AUDIO-VISUAL SCENE-AWARE DIALOG AND REASONING USING AUDIO-VISUAL TRANSFORMERS WITH JOINT STUDENT-TEACHER LEARNING

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

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

ABSTRACT

In previous work, we have proposed the Audio-Visual Scene-Aware Dialog (AVSD) task, collected an AVSD dataset, developed AVSD technologies, and hosted an AVSD challenge track at both the 7th and 8th Dialog System Technology Challenges (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, we proposed a third AVSD challenge, 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 our new extension of the AVSD dataset for DSTC10, for which we collected human-generated temporal reasoning data. We also introduce a baseline system built using an AV-transformer, which we released along with the new dataset. Finally, this paper introduces a new system that extends our baseline system with attentional multimodal fusion, joint student-teacher learning (JSTL), and model combination techniques, achieving state-of-the-art performances on the AVSD datasets for DSTC7, DSTC8, and DSTC10. We also propose two temporal reasoning methods for AVSD: one attention-based, and one based on a time-domain region proposal network.

Index Terms— Audio-visual scene-aware dialog, Video description, Temporal reasoning, End-to-end modeling, Audio-visual Transformer

1. INTRODUCTION

To encourage development of dialog system technologies that enable an agent to discuss audio-visual scenes with humans, we held two challenges on audio-visual Scene-Aware Dialog (AVSD) at DSTC7 and DSTC8 [1, 2] using a dataset we collected based on the videos from the Charades dataset [3]. The AVSD task we defined and dataset we prepared were the first attempt to promote the combination of audio-visual question-answering systems and conversation systems into a single framework [4, 5]. This task that we proposed is to generate a system response to a query, where the query is part of a multi-turn dialog about a video. Challenge participants used the video, its associated audio, and the dialog text to train end-to-end deep learning models to produce the answers. In addition, the systems had access to human-created video captions. The AVSD task can be seen as an extension to video data of both the *visual question answering* (VQA) task [6–9], in which the goal is to generate answers to questions about a scene in a static image, and the *visual dialog* task [10], in which an AI agent holds a meaningful dialog with humans about a static image using natural, conversational language [11]. While, the conventional Video Question Answering tasks [12] select phrase or sentence for frame-filling one-shot QA based on question types using video captions or synopsis.

Another progenitor to AVSD is the task of *video description* (text summarization of videos), which [13] addressed utilizing multimodal attention mechanisms, which selectively attend to different input modalities (feature types) such as spatiotemporal motion features and audio features, in addition to temporal attention. Combining video description technologies like these with end-to-end dialog systems enables scene-aware dialog systems that make use of multimodal information, such as audio and visual features. In a more recent work, spatio-temporal reasoning has been shown to improve performance on AVSD tasks [14]. Recently, Transformer-based AVSD systems outperform LSTM-based ones [15, 16].

The task setup for AVSD in DSTC7–8 allowed participants to use human-created video captions to help generate answers for the dialog questions, and systems that used these human-generated captions significantly outperformed systems that did not. However, since such human-created descriptions are not available in real-world applications of an AVSD system, in practice a system needs to learn to produce the answers without the captions. 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). The results from AVSD in DSTC7–8 suggest there is still an opportunity to design better audio-visual reasoning methods to approach the performance achieved when using manual video descriptions, but without using these descriptions at test time. Furthermore, real systems should ideally be able to show the evidence supporting their generated answers, by pointing to the relevant segments of the video. To encourage progress towards this end, we propose a third AVSD challenge in DSTC10[1].

In this paper, we introduce the DSTC10-AVSD challenge task, the goals of which are: 1) answer generation without human-created captions at inference time, and 2) temporal reasoning (providing evidence) for the generated answers. Furthermore, we develop an AVSD baseline system using an AV-transformer [17]. In addition, we propose a novel system that extends this AV-transformer using attentional multimodal fusion [13], joint student-teacher learning

[1] https://github.com/dialogtekgeek/AVSD-DSTC10_Official

Table 1. Audio-Visual Scene-aware Dialog data set 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

(JSTL) [18], and model combination techniques. We also propose two temporal reasoning methods for AVSD: one attention-based, and one based on a region proposal network (RPN). Results show that our extended AV-transformer achieves state-of-the-art on DSTC 7, 8, and 10 when combined with our LSTM-based AVSD system [18].

