External Commonsense Knowledge as a Modality for Social Intelligence Question-Answering

Sanika Natu
Carnegie Mellon University
Pittsburgh, USA
natu.sanika@gmail.com

Shounak Sural
Carnegie Mellon University
Pittsburgh, USA
ssural@andrew.cmu.edu

Sulagna Sarkar
Carnegie Mellon University
Pittsburgh, USA
sulagnas@andrew.cmu.edu

Abstract

Artificial Social Intelligence (ASI) refers to the perception and understanding of social interactions. It involves the usage of contextual information about social cues to perform tasks such as Question-Answering (QA) in social situations. In this work, the social intelligence-based Social-IQ dataset consisting of videos with visual, audio, and textual modalities is used for QA in such social contexts. Our approach involves the incorporation of external commonsense knowledge to deal with the lack of reasoning in multimodal machine learning models in the context of question answering. In this work, we use Commonsense Transformers (COMET) to generate contextual information from the textual modality along VisualCOMET for the visual modality. These are incorporated into our model to improve binary QA accuracy over state-of-the-art methods and highlight the need for commonsense understanding in question-answering tasks.

1. Introduction

As artificial intelligence gets integrated increasingly into our day-to-day lives, AI needs to communicate with humans effectively to solve our problems. Human-computer interaction has to happen at a level where AI is able to understand human emotions and is capable of reasoning about topics of importance. For AI to be proficient at such tasks, a branch of artificial intelligence called Artificial Social Intelligence (ASI) is being developed rapidly. This involves the use of machine learning models that have an understanding of social interactions and can reason about them in the context of question-answering (QA) tasks. Humans primarily use vision and audio modalities to understand social scenes. For machine learning models, an additional text modality is often helpful owing to the large-scale development of text-based natural language processing.

In this paper, we perform QA on a social intelligence-based task. Toward this end, we use videos with visual, textual and audio modalities. Videos are naturally composed of a visual and an audio component. Additional transcripts or subtitles are often available based on speaker turns for these videos which add an informative third modality. Datasets such as TVQA [10], Social-IQ [22] and HOW2QA [12] are popular datasets in the broad domain of video-based QA. Social-IQ (Social Intelligence Queries) is a popular dataset in the ASI domain that focuses on situations involving social interactions without much prior context. Our interest lies in question-answering in the multi-modal domain wherein visual, text and audio modalities for each video can be combined effectively for answering questions. The questions involve reasoning and capturing causal relationships between the actors in each video. Emotional aspects of the speakers can play a major role in understanding the context of these conversations. Images extracted from these videos can provide visual context such as facial expressions that can help shape our opinion about people's behaviors and actions. Text is crucial in understanding details of the conversation that is happening between characters in each video to make meaningful choices in QA tasks. Audio conveys information about intonation, pauses and emphasis on words that are lacking in the other modalities. All of these together contain a wide range of information that can contribute towards answering questions based on the videos.

For effectively combining these sources of information in a multi-modal setting, proper alignment of information is required. Additionally, while these videos have a lot of useful information, models trained on these videos do not have information about general social understanding of the world outside the context of these videos.

Our key contributions are as follows:

- We propose two pipelines (COMET-based [2] and VisualCOMET-based [17]) for generating additional commonsense knowledge based on text and image modalities.
- We incorporate generated external knowledge modali-

ties using a knowledge-based loss function and a textual loss function.

- We obtain improvements over the state-of-the-art methods using both our COMET and VisualCOMET-based approaches.

The rest of the paper is structured as follows. Section 2 highlights some prior literature in this area of research. This is followed by Section 3 which delineates our approach of using additional commonsense knowledge-based modalities. Section 4 describes the experiments performed and the dataset used along with notable results. Finally, Section 5 provides a summary of our findings and provides directions for future research.

2. Related work

Video-based question-answering (VQA) has been looked at as a classification problem in datasets such as Social-IQ where the task is to choose one out of two or more plausible answer choices. Multi-modal approaches such as Tensor-MFN [22], MAC-X [18] and F2FCL [19] have looked at transformer and graph-based approaches to solve the problem for social interaction videos. More generalized approaches for language-vision datasets such as VisualBERT [13] and ClipBERT [9] have been proposed that provide end-to-end models for effective modality alignment and representation.

