Improving Attention Effect for Visual Question Answering

Yazhi Gao*
Language Technology Institute
Carnegie Mellon University
Pittsburgh, PA 15213
yazhig@andrew.cmu.edu

Weidong Yuan*
Language Technology Institute
Carnegie Mellon University
Pittsburgh, PA 15213
weidongy@andrew.cmu.edu

Abstract

Visual question answering(VQA) has recently been a very popular topic for multi-modality reasoning. As an AI-complete task, VQA is to answer a natural language query about an image[1]. Deep learning models have been well applied to this task[2, 3, 4]. Especially, attention mechanism[5] has been heavily exploited to test its capacity to enhance the multi-modal relevance within this problem. Initially people only focus on visual attention on image and recently researchers use both visual and text attention to improve VQA models[6]. In this project, we explore beyond the co-attention model which jointly learn visual and question attention to present multi-glimpse co-attention and stacked co-attention models which are built on the vanilla co-attention model[7]. Our stacked co-attention model improves the performance on the VQA test-dev from 61.8% to 62.08% and the multi-glimpse co-attention model improves the benchmark to 62.32%. Also for the test-std, the stacked co-attention model improves the accuracy performance from 62.06% to 62.26% and the multi-glimpse co-attention model improves the accuracy to 62.49%.

1 Introduction

As deep neural networks have been demonstrating impressive impact improvement over natural language processing and computer vision field, people have been investing deep learning to tackle more complex problems in reasoning related field, especially visual question answering. Visual Question Answering is to form a natural language sentence relevant to a language query to an image. By helping people understanding queries about a given image, this research track can bring significant impacts in real-life applications.

When solving vision and text or even multi-modal problems such as VQA and image captioning[6], people use attention mechanism to guide neural models to generate the importance distribution of context embeddings with respect to a query input of asked questions or previously generated text. This attention mechanism is earlier introduced in machine translation field[5]. It improves model performance by discovering more relevant representations learning for vision and language problems.

Attention mechanism has become an extensive research in vision and language learning tasks. Even pure attention-based architecture has been remarkably powerful at translation system in Google production[5]. Stacked and bi-directional attention [4, 6] has been proposed to further address the more important representations because stacked attention fine-tunes attention maps to

*both authors contributed equally to this work.

10707 Class Project (2017).

better distribution and co-attention also deals with text distribution. Multi-glimpse(multi-headed) attention[8] also proved to capture better multi-modality representation by sampling multiple attention map candidates. To fully testify the powerfulness of various attention mechanism, we explore the mixture of the attention mechanism mentioned above in our work to introduce new architectures that embrace the techniques above.

In this report, we present the following two new architectures which combine attention mechanisms on a vanilla co-attention model(HierCoAtt[6]) to improve the visual question answering model performance.

Stacked co-attention: By leveraging bi-directional co-attention for both question and image as well as stacked attention to arrive at more precise attention distributions.

Multi-glimpse co-attention: Sampling alternative attention distributions on both images and text questions to fuse multiple view on the multi-modal representations.

Overall, the contributions of our work are:

- We propose two new architectures which are extended from the HierCoAtt[6] model. We explore multi-glimpse and stacked attention mechanism, which are described in 3.3.2 and 3.3.1.
- We compared multiple VQA architectures and discard complex and minimally helpful text feature modules in our architecture to simplify the model design.
- Finally we evaluate our models on the VQA dataset open-ended track to test their capacity. Along with improved performance benchmark, we present qualitative study on what our attention mechanism have learned in the task.

2 Related Work

Two typical VQA baseline models are bag-of-words and LSTM models[1]. These architectures use image features from pre-trained VGGnet or resent and process the question embeddings by bag-of-words or LSTM approaches. They set a baseline for a typical VQA solution. However, VQA is a frequently revolutionized field because some simple yet powerful models outperform big neural structures with very complex and full-of-trick design. The iBOWIMG model [2] fine-tunes a bag-of-words model to match and even beat some state-of-the-art VQA works back then. Simply adding up the word embeddings and feed the concatenation of image and word features to a fully connected classifier. By training the model carefully, this small model only slightly deviates from state-of-the-art result from much larger model with millions of parameters. The authors show that simple fully connected layers' weights even capture attention distribution over the image locations.