2. AUDIO-VISUAL SCENE-AWARE DIALOG DATA SET

We base the new Audio-Visual Scene-Aware Dialog (AVSD) task for DSTC10 on the AVSD dataset from DSTC7–8 [1, 2]. For the AVSD data, we collected text-based dialogs on short videos from the popular Charades dataset [3], which consists of untrimmed and multi-action videos (each video also has an audio track) and comes with human-generated descriptions of the scene. 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 [5]. Table 1 shows the size of the data used for DSTC10. For this year's challenge (DSTC10), we collected additional data for temporal reasoning, in which humans watched the videos and read the dialogues, then identified segments of the video containing evidence to support a given answer.

3. BASELINE MODEL

Our DSTC10-AVSD baseline model is an AV-transformer architecture [17], shown in Fig. 1. The system employs a transformer-based encoder-decoder, including a bimodal attention mechanism [19, 20] that lets it learn interdependencies between audio and visual features.

Given a video stream, the audio-visual encoder extracts VGGish [21] and I3D [22] 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., $N \geq 6$. More formally, let X^A and X^V denote audio and visual signals. First, the feature extraction module extracts VGGish and I3D feature vector sequences from the input signals:

$$A^0 = \mathrm{VGGish}(X^A), \;\; V^0 = \mathrm{I3D}(X^V). \tag{1}$$

The nth encoder block computes hidden vector sequences as:

$$\bar{A}^n = A^{n-1} + \mathrm{MHA}(A^{n-1}, A^{n-1}, A^{n-1}), \tag{2}$$
$$\bar{V}^n = V^{n-1} + \mathrm{MHA}(V^{n-1}, V^{n-1}, V^{n-1}), \tag{3}$$
$$\tilde{A}^n = \bar{A}^n + \mathrm{MHA}(\bar{A}^n, \bar{V}^n, \bar{V}^n), \tag{4}$$
$$\tilde{V}^n = \bar{V}^n + \mathrm{MHA}(\bar{V}^n, \bar{A}^n, \bar{A}^n), \tag{5}$$
$$A^n = \tilde{A}^n + \mathrm{FFN}(\tilde{A}^n), \tag{6}$$
$$V^n = \tilde{V}^n + \mathrm{FFN}(\tilde{V}^n), \tag{7}$$

where MHA and FFN denote multi-head attention and feed-forward network, respectively. Layer normalization [23] is applied before every MHA and FFN layer, but it is omitted from the equations for simplicity. MHA takes three arguments: query, key, and value vector sequences [24]. The self-attention layer extracts temporal dependency within each modality, where the arguments for MHA are all the same, i.e., A^{n-1} or V^{n-1}, as in (2) and (3). The bimodal attention layers further extract cross-modal dependency between audio and visual features, taking the keys and values from the other modality as in (4) and (5). After that, the feed-forward layers are applied in a point-wise manner. The encoded representations for audio and visual features are obtained as A^N and V^N.

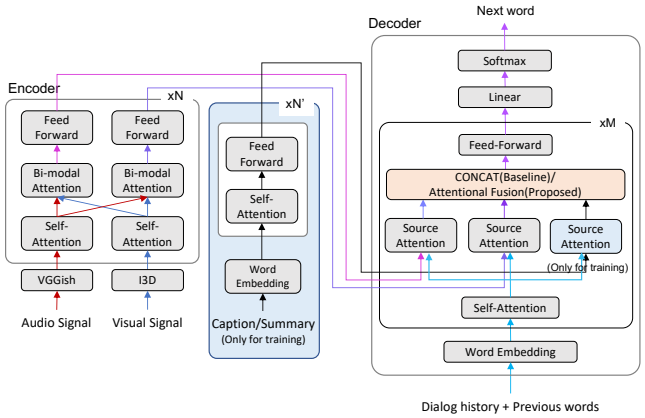

Fig. 1. Baseline and extended AV-transformer. Our extended system adds the JSTL modules (blue and orange boxes) to the baseline.

