Global Journal of Computer Science and Technology: D
Neural & Artificial Intelligence
Volume 17 Issue 1 Version 1.0 Year 2017
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Multimodal Attention in Recurrent Neural Networks for Visual Question Answering

By Lorena Kodra & Elinda Kajo Meçe
Polytechnic University of Tirana

Abstract- Visual Question Answering (VQA) is a task for evaluating image scene understanding abilities and shortcomings and also measuring machine intelligence in the visual domain. Given an image and a natural question about the image, the system must ground the question into the image and return an accurate answer in a natural language. A lot of progress has been done to address the challenges of this task by combining latest advances in image representation and natural language processing. Several recently proposed solutions include attention mechanisms designed to support "reasoning". These mechanisms allow models to focus on specific part of the input in order to generate the answer and improve its accuracy. In this paper we present a novel LSTM architecture for VQA that uses multimodal attention to focus over specific parts of the image and also on specific question words to generate the answer. We evaluate our model on the VQA dataset and demonstrate that it performs better than state of the art. We also make a qualitative analysis of the results and show the abilities and shortcomings of our model.

Keywords: visual question answering (VQA), multimodal attention mechanism, convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory (LSTM).

GJCST-D Classification: F.1.1

MULTIMODALATTENTIONINRECURRENTNEURALNETWORKSFORVISUALQUESTIONANSWERING

Strictly as per the compliance and regulations of:

Open Association of Research Society, USA
Research | Diversity | Ethics

© 2017. Lorena Kodra & Elinda Kajo Meçe. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Multimodal Attention in Recurrent Neural Networks for Visual Question Answering

Lorena Kodra [α] & Elinda Kajo Meçe [σ]

Abstract- Visual Question Answering (VQA) is a task for evaluating image scene understanding abilities and shortcomings and also measuring machine intelligence in the visual domain. Given an image and a natural question about the image, the system must ground the question into the image and return an accurate answer in a natural language. A lot of progress has been done to address the challenges of this task by combining latest advances in image representation and natural language processing. Several recently proposed solutions include attention mechanisms designed to support "reasoning". These mechanisms allow models to focus on specific part of the input in order to generate the answer and improve its accuracy. In this paper we present a novel LSTM architecture for VQA that uses multimodal attention to focus over specific parts of the image and also on specific question words to generate the answer. We evaluate our model on the VQA dataset and demonstrate that it performs better than state of the art. We also make a qualitative analysis of the results and show the abilities and shortcomings of our model.

Keywords: visual question answering (VQA), multimodal attention mechanism, convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory (LSTM).

I. Introduction

Visual question answering has emerged as a multidisciplinary research problem at the intersection of artificial intelligence, natural language processing and computer vision. This task requires an intelligent system to answer a question about an image. Both question and answer are in a natural language. The system must ground the question into the image; hence it requires a deep understanding of the image scene. It is a complex research problem and puts a lot of focus on artificial intelligence, and especially the inference process needed to generate the answer because different question types (e.g. color, number, location, etc.) require different answers. There are also questions requiring some commonsense reasoning such as "Do the people look happy?". With the advancement of image representation, language processing and deep learning, the most promising solutions use a combination of Convolutional Neural Networks (CNNs) to process the image and extract image features and Recurrent Neural Networks (RNNs) to model word sequences. The output of each network is later combined in order to generate the final answer as output [11]. One of the latest concepts introduced in VQA is the attention mechanism. It enables the model to focus on specific parts of the input in order to infer the answer. Recently, the idea of dual attention has been introduced in VQA [6], [14]. It allows the model to focus on specific question words, as well as specific image regions before inferring the answer.

In this paper we propose a novel architecture for long short-term memory (LSTM) networks, which includes image attention and question attention. We refer to the combined attention as multimodal attention. The standard LSTM architecture [8] has been modified in order to include multimodal attention. We evaluate our proposed solution on the VQA [9] dataset and show that it performs better compared with state of the art models. The main contributions of our work are as follows:

- We propose a novel LSTM model with multimodal attention.
- Our model uses image attention guided by the correlation between the current context and image regions, as well textual attention guided by the relevance and importance of distinct question words in relation to the whole question.
- We evaluate our proposed model on the VQA dataset [9].
- We analyze the results qualitatively and show the abilities and shortcomings of our model.

