Attention Driven Multi-modal Similarity Learning

Xinjian Gao[1], Tingting Mu[2], John Y. Goulermas[3], Meng Wang[1]

Abstract

To learn a function for measuring similarity or relevance between objects is an important machine learning task, referred to as similarity learning. Conventional methods are usually insufficient for processing complex patterns, while more sophisticated methods produce results supported by parameters and mathematical operations that are hard to interpret. To improve both model robustness and interpretability, we propose a novel attention driven multi-modal algorithm, which learns a distributed similarity score over different relation modalities and develops an interaction-oriented dynamic attention mechanism to selectively focus on salient patches of objects of interest. Neural networks are used to generate a set of high-level representation vectors for both the entire object and its segmented patches. Multi-view local neighboring structures between objects are encoded in the high-level object representation through an unsupervised pre-training procedure. By initializing the relation embeddings with object cluster centers, each relation modality can be reasonably interpreted as a semantic topic. A layer-wise training scheme based on a mixture of unsupervised and supervised training is proposed to improve generalization. The effectiveness of the proposed method and its superior performance compared against state-of-the-art algorithms are demonstrated through evaluations based on different image retrieval tasks.

Keywords: Multi-modal similarity, attention mechanism, representation learning, multi-view, neural network.

[1]School of Computer and Information, Hefei University of Technology, China.
[2]School of Computer Science, University of Manchester, Manchester, M1 7DN, United Kingdom.
[3]Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, United Kingdom.

1. Introduction

To learn a function that accurately calculates the similarity or relevance between objects is one of the most significant machine learning tasks, and is known as similarity learning. It is closely related to other fundamental machine learning paradigms, including clustering, ranking, classification and regression, and plays an important role in many real-world applications, such as image annotation and retrieval [48], intelligent recommendation systems [37] and knowledge graph completion [20]. Conventional similarity learning methods often learn a distance metric (e.g., Mahalanobis distance) [41] or use a kernel function [12] to measure the (dis)similarity between objects, where the metric (or kernel) formulation is adjusted by function parameters. These methods are mostly based on single modality. Although they are capable of measuring relevance in a standard environment, they may not be able to deal with tasks of more complex nature. For example, to retrieve images relevant to the query image of an apple fruit, images of apple juice (or the company Apple), which are related to the query in other relation types, can also be of interest. Therefore, this requires more sophisticated similarity learning models to encode multiple relation types.

Multi-modal similarity learning takes into account multiple types of relevance patterns between objects. For example, image relevance reflected by their shape and colour appearance. Multi-modal extensions have been developed for conventional similarity learning based on distance metrics and kernel functions. For instance, multiple kernel similarity learning [39] is proposed to facilitate image ranking, where the multiple modalities of image connections are realized by multiple kernel functions and the overall similarity is computed as a weighted sum of these functions. Transfer distance metric learning [24] is developed to overcome the lack of available information in the target task and discovers multiple alternative connections between objects in relevant source tasks. These correspond to multiple modalities characterized by different base metrics combined to form the final metric. In general, the intermediate results of these methods, such as the parameters or learned relation types, are hard to interpret and the whole learning procedure is usually treated as a black box. Intelligent similarity learning methods that exhibit not only excellent performance but also good model in-

terpretability are in demand.

To extract information from visual objects, primate visual systems employ attention mechanisms to dynamically focus on important information that is relevant to the current behavior or visual tasks [32]. Using the image retrieval task as an example, if the query is the image of a beach, the users could move their focus from the whole scene, to certain parts of the image, e.g., boat, people who swim, or sea. Recent advances in attention mechanisms use a set of dynamic attention weights to control the contribution of different parts [1]. Such techniques have been successful in tasks, such as machine translation [1] and image caption generation [42]. Taking the attention based image caption generation model [42] as an example, it works with high-level representations extracted from image patches using a convolutional neural network (CNN). The model learns the attention weights for each patch to construct a weighted context vector that represents relevant parts of an image based on which a long short-term memory (LSTM) network is used to generate text captions. Inspired by the recent success of attention learning in language and vision, we propose an interaction-oriented attention mechanism to improve the accuracy of similarity learning, and meanwhile show that the attention weights returned by the mechanism are able to improve the model interpretability.

In addition to multi-modal similarity analysis and attention mechanisms that can potentially improve the robustness and interpretability of a learning model, it is also important to improve the model performance. When dealing with complex real-world tasks, features that exhibit heterogeneous properties should be considered. For example, in image retrieval, shape feature is more important for measuring similarity between a brown bear and a polar bear, while the color feature is more important for examining brown bears in different poses. One representative work that deals with this problem is [6], which leverages shared knowledge from multiple related tasks to improve the performance of feature selection. Another commonly used technique for combining multi-view information is multi-view embedding. It aims at mixing and refining information provided by different types of features within a low-dimensional embedding space [28, 36, 43, 50]. Recent developments on multi-view learning have shown that complementary information across different views has the potential of im-

proving the performance of many machine learning tasks [10]. To further improve similarity learning, we take into account multi-view local structures in similarity formulation.

In summary, this work proposes a powerful similarity measure by exploring multiple hidden relationships between image objects that suit the multi-modal nature of real-world tasks. To improve model robustness and interpretability, dynamic attentions are incorporated to selectively capture salient parts that contribute to the object interactions. To deal with heterogeneous object properties, we encode multi-view information that improves the object representation. These result in proposing a novel attention-driven multi-modal similarity (AMoS) model possessing a multi-layered architecture. Neural networks are used to compute representation vectors of a given object and its corresponding patches. Different relation modalities are encoded as different hidden neurons in the relation layer. Dynamic attention weights are modeled as functions receiving the entire image for their patch representation, and multi-view information provided by different feature extraction methods are used to enhance the image representation in pre-training. The effectiveness of the proposed model is compared with various state-of-the-art methods evaluated through image retrieval tasks. The remaining paper is organized as follows. Section 2 briefly introduces related works. Section 3 delivers the proposed algorithm, while Section 4 contains experimental results and comparative analyses. Finally, Section 5 concludes the work.

2. Related Works

2.1. Multi-modal Similarity Learning

Multi-modal similarity learning is a type of learning that relies on measuring the relevance between objects from multiple aspects. It has been shown to be effective in many real-world applications. One example is person identification over camera networks, using multiple Mahalanobis distance metrics designed to characterize different cameras that contain different types of noise [26]. These metrics are connected by enforcing joint regularization that reduce overfitting. Another work mines complementary information among features that exhibit heterogeneous properties by optimizing

different distance metrics in different feature spaces [38] . To facilitate tasks such as inter-modal label transfer and zero-shot learning, multi-modal models are developed to formulate the relations between text and image features [31]. In social media network analysis, a latent semantic space is computed to encode multi-modal links, e.g., context and content links between the multimedia and context objects [30]. Another example work in data retrieval [19], develops multi-modal algorithms to achieve cross-modal hashing. For instance, linear subspace ranking hashing maps data from different modalities into a common binary space, so that the cross-modal similarity can be measured using Hamming distance, where different modalities are modeled as groups of linear subspaces. Multi-modal deep hashing [21] has also been proposed, where the data modalities are encoded by multiple hierarchical nonlinear transformations and constraints are incorporated at the top layer of the network to exploit nonlinear relations between samples. However, its intermediate results, e.g., the meaning of the learned relation types and their controlling parameters, are difficult to interpret.

