Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Cross-Modal Similarity Learning via Pairs, Preferences, and Active Supervision

Yi Zhen
Georgia Institute of Technology
Atlanta, GA 30332
yzhen@cc.gatech.edu

Piyush Rai
Duke University
Durham, NC 27708
piyush.rai@duke.edu

Hongyuan Zha
Georgia Institute of Technology
Atlanta, GA 30332
zha@cc.gatech.edu

Lawrence Carin
Duke University
Durham, NC 27708
lcarin@duke.edu

Abstract

We present a probabilistic framework for learning pairwise similarities between objects belonging to different modalities, such as drugs and proteins, or text and images. Our framework is based on learning a binary code based representation for objects in each modality, and has the following key properties: (i) it can leverage both pairwise as well as easy-to-obtain *relative preference* based cross-modal constraints, (ii) the probabilistic framework naturally allows querying for the most useful/informative constraints, facilitating an active learning setting (existing methods for cross-modal similarity learning do not have such a mechanism), and (iii) the binary code length is learned from the data. We demonstrate the effectiveness of the proposed approach on two problems that require computing pairwise similarities between cross-modal object pairs: cross-modal link prediction in bipartite graphs, and hashing based cross-modal similarity search.

Introduction

Many real-world datasets are inherently multimodal and heterogeneous in nature. For example, a web corpora may consist of data from multiple modalities, such as text, images, video and audio; multi-lingual text corpora may consist of documents from different languages; and medical imaging data may consist of images from multiple imaging modalities such as fMRI, CT, and PET. Often, we are interested in computing the similarities between objects that belong to different modalities. This has applications in a variety of problems such as similarity search, object-alignment, and link prediction (Bronstein et al. 2010; Whang, Rai, and Dhillon 2013), involving cross-modal data. This task requires feature representations that are (i) compact and interpretable, and (ii) *comparable* across the different modalities.

Algorithms for learning cross-modal similarities usually rely on some form of supervision, such as pairwise constraints between objects *across* different modalities. These constraints may be generated using a small amount of *supervised* side-information about some of the cross-modal pairs of objects (e.g., using their class labels, or some user-definied *subjective* notion of similarity). These constraints are subsequently used by the algorithm to learn a mapping of data to a new feature space, such that the mapping respects the cross-domain pairwise constraints (Bronstein et al. 2010; Zhen and Yeung 2012a; 2012b). In many cases, however, explicit pairwise constraints may be difficult and/or expensive to obtain. It is therefore desirable to (i) exploit *weaker* but *easier-to-obtain* constraints, and/or (ii) select the constraints based on some active sampling scheme that returns the most useful/informative constraints.

Driven by this motivation, in this paper we present a cross-modal similarity learning framework which fulfils both the aforementioned desiderarta. Our framework is based on learning compact *binary* codes for objects in each modality, which can be used to compute cross-modal similarities. The key aspects of our framework are: (i) both cross-modal pairwise and cross-modal *triplet* constraints (*relative preferences* defined w.r.t. two cross-modal object pairs) can be incorporated in a coherent way. In particular, to the best of our knowledge, none of the existing methods for learning *cross-modal* binary codes incorporate implicit constraint such as triplets, and (ii) An efficient active learning strategy for querying the most useful pairwise and triplet constraints. Moreover, the binary code size for each modality is learned adaptively from the data, taking a nonparametric Bayesian approach (Griffiths and Ghahramani 2011). This is particularly useful for dynamically extending existing hash codes when new objects (from any modality) or new constraints become available (Quadrianto et al. 2013).

Learning Cross-Modal Similarities via Pairs and Preferences

For simplicity of exposition, we focus on the bi-modal case, for examples texts and images (our framework can be extended for more than two modalities in a straightforward

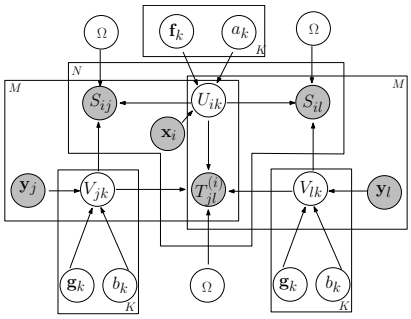


Figure 1: Graphical model representation of CSLP

manner). Subsequently, we will refer to our framework as CSLP (abbreviated for **C**ross-Modal **S**imilarity **L**earning via **P**airs and Preferences). The graphical model representation of CSLP is shown in Figure 1, where the gray circles denote the observed data and the white circles denote the latent variables, which we want to infer using the observed data. As shown in Figure 1, we are given two sets of data points from two modalities. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{D_x}$ be a set of N points from modality $\mathcal{X}$, and $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\} \in \mathbb{R}^{D_y}$ be a set of M points from modality $\mathcal{Y}$. Moreover, we are given two types of cross-modal constraints: (i) pairwise constraint matrix $\mathbf{S}$ of size $N \times M$ where S_{ij} indicates the pairwise constraint between $\mathbf{x}_i$ and $\mathbf{y}_j$. Specifically, $S_{ij} = 1$ if the two points are similar, and $S_{ij} = 0$ otherwise. (ii) Triplet constraints $\mathcal{T} = \{\mathbf{T}^{(i)}, i = 1, \ldots, N\}$ given as a set of binary matrices encoding cross-modal *relative* comparisons. $T^{(i)}_{jl} = 1$ indicates that a modality $\mathcal{X}$ object $\mathbf{x}_i$ is more similar to modality $\mathcal{Y}$ object $\mathbf{y}_j$ than $\mathbf{y}_l$.[1] Also note that, in practice, such constraints may only be available for a very small fraction of the cross-modal pairs and the cross-modal triplets (or, in an active learning setting, there may be a budget on the number of constraints that can be acquired).

For each object in modality $\mathcal{X}$, we associate a latent variable $\mathbf{u} \in \mathbb{H}^K$ where $\mathbb{H}^K = \{1, -1\}^K$ is a K-dimensional Hamming space. Similarly, we associate a latent variable $\mathbf{v} \in \mathbb{H}^L$ for each object in modality $\mathcal{Y}$. For simplicity of exposition, in Figure 1 and elsewhere in the text, we have shown the case of $K = L$, but the code lengths in different modalities can be the same or different, depending on the specific application. In either case, we will infer the appropriate code lengths adaptively from the data, taking a nonparametric Bayesian approach (Griffiths and Ghahramani 2011).

