The 7th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2023)

Utilization of Image-Text Embeddings for Audio-Visual Scene Dialogue

Jungwoo Lim[1], Sungmin Ahn[2], Kinam Park[3‡]
[1]Department of Computer Science and Engineering, Korea University,
[2]O2O Inc., [3]Human-inspired AI Research
{wjddn803,spknn}@korea.ac.kr, smahn@o2o.kr

Abstract

Current dialogue systems in real-world applications like smartphone assistants and car navigations are limited to text-based interactions and single modality responses. They are not capable of understanding multi-modal inputs, such as visuals or audio, limiting their utility in context-aware conversations. Although various video-based dialogue systems have been developed using the AVSD dataset, most have focused on merely synthesizing individual visual, image, and audio features. Research suggests that pre-training models to align image and text features before applying them to specific tasks leads to better performance. Recognizing this gap and the need for more advanced capabilities, the proposed model aims to perform AVSD using pre-trained image-text features, aiming to enhance the system's understanding of real-world contexts.

Keywords: Video, Dialogue, CLIP

1. Introduction

Recently, various dialogue systems are being applied to real-world human-machine interfaces like smartphone assistants, car navigation, voice-controlled speakers, and human-centric robots. However, in these applications, all dialogue is entered as text, and the content of the system's response is limited to a single modality. Current dialogue systems are unable to understand multi-modality based inputs such as visuals or voice audio, which means machines using these dialogue systems can't have conversations about what's happening around them. To develop a system capable of conversing about events occurring around the user, video-based dialogue systems that are key in multi-modality scene recognition are essential.

To this end, various video-based dialogue systems have emerged using the AVSD (Audio Visual Scene-aware Dialog) dataset, but existing video-based dialogue systems have focused on synthesizing individual visual, image, and audio features [1, 2, 3]. However, it has been confirmed that models that first undergo pre-training to align image-text well, and then use these features for specific tasks, show better understanding capabilities in more tasks compared to simply synthesizing image, audio, and text features for a specific task [4, 5]. Due to the fact that a model that performs AVSD using these embeddings is absolutely necessary, we therefore propose a model for performing AVSD (Audio Visual Scene-aware Dialog) using image-text features.

2. Related Work

The CLIP (Contrastive Language-Image Pretraining) model is a multimodal machine learning model developed by OpenAI [4]. This model is designed to recognize images related to text descriptions or conversely, generate text descriptions appropriate for given images. CLIP is trained using a contrastive loss function that strengthens the positive correlation between text and images while weakening the correlation with other text-image pairs. This allows the model to find relevant images when a text description is given or generate appropriate text descriptions when an image is provided. It is pre-trained on a large dataset of text and image pairs and is designed to perform well in various visual tasks. Similarly, BLIP-2 [5] bootstraps vision-language pretraining using existing pre-trained image encoders and large language models. BLIP-2 employs a lightweight Querying Transformer that has been pre-trained in two stages to bridge the gap between modalities. In the first stage, vision-language representation learning is bootstrapped using a pre-trained image encoder. In the second stage, generative learning from

vision to language is bootstrapped using a pre-trained language model.

3. Proposed Method

The proposed model enhances video understanding by utilizing pre-trained embeddings that are aligned for image-text, and also leverages behavioral cues necessary for understanding the video. The proposed model works as follows: 1) It acquires text embeddings by encoding the dialogue using an image-text pre-training model, and obtains image embeddings by encoding frames from the video. It also identifies what kinds of sounds are present in the audio using the cosine similarity between embeddings. 2) It calculates the similarity between the encoded dialogues and frames to extract relevant frames. 3) It predicts salient behavioral cues from these frames. Finally, 4) it generates utterances using the dialogue, predicted behavioral cues, and auditory cues.

In other words, the proposed model performs encoding for text, video, and audio, and calculates the frames most relevant to the conversation using CLIP [4]. It then predicts prominent behavioral cues from the frames with the highest similarity. Additionally, it predicts what kinds of sounds are present in the video to acquire auditory cues with BEATs [6]. Using these acquired cues along with the existing conversation history, it generates the next utterance utilizing BART [7]. In this process, the pre-trained image-text and audio models are both frozen in terms of parameters and are not fine-tuned.

4. Conclusion

Although various dialogue systems are currently being applied in the real world, most are text-based and unable to handle multi-modality, missing crucial behavioral and auditory cues. In this paper, we propose a new video-based dialogue system to address this issue, utilizing image-text alignment and behavioral cues.

5. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1A2C1007616). Following are results of a study on the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea

Reference

[1] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das *et al.*, "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2352–2356, 2019.

[2] I. Schwartz, A. G. Schwing, and T. Hazan, "A simple baseline for audio-visual scene-aware dialog," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12 548–12 558, 2019.

[3] X. Lin, G. Bertasius, J. Wang, S.-F. Chang, D. Parikh, and L. Torresani, "Vx2text: End-to-end learning of video-based text generation from multimodal inputs," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7005–7015, 2021.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *International conference on machine learning*, pp. 8748–8763, 2021.

[5] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[6] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.