TINA: Cross-modal Correlation Learning by Adaptive Hierarchical Semantic Aggregation

Yan (Tina) Hua*, Shuhui Wang†, Siyuan Liu‡, Qingming Huang†§, Anni Cai*
*Beijing University of Posts and Telecommunications, Beijing, China, huayan@bupt.edu.cn, annicai@bupt.edu.cn
†Key Lab of Intell. Info. Process. (CAS), Inst. of Comput. Tech., CAS, Beijing, China, wangshuhui@ict.ac.cn
‡Heinz College, Carnegie Mellon University, siyuan@cmu.edu
§University of Chinese Academy of Sciences, Beijing, China, qmhuang@ucas.ac.cn

***Abstract*—With the explosive growth of web data, effective and efficient technologies are in urgent needs for retrieving semantically relevant contents of heterogeneous modalities. Previous studies construct global transformations to project the heterogeneous data into a measurable subspace. However, global projections cannot appropriately adapt to diverse contents, and the naturally existing multi-level 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 queries. 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 multi-level semantic relevance. Second, with a set of local linear projections aggregated by gating functions, we optimize the structure risk objective function that involves semantic coherence measurement, local projection consistency and the complexity penalty of local projections. Therefore, semantic coherence and a better bias-variance trade-off can be achieved by TINA. Extensive experiments on widely used NUS-WIDE and ICML-Challenge datasets demonstrate that TINA outperforms state-of-the-art, and achieves better adaptation to the multi-level semantic relation and content divergence.**

***Keywords*-Cross-modal retrieval; Semantic hierarchy; Local correlation learning.**

I. Introduction

The multi-modal data refer to contents from heterogeneous modalities describing the same or relevant topics, *e.g.*, web images and their surrounding texts, video frames and their accompanied voice messages. When online users input queries for seeking complementary information of other modalities about certain topics, the retrieval results that are ranked according to the semantic relevance are expected. Therefore, the need for effective and efficient cross-modal retrieval techniques has arisen along with the proliferation of multi-modal data and diversified user demands.

Previous text-based techniques compare the similarities of the textual query and the surrounding texts of web images, while they suffer from the mismatch between textual descriptions and web images. Semantic-based techniques rely on an intermediate risky semantic annotation process, which in turn delivers the cross-modal retrieval problem into a "chicken-egg" dilemma. A goal-oriented solution routine is to transform the heterogeneous modalities into measureable low-dimensional representations, hence semantically similar cross-modal documents can be directly retrieved as nearest neighbor search. Canonical Correlation Analysis (CCA) [15] and its variants [6] [12] [28] which conduct dimensionality reduction by maximizing the cross-modal correlation, have been the workhorse in cross-modal retrieval [25]. Correlation learning is investigated with various models such as regression [26], graph embedding [29] and boosting [5]. In this paper, we address the challenges that hinder existing correlation learning models from real applications.

First, existing approaches learn the cross-modal statistical dependencies by inter-modal correspondence [15] and intra-modal similarity (dissimilarity) [21] [22]. However, there exists multi-level semantic relation among large scale real data. For example, 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. A natural way to represent the multi-level semantic relation is hierarchical category structures [11] [20], which can serve as knowledge ontology to boost the end tasks such as image retrieval [8] [10] and recognition [16] [20] [32]. However, existing semantic structures [11] [13] are not constructed towards cross-modal correlation learning.

To achieve semantic coherent retrieval, an appropriate multi-level semantic structure should be constructed to model the general-to-specific semantic relationship [20] on multi-modal data. In this paper, we propose to build an adaptive semantic hierarchy by integrating multi-modal content and semantic information. Based on the hierarchy, we organize the cross-modal data into training pairs with multi-level semantic relevances. Accordingly, we model the multi-level semantics by large margin bilateral constraints like Support Vector Regression (SVR). The distances between the training pairs are optimized towards their semantic relevances, which makes the learned distance measure better reflect the

multi-level semantics of cross-modal data.

Besides the multi-level semantics, real world multi-modal data are diversified in content. Existing global transformation strategies [15] [26] are not well adapted to cluttered data distribution. By exploiting the local property in single modality, localized expert strategies [37] [36] [35] are effective to deal with the intra-class divergence in classification, instance search [36] and manifold learning [35]. However, how local experts are adapted to content divergence and multi-level semantics has not been well addressed in correlation learning. In this paper, we construct local experts by learning multiple projections. The projected representations are softly weighted and aggregated by gating functions. The parameters of the gating functions are optimized to minimize the empirical loss and maximize the local projection consistency, so that each learned local projection will play a dominant role in constructing correlation among a subset of similar and semantically relevant data. Consequently, more robustness and consistency can be achieved.

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 gating functions are constructed for both modalities. Our key contributions include:

- We encode the multi-level sematic relevance by a large margin regression framework into the cross-modal distance. To our best knowledge, our work is the first to study semantic coherence in cross-modal retrieval.
- We construct a semantic hierarchy for cross-modal retrieval with joint modeling of visual, textual and ontology similarities. It appropriately encodes the relation of cross-modal documents from both content and semantic perspectives.
- We propose a structure risk objective function to learn the local projections which are probabilistically aggregated by gating functions. Our model better adapts to the cross-modal content divergence and multi-level semantic relation, then semantic coherent retrieval can be performed by simple ranking with the distances.
- Extensive cross-modal 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

The aim of correlation learning is to construct measureable representations on heterogeneous modalites. Existing works can be categorized as follow.

The *subspace learning* learns a pair of transformations to project data into a measureable low dimensional subspace. CCA [15] and its variants [6] [14] provide direct solutions. Partial least square (PLS) [26] formulates the problem with a bilateral regression model. As a supervised extension of CCA, Sharma *et al.* [28] proposed a generalized multi-view analysis model to learn (non-)linear subspace using label information. A boosting based hashing method is proposed in [5]. Graph based methods [18] [29] encode intra-modal similarity and inter-modal co-occurrence into a unified graph representation.

The subspace learning is closely related with the *probabilistic graphical methods*, when we build connections between subspaces and latent topics. Archambeau *et al.* [1] and Virtanen *et al.* [33] have provided Bayesian interpretation of CCA-based models. Correspondence LDA (Corr-LDA) [4] captures the topic-level relations between images and texts. The model of [17] can be seen as Markov random field over LDA topic model. Zhen *et al.* [34] developed a latent binary embedding approach.

