Social-IQ: A Question Answering Benchmark
for Artificial Social Intelligence

Amir Zadeh[1], Michael Chan[1], Paul Pu Liang[2], Edmund Tong[1], Louis-Philippe Morency[1]

[1] Language Technologies Institute, [2] Machine Learning Department

School of Computer Science, Carnegie Mellon University

multicomp.cs.cmu.edu/social-iq
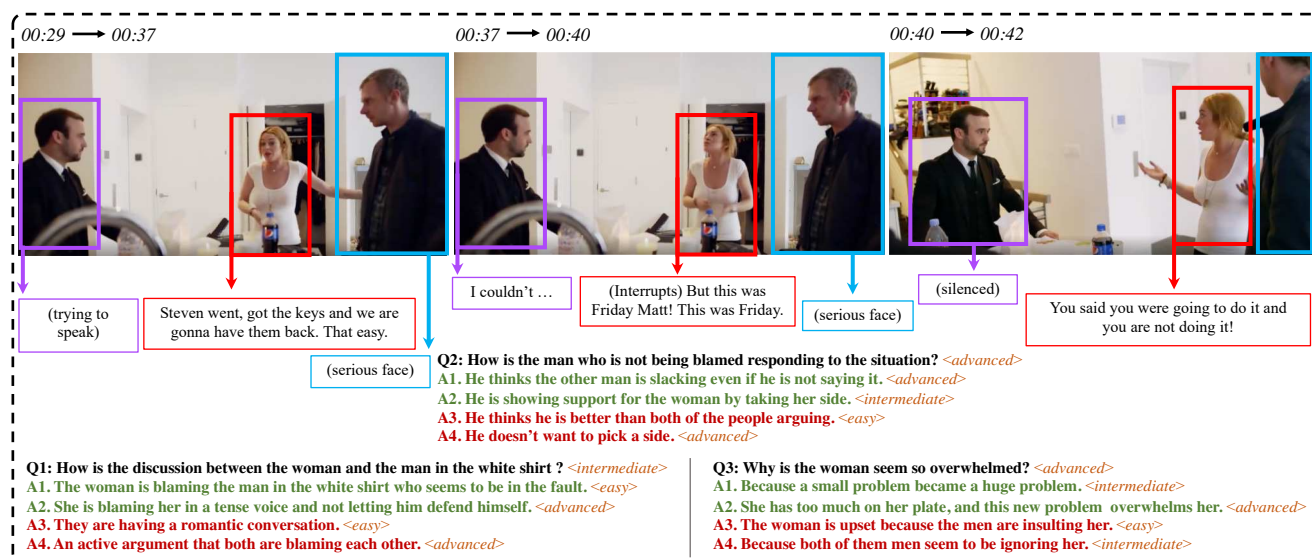
{abagherz,mkchan,pliang,edtong,morency}@cs.cmu.edu

Figure 1: Best viewed zoomed in and in color. An overview of the Social-IQ dataset questions and videos. Social-IQ evaluates artificial social intelligence through question answering. The dataset contains $1,250$ videos, 7500 questions, $30,000$ correct answers and $22,500$ incorrect answers. Questions and answers are annotated for complexity levels: easy, intermediate and advanced. Q indicates questions and A indicates answers. Answers in **green** are correct and answers in **red** are incorrect.

Abstract

As intelligent systems increasingly blend into our everyday life, artificial social intelligence becomes a prominent area of research. Intelligent systems must be socially intelligent in order to comprehend human intents and maintain a rich level of interaction with humans. Human language offers a unique unconstrained approach to probe **through questions** and reason **through answers** about social situations. This unconstrained approach extends previous attempts to model social intelligence through numeric supervision (e.g. sentiment and emotions labels). In this paper, we introduce the Social-IQ, an unconstrained benchmark specifically designed to train and evaluate socially intelligent technologies. By providing a rich source of open-ended questions and answers, Social-IQ opens the door to explainable social intelligence. The dataset contains rigorously annotated and validated videos, questions and answers, as well as annotations for the complexity level of each question and answer. Social-IQ contains $1,250$ natural in-the-wild social situations, $7,500$ questions and $52,500$ correct and incorrect answers. Although humans can reason about social situations with very high accuracy (95.08%), existing state-of-the-art computational models struggle on this task. As a result, Social-IQ brings novel challenges that will spark future research in social intelligence modeling, visual reasoning, and multimodal question answering (QA).

1. Introduction

Definition and studies of social intelligence have a rich history in psychology, sociology and psycholinguistics [44, 47]. These studies aim to evaluate the cognitive process behind understanding social situations; a hidden cognitive process which often goes beyond explicit understanding of meanings and structures [23]. As intelligent systems increasingly become a reality in our every day lives, social intelligence becomes a key part of future artificial intelligence (AI) systems.

Unlike traditional AI systems that can measure a phenomena based on numerical labels, psychometric evaluation of social intelligence requires probes that go beyond numeric labels. To this end, we present the Social-IQ (Social Intelligence Queries) dataset. Social-IQ opens the door to unconstrained and explainable social evaluation and understanding for AI. It contains a rigorously annotated and manually validated set of $7,500$ questions, $52,500$ answers ($30,000$ correct and $22,500$ incorrect) over a broad range of $1,250$ social in-the-wild videos.

