

# Ask Your Neurons: A Deep Learning Approach to Visual Question Answering

Mateusz Malinowski · Marcus Rohrbach · Mario Fritz

Received: date / Accepted: date

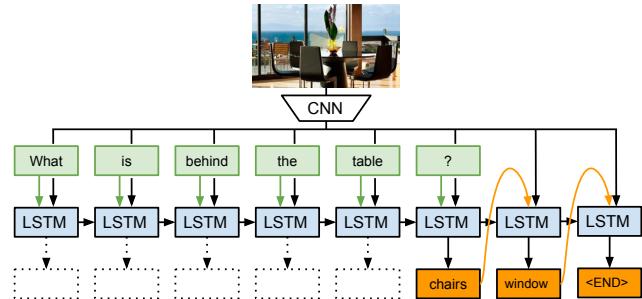
**Abstract** We propose a Deep Learning approach to the visual question answering task, where machines answer to questions about real-world images. By combining latest advances in image representation and natural language processing, we propose Ask Your Neurons, a scalable, jointly trained, end-to-end formulation to this problem. In contrast to previous efforts, we are facing a multi-modal problem where the language output (answer) is conditioned on visual and natural language inputs (image and question). We evaluate our approaches on the DAQUAR as well as the VQA dataset where we also report various baselines, including an analysis how much information is contained in the language part only. To study human consensus, we propose two novel metrics and collect additional answers which extend the original DAQUAR dataset to DAQUAR-Consensus. Finally, we evaluate a rich set of design choices how to encode, combine and decode information in our proposed Deep Learning formulation.

**Keywords** Computer Vision · Scene Understanding · Deep Learning · Natural Language Processing · Visual Turing Test · Visual Question Answering

Mateusz Malinowski  
Max Planck Institute for Informatics  
Saarbrücken, Germany  
E-mail: mmalinow@mpi-inf.mpg.de

Marcus Rohrbach  
UC Berkeley EECS and ICSI  
Berkeley, CA, United States  
E-mail: rohrbach@berkeley.edu

Mario Fritz  
Max Planck Institute for Informatics  
Saarbrücken, Germany  
E-mail: mfritz@mpi-inf.mpg.de



**Fig. 1** Our approach *Ask Your Neurons* to question answering with a Recurrent Neural Network using Long Short Term Memory (LSTM). To answer a question about an image, we feed in both, the image (CNN features) and the question (green boxes) into the LSTM. After the (variable length) question is encoded, we generate the answers (multiple words, orange boxes). During the answer generation phase the previously predicted answers are fed into the LSTM until the *<END>* symbol is predicted. See Section 3.1 for more details.

## 1 Introduction

With the advances of natural language processing and image understanding, more complex and demanding tasks have become within reach. Our aim is to take advantage of the most recent developments to push the state-of-the-art for answering natural language questions on real-world images. This task unites inference of a question intent and visual scene understanding with an answer generation task.

Recently, architectures based on the idea of layered, end-to-end trainable artificial neural networks have improved the state of the art across a wide range of diverse tasks. Most prominently Convolutional Neural Networks have raised the bar on image classification tasks [Krizhevsky et al., 2012] and Long Short Term Memory Networks [Hochreiter and Schmidhuber, 1997] are dominating performance on a range of sequence prediction tasks such as machine translation [Sutskever et al., 2014].

Most recently, these two trends of employing neural architectures have been combined fruitfully with methods that can generate image [Karpathy and Fei-Fei, 2015] and video descriptions [Venugopalan et al., 2015a]. Both are conditioned on the visual features that stem from deep learning architectures and employ recurrent neural network approaches to produce descriptions.

To further push the boundaries and explore the limits of deep learning architectures, we propose an architecture for answering questions about images. In contrast to prior work, this task needs conditioning on language as well visual input. Both modalities have to be interpreted and jointly represented as an answer depends on inferred meaning of the question and image content.

While there is a rich body of work on natural language understanding that has addressed textual question answering tasks based on semantic parsing, symbolic representation and deduction systems, which also has seen applications to question answering about images [Malinowski and Fritz, 2014a], there is evidence that deep architectures can indeed achieve a similar goal [Weston et al., 2014]. This motivates our work to seek end-to-end architectures that learn to answer questions in a single, holistic model.

We propose *Ask Your Neurons*, an approach to question answering with a recurrent neural network at its core. An overview is given in Figure 1. The image information is encoded via a Convolutional Neural Network (CNN) and the question together with the visual representation is fed into a Long Short Term Memory (LSTM) network. The system is trained to produce the correct answer to the question about the image. CNN and LSTM are trained jointly and end-to-end starting from words and pixels.

*Outline.* In Section 3, we present our novel approach based on recurrent neural networks for the challenging task of answering questions about images, which we presented originally in Malinowski et al. [2015]. The approach combines a CNN with an LSTM into an end-to-end architecture that predicts answers conditioning on a question and an image. Section 4 shows that the proposed approach doubles performance compared to a prior symbolic approach on this task. We collect additional data to study human consensus on this task, propose two new metrics sensitive to these effects, and provide a new baseline, by asking humans to answer the questions without observing the image. We demonstrate a variant of our system that also answers question without accessing any visual information, which beats the human baseline. We also frame the multimodal approach to answer questions about images that combines LSTM with CNN [Malinowski et al., 2015] as a special instance of an encoder-decoder framework. This modular perspective, shown in Section 3.2, allows us to study different design choices on a large scale visual question answering dataset VQA.

Section 6 shows our analysis that leads to an improved visual question answering architecture. A deeper visual encoding together with several important design choices lead to a model that achieves stronger performance on VQA and DAQUAR datasets.

## 2 Related Work

Since we have proposed a challenge and first methods for answering questions about real-world images [Malinowski and Fritz, 2014a,b, 2015, Malinowski et al., 2015], frequently also referred to as “Visual Question Answering”, numerous follow up works have appeared. In the following we first discuss related tasks and subtasks, then early approaches to tackle a broader Visual Turing Test and datasets proposed for it. Finally, we discuss the relations to our work.

### 2.1 Convolutional neural networks for visual recognition

One component to answer questions about images is to extract information from visual content. Since the proposal of AlexNet [Krizhevsky et al., 2012], Convolutional Neural Networks (CNNs) have become dominant and most successful approaches to extract relevant representation from the image. CNNs directly learn the representation from the raw image data and are trained on large image corpora, typically ImageNet [Russakovsky et al., 2014]. Interestingly, after these models are pre-trained on ImageNet, they can typically be adapted for other tasks. In this work, we evaluate how well the most dominant and successful CNN models can be adapted for the Visual Turing Test. Specifically, we evaluate *AlexNet* [Krizhevsky et al., 2012], *VGG* [Simonyan and Zisserman, 2014], *GoogleNet* [Szegedy et al., 2014], and *ResNet* [He et al., 2015]. These models, reportedly, achieve more and more accurate results on the ImageNet dataset, and hence, arguably, serve as increasingly stronger models of visual perception.

### 2.2 Encodings for text sequence understanding

The other important component to answer a question about an image is to understand the natural language question, which means here building a representation of a variable length sequence of words (or characters, but we will focus only on the words in this work). The first approach is to encode all words of the question as a Bag-Of-Words [Manning and Schütze, 1999], and hence ignoring an order in the sequence of words. Another option is to use, similar to the image encoding, a CNN with pooling to handle variable length input [Kim, 2014, Kalchbrenner et al., 2014]. Finally, Recurrent Neural Networks (RNNs) are methods developed to

directly handle sequences, and have shown recent success on natural language tasks such as machine translation [Cho et al., 2014, Sutskever et al., 2014]. In this work we investigate a Bag-Of-Words (BOW), a CNN, and two RNN variants (LSTM [Hochreiter and Schmidhuber, 1997] and GRU [Cho et al., 2014]) to encode the question.

### 2.3 Combining RNNs and CNNs for description of visual content

The task of describing visual content like still images as well as videos has been successfully addressed with a combination of encoding the image with CNNs and decoding, i.e. predicting the sentence description with an RNN [Donahue et al., 2015, Karpathy and Fei-Fei, 2015, Venugopalan et al., 2015b, Vinyals et al., 2014, Zitnick et al., 2013]. This is achieved by using the RNN model that first gets to observe the visual content and is trained to afterwards predict a sequence of words that is a description of the visual content. Our work extends this idea to question answering, where we formulate a model trained to either generate or classify an answer based on visual as well as natural language input.

### 2.4 Grounding of natural language and visual concepts

Dealing with natural language input does involve the association of words with meaning. This is often referred to as the grounding problem - in particular if the “meaning” is associated with a sensory input. While such problems have been historically addressed by symbolic semantic parsing techniques [Krishnamurthy and Kollar, 2013, Matuszek et al., 2012], there is a recent trend of machine learning-based approaches [Kong et al., 2014, Karpathy et al., 2014, Karpathy and Fei-Fei, 2015, Akata et al., 2016, Hu et al., 2016b, Rohrbach et al., 2015a, Mao et al., 2016, Wang et al., 2016, Hu et al., 2016a, Plummer et al., 2016] to find the associations. These approaches have partially been enabled by recently proposed larger scale datasets [Kazemzadeh et al., 2014, Plummer et al., 2015, Yu et al., 2016], providing phrases or referential expressions which are associated with their corresponding image location. Answering questions about images can be interpreted as first grounding the question in the image and then predicting an answer. Our approach thus is similar to the latter approaches in that we do not enforce or evaluate any particular representation of “meaning” on the language or image modality. We treat this as latent and leave it to the joint training approach to establish an appropriate hidden representation to link the visual and textual representations.

### 2.5 Textual question answering

Answering on purely textual questions has been studied in the NLP community [Berant and Liang, 2014, Liang et al., 2013] and state of the art techniques typically employ semantic parsing to arrive at a logical form capturing the intended meaning and infer relevant answers. Only recently, the success of the previously mentioned neural sequence models, namely RNNs, has carried over to this task [Iyyer et al., 2014, Weston et al., 2014]. More specifically Iyyer et al. [2014] use dependency-tree Recursive NN instead of LSTM, and reduce the question-answering problem to a classification task. Weston et al. [2014] propose different kind of network - memory networks - that is used to answer questions about short stories. In their work, all the parts of the story are embedded into different “memory cells”, and next a network is trained to attend to relevant cells based on the question and decode an answer from that. A similar idea has also been applied to question answering about images, for instance by Yang et al. [2015].

### 2.6 Towards a Visual Turing Test

Recently, a large number architectures have been proposed to approach the Visual Turing Test [Malinowski and Fritz, 2014b], in which they mainly tackle a particularly important subtask that tests machines regarding their abilities to answer to questions about real-world images. Such methods range from symbolic to neural based. There are also architectures that combine both symbolic and neural paradigms together. Some approaches use explicit visual representation in the form of bounding boxes surrounding objects of interest, while other use global full frame image representation, or soft attention mechanism. Yet others use an external knowledge base that helps in answering questions.

*Symbolic based approaches.* In our first work towards a Visual Turing Test [Malinowski and Fritz, 2014a], we present a question answering system based on a semantic parser on a varied set of human question-answer pairs. Although it is the first attempt to handle question answering on DAQUAR, and despite its introspective benefits, it is a rule-based approach that requires a careful schema crafting, is not that scalable, and finally it strongly depends on the output of visual analysis methods as joint training in this model is not yet possible. Due to such limitations, the community has rather shifted towards either neural based or combined approaches.

