Cross-media Relevance Computation for Multimedia Retrieval

Jianfeng Dong
College of Computer Science and Technology, Zhejiang University
danieljf24@zju.edu.cn

ABSTRACT

In this paper, we summarize our works for cross-media retrieval where the queries and retrieval content are of different media types. We study cross-media retrieval in the context of two applications, *i.e.,* image retrieval by textual queries, and sentence retrieval by visual queries, two popular applications in multimedia retrieval. For image retrieval by textual queries, we propose *text2image* which converts computing cross-media relevance between images and textual queries to comparing the visual similarity among images. We also propose *cross-media relevance fusion*, a conceptual framework that combines multiple cross-media relevance estimators. These two techniques have resulted in a winning entry in the Microsoft Image Retrieval Challenge at ACM MM 2015. For sentence retrieval by visual queries, we propose to compute cross-media relevance in a visual space exclusively. We contribute *Word2VisualVec*, a deep neural network architecture that learns to predict a visual feature representation from textual input. With proposed Word2VisualVec model, we won the Video to Text Description task at TRECVID 2016.

KEYWORDS

Cross-media retrieval, Image retrieval by textual queries, Sentence retrieval by visual queries

1 INTRODUCTION

With the rapid development of Internet techniques, smart mobile devices and social media, people can readily create multimedia contents by themselves, which lead to the deluge of multimedia data. Hence, efficient and effective multimedia retrieval tools become a big demanding for people. In my doctoral research, we focus on text, images, and videos, three types of widely used media. We aim to attack the challenging problem of cross-media retrieval where the queries and retrieval content are of different media type. For example, given an image, find sentences relevant to the image. The key of cross-media retrieval is computing the cross-media relevance between queries and retrieval content. As the representations of different media types are inconsistent and reside in different feature spaces, they are not directly comparable. So it is extremely challenging to compute cross-media relevance among them. Hence, the fundamental question we try to answer during the PhD study is: *"What determines the cross-media relevance for cross-media retrieval?"*. We answer the question in the context of image retrieval by textual query, and sentence retrieval by visual queries, two popular applications in multimedia retrieval. In [7, 9], we propose a text2image model and cross-media relevance fusion using click-through data for image retrieval by textual query. In [8, 29], we propose to compute cross-media relevance in a deep visual space for sentence retrieval by visual queries.

In order to train a cross-media relevance computation model, labeled data play the crucial role. Typically, the labeled data is annotated by humans. However, manual annotation is labor intensive and time-consuming, which makes well-labeled data expensive. The lack of high-quality labeled data limits the quality of cross-media retrieval systems. However, there are massive amount of click-through data from commercial search engines. Taking click-through data from the image search engine as an example [15], it is comprised of triads of $(textual\ query, image, click)$, where *click* is the accumulated amount of user clicks a specific image has received with respect to a given query. And the click reflects to some extent the relevance of the image with respect to the query, which opens a new way for multimedia retrieval study. So the first question that we try to investigate is: *"What is the value of click-through data for cross-media retrieval?"*. Given a large amount of click-through data from the commercial image search engine, we propose a text2image model for image retrieval by textual queries, which compares favorably to recent deep learning based alternatives. Moreover, different cross-media relevance estimators have their own different mechanisms to compute the relevance, so they may complement each other. Hence, computing cross-media relevance using one estimator tends to be limited. We propose cross-media relevance fusion which combines relevance scores from multiple cross-media relevance estimators.

As text, images and videos are three distinct modalities, they have to be represented in a common space wherein the cross-media relevance between them can be computed. Hence, the choice of the common space is an important point for cross-media relevance computation. Previous works [16, 17, 20, 33, 36] for multimedia retrieval prefer to represent the visual and lingual modalities in a common latent subspace, before computing their relevance. However, there are other alternatives of common space. So our second question arises as: *"What common space is suited for cross-media relevance computation?"*. In [8], we explore utilizing a deep visual feature space as the common space for sentence retrieval by visual queries. We propose Word2VisualVec that predicts a deep visual feature representation from textual input, and compute the cross-media relevance in the visual feature space exclusively. To the best of our knowledge, we are the first to solve the sentence retrieval problem in the visual space only.