Question answering is an effective way of probing the level of understanding of an underlying phenomena [27, 6]. In machine learning, this form of probing has a well established precedence in multiple different areas ranging from understanding books and text [25], to understanding events in the movies [26]. To build a suitable question answering resource for social understanding, Social-IQ strives to analyze social situations as they happen in the wild. Naturalistic interactions are captured by cameras and uploaded to social media on a daily basis from different aspects of life; such as birthday parties or a basketball game. Using an extensive set of YouTube videos, Social-IQ covers a broad range of social and behavioral situations. Furthermore, Social-IQ is diverse in question types and how each question probes social intelligence. The questions also cover a broad range of complexity (advanced, intermediate and easy).

Our contributions in this paper are as follows: 1) We formalize an open-ended question answering task for measuring social intelligence for current and future AI systems. 2) We present the first dataset in this area, called Social-IQ, that focuses on psychometric measurement of social intelligence and operationalizes this measurement through question answering. 3) We analyze the performance of state of the art in multimodal QA over the Social-IQ dataset. Through our experiments, we observe that Social-IQ is a challenging dataset; Humans can achieve very high level of accuracy (95.08%) while state of the art in machine learning (64.82%) trails by a large margin (on a task with 50% random performance). This gap highlights the value of a resource such as Social-IQ; a dataset which enables unconstrained probing of social intelligence.

2. Related Works

The dataset and experiments in this paper are connected to the following areas:

2.1. Question Answering

Intelligent question answering, one of the most ambitious goals of AI, has roots in decades of research in artificial intelligence [17, 54]. In the past few years, there has been a surge of interest in using neural models for intelligent question answering. Recently, question answering has evolved into a multimodal framework. Datasets in this domain started with DAQUAR [33], where image and questions were paired together. Subsequently, four other successful and influential datasets followed which are as follows: COCO-QA [37], VQA [6], FM-IQA [15], Visual7w [63]. In all aforementioned datasets, questions are asked about a single image. More recently, the idea of visual question answering has extended to videos. MovieQA [43] focused on understanding the events in a movie as well as their ordering from movie frames, scripts and plot. Close to this idea, TVQA [26] presented an alternative dataset for the task of understanding movies and plots. In general, compared to visual question answering and textual question answering [59, 11], there is a lack of resources specifically designed to benchmark social intelligence in current and future AI systems.

Social-IQ builds upon lessons learned from previous multimodal datasets and includes some key components: 1) *unconstrained and unscripted environment*: Social-IQ videos come from a diverse set of in-the-wild videos on YouTube. There are diverse sets of distinct characters across these videos. Social situations in these videos are rarely scripted and events are more volatile than movies. 2) *multimodal stimuli*: all questions directly relate to events in the videos and require information from multiple modalities to correctly answer. The question are grounded in variety of manners across video, dialogues, and audio. 3) *annotator bias*: unlike famous movies, arbitrary social online videos are less likely to be seen by annotators prior to the annotation. Furthermore, multiple validation stages are devised for Social-IQ to remove annotation bias and make sure the quality of videos, questions, and answers remains high. 4) *explainability*: annotators of Social-IQ accompany their answers with sufficient reasoning, going beyond short answers consisting of only a few words. Social-IQ answers are longer than previous datasets by nearly a factor of 100% in average length.

2.2. Multimodal Machine Learning

Multimodal machine learning has been among the most successful recent trends in machine learning [7]. Powered by advances in deep learning, multimodal models are

creatively used by research communities centered around tasks such as multimodal language analysis [58], sentiment analysis [46, 31], emotion recognition [29], personality traits recognition [56], image captioning [5, 4, 34], multimedia description [50, 51, 62], and video comprehension [14, 32, 19, 60].

3. Measurement of Social Intelligence

Inspired by past psychological and sociological studies in measuring social intelligence [21, 45, 35, 48, 39, 53], we design the guidelines of Social-IQ according to the following four criteria: 1) Judgment in Social Situations, 2) Processing Human Intelligent Behavior, 3) Understanding Mental State, Trait, Attitude, and Attributes 4) Memory for Referencing and Grounding. Questions in Social-IQ relate to at least one or more of the above social intelligence criteria. What follows is the detailed definition of each of the above criteria with examples:

Judgment in Social Situations: Aligned with sociological definitions developed by Piotr Sztompka [42] and Max Weber [52], we define a social situation as a social exchange or behavior involving two (dyadic) or more individuals. More formally, a social situation involves human physical movement, intentions, and a set of unique interactions in response to one another. Social situations can occur through communications from both verbal and nonverbal channels. Intelligence in this areas includes understanding the causes and intentions behind a social situation. Example acceptable questions for this criteria are: *"Are the people in this group getting along?" (yes, the group seems to be laughing together)*, or *"How is the atmosphere of the room?" (it is tense since the people involved seem to argue and disagree)*. In both cases, questions target the core of a particular social interaction.

