2D-Convolution Based Feature Fusion for Cross-Modal Correlation Learning

Jingjing Guo[1,2], Jing Yu[1,2(✉)], Yuhang Lu[1,2], Yue Hu[1], and Yanbing Liu[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{guojingjing,yujing02,luyuhang,huyue,liuyanbing}@iie.ac.cn

Abstract. Cross-modal information retrieval (CMIR) enables users to search for semantically relevant data of various modalities from a given query of one modality. The predominant challenge is to alleviate the "heterogeneous gap" between different modalities. For text-image retrieval, the typical solution is to project text features and image features into a common semantic space and measure the cross-modal similarity. However, semantically relevant data from different modalities usually contains imbalanced information. Aligning all the modalities in the same space will weaken modal-specific semantics and introduce unexpected noise. In this paper, we propose a novel CMIR framework based on multi-modal feature fusion. In this framework, the cross-modal similarity is measured by directly analyzing the fine-grained correlations between the text features and image features without common semantic space learning. Specifically, we preliminarily construct a cross-modal feature matrix to fuse the original visual and textural features. Then the 2D-convolutional networks are proposed to reason about inner-group relationships among features across modalities, resulting in fine-grained text-image representations. The cross-modal similarity is measured by a multi-layer perception based on the fused feature representations. We conduct extensive experiments on two representative CMIR datasets, i.e. English Wikipedia and TVGraz. Experimental results indicate that our model outperforms state-of-the-art methods significantly. Meanwhile, the proposed cross-modal feature fusion approach is more effective in the CMIR tasks compared with other feature fusion approaches.

Keywords: 2D-convolutional network · Inner-group relationship · Feature fusion · Cross-modal correlation · Cross-modal information retrieval

Y. Hu—Co-corresponding author.

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

J. M. F. Rodrigues et al. (Eds.): ICCS 2019, LNCS 11537, pp. 131–144, 2019.
https://doi.org/10.1007/978-3-030-22741-8_10

1 Introduction

With the explosive increase of online multimedia data, such as image, text, video, and audio, there is a great demand for intelligent search across different modalities. Cross-modal information retrieval (CMIR), which aims to enable queries of one modality to retrieve semantically relevant information of other modalities, plays an important role for realizing intelligent search. Since data from different modalities are heterogenous in feature representations, it's challenging to directly measure the cross-modal similarity. The typical solution for the problem of "heterogeneous gap" is to learn a common semantic space and align the features of different modalities in such space [13,15,17]. The cross-modal similarity can be directly measured by computing the feature distance in the common semantic space.

However, the common semantic space is equal for all the modalities, which ignores the fact that semantically relevant data from different modalities contains imbalanced information. For example, an image and its textual description convey the same semantics, but containing unequal amount of information. The image usually contains complex visual scenes which cannot be completely described by a few words. On the other hand, the text describes more background content beyond the visual information. Not all the fine-grained information between text and image can be aligned exactly. Therefore, projecting different modalities into the same space will loss some important modality-specific information and introduce some extra noise.

Recently, a few novel works explore to capture the cross-modal correlations based on feature fusion and compute the cross-modal similarity directly without common space learning [10,22,24]. Yu *et al.* [24] have demonstrated the advantages of element-wise product feature fusion for cross-modal correlation modeling. They propose a dual-path neural networks to learn visual features and textual features respectively. Then element-wise product is utilized to fuse two modal features for the successive distance metric learning. However, such simple feature fusion approach is not capable to generate expressive text-image features that can fully capture the complex correlations across modalities. More effective feature fusion approaches [3,7] have been proposed in the visual question answering (VQA) tasks and proved to be effective to improve the answer accuracy. Cross-modal information retrieval also need specific feature fusion approach to capture complex relationship between multi-modal features. How to design effective feature fusion approach for cross-modal information retrieval tasks has not been well studied.