Some research has been conducted where VQA is treated as a generation task instead of a classification task (also known as Extractive QA) [15, 14]. However, due to the complexity of a generation task [15], increased data requirement [16], and the bias of evaluation techniques such as BLEU and ROUGE towards lexical overlap [21], VQA is better suited as a classification task. Current state-of-the-art multimodal models for VQA have focused mostly on transformer-based approaches in order to encode the relationships between modalities. MERLOT Reserve represents videos jointly over time through a contrastive learning approach where the model is expected to predict audio and text tokens that have been masked [23]. MCQA is another multi-modal QA approach that fuses multi-modal input and then uses co-attention between the inputs and the Q&A in order to align multi-modal context to the relevant query [8]. Other methods have focused on incorporating external knowledge into transformer-based models to aid in reasoning tasks. Specifically, Concept-Bert looks at creating a pipeline that fuses an external knowledge base with the vision-language representation in order to find the right answer [5]. Various types of knowledge graphs have emerged as well. VisualSem is a multi-modal knowledge graph consisting of images and sentences based on Wikipedia articles and WordNet synsets [1] whereas COMET is a generative knowledge graph for commonsense reasoning and inferences describing people, entities, and relationships [2]. Current methods add knowledge to the textual modality and focus on literal knowledge as opposed to knowledge specific to social relationships. The novelty of our method is that we focus on commonsense generative knowledge methods for social relationships and introduce knowledge as a separate modality in order to better align relationships across all modalities.

3. Methodology

In this section, we present our proposed approach for integrating external commonsense knowledge as an additional modality into a question-answering task in the ASI domain.

Current multi-modal transformer-based deep-learning models lack inherent intuition or 'commonsense' to aid social reasoning tasks. Hence, integrating an external knowledge graph has become a growing approach in VQA, as it enables the model to leverage commonsense understanding beyond the training dataset to answer questions.

Video-based social intelligence datasets encompass multiple modalities, including images, audio and language, which are encoded to obtain individual representations. In addition to these existing modalities, we introduce external knowledge into the pipeline. Incorporating external knowledge into transformer-based models involves two key considerations: effectively querying relevant commonsense knowledge using the dataset inputs and determining the method of knowledge incorporation. Simpler techniques may involve concatenating knowledge with textual inputs, while more complex approaches treat knowledge as a distinct modality. Our proposed approach explores two methods of querying knowledge and three methods of adding knowledge during model training. By implementing these techniques and experimenting with various combinations of querying and knowledge incorporation, we can evaluate approaches that lead to performance improvement. The external knowledge can be encoded along with the text or treated as a separate modality. The representations of these modalities are then combined into a joint representation, which is used to predict the answer. The proposed architecture is illustrated in Fig.1.

3.1. Querying for External Knowledge

To enhance our baseline model with commonsense reasoning, we leverage COMET[2], which can generate novel commonsense information based on a textual input and a specified attribute (Causes, Intents, Needs). In order to use textual input as a query for COMET, we used the correct answer instead of subtitles since subtitles would not effectively serve as the right kind of descriptions or events that COMET has been trained on. As we cannot use the correct answer at test time, we propose this method as a proof of

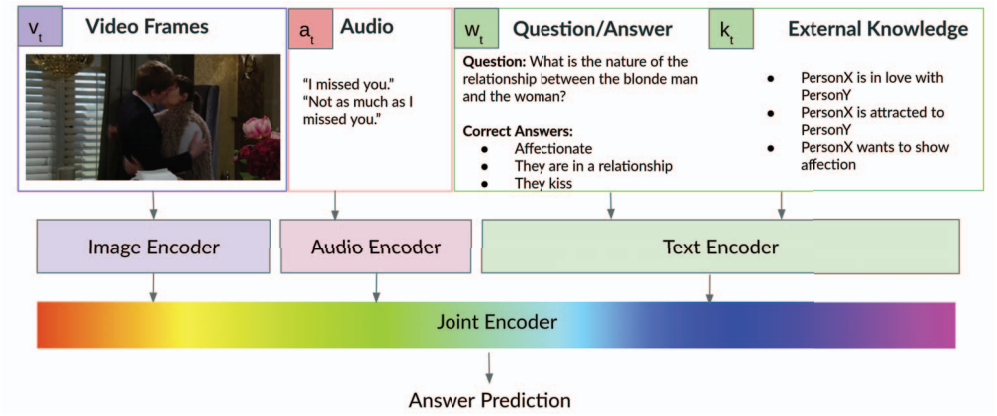

Figure 1. Overview of our proposed architecture for social intelligence QA

concept method where given "correct" knowledge, we assess how the model performs. Hence, the correct answer is used as a text query during training to generate 5 causes that work as external knowledge in the model. We evaluate this method without knowledge at test time in order to assess whether our baseline model has encoded knowledge in its parameters for better QA on unseen samples in the domain.