2.2. Attention Learning

Attention learning constitutes a recent kind of machine learning, where the core idea is to assign different weights to different components or parts of an object according to different requirements in learning. The technique has been successful in neural machine translation [1], which can automatically translate sentences to a target language by first encoding the source sentence into a fixed-length vector and then generating a translation from the vector through a decoder. Unlike conventional translation algorithms, the attention based neural translation extends the encoder-decoder model by introducing a dynamic context vector to focus on salient information that is relevant to the generation of the next target word. This yields good results in translating long sentences [1]. Another attention based translation work [25] explores better the architecture of attention mechanisms. It proposes two simple and effective mechanisms, one of which is a global approach that takes into account all source positions and the other a local approach that pays attention to just a subset of source positions at a time. Attention mechanisms have also been successful in speech recognition. To deal with long and noisy speech input, an attention learning model is developed to combine both con-

tent and location information, so that the most relevant position in a sequence can be selected for further decoding [9]. Another application is video description generation [45]. Unlike images, video description requires the consideration of dynamic temporal structure to produce descriptions. The temporal attention mechanism developed in [45] selectively focuses on a small set of salient frames and lets the generator describe only objects and activities in this set. This mechanism not only improves the quality of the generated descriptions, but also effectively eliminates redundant information through the use of salient frames. To improve image segmentation, the work in [8] proposes an attention mechanism that softly weights the multi-scale features at each pixel location. For image captioning, the work in [46] learns to selectively attend to a semantically important concept (or a region of interest in an image) by weighting the relative attention strengths. Additionally, the proposed attention mechanism is able to dynamically switch attention between concepts according to task status.

2.3. Representation Learning

Representation learning refines the input raw data by highlighting useful information and eliminating redundant information and noise. It is one of the most important techniques in computer vision and multimedia [5, 7], and so far deep learning is the most successful representation learning technique [16, 3, 34]. One of the most commonly used deep representation learning methods is the convolutional neural network (CNN) [17], which is widely used [29, 13, 44, 18]. For instance, a CNN-based mapping function is learned to transfer mid-level representations obtained from large source datasets to a target image recognition task with limited training data in [29]. Robust unsupervised representation is computed by first automatically generating surrogate tasks through data augmentation and then training a CNN-based classifier over these tasks in [13]. In video surveillance, it is assumed that two video patches connected by the same track should share similar representations in a high-level space. An unsupervised CNN is trained to draw frames from the same track to be closer to each other than to random frames from other tracks [44]. To compute visual saliency from multi-scale features, the input image is first decomposed into non-overlapping regions, then features are extracted from these regions by a CNN and fed into a fully connected neural network to

generate saliency map [18]. Apart from CNN, auto-encoder is another commonly used algorithm for unsupervised representation learning [3]. In speech emotion recognition, different scales of kernels are learned by auto-encoder to extract local features from entire spectrogram fragments [27]. To develop 3D shape feature descriptors, the Fisher criterion is employed as an extra constraint added to the conventional auto-encoder, so that the learned hidden features can be discriminative and insensitive to geometric structure variations [40].

3. Proposed Method

The information carried by a color image is stored as matrices of pixel values each corresponding to a color (e.g., R, G, B channel). The semantic content of an image is characterized by its high-level representation learned from the entire image using, for instance, a CNN [3]. In addition to collecting information from the entire image, the human visual system is able to pay attention to different parts of the image under different circumstances. To simulate such attention function, models have been developed to partition an image into different patches and dynamically allow each patch to come to the forefront as needed [42]. Motivated by these, given a collection of n images $\{x_i\}_{i=1}^n$, we represent each image using not only its k-dimensional high-level representation vector computed from the entire image by a CNN network, denoted as $\boldsymbol{\phi}_i = \boldsymbol{\phi}(x_i, \boldsymbol{w})$, but also a set of k-dimensional patch representation vectors $\{\boldsymbol{p}_i^{(j)}\}_{j=1}^d$. These patch vectors are computed by first partitioning the entire image to a set of d image patches[4] and then converting each patch $x_i^{(j)}$ to a k-dimensional vector by another neural network, denoted as $\boldsymbol{p}_i^{(j)} = \boldsymbol{\phi}_p(x_i^{(j)}, \boldsymbol{w}_p)$. The vector $\boldsymbol{w}$ and $\boldsymbol{w}_p$ store the network weights for computing the entire image representation and the patch representation, respectively.

[4]One way to obtain image patches is to apply the k-means clustering algorithm to group the pixels of a given image, so that the image is segmented according to its color (or grayscale) distribution. The cluster centers can be used as the central points of different image patches and each patch can be set to contain for example $l \times l$ pixels. Both the cluster number d and the patch size l can be adjusted by the user. The obtained patches are denoted by $\{x_i^{(j)}\}_{j=1}^d$ for the i-th image.

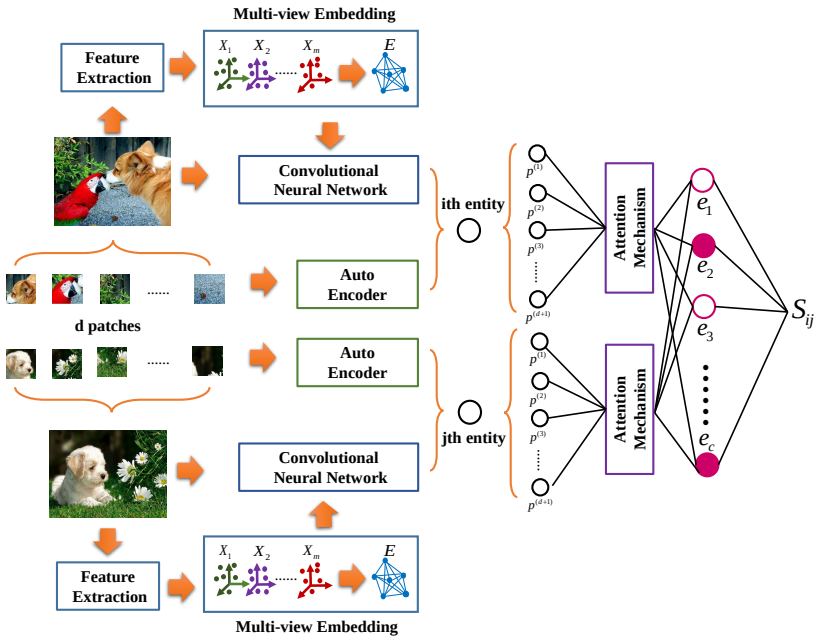

Figure 1: Overall architecture of the proposed similarity learning model AMoS.

We propose a robust similarity model to evaluate the relevance between each i-th and j-th image, based not only on their entire image representations but also their patch representations. This is defined as

$$S_{ij} = f\left(\boldsymbol{\phi}(x_i, \boldsymbol{w}), \left\{\boldsymbol{\phi}_p(x_i^{(s)}, \boldsymbol{w}_p)\right\}_{s=1}^{d}, \boldsymbol{\phi}(x_j, \boldsymbol{w}), \left\{\boldsymbol{\phi}_p(x_j^{(h)}, \boldsymbol{w}_p)\right\}_{h=1}^{d}, \boldsymbol{\theta}\right), \quad (1)$$

where $\boldsymbol{\theta}$ stores the collection of similarity parameters to be learned. The architecture of the proposed model is illustrated in Figure 1. Other than working with a static image representation and simulating single-modal relations as most existing works do, the proposed model dynamically pays attention to interactions between different image parts (entire image and image patches) and accentuate the influence of individual parts when computing similarities over different relation modalities. The robustness of the proposed similarity is supported by its unique way of encoding multi-modal relations and dynamic attention. In the following, we explain the proposed attention-driven multi-modal similarity model, referred to as AMoS, in detail.

