END-TO-END AUDIO VISUAL SCENE-AWARE DIALOG USING MULTIMODAL ATTENTION-BASED VIDEO FEATURES

Chiori Hori[†], *Huda Alamri*[*†], *Jue Wang*[†], *Gordon Wichern*[†], *Takaaki Hori*[†], *Anoop Cherian*[†], *Tim K. Marks*[†], *Vincent Cartillier*[*], *Raphael Gontijo Lopes*[*], *Abhishek Das*[*], *Irfan Essa*[*], *Dhruv Batra*[*] *Devi Parikh*[*]

[†]Mitsubishi Electric Research Laboratories (MERL) [*]Georgia Institute of Technology

ABSTRACT

In order for machines interacting with the real world to have conversations with users about the objects and events around them, they need to understand dynamic audiovisual scenes. The recent revolution of neural network models allows us to combine various modules into a single end-to-end differentiable network. As a result, Audio Visual Scene-Aware Dialog (AVSD) systems for real-world applications can be developed by integrating state-of-the-art technologies from multiple research areas, including end-to-end dialog technologies, visual question answering (VQA) technologies, and video description technologies. In this paper, we introduce a new data set of dialogs about videos of human behaviors, as well as an end-to-end Audio Visual Scene-Aware Dialog (AVSD) model, trained using this new data set, that generates responses in a dialog about a video. By using features that were developed for multimodal attention-based video description, our system improves the quality of generated dialog about dynamic video scenes.

Index Terms— Audio visual scene-aware dialog, Visual QA, Video description, End-to-end modeling

1. INTRODUCTION

Recently, end-to-end approaches have been shown to better handle flexible conversations between the user and the system by training the model on large conversational data sets [1, 2]. However, current dialog systems cannot understand dynamic scenes captured using multimodal sensor-based input such as vision and non-speech audio. As a result, 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. To interact with humans about audiovisual information, systems need to understand both audiovisual scenes and natural language inputs. The recent revolution of neural network models allows us to combine various modules into a single end-to-end differentiable network. Thus, we can simultaneously input visual features and user utterances into an encoder-decoder-based system whose outputs are natural-language responses.

Using this end-to-end framework, *visual question answering* (VQA) has been intensively researched in the field of computer vision [3–6]. to generate answers to questions about an imaged scene. As a further step towards conversational visual AI, the new task of *visual dialog* was introduced [7], in which an AI agent holds a meaningful dialog with humans about a static image using natural, conversational language [8]. While VQA and visual dialog take significant steps towards human-machine interaction, they only consider a single static image. In contrast, many real-world scenarios involve dynamic scenes that will require a system to understand the content and temporal dynamics of a scene, such as that contained in video data. To capture the semantics of dynamic scenes, recent research has focused on *video description*. The state of the art in video description uses a multimodal attention mechanism that selectively attends to different input modalities (feature types), such spatiotemporal motion features and audio features, in addition to temporal attention [9]. This framework allows us to build scene aware dialog systems using multimodal information, such as audio and visual features, by combining end-to-end dialog and video description technologies.

In this paper, we propose a new research target: a dialog system that can discuss dynamic scenes with humans. This goal lies at the intersection of multiple avenues of research in natural language processing, computer vision, and audio processing. To advance this goal, we introduce a new model that incorporates technologies for multimodal attention-based video description into an end-to-end dialog system. We also introduce a new data set of human dialogs about videos. We are making our data set, code, and model publicly available for a new Audio Visual Scene-Aware Dialog (AVSD) Challenge at the 7th Dialog System Technology Challenge (DSTC7) [10].

2. AUDIO VISUAL SCENE-AWARE DIALOG DATA SET

For this Audio Visual Scene-Aware Dialog (AVSD) data set, we collected text-based conversations about short videos from the Charades [11] video description data set, as described in [12]. In the AVSD track of the Dialog System Technology Challenges 7th edition (DSTC7)[1], the target is to generate human responses in dialogs using Natural Language Generation (NLG) technologies [10]. In AVSD data collection, two humans, a questioner and an answerer, had a discussion about the events in a video. The answerer, who had already watched the video, answered questions asked by the questioner. The questioner was not allowed to watch the video, but was shown the first, middle, and last frames of the video (three static images) to provide some fundamental grounding in the scene. After 10 rounds of Question (by the questioner) and Answer (by the answerer), the questioner was required to write a description summarizing the events in the video. Currently, we have collected dialogs for most of the Charades training set and all of the validation set, which form the prototype data set described in Table 1. The final data set for the AVSD track of DSTC7 will include the entire Charades data set. In this experiment, we split the official validation set for the Charades challenge in half, using the two halves as our validation and test sets.

