# EAGLE: An Enhanced Attention-based Strategy by Generating Answers from Learning Questions to a Remote Sensing Image

Yeyang Zhou, Yixin Chen (✉), Yimin Chen, Shunlong Ye, Mingxin Guo,
Ziqi Sha, Heyu Wei, Yanhui Gu, Junsheng Zhou, and Weiguang Qu

School of CS and Technology, Nanjing Normal University, Nanjing, China
22161917@stu.njnu.edu.cn

**Abstract.** Image understanding is an essential research issue for many applications, such as text-to-image retrieval, Visual Question Answering, visual dialog and so forth. In all these applications, how to comprehend images through queries has been a challenge. Most studies concentrate on general images and have obtained desired results, but not for some specific real applications, e.g., remote sensing. To tackle this issue, in this paper, we propose an enhanced attention-based approach, entitled EAGLE, which seamlessly integrates the property of aerial images and NLP. In addition, we contribute the first large-scale remote sensing question answering corpus [1]. Extensive experiments conducted on real data demonstrate that EAGLE outperforms the-state-of-the-art approaches.

## 1 Introduction

Image understanding is expected to investigate semantic information in images, and further helps to make decisions. Great progress has been achieved in its downstream applications including image captioning [4], visual question generation [9] and text-to-image retrieval [15]. The Visual Question Answering (VQA) [10,5,13] task has recently emerged as a more elusive task. It is required to answer a textual question according to an image. Since questions are open-ended, VQA includes many challenges in language representation, object recognition, reasoning and specialized tasks like counting. Typically, VQA systems apply LSTM and CNN individually to extract features from input questions and images, then project the two modalities into a joint semantic space for answer prediction. Existing research mainly focuses on common images. In contrast, few studies focus on specific application scenarios, e.g., remote sensing.

Remote sensing is the science of acquiring information from a location that is distant from the data source [6]. In this paper, we only take into consideration visible photographs of Earth's surface. Investigating concrete information from remote sensing images is an arduous but fundamental task. On the one hand, remote sensing images are demanding to comprehend. To begin with, remote sensing images are typically captured from aeroplanes or satellites, and

[1] https://github.com/rsqa2018/corpus

variable factors thus should be meditated, e.g., position, depression angle, rotation and illumination. Additionally, given equal image size, the smaller the scale is, the wider range the image will cover, and the coarser content it will have. As a result, remote sensing images may contain an enormous amount of local objects in any size, including tiny ones with low resolution and vague contour. Such uncertainty makes the content of remote sensing images even more intricate. Furthermore, plenty of characteristics in remote sensing images may include uncertain semantics from distinct perspectives. It is likely that there is significant variation among queries even about the same image due to their different purposes. But on the other hand, it is vital to understand remote sensing images automatically from desired angles, since the semantic understanding of aerial images has many useful applications in civilian and even military fields, such as disaster monitoring [1], scene classification [2] and military intelligence generation [12].

Inspired by human attention mechanisms, current VQA systems [3,8,7] attempt to focus on parts of interest and have achieved promising performance on general images. However, the resulting algorithms still fail to know where to look when images are informative and tend to struggle in understanding questions for which specialized knowledge is required. For better understanding the question and image contents and their relations, a good VQA algorithm needs to identify global scene attributes, locate objects, identify object attributes, quantity and categories to make appropriate inference. Therefore, advanced attentive approaches ought to be explored in remote sensing settings.

In this paper, we propose an **E**nhanced **A**ttention-based strategy by **G**enerating answers from **L**earning questions to a remote sensing imag**E**, entitled **EAGLE**. It could be thought of as a metaphor of eagle hunting. Features of the image are extracted at first. Just imagine an eagle is hovering at high altitude, locating its potential preys. Then the question is encoded word by word. After every word embedding is encoded, the co-attention unit helps to dynamically focus on the part-of-interest of both visual and textual information. After the eagle begins swooping down, preys are aware of danger and are fleeing in all directions, yet the penetrating eagle adjusts its subduction angle. After the last word of the question is processed, we combine hindmost attended question and image features, modeling answer prediction as a classification task. Eventually big claws grab preys and pull them up.

