Learning to Disambiguate by Asking Discriminative Questions

Yining Li[1] Chen Huang[2] Xiaoou Tang[1] Chen Change Loy[1]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Robotics Institute, Carnegie Mellon University
{ly015, xtang, ccloy}@ie.cuhk.edu.hk, chenh2@andrew.cmu.edu

Abstract

The ability to ask questions is a powerful tool to gather information in order to learn about the world and resolve ambiguities. In this paper, we explore a novel problem of generating discriminative questions to help disambiguate visual instances. Our work can be seen as a complement and new extension to the rich research studies on image captioning and question answering. We introduce the first large-scale dataset with over 10,000 carefully annotated images-question tuples to facilitate benchmarking. In particular, each tuple consists of a pair of images and 4.6 discriminative questions (as positive samples) and 5.9 non-discriminative questions (as negative samples) on average. In addition, we present an effective method for visual discriminative question generation. The method can be trained in a weakly supervised manner without discriminative images-question tuples but just existing visual question answering datasets. Promising results are shown against representative baselines through quantitative evaluations and user studies.

1. Introduction

Imagine a natural language dialog between a computer and a human (see Fig. 1):

Kid : *"What sport is the man playing?"*.
Computer : *"What is the color of his shirt?"*
Kid : *"Blue."*
Computer : *"He is playing tennis."*

In the conversation, the kid refers to an object but the description is linguistically underspecified. The computer analyzes the visual differences of plausible targets and then reacts by asking a discriminative question *"What is the color of his shirt?"* to resolve the reference.

We define the aforementioned problem as *Visual Discriminative Question Generation* (VDQG). Specifically, the computer is given with two visual instances, and the goal is to ask a good question to distinguish or disambiguate them. In this study, we call the pair images as ambiguous pairs

Good	*What color is the man's shirt?*
Good	*What sport is the man playing?*
Bad	*How many people are there?*
Bad	*Where is the man?*

Figure 1: Example ambiguous image pair and both good and bad discriminative questions.

– the ambiguity may not necessarily be due to their subtle visual differences. They may just belong to the same object class with close proximity in their deep representation. Although such ambiguity can be easily resolved by human, they can be difficult to a machine. Distinguishing different image pairs require asking different types of questions, ranging from color, action, location, and number. Akin to the classic "Twenty Questions" game, a careful selection of questions can greatly improve the odds of the questioner to narrow down the answer. A bad question would fail to eliminate ambiguities. Figure 1 gives good and bad examples of questions. This questioning capability can subsequently be extended to generating a sequence of discriminative questions and prompting a human-in-the-loop to answer them. In the process, the machine accumulates evidence that can gradually refine the language expression from humans and finally distinguish the object of interest.

Such VDQG ability allows a machine to play a more natural and interactive role in Human-Computer Interaction (HCI), or improve a robot to bind the references made by a speaker more accurately to objects in a scene. While there have been various attempts to build a system that can provide explanations [14] or ask questions [30, 33] based on visual instances, the problem of VDQG has not been ex-

plored. The goal of VDQG is to resolve inter-object ambiguities through asking questions. It is thus differs from image captioning that aims at generating a literal description based on a single visual instance. It also differs from Visual Question Answering (VQA), which takes an image and a question as inputs and provides an answer. A closer work is Visual Question Generation (VQG) [30, 33]. Unlike the setting of generating one possible question from an image, VDQG operates on two visual instances and generates a discriminating question for them. The most relevant work to ours is Yu *et al*. [50], which generates unambiguous referring expressions for an object by incorporating visual comparison to other objects in an image. Our problem differs in that we generate one single question to distinguish multiple objects instead of referring expressions for all objects.

It is non-trivial to train a machine to ask discriminative questions in an automatic and human understandable way. Firstly, it should ask a natural and object-focused question. Secondly, and importantly, the machine is required to pinpoint the most distinguishing characteristics of two objects to perform a comparison. Addressing the problem is further compounded by the lack of data. In particular, there are no existing datasets that come readily with pair images annotated with discriminative questions. Thus we cannot perform a direct supervised learning.

