

Twitter100k: A Real-world Dataset for Weakly Supervised Cross-Media Retrieval

Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang

Abstract—This paper contributes a new large-scale dataset for weakly supervised cross-media retrieval, named Twitter100k. Current datasets, such as Wikipedia, NUS Wide and Flickr30k, have two major limitations. First, these datasets are lacking in content diversity, *i.e.*, only some pre-defined classes are covered. Second, texts in these datasets are written in well-organized language, leading to inconsistency with realistic applications. To overcome these drawbacks, the proposed Twitter100k dataset is characterized by two aspects: 1) it has 100,000 image-text pairs randomly crawled from Twitter and thus has no constraint in the image categories; 2) text in Twitter100k is written in informal language by the users.

Since strongly supervised methods leverage the class labels that may be missing in practice, this paper focuses on weakly supervised learning for cross-media retrieval, in which only text-image pairs are exploited during training. We extensively benchmark the performance of four subspace learning methods and three variants of the Correspondence AutoEncoder, along with various text features on Wikipedia, Flickr30k and Twitter100k. Novel insights are provided. As a minor contribution, inspired by the characteristic of Twitter100k, we propose an OCR-based cross-media retrieval method. In experiment, we show that the proposed OCR-based method improves the baseline performance.

Index Terms—cross-media retrieval, Twitter100k dataset, weakly supervised method, benchmark

I. INTRODUCTION

CROSS-media retrieval has extensive applications. In this paper, we mainly discuss image-text retrieval. In this task, we aim to search relevant images (texts) from a large gallery that depict similar content with a text (image) query. The primary challenge in cross-media retrieval consists in eliminating the heterogeneity and mining the semantic correlation between modalities [1]–[6]. Another challenge is to speed up the process of finding relevant data to a query, which can be met by indexing techniques [7]–[9]. We focus on the first challenge in this work.

In the community, previous works can be categorized into two classes: weakly supervised and strongly supervised methods. For the former, only image-text pairs are available during training, *e.g.*, multimodal DBM [10], DCCA [11], MSAE [12], MSDS [13] and CFA [14]; for the latter, class labels are

provided for each modality, such as TINA [15], LCFS [16], JFSSL [17], cluster-CCA [18] and CAMH [19]. This paper is concentrated on weakly supervised methods, which have critical research and application significance in realistic settings due to the expensive data annotation process [20].

This paper is motivated in two aspects. On the one hand, a majority of recent methods lay emphasis on leveraging fully supervised information such as the class labels. Usually, it is assumed that the training classes and testing classes are identical. However, this practice may be problematic in two aspects: 1) it seems a strong assumption that a query falls into a pre-defined class during training; 2) it might well be the case that labeled data is not available due to the annotation cost.

On the other hand, currently available cross-media datasets have some limitations. First, they are usually deficient in content diversity. For example, the Pascal VOC 2012 dataset [21] has 20 different classes such as dog, horse, aeroplane *etc.* However, retrieval involves multiple domains under realistic Internet circumstances. Retrieval methods trained on datasets of scanty domains may have difficulties in handling queries from unknown domains. Second, texts in current text-image datasets are written in well-organized language, using standard grammar and spelling as well as proper words. However, texts may be written in a casual way and contain informal expressions in practice. Third, each image in existing dataset is associated with tags or a highly-related text description. But in more realistic scenarios, images and texts are loosely correlated. Not all the words in texts have a visual interpretation in images. Fourth, popular cross-media datasets consisting of sentences and images may be flawed in dataset scale, such as the IAPR TC-12 dataset [22] (20,000 samples) and the Wikipedia dataset [23] (2,866 samples). The lack of data makes it challenging to evaluate the robustness of retrieval methods in large-scale galleries.

Considering the above-mentioned problems, this paper makes two major contributions. Our first contribution is the collection of a new large-scale cross-media dataset, named Twitter100k. It contains 100,000 image-text pairs collected from Twitter¹. It is distinguished from existing datasets in two aspects: varied domains and informal text language. In result, this dataset provides a more realistic benchmark for cross-media analysis. Another contribution is that we provide extensive benchmarking experiments of the weakly supervised methods by testing the performance of four subspace learning methods [23]–[26] and three variants of Correspondence Autoencoder [27] on the new dataset along with the Wikipedia

Y. Hu and Y. Huang are with Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. Email: huyt16@mails.tsinghua.edu.cn, yfhuang@tsinghua.edu.cn.

L. Zheng and Y. Yang is with the Center of AI, University of Technology, Sydney. Email: liangzheng06@gmail.com, yi.yang@uts.edu.au.

The Twitter100k dataset together with the codes, feature files, training/testing set split and benchmarking results are available at <http://ngn.ee.tsinghua.edu.cn/members/yuting-hu/>.

¹twitter.com

and Flickr30k datasets. Under the circumstances where class labels are unknown, pairs of texts and images are the only attainable training data. We propose to employ the cumulative match characteristic (CMC) curve rather than mean average precision (mAP) for accuracy evaluation.

As a minor contribution, inspired by the characteristics of Twitter100k that nearly 1/4 of the images contain texts which are highly correlated to the paired tweets, we make use of the texts in images and propose an OCR-based retrieval method. The effectiveness of the method is verified by the experiment.

The rest of this paper is organized as follows: we first present an overview on commonly used methods and datasets for cross-media retrieval in Section II. Then, we introduce the Twitter100k dataset in detail in Section III. Inspired by the characteristic of the new dataset, we proposed an OCR-based retrieval method in Section IV. The benchmarking methods and evaluation protocol are described in Section V. The benchmarking results and experimental analysis are presented in Section VI. Finally, Section VII concludes this paper.

II. RELATED WORK

A. Cross-media Retrieval Methods

Subspace learning methods. Subspace learning methods learn a common space for cross-media data, in which the similarity between the two modalities can be measured by L_2 distance, cosine distance, *etc.* Canonical Correlation analysis (CCA) [23] learns subspaces in which the correlation between the two modalities is maximized. Partial least square (PLS) [25] learns latent vectors by maximizing the covariance between the two modalities. Bilinear Model (BLM) [24] learns a set of basis in which data of the same content and different modality will project to the same coordinates. CCA, PLS and BLM are united in a framework called Generalized Multiview Analysis (GMA) by defining a joint optimization of two objective functions over two different vector spaces [26]. A multi-view extension of Marginal Fisher Analysis (MFA) is derived, called generalized multiview marginal fisher analysis (GMMFA).

