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

Joint Student-Teacher Learning for Audio-Visual Scene-Aware Dialog

Hori, C.; Cherian, A.; Marks, T.; Hori, T.

TR2019-097 September 18, 2019

Abstract

Multimodal fusion of audio, vision, and text has demonstrated significant benefits in advancing the performance of several tasks, including machine translation, video captioning, and video summarization. Audio-Visual Scene-aware Dialog (AVSD) is a new and more challenging task, proposed recently, that focuses on generating sentence responses to questions that are asked in a dialog about video content. While prior approaches designed to tackle this task have shown the need for multimodal fusion to improve response quality, the best-performing systems often rely heavily on human-generated summaries of the video content, which are unavailable when such systems are deployed in real-world. This paper investigates how to compensate for such information, which is missing in the inference phase but available during the training phase. To this end, we propose a novel AVSD system using studentteacher learning, in which a student network is (jointly) trained to mimic the teacher's responses. Our experiments demonstrate that in addition to yielding state-of-the-art accuracy against the baseline DSTC7-AVSD system, the proposed approach (which does not use human-generated summaries at test time) performs competitively with methods that do use those summaries

Interspeech

Joint Student-Teacher Learning for Audio-Visual Scene-Aware Dialog

Chiori Hori, Anoop Cherian, Tim K. Marks, Takaaki Hori

Mitsubishi Electric Research Laboratories (MERL), USA.

chori@merl.com, cherian@merl.com, tmarks@merl.com, thori@merl.com

Abstract

Multimodal fusion of audio, vision, and text has demonstrated significant benefits in advancing the performance of several tasks, including machine translation, video captioning, and video summarization. Audio-Visual Scene-aware Dialog (AVSD) is a new and more challenging task, proposed recently, that focuses on generating sentence responses to questions that are asked in a dialog about video content. While prior approaches designed to tackle this task have shown the need for multimodal fusion to improve response quality, the best-performing systems often rely heavily on human-generated summaries of the video content, which are unavailable when such systems are deployed in real-world. This paper investigates how to compensate for such information, which is missing in the inference phase but available during the training phase. To this end, we propose a novel AVSD system using student-teacher learning, in which a student network is (jointly) trained to mimic the teacher's responses. Our experiments demonstrate that in addition to yielding state-of-the-art accuracy against the baseline DSTC7-AVSD system, the proposed approach (which does not use human-generated summaries at test time) performs competitively with methods that do use those summaries.

Index Terms: dialog system, end-to-end conversation model, question answering, audio-visual scene-aware dialog

1. Introduction

Human-machine interfaces that can process spoken dialogs have revolutionized the way we interact with smart phone digital assistants, car navigation systems, voice-controlled smart speakers, and human-facing robots. Going forward, such systems will need capabilities to accommodate other input modalities, including vision, to generate adequate responses in varied user contexts or process novel situations that were not available during training. However, the current state-of-the-art dialog systems lack efficient models for processing multimodal sensory inputs (e.g., vision, audio, and text) that are required to handle such dynamic scenes, and thus may not be able to generate suitable responses in conversations.

Recently, through the advances in deep learning, there have been efforts to build end-to-end dialog systems that can be trained to map directly from a user utterance to a system response. Such systems allow us to combine different modules into a single end-to-end differentiable network, simultaneously taking video features and user utterances as inputs to an encoder-decoder-based system whose outputs are natural-language responses. 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 [1, 2]. Using such end-to-end frameworks, *visual question answering* (VQA) [3] and *visual dialog* [4] have been proposed to directly answer questions about a scene using information present in a single static image. These are significant steps towards enabling more natural human-machine interaction.

As a further step towards conversational visual AI, a new dialog task using multimodal information processing has been proposed, called Audio-Visual Scene-aware Dialog (AVSD) [5, 6, 7]. AVSD focuses on response sentence generation for dialog systems aimed at answering a user's questions about a provided video, in which the system can use audio-visual information in the video as well as the dialog history up to the user's last question. Optionally, short human-generated summaries that explain the video clip are also available as input to the system. Recent approaches to the AVSD task (proposed in the 7th Dialog System Technology Challenge (DSTC7)) have shown that multimodal fusion of audio, visual, and text information is effective to enhance the response quality. Further, it is found that the best performance is achieved when including text features extracted from the available summaries. Surprisingly, systems using such manual descriptions enable performance close to the best system, even without using the audio-visual features. However, such summaries are unavailable in the real world, posing challenges during deployment.