To deal with the diversified content, the *complicated function learning* has been investigated thenceforth. Wang *et al.* [19] proposed a locally aligned multi-view transformation approach. Deep structure has been applied to correlation learning recently. Galen *et al.* [12] proposed Deep CCA to learn complex nonlinear transformations. The multi-modal auto-encoders [24] and multi-modal restricted Boltzmann machines [30] are used as building blocks for shared representation learning. Masci *et al.* [21] [22] constructed multi-layered neuro-networks with both intra-modal similarity and inter-modal correlation. Cross-modal topic classifiers [25] are constructed on the CCA representations, and map heterogeneous data into a unified semantic space.

TINA models the hierarchical semantic relation in cross-modal data which are neglected by the above-mentioned approaches on feature level or semantic level. The learned local projections are endowed with better adaptation to complicated semantic relations.

B. Modeling Semantic Hierarchy

Semantic hierarchy [11] is a formally defined taxonomy or ontology structure in natural language processing. It has been used to other domains, such as image and multimedia [9]. It can be general or domain specific. WordNet [11] defines a general lexical database for English language. A large scale image database ImageNet [9] has been constructed by collecting images for each semantic concept in WordNet. Based on the inter-category classification confusion [13], 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.* [20] integrated tags to automatically build the "semantivisual" hierarchy, which encodes the general-to-specific semantic and visual relationship. Marszalek *et al.* [23] constructed the class hierarchies that postpone decisions in the presence of uncertainty. Sivic [31] proposed to automatically discover a hierarchical structure from an unlabeled image collection.

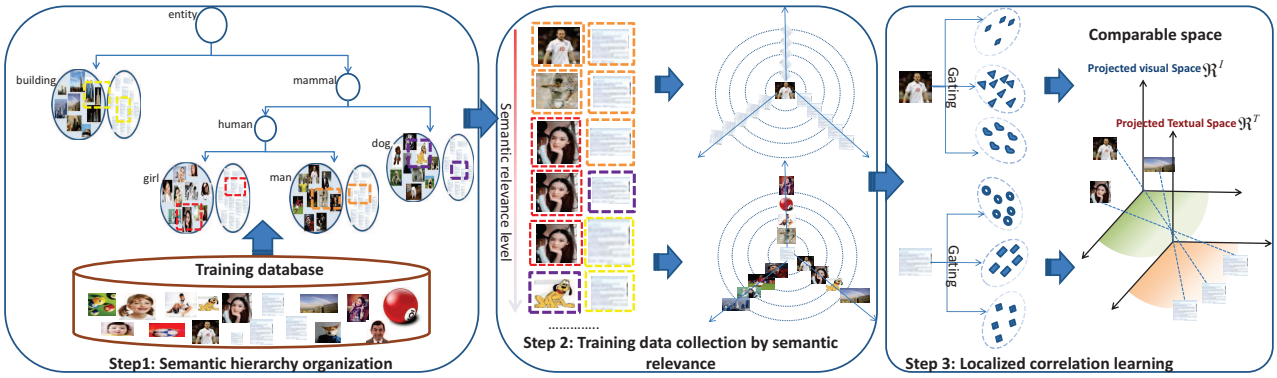


Figure 1. The framework of TINA. First, we build an adaptive semantic hierarchy by top-down hierarchical clustering in Step 1. Based on the semantic hierarchy, a set of training pairs are sampled from the database, where each document is associated with the cross-modal documents ranked by their semantic relevances, as shown in Step 2. By optimizing a structure risk objective function, we learn a set of local linear projections and gating functions for semantic coherent cross-modal retrieval, 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 *et al.* [10] computed the semantic distance between images by measuring the divergence of concept distribution of their neighborhoods. Deng *et al.* [8] developed a semantic vector representation, and constructed a hierarchical bilinear similarity function based on the pairwise semantic affinity. Verma *et al.* [32] associated the separated visual similarity metric for every concept in the hierarchy, and the metrics are learned jointly through hierarchical aggregation for nearest neighbor classifiers. A tree of metrics is learned by Hwang *et al.* [16], which imposes the appropriate (dis)similarity constraints among its subtree members. However, to the best of our knowledge, the hierarchical semantics have not been well-studied for cross-modal correlation learning. In this paper, to deal with multi-level semantics, we propose a large margin regression framework for local projection learning. Our method better adapts to the cross-modal content divergence and multi-level semantic relation by sub-model aggregation.

III. Problem and Solution Framework

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

For example, on NUS-WIDE data (Figure 2), given an image query of "statue", the idealized retrieval result should be: the top ranked document is the corresponding textual description of the query image, 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, the proposed TINA consists of the following key steps (see Figure 1):

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

Step 2: Training data collection (Section V-A). 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 on the selected data pairs for the subsequent model training.

Step 3: Localized correlation learning (Section V). We utilize the training pairs with multi-level semantic relations collected in Step 2, and construct the large-margin regularized regression learning framework. With a set of local projections and gating functions learned for both modalities, semantic coherent cross-modal retrieval can be easily performed by nearest neighbor search on the projected representations.

We begin by introducing the method of semantic hierarchy organization in the next section.

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 the semantic category hierarchy $\mathbb{H}$ on $\mathbb{D}$ by combining the similarity modeling from visual domain, textual domain and the ontology relatedness.

A. Visual Similarity

The appearances of visual categories are divergent. The intra-class visual divergence is even larger than that of inter-class. We develop a simple and effective method based on visual subcategory similarity. First, we use K-means to divide category c into k_c subcategories, $X_{c,k} = \{x_i\}, c_i =$

$c, k = 1, ..., k_c$, where each subcategory contains images with more visual cohesiveness. Then, the 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:

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

where σ_v is a bandwidth parameter that controls the sensitivity to the distance range. The $\min$ rule identifies the similarity between categories with their most similar images. As a result, the scheme guarantees that the category level visual similarity is robust to uncontrolled variations brought by light, angle and occlusion.

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 [11] [3] [2] [27]. In our ontology similarity computation, we adopt the normalized similarity with the depth of their parent node [39]. The ontology similarity matrix is denoted as S_o.

D. Tree Hierarchy

We linearly combine the above three similarities to get the semantic similarity matrix:

$$S = \alpha_1 S_v + \alpha_2 S_t + \alpha_3 S_o \quad (4)$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1, \alpha_1, \alpha_2, \alpha_3 \geq 0$, which are optimally tuned on a validation set. Based on the combined similarity matrix, we seek to exploit a new hierarchical category structure $\mathbb{H}$ in a top-down partition fashion. 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. Compared to WordNet, the hierarchy structure $\mathbb{H}$ is more balanced and coherent in semantic and feature level. Two examples are shown in Figure 2 and Figure 3, where the hierarchies are built on NUS-WIDE ($n = 3$) and ICML-Challenge ($n = 5$), respectively.