The rest of the paper is organized as follows: In section 2 we describe related work in this research area. Section 3 describes in detail our proposed model. In section 4 we describe the experimental setup and show the evaluation results. Finally in section 5 we discuss the results and conclusions.

II. Related Work

a) Visual Question Answering

Deep learning based approaches have demonstrated competitive performance in the VQA task [21], [26], [24] [25], [23]. For processing the image, most approaches extract features from images using CNNs which have shown to work best in representing images [1]. On the sentence side, most approaches use RNNs to model word sequences [22], [18], [19], [14], [7], [20], [21], [15]. Other approaches include Bag-of-Words question embedding [17] or multilayer

Author α σ: Polytechnic University of Tirana.
e-mails: lorena.kodra@gmail.com, ekajo@fti.edu.al

© 2017 Global Journals Inc. (US)

perceptrons (MLP) [16] to predict the answer. All approaches treat question answering as a classification problem and learn a softmax classifier to generate the answer.

Several mechanisms and techniques have been proposed for the process of question answering. The authors in [15] use a dynamic parameter prediction RNN whose parameters are determined adaptively based on input questions. In this way the system reasons differently for each question. The motivation behind this approach is the fact that different questions require different types and levels of understanding of an image to find correct answers. Another proposed model [20] is a neural reasoner based on a MLP that is able to update the question representation iteratively by inferring image information. The model achieves this by selecting image regions relevant to the question and learns to give the correct answer by interacting it with supporting facts through multiple reasoning layers. With this technique, it is possible to make questions more specific than the original ones focusing on important image information automatically. The authors in [22] propose a multimodal compact bilinear pooling method to combine multimodal features extracted from a CNN for the image and a LSTM for the question. This mechanism reduces the dimensionality of the joint representation of the image and question and produces a model with less parameters and hence easier to train. Another alternative are multimodal systems composed of CNN and RNN that are trained end-to-end to extract question information, visual representation, store the linguistic context of the answer and combine this information into generating a relevant answer to a free language question [21].

b) Neural Attention Mechanisms

Attention mechanisms allow neural network models to use a question to selectively focus on specific inputs. This idea has been recently successfully implemented in various areas such as image captioning [2], [27], [28], [29], [30], neural machine translation [3], [4], [5], and visual question answering [6], [7], [14], [19], [18], [17]. In the case of visual question answering, attention mechanisms allow models to focus on specific parts of visual or textual inputs that are relevant to the context of the answer, at each step of the process. Instead of looking at the whole image, visual attention models selectively pay attention to specific regions in an image to extract image features that are relevant to the question as well as reduce the amount of information to process. On the other hand, textual attention mechanisms find semantic or syntactic input-output alignments under an encoder-decoder framework.

In order to tackle the VQA task, several works perform image attention multiple times in a stacked manner. In [18] the authors propose a stacked attention network which queries the image multiple times to infer the answer progressively. It uses semantic representation of a question as a query to identify the regions of the image that are related to the answer. The authors in [17] propose a multi-hop visual attention scheme. In the first hop, it aligns words to image regions while in the second hop it uses the entire question representation to obtain image attention maps.

The idea of incorporating attention into the standard RNN architecture has been explored in [7] and [19]. Xiong et al. [19] augment dynamic memory networks with a new input fusion layer that uses bidirectional gated recurrent units (GRU). They also propose an attention based GRU to retrieve the answer. Zhu et al [7] add visual attention to the standard LSTM architecture for pointing and grounded QA. However, the models mentioned above model only visual attention and do not model textual attention. Hyeonseob et al [6] propose dual attention networks which attend to specific regions in images and words in text through multiple steps and gather essential information from both modalities. Lu et al [14] propose hierarchical co-attention that jointly reasons about visual attention and question attention. Following this line of research and the idea explored in [7] and [19], we propose a novel LSTM architecture by incorporating visual and question attention in the gates of the LSTM network. Each step of the attention distribution depends on the previous LSTM state and the current focus on specific question words and image regions.

