Cross-Modal Correlation Learning by Adaptive Hierarchical Semantic Aggregation

Yan Hua, Shuhui Wang, Siyuan Liu, Anni Cai, and Qingming Huang

***Abstract*—With the explosive growth of web data, effective and efficient technologies are in urgent need for retrieving semantically relevant contents of heterogeneous modalities. Previous studies devote efforts to modeling simple cross-modal statistical dependencies, and globally projecting the heterogeneous modalities into a measurable subspace. However, global projections cannot appropriately adapt to diverse contents, and the naturally existing multilevel semantic relation in web data is ignored. We study the problem of semantic coherent retrieval, where documents from different modalities should be ranked by the semantic relevance to the query. Accordingly, we propose TINA, a correlation learning method by adaptive hierarchical semantic aggregation. First, by joint modeling of content and ontology similarities, we build a semantic hierarchy to measure multilevel semantic relevance. Second, with a set of local linear projections and probabilistic membership functions, we propose two paradigms for local expert aggregation, i.e., local projection aggregation and local distance aggregation. To learn the cross-modal projections, we optimize the structure risk objective function that involves semantic coherence measurement, local projection consistency, and the complexity penalty of local projections. Compared to existing approaches, a better bias-variance tradeoff is achieved by TINA in real-world cross-modal correlation learning tasks. Extensive experiments on widely used NUS-WIDE and ICML-Challenge for image–text retrieval demonstrate that TINA better adapts to the multilevel semantic relation and content divergence, and, thus, outperforms state of the art with better semantic coherence.**

***Index Terms*—Cross-modal retrieval, localized correlation learning, semantic hierarchy.**

Manuscript received May 15, 2015; revised September 29, 2015 and December 11, 2015; accepted February 14, 2016. Date of publication February 29, 2016; date of current version May 13, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wolfgang Hürst. *(Corresponding author: Shuhui Wang.)*

Y. Hua was with the Key Laboratory of Intellectual Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China. She is now with the School of Information Engineering, Communication University of China, Beijing 100024, China (e-mail: huayan@cuc.edu.cn).

S. Wang is with the Key Laboratory of Intellectual Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and the School of Information Engineering, Communication University of China, Beijing 100024, China (e-mail: wangshuhui@ict.ac.cn).

S. Liu is with the Smeal College of Business, Pennsylvania State University, University Park, PA 16801 USA, and also with the Center of Cloud Computing, Institute of Advanced Computing and Digital Engineering, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China (e-mail: siyuan@psu.edu).

A. Cai is with the Department of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: annicai@bupt.edu.cn).

Q. Huang is with the Key Laboratory of Intellectual Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qmhuang@ucas.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2535864

I. Introduction

MULTI-MODAL data refer to contents from heterogeneous modalities describing the same or relevant topics, e.g., web images with their surrounding texts and videos with vocal sequences. When online users input queries for seeking complementary multimedia documents of the same topics from other modalities, the retrieved results are expected to be ranked according to the semantic relevance to the queries. With the proliferation of multi-modal data and diversified user demands, effective and efficient cross-modal retrieval techniques are necessitated in multimedia research domain.

Previous text-based techniques compare the similarities of textual query and the surrounding texts of web images, while they suffer from the mismatch between textual descriptions and web images. Classification-based techniques [1], [2] rely on an intermediate risky semantic classification and annotation process, which in turn deliver the cross-modal retrieval problem into a "chicken-egg" dilemma. Correlation learning [3]–[10] is goal-oriented solution. They transform the heterogeneous modalities into measurable low-dimensional representations, hence semantically similar cross-modal documents can be directly retrieved by nearest neighbor search. In this paper, we address two main challenges that hinder existing correlation learning approaches from real applications, i.e., the structured semantic relation and content divergence in cross-modal feature representations.

First, existing correlation learning approaches employ inter-modal correspondence [3], [4], [6], [11] and category information in a single modality [7], [12] for cross-modal retrieval. Both of them are proved to be indispensable for enhancing the retrieval performance. However, these approaches ignore multi-level semantic relation, which is naturally existed among real multi-modal data. For cross-modal retrieval task, if the query text describes "dog," a *semantic coherent* retrieval result should be "dog" images that are naturally co-occurred in certain webpages with the text document, and followed by other "dog" images; "cat" images are also relevant as they belong to "carnivore"; "horse" images are somehow relevant since they are four-leg mammals. But "building" images are unexpected so that they should be positioned behind. In this paper, we study the multi-level semantics in correlation learning and utilize the information to improve the cross-modal retrieval performances and user experiences.

The hierarchical category tree [13], [14] explicitly represents the top-down general-to-specific category relation of real world data. It is widely used to represent multi-level semantics and served as a well-established ontology structure to boost the end tasks such as image retrieval [15], [16] and recognition [14], [17], [18]. However, existing hierarchies [13], [19] are not

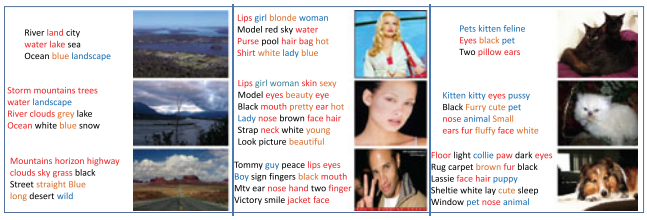


Fig. 1. Illustrations of complementary textual and visual examples. Images and text in the first, second, and third column are on "wild outdoor," "human," and "animal," respectively. Text in red and orange denotes the textual description of objects (parts) and the attributes in the images. Text in blue indicates the topics of the images.

specially constructed for cross-modal learning and semantic coherent retrieval.

Thus, an appropriate semantic hierarchy is needed to model the general-to-specific relationship on multi-modal data. In this paper, by integrating multi-modal content and ontology similarity, we propose to build a balanced category hierarchy with a hierarchical clustering process. The combination of content and ontology similarity endows us with a good measurement on the true semantic relevance among cross-modal data, thus the semantic distances between cross-modal data can be calculated to reflect a more perceptually consistent multi-level semantic relationship.

Second, the content of diversified topics is expressed in heterogenous high-dimensional spaces with cluster-specific and topic-specific cross-modal correlation between feature dimensions. As shown in Fig. 1, the textual words can be roughly divided into objects, attributes and image topics for describing the visual content in cross-modal data. In the first column, the visual characteristics of "wild outdoor" images are represented in color features and scene features, e.g., the skyline. They are relevant to the object description such as "lake," "mountain" and "sky" and the attribute description such as "blue" and "straight" in the "wild outdoor" texts. "Human" and "animal" are semantically relevant categories, as the visual objects in both of them have common five-sense organs described in shapes and textures, see the second and the third columns in Fig. 1. Therefore, they both can be easily discriminated from "wild outdoor." They are also somewhat similar in the textual description of objects and attributes marked in red and orange, respectively. Despite of that, content of "animal" possesses distinguished attributes such as "furry" and "cute" in both visual appearances and textual keywords. Existing correlation learning approaches usually adopt global linear projection [3], [4], [20], [21], hashing function [9], [11], [22], [23] and deep neural network [24]–[26]. However, such a unified projection schema can only capture the principal statistical correlation rather than the cluster-specific correlation. Therefore, they lack adaptability to the complicated semantics and correlations in real world data.

To deal with the content divergence and complicated correlations, we construct local experts, where multiple cross-modal projections are learned simultaneously. The projected coordinates of local projections for data items are weighted combined at either representation level or distance level to measure the distances of cross-modal document pairs. Each local projection is responsible for modeling the content and cross-modal relations among a group of data with content similarity and semantic closeness in either textual or visual modality. The weights with respect to the local experts are determined by the data itself and the probabilistic membership functions defined in our model. Specifically, based on the projected coordinates using the local projection matrices, two aggregation paradigms are investigated, i.e., local projection aggregation and local distance aggregation, see Fig. 5. Local projection aggregation is a combination of locally projected coordinates in each modality. It encodes the local property of each modality to learn the comparable semantic representations. Local distance aggregation is to aggregate a set of local distances calculated on the projected coordinates of local projection pairs from both modalities. The former performs local expert aggregation at representation level, while the latter at distance level. Consequently, we obtain two types of distances between a pair of cross-modal examples. The two strategies better capture the subtle cluster-specific cross-modal correlations than global-projection-based methods [3], [4], [9], [11], [20]–[23], and they are effective to deal with real world cross-modal data in our experiments.

Based on the semantic hierarchy, we organize the cross-modal data into training pairs with multi-level semantic relevance. Accordingly, we propose semantic coherence measurement to model the multi-level semantic relation by max-margin bilateral constraints similar to support vector regression (SVR). Moreover, we design a local projection consistency constraint to regularize the probabilistic membership functions. It enforces that the probabilistic local expert memberships of groups of similar and semantically relevant data items should be similar. The benefit is two-folds. First, each local projection plays a dominant role in correlation measurement only among a group of data items, thus our model better adapts to content divergence in a way analogous to the "divide-and-conquer" strategy. Second, if too many local experts contribute to the projected representation of single data pair, the model may have high overfitting risk. Such kind of inappropriate local expert entanglement may be avoided by imposing local projection consistency constraint. The model parameters of local projections and probabilistic membership functions are jointly learned, and the aggregated local distances between training pairs are learned towards their pair-wise semantic relevance with controllable local expert behavior.

In summary, we propose TINA, a cross-modal correlation learning method by adap**T**ive h**I**erarchical sema**N**tic **A**ggregation. By optimizing the structure risk objective function that involves semantic coherence measurement, local projection consistency and model complexity penalty, a set of local projections and probabilistic membership functions are constructed for both modalities. Our key contributions include the following.

1) We encode multi-level sematic relation into cross-modal distance with a max-margin regression learning framework. To our best knowledge, our work is the first to study the semantic coherence in cross-modal retrieval.
2) We construct a semantic hierarchy for cross-modal retrieval by joint modeling of visual, textual and ontology

similarities. It appropriately encodes the relation of cross-modal documents from both content and semantic perspectives.

3) We propose a structure risk objective function to learn multiple local experts which are probabilistically aggregated towards the semantic coherent cross-modal representations and distances. Our model adapts to the cross-modal content divergence and multi-level semantic correlation. With the learned projections, semantically coherent retrieval on large scale cross-modal documents can be performed by simple ranking on distances.
4) Extensive image-text retrieval experiments on large scale NUS-WIDE and ICML-Challenge data show that TINA outperforms state-of-the-art approaches.

II. Related Work

A. Correlation Learning on Heterogeneous Modalities

From the methodology perspective, existing correlation learning methods can be categorized as follows.

Subspace learning, e.g., canonical correlation analysis (CCA) and its variants, provides direct solutions for cross-modal retrieval by learning a pair of transformations to project heterogenous data into a measurable low dimensional subspace. Rasiwasia *et al.* [1] proposed cross-modal topic classifiers constructed on the CCA representations to map heterogeneous data into a unified semantic space. Partial least square (PLS) [4] formulates the problem with a bilateral regression model. Cross-modal factor analysis (CFA) [21] minimizes the projected distances of coupled samples with linear orthogonal transformations. Multi-view CCA (MCCA) [20] incorporates a third view capturing high-level image semantics into the same latent space besides the visual and textual views. The category information of single modality is incorporated by local discrimination CCA [27]. Locality preserving CCA [28] is proposed to incorporate local structure information of the single modality into CCA. Sun *et al.* [29] proposed discriminant CCA utilizing category information of multi-views and performed discriminative feature extraction. Sharma *et al.* [7] proposed a generalized multi-view analysis model to learn the (non-)linear subspace using side information. Sparse CCA (SCCA) methods [5], [30] are proposed to deal with the sparsity issue of textual modality. Archambeau and Bach [31] and Virtanen *et al.* [32] provided Bayesian interpretation of CCA-based models. Yang *et al.* [33] proposed a new semi-supervised method with local regression and global alignment to learn a Laplacian matrix for data ranking, which can be used to learn the representation of co-occurring multimedia objects.

