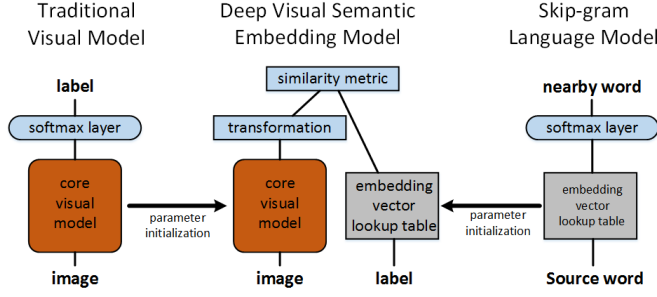


Fig. 7. The DeVISE model (adapted from [255]), which is initialized with parameters pre-trained at the lower layers of the visual object categorization network and the skip-gram language model.

initialized with a pre-trained neural network language model [259] and a pre-trained deep visual-semantic model [124]. Consequently, a linear transformation is exploited to map the representation at the top of the core visual model into the learned dense vector representations by the neural language model.

Following the setup of loss function in [212], DeVISE employs a combination of dot-product similarity and hinge rank loss so that the model has the ability of producing a higher dot-product similarity between the visual model output and the vector representation of the correct label than between the visual output and the other randomly chosen text terms. The per training example hinge rank loss is defined as follows:

$$\sum_{j \neq \text{label}} \max \left[0, \text{margin} - \vec{t}_{\text{label}} M \vec{v}(\text{image}) + \vec{t}_j M \vec{v}(\text{image}) \right] \quad (56)$$

where $\vec{v}(\text{image})$ is a column vector denoting the output of the top layer of the core visual network for the given image, M is the mapping matrix of the linear transformation layer, \vec{t}_{label} is a row vector denoting the learned embedding vector for the provided text label, and \vec{t}_j are the embeddings of the other text terms. This DeVISE model is trained by asynchronous stochastic gradient descent on a distributed computing platform [260].

Inspired by the success of DeVISE, Norouzi et al. [261] propose a convex combination of semantic embedding model (ConSE) for mapping images into continuous semantic embedding spaces. Unlike DeVISE, this ConSE model keeps the softmax layer of the convolutional net intact. Given a test image, ConSE simply runs the convolutional classifier and considers the convex combination of the semantic embedding vectors from the top T predictions as its corresponding semantic embedding vector. Further, Fang et al. [262] develop a deep multi-modal similarity model that learns two neural networks to map images and text fragments to a common vector representation.

With the development of multi-modal distributional semantic models [263–265], deep cross-modal mapping is naturally exploited to learn the improved multimodal distributed semantic representation. Lazaridou et al. [258] introduce multimodal skip-gram models to extend the skip-gram model of [259] by taking visual information into account. In this extension, for a subset of the target words, relevant visual evidence from natural images is presented together with the corpus contexts. In particular, the model is designed to encourage the propagation of visual information to all words via a deep cross-modal ranking so that it can improve image labeling and retrieval in the zero-shot setup,

where the test concepts are never seen during model training. Further, Dong et al. [256] present Word2VisualVec, a deep neural network architecture that learns to predict a deep visual encoding of textual input, and thus enables cross-media retrieval in a visual space.

In addition, Xu et al. [266] propose a unified framework that jointly models video and the corresponding text sentence. In this joint architecture, the goal is to learn a function $f(\mathcal{V}) : \mathcal{V} \rightarrow \mathcal{T}$, where \mathcal{V} represents the low-level features extracted from video, and \mathcal{T} is the high-level text description of the video. A joint model \mathcal{P} is designed to connect these two levels of information. It consists of three parts: compositional language model $M_L : T \rightarrow T_f$, deep video model $M_V : V \rightarrow V_f$, and a joint embedding model $E(V_f, T_f)$, such that

$$\mathcal{P} : M_V(V) \rightarrow V_f \leftrightarrow E(V_f, T_f) \leftrightarrow T_f \leftarrow M_L(T) \quad (57)$$

where V_f and T_f are the output of the deep video model and the compositional language model, respectively. In this joint embedding model, the distance of the outputs of the deep video model and the compositional language model in the joint space is minimized to make them alignment.

Along this line of compositional cross-modal semantic representation learning, Yu et al. [267] present a unified deep neural network model called cross space mapping, in which the image and query are mapped to a common vector space via a convolution part and a query-embedding part, respectively. Jiang et al. [268] also introduce a deep cross-modal retrieval method, called deep compositional cross-modal learning to rank (C^2MLR). C^2MLR considers learning a multi-modal embedding from the perspective of optimizing a pairwise ranking problem while enhancing both local alignment and global alignment. In addition, Wei et al. [269] introduce a deep semantic matching method, in which two independent deep networks are learned to map image and text into a common semantic space with a high level abstraction. Wu et al. [270] consider learning multi-modal representation from the perspective of encoding the explicit/implicit relevance relationship between the vertices in the click graph, in which vertices are images/text queries and edges indicate the clicks between an image and a query.

4.2.3 Deep Multi-Modal Hashing

Motivated by the recent advance of deep learning for multi-modal data [17, 18, 255], deep neural networks have gradually been exploited to learn hash functions. Kang et al. [271] introduce a deep multi-view hashing method in which each layer of hidden nodes consists of view-specific and shared hidden nodes to learn individual and shared hidden spaces from multiple views of data.

Consequently, Zhuang et al. [272] propose a cross-media hashing approach based on a correspondence multi-modal neural network, referred as Cross-Media Neural Network Hashing (CMNNH). The network structure of CMNNH can be considered as a combination of two modality-specific neural networks with an additional correspondence layer as shown in Figure 8.

Denote the two neural networks corresponding to multi-modal input $\{X, Y\}$ as NN^x and NN^y , respectively. For each $x \in X$ ($y \in Y$ in a similar way), x is forwarded layer-by-layer through NN^x to generate the representation of each layer. The l^{th} layer takes h^l as the input and uses a projection function to transform it to h^{l+1} in the next layer:

$$h^{l+1} = f^l(W^l h^l) \quad (58)$$

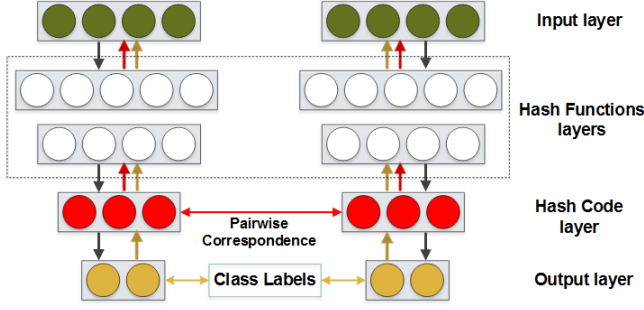


Fig. 8. The illustration of cross-media neural network hashing architecture (adapted from [272]).

where h^l and h^{l+1} are the feature representations in the l^{th} and $l+1^{th}$ hidden layer, respectively. W^l is the projection matrix and f^l is the activation function.

From the perspective of the hash function H^x , it takes x as input, forwards x to the hash code layer, and outputs the k -dimensional binary hash codes:

$$H^x(x) = \text{sign}(h^x) \quad (59)$$

where $h^x \in \mathbb{R}^k$ is a k -dimensional, real-value vector of the hash code layer, and it can be converted to a binary hash code by the sign function. The hash function H^y is formulated by analogy. However, the sign function is not differentiable, and thus is hard to optimize directly. By following the setup of most of the existing hashing methods [222, 235], the sign function can be removed at the hash function learning stage and added at the testing stage.

Further, to preserve the inter-modal pair-wise correspondence between NN^x and NN^y , a loss function is defined based on the least square error of pairwise inter-modal correspondence:

$$\mathcal{L}_1(x, y) = \frac{1}{2} \|\tilde{h}^x - \tilde{h}^y\|_F^2 \quad (60)$$

Besides, to preserve the intra-modal discriminative capability, softmax regression function is employed as the loss function on the output layer as follows:

$$\mathcal{L}_2(x, y, c) = \text{KL}(\hat{t}_x, t) + \text{KL}(\hat{t}_y, t) \quad (61)$$

where t is the class label for x and y , \hat{t}_x and \hat{t}_y are the estimated values by NN^x and NN^y , respectively, and $\text{KL}(\cdot)$ is the KL -divergence function. Further, the two loss functions for all the data points in X and Y are integrated and the overall loss function of CMNNH is minimized as follows:

$$\mathcal{J} = \sum_{i=1}^n \mathcal{L}_1(x_i, y_i) + \lambda \sum_{i=1}^n \mathcal{L}_2(x_i, y_i, t_i) \quad (62)$$

where λ is a hyper-parameter to balance the two losses. The training of CMNNH is conducted by the classical back-propagation method.

