2018 IEEE Winter Conference on Applications of Computer Vision

Semantically Guided Visual Question Answering

Handong Zhao
Northeastern University
hdzhao@ece.neu.edu

Quanfu Fan
IBM T. J. Watson Research Center
qfan@us.ibm.com

Dan Gutfreund
IBM T. J. Watson Research Center
dgutfre@us.ibm.com

Yun Fu
Northeastern University
yunfu@ece.neu.edu

Abstract

We present a novel approach to enhance the challenging task of Visual Question Answering (VQA) by incorporating and enriching semantic knowledge in a VQA model. We first apply Multiple Instance Learning (MIL) to extract a richer visual representation addressing concepts beyond objects such as actions and colors. Motivated by the observation that semantically related answers often appear together in prediction, we further develop a new semantically-guided loss function for model learning which has the potential to drive weakly-scored but correct answers to the top while suppressing wrong answers. We show that these two ideas contribute to performance improvement in a complementary way. We demonstrate competitive results comparable to the state of the art on two VQA benchmark datasets.

1. Introduction

Visual Question Answering (VQA) is the task of providing a text-based answer to a text-based question related to a given image [2]. The task is particularly challenging as there is little restriction on the question itself, which could be free-form and open-ended. The problem requires not only to tackle classic challenges in computer vision such as object detection and classification, but also to understand where to focus attention in the image based on a question presented in free form text. Malinowski and Fritz consider VQA as a step towards a full fledged visual Turing test [20, 21].

VQA is emerging as a new frontier in the effort to jointly understand images and natural language. Recent years have witnessed a surge of research on this problem, largely due to the success of deep learning in the fields of computer vision and NLP [31, 10, 11]. One common theme in most of these works is to treat the problem as a classification task where CNN-based image representations are combined with RNN- or CNN- based question representations for answer prediction [2, 19, 20, 18]. Many approaches on how to extract the features and how to combine them were suggested. For example, attention-based models learn, based on the question, which region in the image is the most relevant to the question. Other approaches introduce external knowledge to improve the question representation [34]. In Section 2, we discuss these approaches in detail.

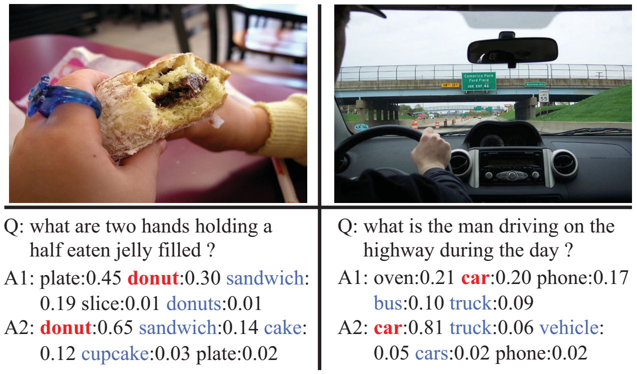

Q: what are two hands holding a half eaten jelly filled ?
A1: plate:0.45 donut:0.30 sandwich: 0.19 slice:0.01 donuts:0.01
A2: donut:0.65 sandwich:0.14 cake: 0.12 cupcake:0.03 plate:0.02
Q: what is the man driving on the highway during the day ?
A1: oven:0.21 car:0.20 phone:0.17 bus:0.10 truck:0.09
A2: car:0.81 truck:0.06 vehicle: 0.05 cars:0.02 phone:0.02

Figure 1. Examples of our proposed semantically guided VQA. The images are from COCO-QA dataset [17], and the questions/answers from [29]. We produce two sets of results for each question. A1 are generated by a baseline approach, and A2 are enhanced by our proposed approach. All the predictions are followed by their confidence scores. Ground-truth labels are marked in red and semantically related words are colored in blue for the ease of illustration. Our proposed approach successfully predicts the correct answers while the baseline does not.

In this paper we propose two new ideas to enhance a VQA model and enrich it with external knowledge. The first idea injects external knowledge into the model by considering the semantics of the answers at the **prediction level**. Note that this is different in principal than previous works such as [34] which introduced semantic knowledge at the representation level. The second idea continues the line of work on improving visual representations. We apply Multiple Instance Learning (MIL) [22, 32] to enrich concept representations beyond objects in the visual model, in a similar spirit to [6].

We start by describing our main contribution of this paper. For motivation, consider the example on the right in Figure 1. A1 lists the top predictions of a baseline model while A2 are the predictions of our proposed model. We can see that the baseline model assigns a higher probabil-

978-1-5386-4886-5/18/$31.00 ©2018 IEEE
DOI 10.1109/WACV.2018.00205

ity to the wrong answer *oven* over the correct answer *car*. Note however, that more predictions (i.e. *bus* and *truck*) in the example are semantically related to *car*, and not to *oven*. We observe many examples like this in the baseline results especially when the model is not confident about the correct answer. Motivated by this, we design a new loss function that favors predictions that are semantically related to the ground truth while penalizing those less relevant. Minimizing such a loss can help guide the model in the right direction for boosting up the likelihood of the correct answer and suppressing wrong predictions. As illustrated in Figure 1, our proposed approach correctly picks *car* as the answer while pushing *oven* away in the prediction. Worthy to mention is that even if our approach fails to provide a correct answer, it is more likely to choose an answer related to the correct one (See the first example of Figure 6.), reducing the risk of making embarrassing errors. In this work we compute semantic similarity between concepts based on Google's pre-trained Word Embeddings [23], which was learned from a large Google News corpus.

