Computer Vision and Image Understanding 185 (2019) 24–30

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

DRAU: Dual Recurrent Attention Units for Visual Question Answering ☆

Ahmed Osman *, Wojciech Samek *

Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, Berlin 10587, Germany

ARTICLE INFO

Communicated by D. Parikh

MSC:
68T45
68T50
68T30
68T10

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

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 perform two ablation studies to evaluate recurrent attention. First, we introduce a baseline VQA model with visual attention and test the performance difference between convolutional and recurrent attention on the VQA 2.0 dataset. Secondly, we design an architecture for VQA which utilizes dual (textual and visual) Recurrent Attention Units (RAUs). Using this model, we show the effect of all possible combinations of recurrent and convolutional dual attention. 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. Moreover, we experiment replacing attention mechanisms in state-of-the-art models with our RAUs and show increased performance.

1. Introduction

Although convolutional neural networks (CNNs) and 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 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 pertaining to certain local parts of the input. This motivation 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. According to the literature in human cognition, humans process cognitive attention both spatially and temporally (Rensink, 2000). Consequently, recurrent attention mechanisms come to mind. 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 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 competitive results compared to state-of-the-art models.

Our main contributions are the following:

- We introduce a novel approach to generate soft attention for VQA. To the best of our knowledge, this is the first attempt to generate attention maps using recurrent neural networks in the VQA domain.

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.cviu.2019.05.001.

* Corresponding author.
E-mail addresses: ahmed.osman@hhi.fraunhofer.de (A. Osman), wojciech.samek@hhi.fraunhofer.de (W. Samek).

https://doi.org/10.1016/j.cviu.2019.05.001
Received 24 October 2018; Received in revised form 16 March 2019; Accepted 12 May 2019
Available online 20 May 2019

- 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).
- We propose a network that utilizes two RAUs to co-attend the multi-modal inputs. We perform an ablation study to further test the effect of recurrent attention in our model. The results show a significant improvement when using both RAUs compared to using convolutional attention.
- RAUs 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 show that our network outperforms the VQA 2016 and 2017 challenge winners and performs close to the current state-of-the-art single models. Additionally, we provide qualitative results showing subjective improvements over the default attention used in most VQA models.

In Section 2, we review related work for recurrent attention and VQA methods. In Section 3, we break down the components of the DRAU network and explain the details of a RAU. In Section 4, we compare convolutional and recurrent attention in a baseline model, conduct ablation experiments using DRAU, and report the results of substituting attention mechanisms of state-of-the-art models with RAUs on the VQA 2.0 dataset (Goyal et al., 2017). Then we compare our model against the state-of-the-art on the VQA 1.0 (Antol et al., 2015) and 2.0 datasets. Furthermore, 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 5 and discuss future work.

2. Related work

This section discusses related recurrent attention mechanisms and common methods that have been explored in the past for VQA.

Recurrent attention. RNN-based attention models have been used outside of the VQA domain. Mnih et al. (2014) train a recurrent neural network model for digit recognition. Their model selects a sequence of regions of an image and processes the selected regions at high resolutions to save computational power. However, the whole architecture is a monolithic recurrent model that limits its use in existing VQA models since it cannot easily substitute VQA attention mechanisms. Moreover, the model is not differentiable and requires training with reinforcement learning which is highly inefficient to train in practice due to training instability (Lanctot et al., 2017) and complexity of reward function design (Paulus et al., 2017). Closer to our work, Homayounfar et al. (2018) uses a recurrent attention mechanism to attend to lane boundaries for traffic lane counting. There exists some differences compared to our attention unit. Their attention mechanism uses multi-step training to train the attention unit while RAUs do not require any attention-specific training. It is worth noting that Homayounfar et al. (2018) use a vanilla Convolutional RNN for attention. Furthermore, they sample the input feature maps from different scales to provide information from different granularities. Trying different recurrent architectures like GRU (Cho et al., 2014), Grid-LSTM (Kalchbrenner et al., 2015), and Conv-LSTM (Shi et al., 2015) and their effect on attention could be an interesting research direction in the future.

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. Benyounes 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).

