Leveraging Topics and Audio Features with Multimodal Attention for Audio Visual Scene-Aware Dialog

Shachi H Kumar
Intel Labs
Anticipatory Computing Lab
Santa Clara, CA 95054
shachi.h.kumar@intel.com

Eda Okur
Intel Labs
Anticipatory Computing Lab
Hillsboro, OR 97124
eda.okur@intel.com

Saurav Sahay
Intel Labs
Anticipatory Computing Lab
Santa Clara, CA 95054
saurav.sahay@intel.com

Jonathan Huang
Intel Labs
Anticipatory Computing Lab
Santa Clara, CA 95054
jonathan.huang@intel.com

Lama Nachman
Intel Labs
Anticipatory Computing Lab
Santa Clara, CA 95054
lama.nachman@intel.com

Abstract

With the recent advancements in Artificial Intelligence (AI), Intelligent Virtual Assistants (IVA) such as Alexa, Google Home, etc., have become a ubiquitous part of every home. Currently, such IVAs are mostly audio-based, but going forward, we are witnessing a confluence of vision, speech and dialog system technologies that are enabling the IVAs to learn audio-visual groundings of utterances. This will enable agents to have conversations with users about the objects, activities and events surrounding them. In this work, we present three main architectural explorations: 1) investigating 'topics' of the dialog as an important contextual feature for the conversation, 2) exploring several multimodal attention mechanisms during response generation and 3) incorporating an end-to-end audio classification ConvNet, AclNet, into our architecture. We discuss detailed analysis of the experimental results and show that our model variations outperform the baseline system presented for the Audio Visual Scene-Aware Dialog (AVSD) task.

1 Introduction

We are witnessing a confluence of vision, speech and dialog system technologies that are enabling the IVAs to learn audio-visual groundings of utterances and have conversations with users about the objects, activities and events surrounding them. Recent progress in visual grounding techniques [3, 6] and audio understanding [7] are enabling machines to understand shared semantic concepts and listen to the various sensory events in the environment. With audio and visual grounding methods [17, 8], end-to-end multimodal Spoken Dialog Systems (SDS) [14] are now being trained to meaningfully communicate in natural language about the real dynamic audio-visual sensory world around us. In this work, we explore the role of 'topics' of the dialog as the context of the conversation along with multimodal attention into an end-to-end audio-visual scene-aware dialog system architecture. We also incorporate an end-to-end audio classification ConvNet, AclNet, into our models. We develop and test our approaches on the Audio Visual Scene-Aware Dialog (AVSD) dataset [1, 2] released as part of the 7th Dialog System Technology Challenges (DSTC7) task, showing that some of our model variations outperform the AVSD baseline model [9].

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

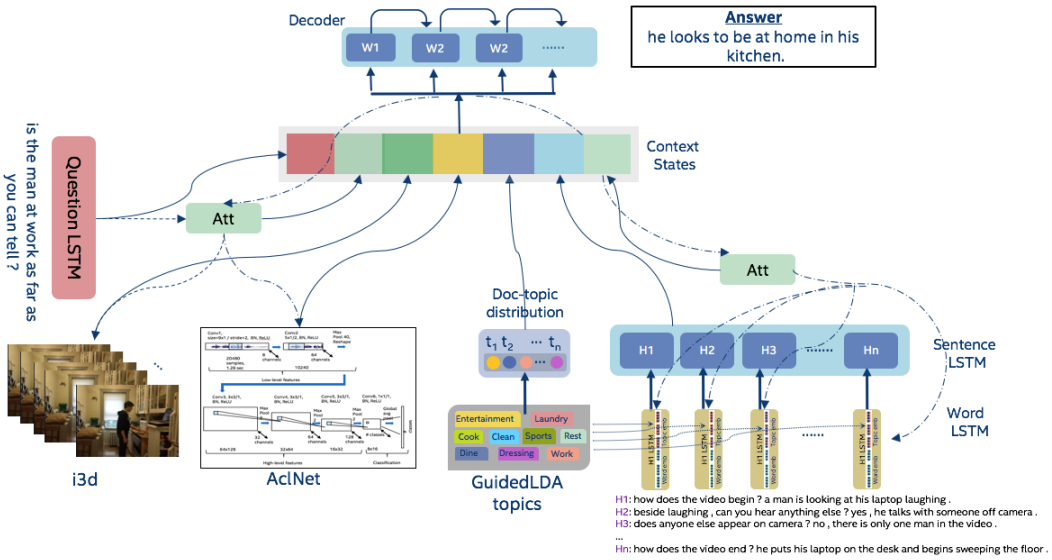

Figure 1: Architecture of Our System

2 Model Description

In this section, we describe the main architectural explorations of our work as shown in Figure 1.

