Just Ask Plus: Using Transcripts for VideoQA

Mohammad Javad Pirhadi Motahhare Mirzaei
Sauleh Eetemadi
Iran University of Science and Technology
Tehran, Iran
{mohammad_pirhadi, m_mirzaei96}@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

Social-IQ 2.0 challenge is designed to benchmark recent AI technologies' skills to reason about social interactions, which is referred as Artificial Social Intelligence in the form of a VideoQA task. In this work, we use Just Ask and SpeechT5 models as feature extractors, and reason by adding one attention layer and two transformer encoders. Our best configuration reaches 53.35% accuracy on the validation set. The code is publicly available on GitHub.

1. Introduction

VideoQA (Video Question Answering) is a multi-modal task that aims to answer a question based on a video. In the Social-IQ 2.0 challenge (Wilf *et al*. [8]), given a video, a question and four possible answers, the correct must be chosen. The dataset contains over 1000 videos gathered from YouTube and over 6000 multiple-choice questions.

Yang *et al*. [9, 10] introduced a pretrained model using contrastive learning between a multi-modal video-question transformer and an answer transformer. Yang *et al*. report this model performs well in a zero-shot setting for VideoQA. This model has been pretrained using a large amount of videos including YouTube videos which is very close to the challenge dataset. Therefore we use the aforementioned model to solve the challenge in a zero-shot manner. In addition, for each video, we extract features from the provided question and suggested answers.

On the other hand, Ao *et al*. [1] propose SpeechT5 which is a unified-modal framework that explores the encoder-decoder pretraining for self-supervised speech/text representation learning. This model is pretrained on speech audio-text pairs which are very close to transcripts. We use this model to extract features from transcripts.

A multi-headed attention layer and a transformer encoder [7] are used to get the representation of the question with respect to the transcript. Next, another transformer encoder is used to get the representation of the question and answers with respect to each other. Finally, we calculate the similarity between the question and answers to pick the correct answer.

The official dataset consists of 6159, 943, and 1715 samples as training, validation, and test data respectively but we use a subset of it due to the unavailability of some videos at the time of download. The used dataset consists of 5618, 881, and 1577 samples as training, validation, and test data respectively (available on GitHub).

2. System Description

Overall we calculate two representations for the question (one with respect to the video and one with respect to the transcript) and one representation for each answer. These six vectors are fed into a transformer to produced six corresponding learned representations. The similarity between these representations are used to choose the correct answer. Figure 1 shows an overview of this model.

2.1. Feature Extraction

Just Ask model is composed of a frozen video encoder (S3D [11]), a question encoder, an answer encoder (both text encoders are DistillBERT [5]), and a unified-modal encoder for video-question joint encoding on top of the video encoder and question encoder. We extract and save 3 types of features using the Just Ask model: 1) Question representation 2) Answers representation and 3) Question representation after video-question joint encoding.

The SpeechT5 model has two parts: an encoder and a decoder. The encoder itself consists of 3 parts: speech encoder pre-net, text encoder pre-net, and joint speech/text encoder. We use text encoder pre-net and joint speech/text encoder to encode the transcripts. To this end we first add punctuation and capitalization to captions of each clip using Nvidia NeMo Toolkit [6], then split the sentences using NLTK [2], and finally encode each sentence of the transcript using SpeechT5. Note that the SpeechT5 model does not have a CLS token and the sentence representations are in the shape of $sentetnce\ length \times features\ count$. We chose