We argue that bimodal attentions in EAGLE provide complementary information and are effectively integrated in a unified framework. In order to demonstrate the efficacy of EAGLE, we construct the first large-scale remote sensing question answering corpus. Extensive experiments on real data demonstrate that our proposal outperforms the-state-of-the-art approaches.

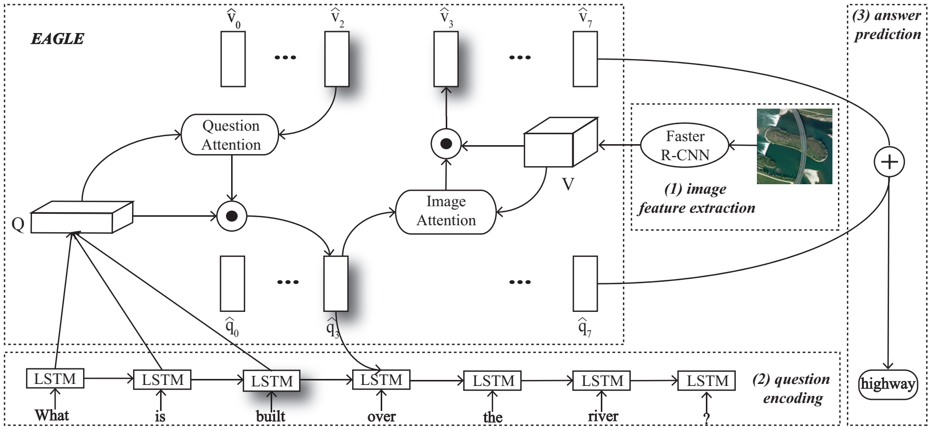

**Fig. 1.** The framework of EAGLE with fused question attention and image attention. EAGLE comprises (1) image feature extraction, (2) question encoding and (3) answer prediction. We incorporate attentive question features $\widehat{\boldsymbol{q}}$ and attentive image features $\widehat{\boldsymbol{v}}$ to jointly reason about visual attention and question attention in a series of steps iteratively. (e.g., "built" in "What is built over the river?" is being encoded.)

## 2 Methodology

### 2.1 Problem Formulation

The remote sensing question answering task considered in this paper, is defined as follows. Before the task, we predefine an answer candidate set with $M$ classes $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_M\}$. Given an aerial image, with its features for $N$ spatial regions represented by $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$ and the question about it with $K$ word embeddings $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_K\}$, output an answer class with the highest prediction score among the answer set $\boldsymbol{A}$. The weights in different modules/layers are denoted with $\boldsymbol{W}$, with appropriate sub/super-scripts as necessary. In the exposition that follows, we omit the bias term $\boldsymbol{b}$ to avoid notational clutter.

### 2.2 EAGLE: an Enhanced Attention-based Strategy

Traditional attention methods have obtained fantastic achievement on general image. Yet, it is not the case when there are enormous amount of rich details in the image, e.g., remote sensing photographs. However, traditional attention may assign false weights and thereupon is likely to hinder the overall performance. It shows that more sophisticated attention methods should be explored to attack this issue.

Let us shift to a canyon. An eagle takes off from a cliff to forage. It hovers at high altitude. The panoramic view comes into its eyes, and it starts locating its potential preys. All of a sudden, the eagle specifies objectives and begins swooping down on it. Soon, preys are aware of danger and try to flee, but the eagle discerns such situation clearly and adjusts its subduction angle. Eventually the eagle seizes its preys and plays knife and fork. Inspired by such observation, we propose an enhanced attention-based strategy by generating answers from learning questions to a remote sensing image, entitled EAGLE (see Fig. 1 for its framework). We incorporate attentive question features $\widehat{\boldsymbol{q}}$ and attentive image features $\widehat{\boldsymbol{v}}$ to jointly reasons about visual attention and question attention in a series of steps in an alternative manner, in the sense that the image representations are applied to guide the question attention and the question representations are applied to guide image attention.