In addition to COMET which uses a text-based input, we also use VisualCOMET [17] which generates 5 potential "intents" based on image frames from the video. An additional advantage of this approach over COMET is that this method does not require the correct answer (which is the target variable) for querying. Hence, this method can provide extra insight in the form of commonsense external knowledge during test time. VisualCOMET relies on instance information and event descriptions from the visual component of videos.

To ensure that VisualCOMET can be extended to work on images from VQA datasets, we go through a few key steps. Since these datasets (e.g. Social-IQ) typically do not come with visual annotations, we use a state-of-the-art Mask R-CNN network [6] from Detectron2 [20] to generate bounding boxes for instances in each frame. More specifically, the instance information for all the humans and objects of interest present in the image is key to prediction tasks in VisualCOMET. The bounding boxes are generated along with segmentation masks and confidence scores that serve as annotations for the VisualCOMET pipeline. In addition to images, VisualCOMET also requires some additional information as input including an event description and a location. Since these are typically not available for VQA datasets, we use an additional model to get around this problem. Specifically, we use the BLIP model [11] to generate an event description corresponding to each image. An example of a generated description created by the BLIP model for a scene is "two men and a woman are talking to each other sitting in an interview set". In addition to this, we use the VQA head of BLIP to predict a location for the scene such as "indoors", "outdoors" and so on. Once we have the annotated images, the scene description and the location information, we feed these into VisualCOMET to generate commonsense knowledge. This is eventually incorporated as external knowledge in our multi-modal QA model.

3.2. Incorporating External Knowledge

The initial approach to incorporate external knowledge adopts a multitask learning strategy. By sharing representations between two tasks - predicting the correct answer and predicting relevant external knowledge - we aim to enhance the model's ability to generalize in cases requiring external knowledge. In this setting, we trained the baseline model to predict the correct answer 50% of the time and the correct knowledge 50% of the time. This approach encourages the model to develop representations that excel in commonsense reasoning, influenced by the external knowledge integrated during training.

Additionally, we introduce knowledge as an input to the model in two ways: concatenating the knowledge with the textual modality or treating it as a separate modality. In the former approach, the model predicts the correct answer while considering the transcript with the external knowledge concatenated within. In the latter approach, the model aligns the knowledge with the transcript and audio, using it as an additional modality to enhance model prediction.

4. Experiments and analysis

In this section, we examine a social-interaction dataset, analyze the baseline method, conduct error analysis, and explore the impact of adding external knowledge on prediction.

Figure 2. Image sampled from a video in the Social-IQ dataset

4.1. Dataset

We selected the Social-IQ 1.0 dataset as our choice for the topic of artificial social intelligence due to its emphasis on understanding social intelligence in a multi-modal context. The Social-IQ 1.0 dataset [22] serves as a QA benchmark, specifically designed to assess machine understanding of social intelligence within real-life situations. This dataset comprises 1,250 videos, 7,500 questions, and provides 4 correct and 3 incorrect answers per question, amounting to a total of 52,500 answers. What sets the Social-IQ dataset apart from other multi-modal QA datasets is the prevalence of "why" and "how" questions, which demand causal reasoning for accurate answers. Additionally, the answers in this dataset tend to be more extensive, indicating a higher level of detail. An image sampled from a video in the Social-IQ dataset is shown in Figure. 2. It can be seen that a man is performing a stunt show with some knives in his hand while an audience watches him. The Social-IQ dataset has various questions for each such video based on the social interaction happening between people in the video with a focus on causal reasoning.

For our experiments, we use the train and validation split provided by Zadeh et al. [22]. The dataset comprises of 833 training videos, 61 validation videos, and 356 test videos. Each video has 6 questions, each with 7 answer options, consisting of 4 correct and 3 incorrect answers. In the binary setting, the 4 correct and 3 incorrect answers are transformed into 12 correct-incorrect answer pairings, resulting in a total of 72 question-answer pairs per video. As the test data is private, we have reported accuracy on the validation dataset.