Our framework further consists of a set of latent variables Ω for generating the constraints, and their specific form depends on the likelihood functions for the constraints that we will introduce later. As shown in Figure 1, we assume that the *probability* of bit k of $\mathbf{u}$ (resp., $\mathbf{v}$) being 1 depends on a regression coefficient vector $\mathbf{f}_k \in \mathbb{R}^{D_x}$ applied on the raw features $\mathbf{x}$, and a positive scalar a_k (resp., on a regression coefficient vector $\mathbf{g}_k \in \mathbb{R}^{D_y}$ applied on the raw features $\mathbf{y}$, and a positive scalar b_k). We denote by $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]^T \in \mathbb{H}^{N \times K}$ the binary codes for the N objects in modality $\mathcal{X}$, and denote by $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]^T \in \mathbb{H}^{M \times K}$ the binary codes for the M objects in modality $\mathcal{Y}$. We group the regression coefficient vectors in a column-wise fashion and represent them using two matrices $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_K] \in \mathbb{R}^{D_x \times K}$ and $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_K] \in \mathbb{R}^{D_y \times K}$. We use vectors $\mathbf{a}$ and $\mathbf{b}$ to denote the sets $\{a_k\}_{k=1}^K$ and $\{b_k\}_{k=1}^K$, respectively.

[1] Although we only consider triplets consisting of one object from $\mathcal{X}$ and two objects from $\mathcal{Y}$, it is straightforward to incorporate triplets consisting of two objects from $\mathcal{X}$ and one object from $\mathcal{Y}$.

Data-Dependent Priors for Binary Codes

We use a data-dependent variant of the Indian Buffet Process (IBP) (Griffiths and Ghahramani 2011) as the prior distribution on the latent binary matrices $\mathbf{U}$ and $\mathbf{V}$ (the binary codes). The choice of this prior accomplishes two things: (i) inferring the apprropriate code length from the data using the IBP, and (ii) enabling us to introduce the raw features of the objects in the IBP prior distribution (Quadrianto et al. 2013) over the binary matrices $\mathbf{U}$ and $\mathbf{V}$; this allows predicting the binary codes for *out-of-sample* objects from either modality $\mathcal{X}/\mathcal{Y}$ based *solely* on their features $\mathbf{x}_i/\mathbf{y}_j$. The data-dependent prior is defined as: $p(\mathbf{U}) = \prod_{i=1}^N \prod_{k=1}^K \text{Bern}(U_{ik} \mid \Phi_{0,1}(\mathbf{f}_k^T \mathbf{x}_i + \Phi_{0,1}^{-1}(a_k)))$, and $p(\mathbf{V}) = \prod_{j=1}^M \prod_{k=1}^K \text{Bern}(V_{jk} \mid \Phi_{0,1}(\mathbf{g}_k^T \mathbf{y}_j + \Phi_{0,1}^{-1}(b_k)))$, where $\Phi_{0,1}(\cdot)$ is the cumulative density function of univariate Gaussian distribution with mean 0 and variance 1, and $\Phi_{0,1}^{-1}(\cdot)$ denotes the inverse function of $\Phi_{0,1}(\cdot)$. The above equations essentially define a probit-model for the binary variables in $\mathbf{U}$ and $\mathbf{V}$, such that the probabilities of these binary variables taking a value 1 depends on the features $\mathbf{x}$ and $\mathbf{y}$. Without the features, the prior reduces to the standard stick-breaking representation of the IBP (Teh, Görür, and Ghahramani 2007). In fact, we can think of the random variables a_k and b_k as the prior probabilities of bit k of the binary code being 1 for an object of modality $\mathcal{X}$ and $\mathcal{Y}$, respectively, in the absence of the raw features. We use the stick-breaking representation of the IBP to define the priors on a_k and b_k that can be written as $\forall k = 1, \ldots, K$: $a_k = \prod_{t=1}^k \mu_t, \quad \mu_t \sim \text{Beta}(\alpha_a, 1)$ and $b_k = \prod_{t=1}^k \nu_t, \quad \nu_t \sim \text{Beta}(\alpha_b, 1)$, and impose Gaussian priors on $\mathbf{F}$ and $\mathbf{G}$: $p(\mathbf{F}) = \prod_{k=1}^K \mathcal{N}(\mathbf{f}_k \mid \mathbf{0}, \sigma_{\mathbf{f}}^2 \mathbf{I})$ and $p(\mathbf{G}) = \prod_{k=1}^K \mathcal{N}(\mathbf{g}_k \mid \mathbf{0}, \sigma_{\mathbf{g}}^2 \mathbf{I})$ where $\alpha_a, \alpha_b, \sigma_{\mathbf{f}}$ and $\sigma_{\mathbf{g}}$ are hyperparameters which were set to 1 for our experiments and it worked well in practice. Alternatively, we can put hyperpriors on these and learn them from data.

Modeling Pairwise and Preference Constraints

For our framework, we propose three types of likelihood functions to model the cross-modal pairwise and cross-modal triplet constraints.

Logistic Model Likelihood Given the binary codes $\mathbf{u}_i$ and $\mathbf{u}_j$ of a cross-modal pair, the logistic model defines the *probability* μ_{ij} of this pair to be similar as: $\mu_{ij} = \sigma(\mathbf{u}_i^T \mathbf{W} \mathbf{v}_j)$, where $\sigma(x) = 1/(1 + \mathrm{e}^{-x})$ is the logistic sigmoid function. The $K \times K$ matrix $\mathbf{W}$ plays the role of a weighting matrix. The logistic model also allows the binary code sizes of the two modalities to be different (say, K and L) using

a rectangular weighting matrix $\mathbf{W}$ of size $K \times L$. The joint probability of the cross-modal pairwise constraint matrix $\mathbf{S}$ (assuming iid constraints) is:

$$p(\mathbf{S}) = \prod_{i=1}^{N} \prod_{j=1}^{M} [\text{Bern}(S_{ij} \mid \mu_{ij})]^{\phi_{ij}}, \quad (1)$$