Processing Human Intelligent Behavior: This criteria refers to both how and why humans act or react in a certain manner [40]. Example questions to probe human behavior include: *"How did the two men demonstrate that they forgive each other?" (by hugging for a long time)*, or *"Why does the woman pretend to not hear the man?" (she is acting this way because her feelings were hurt by him)*. It is noteworthy that direct action questions are not acceptable based on this criteria. Examples such as *"Is the man lifting weights?" (yes, he is doing so in a gym)* are not acceptable as questions for Social-IQ, because they do not probe social intelligence.

Understanding Mental State, Trait, Attitude and Attributes: We define traits as stable characteristics of personality, while states are temporary behaviors or feelings that depend on a persons situation and motives at a particular time [8]. Both traits and states are manifest through com-

munication and inferable by humans [16]. Furthermore, we define attitude as a person's (or a group's) opinion towards a specific topic [18]. Example acceptable questions for this criteria include: *"Does the man in the black robe seem like he can manage high stress?" (no because a simple problem with his laptop made him panic more than he should)*, *"Why did the woman in the purple skirt call the man in the suit a psychopath?" (she thinks he has no remorse for what he did to her)*. We define human attributes as demonstrating certain manners or consistent behaviors (for example bravery, justness). Example questions for attributes include: *"How did the man in blue shirt show his bravery?" (by standing up to the crowd who were bullying the silent man)*.

Memory for Referencing and Grounding: Aside the above criteria, social intelligence includes comprehending a variety of references through multimodal grounding. This form of grounding goes beyond simple references from one modality (i.e. individual names or appearances). In social situations, even if the identity of a character is not known, humans establish a common grounding to point to entities. For example *"the man with a tense voice"*, or *"the woman who was sad when coming into the house"*. Social-IQ strives to diversify references. Since humans understand these references (as long as the references are deterministic), we encourage a broad range of referencing methods for entities. It is noteworthy that references should be contained within the respective video. As an example, individual names that cannot be inferred from videos are not acceptable.

Beyond the above criteria, it is required that all questions in Social-IQ focus on humans. Therefore, questions focusing on inanimate entities, objects and animals are rejected. For example: *"What is the man picking up?" (a big wooden box)* is not acceptable. However, a question such as *"Is the man lifting the box under pressure?" (Yes, the box is too heavy for him)* is accepted.

Understanding and answering questions in Social-IQ may require different levels of social intelligence. We add a complexity measure for the questions and answers as a subjective approximation of the level of intelligence and reasoning required to process them. Each question or answer (correct or incorrect) is assigned a complexity level. The complexity level is defined as a Likert scale based on 3 levels (easy, intermediate, and advanced) outlining a vote for the level of social intelligence deemed required for answering the question, accepting (as correct), or rejecting (as incorrect) the answer. The easy complexity level is assigned to questions and answers which require simple social intelligence and understanding of the video. For example *"Who is the dominant person in the group?" (the woman in red dress)*, which may require simple understanding of who is speaking and their tone of voice. For the advanced com-

(a) Distribution of question lengths
(b) Distribution of correct (green) and incorrect (red) answer lengths
(c) Distribution of question types
(d) Distribution of # of active characters in videos
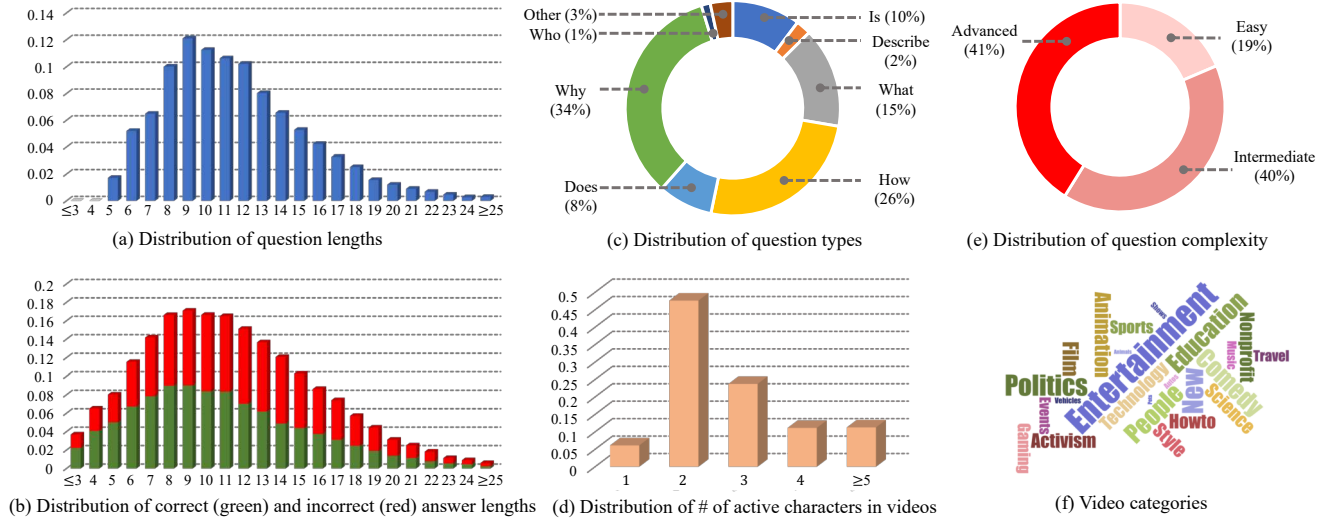(e) Distribution of question complexity
(f) Video categories

