10-2013

Online Multimodal Deep Similiarity Learning with Application to Image Retrieval

Pengcheng WU
Nanyang Technological University

Steven C. H. HOI
Singapore Management University, CHHOI@smu.edu.sg

Hao XIA
Nanyang Technological University

Peilin ZHAO
Nanyang Technological University

Dayong WANG
Nanyang Technological University

See next page for additional authors

DOI: https://doi.org/10.1145/2502081.2502112

Citation

WU, Pengcheng; HOI, Steven C. H.; XIA, Hao; ZHAO, Peilin; WANG, Dayong; and MIAO, Chunyan. Online Multimodal Deep Similiarity Learning with Application to Image Retrieval. (2013). *MM '13: Proceedings of the 21st ACM International Conference on Multimedia: October 21-25, Barcelona, Spain*. 153-162. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/2333

Author

Pengcheng WU, Steven C. H. HOI, Hao XIA, Peilin ZHAO, Dayong WANG, and Chunyan MIAO

Online Multimodal Deep Similarity Learning with Application to Image Retrieval

Pengcheng Wu, Steven C.H. Hoi, Hao Xia, Peilin Zhao, Dayong Wang, Chunyan Miao
School of Computer Engineering, Nanyang Technological University, Singapore 639798
{wupe0003,chhoi,xiah0002,zhao0106,s090023,ascymiao}@ntu.edu.sg

ABSTRACT

Recent years have witnessed extensive studies on distance metric learning (DML) for improving similarity search in multimedia information retrieval tasks. Despite their successes, most existing DML methods suffer from two critical limitations: (i) they typically attempt to learn a linear distance function on the input feature space, in which the assumption of linearity limits their capacity of measuring the similarity on complex patterns in real-world applications; (ii) they are often designed for learning distance metrics on uni-modal data, which may not effectively handle the similarity measures for multimedia objects with multimodal representations. To address these limitations, in this paper, we propose a novel framework of online multimodal deep similarity learning (OMDSL), which aims to optimally integrate multiple deep neural networks pretrained with stacked denoising autoencoder. In particular, the proposed framework explores a unified two-stage online learning scheme that consists of (i) learning a flexible nonlinear transformation function for each individual modality, and (ii) learning to find the optimal combination of multiple diverse modalities simultaneously in a coherent process. We conduct an extensive set of experiments to evaluate the performance of the proposed algorithms for multimodal image retrieval tasks, in which the encouraging results validate the effectiveness of the proposed technique.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning; I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement

General Terms

Algorithms, Experimentation

Keywords

deep learning; similarity learning; distance metric learning; online learning; image retrieval

1. INTRODUCTION

Similarity search is a fundamental issue of multimedia information retrieval research, which has been actively studied for years across several communities, including multimedia, signal processing, computer vision, and machine learning [28, 22]. The crux of similarity search lies in two key components: (i) effective feature representation and (ii) proper similarity functions over the feature space. On one hand, researchers in multimedia signal processing and computer vision have spent a few decades in searching and designing effective feature representations for content-based image retrieval (CBIR) [41, 36], ranging from early global features [31] (e.g., color, shape, and texture) to recent local features (e.g., SIFT and SURF features [30, 3] and their bag of visual words representations [49, 48]). On the other hand, various similarity/distance functions have been proposed for multimedia retrieval, such as typical Euclidean distance or cosine similarity. The key limitation of these early studies is that they often assume a fixed rigid similarity/distance function (e.g., Euclidean) for all applications, which may not be always optimal. Recent years have witnessed a surge of research efforts in Distance Metric Learning (DML) [50, 16, 15], which applies machine learning techniques to optimize distance metrics from training data or side information (e.g., logs of user relevance feedback in CBIR [16]) for multimedia applications.

Although a variety of DML algorithms have been proposed [50], most existing DML methods suffer from two critical limitations: (i) They usually learn a linear distance metric in the forms of Mahalanobis distances, which can be viewed as learning a linear projection to map feature vectors into another feature space; their linearity assumption however limits their capacity of measuring similarity for complex patterns; (ii) They are often designed for learning metrics on uni-modal data, i.e., either a single type of features or a combined feature space by a simple concatenation of multiple types of features, which could be sub-optimal for measuring similarity of multimedia objects with multimodal representations (e.g., it may fail to fully exploit the potential of all modalities whenever some modality significantly dominates the rest).

To tackle the above challenges, we propose a novel framework of Online Multimodal Deep Similarity Learning (OMDSL), which applies deep learning techniques [13] to learn a flexible nonlinear similarity function from images of multimodal feature representation via an efficient and scalable online learning scheme. Unlike conventional methods, OMDSL applies machine learning to address three key concerns in multimedia: (i) multimodal learning — a unique and key challenge for learning with multimodal contents in multimedia applications; (ii) deep learning — a powerful technique for learning a nonlinear similarity function, going beyond conventional linear/shallow machine learning approaches in multimedia retrieval studies; (iii) online learning — a family of efficient and scalable machine learning algorithms that are promising for mining large-scale multimedia data. Thus, OMDSL is partic-

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502112.

Area Chair: Matthew Cooper

ularly suitable for multimedia applications where data are multimodal, complex, large in size, and growing rapidly.

To the best of our knowledge, this is the first work that attempts to explore deep neural networks together with online learning for learning a nonlinear similarity function for multimodal image retrieval. As a summary, the main contributions of this paper include:

- a novel framework of Online Multimodal Deep Similarity Learning (OMDSL), which learns both the optimal metrics for each modality and the optimal combination of multiple modalities from training data streams;
- an online learning algorithm to solve the multimodal deep similarity learning task for multimodal image retrieval;
- theoretical bound analysis on the proposed algorithm;
- extensive empirical performance evaluation of the proposed technique for CBIR.

The rest of this paper is organized as follows. Section 2 reviews related work, Section 3 gives problem setting, Section 4 presents the proposed framework and algorithm, Section 5 discusses experimental results, and section 6 concludes this work.

2. RELATED WORK

Our research work lies in the interplay between multimedia information retrieval and machine learning. We refer readers to some classical surveys of multimedia retrieval in [41, 28]. In the following, we focus on reviewing several categories of key related works.

2.1 Content-based Image Retrieval

CBIR has been extensively studied for several decades [28, 41], where conventional approaches usually choose rigid distance functions on some low-level features for similarity search, such as Euclidean distance or cosine similarity. The fixed rigid similarity or distance functions often fail to tackle the CBIR due to the complexity of visual image representation and the semantic gap challenge between the low-level visual features and high-level human perception. Recent years have witnessed a surge of active research efforts in designing various distance/similarity measures on some low-level features using machine learning techniques, among which some works focus on learning to hashing or compact codes [39, 34, 20, 7, 23], and some others can be categorized into distance metric learning (DML). Our work is also related to multimodal/multiview studies, which have been widely studied on image classification and object recognition [51, 1, 12]. However, it is usually hard to apply these techniques directly to CBIR because (i) image classes typically will not be given explicitly in CBIR tasks, (ii) even if classes can be given, the number can be very large, (iii) image datasets for CBIR tend to be much larger than those for classification tasks. We thus exclude the direct comparisons to such existing works.

2.2 Distance Metric Learning

Distance metric learning has been extensively studied in both machine learning and multimedia communities [50]. Existing DML studies can be grouped into different categories according to varied learning settings and methodologies. For example, in terms of training data formats, many DML studies in machine learning typically learn metrics directly from explicit class labels [47], while DML studies in multimedia mainly learn metrics from side information in the forms of either pairwise constraints [16, 15] or triplet constraints [8]. In terms of learning methodology, most existing DML studies generally follow batch machine learning methods, except that some recent DML studies begin to explore online learning techniques [19, 21]. All these studies generally address uni-modal DML, which differs from our multi-modal learning scheme. From another learning perspective, DML aims to learn an optimal linear projection that maps the input features into another feature space; by contrast, we aim to learn a nonlinear projection via deep neural networks. We also note that our work is different from some existing distance learning studies that learn nonlinear distance functions using kernel methods [16, 32], which are often poorly scalable and simply cannot scale up even for medium-scale applications. We thus cannot include the direct empirical comparisons to these existing works. Finally, we note that our work is very different from some existing multiview DML study [51] which is concerned with classification tasks by learning metrics on training data with explicit class labels, making it difficult to be directly compared with our method in experiments.

2.3 Multimodal Deep Learning

Our work falls into the category of deep learning (a.k.a. deep neural networks) methodology, which has been actively studied in machine learning recently. The goal of deep learning is to learn multiple levels of representations through a hierarchy of neural network architectures, where higher-level representation is expected to help define higher-level concepts. Many recent studies [14, 13] have found promising empirical results by applying deep learning to tackle a variety of tasks across different application domains, ranging from speech recognition [10], face recognition, to generic object recognition [27] and image classification [53, 26], etc. Some pioneering and representative studies of deep learning include Deep Belief Nets (DBN) [13], Auto Encoder (AE) [5], Stacked Denoising Autoencoder (SDA) [45], Deep Boltzmann Machine (DBM) [40] and Deep Energy Model (DEM) [33]. A more comprehensive introduction to deep learning can be found in [4] and references therein.

