Multi-modal Circulant Fusion for Video-to-Language and Backward

Aming Wu and Yahong Han
School of Computer Science and Technology, Tianjin University, Tianjin, China
{tjwam, yahong}@tju.edu.cn

Abstract

Multi-modal fusion has been widely involved in focuses on the modern artificial intelligence research, e.g., from visual content to languages and backward. Common-used multi-modal fusion methods mainly include element-wise product, element-wise sum, or even simply concatenation between different types of features, which are somewhat straightforward but lack in-depth analysis. Recent studies have shown fully exploiting interactions among elements of multi-modal features will lead to a further performance gain. In this paper, we put forward a new approach of multi-modal fusion, namely Multi-modal Circulant Fusion (MCF). Particularly, after reshaping feature vectors into circulant matrices, we define two types of interaction operations between vectors and matrices. As each row of the circulant matrix shifts one element, with newly-defined interaction operations, we almost explore all possible interactions between vectors of different modalities. Moreover, as only regular operations are involved and defined a priori, MCF avoids increasing parameters or computational costs for multi-modal fusion. We evaluate MCF with tasks of video captioning and temporal activity localization via language (TALL). Experiments on MSVD and MSRVTT show our method obtains the state-of-the-art performance for video captioning. For TALL, by plugging into MCF, we achieve a performance gain of roughly 4.2% on TACoS.

1 Introduction

Multi-modal data are widely involved in recent focuses on artificial intelligence research, e.g., visual to languages [Yao et al., 2015] and the backward of visual localization via language queries [Gao et al., 2017]. Though many efforts have been made towards feature learning of images and texts, e.g., via convolutional neural networks (CNN) [He et al., 2016] and recurrent neural networks (RNN) [Kiros et al., 2015], respectively, in-depth analysis of multi-modal fusion has been unintentionally ignored. Common-used or somewhat straightforward multi-modal fusion methods mainly include element-wise sum (Fig. 1 (a)), element-wise product (Fig. 1 (b)), or even simply concatenation (Fig. 1 (c)) between different types of features [Fukui et al., 2016; Yu et al., 2017]. As feature vectors of different modalities lie in different feature spaces, interactions or correlations might not just exist among corresponding dimensions of multi-modal vectors. Besides no interactions with concatenation, element-wise sum or product only partially explores interactions or correlations among multi-modal features, which may burden the fusion performance.

Recent studies [Fukui et al., 2016; Yu et al., 2017] have shown fully exploiting interactions among elements of multi-modal features will lead to a further performance gain. Authors in [Fukui et al., 2016; Yu et al., 2017] develop multi-modal bilinear pooling to capture pairwise interactions between multi-modal feature dimensions. As bilinear pooling defines p parameterized projection matrices, where p is the dimension of the output fused features, additional huge number of parameters are introduced into the model. Although count sketch [Fukui et al., 2016] or matrix factorization [Yu et al., 2017] were employed to shrink projection matrices, more computational costs were introduced again. Moreover, empirical studies in [Fukui et al., 2016] show that the performance gain from multi-modal fusion is guaranteed only when
the dimension $p$ of the output fused features is high, which means there still needs a large number of parameterized projection matrices. Therefore, if we skipped the parameterized projection matrices and defined a fully-interacted module of multi-modal fusion a priori, e.g., constituted with only regular operations like element-wise sum or product, we expect to get a good fusion performance with fewer parameters or lower computational costs.

In this paper, we propose a new module of Multi-modal Circulant Fusion (MCF) to fully exploit interactions among multi-modal features. In Fig. 1 (d) we show the main idea of MCF by taking the fusion of visual and textual vectors as an example. In particular, after reshaping visual or textual vectors into circulant matrices, respectively, we define two types of interaction operations between original feature vectors and the reshaped circulant matrices. Finally, we use element-wise sum to obtain the joint representation of these two cross-fused vectors. As each row of the circulant matrix shifts one elements, with newly-defined interaction operations, we almost explore all possible interactions between vectors of different modalities. Note that, as only regular operations are involved in the MCF and the proposed MCF is defined a priori, we avoid introducing new parameters or increasing computational costs for multi-modal fusion.

