MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Overview of Audio Visual Scene-Aware Dialog with Reasoning Track for Natural Language Generation in DSTC10

Hori, Chiori; Shah, Ankit Parag; Geng, Shijie; Gao, Peng; Cherian, Anoop; Hori, Takaaki; Le Roux, Jonathan; Marks, Tim K.

TR2022-016 March 03, 2022

Abstract

The Audio-Visual Scene-Aware Dialog (AVSD) task was proposed in the Dialog System Technology Challenge (DSTC), where an AVSD dataset was collected and AVSD technologies were developed. An AVSD challenge track was hosted at both the 7th and 8th DSTCs (DSTC7, DSTC8). In these challenges, the best-performing systems relied heavily on human-generated descriptions of the video content, which were available in the datasets but would be unavailable in real-world applications. To promote further advancements for real-world applications, a third AVSD challenge is proposed, at DSTC10, with two modifications: 1) the human-created description is unavailable at inference time, and 2) systems must demonstrate temporal reasoning by finding evidence from the video to support each answer. This paper introduces the new task that includes temporal reasoning and the new extension of the AVSD dataset for DSTC10, for which humangenerated temporal reasoning data were collected. A baseline system was built using an AV-transformer and the new datasets were released for the challenge. Finally, this paper reports the challenge results of 12 systems submitted to the AVSD task in DSTC10. The two systems using GPT-2 based multimodal transformer have achieved the best performance for human rating, BLEU4 and CIDEr. The temporal reasoning performed by those systems has outperformed the baseline method with temporal attention.

The 10th Dialog System Technology Challenge Workshop at AAAI 2022

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Overview of Audio Visual Scene-Aware Dialog with Reasoning Track for Natural Language Generation in DSTC10

Chiori Hori[†], Ankit P. Shah[†§], Shijie Geng[‡], Peng Gao[⋆], Anoop Cherian[†], Takaaki Hori[†], Jonathan Le Roux[†], Tim K. Marks[†]

[†]Mitsubishi Electric Research Laboratories (MERL)
[§]Carnegie Mellon University [‡]Rutgers University [⋆]The Chinese University of Hong Kong

Abstract

The Audio-Visual Scene-Aware Dialog (AVSD) task was proposed in the Dialog System Technology Challenge (DSTC), where an AVSD dataset was collected and AVSD technologies were developed. An AVSD challenge track was hosted at both the 7th and 8th DSTCs (DSTC7, DSTC8). In these challenges, the best-performing systems relied heavily on human-generated descriptions of the video content, which were available in the datasets but would be unavailable in real-world applications. To promote further advancements for real-world applications, a third AVSD challenge is proposed, at DSTC10, with two modifications: 1) the human-created description is unavailable at inference time, and 2) systems must demonstrate temporal reasoning by finding evidence from the video to support each answer. This paper introduces the new task that includes temporal reasoning and the new extension of the AVSD dataset for DSTC10, for which human-generated temporal reasoning data were collected. A baseline system was built using an AV-transformer and the new datasets were released for the challenge. Finally, this paper reports the challenge results of 12 systems submitted to the AVSD task in DSTC10. The two systems using GPT-2 based multimodal transformer have achieved the best performance for human rating, BLEU4 and CIDEr. The temporal reasoning performed by those systems has outperformed the baseline method with temporal attention.

Introduction

Recent artificial intelligence (AI) research activities have accelerated the development of technologies required for advanced human-like capabilities in machines, such as robots. For instance, current computer vision technologies can accurately perceive visual scenes, and spoken dialog systems can transcribe speech and understand speakers' intention. However, one important piece of technology is missing: natural and context-aware human-machine interaction, where machines understand their surrounding scene from the human perspective, and they are able to share their understanding with humans using natural language. Currently, there are no mechanisms for machines to have a conversation about a surrounding event or experience with humans using natural language.

To invent machines that can communicate with humans about objects and events in surrounding scenes, the project to work on Audio-visual Scene-aware Dialog (AVSD) was kicked-off in the track proposal for DSTC7 at DSTC6 in 2017 (Hori et al. 2019c). 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. The goal of AVSD in DSTC is to have question-answering based conversations on videos from daily life. To this end, the AVSD challenge task was designed based on the popular Charades dataset (Sigurdsson et al. 2016), with the goals: (1) generate answers to questions about objects and events in the video clips and (2) hold a meaningful dialog with humans about objects and events using natural, conversational language in an end-to-end framework.

