Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Cross-modal Bidirectional Translation via Reinforcement Learning

Jinwei Qi and **Yuxin Peng***
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
pengyuxin@pku.edu.cn

Abstract

The inconsistent distribution and representation of image and text make it quite challenging to measure their similarity, and construct correlation between them. Inspired by neural machine translation to establish a corresponding relationship between two entirely different languages, we attempt to treat images as a special kind of language to provide visual descriptions, so that translation can be conduct between bilingual pair of image and text to effectively explore cross-modal correlation. Thus, we propose Cross-modal Bidirectional Translation (CBT) approach, and further explore the utilization of reinforcement learning to improve the translation process. First, a **cross-modal translation mechanism** is proposed, where image and text are treated as bilingual pairs, and cross-modal correlation can be effectively captured in both feature spaces of image and text by bidirectional translation training. Second, **cross-modal reinforcement learning** is proposed to perform a bidirectional game between image and text, which is played as a round to promote the bidirectional translation process. Besides, both inter-modality and intra-modality reward signals can be extracted to provide complementary clues for boosting cross-modal correlation learning. Experiments are conducted to verify the performance of our proposed approach on cross-modal retrieval, compared with 11 state-of-the-art methods on 3 datasets.

1 Introduction

Heterogeneous data of different modalities, such as image and text, have been widely available with huge quantity on the Internet, and they commonly coexist. For example, an image often co-occurs with its corresponding text description on a web page to describe the same semantics such as objects or events. While cross-modal correlation naturally exists between image and text data to describe specific kinds of statistical dependencies. However, the inconsistent representations of different modalities make it very challenging to establish correlation and measure the semantical relevance between them. For addressing the above issue, some works have been done for bridging the gap between heterogeneous data, such as cross-modal retrieval [Rasiwasia *et al.*, 2010; Peng *et al.*, 2017a], where the data of different modalities can be retrieved flexibly by giving a query of any modality at the same time, which is extremely useful for users to retrieve whatever they want across multiple media types.

Most of existing methods [Feng *et al.*, 2014; Peng *et al.*, 2016; Rasiwasia *et al.*, 2010] follow an intuitive idea to map the data of different modalities into one common space to learn the common representation, so that the general distance metrics can be directly adopted to calculate the similarities between them, and further establish correlation among the heterogeneous data. Traditional methods [Hardoon *et al.*, 2004; Rasiwasia *et al.*, 2010] learn mapping matrices by optimizing the statistical values to project the data of different modalities into the common space. Recently, with the great progress of deep learning, many methods [Feng *et al.*, 2014; Peng *et al.*, 2017b] utilize the strong learning ability of deep neural network to perform cross-modal correlation learning. However, the aforementioned methods mainly project the feature of each modality into one common space, which cannot fully capture the complex cross-modal correlation with such unidirectional projections. Thus, we attempt to conduct bidirectional transformation between image and text, which not only transforms from image to text, but also transforms back to text from image, so as to mine the intrinsic characteristic in each modality and further enhance the cross-modal correlation through the bidirectional learning process.

Inspired by the recent progress of neural machine translation [Cho *et al.*, 2014; He *et al.*, 2016], whose key problem is to establish a corresponding relationship and make arbitrary conversion between two or more entirely different languages, we can intuitively treat images as a special kind of language, where each pixel or each region in one image can be taken as a visual word, and all of them weave together to provide rich visual descriptions. Therefore, we can conduct bidirectional translation between the bilingual pair of image and text to exploit the intrinsic characteristic in each modality and further learn the cross-modal correlation. Besides, deep reinforcement learning has recently attracted much attention. However, most of them mainly focus on video or board games [Mnih *et al.*, 2015], it is still a challenging problem to apply it

*Corresponding author.

into other scenarios with no pre-defined rules and explicit reward signals. While the bidirectional cross-modal translation process can be treated as a bidirectional game between image and text, which is a promising way to obtain reward signals for reinforcement learning. Thus we can utilize the power of reinforcement learning to promote cross-modal correlation modeling. Overall, in this paper, we propose Cross-modal Bidirectional Translation (CBT) approach with the following contributions:

- ***Cross-modal translation mechanism***. We treat image and text as bilingual sequence pairs, and utilize recurrent neural network to exploit both fine-grained local and context information within each modality. Furthermore, bidirectional translation training strategy is adopted to translate from one modality to another and also translate back, which can fully capture the intrinsic characteristic in each modality and further enhance the cross-modal correlation through the bidirectional learning process.
- ***Cross-modal reinforcement learning***. We construct a bidirectional game between image and text, which can be played as a round to promote the cross-modal bidirectional translation process through reinforcement procedure. Two kinds of reward signals are designed to extract from inter-modality correlation learning error and intra-modality reconstruction error in both two feature spaces of image and text, which can mutually boost for the cross-modal correlation learning.