In order to focus on regions relevant to the question in the input image, we exploit attented question features $\widehat{\boldsymbol{q}}$ as guidance. To put it more specifically, the attentive image feature $\widehat{\boldsymbol{v}}_k$ is a weighted arithmetic mean of all convolutional features $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$. Dynamic weights in attention matrices indicate the significance of different regions in an image according to $\widehat{\boldsymbol{q}}_k$. Formally at time step $k$, we first compute attention matrics $\boldsymbol{H}_k^v$ associated with region features $\boldsymbol{V}$ corresponding to the feature vector $\widehat{\boldsymbol{q}}_k$:

$$\boldsymbol{H}_k^v = \tanh(\boldsymbol{W}_v \boldsymbol{V} + (\boldsymbol{W}_{\widehat{q}} \widehat{\boldsymbol{q}}_k) \mathbb{1}^{\mathrm{T}}) \tag{1}$$

where $\mathbb{1}$ is a vector with all elements to be "1". Then, we rescale each weight using a softmax:

$$\boldsymbol{a}_k^v = \mathrm{softmax}(\boldsymbol{w}_v^{\mathrm{T}} \boldsymbol{H}_k^v) \tag{2}$$

The image feature $\boldsymbol{v}_n$ at the n'th location is then weighted by the attention value $a_{k,n}^v$, and we obtain the attended feature vectors for images $\widehat{\boldsymbol{v}}_k$:

$$\widehat{\boldsymbol{v}}_k = \sum_{n=1}^{N} (a_{k,n}^v \boldsymbol{v}_n) \tag{3}$$

Attentive question features $\widehat{\boldsymbol{q}}$ are a natural symmetry against attentive image features $\widehat{\boldsymbol{v}}$. Unfortunately, there is a major difference between image processing and question encoding: image features are extracted in a parallel manner, yet questions are encoded word by word. In consequence, at a different time step, the number of updated hidden state of LSTM varies, and we fail to obtain a fixed-size matrix to represent information in the question. It will have a bad effect on training of parameters. To circumvent such issue, we define a fixed-size matrix $\boldsymbol{Q}$ by:

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{q}_1 \\ \boldsymbol{q}_2 \\ \cdots \\ \boldsymbol{q}_K \end{pmatrix}, \text{ where } \boldsymbol{q}_p = \begin{cases} \boldsymbol{h}_p, & \text{if } p \in [1, k] \text{ and } p \in \mathbb{Z} \\ \boldsymbol{0}, & \text{else} \end{cases} \tag{4}$$

It is worth noting that at time step $k$, we compute softmax only for the first $k$ question features. The rest of the computation is similar to that of image attention. Therefore we only enumerate the formulas without further discussion:

$$\begin{aligned} \boldsymbol{H}_k^q &= \tanh(\boldsymbol{W}_q \boldsymbol{Q} + (\boldsymbol{W}_{\widehat{v}} \widehat{\boldsymbol{v}}_{k-1}) \mathbb{1}^{\mathrm{T}}) \\ \boldsymbol{a}_k^q &= \underset{1,2,\cdots,k}{\mathrm{softmax}}(\boldsymbol{w}_q^{\mathrm{T}} \boldsymbol{H}_k^q) \\ \widehat{\boldsymbol{q}}_k &= \sum_{p=1}^{k} (a_{k,p}^q \boldsymbol{q}_p) \end{aligned} \tag{5}$$

We argue that the aforementioned enhanced attention-based strategy EAGLE provide complementary information and is effectively integrated in a unified framework. Therefore, EAGLE is beneficial for thorough understanding the aerial images, where there are numerous characteristics.

### 2.3 Overall Framework

EAGLE comprises three major components: (1) image feature extraction, (2) question encoding and (3) answer prediction. These are the standard building modules for most, if not all, existing visual question answering models. Algorithm 1 describes full details of the forward procedure to answer a given question according to a remote sensing image.

