Abstract

Methods for teaching machines to answer visual questions have made significant progress in the last few years, but although demonstrating impressive results on particular datasets, these methods lack some important human capabilities, including integrating new visual classes and concepts in a modular manner, providing explanations for the answer and handling new domains without new examples. In this paper we present a system that achieves state-of-the-art results on the CLEVR dataset without any questions-answers training, utilizes real visual estimators and explains the answer. The system includes a question representation stage followed by an answering procedure, which invokes an extendable set of visual estimators. It can explain the answer, including its failures, and provide alternatives to negative answers. The scheme builds upon a framework proposed recently, with extensions allowing the system to deal with novel domains without relying on training examples.

1. Introduction

Visual question answering is inspired by the remarkable human ability to answer specific questions on images, which may require analysis of subtle cues, along with the integration of prior knowledge and experience. The learning of new visual classes, properties and relations, can be easily integrated into the question-answering process. Humans can elaborate on the answers they give, explain how they were derived, and why they failed to produce an adequate answer. Current approach to handle this problem by a machine \cite{35, 32, 40, 20} takes a different path, where most answering systems are trained directly to select an answer from common answers of a training set based on fused image features (mostly using a pre-trained CNN \cite{15}) and question features (mostly using an RNN, e.g. LSTM \cite{25}).

The approach we take below does not rely on any question-answering training, but uses instead a process composed according to the question's structure, and applying a sequence of 'visual estimators' for object detection and identifying a set of visual properties and relations. Answering is divided into two stages. First, a graph representation is generated for the question, in terms of objects classes, properties and relations, supplemented with quantifiers and logical connectives. An answering procedure then follows the question graph, and seeks either a single or multiple assignments of the classes, properties and relations in the graph to the image (Section 3.3). The method...

---

Figure 1. Examples for UnCoRd full answers to visual questions. Intermediate results are plotted.
builds upon the general scheme used by UnCoRd (Understand, Compose and Respond) [37] framework introduced recently.

Our work includes several novel contributions. First, a method that produces state-of-the-art results on the CLEVR visual question answering dataset [18] without any questions-answers training. Second, we developed sequence-to-sequence based method to map questions into their graph representation. Third, the method deals with questions that include quantifies and logical connectives. Fourth, we present a model that can both perform well on CLEVR, as well as generalize to novel domains by adding visual estimators (for objects, properties and relations) but without QA examples.

Examples for UnCoRd’s answers are given in Figure 1. Using the UnCoRd method we demonstrate that a visual question answering system, without any questions-answers training, provides state-of-the-art results on a challenging dataset. It is modular, extendable, utilizes external knowledge and provides elaborated answers, including alternatives when an answer is not found, and notifying about unsupported categories. In addition we provide models that can represent questions beyond the domain of a particular dataset. This demonstrates the potential to build a general answering scheme, not coupled to a dataset. Current visual question answering methods require many question-answer examples and are fitted to specific datasets (including specific answers) and exploit their biases. They lack the UnCoRd abilities to generalize across datasets, provide explicit reasoning, and the ability to easily add additional visual estimators (e.g. novel object classes) without changes to the answering scheme.

2. Related Work

Much work has been done on visual question answering in recent years [35, 32, 40, 20], developing several methods applied to a number of datasets [7, 11, 18, 23, 13, 38]. With some variations, most of this work shares the approach of handling the problem as a multi-class classification problem, selecting answers from the training set. A fusion of image features (based on Convolutional Neural Network) and question features (mostly based on Recurrent Neural Network) is used to predict the answer. This approach provides the ability to obtain successful results for a target dataset without the need to “understand” the question explicitly. However, the approach lacks desired human skills such as dealing with novel domains without question-answering training, or provide explanations and alternative suggestions when answers are not found.

As end-to-end training dominates current answering schemes, many works focused on improving the image-question fused features [10, 8, 45], various attention mechanisms for selecting important features [43, 17, 31, 29, 4] and incorporating outputs of other visual tasks [12, 3, 9, 36].

Some methods provide reasoning by using “facts” extraction (e.g. scene type) [39] or image caption results [28, 1, 27]. Other methods focused on integrating external prior knowledge, mostly by producing a query to a knowledge database using the question and the image [38]. Extracted external knowledge was also fused with question and image representations [41, 26].

