TA-Student VQA: Multi-Agents Training by Self-Questioning

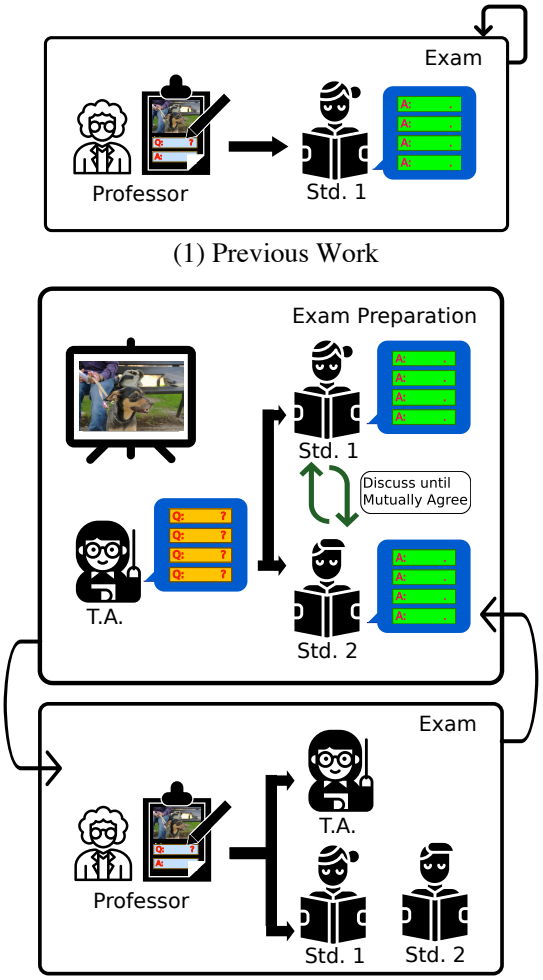
Peixi Xiong and Ying Wu
Northwestern Univeristy
2145 Sheridan Road, Evanston, IL 60208
{peixixiong2018,yingwu}@u.northwestern.edu

Abstract

There are two main challenges in Visual Question Answering (VQA). The first one is that each model obtains its strengths and shortcomings when applied to several questions; what is more, the "ceiling effect" for specific questions is difficult to overcome with simple consecutive training. The second challenge is that even the state-of-the-art dataset is of large scale, questions targeted at a single image are off in format and lack diversity in content. We introduce our self-questioning model with multi-agent training: TA-student VQA. This framework differs from standard VQA algorithms by involving question-generating mechanisms and collaborative learning between question-answering agents. Thus, TA-student VQA overcomes the limitation of the content diversity and format variation of questions and improves the overall performance of multiple question-answering agents. We evaluate our model on VQA-v2 [1], which outperforms algorithms without such mechanisms. In addition, TA-student VQA achieves a greater model capacity, allowing it to answer more generated questions in addition to those in the annotated datasets.

1. Introduction

In recent years, Visual Question Answering (VQA) has garnered significant attention [45, 16, 35, 51], as it relates to multidisciplinary research such as natural language understanding [58], visual information retrieval [19, 62], and multi-modal reasoning [4, 39]. Many methods [56, 18] have been developed in this field with datasets [36, 65, 59, 1, 15, 21, 25] for different purposes. However, for each high-performing algorithm focusing on one aspect, the algorithm obtains respective drawbacks in other aspects (e.g., some algorithms that are good at color-related questions but are not experts at reasoning, whereas other algorithms are proficient at the latter and vice versa). Meanwhile, there is a crucial point where, during their training phase, even when the question datasets are of a large scale, the number of Q/A



(1) Previous Work

(2) Our Method

Figure 1: (1) Previous work only has one student agent to train on the annotated data. (2) The TA-Student VQA paradigm. We proposed a framework for Visual Question Answering (VQA) in which the TA agent generate questions by the given image, and two student agents answer them. As their answers converge, question-answer pairs from annotated dataset will be used to evaluate these agents, and update them. This method breaks through the barriers caused by limited problems for one image, and takes advantage of the strengths for two student agents.

pairs for one image is limited. Such insufficiency is not only found in the format of questions with similar semantics but also in the lack of image-targeted questions.

Regarding the training aspect of visual questions as exam preparation and the testing part as the final exam (Figure 1), previous work used the annotated data to train one model and updated the parameters using the result. This scenario is similar to when a student prepares for an exam along. The student can only improve himself by taking the exam; however, empirically, the exam questions are limited, which leads to a difficulty in reinforcing learning and a lack of generalization. In our work, we involve an agent that plays a similar role as a teaching assistant (TA), therein generating questions based on the given image. In addition, we introduce the idea of using two Question-Answering Agents (Agts) to realize collaborative learning, similar to two students preparing for their exams without a solution manual; they can only make progress by discussing the topic with each other.

Our main contributions are the following:

1. We overcome the barrier whereby the training Q/A pairs for each image are limited in addressing the deficiency in the model capacity by adding a TA model to the system.
2. We utilize two Question-Answering Agents (Agts) in the self-questioning stage in a collaborative learning manner to combine the advantages of the two methods.
3. We obtain a better understanding of the given image through the process of self-QA, therein applying a strategy to select the most informative questions/content to best improve the visual question-answering performance.

