Improved Multimodal Deep Learning with Variation of Information

Kihyuk Sohn **Wenling Shang** **Honglak Lee**
University of Michigan Ann Arbor, MI, USA
{kihyuks,shangw,honglak}@umich.edu

Abstract

Deep learning has been successfully applied to multimodal representation learning problems, with a common strategy of learning joint representations that are shared across multiple modalities on top of layers of modality-specific networks. Nonetheless, there still remains a question about how to effectively learn associations between heterogeneous data modalities; in particular, a good generative model of multimodal data should be able to reason about missing data modality given the rest of data modalities. In this paper, we propose a novel multimodal representation learning framework that explicitly aims at this goal by training the model to minimize the *variation of information* rather than maximizing likelihood. We provide a theoretical insight into why the proposed learning objective is sufficient to estimate the data-generating joint distribution of multimodal data. We apply our method to restricted Boltzmann machines and introduce learning algorithms based on contrastive divergence and multi-prediction training. Further, we extend our method to deep networks with recurrent encoding for finetuning. In experiments, we demonstrate the state-of-the-art visual recognition performance on MIR-Flickr and PASCAL VOC2007 database with and without text observations.

1 Introduction

Different types of multiple data modalities can be used to describe the same event. For example, images, which are often represented with pixels or image descriptors, can also be described with accompanying text (e.g., user tags or subtitles) or audio data (e.g., human voice or natural sound). There have been several applications of multimodal learning from multiple domains such as emotion and speech recognition with audio-visual data [16, 24, 13], robotics applications with visual and depth data [18, 20, 34, 26], and medical applications with visual and temporal data [29]. For each application, data from multiple sources are *semantically* correlated, and sometimes provide complementary information about each other. To facilitate information exchange, it is important to capture a high-level association between data modalities with a compact set of latent variables. However, learning associations between multiple *heterogeneous* data distributions is a challenging problem.

A naive approach is to concatenate the data descriptors from different input sources to construct a single high-dimensional feature vector and use it to solve a unimodal representation learning problem. However, the correlation between features in each data modality is much stronger than that between data modalities. As a result, the learning algorithms are easily tempted to learn dominant patterns in each data modality separately while giving up learning patterns that occur simultaneously in multiple data modalities, as suggested by [24]. To resolve this issue, deep learning methods, such as deep autoencoders [11] or deep Boltzmann machines (DBM) [27], have been adapted [24, 30], where the common strategy is to learn joint representations that are shared across multiple modalities at the higher layer of the deep network, after learning layers of modality-specific networks. The rationale is that the learned features may have less within-modality correlation than raw features, and this makes it easier to capture patterns across data modalities. This has shown promise, but there still remains the challenging question of how to learn associations between multiple heterogeneous data modalities so that we can effectively deal with missing data modalities at testing time.

One necessary condition for a good generative model of multimodal data is the ability to predict or reason about missing data modalities given partial observation. To this end, we propose a novel

multimodal representation learning framework that explicitly aims at this goal. The key idea is to minimize the information distance between data modalities through the shared latent representations. More concretely, we train the model to minimize the *variation of information* (VI), an information theoretic measure that computes the distance between random variables, i.e., multiple data modalities. Note that this is in contrast to previous approaches on multimodal deep learning, which are based on maximum (joint) likelihood (ML) learning [24, 30]. We explain as to how our method could be more effective in learning the joint representation of multimodal data than ML learning, and show theoretical insights why the proposed learning objective is sufficient to estimate the data-generating joint distribution of multimodal data. We apply the proposed framework to multimodal restricted Boltzmann machine (MRBM) and propose two learning algorithms, based on contrastive divergence [23] and multi-prediction training [7]. Finally, we extend to multimodal deep recurrent neural network (MDRNN) for unsupervised finetuning of whole network. In experiments, we demonstrate the state-of-the-art visual recognition performance on MIR-Flickr database and PASCAL VOC2007 database with and without text observations at testing time.

2 Multimodal Learning with Variation of Information

In this section, we propose a novel training objective based on the VI. We make a comparison to the ML objective, a typical learning objective for training generative models of multimodal data, to give an insight as to how our proposed method can be better for multimodal data. Finally, we establish a theorem showing that the proposed learning objective is sufficient to obtain a good generative model that fully recovers the joint data-generating distribution of multimodal data.

Notation. We use uppercase letters X, Y to denote random variables, lowercase letters x, y for realizations. Let $P_{\mathcal{D}}$ be the data-generating distribution and P_θ the model distribution parametrized by θ. For presentation clarity, we slightly abuse the notation for Q to denote conditional ($Q(x|y), Q(y|x)$), marginal ($Q(x), Q(y)$), as well as joint distributions ($Q(x, y)$). The type of distribution of Q should be clear from the context.

2.1 Minimum Variation of Information Learning

Motivated by the necessary condition for good generative models to reason about the missing data modality, it seems natural to learn to maximize the amount of information that one data modality has about the others. We quantify such an amount of information between data modalities using variation of information. The VI is an information theoretic measure that computes the information distance between two random variables (e.g., data modalities), and is written as follows:[1]

$$\mathrm{VI}_Q(X, Y) = -\mathbb{E}_{Q(X,Y)}\big[\log Q(X|Y) + \log Q(Y|X)\big] \tag{1}$$

where $Q(X, Y) = P_\theta(X, Y)$ is any joint distribution on random variables (X, Y) parametrized by θ. Informally, VI is small when the conditional likelihoods $Q(X|Y)$ and $Q(Y|X)$ are "peaked", meaning that X has low entropy conditioned on Y and vice versa. Following the intuition, we define new multimodal learning criteria, a *minimum variation of information* (MinVI) learning, as follows:

$$\textbf{\textit{MinVI}}\textbf{:} \quad \min_\theta \mathcal{L}^{\mathrm{VI}}(\theta), \quad \mathcal{L}^{\mathrm{VI}}(\theta) = -\mathbb{E}_{P_{\mathcal{D}}(X,Y)}\big[\log P_\theta(X|Y) + \log P_\theta(Y|X)\big] \tag{2}$$

Note the difference that we take the expectation over $P_{\mathcal{D}}$ in $\mathcal{L}^{\mathrm{VI}}(\theta)$. Furthermore, we observe that the MinVI objective can be decomposed into a sum of two negative conditional LLs. This indeed aligns well with our initial motivation of reasoning about missing data modality. In the following, we provide more insight into our MinVI objective in relation to the ML objective, which is a standard learning objective in generative models.

