Answer-Aware Attention on Grounded Question Answering in Images

Junjie Hu, Desai Fan, Shuxin Yao, Jean Oh
School of Computer Science, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
{junjieh, dfan, shuxiny}@cs.cmu.edu, jeanoh@nrec.ri.cmu.edu

Abstract

Grounding natural language expressions to visual context in an image is essential to understanding the semantic meaning of an image. Recent attention approaches on the task of grounded question answering in images simply rely on either attention over arbitrary regions in an image or attention over words in a question, which have not exploited the information behind candidate answers when encoding the question. To address this limitation, we propose two Answer-Aware Attention (AAA) models which use attention over candidate answers, i.e., global and local attention over answers, each of which learns an answer-aware summarization vector of a question. Our proposed attention model leverages information from both textual and visual modalities, which boosts the prediction accuracy in the grounded question answering task. Extensive experiments show that our proposed attention model performs comparably to the state-of-the-art models with much fewer learning parameters.

Introduction

The task of visual question answering (Antol et al. 2015; Wu et al. 2017) has gained significant popularity over the past few years in both the computer vision and natural language processing communities. Grounded question answering in images (Zhu et al. 2016) is a new type of visual question answering task in which answers to textual questions are image regions. Searching for a corresponding image region in an image based on a text entity is known as *grounding*. This requires the artificial intelligence system to learn semantic links between textual expressions and image regions. However, learning the similarity or correspondence of data in textual modality and visual modality remains far from solved due to the intrinsic difference between symbolic representations of words and continuous representations of images at the pixel-level.

Attention mechanism in neural network models mimics the attention behavior of human's visual system–when human perceive an image, attention allows for distilling information down to most salient objects rather than the entire image. Fusing and encoding information from visual and textual data via attention mechanism have achieved great success in jointly modeling vision and language tasks, such as image captioning (Xu et al. 2015), image generation from text (Mansimov et al. 2015), and grounded question answering (Hu et al. 2016a; Zhu et al. 2016). Pioneering work in grounded question answering in images (Zhu et al. 2016; Hu et al. 2016a) applied attention mechanism to summarize the textual question either over the question itself or over arbitrary regions extracted from convolution neural networks. These attention models suffer from missing information from the candidate visual answers during the summarization of the textual question.

Motivated by the goal of encoding a textual question via attending to its candidate answers, we present two Answer-Aware Attention (AAA) models that explicitly model the attention either *globally* over all candidate answers or *locally* over each candidate answer, and learn an answer-aware summarization vector of a question. To perform *global* attention over answers, we use a pooling function to integrate information from all candidate answers before encoding the question, whereas we summarize a specific question vector conditional on the *local* attention over each candidate answer. We evaluate our models on Visual7W dataset. Experiemtal results show that our models are comparable to the state-of-the-art model in terms of the grounding accuracy, while our models have fewer learning parameters thus resulting in faster training.

Related Work

Grounded Question Answering in images refers to the task of retrieving an image bounding box from a pool of candidates in an image according to a textual question (Hu et al. 2016b; Mao et al. 2016; Rohrbach et al. 2016; Yu et al. 2016; Nagaraja, Morariu, and Davis 2016). First of all, the pool of candidate bounding boxes can be obtained by object proposal networks (Arbeláez et al. 2014; Krähenbühl and Koltun 2014; Uijlings et al. 2013; Zitnick and Dollár 2014), or provided by human annotations. An encoding model is first learned to summarize the context of a given textual question, and a scoring function is then learned to score each candidate based on the question embedding. Finally, the candidate bounding box with the highest score is retrieved as the grounding prediction. Visual7w pointing dataset (Zhu et al. 2016) is a benchmark dataset for grounded question answering given a pool of annotated bounding boxes in an image. In this paper, we focus on

learning the encoding model of the textual question, and the scoring function which capture the relevance between the visual and textual modalities.