To advance research into multimodal reasoning-based dialog generation, the AVSD dataset was collected and the AVSD challenge was held in DSTC7 (Alamri et al. 2019; Hori et al. 2019a). The DSTC7 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) against the baseline model. This large margin suggested that there was perhaps more potential in store for advancing this new research area. Towards this end, a second edition of our AVSD challenge was held in DSTC8 (Kim et al. 2021). In DSTC8, the results (averaged across the test set) for each team's entries were evaluated using both word-overlap-based objective measures and subjective human ratings. Although the language-based transformer models such as BERT (Devlin et al. 2019) and GPT-2 (Radford et al. 2019) demonstrated state-of-the-art performances on the DSTC tasks, these systems required features extracted from manually-generated video descriptions (captions and scripts provided with the dataset). However, such text modalities may be unavailable in real-world deployment scenarios. There are two other design difficulties that such text-based descriptions introduce that may skew the evaluation: (i) some descriptions already include parts of the answers that are used in the evaluations, making audio-visual inference redundant, and (ii) language models trained using a simple (and limited) QA dataset may generate answers using frequently-occurring text patterns in the training data, without needing to use audio-visual cues (e.g., Q:

How many people are in the scene? A: Two people). These observations are empirically supported by the results: without providing human-generated descriptions, the best performing model achieves only 0.387 in BLEU score, which is a relative reduction of 12% from its score when using human descriptions. This result suggests that there is still opportunity to design better audio-visual reasoning approaches that can match the performance achieved when using manual video descriptions.

To promote further advancements into real-world applications of the AVSD setup, a third challenge was proposed in DSTC10, progressively improving the challenge from the previous video-based scene-aware dialog tracks. The new task is to generate sentences for a system response to a query that occurs during a dialog about a video using reasoning features without using the human-created video description. Participants used the video, audio, and dialog text data to train end-to-end models without the manual descriptions. This challenge used the AVSD datasets that were collected and used in the previous challenges. The additional datasets for temporal reasoning for QA datasets were collected and used in DSTC10.

Audio-Visual Scene-Aware Dialog data set

The AVSD in DSTC10, the same AVSD data collected by (Alamri et al. 2019) have been used. Table 1 shows the size of the data used for DSTC10. For this year's AVSD challenge, additional data for temporal reasoning were collected, in which humans watched the videos and read the dialogues, then identified segments of the video containing evidence to support each given answer. Figure 1 shows the annotation tool for reasoning. With this tool, humans identified temporal segments based on visual evidence and/or audio evidence and filled in the appropriate fields with begin and end timestamps to provide temporal reasoning.

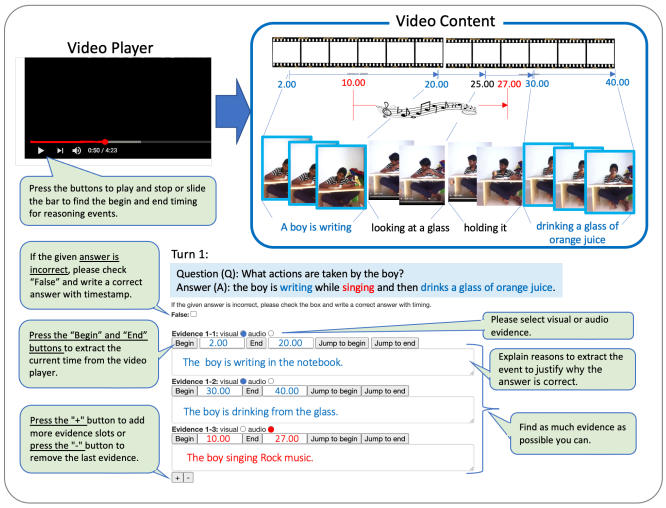


Figure 1: Temporal reasoning data collection tool for AVSD in DSTC10.

