Multi-Modal Correlated Network with Emotional Reasoning Knowledge for Social Intelligence Question-Answering

Baijun Xie[1] and Chung Hyuk Park[1,2]

[1]Department of Biomedical Engineering, [2]Department of Computer Science

School of Engineering and Applied Science, The George Washington University

Washington, DC, USA

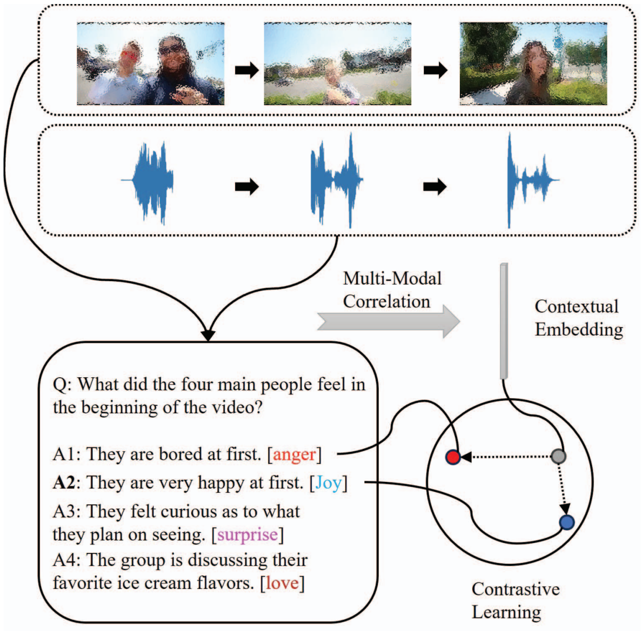
bdxie@gwu.edu, chpark@gwu.edu

Abstract

The capacity for social reasoning is essential to the development of social intelligence in humans, which we easily acquire through study and experience. The acquisition of such ability by machines, however, is still challenging, even with the diverse deep learning models that are currently available. Recent artificial social intelligence models have achieved state-of-the-art results in question-answering tasks by employing a variety of methods, including self-supervised setups, multi-modal inputs, and so on. However, there is still a gap in the literature regarding the introduction of commonsense knowledge when training the model in social intelligence tasks. In this paper, we propose a Multi-Modal Temporal Correlated Network with Emotional Social Cues (MMTC-ESC). In order to model cross-modal correlations, an attention-based mechanism is used, and contrastive learning is achieved using emotional social cues. Our findings indicate that combining multimodal inputs and using contrastive loss is advantageous for the performance of social intelligence learning.

1. Introduction

Humans can develop intelligence by understanding different modalities of verbal and nonverbal social cues to reason about the underlying mental feelings, thoughts, and intentions of others [1]. When a person being questioned by someone looks around in a hesitant manner, for instance, we can recognize that the person may not be certain of the answer based on their own experiences and thought processes. This is an example of a well-known term in psychology known as the Theory of Mind [15], which is the ability to comprehend others by attributing mental states to them.



Q: What did the four main people feel in the beginning of the video?

A1: They are bored at first. [anger]

A2: They are very happy at first. [Joy]

A3: They felt curious as to what they plan on seeing. [surprise]

A4: The group is discussing their favorite ice cream flavors. [love]

Figure 1. A scenario of using common sense knowledge for modeling the multimodal social intelligence challenge. The process of cross-modal correlation will enrich the information of the overall contextual feature, while the underlying emotional and social clues are employed for contrastive learning. The purpose of introducing emotional cues is to pull the contextual feature vector away from the incorrect answer with atypical affective states.

Theory of Mind is also related to empathy [25], where empathy is the capacity to understand another person's feelings and emotions. Although such social reasoning abilities are easily learned and developed by humans, the current benchmarks for the task of understanding social interaction are comparatively low-resource [31]. Consequently, this raises the question of whether we can use the characteristics of social reasoning abilities to enhance the effectiveness of supervised learning models for the task of enhancing social intelligence.