[1] http://workshop.colips.org/dstc7/index.html

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

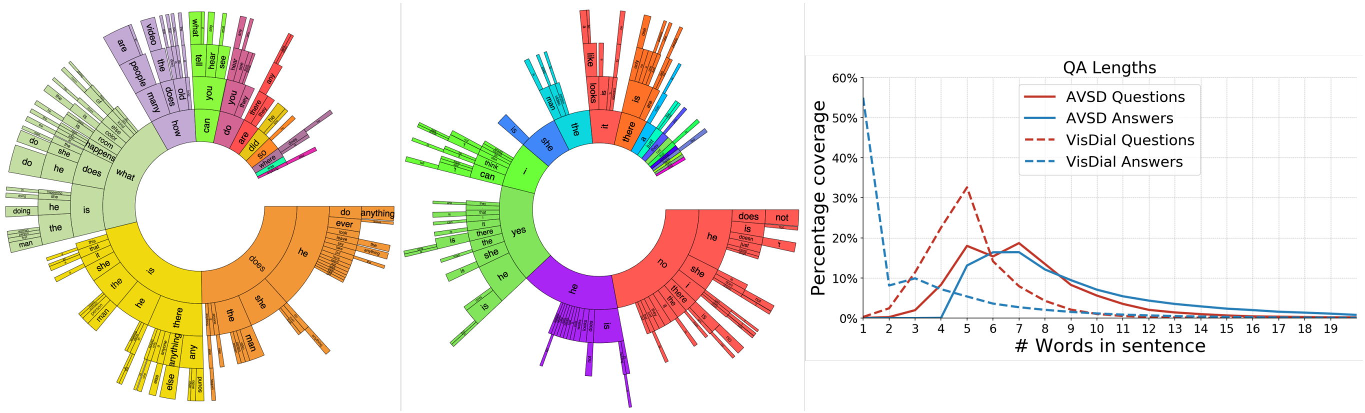


Fig. 1. The distributions of word 4-grams in the questions (left) and answers (middle) of the prototype data set of the AVSD, and the average length (right) of the sentences of the VQA and the prototype data set of the AVSD. The actions were mainly asked by the questioners. There are some questions regarding audio information. Half of the answers are Yes/No. The questions and answers of AVSD mainly consists of 5 to 8 words and longer than those of VQA. More descriptive sentences were generated for AVSD.

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

	training	validation	test
#dialogs	6,172	732	733
#turns	123,480	14,680	14,660
#words	1,163,969	138,314	138,790

3. AUDIO VISUAL SCENE-AWARE DIALOG SYSTEM

We built an end-to-end dialog system that can generate answers in response to user questions about events in a video sequence. Our architecture is similar to the Hierarchical Recurrent Encoder in Das *et al.* [7]. 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 attention-based video features into the LSTM encoder instead of single static image features. Fig. 2 shows the architecture of our audio visual scene-aware dialog system.

3.1. End-to-End Conversation Modeling

This section explains the neural conversation model of [1], which is designed as a sequence-to-sequence mapping process using recurrent neural networks (RNNs). Let X and Y be input and output sequences, respectively. The model is used to compute posterior probability distribution $P(Y|X)$. For conversation modeling, X corresponds to the sequence of previous sentences in a conversation, and Y is the system response sentence we want to generate. In our model, both X and Y are sequences of words. X contains all of the previous turns of the conversation, concatenated in sequence, separated by markers that indicate to the model not only that a new turn has started, but which speaker said that sentence. The most likely hypothesis of Y is obtained as

$$\hat{Y} = \arg\max_{Y\in\mathcal{V}^*} P(Y|X) \tag{1}$$

$$= \arg\max_{Y\in\mathcal{V}^*} \prod_{m=1}^{|Y|} P(y_m|y_1,\ldots,y_{m-1},X), \tag{2}$$

where $\mathcal{V}^*$ denotes a set of sequences of zero or more words in system vocabulary $\mathcal{V}$.