The Social-IQ dataset features various input modalities, including language in questions, answers, and audio transcripts, acoustic information extracted from the video's audio, and visual data from the video frames. To encode these modalities, Zadeh et al. employed BERT [4] for generating language embeddings, DenseNet161 [7] for video embeddings, and COVAREP [3] for acoustic representations during their preliminary exploration of modality bias. Furthermore, the introductory Social-IQ paper's baselines utilized LSTMs to encode each input modality separately, followed by concatenating them together [22].

4.2. Modality analysis

To gain a better understanding of the dataset, we conducted uni-modal and multi-modal analyses to assess each modality's effectiveness in identifying social cues. For both multi-modal and uni-modal analyses, 5 annotators watched 3 videos each and answered all 6 questions pertaining to each video. We achieve an accuracy of 92.8% for multi-modal analysis. Our observations revealed the challenge of establishing relations between questions, answers, and speakers, particularly in aligning sentiments and timestamps for accurate responses. Most questions required a combination of modalities, including interpreting facial expressions, intonation, body language, and speaker relationships. We also found that incorporating external knowledge and commonsense reasoning can enhance QA, especially in understanding social settings and human behavior for certain questions. In the human-based uni-modal analysis, we noted that the language modality provides the most information for the QA task, achieving an accuracy of 88%. However, it is heavily affected by the human analyst's implicit external knowledge and educated inferences.

4.3. Multi-modal baseline model

Using the Social-IQ dataset, our objective is a binary QA task, where the models are given an incorrect answer and a correct answer and are expected to predict which is the correct answer using the video about the corresponding social situation as context. For each example *i*, our models are expected to predict the correctness of the provided answer, where each example *i* is decomposed into modality-specific inputs: $x^i_{text}, x^i_{audio}, x^i_{frame}$, and h is the model-specific scoring function.

$$y^i = h(x^i_{text}, x^i_{audio}, x^i_{frame}) \tag{1}$$

The accuracy of the QA task for the binary case is given by Eq.2. Here, $y1$ and $y2$ are the correct and incorrect answer predictions respectively and M is the total number of samples in the set.

$$Accuracy = \frac{1}{M}\sum_{i=1}^{M}(y^i_1 > y^i_2) \tag{2}$$

We investigate several baseline models for the binary QA task on the Social-IQ dataset. Firstly, we utilize the TMFN model, as proposed in the introductory Social-IQ paper [22], achieving an accuracy of 64.82%. Additionally, we employ the F2F-CL model, which utilizes a graph neural network based contrastive-learning approach [19], achieving an accuracy of 78.97%. We explore the MAC-X model,

which combines memory and attention for multiple-choice QA and achieve an accuracy of 71.88%. We achieve the maximum binary prediction accuracy of 83.46% with the MERLOT Reserve model fine-tuned on the Social IQ dataet [23]. We employ the MERLOT Reserve as the baseline model, which utilizes video, text, and audio inputs independently encoded and then fused over time using a joint encoder to achieve optimal performance in QA tasks. The model employs contrastive span training, wherein a video frame and either text or audio are presented, and 25% of the tokens in text and audio segments are MASKed, requiring the model to maximize the similarity between the encoded MASKed and the encoded audio or text segment. The objective is to minimize the cross entropy between the predicted MASK $\hat{w_t}$ and the associated encoded representation w_t from all representations in batch W, as shown in Eq.3

$$L_{mask \rightarrow text} = \frac{1}{|W|} \sum_{w_t \in W} (\log \frac{exp(\sigma \hat{w_t} \bullet w_t)}{\sum_{w \in W} exp(\sigma \hat{w_t} \bullet w_t)}) \quad (3)$$

L_{text} is obtained by normalizing w and $\hat{w}$, scaling the dot product with parameter σ, and adding to the transposed $L_{text \rightarrow mask}$. A similar approach is used to obtain L_{audio}. L_{video} is obtained by maximizing the similarity of the video vectors from the frames to the hidden representation of the entire video based on the video's transcript. The final loss is a summation of all these individual losses as shown in Eq.4

$$Loss = L_{text} + L_{audio} + L_{frame} \quad (4)$$

The MERLOT Reserve is fine-tuned on the Social-IQ dataset, where each instance consists of a question, an answer, a MASK token, video frames, and either subtitles or audio. The model predicts the MASK token to score the correctness of the question-answer pair and is optimized using softmax-crossentropy. During training, we use a learning rate of 5e-6, a batch size of 8, and a weight decay rate of 0.1, training for a single epoch. The baseline model has an accuracy of 83.84% in binary answer prediction task.