*Deep Neural Approaches with full frame CNN.* Several contemporary approaches use a global image representation, i.e. they encode the whole image with a CNN. Questions are then encoded with an RNN [Malinowski et al., 2015, Ren et al., 2015, Gao et al., 2015] or a CNN [Ma et al., 2016].

In contrast to symbolic based approaches, neural based architectures offer scalable and joint end-to-end training that liberates them from ontological commitment that would otherwise be introduced by a semantic parser. Moreover, such approaches are not ‘hard’ conditioned on the visual input and therefore can naturally take advantage of different language biases in question answer pairs, which can be interpret as learning common sense knowledge.

*Attention-based Approaches.* Following Xu et al. [2015], who proposed to use spatial attention for image description, Yang et al. [2015], Xu and Saenko [2015], Zhu et al. [2016], Chen et al. [2015], Shih et al. [2016], Fukui et al. [2016] predict a latent weighting (attention) of spatially localized images features (typically a convolutional layer of the CNN) based on the question. The weighted image representation rather than the full frame feature representation is then used as a basis for answering the question. In contrast to the previous models using attention, Dynamic Memory Networks (DMN) [Kumar et al., 2015, Xiong et al., 2016] first pass all spatial image features through a bi-directional GRU that captures spatial information from the neighboring image patches, and next retrieve an answer from a recurrent attention based neural network that allows to focus only on a subset of the visual features extracted in the first pass. Another interesting direction has been taken by Ilievski et al. [2016] who run state-of-the-art object detector of the classes extracted from the key words in the question. In contrast to other attention mechanisms, this approach offers a focused, question dependent, “hard” attention.

*Answering with an external knowledge base.* Wu et al. [2016] argue for an approach that first represents an image as an intermediate semantic attribute representation, and next query external knowledge sources based on the most prominent attributes and relate them to the question. With the help of such external knowledge bases, this approach captures richer semantic representation of the world, beyond what is directly contained in images.

*Compositional approaches.* A different direction is taken by Andreas et al. [2016b,a] who predict the most important components to answer the question with a natural language parser. The components are then mapped to neural modules, which are composed to a deep neural network based on the parse tree. While each question induces a different network, the modules are trained jointly across questions. This work compares to Malinowski and Fritz [2014a] by exploiting explicit assumptions about the compositionality of natural language sentences. Related to the Visual Turing Test, Malinowski and Fritz [2014c] have also combined a neural based representation with the compositionality of the language for the text-to-image retrieval task.

*Dynamic parameters.* Noh et al. [2015] have an image recognition network and a Recurrent Neural Network (GRU) that dynamically change the parameters (weights) of visual representation based on the question. More precisely, the parameters of its second last layer are dynamically predicted from the question encoder network and in this way changing for each question. While question encoding and image encoding is pre-trained, the network learns parameter prediction only from image-question-answer triples.

## 2.7 Datasets for visual question answering

Datasets are a driving force for the recent progress in visual question answering. A large number of visual question answering datasets have recently been proposed. The first proposed datasets is DAQUAR [Malinowski and Fritz, 2014a], which contains about 12.5 thousands manually annotated question-answer pairs about 1449 indoor scenes [Silberman et al., 2012]. While the dataset originally contained a single answer (that can consist of multiple words) per question, in this work we extend the dataset by collecting additional answers for each questions. This captures uncertainties in evaluation. We evaluate our approach on this dataset and discuss several consensus evaluation metrics that take the extended annotations into account. In parallel to our work, Geman et al. [2015] developed another variant of the Visual Turing Test. Their work, however, focuses on sequences of yes/no type of questions, and provides detailed object-scene annotations.

Shortly after the introduction of DAQUAR, three other large-scale datasets have been proposed. All are based on MS-COCO [Lin et al., 2014]. Gao et al. [2015] have annotated about 158k images with 316k Chinese question answer pairs together with their corresponding English translations. Ren et al. [2015] have taken advantage of the existing annotations for the purpose of image description generation task and transform them into question answer pairs with the help of a set of hand-designed rules and a syntactic parser [Klein and Manning, 2003]. This procedure has approximately generated 118k question answer pairs. Finally and currently the most popular, large scale dataset on question answering about images is VQA [Antol et al., 2015]. It has approximately 614k questions about the visual content of about 205k real-world images (the whole VQA dataset also contains 150k questions about 50k abstract scenes that are not considered in this work). Similarly to our Consensus idea, VQA provides 10 answers per each image. For the purpose of the challenge the test answers are not publicly available. We perform one part of the experimental analysis in this paper on the VQA dataset, examining different variants of our proposed approach.

Although simple, automatic performance evaluation metrics have been a part of building first visual question answer-

ing datasets [Malinowski and Fritz, 2014a,b, 2015], Yu et al. [2015] have simplified the evaluation even further by introducing Visual Madlibs - a multiple choice question answering by filling the blanks task. In this task, a question answering architecture has to choose one out of four provided answers for a given image and the prompt. Formulating question answering task in this way has wiped out ambiguities in answers, and just a simple accuracy metric can be used to evaluate different architectures on this task. Yet, the task requires holistic reasoning about the images, and despite of simple evaluation, it remains challenging for machines.

The Visual7W [Zhu et al., 2016] extends canonical question and answer pairs with additional groundings of all objects appearing in the questions and answers to the image by annotating the correspondences. It contains natural language answers, but also answers which require to locate the object, which is then similar to the task of explicit grounding discussed above. Visual7W builds question answer pairs based on the Visual Genome dataset [Krishna et al., 2016], and contains about 330k questions. In contrast to others such as VQA [Antol et al., 2015] or DAQUAR [Malinowski and Fritz, 2014a] that has collected unconstrained question answer pairs, the Visual Genome focuses on the six, so called, Ws: *what, where, when, who, why, and how*, which can be answered with a natural language answer. An additional 7th question – *which* – requires a bounding box location as answer. Similarly to Visual Madlibs [Yu et al., 2015], Visual7W also contains multiple-choice answers.

Related to Visual Turing Test, Chowdhury et al. [2016] have proposed collective memories and Xplore-M-Ego - a dataset of images with natural language queries, and a media retrieval system. This work focuses on a user centric, dynamic scenario, where the provided answers are conditioned not only on questions but also on the geographical position of the questioner.

Moving from asking questions about images to questions about video enhances typical questions with temporal structure. Zhu et al. [2015] propose a task which requires to fill in blanks the captions associated with videos. The task requires inferring the past, describing the present and predicting the future in a diverse set of video description data ranging from cooking videos [Regneri et al., 2013] over web videos [Trecvid, 2014] to movies [Rohrbach et al., 2015b]. Tapaswi et al. [2016] propose MovieQA, which requires to understand long term connections in the plot of the movie. Given the difficulty of the data, both works provide multiple-choice answers.

## 2.8 Relations to our work

The original version of this work [Malinowski et al., 2015] belongs to the category of “Deep Neural Approaches with full frame CNN”, and is among the very first methods of

this kind (Section 3.1). We extend [Malinowski et al., 2015] by introducing a more general and modular encoder-decoder perspective (Section 3.2) that encapsulates a few different neural approaches. Next, we broaden our original analysis done on DAQUAR (Section 4) to the analysis of different neural based approaches on VQA showing the importance of getting a few details right together with benefits of a stronger visual encoder (Section 6). Finally, we transfer lessons learnt from VQA [Antol et al., 2015] to DAQUAR [Malinowski and Fritz, 2014a], showing a significant improvement on this challenging task (Section 6).

## 3 Ask Your Neurons

Answering questions about images can be formulated as the problem of predicting an answer  $a$  given an image  $x$  and a question  $q$  according to a parametric probability measure:

$$\hat{a} = \arg \max_{a \in \mathcal{A}} p(a|x, q; \theta) \quad (1)$$

where  $\theta$  represent a vector of all parameters to learn and  $\mathcal{A}$  is a set of all answers. The question  $q$  is a sequence of words, i.e.  $q = [q_1, \dots, q_n]$ , where each  $q_t$  is the  $t$ -th word question with  $q_n = \text{?}$  encoding the question mark - the end of the question. In the following we describe how we represent  $x$ ,  $a$ ,  $q$ , and  $p(\cdot|x, q; \theta)$  in more details.

In a scenario of **multiple word answers**, we consequently decompose the problem to predicting a set of answer words  $a_{q,x} = \{a_1, a_2, \dots, a_{N(q,x)}\}$ , where  $a_t$  are words from a finite vocabulary  $\mathcal{V}$ , and  $N(q, x)$  is the number of answer words for the given question and image. In our approach, named Ask Your Neurons, we propose to tackle the problem as follows. To predict multiple words we formulate the problem as predicting a sequence of words from the vocabulary  $\mathcal{V} := \mathcal{V}' \cup \{\$\}$  where the extra token  $\$$  indicates the end of the answer sequence, and points out that the question has been fully answered. We thus formulate the prediction procedure recursively:

$$\hat{a}_t = \arg \max_{a \in \mathcal{V}} p(a|x, q, \hat{A}_{t-1}; \theta) \quad (2)$$

where  $\hat{A}_{t-1} = \{\hat{a}_1, \dots, \hat{a}_{t-1}\}$  is the set of previous words, with  $\hat{A}_0 = \{\}$  at the beginning, when our approach has not given any answer word so far. The approach is terminated when  $\hat{a}_t = \$$ . We evaluate the method solely based on the predicted answer words ignoring the extra token  $\$$ . To ensure uniqueness of the predicted answer words, as we want to predict the *set* of answer words, the prediction procedure can be trivially changed by maximizing over  $\mathcal{V} \setminus \hat{A}_{t-1}$ . However, in practice, our algorithm learns to not predict any previously predicted words.

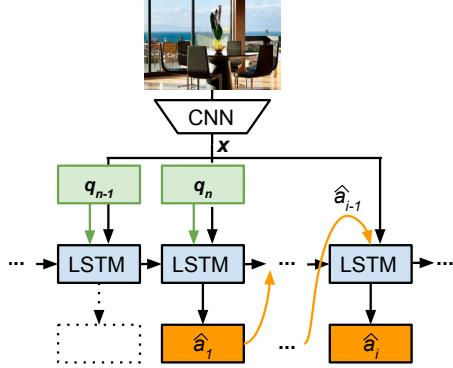


Fig. 2 Our approach *Ask Your Neurons*, see Section 3 for details.

If we only have **single word answers**, or if we model each multi-word answer as a different answer (i.e. vocabulary entry), we use Equation 1 only once to pick the most likely answer.

In the following we first present a Ask Your Neurons that models multi-word answers with a single recurrent network for question and image encoding and answer prediction (Section 3.1) and then present a more general and modular framework with question and image encoders, as well as answer decoder as modules (Section 3.2).

### 3.1 Method

As shown in Figure 1 and Figure 2, we feed our approach Ask Your Neurons with a question as a sequence of words. Since our problem is formulated as a variable-length input-output-sequence, we decide to model the parametric distribution  $p(\cdot|x, q; \theta)$  of Ask Your Neurons with a recurrent neural network and a softmax prediction layer. More precisely, Ask Your Neurons is a deep network built of CNN [LeCun et al., 1998] and Long-Short Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997]. We decide on LSTM as it has been recently shown to be effective in learning a variable-length sequence-to-sequence mapping [Donahue et al., 2015, Sutskever et al., 2014].