In this paper, we investigate how to compensate during the inference (test) phase for this information, which is missing in the inference phase but available in training phase, to improve AVSD responses without using manual descriptions. We could use automatic video description to generate descriptive sentences from the video clips, but it is not easy to build a video description system that generates accurate descriptions. Instead, we propose a new AVSD system, which is trained through a student-teacher learning approach. The teacher model is first trained with manual descriptions, then a student model is trained without the descriptions to mimic the teacher's output. The student model is used in the inference phase. We also extend this framework to joint student-teacher learning, where the both models are trained together not only to reduce their own loss functions but also to have similar hidden representations of context vectors with each other. In this learning, the teacher model is updated to be mimicked more easily by the student model since the context vector of the teacher model approaches to that of the student model. The new system achieves better performance than prior approaches. It is competitive to those trained with manual descriptions including the best DSTC7-AVSD system [8].

2. Related Work

Student-teacher learning is a technique of transfer learning, in which the knowledge in a teacher model is transferred to a student model. This is typically used for model compression [9, 10, 11], where a small model is trained to mimic the output of a large model that has higher prediction accuracy. Student-teacher learning can bring the performance of the small model closer to that of the large model, while preserving the small model's benefits of reduced computational cost and memory consumption.

Student-teacher learning can also be used to compensate

for missing information in the input. In this case, the teacher model is trained to predict target labels using additional information, but the student model is trained to mimic the teacher's output without that information. In automatic speech recognition (ASR), for example, a teacher model is trained with enhanced speech obtained through a microphone array, while a student model is trained to mimic the teacher's output for the same speech but only using single-channel-recorded noisy speech [12]. With this method, the student model can improve the performance without the microphone array at test time. This technique was also used for domain adaption between child and adult speech [13]. The proposed AVSD system takes this approach to compensate for a missing video description. The student model can generate better responses without description features. We further extend this framework to joint student-teacher learning, aiming at improving the teacher model to be a better teacher for the student model.

3. System Architecture

The architecture of the proposed AVSD system is shown in Figure 1. The system employs an attention-based encoder-decoder [14, 15], which enables the network to emphasize features from specific time frames depending on the current context, enabling the next word to be generated more accurately. The efficacy of attention models has been shown in many tasks such as machine translation [14] and video description [16, 17].

The attention-based encoder-decoder 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 computes the posterior probability distribution $P(Y|X)$. For the AVSD task, X includes all the input information such as the user's question, audio-visual features, and dialog context (dialog history). Y is the system response to be generated, which answers the user's question. The most likely hypothesis of Y is obtained as:

$$\hat{Y} = \underset{Y \in \mathcal{V}^*}{\operatorname{argmax}} P(Y|X) \tag{1}$$

$$= \underset{Y \in \mathcal{V}^*}{\operatorname{argmax}} \prod_{i=1}^{|Y|} P(y_i|y_1, \ldots, y_{i-1}, X), \tag{2}$$

where $\mathcal{V}^*$ denotes a set of sequences of zero or more words in system vocabulary $\mathcal{V}$, and each y_i is a word in the response.

Let $X = \{X_1, \ldots, X_K\}$ be a set of input sequences, where X_k is the kth input sequence, which can represent the user's question, a feature vector sequence extracted from the target video clip, or dialog history that includes all of the previous questions and answers in the dialog about the video clip. To generate system response Y, each input sequence in X is first encoded to a better representation using a corresponding encoder.

If X_k is a user's question, the sentence $Q = w_{Q,1}, \ldots, w_{Q,N}$ is encoded with word embedding and BLSTM layers. If X_k is a video feature sequence $X_k = x_{k1}, x_{k2}, \ldots, x_{kL_k}$, it can be extracted from the image sequence of the video clip using a pretrained CNN, such as VGG-16 [18], C3D [19], or I3D [20], that was originally trained for an image or video classification task. In the case of C3D and I3D, multiple images are fed to the network at once to capture dynamic features in the video. The audio features can also be extracted in a similar way using a pretrained CNN such as SoundNet [21] or VGGish [22]. Each feature vector sequence is encoded to an appropriate representation $X'_k = x'_{k1}, x'_{k2}, \ldots, x'_{kL_k}$ using a single projection layer for dimensionality reduction. If X_k is the dialog history, it can be a sequence of question-answer pairs $\mathbf{H} = H_1, \ldots, H_J$ that appear before the current question in the dialog. $\mathbf{H}$ is encoded using a hierarchical LSTM encoder, where each question-answer pair is first encoded to a fixed dimensional vector H_j using a sentence-embedding LSTM, and the sequence of sentence embeddings is further embedded using additional BLSTM layers.