III. Multimodal Attention Model

The idea of using multimodal attention for the task of VQA has been recently explored in [6] and [14]. The main difference between these models and ours is that we include attention as a component of each LSTM gate as illustrated in Fig. 2. The intuition behind this is that by simultaneously focusing on specific image regions and specific question words, the model can decide how to change its current state and what answer word to generate next. Using the actual context (previous LSTM hidden state) helps to guide attention correctly and improve answer accuracy. We choose LSTM models because they have shown to achieve state-of-the-art results in several sequence processing tasks [30], [32] including VQA [24], [21], [25].

The input of our model is an image of size 224x224 pixels and a question comprised of a variable-length set of words. Each word is first transformed into its one-hot representation, a column vector the size of the vocabulary where there is a single one at the index of the token in the vocabulary. Each word is then embedded into a real-valued word vector $Q = \{q_j | q_j \in R^D, j = 1, \dots, N\}$ where N is the number of question words, D is the dimensionality of the embedding space and$Q \in R^{DxT}$ for the image representation we extract the activations from the last fully connected layer (fc7) of

© 2017 Global Journals Inc. (US)

VGG-16, a pretrained CNN model [31]. Given the image I, this model transforms it into a 4096-dimensional feature representation. We also learn the embedding of the input image where 4096-dimensional image features are transformed into a D dimensional embedding space denoted by $V = \{v_i | v_i \in R^D, i = 1, \ldots, M\}$ where *M* is the number of image features, *D* is the dimensionality of the embedding space and$V \in R^{DxM}$. Both embedding modalities are 512 dimensional and are learnt end-to-end.

We treat the image as the first input token and the image embedding vectors are fed one by one to the LSTM model. Afterwards we feed the tokens of the question embedding. The update rules of our LSTM model are:

$$i_t = \sigma(W_{iv} v_t + W_{ih} h_{t-1} + W_{it}^{txt} a_t^{txt} + W_{it}^{img} a_t^{img} + b_i) \quad (1)$$

$$f_t = \sigma\left(W_{fv} v_t + W_{fh} h_{t-1} + W_{ft}^{txt} a_t^{txt} + W_{ft}^{img} a_t^{img} + b_f\right) \quad (2)$$

$$o_t = \sigma(W_{ov} v_t + W_{oh} h_{t-1} + W_{ot}^{txt} a_t^{txt} + W_{ot}^{img} a_t^{img} + b_o) \quad (3)$$

$$g_t = \tanh(W_{cv} v_t + W_{ch} h_{t-1} + W_{ct}^{txt} a_t^{txt} + W_{ct}^{img} a_t^{img} + b_c) \quad (4)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ g_t \quad (5)$$

$$h_t = o_t \circ \tanh(c_t) \quad (6)$$

Where σ is the sigmoid activation function and ◦ is the element-wise product.

Different from [7] which only use image attention, we integrate also textual (question) attention in the LSTM gates. The image and textual attention features are represented by the term a_t^{img} and a_t^{txt}respectively. These features are learnt end-to-end. The authors in [14] use the dot product of question and image representation to produce an affinity matrix. This matrix is then added to image or question representation and used to guide both the textual and image attention respectively. Different from their approach, we use the previous LSTM hidden state (h_{t-1}) and question or image representation to guide question and image attention respectively. We calculate image attention as follows:

$$l_t^{img} = tanh(W_{lh}^{img} h_{t-1} + W_{lq}^{img} CNN(I) + b_{img}) \quad (7)$$

$$r_t^{img} = softmax\,(W_{img}^T l_t^{img}) \quad (8)$$

$$a_t^{img} = r_t^{img} CNN(I) \quad (9)$$

Following [7], for generating the image attention we use the fourth convolutional layer of VGG-16 [31]. This layer returns a 196 512-dimensional convolutional feature map of image *I* represented by the term *CNN (I)* in equations (7) and (9). The termr_t^{img} represents the attention probabilities of each image region. Based on these attention probabilities the image attention vector is calculated as the weighted sum of the attention probabilities. The attention terma_t^{img} is a 196-dimensional vector that decides the contribution of each image feature at the t-th step. The W and b coefficients are learnable parameters.

