Multi-word Answer and Rationale Generation to Visual Questions

Radhika Dua
KAIST
radhikadua@kaist.ac.kr

Abstract

Visual Question Answering is a multi-modal task that aims to measure high-level visual understanding. Contemporary VQA models are restrictive in the sense that answers are obtained via classification over a limited vocabulary (in the case of open-ended VQA), or via classification over a set of multiple-choice-type answers. In this work, we present a completely generative formulation where a multi-word answer is *generated* for a visual query. To take this a step forward, we introduce a new task: ViQAR (Visual Question Answering and Reasoning), wherein a model must generate the complete answer and a rationale that seeks to justify the generated answer. We propose an end-to-end architecture to solve this task and describe how to evaluate it. We show that our model generates strong answers and rationales through qualitative and quantitative evaluation, as well as through a human Turing Test.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015) is a vision-language task that has seen a lot of attention in recent years. In general, the VQA task consists of either open-ended or multiple choice answers to a question about the image. There are an increasing number of models that obtain the best possible performance on benchmark VQA datasets, which intend to measure visual understanding based on visual questions. However, answers in existing VQA datasets and models are largely one-word answers (average length is 1.1 words), which gives existing models the freedom to treat answer generation as a classification task. For the open-ended VQA task, the top-K answers are chosen, and models perform classification over this vocabulary.

However, many questions which require commonsense reasoning cannot be answered in a single word. A textual answer for a sufficiently complicated question may need to be a sentence. For example, a question of the type "What will happen...." usually cannot be answered completely using a single word. Figure 1 shows examples of such questions where multi-word answers are required (the answers and rationales in this figure are generated by our model in this work). Current VQA systems are not well-suited for questions of this type. To reduce this gap, more recently, the Visual Commonsense Reasoning (VCR) task (Zellers et al., 2018; Lu et al., 2019; Dua et al., 2019; Zheng et al., 2019; Talmor et al., 2018) was proposed, which requires a greater level of visual understanding and an ability to reason about the world. More interestingly, the VCR dataset features multi-word answers, with an average answer length of 7.55 words. However, VCR is still a classification task, where the correct answer is chosen from a set of four answers. Models which solve classification tasks simply need to pick an answer in the case of VQA, or an answer and a rationale for VCR. However, when multi-word answers are required for a visual question, options are not sufficient, since the same 'correct' answer can be paraphrased in a multitude of ways, each having the same semantic meaning but differing in grammar. Figure 2 shows an image from the VCR dataset, where the first highlighted answer is the correct one among a set of four options provided in the dataset. The remaining three answers in the figure are included by us here (not in the dataset) as other plausible correct answers. Existing VQA models are fundamentally limited by picking a right option, rather to answer in a more natural manner. Moreover, since the number of possible 'correct' options in multi-word answer settings can be large (as evidenced by Figure 2), we propose that for richer answers, one would need to move away from the traditional classification setting, and instead let our model *generate the answer* to a given question.

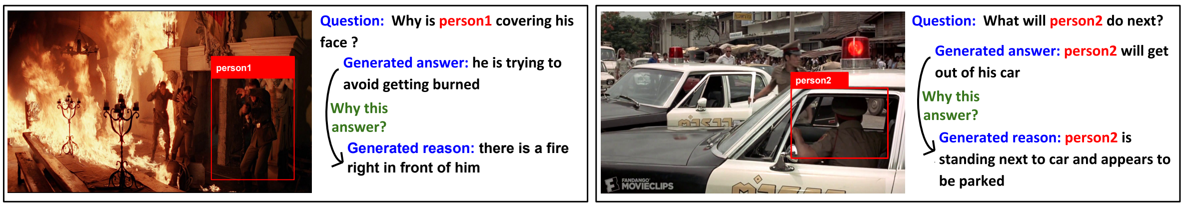

Figure 1: Given an image and a question about the image, we **generate** a natural language answer and reason that explains why the answer was generated. The images shown above are examples of outputs that our proposed model generates. These examples also illustrate the kind of visual questions for which a single-word answer is insufficient. Contemporary VQA models handle even such kinds of questions only in a classification setting, which is limiting.

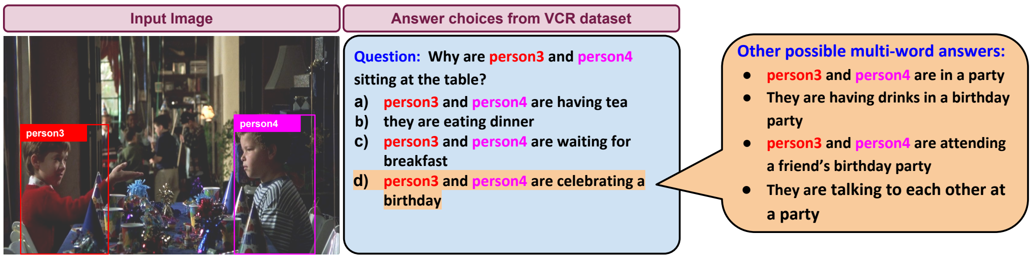

Figure 2: An example from the VCR dataset shows that there can be many correct multi-word answers to a question that makes classification setting restrictive. The highlighted option is the correct option present in the VCR dataset, the rest are plausible correct answers.

