TMT: A Transformer-based Modal Translator for Improving Multimodal Sequence Representations in Audio Visual Scene-aware Dialog

Wubo Li, Dongwei Jiang, Wei Zou, Xiangang Li

Didi Chuxing, Beijing, China

{liwubo, jiangdongwei, zouwei, lixiangang}@didiglobal.com

Abstract

Audio Visual Scene-aware Dialog (AVSD) is a task to generate responses when discussing about a given video. The previous state-of-the-art model shows superior performance for this task using Transformer-based architecture. However, there remain some limitations in learning better representation of modalities. Inspired by Neural Machine Translation (NMT), we propose the Transformer-based Modal Translator (TMT) to learn the representations of the source modal sequence by translating the source modal sequence to the related target modal sequence in a supervised manner. Based on Multimodal Transformer Networks (MTN), we apply TMT to video and dialog, proposing MTN-TMT for the video-grounded dialog system. On the AVSD track of the Dialog System Technology Challenge 7, MTN-TMT outperforms the MTN and other submission models in both Video and Text task and Text Only task. Compared with MTN, MTN-TMT improves all metrics, especially, achieving relative improvement up to 14.1% on CIDEr.

Index Terms: multimodal learning, audio-visual scene-aware dialog, neural machine translation, multi-task learning

1. Introduction

Recently, a considerable amount of literature has been published on multimodal learning of visual and language. These studies concern the tasks of media description and question answering, such as video caption [1, 2, 3], video question answering [4, 5, 6] and video dialog [7]. Audio Visual Scene-aware Dialog (AVSD) is the task of generating a response for a question with a given scene, video, audio, and the history of previous turns in the dialog. The AVSD is considered one of the most challenging tasks because the system needs to recognize the history of dialog along with the visual and acoustic data for accurately answering the question [8].

One of the challenges in the AVSD task: acquiring accurate representations of the multiple modalities. To tackle this, some previous studies focus on pre-training a modal extractor. In [9, 10, 11], features are extracted from the fix-frame video and audio by pre-trained I3D [12], VGG [13] and 3D ResNeXt [14]. Sanabria et al. [10] use 2000 hours of how-to videos [15] to pre-train a video modal extractor and treat the video-grounded dialog task as a video caption task to generate answers. In the textual question answering task, lots of studies use questions to figure out useful information from documents [16, 17]. This method can be transferred to the multimodal dialog system as well. The work in [18] uses a simple but effective 1×1 convolution to fuse multimodal features and proposes a multi-stage fusion mechanism to thoroughly understand the question. Inspired by FiLM, FiLM Attention Hierarchical Recurrent Encoder-Decoder (FA-HERD) [19] learns the representation of multiple modalities by using FiLM blocks to condition video and audio on the embeddings of the current question.

[20] proposed query-aware auto-encoder to obtain accurate representations from visual and audio modalities.

Recently, the evidence [21] reports that translating cross-modalities is beneficial to capture correlated signals across asynchronous modalities. Along with this motivation, we try to learn the representation of modality by the consistency of related modalities.

In this paper, we propose the Transformer-based Modal Translator (TMT) for learning representation of modalities. In the TMT, the source modal sequence is translated to other related modal sequences in a supervised learning manner. The key to our method is that we effectively make use of correlative multiple modalities to express source modality. Based on MTN, we apply TMT to video and dialog, proposing MTN-TMT for the video-grounded dialog system. We evaluated MTN-TMT on the AVSD track of the 7th Dialog System Technology Challenge, which generates dialog responses considering multiple modalities. MTN-TMT outperforms the MTN and other submission models in both Video and Text task and Text Only task. Compared with MTN, MTN-TMT improves all metrics, especially, achieving relative improvement up to 14.1% on CIDEr.