To verify the performance of cross-modal correlation learning, we conduct extensive experiments on the cross-modal retrieval paradigm, and our proposed approach achieves the best retrieval accuracy compared with totally 11 state-of-the-art methods on 3 cross-modal datasets.

2 Related Works

2.1 Cross-modal Correlation Learning

Traditional cross-modal correlation learning methods [Rasiwasia *et al.*, 2010; Li *et al.*, 2003; Zhai *et al.*, 2014] mainly learn linear projections to map the features of different modalities into one common space, where the similarity of heterogeneous data can be directly calculated by general distance metric on the learned common representations. A class of representative methods utilize canonical correlation analysis (CCA) to optimize the statistical values for the cross-modal correlation learning [Rasiwasia *et al.*, 2010]. Besides, another kind of methods construct graphs to correlate the heterogeneous data in the common space, such as joint representation learning (JRL) proposed by Zhai *et al.* [Zhai *et al.*, 2014] to adopt graph regularization as well as utilize semi-supervised information.

Recently, deep learning based methods [Feng *et al.*, 2014; Peng *et al.*, 2017b] have become mainstream for cross-modal correlation learning. Correspondence autoencoder (Corr-AE) [Feng *et al.*, 2014] is proposed to jointly model the cross-modal correlation and reconstruction information. Andrew *et al.* [Andrew *et al.*, 2013] integrate CCA with deep network to propose deep canonical correlation analysis (DCCA). Wei *et al.* [Wei *et al.*, 2017] utilize convolutional neural network to learn strong representation for image and perform deep semantic matching (Deep-SM). Besides, Peng *et al.* [Peng *et al.*, 2016; 2017b] propose cross-modal multiple deep networks (CMDN) and cross-modal correlation learning (CCL) methods to fully exploit inter-modality and intra-modality correlation and further model fine-grained information for better performance. Inspired by the recent progress of generative adversarial networks, there are some attempts [Wang *et al.*, 2017] to adopt adversarial learning for cross-modal correlation modeling.

2.2 Neural Machine Translation

As a classical research topic in natural language process, machine translation aims to establish a corresponding relationship between different languages with both structural and vocabulary differences. Most of the recent works adopt deep neural network to achieve promising results in neural machine translation. Cho *et al.* [Cho *et al.*, 2014] propose recurrent neural network (RNN) based encoder-decoder architecture, where one RNN encodes a sequence of symbols into a intermediate representation, and the other decodes it into another sequence of symbols. Similarly, Sutskever *et al.* [Sutskever *et al.*, 2014] propose sequence to sequence learning with neural networks with a general end-to-end method that makes minimal assumptions on the sequence structure. Bahdanau *et al.* [Bahdanau *et al.*, 2015] improve the basic encoder-decoder architecture by joint learning to align and translate. He *et al.* [He *et al.*, 2016] adopt dual learning mechanism with reinforcement learning process to automatically learn from unlabeled data. Inspired by the recent progress in neural machine translation, we take images as a special kind of language to conduct bidirectional translation between image and text for cross-modal correlation learning.

2.3 Reinforcement Learning

Reinforcement learning generally address the problem of how the agents learn to optimize their control that maximizes cumulative reward through interactions with the environment. Mnih *et al.* [Mnih *et al.*, 2015] integrate traditional Q-learning algorithm with multi-layer network to propose deep Q network (DQN). However, most of the existing methods mainly focus on video or board games [Mnih *et al.*, 2013]. It is still quite challenging to apply it into other scenarios. There are some attempts to perform object detection [Caicedo and Lazebnik, 2015] or image caption [Ren *et al.*, 2017]. Inspired by these, we treat the cross-modal bidirectional translation process as a bidirectional game between image and text to obtain the reward signals, and utilize policy gradient methods for reward maximization, which is widely used in reinforcement learning tasks [Sutton *et al.*, 1999].

