DSTC7-AVSD: Scene-Aware Video-Dialogue Systems with Dual Attention

Ramakanth Pasunuru **Mohit Bansal**
Department of Computer Science
University of North Carolina at Chapel Hill
{ram, mbansal}@cs.unc.edu

Abstract

Scene-aware dialogue systems are designed to have conversations about surrounding objects and events. We approach this challenge by building an end-to-end multimodal dialogue system with video (non-audio) and chat history as the context with novel ways of grounding through effective alignment and cross-attention approaches. For this, we use the Audio Visual Scene-Aware Dialog (AVSD) dataset to evaluate the performance of our models and also study the importance of each of the modality and component to the overall performance of our multimodal dialogue models. This achieves the third-rank system in the competition. Further, we also discuss the various other approaches that we tried to improve the performance of our models, e.g., reinforcement learning, contextual embeddings, pointer-generator copy models, and external data.

1 Introduction

Building end-to-end scene-aware dialogue systems is an important step in enabling the grounding of objects and events for having natural conversations with robots in a collaborative environment. To enable such interactions, systems need to understand both dynamic visual scenes and the natural language inputs. Such systems play a crucial role in the applications of virtual assistants, intelligent tutoring, and human robot collaboration.

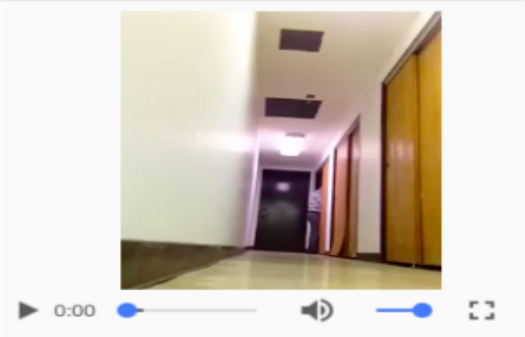
In the direction of grounding, previous works have explored the translation of visual information in the form of image/video captioning (Karpathy and Fei-Fei 2015; Xu et al. 2015; Venugopalan et al. 2015; Pasunuru and Bansal 2017a), image/video question answering and reasoning (Antol et al. 2015; Jang et al. 2017; Lei et al. 2018). Also, recent works have explored the visual context in dialogue systems in the form of static-image based context (Das et al. 2017; Mostafazadeh et al. 2017; de Vries et al. 2017). Very recently, Pasunuru and Bansal (2018) introduced game-based video context dialogue with multi-speaker dialogue. Introducing video and audio context for dialogue systems, audio visual scene-aware dialog (AVSD) was proposed by Alamri et al. (2018) (see Fig. 1 for an example). In this work, we propose end-to-end neural network based multimodal video context dialogue models for this AVSD dataset.

Several end-to-end neural network based models have been explored for dialogue systems with textual, speech, gaze, and gesture as context (Lowe et al. 2015; Serban et al. 2016; Johnston et al. 2002; Cassell 1999). At the same time, different multimodal end-to-end dialogue models were developed in the domain of image-based visual question and answering type dialogue (Das et al. 2017). However, models for video-based context are less explored. To this end, we present a multimodal-context (video and chat history) based question-answering-style dialogue model exploring the recent AVSD dataset. In this work, we primarily focus on encoding and aligning multiple modalities (video, chat history, and summary, but no audio) w.r.t. the given question. For this, we propose a dual attention mechanism, where we use cross-modality attention to better align different video and question modalities and as well use general attention at the answer decoder to allow the model to attend to important parts of each of the modalities to answer the given question.

In our empirical studies, we found that video and chat history modalities are both important for improving the performance of the question answering based multimodal dialogue model. Also, we use the summary information to further improve the results. We show that cross-modality attention improves the results showing that it is important to align the modalities (video and question) to answer the question. Further, we also experimented with various other advanced techniques such as reinforcement learning based policy gradient approach with task-specific rewards, adding contextual word embedding representations (ELMo), using additional external data, as well as joint pointer-generator models. However, these methods did not perform well with the AVSD dataset. We describe each of these approaches and discuss the possible reasons for their lack of impact on the model performance.

