

Audio Visual Scene-aware dialog (AVSD) Track for Natural Language Generation in DSTC7

Huda Alamri^{*†}, Chiori Hori[†], Tim K. Marks[†], Dhruv Batra^{*}, Devi Parikh^{*},

[†]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

^{*}School of Interactive Computing, Georgia Tech

Abstract

Dialog systems need to understand dynamic visual scenes in order to have conversations with users about the objects and events around them. In this Audio Visual Scene-aware Dialog (AVSD) track for Natural Language Generation (NLG), we challenge a new research target, a dialog system that can have conversations with users about the objects and events around them, which lies at the intersection of multiple avenues of research in natural language processing, computer vision, and audio processing. We introduce a new dataset of dialogs about videos of human behaviors. Each dialog is a typed conversation that consists of a sequence of 10 question-and-answer (QA) pairs between two parties. In total, we collected dialogs on 11,156 videos. In this overview, we describe the task design and data sets, and review the submitted systems and applied techniques for conversation modeling. The AVSD track for Natural Language Generation (NLG) received 31 system submission from a total of 9 teams, and evaluated them based on several objective measures such as BLEU, METEOR, CIDEr and a human-rating based subjective measure. Finally, we discuss technical achievements and remaining problems related to this challenge.

Introduction

Spoken dialog technologies have been applied in real-world human-machine interfaces including smart phone digital assistants, car navigation systems, voice-controlled smart speakers, and human-facing robots (McTear 2002; Young 2000; Zue et al. 2000). Generally, a dialog system consists of a pipeline of data processing modules, including automatic speech recognition, spoken language understanding, dialog management, sentence generation, and speech synthesis. However, all of these modules require significant hand engineering and domain knowledge for training. Recently, end-to-end dialog systems have been gathering attention, and they obviate this need for expensive hand engineering to some extent. In end-to-end approaches, dialog models are trained using only paired input and output sentences, without relying on pre-designed data processing modules or intermediate internal data representations such as concept tags and slot-value pairs. End-to-end systems can be trained to directly map from a user's utter-

ance to a system response sentence and/or action. This significantly reduces the data preparation and system development cost. Several types of sequence-to-sequence models have been applied to end-to-end dialog systems, and it has been shown that they can be trained in a completely data-driven manner. End-to-end approaches have also been shown to better handle flexible conversations between the user and the system by training the model on large conversational datasets (Vinyals and Le 2015; Lowe et al. 2015; Hori et al. to appear in 2018). In these applications, however, all conversation is triggered by user speech input, and the contents of system responses are limited by the training data (a set of dialogs). Current dialog systems cannot understand dynamic scenes using multimodal sensor-based input such as vision and non-speech audio, so machines using such dialog systems cannot have a conversation about what's going on in their surroundings. To develop machines that can carry on a conversation about objects and events taking place around the machines or the users, dynamic scene-aware dialog technology is essential.

Using this end-to-end framework, *visual question answering* (VQA) has been intensively researched in the field of computer vision (Antol et al. 2015; Zhang et al. 2016; Goyal et al. 2017). The goal of VQA is to generate answers to questions about an imaged scene, using the information present in a single static image. As a further step towards conversational visual AI, the new task of *visual dialog* was introduced (Das et al. 2016), in which an AI agent holds a meaningful dialog with humans about an image using natural, conversational language (Das et al. 2017). While VQA and visual dialog take significant steps towards human-machine interaction, they only consider a single static image. Most real-world scenarios such as helping visually-impaired users, intelligent home assistants involve videos specifically, understanding the content and temporal dynamics of a scene. To capture the semantics of dynamic scenes, recent research has focused on *video description* (natural-language descriptions of videos). The state of the art in video description uses a multimodal attention mechanism that selectively attends to different input modalities (feature types), such as spatiotemporal motion features and audio features, in addition to temporal attention (Hori et al. 2017). This framework allows us to build scene aware dialog systems using multimodal information such as audio and visual features.