Our second contribution in this work is to improve semantic concept representations in visual modeling for VQA. The majority of the previous approaches extract visual features from CNNs such as VGG [31], GoogLeNet [33] and ResNet [10]. While these models show promising results, it is worth noting that they were pre-trained on ImageNet with 1000 object categories, only a very small coverage of the broad semantic domain. In VQA, on the other hand, since the questions are free formed, an ideal model would be required to go beyond objects to recognize a substantially larger set of concepts like humans do, including numbers, colors, locations and attributes. To address this issue, we adopt a weakly supervised model based on MIL as the visual feature representation in our work [6]. This model was trained on keywords from image captions on the MS COCO dataset [17] including nouns, verbs and adjectives. It has demonstrated better performance in detecting and localizing nouns and adjectives than CNN-based classification models. Note that Attribute-based models for VQA have been explored in [34]. However, our model is based on MIL, thus is expected to provide better concept localization [32], which is important for attention-based VQA approaches.

To sum up, the major contributions of our proposed method are two-fold:

- Enforcing semantic relatedness during training through the loss function by incorporating external semantic knowledge via word2vc embeddings. This not only improves the performance, but also makes it less likely to provide *embarrassingly* wrong answers.
- Enriching feature representation from an MIL model. This representation not only enables better recognition of attributes beyond objects, but also improves the localization of these concepts which is in particular important for attention-based models.

2. Related Work

One of the first works on VQA by Antol et al. [2], presents a number of models as well as the VQA dataset which is the largest to date. Driven by VQA and other datasets, a number of approaches indexing global visual and question representations are proposed [21, 29, 19, 1, 13]. Soon researchers find that visual representations that are based on the entire image are often too coarse to attend to the region of interest, which is crucial for accurate question answering. Motivated by this, recent studies focus on attention models. Yang et al. [37] present a multiple-layer stacked attention network, which extracts the attended region progressively by querying the questions multiple times. Similarly, Shih et al. [30] project the question features and image features into a common space and compute the relevance of each sub-region via inner product. Xiong et al. [35] propose several improvements to the memory and input modules of dynamic memory network. Their best performed model includes a two-layer encoder with sentence reader and input fusion layer to allow for information flow between sentences. Xu and Saenko [36] present a spatial memory network, which is a recurrent neural network with an explicit attention mechanism. The spatial attention architecture is able to align words with image patches information stored in memory. Ilievski et al [12] propose a different attention model. They use an off-the-shelf object detector to determine the region of interest. Then LSTM is applied to embed the information from regions together with global features, which is then combined with the question representation. Noh et al. [26] present a dynamic parameter prediction network built on gated recurrent unit (GRU). To solve large amount of parameters problem in their network, they apply a hashing trick in the dynamic parameter layer. Lu et al. [18] jointly learn a hierarchical attention mechanism on both image and the text, based on three levels: word, phrase and question. Such hierarchical image-text co-attention mechanism also appears in image captioning literature [24].

Improvements on top of attention models. Attention models have been proven effective in solving the VQA problem. We now review works which, similar to our work, explored different ideas on top of such models, such as multimodal data fusion and external knowledge integration.

Fukui et al. [7] propose the use of bilinear pooling for combining multi-modal information and suggest an efficient compact implementation. Kim et al. [15] introduce a low-rank bilinear pooling using Hadamard product. Kim et al. [14] present another multimodal feature fusion model based on deep residual learning. They use element-wise multiplication to learn the joint residual mappings for both visual and textual features.

Wu et al. [34] use external knowledge from DBpedia [3] and apply it to a LSTM model to improve answers prediction. Gao et al. [9] propose a similar architecture but decompose the LSTMs into encoding and decoding.

As an interesting study on VQA, Ray et al. [28] explore the relevance of questions to images. Specifically, their proposed LSTM-based method first determines whether the question is "visual" or not. If visual, it further determines whether the question is relevant to the given image or not.

3. Our Approach

Since our method can be easily incorporated with any classification based approach to VQA, we implement our ideas on top of a recent approach [18] which achieved state of the art performance, Hierarchical Co-Attention, as the baseline to illustrate our idea. Let us start by describing this approach at a high level.

3.1. Hierarchical Co-Attention

The basic idea of co-attention models is to learn jointly from the question and the image which regions in the image and the question are the most relevant to provide an answer. The Hierarchical co-attention (HieCoAtten) model proposed by Lu et. al. [18] perform co-attention based on three different levels in the text: word level, phrase level and sentence level. Roughly speaking, word-level features are extracted from words embedding. Phrase-level features are computed by a convolutional layer that receives the word embeddings as inputs, with three different window sizes: unigram, bigram, and trigram. Question-level features are computed by a LSTM receiving the phrase-level features as inputs. Correspondingly, three levels of visual features are produced by the co-attention mechanism. The visual features are extracted from a CNN (VGGnet [31] or ResNet [10]) that was pre-trained on ImageNet. The final feature vector is computed recursively using a multi-layer perceptron (MLP) from word-level to question-level. We refer the reader to [18] for more details.

3.2. Multiple Instance Learning (MIL)

Models pre-trained on the ImageNet dataset [5] such as AlexNet [16], VGGnet [31], GoogLeNet [33] and ResNet [10] have been widely applied to extract features from images in a plethora of applications. These networks specialize in object classification and recognition. We argue that for VQA this is not enough. Consider the right most example in Figure 5. The above mentioned networks are trained to recognize *horses* but not the color *white* or the action *pulling*. Given the free form and open ended nature of the questions, feature representations of verbs and adjectives are just as critical to solve VQA as objects do.