Figure 2: Best viewed zoomed in and in color. Social-IQ dataset statistics. (a) demonstrates the distribution of question length in terms of number of words. The average number of words in questions is 10.87. (b) demonstrates the distribution of answer length in terms of number of words with an average of 10.46 words per answer. Both correct (green) and incorrect (red) follow the same distribution. (c) various question types in Social-IQ dataset. (d) distribution of number of active characters in videos. (e) distribution of question complexity in Social-IQ with majority of questions being intermediate and advanced. (f) variety of topics in Social-IQ dataset.

plexity level, questions and answers require in-depth understanding and analysis of the video, characters and their interactions, as well as potential multi-hop inferences and reference resolution. For example: *"What strategy did the woman who disagrees the most with the man choose to confront him?" (she decided to first blame him for being naive, after which she conjectured he is immoral)*. This question answer pair requires understanding the interactions between the characters in the video, as well as how the interaction develops over time.

4. Social-IQ Dataset

In this section we present the details of the Social-IQ (Social Intelligence Queries) dataset which follows the guidelines for measuring social intelligence outlined in the previous section (Section 3). Social-IQ is a question answering (QA) benchmark for assessment of social intelligence in naturalistic social situations. Social-IQ presents $1,250$ videos, $7,500$ questions, and $52,500$ answers ($30,000$ correct and $22,500$ incorrect).

We first comprehensively outline the statistics of the Social-IQ dataset. Afterwards, we discuss the rigorous annotation procedure and multiple validation stages.

4.1. Dataset Statistics

In this subsection we present the main statistics of the Social-IQ dataset. We split the statistics into three parts: a)

questions statistics, b) answers statistics, and c) multimedia statistics.

Question Statistics: The Social-IQ dataset contains a total of 7500 questions (6 per video). Figure 2 (a) demonstrates the distribution of question length in terms of number of words. The average length of questions in Social-IQ is 10.87 words. Figure 2 (c) shows the different question types in the Social-IQ dataset. Questions starting with *why* and *how*, which often require causal reasoning, are the largest group of questions in Social-IQ. This is a unique feature of the Social-IQ dataset and a distinguishing factor of Social-IQ from other multimodal QA datasets (which commonly have *what (object)* and *who* questions as the most common [26, 43]).

Figure 2 (e) demonstrates the distribution of complexity across questions of the Social-IQ. Majority of the dataset consists of advanced and intermediate questions (with almost equal share between the two) while easy questions share a small portion of the dataset. The distribution of question types and complexity levels in Social-IQ demonstrates the challenging nature of the dataset.

Answer Statistics: Social-IQ contains a total of $30,000$ correct (4 per question) and $22,500$ (3 per question) incorrect answers. Figure 2 (b) demonstrates the distribution of word length for answers in the Social-IQ dataset. Both the correct (green) and incorrect (red) answers follow similar distribution. On average, there are a total of 10.46 words

per answer in Social-IQ. This is also a unique characteristic of the Social-IQ dataset since the average answer length is longer than other multimodal QA datasets (with average length between 1.24 to 5.3 words [6, 38, 33, 26, 43]). The long average length demonstrates the level of detail included in Social-IQ answers.

Presence of multiple correct answers in the Social-IQ dataset allows for modeling diversity and subjectivity across annotators in cases where multiple explanations are correct for a certain question. Furthermore, having multiple correct answers enables answer generation tasks (which often require multiple correct answers for successful evaluation [36, 30]).

Multimedia Statistics: Social-IQ dataset consists of a total of $1,250$ videos from YouTube. Figure 2 (f) demonstrates an overview of categories of the videos in Social-IQ. There is a total of $1,239$ minutes of annotated video content (across $10,529$ minutes of full videos). Figure 2 (d) shows the distribution of number of characters in videos. All the videos in the Social-IQ dataset contain manual transcriptions with detailed timestamps.

4.2. Annotation Procedure

Annotation of the Social-IQ dataset is carried out in 6 distinct stages (Figure 3). A total of 50 annotators [1] worked across these multiple stages over a period of 14 months (across three annotation seasons). Before annotating, the annotators went through several training sessions (discussed in Subsection 4.3) for this task to build a proper understanding of measuring social intelligence as defined in Section 3. The details of these 6 stages is as follows:

Video Acquisition Stage: Online social media platforms, including YouTube, contain a large cache of in-the-wild videos with variety of social situations. As a first step, a set of $2,000$ videos were harvested from YouTube [2] using a broad set of search terms. The choice of these search terms followed precedence previously established by the CMU-MOSEI dataset [58] which contains 250 diverse search terms. We required all the videos to maintain at least one face (detected using MTCNN [61]) in 80% of the frames. A set of $2,000$ videos were acquired using this strategy.