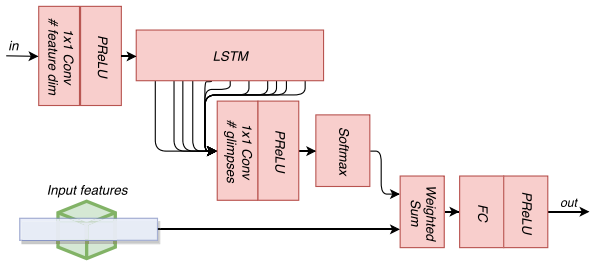

Fig. 1. Recurrent attention unit.

Attention-based. Xu and Saenko (2016) propose a memory network that predicts visual attention based on a dot product between image features and word embedding. They utilize a separate embedding that predicts visual "evidence" related to the question. The evidence embedding can be computed iteratively (2-hops) to collect more evidence. 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 guide each other in a co-dependent manner. Schwartz et al. (2017) use high-order correlations between the multimodal input and multiple-choice answers to guide the model's attention. It is worth noting that this model processes attention not only for both the question and image, but also for the answer. However, their VQA model utility is gravely weakened by their dependence on multiple-choice answer embedding which renders their model computationally infeasible for standard open-ended VQA tasks.

RNNs for VQA. Using 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 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 explicitly 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. All modules are annotated in Fig. 2 for reference.

3.1. Recurrent Attention Units (RAUs)

The RAU receives a multi-modal multi-channel representation of the inputs, X. To scale down the input representation, a RAU starts with a

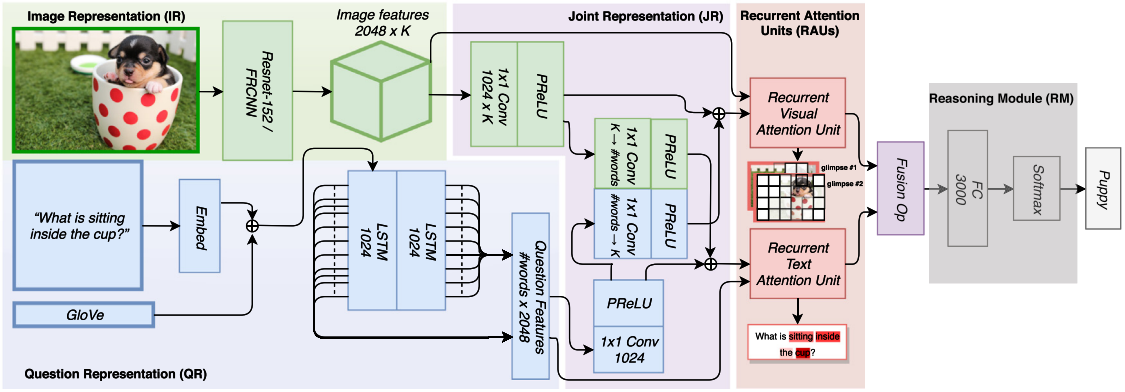

Fig. 2. The proposed network. ⊕ denotes concatenation.

1×1 convolution and PReLU activation (He et al., 2015):

$$c_a = \text{PReLU}\left(W_a\ X\right), \quad X \in \mathbb{R}^{K \times \phi} \tag{1}$$

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 pixels/number of visual objects or question length), and ϕ is the number of channels in the input.

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

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

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.

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

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 and a PReLU activation. Fig. 1 illustrates the structure of a RAU.

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

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 (IR). In our baseline and ablation studies, we use two types of image representations. First, a 152-layer "ResNet" pretrained CNN from He et al. (2015) to extract image features. Similar to Fukui et al. (2016) and Nam et al. (2017), we resize the images to 448×448 and extract the last layer before the final pooling layer (res5c) with size $2048 \times 14 \times 14$. 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 use Faster R-CNN features (Ren et al., 2015) with a fixed number of proposals per image ($K = 36$) for the DRAU model and its variants.

Question representation (QR). 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. Use of 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. This is also true for the question representation to a lesser degree (GloVe vectors are trained on co-occurrence statistics) Second, they are used to generate a common representation size. To obtain a joint representation (JR), we apply 1×1 convolutions followed by PReLU activations 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 the supplementary material.

3.4. Fusion operation

A fusion operation is used to merge the textual and visual branches. For DRAU, we use MCB (Fukui et al., 2016; Gao et al., 2016). We experiment with using element-wise multiplication (Hadamard product) in the supplementary document.