3.1. Model Construction

3.1.1. Multi-modal Similarity

We formulate the similarity measure so that it reflects the validities of multiple hidden relations (relation multi-modality). In image retrieval, the query image can be related to the searched imagery under different relation types, e.g., the query of an apple image of fruit can be related to images of apple juice or the company's logo. In order to measure similarity under different aspects of relations, we employ a set of hidden neurons, each representing a hidden relationship between the two input objects. Letting c denote the neuron cardinality, a c-dimensional multi-modal similarity vector $\boldsymbol{s}_{ij} = [s_{ij}^{(1)}, \ldots, s_{ij}^{(c)}]$ is learned.

We first consider the formulation of each similarity dimension based only on the image representation vector $\boldsymbol{\phi}_i$. This leads to

$$s_{ij}^{(t)} = \left(\boldsymbol{\phi}_i^T \boldsymbol{e}_t\right)\left(\boldsymbol{\phi}_j^T \boldsymbol{e}_t\right) + \boldsymbol{\alpha}^T \boldsymbol{\phi}_i + \boldsymbol{\beta}^T \boldsymbol{\phi}_j, \tag{2}$$

for $t = 1, \ldots, c$, where $\boldsymbol{e}_t$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are column vectors of the same dimensionality as $\boldsymbol{\phi}_i$. The relation embedding vector $\boldsymbol{e}_t$ is used to parameterize the unique character of the relation type t. The operation $\left(\boldsymbol{\phi}_i^T \boldsymbol{e}_t\right)\left(\boldsymbol{\phi}_j^T \boldsymbol{e}_t\right)$ constructs a bilinear interaction of the image pair that is $\boldsymbol{\phi}_i^T \mathbf{W}_t \boldsymbol{\phi}_j$ based on the rank-1 interaction matrix $\mathbf{W}_t = \boldsymbol{e}_t \boldsymbol{e}_t^T$, and the use of vector $\boldsymbol{e}_t$ instead of an arbitrary matrix $\mathbf{W}_t$ reduces the number of variables used to parameterize the bilinear mixing. The linear weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ incorporate the properties of individual images to further enrich the similarity formulation. To compute the final similarity, the following accumulated scalar score is used

$$S_{ij} = \frac{1}{c} \sum_{t=1}^{c} \text{sig}\left(s_{ij}^{(t)}\right), \tag{3}$$

where the sigmoid function $\text{sig}(\cdot)$ acts as an activation function, rescaling the similarity value within the range of $[0, 1]$.

3.1.2. Attention Incorporation

Building upon Eq. (2), we further incorporate attention into the model. Our core idea is that, instead of characterizing interactions between images with static representations, attention can be modeled via dynamic weights to allow different image parts to selectively contribute to different relation modalities given different input image pairs.

Firstly, we formulate a similarity model that pays equal attention to different image parts and the entire image based on Eq. (2). This can be achieved by simply accumulating similarities between image patches and adding the accumulated similarity to the previously computed similarity score $s_{ij}^{(t)}$. Letting the $d \times k$ matrix $\mathbf{Z}_i = [\boldsymbol{p}_i^{(1)}, \ldots, \boldsymbol{p}_i^{(d)}]^T$ denote the local context feature matrix for the i-th image, the modified model becomes

$$\begin{aligned}\tilde{s}_{ij}^{(t)} &= \left|s_{ij}^{(t)}\right| + \left\|\mathbf{Z}_i \boldsymbol{e}_t \boldsymbol{e}_t^T \mathbf{Z}_j^T\right\|_1 + \|\mathbf{Z}_i \boldsymbol{\alpha}\|_1 + \|\mathbf{Z}_j \boldsymbol{\beta}\|_1, \\ &= \left|s_{ij}^{(t)}\right| + \sum_{s=1}^{d}\sum_{h=1}^{d}\left|\boldsymbol{p}_i^{(s)T} \boldsymbol{e}_t \boldsymbol{e}_t^T \boldsymbol{p}_j^{(h)}\right| + \sum_{s=1}^{d}\left|\boldsymbol{\alpha}^T \boldsymbol{p}_i^{(s)}\right| + \sum_{h=1}^{d}\left|\boldsymbol{\beta}^T \boldsymbol{p}_j^{(h)}\right|, \end{aligned} \tag{4}$$

where $\|\cdot\|_1 = \|\text{vec}(\cdot)\|$ is the entrywise p-norm in the case of $p = 1$.

To identify the salient features that contribute more to the similarity score under a relation modality, we add dynamic attention weights to Eq. (4). For an input image pair, these attention weights distinguish contributions of their different image parts to the target relation type t. Introducing the new notation $\boldsymbol{p}_i^{(d+1)} = \boldsymbol{\phi}_i$ to treat the entire image as the $(d+1)$-th patch, we construct the following attention-driven similarity function over the t-th relation modality:

$$\tilde{s}_{ij}^{(t)} = \sum_{s=1}^{d+1}\sum_{h=1}^{d+1} e_{ij}^{(t)}(s,h)\left|\boldsymbol{p}_i^{(s)T} \boldsymbol{e}_t \boldsymbol{e}_t^T \boldsymbol{p}_j^{(h)}\right| + \sum_{s=1}^{d+1} a_i^{(t)}(s)\left|\boldsymbol{\alpha}^T \boldsymbol{p}_i^{(s)}\right| + \sum_{h=1}^{d+1} a_j^{(t)}(h)\left|\boldsymbol{\beta}^T \boldsymbol{p}_j^{(h)}\right|. \tag{5}$$

The non-negative attention weights $\{e_{ij}^{(t)}(s,h)\}_{s,h=1}^{d+1}$ control the degree of importance of image patch pairs, and the non-negative attention weights $\{a_i^{(t)}(s)\}_{s=1}^{d+1}$ control the importance of image patches that contribute to the similarity computation. These weights are dynamically renewed given different input images, guided by the following weighting functions:

$$e_{ij}^{(t)}(s,h) = \frac{\exp\left(\gamma_{ij}^{(t)}(s,h)\right)}{\sum_{s=1}^{d+1}\sum_{h=1}^{d+1}\exp\left(\gamma_{ij}^{(t)}(s,h)\right)}, \tag{6}$$

$$a_i^{(t)}(s) = \frac{\exp\left(\mu_i^{(t)}(s)\right)}{\sum_{s=1}^{d+1}\exp\left(\mu_i^{(t)}(s)\right)}, \tag{7}$$

where

$$\gamma_{ij}^{(t)}(s,h) = \text{sig}\left(\boldsymbol{\eta}_t^T \boldsymbol{p}_i^{(s)} + \boldsymbol{\eta}_t^T \boldsymbol{p}_j^{(h)}\right), \tag{8}$$

$$\mu_i^{(t)}(s) = \text{sig}\left(\boldsymbol{\tau}_t^T \boldsymbol{p}_i^{(s)}\right). \tag{9}$$

The column vectors $\boldsymbol{\eta}_t$ and $\boldsymbol{\tau}_t$ store the attention parameters and contribute to the t-th relation type, and they possess the same dimensionality k as the high-level representation of an image patch. In Eqs. (6,7), the softmax function is used to generate positive attention weights that sum to 1. Finally, by restricting each weighted similarity quantity in Eq. (5) within $[0, 1]$ through the smooth sigmoid function, the following attention-driven similarity function is proposed, given as

$$\begin{aligned}\tilde{s}_{ij}^{(t)} =& \sum_{s=1}^{d+1}\sum_{h=1}^{d+1} \text{sig}\left(e_{ij}^{(t)}(s,h)\boldsymbol{p}_i^{(s)T}\boldsymbol{e}_t\boldsymbol{e}_t^T\boldsymbol{p}_j^{(h)}\right) + \sum_{s=1}^{d+1}\text{sig}\left(a_i^{(t)}(s)\boldsymbol{\alpha}^T\boldsymbol{p}_i^{(s)}\right) \\ &+ \sum_{h=1}^{d+1}\text{sig}\left(a_j^{(t)}(h)\boldsymbol{\beta}^T\boldsymbol{p}_j^{(h)}\right).\end{aligned} \tag{10}$$

