The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions

Peng Wang*[1,2], Qi Wu*[3], Chunhua Shen[2,3], and Anton van den Hengel[2,3]

[1]Northwestern Polytechnical University, China
[2]The University of Adelaide, Australia
[3]Australian Centre for Robotic Vision

Abstract

One of the most intriguing features of the Visual Question Answering (VQA) challenge is the unpredictability of the questions. Extracting the information required to answer them demands a variety of image operations from detection and counting, to segmentation and reconstruction. To train a method to perform even one of these operations accurately from {image,question,answer} tuples would be challenging, but to aim to achieve them all with a limited set of such training data seems ambitious at best. We propose here instead a more general and scalable approach which exploits the fact that very good methods to achieve these operations already exist, and thus do not need to be trained. Our method thus learns how to exploit a set of external off-the-shelf algorithms to achieve its goal, an approach that has something in common with the Neural Turing Machine [10]. The core of our proposed method is a new co-attention model. In addition, the proposed approach generates human-readable reasons for its decision, and can still be trained end-to-end without ground truth reasons being given. We demonstrate the effectiveness on two publicly available datasets, Visual Genome and VQA, and show that it produces the state-of-the-art results in both cases.

1. Introduction

Visual Question Answering (VQA) is an AI-complete task lying at the intersection of computer vision (CV) and natural language processing (NLP). Current VQA approaches are predominantly based on a joint embedding [3, 9, 18, 22, 34, 40] of image features and question representations into the same space, the result of which is used to predict the answer. One of the advantages of this approach is its ability to exploit a pre-trained CNN model.

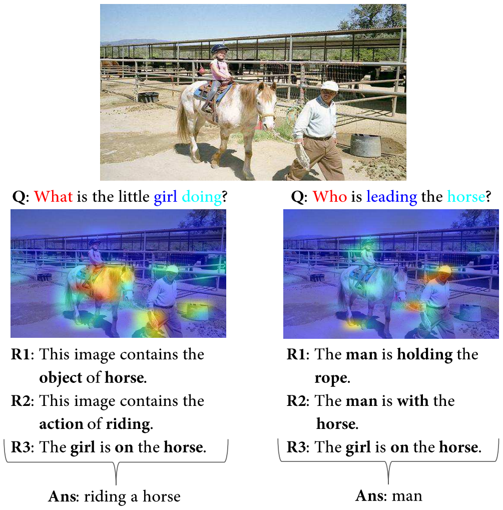

Q: What is the little girl doing?
Q: Who is leading the horse?
R1: This image contains the object of horse.
R2: This image contains the action of riding.
R3: The girl is on the horse.
Ans: riding a horse
R1: The man is holding the rope.
R2: The man is with the horse.
R3: The girl is on the horse.
Ans: man

Figure 1: Two real example results of our proposed model. Given an image-question pair, our model generates not only an answer, but also a set of reasons (as text) and visual attention maps. The colored words in the question have Top-3 weights, ordered as red, blue and cyan. The highlighted area in the attention map indicates the attention weights on the image regions. The Top-3 weighted visual facts are re-formulated as human readable reasons. The comparison between the results for different questions relating to the same image shows that our model can produce highly informative reasons relating to the specifics of each question.

In contrast to this joint embedding approach we propose a co-attention based method which learns how to use a set of off-the-shelf CV methods in answering image-based questions. Applying the existing CV methods to the image generates a variety of information which we label the image facts. Inevitably, much of this information would not be relevant to the particular question asked. So part of the role of the attention mechanism is to determine which types of facts are useful in answering a question.

The fact that the VQA-Machine is able to exploit a set of off-the-shelf CV methods in answering a question means that it does not need to learn how to perform these functions itself. The method instead learns to predict the appropriate

*The first two authors contributed to this work equally.

combination of algorithms to exploit in response to a previously unseen question and image. It thus represents a step towards a Neural Network capable of learning an *algorithm* for solving its problem. In this sense it is comparable to the Neural Turning Machine [10] (NTM) whereby an RNN is trained to use an associative memory module in solving its larger task. The method that we propose does not alter the parameters of the external modules it uses, but then it is able to exploit a much wider variety of module types.

In order to enable the VQA-Machine to exploit a wide variety of available CV methods, and to provide a compact, but flexible interface, we formulate the visual facts as triplets. The advantages of this approach are threefold. Firstly, many relevant off-the-shelf CV methods produce outputs that can be reformulated as triplets (see Tab.1). Secondly, such compact formats are human readable, and interpretable. This allows us to provide human readable reasons along with the answers. See Fig. 1 for an example. At last, the proposed triplet representation is similar to the triplet representations used in some Knowledge Bases, such as $<$ *cat, eat, fish* $>$. The method might thus be extendable to accept information from these sources.

To select the facts which are relevant in answering a specific question, we employ a co-attention mechanism. This is achieved by extending the approach of Lu *et al.* [16], which proposed a co-attention mechanism that jointly reasons about the image and question, to also reason over a set of facts. Specifically, we design a sequential co-attention mechanism (see Fig.3) which aims to ensure that attention can be passed effectively between all three forms of data. The initial question representation (without attention) is thus first used to guide facts weighting. The weighted facts and the initial question representation are then combined to guide the image weighting. The weighted facts and image regions are then jointly used to guide the question attention mechanism. All that remains is to run the fact attention again, but informed by the question and image attention weights, as this completes the circle, and means that each attention process has access to the output of all others. All of the weighted features are further fed into a multi-layer perceptron (MLP) to predict the answer.

One of the advantages of this approach is that the question, image, and facts are interpreted together, which particularly means that information extracted from the image (and represented as facts) can guide the question interpretation. A question such as 'Who is looking at the man with a telescope?' means different things when combined with an image of a man holding a telescope, rather than an image of a man viewed through a telescope.

Our main contributions are as follows:

- We propose a new VQA model which is able to learn to adaptively combine multiple off-the-shelf CV methods to answer questions.
- To achieve that, we extend the co-attention mechanism to a higher order which is able to jointly process questions, image, and facts.
- The method that we propose generates not only an answer to the posed questions, but also a set of supporting information, including the visual (attention) reasoning and human-readable textual reasons. To the best of our knowledge, this is the first VQA model that is capable of outputting human-readable reasons on free-form open-ended visual questions.
- Finally, we evaluate our proposed model on two VQA datasets. Our model achieves the state-of-art in both cases. A human agreement study is conducted to evaluate the reason generation ability of our model.