Table 1: An example dialog of the AVSD dataset.

	Questioner	Answerer
QA1	What kind of room does this appear to be?	He appears to be in the bedroom.
QA2	How does the video begin?	By him entering the room.
QA3	Does he have anything in his hands?	He pick up a towel and folds it.
QA4	What does he do with it ?	He just folds them and leaves them on the chair.
QA5	What does he do next?	Nothing much except this activity.
QA6	Does he speak in the video?	No he did not speak at all.
QA7	Is there anyone else in room at all?	No he appears alone there.
QA8	Can you see or hear any pets in the video?	No pets to see in this clip.
QA9	Is there any noise in the video of importance?	Not any noise important there.
QA10	Are there any other actions in the video?	Nothing else important to know.

In this track, we propose a new research target, a dialog system that can discuss dynamic scenes with humans, which lies at the intersection of multiple avenues of research in natural language processing, computer vision, and audio processing. Since the recent revolution of neural network models allows us to combine different modules into a single end-to-end differentiable network, this framework allow us to build scene aware dialog systems by combining end-to-end dialog and multimodal video description approaches. We can simultaneously input video features and user utterances into an encoder-decoder-based system whose outputs are natural-language responses. To advance this goal, we introduce a new dataset of human dialogues about videos. We used the short video clips of the Charades dataset videos (Sigurdsson et al. 2016) for Audio Visual Scene-aware Dialog (AVSD) which is simple videos of real people performing everyday actions in real-world settings, with natural audio. In this track, system responses are generated using Natural language Generation (NLG) technologies which challenged by the DSTC6 Trak2 modeling end-to-end conversation for Twitter customer service. We used the baseline system that incorporated technologies for video description into an end-to-end dialog system (Hori et al. 2018). We are making our dataset, code, and model publicly available for a new Audio Visual Scene-Aware Dialog (AVSD) Challenge.

Task definition

The system must generate responses to a user input in the context of a given dialog in this track. 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):

Task 1: Video and Text (a) Use the video and text training data provided but no external data sources, other than publicly available pre-trained feature extraction models. (b) External data may also be used 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 also be used for training.

Data Collection

To setup the Audio Visual Scene-Aware Dialog (AVSD) track, we collected text-based dialog data discussing about

Table 2: The dialog data for the DSTC7 AVSD track. The test videos for this challenge was partially selected from the official test data of the Charades challenge.

	training	validation	test
# of dialogs	7,659	1,787	1,710
# of turns	153,180	35,740	13,490
# of words	1,450,754	339,006	110,252

short videos of Charades by (Sigurdsson et al. 2016)¹, which contains untrimmed and multi-action videos, along with video descriptions in (Alamri et al. 2018). The data collection paradigm for dialogs was similar to the one described in (Das et al. 2016), in which for each image, two party interacted via a text interface to yield a dialog. In (Das et al. 2016), each dialog consisted of a sequence of questions and answers about an image. In the video scene-aware dialog case, two party had a discussion about events in a video. One of the two-party played the role of an answerer who had already watched the video. The answerer answered questions asked by the counterpart – the questioner. The questioner was not allowed to watch the whole video but only the first, middle and last frames of the video which were single static images. After having a conversation to ask about the events that happened between the frames through 10 rounds of QA, the questioner summarized the events in the video as a description. The sample dialog is show in Table 1. The DSTC7 AVSD official dataset for NLG contains 7,659, 1,787 and 1,710 dialogs for training, validation and test sets, respectively. The questions and answers of the AVSD dataset mainly consists of 5 to 8 words and longer than those of VQA. More descriptive sentences were generated. The dialog contains questions asking objects, actions and audio information in the videos. Although we tried to collect the event relevant questions, some questions ask abstract information of the video such as how to begin the videos, the duration of the videos other than the events. Table 2 shows the statistics of the data set.