3 Our CBT Approach

As shown in Figure 1, we propose cross-modal translation mechanism to effectively model cross-modal correlation by bidirectional translation training, taking the translation process as a bidirectional game between image and text, and inter-modality and intra-modality reward signals can be extracted from correlation learning error and reconstruction error to utilize the power of reinforcement learning. We first

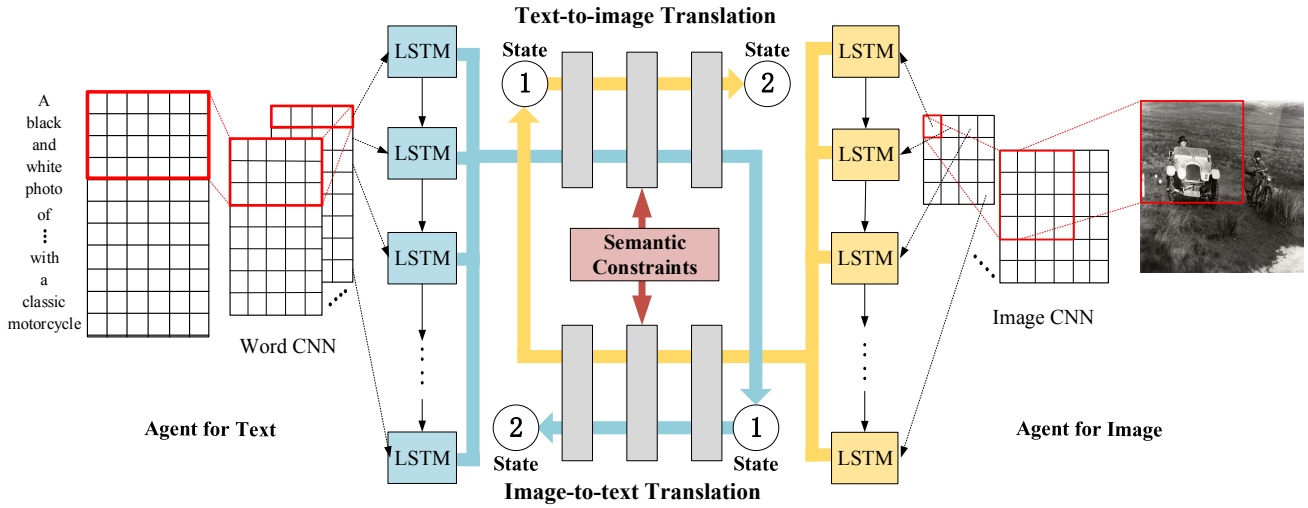


Figure 1: An overview of our proposed CBT approach. Cross-modal correlation can be modeled by bidirectional translation training, where the translation process is treated as a bidirectional game between image and text to perform reinforcement learning.

introduce the formal definition on the multimodal dataset, denoted as $D = \{I, T\}$ to conduct correlation learning, where $I = \{i_p\}_{p=1}^n$ and text $T = \{t_p\}_{p=1}^n$ with totally n instances in each modality. i_p and t_p are the p-th instance of image and text with the semantic category label c_p.

3.1 Cross-modal Translation Framework

Inspired by the sequence to sequence model in neural machine translation [Cho *et al.*, 2014], we construct cross-modal sequence model with convolutional recurrent network for both image and text, which can fully exploit the fine-grained local and spatial context information simultaneously.

For the image data, each image i_p is fed into 19-layer VGGNet [Simonyan and Zisserman, 2014] to generate separate feature vectors for different regions that contain fine-grained local information, which are obtained from the response of each filter over the a 7×7 mapping in the last pooling layer (pool5). These regions are composed as a sequence, which can be regarded as the eye movement result when glancing at the image, and denoted as $\{v_1^i, ..., v_n^i\}$ with totally n regions. Then, long short term memory (LSTM) network is adopted on these image sequence to model the fine-grained spatial context information of image. The LSTM is updated recursively with the following equations:

$$\begin{Bmatrix} i_t \\ f_t \\ o_t \end{Bmatrix} = \sigma \left(\begin{Bmatrix} W_i \\ W_f \\ W_o \end{Bmatrix} x_t + \begin{Bmatrix} U_i \\ U_f \\ U_o \end{Bmatrix} h_{t-1} + \begin{Bmatrix} b_i \\ b_f \\ b_o \end{Bmatrix} \right) \quad (1)$$

$$c_t = c_{t-1} \odot f_t + \tanh(W_u x_t + U_u h_{t-1} + b_u) \odot i_t \quad (2)$$

$$h_t = o_t \odot \tanh(c_t) \quad (3)$$

where i, f, c and o denote the activation vectors of input, forget, memory cell and output respectively. x is the input image sequences, and h is the output from the hidden units. W and U are the weight matrices and b is the bias term. $\odot$ denotes the element-wise multiplication. And σ is the sigmoid nonlinearity to activate the gate. Then, the output sequence can be obtained from LSTM and averaged as $H_i = 1/n \sum_{k=1}^n h_k^i$. Besides, we also generate image feature representation G_i from the last fully-connected layer of VGGNet to exploit the global information of image. The final encoded image representation is the averaged outputs H_i concatenated with global image representation G_i, denoted as $S^i = \{s_p^i\}$.