Table 1: Audio-Visual Scene-aware Dialog dataset for DSTC10.

	training	validation	test
#dialogs	7,659	1,787	1,804
#turns	153,180	35,740	28,406
#words	1,450,754	339,006	272,606

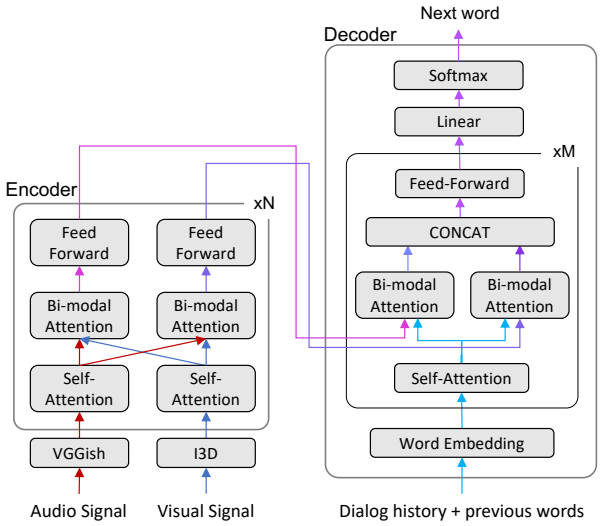


Figure 2: Baseline AV-transformer used for AVSD in DSTC10.

Baseline Model

A baseline system has been built for the DSTC10 AVSD track, which utilizes an AV-transformer architecture (Iashin and Rahtu 2020) shown in Fig. 2. The system employs a transformer-based encoder-decoder, including a bimodal attention mechanism (Bahdanau, Cho, and Bengio 2014; Chorowski et al. 2015) that lets it learn interdependencies between audio and visual features.

Given a video stream, the audio-visual encoder extracts VGGish (Hershey et al. 2017) and I3D (Carreira and Zisserman 2017) features from the audio and video tracks, respectively, and encodes these using self-attention, bimodal attention, and feed-forward layers. Typically, this encoder block is repeated N times, e.g., $N \geq 6$. The self-attention layer extracts temporal dependency within each modality, and the bimodal attention layers further extract cross-modal dependency between audio and visual features, taking the keys and values from the other modality. After that, the feed-forward layers are applied in a point-wise manner, and the encoded representations for audio and visual features are obtained.

The decoder receives the encoder outputs and the dialog history until the current question, and starts generating the answer sentence from the beginning token (<sos>) placed at the end of the last question. At each iteration step, it receives the preceding word sequence and predicts the next word by applying M decoder blocks and a prediction network.In each decoder block, the encoded audio-visual features are combined with each word using the bimodal atten-

tion layers.

The self-attention layer converts the word vectors to high-level representations considering their temporal dependency. The bimodal source attention layers update the word representations based on the relevance to the encoded multimodal representations. A feed-forward layer is then applied to the outputs of the bimodal attention layers. Finally, a linear transform and softmax operation are applied to the output of the M-th decoder block to obtain the probability distribution of the next word. The answer sentence is extended by adding the word with the highest probability to the already generated word sequence by a greedy search process that ends if the word is end token (<eos>). It is also possible to pick multiple words with highest probabilities and consider multiple candidates for the answer sentence using the beam search.

Temporal Reasoning

Temporal reasoning is the task of finding evidence supporting the generated answers, where the evidence corresponds to human-annotated time regions of the video that have been identified as supporting each ground-truth answer. Human annotators were allowed to choose multiple time regions for each question-answer pair, but most of the reasons consist of a single region.

The baseline system performs temporal reasoning based on attention weights obtained during decoding to generate the answer. The attention weights are computed to predict each word, where each attention weight corresponds to a certain time frame of input audio/visual features. Thus, a high weight means that the corresponding time frame is strongly correlated to a word in the generated answer. Given an attention weight distribution, mean μ and standard deviation σ of the attention distribution can be computed, and the time region can be estimated as $\mu \pm \nu\sigma$, where ν is a hyperparameter. Since there exist multiple attention distributions over the word sequence, attention heads, and layers, their averaged distribution is used. This method finds only one time region for each answer, and it requires no special training to select time regions.

Submitted Systems

The AVSD Task received 12 system submissions from 5 teams. This section summarizes the techniques used in the submitted systems to the AVSD challenge, including the baseline system. The individual system description papers contain more details about the systems of Team 1, Team 2, Team 3, Team 4, Team 5.

