Reciprocal Attention Fusion for Visual Question Answering

Moshiur R Farazi[1,2]
moshiur.farazi@anu.edu.au
Salman H Khan[2,1,3]
salman.khan@anu.edu.au

[1] Australian National University, Australia
[2] Data61 - CSIRO, Australia
[3] Inception Institute of AI, UAE

Abstract

Existing attention mechanisms either attend to local image-grid or object level features for Visual Question Answering (VQA). Motivated by the observation that questions can relate to both object instances and their parts, we propose a novel attention mechanism that jointly considers reciprocal relationships between the two levels of visual details. The bottom-up attention thus generated is further coalesced with the top-down information to only focus on the scene elements that are most relevant to a given question. Our design hierarchically fuses multi-modal information i.e., language, object- and grid-level features, through an efficient tensor decomposition scheme. The proposed model improves the state-of-the-art single model performances from 67.9% to 68.2% on VQAv1 and from 65.7% to 67.4% on VQAv2, demonstrating a significant boost.

1 Introduction

An AI agent equipped with visual question answering ability can respond to intelligent questions about a complex scene. This task bridges the gap between visual and language understanding to realize the longstanding goal of highly intelligent machine vision systems. Recent advances in automatic feature learning with deep neural networks allow joint processing of both visual and language modalities in a unified framework, leading to significant improvements on the challenging VQA problem [3, 12, 16, 20, 40].

To deduce the correct answer, an AI agent needs to correlate image and question information. A predominant focus in the existing efforts has remained on attending to local regions on the image-grid based on language input [15, 21, 26, 34, 35]. Since these regions do not necessarily correspond to representative scene elements (objects, attributes and actions), there exists a "semantic gap" in such attention mechanisms. To address this issue, Anderson *et al.* [1] proposed to work at the object level, where model attention is spread over a set of possible object locations. However, the object proposal set considered in this way is non-exhaustive and can miss important aspects of a scene. Furthermore, language questions can pertain to local details about objects parts and attributes, which are not encompassed by the object-level scene decomposition.

In this work, we propose to simultaneously attend to both low-level visual concepts as well as the high-level object based scene representation. Our intuition is based on the fact that

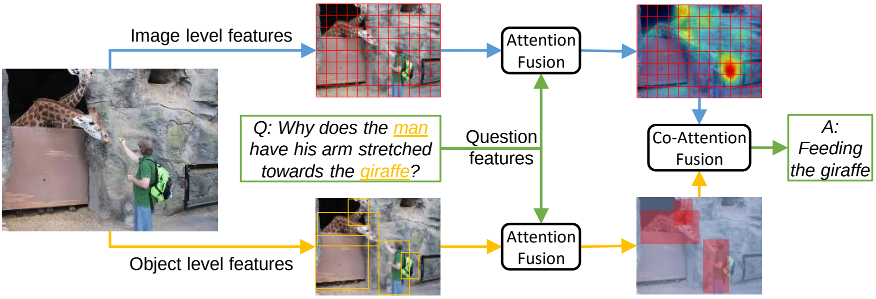


Figure 1: Applying attention to reciprocal visual features allow a VQA model to obtain the most relevant informations required to answer a given visual question.

the questions can be related to objects, object-parts and local attributes, therefore focusing on a single scene representation can degrade model capacity. To this end, we jointly attend to two reciprocal scene representations that encompass local information on the image-grid and the object-level features. The bottom-up attention thus generated is further combined with the top-down attention driven by the linguistic input. Our design draws inspiration from the human cognitive psychology, where attention mechanism is known to be a combination of both exogenous (bottom-up) and endogenous (top-down) factors [6, 9].

Given the multi-modal inputs, a critical requirement is to effectively model complex interactions between the multi-level bottom-up and top-down factors. For this purpose, we propose a multi-branch CNN architecture that hierarchically fuses visual and linguistic features by leveraging an efficient tensor decomposition mechanism [5, 30]. Our experiments and extensive ablative study proves that a language driven attention on both image-grid and object level representation allows a deep network to model the complex interaction between vision and language as our model outperforms the state-of-the-art models in VQA tasks.