To solve the redundancy problem of multi-modal hashing representation learning by deep models, Wang et al. [273] introduce a deep multi-modal hashing model with orthogonal regularization for mapping multi-modal data into a common hamming space. On one hand, this model captures intra-modality and inter-modality correlations to extract useful information from multi-modal data; on the other hand, to address the redundancy problem, orthogonal regularizers are imposed on the weighting matrices.

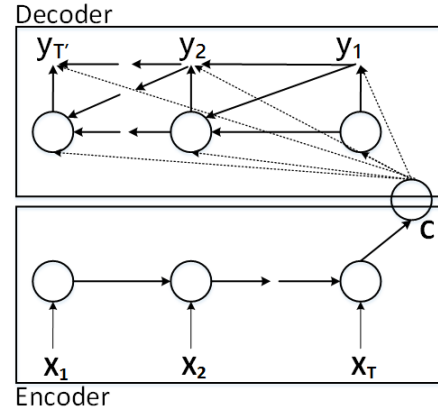


Fig. 9. The illustration of the RNN encoder-decoder (adapted from [277]).

Recently, Jiang and Li [274] propose a deep cross-modal hashing method (DCMH) by integrating feature learning and hash-code learning into the same framework. Unlike most existing methods that typically solve the discrete optimization problem of hash-code learning via relaxing the original discrete learning problem into a continuous learning problem, DCMH directly learns the discrete hash codes without relaxation. In addition, Cao et al. present a fully correlation autoencoder hashing method by extending the correspondence autoencoder by [19].

4.2.4 Multi-Modal Recurrent Neural Network

A recurrent neural network (RNN) [275] is a neural network which processes a variable-length sequence $\mathbf{x} = (x_1, \dots, x_T)$ through hidden state representation \mathbf{h} . At each time step t , the hidden state h_t of the RNN is estimated by

$$h_t = f(h_{t-1}, x_t) \quad (63)$$

where f is a non-linear activation function and is selected based on the requirement of data modeling. For example, a simple case may be a common element-wise logistic sigmoid function and a complex case may be a long short-term memory (LSTM) unit [276].

An RNN is known as its ability of learning a probability distribution over a sequence by being trained to predict the next symbol in a sequence. In this training, the prediction at each time step t is decided by the conditional distribution $p(x_t | x_{t-1}, \dots, x_1)$. For example, a multinomial distribution can be learned as output with a softmax activation function

$$p(x_{t,j} = 1 | x_{t-1}, \dots, x_1) = \frac{\exp(w_j h_t)}{\sum_{j'=1}^K \exp(w_{j'} h_t)} \quad (64)$$

where $j = 1, \dots, K$ denotes the possible symbol components and w_j are the corresponding rows of a weight matrix W . Further, based the above probabilities, the probability of the sequence \mathbf{x} can be computed as

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t | x_{t-1}, \dots, x_1). \quad (65)$$

With this learned distribution, it is straightforward to generate a new sequence by iteratively generating a symbol at each time step.

Cho et al. [277] propose a RNN encoder-decoder model by exploiting RNN to connect multi-modal sequence. As shown in Figure 9, this neural network first encodes a variable-length source sequence into a fixed-length vector representation and then decodes this fixed-length vector representation back into a variable-length target sequence. In fact, it is a general method to learn the conditional distribution over an output sequence conditioned on another input sequence, e.g., $p(y_1, \dots, y_{T'} | x_1, \dots, x_T)$, where the input and output sequence lengths T and T' can be different. In particular, the encoder of the proposed model is an RNN which sequentially encodes each symbol of an input sequence \mathbf{x} into the corresponding hidden state according to Eq.(63). After reading the end of the input sequence, a summary hidden state of the whole source sequence \mathbf{c} is acquired. The decoder of the proposed model is another RNN which is exploited to generate the target sequence by predicting the next symbol y_t with the hidden state h_t . Based on the recurrent property, both y_t and h_t are also conditioned on y_{t-1} and on the summary \mathbf{c} of the input sequence. Thus, the hidden state of the decoder at time t is computed by,

$$h_t = f(h_{t-1}, y_{t-1}, \mathbf{c}) \quad (66)$$

and the conditional distribution of the next symbol is

$$p(y_t | y_{t-1}, y_{t-2}, \dots, y_1, \mathbf{c}) = g(h_t, y_{t-1}, \mathbf{c}) \quad (67)$$

where g is an activation function and produces valid probabilities with a softmax. The main idea of the RNN-based Encoder-Decoder framework can be summarized by jointly training two RNNs to maximize the conditional log-likelihood

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^N \log p_{\theta}(\mathbf{y}_n | \mathbf{x}_n) \quad (68)$$

where θ is the set of the model parameters and each pair $(\mathbf{x}_n, \mathbf{y}_n)$ consists of an input sequence and an output sequence from the training set. The model parameters can be estimated by a gradient-based algorithm.

Further, Sutskever et al. [278] also present a general end-to-end approach for multi-modal sequence to sequence learning based on deep LSTM networks, which is very useful for learning problems with long range temporal dependencies [276, 279]. The goal of this method is also to estimate the conditional probability $p(y_1, \dots, y_{T'} | x_1, \dots, x_T)$. Similar to [277], the conditional probability is computed by first obtaining the fixed dimensional representation v of the input sequence (x_1, \dots, x_T) with the encoding LSTM-based networks, and then computing the probability of $y_1, \dots, y_{T'}$ with the decoding LSTM-based networks whose initial hidden state is set to the representation v of x_1, \dots, x_T :

$$p(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \dots, y_{t-1}) \quad (69)$$

where each $p(y_t | v, y_1, \dots, y_{t-1})$ distribution is represented with a softmax over all the words in the vocabulary.

Besides, multi-modal RNNs have been widely applied in image captioning [21, 22, 280], videos captioning [23, 281, 282], and visual question answering [283]. Karpathy and Li [21] propose a multi-modal recurrent neural network architecture to generate new descriptions of image regions. Chen and Zitnick [284] explore the bi-directional mapping between images and their sentence-based descriptions with RNNs. Venugopalan et al. [282] introduce an end-to-end sequence model to generate captions for videos.

By applying attention mechanism [285] to visual recognition [286, 287], Xu et al. [288] introduce an attention based multi-modal RNN model, which trains the multi-modal RNN in a deterministic manner using the standard back propagation. In particular, it incorporates a form of attention with two variants: a "hard" attention mechanism and a "soft" attention mechanism. The advantage of the proposed model lies in attending to salient part of an image while generating its caption.

5 MULTI-VIEW REPRESENTATION LEARNING AS MANIFOLD ALIGNMENT

Another important perspective on multi-view representation learning is based on the manifold alignment. The key idea underlying this approach is to embed inputs from different domains to a new latent common space, simultaneously preserving topology of each input domain. Its premise is the manifold hypothesis, according to which high-dimensional real-world data are expected to concentrate in the vicinity of a low-dimensional manifold, embedded in high dimensional input space. In particular, this prior is well suited for datasets such as images, sounds and texts, as this information are not natural signals. Since most data sources can be modeled by manifolds, manifold alignment can be used to align the underlying structures across different data views. With this perspective, the multi-view representation learning task can be seen as finding the relationship among the structures of manifolds from the different views of data. The associated multi-view embeddings being learned can be associated with the intrinsic coordinate systems on the embedded manifolds. Further, manifold alignment makes full use of paired and unpaired data and this capability is particularly useful for multi-view feature learning in certain applications, where the paired multi-view instances are limited.