Both question and answer words are represented with one-hot vector encoding (a binary vector with exactly one non-zero entry at the position indicating the index of the word in the vocabulary) and embedded in a lower dimensional space, using a jointly learnt latent linear embedding. In the training phase, we augment the question words sequence  $q$  with the corresponding ground truth answer words sequence  $a$ , i.e.  $\hat{q} := [q, a]$ . During the test time, in the prediction phase, at time step  $t$ , we augment  $q$  with previously predicted answer words  $\hat{a}_{1..t} := [\hat{a}_1, \dots, \hat{a}_{t-1}]$ , i.e.  $\hat{q}_t := [q, \hat{a}_{1..t}]$ . This means the question  $q$  and the previous answer words are encoded implicitly in the hidden states of the LSTM, while the latent hidden representation is learnt. We encode the image  $x$  using a CNN and provide it at every

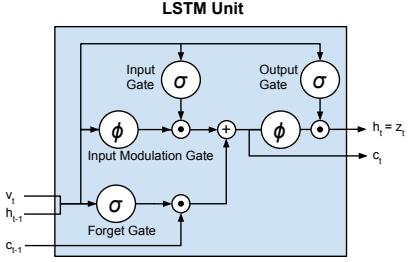


Fig. 3 LSTM unit. See Section 3.1.1, Equations (3)-(8) for details.

time step as input to the LSTM. We set the input  $v_t$  as a concatenation of  $[\Phi(x), \hat{q}_t]$ , where  $\Phi(\cdot)$  is the CNN encoding.

#### 3.1.1 Long-Short Term Memory (LSTM)

As visualized in detail in Figure 3, the LSTM unit takes an input vector  $v_t$  at each time step  $t$  and predicts an output word  $z_t$  which is equal to its latent hidden state  $h_t$ . As discussed above,  $z_t$  is a linear embedding of the corresponding answer word  $a_t$ . In contrast to a simple RNN unit, the LSTM unit additionally maintains a memory cell  $c$ . This allows to learn long-term dynamics more easily and significantly reduces the vanishing and exploding gradients problem [Hochreiter and Schmidhuber, 1997]. More precisely, we use the LSTM unit as described in Zaremba and Sutskever [2014]. With the *sigmoid* nonlinearity  $\sigma : \mathbb{R} \mapsto [0, 1]$ ,  $\sigma(v) = (1 + e^{-v})^{-1}$  and the *hyperbolic tangent* nonlinearity  $\phi : \mathbb{R} \mapsto [-1, 1]$ ,  $\phi(v) = \frac{e^v - e^{-v}}{e^v + e^{-v}} = 2\sigma(2v) - 1$ , the LSTM updates for time step  $t$  given inputs  $v_t$ ,  $h_{t-1}$ , and the memory cell  $c_{t-1}$  as follows:

$$i_t = \sigma(W_{vi}v_t + W_{hi}h_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{vf}v_t + W_{hf}h_{t-1} + b_f) \quad (4)$$

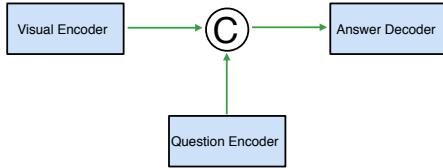
$$o_t = \sigma(W_{vo}v_t + W_{ho}h_{t-1} + b_o) \quad (5)$$

$$g_t = \phi(W_{vg}v_t + W_{hg}h_{t-1} + b_g) \quad (6)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (7)$$

$$h_t = o_t \odot \phi(c_t) \quad (8)$$

where  $\odot$  denotes element-wise multiplication. All the weights  $W$  and biases  $b$  of the network are learnt jointly with the cross-entropy loss. Conceptually, as shown in Figure 3, Equation 3 corresponds to the input gate, Equation 6 the input modulation gate, and Equation 4 the forget gate, which determines how much to keep from the previous memory  $c_{t-1}$  state. As Figures 1 and 2 suggest, all the output predictions that occur before the question mark are excluded from the loss computation, so that the model is penalized solely based on the predicted answer words.



**Fig. 4** Our *Refined Ask Your Neurons* architecture for answering questions about images that includes the following modules: visual and question encoders, and answer decoder. A multimodal embedding  $C$  combines both encodings into a joint space that the decoder decodes from. See Section 3.2 for details.

### 3.2 Refined Ask Your Neurons

In the previous section, we have described how to achieve visual question answering with a single recurrent network for question and image encoding and answering. In this section, we abstract away from the particular design choices taken and describe a modular framework, where a question encoder has to be combined with a visual encoder in order to produce answers with an answer decoder (Figure 4). This modular representation allows us to systematically investigate and evaluate a range of design choices of different encoders, multimodal embeddings, and decoders.

#### 3.2.1 Question encoders

The main goal of a question encoder is to capture a meaning of the question, which we write here as  $\Psi(q)$ . Such an encoder can range from a very structured ones like Semantic Parser used in Malinowski and Fritz [2014a] and Liang et al. [2013] that explicitly model compositional nature of the question, to orderless Bag-Of-Word (BOW) approaches that merely compute a histogram over the question words (Figure 6). In this work, we investigate a few encoders within such spectrum that are compatible with the proposed Deep Learning approach: Two recurrent question encoders, LSTM [Hochreiter and Schmidhuber, 1997] (see Section 3.1.1) and GRU [Cho et al., 2014], that assume a temporal ordering in questions, as well as the aforementioned BOW.

**Gated Recurrent Unit (GRU).** GRU is a simplified variant of LSTM that also uses gates (a reset gate  $r$  and an update gate  $u$ ) in order to keep long term dependencies. GRU is expressed by the following set of equations:

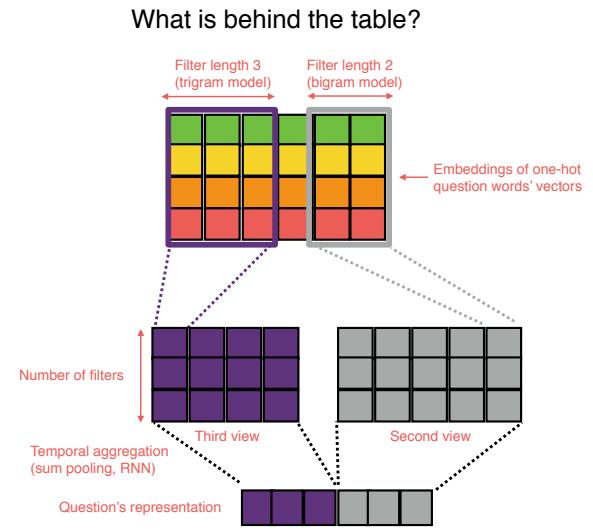
$$r_t = \sigma(W_{vr}v_t + W_{hr}h_{t-1} + b_r) \quad (9)$$

$$u_t = \sigma(W_{vu}v_t + W_{hu}h_{t-1} + b_u) \quad (10)$$

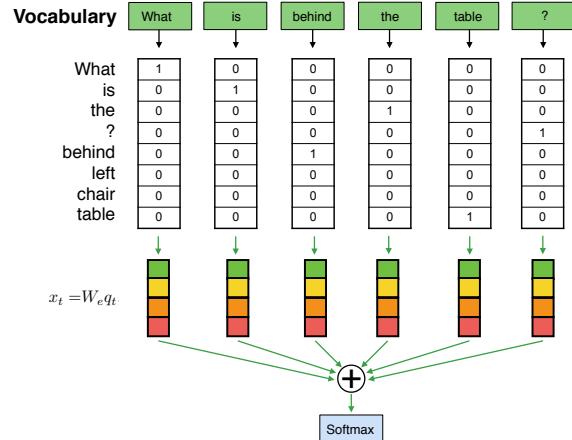
$$c_t = W_{vc}v_t + W_{hc}(r_t \odot h_{t-1}) + b_c \quad (11)$$

$$h_t = u_t \odot h_{t-1} + (1 - u_t) \odot \phi(c_t) \quad (12)$$

where  $\sigma$  is the sigmoid function,  $\phi$  is the hyperbolic tangent, and  $v_t$ ,  $h_t$  are input and hidden state at time  $t$ . The representation of the question  $q$  is the hidden vector at last time step, i.e.  $\Psi_{\text{RNN}}(q) := h_T$ .



**Fig. 5** CNN for encoding the question that convolves word embeddings (learnt or pre-trained) with different kernels, second and third views are shown, see Section 3.2.1 and Yang et al. [2015] for details.



**Fig. 6** Bag-Of-Words (BOW) for encoding the question, see Section 3.2.1 for details.

**Bag-Of-Word (BOW).** Conceptually the simplest encoder, the BOW approach (Figure 6) sums over the words embeddings:

$$\Psi_{\text{BOW}}(q) := \sum_t^n W_e(q_t). \quad (13)$$

where  $W_e$  is a matrix and  $q_t$  is one-hot binary vector of the word with exactly one 1 pointing to a place of the 'word' in the vocabulary (Figure 6). BOW does not encode the ordering of words in the question, so that especially questions with swapped arguments of spatial prepositions become indistinguishable, i.e.  $\Psi_{\text{BOW}}(\text{red chair left of sofa}) = \Psi_{\text{BOW}}(\text{red sofa left of chair})$  in the BOW sentence representation.

*Convolutional Neural Network (CNN).* Convolutional Neural Networks (CNN) have been proposed to encode language [Kim, 2014, Kalchbrenner et al., 2014, Ma et al., 2016, Yang et al., 2015] and since have shown to be fast to compute and result in good accuracy. Since they consider a larger context, they arguably maintain more structure than BOW, but do not model such long term dependencies as recurrent neural networks. Figure 5 depicts our CNN architecture, which is very similar to Ma et al. [2016] and Yang et al. [2015], that convolves word embeddings with three convolutional kernels of length 1, 2 and 3. For the sake of clarity, we only show two kernels in the figure. We either learn them jointly with the whole model or use GLOVE [Pennington et al., 2014] in our experiments. We call such architecture with 1, ...,  $n$  kernel lengths  $n$  views CNN. At the end, the kernel’s outputs are temporarily aggregated for the final question’s representation. We use either sum pooling or a recurrent neural network (CNN-RNN) to accomplish this step.

### 3.2.2 Visual encoders

The second important component of the encoder-decoder architectures is the visual representation. Convolutional Neural Networks (CNNs) have become the state-of-the-art framework that provide features from images. The typical protocol of using the visual models is to first pre-train them on the ImageNet dataset [Russakovsky et al., 2014], a large scale recognition dataset, and next use them as an input for the rest of the architecture. Fine-tuning the weights of the encoder to the task at hand is also possible. In our experiments, we use chronologically the oldest CNN architecture fully trained on ImageNet – a Caffe implementation of AlexNet [Jia et al., 2014, Krizhevsky et al., 2012] – as well as the recently introduced deeper networks – Caffe implementations of GoogLeNet and VGG [Szegedy et al., 2014, Simonyan and Zisserman, 2014] – to the most recent extremely deep architectures – a Facebook implementation of 152 layered ResidualNet [He et al., 2015]. As can be seen from our experiments in Section 6, a strong visual encoder plays an important role in the overall performance of the architecture.