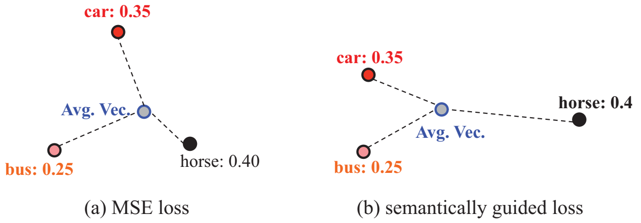


Figure 2. An illustration of comparison between traditional MSE loss (a) and the proposed semantically guided loss (b). In MSE, categories representations are pairwise equidistant. As shown in (a), if the ground-truth label *car* is predicted with probability 0.35 and *horse* is predicted with probability 0.4, the weighted average vector is closer to *horse*. In (b) on the other hand, *car* and *bus* are closer to each other than each of them to *horse*. As a result, they together pull in the direction of the cluster of word representations that are semantically related to vehicles and as a result, the weighted average vector is closer to *car* than to *horse*.

Motivated by this, we replace the visual feature extraction mechanism in the co-attention model of [18] with a MIL model from [6]. This model is based on VGGnet architecture and it was trained on Microsoft COCO captions dataset. The 1000 categories of the model are the most common words in the training captions and it includes nouns, verbs and adjectives. As such, it is richer than models that were pre-trained on ImageNet for object classification and hence more suitable for VQA. In addition it tends to have better localization performance than classification models which is important for attention-based models.

3.3. Semantically Guided Loss

Most classification neural networks based models for VQA have a softmax layer at the top outputting a probability distribution $\mathbf{p} \in \mathbb{R}^C$ for each image/question pair: $p_i = \frac{\exp(h_i)}{\sum_i \exp(h_i)}$, $i = 1, \ldots, C$, where C is the number of categories. During training a standard loss function is applied to quantify the loss between the prediction distribution $\mathbf{p}$ and the ground truth distribution over categories $\mathbf{t} \in \mathbb{R}^C$. In the case of a single correct label, $\mathbf{t}$ has all the probability weight on the one ground truth category. Typically, cross entropy is the loss function of choice: $\mathcal{L}_{\mathbf{CE}}(\mathbf{p}, \mathbf{t}) = -\sum_i t_i \log(\mathbf{p}_i)$. [1]

When we observe the failure cases generated by the classification model A1 in Figure 1, we find interesting phenomena: (1) For many cases, the ground-truth label is one of the top ranking predictions. For example, in the right example, the ground-truth *car* is ranked second with only 1% probability less than *oven*. (2) A number of predictions that are semantically related to the ground-truth are also scattered among the high ranking predictions. This often happens because of the nature of a classification model in which

[1] Throughout we use bold letters to represent vectors and matrices.

each category's contribution to the loss is equally weighted. That is, the model detects *vehicle* related elements in the image (e.g. *steering wheel*) and in the question (e.g. *driving*), but it is not confident enough regarding which type of vehicle is the correct one and therefore spreads the probability weights among several vehicle related categories.

Motivated by this, we propose a loss function which in addition to the standard cross entropy loss, has a component that measures the loss based on the semantic relatedness of the predictions to the ground truth. Here semantic relatedness is based on distances between word2vec representations of the predictions.

We define the *Weighted Average Vector (WAV)* loss function for a given image-question example as follows:

$$\mathcal{L}_{\mathbf{WAV}}(\mathbf{p},\mathbf{t}) = \mathcal{L}_c(\mathbf{p},\mathbf{t}) + \lambda \|\mathbf{V}\mathbf{p} - \hat{\mathbf{v}}_{\mathbf{t}}\|_2^2 \qquad (1)$$

where $\mathbf{V} \in \mathbb{R}^{d \times C}$ is a matrix whose columns are the word2vec embeddings of the word categories in a d-dimensional vector space. $\hat{\mathbf{v}}_{\mathbf{t}} \in \mathbb{R}^d$ is the word2vec embedding of the ground-truth label $\mathbf{t}$, and $\mathbf{p} \in \mathbb{R}^C$ is the probability vector after the softmax activation function. λ is a balancing parameter between the two components. [2]

In short, the loss function has a standard cross entropy loss component as well as a semantic component which penalizes the model during training in proportion to the Euclidean distance between the vector which is the weighted average of the word2vec representations of all the categories and the representation of the ground truth. The rational is that if the model gives high probability to semantically unrelated categories, it is severely penalized by the semantic component of the loss. The cross entropy component plays a role in determining the correct answer among semantically related categories. The combination of the two gives the best performance.

Comparison to Mean Squared Error (MSE) loss. We note that there is an interesting correspondence between the semantic component of WAV and the MSE loss in the case of a single correct label. In such a case, MSE is written as:

$$\mathcal{L}_{\mathbf{MSE}}(\mathbf{p},\mathbf{t}) = \|\mathbf{E}\mathbf{p} - \mathbf{e}_t\|_2^2 \qquad (2)$$

where $\mathbf{t}$ is the ground truth and $\mathbf{p} \in \mathbb{R}^C$ is the probability vector after the softmax activation function. $\mathbf{E} \in \mathbb{R}^{C \times C}$ is the diagonal matrix whose diagonal entries are 1 and it is 0 elsewhere. $\mathbf{e}_t \in \mathbb{R}^C$ is the vector that is 1 in the t'th entry and 0 elsewhere, i.e. it is the indicator vector (a.k.a the one-hot vector) of category t. In words, MSE penalizes the model during training in proportion to the Euclidean distance between the vector which is the weighted average of the indicator vectors of all the categories and the indicator vector of the ground truth.

Our semantic loss replaces the indicator vector representations (in *categorical* space) in which each pair of representations has the same distance, with a word2vec representation (in *semantic* space) in which pairwise distances are semantically driven. In Figure 2 we illustrate the effect of this on the model's predictions.