The remaining sections are organized as follows. We describe the current state-of-the-art in Section 2, followed by our proposed models and experimental results in Section 3. Conclusions and further work are given in Section 4.

MM '17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10.
DOI: https://doi.org/10.1145/3123266.3123963

2 RELATED WORK

The key of cross-media relevance computation is to represent distinct modalities into a common space. What matters are forms of the embeddings and objectives to be optimized. So we review the state of the art in these two aspects.

Regarding the forms, the main stream is to place affine transformations on the different modality sides to construct a latent space [2, 5, 27]. Depending on the choice of objectives, the embedding technique is known as Canonical Correlation Analysis (CCA) if one aims to maximize the correlation between embedding vectors of distinct modalities [25, 27], or as Polynomial Semantic Indexing (PSI)[2] if a marginal ranking loss is minimized. In [24], Pan *et al.* propose to minimize the distance of relevant pairs in the latent space, with regularization terms to preserve the inherent structure in each original space. A recent work by Yao *et al.* [36] considers a joint use of CCA and PSI, achieved by firstly finding a latent space by CCA and then re-adjusting the space to incorporate ranking preferences from click-through data. Habibian *et al.* [13] leverage a textual projection matrix and a visual projection matrix to embed both videos and sentences into a latent subspace.

For the success of deep learning in computer vision and natural language processing, we observe an increasing use of such techniques as an alternative to the affine transformation. In [38], for instance, Yu *et al.* use a deep Convolutional Neural Network (CNN) for image embedding, while keep the transformation at the text side. In the DeViSE model developed by Frome *et al.* [12], the common space is formed by a pre-trained word2vec model [22], where the embedding vector of a text is obtained by average pooling of the vectors of its words. In a follow-up work, Norouzi *et al.* employ word2vec for both text and image embedding [23]. In their ConSE model, an image is embedded into the word2vec space, achieved by a convex combination of the word embedding vectors of the visual labels predicted to be most relevant to the image. Kiros *et al.* [16] use Long Short Term Memory (LSTM) and CNN to embed sentence and image with a ranking loss which ensures that relevant sentences for an image rank above irrelevant ones, also ensures relevant images for a sentence rank above irrelevant images. Wang *et al.* [32] introduce a two-branch neural network to project images and sentences into a latent subspace, using a ranking based loss similar to [16].

Previous works [3, 24, 36] aim to propose a new model to compute relevance among different multimedia data, while we propose cross-media relevance fusion which is designed to combine different models as an extension to methods for cross-media relevance computation. Moreover, different from the existing works [16, 16, 32, 39] that rely on a joint subspace, we propose to perform sentence retrieval by visual queries directly in the visual space. This change is important as it allows us to completely remove the learning part from the visual side and focus our energy on learning an effective mapping from natural language text to the visual feature space.

3 WORK IN PROGRESS

3.1 Image Retrieval by Textual Queries

Given an unlabeled image x and a textual query q, we aim to construct a real-valued function $f(x, q)$ that computes the cross-media relevance for the given image-query pair. Similar to previous works

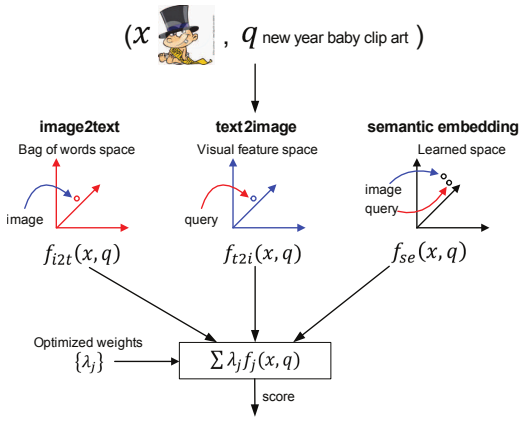

Figure 1: A conceptual diagram of the proposed cross-media relevance fusion.