### 3.2.3 Multimodal embedding

The presented neural question encoders transform linguistic question into a vector space. Similarly visual encoders encode images as vectors. A multimodal fusion module combines both vector spaces into another vector based on which the answer is decoded. Let  $\Psi(\mathbf{q})$  be a question representation (BOW, CNN, LSTM, GRU), and  $\Phi(\mathbf{x})$  be a representation of an image. Then  $C(\Psi(\mathbf{q}), \Phi(\mathbf{x}))$  is a function which embeds both vectors. In this work, we investigate three multimodal embedding techniques: Concatenation, element-wise multiplication, and summation. Since the last two techniques require compatibility in the number of feature components, we

use additional visual embedding matrix  $W_{ve} \in \mathbb{R}^{|\Psi(\mathbf{q})| \times |\Phi(\mathbf{x})|}$ . Let  $W$  be weights of an answer decoder. Then we have  $WC(\Psi(\mathbf{q}), \Phi(\mathbf{x}))$ , which is

$$W_q \Psi(\mathbf{q}) + W_v \Phi(\mathbf{x}) \quad (14)$$

$$W(\Psi(\mathbf{q}) \odot W_{ve} \Phi(\mathbf{x})) \quad (15)$$

$$W\Psi(\mathbf{q}) + WW_{ve}\Phi(\mathbf{x}) \quad (16)$$

in concatenation, element-wise multiplication, and summation fusion techniques respectively. In Equation 14, we decompose  $W$  into two matrices  $W_q$  and  $W_v$ , that is  $W = [W_q; W_v]$ . In Equation 15,  $\odot$  is an element-wise multiplication. Similarity between Equation 14 and Equation 16 is interesting as the latter is the former with weight sharing and additional decomposition into  $WW_{ve}$ .

### 3.2.4 Answer decoders

We consider two approach to decode the internal representation of our model into an answer.

*Answer words generation.* The last component of our architecture (Figure 4) is an answer decoder. Inspired by the work on the image description task [Donahue et al., 2015], we uses an LSTM as decoder that shares the parameters with the encoder.

*Classification.* As alternative, we can cast the answering problem as a classification task, with answers as different classes. This approach has been widely explored, especially on VQA [Antol et al., 2015].

## 4 Analysis on DAQUAR

In this section, we benchmark our method on a task of answering questions about images. We compare different variants of our proposed model to prior work in Section 4.1. In addition, in Section 4.2, we analyze how well questions can be answered without using the image in order to gain an understanding of biases in form of prior knowledge and common sense. We provide a new human baseline for this task. In Section 4.3 we discuss ambiguities in the question answering tasks and analyze them further by introducing metrics that are sensitive to these phenomena. In particular, the WUPS score [Malinowski and Fritz, 2014a] is extended to a consensus metric that considers multiple human answers. All the material is available on our project webpage<sup>1</sup>.

<sup>1</sup> [http://mpii.de/visual\\_turing\\_test](http://mpii.de/visual_turing_test)

	Accu- racy	WUPS @0.9	WUPS @0.0
Malinowski and Fritz [2014a]	7.86	11.86	38.79
Ask Your Neurons (ours)			
- multiple words	17.49	23.28	57.76
- single word	<b>19.43</b>	<b>25.28</b>	<b>62.00</b>
Human answers 2014a	50.20	50.82	67.27
Question-only (ours)			
- multiple words	17.06	22.30	56.53
- single word	17.15	22.80	58.42
Human answers, no images	7.34	13.17	35.56

**Table 1** Results on DAQUAR, all classes, single reference, in %.

*Experimental protocol.* We evaluate our approach from [Section 3](#) on the DAQUAR dataset [Malinowski and Fritz, 2014a], which provides 12,468 human question answer pairs on images of indoor scenes [Silberman et al., 2012] and follow the same evaluation protocol by providing results on accuracy and the WUPS score at {0.9, 0.0}. We run experiments for the full dataset as well as their proposed reduced set that restricts the output space to only 37 object categories and uses 25 test images. In addition, we also evaluate the methods on different subsets of DAQUAR where only 1, 2, 3 or 4 word answers are present.

We use default hyper-parameters of LSTM [Donahue et al., 2015] and CNN [Jia et al., 2014]. All CNN models are first pre-trained on the ImageNet dataset [Russakovsky et al., 2014], and next we randomly initialize and train the last layer together with the LSTM network on the task. We find this step crucial to obtain good results. We have explored the use of a 2 layered LSTM model, but have consistently obtained worse performance. In a pilot study, we have found that *GoogleNet* architecture [Jia et al., 2014, Szegedy et al., 2014] consistently outperforms the *AlexNet* architecture [Jia et al., 2014, Krizhevsky et al., 2012] as a CNN model for our task and model.

*WUPS scores.* We base our experiments and the consensus metrics on WUPS scores [Malinowski and Fritz, 2014a]. The metric is a generalization of the accuracy measure that accounts for word-level ambiguities in the answer words. For instance ‘carton’ and ‘box’ can be associated with a similar concept, and hence models should not be strongly penalized for this type of mistakes. Formally:

$$\text{WUPS}(A, T) = \frac{1}{N} \sum_{i=1}^N \min \left\{ \prod_{a \in A^i} \max_{t \in T^i} \mu(a, t), \prod_{t \in T^i} \max_{a \in A^i} \mu(a, t) \right\}$$

To embrace the aforementioned ambiguities, Malinowski and Fritz [2014a] suggest using a thresholded taxonomy-based

Wu-Palmer similarity [Wu and Palmer, 1994] for  $\mu$ . Smaller thresholds yield more forgiving metrics. As in Malinowski and Fritz [2014a], we report WUPS at two extremes, 0.0 and 0.9.

#### 4.1 Evaluation of Ask Your Neurons

We start with the evaluation of our Ask Your Neurons on the full DAQUAR dataset in order to study different variants and training conditions. Afterwards we evaluate on the reduced DAQUAR for additional points of comparison to prior work.

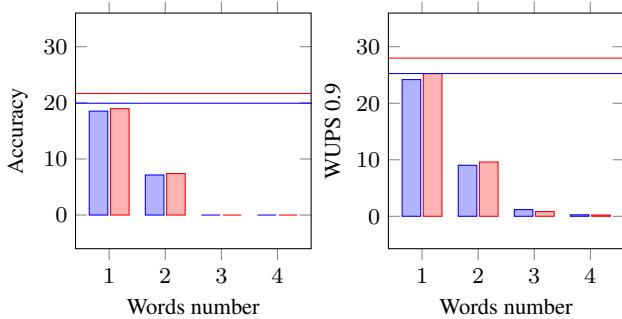
*Results on full DAQUAR.* [Table 1](#) shows the results of our Ask Your Neurons method on the full set (“multiple words”) with 653 images and 5673 question-answer pairs available at test time. In addition, we evaluate a variant that is trained to predict only a single word (“single word”) as well as a variant that does not use visual features (“Question-only”). Note, however, that “single word” refers to a training procedure. All the methods in [Table 1](#) are evaluated on the full DAQUAR dataset at test time that also contains multi-word answers. In comparison to the prior work [Malinowski and Fritz, 2014a] (shown in the first row in [Table 1](#)), we observe strong improvements of over 9% points in accuracy and over 11% in the WUPS scores (second row in [Table 1](#) that corresponds to “multiple words”). Note that, we achieve this improvement despite the fact that the only published number available for the comparison on the full set uses ground truth object annotations [Malinowski and Fritz, 2014a] – which puts our method at a disadvantage. Further improvements are observed when we train only on a single word answer, which doubles the accuracy obtained in prior work. We attribute this to a joint training of the language and visual representations and the dataset bias, where about 90% of the answers contain only a single word.

We further analyze this effect in [Figure 7](#), where we show performance of our approach (“multiple words”) in dependence on the number of words in the answer (truncated at 4 words due to the diminishing performance). The performance of the “single word” variants on the one-word subset are shown as horizontal lines. Although accuracy drops rapidly for longer answers, our model is capable of producing a significant number of correct two words answers. The “single word” variants have an edge on the single answers and benefit from the dataset bias towards these type of answers. Quantitative results of the “single word” model on the one-word answers subset of DAQUAR are shown in [Table 2](#). While we have made substantial progress compared to prior work, there is still a 30% points margin to human accuracy and 25 in WUPS score (“Human answers” in [Table 1](#)).

Later on, in [Section 7](#), we will show improved results on DAQUAR with a stronger visual model and a pre-trained word embedding, with ADAM [Kingma and Ba, 2014] as

	Accu- racy	WUPS @0.9	WUPS @0.0
Ask Your Neurons (ours)	21.67	27.99	65.11
Question-only (ours)	19.13	25.16	61.51

**Table 2** Results of the single word model on the one-word answers subset of DAQUAR, all classes, single reference, in %.



**Fig. 7** Question-only (blue bar) and Ask Your Neurons (red bar) “multi word” models evaluated on different subsets of DAQUAR. We consider 1, 2, 3, 4 word subsets. The blue and red horizontal lines represent “single word” variants evaluated on the answers with exactly 1 word.

the chosen optimization technique. We also put the method in a broader context, and compare with other approaches.

*Results on reduced DAQUAR.* In order to provide performance numbers that are comparable to the proposed Multi-World approach in [Malinowski and Fritz \[2014a\]](#), we also run our method on the reduced set with 37 object classes and only 25 images with 297 question-answer pairs at test time.

[Table 3](#) shows that Ask Your Neurons also improves on the reduced DAQUAR set, achieving 34.68% Accuracy and 40.76% WUPS at 0.9 substantially outperforming [Malinowski and Fritz \[2014a\]](#) by 21.95 percent points of Accuracy and 22.6 WUPS points. Similarly to previous experiments, we achieve the best performance using the “single word” variant of our method. Note that [Ren et al. \[2015\]](#) reported 36.94%, 48.15%, and 82.68% Accuracy, WUPS at 0.9, and WUPS at 0.0 respectively. However, they use another variant of our reduced DAQUAR dataset, where all the multiple word answers are removed. This roughly accounts for 98% of the original reduced DAQUAR dataset.

#### 4.2 Answering questions without looking at images

In order to study how much information is already contained in questions, we train a version of our model that ignores the visual input. The results are shown in [Table 1](#) and [Table 3](#) under “Question-only (ours)”. The best “Question-only” models with 17.15% and 32.32% compare very well in terms of accuracy to the best models that include vision. The latter

	Accu- racy	WUPS @0.9	WUPS @0.0
Malinowski and Fritz [2014a]	12.73	18.10	51.47
Ask Your Neurons (ours)			
- multiple words	29.27	36.50	79.47
- single word	<b>34.68</b>	<b>40.76</b>	79.54
Question-only (ours)			
- multiple words	32.32	38.39	80.05
- single word	31.65	38.35	<b>80.08</b>

**Table 3** Results on reduced DAQUAR, single reference, with a reduced set of 37 object classes and 25 test images with 297 question-answer pairs, in %

achieve 19.43% and 34.68% on the full and reduced set respectively.

