Semantic Role Aware Correlation Transformer for Text to Video Retrieval

Burak Satar1,2, Zhu Hongyuan1,4, Xavier Bresson3, Joo Hwee Lim1,2

Overview

Introduction. Retrieving related video on a textual query gets harder since the number of videos on the internet increases. Most works use one joint embedding space for text-to-video retrieval task without fully exploiting cross-modal features. We propose a hierarchical model representing complex textual and visual features with three joint embedding spaces by utilizing self-attention and cross-modal attention to exploit the modality-specific and modality-complement visual embeddings. Preliminary results show that our approach surpasses a current state-of-the-art method, with a high margin in all metrics. It also overpasses two SOTA methods in terms of two metrics.

Related Work. Conventional models are based on keywords query, which is insufficient to retrieve fine-grained and compositional events [1].

Most works [4] embeds whole videos and texts into flat vectors to exploit global features, while the others focus only on local features. Recently, Chen et al. [3] propose decomposing text into three semantic roles (events, actions and entities) and then embedding 2D video features into these three spaces accordingly for matching. Another line of research uses a BERT-like transformer [5] to learn the text-video correspondence, based on recent mixture-of-expert embedding [6], which requires a large-scale dataset for pre-training.

Experiments. We evaluate our model on YouCook2 [2], which is a video dataset on cooking gathered from YouTube. The task is retrieving video clips based on text queries. R@1, R@5, R@10 and median rank are used as evaluation metrics.

Method

We propose a novel transformer architecture for video-text matching inspired by [3] and [5]. Different from [3], which only considers multi-head embedding of the spatial frame and ignores the interaction between different visual contexts, our method explicitly considers more fine-grained visual encoding of object, spatial and temporal contexts by embedding RoI regions, 2D frames and video sequences into the corresponding space with their interactions. Different from [5], which only uses self-attention to discover modality-specific information, our method uses a self-attention scheme to discover modality-specific discriminative features. Also, our model utilizes cross-modal attention to consider the interactions between object, spatial and temporal contexts to discover modality-complement features for better align video and text.

Visual & Textual Encoding

Textual encoding. We follow Chen et al. [3] to disentangle text embeddings. First, semantic role labelling, then GCN is applied. Visual encoding. Our aim is to calculate final embeddings for each level. This formula only shows our implementation on the spatial level.

Results

Table 1. Text-to-video retrieval comparison with SOTA approaches on YouCook2 validation set. Our method surpasses the SOTA methods in the first two parameters when pre-training.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-training</th>
<th>Visual Backbone</th>
<th>Batch Size</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>MedR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>No</td>
<td>-</td>
<td>-</td>
<td>0.03</td>
<td>0.15</td>
<td>0.3</td>
<td>1675</td>
</tr>
<tr>
<td>Miech et al</td>
<td>No</td>
<td>ResNeXt-101</td>
<td>-</td>
<td>4.2</td>
<td>13.7</td>
<td>21.5</td>
<td>65</td>
</tr>
<tr>
<td>HGLMM [5]</td>
<td>No</td>
<td>ResNeXt-101</td>
<td>32</td>
<td>4.6</td>
<td>14.3</td>
<td>21.6</td>
<td>75</td>
</tr>
<tr>
<td>HGR [5]</td>
<td>No</td>
<td>ResNeXt-101</td>
<td>32</td>
<td>4.7</td>
<td>14.1</td>
<td>20.0</td>
<td>87</td>
</tr>
<tr>
<td>Ours</td>
<td>No</td>
<td>ResNeXt-101</td>
<td>32</td>
<td>5.3</td>
<td>14.5</td>
<td>20.8</td>
<td>77</td>
</tr>
<tr>
<td>Miech et al+FT</td>
<td>No</td>
<td>HowTo100M</td>
<td>-</td>
<td>8.2</td>
<td>24.5</td>
<td>35.3</td>
<td>24</td>
</tr>
<tr>
<td>ActBet [7]</td>
<td>No</td>
<td>HowTo100M</td>
<td>HowNet-3D</td>
<td>9.6</td>
<td>26.7</td>
<td>38.0</td>
<td>19</td>
</tr>
<tr>
<td>MMV FAC [8]</td>
<td>No</td>
<td>HowTo100M+AudioSet</td>
<td>TSM-50</td>
<td>4096</td>
<td>11.5</td>
<td>30.2</td>
<td>41.5</td>
</tr>
<tr>
<td>MIL-NE [9]</td>
<td>No</td>
<td>HowTo100M</td>
<td>83D</td>
<td>8192</td>
<td>15.1</td>
<td>38.0</td>
<td>51.2</td>
</tr>
</tbody>
</table>

Table 2. Ablation studies to investigate the contributions of various feature experts at different levels. This confirms our insight that inter-modal correlation can be exploited with our proposed cross-modal attention mechanism to achieve better results.

Conclusion. Our model surpasses a strong baseline with a high margin, and it also overpasses other SOTA methods in R@1, R@5 metrics. We think that modality-specific and modality-complement features improve accuracy at R@1 and R@5, which are more demanding and useful for real-world applications.

Authors

1 Institute for Infocomm Research, A*STAR, Singapore
2 School of Computer Science and Engineering, NTU, Singapore
3 Department of Computer Science, National University of Singapore
4 Institute of Infocomm Research, A*STAR, Singapore

Acknowledgements

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project A18A260046). Xavier Bresson is supported by the NRF Fellowship NRF2017-10.

References