2. Related Work

2.1. Visual Question Answering

Many visual question-answering algorithms have been proposed recently. These algorithms can be divided into four main categories. The first category is **standard deep learning models**, which typically use convolutional neural networks (CNNs) [26] to embed the image and implement recurrent neural networks (RNNs) (*e.g.* , long short-term memory units (LSTMs) [20] and gated recurrent units (GRUs) [8]) to embed natural language. Methods such as [63, 13] merge visual and textural features directly, while [46, 34, 36] put these features into a new network to achieve the combination. Although such standard deep learning methods do not always obtain excellent performance, standard methods have established milestones in the VQA task and preserve intuition in discovering the relationships between these two types of features. The second category is **attention-based deep learning techniques**. Such mechanisms are often applied by focusing on key parts of images, questions, or both [32], thereby effectively targeting output answers. Methods such as [48, 64, 57, 60] achieve good results for the task. However, they suffer from certain shortcomings, similar to standard deep learning methods. Utilizing human attention or performing direct learning does not eliminate the problem whereby deep learning methods lack good interpretable reasoning and can obtain a fairly good result just by simply memorizing statistics about the Q/A pairs. This leads to the deficiency in model capacity, that is, if you ask a question in a tone that is different from the questions in the training and testing sets in terms of format or content, even if the question is related to the image, the model will produce poor results. Despite deep learning approaches dominating the VQA field, **non-deep learning approaches** often create innovation by offering interpretative features or intermediate results. Related work, such as [23, 35, 28], attempts to build a probability model and infer hidden information to complete the model. Such methods are unlikely to suffer from overfitting and can preserve generalization. Despite this, they require feature engineering, and empirically, it takes time to choose a model adapted to new problems. **Knowledge base support methods** are another type of algorithm, including [54, 50, 37, 52, 33, 40], and utilize facts about objects in the image and their relationships. These methods show their strength in difficult cases that need the assistance of external knowledge. However, such strength requires extra time and effort for building the knowledge bases.

Currently, VQA methods retain their strengths but lack image understanding. The limitation of Q/A pairs in the dataset also leads to a lack of generalization of the model. Our self-QA method overcomes the limitation of the annotated dataset, thereby achieving the generation of informative questions based on the given image and avoiding the influence of the deficiency in the model capacity.

2.2. Visual Question Generation

As an interdisciplinary direction of image captioning [27, 11, 9, 7, 10], visual question generation [24, 61, 42] has been recently proposed as a method to generate questions rather than captions based on a given image. A good generated question should be tightly focused on images rather than on general statements such as "What is in this picture?". Our TA-Student VQA system is highly relevant. However, the main difference with our method is that, instead of generating questions that maximize the mutual information between the image and the question, our method focuses more on generating questions that will most benefit the Question-Answering Agents. More specifically, if a question such as "What is in this picture?" helps improve the Question-Answering Agents' performance, then it will be a good question for our model.

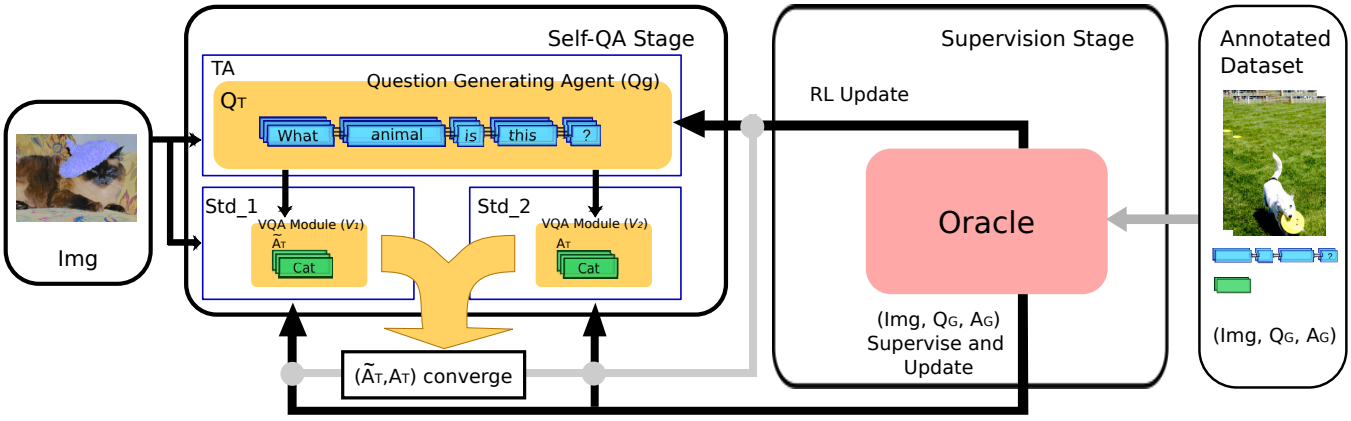


Figure 2: Overview of our approach. The system consists of two stages, Self-QA stage and Supervision Stage, and these two stages will execute iteratively. In the first stage, there are three agents, one question proposal agent TA and two question answering agent Std_1 and Std_2. Once TA agent proposes questions Q_T, Std_1 and Std_2 will output corresponding answer $\tilde{A_T}$ and A_T. Once the answers converge, the second stage begins. $Oracle$ will supervise models in Self-QA block, based on the result to update the parameters of Std_1 and Std_2, and use reinforcement learning method to update the parameters of TA.

