Generating Question Relevant Captions to Aid Visual Question Answering

Jialin Wu, Zeyuan Hu and Raymond J. Mooney
Department of Computer Science
University of Texas at Austin
{jialinwu, iamzeyuanhu, mooney}@cs.utexas.edu

Abstract

Visual question answering (VQA) and image captioning require a shared body of general knowledge connecting language and vision. We present a novel approach to improve VQA performance that exploits this connection by jointly generating captions that are targeted to help answer a specific visual question. The model is trained using an existing caption dataset by automatically determining question-relevant captions using an online gradient-based method. Experimental results on the VQA v2 challenge demonstrate that our approach obtains state-of-the-art VQA performance (e.g. 68.4% on the Test-standard set using a single model) by simultaneously generating question-relevant captions.

1 Introduction

In recent years, visual question answering (VQA) (Antol et al., 2015) and image captioning (Donahue et al., 2015; Rennie et al., 2017) have been widely studied in both the computer vision and NLP communities. Most recent VQA research (Lu et al., 2017; Pedersoli et al., 2017; Anderson et al., 2018; Lu et al., 2018) concentrates on directly utilizing visual input features including detected objects, attributes, and relations between pairs of objects.

However, little VQA research works on exploiting textual features from the image which are able to tersely encode the necessary information to answer the questions. This information could be richer than the visual features in that the sentences have fewer structural constraints and can easily include the attributes of and relation among multiple objects. In fact, we observe that appropriate captions can be very useful for many VQA questions. In particular, we trained a model to answer visual questions for the VQA v2 challenge (Antol et al., 2015) only using the human annotated captions without images and achieved a score of 59.6%, outperforming a large number of VQA models that use image features. Existing work using captions for VQA has generated question-agnostic captions using a pretrained captioner (Li et al., 2018a). This approach can provide additional general information; however, this information is not guaranteed to be relevant to the given VQA question.

We explore a novel approach that generates question-relevant image descriptions, which contain information that is directly relevant to a particular VQA question. Fig. 1 shows examples of our generated captions given different questions. In order to encourage the generation of relevant captions, we propose a novel greedy algorithm that aims to minimize the cross entropy loss only for
the most relevant and helpful gold-standard captions. Specifically, helpfulness is measured using the inner-product of the gradients from the caption generation loss and the VQA answer prediction loss. A positive inner-product means the two objective functions share some descent directions in the optimization process, and therefore indicates that the corresponding captions help the VQA training process.

In order to incorporate the caption information, we propose a novel caption embedding module that, given the question and image features for a visual question, recognizes important words in the caption, and produces a caption embedding tailored for answer prediction. In addition, the caption embeddings are also utilized to adjust the visual top-down attention weights for each object.

Furthermore, generating question-relevant captions ensures that both image and question information is encoded in their joint representations, which reduces the risk of learning from question bias (Li et al., 2018a) and ignoring the image content when high accuracy can be achieved from the questions alone.

Experimental evaluation of our approach shows significant improvements on VQA accuracy over our baseline Up-Down (Anderson et al., 2018) model on the VQA v2 validation set (Antol et al., 2015), from 63.2% to 67.1% with gold-standard human captions from the COCO dataset (Chen et al., 2015) and 65.8% with automatically generated question-relevant captions. Our single model is able to score 68.4% on the test-standard split, and an ensemble of 10 models scores 69.7%.

2 Background Related Work

2.1 Visual Question Answering

Recently, a large amount of attention-based deep-learning methods have been proposed for VQA, including top-down (Ren et al., 2015a; Fukui et al., 2016; Wu et al., 2016; Goyal et al., 2017; Li et al., 2018a) and bottom-up attention methods (Anderson et al., 2018; Li et al., 2018b; Wu and Mooney, 2019). Specifically, a typical model first extracts image features using a pre-trained CNN, and then trains an RNN to encode the question, using an attention mechanism to focus on specific features of the image. Finally, both question and attended image features are used to predict the final answer.