In summary, this paper makes the following key contributions:

- A hierarchical architecture incorporating both the bottom-up and top-down factors pertaining to meaningful scene elements and their parts.
- Co-attention mechanism enhancing scene understanding by combining local image-grid and object-level visual cues.
- Extensive evaluation and ablation on both balanced and imbalanced versions of the large-scale VQA dataset achieving single model state-of-the-art performance in both.

2 Related Works

Deep Networks: Given the success of deep learning, one common approach to address the VQA problem is by generating image features using pretrained Convolutional Neural Networks (CNNs), e.g., VGGNet [27], ResNet [13], and language features using word-embeddings or Long Short-Term Memory (LSTM) [3]. After generating image and language features, some approaches train RNNs to generate top-K candidate answers and use a multi-class classifier to choose the best answer [3, 38, 40]. A number of attention mechanisms have been incorporated within deep networks to automatically focus on specific details in an image based on the given question [15, 21, 35]. Memory networks have also been incorporated in many top performing models [28, 33, 39] where the questions required the system to compare attributes or use a long reasoning chain. While robust features and memory modules help capture some aspects of the semantics present in the scene, modeling the complex interplay between image-grid and objects level features can complement the understanding of the rich scene semantics.

Attention Models: The incorporation of spatial attention on the image and/or the text features has been investigated to capture the most important parts required to answer a given question [15, 21, 26, 34, 35]. Different pooling methods have been used previously to compute the attention maps such as soft attention, bilinear pooling and tucker fusion [5, 11, 34]. All these techniques explore top-down attention and only focus on the image-grid. Different to these works, based on the observation that questions pertain to objects, their parts and attributes, we propose to work jointly at the spatial grid of image regions and the object-level. The closest to our work is [1], which attends to salient objects in an image for improved VQA. However, they ignore two key aspects of visual reasoning i.e., the image level visual features and an effective fusion mechanism to combine the bimodal interaction between visual and language features. Another recent effort [22] co-attends to both image regions and objects, but uses a simplistic fusion mechanism based on element-wise multiplication that is outperformed by our bilinear feature encoding. Besides, our multi-level attention mechanism effectively uses object features and scene context based on the natural language queries.

3 Methods

The VQA task requires an AI agent to generate a natural language response, given a visual (i.e. image, video) and natural language input (i.e. questions, parse). We formulate VQA task as a classification task, where the model predicts the correct answer ($\hat{a}$) from all possible answers for a given image ($\mathbf{v}$) and question ($\mathbf{q}$) pair:

$$\hat{a} = \operatorname*{argmax}_{a \in A} p(a|\mathbf{v}, \mathbf{q}; \theta), \tag{1}$$

where θ denotes the set of parameters used to predict the best answer from the set of all possible answers A.

Our proposed architecture to perform VQA task is illustrated in Figure 2. The key highlights of our proposed architecture include a hierarchical attention mechanism that focuses on complementary levels of scene details i.e., grid of image regions and object proposals. The relevant co-attended features are then fused together to perform final prediction. We name our model as the *'Reciprocal Attention Fusion'* because it simultaneously attends to two complementary scene representations i.e., image-grid and object proposals. Our experimental results demonstrate that both levels of scene details are reciprocal and reinforce each other to achieve the best single-model performance on challenging VQA task. Before elaborating on the hierarchical attention and feature fusion, we first discuss the joint feature embedding in Section 3.1.

3.1 Joint Feature Embedding

Let V be the collection of all visual features extracted from an image and Q be the language features extracted from the question. The objective of joint embedding is to learn the language feature representation $q = \chi(Q)$ and multilevel visual features $v_k = \zeta(V)$. These feature representations are used to encode the multilevel relationships between question and image which in turn is used to train the classifier to select the correct answer.