In Figure 2 and 3, the semantic hierarchies we build are consistent in both content and semantic perspectives.

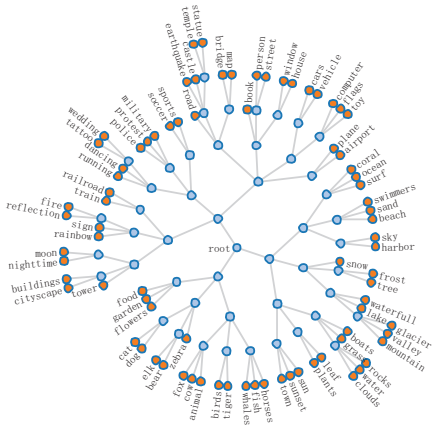


Figure 2. Semantic hierarchy on NUS-WIDE with $n = 3$.

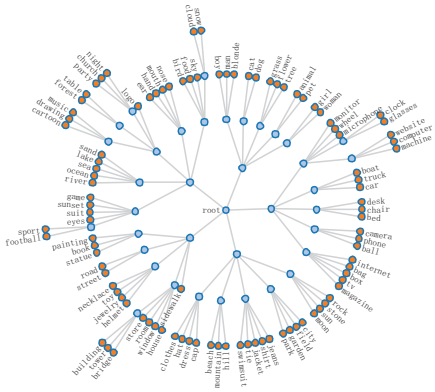


Figure 3. Semantic hierarchy on ICML-Challenge with $n = 5$.

For example, in Figure 2, "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.*, furry and gray. "zebra" with obvious stripes is the sibling node of the parent of "elk" and "bear". In Figure 3, "boat", "truck" and "car" are transportations. "desk", "chair" and "bed" are furnitures. They belong to instruments, so that their parent nodes are siblings of each other.

On a hierarchy tree, the length of shortest path does not appropriately indicate the specificity of ontology [39] [8]. Intuitively, the depth of parent node of two concepts represents the category specificity [39]. To encode such knowledge, we define a normalized distance of two nodes on $\mathbb{H}$ as follows:

$$h(c_1, c_2) = \frac{p(c_1, c_2)}{L_h \cdot d_e(p_a(c_1, c_2))} \quad (5)$$

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, and L_h is the length of the longest path on $\mathbb{H}$. The larger distance between the two categories is, they will be more semantically distinct. The normalized category distance is used to calculate the *semantic distance* d_{ij} in each empirical training pair, as will be introduced in Section V and Eqn. 10.

V. Correlation Learning

A. Model

Given the cross-modal data $\mathbb{D}$, and its semantic hierarchy tree $\mathbb{H}$, we want to learn local linear transformations f_x and f_y to project cross-modal data into a comparable space, where the correlations of cross-modal data are consistent with their relations on $\mathbb{H}$, so that the *semantic coherence* of cross-modal retrieval is better achieved.

Local linear projection. We learn a set of local projection functions $\mathbf{W} = \{\mathbf{W}_k^x \in \mathbb{R}^{d\times d_x}, \mathbf{W}_k^y \in \mathbb{R}^{d\times d_y}\}_{k=1}^{\mathcal{K}}$. x_i and y_j are projected using the $\mathcal{K}$ local projections as:

$$f_x(x_i) = \sum_k g_i^x(k)\mathbf{W}_k^x x_i,\ f_y(y_j) = \sum_k g_j^y(k)\mathbf{W}_k^y y_j \quad (6)$$

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 local projection aggregation weights for $f_x(x_i)$ and $f_y(y_j)$, define as:

$$g_i^x(k) = \frac{\exp(\phi_k^\top x_i)}{\sum_{k'}\exp(\phi_{k'}^\top x_i)},\ g_j^y(k) = \frac{\exp(\psi_k^\top y_j)}{\sum_{k'}\exp(\psi_{k'}^\top y_j)} \quad (7)$$

where $\{\phi_k \in \mathbb{R}^{d_x}\}$ and $\{\psi_k \in \mathbb{R}^{d_y}\}$ are parameters of the kth gating function. Then we define distance in the projected space as:

$$D(x_i, y_j) = ||f_x(x_i) - f_y(y_j)||^2 \quad (8)$$

There are several other types of projection functions. The unified correlation model [15] lacks the adaptability to content divergence and complicated semantic relation. The sample specific projection [35] learns one projection for each data point, which involves intensive computational burdens, and is sensitive when the tangent structure is influenced by outliers. Our local linear projection achieves a better bias-variance trade-off between the two existing approaches. The local projections are probabilistically aggregated by gating functions that are adaptively fit to each cross-modal training datum. Consequently, the complicated non-linear cross-modal semantic relation can be well approximated by our local linear projections.

Local projection consistency. We impose local projection consistency on the gating functions g_i^x and g_j^y, which manifests that the adjacent and semantically related data should possess similar gating values. The consistency measurement is calculated by *intra-modal similarity* as:

$$\begin{aligned} L_\phi = \textstyle\sum_{ij} s_{ij}^x (g_i^x - g_j^x)^2, L_\psi = \textstyle\sum_{ij} s_{ij}^y (g_i^y - g_j^y)^2 \\ s_{ij}^x = \exp(-\tfrac{d_{ij}}{2\sigma_d^2})\exp(-\tfrac{(x_i - x_j)^2}{2\sigma_x^2}) \\ s_{ij}^y = \exp(-\tfrac{d_{ij}}{2\sigma_d^2})\exp(-\tfrac{(y_i - y_j)^2}{2\sigma_y^2}) \end{aligned} \quad (9)$$

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. By imposing the 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 similar in feature representation as they are "four-leg" in shape and fluffy, and they are discussed as pets. They are also semantically relevant on both WordNet and $\mathbb{H}$. By optimizing the local projection consistency in Eqn. 9, the correlations among documents of two similar categories will be encoded by some specific local experts, where local projection selectivity is learned and performed by the gating functions.

Additionally, to deal with multi-label that widely existing in real-world data, we revise the *semantic distance* as:

$$d_{ij} = min_{x_i(y_i)\in c_1, x_j(y_j)\in c_2} h(c_1, c_2) \quad (10)$$

where $h(c_1, c_2)$ is the distance of concepts c_1 and c_2 on $\mathbb{H}$ as in Eqn. 5.