Topic models. Topic models are widely applied to cross-media problem. Correspondence Latent Dirichlet Allocation (Corr-LDA) is proposed to effectively model the joint distribution of the two modalities and the conditional distribution of the text given the image [28]. A set of latent topics serves as latent variable and is shared between the two modalities. In topic-regression multi-modal Latent Dirichlet Allocation (trmmLDA) model [29], two separate sets of topics are learned and correlated by a regression module. Multi-modal Document Random Field (MDRF) model [30] defines a Markov random field over LDA topic model and learns topics shared across connected documents to encode the relations between different modalities.

Deep learning methods. Some cross-media retrieval methods are based on deep learning. Deep Restricted Boltzmann Machine (Deep RBM) is utilized to model the joint representations for the two modalities by learning a probability density over the space of multimodal inputs [31]. Deep Canonical Correlation Analysis (DCCA) [32] finds complex nonlinear

TABLE I: Summary of some popular cross-media datasets.

dataset	modality	# pairs	avg. text len.	# classes
Wikipedia	img./text	2,866	640.5 words	10
Flickr30k	img./sentence	31,783	5 sentences	n.a
IAPR TC-12	img./text	20,000	23.1 words	n.a
Pascal VOC'12	img./tags	11,540	1.4 tags	20
NUS-WIDE	img./tags	269,648	2.4 tags	81

transformations of two modalities such that the resulting representations are highly linearly correlated. A Relational Generative Deep Belief Nets (RGGDN) model [33] computes the latent features for social media that best embed both the content and observed relationships by integrating the Indian buffet process into the modified Deep Belief Nets. Inspired by the efficiency of convolutional neural networks (CNN) for image [34] and text [35], a method called Deep and Bidirectional Representation Learning Model (DBRLM) uses two types of convolutional neural networks to represent text and image [36]. Correspondence Autoencoder (Corr-AE) [27] methods adopt two autoencoders to reconstruct the input text and image features while adding a connective layer to two hidden layers such that representation learning and common space learning can be accomplished in a single process.

B. Cross-media Datasets

We introduce five commonly used datasets in cross-media retrieval. A brief summary of these datasets is provided in Table I.

The Wikipedia dataset. The Wikipedia dataset² [23] contains 2,866 image-text pairs spread over 10 categories collected from Wikipedia's featured articles. The dataset is pruned to keep text that contains at least 70 words and the average text length is 640.5 words. The articles are determined by Wikipedia's editors and written in formal language. The shortcoming of this dataset is the insufficiency of the image-text pairs. We present two examples in this dataset in Fig. 1a.

The Flickr30k dataset. The Flickr30k dataset³ [37] comprises 31,783 images from Flickr. Each image is associated with five descriptive sentences independently written by native English speakers from Mechanical Turk. Different annotators use different levels of specificity, from describing the overall situation to specific actions. The images and texts focus on people involved in everyday activities, events and scenes. No category information of this dataset is available. Examples in the dataset are given in Fig. 1b.

The IAPR TC-12 dataset. The IAPR TC-12 dataset⁴ [22] is made up of 20,000 image-text pairs including domains of sports, actions, people, animals, cities, landscapes, *etc.* The images are taken from locations around the world. Each text is composed of a title, a short description of the image, location and date of creation. The texts are written in English, German, Spanish or Portuguese. Each English text has 5.4 words in title and 23.1 words in the description averagely. We provide several examples in Fig. 1c.

²<http://www.svcl.ucsd.edu/projects/crossmodal/>

³<http://shannon.cs.illinois.edu/DenotationGraph/>

⁴<http://imageclef.org/photodata>



Fig. 1: Examples in common-used cross-media datasets. (a)-(e) are examples in the Wikipedia, Flickr30k, IAPR TC-12, Pascal VOC 2012 and NUS-WIDE datasets, respectively.

The Pascal VOC 2012 dataset. The Pascal VOC 2012⁵ [21] is a dataset designed for classification and detection tasks. It consists of 11,540 image-tag pairs in 20 different classes. Tags are the objects in the images. Each image has 1.4 tags on average and 7,567 images are labeled with only one tag. Examples in this dataset are show in Fig. 1d.

The NUS-WIDE dataset. The NUS-WIDE dataset⁶ [38] contains 269,648 images and the associated tags from Flickr. Six types of low-level features are extracted from these images including color histogram, edge direction histogram, wavelet texture, block-wise color moments and bag of words based on SIFT descriptions. Ground-truth for 81 concepts can be used for evaluation. Each image holds 2.4 concepts on average. Some examples in this dataset are illustrated in Fig. 1e.

III. TWITTER100K: A MULTI-DOMAIN DATASET

In this section, we introduce a multi-domain dataset called Twitter100k, which is comprised of 100,000 image-text pairs collected from Twitter.

A. Dataset Collection

Individual steps in data collection are described below.

Seed user gathering. To ensure the diversity of the collected data, we obtain seed users by sending queries to Twitter with various topic words, such as trip, meal, fitness, sports, *etc.* These randomly selected users serve as seed to acquire more user candidates.

User candidate generation. A web spider is developed to crawl the accounts of the users who are following the seed users. This step iterates several times until we get a long list of

user candidates. The domains covered by the users are further enlarged with the iteration.

Tweet collection. Another web spider collects tweets with the corresponding images by visiting the homepages of all the users in the candidate list. We find that around 1/3 of the tweets are accompanied with images.

Data pruning. The image-tweet pair is pruned under any of the following situations.

- Messy codes in tweets;
- Tweets without words;
- Tweets not written in English;
- Reduplicate tweets with same ID;
- Error images.

We finally obtain 100,000 image-text pairs in total. An image and text appearing in one piece of tweet are considered as a pair. Some examples in this new dataset are presented in Fig. 2.