3.2. Modal Training

Taking the computed attention-driven similarities $\left\{\tilde{s}_{ij}^{(t)}\right\}_{t=1}^{c}$ over different relation modalities as input, we obtain the final similarity score between two images by $S_{ij} = \sum_{t=1}^{c} \tilde{s}_{ij}^{(t)}$. Model variables that participate in the similarity computation include weights $\boldsymbol{w}$ and $\boldsymbol{w}_p$ of the two neural networks for processing the entire image and image patches, and those involved in the relation formulation $\boldsymbol{\theta} = \{\{\boldsymbol{e}_t, \boldsymbol{\eta}_t, \boldsymbol{\tau}_t\}_{t=1}^{c}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \}$. The parameter vector $\boldsymbol{\theta}$ includes the relation embeddings $\{\boldsymbol{e}_t\}_{t=1}^{c}$ and the weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for formulating the multi-modal relations, as well as $\{\boldsymbol{\eta}_t, \boldsymbol{\tau}_t\}_{t=1}^{c}$ for computing the dynamic attention weights under different relation modalities. The proposed model contains a substantial amount of variables to be optimized. The optimization objective function is non-convex and it is difficult to obtain a good local optimal solution with the traditional training approach. We thus adapt a stage-wise training scheme that first performs the representation training and the relation training separately, and then fine-tunes all the pre-trained parameters. This aims to obtain a more robust model solution with improved generalization ability.

3.2.1. Unsupervised Pre-training of Image and Patch Representation

The image representation is computed by a CNN receiving the entire image as input. Its network weights $\boldsymbol{w}$ are first pre-trained in an unsupervised manner. We set the pre-training objective function as a penalized distance error sum between images, given by

$$\min_{\boldsymbol{w}} \; O = \sum_{ij} \text{sig}\left(\sigma_{ij}^{(\boldsymbol{w})} \left\|\boldsymbol{\phi}(x_i, \boldsymbol{w}) - \boldsymbol{\phi}(x_j, \boldsymbol{w})\right\|_2^2\right), \tag{11}$$

where $\|\cdot\|_2$ denotes the l_2-norm and the sigmoid function is used to smoothen and bound the error. The weight $\sigma_{ij}^{(\boldsymbol{w})}$ quantifies the similarity and neighboring information between two images. By minimizing Eq. (11), the CNN weights $\boldsymbol{w}$ support the generation of high-level features that preserve a desired neighborhood structure captured by $\left\{\sigma_{ij}^{(\boldsymbol{w})}\right\}_{i,j=1}^{n}$.

To compute the neighborhood weights $\left\{\sigma_{ij}^{(\boldsymbol{w})}\right\}_{i,j=1}^{n}$ so that they reflect more accurate proximity structure between images, we utilize multi-view information offered by different feature extraction methods, such as color histogram, color correlogram, edge direction histogram, wavelet texture, block-wise colour moments and bag of words based on the scale-invariant feature transform (SIFT) descriptions. Letting $\boldsymbol{\Omega}_s = \left[\omega_{ij}^{(s)}\right]$ denote the local proximity matrix computed based on the sth feature view, its nonzero elements indicate the similarities between neighboring objects computed under the sth feature view, whereas its zero elements indicate the non-neighboring pairs. The penalty weights $\left\{\sigma_{ij}^{(\boldsymbol{w})}\right\}_{i,j=1}^{n}$ are computed as the averaged local proximities such that $\sigma_{ij}^{(\boldsymbol{w})} = \frac{1}{m}\sum_{s=1}^{m}\omega_{ij}^{(s)}$, where m denotes the total number of used feature views. This assumes that when there are more views agreeing on the neighborhood relation between two objects, then that object pair is considered to be more reliable and it is awarded a higher weight [10]. This is because these averaged weights can be written as $\sigma_{ij}^{(\boldsymbol{w})} = \frac{m_a}{m}\bar{\sigma}_{ij}^{(\boldsymbol{w})}(m_a)$, where

$$\bar{\sigma}_{ij}^{(\boldsymbol{w})}(m_a) = \begin{cases} 0, & \text{if } I_{ij}^{(m_a)} = \emptyset, \\ \sum_{s \in I_{ij}^{(m_a)}} \frac{1}{m_a}\omega_{ij}^{(s)}, & \text{otherwise.} \end{cases} \tag{12}$$

Here, m_a denotes the number of the views in agreement and the set $I_{ij}^{(m_a)}$ records the indices of the m_a views agreeing that the ith and the jth objects are neighbors through

proximity comparison using the features of the corresponding views. In this way, $\frac{m_a}{m}$ can be viewed as the confidence degree for weighting the voted similarity $\bar{\sigma}_{ij}^{(\boldsymbol{w})}(m_a)$.

Let the i-th row of the $n \times k$ matrix $\mathbf{\Phi}$ store the k-dimensional high-level feature vector $\boldsymbol{\phi}(x_i, \boldsymbol{w})$, the ij-th element of the $n \times n$ matrix $\mathbf{W}$ store the penalty weight $\sigma_{ij}^{(\boldsymbol{w})}$, and the diagonal elements of the $n \times n$ diagonal matrix $\mathbf{D}(\mathbf{W})$ store the row sum of $\mathbf{W}$. We compute the Laplacian matrix of $\mathbf{W}$ as $\mathbf{L} = \mathbf{D}(\mathbf{W}) - \mathbf{W}$ [2]. The derivate of the CNN output is then given by

$$\frac{\partial O}{\partial \mathbf{\Phi}} = \gamma \mathbf{\Phi} \left(\mathbf{L} + \mathbf{L}^T \right), \tag{13}$$

where

$$\gamma = \text{sig}\left(\text{tr}\left(\mathbf{\Phi}\mathbf{L}\mathbf{\Phi}^T\right)\right)\left[1 - \text{sig}\left(\text{tr}\left(\mathbf{\Phi}\mathbf{L}\mathbf{\Phi}^T\right)\right)\right].$$

The CNN weights are updated through backpropagation and gradient descent.

Different from the entire image, image patches are segmented parts and contain less rich local structural information. Instead of using convolutional kernels, we compute the patch representation using a fully connected neural network, and pre-train the network weights $\boldsymbol{w}_p$ in an unsupervised manner by following the auto-encoder training scheme. Specifically, the network weights $\boldsymbol{w}_p$ are computed by minimizing the reconstruction error between the input patch $x_i^{(s)}$ and the decoded output computed from the patch representation $\boldsymbol{p}_i^{(s)}$ [35].

3.2.2. Supervised Training of Relation Parameters

By fixing the pre-trained weights $\boldsymbol{w}$ and $\boldsymbol{w}_p$, we further optimize the relation embeddings $\{\boldsymbol{e}_t\}_{t=1}^c$, attention parameters $\{\boldsymbol{\eta}_t, \boldsymbol{\tau}_t\}_{t=1}^c$ and linear weights $\boldsymbol{\alpha}, \boldsymbol{\beta}$ in a supervised manner. The following ranking loss based on the stochastic margin error [4] is minimized

$$L(\{\boldsymbol{e}_t, \boldsymbol{\eta}_t, \boldsymbol{\tau}_t\}_{t=1}^c, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i, j_+ \in I_+} \sum_{i, j_- \in I_-} \max(S_{ij_-} - S_{ij_+} + 1, 0), \tag{14}$$

where the index set I_+ contains the truly related object pairs in the training set referred to as the positive training pairs, while I_- the truly unrelated object pairs referred to as the negative training pairs. The cost function evaluates the difference between the

similarity scores of the positive and negative pairs. The optimization drives the positive pairs to have higher similarity scores than the negative pairs.