[7, 24, 38], we also compute the relevance based on large-scale click-through data, denoted as $\mathcal{D} = \{(textual\ query, image, click)\}$. We propose a text2image model, inspired by [11, 31] but redesigned to better exploit click-through data and cross-media relevance fusion that combines distinct cross-media relevance estimators [7, 9].

text2image. text2image presents a novel test query by a set of images selected from large-scale click-through data, thus computing cross-media similarity between the test query and a given image boils down to comparing the visual similarity between the given image and the selected images. Specially, given a test query q, we first retrieve the top k most similar textual queries, denoted as $\{q_1, \ldots, q_k\}$, from $\mathcal{D}$. In order to represent the test query q by a set of truly relevant images, for each candidate image x_i from the j-th neighbor query q_j, we estimate the relevance between the test query and the candidate image by jointly considering the relevance between x_i and q_j and the relevance between q_j and q, *i.e.,*

$$sim_{t2i}(x_i, q) := \log(click_{i,j}) \cdot sim(q, q_j). \tag{1}$$

Accordingly, we sort all the candidate images in descending order by $sim_{t2i}(x_i, q)$, obtaining an ordered list of images $\{x_1, \ldots, x_{k'}\}$. Note that for a candidate image that is associated with multiple queries, its sim_{t2i} score is accumulated over the queries. Consequently, cross-media relevance between the image x and the textual query q is computed as a weighted sum of the visual similarity between x and $\{x_1, \ldots, x_{k'}\}$. That is,

$$f_{t2i}(x, q) := \frac{1}{k'} \sum_{i=1}^{k'} sim(x, x_i) \cdot sim_{t2i}(x_i, q). \tag{2}$$

Cross-media relevance fusion. A conceptual diagram of cross-media relevance fusion is illustrated in Fig. 1. Given some existing cross-media relevance estimators, such as image2text [26], semantic embedding models [12, 23], we can obtain a series of relevance scores for each given image-query pair. For a given image-query pair, let $\{f_i(x, q) | i = 1, \ldots, d\}$ be cross-media relevance scores computed by d distinct models. We consider the following late fusion strategy, for its simplicity and flexibility to employ a number of

Table 1: Comparison to the State-of-the-art for image retrieval by textual query on Clickture-dev.

Method	$NDCG_{25}$
CCA [25]	0.5055
CCL [24]	0.5059
BoWDNN [3]	0.5089
MRW-NN [33]	0.5104
RCCA [36]	0.5112
text2image (this work)	0.5153

off-the-shelf learning to rank techniques:

$$f_\Lambda(x,q) := \sum_{i=1}^{d} \lambda_i \cdot \varphi(f_i(x,q)), \tag{3}$$

where $\Lambda = \{\lambda_i\}$ are weights to be optimized, and $\varphi(\cdot)$ is a sigmoid function for rescaling the input.

For the fusion weights, the simplest choice is to take uniform weights. Despite its simplicity, this choice often works well in practice when the similarity functions to be fused are relatively close in terms of their performance and complementary to each other. Once some ground truth data are provided, a range of learning to rank algorithms can be employed to find better weights. We utilize Coordinate Ascent [21], a greedy algorithm capable of directly optimizing (non-differentiable) performance metrics such as Normalized Discounted Cumulated Gain (NDCG) adopted in our experiments.

Experimental Results. We compare our proposed text2image with a number of state-of-the-art models on Clickture-dev [15]. As shown in Table 1, our simple text2image model gives a larger $NDCG_{25}$ score than all other advanced heavily trained models. Moreover, we combine the text2image with four models (PSI[2], DeViSE[12], ConSE[23], image2text[26]) by our proposed cross-media relevance fusion, which further improves the performance. The detailed results on can be referred to [9].