Multi-modal deep models have been investigated to address the cross-modal retrieval problem. Andrew *et al.* [6] proposed deep CCA to learn stacked nonlinear transformations. Multimodal auto-encoders [24] and multi-modal restricted Boltzmann machines [25] are used as building blocks for deep-model-based shared representation learning. Masci *et al.* [22], [23] constructed multi-layered neuro-networks with both intra-modal similarity and inter-modal correlation. More recently, Karpathy *et al.* [26] embeded fragments of images (objects detected with RCNN) and fragments of sentences (typed dependency tree relations) into a common space.

Cross-modal hashing methods learn a common hamming space for efficient indexing and bit-wise distance calculation of cross-modal data. Bronstein *et al.* [9] proposed a boosting-based hashing method (CMSSH) for two modalities. Graph-based hashing methods (e.g., CVH [34] and inter-media hashing (IMH) [8]) encode intra-modal similarity and inter-modal co-occurrence into a unified graph representation, and the subspaces can be obtained through eigen-decomposition on the equivalent graph Laplacian. Zhen and Yeung [35] proposed multi-modal latent binary embedding approach to learn binary latent factors on the Gaussian distribution component sets.

Our model possesses distinctive characteristics from state-of-the-art. Compared to CCA-based models, which usually employ global linear transformations, our local expert learning strategy adapts to the divergence in both content and correlation with a better bias-variance tradeoff. Compared to deep models [6], [24], [25], [36], which gain model capacity by increasing the number of layers, TINA increases the number of local projections to gain more flexibility in processing real data. Nevertheless, the visual features extracted by deep models, e.g., convolutional neural network [37], can also be utilized in our framework to achieve high retrieval performances. The comparable representations learned by our approach could also be further binarized to improve retrieval efficiency on big data. A related work to ours is a locally aligned multi-view transformation approach [38] for binary classification. In contrast, we impose guidance on how data is associated with the local projections by both content and semantic similarity, and the proposed max-margin bilateral loss function is specially designed for our semantic coherence cross-modal retrieval task. In summary, TINA models the hierarchical semantic relation on cross-modal data that is neglected by the above-mentioned approaches. The learned local projections are endowed with more adaptation to divergent content and complicated semantic relations.

B. Modeling Semantic Hierarchy

Semantic hierarchy [13] is a formally defined taxonomy or ontology structure. For image and multimedia analytics [39], it can be general or domain specific. A large scale image database ImageNet [39] has been constructed by collecting images for each concept node on WordNet [13]. Based on the inter-category confusion matrix [19], the hierarchical structure of object categories can be automatically created by top-down or bottom-up recursive clustering processes. Besides the visual features, Li *et al.* [14] integrated tags to automatically build the "semantivisual" hierarchy, which encodes the general-to-specific semantic and visual relationship. Marszalek and Schmid [40] constructed the category hierarchy that postpones decisions in the presence of uncertainty. Sivic *et al.* [41] proposed to automatically discover a visual feature hierarchical structure from an unlabeled image collection. Mazloom *et al.* [42] discovered the informative concept detectors to different types of multimedia events by cross-entropy optimization.

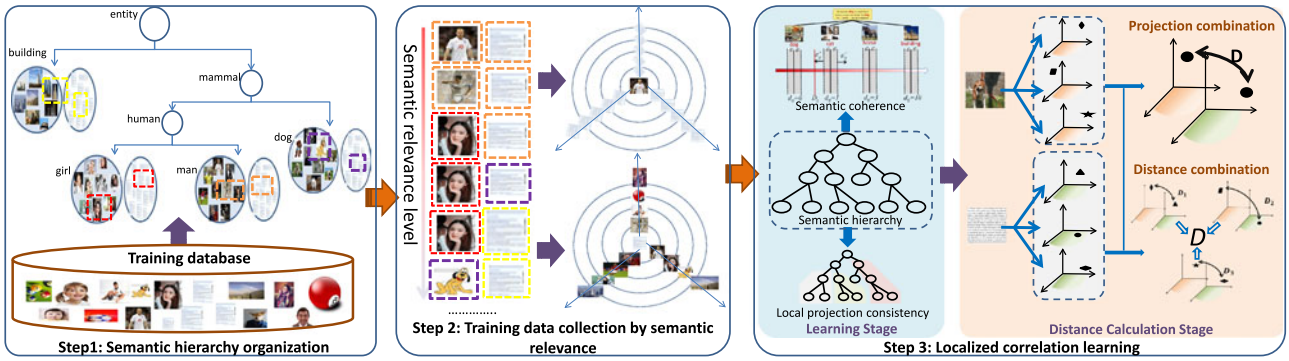


Fig. 2. Framework of TINA. We build a balanced semantic hierarchy by top-down hierarchical clustering in Step 1. Then a set of training pairs are sampled from the database, where each document is associated with the cross-modal documents ranked by their semantic relevance, as shown in Step 2. We learn a set of local linear projections and probabilistic membership functions by optimizing a structure risk objective function, as shown in Step 3.

Organizing semantic concepts from general to specific has been shown to be effective in boosting the performance of real world applications. Deselaers and Ferrari [16] computed the semantic distance between images by measuring the divergence of concept distribution of their neighborhoods. Deng *et al.* [15] developed an image retrieval framework with semantic visual representations and hierarchical bilinear similarity function based on the pairwise semantic affinity. Verma *et al.* [18] constructed similarity metric hierarchy to associate the metric of a certain node (category) and that of its parent nodes. Then all the metrics are learned jointly through hierarchical aggregation for nearest neighbor classifiers. A tree of metrics is also learned by Grauman *et al.* [17], where the appropriate (dis)similarity constraints on its subtree members are imposed to enforce the metrics on different tree levels to be different.

In the context of cross-modal retrieval, the "hierarchy" used in [43] refers to a two-level hierarchical manifold, i.e., manifold for each modality and a multimedia document semantic graph. It focuses on the manifold property of different modalities to be harmonized to the semantic space defined on multi-media documents. In contrast, the semantic hierarchy, defined in this paper, means the general-to-specific category relation that has not been well-studied in cross-modal correlation learning. We combine the category level content similarity and ontology similarity to construct a balanced semantic hierarchy based on a top-down recursive category partition process. Such a hierarchy is then seamlessly embedded in our max-margin regression framework to enhance the semantic coherence of the learned cross-modal distance.

III. Problem and Solution Framework

Definition 1: The semantic coherence of cross-modal retrieval means that the retrieved documents from heterogenous modalities are ranked according to the semantic relevance of their category labels to that of the given query document.

For example, on NUS-WIDE data (see Fig. 3), given an image query of "statue," the idealized retrieval result should be: the corresponding textual description of the query image top ranked, followed by text documents of "temple" (the sibling of "statue"), then followed by "castle" (the sibling of the second upper layer), and then by "road" (the sibling of the third upper layer), etc.

To this end, our proposed TINA consists of the following key steps (see Fig. 2).

Step 1: Semantic hierarchy organization (see Section IV). We propose a data-driven semantic hierarchy organization approach by joint visual, textual and ontology similarity modeling and top-down recursive hierarchical clustering. The nodes on the hierarchy share different levels of content and ontology relations with respect to their tree path distances and the depth of their parent nodes on the tree.

Step 2: Training data collection. Based on the hierarchy, a set of training data pairs are sampled from the database, where each document is associated with its correspondence and documents from different semantic levels. Accordingly, we calculate the *semantic distance* and *intra-modal similarity* matrices (see Section V) on the selected data pairs for subsequent model training.

Step 3: Localized correlation learning (see Section V). We utilize the training pairs with multi-level semantic relevance collected in Step 2 for max-margin regularized regression learning. With the learned model, semantic coherent cross-modal retrieval is performed by nearest neighbor search.

IV. Semantic Hierarchy

We are given a cross-modal dataset $\mathbb{D} = \{x_i, y_i, c_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{d_x}$ and $y_i \in \mathbb{R}^{d_y}$ denote the ith training data pair from X and Y modalities, respectively. $c_i \in \{1, 2, \ldots, C\}$ denotes the category index of the ith training pair. Since there is complicated semantic relation among the categories, we construct a semantic category hierarchy $\mathbb{H}$ on $\mathbb{D}$ by combining the similarity modeling from visual domain, textual domain and ontology relatedness.

A. Visual Similarity

The appearances of visual categories are divergent. In real world images, the intra-class visual divergence may be even larger than inter-class divergence, which results in clutter and overlap among visual categories. To alleviate these adverse effects, we develop a simple and effective method based on visual

subcategory similarity. We use K-means to divide category c into k_c subcategories, $X_{c,k} = \{x_i\}, c_i = c, k = 1, \ldots, k_c$, where each subcategory contains images with higher visual cohesiveness. Then, average visual feature in each subcategory $v_c(k) = \frac{\sum x_i}{n_c^k}, x_i \in X_{c,k}$ is calculated, where n_c^k is the number of samples in $X_{c,k}$. We define the distance between two categories as

$$D_v(c_1, c_2) = \min_{k_1, k_2} ||v_{c_1}(k_1) - v_{c_2}(k_2)||_2^2. \quad (1)$$

The similarity of two categories is calculated as

$$S_v(c_1, c_2) = \exp\left(-\frac{D_v(c_1, c_2)}{\sigma_v^2}\right) \quad (2)$$

where σ_v is a bandwidth parameter that controls the sensitivity to the distance range. Its value is the average of all category distances. The min rule identifies the similarity between categories with their most similar images. The scheme guarantees that the category level visual similarity is robust to uncontrolled appearance variations.

B. Textual Similarity

Textual information is represented by a constant set of lexical terms. We use the average of the BOW features as the textual description of category c, denoted by t_c. We define the similarity between c_1 and c_2 as

$$S_t(c_1, c_2) = \frac{t_{c_1}^\top t_{c_2}}{\|t_{c_1}\| \, \|t_{c_2}\|}. \quad (3)$$

C. Ontology Similarity

Ontology similarity reflects the semantic relatedness between two concepts from a taxonomic point of view. Most of conceptual similarity between two concepts is computed according to their shortest path on WordNet [13], [44]–[46]. In our ontology similarity computation, we adopt the similarity normalized with the depth of their parent node [47]. The similarity of concepts c_1 and c_2 in the ontology similarity matrix S_o is calculated as

$$S_o(c_1, c_2) = \frac{\beta d_e(p_a(c_1, c_2))}{p(c_1, c_2) + \beta d_e(p_a(c_1, c_2))} \quad (4)$$

where $d_e(p_a(c_1, c_2))$ is the depth of the parent node of c_1 and c_2, and $p(c_1, c_2)$ is the length of shortest path on WordNet. In (4), the deeper the parent node is, the more similar the two concepts will be ($\to 1$). The longer the path between the two concepts is, the more semantically distinct they will be.

D. Tree Hierarchy

We linearly combine the above three similarity matrices to get a combined similarity matrix

$$S = \alpha_1 S_v + \alpha_2 S_t + \alpha_3 S_o \quad (5)$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1, \alpha_1, \alpha_2, \alpha_3 \geq 0$. Based on the combined similarity matrix, we seek to exploit a new hierarchical category structure $\mathbb{H}$ in a top-down partition process. From the root node, we divide each internode into n child nodes using spectral clustering successively, until we reach the tree layer whose internodes include no more than n leaf nodes.