Video Validation Stage: After acquiring the initial set of videos, for each video, two trained annotators inspected the video to make sure a social situation exists. Specifically, annotators looked at presence of social interactions, opinion sharing and communication. A total of $1,250$ videos were selected this way.

Question Creation Stage: During this stage, expert anno-
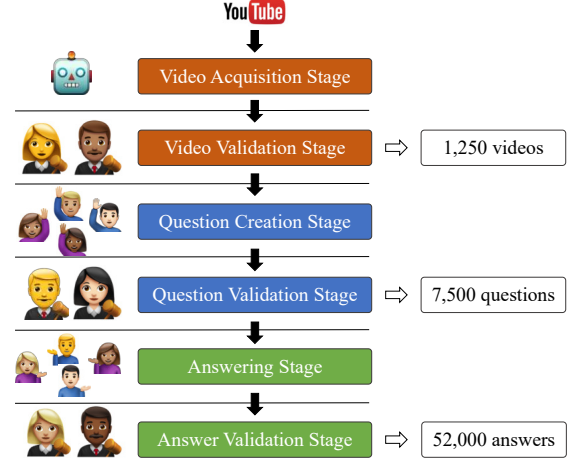
---

¹

Figure 3: The 6 stages used to create the Social-IQ dataset. Video Acquisition and Validation Stage produce $1,250$ videos with social situations in them. Question Creation and Validation Stages produce $7,500$ questions. Answering and Answer Validation Stage produce $52,500$ answers ($30,000$ correct and $22,500$ incorrect).

tators were tasked to ask questions that probe social intelligence as defined by Section 3. Given a video, two trained annotators were instructed to each ask 3 questions. The annotators were also instructed to keep their questions diverse with high level of complexity. They proposed one correct and one incorrect answer for each question. Furthermore, they labelled their questions and answers with complexity labels. After this stage, each video consists of 6 questions, 6 correct answers and 6 incorrect answers.

Question Validation Stage: Given the set of 6 questions, we ask a separate set of 2 annotators to validate the questions (whether or not they comply with definitions in Section 3). If an annotator disputes the validity of a question, the question is removed and passed to Question Creation Stage for re-annotation. A similar procedure is performed on the answer. Furthermore, the two annotators label each question and answer with complexity labels.

Answering Stage: A set of two annotators answer the 6 questions for each video (3 for each annotator). These annotators are different than the annotators who asked the questions in Question Creation Phase and validated the question in Question Validation Stage. Each annotator creates 3 correct and 2 incorrect answers for each question (without knowledge of any prior correct or incorrect answers from Question Creation Stage). Similar to Question Creation Stage, the annotators are encouraged to keep their answers diverse. Annotators also label each answer with complexity levels. After this stage, each question contains 4 correct an-

---

¹Hired and trained undergraduate students from Carnegie Mellon University.

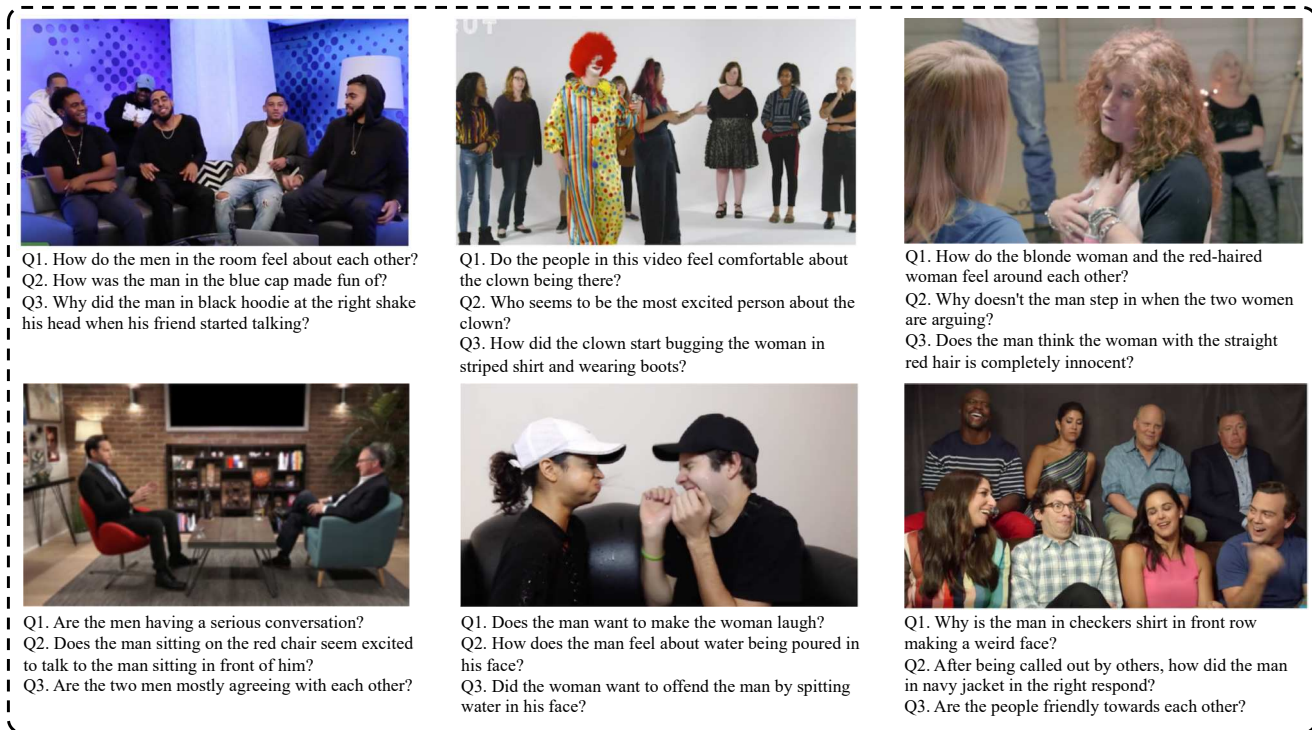²Videos followed creative commons license.