In the experiments, we extensively evaluate the proposed MCF with tasks of video captioning and temporal activity localization via language (TALL) [Gao et al., 2017]. Experimental results on MSVD [Chen and Dolan, 2011] and MSRVTT [Chen and Dolan, 2011] show that our method could obtain the state-of-the-art performance of video captioning. For TALL, by plugging into the MCF, we achieve a performance improvement of roughly 4.2% on TACoS [Regneri et al., 2013].

2 Related Works

As the original bilinear method [Tenenbaum and Freeman, 1997] needs a huge number of parameters and is not applied into a multi-modal task [Peng et al., 2018; Zhang et al., 2018], the method MCB [Fukui et al., 2016] uses the count sketch to reduce the number of parameters of bilinear pooling and employs bilinear pooling to solve multi-modal fusion. However, MCB always outputs a high-dimensional fused feature which increases computational cost of following processes. Whereas the work in [Yu et al., 2017] proposes a bilinear pooling method based on matrix factorization to combine multi-modal features.

Recent advances in video captioning mainly follow the encoding-decoding framework and generate captions via RNN [Zhu et al., 2017; Yang et al., 2017]. Though strengths in representation ability, CNN has not been well exploited in video captioning. In this paper, we propose a coarse-to-fine multi-stage convolutional network which includes a MCF and many convolutional decoders that generate finer video descriptions. In order to reduce the risk of vanishing gradients, inspired by the work [Zhang et al., 2016], we enforce intermediate supervisions for each stage. Experimental results on MSVD and MSRVTT show the effectiveness of our method.

As a backward direction of video captioning, Temporal Activity Localization via Language (TALL) [Gao et al., 2017] is given a temporally untrimmed video and a natural language query, and then determines the start and end moments for the described activity inside the video. By plugging MCF into the architecture of [Gao et al., 2017], we obtain a performance improvement owing to a better multi-modal fusion.

3 Multi-modal Circulant Fusion

The detailed procedures of MCF are illustrated in Fig. 2. Given two feature vectors in different modalities, e.g., the visual features $x \in \mathbb{R}^o$ and the textual features $y \in \mathbb{R}^n$, to reduce computational cost, we first utilize two projection matrices $W_1 \in \mathbb{R}^{d \times o}$ and $W_2 \in \mathbb{R}^{d \times n}$ $(d \leq \min(o, n))$ to project the two input features to a lower dimensional space.

$$V = xW_1^T$$

$$C = yW_2^T$$

where $W_1^T$ and $W_2^T$ are the transpose of $W_1$ and $W_2$.

Then we use the projection vector $V \in \mathbb{R}^d$ and $C \in \mathbb{R}^d$ to construct circulant matrix $A \in \mathbb{R}^{d \times d}$ and $B \in \mathbb{R}^{d \times d}$.

$$A = \text{circ}(V)$$

$$B = \text{circ}(C)$$

where $\text{circ}(b)$ denotes converting b to a circulant matrix.

In order to make elements in projection vector and circulant matrix fully interact, we explore two different multiplication operations. The first is shown in Eq. (3) and uses matrix multiplication between circulant matrix and projection vector.

$$F = CA$$

$$G = VB$$

The second is to have projection vector and each row vector of circulant matrix do element-wise product. The procedures are shown as follows:

$$F = \frac{1}{d}\sum_{i=1}^{d}a_i \odot C$$

$$G = \frac{1}{d}\sum_{i=1}^{d}b_i \odot V$$

Figure 2: The flowchart of Multimodal Circulant Fusion (MCF).
where $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}^d$ are row vector of circulant matrix $A$ and $B$. $\odot$ denotes operation of element-wise product. It is noted that we do not introduce new parameters in the multiplication operation.