where $\text{Bern}(\cdot \mid \mu)$ denotes the Bernoulli distribution with parameter μ, ϕ_{ij} is an indicator variable which is equal to 1 if the constraint S_{ij} is observed and 0 otherwise, and Ω consists of the other model parameters such as $\mathbf{W}$. Likewise, the joint probability of the cross-modal triplet-based constraint matrices $\mathcal{T} = \{\mathbf{T}^{(i)}, i = 1, \ldots, N\}$ is defined as:

$$p(\mathcal{T}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \prod_{l=1}^{M} \left[\text{Bern}(T_{jl}^{(i)} \mid \gamma_{jl}^{(i)})\right]^{\psi_{ijl}}, \quad (2)$$

where $\gamma_{jl}^{(i)} = \mu_{ij} - \mu_{il}$ if $\mu_{ij} - \mu_{il} > 0$ and $\gamma_{jl}^{(i)} = 0$ otherwise, and ψ_{ijl} is an indicator variable, which is 1 if $T_{jl}^{(i)}$ is observed, and 0 otherwise. It is also possible to replace this simple form of preference with other more sophisticated forms such as preference based on probit and logit models (Bonilla, Guo, and Sanner 2010; Houlsby et al. 2012). For this model, we assume positive-valued weighting matrix $\mathbf{W}$, and impose a half-normal prior on $\mathbf{W}$: $p(\mathbf{W}) = \prod_{k=1}^{K} \prod_{k'=k}^{K} \mathcal{HN}(W_{kk'} \mid 0, \sigma_{\mathbf{W}}^2)$, where $\mathcal{HN}(x \mid 0, \sigma_{\mathbf{W}}) = \sqrt{2/(\pi\sigma_{\mathbf{W}})}\text{e}^{-x^2/(2\sigma_{\mathbf{W}}^2)}$. For real-valued $\mathbf{W}$, we can instead use a Gaussian prior.

Choice Model Likelihood Our choice model likelihood function for cross-modal constraints is inspired by the probabilistic choice model (Görür, Jäkel, and Rasmussen 2006). The basic idea is to use a weighted sum of the *matching bits* of the binary codes of a cross-modal object pair, and use the score of the match to define the cross-modal pairwise probabilities. Specifically, given $\mathbf{U}, \mathbf{V}$ and Ω, the joint probability of the cross-modal pairwise constraints $\mathbf{S}$ can be defined similarly as in Eqn. (1) but with different $\mu_{ij} = \frac{1}{C_1} \sum_{k=1}^{K} w_k \mathbb{I}[U_{ik} = V_{jk}]$, where $C_1 = \sum_{k=1}^{K} w_k$ is a normalization term and $\mathbb{I}[\cdot]$ returns 1 if its argument is true and 0 otherwise. Note that Ω in this case becomes a weighting vector $\mathbf{w} \in \mathbb{R}_+^K$. Likewise, the joint probability of the cross-modal triplets is defined as in Eqn. (2) but $\gamma_{jl}^{(i)}$ is defined as $\gamma_{jl}^{(i)} = \frac{1}{\tilde{C}_1} \sum_{k=1}^{K} w_k \mathbb{I}[U_{ik} = V_{jk}](1 - \mathbb{I}[U_{ik} = V_{lk}])$, and $\tilde{C}_1 = \sum_{k=1}^{K} w_k \mathbb{I}[U_{ik} = V_{jk}](1 - \mathbb{I}[U_{ik} = V_{lk}]) + \sum_{k=1}^{K} w_k (1 - \mathbb{I}[U_{ik} = V_{jk}])\mathbb{I}[U_{ik} = V_{lk}]$. Intuitively, we can interpret this likelihood function as follows: for each triplet $\{i, j, l\}$, if the binary code of object i in modality $\mathcal{X}$ shares more bits with the code of object j of modality $\mathcal{Y}$ than with the code of object l of modality $\mathcal{Y}$, then probability of $T_{jl}^{(i)}$ being 1 will be high. For this model, we put a Gamma prior on each entry of the weighting vector $\mathbf{w}$: $p(\mathbf{w}) = \prod_{k=1}^{K} \text{Gam}(w_k \mid \gamma_{\mathbf{w}}, \theta_{\mathbf{w}})$.

Margin based Likelihood For many applications, such as hashing based approximate similarity search (Grauman and Fergus 2013), semantic similarities are typically defined using pairwise Hamming distances. To reflect this aspect in a probabilistic framework like ours, we propose a margin-based likelihood function which directly models the relationship between Hamming distance and the associated constraint probabilities. Specifically, given $\mathbf{U}, \mathbf{V}$ and Ω, the joint probability of the cross-modal pairwise constraints is defined in Eqn. (1), where $\mu_{ij} = \frac{1}{C_2}\text{e}^{-[\|\mathbf{u}_i - \mathbf{v}_j\|_H - \rho_1 + 1]_+}$, $C_2 = \text{e}^{-[\|\mathbf{u}_i - \mathbf{v}_j\|_H - \rho_1 + 1]_+} + \text{e}^{-[\rho_2 - \|\mathbf{u}_i - \mathbf{v}_j\|_H + 1]_+}$, and Ω consists of two random variables ρ_1 and ρ_2 that control the margins. $[a]_+$ returns a if $a > 0$ and 0 otherwise. The joint probability of the cross-modal triplets constraints is defined in Eqn. (2) where $\gamma_{jl}^{(i)} = \text{e}^{-[\rho_3 + \|\mathbf{u}_i - \mathbf{v}_j\|_H - \|\mathbf{u}_i - \mathbf{v}_l\|_H + 1]_+}$ and $\Omega = \{\rho_3\}$ controls the relative margins. For the margin based likelihood functions, we simply use the uniform distribution as the prior of each margin variable: $\rho_1 \sim \text{Unif}[0, K]$, $\rho_2 \sim \text{Unif}[0, K]$, and $\rho_3 \sim \text{Unif}[0, K]$, where $\text{Unif}[a, b]$ denotes the Uniform distribution in $[a, b]$.

Active Sampling for Constraints

Existing methods for learning cross-modal similarities (Bronstein et al. 2010; Zhen and Yeung 2012a; 2012b) assume that the learner has no control over the constraints it gets to see. Since not all constraints are equally informative and, moreover, since constraints are often costly to acquire, we present an active sampling scheme which allows the learner to decide which constraints to acquire. To the best of our knowledge, this has not been done before for cross-modal similarity learning problems.