Figure 4: Example videos and questions in Social-IQ dataset, a benchmark for assessment of social intelligence in naturalistic social situations. In-the-wild online videos exhibit various social situations which form the basis of the Social-IQ dataset. Social-IQ presents 1250 videos, 7500 questions, and $52,500$ answers ($30,000$ correct and $22,500$ incorrect).

swers and 3 incorrect answers (including the 1 correct and 1 incorrect from Question Creation Stage).

Answer Validation Stage: Similar to Question Validation Phase, a set of 2 annotators (different than annotators in Question Creation and Answering Stages) validate each answer. Answers are validated for diversity and whether or not they are correctly labelled (if correct/incorrect answers are indeed correct/incorrect). Furthermore, they label the answers with complexity level.

After the above stages, the set of $1,250$ videos, $7,500$ questions, and $52,500$ correct and incorrect answers shape the Social-IQ dataset. Figure 4 shows examples of some videos in the Social-IQ dataset along with the annotations for questions, correct answers and incorrect answers.

4.3. Annotator Training

Due to the rigorous nature of the guidelines in Section 3, a detailed annotator selection and training process is required to achieve high quality annotations. The training process was split into 3 stages:

Initial Training Stage: The first stage of training involves in-depth understanding of the crtieria in Section 3. Annotators were trained during a single training session where the Social-IQ criteria were defined. Annotators also learned how to annotate the data through a designated online annotation system built for Social-IQ. A generic implementation of this annotation system called CMU-Crowd: https://github.com/A2Zadeh/CMU-Crowd is available for academic use.

Secondary Training Stage: Before using the online annotation system, annotators were given training videos for each of the annotation stages in Subsection 4.2. After watching the videos, annotators finished a set of 10 training examples from Question Creation Stage before beginning Social-IQ annotations.

Continuous Supervision Stage: The performance of annotators was continuously monitored on a weekly basis by the authors. A set of 8 annotation workshops were held throughout a period of a year. Annotators with low quality of work were asked to attend individual meetings for retraining. It is noteworthy that the training and supervision throughout the annotation timeline was designed to encourage creativity and diversity of questions and answers. None of our measures stopped annotators from exploring new directions of asking and answering questions. In fact, annotators were incentivized through monetary gifts based on

the creativity of their annotations and their ability to bring questions and answers from areas that were unexplored previously.

5. Experiments

The first goal of our experiments is to analyze the performance of state of the art on Social-IQ. We conduct extensive evaluation of top performing models across scoreboards of MovieQA [43][3], TVQA [26] [4], and CMU-MOSEI [58]. We compare the performance of these models with each other and human level performance in binary and multiple choice setups. In binary case, models are given an answer and are expected to predict whether or not the answer is correct or incorrect. In multiple choice case, models should pick the correct answer from a set of 4 answers (3 of which are incorrect).

The second goal of our experiments is to identify any potential biases in the Social-IQ. Models that target biases are by design simple models that demonstrate whether or not there is any trivial yet frequently occurring pattern in the data that can be exploited during training. The following baselines aim to explore these biases. In all these baselines, a LSTM model is used to encode sequential information for each input modality, and answers are conditioned on the concatenation of input encodings similar to [26]. We first outline the models for exploiting biases, followed by state-of-the-art performing models on relevant tasks.

Q+A: We study the predictability of correct and incorrect answers given only question and answers (no video, audio or transcript). This baseline demonstrates whether or not there exists a pattern across correct or incorrect answers which can lead to identifying the correctness without any context from videos. We use BERT embeddings [11] as contextual distributed word representations for language. BERT embeddings have shown to be suitable representations for both common sense reasoning and question answering.

Q+A+T: This bias demonstrates the usefulness of transcripts (T) in predicting the correct and incorrect answers. Similar to Q+A baselines, distributional features of T are also extracted using BERT embeddings [5]. The sequence of embeddings for T are then encoded using an LSTM.