The decoder predicts the next word iteratively beginning with the start-of-sentence token, <sos>, until it predicts the end-of-sentence token, <eos>. Given decoder state s_{i-1}, the decoder network λ_D infers the next-word probability distribution as

$$\begin{aligned} &P(y|y_1, \ldots, y_{i-1}, X) \\ &\quad \approx P(y|s_{i-1}, g_i) \\ &\quad = \text{softmax}\left(W_s^{(\lambda_D)}[s_{i-1}, g_i] + b_s^{(\lambda_D)}\right), \end{aligned} \tag{3}$$

and generates the word y_i that has the highest probability according to

$$y_i = \underset{y \in \mathcal{V}}{\operatorname{argmax}} P(y|s_{i-1}, g_i). \tag{4}$$

The decoder state is updated using the LSTM network of the decoder as

$$s_i = \text{LSTM}\big(s_{i-1}, [y'_i, g_i]; \lambda_D\big), \tag{5}$$

where y'_i is a word-embedding vector of y_i, and g_i is a *context vector* including the input information relevant to the previous decoder state. λ_D denotes the set of decoder parameters.

The context vector is obtained by a hierarchical attention mechanism that first aggregates frame-level hidden vectors for each input sequence into modality-wise context vector $c_{k,i}$, and then fuses the context vectors $c_{1,i}, \ldots, c_{K,i}$ into a single context vector g_i. The attention mechanism is realized by using *attention weights* to the hidden activation vectors throughout the input sequence. These weights enable the network to emphasize features from those time steps that are most important for predicting the next output word.

Let $\alpha_{k,i,t}$ be an attention weight between the ith output word and the tth input feature vector from the kth modality. For the ith output, the vector representing the relevant content of the input sequence is obtained as a weighted sum of hidden unit activation vectors:

$$c_{k,i} = \sum_{t=1}^{L_k} \alpha_{k,i,t} h_{k,t}, \tag{6}$$

where $h_{k,t}$ is the tth output vector of the kth encoder. The attention weights are computed in the same manner as in [14]:

The model also utilizes a multimodal attention mechanism. To fuse multimodal information, prior work [17] proposed a method that extends the attention mechanism from temporal attention (attention over time) to attention over modalities. The following equation shows an approach to perform the attention-based feature fusion:

$$g_i = \tanh\left(\sum_{k=1}^{K} \beta_{k,i} d_{k,i}\right), \tag{7}$$

where

$$d_{k,i} = W_{ck}^{(\lambda_D)} c_{k,i} + b_{ck}^{(\lambda_D)}, \tag{8}$$

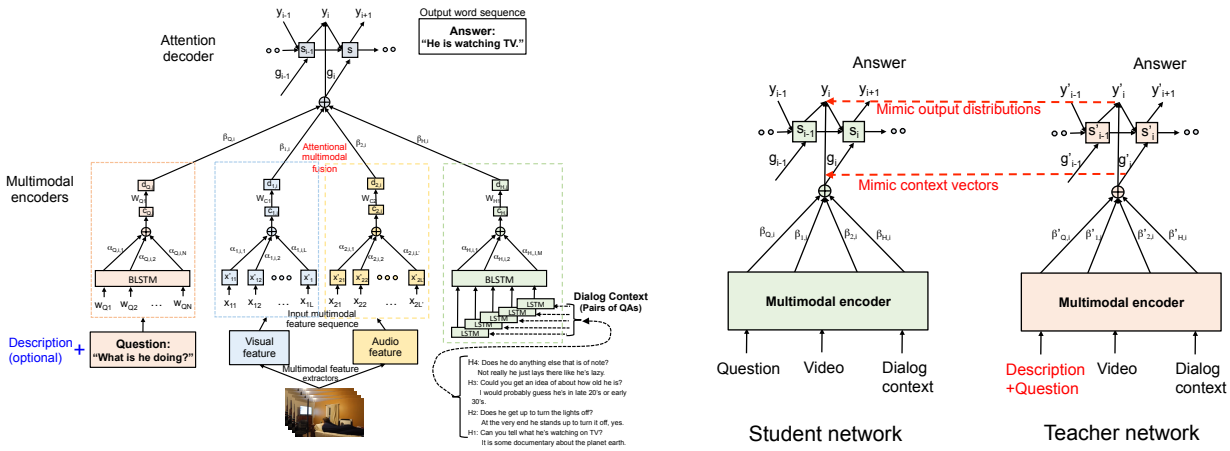

Figure 1: *Left: Multimodal-attention based audio-visual scene-aware dialog system, Right: Student-teacher learning of AVSD system.*

and $c_{k,i}$ is a context vector obtained using the kth input modality. A similar mechanism for temporal attention is applied to obtain the multimodal attention weights $\beta_{k,i}$ [17]. These weights can change according to the decoder state and the context vector from each encoder. This enables the decoder network to attend to a different set of features and/or modalities when predicting each subsequent word in the description.

4. Student-Teacher Learning

Figure 1 (right) depicts the concept of student-teacher learning for the AVSD system. The goal of this step is to obtain a student model that does not make use of video description text, which is trained to mimic a teacher model that has already been trained using video description text. Accordingly, the student model can be used to generate system responses without relying on description text, while hopefully achieving similar performance to the teacher model.

Following the best system in DSTC7-AVSD track [8], we insert the description text at the beginning of each question. This means that the same description is always fed to the encoder together with a new question, at every turn of the dialog about the target video clip. The student network is trained to reduce the cross entropy loss, by using the output of the teacher network as a soft target to make the output distribution of the student model closer to that of the teacher model.

In this paper, we investigate three loss functions for student-teacher learning. The first one is a cross entropy loss with *soft* targets:

$$\mathcal{L}_{\text{ST}}(X,Y) = -\sum_{i=1}^{|Y|}\sum_{y\in\mathcal{V}} \hat{P}(y|\hat{s}_{i-1},\hat{g}_i)\log P(y|s_{i-1},g_i), \tag{9}$$

where $\hat{P}(y|\hat{s}_{i-1},\hat{g}_i)$ denotes the probability distribution for the ith word obtained by the teacher network, and $\hat{s}_{i-1}$ and $\hat{g}_i$ are state and context vectors generated by the teacher network for training sample (X,Y). Here, $P(y|s_{i-1},g_i)$ is the posterior distribution from the current student network (which is being trained), which is predicted without the description text.

The second loss function further incorporates the context vector similarity as

$$\mathcal{L}'_{\text{ST}}(X,Y) = \mathcal{L}_{\text{ST}}(X,Y) + \lambda_c\mathcal{L}_{\text{MSE}}(X,Y) \tag{10}$$

Table 1: *Video Scene-aware Dialog Dataset on Charades*

	training	validation	trial	test
#dialogs	7,659	1,787	733	1,710
#turns	153,180	35,740	14,660	13,490
#words	1,450,754	339,006	138,790	110,252

where $\mathcal{L}_{\text{MSE}}(X,Y) = \sum_{i=1}^{|Y|}\text{MSE}(g_i,\hat{g}_i)$, where $\text{MSE}(\cdot,\cdot)$ denotes the mean square error between two context vectors, and λ_c denotes a scaling factor. We aim here to compensate for missing input features at the context vector level, which hopefully exploits other modalities more actively.

The last loss function we consider is joint student-teacher learning. The parameters of the teacher network are typically kept fixed throughout the training phase. However, in the joint training approach, we update not only the student network but also the teacher network. The loss function is computed as

$$\mathcal{L}_{\text{JST}}(X,Y) = \mathcal{L}^{(S)}_{\text{ST}}(X,Y) + \mathcal{L}^{(T)}_{\text{CE}}(X,Y) + \lambda_c\mathcal{L}^{(ST)}_{\text{MSE}}(X,Y), \tag{11}$$

where $\mathcal{L}^{(T)}_{\text{CE}}$ is the standard cross entropy for *hard* target Y, which is used only for the teacher network in the backpropagation process. Likewise, $\mathcal{L}^{(S)}_{\text{ST}}$ is used only for the student network, while $\mathcal{L}^{(ST)}_{\text{MSE}}$ is used for the both networks.