To overcome the challenges, we utilize the Long Short-Term Memory (LSTM) [13] network to generate natural language questions. To generate discriminative questions, which are object-focus, we condition the LSTM with a visual deep convolutional network that predicts fine-grained attributes. Here visual attributes provide a tight constraint on the large space of possible questions that can be generated from the LSTM. We propose a new method to identify the most discriminative attributes from noisy attribute detections on the two considered objects. Then we feed the chosen attributes into the LSTM network, which is trained end-to-end to generate an unambiguous question. To address the training data problem, we introduce a novel approach to training the LSTM in a weakly-supervised manner with rich visual questioning information extracted from the Visual Genome dataset [23]. In addition, a large-scale VDQG dataset is proposed for evaluation purposes.

Contributions: We present the first attempt to address the novel problem of Visual Discriminative Question Generation (VDQG). To facilitate future benchmarking, we extend the current Visual Genome dataset [23] by establishing a large-scale VDQG dataset of over $10,000$ image pairs with over $100,000$ discriminative and non-discriminative questions. We further demonstrate an effective LSTM-based method for discriminative question generation. Unlike existing image captioning and VQG methods, the proposed LSTM is conditioned on discriminative attributes selected through a discriminative score function. We conduct both quantitative and user studies to validate the effectiveness of our approach.

2. Related Work

Image Captioning. The goal of image captioning is to automatically generate natural language description of images [9]. The CNN-LSTM framework has been commonly adopted and shows good performance [7, 19, 29, 43, 45]. Xu *et al*. [47] introduce attention mechanism to exploit spatial information from image context. Krishna *et al*. [23] incorporate object detection [38] to generate descriptions for dense regions. Jia *et al*. [16] extracts semantic information from images as extra guide to caption generation. Krause *et al*. [22] uses hierarchical RNN to generates entire paragraphs to describe images, which is more descriptive than single sentence caption. In contrast to these studies, we are interested in generating a question rather than a caption to distinguish two objects in images.

Visual Question Answering (VQA). VQA aims at generating answer given an input image and question. It differs from our task of generating questions to disambiguate images. Deep encoder-decoder framework [27] has been adopted to learn a joint representation of input visual and textual information for answer prediction (multiple-choice) or generation (open-ended). Visual attention [26, 41, 46, 48] and question conditioned model [2, 35] have been explored to capture most answer-related information from images and questions. To facilitate VQA research, a number of benchmarks has been introduced [3, 23, 32, 37, 49, 53]. Johnson *et al*. [17] introduce a diagnostic VQA dataset by mitigating the answer biases which can be exploit to achieve inflated performance. Das *et al*. [5] extend VQA to a dialog scenario. Zhang *et al*. [52] build a balanced binary VQA dataset on abstract scenes by collect counterpart images that yield opposite answers to the same question. A concurrent work to ours is [12], which extends the popular VQA dataset [3] by collecting complementary images such that each question will be associated to a pair of similar images that result in to different answers. Both [52] and [12] contribute a balanced VQA dataset do not explore the VDQG problem. Although our model can be trained on balanced VQA data, we show that it performs reasonably well by just learning from unbalanced VQA datasets.

Referring Expression Generation (REG). A closely related task to VDQG is REG, where the model is required to generate unambiguous object descriptions. Referring expression has been studied in Natural Language Processing (NLP) [11, 21, 44]. Kazemzadeh *et al*. [20] introduce the first large-scale dataset for the REG in real-world scenes. They use images from the ImageCLEF dataset [8], and collect referring expression annotations by developing a ReferIt game. The authors of [28, 50] build two larger

REG datasets by using similar approaches on top of MS COCO [25]. CNN-LSTM model has been shown effective in both generation [28, 50] and comprehension [15, 34] of REG. Mao *et al*. [28] introduce a discriminative loss function based on Maximum Mutual Information. Yu *et al*. [50] study the usage of context in REG task. Yu *et al*. [51] propose a speaker-listener-reinforcer framework for REG, which is end-to-end trainable by reinforcement learning.