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 attention layers. Let Y_i be a dialog history plus preceding word sequence $h_1, ..., h_L, \texttt{<sos>}, y_1, ..., y_i$ after i iterations and Y_i^0 be a word embedding vector sequence given by $Y_i^0 = \mathrm{Embed}(Y_i)$.

Each decoder block has self-attention, bimodal source attention, and feed-forward layers. Computations within the m-th block are as follows:

$$\bar{Y}_i^m = Y_i^{m-1} + \mathrm{MHA}(Y_i^{m-1}, Y_i^{m-1}, Y_i^{m-1}), \tag{8}$$
$$\bar{Y}_i^{Am} = \bar{Y}_i^m + \mathrm{MHA}(\bar{Y}_i^m, A^N, A^N), \tag{9}$$
$$\bar{Y}_i^{Vm} = \bar{Y}_i^m + \mathrm{MHA}(\bar{Y}_i^m, V^N, V^N), \tag{10}$$
$$\tilde{Y}_i^m = \mathrm{Concat}(\bar{Y}_i^{Am}, \bar{Y}_i^{Vm}), \tag{11}$$
$$Y_i^m = \tilde{Y}_i^m + \mathrm{FFN}(\tilde{Y}_i^m). \tag{12}$$

The self-attention layer converts the word vectors to high-level representations considering their temporal dependency in (8). The bimodal source attention layers update the word representations based on the relevance to the encoded multi-modal representations in (9) and (10). A feed-forward layer is then applied to the outputs of the bimodal attention layers in (11) and (12). 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 as

$$P(\mathbf{y}_{i+1}|Y_i, X^A, X^V) = \mathrm{Softmax}(\mathrm{Linear}(Y_i^M)). \tag{13}$$

At inference time, we can pick the one-best word $\hat{y}_{i+1}$ for y_{i+1} as

$$\hat{y}_{i+1} = \underset{y \in \mathcal{V}}{\mathrm{argmax}} \, P(\mathbf{y}_{i+1} = y|Y_i, X^A, X^V), \tag{14}$$

where $\mathcal{V}$ denotes the vocabulary, and the answer sentence is extended by adding the selected word to the already generated word sequence as $Y_{i+1} = Y_i, \hat{y}_{i+1}$. This is a greedy search process that ends if $\hat{y}_{i+1} = \texttt{<eos>}$, which represents an end token. It is also possible to

pick multiple words with highest probabilities and consider multiple candidates for the answer sentence using the beam search.

4. EXTENDED AV-TRANSFORMER

We extend the baseline AV-transformer by applying attentional multimodal fusion [13] and joint student-teacher learning (JSTL) [18], which have successfully been applied to an LSTM-based AVSD system [4] but have not previously been applied to transformer-based systems. In this paper, we propose to extend the AV-transformer with these techniques and test their effectiveness.

Fig. 1 shows the teacher model of the extended AV-transformer, which has a caption/summary encoder in the encoder and an attentional fusion layer in the decoder. In student-teacher learning, a student model without the caption/summary encoder and its attention module in the decoder is trained using the teacher model output as the target distribution.

To further improve the performance, we combine the extended AV-transformer with the LSTM-based model trained with student-teacher learning as well, where the two decoder outputs are averaged in the log domain during the beam search.

4.1. Attentional Multimodal Fusion

The baseline AV-transformer in Fig. 1 concatenates multi-modal encoder outputs in each decoder block, assuming that the audio and visual features have equal contribution to the next word prediction regardless of the given question and the generated answer. However, prior work has shown that attentional multimodal fusion is effective for LSTM-based systems. In this work, we apply the attentional fusion technique to the AV-transformer. In the case of Transformer, we can use single-head attention (SHA) in each decoder block as

$$\tilde{\bar{Y}}_i^m = \mathrm{SHA}(\bar{Y}_i^m, \tilde{Y}_i^m, \tilde{Y}_i^m), \tag{15}$$

where $\tilde{Y}_i^m$ is here a concatenation of $\bar{Y}_i^{Am}$ and $\bar{Y}_i^{Vm}$. If the model has a caption/summary encoder, its output $\bar{Y}_i^{Cm}$ is also concatenated. In this case, $\tilde{Y}_i^m$ is a $3 \times D$ tensor including three modalities, each of which has a D-dimensional vector. Then, the fused vector $\tilde{\bar{Y}}_i^m$ is fed to the feed-forward layer.