Our goal is to select the most informative pairwise or triplet constraints from the pool of all *potential* constraints. One common principle to select the most informative constraints is the uncertainty principle, i.e., the most informative constraints should be ones that the current learner is the most uncertain about. At the same time, we wish to have a *diverse* set of constraints to avoid information redundancy.

For our CSLP framework with logistic model likelihood for the constraints,[2] we can define the informativeness of a pairwise constraint as $\varepsilon(\mathbf{x}_i, \mathbf{y}_j) = \Delta_x(\mathbf{x}_i) + \Delta_y(\mathbf{y}_j) + \Pi(\mathbf{x}_i, \mathbf{y}_j)$. We define $\Delta_x(\mathbf{x}) = \sum_{k=1}^{K} -p_k \log(p_k) - (1 - p_k)\log(1 - p_k)$ where $p_k = \mathbb{E}_{\mathbf{f}_k, a_k}\left[\Phi(\mathbf{f}_k^T \mathbf{x} + \Phi^{-1}(a_k))\right]$, $\Delta_y(\mathbf{y}) = \sum_{k=1}^{K} -q_k \log(q_k) - (1 - q_k)\log(1 - q_k)$ where $q_k = \mathbb{E}_{\mathbf{g}_k, b_k}\left[\Phi(\mathbf{g}_k^T \mathbf{y} + \Phi^{-1}(b_k))\right]$, and $\Pi(\mathbf{x}, \mathbf{y}) = -r \log(r) - (1 - r)\log(1 - r)$ where $r = \mathbb{E}_{\mathbf{u}, \mathbf{v}, \mathbf{W}}\left[\sigma(\mathbf{u}^T \mathbf{W} \mathbf{v})\right]$, and $\mathbf{u}$ and $\mathbf{v}$ are binary codes for $\mathbf{x}$ and $\mathbf{y}$, respectively. The informativeness of a triplet constraint is similarly defined: $\hat{\varepsilon}(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) = \Delta_x(\mathbf{x}) + \Delta_y(\mathbf{y}_1) + \Delta_y(\mathbf{y}_2) + \hat{\Pi}(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ where $\hat{\Pi}(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) = -t \log(t) - (1 - t)\log(1 - t)$, and $t = \mathbb{E}_{\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2, \mathbf{W}}\left[\max(0, \sigma(\mathbf{u}^T \mathbf{W} \mathbf{v}_1) - \sigma(\mathbf{u}^T \mathbf{W} \mathbf{v}_2))\right]$.

We first describe the selection criteria for the pairwise constraints. Given a candidate set of C pairwise constraints $\mathcal{C} = \{(\mathbf{x}_{i_1}, \mathbf{y}_{j_1}), (\mathbf{x}_{i_2}, \mathbf{y}_{j_2}), \ldots, (\mathbf{x}_{i_C}, \mathbf{y}_{j_C})\}$, we want to select a group of P most informative pairwise constraint, based on the *uncertainty* and *diversity* principle:

$$\min_{\boldsymbol{\mu}} \frac{\lambda}{P} \boldsymbol{\mu}^T \mathbf{K} \boldsymbol{\mu} - \boldsymbol{\mu}^T \boldsymbol{\varepsilon}, \text{ s.t. } \boldsymbol{\mu} \in \{0, 1\}^C, \boldsymbol{\mu}^T \mathbf{1} = P, \quad (3)$$

[2] Note that definitions for other likelihood functions can be derived similarly.

where $\boldsymbol{\varepsilon} \in \mathbb{R}^C$ consists of information values of the C candidate pairs, $\mathbf{K}$ is a $C \times C$ matrix whose (c, c')th element measures the similarity between pairwise constraints c and c', and $\boldsymbol{\mu} \in \{0,1\}^C$ is an indicator vector denoting which pairwise constraints are selected. The objective function in Problem (3) has an intuitive meaning: it promotes selecting constraints that have have high information value (via the term $-\boldsymbol{\mu}^T\boldsymbol{\varepsilon}$) and, at the same time, forces μ_c and $\mu_{c'}$ to take different values (one to be 0 and the other to be 1) if the constraints c and c' are deemed similar based on $K_{cc'}$ (so encouraging only one of them to be selected). The parameter λ controls the trade-off between informativeness and redundancy. To measure the redundancy, we define the (c, c')th element of $\mathbf{K}$ as $K_{cc'} = \mathrm{e}^{-(\varphi_x(\mathbf{x}_{c_i}, \mathbf{x}_{c'_i}) + \varphi_y(\mathbf{y}_{c_i}, \mathbf{y}_{c'_i}))/\pi^2}$, and $D_{\mathrm{KL}}(p\|q)$ denotes the KL divergence between two distributions p and q. $p(\mathbf{u}_{c_i})$ is the posterior distribution of $\mathbf{u}_{c_i}$, and each of its entry follows a Bernoulli distribution $p(u_{c_i k} = 1 \mid \Phi(\mathbf{f}_k^T \mathbf{x}_{c_i} + \Phi^{-1}(a_k))), \forall k = 1, \ldots, K$. $q(\mathbf{v}_{c_j})$ is the posterior distribution of $\mathbf{v}_{c_j}$, and each of its entry follows a Bernoulli distribution $q(v_{c_j k} = 1 \mid \Phi(\mathbf{g}_k^T \mathbf{y}_{c_j} + \Phi^{-1}(b_k))), \forall k = 1, \ldots, K$. Specifically, we have $D_{\mathrm{KL}}(p(\mathbf{u}_{c_i})\|p(\mathbf{u}_{c'_i})) = \sum_{k=1}^{K} \sum_{b \in \{1,-1\}} \left(\ln \frac{p(u_{c_i k} = b)}{p(u_{c'_i k} = b)} p(u_{c_i k} = b) \right)$. Problem (3) is hard to solve since it is an integer program, but we can relax it to a QP problem and solve it approximately: $\min_{\boldsymbol{\mu}} \ \frac{\lambda}{P} \boldsymbol{\mu}^T \mathbf{K} \boldsymbol{\mu} - \boldsymbol{\mu}^T \boldsymbol{\varepsilon}$, s.t. $\boldsymbol{\mu} \in \mathbb{R}^C, \boldsymbol{\mu}^T \mathbf{1} = P$.