Visual Question Generation (VQG). Natural-language question generation from text corpus has been studied for years [1, 4, 18, 40]. The task of generating question about images, however, has not been extensively studied. A key problem is the uncertainty of the questions' query targets, which makes the question generation subjective and hard to evaluate. Masuda-Mora *et al*. [30] design a question-answer pair generation framework, where a CNN-LSTM model is used to generate image-related questions, and a following LSTM will decode the hidden representation of the question into its answer. Their model is trained using VQA annotations [3]. Mostafazadeh *et al*. [33] introduce the first VQG dataset. Mostafazadeh *et al*. [32] further extend the scenario to image-grounded conversation generation, where the model is repurposed for generating a sequence of questions and responses given image contexts. These tasks are essentially same as image captioning, because the goal is to model the joint distribution of image and language (questions), without explicitly considering the query target of the generated question. A concurrent work [6] proposes to use yes-no question sequences to locate unknown objects in images, and introduces a large-scale dataset. This work strengthens our belief on the importance of visual disambiguation by natural-language questions. The differences between this work and ours are: 1) We do not restrict a question to be yes-no type but more open-ended. 2) We explore the usage of semantic attributes in discriminative question generation. 3) No training data is available for training our VDQG. We circumvent this issue through a weakly-supervised learning method, which learns discriminative question generation from general VQA datasets.

3. VDQG Dataset for Evaluation

Existing VQG and VQA datasets [3, 17, 23, 33, 37, 53] only contain questions annotated on single image[1], which is inadequate for quantitative evaluation and analysis of VDQG methods. To fill the gap, we build a large-scale dataset that contains image pairs with human-annotated questions. We gather images from the Visual Genome dataset [23] and select image pairs as those that possess the same category label and high CNN feature similarity. Finally we employ crowd-sourcing to annotate discriminative and non-discriminative questions on these pairs. Some of the example pairs and the associated questions are shown in Fig. 2. As can be observed, many of these pairs are ambiguous not only because they are of the same object class, but also due to their similar visual appearances. We detail the data collection process as follows.

[1]Apart from the concurrent work [12], which released a large-scale balanced VQA dataset. Unfortunately the dataset was released in late March so we were not able to train/test our model on this data.

Table 1: Statistics of VDQG dataset. The length of a question is given by the number of tokens.

No. of images	8,058
No. of objects	13,987
No. of ambiguous image pairs	11,202
No. of questions	117,745
Avg. pos-question number per object pair	4.57
Avg. neg-question number per object pair	5.94
Avg. token number per question	5.44

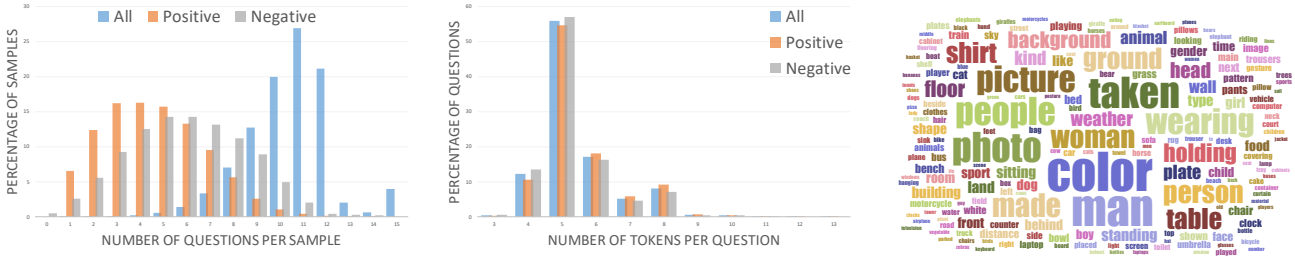
Ambiguous Pair Collection. The Visual Genome dataset provides object annotations with their category labels and bounding boxes. We select 87 object categories that contain rich and diverse instances. Incorrect labeled and low-quality samples are discarded. Subsequently, we cluster image instances in each object category by their features extracted with Inception-ResNet [42]. Image pairs are randomly sampled from a cluster to form the ambiguous pairs.

Question Annotation. Question annotation is a laborious process. We therefore adopt a two-step approach to collect annotations by crowd-sourcing, and augment with more questions automatically followed by human verification. In the first step, the workers are prompted to ask questions that can tell the differences between two images in an ambiguous pair. In this way we collect 2 to 3 discriminative questions for pair. It is worth pointing out that we collect '7W' questions, consistent with protocol adopted by the Visual Genome dataset [23]. This is the major difference between our dataset and [6], which only contains 'yes-no' questions.

