DRAU: Dual Recurrent Attention Units for Visual Question Answering

Ahmed Osman[a,*], Wojciech Samek[a,*]

[a] *Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, Berlin 10587, Germany*

Abstract

Visual question answering (VQA) requires AI models to comprehend data in two domains, vision and text. Current state-of-the-art models use learned attention mechanisms to extract relevant information from the input domains to answer a certain question. Thus, robust attention mechanisms are essential for powerful VQA models. In this paper, we propose a recurrent attention mechanism and show its benefits compared to the traditional convolutional approach. We introduce a baseline VQA model with visual attention and compare the performance difference between convolutional and recurrent attention on the VQA 2.0 dataset. Furthermore, we experiment replacing attention mechanisms in state-of-the-art models with our recurrent attention units (RAUs) and show increased performance. Additionally, we design an architecture for VQA which utilizes recurrent attention units to highlight the benefits of RAUs. Our single model outperforms the first place winner on the VQA 2016 challenge and to the best of our knowledge, it is the second best performing single model on the VQA 1.0 dataset. Furthermore, our model noticeably improves upon the winner of the VQA 2017 challenge.

Keywords: Visual Question Answering, Attention Mechanisms, Multi-modal Learning, Machine Vision, Natural Language Processing

1. Introduction

Although convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been successfully applied to various image and natural language processing tasks (cf. He et al. (2015); Bosse et al. (2018); Bahdanau et al. (2015); Nallapati et al. (2016)), these breakthroughs only slowly translate to multimodal tasks such as visual question answering (VQA) where the model needs to create a joint understanding of the image and question. Such multimodal tasks require designing highly expressive joint visual and textual representations. On the other hand, a highly discriminative multi-modal feature fusion method is not sufficient for all VQA questions, since global features can contain noisy information for answering questions about certain local parts of the input. This has led to the use of attention mechanisms in VQA.

Attention mechanisms have been extensively used in VQA recently (Anderson et al., 2017; Fukui et al., 2016; Kim et al., 2017). They attempt to make the model selectively predict based on segments of the spatial or lingual context. However, most attention mechanisms used in VQA models are rather simple, consisting of two convolutional layers and a softmax function to generate the attention weights which are summed over the input features. These shallow attention mechanisms could fail to select the relevant information from the joint representation of the question and image for complex questions. Creating attention for complex questions, particularly sequential or relational reasoning questions, requires processing information in a sequential manner which recurrent layers are better suited due to their intrinsic ability to capture relevant information over an

[*] Corresponding authors: Tel.: +49-17-65965-048; +49-30-31002-417;

Email addresses: `ahmed.osman@hhi.fraunhofer.de` (Ahmed Osman), `wojciech.samek@hhi.fraunhofer.de` (Wojciech Samek)

input sequence.

In this paper, we propose a RNN-based attention mechanism for visual and textual attention. We argue that embedding an RNN in the attention mechanism helps the model process information in a sequential manner and determine what is relevant to solve the task. We refer to the combination of RNN embedding and attention as Recurrent Textual Attention Unit (RTAU) and Recurrent Visual Attention Unit (RVAU) respective of their purpose. Furthermore, we employ these units in a fairly simple network, referred to as Dual Recurrent Attention Units (DRAU) network, and show improved results over several baselines. Finally, we enhance state-of-the-art models by replacing their default attention mechanism with RAUs.

Our main contributions are the following:

- We introduce a novel approach to generate soft attention in VQA. To the best of our knowledge, this is the first attempt to generate attention maps using recurrent neural networks for VQA.
- We conduct a direct comparison between two identical models except for their attention mechanism. In this controlled environment, the recurrent attention outperforms the convolutional attention significantly (4% absolute difference). Moreover, we provide qualitative results showing subjective improvements over the default attention used in most VQA models.
- Our attention units are modular, thus, they can substitute existing attention mechanisms in most models fairly easily. We show that state-of-the-art models with RVAU or RTAU "plugged-in" perform consistently better than their standard counterparts.
- We propose a network that utilizes RAUs to co-attend the multi-modal input. We show that our network outperforms the VQA 2016 and 2017 challenge winners and performs close to the current state-of-the-art single models.