Inspired by the result that simple text2image compares favorably to the heavily trained models, our ongoing work is trying to reveal the merit and limit of the current image retrieval models. To this end, we introduce query visualness, a quantifiable property for categorizing large-scale queries. Query visualness is found to be correlated with image retrieval performance. We initially have observed that the advanced models are good at answering queries of high visualness, such as the query of *wolf*, while it performs worse on queries of low visualness, such as the query of *quote and saying*.

3.2 Sentence Retrieval by Visual Queries

Inspired from the good performance of our text2image model compared to the state-of-the-art, we propose *Word2VisualVec*, a deep neural network architecture that learns to predict a visual feature representation from textual input [8, 29]. Different from text2image that depends on utilizing a set of images to represent textual queries, Word2VisualVec directly predicts a visual feature representation for text. We apply Word2VisualVec to sentence retrieval by visual

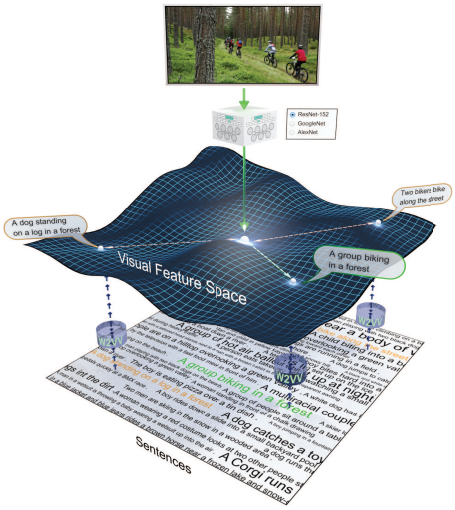

Figure 2: We propose to perform sentence retrieval by visual queries in a visual feature space exclusively, which is achieved by Word2VisualVec (W2VV).

queries which ranks a set of sentences in terms of their relevance to a given image or video. And we propose to compute their relevance in a visual feature space exclusively, as illustrated in Fig. 2.

Word2VisualVec. For the ease of reference, let $\phi(x) \in \mathbb{R}^d$ be a d-dimensional feature vector of visual instance x and q be textual input. To handle sentences of varied length, we choose to first vectorize each sentence as vector $s(q)$. In particular, we consider two common text vectorization strategies, *i.e.,* bag-of-words and word2vec. Bag-of-words is a classical representation used in text analysis. Each dimension in a bow vector corresponds to the occurrence of a specific word in the input text. For word2vec, we obtain the embedding vector of the sentence by mean pooling over its words. We train a word2vec model on Flickr tags, as recent studies suggested that it better captures visual relationships than its counterpart learned from web documents [4, 19]. The sentence vector $s(q)$ goes through subsequent hidden layers until it reaches the output layer $r(q)$, which resides in the visual feature space. More concretely, by applying l affine transformations on $s(q)$, followed by an element-wise ReLU activation $\sigma(z) = \max(0, z)$, we obtain the output layer $r(q)$ of Word2VisualVec by:

$$\begin{aligned} h_1(q) &= \sigma(W_1 s(q) + b_1), \\ h_i(q) &= \sigma(W_i h_{i-1}(q) + b_i), i = 2, ..., l-2, \\ r(q) &= \sigma(W_l h_{l-1}(q) + b_l), \end{aligned} \tag{4}$$

where $W_.$ and $b_.$ indicate affine transformation and bias terms respectively. Putting all trainable parameters together, we represent it as $\theta = [W_., b_.]$. In order to reconstruct the visual feature $\phi(x)$ directly from q, we use Mean Squared Error (MSE) as our objective function. We train Word2VisualVec to minimize the overall MSE loss on a given training set $\mathbb{D} = \{(x, q)\}$, containing a number of relevant image/video-sentence pairs:

$$\underset{\theta}{\operatorname{argmin}} \sum_{(x,q)\in\mathbb{D}} (r(q) - \phi(x))^2. \tag{5}$$

Table 2: Comparison to the State-of-the-art for sentence retrieval by image queries on Flickr8k and Flickr30k.