Faster R-CNN [11] holds state-of-the-art results in object detection. We apply it to extract convolutional features from input aerial images. Bottom-up attention within the region proposal network aims to focus on specific elements in the given image. The Faster R-CNN is pretrained and its parameters are fixed during training of our model. After the image is passed through Faster R-CNN, we obtain a matrix of its features $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$, where $N$ denotes the number of image regions and the vector $\boldsymbol{v}_n$ represents the feature extracted from the n'th region.

Then, the input question $\boldsymbol{R}$ with $K$ words is ready. Each word is turned into a dense vector representation according to a look-up dictionary. These vectors are initialized with pretrained word embeddings. The resulting embedding vectors $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_K\}$ are fed into our enhanced LSTM model one by one. Our LSTMs encompass a memory cell vector $\boldsymbol{c}$ that can store information for a long period of time, as well as three types of gates that control the flow of information into and out of these cells: input gates $\boldsymbol{i}$, forget gates $\boldsymbol{f}$ and output gates $\boldsymbol{o}$. Given word vector $\boldsymbol{r}_k$ at time step $k$, the previous hidden state $\boldsymbol{h}_{k-1}$, attentive question feature $\widehat{\boldsymbol{q}}_{k-1}$ and cell state $\boldsymbol{c}_{k-1}$, the update rules of our LSTM model can be defined as follows:

$$\begin{aligned}
\boldsymbol{i}_k &= \text{sigmoid}(\boldsymbol{W}_{ri}\boldsymbol{r}_k + \boldsymbol{W}_{hi}\boldsymbol{h}_{k-1} + \boldsymbol{W}_{\widehat{q}i}\widehat{\boldsymbol{q}}_{k-1}) \\
\boldsymbol{f}_k &= \text{sigmoid}(\boldsymbol{W}_{rf}\boldsymbol{r}_k + \boldsymbol{W}_{hf}\boldsymbol{h}_{k-1} + \boldsymbol{W}_{\widehat{q}f}\widehat{\boldsymbol{q}}_{k-1}) \\
\boldsymbol{o}_k &= \text{sigmoid}(\boldsymbol{W}_{ro}\boldsymbol{r}_k + \boldsymbol{W}_{ho}\boldsymbol{h}_{k-1} + \boldsymbol{W}_{\widehat{q}o}\widehat{\boldsymbol{q}}_{k-1}) \\
\boldsymbol{g}_k &= \tanh(\boldsymbol{W}_{rg}\boldsymbol{r}_k + \boldsymbol{W}_{hg}\boldsymbol{h}_{k-1} + \boldsymbol{W}_{\widehat{q}g}\widehat{\boldsymbol{q}}_{k-1}) \\
\boldsymbol{c}_k &= \boldsymbol{f}_k \odot \boldsymbol{c}_{k-1} + \boldsymbol{i}_k \odot \boldsymbol{g}_k \\
\boldsymbol{h}_k &= \boldsymbol{o}_k \odot \tanh(\boldsymbol{c}_k)
\end{aligned} \tag{6}$$

where $\boldsymbol{g}_k$ is the activation term at time step $k$, and $\odot$ is the element-wise multiplication of two vectors.

Next, we fuse hindmost attentive question feature $\widehat{\boldsymbol{q}}_K$ and attentive image feature $\widehat{\boldsymbol{v}}_K$ using:

$$\boldsymbol{h} = \tanh(\boldsymbol{W}_Q\widehat{\boldsymbol{q}}_K + \boldsymbol{W}_V\widehat{\boldsymbol{v}}_K) \tag{7}$$

It can be interpreted as projecting representation of the question and that of the image into a joint semantic space.