In this paper, we propose a **2D-Conv**olution based **F**eature **F**usion (**2D-ConvFF** for short) approach to explore more complex interactions between multi-modal features for enhancing their semantic correlations and improving the performance of cross-modal information retrieval. A dual-path neural network is proposed to respectively extract the image features and text features. Then 2D-ConvFF is utilized for multi-modal feature fusion. Specifically, the learnt multi-modal features are preliminarily fused by a cross-modal feature matrix. Then we propose the 2D-convolutional networks to reason about inner-group relationships across different modal features for fine-grained text-image

feature fusion. Cross-modal similarity is measured by a multi-layer perception given the fused features. The model is trained by a ranking-based pairwise loss function. Compared with previous works, our proposed 2D-ConvFF can leverage the advantages of convolutional networks to capture inner-group relations across multi-modal features with small amount of parameters.

The main contribution of this work is the proposed 2D-convolutional networks for enhancing multi-modal feature fusion and eventually improving the cross-modal information retrieval performance. This also leads to the views of how to nicely capture more complex correlations across multi-modal data to learn more informative representations for cross-modal information retrieval, which are problems not fully explored yet.

2 Related Work

The mainstream solution for cross-modal information retrieval is to learn a common semantic space for multi-modal data, and directly measure their similarity. Traditional statistical correlation analysis methods, such as CCA [17] and its variants [15,16], are representative solutions for such case, which learn a subspace that maximizes the pairwise correlations between two modalities. Considering the importance of semantic information, some methods explore the semantics of category to constrain the common semantic space learning, such as semi-supervised methods [20,25] and supervised methods [18,21]. With the advance of deep learning in multimedia applications, deep neural networks (DNN) are used in cross-modal information retrieval [1,23]. This kind of methods generally construct a dual-path neural network to represent different modalities and jointly learn their correlations in an end-to-end mode. He *et al.* [4] apply two convolution networks to map images and texts into a common space and measure the cross-modal similarity by cosine distance.

Apart from common semantic space learning, another kind of solutions are based on correlation measurement. These methods are designed to train an end-to-end network to directly predict cross-modal correlations between different modal data [10,22,24] by multi-modal feature fusion. Wang *et al.* [22] propose to use two branch networks to project the images and texts to the same dimension. Then they fuse the two branches via element-wise product and the model is trained with an regression loss to predict the similarity score. Concatenation, element-wise addition and element-wise product are the basic approaches in cross-modal feature fusion [10,22,24]. It is very difficult for such a simple model to mine complex fine-grained relationships between multi-modal features.

Recently, some works exploit to fuse multi-modal features via more complex interactions. Huang *et al.* [5] measure multiple local similarities within a few timesteps, and aggregate them with hidden states to obtain a final matching score as the desired global similarity. Bilinear models, such as Multi-modal Compact Bilinear (MCB) [3], Multi-modal Low-rank Bilinear (MLB) [7] and Multi-modal Factorized Bilinear (MFB) [7], have been proposed for multi-modal feature fusion, especially for the visual question answering (VQA) tasks. Despite the remarkable progress of previous studies, these methods mainly capture the

