Semantic Modeling of Textual Relationships in Cross-modal Retrieval

Jing Yu[1], Chenghao Yang[2], Zengchang Qin[2(✉)], Zhuoqian Yang[2], Yue Hu[1], and Zhiguo Shi[3]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{yujing02,huyue}@iie.ac.cn
[2] Intelligent Computing and Machine Learning Lab, Beihang University, Beijing, China
{alanyang,zcqin,yzhq97}@buaa.edu.cn
[3] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
szg@ustb.edu.cn

Abstract. Feature modeling of different modalities is a basic problem in current research of cross-modal information retrieval. Existing models typically project texts and images into one embedding space, in which semantically similar information will have a shorter distance. Semantic modeling of textural relationships is notoriously difficult. In this paper, we propose an approach to model texts using a featured graph by integrating multi-view textual relationships including semantic relationships, statistical co-occurrence, and prior relationships in knowledge base. A dual-path neural network is adopted to learn multi-modal representations of information and cross-modal similarity measure jointly. We use a Graph Convolutional Network (GCN) for generating relation-aware text representations, and use a Convolutional Neural Network (CNN) with non-linearities for image representations. The cross-modal similarity measure is learned by distance metric learning. Experimental results show that, by leveraging the rich relational semantics in texts, our model can outperform the state-of-the-art models by 3.4% on 6.3% in accuracy on two benchmark datasets.

Keywords: Textual relationships · Relationship integration · Cross-modal retrieval · Knowledge graph · Graph Convolutional Network

1 Introduction

Cross-modal information retrieval (CMIR), which enables queries from one modality to retrieve information in another, plays an increasingly important role in intelligent searching and recommendation systems. A typical solution of CMIR is to project features from different modalities into one common semantic space in order to measure cross-modal similarity directly. Therefore, feature

C. Douligeris et al. (Eds.): KSEM 2019, LNAI 11775, pp. 24–32, 2019.
https://doi.org/10.1007/978-3-030-29551-6_3

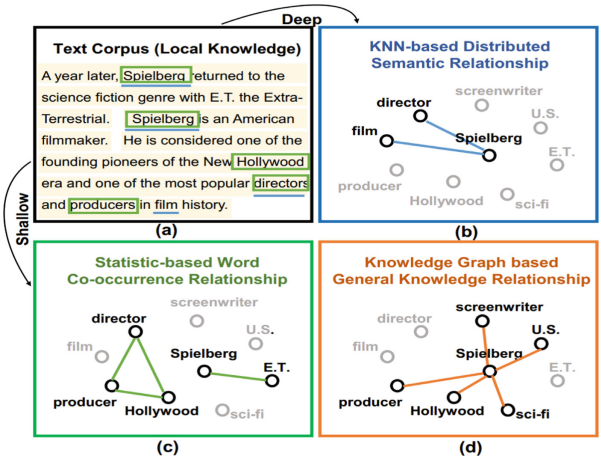

Fig. 1. (a) The original text; (b) distributed semantic relationship; (c) word co-occurrence relationship; and (d) general knowledge relationship.

representation is fundamental for CMIR research and has great influence on the retrieval performance. Recently, Deep Neural Networks (DNN) achieve superior advances in cross-modal retrieval [7,17]. For text-image retrieval, much effort has been devoted to vector-space models, such as the CNN-LSTM network [7], to represent multimodal data as "flat" features for both irregular-structured text data and grid-structured image data. For image data, CNN can effectively extract hierarchies of visual feature vectors. However, for text data, the "flat" features are seriously limited by their inability to capture complex structures hidden in texts [9] – there are many implicit and explicit textual relationships that characterize syntactic rules in text modeling. Nevertheless, the possibility of infusing prior facts or relationships (e.g., from a knowledge graph) into deep textual models is excluded by the great difficulty it imposes.