To generate an answer, we treat VQA output as a multi-lable classification task. Before we train the model, we predetermine a candidate answer set $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_M\}$, which is derived from the training answer set. The joint representation $\boldsymbol{h}$ is passed to compute a prediction score for each class, and we output the answer class with the highest prediction score among the candidate answer set $\boldsymbol{A}$. The exact formulation is as follows:

$$\boldsymbol{p}^A = \mathrm{softmax}(\boldsymbol{W}_h \boldsymbol{h}) \tag{8}$$

**Algorithm 1:** Forward Pass for One Question

```
Input: an aerial image and a natural language question about it;
Output: an answer;
Extract image features with Faster R-CNN;
$\widehat{\boldsymbol{q}}_0 = \mathbf{0}$;
Compute initial attentive image feature $\widehat{\boldsymbol{v}}_0$ according to eqs. (1) to (3);
for $k = 1$ to $K$ do
    Encode $\boldsymbol{r}_k$ with LSTMs according to eq. (6);
    Compute attentive question feature $\widehat{\boldsymbol{q}}_k$ according to eqs. (4) and (5);
    Compute attentive image feature $\widehat{\boldsymbol{v}}_k$ according to eqs. (1) to (3);
end
Fuse ultimate question and image features to predict an answer according to
  eqs. (7) and (8);
```

## 3 Remote Sensing Question Answering Corpus

To evaluate the performance of our proposed method, we construct a large-scale remote sensing question answering corpus. To the best of our knowledge, our corpus is the first one for remote sensing question answering. The current version of the corpus contains 6,921 images with 8,763 open-ended question-answer pairs. Most of the questions are about object, color, shape, number and location.

### 3.1 Creation Procedure

We invite experts in the remote sensing field to collect pictures with desirable quality from Google Earth. After that, we obtain 6,921 aerial images fixed to 224 × 224 pixels with various resolutions: 921 of them are typical urban shot, and the rest are divided into 30 object classes according to land use. The object classes include airport, bareland, baseball field, beach and so on. Each class has 200 images. We choose different land use of images to simulate complicated geographical scenes. We claim that the practice effectively reduces bias that remote sensing question answering models can exploit, and thereupon enhances our corpus's ability to benchmark VQA algorithms.

As for image annotations, we resort to crowdsourcing. Remote sensing images are complicated for ordinary people to describe. In consequence, we screen annotators only from volunteers with related knowledge of remote sensing field and annotation experience. The 921 pictures of typical urban shot are annotated with three questions per picture, and the rest 6,000 pictures one question for each. Each question is tagged with 10 answers. To sum up, we collect 8,763 QA pairs.

To get freestyle, interesting and diversified question-answer pairs, the annotators are free to raise any questions. The only limitation is that questions should be answered by the visual content and commonsense. Therefore, our corpus contains a wide range of AI related questions, such as object recognition (e.g., "What is there in green?"), positions and interactions among objects in the image (e.g. "Where is the centre?") and reasoning based on commonsense and visual content (e.g. "Why does the car park here?"). We give preference to the annotators who provide interesting questions that require high level reasoning to give the answer.

Yet, the freedom we give to the annotators makes it harder to control the quality of the annotation, compared to a more detailed instruction. To monitor the annotation quality, we conduct a pilot quality filtering task to select promising annotators before the formal annotation task. Specifically, we firstly randomly sample ten images from our image set as a quality monitoring corpus. Then we ask participants to view these images and write a QA pair for each one. After each annotator finishes labeling on the quality monitoring corpus, we examine the results and eliminate participants whose annotations are far from satisfactory (i.e. the questions are related to the content of the image and the answers are correct). Finally, we select a number of annotators with CV and NLP background (50 individuals) to participate in the formal annotation task.

We pick a set of good and bad examples of the annotated question-answer pairs from the pilot quality monitoring task, and show them to the selected annotators as references. We also provide reasons for selecting these examples. We post tasks in small batches over the course of two months to prevent participants from working long hours. Despite these measures, there are still annotations with general, uninformative QA pairs. After the annotation of all the images is completed, we further refine the corpus and remove a small portion of the images with badly labeled questions and answers.

### 3.2 Corpus Statistics

In our remote sensing corpus, the total number of question-answers pairs is 8,763. For each question, there are around 4.98 words and the sum of questions those are longer than 5 words is 3,801, making up 43.37% of the total. Meanwhile, the average length of each answer is about 2.00 words for 12.01% answers longer than 3 words.