A compositional approach that builds a dynamic network out of trained modules is proposed by the Neural module Network (NMN) works. The modules structure was originally based on the dependency parsing of the question [6, 5]. The following versions included supervised learning of the modules arrangement [19, 16] according to annotations of the answering programs, that are available for the CLEVR dataset [18]. While the assignment of modules is according to a meaningful learned program, the modules are trained only as components of an answering network for a specific dataset. They do not function as independent visual estimators and hence could not be modified or replaced by exiting methods, and consequently modular addition of independent modules is not possible. As other methods, a large amount of question-answer examples is required for training the system, in addition to question-program training. The answers are selected by classification, with no means for providing explanations or proposing alternatives. Our approach, in contrast, allows flexible integration of additional visual capabilities (e.g. novel object class), providing elaborated answers, propose alternatives, and use external knowledge. This is obtained without any question-answer examples.

In parallel to our work, a method that learns to generate a program and carry it out according to full scene analysis (object detection and properties classification) was proposed [44]. This method uses questions-answers training to learn the programs. It performs full scene analysis which may become infeasible for data sets that are less restricted than CLEVR. In our method, the answering process is guided by the question and does not perform a full scene analysis in order to produce the answer.

The framework that we follow [37] splits the answering task into question-to-graph mapping, followed by a recursive answering procedure. Mapping to graph representation utilizes the START parser [21, 22] to obtain a representation, where the nodes represent objects and their required information (e.g. properties and quantifiers), and edges represent relations between objects. The answering procedure utilizes several visual estimators for detecting objects and classifying properties and relations between them. External knowledge database [33] was used to extract information on question concepts that relates them to recognizable classes (e.g. finding synonyms). In our work, we train novel question-to-graph mappers, and apply them to differ-
Figure 2. A schematic illustration of the answering process. The first stage maps the question into a graph representation using a sequence-to-sequence LSTM based model. At the second stage, the recursive answering procedure follows the graph, searching for a valid assignment in the image. At each step, the current node (cur_node) is set and the objects are examined according to node’s requirements. If succeeded, a new cur_node is set (according to a relation, or next root of a rooted subgraph) and the function is called again to handle the subgraph defined by the original graph excluding the assigned nodes and edges. The child object detection is activated only when no corresponding object was detected in previous stages.

1. **Question mapping into a graph representation** - First, a representation of the question as a directed graph is generated, where nodes represent objects and edges represent relations between objects. Graph components include object classes, properties and relations. The node representation includes all the object visual requirements needed to answer the question, which is a combination of the following:
   - Object class \(c\) (e.g., ‘horse’).
   - Object property \(p\) (e.g., ‘red’).
   - Queried object property \(f\) (e.g., ‘color’).
   - Queried set property \(g\) (e.g., ‘number’).
   - Quantifiers (e.g., ‘all’, ‘two’).
   - Quantity relative to another node (e.g., same).
   - Node type: regular or SuperNode: union of a few nodes (with optional additional requirements).

2. **Answering procedure** - In this stage, a recursive procedure finds valid assignments of the graph in the image. The number of required assignments for each node is determined by its quantifiers. The procedure follows the graph, invoking relevant sub-procedures and integrates the information to provide the answer. It depends only in the structure of the question graph, where the particular object classes, properties and relations are parameters, used to apply the corresponding visual estimators (e.g., which property to extract). The invoked sub-procedures are selected from a pool of the following basic procedures, which are simple visual procedures used to compose the entire answering procedure:
   - Detect object of a certain class \(c\).
   - Check the existence of object property \(p\).
   - Return an object property of type \(f\).
   - Return an object’s set property of type \(g\).
   - Check the existence of relation \(r\) between two objects (e.g., ‘looking at’).
3.2. Question to Graph Mapping

We've performed an LSTM based sequence to sequence training [34] from question to graph representation. Graph annotations are based on the CLEVR dataset [18] programs corresponding to the dataset’s questions. The programs can be described as trees, where nodes are functions performing visual evaluations for object classes, properties and relations, and can be transferred to our graph representation, providing annotations for our mappers training. Our models configuration is based on Google’s Neural Machine Translation model [42] trained using tensorflow implementation [30]. The graph was serialized (using DFS traversal) and represented as a sequence of strings (including special tokens for graph fields), so the model task is to translate the question sequence into the graph sequence. The initial training was done using questions used in CLEVR. To generalize the scheme to a larger range of visual elements (classes, properties and relations) beyond the limited set used in CLEVR, we trained the mapper on modified sets of questions, in which CLEVR visual elements were replaced by visual elements from a larger set. Note that as this stage deals with question mapping and not questions answering, the questions do not have to be meaningful (e.g. "What is the age of the water?") as long as they have a proper mapping, preserving the role of each visual element (see Figure 3). We make sure that graph’s replacements correspond to question’s by a preliminary modification of each visual element’s synonyms into one form (e.g. ‘ball’ is replaced with ‘sphere’). In addition, all appearances of a particular visual element in a question are replaced with one destination term (for the same question). We used four ‘modes’ of replacing visual elements, differing in the set of visual elements, as described below.