In Section 2, we review related work for VQA methods. In Section 3, we break down the components of the DRAU network and explain the components of a RAU. In Section 4, we compare convolutional and recurrent attention in a baseline model, conduct ablation experiments, and report our models' performance on the VQA datasets (Antol et al., 2015; Goyal et al., 2017). Then, we report the results of substituting attention mechanisms of state-of-the-art models with our RAUs. Furthermore, we compare our model against the state-of-the-art on the VQA 1.0 and 2.0 datasets. In Section 5, we compare the difference in attention maps between standard and recurrent attention with qualitative examples to illustrate the effect of RAUs. Finally, we conclude the paper in Section 6.

2. Related Work

This section discusses common methods that have been explored in the past for VQA.

Bilinear pooling representations. Fukui et al. (2016) use compact bilinear pooling to attend over the image features and combine it with the language representation. The basic concept behind compact bilinear pooling is approximating the outer product by randomly projecting the embeddings to a higher dimensional space using Count Sketch projection (Charikar et al., 2004) and then exploiting Fast Fourier Transforms to compute an efficient convolution. An ensemble model using MCB won first place in VQA (1.0) 2016 challenge. Kim et al. (2017) argues that compact bilinear pooling is still expensive to compute and shows that it can be replaced by element-wise product (Hadamard product) and a linear mapping (i.e. fully-connected layer) which gives a lower dimensional representation and also improves the model accuracy. Ben-younes et al. (2017) proposed using Tucker decomposition (Tucker, 1966) with a low-rank matrix constraint as a bilinear representation. Yu et al. (2017a) utilize matrix factorization tricks to create a multi-modal factorized bilinear pooling method (MFB). Later, Yu et al. (2017b) generalizes the factorization for higher-order factorized pooling (MFH).

Attention-based. Lu et al. (2016) were the first to feature a co-attention mechanism that applies attention to both the question and image. Nam et al. (2017) use a Dual Attention Network (DAN) that employs attention on both text and visual features iteratively to predict the result. The goal behind this is to allow the image and question attentions to iteratively guide each other in a synergistic manner. Both methods use convolutional layers to learn the attention mechanisms.

RNNs for VQA. Using recurrent neural networks (RNNs) for VQA models has been explored in the past, but, to the best of our knowledge, has never been used as an attention mechanism. Xiong et al. (2016) build upon the dynamic memory network from Kumar and Varaiya (2015) and proposes DMN+. DMN+ uses episodic modules which contain attention-based Gated Recurrent Units (GRUs). Note that this is not the same as what we propose; Xiong et al. generate soft attention using *convolutional layers* and uses the output to substitute the update gate of the GRU. In contrast, our approach uses the *recurrent layers* to generate the attention. Noh and Han (2016) propose recurrent answering units in which each unit is a complete module that can answer a question about an image. They use joint loss minimization to train the units. However during testing, they use the first answering unit which was trained from other units through backpropagation.

3. Dual Recurrent Attention in VQA

In this section, we define our attention mechanism. Then, we describe the components of our VQA model in this section.

3.1. Recurrent Attention Units

The recurrent attention unit (RAU)receives a multi-modal multi-channel representation of the inputs, X. To reduce the input representation, a RAU starts with a 1×1 convolution and PReLU activation:

$$\begin{aligned} c_a &= \text{PReLU}\,(W_a\, X), \\ X &\in \mathbb{R}^{K \times \phi} \end{aligned} \tag{1}$$

where W_a is the 1×1 convolution weights, X is the multimodal input to the RAU, K is the shape of the target attention (e.g. image size or question length), and ϕ is the number of channels in the input.

Furthermore, we feed the previous output into an unidirectional LSTM:

$$h_{a,n} = \text{LSTM}\,(c_{a,n}) \tag{2}$$

where $h_{a,n}$ is the hidden state at time n. Each hidden state processes the joint features at each location/word of the input and decides which information should be kept and propagated forward and which information should be ignored.

To generate the attention weights, we feed all the hidden states of the previous LSTM to a 1×1 convolution layer followed by a softmax function. The 1×1 convolution layer could be interpreted as the *number of glimpses* the model sees.

$$W_{att,n} = \text{softmax}\left(\text{PReLU}\,(W_g\; h_{a,n})\right) \tag{3}$$

where W_g is the glimpses' weights and $W_{att,n}$ is the attention weight vector.

Next, we use the attention weights to compute a weighted average of the image and question features.

$$att_{a,n} = \sum_{n=1}^{N} W_{att,n}\; f_n \tag{4}$$

where f_n is the input representation and $att_{a,n}$ is the attention applied on the input. Finally, the attention maps are fed into a fully-connected layer followed by a PReLU activation. Figure 2 illustrates the structure of a RAU.

$$y_{att,n} = \text{PReLU}\,(W_{out}\; att_{a,n}) \tag{5}$$

where W_{out} is a weight vector of the fully connected layer and $y_{att,n}$ is the output of the RAU.