2.3. Boosting Method

Boosting refers to a group of algorithms that turn weak learners into strong learners; more specifically, it is an ensemble method for improving the model predictions of a learning algorithm. The idea of boosting is to train weak learners sequentially, each attempting to correct its predecessor. Boosting was first introduced in [12], and later works, such as [2, 6], have also followed the same path to achieve better training of weak classifiers. Our method converts this idea into collaborative learning, and two Question-Answering Agents are trained consecutively to improve the prediction performance.

2.4. Generative Adversarial Network

Generative adversarial networks (GANs) were first proposed in [17]; they estimate generative models via an adversarial process. The main idea is to simultaneously train the generative model (G) and the discriminative model (D). G is responsible for capturing the data distribution, while D is responsible for estimating the probability that a sample came from the training data rather than G. Many works [53, 41, 49] have used GANs to perform image synthesis and image retrieval. We borrow this adversarial idea for our TA-Student VQA system by adversarially generating questions for a given image and evaluating the generated questions by our Question-Answering Agents and Oracle.

3. Approach

3.1. Overview

We now formally introduce a new approach called TA-Student VQA. The test phase of the VQA task can be formatted as finding the correct answer a in the space of candidate answer words Ans by $\text{argmax}_a P(a \mid Img, Q_G), a \in Ans$, where Img is the given image from the dataset and Q_G is the corresponding question. However, for the training phase, unlike previous methods of VQA, we alter the strategy by involving a self-QA stage. A TA model is responsible for raising questions (Q_T) for the given image Img, and two VQA models are set to output corresponding answers (A_T and $\tilde{A_T}$). Once A_T and $\tilde{A_T}$ converge, the supervision stage is performed. There is an oracle O that asks questions (Q_G) from the dataset, while the ground-truth answers (A_G) combined with Q_G and Img are used to update the two VQA models and TA model.

To facilitate self-questioning, it is possible to build two agents, each responsible for both generating questions and answering them; however, to obtain a tight structure and clear delineation of responsibilities, we design the system as illustrated in Figure 2. Our proposed method consists of two stages, the self-QA stage and the supervision stage. The first stage includes the Question Generating Agent (Q_g) and two Visual Question Answering Agents (Agt), which will be introduced in detail in Section 3.2 and Section 3.3, respectively. The second stage concerns how the oracle (O) updates the parameters of Agts and Q_g, which will be explained in Section 3.4.

3.2. Question Generating Agent (Q_g)

The Question Generating Agent (Q_g) acts as a TA, which is designed to generate a set of questions with diversity in format and content under the condition that they are related to the given image Img. To obtain these properties, three sub-models are built and combined: (1) the question generation model (g), which is responsible for proposing questions based on the given image Img; (2) the question validation model (v), which checks if the generated questions are grammatically correct and relevant to the content; and (3)

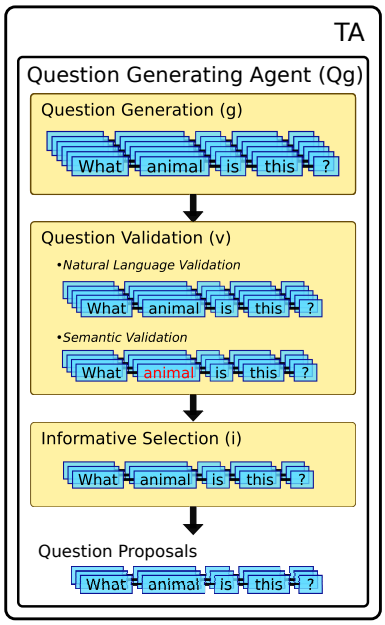


Figure 3: Question Generating Agent (Q_g). A set of questions are proposed by Question Generation Model (g), then these questions are filtered by Question Validation Model (v), to achieve the grammarly validation and the content relevance of questions. The Informative Selection Model (i) is used to select questions that contribute most to improve the Question Answering Agents (Agts).

the informative selection model (i), which selects the most informative questions from previous questions to improve the training efficiency. Figure 3 demonstrates the structure of Q_g.

– **Question Generation Model** (g)
This model generates questions Q_{T0} based on a given image Img, denoted as $Q_{T0} = g(Img)$. We build a structure similar to [38]. To better handle large-scale data [3], we use long short-term memory (LSTM) instead of gated recurrent units (GRUs). In addition, to improve the diversity of Q_{T0}, we use the question type as the first word of the generated question and randomly sample it before each generation. Here, we use the 64 types of question categories defined in [1].

– **Question Validation Model** (v)
To further filter the generated questions Q_{T0}, we design two mechanisms. The first mechanism is a grammar checker that ensures that the generated questions are valid in terms of natural language. The second mechanism checks whether the main components referred to in the question are present in the image Img, which avoids asking about an invalid object. To locate the subject of interest in the sentence, [5] is used to parse the question with grammatical relations. These two checkers serve as filters to retain Q_{T1} from Q_{T0} with $Q_{T1} = v(Q_{T0}, Img)$ and preserve more effective information.