3.4. Implementation details

To implement our idea, we use Torch deep learning package [4]. We train the model with stochastic gradient descent using Rmsprop algorithm. We set the base learning rate 4e-4, momentum 0.99 and weight decay of 1e-8. The MIL features, are extracted from the model of [6] that was trained on COCO captions dataset. We extract features from the last pooling layer (i.e. pool-5 layer) before the fully-connected layers with dimension of $18 \times 18 \times 512$, where 18×18 is the number of patches and 512 is the feature dimension. The MIL features are fixed and are not trained further.

The word2vec embedding is taken from [23]. It was computed on top of the Google News corpus and embeds words into a space of dimension $d = 300$. Note that the model only embeds single words. When class label is not a single word but rather a phrase, e.g. *black and white*, we take the representation of the first word to represent the label. This practice does harm the representation accuracy of labels to some extent. Fortunately, the number of labels with multiple words is much smaller than single words. As reported in [2], about 90% of the answers have single words and 98% of answers do not exceed three words. One could use several other more sophisticated strategies, e.g. doc2vec [27] to handle it, which is not the focus of this paper.

4. Experiments

4.1. Evaluation data and metric

We evaluated our approach on two datasets: Toronto **COCO-QA** [29] and **VQA** [2]. Currently these are two of the widely used benchmarks for VQA evaluation.

COCO-QA is based on the Microsoft COCO dataset. The ground truth annotations of this dataset were automatically generated by running a text parser to parse the image captions and then replacing the keywords with corresponding question words to form question/answer pairs. There are a total of 78,736 questions for training and 38,948 test questions in COCO-QA, including four types of questions: object (70%), number (7%), color (17%), and location (6%). All the answers are single words. Top-1 accuracy is the most widely applied evaluation metric for COCO-QA. We report it in Table 1 for all the methods. Besides, for the analytical study of the proposed method, we also report top-5 accuracy in Figure 3.

VQA [2] is a benchmark dataset for visual question an-

[2] Due to the limited page length, a sensitivity study on parameter λ is moved to the supplementary material.

swering[3]. Unlike **COCO-QA**, **VQA** uses human-annotated questions and answers. It contains a total of 6,141,630 question-answers pairs, which are split into three subsets, 248,349 for training, 121,512 for validation, and 244,302 for testing. There are three major sub-categories for questions in this dataset, including yes/no, number and others. Each question has 10 free-response answers from different annotators. Two settings are provided for evaluation on **VQA** , open-ended and multiple-choice. In the open-ended setting, there are possibly multiple correct answers to the same question while in the multiple-choice case, the answer is limited to 18 pre-specified possible candidates only. We followed the most widely used experimental setting in [2, 37, 18], and used the top 1000 most frequent answers in training and testing. Note that these 1000 answers only cover 86.54% of the total number of questions in the development set (train+val), thus the highest performance one could expect for our approach, as well as most previous works, on the validation set would be at most 86.54%. In this work, we trained our models on the development set and reported the performance on the test-dev and test-standard datasets using the VQA evaluation server.

Following previous works, we use a different accuracy evaluation metric. Each question has ten answers from ten different annotators. A prediction is considered correct if at least three annotators suggested it as their answer. In addition, a prediction receives a partial score even if only one or two annotators suggested it. Specifically, for a predicted answer a and a set of answers T for a given image-question pair, the classification accuracy is defined as

$$\mathrm{Acc}(a, T) = \min\left\{\frac{\sum_{t\in T}\mathcal{I}[a=t]}{3},\ 1\right\} \tag{3}$$

where $\mathcal{I}$ is the indicator function. The accuracy of the model is the average over all the image-query pairs in the test set.

4.2. Results

We compared our approach to a number of recently developed techniques for VQA. These include *2-VIS+BLSTM* [29], *IMG-CNN* [19], *LSTM Q+I* [2], *Region Sel.* [30], *SMem* [36], *SAN* [37], *FDA* [12], *DMN+* [35], *DPPnet* [26], RAU [25] and *HCoAtten* [18].

Among them, the last seven approaches starting from *Region Sel.* are attention-based, thus in general outperforming the previous models, such as *VIS+BOW* [29] and *VIS+LSTM* [29], which are based on global visual features. We consider *HieCoAtten* as the baseline in our comparison as it is one of the most competitive approaches on both COCO-QA and VQA, and our approach is built on top of it. Similar to [18] we trained two models, one based on a VGG architecture (denoted *CoAtten+WAV*(MIL)) and the other on ResNet (denoted *CoAtten+WAV*(ResNet)). Thus in order to investigate the added value of our approach it is instructive to compare the respective performances of the VGG based models and the ResNet based models. The only model available to us that was trained using the MIL paradigm was VGG based. The ResNet model was trained using standard supervised learning. To isolate the contribution of the MIL approach from the semantically guided loss idea, we also trained the model *CoAtten*(MIL) with a standard cross entropy loss.

Table 1. Top-1 accuracy results on the COCO-QA dataset. "-" indicates that no results are available.

