Cognitive Visual Commonsense Reasoning Using Dynamic Working Memory

Xuejiao Tang[1], Xin Huang[2], Wenbin Zhang[2]
Travers B. Child[3], Qiong Hu[4], Zhen Liu[5], and Ji Zhang[6]

[1]Leibniz University of Hannover, Germany [2]University of Maryland, Baltimore County, USA
[3]China Europe International Business School, China [4]Auburn University, USA
[5]Guangdong Pharmaceutical University, China [6]University of Southern Queensland, Australia
[1]xuejiao.tang@stud.uni-hannover.de,[2]{xinh1,wenbinzhang}@umbc.edu
[3]t.b.child@ceibs.edu,[4]qzh0011@auburn.edu
[5]liu.zhen@gdpu.edu.cn,[6]ji.zhang@usq.edu.au

Abstract. Visual Commonsense Reasoning (VCR) predicts an answer with corresponding rationale, given a question-image input. VCR is a recently introduced visual scene understanding task with a wide range of applications, including visual question answering, automated vehicle systems, and clinical decision support. Previous approaches to solving the VCR task generally rely on pre-training or exploiting memory with long dependency relationship encoded models. However, these approaches suffer from a lack of generalizability and prior knowledge. In this paper we propose a dynamic working memory based cognitive VCR network, which stores accumulated commonsense between sentences to provide prior knowledge for inference. Extensive experiments show that the proposed model yields significant improvements over existing methods on the benchmark VCR dataset. Moreover, the proposed model provides intuitive interpretation into visual commonsense reasoning. A Python implementation of our mechanism is publicly available at https://github.com/tanjatang/DMVCR

1 Introduction

Reflecting the success of Question Answering (QA) [1] research in Natural Language Processing (NLP), many practical applications have appeared in daily life, such as Artificial Intelligence (AI) customer support, Siri, Alex, etc. However, the ideal AI application is a multimodal system integrating information from different sources [2]. For example, search engines may require more than just text, with image inputs also necessary to yield more comprehensive results. In this respect, researchers have begun to focus on multimodal learning which bridges vision and language processing. Multimodal learning has gained broad interest from the computer vision and natural language processing communities, resulting in the study of Visual Question Answering (VQA) [3]. VQA systems predict answers to language questions conditioned on an image or video. This is challenging for the visual system as often the answer does not directly refer to the image or video in question. Accordingly, high demand has arisen for AI models with cognition-level scene understanding of the real world. But presently, cognition-level scene understanding remains an open, challenging problem. To tackle this problem,

Rowan Zeller et al. [4] developed Visual Commonsense Reasoning (VCR). Given an image, a list of object regions, and a question, a VCR model answers the question and provides a rationale for its answer (both the answer and rationale are selected from a set of four candidates). As such, VCR needs not only to tackle the VQA task (i.e., to predict answers based on a given image and question), but also provides explanations for why the given answer is correct. VCR thus expands the VQA task, thereby improving cognition-level scene understanding. Effectively, the VCR task is more challenging as it requires high-level inference ability to predict rationales for a given scenario (i.e., it must infer deep-level meaning behind a scene).

The VCR task is challenging as it requires higher-order cognition and commonsense reasoning ability about the real world. For instance, looking at an image, the model needs to identify the objects of interest and potentially infer people's actions, mental states, professions, or intentions. This task can be relatively easy for human beings in most situations, but it remains challenging for up-to-date AI systems. Recently, many researchers have studied VCR tasks (see, e.g., [4–8]). However, existing methods focus on designing reasoning modules without consideration of prior knowledge or pre-training the model on large scale datasets which lacks generalizability. To address the aforementioned challenges, we propose a Dynamic working Memory based cognitive Visual Commonsense Reasoning network (DMVCR), which aims to design a network mimicking human thinking by storing learned knowledge in a dictionary (with the dictionary regarded as prior knowledge for the network). In summary, our main contributions are as follows. First, we propose a new framework for VCR. Second, we design a dynamic working memory module with enhanced run-time inference for reasoning tasks. And third, we conduct a detailed experimental evaluation on the VCR dataset, demonstrating the effectiveness of our proposed DMVCR model.

The rest of this paper is organized as follows. In section 2 we review related work on QA (and specifically on VCR). Section 3 briefly covers notation. In section 4 we detail how the VCR task is tackled with a dictionary, and how we train a dictionary to assist inference for reasoning. In section 5 we apply our model to the VCR dataset. Finally, in section 6 we conclude our paper.