B. Dataset Characteristics

The Twitter100k dataset is featured in the following five aspects. First, this dataset is collected from social media, hence it covers a wide range of domains, such as sport, architecture, food, animal, news, plant, person, poster and so forth.

Second, since informal language is usually used by Internet users when posting tweets, texts in Twitter100k are distinguished from other datasets in grammar and vocabularies.

- **Abbreviations.** An abbreviation is a shortened form of a word or word group. Take the texts marked with blue panes in the third row of Fig. 2 as example, *ur* and *ppl* are short for *you are* and *people*, respectively.
- **Initialisms.** An initialism is a term formed from the initial letter or letters of several words. In the texts marked with green panes in the third row of Fig. 2, *idc* and *lol* are initialisms for *I don't care* and *laughing out loud* in several.
- **Omission of subject and verb.** Both subject and verb are omitted when tweet is description or exclamation of an activity or something. The texts marked with purple angles in the third row of Fig. 2 are examples.
- **Hashtags.** A hashtag is a type of label used on social network to indicate a specific theme. It consists of a hash character # and a word or unspaced phrase. In the texts marked with yellow lines in the third row of Fig. 2, hashtags are highly correlated to the images.

Third, the correlation between the image and text is often very loose. In the examples in the first row of Fig. 2, some words in texts are direct descriptions of images while other words are out of semantic consistence and have no visual interpretations. Such is the case in practice.

Fourth, Twitter100k is a large-scale dataset, comprising 100,000 image-text pairs. A wealth of data can avoid over-fitting during training. Moreover, it can be exploited to test the robustness of retrieval methods under massive data.

Last, approximately 1/4 of the images in this dataset contain text which are highly correlated to the paired tweets. As is presented by the examples in the second row of Fig. 2, some words in tweets are corresponding to the text located in

⁵<http://host.robots.ox.ac.uk/pascal/VOC/>

⁶<http://ms.comp.nus.edu.sg/research/NUS-WIDE.htm>

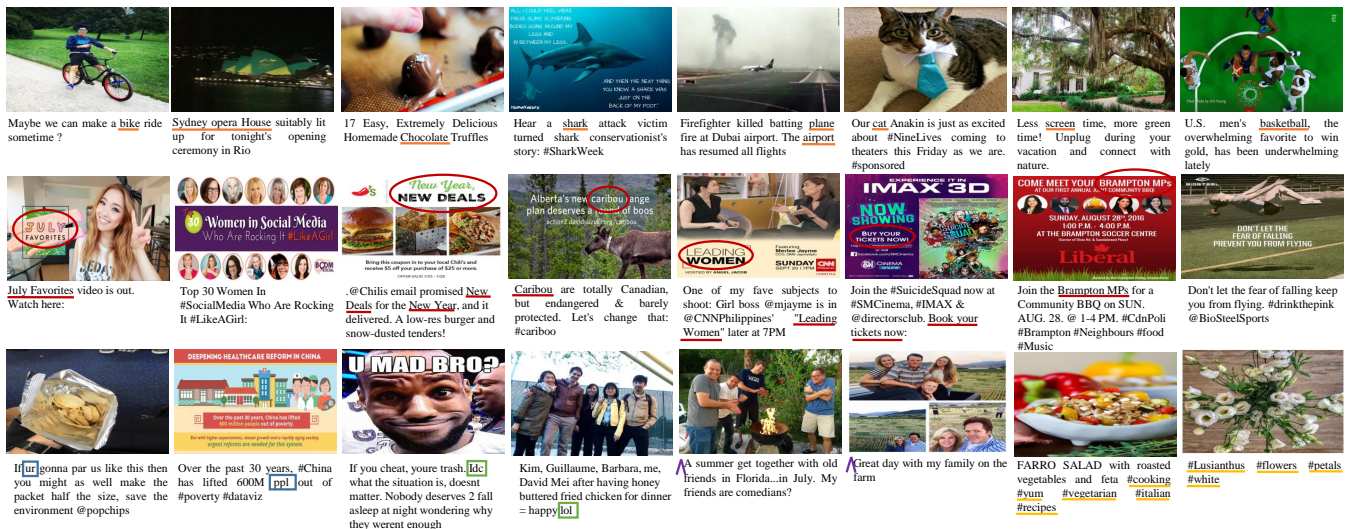


Fig. 2: Examples in Twitter100k. In the first row are presented the examples whose images are visual interpretations of texts marked with orange lines. In the second row are presented the examples whose images contain words marked with red circles which are highly correlated to texts marked with red lines. In the third row are presented examples whose texts are written in informal language: abbreviations marked with blue panes, initialisms marked with green panes, omission of subject and verb marked with purple angles, hashtags marked with yellow lines.

images. In extreme cases, tweets can be identical to the text in images. To our knowledge, Twitter100k is the only cross-media dataset with this characteristic.

C. Potential Application Scenarios

The Twitter100k dataset provides a more realistic benchmark for cross-media retrieval. Since it is collected from Twitter, cross-media retrieval on this dataset is promising for application scenarios to be detailed below.

- Social media platforms such as Twitter only provide pre-defined emoticons for users to choose when posting a tweet. In fact, by cross-media retrieval, it can be more convenient and interesting to extend the range of emoticons and recommend suitable images for users according to the contents of the tweets.
- Current content-based user recommendation and interest-group mining only take into consideration single-media data such as text. Adding correlation information between texts and images can improve the effectiveness of user recommendation systems and interest-group mining.

IV. PROPOSED OCR-BASED RETRIEVAL METHOD FOR TWITTER100K DATASET

It can be found out from the Twitter100k dataset that about 1/4 of the images contain text which may be correlated to the corresponding tweets and several of these images even involve no objects except text. It is challenging for current methods in cross-media retrieval to find the correlation between the tweets and these kind of images when optical character recognition (OCR) is not employed. Various features are used for image representation. For example, SIFT [39] and color features are exploited in [40]; GIST [41] and HOG [42] are adopted in

[43]. These features are effective in extracting color and shape of images, but defect in representing the words contained in the images. In order to tackle the problem, we propose an OCR-based cross-media retrieval method.