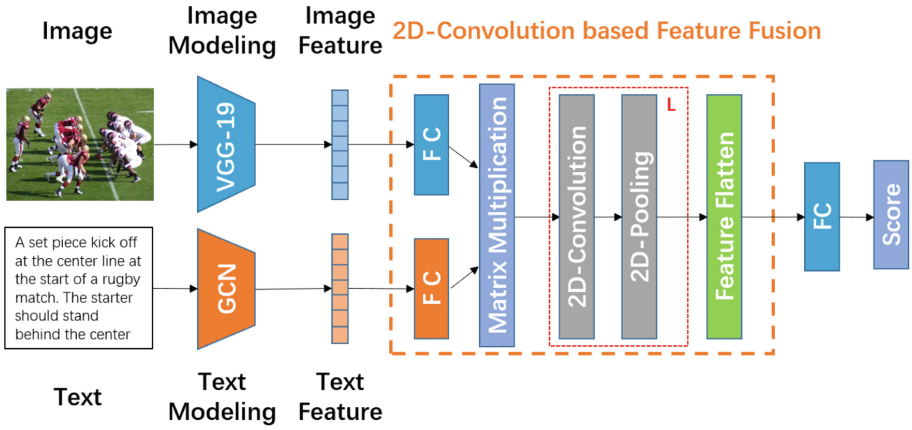

Fig. 1. This overall framework of the 2D-Convolution based Feature Fusion (2D-ConvFF). Given a image and a text, we use VGG-19 to get the image feature vector and use GCN to get the text feature vector. FC represents the fully connected layer. After that, we get two feature vectors which have the same dimensions. Matrix Multiplication denotes that we use two feature vectors get a feature matrix by matrix multiplication. Next, we use L successive convolution layers and max-pooling layers mine internal relations of feature groups between multi-modal features. Feature Flatten flattens the above fused feature matrix. Finally, we use a fully connected layer to attain the similarity score.

pairwise correlations between multi-modal features with high computational complexity. In our work, we propose a 2D-Convolution based Feature Fusion approach to explore inner-group interactions between multi-modal features with relatively small amount of parameters for improving cross-modal information retrieval.

3 Methodology

In this section, we introduce the novel cross-modal information retrieval framework based on cross-modal feature fusion. The overall framework is illustrated in Fig. 1. It is a dual-path neural network to simultaneously learn the feature representations of the texts and images. Then the proposed 2D-convolution based feature fusion method is applied to fuse the text features and image features. The cross-modal similarity score is measured by a multi-layer perception given the fused features. Our framework mainly contains four components: text modeling, image modeling, 2D-convolution based feature fusion and the objective function. We will describe each component in detail in the following sections.

3.1 Text Modeling

Most of existing deep neural network methods for cross-modal information retrieval use word-level features, e.g. bag-of-words and word embeddings [2,11], or

sequential features, e.g. RNN [12], to represent texts. Besides word-level semantics, the semantic relations between words are also informative. Kipf *et al.* [8] propose Graph Convolutional Network (GCN), which has good performance in modeling the relational semantics of texts and improve the accuracy of text classification. We follow the extended work [24], which combines the structural information and semantic information, to represent a text by a word-level featured graph (bottom in Fig. 1). We firstly extract the most common words from the text corpus and use a pre-trained word2vec [11] embedding to represent each word. Then each extracted word is corresponding to a vertex in the graph. We construct the edge set by computing k-nearest neighbors of each vertex based on the cosine similarity between word word2vec embeddings. Each text is represented by a bag-of-words vector and the word frequency serves as the 1-dimensional feature of the corresponding vertex. Then, we adopt Graph Convolutional Network (GCN) to enhance the text features based on the word-level relational connections and obtain the text representations. Given a text T, we learn the text representation f_T by the GCN model $V_T(\cdot)$, formally denoted as: $f_T = V_T(T)$.

3.2 Image Modeling

Convolutional neural network has achieved a great success in image recognition and other computer vision tasks. Simonyan *et al.* [8] show the increase of convolutional layer depth is beneficial to the improvement of accuracy. Based on this study, in the image modeling path (top in Fig. 1), we utilize the pre-trained VGG-19 [19] to extract the features from output of the last fully connected layer as the image features. Given the fixed image features, a fully connected layer is used to fine-tune the feature representations according to the downstream task. Given an image I, we learn the image representation f_I based on the pre-training VGG-19 model with fine-tuning stage $V_I(\cdot)$ denoted as: $f_I = V_I(I)$.

3.3 2D-Convolution Based Feature Fusion

The proposed 2D-Convolution based feature fusion method (middle in Fig. 1) mainly consists of two parts: feature matrix construction and feature fusion. The detailed procedure is illustrated in Fig. 2.