Table 2 lists the baseline and submitted systems with brief specifications including the encoder-decoder model type, multimodal fusion type, audio-visual video features used, and additional techniques or data sets. Whereas DSTC7 systems mainly employed LSTM or GRU-based models, in DSTC8 most systems (Teams 1–6) employed a transformer-based architecture (Vaswani et al. 2017) for the encoder-decoder model. These transformer-based systems outperformed the systems based on LSTMs and GRUs. In DSTC10, most systems have also employed transformers. For multimodal fusion, some systems (Teams 1 and 2) utilized attentional fusion of multimodal features through the cross attention mechanism. Some other systems (Teams 3 and 4) utilized a GPT-2 model, which was fine-tuned to take serialized multimodal sequences. The last system (Team 5) used UniVL (a unified video and language pre-training model for multimodal understanding and generation) to fuse multimodal information. For feature extraction, Teams 1 and 2 used I3D and VGGish provided by the challenge organizer. Team 3 introduced action recognition and acoustic event detection to utilize predicted labels as audio-visual information instead of using the audio-visual features, which is a distinctive approach compared to other systems. Team 4 employed TimeSformer to extract visual features. Team 5 also extracted object features using a D2Det object detector.

In addition, Teams 1 and 2 applied student-teacher learning (STL) with caption/summary text to perform AVSD without caption/summary at inference stage (Hori et al. 2019b). Furthermore, Team 2 combined LSTM and Transformer by linear combination of word probabilities, where both models were trained by STL.

For temporal reasoning, Team 3 exploited a publicly available 2D-TAN network to obtain the timestamp results.

Evaluation

In this challenge, the quality of a system's automatically generated sentences is evaluated using objective measures to determine the level of similarity between the system-generated responses and ground-truth responses provided by humans. For this purpose, we needed to collect more human-generated responses to each test question (the original dialog, of course, contains only a single human response to each question). To collect more possible human answers in response to the test question for each test video, we asked 5 humans to watch the video, read a dialogue (up to the test question) about the video between a questioner and an answerer, and then provide an answer in response to the test question.

To evaluate the systems' automatically generated answers, we compared them with 6 ground-truth human answers, which consisted of the one original answer and these 5 newly collected answers. We used the MSCOCO evaluation tool for objective evaluation of system outputs[1]. The supported metrics include metrics based on word overlap, such as BLEU, METEOR, ROUGE_L, and CIDEr.

In addition, we collected human ratings for each system response using a 5-point Likert scale, in which humans rated system responses given a dialog context as follows: 5 for very good, 4 for good, 3 for acceptable, 2 for poor, 1 for very poor. we asked the human raters to consider correctness of the answers as well as naturalness, informativeness, and appropriateness of the response according to the given context.

The reasoning performance was measured by Intersection over Union (IoU), which indicates the ratio of overlap between the predicted and ground-truth time regions (higher is better). Since there may be multiple valid reasons for each

[1] https://github.com/tylin/coco-caption

Table 2: Submitted systems to the DSTC10-AVSD track.

Team	Encoder-decoder type	Multimodal fusion type	Features	Additional techniques/data
Baseline	Transformer	Audio-visual bi-modal attention	I3D, VGGish, QA history	
Team 1: (No paper)	Transformer	(1) Triple cross attention (2) (1) + Attentional fusion	I3D, VGGish, QA history	Student-teacher learning (STL)
Team 2: (Shah et al. 2022)	LSTM & Transformer	Attentional multimodal fusion	I3D, VGGish, last question	(1) Linear comb. of LSTM & Transformer + STL (2) (1) + Cross student-teacher loss
Team 3: (Heo 2022)	Transformer	Input multimodal labels and text to GPT-2	Action/event labels, QA history	GPT-2 with Action/Event labels by Video Swim/Audio Spectrum Transformers. Reasoning with 2D-TAN network. (1) No Caption, (2) Use Caption, (3) No Caption / No Audio, (4) Use Caption (FRONT)
Team 4: (Yamazaki et al. 2022)	Transformer	Input video features and text to GPT-2	TimeSformer video feature, QA history	GPT-2 + TimeSformer video features with (1) fixed 32 frames or (2) variable-length input
Team 5: (Huang et al. 2022)	(1) Attention-based encoder decoder (2) UniVL model	(1) Concatenation of encoder states from different modalities (2) Cross-attention in early fusion of multimodal features	(1) S3D, VGGish, BERT-encoded QA history (2) (1) + object feature	(2) UniVL model + D2Det object detector

answer, two IoU measures have been designed, where IoU-1 is obtained as an average IoU computed between each ground truth and the predicted region that gives the highest IoU to the ground truth. IoU-2 is computed by frame-level matching among all predicted and ground-truth regions for each answer, i.e., frames included in both predicted and ground-truth regions are counted as intersections while those included in both or either of them are counted as union.