Our method consists of five components, which will be described below.

OCR-text extraction. We extract words on each image using python wrapper for Tesseract⁷.

OCR-text pruning. As a revision of Tesseract, we prune the OCR-texts based on word frequency and text length. We first generate a vocabulary of the most frequent 5,000 words using the tweets in Twitter100k. After removing the words not in this vocabulary, we keep those OCR-texts which have at least two words remained. A total of 21,579 OCR-texts are obtained.

Distance between tweet and OCR-text. We adopt Jaccard Distance to measure the similarity between the tweet and OCR-text. The Jaccard Distance is defined as:

$$J(T, O) = \frac{M_{10} + M_{01}}{M_{10} + M_{01} + M_{11}}, \quad (1)$$

where T and O denote tweet and the OCR-text, respectively; M_{11} is the number of words contained in both the tweet and OCR-text, M_{10} is the number of words contained in tweet but not in OCR-text, and M_{01} is the number of words contained in OCR-text but not in tweet.

Distance in common subspace. We use T and I to denote tweet and image. $C(T, I)$ represents the cosine distance between the two modalities in the common subspace learned by retrieval methods such as CCA, PLS, BLM, GMMFA and Corr-AE methods.

⁷Tesseract is an optical character recognition engine and considered as one of the most accurate open-source OCR engines. The code can be found at <https://github.com/tesseract-ocr/tesseract>

Hybrid distance. We define a hybrid distance between the tweet and image as follows:

$$\text{dist}(T, I) = \begin{cases} \alpha J(T, O) + (1 - \alpha)C(T, I), & \text{Ind}(I) = 1 \\ C(T, I), & \text{Ind}(I) = 0 \end{cases} \quad (2)$$

where α is a weight parameter and $\text{Ind}(I)$ is a boolean-valued function to indicate whether image I has a corresponding OCR-text. The influence of weight factor α is evaluated in the experiments.

Ranking. We rank the candidates in the gallery based on the hybrid distance between the query and candidate.

V. BENCHMARKING METHODS AND EVALUATION PROTOCOL

This paper mainly discuss the weakly supervised retrieval methods, in which only the co-occurrence information of the image and text is exploited. We will first describe some popular methods that will be adopted in the benchmarking experiment and then introduce the evaluation protocol.

A. Compared Methods

The cross-media retrieval methods we will compare in the paper are listed below.

- **CCA.** Canonical correlation analysis (CCA) [23] finds a basis of canonical components, *i.e.*, directions, along which the data is maximally correlated.
- **BLM.** The bilinear model (BLM) [24] can separate style and content by using singular value decomposition (SVD). It learns a shared subspace in which data of the same content and different modality is projected to the same coordinates.
- **PLS.** Partial least square (PLS) [25] uses the least square method to correlate the subspaces of CCA in order to avoid information dissipation in the process of different modal correlations.
- **GMMFA.** Generalized multi-view marginal Fisher analysis (GMMFA) [26] is a multi-view extension of Marginal Fisher Analysis. It tries to separate different-class and compress same-class samples in the feature space. We use GMMFA in a weakly supervised way by regarding every image-text pairs as an independent category.
- **Corr-AE.** Correspondence autoencoder (Corr-AE) [27] uses two autoencoders to reconstruct the input text and image features. It minimizes the weighted sum of the reconstruction loss and the distance between the hidden vectors of the two modalities.
- **cross Corr-AE.** Cross Corr-AE is a variant of Corr-AE. It takes one modality as input to reconstruct the other modality.
- **full Corr-AE.** Full Corr-AE is another variant of Corr-AE. It takes one modality as input to simultaneously reconstruct the two modalities.

Among the above-mentioned methods, CCA, BLM, PLS and GMMFA are subspace learning methods, while Corr-AE, cross Corr-AE and full Corr-AE are collectively called Corr-AE methods.

B. Implement Details

In this section, we describe the implementations and settings of the compared methods in detail.

- **Subspace learning methods.** For subspace learning methods, we adopt the matlab implementation provided by [26]⁸ to compute the linear projection matrix. Since no label information is considered in this paper, we set every text-image pair with an independent label. The options are set default.
- **Corr-AE methods.** For Corr-AE methods, we adopt the GPU-based python implementation provided by [27]⁹ to compute the hidden vectors of the two modalities. According to the experiment results presented in [27], 1024-dimensional hidden layer is employed. The weight factor of reconstruction error and correlation distance is set to 0.8, 0.2, 0.8 for Corr-AE, cross Corr-AE and full Corr-AE, respectively.

C. Modality Representations

1) Text Representations:

- **LDA feature.** For subspace learning methods, the representation of text is derived from a latent Dirichlet allocation (LDA) [44] model. For each dataset, we first train a LDA model with 50 topics. Then each text is represented as a 50-dimensional LDA feature by the topic assignment probability distributions.
- **BoW feature.** For Corr-AE methods, text is represented by Bag-of-Word (BoW) feature. We first convert texts to lower case and remove stop-words. We adopt unigram model and select the most frequent 5,000 words to form a vocabulary. A 5000-dimensional BoW feature based on this vocabulary is generated for each text.
- **1024-dimensional WE-BoW feature.** Word embedding [45] (WE) is a language modeling and feature learning technique in natural language processing (NLP). It attempts to learn distributed representation of word, which is called word vector. Word vector contains semantic information of word and is applied to significantly improve many NLP applications [46]. Therefore, we propose to utilize a WE-BoW feature. A codebook of 1024 word vectors is first built with the k-means clustering algorithm using all the 400,000 300-dimensional word vectors pre-trained by GloVe¹⁰ [47]. Then word vectors in each text are quantized with this codebook, and text is represented by the L_2 -normalized word vector histogram that results from this quantization.

2) Image Representations:

- **4096-dimensional CNN feature.** We first resize each image to 224*224 and extract the fc7 CNN feature using VGG16 model [48] with the implementation of CAFFE¹¹ [49].