2. Related Work

Joint Embedding Most recent methods are based on a joint embedding of the image and the question using a deep neural network. Practically, image representations are obtained through pre-trained CNNs. The question is typically passed through an RNN, which' produces a fixed-length vector representation. These two representations are jointly embedded into the same space and fed into a classifier which predicts the final answer. Many previous works [3, 9, 18, 22, 32, 40] adopted this approach, while some [8, 13, 23] have proposed modifications of this basic idea. However, these modifications are either focused on developing more advanced embedding techniques or employing different question encoding methods, with only very few methods actually aim to improve the visual information available [20, 29, 34]. This seems a surprising situation given that VQA can be seen as encompassing the vast majority of CV tasks (by phrasing the task as a question, for instance). It is hard to imagine that a single pre-trained CNN model (such as VGG [25] or ResNet [11]) would suffice for all such tasks, or be able to recover all of the required visual information. The approach that we propose here, in contrast, is able to exploit the wealth of methods that already exist to extract useful information from images, and thus does not need to learn to perform these operations from a dataset that is ill suited to the task.

Attention Mechanisms Instead of directly using the holistic global-image embedding from the fully connected layer of a CNN, several recent works [12, 16, 24, 37, 38, 41] have explored image attention models for VQA. Specifically, the feature map (normally the convolutional layer of a pre-trained deep CNN) is used with the question to determine spatial weights that reflect the most relevant regions of the image. Recently, Lu *et al.* [16] determine attention weights on both image regions and question words. In our work, we extend the co-attention to a higher order so that the image, question and facts can be jointly weighted.

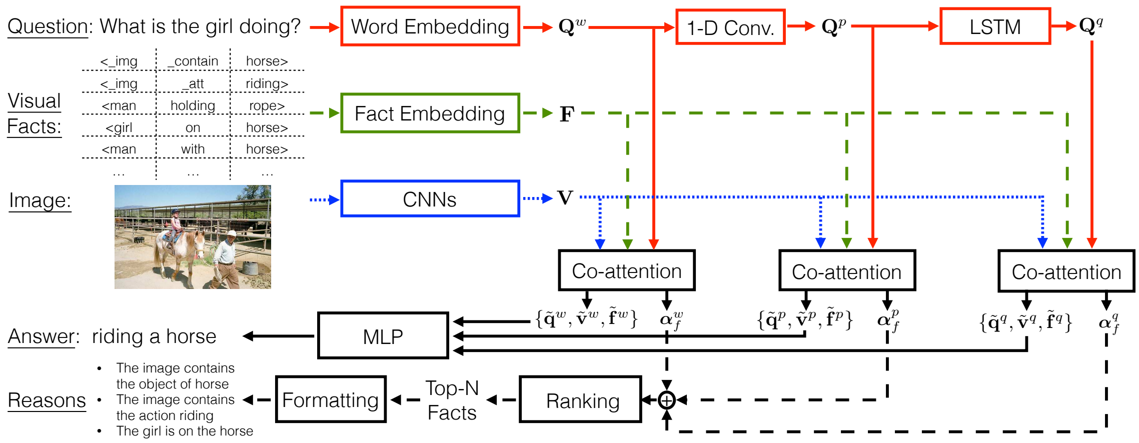

Question: What is the girl doing?
Word Embedding
$\mathbf{Q}^w$
1-D Conv.
$\mathbf{Q}^p$
LSTM
$\mathbf{Q}^q$
Visual Facts:
<_img _contain horse>
<_img _att riding>
<man holding rope>
<girl on horse>
<man with horse>
... ...
Fact Embedding
$\mathbf{F}$
Image:
CNNs
$\mathbf{V}$
Co-attention
Co-attention
Co-attention
$\{\tilde{\mathbf{q}}^w, \tilde{\mathbf{v}}^w, \tilde{\mathbf{f}}^w\}$
$\boldsymbol{\alpha}_f^w$
$\{\tilde{\mathbf{q}}^p, \tilde{\mathbf{v}}^p, \tilde{\mathbf{f}}^p\}$
$\boldsymbol{\alpha}_f^p$
$\{\tilde{\mathbf{q}}^q, \tilde{\mathbf{v}}^q, \tilde{\mathbf{f}}^q\}$
$\boldsymbol{\alpha}_f^q$
Answer: riding a horse
MLP
Reasons
• The image contains the object of horse
• The image contains the action riding
• The girl is on the horse
Formatting
Top-N Facts
Ranking
$\oplus$

Figure 2: The proposed VQA model. The input question, facts and image features are weighted at three question-encoding levels. Given the co-weighted features at all levels, a multi-layer perceptron (MLP) classifier is used to predict answers. Then the ranked facts are used to generate reasons.

Modular Architecture and Memory Networks Neural Module Networks (NMNs) were introduced by Andreas *et al.* in [1, 2]. In NMNs, the question parse tree is turned into an assembly of modules from a predefined set, which are then used to answer the question. Dynamic Memory Networks (DMN) [36] retrieves the 'facts' required to answer the question, where the 'facts' are simply CNN features calculated over small image patches. In comparison to NMNs and DMNs, our method uses a set of external algorithms that does not depend on the question. The set of algorithms used is larger, however, the method for combining their outputs is more flexible, and varies in response to the question, the image, and the facts.

Explicit Reasoning One of the limitations of most VQA methods is that it impossible to distinguish between an answer which has arisen as a result of the image content, and one selected because it occurs frequently in the training set [27]. This is a significant limitation to the practical application of the technology, particularly in Medicine or Defence, as it makes it impossible to have any faith in the answers provided. One solution to this problem is to provide human readable reasoning to justify or explain the answer, which has been a long-standing goal in Neural Networks (see [7, 28], for example). Wang *et al.* [30] propose a VQA framework named "Ahab" that uses explicit reasoning over an RDF (Resource Description Framework) Knowledge Base to derive the answer, which naturally gives rise to a reasoning chain. This approach is limited to a handcrafted set of question templates, however. FVQA [31] used an LSTM and a data-driven approach to learn the mapping of images/questions to RDF queries, but only considers questions relating to specified Knowledge Bases. In this work, we employ attention mechanisms over facts provided by multiple off-the-shelf CV methods. The facts are formulated as human understandable structural triplets and are further processed into human readable reasons. To the best of our knowledge, this is the first approach that can provide human readable reasons for the open-ended VQA problem. A human agreement study is reported in Sec. 4.2.3 which demonstrates the performance of this approach.