Table 3 reports the numerical results of all qualifying submitted systems (entries) from all teams. The subjective human ratings described above are given in the rightmost column of the table, and the others are the objective scores that were computed using word-overlap metrics (Bleu, METEOR, ROUGE_L, and CIDEr) and reasoning metrics (IoU-1 and IoU-2). Team 2 shows higher objective scores in most metrics (Bleu-1...3, METEOR, ROUGE_L), while Teams 4 and 5 achieve highest human rating scores 3.567 and 3.569, where the difference is negligible. Regarding the reasoning result, Team 3 show highest IoU scores 0.516 and 0.544.

Figure 3 plots the human ratings for each system in several ways. In all three figures, the systems are shown in the same order on the x-axis. Figure 3a plots the mean and standard deviation of the human ratings for each system (across all responses and all raters for that system). Figure 3b shows the distribution over the sentences in the test set of the mean human rating score for each sentence. Figure 3c shows each system's distribution over rating scores $(1, \ldots, 5)$ across all sentences and all raters. In this figure, the area of the violinplot for each score indicates the number of scores at each level on the Likert scale. It may be observed from this figure that the distribution of human rating scores across all systems appears to be bimodal—most answers are rated either highly or poorly, with few examples in the middle. This is because the human ratings of each answer depends strongly on whether the answer is a correct response to the question: correct answers generally receive high ratings, but incorrect answers receive low human ratings. The best systems generated mostly correct answers, while the worst systems generated more incorrect answers.

The Reference system (labeled "Ref" at the far left of each figure) shows the ratings for the ground truth answers extracted from the original dialogs of the AVSD dataset. The Reference system had the best human ratings: it had the highest mean rating in Figure 3a, the highest median sentence rating in Figure 3b and the most sentences rated as level 5 ("very good") in Figure 3c. The worst system (at the right), which was the baseline system, had a much lower mean rating and a long tail of poorly rated sentences.

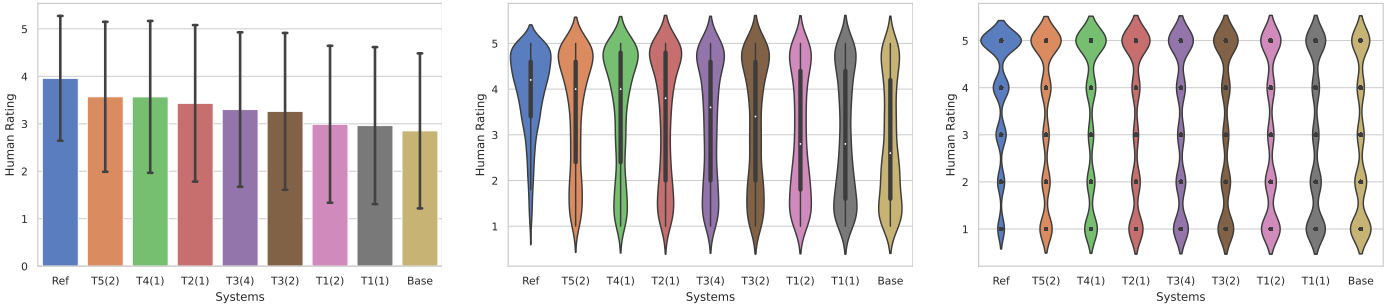
Figures 4 and 5 show examples of temporal reasoning obtained by the baseline and the best reasoning system (Team 3(2)) in comparison with the ground truth. These examples clearly show that the best system provided much better reasoning than the attention method of the baseline system.

What We Learned from DSTC10

We now discuss what we learned from the AVSD challenges in DSTC10. Most of the DSTC10 systems employed transformers, rather than recurrent networks using LSTM or GRU. The inclusion of transformers drastically improved performance the AVSD task from DSTC7 to DSTC8, similar to the improvements that have been observed in other applications such as machine translation and speech recognition. Two of the most successful systems extracted semantic features of the word sequences by initializing network weights using a pre-trained model such as BERT and GPT-2, then fine-tuning the weights on the AVSD dataset. Furthermore, in DSTC10, different pre-trained models were utilized in different ways to extract features, generate relevant labels, and enhance temporal reasoning, which substantially improved the AVSD performance. This is probably because publicly available code and models have increased and become more powerful and versatile for a wide variety of audio-visual tasks.