The question attention is calculated as follows:

$$l_t^{txt} = \tanh(W_{lh}^{txt} h_{t-1} + W_{lq}^{txt} Q + b_{txt}) \quad (10)$$

$$r_t^{txt} = softmax\,(W_{txt}^T l_t^{txt}) \quad (11)$$

$$a_t^{txt} = r_t^{txt} Q \quad (12)$$

The term r_t^{txt}represents the attention probabilities of each question word. Based on these attention probabilities the question attention vector is calculated as the weighted sum of the attention probabilities. The attention term a_t^{txt} is a N-dimensional vector that decides the contribution of each word at the t-th step. Fig.1 illustrates the dataflow for generating each attention modality.

In each step, the LSTM generates new image and textual attention vectors based on the current context (previous LSTM hidden state) and the respective embeddings. The intuition behind this is that the model might need to focus on different parts of the image or different question words in order to generate the next answer word. The authors in [6] introduce accumulative attention to their model to keep track of the attended parts and guide future attention. An accumulative attention may suffer from the introduction of errors in earlier steps that might be propagated into future steps. In contrast, our model generates independent attention each step and does not suffer from this kind of problem. As in [7] the question words are feed one by one until reaching the end token of the question sequence. The model generates attention and leverages it together with the question and input image to generate the answer (Fig.2). We treat question answering as a classification task and use a softmax classifier to generate the answer. During training we also feed the ground truth answer tokens into the model and maximize their log-likelihood.

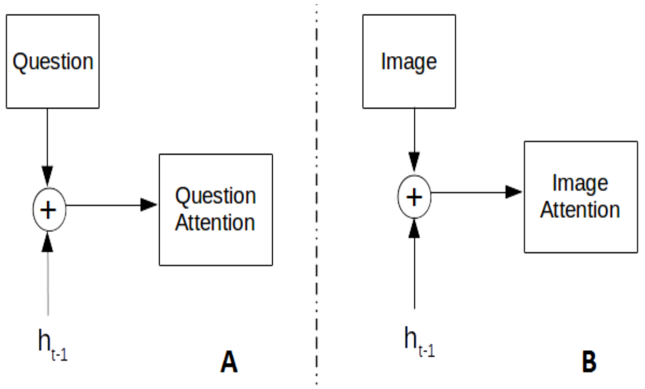


Fig. 1: Attention generation. (A) At each step, question attention is generated by combining the current context (previous LSTM cell hidden state h_{t-1}) and question representation. (B) At each step, image attention is generated by combining the current context (previous LSTM hidden stateh_{t-1}) and image representation.

© 2017 Global Journals Inc. (US)

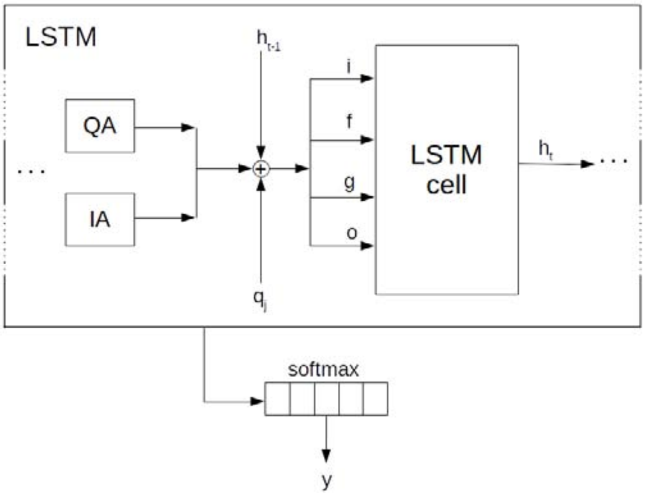


Fig. 2: Data flow for LSTM cells inside the LSTM network. Question attention (QA), image attention (IA), previous LSTM state (h_{t-1}) and current question token (q_j) are used in each LSTM cell gate to generate the context (h_t) that will be used by the next LSTM cell. A *softmax* classifier is used at the end as the output of the LSTM network to generate one by one each answer word *y*