We hence propose a new task which takes a generative approach to multi-word VQA in this work.

Humans when answering questions often use a rationale to justify the answer. In certain cases, humans answer directly from memory (perhaps through associations) and then provide a post-hoc rationale, which could help improve the answer too - thus suggesting an interplay between an answer and its rationale. Following this cue, we also propose to generate a rationale along with the answer which serves two purposes: (i) it helps justify the generated answer to end-users; and (ii) it helps generate a better answer. Going beyond contemporary efforts in VQA, we hence propose, for the first time to the best of our knowledge, an approach that automatically generates both multi-word answers and an accompanying rationale, that also serves as a textual justification for the answer. We term this task `Visual Question Answering and Reasoning(ViQAR)` and propose an end-to-end methodology to address this task. This task is especially important in critical AI tasks such as VQA in the medical domain, where simply answering questions about the medical images is not sufficient.

In addition to formalizing this new task, we provide a simple yet reasonably effective model consisting of four sequentially arranged recurrent networks to address this challenge. The model can be seen as having two parts: a generation module (GM), which comprises of the first two sequential recurrent networks, and a refinement module (RM), which comprises of the final two sequential recurrent networks. The GM first generates an answer, using which it generates a rationale that explains the answer. The RM generates a *refined* answer based on the rationale generated by GM. The refined answer is further used to generate a refined rationale. Our overall model design is motivated by the way humans think about answers to questions, wherein the answer and rationale are often mutually dependent on each other. We seek to model this dependency by first generating an answer-rationale pair and then using them as priors to regenerate a refined answer and rationale. We train our model on the VCR dataset, which contains open-ended visual questions along with answers and rationales. Considering this is a generative task, we evaluate our methodology by comparing our generated answer/rationale with the ground truth answer/rationale on correctness and goodness of the generated content using generative language metrics, as well as by human Turing Tests.

Our main contributions in this work can be summarized as follows: (i) We propose a new task

ViQAR that seeks to open up a new dimension of Visual Question Answering tasks, by moving to a completely generative paradigm; (ii) We propose a simple and effective model based on generation and iterative refinement for ViQAR (which could serve as a baseline to the community); (iii) Considering generative models in general can be difficult to evaluate, we provide a discussion on how to evaluate such models, as well as study a comprehensive list of evaluation metrics for this task; (iv) We conduct a suite of experiments which show promise of the proposed model for this task, and also perform ablation studies of various choices and components to study the effectiveness of the proposed methodology on ViQAR. We believe that this work could lead to further efforts on commonsense answer and rationale generation in vision tasks in the near future. To the best of our knowledge, this is the first such effort of automatically generating a multi-word answer and rationale to a visual question.

2 Related Work

VQA and Image Captioning. A lot of work in VQA is based on attention-based models that aim to 'look' at the relevant regions of the image in order to answer the question (Anderson et al., 2018; Lu et al., 2016; Yu et al., 2017a; Yi et al., 2018). Other recent work has focused on better multimodal fusion methods (Kim et al., 2018, 2017; Fukui et al., 2016; Yu et al., 2017b), the incorporation of relations (Norcliffe-Brown et al., 2018; Li et al., 2019; Santoro et al., 2017), the use of multi-step reasoning (Cadene et al., 2019), and neural module networks for compositional reasoning (Johnson et al., 2017b; Chen et al., 2021; Hu et al., 2017). Visual Dialog (Das et al., 2017; Zheng et al., 2019) extends VQA but requires an agent to hold a meaningful conversation with humans in natural language based on visual questions. Image captioning (Xu et al., 2015; You et al., 2016; Lu et al., 2017; Anderson et al., 2018; Rennie et al., 2016) is a global description of an image and hence different from ViQAR which is concerned with answering a question about understanding of a local region in the image.

The efforts closest to ours are those that provide justifications along with answers (Li et al., 2018b; Hendricks et al., 2016; Li et al., 2018a; Park et al., 2018; Wu et al., 2019; Park et al., 2018), each of which however also answers a question as a classification task (and not in a generative manner) as described below. Li et al. (Li et al., 2018b) created the VQA-E dataset that has an explanation along with the answer to the question. Wu et al. (Wu et al., 2019) provide relevant captions to aid in solving VQA, which can be thought of as weak justifications. More recent efforts (Park et al., 2018; Patro et al., 2020) attempt to provide visual and textual explanations to justify the predicted answers. Datasets have also been proposed for VQA in the recent past to test visual understanding (Zhu et al., 2016; Goyal et al., 2017; Johnson et al., 2017a); for e.g., the Visual7W dataset (Zhu et al., 2016) contains a richer class of questions about an image with textual and visual answers. However, all these aforementioned efforts continue to focus on answering a question as a classification task (often in one word, such as Yes/No), followed by simple explanations. We however, in this work, focus on *generating* multi-word answers with a corresponding multi-word rationale, which has not been done before.