For the text data, each input text t_p is represented as an $n \times k$ matrix, where n is the number of words in t_p, and each word has a k-dimensional vector extracted by Word2Vec model, which is pre-trained on billions of words in Google News. Then Word CNN is adopted on the input matrix following [Kim, 2014], which is similar with the CNN for image except the 2D convolution and spatial max-pooling are replaced by temporal (1D) convolution and temporal max-pooling. We still generate the features of text fragments from the activation of last pooling layer, and split as a sequence denoted as $\{v_1^t, ..., v_n^t\}$, which contains rich fine-grained local information of text. To further exploit the context information, we also adopt LSTM to model the temporal dependency along the input text sequence, which is updated following the equations (1) to (3), where x denotes the text sequence $\{v_1^t, ..., v_n^t\}$. Similarly, the output sequence from LSTM is averaged as $H_t = 1/n \sum_{k=1}^n h_k^t$, and we also extract the global text representation G_t from the last fully-connected layer of the above Word CNN. They are concatenated as the final encoded text representation, denoted as $S^t = \{s_p^t\}$.

To conduct cross-modal translation between image and text, we construct two-pathway networks, which consist of several fully-connected layers on each pathway. Specifically, image-to-text pathway translates image representation s_p^i to its corresponding text representation s_p^t, which aims to make the translated representation from image as far as possible to be similar with s_p^t. While text-to-image pathway tends to translate text representation s_p^t back to image, which generates the translated representation close to s_p^i. Besides, the

translated representations of image and text are also translated back to their original feature space through the other pathway. Furthermore, we also connect the two-pathway networks at the middle layer, which have the shared weights and are followed by a softmax loss layer as semantic constraints, aiming to make image-to-text pathway and text-to-image pathway fully interact and keep semantic consistency.

Algorithm 1 Reinforcement training process of CBT

Require: Image training data I_{tr}, text training data T_{tr}, batchsize N, hyper-parameter α, learning rate γ.

1: **repeat**
2: Sample N encoded image representations from the CNN-RNN based network.
3: Generate N translated representations for each image s_p^i with $P(\cdot|s;\theta_{IT})$ as $s^i_{mid,1}, ..., s^i_{mid,N}$, and translate back with $P(\cdot|s;\theta_{TI})$ as $s^i_{ori,1}, ..., s^i_{ori,N}$.
4: **for** $k = 1, ..., N$ **do**
5: Set inter-modality reward r_p^{inter} for the k-th sample with equation (7).
6: Set intra-modality reward r_p^{intra} for the k-th sample with equation (8).
7: Set the total reward of the k-th sample r_p.
8: **end for**
9: Compute stochastic gradient of θ_{IT} by equation (9)
10: Compute stochastic gradient of θ_{TI} by equation (10)
11: Model updates:
$\theta_{IT} \leftarrow \theta_{IT} + \gamma \bigtriangledown_{\theta_{IT}} E(r)$,
$\theta_{TI} \leftarrow \theta_{TI} + \gamma \bigtriangledown_{\theta_{TI}} E(r)$.
12: Go through the above process from step 2 to 11 symmetrically for the game beginning from text s_p^t.
13: **until** CBT converges
14: **return** Optimized CBT model.

3.2 Reinforcement Learning Procedure

We design a bidirectional game between image and text, which is played as a round to realize the bidirectional translation process with reinforcement learning. Specifically, the two-pathway networks mentioned in Section 3.1 for cross-modal translation are denoted as $P(\cdot|s;\theta_{IT})$ and $P(\cdot|s;\theta_{TI})$ respectively, where θ_{IT} and θ_{TI} are their parameters.