3. Models

In this section, we introduce the proposed VQA model that takes questions, images and facts as inputs and outputs a predicted answer with ranked reasons. The overall framework is described in Sec. 3.1, while Sec. 3.2 demonstrates how the three types of input are jointly embedded using the proposed sequential co-attention model. Finally, the module used for generating answers and reasons is introduced in Sec. 3.3.

Notation In the following, matrices are represented by bold capital letters and column vectors are denoted by bold lower-case letters. $[\mathbf{a}; \mathbf{b}]$ vertically concatenates the vectors $\mathbf{a}$ and $\mathbf{b}$, while $[\mathbf{a},\mathbf{b}]$ stacks $\mathbf{a}$ and $\mathbf{b}$ horizontally. We omit bias terms in neural networks.

3.1. Overall Framework

The entire model is shown in the Fig. 2. The first step sees the input question encoded at three different levels. At each level, the question features are embedded jointly with images and facts via the proposed sequential co-attention model. Finally, a multi-layer perceptron (MLP) is used to predict answers based on the outputs (*i.e.*, the weighted question, image and fact features) of the co-attention models at all levels. Reasons are generated by ranking and reformulating the weighted facts.

Hierarchical Question Encoding We apply a hierarchical question encoding [16] to effectively capture the information from a question at multiple scales, *i.e.* word, phase and sentence level. Firstly, the one-hot vectors of question words $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_T]$ are embedded individually to continuous vectors $\mathbf{Q}^w = [\mathbf{q}_1^w, \dots, \mathbf{q}_T^w]$, using a linear transfor-

Triplet	Example
`(_img,_scene,img_scn)`	`(_img,_scene,office)`
`(_img,_att,img_att)`	`(_img,_att,wedding)`
`(_img,_contain,obj)`	`(_img,_contain,dog)`
`(obj,_att,obj_att)`	`(shirt,_att,red)`
`(obj1,rel,obj2)`	`(man,hold,umbrella)`

Table 1: Facts represented by triplets. `_img`, `_scene`, `_att` and `_contain` are specific tokens. While `img_scn`, `obj`, `img_att`, `obj_att` and `rel` refer to vocabularies describing image scenes, objects, image/object attributes and relationships between objects.

mation followed by a tanh function. Then 1-D convolutions with different filter sizes (unigram, bigram and trigram) are applied to the word-level embeddings $\mathbf{Q}^w$, followed by a max-pooling over different filters at each word location, to form the phrase-level features $\mathbf{Q}^p = [\mathbf{q}_1^p, \ldots, \mathbf{q}_T^p]$. Finally, the phrase-level features are further encoded by an LSTM, resulting in the question-level features $\mathbf{Q}^q = [\mathbf{q}_1^q, \ldots, \mathbf{q}_T^q]$.

Encoding Image Regions Following [16, 38], the input image is resized to 448×448 and divided to 14×14 regions. The corresponding regions of the last pooling layer of VGG-19 [25] or ResNet-100 [25] networks are extracted and further embedded using an end-to-end learned linear transformation followed by a tanh function. The output is $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$ ($N = 196$), are taken as image features.

Encoding Facts In this work, we use triplets of the form `(subject,relation,object)` to represent facts in an image, where `subject` and `object` denote two visual concepts and `relation` represents a relationship between these two concepts. This format of triplets is very general and widely used in large-scale structured knowledge graphs (such as DBpedia [4], Freebase [5], YAGO [17]) to record a surprising variety of information. In this work, we consider 5 types of visual facts as shown in Table 1, which respectively records information about the scene, objects, object attributes, image attributes, and relationships between two objects. A fact (triplet) is encoded as $\mathbf{f} = \tanh([\mathbf{W}_s\mathbf{e}_s; \mathbf{W}_r\mathbf{e}_r; \mathbf{W}_o\mathbf{e}_o])$, where $\mathbf{e}_s$, $\mathbf{e}_r$ and $\mathbf{e}_o$ are one-hot vectors representing `subject`, `relation` or `object` respectively. $\mathbf{W}_s$, $\mathbf{W}_r$ and $\mathbf{W}_o$ are linear transformation weights to be learned. By applying different types of visual models, we achieve a list of encoded fact features $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_M]$, where M is the number of facts. Note that this approach may be easily extended to using any existing vision methods to extract image information that might usefully be recorded as a triplet, or even more generally, an n-tuple that includes additional information, such as confidence score and data source.

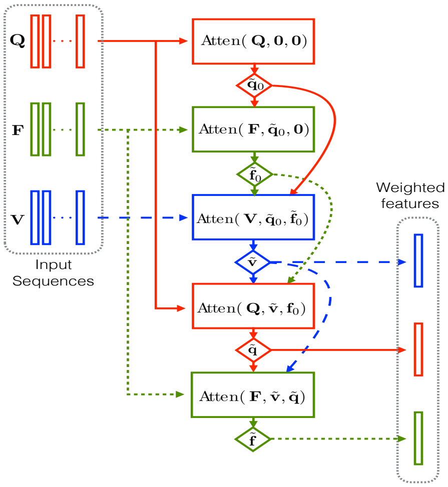

Q
F
V
Input Sequences
Atten($\mathbf{Q}, \mathbf{0}, \mathbf{0}$)
$\tilde{\mathbf{q}}_0$
Atten($\mathbf{F}, \tilde{\mathbf{q}}_0, \mathbf{0}$)
$\tilde{\mathbf{f}}_0$
Atten($\mathbf{V}, \tilde{\mathbf{q}}_0, \tilde{\mathbf{f}}_0$)
$\tilde{\mathbf{v}}$
Atten($\mathbf{Q}, \tilde{\mathbf{v}}, \tilde{\mathbf{f}}_0$)
$\tilde{\mathbf{q}}$
Atten($\mathbf{F}, \tilde{\mathbf{v}}, \tilde{\mathbf{q}}$)
$\tilde{\mathbf{f}}$
Weighted features

Figure 3: The sequential co-attention module. Given the feature sequences for the question ($\mathbf{Q}$), facts ($\mathbf{F}$) and image ($\mathbf{V}$), this module sequentially generates weighted features ($\tilde{\mathbf{v}}$, $\tilde{\mathbf{q}}$, $\tilde{\mathbf{f}}$).