Q+A+V: This bias demonstrates the usefulness of visual modality (V) through using holistic visual embeddings in predicting the correct and incorrect answers. We use representations extracted from DenseNet161 [20] (last mean

---

Baseline	Accuracy	
Metric	A2 ↑	A4 ↑
Random	50.00	25.00
Q+A (BERT) [11]	57.02	28.61
Q+A+T (BERT) [11]	57.87	29.36
Q+A+Ac (BERT+COVAREP)	57.22	29.58
Q+A+V (BERT+DenseNet161)	63.91	32.62
LMN [49]	61.12	31.81
FVTA [28]	60.88	31.01
E2EMemNet [41]	62.58	31.46
MDAM [24]	60.23	30.71
MSM [26]	59.96	29.89
TFN [55]	63.15	29.82
MFN [56]	62.78	30.86
Tensor-MFN	**64.82**	**34.14**
Human	95.08	-

Table 1: Performance of various models including state of the art in MovieQA [43], TVQA [26], and CMU-MOSEI [57]. A2 ↑ demonstrates the binary accuracy and A4 ↑ demonstrates multiple (four) choice (higher is better). There is a large discrepancy between human level performance and neural state of the art, a gap of 30.26% in binary question answering task.

pooling layer, 2208 dimensions) for each frame. While videos are originally in 30fps sampling rate, we only use 1fps for baseline experiments.

Q+A+Ac: This bias demonstrates the usefulness of acoustic (Ac) modality in predicting the correct and incorrect answers. We use low and high level acoustic representations from COVAREP [10] including 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features [12], glottal source parameters [9, 13, 1, 3, 2], peak slope parameters and maxima dispersion quotients [22].

The following baselines are among the state of the art for TVQA [26], MovieQA [43] and CMU-MOSEI [58] dataset. We pick these baselines based on their performance and structural diversity.

End2End Multimodal Memory Network (E2EMMemNet): This baseline has shown promising performance on MovieQA dataset. We implement this baseline based on the original implementation [41] and multimodal extensions using DenseNet161 features and COVAREP.

Multimodal Dual Attention Memory (MDAM) [24]: This baseline is also among the top performing for MovieQA dataset. MDAM uses two attentions: 1) self-attention (temporal) based on visual frames and 2) cross-attention based on question. Afterwards, the answering is done using a deep recurrent neural network.

Layered Memory Network (LMN) [49]: This baseline is the winner of "The Joint Video and Language Understanding Workshop" in ICCV 2017 [6] and still a strong performing model for MovieQA. The baseline has two main modules: Static Word Memory Module, which builds a representation of the transcription words based on the visual frames, and Dynamic Subtitle Memory Module, which builds a representation of the transcription sentences based on the high level descriptors of the frames.

Focal Visual-Text Attention (FVTA) [28]: FVTA is a strong baseline for MovieQA. This baseline proposed a new form of attention called Focal Visual-Text (FVT); an extension of attention which uses outer-product to build a joint multimodal space.

Multi-stream Memory (MSM) [26]: This baseline is the top performing baseline of TVQA dataset. Multiple streams of data from visual, acoustic and language are fused together to answer questions. All the modalities are embedded using recurrent networks and fused together in subsequent stages to answer questions.

Tensor Fusion Network (TFN) [55]: Originally proposed for multimodal sentiment analysis, we extend this model for question answering by conditioning the answer based on an outer tensor-product of embeddings of transcript, visual and acoustic modalities. A strong aspect of TFN is performing fusion on unimodal, bimodal and trimodal components of the data. Before fusion, the modalities are summarized using three LSTMs. The output of fusion is added to the question and answer to make a final prediction.

Memory Fusion Network (MFN) [56]: This models is used for the tasks of sentiment analysis, emotion recognition and personality traits recognition. It uses a delta-memory attention which stores the sequntial changes of memory across multiple LSTMs. Afterwards, it performs multimodal fusion over the changes in modalities and stores the information in a separate memory. MFN model uses alignment information between transcript, audio and video which is an important component specifically in understanding multimodal language [58].

Tensor-MFN is a baseline created by performing architecture and hyperparameter search on TFN and MFN models and combining them into a joint model. In simple terms, Tensor-MFN uses DenseNet161 scene embeddings and Tensor Fusion for multimodal fusion in the recurrent stages of MFN.

Human Performance demonstrates the human performance (annotators did not see the question and the video prior) in picking correct answer for question-answer pair in

binary format, similar to the setup used for all the baselines.

6. Results and Discussion

Table 1 demonstrates the performance of the baselines in Subsection 5. At a first glance, our bias analysis experiments demonstrate minimal bias in the Social-IQ dataset coming from Q+A. BERT embeddings, commonly known for their success in common-sense reasoning, show slightly higher performance than random. This essentially demonstrates that common sense reasoning purely by looking at question and answers is not enough for answering the questions in Social-IQ. Answering questions in the Social-IQ dataset requires both common sense and context. Contextual information from T, Ac, and V are able to improve the answering performance. Specifically, the improvement is highest by adding visual information from DenseNet161.

Aside the performance of bias analysis models, results of state of the art models from MovieQA, TVQA, and CMU-MOSEI are reported in Table 1. Human performance (calculated during the validation stage) is reported 95.08% for the binary task. The gap between state of the art model and human performance remains large. This signifies the challenging nature of Social-IQ dataset and the necessity of further research in this direction.