2. Related Work

The work relevant to ours is MTN proposed by Le et al. [20], which is a Transformer-based encoder-decoder framework that has several attention blocks to incorporate multiple modalities such as video, audio, and text. The MTN is composed of three major components: encoder, decoder, and query-aware auto-encoder. In the encoder, the text sequence and video features are mapped to a sequence of continuous representations $z = (z_1, ..., z_n) \in \mathbb{R}^d$, f_a and f_v, where d is the dimension of word embeddings. The aim of query-aware auto-encoder is to learn the representations of non-textual modalities that are relevant to the query, such as audio and visual. Decoder consists of a stack of Transformer layers to fuse all modalities and generate an answer $y = (y_1, ..., y_m)$ including an end-of-sentence token (eos). We refer the reader to [20] for a more detailed explanation of the MTN. One of the contributions of MTN is that replacing RNN with Transformer enhances the context-dependence of modality. Another contribution is that they propose a query-aware attention encoder to learn the representations of non-text modalities. Compared with query-aware auto-encoder, TMT learns representations of modalities by other related modalities instead of query.

Multimodal Transformer (MulT) [21] is another work related to ours which uses decoder Transformer in NMT to align cross-modalities. They hypothesize a good way to fuse cross-modal information is providing a latent adaptation across modalities. For human multimodal sentiment analysis and human emotion analysis tasks, they use six Transformer-based translators to align each pair of modalities among visual, au-

dio and language. Similar to MulT, TMT uses a machine-translation-like method to translate the source modality to the related target modality. In addition, we treat target sequence modality as a label for supervised learning.

3. Our Approach

3.1. Transformer-based modal translator

Motivated by that translating related modalities could deeply capture the relevance between modalities, we introduce Transformer-based Modal Translator (TMT) to translate the source modal sequence to the related target modal sequence to learn better source modal sequence. As shown in Figure 1, TMT is made up of multi-layer Transformer encoder (left) and Transformer decoder (right). In this work, TMT with M layers indicates that the depth of Transformer encoder and Transformer decoder are both M. Given the source modal sequence $X = \{x_1, x_2, ..., x_T\}$ and target modal sequence $Y = \{y_1, y_2, ..., y_N\}$, where T and N is the length of the source modality and target modality respectively. TMT mainly consists of two steps. At the first step, a Transformer encoder is used to encode the source modal sequence. Then, a Transformer decoder is used to translate the source modal sequence to the target modal sequence. In order to learn the mapping between the source modal sequence and the target modal sequence, we use supervised learning to fit the output sequence $\widehat{Y} = \{\widehat{y_1}, \widehat{y_2}, ..., \widehat{y_N}\}$ of the TMT and the target modal sequence Y. When the target modality is text, we employ cross-entropy as the loss function. As for dense modalities, such as speech and images, we use the $L1$ loss or similarity loss [22].
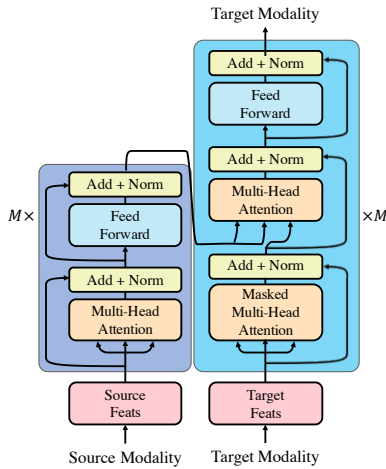

Figure 1: *Overall Architecture of TMT.*

3.2. MTN-TMT

Recently, [20] proposed Multimodal Transformer Networks (MTN), which is the state-of-the-art system for DSTC7-AVSD. Based on MTN, we propose a novel MTN-TMT applying TMT to video and dialog. Figure 2 shows the overall architecture of MTN-TMT, which includes encoder, auto-encoder, decoder, video-caption translator and dialog-summary translator. We follow the pre-processing and structure of MTN [20], except adding TMT-based video-caption translator and dialog-summary translator after encoder for learning representations of video and dialog. Next, we will briefly introduce the video-

caption translator, dialog-summary translator and loss function.