But iBOWIMG's performance is still capped for its simple structure cannot further extract relevance between question query and the images. For advanced models, attention are introduced. Stacked attention network(SAN)[4] use multiple hop of attention generation to produce a more precise attention map over the image to generate attended image features. SAN has great performance over multiple VQA related dataset. This technique is adopted in one of our model.

Afterwards, question attention is incorporated to generate better features[6]. By using bilinear methods, the model extract question-guided visual attentions and also image-guided question attentions. The co-attention mechanism combined with hierarchical text feature extraction sets a better benchmark than stacked attentions. One of the drawbacks of HierCoAtt is its complex text feature processing pipeline called Question Hierarchy. The model used multiple level text features from word, phrase and sentence to better preserve semantics in the question embedding. However, as the iBOWIMG and many other models uses proves the embedding from word-level representation is powerful enough to model the question text, we consider remove the question hierarchy from the model setting based on that its parameter addition cost is huge while the performance gain is limited by the author's ablation study over the model architecture.

There is also another type of attention mechanism which is multi-headed(Multi-glimpse)attention[9].

The Google model[9] generates multiple attention maps from different parameter initialization. The multiple glimpses of the image can act as complementary visual guidance for each other on the image. Also many engineering tricks are introduced to train a model with multiple components well in this work. The multi-glimpse attention is also adopted in our model which ultimately outperform the HierCoAtt baseline by the most performance boost.

3 Method

Visual question answering is currently modeled as classification task. We are learning answer with highest conditional probability on an image V and a question Q:

$$argmax_A P(A|V,Q) \tag{1}$$

Instead of a sentence, we search for the best answer word $\mathcal{A}$ in the vocabulary $\mathcal{D}$:

$$argmax_{\mathcal{A}\in\mathcal{D}} P(\mathcal{A}|V,Q) \tag{2}$$

Usually V is learned from computer vision architecture such as convolutional neural network. Q is learned from neural language models or sequence learning models such as recurrent neural network.

Our approaches are based on the hierarchical co-attention method[6]. We designed two networks that respectively utilize multi-glimpse and stacked attention mechanism. Both of them generate attention feature map and weighted sum of image and question embeddings with respect to the attention distribution. The attended question and image embeddings are then concatenated and fed to the multilayer classifier in the end.

3.1 Question Features

While the original work of co-attention network use hierarchical question features of various n-grams, many successful work use simpler word-level representation to form the sentence representations and the paper's ablation study shows the hierarchical feature provides marginal improvement for the model. Thus we want to focus on precisely attended image features or multi-glimpse attention as our key to model improvement rather than a complicated way of pooling convnet features which significantly increase the number of parameters. For our question embeddings, we are using simple LSTM generated hidden states at each sequence token location.

Given a question input of length T, $\mathbf{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_T\}$, we transform the question into vector space by a word embedding table so that every word is a vector $\mathbf{Q}_i^w \in R^h$. The question is embedded into $\mathbf{Q}^w = \{\mathbf{Q}_1^w, \mathbf{Q}_2^w, ..., \mathbf{Q}_T^w\}$. We feed the embeddings into LSTM and get the LSTM hidden states at all time-step as the question features $\mathbf{Q}^{lstm} \in R^{d1\times T}$. We use Q as notation for the later section as LSTM question features.

3.2 Image Features

We use pre-trained resnet-152 model[9] as our image feature generator. We fix the parameter of the resnet-152 model to generate $V \in R^{d1\times N}$. While the original HierCoAtt has two variants of image feature generator: VGG[7] and resnet, we choose resnet because it proves to better generalize about image semantics and is widely used in recently published VQA research.