Recent studies have demonstrated the utilization of large-scale pre-trained models for the downstream task of question-answering (QA) [5, 29]. Despite these significant advancements achieved by large pre-trained models, these models remain deficient in the ability to reason about social situations when performing various tasks [14, 21]. This is partially due to the fact that the training text corpora include inherent biases, which restrict the depth of commonsense knowledge [6]. To overcome this, a recent study investigated capturing relationships among different modalities using commonsense reasoning [36]. Other research has also shown that integrating multi-modal inputs can enhance the performance of different tasks [30, 34]. Besides leveraging multi-modal information, deeper reasoning skills, including evidence and commonsense reasoning, were also explored to advance the video QA task [11]. A language model could also improve their understanding of the underlying commonsense knowledge through the strategy of introducing domain knowledge and semantic information [35].

In this paper, inspired by the concept of incorporating multi-modal inputs and the strategy of utilizing domain knowledge for improving the understanding of commonsense knowledge, we propose to develop a novel framework named **Multi-Modal Temporal Correlated Network with Emotional Social Cues** (MMTC-ESC). MMTC-ESC exhibits an attention-based mechanism to model cross-modal correlations and utilizes contrastive learning for reasoning about emotional and social cues.

2. Related Work

2.1. Video Question Answering

Video QA is a popular vision-language task that has been researched for years. Some of the previous datasets [7, 28] collected short video clips regarding everyday activities that are performed by humans. Additionally, other datasets also collected long-term videos from movies or TV series, such as TVQA [10] and MVQA [22], to understand the underlying meaning of dialogues from the videos. Nevertheless, the Social-IQ [18] dataset contains various videos with a greater focus on social interaction, which aims to evaluate the model's social intelligence abilities. The primary focus of this study is on the development of a model that incorporates social intelligence and the evaluation of that model with Social-IQ 2.0 [26], which is the second generation of the Social-IQ dataset [31].

2.2. Multi-Modal Question Answering Models

In multiple previous studies, the long short-term memory (LSTM) model has been investigated for its capacity to summarize or fuse multi-modal information in order to provide answers to questions in videos [32, 10, 33, 9]. However, LSTM training is hard, and there is no capacity for transfer learning; which restricts its applicability to a variety of different tasks. Additionally, the attention mechanism [24] has also received a considerable amount of interest in previous studies [8, 20, 5]. Dual attention combining late fusion on the latent representations of frames and captions demonstrated increased performance on early fusion [8]. Human-like attention signals were also utilized to apply attention mechanism with questions and images during training video QA models [20]. A model with a spatial-temporal transformer was also found to be better suited in the pre-trained models for long-form video QA tasks [5]. However, those studies are still insufficient in taking into account commonsense knowledge reasoning. Other studies [11, 35] that consider the understanding of commonsense knowledge are not yet fully adapted to the task of social intelligence. Therefore, we aim to include commonsense knowledge in the procedure of the training by making use of the latent representation of the multi-modal outputs.

3. Method

Based on the Social-IQ 2.0 dataset [26], the goal of the social intelligence task in this study is to predict the answer y for a given media $\mathcal{M}$ with video and audio inputs, and a question q, which is formulated as follows:

$$\widetilde{y} = \underset{y \in \mathcal{A}}{\arg\max} \mathcal{F}_{\theta}(y \mid q, \mathcal{M}, \mathcal{A}), \tag{1}$$

where $\widetilde{y}$ is the predicted answer chosen from multiple choices, which is denoted by $\mathcal{A}$, and θ is the trainable parameter of the inference model $\mathcal{F}$.

Our approach uses three primary components to accomplish the goal: (1) feature extraction from multi-modal inputs; (2) cross-modal correlation modeling via attention mechanisms; and (3) emotional social cues with contrastive learning loss.

3.1. Multi-Modal Input Representation

Specifically, our multi-modal inputs for the task of social intelligence QA include textual input T, audio input A, and visual input V. The current models have the ability to extract distinguished features from short-term clips of video or audio data. To adapt the long-term media inputs, we first divide the media into N uniform-length segments, where each segment contains an equal data length L.