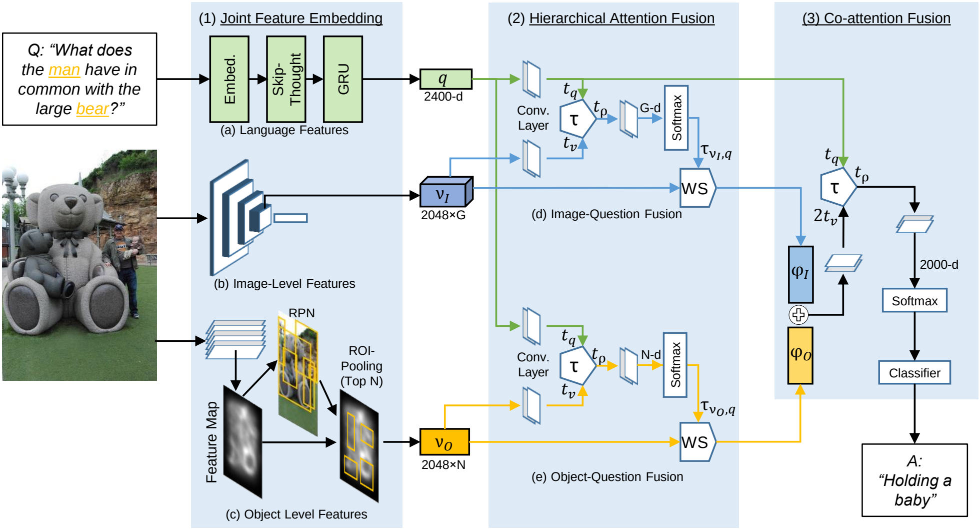


Figure 2: Given an image-question pair, our model employs (1) Joint Feature Embedding (Sec.3.1) to embed (a) Language Feature q, (b) Image-Level Feature v_I and (c) Object-Level Feature v_O. Further, these embeddings undergo (2) Hierarchical Attention Fusion (Sec.3.2) which consists of (d) Image-Question and (e) Object-Question Fusion followed by top-down attention. These multi-modal representations are combined together by (3) Co-attention Fusion (Sec.3.3) that predicts an answer for the given Image-Question pair. Overall, the proposed model attends to complementary levels of scene details and fuses multi-modal information to predict highly accurate answers.

Multilevel visual features: The multilevel visual embedding v_k consists of image level features v_I and object level features v_O. Our model employs ResNeXt [32] to obtain image level features, $v_I \in \mathbb{R}^{n_v \times G}$ by taking the output of convolution layer before the final pooling layer, where G denotes the number of spatial grid locations of the extracted visual feature with n_v dimensions. This convolution layer retains the spatial information of the original image and enable the model to apply attention on the image-grid. On the other hand, our model employs object detectors to localize object instances and pass them through another deep CNN to generate object level features $v_O \in \mathbb{R}^{n_v \times N}$ for N object proposals. We use Faster R-CNN [25] with ResNet-101 [13] backbone and pretrain the object detector on ImageNet [8] and again retrain it on Visual Genome Dataset [20] with class label and attribute features similar to [1].

Bottom-up (BU) Attention: In order to focus on the most relevant features, two bottom-up attention mechanisms are applied during multilevel feature extraction. The image-grid attention is generated using ResNeXt [32] pretrained on ImageNet [8] to obtain $v_I \in \mathbb{R}^{2048 \times G}$, which represents 2048 dimensional features vectors for $G = 14 \times 14$ image-grid over the visual input. The size and scale of the image-grid can be changed by using different CNN architecture or taking the output of a different convolutional layer to generate a different sized BU attention. Meanwhile, object proposals are generated in a bottom up fashion to encode object level visual features v_O. We select a total of top $N = 36$ object proposals whose $n_v = 2048$ dimensional feature vectors are obtained from the ROI pooling layer in the Region Proposal Network.

Language features: To represent the questions embedding in an end-to-end framework, GRUs [7] are used in a manner similar to [5, 10]. The words in questions are encoded using one-hot-vector representation and embedded into vector space by using a word embedding matrix. The embedded word vectors are fed to the GRU with n_q units initialized with pre-trained Skip-thought Vector model [19]. The output of the GRU is fine-tuned to get the language feature embedding $q = \{q_i \ : i \in [1, n_q]\}$ where $n_q = 2400$. The language feature embedding is used to further refine the spatial visual features (i.e. image-grid and object level) by incorporating top-down attention discussed in Section 3.2.

3.2 Hierarchical Attention Fusion

The hierarchical attention mechanism takes spatial visual features v_I, v_O and language feature q as input and a learns multi-modal representation W to predict answer embedding ρ. This step can be formulated as an outer product of the multi-modal representation, visual and language embeddings as follows:

$$\rho = W \times_1 \ q \ \times_2 \ v, \tag{2}$$

where, $\times_n$ denotes n-mode tensor-matrix product. However, this approach has some serious practical limitations in terms of learnable parameters for $W \in \mathbb{R}^{n_v \times n_q \times n_\rho}$ as the visual and language feature are very high dimensional, which results in huge computational and memory requirements. To counter this problem, our model employs a multi-modal fusion operation $\tau(v, q)$ to encode the relationships between these two modalities, which is discussed next.