Visual Commonsense Reasoning (VCR). VCR (Zellers et al., 2018) is a vision-language dataset, which involves choosing a correct answer (among four provided options) for a given question about the image, and then choosing a rationale to justify the answer. The task associated with the dataset aims to test for visual commonsense understanding and provides images, questions and answers of a higher complexity than other datasets such as CLEVR (Johnson et al., 2017a). The dataset has attracted various methods (Zellers et al., 2018; Lu et al., 2019; Dua et al., 2019; Zheng et al., 2019; Talmor et al., 2018; Lin et al., 2019; Brad, 2019), each of which however follow the dataset's task and treat this as a classification problem. None of these efforts attempt to answer and reason using generated sentences.

In contrast to all the aforementioned efforts, our work, ViQAR, focuses on automatic complete *generation* of the answer, and of a rationale, given a visual query. This is a challenging task, since the generated answers must be correct (with respect to the question asked), be complete, be natural, and also be justified with a well-formed rationale.

3 ViQAR: Task description

Let $\mathcal{V}$ be a given vocabulary of size $|\mathcal{V}|$ and $\mathbf{A} = (a_1, \dots, a_{l_a}) \in \mathcal{V}^{l_a}$, $\mathbf{R} = (r_1, \dots, r_{l_r}) \in \mathcal{V}^{l_r}$ represent answer sequences of length l_a and rationale sequences of length l_r respectively. Let $\mathbf{I} \in \mathbb{R}^D$ represent the image representation, and $\mathbf{Q} \in \mathbb{R}^B$ be the feature representation of a given question. We also allow the use of an image caption, if available, in this framework given by a feature representation $\mathbf{C} \in \mathbb{R}^B$. Our task is to compute a function $\mathcal{F} : \mathbb{R}^D \times \mathbb{R}^B \times \mathbb{R}^B \rightarrow \mathcal{V}^{l_a} \times \mathcal{V}^{l_r}$ that maps the input image, question and caption features to a large space of generated answers $\mathbf{A}$ and rationales $\mathbf{R}$, as given below:

$$\mathcal{F}(\mathbf{I}, \mathbf{Q}, \mathbf{C}) = (\mathbf{A}, \mathbf{R}) \tag{1}$$

Note that the formalization of this task is different from other tasks in this domain, such as Visual Question Answering (Antol et al., 2015) and Visual Commonsense Reasoning (Zellers et al., 2018). The VQA task can be formulated as learning a function $\mathcal{G} : \mathbb{R}^D \times \mathbb{R}^B \rightarrow C$, where C is a discrete, finite set of choices (classification setting). Similarly, the Visual Commonsense Reasoning task provided in (Zellers et al., 2018) aims to learn a function $\mathcal{H} : \mathbb{R}^D \times \mathbb{R}^B \rightarrow C_1 \times C_2$, where C_1 is the set of possible answers, and C_2 is the set of possible reasons. The generative task, proposed here in `ViQAR`, is harder to solve when compared to VQA and VCR. One can divide `ViQAR` into two sub-tasks:

- **Answer Generation.** Given an image, its caption, and a complex question about the image, a multi-word natural language answer is generated:
  $(\mathbf{I}, \mathbf{Q}, \mathbf{C}) \rightarrow \mathbf{A}$
- **Rationale Generation.** Given an image, its caption, a complex question about the image, and an answer to the question, a rationale to justify the answer is generated: $(\mathbf{I}, \mathbf{Q}, \mathbf{C}, \mathbf{A}) \rightarrow \mathbf{R}$

We also study variants of the above sub-tasks (such as when captions are not available) in our experiments. Our experiments suggest that the availability of captions helps a model achieve better performance on our task. We now present a methodology built using known basic components to study and show that the proposed, seemingly challenging, new task can be solved with existing architectures. In particular, our methodology is based on the understanding that the answer and rationale can help each other, and hence needs an iterative refinement procedure to handle such a multi-word multi-output task. We consider the simplicity of the proposed solution as an aspect of our solution by design, more than a limitation, and hope that the proposed architecture will serve as a baseline for future efforts on this task.

4 Proposed methodology

We present an end-to-end, attention-based, encoder-decoder architecture for answer and rationale generation which is based on an iterative refinement procedure. The refinement in our architecture is motivated by the observation that answers and rationales can influence one another mutually. Thus, knowing the answer helps in generation of a rationale, which in turn can help in the generation of a more refined answer. The encoder part of the architecture generates the features from the image, question and caption. These features are used by the decoder to generate the answer and rationale for a question.

Feature Extraction. We use spatial image features as proposed in (Anderson et al., 2018), which are termed bottom-up image features. We consider a fixed number of regions for each image, and extract a set of k features, V, as defined below:

$$V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\} \quad \text{where} \quad \mathbf{v}_i \in \mathbb{R}^D. \tag{2}$$

We use BERT (Devlin et al., 2019) representations to obtain fixed-size embeddings for the question and caption, $\mathbf{Q} \in \mathbb{R}^B$ and $\mathbf{C} \in \mathbb{R}^B$ respectively. The question and caption are projected into a common feature space $\mathbf{T} \in \mathbb{R}^L$ given by:

$$\mathbf{T} = g(W_t^T(\tanh(W_q^T\mathbf{Q}) \oplus \tanh(W_c^T\mathbf{C}))), \tag{3}$$

where g is a non-linear function, $\oplus$ indicates concatenation and $W_t \in \mathbb{R}^{L \times L}$, $W_q \in \mathbb{R}^{B \times L}$ and $W_c \in \mathbb{R}^{B \times L}$ are learnable weight matrices of the layers (we use two linear layers in our implementation in this work).