3.2.1. Video-caption translator

We consider caption as a relevant modality to video which summarizes the video's content. To enhance the representations of video, we propose a TMT-based video-caption translator to translate the video to caption. Figure 2 shows the data flow from encoder to the video-caption translator. We feed visual representations f_v as source modality and tokenized caption z_{cap} as target modality into video-caption translator. Video-caption translator consists of M layers of TMT. The output of the video-caption translator are caption-related visual representations f_v^{cap}.

3.2.2. Dialog-summary translator

Although dialog history usually contains a wealth of details, it remains a challenge to capture useful information from a long dialog history. Given that dialog history and summary are describing the same content in different ways, we introduce a dialog-summary translator to translate the dialog history into a summary to enhance representations of dialog history. Dialog-summary translator consists of N layers of TMT. As shown in Figure 2, the input modalities of the dialog-summary translator are sentence-level dialog history z_{hiss} and summary representations z_{sum}. Inspired by work in [23], before translating, we use hierarchical Transformer layer to preprocess tokenized dialog history sequences z_{his} to get the sentence-level dialog history z_{hiss}. The hierarchical Transformer consists of N Transformer encoder layer [24]. The sentence-level dialog history z_{hiss} are outputs of the eos tokens in dialog history. Then summary representations z_{sum} and the sentence-level dialog history z_{hiss} are fed into dialog-summary translator to generate summary-related dialog representations z_{hiss}^{sum}.

3.2.3. Loss

Given the dialog history (H), question (Q), video features (video V and audio A), video caption (C) and dialog summary (S), we use the log-likelihood as the objective function for target sequences answer (Ans), video caption (C) and dialog summary (S). The log-likelihood function consists of three parts respectively:

$$\begin{aligned} L =& L(Ans) + \alpha L(C) + \beta L(S) \\ =& logP(Ans|H, Q, V, A, C, S) + \alpha logP(C|V) + \quad (1) \\ & \beta logP(S|H). \end{aligned}$$

where α and β are learning weights for the video-caption translator and dialog-summary translator in loss function.

4. Experiments

4.1. Datasets

We evaluated MTN-TMT on the Audio Visual Scene-aware Dialog track of the 7th Dialog System Technology Challenge

Table 1: *DSTC7 Audio Visual Scene-aware Dialog Dataset*

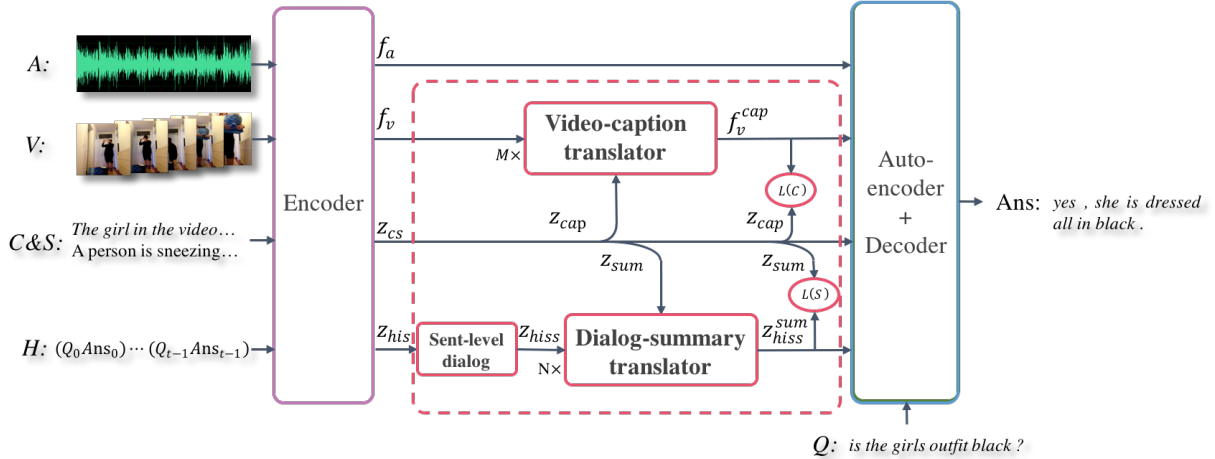|	Training	Dev	Test
num of Dialogs	7,659	1,787	1,710
num of Turns	153,180	35,740	13,490
num of Words	1,450,754	339,006	110,252

