T2V2T: Text-to-Video-to-Text Fusion for Text-to-Video Retrieval

Jonghee Kim1  Youngwan Lee1,2  Jinyoung Moon1,3
1Electronics and Telecommunications Research Institute (ETRI), South Korea
2Korea Advanced Institute of Science and Technology (KAIST), South Korea
3University of Science and Technology (UST), South Korea
{jhkim27, yw.lee, jymoon}@etri.re.kr

Abstract

Video-language transformers for text-to-video retrieval typically consist of a video encoder, a text encoder, and a joint encoder. The joint encoder can be categorized into 1) self-attention-based fusion and 2) unidirectional fusion based on cross-attention. The former approach performs self-attention on the concatenation of video and text embeddings. Although it allows complete interaction between text and video, the length of the input sequences makes it computationally intensive. Instead, unidirectional fusion employs rather efficient cross-attention to fuse video embeddings into text embeddings while ignoring text-to-video interaction. The text-to-video fusion is not well explored because of the information imbalance between text and video, which makes it difficult to determine which video patches can be used as queries in cross-attention. i.e., a text embedding corresponds to one or more patch embeddings, while a video patch embedding may not correspond to any text embeddings. In order to address this challenge, we devise a Bypass cross-attention (Bypass CA) which prevents matching between irrelevant video and text embedding pairs in the cross-attention. Using Bypass CA, we propose a novel bidirectional interaction approach, Text-to-Video-to-Text (T2V2T) fusion. The proposed T2V2T uses two unidirectional fusions with opposite directions, i.e., text-to-video fusion followed by video-to-text fusion. As a result, the proposed T2V2T fusion yields state-of-the-art results on MSR-VTT, DiDeMo, and ActivityNet Captions.

1. Introduction

Recently, text-to-video retrieval methods have leveraged video-language pre-training (VLP) [2, 5, 7, 9, 12, 17, 24–26, 33] since VLP has shown superior performance in downstream tasks such as text-to-video retrieval [1, 11, 28], video question answering [14, 27, 30, 32], and video captioning [11, 28, 34]. Video-language models (VLMs) for VLP commonly employ a text encoder, a video encoder, and a joint encoder. Whether or not the joint encoder is employed, we could categorize VLMs into dual and joint encoder-based models. Dual encoder-based models [2, 9, 24] employ text and video encoders to learn shared representation using contrastive learning [15]. Joint encoder-based models [5, 7, 12, 17, 26, 33] embed each modality into video and text embeddings separately and then merge the embeddings with a joint encoder for downstream tasks. Although dual encoder-based models are efficient, they performed worse than joint encoder-based models in a variety of tasks due to the lack of video-language interaction.

Existing joint encoder-based models for fusing video and text could be categorized into self-attention-based fusion (Fig. 1a) and unidirectional fusion based on cross-attention (Fig. 1b). Since self-attention-based approaches [7, 17, 25, 33] take concatenated video and text embeddings as input, they require a huge computational cost since the computational cost is proportional to the square of the input sequence length. In order to reduce the computational cost, unidirectional fusion methods [2, 5, 12, 26] fuse video embeddings into text embeddings by using cross-attention followed by self-attention on the fused text em-
beddings as shown in Fig. 1b. Although applying self-attention with only text embeddings is efficient, it results in limited unidirectional interaction, i.e., video-to-text interaction without text-to-video interaction. As a straightforward extension of unidirectional fusion, Cheng et al. [5] explore a bidirectional fusion method (Fig. 1c), where two cross-attentions transfer each video and text embedding to each other. However, the bidirectional method shows similar performance even with extra text-to-video fusion. We assume that it is due to the property of cross-attention in text-to-video fusion, which associates all patch embeddings as queries with the text embeddings, regardless of the correlation between each patch and the text embedding. For example, the number of text tokens is less than 100 in our experiments, while the number of patch tokens is up to 6,272 for 32 frames. Due to the imbalance, a text token can correspond to one or more patch tokens, while a patch token may not correspond to any text tokens. It could cause improper modification of patch tokens by using irrelevant text tokens.

To resolve this limitation, we devise Bypass cross-attention (Bypass CA), which prevents video embeddings from being modified by the use of irrelevant text. To this end, we copy query video embeddings into the key and value in addition to text features in text-to-video cross-attention to make video embeddings associate with themselves in the key instead of text features if the given sentence and each frame are not relevant to each other. With Bypass CA, we propose an effective bidirectional fusion method (Fig. 1d), called Text-to-Video-to-Text (T2V2T) fusion, by enabling text-to-video interaction. Concretely, T2V2T fusion first transfers text embeddings to video embeddings using Bypass CA and then transfers the modulated video embeddings back to text embeddings using cross-attention. For video-to-text fusion, we follow the previous unidirectional fusion methods [5, 12]. The video-fused text features are then re-integrated by self-attention. The T2V2T fusion is performed recursively for iterative interaction.