2.2 Relation to Maximum Likelihood Learning

The ML objective function can be written as a minimization of the negative LL (NLL) as follows:

$$\textbf{\textit{ML}}\textbf{:} \quad \min_\theta \mathcal{L}^{\mathrm{NLL}}(\theta), \quad \mathcal{L}^{\mathrm{NLL}}(\theta) = -\mathbb{E}_{P_{\mathcal{D}}(X,Y)}\big[\log P_\theta(X, Y)\big], \tag{3}$$

and we can show that the NLL objective function is reformulated as follows:

$$2\mathcal{L}^{\mathrm{NLL}}(\theta) = \underbrace{KL\left(P_{\mathcal{D}}(X)\|P_\theta(X)\right) + KL\left(P_{\mathcal{D}}(Y)\|P_\theta(Y)\right)}_{(a)} +$$

$$\underbrace{\mathbb{E}_{P_{\mathcal{D}}(X)}\big[KL\left(P_{\mathcal{D}}(Y|X)\|P_\theta(Y|X)\right)\big] + \mathbb{E}_{P_{\mathcal{D}}(Y)}\big[KL\left(P_{\mathcal{D}}(X|Y)\|P_\theta(X|Y)\right)\big]}_{(b)} + C, \tag{4}$$

[1] In practice, we use finite samples of the training data and use a regularizer (e.g., l_2 regularizer) to avoid overfitting to the finite sample distribution.

where C is a constant which is irrelevant to θ. Note that (b) is equivalent to $\mathcal{L}^{\text{VI}}(\theta)$ in Equation (2) up to a constant. We provide a full derivation of Equation (4) in Appendix A.

Ignoring the constant, the NLL objective has four KL divergence terms. Since KL divergence is non-negative and is zero only when two distributions match, the ML learning in Equation (3) can be viewed as a distribution matching problem involving (a) marginal likelihoods and (b) conditional likelihoods. Here, we argue that (a) is more difficult to optimize than (b) because there are often too many modes in the marginal distribution. Compared to the marginal distribution, the number of modes can be dramatically reduced in the conditional distribution since the conditioning variables may restrict the support of random variable effectively. Therefore, (a) may become a dominant factor to be minimized during the optimization process and as a trade-off, (b) will be easily compromised, which makes it difficult to learn a good association between data modalities. On the other hand, the MinVI objective focuses on modeling the conditional distributions (Equation (4)), which is arguably easier to optimize. Indeed, similar argument has been made for generalized denoising autoencoders (DAEs) [3] and generative stochastic networks (GSNs) [2], which focus on learning the transition operators (e.g., $P_\theta(X|\tilde{X})$, where $\tilde{X}$ is a corrupted version of data X, or $P_\theta(X|H)$, where H can be arbitrary latent variables) to bypass an intractable problem of learning density model $P_\theta(X)$.

2.3 Theoretical Results

Bengio et al. [3, 2] proved that learning transition operators of DAEs or GSNs is sufficient to learn a good generative model that estimates a data-generating distribution. Under similar assumptions, we establish a theoretical result that we can obtain a good density estimator for joint distribution of multimodal data by learning the transition operators derived from the conditional distributions of one data modality given the other. In the multimodal learning framework, we define the transition operators $T_n^{\mathcal{X}}$ and $T_n^{\mathcal{Y}}$ for Markov chains of data modalities X and Y, respectively. Specifically, $T_n^{\mathcal{X}}\left(x[t]|x[t-1]\right) = \sum_{y\in\mathcal{Y}} P_{\theta_n}\left(x[t]|y\right) P_{\theta_n}\left(y|x[t-1]\right)$, where $P_{\theta_n}\left(X|Y\right)$ and $P_{\theta_n}\left(Y|X\right)$ are model conditional distributions after learning from the training data of size n. $T_n^{\mathcal{Y}}$ is defined in a similar way. Note that we do not require that the model conditionals are derived from an analytically defined joint distribution. Now, we formalize the theorem as follows:

Theorem 2.1. *For finite state space $\mathcal{X}, \mathcal{Y}$, if, $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$, $P_{\theta_n}(\cdot|y)$ and $P_{\theta_n}(\cdot|x)$ converges in probability to $P_{\mathcal{D}}(\cdot|y)$ and $P_{\mathcal{D}}(\cdot|x)$, respectively, and $T_n^{\mathcal{X}}$ and $T_n^{\mathcal{Y}}$ are ergodic Markov chains, then, as the number of examples $n \to \infty$, the asymptotic distribution $\pi_n(X)$ and $\pi_n(Y)$ converge to data-generating marginal distributions $P_{\mathcal{D}}(X)$ and $P_{\mathcal{D}}(Y)$, respectively. Moreover, the joint probability distribution $P_{\theta_n}\left(X, Y\right)$ converges to $P_{\mathcal{D}}\left(X, Y\right)$ in probability.*

The proof is provided in Appendix B. The theorem ensures that the MinVI objective can lead to a good generative model estimating the joint data-generating distribution of multimodal data. The theorem holds under two assumptions: consistency of density estimators and ergodicity of transition operators. The ergodicity condition is satisfied for a wide variety of neural networks, such as RBM or DBM.[2] The consistency assumption is more difficult to satisfy, and the aforementioned deep energy-based models or RNN may not satisfy the condition due to the model capacity limitation or approximated posteriors (e.g., factorial distribution). However, deep architectures are arguably among the most promising models for approximating the true conditionals from multimodal data. We expect that more accurate approximation of the true conditional distributions would lead to better performance in our multimodal learning framework, and we leave it for future work.

We note that our Theorem 2.1 is related to composite likelihood methods [21] and dependency networks [9]. For composite likelihood, the consistency result is derived upon a well-defined graphical model (e.g., Markov network) and the joint distribution converges in the sense that the maximum composite likelihood estimators are consistent for the parameters associated with the graphical model. However, in Theorem 2.1, it is not necessary to design a full graphical model (e.g., of the joint distribution) with analytical forms; for example, the two conditionals can be defined by neural networks with different parameters. In this case, the joint distribution is defined implicitly, and the setting is similar to general dependency networks [9]. However, [9] uses ordered pseudo-Gibbs samplers which may be unstable (i.e., inconsistencies between the local conditionals and the true conditionals can be amplified to a large inconsistency between the model joint distribution and the true joint distribution). In our case, we prove that the implicit model joint distribution will converge to the true joint distribution under assumptions that can plausibly hold for deep architectures.

[2] For energy-based models like RBM and DBM, it is straightforward to see that every state has non-zero probability and can be reached from any other state. However, the mixing of the chain might be slow in practice.

3 Application to Multimodal Deep Learning

In this section, we describe the MinVI learning in multimodal deep learning framework. To overview our pipeline, we use the commonly used network architecture that consists of layers of modality-specific deep networks followed by a layer of neural network that jointly models the multiple modalities [24, 30]. The network is trained in two steps: In layer-wise pretraining, each layer of modality-specific deep network is trained using restricted Boltzmann machines (RBMs). For the top-layer shared network, we train MRBM with MinVI objective (Section 3.2). Then, we finetune the whole deep network by constructing multimodal deep recurrent neural network (MDRNN) (Section 3.3).