The selection criteria for triplets is similar: for a set of $\hat{C}$ triplets, we choose a group of $\hat{P}$ triplets by solving:

$$\min_{\boldsymbol{\mu}} \frac{\hat{\lambda}}{\hat{P}} \boldsymbol{\mu}^T \hat{\mathbf{K}} \boldsymbol{\mu} - \boldsymbol{\mu}^T \hat{\boldsymbol{\varepsilon}}, \ \text{s.t.} \, \boldsymbol{\mu} \in \{0,1\}^{\hat{C}}, \boldsymbol{\mu}^T \mathbf{1} = \hat{P}, \quad (4)$$

where $\hat{K}_{cc'} = \mathrm{e}^{-\left(\varphi_x(\mathbf{x}_{c_1}, \mathbf{x}_{c'_1}) + \varphi_y(\mathbf{y}_{c_2}, \mathbf{y}_{c'_2}) + \varphi_y(\mathbf{y}_{c_3}, \mathbf{y}_{c'_3})\right)/\hat{\pi}^2}$, $\hat{\boldsymbol{\varepsilon}} \in \mathbb{R}^{\hat{C}}$ consists of the information values of all candidate triplets, $\hat{\mathbf{K}} \in \mathbb{R}^{\hat{C} \times \hat{C}}$ measures the similarity between the triplets, and $\hat{\lambda}$ is a user provided parameter controlling the trade-off between informativeness and redundancy. A relaxation similar to the one used for the pairwise constraints case is used to solve this problem.

Inference

As exact inference is computationally intractable, we use Markov Chain Monte Carlo (MCMC), combined with slice sampling, to draw samples from posterior distribution (5).

$$p(\mathbf{U}, \mathbf{V}, \mathbf{F}, \mathbf{G}, \mathbf{a}, \mathbf{b}, \Omega \mid \mathbf{S}, \mathcal{T}, \Sigma) \propto p(\mathbf{S}) p(\mathcal{T}) p(\mathbf{U}) p(\mathbf{V}) p(\Omega) p(\mathbf{F}) p(\mathbf{G}), \quad (5)$$

where we use Σ to denote all the hyperparameters.

Inferring the Code Length Since both $\mathbf{U}$ and $\mathbf{V}$ are given the data-dependent IBP priors, the code length of each modality can be automatically inferred from data. Borrowing the basic idea of stick-breaking representation of IBP (Teh, Görür, and Ghahramani 2007), we assume that the sequence $\{a_k\}$ (resp., $\{b_k\}$) is decreasing where $k = 1, \ldots, K$, and $\mathbf{U}$ (resp., $\mathbf{V}$) only contains *active* columns (i.e., columns that have at least a single bit equal to 1). We first sample a slice variable s from $s \sim \mathrm{Unif}(0, a^*)$, where $a^* = \min\{1, \min_{k:\exists i, U_{ik}=1} a_k\}$ is the minimal value among the active features. Then, we sample $\hat{a}$ using slice sampling from the following distribution: $p(\hat{a}) \propto \mathrm{e}^{\alpha_a \sum_{i=1}^{N} \frac{1}{i}(1-\hat{a})^i} \hat{a}^{\alpha_a - 1}(1-\hat{a})^N \mathbb{I}[0 \le \hat{a} \le a_K]$. If $\hat{a} >= s$, we add a new $+1$ bit for the data point being processed (by adding a new column) to $\mathbf{U}$ and set the value of this bit to -1 for all other points. Besides, we sample corresponding variables for this new bit. Specifically, let $K = K + 1$, we sample $\mathbf{f}_K, \mathbf{g}_K$, and Ω from their prior distributions, respectively. We continue above procedure until $\hat{a} < s$. Also note that in the cases when we require the binary codes to be the same in both modalities, the code length is determined by the IBP for the modality that has richer information and the other modality uses the same code length (the IBP sampler for this modality simply uses a truncation level equal to the code length of the first modality).

Sampling U and V: We sample each element of $\mathbf{U}$ according to its posterior distribution. Specifically, for the ith point and kth bit, we sample U_{ik} from the following distribution: $\Pr(U_{ik} = 1 \mid \text{rest}) \propto \Pr(U_{ik} = 1 \mid \mathbf{x}_i, \mathbf{f}_k, a_k) \times p(\mathbf{S} \mid \mathbf{U}_{-ik}, U_{ik} = 1, \mathbf{V}, \Omega) \times p(\mathcal{T} \mid \mathbf{U}_{-ik}, U_{ik} = 1, \mathbf{V}, \Omega)$, where $\Pr(U_{ik} = 1 \mid \mathbf{x}_i, \mathbf{f}_k, a_k) = \Phi_{0,1}(\mathbf{f}_k^T \mathbf{x}_i + \Phi_{0,1}^{-1}(a_k))$ and $\mathbf{U}_{-ik}$ denotes all the elements in $\mathbf{U}$ except U_{ik}. The matrix $\mathbf{V}$ is sampled in an analogous way.

Sampling F and G: We sample $\mathbf{f}_k, k = 1, \ldots, K$, sequentially using elliptical slice sampling (Murray, Adams, and MacKay 2010) from the following distribution: $p(\mathbf{f}_k \mid \text{rest}) \propto \frac{1}{\sigma_{\mathbf{f}}} \mathrm{e}^{-\frac{\mathbf{f}_k^T \mathbf{f}_k}{2\sigma_{\mathbf{f}}^2}} \prod_{i=1}^{N} \mathrm{Bern}(U_{ik} \mid \Phi_{0,1}(\mathbf{f}_k^T \mathbf{x}_i + \Phi_{0,1}^{-1}(a_k)))$. The matrix $\mathbf{G}$ is sampled in an analogous way.