– **Informative Selection Model** (i)
To select the most informative questions Q_{T2}, we propose the Informative Selection Model, which is a policy $\pi(Q_{T2}|Img, Q_{T1}, \theta)$ used to select Q_{T2} by the given image Img, Q_{T1} are the questions from the last model, and θ are the parameters of this model.

The task now becomes a policy-learning problem. Given an image Img and a set of question candidates $\{Q_{T2}(i) \in Q_{T2} : 1 \leq i \leq n\}$, we output a policy containing a sequence of actions $[\mathsf{a}_1, \mathsf{a}_2, \mathsf{a}_3, \ldots, \mathsf{a}_n]$. a_i is a binary value used to determine whether the question candidate $Q_{T2}(i)$ is informative. There is no ground truth for each action but rather only a final reward indicating under such policy π, in other words, under the selected proposed questions Q_{T2} in this round, whether the prediction result in the supervision stage is significantly improved. The details will be provided in Section 3.4. We use Monte Carlo concepts to learn the policy, which will guide the question selection. Such a policy network requires an extra reward value in a loss.

$$L_{policy}(\theta) = \sum\nolimits_{i \in \|Q_{T2}\|} log\pi(\mathsf{a}_i|Img, Q_{T1}, \theta)\ell(Q, \mathsf{A}) \tag{1}$$

where a_i is the action taken based on the current status, $\pi(\cdot)$ is the policy function that maps the status to actions, where the policy is the probability of outputting the next action module a_i based on the current status, and $\ell(\cdot)$ is the softmax loss in the supervision stage based on the overall action module sequence $[\mathsf{a}_1, \mathsf{a}_2, \mathsf{a}_3, \ldots, \mathsf{a}_n]$. Because all actions are discrete, which leads to a non-differentiable problem, back-propagation will not work. The policy gradient [29] is used instead during training.

We iteratively generate questions Q_{T2} until there are 100 question proposals, which are regarded as the output of this Question Generating Agent, denoted as Q_T. Model Q_g outputs question proposals that relieve the situation whereby the questions asked for images are limited, and the previously shown mechanisms guarantee that the questions are related and informative.

3.3. Question-Answering Agents (Agt)

Two Visual Question Answering Agents (Agts) act as two students to answer the questions Q_T generated by the TA, which is Q_g. Their outputs, A_t and $\tilde{A_T}$, are the results from two heterogeneous-structured Visual Question Answering models (the details of the two models will be provided in Section 3.5). Once A_t and $\tilde{A_T}$ "softly" converge

(which means that they obtain semantic similarity, as detected by [43]), this implies that after several rounds of discussion, these two students finally have come to an agreement. Then, it is time for the real exam, rather than the questions provided by the TA. In other words, the Oracle will act in supervision stage (Section 3.4), using (Img, Q_G, A_G) pairs from the annotated dataset to update the parameters in Q_g and the two Agts.

3.4. Oracle Check Model (O)

The Oracle Check Model (O) works similarly as the model used to activate learning [47]; however, it does not label the answer of the generated questions Q_T. Instead, it computes its informative score, which is shown in Section 3.2, to decide if these questions help most during the training phase.

- **Reinforcement Learning Update for Question Informative**
  Once A_t and $\tilde{A_T}$ softly converge, the Oracle Check Model (O) selects a few questions from the annotated dataset (Q_G), and two Agts output their corresponding answers; the answer with the higher confidence will be their "answer in consent". With the answer from the Agts and the ground-truth answer A_G, the softmax loss $\ell(\cdot)$ in Equation 1 can be calculated.

- **Parameters Update for Agts**
  In addition to updating the model, which determines whether the proposed questions are informative, the Oracle O is also responsible for supervising and updating the parameters of the two Agts for a few iterations. This will prevent two Agts from converging to a local optimum.

3.5. Implementation Details

3.5.1 Model Configuration

For the Question Generation Model g, we follow a similar structure to [38] and substitute GRU into LSTM to better process large-scale data. Moreover, we add a discrete variable to represent the question type and regard it as the first token of the model to improve the diversity of the questions.

To check the grammatical correctness of questions, we apply [30]. In addition, [5] is used to parse the sentence to extract the relevant objects and subjects so that it will check if the target components are present in Img.

For two Agts that answer visual questions, we choose two heterogeneous-structured models. The first model is [55], while the second model is [44]. The main difference is that after converting the input into internal representations, the first model iteratively retrieves the related facts, while the second model directly uses the image feature as the first word of the question, subsequently feeding them into the LSTM. Thus, there is no dynamic iteration. Choosing these models presents a significant variance; however, the only constraint is that they should be different in structure, and we do not regard such VQA models as our contribution. In addition, we test homogeneous-structured models in Section 4 for verification.