3.1 Restricted Boltzmann Machines for Multimodal Learning

The restricted Boltzmann machine (RBM) is an undirected graphical model that defines the distribution of visible units using hidden units. For multimodal input, we define the joint distribution of multimodal RBM (MRBM) [24, 30] as $P(x, y, h) = \frac{1}{Z} \exp\big(-E(x, y, h)\big)$ with the energy function:

$$E(x, y, h) = -\sum_{i=1}^{N_x}\sum_{k=1}^{K} x_i W^x_{ik} h_k - \sum_{j=1}^{N_y}\sum_{k=1}^{K} y_j W^y_{jk} h_k - \sum_{k=1}^{K} b_k h_k - \sum_{i=1}^{N_x} c^x_i x_i - \sum_{j=1}^{N_y} c^y_j y_j, \quad (5)$$

where Z is the normalizing constant, $x \in \{0,1\}^{N_x}$, $y \in \{0,1\}^{N_y}$ are the binary visible units of multimodal input (i.e., observations), and $h \in \{0,1\}^K$ are the binary hidden units (i.e., latent variables). $W^x \in \mathbb{R}^{N_x \times K}$ defines the weights between x and h, and $W^y \in \mathbb{R}^{N_y \times K}$ defines the weights between y and h. $c^x \in \mathbb{R}^{N_x}$, $c^y \in \mathbb{R}^{N_y}$, and $b \in \mathbb{R}^K$ are bias vectors corresponding to x, y, and h, respectively. Note that the MRBM is equivalent to an RBM whose visible units are constructed by concatenating the visible units of multiple input modalities, i.e., $v = [x\,; y]$.

Due to bipartite structure, units in the same layer are conditionally independent given the units of the other layer, and the conditional probabilities are written as follows:

$$P(h_k = 1 \mid x, y) = \sigma\big(\sum_i W^x_{ik} x_i + \sum_j W^y_{jk} y_j + b_k\big), \quad (6)$$

$$P(x_i = 1 \mid h) = \sigma\big(\sum_k W^x_{ik} h_k + c^x_i\big), \quad P(y_j = 1 \mid h) = \sigma\big(\sum_k W^y_{jk} h_k + c^y_j\big), \quad (7)$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$. Similar to the standard RBM, the MRBM can be trained to maximize the joint LL $(\log P(x, y))$ using stochastic gradient descent (SGD) while approximating the gradient with contrastive divergence (CD) [10] or persistent CD (PCD) [32]. In our case, however, we train the MRBM in MinVI criteria. We will discuss the inference and training algorithms in Section 3.2.

When we have access to all data modalities, we can use Equation (6) for exact posterior inference. On the other hand, when some of the input modalities are missing, the inference is intractable, and we resort to the variational method. For example, when we are given x but not y, the true posterior can be approximated with a fully factorized distribution $Q(y, h) = \prod_j \prod_k Q(y_j)Q(h_k)$ by minimizing the $KL\big(Q(y, h) \| P_\theta(y, h|x)\big)$. This leads to the following fixed-point equations:

$$\hat{h}_k = \sigma\big(\sum_i W^x_{ik} x_i + \sum_j W^y_{jk} \hat{y}_j + b_k\big), \; \hat{y}_j = \sigma\big(\sum_k W^y_{jk} \hat{h}_k + c^y_j\big), \quad (8)$$

where $\hat{h}_k = Q(h_k)$ and $\hat{y}_j = Q(y_j)$. The variational inference proceeds by alternately updating the mean-field parameters $\hat{h}$ and $\hat{y}$ that are initialized with all zeros.

3.2 Training Algorithms

CD-PercLoss. As in Equation (2), the objective function can be decomposed into two conditional LLs, and the MRBM with MinVI objective can be trained equivalently by training the two conditional RBMs (CRBMs) while sharing the weights. Since the objective functions are the sum of two conditional LLs, we compute the (approximate) gradient of each CRBM separately using CD-PercLoss [23] and accumulate them to update parameters.[3]

[3] In CD-PercLoss learning, we run separate Gibbs chains for different conditioning variables and select the negative particles with the lowest free energy among sampled particles. We refer [23] for further details.

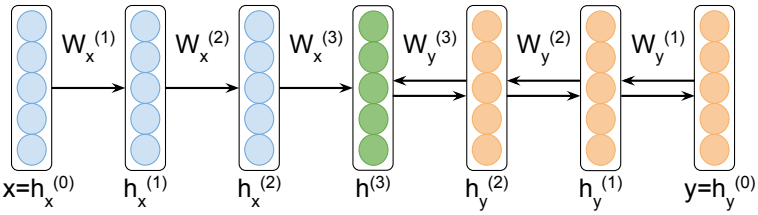


Figure 1: An instance of MDRNN with target y given x. Multiple iterations of bottom-up updates ($y \rightarrow h^{(3)}$; Eqs. (11) & (12)) and top-down updates ($h^{(3)} \rightarrow y$; Eq. (13)) are performed. The arrow indicates encoding direction.

Multi-Prediction. We found a few practical issues of CD-PercLoss training in our application. In particular, there exists a difference between the encoding process of training and testing, especially when the unimodal query (e.g., when one of the data modalities is missing) is considered for testing. As an alternative objective, we propose multi-prediction (MP) training of MRBM in MinVI criteria. The MP training was originally proposed to train deep Boltzmann machines [7] as an alternative to the stochastic approximation learning [27]. The idea is to train the model to be good at predicting any subset of input variables given the rest of them by constructing the recurrent network with encoding function derived from the variational inference problem.

The MP training can be adapted to learn MRBM with MinVI objective with some modifications. For example, the CRBM with an objective $\log P(y|x)$ can be trained by randomly selecting the subset of variables to be predicted only from the target modality y, but the conditioning modality x is assumed to be given in all cases. Specifically, given an arbitrary subset $S \subset \{1, \cdots, N_y\}$ drawn from the independent Bernoulli distribution P_S, the MP algorithm predicts $y_S = \{y_j : j \in S\}$ given x and $y_{\setminus S} = \{y_j : j \notin S\}$ through the iterative encoding function derived from fixed-point equations:

$$\hat{h}_k = \sigma\big(\sum_i W^x_{ik} x_i + \sum_{j \in S} W^y_{jk} \hat{y}_j + \sum_{j \notin S} W^y_{jk} y_j + b_k\big), \ \hat{y}_j = \sigma\big(\sum_k W^y_{jk} \hat{h}_k + c^y_j\big), j \in S, \quad (9)$$

which is a solution to the variational inference problem $\min_Q KL\left(Q(y_S, h) \| P_\theta(y_S, h | x, y_{\setminus S})\right)$ with factorized distribution $Q(y_S, h) = \prod_{j \in S} \prod_k Q(y_j) Q(h_k)$. Note that Equation (9) is similar to the Equation (8) except that only $y_j, j \in S$ are updated. Using an iterative encoding function, the network parameters are trained using SGD while computing the gradient by backpropagating the error between the prediction and the ground truth of y_S through the derived recurrent network. The MP formulation (e.g., encoding function) of the CRBM with $\log P(x|y)$ can be derived similarly, and the gradients are simply the addition of two gradients that are computed individually.