Baseline System

We provided a baseline end-to-end dialog system that can generate answers in response to user questions about events

¹<http://allenai.org/plato/charades/>

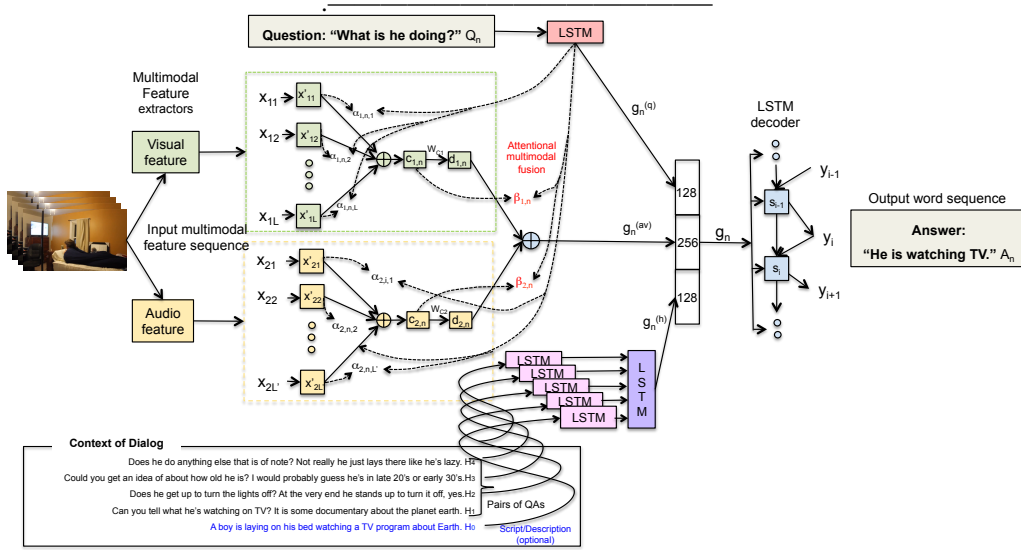


Figure 1: Attentional multimodal fusion based video scene-aware dialog system (Hori et al. 2018)

in a video sequence. Our architecture is based on the Track2 of DSTC6 (Hori et al. to appear in 2018) which similar to the Hierarchical Recurrent Encoder in Das *et al.* (Das et al. 2016). The question, visual features, and the dialog history are fed into corresponding LSTM-based encoders to build up a context embedding, and then the outputs of the encoders are fed into a LSTM-based decoder to generate an answer. The history consists of encodings of QA pairs. We feed multimodal video features into the LSTM encoder instead of single static image features as described in (Hori et al. 2018). Figure 1 shows the architecture of the baseline system. In the AVSD challenge in DSTC7, we provide a Naïve fusion multimodal fusion system (Alamri et al. 2018).

Video Processing

We adopted the state-of-the-art I3D features (Carreira and Zisserman 2017), spatiotemporal features that were developed for action recognition. The I3D model inflates the 2D filters and pooling kernels in the Inception V3 network along their temporal dimension, building 3D spatiotemporal ones. We used the output from the "Mixed_5c" layer of the I3D network to be used as video features in our framework. As a pre-processing step, we normalized all the video features to have zero mean and unit norm; the mean was computed over all the sequences in the training set for the respective feature.

In the experiments in this paper, we treated I3D-rgb (I3D features computed on a stack of 16 video frame images) and I3D-flow (I3D features computed on a stack of 16 frames of optical flow fields) as two separate modalities that are input to our multimodal attention model. To emphasize this, we refer to I3D in the results tables as I3D (rgb-flow).

Audio Processing

In this track, we used features extracted using a new state-of-the-art model, Audio Set VGGish (Hershey et al. 2017).