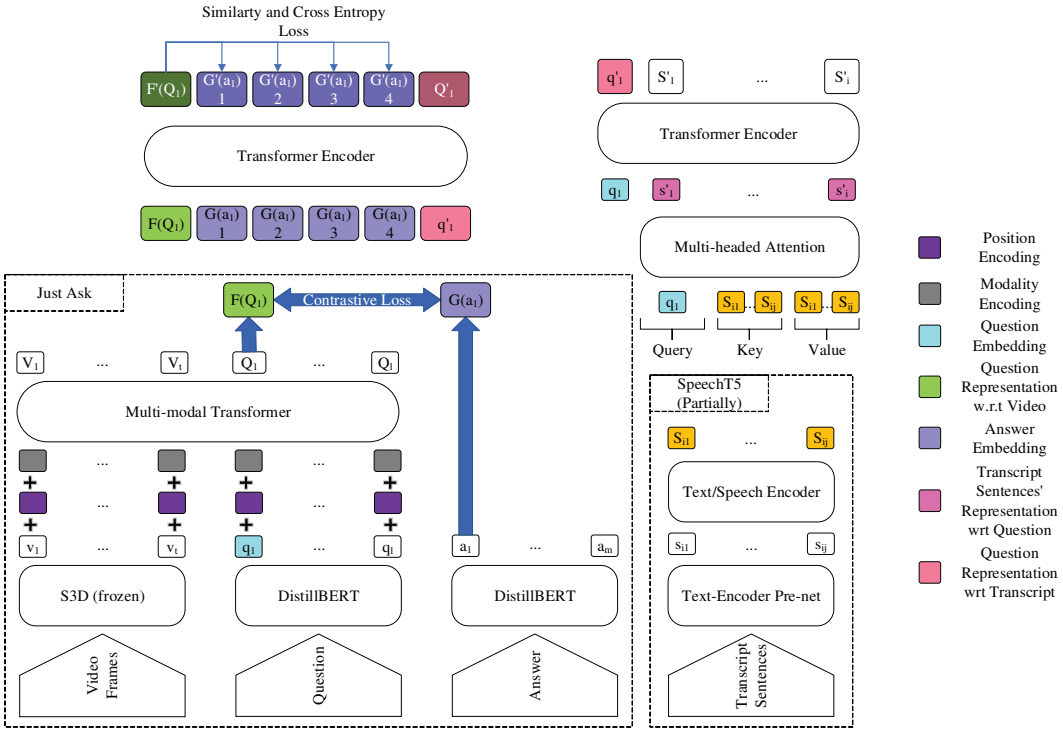

Figure 1. The overview of the proposed model. A multi-headed attention layer and a transformer encoder is used to get the representation of the question with respect to the transcript. Another transformer encoder calculates the representation of the question ($F(Q_1)$ and q'_1) and answers with respect to each other. To pick the correct answer, the similarities between the question ($F(Q_1)$) and the proposed answers are used.

SpeechT5 instead of other text encoders like RoBERTa [3] because it has been trained on speech text which is closer to transcripts, we explore using RoBERTa to prove this.

2.2. Zero-Shot setting

As part of the challenge we focus on using existing pretrained models without training or fine-tuning any additional or existing models. For this part of the challenge, we calculate question and suggested answer representations using the Just Ask pretrained model resulting in 5 vectors. We use the similarity between the question vector and the 4 suggested answer vectors to choose the correct answer.

2.3. Fusion and reasoning

The second focus of the challenge is on fusion and reasoning. For this part, we explore various methods of fusing vector representations of multi-modal data from pretrained models without any fine-tuning. For this goal, we first calculate a single representation for each sentence in the transcript. This is done using a multi-headed attention layer to extract the representation of the sentence from the representation of its tokens. The query is the question vector and key and value are the token vectors for each sentence. After this layer, we have a representation for each sentence in the transcript with respect to the question. Next, we jointly encode the question and transcript using a transformer encoder. Following [4] another transformer encoder encodes the two question vectors (with respect to video and transcript) and the suggested answer vectors together. This encoder helps the model to relatively compare the suggested answers and the question. Finally, we pick the most similar answer to the question as the correct answer. Note that we use the question vector with respect to the video for calculating the similarity because the questions are mostly about the visual content.

2.4. Transcript omission

Some clips do not have transcript. In order to have a robust model which can act correctly with and without the presence of a transcript, we skip using the transcript of a clip with some probability p which is a hyper-parameter. We examine the effect of different values for this probability. See experiments section for more details.

2.5. Loss Function

The correct answer is chosen using similarity between the question vector and the suggested answer vectors. Two loss functions are usually used in this context: cosine

Checkpoint	Accuracy
HowTo	26.67%
WebVid	30.53%
HowTo + WebVid	28.72%
HowTo + iVQA	27.13%
HowTo + MSRVTT-QA	30.08%
HowTo + MSVD-QA	26.45%
HowTo + ActivityNet-QA	32.80%
HowTo + How2QA	% 36.78
HowTo + WebVid + MSRVTT-QA	30.76%
HowTo + WebVid + MSVD-QA	31.44%
HowTo + WebVid + ActivityNet-QA	35.30%
HowTo + WebVid + How2QA	**39.39%**

Table 1. The Just Ask model accuracy on the validation set in a zero-shot manner. HowTo stands for HowToVQA69M and WebVid stands for WebVidVQA3M.

p	With transcripts	Without transcripts
0.00	**53.35%**	52.89%
0.33	52.67%	51.53%
0.66	52.89%	52.55%
1.00	50.51%	51.65%

Table 2. The Just Ask model accuracy on the validation set with and without transcript for the different probabilities of skipping the transcript.

embedding loss and cross-entropy loss. We chose cross-entropy loss due to time and resource constraints and to avoid tuning an extra hyper-parameter for cosine embedding loss.