We have two additional hyper parameters, the number of mean-field updates and the sampling ratio of a subset S to be predicted from the target data modality. In our experiments, it was sufficient to use $10 \sim 20$ iterations until convergence. We used a sampling ratio of 1 (i.e., all the variables in the target data modality are to be predicted) since we are already conditioned on one data modality, which is sufficient to make a good prediction of variables in the target data modality.

3.3 Finetuning Multimodal Deep Network with Recurrent Neural Network

Motivated from the MP training of MRBM, we propose a multimodal deep recurrent neural network (MDRNN) that tries to predict the target modality given the input modality through the recurrent encoding function. The MDRNN iteratively performs a full pass of bottom-up and top-down encoding from bottom-layer visible variables to top-layer joint representation back to bottom-layer through the modality-specific deep network corresponding to the target. We show an instance of $L = 3$ layer MDRNN in Figure 1, and the encoding functions are written as follows:[4]

$$x \rightarrow h_x^{(L-1)}: \quad h_x^{(l)} = \sigma\left(W^{x,(l)\top} h_x^{(l-1)} + b^{x,(l)}\right), l = 1, \cdots, L-1 \quad (10)$$

$$y \rightarrow h_y^{(L-1)}: \quad h_y^{(l)} = \sigma\left(W^{y,(l)\top} h_y^{(l-1)} + b^{y,(l)}\right), l = 1, \cdots, L-1 \quad (11)$$

$$h_x^{(L-1)}, h_y^{(L-1)} \rightarrow h^{(L)}: \quad h^{(L)} = \sigma\left(W^{x,(L)\top} h_x^{(L-1)} + W^{y,(L)\top} h_y^{(L-1)} + b^{(L)}\right) \quad (12)$$

$$h^{(L)} \rightarrow y: \quad h_y^{(l-1)} = \sigma\left(W^{y,(l)} h_y^{(l)} + b^{y,(l-1)}\right), l = L, \cdots, 1. \quad (13)$$

[4]There could be different ways of constructing MDRNN; for instance, one can construct the RNN with DBM-style mean-field updates. In our empirical evaluation, however, running full pass of bottom-up and top-down updates performed the best, and DBM-style updates didn't give competitive results.

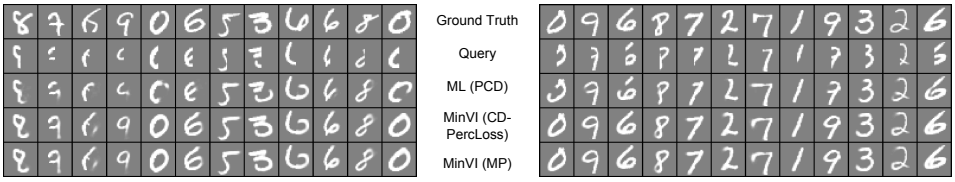


Figure 2: Visualization of samples with inferred missing modality. From top to bottom, we visualize ground truth, left or right halves of digits, generated samples with inferred missing modality using MRBM with ML objective, MinVI objective using CD-PercLoss and MP training methods.

Input modalities at test time	Left+Right	Left	Right
ML (PCD)	**1.57**%	14.98%	18.88%
MinVI (CD-PercLoss)	1.71%	9.42%	11.02%
MinVI (MP)	1.73%	**6.58**%	**7.27**%

Table 1: Test set errors on handwritten digit recognition dataset using MRBMs with different training objectives and learning methods. The joint representation was fed into linear SVM for classification.

Here, we define $h_x^{(0)} = x$ and $h_y^{(0)} = y$, and the visible variables of the target modality are initialized with zeros. In other words, in the initial bottom-up update, we compute $h^{(L)}$ only from x while setting $y = 0$ using Equations (10), (11), & (12). Then, we run multiple iterations of top-down (Equation (13)) and bottom-up updates (Equations (11) & (12)). Finally, we compute the gradient by backpropagating the reconstruction error of target modality through the network.

4 Experiments

4.1 Toy Example on MNIST

In our first experiment, we evaluate the proposed learning algorithm on the MNIST handwritten digit recognition dataset [19]. We consider left and right halves of the digit images as two input modalities and report the recognition performance with different combinations of input modalities at the test time, such as full (left + right) or missing (left or right) data modalities. We compare the performance of the MRBM trained with 1) ML objective using PCD [32], or MinVI objectives with 2) CD-PercLoss or 3) MP training. The recognition errors are provided in Table 1. Compared to ML training, the recognition errors for unimodal queries are reduced by more than a half with MP training of MinVI objective. For multimodal queries, the model trained with ML objective performed the best, although the performance gain was incremental. CD-PercLoss training of MinVI objective also showed significant improvement over ML training, but the errors were not as low as those obtained with MP training. We hypothesize that, although it is an approximation of MinVI objective, the exact gradient for MP algorithm makes learning more efficient than CD-PercLoss. For the rest of the paper, we focus on MP training method.

In Figure 2, we visualize the generated samples conditioned on one input modality (e.g., left or right halves of digits). There are many samples generated by the models with MinVI objective that look clearly better than those generated by the model with ML objective.

4.2 MIR-Flickr Database

In this section, we evaluate our methods on MIR-Flickr database [14], which is composed of 1 million examples of images and their user tags collected from the social photo-sharing website Flickr. Among those, 25000 examples were annotated with 24 potential topics and 14 regular topics, which leads to 38 classes in total with distributed class membership. The topics included object categories such as dog, flower, and people, or scenic concepts such as sky, sea, and night.

We used the same visual and text features as in [30].[5] Specifically, the image feature was a 3857 dimensional vector composed of Pyramid Histogram of Words (PHOW) features [4], GIST [25], and MPEG-7 descriptors [22]. We preprocessed the image features to have zero mean and unit variance for each dimension across all examples. The text feature was a word count vector of 2000 most frequent tags. The number of tags varied from 0 to 72, with 5.15 tags per example in average.

Following the experimental protocol [15, 30], we randomly split the labeled examples into 15000 for training and 10000 for testing, and used 5000 from training set for validation. We iterated the procedure for 5 times and report the mean average precision (mAP) averaged over 38 classes.