Recently, some studies [33, 43] have applied DBN or DBM to learn cross-modality/multimodal features, e.g., audio versus visual or textual versus visual. Our work differs considerably from these studies, at least in two aspects: (i) their approaches aim to learn a shared representation between modalities so as to infer some missing modality from others, e.g., to infer text from images; by contrast, our method aims to learn both the optimal nonlinear similarity function for each modality and their optimal combination; (ii) our problem setting is very different from the former studies [33, 43] which usually deal with classification tasks and fine-tune the deep neural networks using training data with explicit class labels; by contrast, we address retrieval tasks by fine-tuning the deep neural networks using triplet constraints. Due to the very different formulations, it is difficult and nontrivial to apply the existing techniques in our problem setting, and thus cannot compare with them directly in our empirical study. Finally, we note that our work is also different from some recent deep learning studies [34, 24] that apply deep learning to optimize hashing codes, making it difficult and unfair to be directly compared with our method.

2.4 Online Learning

In this paper, we explore online learning techniques to learn the multimodal similarity functions from the streams of triplet constraints. Online learning, a family of efficient and scalable machine learning algorithms, has been extensively studied for years. Unlike batch learning methods that usually suffer from expensive re-training cost whenever new training data arrive, online learning works in a sequential fashion by performing highly efficient (typically constant) updates for each new training data sequentially, making it highly scalable for large-scale applications. In literature, a variety of algorithm [6, 37, 9, 11, 52, 46] have been proposed for online learning. The most well-known algorithms is the Perceptron algorithm [37], which simply updates the weight vector of the learning model by adding the incoming instance with a constant weight whenever it is misclassified. In recent years, various algorithms [29, 9] have been proposed to improve the Perceptron,

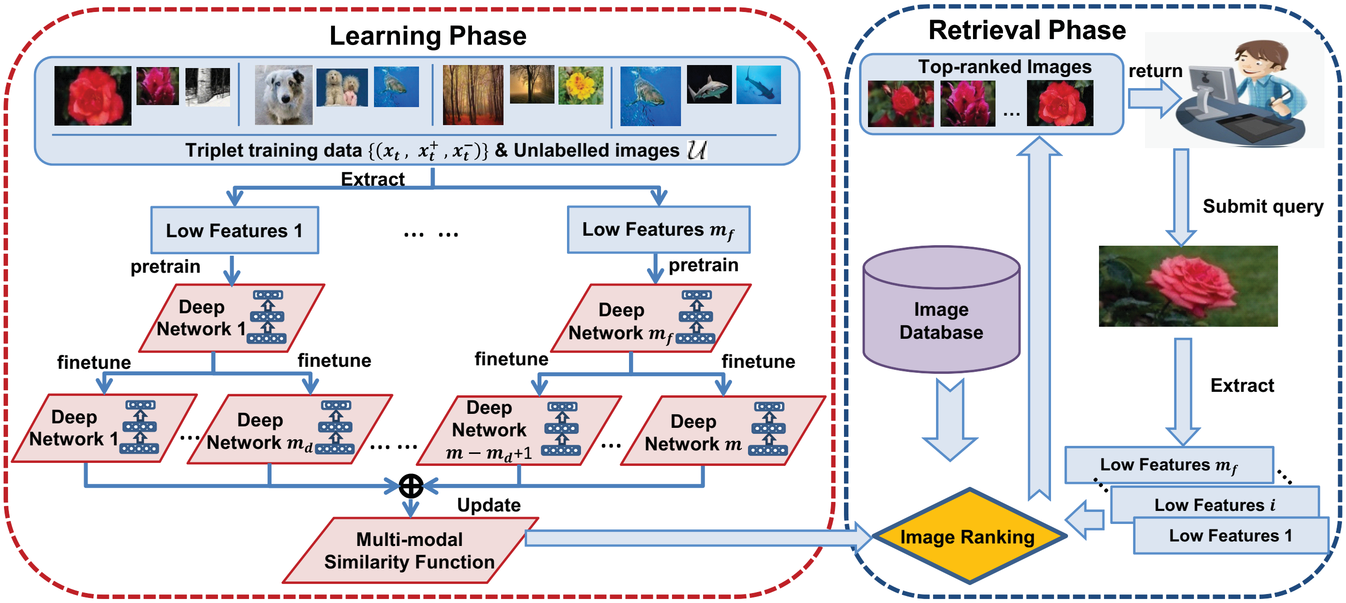


Figure 1: Overview of the proposed multimodal deep similarity learning scheme for image retrieval

many of them usually follow the principle of large margin learning. One notable approach is the Exponentiated Gradient online learning algorithm [6], which updates the classifier whenever it fails to produce a large margin on the received instance and meanwhile attempts to avoid too aggressive updates. In general, most existing online learning algorithms are designed for online classification tasks by learning models from a sequence of training data with class labels [18]. In this work, we apply the Exponentiated Gradient learning methodology to tackle the task of learning similarity functions from a sequence of triplet constraints.

3. PROBLEM SETTING

We address the fundamental problem of learning a pairwise similarity function for image retrieval from side information of pairwise/tripletwise image relationship. To formulate the learning task, we define the similarity function $S(\mathbf{x}_1, \mathbf{x}_2)$ for any two images $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$, and assume a collection of training data instances is given (sequentially) in the forms of triplet, i.e.,

$$\{(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-), t = 1, \ldots, T\},$$

where each triplet indicates the tripletwise relationship of three images, i.e., image $\mathbf{x}_t$ is more similar to image $\mathbf{x}_t^+$ in contrast to image $\mathbf{x}_t^-$, and T is the total number of triplets. The goal is to learn a similarity function $S(\cdot, \cdot)$ that can always produce the similarity values satisfying the tripletwise constrain as follows:

$$S(\mathbf{x}_t, \mathbf{x}_t^+) > S(\mathbf{x}_t, \mathbf{x}_t^-), \quad (\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \forall t; \tag{1}$$

The above discussion generally assumes similarity learning is performed on uni-modal data. We aim to generalize it for multi-modal data, where each image can be represented by different types of low-level feature descriptors (e.g., color, shape, or texture) and the similarity of any two images can be computed by defining multiple kinds of distance measures (e.g., cosine similarity, and Euclidean distance, etc.). In particular, we assume there are m_f types of feature descriptors and m_d kinds of measures, leading to a total of $m = m_f \times m_d$ modalities, each of which applies one kind of distance measure to compute the similarity of two images based on one type of feature descriptor.

The general idea of our multimodal similarity learning scheme is to learn optimal similarity functions $S_i(\cdot, \cdot)$ for each individual modality (with respect to one specific distance measure), and meanwhile identify an optimal combination of the m different modalities to obtain the final multimodal similarity function:

$$S(\mathbf{x}_1, \mathbf{x}_2) \equiv \sum_{i=1}^{m} \theta_i S_i(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}), \tag{2}$$

where $\theta_i \in [0, 1]$ denotes the combination weight for the i-th modality and $\mathbf{x}_1^{(i)}$ denotes the feature space of the i-th modality. In the following, without loss of clarity, we will simplify notation $S_i(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)})$ as $S_i(\mathbf{x}_1, \mathbf{x}_2)$ by removing the superscript.

To simultaneously learn both the optimal combination weights $\theta = (\theta_1, \ldots, \theta_m)$ and the optimal individual similarity functions $\{S_i | i = 1 \ldots, m\}$, we cast the multimodal similarity learning problem into the following optimization task:

$$\min_{\theta \in \Delta} \min_{S_i} \frac{1}{2} \sum_{i=1}^{m} \|S_i\|^2 + C \sum_{t=1}^{T} \mathcal{L}_t((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); S), \tag{3}$$

where $\Delta = \{\theta | \sum_{i=1}^{m} \theta_i = 1, \theta_i \in [0, 1], \forall i\}$, $\|S_i\|$ is a regularization term that limits the model complexity, $C > 0$ and

$$\mathcal{L}_t((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); S) = \max(0, S(\mathbf{x}_t, \mathbf{x}_t^-) - S(\mathbf{x}_t, \mathbf{x}_t^+) + 1),$$

which implicitly impose the constraints in (1) by introducing a margin factor $+1$ to ensure a sufficiently large difference. The crux of the above multimodal similarity learning framework lies in how to effectively define or construct the set of individual similarity functions S_i and how to efficiently resolve the resulting and often challenging optimization task.

4. MULTIMODAL DEEP SIMILARITY LEARNING FRAMEWORK

4.1 Overview

Figure 1 illustrates the system flow of the proposed multimodal deep similarity learning scheme for image retrieval. The goal is to learn the similarity functions in the learning phase in order to facilitate the image ranking task in the retrieval phase. During the learning phase, we assume triplet training data instances arrive sequentially, which is natural for a real-world CBIR system where users' online relevance feedback are usually collected in a sequential manner [17].