2 Related Work

It is important for grounding objects and events to enable human-robot interactions. Several previous works have explored the visual domain (both image and video) for translation, question answering, and summarization using deep neural network models (Venugopalan et al. 2015; Pasunuru and Bansal 2017a; Antol et al. 2015; Lei et al. 2018; Yu, Bansal, and Berg 2017). Recently, some previous works have extended this for question answering based di-



Person A (Questioner)

1. How many people are in the video?
2. Is he speaking with anyone?
3. What room is he in?
4. What is the man doing?
5. Does he start the video in the hallway?
6. Where does he put the tie and shirt?
7. Does he leave the hallway?
8. Does he open the closet door?
9. Can you tell what he grabs from the closet?
10. Is there anything else I should know?

Person B (Answerer)

1. There is only one man in the video
2. No there is no sound
3. He is in a hallway
4. He is taking off his tie and shirt
5. Yes he does start in the hallway
6. He puts it in a closet
7. After he puts his stuff in the closet he grabs something out of the closet
8. No it is already open
9. He grabs a box and then starts walking toward the camera
10. No that is it from start to finish

Figure 1: A sample from the DSTC7-AVSD dataset.

alogue systems (Das et al. 2017; Mostafazadeh et al. 2017; de Vries et al. 2017). Recently, Pasunuru and Bansal (2018) introduced game-based video-context dialogue systems with multiple speaker interactions. However, there is a very little work in terms of using video-based multimodal information for this question answering based dialogue systems. To this end, a very recent work has introduced audio-visual scene-aware dialog dataset (Alamri et al. 2018), and our focus is on developing better models for this dataset.

It has been shown that attention plays an important role in improving the performance of end-to-end sequence-to-sequence based models for several tasks, e.g., machine translation, summarization, video captioning, etc. (Xu et al. 2015; Sutskever, Vinyals, and Le 2014). Also, previous work has shown that cross-modality attention is important for improving question answering based tasks, for example, reading comprehension, visual reasoning, and question answering (Seo et al. 2017; Tan and Bansal 2018). In this work, we use both the general attention and cross-modality attention for scene-aware dialogue and show that both are important for improving the performance of the model.

3 Models

In this section, we describe our various modeling approaches for the multimodal question and answering based dialogue systems. We will first introduce the task of video dialogue (AVSD). Next, we will present our basic sequence-to-sequence model where the answer decoder attends to multiple encoders at the same time. Later, we incorporate the cross-attention mechanism to align different modalities for better performance.

Task Formulation Let the video be represented as v and the corresponding frames in the video be $\{f_1, \ldots, f_m\}$, where m is the number of frames. The question is represented with a sequence of words by $q = \{w_1^q, \ldots, w_n^q\}$, and $a = \{w_1^a, \ldots, w_p^a\}$ is the generated answer, where n and p are the sentence lengths of question and answer, respectively. Let $b = \{w_1^b, \ldots, w_r^b\}$ be the summary sequence with length r. The task is to generate the answer a for the given question q using one or more modalities/information from video (v), chat history ($\{q_i, a_i\}$), and summary (b). Next, we describe the models for these various choices and combination of these modalities/information.

3.1 Seq2Seq with Attention Model

First, we describe the sequence-to-sequence (seq2seq) model with attention mechanism which is further used to fuse multiple modalities together to generate an answer for the given question. We use the standard seq2seq model similar to the standard machine translation encoder-decoder RNN model (Bahdanau, Cho, and Bengio 2015), where the RNN is based on Long Short-Term Memory (LSTM) units, which are good at memorizing long sequences due to forget-style gates. Let $x = \{x_1, x_2, ..., x_m\}$ be the input sequence and $y = \{y_1, y_2, ..., y_n\}$ be the target sequence. The conditional probability of the target sequence given the input sequence is parameterized with the chain rule:

$$P(y|x;\theta) = \prod_{t=1}^{n} p(y_t|y_{1:t-1}, x; \theta) \quad (1)$$

where θ denotes the model parameters. At each time step t, the decoder LSTM hidden state s_t is a non-linear recurrent function of the previous decoder hidden state s_{t-1}, the previous time-steps generated token y_{t-1}, and the context vector c_t, which is defined as follows:

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \quad (2)$$

where c_t is the weighted sum of the encoder hidden states $\{h_t\}^m$:

$$c_t = \sum_{i=i}^{m} \alpha_{t,i} h_i \quad (3)$$

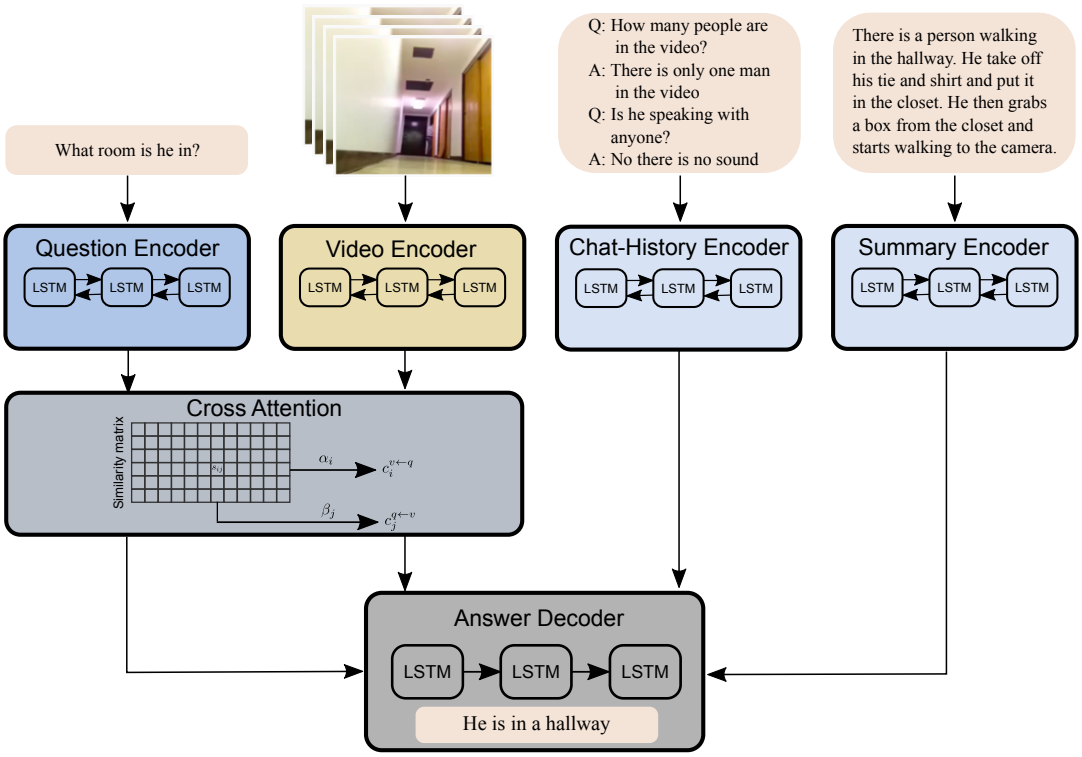


Figure 2: End-to-end multimodal dialogue model with video, summary, chat history, and question encoders (with cross-attention between question and video encoders).

These attention weights $\{\alpha_{t,i}\}$ are computed as following:

$$\alpha_{t,i} = \frac{\exp\left(e_{t,i}\right)}{\sum_{k=1}^{m} \exp\left(e_{t,k}\right)} \tag{4}$$

where the attention function is defined as follows:

$$e_{t,i} = V_a^T \tanh(W_a h_i + U_a s_{t-1} + b_a) \tag{5}$$

where V_a, W_a, U_a, and b_a are the trainable model parameters.

Video-based Model In this simpler model, we only use two encoders, one for the question encoding and another for the video encoding.[1] Next, we use two attention mechanisms (as described in the above section), one for the question and another for the video to allow the decoder separately attend to each of these encoders. At time step t of the decoder, let c_t^v and c_t^q be the context vectors from video and question encoders, respectively, then we concatenate these two vectors along with the embedding representation of the previously generated word and give it as input to the current time step of the answer decoder.

[1]Note that we only use the visual information from the video in our work, i.e., we do not use the audio information, which we will pursue in future work.

Video- and Chat History-based Model Chat history information contains the co-referencing and other useful information to answer the current question. Hence, additionally, we further add the chat history information to the model as another separate LSTM (see Fig. 2), where we also consider the attention from the chat history encoder along with the video and question encoder attention. Let c_t^v, c_t^h, and c_t^q be the context vectors from video, chat history, and question encoders respectively at time step t for the decoder, where we concatenate all these vectors along with the embedding representation of the previously generated word and give it as input to the current time step of the decoder.