Adding Topics of Conversations: Topics form a very important source of context in a dialog. Charades dataset [15] contains videos on common household activities such as watching TV, eating, cleaning, using a laptop, sleeping, and so on. We train Latent Dirichlet Allocation (LDA) [5] and Guided LDA [11] models on questions, answers, QA pairs, captions and dialog history. Since we are interested in identifying domain-specific topics such as entertainment, cooking, cleaning, resting, etc., we use Guided LDA to generate topics via seed words. A detailed list of sample seed words provided to Guided LDA for the 9-topics configuration is presented in Table 1. These seed words are constructed by identifying a set of most common nouns (objects), verbs, scenes, and actions from the Charades dataset analysis [15]. Generated topic distributions are incorporated as features into our models or used to learn topic embeddings.

Attention Explorations: We explore several configurations of the attention-based model where at every step, the decoder attends to the dialog history representations and audio/video (AV) features to selectively focus on relevant parts of the dialog history and AV. We calculate the attention weights [4, 16] corresponding to every dialog history turn, multimodal features and the decoder representation, and apply the weights to the history and multimodal features to compute the relevant representations. These help create a combination of the dialog history and multimodal context that is richer than the single context vectors of the individual modalities. We append the input encoding along with the AV multimodal feature encodings and pass that to the decoder LSTM for learning the output encodings.

Table 1: Sample of Seed Words for 9 Topics

Topic	Seed Words
Entertainment/LivingRoom	living, room, recreation, garage, basement, entryway, television, tv, phone, laptop, sofa, chair, couch, armchair, seat, picture, sit ...
Cooking/Kitchen	kitchen, pantry, food, water, dish, sink, refrigerator, fridge, stove, microwave, toaster, kettle, oven, stewpot, saucepan, cook, wash ...
Eating/Dining	dining, room, table, chair, plate, fork, knife, spoon, bowl, glass, cup, mug, coffee, tea, sandwich, meal, breakfast, lunch, dinner ...
Cleaning/Bath	bathroom, hallway, entryway, stairs, restroom, toilet, towel, broom, vacuum, floor, sink, water, mirror, cabinet, hairdryer, clean ...
Dressing/Closet	walk-in, closet, clothes, wardrobe, shoes, shirt, pants, trousers, skirt, jacket, t-shirt, underwear, sweatshirt, coat, rack, dress, wear ...
Laundry	laundry, room, basement, clothes, clothing, cloth, basket, bag, box, towel, shelf, dryer, washer, washing, machine, do, wash, hold ...
Rest/Bedroom	bedroom, room, bed, pillow, blanket, mattress, bedstand, nightstand, commode, dresser, bedside, lamp, nightlight, night, light, lie ...
Work/Study	home, office, den, workroom, garage, basement, laptop, computer, pc, screen, mouse, keyboard, phone, desk, chair, light, work, study ...
Sports/Exercise	recreation, room, garage, basement, hallway, stairs, gym, fitness, floor, bag, towel, ball, treadmill, bike, rope, mat, run, walk, exercise ...

Audio Feature Explorations: We used an end-to-end audio classification ConvNet, called AclNet [10]. AclNet takes raw, amplitude-normalized 44.1 kHz audio samples as input, and produces classification output without the need to compute spectral features. AclNet is trained using the ESC-50 [13] corpus, a dataset of 50 classes of environmental sounds organized in 5 semantic categories (animals, interior/domestic, exterior/urban, human, natural landscapes).

3 Dataset

We use the dialog dataset consisting of conversations between two parties about short videos (from Charades human action dataset [15]), which was released as part of the AVSD challenge track of DSTC7 [1]. The two parties in the conversation discuss about events in the video, where one plays the role of a questioner and the other is the answerer [2]. For the results presented in this work, we use the official training and validation sets to train and optimize our models, which are evaluated on the official test set. Table 2 shows the distribution of DSTC7 AVSD data across different sets.