Methods	Obj.	Num.	Col.	Loc.	All
2-VIS+BLSTM [29]	58.2	44.8	49.5	47.3	55.1
IMG-CNN [19]	-	-	-	-	58.4
DPPnet [26]	-	-	-	-	61.2
SAN(2, CNN) [37]	64.5	48.6	57.9	54.0	61.6
HieCoAtten$^{\alpha}$(VGG)[18]	65.6	48.9	59.8	56.7	62.9
HieCoAtten$^{\alpha}$(ResNet)[18]	68.0	51.0	62.9	58.8	65.4
CoAtten(MIL)	65.6	51.7	64.7	56.3	63.9
CoAtten+WAV(MIL)	66.4	51.7	64.8	56.4	64.5
CoAtten+WAV(ResNet)	68.6	51.0	64.6	57.8	66.1

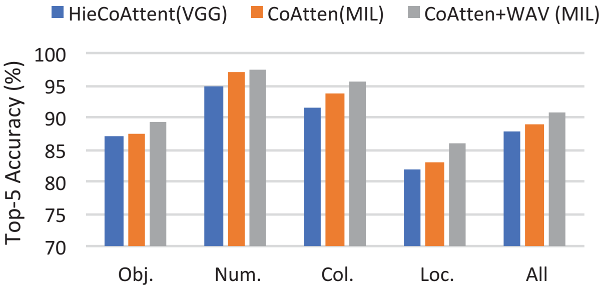


Figure 3. Top-5 accuracy results of HieCoAtten$^{\alpha}$(VGG)[18], *CoAtten*(MIL), and *CoAtten+WAV*(MIL) on COCO-QA.

Evaluation on COCO-QA. Table 1 shows the results of different approaches on COCO-QA. We first notice that by applying MIL-based features to *HieCoAtten*, i.e. the model CoAtten(MIL), we already achieve better overall accuracy in comparison to the VGG based baseline. In particular, this model performs well on *number* and *color*, improving *HieCoAtten* by almost 3 points and 5 points respectively. This confirms that improving the visual representation by models that are semantically richer and have better localization abilities is critical for VQA systems. By further integrating the semantically guided loss, CoAtten+WAV(MIL) gains roughly another additional 0.6 point improvement overall. It's interesting to note that the semantic model makes the biggest contribution on the *object* category, suggesting that it is complementary to the MIL model which is beneficial for the *color* and *number* categories. This is un-

[3] Now, VQA has two versions. We started preparing this manuscript before the release of VQA 2.0. All the results are reported on VQA 1.0.

derstandable because word2vec provides better modeling of semantic relatedness on object concepts than on other types of word categories. For example, it is evident that *car* is semantically closer to *road* than to *donut* while it is not so certain that *red* is closer to *black* than to *green*.

Our ResNet based model CoAtten+WAV(ReseNet) achieves the best performance in comparison to our other models as well as to *HieCoAtten*. This is not surprising as this architecture has proven to be superior to other architectures on many computer vision tasks. We speculate that ResNet features trained in the MIL paradigm would perform even better.

Note that *A-C-VQA* outperforms our approach (which is the second best) by a large margin on COCO-QA. Since the attribute-based representation used by *A-C-VQA* was learned directly from MS COCO image captions, it is not surprising that it is more effective on COCO-QA questions and answers which are automatically generated from the same captions. However, as shown in Table 2, the same model, when applied to the VQA dataset, gives inferior performance in comparison to several other recently developed approaches including ours. It is worth noting that our MIL model was also trained on MS COCO captions. However, our model generalizes well to the VQA dataset.

We also evaluated our models on the top-5 candidate answers, as this measure seems more suitable to capture the intuition behind the WAV loss. It is observed from Figure 3, the overall accuracy improves on COCO-QA from 87.9% HieCoAtten(VGG) to 90.8% HieCoAtten+WAV(MIL), showing a more significant improvement than we get for top-1. It means that for those challenging cases that all the methods fail, the proposed method makes more good estimations than the baseline. The illustrative examples are shown in Figure 6.

Evaluation on VQA. In Table 2 we compare the results of the different models on the VQA test sets for both open-ended and multiple-choice settings. Our models outperform the corresponding VGG and ResNet *HieCoAtten* models respectively. Specifically, comparing the two VGG based models, our CoAtten+WAV(MIL) model obtains 1.3 point accuracy improvement on test-dev in the open-ended setting. Also, as expected, compared to *HieCoAtten*, most of the improvement is in the *other* category, which includes questions related to objects and attributes such as color.

We emphasize that our approach is general and can be combined with any supervised learning approach simply by adding a semantically guided component to the loss function. In this paper we combined our approach with *HieCoAtten* [18] which was the top performer when we initiated this work. Hence the most meaningful comparison is to that model. However, since then, the recent RAU model [25] obtained superior performance, reporting 81.9%, 39.0%, 53.0%, 63.3% for "Y/N", "Num", "Other" and "All" in open-ended test-dev setting, respectively. We combine our approach with RAU to further boost the performance to 82.2%, 39.3%, 54.4% and 64.2%. RAU is not listed in Table 2, as it is not official published. We also point out that two recent papers [8] and [15], are not listed in Table 2 as they measure performance on the 3000 and 2000 most frequent answers in the dataset respectively. Hence they are incomparable to all previous works as well as ours which look at the most 1000 frequent answers.