3.3 Co-attention

In the vanilla HierCoAtt model, two co-attention generation methods are proposed: parallel and alternating co-attention generation. The parallel generation uses a bilinear model to generate attention maps for image and question at the same time while the alternating generation is to iteratively generate one kind of attention at a time. In our implementation, we will use the reportedly better co-attention generation mechanism - parallel co-attention.

Given the image feature map $V \in R^{d1\times N}$ (N is the number of spatial location of the features) and the question representation $Q \in R^{d2\times T}$, we calculate the affinity matrix $C \in R^{T\times N}$ which stores the attention information:

$$C = \tanh(Q^T W_b V) \tag{3}$$

To predict the image and question attention, we generate hidden features H^v, H^q and further get the normalized attention distributions $a^v \in R^N, a^q \in R^T$:

$$H^v = \tanh(W_v V + (W_q Q)C), H^q = \tanh(W_q Q + (W_v V)C^T) \tag{4}$$

$$a^v = softmax((w_{hv})^T H^v), a^q = softmax((w_{hq})^T H^q) \tag{5}$$

where $W^v \in R^{k \times d1}, W^q \in R^{k \times d2}, w_{hv} \in R^k, w_{hq} \in R^k$ are the weights parameters.

Then we sum the question and image embeddings over the spatial locations weighted by its attention distribution to get the final question and images features to feed to the classifier:

$$\hat{v} = \sum_{n=1}^{N} a_n^v v_n, \hat{q} = \sum_{t=1}^{T} a_t^q q_t \tag{6}$$

The HierCoAtt is illustrated in Figure 1

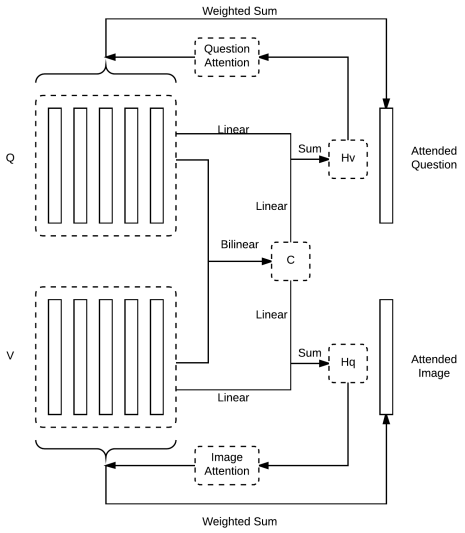


Figure 1: HierCoAtt Model

3.3.1 Multi-glimpse Co-attention

We add the multi-glimpse mechanism[8] to our model, instead of applying one attention, we apply multiple attention maps to the question and image feature maps and then do concatenation on the attended question and image embeddings which are the sums of the original embedding weighted by attention distribution at spatial locations.

We rely on different linear transformation on C to generate attention features H_v and H_q and on the different attention features to generate distinct attention maps. In our paper, we only explore 2 glimpses of attention because [8] has reported glimpses of 3 or more do not give meaningful improvement to the model. The formula for computing image attention will become:

$$\begin{aligned} H_c^v &= \tanh(W_c^v V + W_c^q QC), a_c^v = softmax((w_c^{hv})^T H_c^v) \\ \hat{v}_c &= \sum_{n=1}^{N} (a_c^v)_n v_n, \hat{v} = concat(\hat{v}_1, \hat{v}_2, ..., \hat{v}_C) \end{aligned} \tag{7}$$

where $c \in 1, 2, 3, ..., C$ indicating the number of the glimpse.
We can apply the same operation to question attention and get the question feature vector $\hat{q}_c$. The model is illustrated in Figure 2

