HowToVQA69M: large-scale VideoQA training dataset

- Generated from HowTo100M
- 69M video-question-answer triplets

**Motivation**
- Manual annotation of visual data is expensive
- Text-only annotations are easier to obtain

**Goal**
- Tackle Video Question Answering (VideoQA) without any manual supervision of visual data

**Idea**
- Automatically generate VideoQA training data from narrated videos
- Rely on cross-modal supervision and language models trained on text-only annotations

**Generating VideoQA data**

**Text-only supervision**
- **Assumption**: weak correlation between video and speech
- **Punctuator** $p$ is trained on a punctuated corpus
- **Transformers** $T_v$ and $T_s$ are trained on question-answers

**Generation procedure**

- **Input**: video with raw speech $x$
- **Output**: $(v, q, a)$ triplet
  1. Punctuation: extract speech sentence $p(s)$
  2. Video extraction: extract clip $v$ temporally aligned with $p(s)$
  3. Answer extraction: extract answer $a = T_v(p(s))$
  4. Question generation: generate question $q = T_s(a, p(s))$

**HowToVQA69M**

- 10,000 videos from HowTo100M
- Manually annotated
- 10,000 open-ended questions
- 5 correct answers per question for a detailed evaluation
- Exclusion of non-visual questions to reduce language bias

**IVQA: new dataset for VideoQA evaluation**

- Question: What shape is the handicraft item in the end?
  - 2 annotators: spiral
  - 1 annotator: heart

**Zero-shot VideoQA**

**Definition**
- No manual supervision of visual data

**Quantitative results**
- Our model trained on HowToVQA69M (ii) outperforms its language-only variant (i) and its variant trained on HowTo100M (iii)

**Qualitative results**

**Results after finetuning**
- Our model pretrained on HowToVQA69M (iii) improves over its variant trained from scratch (i) and its variant pretrained on HowTo100M (ii)

**State-of-the-art results on 4 existing VideoQA datasets**

**Table**

<table>
<thead>
<tr>
<th>Method</th>
<th>IVQA</th>
<th>MSRv7-TQA</th>
<th>MSVD-QA</th>
<th>ActivityNet-QA</th>
<th>How2QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretraining data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HowTo100M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HowToVQA69M (i)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HowToVQA69M (ii)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HowToVQA69M (iii)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HowTo100M + TV</td>
<td>38.2</td>
<td>39.4</td>
<td>37.6</td>
<td>36.1</td>
<td>38.1</td>
</tr>
<tr>
<td>How2QA</td>
<td>26.2</td>
<td>27.5</td>
<td>26.2</td>
<td>25.9</td>
<td>27.5</td>
</tr>
</tbody>
</table>

**Table**

<table>
<thead>
<tr>
<th>Question</th>
<th>GT Answer</th>
<th>Other answers</th>
<th>Model Answer</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the man's hand?</td>
<td>shovel (3)</td>
<td>spa (2)</td>
<td>shovel (2)</td>
<td>Contrastive loss</td>
</tr>
<tr>
<td>What fruit is shown in the man's hand?</td>
<td>watermelon (5)</td>
<td>-</td>
<td>watermelon (5)</td>
<td>Contrastive loss</td>
</tr>
<tr>
<td>What animal did I put up pictures of him with?</td>
<td>monkey (4)</td>
<td>-</td>
<td>monkey (4)</td>
<td>Contrastive loss</td>
</tr>
</tbody>
</table>