Investigation on Transformer-based Multi-modal Fusion for Audio-Visual Scene-Aware Dialog

Xin Huang, Hui Li Tan, Mei Chee Leong, Ying Sun, Liyuan Li, Ridong Jiang, Jung-jae Kim

{huangx2, hltan, Leong_Mei_Chee, suny, lyli, rjiang, jjkim}@i2r.a-star.edu.sg

Abstract

In this report, we present our submissions to the DSTC10 Audio Visual Scene Dialog (AVSD) challenge. We investigated variants of an encoder-decoder model, including those with multi-modal cross-attention and those with various fusion strategies to aggregate the multi-modal inputs (audio, visual, text, object). Our submissions achieved competitive results in the two tasks of the AVSD challenge. For the first task (video Q&A dialog), our submissions achieved BLEU-4, METEOR, ROUGE, CIDEr, and human rating of 37.2%, 24.3%, 53.0%, 91.2%, and 3.57 respectively. For the second task (reasoning for Q&A), our submissions achieved IoU-1 and IoU-2 of 48.5% and 51.0% respectively. Our submissions, under Team Anonymous, achieved the top rank for the human rating, the third rank for the automatic evaluation of the first task, and the second rank for the second task.

Introduction

The DSTC10 AVSD challenge comprises two tasks. The first task (video Q&A dialog) (Antol et al. 2015; Zhang et al. 2016; Goyal et al. 2017; Tapaswi et al. 2016) is to generate text response (answer) to user input (question) about a video. The videos used in the AVSD 10 challenge are based on the Charades dataset (Sigurdsson et al. 2016). The training data of the task include manual descriptions of videos, which can be used to train models, but the manual descriptions are not given in the testing data. This task is identical to the previous AVSD challenges in DSTC 7 and 8 (D'Haro et al. 2020; Kim et al. 2021; Hori et al. 2020). The second task (reasoning for Q&A) is to identify video-grounded evidence to the answer of the first task, in form of the begin/end timing in the video. For example, when a system-generated answer is "A dog is barking", the sound of the dog's barking and the dog must be grounded in the video as evidence. The begin and end timing of the grounding/evidence are provided in the training data of the task.

We made two submissions to the DSTC10 AVSD challenge. The UniVL system is based on the pre-trained UniVL model (Luo et al. 2020) and extended for audio, visual, text features and object features. The MED-CAT system is a Transformer-based multi-modal encoder-decoder model (Vaswani et al. 2017) which simply concatenates encoder embeddings of the different modalities.

Methods

UniVL

We adapted the **Uni**fied **V**ideo and **L**anguage (UniVL) model (Luo et al. 2020) for the AVSD challenge. The UniVL model comprises a single-modal text encoder, a single-modal video encoder, a cross encoder, and a text decoder. All encoders are Transformer-based. The cross encoder combines the text encoding and video encoding to get a unified multi-modal encoding. The model is pre-trained on the HowTo100M dataset (Miech et al. 2019). UniVL has been shown to learn strong video-text representations for multi-modal generation (e.g. video captioning) and understanding (e.g. text-based video retrieval) downstream tasks.

We apply the text encoder for the following two text inputs: a given question q_t and the concatenation of its past question/answer pairs in a dialogue $q_1 a_1 \ldots q_{t-1} a_{t-1}$. Each text input is tokenized into a token sequence ($tokens$), and encoded by the text encoder, which is a BERT model (Devlin et al. 2019), i.e.,

$$T = BERT(tokens), \tag{1}$$

where $T \in \mathbb{R}^{l_t \times d}$ is the text encoding of $tokens$. l_t is the length of the token sequence, and $d = 768$ is the hidden size of the text representation. T_q represents the text encoding of the question q_t, and T_h represents the text encoding of the conversation history $q_1 a_1 \ldots q_{t-1} a_{t-1}$.

