Audio Visual Scene-Aware Dialog Track in DSTC8

Chiori Hori*, **Anoop Cherian***, **Tim K. Marks***, **and Florian Metze****

*Mitsubishi Electric Research Laboratories
Cambridge, MA, USA
{chori, cherian, tmarks}@merl.com
**Carnegie Mellon University
{fmetze}@cmu.edu

Abstract

Dialog systems need to understand scenes in order to have conversations with users about the objects and events in them. We introduced a new challenge task and dataset for Audio Visual Scene-Aware Dialog (AVSD) in DSTC7, which was the first attempt to combine conversation and multimodal video description into a single end-to-end differentiable network to build scene-aware dialog systems. The winning system of the challenge applied hierarchical attention mechanisms to combine text and visual information, yielding a relative improvement of 22% in the human rating of the output of the winning system vs. that of the baseline system. The language models trained from QA contributed most to this improvement, while the benefits from visual models (such as object or event recognition from videos) were limited. This means that there is still more opportunity to boost performance on the AVSD challenge using video features. To encourage such progress, we propose a second challenge for DSTC8 as a follow-up to the video-based scene-aware dialog track from DSTC7. The task is to generate or select a system response to a query that occurs during a dialog about a video. Participants will use the video, audio, and dialog text data to train end-to-end models. We will again use the AVSD data sets that we collected and used at DSTC7. We may also include additional data sets, such as the How2 and Dense-Captioning datasets, to provide example dialogs that have more long-term history dependence.

1 Introduction

An automated system that can converse with humans on video scenes via natural dialogs is a challenging research problem that lies at the intersection of natural language processing, computer vision, and audio processing. As seen at DSTC6 and DSTC7, end-to-end dialog modeling using paired input and output sentences is a way to reduce the cost of data preparation and system development to generate reasonable dialogs in many situations. Such end-to-end approaches have been shown to better handle flexible conversations by enabling model training on large conversational datasets [1, 2, 3]. In the field of computer vision, interaction with humans about visual information has been explored in *visual question answering* (VQA) by [4] and *visual dialog* by [5]. These tasks have been the focus of intense research recently, aiming to (1) generate answers to questions about objects and events in a single static image and (2) hold a meaningful dialog with humans about an image using natural, conversational language in an end-to-end framework. To capture the semantics of dynamic scenes, recent research has focused on *video description*[6] . The state-of-the-art in video description uses multimodal fusion to combine different input modalities (feature types), such as spatio-temporal motion features and audio features [7]. Since the recent revolution of neural network models allows us to combine different modules into a single end-to-end differentiable network, this framework allows us to build scene aware dialog systems by combining dialog and multimodal video description

approaches. That is, we can simultaneously use video features and user utterances as input to an encoder-decoder-based system whose outputs are natural-language responses.

To advance research into multi-modal reasoning-based dialog generation, we developed the AVSD dataset and proposed a challenge in DSTC7. The goal was to design systems to generate responses in a dialog about a video, given the dialog history and audio-visual content of the video. The winning system of the challenge applied hierarchical attention mechanisms to combine text and visual information, yielding a relative improvement of 22% in the human rating of the output of the winning system vs. that of the baseline system. This suggests that there is perhaps significantly more potential in-store for advancing this new research area. Towards this end, we propose a second edition of our AVSD challenge in DSTC8.

2 Data

To set up the Audio Visual Scene-Aware Dialog (AVSD) track, we will use the AVSD data used in DSTC7. We collected text-based dialogs on short videos from the popular Charades dataset [8], which consists of untrimmed and multi-action videos (each video also has an audio track) and comes with human-generated descriptions of the scene [9]. The data collection paradigm for dialogs was similar to the one described in [10] for image-based dialog (Visual Dialog), in which for each image, two parties interacted via a text interface to yield a dialog. In [10], each dialog consisted of a sequence of questions and answers about an image.

In our video scene-aware dialog case, two parties, dubbed *questioner* and *answerer*, have a dialog about events in the provided video. The job of the answerer, who has already watched the video, is to answer questions asked by the questioner. The questioner, who is not permitted to see the video, is only shown the first, middle and last frames of the video as static images (see Figure 1). The two parties have 10 rounds of QA, in which the questioner asks about the events that happened in the video. At the end of the dialog, the questioner uses the knowledge gained from the dialog (and the three static frames) to write a video description summarizing the events in the video.

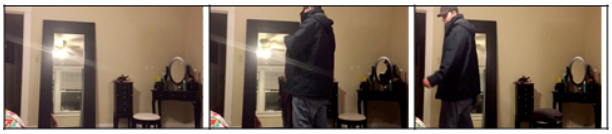

Figure 1: An example a set of images of first, middle, and last frames for a questioner.

The DSTC7 AVSD official dataset contains dialogs about 11,156 videos (7,659 training, 1,787 validation, and 1,710 test). The questions and answers of the AVSD dataset mainly consists of 5 to 8 words, making them longer and more descriptive than those in VQA. The dialog contains questions asking about objects, actions, and audio information in the videos. Table 2 shows an example dialog from the data set. Wheres the questions in the VQA data set are mainly about objects in images, the AVSD dataset has questions about objects, actions, and sounds in the videos. For DSTC8, we will prepare new test data similar to the test data used in the DSTC7 AVSD track.