5. Experiments

5.1. Dialog Data

The AVSD data set is a collection of text-based conversations about short videos [5, 6, 7][1]. The video clips were originally from Charades [23] data set, which is an untrimmed and multi-action dataset, containing 11,848 videos split into 7,985 for training, 1,863 for validation, and 2,000 for testing. It has 157 action categories, with several fine-grained actions. Further, this dataset also provides 27,847 textual descriptions for the videos; each video is described using 1–3 sentences. For each video in the Charades dataset, the AVSD dataset contains a text dialog between two people discussing the video. See Table 1 for statistics of the dataset.

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

Table 2: *Evaluation results on the AVSD trial set with single references*

System	Description train	Description test	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
AVSD baseline [5]	—	—	0.273	0.173	0.118	0.084	0.117	0.291	0.766
AVSD best system [8]	man.	man.	0.306	0.209	0.150	0.112	0.144	0.338	1.161
+ How2 data	man.	man.	0.311	0.212	0.152	0.114	0.146	0.337	1.169
Our system	man.	man.	**0.311**	**0.214**	**0.156**	**0.117**	**0.150**	**0.345**	**1.234**
Our system	man.	—	0.272	0.186	0.135	0.102	0.132	0.325	1.105
Our system	man.	auto	0.285	0.193	0.140	0.106	0.135	0.329	1.121
Our system	—	—	0.283	0.192	0.139	0.105	0.135	0.327	1.119
Student-teacher $\mathcal{L}_{\mathrm{ST}}$	man.	—	0.313	0.212	0.152	0.113	0.143	0.334	1.138
Student-teacher $\mathcal{L}'_{\mathrm{ST}}$	man.	—	**0.314**	0.212	0.152	0.113	0.143	0.334	1.139
Student-teacher $\mathcal{L}_{\mathrm{JST}}$	man.	—	**0.314**	**0.213**	**0.153**	**0.115**	**0.144**	**0.335**	**1.148**

Table 3: *Evaluation results on the AVSD official test set with six references*

System	Description train	Description test	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
AVSD baseline [5]	—	—	0.621	0.480	0.379	0.305	0.217	0.481	0.733
AVSD best system [8]	man.	man.	0.718	0.584	0.478	0.394	0.267	0.563	1.094
+ How2 data	man.	man.	0.723	0.586	0.476	0.387	0.266	0.564	1.087
Our system	man.	man.	**0.727**	**0.593**	**0.488**	**0.405**	**0.273**	**0.566**	**1.118**
Our system	—	—	0.675	0.543	0.446	0.371	0.248	0.527	0.966
Student-teacher $\mathcal{L}_{\mathrm{ST}}$	man.	—	**0.686**	0.556	0.457	0.380	**0.254**	0.535	0.995
Student-teacher $\mathcal{L}_{\mathrm{JST}}$	man.	—	**0.686**	**0.557**	**0.458**	**0.382**	**0.254**	**0.537**	**1.005**

5.2. AVSD system

We trained the AVSD system in Fig. 1. The question encoder had a word embedding layer (200 dim.) and two BLSTM layers (256 dim. for each direction). Audio-visual features consisting of I3D-rgb (2048 dim.), I3D-flow (2048 dim.), and VGGish (128 dim.) were extracted from video frames using pre-trained deep CNNs. Those feature sequences were then fed to the multimodal encoders with single projection layers, which converted them to 512, 512, and 64 dimensional vectors, respectively. The history encoder had a word embedding layer (200 dim.) and two LSTM layers for QA-pair embedding (256 dim.) and a 1-layer BLSTM for embedding the history (256 dim. for each direction). We used ADAM optimizer for training, where the learning rate was halved if the validation perplexity did not decrease after each epoch, and continued training up to 20 epochs. The vocabulary size was 3910, where we kept only the words that appeared at least four times in the training set.

5.3. Results and Discussion

Table 2 shows evaluation results on the AVSD trial test set with single references. The quality of system responses was measured using objective scores such as BLEU, METEOR, ROUGE-L, and CIDEr, which were based on the degree of word overlapping with references. The baseline system provided by DSTC7-AVSD track organizers, which was a simple LSTM-based encoder decoder [5] utilizing the same audio-visual features as ours, was also evaluated. We also show the results of the AVSD best system [8]. That system had a similar architecture to ours, but it had only two encoders: one for questions, and the other for video features obtained by a 3D ResNet. That network was additionally pretrained using the How2 data set, while our model was trained with only the AVSD data set.