Inspired by the VGG image classification architecture (Configuration A without the last group of convolutional/pooling layers), the Audio Set VGGish model operates on 0.96 sec log Mel spectrogram patches extracted from 16 kHz audio, and outputs a 128-dimensional embedding vector. The model was trained to predict an ontology of labels from only the audio tracks of millions of YouTube videos. In this work, we overlap frames of input to the VGGish network by 50%, meaning an Audio Set VGGish feature vector is output every 0.48 sec.

Submitted Systems

We received 32 sets of system outputs for the AVSD task, from 9 teams, and eight system description papers were accepted (Sanabria, Palaskar, and Metze 2019) (Nguyen et al. 2019) (Pasunuru and Bansal 2019) (Yeh et al. 2019) (Zhuang, Wang, and Shinozaki 2019) (Kumar et al. 2019) (Lin et al. 2019) (Le et al. 2019).

In this section, we summarize the techniques used in the systems, including the baseline system for the challenge track. The baseline system is an LSTM-based encoder-decoder with Naïve multimodal fusion in (Alamri et al. 2018). This is a simplified version of (Hori et al. 2018), in which multimodal fusion is performed without attention between modalities such as audio and video features. The full set of the test data was used in (Hori et al. 2018), on the other hand the AVSD challenge at DSTC7 selected 2,000 responses from the full set.

Table 3 shows the baseline and submitted systems with their brief specifications including Encoder-decoder Model type, Multimodal fusion type, and Additional techniques, models and data sets. Most systems employed an LSTM, Bi-LSTM or GRU encoder/decoder. Some systems used a hierarchical and attention frameworks. Furthermore, several additional techniques are introduced to improve the response quality such as MMI, Episodic Memory Module.

Table 3: Submitted systems to the AVSD Track.

Team	Encoder-decoder type	Multimodal fusion type	Additional techniques/data
baseline	LSTM	Naïve fusion	
team_1	Bidirectional Gated Recurrent Units (GRU) based encode, Conditional Gated Recurrent Units (CGRU) based decoder	Hierarchical attention	ResNeXt, Transfer learning using How2 dataset
team_2	FiLM Attention Hierarchical Recurrent Encoder Decoder (FA-HRED), LSTM	Naïve fusion	FiLM
team_3	Dual attention LSTM encoder,	Cross-attention fusion	Similarity matrix
team_4	LSTM/GRU encoder, Top-down Attention LSTM/GRU decoder	Muti-stage fusion, 1x1 Convolution fusion, Multi-head Attention	
team_5	Bi-LSTM and LSTM encoder, LSTM decoder	Attentional multimodal fusion	MMI objective
team_6	LSTM encoder-decoder	Attentional multimodal fusion	Topic-base Conceptual model, ConvNet, AclMet
team_7	–	–	–
team_8	Bi-LSTM/LSTM encoder, Attention-based GRU encoder, LSTM decoder	Entropy-enhanced Dynamic Memory Network (DMN)	Episodic Memory Module
team_9	GRU encoder-decoder	Question-to-Caption/Multimodal attention	

⁺ Team 7 did not submit a system description paper to the DSTC7 workshop.

Evaluation

In this challenge, the quality of a system’s automatically generated sentences is evaluated using objective measures to determine how similar generated responses and ground truths by humans and how much natural and informative as responses. To collect more possible answers in response to the questions for the test videos, we asked 5 humans to watch a video and read a dialogue between a questioner and an answer about the video, and then to generate an answer in response to the question. We evaluated the automatic generated answers by comparing with the 6 ground truths consisting of one original answer and 5 newly collected answers. We used the MSCOCO evaluation tool for objective evaluation of system outputs². The supported metrics include word-overlap-based metrics such as BLEU, METEOR, ROUGE.L, and CIDEr.

Furthermore, we collected human ratings for each system response using 5 point Likert Scale, where humans rated system responses given a dialog context by 5 level scores as: 5 for Very good, 4 for Good, 3 for Acceptable, 2 for Poor, 1 for Very poor. Since we used a question answering dialog dataset, we asked humans to consider correctness of the answers and also naturalness, informativeness, and appropriateness of the response according to the given context.

