Just Ask: Learning to Answer Questions from Millions of Narrared Videos

Project page: https://antoyang.github.io/just-ask.html
Computer Vision - Video Understanding

• **Goal:** automating tasks that the human visual system can do

• **Motivation:** video data is plentiful

• **Applications:**
  - Video search engines
  - Video surveillance
  - Self-driving cars
  - Describing videos for visually impaired people
  - Generation of video content for entertainment

• **Challenge:** How to evaluate video understanding?
Video QA is a promising proxy task to evaluate video understanding.

**Open-Ended Question:** Where are the men?
**Answer:** Track

**Multiple-Choice Question:** What are the lined up men doing?
**Proposal 1:** Running
**Proposal 2:** Talking
**Proposal 3:** Shaving
VideoQA Challenges

• VideoQA is a difficult task because of the diversity of questions that one may ask, requiring the ability to recognize actions, objects, colors at different spatio-temporal granularities

• Learning is currently the only known approach to handle variability in the data, but it requires lots of training data, and obtaining manually annotated VideoQA data is expensive and not scalable

**Question:** How many times does the cat lick?
**Answer:** 7 times

**Question:** What does the cat do 3 times?
**Answer:** put head down

**Question:** What is the color of the bulldog?
**Answer:** brown
・We automatically generate large-scale VideoQA data from narrated videos, relying on language models trained on text-only annotations

・We show how VideoQA models can benefit from such data, by tackling VideoQA without any manual supervision of visual data (zero-shot) or by finetuning our pretrained model
Weak supervision

- Narrated videos contain speech, therefore paired (video, speech) data is easy to obtain and abundant.
- The weak correlation between the visual content and speech in narrated videos helped improve on other tasks [Miech 2019]
Text-only supervision for automatic generation of VideoQA data

To generate VideoQA data, we rely on cross-modal supervision and language models [Raffel 2020] trained on text-only annotations.
Generating video-question-answer triplets

Raw narration \( S \)

"to dry before you stick him on a kick I"

"put up some pictures of him with another"

"monkey as well so you can make many"

"as you like thank you for watching"

Extracted sentence \( p(s) \)

"I put up some pictures of him with another monkey."

Sentences extractor \( T_s \)

Answer extractor \( T_a \)

Question generator \( T_q \)

Extracted answer \( a \)

"What animal did I put up pictures of him with?"

Generated question \( q \)

Outputs

Sentence-aligned video \( V \)

"Thank you for watching"
HowToVQA69M: a large-scale VideoQA training dataset

We apply our generation pipeline to the videos from HowTo100M [Miech 2019] and obtain HowToVQA69M, a large-scale and noisy VideoQA dataset.

Speech: So you bring it to a point and we'll, just cut it off at the bottom.
Generated question: What do we do at the bottom?
Generated answer: cut it off

Speech: Do it on the other side, and you've peeled your orange.
Generated question: What color did you peel on the other side?
Generated answer: orange

Speech: You can’t miss this…
Generated question: What can’t you do?
Generated answer: miss

Wrong QA Generation

Weak video-speech correlation
VideoQA model and training procedure on HowToVQA69M

**Video:**

**Question:** Where are the men?

**Answer:** Track
iVQA: a new VideoQA evaluation benchmark

• We manually collected an open-ended VideoQA dataset based on HowTo100M narrated videos
• It contains 10K videos, each annotated with 1 question and 5 corresponding correct answers

**Question:** What shape is the handcraft item in the end?  

**Answers:**  
- shell ✔️ 2 annotators  
- spiral ✔️ 2 annotators  
- heart ✔️ 1 annotator
Zero-shot VideoQA with no manual supervision of visual data

We evaluate our VideoQA model VQA-T pretrained on HowToVQA69M with the following baselines:
• QA-T pretrained on HowToVQA69M: language-only variant, not using the visual modality
• VQA-T pretrained on HowTo100M: common pretraining approach for multi-modal transformers

Quantitative results on 5 VideoQA datasets:

<table>
<thead>
<tr>
<th>Method</th>
<th>Pretraining Data</th>
<th>iVQA</th>
<th>MSRVTT-QA</th>
<th>MSVD-QA</th>
<th>ActivityNet-QA</th>
<th>How2QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>Ø</td>
<td>0.09</td>
<td>0.02</td>
<td>0.05</td>
<td>0.05</td>
<td>25.0</td>
</tr>
<tr>
<td>QA-T</td>
<td>HowToVQA69M</td>
<td>4.4</td>
<td>2.5</td>
<td>4.8</td>
<td>11.6</td>
<td>38.4</td>
</tr>
<tr>
<td>VQA-T</td>
<td>HowTo100M</td>
<td>1.9</td>
<td>0.3</td>
<td>1.4</td>
<td>0.3</td>
<td>46.2</td>
</tr>
<tr>
<td>VQA-T</td>
<td>HowToVQA69M</td>
<td>12.2</td>
<td>2.9</td>
<td>7.5</td>
<td>12.9</td>
<td>51.1</td>
</tr>
</tbody>
</table>
Zero-shot VideoQA with no manual supervision of visual data