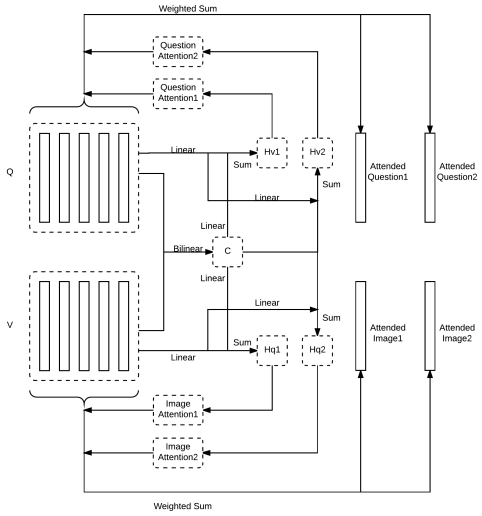


Figure 2: multi-glimpse co-attention model for 2 glimpses

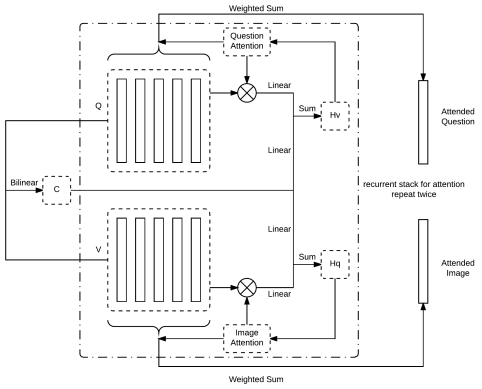


Figure 3: stacked co-attention model[2]

3.3.2 Stacked Co-attention

Another innovation we propose is to incorporate stacked attention when generating question and images attention maps. When producing the hidden feature for attention map, we mask the image and question with last attention we generated to arrive at better attention. The attention map is initialized as a uniform distribution over the spatial locations for images and questions. We set a stack of 2 in our model which distills better attention from the initial generation. The following are the derivations for stacked co-attention generation:

$$H_t^v = \tanh(W_v((a_{t-1}^v \cdot V)) + (W_q(a_{t-1}^q \cdot Q))C), H_t^q = \tanh(W_q(a_{t-1}^q \cdot Q) + (W_v(a_{t-1}^v \cdot V))C^T) \tag{8}$$

$$a_t^v = softmax((w_{hv})^T H_t^v), a_t^q = softmax((w_{hq})^T H_t^q) \tag{9}$$

The model is illustrated in Figure 3

3.4 Predicting Answer

We predict the answer using the co-attended images and question feature vectors. We use the MLP to generate the probabilities for the answer words. Like[8, 10], we use two fully connected layers to build the classifier.

$$\begin{aligned} h_{hidden} &= ReLU(W_{hidden}(concat(\hat{q}, \hat{v})) + b_{hidden}) \\ p &= softmax(W_{out} h_{hidden} + b_{out}) \end{aligned} \tag{10}$$

where p is the probability distribution over the all the answer words.

4 Experiment

4.1 Dataset

We are using VQA dataset 1.0[1] which has MSCOCO images[11] paired with questions solved by users on Amazon Mechanical Turk. The dataset has 82,783 training images, 40,504 validation images and 443,757 training questions, 214,354 validation questions. A testset is also provided for prediction challenge result submission. We run evaluation on test-dev and test-std to show the performance of our model on the task.

4.2 Evaluation Metric

The evaluation metric we use is the same as the method VQA challenge[1] provided:

$$Acc(a) = \frac{1}{K}\sum_{k=1}^{K}\min(\frac{\sum_{1\leq j\leq K, j\neq k} 1(a = a_j)}{3}, 1) \tag{11}$$

where a_i is the correct answer the user gives. K=10. We define the correct answer as the answer which three answer annotators thinks it is correct. The accuracy of over all 10 choose 9 subsets of true answers is calculated and averaged.

4.3 Setup

We use PyTorch to implement our models and use some preprocessing functions from open source[3] to deal with image cropping, scaling and text normalization for the experiment. We predict the probability for the top 3000 most frequent answer words which covers 92% answers in the validation set. We have implemented iBOWIMG as a simple baseline and the stacked co-attention and multi-glimpse attention model to compare with other candidate methods. Here we list our model setup for the models in experiments.