- No replacement: CLEVR categories (3 object classes, 12 properties, 4 property types and 4 relations).
- Minimal replacement: Visual elements are selected from a pool that includes UnCoRd’s real world recognizable categories (100 object classes, 32 properties, 7 property types and 82 relations).
- Extended replacement: Visual elements are selected from enlarged lists of real world categories (230 object classes, 200 properties, 53 property types and 160 relations).
- VG replacement: Visual elements are selected from the categories of the Visual Genome dataset [24] (65,178 object classes, 53,498 properties, 53 property types and 47,448 relations), which include many inaccuracies, such as mixed categories (e.g. ‘fat fluffy clouds’) and irrelevant concepts (e.g. object classes: ‘there are white’, properties: ‘an elephant’, relations: ‘wheels’). Using these concepts results with an inconsistent mapping (as ‘fat fluffy clouds’ is listed as an object class but actually includes two properties: ‘fat’ and ‘fluffy’). The replacement is done with probability corresponding to the statistics of the dataset, hence probability of noisy elements is expected to be low.

The vocabulary we use for training all sets is the same. It is a 56,000 words vocabulary that was composed of the union of a standard English vocabulary and all the used object classes, properties and relations. Both the question and the graph representations are based on the same vocabulary where the graph has additional tokens to mark graph nodes and fields (e.g. <NewNode>, <p>).

To further increase the representation scope, the diversity of questions should also be addressed. CLEVR questions contain long sequences of various requirements. However it lacks some basic question elements (e.g. quantifiers and difference queries) and does not include simple questions, i.e. questions with very few explicit requirements. This bias also biases the mapping results of simpler questions (e.g. "Is there a book on a shelf?"). To address this, we created enhanced sets where additional examples were added to each of the above sets. These examples include:

- Questions where ‘same’ is replaced by ‘different’. This is performed on questions that include ‘same’ only as part of the relation ‘same p’ (where p is a property).
- Questions with added quantifiers (‘all’ and numbers). The quantifiers are added in questions that fit this addition (a few quantifiers may be added to a question).
- Basic questions that include existence and count for: class, class and property, class and 2 properties, 2 objects and a relation, as well as queries for objects class (in a relation) and property types (including various WH questions). Each group of questions is divided to questions with and without quantifiers.

An example for a graph, mapped using the ‘Enhanced-Extended replace’ model is given in Figure 3. It is evident that although the question does not have a meaning, it has a structure in terms of objects, properties and relations that can be mapped into a question graph, in the same way that the original question was mapped. All visual elements are mapped properly representing the same structure of the original question, only with the replaced visual elements. This means that the same answering procedure will be carried out, fulfilling our intent to apply the same procedure to similar structured questions.

3.3. Answering Procedure

In this stage we follow the general UnCoRd scheme [37], where a recursive procedure seeks valid assignments (Section 3.1) between the question graph and the image. The question graph, the image and the mask R-CNN [14] output (activated on the image) are fed into the procedure that
Q: What is the fabric of the object that is both walking towards all the light blue girls and next to the sixteen full tied-up tiled objects?

Figure 3. An example of a question and a corresponding graph, mapped using Extended-Enhanced model. The original CLEVR question is: ‘What is the size of the object that is both right of the cyan sphere and left of the tiny red metallic object?’ The accuracy of the representation can be confirmed by the accurate representation of the original question, when graph concepts are replaced with the corresponding original ones.

3.3.1 CLEVR Visual Estimators

In order to find a valid assignment of a question graph in the image, and provide the answer, corresponding visual estimators need to be trained. Object locations are not explicitly provided for CLEVR images, however they can be recovered using the provided scene annotations. This provided approximated contour annotations for CLEVR objects (see Figure 4), which were sufficient for training decent estimators. Mask R-CNN [14] was used for instance segmentation. For property classifiers, simple CNN models (3 convolutional layers and 3 fully connected layers) were trained to classify color and material, while size was estimated according object’s bottom coordinates and its largest edge. Relations are classified by rule based functions.