[5] http://www.cs.toronto.edu/~nitish/multimodal/index.html

Model Architecture. We used the network composed of $[3857, 1024, 1024]$ variables for visual pathway, $[2000, 1024, 1024]$ variables for text pathway, and 2048 variables for top-layer MRBM, as used in [30]. As described in Section 3, we pretrained the modality-specific deep networks in a greedy layerwise way, and finetuned the whole network by initializing MDRNN with the pretrained network. Specifically, we used gaussian RBM for the bottom layer of visual pathway and binary RBM for text pathway.[6] The intermediate layers were trained with binary RBMs, and the top-layer MRBM was trained using MP training algorithm. For the layer-wise pretraining of RBMs, we used PCD [32] to approximate the gradient. Since our algorithm requires both data modalities during training, we excluded examples with too sparse or no tags from unlabeled dataset and used about 750K examples with at least 2 tags. After unsupervised training, we extracted joint feature representations of the labeled training data and use them to train multiclass logistic regression classifiers.

Recognition Tasks. For recognition tasks, we trained multiclass logistic regression classifiers using joint representations as input features. Depending on the availability of data modalities at testing time, we evaluated the performance using multimodal queries (i.e., both visual and text data are available) and unimodal queries (i.e., visual data is available while the text data is missing). In Table 2, we report the test set mAPs of our proposed model and compared to other methods. The proposed MDRNN outperformed the previous state-of-the-art in multimodal queries by 4.5% in mAP. The performance improvement becomes more significant for unimodal queries, achieving 7.6% improvement in mAP over the best published result. As we used the same input features in [30], the results suggest that our proposed algorithm learns better representations shared across multiple modalities.

Model	Multimodal query
Autoencoder	0.610
Multimodal DBM [30]	0.609
Multimodal DBM† [31]	0.641
MK-SVM [8]	0.623
TagProp [33]	0.640
MDRNN	$\mathbf{0.686} \pm 0.003$
Model	**Unimodal query**
Autoencoder	0.495
Multimodal DBM [30]	0.531
MK-SVM [8]	0.530
MDRNN	$\mathbf{0.607} \pm 0.005$

Table 2: Test set mAPs on MIR-Flickr database. We implemented autoencoder following the description in [24]. Multimodal DBM† is supervised finetuned model. See [31] for details.

For a closer look into our model, we performed an additional control experiment to explore the benefit of recurrent encoding of MDRNN. Specifically, we compared the performance of the models with different number of mean-field iterations.[7] We report the validation set mAPs of models with different number of iterations ($0 \sim 10$) in Table 3. For multimodal query, the MDRNN with 10 iterations improves the recognition performance by only 0.8% compared to the model with 0 iterations. However, the improvement becomes significant for unimodal query, achieving 5.0% performance gain. In addition, the largest improvement was made when we have at least one iteration (from 0 to 1 iteration, 3.4% gain; from 1 to 10 iteration, 1.6% gain). This suggests that a crucial factor of improvement comes from the inference with reconstructed missing data modality (e.g., text features), and the quality of inferred missing modality improves as we increase the number of iterations.

# iterations	0	1	2	3	5	10
Multimodal query	0.677	0.678	0.679	0.680	0.682	**0.685**
Unimodal query	0.557	0.591	0.599	0.602	0.605	**0.607**

Table 3: Validation set mAPs on MIR-Flickr database with different number of mean-field iterations.

Retrieval Tasks. We performed retrieval tasks using multimodal and unimodal input queries. Following [30], we selected 5000 image-text pairs from the test set to form a database and use 1000 disjoint set of examples from the test set as queries. For each query example, we computed the relevance score to the data points as a cosine similarity of joint representations. The binary relevance labels between query and the data points are determined 1 if any of the 38 class labels are overlapped. Our proposed model achieves **0.633** mAP with multimodal query and **0.638** mAP with unimodal query. This significantly outperforms the performance of multimodal DBM [30], which reported 0.622 mAP with multimodal query and 0.614 mAP with unimodal query. We show retrieved examples with multimodal queries in Figure 3.

[6] We assumed text features as binary, which is different from [30] where they modeled using replicated-softmax RBM [28]. The rationale is that the tags are not likely to be assigned more than once for single image.

[7] In [24], Ngiam et al. proposed the "video-only" deep autoencoder whose objective is to predict audio data and reconstruct video data when only video data is given as an input during the training. Our baseline model (MDRNN with 0 iterations) is similar, but different since we don't have a reconstruction training objective.

Figure 3: Retrieval results with multimodal queries. The leftmost image-text pairs are multimodal query samples and those in the right side of the bar are retrieved samples with the highest similarities to the query sample from the database. We include more results in Appendix C.

4.3 PASCAL VOC 2007

We evaluate the proposed algorithm on PASCAL VOC 2007 database. The original dataset does not contain user tags, but Guillaumin et al. [8] have collected user tags from Flickr website.[8]

Motivated by the success of convolutional neural networks (CNNs) on large-scale visual object recognition [17], we used the DeCAF$_7$ features [6] as an input features for visual pathway, where DeCAF$_7$ is 4096 dimensional feature extracted from the CNN trained on ImageNet [5]. For text features, we used the vocabulary of size 804 suggested by [8]. For unsupervised feature learning of MDRNN, we used unlabeled data of MIR-Flickr database while converting the text features using the new vocabulary from PASCAL database. The network architecture used in this experiment was as follows: $[4096, 1536, 1536]$ variables for the visual pathway, $[804, 512, 1536]$ variables for the text pathway, and 2048 variables for top-layer joint network.

Following the standard practice, we reported the mAP over 20 object classes. The performance improvement of our proposed method was significant, achieving $\mathbf{81.5}\%$ mAP with multimodal queries and $\mathbf{76.2}\%$ mAP with unimodal queries, whereas the performance of the baseline model was 74.5% mAP with multimodal queries (DeCAF$_7$ + Text) and 74.3% mAP with unimodal queries (DeCAF$_7$).

5 Conclusion

Motivated by the property of good generative models of multimodal data, we proposed a novel multimodal deep learning framework based on variation of information. The minimum variation of information objective enables to learn good shared representations of multiple heterogeneous data modalities with a better prediction of missing input modality. We demonstrated the effectiveness of our proposed method on multimodal RBM and its deep extensions and showed state-of-the-art recognition performance on MIR-Flickr database and competitive performance on PASCAL VOC 2007 database with multimodal (visual + text) and unimodal (visual only) queries.

Acknowledgments This work was supported in part by ONR N00014-13-1-0762, Toyota Technical Center, and the Google Faculty Research Award. We thank Yoshua Bengio, Pedro Domingos, Francis Bach, Nando de Freitas, Max Welling, Scott Reed, and Yuting Zhang for helpful comments.

References

[1] G. Bachman and L. Narici. *Functional Analysis*. Dover Publications, 2012.