- iBOWIMG: The word embedding size is 500. The image feature we use is extracted from the final fully connected layer of resnet with a dimension of 1x2048. The learning rate is 0.0001 for the model and 0.001 for word embedding with Adam optimizer. The learning rate decay step is 10 epochs with the decay size of 0.1.
- Stacked co-attention & Multi-glimpse co-attention: The word embedding size is 300. The image feature we use is the 14 x 14 x 2048 features from the pooling layer input in resnet. The learning rate is 0.001 for the whole model with Adam optimizer. The learning rate decay step is 10 epochs with the decay size of 0.1.

4.4 Result

We report the models' performance in the following table. On the VQA 1.0 dataset, we list the performance of the iBOWIMG as the simple baseline and we also compare our co-attention models with the HierCoAtt and other VQA architectures. The iBOWIMG works reasonably well considering its simplicity in design and does not fall much behind the advanced models. In terms of our co-attention models, the stacked co-attention model reaches 62.26% accuracy and the accuracy of co-attention multi-glimpse is 62.49%. They outperform the HierCoAtt Model[6] based on which we design our models. The multi-glimpse co-attention is the best model. Based on model analysis in the next section, we believe the multiple views that the multi-glimpse model provides is the key to this performance gap. Stacked Attention might start from a bad attention and get stuck in it resulting in failure to capture the important entities in the multi-modal representations.

	test-dev				test-std			
Model	y/n	num	other	all	y/n	num	other	all
iBOWIMG[2]	76.55	35.03	42.62	55.72	76.76	34.98	42.62	55.89
SAN[4]	79.3	36.6	46.1	58.7	-	-	-	58.9
NMN[12]	81.2	38.0	44.0	58.6	-	-	-	58.7
DMN+[13]	80.5	48.3	36.8	60.3	-		-	60.4
HierCoAtt[6]	79.7	38.7	51.7	61.8	79.95	38.22	51.95	62.06
our stacked co-att	80.34	36.70	**52.21**	62.08	80.57	36.02	52.32	62.26
our multi-glimpse co-att	**81.15**	36.73	52.03	**62.32**	81.25	35.87	**52.23**	**62.49**

Table 1: Performance of Different Methods on VQA Dataset

4.5 Qualitative Analysis and Discussion

We analyze our models' strength and weakness by visualizing the attention distribution of the image and question along with the answer we predict for the question entry. We will demonstrate more

[3] `https://github.com/Cyanogenoid/pytorch-vqa`

samples in the Appendix. In the following visualizations, we list the picture, question, answers along with the multiple visualized attention maps on the modalities themselves. For multi-glimpse co-attention, the two image features maps are multiple views to the problem; for stacked co-attention the upper attention is the first attention and the lower is the refined one. For text, the more a token is attended, the more we colorize it with yellow hue.