4. Experiments

Testing the full UnCoRd answering method includes two parts: one is creating a correct graph representation of the question, including for questions out of the original domain, and the second is whether, given the question representation, and assuming the availability of the necessary visual estimators, the general procedure used in UnCoRd produces the correct answer. If UnCoRd performs correctly, then question answering in a new domain would not require any specific training, and will only require visual estimators applicable to the domain.

For our evaluations, we trained 8 question-to-graph models that include all replacement modes (no-replacement, minimal, extended and VG), each trained in two forms: Basic, i.e. no added question examples (∼700K examples) and enhanced, i.e. with additional examples (∼1.4M examples). See Section 3.2 for further details.

In the tests below we first analyze representation results of the different question-to-graph models for their corresponding validation sets, as well as for validations sets of other models. This evaluates the generalization capabilities of the different models. The representation is also evaluated for a sample of free questions asked on real world images (sampled from the VQA dataset [7]). We next examine the visual estimators quality, and the combined quality to answer questions on the CLEVR dataset and for freely asked questions of the CLEVR-Humans data set [19]. Unless stated, system was configured to provide short answers; markings on images are related to intermediate results and calculations.

4.1. Question to Graph

We first report the results for mapping questions to their graphs, each model for its corresponding validation set. We report results of the BLEU scores (commonly used in machine translation) and accuracy for the various mapping models in Table 1.

| Replace Type | Basic BLEU | Acc. | Enhanced BLEU | Acc. |
|--------------|------------|------|---------------|------|
| None         | 100        | 100  | 100           | 99.8 |
| Minimal      | 99.8       | 98.4 | 99.8          | 97.7 |
| Extended     | 99.6       | 96.2 | 99.5          | 95.8 |
| VG           | 96.9       | 76.9 | 96.9          | 77.1 |

Table 1. Question-to-graph mapping results on validation sets
We use the harsh accuracy measure for this task, since unlike most other translation tasks (e.g. language translation), dislocation of even one word in the result, may cause a failure in question representation and in answering it, where such errors in language translation may be acceptable.

For checking the generalization of the question-to-graph mapping across the different sets, we tested each model on the validation sets of all 8 models. Results are given in Table 2. Note that there is a difference between the "None" models and the others. The "None" data includes mapping from concepts to their synonyms, where terms like "ball" and "block" in the question are mapped to "sphere" and "cube" in the graph, respectively. In the other models, for each category there is a wide range of terms and they are mapped directly to the graph, leaving synonyms identification for the external knowledge queries and additional processing.

In this evaluation, the results for "None" data predicted by the other models include a preprocessing stage transforming concept synonyms to a single form.

### Table 2. Accuracy results of question-to-graph mapping, evaluated on validations sets, for all data types (B: Basic, E: Enhanced)

| Test Type | Basic | Enhanced |
|-----------|-------|----------|
| None      |       |          |
| Minimal   |       |          |
| Extended  |       |          |
| VG        |       |          |

Results demonstrate that models do not generalize well on new elements and perform poorly on data that include visual categories and question phrasings "unseen" during training. However, when trained on richer vocabulary and question types, results accuracy decreases, but generalization over data with "reduced" vocabulary is high. Increasing vocabulary size and diversity of training data appears beneficial, as the Extended-Enhanced model obtains very high accuracy practically on all sets of data, other than VG. VG is different as other than including a very rich vocabulary, its data includes many incompatible elements (see Section 3.2). Additional tests are required to check possible advantages of VG models on representing different domains. We report such a test next.

### 4.2. VQA representation

In order to check the representation generality of the graph mapping, we would like to examine its results for different domains. Since VQA datasets (except CLEVR) do not include annotations corresponding to our graph representation, we sampled 100 questions of the VQA validation set [7] and manually examined the results for all the models.

The results in Table 3 demonstrate the large gaps in the abilities of models to represent new domains. Examples for several models are given in Figure 5. models trained specifically on CLEVR do not generalize at all to the untrained domain. As the models are trained on a more diverse data results improve substantially, peaking clearly for VG-Enhanced model by a large margin from other models. This result is interesting as answering CLEVR questions using this model is also performed with high accuracy (see Table 5). It means that structured description of questions provides a promising direction for answering systems of visual questions. An interesting direction would be to further investigate means to enrich question description examples and produce further significant improvements.