Attention Models have recently attracted lots of research interest in the fields of the textual question answering and visual question answering. A large batch of works on question answering (Seo et al. 2016; Fukui et al. 2016; Hu et al. 2016a) have demonstrated significant improvements by integrating attention mechanisms in neural models. The baseline method in (Zhu et al. 2016) follows the similar idea of the image captioning model in (Xu et al. 2015) which extracts a large number of feature maps from the convolution layer of a pre-trained convolution neural network, and performs soft attention over the features maps during the encoding of the question. This method suffers from a large amount of computation on the attention over the extracted feature maps which attend to arbitrary regions in an image. Compositional Modular Networks proposed in (Hu et al. 2016a) perform soft attention over the question words to obtain three embedding vectors for the *subject, relation, object* respectively in a relation triple. The localization module then grounds the bounding boxes with the text embedding of the *subject* and *object*, while the relationship module grounds the spatial embedding of two bounding boxes with the text embedding of the *relation*. These models encodes the question via attention either on the text or the image without explicitly fusing the information from the candidate bounding boxes. Our answer-aware attention model summarizes the question embedding conditional on either each the candidate bounding box locally or all the candidate bounding boxes globally.

Answer-Aware Attention Model

In this section, we first define the problem formally, and then describe two novel answer-aware attention models for the task of grounded question answering in images. Last, we introduce the learning objective to optimize the models.

Problem Definition

Given an image I and a question $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_M\}$, where $\mathbf{q}_i$ is the vector representation of the i-th words in the question with M words, we aim at learning a decision function to predict the correct answer out of N candidate answers $\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N\}$ which are N bounding boxes in the image.

Model Description

Figure 1 shows the general architecture of our proposed method, in which every dot in the orange area denotes a similarity score between a bounding box and a question word.

1. Embedding Layer is responsible for mapping the question and answer to the fixed-size vector spaces.

Question Encoding. The fixed-size vector representation $\mathbf{q}_i$ of each word in the question can be obtained from pre-trained word embeddings, e.g., GloVe (Pennington, Socher, and Manning 2014). We further use a Long Short-Term Memory Network (LSTM) (Hochreiter and Schmidhuber 1997) on top of the word embeddings to model

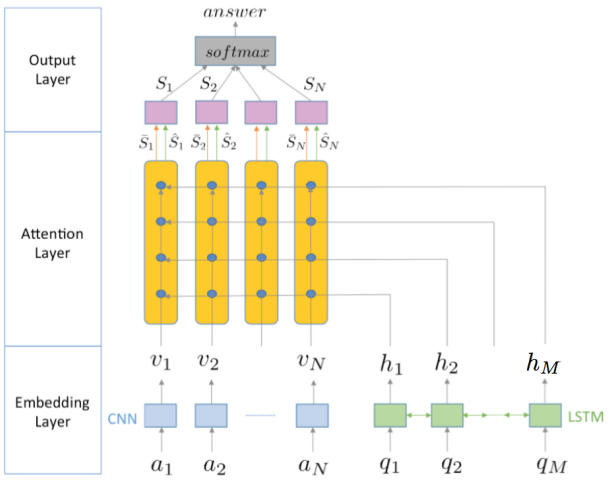

Figure 1: Illustration of the network architecture.

the interactions between words. Hence we obtain $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_M\}$ where $\mathbf{H} \in \mathbb{R}^{d \times M}$. M is the length of question. Notice that here we can also use any advanced recurrent neural network, e.g., stack LSTM or Bi-directional LSTM.

Answer Encoding. We use a Convolution Neural Network (CNN), i.e., VGG16 (Simonyan and Zisserman 2014) pre-trained on ImageNet dataset, to extract a fixed-size vector representation $\mathbf{v}_{\text{CNN}}$ for each bounding box in the image. We obtain the vector representation in 4,096 dimensions from the fc7 layer of the VGG16. Since each candidate answer is a bounding box in the image, we follow (Hu et al. 2016a) to extract a 5-dimensional spatial feature $\mathbf{v}_{\text{spatial}}$ for each answer, i.e., the vertical and horizontal coordinates of the top left point, the width, height and area of the box. We then concatenate these two vectors to represent each bounding box, i.e., $\mathbf{v}' = [\mathbf{v}_{\text{spatial}}; \mathbf{v}_{\text{CNN}}] \in \mathbb{R}^{d'}$, where $[;]$ denotes the concatenation of two column vectors throughout this paper. We further map the feature vector of each bounding box to the word embedding space by a linear transformation, i.e., $\mathbf{v} = \mathbf{W}_{(\mathbf{v})}^T \mathbf{v}' \in \mathbb{R}^d$.