To optimize the relation embeddings $\{\boldsymbol{e}_t\}_{t=1}^{c}$, we propose to initialize each $\boldsymbol{e}_t$ with the center vector $\boldsymbol{c}_t$ of an image cluster instead of random initialization. For an image retrieval task, these image clusters can be manually defined according to the training data. For instance, a collection of real animal (or toy, or logo) bear images can be defined as a cluster. In this case, relevance between a bear animal image ($\boldsymbol{\phi}_i$) and a bear logo image ($\boldsymbol{\phi}_j$) is increased over the two neurons initialized by the bear animal cluster center ($\boldsymbol{e}_1$) and the bear logo cluster center ($\boldsymbol{e}_2$) because of the increased values of $\boldsymbol{\phi}_i^T\boldsymbol{e}_1$ and $\boldsymbol{\phi}_j^T\boldsymbol{e}_2$. Another way to obtain the image clusters is to perform k-means clustering over the training images based on their high-level image representations $\{\boldsymbol{\phi}_i\}_{i=1}^{n}$. This procedure automatically discovers c different clusters representing the semantic topic structure of the images, and uses these topics to drive different relation modalities when formulating the similarity measure.

To minimize Eq. (14), the stochastic gradient descent algorithm is used. The three quantities of $D_1^{(i,j)}(s,h)$, $D_2^{(i)}(s)$ and $D_3^{(j)}(h)$ are used to simplify the gradient formulas, given as:

$$D_1^{(i,j)}(s,h) = \text{sig}\left(e_{ij}^{(t)}(s,h)\boldsymbol{p}_i^{(s)T}\boldsymbol{e}_t\boldsymbol{e}_t^T\boldsymbol{p}_j^{(h)}\right)\left(1-\text{sig}\left(e_{ij}^{(t)}(s,h)\boldsymbol{p}_i^{(s)T}\boldsymbol{e}_t\boldsymbol{e}_t^T\boldsymbol{p}_j^{(h)}\right)\right), \tag{15}$$

$$D_2^{(i)}(s) = \text{sig}\left(a_i^{(t)}(s)\boldsymbol{\alpha}^T\boldsymbol{p}_i^{(s)}\right)\left(1-\text{sig}\left(a_i^{(t)}(s)\boldsymbol{\alpha}^T\boldsymbol{p}_i^{(s)}\right)\right), \tag{16}$$

$$D_3^{(j)}(h) = \text{sig}\left(a_j^{(t)}(h)\boldsymbol{\beta}^T\boldsymbol{p}_j^{(h)}\right)\left(1-\text{sig}\left(a_j^{(t)}(h)\boldsymbol{\beta}^T\boldsymbol{p}_j^{(h)}\right)\right). \tag{17}$$

In the following, we list the related differentiations with respect to the relation param-

eters:

$$\frac{\partial \tilde{s}_{ij}^{(t)}}{\partial \boldsymbol{\alpha}} = \sum_{s=1}^{d+1} D_2^{(i)}(s) a_i^{(t)}(s) \boldsymbol{p}_i^{(s)}, \tag{18}$$

$$\frac{\partial \tilde{s}_{ij}^{(t)}}{\partial \boldsymbol{\beta}} = \sum_{h=1}^{d+1} D_3^{(j)}(h) a_j^{(t)}(h) \boldsymbol{p}_j^{(h)}, \tag{19}$$

$$\frac{\partial \tilde{s}_{ij}^{(t)}}{\partial \boldsymbol{e}_t} = \sum_{s=1}^{d+1} \sum_{h=1}^{d+1} D_1^{(i,j)}(s,h) e_{ij}^{(t)}(s,h) \left(\boldsymbol{p}_i^{(s)} \left(\boldsymbol{e}_t^T \boldsymbol{p}_j^{(h)} \right) + \left(\boldsymbol{p}_i^{(s)T} \boldsymbol{e}_t \right) \boldsymbol{p}_j^{(h)} \right), \tag{20}$$

$$\frac{\partial \tilde{s}_{ij}^{(t)}}{\partial \boldsymbol{\eta}_t} = \frac{\partial \tilde{s}_{ij}^{(t)}}{\partial e_{ij}^{(t)}(s,h)} \frac{de_{ij}^{(t)}(s,h)}{d\gamma_{ij}^{(t)}(s,h)} \frac{d\gamma_{ij}^{(t)}(s,h)}{d\boldsymbol{\eta}_t}, \tag{21}$$

$$\frac{\partial \tilde{s}_{ij}^{(t)}}{\partial e_{ij}^{(t)}(s,h)} = D_1^{(i,j)}(s,h) \boldsymbol{p}_i^{(s)T} \boldsymbol{e}_t \boldsymbol{e}_t^T \boldsymbol{p}_j^{(h)}, \tag{22}$$

$$\frac{de_{ij}^{(t)}(s,h)}{d\gamma_{ij}^{(t)}(s,h)} = e_{ij}^{(t)}(s,h) \left(1 - e_{ij}^{(t)}(s,h) \right), \tag{23}$$

$$\frac{de_{ij}^{(q)}(s,h)}{d\gamma_{ij}^{(t)}(s,h)} = -e_{ij}^{(q)}(s,h) e_{ij}^{(t)}(s,h), \text{for } q \neq t, \tag{24}$$

$$\frac{d\gamma_{ij}^{(t)}(s,h)}{d\boldsymbol{\eta}_t} = \left(\boldsymbol{p}_i^{(s)} + \boldsymbol{p}_j^{(h)} \right) \gamma_{ij}^{(t)}(s,h) \left(1 - \gamma_{ij}^{(t)}(s,h) \right), \tag{25}$$

$$\frac{\partial \tilde{s}_{ij}^{(t)}}{\partial \boldsymbol{\tau}_t} = \frac{\partial \tilde{s}_{ij}^{(t)}}{\partial a_i^{(t)}(s)} \frac{da_i^{(t)}(s)}{d\mu_i^{(t)}(s)} \frac{d\mu_i^{(t)}(s)}{d\boldsymbol{\tau}_t} + \frac{\partial \tilde{s}_{ij}^{(t)}}{\partial a_j^{(t)}(h)} \frac{da_j^{(t)}(h)}{d\mu_j^{(t)}(h)} \frac{d\mu_j^{(t)}(h)}{d\boldsymbol{\tau}_t}, \tag{26}$$

$$\frac{\partial \tilde{s}_{ij}^{(t)}}{\partial a_i^{(t)}(s)} = D_2^{(i)}(s) \boldsymbol{\alpha}^T \boldsymbol{p}_i^{(s)}, \tag{27}$$

$$\frac{\partial \tilde{s}_{ij}^{(t)}}{\partial a_j^{(t)}(h)} = D_3^{(j)}(h) \boldsymbol{\beta}^T \boldsymbol{p}_j^{(h)}, \tag{28}$$

$$\frac{da_i^{(t)}(s)}{d\mu_i^{(t)}(s)} = a_i^{(t)}(s) \left(1 - a_i^{(t)}(s) \right), \tag{29}$$

$$\frac{da_i^{(q)}(s)}{d\mu_i^{(t)}(s)} = -a_i^{(q)}(s) a_i^{(t)}(s), \text{for } q \neq t, \tag{30}$$

$$\frac{da_j^{(t)}(h)}{d\mu_j^{(t)}(h)} = a_j^{(t)}(h) \left(1 - a_j^{(t)}(h) \right), \tag{31}$$

$$\frac{da_j^{(q)}(h)}{d\mu_j^{(t)}(h)} = -a_j^{(q)}(h) a_j^{(t)}(h), \text{for } q \neq t, \tag{32}$$

$$\frac{d\mu_i^{(t)}(s)}{d\boldsymbol{\tau}_t} = \boldsymbol{p}_i^{(s)} \mu_i^{(t)}(s) \left(1 - \mu_i^{(t)}(s) \right), \tag{33}$$

$$\frac{d\mu_j^{(t)}(h)}{d\boldsymbol{\tau}_t} = \boldsymbol{p}_j^{(h)} \mu_j^{(t)}(h) \left(1 - \mu_j^{(t)}(h) \right). \tag{34}$$