One of the pioneering efforts in manifold alignment with given coordinates is the semi-supervised alignment by Ham et al. [24]. Given certain labeled samples by the ordinal index l , semi-supervised alignment aims to find a map defined on the vertices of the graph $f : V \mapsto R$ that matches known target values for the labeled vertices. This can be solved by directly finding $\arg \min_f |f_i - s_i|$ ($i \in l$) where s is the vector of target value. Since only a small number of labeled examples are provided, it is important to exploit manifold structure in the data when constructing the mapping function. In particular, graph Laplacian matrix L provides this structural information. The semi-supervised alignment problem can be solved by minimizing the following objective function:

$$C(f) = \sum_i \mu |f_i - s_i|^2 + f^T L f \quad (70)$$

where the first term is the loss function, and the second term enforces smoothness along the manifold. The optimum mapping function f is easily obtained by a linear transform.

This semi-supervised alignment is similar to manifold ranking [290], which learns to rank on data manifolds. They both take advantage of the manifold learning to do alignment between different types of data. To exploit the high level geometry structure embedded in the data samples, Mao et al. [289] extend the original semi-supervised alignment and proposed a method called parallel field alignment retrieval (PFAR), which investigates alignment framework from the perspective of parallel vector fields. An example of parallel vector alignment is shown in Figure 10.

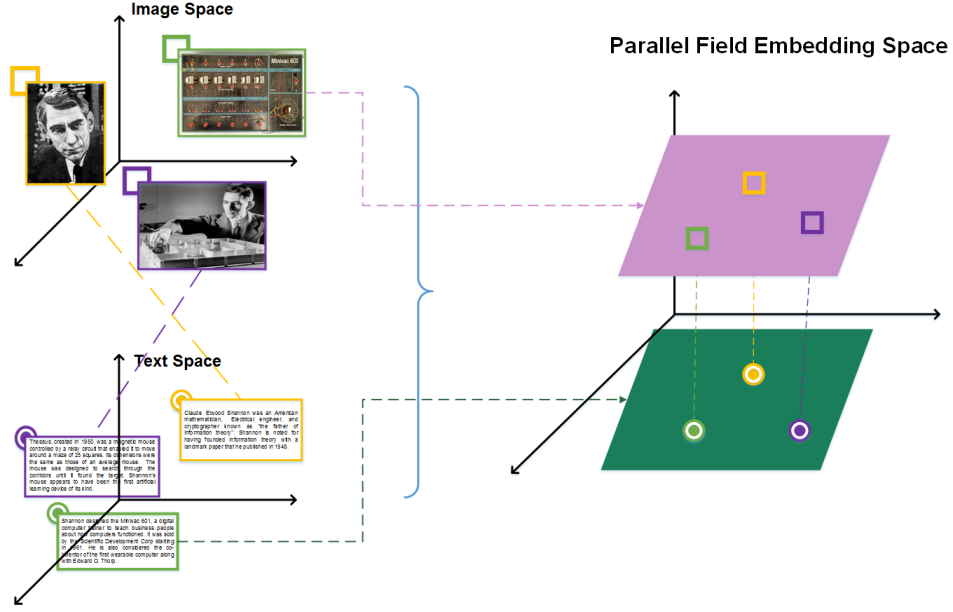


Fig. 10. An illustrative example of parallel field alignment (adapted from [289]).

In addition, Ham et al. [24] also introduce manifold alignment with pairwise correspondence, which builds connections between multiple data sets by aligning their underlying manifolds. In particular, multi-view input data sets $\{X, Y\}$ have subsets X_l and Y_l which are in pairwise alignment by the indices $x_i \leftrightarrow y_i$, ($i \in l$). Then we intend to determine how to match the remaining examples using aligned manifold embeddings. Considering that embedding coordinates for each data set should take similar values for corresponding pairs, the multi-view manifold embedding can be defined by generalizing the single graph embedding algorithm as follows:

$$C(f, g) = \mu \sum_{i \in l} |f_i - g_i| + f^T L^x f + g^T L^y g \quad (71)$$

where f and g denote real-valued functions defined on the respective graphs of X and Y , L^x and L^y are the graph Laplacian matrices of X and Y . The first term penalizes the difference between f and g on the corresponding vertices, and the second and third terms impose smoothness constraints on f and g based on the respective graphs.

This graph algorithm is able to robustly align the underlying manifold structure across multi-view data sets. Even with a small number of paired samples provided, the algorithm is capable of estimating a common low-dimensional embedding space which can be used in cross-view retrieval task. The main concern about this algorithm is the computational cost, which lies in finding the spectral decomposition of a large matrix. Methods for calculating eigenvectors of large sparse matrices can be employed to speed up the computation of the embeddings.

In addition, since the above manifold alignment is defined only on the known data points and it is hard to handle the new test points, Lafon et al. [291] propose a two-step manifold alignment to solve this problem. Diffusion maps are first used to embed the nodes of the graphs corresponding to the aligned multi-modal data sets, followed by an affine matching to align the obtained clouds of points. Further, Wang and Mahadevan [292] also introduce a two-stage approach for manifold alignment using procrustes analysis.

In the first stage, it uses a standard dimensionality reduction method to map the data sets to low dimensional spaces which reflect their intrinsic geometries. In the second stage, procrustes analysis is applied to align the two low dimensional embeddings of the data sets based on a number of landmark points. This two-stage based manifold alignment method obtains mappings that are defined everywhere rather than only training data points. Consequently, Wang and Mahadevan [293] also propose a manifold alignment based approach for heterogeneous domain adaptation.

6 CONCLUSION

Multi-view representation learning has attracted much attention in machine learning and data mining areas. This paper first reviews the root methods and theories on multi-view representation learning, especially on canonical correlation analysis (CCA) and its extensions. And then we investigate the advances of multi-view representation learning that ranges from shallow methods including multi-modal topic learning, multi-view sparse coding, and multi-view latent space Markov networks, to deep methods including multi-modal restricted Boltzmann machines, multi-modal autoencoders, and multi-modal recurrent neural networks. Further, we also provide an important perspective for multi-view representation learning from manifold alignment. This survey aims to provide an insightful picture of the theoretical basis and the current development in the field of multi-view representation learning and to help researchers find the most appropriate methodologies for particular applications.

ACKNOWLEDGMENTS

This work is supported in part by the National Basic Research Program of China (2012CB316400), Zhejiang University — Alibaba Financial Joint lab, and Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis. ZZ is also supported in part by US NSF (IIS-0812114, CCF-1017828).

REFERENCES

- [1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.
- [2] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.
- [3] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.
- [4] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2031–2038, 2013.
- [5] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*, 2008, pp. 1096–1103.
- [6] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines," in *AISTATS*, vol. 1, 2009, p. 3.
- [7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [8] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.
- [9] D. A. Cohn and T. Hofmann, "The missing link - A probabilistic model of document content and hypertext connectivity," in *NIPS*, 2000, pp. 430–436.
- [10] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, Mar. 2003.
- [11] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR*, 2003, pp. 127–134.
- [12] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *NIPS*, 2010, pp. 982–990.
- [13] T. Cao, V. Jovic, S. Modla, D. Powell, K. Czymmek, and M. Niethammer, "Robust multimodal dictionary learning," in *MICCAI*, 2013, pp. 259–266.
- [14] W. Liu, D. Tao, J. Cheng, and Y. Tang, "Multiview hessian discriminative sparse coding for image annotation," *Computer Vision and Image Understanding*, vol. 118, pp. 50–60, 2014.
- [15] E. P. Xing, R. Yan, and A. G. Hauptmann, "Mining associated text and images with dual-wing harmoniums," in *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Edinburgh, Scotland, July 26-29, 2005*, 2005, pp. 633–641.
- [16] N. Chen, J. Zhu, and E. P. Xing, "Predictive subspace learning for multi-view data: a large margin approach," in *NIPS*, 2010, pp. 361–369.
- [17] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *NIPS*, 2012, pp. 2231–2239.
- [18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011, pp. 689–696.
- [19] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *ACM Multimedia*, 2014, pp. 7–16.
- [20] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "On deep multi-view representation learning," in *ICML*, 2015, pp. 1083–1092.
- [21] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," *CoRR*, vol. abs/1412.2306, 2014.
- [22] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.
- [23] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.
- [24] J. Ham, D. Lee, and L. Saul, "Semisupervised alignment of manifolds," in *10th International Workshop on Artificial Intelligence and Statistics*, 2005, pp. 120–127.
- [25] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM Multimedia*, 2010, pp. 251–260.
- [26] E. Kidron, Y. Y. Schechner, and M. Elad, "Pixels that sound," in *CVPR (1)*. IEEE Computer Society, 2005, pp. 88–95.
- [27] L. Sun, S. Ji, and J. Ye, "A least squares formulation for canonical correlation analysis," in *ICML*, 2008, pp. 1024–1031.
- [28] D. P. Foster, R. Johnson, and T. Zhang, "Multi-view dimensionality reduction via canonical correlation analysis," Tech. Rep., 2008.
- [29] L. Sun, B. Ceran, and J. Ye, "A scalable two-stage approach for a class of dimensionality reduction techniques," in *KDD*, 2010, pp. 313–322.
- [30] H. Avron, C. Boutsidis, S. Toledo, and A. Zouzias, "Efficient dimensionality reduction for canonical correlation analysis," in *ICML*, 2013, pp. 347–355.
- [31] X. Z. Fern, C. E. Brodley, and M. A. Friedl, "Correlation clustering for learning mixtures of canonical correlation models," in *SDM*, 2005, pp. 439–448.
- [32] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *CVPR*, 2008.
- [33] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *ICML*, 2009, pp. 129–136.
- [34] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *COLT*, 2007, pp. 82–96.
- [35] B. McWilliams, D. Balduzzi, and J. M. Buhmann, "Correlated random features for fast semi-supervised learning," in *NIPS*, 2013, pp. 440–448.
- [36] P. S. Dhillon, D. P. Foster, and L. H. Ungar, "Multi-view learning of word embeddings via CCA," in *NIPS*, 2011, pp. 199–207.
- [37] P. S. Dhillon, J. Rodu, D. P. Foster, and L. H. Ungar, "Using CCA to improve CCA: A new spectral method for estimating vector models of words," in *ICML*, 2012.
- [38] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.
- [39] T. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical corre-