For the game begins with one image s_p^i in S^i, the first state is to translate it into text feature space to get the translated representation $s^i_{mid,p}$, and we can extract the inter-modality reward r_p^{inter}, which can be obtained from correlation error that indicates the similarity between $s^i_{mid,p}$ and a sampled text instance s_p^t. Then, the second state is to conduct translation from $s^i_{mid,p}$ back to original image space as $s^i_{ori,p}$, which can generate intra-modality reward r_p^{intra} from reconstruction error. Thus, the total reward is calculated as follows:

$$r_p = \alpha r_p^{inter} + (1-\alpha) r_p^{intra} \tag{4}$$

where α is the parameter to balance the two rewards. The two-pathway networks are trained through policy gradient methods for maximizing reward, which is widely used in reinforcement learning. The reinforcement learning process is defined as follows:

$$\begin{aligned}\max_{\theta_{IT},\theta_{TI}} E(r) &= \max_{\theta_{IT},\theta_{TI}} E_{p\sim P(s^i_{mid,p}|S^i;\theta_{IT})} r_p \\ &= \max_{\theta_{IT},\theta_{TI}} \sum_{p=1}^{N} P(s^i_{mid,p}|S^i;\theta_{IT}) \\ &\quad \times \left(\alpha r_p^{inter} + (1-\alpha) r_p^{intra}\right)\end{aligned} \tag{5}$$

$$P(s^i_{mid,p}|S^i;\theta_{IT}) = \frac{\exp(r_p^{inter})}{\sum_{k=1}^{N} \exp(r_k^{inter})} \tag{6}$$

Note that we translate image to text and generate N candidate translated representations from S^i to form a mini-batch, and each of them is sampled with one text representation. The probability $P(s^i_{mid,p}|S^i;\theta_{IT})$ indicates the relevance between the p-th candidate and its sampled target text, which means those candidate pairs consisting relevant image and text in same category can get higher rewards compared with other irrelevant pairs. Similarly, we conduct the game beginning with text s_p^t with two states, namely to translate into image first, and then translate back to text. Thus the reinforcement learning process is symmetric with equation (5).

Then, taking the game beginning from image as an example, the details of objective function are introduced in the following parts. First, two kinds of rewards r_p^{inter} and r_p^{intra} in the above equation are defined as:

$$r_p^{inter} = \log(\text{norm}(\frac{s^i_{mid,p} \cdot s_p^t}{\left\|s^i_{mid,p}\right\|_2 \left\|s_p^t\right\|_2})) \tag{7}$$

$$r_p^{intra} = \log(\text{norm}(\frac{s_p^i \cdot s^i_{ori,p}}{\left\|s_p^i\right\|_2 \left\|s^i_{ori,p}\right\|_2})) \tag{8}$$

where $\|\cdot\|_2$ denotes the 2-norm, and "norm" means to normalize the similarity score to $[0, 1]$, which can be represented as the probability that indicates how similar the translated representation is with its corresponding sample, and those relevant candidate pairs would have larger similarities than others. With the objective function defined in equation (5), we calculate the stochastic gradient of θ_{IT} and θ_{TI} according to policy gradient theorem as follows:

$$\begin{aligned}\bigtriangledown_{\theta_{IT}} E(r) &= \bigtriangledown_{\theta_{IT}} E_{p\sim P(s^i_{mid,p}|S^i;\theta_{IT})} r_p \\ &= \sum_{p=1}^{N} \bigtriangledown_{\theta_{IT}} P(s^i_{mid,p}|S^i;\theta_{IT}) r_p \\ &\approx \frac{1}{N} \sum_{p=1}^{N} \bigtriangledown_{\theta_{IT}} \log\left(P(s^i_{mid,p}|S^i;\theta_{IT})\right) r_p\end{aligned} \tag{9}$$

$$\begin{aligned}\bigtriangledown_{\theta_{TI}} E(r) &= \bigtriangledown_{\theta_{TI}} E_{p\sim P(s^i_{mid,p}|S^i;\theta_{IT})} r_p \\ &= E_{p\sim P(s^i_{mid,p}|S^i;\theta_{IT})} \bigtriangledown_{\theta_{TI}} (1-\alpha) r_p^{intra} \\ &\approx \frac{1}{N} \sum_{p=1}^{N} \bigtriangledown_{\theta_{TI}} (1-\alpha) r_p^{intra}\end{aligned} \tag{10}$$

Finally, we summarize the reinforcement training process of proposed CBT in Algorithm 1. Besides, the gradient from semantic constraint is summed with policy gradient to preserve

semantic consistency during reinforcement learning process. Specifically, we obtain the intermediate representation from the middle shared layer to establish correlation between image and text, which can preserve the semantic constraints and get benefit from both inter-modality and intra-modality rewards from the cross-modal reinforcement learning process.