Finally, through a projection matrix $W_3 \in \mathbb{R}^{d \times k}$, we convert the element-wise sum vector of $F \in \mathbb{R}^d$ and $G \in \mathbb{R}^d$ to target vector $M \in \mathbb{R}^k$.

4 MCF for Video Captioning

In this section, we develop a new framework for video captioning, in which we construct convolutional encoder and decoder for video-to-language translation. Note that, in the decoder, we take MCF as a base layer for coarse decoding, stacked on which with layered dilations for refined and final decoding. Thus we construct a Multi-stage Decoder with MCF.

4.1 Convolutional Encoding Network

Feature Extraction. We use pre-trained convolutional networks to extract feature for each of the $m$ video frames, which results in a vector $X_i \in \mathbb{R}^d$ for the $i$-th frame.

Discriminative Enhancing. Given two consecutive frame features $X_i$ and $X_{i+1}$ ($i = 0, 2, \cdots$), we first compute inter-frame differences $diff$. Then by a $ReLU$ operation, we add the positive values of $diff$, to $X_{i+1}$ and the absolute value of negative values in $diff$, to $X_i$. Thus, we enlarge the discrimination gap between $X_i$ and $X_{i+1}$.

\[
\text{diff} = X_{i+1} - X_i, \tag{5}
\]

\[
V_{i+1} = X_{i+1} + \text{ReLU}(\text{diff}), \tag{6}
\]

\[
V_i = X_i + \text{ReLU}(-\text{diff}), \tag{7}
\]

\[
\text{ReLU}(x) = \begin{cases} x, & \text{if } x > 0; \\ 0, & \text{else} \end{cases} \tag{8}
\]

where $i = 0, 2, \cdots$ and $V_i \in \mathbb{R}^q$ is the enhancing result of $X_i$.

Reconstruction Network. In this work, we construct a reconstruction network to learn a compact representation $Z_t$ for each video frame. The procedures are shown as follows:

\[
Z_i = \text{ReLU}(W_E \ast V_i + b_E) \tag{9}
\]

\[
R_i = W_D \ast Z_i + b_D \tag{9}
\]

\[
L_{\text{recon}} = \|X_i - R_i\|_2^2 \tag{9}
\]

where $i = 0, 1, \cdots$. $W_E \in \mathbb{R}^{1 \times q \times r}$ denotes convolutional filter and $r \leq q$. $b_E \in \mathbb{R}^q$ and $b_D \in \mathbb{R}^q$ are the bias parameters. $Z_i \in \mathbb{R}^q$ is the learned compact representation. $W_D \in \mathbb{R}^{1 \times r \times q}$ is the reconstruction filter and $R_i \in \mathbb{R}^q$ denotes reconstruction result. $L_{\text{recon}}$ denotes reconstruction loss between $R_i$ and $X_i$ and $\| \cdot \|_2$ denotes an $\ell_2$-norm. $*$ denotes a convolutional operator.

4.2 Multi-stage Convolutional Decoder with MCF

As introduced above, we stack many dilated convolutional layers [Chen et al., 2017] to form a coarse-to-fine multi-stage decoder (Fig. 3). In the following, we denote by $\hat{Y}^j = \{\hat{Y}_0^j, \ldots, \hat{Y}_{T-1}^j\}$ the predicted word sequence of the $j$-th stage of decoder, where $j \in \{0, \ldots, N_f\}$ and $(N_f + 1)$ is the number of stages. We denote by $Y = \{Y_0, \ldots, Y_{T-1}\}$ the target word sequence, where $T$ denotes sequence length. Note that we treat the initial stage $j = 0$ as coarse decoder and above stages as refined decoders. $[a, b]$ represents the concatenation of $a$ and $b$. And $\odot$ is an element-wise product operation, $\ast$ denotes a convolutional operator. $Z = \{Z_0, Z_1, \ldots\}$. The $\sigma(\cdot)$ is a sigmoid function.