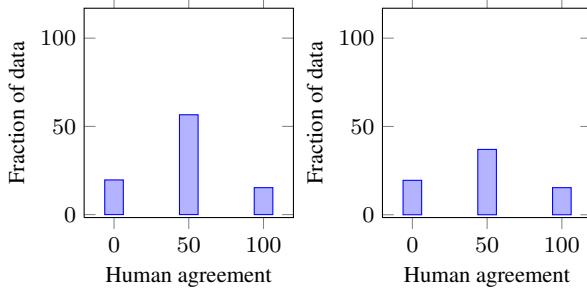
In order to further analyze this finding, we have collected a new human baseline “Human answer, no image”, where we have asked participants to answer on the DAQUAR questions without looking at the images. It turns out that humans can guess the correct answer in 7.86% of the cases by exploiting prior knowledge and common sense. Interestingly, our best “Question-only” model outperforms the human baseline by over 9 percent points. A substantial number of answers are plausible and resemble a form of common sense knowledge employed by humans to infer answers without having seen the image.

#### 4.3 Human Consensus

We observe that in many cases there is an inter human agreement in the answers for a given image and question and this is also reflected by the human baseline performance on the question answering task of 50.20% (“Human answers” in [Table 1](#)). We study and analyze this effect further by extending our dataset to multiple human reference answers in [Section 4.3.1](#), and proposing a new measure – inspired by the work in psychology [[Cohen et al., 1960](#), [Fleiss and Cohen, 1973](#), [Nakashole et al., 2013](#)] – that handles agreement in [Section 4.3.2](#), as well as conducting additional experiments in [Section 4.3.3](#).

##### 4.3.1 DAQUAR-Consensus

In order to study the effects of consensus in the question answering task, we have asked multiple participants to answer the same question of the DAQUAR dataset given the respective image. We follow the same scheme as in the original data collection effort, where the answer is a set of words or numbers. We do not impose any further restrictions on the answers. This extends the original data [[Malinowski and Fritz, 2014a](#)] to an average of 5 test answers per image and



**Fig. 8** Study of inter human agreement. At  $x$ -axis: no consensus (0%), at least half consensus (50%), full consensus (100%). Results in %. Left: consensus on the whole data, right: consensus on the test data.

question collected from 5 in-house annotators. The annotators were first tested for their English proficiency so they would be able to accomplish the task. They were instructed verbally and were given the image and entered an answer for a given question in a text editor. Regular quality checks were performed with a random set of question-answer-image triplets. We refer to this dataset as DAQUAR-Consensus.

#### 4.3.2 Consensus Measures

While we have to acknowledge inherent ambiguities in our task, we seek a metric that prefers an answer that is commonly seen as preferred. We make two proposals:

*Average Consensus.* We use our new annotation set that contains multiple answers per question in order to compute an expected score in the evaluation:

$$\frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K \min \left\{ \prod_{a \in A^i} \max_{t \in T_k^i} \mu(a, t), \prod_{t \in T_k^i} \max_{a \in A^i} \mu(a, t) \right\} \quad (17)$$

where for the  $i$ -th question  $A^i$  is the answer generated by the architecture and  $T_k^i$  is the  $k$ -th possible human answer corresponding to the  $k$ -th interpretation of the question. Both answers  $A^i$  and  $T_k^i$  are sets of the words, and  $\mu$  is a membership measure, for instance WUP [Wu and Palmer, 1994]. We call this metric ‘‘Average Consensus Metric (ACM)’’ since, in the limits, as  $K$  approaches the total number of humans, we truly measure the inter human agreement of every question.

*Min Consensus.* The Average Consensus Metric puts more weights on more ‘‘mainstream’’ answers due to the summation over possible answers given by humans. In order to measure if the result was at least with one human in agreement, we propose a ‘‘Min Consensus Metric (MCM)’’ by replacing the averaging in Equation 17 with a max operator. We call such metric Min Consensus and suggest using both

	Accu- racy	WUPS @0.9	WUPS @0.0
<b>Subset: No agreement</b>			
Question-only (ours)			
- multiple words	8.86	12.46	38.89
- single word	8.50	12.05	40.94
<b>Ask Your Neurons (ours)</b>			
- multiple words	<b>10.31</b>	<b>13.39</b>	40.05
- single word	9.13	13.06	<b>43.48</b>
<b>Subset: <math>\geq 50\%</math> agreement</b>			
Question-only (ours)			
- multiple words	21.17	27.43	66.68
- single word	20.73	27.38	67.69
<b>Ask Your Neurons (ours)</b>			
- multiple words	20.45	27.71	67.30
- single word	<b>24.10</b>	<b>30.94</b>	<b>71.95</b>
<b>Subset: Full Agreement</b>			
Question-only (ours)			
- multiple words	27.86	35.26	78.83
- single word	25.26	32.89	79.08
<b>Ask Your Neurons (ours)</b>			
- multiple words	22.85	33.29	78.56
- single word	<b>29.62</b>	<b>37.71</b>	<b>82.31</b>

**Table 4** Results on DAQUAR, all classes, single reference in % (the subsets are chosen based on DAQUAR-Consensus).

metrics in the benchmarks. We will make the implementation of both metrics publicly available.

$$\frac{1}{N} \sum_{i=1}^N \max_{k=1}^K \left( \min \left\{ \prod_{a \in A^i} \max_{t \in T_k^i} \mu(a, t), \prod_{t \in T_k^i} \max_{a \in A^i} \mu(a, t) \right\} \right) \quad (18)$$

Intuitively, the max operator uses in evaluation a human answer that is the closest to the predicted one – which represents a minimal form of consensus.

#### 4.3.3 Consensus results

Using the multiple reference answers in DAQUAR-Consensus we can show a more detailed analysis of inter human agreement. Figure 8 shows the fraction of the data where the answers agree between all available questions (“100”), at least 50% of the available questions and do not agree at all (no agreement - “0”). We observe that for the majority of the data, there is a partial agreement, but even full disagreement is possible. We split the dataset into three parts according to the above criteria ‘‘No agreement’’, ‘‘ $\geq 50\%$  agreement’’ and ‘‘Full agreement’’ and evaluate our models on these splits (Table 4 summarizes the results). On subsets with stronger agreement, we achieve substantial gains of up to 10% and 20% points in accuracy over the full set (Table 1) and the **Subset: No agreement** (Table 4), respectively. These splits can be seen as curated versions of DAQUAR, which allows studies with factored out ambiguities.

	Accu- racy	WUPS @0.9	WUPS @0.0
<b>Average Consensus Metric</b>			
Question-only (ours)			
- multiple words	11.60	18.24	52.68
- single word	11.57	18.97	54.39
<b>Ask Your Neurons (ours)</b>			
- multiple words	11.31	18.62	53.21
- single word	<b>13.51</b>	<b>21.36</b>	<b>58.03</b>
<b>Min Consensus Metric</b>			
Question-only (ours)			
- multiple words	22.14	29.43	66.88
- single word	22.56	30.93	69.82
Ask Your Neurons (ours)			
- multiple words	22.74	30.54	68.17
- single word	<b>26.53</b>	<b>34.87</b>	<b>74.51</b>

**Table 5** Results on DAQUAR-Consensus, all classes, consensus in %.

	Accuracy	WUPS @0.9	WUPS @0.0
WUPS [Malinowski and Fritz, 2014a]	50.20	50.82	67.27
ACM (ours)	36.78	45.68	64.10
MCM (ours)	60.50	69.65	82.40

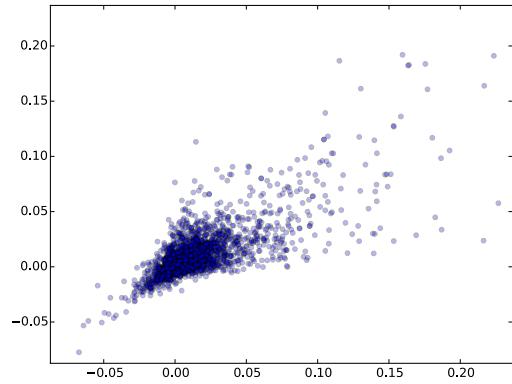
**Table 6** Min and Average Consensus on human answers from DAQUAR, as reference sentence we use all answers in DAQUAR-Consensus which are not in DAQUAR, in %

The aforementioned “Average Consensus Metric” generalizes the notion of the agreement, and encourages predictions of the most agreeable answers. On the other hand “Min Consensus Metric” has a desired effect of providing a more optimistic evaluation. **Table 5** shows the application of both measures to our data and models.

Moreover, **Table 6** shows that “MCM” applied to human answers at test time captures ambiguities in interpreting questions by improving the score of the human baseline from [Malinowski and Fritz \[2014a\]](#) (here, as opposed to **Table 5**, we exclude the original human answers from the measure). It cooperates well with WUPS at 0.9, which takes word ambiguities into account, gaining an 18% higher score.

#### 4.4 Qualitative results

We show predicted answers of different architecture variants in Tables 19, 20, and 21. We chose the examples to highlight differences between Ask Your Neurons and the “Question-only”. We use a “multiple words” approach only in **Table 20**, otherwise the “single word” model is shown. Despite some failure cases, “Question-only” makes “reasonable guesses” like predicting that the largest object could be table or an object that could be found on the bed is a pillow or doll.



**Fig. 9** Figure showing correlation between question and answer words of the “Question-only” model (at x-axis), and a similar correlation of the “Human-baseline” [Malinowski and Fritz, 2014a] (at y-axis).

#### 4.5 Failure cases

While our method answers correctly on a large part of the challenge (e.g.  $\approx 35$  WUPS at 0.9 on “what color” and “how many” question subsets), spatial relations ( $\approx 21$  WUPS at 0.9) which account for a substantial part of DAQUAR remain challenging. Other errors involve questions with small objects, negations, and shapes (below 12 WUPS at 0.9). Too few training data points for the aforementioned cases may contribute to these mistakes. **Table 21** shows examples of failure cases that include (in order) strong occlusion, a possible answer not captured by our ground truth answers, and unusual instances (red toaster).

#### 4.6 Common Sense Knowledge

Although “Question-only” ignores the image, it is still able to make “reasonable guesses” by exploiting biases captured by the dataset. Some of such biases we interpret as a type of a common sense knowledge. For instance, “tea kettle” often sits on the oven, cabinets are usually “brown”, “chair” is typically placed in front of a table, and we commonly keep a “photo” on a cabinet (**Table 22, 24, 25, 28**). On the other hand, some other biases are hardly related to the common sense knowledge. For instance, “11” as the answer to the question “How many bottles are on the desk?” or a “clock” as the answer to the question “what is the black and white object on the top right of the brown board?”. This effect is analysed in **Figure 9**. Each data point in the plot represents the correlation between a question and a predicted answer words for our “Question-only” model (x-axis) versus the correlation in the human answers (y-axis). Despite the reasonable guesses of the “Question-only” architecture, the “Ask Your Neurons” predicts in average better answers (shown in **Table 1**). For instance in **Table 26** the “Question-only” model incorrectly answers “6” on the question “How

many burner knobs are there ?” because it has seen only this answer during the training with exactly the same question but on different image.

## 5 Extended Experiments on DAQUAR

In this section, we first extend our experiments with other baseline methods, and next, guided by our findings on VQA shown in [Section 7](#), we show the results with the refined model in the context of a larger body of results on DAQUAR.