Let the mean of the extracted spatial image features (as in Equation 2) be denoted by $\bar{\mathbf{V}} \in \mathbb{R}^D$. These are concatenated with the projected question and caption features to obtain $\mathbf{F}$, which is the common input feature vector to all the LSTMs in our architecture:

$$\mathbf{F} = \bar{\mathbf{V}} \oplus \mathbf{T} \tag{4}$$

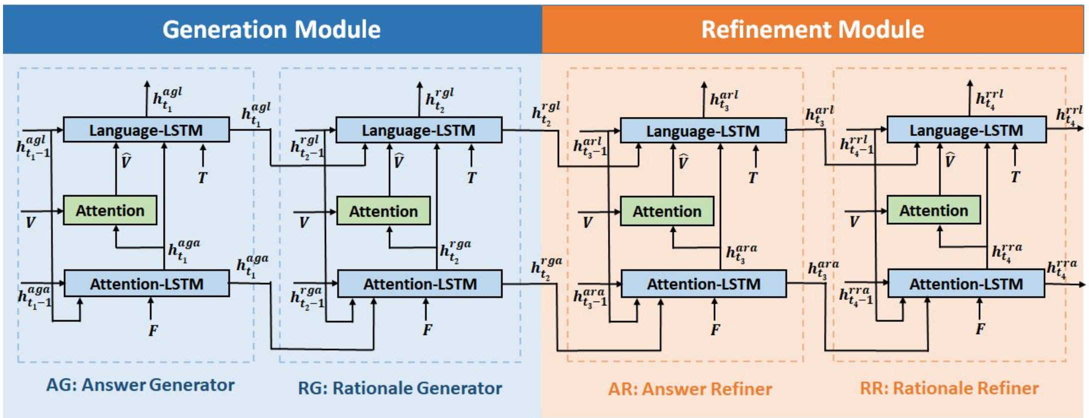

Figure 3: The decoder of our proposed architecture: For simplicity we only show the last time-step of each unrolled LSTM. Here $t_1 = l_a, t_2 = l_a + l_r, t_3 = 2l_a + l_r$ and $t_4 = 2l_a + 2l_r$. Given an image and a question on the image, the model must generate an answer to the question and a rationale to justify why the answer is correct.

Architecture. Figure 3 shows our end-to-end architecture to address `ViQAR`. As stated earlier, our architecture has two modules: *generation* (GM) and *refinement* (RM). The GM consists of two sequential, stacked LSTMs, henceforth referred to as answer generator (AG) and rationale generator (RG) respectively. The RM seeks to refine the generated answer as well as rationale, and is an important part of the proposed solution as seen in our experimental results. It also consists of two sequential, stacked LSTMs, which we denote as answer refiner (AR) and rationale refiner (RR).

Each sub-module (presented inside dashed lines in the figure) is a complete LSTM. Given an image, question, and caption, the AG sub-module unrolls for l_a time steps to generate an answer. The hidden state of Language and Attention LSTMs after l_a time steps is a representation of the generated answer. Using the representation of the generated answer from AG, RG sub-module unrolls for l_r time steps to generate a rationale and obtain its representation. Then the AR sub-module uses the features from RG to generate a refined answer. Lastly, the RR sub-module uses the answer features from AR to generate a refined rationale. Thus, a refined answer is generated after $l_a + l_r$ time steps and a refined rationale is generated after l_a further time steps. The complete architecture runs in $2l_a + 2l_r$ time steps.

The LSTMs. The two layers of each stacked LSTM (Hochreiter and Schmidhuber, 1997) are referred to as the Attention-LSTM ($\mathcal{L}_a$) and Language-LSTM ($\mathcal{L}_l$) respectively. We denote h_t^a and x_t^a as the hidden state and input of the Attention-LSTM at time step t respectively. Analogously, h_t^l and x_t^l denote the hidden state and input of the Language-LSTM at time t. Since the four LSTMs are identical in operation, we describe the attention and sequence generation modules of one of the sequential LSTMs below in detail.

Spatial visual attention. We use a soft, spatial-attention model, similar to (Anderson et al., 2018) and (Lu et al., 2017), to compute attended image features $\hat{\mathbf{V}}$. Given the combined input features $\mathbf{F}$ and previous hidden states h_{t-1}^a, h_{t-1}^l, the current hidden state of the Attention-LSTM is given by:

$$\begin{aligned} x_t^a &\equiv h^p \oplus h_{t-1}^l \oplus \mathbf{F} \oplus \pi_t, \\ h_t^a &= \mathcal{L}_a(x_t^a, h_{t-1}^a), \end{aligned} \tag{5}$$

where $\pi_t = W_e^T \mathbf{1}_t$ is the embedding of the input word, $W_e \in \mathbb{R}^{|\mathcal{V}| \times E}$ is the weight of the embedding layer, and $\mathbf{1}_t$ is the one-hot representation of the input at time t. h^p is the hidden representation of the previous LSTM (answer or rationale, depending on the current LSTM).