	Flickr8k			Flickr30k		
	R@1	R@5	R@10	R@1	R@5	R@10
Ma *et al.* [20]	24.8	53.7	67.1	33.6	64.1	74.9
Kiros *et al.* [16]	23.7	53.1	67.3	32.9	65.6	77.1
Klein *et al.* [17]	31.0	59.3	73.7	35.0	62.0	73.8
Lev *et al.* [18]	31.6	61.2	74.3	35.6	62.5	74.2
Plummer *et al.* [28]	–	–	–	39.1	64.8	76.4
Wang *et al.* [32]	–	–	–	40.3	68.9	**79.9**
Word2VisualVec	**35.2**	**64.6**	**76.7**	**41.6**	**69.2**	79.2

Sentence retrieval by image queries. After Word2VisualVec trained on image-sentence pairs, sentences can be directly mapped into a deep visual feature space by efficient forward computation through the Word2VisualVec network. Given a test image, we rank all candidate sentences in terms of their cosine similarity with the given image in the visual space. In this experiment, we use bag-of-words for sentence vectorization which yields better performance than word2vec. Table 2 presents the performance of the State-of-the-art models on two popular benchmark sets, Flickr8k [14] and Flickr30k [37]. Word2VisualVec compares favorably against the state-of-the-art. Notice that Plummer *et al.* [28] employ extra bounding-box level annotations. Still our results are better, which indicates that we can expect further gains by including locality in the Word2VisualVec representation. As all the competitor models use joint subspaces, the results justify the viability of directly using the deep visual feature space as common space for sentence retrieval by image queries. Additionally, Word2VisualVec is designed to predict a visual feature representation of text, so it can also be used for text embedding. In [6], we employ Word2VisualVec for tag embedding to enrich the current low-level input to LSTM on the top of a neural image captioning model [30], which leads to better performance.

Sentence retrieval by video queries. Word2VisualVec is used in a principled manner for sentence retrieval by video queries, transforming an input sentence to a video feature vector, let it be visual or visual-audio. For the sake of clarity, we term the video variant *Word2VideoVec*. The visual feature vector of each video is obtained by averaging the feature vectors of its frames. For the audio feature, we extract a bag of quantized Mel-frequency Cepstral Coefficients (MFCC) [10] and concatenate it with the previous visual feature. Word2VideoVec is trained on relevant video-sentence pairs to predict such a visual-audio feature, as a whole, from textual input. In order to verify the viability of Word2VideoVec, we participated in the Video to Text Description task at TRECVID 2016 organized by NIST [1]. The test set consists of 1,915 videos collected from Twitter Vine. For each test video, participants are asked to rank all sentences in the two provided sets, denoted as set *A* and set *B*. Due to space limit, we only report results on the set *A*. Results on the set *B* are similar. NIST also provides a training set of 200 videos, which we consider insufficient for training Word2VideoVec. Instead, we learn the network parameters on MSR-VTT [34], with hyper-parameters tuned on the provided TrecVid training set. We use word2vec to vectorize sentence as the training data is limited. The performance metric is Mean Inverted Rank at which the annotated item is found. As shown in Fig. 3, Word2VideoVec leads the evaluation in the context of 21 submissions from seven teams worldwide. Moreover, the results can be further improved by predicting the visual-audio feature.