We use modality-specific models to extract unimodal features from different modalities of input. The Video-

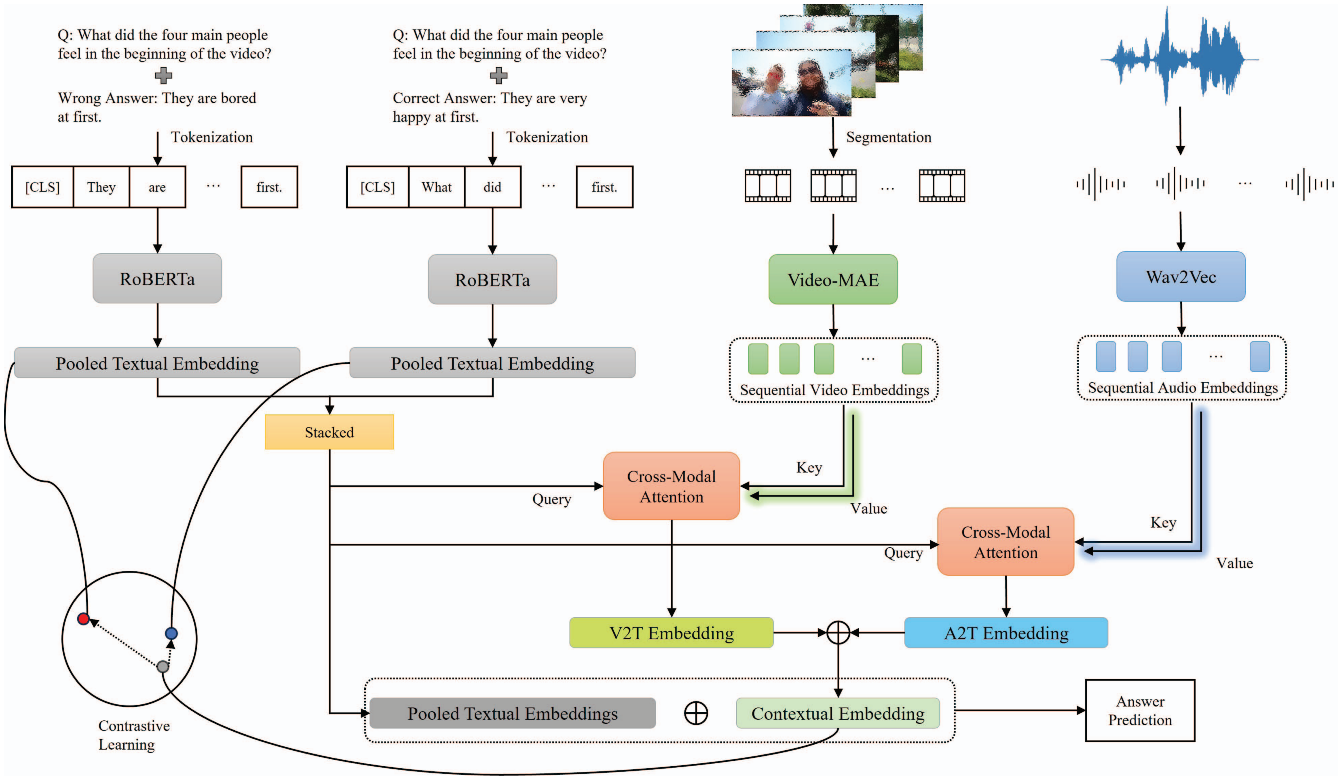


Figure 2. The overall framework of the MMTC-ESC. When given text with video or audio sequences as inputs, the cross-modal attention module generates multimodal representations, V2T and A2T. In addition, a contrastive loss is created by bringing the contextual embedding close to the textual embedding containing the correct answer.

MAE model [23], with frozen parameters, extracts the features of all video segments from the video inputs, denotes by $\boldsymbol{x}_V = \{x_V^1, x_V^2, \ldots, x_V^n\}$, where $x^n \in \mathbb{R}^{N \times L_V \times D_V}$, and D is the dimension of the extracted features. Similarly, the Wav2vec model [2], with frozen parameters, extracts the features of all audio segments from the audio inputs, denotes by $\boldsymbol{x}_A = \{x_A^1, x_A^2, \ldots, x_A^n\}$, where $x^n \in \mathbb{R}^{N \times L_A \times D_A}$.

For the textual information, we generate QA tuples from the dataset. First, given the transcript from the dataset, we use the T5 model [17] to generate summaries from each video's transcript and denote the summary as s. Then, given a sample with a summary, a question q and multiple-choice answers $\boldsymbol{a} = \{a_1, a^2, \ldots, a^k\}$, we tokenize each summary, question, and answer tuple as $\boldsymbol{QA} = [\{s, q, a_1\}, \{s, q, a_2\}, \ldots, \{s, q, a_k\}]$, where $i \in \mathbb{R}^K$, and K is the number of answers. Then, we employ the pretrained model RoBERTa [12] to extract the feature vectors for each QA pair. Then, the pooled outputs from the RoBERTa model, the last layer hidden state of the first token of the sequence, are used for the downstream task of predicting the correct answer.