Feature Matrix Construction. According to Sects. 3.1 and 3.2, we obtain the text feature $f_T \in \mathbb{R}^m$ and the image feature $f_I \in \mathbb{R}^n$. Since f_T and f_I have different dimensions, we apply two fully connected layer W_t and W_i to project f_T and f_I to the same dimension. The feature mapping of the text is formally defined as:

$$\tilde{f_T} = (f_T)^T W_t \tag{1}$$

where $W_t \in \mathbb{R}^{m \times d}$ is the trained parameter of textual fully connected layer and $\tilde{f_T} \in \mathbb{R}^d$ is the refined text feature. Similarly, we define the feature mapping of the image as follows:

$$\tilde{f_I} = (f_I)^T W_i \tag{2}$$

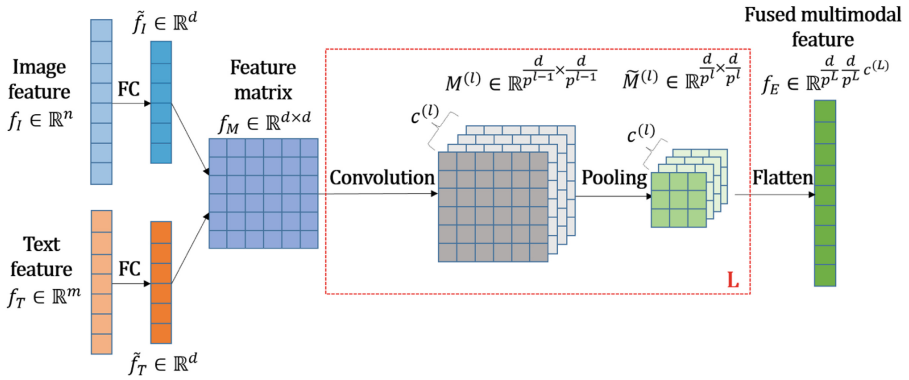

Fig. 2. Illustration of 2D-Convolution based Feature Fusion. We firstly use two fully connected layers to map the image features and text features to the same dimension. The two feature vectors are multiplied to form a feature matrix. Then the 2D-convolutional network with multiple convolutional and pooling operations are applied to enhance the cross-modal correlations. Finally, we flatten the enhanced multi-modal feature matrix and obtain the fused text-image features.

where $W_i \in \mathbb{R}^{n \times d}$ is the trained parameter of visual fully connected layer, and $\tilde{f}_I \in \mathbb{R}^d$ is the refined image feature. We utilize vector multiplication to form the feature matrix, the operation is defined as follows:

$$f_M = (\tilde{f}_T)^T \tilde{f}_I \tag{3}$$

where $f_M \in \mathbb{R}^{d \times d}$ is the feature matrix, which captures pairwise relationships between each dimension of the text feature and the image feature.

2D-Convolutional Network. We introduce a multi-layer convolutional network for multi-modal feature fusion. The constructed feature matrix can only capture pairwise relationships. Except pairwise relationships, there exists inner-group relationships between the image feature and the text feature which conveys more complex cross-modal correlations. A straightforward solution is to fuse inner-group features in the feature matrix by convolutional operations.

Given the feature matrix f_M, we construct L convolutional layers each followed by a ReLU activation and a max-pooling layer. The l_{th} convolution layer has $c^{(l)}$ convolution kernels with the kernel size $n^{(l)} \times n^{(l)}$. The strides of all convolutions are set to 1 and zero-padding is adopted to guarantee that the output after each convolution has the same size as the input. The output feature matrix of the last convolution layer is flattened and forms the fused text-image feature. The feature fusion procedure is formally defined as follows.

For the first convolution layer, the k_{th} kernel, denoted as $w^{(1,k)}$ scans over the feature matrix $M^{(0)} = f_M$ to generate a feature map $M^{(1,k)}$ defined by:

$$M_{i,j}^{(1,k)} = ReLU(\sum_{x=0}^{n^{(1)}-1}\sum_{y=0}^{n^{(1)}-1} w_{x,y}^{(1,k)} \cdot M_{i+x,j+y}^{(0)} + b^{(1,k)}) \tag{4}$$

where $n^{(1)}$ denotes the size of the kernel in the first convolution layer. The size of all convolution kernels in the same convolutional layer are the same. $b^{(1,k)}$ denotes the bias of the k_{th} kernel.