Following the optimization framework in (3), one possible approach is to define S_i using traditional distance metric learning for optimizing the Mahalanobis distances:

$$S_i(\mathbf{x}_1, \mathbf{x}_2) = -(\mathbf{x}_1 - \mathbf{x}_2)^\top M^{(i)} (\mathbf{x}_1 - \mathbf{x}_2), \qquad (4)$$

The optimization thus can be turned into the following:

$$\min_{\theta \in \Delta} \min_{M^{(i)} \succeq 0} \frac{1}{2} \sum_{i=1}^{m} \|M^{(i)}\|_F^2 + C \sum_{t=1}^{T} \mathcal{L}_t((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); S), \qquad (5)$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm and $M^{(i)} \succeq 0$ enforce $M^{(i)}$ to be positive semi-definite (PSD) matrix.

By decomposing $M^{(i)} = A_i^\top A_i$, we reformulate Eqn.(4) as:

$$\begin{aligned} S_i(\mathbf{x}_1, \mathbf{x}_2) &= -(\mathbf{x}_1 - \mathbf{x}_2)^\top A_i^\top A_i (\mathbf{x}_1 - \mathbf{x}_2) \\ &= -(A_i\mathbf{x}_1 - A_i\mathbf{x}_2)^\top (A_i\mathbf{x}_1 - A_i\mathbf{x}_2) \\ &= -\text{Euclidean}(\phi_i(\mathbf{x}_1), \phi_i(\mathbf{x}_2))^2 \end{aligned}$$

where $\phi_i(\mathbf{x}) = A_i\mathbf{x}$. From this perspective, DML is equivalent to learning a *linear* projection A_i, which maps $\mathbf{x}$ into a new representation. Unlike the regular DML, in this paper, we aim to go beyond the linear projection by exploring deep neural networks to learn a non-linear projection. Figure 2 shows the proposed architecture of three-layer neural network, which works as follows: (i) we first adopt Stacked Denoising Autoencoder to pretrain neural networks for each feature space with unlabelled data for each modal, (ii) we then finetune these neural networks by investigating a novel unified twofold online learning scheme: (a) learning to optimize the parameters of neural networks on each individual modality; and (b) meanwhile learning to find the optimal combination of multiple diverse types of modality.

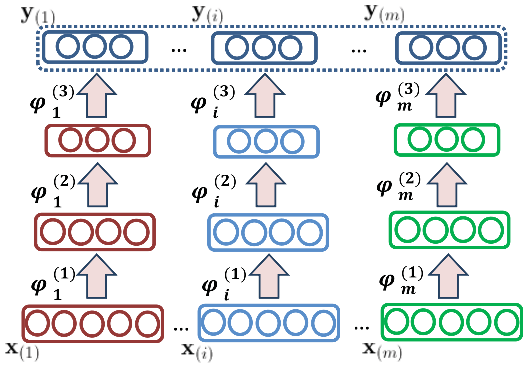


Figure 2: The proposed architecture of multimodal deep neural networks for learning multimodal similarity

4.2 Stacked Denoising Autoencoder

In this section, we briefly review Auto Encoder (AE) [5] and it extensions, including Denoising Autoencoder (DAE) [44] and Stacked Denoising Autoencoder (SDA) [45]. In our approach, we explore SDA to pretrain deep neural networks with unlabelled data.

Auto Encoder (AE), also called Auto Associator or Diabolo network, is trained to encode/map an input $\mathbf{x}$ to a hidden representation $\mathbf{y}$ such that $\mathbf{x}$ can be decoded/reconstructed from $\mathbf{y}$. Considering the hidden layer is encoded by a nonlinear one-layer neural network, we can set $\mathbf{y} = \varphi(\mathbf{x}) = s(W\mathbf{x} + b)$ and denote by $\mathbf{z} = \psi(\mathbf{y}) = s(W'\mathbf{y} + c)$ the reconstruction from $\mathbf{y}$, where $\mathbf{z}$ has the same shape as $\mathbf{x}$, and $s(\cdot)$ is a nonlinear function, e.g. the sigmoid. Depending on the distributional assumptions on the input given the code, we can apply diverse loss functions to measure the reconstruction error, e.g., if assuming the distribution of $p(\mathbf{x}|\mathbf{z})$ is Gaussian, we can adopt *squared error* as:

$$\mathcal{L}_{sq}(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|^2;$$

if we assume $\mathbf{x}$ is either binary or binomial probability, we can pose the loss function as the *cross-entropy* function:

$$\mathcal{L}_{ce}(\mathbf{x}, \mathbf{z}) = -\log P(\mathbf{x}|\mathbf{z}) = -\sum_i \mathbf{x}_i \log \mathbf{z}_i + (1 - \mathbf{x}_i)\log(1 - \mathbf{z}_i),$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the i-th entry of $\mathbf{x}$ and $\mathbf{z}$ individually. To minimize the above reconstruction error, AE seeks to learn the parameters W, W', b and c on the training dataset. In particular, the parameters are initialized randomly and then optimized by stochastic gradient descent. The part $(\mathbf{A})$ of Figure 3 illustrates the idea for auto encoder.

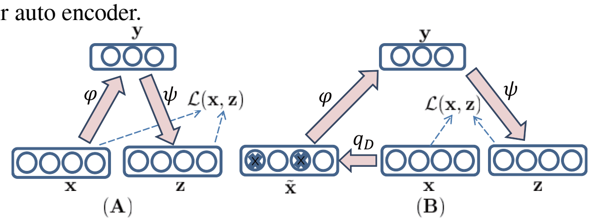


Figure 3: (A) Autoencoder; (B) Denoising Autoencoder

Denoising Autoencoder (DAE) is a simple variant of the basic autoencoder, which is trained to reconstruct a clean "repaired" input from its noisy version. Formally, we firstly construct $\mathbf{x}$'s noisy version $\tilde{\mathbf{x}}$ through a stochastic mapping $\tilde{\mathbf{x}} \sim q_D(\tilde{\mathbf{x}}|\mathbf{x})$, where q_D is a function to add noise into the original $\mathbf{x}$, which will be specified later. The noisy version $\tilde{\mathbf{x}}$ will be then mapped through AE to a hidden representation $\mathbf{y} = \varphi(\tilde{\mathbf{x}})$, where $\mathbf{y}$ is then used to reconstruct a clean version of $\tilde{\mathbf{x}}$ by $\mathbf{z} = \psi(\mathbf{y})$. Noted that the reconstruction error $\mathcal{L}(\mathbf{x}, \mathbf{z})$ but not $\mathcal{L}(\tilde{\mathbf{x}}, \mathbf{z})$ is minimized in DAE. The part $(\mathbf{B})$ of Figure 3 illustrates the idea for denoising autoencoder. To construct the map $\tilde{\mathbf{x}} \sim q_D(\tilde{\mathbf{x}}|\mathbf{x})$, Masking Noise strategy is exploited, where a fraction ν of the elements of $\mathbf{x}$ (chosen randomly for each example) is forced to 0.

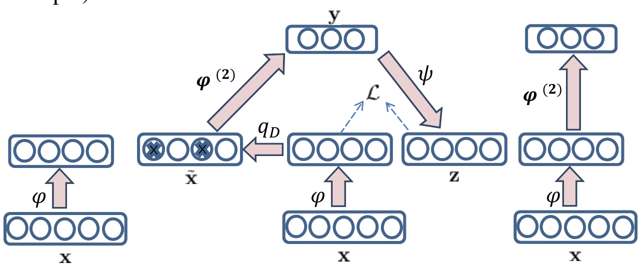


Figure 4: Stacking Denoising Autoencoder

Stacked Denoising Autoencoder (SDA) is utilized to initialize a deep network in the same way as stacking RBMs in deep belief networks. It is important to notice that input corruption is only used for the initial denoising-training of each individual layer, so that it may learn useful feature extractors. Once the mapping φ has thus been learnt, it will henceforth be used on uncorrupted inputs. In particular no corruption is applied to produce the representation that will serve as clean input for training the next layer. Figure 4 summarizes the complete procedure for learning and stacking serval layers of denoising autoencoders. Once a stack of encoders has thus been built, its highest level output representation can be used as input to a standalone supervised learning algorithm. The parameters of all layers can then be simultaneously fine-tuned using a gradient-based procedure such as stochastic gradient descent.

4.3 Online Learning Algorithms

After introducing Stacked Denoising Autoencoder, we now propose a novel method to tackle the optimization of the objective (3) with multiple deep models. Instead of directly solving the optimization task of Eqn. (3) in a batch learning fashion, in this section, we present an online learning algorithm to tackle the multi-modal similarity learning task. The key motivations are twofold: (i) solving Eqn.(3) directly can be computationally highly expensive for a large amount of training data; and (ii) more seriously, such a batch training process suffers from an extremely high re-training cost as whenever there is a new training data, the entire model has to be completely re-trained, making it non-scalable for a real-world CBIR application.