However, answering visual questions requires not only information about the visual content but also common knowledge, which is usually too hard to directly learn from only a limited number of images with human annotated answers as supervision. However, comparatively little previous VQA research has worked on enriching the knowledge base. We are aware of two related papers. Li et al. (2018a) use a pre-trained captioner to generate general captions and attributes with a fixed annotator and then use them to predict answers. Therefore, the captions they generate are not necessarily relevant to the question, and they may ignore image features needed for answer prediction. Narasimhan et al. (2018) employed an out-of-the-box knowledge base and trained their model to filter out irrelevant facts. After that, graph convolutional networks use this knowledge to build connections to the relevant facts and predict the final answer. Unlike them, we generate captions to provide information that is directly relevant to the VQA process.

2.2 Image Captioning

Most recent image captioning models are also attention-based deep-learning models (Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Luo et al., 2018; Liu et al., 2018). With the help of large image description datasets (Chen et al., 2015), these models have demonstrated remarkable results. Most of them encode the image using a CNN, and build an attentional RNN (i.e. GRU (Cho et al., 2014), LSTM (Hochreiter and Schmidhuber, 1997)) on top of the image features as a language model to generate image captions.

However, deep neural models still tend to generate general captions based on the most significant objects (Vijayakumar et al., 2016). Although previous works (Luo et al., 2018; Liu et al., 2018) build captioning models that are encouraged to generate different captions with discriminability objectives, the captions are usually less informative and fail to describe most of the objects and their relationships diversely. In this work, we develop an approach to generating captions that directly focus on the critical objects in the VQA process and provide information that can help the VQA module predict the answer.

3 Approach

We first describe the overall structure of our joint model in Sec. 3.1 and explain the foundational
feature representations (i.e., image, question and caption embeddings) in Sec. 3.2. Then, the VQA module is presented in Sec. 3.3, which takes advantage of the generated image captions to improve the VQA performance. In Sec. 3.4, we explain the image captioning module which generates question-relevant captions. Finally, the training and implementation details are provided in Sec. 3.5.

3.1 Overview
As illustrated in Fig. 2, the proposed model first extracts image features \( V = \{ v_1, v_2, \ldots, v_K \} \) using bottom-up attention and question features \( q \) to produce their joint representation and then generates question-related captions. Next, our caption embedding module encodes the generated captions as caption features \( c \) as detailed in Sec. 3.2. After that, both question features \( q \) and caption features \( c \) are utilized to generate the visual attention \( A^{qv} \) to weight the images’ feature set \( V \), producing attended image features \( V^q \). Finally, we add \( V^q \) to the caption features \( c \) and further perform element-wise multiplication with the question features \( q \) (Anderson et al., 2018) to produce the joint representation of the question, image and caption, which is then used to predict the answer.

3.2 Feature Representation
In this section, we explain the details of this joint representation. We use \( f(x) \) to denote fully-connected layers, where \( f(x) = \text{LReLU}(Wx + b) \) with input features \( x \) and ignore the notation of weights and biases for simplicity, where these \( fc \) layers do not share weights. LReLU denotes a Leaky ReLU (He et al., 2015).

Image and Question Embedding
We use object detection as bottom-up attention (Anderson et al., 2018), which provides salient image regions with clear boundaries. In particular, we use a Faster R-CNN head (Ren et al., 2015b) in conjunction with a ResNet-101 base network (He et al., 2016) as our detection module. The detection head is first pre-trained on the Visual Genome dataset (Krishna et al., 2017) and is capable of detecting 1,600 objects categories and 400 attributes. To generate an output set of image features \( V \), we take the final detection outputs and perform non-maximum suppression (NMS) for each object category using an IoU threshold of 0.7. Finally, a fixed number of 36 detected objects for each image are extracted as the image features (a 2,048 dimensional vector for each object) as suggested by Teney et al. (2017).

For the question embedding, we use a standard GRU (Cho et al., 2014) with 1,280 hidden units and extract the output of the hidden units at the final time step as the question features \( q \). Following Anderson et al. (2018), the question features \( q \) and image feature set \( V \) are further embedded together to produce a question-attained image feature set \( V^q \) via question visual-attention \( A^{qv} \) as illustrated in Fig. 2.