Although our system outperformed the best AVSD system when using manual descriptions for both training and testing ("man. man." in the second column), the performance significantly degraded when the description was not fed to the network in the test phase ("man. —"). When we provided automatic description instead of manual one ("man. auto"), where we used a video description model trained with the same AVSD data set, the improvement was limited. The model trained without descriptions ("— —") was slightly better than other conditions.

Next, we applied student-teacher learning with loss $\mathcal{L}_{\mathrm{ST}}$. The trained model provided significant gains in all the objective metrics (e.g., BLEU4: 0.105 $\rightarrow$ 0.113, METEOR: 0.135 $\rightarrow$ 0.143), which were closer to those obtained using the manual descriptions (e.g., BLEU4: 0.117, METEOR: 0.150). We also applied loss function $\mathcal{L}'_{\mathrm{ST}}$ that considered context vector similarity, but the response quality was almost the same as $\mathcal{L}_{\mathrm{ST}}$. Finally, we conducted joint student-teacher learning with $\mathcal{L}_{\mathrm{JST}}$, and obtained further improvements in most objective measures (e.g., BLEU4: 0.113 $\rightarrow$ 0.115, METEOR: 0.143 $\rightarrow$ 0.144).

Table 3 shows evaluation results on the AVSD official test set with six references for each response. Similar to Table 2, our system outperformed the other ones including the best system of DSTC7. The student-teacher framework also provided significant gains for the official test set.

6. Conclusion

This paper investigated how to compensate, at test time, for the lack of video description features that were available during training. We proposed a student-teacher learning framework for Audio-Visual Scene-aware Dialog (AVSD). Our AVSD system achieved better performance than previous methods, which is competitive to systems trained with manual descriptions, and further outperformed the best DSTC7-AVSD system. The trained model can answer questions about video content by fusing audio, visual, and text information about the video, and generates high quality responses without relying on manual video descriptions. We also proposed a joint student-teacher learning approach, which provided further gains in most objective metrics.

7. References

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

[2] R. Lowe, N. Pow, I. Serban, and J. Pineau, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," *arXiv preprint arXiv:1506.08909*, 2015.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[4] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," *CoRR*, vol. abs/1611.08669, 2016. [Online]. Available: http://arxiv.org/abs/1611.08669

[5] H. Alamri, C. Hori, T. K. Marks, D. Batra, and D. Parikh, "Audio visual scene-aware dialog (AVSD) track for natural language generation in DSTC7," in *DSTC7 at AAAI2019 Workshop*, 2019.

[6] C. Hori, H. AlAmri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh, "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," *CoRR*, vol. abs/1806.08409, 2018. [Online]. Available: http://arxiv.org/abs/1806.08409

[7] H. AlAmri, V. Cartillier, A. Das, J. Wang, S. Lee, P. Anderson, I. Essa, D. Parikh, D. Batra, A. Cherian, T. K. Marks, and C. Hori, "Audio-visual scene-aware dialog," *CoRR*, vol. abs/1901.09107, 2019. [Online]. Available: http://arxiv.org/abs/1901.09107

[8] R. Sanabria, S. Palaskar, and F. Metze, "Cmu sinbads submission for the dstc7 avsd challenge," in *DSTC7 at AAAI2019 workshop*, 2019.

[9] C. Bucilu, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 535–541.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[11] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1317–1327.

[12] S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "Student-teacher network learning with enhanced features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5275–5279.

[13] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," *Proc. Interspeech 2017*, pp. 2386–2390, 2017.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[15] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 577–585. [Online]. Available: http://papers.nips.cc/paper/5847-attention-based-models-for-speech-recognition.pdf

[16] L. Yao, A. Torabi, K. Cho, N. Ballas, C. J. Pal, H. Larochelle, and A. C. Courville, "Describing videos by exploiting temporal structure," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 4507–4515. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.512

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

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[19] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. 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. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.510

[20] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[21] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in neural information processing systems*, 2016, pp. 892–900.

[22] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[23] G. A. Sigurdsson, G. Varol, X. Wang, I. Laptev, A. Farhadi, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *ArXiv*, 2016. [Online]. Available: http://arxiv.org/abs/1604.01753