Let X be word sequence $x_1,\ldots,x_T$ and Y be word sequence $y_1,\ldots,y_M$. The encoder network is used to obtain hidden states h_t for $t=1,\ldots,T$ as:

$$h_t = \text{LSTM}\left(x_t, h_{t-1}; \theta_{enc}\right), \tag{3}$$

where h_0 is initialized with a zero vector. LSTM$(\cdot)$ is a LSTM function with parameter set θ_{enc}.

The decoder network is used to compute probabilities $P(y_m|y_1,\ldots,y_{m-1},X)$ for $m=1,\ldots,M$ as:

$$s_0 = h_T \tag{4}$$

$$s_m = \text{LSTM}\left(y_{m-1}, s_{m-1}; \theta_{dec}\right) \tag{5}$$

$$P(\mathbf{y}|y_1,\ldots,y_{m-1},X) = \text{softmax}(W_o s_m + b_o), \tag{6}$$

where y_0 is set to <eos>, a special symbol representing the end of sequence. s_m is the mth decoder state. θ_{dec} is a set of decoder parameters, and W_o and b_o are a matrix and a vector. In this model, the initial decoder state s_0 is given by the final encoder state h_T as in Eq. (4), and the probability is estimated from each state s_m. To efficiently find $\hat{Y}$ in Eq. (1), we use a beam search technique since it is computationally intractable to consider all possible Y.

In the scene-aware dialog scenario, a scene context vector including audio and visual features is also fed to the decoder. We modify the LSTM in Eqs. (4)–(6) as

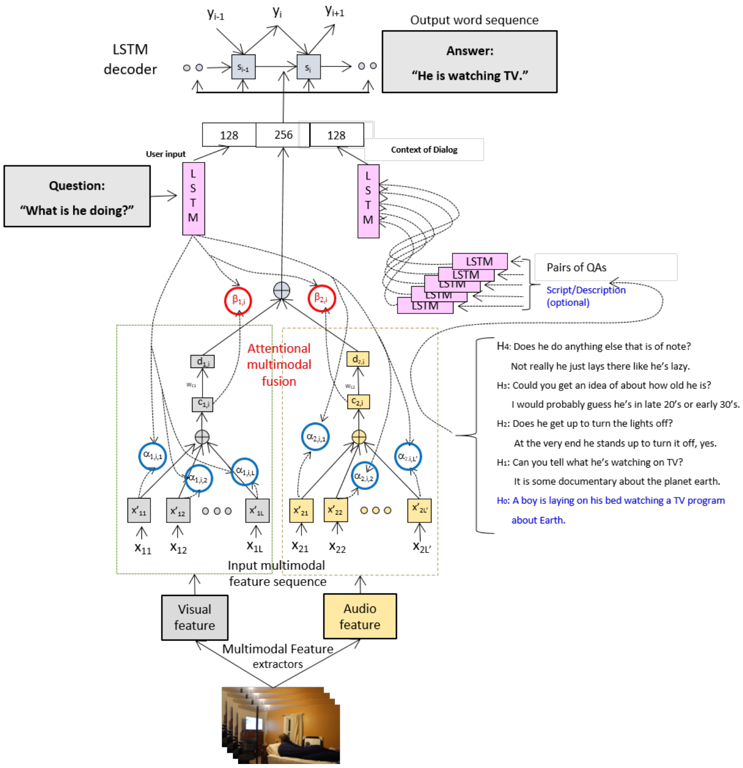
$$s_{n,0} = \bar{\mathbf{0}} \tag{7}$$

$$s_{n,m} = \text{LSTM}\left([y_{n,m-1}^{\intercal}, g_n^{\intercal}]^{\intercal}, s_{n,m-1}; \theta_{dec}\right), \tag{8}$$

$$P(\mathbf{y_n}|y_{n,1},\ldots,y_{n,m-1},X) = \text{softmax}(W_o s_{n,m} + b_o), \tag{9}$$

where g_n is the concatenation of question encoding $g_n^{(q)}$, audio-visual encoding $g_n^{(av)}$ and history encoding $g_n^{(h)}$ for generating the nth answer $A_n = y_{n,1},\ldots,y_{n,|Y_n|}$. Note that unlike Eq. (4), we feed all contextual information to the LSTM at every prediction step. This architecture is more flexible since the dimensions of encoder and decoder states can be different.