2. Attention Layer is responsible for connecting and fusing information from both textual and visual modalities. Unlike previous attention methods which ignore attentions over the candidate answers during the encoding of the question sequence, we encode an answer-aware representation of the question after reading all the candidate answers. More specifically, we compute the similarity matrix $\boldsymbol{\beta}$ where each element $\beta_{ij} = g(\mathbf{h}_i, \mathbf{v}_j)$ measures the relevance of the i-th question word and the j-th candidate answer and $g(\cdot, \cdot)$ can be any similarity function. In this paper, we use $g(\mathbf{h}, \mathbf{v}) = \mathbf{w}_{(\boldsymbol{\beta})}^T [\mathbf{h}; \mathbf{v}; \mathbf{h} \odot \mathbf{v}] + b_{(\boldsymbol{\beta})}$ where $\odot$ is the elementwise multiplication operator. Next, we perform attention over candidate answers globally and locally.

Global Attention Over Answers (GAOA) first compute the attention weight over the entire question sequence by

integrating the relevant information from all the candidate answers in Equation 1.

$$\boldsymbol{\alpha} = \text{softmax}(\text{pool}_{\text{row}}(\boldsymbol{\beta})), \tag{1}$$

where $\text{pool}_{\text{row}}(\boldsymbol{\beta})$ is a pooling function of an input matrix over rows, e.g., max-pooling, mean-pooling. In this paper, we use the mean-pooling function which returns the mean value of each row in the matrix. Hence we can obtain an answer-aware summarization of the question $\hat{\mathbf{h}} \in \mathbb{R}^d$, and compute the relevance score between the question vector and each candidate answer by Equation 2 and 3.

$$\hat{\mathbf{h}} = \sum_{i}^{M} \alpha_i \mathbf{h}_i, \tag{2}$$

$$\hat{S} = \text{softmax}([\phi(\hat{\mathbf{h}}, \mathbf{v}_1); \cdots; \phi(\hat{\mathbf{h}}, \mathbf{v}_M)]) \tag{3}$$

where $\phi(\cdot, \cdot)$ can be any similarity function. Specifically in this paper, we use $\phi(\hat{\mathbf{h}}, \mathbf{v}_j) = \mathbf{w}_{(\phi)}^T \frac{\hat{\mathbf{h}} \odot \mathbf{v}_j}{\|\hat{\mathbf{h}} \odot \mathbf{v}_j\|_2} + b_{(\phi)}$.

Local Attention Over Answers (LAOA) focuses on each candidate bounding box and compute the attention weight over the entire question sequence conditional on each box locally by Equation 4.

$$\boldsymbol{\gamma}_j = \text{softmax}(\boldsymbol{\beta}_{:,j}) \tag{4}$$

where $\boldsymbol{\beta}_{:,j}$ denotes the j-th column of $\boldsymbol{\beta}$. Hence we can also obtain an answer-aware summarization of the question $\bar{\mathbf{h}}$ conditional on a specific answer, and compute the relevance score between the question vector and each candidate answer in Equation 5 and 6.

$$\bar{\mathbf{h}}_j = \sum_{i}^{M} \boldsymbol{\gamma}_i \mathbf{h}_i, \tag{5}$$

$$\bar{S} = \text{softmax}([\varphi(\bar{\mathbf{h}}_1, \mathbf{v}_1); \cdots; \varphi(\bar{\mathbf{h}}_M, \mathbf{v}_M)]) \tag{6}$$

where $\varphi(\cdot, \cdot)$ can be any similarity function. In this paper, we use $\varphi(\bar{\mathbf{h}}_j, \mathbf{v}_j) = \mathbf{w}_{(\varphi)}^T \frac{\bar{\mathbf{h}}_j \odot \mathbf{v}_j}{\|\bar{\mathbf{h}}_j \odot \mathbf{v}_j\|_2} + b_{(\varphi)}$

3. Output Layer is responsible for selecting one bounding box as the grounding result. We can choose the bounding box with the highest score defined in Equation 3 and 6 respectively. We can also ensemble the two models by linearly interpolating the scores from both attentions, and place a softmax function to estimate the probability of predicting the correct answers in Equation 7.

$$\mathbf{S} = \text{softmax}((1-\lambda)\hat{\mathbf{S}} + \lambda\bar{\mathbf{S}}) \in \mathbb{R}^N, \;\; j^* = \arg\max(\mathbf{S}) \tag{7}$$

where λ is a hyper-parameter that trade-offs the global attention and local attention over answers.