4.2. Student-Teacher Learning

The goal of student-teacher learning is to obtain a student model that does not make use of the video caption or summary, which is trained to mimic a teacher model that has already been trained using the caption/summary text. Accordingly, the student model can be used to generate system responses without relying on the caption text, while hopefully achieving similar performance to the teacher model.

The student-teacher loss is a cross entropy loss with *soft* targets:

$$\mathcal{L}_{\mathrm{ST}} = -\sum_{i=1}^{|Y|} \sum_{y \in \mathcal{V}} \hat{P}(y|Y_{i-1}, X^A, X^V, X^C) \log P(y|Y_{i-1}, X^A, X^V), \tag{16}$$

where $\hat{P}(y|Y_{i-1}, X^A, X^V, X^C)$ denotes the probability distribution for the ith word obtained by the teacher network. Here, $P(y|Y_{i-1}, X^A, X^V)$ is the posterior distribution from the current student network (which is being trained), which is predicted without the caption text X^C.

Following our prior work, we also incorporate a decoder state similarity loss and a cross-entropy loss on the teacher for joint student-teacher learning as

$$\mathcal{L}_{\mathrm{JST}} = \mathcal{L}_{\mathrm{ST}} + \lambda_c \mathcal{L}_{\mathrm{MSE}} + \mathcal{L}_{\mathrm{CE}}^{(T)}, \tag{17}$$

where $\mathcal{L}_{\mathrm{MSE}} = \sum_{i=1}^{|Y|} \mathrm{MSE}(Y_i^m, \hat{Y}_i^m)$. Here, $\mathrm{MSE}(\cdot,\cdot)$ denotes the mean square error between two vectors, λ_c denotes a scaling factor, and $m = M/2$, i.e., we use only one intermediate layer. We aim here to compensate for missing input features at the decoder state level, so that the student model can hopefully exploits other modalities more actively. Furthermore, joint student-teacher learning updates not only the student network but also the teacher network. We use the standard cross entropy $\mathcal{L}_{\mathrm{CE}}^{(T)}$ for a *hard* target, only for the teacher network. Likewise, $\mathcal{L}_{\mathrm{ST}}$ is used only for the student network, while $\mathcal{L}_{\mathrm{MSE}}$ is used for both networks.

5. 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.

5.1. Attention-based method

We built a baseline method for temporal reasoning based on attention weights obtained during decoding. 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, we can compute mean μ and standard deviation σ of the distribution, and roughly estimate the time region as $\mu \pm \nu\sigma$, where ν is a hyper parameter. Since we have multiple attention distributions over the word sequence, attention heads, and layers, we use their averaged distribution. This method finds only one time region for each answer, and it requires no special training to select time regions.

5.2. RPN-based time region detection

We also built a CNN-based temporal reasoning model, which accepts encoder outputs of the AV-transformer and an embedded QA pair to predict temporal regions that support the answer. The model employs a time-domain region proposal network (RPN) [17, 25], where Conv1D modules with different kernel sizes accept frame-level outputs of the multimodal encoders, each of which is concatenated with the QA pair embedded by the decoder followed by mean pooling. It predicts the center position, the region length, and the confidence score of each region candidate. We pick high-confidence regions from the candidates using a predetermined threshold.

6. EXPERIMENTS

We evaluate our AV-transformer using the AVSD datasets from DSTC7, DSTC8, and DSTC10. Training and validation sets are common across the three challenges, but the test sets are different.

6.1. Conditions

We extracted VGGish audio features [21] and I3D video features [22] from each video clip, where I3D features consisted of sequences fo 2040-dimensional RGB and flow vector, and VGGish

Table 2. Evaluation results on DSTC7-AVSD test set.

Model	BLEU4	METEOR	ROUGE_L	CIDEr
Baseline AV-transformer	0.296	0.214	0.485	0.771
+ Hyperparam. tuning.	0.362	0.237	0.522	0.974
+ Beam search	0.380	0.239	0.530	0.998
+ Attentional MM fusion	0.391	0.248	0.536	1.013
+ JST learning	0.401	0.256	0.549	1.051
+ Comb. with LSTM	**0.406**	**0.262**	**0.554**	**1.079**
LSTM + JST learning [18]	0.382	0.254	0.537	1.005
DSTC7 best [27] w/ cap.*	0.394	0.267	0.563	1.094

*DSTC7 best system does not have results without captions.

