Video Question Answering with Iterative Video-Text Co-Tokenization Supplementary material

AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S. Ryoo, and Anelia Angelova

Google Research

1 Additional experiments

Table 1 shows the results on the TGIF-QA dataset [3]. It is an interesting result, where our method, using a medium size pretrained text model T5 (T5-Base) [8], is able to accomplish close to 100% accuracy on the multiple-choice questions which are for the 'Actions' and 'Transition' types of questions. This is due to the fact that the limited selection of answers is easy to guess even without video[1]. The questions from the 'FrameQA' type are not multiple choice and thus are harder to guess. We acknowledge that this contemporary text model is stronger than the previous text models, but is important to note that these two tasks likely had been mostly language understanding ones. We also note that the datasets evaluated in the main paper, MSRVTT-QA, MSVD-QA, IVQA are much more challenging and do not suffer from this problem. We also compare the results of using our approach with the open vocabulary output (last row of Table 1). Note that when using the open vocabulary text-generative setting on TGIF-QA, the tasks are significantly harder, even when using both video and text, with much accuracy lower numbers showing the difficulty of this setting.

2 Implementation and experimentation details

Model Training. For pretraining, we train the model with a batch size of 256 for 500,000 steps. The learning rate was set to 0.001 When finetuning, the batch size was 128 and trained for 10,000 steps, with a learning rate of 1×10^{-6}. We used the Adam optimizer, with weight decay set to 0.01.

2.1 Metrics

For the IVQA dataset, we use accuracy as defined in the paper:

$$Acc = \min\{\frac{\#\text{humans with ans}}{2}, 1\} \quad (1)$$

[1] The Text-only method performs randomly for the action counting category as it definitely needs the video input to correctly answer how many times an action is performed.

Table 1. TGIF-QA. We note that when using a contemporary pre-trained text model (T5-Base), our results achieve almost 100% on the multiple-choice questions, alluding the answers are easy to guess.

Model	Action	Transition	FrameQA
ST-VQA(R+C)	60.8	67.1	49.3
Co-Memory(R+F)	68.2	74.3	51.5
PSAC(R)	70.4	76.9	55.7
Heterogeneous Memory(HME)(R+C) [1]	73.9	77.8	53.8
Location Aware GCN [2]	74.3	81.1	56.3
HCRN [5]	75.0	81.4	55.9
Bridge2Answer [7]	75.9	82.6	57.5
QueST [4]	75.9	81.0	59.7
ClipBERT [6] 1x1 (Ntest=1)	82.9	87.5	59.4
ClipBERT [6] 1x1	82.8	87.8	60.3
Ours (Text-only, T5 pretrained [8])	98.4	97.3	62.5
Ours (Video+Text, Open Vocabulary)	11.8	12.5	27.3

averaged over 5 choose 4 of the ground truth (GT) answers.

For MSRVTT-QA and MSVD-QA, we use the standard accuracy metric. In the open-ended text generative setting, we check that the generated string is exactly equal to the ground truth string. The output string is the result of using beam search on the trained model. This is the hardest setting. With the restricted vocabulary, this check is easier, as there are fewer output options.

3 Additional visualizations

Figure 1 shows additional visualizations of our approach.

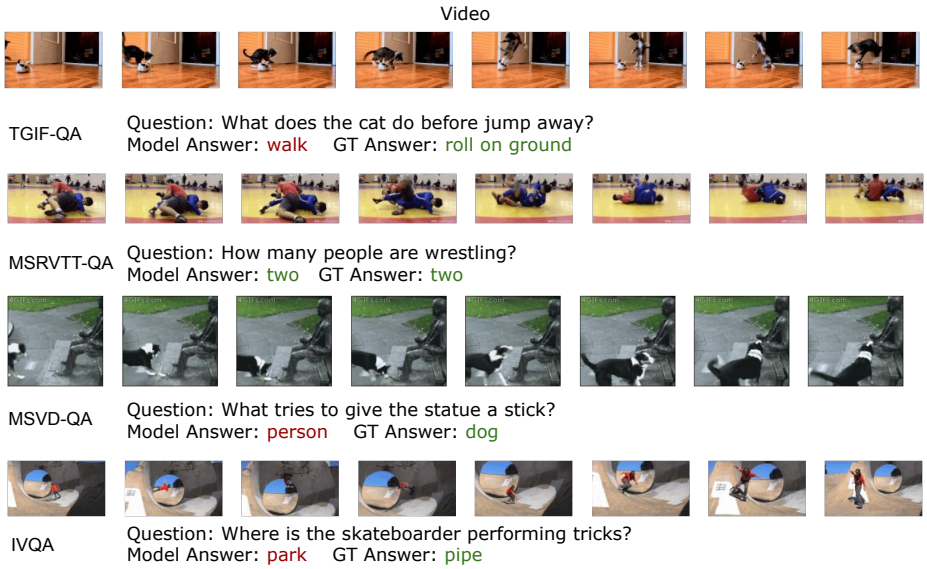

Fig. 1. Examples of our method. Examples are from TGIF-QA, MSRVTT, MSVD, and IVQA datasets, in this order.

References

1. Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., Huang, H.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: CVPR (2019)
2. Huang, D., Chen, P., Zeng, R., Du, Q., Tan, M., Gan, C.: Location-aware graph convolutional networks for video question answering. In: AAAI (2020)
3. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: CVPR (2017)
4. Jiang, J., Chen, Z., Lin, H., Zhao, X., Gao, Y.: Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In: AAAI (2020)
5. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. In: CVPR (2020)
6. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: CVPR (2021)
7. Park, J., Lee, J., Sohn, K.: Bridge to answer: Structure-aware graph interaction network for video question answering. In: CVPR (2021)
8. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. In: Journal of Machine Learning Research (2020)