Figures 2-4 show the human ratings for each system in several ways. The systems are shown in the same order on the X axis for all three figures. Figure 2 shows the mean and the standard deviation of the human ratings for each system (across all responses and all raters for that system). Figure 3 shows the distributions of the mean human rating score for each sentence for each system. Figure 4 shows the distribution of all human rating scores for each system across all sentences. In this Figure, the area for each score of the violinplot shows a count of the number of scores of each level on the Likert scale. The "Reference" system (at the

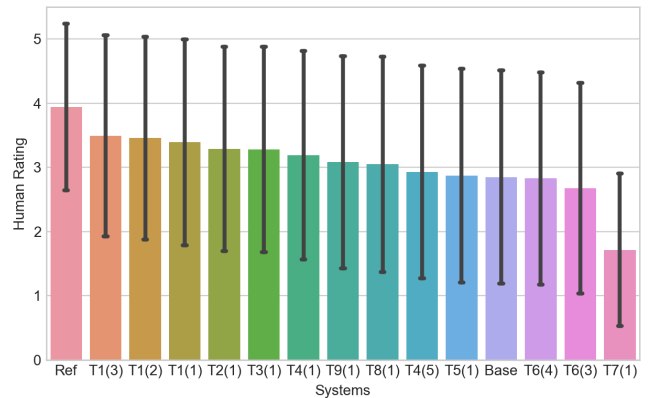


Figure 2: Mean and standard deviation of human rating score.

far left of each figure) is ratings for the sentences extracted from the original QA data of the AVSD dataset. The Reference system had the best human ratings: it had the highest mean rating in Fig. 2, the highest median sentence rating in Fig. 3 and the most sentences rated as level 5 ("Very good") in Fig. 4. The worst system (at the right) had a much lower mean rating, and a long tail of poorly rated sentences.

The human rating for end-to-end conversation models for Twitter customer services reported in (Hori et al. to appear in 2018) shows various quality from 1 to 5. On the other hand, the human rating for the AVSD track split into the binary class such as "good" and "bad". This is because the quality of the answers depends on the answer correctness in response to the questions and the incorrect answers result in more drastic changes of the human rating scores. The best system generated more correct answers and the worst system generated incorrect answers mostly.

²<https://github.com/tylin/coco-caption>

Table 4: Evaluation results with word-overlapping-based objective measures based on 6 references and a subjective measure based on 5-level ratings for the AVSD track. The details of the system description will be added after getting the system description papers from the system submitters.