Table 2: Audio Visual Scene-Aware Dialog Dataset

	Training	Validation	Test
# of Dialogs	7,659	1,787	1,710
# of Turns	153,180	35,740	13,490
# of Words	1,450,754	339,006	110,252

4 Experiments and Results

Topic Modeling Experiments: We use separate topic models trained on questions (Q), answers (A), QA pairs, captions (C), history and history+captions to generate topics for samples from each category. The generated topic vectors are incorporated as features for questions and dialog history. The question topics are added to the decoder state directly. In one variation, the dialog history topics (QA and C, or all topics) are copied to all the decoder states directly. In another variation, the dialog history topics are added as features to the history encoder LSTM (HLSTM). We learn topic embeddings from topics generated for the questions, QA pairs and captions as well. In addition, GloVe embeddings [12] (200-dim) are incorporated with fine-tuning for questions and history.

Table 3: Topic Modeling Experiments

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr
Baseline	0.621	0.480	0.379	0.305	0.217	0.481	0.733
GuidedLDA (Q,QA,C)	0.614	0.475	0.374	0.299	0.215	0.474	0.695
GuidedLDA (Q,QA,C) + GloVe	0.629	0.491	0.390	0.315	0.219	0.484	0.731
StandardLDA (All topics)	0.621	0.480	0.380	0.306	0.221	0.483	0.753
GuidedLDA (All topics)	0.619	0.480	0.378	0.303	0.217	0.476	0.701
GuidedLDA (All topics) + GloVe	**0.631**	**0.493**	**0.390**	**0.315**	**0.224**	**0.492**	**0.773**
HLSTM with topics	**0.627**	**0.489**	**0.387**	**0.311**	**0.218**	0.480	0.723
Topic Embeddings	0.623	0.488	0.387	0.311	0.217	0.479	0.701
Topic Embeddings + GloVe	**0.632**	**0.499**	**0.402**	**0.329**	**0.223**	**0.488**	**0.762**

Table 4: Topic Model Performances on Binary/Non-binary Answers

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr
Binary							
Baseline	0.626	0.479	0.371	0.294	0.214	0.474	0.676
GuidedLDA (Q,QA,C) + GloVe	0.616	0.476	0.374	0.301	0.215	0.474	0.673
GuidedLDA (All topics) + GloVe	0.629	0.486	0.381	0.306	0.223	0.488	0.728
HLSTM with topics	0.623	0.480	0.375	0.297	0.214	0.473	0.696
Topic Embeddings + GloVe	**0.635**	**0.497**	**0.398**	**0.325**	**0.224**	**0.491**	**0.746**
Non-binary							
Baseline	0.624	0.486	0.387	0.312	0.219	0.482	0.726
GuidedLDA (Q,QA,C) + GloVe	**0.633**	0.497	0.396	0.320	0.220	0.487	0.759
GuidedLDA (All topics) + GloVe	0.632	0.495	0.394	0.318	**0.225**	**0.494**	**0.796**
HLSTM with topics	0.629	0.492	0.392	0.316	0.220	0.483	0.740
Topic Embeddings + GloVe	0.630	**0.499**	**0.403**	**0.330**	0.223	0.487	0.776

Table 3 compares the baseline model [9] with the topic-based model variations. GuidedLDA (All topics) + GloVe performs better than the baseline in all metrics. Adding topics as part of the HLSTM also slightly improves performance compared to the baseline. Learning topic embeddings along with the word embeddings (+GloVe fine-tuning) achieves the best performance in most of the metrics (BLEU-scores), whereas GuidedLDA (All topics) + GloVe succeeds in other metrics. We also evaluated topic-based models on subsets having binary and non-binary answers. As shown in Table 4, for the non-binary subset, all topic-based models perform better than the baseline in all metrics, which shows that these models can generate better responses for the more complex, non-binary answers.