We split the corpus into training set with 7,170 question-answers pairs and test set with 1,593. We can also see that the questions are mostly about object, color, shape, number and location from the statistical data. We provide an analysis on our remote sensing corpus in terms of both questions and answers.

**Statistics of Questions**

- **Types of Questions.** From our statistical data we can see that the questions are mostly about object, color, shape, number and location, which are in accordance with attributes of remote sensing images. Additionally, there exists a variety of question types including "What is", "What color", "How many", "Is there?", and questions like "What is around the playground?", "What color of the tree?", "What is on the bridge?" are very common in our remote sensing corpus.
- **Lengths.** For each questions, there are around 4.98 words and the sum of questions those are longer than 5 words is 3,801, making up 43.37% of the total questions.

**Statistics of Answers**

- **The Most Frequent Answers.** For lots of questions beginning with "What is", "What color", "How many", answers are mostly about object, color and number.
- **Lengths.** Most answers consist of a single word. The average length of each answer is about 2 words, and there are about 12.01% answers longer than 3 words. Though it may be tempt to believe that brief answers can make question-answering much easier, these question-answers pairs are open-ended and require complicated reasoning.

## 4 Experimental Evaluation

### 4.1 Models including Ablative Ones

We choose LSTM+CNN model without attention as Baseline. We refer to our Eagle Attention as EAGLE for conciseness. We also perform ablation studies to quantify the roles of each component in our model. The goal of the comparison is to verify that our improvements derive from synergistic effect of the two intertwining attention approaches. Specifically, we re-train our approach by ablating certain components:

- Question Attention alone (QuesAtt for short), where no image attention is performed.
- Image Attention alone (ImgAtt for short), where we do not apply any question attention.

### 4.2 Dataset

All experiments are conducted on our remote sensing corpus. We split the QA pairs into a training set with 7,170 QA pairs and a test set with 1,593 QA pairs. We ensure that all images are exclusive to either the training set or the test set.

### 4.3 Evaluation Metrics

We evaluate the predicted answers with accuracy:

$$\text{Accuracy} = \frac{\text{Num. of correctly classified questions}}{\text{Num. of questions}} \times 100\%$$

### 4.4 Implementation Details

For question encoding module, we select two layers of LSTM model with the hidden size of 512. Moreover the length of questions is fixed to 26 and each word embedding is a vector of size 300 and 2,400 dimension hidden states. For image feature extraction module, the pretrained VGG-16 model [14] is applied as initialization and only the fully-connected layers are fine-tuned. Then We take the activation from the last pooling layer of VGGNet as its feature. To prevent overfitting, dropouts are used after fully connected layers with dropout ratio 0.5.

### 4.5 Results and Analysis

We perform experimental evaluations to demonstrate the effectiveness of the proposed model. The experimental results in Table 1 show that EAGLE model performs the best, and the ImgAtt model and QuesAtt model both perform better than the standard LSTM-CNN VQA model. We argue that both image and question attention are beneficial for feature extraction, and the fusion of the two modal attentions by means of EAGLE is more suitable for remote sensing settings.

**Table 1.** Results on Remote Sensing Question Answering Corpus, in percentage.

| Models | Types of Questions | | | | | | All |
|---|---|---|---|---|---|---|---|
| | Color | How | Inference | Number | Object | Location | |
| Baseline | 63.9 | 18.8 | 82.0 | 62.3 | 38.3 | 8.8 | 25.5 |
| QuesAtt | 54.6 | 20.0 | 82.0 | 52.3 | 38.2 | 9.0 | 27.7 |
| ImgAtt | 64.7 | 19.4 | 82.8 | **62.7** | 38.8 | 9.3 | 27.1 |
| EAGLE | **66.2** | **20.1** | **83.7** | 50.9 | **51.0** | **9.9** | **31.6** |

