Zero-Shot Video Question Answering via Frozen Bidirectional Language Models

Antoine Yang[1,2], Antoine Miech[3], Josef Sivic[4], Ivan Laptev[1,2], Cordelia Schmid[1,2]

[1]Inria Paris [2]Ecole Normale Supérieure, PSL [3]DeepMind [4]CIIRC, CTU in Prague

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

Zero-Shot VideoQA

➤ Answer questions about videos without training on manual annotation of visual data.

Motivation

➤ A strong approach relies on frozen autoregressive language models [1] but requires large models.

➤ Recent work in natural language [2] shows strong zero-shot results with lighter bidirectional masked language models (BiLM).

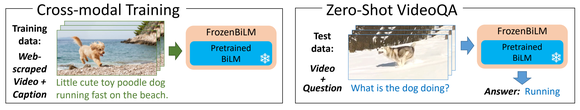
Contributions

➤ **FrozenBiLM:** a framework that handles multi-modal inputs using frozen BiLM and enables zero-shot VideoQA through masked language modeling.

➤ Extensive ablation studies and improvements over previous autoregressive models.

➤ SoTA on 8 zero-shot benchmarks, competitive performance on fully-supervised benchmarks and promising few-shot results.

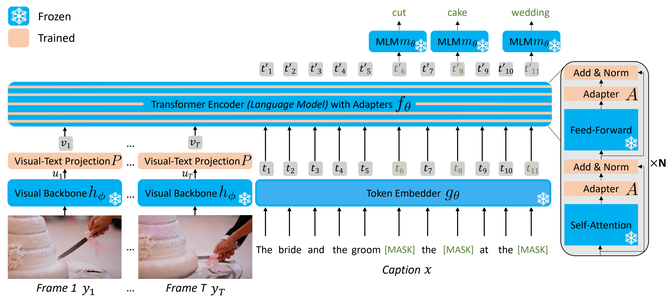
➤ **Code and models:**
https://github.com/antoyang/FrozenBiLM



References

[1] M. Tsimpoukelli et al., Multimodal few-shot learning with frozen language models. In NeurIPS 2021.
[2] T. Shick, et. al., It's not just size that matters: small language models are also few-shot learners. In NAACL 2021.
[3] A. Radford, et. al., Learning transferable visual models from natural language supervision. In arXiv, 2021.
[4] A. Yang, et. al., Learning to answer visual questions from Web videos. In TPAMI, 2022..
[5] R. Zellers, et. al., Merlot reserve: neural script knowledge through vision, language and sound. In CVPR, 2022.
[6] TJ. Fu et al., Violet: end-to-end video-language transformers with masked visual token modeling. In arXiv 2021.
[7] AJ. Wang et al, All in one: exploring unified video-language pretraining. In arXiv 2022.
[8] R. Zellers et al. Merlot: multi-modal neural script knowledge models. In NeurIPS 2021.

Architecture



➤ **Frozen parameters:** pretrained BiLM and visual backbone.

➤ **Trained parameters:** visual-to-text projection, adapters, layer norm.

Cross-modal Training

➤ **Loss:** visually-conditioned masked language modeling.

➤ **Data:** videos with alt-text description from WebVid10M.

Downstream Task Adaptation

➤ **Input prompts:**

Open-ended VideoQA: "`[CLS] Question: <Question>? Answer: [MASK]. Subtitles: <Subtitles> [SEP]`"

Multiple-choice VideoQA: "`[CLS] Question: <Question>? Is it '<Answer Candidate>'? [MASK]. Subtitles: <Subtitles> [SEP]`"

Video-conditioned fill-in-the-blank: "`[CLS] <Sentence with a [MASK] token>. Subtitles: <Subtitles> [SEP]`"

➤ **Answer embedding module:** maps a masked token to an answer using the frozen masked language modeling head.

➤ **Fully-supervised finetuning:** standard cross-entropy loss while keeping the same weights frozen.

BiLM vs Autoregressive LM

Method	Params	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA
Autoregressive (GPT-J)	6B	21.4	9.6	26.7	24.5	37.3
Bidirectional (BERT-Base)	**110M**	12.4	6.4	11.7	16.7	23.1
Bidirectional (DeBERTa-V2-XLarge)	890M	**27.3**	**16.8**	**32.2**	**24.7**	**41.0**

Comparison to SoTA

➤ Zero-shot results.

Method	LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
SoTA	31.0 [5]	13.3 [4]	5.8 [5]	13.5 [4]	12.3 [4]	3.6 [3]	53.1 [4]	26.1 [3]
FrozenBiLM	**51.5**	**26.8**	**16.7**	**33.8**	**25.9**	**41.9**	**58.4**	**59.7**


Question: What is the man holding at the start of the video?
GT Answer: guitar, electric guitar
Just Ask: typewriter
UnFrozenBiLM: beer
FrozenBiLM (text-only): scissors
FrozenBiLM (ours): guitar


Question: What item hanging on the wall features a tree?
GT Answer: quilt
Just Ask: christmas sock
UnFrozenBiLM: fabric
FrozenBiLM (text-only): tree
FrozenBiLM (ours): quilt


Question: What is the sitting man doing?
GT Answer: knit sweater
Just Ask: tie cow
UnFrozenBiLM: swimming
FrozenBiLM (text-only): eating
FrozenBiLM (ours): knit sweater


Question: Where is the woman sitting on?
GT Answer: camel
Just Ask: horse yard
UnFrozenBiLM: desert
FrozenBiLM (text-only): chair
FrozenBiLM (ours): camel


Question: What is the color of the cabinet door in the video?
GT Answer: red
Just Ask: dresser
UnFrozenBiLM: blue
FrozenBiLM (text-only): black
FrozenBiLM (ours): red

➤ Fully-supervised benchmarks. Note that FrozenBiLM only trains 30M parameters.

Method	LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
SoTA	53.7 [6]	35.4 [4]	46.8 [7]	48.3 [7]	41.4 [8]	**69.5** [8]	85.3 [4]	**86.1** [5]
FrozenBiLM	**63.5**	**39.6**	**47.0**	**54.8**	**43.2**	68.6	**86.7**	82.0

Few-Shot results

➤ New setting using a small fraction of the downstream dataset for finetuning.

Supervision	LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
0% (zero-shot)	51.5	26.8	16.7	33.8	25.9	41.0	58.4	59.7
1% (few-shot)	56.9	31.1	36.0	46.5	33.2	55.1	71.7	72.5
10% (few-shot)	59.9	35.3	41.7	51.0	37.4	61.2	75.8	77.6
100% (fully-supervised)	**63.5**	**39.6**	**47.0**	**54.8**	**43.2**	**68.6**	**86.7**	**82.0**