### Table 3. Accuracy of graph representation for VQA [7] sample

| Replace Type | Basic | Enhanced |
|--------------|-------|----------|
| None         | 1     | 0        |
| Minimal      | 12    | 12       |
| Extended     | 22    | 22       |
| VG           | 34    | 50       |

### 4.3. Visual Estimators Performance

Results for CLEVR estimators are given in Table 4. As the visual elements are quite constrained, estimators accuracy is very high and should suffice to provide accurate answers. Estimating CLEVR relations is based on simple rules using the coordinates of the objects.

### Table 4. CLEVR estimators results on CLEVR validation set

### 4.4. Answering CLEVR questions

We evaluated the UnCoRd system with the various question-to-graph mapping models on the CLEVR test set. The results are given in Table 5.
As can be seen, using the mapper models trained specifically on CLEVR data (the two None models) achieve state-of-the-art results with results close to perfect. When checking 10,000 examples of the validation set, all wrong answers were due to wrong estimations of the visual estimators, mainly miss detection of a highly occluded object. Hence, accurate annotation of object coordinates may even further reduce the small amount of remaining errors. Other models, which were trained on a much wider vocabulary and question types still perform well, mostly with only minor accuracy reduction. This demonstrates that our approach can achieve state-of-the-art results without using any question-answer examples, and at the same time it offers additional "human-like" advantages of modularity, elaborations and explanations of answers and failures, and use of external knowledge.

Examples for questions on the CLEVR data (both CLEVR questions and others) are shown in Figure 6. Question-to-graph mapping was done by the None-Enhanced model. Results for the IEP-strong model [19] are given as well. As expected, the end-to-end model provides accurate answer to the question from the original CLEVR data, but much less accurate for questions from unseen types.

4.5. Extensibility and different domains

Another demonstration of the UnCoRd system robustness and modularity can be obtained by creating a new dataset and testing the results, given the corresponding detectors. Simple extensions may only add questions with new properties or relations that can be used with existing images (which means using the same object categories). For standard end-to-end models, adding even simple relations, such as 'bigger than', would require a tuning of the entire model, whereas UnCoRd needed only a simple plugging in of this relation detector, with no further modifications. Moreover, the system can handle entirely different domains by incorporating the relevant estimators. Many of the estimators are general and can be used regardless of the type of data (e.g. 'to the left of') or be available and invoked according to the need of each domain (e.g. 'looking at'). Examples for a simple extension and for a different domain (each using a different model) are given in Figure 7, including comparison to the IEP-strong model [19].

4.6. CLEVR Humans

An example for using the CLEVR images with different questions is the CLEVR-Humans [19], where people were asked to provide challenging questions for CLEVR images. The questions vary from simple questions (e.g. 'What color is the ball?'), and questions similar to the original CLEVR form, to questions that are phrased differently and require prior knowledge (e.g. 'How many of these things could be stacked on top of each other?'). Results for the CLEVR-Humans test set (7145 questions) are given in Table 6, including comparison to the IEP models [19].

Results demonstrate that for models without finetuning, our "None-Enhanced" model provide state-of-the-art results (without any answer examples). The "None" models are
Figure 7. Examples for the extensibility of our method. Left: 'smaller than' relation is supported by integrating a simple estimator for it. Right: Entirely different domain of real-world properties and relations is supported by just having the required visual estimators. Different mask R-CNN model is used. Otherwise, all available estimators were unchanged.

Table 6. Accuracy of CLEVR-Humans dataset question answering for the different question-to-graph models and two IEP models: one trained on CLEVR and finetuned to CLEVR-Humans

| Method       | No Finetune | Finetune |
|--------------|-------------|----------|
| IEP-18k      | 54.0        | 66.6     |
| UnCoRd-None  | Basic       | 60.46    | 60.59    |
|              | Enhanced    |          |          |
| UnCoRd-Min   | Basic       | 48.24    |          |
|              | Enhanced    | 52.23    |          |
| UnCoRd-Ext   | Basic       | 43.97    |          |
|              | Enhanced    | 52.83    |          |
| UnCoRd-VG    | Basic       | 43.47    |          |
|              | Enhanced    | 48.71    |          |

Figure 8. Examples for CLEVR-Humans questions including answers of the IEP [19] models (IEP-Strong: trained on CLEVR and IEP-Hum: finetuned for CLEVR-Humans). No-E and VG-E refer to the None-Enhanced and VG-Enhance models respectively.

5. Conclusion and Future Directions

Unlike end-to-end methods for VQA, in the proposed approach the system first produces an explicit representation of the question’s meaning, in terms of a graph that needs to be searched in the image. The answering algorithm then proceeds to match the question graph to the image, guided by the graph structure, by applying sequentially visual estimators. Based on this approach, the UnCoRd system achieves near perfect results on a challenging dataset, without using any question-answer examples. It can also explain its answers and suggest alternatives when answers are not found. We have demonstrated that the representation capabilities of questions can be extended outside the scope of the trained dataset, preserving good results for the original domain.