Qualitative examples on iVQA:

**Question:** What is the man cutting?
**GT answer:** pipe
**QA-T (HowToVQA69M):** onion
**VQA-T (HowTo100M):** knife holder
**Ours:** pipe

**Question:** What is the largest object at the right of the man?
**GT answer:** wheelbarrow
**QA-T (HowToVQA69M):** statue
**VQA-T (HowTo100M):** trowel
**Ours:** wheelbarrow

**Question:** What fruit is shown in the end?
**GT answer:** watermelon
**QA-T (HowToVQA69M):** pineapple
**VQA-T (HowTo100M):** slotted spoon
**Ours:** watermelon
Zero-shot VideoQA: failure cases

Qualitative examples on iVQA:

**Question:** What are standing up behind the man on his right?
**GT answer:** guitars
**Ours:** strings

**Question:** In what room does the video take place?
**GT answer:** kitchen
**Ours:** dining room
# Benefits of HowToVQA69M pretraining

Comparison with state-of-the-art on 4 VideoQA datasets:

<table>
<thead>
<tr>
<th>Method</th>
<th>Pretraining Data</th>
<th>MSRVTT-QA</th>
<th>MSVD-QA</th>
<th>ActivityNet-QA</th>
<th>How2QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCRN [Le 2020]</td>
<td>∅</td>
<td>35.6</td>
<td>36.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SSML [Amrani 2020]</td>
<td>HowTo100M</td>
<td>35.1</td>
<td>35.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HERO [Li 2020]</td>
<td>HowTo100M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>74.1</td>
</tr>
<tr>
<td>ClipBERT [Lei 2021]</td>
<td>COCO + VG</td>
<td>37.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CoMVT [Seo 2021]</td>
<td>HowTo100M</td>
<td>39.5</td>
<td>42.6</td>
<td>38.8</td>
<td>82.3</td>
</tr>
<tr>
<td>Ours (∅)</td>
<td>∅</td>
<td>39.6</td>
<td>41.2</td>
<td>36.8</td>
<td>80.8</td>
</tr>
<tr>
<td>Ours</td>
<td>HowToVQA69M</td>
<td>41.5</td>
<td>46.3</td>
<td>38.9</td>
<td>84.4</td>
</tr>
</tbody>
</table>
Results for rare answers

Results on subsets of iVQA corresponding to four quartiles with Q1 and Q4 corresponding to samples with most frequent and least frequent answers:

<table>
<thead>
<tr>
<th>Pretraining Data</th>
<th>Finetuning</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>∅</td>
<td>✓</td>
<td>38.4</td>
<td>16.7</td>
<td>5.9</td>
<td>2.6</td>
</tr>
<tr>
<td>HowTo100M</td>
<td>✓</td>
<td>46.7</td>
<td>22.0</td>
<td>8.6</td>
<td>3.6</td>
</tr>
<tr>
<td>HowToVQA69M</td>
<td>✗</td>
<td>9.0</td>
<td>8.0</td>
<td>9.5</td>
<td>7.7</td>
</tr>
<tr>
<td>HowToVQA69M</td>
<td>✓</td>
<td>47.9</td>
<td>28.1</td>
<td>15.6</td>
<td>8.5</td>
</tr>
</tbody>
</table>

=> VideoQA specific pretraining on additional large-scale, diverse data helps improve generalization
Open research directions

• Reduce the domain gap between the question-answer generator trained on text-only data (SQuADv1 in our case) and the text data used for VideoQA generation (speech in our case)

• Automatic cleaning of generated data

• Generalization of VideoQA models to other tasks

• Creation of VideoQA datasets that are closer to potential applications

• End-to-end learning of VideoQA models
Conclusion

• We automatically generate a large-scale VideoQA dataset, HowToVQA69M, using text-only supervision and videos with readily-available narration

• We manually collect iVQA, a new VideoQA benchmark with redundant annotations and reduced language bias

• We show that our VideoQA model highly benefits from training on HowToVQA69M in a new zero-shot VideoQA setting; additionally, after finetuning, our model improves the state-of-the-art on 4 VideoQA datasets