Figure 2: *Overall Architecture of MTN-TMT. Video-caption translator and dialog-summary translator are highlighted by red dotted rectangle. The A, V, C, S, H, Q and Ans stand for audio, video, caption, summary, dialog history, question, and answer, respectively.*

Table 2: *Automatic evaluation metrics on DSTC7-AVSD. The VCT and DST stand for video-caption translator and dialog-summary translator, respectively.*

Description	BLEU-4	METEOR	ROUGE-L	CIDEr
Text Only				
MTN [20]	0.307	0.283	0.553	0.973
MTN-TMT	**0.346**	**0.289**	**0.571**	**1.107**
Video and Text				
MTN[20]	0.392	0.278	0.571	1.128
MTN+VCT	0.379	0.286	0.578	1.146
MTN+DST	0.398	0.291	0.583	1.176
MTN-TMT	**0.402**	**0.293**	**0.587**	**1.190**

Table 3: *An ablation study on the influence of supervised learning on TMT. WD indicates learning weights decay mentioned in section 4.4*

Description	BLEU-4	METEOR	ROUGE-L	CIDEr
MTN [23]	0.392	0.278	0.571	1.128
MTN-TMT (WD)	0.375	0.286	0.574	1.135
MTN-TMT	**0.393**	**0.296**	**0.587**	**1.169**

(DSTC7-AVSD) proposed on [23]. This dataset consists of Q&A conversations regarding short videos obtained from two Amazon Mechanical Turk (AMT) workers, who discuss events in a video. In each dialog, one of the workers takes the role of an answerer who had already watched the video. The answerer replies to questions asked by another AMT worker, the questioner. The dataset contains 9,848 videos taken from CHARADES, which is a multi-action dataset including 157 action categories [25].

We used 7,659 dialogs for the training set and 1787 dialogs for dev set, and evaluated our model on the official test set which contains 1710 dialogs [23]. Table 1 summarizes the dataset.

4.2. Metrics

In the DSTC7-AVSD, the metrics are commonly used in natural language process tasks, such as BLEU [26], METEOR [27], ROUGE-L [28] and CIDEr [29]. Note that higher scores are better. The scores of the metrics are obtained by the toolkit provided by the organizer.

4.3. Experiment Settings

Except for the hyper-parameters mentioned below, we follow the training settings of MTN [20] in DSTC7-AVSD. We use dropout [30] with keeping probability 0.5, and warm up the schedule with $warmup\_step$ 13000 for modifying the learning rate [24]. We adopt the Adam optimizer [31] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. In all experiments, we select a batch

size of 32 and train models up to 35 epochs. While training MTN-TMT, we use the grid search method among hyperparameters α and β on dev set. For all models, we evaluated on the test set of the DSTC7-AVSD by loading models with the lowest perplexity on the dev set. We have carried out multiple experiments, and the reported metrics of each model are the average of three experiments.

We evaluate MTN-TMT model in DSTC7-AVSD 's Video and Text and Text Only tasks [23]. In Video and Text task, we use multiple modal input, including dialog history, audio, visual, caption and summary. Text Only task allows only dialog history and caption as input. Compared with the caption, we consider that summary is more relevant to dialog history. As a result, in the Text Only task, we add the summary as an extra input.

4.4. Results

Text Only task. We first evaluate MTN-TMT in Text Only task. The upper part of the Table 2 shows the result of MTN and MTN-TMT. For Text Only task, we only add dialog-summary translator to MTN and note it as MTN-TMT. We observe MTN-TMT outperforms MTN, where BLEU-4, METEOR, ROUGE-L, and CIDEr are improved by 0.039, 0.006, 0.018 and 0.134.