The hidden state h_t^a and visual features V are used by the attention module (implemented as a two-layered MLP in this work) to compute the normalized set of attention weights $\boldsymbol{\alpha}_t = \{\alpha_{1t}, \ldots, \alpha_{kt}\}$ (where α_{it} is the normalized weight of image feature $\mathbf{v}_i$) as below:

$$\begin{aligned} y_{i,t} &= W_{ay}^T(\tanh(W_{av}^T \mathbf{v}_i + W_{ah}^T h_t^a)), \\ \boldsymbol{\alpha}_t &= \text{softmax}(y_{1t} \ldots, y_{kt}). \end{aligned} \tag{6}$$

In the above equations, $W_{ay} \in \mathbb{R}^{A \times 1}$, $W_{av} \in \mathbb{R}^{D \times A}$ and $W_{ah} \in \mathbb{R}^{H \times A}$ are weights learned by the attention MLP, H is the hidden size of the LSTM and A is the hidden size of the attention MLP. The attended image feature vector $\hat{\mathbf{V}}_t = \sum_{i=1}^{k} \boldsymbol{\alpha}_{it}\mathbf{v}_i$ is the weighted sum of all visual features.

Sequence generation. The attended image features $\hat{\mathbf{V}}_t$, together with $\mathbf{T}$ and h_t^a, are inputs to the language-LSTM at time t. We then have:

$$\begin{aligned} x_t^l &\equiv h^p \oplus \hat{\mathbf{V}}_t \oplus h_t^a \oplus \mathbf{T} \\ h_t^l &= \mathcal{L}_l(x_t^l, h_{t-1}^l) \\ y_t &= W_{lh}^T h_t^l + b_{lh} \\ p_t &= \mathrm{softmax}(y_t) \end{aligned} \tag{7}$$

where h^p is the hidden state of the previous LSTM, h_t^l is the output of the Language-LSTM, p_t is the conditional probability over words in $\mathcal{V}$ at time t. The word at time step t is generated by a single-layered MLP with learnable parameters: $W_{lh} \in \mathbb{R}^{H\times|\mathcal{V}|}$, $b_{lh} \in \mathbb{R}^{|\mathcal{V}|\times 1}$. The attention MLP parameters W_{ay}, W_{av} and W_{ah}, and embedding layer's parameters W_e are shared across all four LSTMs.

Loss Function. For a better understanding of our approach, Figure 4 presents a high-level illustration of our proposed generation-refinement model.

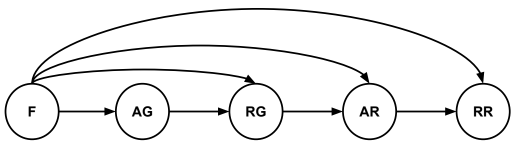

Figure 4: High-level illustration of our proposed Generation-Refinement model

Let $A_1 = (a_{11}, a_{12}, ..., a_{1l_a})$, $R_1 = (r_{11}, r_{12}, ..., r_{1l_r})$, $A_2 = (a_{21}, a_{22}, ..., a_{2l_a})$ and $R_2 = (r_{21}, r_{22}, ..., r_{2l_r})$ be the generated answer, generated rationale, refined answer and refined rationale sequences respectively, where a_{ij} and r_{ij} are discrete random variables taking values from the common vocabulary $\mathcal{V}$. Given the common input F, our objective is to maximize the likelihood $P(A_1, R_1, A_2, R_2|F)$ given by:

$$P(A_1, R_1, A_2, R_2|F) \tag{8}$$

$$\begin{aligned} &= P(A_1, R1|F)P(A_2, R_2|F, A_1, R_1) \\ &= P(A_1|F)P(R_1|F, A_1) \\ &= P(A_2|F, A_1, R_1)P(R_2|F, A_1, R_1, A_2) \end{aligned} \tag{9}$$

In our model design, each term in the RHS of Eqn 9 is computed by a distinct LSTM. Hence, minimizing the sum of losses of the four LSTMs becomes equivalent to maximizing the joint likelihood. Our overall loss is the sum of four cross-entropy losses, one for each LSTM, as given below:

$$\mathcal{L} = -\left(\sum_{t=1}^{l_a} \log p_t^{\theta_1} + \sum_{t=1}^{l_r} \log p_t^{\theta_2} + \sum_{t=1}^{l_a} \log p_t^{\theta_3} + \sum_{t=1}^{l_r} \log p_t^{\theta_4}\right) \tag{10}$$

where θ_i represents the i^{th} sub-module LSTM, p_t is the conditional probability of the t^{th} word in the input sequence as calculated by the corresponding LSTM, l_a indicates the ground-truth answer length, and l_r the ground truth rationale length. Other loss formulations, such as a weighted average of the cross entropy terms did not perform better than a simple sum. We tried weights from $0.0, 0.25, 0.5, 0.75, 1.0$ for the loss terms.