⁸<https://www.cs.umd.edu/~bhokaal/>

⁹<https://github.com/fangxiangfeng/deepnet>

¹⁰<http://nlp.stanford.edu/projects/glove/>

¹¹<http://caffe.berkeleyvision.org>

D. Dataset Split

In this paper, we benchmark the retrieval methods on the Wikipedia, Flickr30k and Twitter100k datasets. Each dataset is split into a training set, a validation set and a test set. The dataset split is described as follows:

- **The Wikipedia dataset.** We use 2,173 image-text pairs for training and 500 pairs for testing. For Corr-AE methods, additional 193 pairs serve as validation set. All the data in test set is utilized as query.
- **The Flickr30k dataset.** The amounts of the training set and test set are both 15,000 image-text pairs. Extra 1,783 pairs are employed for validation in Corr-AE methods. We select 2,000 images and texts from the test set randomly to function as query. Since each image has 5 matched sentences, when taking image as query, the text gallery contains 75,000 sentences and any one of the 5 matched sentences is considered correct.
- **The Twitter100k dataset.** 50,000 and 40,000 image-text pairs are exploited for training and testing, respectively. For Corr-AE methods, 10,000 pairs are used as validation set. 2,000 images and texts are selected randomly from the test set to serve as query.

E. Evaluation Metrics

Since no pre-defined category labels are available, text and image in a pair are considered as a ground-truth match. That is, given a query text (image), only one ground-truth image (text) exists in the gallery. As a consequence, we take the following two evaluation metrics.

- **CMC.** Cumulative match characteristic (CMC) is frequently used as a metric in the field of face recognition [50] and person re-identification [51] [52]. It measures how well an identification system ranks the identities in the enrolled database with respect to "unknown" probe image. For cross-media retrieval, CMC represents the expectation of finding the correct match in the top n matches and can be described by a curve of average retrieval accuracy with respect to rank.
- **Mean rank.** Mean rank is the average of the ranks of the correct matches for a series of queries Q .

$$\text{Mean Rank} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \text{rank}_i, \quad (3)$$

where rank_i refers to the rank position of the correct match for the i -th query.

VI. EXPERIMENT RESULTS

In this section, we present the experiment results and discuss the impact of several factors on the retrieval performance, including datasets, the amount of training data, text features and retrieval methods.

A. Comparison of Retrieval Methods on Various Datasets

The benchmarking results on Wikipedia, Flickr30k and Twitter100k are presented in Fig. 3. Two findings can be drawn.

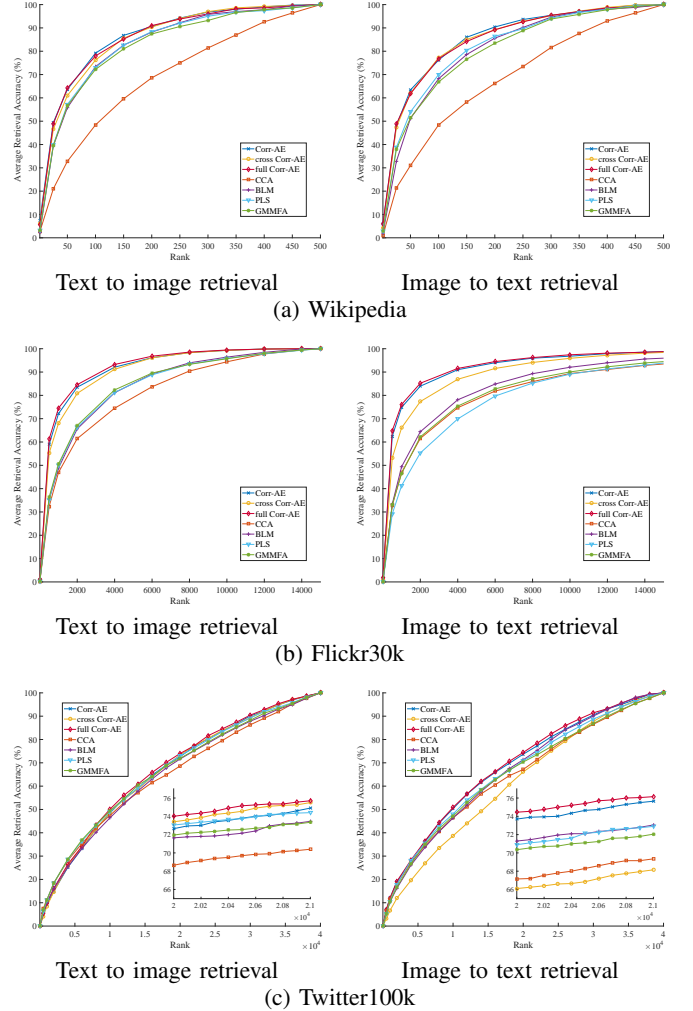


Fig. 3: CMC curves on the (a) Wikipedia, (b) Flickr30k, and (c) Twitter100k datasets. Various Corr-AE and subspace learning methods are compared.

First, Corr-AE methods surpass the subspace learning methods (CCA, BLM, PLS and GMMFA), especially on the Wikipedia and Flickr30k dataset. The reason is that the subspace learning methods are based on a two-stage framework, which first extracts features for each modality separately and then finds a linear matrix to project two modalities into a shared space. In other words, correlation learning is separated from representation learning. By contrast, Corr-AE methods incorporates representation learning and correlation learning into a single process by defining the loss function as the weighted sum of reconstruction error and correlation loss. Thus Corr-AE methods can achieve superior performance to linear-projection methods. Among the three variants of Corr-AE methods, full Corr-AE achieves the best performance on the whole. In full Corr-AE, two modalities are reconstructed with one modality as input, hence the correlation between images and texts is well embedded in the hidden feature of the autoencoder. Moreover, cross Corr-AE performs poorly in image to text retrieval, which manifests the heterogeneous gap between the two modalities.