2 Related Work

Question answering (QA) has become an increasingly important research theme in recent publications. Due to its broad range of applications in customer service and smart question answering, researchers have devised several QA tasks (e.g., Visual Question Answering (VQA) [3], Question-Answer-Generation [9]). Recently, a new QA task named VCR [4] provides answers with justifications for questions accompanied by an image. The key step in solving the VCR task is to achieve inference ability. There exists two major methods of enhancing inference ability. The first focuses on encoding the relationship between sentences using sequence-to-sequence based encoding methods. These methods infer rationales by encoding the long dependency relationship between sentences (see, e.g., R2C [4] and TAB-VCR [6]). However, these models face difficulty reasoning with prior knowledge, and it is hard for them to infer reason based on commonsense about the world. The second method focuses on pre-training [7, 10, 8].

Such studies typically leverage pre-training models on more than three other image-text datasets to learn various abilities like masked multimodal modeling, and multimodal alignment prediction [8]. The approach then regards VCR as a downstream fine-tuning task. This method however lacks generalizability.

Considering the disadvantages of either aforementioned approach, we design a network which provides prior knowledge to enhance inference ability for reasoning. The idea is borrowed from human beings' experience – prior knowledge or commonsense provides rationale information when people infer a scene. To achieve this goal, we propose a working memory based dictionary module for run-time inference. Recent works such as [11–13] have successfully applied the working memory into QA, VQA, and image caption. Working memory provides a dynamic knowledge base in these studies. However, existing work focuses on textual question answering tasks, paying less attention to inference ability [11, 12]. Concretely, the DMN network proposed in [14] uses working memory to predict answers based on given textual information. This constitutes a step forward in demonstrating the power of dynamic memory in QA tasks. However, that approach can only tackle textual QA tasks. Another work in [12] improves upon DMN by adding an input fusion layer (containing textual and visual information) to be used in VQA tasks. However, both methods failed to prove the inference ability of dynamic working memory. Our paper proposes a dictionary unit based on dynamic working memory to store commonsense as prior knowledge for inference.

3 Notations and Problem Formulation

The VCR dataset consists of millions of labeled subsets. Each subset is composed of an image with one to three associated questions. Each question is then associated with four candidate answers and four candidate rationales. The overarching task is formulated as three subtasks: (1) predicting the correct answer for a given question and image ($Q \rightarrow A$); (2) predicting the correct rationale for a given question, image, and correct answer ($QA \rightarrow R$); and (3) predicting the correct answer and rationale for a given image and question ($Q \rightarrow AR$). Additionally, we defined two language inputs - query $q\{q_1, q_2, \cdots, q_n\}$ and response $r\{r_1, r_2, \cdots, r_n\}$, as reflected in Figure 1. In the $Q \rightarrow A$ subtask, query q is the question and response r is the answers. In the $QA \rightarrow R$ subtask, query q becomes the question together with correct answer, while rationales constitute the response r.

4 Proposed Framework

As shown in Figure 1, our framework consists of four layers: a feature representation layer, a multimodal fusion layer, an encoder layer, and a prediction layer. The first layer captures language and image features, and converts them into dense representations. The represented features are then fed into the multimodal fusion layer to generate meaningful contexts of language-image fused information. Next, the fused features are fed into an encoder layer, which consists of a long dependency encoder [15] RNN module along with a dictionary unit. Finally, a prediction layer is designed to predict the correct answer or rationale.