*Baseline methods.* To gain a better understanding of the effectiveness of our neural-based approach, we relate the obtained results to other baseline techniques. Similar baselines were also introduced in [Antol et al. \[2015\]](#) and [Ren et al. \[2015\]](#), although they target either a different dataset or another variant of DAQUAR. The *Constant* technique uses the most frequent answer in the training set to answer to every question in the test set. In *Constant, per question type*, we first break all questions into a few categories and then use the most frequent answer per category to answer to every question in that category at the test time. [Table 7](#) provides more details on the chosen categories. *Look-up table* builds a hash map from a textual question into the most frequent answer for that question at the training time. At the test time, the method just looks up the answer for the question in the hashmap. If the question exists then the most popular answer for that question is provided, otherwise an ‘empty’ answer is given. In addition, we also remove articles, such as ‘the’ and ‘a’, from all the questions. However, this brings only a minor improvement. Finally, we experiment with two nearest-neighbor methods. Both methods rely on a Bag-of-Words representation of questions, where every question word is encoded by GLOVE [[Pennington et al., 2014](#)]. In the following, we call this the representation the semantic space. *Nearest Neighbor, Question-only* searches at test time for the most similar question in the semantic space from the training set. Then it takes the answer that corresponds to this question. *Nearest Neighbor* is inspired by a similar baseline introduced in [Antol et al. \[2015\]](#). At the test time we first search for the 4 most similar questions in the semantic space available in the training set. Next, we form candidate images that correspond to the aforementioned 4 questions. At the last step, we choose an answer that is associated with the best match in the visual space. The latter is done by a cosine similarity between global CNN representations of the test image and every candidate image. We experiment with several CNN representations (VGG-19, GoogLeNet, ResNet-152) but to our surprise there is little performance difference between them. We decide to use GoogLeNet as the results are slightly better.

*Baseline results.* *Constant* shows how the dataset is biased w.r.t. the most frequent answer. This answer turns out to be the number “2”, which also explain a good performance of Ask Your Neurons on the “how many” question subset in DAQUAR. *Constant, per question type* shows that question types provide quite strong clues for answering questions. [Kafle and Kanan \[2016\]](#) take advantage of such clues in their Bayesian and Hybrid models (in [Table 9](#), we only show a better performing Hybrid model). Our next baseline, *Look-up table* can be seen as an extreme case of the *Constant, per question type* model. It also gives surprisingly good results even though it cannot answer to novel questions. This result also confirms our intuitions that “Question-only” biases are important in the ‘question answering about images’ datasets. Finally, next nearest-neighbor baselines show that a visual representation still helps.

*State-of-the-Art.* Based on our further analysis on VQA (for more details we refer to [Section 6](#) and [Section 7](#)), we have also applied the improved model to DAQUAR, and we significantly outperform [Malinowski et al. \[2015\]](#) presented in [Section 4](#). In the experiments, we first choose last 10% of training set as a validation set in order to determine a number of training epochs  $K$ , and next we train the model for  $K$  epochs. We evaluate model on two variants of DAQUAR: all data points (‘all’ in [Table 9](#)), and a subset (‘single word’ in [Table 9](#)) containing only single word answers, which consists of about 90% of the original dataset. As [Table 9](#) shows, our model, Vision + Language with GLOVE and Residual Net that sums visual and question representations, outperforms the model of [Malinowski et al. \[2015\]](#) by 5.05, 4.50, 0.80 of Accuracy, WUPS at 0.9, and WUPS at 0.0 respectively. This shows how important a strong visual model is, as well as the aforementioned details used in training. Likewise to our conclusions on VQA, we are also observing an improvement with attention based models (comparison in *Attention* and *Global* sections in [Table 9](#)).

## 6 Analysis on VQA

While [Section 4](#) analyses our original architecture [[Malinowski et al., 2015](#)] on the DAQUAR dataset, in this section, we analyze different variants and design choices for neural question answering on the large-scale Visual Question Answering (VQA) dataset [[Antol et al., 2015](#)]. It is currently one of the largest and most popular visual question answering dataset with human question answer pairs. In the following, after describing the experimental setup ([Section 6.1](#)), we first describe several experiments which examine the different variants of question encoding, only looking at language input to predict the answer ([Section 6.1](#)), and then, we examine the full model ([Section 6.3](#)).

Type	Regular expression
color	what \is \b the \)?\(\the \)?colo[u]?
count	how many
size	\blargest\b \bsmallest\b \blarge\b \bsmall\b \bbig\b \bbiggest\b
spatial	\b\bfront\b \bleft\b \bright\b \bbelow\b \babove\b \bbeneath\b \b\border\b \bbehind\b \bbeside\b \bacross\b \bahead\b \baround\b
other	-

**Table 7** We start with a regular expression from the top of this table to define the question type and extract the corresponding subset of questions. Next, such a subset is removed from the set of all questions to form the remaining set. Subsequently, we replace the whole set with the remaining set, and continue the procedure. Therefore, a subset of questions corresponding to the “count” type does not contain the “color” question. We will provide the splits in our project web-page under: [http://mpii.de/visual\\_turing\\_test](http://mpii.de/visual_turing_test).

	Accuracy	WUPS@0.9	WUPS@0 on subset
Ask Your Neurons [Malinowski et al., 2015]	19.43	21.67	25.28
Question-only of Malinowski et al. [2015]	17.15	22.80	58.42
Constant	4.97	7.13	26.89
Constant, per question type	11.35	19.55	61.66
Look-up table	15.44	18.45	36.87
Nearest Neighbor, Question-only	13.15	18.39	51.80
Nearest Neighbor	14.58	19.65	54.01

**Table 8** Results of various baselines on DAQUAR. All the results are shown on the whole dataset.

	Accuracy on subset		WUPS@0.9 on subset		WUPS@0 on subset	
	all	single word	all	single word	all	single word
<b>Global</b>						
Ask Your Neurons [Malinowski et al., 2015]	19.43	21.67	25.28	27.99	62.00	65.11
<b>Refined Ask Your Neurons</b>	24.48	26.67	29.78	32.55	62.80	66.25
<b>Refined Ask Your Neurons *</b>	25.74	27.26	31.00	33.25	63.14	66.79
IMG-CNN [Ma et al., 2016]	21.47	24.49	27.15	30.47	59.44	66.08
<b>Attention</b>						
SAN (2, CNN) [Yang et al., 2015]	-	29.30	-	35.10	-	68.60
DMN+ [Xiong et al., 2016]	-	28.79	-	-	-	-
ABC-CNN [Chen et al., 2015]	-	25.37	-	31.35	-	65.89
Comp. Mem. [Jiang et al., 2015]	24.37	-	29.77	-	62.73	-
<b>Question prior</b>						
Hybrid	-	28.96	-	34.74	-	67.33

**Table 9** Comparison with state-of-the-art on DAQUAR. Refined Ask Your Neurons architecture: LSTM + Vision with GLOVE and ResNet-152. Ask Your Neurons architecture: originally presented in Malinowski et al. [2015], results in %. In the comparison, we use original data (all), or a subset with only single word answers (single word) that covers about 90% of the original data. Asterisk ‘\*’ after the method denotes using a box filter that smooths the otherwise noisy validation accuracies. Dash ‘-’ denotes lack of data.

## 6.1 Experimental setup

We evaluate on the VQA dataset [Antol et al., 2015], which is built on top of the MS-COCO dataset [Lin et al., 2014]. Although VQA offers a different challenge tasks, we focus our efforts on the Real Open-Ended Visual Question Answering challenge. The challenge consists of 10 answers per question with about 248k training questions, about 122k validation questions, and about 244k test questions.

As VQA consist mostly of single word answers (over 89%), we treat the question answering problem as a classifi-

cation problem of the most frequent answers in the training set. For the evaluation of the different model variants and design choices, we train on the training set and test on the validation set. Only the final evaluations (Table 17) are evaluated on the test set of the VQA challenge, we evaluate on both parts test-dev and test-standard, where the answers are not publicly available. As a performance measure we use a consensus variant of accuracy introduced in Antol et al. [2015], where the predicted answer gets score between 0 and 1, with 1 if it matches with at least three human answers. We use ADAM [Kingma and Ba, 2014] throughout our ex-

kernel length $k$	single view $= k$	multi view $\leq k$
1	47.43	47.43
2	48.11	48.06
3	<b>48.26</b>	48.09
4	<b>48.27</b>	47.86

**Table 10** Results on VQA validation set, “Question-only” model: Analysis of CNN questions encoders with different filter lengths, accuracy in %, see [Section 6.2.1](#) for discussion.

Question encoder	Word embedding learned	Word embedding GLOVE
BOW	47.41	47.91
CNN	48.26	48.53
GRU	47.60	48.11
LSTM	47.80	<b>48.58</b>

**Table 11** Results on VQA validation set, “Question-only” model: Analysis of different questions encoders, accuracy in %, see [Section 6.2](#) for discussion.

periments as we found out it performs better than SGD with momentum. We keep default hyper-parameters for ADAM. Employed Recurrent Neural Networks maps input question into 500 dimensional vector representation. All the CNNs for text are using 500 feature maps in our experiments, but the output dimensionality also depends on the number of views. In preliminary experiments, we found that removing question mark ‘?’ in the questions slightly improves the results, and we report the numbers only with this setting. Since VQA has 10 answers associated with each question, we need to consider a suitable training strategy that takes this into account. We have examined the following strategies: (1) picking an answer randomly, (2) randomly but if possible annotated as confidently answered, (3) all answers, or (4) choosing the most frequent answer. In the following, we only report the results using the last strategy as we have found out little difference in accuracy between the strategies. To allow training and evaluating many different models with limited time and computational power, we do not fine-tune the visual representations in these experiments, although our model would allow us to do so. All the models, which are publicly available together with our tutorial [[Malinowski and Fritz, 2016](#)] <sup>2</sup>, are implemented in Keras [[Chollet, 2015](#)] and Theano [[Bastien et al., 2012](#)].

## 6.2 Question-only

We start our analysis from “Question-only” models that do not use images to answer on questions. Note that the “Question-only” baselines play an important role in the question answering about images tasks since how much performance is

Encoder	top frequent answers		
	1000	2000	3000
BOW	47.91	48.13	47.94
CNN	48.53	48.67	48.57
LSTM	48.58	<b>48.86</b>	48.65

**Table 12** Results on VQA validation set, “Question-only” model: Analysis of the number of top frequent answer classes, with different question encoders. All using GLOVE; accuracy in %; see [Section 6.2.4](#) for discussion.

added by vision. Hence, better overall performance of the model is not obscured by a better language model. To understand better different design choices, we have conducted our analysis along the different ‘design’ dimensions.

### 6.2.1 CNN questions encoder

We first examine different hyper-parameters for CNNs to encode the question. We first consider the filter’s length of the convolutional kernel. We run the model over different kernel lengths ranging from 1 to 4 ([Table 10](#), left column). We notice that increasing the kernel lengths improves performance up to length 3 were the performance levels out, we thus use kernel length 3 in the following experiments for, such CNN can be interpreted as a trigram model. We also tried to run simultaneously a few kernels with different lengths. In [Table 10](#) (right column) one view corresponds to a kernel length 1, two views correspond to two kernels with length 1 and 2, three views correspond to length 1, 2 and 3, etc. However, we find that the best performance still achieve with a single view and kernel length 3 or 4.