- lations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, 2007.
- [40] Y. Zhang and J. G. Schneider, “Multi-label output codes using canonical correlation analysis,” in *AISTATS*, 2011, pp. 873–882.
- [41] Y. Su, Y. Fu, X. Gao, and Q. Tian, “Discriminant learning through multiple principal angles for visual recognition,” *IEEE Trans. Image Processing*, vol. 21, no. 3, pp. 1381–1390, 2012.
- [42] G. H. Golub and H. Zha, “The canonical correlations of matrix pairs and their numerical computation,” Stanford, CA, USA, Tech. Rep., 1992.
- [43] J. R. Kettenring, “Canonical analysis of several sets of variables,” *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.
- [44] H. Avron, C. Boutsidis, S. Toledo, and A. Zouzias, “Efficient dimensionality reduction for canonical correlation analysis,” *SIAM J. Scientific Computing*, vol. 36, no. 5, 2014.
- [45] J. A. Tropp, “Improved analysis of the subsampled randomized hadamard transform,” *CoRR*, vol. abs/1011.1595, 2010.
- [46] Y. Lu and D. P. Foster, “large scale canonical correlation analysis with iterative least squares,” in *NIPS*, 2014, pp. 91–99.
- [47] R. Arora, A. Cotter, K. Livescu, and N. Srebro, “Stochastic optimization for PCA and PLS,” in *50th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2012, Allerton Park & Retreat Center, Monticello, IL, USA, October 1-5, 2012*, 2012, pp. 861–868.
- [48] Z. Ma, Y. Lu, and D. P. Foster, “Finding linear structure in large datasets with scalable canonical correlation analysis,” in *ICML*, 2015, pp. 169–178.
- [49] W. Wang, J. Wang, and N. Srebro, “Globally convergent stochastic optimization for canonical correlation analysis,” *CoRR*, vol. abs/1604.01870, 2016.
- [50] R. Ge, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford, “Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis,” *CoRR*, vol. abs/1604.03930, 2016.
- [51] F. R. Bach and M. I. Jordan, “A probabilistic interpretation of canonical correlation analysis,” Department of Statistics, University of California, Berkeley, Tech. Rep., 2005.
- [52] M. W. Browne, “The maximum-likelihood solution in inter-battery factor analysis,” *British Journal of Mathematical and Statistical Psychology*, vol. 32, no. 1, pp. 75–86, 1979.
- [53] C. Archambeau, N. Delannay, and M. Verleysen, “Robust probabilistic projections,” in *ICML*, 2006, pp. 33–40.
- [54] C. Fyfe and G. Leen, “Stochastic processes for canonical correlation analysis,” in *ESANN*, 2006, pp. 245–250.
- [55] A. Klami and S. Kaski, “Local dependent components,” in *ICML*, 2007, pp. 425–432.
- [56] C. Wang, “Variational bayesian approach to canonical correlation analysis,” *IEEE Trans. Neural Networks*, vol. 18, no. 3, pp. 905–910, 2007.
- [57] J. Viinikanoja, A. Klami, and S. Kaski, “Variational bayesian mixture of robust CCA models,” in *ECML PKDD*, 2010, pp. 370–385.
- [58] M. Collins, S. Dasgupta, and R. E. Schapire, “A generalization of principal components analysis to the exponential family,” in *NIPS*, 2001, pp. 617–624.
- [59] S. Mohamed, K. A. Heller, and Z. Ghahramani, “Bayesian exponential family PCA,” in *NIPS*, 2008, pp. 1089–1096.
- [60] A. Klami, S. Virtanen, and S. Kaski, “Bayesian exponential family projections for coupled data sources,” in *UAI*, 2010, pp. 286–293.
- [61] C. Archambeau, S. Guo, and O. Zoeter, “Sparse bayesian multi-task learning,” in *NIPS*, 2011, pp. 1755–1763.
- [62] B. Lakshminarayanan, G. Bouchard, and C. Archambeau, “Robust bayesian matrix factorisation,” in *AISTATS*, 2011, pp. 425–433.
- [63] Y. Mukuta and T. Harada, “Probabilistic partial canonical correlation analysis,” in *ICML*, 2014, pp. 1449–1457.
- [64] B. R. RAO, “Partial canonical correlations,” vol. 20, no. 2, 1969, pp. 211–219.
- [65] C. Kamada, A. Kanezaki, and T. Harada, “Probabilistic semi-canonical correlation analysis,” in *ACM Multimedia*, 2015, pp. 1131–1134.
- [66] I. Huopaniemi, T. Suvitaival, J. Nikkilä, M. Oresic, and S. Kaski, “Two-way analysis of high-dimensional collinear data,” *Data Min. Knowl. Discov.*, vol. 19, no. 2, pp. 261–276, 2009.
- [67] —, “Multivariate multi-way analysis of multi-source data,” *Bioinformatics*, vol. 26, no. 12, pp. 391–398, 2010.
- [68] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society (Series B)*, vol. 58, pp. 267–288, 1996.
- [69] D. R. Hardoon and J. Shawe-Taylor, “Sparse canonical correlation analysis,” Department of Computer Science, University College London, Tech. Rep., 2007.
- [70] S. Waaijenborg, P. C. Verselewe de Witt Hamer, and A. H. Zwinderman, “Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis,” *Statistical applications in genetics and molecular biology*, vol. 7, no. 1, 2008.
- [71] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *Annals of Statistics*, vol. 32, pp. 407–499, 2004.
- [72] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,” *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Dec. 2006.
- [73] B. K. Sriperumbudur, D. A. Torres, and G. R. G. Lanckriet, “Sparse eigen methods by D.C. programming,” in *ICML*, 2007, pp. 831–838.
- [74] A. d’Aspremont, F. R. Bach, and L. E. Ghaoui, “Full regularization path for sparse principal component analysis,” in *ICML*, 2007, pp. 177–184.
- [75] A. d’Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, “A direct formulation for sparse PCA using semidefinite programming,” *SIAM Review*, vol. 49, no. 3, pp. 434–448, 2007.
- [76] D. Torres, D. Turnbull, L. Barrington, B. Sriperumbudur, and G. Lanckriet, “Finding musically meaningful words by sparse cca,” *NIPS WMBC ’07*, December 2007.
- [77] A. Wiesel, M. Kliger, and A. O. Hero, III, “A greedy approach to sparse canonical correlation analysis,” *ArXiv e-prints*, 2008.
- [78] D. M. Witten, T. Hastie, and R. Tibshirani, “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” *Biostatistics*, 2009.
- [79] X. Chen, H. Liu, and J. G. Carbonell, “Structured sparse