Table 3. Evaluation results on DSTC8-AVSD test set.

Model	BLEU4	METEOR	ROUGE_L	CIDEr
Baseline AV-transformer	0.281	0.203	0.468	0.701
Extended AV-transformer	0.380	0.242	0.535	0.957
+ Comb. with LSTM	**0.394**	**0.250**	**0.545**	0.997
DSTC8 best [16] w/o cap.	0.387	0.249	0.544	**1.022**

features were sequences of 128-dimensional vectors. The RGB and flow features were concatenated before feeding them to the encoder.

The baseline AV-Transformer has projection layers before encoder blocks, where the audio and visual features are projected to 64 and 128 dimensional vectors, respectively. The encoder has 2 encoder blocks, in which the audio and visual attention layers have 64 and 128 dimensions, and their feed-forward layers have 256 and 512 dimensions, respectively. The decoder has 2 decoder blocks, in which 300-dimensional GloVe word vectors [26] are projected to 256-dimensional embedding vectors and fed to 256-dimensional attention layers followed by 1024-dimensional feed-forward layers. The baseline system employs greedy search to generate the answers.

We used the evaluation code for MS COCO caption generation[2] for objective evaluation of system outputs, which supports automated metrics such as BLEU, METEOR, ROUGE_L, and CIDEr.

6.2. Results and Discussion

Table 2 shows the evaluation results on the DSTC7 test set. To improve the performance from the baseline, we first tuned the hyperparameters using the validation set, where we made the decoder network deeper to 6 blocks and reduced the dimension of the attention layers to 200. We shrank the dialog history given to the decoder into just the previous question. In addition, we applied a learning rate control that halves the learning rate of Adam optimizer if the validation loss did not decrease after each training epoch. With this tuning, we obtained substantial improvement, e.g., $0.296 \rightarrow 0.332$ in BLEU4. Then, we applied the beam search technique with beam size 5, which further improved the performance.

We extend the AV-transformer by adding attentional multimodal (MM) fusion and joint student-teacher (JST) learning, achieving further performance improvement. Finally, we combine our AV-transformer with our LSTM-based model from [18], which also employed attentional MM fusion and JST learning. When we combine the word posterior probabilities of the two decoders in the log domain, we obtain the best results, which outperform the prior method [18] and even achieve competitive performance to the best DSTC7 system that used the caption/summary information.

Table 3 shows the evaluation results on the DSTC8 test set. As in the DSTC7 results, the AV-transformer including all the extensions shows substantial improvements on all the performance metrics. Furthermore, the table also shows that combination of the AV-transformer and the LSTM model achieves the state-of-the-art performance in BLEU4, METEOR, and ROUGE_L in comparison with the DSTC8 best system [16] based on a large-scale Transformer initialized with GPT-2 [28], for the condition in which caption/summary information were not available.

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

Table 4. Evaluation results on DSTC10-AVSD test set.

Model	BLEU4	METEOR	ROUGE_L	CIDEr
Baseline AV-transformer	0.247	0.191	0.437	0.566
Extended AV-transformer	0.371	0.245	0.535	0.869
+ Comb. with LSTM	**0.385**	**0.247**	**0.539**	**0.888**

Table 5. Evaluation results on temporal reasoning for DSTC10-AVSD test set.

Model	IoU-1	IoU-2
Attention method	0.361	0.380
Region Proposal Net (RPN)	**0.521**	**0.550**

Finally, we evaluated our model with the DSTC10-AVSD test set. The sentence generation performance is shown in Table 4, and we see improvements similar to the ones in the DSTC7 and DSTC8 results. We also evaluated the reasoning performance of the attention-based and RPN-based methods introduced in Section 5. The RPN had 3-layer Conv1D modules with 10 different kernel sizes for each modality and 256 dimensions in each internal layer. Table 5 shows the reasoning performance 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 answer, we designed two IoU measures, 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 5 shows that the RPN outperforms the naive attention-based approach, which suggests that model training with ground-truth annotations for temporal reasoning is important for temporal reasoning in the AVSD task [3].