MCF as a Coarse Decoder. In the bottom stage ($j = 0$), we use one dilated convolutional layer to learn a coarse decoder. At each time step $t \in [0, T - 1]$, the input to coarse decoder consists of previous target word $Y_{t-1}$ and mean vector $Z_{\text{mean}}$ of encoding output $Z$. As the inputs are two different modalities, we first use MCF to obtain the joint representation $M^0_t$.

The operation of coarse decoder can be described as follows:

\[
M^0_t = MCF(Z_{\text{mean}}, Y_{t-1}) \tag{10}
\]

\[
h^0_t = \text{tanh}(w^0_t \ast H^0_t + b^0_t) \odot \sigma(w^0_t \ast H^0_t + b^0_t) \tag{10}
\]

where $H^0_t = [M^0_{t-1}, M^0_t]$. $MCF(a, b)$ represents using MCF to fuse $a$ and $b$. $w^0_t$ and $w^0_t$ denote convolutional filters on the 0th layer. $b^0_t$ and $b^0_t$ denote bias on the 0th layer.

Refined Decoder. In this paper, our refined decoder consists of two stages. The first stage contains three dilated convolutional layers. And the second stage only includes one dilated convolutional layer which stacks on top of first stage. We take prediction of the second stage as final description.

For the first refined decoder, we first use the output $h^0_t$ of coarse decoder to compute visual attention $\varphi^0_t(Z)$. The operation of first layer in this refined decoder is shown as follows:

\[
M^1_t = w_1 \ast [h^0_t \ast \varphi^0_t(Z)] + b_1 \tag{11}
\]

\[
H^1_t = [M^0_{t-1}, M^1_t] \tag{11}
\]

\[
h^1_t = \text{tanh}(w^1_t \ast H^1_t + b^1_t) \odot \sigma(w^1_t \ast H^1_t + b^1_t) \tag{11}
\]

where $w_1$ is learnable filter to convert the channel of concatenated representation. $w^1_t$ and $w^1_t$ denote convolutional filters on the 1th layer. $b^1_t$ and $b^1_t$ denote bias on the 1th layer.
Then, operations of the next two layers in first refined decoder are shown as follows:

\[
H_t^1 = [h_t^{l-1}, h_t^{l-1}]
\]

\[
h_t^1 = \tanh(w_f^1 \ast H_t^1 + b_f^1) \odot \sigma(w_g^1 \ast H_t^1 + b_g^1)
\]

(12)

where \(r\) represents dilated rate of the layer \(l\). \(h_t^{l-1}\) denotes the output of \((l-1)\)-th layer at time step \(t\). \(w_f^1\) and \(w_g^1\) denote convolutional filters on the layer \(l\). \(b_f^1\) and \(b_g^1\) are bias.

For the second refined decoder, the procedures can be described as follows:

\[
M_{L+1}^t = h_t^L + \varphi_t^1(Z)
\]

\[
H_{L+1}^t = [M_{L+1}^t, M_{L+1}^{t+1}]
\]

\[
\text{Mid1} = \tanh(w_f^{L+1} \ast H_t^{L+1} + b_f^{L+1})
\]

\[
\text{Mid2} = \sigma(w_g^{L+1} \ast H_t^{L+1} + b_g^{L+1})
\]

\[
h_t^{L+1} = \text{Mid1} \odot \text{Mid2}
\]

(13)

where \(L\) denotes the number of layers of first refined stage. \(h_t^L\) denotes the output of layer \(L\) at time step \(t\). \(w_f^{L+1}\) and \(w_g^{L+1}\) denote convolutional filters of layer \(L+1\). \(b_f^{L+1}\) and \(b_g^{L+1}\) are bias. \(\varphi_t^1(Z)\) represents attention computed by \(h_t^L\).

In this section, we use the method [Yao et al., 2015] to compute attention (Fig. 3). Besides, based on different size of filter and dilated rate, we use different number of zero vectors to pad the input of each layer.