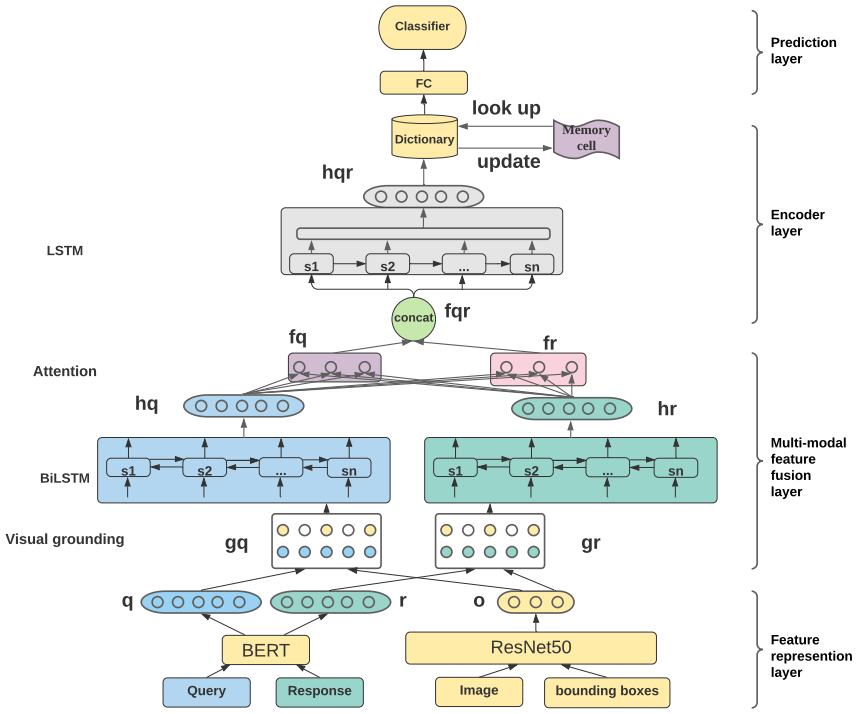


Fig. 1. High-level overview of the proposed DMVCR consisting of four modules: feature representation layer to extract visual and textual features; multimodal feature fusion to contextualize multimodal representations; encoder layer to encode rich visual commonsense; and prediction layer to select the most related response.

4.1 Feature Representation Layer

The feature representation layer converts features from images and language into dense representations. For the language, we learn embeddings for the query $q\{q_1, q_2, \cdots, q_n\}$ and response $r\{r_1, r_2, \cdots, r_n\}$ features. Additionally, the object features $o\{o_1, o_2, \cdots, o_n\}$ are extracted from a deep network based on residual learning [16].

Language embedding. The language embeddings are obtained by transforming raw input sentences into low-dimensional embeddings. The query represented by $q\{q_1, q_2, \cdots, q_n\}$ refers to a question in the question answering task ($Q \rightarrow A$), and a question paired with correct answer in the reasononing task ($QA \rightarrow R$). Responses $r\{r_1, r_2, \cdots, r_n\}$ refer to answer candidates in the question answering task ($Q \rightarrow A$), and rationale candidates in the reasoning task ($QA \rightarrow R$). The embeddings are extracted using an attention mechanism with parallel structure [17]. Note that the sentences contain tags related to objects in the image. For example, see Figure 3(a) and the question "Are [0,1] happy to be here?" The [0,1] are tags set to identify objects in the image (i.e., the object features of person 1 and person 2).

Object embedding. The images are filtered from movie clips. To ensure images with rich information, a filter is set to select images with more than two objects each [4]. The object features are then extracted with a residual connected deep network [16].

The output of the deep network is object features with low-dimensional embeddings $o\{o_1, o_2, \cdots, o_n\}$.

4.2 Multimodal Feature Fusion Layer

The multimodal feature fusion layer consists of three modules: a visual grounding module, an RNN module, and an attention module.

Visual grounding. Visual grounding aims at finding out target objects for query and response in images. As mentioned in section 4.1, tags are set in query and responses to reference corresponding objects. The object features will be extracted and concatenated to language features at the visual grounding unit to obtain the representations with both image and language information. As shown in Figure 1, the inputs of visual grounding consist of language ($q\{q_1, q_2, \cdots, q_n\}$ and $r\{r_1, r_2, \cdots, r_n\}$) along with related objects features ($o\{o_1, o_2, \cdots, o_m\}$). The output contains aligned language and objects features (g_q and g_r). The white unit at visual grounding is grounded representations, which contains image and text information. It can be formulated as follows (where $concat$ represents the concatenate operation):

$$g_r = concat(o, r) \tag{1}$$

$$g_q = concat(o, q) \tag{2}$$

RNN module. The grounded language and objects features g_q and g_r at the visual grounding stage contain multimodal information from images and text. However, they cannot understand the semantic dependency relationship around each word. To obtain language-objects mixed vectors with rich dependency relationship information, we feed the aligned language features g_q and g_r into BiLSTM [15], which exploits the contexts from both past and future. In details, it increases the amount of information by means of two LSTMs, one taking the input in a forward direction with hidden layer $\overrightarrow{h_{lt}}$, and the other in a backwards direction with hidden layer $\overleftarrow{h_{lt}}$. The query-objects representations and response-objects $\underset{l\in\{q,r\}}{h_l}$ output at each time step t is formulated as:

$$h_{lt} = \overrightarrow{h_{lt}} \oplus \overleftarrow{h_{lt}} \tag{3}$$

$$\overrightarrow{h_{lt}} = o_t \odot \tanh(c_t) \tag{4}$$

where c_t is the current cell state and formulated as:

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c \cdot [c_{t-1}, h_{l(t-1)}, x_t] + b_c) \tag{5}$$

$$i_t = \sigma(W_i \cdot [c_{t-1}, h_{l(t-1)}, x_t] + b_i) \tag{6}$$

$$o_t = \sigma(W_o \cdot [c_{t-1}, h_{l(t-1)}, x_t] + b_o) \tag{7}$$

$$f_t = \sigma(W_f \cdot [c_{t-1}, h_{l(t-1)}, x_t] + b_f) \tag{8}$$

where i, o, f represent input gate, output gate, and forget gate, respectively, and x_t is the t^{th} input of a senquence. In addition, $W_i, W_o, W_f, W_c, b_c, b_i, b_o, b_f$ are trainable parameters with σ representing the sigmoid activation function [18].

Attention module. Despite the good learning of BiLSTM in modeling sequential transition patterns, it is unable to fully capture all information from images and languages. Therefore, an attention module is introduced to enhance the RNN module, picking up object features which are ignored in the visual grounding and RNN modules. The attention mechanism on object features $o\{o_1, o_2, \cdots, o_n\}$ and response-objects representations h_r is formulated as:

$$\alpha_{i,j} = softmax(o_i W_r h_{rj}) \tag{9}$$

$$\hat{fr}_i = \sum_j \alpha_{i,j} h_{rj} \tag{10}$$

where i and j represent the position in a sentence, W_r is trainable weight matrix. In addition, this attention step also contextualizes the text through object information.

Furthermore, another attention module is implemented between query-objects representations h_q and response-objects representations h_r, so that the output fused query-objects representation contains weighted information of response-objects representations. It can be formulated as:

$$\alpha_{i,j} = softmax(h_{ri} W_q h_{qj}) \tag{11}$$

$$\hat{fq}_i = \sum_j \alpha_{i,j} h_{qj} \tag{12}$$

where W_q is the trainable weight matrix, i and j denote positions in a sentence.

4.3 Encoder Layer

The encoder layer aims to capture the commonsense between sentences and use it to enhance inference. It is composed of an RNN module and a dictionary module.

RNN module. An RNN unit encodes the fused queries and responses by long dependency memory cells [19], so that relationships between sentences can be captured. The input is fused query ($\hat{f}_q$) and response ($\hat{f}_r$) features. To encode the relationship between sentences, we concatenate $\hat{f}_q$ and $\hat{f}_r$ at sentence length dimensions as the input of LSTM. Its last output hidden layer contains rich information about commonsense between sentences. At time step t, the outputting representations can be formulated as:

$$h_t = o_t \odot tanh(c_t) \tag{13}$$

where the c_t is formulated the same as in Equation (5). The difference is that $x_t = concat(\hat{f}_q, \hat{f}_q)$, where concat is the concatenate operation. In addition, the outputting representations h_t is the last hidden layer of LSTM, while the outputting in Equation (3) is every time step of BiLSTM.

Dictionary module. Despite effective learning of the RNN unit in modeling the relationship between sentences, it is still limited for run-time inference. We therefore propose a dictionary unit to learn dictionary D, and then use it to look up commonsense for inference. The dictionary is a dynamic knowledge base and is being updated during training. We denote the dictionary as a $d \times k$ matrix $D\{d_1, d_2, ..., d_k\}$, where k is the size of dictionary. The given encoded representation h from RNN module will be encoded using the formulations:

$$\hat{h} = \sum_{k=1}^{K} \alpha_k d_k, \alpha = softmax(D^T h) \tag{14}$$

where α can be viewed as the "key" idea in memory network [13].

4.4 Prediction Layer

The prediction layer generates a probability distribution of responses from the high-dimension context generated in the encoder layer. It consists of a multi-layer perceptron. VCR is a multi-classification task in which one of the four responses is correct. Therefore, multiclass cross-entropy [20] is applied to complete the prediction.