3.2.3. Supervised Model Fine-tuning

So far, we first perform unsupervised pre-training of the network weights $\boldsymbol{w}$ and $\boldsymbol{w}_p$ to generate high-level features for entire images and image patches, respectively. After that, by fixing these weights, we continue to perform supervised training of the relation parameters. This divides the model into two independent networks: (1) unsupervised representation learning and (2) supervised relational learning. In order to connect these two components and seek a better solution for the entire model, a supervised fine-tuning procedure that optimizes all the model variables based on the ranking loss in Eq. (14) is carried out. Gradient formulations for updating $\boldsymbol{w}$ and $\boldsymbol{w}_p$ based on the ranking loss are realized as

$$\frac{\partial L}{\partial \boldsymbol{w}} = \frac{\partial L}{\partial S_{ij}} \frac{\partial S_{ij}}{\partial \tilde{s}_{ij}^{(t)}} \frac{\partial \tilde{s}_{ij}^{(t)}}{\partial \boldsymbol{\Phi}} \frac{d\boldsymbol{\Phi}}{d\boldsymbol{w}}, \tag{35}$$

$$\frac{\partial L}{\partial \boldsymbol{w}_p} = \frac{\partial L}{\partial S_{ij}} \frac{\partial S_{ij}}{\partial \tilde{s}_{ij}^{(t)}} \frac{\partial \tilde{s}_{ij}^{(t)}}{\partial \boldsymbol{\Phi}} \frac{d\boldsymbol{\Phi}}{d\boldsymbol{w}_p}. \tag{36}$$

4. Experimental Analysis and Results

4.1. Datasets and Experimental Settings

The proposed work is evaluated and compared with state-of-the-art methods using four challenging image datasets of CIFAR-10[5], NUS-WIDE[6], Places2[7] and ImageNet[8] for image retrieval tasks. In more detail, CIFAR-10 is a large collection of color images collected from Flicker, containing 60,000 images from to 10 object classes, such as airplane, truck, bird, cat, deer, horse, etc. We randomly select 1,000 images per class as query images, 1000 as training images, and the remaining ones as testing images. NUS-WIDE is a larger collection of Flicker web images containing 269,648 images belonging to 81 concepts, such as garden, street, tower, dancing, tree, etc. We randomly select 2,000 images as queries, 8,000 as training images and the remaining as the testing

[5]The CIFAR-10 dataset is available on: https://www.cs.toronto.edu/ kriz/cifar.html

[6]The NUS-WIDE dataset is available on: http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

[7]The Places2 dataset is available on: http://places2.csail.mit.edu

[8]The ImageNet dataset is available on: http://www.image-net.org/

ones. Places2 is a large image collection containing more than 10 million images belonging to over 400 unique scene categories, such as underwater, park, museum, mountain, etc. We randomly select 500 images per class as query, 1,000 images per class for training, and 1,500 images per class for testing, and report the performance for 10 randomly chosen classes. ImageNet is a large image dataset that includes 14 million images organized according to WordNet hierarchy and representing concepts, such as dog, cat, swimming, vehicle, etc. We randomly select 200 images per concept as query, 300 images per concept for training and 1,000 images per concept for testing, and report the performance for 10 randomly chosen concepts.

For the proposed method AMoS, to compute the multi-view neighboring weights of $\left\{\sigma_{ij}^{(\boldsymbol{w})}\right\}_{i,j=1}^{n}$, different feature extraction methods are used. These include 900-D local binary pattern (LBP), 256-D color histogram (CH), 324-D histogram of gradient (HoG) and 1024-D wavelet texture (WT) features for CIFAR-10, Places2 and ImageNet. Specifically, to generate the LBP features, the center pixel is compared with its 8 neighboring pixels in each cell. To generate the CH features, RGB images are first transformed to HSV images, and then 16 bins for the hue space, 4 bins for the saturation space and 4 bins for the value space are used, resulting in features of pixel counts in $16 \times 4 \times 4 = 256$ bins. To generate the HoG features, the input image is divided into 8 cells, and then gradients from 36 angles within 180 degrees are computed in each cell with 9 bins used for each angle. This results in feature of gradient counts in $9 \times 36 = 324$ bins. To generate the WT features, the 20×20 Gaussian filter is used on the input image. For NUS-WIDE, six types of low-level features are readily provided by the dataset, including 64-D CH,144-D color correlogram (CORR) , 73-D edge direction histogram (EDH), 128-D WT, 225-D block-wise color moments (CM) and the 500-D bag-of-word model based on SIFT descriptions.

To compute the high-level representation for entire images, a CNN network with an architecture of C-S-C-S-F is used, where C stands for convolutional layer, S for subsampling layer, and F for fully connected layer. A total of 20 convolutional kernels of size 5×5 are used in each of the two convolutional layers. To compute the image patch representation, its auto-encoder training includes three layers (input, encode, decode) with 300, 100 and 300 neurons. To implement the attention scheme, we extract

	(a) High-level representation (✓) (1) Concatenated multi-view feature (×) (2) Caffe features (□)	(b) Multi-modal relation mechanism (✓) (3) Single-modal relation via LSH [14] (×)	(c) Attention mechanism (✓) (4) No attention (×)	(d) Fine-tuning (✓) (5) No fine-tuning (×)
Setting 1	×	×	×	×
Setting 2	✓	×	×	×
Setting 3	✓	✓	×	×
Setting 4	✓	✓	✓	×
Setting 5	□	✓	✓	✓
Setting 6	✓	✓	✓	✓

Table 1: List of settings alternative to the proposed model design.

$d = 1, 2, 3, 4$ patches for each input image and the size of each patch is set as 11×11. For the multi-modal scheme, we initialize the relation embeddings using cluster centers estimated by k-means clustering, for which the setting of $c = 5, 10, 15, 20$ is experimented with. To optimize the network, batch based stochastic gradient descent is used with the setting of Batchsize $= 15$ and Learning rate $= 0.1$. The precision of top 500 retrieved images (500AP) and mean average precision (mAP) are used to assess the retrieval performance.

4.2. Empirical Analysis of AMoS

The robustness of the proposed AMoS model is supported by the four design components: (a) high-level image and patch representation, (b) multi-modal relation mechanism, (c) dynamic attention mechanism, and (d) supervised model fine tuning. To illustrate the importance and consequent necessity of each design, we compare various alternative design options: (1) Replacing (a) with a long vector concatenating all the features extracted by different feature extraction methods. (2) Replacing (a) with features extracted from raw image pixels by a pre-trained Caffe network [15]. (3) Replacing (b) with a commonly used single-modal similarity learning method based on local sensitive hash (LSH) [14]. (4) Removing (c) from the model. (5) Removing (d) from the model. Table 1 lists these different examined settings, where (✓) indicates the use of the proposed design options as in AMoS, while (×) or (□) indicates the use of the alternative options.