3.2. Sequential Co-attention

To accommodate multiple sources of information (question, image, fact), the proposed co-attention approach sequentially generates attention weights for each feature type using others as guidance, as shown in Fig 3. The operation in each of the five attention modules is denoted $\tilde{\mathbf{x}} = \text{Atten}(\mathbf{X}, \mathbf{g}_1, \mathbf{g}_2)$, which can be expressed as follows:

$$\mathbf{H}_i = \tanh(\mathbf{W}_x\mathbf{x}_i + \mathbf{W}_{g_1}\mathbf{g}_1 + \mathbf{W}_{g_2}\mathbf{g}_2), \tag{1a}$$

$$\alpha_i = \text{softmax}(\mathbf{w}^\top \mathbf{H}_i), \quad i = 1, \ldots, N, \tag{1b}$$

$$\tilde{\mathbf{x}} = \textstyle\sum_{i=1}^{N} \alpha_i \mathbf{x}_i, \tag{1c}$$

where $\mathbf{X}$ is the input sequence (*i.e.*, $\mathbf{Q}$, $\mathbf{F}$ or $\mathbf{V}$), and $\mathbf{g}_1$, $\mathbf{g}_2 \in \mathbb{R}^d$ represent guidances that are outputs of previous attention modules. $\mathbf{W}_x$, $\mathbf{W}_{g_1}$, $\mathbf{W}_{g_2} \in \mathbb{R}^{h \times d}$ and $\mathbf{w} \in \mathbb{R}^h$ are linear embedding parameters to be learned. Here h denotes the size of hidden layers of the attention module. In this work, all the question, image and fact features are embedded to d-dimensional vectors, *i.e.*, $\mathbf{Q} \in \mathbb{R}^{d \times T}$, $\mathbf{V} \in \mathbb{R}^{d \times N}$, $\mathbf{F} \in \mathbb{R}^{d \times M}$. $\boldsymbol{\alpha}$ is the attention weights of the input sequence and $\tilde{\mathbf{x}}$ is the weighted sum of features.

In the proposed co-attention approach, the encoded question/image/fact features (see Sec. 3.1) are sequentially fed into the attention module (Eqn. 1) as input sequences, and the weighted features from the previous two steps are used as guidance. Firstly, the question features are summarized without any guidance ($\tilde{\mathbf{q}}_0 = \text{Atten}(\mathbf{Q}, \mathbf{0}, \mathbf{0})$). At the second step, the fact features are weighted based on the summarized question features ($\tilde{\mathbf{f}}_0 = \text{Atten}(\mathbf{F}, \tilde{\mathbf{q}}_0, \mathbf{0})$). Next, the weighted image features are generated with the weighted fact features and summarized question features as guidances ($\tilde{\mathbf{v}} = \text{Atten}(\mathbf{V}, \tilde{\mathbf{q}}_0, \tilde{\mathbf{f}}_0)$). In step 4 ($\tilde{\mathbf{q}} = \text{Atten}(\mathbf{Q}, \tilde{\mathbf{v}}, \tilde{\mathbf{f}}_0)$) and step 5 ($\tilde{\mathbf{f}} = \text{Atten}(\mathbf{F}, \tilde{\mathbf{v}}, \tilde{\mathbf{q}})$), the question and fact features are re-weighted based on the outputs of the previous steps. Finally, the weighted question/image/fact features ($\tilde{\mathbf{q}}$, $\tilde{\mathbf{f}}$, $\tilde{\mathbf{v}}$) are further used for answer prediction and the attention weights of the last attention module $\boldsymbol{\alpha}_f$ are used for reasons generation.

3.3. Answer Prediction and Reason Generation

Similar to many previous VQA models [16, 21, 22, 40], the answer prediction process is treated as a multi-class classification problem, in which each class corresponds to a distinct answer. Given the weighted features generated from the word/phrase/question levels, a multi-layer perceptron (MLP) is used for classification:

$$\mathbf{h}^w = \tanh\big(\mathbf{W}_w(\tilde{\mathbf{q}}^w + \tilde{\mathbf{v}}^w + \tilde{\mathbf{f}}^w)\big), \quad (2a)$$

$$\mathbf{h}^p = \tanh\big(\mathbf{W}_p\big[(\tilde{\mathbf{q}}^p + \tilde{\mathbf{v}}^p + \tilde{\mathbf{f}}^p); \mathbf{h}^w\big]\big), \quad (2b)$$

$$\mathbf{h}^q = \tanh\big(\mathbf{W}_q\big[(\tilde{\mathbf{q}}^q + \tilde{\mathbf{v}}^q + \tilde{\mathbf{f}}^q); \mathbf{h}^p\big]\big), \quad (2c)$$

$$\mathbf{p} = \mathrm{softmax}(\mathbf{W}_h \mathbf{h}^q), \quad (2d)$$

where $\{\tilde{\mathbf{q}}^w, \tilde{\mathbf{v}}^w, \tilde{\mathbf{f}}^w\}$, $\{\tilde{\mathbf{q}}^p, \tilde{\mathbf{v}}^p, \tilde{\mathbf{f}}^p\}$ and $\{\tilde{\mathbf{q}}^q, \tilde{\mathbf{v}}^q, \tilde{\mathbf{f}}^q\}$ are weighted features from all three levels. $\mathbf{W}_w$, $\mathbf{W}_p$, $\mathbf{W}_q$ and $\mathbf{W}_h$ are parameters and $\mathbf{p}$ is the probability vector.

As shown in Fig. 2, the attention weights of facts from all three levels ($\boldsymbol{\alpha}_f^w$, $\boldsymbol{\alpha}_f^p$, $\boldsymbol{\alpha}_f^q$) are summed together. Then the top-3 ranked facts are automatically formulated (by simple rule-based approaches, see supplementary) to human readable sentences and are considered as reasons.

4. Experiments

We evaluate our models on two datasets, Visual Genome QA [14] and VQA-real [3]. The Visual Genome QA contains 1,445,322 questions on 108,077 images. Since the official split of the Visual Genome dataset is not released, we randomly generate our own split. In this split, we have 723,060 training, 54,506 validation and 667,753 testing question/answer examples, based on 54,038 training, 4,039 validation and 50,000 testing images. VQA dataset [3] is one of the most widely used datasets, which comprises two parts, one using natural images, and a second using cartoon images. In this paper, we only evaluate our models on the real image subset, which we have labeled VQA-real. VQA-real comprises 123,287 training and 81,434 test images.