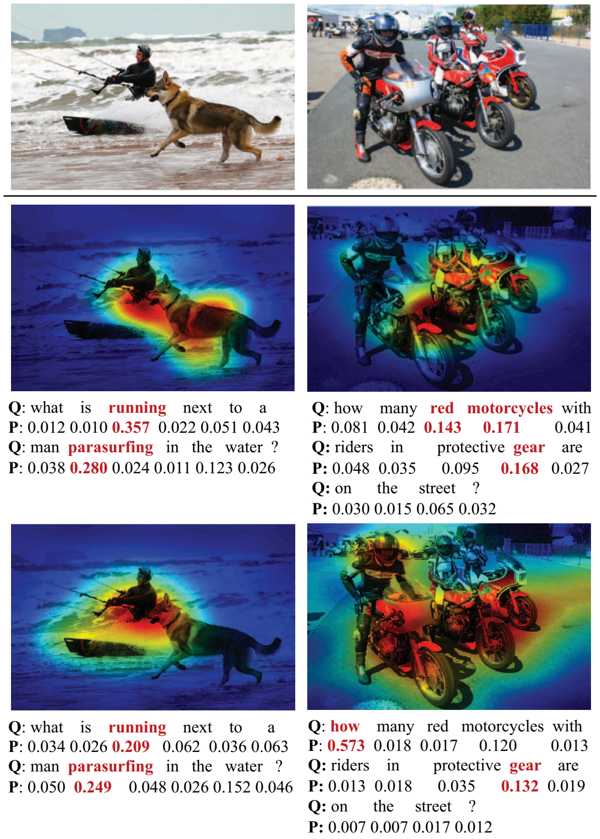

Q: what is running next to a
P: 0.012 0.010 0.357 0.022 0.051 0.043
Q: man parasurfing in the water ?
P: 0.038 0.280 0.024 0.011 0.123 0.026
Q: how many red motorcycles with
P: 0.081 0.042 0.143 0.171 0.041
Q: riders in protective gear are
P: 0.048 0.035 0.095 0.168 0.027
Q: on the street ?
P: 0.030 0.015 0.065 0.032
Q: what is running next to a
P: 0.034 0.026 0.209 0.062 0.036 0.063
Q: man parasurfing in the water ?
P: 0.050 0.249 0.048 0.026 0.152 0.046
Q: how many red motorcycles with
P: 0.573 0.018 0.017 0.120 0.013
Q: riders in protective gear are
P: 0.013 0.018 0.035 0.132 0.019
Q: on the street ?
P: 0.007 0.007 0.017 0.012

Figure 4. Visualization of attention maps on the COCO-QA dataset. The top row is the original testing images. The second and third rows are the corresponding attention maps and words from our *CoAtten+WAV*(MIL) model and HieCoAtten^a(VGG) [18], respectively. Best viewed in color.

Attention Visualization. In Figure 4 we visualize the attention maps generated on examples from the COCO-QA test set by our model, CoAtten+WAV(MIL), as well as the HieCoAtten(VGG) model. The images on the left are the original ones. Images on the bottom two rows are the attention heat maps, where red denotes the high probability region of attention, and blue represents the low attended regions. For each example, we also write the question with the probability associated with each word. We highlight the few steering words in each question, that match the attention map. As observed in the first example in Figure 4, the steering words are the verbs "running" and "parasurfing" with

Table 2. Results on the VQA dataset under open-ended (left) and multiple-choice (right) settings. “-” indicates no results are available.

	Open-Ended					Multiple-Choice				
	test-dev				test-std	test-dev				test-std
Methods	Y/N	Num	Other	All	All	Y/N	Num	Other	All	All
Region Sel. [30]	-	-	-	-	-	77.6	34.3	55.8	62.4	-
SMem [36]	80.9	37.3	43.1	58.0	58.2	-	-	-	-	-
SAN [37]	79.3	36.6	46.1	58.7	58.9	-	-	-	-	-
FDA [12]	81.1	36.2	45.8	59.2	59.5	81.5	39.0	54.7	64.0	64.2
DMN+ [35]	80.5	36.8	48.3	60.3	60.4	-	-	-	-	-
DPPnet [26]	80.7	37.2	41.7	57.2	57.4	80.8	38.9	52.2	62.5	62.7
C-A-VQA [34]	81.0	38.4	45.2	59.4	59.2	81.1	37.1	45.8	59.4	-
HieCoAtten^a(VGG) [18]	79.6	38.4	49.1	60.5	-	79.7	40.1	57.9	64.9	-
HieCoAtten^a(ResNet) [18]	79.7	38.7	51.7	61.8	62.1	79.7	40.0	59.8	65.8	66.1
CoAtten (MIL, ours)	79.6	38.6	50.6	61.2	61.5	79.6	40.4	58.3	65.2	65.4
CoAtten+WAV (MIL, ours)	79.8	38.6	51.3	61.8	62.0	79.9	39.8	59.4	65.6	65.9
CoAtten+WAV (ResNet, ours)	80.2	38.5	52.4	62.3	62.2	80.2	40.3	60.5	66.4	66.3

the highest probabilities of 0.357 and 0.280, respectively. In the heat map generated by HieCoAtten(VGG) (bottom left corner), the attention lies on the noun “man”. Even though the verb “running” gets relatively high probability of 0.209, there is no response on the running dog region. Due to the MIL in modeling actions, our model responds intensively to the verb ”running”. Similarly, in the second example, the adjective word “red” as well as the nouns “motorcycles” and “gear” get the highest response in the question. It is observed that our attention heat map is more accurate, concentrated on the red motorcycles, compared to the one generated by HieCoAtten(VGG). In this example, our model successfully attends the region of interest, and as a result predicts the correct answer “three”.

Sensitivity Analysis. An important parameter in our approach is λ in Eq. (1), which balances the classification loss and the semantically guided loss. We performed sensitivity analysis of λ on COCO-QA to understand how it influences the performance of our approach. As seen in Figure 7, either a too small λ or a too larger λ results in a noticeable performance drop while a value between 0.1 and 10 works reasonably well with our approach. The best accuracy (64.5%) of our approach is achieved when λ is set to 0.1.

Illustrative Examples. To better understand our approach, we provide more examples in Figure 5 and Figure 6, and visualize the top-5 predictions as well as their prediction scores from the baseline and our proposed approach, i.e.,*HieCoAtten*(A1), *MIL-CoAtten* (A2) and *CoAtten+WAV*(MIL) (A3). Ground-truth labels (if predicted) are colored in red while predicted answers semantically related to the correct answers are marked as blue.