Table 3: DSTC10-AVSD evaluation results with word-overlap-based objective measures based on 6 references, a subjective measure based on 5-level ratings by humans (HR), and reasoning performance measures based on Intersection-over-Union (IoU).

System	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L	CIDEr	IoU-1	IoU-2	HR
Baseline	0.572	0.422	0.320	0.247	0.191	0.439	0.566	0.361	0.380	2.851
Team 1 (1)	0.601	0.451	0.347	0.270	0.196	0.456	0.607	0.360	0.378	2.962
Team 1 (2)	0.598	0.449	0.345	0.270	0.198	0.458	0.613	0.362	0.380	2.990
Team 2 (1)	**0.695**	**0.564**	**0.462**	0.381	**0.248**	**0.540**	0.888	-	-	3.431
Team 2 (2)	0.692	0.563	**0.462**	0.381	0.246	0.537	0.880	-	-	-
Team 3 (1)	0.641	0.489	0.379	0.298	0.225	0.502	0.804	0.506	0.534	-
Team 3 (2)	0.624	0.475	0.366	0.286	0.231	0.503	0.786	**0.516**	**0.544**	3.262
Team 3 (3)	0.651	0.490	0.376	0.295	0.227	0.502	0.789	0.505	0.533	-
Team 3 (4)	0.646	0.489	0.380	0.299	0.225	0.499	0.787	0.505	0.533	3.300
Team 4 (1)	0.680	0.558	0.461	**0.385**	0.247	0.539	**0.957**	-	-	3.567
Team 4 (2)	0.679	0.554	0.456	0.379	0.246	0.536	0.945	-	-	-
Team 5 (1)	0.670	0.541	0.441	0.365	0.241	0.526	0.906	0.485	0.510	-
Team 5 (2)	0.673	0.545	0.448	0.372	0.243	0.530	0.912	0.479	0.505	**3.569**
Reference	-	-	-	-	-	-	-	-	-	3.958



(a) Mean and standard deviation.

(b) Distribution over sentence scores.

(c) Distribution over score values.

Figure 3: Statistics of human rating scores.

Conclusion

This paper described the DSTC10-AVSD challenge task, the baseline system, and the results provided by submitted systems. In the previous AVSD challenges, DSTC7 and DSTC8, the best-performing systems relied heavily on human-generated descriptions of the video content, which were available in the datasets but would be unavailable in real-world applications. The third AVSD challenge promoted further advancements for real-world applications, where 1) the human-created description is unavailable at inference time, and 2) systems must demonstrate temporal reasoning by finding evidence from the video to support each answer. The submitted systems provided high-quality answers and reasoning even without human-generated descriptions at inference time. The DSTC10 winning system achieved 90.2% of the human performance based on human ratings. The result is considerable, but the gap with human performance is actually larger than the DSTC8 result (98.4%). This is obviously because the real-world condition made the AVSD task more difficult. This shows that continued research is still needed to achieve human performance. After the workshop, data setup, baseline system, and evaluation tools will be released, which will facilitate continuous improvement by the community in the future.

References

Alamri, H.; Cartillier, V.; Das, A.; Wang, J.; Cherian, A.; Essa, I.; Batra, D.; Marks, T. K.; Hori, C.; Anderson, P.; et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7558–7567.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proc. CVPR*.

Chorowski, J. K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; and Bengio, Y. 2015. Attention-Based Models for Speech Recognition. In Cortes, C.; Lawrence, N. D.; Lee, D. D.;

(1) Reasoning for the entire video: Answers need to be generated based on the information in all frames.

In [ ]:
```
view('N04QT')
```

```
Q1: is this a male ?
A1: as far as i can tell , yes , the person is male .
Q2: what is he doing ?
A2: he 's watching the movie on the tv .
Q3: does he make any sounds ?
A3: he doesn 't , but the sound from the movie is playing .
Q4: does he seem to be enjoying the movie ?
A4: it 's hard to tell . he doesn 't have much of an expression .
Q5: does he do anything else but watch ?
A5: he briefly picks up a blue cup on the table , then places it back dow .
Q6: does he drink from it ?
A6: no , he never drinks from it .
Q7: is there anyone else in the room ?
A7: no , he is alone in the room .
Q8: are there any pets ?
---------------
(1) Groundtruth - A8: no pets are visible in the video .
(2) Baseline    - A8: no , there are no pets in the room .
(3) Team_3(2)   - A8: no, there are no pets in the room.
```

