Comparing Classification and Generation Approaches to Situated Reasoning with Vision-language Pre-trained Models

Xin Huang, Hui Li Tan, Jung-jae Kim

Institute for Infocomm Research, A*STAR, Singapore
huangx2, hltan, jjkim@i2r.a-star.edu.sg

Abstract. Situated Reasoning in Real-World Videos (STAR) is a new benchmark for evaluating situated reasoning ability through situation abstraction and logic-grounded question answering on real-world videos. In this paper, we present our submission to the STAR challenge which achieves the top-1 result for the situated question reasoning. We investigated two approaches to utilizing a vision-language pre-trained model, including classification and generation methods, and we show that the generation method outperforms the classification method for all question types of the challenge. We also compared the methods with other baselines, including another vision-language pre-trained model, and discuss why different vision-language pre-trained models show significant performance gap for the STAR challenge.

1 Introduction

Situated Reasoning in Real-World Videos (STAR) is a new benchmark for evaluating situated reasoning ability through situation abstraction and logic-grounded question answering on real-world videos [17]. Built upon Charades dataset [15] of real-world videos associated with dynamic, compositional, and logical human actions and interactions, the dataset comprises four types of questions - interaction, sequence, prediction, and feasibility. Each question is associated with a video and its 'situation', represented by a hypergraph connecting atomic entities (e.g. actions, persons, objects) with relations. Situations of interaction and sequence questions contain complete action segments while those of prediction and feasibility questions contain incomplete action segments to deduce the answers of the questions. Furthermore, each question is associated with four choices in text form, where one of the four choices is the answer of the question.

The benchmark dataset is multi-modal, where the inputs consist of text (question) and video, and the outputs are of hypergraph (optional output) and text (final answer). To analyze the multi-modal inputs, we employ a pre-trained multi-modal (text+video) model, called Unified Video and Language (UniVL) [12]. UniVL was pre-trained with the instructional video dataset HowTo100M [13], using 5 pre-training objectives, including video-text joint, conditioned masked language model, conditioned masked frame model, video-text alignment, and language reconstruction. An instance of the benchmark dataset, pair of a question

and a video, is segmented into text tokens and video frames before inputted to the UniVL model, which in return selects the answer of the question among the given choices. We use the model for the answer selection with both classification and text generation approaches: In the classification approach, the model takes each choice as additional input and outputs the likelihood of the choice as answer. In the generation approach, the model learns to generate the answer as string. In this paper, we describe our experiments for comparing the two approaches for the video Q&A task of the STAR benchmark.

2 Literature Review

Since the seminal paper [16] shows that Transformer has good ability to model long-distance relationships, they have been used in language modelling, and extended to vision modelling and vision-language modelling [12, 10, 3]. The key factors for the success of Transformer for video Q&A are cross-modal attention and pre-training. For cross-modal attention, various methods of self-attention [12] and co-attention [11] for cross-modal modelling have been investigated. Cross-modal pre-training makes use of noisy but large-scale visual-text datasets, and the pre-trained model is transferred to downstream vision-language tasks (e.g. video Q&A) by fine-tuning it on small-scale labelled datasets.

Graph neural networks [6] and modular networks [9, 18] have also been investigated for video Q&A. As relation information is important for the reasoning ability for video Q&A, the graph neural networks are used to model relation information in form of graphs. Modular networks with reusable neural units are used for their better generalization ability. Neural symbolic networks [20, 2], which combine neural networks for pattern recognition and dynamic prediction, with symbolic logic for causal reasoning, have also been investigated for video Q&A.

In this paper, we explore leveraging pre-trained transformer-based visual-language model for the situated reasoning task. We adapted UniVL [12] for the STAR challenge. The UniVL has a text encoder, a video encoder, a cross encoder and task specific decoders such as a text decoder for generation tasks and a classification head for classification tasks. The cross encoder first concatenates latent representations of the video and text features, and then leverages a transformer encoder to perform cross attention between the two modalities. The UniVL model is pretrained on HowTo100M dataset. The model shows strong performance for visual language understanding and text generation task [7].

3 Methodology

Given a question q, a video clip v, and a list of candidate choices $C = \{c_1, ..., c_n\}$, the goal of the situated question answering task is to identify the correct answer c_{ans} among the candidates. We investigate two approaches for the situated question answering task: classification and generation.

Classification Model: The input to our classification model for the task is generated by merging the question q with each candidate choice c_i and then combining their tokenization results with the frame features of the video clip v. The output of the model is expected to be close to 1 if the given candidate choice is the correct answer, or to 0 otherwise.