Team	Entry	text only	video	caption and/or summary	extra	prototype	Bleu_4	METEOR	ROUGE_L	CIDEr	Human rating
Team 1	(1)	✓		✓	✓		0.376	0.264	0.554	1.076	3.394
	(2)		✓	✓	✓		0.387	0.266	0.564	1.087	3.459
	(3)		✓	✓			0.394	0.267	0.563	1.094	3.491
	(4)	✓		✓			0.364	0.254	0.543	1.006	-
Team 2	(1)		✓	✓			0.360	0.249	0.544	0.997	3.288
	(2)	✓	✓	✓			0.323	0.231	0.510	0.843	
	(3)	✓		✓			0.343	0.243	0.536	0.920	
	(4)	✓		✓			0.340	0.228	0.518	0.851	
	(5)			✓		✓	0.349	0.242	0.536	0.947	
	(6)		✓	✓		✓	0.316	0.224	0.505	0.795	
	(7)		✓	✓		✓	0.319	0.228	0.513	0.836	
	(8)	✓		✓		✓	0.323	0.220	0.501	0.799	
Team 3	(1)		✓	✓			0.337	0.242	0.532	0.957	3.279
Team 4	(1)		✓	✓		✓	0.342	0.223	0.504	0.837	3.188
	(2)		✓	✓			0.345	0.224	0.505	0.877	
	(3)		✓	✓		✓	0.342	0.223	0.504	0.836	
	(4)	✓		✓			0.304	0.207	0.477	0.731	
	(5)	✓		✓			0.304	0.206	0.475	0.729	2.928
Team 5	(1)		✓			✓	0.293	0.221	0.486	0.761	2.869
	(2)		✓			✓	0.302	0.222	0.488	0.770	
	(3)		✓			✓	0.302	0.222	0.487	0.769	
	(4)		✓			✓	0.296	0.219	0.484	0.745	
	(5)		✓			✓	0.283	0.217	0.480	0.731	
Team 6	(1)	✓	✓	✓		✓	0.307	0.213	0.469	0.701	
	(2)	✓	✓	✓		✓	0.307	0.215	0.479	0.733	
	(3)	✓	✓	✓		✓	0.278	0.198	0.442	0.614	2.675
	(4)	✓	✓	✓		✓	0.310	0.217	0.483	0.718	2.827
Team 7	(1)	✓		✓			0.056	0.096	0.236	0.085	1.715
Team 8	(1)		✓	✓			0.310	0.241	0.527	0.912	3.048
	(2)		✓	✓			0.307	0.239	0.525	0.915	
Team 9	(1)	✓		✓			0.310	0.242	0.515	0.856	3.080
	(2)		✓				0.315	0.239	0.509	0.848	
Reference											3.938
Baseline w/o audio			✓				0.305	0.217	0.481	0.733	
Baseline			✓				0.309	0.215	0.487	0.746	2.848

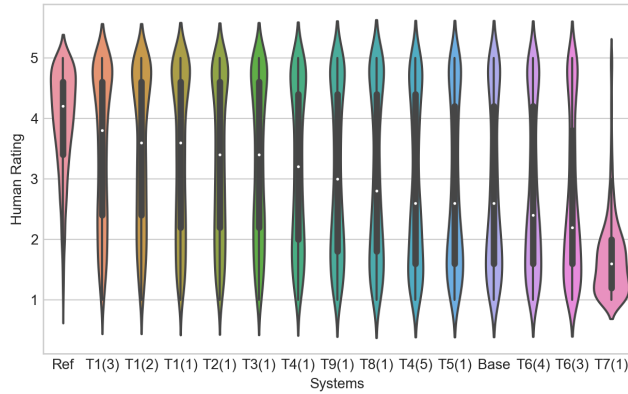


Figure 3: Distribution of human scores averaged sentence by sentence.

Conclusion

We introduced a new challenge task and dataset for Audio Visual Scene-Aware Dialog (AVSD) in DSTC7. This is the first attempt to combine end-to-end conversation and end-to-end multimodal video description models into a single

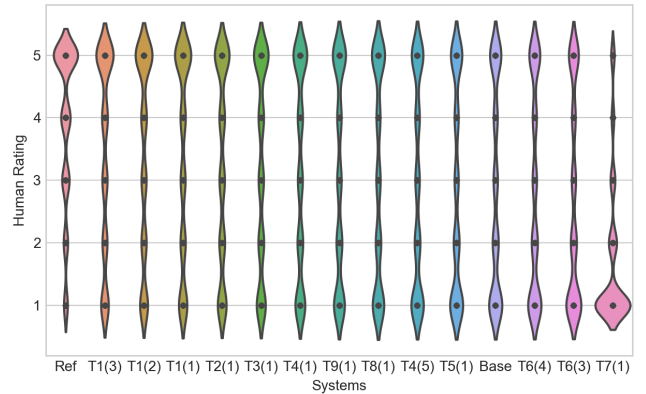


Figure 4: Distribution of human rating score for each level of scores.

end-to-end differentiable network to build scene-aware dialog systems. The best system using "Hierarchical Attention mechanisms to combine text and vision" improved by 22% of the human rating score from the baseline system. The language models trained from the QA are still strong and the

power to capture ques to predict the objects and events in the video is not sufficient to answer the questions accurately. Future work includes more detailed analysis of the correlation between the QA text and the video scenes.