3. Experiments

We used Google Colab infrastructure for all our training and experimentation. Since the test set is not released, all evaluations are reported on the validation set.

3.1. Zero-Shot

The Just Ask model has multiple checkpoints, to choose the best one we evaluate them on the validation set (Table 1). The HowToVQA69M and WebVidVQA3M datasets are used in the pretraining stage and several other datasets for fine-tuning. As expected, the best accuracy is for the HowToVQA69M + WebVidVQA3M + How2QA checkpoint which has the most data during the pretraining stage and the most similar dataset (relative to the challenge dataset) during the fine-tuning stage. We also use this checkpoint to extract features for the fusion and reasoning part of our work.

p	With transcripts	Without transcripts
0.00	52.21%	52.55%
0.33	52.89%	52.67%
0.66	52.21%	52.10%
1.00	52.78%	**53.01%**

Table 3. The Just Ask model accuracy on the validation set using RoBERTa as transcript feature extractor for the different probabilities of skipping the transcript.

3.2. Transcript omission

We evaluate the model on the validation set with and without transcripts for $p \in \{0, 33, 66, 100\}$. You can see the results in Table 2. Note that with $p = 0$, there are still some videos that do not have transcripts.
The best result with and without transcripts is for using all available transcripts which means not only the usage of the transcripts are effective, but also they act like a regularization as well.

3.3. RoBERTa instead of SpeechT5

In this work, we use a method to get the representation of transcripts' sentences which is not common. We use a multi-headed attention layer on tokens' features to get the whole sentence representation. To prove that this method works well, we replace the SpeechT5 model with RoBERTa base and omit the attention layer to see what happens. You can see the results in Table 3.
The second method (RoBERTAa) only acts better when we do not have any transcripts. This means that not only does the proposed method work, but also the pretraining stage of SpeechT5 improves it over RoBERTa base.
Another interesting point is that using RoBERTa for feature extraction can not help the model to improve, so choosing the right feature extractor matters.

4. Future work

By not considering audio information in the VideoQA task, the proposed model lacks a significant modality that is often crucial for understanding social interactions, emotions, and context in videos but it can be added just like how we add transcript information. Also, the results can be improved by automatic data generation using a method similar to the Just Ask paper. This data can be used to pretrain our proposed model and fine-tune on the challenge dataset to further improve the results. Further more, using a single transformer encoder to encode the video, question, and transcript together can be helpful.

5. Conclusion

In this work, we proposed a model to use the transcripts of the videos to solve the Social-IQ 2.0 Challenge in addition to the video (Just Ask paper only uses the video). The results show that using the transcripts can help the model when we choose the right feature extractor. Our best configuration reaches 53.35% accuracy on the validation set.

References

[1] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. 2021.

[2] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[4] Mohammad Javad Pirhadi, Motahhare Mirzaei, Mohammad Reza Mohammadi, and Sauleh Eetemadi. PMCoders at SemEval-2023 task 1: RAltCLIP: Use relative AltCLIP features to rank. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1751–1755, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[6] Monica Sunkara, Srikanth Ronanki, Dhanush Bekal, Sravan Bodapati, and Katrin Kirchhoff. Multimodal Semi-Supervised Learning Framework for Punctuation Prediction in Conversational Speech. In *Proc. Interspeech 2020*, pages 4911–4915, 2020.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[8] Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. `https://github.com/abwilf/Social-IQ-2.0-Challenge`, 2023.

[9] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021.

[10] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Learning to answer visual questions from web videos. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022.

[11] Da Zhang, Xiyang Dai, Xin Wang, and Yuan-Fang Wang. S3D: single shot multi-span detector via fully 3d convolutional networks. *CoRR*, abs/1807.08069, 2018.