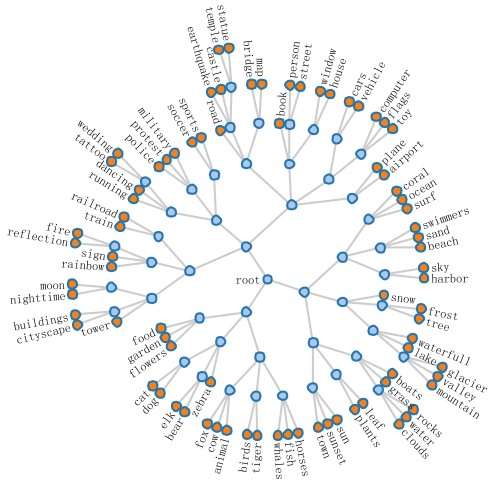


Fig. 3. Semantic hierarchy on NUS-WIDE with $n = 3$.

The parameters $\{\alpha_1, \alpha_2, \alpha_3\}$ are selected by a validation set with subjective scoring on the resulted hierarchy tree. We design 15 sets of values, resulting in 15 hierarchies on the validation dataset. Then, ten non-expert subjects are asked to evaluate the trees with excellent (1), good (0) and bad (-1). The optimal weights $\{\alpha_1 = 0.333, \alpha_2 = 0.5, \alpha_3 = 0.167\}$ produce the structure with the highest average score among the 15 trees on NUS-WIDE dataset. The optimal weights are $\{\alpha_1 = 0.2, \alpha_2 = 0.5, \alpha_3 = 0.3\}$ on ICML-Challenge dataset. Textual similarity contributes the most to the hierarchy establishment. The categories on ICML-Challenge are the most frequently occurred ones in the tag vocabulary, while categories on NUS-WIDE are predefined object classes. On WordNet hierarchy, categories of NUS-WIDE tend to be more widely distributed than those of ICML-Challenge, and their ontology relationships are more consistent with the true semantic relation than those of NUS-WIDE [48]. Therefore, the ontology similarity on ICML-Challenge contributes more to the combined similarity than on NUS-WIDE.

Two exemplar hierarchical structures built on NUS-WIDE ($n = 3$) and ICML-Challenge ($n = 5$) are shown in Figs. 3 and 4. In Figs. 3 and 4, the semantic hierarchies we build are more explicable and commonsensical from both content and semantic perspectives. For example, in Fig. 3, "elk," "bear" and "zebra" are close on the hierarchy since they are all mammals. "elk" and "bear" are more consistent in visual appearance and textual description, e.g., they are both furry and gray. "zebra" with obvious stripes is the sibling node of the parent of "elk" and "bear." In Fig. 4, "boat," "truck" and "car" are vehicles for transportation, while "desk," "chair" and "bed" are furniture. Both vehicles and furniture belong to instruments. Therefore, the parent nodes of "truck" and "desk" are siblings to each other.

On the hierarchy, the length of shortest path does not appropriately indicate the specificity of ontology [15], [47]. Intuitively, the depth of parent node of two concepts represents the category specificity [47]. To encode such knowledge, we define a

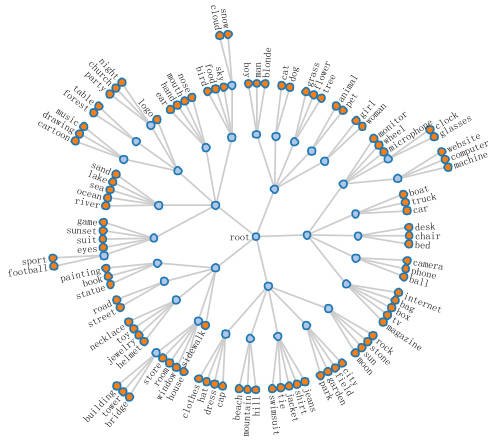


Fig. 4. Semantic hierarchy on ICML-Challenge with $n = 5$.

normalized distance of two nodes on $\mathbb{H}$ as follows:

$$h(c_1, c_2) = \frac{p(c_1, c_2)}{L_h \cdot d_e(p_a(c_1, c_2))} \tag{6}$$

where $d_e(p_a(c_1, c_2))$ is the depth of the parent node of c_1 and c_2, $p(c_1, c_2)$ is the length of the shortest path on $\mathbb{H}$, and L_h is the length of the longest path on $\mathbb{H}$. The larger the distance between the two categories is, the more semantically distinct they will be. The normalized category distance is used to calculate the *semantic distance* d_{ij} in each empirical training pair, which will be used in (13) in Section V.

Discussion: The visual similarity reflects the category similarity from visual appearance perspective, the text similarity from linguistic perspective, and the ontology similarity from semantic perspective. A combination of three similarities produces a more perceptually consistent hierarchical structure than using any single metric. Removing any of them will discount the perceptual consistency according to the survey on the involved subjects. The subjective evaluation sufficiently reflects the semantic consistency from user perspective, thus it provides direct guidance on the construction of semantic hierarchy and the weights of visual, textual and ontology similarity on general cross-modal dataset. Besides subjective evaluation, other possible technique can be employed. For example, if part of the ground-truth category similarity matrix can be obtained, the parameters can be learned by minimizing the overall element-wise difference between the combined similarity and the ground-truth similarity. However, this is beyond the scope of this paper and left for future study.

V. Correlation Learning

Given the cross-modal data $\mathbb{D}$, and its semantic hierarchy tree $\mathbb{H}$, we want to learn a set of local linear projections and probabilistic membership functions. They are jointly optimized to guarantee that the correlations of data $\mathbb{D}$ are consistent with their relations on $\mathbb{H}$, so that the *semantic coherence* of cross-modal retrieval is better achieved.

We define a structure risk objective function including semantic coherence measurement, local projection consistency and parameter regularization. The empirical loss term, i.e., semantic coherence measurement, guarantees the minimization of differences between feature distances and semantic distances of cross-modal data. Local projection consistency expects that a group of similar data items select the same local experts to learn comparable representations. We list important parameters and notations used in this paper in Table I for quick reference. The structure risk objective function is minimized with respect to the local projection matrices ($\mathbf{W}_k^x$ and $\mathbf{W}_k^y$) and the parameters (ϕ_k and ψ_k) of the probabilistic membership functions.

TABLE I
Notations Used in This Paper

Notation	Meaning
$\mathbf{W}_k^x$	the kth local projection for X modality
$\mathbf{W}_k^y$	the kth local projection for Y modality
$g_i^x(k)$	the kth probabilistic weight for the ith example in X
$g_i^y(k)$	the kth probabilistic weight for the jth example in Y
$q_{\mathrm{ij}}(k)$	probabilistic weight for the kth local distance
ϕ_k	parameters of p.m.f.[1] for X modality
ψ_k	parameters of p.m.f. for Y modality
$D_p(x_i, y_j)$	distance of x_i and y_j with local projection aggregation
$D_d(x_i, y_j)$	distance of x_i and y_j with local distance aggregation
$D(x_i, y_j)$	represents $D_p(x_i, y_j)$ or $D_d(x_i, y_j)$
d_{ij}	semantic distance for the ith and jth examples

[1]p.m.f. = probabilistic membership functions.

A. Local Expert Aggregation

We propose two strategies for aggregating local experts in different modalities. The first is aggregation of the local projected representations, called *local projection aggregation* [see Fig. 5(a)]. Cross-modal data are projected into a comparable low-dimensional space with a number of local projections in each modality. The second is aggregation of local distances measured by pairs of local experts, called *local distance aggregation* [see Fig. 5(b)]. It organizes the local projections of different modalities in a pairwise manner. The projected representations or the pair-wise distances are aggregated where the weights are determined by the probabilistic membership functions. Intuitively, the former performs local expert aggregation at representation level, and the latter at distance level.

Local projection aggregation: We learn a set of local projection functions $\{\mathbf{W}_k^x \in \mathbb{R}^{d\times d_x}\}_{k=1}^{\mathcal{K}_x}$ and $\{\mathbf{W}_k^y \in \mathbb{R}^{d\times d_y}\}_{k=1}^{\mathcal{K}_y}$ for both modalities. Examples x_i and y_j from two modalities are projected using $\mathbf{W}_k^x, k = 1, \ldots, \mathcal{K}_x$ and $\mathbf{W}_k^y, k = 1, \ldots, \mathcal{K}_y$, respectively. In this paper, we set $\mathcal{K} = \mathcal{K}_x = \mathcal{K}_y$ for simplicity. We define the distance in the aggregated projected space as

$$D_p(x_i, y_j) = ||f_x(x_i) - f_y(y_j)||^2 \tag{7}$$

where $f_x(x_i)$ for X and $f_y(y_j)$ for Y are defined as

$$f_x(x_i) = \sum_{k=1}^{\mathcal{K}} g_i^x(k)\mathbf{W}_k^x x_i, f_y(y_j) = \sum_{k=1}^{\mathcal{K}} g_j^y(k)\mathbf{W}_k^y y_j \tag{8}$$

where $\mathbf{W}_k^x$ is the kth local projection for X modality, and $\mathbf{W}_k^y$ is the kth local projection for Y modality. $g_i^x(k)$ and $g_j^y(k)$ denote the non-negative probabilistic weights of local projection

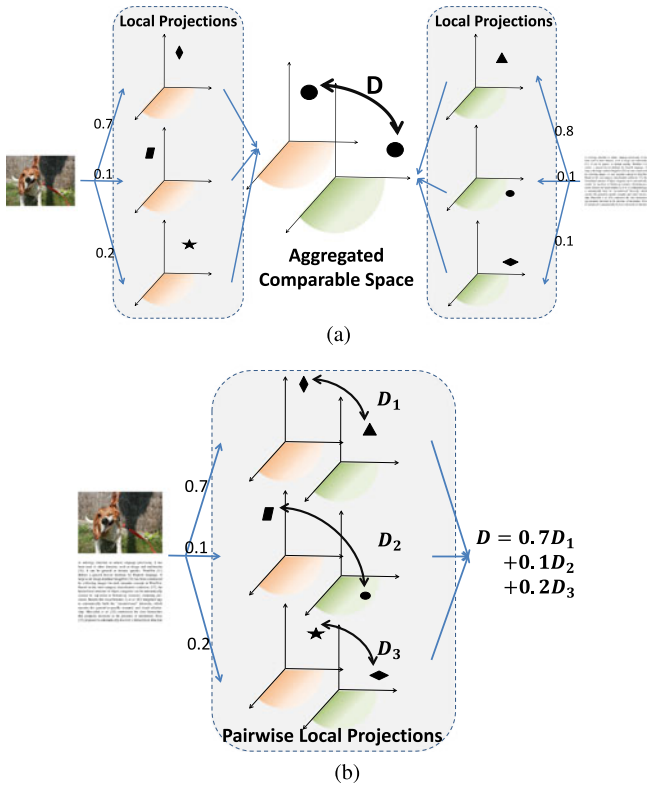


Fig. 5. Two paradigms of local expert aggregation in this paper. (a) Local projection aggregation. (b) Local distance aggregation. In (a), D is the distance of the two aggregated representations, and the aggregated representation is the weighted combination of the projected coordinates of the local projections. In (b), D_1, D_2, and D_3 are the distances computed on the representations projected by three pairs of local experts. D denotes the weighted aggregated cross-modal distance with D_1, D_2, and D_3.

aggregations for $f_x(x_i)$ and $f_y(y_j)$, defined as

$$g_i^x(k) = \frac{\exp(\phi_k^\top x_i)}{\sum_{k'} \exp(\phi_{k'}^\top x_i)}, \quad g_j^y(k) = \frac{\exp(\psi_k^\top y_j)}{\sum_{k'} \exp(\psi_{k'}^\top y_j)} \tag{9}$$

where $\{\phi_k \in \mathbb{R}^{d_x}\}$ and $\{\psi_k \in \mathbb{R}^{d_y}\}$ are parameters of the kth probabilistic membership function. The definitions of $g_i^x(k)$ and $g_j^y(k)$ ensure that $\sum_{k=1}^{\mathcal{K}} g_i^x(k) = 1$ and $\sum_{k=1}^{\mathcal{K}} g_j^y(k) = 1$.