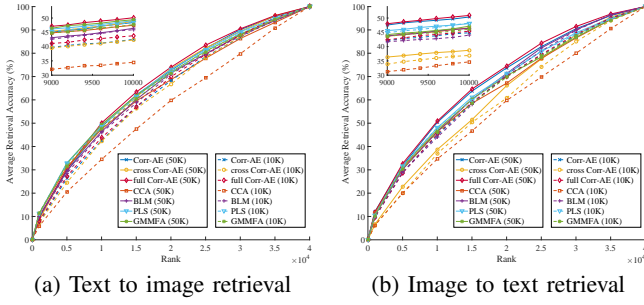


Fig. 4: CMC curves on the Twitter100k dataset with 10k or 50k image-text pairs as training data.

Second, generally speaking, the retrieval performance on Flickr30k is the highest among the three datasets while the performance on Twitter100k is the lowest. In other words, the Twitter100k dataset can be viewed as the most challenging one among the three datasets. We speculate that the relatively high performance of the Flickr30k dataset can be attributed to the fact that the texts are written by native English speakers to directly describe the images; the texts and the images are highly correlated. For the Wikipedia dataset, the text-image pairs are collected from Wikipedia’s featured articles and the long texts introduce abundant aspects about the images, including background, history, episode and so on. Consequently, majority words in the texts are not visual interpretations of the images. The reason why Twitter100k has lowest retrieval accuracy lies in three aspects. First, this dataset covers a diversity of domains. Second, informal expressions often exist in tweets. Third, the texts and images have a relatively loose correlation, and the correlation is even various from user to user. As a result, it is challenging to learn the correlation between the texts and images for Twitter100k.

B. The Impact of Different Amount of Training Data

In this section, we explore whether retrieval performance can be improved with larger amount of training data. To this end, we use 50,000 and 10,000 image-text pairs for training, respectively. The experiment results on Twitter100k dataset with different quantity of training data are given in Fig. 4.

As is shown in Fig. 4, all the dotted lines (10k training data) are below the corresponding solid line (50k training data). For text to image retrieval with full Corr-AE, for instance, average retrieval accuracy improves from 42.55% to 48.65% (+6.10%) at rank= 9500 when the available training data increases from 10k to 50k.

Abundant training data can lead to high ability for generalization and effective correlation learning. The improvement brought by the massive data verifies that the large scale of Twitter100k is a crucial advantage for cross-media retrieval.

C. Comparison of Various Text Features

To evaluate the influence of different text features, we test the retrieval performance on the Wikipedia, Flickr30k and Twitter100k datasets using various features, *i.e.*, LDA feature,

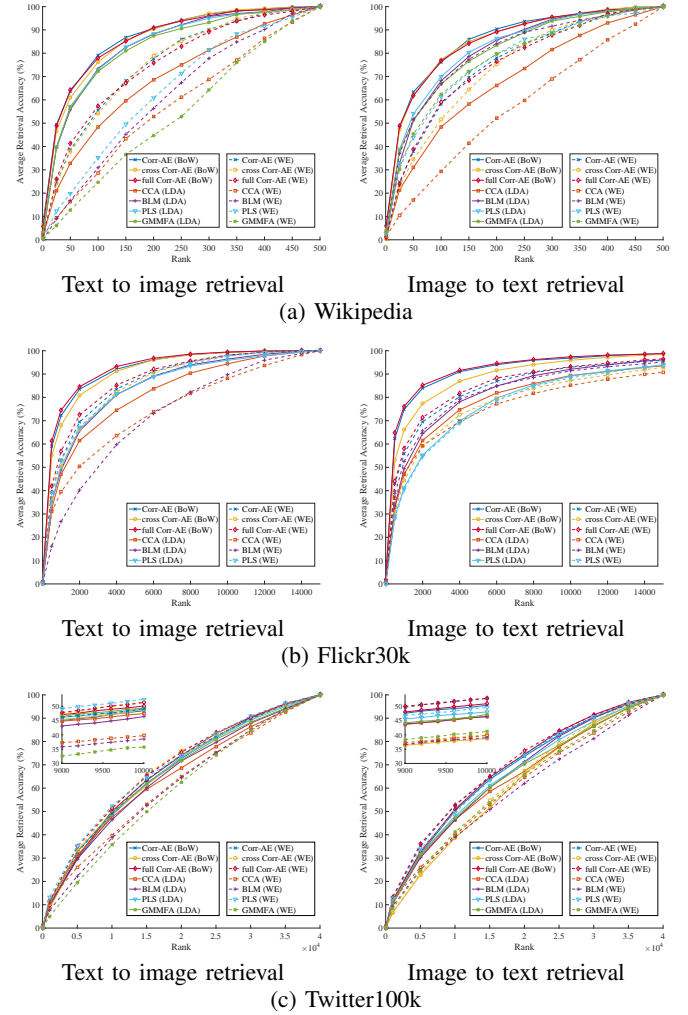


Fig. 5: CMC curves with various text features on the (a) Wikipedia, (b) Flickr30k, and (c) Twitter100k datasets.

BoW feature and WE-BoW feature. The benchmarking results with different text features are shown in Fig. 5.

The impact of WE-BoW feature varies on datasets and retrieval methods. Take text to image retrieval as example. Similar findings can be obtained in image to text retrieval. First, the results in Fig. 5a and Fig. 5b demonstrate that the WE-BoW feature is secondary to the LDA feature and BoW feature for cross-media retrieval on Wikipedia and Flickr30k. For example, after exploiting the WE-BoW feature, average retrieval accuracy decreases from 90.60% to 77.60% (-13.0%) for Corr-AE at rank= 200 on Wikipedia. Even more serious deterioration can be seen for BLM at rank= 2000 on Flickr30k, from 65.65% to 40.30% (-25.35%). Second, similar degradations in performance can be observed on Twitter100k when CCA, BLM and GMMFA are adopted, *e.g.*, the accuracy is lower by 7.60% at rank= 9200 for CCA. Third, the usage of the WE-BoW feature ameliorates the retrieval accuracy for Corr-AE methods and PLS on Twitter100k. For instance, the accuracy exceeds by +2.65% and +3.3% for cross Corr-AE at rank= 9400 and PLS at rank= 9200, respectively.