References

- Alamri, H.; Hori, C.; Marks, T. K.; Batra, D.; and Parikh, D. 2018. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 at AAAI2019 Workshop*.
- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.
- Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.
- Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2016. Visual dialog. *CoRR* abs/1611.08669.
- Das, A.; Kottur, S.; Moura, J. M.; Lee, S.; and Batra, D. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *International Conference on Computer Vision (ICCV)*.
- Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hershey, S.; Chaudhuri, S.; Ellis, D. P. W.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; Slaney, M.; Weiss, R. J.; and Wilson, K. 2017. CNN architectures for large-scale audio classification. In *ICASSP*.
- Hori, C.; Hori, T.; Lee, T.-Y.; Zhang, Z.; Harsham, B.; Hershey, J. R.; Marks, T. K.; and Sumi, K. 2017. Attention-based multimodal fusion for video description. In *ICCV*.
- Hori, C.; Alamri, H.; Wang, J.; Winchern, G.; Hori, T.; Cherian, A.; Marks, T. K.; Cartillier, V.; Lopes, R. G.; Das, A.; et al. 2018. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. *arXiv preprint arXiv:1806.08409*.
- Hori, C.; Perez, J.; Higashinaka, R.; Hori, T.; Boureau, Y.-L.; Inaba, M.; Tsunomori, Y.; Takahashi, T.; Yoshino, K.; and Kim, S. to appear in 2018. Overview of the sixth dialog system technology challenge: DSTC6. *Computer Speech and Language* Special issue on DSTC6.
- Kumar, S. H.; Okur, E.; Sahay, S.; Leanos, J. J. A.; Huang, J.; and Nachman, L. 2019. Context, attention and audio feature explorations for audio visual scene-aware dialogue. In *DSTC7 at AAAI2019 workshop*.
- Le, H.; Hoi, S.; Sahoo, D.; and Chen, N. 2019. End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In *DSTC7 at AAAI2019 workshop*.
- Lin, K.-Y.; Hsu, C.-C.; Chen, Y.-N.; and Ku, L.-W. 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. In *DSTC7 at AAAI2019 workshop*.
- Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.
- McTear, M. F. 2002. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)* 34(1):90–169.
- Nguyen, D.; Sharma, S.; Schulz, H.; and Asri, L. E. 2019. From film to video: Multi-turn question answering with multi-modal context. In *DSTC7 at AAAI2019 workshop*.
- Pasunuru, R. R., and Bansal, M. 2019. Dstc7-avsd: Scene-aware video-dialogue systems with dual attention. In *DSTC7 at AAAI2019 workshop*.
- Sanabria, R.; Palaskar, S.; and Metze, F. 2019. Cmu sinbad submission for the dstc7 avsd challenge. In *DSTC7 at AAAI2019 workshop*.
- Sigurdsson, G. A.; Varol, G.; Wang, X.; Laptev, I.; Farhadi, A.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. *ArXiv*.
- Vinyals, O., and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.
- Yeh, Y.-T.; Lin, T.-C.; Cheng, H.-H.; Deng, Y.-H.; Su, S.-Y.; and Chen, Y.-N. 2019. Reactive multi-stage feature fusion for multimodal dialogue modeling. In *DSTC7 at AAAI2019 workshop*.
- Young, S. J. 2000. Probabilistic methods in spoken-dialogue systems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 358(1769):1389–1402.
- Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2016. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhuang, B.; Wang, W.; and Shinozaki, T. 2019. Investigation of attention-based multimodal fusion and maximum mutual information objective for dstc7 track3. In *DSTC7 at AAAI2019 workshop*.
- Zue, V.; Seneff, S.; Glass, J. R.; Polifroni, J.; Pao, C.; Hazen, T. J.; and Hetherington, L. 2000. Juplter: a telephone-based conversational interface for weather information. *IEEE Transactions on speech and audio processing* 8(1):85–96.