Local distance aggregation: We learn $\mathcal{K}$ pairs of local projection functions $\mathbf{W} = \{\mathbf{W}_k^x \in \mathbb{R}^{d\times d_x}, \mathbf{W}_k^y \in \mathbb{R}^{d\times d_y}\}_{k=1}^{\mathcal{K}}$. The examples x_i and y_j are projected using the $\mathcal{K}$ pairs of local projections. Then the distance between the two examples is the weighted sum of the distance calculated with one pair local projections as

$$D_d(x_i, y_j) = \sum_{k=1}^{\mathcal{K}} q_{\mathrm{ij}}(k) D_k(x_i, y_j) \tag{10}$$

where $D_k(x_i, y_j)$ is defined as

$$D_k(x_i, y_j) = ||\mathbf{W}_k^x x_i - \mathbf{W}_k^y y_j||^2 \tag{11}$$

where $\mathbf{W}_k^x$ and $\mathbf{W}_k^y$ are the kth local projection for X and Y modality, respectively. $q_{\mathrm{ij}}(k)$ denotes the non-negative weight of the kth local distance, which is defined as

$$q_{\mathrm{ij}}(k) = \frac{\exp(\phi_k^\top x_i)\exp(\psi_k^\top y_j)}{\sum_{k'} \exp(\phi_{k'}^\top x_i)\exp(\psi_{k'}^\top y_j)}, \quad k = 1, \ldots \mathcal{K} \tag{12}$$

where $\{\phi_k \in \mathbb{R}^{d_x}\}$ and $\{\psi_k \in \mathbb{R}^{d_y}\}$ are parameters of the kth probabilistic membership function. The definition of $q_{\mathrm{ij}}(k)$ ensures that $\sum_{k=1}^{\mathcal{K}} q_{\mathrm{ij}}(k) = 1$. The formulations of our methods are similar to CFA [21], which calculates the distance between features from two modalities by linear orthogonal transformations. Regardless of the orthogonality, the distance definition in CFA is a special case (i.e., $\mathcal{K} = 1$) of our methods.

Discussion: There are several classic types of projection function in existing models with different sensitivity levels on the cross-modal data distribution. For example, unified projection in CCA-based model [3] tends to be over-smooth on large scale data, thus it lacks the adaptability to content divergence and complicated semantic relation. One projection for each data point is learned in sample specific projection [49]. It is too sensitive to outliers and local tangent structure. It also involves intensive computational burdens which is prohibitive for processing large scale high-dimensional data. Gavves *et al.* [50] proposed reduced linear kernels that use only a portion of the dimensions to reconstruct a linear kernel. It maintains the accuracy benefits from non-linear embedding methods that mimic non-linear SVMs.

Our local experts perform information averaging among a group of samples in each modality, which achieves a better bias-variance trade-off between the two typical projection functions. In local projection aggregation, the local experts are probabilistically aggregated at representation level by membership functions that are adaptively fit to each cross-modal training datum. In local distance aggregation, the local experts are probabilistically aggregated at distance level. Pairs of local experts collaboratively adapt to the semantic correlation of a group of cross-modal data pairs. Compared to local projection aggregation, local distance aggregation avoids entanglement among local experts within one modality. Both of them have the same model complexity level. In addition, our method achieves a good approximation to the nonlinear semantic distance by performing dimension reduction and cluster-specific linear embedding. Different from method in [50], which only works on additive kernels, the embedding and aggregation function are explicitly given in our framework, and arbitrary nonlinear semantic distance can be firmly approximated by minimizing the empirical loss defined in (14).

In consequent sections, we use $D(x_i, y_j)$ to represent the distance computed by local expert aggregation. We denote our method as $\texttt{TINA}_p$ and $\texttt{TINA}_d$ when the local experts are aggregated at representation level and distance level, respectively.

B. Semantic Coherence Measurement

The learned cross-modal distance $D(x_i, y_j)$ on training data $\mathbb{D}$ is expected to be consistent with the semantic distance d_{ij} defined on hierarchy $\mathbb{H}$. A normalized distance of two semantic categories is defined in (6). Additionally, to deal with multi-label information in real-world data, we calculate the *semantic*

distance as

$$d_{\mathrm{ij}} = \min_{x_i(y_i)\in c_1, x_j(y_j)\in c_2} h(c_1, c_2) \tag{13}$$

where $h(c_1, c_2)$ is the distance of concepts c_1 and c_2 on $\mathbb{H}$ as in (6).

The consistency is considered to be achieved if the absolute differences between the semantic distances and the learned distances of the training pairs are less than a predefined tolerance ε. Unfortunately, not all the training pairs perfectly satisfy the semantic consistency in practical situations. Therefore, we introduce two slack variables $\varepsilon_{\mathrm{ij}}^{+}$ and $\varepsilon_{\mathrm{ij}}^{-}$ for each training pair to measure the inconsistency out of ε. The relaxed formulation is defined as

$$\begin{aligned} D(x_i, y_j) - d_{\mathrm{ij}} &\le \varepsilon + \varepsilon_{\mathrm{ij}}^{+}, \varepsilon_{\mathrm{ij}}^{+} \ge 0 \\ d_{\mathrm{ij}} - D(x_i, y_j) &\le \varepsilon + \varepsilon_{\mathrm{ij}}^{-}, \varepsilon_{\mathrm{ij}}^{-} \ge 0 \end{aligned} \tag{14}$$

where $\varepsilon_{\mathrm{ij}}^{+}$ and $\varepsilon_{\mathrm{ij}}^{-}$ are the slack variables of the positive side and negative side, respectively. The constraint in (14) is similar in spirit with SVR. ε controls the precision in distance fitting. A small ε results in a good fitting on training data but with too many "support vectors," while a large ε leads to over-smoothness and under-fitting. To achieve a better trade-off, we empirically set $\varepsilon = 0.01$.

The learned distances among data pairs are optimized towards the multi-level semantic distances on $\mathbb{H}$. For example, if the semantic distances d_{ij} between the ith image and the j_1th, j_2th and j_3th texts are 0.1, 0.5 and 0.9, respectively, then the learned distances $D(x_i, y_{j_1})$, $D(x_i, y_{j_2})$, $D(x_i, y_{j_3})$ should not deviate too much from 0.1, 0.5 and 0.9, respectively. By allowing a small tolerance ε, our model minimizes the total deviation reflected by the summation of $\varepsilon_{\mathrm{ij}}^{+}$ and $\varepsilon_{\mathrm{ij}}^{-}$ over all the training pairs, so that the relevance ranking order of j_1, j_2 and j_3 is less likely to change given x_i as the query. A possible alternative is to employ the relative measurement [51] for cross-modal retrieval. But it involves a semantic relevance comparison among different training pairs. Thus it results in a more complicated model.

C. Local Projection Consistency

We impose local projection consistency on probabilistic membership functions of both $\mathtt{TINA}_p$ and $\mathtt{TINA}_d$, which manifests that the adjacent and semantically related data in each modality should possess similar weighting values. For both cases, the consistency is measured by intra-modal content and semantic similarity as

$$L_\phi = \sum_{\mathrm{ij}} s_{\mathrm{ij}}^{x} (\overline{g}_i^x - \overline{g}_j^x)^2, \quad L_\psi = \sum_{\mathrm{ij}} s_{\mathrm{ij}}^{y} (\overline{g}_i^y - \overline{g}_j^y)^2 \tag{15}$$

where

$$\begin{aligned} s_{\mathrm{ij}}^{x} &= \exp\left(-\frac{d_{\mathrm{ij}}}{2\sigma_d^2}\right) \exp\left(-\frac{(x_i - x_j)^2}{2\sigma_x^2}\right) \\ s_{\mathrm{ij}}^{y} &= \exp\left(-\frac{d_{\mathrm{ij}}}{2\sigma_d^2}\right) \exp\left(-\frac{(y_i - y_j)^2}{2\sigma_y^2}\right) \end{aligned} \tag{16}$$

where d_{ij} denotes the *semantic distance* for training pair of $x_i(y_i)$ and $x_j(y_j)$. The intra-modal similarity s_{ij}^{x} and s_{ij}^{y} jointly consider semantic similarity and feature similarity, where σ_d and σ_x (σ_y) represent the sensitivity parameters for semantic distance and feature distance. σ_d^2 and σ_x^2 (σ_y^2) are the average values of the semantic distances and feature distances among training data pairs, respectively. For $\mathtt{TINA}_p$, $\overline{g}_i^x = g_i^x$ and $\overline{g}_i^y = g_i^y$. For $\mathtt{TINA}_d$, $\overline{g}_i^x(k) = \frac{\exp(\phi_k^\top x_i)}{\sum_{k'} \exp(\phi_{k'}^\top x_i)}$ and $\overline{g}_j^y(k) = \frac{\exp(\psi_k^\top y_j)}{\sum_{k'} \exp(\psi_{k'}^\top y_j)}$. After the model learning, $q_{\mathrm{ij}}(k)$ is calculated and normalized by using ϕ_k and ψ_k, $k = 1, \ldots, \mathcal{K}$.

By optimizing local projection consistency, each local expert $\mathbf{W}_k^x$ ($\mathbf{W}_k^y$) will play a dominant role in constructing correlation among a subset of similar and semantically relevant data, thus more robustness and consistency can be achieved. For example, "dog" and "cat" are recognized as "pets." They are similar in feature representation as they are "four-leg" in shape and fluffy. They are also semantically relevant on both WordNet and $\mathbb{H}$ (see Figs. 3 and 4). Accordingly, the correlations among documents of two similar categories will be encoded by some specific local experts, where local projection selectivity is performed by probabilistic membership functions with the learned parameters ϕ and ψ.