After the ReLU activation function, we apply $p \times p$ max-pooling to form the feature map $\tilde{M}^{(1,k)}$:

$$\tilde{M}_{i,j}^{(1,k)} = \max_{0\le x<p}\max_{0\le y<p} M_{i\cdot p+x,j\cdot p+y}^{(1,k)} \tag{5}$$

where p denotes the size of the pooling kernel. After the first convolution layer, the output feature maps are feeded to the next convolution layer as the input, denoted as $M^{(l)}$, where l is the number of the convolution layer. The $(l+1)_{th}$ layer of convolution is formally defined as:

$$M_{i,j}^{(l+1,k')} = ReLU(\sum_{k'=0}^{c^{(l+1)}-1}\sum_{x=0}^{n^{(l+1)}-1}\sum_{y=0}^{n^{(l+1)}-1} w_{x,y}^{(l+1,k')} \cdot \tilde{M}_{i+x,j+y}^{(l,k)} + b^{(l+1,k')}) \tag{6}$$

$$\tilde{M}_{i,j}^{(l+1,k')} = \max_{0\le x<p}\max_{0\le y<p} M_{i\cdot p+x,j\cdot p+y}^{(l+1,k')} \tag{7}$$

where $c^{(l)}$ denotes the number of feature maps in the l_{th} convolution layer. $M^{(l)} \in \mathbb{R}^{\frac{d}{p^{l-1}}\times\frac{d}{p^{l-1}}}$ and $\tilde{M}^{(l)} \in \mathbb{R}^{\frac{d}{p^{l}}\times\frac{d}{p^{l}}}$. The number of the convolution layers is set to L. Finally, We flatten $\tilde{M}^{(L)}$ to get the fused text-image feature $f_E \in \mathbb{R}^{\frac{d}{p^L}\frac{d}{p^L}c^{(L)}}$ as the output of feature fusion process.

3.4 Objective Function

After 2D-convolutional networks, we obtain a fused text-image feature vector. The fused feature vector is fed into a fully connected layer to get the similarity score. The objective function is a rank-based pairwise loss function [9], which maximizes the mean similarity score u^+ between text-image pairs from the same semantic concept and minimize the mean similarity score u^- between pairs from different semantic concepts. At the same time, our model also minimises the variance of pairwise similarity score for both matching σ^{2+} and non-matching σ^{2-} pairs. The loss function is as follows:

$$Loss = (\sigma^{2+} + \sigma^{2-}) + \lambda \max(0, m - (u^+ - u^-)) \tag{8}$$

where m is the threshold between the mean distributions of matching similarity and non-matching similarity, and λ is used to adjust the influences of mean and variance.

4 Experiment

4.1 Datasets

To evaluate the performance of our proposed model, we accomplish the general cross-modal retrieval tasks: text-query-images and image-query-texts. Experiments are conducted on two benchmark datasets: English Wikipedia (Eng-Wiki for short) [17] and TVGraz [13]. Eng-Wiki, collected from Wikipedia's "featured articles", contains 2,866 text-image pairs, where 2,173 pairs for training and 693 pairs for test. TVGraz is a collection of webpages, which contains 2,360 text-image pairs, where 1,885 pairs for training and 475 pairs for test. Both of the two datasets contains 10 categories and each text-image pair belongs to one category. We utilize VGG-19 [19] to model the images and represent each image as a 4,096-dimensional vector. Meanwhile, we adopt GCN model with the same structure as [24] to model the texts and represent each text as a 10,055-dimensional vector in Eng-Wiki and a 8,172-dimensional vector in TVGraz, respectively.

4.2 Evaluation and Implementation

In all the experiments, mean average precision (MAP) [17] is used to evaluate the retrieval results. Higher MAP indicates better retrieval performance. We implement our framework in Tensorflow. Following the same strategy as [24] for constructing training set, we selected 40,960 text-image pairs as positive samples and 40,960 text-image pairs as negative samples from the training set of the Eng-Wiki dataset. For the TVGraz dataset, 40,000 positive samples and 40,000 negative samples are randomly selected from the corresponding training set. We set the regularization weight 0.005, learning rate 0.001 with an Adam optimization, and 50 epochs for training. The dropout ratio at the input of the last fully connected layer is 0.4 for the Eng-Wiki dataset and 0.2 for the TVGraz dataset. We set the same parameters for loss function as in [24]. In the text modeling and image modeling paths, the feature dimension after fully connected layer is set to 128 for both modalities on the two datasets.