5 Experimental Results

In this section, we conduct extensive experiments to demonstrate the effectiveness of our proposed DMVCR network for solving VCR tasks. We first introduce the datasets, baseline models, and evaluation metrics of our experiments. Then we compare our model with baseline models, and present an analysis of the impact of the different strategies. Finally, we present an intuitive interpretation of the prediction.

5.1 Experimental Settings

Dataset. The VCR dataset [4] is composed of 290k multiple-choice questions in total (including 290k correct answers, 290k correct rationales, and 110k images). The correct answers and rationales labeled in the dataset are met with 90% of human agreements. An adversarial matching approach is adopted to obtain counterfactual choices with minimal bias. Each answer contains 7.5 words on average, and each rationale contains 16 words on average. Each set consists of an image, a question, four available answer choices, and four candidate rationales. The correct answer and rationale are given in the dataset.

The distribution of inference types is shown in Figure 2. Thirty-eight percent of the inference types are regarding explanation, and 24% of them are about the activity. The rest are related to temporal, mental, role, scene, and hypothetical inference problems.

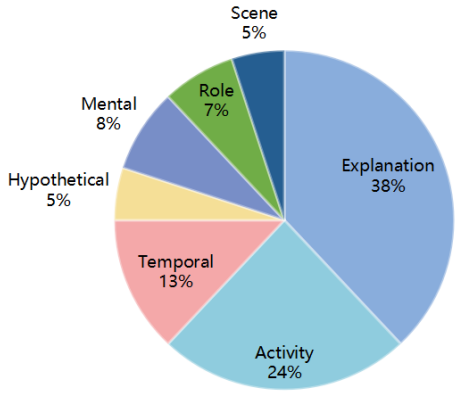


Fig. 2. Overview of the types of inference required by questions in VCR.

Hyperparameters. The image features are projected to 512 dimension. The word embedding dimension is 768. The dictionary is a [512,800] matrix, where 512 is the embedding dimension, and 800 is the dictionary size. We separately set the learning rate for the memory cell (the dictionary cell) to 0.02, and others to 0.0002. In addition, for the $Q \rightarrow A$ subtask, we set the hidden size of LSTM encoder to 512. For the $QA \rightarrow R$ subtask, we set the hidden size of LSTM encoder to 64. The model was trained with the Adam algorithm [21] using PyTorch on NVIDIA GPU GTX 1080.

Metric. The VCR task can be regarded as a multi-classification problem. We use mAp [22] to evaluate the performance, which is a common metric for evaluating prediction accuracy in multi-classification areas.

Approach for comparison. We compare the proposed DMVCR with recent deep learning-based models for VCR. Specifically, the following baseline approaches are evaluated:

- **RevisitedVQA [23]:** Different from the recently proposed systems, which have a reasoning module that includes an attention mechanism or memory mechanism, RevisitedVQA focuses on developing a "simple" alternative model, which reasons the response using logistic regressions and multi-layer perceptrons (MLP).
- **BottomUpTopDown [24]:** Proposed a bottom-up and top-down attention method to determine the feature weightings for prediction. It computes a weighted sum over image locations to fuse image and language information so that the model can predict the answer based on a given scene and question.
- **MLB [25]:** Proposed a low-rank bilinear pooling for the task. The bilinear pooling is realized by using the Hadamard product for attention mechanism and has two linear mappings without biases for embedding input vectors.

– **MUTAN [26]:** Proposed a multimodal fusion module tucker decomposition (a 3-way tensor), to fuse image and language information. In addition, multimodal low-rank bilinear (MLB) is used to reason the response for the input.
– **R2C [4]:** Proposed a fusion module, a contextualization module, and a reasoning module for VCR. It is based on the sequence relationship model LSTM and attention mechanism.

5.2 Analysis of Experimental Results

Task description. We implement the experiments separately in three steps. We firstly conducted $Q \rightarrow A$ evaluation, and then $QA \rightarrow R$. Finally, we join the $Q \rightarrow A$ result and $QA \rightarrow R$ results to obtain the final $Q \rightarrow AR$ prediction result. The difference between the implementation of $Q \rightarrow A$ and $QA \rightarrow R$ tasks is the input query and response. For the $Q \rightarrow A$ task, the query is the paired question, image, four candidate answers; while the response is the correct answer. For the $QA \rightarrow R$ task, the query is the paired question, image, correct answer, and four candidate rationales; while the response is the correct rationale.