3.2. Input Representation

Image representation. We use the 152-layer "ResNet" pretrained CNN from He et al. (2015) to extract image features. Similar to (Fukui et al., 2016; Nam et al., 2017), we resize the images to 448×448 and extract the last layer before the final pooling layer (res5c) with size 2048×14×14. Finally, we use l_2 normalization on all dimensions. Recently, Anderson et al. (2017) have shown that object-level features can provide a significant performance uplift compared to global-level features from pretrained CNNs. Therefore, we experiment with replacing the ResNet features with FRCNN Ren et al. (2015) features with a fixed number of proposals per image ($K = 36$) for our final DRAU model.

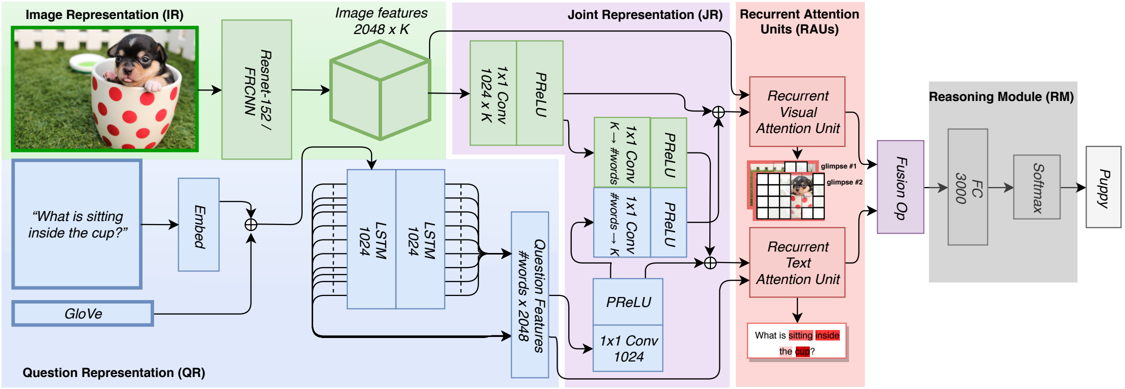
Image Representation (IR)
Resnet-152 / FRCNN
Image features 2048 x K
Joint Representation (JR)
1x1 Conv 1024 x K
PReLU
1x1 Conv K → #words
PReLU
1x1 Conv #words → K
PReLU
PReLU
1x1 Conv 1024
Recurrent Attention Units (RAUs)
Recurrent Visual Attention Unit
glimpse #1
glimpse #2
Recurrent Text Attention Unit
What is sitting inside the cup?
Fusion Op
Reasoning Module (RM)
FC 3000
Softmax
Puppy
"What is sitting inside the cup?"
Embed
GloVe
LSTM 1024
LSTM 1024
Question Features #words x 2048
Question Representation (QR)

Figure 1: The proposed network. ⊕ denotes concatenation.

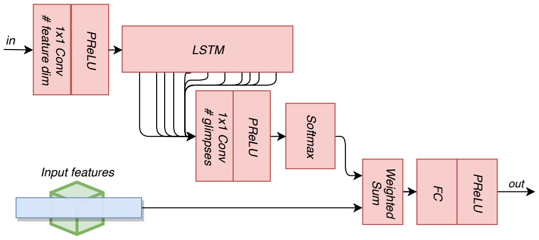

Figure 2: Recurrent Attention Unit.

Question representation. We use a fairly similar representation as Fukui et al. (2016). In short, the question is tokenized and encoded using an embedding layer followed by a tanh activation. We also exploit pretrained GloVe vectors (Pennington et al., 2014) and concatenate them with the output of the embedding layer. The concatenated vector is fed to a two-layer unidirectional LSTM that contains 1024 hidden states each. In contrast to Fukui et al., we use all the hidden states of both LSTMs rather than concatenating the final states to represent the final question representation.

3.3. 1×1 Convolution and PReLU

We apply multiple 1×1 convolution layers in the network for mainly two reasons. First, they learn weights from the image and question representations in the early layers. This is important especially for the image representation, since it was originally trained for a different task. Second, they are used to generate a common representation size. To obtain a joint representation, we apply 1×1 convolutions followed by PReLU activations (He et al., 2015) on both the image and question representations. Through empirical evidence, PReLU activations were found to reduce training time significantly and improve performance compared to ReLU and tanh activations. We provide these results in Section 4.