5 Experiments and results

In this section, we describe the dataset used for this work, implementation details of out model, and present the results of the proposed method and its variants.

5.1 Experimental setup

Dataset. Considering there has been no dataset explicitly built for this new task, we study the performance of the proposed method on the recently introduced VCR (Zellers et al., 2018) dataset, which has all components needed for our approach. We train our proposed architecture on VCR, which contains ground truth answers and ground truth rationales against which we compare our generated answers and rationales.

Table 4: Statistical comparison of VCR with VQA-E, and VQA-X datasets. VCR dataset is highly complex as it is made up of complex subjective questions.

Dataset	**Avg. A length**	**Avg. Q length**	**Avg. R length**	**Complexity**
VCR	7.55	6.61	16.2	High
VQA-E	1.11	6.1	11.1	Low
VQA-X	1.12	6.13	8.56	Low

VQA-E (Li et al., 2018b) and VQA-X (Park et al., 2018) are competing datasets that contains explanations along with question-answer pairs. Table 4 shows the high-level analysis of the three datasets. Since VQA-E and VQA-X are derived

Table 1: Quantitative evaluation on VCR dataset; we compare against a basic two-stage LSTM model and a VQA model as baselines; remaining columns are proposed model variants.[CS = cosine similarity]

Metrics	VQA-Baseline	Baseline	Q+I+C (Ours)	Q+I (Ours)	Q+C (Ours)
Univ Sent Encoder CS	0.419	0.410	**0.455**	0.454	0.440
Infersent CS	0.370	0.400	0.438	**0.442**	0.426
Embedding Avg CS	0.838	0.840	0.846	**0.853**	0.845
Vector Extrema CS	0.474	0.444	**0.493**	0.483	0.475
Greedy Matching Score	0.662	0.633	**0.672**	0.661	0.657
METEOR	0.107	0.095	**0.116**	0.104	0.103
Skipthought CS	0.430	0.359	**0.436**	0.387	0.385
RougeL	0.259	0.206	**0.262**	0.232	0.236
CIDEr	0.364	0.158	**0.455**	0.310	0.298
F-BERTScore	0.877	0.860	**0.879**	0.867	0.868

Table 2: Comparison of proposed Generation-Refinement architecture with variations in number of refinement modules. [CS: cosine similarity]

Metrics	#Refine Modules		
	0	1	2
Univ Sent Encoder CS	0.453	**0.455**	0.430
Infersent CS	0.434	**0.438**	0.421
Embedding Avg CS	0.850	0.846	0.840
Vector Extrema CS	0.482	**0.493**	0.462
Greedy Matching Score	0.659	**0.672**	0.639
METEOR	0.101	**0.116**	0.090
Skipthought CS	0.384	**0.436**	0.375
RougeL	0.234	**0.262**	0.198
CIDEr	0.314	**0.455**	0.197
F-BertScore	0.868	**0.879**	0.861

Table 3: Results of the Turing test on VCR and Visual7W dataset performed with 30 people who had to rate samples consisting of a question and its corresponding answer and rationales on five criteria. For each criterion, a rating of 1 to 5 was given. The table gives the mean score and standard deviation for each criterion for the generated and ground truth samples.

Criteria	VCR		Visual7W	
	Generated	Ground-truth	Generated	Ground-truth
How well-formed and grammatically correct is the answer?	4.15±1.05	4.40±0.87	3.98±1.08	–
How well-formed and grammatically correct is the rationale?	3.53±1.26	4.26±0.92	3.80±1.04	–
How relevant is the answer to the image-question pair?	3.60±1.32	4.08±1.03	4.11±1.17	–
How well does the rationale explain the answer with respect to the image-question pair?	3.04±1.36	4.05±1.10	3.83±1.23	–
Irrespective of the image-question pair, how well does the rationale explain the answer ?	3.46±1.35	4.13±1.09	3.83±1.28	–

from VQA-2 (Goyal et al., 2017), many of the questions can be answered in one word (a yes/no answer or a number). In contrast, VCR asks open-ended questions and has longer answers. Since our task aims to generate rich answers, the VCR dataset provides a richer context for this work. CLEVR (Johnson et al., 2017a) is another VQA dataset that measures the logical reasoning capabilities by asking the question that can be answered when a certain sequential reasoning is followed. This dataset however does not contain reasons/rationales on which we can train. Also, we do not perform a direct evaluation on CLEVR because our model is trained on real-world natural images while CLEVR is a synthetic shapes dataset. In order to study our method further, we also study the transfer of our learned model to another challenging dataset, Visual7W (Zhu et al., 2016), by generating an answer/rationale pair for visual questions in Visual7W (please see Section 5.3 more more details).

Implementation Details. We use spatial image features generated from (Anderson et al., 2018) as our image input. Fixed-size BERT representations of questions and captions are used. Hidden size of all LSTMs is set to 1024 and hidden size of the attention MLP is set to 512. We trained using the ADAM optimizer with a decaying learning rate starting from $4e^{-4}$, using a batch size of 64. Dropout is used as a regularizer.