Substantial work is required to obtain a system that will be able to perform well on entirely general images and questions. The main immediate bottleneck is obtaining question-to-graph mapping with general representation capabilities for a broad range of questions. Question graph representation may also be enhanced to support questions with more complex logic, as well as extending the scope of the supported visual categories (e.g. including global scene types). Additional basic areas that current schemes, including ours, have only begun to address, are the use of external, non-visual knowledge in the answering process, and the composition of detailed, informative answers, integrating the language and visual aspects of VQA.
References

[1] S. Aditya, Y. Yang, and C. Baral. Explicit reasoning over end-to-end neural architectures for visual question answering. In AAAI, 2018. 2

[2] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. In Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, Texas, USA, 2016. 3

[3] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2

[4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998, 2017. 2

[5] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 2016. 2

[6] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 39–48, 2016. 2

[7] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In International Conference on Computer Vision (ICCV), 2015. 2, 5, 6

[8] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. In The IEEE International Conference on Computer Vision (ICCV), 2017. 2

[9] M. T. Desta, L. Chen, and T. Kornuta. Object-based reasoning in vqa. In Winter Conference on Applications of Computer Vision, WACV, IEEE, 2018. 2

[10] 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 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, Texas, USA, 2016. 2

[11] Y. Goyal, T. Khot, D. Summers-Stuy, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2

[12] T. Gupta, K. Shih, S. Singh, and D. Hoiem. Aligned image-word representations improve inductive transfer across vision-language tasks. In The IEEE International Conference on Computer Vision (ICCV), Oct 2017. 2

[13] 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In Proceedings of the International Conference on Computer Vision (ICCV), 2017. 4, 5

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997. 1

[16] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 804–813, 2017. 2

[17] I. Ilievski, S. Yan, and J. Feng. A focused dynamic attention model for visual question answering. arXiv preprint arXiv:1604.01485, 2016. 2

[18] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, 2017. 2, 3, 4

[19] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. In ICCV, 2017. 2, 5, 7, 8

[20] K. Kafle and C. Kanan. Visual question answering: Datasets, algorithms, and future challenges. arXiv preprint arXiv:1610.01465, 2016. 1, 2

[21] B. Katz. Using english for indexing and retrieving. In Proceedings of the 1st RIAO Conference on User-Oriented Content-Based Text and Image Handling (RIAO ’88), 1988. 2

[22] B. Katz. Annotating the world wide web using natural language. In RIAO, pages 136–159, 1997. 2

[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332, 2016. 2

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidies, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1):32–73, 2017. 4

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998. 1

[26] G. Li, H. Su, and W. Zhu. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. arXiv preprint arXiv:1712.00733, 2017. 2

[27] Q. Li, J. Fu, D. Yu, T. Mei, and J. Luo. Tell-and-answer: Expanding, elaborating, and enhancing your answers for visual questions. arXiv preprint arXiv:1803.07464, 2018. 2

[28] P. Lu, H. Li, W. Zhang, J. Wang, and X. Wang. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In AAAI, 2018. 2
[30] M. Luong, E. Brevdo, and R. Zhao. Neural machine translation (seq2seq) tutorial. https://github.com/tensorflow/nmt, 2017. 4

[31] D.-K. Nguyen and T. Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. arXiv preprint arXiv:1804.00775, 2018. 2

[32] S. Pandhre and S. Sodhani. Survey of recent advances in visual question answering. arXiv preprint arXiv:1709.08203, 2017. 1, 2

[33] R. Speer and C. Havasi. Conceptnet 5: A large semantic network for relational knowledge. In The People’s Web Meets NLP: Collaboratively Constructed Language Resources, pages 161–176. Springer Berlin Heidelberg, 2013. 2

[34] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems 27, pages 3104–3112. Curran Associates, Inc., 2014. 4

[35] D. Teney, P. Anderson, X. He, and A. v. d. Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. arXiv preprint arXiv:1708.02711, 2017. 1, 2

[36] D. Teney, L. Liu, and A. van den Hengel. Graph-structured representations for visual question answering. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017. 2

[37] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. 2

[43] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bi-linear pooling with co-attention learning for visual question answering. In Proc. IEEE Int. Conf. Comp. Vis, volume 3, 2017. 2