Multi-modal Fusion: Multi-modal fusion aims to reduce the number of free parameters in tensor $W \in \mathbb{R}^{n_v \times n_q \times n_\rho}$ for a fully parameterized VQA bilinear model. Our model achieves this by using Tucker Decomposition [30] which is a special case of higher-order principal component analysis to express W as a core tensor T_c multiplied by a matrix along input mode. The decomposed tensors are fused in a manner similar to [5] that encompass the multi-modal relationship between language and vision domain. The tensor W can be approximated as:

$$W \approx T_c \times_1 T_q \times_2 T_v \times_3 T_\rho = [\![T_c; T_q, T_v, T_\rho]\!] \tag{3}$$

where $T_v \in \mathbb{R}^{n_v \times t_v}$, $T_q \in \mathbb{R}^{n_q \times t_q}$ and $T_\rho \in \mathbb{R}^{n_\rho \times t_\rho}$ are factor matrices similar to principal components along each input and output embeddings and $T_c \in \mathbb{R}^{t_v \times t_q \times t_\rho}$ is the core tensor which encapsulates interactions between the factor matrices. The notation $[\![\cdot]\!]$ represents the shorthand for Tucker decomposition. In practice, the decomposed version of W is significantly smaller number of parameters than the original tensor [4].

After reducing the parameter complexity of W with tucker decomposition, the fully parametrized outer product representation in Eq. 2 can be rewritten as:

$$\rho = T_c \ \times_1 \ \tilde{q} \ \times_2 \ \tilde{v} \ \times_3 \ T_\rho \tag{4}$$

where $\tilde{v} = v^\intercal \ T_v \in \mathbb{R}^{t_v}$ and $\tilde{q} = q^\intercal \ T_q \in \mathbb{R}^{t_q}$. We define a prediction space $\rho = \tau^\intercal \ T_\rho \in \mathbb{R}^{n_\rho}$ where the multi-modal fusion τ is:

$$\tau = T_c \ \times_1 \ \tilde{q} \ \times_2 \ \tilde{v} \ \in \mathbb{R}^{t_\rho} \tag{5}$$

The Tucker decomposition allows our model to decompose W into a core tensor T_c and three matrices. The first two matrices, T_q and T_v project the question and visual embeddings to lower t_q and t_v dimensional space that learns to model the multi-modal interaction and projects the resulting output to t_ρ dimensional vector. We set the input projections dimension to $t_q = t_v = 310$ and output projection dimension as $t_\rho = 510$. The input and output tensor projection dimensions determine the complexity of the model and the degree of multi-modal interaction which in turn affects the performance of the model. These values are set empirically by testing them on VQAv1 validation dataset. It has been reported in the literature [6, 10] that applying nonlinearity to the input feature embeddings improve performance of multi-modal fusion. Therefore, we encode $\tilde{v}$ and $\tilde{q}$ with tanh nonlinearity during fusion. The output of the multi-modal fusion $\tau \in \mathbb{R}^{t_\rho}$ passes through convolution and softmax layers to create $1 \times G$ and $1 \times N$ dimensional representation for image-question and object-question embedding respectively. Thus, by employing hierarchical attention fusion, we embed question with spatial visual features to generate image-question $\tau_{v_I,q} \in \mathbb{R}^{1\times G}$ and object-question $\tau_{v_O,q} \in \mathbb{R}^{1\times N}$ embedding.