Evaluation metrics. We use multiple objective evaluation metrics to evaluate the goodness of answers and rationales generated by our model. Since our task is generative, evaluation is done by comparing our generated sentences with ground-truth sentences to assess their semantic correctness as well as structural soundness. To this end, we use a combination of multiple existing evaluation metrics. Word overlap-based metrics such as METEOR (Lavie and Agarwal, 2007), CIDEr (Vedantam et al., 2015) and ROUGE (Lin, 2004) quantify the structural closeness of the generated sentences to the ground-truth. While such metrics give a sense of the structural correctness of the generated sentences, they may be insufficient for evaluating generation tasks, since there could be many valid generations which are correct, but not share the same words as a single ground truth answer. Hence, in order to measure how close the generation is to the ground-truth in meaning, we additionally use embedding-based metrics (which calculate the cosine similarity between sentence embeddings for generated and ground-truth sentences) including SkipThought cosine similarity (Kiros et al., 2015), Vector Extrema cosine similarity (Forgues and Pineau, 2014), Universal sentence encoder (Cer et al., 2018), Infersent (Conneau et al., 2017) and BERTScore (Zhang* et al., 2020). We use a comprehensive suite of all the aforementioned metrics to study the performance of our model. We further provide details of performance of our model on the VCR classification task in the supplementary.

5.2 Performance evaluation of `ViQAR`

Quantitative results. Quantitative results on the suite of evaluation metrics stated earlier are shown in Table 1. Since this is a new task, there are no known methods to compare against. We compare our model against a baseline (called *Baseline* in Table 1) composed of two separate two-stage LSTMs, one for answer and one for the rationale, and a VQA-based method (Anderson et al., 2018) that extracts multi-modal features to generate answers and rationales parallelly in an end-to-end manner (called *VQA-Baseline* in Table 1). (Comparison with other standard VQA models is not relevant in this setting, since we perform a generative task, unlike existing VQA models.) We show results on three variants of our proposed Generation-Refinement model: Q+I+C (when question, image and caption are given as inputs), Q+I (question and image alone as inputs), and Q+C (question and caption alone as inputs). Evidently, our Q+I+C performed the most consistently across all the evaluation metrics. Importantly, our model outperforms both baselines, including the VQA-based one, on every single evaluation metric, showing the utility of the proposed architecture.

Qualitative results. Figure 5 (Top) shows an example where the proposed model (Q+I+C setting) generates a meaningful answer with a supporting rationale. Given the question, "What does person2 do?", the generated rationale: "person2 is wearing a school uniform" actively supports the generated answer: "person2 is a student", justifying the choice of a two-stage generation-refinement architecture.

For completeness of understanding, we present an example, Figure 5 (Bottom), on which our model fails to generate the semantically correct answer. Even on this result, we observe the generated answer and rationale are grammatically correct and complete (where rationale supports the answer). Improving the semantic correctness of the generations will be an important direction of future work. Qualitative results indicate that our model is capable of generating answer-rationale pairs to complex subjective questions starting with 'Why', 'What', 'How', etc. More qualitative results are presented in the supplementary owing to space constraints.

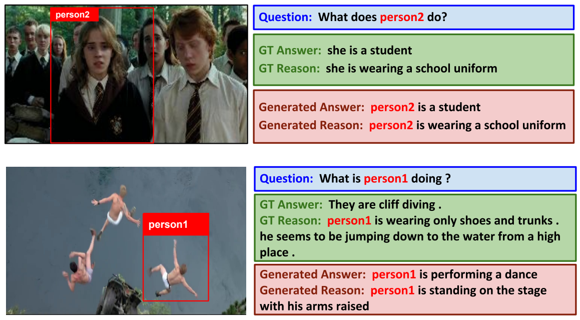

Figure 5: *(Best viewed in color)* **Top:** Example output from our proposed Generation-Refinement architecture. **Bottom:** A challenging input for which our model fails.

Human Turing test. In addition to the study of the objective evaluation metrics, we also performed a human Turing test on the generated answers and rationales. 30 human evaluators were presented each with 50 randomly sampled image-question pairs, each containing an answer to the question and its rationale. The test aims to measure how humans score the generated sentences w.r.t. ground truth sentences. Sixteen of the fifty questions had ground truth answers and rationales, while the rest were generated by our proposed model. For each sample, the evaluators had to give a rating of 1 to 5 for five different criteria, with 1 being very poor and 5 being very good. The results are presented in Table 3. Despite the higher number of generated answer-rationales judged by human users, the answers and rationales produced by our method were deemed to be fairly correct grammatically. The evaluators also agreed that our answers were relevant to the question and the generated rationales are acceptably relevant to the generated answer.