- canonical correlation analysis,” in *AISTATS*, 2012, pp. 199–207.
- [80] C. Fyfe and G. Leen, “Two methods for sparsifying probabilistic canonical correlation analysis,” in *ICONIP*, 2006, pp. 361–370.
- [81] C. Archambeau and F. R. Bach, “Sparse probabilistic projections,” in *NIPS*, 2008, pp. 73–80.
- [82] P. Rai and H. D. III, “Multi-label prediction via sparse infinite CCA,” in *NIPS*, 2009, pp. 1518–1526.
- [83] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani, “Estimating image bases for visual image reconstruction from human brain activity,” in *NIPS*, 2009, pp. 576–584.
- [84] K. A. Heller and Z. Ghahramani, “A nonparametric bayesian approach to modeling overlapping clusters,” in *AISTATS*, 2007, pp. 187–194.
- [85] L. R. Tucker, “An inter-battery method of factor analysis,” *Psychometrika*, vol. 23, no. 2, pp. 111–136, 1958.
- [86] A. Klami, S. Virtanen, and S. Kaski, “Bayesian canonical correlation analysis,” *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 965–1003, 2013.
- [87] R. McDonald, “Three common factor models for groups of variables,” *Psychometrika*, vol. 37, no. 1, pp. 173–178, 1970.
- [88] M. W. Browne, “Factor analysis of multiple batteries by maximum likelihood,” *British Journal of Mathematical and Statistical Psychology*, vol. 33, pp. 184–199, 1979.
- [89] A. Klami, S. Virtanen, E. Leppäaho, and S. Kaski, “Group factor analysis,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 26, no. 9, pp. 2136–2147, 2015.
- [90] C. Xu, D. Tao, and C. Xu, “A survey on multi-view learning,” *arXiv preprint arXiv:1304.5634*, 2013.
- [91] P. L. Lai and C. Fyfe, “Kernel and nonlinear canonical correlation analysis,” in *IJCNN (4)*, 2000, p. 614.
- [92] S. Akaho, “A kernel method for canonical correlation analysis,” in *International Meeting of the Psychometric Society*, 2001.
- [93] T. Melzer, M. Reiter, and H. Bischof, “Nonlinear feature extraction using generalized canonical correlation analysis,” in *ICANN*, 2001, pp. 353–360.
- [94] R. Socher and F. Li, “Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora,” in *CVPR*, 2010, pp. 966–973.
- [95] S. J. Hwang and K. Grauman, “Learning the relative importance of objects from tagged images for retrieval and cross-modal search,” *International Journal of Computer Vision*, vol. 100, no. 2, pp. 134–153, 2012.
- [96] Y. Yamanishi, J. Vert, A. Nakaya, and M. Kanehisa, “Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis,” in *ICISMB*, 2003, pp. 323–330.
- [97] M. B. Blaschko, J. A. Shelton, A. Bartels, C. H. Lampert, and A. Gretton, “Semi-supervised kernel canonical correlation analysis with application to human fmri,” *Pattern Recognition Letters*, vol. 32, no. 11, pp. 1572–1583, 2011.
- [98] A. Trivedi, P. Rai, S. L. DuVall, and H. Daumé, III, “Exploiting tag and word correlations for improved webpage clustering,” in *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, ser. SMUC ’10, 2010, pp. 3–12.
- [99] R. Arora and K. Livescu, “Kernel CCA for multi-view learning of acoustic features using articulatory measurements,” in *MLSLP*, 2012, pp. 34–37.
- [100] —, “Multi-view cca-based acoustic features for phonetic recognition across speakers and domains,” in *ICASSP*, 2013, pp. 7135–7139.
- [101] A. Gretton, R. Herbrich, and A. J. Smola, “The kernel mutual information,” in *ICASSP*, 2003, pp. 880–884.
- [102] A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf, “Kernel methods for measuring independence,” *Journal of Machine Learning Research*, vol. 6, pp. 2075–2129, 2005.
- [103] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2001.
- [104] M. Kuss and T. Graepel, “The geometry of kernel canonical correlation analysis,” Max Planck Institute for Biological Cybernetics, Tübingen, Germany, Tech. Rep. 108, may 2003.
- [105] K. Fukumizu, F. R. Bach, and A. Gretton, “Statistical consistency of kernel canonical correlation analysis,” *Journal of Machine Learning Research*, vol. 8, pp. 361–383, 2007.
- [106] D. R. Hardoon and J. Shawe-Taylor, “Convergence analysis of kernel canonical correlation analysis: theory and practice,” *Machine Learning*, vol. 74, no. 1, pp. 23–38, 2009.
- [107] J. Cai and H. Sun, “Convergence rate of kernel canonical correlation analysis,” *Science China Mathematics*, vol. 54, no. 10, pp. 2161–2170, 2011.
- [108] M. Brand, “Incremental singular value decomposition of uncertain data with missing values,” in *ECCV*, 2002, pp. 707–720.
- [109] C. K. I. Williams and M. W. Seeger, “Using the nystrom method to speed up kernel machines,” in *NIPS*, 2000, pp. 682–688.
- [110] T. Yang, Y. Li, M. Mahdavi, R. Jin, and Z. Zhou, “Nyström method vs random fourier features: A theoretical and empirical comparison,” in *NIPS*, 2012, pp. 485–493.
- [111] Q. V. Le, T. Sarlós, and A. J. Smola, “Fastfood - computing hilbert space expansions in loglinear time,” in *ICML*, 2013, pp. 244–252.
- [112] D. Lopez-Paz, S. Sra, A. J. Smola, Z. Ghahramani, and B. Schölkopf, “Randomized nonlinear component analysis,” in *ICML*, 2014, pp. 1359–1367.
- [113] W. Wang and K. Livescu, “Large-scale approximate kernel canonical correlation analysis,” *CoRR*, vol. abs/1511.04773, 2015.
- [114] Y. Li and J. Shawe-Taylor, “Using KCCA for japanese-english cross-language information retrieval and document classification,” *J. Intell. Inf. Syst.*, vol. 27, no. 2, pp. 117–133, 2006.
- [115] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, “Two view learning: Svm-2k, theory and practice,” in *NIPS*, 2005, pp. 355–362.
- [116] D. R. Hardoon, C. Saunders, S. Szedmak, and J. Shawe-Taylor, “A correlation approach for automatic image annotation,” in *ADMA*, 2006, pp. 681–692.
- [117] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *J. Artif. Intell. Res. (JAIR)*, vol. 47, pp. 853–899, 2013.
- [118] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,” *Journal of Machine Learning*