IV. Experiments and Results

In this section we describe model implementation details, evaluation results and analyze them quantitatively. The results of the evaluation are shown in section IV.C.

a) Datasets and Evaluation Metrics

We evaluate the proposed model on the Visual Question Answering version 1 (VQA-v1) dataset [9]. The VQA dataset was used because it is the largest and most complex dataset for the visual question answering task. VQA-v1 was selected for fairness of comparison with other models.

The VQA-v1 dataset was constructed using the Microsoft COCO dataset [33] which contains 123,287 training/validation images and 81,434 test images. Each image has several related questions and each question is answered by multiple people. This dataset contains 248,349 training questions, 121,512 validation questions, and 244,302 testing questions. The total number of images, questions and answers are as follows: 204,721 COCO images (all of current train/val/test) 614,163 questions, 6,141,630 ground truth answers, 1,842,489 plausible answers.

Since we formulate VQA as a classification task, classification accuracy is used to measure the performance of our model and to compare it with state-of-the-art models.

b) Setup and Implementation Details

We use Torch [10] to develop our model. Before training, all questions are normalized to lower case and the question marks are removed. The model is initialized with Xavier initialization [13] except for the embeddings which used random uniform initialization. We train the model with Adam update rule [12] with a global learning rate of 10^{-4}. We train the model with back propagation and use cross-entropy as the loss function. During testing we select the candidate answer with the largest log-likehood. We set batch size to 128 and train for up to 256 epochs with early stopping if the validation accuracy has not improved in the last 5 epochs. The dimension of the LSTM network is 512 for all experiments. All embeddings are vectors of size 512. We apply dropout with probability 0.5 on each layer and also gradient clipping to regularize the training process. We rescale the images to 224 × 224. Following [7] we use the activations from the last fully connected layer (fc) of VGG-16 [31] to learn the image embeddings and the activations from the fourth convolutional layer of the same CNN for calculating image attention.

c) Quantitative Results and Analysis

The VQA dataset includes two test scenarios: open-ended and multiple-choice. We evaluate our model on both scenarios. The full release (V1.0) of this dataset contains a train set and a validation set. Following standard practice, we choose the top 1,000 most frequent answers in train and validation sets as candidate answers. We only keep the examples whose answers belong to these 1,000 answers as training data, which constitutes 86.54% of the train and validation answers. The question vocabulary size is 7477 with the word frequency of at least three.

Table 1: Open-ended results on VQA test set compared with state-of-the-art: accuracy in %. We denote with “-“the cases with lack of data

Method	Test-dev				Test-standard			
	Y/N	Num	Other	All	Y/N	Num	Other	All
HieCo[14]	79.5	38.7	48.3	60.1	-	-	-	-
D-NMN[16]	80.5	37.4	43.1	57.9	-	-	-	58
SAN(2, LSTM)[18]	79.3	36.6	46.1	58.7	-	-	-	58.9
SMem-VQA[17]	80.87	37.32	43.12	57.99	80.8	37.53	43.48	56.24
Ours-MAVQA	81.9	37.51	49.1	61.08	81.8	37.5	49.05	61

© 2017 Global Journals Inc. (US)

We compare the performance of our model with current state-of-the-art models and show the experimental results on free-form answers in Table 1. We also report the accuracy in each category to show the strength and weakness of our model.