[2] Y. Bengio, E. Thibodeau-Laufer, G. Alain, and J. Yosinski. Deep generative stochastic networks trainable by backprop. In *ICML*, 2014.

[3] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *NIPS*, 2013.

[4] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[8] http://lear.inrialpes.fr/people/guillaumin/data.php

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[7] I. Goodfellow, M. Mirza, A. Courville, and Y. Bengio. Multi-prediction deep Boltzmann machines. In *NIPS*, 2013.

[8] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010.

[9] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *The Journal of Machine Learning Research*, 1:49–75, 2001.

[10] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[11] G. E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[12] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge university press, 2012.

[13] J. Huang and B. Kingsbury. Audio-visual deep learning for noise robust speech recognition. In *ICASSP*, 2013.

[14] M. J. Huiskes and M. S. Lew. The MIR Flickr retrieval evaluation. In *ICMIR*, 2008.

[15] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: The MIR Flickr retrieval evaluation initiative. In *ICMIR*, 2010.

[16] Y. Kim, H. Lee, and E. M. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *ICASSP*, 2013.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[18] K. Lai, L. Bo, X. Ren, and D. Fox. RGB-D object recognition: Features, algorithms, and a large scale benchmark. In *Consumer Depth Cameras for Computer Vision*, pages 167–192. Springer, 2013.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[20] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. In *RSS*, 2013.

[21] B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–39, 1988.

[22] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[23] V. Mnih, H. Larochelle, and G. E. Hinton. Conditional restricted Boltzmann machines for structured output prediction. In *UAI*, 2011.

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[25] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[26] D. Rao, M. D. Deuge, N. Nourani-Vatani, B. Douillard, S. B. Williams, and O. Pizarro. Multimodal learning for autonomous underwater vehicles from visual and bathymetric data. In *ICRA*, 2014.

[27] R. Salakhutdinov and G. E. Hinton. Deep Boltzmann machines. In *AISTATS*, 2009.

[28] R. Salakhutdinov and G. E. Hinton. Replicated softmax: an undirected topic model. In *NIPS*, 2009.

[29] H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1930–1943, 2013.

[30] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep Boltzmann machines. In *NIPS*, 2012.

[31] N. Srivastava and R. Salakhutdinov. Discriminative transfer learning with tree-based priors. In *NIPS*, 2013.

[32] T. Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML*, 2008.

[33] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid. Image annotation with tagprop on the MIR Flickr set. In *ICMIR*, 2010.

[34] A. Wang, J. Lu, G. Wang, J. Cai, and T.-J. Cham. Multi-modal unsupervised feature learning for RGB-D scene labeling. In *ECCV*. Springer, 2014.

Appendix

A Derivation of Equation (4)

The NLL objective function can be written as

$$
\begin{aligned}
2\mathcal{L}^{\text{NLL}}(\theta) &= -2\mathbb{E}_{P_{\mathcal{D}}}\big[\log P_\theta(X,Y)\big] \\
&= -\mathbb{E}_{P_{\mathcal{D}}}\big[\log P_\theta(X|Y) + \log P_\theta(Y)\big] - \mathbb{E}_{P_{\mathcal{D}}}\big[\log P_\theta(Y|X) + \log P_\theta(X)\big] \\
&= -\mathbb{E}_{P_{\mathcal{D}}}\big[\log P_\theta(X|Y) + \log P_\theta(Y|X)\big] - \mathbb{E}_{P_{\mathcal{D}}}\big[\log P_\theta(X) + \log P_\theta(Y)\big] \\
&= \mathcal{L}^{\text{VI}}(\theta) - \mathbb{E}_{P_{\mathcal{D}}}\big[\log P_\theta(X)\big] - \mathbb{E}_{P_{\mathcal{D}}}\big[\log P_\theta(Y)\big] \qquad (14) \\
&= \mathcal{L}^{\text{VI}}(\theta) + \underbrace{\mathbb{E}_{P_{\mathcal{D}}}\left[\log \frac{P_{\mathcal{D}}(X)}{P_\theta(X)}\right]}_{KL(P_{\mathcal{D}}(X)\|P_\theta(X))} + \underbrace{\mathbb{E}_{P_{\mathcal{D}}}\left[\log \frac{P_{\mathcal{D}}(Y)}{P_\theta(Y)}\right]}_{KL(P_{\mathcal{D}}(Y)\|P_\theta(Y))} \qquad (15) \\
&\quad \underbrace{-\mathbb{E}_{P_{\mathcal{D}}}\big[\log P_{\mathcal{D}}(X)\big] - \mathbb{E}_{P_{\mathcal{D}}}\big[\log P_{\mathcal{D}}(Y)\big]}_{C_1} \\
&= \mathcal{L}^{\text{VI}}(\theta) + KL\left(P_{\mathcal{D}}(X)\|P_\theta(X)\right) + KL\left(P_{\mathcal{D}}(Y)\|P_\theta(Y)\right) + C_1 \qquad (16)
\end{aligned}
$$

where Equation (14) holds by the definition of $\mathcal{L}^{\text{VI}}(\theta)$. Note that C_1 is independent of θ. Similarly, we can rewrite the MinVI objective as

$$
\begin{aligned}
\mathcal{L}^{\text{VI}}(\theta) &= -\mathbb{E}_{P_{\mathcal{D}}}\big[\log P_\theta(X|Y) + \log P_\theta(Y|X)\big] \qquad (17) \\
&= \mathbb{E}_{P_{\mathcal{D}}}\left[\log \frac{P_{\mathcal{D}}(X|Y)}{P_\theta(X|Y)}\right] + \mathbb{E}_{P_{\mathcal{D}}}\left[\log \frac{P_{\mathcal{D}}(Y|X)}{P_\theta(Y|X)}\right] \qquad (18) \\
&\quad \underbrace{-\mathbb{E}_{P_{\mathcal{D}}}\big[\log P_{\mathcal{D}}(X|Y)\big] - \mathbb{E}_{P_{\mathcal{D}}}\big[\log P_{\mathcal{D}}(Y|X)\big]}_{C_2}
\end{aligned}
$$

where in Equation (18), we have

$$
\begin{aligned}
\mathbb{E}_{P_{\mathcal{D}}}\left[\log \frac{P_{\mathcal{D}}(X|Y)}{P_\theta(X|Y)}\right] &= \sum_y P_{\mathcal{D}}(y)\mathbb{E}_{P_{\mathcal{D}}(X|y)}\left[\log \frac{P_{\mathcal{D}}(X|y)}{P_\theta(X|y)}\right] \qquad (19) \\
&= \mathbb{E}_{P_{\mathcal{D}}(Y)}\big[KL\left(P_{\mathcal{D}}(X|Y)\|P_\theta(X|Y)\right)\big] \qquad (20)
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
\mathcal{L}^{\text{VI}}(\theta) = \mathbb{E}_{P_{\mathcal{D}}(X)}\big[KL\left(P_{\mathcal{D}}(Y|X)\|P_\theta(Y|X)\right)\big] + \\
\mathbb{E}_{P_{\mathcal{D}}(Y)}\big[KL\left(P_{\mathcal{D}}(X|Y)\|P_\theta(X|Y)\right)\big] + C_2. \qquad (21)
\end{aligned}
$$

C_2 is independent of θ and by setting $C = C_1 + C_2$, we derive the Equation (4).