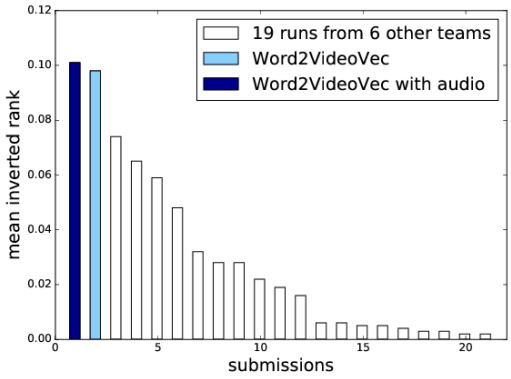

Figure 3: Comparison to the State-of-the-art for sentence retrieval by video queries on TrecVid 2016 benchmark.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we summarize our study on cross-media relevance computation. For image retrieval by textual query, we propose text2image using a large amount of click-through data, which compares favorably to deep learning based alternatives. And the results can be further improved by our proposed cross-media relevance fusion. For sentence retrieval by visual queries, we propose to compute cross-media relevance in a deep visual feature space. Our proposed Word2VisualVec outperforms the state-of-the-art, which justifies the viability of directly using the deep visual feature space as common space. Moreover, the models described in this paper have resulted in a winning entry in the Microsoft Image Retrieval Challenge at ACM MM 2015 and Video to Text Description task at TRECVID 2016, which further shows the viability of our proposed models.

Based on the current works, we consider the following directions important for future research: 1) As humans, we usually focus on specific regions of an image when looking at it. Hence, integrating attention mechanism [35] into cross-media retrieval model is meaningful. 2) As videos used in our experiments are short, we adopt average pooling on video frames to obtain the video feature. However, it leads to losing some temporal clues in the video. Exploiting temporal order of video frames is important for video-related retrieval. 3) According to our ongoing work, we observe that current advanced image retrieval models are not good at address queries of low visualness, and the majority of the real-user queries are of this type. Addressing these types of queries will be a valuable topic for future work.

Acknowledgments. This work was supported by the National Social Science Foundation of China (No. 12&ZD141). The author thanks Xirong Li for many helpful comments to improve this paper.

REFERENCES

[1] G. Awad, J. Fiscus, D. Joy, M. Michel, A. Smeaton, W. Kraaij, G. Quenot, M. Eskevich, R. Aly, R. Ordelman, G. Jones, B. Huet, and M. Larson. 2016. TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking. In *TRECVID*.

[2] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, C. Cortes, and M. Mohri. 2009. Polynomial Semantic Indexing. In *NIPS*.

[3] Y. Bai, W. Yu, T. Xiao, C. Xu, K. Yang, W. Ma, and T. Zhao. 2014. Bag-of-Words based Deep Neural Network for Image Retrieval. In *MM*.

[4] S. Cappallo, T. Mensink, and C. G. M. Snoek. 2015. Image2Emoji: Zero-shot Emoji Prediction for Visual Media. In *MM*.

[5] C. Deng, X. Tang, J. Yan, and W. Liu. 2016. Discriminative Dictionary Learning With Common Label Alignment for Cross-Modal Retrieval. *TMM* 18, 2 (2016), 208–218.

[6] J. Dong, X. Li, W. Lan, Y. Huo, and C. G. M. Snoek. 2016. Early Embedding and Late Reranking for Video Captioning. In *MM*.

[7] J. Dong, X. Li, S. Liao, J. Xu, D. Xu, and X. Du. 2015. Image Retrieval by Cross-Media Relevance Fusion. In *MM*.

[8] J. Dong, X. Li, and C. G. M. Snoek. 2016. Word2VisualVec: Image and Video to Sentence Matching by Visual Feature Prediction. *arXiv preprint arXiv:1604.06838* (2016).

[9] J. Dong, X. Li, and D. Xu. 2017. Cross-Media Similarity Evaluation for Web Image Retrieval in the Wild. TMM (under review), (2017).

[10] F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *MM*.

[11] Q. Fang, H. Xu, R. Wang, S. Qian, T. Wang, J. Sang, and C. Xu. 2013. Towards MSR-Bing Challenge: Ensemble of Diverse Models for Image Retrieval. In *MSR-Bing IRC 2013 Workshop*.

[12] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. 2013. DeViSE: A Deep Visual-semantic Embedding Model. In *NIPS*.

[13] A. Habibian, T. Mensink, and C. G. M. Snoek. 2014. VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events. In *MM*.