3.5. Reasoning module (RM)

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:

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

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:

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

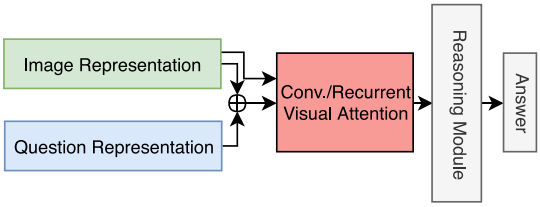

Fig. 3. Simple Net architecture. This model uses components from the DRAU model (Fig. 2) to be used as a baseline for attention evaluation. Multimodal features are concatenated (⊕) and fed directly to the attention mechanism.

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. 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:

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

We evaluate our results on the *validation*, *test-dev*, *test-std* splits of each dataset. We test on multiple splits due to the fact that submissions to the test server are limited (for test and test-dev splits). Thus, we mainly evaluate models on the validation split unless comparing results with the state-of-the-art.

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 models are trained with a small batch size of 32 for 400 K 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 randomly sample one answer at each training iteration.

4.1. Attention ablation studies

Convolutional versus recurrent attention. We compare using convolution against our recurrent attention in a simple baseline model. This baseline 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 a visual attention mechanism. The processed attention is fed to the reasoning module. We refer to this model as *Simple Net* (illustrated in Fig. 3). Simple Net was trained twice: once with convolutional attention and once with RVAU. To avoid any type of parameter advantage, both models were designed to have the same number of parameters approximately. Both baselines use the *train* split and Visual Genome (Krishna et al., 2017) for training, and were evaluated on the VQA 2.0 validation split. The results of Simple Net in Table 1 show a clear advantage of recurrent attention outperforming convolutional attention by over 4% absolute overall accuracy.

In second ablation study, we aim to examine the effect of different combinations of recurrent attention by training 4 different variants of the DRAU model that only differ in the type of attention used. The models are the following:

- Dual Convolution Attention (DCA): DRAU with dual convolutional attention, i.e. text att.(Convolution) – image att.(Convolution).
- DCA w/RVAU: text att.(Convolution) – image att.(RVAU).
- DCA w/RTAU: text att.(RTAU) – image att.(Convolution).
- DRAU: text att.(RTAU) – image att.(RVAU).

Similar to the baseline models, we match the number of parameters in the attention units such that all models have roughly the same number of parameters to ensure that differences are not from a higher number of parameters. All models were trained using Faster R-CNN image features on the *train* split. Additionally, we train each variant with 3 different initial seeds to show the effect of parameter initialization.

The results in Table 1 show the evaluation results of this ablation study on the VQA 2.0 validation split. We report the mean and standard deviation of the each model. Using RVAU improves the baseline DCA model by 1.48%. While swapping convolutional text attention with RTAU reduces performance by 0.66%, using both RTAU and RVAU in the same model improves the overall performance by almost 2%. This might indicate that multimodal recurrent text attention thrives with certain architectural designs since it improves performance in our DRAU model.

In conclusion, using RVAU significantly improves network accuracy. RTAU requires more careful use, but in an appropriate model, improves performance significantly.

4.2. Using RAUs in other models

To verify the effectiveness of the recurrent attention units, we replace the convolutional 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). 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. Note that the text attention in MFH is "self-attending" which means that the textual attention does not interact with visual content in this setting. This is different from our DRAU model where RTAU uses a joint representation of the question and image to predict the textual attention. This gives insight about the difference of performance between the ablation study and the modified MFH results. While the performance improvement might not look large for some models, it is consistent which shows that RAUs can reliably improve existing state-of-the-art models with different architectures.

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

In this section, we present the results of our model in the scope of state-of-the-art models. Although our model does not utilize some extra knowledge and performance optimization techniques (Visual Genome augmentation, model ensembles, hyperparameter tuning), we choose to show models that do so in order to present our results from a practical point of view.