We first obtain the text and video latent representations with a BERT encoder and a video transformer encoder. For the input of the video transformer encoder, we use the S3D embedding which can be extracted using the s3d feature extractor [19]. The learning rate of the classification model is $5e^{-5}$. There are 12 layers for the BERT model, 1 layer for the video transformer encoder, and 2 layers for the cross encoder. We use the hidden representations of the last layers of the transformer encoder and of the BERT encoder as latent representations (h_{T_i}, h_V), respectively, as follows:

$$h_{T_i} = BERT([q, c_i]),$$

$$h_V = TransformerEncoder(v).$$

And then we concatenate both hidden representations and use a cross encoder to obtain the fused latent representation for both video and text inputs, i.e.,

$$h_c = CrossEncoder([h_{T_i}, h_V]).$$

Finally, we use a classification head to obtain the logit of whether the choice is the answer. We used the cross entrophy as the loss function. We consider the choice with the highest score as our answer, i.e.

$$logits = Dense(h_c),$$

$$score = sigmoid(logits),$$

$$loss = CrossEntrophy(logits, c_{ans}).$$

Generation Model: The input to our generation model is generated by merging q with all the candidate choices C as a single string, before combined with the visual inputs like the classification model. The output of the model is expected to be the string of the correct answer.

We use the UniVL encoders for understanding the inputs of both classification and generation models and the UniVL text decoder for the generation model. We use the same hyperparameter values from the classification model. The decoder is a 1 layer transformer decoder with a cross attention layer $CrossAttn$.

After we obtain the concatenated hidden representation h_c, we use a transformer decoder to generate the answer text $A = w1, ..., w_d$ among the choices where d is the length of the answer text.

The predicted hidden representation of the decoder is used to proceed with answer generation. We used the beam search for the answer generation where the number of beams is 5.

4 Experiments

Dataset STAR consists 22k trimmed situation video clips, 144K situation hypergraphs, 60K situated reasoning questions with programs and answers, and 240K candidate choices. The interaction questions are to evaluate the understanding of interactions between human and objects in a situation (*e.g., "What did a person do ..."*). The sequence questions are to evaluate temporal relationship reasoning (*e.g., "What did the person do before/after ..."*). The prediction questions are to evaluate the forecasting of future actions based on the current situation (*e.g., "What will the person do next with ..."*). Lastly, the feasibility questions are to evaluate the ability to infer feasible actions under particular situation (*e.g., "What is the person able to do ..."*).

Baselines: We compared our proposed methods with the following baseline models for the video Q&A task:

- L-GCN (location-aware graph convolutional networks) [6]: It aims to model the interaction between objects related to questions. It represents the content in a video as a graph of objects detected by an off-the-shelf object detector and identifies actions through graph convolution networks. It represents a text with bi-LSTM that analyzes character embeddings and word embeddings. It combines the two modal representations with a complex module consisting of cross-modal attention, bi-LSTM and MaxPooling.
- HCRN (hierarchical conditional relation networks) [9]: Conditional Relation Network (CRN) is a reusable neural unit that encapsulates and transforms an array of objects into a new array conditioned on a contextual feature. HCRN is stacked CRNs. It represents text tokens with GloVe word embeddings [14] and video frames with ResNet [5] and ResNeXt-101 [4].
- ClipBERT [10]: It is a model pre-trained for video-language representation. It sparsely samples frames from a given video and encodes them with ResNet-50 and temporal fusion. It fuses the visual representation with text representation by BERT's word embedding layer, by training a transformer cross-encoder. It is pre-trained with COCO Captions [1] and Visual Genome Captions [8]. The model is fine-tuned with the task training data.

Experimental Results: A model is evaluated on the four question types of the situated reasoning task in terms of answer accuracy, individually, and the arithmetic mean of the four accuracies is also reported. Table 1 shows our evaluation results and the results of the baselines from [17]. The results show that our UniVL-based methods outperform all the baselines significantly. In particular, the UniVL-based methods outperform another vision-language pre-trained model ClipBERT. There can be multiple reasons for the performance difference between the two pre-trained models. First, UniVL extracts S3D features from video, while ClipBERT uses 2D features extracted by the ResNet-50. Second, ClipBERT uses sparse sampling of video frames and average-pooling of the sparsely sampled frames' representations, which may lose crucial temporal information in video. Third, UniVL is pre-trained on large-scale *video*-text dataset (HowTo100M), while ClipBERT is pre-trained on large-scale *image*-text dataset

(COCO Captions and Visual Genome), which may give UniVL advantage on video content.

Table 1: Experimental results on the official test-set for the situated reasoning Q&A task.

Model	Interaction	Sequence	Prediction	Feasibility	Mean
L-GCN [6]	39.01	37.97	28.81	26.98	33.19
HCRN [9]	39.10	38.17	28.75	27.27	33.32
ClipBERT [10]	39.81	43.59	32.34	31.42	36.79
UniVL-Classification	58.05	60.27	53.77	43.83	53.98
UniVL-Generation	**60.93**	**62.75**	**56.56**	**50.78**	**57.76**

Also, the proposed generation method shows higher performance for all question types than the proposed classification method. We argue that it is probably due to the difference of inputs to the two methods, where the input to the generation method includes all candidate choices of given question, while the input to the classification method includes only one candidate choice in order to test if the candidate choice answers the given question.

Table 2: Experimental results on the dev-set for the situated reasoning Q&A task with different setting.