Table 1 also presents models' performance in terms of question type. The "color" and "number" category corresponds to the question type "what color" and "how many", while the "object" category contains the questions types starting with "What is/type/kind/field/building/bridge...". The "inference" questions, of which corresponding answers are "Yes/No". The "How" questions customarily require general knowledge and commonsense of people's motivations, and "Where" questions also plenty of external knowledge, particularly knowledge of relevant location. We observe that VQA systems excel in answering questions requiring inference, where accuracies exceed 80.0%. However, they always generate inapt responses about commonsense and location. As we can see, EAGLE performs best among four models in all question types except "number" questions. Surprisingly EAGLE improves near 13 percentage points in questions about object. Indeed, the other models have observed counting ability in images with single object type but their ability to focus is weak since the remote sensing images contain amount of characteristics. For some questions like "What is next to the bare land?", EAGLE can locate the target object and recognize it. The superior performance of EALGE demonstrates the effectiveness of using hybrid attention mechanisms.

In order to validate EAGLE's achievement in real scenes, we report EAGLE's performance on large quantities of real data in Table 2. Although the average accuracy is 31.6%, there is a huge difference among all scenes. According to the result, EAGLE behaves best in the scene of forest. It achieves 59.5% in accuracy, which is comparable to the results in general images. We argue that such good performance is due to the fact that green is so prevalent in these images that only a little insightful information can be investigated. In light of the large percentage of "green" and "tree" in the corpus, VQA systems can perform well even if it only yields the two answers without reference to the picture. By contrast, EAGLE only achieves 3.0% in airport settings. We assume that it is likely for people to ask questions about numbers in the scene. Unfortunately, EAGLE does not explicitly incorporate modules to handle the task of counting and is inclined to answer these questions arbitrarily. It shows that there is still a long way to go for a general solution to visual questions in specific application scenarios. In the future, we will build some bigger remote sensing question answering corpus in order to train better models for semantic understanding of aerial images.

The main purpose of the proposed EAGLE method is to obtain an enhanced co-attention that performs more than naive combination of question attention

**Table 2.** EAGLE's accuracy per scene, in percentage.

| Category | airport | bareland | baseball field | beach | bridge | center |
|---|---|---|---|---|---|---|
| Accuracy | **3.0** | 37.0 | 57.0 | 25.5 | 9.0 | 33.5 |
| Category | church | commercial | dense residential | desert | farmland | forest |
| Accuracy | 5.5 | 51.5 | 50.0 | 31.5 | 40.0 | **59.5** |
| Category | industrial | meadow | medium residential | mountain | park | pond |
| Accuracy | 41.5 | 50.5 | 55.0 | 36.5 | 31.0 | 38.5 |
| Category | port | station | sparse residential | square | stadium | storage tank |
| Accuracy | 43.5 | 25.0 | 48.0 | 41.5 | 38.5 | 12.5 |

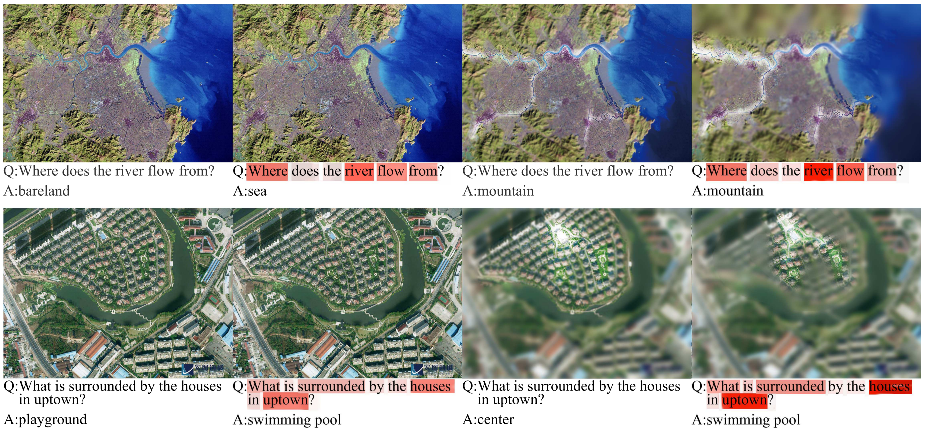
Q:Where does the river flow from?
A:bareland
Q:Where does the river flow from?
A:sea
Q:Where does the river flow from?
A:mountain
Q:Where does the river flow from?
A:mountain
Q:What is surrounded by the houses in uptown?
A:playground
Q:What is surrounded by the houses in uptown?
A:swimming pool
Q:What is surrounded by the houses in uptown?
A:center
Q:What is surrounded by the houses in uptown?
A:swimming pool