7. Conclusion

To conclude, this paper introduced Social-IQ (Social Intelligence Queries), a pioneer real-world unconstrained dataset designed to evaluate the social intelligence and capabilities of existing and future AI technologies. Social-IQ also focuses on the explainability of models by using open-ended answers to model the rationale behind the model's comprehension of social intelligence. The rigorously annotated dataset contains 7, 500 questions with 52, 500 answers spanning across 1, 250 natural social situations. Our experimental results show that although humans can reason about open-ended social intelligence with high accuracy (95.08%), existing QA models struggle on this task. As a result, Social-IQ is a challenging dataset that we hope will instigate future research in social intelligence modeling, visual reasoning, and multimodal QA. The dataset is made publicly available for research purposes alongside provided features.

Acknowledgment

---

[6] http://movieqa.cs.toronto.edu/workshops/iccv2017/

References

[1] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3):109–118, 1992.

[2] Paavo Alku, Tom Bäckström, and Erkki Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

[3] Paavo Alku, Helmer Strik, and Erkki Vilkman. Parabolic spectral parametera new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79, 1997.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[7] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[8] William F Chaplin, Oliver P John, and Lewis R Goldberg. Conceptions of states and traits: dimensional attributes with ideals as prototypes. *Journal of personality and social psychology*, 54(4):541, 1988.

[9] Donald G Childers and CK Lee. Vocal quality factors: Analysis, synthesis, and perception. *the Journal of the Acoustical Society of America*, 90(5):2394–2410, 1991.

[10] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarepa collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE, 2014.

[11] 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.

[12] Thomas Drugman and Abeer Alwan. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, pages 1973–1976, 2011.

[13] Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):994–1006, 2012.

[14] Lijie Fan, Wenbing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, and Junzhou Huang. End-to-end learning of motion representation for video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[15] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304, 2015.

[16] Jennifer M George. State or trait: Effects of positive mood on prosocial behaviors at work. *Journal of applied Psychology*, 76(2):299, 1991.

[17] Bert F Green Jr, Alice K Wolf, Carol Chomsky, and Kenneth Laughery. Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224. ACM, 1961.

[18] Eduard H Hovy. What are sentiment, affect, and emotion? applying the methodology of michael zock to sentiment analysis. In *Language production, cognition, and the Lexicon*, pages 13–24. Springer, 2015.

[19] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[21] Thelma Hunt. The measurement of social intelligence. *Journal of Applied Psychology*, 12(3):317, 1928.

[22] John Kane and Christer Gobl. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179, 2013.

[23] John F Kihlstrom and Nancy Cantor. Social intelligence. *Handbook of intelligence*, 2:359–379, 2000.

[24] Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. Multimodal dual attention memory for video story question answering. *arXiv preprint arXiv:1809.07999*, 2018.

[25] Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association of Computational Linguistics*, 6:317–328, 2018.

[26] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

[27] Willem J.M Levelt and Stephanie Kelter. Surface form and memory in question answering. *Cognitive Psychology*, 14(1):78 – 106, 1982.

[28] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander Hauptmann. Focal visual-text attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6135–6143, 2018.

[29] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, 2018.

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10, 2004.

[31] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[32] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014.

[34] Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] Maureen O'sullivan et al. Measurement of social intelligence. 1965.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[37] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015.

[38] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203. Association for Computational Linguistics, 2013.

[39] David M Romney and Michael C Pyryt. Guilford's concept of social intelligence revisited. *High Ability Studies*, 10(2):137–142, 1999.

[40] Constanze Rossmann. *Theory of reasoned action-theory of planned behavior*. Nomos Verlagsgesellschaft mbH & Co. KG, 2010.

[41] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[42] Piotr Sztompka. Socjologia znak. 2002.

[43] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[44] Edward L Thorndike. Intelligence and its uses. *Harper's magazine*, 1920.

[45] Robert L Thorndike and Saul Stein. An evaluation of the attempts to measure social intelligence. *Psychological Bulletin*, 34(5):275, 1937.

[46] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[47] Philip E Vernon. Some characteristics of the good judge of personality. *The Journal of Social Psychology*, 4(1):42–57, 1933.

[48] Ronald E Walker and Jeanne M Foley. Social intelligence: Its history and measurement. *Psychological Reports*, 33(3):839–864, 1973.

[49] Bo Wang, Youjiang Xu, Yahong Han, and Richang Hong. Movie question answering: Remembering the textual cues for layered visual contents. In *AAAI*, 2018.

[50] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[51] Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. M3: Multimodal memory modelling for video captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[52] Max Weber. The nature of social action. *Weber: Selections in translation*, pages 7–32, 1978.

[53] Susanne Weis and Heinz-Martin Süß. Reviving the search for social intelligence–a multitrait-multimethod study of its structure and construct validity. *Personality and individual differences*, 42(1):3–14, 2007.

[54] William A Woods. Semantics and quantification in natural language question answering. In *Advances in computers*, volume 17, pages 1–87. Elsevier, 1978.

[55] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Empirical Methods in Natural Language Processing, EMNLP*, 2017.

[56] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[57] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[58] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2236–2246, 2018.

[59] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.

[60] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. *CoRR*, abs/1611.04021, 2016.

[61] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[62] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[63] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.