B Proof of Theorem 2.1

Proposition B.1 ([3, 2]). *Assume that $\mathcal{X}$ is a finite state space. Let T_n and T be irreducible transition matrices that have stationary distributions $\pi_n(X)$ and $\pi(X)$, respectively, where $\pi(X) = P_{\mathcal{D}}(X)$ is a data-generating distribution of X. If T_n converges to T entrywise, then $\pi_n(X)$ converges to $P_{\mathcal{D}}(X)$ entrywise.*

Proof. Let $|\mathcal{X}|$ be the number of states of variable X. For simplicity, we denote $\pi = \pi(X)$ and $\pi_n = \pi_n(X)$. Since the transition matrix T is irreducible, the stationary distribution π is unique. In other words, π is characterized by the following equations:

$$
\begin{aligned}
\sum_{k=1}^{|\mathcal{X}|} T_{j,k}\pi_k &= \pi_j, \forall j \in \{1, \cdots, |\mathcal{X}|\} \qquad (22) \\
\sum_{k=1}^{|\mathcal{X}|} \pi_k &= 1, \qquad (23) \\
\sum_{j=1}^{|\mathcal{X}|} T_{j,k} &= 1, \forall j \in \{1, \cdots, |\mathcal{X}|\}. \qquad (24)
\end{aligned}
$$

Here, (24) holds since T is a transition matrix. It is easy to see that one of the equations from (22) is redundant; for example, $\sum_{k=1}^{|\mathcal{X}|} T_{|\mathcal{X}|,k}\pi_k = \pi_{|\mathcal{X}|}$ can be recovered from other equations of (22), (23), and (24). Therefore, we can combine the above system of linear equations in an equivalent form as follows:

$$\underbrace{\begin{bmatrix} T_{1,1}-1 & T_{1,2} & \cdots & T_{1,|\mathcal{X}|} \\ T_{2,1} & T_{2,2}-1 & \cdots & T_{2,|\mathcal{X}|} \\ \vdots & \cdots & \cdots & \vdots \\ T_{|\mathcal{X}|-1,1} & \cdots & \cdots & T_{|\mathcal{X}|-1,|\mathcal{X}|} \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{=\widetilde{T}} \pi = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1, \end{bmatrix} \tag{25}$$

where $\widetilde{T}$ is defined accordingly. Since π exists and is unique, the null space of $\widetilde{T}$ must be empty and $\widetilde{T}$ is invertible. Now we have

$$\pi = \widetilde{T}^{-1} \begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}^\top \tag{26}$$

and similarly,

$$\pi_n = \widetilde{T}_n^{-1} \begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}^\top . \tag{27}$$

Since T_n converges to T entrywise, $\widetilde{T}_n$ converges to $\widetilde{T}$ entrywise, and $\widetilde{T}_n^{-1}$ also converges to $\widetilde{T}^{-1}$ entrywise. Therefore, we conclude π_n converges to $\pi = P_{\mathcal{D}}(X)$ entrywise [12]. Since on a finite-dimensional space, all norms are equivalent [1], the above convergence, in fact, holds for any norm. □

Now, we provide a proof of Theorem 2.1.

Proof of Theorem 2.1. To prove the convergence of marginal distributions, it is sufficient to show the convergence of transition operators. Since $|\mathcal{X}|$ and $|\mathcal{Y}|$ are finite, for any $\epsilon > 0$, $\delta > 0$ there exists N such that $\forall n \geq N$, with probability at least $1-\delta$, $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$,

$$\left|P_{\theta_n}(y|x) - P_{\mathcal{D}}(y|x)\right| < \epsilon,\ \left|P_{\theta_n}(x|y) - P_{\mathcal{D}}(x|y)\right| < \epsilon$$

The transition operators are defined as follows:

$$T_n^{\mathcal{Y}}(y[t]|y[t-1]) = \sum_{x\in\mathcal{X}} P_{\theta_n}(y[t]|x)\, P_{\theta_n}(x|y[t-1]),$$
$$T^{\mathcal{Y}}(y[t]|y[t-1]) = \sum_{x\in\mathcal{X}} P_{\mathcal{D}}(y[t]|x)\, P_{\mathcal{D}}(x|y[t-1]).$$

For data-generating distribution, $P_{\mathcal{D}}(x|y)$ and $P_{\mathcal{D}}(y|x)$ are derived from $P_{\mathcal{D}}(x,y)$. Then, for $\forall n \geq N$, we have, for any $y_t, y_{t-1} \in \mathcal{Y}$, with probability at least $1-\delta$,

$$\begin{aligned} &\left| T_n^{\mathcal{Y}}(y_t|y_{t-1}) - T^{\mathcal{Y}}(y_t|y_{t-1}) \right| \\ &\quad\leq \left| \sum_{x\in\mathcal{X}} P_{\theta_n}(y_t|x)\, P_{\theta_n}(x|y_{t-1}) - P_{\mathcal{D}}(y_t|x)\, P_{\mathcal{D}}(x|y_{t-1}) \right| \\ &\quad\leq |\mathcal{X}| \max_{x\in\mathcal{X}} \left| P_{\theta_n}(y_t|x)\, P_{\theta_n}(x|y_{t-1}) - P_{\mathcal{D}}(y_t|x)\, P_{\mathcal{D}}(x|y_{t-1}) \right| \\ &\quad\leq |\mathcal{X}|\,(2\epsilon) \end{aligned} \tag{28}$$

As we assume finite sets $\mathcal{X}$ and $\mathcal{Y}$, this proves the convergence (in probability) of transition operator $T_n^{\mathcal{Y}}$ to $T^{\mathcal{Y}}$. The same argument holds for the convergence of transition operator $T_n^{\mathcal{X}}$ to $T^{\mathcal{X}}$. Together with Proposition B.1, we have proved the convergence of asymptotic marginal distribution $\pi_n(X)$ and $\pi_n(Y)$ to data-generating marginal distributions $P_{\mathcal{D}}(X)$ and $P_{\mathcal{D}}(Y)$, respectively.