Specifically, the key challenge of our learning task is how to develop an efficient and scalable learning scheme that can optimize both the similarity functions on each individual modality and meanwhile optimize the combinational weights of different modalities. To this end, we propose to explore an online multi-modal similarity learning algorithm. In particular, after initializing a set of deep neural networks for m_f features by applying Denoising Autoencoder, we explore online learning techniques to update similarity function on each modality, i.e., once receiving a new arrived triplet, we finetune individual deep neural network based on the similarity functions, and then apply the Exponentiated Gradient online learning [6] to find the optimal combinational weights. We discuss each of the two learning tasks in detail below.

Let us denote $\mathbf{y} = \phi(\mathbf{x})$ the new representation of the image feature $\mathbf{x}$ in one single feature space, i.e., $\mathbf{y}$ is obtained by applying a nonlinear projection on $\mathbf{x}$ with an l-layer deep neural network, we can apply diverse approaches to define the similarity functions:

$$S_i(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} \frac{\mathbf{y}_1^T \mathbf{y}_2}{\|\mathbf{y}_1\| \|\mathbf{y}_2\|} & \text{(Cosine)} \\ -\|\mathbf{y}_1 - \mathbf{y}_2\|^2 & \text{(Euclidean)} \\ e^{\frac{-\|\mathbf{y}_1 - \mathbf{y}_2\|^2}{2\sigma}} & \text{(RBF Kernel)} \\ \mathbf{y}_1^T \mathbf{y}_2 & \text{(Linear Kernel)} \end{cases}$$

Then, we explore the hinge loss function as follows:

$$\mathcal{L}_t^{\gamma_s}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); S_i) = \max\{0, \gamma_s - S_i(\mathbf{x}_t, \mathbf{x}_t^+) + S_i(\mathbf{x}_t, \mathbf{x}_t^-)\}$$

where γ_s is the parameter of margin and $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$ is the t-th training triplet received. By defining $(\mathbf{y}_t, \mathbf{y}_t^+, \mathbf{y}_t^-)$ as the new representation of $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$ projected by deep neural networks, we can compute the derivation of $\mathcal{L}$ with respect to $\mathbf{y}_t, \mathbf{y}_t^+, \mathbf{y}_t^-$ individually as follows,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{y}_t} = \begin{cases} -\frac{1}{\|\mathbf{y}_t^+\| \|\mathbf{y}_t\|} \mathbf{y}_t^+ + \frac{1}{\|\mathbf{y}_t^-\| \|\mathbf{y}_t\|} \mathbf{y}_t^- & \\ \quad +(\frac{\mathbf{y}_t^T \mathbf{y}_t^+}{\|\mathbf{y}_t^+\| \|\mathbf{y}_t\|^3} - \frac{\mathbf{y}_t^T \mathbf{y}_t^-}{\|\mathbf{y}_t^-\| \|\mathbf{y}_t\|^3}) \mathbf{y}_t & \text{(Cosine)} \\ -2\mathbf{y}_t^+ + 2\mathbf{y}_t^- & \text{(Euclidean)} \\ \frac{\mathbf{y}_t - \mathbf{y}_t^+}{\sigma} e^{\frac{-\|\mathbf{y}_t - \mathbf{y}_t^+\|^2}{2\sigma}} - \frac{\mathbf{y}_t - \mathbf{y}_t^-}{\sigma} e^{\frac{-\|\mathbf{y}_t - \mathbf{y}_t^-\|^2}{2\sigma}} & \text{(RBF Kernel)} \\ -\mathbf{y}_t^+ + \mathbf{y}_t^- & \text{(Linear Kernel)} \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{y}_t^+} = \begin{cases} -\frac{1}{\|\mathbf{y}_t\| \|\mathbf{y}_t^+\|} \mathbf{y}_t + \frac{\mathbf{y}_t^T \mathbf{y}_t^+}{\|\mathbf{y}_t\| \|\mathbf{y}_t^+\|^3} \mathbf{y}_t^+ & \text{(Cosine)} \\ -2\mathbf{y}_t + 2\mathbf{y}_t^+ & \text{(Euclidean)} \\ \frac{\mathbf{y}_t^+ - \mathbf{y}_t}{\sigma} e^{\frac{-\|\mathbf{y}_t - \mathbf{y}_t^+\|^2}{2\sigma}} & \text{(RBF Kernel)} \\ -\mathbf{y}_t & \text{(Linear Kernel)} \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{y}_t^-} = \begin{cases} \frac{1}{\|\mathbf{y}_t\| \|\mathbf{y}_t^-\|} \mathbf{y}_t - \frac{\mathbf{y}_t^T \mathbf{y}_t^-}{\|\mathbf{y}_t\| \|\mathbf{y}_t^-\|^3} \mathbf{y}_t^- & \text{(Cosine)} \\ 2\mathbf{y}_t - 2\mathbf{y}_t^- & \text{(Euclidean)} \\ \frac{\mathbf{y}_t - \mathbf{y}_t^-}{\sigma} e^{\frac{-\|\mathbf{y}_t - \mathbf{y}_t^-\|^2}{2\sigma}} & \text{(RBF Kernel)} \\ \mathbf{y}_t & \text{(Linear Kernel)} \end{cases}$$

We then follow the idea of Online Gradient Descent [54] to update the parameters $(\mathbf{W}^{(l)}, b^{(l)})$ of the last layer of l-layer neural network as follows:

$$\mathbf{W}_{t+1}^{(l)} \leftarrow \mathbf{W}_t^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}} \tag{6}$$

$$b_{t+1}^{(l)} \quad \leftarrow \quad b_t^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial b^{(l)}} \tag{7}$$

where $\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}}$ and $\frac{\partial \mathcal{L}}{\partial b^{(l)}}$ are computed as follows, respectively:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}} = \sum_{i=1}^{d} (\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i} \frac{\partial \mathbf{y}_i}{\partial \mathbf{W}^{(l)}} + \frac{\partial \mathcal{L}}{\partial \mathbf{y}_i^+} \frac{\partial \mathbf{y}_i^+}{\partial \mathbf{W}^{(l)}} + \frac{\partial \mathcal{L}}{\partial \mathbf{y}_i^-} \frac{\partial \mathbf{y}_i^-}{\partial \mathbf{W}^{(l)}})|_{\mathbf{W}^{(l)} = \mathbf{W}_t^{(l)}}$$

$$\frac{\partial \mathcal{L}}{\partial b^{(l)}} = \sum_{i=1}^{d} (\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i} \frac{\partial \mathbf{y}_i}{\partial b^{(l)}} + \frac{\partial \mathcal{L}}{\partial \mathbf{y}_i^+} \frac{\partial \mathbf{y}_i^+}{\partial b^{(l)}} + \frac{\partial \mathcal{L}}{\partial \mathbf{y}_i^-} \frac{\partial \mathbf{y}_i^-}{\partial b^{(l)}})|_{b^{(l)} = b_t^{(l)}}$$

where d is the dimension of the representation space, $(\mathbf{y}, \mathbf{y}^+, \mathbf{y}^-)$ is the t-th arrived triplet, and $\mathbf{y}_i$ is the i-th entry of $\mathbf{y}$. Finally, we can adopt backpropagation [38] to update the parameters of other layers of deep networks.

Another key task of multi-modal DML is to learn the optimal combinational weights $\theta = (\theta^{(1)}, \ldots, \theta^{(m)})$, where $\theta^{(i)}$ is set to $1/m$ at the beginning of the learning task. In our approach, we apply the Exponentiated Gradient online learning algorithm [6] to find the combinational weights sequentially. In particular, we define

$$\mathbf{s}_t = (S_1(\mathbf{x}_t, \mathbf{x}_t^+) - S_1(\mathbf{x}_t, \mathbf{x}_t^-), \ldots, S_m(\mathbf{x}_t, \mathbf{x}_t^+) - S_m(\mathbf{x}_t, \mathbf{x}_t^-))^T,$$

and formulate the optimization problem as follows:

$$\theta_{t+1} = \arg\min_{\theta} \; D_{KL}(\theta \| \theta_t) + \lambda \ell_t(\theta) \;\; s.t. \;\; \theta \in \Delta \tag{8}$$

where $D_{KL}(\mathbf{u} \| \mathbf{v}) = \sum_i u_i \ln(\frac{u_i}{v_i})$ is the KL-divergence, and $\ell_t(\theta)$ is a hinge loss defined as:

$$\ell_t(\theta) = \max(0, \gamma - \theta^T \mathbf{s}_t).$$

The above optimization aims to trade off two major concerns: (i) the updated model should not be deviated too much from the previous model, which is measured by the KL-divergence between the two models; and (ii) the updated model suffers a small loss on the triplet training instance $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$. The trade-off is essentially achieved by introducing the penalty cost parameter $C > 0$. In general, it is hard to derive a closed-form solution for the above optimization. To further simplify the optimization, we approximate the objective function by applying the first-order Taylor expansion of $\ell_t(\theta)$ at θ_t, and thus turn the optimization into the following:

$$\theta_{t+1} = \arg\min_{\theta \in \Delta} \; D_{KL}(\theta \| \theta_t) + \lambda[\ell_t(\theta_t) + \nabla \ell_t(\theta_t)(\theta - \theta_t)] \tag{9}$$

We can derive a closed-form solution in the following proposition.