### 6.2.2 BOW questions encoder

Alternatively to neural network encoders, we consider Bag-Of-Words (BOW) approach where one-hot representations of the question words are first mapped to a shared embedding space, and subsequently summed over ([Equation 13](#)), i.e.  $\Psi(\text{question}) := \sum_{\text{word}} W_e(\text{word})$ . Surprisingly, such a simple approach gives very competitive results (first row in [Table 11](#)) compared to the CNN encoding discussed in the previous section (second row).

*Recurrent questions encoder.* We examine two recurrent questions encoders, LSTM [[Hochreiter and Schmidhuber, 1997](#)] and a simpler GRU [[Cho et al., 2014](#)]. The last two rows of [Table 11](#) show a slight advantage of using LSTM.

### 6.2.3 Pre-trained words embedding

In all the previous experiments, we jointly learn the embedding transformation  $W_e$  together with the whole architecture only on the VQA dataset. This means we do not have

<sup>2</sup> [https://github.com/mateuszmalinowski/visual\\_turing\\_test-tutorial](https://github.com/mateuszmalinowski/visual_turing_test-tutorial)

	no norm	L2 norm
Concatenation	47.21	52.39
Summation	40.67	<b>53.27</b>
Element-wise multiplication	49.50	52.70

**Table 13** Results on VQA validation set, vision and language: Analysis of different multimodal techniques that combine vision with language on BOW (with GLOVE word embedding and VGG-19 fc7), accuracy in %, see Section 6.3.1.

Method	Accuracy
BOW	53.27
CNN	54.23
GRU	54.23
LSTM	<b>54.29</b>

**Table 14** Results on VQA validation set, vision and language: Analysis of different language encoders with GLOVE word embedding, VGG-19, and Summation to combine vision and language. Results in %, see Section 6.3.2 for discussion.

any means for dealing with unknown words in questions at test time apart from using a special token (UNK) to indicate such class. To address such shortcoming, we investigate the pre-trained word embedding transformation GLOVE [Pennington et al., 2014] that encodes question words (technically it maps one-hot vector into a 300 dimensional real vector). This choice naturally extends the vocabulary of the question words to about 2 million words extracted a large corpus of web data – Common Crawl [Pennington et al., 2014] – that is used to train the GLOVE embedding. Since the BOW architecture in this scenario becomes shallow (only classification weights are learnt), we add an extra hidden layer between pooling and classification (without this embedding, accuracy drops by 5%). Table 11 (right column) summarizes our experiments with GLOVE. For all question encoders, the word embedding consistently improves performance which confirms that using a word embedding model learnt from a larger corpus helps. LSTM benefits most from GLOVE embedding, archiving the overall best performance with 48.58% accuracy.

#### 6.2.4 Top most frequent answers

Our experiments reported in Table 12 investigate predictions using different number of answer classes. We experiment with a truncation of 1000, 2000, or 4000 most frequent classes. For all question encoders (and always using GLOVE word embedding), we find that a truncation at 2000 words is best, being apparently a good compromise between answer frequency and missing recall.

#### 6.2.5 Summary Question-only

We achieve the best “Question-only” accuracy with GLOVE word embedding, LSTM sentence encoding, and using the

Method	Accuracy
AlexNet	53.69
GoogLeNet	54.52
VGG-19	54.29
ResNet-152	<b>55.52</b>

**Table 15** Results on VQA validation set, vision and language: Different visual encoders (with LSTM, GLOVE, the summation technique, L2 normalized features). Results in %, see Section 6.3.3 for discussion.

Question encoding ↓	Question only		+ Vision	
	Learnt - GLOVE - word embedding			
	Top 1000 answers	Top 2000 answers		
BOW	47.41	47.91	48.13	54.45
CNN	48.26	48.53	48.67	55.34
LSTM	47.80	48.58	48.86	<b>55.52</b>

**Table 16** Results on VQA validation set, vision and language: Summary of our results, results in %, see Section 6.4 for discussion. Columns denote, from the left to right, word embedding learnt together with the architecture, GLOVE embedding that replaces learnt word embedding, truncating the dataset to 2000 most frequent answer classes, and finally added visual representation to the model (ResNet-152).

top 2000 most frequent answers. This achieves an performance of 48.86% accuracy. In the remaining experiments, we use these settings for language and answer encoding.

### 6.3 Vision and Language

Although Question-only models can answer on a substantial number of questions as they arguably capture common sense knowledge, in order to address the full problem we will now also observing the image the question is based on.

#### 6.3.1 Multimodal fusion

Table 13 investigates different techniques that combine visual and language representations. To speed up training, we combine the last unit of the question encoder with the visual encoder, as it is explicitly shown in Figure 4. In the experiments we use concatenation, summation, and element-wise multiplication on the BOW language encoder with GLOVE word embedding and features extracted from the VGG-19 net. In addition, we also investigate using L2 normalization of the visual features, which divides every feature vector by its L2 norm. The experiments show that the normalization is crucial in obtaining good performance, especially for Concatenation and Summation. In the remaining experiments, we use Summation.

#### 6.3.2 Questions encoders

Table 14 shows how well different questions encoders combine with the visual features. We can see that LSTM slightly

Trained on	Test-dev				Test-standard			
	Yes/No	Number	Other	All	Yes/No	Number	Other	All
Training set	78.06	36.79	44.59	57.48	-	-	-	57.55
Training + Val set	78.39	36.45	46.28	<b>58.39</b>	78.24	36.27	46.32	<b>58.43</b>

**Table 17** Results on VQA test set, our best vision and language model chosen based on the validation set: accuracy in %, from the challenge test server. Dash ‘-’ denotes lack of data

outperforms two other encoders GRU and CNN, while BOW remains the worst, confirming our findings in our language-only experiments with GLOVE and 2000 answers (Table 12, second column).

### 6.3.3 Visual encoders

Next we fix the question encoder to LSTM and vary different visual encoders: Caffe variant of *AlexNet* [Krizhevsky et al., 2012], *GoogLeNet* [Szegedy et al., 2014], *VGG-19* [Simonyan and Zisserman, 2014], and recently introduced 152 layered *ResNet* (we use the Facebook implementation of He et al. [2015]). Table 15 confirms our hypothesis that stronger visual models perform better.

### 6.3.4 Qualitative results

We show predicted answers using our best model on VQA test set in Tables 30, 31, 32, 33. We show chosen examples with ‘yes/no’, ‘counting’, and ‘what’ questions, where our model, according to our opinion, makes valid predictions. Moreover, Table 33 shows predicted compound answers.

## 6.4 Summary VQA results

Table 16 summarises our findings on the validation set. We can see that on one hand methods that use contextual language information such as CNN and LSTM are performing better, on the other hand adding strong vision becomes crucial. Furthermore, we use the best found models to run experiments on the VQA test sets: test-dev2015 and test-standard. To prevent overfitting, the latter restricts the number of submissions to 1 per day and 5 submissions in total. Here, we also study the effect of larger datasets where first we train only on the training set, and next we train for 20 epochs on a joint, training and validation, set. When we train on the joint set, we consider question answer pairs with answers among 2000 the most frequent answer classes from the training and validation sets. Training on the joint set have gained us about 0.9%. This implies that on one hand having more data indeed helps, but arguably we also need better models that exploit the current training datasets more effectively. Our findings are summarized in Table 17.

## 7 State-of-the-art on VQA

In this section, we first put our findings on VQA in a broader context, where we compare our refined version of *Ask Your Neurons* with other, publicly available, approaches. Table 18 compares our *Refined Ask Your Neurons* model with other approaches. Some methods, likewise to our approach, use global image representation, other attention mechanism, yet other dynamically predict question dependent weights, external textual sources, or fuse compositional question’s representation with neural networks. Table 18 shows a few trends. First of all, a better visual representation significantly helps (Table 15). Most of the leading approaches to VQA also uses variants of ResNet, which is among the strongest approaches to the image classification task. It is, however, important to normalize the visual features (Table 13). Additionally, all the best models use an explicit attention mechanism (e.g. *DMN+*, *FDA*, *SAN*, *POSTECH*, *MCB*, *HieCoAtt*)

In this work, however, we focus on the extension of the plain “Ask Your Neurons” model that uses a global, full-frame image representation. A similar representation is used in *Refined Ask Your Neurons*, *iBOWIMG*, *VQA team*, and *LSTM Q+I*. The best performing approaches also use different variants of the Recurrent Neural Networks (LSTM and GRU are the most popular). Such a question encoding outperforms Bag-of-Words representations (*iBOWIMG* in Table 18, and *BOW* in Table 14). As we hypothesize, a multimodal embedding plays an important role. This is not only shown in Table 13, but also emphasized in two leading approaches to VQA (*MCB* and *SNUBI*). Both methods use novel multimodal embedding techniques that build upon the element-wise multiplication. Finally, using external textual resources also seems to be beneficial (*AMA*).

## 8 Conclusions

We have presented a neural architecture for answering natural language questions about images that contrasts with prior efforts based on semantic parsing and outperforms a prior symbolic based approach by doubling performance on this challenging task. A variant of our model, that does not use the image to answer the question, already explains a substantial part of the overall performance and helps to understand the contribution of the visual features in this task. From a comparison with the human baseline where humans are not

	Test-dev				Test-standard			
	Yes/No	Number	Other	All	Yes/No	Number	Other	All
SNUBI [Kim et al., 2016]	-	-	-	-	84.6	39.1	57.8	66.9
MCB [Fukui et al., 2016]	83.4	39.8	58.5	66.7	83.2	39.5	58.0	66.5
DLAIT (unpublished)	83.7	40.7	52.3	63.9	83.2	40.8	54.3	64.8
Naver Labs (unpublished)	83.5	39.8	54.8	64.9	83.3	38.7	54.6	64.8
POSTECH (unpublished)	-	-	-	-	83.3	38.0	53.4	64.1
Brandeis [Prakash and Storer, 2016]	82.6	38.1	51.3	62.7	82.1	37.7	51.9	62.9
HieCoAtt [Lu et al., 2016]	79.7	38.7	51.7	61.8	79.9	38.2	51.9	62.1
DualNet [Saito et al., 2016]	82.0	37.9	49.2	61.5	82.0	37.6	49.7	61.8
klab [Kafle and Kanan, 2016]	81.7	39.1	49.2	61.5	81.5	39.3	49.6	61.7
SHB.1026 (unpublished)	82.3	37.0	47.7	60.7	82.1	36.8	47.8	60.8
DMN+ [Xiong et al., 2016]	80.5	36.8	48.3	60.3	80.7	37.0	48.2	60.4
VT.CV.Jiasen (unpublished)	80.5	38.5	47.5	60.1	80.6	38.1	47.9	60.3
FDA [Ilievski et al., 2016]	81.1	36.2	45.8	59.2	81.3	35.7	46.1	59.5
D-NMN [Andreas et al., 2016b]	81.1	38.6	45.5	59.4	81.0	37.5	45.8	59.4
AMA [Wu et al., 2016]	81.0	38.4	45.2	59.2	81.1	37.1	45.8	59.4
SAN [Yang et al., 2015]	79.3	36.6	46.1	58.7	79.1	36.4	46.4	58.8
NMN [Andreas et al., 2016a]	81.2	38.0	44.0	58.6	81.2	37.7	44.0	58.7
<b>Refined Ask Your Neurons</b>	78.4	36.4	46.3	58.4	78.2	36.3	46.3	58.4
SMem [Xu and Saenko, 2015]	80.9	37.3	43.1	58.0	80.8	37.5	43.5	58.2
VQA team [Antol et al., 2015]	80.5	36.8	43.1	57.8	80.6	36.5	43.7	58.2
DPPnet [Noh et al., 2015]	80.7	37.2	41.7	57.2	80.3	36.9	42.2	57.4
iBOWIMG [Zhou et al., 2015]	76.5	35.0	42.6	55.7	76.8	35.0	42.6	55.9
LSTM Q+I [Antol et al., 2015]	78.9	35.2	36.4	53.7	-	-	-	54.1
Comp. Mem. [Jiang et al., 2015]	78.3	35.9	34.5	52.7	-	-	-	-