Learning

In the training phrase, we have the ground truth label of the correct answer for each question. Hence we can define the cross-entropy loss over all the question-answer pairs:

$$L(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log(\mathbf{S}_{y_t}) + \mu\|\theta\|_2 \tag{8}$$

where y_t denotes the index of the correct answer to the t-th question, θ is the collections of all the learning parameters, and μ is a regularization constant. We apply stochastic gradient descent to update all the parameters.

Experiments

Dataset statistics

We use the visual7W data (Zhu et al. 2016)[1] which is a benchmark dataset for grounded question answering in images. More specifically, we focus on the visual7w pointing task. The visual7w pointing dataset includes 188,068 QA pairs on 25,733 COCO images (Lin et al. 2014), together with 308,407 bounding boxes from 22,594 distinct categories. Follow the same splits in (Zhu et al. 2016), the dataset is split into 12,881/5,072/7,780 images for the train/validation/test sets respectively, resulting in 93,813/36,990/57,265 question-answer pairs.

Experimental Setup

To make fair comparisons, all our proposed methods adopt the same setup. We initialize the learning rate to 0.005, and applies the exponential decay to the learning rate with a base of 0.1 every 8,0000 iterations. We use the GloVe word embeddings with the dimension of 300, and fix the vocabularity size to be 72,704. The hidden dimension of LSTM used in our proposed methods is set to 500. We use the Faster-RCNN VGG-16 network pre-trained on the ImageNet dataset to extract the visual features of the bounding boxes. We set the parameter λ in Equation 7 to be 0.5.

Results and Discussions

Table 1 shows the prediction accuracy on the test set and validation set in the Visual7w pointing task. Several observations can be drawn as follows.

- The attention over arbitrary convolution feature maps proposed in (Zhu et al. 2016) performs worse than the attention over words in the question proposed in (Hu et al. 2016a). This indicates that identifying the key words in the question is essential in learning a good representation of the question, while attention over arbitrary convolution feature maps confuses the sentence encoder in encoding a good representation of a question.
- LAOA performs better than GAOA in terms of the prediction accuracy. Since that LAOA summarizes the question vector conditional on each candidate answer locally without including additional information from other candidate answers. The performance of GAOA relies on selecting a good pooling function to fuse information over all the candidate answers, which we leave as our future work.
- The number of the learning parameters of the LSTM encoder in CMN is 21x larger than that in our two proposed attention models, while the performance of LAOA is comparative to that of CMN. This indicates that our proposed models are more efficient in learning from the question answering pairs, and have the advantage of faster training.

[1] http://web.stanford.edu/~yukez/visual7w/

Table 1: Experimental Results

Method	Test accuracy	LSTM hidden dimension	LSTM layer
Visual Attention Baseline (Zhu et al. 2016)	0.561	1000	1-layer unidirectional LSTM
CMN (Localization) (Hu et al. 2016a)	0.716	1000	1-layer unidirectional LSTM
CMN (full) (Hu et al. 2016a)	0.725	1000	2-layer bi-directional LSTM
GAOA	0.707	500	1-layer unidirectional LSTM
LAOA	0.723	500	1-layer unidirectional LSTM
GAOA+LAOA	0.713	500	1-layer unidirectional LSTM
GAOA w/o LSTM	0.697	-	-
LAOA w/o LSTM	0.642	-	-

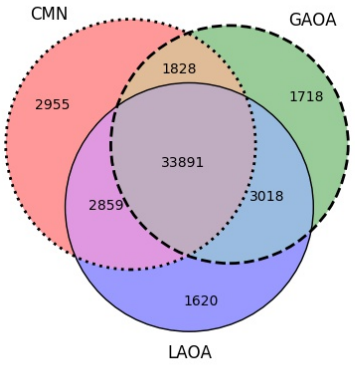

Figure 2: Venn diagram on the number of correct answers predicted by CMN, LAOA and GAOA.

- The last two lines in Table 1 show the performance of our two proposed models directly using GloVe word embeddings for question words rather than encoding them by a single layer unidirectional LSTM. The results show that the performance of both models without using LSTM drops dramatically. This indicates that fine-tuning the representation of question words by a LSTM using the training data performs much better than fixed representations.

Visualization

To further understand the performance benefits of incorporating answer-aware attention mechanisms into the grounded question answering in images, we can take a look at the questions on which models disagree. Figure 2 shows the Venn Diagram on the questions that have been corrected identified by our two proposed attention models and the state-of-the-art CMN model. Here we see that the vast majority of the correctly answered questions are shared across all three models. The rest of them indicating questions that models disagree are distributed fairly evenly.