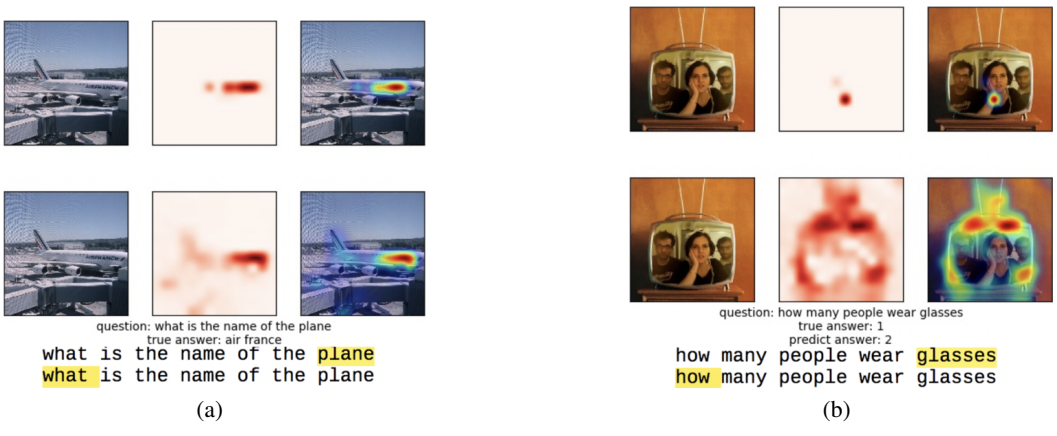
4.5.1 Multi-glimpse Co-attention Model

Figure 4 below shows the attention on the questions and the images for two glimpses respectively in the multi-glimpse co-attention model. We do see 2 glimpses can always capture different information in the image and they are complementary for each other. Sometimes the first glimpse pays attention to the correct part of the image and sometimes the second glimpse does. For the question attention, different glimpse focus on different kinds of information. The first glimpse mainly focuses on the noun phrase and verb phrase in the question. And the second glimpse mainly focuses on the Wh-pronoun.

Figure 5 provides the examples for which our multi-glimpse co-attention model gives the wrong answers. It shows that our model doesn't perform well on for the questions about counting number and the questions about recognizing words in the images which is the performance bottleneck of all VQA works published so far. The counting problem is highly complex in recognizing entities and segmenting instances and incorporating reasoning along the way. This is a topic which requires further research.

(a) (b)

Figure 4: multi-glimpse co-attention model correct answer

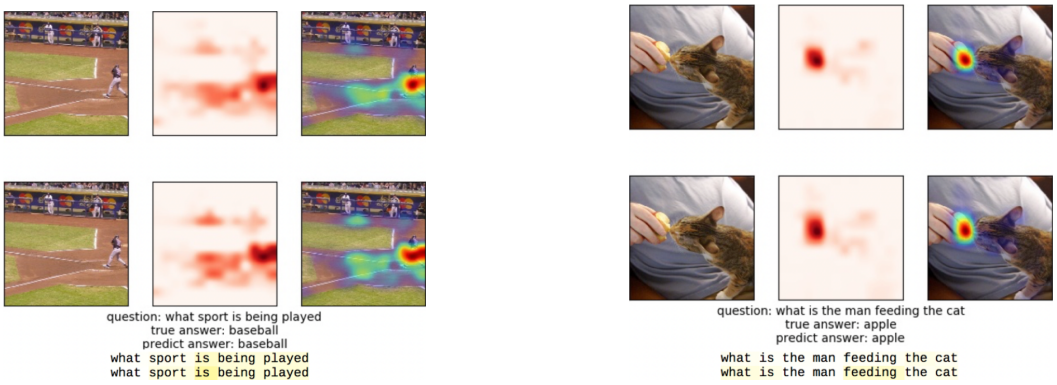


(a) (b)

Figure 5: multi-glimpse co-attention model incorrect answer

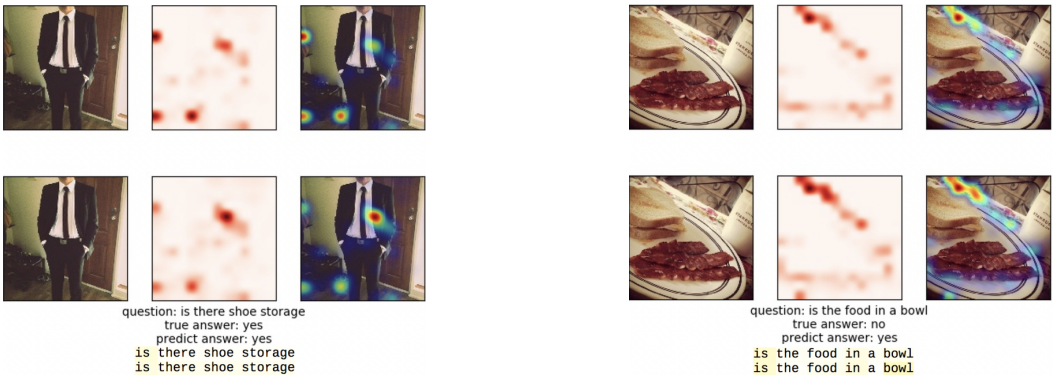
4.5.2 Stacked Co-attention Model

Figure 6 shows some results from the stacked co-attention model.In (b), The model learned to adjust the attention distribution of both image and question to focus on the right image part and question tokens. For example, the second picture shows the model learns to focus on the "feeding" word and a larger part of the activity area in the image. We can also see in Figure 7 from the incorrect sample (a) that that model fails to initially focus on "the shoe storage" and after the second attention, the model diverges further to the human object instead of the left corner of the image. Also the sample (b) fails to focus on the food in the plate twice. These incorrect answers demonstrate the model's drawback in getting stuck in the wrong attention distributions.