4.1. Implementation Details

Facts Extraction Using the training splits of the Visual Genome dataset, we trained three naive multi-label CNN models to extract objects, object attributes and object-object relations. Specifically, we extract triplets of the form `(_img,_contain,obj)`/`(obj,_att,obj_att)`/ `(obj1,rel,obj2)` based on the annotations of Visual Genome and rank them individually according to frequency. The top-5000/10000/15000 triplets are selected respectively and used as class labels. We then formulate them as three separate multi-label classification problems using an element-wise logistic loss function. The pretrained VGG-16 model is used as initialization and only the fully-connected layers are fine-tuned. We also used the scene classification model of [39] and the image attribute model of [32] to extract facts not included in the Visual Genome dataset (*i.e.*, image scenes and image attributes). Thresholds for these visual models' softmax scores are set to 0.4 and on average 76 facts (ranging from 9 to 150) are extracted for each image. The score of predicted fact is added as an additional dimension of fact features $\mathbf{f}$.

Training Parameters In our system, the dimension d of encoded question/image/fact features and the hidden layer size h of co-attention models (see Eqn. 1) are both set to 512. For facts, the `subject`/`relation`/`object` entities are embedded to 128/128/256 dimensional vectors respectively and concatenated to form 512-d vectors. We used two layers of LSTM model with the hidden size of 512. For the MLP in Eq. (2), the dimensions of $\mathbf{h}^w$ and $\mathbf{h}^p$ are also 512, while the dimension of $\mathbf{h}^q$ is set to 1024 for the VQA dataset and 2048 for the Visual Genome dataset. For prediction, we take the top 3000 answers for the VQA dataset and the top 5000 answers for the Visual Genome dataset. The whole system is implemented on the Torch7 [6] and trained end-to-end but with fixed CNN features. For optimization, the `RMSProp` method is used with a base learning rate of 2×10^{-4} and momentum 0.99. The model is trained for up to 256 epochs until the validation error has not improved in the last 5 epochs.

4.2. Results on the Visual Genome QA

Metrics We use the accuracy value and the Wu-Palmer similarity (WUPS) [35] to measure the performance on the Visual Genome QA (see Table 2). Before comparison, all responses are made lowercase, numbers converted to digits, and punctuation & articles removed. The accuracy according to the question types are also reported.

Baselines and State-of-the-art The first baseline method is **VGG+LSTM** from Antol *et al.* in [3], who uses a two layer LSTM to encode the questions and the last hidden layer of VGG [25] to encode the images. The image features are then $\ell 2$ normalized. We use the author provided code[1] to train the model on the Visual Genome QA training split. **VGG+Obj+Att+Rel+Extra+LSTM** uses the same configuration as VGG+LSTM, except that we concatenate additional image features extracted by VGG-16 models (fc7) that have been pre-trained on different CV tasks described in the previous section. **HieCoAtt-VGG** is the original model presented in [16], which is the current state of art. The authors' implementation[2] is used to train the model.

4.2.1 Ablation Studies with Ground Truth Facts

In this section, we conduct an ablation study to evaluate the effectiveness of incorporating different types of facts. To avoid the bias that would be caused by the varying accuracy with which the facts are predicted, we use the ground truth

[1] https://github.com/VT-vision-lab/VQA_LSTM_CNN

[2] https://github.com/jiasenlu/HieCoAttenVQA

Methods	Accuracy (%)							WUPS (%)	
	What	Where	When	Who	Why	How	**Overall**	Overall	
	(60.5%)	(17.0%)	(3.5%)	(5.5%)	(2.7%)	(10.8%)		@0.9	@0.0
VGG+LSTM [3]	35.12	16.33	52.71	30.03	11.55	42.69	32.46	38.30	58.39
VGG+Obj+Att+Rel+Extra+LSTM	36.88	16.85	52.74	32.30	11.65	44.00	33.88	39.61	58.89
HieCoAtt-VGG [16]	39.72	17.53	52.53	33.80	12.62	45.14	35.94	41.75	59.97
Ours-GtFact(Obj)	37.82	17.73	51.48	37.32	12.84	43.10	34.77	40.83	59.69
Ours-GtFact(Obj+Att)	42.21	17.56	51.89	37.45	12.93	43.90	37.50	43.24	60.39
Ours-GtFact(Obj+Rel)	38.25	18.10	51.13	38.22	12.86	43.32	35.15	41.25	59.91
Ours-GtFact(Obj+Att+Rel)	42.86	18.22	51.06	38.26	13.02	44.26	38.06	43.86	60.72
Ours-GtFact(Obj+Att+Rel)+VGG	44.28	18.87	52.06	38.87	12.93	46.08	39.30	44.94	61.21
Ours-PredFact(Obj+Att+Rel)	37.13	16.99	51.70	33.87	12.73	42.87	34.01	39.92	59.20
Ours-PredFact(Obj+Att+Rel+Extra)	38.52	17.86	51.55	34.65	12.87	44.34	35.20	41.08	59.75
Ours-PredFact(Obj+Att+Rel)+VGG	40.34	17.80	52.12	34.98	12.78	45.37	36.44	42.16	60.09
Ours-PredFact(Obj+Att+Rel+Extra)+VGG	40.91	18.33	52.33	35.50	12.88	46.04	36.99	42.73	60.39

Table 2: Ablation study on the Visual Genome QA dataset. Accuracy for different question types are shown. The percentage of questions for each type is shown in parentheses. We additionally calculate the WUPS at 0.9 and 0.0 for different models.

facts provided by the Visual Genome dataset as the inputs to our proposed models for ablation testing.

GtFact(Obj) is the initial implementation of our proposed co-attention model, using only the 'object' facts, which means the co-attention mechanisms only occur between question and the input object facts. The overall accuracy of this model on the test split is 34.77%, which already outperforms the baseline model **VGG+LSTM**. However, there is still a gap between this model and the **HieCoAtt-VGG** (35.94%), which applies the co-attention mechanism to the questions and whole image features. Considering that the image features are still not used at this stage, this result is reasonable.