4.3 The Influence of 2D-Convolution Settings

To evaluate the influence of the 2D-convolutional networks in our proposed framework, we vary the number of convolution layers as well as the size of kernels to form several variant models based on our proposed framework. Except for the convolution part, other structures in the framework (as shown in Fig. 1) are fixed for all the variant models. By default, each convolutional layer is followed by a ReLU activation layer and a 2×2 max-pooling layer. The variant models include:

– **baseline**: we fuse the image feature and text feature by directly multiplying their feature vectors to get the feature matrix, without convolution layers. Then a non-linear mapping is followed to obtain the similarity score.

Table 1. Retrieval results of different 2D-convolution settings on the Eng-Wiki dataset.

Method	Text query	Image query	Average
Baseline	0.780	0.406	0.593
2D-ConvFF (16-1 × 1)	0.827	0.448	0.637
2D-ConvFF (16-3 × 3)	**0.874**	0.394	0.634
2D-ConvFF (16-1 × 1, 32-1 × 1)	0.863	0.419	0.641
2D-ConvFF (16-1 × 1, 32-3 × 3)	0.836	0.449	0.643
2D-ConvFF (16-3 × 3, 32-3 × 3)	0.846	**0.457**	**0.651**

Table 2. Retrieval results of different 2D-convolution settings on the TVGraz dataset.

Method	Text query	Image query	Average
Baseline	0.869	0.913	0.891
2D-ConvFF (16-1 × 1)	0.861	0.921	0.892
2D-ConvFF (16-3 × 3)	0.875	0.906	0.891
2D-ConvFF (16-1 × 1, 32-1 × 1)	0.884	0.921	0.902
2D-ConvFF (16-1 × 1, 32-3 × 3)	0.899	0.926	0.912
2D-ConvFF (16-3 × 3, 32-3 × 3)	**0.938**	**0.933**	**0.935**

- **2D-ConvFF** (16-1 × 1): we add one convolution layer in the 2D-ConvFF module. The convolution layer has 16 kernels and that the kernel size is 1×1, which can capture the pairwise relationships between each dimension of the image feature and the text feature.
- **2D-ConvFF** (16-3 × 3): this model has the same structure as the model 2D-ConvFF (16-1 × 1), only differing in the kernel size. The kernel size in this model is 3×3.
- **2D-ConvFF** (16-1 × 1, 32-1 × 1): we add two convolution layers in the 2D-ConvFF module. The first convolution layer has 16 kernels while the second convolution layer has 32 convolution kernels. The size of all the kernels is 1×1.
- **2D-ConvFF** (16-1 × 1, 32-3 × 3): this model has the same structure as the model 2D-ConvFF (16-1 × 1, 32-1 × 1), only differing in the kernel size. The kernel size of the second convolution layer in this model is 3×3.
- **2D-ConvFF** (16-3 × 3, 32-3 × 3): this model has the same structure as the model 2D-ConvFF (16-1 × 1, 32-1 × 1), only differing in the kernel size. The kernel size in this model is 3×3.

Table 1 shows the retrieval results on the Eng-Wiki dataset. Compared with the baseline model, we obtain about 4% improvement when adding only one convolution layer. The performance of 1×1 kernels is slightly superior than that of 3×3 kernels. Furthermore, we achieve another 1% improvement when adding two convolution layers. The aforementioned results indicate that the convolution layers can effectively capture inner-group relationships of multi-modal features and enhance the expressiveness of the text-image features, thus promoting the

retrieval performance. Two convolution layers with 3×3 kernels achieve the best performance on the Eng-Wiki dataset. Table 2 presents the retrieval results on the TVGraz dataset and we come out with similar conclusion as the Eng-Wiki dataset. We observe that adding convolution layers can achieve higher MAP scores compared with feature fusion by only multiplication. Meanwhile, using two convolution layers are more effective than only one convolution layer. This is because that two convolution layers can capture more complex cross-modal correlations within lager receptive field. In summary, 2D-ConvFF(16-3×3, 32-3×3) achieves the best performance on both datasets. In the following experiments, we use this model to compare with state-of-the-art methods.