Top-down (TD) Attention The image level and object level features are used alongside image-question and object-question embeddings to generate an attention distribution over spatial grid and object proposals respectively. We take weighted sum (WS) of the spatial visual features (i.e. v_I and v_O) vectors using the attention weights (i.e. $\tau_{v_I,q}$ and $\tau_{v_O,q}$) to generate φ_I and φ_O which are top-down attended visual features,

$$\varphi_I = \sum_i^G \tau_{v_I,q}^i \, v_I^i \quad \text{and} \quad \varphi_O = \sum_i^N \tau_{v_O,q}^i \, v_O^i. \tag{6}$$

3.3 Co-attention Fusion

The attended image-question and object-question visual features represent a combination of visual and language features that are most important to generate an answer for a given question. We concatenate these two bimodal representations to create the final visual-question embedding $\varphi = \varphi_I \oplus \varphi_O$. The visual-question embedding, φ and original question embedding q again undergo same multi-modal fusion as Eq. 5. The only difference is now $t_\varphi = 2 \times t_v$ as our model uses two glimpse attention which was found to yield better results [6, 10, 17]. The output of the final fusion is then passed on to the classifier that predicts the best answer $\hat{a}$ from the answer dictionary A given question $\mathbf{q}$ and visual input $\mathbf{v}$.

4 Experiments

4.1 Dataset

We perform experiments on **VQAv1** [3] and **VQAv2** [12] both of which are large scale VQA datasets. VQAv1 contains over 200K images from the COCO dataset with 610K natural language open-ended questions. VQAv2 [12] contains almost twice as many question for the same number of images. VQAv2 has a balanced image-question pair to mitigate the language bias that allows a more realistic evaluation protocol. **Visual Genome** is another larger scale dataset that has image question pair with dense annotation of objects, attributes [20]. We train a pretrained faster RCNN model (on ImageNet) again on Visual Genome dataset with class and attribute labels to extract object level features from the input image.

Methods	Test-dev				Test-standard			
	Y/N	No.	Other	All	Y/N	No.	Other	All
RAF (Ours)	**85.9**	**41.3**	**58.7**	**68.0**	**85.8**	41.4	58.9	**68.2**
ReasonNet[14]	-	-	-	-	84.0	38.7	**60.4**	67.9
MFB+CoAtt+Glove [37]	85.0	39.7	57.4	66.8	85.0	39.5	57.4	66.9
Dual-MFA [22]	83.6	40.2	56.8	66.0	83.4	40.4	56.9	66.1
MLB+VG [17]	84.1	38.0	54.9	65.8	-	-	-	-
MCB+Att+GloVe [10]	82.3	37.2	57.4	65.4	-	-	-	-
MLAN [36]	81.8	41.2	56.7	65.3	81.3	**41.9**	56.5	65.2
MUTAN [5][1]	84.8	37.7	54.9	65.2	-	-	-	-
DAN (ResNet) [24]	83.0	39.1	53.9	64.3	82.8	38.1	54.0	64.2
HieCoAtt [21]	79.7	38.7	51.7	61.8	-	-	-	62.1
A+C+K+LSTM[31]	81.0	38.4	45.2	59.2	81.1	37.1	45.8	59.4
VQA LSTM Q+I [3]	80.5	36.8	43.1	57.8	80.6	36.5	43.7	58.2
SAN[35]	79.3	36.6	46.1	58.7	-	-	-	58.9
AYN [23]	78.4	36.4	46.3	58.4	78.2	37.1	45.8	59.4
NMN [2]	81.2	38.0	44.0	58.6	-	-	-	58.7
DMN+ [33]	60.3	80.5	48.3	56.8	-	-	-	60.4
iBowling [38]	76.5	35.0	42.6	55.7	76.8	35.0	42.6	55.9

Table 1: Comparison of the state-of-the-art methods with our single model performance on VQAv1.0 test-dev and test-standard server.

4.2 VQA Model Architecture

Question Feature Embedding: Our model embeds the question features by first generating the questions and answer dictionary from training and validation set of the VQA datasets. We make the question and answers lower case, remove punctuation and perform other standard preprocessing steps before tokenizing the words, and representing them into one-hot vector representation. As mentioned in Section 3.1, these question embeddings are fed to GRUs pretrained with Skip-thoughts [19] model that generates 2400-d language feature embeddings for the given question. When experimenting with VQAv1 and VQAv2, we parse questions respectively from training and validation sets to create the question vocabulary.

Answer Encoding: We formulate the VQA task as a classification task. We create an answer dictionary from the training data and select the top 2000 answers as the different classes. We pass the output of the final fusion layer through a convolutional layer that outputs a 2000d vector. This vector is passed through the classifier to predict $\hat{a}$.