PROPOSITION 1. *The closed-form solution to the optimization in (9) is:*

$$\theta_{t+1,i} = \frac{\theta_{t,i} e^{-\lambda \nabla \ell_t(\theta_t)_i}}{\sum_j \theta_{t,j} e^{-\lambda \nabla \ell_t(\theta_t)_j}}, i = 1, \ldots, m \tag{10}$$

where $\nabla \ell_t(\theta_t) = -s_t$ *if* $\ell_t(\theta_t) > 0$ *or* $\nabla \ell_t(\theta_t) = 0$ *if* $\ell_t(\theta_t) = 0$.

The above proposition can be obtained by following the idea of Exponentiated Gradient learning in [6]. We omit the detail proof because of the space limitation. From the results, we can see that θ remains unchanged if $\theta^T \mathbf{s}_t \geq \gamma$. That is, we will update the θ whenever the current θ fails to rank the order of $\mathbf{x}_t^+$ and $\mathbf{x}_t^-$ w.r.t. query $\mathbf{x}_t$ correctly at a sufficiently large margin.

Algorithm 1 OMDSL—Online Multimodal Deep Similarity Learning

```
INPUT:
    • Unlabelled data: $\mathcal{U}$
    • Training Triplets: $\{(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)\}_{t=1,2,\ldots}$
Initialize weights: $\theta_{1,i} = 1/m, \forall i = 1, \ldots, m$
Pretrain $m_f$ $l$-layer deep networks with unlabelled data for
each feature space by adopting SDA (Fig. 4)
for $t = 1, 2, \ldots$ do
    Receive: $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$
    for $i = 1, 2, \ldots, m$ do
        Update the parameters $(W_{(i)}^{(l)}, b_{(i)}^{(l)})$ of last layer of deep
        networks by Eqn 8.
        Adopt Backpropagation [38] to finetune other parameters
        $(W_{(i)}^{(1)}, W_{(i)}^{(l-1)}, b_{(i)}^{(1)}, \ldots, b_{(i)}^{(l-1)})$
    end for
    Compute: $s_{t,i} = S_i(\mathbf{x}_t, \mathbf{x}_t^+) - S_i(\mathbf{x}_t, \mathbf{x}_t^-), i = 1, \ldots, m$
    Compute: $\ell_t(\theta_t) = \max(0, \gamma - \theta_t^\top \mathbf{s}_t)$
    if $\ell_t(\theta_t) > 0$ then
        $\theta_{t+1,i} = \frac{\theta_{t,i} e^{-\lambda \nabla \ell_t(\theta_t)_i}}{\sum_{j=1}^m \theta_{t,j} e^{-\lambda \nabla \ell_t(\theta_t)_j}}, i = 1, \ldots, m$
    end if
end for
```

Finally, Algorithm 1 summarizes the details of the proposed Online Multimodal Deep Similarity Learning (OMDSL) algorithm. The space complexity of the algorithm is $\mathcal{O}(\sum_{j=1}^m \sum_{i=1}^l d_{i-1}^{(j)} \times d_i^{(j)})$, where $d_0^{(j)} = D^{(j)}$ is the dimension of j-th feature space, and $d_i^{(j)} (i = 1, \ldots, l)$ is the dimension of i-th hidden layer representation space of j-th deep networks. Because usually we choose $d_i^{(j)} \times d_{i-1}^{(j)} << D^{(j)} \times d_1^{(j)}; (i = 2, \ldots, l)$, the first layer of deep network will dominate the space complexity. By denoting $D = \max(D^1, \ldots, D^m)$ and $d = \max(d_1^{(1)}, \ldots, d_1^{(m)})$, the worst-case space complexity is simply $\mathcal{O}(mDd)$, and the worst-case overall time complexity of the algorithm is $\mathcal{O}(TmDd)$, which is linear with respect to T — the total number of training triplets.

4.4 Theoretical Analysis

We now analyze the theoretical bounds of the proposed OMDSL algorithm. In particular, our goal is to bound the number of mistakes suffered by the proposed algorithm, denote by M, which measures the total number of cases where the model fails to predict the triplet constraints over the entire sequence, formally defined as:

$$M = \sum_{t=1}^{T} \mathbb{I}(S(x_t, x_t^+) < S(x_t, x_t^-)) \tag{11}$$

where $\mathbb{I}(\cdot)$ is an indicator function that output 1 when the statement holds and 0 otherwise. The following gives our theorem.

Theorem 1. *Assume $\|\mathbf{s}_t\|_\infty \leq C$ and the OMDSL algorithm runs with a learning rate of $\lambda = \sqrt{\frac{2 \ln m}{C^2 T}}$ over a sequence of triplets $(\mathbf{x}_1, \mathbf{x}_1^+, \mathbf{x}_1^-), \ldots, (\mathbf{x}_T, \mathbf{x}_T^+, \mathbf{x}_T^-)$. Then, for any multimodal similarity function $S = \sum_{i=1}^m \theta_i S_i$ with $\forall \theta \in \Delta$, we have the following bound:*

$$M \leq \frac{1}{\gamma} \sum_{t=1}^{T} \mathcal{L}_t^\gamma((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); S) + \frac{C}{\gamma} \sqrt{\frac{T \ln m}{2}} \tag{12}$$

Proof. The key idea of this proof follows the principle of similar proof in [6]. In particular, according to the definition of ℓ_t and its convexity property, we have

$$M = \sum_{t=1}^{T} \mathbb{I}(S(x_t, x_t^+) < S(x_t, x_t^-)) \leq \frac{1}{\gamma} \sum_{t=1}^{T} \ell_t(\theta_t) \tag{13}$$

By applying the Taylor's theorem to ℓ_t, we can obtain

$$\ell_t(\theta_t) - \ell_t(\theta) \leq -(\theta - \theta_t) \cdot \nabla \ell_t(\theta_t) \tag{14}$$

The next is to bound the right hand side of the above inequality. To facilitate the analysis, we denote $\mathbf{z} = \lambda \nabla \ell_t(\theta_t)$ and $\mathbf{v} = \theta_t \cdot \mathbf{z} - \mathbf{z}$. We then have the following:

$$\begin{aligned}
-(\theta - \theta_t) \cdot \mathbf{z} &= -\theta \cdot \mathbf{z} + \theta_t \cdot \mathbf{z} - \ln(\sum_{i=1}^{m} \theta_{t,i} e^{v_i}) + \ln(\sum_{i=1}^{m} \theta_{t,i} e^{v_i}) \\
&= -\theta \cdot \mathbf{z} - \ln(\sum_{i=1}^{m} \theta_{t,i} e^{-z_i}) + \ln(\sum_{i=1}^{m} \theta_{t,i} e^{v_i}) \\
&= \sum_{j=1}^{m} \theta_j \ln e^{-z_j} - \ln(\sum_{i=1}^{m} \theta_{t,i} e^{-z_i}) + \ln(\sum_{i=1}^{m} \theta_{t,i} e^{v_i}) \\
&= \sum_{j=1}^{m} \theta_j \ln(\frac{1}{\theta_{t,j}} \frac{\theta_{t,j} e^{-z_j}}{\sum_{i=1}^{m} \theta_{t,i} e^{-z_i}}) + \ln(\sum_{i=1}^{m} \theta_{t,i} e^{v_i}) \\
&= \sum_{j=1}^{m} \theta_j \ln \frac{\theta_{t+1,j}}{\theta_{t,j}} + \ln(\sum_{i=1}^{m} \theta_{t,i} e^{v_i}) \\
&= D_{KL}(\theta \| \theta_t) - D_{KL}(\theta \| \theta_{t+1}) + \ln(\sum_{i=1}^{m} \theta_{t,i} e^{v_i})
\end{aligned}$$

Plugging the above into (14) and summing over all t leads to:

$$\sum_{t=1}^{T} [\ell_t(\theta_t) - \ell_t(\theta)] \leq \frac{1}{\lambda} [D_{KL}(\theta \| \theta_1) + \sum_{t=1}^{T} \ln(\sum_{i=1}^{m} \theta_{t,i} e^{v_i})] \tag{15}$$

where $-D_{KL}(\theta \| \theta_{T+1})$ is omitted. To bound the right hand side of the above inequality, we note that $D_{KL}(\theta \| \theta_1) \leq \ln m$, since $\theta_1 = (1/m, \ldots, 1/m)$. And we need bound the second term in the right hand side. Since $\|\mathbf{s}_t\|_\infty \leq C$, then $|z_i| \leq \lambda C$; by applying Hoeffding's inequality, we can get

$$\ln(\sum_{i=1}^{m} \theta_{t,i} e^{v_i}) \leq \lambda^2 C^2 / 2 \tag{16}$$

As a result, we have the following

$$\sum_{t=1}^{T} [\ell_t(\theta_t) - \ell_t(\theta)] \leq \frac{\ln m}{\lambda} + \frac{\lambda C^2 T}{2} = C \sqrt{\frac{T \ln m}{2}} \tag{17}$$

Re-arranging the above inequality concludes the theorem. □

The above theorem basically shows that the number of mistakes suffered by the proposed algorithm is bounded by the rate of $\mathcal{O}(\sqrt{T})$. Moreover, assume the best similarity function among all modalities is known in hindsight, e.g., assuming the i-th modality, we then have the following corollary.