We performed an error analysis on the dataset to examine specific instances where the MERLOT Reserve succeeded and failed. Our goal was to gain a better understanding of the failure cases and investigate the potential benefits of incorporating external knowledge. For this purpose, we used the COMET knowledge graph [2], which provides high-quality commonsense knowledge. In our study, we focused on the causal relation, where COMET generates possible causes for the provided answers. Initially, we tested if our baseline model could accurately reconstruct external commonsense knowledge. To achieve this, we formulated a binary task, where the MERLOT Reserve was expected to predict the correct external knowledge from randomly sampled external knowledge derived from other questions of the same video. The results of this analysis are presented in Table 1.

Table 1. Binary accuracy by baseline models

Split	Binary Accuracy
Correctly Predicted Answers	67.61
Incorrectly Predicted Answers	50.70

When MERLOT Reserve provided correct answers, its performance in selecting the appropriate external knowledge was 67% accurate, suggesting the presence of some commonsense information within the model that aids in accurate predictions. However, for incorrectly predicted answers, the ability of MERLOT Reserve to predict the correct external knowledge was only marginally better than chance. This highlights that the baseline model lacks commonsense information for incorrectly predicted answers, and the addition of external commonsense information can assist in improving accuracy.

4.4. Adding external knowledge

In our experiments, we integrated external knowledge from COMET [2] and VisualCOMET [17] using various methods. As mentioned in the method section, we integrate knowledge into the model using either a multitask learning approach or as an input modality (concatenation with text or as a separate modality). Since we use MERLOT Reserve as the baseline, when incorporating knowledge into the input, the masking of tokens also includes the external knowledge. The external knowledge can be MASKed as part of the text modality or as a separate knowledge modality. In the former case, the loss factor for knowledge is incorporated in L_{text} in Eq.4, and in the later case, an additional $L_{knowledge}$ is added to the loss function in Eq.4 that depicts the reconstruction loss of the MASKed knowledge.

The results from Table 2 demonstrate that all approaches, except for the multitask learning setting, show improvement over the baseline. The multi-task learning approach involves prediction of both correct answer and correct external knowledge. Hence, the model parameters encodes the external knowledge through the shared representation between the tasks. The poor performance of the multi-task learning approach can be due to task-interference where the external knowledge prediction task dominates the learning process, hindering the model's ability to effectively utilize the external knowledge for the QA task. Also, the shared representation can cause the model's focus being divided across tasks, which may not be optimal for learning complex relationships between the input modalities and the question, answer. Utilizing external knowledge as a separate modality allows the model to independently encode

Table 2. Performance of approaches using external knowledge in the MERLOT Reserve model. A few baselines of interest are provided followed by our results that show various combinations of querying and external knowledge incorporation.

Method	Binary Accuracy
Tensor-MFN[22]	64.82
MAC-X[18]	71.88
F2F-CL[19]	78.97
MERLOT Reserve Baseline [23]	83.46
COMET MultiTask Learning	82.28
COMET Knowledge Loss	84.37
COMET Textual Loss	**84.83**
VisualCOMET Textual Loss	84.07
Human Performance	95.08

and process knowledge, avoiding potential conflicts and ensuring the knowledge is effectively integrated into the prediction. Thus, we see that all the methods that use external knowledge in the input modality show an improvement in prediction accuracy over the baseline model. Among the methods, the COMET textual loss approach yields the best performance, surpassing the baseline MERLOT Reserve model by 1.37%. Interestingly, a simple concatenation in text form performs better than incorporating a separate knowledge loss term. Our hypothesis suggests that the superior performance of simple concatenation, compared to using knowledge as a separate modality, is attributed to better information fusion and simplicity of the architecture which enable better generalization and leads to improved overall performance in the task. Thus, we concatenate external knowledge from VisualCOMET with the textual information and this method too outperforms the baseline by 0.61%.

While VisualCOMET's commonsense knowledge enhances accuracy compared to the baseline model, the textual knowledge incorporated from COMET demonstrates better predictive performance. We attribute this difference to the inclusion of irrelevant external knowledge from VisualCOMET, where the average knowledge length is 788.34 characters, as opposed to 84.75 characters from COMET. Furthermore, the VisualCOMET approach produces 5 textual explanations for intent per image frame, resulting in 35 knowledge generations per video with 7 frames. This abundance of knowledge potentially skews the model towards external information rather than focusing on the text modality.