Our main contributions are summarized as follows. 1) We investigate the text-to-video interaction, which suffers from the imbalance between the number of video and text embeddings. 2) We propose a novel fusion method called T2V2T fusion by introducing Bypass CA. 3) T2V2T achieves SOTA text-to-video retrieval results on MSR-VTT [28], DiDeMo [1], and ActivityNet Captions [11].

2. Proposed Method

The proposed text-to-video retrieval follows previous VLP-based approaches, i.e., pre-training followed by fine-tuning on text-to-video retrieval. In this section, we describe our VLM with T2V2T fusion based on Bypass CA, training objectives and implementation details.

Overall Architecture is shown in Fig. 2. It consists of a text encoder, a video encoder, and a joint encoder (T2V2T fusion). The text encoder takes a sentence of length \( L \) as input and yields text features \( F_T \in \mathbb{R}^{L \times D} \), where \( D \) is an embedding dimension. The video encoder takes \( N \) frames of a video as input and returns video features \( F_V \in \mathbb{R}^{N \times P \times D} \), where \( P \) is the number of patches in a frame. The joint encoder (T2V2T fusion) takes \( F_T \) and \( F_V \) as input and outputs video-text features. The video-text features are employed for text-to-video-retrieval through a fully connected layer.

Text-to-Video-to-Text (T2V2T) Fusion is a bidirectional fusion encoder that extends the unidirectional encoder employed in [5, 12]. In [5, 12], a cross-attention is employed to fuse video features into text features without regard to text-to-video interaction. In order to enable text-to-video interaction, we introduce Bypass cross-attention (Bypass CA) in addition to the unidirectional encoder.

Since cross-attention in the text-to-video fusion of bidirectional fusion [5] (Fig. 1c) associates all frames with the given sentence, video features in each frame are modulated...
by text features without regard to the correlation between the given sentence and each frame. We assume that it is inappropriate to associate all frames with the given sentence since only a subset of the frames is relevant to the given sentence. For example, self-attention-based fusion Fig. 1a can discourage the association between irrelevant words and frames since self-attention is performed using both embeddings. As an efficient alternative, we introduce a bypass mechanism in the cross-attention by including frame features in the key and value in addition to text features, so that frame features (i.e., query) can be associated with themselves in the key instead of text features if the given sentence and each frame are not relevant to each other as follows:

\[ F_{T2V}[i] = x\text{-Attn}(F_V[i], F_T||F_V[i]), \]

where \( F_V[i] \) is video features of \( i \)-th frame. \( x\text{-Attn}(Q,KV) \) is a cross-attention where \( Q \) is a query, \( KV \) is employed as key and value. Note that Bypass CA is performed in a frame-wise manner to avoid a significant increase in computational cost.

Then, \( F_{T2V} \) is fused into text features as follows:

\[ F_{T2V2T} = s\text{-Attn}(x\text{-Attn}(F_T, F_{T2V})), \]

where \( s\text{-Attn}(\cdot) \) is a self-attention to further refine the fused features. Note that we omit a fully connected layer, GELU [10] activation function, and a layer normalization layer after Eq. (1) and Eq. (2) for simplicity. T2V2T fusion is performed \( M \) times recursively for iterative interaction.

**Objectives** In the pre-training phase, we employ three common objectives: Video-Text Contrastive loss (\( L_{VTC} \)) [15] on text and video embeddings, Video-Text Matching loss (\( L_{VTM} \)) [15, 16, 20], and Masked Language Modeling loss (\( L_{MLM} \)) [6]. For finetuning on text-to-video retrieval, we employ \( L_{VTC} \) and \( L_{VTM} \) except \( L_{MLM} \).

**Implementation details.** We follow the setup of [5] to fairly compare proposed T2V2T fusion encoder with the state-of-the-art unidirectional encoder [5]. The video encoder is based on BEiT [3], and temporal attention inspired by TimeFormer [4] is inserted before each spatial attention. The first nine layers of BERT-Base [6] are employed for the text encoder, and the last three layers are employed to initialize self-attention in the T2V2T fusion encoder. We randomly initialize the temporal attention layers in the video encoder and the cross-attention layers in the T2V2T fusion.

### 3. Experiments

In this section, we first describe experiments setup of pre-training (Sec. 3.1) and text-to-video retrieval (Sec. 3.2), then demonstrate experimental results compared to state-of-the-art VLP-based text-to-retrieval methods [2, 5, 9, 12, 13, 15] in Sec. 3.3. In addition, we show an ablation study to verify the effectiveness of Bypass CA in Sec. 3.4.