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

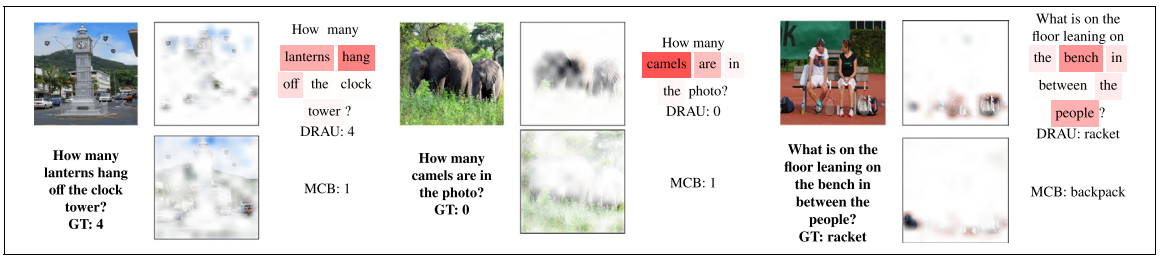

Fig. 4. DRAU vs. MCB Qualitative examples. Attention maps for both models shown. DRAU shows subjectively better attention map quality.

Table 1
Simple Net model and ablation study of the DRAU model results on the VQA 2.0 *validation* split. The number of trainable parameters are shown for each model. IF indicates the type of image features the model uses. VG indicates whether the method uses external data augmentation from the Visual Genome dataset.

VQA 2.0 validation split								
Model	# Parameters	IF	Fusion	VG	All	Y/N	Num.	Other
Simple Net (Conv. Visual Attn.)	37.4×10^9	ResNet	✗	✓	41.06	66.01	28.08	25.51
Simple Net w/RVAU	37.3×10^9	ResNet	✗	✓	45.12	66.24	28.48	33.46
DCA (Conv. Attn)	138.4×10^9	FRCNN	MCB	✗	59.42 ± 0.09	77.80	36.28	51.57
DCA w/RVAU	138.4×10^9	FRCNN	MCB	✗	60.90 ± 0.02	79.04	39.63	52.73
DCA w/RTAU	138.4×10^9	FRCNN	MCB	✗	58.76 ± 0.08	77.56	35.54	50.61
DRAU	138.4×10^9	FRCNN	MCB	✗	61.36 ± 0.02	79.74	40.03	53.03

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

VQA 2.0 Test-dev split				
Model	All	Y/N	Num.	Other
MCB (Fukui et al., 2016)[a]	61.96	78.41	38.81	53.23
MCB w/RVAU	62.33	77.31	40.12	54.64
MUTAN (Ben-younes et al., 2017)	62.36	79.06	38.95	53.46
MUTAN w/RVAU	62.45	79.33	39.48	53.28
VQA 2.0 validation split				
MFH (Yu et al., 2017b)	64.31	82.26	43.49	56.17
MFH w/RTAU	64.38	82.35	43.31	56.3

[a] http://www.visualqa.org/roe_2017.html.

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 (see Table 4). 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. Fig. 4 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.[2] 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 (illustrated in Fig. 2).

In Fig. 4, 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. The first example shows that DRAU attends to the right object (lanterns) both visually and textually. The model also attends to the cars' rear lights as possible "lanterns", but the text attention attends to the word "hang" which disqualifies the car lights and guides the model to count hanging lanterns. Furthermore, DRAU can predict non-existing object(s). The second example in Fig. 4 illustrates that DRAU is not easily fooled by counting whatever animal is present in the image but rather the "camels" that is needed to answer the question. This property also translates to questions that require relational reasoning. The third example in Fig. 4 demonstrates how well DRAU can attend to the relative location required to answer the question based on the textual and visual attention maps compared to MCB.

Attention quality. Fig. 5 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 leftmost 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

[2] 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

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

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

Model	N	WE	VG	Test-dev				Test-standard			
				All	Y/N	Num.	Other	All	Y/N	Num.	Other
VQATeam_MCB (Goyal et al., 2017)	1	✓	✓	61.96	78.41	38.81	53.23	62.27	78.82	38.28	53.36
UPMC-LIP6 (Ben-younes et al., 2017)	5	✓	✓	65.57	81.96	41.62	57.07	65.71	82.07	41.06	57.12
HDU-USYD-UNCC (Yu et al., 2017b)	9	✓	✓	68.02	84.39	45.76	59.14	68.09	84.5	45.39	59.01
Adelaide-Teney (Teney et al., 2017)	1	✓	✓	**65.32**	**81.82**	**44.21**	**56.05**	**65.67**	**82.20**	**43.90**	**56.26**
Adelaide-Teney (Anderson et al., 2017)	30	✓	✓	–	–	–	–	70.34	86.60	48.64	61.15
FAIR A-STAR (Jiang et al., 2018)	1	✓	✓	70.01	–	–	–	70.24	–	–	–
FAIR A-STAR (Jiang et al., 2018)	30	✓	✓	72.12	87.82	51.54	63.41	72.25	87.82	51.59	63.43
$DRAU_{FRCNN+MCB\ fusion}$	1	✓	✗	**66.45**	**82.85**	**44.78**	**57.4**	**66.85**	**83.35**	**44.37**	**57.63**

