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

Antoine Yang[1,2], Antoine Miech[1,2,+], Josef Sivic[3], Ivan Laptev[1,2], Cordelia Schmid[1,2]

[1]Inria Paris [2]Ecole Normale Supérieure, PSL [3]CIIRC, CTU in Prague [+]Now at DeepMind

2021 ICCV OCTOBER 11-17 VIRTUAL

OVERVIEW

Motivation

➤Manual annotation for Video Question Answering is expensive

➤Text-only annotations are easier to obtain

Goal

➤Tackle Video Question Answering (VideoQA) without using any manual supervision of visual data

Idea

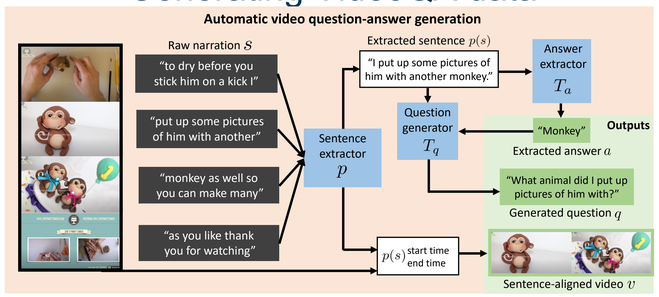
➤Automatically generate VideoQA training data from narrated videos

➤Rely on cross-modal supervision and language models trained on text-only annotations

Speech: Fold them in half again, to make a triangle.

Generated Question: How do you make a triangle?
Generated Answer: fold them in half again

Generating VideoQA data



Text-only supervision

➤**Assumption:** weak correlation between video and speech

➤Punctuator p is trained on a punctuated corpus

➤Transformers $\mathrm{T_a}$ and $\mathrm{T_q}$ are trained on question-answers

Generation procedure

Input: video with raw speech s **Output:** $(\mathrm{v}, \mathrm{q}, \mathrm{a})$ triplet

1. Punctuation: extract speech sentence $\mathrm{p(s)}$
2. Video extraction: extract clip v temporally aligned with $\mathrm{p(s)}$
3. Answer extraction: extract answer $\mathrm{a} = \mathrm{Ta}(\mathrm{p(s)})$
4. Question generation: generate question $\mathrm{q} = \mathrm{Tq}(\mathrm{a},\ \mathrm{p(s)})$

HowToVQA69M: large-scale VideoQA training dataset

➤Generated from HowTo100M

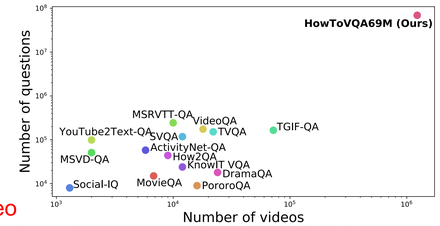
➤69M video-question-answer triplets

➤Noisy:

≃30% correct samples

≃31% question-answer generation errors

≃39% question-answers unrelated to the video



Speech: Add some of your favorite sprinkles give it a mix.
Question: What can you add to the mix?
Answer: sprinkles

Speech: ...I'm going to show you how to unlock your ipod touch.
Question: What will I show you?
Answer: how to unlock your ipod touch

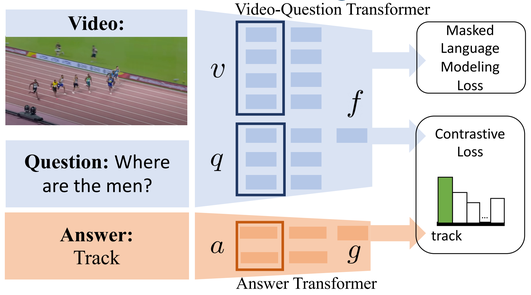
Speech: …do it on the other side, and you've peeled your orange
Question: What color did you peel on the other side?
Answer: orange

Question-answer generation error

Speech: ...I've had over a hundred emails.
Question: How many emails have I had?
Answer: over a hundred

Question-answer unrelated to the video

Training on HowToVQA69M



Training with a large-scale vocabulary of answers

➤Contrastive loss between a video-question transformer and an answer transformer

➤For each triplet $(\mathrm{v_i}, \mathrm{q_i}, \mathrm{a_i})$, we construct negative triplets $(\mathrm{v_i}, \mathrm{q_i}, \mathrm{a_j})$, by sampling other answers $\mathrm{a_j} \neq \mathrm{a_i}$

iVQA: new dataset for VideoQA evaluation

➤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

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

Answers
- shell ✓ 2 annotators
- spiral ✓ 2 annotators
- heart ✓ 1 annotator

Zero-shot VideoQA

Definition

➤No manual supervision of visual data

Quantitative results

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

Method	Pretraining data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
Random	∅	0.09	0.02	0.05	0.05	25.0
QA-T (i)	HowToVQA69M	4.4	2.5	4.8	11.6	38.4
VQA-T (ii)	HowTo100M	1.9	0.3	1.4	0.3	46.2
VQA-T (iii)	HowToVQA69M	**12.2**	**2.9**	**7.5**	**12.2**	**51.1**

Qualitative results

Question: What design are they making?
GT Answer: rose (4), rose flower (1)
QA-T (i): pinwheel
VQA-T (ii): piping bag
VQA-T (iii): rose

Question: What is in the man's hand?
GT Answer: shovel (3), spade (2)
QA-T (i): coin
VQA-T (ii): planting
VQA-T (iii): shovel

Question: What fruit is shown at the end?
GT Answer: watermelon (5)
QA-T (i): pineapple
VQA-T (ii): slotted spoon
VQA-T (iii): watermelon

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

Method	Pretraining data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
HCRN [1]	∅	-	35.6	36.1	-	-
SSML [2]	HowTo100M	-	35.1	35.1	-	-
ClipBERT [3]	COCO + VG	-	37.4	-	-	-
HERO [4]	HowTo100M + TV	-	-	-	-	74.1
CoMVT [5]	HowTo100M	-	39.5	42.6	38.8	82.3
Ours (i)	∅	23.0	39.6	41.2	36.8	80.8
Ours (ii)	HowTo100M	28.1	40.4	43.5	38.1	81.9
Ours (iii)	HowToVQA69M	**35.4**	**41.5**	**46.3**	**38.9**	**84.4**

[1] TM. Le, et. al., Hierarchical conditional relation networks for video question answering. In CVPR, 2020.
[2] E. Amrani, et. al., Noise estimation using density estimation for self-supervised multimodal learning. In AAAI, 2021.
[3] J. Lei, et. al., Less is more: Clipbert for video-and-language learning via sparse sampling. In CVPR, 2021.
[4] L. Li, et. al., HERO: Hierarchical encoder for video+language omni-representation pre-training. In EMNLP, 2020.
[5] PH. Seo, et. al., Look before you speak: Visually contextualized utterances. In CVPR, 2021.