$g_n^{(q)}$ is encoded by another LSTM for the nth question, and $g_n^{(h)}$ is encoded with hierarchical LSTMs, where one LSTM encodes each question-answer pair and then the other LSTM summarizes the question-answer encodings into $g_n^{(h)}$. The audio-visual encoding is obtained by multi-modal attention described in the next section.



H4: Does he do anything else that is of note?
Not really he just lays there like he's lazy.
H3: Could you get an idea of about how old he is?
I would probably guess he's in late 20's or early 30's.
H2: Does he get up to turn the lights off?
At the very end he stands up to turn it off, yes.
H1: Can you tell what he's watching on TV?
It is some documentary about the planet earth.
H0: A boy is laying on his bed watching a TV program about Earth.

Fig. 2. Our multimodal attention-based audio visual scene-aware dialog (AVSD) system

3.2. Multimodal Attention-Based Video Features

To predict a word sequence in video description, prior work [13] extracted content vectors from image features of VGG-16 and spatiotemporal motion features of C3D, and combined them into one vector in the fusion layer as:

$$g_n^{(av)} = \tanh\left(\sum_{k=1}^{K} d_{k,n}\right), \tag{10}$$

where

$$d_{k,n} = W_{ck}^{(\lambda_D)} c_{k,n} + b_{ck}^{(\lambda_D)}, \tag{11}$$

and $c_{k,n}$ is a context vector obtained using the kth input modality. We call this approach Naïve Fusion, in which multimodal feature vectors are combined using projection matrices W_{ck} for K different modalities (input sequences $x_{k1}, \ldots, x_{kL}$ for $k = 1, \ldots, K$).

To fuse multimodal information, prior work [9] proposed a method extends the attention mechanism. We call this fusion approach *multimodal attention*. to predict the word sequence in video description. The number of modalities indicating the number of sequences of input feature vectors is denoted by K.

The following equation shows an approach to perform the attention-based feature fusion:

$$g_n^{(av)} = \tanh\left(\sum_{k=1}^{K} \beta_{k,n} d_{k,n}\right). \tag{12}$$

A similar mechanism for temporal attention is applied to obtain the multimodal attention weights $\beta_{k,n}$:

$$\beta_{k,n} = \frac{\exp(v_{k,n})}{\sum_{\kappa=1}^{K} \exp(v_{\kappa,n})}, \tag{13}$$

where

$$v_{k,n} = w_B^{\intercal} \tanh(W_B g_n^{(q)} + V_{Bk} c_{k,n} + b_{Bk}). \tag{14}$$

Here the multimodal attention weights are determined by question encoding $g_n^{(q)}$ and the context vector of each modality $c_{k,n}$ as well as temporal attention weights in each modality. W_B and V_{Bk} are matrices, w_B and b_{Bk} are vectors, and $v_{k,n}$ is a scalar. The multimodal attention weights can change according to the question encoding and the feature vectors (shown in Fig. 2). This enables the decoder network to attend to a different set of features and/or modalities when predicting each subsequent word in the description. Naïve fusion can be considered a special case of Attentional fusion, in which all modality attention weights, $\beta_{k,n}$, are constantly 1.

4. EXPERIMENTS FOR VIDEO DESCRIPTION

To select the best video features for the audio visual scene-aware dialog system, we first evaluate the performance of video description using multimodal attention-based video features in this paper. We evaluated our proposed feature fusion using the MSVD [14], MSR-VTT [15], and Charades [11] video data sets. Details of textual descriptions are summarized in Table 2.

Table 2. Sizes of textual descriptions in MSVD (YouTube2Text), MSR-VTT and Charades. DSCP: Description.