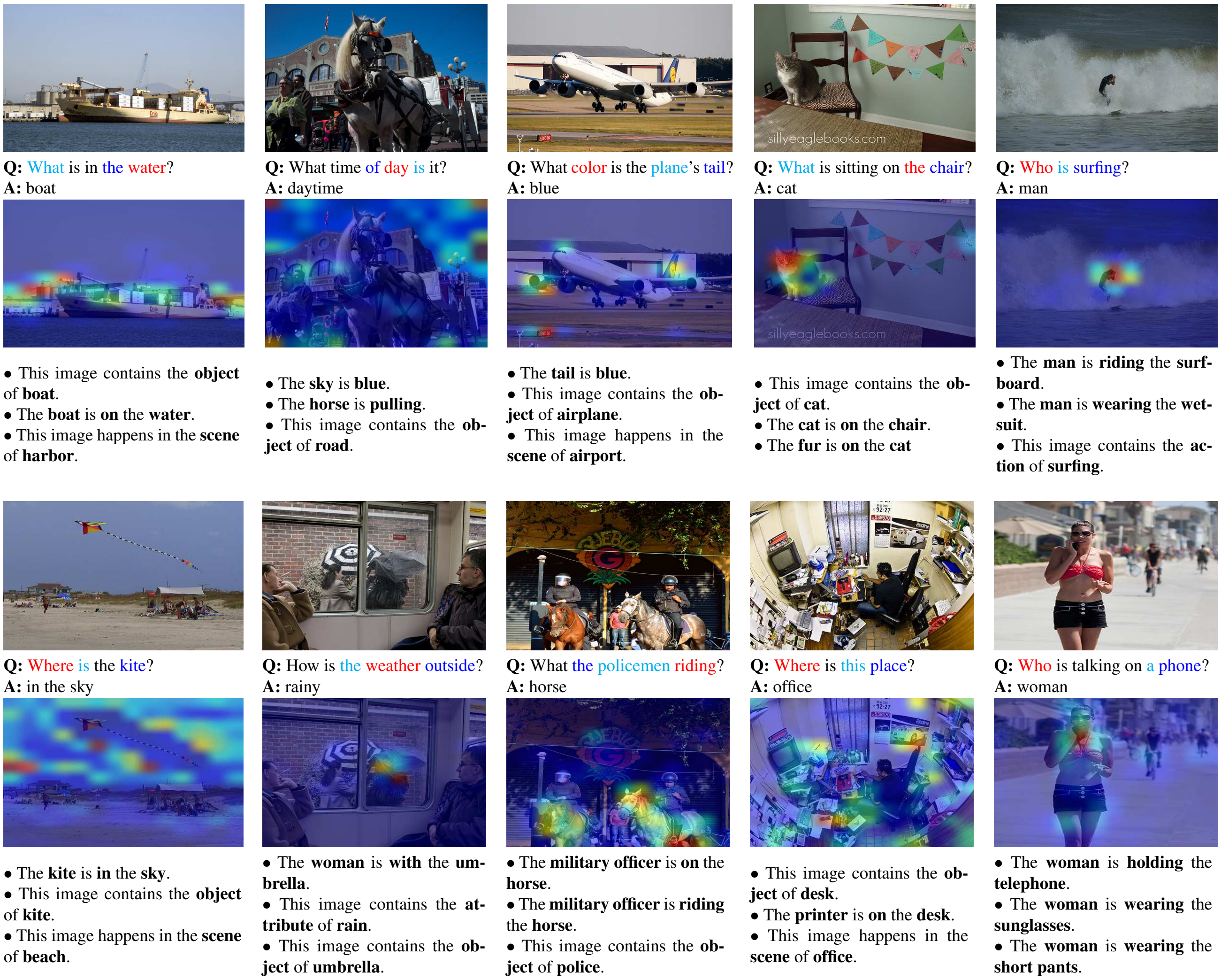
Based on this initial model, we add the 'attribute' and 'relationship' facts separately, producing two models, **GtFact(Obj+Att)** and **GtFact(Obj+Rel)**. Table 2 shows that the former performs better than the latter (37.50% VS. 35.15%), although both outperform the previous 'Object'-only model. This suggests that 'attribute' facts are more effective than the 'relationship' facts. However, when it comes to questions starting with 'where' and 'who', the **GtFact(Obj+Rel)** performs slightly better (18.10% VS. 17.56%, 38.22% VS.37.45%), which suggests that the 'relationship' facts play a more important role in these types of question. This makes sense because 'relationship' facts naturally encode location and identity information, for example $< cat, on, mat >$ and $< boy, holding, ball >$. The **GtFact(Obj+Att)** model exceeds the image features based co-attention model **HieCoAtt-VGG** by a large margin, 1.56%. This is mainly caused by the substantial improvement in the 'what' questions, from 39.72% to 42.21%. This is not surprising because the 'attributes' facts include detailed information relevant to the 'what' questions, such as 'color', 'shape', 'size', 'material' and so on.

In the **GtFact(Obj+Att+Rel)**, all of the facts are plugged into the proposed model, which brings further improvements for nearly all question types, achieving an overall accuracy 38.06%. Compared with the baseline model **VGG+Obj+Att+Rel+Extra+LSTM** that uses a conventional method (concatenating and embedding) to introduce additional features into the VQA, our model outperforms by 4.18%, which is a large gap. However, for the 'when' questions, we find that the performance of our 'facts' based models are always lower than the image-based ones. We observed that this mainly because the annotated facts in the Visual Genome normally do not cover the 'when' related information, such as time and day or night. The survey paper [33] makes a similar observation - *"98.8% of the 'when' questions in the Visual Genome can not be directly answered from the provided scene graph annotations"*. Hence, in the final model **GtFact(Obj+Att+Rel)+VGG**, we add the image features back, allowing for a higher order co-attention between image, question and facts, achieving an overall accuracy 39.30%. It also brings 1% improvements on the 'when' questions. The WUPS evaluation exhibits the same trend as the above results, the question type-specific results can be found in the supplementary material.

4.2.2 Evaluation with Predicted Facts

In a normal VQA setting the ground truth facts are not provided, so we now evaluate our model using predicted facts. All facts were predicted by models that have been pre-trained on different computer vision tasks, *e.g.* object detection, attributes prediction, relationship prediction and scene classification (see Sec. 4.1 for more details).

The **PredFact(Obj+Att+Rel)** model uses all of the predicted facts as the input, while **PredFact(Obj+Att+Rel+Extra)** additionally uses the predicted scene category, which is trained on a different data source, the MIT Places 205 [39]. From Table 2, we see that although all facts have been included, the performance of these two facts-only models is still lower than that of the previous state-of-the-art model **HieCoAtt-VGG**. Considering that errors in fact prediction may pass to the question answering part and the image features are not used, these results are reasonable. If the fact prediction



Q: What is in the water?
A: boat

- This image contains the **object** of **boat**.
- The **boat** is **on** the **water**.
- This image happens in the **scene** of **harbor**.