4.4 Comparison with State-of-the-Art Methods

We compare our model with several state-of-the-art models. These models are widely cited works in this field, which include unsupervised models, supervised models and semi-supervised models. CCA [17] and CM [13] are unsupervised methods which adopts pairwise constrains to maximize the correlation between multi-modal features. AUSL [25] is a semi-supervised model that leverages both labelled and unlabelled data. SM [13], SCM [13], GMLDA [18], GMMFA [18], ml-CCA [15], TCM [14], LCFS [21], LGCFL [6], JFSSL [20] and GIN [24] are supervised models that use the semantic category information to capture the correlation between paired text-image pairs. Compared with GIN, when we process the training data to generate positive and negative samples, we remove some of the repeated samples to guarantee the quality of the training data. For fair comparison, we retrain the GIN model with new data and use the results to compare with our model.

Table 3. Comparison with state-of-the-art methods on the Eng-Wiki dataset.

Method	Text query	Image query	Average
CCA [17]	0.187	0.216	0.202
SCM [17]	0.234	0.276	0.255
TCM [14]	0.293	0.232	0.266
LCFS [21]	0.204	0.271	0.238
LGCFL [6]	0.316	0.378	0.347
ml-CCA [15]	0.287	0.353	0.312
CMLDA [18]	0.289	0.316	0.302
CMMFA [18]	0.296	0.316	0.306
AUSL [25]	0.332	0.397	0.364
JFSSL [20]	0.410	**0.467**	0.439
GIN [24]	0.789	0.432	0.610
2D-ConvFF (ours)	**0.846**	0.457	**0.651**

Table 3 shows the MAP scores on the Eng-Wiki dataset. Our proposed model remarkably outperforms state-of-the-art methods in all measures except for the

Table 4. Comparison with state-of-the-art methods on the TVGraz dataset.

Method	Text query	Image query	Average
CM [13]	0.450	0.460	0.455
SM [13]	0.585	0.619	0.602
SCM [17]	0.696	0.693	0.695
TCM [14]	0.706	0.694	0.695
GIN [24]	0.885	0.903	0.894
2D-ConvFF (ours)	**0.938**	**0.933**	**0.935**

image query, which is slightly inferior to JFSSL. Compared with the second best model GIN, 2D-ConvFF is about 5.7%, 2.5%, and 4.1% improvement on text query, image query, and the average performance, respectively. It proves the effectiveness of our proposed model on the most widely compared cross-modal retrieval dataset. Table 4 shows the results on the TVGraz dataset. We observe that our model achieve the best results on all the measures, which respectively gains about 5.3%, 3%, and 4.1% improvement on text query, image query, and average performance, compared with state-of-the-art model GIN. GIN use element-wise product to fuse features among the text and image, which only can attain simple correlations across modalities. While 2D-ConvFF capture the inner-group fine-grained correlations across different modalities by the convolution networks. We can conclude that our proposed feature fusion method is a great benefit to the cross-modal retrieval tasks. The complex cross-modal interactions can be well modeled by the convolution networks.

4.5 Comparison with Baseline Models

Besides our proposed model, we implement another four baseline models to evaluate the influence of different feature fusion methods on the final cross-modal retrieval performance. All the experiments are conducted on the Eng-Wiki dataset. The other four baseline models are respectively combined with four feature fusion methods, including concatenation, element-wise addition, element-wise product, and vector multiplication.

Table 5. Comparison with baseline models on the Eng-Wiki dataset.