The AVSD dialog data set consists of a sequence of Question and Answer pairs (QAs) about the events in the video that enable the questioner to summarize the video. Video and audio features are important for answering questions about actions and sounds, respectively. Results from the AVSD track of DSTC7 demonstrate the promise of this challenge to promote the development of automatic scene-aware dialog systems. In order to promote additional focus on long-term dialog history, we may additionally use other data sets such as How2 dataset[11] and Dense-Captioning Events in Videos [12].

3 Task definition

In this track, the system must generate responses to a user input in the context of a given dialog. The dialog context consists of a dialog history between the user and the system in addition to the video and audio information in the scene. There are two tasks, each with two versions (a and b):

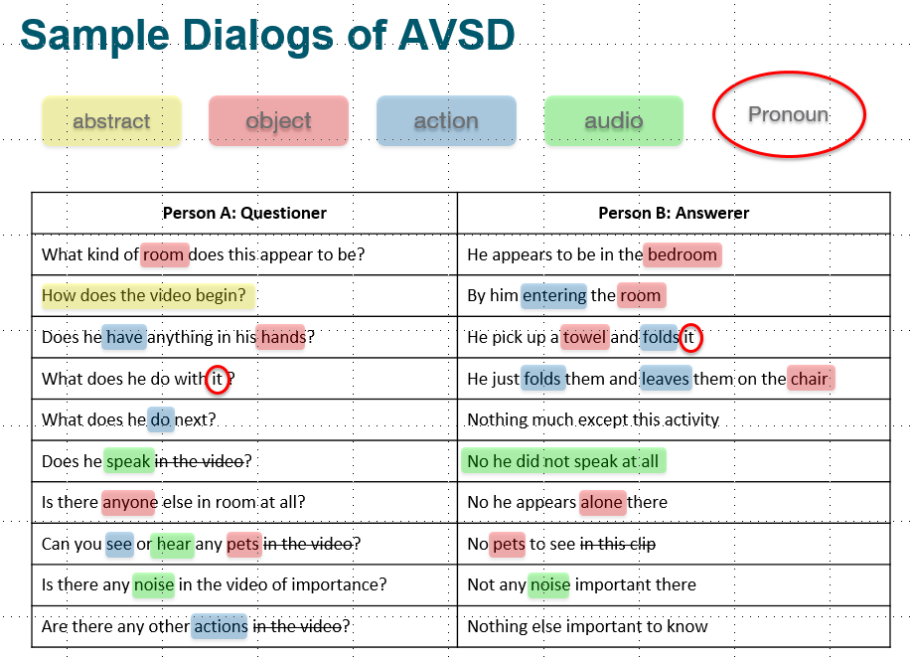

Figure 2: An example dialog from the AVSD dataset.

Task 1: Video and Text (a) Using the provided video and text training data, but no external data sources, other than publicly available pre-trained feature extraction models (b) Also using external data for training.

Task 2: Text Only (a) Do not use the input videos for training or testing. Use only the text training data (dialogs and video descriptions) provided. (b) Any publicly available text data may be used for training.

Challenge participants can select to submit entries in Task 1, Task 2, or both. The training data and a baseline system will be released to all participants of DSTC8.

3.1 Objective evaluation

The quality of the automatically generated sentences will be evaluated using objective measures that capture the similarity between the generated sentences and ground truth sentences. For the challenge track, we will use `nlg-eval`[1] for objective evaluation of system outputs, which is a publicly available supporting various unsupervised automated metrics for natural language generation as we have done at DSTC7. The supported metrics include word-overlap-based metrics such as BLEU, METEOR, ROUGE_L, and CIDEr, etc. Details of these metrics are described in [13].

4 Summary

This article described the proposed Video Scene-Aware Dialog track for the 8th Dialog System Technology Challenges (DSTC8) workshop. The information provided to participants will include a detailed description of the baseline system, instructions for submitting results for evaluation, and details of the evaluation scheme.

References

[1] Oriol Vinyals and Quoc Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[1] `https://github.com/Maluuba/nlg-eval`

[2] Chiori Hori, Julien Perez, Ryuichi Higasinaka, Takaaki Hori, Y-Lan Boureau, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, Koichiro Yoshino, and Seokhwan Kim, "Overview of the sixth dialog system technology challenge: Dstc6," 2018.

[3] Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S. Lasecki, Jonathan Kummerfeld, Michael Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Sean Gao, Tim K. Marks, Devi Parikh, and Dhruv Batra, "The 7th dialog system technology challenge," *arXiv preprint*, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[5] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra, "Learning cooperative visual dialog agents with deep reinforcement learning," in *International Conference on Computer Vision (ICCV)*, 2017.

[6] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko, "Sequence to sequence - video to text," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 4534–4542.

[7] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi, "Attention-based multimodal fusion for video description," in *ICCV*, 2017.

[8] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ivan Laptev, Ali Farhadi, and Abhinav Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *ArXiv*, 2016.

[9] Huda Alamri, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Jue Wang, Irfan Essa, Dhruv Batra, Devi Parikh, Anoop Cherian, Tim K Marks, et al., "Audio visual scene-aware dialog (avsd) challenge at dstc7," *arXiv preprint arXiv:1806.00525*, 2018.

[10] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra, "Visual dialog," *CoRR*, vol. abs/1611.08669, 2016.

[11] "How2: A large-scale dataset for multimodal language understanding," .

[12] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, "Dense-captioning events in videos," in *International Conference on Computer Vision (ICCV)*, 2017.

[13] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer, "Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation," *CoRR*, vol. abs/1706.09799, 2017.