The video inputs (visual only, without audio) are sampled and embedded using the s3d feature extractor (Xie et al. 2018), which are encoded by a Transformer-based visual encoder VE, i.e.,

$$V = VE(F_v), \tag{2}$$

where F_v is the s3d feature, and $V \in \mathbb{R}^{l_v \times d}$ is the visual encoding. l_v is the video frame length.

We extended UniVL with audio and object features. As provided by the challenge organizers, the audio of the video are extracted and embedded using the VGGish feature extractor (Hershey et al. 2017). $A \in \mathbb{R}^{l_a \times a}$, later padded to $A \in \mathbb{R}^{l_a \times d}$, designates the audio embedding, where l_a and $a = 128$ are the audio frame length and the size of audio embedding respectively.

We employ a state-of-the-art object detector D2Det (Cao et al. 2020) for recognizing objects in the video. D2Det is trained on the COCO dataset (Lin et al. 2014) for 80 classes,

including person, items found in kitchen (e.g., bottle, cup, bowl, fork, knife, spoon, dining table, refrigerator, sink), items found in living room (e.g., tv, desk, couch, chair), and items found in room (e.g., bed). The objects are embedded as a matrix where each row indicates a video frame and is associated with the top-five object classes detected by D2Det. Each object class is represented with its text description, which is tokenized into up to 4 tokens (e.g., 'tennis_racket' is tokenized as ['tennis', '_', 'rack', 'et']), and the tokens of the top-five class descriptions are concatenated and then embedded. The class description whose number of tokens is less than 4 is padded to the maximum token length (4) before the concatenation. The resultant object embedding is designated as $O \in \mathbb{R}^{l_v \times 20}$, which is padded to $O \in \mathbb{R}^{l_v \times d}$.

The features/encodings (T_q, T_h, V, A, O) are then passed into a Transformer-based cross encoder (CE). We experimented with early fusion and late fusion of the features/encodings. In the early fusion approach, all the five features are concatenated along the sequence dimension and passed into the cross encoder, i.e.,

$$M_{early} = CE([T_q, T_h, V, A, O]). \quad (3)$$

In the late fusion approach, the encoding features of question T_q and conversation history T_h are individually concatenated with the visual encoding V and passed into the cross encoders. Both of them share the same cross encoder model. The audio features A, object features O, and cross encoder outputs are concatenated along the sequence dimension as input to the decoder, i.e.,

$$M_{late} = [CE([T_q, V]), CE([T_h, V]), A, O]. \quad (4)$$

Finally, the outputs of the fusion (M) together with the target string (Y) for training, are passed into a Transformer decoder, i.e.,

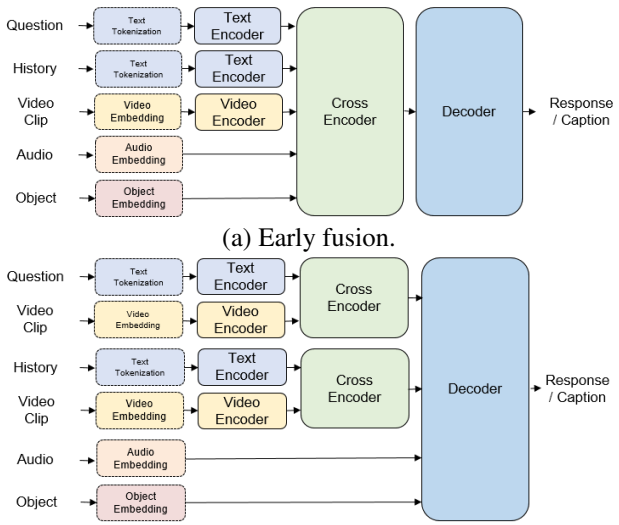
$$D = Decoder(Y, M), \quad (5)$$

where $D \in \mathbb{R}^{l \times v}$ is the decoder output, from which a sequence of words $Y' = y_1, \ldots, y_d$ is generated as the system response (answer). l is the decoder length, and v is the size of the token vocabulary. We employ cross-entropy loss on the generated answers for model training. The early fusion approach is submitted for the AVSD challenge. Figure 1 illustrates the two fusion approaches of the UniVL model extended for AVSD.