3.5.2 Training Details

We pre-trained g and two Agts for the first 80k iterations as a warm-up. Then, Q_g generates question proposals Q_T after filtering out invalid ones and selecting 100 questions based on the initial informative score. Two Agts output their corresponding answers A_t and $\tilde{A_T}$ based on Img and Q_T. Once A_t and $\tilde{A_T}$ softly converge, it comes to the Supervision Stage while O asks questions Q_G from the annotated dataset. The performances of the two Agts will be the basis for updating their parameters; moreover, they will help to calculate the reward value in Equation 1 and update Q_g. After rounds of such training, Agts and Q_g are trained well for further evaluation.

3.5.3 Testing Phase

In the testing phase, we disable the Q_g for its question generation function. And the answer a for question Q and image Img is given by the following equations.

$$a^j = \underset{\{a_i^j, a_i^j \in Ans, j \in ||Agts||\}}{\operatorname{argmax}} P^j(a_i^j \mid Img, Q) \tag{2}$$

$$(\mathbb{A}, \mathbb{P}) = \{(a^j, P^j) \mid j \in ||Agts||, Eq.\ 2\} \tag{3}$$

$$a = \underset{\{a^j, (a^j, P^j) \in (\mathbb{A}, \mathbb{P}), j \in ||Agts||\}}{\operatorname{argmax}} \mathbb{P} \tag{4}$$

Equation 2 and Equation 3 is to predict a set of answer/probability pair (a^j, P^j) for each Agt^j, by selecting the one with the highest probability, while Equation 4 aims to choose the final answer from the answer set a^j of each Agt^j by its confidence score P^j.

4. Experiments

4.1. Dataset

We evaluate our TA-Student VQA system on the VQA-v2 [1] dataset, which includes 82,783 training images. We use 8,000 images and their corresponding Q/A pairs as pre-training data. We evaluate the model on its validation set, which includes 40,504 images.

4.2. Models

We use the four following models as our candidate question-answering agents (Agt).

VIS_CNN [44], which uses a concept that treats an image as a word and inputs it into the LSTM with questions.

DMN [55], which is a neural network architecture that processes input sequences and questions, forms episodic memories, and generates relevant answers with its improved

(a)
Question: How many people are on the field?
Ours: 3
Ground Truth: 3

(b)
Question: Is there a mirror on the wall?
Ours: yes
Ground Truth: yes

(c)
Question: What sport is this?
Ours: baseball
Ground Truth: baseball

(d)
Question: What type of room is this?
Ours: kitchen
Ground Truth: kitchen

(e)
Question: Are there any clouds in the sky?
Ours: no
Ground Truth: no

(1) Successful Cases

(f)
Question: Is the man taking a picture?
Ours: no
Ground Truth: yes

(g)
Question: What is the wooden object?
Ours: cabinet
Ground Truth: dish rack

(h)
Question: Would this be an ideal day to fly a kite?
Ours: yes
Ground Truth: no

(i)
Question: Where is the cat sittinge?
Ours: bed
Ground Truth: couch

(j)
Question: Is there a clock on the wall?
Ours: yes
Ground Truth: no

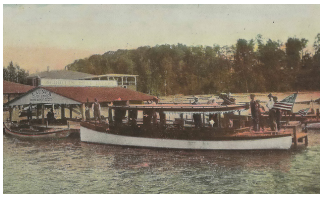
(2) Failure Cases

Figure 4: (1) Successful Cases and (2) Failure Cases of Our Model.

Iteration 100k:
What is the girl riding?
Iteration 200k:
Are there mountains in the background?
Iteration 300k:
Where is the motorcycle?
Iteration 400k:
Are they having a outdoor party?

Iteration 100k:
Is there anyone on the road?
Iteration 200k:
Is the sun out?
Iteration 300k:
What time of the day was the picture taken?
Iteration 400k:
What is the cause of the orange line across the sky?

Iteration 100k:
What is the man in red holding?
Iteration 200k:
Is there any girl wearing a uniform?
Iteration 300k:
Are these people focusing on the game?
Iteration 400k:
Will someone get dirty if he falls?



Iteration 100k:
Is there a boat in the water?
Iteration 200k:
How many boats are on the water?
Iteration 300k:
Do you think it's an early day in the scene?
Iteration 400k:
Is this a commercial boat?

Iteration 100k:
What is the man holding?
Iteration 200k:
Is the player 's pants striped?
Iteration 300k:
What sport is this?
Iteration 400k:
Is this a professional baseball player?

Iteration 100k:
What is the color of the sofa?
Iteration 200k:
Are there any paintings on the wall?
Iteration 300k:
What kind of room is this?
Iteration 400k:
What is the theme of the room?

Figure 5: Question Proposals per Iteration. With the update in the Question Generating Agent, the question proposals are with increasing sophistication. At first, it asks questions that require simple visual tasks (e.g., object detection), then asks location questions that need to detect scene by objects and their connections; further, it will ask questions that need complex inference.

memory and input modules.

LSTM_CNN [31], which encodes images and questions by a CNN and an LSTM, respectively, and then chooses answers from the candidate space with a multi-layer perceptron.

MCB [14], which is a method that utilizes MCB to efficiently and expressively combine visual and textual features.

Unless stated otherwise, we use **VIS_CNN** and **DMN** as our Question-Answering Agents (Agt) because they obtain a heterogeneous structure compared with **LSTM_CNN** and **MCB**. The performance comparison between these settings

is given in Section 4.4.1.