Video and Text task. Next, we evaluated MTN-TMT in the Video and Text task. The result is shown in the bottom part of Table 2. In this task, MTN-TMT improves upon the MTN by 0.004, 0.015, 0.016 and 0.062 on BLEU-4, METEOR, ROUGE-L, and CIDEr. Empirically, we find that MTN-TMT converges more stable to better results after 20 epochs when compared to MTN (see Figure 3). We speculate that such performance improvement occurs because TMT introduces

Table 5: *Automatic evaluation metrics of the submission models to the DSTC7-AVSD. Note that the evaluation of HAN on one reference is not available.*

Description	Official 1 reference				Official 6 references			
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
Text Only								
HAN [10]	-	-	-	-	**0.376**	0.264	0.554	1.076
MTN [20]	0.115	0.173	0.358	1.141	0.307	0.283	0.553	0.973
MTN-TMT (Ours)	**0.131**	**0.174**	**0.368**	**1.302**	0.346	**0.289**	**0.571**	**1.107**
Video and Text								
HAN [10]	-	-	-	-	0.394	0.267	0.563	1.094
MTN [20]	0.128	0.162	0.355	1.249	0.392	0.278	0.571	1.128
MTN-TMT (Ours)	**0.142**	**0.171**	**0.371**	**1.357**	**0.402**	**0.293**	**0.587**	**1.190**

Table 4: *Metrics of MTN-TMT with deeper TMT.*

Depth M	N	BLEU-4	METEOR	ROUGE-L	CIDEr
Text Only					
-	1	0.346	**0.289**	**0.571**	**1.107**
-	2	0.343	0.288	0.568	1.089
-	3	**0.349**	0.288	**0.571**	1.093
Video and Text					
1	1	0.393	**0.296**	**0.587**	1.169
2	1	**0.402**	0.293	**0.587**	**1.190**
3	1	0.394	0.288	0.584	1.167

better modal representations to MTN to accurately answer questions.

Ablation Study. To further study the influence of the individual components in MTN-TMT, we perform comprehensive ablation analysis in the Video and Text task. First, we evaluate the performance of video-caption translator and dialog-summary translator, respectively. The results are shown in the lower part of Table 2. We observe that dialog-summary translator and video-caption translator are beneficial to improve the performance of MTN. In addition, dialog-summary translator help MTN obtain the more competitive scores.

Next, we consider supervised learning as a crucial factor for TMT. Consequently, we conduct experiments on decaying the learning weights α and β simultaneously while training MTN-TMT. Note that we find the best hyperparameters of α and β by grid search method on dev set. In this ablation study, we initialize α and β to 0.3 and 0.8, respectively. Then, we decay the α and β by 10% every 10 epochs. When epoch reaches up to 20, TMT can be regarded as a MulT-like method due to losing the supervision of caption and summary. As shown in Table 3, training TMT by supervised learning is a positive way to improve the representations of the modalities.

Considering deep TMT naturally integrates multiple-level features and would provide better representations for modality [32], we study the influence on depth of TMT. First, we increase the depth of TMT in Text Only task. The upper part of Table 4 shows the results. However, deeper dialog-summary translator does not improve the performance in Text Only task. We suspect that TMT with one layer has sufficient ability to learn representations of dialog modality. Next, we keep the depth of dialog-summary translator to 1 and evaluate deeper video-caption translator on Video and Text task. The lower part of the Table 4 shows results. We find that video-caption translator with 2 layers outperforms the other two, for example, improving CIDEr by 0.021.
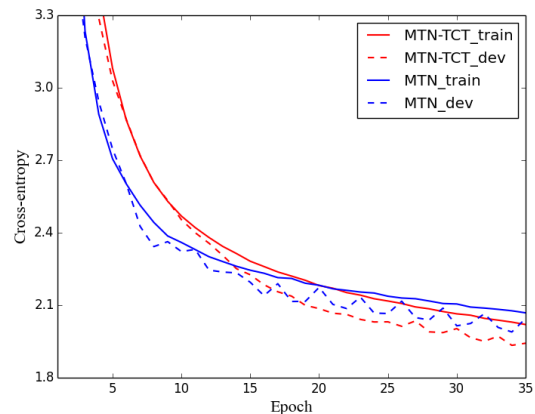