We use Adam solver [18] with base learning rate of 10^{-4} and batch size of 512 for our experiments. We keep the training parameters same for all our experiments. We use NVidia Tesla P100 (SXM2) GPUs to train our models and report our experimentation results on VQAv1 [3] and VQAv2 [12] dataset representing 1500 GPU hours of computation.

5 Results

We evaluate the proposed models' performance on the VQA test servers which ensures blind evaluation on the VQAv1 [3] and v2 [12] test sets (i.e. test-dev, test-standard) following the

[1] Single model performance is evaluated using their publicly available code.

Methods	Test-dev				Test-standard			
	Y/N	No.	Other	All	Y/N	No.	Other	All
RAF (Ours)	**84.1**	**44.9**	**57.8**	**67.2**	**84.2**	**44.4**	**58.0**	**67.4**
BU, adaptive K [29]	81.8	44.2	56.1	65.3	82.2	43.9	56.3	65.7
MFB [37]	-	-	-	64.9	-	-	-	-
ResonNet[14]	-	-	-	-	78.9	42.0	57.4	64.6
MUTAN[5][2]	80.7	39.4	53.7	63.2	80.9	38.6	54.0	63.5
MCB [10, 12]	-	-	-	-	77.4	36.7	51.2	59.1
HieCoAtt [12, 21]	-	-	-	-	71.8	36.5	46.3	54.6
Language only[12]	-	-	-	-	67.1	31.6	27.4	44.3
Common answer[12]	-	-	-	-	61.2	0.4	1.8	26.0

Table 2: Comparison of the state-of-the-art methods with our single model performance on VQAv2.0 test-dev and test-standard server.

VQA benchmark evaluation approach. The accuracy y of the predicted answer $\hat{a}$ is calculated with the following formulation:

$$y = min\left(\frac{\text{\# of humans answered } \hat{a}}{3}, 1\right) \tag{7}$$

which means that answer provided by the model is given 100% accuracy if at least 3 human annotators who helped create the VQA dataset gave the exact answer.

In Table 1, we report VQAv1 test-dev and test-standard accuracies for our proposed RAF model and compare it with other single models found in literature. Remarkably, our model outperforms all other models in the overall accuracy. We report a significant performance boost of 1.2% on the test-dev set and 0.3% on the test-standard set. It is to be noted that using multiple ensembles and data augmentation with complementary training in Visual Genome QA pairs can increase the accuracy performance of the VQA models. For instance, MCB [10], MLB [17], MUTAN [5] and MFB [37] employ similar model ensemble consisting of 7,7,5 and 7 models respectively, and report overall 66.5, 66.9, 67.4 and 69.2 on the test-standard set. It is interesting to note that except for MFB (7) all other ensemble models are $\sim$ 1% less than our reported single model performance. We do not ensemble our model or use data augmentation with complementary dataset as it makes the best results irreproducible and most of the models in the literature do not adopt this strategy.

We also evaluate our model on VQAv2 test-standard dataset and compare it with state-of-the-art single model performance in Table 2, illustrating that our model surpasses the closest method [29] in all question categories and overall by a significant margin of 1.7%. The bottom up, adaptive-k[29] is the same model whose 30-ensemble version [1] reports currently the best performance among on VQAv2 test-standard dataset. This indicates our models superior capability to interpret and incorporate multi-modal relationships for visual reasoning.

In summary, our model achieves state-of-the-art performance on both VQAv1 and VQAv2 dataset which affirms the robustness of our model against language bias without the need of data augmentation or the use of ensemble model. We also show qualitative results in Fig. 4 to demonstrate the efficacy and complimentary nature of attention focused on image-grid and object proposals.

[2] Performance on VQAv2 is evaluated from their publicly available repository.

Cat.	Methods	Val-set
I	RAF-I(ResNet)	53.9
	HieCoAtt [12, 21]	54.6
	RAF-I(ResNeXt)	58.0
	MCB [10, 12]	59.1
	MUTAN [5]	60.1
II	Up-Down[1]	63.2
	RAF-O(ResNet)	63.9
III	RAF-IO(ResNet-ResNet)	64.0
	RAF-IO(ResNeXt-ResNet)	64.2

Table 3: Ablation Study on VQAv2 val-set.