3.2. Cross-Modal Correlations Modeling

To correlate different modalities of features with the QA textual information, we apply a cross-modal attention mechanism to model the correlations between textual modality and other modalities. Given the tokenized sequence $\{[CLS], w_1, w_2, \cdots, w_L\}$ from a pair of QA, which is fed into the RoBERTa, and the output vector, $X_T \in \mathbb{R}^{L_T \times D_T}$, at the position of the $[CLS]$ token is extracted for the latter utilization.

In order to obtain the relevant sequential model outputs for the video modality, we feed the sequential video segments into the video model. Then, we get the pooled outputs from the last hidden state of each model output and stack them to produce the sequential video features, $X_V \in \mathbb{R}^{L_V \times D_V}$. Equivalently, the sequential audio features extracted by the audio model are denoted as $X_A \in \mathbb{R}^{L_A \times D_A}$.

Using the multi-head attention mechanism proposed by Vaswani et al. [24], for the given query Q, key K and value V, the attended output is formulated as below:

$$Y = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{2}$$

In this study, query Q represents the textual information, which is defined as $Q_T = X_T W_{Q_T}$. Key K and value V are the information from another modality, take audio modality as an instance, $K_A = K_A W_{K_A}$ and $V_A = V_A W_{V_A}$. Therefore, the attended contextual output modeling the temporal correlations between text and audio modalities is denoted as Y_{A2T}. It is noted that Y_{A2T} has the same shape as the textual feature X_T, but also a representation vector in the feature space of V_A. Specifically, Y_{A2T} is a linear weighted combination of the value V_A, representing the modeling temporal correlations between the modalities. Applying the multi-head attention mechanism is motivated by the idea that the textual information Q can correspond to the target item, where K and V from the audio or video modality are the sources. Based on the textual input Q, we expect that attention outputs can retrieve the most relevant information from the source modalities.

Finally, the overall representation vector containing the original textual information and the other contextual features vector is represented as:

$$X_{all} = \mathrm{ReLU}(X_T || Y_M), \tag{3}$$

where $||$ denotes the concatenation of feature representations. The multimodal representation $Y_M = Y_{V2T} || Y_{A2T}$ is also the concatenation of the features from two modalities.

3.3. Emotional Causal Cues with Contrastive Learning

To provide a model with commonsense reasoning abilities, we propose employing emotional characteristics as an indicator to generate positive and negative pairs from data, a process known as contrastive learning. Given the multiple-choice QA pairs $\boldsymbol{QA} = [\{q, a_1\}, \{q, a_2\}, \dots, \{q, a_k\}]$, assuming the index of the correct answer is c, we extract the emotion embeddings from each QA pair by using the T5 [17] model fine-tuned on emotion recognition dataset [19], and denote them as E_i, where $i \in \mathbb{R}^K$, and K is the number of answers. Thus, the pair-wise comparison function using cosine similarity to compare the E_c with other emotion embeddings is given as:

$$\bar{e} = \left[\mathrm{cosine}\left(E_c, \hat{E}_{:i}\right)\right]_{i=1, i \neq c}^{K} \in \mathbb{R}^K, \tag{4}$$

where $\bar{e}$ represents the within similarity scores between the emotion embedding from the correct answer and the embeddings from other incorrect answers. We select two negative samples based on the scores alongside the lowest similarity rankings as the negative sample sets.

The intuition behind the “reasoning” of explicit commonsense knowledge is to select the hard negative samples to avoid the confusion of computing and finding correct answers. It is also a common strategy in general question-answer completion on TV shows, where participants have the chance of removing two incorrect answers from the list of four potential answers in order to increase the probability that they will select the correct answer from the remaining options. Regarding training the model, we wish to make the multimodal embedding representation close to its correct QA embedding according to the emotional states of different answers. A contrastive loss objective related to the well-known InfoNCE loss [3] is given as:

$$\mathcal{L}_C = -\log\left(\frac{e^{(Y_M^\top X_T^+)}}{e^{(Y_M^\top X_T^+)} + \sum_i^2 e^{(Y_M^\top X_{T_i}^-)}}\right), \tag{5}$$

where X_T^+ denotes the positive QA embedding for the correct answer, and $X_{T_i}^-$ denotes the negative QA embedding for the incorrect answer.