[14] M. Hodosh, P. Young, and J. Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR* 47, 1 (2013), 853–899.

[15] X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li. 2013. Clickage: Towards Bridging Semantic and Intent Gaps via Mining Click Logs of Search Engines. In *MM*.

[16] R. Kiros, R. Salakhutdinov, and R. S. Zemel. 2015. Unifying Visual-semantic Embeddings with Multimodal Neural Language Models. *TACL* (2015).

[17] B. Klein, G. Lev, G. Sadeh, and L. Wolf. 2015. Associating Neural Word Embeddings with Deep Image Representations using Fisher Vectors. In *CVPR*.

[18] G. Lev, G. Sadeh, B. Klein, and L. Wolf. 2016. RNN Fisher Vectors for Action Recognition and Image Annotation. In *ECCV*.

[19] X. Li, S. Liao, W. Lan, X. Du, and G. Yang. 2015. Zero-shot Image Tagging by Hierarchical Semantic Embedding. In *SIGIR*.

[20] L. Ma, Z. Lu, L. Shang, and H. Li. 2015. Multimodal Convolutional Neural Networks for Matching Image and Sentence. In *ICCV*.

[21] D. Metzler and B. Croft. 2007. Linear Feature-based Models for Information Retrieval. *Inf. Retr.* 10, 3 (2007), 257–274.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*.

[23] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. 2014. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *ICLR*.

[24] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui. 2014. Click-through-based Cross-view Learning for Image Search. In *SIGIR*.

[25] Y. Pan, T. Yao, X. Tian, H. Li, and C.-W. Ngo. 2014. Click-through-based Subspace Learning for Image Search. In *MM*.

[26] Y. Pan, T. Yao, K. Yang, H. Li, C.-W. Ngo, and T. Wang, J.and Mei. 2013. Image Search by Graph-based Label Propagation with Image Representation from Dnn. In *MM*.

[27] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos. 2014. On the Role of Correlation and Abstraction in Cross-modal Multimedia Retrieval. *TPAMI* 36, 3 (2014), 521–535.

[28] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-phrase Correspondences for Richer Image-to-sentence Models. In *ICCV*.

[29] C. G. M. Snoek, J. Dong, X. Li, X. Wang, Q. Wei, W. Lan, E. Gavves, N. Hussein, D. C. Koelma, and A. W. M. Smeulders. 2016. University of Amsterdam and Renmin University at TRECVID 2016: Searching Video, Detecting Events and Describing Video. In *TRECVID Workshop*.

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and Tell: a Neural Image Caption Generator. In *CVPR*.

[31] L. Wang, S. Cen, H. Bai, C. Huang, N. Zhao, B. Liu, Y. Feng, and Y. Dong. 2013. France Telecom Orange Labs (Beijing) at MSR-Bing Challenge On Image Retrieval 2013. In *MSR-Bing IRC 2013 Workshop*.

[32] L. Wang, Y. Li, and S. Lazebnik. 2016. Learning Deep Structure-preserving Image-text Embeddings. In *CVPR*.

[33] F. Wu, X. Lu, J. Song, S. Yan, Z. M. Zhang, Y. Rui, and Y. Zhuang. 2016. Learning of Multimodal Representations with Random Walks On the Click Graph. *TIP* 25, 2 (2016), 630–642.

[34] J. Xu, T. Mei, T. Yao, and Y. Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*.

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

[36] T. Yao, T. Mei, and C.-W. Ngo. 2015. Learning Query and Image Similarities with Ranking Canonical Correlation Analysis. In *ICCV*.

[37] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions. *TACL* 2 (2014), 67–78.

[38] W. Yu, K. Yang, Y. Bai, H. Yao, and Y. Rui. 2015. Learning Cross Space Mapping via DNN using Large Scale Click-Through Logs. *TMM* 17, 11 (2015), 2000–2007.

[39] Y. Yu, H. Ko, J. Choi, and G. Kim. 2017. End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering. In *CVPR*.