Caption Embedding
Our novel caption embedding module takes as in-
put the question-attended image feature set $V^q$, question features $q$, and $C$ captions $W^c_i = \{w^c_{i,1}, w^c_{i,2}, \ldots, w^c_{i,T}\}$, where $T$ denotes the length of the captions and $i = 1, \ldots, C$ are the caption indices, and then produces the caption features $c$.

$$h^1_{i,t} = \text{GRU}(W_c h^1_{i,t-1}, h^1_{i,t-1}) \quad (1)$$

where $W_c$ is the word embedding matrix, and $\Pi^c_{i,t}$ is the one-hot embedding for the word $w^c_{i,t}$.

Then, we design a caption attention module $A^c$ which utilizes the question-attended feature set $V^q$, question features $q$, and $h^1_{i,t}$ to generate the attention weight on the current word in order to indicate its importance. Specifically, the Word GRU first encodes the words embedding $\Pi^c_{i,t}$ in Eq. 1, and then we feed the outputs $h^1_{i,t}$ and $V^q$ to the attention module $A^c$ as shown in Eq. 4.

$$V^q = \sum_{k=1}^{K} v^q_k \quad (2)$$

$$a^c_{i,t} = h^1_{i,t} \circ f(V^q) + h^1_{i,t} \circ f(q) \quad (3)$$

$$\alpha^c_{i,t} = \sigma(a^c_{i,t}) \quad (4)$$

where $\sigma$ denotes the sigmoid function, and $K$ is the number of objects in the bottom-up attention.

Next, the attended words in the caption are used to produce the final caption representation in Eq. 5 via the Caption GRU. Since the goal is to gather more information, we perform element-wise max pooling across the representations of all of the input captions $c_i$ in Eq. 7.

$$h^2_{i,t} = \text{GRU}(\alpha^c_{i,t} W_c, h^2_{i,t-1}) \quad (5)$$

$$c_i = f(h^2_{i,T}) \quad (6)$$

$$c = \text{max}(c_i) \quad (7)$$

where $\text{max}$ denotes the element-wise max pooling across all of caption representations $c_i$ of the image.

### 3.3 VQA Module

This section describes the details of the VQA module. The generated captions are usually capable of capturing relations among the question-relevant objects; however these relations are absent in the bottom-up attention. Therefore, our VQA module utilizes the caption embeddings $c$ to adjust the top-down attention weights in VQA in order to produce the final caption-attended features $\nu^{cv}$ in Eq. 10:

$$a^c_{i,k} = f(f(c) \circ f(V^q_k)) \quad (8)$$

$$\alpha^c_{i,k} = \text{softmax}(a^c_{i,k}) \quad (9)$$

$$\nu^{cv} = \sum_{k} v^q_k \alpha^c_{i,k} \quad (10)$$

where $k$ traverses the $K$ objects features.

To better incorporate the information from the captions into the VQA process, we add the caption features $c$ to the attended image features $\nu^{cv}$, and then element-wise multiply by the question features as shown in Eq. 11:

$$h = q \circ (f(\nu^{cv}) + f(c)) \quad (11)$$

$$\hat{s} = \sigma(f(h)) \quad (12)$$

We frame the answer prediction task as a multi-label regression problem (Anderson et al., 2018). In particular, we use the soft scores in the gold-standard VQA-v2 data (which are used in the evaluation metric), as labels to supervise the sigmoid-normalized predictions as shown in Eq. 13:

$$\mathcal{L}^{vqa} = -\sum_{j=1}^{N} s_j \log \hat{s}_j + (1-s_j) \log(1-\hat{s}_j) \quad (13)$$
where the index \( j \) runs over \( N \) candidate answers and \( s \) are the soft answer scores.

In case of multiple feasible answers, the soft scores capture the occasional uncertainty in the ground-truth annotations. As suggested by Teney et al. (2017), we collect the candidate answers that appear more than 8 times in the training set, which results in 3, 129 answer candidates.