4.3. Qualitative Results

4.3.1 Successful Results

To obtain more in-depth results on the capability of our TA-Student VQA system, we show several representative examples from different image-question pairs in Figure 4 (1). The results indicate how our model is capable of answering questions that required multiple tasks such as counting, finding objects, and performing direct reasoning (wherein objects are required to infer the scene. e.g., Figure 4 (1) c and Figure 4 (1) d).

4.3.2 Failure Cases

Some examples of our failure cases are shown in Figure 4 (2). Our TA-Student VQA system cannot handle an action without an object being acted upon (e.g., Figure 4 (1) a), as well as cases with objects with highly similar appearances (e.g., Figure 4 (1) g and Figure 4 (2) j). For certain questions that necessitate complicated reasoning due to a lack of information, our system always outputs the answer with the highest correlations of the target objects in the question (e.g., Figure 4 (2) i, "cat" is more related to "bed" rather than to "couch" in the training dataset). Additionally, the model tends to give answers intuitively, rather than perform more in-depth reasoning first (e.g., "kite" is in the middle of the image, and the system prefers to produce an output that is an ideal day to fly a kite.)

4.3.3 Question Complexity

Figure 5 shows the generated questions with further iterations. The format of the questions varies, from "is there" to "what" to "how". Additionally, the complexity and diversity increase, from which the questions start from simply asking object attributes and progress to counting and then reasoning.

4.4. Quantitative Results

4.4.1 Overall Performance

	$\mathbf{VIS\_LSTM_v}$	$\mathbf{DMN_v}$	**VIS_LSTM+DMN (Ours)**
Acc.	52.77	57.10	62.86
	$\mathbf{LSTM\_CNN_v}$	$\mathbf{MCB_v}$	**LSTM_CNN+MCB (Homo-structured)**
Acc.	45.82	46.87	48.14

Table 1: Overall Performance Comparison (%)

Table 1 is our overall result. We compare the performance by using a combination of **LSTM_CNN** and **MCB** as the Question-Answering Agents (Agt) with **VIS_CNN** and **DMN**. For the first combination, there is a 5.06% improvement over the standard method $\mathbf{LSTM\_CNN_v}$ (i.e., no self-QA mechanism); however, for the second combination, there is a 10.59% improvement over the standard method $\mathbf{VIS\_LSTM_v}$. This indicates that either combination will improve the result. Furthermore, the more the structure differs, the greater the improvement.

Question Type	$\mathbf{VIS\_LSTM_v}$	$\mathbf{DMN_v}$	$\mathbf{VIS\_LSTM_{sq}}$	$\mathbf{DMN_{sq}}$	**Ours**
are	19.69	18.70	21.74	21.31	**21.79**
are there	**24.61**	11.29	9.00	9.71	10.32
are these	**11.60**	4.20	7.92	5.38	5.37
are they	18.79	10.43	21.98	20.10	**26.71**
can you	15.63	10.49	18.10	14.11	**21.47**
could	10.68	13.33	22.13	24.03	**29.38**
do you	8.79	12.30	18.14	14.97	**20.85**
do/does	**8.73**	4.92	7.13	4.93	5.20
does this	7.78	**15.24**	8.11	10.26	5.37
has	29.35	16.34	18.34	19.01	**29.92**
how	8.52	6.50	18.55	16.80	**20.00**
how many	15.08	5.30	25.09	26.89	**27.63**
is	31.41	21.86	38.85	37.96	**41.71**
is obj.	25.71	12.82	25.86	28.98	**29.87**
was	18.12	7.25	20.02	17.94	**20.49**
what	17.68	22.23	20.26	37.79	**45.22**
what color	9.50	13.92	30.05	26.32	**31.70**
what does	**15.33**	14.63	15.09	15.16	15.32
what is/are	11.26	5.91	19.51	16.28	**22.03**
what numb.	5.57	**8.50**	7.94	7.77	8.13
what obj. is	8.25	11.92	12.24	14.57	**15.79**
what time	3.54	4.37	**4.90**	3.29	4.12
what type	18.37	7.51	21.44	18.30	**22.67**
where	8.39	12.76	30.46	32.16	**34.56**
which	14.03	17.94	15.85	20.72	**20.77**
who	14.51	7.80	14.78	10.05	**14.92**
why	13.84	6.46	19.73	17.56	**20.52**
others	8.69	5.70	**9.58**	4.84	5.02
total	14.41	11.09	17.96	17.76	**20.60**

Table 2: Accuracy per Generated Question in Dataset (%)

4.4.2 Performance per Annotated Question

Figure 6 demonstrates the accuracy of each question in the annotated dataset. The figure shows that our model is typically effective for questions such as "was/is/are", "could", "have/has", "what" and "who" questions. For questions that need in-depth reasoning (e.g., "how" and "why"), the model obtains the opportunity to improve.

4.4.3 Performance per Generated Question

To validate our question generation model, we evaluate our model on our question proposals. For each category, we test 100 generated questions. We evaluate the results through manual inspection. The results are shown in Table 2. Here, $\mathbf{LSTM\_CNN_v}$ and $\mathbf{MCB_v}$ are the standard methods, serving as baselines, with no other mechanism.