#### 3.1. Pre-training

We pre-trained our model on CC3M [23] and WebVid2M [2], which are common for VLP. For a fair comparison, we followed the training setting in [5] as shown in Tab. 3.
### Method Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>MSR-VTT</th>
<th></th>
<th>DiDeMo</th>
<th>ActivityNet Captions</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R5</td>
<td>R10</td>
<td>Avg.</td>
<td>R1</td>
</tr>
<tr>
<td>VindLU [5]</td>
<td>43.8</td>
<td>70.3</td>
<td>79.5</td>
<td>64.5</td>
<td>54.6</td>
</tr>
<tr>
<td>naïve T2V2T</td>
<td>44.3</td>
<td>70.1</td>
<td>79.3</td>
<td>64.6</td>
<td>55.1</td>
</tr>
<tr>
<td>T2V2T</td>
<td>44.4</td>
<td>70.7</td>
<td>79.5</td>
<td>64.9</td>
<td>56.0</td>
</tr>
</tbody>
</table>

Table 2. An ablation study to verify the efficacy of the proposed Bypass CA.

### Hyper-parameters for Pre-training

<table>
<thead>
<tr>
<th>config</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>optimizer</td>
<td>AdamW [19]</td>
</tr>
<tr>
<td>learning rate</td>
<td>1e-4–1e-6 (cosine decay [18])</td>
</tr>
<tr>
<td>#epochs</td>
<td>10 (warmup = 1)</td>
</tr>
<tr>
<td>batch size × #GPUs</td>
<td>64×8</td>
</tr>
<tr>
<td>spatial resolution</td>
<td>224 × 224</td>
</tr>
<tr>
<td>Augmentation</td>
<td>random resize, crop horizontal flip</td>
</tr>
<tr>
<td>#training frames</td>
<td>4</td>
</tr>
</tbody>
</table>

### Hyper-parameters for Text-to-Video Retrieval

<table>
<thead>
<tr>
<th>config</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>learning rate</td>
<td>1e-5–1e-6 (cosine decay [18])</td>
</tr>
<tr>
<td>#epochs</td>
<td>5</td>
</tr>
<tr>
<td>batch size × #GPUs</td>
<td>32×4</td>
</tr>
<tr>
<td>#training frames</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 3. Hyper-parameters for pre-training. We omit parameters which are common in the pre-training (Tab. 3).

### 3.2. Text-to-video retrieval

We evaluated VLP-based text-to-video retrieval methods on the following three datasets:

- **MSR-VTT** [28] contains 10K videos with 200K captions. We trained on 9K videos and evaluated on 1K-A test set following [2, 5, 31].
- **DiDeMo** [1] contains 10K videos with 41K captions. We trained on training set with 8,395 videos and evaluated on test set with 1,004 videos following [5, 13, 21].
- **ActivityNet Captions** [11] contains 20K videos with 100K captions. We trained on training set with 10K videos and evaluated on validation set with 4.9K videos following [5, 12, 21].

We followed the training setting of [5] as shown in Tab. 4 for a fair comparison.

### 3.3. Experimental Results

We compared the proposed method with state-of-the-art VLP-based methods in terms of text-to-video retrieval recall as shown in Tab. 1. Since the number of images and videos used in the pre-training phase affects the performance, we also report the number of images and videos. Our proposed method achieved the best recall under 5.5M pre-training data. Notably, the proposed method obtained gains in recall at 1, +0.6 on MSR-VTT, +1.4 on DiDeMo, and +1.0 on ActivityNet Captions compared to the second best method, VindLU (5.5M) [5]. In addition, the proposed method achieved better results than the methods pre-trained on 17-400M data except VindLU (17M/25M) [5] for DiDeMo and ActivityNet Captions text-to-video retrieval.

### 3.4. Ablation Study

In order to verify the effectiveness of the proposed T2V2T and Bypass CA, we compared three methods in terms of text-to-video retrieval recall on three text-to-video retrieval datasets; baseline unidirectional fusion [5], naïve T2V2T ([5] + cross-attention), and our T2V2T ([5] + Bypass CA). As shown in Tab. 2, naïve T2V2T yielded slightly worse results than its baseline unidirectional fusion. It is the same phenomenon as [5] that bidirectional fusion does not improve over unidirectional fusion. We speculate that it is due to the property of cross-attention in text-to-video fusion, which associates irrelevant frames and text. In contrast, the proposed Bypass CA improved over unidirectional fusion by taking advantage of the bypass mechanism.

### 4. Conclusion

In this paper, we have proposed Bypass cross-attention to deal with the information imbalance between video and text. Using Bypass cross-attention, the proposed T2V2T fusion achieves state-of-the-art results in text-to-video retrieval. We will extend this work to a variety of downstream tasks to verify the generality of Bypass CA and scale up the pre-training data for further performance gains.

### 5. Acknowledgement

This work was supported by IITP grant funded by the Korea government (MSIT) (No. 2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network (80%), and No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training (20%).)
References