3.4. Reasoning layer

A fusion operation is used to merge the textual and visual branches. For DRAU, we experiment with using element-wise multiplication (Hadamard product) and MCB (Fukui et al., 2016; Gao et al., 2016). The result of the fusion is given to a many-class classifier using the top 3000 frequent answers. We use a single-layer softmax with cross-entropy loss. This can be written as:

$$P_a = \text{softmax}\Big(\text{fusion\_op}(y_{text}, y_{vis})W_{ans}\Big) \tag{6}$$

where y_{text} and y_{vis} are the outputs of the RAUs, W_{ans} represents the weights of the multi-way classifier, and P_a is the probability of the top 3000 frequent answers.

The final answer $\hat{a}$ is chosen according to the following:

$$\hat{a} = \arg\max P_a \tag{7}$$

4. Experiments and Results

Experiments are performed on the VQA 1.0 and 2.0 datasets (Goyal et al., 2017; Antol et al., 2015). These datasets use images from the MS-COCO dataset (Lin et al., 2014) and generate questions and labels (10 labels per question) using Amazon's Mechanical Turk (AMT). Compared to VQA 1.0, VQA 2.0 adds more image-question pairs to balance the language prior present in the VQA 1.0 dataset (Goyal et al., 2017). The ground truth answers in the VQA dataset are evaluated using human consensus:

$$\mathrm{Acc}(a) = \min\left(\frac{\sum a \text{ is in human annotation}}{3}, 1\right) \quad (8)$$

We evaluate our results on the *validation*, *test-dev*, *test-std* splits of each dataset. Models evaluated on the validation set use *train* and Visual Genome for training for our baselines, but not for our DRAU model using FRCNN features which only use the *train* split. For the other splits, we include the validation set in the training data.

To train our model, we use Adam (Kingma and Ba, 2014) for optimization with $\beta_1 = 0.9, \beta_2 = 0.999$, and an initial learning rate of $\epsilon = 7 \times 10^{-4}$. The final model is trained with a small batch size of 32 for 400K iterations. We did not fully explore tuning the batch size which explains the relatively high number of training iterations. Dropout ($p = 0.3$) is applied after each LSTM and after the fusion operation. All weights are initialized as described in (Glorot and Bengio, 2010) except LSTM layers which use an uniform weight distribution. Since VQA datasets provide 10 answers per image-question pair, we sample one answer randomly for each training iteration.

4.1. Early baselines

During early experiments, the VQA 2.0 dataset was not yet released. Thus, the baselines and early models were evaluated on the VQA 1.0 dataset.

Baselines. We started by designing three baseline architectures. The first baseline produced predictions solely from the question while totally ignoring the image. The model used the same question representation described in Fukui et al. (2016) and passed the output to a softmax 3000-way classification layer. The goal of this architecture was to assess the extent of the language bias present in VQA.

The second baseline is a simple joint representation of the image features and the language representation. The representations were combined using the compact bilinear pooling from Gao et al. (2016). We chose this method specifically because it was shown to be effective by Fukui et al. (2016). The main objective of this model is to measure how a robust pooling method of multimodal features would perform on its own without a deep architecture or attention. We refer to this model as *Simple MCB*.

For the last baseline, we substituted the compact bilinear pooling from Simple MCB with an LSTM consisting of hidden states equal to the image size. A 1×1 convolutional layer followed by a tanh activation were used on the image features prior to the LSTM, while the question representation was replicated to have a common embedding size for both representations This model is referred to as *Joint LSTM*. Note that this model does not use attention.

We begin by testing our baseline models on the VQA 1.0 validation set. As shown in Table 1, the language-only baseline model managed to get 48.3% overall. More impressively, it scored 78.56% on Yes/No questions. The *Simple MCB* model further improves the overall performance, although little improvement is gained in the binary Yes/No tasks. Replacing MCB with our basic *Joint LSTM* embedding improves performance across the board.

VQA 2.0. After the release of VQA 2.0, we shifted our empirical evaluation towards the newer dataset. We retrain and retest our best performing VQA 1.0 model *Joint LSTM*. Since VQA 2.0 was built to reduce the language prior and bias inherent in VQA, the accuracy of *Joint LSTM* drops significantly as shown in Table 1. Note that all the models that were trained so far do not have explicit visual or textual attention implemented.

Comparing convolutional versus recurrent attention. Due to the widespread use of attention in VQA, we compare using convolution against our recurrent attention in a simple baseline model. We chose to use an even simpler model than our previous baselines as an effort to

Table 1: Evaluation of the baseline models on the VQA 1.0 and 2.0 *validation* splits.