3.3 Implementation Details

Our proposed CBT approach is implemented by TensorFlow. The Word CNN contains 3 convolution layers, followed by ReLU activation and max-pooling. Their parameters are (384,15)→(512,9)→(256,7), where the first means the number of convolution kernels and the second is the kernel width. For image, the pre-trained CNN of 19-layer VGGNet is adopted to obtain the for image representations. The LSTM for image and text have two units in series, whose output has the same dimension with input as 300. Besides, the global image representation has 4,096 dimensions from VGGNet and global text representation has 300 dimensions from Word CNN respectively. Each of them are concatenated with the output from LSTM. Finally, two-pathway network consists of 5 fully-connected layers (4,396→3,000→2,000→1,000→600) from image to text on each pathway. The two pathways are tied at middle layer followed by softmax function for semantic constraints.

4 Experiments

4.1 Datasets

The brief introduction of 3 cross-modal datasets adopted in the experiments is given in the following paragraphs.

Wikipedia dataset [Rasiwasia *et al.*, 2010] has 10 categories with 2,866 image/text pairs. We follow [Peng *et al.*, 2016; Feng *et al.*, 2014] to split it into 3 subsets, namely 2,173 pairs for training, 231 for validation and 462 for testing.

Pascal Sentence dataset [Rashtchian *et al.*, 2010] contains 1,000 images with totally 20 categories, and each image has 5 independent sentences. Following [Peng *et al.*, 2016; Feng *et al.*, 2014], 800 image/text pairs are selected for training, while 100 pairs for testing and 100 pairs for validation.

XMediaNet dataset [Peng *et al.*, 2017a] is a large-scale cross-modal dataset with 200 categories, and has 40,000 image/text pairs, which are divided into 3 subsets, 32,000 pairs for training, 4,000 for testing and 4,000 for validation.

Method	MAP scores		
	Image→Text	Text→Image	Average
Our CBT Approach	**0.516**	**0.464**	**0.490**
CCL	0.505	0.457	0.481
ACMR	0.468	0.412	0.440
CMDN	0.487	0.427	0.457
Deep-SM	0.478	0.422	0.450
LGCFL	0.466	0.431	0.449
JRL	0.479	0.428	0.454
DCCA	0.445	0.399	0.422
Corr-AE	0.442	0.429	0.436
KCCA	0.438	0.389	0.414
CFA	0.319	0.316	0.318
CCA	0.298	0.273	0.286

Table 1: The MAP scores of cross-modal retrieval for our CBT approach and 11 compared methods on **Wikipedia** dataset.

Method	MAP scores		
	Image→Text	Text→Image	Average
Our CBT Approach	**0.602**	**0.583**	**0.592**
CCL	0.576	0.561	0.569
ACMR	0.538	0.544	0.541
CMDN	0.544	0.526	0.535
Deep-SM	0.560	0.539	0.550
LGCFL	0.539	0.503	0.521
JRL	0.563	0.505	0.534
DCCA	0.568	0.509	0.539
Corr-AE	0.532	0.521	0.527
KCCA	0.488	0.446	0.467
CFA	0.476	0.470	0.473
CCA	0.203	0.208	0.206

Table 2: The MAP scores of cross-modal retrieval for our CBT approach and 11 compared methods on **Pascal Sentence** dataset.