Finally, the \(t\)-th generated word \(Y_t^1\) of the first refined stage and \(Y_t^2\) of the second refined stage are computed as follows:

\[
\hat{Y}_t^1 \sim \text{softmax}(w_p(h_t^L + \varphi_t^1(Z)) + b_p)
\]

\[
\hat{Y}_t^2 \sim \text{softmax}(w_p(h_t^{L+1} + \varphi_t^1(Z)) + b_p)
\]

(14)

where \(w_p\) and \(b_p\) are learnable projection matrix and bias.

**Training Loss.** For each stage \(j\), we employ a cross-entropy (XE) loss.

\[
L_{XE}^j(\theta_{0:j}) = - \sum_{t=0}^{T-1} \log(p_{\theta_{0:j}}(Y_t|Y_{0:t-1}, Z))
\]

(15)

where \(Y_t\) is the ground-truth word at time \(t\). \(\theta_{0:j}\) are parameters up to the \(j\)-th stage decoder, and \(p_{\theta_{0:j}}(Y_t|Y_{0:t-1}, Z)\) is the output probability of word \(Y_t\) given by the previous word \(Y_{0:t-1}\) and encoding output \(Z\).

The training loss \(L_{\text{train}}\) is computed as follows:

\[
L_{\text{train}} = \beta_2 L_{\text{recon}} + \beta_1 \sum_{j=0}^{N_j} \lambda_j L_{XE}^j(\theta_{0:j})
\]

(16)

where \(\lambda_j, \beta_1, \) and \(\beta_2\) are hyper-parameters.

---

**5 MCF for TALL**

For temporal activity localization via language query, we plug MCF into the architecture of work [Gao et al., 2017]. Concretely, we first feed visual feature \(f_s\) and textual feature \(f_v\) into MCF (Fig. 4). For MCF, after obtaining \(F \in \mathbb{R}^d\) and \(G \in \mathbb{R}^d\) by Eq. (4), we use a projection matrix \(W_M \in \mathbb{R}^{d \times k}\) to convert \(F\) and \(G\) to \(F_s \in \mathbb{R}^k\) and \(G_v \in \mathbb{R}^k\).

\[
F_s = f_s W_M
\]

\[
G_v = f_v W_M
\]

(17)

The following operations are same as that of the work [Gao et al., 2017].

---

**6 Evaluation on Video Captioning**

We evaluate the benefit of MCF on two video captioning datasets. All results are evaluated by metrics of BLEU, METEOR, and CIDEr [Yao et al., 2015].

**6.1 Dataset and Implementation Details**

**Datasets.** MSVD [Chen and Dolan, 2011] contains 1,970 video clips. We use 1,200 clips for training, 100 clips for validation, and 670 clips for testing. MSRVTT [Xu et al., 2016] contains 10,000 video clips. We use 6,513 clips for training, 497 clips for validation, and 2,990 clips for testing.

**Video Processing.** For the MSVD dataset, we select 40 frames from each video and feed them into GoogLeNet [Szegedy et al., 2015] to extract a 1,024 dimensional representation. For the MSRVTT dataset, we select 20 frames from each video and feed them into GoogLeNet and ResNet-152 [He et al., 2016] to extract 1,024 and 2,048 dimensional representation, respectively.

**Encoding Network.** In the encoder, we set the channel \(r\) (in Eq. (9)) of the encoding temporal output \(Z_t\) to 512.

**Decoding Network.** For the multi-stage decoder, we use five dilated layers with dilated rate 1, 2, 3, 4 and 2. The number of filter channel is set to 512, 256, 256, 512 and 512, respectively. The width of filter is set to 2. For MCF, we set \(W_1 \in \mathbb{R}^{256 \times 512}\), \(W_2 \in \mathbb{R}^{256 \times 512}\) (in Eq. (1)) and \(W_3 \in \mathbb{R}^{512 \times 512}\).