Q: What time of day is it?
A: daytime

- The **sky** is **blue**.
- The **horse** is **pulling**.
- This image contains the **object** of **road**.

Q: What color is the plane's tail?
A: blue

- The **tail** is **blue**.
- This image contains the **object** of **airplane**.
- This image happens in the **scene** of **airport**.

Q: What is sitting on the chair?
A: cat

- This image contains the **object** of **cat**.
- The **cat** is **on** the **chair**.
- The **fur** is **on** the **cat**

Q: Who is surfing?
A: man

- The **man** is **riding** the **surfboard**.
- The **man** is **wearing** the **wetsuit**.
- This image contains the **action** of **surfing**.

Q: Where is the kite?
A: in the sky

- The **kite** is **in** the **sky**.
- This image contains the **object** of **kite**.
- This image happens in the **scene** of **beach**.

Q: How is the weather outside?
A: rainy

- The **woman** is **with** the **umbrella**.
- This image contains the **attribute** of **rain**.
- This image contains the **object** of **umbrella**.

Q: What the policemen riding?
A: horse

- The **military officer** is **on** the **horse**.
- The **military officer** is **riding** the **horse**.
- This image contains the **object** of **police**.

Q: Where is this place?
A: office

- This image contains the **object** of **desk**.
- The **printer** is **on** the **desk**.
- This image happens in the **scene** of **office**.

Q: Who is talking on a phone?
A: woman

- The **woman** is **holding** the **telephone**.
- The **woman** is **wearing** the **sunglasses**.
- The **woman** is **wearing** the **short pants**.

Figure 4: Some qualitative results produced by our complete model on the Visual Genome QA test split. Image, QA pair, attention map and predicted Top-3 reasons are shown in order. Our model is capable of co-attending between question, image and supporting facts. The colored words in the question have Top-3 identified weights, ordered as red, blue and cyan. The highlighted area in the attention map indicates the attention weights on the image regions (from red: high to blue: low). The Top-3 identified facts are re-formulated as human readable reasons, shown as bullets.

models perform better, the final answer accuracy will be higher, as shown in the previous ground-truth facts experiments. The **PredFact(Obj+Att+Rel+Extra)** outperforms the **PredFact(Obj+Att+Rel)** model by 1.2%, because the extra scene category facts are included. This suggests that our model benefits by using multiple off-the-shelf CV methods, even though they are trained on different data sources, and on different tasks. And as more facts are added, our model performs better. Compared with the baseline model **VGG+Obj+Att+Rel+Extra+LSTM**, our **PredFact(Obj+Att+Rel+Extra)** outperforms it by a large margin, which suggests that our co-attention model can more effectively exploit the fact-based information.

Image features are further inputted to our model, producing two variants **PredFact(Obj+Att+Rel)+VGG** and **PredFact(Obj+Att+Rel+Extra)+VGG**. Both of these outperform the state of art, and our complete model **PredFact(Obj+Att+Rel+Extra)+VGG** outperforms it by 1%, *i.e.* more than 6,000 more questions are correctly answered.

Please note that all of these results are produced by using the naive facts extraction models described in Sec. 4.1. We believe that as better facts extraction models (such as the relationship prediction models from [15], for instance) become available, the results will improve further.

4.2.3 Human agreements on Predicted Reasons

A key differentiator of our model is that the attention weights on facts can be used to provide human-interpretable reasons for generated answers. We sampled 1,000 questions that have been correctly answered by our **PredFact(Obj+Att+Rel+Extra)+VGG** model, and conduct a human agreement study on the generated reasons.

Method	Test-**dev**								Test-**std**							
	Open-Ended				Multiple-Choice				Open-Ended				Multiple-Choice			
	Y/N	Num	Other	All	Y/N	Num	Other	All	Y/N	Num	Other	All	Y/N	Num	Other	All
iBOWING [40]	76.6	35.0	42.6	55.7	76.7	37.1	54.4	61.7	76.8	35.0	42.6	55.9	76.9	37.3	54.6	62.0
MCB-VGG [8]	-	-	-	57.1	-	-	-	-	-	-	-	-	-	-	-	-
DPPnet [21]	80.7	37.2	41.7	57.2	80.8	38.9	52.2	62.5	80.3	36.9	42.2	57.4	80.4	38.8	52.8	62.7
D-NMN [1]	80.5	37.4	43.1	57.9	-	-	-	-	-	-	-	58.0	-	-	-	-
VQA team [3]	80.5	36.8	43.1	57.8	80.5	38.2	53.0	62.7	80.6	36.4	43.7	58.2	80.6	37.7	53.6	63.1
SMem [37]	80.9	37.3	43.1	58.0	-	-	-	-	80.8	37.3	43.1	58.2	-	-	-	-
SAN [38]	79.3	36.6	46.1	58.7	-	-	-	-	-	-	-	58.9	-	-	-	-
ACK [34]	81.0	38.4	45.2	59.2	-	-	-	-	81.1	37.1	45.8	59.4	-	-	-	-
DMN+ [36]	80.5	36.8	48.3	60.3	-	-	-	-	-	-	-	60.4	-	-	-	-
MRN-VGG [13]	82.5	38.3	46.8	60.5	82.6	39.9	55.2	64.8	-	-	-	-	-	-	-	-
HieCoAtt-VGG [16]	79.6	38.4	49.1	60.5	79.7	40.1	57.6	64.9	-	-	-	-	-	-	-	-
Re-Ask-ResNet [19]	78.4	36.4	46.3	58.4	-	-	-	-	78.2	36.3	46.3	58.4	-	-	-	-
FDA-ResNet [12]	81.1	36.2	45.8	59.2	-	-	-	-	-	-	-	59.5	-	-	-	-
MRN-ResNet [13]	82.4	38.4	49.3	61.5	82.4	39.7	57.2	65.6	82.4	38.2	49.4	61.8	82.4	39.6	58.4	66.3
HieCoAtt-ResNet [16]	79.7	38.7	51.7	61.8	79.7	40.0	59.8	65.8	-	-	-	62.1	-	-	-	66.1
MCB-Att-ResNet [8]	82.5	37.6	55.6	64.7	-	-	-	69.1	-	-	-	-	-	-	-	-
Ours-VGG	81.2	37.7	50.5	61.7	81.3	39.9	60.5	66.8	81.3	36.7	50.9	61.9	81.4	39.0	60.8	67.0
Ours-ResNet	81.5	38.4	53.0	63.1	81.5	40.0	62.2	67.7	81.4	38.2	53.2	63.3	81.4	39.8	62.3	67.8

Table 3: Single model performance on the VQA-real test set in the open-ended and multiple-choice settings.