Semantic coherence measurement. The learned cross-modal distance on training data $\mathbb{D}$ is expected to be consistent with their *semantic distance* on hierarchy $\mathbb{H}$. The consistency is considered to be matched if their absolute differences on all training pairs are less than ε, a predefined tolerance. Unfortunately, not all the training pairs perfectly satisfy the semantic consistency in practical situations. Therefore, we introduce two slack variables for each training pair to measure the inconsistency out of ε. The relaxed formulation is defined as:

$$\begin{aligned} D(x_i, y_j) - d_{ij} \le \varepsilon + \varepsilon_{ij}^+, \quad \varepsilon_{ij}^+ \ge 0 \\ d_{ij} - D(x_i, y_j) \le \varepsilon + \varepsilon_{ij}^-, \quad \varepsilon_{ij}^- \ge 0 \end{aligned} \quad (11)$$

where d_{ij} is the *semantic distance* of x_i and y_j as in Eqn. 10. ε_{ij}^+ and ε_{ij}^- are the slack variables of the positive side and negative side, respectively. The constraint in Eqn. 11 is similar in spirit with Support Vector Regression. The correlation measures among data pairs are optimized towards the multi-level semantic distances on $\mathbb{H}$. A possible alternative is to employ the relative measurement such as ranking SVM loss, which involves a semantic relevance comparison among different training pairs, and thus results in a more complicated model solution.

Loss function. With the local projection consistency defined in Eqn. 9 and semantic coherence measurement defined in Eqn. 11, we jointly learn local projections $\{\mathbf{W}^x, \mathbf{W}^y\}$ and gating functions $\{\phi, \psi\}$ by minimizing the following structure risk objective function:

$$\begin{aligned} & L_{\mathbf{W},\phi,\psi} = \tfrac{1}{2}\textstyle\sum_k \left(||\mathbf{W}_k^x||^2 + ||\mathbf{W}_k^y||^2\right) \\ & + \tfrac{C_1}{N_1}\textstyle\sum_{i,i}(\varepsilon_{ii}^- + \varepsilon_{ii}^+) + \tfrac{C_2}{N_2}\textstyle\sum_{i,j\neq i}(\varepsilon_{ij}^- + \varepsilon_{ij}^+) \\ & + \tfrac{\beta}{N_3}\textstyle\sum_{i,j} s_{ij}^x (g_i^x - g_j^x)^2 + \tfrac{\gamma}{N_4}\textstyle\sum_{i,j} s_{ij}^y (g_i^y - g_j^y)^2 \\ & s.t.\ \ D(x_i, y_j) - d_{ij} \le \varepsilon + \varepsilon_{ij}^+, \quad \varepsilon_{ij}^+ \ge 0 \\ & \qquad d_{ij} - D(x_i, y_j) \le \varepsilon + \varepsilon_{ij}^-, \quad \varepsilon_{ij}^- \ge 0 \end{aligned} \quad (12)$$

Algorithm 1 Model Optimization of TINA

Initialize $\mathbf{W}^x, \mathbf{W}^y, \phi, \psi, T_1, T_2, T_3, l \leftarrow 0$;
repeat
  Step 1: Fix $\{\phi, \psi\}$**, optimize** $\{\mathbf{W}^x, \mathbf{W}^y\}$**;**
  $t \leftarrow 0$;
  repeat
    Find $\alpha_{ij}^+(t) > 0$ and $\alpha_{ij}^-(t) > 0$;
    Compute the gradient $G(\mathbf{W}^x(t)), G(\mathbf{W}^y(t))$;
    $\mathbf{W}^x(t+1) = \mathbf{W}^x(t) - \lambda_w^t \cdot G(\mathbf{W}^x(t))$;
    $\mathbf{W}^y(t+1) = \mathbf{W}^y(t) - \lambda_w^t \cdot G(\mathbf{W}^y(t))$;
    $t \leftarrow t+1$;
  until $t > T_1$
  Step 2: Fix $\{\mathbf{W}^x, \mathbf{W}^y\}$**, optimize** $\{\phi, \psi\}$**;**
  $t \leftarrow 0$;
  repeat
    Find $\alpha_{ij}^+(t) > 0$ and $\alpha_{ij}^-(t) > 0$;
    Compute the gradient $G(\phi(t))$ and $G(\psi(t))$;
    Find the stepsize λ_g^t with *Armijo* linear search;
    $\phi(t+1) = \phi(t) - \lambda_g^t \cdot G(\phi(t))$;
    $\psi(t+1) = \psi(t) - \lambda_g^t \cdot G(\psi(t))$;
    $t \leftarrow t+1$;
  until $t > T_2$
  $l \leftarrow l+1$
until $l > T_3$

where the first term is the complexity penalty of local projections to avoid over-fitting. ε_{ii}^+ and ε_{ii}^- are the slack variables for the cross-modal pair with correspondence, where x_i and y_i are the complementary description to each other. ε_{ij}^+ and ε_{ij}^-, $i \neq j$, are for the multi-level semantic relevance. And their weights are adjusted by C_1 and C_2. In the constraints, if $i = j$, $d_{ij} = 0$, otherwise, d_{ij} is calculated using the semantic distance defined in Eqn. 5 and 10. 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 gating functions for X and Y modalities, respectively.

$L_{\mathbf{W},\phi,\psi}$ is convex with respect to each model parameter. Therefore, the model can be solved by alternating optimization, until a local optimal solution is achieved.

B. Optimization

We optimize the loss function in Eqn. 12 alternatively on the local projections and the gating functions. The overall objective function in Eqn. 12 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 gating 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 Eqn. 12 w.r.t. $\{\mathbf{W}^x, \mathbf{W}^y\}$ is rewritten as:

$$\begin{aligned} L_{\mathbf{W}} = & \tfrac{1}{2}\textstyle\sum_k(||\mathbf{W}_k^x||^2 + ||\mathbf{W}_k^y||^2) + \\ & \tfrac{C_1}{N_1}\textstyle\sum_{i,i}(\varepsilon_{ii}^- + \varepsilon_{ii}^+) + \tfrac{C_2}{N_2}\textstyle\sum_{i,j\neq i}(\varepsilon_{ij}^- + \varepsilon_{ij}^+) \\ s.t.\ \ & D(x_i, y_j) - d_{ij} \leq \varepsilon + \varepsilon_{ij}^+, \varepsilon_{ij}^+ \geq 0 \\ & d_{ij} - D(x_i, y_j) \leq \varepsilon + \varepsilon_{ij}^-, \varepsilon_{ij}^- \geq 0 \end{aligned} \tag{13}$$