Sampling a and b: Let the $\hat{k}$th bit be an active bit in $\mathbf{U}$, in other words, $\hat{k} \le K$ and $\exists_{i=1,\ldots,N} U_{i\hat{k}} > 0$. We sample $a_{\hat{k}}$ from the following conditional distribution: $p(a_{\hat{k}} \mid \text{rest}) \propto a_{\hat{k}}^{n_{\hat{k}} - 1}(1 - a_{\hat{k}})^{N - n_{\hat{k}}} \mathbb{I}[a_{\hat{k}+1} \le a_{\hat{k}} \le a_{\hat{k}-1}]$, where $n_{\hat{k}} = \sum_{i=1}^{N} \mathbb{I}[U_{i\hat{k}} > 0]$ is the number of points with bit $\hat{k}$ active. In the other hand, let the $\tilde{k}$th bit be an inactive bit and we draw $a_{\tilde{k}}$ from the following distribution: $p(a_{\tilde{k}} \mid \text{rest}) \propto \mathrm{e}^{\alpha_a \sum_{i=1}^{N} \frac{1}{i}(1-a_{\tilde{k}})^i} a_{\tilde{k}}^{\alpha_a - 1}(1 - a_{\tilde{k}})^N \mathbb{I}[0 \le a_{\tilde{k}} \le a_{\tilde{k}-1}]$. The variables in $\mathbf{b}$ are sampled in an analogous way.

Sampling Ω: For the logistic model likelihood, we sample each element of $\mathbf{W}$ using slice sampling from the following distribution: $p(\mathbf{W} \mid \text{rest}) \propto \prod_{k=1}^{K} \prod_{k' \ne k}^{K} \mathcal{N}(W_{kk'} \mid 0, \sigma_{\mathbf{W}}^2) p(\mathbf{S}) p(\mathcal{T})$. For choice model based likelihood functions, we sample each element of $\mathbf{w}$ via slice sampling from: $p(\mathbf{w} \mid \text{rest}) \propto \prod_{k=1}^{K} \mathrm{Gam}(w_k \mid \gamma_{\mathbf{w}}, \theta_{\mathbf{w}}) p(\mathbf{S}) p(\mathcal{T})$. For margin based likelihood functions, ρ_1, ρ_2 and ρ_3 can be sampled sequentially through slice sampling from: $p(\rho_1 \mid \text{rest}) \propto \frac{1}{K+1} p(\mathbf{S})$, $p(\rho_2 \mid \text{rest}) \propto \frac{1}{K+1} p(\mathbf{S})$, and $p(\rho_3 \mid \text{rest}) \propto \frac{1}{K+1} p(\mathcal{T})$.

Out-of-sample Prediction: For new points $\mathbf{x}^\dagger \in \mathcal{X}$ and $\mathbf{y}^\dagger \in \mathcal{Y}$ which do not appear in model training, the binary codes can be obtained from the following conditional distributions: $\Pr(u_k^\dagger = 1 \mid \mathbf{f}_k, b_k, \mathbf{x}^\dagger) = \Phi_{0,1}(\mathbf{f}_k^T \mathbf{x}^\dagger + \Phi_{0,1}^{-1}(a_k))$ and $\Pr(v_k^\dagger = 1 \mid \mathbf{g}_k, b_k, \mathbf{y}^\dagger) = \Phi_{0,1}(\mathbf{g}_k^T \mathbf{y}^\dagger + \Phi_{0,1}^{-1}(b_k))$.

Related Work

The problem of learning multimodal similarities (also sometimes referred to multimodal metric learning and multimodal hashing in some other works) has received a significant amount of interest recently (Jia, Salzmann, and Darrell 2011; Bronstein et al. 2010; Kumar and Udupa 2011; Ou et al. 2013; Zhen and Yeung 2012a; Masci et al. 2013; Zhai et al. 2013; Zhen and Yeung 2012b). Of these, the closest in spirit are methods based on multimodal hashing or binary coding (Bronstein et al. 2010; Kumar and Udupa 2011; Ou et al. 2013; Zhen and Yeung 2012a; Masci et al. 2013; Zhai et al. 2013; Zhen and Yeung 2012b; Rastegari et al. 2013) where the goal is to learn binary code for objects in each modality and use the binary codes to compare objects across modalities. However, most of these methods work assume one-to-one correspondence or fully-paired objects (Kumar and Udupa 2011; Rastegari et al. 2013) across different modalities, or can only leverage pairwise relationships, and cannot incorporate weaker forms of supervision such as *relative preferences* based on triplet constraints. Another limitation is that the constraints are provided in a *passive* manner, i.e., the learner has no control over which constraints it gets to see. Moreover, most of these method are non-probabilistic without a explicit generative model of the data and therefore cannot be used for tasks such as cross-modal link-prediction.

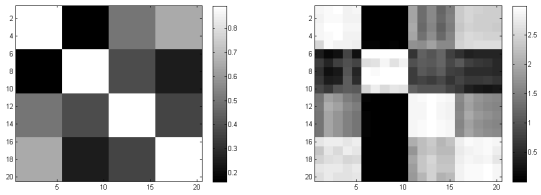
Probabilistic approaches to learning multimodal similarities are relatively few. The only work we are aware of include the Cross-Modal Similarity Learning (Jia, Salzmann, and Darrell 2011) which is however limited to modeling multinomial data (e.g., text, or images represented as bag of *visual* words), and the Multimodal Latent Binary Embedding (MLBE) model (Zhen and Yeung 2012b) which can only leverage *intra-modal* pairwise similarities and *partially-known* cross-modal pairwise constraints. Moreover, unlike our proposed framework, these models do not learn explicit hash functions and therefore predicting the hash codes for out-of-sample data is computationally expensive (Zhai et al. 2013). Also, MLBE and all the other methods discussed above require pre-specifying the binary code length (which is typically chosen via cross-validation). The only hashing method which allows learning the code length adaptively from the data was proposed recently (Quadrianto et al. 2013), but it is limited to the unimodal setting.

Some recent works (Wang et al. 2013b; 2013a; Li et al. 2013; Quadrianto et al. 2013) have begun looking at incorporating triplets or other implicit forms of constraints for learning binary codes, but only in *unimodal* settings. The only other works we are aware of that includes triplet based supervision in cross-modal settings include (Mignon, Jurie, and others 2012; Kuang and Wong 2013). These are however based on learning real-valued low-dimensional projections, requires the length of the projection matrix to be specified, lacks a probabilistic formulation, and is limited to only learning a projection of the data without any explicit model for the pairwise relations between the objects (and therefore cannot be used for cross-modal link-prediction).

Experiments

We perform experiments using the CSLP framework on two tasks: (i) cross-modal link prediction, and (ii) cross-modal hashing based image vs. text retrieval. We then show experiments using the active sampling based variant of CSLP comparing with *passive* CSLP. For the link prediction task, we use the AUC measure (Cortes and Mohri 2003) to evaluate the model performance; for the cross-modal retrieval task, we use Mean Average Precision (mAP) (Yue et al. 2007) as the evaluation measure.