Figure 3 shows a case study of the similarity heatmap between question tokens and candidate answers for exclusively correct examples as well as wrong examples predicted by LAOA. The exclusively correct examples means examples that only LAOA gives the correct predictions while the other two, CMN and GAOA, provide wrong answers. The absolute number and the relative percentage of the exclusively correct examples could be found in the above Venn diagram, Figure 2.

A few observations could be found from the examples:

- It can be seen that our LAOA model has the ability to extract the target object the question is asking, for instance, the word *vehicle* in the first example and the word *bush* in the sixth example.
- LAOA could also find the key properties or relationships to other object that helps distinguish with other answers, for instance, the word *orange* in the second example and the word *underneath* in the fourth example.
- There are also some situations when the LAOA model makes mistakes. In the seventh example, LAOA could not perform a deeper reasoning of figuring out what is "in front of the others", where coreference is involved. In the last example, although LAOA is very confident about the mapping between the word *hanging* and the yellow bounding box, it makes the wrong prediction in the end. The reason could be that the dominating confidence actually obscures the target object the question really asks, which hinders the model to give the correct answer.

From the heatmap we could conclude that our LAOA model is capable of learning reasonable semantic mappings between the visual space and the textual space. This ability of bridging the visual-textual modality gap enables the LAOA model to have such a good performance. On the other hand, the model could still fail at some questions, even if some visual-textual correspondences are correctly found. This is because our model still lacks the deeper reasoning ability.

Conclusion and Future Directions

In this paper, we propose two attention models each of which learns an answer-aware summarization vector of a question after reading candidate answers. Our proposed models fuse information from both visual and textual modalities, and ground key words in a question to the corresponding bounding box in an image. Experimental results show that our proposed local attention model over answers performs compar-