(a) (b)

Figure 6: stacked co-attention model correct answer



(a) (b)

Figure 7: stacked co-attention model incorrect answer

5 Conclusion and Future Work

In this project, we propose new architectures based on the HierCoAtt model by incorporating stacked attention and multi-glimpse and discarding the hierarchical architecture to make the model more powerful and yet simpler. However the models trained on VQA 1.0 is far from satisfactory. The dataset is previously accused of question bias that there are no dual questions for the same image to balance the question semantics distribution and because of this, models may have learned some characteristics of the biased question to infer the answer instead of composing the answering by reasoning through the semantic representation space. We already implemented, trained and tested the baseline iBOWIMG model on the balanced VQA 2.0 and the performance of 52.17% is close to the result proposed on the paper[2] on VQA 1.0. In the future, we are going to train our stacked co-attention model and multi-glimpse co-attention model on VQA 2.0.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[6] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061, 2016.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[8] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *CoRR*, abs/1704.03162, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

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

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Deep compositional question answering with neural module networks. *CoRR*, abs/1511.02799, 2015.

[13] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning*, pages 2397–2406, 2016.

Appendix.1 More examples for multi-glimpse co-attention model

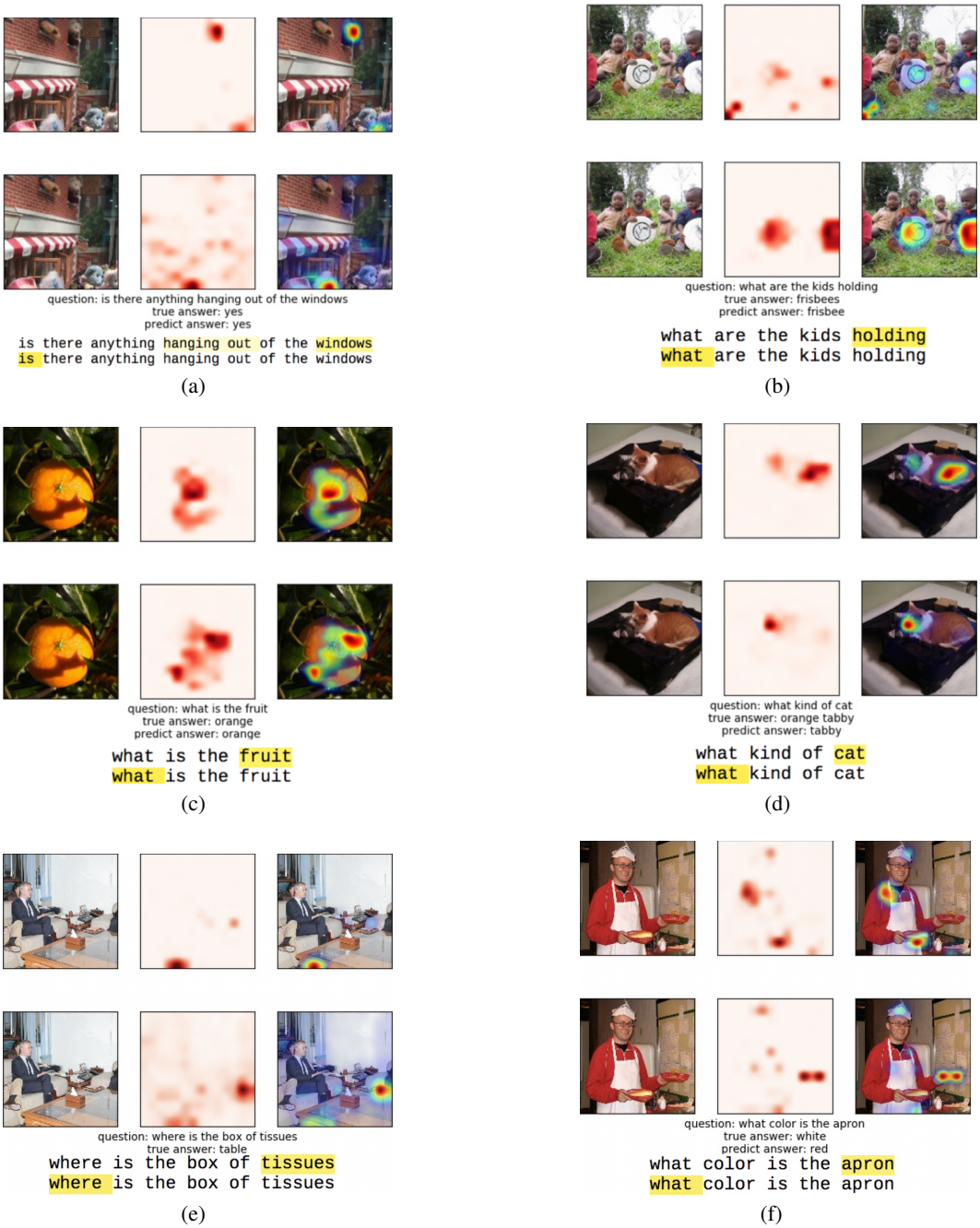
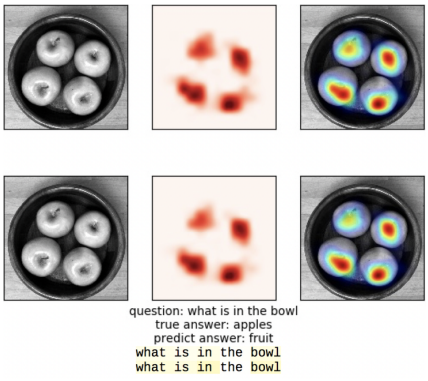