D. Objective Function

With semantic coherence measurement defined in (14) and local projection consistency defined in (15), we jointly learn local projections $\{\mathbf{W}^x, \mathbf{W}^y\}$ and probabilistic membership functions $\{\phi, \psi\}$ by minimizing the following structure risk objective function

$$\begin{aligned} L_{\mathbf{W},\phi,\psi} = \; & \frac{1}{2}\sum_k \left(||\mathbf{W}_k^x||^2 + ||\mathbf{W}_k^y||^2\right) \\ & + \frac{C_1}{N_1}\sum_{i,i} (\varepsilon_{\mathrm{ii}}^{-} + \varepsilon_{\mathrm{ii}}^{+}) + \frac{C_2}{N_2}\sum_{i,j\neq i} (\varepsilon_{\mathrm{ij}}^{-} + \varepsilon_{\mathrm{ij}}^{+}) \\ & + \frac{\beta}{N_3}\sum_{i,j} s_{\mathrm{ij}}^{x} (\overline{g}_i^x - \overline{g}_j^x)^2 + \frac{\gamma}{N_4}\sum_{i,j} s_{\mathrm{ij}}^{y} (\overline{g}_i^y - \overline{g}_j^y)^2 \\ \text{s.t.} \quad & D(x_i, y_j) - d_{\mathrm{ij}} \le \varepsilon + \varepsilon_{\mathrm{ij}}^{+}, \varepsilon_{\mathrm{ij}}^{+} \ge 0 \\ & d_{\mathrm{ij}} - D(x_i, y_j) \le \varepsilon + \varepsilon_{\mathrm{ij}}^{-}, \varepsilon_{\mathrm{ij}}^{-} \ge 0 \end{aligned} \tag{17}$$

where the first term is the complexity penalty of local projections to avoid over-fitting. $\varepsilon_{\mathrm{ii}}^{+}$ and $\varepsilon_{\mathrm{ii}}^{-}$ are the slack variables for the cross-modal pair with correspondence, where x_i and y_i are the document pairs from different modalities. $\varepsilon_{\mathrm{ij}}^{+}$ and $\varepsilon_{\mathrm{ij}}^{-}$, $i \neq j$, are slack variables of the multi-level semantic relevance. Their weights are determined by C_1 and C_2. In the constraints, if $i = j$, $d_{\mathrm{ij}} = 0$, otherwise, d_{ij} is calculated using the semantic distance defined in (6) and (13). N_1 and N_2 denote the numbers of training pairs with correspondence and multi-level semantic relevance, respectively. N_3 and N_4 denote the numbers of intra-modal training pairs used to learn the probabilistic membership functions for X and Y modalities, respectively.

E. Optimization

In (17), $L_{\mathbf{W},\phi,\psi}$ is convex with respect to each model parameter. Therefore, we optimize the loss function alternatively

on the local projections and the probabilistic membership functions, until a local optimal solution is achieved. The objective function in (17) is decomposed into two convex subproblems. First, fixing $\{\phi, \psi\}$, we learn the local projections $\{\mathbf{W}^x, \mathbf{W}^y\}$ with primal-dual coordinate gradient descent. Second, fixing $\{\mathbf{W}^x, \mathbf{W}^y\}$, we learn the probabilistic membership functions $\{\phi, \psi\}$ with gradient descent and line search.

Step 1: Fix $\{\phi, \psi\}$*, optimize* $\{\mathbf{W}^x, \mathbf{W}^y\}$. The loss function of (17) w.r.t. $\{\mathbf{W}^x, \mathbf{W}^y\}$ is rewritten as

$$\begin{aligned} L_{\mathbf{W}} = \; & \frac{1}{2}\sum_k (||\mathbf{W}_k^x||^2 + ||\mathbf{W}_k^y||^2) + \\ & \frac{C_1}{N_1}\sum_{i,i}(\varepsilon_{\mathrm{ii}}^- + \varepsilon_{\mathrm{ii}}^+) + \frac{C_2}{N_2}\sum_{i,j\neq i}(\varepsilon_{\mathrm{ij}}^- + \varepsilon_{\mathrm{ij}}^+) \\ \text{s.t.} \quad & D(x_i, y_j) - d_{\mathrm{ij}} \leq \varepsilon + \varepsilon_{\mathrm{ij}}^+, \varepsilon_{\mathrm{ij}}^+ \geq 0 \\ & d_{\mathrm{ij}} - D(x_i, y_j) \leq \varepsilon + \varepsilon_{\mathrm{ij}}^-, \varepsilon_{\mathrm{ij}}^- \geq 0. \end{aligned} \tag{18}$$

By applying the Karush–Kuhn–Tucker conditions on the Lagrangian $\mathcal{L}$, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{W}_k^x} &= \mathbf{W}_k^x + \sum_{i,j}(\alpha_{\mathrm{ij}}^+ - \alpha_{\mathrm{ij}}^-)\frac{\partial D(x_i, y_j)}{\partial \mathbf{W}_k^x} = 0 \\ \frac{\partial \mathcal{L}}{\partial \mathbf{W}_k^y} &= \mathbf{W}_k^y + \sum_{i,j}(\alpha_{\mathrm{ij}}^+ - \alpha_{\mathrm{ij}}^-)\frac{\partial D(x_i, y_j)}{\partial \mathbf{W}_k^y} = 0 \\ \frac{\partial \mathcal{L}}{\partial \varepsilon_{i,j}^+} &= C_N^{ij} - \alpha_{\mathrm{ij}}^+ - \eta_{\mathrm{ij}}^+ = 0, \eta_{\mathrm{ij}}^+ \varepsilon_{\mathrm{ij}}^+ = 0 \\ \frac{\partial \mathcal{L}}{\partial \varepsilon_{i,j}^-} &= C_N^{ij} - \alpha_{\mathrm{ij}}^- - \eta_{\mathrm{ij}}^- = 0, \eta_{\mathrm{ij}}^- \varepsilon_{\mathrm{ij}}^- = 0 \\ \alpha_{\mathrm{ij}}^+ \alpha_{\mathrm{ij}}^- &= 0 \end{aligned} \tag{19}$$

where $\alpha_{\mathrm{ij}}^+ \geq 0$ and $\alpha_{\mathrm{ij}}^- \geq 0$ denote the *Lagrange* multipliers for positive and negative sides, respectively. η_{ij}^+ and η_{ij}^- denote the *Lagrange* multipliers for $\varepsilon_{\mathrm{ij}}^+ \geq 0$ and $\varepsilon_{\mathrm{ij}}^- \geq 0$, respectively. When $i = j$, we have $C_N^{ij} = \frac{C_1}{N_1}$, otherwise $C_N^{ij} = \frac{C_2}{N_2}$. If $\varepsilon_{\mathrm{ij}}^+ \geq 0$, then $\eta_{\mathrm{ij}}^+ = 0, \alpha_{\mathrm{ij}}^+ = C_N^{ij}$, and $\alpha_{\mathrm{ij}}^- = 0$. If $\varepsilon_{\mathrm{ij}}^- \geq 0$, then $\eta_{\mathrm{ij}}^- = 0$, $\alpha_{\mathrm{ij}}^- = C_N^{ij}$, and $\alpha_{\mathrm{ij}}^+ = 0$.

In the first two equations of (19), $\mathbf{W}_k^x$ and $\mathbf{W}_k^y$ are still involved in $\frac{\partial D(x_i, y_j)}{\partial \mathbf{W}_k^x}$ and $\frac{\partial D(x_i, y_j)}{\partial \mathbf{W}_k^y}$. Therefore, given an intermediate solution $\{\mathbf{W}_k^x(t), \mathbf{W}_k^y(t)\}$, the model can be further optimized by calculating the gradient $G(\mathbf{W}_k^x(t))$ and $G(\mathbf{W}_k^y(t))$ using the support vectors $\alpha_{\mathrm{ij}}^+(t)$ and $\alpha_{\mathrm{ij}}^-(t)$ on the current solution. After the model update, we obtain $\mathbf{W}_k^x(t+1)$ and $\mathbf{W}_k^y(t+1)$. The sub-problem in (18) is minimized until an (local) optimal solution is achieved.

Step 2: Fix $\{\mathbf{W}^x, \mathbf{W}^y\}$*, optimize* $\{\phi, \psi\}$. The sub-problem in (17) with respect to the parameters of the probabilistic membership functions can be represented as

$$\begin{aligned} L_{\phi,\psi} = \; & \frac{C_1}{N_1}\sum_{i,i}(\varepsilon_{\mathrm{ii}}^- + \varepsilon_{\mathrm{ii}}^+) + \frac{C_2}{N_2}\sum_{i,j\neq i}(\varepsilon_{\mathrm{ij}}^- + \varepsilon_{\mathrm{ij}}^+) \\ & + \frac{\beta}{N_3}\sum_{i,j} s_{\mathrm{ij}}^x(\overline{g}_i^x - \overline{g}_j^x)^2 + \frac{\gamma}{N_4}\sum_{i,j} s_{\mathrm{ij}}^y(\overline{g}_i^y - \overline{g}_j^y)^2 \\ \text{s.t.} \quad & D(x_i, y_j) - d_{\mathrm{ij}} \leq \varepsilon + \varepsilon_{\mathrm{ij}}^+, \varepsilon_{\mathrm{ij}}^+ \geq 0 \\ & d_{\mathrm{ij}} - D(x_i, y_j) \leq \varepsilon + \varepsilon_{\mathrm{ij}}^-, \varepsilon_{\mathrm{ij}}^- \geq 0. \end{aligned} \tag{20}$$

Again, by checking the *Lagrangian*, only the empirical training pairs with non-zero support vectors will contribute to the gradient $G(\phi(t))$ and $G(\psi(t))$. Therefore, the support vectors should also be identified before gradient calculation. Based on the gradient, a line search on step size for gradient descent is performed using *Armijo* rule. Since the sub-problem in (20) is convex, it is minimized until an (local) optimal solution is achieved.

Initialization: We use CCA to project the original features into 200-dim representations on the two datasets (see Section VI). The projections learned by CCA are used as the initialization of $\{\mathbf{W}^x, \mathbf{W}^y\}$. We conduct K-means on each modality, and the cluster centers are used as initialized values of $\{\phi, \psi\}$. The retrieval performances are not sensitive to the initializations of K-means. The initializations of probabilistic membership functions are further jointly optimized with local projections towards predefined semantic distances, thus the learning process compromises the stochasticity of K-means on the learned model. Moreover, the experiments show that the retrieval performance is stable against different initialization of probabilistic membership functions using K-means.

VI. Experiments

We conduct extensive experiments to compare $\mathtt{TINA}_p$ and $\mathtt{TINA}_d$ to state-of-the-art approaches on image-to-text and text-to-image retrieval tasks using the following datasets.

NUS-WIDE [48] consists of 269 648 images and the associated tags collected from Flickr. We represent the images with 4096-dim features, which are the $6\mathrm{th}$ layered outputs of deep convolutional neural network trained on ImageNet dataset [37]. The 1000-dim TF-IDF tag vectors are treated as the textual representations and 81-dim category indicator vectors are treated as ground-truth class labels. After removing all the images without tags or textual descriptions, we have 79 659 image-text pairs for training, 5000 for parameter validation, and 48 550 for test.

ICML-Challenge[1] contains 100 000 images and their corresponding textual descriptions. We choose 100 frequently occurred categories from all the tags as label information. We also represent the images with 4096-dim features of the $6\mathrm{th}$ layered output of CNN and represent the texts with 5000-dim TF-IDF tag vectors. We randomly select 10 000 for training, 5 000 for validation and the rest for test.

Evaluation criteria: We adopt mean average precision (MAP) and normalized discount cumulative gain (NDCG) for

[1]"ICML-Challenge dataset," [Online]. Available: https://www.kaggle.com/c/challenges-in-representation-learning-multi-mo dal-learning/data

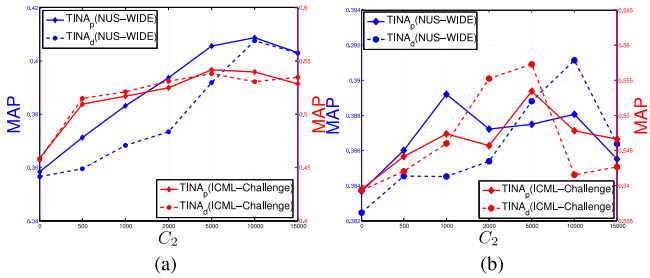


Fig. 6. Parameter validation on hierarchy semantics of $\texttt{TINA}_p$ and $\texttt{TINA}_d$. (a) The influence of hierarchy semantics on image-to-text retrieval. (b) The influence of hierarchy semantics on text-to-image retrieval.