Theorem 2. *Assume $\|\mathbf{s}_t\|_\infty \leq C$ and the OMDSL algorithm runs with a learning rate of $\lambda = \sqrt{\frac{2 \ln m}{C^2 T}}$ over a sequence of triplets $(\mathbf{x}_1, \mathbf{x}_1^+, \mathbf{x}_1^-), \ldots, (\mathbf{x}_T, \mathbf{x}_T^+, \mathbf{x}_T^-)$, we then have*

$$M \leq \frac{1}{\gamma} \sum_{t=1}^{T} \mathcal{L}_t^\gamma((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); S_i) + \frac{C}{\gamma} \sqrt{\frac{T \ln m}{2}} \tag{18}$$

where S_i denotes the similarity function of the best modality.

The above follows immediately by setting $\theta_i = 1$ and $\theta_j = 0$ for $j \neq i$. This above result bounds the relationship between the total number of mistakes suffered by the proposed algorithm and the cumulative loss suffered by the best modality known in hindsight.

5. EXPERIMENTS

In this section, we conduct an extensive set of experiments to evaluate the efficacy of the proposed algorithm for similarity search on images with multimodal representations in CBIR.

5.1 Experimental Testbeds and Setup

We adopt four publicly available image data sets in our experiments, including COREL [16], Caltech101[1], Indoor [35][2], and ImageCLEF[3]. For each database, we randomly split it into three disjoint partitions: a test set of 500 images, a validation set of 500 images, and the rest images for training. Besides, to examine the scalability, we also include a large database "ImageCLEF+Flickr", which includes the ImageCLEF set as a ground-truth subset and additional 1-million Flickr images as background images.

To generate side information in the forms of triplet instances for learning the similarity functions, we sample the sequence of triplets from the images in the training set according to their ground truth class labels. Specifically, we generate a triplet instance by randomly sampling two images belonging to the same class and one image from another different class. In total, we generate 100K triplet instances for each standard dataset (except for the small-scale and large-scale experiments).

To fairly evaluate different algorithms, we choose their parameters using the same cross validation scheme. We adopt three-layer neural networks with $d_1 = 100$, $d_2 = 50$ and $d_3 = 50$ for all modalities on OMDSL, and set the maximum iteration to 500 for LMNN. To evaluate the retrieval performance, we adopt the standard mean Average Precision (mAP) and top-K retrieval accuracy.

5.2 Visual Features and Similarity Measures

We adopt both global and local feature descriptors for representing images. Before that, we have resized all the images to the scale of 500×500 pixels while keeping the aspect ratio unchanged. For global features, we extract five types of features to represent an image, including (1) color histogram and color moments (81 dimensions), (2) edge direction histogram (37 dimensions), (3) Gabor wavelets transformation (120 dimensions), (4) local binary pattern (59 dimensions), and (5) GIST features (512 dimensions). For local features, we extract the bag-of-visual-words representation [49] using two kinds of descriptors: (i) SIFT feature [30] and Hessian-Affine interest region detector with a threshold of 500; and (ii) SURF feature [3] and SURF detector with a threshold of 500. For the clustering step, we adopt a forest of 16 kd-trees and search 2048 neighbors to speed up the clustering task. By combining different descriptors (SIFT/SURF) and vocabulary sizes (200/1000), we extract four types of local features: SIFT200, SIFT1000, SURF200 and SURF1000. Finally, we adopt the TF-IDF weighing scheme to generate the final bag-of-visual-words. For all experiments, we normalize the feature vectors to the range of $[0, 1]$.

For each query-instance pair, we adopt four similarity functions to measure their similarity as described in Section 4.3, where we choose $\gamma_s = 0.25$ for Cosine similarity, $\gamma_s = 1$ for Euclidean, $\sigma = 2, \gamma_s = 0.25$ for RBF kernel, and $\gamma_s = 1$ for Linear kernel. Finally, we explore a total of 36 modalities to measure the overall image similarity for CBIR.

5.3 Comparison Algorithms

To extensively evaluate the efficacy of our algorithms, we compare the proposed OMDSL algorithm for image retrieval, against several representative distance metric learning algorithms, including RCA [2], LMNN [47], OASIS [8]. We also evaluate a heuristic baseline method with squared Euclidean distance, named as "EUCL-*". To adapt the existing DML methods for multimodal image retrieval, we have implemented several variants of each DML algorithm by exploring three fusion strategies [42, 25]: (i) "**B**est" — applying DML for each modality individually and then selecting the best modality, (ii) "**C**oncatenation" — an early fusion approach by concatenating features of all modalities before applying DML, and (iii) "**U**niform combination" — a late fusion approach by uniformly combining all modalities after metric learning.

5.4 Evaluation on Small-Scale Datasets

In this section, we build four small-scale data sets, named "Caltech101(S)", "Indoor(S)", "COREL(S)" and "ImageCLEF(S)", from the corresponding standard datasets by first choosing 10 object categories, and then randomly sampling 50 examples from each category. We adopt 5 global features described above as the multimodal inputs. To construct triplet constraints for online learning approaches, we generate all positive pairs (two images belong to the same class), and for each positive pair we randomly select an image from the other different classes to form a triplet. In total, about 10K triplets are generated for each dataset.

Table 1: Evaluation of the mAP performance.

Alg.	COREL(S)	Caltech101(S)	Indoor(S)	ImageCLEF(S)
Eucl-B	0.4431	0.4299	0.1726	0.4325
RCA-B	0.5097	0.4984	0.1915	0.4492
LMNN-B	0.4876	0.5462	0.1852	0.5231
OASIS-B	0.4445	0.5072	0.1884	0.4424
Eucl-C	0.5220	0.4306	0.1842	0.4431
RCA-C	0.6437	0.6156	0.2078	0.5927
LMNN-C	0.5816	0.5894	0.2027	0.5821
OASIS-C	0.5657	0.5441	0.2017	0.5618
Eucl-U	0.5220	0.4306	0.1842	0.4431
RCA-U	0.5625	0.4860	0.1894	0.4909
LMNN-U	0.6026	0.4282	0.2007	0.4647
OASIS-U	0.5679	0.5419	0.1989	0.5338
OMDSL	**0.7379**	**0.7136**	**0.2430**	**0.7552**

Table 1 summarizes the evaluation results on the small-scale data sets. The postfix "-B" indicates that we learn metrics on each modality (type of features) separately and then select the *best* single metric as the final similarity function via cross validation. The postfix "-C" indicates that we first *concatenate* all types of features together, and then learn the metric on the combined feature space, which is a kind of "early fusion" strategy. The postfix "-U" indicates we first learn an optimal metric for each modality, and then uniformly combine all distance functions for the final ranking, which is a kind of "late fusion" strategy. Table 2 shows the running time cost on the COREL(S) data set.

Table 2: Running time (in sec.) on COREL(S).

RCA-C	LMNN-C	OASIS-C	RCA-U
5.07	1442.66	404.35	2.91
LMNN-U	OASIS-U	OMDSL	—
858.94	376.77	300.27	—

We can draw several observations from the results in Table 1 and 2. First of all, the two kinds of fusion strategies, i.e., early fusion (with postfix"-C") and late fusion (with postfix"-U"), generally tend to outperform the best single metric approaches (with postfix"-B"). This is primarily because combining multiple types of features for retrieval could better explore the potential of all the features, which validates the importance of the proposed technique. Second, some of the uniformly combination algorithms (i.e., the

[1] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

[2] http://web.mit.edu/torralba/www/indoor.html

[3] http://imageclef.org/

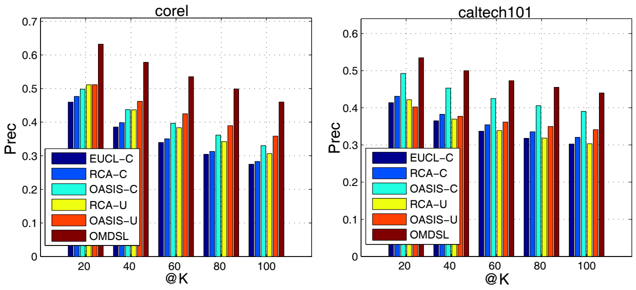
late fusion strategy) failed to outperform the best single metric approach in some cases, e.g., “RCA-U” (compared with “RCA-B”) and “LMNN-U” (compared with “LMNN-B”) on Caltech101(S). This implies that linear concatenation is not optimal to combine different kinds of features. Thus, it is critical to identify the effective features via machine learning and then assign them higher weights. Third, among all the compared algorithms, the proposed OMDSL significantly outperform the other algorithms, which validates that nonlinear learning with deep networks can learn more flexible similarity functions than traditional DML approaches. Finally, it is interesting to observe that OMDSL is even more efficient than most algorithms including OASIS, which is mainly because of the online learning strategy and the hidden layers of deep networks lying in intrinsic lower-dimensional space.