### 3.4 Image Captioning Module

We adopt an image captioning module similar to that of Anderson et al. (2018), which takes the object detection features as inputs and learns attention weights over those objects’ features in order to predict the next word at each step. The key difference between our module and theirs lies in the input features and the caption supervision. Specifically, we use the question-attended image features \( V^q \), as inputs, and only use the most relevant caption, which is automatically determined in an online fashion (detailed below), for each question-image pair to train the captioning module. This ensures that only question-relevant captions are generated.

### Selecting Relevant Captions for Training

Previously, Li et al. (2018b) selected relevant captions for VQA based on word similarities between captions and questions, however, their approach does not take into account the details of the VQA process. In contrast, during training, our approach dynamically determines for each problem, the caption that will most improve VQA. We do this by updating with a shared descent direction (Wu et al., 2018) which decreases the loss for both captioning and VQA. This ensures a consistent target for both the image captioning module and the VQA module in the optimization process.

During training, we compute the cross-entropy loss for the \( i \)-th caption using Eq. 14, and back-propagate the gradients only from the most relevant caption determined by solving Eq. 15.

\[
L^c_i = - \sum_{t=1}^{T} \log(p(w^c_{i,t} | w^c_{i,t-1})) \tag{14}
\]

In particular, we require the inner product of the current gradient vectors from the predicted answer and the human captions to be greater than a positive constant \( \xi \), and further select the caption that maximizes that inner product.

\[
\arg \max_i \sum_{k=0}^{K} \left( \frac{\partial \hat{s}_{\text{pred}}}{\partial W^q_k} \right)^T \frac{\partial \log(p(W^c_i))}{\partial W^q_k} \tag{15}
\]

\[s.t. \quad \sum_{k=0}^{K} \left( \frac{\partial \hat{s}_{\text{pred}}}{\partial W^q_k} \right)^T \frac{\partial \log(p(W^c_i))}{\partial W^q_k} > \xi \]

where the \( \hat{s}_{\text{pred}} \) is the logit\(^1\) for the predicted answer, \( W^c_i \) denotes the \( i \)-th human caption for the image and \( k \) traverses the \( K \) object features.

Therefore, given the solution to Eq. 15, \( i^* \), the final loss of our joint model is the sum of the VQA loss and the captioning loss for the selected captions as shown in Eq. 16. If Eq. 15 has no feasible solution, we ignore the caption loss.

\[
L = L^{vqa} + L^c_{i^*}. \tag{16}
\]

### 3.5 Training and Implementation Details

We train our joint model using the AdaMax optimizer (Kingma and Ba, 2015) with a batch size of 384 and a learning rate of 0.002 as suggested by Teney et al. (2017). We use the validation set for VQA v2 to tune the initial learning rate and the number of epochs, yielding the highest overall VQA score. We use 1, 280 hidden units in the question embedding and attention model in the VQA module with 36 object detection features for each image. For captioning models, the dimension of the LSTM hidden state, image feature embedding, and word embedding are all set to 512. We also use Glove vectors (Pennington et al., 2014) to initialize the word embedding matrix in the caption embedding module.

We initialize the training process with human annotated captions from the COCO dataset (Chen et al., 2015) and pre-train the VQA and caption-generation modules for 20 epochs with the final joint loss in Eq. 16. After that, we generate question-relevant captions for all question-image pairs in the COCO train, validation, and test sets. In particular, we sample 5 captions per question-image pair. We fine-tune our model using the generated captions with 0.25 × learning-rate for another 10 epochs.

### 4 Experiments

We perform extensive experiments and ablation studies to evaluate our joint model on VQA.