(2) Reasoning for how the video ends

In [ ]:
```
view('RAHFS')
```

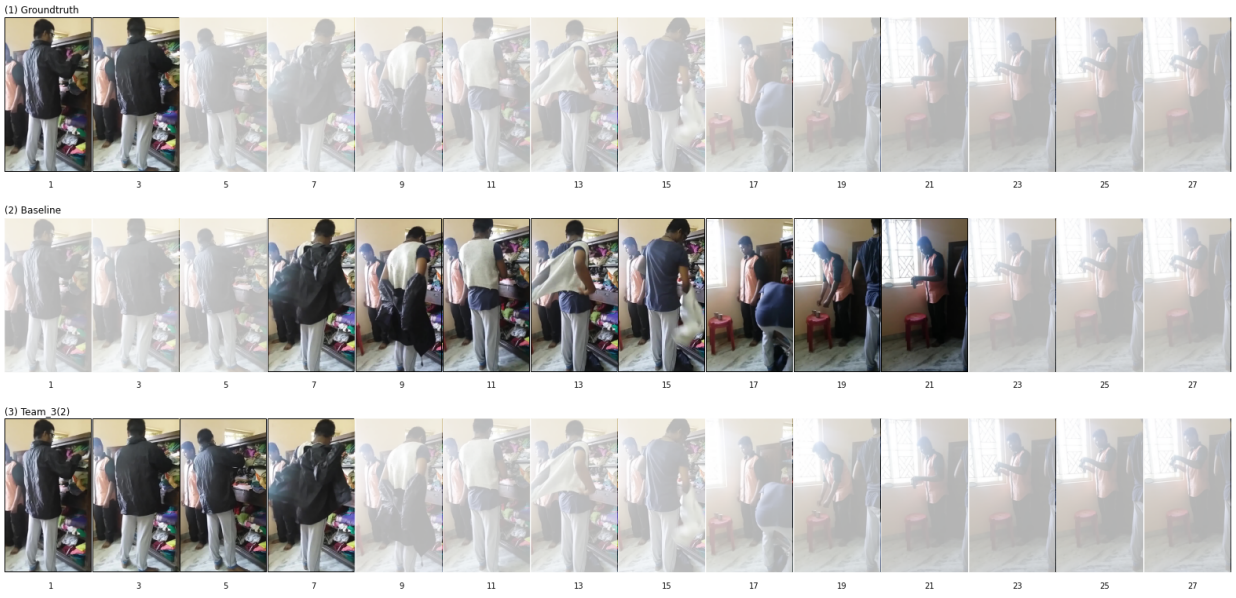
```
Q1: how many people are in the video ?
A1: just one person is in the video .
Q2: is there sound ?
A2: yes there is sound in the video .
Q3: what is he doing when the video starts ?
A3: in the beginning of the video he walks into the frame .
Q4: does he talk or do you hear background noise ?
A4: he doesn 't talk , the only noise you hear is of the plates clattering .
Q5: what does he do once he walks into the frame ?
A5: he grabs a plate out of the pantry , inspects it , and then tosses it back into the pantry .
Q6: does he show any obvious emotion ?
A6: no , the man does not show any obvious emotion .
Q7: what does he do after he tosses the plate back ?
A7: he repeats the same action with a second plate , pulling it out and then tossing it back in .
Q8: does he leave the room or is he visible still at the end of the video ?
A8: he 's still visible at the end of the video . he does something else though , before the video ends .
Q9: what does he do after he tosses the second plate back ?
A9: after he tosses the second plate back he grabs a clear container , which one might assume is alcohol , takes a long s
wig , and then places it back into the pantry .
Q10: does the video end after that ?
---------------
(1) Groundtruth - A10: yes it does end after that .
(2) Baseline    - A10: no , he just stands there the plate , he just stands there .
(3) Team_3(2)   - A10: yes, the video ends with him still in the pantry.
```

Figure 4: Example of reasoning results (1/2)