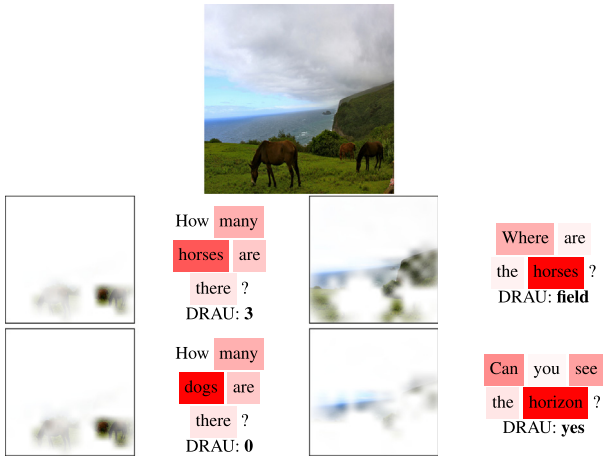

Fig. 5. Four real example results of our proposed model for a single random image. The visual attention, textual attention, and answer are shown. Even on the same image, our model shows rich reasoning capabilities for different question types. The first column shows that the model is able to do two-hop 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.

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 recurrent attention mechanism, 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. Using a simple VQA baseline, we have shown the performance advantage of recurrent attention compared to the traditional convolutional attention used in most VQA models. Furthermore, we performed an ablation study with all possible combinations of recurrent and convolutional attention. The results of the study indicate that recurrent attention can be very beneficial in dual attention VQA models. Then, 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 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.

In future work we will investigate implicit recurrent attention mechanism using recently proposed explanation methods (Arras et al., 2017; Montavon et al., 2018) and explore different recurrent models for attention (Kalchbrenner et al., 2015; Shi et al., 2015).

Acknowledgments

This work was supported by the Fraunhofer Society, Germany 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, Germany under Grant 01IS180371.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2019.05.001.

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 (1), 206–219.

Charikar, M., Chen, K., Farach-Colton, M., 2004. Finding frequent items in data streams. Theoret. Comput. Sci. 312 (1), 3–15.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation, arXiv:1406.1078 [cs, stat].

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.

Homayounfar, N., Ma, W., Lakshmikanth, S.K., Urtasun, R., 2018. Hierarchical recurrent attention networks for structured online maps. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3417–3426. http://dx.doi.org/10.1109/CVPR.2018.00360.

Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.,

Kalchbrenner, N., Danihelka, I., Graves, A., 2015. Grid Long Short-Term Memory, arXiv:1507.01526.

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.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D.A., et al., 2017. Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. 123 (1), 32–73.

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

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., Graepel, T., 2007. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning, arXiv:1711.00832 [cs].

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.

Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K., 2014. Recurrent models of visual attention. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. In: NIPS'14, MIT Press, Cambridge, MA, USA, pp. 2204–2212.

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., et al., 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.

Paulus, R., Xiong, C., Socher, R., 2017. A Deep Reinforced Model for Abstractive Summarization, arXiv:1705.04304 [cs].

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.

Rensink, R.A., 2000. The dynamic representation of scenes. Vis. Cogn. (ISSN: 1350-6285) 7 (1–3), 17–42. http://dx.doi.org/10.1080/135062800394667.

Schwartz, I., Schwing, A.G., Hazan, T., 2017. High-Order Attention Models for Visual Question Answering, arXiv:1711.04323 [cs].

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., Woo, W.-c., 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, arXiv:1506.04214 [cs].

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 (3), 279–311.

Xiong, C., Merity, S., Socher, R., 2016. Dynamic memory networks for visual and textual question answering. In: ICML. pp. 2397–2406.

Xu, H., Saenko, K., 2016. Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In: ECCV. pp. 451–466.

Yu, Z., Yu, J., Fan, J., Tao, D., 2017. Multi-Modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering, ArXiv:170801471 Cs.

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