- Research*, vol. 7, pp. 2399–2434, 2006.
- [119] S. Becker and G. E. Hinton, “Self-organizing neural network that discovers surfaces in random-dot stereograms,” *Nature*, vol. 355, no. 6356, pp. 161–163, 1992.
 - [120] S. Becker, “Mutual information maximization: Models of cortical self-organization,” *Network : Computation in Neural Systems*, vol. 7, pp. 7–31, 1996.
 - [121] P. L. Lai and C. Fyfe, “Canonical correlation analysis using artificial neural networks,” in *ESANN*, 1998, pp. 363–368.
 - [122] —, “A neural implementation of canonical correlation analysis,” *Neural Networks*, vol. 12, no. 10, pp. 1391–1397, 1999.
 - [123] W. W. Hsieh, “Nonlinear canonical correlation analysis by neural networks,” *Neural Networks*, vol. 13, no. 10, pp. 1095–1105, 2000.
 - [124] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012, pp. 1097–1105.
 - [125] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, “Unsupervised learning of acoustic features via deep canonical correlation analysis,” in *ICASSP*, 2015, pp. 4590–4594.
 - [126] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu, “Deep multilingual correlation for improved word embeddings,” in *HLT-NAACL*, 2015, pp. 250–256.
 - [127] W. Wang, R. Arora, K. Livescu, and N. Srebro, “Stochastic optimization for deep CCA via nonlinear orthogonal iterations,” in *Allerton*, 2015, pp. 688–695.
 - [128] F. Yan and K. Mikolajczyk, “Deep correlation for matching images and text,” in *CVPR*, 2015, pp. 3441–3450.
 - [129] T. Hofmann, “Probabilistic latent semantic analysis,” in *UAI*, 1999, pp. 289–296.
 - [130] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.
 - [131] D. Cohn and H. Chang, “Learning to probabilistically identify authoritative documents,” in *ICML*, 2000, pp. 167–174.
 - [132] F. Monay and D. Gatica-Perez, “Plsa-based image auto-annotation: Constraining the latent space,” in *ACM Multimedia*, 2004, pp. 348–351.
 - [133] R. Zhang, Z. M. Zhang, M. Li, W. Ma, and H. Zhang, “A probabilistic semantic model for image annotation and multi-modal image retrieval,” in *ICCV*, 2005, pp. 846–851.
 - [134] F. Monay and D. Gatica-Perez, “Modeling semantic aspects for cross-media image indexing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1802–1817, 2007.
 - [135] Y. Shao, Y. Zhou, X. He, D. Cai, and H. Bao, “Semi-supervised topic modeling for image annotation,” in *ACM Multimedia*, 2009, pp. 521–524.
 - [136] R. Lienhart, S. Romberg, and E. Hörster, “Multilayer plsa for multimodal image retrieval,” in *ACM CIVR*, 2009.
 - [137] Q. Yang, Y. Chen, G. Xue, W. Dai, and Y. Yu, “Heterogeneous transfer learning for image clustering via the socialweb,” in *ACL*, 2009, pp. 1–9.
 - [138] C. Nguyen, N. Kaothanthong, X. H. Phan, and T. Tokuyama, “A feature-word-topic model for image annotation,” in *ACM CIKM*, 2010, pp. 1481–1484.
 - [139] Z. Li, Z. Shi, X. Liu, and Z. Shi, “Automatic image annotation with continuous PLSA,” in *ICASSP*, 2010, pp. 806–809.
 - [140] C. Pulla and C. V. Jawahar, “Multi modal semantic indexing for image retrieval,” in *ACM CIVR*, 2010, pp. 342–349.
 - [141] R. Zhang, L. Zhang, X. Wang, and L. Guan, “Multi-feature plsa for combining visual features in image annotation,” in *ACM Multimedia*, 2011, pp. 1513–1516.
 - [142] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
 - [143] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, 2004.
 - [144] O. Yakhnenko and V. Honavar, “Multi-modal hierarchical dirichlet process model for predicting image annotation and image-object label correspondence,” in *SDM*, 2009, pp. 283–293.
 - [145] H. Ishwaran and L. F. James, “Gibbs sampling methods for stick-breaking priors,” *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 161–173, 2001.
 - [146] Z. Qi, M. Yang, Z. M. Zhang, and Z. Zhang, “Mining partially annotated images,” in *KDD*, 2011, pp. 1199–1207.
 - [147] J. Zhu, L. Li, F. Li, and E. P. Xing, “Large margin learning of upstream scene understanding models,” in *NIPS*, 2010, pp. 2586–2594.
 - [148] C. Wang, D. M. Blei, and F. Li, “Simultaneous image classification and annotation,” in *CVPR*, 2009, pp. 1903–1910.
 - [149] D. M. Blei and J. D. McAuliffe, “Supervised topic models,” in *NIPS*, 2007, pp. 121–128.
 - [150] L. Li and F. Li, “What, where and who? classifying events by scene and object recognition,” in *ICCV*, 2007, pp. 1–8.
 - [151] L. Cao and F. Li, “Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes,” in *ICCV*, 2007, pp. 1–8.
 - [152] L. Li, R. Socher, and F. Li, “Towards total scene understanding: Classification, annotation and segmentation in an automatic framework,” in *CVPR*, 2009, pp. 2036–2043.
 - [153] C. Nguyen, D. Zhan, and Z. Zhou, “Multi-modal image annotation with multi-instance multi-label LDA,” in *IJCAI*, 2013.
 - [154] C. E. Rasmussen, “Gaussian processes for machine learning.” MIT Press, 2006.
 - [155] C. H. Ek, P. H. S. Torr, and N. D. Lawrence, “Gaussian process latent variable models for human pose estimation,” in *MLMI*, 2007, pp. 132–143.
 - [156] N. D. Lawrence, “Gaussian process latent variable models for visualisation of high dimensional data,” in *NIPS*, 2003, pp. 329–336.
 - [157] A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao, “Learning shared latent structure for image synthesis and robotic imitation,” in *NIPS*, 2005, pp. 1233–1240.
 - [158] R. Urtasun and T. Darrell, “Discriminative gaussian process latent variable model for classification,” in *ICML*, 2007, pp. 927–934.
 - [159] S. Eleftheriadis, O. Rudovic, and M. Pantic, “Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition,” *IEEE Trans. Image Processing*, vol. 24, no. 1, pp. 189–204, 2015.
 - [160] C. Xu, D. Tao, Y. Li, and C. Xu, “Large-margin multi-view gaussian process for image classification,” in *ICIMCS*, 2013, pp. 7–12.
 - [161] G. Song, S. Wang, Q. Huang, and Q. Tian, “Similarity

- gaussian process latent variable model for multi-modal data analysis,” in *ICCV*, 2015, pp. 4050–4058.
- [162] A. P. Singh and G. J. Gordon, “Relational learning via collective matrix factorization,” in *SIGKDD*, 2008, pp. 650–658.
- [163] K. Yu, S. Yu, and V. Tresp, “Multi-label informed latent semantic indexing,” in *SIGIR*, 2005, pp. 258–265.
- [164] B. Long, Z. M. Zhang, X. Wu, and P. S. Yu, “Spectral clustering for multi-type relational data,” in *ICML*, 2006, pp. 585–592.
- [165] B. Long, Z. M. Zhang, and P. S. Yu, “A probabilistic framework for relational clustering,” in *KDD*, 2007, pp. 470–479.
- [166] C. Lippert, S. H. Weber, Y. Huang, V. Tresp, M. Schubert, and H.-P. Kriegel, “Relation-Prediction in Multi-Relational Domains using Matrix-Factorization,” in *NIPS 2008 Workshop: Structured Input - Structured Output*, 2008.
- [167] A. Klami, G. Bouchard, and A. Tripathi, “Group-sparse embeddings in collective matrix factorization,” *CoRR*, vol. abs/1312.5921, 2013.
- [168] G. Bouchard, D. Yin, and S. Guo, “Convex collective matrix factorization,” in *AISTATS*, 2013, pp. 144–152.
- [169] S. Gunasekar, M. Yamada, D. Yin, and Y. Chang, “Consistent collective matrix completion under joint low rank structure,” in *AISTATS*, 2015.
- [170] C. Lan, X. Li, Y. Deng, and J. Huan, “Partial collective matrix factorization and its pac bound,” 2016.
- [171] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu, “Image annotation using multi-correlation probabilistic matrix factorization,” in *ACM Multimedia*, 2010, pp. 1187–1190.
- [172] S. Zhu, K. Yu, Y. Chi, and Y. Gong, “Combining content and link for classification using matrix factorization,” in *SIGIR*, 2007, pp. 487–494.
- [173] A. P. Singh and G. J. Gordon, “A bayesian matrix factorization model for relational data,” *CoRR*, vol. abs/1203.3517, 2012.
- [174] Y. Zhen and D. Yeung, “A probabilistic model for multi-modal hash function learning,” in *SIGKDD*, 2012, pp. 940–948.
- [175] G. Ding, Y. Guo, and J. Zhou, “Collective matrix factorization hashing for multimodal data,” in *CVPR*, 2014, pp. 2083–2090.
- [176] Z. Akata, C. Thureau, and C. Bauckhage, “Non-negative Matrix Factorization in Multimodality Data for Segmentation and Label Prediction,” in *16th Computer Vision Winter Workshop*, 2011.
- [177] J. Liu, C. Wang, J. Gao, and J. Han, “Multi-view clustering via joint nonnegative matrix factorization,” in *SDM*, 2013, pp. 252–260.
- [178] M. Zitnik and B. Zupan, “Data fusion by matrix factorization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 41–53, 2015.
- [179] L. Yang, L. Jing, and M. K. Ng, “Robust and non-negative collective matrix factorization for text-to-image transfer learning,” *IEEE Trans. Image Processing*, vol. 24, no. 12, pp. 4701–4714, 2015.
- [180] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
- [181] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang, “Sparse unsupervised dimensionality reduction for multiple view data,” *IEEE Trans. Circuits Syst. Video Techn.*, vol. 22, no. 10, pp. 1485–1496, 2012.
- [182] J. Yu, Y. Rui, and D. Tao, “Click prediction for web image reranking using multimodal sparse coding,” *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2019–2032, 2014.
- [183] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, “Supervised coupled dictionary learning with group structures for multi-modal retrieval,” in *AAAI*, 2013.
- [184] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, “Sparse multi-modal hashing,” *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, 2014.
- [185] M. Welling, M. Rosen-Zvi, and G. E. Hinton, “Exponential family harmoniums with an application to information retrieval,” in *NIPS*, 2004, pp. 1481–1488.
- [186] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
- [187] P. Xie and E. P. Xing, “Multi-modal distance metric learning,” in *IJCAI*, 2013, pp. 1806–1812.
- [188] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, “Distance metric learning with application to clustering with side-information,” in *NIPS*, 2002, pp. 505–512.
- [189] H. Wold, “Soft modeling: the basic design and some extensions,” *Systems under indirect observation*, vol. 2, pp. 589–591, 1982.
- [190] R. Rosipal and N. Krämer, *Overview and Recent Advances in Partial Least Squares*, 2006, pp. 34–51.
- [191] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, “Human detection using partial least squares analysis,” in *ICCV*, 2009, pp. 24–31.
- [192] S. Wold, M. Sjström, and L. Eriksson, “PLS-regression: a basic tool of chemometrics,” *Chemometrics and Intelligent Laboratory Systems*, vol. 58, pp. 109–130, 2001.
- [193] M. Barker and W. Rayens, “Partial least squares for discrimination,” *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.
- [194] G. Guo and G. Mu, “Joint estimation of age, gender and ethnicity: CCA vs. PLS,” in *FG*, 2013, pp. 1–6.
- [195] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, “Multimedia content processing through cross-modal association,” in *ACM Multimedia*, 2003, pp. 604–611.
- [196] Y. Wang, L. Guan, and A. N. Venetsanopoulos, “Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition,” *IEEE Transactions on Multimedia*, vol. 14, no. 3-1, pp. 597–607, 2012.
- [197] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, “On the role of correlation and abstraction in cross-modal multimedia retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, 2014.
- [198] J. Wang, H. Wang, Y. Tu, K. Duan, Z. Zhan, and S. Chekuri, “Supervised cross-modal factor analysis,” *CoRR*, vol. abs/1502.05134, 2015.
- [199] K. Worsley, J.-B. Poline, K. Friston, and A. Evans, “Characterizing the response of PET and fMRI data using multivariate linear models,” *Neuroimage*, vol. 6, no. 4, pp. 305–319, 1997.
- [200] J. Arenas-García and G. Camps-Valls, “Efficient kernel orthonormalized PLS for remote sensing applications,” *IEEE Trans. Geoscience and Remote Sensing*, vol. 46, no. 10, pp. 2872–2881, 2008.

