Learning Fine-Grained Visual Understanding for Video Question Answering via Decoupling Spatial-Temporal Modeling

Hsin-Ying Lee     Hung-Ting Su     Bing-Chen Tsai     Tsung-Han Wu     Jia-Fong Yeh     Winston H. Hsu

Decoupling spatial-temporal modeling into image and video-language models and pretraining to learn temporal relations between events in videos help video question answering.

Introduction
- State-of-the-art approaches to video question answering mostly perform coarse-grained spatial-temporal modeling.
- Image-language (IL) models encode regions and grids, showing great potential for encoding fine-grained spatial semantics for video question answering.
- To answer questions about temporal relations, video-language (VL) models have to recognize events in videos and associate events with time conjunctions in questions.

Method
- We propose Decoupled Spatial-Temporal Encoders (DeST), decoupling spatial-temporal modeling into IL and VL encoders.
- We incorporate a pre-trained IL encoder to encode static spatial semantics by averaging sparsely sampled frame-by-frame predictions at high spatial resolution.
- For questions requiring temporal relations, we train a VL encoder to model temporal dynamics, operating at high temporal but low spatial resolution.
- The VL encoder is pre-trained with our proposed objective, Temporal Referring Modeling (TRM).
- TRM queries absolute and relative positions of events in videos synthesized by concatenating clips sampled from video captioning datasets.

Experiments
- We conduct preliminary analyses to estimate the spatial and temporal modeling capability of previous work.
- Some VL models perform better than VL models on questions about spatial understanding.
- Some IL models perform similarly when taking normal and shuffled videos on questions about temporal understanding.
- DeST outperforms the previous state-of-the-art on two video QA benchmarks, ActivityNet-QA and AGQA. The ablation studies also demonstrate the efficacy of the proposed pipeline DeST and pre-training objective TRM.

Decoupled Spatial-Temporal Encoders

Temporal Referring Modeling

Table: Comparison with previous methods on ActivityNet-QA.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-train</th>
<th>Baseline</th>
<th>ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQA2</td>
<td>56.95</td>
<td>57.64</td>
<td>58.16</td>
</tr>
<tr>
<td>VQA2 (with TRM)</td>
<td>57.35</td>
<td>57.64</td>
<td>58.83</td>
</tr>
</tbody>
</table>

A figure illustrating the decoupling of spatial and temporal modeling in video-language models.