The performance of different settings is compared in Table 2 using the NUS-WIDE and CIFAR-10 datasets, where $c = 5$ relation modalities and $d = 3$ image patches are

Settings	NUS-WIDE 500AP (%)	NUS-WIDE mAP (%)	CIFAR-10 500AP (%)	CIFAR-10 mAP (%)
Setting 1	0.05	0.06	0.09	0.11
Setting 2	0.22	0.23	0.16	0.18
Setting 3 ($c = 5$)	0.65	0.63	0.56	0.58
Setting 4 ($c = 5, d = 3$)	0.70	**0.72**	0.62	0.65
Setting 5 ($c = 5, d = 3$)	0.67	0.66	0.63	0.65
Setting 6 ($c = 5, d = 3$)	**0.71**	**0.72**	**0.67**	**0.67**

Table 2: Performance comparison of different model settings.

used in the evaluation. Performance differences between settings 4 and 6, between settings 3 and 4, and between settings 2 and 3 demonstrate the effectiveness of fine-tuning, the attention mechanism, and the multi-modal mechanism, respectively. The high-level image and patch representation obtained after the unsupervised pre-training is able to offer significantly better retrieval performance than the concatenated multi-view features. Although the Caffe network is trained in a supervised way with a large amount of training examples, it is trained to solve a different task from image retrieval and offers lower retrieval performance than the proposed representation learning method (see the performance difference between settings 5 and 6).

4.3. Comparison with State-of-the-art Methods

We now compare the proposed AMoS with seven state-of-the-art algorithms and report the performance in Table 3. The competing methods include a multi-modal similarity learning algorithm of online multiple kernel similarity learning (OMKS) [39], a deep multi-modal similarity learning algorithm of deep semantic ranking hashing (DSRH) [49], a deep similarity learning algorithm of deep regularized similarity comparison hashing (DRSCH) [47], several hashing based algorithms, such as the kernel based supervised hashing (KSH-CNN) [23], mutliview alignment hashing (MAH) [22] and neighborhood discriminant hashing (NDH) [33], as well as a conventional image retrieval approach of iterative quantization (ITQ) [11]. For these competing methods, the parameter settings recommended in their corresponding published works are adopted in our experiments. It can be seen from Table 3 that AMoS outperforms all

Methods	NUS-WIDE 500AP (%)	NUS-WIDE mAP (%)	CIFAR-10 500AP (%)	CIFAR-10 mAP (%)	PLACES2 500AP (%)	PLACES2 mAP (%)	ImageNet 500AP (%)	ImageNet mAP (%)
AMoS ($c = 5, d = 3$)	**0.71**	**0.72**	**0.67**	**0.67**	**0.60**	**0.63**	**0.92**	**0.88**
AMoS ($c = 10, d = 3$)	0.70	0.71	0.65	0.66	0.58	0.60	0.87	0.87
AMoS ($c = 15, d = 3$)	0.70	0.67	0.65	0.66	0.58	0.57	0.86	0.87
AMoS ($c = 20, d = 3$)	0.68	0.67	0.62	0.62	0.56	0.55	0.85	0.87
AMoS ($c = 5, d = 1$)	0.66	0.65	0.60	0.62	0.52	0.50	0.83	0.80
AMoS ($c = 5, d = 2$)	0.70	0.66	0.63	0.64	0.56	0.60	0.88	0.90
AMoS ($c = 5, d = 3$)	**0.71**	**0.72**	**0.67**	**0.67**	**0.60**	**0.63**	**0.92**	**0.88**
AMoS ($c = 5, d = 4$)	0.68	0.69	0.65	0.66	0.58	0.62	0.86	0.87
OMKS [39]	0.60	0.62	0.58	0.55	0.47	0.50	0.70	0.73
DSRH [49]	0.62	0.63	0.64	0.63	0.53	0.51	0.85	0.84
DRSCH [47]	0.63	0.64	0.65	0.63	0.55	0.56	0.88	0.87
NDH [33]	0.30	0.32	0.26	0.32	0.26	0.22	0.50	0.47
ITQ [11]	0.28	0.28	0.22	0.25	0.18	0.17	0.42	0.33
MAH [22]	0.35	0.32	0.38	0.40	0.28	0.37	0.53	0.48
KSH-CNN [23]	0.62	0.62	0.52	0.47	0.44	0.40	0.64	0.70

Table 3: Performance comparison of different methods for different datasets.

the competing methods satisfactorily and for all the studied datasets in terms of both performance measures 500AP and mAP.

4.4. Output Demonstration

In this section, we provide several examples to illustrate the learning output produced by AMoS. In Figure 2, we illustrate multiple pairs of related images. Within each image pair, two pairs of patches possessing the highest attention weight $(s_t^*, h_t^*) = \arg\max_{s,h=1}^{d} e_{ij}^{(t)}(s,h)$ for the top two relation modalities (with the highest similarity scores) are highlighted. To generate the second example, an AMoS model supporting $c = 5$ relation modalities is trained. Figure 3 highlights the salient patch pair possessing the highest attention weight for each relation modality, where three example pairs of related images are illustrated. From the left to the right side of the figure, the superimposed boxes in each column represents the relation modalities ranked in order of descending similarity scores; specifically, the red (yellow) box indicates the patch pair with the highest attention weight under the (second) most dominating relation modality. It can be seen, from Figures 2 and 3, that the proposed model can successfully learn salient patches that contribute significantly to the relevance between two related

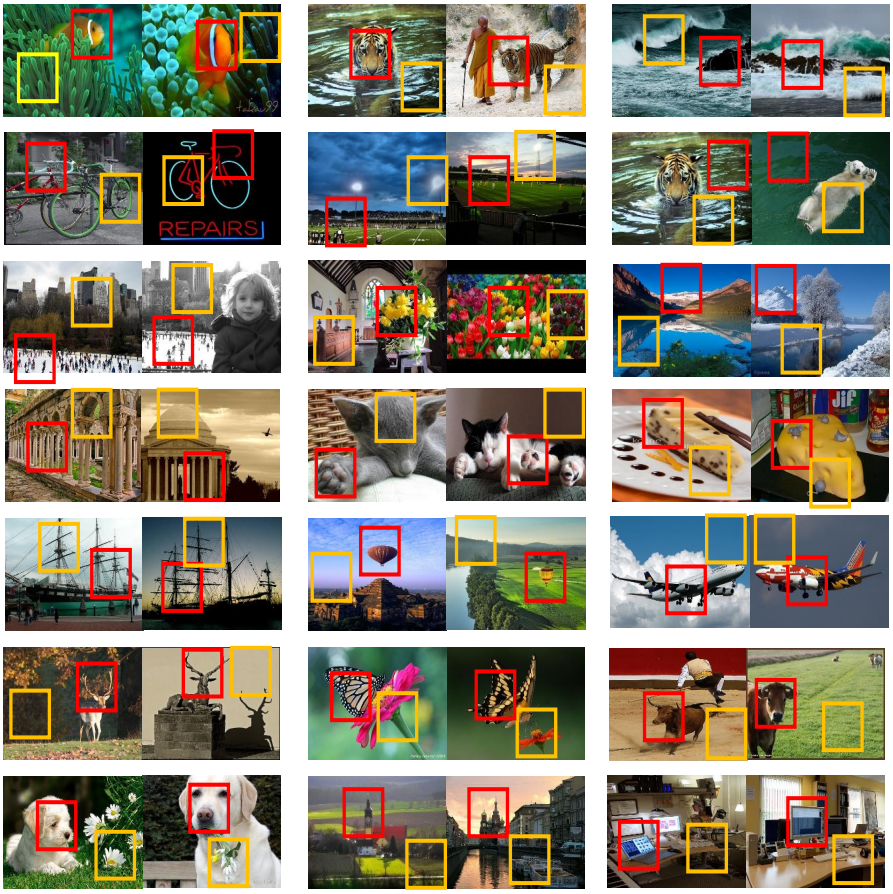

Figure 2: Illustration of the learned salient patch pairs possessing the highest attention weights for the top two relation modalities (red for top one and yellow for top two) with the highest similarity scores.

images. It is also interesting to observe that the way AMoS pays attention to image matching is quite similar to humans.