Since this is the first VQA model that can generate human readable reasons, there is no previous work that we can follow to perform the human evaluation. We have thus designed the following human agreement experimental protocols. At first, an image with the question and our correctly generated answer are given to a human subject. Then a list of human readable reasons (ranging from 20 to 40) are shown. These reasons are formulated from facts that are predicted from the facts extraction model introduced in the Sec.4.1. Although these reasons are all related to the image, not all of them are relevant to the particular question asked. The task of the human subjects is thus to *choose the reasons that are related to answering the question*. The human agreements are calculated by matching the human selected reasons with our model ranked reasons. In order to ease the human subjects' workload and to have an unique guide for them to select the 'reasons', we ask them to select only the top-1 reason that is related to the question answering. And they can choose nothing if they think none of the provided reasons are useful.

Finally, in the evaluation, we calculate the rate at which the human selected reason can be found in our generated top-1/3/5 reasons. We find that **30.1%** of the human selected top-1 reason can be matched with our model ranked top-1 reason. For the top-3 and top-5, the matching rate are **54.2%** and **70.9%**, respectively. This suggests that the reasons generated are both interpretable and informative. Figure 4 shows some example reasons generated by our model.

4.3. Results on the VQA-real

Table 3 compares our approach with state-of-the-art on the VQA-real dataset. Since we do not use any ensemble models, we only compare with the single models on the VQA test leader-board. The *test-dev* is normally used for validation while the *test-standard* is the default test data. The first section of Table 3 shows the state of art methods that use VGG features, except iBOWING [40], which uses the GoogLeNet features [26]. The second section gives the results of models that use ResNet [11] features. Two versions of our complete model are evaluated at the last section, using VGG and ResNet features, respectively.

Ours-VGG produces the best result on all of the splits, compared with models using the same VGG image encoding method. **Ours-ResNet** ranks the second amongst the single models using ResNet features on the test-dev split, but we achieve the state of the art results on the test-std, for both Open-Ended and Multiple-Choice questions. The best result on the test-dev with ResNet features is achieved by the Multimodal Compact Bilinear (MCB) pooling model with the visual attention [8]. We believe the MCB can be integrated within our proposed co-attention model, by replacing the linear embedding steps in Eqs. 1 and 2, but we leave it as a future work.

5. Conclusion

We have proposed a new approach which is capable of adaptively combining the outputs from other algorithms in order to solve a new, more complex problem. We have shown that the approach can be applied to the problem of Visual Question Answering, and that in doing so it achieves state of the art results. Visual Question Answering is a particularly interesting application of the approach, as in this case the new problem to be solved is not completely specified until run time. In retrospect, it seems strange to attempt to answer general questions about images without first providing access to readily available image information that might assist in the process. In developing our approach we proposed a co-attention method applicable to questions, image and facts jointly. We also showed that attention-weighted facts serve to illuminate why the method reached its conclusion, which is critical if such techniques are to be used in practice.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. In *Proc. Conf. of North American Chapter of Association for Computational Linguistics*, 2016.

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural Module Networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[5] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD International Conference on Management of Data*, pages 1247–1250. ACM, 2008.

[6] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *Proc. Advances in Neural Inf. Process. Syst. Workshop*, 2011.

[7] M. Craven and J. W. Shavlik. Using sampling and queries to extract rules from trained neural networks. In *Proc. Int. Conf. Mach. Learn.*, pages 37–45, 1994.

[8] 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 *Proc. Conf. Empirical Methods in Natural Language Processing*, 2016.

[9] 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 Answering. In *Proc. Advances in Neural Inf. Process. Syst.*, 2015.

[10] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

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

[13] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *Proc. Advances in Neural Inf. Process. Syst.*, 2016.

[14] 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.

[15] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *Proc. Eur. Conf. Comp. Vis.*, 2016.

[16] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Proc. Advances in Neural Inf. Process. Syst.*, 2016.

[17] F. Mahdisoltani, J. Biega, and F. Suchanek. YAGO3: A knowledge base from multilingual Wikipedias. In *CIDR*, 2015.

[18] M. Malinowski, M. Rohrbach, and M. Fritz. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.

[19] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A deep learning approach to visual question answering. *arXiv preprint arXiv:1605.02697*, 2016.

[20] A. Mallya and S. Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *Proc. Eur. Conf. Comp. Vis.*, 2016.

[21] H. Noh, P. H. Seo, and B. Han. Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[22] M. Ren, R. Kiros, and R. Zemel. Image Question Answering: A Visual Semantic Embedding Model and a New Dataset. In *Proc. Advances in Neural Inf. Process. Syst.*, 2015.

[23] K. Saito, A. Shin, Y. Ushiku, and T. Harada. Dualnet: Domain-invariant network for visual question answering. *arXiv preprint arXiv:1606.06108*, 2016.

[24] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Representations*, 2015.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015.

[27] D. Teney, L. Liu, and A. van den Hengel. Graph-structured representations for visual question answering. *CoRR*, abs/1609.05600, 2016.

[28] S. Thrun. Extracting rules from artificial neural networks with distributed representations. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 505–512, 1995.

[29] T. Tommasi, A. Mallya, B. Plummer, S. Lazebnik, A. C. Berg, and T. L. Berg. Solving visual madlibs with multiple cues. In *Proc. British Machine Vis. Conf.*, 2016.

[30] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015.

[31] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick. Fvqa: Fact-based visual question answering. *arXiv:1606.05433*, 2016.

[32] Q. Wu, C. Shen, A. v. d. Hengel, L. Liu, and A. Dick. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[33] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel. Visual question answering: A survey of methods and datasets. *arXiv preprint arXiv:1607.05910*, 2016.

[34] Q. Wu, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[35] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proc. Conf. Association for Computational Linguistics*, 1994.

[36] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *Proc. Int. Conf. Mach. Learn.*, 2016.

[37] H. Xu and K. Saenko. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In *Proc. Eur. Conf. Comp. Vis.*, 2016.

[38] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked Attention Networks for Image Question Answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Proc. Advances in Neural Inf. Process. Syst.*, 2014.

[40] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.

[41] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.