As a sanity check for our model CSLP, we conduct an experiment on a two-modality synthetic dataset with 20 points from each modality generated using a linear Gaussian model (Griffiths and Ghahramani 2011): $\mathbf{x}_i = \mathbf{W}_x\mathbf{z}_i + \boldsymbol{\varepsilon}$ and $\mathbf{y}_j = \mathbf{W}_y\mathbf{z}_j + \boldsymbol{\varepsilon}$, where the binary codes $\mathbf{z}_i$ and $\mathbf{z}_j$ are assumed to have one of the 4 possible values $\mathbf{z}_1 = [1,-1,-1,-1]^T, \mathbf{z}_2 = [-1,1,-1,1]^T, \mathbf{z}_3 = [-1,-1,1,-1]^T, \mathbf{z}_4 = [1,1,1,-1]^T$, denoting the 4 underlying classes. Using the class information, we generate 50 pairwise constraints using half of $\mathbf{X}$ and the whole $\mathbf{Y}$; and 100 triplet constraints using the other half of $\mathbf{X}$ and the whole $\mathbf{Y}$ (to ensure no redundancy between pairwise and triplet constraints). As Figure 2 shows, our model can infer the cross-modal link probabilities accurately. On this data, our models also recovered the correct binary code size (4). We also provide a quantitative comparison, in terms of AUC on link prediction, in Table 1.



(a) Ground Truth (b) Inferred

Figure 2: Similarity on Synthetic Data

Real-world Datasets In the experiments, we use three real-world data sets, namely, Drug[3], Wiki[4] and Flickr[5]: (i) **Drug** is a drug-protein interaction network representing interactions between 200 drug molecules and 150 target proteins. In addition, we also have access to features of the drugs and the proteins. (ii) **Wiki** is generated from Wikipedia featured articles and consists of 2,866 image-text pairs. In each pair, the text is an article describing some events or people and the image is closely related to the content of the article. (iii) **Flickr** consists of 186,577 image-tag pairs pruned from the NUS data set by keeping the pairs from the largest 10 classes.

Baselines We denote by CSLP–L, CSLP–C, and CSLP–M the three variants of our framework CLSP (without active sampling), namely the logistic model likelihood, choice

[3] http://www.genome.jp/tools/dinies/help.html

[4] http://www.svcl.ucsd.edu/projects/cross-modal/

[5] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

Table 1: AUC Comparison on Cross–Modal Link Prediction

Method	**Synthetic**	**Drug**	**Wiki**	**Flickr**
MLBE	0.5184 ± 0.0992	0.4566 ± 0.1379	0.5019 ± 0.0130	0.4870 ± 0.0138
B–LFRM	0.5512 ± 0.0531	0.7212 ± 0.0306	0.5342 ± 0.0422	0.5964 ± 0.0354
Sim–B–LFRM	0.7354 ± 0.0250	0.7432 ± 0.0238	0.5512 ± 0.0322	0.5506 ± 0.0364
CSLP–L	$\mathbf{0.8173 \pm 0.0317}$	$\mathbf{0.9232 \pm 0.0216}$	$\mathbf{0.7697 \pm 0.0068}$	$\mathbf{0.7104 \pm 0.0154}$
CSLP–C	0.7884 ± 0.0294	0.9045 ± 0.0343	0.6218 ± 0.0174	0.6981 ± 0.0316
CSLP–M	0.7983 ± 0.0322	0.8744 ± 0.0311	0.7414 ± 0.0197	0.6729 ± 0.0169

Table 2: mAP Comparison on Cross–Modal Retrieval

Methods	**Wiki (Query *vs.* Database)**		**Flickr (Query *vs.* Database)**	
	Image *vs.* Text	Text *vs.* Image	Image *vs.* Text	Text *vs.* Image
CSLP–L	0.2187 ± 0.0141	0.2005 ± 0.0150	0.3887 ± 0.0117	0.3947 ± 0.0091
CSLP–C	0.2240 ± 0.0227	$\mathbf{0.2590 \pm 0.0342}$	$\mathbf{0.4065 \pm 0.0097}$	$\mathbf{0.4099 \pm 0.0090}$
CSLP–M	$\mathbf{0.2257 \pm 0.0142}$	0.2537 ± 0.0288	0.4062 ± 0.0071	0.4082 ± 0.0073
MLBE	0.1766 ± 0.0139	0.1634 ± 0.0011	0.3833 ± 0.0087	0.3883 ± 0.0051
CVH	0.2034 ± 0.0103	0.2063 ± 0.0138	0.3843 ± 0.0079	0.3826 ± 0.0027

model likelihood, and margin based likelihood, respectively. We compare these with the several state-of-the-art cross-modal link prediction and cross-modal hashing methods: (i) **MLBE**: Multimodal latent binary embedding (Zhen and Yeung 2012b) for cross-modal link prediction and cross-modal retrieval; (ii) **Sim–B–LFRM**: Intra-modal similarity informed bipartite latent feature relational model (Whang, Rai, and Dhillon 2013) for the cross-modal link prediction; (iii) **CVH**: Cross-view hashing (Kumar and Udupa 2011) for the cross-modal retrieval.

Cross–Modal Link Prediction In this subsection, we consider the task of cross-modal link prediction to identify if two objects from different modalities (e.g. image and text) should have a link between them. This is essentially a bipartite graph matching problem. For this task, we run the experiments on the synthetic data (described above), drug-protein data (to predict matching drug and protein pairs), and Wiki and Flickr data (to predict text to image associations). Note that the baselines are not able to exploit the triplet constraints.

In Table 1, we observe that all three variants of CSLP achieve higher AUC values than the baselines, indicating the advantages of using both pairwise and triplet cross-modal constraints (in addition to the raw features of objects in each modality) for model learning. Moreover, we observe that CSLP–L performs better than the other two variants, i.e., CSLP–C and CSLP–M. The reason can be attributed to the additional flexibility of CSLP–L in modeling the links by using a dense weighting matrix $\mathbf{W}$ which can model interactions between all pairs of bits in the binary codes of the two modalities.