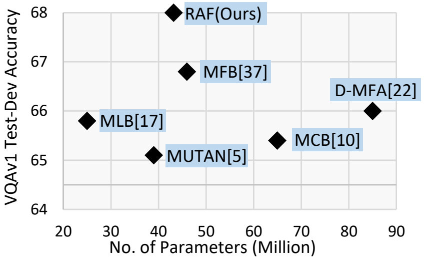


Figure 3: Accuracy vs. Complexity (no. of parameters) comparison.

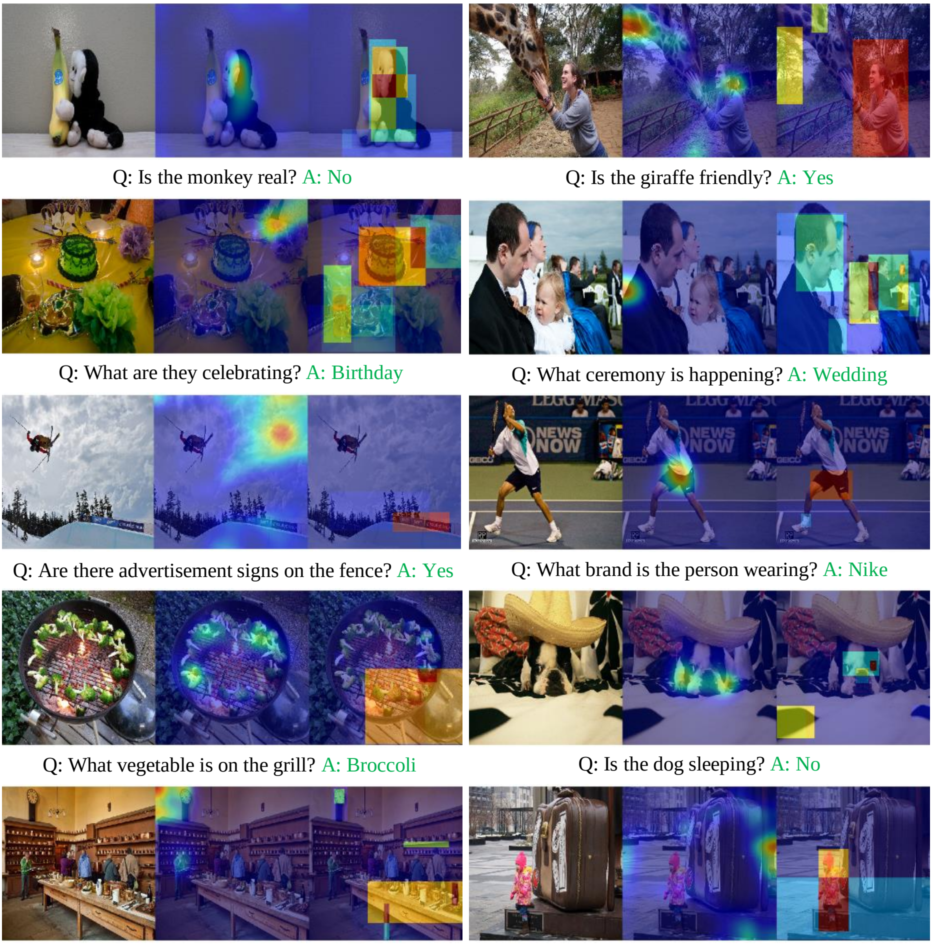
5.1 Ablation Study

We perform an extensive ablation study of the proposed model on VQAv2 [12] validation dataset and compare it with the best performing model in Table 3. This ablation study helps to better understand the contribution of different components of our model towards the overall performance on the VQA task. The objective of this ablation study is to show that when the language features are combined with image- grid and object level visual features, the accuracy of the high level visual reasoning task (i.e. VQA) increases in contrast to only combining language with image or object level features. The models reported in Category I in Table 3 use only image level features extracted with deep CNNs and we compare RAF-I which is a variant of our proposed RAF architecture only using image level features. We observe RAF-I achieve comparable performance in this category. In Category II, RAF-O model extracts only 36 object level features but outperforms the models in Category I. [1] also used only object level features and this variant of our model achieves comparable performance to that model. When we combine image and object level features together in Category III, we observe that the best results are obtained. This proves our hypothesis that the questions relate to both objects, object parts and local attributes, which should be attended for jointly an improved VQA performance.