MED-CAT

We also developed a multi-modal encoder-decoder model for the AVSD challenge (MED-CAT), whose decoder takes the concatenation of multiple encoders' embeddings as input. The MED-CAT model includes a text encoder, a visual encoder, an audio encoder and a text decoder. The structure of the MED-CAT is a variant of UniT (Hu and Singh 2021) model, which we add an audio encoder and use a language modeling head for the decoder to do language generation task. The model is similar to the UniVL model with late fusion, while not including a cross encoder. The model is pre-trained on the HowTo100M dataset.

The inputs of the model include text feature F_t, visual feature F_v and audio feature F_a, and the output of the model



(a) Early fusion.

(b) Late fusion.

Figure 1: Adapted UniVL model.

is system response (answer) which comprises of a sequence of words $Y = y_1, \ldots, y_d$. Figure 2 depicts the architecture of the MED-CAT model.

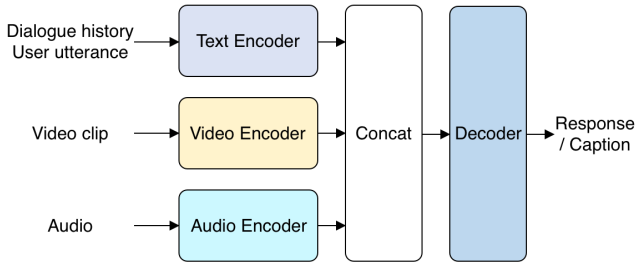


Figure 2: MED-CAT model.

As for text features, the MED-CAT model concatenates the tokens of question q_t and of conversation history $q_1 a_1 \ldots q_{t-1} a_{t-1}$ as input to the BERT model and thus has only one text encoding T instead of two text encodings (T_q, T_h) for the UniVL model.

$$T = BERT([tokens_q, tokens_h]) \quad (6)$$

The MED-CAT model uses the same visual and audio features as in the UniVL model (i.e., F_v, F_a). The same visual encoder VE from UniVL is used to encode the visual features, but an additional transformer encoder AE is used for encoding the audio features, hence obtaining encoded visual features V and encoded audio features A, i.e.,

$$V = VE(F_v) \quad (7)$$
$$A = VA(F_a). \quad (8)$$

A transformer decoder is finally used to predict target utterance $Y = y_1, \ldots, y_d$, where $D \in \mathbb{R}^{l \times d}$ is the output

of the decoder. The decoder includes a multi-head cross-attention layer $CrossAttn$, where the hidden feature of target string h_Y attends to the last layer of hidden representation for visual and audio features.

$$D = Decoder(Y, [T, V, A]) \tag{9}$$

$$h'_Y = CrossAttn(D, [V, A]) \tag{10}$$

The output representation h'_Y is used to proceed the decoding process. The cross-entropy loss is used for model training.

Temporal Localization for Answer Reasoning

The goal of answer reasoning is to find the localization of audio or video features as evidence for each generated answer. We follow the approach of the baseline (Shah et al. 2021; Hori et al. 2018, 2019) which uses the attention based method to calculate the distribution of the time region for the input visual features. Note that both our method and the baseline method select time regions based on the visual features only. We adapted the attention-based baseline method into UniVL and MED-CAT with a few modifications: 1) Our method adapts the beam search to proceed with decoding process, and we use the average encoder-decoder attention scores of each beam to proceed with the calculation of attention distribution, and 2) our method decodes multiple answers at the same time, and we use the length of the video to deal with padding issues.