We notice that all models reach top accuracy for the Yes/No questions. This is justified by the fact that there are only two possible answers and the possibility of giving an incorrect answer is decreased. We can see that our approach performs better and improves the state of the art from 60.1% (HieCo [14]) to 61.08% (Ours-MA VQA) in test-dev. In test-standard the accuracy is improved by 4.76% from 56.24% (SMem-VQA [17]) to 61% (Ours-MA VQA). For Yes/No and Other questions we achieve an improvement of 1.03% and 0.8% respectively. This indicates that our model is able to attend better and benefits from the multimodal attention and the independence of each attention modality from the other and from previous attention steps. For Number questions the counting ability of our model is weakened. This indicates that our model doesn't attend correctly and having a correlated attention like in HieCo [14] helps in achieving better performance at counting objects. We observe that all models perform worst on Number questions. This is justified by the fact that the ability to count objects is still a pervasive computer vision problem.

Table 2 shows results from multiple-choice question. The data was available for comparison only with HieCo [14]. We also report the accuracy in each category to show the strength and weakness of our model. We notice that models perform better for multiple choice questions. This comes from the fact that they exploit and tune to the biases in each of the answer options. However it is debatable whether this is indicative of progress because in realistic applications, answer options are not known beforehand. From Table 2 we see that our multimodal approach performs better and improves the state of the art by 1.48% from 64.6% to 66.08%. We also notice that our model performs 1.03% better than state of the art on Yes/No questions. As in the case of free-form answers, the models reach top accuracy for Yes/No questions and perform the worst on numbering questions. For Number questions, as in free-form answers, having a correlated attention, like the model in HieCo [14], helps the model attend the image better and achieve a higher accuracy.

Table 2: Multiple choice results on VQA test set compared with state-of-the-art: accuracy in %

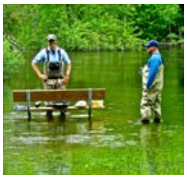
Method	Test-dev			
	Y/N	Num	Other	All
HieCo[14]	79.5	39.8	57.4	64.6
Ours-MA VQA	82.1	38.68	58.61	66.08

d) Qualitative Analysis

In order to gain a better understanding on the behavior and limitations of our model we analyzed the answers generated using multimodal attention. Each question requires different type and level of understanding and attention in order to find the correct answer. Table 3, 4 and 5 show some examples for each question type on the VQA dataset.

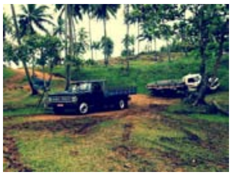
Table 3: Answer examples on the VQA dataset for Yes/No questions. We denote questions with "Q", model answers with "A", and ground truth with "GT"

1. Q: Is the horse eating?
A: No
GT: No



2. Q: Is there a bench?
A: No
GT: Yes

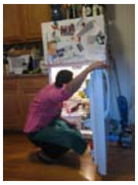
3. Q: Is there a red sandal here?
A: Yes
GT: Yes



4. Q: Is the road paved?
A: Yes
GT: No

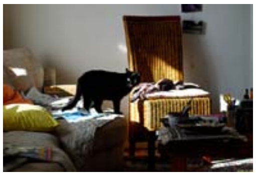
5. Q: Are there lights on in the two buildings?
A: Yes
GT: Yes

6. Q: Are these zebra confined?
A: Yes
GT: Yes



7. Q: Is the kitchen cluttered?
A: No
GT: Yes

8. Q: Is the beach crowded?
A: Yes
GT: Yes



9. Q: Are there any scissors in this picture?
A: No
GT: Yes

10. Q: Is this animal in a zoo?
A: No
GT: No

We noticed the following characteristics in our model for *Yes/No* questions:

+ The model correctly attends, identifies and infers about objects in the foreground and their characteristics. (e.g. images 1,3).
+ Difficulty inferring about background objects. The model cannot identify correctly the objects in the background. The focusing attention is weakened for this kind of objects and the model cannot infer correctly about them (e.g. images 9, 4).
+ Difficulty identifying objects that appear incomplete in the image. Attention is weakened for this kind of objects and the model cannot infer correctly about them (e.g. images 2, 7).