Then we augment the question set of each ambiguous pair by 1) retrieving questions from other visually similar ambiguous pair and 2) automatically generating questions using a CNN-LSTM model trained on Visual Genome VQA annotations. After augmentation each ambiguous pair has over 8 question annotations. The added questions are expected to be related to the given images, but not guaranteed to be discriminative. Thus in the second step, the workers are shown with an ambiguous pair and a question, and they will judge whether the question would provide two different answers respectively to the images pair. Specifically, the worker will rate the question in a range of strong-positive, weak-positive and negative, which will serve as the label of the question.

Statistics. Our dataset contains 13,987 images covering 87 object categories. We annotated 11,202 ambiguous image pairs with 117,745 discriminative and non-discriminative questions. Table 1 summarizes key statistics of our dataset.

Figure 2: Example of ambiguous pairs and the associated positive and negative question annotations in the proposed VDQG dataset. Positive and negative questions are written in blue and red, respectively. More examples in supplementary material.



(a) Distribution of number of questions (b) Distribution of question length (c) Word cloud of questions

Figure 3: Statistics of the VDQG dataset.

We provide an illustration in Fig. 3 to show more statistics of the proposed dataset. Further statistics and examples of this dataset can be found in the supplementary material.

4. Visual Discriminative Question Generation

Our goal is to generate discriminative questions collaboratively from two image regions R^A and R^B. We show the proposed VDQG approach in Fig. 4. The approach can be divided into two steps. The first step is to find discriminative attribute pairs. An attribute recognition and attribute selection components will be developed to achieve this goal. In particular, each region will be described by an attribute, and collectively, they should form a pair that best distinguish the two regions. For instance, as shown in Fig. 4, the 'blue-white' attributes constitute a pair that is deemed more discriminative than the 'tennis-baseball' pair, since the baseball bat is hardly visible. Given the discriminative attributes, the second step is to use the attributes to condition an LSTM to generate discriminative question.

Inspired by [15, 28], the image region is represented by a concatenation of its local feature, image context and relative location/size: $\mathbf{f} = [f_{cnn}(R), f_{cnn}(I), \mathbf{l}_r]$. Specifically, $f_{cnn}(R)$ and $f_{cnn}(I)$ represent the 2048-d region and image features, respectively. The features are extracted using a Inception-ResNet [42] pre-trained on ImageNet [39]. The vector $\mathbf{l}_r = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{S_r}{S_I}]$ denotes the relative location and size of the region.

4.1. Finding Discriminative Attribute Pairs

To find a pair of discriminative attributes, our method first recognizes visual attributes from each region to form an paired attribute pool. The method then applies attribute selection to select a pair of attributes that best distinguish the two regions.

Attribute Recognition: Attributes offer important mid-level cues of objects, usually in the form of a single word [9, 45]. Since we only use attributes for discerning the two images, we extend the notion of 'single-word attribute' to a short phrase to enhance its discriminative power. For example, the attribute of "next to building" is actually frequent in everyday conversation and can be more expressive and discriminative than those single "location" attributes. To this end, we extract the commonly used n-gram expressions ($n \leq 3$) from region descriptions in Visual Genome dataset. We add the part-of-speech constraint to select for descriptive expressions. An additional constraint is added so that the expressions should intersect with the top 1000 most frequent answers in the dataset. This helps filtering expressions that are less frequent or too specific. Examples of expressions chosen to serve as our attributes include "man", "stand", "in white shirt", "on wooden table", "next to tree". More examples can be found in the supplementary material. The top $K = 612$ constrained expressions are collected to form our attribute list $\{att_k\}$.

Next, we can associate each image region with its ground-truth attributes and train a visual attribute recognition model. We cast the learning as a multi-label classifica-