By applying the Karush-Kuhn-Tucker (KKT) conditions on the Lagrangian $\mathcal{L}$, we have:

$$\begin{aligned} \tfrac{\partial \mathcal{L}}{\partial \mathbf{W}_k^x} &= \mathbf{W}_k^x + \textstyle\sum_{i,j}(\alpha_{ij}^+ - \alpha_{ij}^-)\tfrac{\partial D(x_i,y_j)}{\partial \mathbf{W}_k^x} = 0 \\ \tfrac{\partial \mathcal{L}}{\partial \mathbf{W}_k^y} &= \mathbf{W}_k^y + \textstyle\sum_{i,j}(\alpha_{ij}^+ - \alpha_{ij}^-)\tfrac{\partial D(x_i,y_j)}{\partial \mathbf{W}_k^y} = 0 \\ \tfrac{\partial \mathcal{L}}{\partial \varepsilon_{i,j}^+} &= C_N^{ij} - \alpha_{ij}^+ - \eta_{ij}^+ = 0,\ \eta_{ij}^+\varepsilon_{ij}^+ = 0 \\ \tfrac{\partial \mathcal{L}}{\partial \varepsilon_{i,j}^-} &= C_N^{ij} - \alpha_{ij}^- - \eta_{ij}^- = 0,\ \eta_{ij}^-\varepsilon_{ij}^- = 0 \\ & \alpha_{ij}^+\alpha_{ij}^- = 0 \end{aligned} \tag{14}$$

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

In the first two equations of Eqn. 14, $\mathbf{W}_k^x$ and $\mathbf{W}_k^y$ are still evolved 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 $\alpha_{ij}^+(t)$ and $\alpha_{ij}^-(t)$ on the current solution. After the model update, we get $\mathbf{W}_k^x(t+1)$ and $\mathbf{W}_k^y(t+1)$. Based on this update rule, the sub-problem in Eqn. 13 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 Eqn. 12 with respect to the gating functions can be represented as:

$$\begin{aligned} L_{\phi,\psi} = & \tfrac{C_1}{N_1}\textstyle\sum_{i,i}(\varepsilon_{ii}^- + \varepsilon_{ii}^+) + \tfrac{C_2}{N_2}\textstyle\sum_{i,j\neq i}(\varepsilon_{ij}^- + \varepsilon_{ij}^+) \\ & + \tfrac{\beta}{N_3}\textstyle\sum_{i,j} s_{ij}^x(g_i^x - g_j^x)^2 + \tfrac{\gamma}{N_4}\textstyle\sum_{i,j} s_{ij}^y(g_i^y - g_j^y)^2 \\ s.t.\ \ & D(x_i, y_j) - d_{ij} \leq \varepsilon + \varepsilon_{ij}^+, \quad \varepsilon_{ij}^+ \geq 0 \\ & d_{ij} - D(x_i, y_j) \leq \varepsilon + \varepsilon_{ij}^-, \quad \varepsilon_{ij}^- \geq 0 \end{aligned} \tag{15}$$

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 the *Armijo* rule. Since the sub-problem in Eqn. 15 is convex, it is minimized until an (local) optimal solution is achieved. The whole optimization procedure is shown in Algorithm 1.

Complexity. The complexity in computing the gradient $\{G(\mathbf{W}^x(t)), G(\mathbf{W}^y(t))\}$ is $O\left(\widehat{N}\mathcal{K}d(d_x + d_y)\right)$, where $\widehat{N}$ is the average number of support vectors. The complexity in computing the gradient $\{G(\phi(t)), G(\psi(t))\}$ is $O\left((\widehat{N} + 2N_3)\mathcal{K}d_x + (\widehat{N} + 2N_4)\mathcal{K}d_y\right)$. In the iterative optimization process of TINA, see Algorithm 1, the model can be sufficiently optimized by setting $T_1 = 20, T_2 = 5, T_3 = 15 \sim 20$.

VI. EXPERIMENTS

We conduct extensive experiments to compare TINA to state-of-the-art approaches on image-to-text and text-to-image retrieval tasks using the following datasets:

NUS-WIDE [7] consists of 269,648 images and the associated tags collected from Flickr. We represent the images by 500-dim bag-of-visual-word on SIFT. 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, 10,000 for parameter validation, and 43,550 for test.

ICML-Challenge [38] contains 100,000 images and their corresponding textual descriptions. We choose 100 frequently occurred categories from all the tags as label information. We represent the images with 10752-dim spatial pyramid with sparse coding on SIFT, 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). MAP is widely accepted evaluation paradigm. To measure the performance on data with multi-level semantic relevance, we adopt NDCG as a complementary criterion. NDCG is defined as:

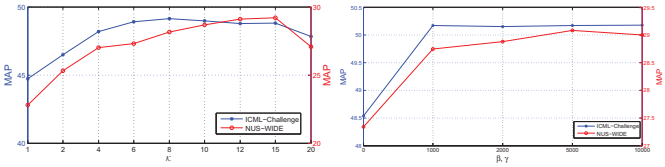
$$NDCG@K = \frac{1}{N^K} \sum_{j=1}^{K} \frac{2^{Rel(j)} - 1}{log(1+j)} \tag{16}$$

where N^K is a normalization constant to ensure that the idealized top K ranking of the query 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 hierarchy semantic similarity $\exp(-\frac{d_{ij}}{2\sigma_d^2})$ of query i and the jth document.

Compared approaches. We compare the following methods: (1) PLS: Partial Least Square [26]; (2) CCA: Canonical Correlation Analysis [15]; (3) SCCA: Sparse Canonical Correlation Analysis [14]; (4) GMLDA: Generalized Multi-view LDA [28]; (5) IMH: Inter-Media Hashing [29]; (6) MMNN: Multi-Modal Neuro-Networks [21]; (7) DeepCCA: Deep Canonical Correlation Analysis [12].

For GMLDA, the category of a training pair is provided with a randomly selected category from its multi-class labels. 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 DeepCCA.

For all the compared approaches except SCCA, we first use CCA to transform the original features into 150-dim representations on NUS-WIDE and 200-dim representations on ICML-Challenge. The projections learned by CCA are used for initializations of $\{\mathbf{W}^x, \mathbf{W}^y\}$. We conduct K-means on each modality, and the cluster centers are used as initializations of $\{\phi, \psi\}$. For other parameters of all the methods, we conduct a validation process to find an optimal setting.



(a) The number of local projections $\mathcal{K}$.

(b) The weights β, γ of local projection consistency.

Figure 4. Parameter validation on $\mathcal{K}, \beta, \gamma$ of TINA.