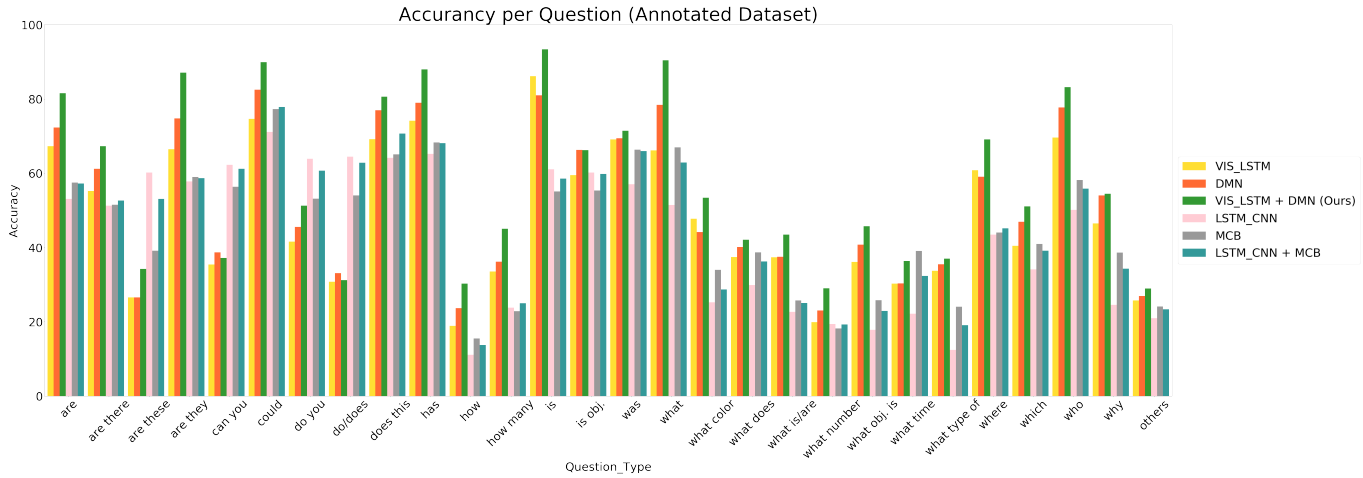


Figure 6: Accuracy per Annotated Question in Dataset

In $\mathbf{LSTM\_CNN_{sq}}$ and $\mathbf{MCB_{sq}}$, the self-questioning mechanism is implemented; however, only one question-answering agent (either **LSTM_CNN** or **MCB**, respectively) is used to obtain the result. Finally, **Ours** is the proposed technique, in which the self-questioning mechanism is implemented and both **LSTM_CNN** and **MCB** are used as question-answering agents; between the results generated by these two agents, the one with the higher confidence score is chosen as the final result.

The table tells us the following: (1) With a larger capacity, our model is able to address questions through generalization. (2) With such a question proposal mechanism, our model obtains improved reasoning questions (e.g., "how", "where" and "why"). (3) With the collaborative learning between two agents, $\mathbf{LSTM\_CNN_{sq}}$ and $\mathbf{MCB_{sq}}$, an overall better performance is obtained compared to the standard method, i.e., $\mathbf{LSTM\_CNN_{v}}$ and $\mathbf{MCB_{v}}$, and the difference between two Agts' total accuracy gets smaller.

4.4.4 Question Generation Strategy

	w/o Q_g	w/ Q_g, $\epsilon = 0.1$	w/ Q_g, $\epsilon = 0.4$	w/ Q_g, $\epsilon = 0.7$
VIS_LSTM+DMN (Ours)	58.92	59.14	61.08	62.86

Table 3: Accuracy with Different Question Generation Strategy (%)

To validate the question generation model (Q_g), we evaluate it with several different settings, as shown in Table 3. From the last three columns, we observe that when a higher ϵ value is used in epsilon-greedy [22], the question proposals will be more diverse, thus improving the overall accuracy.

5. Conclusion and Future Work

This paper introduces a self-QA pattern and proposes a system based on this idea. Our TA-Student VQA system utilizes a TA agent (Q_g), which is responsible for generating informative questions, and two student agents (Agts) that answer the proposed questions. The O model plays the supervision role and updates the previous three models to guarantee that the training phase is efficient. Unlike previous work, our mechanisms overcome the barrier whereby the training Q/A pairs for each image are limited because our system can generate questions diverse formats and content. Additionally, utilizing two question-answering agents (Agts) combines the advantages of the two methods and increases the system capacity. Our results also show that such a self-QA mechanism not only performs better in the annotated dataset but also performed well for questions that are off the distribution of the training data, thereby improving the generalization ability.

The study of this problem is still in its infancy. One issue to be addressed is the number of Question-Answering Agents (Agts). We attempted to use more than two Agts, but the time cost of the training phase makes it less competitive than two Agts because there are more parameters to update and more computations when attempting to converge three or more parties. This provides clear direction for future work: developing an efficient method to reduce the computation costs and improve the parameter update efficiency. Second, with our Question Generating Agent (Q_g), the system achieves improved performance. Thus, another direction for our future work is to develop a system that is not only responsible for generating reverent, diverse, and informative questions but also can produce reliable corresponding answers to achieve self-labeling VQA.