The improvement can be attributed to the semantic informa-

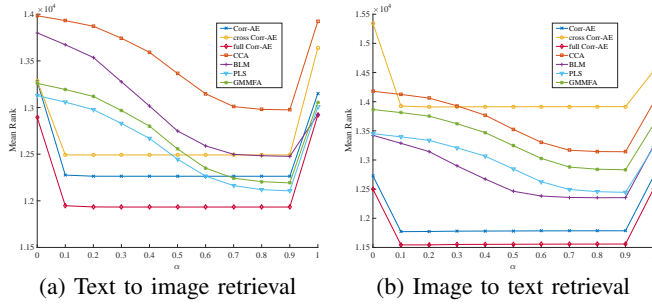


Fig. 6: Mean rank of the correct matches with different values of α on the Twitter100k dataset.

tion brought by the word vectors, which are trained on large corpus of texts. The vectors of similar words are close in space. In contrast, words are considered independent in the BoW representation. Since texts in Twitter100k contain plenty of informal expressions and abbreviations, semantic information embedded in WE-BoW benefits the retrieval.

However, the text in Wikipedia and Flickr30k gives a detailed description of image and is written in formal language. LDA and BoW are effective to represent text in this kind of corpus. Since WE-BoW is derived from the pre-trained word vectors, LDA and BoW offers an advantage over the WE-BoW feature on Wikipedia and Flickr30k.

On Twitter100k, the opposite effect of word embedding on different retrieval methods results from the merit and drawback of WE-BoW. The merit is the semantic information contained by word vectors. The drawback is the high dimension compared with the 50-dimensional LDA feature. For subspace learning methods, 1024-dimensional WE-BoW makes it difficult to find the shared space for the cross-media data, hence retrieval performance suffers from the WE-BoW feature. On the contrary, since WE-BoW is lower in dimension than the 5000-dimensional BoW feature used for Corr-AE methods, retrieval accuracy is improved with the integration of word embedding.

In addition, the word vectors used in this experiments are pre-trained on text corpus written in well-organized language, specific word vectors trained on Twitter corpus will be more beneficial for the Twitter100k dataset.

D. Performance of the Proposed OCR-based Method

This section presents the performance of the proposed OCR-based retrieval method on Twitter100k. Above all, we discuss the impact of the weight parameter α on the retrieval performance. The mean rank of the correct matches with different values of α on Twitter100k are provided in Fig. 6.

These results reveal that the mean rank of the correct matches declines with α when $\alpha < 1$ for all the baselines. It shows that text on image plays a dominant role in retrieval. But the performance deteriorates when α is set to 1, because the information about the color and the shape of the image is lost. We accordingly choose 0.9 as the best value for α .

Then we compare the performance of the proposed OCR-based method with baselines. The experiment results are represented in Fig. 7.

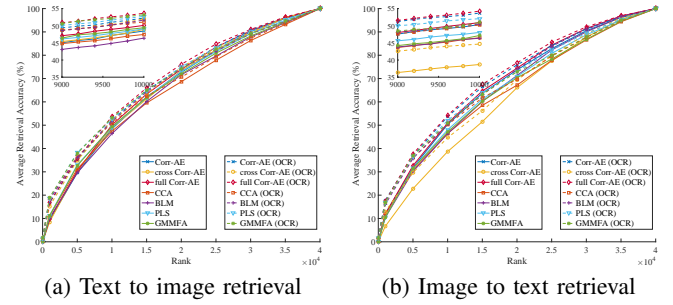


Fig. 7: CMC on Twitter100k dataset with proposed method and baselines.

From Fig. 7 we can find out that all the dotted lines (the proposed method) are above the corresponding full line (baselines). In other words, average retrieval accuracy improves by incorporating OCR when the same retrieval method is used, e.g., for BLM in text to image retrieval, average retrieval accuracy ascends from 43.10% to 48.60% (+5.50%) at rank=9000. This enhancement is ascribed to the facts that the tweets and the texts on the images are highly correlated; the proposed methods can utilize the text information of the images besides the shape and color information.

The proposed OCR-based method may be further developed by utilizing superior similarity metrics and advanced text retrieval methods. Moreover, besides modifying the distance formulation by incorporating Jaccard distance between tweet and OCR-text, OCR-text can be integrated in multiple ways, such as concatenating the OCR-text with other image features and so forth.





E. Retrieved Examples of the Twitter100k Dataset

In this section, we study the retrieval results of the Twitter100k dataset. We adopt full Corr-AE method and BoW feature. OCR-texts and 50,000 training data are employed. For each query, we represent top five retrieved results in the rank list and provide the ground truth match together with its rank for reference.



As is demonstrated in Fig. 8, the ground truth match and top five retrieved results are correlated to the query from two perspectives. One is the content aspect. Take the first text to image retrieval for example. *Dinner* and *baby* are two key words in the query text. The food and drinks in images are the visual interpretations of *dinner*. The people in images are corresponding to *baby* because *baby* can be used as hypocorism to call the person loved by someone besides infant.

Another perspective is opinion or sentiment. For instance, in the first image to text retrieval, the query image consists of a woman in dress, sky, clouds, grass, trees and mountains. It is a scenic image taken during the travel, which can be confirmed by the ground truth text. Opinion words such as *beautiful* and *gorgeous* appear in the top four retrieved texts, which manifest the sentiment of the Internet users upon the scenery and travel.

In Twitter100k dataset, image is more than what it shows. It involves opinion and sentiment of the Internet users in addition

Query text	Ground Truth	Top 5 retrieved images
<code>Dinner</code> is served. Thanks <code>baby</code> @AllisonSky I <code>love</code> you.	 (Rank=9)	
Nice view <code>Sunset</code> looks good too. #MMonday	 (Rank=876)	

(a) Text to image retrieval

Query image	Ground Truth	Top 5 retrieved texts
	A #tbt to all the sun, smiles and views with @inhershoesblog in #StThomas during our #Trav (Rank=37)	One of the officially most <code>#beautiful</code> towns in #Spain #Morella Hassan II Mosque has some of the most <code>beautiful</code> arches I've seen. <code>Gorgeous</code> <code>view</code> of the Swiss Alps as can be seen reflecting from my world cup Escape to the <code>beautiful</code> & summery Corsica via Auguste Herbin #TBT Missing that <code>#Travel</code> glowing water.
	Emilia's baby (my baby) I love her (Rank=9)	my <code>little</code> smiley happy angel <code>love</code> her sooooo much I <code>love</code> you soooooooo much . My <code>Baby</code> Maik #ZeusMaik #Dogs #Yorks #miFamilia So precious. I remember being that <code>little girl</code> too 5 Things Dad Can Do to Help with the New <code>Baby</code> #SelfieForSebnah not you bro, this for my dad Seb Stan <code>love</code> you papi