Figure 3: *Loss of MTN and MTN-TMT on the training set and dev set. Because dropout is activated when training but deactivated when evaluating on the dev set, training loss is slightly higher than dev loss.*

Comparison with Submission Models. Finally, we summarize the performance of the recently proposed models in Text Only task and Video and Text task. The hierarchical attention (HAN) [10] is the systems that rank 1st in the DSTC7-AVSD task adopted. Multimodal Transformer Networks (MTN) [20] is the state-of-the-art system before the DSTC8-AVSD Challenge. As the DSTC7 dataset is organized as two versions (single reference and six references), we report the model performance on both sets. As shown in Table 5, MTN-TMT surpasses MTN and other submission models in both Text Only task and Video and Text task. Compared with MTN, MTN-TMT obtains improvements on all metrics, especially, achieving relative improvement up to 14.1% on CIDEr.

5. Conclusion

In this paper, we propose a Transformer-based Modal Translator (TMT) to learn the representations of multimodal sequence for Audio Visual Scene-aware Dialog. TMT enhance representations of the source modal sequence by translating the source modal sequence to the related target modal sequence. Based on MTN, we apply TMT to video and dialog to learn better representations of modalities in video-grounded dialog systems. On the AVSD track of the 7th Dialog System Technology Challenge, MTN-TMT outperforms the MTN and other submission models in both Video and Text task and Text Only task. Compared with MTN, MTN-TMT improves all metrics, especially, achieving relative improvement up to 14.1% on CIDEr.

6. References

[1] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *International Conference on Computer Vision (ICCV)*, 2017.

[2] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language." IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[3] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* Association for Computational Linguistics, 2011, pp. 190–200.

[4] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *ACM Multimedia*, 2017.

[5] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4631–4640.

[6] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvqa+: Spatio-temporal grounding for video question answering," in *Tech Report, arXiv*, 2019.

[7] H. Alamri, C. Hori, T. K. Marks, D. Batr, and D. Parikh, "Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7," in *DSTC7 workshop at AAAI*, 2019.

[8] H. Lee, S. Yoon, F. Dernoncourt, D. S. Kim, T. Bui, and K. Jung, "Dstc8-avsd: Multimodal semantic transformer network with retrieval style word generator," 2020.

[9] C. Hori, T. Hori, T. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 4203–4212. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/ICCV.2017.450

[10] R. Sanabria, S. Palaskar, and F. Metze, "Cmu sinbad's submission for the dstc7 avsd challenge," in *DSTC7 at AAAI2019 workshop*, vol. 6, 2019.

[11] S. H. Kumar, E. Okur, S. Sahay, J. J. A. Leanos, J. Huang, and L. Nachman, "Context, attention and audio feature explorations for audio visual scene-aware dialog," 2019.

[12] 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.

[13] 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.

[14] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[15] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: A large-scale dataset for multimodal language understanding."

[16] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *International conference on machine learning*, 2016, pp. 1378–1387.

[17] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv:1611.01603*, 2016.

[18] Y.-T. Yeh, T.-C. Lin, H.-H. Cheng, Y.-H. Deng, S.-Y. Su, and Y.-N. Chen, "Reactive multi-stage feature fusion for multimodal dialogue modeling," *arXiv preprint arXiv:1908.05067*, 2019.

[19] D. T. Nguyen, S. Sharma, H. Schulz, and L. El Asri, "From film to video: Multi-turn question answering with multi-modal context," 2019.

[20] H. Le, D. Sahoo, N. Chen, and S. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5612–5623.

[21] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

[22] O. Kovaleva, A. Rumshisky, and A. Romanov, "Similarity-based reconstruction loss for meaning representation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4875–4880. [Online]. Available: https://www.aclweb.org/anthology/D18-1525

[23] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, and et al., "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019. [Online]. Available: http://dx.doi.org/10.1109/icassp.2019.8682583

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[25] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *Lecture Notes in Computer Science*, p. 510–526, 2016. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46448-0_31

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics.* Association for Computational Linguistics, 2002, pp. 311–318.

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

[28] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 2004, p. 605.

[29] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.