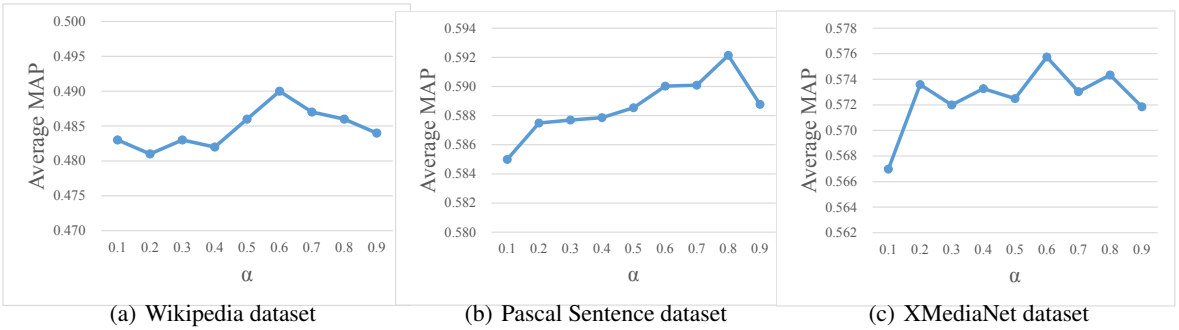
Method	MAP scores		
	Image→Text	Text→Image	Average
Our CBT Approach	**0.577**	**0.575**	**0.576**
CCL	0.537	0.528	0.533
ACMR	0.536	0.519	0.528
CMDN	0.485	0.516	0.501
Deep-SM	0.399	0.342	0.371
LGCFL	0.441	0.509	0.475
JRL	0.488	0.405	0.447
DCCA	0.425	0.433	0.429
Corr-AE	0.469	0.507	0.488
KCCA	0.252	0.270	0.261
CFA	0.252	0.400	0.326
CCA	0.212	0.217	0.215

Table 3: The MAP scores of cross-modal retrieval for our CBT approach and 11 compared methods on **XMediaNet** dataset.

4.2 Evaluation Metric and Compared Methods

To comprehensively evaluate the performance of cross-modal correlation, we perform two cross-modal retrieval tasks as: retrieving text by image query (**Image→Text**) and retrieving image by text query (**Text→Image**). We adopt mean average precision (**MAP**) as the evaluation metric, which is calculated on *all returned results* for comprehensive evaluation. It should be noted that not only *top 50 returned results* are calculated in Corr-AE [Feng *et al.*, 2014] and ACMR [Wang *et al.*, 2017], while the rest returned results are not considered.

The proposed CBT approach is compared with 11 state-of-the-art cross-modal retrieval methods to fully verify its effectiveness, including 5 traditional cross-modal retrieval methods, namely CCA [Rasiwasia *et al.*, 2010], CFA [Li *et al.*, 2003], KCCA [Hardoon *et al.*, 2004], JRL [Zhai *et al.*, 2014] and LGCFL [Kang *et al.*, 2015], and 6 deep learning based methods, namely Corr-AE [Feng *et al.*, 2014], DCCA [Andrew *et al.*, 2013], CMDN [Peng *et al.*, 2016], Deep-SM [Wei *et al.*, 2017], CCL [Peng *et al.*, 2017b] and ACMR [Wang *et al.*, 2017]. For fair comparison, all the compared methods adopt the same CNN features for both image and text, which are extracted from the CNN architectures used in our approach. Specifically, the CNN feature for image is extracted from the fc7 layer in 19-layer VGGNet [Simonyan and Zisserman, 2014] with 4,096 dimensions. While the CNN feature for text is extracted from Word CNN with the same configuration of [Kim, 2014] with 300 dimensions.



(a) Wikipedia dataset (b) Pascal Sentence dataset (c) XMediaNet dataset

Figure 2: Experiments on the influence of the parameter α in the reward function, on Wikipedia, Pascal Sentence and XMediaNet datasets. It should be noted that we report the average MAP score of Image→Text and Text→Image tasks.

Dataset	Method	MAP scores		
		Image→Text	Text→Image	Average
Wikipedia	**CBT**	**0.516**	**0.464**	**0.490**
	CBT-inter	0.505	0.443	0.474
	CBT-intra	0.499	0.434	0.466
	CBT-baseline	0.483	0.426	0.455
Pascal Sentence	**CBT**	**0.602**	**0.583**	**0.592**
	CBT-inter	0.595	0.572	0.583
	CBT-intra	0.589	0.570	0.580
	CBT-baseline	0.577	0.560	0.569
XMediaNet	**CBT**	**0.577**	**0.575**	**0.576**
	CBT-inter	0.572	0.567	0.569
	CBT-intra	0.568	0.561	0.564
	CBT-baseline	0.557	0.554	0.555

Table 4: Baseline experiments on performance of two rewards.