Formally, let b be the beam size, l_v be the video frame length, and d, g, l be the hidden size, the number of heads and the number of layers of the $CrossAttn$ respectively. The weight of the $CrossAttn$ is noted as $W_{Attn_{i,v}} \in \mathbb{R}^{l \times g \times d}$ for i-th beam and v-th frame. We first sample the last layer of the attention weight $W_{last_{i,v}} \in \mathbb{R}^{g \times d}$, and then average the attention score across all heads and beams to obtain $W_{avg_v} \in \mathbb{R}^d$ for v-th frame:

$$W_{avg_v} = \frac{1}{b \times g} \sum_{\forall i} \sum_{\forall j} W^{(j)}_{last_{i,v}}. \tag{11}$$

Then we calculate the mean μ and the standard deviation σ from the attention score W_{avg_v}:

$$\mu = \frac{1}{l_v \times d} \sum_{\forall v} \sum_{\forall z} Softmax(W^{(z)}_{avg_v}) \tag{12}$$

$$\sigma = Std(Softmax(W_{avg_v})). \tag{13}$$

Finally we calculate the predicted time region as $(l_v(\mu - \sigma), l_v(\mu + \sigma))$. Note that we can adjust the time region for different frame lengths to the valid region in order to support batch decoding.

Experimental Results

For UniVL and MED-CAT, we used maximum word length of 120 and maximum frame of 60. For UniVL, we used 12 Transformer layers for text encoder, 6 Transformer layers for visual encoder, 2 Transformer layers for cross encoder and 3 Transformer layers for decoder. We used learning rate of $3e^{-5}$, number of epochs of 5, and batch size of 8. For MED-CAT, we also used 12 Transformer layers for text encoder, 3 layers for visual and audio encoder and 3 layers for decoder. We chose the hyperparameters of learning rate $\alpha = 3 \times 10^{-5}$, number of epochs $E = 10$ and the batch size $B = 32$.

The experimental results on the official test-set for the video Q&A dialog task are shown in Table 1. Our submissions are medcat_multimodal+audio_visual_qa (MED-CAT) and univl_multimodal+audio_visual_qa_objects (UniVL). The submission univl_multimodal+audio_visual_qa_objects is the UniVL model with early fusion of audio (VGGish), visual (s3d), text features (QA history) and object features (objects are extracted using the state-of-the-art D2Det object detector). Our submissions achieved BLEU-4, METEOR, ROUGE, CIDEr, and human rating of 37.2%, 24.3%, 53.0%, 91.2%, and 3.57 respectively. Both MED-CAT and UniVL outperform the baseline model by more than 13% in terms of BLEU-4. MED-CAT and UniVL also achieve high CIDEr of 90%, almost 30% above the baseline model. Our automatic evaluation results are close to those of the top performing methods, and our human evaluation results outperform them.

Tables 2 and 3 show the experimental results on the official test-set for the DSTC 7 AVSD and DSTC 8 AVSD challenges, respectively. CMU Sinbad's method (Ramon Sanabria and Metze 2019) is the previous SOTA in the DSTC 7 AVSD challenge, which uses a GRU based encoder, a conditional GRU based decoder and the hierarchical attention for the multimodal fusion. The method from Li et al. (2020) is the the previous SOTA in the the DSTC 8 AVSD challenge, which concatenates multimodal features to a long sequence and uses a pretrained GPT-2 model to fine-tune on the AVSD dataset.

We explored with MED-CAT various intra-encoder fusion strategies to combine the audio, visual, text and object features. Early fusion of MED-CAT fuses the intermediate encodings of all the modalities by using cross-attention at every layer level. Mid-level fusion of MED-CAT employs the fusion only at higher layer levels (Nagrani et al. 2021). Late fusion is the default setting of MED-CAT where the last layer hidden representations of different encoders are concatenated before the decoding process. Table 4 shows the experimental results on the validation set, including the results of the two fusion approaches of the UniVL model for comparison purpose. The better performing models are submitted to the challenge.