Additional Summary-based Model The AVSD dataset also provides the text summary of the video. Some of the answers might already be available from this summary information. Hence, we also use the summary information by encoding it through a separate LSTM, and concatenate its context vector c_t^b with the context vectors of video, chat history, and question, along with the embedding representation of the previously generated work and give it as input to the answer decoder.

3.2 Cross-Attention Model

In order to learn a strong joint-aligned space between the video modality and the question, so has to only focus on video frames relevant to the question, we use bidirectional attention mechanism between video context and the ques-

Model	METEOR	CIDEr	BLEU-4	ROUGE-L
Video Only	12.43	95.54	8.83	34.23
Video + Chat History	14.13	105.39	10.58	36.54
Video + Chat History + Summary	14.94	112.80	11.22	37.53
Video + Chat History + Summary + Cross-attention	14.95	115.82	11.38	37.87

Table 1: Our models' performance on AVSD dataset's public test set. All of these models use the question information.

tion, following the previous work from reading comprehension (Seo et al. 2017). Let h_i^v and h_j^q represent the video encoder and question encoder hidden state representations at time steps i and j respectively. The bidirectional attention mechanism is based on a similarity score which is defined as follows:

$$S_{i,j}^{(v,q)} = w_s^T[h_i^v; h_j^q; h_i^v \odot h_j^q] \quad (6)$$

where w_s is a trainable parameter, $[x; y]$ represents concatenation, and $\odot$ represents the element-wise product. The attention distribution from question to video context is defined as $\alpha_{i:} = softmax(S_{i:})$, hence the question-to-video context vector is defined as $c_i^{v \leftarrow q} = \sum_j \alpha_{i,j} h_j^q$. Similarly, the attention distribution from the video context to question is defined as $\beta_{j:} = softmax(S_{:j})$, and the video to question context vector is defined as $c_j^{q \leftarrow v} = \sum_i \beta_{j,i} h_i^v$. Finally, we concatenate the hidden state and the corresponding context vector from the two modalities. $\hat{h}_i^v = [h_i^v; c_i^{v \leftarrow q}]$ is the final hidden state representation for the video encoder. Similarly, $\hat{h}_j^q = [h_j^q; c_j^{q \leftarrow v}]$ is the final hidden state representation for the question encoder. Let $\hat{c}_t^v$ and $\hat{c}_t^q$ be the new context vectors based on general attention from video and question encoders, respectively, at time step t of the decoder. Finally, we concatenate the context vectors from video ($\hat{c}_t^v$), question ($\hat{c}_t^q$), chat history (c_t^h), and summary (c_t^b), along with the embedding representation of the previously generated word and give it as input to the current time step of the decoder.

4 Results

4.1 Experimental Setup

Dataset We use Audio Visual Scene-Aware Dialog (AVSD) dataset (Alamri et al. 2018) for our video and chat context based question answering dialogue systems, where we use the visual and text features but not the audio features. This dataset has 11,156 dialogues, out of which 7,659 are used for training, 1,787 are used for validation, and 1,710 are used for testing. We use this official split as described above in all our experiments.

Evaluation Metrics For evaluation of our models, we use four diverse automatic evaluation metrics that are popular for image/video captioning and language generation in general: METEOR (Denkowski and Lavie 2014), BLEU-4 (Papineni et al. 2002), CIDEr-D (Vedantam, Lawrence Zitnick, and Parikh 2015), and ROUGE-L (Lin 2004). We use the standard evaluation toolkit (Chen et al. 2015) to obtain these four metrics. The AVSD dataset challenge also uses these four automatic metrics for the evaluation.

Training Details All training parameters are tuned on the validation set. We use a learning rate of 0.0001 with Adam optimizer (Kingma and Ba 2015). For video context, we unroll the encoder LSTM to a maximum of 400 time steps. We use a maximum of 200 time steps for the chat history encoder and 50 time steps for both question encoder and answer decoder. We use a batch size of 16. We use LSTM hidden size of 1024 dimension and word embedding size of 512 dimension. We use a vocabulary size of 5,398, replacing the less frequent words with UNK token. We clip the gradient to a maximum absolute value of 10.0. We apply a dropout with a probability of 0.5 to the vertical connections in LSTM.