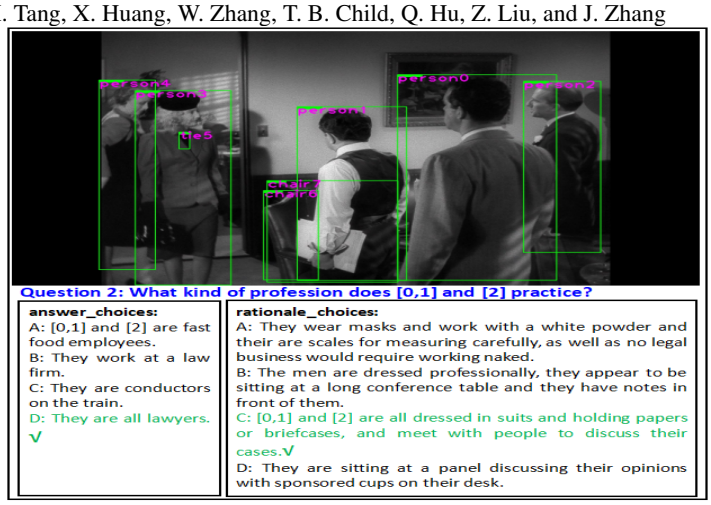
Models	$Q \rightarrow A$	$QA \rightarrow R$	$Q \rightarrow AR$
RevisitedVQA [23]	39.4	34.0	13.5
BottomUpTopDown [24]	42.8	25.1	10.7
MLB [25]	45.5	36.1	17
MUTAN [26]	44.4	32.0	14.6
R2C (Baseline) [4]	**61.9**	**62.8**	**39.1**
DMVCR	**62.4** (+0.8%)	**67.5** (+7.5%)	**42.3** (+8.2%)

Table 1. Comparison of results between our methods and other popular methods using the VCR Dataset. The best performance of the compared methods is highlighted. Percentage in parenthesis is our relative improvement over the performance of the best baseline method.

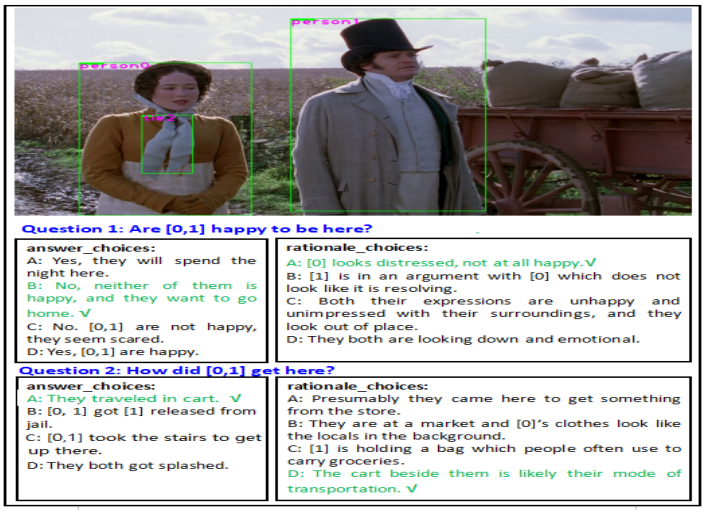
Analysis. We evaluated our method on the VCR dataset and compared the performance with other popular models. As the results in Table 1 show, our approach outperforms in all of the subtasks: $Q \rightarrow A$, $QA \rightarrow R$, and $Q \rightarrow AR$. Specifically, our method outperforms MUTAN and MLB by a large margin. Furthermore, it also performs better than R2C.

5.3 Qualitative Results

We evaluate qualitative results on the DMVCR model. The qualitative examples are provided in Figure 3. The candidate in green represents the correct choice; the candidate with a checkmark ✓ represents the prediction result by our proposed DMVCR model. As the qualitative results show, the DMVCR model improves its power in inference.



(a) Qualitative example 1. The model predicts the correct answer and rationale.



(b) Qualitative example 2. The model predicts the correct answer and rationale.

(c) Qualitative example 3. The model predicts the correct answer but incorrect rationale in Question 1. The model predicts an incorrect answer but correct rationale in Question 2.

Fig. 3. Qualitative examples. Prediction from DMVCR is marked with a ✓ while correct results are highlighted in green.