(3) Reasoning for scenes and actions

In [ ]:
```
view('JPTQC')
```

```
Q1: how many people are in the video ?
A1: there are two people in the video .
Q2: what room are they in ?
A2: they look to be in a living room or bedroom .
Q3: are they both visible when the video begins ?
---------------
(1) Groundtruth - A3: yes they are both there when the video starts in the beginning .
(2) Baseline    - A3: yes , they are in the living room .
(3) Team_3(2)   - A3: yes, they are both visible when the video begins.
```



In [ ]:
```
view('UKVJ6')
```

```
Q1: is this in a kitchen ?
A1: yes , a very small one .
Q2: ok , is the man holding a paper in his hand ?
A2: yes , several papers that he looks like he is reading .
Q3: and then he sets them on a counter ?
A3: yes , he sets the down by the sink .
Q4: and then does he go kneel down by a cupboard ?
A4: yes , he kneels down and looks at the cabinet door .
Q5: does he open the cabinet ?
A5: yes , he opens it and looks at it .
Q6: does he take anything out ?
A6: no , he is reaching inside and fiddling with something .
Q7: does he close it when he is done ?
---------------
(1) Groundtruth - A7: yes , he closes it as soon as he is done .
(2) Baseline    - A7: no , he doesn 't close the cabinet door .
(3) Team_3(2)   - A7: yes, he closes the cabinet door.
```

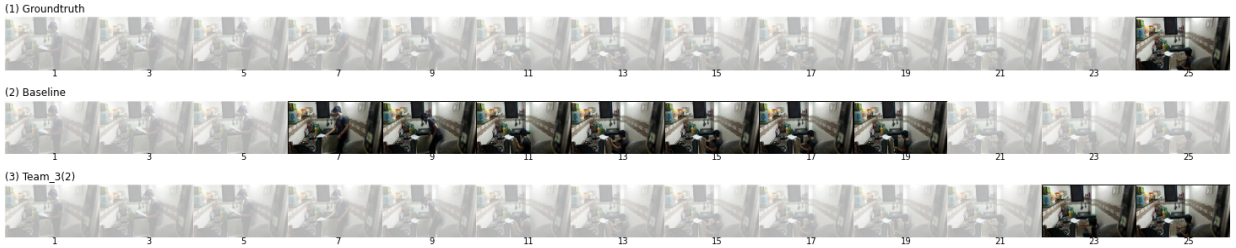


Figure 5: Example of reasoning results (2/2)

Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28*, 577–585. Curran Associates, Inc.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Heo, Y. 2022. Interpretable Multimodal Dialogue System with Natural Language-Based Multimodal Integration. In *Proceedings of DSTC10 Workshop at AAAI-2022*.

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 *Proc. ICASSP*.

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

Hori, C.; Cherian, A.; Marks, T. K.; and Hori, T. 2019b. Joint Student-Teacher Learning for Audio-Visual Scene-Aware Dialog.

Hori, C.; Perez, J.; Higashinaka, R.; Hori, T.; Boureau, Y.-L.; Inaba, M.; Tsunomori, Y.; Takahashi, T.; Yoshino, K.; and Kim, S. 2019c. Overview of the sixth dialog system technology challenge: DSTC6. *Computer Speech & Language*, 55: 1–25.

Huang, X.; Tan, H. L.; Leong, M. C.; Sun, Y.; Li, L.; Jiang, R.; and Kim, J. J. 2022. Investigation on Transformer-based Multi-modal Fusionfor Audio-Visual Scene-Aware Dialog. In *Proceedings of DSTC10 Workshop at AAAI-2022*.

Iashin, V.; and Rahtu, E. 2020. A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer. In *Proc. BMVC*.

Kim, S.; Galley, M.; Gunasekara, C.; Lee, S.; Atkinson, A.; Baolin, P.; Schulz, H.; Gao, J.; Li, J.; Adada, M.; et al. 2021. Overview of the Eighth Dialog System Technology Challenge: DSTC8. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Shah, A. P.; Hori, T.; Le Roux, J.; and Hori, C. 2022. DSTC10-AVSD Submission System with Reasoning using Audio-Visual Transformers with Joint Student-Teacher Learning. In *Proceedings of DSTC10 Workshop at AAAI-2022*.

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*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. NeurIPS*, 5998–6008.

Yamazaki, Y.; Orihashi, S.; Masumura, R.; Uchida, M.; and Takashima, A. 2022. Audio Visual Scene-Aware Dialog Generation withTransformer-based Video Representations. In *Proceedings of DSTC10 Workshop at AAAI-2022*.