The recent Dual-MFA [22] model also uses complementary image and object-level features. In contrast, our model uses more efficient bimodal attention fusion mechanism and exhibit robustness on balanced VQAv2 [12] dataset. We also study the accuracy vs. complexity (no. of parameters) trade off in Fig. 3 on VQAv1 test-dev set as most of the bilinear models do not report performance on VQAv2. Remarkably, our RAF model achieves significant performance boost over Dual-MFA (66% to 68%) with around half the complexity.

6 Conclusion

We build our proposed model based on the hypotheses that multi-level visual features and associated attention can provide an AI agent additional information pertinent for deep visual understanding. As VQA is a standard measure of image understanding and visual reasoning, we propose a VQA model that learns to capture the bimodal feature representation from visual and language domain. To this end, we employ state of the art CNN architectures to obtain visual features for local regions on the image-grid and object proposals. Based on these feature encodings, we develop a hierarchical co-attention scheme that learns the mutual rela-



Q: Is the monkey real? A: No

Q: Is the giraffe friendly? A: Yes

Q: What are they celebrating? A: Birthday

Q: What ceremony is happening? A: Wedding

Q: Are there advertisement signs on the fence? A: Yes

Q: What brand is the person wearing? A: Nike

Q: What vegetable is on the grill? A: Broccoli

Q: Is the dog sleeping? A: No

Q: What time is on the clock? A: 1:50

Q: Is this the girl's suitcase? A: Yes

Figure 4: Qualitative results of the proposed Reciprocal Attention Fusion mechanism for Visual Question Answering. Given a question and an image (columns: $1,4$), attention based on image-grid (columns: $2,5$) and object proposals (columns: $3,6$) is shown above. Correct and incorrect answers are shown in green and red, respectively. Remarkably, the two attention levels provide complementary information about localized regions and objects that in turn help in obtaining the correct answer (rows: $1,2,3,4$). In some failure cases of our technique, ambiguous attention maps lead to incorrect predictions (row: 5).

tionships between objects, object-parts and given questions to predict the best response. We validate our hypotheses by evaluating the proposed model on two large scale VQA dataset servers followed by an extensive ablation study reporting state-of-the art performance.

Acknowledgements

The authors would like to thank CSIRO Scientific Computing team, especially Peter Campbell and Ondrej Hlinka, for their assistance in optimizing the work-flow on GPU clusters.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[3] 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*, pages 2425–2433, 2015.

[4] Brett W Bader and Tamara G Kolda. Efficient matlab computations with sparse and factored tensors. *SIAM Journal on Scientific Computing*, 30(1):205–231, 2007.

[5] Hedi Ben-Younes, Rémi Cadène, Nicolas Thome, and Matthieu Cord. Mutan: Multimodal tucker fusion for visual question answering. *ICCV*, 2017. URL http://arxiv.org/abs/1705.06676.

[6] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2013.

[7] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[9] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1): 193–222, 1995.

[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. *arXiv preprint arXiv:1606.01847*, 2016.

[11] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326, 2016.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *arXiv preprint arXiv:1612.00837*, 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] Ilija Ilievski and Jiashi Feng. Multimodal learning and reasoning for visual question answering. In *Advances in Neural Information Processing Systems*, pages 551–562, 2017.

[15] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *European Conference on Computer Vision*, pages 727–739. Springer, 2016.

[16] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, 2016.

[17] 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] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016.

[21] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[22] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. *arXiv preprint arXiv:1711.06794*, 2017.

[23] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A deep learning approach to visual question answering. *International Journal of Computer Vision*, 125(1-3):110–135, 2017.

[24] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[26] 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*, pages 4613–4621, 2016.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[29] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017.

[30] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[31] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 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*, pages 4622–4630, 2016.

[32] Saining Xie, Ross Girshick, Piotr DollÃąr, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[33] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *arXiv*, 1603, 2016.

[34] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[35] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[36] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *Conf. on Computer Vision and Pattern Recognition*, 2017.

[37] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[38] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.

[39] Yuke Zhu, Joseph J Lim, and Li Fei-Fei. Knowledge acquisition for visual question answering via iterative querying.

[40] 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*, pages 4995–5004, 2016.