7. CONCLUSIONS

In this paper, we introduced the DSTC10-AVSD task and dataset, which promote further advancements into real-world applications of the AVSD, in which human-created descriptions are not available at inference time and where temporal reasoning is required to provide evidence supporting the answers. We developed an AV-transformer as a baseline system for DSTC10-AVSD. We also proposed extending it with attentional multimodal fusion, joint student-teacher learning, and model combination techniques, achieving state-of-the-art performance. Our experiments compared the performance of the baseline system and our extended system with the previous state of the art, testing on the AVSD test sets for DSTC7, DSTC8, and DSTC10. We have just released the temporal reasoning dataset and the baseline system for open competition as the AVSD challenge in DSTC10.

[3] https://github.com/dialogtekgeek/AVSD-DSTC10_Official/tree/main/baseline

8. REFERENCES

[1] L. F. D'Haro, K. Yoshino, C. Hori, T. K. Marks, L. Polymenakos, J. K. Kummerfeld, M. Galley, and X. Gao, "Overview of the seventh dialog system technology challenge: DSTC7," *Comput. Speech Lang.*, vol. 62, p. 101068, 2020.

[2] S. Kim, M. Galley, C. Gunasekara, S. Lee, A. Atkinson, B. Peng, H. Schulz, J. Gao, J. Li, M. Adada, M. Huang, L. Lastras, J. K. Kummerfeld, W. S. Lasecki, C. Hori, A. Cherian, T. K. Marks, A. Rastogi, X. Zang, S. Sunkara, and R. Gupta, "Overview of the eighth dialog system technology challenge: DSTC8," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2529–2540, 2021.

[3] G. A. Sigurdsson, G. Varol, X. Wang, I. Laptev, A. Farhadi, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *arXiv preprint arXiv:1604.01753*, 2016.

[4] 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," in *Proc. ICASSP*, May 2019, pp. 2352–2356.

[5] H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson, S. Lee, and D. Parikh, "Audio visual scene-aware dialog," in *Proc. CVPR*, Jun. 2019.

[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *Proc. ICCV*, 2015.

[7] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and Yang: Balancing and answering binary visual questions," in *Proc. CVPR*, 2016.

[8] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *Proc. CVPR*, 2017.

[9] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proc. CVPR*, 2016, pp. 4631–4640.

[10] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," *arXiv preprint arXiv:1611.08669*, 2016.

[11] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra, "Learning cooperative visual dialog agents with deep reinforcement learning," in *Proc. ICCV*, 2017.

[12] K. Khurana and U. Deshpande, "Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: A comprehensive survey." *IEEE Access*, 2021.

[13] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proc. ICCV*, Oct. 2017.

[14] S. Geng, P. Gao, M. Chatterjee, C. Hori, J. Le Roux, Y. Zhang, H. Li, and A. Cherian, "Dynamic graph representation learning for video dialog via multi-modal shuffled transformers," in *Proc. AAAI*, 2021.

[15] H. Le, D. Sahoo, N. Chen, and S. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," in *Proc. ACL*, 2019.

[16] Z. Li, Z. Li, J. Zhang, Y. Feng, and J. Zhou, "Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2021.

[17] V. Iashin and E. Rahtu, "A better use of audio-visual cues: Dense video captioning with bi-modal transformer," in *Proc. BMVC*, 2020.

[18] C. Hori, A. Cherian, T. K. Marks, and T. Hori, "Joint student-teacher learning for audio-visual scene-aware dialog," in *Proc. Interspeech*, 2019.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[20] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015, pp. 577–585.

[21] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, Mar. 2017.

[22] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. CVPR*, Jul. 2017.

[23] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *Proc. NIPS Deep Learning Symposium*, 2016.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, Dec. 2017, pp. 5998–6008.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, vol. 28, 2015, pp. 91–99.

[26] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[27] S. Palaskar, R. Sanabria, and F. Metze, "Transfer learning for multimodal dialog," *Comput. Speech Lang.*, vol. 64, p. 101093, 2020.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.