Data set	#Clips	#DSCP	#DSCP per clip	#Word	Vocabulary size
MSVD	1,970	80,839	41.00	8.00	13,010
MSR-VTT	10,000	200,000	20.00	9.28	29,322
Charades	9,848	16,140	1.64	13.04	2,582

The quality of the automatically generated sentences was evaluated with objective measures to compare the similarity between the generated sentences and the ground truth sentences. We used the evaluation code for MS COCO caption generation[2] for objective evaluation of system outputs, which is a publicly available tool supporting various automated metrics for natural language generation such as BLEU, METEOR, ROUGE_L, and CIDEr.

4.1. Audio and Video Processing

In our previous work on multimodal attention for video description [9] [16], we used two different types of audio features: concatenated mel-frequency cepstral coefficient (MFCC) features, and SoundNet [17] features. In this paper, we also evaluate features extracted using a new state-of-the-art model, Audio Set VGGish [18]. In this paper, we applied the VGGish model which was trained to predict an ontology of more than 600 audio event classes from only the audio tracks of 2 million human-labeled 10-second YouTube video soundtracks [18]. 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 seconds.

To understand visual context, the pretrained VGG-16 [19] and the pretrained C3D [20] models were used to generate features for object recognition and short-term spatiotemporal activity. In this experiment, we also adopted the state-of-the-art I3D features [21], 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 the experiments in

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

Table 3. Video description evaluation results on the MSVD (YouTube2Text), MSR-VTT Subset [9] and Charades.

MSVD (YouTube2Text) Full data set

Modalities (feature types)			Evaluation metric		
Image	Spatiotemporal	Audio	BLEU4	METEOR	CIDEr
VGG-16	C3D		0.524	0.320	0.688
VGG-16	C3D	MFCC	0.539	0.322	0.674
	I3D (rgb-flow)		0.525	0.330	0.742
	I3D (rgb-flow)	MFCC	0.527	0.325	0.702
		SoundNet	0.529	0.319	0.719
		VGGish	**0.554**	**0.332**	**0.743**

MSR-VTT Subset

Modalities (feature types)			Evaluation metric		
Image	Spatiotemporal	Audio	BLEU4	METEOR	CIDEr
VGG-16	C3D	MFCC	**0.397**	0.255	0.400
	I3D (rgb-flow)		0.347	0.241	0.349
	I3D (rgb-flow)	MFCC	0.364	0.253	0.393
		SoundNet	0.366	0.246	0.387
		VGGish	0.390	**0.263**	**0.417**

Charades data set

Modalities (feature types)			Evaluation metric		
Image	Spatiotemporal	Audio	BLEU4	METEOR	CIDEr
	I3D (rgb-flow)		0.094	0.149	0.236
	I3D (rgb-flow)	MFCC	0.098	0.156	0.268
		SoundNet	-	-	-
		VGGish	**0.100**	**0.157**	**0.270**

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). We used the same encoder-decoder network used in [9].

4.2. Results and Discussion

Table 3 shows the I3D spatiotemporal features outperformed the combination of VGG-16 image features and C3D spatiotemporal features. We believe this is because I3D features already include enough image information for the video description task, since they use the more powerful Inception-V3 network architecture and were trained on the larger (and cleaner) Kinectics [22] data set. As a result, I3D has demonstrated state-of-the-art performance for the task of human action recognition in video sequences [21]. Further, the Inception-V3 architecture has significantly fewer network parameters than the VGG-16 network, making it more efficient. In terms of audio features, the Audio Set VGGish model provided the best performance. First, the VGGish model was trained on more data, and had audio specific labels, whereas SoundNet used pre-trained image classification networks to provide labels for training the audio network. Second, the large Audio Set ontology used to train VGGish likely provides the ability to learn features more relevant to text descriptions than the broad scene/object labels used by SoundNet.

Since it is intractable to enumerate all possible word sequences in vocabulary $\mathcal{V}$, we usually limit them to the n-best hypotheses generated by the system. Although in theory the distribution $P(Y'|X)$ should be the true distribution, we instead estimate it using the encoder-decoder model.