We also explored different combinations of the inputs to test the performance of MED-CAT. In the first setting, only a text feature (T) is allowed, and thus $BERT$ is the only encoder of MED-CAT. In the second setting, both the text feature (T) and visual feature (V) are allowed, and the model includes both $BERT$ and VE as encoders for text and visual features respectively. In the third setting, text feature (T), visual feature (V) and audio feature (A) are allowed which is the default setting for MED-CAT. Table 5 shows the experimental results on the validation set for different settings. we can see the performance improvement for MED-CAT with the additional visual and audio feature and the model is able to learn from multimodal features.

Method (Submission Name)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	human rating
medcat_multimodal+audio_visual_qa	0.670	0.541	0.441	0.365	0.241	0.526	0.906	-
univl_multimodal+audio_visual_qa_objects	**0.673**	**0.545**	**0.448**	**0.372**	**0.243**	**0.530**	**0.912**	**3.569**
baseline	0.572	0.422	0.320	0.243	0.191	0.439	0.566	2.851

Table 1: Experimental results on the official test-set for the video Q&A dialog task.

Method	TEST (AVSD7) w/o caption / summary Last_only			
	BLEU-4	METEOR	ROUGE_L	CIDEr
Baseline AV-transformer	0.296	0.214	0.485	0.771
CMU Sinbad's Method	0.394	**0.267**	**0.563**	1.094
Hung Le (2019)	0.315	0.239	0.509	0.848
UniVL + objects	**0.409**	0.265	0.561	1.115
MED-CAT	0.401	0.260	0.559	**1.117**

Table 2: Experimental results on the DSTC7-AVSD test set.

Method	TEST (AVSD8) w/o caption / summary Last_only			
	BLEU-4	METEOR	ROUGE_L	CIDEr
Baseline	0.289	0.210	0.480	0.651
Li et al. (2020)	**0.387**	0.249	0.544	**1.022**
UniVL + objects	0.383	**0.253**	**0.555**	1.006
MED-CAT	0.376	0.247	0.547	0.982

Table 3: Experimental results on the DSTC8-AVSD test set.

Method	VAL (AVSD10) Last_only			
	BLEU-4	METEOR	ROUGE_L	CIDEr
UniVL (early fusion)	**0.125**	**0.153**	**0.383**	**1.292**
UniVL (late fusion)	0.120	0.149	0.381	1.251
MED-CAT (early fusion)	0.120	0.149	0.380	1.243
MED-CAT (mid fusion)	0.121	0.149	0.379	1.239
MED-CAT (late fusion)	**0.124**	**0.151**	**0.382**	**1.272**

Table 4: Fusion approaches. Experiments results on validation set for the video Q&A dialog task.

Method	VAL (AVSD10) Last_only			
	BLEU-4	METEOR	ROUGE_L	CIDEr
MED-CAT (T)	0.119	0.147	0.376	1.230
MED-CAT (T + V)	0.121	0.149	0.382	1.265
MED-CAT (T + V + A)	**0.124**	**0.151**	**0.382**	**1.272**

Table 5: Ablation studies of MED-CAT on different combinations of input features.

The experimental results on the official test-set for the reasoning for Q&A task are shown in Table 6. Our submissions achieved IoU-1 and IoU-2 of 48.5% and 51.0% respectively, outperforming the baseline, and are close to the top performing method.

Method	IoU-1	IoU-2
medcat_multimodal+audio_visual_qa	0.485	0.510
univl_multimodal+audio_visual_qa_objects	0.479	0.505
AV_CAPTION(FRONT)_QA_GPT2_greedy	**0.518**	**0.544**
final4	0.362	0.380
baseline	0.361	0.380

Table 6: Experimental results on the official test-set for the reasoning for Q&A task.