**Fig. 2.** Visualization of attention to question and/or image, if any. From left to right: Baseline, QuesAtt, ImgAtt and EAGLE. Specifically, with regard to rearmost attented question vector $\widehat{\boldsymbol{q}}_K$, questions attentions are scaled (from scarlet:high to transparent:low); as for terminal attended image vector $\widehat{\boldsymbol{v}}_K$, the lighter part denotes emphasized locations and by contrast other regions become blear.

and image attention. Hence we visualize the attention regions in both two modal attentions in Fig. 2. In attention visualization we overlay attention probability distribution matrices, denoting the most prominent part of a(n) image/question based on the question/image. Notice that in the first example, while the QuesAtt model identifies the target region river according to the query, it concerns many other irrelevant parts as well. By contrast, our EAGLE model is in single concentration on the river. It is also the case for question attention. As it is possible to see, EAGLE model is capable of more accurately paying attention to the area of interest in both questions and images, and the predicted answers yield better results. The comparison vividly illustrates that our improvements derive from synergistic effect of the two intertwining attention approaches.

## 5 Conclusion

We propose the task of mining details from remote sensing images. It necessitates thorough comprehension of language and visual data. While much research on general images obtains desired results, traditional approaches are not directly applicable in remote sensing setting. In this paper, we propose an enhanced attention-based approach which integrates seamlessly the property of remote sensing images and NLP. In addition, we present a large-scale remote sensing question answering corpus. To the best of our knowledge, it is first such corpus. Extensive experiments conducted on real data demonstrate that our proposal outperforms the-state-of-the-art approaches. Last but not least, we perform ablation studies so as to quantify the roles of different components in our model.

## Acknowledgements

We would like to thank the anonymous annotators for their professional annotations. This work is supported by the Natural Science Research of Jiangsu Higher Education Institutions of China under Grant 15KJA420001, National Natural Science Foundation of China under Grant 61772278, and Open Foundation for Key Laboratory of Information Processing and Intelligent Control in Fujian Province under Grant MJUKF201705.

## References

1. Chen, K., Crawford, M.M., Gamba, P., Smith, J.S.: Introduction for the special issue on remote sensing for major disaster prevention, monitoring, and assessment. IEEE Trans. Geoscience and Remote Sensing **45**(6-1), 1515–1518 (2007)
2. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
3. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? In: EMNLP (2016)
4. Karpathy, A., Li, F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
5. Li, Q., Fu, J., Yu, D., Mei, T., Luo, J.: Tell-and-answer: Towards explainable visual question answering using attributes and captions. In: EMNLP (2018)
6. Lillesand, T., Kiefer, R.W., Chipman, J.: Remote sensing and image interpretation (2014)
7. Lin, Y., Pang, Z., Wang, D., Zhuang, Y.: Feature enhancement in attention for visual question answering. In: IJCAI (2018)
8. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: NIPS (2016)
9. Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L.: Generating natural questions about an image. In: ACL (2016)
10. Rajani, N.F., Mooney, R.J.: Stacking with auxiliary features for visual question answering. In: NAACL-HLT (2018)
11. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015)
12. Shi, Z., Zou, Z.: Can a machine generate humanlike language descriptions for a remote sensing image? IEEE Trans. Geoscience and Remote Sensing **55**(6), 3623–3634 (2017)
13. Shimizu, N., Rong, N., Miyazaki, T.: Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In: COLING (2018)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
15. Xie, L., Shen, J., Zhu, L.: Online cross-modal hashing for web image retrieval. In: AAAI (2016)