VQA 1.0 Validation Split				
Baselines	Y/N	Num.	Other	All
Language only	78.56	27.98	30.76	48.3
Simple MCB	78.64	32.98	39.79	54.82
Joint LSTM	**79.90**	**36.96**	49.58	59.34
VQA 2.0 Validation Split				
Baselines	Y/N	Num.	Other	All
Joint LSTM	72.04	37.95	48.58	56.00
Simple Net (Conv. Attn.)	*66.01*	*28.08*	*25.51*	*41.06*
Simple Net (Recurrent Attn.)	*66.24*	*28.48*	*33.46*	*45.12*

reduce the influence of complex network components on the overall model performance.

This simple VQA model uses the same question representation as described in the previous section and the ResNet global image features. The input features are simply concatenated and sent to the attention mechanism. The processed attention is fed to the reasoning module. We refer to this model as *Simple Net*. The results of Simple Net in Table 1 show a clear advantage of recurrent attention outperforming convolutional attention by over 4% absolute overall accuracy.

4.2. Transplanting RAU in other models

To verify the effectiveness of the recurrent attention units, we replace the attention layers in MCB (Fukui et al., 2016) and MUTAN (Ben-younes et al., 2017) with RVAU (visual attention). Additionally, we replace the textual attention in MFH (Yu et al., 2017b) with recurrent attention.

For MCB we remove all the layers after the first MCB operation until the first 2048-d output and replace them with RVAU. Due to GPU memory constraints, we reduced the size of each hidden unit in RVAU's LSTM from 2048 to 1024. In the same setting, RVAU significantly helps improve the original MCB model's accuracy as shown in Table 2.

Furthermore, we test RVAU in the MUTAN model. The authors use a multimodal vector with dimension size of 510 for the joint representations. For coherence, we change the usual dimension size in RVAU to 510. At the time of this writing, the authors have not released results on VQA 2.0 using a single model rather than a model ensemble. Therefore, we train a single-model MUTAN using the authors' implementation.[1] The story does not change here, RVAU improves the model's overall accuracy.

Finally, we replace the convolution text attention in MFH with RTAU (text attention). Note that the text attention in MFH is "self-attending" which means that the attention is only predicted by just looking at the question. This is different from our DRAU model where RTAU uses a joint representation of the question and image to predict the textual attention. We train two networks, the standard MFH network and MFH with RTAU, on the VQA 2.0 train split and test on the validation split. It is apparent that RTAU improves the overall accuracy of MFH from Table 2. While the performance improvement might not look large for all the tested models, it is consistent which shows that RAUs can reliably improve existing state-of-the-art models with different architectures.

Table 2: Results of state-of-the-art models with RAUs.

VQA 2.0 Test-dev Split				
Model	Y/N	Num.	Other	All
MCB [2]	78.41	38.81	53.23	61.96
MCB w/RVAU	77.31	40.12	54.64	62.33
MUTAN	79.06	38.95	53.46	62.36
MUTAN w/RVAU	79.33	39.48	53.28	62.45
VQA 2.0 Validation Split				
Model	Y/N	Num.	Other	All
MFH	82.26	43.49	56.17	64.31
MFH w/RTAU	82.35	43.31	56.3	64.38

[1] `https://github.com/Cadene/vqa.pytorch`

[2] `http://www.visualqa.org/roe_2017.html`

4.3. DRAU versus the state-of-the-art

VQA 1.0. Table 3 shows a comparison between DRAU and other state-of-the-art models. Excluding model ensembles, DRAU performs favorably against other models. To the best of our knowledge, Yu et al. (2017b) has the best reported single model performance of 67.5% on the *test-std* split. Our single model (DRAU) comes a very close second to the current state-of-the-art single model.

VQA 2.0. The first place submission Anderson et al. (2017) reports using an ensemble of 30 models. In their report, the best single model that also uses FRCNN features achieves 65.67% on the *test-standard* split which is outperformed by our single model (DRAU).

Recently, the VQA 2018 challenge results have been released. It uses the same dataset as the previous VQA 2017 challenge (VQA 2.0). While we have not participated in this challenge, we include the challenge winners results (Jiang et al., 2018) for the sake of completeness. Jiang et al. builds upon the VQA 2017 challenge winners model by proposing a number of modifications. First, they use weight normalization and ReLU instead of gated hyperbolic tangent activation. For the learning schedule, the Adam optimizer was swapped for Adamax with a warm up strategy. Moreover, the Faster-RCNN features have been replaced by the state-of-the-art Feature Pyramid Networks (FPN) object detectors. Lastly, they use more additional training data from the common Visual Genome and the new Visual Dialog (VisDial) datasets.