Attention Experiments: The baseline architecture [9] only leverages the last hidden state information from the sentence LSTM in the dialog history encoder. In our experiments, we have modified the baseline architecture and added attention layer for the answer decoder to leverage information directly from the dialog history LSTMs and multimodal audio/video features, with 4 different configurations described below. To evaluate the performance of attention solely for questions that could benefit from dialog history, we isolate the questions containing coreferences. Table 5 shows the performance of our models on this coreference-subset. To compare the results at a more semantic level, we further performed quantitative analysis on dialogs that contained binary answers. We evaluate our models on their ability to predict these binary answers correctly (using precision, recall and F1-scores) as presented in Figure 2. The results show that the configuration where decoder attends to all of the sentence-LSTM output states performs better than the baseline.

Table 5: Decoder Attention over Dialog History and Multimodal Features on Coreference-subset

	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	METEOR	$ROUGE_L$	CIDEr
Baseline	0.611	0.475	0.374	0.297	0.210	0.467	0.704
Word LSTM (all output states)	0.594	0.447	0.336	0.262	0.190	0.432	0.553
Word LSTM (last hidden states)	**0.627**	**0.485**	**0.379**	0.297	0.208	**0.468**	0.701
Sentence LSTM (all output states)	**0.619**	**0.484**	**0.384**	**0.307**	**0.213**	**0.472**	**0.749**
Sentence LSTM (all outputs) + AV	0.598	0.464	0.360	0.284	0.209	0.458	0.685

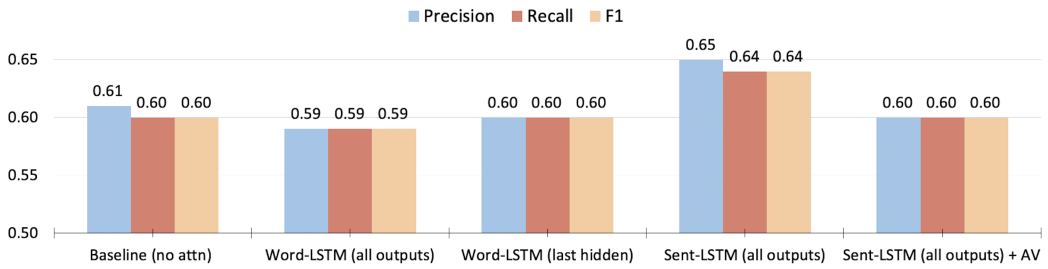

Figure 2: Precision, Recall, F1-scores for Attention Experiments on Coreference-Binary-subset

1. *Attention on Dialog History Word LSTMs, all output states*: In this configuration, we remove the sentence level dialog history LSTM and the decoder computes the attention scores directly between the decoder state and the word level output states for all dialog history. We first padded the Word LSTM outputs from Dialog History LSTMs (see Word LSTM in History in Figure 1) to the maximum sentence length of all the sentences. We summed up all the attention scores from each of the sentence context vectors with the query decoder state. Using this kind of attention, we had hoped that the system could remember answers that were already given (directly or indirectly) in the earlier turns of the dialog. Directly attending to the output states of the word LSTMs in the dialog history encoder did not perform well compared to the baseline. This attention mechanism possibly attended to way more information than needed.
2. *Attention on Dialog History Word LSTMs, last hidden states*: This configuration is similar to the previous configuration with the difference that we only use the last hidden state output representations of the word LSTMs corresponding to the different turns in the dialog. Simpler than the previous setup, we stack up the hidden states from the history sentences for attention computation. This configuration performed better that the baseline on the coreference-subset in most of the evaluation metrics.

3. *Attention on Sentence LSTM, all output states*: The baseline architecture only leverages the last hidden state information from the sentence LSTM in the dialog history encoder. Instead, we extract the output states from all timesteps of the LSTM corresponding to n turns of the dialog history. This variation helps the decoder consider all the dialog turn compressed sentence representations via the attention mechanism. This model performed better than the baseline in all metrics on both coreference-subset (Table 5) and binary answers (Figure 2).
4. *Attention on Sentence LSTM, all output states and Multimodal Audio/Video Features*: This configuration is similar to the last one with the difference that we add multimodal audio/video features as additional state to the attention module. This mechanism allows the decoder to selectively focus on the multimodal features along with the dialog history sentences. This configuration did not really help improve the evaluation metrics compared to the baseline.