For instance, see Figure 3(a). The question listed is: "What kind of profession does [0,1] and [2] practice?". The predicted answer is D - "They are all lawyers." Furthermore, the model offers rationale C - "[0,1] and [2] are all dressed in suits and holding papers or briefcases, and meet with people to discuss their cases." DMVCR correctly infers the rationale based on dress and activity, even though this task is difficult for humans.

DMVCR can also identify human beings' expressions and infer emotion. See for example the result in Figure 3(b). Question 1 is: "Are [0,1] happy to be there?". Our model selects the correct answer B along with reason A: "No, neither of them is happy, and they want to go home"; because "[0] looks distressed, not at all happy."

Finally, there are also results which predict the correct answer but infer the wrong reason. For instance, see question 1 in Figure 3(c): "What are [3,1] doing?" DMVCR predicts the correct answer A - "They are preparing to run from the fire." But it infers the wrong reason A - "They are turned towards the direction of the fire." The correct answer is of course B - "They are in motion, and it would be logistical to try to leave." It is also possible for the model to predict a wrong answer but correct rationale. This appears in question 2 of Figure 3(c). The model predicts the wrong answer D - "[0] is afraid that he will be seen." The correct reason is B - "The building is on fire and he is vulnerable to it."

6 Conclusion

This paper has studied the popular visual commonsense reasoning (VCR). We propose a working memory based model composed of a feature representation layer to capture multiple features containing language and objects information; a multimodal fusion layer to fuse features from language and images; an encoder layer to encode commonsense between sentences and enhance inference ability using dynamic knowledge from a dictionary unit; and a prediction layer to predict a correct response from four choices. We conduct extensive experiments on the VCR dataset to demonstrate the effectiveness of our model and present intuitive interpretation. In the future, it would be interesting to investigate multimodal feature fusion methods as well as encoding commonsense using an attention mechanism to improve the performance of VCR.

References

1. L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," *natural language engineering*, vol. 7, no. 4, p. 275, 2001.
2. P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.
3. S. Antol, A. Agrawal *et al.*, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.
4. R. Zellers, Y. Bisk *et al.*, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE Conference on CVPR*, 2019.

5. H. Ben-younes, R. Cadène, M. Cord, and N. Thome, "MUTAN: multimodal tucker fusion for visual question answering," *CoRR*, 2017.
6. J. Lin, U. Jain *et al.*, "TAB-VCR: tags and attributes based VCR baselines."
7. F. Yu, J. Tang *et al.*, "Ernie-vil: Knowledge enhanced vision-language representations through scene graph," 2020.
8. J. Lu, D. Batra *et al.*, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, 2019, pp. 13–23.
9. D. B. Lee, S. Lee, W. T. Jeong, D. Kim, and S. J. Hwang, "Generating diverse and consistent qa pairs from contexts with information-maximizing hierarchical conditional vaes," *arXiv preprint arXiv:2005.13837*, 2020.
10. Y.-C. Chen, L. Li *et al.*, "Uniter: Learning universal image-text representations," 2019.
11. P. N. Sabes and M. I. Jordan, "Advances in neural information processing systems," in *In G. Tesauro & D. Touretzky & T. Leed (Eds.), Advances in Neural Information Processing Systems*. Citeseer, 1995.
12. C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *International conference on machine learning*. PMLR, 2016, pp. 2397–2406.
13. X. Yang, K. Tang *et al.*, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE Conference on CVPR*, 2019, pp. 10 685–10 694.
14. A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *International conference on machine learning*. PMLR, 2016, pp. 1378–1387.
15. Z. Huang, Xu *et al.*, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.
16. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on CVPR*, 2016, pp. 770–778.
17. J. Devlin, M.-W. Chang *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
18. X. Yin, J. Goudriaan, E. A. Lantinga, J. Vos, and H. J. Spiertz, "A flexible sigmoid function of determinate growth," *Annals of botany*, vol. 91, no. 3, pp. 361–371, 2003.
19. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
20. R. Rubinstein, "The cross-entropy method for combinatorial and continuous optimization," *Methodology and computing in applied probability*, vol. 1, no. 2, pp. 127–190, 1999.
21. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
22. P. Henderson and V. Ferrari, "End-to-end training of object class detectors for mean average precision," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 198–213.
23. A. Jabri, A. Joulin, and L. Van Der Maaten, "Revisiting visual question answering baselines," in *European conference on computer vision*. Springer, 2016, pp. 727–739.
24. P. Anderson, X. He *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.
25. J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," 2016.
26. H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2612–2620.