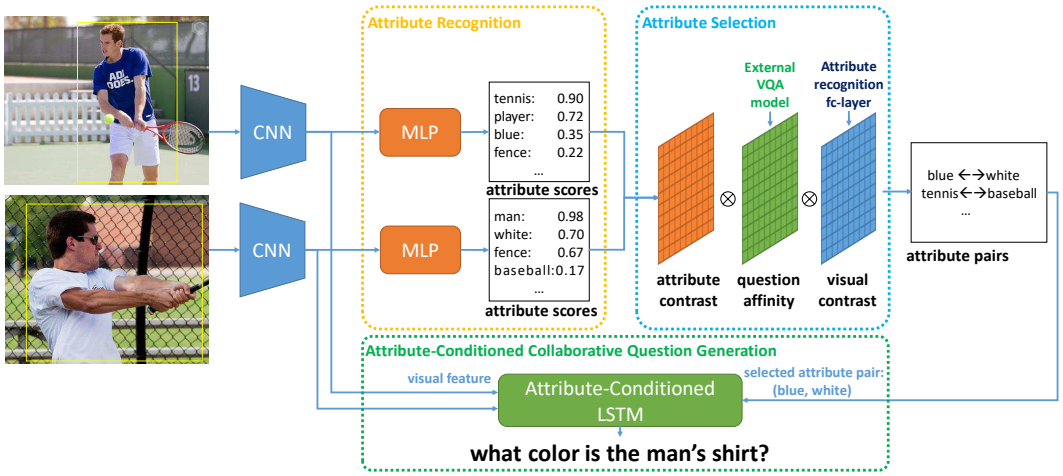


Figure 4: Overview of the attribute-conditioned question generation process. Given a pair of ambiguous images, we first extract semantic attributes from the images respectively. The attribute scores are sent into a selection model to select the distinguishing attributes pair, which reflects the most obvious difference between the ambiguous images. Then the visual feature and selected attribute pair are fed into an attribute-conditioned LSTM model to generate discriminative questions.

tion problem. Specifically, we feed the visual representation $\mathbf{f}$ of each region into Multi-layer Perceptions (MLP) with a sigmoid layer to predict a K-d attribute score vector, $\mathbf{v}$. The MLP parameters are trained under a cross-entropy loss.

Attribute Selection: Given the attribute score vectors $\mathbf{v}^A$, $\mathbf{v}^B \in \mathbb{R}^K$ extracted from two image regions R^A and R^B, we want to choose an attribute pair (att_i, att_j) that best distinguishes them. The chosen attributes should possess the following three desired properties:

1) Each attribute in the chosen pair should have highly contrasting responses on two regions. For examples, two regions with "red" and "green" attributes respectively would fulfill this requirement.

2) The chosen pair of attributes should be able to serve as a plausible answer for a single identical question. For instance, the "red" and "green" attributes both provide plausible answers to the question of "What color is it?".

3) The chosen pair of attributes should be easily distinguished by visual observations. We define the visual dissimilarity as an intrinsic property of attributes independent to particular images.

We integrate these constraints into the following score function. Here we use a shorthand (i, j) to represent (att_i, att_j).

$$s(i,j) = \underbrace{v_i^A(1-v_i^B)\cdot v_j^B(1-v_j^A)}_{attribute\ score\ contrast} \cdot \underbrace{e^{\alpha s_q(i,j)}}_{question\ similarity} \cdot \underbrace{e^{-\beta s_f(i,j)}}_{visual\ dissimilarity}, \quad (1)$$

where α, β are the balancing weights among the three constraints, and $s_q(\cdot,\cdot)$ and $s_f(\cdot,\cdot)$ encode the question and feature similarities, respectively. We use the full score in Eq. (1) to rank all K^2 attribute pairs in an efficient way, and select the top scoring pair to guide our VDQG. Next we explain each term in the score function:

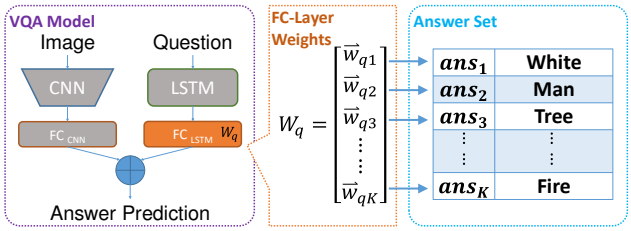


Figure 5: Question similarity scoring. We train a VQA model (left) using question-answer annotations of Visual Genome [23]. Since the answers overlap with our defined attributes, question similarity between two attributes att_i and att_j can be computed as the inner product of the corresponding i-th and j-th row vectors in the weight matrix of the $\mathrm{FC}_{\mathrm{LSTM}}$ layer.

Attribute score contrast. This term computes the score contrast of attributes between two image regions, where $v_i^A \in \mathbf{v}^A$ represents the score/response of i-th attribute on region R^A. Similar notational interpretation applies to other variables in this term. The score contrast of a discriminative attribute pair should be high.