5.5 Evaluation on the Standard Datasets

We further evaluate the algorithms on standard-sized datasets. We exclude LMNN because of its extremely high computational cost. Following the standard setup with 5 global features and 4 local features, Table 3 summarizes the mAP results, and Figure 5 presents the top-K precisions on two randomly selected datasets.

Table 3: Evaluation of the mAP performance.

Alg.	COREL	Caltech101	Indoor	ImageCLEF
Eucl-B	0.1877	0.2187	0.0469	0.5523
RCA-B	0.2305	0.2837	0.0499	0.6010
OASIS-B	0.1958	0.3025	0.0522	0.6723
Eucl-C	0.2628	0.2259	0.0559	0.5752
RCA-C	0.2714	0.2473	0.0604	0.6272
OASIS-C	0.3202	0.3660	0.0726	0.7394
Eucl-U	0.2628	0.2259	0.0559	0.5752
RCA-U	0.2992	0.2413	0.0565	0.6161
OASIS-U	0.3594	0.3243	0.0705	0.6891
OMDSL	**0.4769**	**0.4476**	**0.0990**	**0.8387**



(a) *Corel* (b) *Caltech101*

Figure 5: Precision at Top-K results

From the results, we observed that the proposed OMDSL algorithm considerably surpasses all the other approaches for all cases, not only on mAP performance but also on top-K retrieval accuracy. This clearly validates the efficacy of the proposed algorithm for learning an effective similarity function on multi-modal data.

Table 4: Running time (in sec.) on “COREL”.

RCA-C	OASIS-C	RCA-U	OASIS-U	OMDSL
468.19	65060.93	184.3	8781.54	3976.07

Finally, Table 4 summarizes the average running time cost on the COREL dataset with 100K triplet instances. The result shows that OMSDL is considerably more efficient and scalable than OASIS, making it practical for large-scale applications.

5.5.1 Deep Networks with Varied Numbers of Layers

To evaluate the impact of the number of layers in the deep networks for the proposed OMDSL algorithm, we construct five diverse deep networks with different numbers of layers, i.e., from 1-layer to 5-layer. Table 5 lists the dimensionality settings of the hidden layers configured for each of them individually.

Table 5: Dimensionality settings of the hidden layers.

1-layer	2-layer	3-layer	4-layer	5-layer
50	100, 50	100, 50, 50	100, 50, 50, 50	100, 50, 50, 50, 50

We evaluate mAP performance on four datasets in Table 6. It is interesting to observe that 3-layer and 4-layer deep networks tend to achieve the best, primarily because too high layer structure leads to overly flexible structure that might overfit the training data.

Table 6: Evaluation of the mAP of the proposed OMDSL with diverse deep networks using varied numbers of hidden layers.

Alg.	Caltech101	Indoor	ImageCLEF	COREL
1-layer	0.4374	0.0891	0.8155	0.4427
2-layer	0.4445	0.0982	0.8302	0.4698
3-layer	**0.4476**	0.0990	0.8387	0.4769
4-layer	0.4393	**0.0993**	**0.8395**	**0.4787**
5-layer	0.4390	0.0952	0.8381	0.4719

5.5.2 Evaluation of Pretraining

To examine the importance of pretraining, we conduct experiments by randomly initializing the parameters of deep networks with standard normal distribution, and then performing the same finetune stage to refine deep networks. Table 7 shows the results.

Table 7: mAP on OMDSL without pretraining stage.

Alg.	Caltech101	Indoor	ImageCLEF	COREL
2-layer w/o pre	0.3892	0.0786	0.7635	0.4123
3-layer w/o pre	0.3974	0.0774	0.7685	0.4203

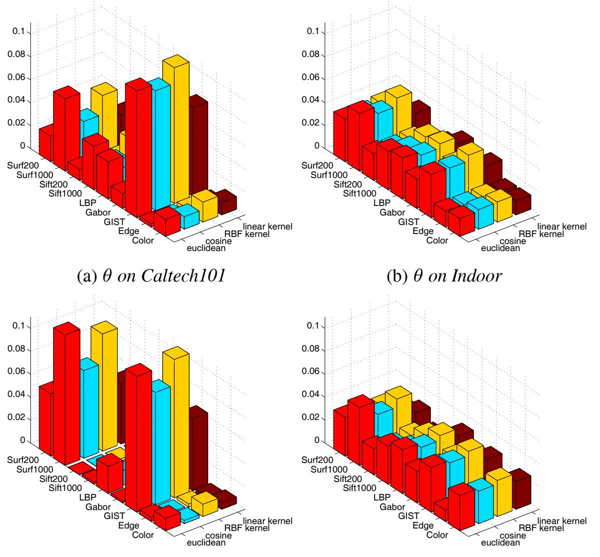
Comparing Table 7 with Table 6, we observe that the performance drops significantly without pretraining, which proves stacked denoising autoencoder is helpful to initialize deep networks.

5.5.3 Analysis of Modality Weights

We now examine the combination weights of different modalities learned by the proposed algorithm. Figure 6 visualizes the results of weights θ for diverse modality learned on four different datasets. From the results, we observe that GIST and Surf1000 features are assigned much higher weights than the others on “Caltech101” and “ImageCLEF”. This is primarily because “Caltech101” is an object dataset and “ImageCLEF” is a medical image dataset, which could be more sensitive to texture-related descriptors. By contrast, since “Indoor” and “Corel” contain many pictures about natural scenes and diverse categories, different modalities can make more or less contributions for different scenarios, further validating the importance of finding the optimal combination by our method.

5.6 Evaluation on the Large-scale Dataset

To test its scalability, we apply the proposed algorithm on a large-scale image retrieval application on “ImageCLEF+Flickr”, which has over 1-million images and 300K triplet training data. Table 8 shows the mAP performance of the five algorithms. Clearly, our proposed algorithm OMDSL achieves the best mAP. Figure 7 presents the top-K precisions on ImageCLEF+Flickr. We can have the similar observation that our proposed methods significantly outperform the state of the art, in terms of precision. In short, the proposed algorithm significantly outperforms the state of the art, in terms of both mAP and retrieval accuracy performance measures.



(a) *θ on Caltech101* (b) *θ on Indoor*

(a) *θ on ImageCLEF* (b) *θ on Corel*

Figure 6: The weights θ learned by OMDSL.

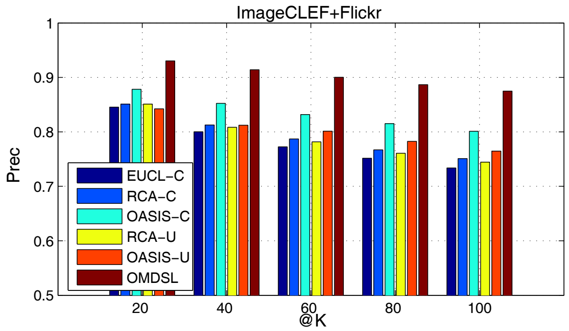


Figure 7: Precision at Top-K on "ImageCLEF+Flickr"

Table 8: mAP on "ImageCLEF+Flickr"

Eucl-C	RCA-C	OASIS-C	RCA-U	OASIS-U	OMDSL
0.5766	0.6163	0.7161	0.6219	0.7028	0.7592

5.7 Qualitative Comparison

Finally, to examine the qualitative retrieval performance, we randomly sample some query images from the query set, and compare the qualitative retrieval results. Figure 9 shows the qualitative results on "COREL" and "Caltech101" datasets. In each block, the first image is the query, and the results from the first to the fourth line represents "Eucl-C", "RCA-C", "OASIS-C", "RCA-U", "OASIS-U" and "OMDSL" respectively. The results show that OMDSL generally returns more relevant results than the others.

There are also some failure cases, as shown in Figure 8, where although all the returned images by "OMDSL" are facial images, only the second belongs to the query person, while Euclidean distance returns three correct images (1st, 2nd and 5th place of the top-5 list). We conjecture that the possible reasons include (i) the adopted features are generic but not face-specific features; (ii) the labeled data assume any two facial images belong to the same category, which cannot differentiate two different persons.

6. CONCLUSIONS

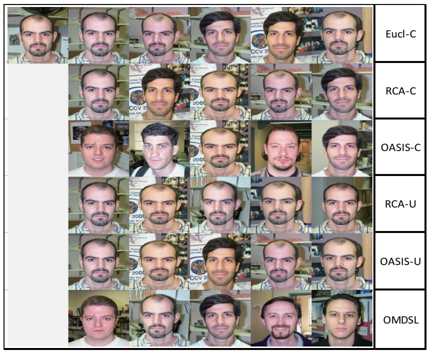


Figure 8: Example of a failure case from "Caltech101"

This paper investigated the fundamental problem of learning similarity functions for multimodal image retrieval. To address the limitations of regular DML approaches, we proposed a novel framework of Online Multimodal Deep Similarity Learning (OMDSL), which integrates with multiple deep networks optimally through an efficient and scalable online learning scheme. In particular, OMDSL explores a unified two-stage online learning scheme by (i) learning a flexible nonlinear similarity function for each individual modality, and (ii) learning to find the optimal combination of multiple modalities simultaneously via a coherent online learning process. Extensive experiments were conducted on several public image datasets, in which the encouraging results showed that OMDSL is promising for multimodal similarity search. Although our current experiments were focused on image retrieval, this work is rather generic for any multimedia retrieval tasks. We plan to extend it to resolve other multimedia applications in the future.