Now, let's look at the joint probability distributions $P_{\theta_n}(x,y) = P_{\theta_n}(x|y)P_{\theta_n}(y)$ and similarly, $P_{\mathcal{D}}(x,y) = P_{\mathcal{D}}(x|y)P_{\mathcal{D}}(y)$. From a similar argument as above, with probability at least $1-\delta$, there exists N' such that the following inequalities hold $\forall n \geq N', \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$:

$$\left|P_{\theta_n}(y) - P_{\mathcal{D}}(y)\right| < \epsilon,\ \left|P_{\theta_n}(x|y) - P_{\mathcal{D}}(x|y)\right| < \epsilon \tag{29}$$

Therefore, using the similar argument in Equation (28), we have

$$\left|P_{\theta_n}(x,y) - P_{\mathcal{D}}(x,y)\right| < 2\epsilon \tag{30}$$

and this completes the proof. □

C Retrieval Task

We provide more results of retrieval task with multimodal queries on MIR-Flickr database.

sky, night, clouds, space

sky, night, stars

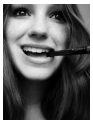
sky, night, mountains, stars, fab

2007

nikon, nature, sky, night, landscape, impressedbeauty, d300, dark, longexposure, colorado, stars

2007, beauty, hair, friend

bw, portrait, blackandwhite, girl, nikon80

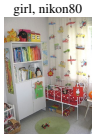


blackandwhite, selfportrait, happy, mac, makeup

bw, portrait, photo

2007, may

studio, craft, room

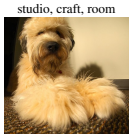


home, toys, interior, bed, books, decor

chair

design, studio

studio, craft



puppy

cute, puppy

puppy

dog

explore

skyline, indiana, 1855mm

night, city, river, dark, buildings, skyline

night, reflection, longexposure, buildings, massachusetts, boston

city, lights, buildings, fireworks, skyscrapers

nikon, night, d80, asia, skyline, hongkong, harbour

portrait, me

portrait, man, colours

selfportrait, me, 365days, 365, self

selfportrait, 365days, 365

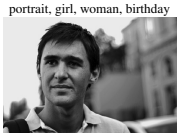
portrait, girl, woman, birthday

portrait, explore, blackwhite, portraits

portrait

london, uk

bw, halloween

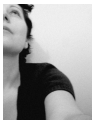


blackandwhite, milan

white, me

bw, selfportrait, me, layers

bw, blackandwhite, selfportrait, me, 365days, photoshop, self, face, head, myself, me



de

bw, self

blue, night, city, explore, nyc, newyork, lights, newyorkcity, manhattan, ny, skyline, cityscape, twilight, skyscrapers

nyc, newyorkcity

city

night, lights, new, york, exposure, noche, long, pier

sunset, chicago, tower, skyline, dusk, skyscraper

sunset, explore, sun

sunset, trees, platinumphoto, silhouette

atardecer, sunset, abigfave, searchthebest, nubes, sol

sunset

canon, naturesfinest, 30d

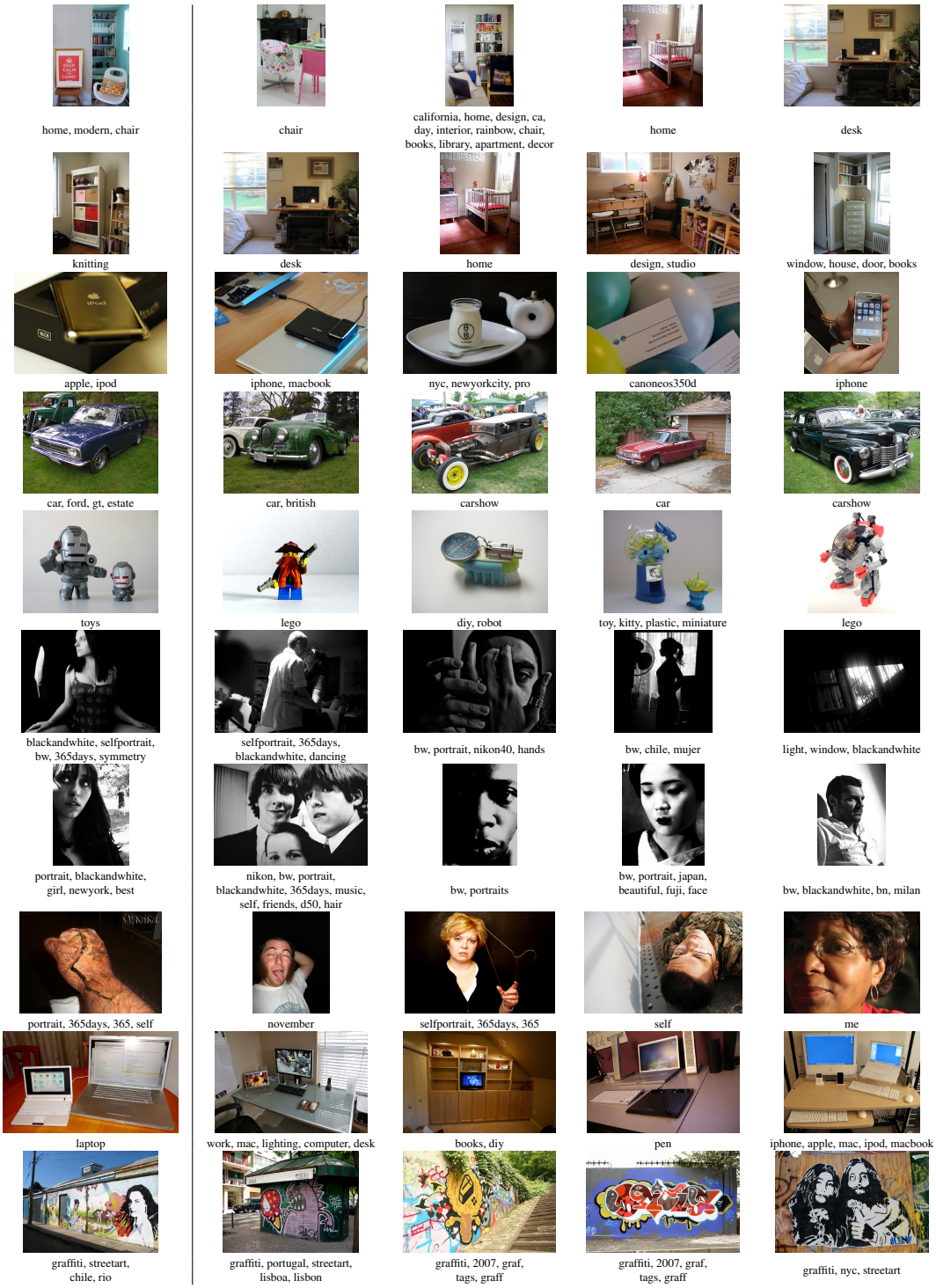


Figure 4: Retrieval results with multimodal queries on MIR-Flickr database. The leftmost image-text pairs are multimodal queries and those in the right side of the bar are retrieved samples with the highest similarities to the query.