Question similarity $s_q(i,j)$. The question similarity score of a discriminative attribute pair should be large because they are intended to respond to the same identical question. Finding this similarity is non-trivial. To compute the question similarity $s_q(i,j)$ of attributes att_i and att_j, we train a small VQA model (see Fig. 5) that is capable of providing an answer given an input question and image. The model is trained using question-answer annotations from the Visual

Genome dataset [23]. Note that we only train the model using question-answer annotations of which the answer is one of the attributes in $\{att_k\}$ that we define earlier (recall that our attribute set overlaps with the answer set). Thus the answer output of the VQA model is actually our attribute set and the model captures the question-attribute relations.

As illustrated in Fig. 5, the fully-connected layer after LSTM (FC_{LSTM}) contains a weight matrix W_q, of which the i-th row vector, denoted as $\overrightarrow{w}_{qi}$, is trained for prediction of attribute att_i. In other words, this vector $\overrightarrow{w}_{qi}$ could serves as the representation of attribute att_i in the question space. Hence, the question similarity between att_i and att_j can be computed as the inner product of $\overrightarrow{w}_{qi}$ and $\overrightarrow{w}_{qj}$, denoted as $\langle \overrightarrow{w}_{qi}, \overrightarrow{w}_{qj} \rangle$.

Visual similarity $s_f(i,j)$. The visual similarity score of a discriminative attribute pair should be small. To determine the visual similarity $s_f(i,j)$ between attribute att_i and att_j, we use the technique which we compute the question similarity. Specifically, the fully-connected layer of our attribute recognition model contains a weight matrix W_f, of which the i-th row vector, denoted as $\overrightarrow{w}_{fi}$, is trained for prediction of attribute att_i. Consequently, the visual similarity between att_i and att_j can be computed as the inner product of $\overrightarrow{w}_{fi}$ and $\overrightarrow{w}_{fj}$, denoted as $\langle \overrightarrow{w}_{fi}, \overrightarrow{w}_{fj} \rangle$.

4.2. CNN-LSTM with Attribute Conditions

In this section, we describe the formulation of the attribute-conditioned LSTM. We start with a brief review of conventional CNN-LSTM.

Conventional CNN-LSTM. In the typical CNN-LSTM language generation framework, CNN features $\mathbf{f}$ are first extracted from an input image. The features are then fed into the LSTM to generate language sequences. The model is trained by minimizing the negative log likelihood:

$$\begin{aligned} L &= \sum\nolimits_n -\log\ p(Q_n|\mathbf{f}_n) \\ &= \sum\nolimits_n \sum\nolimits_t -\log\ p(q_t^n|q_{t-1,\ldots,1}^n, \mathbf{f}_n), \end{aligned} \tag{2}$$

where each question Q_n comprises of a word sequence $\{q_t^n\}$.

Attribute-Conditioned LSTM. To generate questions with specific intent, we utilize semantic attributes as an auxiliary input of the LSTM to condition the generation process. Ideally, when the model takes a "red" attribute, it would generate question like "What is the color?". We train such conditioned LSTM using the tuple $(\mathbf{f}, Q, att_i)$, where att_i is made out of the groundtruth answer of Q. Similar to Eq. (2), we minimize the negative log likelihood as follows:

$$L = \sum\nolimits_n -log\ p(Q_n|\mathbf{f}_n, \sigma(att_i^n)), \tag{3}$$

where $\sigma(\cdot)$ is a feature embedding function for attribute input. We use Word2Vec [31] as the embedding function that can generalize across natural language answers and attributes.

Our goal is to generate one discriminative question collaboratively from two image regions R^A and R^B with the selected attribute pair (att_i, att_j). Thus we duplicate the attribute-conditioned LSTM for each region and compute a joint question probability $p(Q|\mathbf{f}^A, \mathbf{f}^B, \sigma(att_i), \sigma(att_j))$, which can be expressed as

$$\begin{aligned} &p(q_t|q_{t-1,\ldots,1}, \mathbf{f}^A, \mathbf{f}^B, \sigma(att_i), \sigma(att_j)) = \\ &\frac{p(q_t|q_{t-1,\ldots,1}, \mathbf{f}_A, \sigma(att_i)) \cdot p(q_t|q_{t-1,\ldots,1}, \mathbf{f}_B, \sigma(att_j))}{\sum_{q\in\mathcal{V}} p(q|q_{t-1,\ldots,1}, \mathbf{f}_A, \sigma(att_i)) \cdot p(q|q_{t-1,\ldots,1}, \mathbf{f}_B, \sigma(att_j))}, \end{aligned} \tag{4}$$

where $\mathcal{V}$ is the whole vocabulary. We use beam search to find the most probable questions according to Eq. (4).