4.2 Empirical Results

Video-only Context First, we performed experiments studying the importance of using only video information without any chat history for answering the given question. Table 1 shows that the performance of this model on various automatic evaluation metrics. For the rest of this section, we consider this model as baseline reference and show improvements to the model upon adding more modalities/contexts.

Chat History Context Next, we add the chat history context along with the video information to the question answering model enabling us to create a dialogue style model. Here, we encode the previous questions and answers as a single long sequence and encode with an LSTM-RNN. From Table 1, it is clear that adding the chat context significantly improves the performance of the model w.r.t. the baseline, showing that chat context is important in answering the questions.

Summary Context Also, given the summary context of the video, it might already have the answer to the given question. In such a scenario, using this information will be very helpful. We observe that using the summary context helps the model to perform better (see Table 1).

Cross-Attention Model Finally, we also consider the cross-attention between the video context and the question, because it is important to focus on the salient parts of the video which are relevant and useful for answering the given question. We model the cross-attention as described in Sec. 3.2 between the video context and question, and the results are as shown in Table 1. This result suggests that cross-attention plays an important role in aligning the video context with the given question.

4.3 Our Other Approaches and Analysis

Apart from the current approaches that we discussed above, we also experimented with various other techniques such as reinforcement learning based policy gradient approach, adding contextual embedding representations (ELMo), using external data, and pointer-generator model. For the rest of this section, we describe each of these approaches and discuss the possible reasons for their low impact on results.

Reinforcement Learning with Policy Gradient Rewards Policy gradient approaches allows us to directly optimize the model on the evaluation metrics instead of the cross-entropy loss, and has shown promising improvement in a number of generation tasks like machine translation, summarization, and image/video captioning (Ranzato et al. 2016; Paulus, Xiong, and Socher 2017; Rennie et al. 2016; Pasunuru and Bansal 2017b). In order to directly optimize the sentence-level test metrics (e.g, CIDEr), we use policy gradient approach, where our cross-entropy baseline model acts as an agent and interacts with the environment and samples a word at each time step of the decoder, thus forming an answer. At the end of this answer generation, we achieve a reward for this answer w.r.t. the reference answer. Our training objective is to minimize the negative expectation of this reward, which is defined as follows:

$$L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s)] \tag{7}$$

where w^s is the word sequence sampled from the model. For this, we use the REINFORCE algorithm (Williams 1992) where the gradients of this non-differentiable reward-based loss function are:

$$\nabla_\theta L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s) \cdot \nabla_\theta \log p_\theta(w^s)] \tag{8}$$

We approximate the above gradients via a single sampled word sequence (Ranzato et al. 2016).

In our experiments, we tested with various automatic evaluation metrics (CIDEr, ROUGE-L, and BLEU) as reward functions.[2] Unlike the video/image captioning datasets (MSR-VTT (Xu et al. 2016) and MS-COCO (Lin et al. 2014)) which have multiple references, here we are limited to a single answer for each question and hence the reward is noisy. We observe that ROUGE-L is relatively a better choice for the reinforcement learning approach. However, overall, we did not see much improvement with the RL approach and also readability of the answers went down.[3] The possible reason for these negative results are due to the nature of the dataset and the answers, since most of the answers in this dataset are yes/no type and flipping these words during the RL exploration do not bring much change in the phrase-matching metrics but visually its confusing to the model. Further, we explored these yes/no type questions by giving a reward of 1 when the reference answer and the generated answer are both positive (yes type) or both negative (no type), and a reward of 0 in all other cases.

[2]We did not try the METEOR as a reward, because METEOR calculation is very slow and hence the RL training process will be very slow.

[3]Note that we also tried the mixed cross-entropy and reinforce loss for better language modeling and fluency.

Contextualized ELMo Word Embeddings We also experimented with the deep contextualized words representations (ELMo) (Peters et al. 2018). First, we get the ELMo embeddings for the chat history, summary and question. Next, we use these embedding representations as input to their respective encoders. We did not see any improvement in the results, probably because our models on this video-chat dataset might not need this extra information or might have a mismatch with it.