3.4. Supervised Training for Answer Prediction

After the feature extraction and modeling of correlations are complete, we fine-tune the language model and train the attention layers and subsequent feed-forward layers to obtain the final predictions. The cross-entropy loss is used to optimize the predictions:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{K} y_{s,c} \log(p_{s,c}), \tag{6}$$

where K is the number of answers, and y indicating if class label c is the correct answer for sample s, and p is the predicted probability.

In the case of multiple-choice tasks, the set of negative training samples represents the false classes. The final loss is the combination of the cross-entropy loss and the contrastive loss:

$$\mathcal{L}_{all} = \alpha \mathcal{L}_{CE} + \beta \mathcal{L}_C, \tag{7}$$

where α and β are the hyperparameters for weighting both loss objectives, and their choices will be discussed in an ablation analysis of the Results section.

4. Experiments

4.1. Dataset

We used the Social-IQ 2.0 dataset [26] to train the model, which is the second generation of the Social-IQ dataset [31] and the benchmark for evaluating social intelligence via the question-answering tasks in videos. Similarly, videos of Social-IQ 2.0 also contain various social situations where people interact with each other and come with several questions asking about the interactions based on the social scene.

4.2. Implementation Details

We trained our language model with the basic configurations and pre-trained weights of RoBERTa [12]. The pre-trained weights of Video-MAE [23] and Wav2vec [2] are

Table 1. The comparison results on the validation accuracy of the Social-IQ 2.0 dataset for text-modality only (upper) and multi-modal (lower) results. T, A, and V represent the modalities of text, audio, and video. For the multi-modal results, T is the RoBERTa model, and A and V are the corresponding modality-specific models as discussed in Section 3.1.

Model	Accuracy (%)
Random Baseline	55.54
GPT	70.26
RoBERTa-base	71.41
RoBERTa-large	73.55
T+V	74.35
T+A	74.01
T+A+V	74.91
MMTC-ESC	**75.94**

frozen for feature extraction only. We used the AdamW [13] optimizer with layer-wise learning rates. We used a learning rate of $5e^{-6}$ to train the model and used a linear decrease scheduler. We trained with a batch size of 16 and 10 maximum epochs, with an early stop after 3 epochs. We divided each video or audio input into 30 segments to generate sequential embeddings. We used PyTorch and Huggingface Transformer [27] to train the model. The training was performed on a workstation with an Intel i9-10980XE core, four NVIDIA RTX A5000 GPUs with NVLink, and 256 GB of memory over 8 hours.

4.3. Evaluation

In light of the fact that the Social-IQ 2.0 dataset is structured as a multi-choice QA task and that each question contains the candidate answers, we report the accuracy for the purpose of evaluating the model. As a result, making an accurate prediction can be considered as selecting the correct answer from all possible answers. We also used the following baselines for the comparison.

T5 [17] is an encoder-decoder model that has already been trained on a variety of tasks that are both supervised and unsupervised and are each adapted into a text-to-text format. We fed the textual information from the dataset to the model and got the predicted answers, which served as the random baseline.

GPT [16] is a causal transformer that has been pre-trained on a substantial corpus using language modeling. We used it in the same settings as our previously introduced language model training method, but without multimodal inputs and contrastive loss.

RoBERTa [12] is built on BERT [4] but without next-sentence pre-training objectives and training with significantly larger mini-batches and learning rates. We used it in the same setting as our previously introduced language model training method, but without multimodal inputs and contrastive loss.

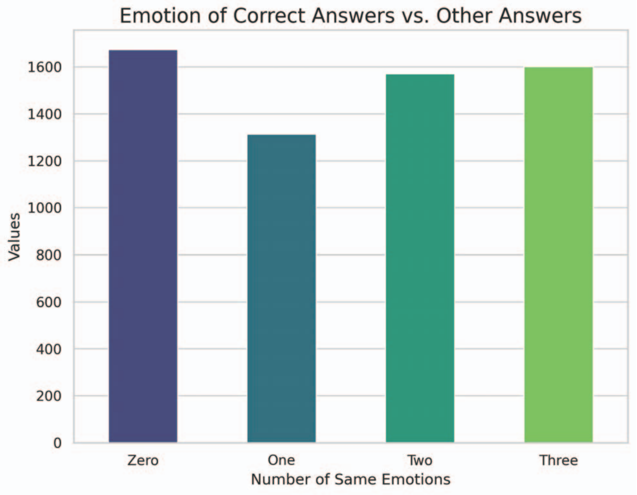


Figure 3. Histogram of the emotional difference between correct and incorrect answers is shown by this bar graph. We documented the number of times the emotions from the incorrect responses differed from the emotions of the correct answers for each QA pair.