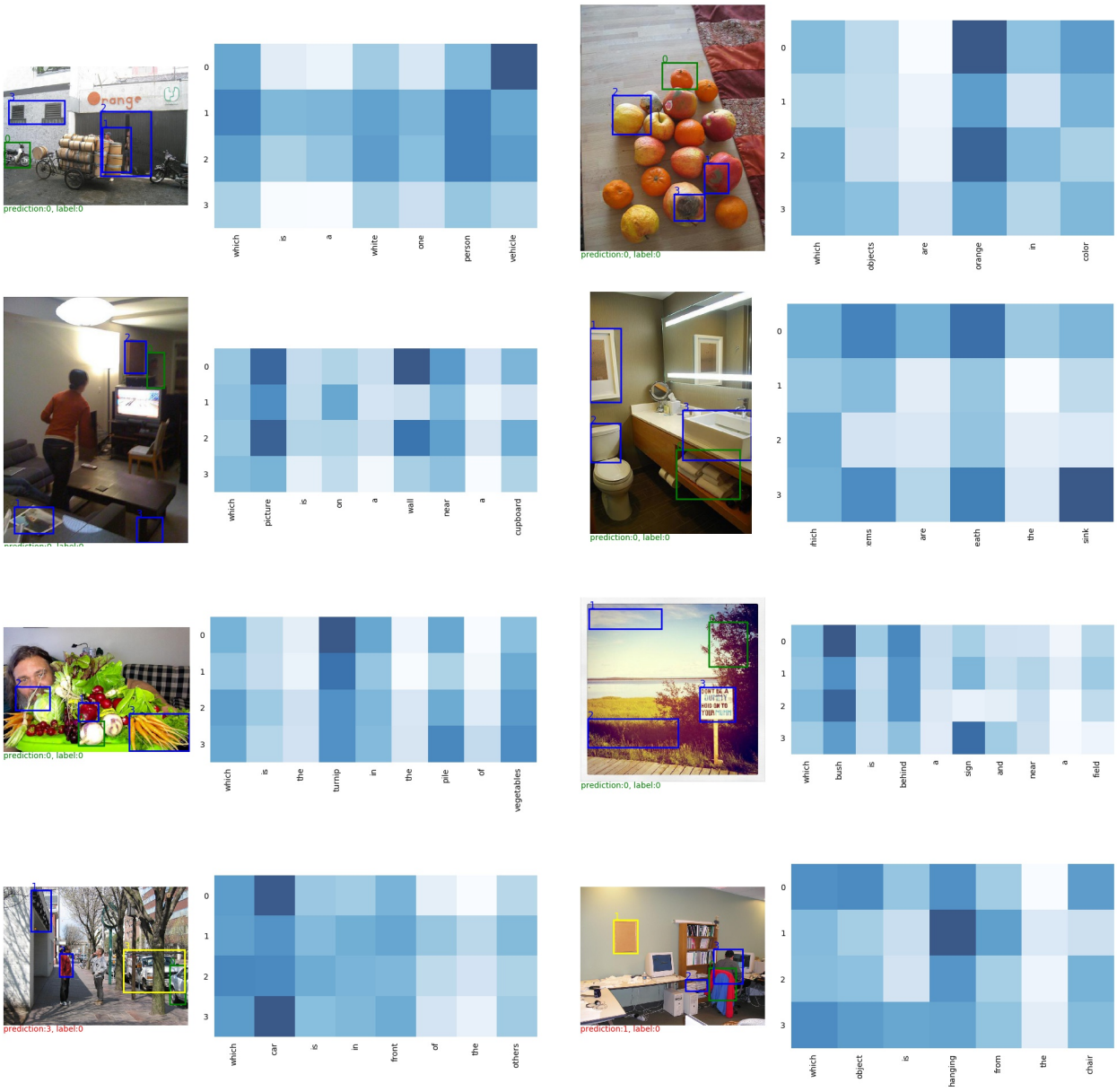

Figure 3: Visualization of the similarity heatmap between question tokens and candidate answers for exclusively correct examples and wrong examples predicted by LAOA. For each example, the image and candidate answers (label is green, prediction is yellow, others are blue) are shown on the left, and the similarity heatmap (darker is higher) between question tokens (x axis) and candidate answers (y axis) is on the right. The last two examples are wrong predictions made by LAOA.

atively to the state-of-the-art method on the Visual7w pointing task while having much fewer learning parameters. Future directions can be further explored in the following aspects: (1) Selecting a good pooling function in the GAOA model to fuse relevant information from all the candidate answers is challenging, and more advanced strategies can be further explored weight over answers. (2) More effective similarity functions can be designed to capture the correspondence at different granularity levels, e.g., attention at the word, phrase and sentence level. (3) Extra image bounding boxes with correspondent text annotations are helpful to fine-tune the mapping function that projects the visual embedding vectors to the textual space. Obtaining a good projection to an appropriate textual space allows the represen-

tation of candidate bounding boxes to be distinguishable.

Acknowledgments

This work was conducted in part through collaborative participation in the Robotics Consortium sponsored by the U.S Army Re- search Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[Antol et al. 2015] Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433. 1

[Arbeláez et al. 2014] Arbeláez, P.; Pont-Tuset, J.; Barron, J. T.; Marques, F.; and Malik, J. 2014. Multiscale combinatorial grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 328–335. 1

[Fukui et al. 2016] Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*. 2

[Hochreiter and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780. 2

[Hu et al. 2016a] Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2016a. Modeling relationships in referential expressions with compositional modular networks. *arXiv preprint arXiv:1611.09978*. 1, 2, 3, 4

[Hu et al. 2016b] Hu, R.; Xu, H.; Rohrbach, M.; Feng, J.; Saenko, K.; and Darrell, T. 2016b. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4555–4564. 1

[Krähenbühl and Koltun 2014] Krähenbühl, P., and Koltun, V. 2014. Geodesic object proposals. In *European Conference on Computer Vision*, 725–739. Springer. 1

[Lin et al. 2014] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755. Springer. 3

[Mansimov et al. 2015] Mansimov, E.; Parisotto, E.; Ba, J. L.; and Salakhutdinov, R. 2015. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*. 1

[Mao et al. 2016] Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11–20. 1

[Nagaraja, Morariu, and Davis 2016] Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, 792–807. Springer. 1

[Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 2

[Rohrbach et al. 2016] Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, 817–834. Springer. 1

[Seo et al. 2016] Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*. 2

[Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2

[Uijlings et al. 2013] Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision* 104(2):154–171. 1

[Wu et al. 2017] Wu, Q.; Teney, D.; Wang, P.; Shen, C.; Dick, A.; and van den Hengel, A. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*. 1

[Xu et al. 2015] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2048–2057. 1, 2

[Yu et al. 2016] Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, 69–85. Springer. 1

[Zhu et al. 2016] Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 4995–5004. IEEE. 1, 2, 3, 4

[Zitnick and Dollár 2014] Zitnick, C. L., and Dollár, P. 2014. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, 391–405. Springer. 1