5. EXPERIMENTS FOR AVSD

In this paper, we extended an end-to-end dialog system to scene-aware dialog with multimodal fusion, as described in Section 3 and shown in Fig. 2. The decision of which video and audio features to extract was based on the results in Section 4. We evaluated our proposed system using the AVSD data set on Charades that we collected (see Table 1 for details of the data set size). We compared the performance between models trained from various combinations of the QA text, visual features, and audio features. In addition, we tested the efficacy of our multimodal attention mechanism for dialog response generation. We employed an ADAM optimizer [23] with the cross-entropy criterion and iterated the training process up to 20 epochs. For each of the encoder-decoder model types, we selected the model with the lowest perplexity on the expanded development set. We used LSTMs with parameter values #layer=2 and #cells=128 to encode dialog history and question sentences. Video features were projected to 256-dimensional feature space before modality fusion. The decoder LSTM also had a structure with #layer=2 and #cells=128.

Table 4. AVSD System response generation evaluation results with objective measures. Note that I3D-rgb and I3D-flow have different attention weights separately. In this experiment, we tested the 733 dialogs by comparing with one groundtruth. The AVSD challenge at DSTC7 compared one response in each dialog with 5 groundtruths for the full set of 1,710 dialogs [10].

Input features	Attentional fusion	BLEU4	METEOR	ROUGE_L	CIDEr
QA	-	0.065	0.101	0.257	0.595
QA + Captions	-	0.073	0.109	0.271	0.705
QA + VGG16	-	0.067	0.102	0.259	0.618
QA + I3D	no	0.073	0.109	0.269	0.680
QA + I3D	yes	0.077	0.110	0.274	0.724
QA + I3D + VGGish	no	0.075	0.110	0.275	0.701
QA + I3D + VGGish	yes	**0.078**	**0.113**	**0.277**	**0.727**

5.1. Evaluation Results

Table 4 evaluates the performance of our models at generating response sentences using objective measures. We investigated different input features including question-answering dialog history plus last question (QA), human-annotated captions (Captions), video features of VGG16 or I3D rgb+flow features (I3D), and audio features (VGGish). We tested these both with and without our multimodal attention (Attentional fusion). All of the objective metrics show that the attentional fusion of I3D and VGGish outperformed other combination of modalities. These results on the Audio Visual Scene-aware Dialog task are entirely consitent with our results from the video description task (Section 4 and Table 3)

6. CONCLUSION

In this paper, we propose a new research target: a dialog system that can discuss dynamic scenes with humans. This task lies at the intersection of multiple avenues of research in natural language processing, computer vision, and audio processing. To advance this goal, we introduce a new model that incorporates technologies for multimodal attention-based video description into an end-to-end dialog system. We also introduce a new data set of human dialogs about videos. Our experiments demonstrate that using multimodal features that were developed for multimodal attention-based video description enhances the quality of generated dialog about dynamic scenes. Future work includes (1) finding additional features that can make the word distributions in the audiovisual semantic vector space more distinguishable and (2) applying open-domain language models to video description.

7. REFERENCES

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

[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," *Computer Speech and Language*, vol. Special issue on DSTC6, to appear in 2018.

[3] 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.

[4] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, "Yin and Yang: Balancing and answering binary visual questions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, "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)*, 2017.

[6] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, "MovieQA: Understanding Stories in Movies through Question-Answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] 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.

[8] 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.

[9] 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 *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] Huda Alamri, Chiori Hori, Tim K. Marks, Dhruv Batr, and Devi Parikh, "Audio Visual Scene-aware dialog (AVSD) Track for Natural Language Generation in DSTC7," in *DSTC7 workshop at AAAI*, 2019.

[11] 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.

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

[13] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu, "Video paragraph captioning using hierarchical recurrent neural networks," *CoRR*, vol. abs/1510.07712, 2015.

[14] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2712–2719.

[15] Jun Xu, Tao Mei, Ting Yao, and Yong Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] Chiori Hori, Takaaki Hori, Tim K Marks, and John R Hershey, "Early and late integration of audio features for automatic video description," in *ASRU*, 2017.

[17] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NIPS*, 2016.

[18] 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 *ICASSP*, 2017.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[20] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 4489–4497.

[21] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv*, 2017.

[23] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.