4.3 Comparisons with State-of-the-art Methods

The experimental results are shown in Tables 1, 2 and 3, which include the MAP scores of two retrieval tasks and their average results on 3 datasets. Obviously, our proposed CBT approach achieves the best retrieval accuracies. Among all the compared methods, we can draw the follow observations: First, most deep learning based methods achieve better retrieval accuracies than the traditional methods, where CCL has the best performance, which verifies the effectiveness of deep network. Second, traditional methods benefit from the CNN feature to get better performance than their original works with hand-crafted features, and even some of them have close accuracy with deep learning based methods, such as JRL and LGCFL. Compared with the state-of-the-art methods, our proposed CBT approach achieves promising improvement with following 2 reasons: (1) Cross-modal translation strategy to conduct bidirectional transformation between image and text to capture the cross-modal correlation in the original feature space of each modality, while the compared methods mainly model the cross-modal correlation through unidirectional projections that limit their performance. (2) Cross-modal reinforcement learning is adopted to extract intra-modality and inter-modality rewards in a bidirectional game, which can model correlation and reconstruction information in both two feature spaces of image and text simultaneously, while the compared methods only model them in single intermediate transformation.

4.4 Parameter Analysis and Baseline Comparisons

We conduct parameter experiment on the effect of key parameter α in reward function 4. The value of α ranges from 0.1 to 0.9, and results are shown in Figure 2. We further evaluate the performance with only single reward in Table 4, where "CBT-inter" means that only inter-modality reward extracted from correlation learning error is adopted, while "CBT-intra" means only intra-modality reward from reconstruction error.

From the above results, we have the following observations: (1) Compared with "CBT-baseline" which only adopts semantic constraint between two-pathway network, the two rewards can further promote the cross-modal correlation learning. (2) The retrieval accuracy becomes highest when α is larger than 0.5, and "CBT-inter" also has better performance than "CBT-intra", which indicates that correlation learning plays a more important role than modeling reconstruction information. (3) Compared with "CBT-inter" that only considers the unidirectional translation from one to another, CBT outperforms it with bidirectional translation to fully capture the cross-modal correlation. (4) CBT outperforms all the baseline methods, which verifies the effectiveness on the integration of two rewards to further promote the accuracy of cross-modal retrieval.

5 Conclusion

In this paper, we have proposed Cross-modal Bidirectional Translation (CBT) approach to conduct bidirectional translation between image and text. First, a cross-modal translation mechanism is designed to model the cross-modal correlation as well as exploit the fine-grained local and context information in original feature space of each modality. Second, cross-modal reinforcement learning is proposed to jointly model the correlation and reconstruction information as two kinds of rewards in the bidirectional game played as a round between image and text. Extensive experiments verify the effectiveness of our proposed CBT approach. In the future work, we attempt to perform unsupervised learning to exploit unlabeled data for practical applications.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 61771025 and Grant 61532005.

References

[Andrew *et al.*, 2013] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, pages 1247–1255, 2013.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

[Caicedo and Lazebnik, 2015] Juan C. Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2488–2496, 2015.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, and et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

[Feng *et al.*, 2014] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *ACM Conference on Multimedia (ACM-MM)*, pages 7–16, 2014.

[Hardoon *et al.*, 2004] David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[He *et al.*, 2016] Di He, Yingce Xia, Tao Qin, Liwei Wang, and et al. Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 820–828, 2016.

[Kang *et al.*, 2015] Cuicui Kang, Shiming Xiang, Shengcai Liao, Changsheng Xu, and Chunhong Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Transactions on Multimedia (TMM)*, 17(3):370–381, 2015.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[Li *et al.*, 2003] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. Multimedia content processing through cross-modal association. In *ACM Conference on Multimedia (ACM-MM)*, pages 604–611, 2003.

[Mnih *et al.*, 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, and et al. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, and et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Peng *et al.*, 2016] Yuxin Peng, Xin Huang, and Jinwei Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3846–3853, 2016.

[Peng *et al.*, 2017a] Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2017.

[Peng *et al.*, 2017b] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network. *IEEE Transactions on Multimedia (TMM)*, 2017.

[Rashtchian *et al.*, 2010] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010.

[Rasiwasia *et al.*, 2010] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Conference on Multimedia (ACM-MM)*, pages 251–260, 2010.

[Ren *et al.*, 2017] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2014.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112, 2014.

[Sutton *et al.*, 1999] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1057–1063, 1999.

[Wang *et al.*, 2017] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Shen Hengtao. Adversarial cross-modal retrieval. In *ACM Conference on Multimedia (ACM-MM)*, pages 154–162, 2017.

[Wei *et al.*, 2017] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Transactions on Cybernetics (TCYB)*, 47(2):449–460, 2017.

[Zhai *et al.*, 2014] Xiaohua Zhai, YuXin Peng, and Jianguo Xiao. Learning cross-media joint representation with sparse and semi-supervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 24:965–978, 2014.