Conclusion

In this report, we present our submissions to the DSTC10 AVSD challenge. We investigated encoder-decoder models with variants of cross attention and multi-modal fusion. Our methods achieve competitive results on the challenge, and serve as our baseline for further investigation of visual-linguistic learning.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.; and Parikh, D. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2425–2433. Los Alamitos, CA, USA: IEEE Computer Society.

Cao, J.; Cholakkal, H.; Anwer, R. M.; Khan, F. S.; Pang, Y.; and Shao, L. 2020. D2Det: Towards High Quality Object Detection and Instance Segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 4171–4186.

D'Haro, L. F.; Yoshino, K.; Hori, C.; Marks, T. K.; Polymenakos, L.; Kummerfeld, J. K.; Galley, M.; and Gao, X. 2020. Overview of the seventh Dialog System Technology Challenge: DSTC7. *Computer Speech Language*, 62: 101068.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6325–6334.

Hershey, S.; Chaudhuri, S.; Ellis, D. P. W.; Gemmeke, J. F.; Jansen, A.; Moore, C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; Slaney, M.; Weiss, R.; and Wilson, K. 2017.

CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Hori, C.; Alamri, H.; Wang, J.; Winchern, G.; Hori, T.; Cherian, A.; Marks, T. K.; Cartillier, V.; Lopes, R. G.; Das, A.; et al. 2018. End-to-End Audio Visual Scene-Aware Dialog using Multimodal Attention-Based Video Features. *arXiv preprint arXiv:1806.08409*.

Hori, C.; Cherian, A.; Hori, T.; and Marks, T. K. 2020. Audio visual scene-aware dialog (AVSD) track for natural language generation in DSTC8. In *Proceedings of DSTC8 Workshop at AAAI-2020*.

Hori, C.; Cherian, A.; Marks, T. K.; and Hori, T. 2019. Joint Student-Teacher Learning for Audio-Visual Scene-Aware Dialog. In *INTERSPEECH*.

Hu, R.; and Singh, A. 2021. UniT: Multimodal Multitask Learning with a Unified Transformer. arXiv:2102.10772.

Hung Le, D. S. N. F. C., Steven C.H. Hoi. 2019. End-to-End Multimodal Dialog Systems with Hierarchical Multimodal Attention on Video Features. In *AAAI2019 workshop*, DSTC7.

Kim, S.; Galley, M.; Gunasekara, C.; Lee, S.; Atkinson, A.; Peng, B.; Schulz, H.; Gao, J.; Li, J.; Adada, M.; Huang, M.; Lastras, L.; Kummerfeld, J. K.; Lasecki, W. S.; Hori, C.; Cherian, A.; Marks, T. K.; Rastogi, A.; Zang, X.; Sunkara, S.; and Gupta, R. 2021. Overview of the Eighth Dialog System Technology Challenge: DSTC8. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 2529–2540.

Li, Z.; Li, Z.; Zhang, J.; Feng, Y.; Niu, C.; and Zhou, J. 2020. Bridging Text and Video: A Universal Multimodal Transformer for Video-Audio Scene-Aware Dialog. arXiv:2002.00163.

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 Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. *arXiv preprint arXiv:2002.06353*.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2630–2640.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention Bottlenecks for Multimodal Fusion. arXiv:2107.00135.

Ramon Sanabria, S. P.; and Metze, F. 2019. CMU Sinbad's Submission for the DSTC7 AVSD Challenge. In *AAAI2019 workshop*, DSTC7.

Shah, A. P.; Geng, S.; Gao, P.; Cherian, A.; Hori, T.; Marks, T. K.; Roux, J. L.; and Hori, C. 2021. Audio-Visual Scene-Aware Dialog and Reasoning using Audio-Visual Transformers with Joint Student-Teacher Learning. arXiv:2110.06894.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. arXiv:1604.01753.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. MovieQA: Understanding Stories in Movies Through Question-Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, 305–321.

Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2016. Yin and Yang: Balancing and Answering Binary Visual Questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.