5. Conclusion

We have proposed the novel similarity learning model AMoS, capable of processing complex relevance patterns exhibiting multi-modal properties. AMoS possesses very good model interpretability and its unique attention mechanism enables the model to dynamically capture salient patches contributing to image relevance. Each of the

Figure 3: Examples of salient patch pairs highlighted in boxes with different colours for $c = 5$ relation modalities. From left to right, the superimposed patch pair corresponds to relation modalities ranked in descending order based on the corresponding similarity scores.

learned relation modalities, according to its relation embedding initialization scheme, can be viewed as either a semantic topic contained by the training images or a user-identified relation type. Properties of the learned relation modalities can be visualized by its signature salient patch pair possessing the highest attention weight over the targeted modality. Model robustness is enhanced by its layer-wise training containing a mixture of unsupervised and supervised training schemes. Quantitative evaluation with the image retrieval task demonstrates the effectiveness of the learned similarity function. Demonstration of output examples from salient patch pairs and relation types indicate some relation between machine intelligence and human vision understanding.

Currently, AMoS learns similarity functions distributed on flat networks (e.g., flat relation modalities encoded as neurons in a flat layer) and it focuses on image objects. In many real-word tasks, objects are connected by hierarchical relations. A potential future direction is to pursue ways of formulating hierarchical multi-modalities to explore more complex relation patterns. Additionally, it is interesting to explore object relations in a cross-modal manner, where different modalities correspond to types of information resources, such as image, text and knowledge graphs, to suit the increasing needs of jointly analyzing multi-media and network data.

Acknowledgements

This work was funded by a visiting Ph.D. studentship from the Chinese Scholarship Council.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. 2014. arXiv preprint arXiv:1409.0473.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[4] A. Bordes, X. Glorot, J. Weston, and Y. Bengio. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, 94(2):233–259, 2014.

[5] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann. Bi-level semantic representation analysis for multimedia event detection. *IEEE Transactions on Cybernetics*, 2016. doi:10.1109/TCYB.2016.2539546.

[6] X. Chang and Y. Yang. Semisupervised feature analysis by mining correlations among multiple tasks. *IEEE Transactions on Neural Networks and Learning Systems*, 2017. doi:10.1109/TNNLS.2016.2582746.

[7] X. Chang, YL. Yu, Y. Yang, and E. P. Xing. Semantic pooling for complex event analysis in untrimmed videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. doi:10.1109/TPAMI.2016.2608901.

[8] LC. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. 2015. arXiv preprint arXiv:1511.03339.

[9] J. K. Chorowskiand, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, pages 577–585, 2015.

[10] X. Gao, T. Mu, and M. Wang. Local voting based multi-view embedding. *Neurocomputing*, 171:901–909, 2016.

[11] Y. Gong, S. Lazebnik, and A. Gordo. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2916–2929, 2013.

[12] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1371–1384, 2008.

[13] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*, pages 2042–2050, 2014.

[14] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *ACM Symposium on Theory of Computing*, pages 604–613. ACM, 1998.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. 2014. arXiv preprint arXiv:1408. 5093.

[16] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[17] 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.

[18] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5455–5463, 2015.

[19] K. Li, G. Qi, J. Ye, and K. Hua. Linear subspace ranking hashing for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. doi:10.1109/TPAMI.2016.2610969.

[20] Y. Lin, Z. Liu, M. Sun, and Y. Liu. Learning entity and relation embeddings for knowledge graph completion. In *Association for the Advancement of Artificial Intelligence*, pages 2181–2187, 2015.

[21] V. E. Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou. Deep hashing for compact binary codes learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2475–2483, 2015.

[22] L. Liu, M. Yu, and L. Shao. Multiview alignment hashing for efficient image search. *IEEE Transactions on Image Processing*, 24(3):956–966, 2015.

[23] W. Liu, J. Wang, R. Ji, YG. Jiang, and SF. Chang. Supervised hashing with kernels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2074–2081. IEEE, 2012.

[24] Y. Luo, T. Liu, D. Tao, and C. Xu. Decomposition-based transfer distance metric learning for image classification. *IEEE Transactions on Image Processing*, 23(9):3789–3801, 2014.

[25] MT. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. 2015. arXiv preprint arXiv:1508.04025.

[26] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, 2014.

[27] Q. Mao, M. Dong, Z. Huang, and Y. Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213, 2014.

[28] P. Merkle, A. Smolic, K. Müller, and T. Wiegand. Multi-view video plus depth representation and coding. In *IEEE International Conference on Image Processing*, volume 1, pages I–201, 2007.

[29] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, 2014.

[30] GJ. Qi, C. Aggarwal, Q. Tian, H. Ji, and T. Huang. Exploring context and content links in social media: A latent space method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):850–862, 2012.

[31] GJ. Qi, W. Liu, C. Aggarwal, and T. S. Huang. Joint intermodal and intramodal label transfers for extremely rare or unseen classes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. doi:10.1109/TPAMI.2016.2587643.

[32] Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1):77–123, 2003.

[33] J. Tang, Z. Li, M. Wang, and R. Zhao. Neighborhood discriminant hashing for large-scale image retrieval. *IEEE Transactions on Image Processing*, 24(9):2827–2840, 2015.

[34] N. A. Tu, DL. Dinh, M. K. Rasel, and YK. Lee. Topic modeling and improvement of image representation for large-scale image retrieval. *Information Sciences*, 366:99–120, 2016.

[35] 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. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[36] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015.

[37] D. Wu, G. Zhang, and J. Lu. A fuzzy preference tree-based recommender system for personalized business-to-business e-services. *IEEE Transactions on Fuzzy Systems*, 23(1):29–43, 2015.

[38] P. Wu, S. CH. Hoi, P. Zhao, C. Miao, and ZY. Liu. Online multi-modal distance metric learning with application to image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):454–467, 2016.

[39] H. Xia, S. Hoi, R. Jin, and P. Zhao. Online multiple kernel similarity learning for visual search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):536–549, 2014.

[40] J. Xie, Y. Fang, F. Zhu, and E. Wong. Deepshape: Deep learned shape descriptor for 3d shape matching and retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1275–1283, 2015.

[41] E. Xing, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pages 505–512, 2002.

[42] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.

[43] X. Xu, W. Li, D. Xu, and I. W. Tsang. Co-labeling for multi-view weakly labeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(6):1113–1125, 2016.

[44] X.Wang and A.Gupta. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[45] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *IEEE International Conference on Computer Vision*, pages 4507–4515, 2015.

[46] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. 2016. arXiv preprint arXiv:1603.03925.

[47] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Transactions on Image Processing*, 24(12):4766–4779, 2015.

[48] S. Zhang, M. Yang, T. Cour, and K. Yu. Query specific rank fusion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4):803–815, 2015.

[49] F. Zhao, Y. Huang, L. Wang, and T. Tan. Deep semantic ranking based hashing for multi-label image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1556–1564, 2015.

[50] J. Zou, W. Li, C. Chen, and Q. Du. Scene classification using local and global features with collaborative representation fusion. *Information Sciences*, 348:209–226, 2016.