**Table 18** Results on VQA test set, comparison with state-of-the-art: accuracy in %, from the challenge test server. Dash '-' denotes lack of data. The leaderboard can be found under <http://www.visualqa.org/roe.html>. Baselines from Antol et al. [2015] are not considered.

shown the image to answer the question, we conclude that this language-only model has learnt biases and patterns that can be seen as forms of common sense and prior knowledge that are also used by humans to accomplish this task. We have extended our existing DAQUAR dataset to the new DAQUAR-Consensus, which now provides multiple reference answers which allows to study inter-human agreement and consensus on the question answering task. We propose two new metrics: “Average Consensus”, which takes into account human disagreement, and “Min Consensus” that captures disagreement in human question answering. Finally, we extend our analysis to the large-scale VQA dataset showing competitive performance, yet still using global visual model, and training the model solely on the provided question answer image triples. A broader analysis of the different Deep Learning components and design choices in our model has led to improved results and highlights the importance of a strong visual model.

**Acknowledgements** Marcus Rohrbach was supported by a fellowship within the FITWeltweit-Program of the German Academic Exchange Service (DAAD). The project was in part supported by the Collaborative Research Center (CRC) 1223 from the German Research Foundation (DFG).

What is on the right side of the cabinet?	How many drawers are there?	What is the largest object?
Ask Your Neurons: bed	3	bed
Question-only: bed	6	table

**Table 19** Examples of questions and answers on DAQUAR. Correct predictions are colored in green, incorrect in red.

What is on the refrigerator?	What is the colour of the comforter?	What objects are found on the bed?
Ask Your Neurons: magnet, paper	blue, white	bed sheets, pillow
Question-only: magnet, paper	blue, green, red, yellow	doll, pillow

**Table 20** Examples of questions and answers on DAQUAR with multiple words. Correct predictions are colored in green, incorrect in red.

How many chairs are there?	What is the object fixed on the window?	Which item is red in colour?
Ask Your Neurons: 1	curtain	remote control
Question-only: 4	curtain	clock
Ground truth answers: 2	handle	toaster

**Table 21** Examples of questions and answers on DAQUAR - failure cases.

What are the objects close to the wall?	What is on the stove?	What is left of sink?
Ask Your Neurons: wall decoration	tea kettle	tissue roll
Question-only: books	tea kettle	towel
Ground truth answers: wall decoration	tea kettle	tissue roll

**Table 22** Examples of compound answer words on DAQUAR.

How many lamps are there?	How many pillows are there on the bed?	How many pillows are there on the sofa?
Ask Your Neurons: 2	2	3
Question-only: 2	3	3
Ground truth answers: 2	2	3

**Table 23** Counting questions on DAQUAR.

What color is the towel?	What color are the cabinets?	What is the colour of the pillows?
Ask Your Neurons: brown	brown	black, white
Question-only: white	brown	blue, green, red
Ground truth answers: white	brown	black, red, white

**Table 24** Questions about color on DAQUAR.

		
What is hanged on the chair?	What is the object close to the sink?	What is the object on the table in the corner?
Ask Your Neurons: clothes	faucet	lamp
Question-only: jacket	faucet	plant
Ground truth answers: clothes	faucet	lamp

**Table 25** Correct answers by our “Ask Your Neurons” architecture on DAQUAR.

		
What are the things on the cabinet?	What is in front of the shelf?	How many burner knobs are there?
Ask Your Neurons: photo	chair	4
Question-only: photo	basket	6
Ground truth answers: photo	chair	4

**Table 26** Correct answers by our “Ask Your Neurons” architecture on DAQUAR.

		
What is the object close to the counter?	What is the colour of the table and chair?	How many towels are hanged?
Ask Your Neurons: sink	brown	3
Question-only: stove	brown	4
Ground truth answers: sink	brown	3

**Table 27** Correct answers by our “Ask Your Neurons” architecture on DAQUAR.

What is on the right most side on the table?	What are the things on the coffee table?	What is in front of the table?
Ask Your Neurons: lamp	books	chair
Question-only: machine	jacket	chair
Ground truth answers: lamp	books	chair

**Table 28** Correct answers by our “Ask Your Neurons” architecture on DAQUAR.

What is on the left side of the white oven on the floor and on right side of the blue armchair?	What are the things on the cabinet?	What color is the frame of the mirror close to the wardrobe?
Ask Your Neurons: oven	chair, lamp, photo	pink
Question-only: exercise equipment	candelabra	curtain
Ground truth answers: garbage bin	lamp, photo, telephone	white

**Table 29** Failure cases on DAQUAR.

Are the dogs tied?	Is it summer time?	Is this a real person?
Ask Your Neurons: yes	no	no

**Table 30** Examples of ‘yes/no’ questions and answers produced by our the best model on test VQA.

		
How many kites are only yellow?	How many taps are on the sink?	How many windows are lit?
Ask Your Neurons: 1	2	12

**Table 31** Examples of 'counting' questions and answers produced by our the best model on test VQA.

		
What is the man holding to his ear?	What sport is this man enjoying?	What brand is the laptop?
Ask Your Neurons: phone	snowboarding	apple

**Table 32** Examples of 'what' questions and answers produced by our the best model on test VQA.

		
Color of cow?	What is the man doing?	Where is the TV control?
Ask Your Neurons: brown and white	playing wii	on table

**Table 33** Examples of 'compound answers' questions and answers produced by our the best model on test VQA.

## References

- Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3
- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a. 4, 18
- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan 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 (NAACL)*, 2016b. 4, 18
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 4, 5, 8, 13, 14, 18
- Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012. 15
- Jonathan Berant and Percy Liang. Semantic parsing via paraphrasing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014. 3
- Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv:1511.05960*, 2015. 4, 14
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, Dzmitry Bahdanau, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 3, 7, 15
- François Chollet. keras. <https://github.com/fchollet/keras>, 2015. 15
- Sreyasi Nag Chowdhury, Mateusz Malinowski, Andreas Bulling, and Mario Fritz. Xplore-m-ego: Contextual media retrieval using natural language queries. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2016. 5
- Jacob Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 1960. 10
- Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3, 6, 8, 9
- Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 1973. 10
- Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 4, 18
- Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 3, 4
- Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. In *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 2015. 4
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015. 2, 8, 17
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 1, 3, 6, 7, 15
- Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016a. 3
- Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016b. 3
- Ilija Ilievski, Shuicheng Yan, and Jiashi Feng. A focused dynamic attention model for visual question answering. *arXiv:1604.01485*, 2016. 4, 18
- Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 3
- Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014. 8, 9
- Aiwen Jiang, Fang Wang, Fatih Porikli, and Yi Li. Compositional memory for visual question answering. *arXiv:1511.05676*, 2015. 14, 18
- Kushal Kafle and Christopher Kanan. Answer-type prediction for visual question answering. In *Proceedings of the*

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 13, 18
- Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014. 2, 8
- Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 3
- Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 3
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 3
- Jin-Hwa Kim, Kyoung Woon On, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 18
- Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2, 8
- Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 9, 14
- Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003. 4
- Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv:1602.07332*, 2016. 5
- Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. In *Transactions of the Association for Computational Linguistics (TACL)*, 2013. 3
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. 1, 2, 8, 9, 17
- Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. *arXiv preprint arXiv:1506.07285*, 2015. 4
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 6
- Percy Liang, Michael I Jordan, and Dan Klein. Learning dependency-based compositional semantics. *Computational Linguistics*, 2013. 3, 7
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 4, 14
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical Co-Attention for Visual Question Answering. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 18
- Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2016. 3, 8, 14
- Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems (NIPS)*, 2014a. 2, 3, 4, 5, 7, 8, 9, 10, 12
- Mateusz Malinowski and Mario Fritz. Towards a visual turing challenge. In *Learning Semantics (NIPS workshop)*, 2014b. 2, 3, 5
- Mateusz Malinowski and Mario Fritz. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv:1411.5190*, 2014c. 4
- Mateusz Malinowski and Mario Fritz. Hard to cheat: A turing test based on answering questions about images. *AAAI Workshop: Beyond the Turing Test*, 2015. 2, 5
- Mateusz Malinowski and Mario Fritz. Tutorial on answering questions about images with deep learning. *arXiv preprint arXiv:1610.01076*, 2016. 15
- Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2015. 2, 3, 5, 13, 14
- Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999. 2
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vi-*

- sion and Pattern Recognition (CVPR)*, 2016. 3
- Cynthia Matuszek, Nicholas Fitzgerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012. 3
- Ndapandula Nakashole, Tomasz Tylenda, and Gerhard Weikum. Fine-grained semantic typing of emerging entities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013. 10
- Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. *arXiv:1511.05756*, 2015. 4, 18
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 8, 13, 16
- Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 3
- Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *arXiv:1505.04870*, 2016. 3
- Aaditya Prakash and James Storer. Highway networks for visual question answering. 2016. 18
- Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1, 2013. 5
- Mengye Ren, Ryan Kiros, and Richard Zemel. Image question answering: A visual semantic embedding model and a new dataset. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 3, 4, 10, 13
- Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2015a. 3
- Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015b. 5
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *arXiv:1409.0575*, 2014. 2, 8, 9
- Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. Dualnet: Domain-invariant network for visual question answering. *arXiv preprint arXiv:1606.06108*, 2016. 18
- Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4
- Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 4, 9
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 2, 8, 17
- Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 1, 3, 6
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014. 2, 8, 9, 17
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5
- Trecvid. Trecvid med 14. <http://nist.gov/itl/iad/mig/med14.cfm>, 2014. 5
- Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015a. 2
- Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015b. 3
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv:1411.4555*, 2014. 3
- Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3
- Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv:1410.3916*, 2014. 2, 3

- Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 18
- Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1994. 9, 11
- Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *arXiv preprint arXiv:1603.01417*, 2016. 4, 14, 18
- Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *arXiv:1511.05234*, 2015. 4, 18
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the International Conference on Machine Learning (ICML)*, 2015. 4
- Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. *arXiv:1511.02274*, 2015. 3, 4, 7, 8, 14, 18
- Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2461–2469, 2015. 5
- Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 3
- Wojciech Zaremba and Ilya Sutskever. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014. 6
- Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv:1512.02167*, 2015. 18
- Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering temporal context for video question and answering. *arXiv:1511.04670*, 2015. 5
- Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 5
- C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 3