\(^1\)The input to the softmax function.
<table>
<thead>
<tr>
<th></th>
<th>Test-standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes/No</td>
</tr>
<tr>
<td>Prior (Goyal et al., 2017)</td>
<td>61.20</td>
</tr>
<tr>
<td>Language-only (Goyal et al., 2017)</td>
<td>67.01</td>
</tr>
<tr>
<td>MCB (Fukui et al., 2016)</td>
<td>78.82</td>
</tr>
<tr>
<td>Up-Down (Anderson et al., 2018)</td>
<td>82.20</td>
</tr>
<tr>
<td>VQA-E (Li et al., 2018b)</td>
<td>83.22</td>
</tr>
<tr>
<td>Ours (single)</td>
<td><strong>84.69</strong></td>
</tr>
<tr>
<td>Ours (Ensemble-10)</td>
<td><strong>86.15</strong></td>
</tr>
</tbody>
</table>

Table 1: Comparison of our results on VQA with the state-of-the-art methods on the test-standard data. Accuracies in percentage (%) are reported.

4.1 Datasets and Evaluation Metrics

**VQA Dataset**

We use the VQA v2.0 dataset (Antol et al., 2015) for the evaluation of our proposed joint model, where the answers are balanced in order to minimize the effectiveness of learning dataset priors. This dataset is used in the VQA 2018 challenge and contains over 1.1M questions from the over 200K images in the MSCOCO 2015 dataset (Chen et al., 2015).

Following Anderson et al. (2018), we perform standard text pre-processing and tokenization. In particular, questions are first converted to lower case and then trimmed to a maximum of 14 words, and the words that appear less than 5 times are replaced with an “<unk>” token. To evaluate answer quality, we report accuracies using the official VQA metric using soft scores, which accounts for the occasional disagreement between annotators for the ground truth answers.

**Image Captioning Dataset**

We use the MSCOCO 2014 dataset (Chen et al., 2015) for the image caption module. To maintain consistency with the VQA tasks, we use the dataset’s official configuration that includes 82,372 images for training and 40,504 for validation. Similar to the VQA question pre-processing, we first convert all sentences to lower case, tokenizing on white spaces, and filtering words that do not occur at least 5 times.

4.2 Results on VQA

We first report the experimental results on the VQA task and compare our results with the state-of-the-art methods in this section. After that, we perform ablation studies to verify the contribution of additional knowledge from the generated captions, and the effectiveness of using caption representations to adjust the top-down visual attention weights.

As demonstrated in Table 1, our single model outperforms other state-of-the-art single models by a clear margin, i.e., 2.06%, which indicates the effectiveness of including caption features as additional inputs. In particular, we observe that our single model outperforms other methods, especially in the ‘Num’ and ‘Other’ categories. This is because the generated captions are capable of providing more numerical clues for answering the ‘Num’ questions, since the captions can describe the number of relevant objects and provide general knowledge for answering the ‘Other’ questions. Furthermore, an ensemble of 10 models with different initialization seeds results in a score of 69.7% for the test-standard set.

Fig. 4 shows several examples of our generated question-relevant captions. These examples illustrate how different captions are generated for the same image when the question is changed. They also show how the objects in the image that are important to answering the question are described in the question-relevant captions.

**Comparison Between Using Generated and Human Captions**

Next, we analyze the difference between using automatically generated captions and using those provided by human annotators. In particular, we train our model with generated question-agnostic captions using the Up-Down (Anderson et al., 2018) captioner, question-relevant captions from our caption generation module, and human annotated captions from the COCO dataset.

As demonstrated in Table 2, our model gains
Table 2: Comparison of the performance using generated and human captions. Both of them provide significant improvements to the baseline model. However, there is still a reasonable gap between generated and human captions.

<table>
<thead>
<tr>
<th></th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-Down (Anderson et al., 2018)</td>
<td>63.2</td>
</tr>
<tr>
<td>Ours with Up-Down captions</td>
<td>64.6</td>
</tr>
<tr>
<td>Ours with our generated captions</td>
<td>65.8</td>
</tr>
<tr>
<td>Ours with human captions</td>
<td><strong>67.1</strong></td>
</tr>
</tbody>
</table>

Figure 4: Examples of our generated question-relevant captions. The influential objects with attention weights greater than 0.1 are indicated by bounding boxes (annotated with their visual attention weights in the blue box), and the gray-scale levels in the caption words indicate the word attentions from the caption embedding module.