Acknowledgments

This research was supported in part by MOE tier-1 project grant (RG33/11), and IDM/NRF Research Grant No.: MDA/IDM/2012/8/8-2 VOL 01.

7. REFERENCES

[1] S. Akaho. A kernel method for canonical correlation analysis. In *IMPS*. Springer-Verlag, 2001.

[2] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *ICML*, pages 11–18, 2003.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417. Springer, 2006.

[4] Y. Bengio. Learning deep architectures for AI. *FTML*, 2(1):1–127, 2009.

[5] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, pages 153–160. MIT Press, 2007.

[6] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[7] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, pages 1–12, 2011.

[8] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11:1109–1135, 2010.

[9] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *JMLR*, 7:551–585, 2006.

[10] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *TASLP*, 20(1):30–42, 2012.

[11] M. Dredze, K. Crammer, and F. Pereira. Confidence-weighted linear classification. In *ICML*, pages 264–271, 2008.

[12] J. D. R. Farquhar, H. Meng, S. Szedmak, D. R. Hardoon, and J. Shawe-taylor. Two view learning: Svm-2k, theory and practice. In *NIPS*. MIT Press, 2006.

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

[14] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *NC*, 18(7):1527–1554, July 2006.

[15] S. C. Hoi, W. Liu, and S.-F. Chang. Semi-supervised distance metric learning for collaborative image retrieval. In *CVPR*, June 2008.

[16] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *CVPR*, June 17–22 2006.

[17] S. C. Hoi, M. R. Lyu, and R. Jin. A unified log-based relevance feedback scheme for image retrieval. *TKDE*, 18(4):509–204, 2006.

[18] S. C. Hoi, J. Wang, and P. Zhao. *LIBOL: A Library for Online Learning Algorithms*. Nanyang Technological University, 2012.

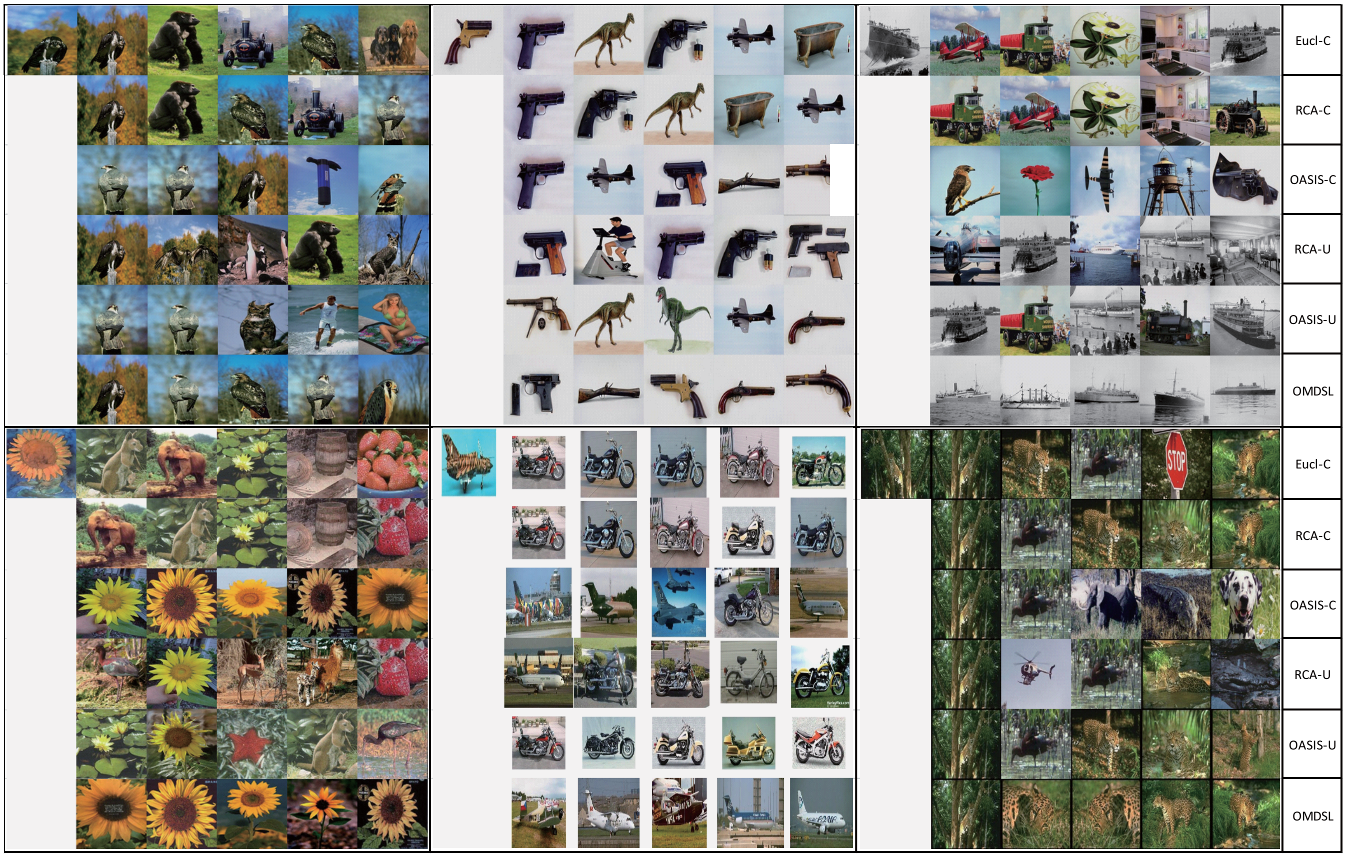


Figure 9: Qualitative evaluation of top-5 retrieved images. The first row is from "COREL" and the second is from "Caltech101".

[19] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman. Online metric learning and fast similarity search. In *NIPS*, pages 761–768, 2008.

[20] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *TPAMI*, 34(9):1704–1716, Sept. 2012.

[21] R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. In *NIPS*, pages 862–870, 2009.

[22] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *TPAMI*, 30(11):1877–1890, 2008.

[23] A. Joly and O. Buisson. Random maximum margin hashing. In *CVPR*, pages 873–880, Washington, DC, USA, 2011.

[24] Y. Kang, S. Kim, and S. Choi. Deep learning to hash with multiple representations. In *ICDM*, pages 930–935, 2012.

[25] J. Kludas, E. Bruno, and S. Marchand-Maillet. Information fusion in multimedia information retrieval. *AMR*, pages 147–159, 2008.

[26] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114. 2012.

[27] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

[28] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP*, 2(1):1–19, 2006.

[29] Y. Li and P. M. Long. The relaxed online maximum margin algorithm. In *NIPS*, pages 498–504, 1999.

[30] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[31] B. S. Manjunath and W.-Y. Ma. Texture features for browsing and retrieval of image data. *TPAMI*, 18(8):837–842, 1996.

[32] B. McFee and G. Lanckriet. Learning multi-modal similarity. *JMLR*, 12:491–523, 2011.

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

[34] M. Norouzi, D. Fleet, and R. Salakhutdinov. Hamming distance metric learning. In *NIPS*. 2012.

[35] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[36] R. Rahmani, S. A. Goldman, H. Zhang, S. R. Cholleti, and J. E. Fritts. Localized content-based image retrieval. *TPAMI*, 30(11):1902–1912, 2008.

[37] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.

[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Neurocomputing: foundations of research*. MIT Press, Cambridge, MA, USA, 1988.

[39] R. Salakhutdinov and G. Hinton. Semantic hashing. *IJAR*, 50(7), July 2009.

[40] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. *AISTATS*, 5:448–455, 2009.

[41] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 22(12):1349–1380, 2000.

[42] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In *MM*, pages 399–402. ACM, 2005.

[43] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*. 2012.

[44] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008.

[45] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11:3371–3408, Dec. 2010.

[46] J. Wang, P. Zhao, and S. C. H. Hoi. Exact soft confidence-weighted learning. In *ICML*, 2012.

[47] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2006.

[48] L. Wu, S. C. Hoi, and N. Yu. Semantics-preserving bag-of-words models and applications. *TIP*, 19(7):1908–1920, 2010.

[49] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *MIR*, pages 197–206, 2007.

[50] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, pages 1–51, 2006.

[51] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao. Multiview metric learning with global consistency and local smoothness. *TIST*, 3(3):53, 2012.

[52] P. Zhao, S. C. H. Hoi, and R. Jin. Double updating online learning. *JMLR*, 12:1587–1615, 2011.

[53] S.-h. Zhong, Y. Liu, and Y. Liu. Bilinear deep learning for image classification. In *MM*, pages 343–352, Scottsdale, Arizona, 2011.

[54] M. Zinkevich. Online convex programming and generalized infinitesimal gradientascent. In *ICML*, pages 928–936, 2003.