Method	Text query	Image query	Average
Concatenation	0.103	0.088	0.096
Element-wise addition	0.111	0.083	0.097
Element-wise product	0.789	0.432	0.610
Vector multiplication	0.780	0.406	0.593
2D-ConvFF (ours)	**0.846**	**0.457**	**0.651**

The MAP scores are listed in Table 5. It's obvious that simply using linear models, i.e. concatenation and element-wise addition, to fuse the image feature and the text feature leads to extremely low MAP scores. This is due to the fact that the feature distributions of different modalities vary dramatically. Linear models may not be able to capture such complex correlations across modalities and construct informative text-image features. Compared with linear models, non-linear models (i.e. element-wise product), vector multiplication, and 2D-ConvFF obtain significant improvement on the CMIR performance. The performance of element-wise product is comparable with vector multiplication. Our proposed 2D-ConvFF further outperforms element-wise product by about 4% and achieves the best performance on all the measures. It indicates that 2D-ConvFF achieves more effective fusion of multi-modal features by capturing their complex relationships sufficiently. As a result, our 2D-ConvFF can generate more informative image-text feature representations and improve the retrieval results remarkably.

5 Conclusion

In this paper, we propose a 2D-Convolution based Feature Fusion (2D-ConvFF) framework for cross-modal information retrieval. By introducing convolution operations on the cross-modal feature matrix, we capture more complex correlations between the features of multi-modal data, resulting in more informative fused text-image representations. Experimental results on both Eng-Wiki and TVGraz datasets indicate that our proposed 2D-ConvFF can achieve significant improvements on the general cross-modal information retrieval tasks comparing with state-of-the-art methods. Besides, baseline study further proves that our 2D-ConvFF is superior to the existing feature fusion methods. In the future work, we will further explore multi-modal correlations by fusing more explicit features, such as image regions and keywords, to make the model more explainable.

References

1. Castrejon, L., Aytar, Y., Vondrick, C., Pirsiavash, H., Torralba, A.: Learning aligned cross-modal representations from weakly aligned data. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2940–2949 (2016)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805 (2018)
3. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 457–468 (2016)
4. He, Y., Xiang, S., Kang, C., Wang, J., Pan, C.: Cross-modal retrieval via deep and bidirectional representation learning. IEEE Trans. Multimedia (TMM) **18**(7), 1363–1377 (2016)
5. Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal LSTM. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2310–2318 (2017)

6. Kang, C., Xiang, S., Liao, S., Xu, C., Pan, C.: Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans. Multimedia (TMM) **17**(3), 276–288 (2017)
7. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. In: International Conference on Learning Representations (ICLR) (2017)
8. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR) (2017)
9. Kumar, B.G.V., Carneiro, G., Reid, I.: Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5385–5394 (2016)
10. Lu, Y., Yu, J., Liu, Y., Tan, J., Guo, L., Zhang, W.: Fine-grained correlation learning with stacked co-attention networks for cross-modal information retrieval. In: Liu, W., Giunchiglia, F., Yang, B. (eds.) KSEM 2018. LNCS (LNAI), vol. 11061, pp. 213–225. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99365-2_19
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (ICLR), pp. 1–12 (2013)
12. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S.: Recurrent neural network based language model. In: Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1045–1048 (2010)
13. Pereira, J.C., et al.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **36**(3), 521–535 (2014)
14. Qin, Z., Yu, J., Cong, Y., Wan, T.: Topic correlation model for cross-modal multimedia information retrieval. Pattern Anal. Appl. (PAA) **19**(4), 1007–1022 (2016)
15. Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. In: IEEE International Conference on Computer Vision (ICCV), pp. 4094–4102 (2015)
16. Rasiwasia, N., Mahajan, D., Mahadevan, V., Aggarwal, G.: Cluster canonical correlation analysis. In: International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 823–831 (2014)
17. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: ACM International Conference on Multimedia (ACM MM), pp. 251–260 (2010)
18. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: a discriminative latent space. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2160–2167 (2012)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
20. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **38**(10), 2010–2023 (2016)
21. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: IEEE International Conference on Computer Vision (ICCV), pp. 2088–2095 (2013)
22. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **41**(2), 394–407 (2018)

23. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3441–3450 (2015)
24. Yu, J., et al.: Modeling text with graph convolutional network for cross-modal information retrieval. In: Pacific-Rim Conference on Multimedia (PCM), pp. 862–871 (2005)
25. Zhang, L., Ma, B., He, J., Li, G., Huang, Q., Tian, Q.: Adaptively unified semi-supervised learning for cross-modal retrieval. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 3406–3412 (2017)