Table 4: Answer examples on the VQA dataset for Number questions. We denote questions with "Q", model answers with "A", and ground truth with "GT"

1. Q: How many street signs are shown?

A: Two
GT: Four

2. Q: How many horses are there?

A: Two
GT: Two

3. Q: How many bikes are there?

A: Two
GT: One

4. Q: How many people do you see?

A: Two
GT: None

5. Q: How many people are there?

A: None
GT: None

6. Q: How many yellow planes are there?

A: One
GT: Three

7. Q: How many giraffes are in this picture?

A: Two
GT: Two

8. Q: How many jets are there?

A: One
GT: Two

9. Q: How many birds?

A: Two
GT: None

10. Q: How many buses are there?

A: One
GT: One

We noticed the following characteristics in our model for *Number* questions:

+ The model correctly attends and identifies objects in foreground and their characteristics (e.g. images 1, 7, 10).
+ The model correctly attends and identifies objects in background that do not appear blended in the image but are clearly distinct from each-other. (e.g. image 2).
+ Difficulty identifying objects in background. Attention is weakened for background objects and the model cannot infer and count them correctly (e.g. images 3, 4, 9).
+ Difficulty differentiating objects in background that appear blended with each-other. Attention is weakened in this case and the model cannot infer and count them correctly (e.g. images 1, 8).

Table 5: Answer examples on the VQA dataset for other questions. We denote questions with "Q", model answers with "A", and ground truth with "GT"

1. Q: Who is with the giraffes?

A: No one
GT: No one

2. Q: What is the woman in front sitting on?

A: A bicycle
GT: A bicycle

3. Q: What color are the walls?

A: yellow
GT: yellow

4. Q: Where are the engines?

A: in the middle of the plane
GT: Behind the wings toward the back of the fuselage.

© 2017 Global Journals Inc. (US)

5. Q: What has a purple border?

A: The window

GT: The box truck.

6. Q: What kind of flooring is in the room?

A: White tile.
GT: Gray marble tile.

7. Q: What angle was the picture taken from?

A: From the left side of the sign
GT: Below the sign, looking up at it

8. Q: Where was this photo taken?

A: At a tennis court
GT: At a tennis court

9. Q: How is the food served?

A: In a basket
GT: In a basket

10. Q: Where was this photo taken?

A: At the park
GT: At the park

For the *Other* type of question our model has the following behavior:

+ Correctly attends, identifies and infers about objects in foreground and their characteristics. The results show that attention works correctly for this kind of objects (e.g. images 1, 9, 2).
+ Correctly attends, identifies and infers about background objects that are clearly distinct from each-other and from foreground (e.g. images 3, 8, 10).
– Difficulty inferring about objects that appear blended with each-other (e.g. images 4, 5, 6).

V. Conclusions

In this paper we proposed a novel LSTM architecture that uses multimodal attention for the task of visual question answering. Our model leverages both textual and visual attention simultaneously in order to identify question entities and ground them in the image. It learns to answer questions by generating independent visual and textual attention over the input. We evaluated our model on the VQA dataset and results show that it performs better than current state of the art. This indicates that integrating multimodal attention inside the LSTM architecture helps improving answer accuracy. Results also show that having independent attention modalities helps with overall accuracy and with questions of type other than counting. We analyzed the answers qualitatively and results show that our model is able to use multimodal attention correctly to: 1) Attend, identify and infer about foreground objects and their characteristics 2) Attend, identify and infer about background objects that are distinct from each-other and from foreground. Our attentive model has also some limitations like: 1) Difficulty inferring about incomplete objects, 2) Difficulty inferring about objects that appear blended with each-other and/or foreground/background, 3) Difficulty inferring about background objects that are not distinct from each-other and from foreground. These difficulties also weaken the counting ability of our model. These problems are indicative of the need to improve the attention mechanisms and solving them is subject to future work. Future research directions also include introducing common sense knowledge into our model and leveraging it to improve answer accuracy.