4.4. Discussion

DRAU versus MCB. The strength of RAUs is notable in tasks that require sequentially processing the image or relational/multi-step reasoning. Figure 3 shows some qualitative results between DRAU and MCB. For fair comparison we compare the first attention map of MCB with the second attention map of our model. We do so because the authors of MCB (Fukui et al., 2016) visualize the first map in their work[3]. Furthermore, the first glimpse of our model seems to be the complement of the second attention, i.e. the model separates the background and the target object(s) into separate attention maps. We have not tested the visual effect of more than two glimpses on our model.

In Figure 3, it is clear that the recurrence helps the model attend to multiple targets as apparent in the difference of the attention maps between the two models. DRAU seems to also know how to count the right object(s). The top right example in Figure 3 illustrates that DRAU is not easily fooled by counting whatever object is present in the image but rather the object that is needed to answer the question. This property also translates to questions that require relational reasoning. The second column in Figure 3 demonstrates how well DRAU can attend to the location required to answer the question based on the textual and visual attention maps compared to MCB.

Attention Quality. Figure 4 shows the model's prediction as well as its attention maps for four questions on the same image. It highlights how DRAU can shift the attention intelligently based on different multi-step reasoning questions. To answer the two left questions, a VQA model needs to sequentially process the image and question. First, the model filters out the animals in the picture. Then, the animal in the question is matched to the visual features and finally counted. The two right-most attention maps give a glimpse on how the model filters out the irrelevant parts in the input. Interestingly, inspecting the visual attention for the top question might indicate a bias in the VQA model. Even though the question asks about "horses", the visual attention filters out all objects and leaves out the two different backgrounds: sea and field. Since "horses" are often found on land, the model predicts "field" without any direct attention on the horses in the image.

5. Conclusion

We proposed an architecture for VQA with a novel attention unit, termed the Recurrent Attention Unit (RAU). The recurrent layers help guide the textual and visual attention since the network can reason relations between several parts of the image and question. We provided quantitative and qualitative results indicating the usefulness of a recurrent attention mechanism.

[3] `https://github.com/akirafukui/vqa-mcb/blob/master/server/server.py\#L185`

Table 3: DRAU compared to the state-of-the-art on the VQA 1.0 dataset. N corresponds to the number of models used for prediction. WE indicates whether the method uses a pre-trained word embedding. VG indicates whether the method uses external data from the Visual Genome dataset.

VQA 1.0 Open Ended Task											
				Test-dev				Test-standard			
Model	N	WE	VG	Y/N	Num.	Other	All	Y/N	Num.	Other	All
DMN+ (Xiong et al., 2016)	1	-	-	80.5	36.8	48.3	60.3	-	-	-	60.4
HieCoAtt (Lu et al., 2016)	1	-	-	79.7	38.7	51.7	61.8	-	-	-	62.1
RAU (Noh and Han, 2016)	1	-	-	81.9	39.0	53.0	63.3	81.7	38.2	52.8	63.2
DAN (Nam et al., 2017)	1	-	-	83.0	39.1	53.9	64.3	82.8	38.1	54.0	64.2
MCB (Fukui et al., 2016)	7	✓	✓	83.4	39.8	58.5	66.7	83.24	39.47	58.00	66.47
MLB (Kim et al., 2017)	1	✓	✗	-	-	-	-	84.02	37.90	54.77	65.07
MLB (Kim et al., 2017)	7	✓	✓	84.57	39.21	57.81	66.77	84.61	39.07	57.79	66.89
MUTAN (Ben-younes et al., 2017))	5	✓	✓	**85.14**	39.81	58.52	67.42	84.91	**39.79**	58.35	67.36
MFH (Yu et al., 2017b)	1	✓	✓	84.9	**40.2**	**59.2**	**67.7**	**84.91**	39.3	**58.7**	**67.5**
DRAU	1	✓	✗	**84.92**	**39.16**	**57.70**	**66.86**	**84.87**	**40.02**	**57.91**	**67.16**