**Training Details.** The vocabulary size is 12,596 for MSVD and 23,308 for MSRVTT, respectively. We use Adam optimizer with an initial learning rate of \(1 \times 10^{-3}\). We empirically set \(\beta_1\) and \(\beta_2\) to 0.9 and 0.1, respectively. And \(\lambda_0, \lambda_1\)
and $\lambda_2$ are set to 0.2, 0.2, and 0.6, respectively. Note that we do not conduct beam search in testing.

### 6.2 Experimental Results

#### MSVD Dataset

On MSVD dataset, we compare our method with other methods. The results are shown in Table 1.

It can be seen that our method outperforms all other methods on the metric of METEOR and CIDEr. Particularly, for the work [Song et al., 2017], they use multi-layer LSTM as the decoder. Our convolutional sequential decoder outperforms the performance of hLSTMmat [Song et al., 2017]. This shows that our method is effective.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2VT [Venugopalan et al., 2015]</td>
<td>41.70</td>
<td>30.90</td>
<td>-</td>
</tr>
<tr>
<td>CD+LSTM-E [Pan et al., 2016b]</td>
<td>44.30</td>
<td>31.10</td>
<td>62.10</td>
</tr>
<tr>
<td>VGG+RNN [Yu et al., 2016]</td>
<td>41.92</td>
<td>29.60</td>
<td>51.67</td>
</tr>
<tr>
<td>Tempor-attention [Yao et al., 2015]</td>
<td>48.42</td>
<td>31.70</td>
<td>65.38</td>
</tr>
<tr>
<td>GR+GRU-RCN1 [Ballas et al., 2016]</td>
<td>43.80</td>
<td>33.10</td>
<td>-</td>
</tr>
<tr>
<td>G+HRNE [Pan et al., 2016a]</td>
<td>41.40</td>
<td>32.20</td>
<td>53.90</td>
</tr>
<tr>
<td>MAMRNN [Li et al., 2017]</td>
<td>42.50</td>
<td>32.40</td>
<td>63.50</td>
</tr>
<tr>
<td>Boundary [Baraldi et al., 2017]</td>
<td>48.50</td>
<td>31.90</td>
<td>-</td>
</tr>
<tr>
<td>M+LSTM [Xu et al., 2017]</td>
<td>49.45</td>
<td>33.39</td>
<td>75.45</td>
</tr>
<tr>
<td>G+MCNN+MCF-element-wise product</td>
<td>43.65</td>
<td>32.56</td>
<td>73.80</td>
</tr>
<tr>
<td>G+MCNN+MCF-matrix multiply</td>
<td>46.46</td>
<td>33.72</td>
<td>75.46</td>
</tr>
</tbody>
</table>

Table 1: Comparison with other models on MSVD. Here ‘G’ denotes GoogLeNet. ‘MCNN’ represents our multi-stage CNN. ‘MCF-element-wise product’ and ‘MCF-matrix multiply’ mean we respectively use element-wise product and matrix multiplication in our MCF. All values are measured by percentage (%).

#### MSRVTT dataset

On MSRVTT dataset, we compare our method with representative methods. Results are shown in Table 2. Compared with all above methods which use a single kind of visual feature as input, our method obtains the best performance on METEOR and CIDEr metric. Besides, compared with the work [Xu et al., 2017] which uses many kinds of features as input, our method outperforms its performance. This shows that the performance of our method on a single kind of visual feature is valid.