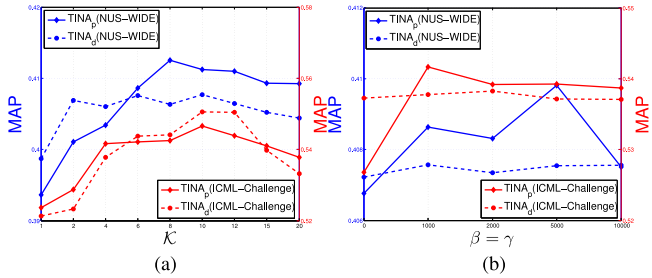


Fig. 7. Parameter validation on $\mathcal{K}$, β and γ of TINA. (a) Parameter sensitivity on the number of local projections $\mathcal{K}$ on image-to-text retrieval task. (b) Parameter sensitivity on local projection consistency with respect to the weight β and γ.

performance evaluation. MAP is widely accepted evaluation paradigm. To measure the performance on data with multi-level semantic relevance, we adopt NDCG as a complementary criterion, defined as

$$\text{NDCG}\,(K) = \frac{1}{N^K} \sum_{j=1}^{K} \frac{2^{\text{Rel}(j)} - 1}{\log(1+j)} \tag{21}$$

where N^K is a normalization constant to ensure that the performance of idealized top K ranking is 1. K is called a truncation or threshold level. In our evaluation, the relative gain (i.e., $Rel(j)$) for the jth cross-modal document is the hierarchical semantic similarity $\exp(-\frac{d_{\text{ij}}}{2\sigma_d^2})$ of query i and the jth document.

Compared approaches: We compare the following methods: (1) PLS: Partial Least Square [4]; (2) CFA: Cross-modal Factor Analysis [21]; (3) CCA: Canonical Correlation Analysis [3]; (4) SCCA: Sparse Canonical Correlation Analysis [30]; (5) MCCA: Multi-view CCA [20]; (6) IMH: Inter-Media Hashing [8]; (7) MMNN: Multi-modal neuro-networks [22]; (8) DCCA: Deep canonical correlation analysis [6].

For MMNN, we set document pairs as the similar pairs when their labels are the same, and pairs with different labels as the dissimilar pairs. For fair comparison, we ignore the binarization step for hashing methods IMH and MMNN. For deep models, the number of layers is set to 2 for MMNN and 3 for DCCA. For the compared methods, we conduct a validation process to find an optimal setting for other parameters.

A. Influence of Hierarchical Semantics

We evaluate the influence of hierarchy semantics on retrieval performance. We vary the value of C_2 from 0 to 15 000 and set $C_1 = 1000$. Other parameters are fixed as $\beta = 1000$, $\gamma = 1000$, $\mathcal{K} = 6$ and $d = 8$. MAP at 100 is reported on the validation sets of NUS-WIDE and ICML-Challenge on image-to-text and text-to-image retrieval tasks (see Fig. 6). The results show that $\texttt{TINA}_d$ and $\texttt{TINA}_p$ obtain the best performances on NUS-WIDE when $C_2 = 10\,000$. On ICML-Challenge, the best performances are obtained when $C_2 = 5000$. The experiment results indicate that by appropriately setting the relative importance of the empirical training data pairs with hierarchical semantic relation, our methods achieve better semantic consistency in cross-modal retrieval.

B. Number of Local Projections

We evaluate the performance of TINA on the number of local projections $\mathcal{K}$. We set $C_1 = 1000, \beta = 1000, \gamma = 1000$, and the dimension of the comparable space as $d = 8$. C_2 is set to 10 000 on NUS-WIDE and 5000 on ICML-Challenge. MAP at 100 is reported on the validation sets of NUS-WIDE and ICML-Challenge on image-to-text retrieval task. As shown in Fig. 7(a), $\texttt{TINA}_p$ and $\texttt{TINA}_d$ achieve the highest performances on NUS-WIDE at $\mathcal{K} = 8$ and 10, respectively. They both obtain the highest performances on ICML-Challenge at $\mathcal{K} = 10$. The performances of our models reach the peak by appropriately setting the number of local projections. When $\mathcal{K}$ is small, our models tend to be smoother, achieving smaller variance by increasing the model bias. When $\mathcal{K}$ is large, our models better adapt to the cross-modal data, but they may also suffer from high over-fitting risk. Similar observations are also obtained on text-to-image task on both datasets.

The results show that TINA achieves a better trade-off between global projection and sample specific projection by setting an appropriate number of local projections. Moreover, heavier computation burden is required using a larger number of local projections. By jointly considering the effectiveness and efficiency, we set $\texttt{TINA}_p$ and $\texttt{TINA}_d$ with $\mathcal{K} = 10$ on both NUS-WIDE and ICML-Challenge datasets for image-to-text and text-to-image retrieval tasks compared to state-of-the-art.

C. Local Projection Consistency

We evaluate the influence of imposing local projection consistency on local expert aggregation. For both $\texttt{TINA}_p$ and $\texttt{TINA}_d$, we set $d = 4, C_1 = 1000, \mathcal{K} = 6$ on both datasets. We set $C_2 = 10\,000$ on NUS-WIDE and $C_2 = 5000$ on ICML-Challenge. We report the performances of MAP at 100 on the validation sets of NUS-WIDE and ICML-Challenge on image-to-text task using different β and γ, as shown in Fig. 7(b). From the results we can see that, the performance of $\texttt{TINA}_p$ is more sensitive with the values of β and γ, while the performance of $\texttt{TINA}_d$ is quite stable. Similar observations are also obtained on text-to-image task on both datasets.

This phenomenon can be explained by the fact that local projection consistency is imposed to enforce the intra-modal consistency on the probabilistic membership functions. $\texttt{TINA}_p$ aggregates local projections via independent probabilistic membership functions to produce a representation in each modality,

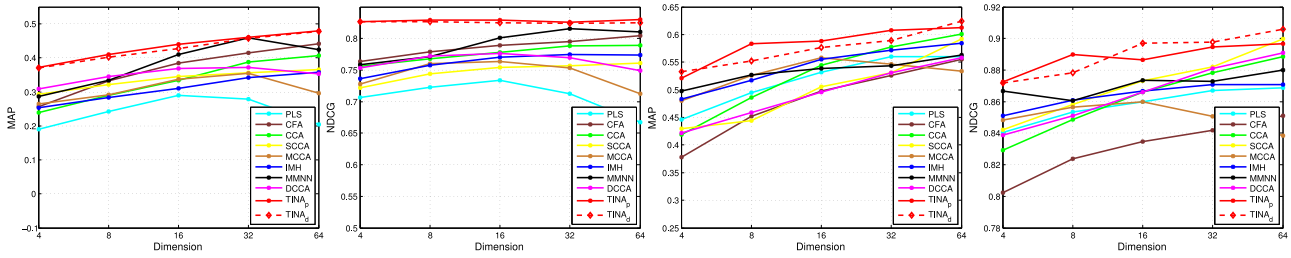


Fig. 8. MAP at 100 and NDCG at 100 of image-to-text retrieval. (a) Image-to-text on NUS-WIDE. (b) Image-to-text on NUS-WIDE. (c) Image-to-text on ICML-Challenge. (d) Image-to-text on ICML-Challenge.

TABLE II
PERFORMANCE OF IMAGE-TO-TEXT RETRIEVAL

Dataset	Criterion	Methods									
		PLS	CFA	CCA	SCCA	MCCA	IMH	MMNN	DCCA	$\texttt{TINA}_p$	$\texttt{TINA}_d$
NUS-WIDE	MAP	28.9479	44.2198	40.6219	36.5851	35.45	35.7789	45.929	37.2243	**48.0109**	47.8563
	NDCG	73.3785	80.4543	78.9201	76.134	76.3537	77.4256	81.5363	77.6613	**82.9665**	82.6816
ICML-Challenge	MAP	56.0114	55.5095	60.0985	59.1933	55.8392	58.4467	56.3314	55.7737	61.2548	**62.4447**
	NDCG	86.8839	85.1029	88.8429	89.9808	85.9758	87.0757	87.9841	89.0969	89.6556	**90.5727**

while $\texttt{TINA}_d$ aggregates the distances of the local projection pairs via multiplication of probabilistic membership functions from two modalities. Compared to $\texttt{TINA}_p$, the membership function of $\texttt{TINA}_d$ has less degree of freedom. Therefore, the influence of local projection consistency on $\texttt{TINA}_p$ tends to be more significant.

Furthermore, the probabilistic membership functions are also learned to minimize the empirical loss, i.e., the semantic coherence measurement. When β and γ in $\texttt{TINA}_p$ are set with a relatively large value (e.g., 1000), their contributions to the final performance have been sufficiently emphasized. For $\texttt{TINA}_d$, the performance is also slightly improved when using a moderate setting of β and γ, e.g., $\beta, \gamma = 1000$ on NUS-WIDE and $\beta, \gamma = 2000$ on ICML-Challenge. Therefore, local projection consistency regularizes the local expert aggregation behavior, thus it improves the cross-modal retrieval performance. For $\texttt{TINA}_p$, we set $\beta = 5000, \gamma = 5000$ on NUS-WIDE and $\beta = 1000, \gamma = 1000$ on ICML-Challenge in the subsequent experiments. For $\texttt{TINA}_d$, we set $\beta = 1000, \gamma = 1000$ on NUS-WIDE and $\beta = 2000, \gamma = 2000$ on ICML-Challenge in the subsequent experiments.

D. Image-to-Text Retrieval

We evaluate the performance of all the compared approaches on image-to-text retrieval. We set $C_1 = 1000, C_2 = 10\,000, \mathcal{K} = 10$ on NUS-WIDE and $C_1 = 1000, C_2 = 5000, \mathcal{K} = 10$ on ICML-Challenge for both $\texttt{TINA}_p$ and $\texttt{TINA}_d$. The optimal settings of β and γ have been described in Section VI-C. We conduct correlation learning on different dimensions of the projected space with $d = \{4, 8, 16, 32, 64\}$ on NUS-WIDE and ICML-Challenge. The experimental results are shown in Fig. 8 in terms of MAP at 100 and NDCG at 100. From the curves, we can see that the performances of $\texttt{TINA}_p$ and $\texttt{TINA}_d$ keep pace with each other on NUS-WIDE. On ICML-Challenge dataset, the performances of both methods show fluctuated rising trends on both evaluation criteria when increasing the dimensions. $\texttt{TINA}_p$ and $\texttt{TINA}_d$ consistently outperform other approaches on both datasets. By increasing the dimension, the performances of CFA, CCA, SCCA and IMH are monotonously increased. However, the performance curves of PLS, MCCA, MMNN and DCCA show obvious knee points on NUS-WIDE, as shown in Fig. 8(a) and (b). The performance degradation at larger d indicates that these models may be suffering from over-fitting.

Our approaches show larger performance gains over other approaches at a lower dimensional space. In other words, our approaches possess better model capacity even with smaller d. It can be explained by the local property of our approaches, where each local expert adapts to a subset of similar data. Therefore, a lower dimension is sufficient to encode the cross-modal correlation for $\texttt{TINA}_p$ and $\texttt{TINA}_d$. In Fig. 8(b), the performances at low dimensions of our approaches are comparable to or even slightly better than the performances at larger d. The results further demonstrate that by constructing local projections and the adaptive projection aggregation mechanisms, $\texttt{TINA}_p$ and $\texttt{TINA}_d$ better adapt to the content divergence and multi-level semantic relation with parsimonious output dimensions. The best performances of all the methods at the optimal setting of d are recorded in Table II. Our models achieve the best performances among all the compared approaches. $\texttt{TINA}_p$ performs the best on NUS-WIDE dataset, while $\texttt{TINA}_d$ performs the best on ICML-Challenge dataset. Nevertheless, the performance gap between $\texttt{TINA}_d$ and $\texttt{TINA}_p$ under any experimental condition is small.