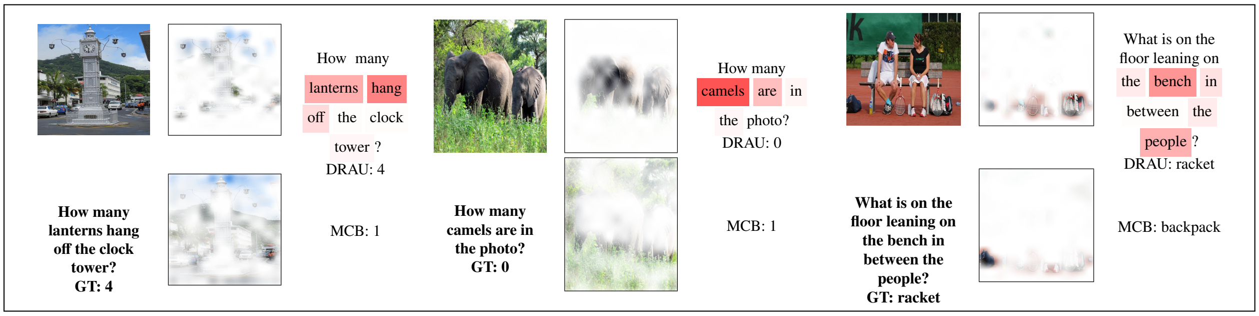
How many lanterns hang off the clock tower ?
DRAU: 4
How many lanterns hang off the clock tower?
GT: 4
MCB: 1
How many camels are in the photo?
DRAU: 0
How many camels are in the photo?
GT: 0
MCB: 1
What is on the floor leaning on the bench in between the people ?
DRAU: racket
What is on the floor leaning on the bench in between the people?
GT: racket
MCB: backpack

Figure 3: DRAU vs. MCB Qualitative examples. Attention maps for both models shown. DRAU shows subjectively better attention map quality.

Using a simple VQA model, we have shown the performance advantage of recurrent attention compared to the traditional convolutional attention used in most VQA attention mechanisms. In VQA 1.0, we come a very close second to the state-of-the-art model. While using the same image features, our DRAU network outperforms the VQA 2017 challenge winner Anderson et al. (2017) in a single-model scenario. Furthermore, we demonstrated that substituting the visual attention mechanism in other networks, MCB (Fukui et al., 2016), MUTAN (Ben-younes et al., 2017), and MFH (Yu et al., 2017b), consistently improves their performance.

In future work we will investigate implicit recurrent attention mechanism using recently proposed explanation methods (Arras et al., 2017; Montavon et al., 2018).

Acknowledgments

This work was supported by the Fraunhofer Society through the MPI-FhG collaboration project ”Theory and Practice for Reduced Learning Machines”. This research was also supported by the German Ministry for Education and Research as Berlin Big Data Center under Grant 01IS14013A and the Berlin Center for Machine Learning under Grant 01IS180371.

Table 4: DRAU compared to the current submissions on the VQA 2.0 dataset. N corresponds to the number of models used for prediction. WE indicates whether the method uses a pre-trained word embedding. VG indicates whether the method uses external data from the Visual Genome dataset.

VQA 2.0 Open Ended Task											
				Test-dev				Test-standard			
Model	N	WE	VG	Y/N	Num.	Other	All	Y/N	Num.	Other	All
VQATeam_MCB (Goyal et al., 2017)	1	✓	✓	78.41	38.81	53.23	61.96	78.82	38.28	53.36	62.27
UPMC-LIP6(Ben-younes et al., 2017)	5	✓	✓	81.96	41.62	57.07	65.57	82.07	41.06	57.12	65.71
HDU-USYD-UNCC(Yu et al., 2017b)	9	✓	✓	84.39	45.76	59.14	68.02	84.5	45.39	59.01	68.09
Adelaide-Teney (Teney et al., 2017)	1	✓	✓	**81.82**	**44.21**	**56.05**	**65.32**	**82.20**	**43.90**	**56.26**	**65.67**
Adelaide-Teney (Teney et al., 2017)	30	✓	✓	85.24	48.19	59.88	69.00	85.54	47.45	59.82	69.13
FAIR A-STAR (Jiang et al., 2018)	1	✓	✓	-	-	-	70.01	-	-	-	70.24
FAIR A-STAR (Jiang et al., 2018)	30	✓	✓	87.82	51.54	63.41	72.12	87.82	51.59	63.43	72.25
DRAU	1	✓	✗	**82.85**	**44.78**	**57.4**	**66.45**	**83.35**	**44.37**	**57.63**	**66.85**