Audio Experiments: Table 6 shows the comparison of the baseline (B) model without audio features, B+VGGish (provided as a part of the AVSD task), and B+AclNet features. We investigate the effects of audio features on the overall dataset as well as on the subset of audio-related questions. We observe that B+AclNet shows improved performances as compared to the baseline and B+VGGish, both on the overall dataset and audio-related subset. Table 7 presents a qualitative analysis of the addition of the VGGish and AclNet features to the baseline model. For these audio-related examples (e.g., 'oscillating', 'eating', 'sneeze'), baseline and B+VGGish models generate irrelevant responses, whereas the answers generated by B+AclNet are in accordance with the ground truth.

Table 6: Audio Feature Performances on Overall vs. Audio-related Questions

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr
Overall							
Baseline (B)	0.621	0.480	0.379	0.305	0.217	0.481	0.733
B + VGGish	0.622	0.487	0.389	0.315	0.216	0.481	0.732
B + AclNet	**0.625**	**0.491**	**0.391**	**0.316**	**0.218**	**0.484**	**0.736**
Audio-related							
Baseline (B)	**0.666**	0.526	0.413	0.329	0.230	0.504	0.767
B + VGGish	0.657	0.519	0.408	0.324	0.230	0.500	0.754
B + AclNet	0.659	**0.527**	**0.424**	**0.348**	**0.236**	**0.507**	**0.796**

Table 7: Audio Examples (VGGish vs. AclNet)

Question:	***is the fan oscillating ?***	***is he eating something ?***	***how many times does she sneeze ?***
Ground Truth	*the fan is on but is still .*	*yes he appears to be eating something*	*she sneezes a few times in the video .*
Baseline	*yes it is very well lit*	*no he is not drinking anything*	*can only see her face*
Baseline + VGGish	*no don 't see any music*	*no he is not drinking anything*	*she laughs at the end of the video*
Baseline + AclNet	*no it is hard to tell*	*yes he is eating sandwich*	*she sneezes at the end of the video*

5 Conclusion

In this paper, we present our explorations towards architectural extensions for contextual and multimodal end-to-end audio-visual scene-aware dialog system. We incorporate context of the dialog in the form of topics, investigate various attention mechanisms to enable the decoder to focus on relevant parts of the dialog history and audio/video features, and incorporate audio features from an end-to-end audio classification architecture, AclNet. We validate our approaches on the AVSD dataset and show that some of the explored techniques yields in improved performances compared to the baseline system for AVSD task.

References

[1] H. AlAmri, V. Cartillier, R. G. Lopes, A. Das, J. Wang, I. Essa, D. Batra, D. Parikh, A. Cherian, T. K. Marks, and C. Hori. Audio visual scene-aware dialog (AVSD) challenge at DSTC7. *CoRR*, abs/1806.00525, 2018. URL `http://arxiv.org/abs/1806.00525`.

[2] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Dec 2015. doi: 10.1109/ICCV.2015.279.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL `http://arxiv.org/abs/1409.0473`.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001]*. MIT Press, 2001. URL `http://papers.nips.cc/paper/2070-latent-dirichlet-allocation`.

[6] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra. Visual dialog. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.121. URL `http://dx.doi.org/10.1109/CVPR.2017.121`.

[7] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference*. IEEE.

[8] 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. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi: 10.1109/iccv.2017.450. URL `http://dx.doi.org/10.1109/ICCV.2017.450`.

[9] 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*, abs/1806.08409, 2018. URL `http://arxiv.org/abs/1806.08409`.

[10] J. J. Huang and J. J. A. Leanos. Aclnet: Efficient end-to-end audio classification cnn. *arXiv preprint arXiv:1811.06669*, 2018.

[11] J. Jagarlamudi, H. D. III, and R. Udupa. Incorporating lexical priors into topic models. In W. Daelemans, M. Lapata, and L. Màrquez, editors, *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics*. The Association for Computer Linguistics, 2012. ISBN 978-1-937284-19-0. URL `http://aclweb.org/anthology/E/E12/E12-1021.pdf`.

[12] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[13] K. J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, 2015. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL `http://dl.acm.org/citation.cfm?doid=2733373.2806390`.

[14] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784, 2016.

[15] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Proceedings*, 2016. ISBN 978-3-319-46447-3. doi: 10.1007/978-3-319-46448-0_31. URL `https://doi.org/10.1007/978-3-319-46448-0_31`.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need`.

[17] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016.