Figure 8: multi-glimpse co-attention model examples

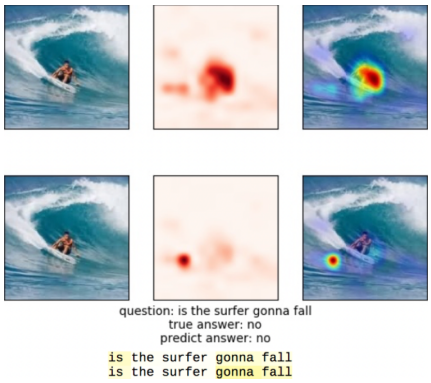
Appendix.2 More examples for stacked co-attention model



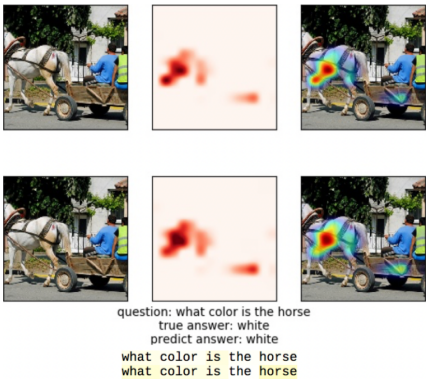
(a)

(b)

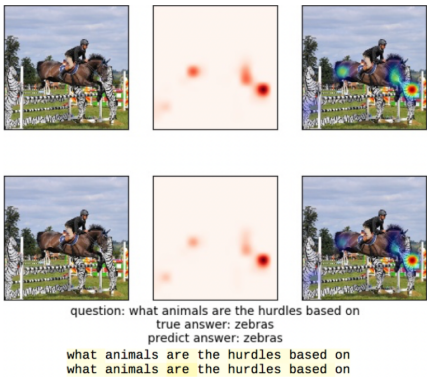
(c)



(d)



(e)



(f)

Figure 9: stacked co-attention model examples