We first look at a few examples where our approach works well. Figure 5(a, b) illustrate that our approach (A2) correctly predicts *purple* and *four* as the answers while the baseline does not, clearly demonstrating the advantage of the semantically enriched MIL representation over a general CNN representation. Figure 5(c, d) further show that our semantic-guided model (A3) can push up weak predictions to the top with support from relevant predictions (i.e. *donut* → *cake* and *street* → *cart*). Note that the more top predictions in the semantically guided approach become relevant, clearly demonstrating the efficacy of our proposed ideas.

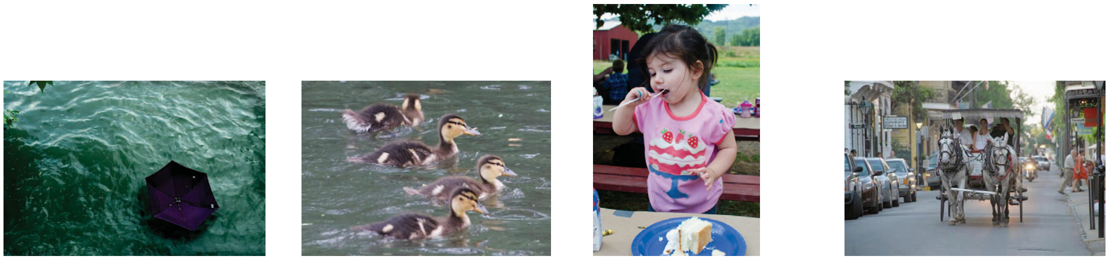
For typical failure cases, We break them down into 4 categories, as shown in Figure 6. From left to right, we summarize the causes of failure as (1) *plausible answers*, (2) *attention errors* (3) *lack of semantic support in prediction* (4) *plural/singular words*. We will describe each case below in details.

The first case refers to the model giving a wrong but plausibly correct answer. As mentioned in Section 3, one desired feature that our model owns is that it tends to predict a semantically plausible answer when having problem identifying the correct one. This is illustrated in Figure 6(a) where our semantically guided model (A3) chooses *suitcase* as the correct answer, which is not bad in comparison to the ground truth answer *bags*.

The second type of errors results from the baseline when it fails to focus attention on the right region(s) in the images. In such a case, the approach tends to produce random answers less relevant to the questions. As seen in Figure 6(b), none of the top-4 predictions are actually related to the correct answer *skateboard*.

The third type of errors occurs when the correct answer, though predicted as one of the top answers, cannot possess sufficient support from other top answers. An example is given in Figure 6(c) where *scissors* appear to be semantically distant from other predictions. In the future we plan to identify relevant concepts in the question along the same direction to address this issue.

Lastly, our approach occasionally switches an answer of singular word to its plural form or vice versa, as shown in Figure 6(d). This confusion is largely because singular-



Q: what is the color of the upside-down ?

A1: red:0.89 **purple**:0.08 green:0.01 blue: 0.01 black:0.00
A2: **purple**:0.35 red:0.35 black:0.13 blue: 0.12 green:0.02
A3: **purple**:0.44 blue:0.22 red:0.18 black: 0.08 green:0.04

Q: how many ducks are swimming on top of the water ?

A1: three:0.76 **four**:0.13 two:0.11 five: 0.00 one:0.00
A2: **four**:0.59 five:0.25 three:0.08 seven: 0.04 six:0.03
A3: **four**:0.48 three:0.48 five:0.03 two: 0.01 six:0.00

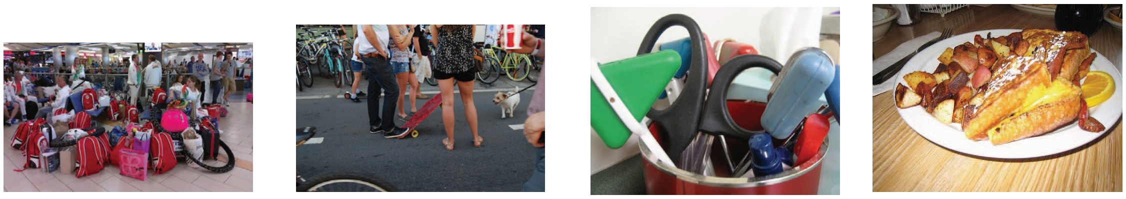
Q: what is the girl wearing a t-shirt with a cake decoration on it eats ?

A1: donut:0.63 fork:0.15 sandwich: 0.05 **cake**:0.03 pastry:0.03
A2: **cake**:0.32 fork:0.15 donut:0.14 toothbrush:0.05 plate:0.05
A3: **cake**:0.70 donut:0.15 toothbrush: 0.05 pastry:0.02 candles:0.02

Q: where are two white horses pulling some people ?

A1: street:0.41 **cart**:0.30 carriage:0.19 wagon:0.07 road:0.03
A2: street:0.75 **cart**:0.08 car:0.05 road:0.02 wagon:0.02
A3: **cart**:0.51 street:0.27 wagon:0.05 carriage:0.04 road:0.04

Figure 5. Successful examples of our proposed approach on COCO-QA dataset [29]. In text, we index each example (a) to (d) from left to right. Followed by the question (start by "Q"), we list three sets of top-5 predicted answers marked as "A1", "A2" and "A3". They represent the co-attention baseline (*HieCoAtten*), the MIL-based cross entropy method (*MIL-CoAtten*) and the MIL-based average vector method (*CoAtten+WAV*(MIL)), respectively. All the predicted answers are followed by their prediction probabilities. Ground-truth labels are colored in red. The related predictions are colored in blue for the ease of illustration.



Q: what are grouped together in the waiting area ? (GT: **bags**)

A1: bicycle:0.99 motorcycle:0.01 bicycles: 0.00 horse:0.00 scooter:0.00
A2: toys:0.61 motorcycles:0.13 scooters: 0.11 dogs:0.08 hats:0.07
A3: suitcases:0.77 luggage:0.09 bags:0.01 boards:0.01 surfboards:0.01

Q: what is the man riding down the street ?