Early works attempt to learn shallow statistical relationships, such as co-occurrence [11] or location [8]. Later on, semantic relationship based on syntactic analysis [4] or semantic rules between conceptual terms are explored. Besides, semantic relationship derived from knowledge graphs (e.g., Wikidata [14]) has attracted increasing attention. A most recent work [17] models text as featured graphs with semantic relationships. However, the performance of this practice heavily relies on the generalization ability of the word embeddings. It also fails to incorporate general human knowledge and other textual relationships. To illustrate the above point, a text modeled by different types of relationships is shown in Fig. 1. It can be observed in the KNN graph (Fig. 1-b) that *Spielberg* is located relatively far away from *Hollywood* as compared to the way *director* is to *film*, whereas in the common sense knowledge graph given in (Fig. 1-d), these two words are closely related to each other as they should be. Figure 1-c shows the less-frequent

subject-predicate relation pattern (e.g. *Spielberg* and *E.T.*) which is absent in the KNN-based graph. The above analysis indicates that graph construction can be improved by fusing different types of textual relationships, which is the underlying motivation of this work.

In this paper, we propose a GCN-CNN model to learn textual and visual features for similarity matching. The novelty is on the in-depth study of textual relationship modeling for enhancing the successive correlation learning. The key idea is to explore the effects of multi-view relationships and propose a graph-based integration model to combine complementary information from different relationships. Specifically, besides semantic and statistic relationships, we also exploit fusion with the relational knowledge bases for acquiring common sense about entities and their semantic relationships, thus resulting in a knowledge-driven model. TensorFlow implementation of the model is available at https://github.com/yzhq97/SCKR.

2 Methodology

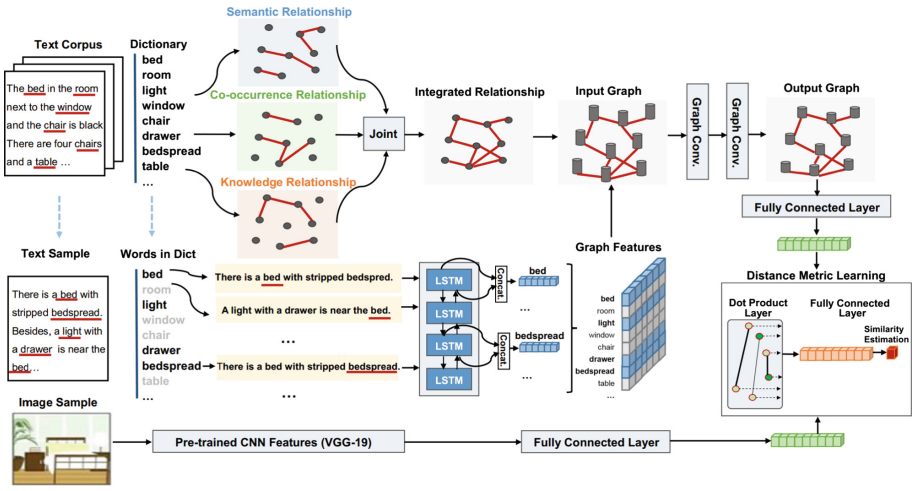

Fig. 2. The schematic illustration of our proposed framework for cross-modal retrieval.

In this paper, a dual-path neural network (as shown Fig. 2) is proposed to learn multimodal features and cross-modal similarity in an end-to-end mode. It mainly consists of three parts: (1) *Text Modeling* (top in Fig. 2): each text is represented by a featured graph by combining multi-view relationships, that is also the key idea and will be elaborated later. Graph construction is performed off-line and the graph structure is identical for all the texts in the dataset. Then we adopt Graph Convolutional Network (GCN) [2], containing two layers of convolution modules, to progressively enhance the textual representations over the

constructed graph. The last FC layer projects the text features to the common semantic space; (2) *Image Modeling* (bottom in Fig. 2): we use pre-trained Convolutional Neural Network (CNN), i.e., VGGNet [13], for visual feature learning. Similar to text modeling, the last FC layer is fine-tuned to project visual features to the same semantic space as the text; (3) *Distance Metric Learning* (right in Fig. 2): the similarity between textual and visual features are measured via distance metric learning. An inner product layer is used to combine these two kinds of features, followed by a FC layer with a sigmoid activation to output the similarity scores. We use ranking-based pairwise loss function formalized in [6] for training, which can maximize the similarity of positive text-image pairs and minimizes the similarity of negative ones.

2.1 Fine-Grained Textual Relationship

In this section, we introduce the construction of graph structure to represent each text. As is mentioned above, all the texts share the same graph. Given the training texts, we extract all the nouns to form a dictionary and each noun corresponds to a vertex in the graph. The vertex set is denoted as V. Edges are the integration of the following relationships from different views.