Table 3 indicates that all methods incorporating external knowledge can equally retain 93% of correctly predicted answers from the baseline. Additionally, these methods demonstrate substantial improvement in handling incorrectly predicted answers compared to the baseline, with over 30% of the incorrect answers being correctly predicted upon the addition of external knowledge. Among the methods, the COMET Textual Loss approach achieves the most significant enhancement.

Table 3. Percent of correctly predicted examples by the baseline and incorrectly predicted examples by the baseline that each external knowledge approach also got correct

	% of Correct	% of Incorrect
COMET Textual	93.56	39.04
COMET Knowledge	93.89	35.93
VisualCOMET Textual	93.37	35.11

To explore how commonsense is integrated and impacts prediction, the annotators analyze 20 examples with VisualCOMET, incorporating knowledge during test time. For instance, we consider a specific example question: `"Why doesn't the older woman ever admit that she isn't okay?"` with answer options `"She doesn't care about the other woman."` and `"She is headstrong."`. We observe that VisualCOMET captures an individual's intent and personality traits through the knowledge it generates for each example, effectively assisting in QA based on the added knowledge. In this particular example, VisualCOMET generates knowledge like `"know why 1 was there, greet 1 as she approaches, see what 1 wants, hear what 1 has to say"`. From this generated knowledge, we discern that there is respect and care for the other woman, which correctly leads the model to predict the second answer as the correct response.

5. Conclusion and future directions

In this paper, we have proposed an external commonsense knowledge-based approach for QA in social interactions. Commonsense knowledge has been incorporated in the form of textual and visual COMET pipelines that both generate useful knowledge in the form of text. We have used a knowledge-based loss function and a textual loss function to incorporate this knowledge. Our approach has been evaluated on the Social-IQ dataset and we have obtained a binary QA accuracy of 84.83%, which is an improvement over the state-of-the-art MERLOT Reserve model that we use as our baseline. Our approach is general and can be integrated into most multi-modal pipelines that attempt to model social interactions. In future, our approach can be extended to study more effective ways to incorporate the generated commonsense knowledge, or a part of it, into multi-modal models so that they can be used to make more informed decisions.

References

[1] Houda Alberts, Teresa Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. Visualsem: a high-quality knowledge graph for vision and language. *arXiv preprint arXiv:2008.09150*, 2020. 2

[2] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*, 2019. 1, 2, 5

[3] Gilles Degottex. Covarep – a collaborative voice analysis repository for speech technologies. 2014. 4

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 4

[5] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online, Nov. 2020. Association for Computational Linguistics. 2

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 3

[7] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. 4

[8] Abhishek Kumar, Trisha Mittal, and Dinesh Manocha. Mcqa: Multimodal co-attention based network for question answering. *ArXiv*, abs/2004.12238, 2020. 2

[9] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 2

[10] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 1

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[12] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online, Nov. 2020. Association for Computational Linguistics. 1

[13] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[14] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8658–8665, Jul. 2019. 2

[15] Man Luo, Kazuma Hashimoto, Semih Yavuz, Zhiwei Liu, Chitta Baral, and Yingbo Zhou. Choose your qa model wisely: A systematic study of generative and extractive readers for question answering, 2022. 2

[16] Dheeraj Mekala, Tu Vu, Timo Schick, and Jingbo Shang. Leveraging qa datasets to improve generative data augmentation, 2022. 2

[17] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer, 2020. 1, 3, 5

[18] Christina Sartzetaki, Georgios Paraskevopoulos, and Alexandros Potamianos. Extending compositional attention networks for social reasoning in videos. In *Interspeech 2022*. ISCA, sep 2022. 2, 6

[19] Alex Wilf, Martin Q Ma, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Face-to-face contrastive learning for social intelligence question-answering. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7. IEEE, 2023. 2, 4, 6

[20] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[21] An Yang, Kai Liu, Jing Liu, Yajuan Lyu, and Sujian Li. Adaptations of rouge and bleu to better evaluate machine reading comprehension task, 2018. 2

[22] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8799–8809, 2019. 1, 2, 4, 6

[23] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 2, 5, 6