References Références Referencias

1. K. He, X. Zhang, Sh. Ren, and J. Su, "Deep residual learning for image recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
2. L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," In CVPR, 2016.
3. J. B. Delbrouck, S. Dupont, "Multimodal compact bilinear pooling for multimodal neural machine translation," In ICLR, 2017.
4. P-Y. Huang, F. Liu, Sz-R. Shiang, J. Oh, and C. Dyer, "Attention-based multimodal neural machine translation," In Proceedings of the First Conference on Machine Translation, 2016.
5. O. Caglayan, W. Aransa, Y. Wang, M. Masana, M. García-Martínez, F. Bougares, L. Barrault, and J. van de Weijer, "Does multimodality help human and machine for translation and image captioning?" arXiv preprint arXiv: 1605.09186, 2016.
6. N. Hyeonseob, H. Jung-Woo, K. Jeonghee, "Dual Attention Networks for Multimodal Reasoning and Matching," arXiv: 1611.00471, 2017.
7. Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," In CVPR, 2016.
8. S. Hochreiter and J. Schmidhuber, "Long short-term memory. Neural computation," 9(8): 1735–1780, 1997.
9. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," arXiv preprint arXiv: 1505.00468, 2015.

© 2017 Global Journals Inc. (US)

10. R. Collobert, K. Kavukcuoglu, and C. Farabet, “Torch7: A matlab-like environment for machine learning,” BigLearn, NIPS Workshop, 2011.
11. A.K Gupta, “Survey of Visual Question Answering: Datasets and Techniques,” arXiv: 1705.03865, 2017.
12. *Kingma and J. Ba. Adam, “A method for stochastic optimization,” arXiv preprint arXiv: 1412.6980, 2014.*
13. X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feed forward neural networks,” in International conference on artificial intelligence and statistics, pp. 249–256, 2010.
14. J. Lu, J. Yang, Dh. Batra, and D. Parikh, “Hierarchical Question-Image Co-Attention for Visual Question Answering,” 30th Conference on Neural Information Processing Systems, NIPS, 2016.
15. H. Noh, P. HongsuckSeo, and B. Han, “Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction,” 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
16. J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Learning to Compose Neural Networks for Question Answering,” Proceedings of NAACL-HLT, pp. 1545–1554, 2016.
17. H. Xu, and K. Saenko, “Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering,” European Conference on Computer Vision (ECCV), 2016.
18. Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked Attention Networks for Image Question Answering,” 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
19. C. Xiong, S. Merity, and R. Socher, “Dynamic Memory Networks for Visual and Textual Question Answering,” International Conference on Machine Learning, (ICML), 2016.
20. R. Li, and J. Jia, “Visual Question Answering with Question Representation Update (QRU),” 30th Conference on Neural Information Processing Systems (NIPS), 2016.
21. M. Malinowski, M. Rohrbach, and M. Fritz, “Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images,” IEEE International Conference on Computer Vision (ICCV), 2015.
22. A. Fukui, D. Huk Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding,” Empirical Methods in Natural Language Processing (EMNLP), 2016.
23. M. Ren, R. Kiros, and R. S. Zemel, “Exploring models and data for image question answering,” In NIPS, 2015.
24. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: visual question answering,” In ICCV, 2015.
25. H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, “Are you talking to a machine? Data set and methods for multilingual image question answering,” In NIPS, 2015.
26. L. Ma, Z. Lu, and H. Li, “Learning to answer questions from image using convolutional neural network,” In AAAI, 2016.
27. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” In CVPR, pp. 3156–3164, 2015.
28. K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” arXiv preprint arXiv:1502.03044, 2015.
29. H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al, “From captions to visual concepts and back,” In CVPR, pages 1473–1482, 2015.
30. A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” In CVPR, pp. 3128–3137, 2015.
31. K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” ICLR, 2014.
32. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” In CVPR, 2015.
33. T-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Lawrence Zitnick, “Microsoft COCO: Common objects in context,” European conference on computer vision (ECCV), pp. 740-755, 2014.

© 2017 Global Journals Inc. (US)