Distributed Semantic Relationship (SR). Following the distributional hypothesis [3], words appear in similar context may share semantic relationship, which is critical for relation modeling. To model such semantic relationship, we build a semantic graph denoted as $G_{SR} = (V, E_{SR})$. Each edge $e_{ij(SR)} \in E_{SR}$ is defined as follows:

$$e_{ij(\mathrm{SR})} = \begin{cases} 1 & \text{if } w_i \in N_k(w_j) \text{ or } w_j \in N_k(w_i) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $N_k(\cdot)$ is the set of k-nearest neighbors computed by the cosine similarity between words using *word2vec* embedding and k is the neighbor numbers, which is set to 8 in our experimental studies.

Word Co-occurrence Relationship (CR). Co-occurrence statistics have been widely used in many tasks such as keyword extraction and web search. Although the appearance of word embeddings seems to eclipse this method, we argue that it can serve as effective backup information to capture infrequent but syntax-relevant relationships. Each edge $e_{ij(CR)} \in E_{CR}$ in the graph $G_{CR} = (V, E_{CR})$ indicates that the words w_i and w_j co-occur at least ϵ times. The CR model can be formulated as below:

$$e_{ij(\mathrm{CR})} = \begin{cases} 1 & \text{if } Freq(w_i, w_j) \geq \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $Freq(w_i, w_j)$ denotes the frequency that w_i and w_j appear in the same sentence in the dataset, we define ϵ as the threshold to rule out noise, which aims to achieve better generalization ability and improve computation efficiency. We empirically set ϵ to be 5.

General Knowledge Relationship (KR). General knowledge can effectively support decision-making and inference by providing high-level expert knowledge as complementary information to training corpus. However, it is not fully covered by task-specific text. In this paper, we utilize the triples in Knowledge Graphs (KG), i.e. (Subject, Relation, Predicate), which well represent various relationships in human commonsense knowledge. To incorporate such real-world relationships, we construct the graph $G_{KR} = (V, E_{KR})$ and each edge $e_{ij(KR)} \in E_{KR}$ is defined as below:

$$e_{ij(\mathrm{KR})} = \begin{cases} 1 & \text{if } (w_i, relation(w_i, w_j), w_j) \in D \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where D refers to a given knowledge graph. In this paper, we adopt wikidata [14] in our experiments. For simplification, we ignore the types of relationships in KG and leave it for the future work.

Graph Integration. Different textual relationships capture information from different perspectives. It is conceivable that the relationship integration will fuse semantic information. We simply utilize the union operation to obtain multi-view relationships. $G = (V, E)$, where the edge set E satisfying:

$$E = E_{SR} \cup E_{CR} \cup E_{KR} \tag{4}$$

2.2 Graph Feature Extraction

Previous work [17] adopts Bag-of-Words (BoW), i.e., the word frequency, as the feature of each word in the text. However, this kind of feature is not informative enough to capture the rich semantic information. In this paper, we propose a kind of context-aware features for word-level representations. We first pretrain a Bi-LSTM in the text parts of the training set to predict the corresponding category labels, then sum up the concatenated outputs of Bi-LSTM of each word over every mention in the text to obtain the word representation. Such representation is context-relevant and can better incorporate the content-specific semantics in the text. From our experiment observation, our proposed context-aware graph features can achieve +2% overall retrieval performance lift compared with traditional BoW features. Due to the space limitation, we omit the BoW experimental results and focus on our proposed Bi-LSTM features.

3 Experimental Studies

Datasets. In this section, we test our models on two benchmark datasets: Cross-Modal Places [1] (CMPlaces) and English Wikipedia [10] (Eng-Wiki). CMPlaces is one of the largest cross-modal datasets providing weakly aligned data in five modalities divided into 205 categories. We follow the way in [17] for sample generation, resulting in 204,800 positive pairs and 204,800 negative pairs for training, 1,435 pairs for validation and 1,435 pairs for test. Eng-Wiki is the most

widely used dataset in literature. There are 2,866 image-text pairs divided into 10 categories. We generate 40,000 positive samples and 40,000 negative samples respectively from the given 2,173 pairs for training. The remaining 693 pairs are for test. We use MAP@100 to evaluate the performance. The density for all models over two datasets is much less than 1%, indicating that our models are not trivial dense matrix.