Cross–Modal Retrieval We next compare CSLP with a probabilistic multi-modal hashing method, MLBE, and a non-probabilistic multi-modal hashing method, CVH, on a cross-modal retrieval task. Given an image (text) as a query, the task is to find the nearest neighbors from a text (image) database. To generate the observations, we randomly select 5000 pairs and 5000 triplets on Wiki, and 1000 pairs and 6000 triplets on Flickr. For Wiki, we use 20% data as the query set and 80% the database set; for Flickr, 1% data are chosen to form the query set and the remaining 99% the database set. All the methods are trained on the same 5 random training sets, and then applied to the same query and database sets. For fair comparison, we train all the methods using random initialization and report the results averaged over 5 repeats.

In Table 2, we note that CSLP in general performs significantly better than the baselines and CSLP–C achieves the highest mAP, on both datasets. CVH performs better than MLBE, as is reasonable because CVH uses aligned data whereas MLBE only uses an extremely small fraction of pairwise constraints. The improved performance achieved by CSLP clearly validates the advantages of taking both pairwise and triplet constraints into account simultaneously.

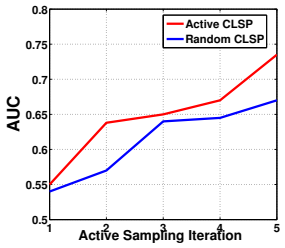


Figure 3: Comparison of active vs random selection

Active Sampling for Constraints We finally experiment with our active sampling scheme for the *informed* selection of pairwise and triplet constraints. For this experiment, we compare our *active* supervision selection strategy with randomly selected constraints on the Drug dataset. Figure 3 shows the performance curves of the two strategies (both using the logistic variant of CLSP) with increasing number of constraint sampling iterations. Each method was initially given 100 pairwise and 100 triplet constraints, and at each

constraint sampling iteration, 50 pairs and 50 triplets selected by each method (active and random) were added into the training set and the models were retrained and evaluated. As the figure shows, active sampling performs significantly better than random sampling, validating the effectiveness of our constraint selection mechanism.

Conclusion

We have presented a flexible probabilistic, nonparametric Bayesian framework for learning cross-modal similarities using weak forms of supervision (relative preferences and pairwise constraints) and proposed an extension of this framework which allows active sampling to select the most informative constraints (existing methods for cross-modal similarity learning do not have such a mechanism). Our framework based on learning binary codes for multimodal data is applicable to a wide-range of applications that require computing pairwise similarities between objects belonging to different modalities, such as cross-modal object matching, cross-modal link prediction, and cross-modal hashing based similarity search and retrieval.

Acknowledgments

This work is supported in part by NSF IIS–1116886, AOR, DARPA, DOE, NGA, and ONR.

References

Bonilla, E. V.; Guo, S.; and Sanner, S. 2010. Gaussian process preference elicitation. In *NIPS*.

Bronstein, M. M.; Bronstein, A. M.; Michel, F.; and Paragios, N. 2010. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*.

Cortes, C., and Mohri, M. 2003. AUC optimization vs. error rate minimization. In *NIPS*.

Görür, D.; Jäkel, F.; and Rasmussen, C. E. 2006. A choice model with infinitely many latent features. In *ICML*.

Grauman, K., and Fergus, R. 2013. Learning binary hash codes for large-scale image search. In *Machine Learning for Computer Vision*. Springer. 49–87.

Griffiths, T. L., and Ghahramani, Z. 2011. The Indian buffet process: An introduction and review. *JMLR* 12:1185–1224.

Houlsby, N.; Hernandez-Lobato, J. M.; Huszar, F.; and Ghahramani, Z. 2012. Collaborative gaussian processes for preference learning. In *NIPS*.

Jia, Y.; Salzmann, M.; and Darrell, T. 2011. Learning cross-modality similarity for multinomial data. In *ICCV*.

Kuang, Z., and Wong, K. K. 2013. Relatively-paired space analysis. In *BMVC*.

Kumar, S., and Udupa, R. 2011. Learning hash functions for cross-view similarity search. In *IJCAI*.

Li, X.; Lin, G.; Shen, C.; Van den Hengel, A.; and Dick, A. 2013. Learning hash functions using column generation. In *ICML*.

Masci, J.; Bronstein, M. M.; Bronstein, A. M.; and Schmidhuber, J. 2013. Multimodal similarity-preserving hashing. *TPAMI*.

Mignon, A.; Jurie, F.; et al. 2012. CMML: A new metric learning approach for cross modal matching. In *ACML*.

Murray, I.; Adams, R. P.; and MacKay, D. J. C. 2010. Elliptical slice sampling. In *AISTATS*.

Ou, M.; Cui, P.; Wang, F.; Wang, J.; Zhu, W.; and Yang, S. 2013. Comparing apples to oranges: A scalable solution with heterogeneous hashing. In *KDD*.

Quadrianto, N.; Sharmanska, V.; Austria, I.; Knowles, D. A.; and Ghahramani, Z. 2013. The supervised IBP: Neighbourhood preserving infinite latent feature models. In *UAI*.

Rastegari, M.; Choi, J.; Fakhraei, S.; Daumé, H.; and Davis, L. S. 2013. Predictable dual-view hashing. In *ICML*.

Teh, Y. W.; Görür, D.; and Ghahramani, Z. 2007. Stick-breaking construction for the Indian buffet process. In *AISTATS*.

Wang, J.; Wang, J.; Yu, N.; and Li, S. 2013a. Order preserving hashing for approximate nearest neighbor search. In *ACM MM*.

Wang, J.; Liu, W.; Sun, A. X.; and Jiang, Y.-G. 2013b. Learning hash codes with listwise supervision. In *ICCV*.

Whang, J. J.; Rai, P.; and Dhillon, I. S. 2013. Stochastic blockmodel with cluster overlap, relevance selection, and similarity-based smoothing. In *ICDM*.

Yue, Y.; Finley, T.; Radlinski, F.; and Joachims, T. 2007. A support vector method for optimizing average precision. In *SIGIR*.

Zhai, D.; Chang, H.; Zhen, Y.; Liu, X.; Chen, X.; and Gao, W. 2013. Parametric local multimodal hashing for cross-view similarity search. In *IJCAI*.

Zhen, Y., and Yeung, D.-Y. 2012a. Co-regularized hashing for multimodal data. In *NIPS*.

Zhen, Y., and Yeung, D.-Y. 2012b. A probabilistic approach to multimodal hash function learning. In *KDD*.