Ablation Analysis. We respectively use bilinear pooling [Fukui et al., 2016], concatenation, element-wise sum and element-wise product to replace MCF used in our multi-stage CNN and keep other components of multi-stage CNN and their parameter settings unchanged. The output dimension of bilinear pooling is set to 512. The results are shown in Table 3. On MSVD and MSRVTT dataset, we can see that MCF outperforms all above fusion methods on the metric of METEOR and CIDEr. This shows that based on our convolutional architecture, MCF is effective for video captioning. Besides, through Table 1 and Table 2, we find that for our method, using matrix multiplication in MCF outperforms the performance of using element-wise product.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA-LSTM [Xu et al., 2017]</td>
<td>36.5</td>
<td>26.5</td>
<td>41.0</td>
</tr>
<tr>
<td>G+LSTM [Venugopalan et al., 2014]</td>
<td>34.6</td>
<td>25.6</td>
<td>-</td>
</tr>
<tr>
<td>C3D+SA [Yao et al., 2015]</td>
<td>36.1</td>
<td>25.7</td>
<td>35.2</td>
</tr>
<tr>
<td>R+S2VT [Venugopalan et al., 2015]</td>
<td>31.4</td>
<td>25.7</td>
<td>35.2</td>
</tr>
<tr>
<td>R+hLSTM [Song et al., 2017]</td>
<td>38.3</td>
<td>26.3</td>
<td>-</td>
</tr>
<tr>
<td>G+MCNN+MCF-matrix multiply</td>
<td>36.1</td>
<td>26.4</td>
<td>39.2</td>
</tr>
<tr>
<td>R+MCNN+MCF-matrix multiply</td>
<td>38.1</td>
<td>27.2</td>
<td>42.1</td>
</tr>
<tr>
<td>R+MCNN+MCF-element-wise product</td>
<td>37.7</td>
<td>27.1</td>
<td>41.3</td>
</tr>
</tbody>
</table>

Table 2: Comparison with other models on MSRVTT. Here ‘G’, ‘R’, and ‘C’ denote GoogLeNet, ResNet, and C3D. ‘MCF-matrix multiply’ and ‘MCF-element-wise product’ represent we respectively use matrix multiplication and element-wise product in our MCF. ‘MCNN’ represents our multi-stage CNN for video captioning.

In Fig. 5, we show some video captioning examples generated by multi-stage CNN which uses different fusion method. We can see that the captions generated by multi-stage CNN using MCF are better than that generated by multi-stage CNN using other fusion methods. Particularly, taking the first and third results as examples, our method successfully identifies ‘pasta’ and ‘makeup’, while are better than above fusion
methods. This also demonstrates that based on our architecture, MCF is an effective fusion method for video captioning.

<table>
<thead>
<tr>
<th>Fusion Method (dataset)</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element-wise sum (MSVD)+G</td>
<td>47.15</td>
<td>32.71</td>
<td>71.29</td>
</tr>
<tr>
<td>Element-wise product (MSVD)+G</td>
<td>46.21</td>
<td>32.67</td>
<td>71.92</td>
</tr>
<tr>
<td>Concatenation (MSVD)+G</td>
<td>45.25</td>
<td>32.90</td>
<td>70.64</td>
</tr>
<tr>
<td>Bilinear Pooling (MSVD)+G</td>
<td>43.61</td>
<td>32.46</td>
<td>71.11</td>
</tr>
<tr>
<td>MCF-matrix multiply (MSVD)+G</td>
<td>46.46</td>
<td>33.72</td>
<td>75.46</td>
</tr>
<tr>
<td>Element-wise sum (MSRVTT)+R</td>
<td>37.3</td>
<td>26.4</td>
<td>40.0</td>
</tr>
<tr>
<td>Element-wise product (MSRVTT)+R</td>
<td>36.8</td>
<td>26.6</td>
<td>39.3</td>
</tr>
<tr>
<td>Concatenation (MSRVTT)+R</td>
<td>36.2</td>
<td>26.3</td>
<td>39.9</td>
</tr>
<tr>
<td>Bilinear Pooling (MSRVTT)+R</td>
<td>37.5</td>
<td>26.7</td>
<td>40.9</td>
</tr>
<tr>
<td>MCF-matrix multiply (MSRVTT)+R</td>
<td>38.1</td>
<td>27.2</td>
<td>42.1</td>
</tr>
</tbody>
</table>

Table 3: The results using different fusion method in our multi-stage CNN. Here ‘G’ and ‘R’ denote GoogleNet and ResNet feature. ‘MCF-matrix multiply’ denote using matrix multiplication in MCF.