- [201] S. Muñoz Romero, J. Arenas-García, and V. Gómez-Verdejo, “Sparse and kernel ops feature extraction based on eigenvalue problem solving,” *Pattern Recognition*, vol. 48, no. 5, pp. 1797–1811, May 2015.
- [202] L. Sun, S. Ji, S. Yu, and J. Ye, “On the equivalence between canonical correlation analysis and orthonormalized partial least squares,” in *IJCAI*, 2009, pp. 1230–1235.
- [203] S. Li, J. Gao, J. O. Nyagilo, and D. P. Dave, “Eigenspectra, a robust regression method for multiplexed raman spectra analysis,” in *BIBM*, 2010, pp. 525–530.
- [204] —, “Probabilistic partial least square regression: A robust model for quantitative analysis of raman spectroscopy data,” in *BIBM*, 2011, pp. 526–531.
- [205] T. Diethe, D. R. Hardoon, and J. Shawe-taylor, “Multiview fisher discriminant analysis,” in *In NIPS Workshop on Learning from Multiple Sources*, 2008.
- [206] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor, “Constructing nonlinear discriminants from multiple data views,” in *ECML PKDD*, 2010, pp. 328–343.
- [207] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, “Multiview discriminant analysis,” in *ECCV*, 2012, pp. 808–821.
- [208] S. Mika, A. J. Smola, and B. Schölkopf, “An improved training algorithm for kernel fisher discriminants,” in *AISTATS*, 2001.
- [209] J. B. Tenenbaum and W. T. Freeman, “Separating style and content with bilinear models,” *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.
- [210] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, “Graph embedding and extensions: A general framework for dimensionality reduction,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, 2007.
- [211] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs, “Generalized multiview analysis: A discriminative latent space,” in *CVPR*, 2012, pp. 2160–2167.
- [212] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Q. Weinberger, “Learning to rank with (a lot of) word features,” *Inf. Retr.*, vol. 13, no. 3, pp. 291–314, 2010.
- [213] J. Weston, S. Bengio, and N. Usunier, “Large scale image annotation: learning to rank with joint word-image embeddings,” *Machine Learning*, vol. 81, no. 1, pp. 21–35, 2010.
- [214] D. Grangier and S. Bengio, “A discriminative kernel-based approach to rank images from text queries,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, 2008.
- [215] R. Herbrich, T. Graepel, and K. Obermayer, “Large margin rank boundaries for ordinal regression,” in *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 2000.
- [216] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender, “Learning to rank using gradient descent,” in *ICML*, 2005, pp. 89–96.
- [217] M. Collins and N. Duffy, “New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron,” in *ACL*, 2002, pp. 263–270.
- [218] R. Collobert and S. Bengio, “Links between perceptrons, mlps and svms,” in *ICML*, 2004.
- [219] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang, “A low rank structural large margin method for cross-modal ranking,” in *SIGIR*, 2013, pp. 433–442.
- [220] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, “Cross-media semantic representation via bi-directional learning to rank,” in *ACM Multimedia*, 2013, pp. 877–886.
- [221] F. Wu, X. Jiang, X. Li, S. Tang, W. Lu, Z. Zhang, and Y. Zhuang, “Cross-modal learning to rank via latent joint representation,” *IEEE Trans. Image Processing*, vol. 24, no. 5, pp. 1497–1509, 2015.
- [222] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, “Data fusion through cross-modality metric learning using similarity-sensitive hashing,” in *CVPR*, 2010, pp. 3594–3601.
- [223] S. Kumar and R. Udupa, “Learning hash functions for cross-view similarity search,” in *IJCAI*, 2011, pp. 1360–1365.
- [224] Y. Gong and S. Lazebnik, “Iterative quantization: A procrustean approach to learning binary codes,” in *CVPR*, 2011, pp. 817–824.
- [225] Y. Zhen and D. Yeung, “Co-regularized hashing for multimodal data,” in *NIPS*, 2012, pp. 1385–1393.
- [226] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, “Linear cross-modal hashing for efficient multimedia search,” in *ACM Multimedia Conference, MM ’13, Barcelona, Spain, October 21-25, 2013*, 2013, pp. 143–152.
- [227] D. Zhai, H. Chang, Y. Zhen, X. Liu, X. Chen, and W. Gao, “Parametric local multimodal hashing for cross-view similarity search,” in *IJCAI*, 2013, pp. 2754–2760.
- [228] X. Liu, J. He, C. Deng, and B. Lang, “Collaborative hashing,” in *CVPR*, 2014, pp. 2147–2154.
- [229] D. Zhang and W. Li, “Large-scale supervised multimodal hashing with semantic correlation maximization,” in *AAAI*, 2014, pp. 2177–2183.
- [230] Z. Lin, G. Ding, M. Hu, and J. Wang, “Semantics-preserving hashing for cross-view retrieval,” in *CVPR*, 2015, pp. 3864–3872.
- [231] Y. Zhen, Y. Gao, D. Yeung, H. Zha, and X. Li, “Spectral multimodal hashing and its application to multimedia retrieval,” *IEEE Trans. Cybernetics*, vol. 46, no. 1, pp. 27–38, 2016.
- [232] G. Shakhnarovich, “Learning task-specific similarity,” Ph.D. dissertation, Cambridge, MA, USA, 2005.
- [233] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
- [234] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, “Discriminative coupled dictionary hashing for fast cross-media retrieval,” in *SIGIR*, 2014, pp. 395–404.
- [235] Y. Weiss, A. Torralba, and R. Fergus, “Spectral hashing,” in *NIPS*, 2008, pp. 1753–1760.
- [236] W. Liu, J. Wang, S. Kumar, and S. Chang, “Hashing with graphs,” in *ICML*, 2011, pp. 1–8.
- [237] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, “Inter-media hashing for large-scale retrieval from heterogeneous data sources,” in *SIGMOD*, 2013, pp. 785–796.
- [238] J. Zhou, G. Ding, and Y. Guo, “Latent semantic sparse hashing for cross-modal similarity search,” in *SIGIR*, 2014, pp. 415–424.
- [239] B. Wu, Q. Yang, W. Zheng, Y. Wang, and J. Wang, “Quantized correlation hashing for fast cross-modal search,” in *IJCAI*, 2015, pp. 3946–3952.
- [240] D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Found-*