Effectiveness of Adjusting Top-Down Attention
In this section, we quantitatively analyze the effectiveness of utilizing captions to adjust the top-down attention weights, in addition to the advantage of providing additional information. In particular, we compare our model with a baseline version where the top-down attention-weight adjustment factor $A_{cv}$ is manually set to 1.0 (resulting in no adjustment).

As demonstrated in Tables 3 and 4, we observe an improvement when using caption features to adjust the attention weights. This indicates that the caption features help the model to more robustly locate the objects that are helpful to the VQA pro-
cess. We use $w\ C.A.A$ to indicate with caption attention adjustment and $w/o\ C.A.A$ to indicate without it. Fig. 5 illustrates an example of caption attention adjustment. Without $C.A.A$, the top-down visual attention focuses on both the yellow surfboard and the blue sail, generating the incorrect answer “yellow and blue.”. However, with “yellow board” in the caption, the caption attention adjustment ($C.A.A$) helps the VQA module focus attention just on the yellow surfboard, thereby generating the correct answer “yellow and red” (since there is some red coloring in the surfboard).

Table 3: Evaluation of the effectiveness of caption-based attention adjustment ($C.A.A$) on the test-standard data. Accuracies in percentage (%) are reported.

<table>
<thead>
<tr>
<th>Test-standard</th>
<th>All</th>
<th>Yes/No</th>
<th>Num</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-Down</td>
<td>65.3</td>
<td>82.2</td>
<td>43.9</td>
<td>56.3</td>
</tr>
<tr>
<td>Ours $w/o\ C.A.A$</td>
<td>67.4</td>
<td>84.0</td>
<td>44.5</td>
<td>57.9</td>
</tr>
<tr>
<td>Ours $w\ C.A.A$</td>
<td>68.4</td>
<td>84.7</td>
<td>46.8</td>
<td>59.3</td>
</tr>
</tbody>
</table>

Table 4: Evaluation of the effectiveness of $C.A.A$ on the validation data. Accuracies in percentage (%) are reported.

<table>
<thead>
<tr>
<th>Validation</th>
<th>All</th>
<th>Yes/No</th>
<th>Num</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-Down</td>
<td>63.2</td>
<td>80.3</td>
<td>42.8</td>
<td>55.8</td>
</tr>
<tr>
<td>Ours $w/o\ C.A.A$</td>
<td>65.2</td>
<td>82.1</td>
<td>43.6</td>
<td>55.8</td>
</tr>
<tr>
<td>Ours $w\ C.A.A$</td>
<td>65.8</td>
<td>82.6</td>
<td>43.9</td>
<td>56.4</td>
</tr>
</tbody>
</table>

Table 5: EMD results comparing the top-down attention weights (with or without caption attention adjustments) to human attention-annotation from the VQA-X dataset. Results are shown for both automatically generated captions and human captions. Lower EMD indicates a closer match to human attention.

The results indicate that caption attention adjustment improves the match between automated attention and human-annotated attention, even though the approach is not trained on supervised data for human attention. Not surprisingly, human captions provide a bit more improvement than automatically generated ones.

5 Conclusion

In this work, we have explored how generating question-relevant image captions can improve VQA performance. In particular, we present a model which jointly generates question-related captions and uses them to provide additional information to aid VQA. This approach only utilizes existing image-caption datasets, automatically determining which captions are relevant to a given question. In particular, we design the training algorithm to only update the network parameters in the optimization process when the caption generation and VQA tasks agree on the direction of change. Our single model joint system outperforms the current state-of-the-art single model for VQA.
Acknowledgement
This research was supported by the DARPA XAI program under a grant from AFRL.

References


Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018b. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. ECCV.


Jialin Wu, Dai Li, Yu Yang, Chandrajit Bajaj, and Xiangyang Ji. 2018. Dynamic Filtering with Large Sampling Field for Convnets. ECCV.