Using External Data We also further experimented with using external data. We used the MSR-VTT (Xu et al. 2016) video captioning dataset, where given the video with no question, we want to generate the caption (otherwise answer). However, this approach also did not improve the overall performance of our final model. Again, the possible reason for this might be because of the different domains of these two datasets (MSR-VTT versus AVSD), or the fact that the MSR-VTT dataset is not a question-answer setup, or we may not have matched the exact sampling or I3D visual feature extraction setup of the AVSD data.

Pointer-Generator Copy Model Pointer mechanism (Vinyals, Fortunato, and Jaitly 2015) allows to directly copy the words from the input sequence (such as chat history or summary or question) during the answer generation. Pointer generator is a good fit to the AVSD dataset because lot of words in the question can also be present in the answer. For this pointer mechanism, we follow See, Liu, and Manning (2017), where we use a soft switch based on the generation probability p_g:

$$p_g = \sigma(W_g c_t + U_g s_t + W_g e w_{t-1} + b_g) \tag{9}$$

where $\sigma(\cdot)$ is a sigmoid function, and W_g, U_g, V_g, and b_g are trainable parameters. Here, ew_{t-1} is the previous time step output word embedding. The final word distribution is a weighted combination of the vocab distribution and attention distribution, where the weight is based on p_g. In our experiments, question-based pointer generator did not improve the performance of our final model. We also tried joint pointer from question and summary, since the answer is usually a combination of the question words and an answer word from the summary. This performed better than the question-based pointer, but not over the non-pointer model. This is probably because of our strong dual attention mechanism and also omitting the less frequent words during the training.

In future work, we plan to further analyze and improve these promising approaches with specific RL rewards, contextualized large language models, and joint copy models.

5 Conclusion

We presented an end-to-end multimodal dialogue system with dual attention (general attention and cross-attention). We showed the usefulness of each of the modalities for improving the model performance. We further discussed various other approaches for improving the performance of the model and the possible reasons for their negative results.

Acknowledgments

We thank the reviewers for their helpful comments. This work was supported by DARPA (YFA17-D17AP00022), and faculty awards from Google, Facebook, and Salesforce. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the funding agency.

References

Alamri, H.; Cartillier, V.; Lopes, R. G.; Das, A.; Wang, J.; Essa, I.; Batra, D.; Parikh, D.; Cherian, A.; Marks, T. K.; et al. 2018. Audio visual scene-aware dialog (avsd) challenge at dstc7. *arXiv preprint arXiv:1806.00525*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *CVPR*, 2425–2433.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Cassell, J. 1999. Embodied conversation: integrating face and gesture into automatic spoken dialogue systems.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *CVPR*.

de Vries, H.; Strub, F.; Chandar, S.; Pietquin, O.; Larochelle, H.; and Courville, A. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2680–8.

Johnston, M.; Bangalore, S.; Vasireddy, G.; Stent, A.; Ehlen, P.; Walker, M.; Whittaker, S.; and Maloor, P. 2002. Match: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 376–383. Association for Computational Linguistics.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. In *EMNLP*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *ECCV*, 740–755. Springer.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Mostafazadeh, N.; Brockett, C.; Dolan, B.; Galley, M.; Gao, J.; Spithourakis, G. P.; and Vanderwende, L. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 311–318. Association for Computational Linguistics.

Pasunuru, R., and Bansal, M. 2017a. Multi-task video captioning with video and entailment generation. In *ACL*.

Pasunuru, R., and Bansal, M. 2017b. Reinforced video captioning with entailment rewards. In *EMNLP*.

Pasunuru, R., and Bansal, M. 2018. Game-based video-context dialogue. In *EMNLP*.

Paulus, R.; Xiong, C.; and Socher, R. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2016. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 3776–3784.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*, 3104–3112.

Tan, H., and Bansal, M. 2018. Object ordering with bidirectional matchings for visual reasoning. In *AAAI*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015.

CIDEr: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015. Sequence to sequence-video to text. In *CVPR*, 4534–4542.

Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, 2692–2700.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 5288–5296.

Yu, L.; Bansal, M.; and Berg, T. 2017. Hierarchically-attentive rnn for album summarization and storytelling. In *EMNLP*, 966–971.