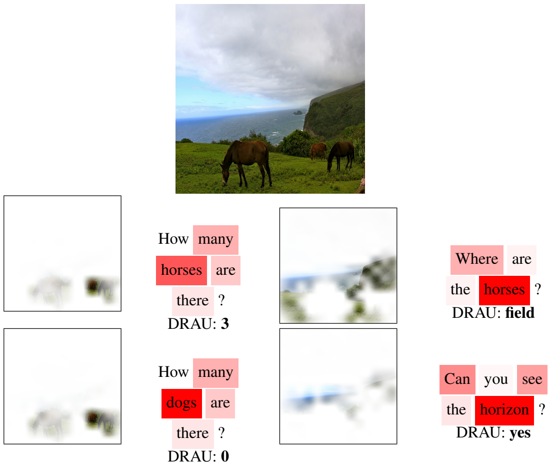
How many horses are there ?
DRAU: 3
Where are the horses ?
DRAU: field
How many dogs are there ?
DRAU: 0
Can you see the horizon ?
DRAU: yes

Figure 4: Four real example results of our proposed model for a single random image. The visual attention, textual attention, and answer are shown. Even on the same image, our model shows rich reasoning capabilities for different question types. The first column shows that the model is able to do two-hop reasoning, initially identifying the animal in the question and then proceed to correctly count it in the image. The second column results highlights the model's ability to shift its attention to the relevant parts of the image and question. It is worth noting that all the keywords in the questions have the highest attention weights.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2017. Bottom-Up and Top-Down Attention for Image Captioning and VQA. arXiv:1707.07998 .

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Zitnick, C.L., Batra, D., Parikh, D., 2015. VQA: Visual Question Answering, in: CVPR, pp. 2425–2433.

Arras, L., Montavon, G., Müller, K.R., Samek, W., 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis, in: EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA), pp. 159–168.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural Machine Translation by Jointly Learning to Align and Translate, in: ICLR.

Ben-younes, H., Cadene, R., Cord, M., Thome, N., 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. arXiv:1705.06676 .

Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W., 2018. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. IEEE Trans. Image Process. 27, 206–219.

Charikar, M., Chen, K., Farach-Colton, M., 2004. Finding frequent items in data streams. Theor. Comput. Sci. 312, 3–15.

Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M., 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding, in: EMNLP, pp. 457–468.

Gao, Y., Beijbom, O., Zhang, N., Darrell, T., 2016. Compact bilinear pooling, in: CVPR, pp. 317–326.

Glorot, X., Bengio, Y., 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks, in: AISTATS, pp. 249–256.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D., 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, in: CVPR, pp. 6904–6913.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, in: ICCV, pp. 1026–1034.

Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D., 2018. Pythia v0.1: The Winning Entry to the VQA Challenge 2018. ArXiv180709956 Cs arXiv:1807.09956.

Kim, J.H., On, K.W., Kim, J., Ha, J.W., Zhang, B.T., 2017. Hadamard Product for Low-Rank Bilinear Pooling, in: ICLR.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 .

Kumar, P.R., Varaiya, P., 2015. Stochastic Systems: Estimation, Identification, and Adaptive Control. SIAM.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft Coco: Common Objects in Context, in: ECCV, pp. 740–755.

Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering, in: NIPS, pp. 289–297.

Montavon, G., Samek, W., Müller, K.R., 2018. Methods for Interpreting and Understanding Deep Neural Networks. Digit. Signal Process. 73, 1–15.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., others, 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv:1602.06023 .

Nam, H., Ha, J.W., Kim, J., 2017. Dual Attention Networks for Multimodal Reasoning and Matching, in: CVPR, pp. 299–307.

Noh, H., Han, B., 2016. Training Recurrent Answering Units with Joint Loss Minimization for VQA. arXiv:1606.03647 .

Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global Vectors for Word Representation, in: EMNLP, pp. 1532–1543.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 .

Teney, D., Anderson, P., He, X., van den Hengel, A., 2017. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. arXiv:1708.02711 .

Tucker, L.R., 1966. Some Mathematical Notes on Three-Mode Factor Analysis. Psychometrika 31, 279–311.

Xiong, C., Merity, S., Socher, R., 2016. Dynamic Memory Networks for Visual and Textual Question Answering, in: ICML, pp. 2397–2406.

Yu, Z., Yu, J., Fan, J., Tao, D., 2017a. Multi-Modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. ArXiv170801471 Cs arXiv:1708.01471.

Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D., 2017b. Beyond Bilinear: Generalized Multi-Modal Factorized High-Order Pooling for Visual Question Answering. arXiv:1708.03619 .