E. Text-to-Image Retrieval

We evaluate the performance of text-to-image retrieval on both datasets. The parameter setting of our methods is the same as Section VI-D. The experimental results are shown in Fig. 9

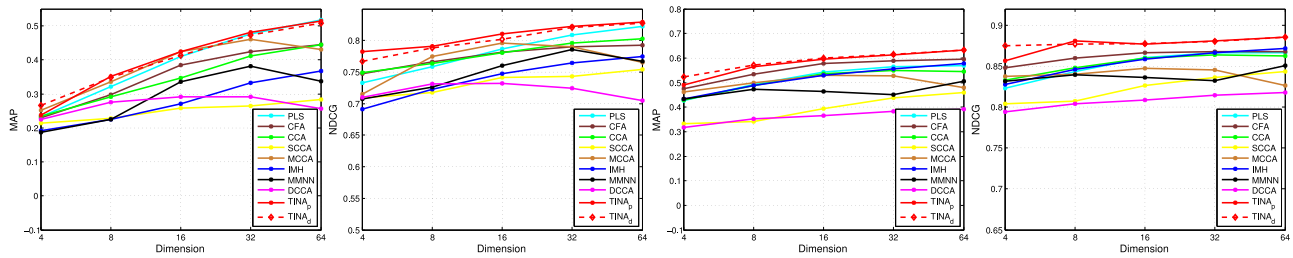


Fig. 9. MAP at 100 and NDCG at 100 of text-to-image retrieval. (a) Text-to-image on NUS-WIDE. (b) Text-to-image on NUS-WIDE. (c) Text-to-image on ICML-Challenge. (d) Text-to-image on ICML-Challenge.

TABLE III
PERFORMANCE OF TEXT-TO-IMAGE RETRIEVAL

Dataset	Criterion	Methods									
		PLS	CFA	CCA	SCCA	MCCA	IMH	MMNN	DCCA	$\texttt{TINA}_p$	$\texttt{TINA}_d$
NUS-WIDE	MAP	**51.7249**	44.5136	44.4325	28.3257	46.0192	36.7323	38.0984	29.0859	51.4209	50.7368
	NDCG	82.2573	79.2376	80.2675	75.4297	79.5794	77.5016	78.5467	73.2316	**82.9159**	82.716
ICML-Challenge	MAP	56.8252	59.4224	54.9115	45.8497	52.8668	57.7358	50.5538	39.2238	**63.2246**	63.1879
	NDCG	86.6611	86.7746	86.3571	84.3486	84.7416	87.1779	85.0462	81.7827	**88.5263**	88.5217

in terms of MAP at 100 and NDCG at 100. From the curves, we can see that $\texttt{TINA}_p$ and $\texttt{TINA}_d$ generally outperform other approaches on both datasets except when $d = 64$ in Fig. 9(a), where PLS achieves the highest performance in MAP. However, its NDCG is worse than those of our methods, see Fig. 9(b). Our approaches obtain remarkable performance gains in NDCG on both datasets. The results indicate that by encoding multi-level semantic relation, TINA achieves better semantic coherence in cross-modal retrieval.

Different from the observation that promising image-to-text retrieval performance is achieved at smaller d, our approaches achieve better performances at larger d on text-to-image retrieval task. This can be explained by the fact that the content divergence in visual modality is more significant. It requires high dimensions to encode the content and semantic information of the retrieval database in text-to-image retrieval task. The best performances of all the methods at the optimal setting of d are recorded in Table III. Our models achieve the best performances among all approaches on three out of four conditions. $\texttt{TINA}_p$ performs the best on both datasets in terms of NDCG, and performs the best on ICML-Challenge in MAP. PLS performs the best on NUS-WIDE in MAP, but its NDCG under-performs both $\texttt{TINA}_p$ and $\texttt{TINA}_d$. The performance gap between $\texttt{TINA}_d$ and $\texttt{TINA}_p$ under any experimental condition is still very small.

F. Findings and Discussions

The learned local experts: We illustrate the selectivity of local experts with respect to $g_i^x(k), k = 1, \ldots, \mathcal{K}$ for TINA_p on NUS-WIDE in Fig. 10. The model is learned with $\mathcal{K} = 4$, $d = 64, C_1 = 1000, C_2 = 10\,000, \beta = 1000$ and $\gamma = 1000$. By analyzing $g_i^x(k)$, we find that a dominant local expert ($\mathbf{W}_1^x$ in Fig. 10) exists which is selected by all the images with non-zero

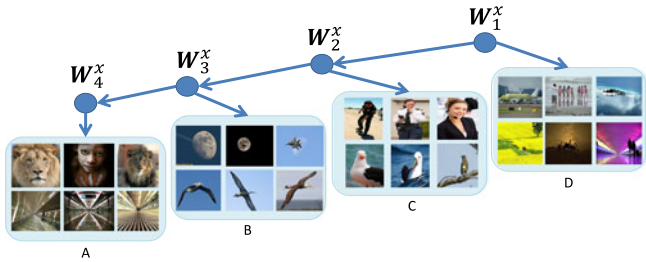


Fig. 10. Illustration of local projection effect on NUS-WIDE. The figure shows the selectivity of TINA_p (with $\mathcal{K} = 4$) on local projections for images. Images are projected by the local projections at their parent node and the upper parent node, i.e., images in group A are projected by all the four local experts, and images in group D are projected only by $\mathbf{W}_1^x$.

weights. Then only a part of images will be associated to $\mathbf{W}_2^x$ and so on so forth. Different subsets of images with different levels of semantic and visual similarity are projected by different local experts, while the most similar subset is associated to the local projection ($\mathbf{W}_4^x$) at the bottom node in Fig. 10. The example shows that the correlations among different image subsets with semantic and visual consistency are captured in different local experts. The more specific semantics the image delivers, the more local experts will be involved to produce a better aggregated projection. Consequently, our model is more adaptive to content and correlation divergence.

On the compared approaches: In Figs. 8 and 9 and Tables II and III, the performances of PLS, CFA, CCA and MCCA are diversified with respect to different tasks, datasets and performance measurements. The results indicate that simple global correlation models are unstable to deal with cross-modal content divergence. Despite of using simple intra-class similarity and inter-class correspondence, MMNN and IMH under-perform $\texttt{TINA}_p$ and $\texttt{TINA}_d$, and sometimes even under-perform PLS

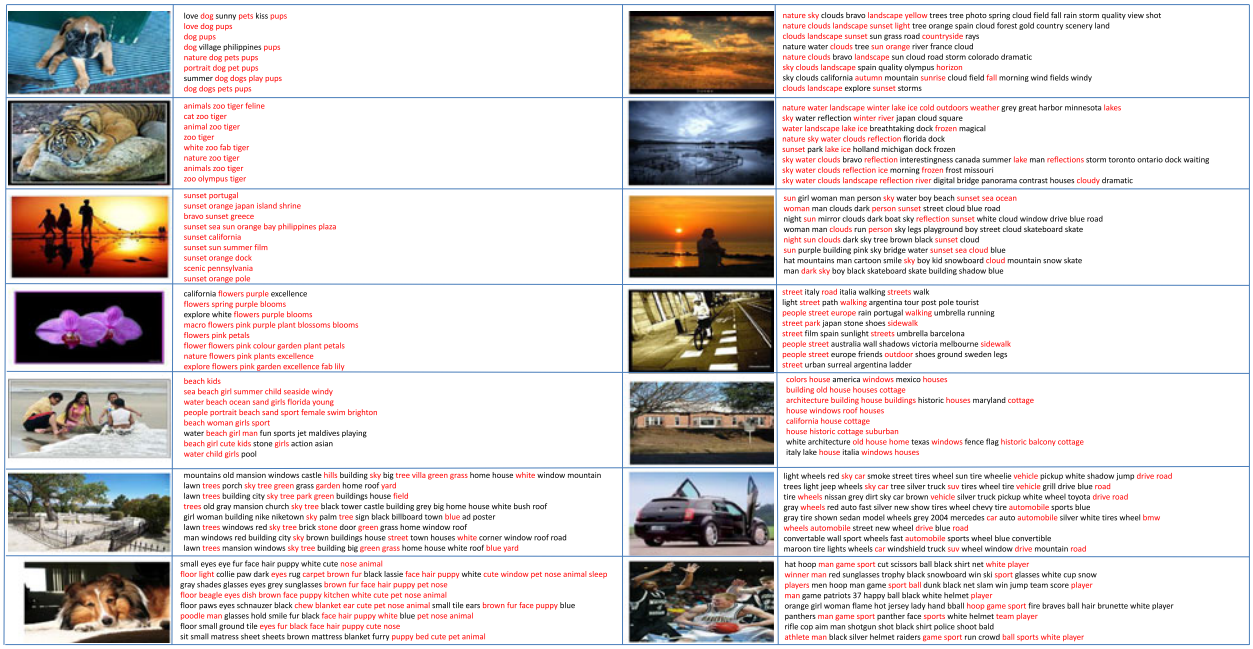


Fig. 11. Examples of top eight results of image-to-text retrieval on NUS-WIDE and ICML-Challenge. Each row of text denotes a retrieved textual document. The words in red text are strongly relevant to the query images.

and CCA. The observation indicates that multi-level semantic modeling is necessary for correlation learning on real data with a large number of categories. SCCA performs poorly on image-to-text retrieval, and even worse on text-to-image retrieval. The observation can be explained by the fact that unilateral sparse constraint imposed on textual modality may not be suitable for both sparse visual and textual features used in the experiments. Deep CCA does not always outperform other CCA-based methods. The reason may be that visual features are the output of deep CNN and they turn out to be sparse. It is likely to over-fit by using the stacked subspace on sparse visual and textual features. Besides, deep CCA does not consider the intra-modal similarity and multi-level semantic relation in correlation learning, so that it may suffer from the conflict brought by correlation information insufficiency and large model capacity. TINA outperforms other approaches because: 1) the learned semantic coherence representations by encoding multi-level semantic relation; 2) the local projections that better adapt to the content divergence.

The robustness of TINA *:* The optimal values of parameters C_2, $\mathcal{K}$, β and γ in our model are not sensitive to the target datasets, i.e., the optimal setting on NUS-WIDE dataset is also a reasonable setting for model learning on ICML-Challenge dataset. A universally acceptable parameter setting for experiments on other datasets can be $C_2 = 5000$, $\mathcal{K} = 10$ and $\beta = \gamma = 1000$. In Figs. 8 and 9, we see that the performances of our approaches are kept at a comparatively high level with respect to different settings of d, retrieval task types and the performance measurements. Our model achieves a better trade-off between model bias and variance by constructing local experts on real data. Moreover, TINA demonstrates more noise tolerance on real applications, since both NUS-WIDE and ICML-Challenge datasets contain a certain level of noise. For example, there are about 10% incorrect tag information on NUS-WIDE.

Retrieval examples: Some top results of image-to-text and text-to-image retrieval obtained by TINA are shown in Figs. 11 and 12 on NUS-WIDE and ICML-Challenge, respectively. In Fig. 11, red marked words are semantically relevant descriptions to the visual queries. The top left five examples of NUS-WIDE are with relatively less textual descriptions than the right five examples. Take the two challenging "sunset" queries in the third row as an example, each of the top eight retrieved texts contains one or more relevant "sunset" words to the visual queries. Similar phenomenon can be observed on other examples. Our approach returns satisfied results on visual queries with single salient object (the 1st, 3rd, 12th and 13th in top-down and left-to-right orders), multi salient objects (the 5th, 9th and 14th), natural scenes (the 2nd, 4th, 6th, 8th, 10th and 11th) and even artifact images (the 7th). The observations further validate the fact that TINA better adapts to cross-modal content divergence.