A1: bicycle:0.99 motorcycle:0.01 bicycles: 0.00 horse:0.00 scooter:0.00
A2: bicycle:1.00 motorcycle:0.00 bicycles: 0.00 scooter:0.00 **skateboard**:0.00
A3: bicycle:0.98 motorcycle:0.01 bicycles: 0.00 scooter:0.00 **skateboard**:0.00

Q: what resting in the cup with markers and other tools ?

A1:container:8.70 cup:7.54 bucket:7.53 toothbrush:7.00 case:6.29 device:5.98
A2: spoon:0.32 cup:0.24 cups:0.04 container:0.04 **scissors**:0.04
A3: spoon:0.08 cup:0.07 tray:0.06 device:0.05 **scissors**:0.04

Q: what filled with different food on a table ?

A1: plates:0.84 **plate**:0.16 tray:0.00 bowl:0.00 platter:0.00
A2: **plate**:0.70 plates:0.30 tray:0.00 platter:0.00 bowl:0.00
A3: plates:0.83 **plate**:0.16 tray:0.00 containers:0.00 bowl:0.00

Figure 6. Failure examples of our proposed approach on COCO-QA dataset [29]. In text, we index each example (a) to (d) from left to right. Followed by the question (start by "Q"), we list three sets of top-5 predicted answers marked as "A1", "A2" and "A3". They represent the co-attention baseline (*HieCoAtten*), the MIL-based cross entropy method (*MIL-CoAtten*) and the MIL-based average vector method (*CoAtten+WAV*(MIL)), respectively. Ground-truth labels are colored in red. The semantically related answers are colored in blue.

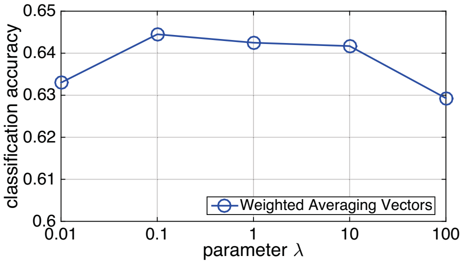


Figure 7. Sensitivity analysis of parameter λ on COCO-QA.

plural relationships are contextually dependent and hard to be accurately described by word2vec. Answers based on singular or plural words are not uncommon in our data, especially in COCO-QA. Fortunately, their impact is not just one way, leaving it out as a major concern.

5. Conclusion

In this paper we suggested two new ideas in the context of visual question answering. The first is to inject external semantic knowledge during training by designing a loss function which takes into account the semantic relatedness between the predictions and the ground truth. The second is to enrich the visual representation of attention models by extracting visual features from a MIL model. This provides representations for concepts beyond objects such as actions, colors etc. We applied these ideas on two VQA benchmarks and obtained competitive performance comparable to the state of the art.

Acknowledgment: This work was done when Handong Zhao was an intern at IBM research. This research is supported in part by the NSF IIS award 1651902, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Award W911NF-17-1-0367. The authors also thank Daniel M. Gruen for the valuable discussions.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. In *NAACL*, pages 1545–1554, 2016.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. In *ICCV*, pages 2425–2433, 2015.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. Dbpedia: A nucleus for a web of open data. In *The 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, pages 722–735, 2007.

[4] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[6] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015.

[7] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *CoRR*, abs/1606.01847, 2016.

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

[9] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *NIPS*, pages 2296–2304, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

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

[12] I. Ilievski, S. Yan, and J. Feng. A focused dynamic attention model for visual question answering. *CoRR*, abs/1604.01485, 2016.

[13] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. *CoRR*, abs/1606.08390, 2016.

[14] J. Kim, S. Lee, D. Kwak, M. Heo, J. Kim, J. Ha, and B. Zhang. Multimodal residual learning for visual QA. *CoRR*, abs/1606.01455, 2016.

[15] J. Kim, K. W. On, J. Kim, J. Ha, and B. Zhang. Hadamard product for low-rank bilinear pooling. *CoRR*, abs/1610.04325, 2016.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[17] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.

[18] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061, 2016.

[19] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In *AAAI*, pages 3567–3573, 2016.

[20] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, pages 1682–1690, 2014.

[21] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, pages 1–9, 2015.

[22] O. Maron and T. Lozano-Perez. A framework for multiple-instance learning. In *NIPS*, 1998.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[24] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Hierarchical multimodal LSTM for dense visual-semantic embedding. In *ICCV*, pages 1899–1907, 2017.

[25] H. Noh and B. Han. Training recurrent answering units with joint loss minimization for vqa. *arXiv preprint arXiv:1606.03647*, 2016.

[26] H. Noh, P. H. Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. *arXiv preprint arXiv:1511.05756*, 2015.

[27] M. Pagliardini, P. Gupta, and M. Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. *CoRR*, abs/1703.02507, 2017.

[28] A. Ray, G. Christie, M. Bansal, D. Batra, and D. Parikh. Question relevance in VQA: identifying non-visual and false-premise questions. In *EMNLP*, pages 919–924, 2016.

[29] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *NIPS*, pages 2953–2961, 2015.

[30] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. *CoRR*, abs/1511.07394, 2015.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[32] H. Song, J. M. R. Girshick anf S. Jegelka, Z. Harachaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *ICML*, 2014.

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[34] Q. Wu, C. Shen, A. van den Hengel, P. Wang, and A. R. Dick. Image captioning and visual question answering based on attributes and their related external knowledge. *CoRR*, abs/1603.02814, 2016.

[35] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, pages 2397–2406, 2016.

[36] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, pages 451–466, 2016.

[37] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015.