- dations. Cambridge, MA, USA: MIT Press, 1986.
- [241] G. E. Hinton and R. R. Salakhutdinov, "Replicated softmax: an undirected topic model," in *NIPS*, 2009, pp. 1607–1614.
 - [242] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *ICASSP*, 2013, pp. 7596–7599.
 - [243] H. Hu, B. Liu, B. Wang, M. Liu, and X. Wang, "Multimodal dbn for predicting high-quality answers in cqa portals," in *ACL (2)*, 2013, pp. 843–847.
 - [244] L. Ge, J. Gao, X. Li, and A. Zhang, "Multi-source deep learning for information trustworthiness estimation," in *KDD*, 2013, pp. 766–774.
 - [245] L. Pang and C.-W. Ngo, "Multimodal learning with deep boltzmann machine for emotion prediction in user generated videos," in *ACM ICMR*, 2015, pp. 619–622.
 - [246] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *NIPS*, 2014, pp. 2141–2149.
 - [247] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer, "Facial attributes classification using multi-task representation learning," in *CVPR Workshops*, 2016, pp. 47–55.
 - [248] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *ICML*, 2008, pp. 536–543.
 - [249] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, "Multimodal deep autoencoder for human pose recovery," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5659–5670, 2015.
 - [250] B. S. Riggan, C. Reale, and N. M. Nasrabadi, "Coupled auto-associative neural networks for heterogeneous face recognition," *IEEE Access*, vol. 3, pp. 1620–1632, 2015.
 - [251] H. Amiri, P. Resnik, J. Boyd-Graber, and H. D. III, "Learning text pair similarity with context-sensitive autoencoders," 2016.
 - [252] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2," in *NIPS*, 2008, pp. 873–880.
 - [253] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked autoencoders," *Proceedings of the VLDB Endowment*, vol. 7, no. 8, pp. 649–660, 2014.
 - [254] S. Rastegar, M. Soleymani, H. R. Rabiee, and S. Mohsen Shojaei, "Mdl-cw: A multimodal deep learning framework with cross weights," in *CVPR*, 2016, pp. 2601–2609.
 - [255] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013, pp. 2121–2129.
 - [256] J. Dong, X. Li, and C. G. Snoek, "Word2visualvec: Cross-media retrieval by visual feature prediction," *arXiv preprint arXiv:1604.06838*, 2016.
 - [257] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Multimodal neural language models," in *ICML*, vol. 14, 2014, pp. 595–603.
 - [258] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," *arXiv preprint arXiv:1501.02598*, 2015.
 - [259] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
 - [260] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le *et al.*, "Large scale distributed deep networks," in *NIPS*, 2012, pp. 1223–1231.
 - [261] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," *arXiv preprint arXiv:1312.5650*, 2013.
 - [262] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *CVPR*, 2015, pp. 1473–1482.
 - [263] Y. Feng and M. Lapata, "Visual information in semantic representation," in *HLT-NAACL*, 2010, pp. 91–99.
 - [264] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics," *J. Artif. Intell. Res. (JAIR)*, vol. 49, no. 1–47, 2014.
 - [265] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *EMNLP*, 2014, pp. 36–45.
 - [266] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *AAAI*, 2015, pp. 2346–2352.
 - [267] W. Yu, K. Yang, Y. Bai, H. Yao, and Y. Rui, "Learning cross space mapping via dnn using large scale click-through logs," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2000–2007, 2015.
 - [268] X. Jiang, F. Wu, X. Li, Z. Zhao, W. Lu, S. Tang, and Y. Zhuang, "Deep compositional cross-modal learning to rank via local-global alignment," in *ACM Multimedia*, 2015, pp. 69–78.
 - [269] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with cnn visual features: A new baseline," 2016.
 - [270] F. Wu, X. Lu, J. Song, S. Yan, Z. M. Zhang, Y. Rui, and Y. Zhuang, "Learning of multimodal representations with random walks on the click graph," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 630–642, 2016.
 - [271] Y. Kang, S. Kim, and S. Choi, "Deep learning to hash with multiple representations," in *ICDM*, 2012, pp. 930–935.
 - [272] Y. Zhuang, Z. Yu, W. Wang, F. Wu, S. Tang, and J. Shao, "Cross-media hashing with neural networks," in *ACM Multimedia*, 2014, pp. 901–904.
 - [273] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *IJCAI*, 2015.
 - [274] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," *arXiv preprint arXiv:1602.02255*, 2016.
 - [275] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Interspeech*, vol. 2, 2010, p. 3.
 - [276] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
 - [277] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP*, 2014, pp. 1724–1734.
 - [278] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014, pp. 3104–3112.
 - [279] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE*

Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994.

- [280] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” *arXiv preprint arXiv:1411.2539*, 2014.
- [281] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, “Translating videos to natural language using deep recurrent neural networks,” *arXiv preprint arXiv:1412.4729*, 2014.
- [282] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence-video to text,” in *ICCV*, 2015, pp. 4534–4542.
- [283] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *ICCV*, 2015, pp. 2425–2433.
- [284] X. Chen and C. L. Zitnick, “Learning a recurrent visual representation for image caption generation,” *arXiv preprint arXiv:1411.5654*, 2014.
- [285] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *CoRR*, vol. abs/1409.0473, 2014.
- [286] J. Ba, V. Mnih, and K. Kavukcuoglu, “Multiple object recognition with visual attention,” *CoRR*, vol. abs/1412.7755, 2014.
- [287] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, “Recurrent models of visual attention,” *CoRR*, vol. abs/1406.6247, 2014.
- [288] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” *CoRR*, vol. abs/1502.03044, 2015.
- [289] X. Mao, B. Lin, D. Cai, X. He, and J. Pei, “Parallel field alignment for cross media retrieval,” in *ACM Multimedia*, 2013, pp. 897–906.
- [290] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, “Ranking on data manifolds,” in *NIPS*, 2003, pp. 169–176.
- [291] S. Lafon, Y. Keller, and R. R. Coifman, “Data fusion and multicue data matching by diffusion maps,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1784–1797, 2006.
- [292] C. Wang and S. Mahadevan, “Manifold alignment using procrustes analysis,” in *ICML*, 2008, pp. 1120–1127.
- [293] —, “Heterogeneous domain adaptation using manifold alignment,” in *IJCAI*, 2011, pp. 1541–1546.

PLACE
PHOTO
HERE

Ming Yang received the BS, MS, and PhD degrees in information science and electronic engineering from Zhejiang University, Hangzhou, China. He had been a visiting scholar in computer science at the State University of New York (SUNY) at Binghamton.

PLACE
PHOTO
HERE

Zhongfei (Mark) Zhang received the BS degree in electronics engineering, the MS degree in information science, both from Zhejiang University, Hangzhou, China, and the PhD degree in computer science from the University of Massachusetts at Amherst. He is a QiuShi Chaired Professor at Zhejiang University, China, and directs the Data Science and Engineering Research Center at the university while he is on leave from State University of New York (SUNY) at Binghamton, USA, where he is a professor at the Computer Science Department and directs the Multimedia Research Laboratory in the Department. He has published more than 100 peer-reviewed academic papers in leading international journals and conferences and several invited papers and book chapters, has authored or co-authored two monographs on multimedia data mining and relational data clustering, respectively.

PLACE
PHOTO
HERE

Yingming Li received the BS and MS degrees in automation from University of Science and Technology of China, Hefei, China, and the PhD degree in information science and electronic engineering from Zhejiang University, Hangzhou, China. He is currently an assistant professor with the College of Information Science and Electronic Engineering at Zhejiang University, China.