In Fig. 12, the first five examples are from NUS-WIDE and the last five from ICML-Challenge. Take the first row as an example, when querying with text on "tiger," TINA returns the nearly perfect semantically coherent results, where the top eight results are all about "tiger" images. At the second row, using the "Cat fun action jump play" as query text, the 1st and the 2nd returned items are perfectly matched images, while the 3rd∼7th returned images are all about jumping people. However, "animal" and "human" are relevant semantic topics, so that "jumping person" images are more acceptable results for the textual query than other irrelevant images. We can see that almost all the illustrated examples are semantically matched textual queries and retrieved images from topics such as "animal," "human," "sports," "event," "scene," "building,"

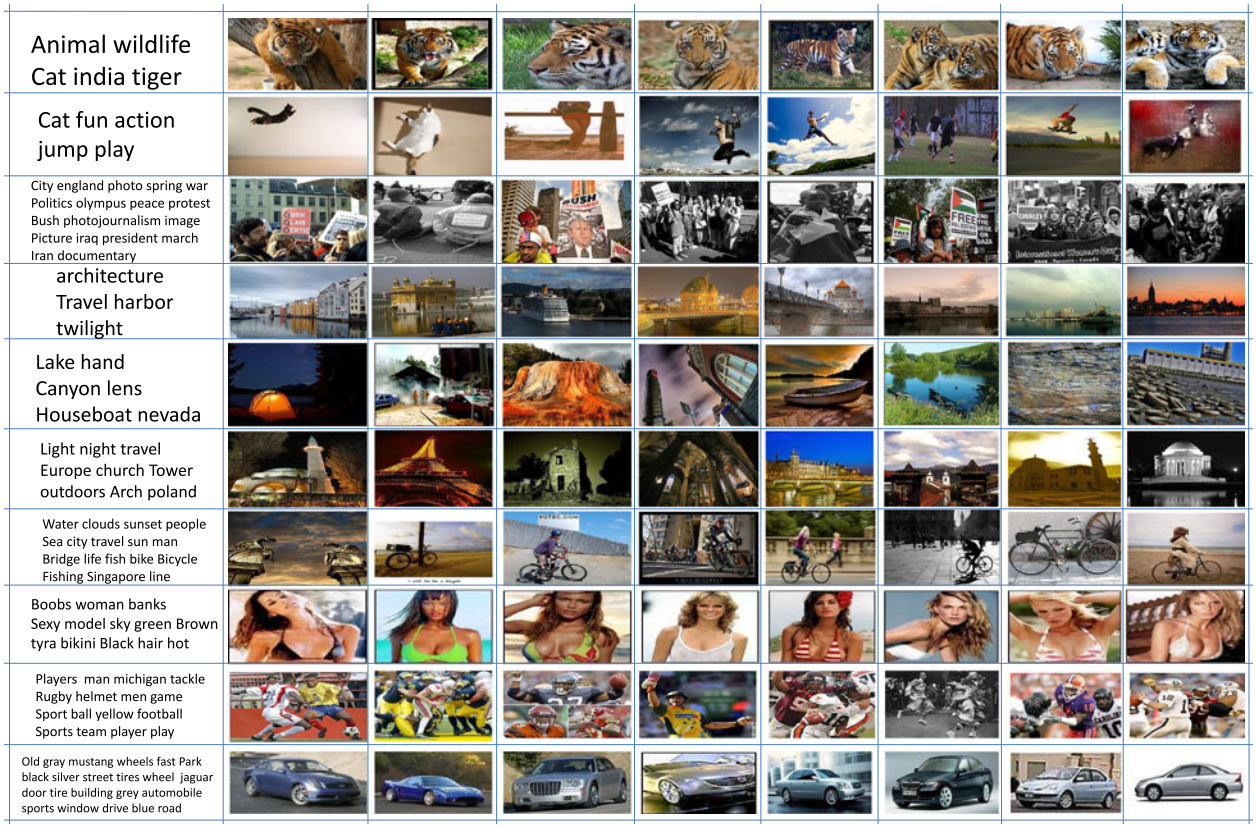


Fig. 12. Examples of top eight results of text-to-image retrieval on NUS-WIDE and ICML-Challenge.

etc. The retrieved images contain richer visual information but consistent semantic meanings, leading to remarkable user experience compared to near-duplicated content retrieved by traditional content-based approaches. The observation further indicates that the semantic hierarchy ensures that the retrieved results are more consistent with human perception.

VII. Conclusion

We propose TINA, a cross-modal correlation learning method by adaptive hierarchical semantic aggregation. Our approach utilizes cross-modal training data with different levels of semantic relation. We propose two strategies for local expert aggregation, i.e., local projection aggregation and local distance aggregation. The structure risk objective function that involves semantic coherence measurement, local projection consistency and the complexity penalty of local projections is optimized. Experiments on two large scale cross-modal datasets demonstrate that TINA achieves a better semantic coherence by effectively adapting to the content divergence and complicated semantic relation. In future work, we will study domain specific feature extraction mechanism (e.g., the stacked convolution layers), or combine stacked auto-encoders with the localized correlation learning.

References

[1] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 251–260.
[2] C. G. Snoek *et al.*, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 975–986, Aug. 2007.
[3] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.
[4] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection*. New York, NY, USA: Springer-Verlag, 2006, pp. 34–51.
[5] X. Chen, H. Liu, and J. G. Carbonell, "Structured sparse canonical correlation analysis," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 199–207.
[6] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[7] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2160–2167.
[8] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 785–796.
[9] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3594–3601.
[10] H. Zhang, Y. Zhuang, and F. Wu, "Cross-modal correlation learning for clustering on image-audio dataset," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 273–276.
[11] F. Wu *et al.*, "Sparse multi-modal hashing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.
[12] Z. Yu *et al.*, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. ACM SIGIR 37th Int. Conf. Res. Develop. Inform. Retrieval*, 2014, pp. 395–404.
[13] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.
[14] L.-J. Li, C. Wang, Y. Lim, D. M. Blei, and L. Fei-Fei, "Building and using a semantivisual image hierarchy," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3336–3343.
[15] J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 785–792.
[16] T. Deselaers and V. Ferrari, "Visual and semantic similarity in ImageNet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1777–1784.

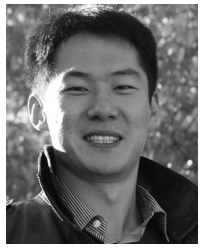
[17] K. Grauman, F. Sha, and S. J. Hwang, "Learning a tree of metrics with disjoint visual features," in *Proc. Adv. Neural Inform. Process. Syst.*, 2011, pp. 621–629.
[18] N. Verma, D. Mahajan, S. Sellamanickam, and V. Nair, "Learning hierarchical similarity metrics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2280–2287.
[19] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.
[20] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2012.
[21] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 604–611.
[22] J. Masci, M. Bronstein, A. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, Apr. 2014.
[23] J. Masci, A. M. Bronstein, M. M. Bronstein, P. Sprechmann, and G. Sapiro, "Sparse similarity-preserving hashing," *CoRR*, 2013. [Online]. Available: http://arxiv.org/abs/1312.5479
[24] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 689–696.
[25] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proc. Adv. Neural Inform. Process. Syst.*, 2012, pp. 2222–2230.
[26] A. Karpathy, A. Joulin, and F.-F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inform. Process. Syst.*, 2014, pp. 1889–1897.
[27] Y. Peng, D. Zhang, and J. Zhang, "A new canonical correlation analysis algorithm with local discrimination," *Neural Process. Lett.*, vol. 31, no. 1, pp. 1–15, 2010.
[28] T. Sun and S. Chen, "Locality preserving cca with applications to data visualization and pose estimation," *Image Vis. Comput.*, vol. 25, no. 5, p. 531–543, 2007.
[29] T. Sun, S. Chen, J. Yang, and P. Shi, "A novel method of combined feature extraction for recognition," in *Proc. Int. Conf. Data Mining*, 2008, pp. 1043–1048.
[30] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Mach. Learn.*, vol. 83, no. 3, pp. 331–353, 2011.
[31] C. Archambeau and F. R. Bach, "Sparse probabilistic projections," in *Proc. 21th Annu. Conf. Neural Inform. Process. Syst.*, 2009, pp. 73–80.
[32] S. Virtanen, A. Klami, and S. Kaski, "Bayesian cca via group sparsity," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 457–464.
[33] Y. Yang *et al.*, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.
[34] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, vol. 22, no. 1, pp. 1360–1365.
[35] Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," in *Proc. ACM SIGKDD 18th Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 940–948.
[36] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.
[37] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[38] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3594–3601.
[39] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 248–255.
[40] M. Marszałek and C. Schmid, "Constructing category hierarchies for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2008, pp. 479–491.
[41] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, "Unsupervised discovery of visual object class hierarchies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.
[42] M. Mazloom, E. Gavves, and C. G. M. Snoek, "Conceptlets: Selective semantics for classifying video events," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2214–2228, Dec. 2014.
[43] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.
[44] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Comput. Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.
[45] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness," in *Proc. 18th Int. Joint Conf. Artif. Intell.*, 2003, vol. 3, pp. 805–810.
[46] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th Int. Joint Conf. Artif. Intell.*, 1995, pp. 448–453.
[47] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. 32nd Annu. Meet. Assoc. Comput. Linguistics*, 1994, pp. 133–138.
[48] T.-S. Chua *et al.*, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Conf. Image Video Retrieval*, 2009, Article 48, pp. 1–9.
[49] Y. Yang, F. Nie, S. Xiang, Y. Zhuang, and W. Wang, "Local and global regressive mapping for manifold learning with out-of-sample extrapolation," in *Proc. 25th Amer. Assoc. Artif. Intell. Conf.*, 2010, pp. 649–654.
[50] E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders, "Convex reduction of high-dimensional kernels for visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3610–3617.
[51] X. Lu *et al.*, "A low rank structural large-margin method for cross-modal ranking," in *Proc. ACM SIGIR 36th Int. Conf. Res. Develop. Inform. Retrieval*, 2013, pp. 433–442.

Yan Hua received the B.S. degree in communication engineering from China Agricultural University, Beijing, China, in 2010, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2015.

She is currently an Assistant Professor with the Communication University of China, Beijing, China. Her research interests include semantic image analysis, image and text retrieval, and visual attention.

Shuhui Wang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in China, 2012.

He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include large-scale Web data mining, visual semantic analysis, and machine learning.



Siyuan Liu received the Ph.D. degree in computer science and engineering from the Hong Kong University of Science and Technology, Hong Kong, China, in 2011, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2014.

He is currently an Assistant Professor with Pennsylvania State University, State College, PA, USA. He is also an Adjunct Research Fellow with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include mobile data analytics and heterogeneous social networks mining.

Anni Cai received the Ph.D. degree from the University of California, Santa Barbara, CA, USA, in 1988.

She is currently a Professor with the Beijing University of Posts and Telecommunications, Beijing, China. She has published 6 monographs, and authored or coauthored about 70 academic papers in international journals and conferences. Her research interests include multimedia communication, cognitive computing of visual information, video search technologies, and pattern recognition.

Qingming Huang received the B.S. degree in computer science and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the University of Chinese Academy of Sciences, Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 300 academic papers in international journals and conferences. His research interests include multimedia content analysis, image processing, computer vision, pattern recognition, and machine learning.