Experiment environment. Experiments are conducted on 2 standard desktop computers (Windows), with Intel (R) Processor I7-4770K (8M Cache, 3.50 GHz, 4 cores), 32 GB main memory and 7,200RPM hard disks. Our model is implemented on C++ platform.

A. The Number of Local Projections

We evaluate the performance of TINA on the number of local projections $\mathcal{K}$. We set $C_1 = 1000, C_2 = 1000, \beta = 1000, \gamma = 1000$, and the dimension of the comparable space $d = 8$. MAP@100 is reported on the validation sets of NUS-WIDE and ICML-Challenge on image-to-text task. As shown in Figure 4.(a), the MAP performances achieve the highest on NUS-WIDE when $\mathcal{K} = 15$, and on ICML-Challenge when $\mathcal{K} = 8$. Similar observations are also obtained on text-to-image task on both datasets. This shows that TINA achieves a better trade-off between global projection and sample specific projection by setting an appropriate number of local projections. However, the larger the number of local projections is, the heavier computation burden becomes. Considering the effectiveness and efficiency, we set TINA with $\mathcal{K} = 15$ on NUS-WIDE and $\mathcal{K} = 8$ on ICML-challenge in the subsequent experiments for both image-to-text and text-to-image tasks.

B. Local Projection Consistency

We evaluate the influence of applying local projection consistency on local projection aggregation. We set $d = 4, C_1 = 1000, C_2 = 1000$, and report the performances of MAP@100 on the validation sets of NUS-WIDE and ICML-challenge on image-to-text task in Figure 4.(b). From the results we can see that, the performance becomes better with larger values of β and γ, but the performance becomes stable when β and γ are larger than 1000. Similar observations are also obtained on text-to-image task on both datasets. This may be explained by the fact that the gating functions are also optimized to minimizing the empirical loss, *i.e.*, the semantic coherence measurement. When β and γ are set with a relatively large value (*e.g.*, 1000), their contributions to the final performance have been sufficiently emphasized. Therefore, local projection consistency can also improve

the performance of cross-modal retrieval, which plays an important supplementary role of local projection learning. In the subsequent experiments, we set $\beta = 5000, \gamma = 5000$ for TINA on both datasets.

C. Image-to-text Retrieval

We evaluate the performance of image-to-text retrieval on all the approaches. For TINA, we conduct parameter validation on C_1 and C_2, and find that the correspondence and multi-level semantic relation are equally important for model learning. Therefore, by setting $C_1 = 1000$ and $C_2 = 1000$, a good performance can be guaranteed. 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 Table I in terms of MAP@100 and NDCG@100. TINA generally outperforms others except IMH ($d = 64$) on NUS-WIDE by MAP, which means that TINA achieves the best results at a lower dimensional space. This can be explained by exploring the local property of TINA, where each local projection is responsible for only a subset of similar data. Therefore, using a lower dimension is appropriate to capture the correlation of the data subset.

When performances are measured with NDCG, our approach outperforms all the other approaches. The performance gains are even more significant when using a lower dimension d. The results further demonstrate that by constructing local projections and the adaptive projection aggregation mechanism, TINA is more adaptive to the content divergence and multi-level semantic relation with parsimonious output dimensions.

D. Text-to-image Retrieval

For text-to-image retrieval, the validation experiments show that the cross-modal correspondence is more important than multi-level semantic relation for TINA. Therefore, we set $C_1 = 1000, C_2 = 200$ on NUS-WIDE and $C_1 = 1000, C_2 = 500$ on ICML-Challenge. The experimental results are shown in Table II. Our model outperforms other approaches under all the settings. Different from image-to-text retrieval, the best performances of TINA are achieved on larger d, which can be explained by the fact that the content divergence in visual modality is significantly larger than textual modality. Therefore, it requires more dimensions to encode the information of the retrieval database in text-to-image retrieval task.

When performances are measured with MAP on text-to-image retrieval, the performance gains of TINA compared to others are more significant on NUS-WIDE dataset. When performances are measured with NDCG, TINA also outperforms all the other approaches, and the performance gains on NUS-WIDE are also more significant.

Figure 5. Examples of top 5 results on image-to-text on ICML-Challenge (left column) and NUS-WIDE (right column). Each row of text denotes a retrieved textual document. The red marked words are strongly relevant to the query images.

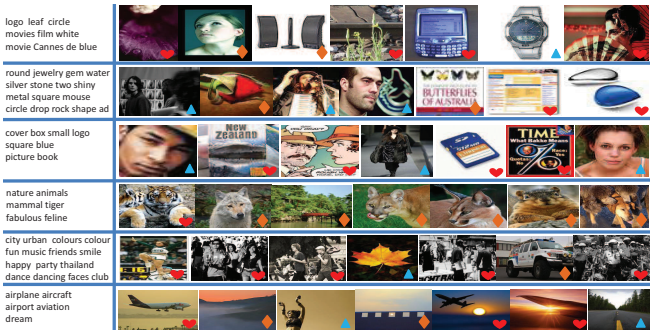


Figure 6. Examples of top 7 results on text-to-image on ICML-Challenge (the top three rows) and NUS-WIDE (the bottom three rows). The red heart shapes marked on the images represent perfectly matched images of the textual query, and the brown prisms represent relevant images, and the blue triangles represent "somewhat" relevant images.

E. Findings and Discussions

On the compared approaches. In Table I and II, the performances of PLS and CCA are diversified with respect to different tasks, datasets and performance measures. The results indicate that the overly simple global correlation models are not capable of dealing with the content divergence. Despite of using simple intra-class similarity and inter-class correspondence, the performances of GMLDA, MMNN and IMH under-perform TINA, and sometimes even under-perform PLS 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 well on image-to-text at low-dimension subspace, but poorly on text-to-image. The performance inconsistency can be explained by the unilateral sparse constraint imposed on textual modality. TINA outperforms other approaches because: 1) the *learned comparable space* by using multi-level semantic relation; 2) the *local linear projections* that better fit the content divergence.

The robustness of TINA. From Table I and II, we see that the performances of TINA are kept at a comparatively high level on different settings of d, the types of retrieval task, and the performance measurements. Our model achieves a better trade-off between model bias and variance by constructing the 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.