5. Results

Table 1 shows the results of our MMTC-ESC network performance with Social-IQ 2.0 based on the validation set. We first found that the GPT showed lower performance during the experiments compared with RoBERTa so we used RoBERTa as the backbone model and the language model for the latter experiments. This finding might be explained by the fact that the pre-trained datasets for GPT and RoBERTa are distinct from one another, and the sources used by RoBERTa contain more information about social interactions. Second, we found that the performance could be improved by including either the video or the audio modalities, and that the performance could be enhanced even further by integrating both of these modalities jointly.

We also looked into the feasibility of employing emotional features as social cues to distinguish between correct and incorrect answers. We counted the number of emotions from the incorrect answer choices that were not identical to the emotions of the correct answer. As shown in Figure 3, it should be noticed that 72.82% of the samples had at least one incorrect response that did not belong to the same emotional category as the correct response, and these incorrect responses could have contributed as effective negative samples. An ablation analysis of choosing the ratios of β and α for the weights of cross-entropy and contrastive loss is

Table 2. An ablation analysis of the effect of choosing different ratios of α and β for cross-entropy and contrastive losses. The main objective of the model is to predict the correct answer, so we set $\alpha = 1$ and tune the ratio of β and α for analyzing the effects.

β:α	Accuracy (%)
0.1	75.71
0.3	**75.94**
0.5	73.96
0.7	74.12
1	73.67

given in Table 2. When $\beta : \alpha$ is greater than 0.3, which demonstrates the sensitivity toward the final performance, it can be seen that increasing the weight of the contrastive loss could result in a degradation of the performance.

6. Conclusions

In this paper, we demonstrate our model, Multi-Modal Temporal Correlated Network with Emotional Social Cues (MMTC-ESC), for the task of social intelligence question-answering. Our results indicate that combining multi-modal inputs can enhance overall performance, and introducing contrastive loss based on commonsense knowledge can further improve performance. The results also imply that the model's performance is dependent on the selection of the weight for contrastive loss. In this current work, we have only considered manual tuning of loss weights; however, our future work will take into account adaptive tuning of the weights over the course of training.

References

[1] OU Avlaev. The role of social intelligence in personal development. *JournalNX*, pages 692–698, 2020.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783, 2023.

[6] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30, 2013.

[7] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

[8] Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. Multimodal dual attention memory for video story question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 673–688, 2018.

[9] Abhishek Kumar, Trisha Mittal, and Dinesh Manocha. Mcqa: Multimodal co-attention based network for question answering. *arXiv preprint arXiv:2004.12238*, 2020.

[10] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[11] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21273–21282, 2022.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[14] Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L Griffiths. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*, 2018.

[15] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[16] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[18] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[19] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697, 2018.

[20] Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. Multimodal integration of human-like attention in visual question answering.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2647–2657, 2023.

[21] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

[22] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[23] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Birgit A Völlm, Alexander NW Taylor, Paul Richardson, Rhiannon Corcoran, John Stirling, Shane McKie, John FW Deakin, and Rebecca Elliott. Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. *Neuroimage*, 29(1):90–98, 2006.

[26] Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. `https://github.com/abwilf/Social-IQ-2.0-Challenge`, 2023.

[27] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[28] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[29] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021.

[30] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021.

[31] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.

[32] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[33] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[34] Shuai Zhang, Xingfu Wang, Ammar Hawbani, Liang Zhao, and Saeed Hamood Alsamhi. Multimodal graph reasoning and fusion for video question answering. In *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1410–1415. IEEE, 2022.

[35] Ruiying Zhou, Keke Tian, Hanjiang Lai, and Jian Yin. Incorporating domain knowledge and semantic information into language models for commonsense question answering. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1160–1165. IEEE, 2021.

[36] Daoming Zong and Shiliang Sun. Mcomet: Multimodal fusion transformer for physical audiovisual commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6621–6629, 2023.