7 Evaluation on TALL

7.1 Dataset and Implementation Details

We evaluate the benefit of MCF on TACoS dataset [Regneri et al., 2013]. This dataset contains 127 videos. Each video includes two types of annotations. The first is activity labels with temporal location (start and end frame). The second is language descriptions. In total, there are 17,344 pairs of sentence and video clips. We split it in 50% for training, 25% for validation and 25% for test. In experiment, we set $W_1 \in \mathbb{R}^{96 \times 1024}$, $W_2 \in \mathbb{R}^{96 \times 1024}$ (in Eq. (1)), and $W_M \in \mathbb{R}^{96 \times 1024}$ (in Eq. (17)). All other parameter settings are same as those of the work [Gao et al., 2017].

7.2 Experimental Results

Evaluation Metric. We employ the metric used by [Regneri et al., 2013; Gao et al., 2017] to compute ‘R@n, IoU=m’.

Comparison with other methods. We compare our method with other methods on TACoS and report the result for IoU $\in \{0.1, 0.3, 0.5\}$ and Recall@{1, 5}. The results are shown in Table 4. ‘Random’, ‘Verb’ and ‘Verb+Obj’ are used by work [Gao et al., 2017] to make contrast. ‘VSA-RNN’ and ‘VST-STV’ [Karpathy and Fei-Fei, 2015] leverage images and their descriptions to learn about their correspondences. ‘CTRL-p’ and ‘CTRL-np’ denote using parameterized and non-parameterized regression loss to train CTRL [Gao et al., 2017]. And we use element-wise product in MCF.

It can be seen from Table 4 that our method outperforms all above methods on most metrics. Particularly, our method achieves a relative improvement of 4.2% on the metric R@5 with IoU=0.1. This shows that MCF is an effective fusion method for this task.

<table>
<thead>
<tr>
<th>Method</th>
<th>R@5 IoU=0.1</th>
<th>R@5 IoU=0.3</th>
<th>R@5 IoU=0.5</th>
<th>R@5 IoU=0.3</th>
<th>R@5 IoU=0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>1.62</td>
<td>1.81</td>
<td>3.28</td>
<td>3.57</td>
<td>7.03</td>
</tr>
<tr>
<td>Verb</td>
<td>8.25</td>
<td>11.24</td>
<td>14.69</td>
<td>16.46</td>
<td>21.50</td>
</tr>
<tr>
<td>Verb+Obj</td>
<td>25.74</td>
<td>24.73</td>
<td>37.13</td>
<td>48.86</td>
<td>37.13</td>
</tr>
<tr>
<td>VSA-RNN</td>
<td>7.56</td>
<td>10.77</td>
<td>15.01</td>
<td>15.50</td>
<td>23.92</td>
</tr>
<tr>
<td>VSA-STV</td>
<td>13.80</td>
<td>18.32</td>
<td>24.32</td>
<td>25.42</td>
<td>36.69</td>
</tr>
<tr>
<td>CTRL-p</td>
<td>13.05</td>
<td>17.08</td>
<td>23.16</td>
<td>25.74</td>
<td>35.62</td>
</tr>
<tr>
<td>CTRL+MCF-p</td>
<td>12.53</td>
<td>18.64</td>
<td>25.84</td>
<td>24.73</td>
<td>37.13</td>
</tr>
</tbody>
</table>

Table 4: Comparison of different methods on TACoS. ‘MCF-m’ and ‘MCF-p’ denote using matrix multiplication and element-wise product in MCF. All values are measured by percentage (%).

In Fig. 6, we show some examples about TALL task. Particularly, taking the second result as example, our method accurately determines the start and end time for the query which includes two actions.

8 Conclusion

In this paper, we propose the Multi-modal Circulant Fusion (MCF) to combine visual and text representations. We test the MCF on video captioning and TALL tasks. Experimental results on three datasets demonstrate the effectiveness of our method.

Acknowledgments

This work was supported by the NSFC (under Grant U1509206, 61472276, U1736219).
References