Setting	Model	Mean
w/choices	UniVL-Generation	59.78
w/o choices	UniVL-Generation	39.14
w/choices	UniVL-Classification	58.21
w/o choices	UniVL-Classification	58.34

Table 2 shows our results on the dev-set with and without candidate choices as part of inputs to our models. We can see that the generation model requires the choices as part of input, since the choices guide the model to generate the answer string. For the classification model with choice, the input to the model is question plus each candidate choice string, the output is a binary value of whether the candidate is the answer, during the inference the choice with highest probability will be considered as answer. For the classification without choice, the input to the model is question only, and we merge all the candidate choices from the data-set of the STAR challenge into a set. The output is to classify which element of candidate set is the answer.We find that the classification model is not sensitive to the choices as feature.

We also evaluated our model performance with different percentage of training data. Figure 1 shows our model is also able to achieve good performance without full training data.

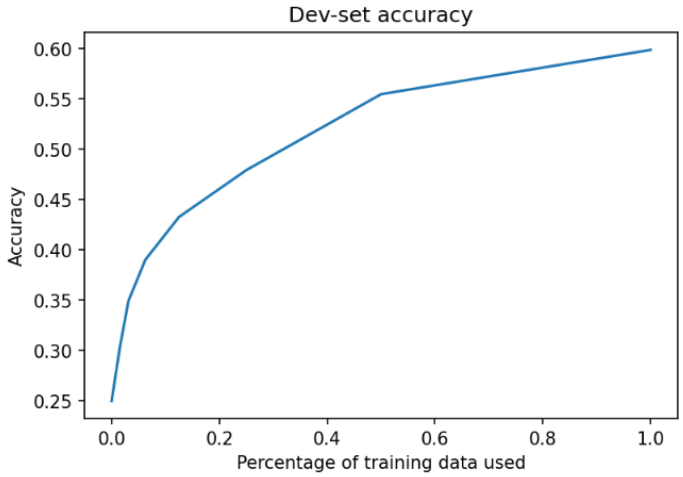

Fig. 1: The model mean accuracy on the dev-set achieved with different percentage of training data

5 Conclusion

We present our submission to the STAR challenge, which is based on the video-language pre-trained model UniVL. We utilize the model in two approaches: classification and generation. We show that the generation method outperforms the classification method for all the question types of the challenge.

References

1. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. CoRR **abs/1504.00325** (2015), http://dblp.uni-trier.de/db/journals/corr/corr1504.htmlChenFLVGDZ15
2. Ding, M., Chen, Z., Du, T., Luo, P., Tenenbaum, J.B., Gan, C.: Dynamic visual reasoning by learning differentiable physics models from video and language. CoRR **abs/2110.15358** (2021), https://arxiv.org/abs/2110.15358
3. Fu, T., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: VIOLET: End-to-end video-language transformers with masked visual-token modeling. CoRR **abs/2111.12681** (2021), https://arxiv.org/abs/2111.12681

4. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? (06 2018). https://doi.org/10.1109/CVPR.2018.00685
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015). https://doi.org/10.48550/ARXIV.1512.03385, https://arxiv.org/abs/1512.03385
6. Huang, D., Chen, P., Zeng, R., Du, Q., Tan, M., Gan, C.: Location-aware graph convolutional networks for video question answering. CoRR **abs/2008.09105** (2020), https://arxiv.org/abs/2008.09105
7. Huang, X., Tan, H.L., Leong, M.C., Sun, Y., Li, L., Jiang, R., jae Kim, J.: Investigation on transformer-based multi-modal fusion for audio-visual scene-aware dialog. In: Dialog System Technology Challenge (DSTC10) Workshop of AAAI 2021 (2021)
8. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vision **123**(1), 32–73 (may 2017). https://doi.org/10.1007/s11263-016-0981-7, https://doi.org/10.1007/s11263-016-0981-7
9. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. CoRR **abs/2002.10698** (2020), https://arxiv.org/abs/2002.10698
10. 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)
11. Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., Gan, C.: Beyond rnns: Positional self-attention with co-attention for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8658–8665 (2019)
12. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
13. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019)
14. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1162, https://aclanthology.org/D14-1162
15. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.K.: Hollywood in homes: Crowdsourcing data collection for activity understanding. ArXiv **abs/1604.01753** (2016)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017). https://doi.org/10.48550/ARXIV.1706.03762, https://arxiv.org/abs/1706.03762
17. Wu, B., Yu, S., Chen, Zhenfang, T.J.B., Gan, C.: STAR: A benchmark for situated reasoning in real-world videos. In: Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) (2021)
18. Xiao, J., Yao, A., Liu, Z., Li, Y., Ji, W., Chua, T.: Video as conditional graph hierarchy for multi-granular question answering. CoRR **abs/2112.06197** (2021), https://arxiv.org/abs/2112.06197
19. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: ECCV. pp. 318–335 (2018)

20. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: CLEVRER: Collision events for video representation and reasoning. CoRR **abs/1910.01442** (2019), http://arxiv.org/abs/1910.01442