(b) image to text retrieval

Fig. 8: Top retrieved examples of the full Corr-AE method on the Twitter100k dataset. The key words which are correlated to the images from the perspective of content and opinion are marked with rectangles.

to the content. Cite the second image to text retrieval as an example, with the query image of a girl, in addition to *little* and *baby*, *love* is a high frequency word in the retrieved texts. This implies that opinion aspect has an influence on multimedia retrieval although the sentiment of an image is hidden in a high semantic level.

VII. CONCLUSION

In this paper, we introduce a large-scale cross-media dataset called Twitter100k, which provides a more realistic benchmark towards weakly supervised text-image retrieval. A number of learning methods and text features are evaluated in our experiments. Considering the characteristic of the new dataset, we propose to improve the retrieval performance based on OCR-texts of images.

There is still a long way to go before achieving a satisfying retrieval performance on the new dataset. In future work, we plan to design a better evaluation protocol for this dataset and consider the correlation between the images and texts from high-level semantic perspectives, such as opinion and sentiment views. We expect this paper and the new dataset will motivate more insightful works.

ACKNOWLEDGMENT

This research is supported by the Key Program of National Natural Science Foundation of China (Grant No. U1536201 and No. U1536207).

REFERENCES

- [1] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 370–381, 2015. 1
- [2] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 208–218, 2016. 1
- [3] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 437–446, 2008. 1
- [4] Y. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 221–229, 2008. 1
- [5] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 175–184. 1
- [6] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 723–742, 2012. 1

- [7] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1997–2008, 2013. 1
- [8] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, "Sparse multimodal hashing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 427–439, 2014. 1
- [9] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua, "Interactive video indexing with statistical active learning," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 17–27, 2012. 1
- [10] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015. 1
- [11] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3441–3450. 1
- [12] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," *Proceedings of the VLDB Endowment*, vol. 7, no. 8, pp. 649–660, 2014. 1
- [13] J. Wang, Y. He, C. Kang, S. Xiang, and C. Pan, "Image-text cross-modal retrieval via modality-specific feature learning," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 347–354. 1
- [14] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 2003, pp. 604–611. 1
- [15] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1201–1216, 2016. 1
- [16] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2088–2095. 1
- [17] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2010–2023, 2016. 1
- [18] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *AISTATS*, 2014, pp. 823–831. 1
- [19] R. Liu, Y. Zhao, S. Wei, and Z. Zhu, "Cross-media hashing with centroid approaching," in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6. 1
- [20] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2494–2502, 2016. 1
- [21] S. J. Hwang and K. Grauman, "Reading between the lines: Object localization using implicit cues from image tags," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 6, pp. 1145–1158, 2012. 1, 2
- [22] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in *International workshop on Image*, vol. 5, 2006, p. 10. 1, 2
- [23] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 251–260. 1, 2, 5
- [24] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Advances in neural information processing systems*, pp. 662–668, 1997. 1, 2, 5
- [25] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, latent structure and feature selection*. Springer, 2006, pp. 34–51. 1, 2, 5
- [26] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2160–2167. 1, 2, 5
- [27] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 7–16. 1, 2, 5
- [28] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 127–134. 2
- [29] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3408–3415. 2
- [30] Y. Jia, M. Salzman, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2407–2414. 2
- [31] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230. 2
- [32] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML (3)*, 2013, pp. 1247–1255. 2
- [33] Z. Yuan, J. Sang, Y. Liu, and C. Xu, "Latent feature learning in social media network," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 253–262. 2
- [34] L. Zheng, Y. Yang, and Q. Tian, "Sift meets cnn: a decade survey of instance retrieval," *arXiv preprint arXiv:1608.01807*, 2016. 2
- [35] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014. 2
- [36] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016. 2
- [37] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014. 2
- [38] 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," in *Proceedings of the ACM international conference on image and video retrieval*. ACM, 2009, p. 48. 3
- [39] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. 4
- [40] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1939–1946. 4
- [41] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001. 4
- [42] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893. 4
- [43] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International journal of computer vision*, vol. 106, no. 2, pp. 210–233, 2014. 4
- [44] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003. 5
- [45] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119. 5
- [46] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011. 5
- [47] PenningtonJeffrey, SocherRichard, and M. D., "Glove: Global vectors for word representation," in *EMNLP*, vol. 14, 2014, pp. 1532–1543. 5
- [48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 5
- [49] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678. 5
- [50] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000. 6
- [51] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 868–884. 6
- [52] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016. 6



Yuting Hu received the B.E. degree in electronic engineering in 2016 from Tsinghua University, Beijing, China, where she is currently working toward Ph.D. degree. Her current research interests include multimedia information retrieval and natural language processing.



Liang Zheng received the B.E. degree in life science from Tsinghua University, Beijing, China, in 2010, and the Ph.D. degree in electronic engineering from Tsinghua University in 2015. From August 2015 to June 2016, he was a Postdoctoral Fellow in University of Texas at San Antonio. He is currently a Postdoctoral Fellow in University of Technology Sydney. His research interests include multimedia information retrieval and computer vision.



Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently an associate professor with University of Technology Sydney, Australia. He was a Post-Doctoral Research with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interest includes machine learning and its applications to multimedia content analysis and computer vision.



Yongfeng Huang received the Ph.D degree from Huazhong University of Science and Technology, Wuhan, China, in 2000. From 2000 to 2002, he was a Postdoctoral Fellow with the Department of Electronic Engineering, Tsinghua University, Beijing, China, where he is currently a professor. His research interests include information retrieval, data mining, multimedia network security and next-generation Internet.