Table I
THE PERFORMANCE OF IMAGE-TO-TEXT RETRIEVAL

Dataset	Methods	MAP@100					NDCG@100				
		The dimension d					The dimension d				
		4	8	16	32	64	4	8	16	32	64
NUS-WIDE	PLS	14.3525	16.8508	18.8091	26.1502	28.9665	66.7542	68.1409	68.9477	73.4477	74.8207
	CCA	17.2468	19.6656	21.2451	22.5493	25.1861	70.0403	70.4156	70.5933	71.0924	72.7897
	SCCA	27.7999	26.1951	24.463	20.0961	14.5372	72.7534	72.5648	72.1037	70.6252	66.5929
	GMLDA	16.1874	17.6116	22.4797	24.8820	24.8993	69.0176	69.1386	71.6559	71.1641	71.8367
	IMH	16.4702	18.505	21.5829	25.453	**29.4325**	69.2561	70.3686	71.8701	73.8712	76.2639
	MMNN	27.5603	23.5747	26.6726	28.5282	28.1334	73.8093	72.0716	72.7696	73.906	74.0629
	DeepCCA	26.7495	24.3317	23.0248	24.9704	27.7242	72.9711	72.1691	71.9383	72.5075	73.5949
	TINA ($\mathcal{K} = 15$)	**29.3755**	**28.9975**	**28.9244**	**30.7638**	29.0345	**76.8121**	**76.1071**	**76.3339**	**77.2096**	**76.617**
ICML-Challenge	PLS	38.0154	42.2151	44.1336	44.5546	44.8203	81.1398	82.3662	82.8553	82.8834	83.1322
	CCA	37.7149	40.6503	41.1134	42.384	44.0964	81.0545	81.9269	82.2715	82.7781	83.6478
	SCCA	31.1677	33.5838	32.8331	33.9194	36.1676	80.0999	81.2869	80.9868	81.335	81.5961
	GMLDA	31.1029	32.4329	36.3549	38.2388	39.7494	80.0147	80.3078	81.092	82.3956	82.6069
	IMH	29.5054	33.1312	41.185	45.8109	48.9123	78.76	80.3451	81.5703	84.0686	85.2624
	MMNN	39.1176	41.9273	46.5337	48.2685	50.5153	81.4853	82.2668	83.8124	84.5703	85.1592
	DeepCCA	39.8027	42.1485	45.3101	46.7567	48.6981	81.4942	81.9476	83.1349	83.9704	84.3913
	TINA ($\mathcal{K} = 8$)	**51.5461**	**50.7976**	**50.0178**	**50.9461**	**51.3821**	**87.7468**	**86.9828**	**86.6715**	**86.7442**	**87.165**

Table II
THE PERFORMANCE OF TEXT-TO-IMAGE RETRIEVAL

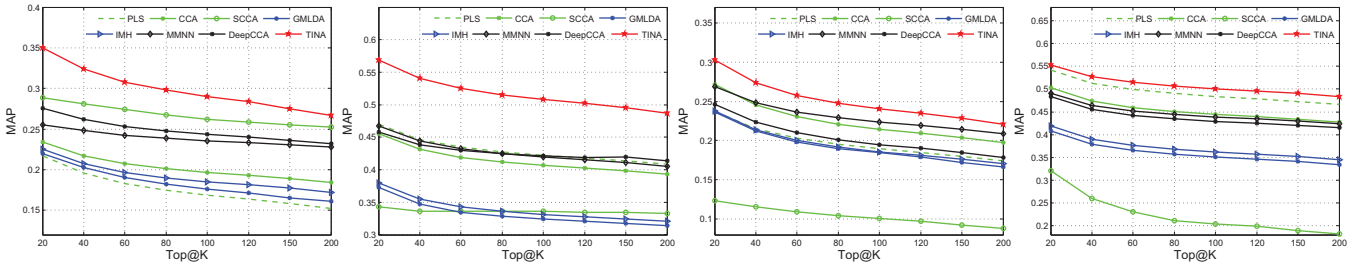
Dataset	Methods	MAP@100					NDCG@100				
		The dimension d					The dimension d				
		4	8	16	32	64	4	8	16	32	64
NUS-WIDE	PLS	16.8807	18.9383	20.922	22.0693	21.9966	68.493	69.0937	69.4116	69.766	69.5334
	CCA	18.2331	21.4304	24.2351	25.0137	25.0741	70.2249	71.1121	71.9574	71.9875	71.8551
	SCCA	11.0013	10.0595	9.27424	9.25333	10.1815	62.4884	61.9359	61.4538	61.3893	62.1687
	GMLDA	18.0606	18.4597	20.4665	20.8271	21.0703	69.0673	69.411	69.7304	69.8903	70.0237
	IMH	16.8332	18.5639	20.2653	21.6111	22.7207	69.467	70.2867	70.8587	71.2365	71.3955
	MMNN	20.2692	22.3745	24.0064	24.7844	25.2179	70.9039	71.593	71.647	71.7685	71.8361
	DeepCCA	20.6593	19.4891	17.6691	18.2436	21.2359	70.7815	70.8711	70.4611	70.5576	70.9103
	TINA ($\mathcal{K} = 15$)	**20.8468**	**24.0288**	**26.3953**	**27.664**	**27.2361**	**71.6631**	**72.1929**	**72.4593**	**72.7026**	**72.0914**
ICML-Challenge	PLS	42.7854	48.4052	51.8847	52.1931	51.5655	82.934	84.4133	85.3549	85.2897	84.9611
	CCA	41.2968	44.4431	44.6885	45.0875	46.0439	82.4765	83.2784	83.3199	83.3468	83.549
	SCCA	20.4633	20.3793	20.6187	20.9811	20.5757	73.6916	73.6748	73.8711	73.6928	73.7016
	GMLDA	34.2822	35.0631	38.3249	40.3751	41.778	80.9744	81.2764	81.889	82.3867	82.5794
	IMH	31.7634	36.1672	46.0654	51.4419	52.3521	79.7613	81.3203	83.9344	85.1238	85.3717
	MMNN	40.3599	43.8567	48.0367	47.0989	51.8498	82.2807	83.2779	84.4745	83.8356	84.9885
	DeepCCA	39.8991	42.9329	45.0924	49.7908	51.1065	81.9029	82.7436	83.1438	84.4462	84.7896
	TINA ($\mathcal{K} = 8$)	**43.4969**	**50.0776**	**54.846**	**55.2927**	**54.3737**	**83.3368**	**85.1637**	**86.3986**	**86.2669**	**85.9262**

MAP@K. The MAP curves of the top 20 to 200 results are plotted in Figure 7 on both datasets, where the dimension d is set to 8 for all the approaches. TINA outperforms other approaches on MAP@20 to MAP@200. Despite that only documents of the same class are treated as correct for MAP evaluation, learning with multi-level semantic relation still boosts the MAP of TINA. This observation further proves the necessity of modeling multi-level semantic relation with local projections for cross-modal retrieval.