Implementation Details. We set the dropout ratio 0.2 at the input of the last fully connected layer, learning rate 0.001 with an Adam optimization, and regularization weight 0.005. The parameters setting for loss function follows [17]. In the final semantic mapping layers of both text path and image path, the reduced dimensions are set to 1,024 for both datasets. The Bi-LSTM model is pretrained on classification task on Eng-wiki and CMPlaces, respectively.

Table 1. MAP score comparison on two benchmark datasets.

Method	Q_T	Q_I	Avg.	Dataset
CCA [10]	18.7	21.6	20.2	Eng-Wiki
SCM [10]	23.4	27.6	25.5	
LCFS [16]	20.4	27.1	23.8	
LGCFL [5]	31.6	37.8	34.7	
GMLDA [12]	28.9	31.6	30.2	
GMMFA [12]	29.6	31.6	30.6	
AUSL [18]	33.2	39.7	36.4	
JFSSL [15]	41.0	**46.7**	43.9	
GIN [17]	76.7	45.3	61.0	
SR [ours]	83.5	41.4	62.4	
SCR [ours]	84.3	42.6	63.4	
SKR [ours]	83.9	42.0	62.9	
SCKR [ours]	**84.9**	44.0	**64.4**	
BL-ShFinal [1]	3.3	12.7	8.0	CMPlaces
Tune(Free) [1]	5.2	18.1	11.7	
TuneStatReg [1]	15.1	22.1	18.6	
GIN [17]	19.3	16.1	17.7	
SR [ours]	18.6	15.8	17.2	
SCR [ours]	25.4	20.3	22.8	
SKR [ours]	24.8	20.5	22.6	
SCKR [ours]	**28.5**	**21.3**	**24.9**	

Comparison with State-of-the-Art Methods. In the Eng-Wiki dataset, we compare our model to some state-of-the-art (SOTA) retrieval models, which are listed in Table 1. We observe that SCKR achieves the best performance on the average MAP scores and slightly inferior to JFSSL on the image query (Q_I),

which confirms that our relation-aware model can bring an overall improvement over existing CMIR models. Especially, text query (Q_T) gains remarkable 8.2% increase over the SOTA model GIN, which proves that our model leads to better representation and generalization ability for the text query. In the large CMPlaces dataset, compared with the previous SOTA models, SCKR also achieves 6.3% improvement compared to TuneStatReg [1].

Ablation Study. In this section, we conduct ablation experiments to evaluate the influence of the components in our proposed SCKR model. We compare SCKR model to three ablated versions, i.e., SR, SCR and SKR. The retrieval performance is also listed in Table 1. Compared to SR, both SCR and SKR achieve a significant improvement on both datasets (i.e., +5% on CMPlaces and +2% on Eng-Wiki). It indicates that either co-occurrence or the commonsense knowledge could provide complementary information to the distributed semantic relationship modeling. By integrating all kinds of textual relationships (SCKR), we obtain further promotion on MAP scores, especially on the relation-rich CMPlaces dataset. It is because that SR, CR or KR alone focuses on different views of relationships and their integration could bring more informative connections to the relational graph, thus facilitating information reasoning.

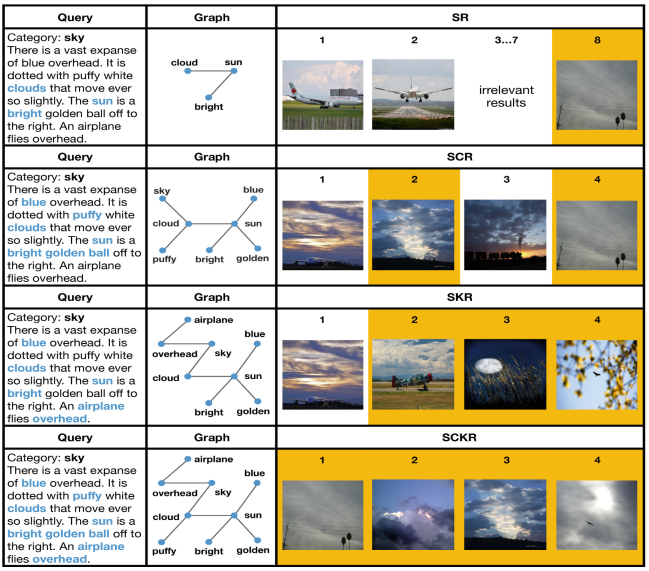

Fig. 3. Some samples of text query results using four of our models on the CMPlaces dataset. The corresponding relation graphs are shown in the second column. The retrieval results are given in the third column.