5.3 Further analysis of `ViQAR`

Ablation studies on refinement module. We evaluate the performance of variations of our proposed generation-refinement architecture M: (i)

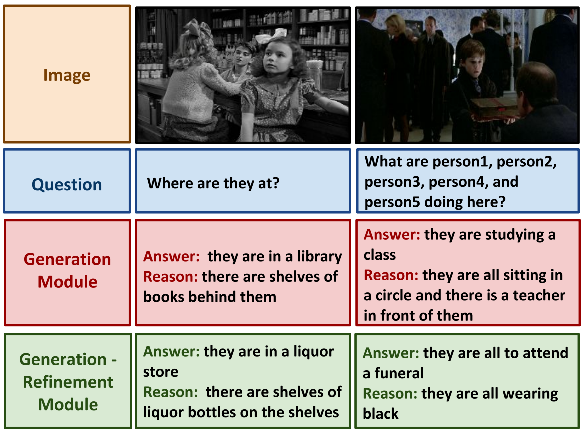

Figure 6: Qualitative results for our model with (in green, last row) and without (in red, penultimate row) Refinement module

Figure 7: Qualitative results on Visual7W dataset (note that there is no rationale provided in this dataset, and all above rationales were generated by our model)

$M - RM$: where the refinement module is removed; (ii) $M + RM$: where a second refinement module is added, i.e. the model has one generation and two refinement modules (to see if further refinement of answer and rationale helps). Table 2 shows the quantitative results. We observe that our proposed model, which has one refinement module has the best results. Adding additional refinement modules causes the performance to go down.

We hypothesize that the additional parameters (in a second Refinement module) in the model makes it harder for the network to learn. Removal of the refinement module also causes performance to drop, supporting our claim on the usefulness for a Refinement module too. Figure 6 provides a few qualitative results with and without the refinement module, supporting our claim. More results are presented in the supplementary.

Transfer to other datasets. We also studied whether the proposed model, trained on the VCR dataset, can provide answers and rationales to visual questions in other VQA datasets (which do not have ground truth rationale provided). To this end, we tested our trained model on the widely used Visual7W (Zhu et al., 2016) dataset without any additional training.

Figure 7 presents qualitative results for `ViQAR` task on the Visual7W dataset. We also perform a Turing test on the generated answers and rationales to evaluate the model's performance on Visual7W. Thirty human evaluators were presented each with twenty five hand-picked image-question pairs, each of which contains a generated answer to the question and its rationale. The results, presented in Table 3 show that our algorithm generalizes reasonably well to another VQA dataset and generates answers and rationales relevant to the image-question pair, without any explicit training for this dataset. This adds a promising dimension to this work. More results are presented in the supplementary.

`ViQAR` is a completely generative task and objective evaluation is a challenge, as in any other generative method. For comprehensive evaluation, we use a suite of objective metrics typically used in related vision-language tasks, and perform a Human Turing Test on the generated sentences. We perform a detailed analysis in the supplementary and show that even our successful results (qualitatively speaking) may have low scores on objective evaluation metrics at times, since generated sentences may not match a ground truth sentence word-by-word. We hope that opening up this dimension of generated explanations will only motivate a better metric in the near future.

6 Conclusion

In this paper, we propose `ViQAR`, a novel task for generating a multi-word answer and a rationale given an image and a question. Our work aims to go beyond classical VQA by moving to a completely generative paradigm. To solve `ViQAR`, we present an end-to-end generation-refinement architecture which is based on the observation that answers and rationales are dependent on one another. We showed the promise of our model on the VCR dataset both qualitatively and quantitatively, and our human Turing test showed results comparable to the ground truth. We also showed that this model can be transferred to tasks without ground truth rationale. We hope that our work will open up a broader discussion around generative answers in VQA and other deep neural network models in general.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*.

Florin Brad. 2019. Scene graph contextualization in visual commonsense reasoning. In *ICCV*.

Remi Cadene, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. 2019. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. 2021. Meta module network for compositional visual reasoning. In *WACV*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL-HLT*.

Gabriel Forgues and Joelle Pineau. 2014. Bootstrapping dialog systems with word embeddings.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *ECCV*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017a. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017b. Inferring and executing programs for visual reasoning supplementary material.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *NeurIPS*.

Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *ICLR*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *WMT@ACL*.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *ICCV*.

Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. 2018a. Tell-and-answer: Towards explainable visual question answering using attributes and captions. In *EMNLP*.

Qing Li, Qingyi Tao, Shafiq R. Joty, Jianfei Cai, and Jiebo Luo. 2018b. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *ECCV*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL*.

Jingxiang Lin, Unnat Jain, and Alexander G. Schwing. 2019. Tab-vcr: Tags and attributes based vcr baselines. In *NeurIPS*.

Jiasen Lu, Dhruv Batra, D. Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*.

Will Norcliffe-Brown, Efstathios Vafeias, and Sarah Parisot. 2018. Learning conditioned graph structures for interpretable visual question answering. In *NeurIPS*.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*.

Badri N. Patro, Shivansh Pate, and Vinay P. Namboodiri. 2020. Robust explanations for visual question answering. *ArXiv*, abs/2001.08730.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2016. Self-critical sequence training for image captioning. *CVPR*.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Timothy P. Lillicrap. 2017. A simple neural network module for relational reasoning. In *NIPS*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*.

R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Jialin Wu, Zeyuan Hu, and Raymond J. Mooney. 2019. Generating question relevant captions to aid visual question answering. In *ACL*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *NeurIPS*.

Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. 2016. Image captioning with semantic attention. In *CVPR*.

Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017a. Multi-level attention networks for visual question answering. *CVPR*.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017b. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *ICCV*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019. Reasoning visual dialogs with structural and partial observations. In *CVPR*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. In *CVPR*.