Retrieval examples. Some top results of image-to-text and text-to-image are shown in Figure 5 and 6, respectively. In Figure 5, take the bottom right challenging query as an example, each of the top 5 retrieved texts contains one or more relevant words to the visually cluttered query. Similar phenomenons can be observed on other examples. In Figure 6, see the 4th row, when querying with text about "tiger", TINA returns the nearly perfect semantically coherent results, where the top result is tiger image, followed by other feline animals, canine animals and forest scene images. At the bottom row, using the "airplane" query text, the 1st, 5th and 6th retrieved images are perfectly matched "airplane" images. However, visual contents in the 2nd, 3rd, 4th and 7th images are either similar to visual appearances of "airplane" images, or they contain "sky" background that can also be found in almost all "airplane" images.

VII. CONCLUSION

We propose TINA, a cross-modal correlation learning method by adaptive hierarchical semantic aggregation. Our approach utilizes the cross-modal training data from different levels of semantic relation, *i.e.*, the correspondence and multi-level semantic relation. The structure risk objective function that involves semantic coherence measurement, local projection consistency and the complexity penalty of local projections is optimized. Consequently, a set of local projections and gating functions are constructed for both



(a) Image-to-text on NUS-WIDE (b) Image-to-text on ICML-Challenge (c) Text-to-image on NUS-WIDE (d) Text-to-Image on ICML-Challenge

Figure 7. MAP of the top K retrieved results.

modalities. 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 more types of projection function for modeling cross-modal correlation, *e.g.*, constructing a domain specific feature extraction mechanism (e.g., the stacked convolution layers), or combining stacked auto-encoders with the localized correlation learning.

ACKNOWLEDGMENT

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61332016, 61303160, 61390511, 61101212, 61372169, 61471049 and 90920001, in part by China 863 program: 2014AA015202 and 2012AA012505, and in part by China Postdoctoral Science Foundation: 2014T70126. It was also supported in part by the NRF under its IRC @ Singapore Funding Initiative and administered by the IDM Programme Office, MDA and the Pinnacle Lab at SMU.

REFERENCES

[1] C. Archambeau and F. Bach. Sparse probabilistic projections. *NIPS*, 2009.
[2] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. *IJCAI*, 2003.
[3] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, 2006.
[4] D. Blei and M. Jordan. Modeling annotated data. *SIGIR*, 2003.
[5] M. M. Bronstein, A. M. Bronstein, F. Michel and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. *CVPR*, 2010.
[6] X. Chen, H. Liu and J. G. Carbonell. Structured Sparse Canonical Correlation Analysis. *AISTATS*, 2012.
[7] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo and Y. Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. *CIVR*, 2009.
[8] J. Deng, A. C. Berg and F. F. Li. Hierarchical Semantic Indexing for Large Scale Image Retrieval. *CVPR*, 2011.
[9] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.
[10] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. *CVPR*, 2011.
[11] C. Fellbaum. WordNet: An Electronic Lexical Database. *MIT Press*, 1998.
[12] A. Galen, A. Raman, B. Jeff and L. Karen. Deep Canonical Correlation Analysis. *JMLR W&CP*, 2013.
[13] G. Griffin and P. Perona. Learning and Using Taxonomies For Fast Visual Categorization. *CVPR*, 2008.
[14] D. R. Hardoon and J. Shawe-Taylor. Sparse Canonical Correlation Analysis. *Machine Learning*, 2011.
[15] H. Hotelling. Relations between Two Sets of Variates. *Biometrika*, 1936.
[16] S. J. Hwang, K. Grauman and F. Sha. Learning a Tree of Metrics with Disjoint Visual Features. *NIPS*, 2011.
[17] Y. Jia, M. Salzmann and T. Darrell. Learning Cross-modality Similarity for Multinominal Data. *ICCV*, 2011.
[18] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. *IJCAI*, 2011.
[19] W. Li and X. Wang. Locally Aligned Feature Transforms across Views. *CVPR*, 2013.
[20] L. J. Li, C. Wang, Y. Lim, D. Blei, and F. F. Li. Building and using a semantivisual image hierarchy. *CVPR*, 2010.
[21] J. Masci, M. M. Bronstein, A. A. Bronstein and J. Schmidhuber. Multimodal similarity preserving hashing. *TPAMI*, 2013.
[22] J. Masci, A. M. Bronstein, M. M. Bronstein, P. Sprechmann and G. Sapiro. Sparse similarity-preserving hashing. *arXiv*, 2013.
[23] M. Marszalek and C. Schmid. Constructing Category Hierarchies for Visual Recognition. *ECCV*, 2008.
[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A. Y. Ng. Multimodal deep learning. *ICML*, 2011.
[25] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. *ACM MM*, 2010.
[26] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. *LNCS*, 2006.
[27] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *IJCAI*, 1995.
[28] A. Sharma, A. Kumar, H. Daume III and D. W. Jacobs. Generalized Multiview Analysis: A Discriminative Latent Space. *CVPR*, 2013.
[29] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. *SIGMOD*, 2013.
[30] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep Boltzmann machines. *NIPS*, 2012.
[31] J. Sivic, B. Russell, A. Zisserman, W. Freeman and A. Efros. Unsupervised discovery of visual object class hierarchies. *CVPR*, 2008.
[32] N. Verma, D. Mahajan, S. Sellamanickam and V. Nair. Learning Hierarchical Similarity Metrics. *CVPR*, 2012.
[33] S. Virtanen, A. Klami and S. Kaski. Bayesian CCA via Group Sparsity. *ICML*, 2011.
[34] Y. Zhen and D. Y. Yeung. A Probabilistic Model for Multimodal Hash Function Learning. *KDD*, 2012.
[35] Y. Yang, F. P. Nie, S. M. Xiang, Y. T. Zhuang and W. H. Wang. Local and Global RegressiveMapping for Manifold Learning with Out-of-Sample Extrapolation. *AAAI*, 2010.
[36] D. C. Zhan, M. Li, Y. F. Li and Z. H. Zhou. Learning instance specific distances using metric propagation. *ICML*, 2009.
[37] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton. Adaptive mixtures of local experts. *Neural Comput.*, 1991.
[38] https://www.kaggle.com/c/challenges-in-representation-learning-multi-modal-learning/data
[39] Z. Wu, M. Palmer. Verbs semantics and lexical selection. *ACL*, 1994.