Acknowledgement

This work was supported in part by National Science Foundation grant IIS-1619078, IIS-1815561, and the Army Research Office ARO W911NF-16-1-0138.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433, 2015. 1, 4, 5

[2] L. Breiman. Arcing the edge. Technical report, Technical Report 486, Statistics Department, University of California at . . . , 1997. 3

[3] D. Britz, A. Goldie, M.-T. Luong, and Q. Le. Massive exploration of neural machine translation architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1451, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 4

[4] R. Cadène, H. Ben-younes, M. Cord, and N. Thome. MUREL: multimodal relational reasoning for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1989–1998, 2019. 1

[5] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 740–750, 2014. 4, 5

[6] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. 3

[7] X. Chen and C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2422–2431, 2015. 2

[8] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734, 2014. 2

[9] W. Daelemans, M. Lapata, and L. Màrquez, editors. *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*. The Association for Computer Linguistics, 2012. 2

[10] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1473–1482, 2015. 2

[11] A. Farhadi, S. M. M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, pages 15–29, 2010. 2

[12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory, Second European Conference, EuroCOLT '95, Barcelona, Spain, March 13-15, 1995, Proceedings*, pages 23–37, 1995. 3

[13] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 457–468, 2016. 2

[14] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv:1606.01847*, 2016. 6

[15] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. 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. 1

[16] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *Proc. Natl. Acad. Sci. U.S.A.*, 112(12):3618–3623, 2015. 1

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. 3

[18] D. Gurari and K. Grauman. Crowdverge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017.*, pages 3511–3522, 2017. 1

[19] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617, 2018. 1

[20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[21] D. A. Hudson and C. D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709, 2019. 1

[22] J. D. Johnson, J. Li, and Z. Chen. Reinforcement learning: An introduction: R.S. sutton, A.G. barto, MIT press, cambridge, MA 1998, 322 pp. ISBN 0-262-19398-1. *Neurocomputing*, 35(1-4):205–206, 2000. 8

[23] K. Kafle and C. Kanan. Answer-type prediction for visual question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4976–4984, 2016. 2

[24] R. Krishna, M. Bernstein, and L. Fei-Fei. Information maximizing visual question generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2008–2018, 2019. 2

[25] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017. 2

[27] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1601–1608, 2011. 2

[28] X. Lin and D. Parikh. Active learning for visual question answering: An empirical study. *CoRR*, abs/1711.01732, 2017. 2

[29] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 4

[30] E. Loper and S. Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002. 5

[31] J. Lu, X. Lin, D. Batra, and D. Parikh. Deeper lstm and normalized cnn visual question answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN, 2015. 6

[32] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 289–297, 2016. 2

[33] P. Lu, L. Ji, W. Zhang, N. Duan, M. Zhou, and J. Wang. R-VQA: learning visual relation facts with semantic attention for visual question answering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1880–1889, 2018. 2

[34] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3567–3573, 2016. 2

[35] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc., 2014. 1, 2

[36] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1–9, 2015. 1, 2

[37] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204, 2019. 2

[38] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 4, 5

[39] H. Nam, J. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2156–2164, 2017. 1

[40] M. Narasimhan and A. G. Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, pages 460–477, 2018. 2

[41] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2642–2651, 2017. 3

[42] B. N. Patro, S. Kumar, V. K. Kurmi, and V. P. Namboodiri. Multimodal differential network for visual question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4002–4012, 2018. 2

[43] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en. 5

[44] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015. 5

[45] M. Ren, R. Kiros, and R. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Advances in Neural Inf. Process. Syst*, 1(2):5, 2015. 1

[46] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2953–2961, 2015. 2

[47] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 5

[48] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4613–4621, 2016. 2

[49] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen. Binary generative adversarial networks for image retrieval. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 394–401, 2018. 3

[50] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, and J. Li. Learning visual knowledge memory networks for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7736–7745, 2018. 2

[51] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S. C. Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 21(2):42–70, 2014. 1

[52] P. Wang, Q. Wu, C. Shen, A. R. Dick, and A. van den Hengel. FVQA: fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427, 2018. 2

[53] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 82–90, 2016. 3

[54] Q. Wu, P. Wang, C. Shen, A. R. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4622–4630, 2016. 2

[55] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2397–2406. JMLR.org, 2016. 5

[56] C. Yang, K. Grauman, and D. Gurari. Visual question answer diversity. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018.*, pages 184–192, 2018. 1

[57] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 21–29, 2016. 2

[58] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 1039–1050, 2018. 1

[59] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2461–2469, 2015. 1

[60] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian. Deep modular co-attention networks for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6281–6290, 2019. 2

[61] J. Zhang, Q. Wu, C. Shen, J. Zhang, J. Lu, and A. van den Hengel. Goal-oriented visual question generation via intermediate rewards. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, pages 189–204, 2018. 2

[62] X. Zhao, H. Li, X. Shen, X. Liang, and Y. Wu. A modulation module for multi-task learning with applications in image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–416, 2018. 1

[63] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *CoRR*, abs/1512.02167, 2015. 2

[64] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016. 2

[65] Y. Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004, 2016. 1