Qualitative Analysis. Fig. 3 gives an example for the text-query task on SCKR and three baseline models. We show the corresponding relation graphs and the retrieved results. We observe that SR captures the least relationships and the

results are far from satisfaction, which necessitates the exploration of the richer textual relationship. SCR can effectively emphasize the descriptive textual relationship (e.g. "*sun-ball*" and "*sun-bright*"), which is infrequent but informative for better understanding the content. Notice that, only SKR incorporates the relationship between "*overhead*" and "*airplane*" through "*sky-overhead-airplane*" inference path, which indicates that general knowledge is beneficial in relation inference and information propagation. The SCKR model leverages the advantages of different models and achieves the best performance.

4 Conclusions

In this paper, we proposed a graph-based approach to integrate multi-view textual relationships, including the semantic relationship, statistical co-occurrence, and pre-defined knowledge graph, for text modeling in the CMIR tasks. A GCN-CNN framework is proposed for feature learning and cross-modal correlation modeling. Experimental results on both two benchmark datasets show that our model can significantly outperforms the state-of-the-art models, especially for text queries. In the future work, we can extend this model to other cross-modal areas such as automatic image captioning and video captioning.

Acknowledgement. This work is supported by the National Key Research and Development Program (Grant No. 2017YFB0803301).

References

1. Castrejon, L., Aytar, Y., Vondrick, C., Pirsiavash, H., Torralba, A.: Learning aligned cross-modal representations from weakly aligned data. In: CVPR (2016)
2. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: NIPS, pp. 3837–3845 (2016)
3. Harris, Z.S.: Distributional structure. Word **10**(2–3), 146–162 (1954)
4. Jiang, C., Coenen, F., Sanderson, R., Zito, M.: Text classifcation using graph mining-based feature extraction. Knowl. Based Syst. **23**(4), 302–308 (2010)
5. Kang, C., Xiang, S., Liao, S., Xu, C., Pan, C.: Learning consistent feature representation for cross-modal multimedia retrieval. TMM **17**(3), 370–381 (2015)
6. Kumar, V.B.G., Carneiro, G., Reid, I.: Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions. In: CVPR, pp. 5385–5394 (2016)
7. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: ECCV, pp. 1908–1917 (2017)
8. Mihalcea, R., Tarau, P.: Textrank: bringing order into text. In: EMNLP, pp. 404–411 (2004)
9. Qin, Z., Yu, J., Cong, Y., Wan, T.: Topic correlation model for cross-modal multimedia information retrieval. Pattern Anal. Appl. **19**(4), 1007–1022 (2016)
10. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: ACMMM, pp. 251–260. ACM (2010)
11. Rousseau, F., Vazirgiannis, M.: Graph-of-word and TWIDF: new approach to ad hoc IR. In: CIKM, pp. 59–68 (2013)

12. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: a discriminative latent space. In: CVPR, pp. 2160–2167 (2012)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
14. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
15. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. PAMI **38**(10), 2010–2023 (2016)
16. Wang, K., He, R., Wang, W., Wang, L.: Learning coupled feature spaces for cross-modal matching. In: ICCV, pp. 2088–2095 (2013)
17. Yu, J., et al.: Modeling text with graph convolutional network for cross-modal information retrieval. In: Hong, R., Cheng, W.-H., Yamasaki, T., Wang, M., Ngo, C.-W. (eds.) PCM 2018. LNCS, vol. 11164, pp. 223–234. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00776-8_21
18. Zhang, L., Ma, B., He, J., Li, G., Huang, Q., Tian, Q.: Adaptively unified semi-supervised learning for cross-modal retrieval. In: IJCAI, pp. 3406–3412 (2017)