Learning from Weak Supervision. As mentioned before, there are no public available paired-image datasets annotated with discriminative questions for fully-supervised learning. Fortunately, due to the unique formulation of our approach, which extends CNN-LSTM to generate questions collaboratively from two image regions (see Eq. 4), our method can be trained by just using 'single image + single question' dataset. We choose to utilize the rich information from Visual Genome dataset [23]. In particular, we extract 1445k image-related question-answer pairs and their grounding information, *i.e.*, region bounding box. We randomly split the question-answer pairs into training (70%), validation (15%) and testing (15%) sets, where questions referring to the same image will only appear in the same set. We also utilize the associated region descriptions to enrich the textual information for our attribute-conditioned model (Sec. 4.1). It is worth noting that the training and validation sets are only used for our model training in a weakly-supervised manner, while the testing set is used to construct the VDQG dataset as introduced in Sec. 3.

5. Experiments

Methods. We perform experiments on the proposed VDQG datasets and evaluate the following methods:

1) *Our Approach* (ACQG). We call our approach as Attribute-Conditioned Question Generation (ACQG). We establish a few variants based on the way discriminative attributes are selected. ACQG_{ac} only uses the attribute score contrast in Eq. (1). ACQG_{ac+qs} uses both attribute score contrast and question similarity. Lastly, ACQG_{full} uses all the terms for attribute selection. For each sample,we select top-5 attribute pairs and generate questions for each pair. The final output is the question with the highest score, which is the product of its attribute score (Eq. 1) and question probability (Eq. 4). This achieves a better performance than only using top-1 attribute pair.

2) *CNN-LSTM*. We modify the state-of-the-art image captioning CNN-LSTM model [7] for the VDQG task. Specif-

ically, we adopt Inception-ResNet [42] as the CNN part, followed by two stacked 512-d LSTMs. We also extend the framework to accommodate image pair input following Eq. (4) without using pair attributes as the condition.

3) *Retrieval-based Approach* (Retrieval). It is shown in [33] that carefully designed retrieval approaches can be competitive with generative approaches for their VQG task. Inspired by [33], we prepare a retrieval-based baseline for the VDQG task. Our training set consists of questions annotated on image regions. Given a test image pair, we first search for the k nearest neighbor ($k = 100$) training image regions for the pair, and use the training questions annotated on these retrieved regions to build a candidate pool. For each question in the candidate pool, we compute its similarity to the other questions using BLEU [36] score. The candidate question with the highest score will be associated with the input image pair.

Evaluation Metrics. To evaluate a generated question, we hope to reward a match with the positive ground-truth questions, and punish a match with the negative ground-truth questions. To this end, we use ΔBLEU [10] as our main evaluation metric, which is tailored for text generation tasks that admit a diverse range of possible outputs. Mostafazadeh *et al*. [33] show that ΔBLEU has a strong correlation with human judgments in visual question generation task. In particular, given a reference (annotated question) set $\{r_{i,j}\}$ and the hypothesis (generated question) set$\{h_i\}$, where i is the sample index and j is the annotated question index of i-th sample, ΔBLEU score is computed as:

$$\Delta\text{BLEU} = \text{BP} \cdot \exp(\sum\nolimits_n \log p_n) \tag{5}$$

The corpus-level n-gram precision is defined as:

$$p_n = \frac{\sum_i \sum_{g \in \text{n-grams}(h_i)} \max_{j:g \in r_{i,j}} \{w_{i,j} \cdot \#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in \text{n-grams}(h_i)} \max_j \{w_{i,j} \cdot \#_g(h_i)\}}, \tag{6}$$

where $\#_g(\cdot)$ is the number of occurrences of n-gram g in a given question, and $\#_g(u, v)$ is the shorthand for $\min\{\#_g(u), \#_g(v)\}$. And the brevity penalty coefficient BP is defined as:

$$\text{BP} = \begin{cases} 1 & \text{if } \rho > \eta \\ e^{1-\eta/\rho} & \text{if } \rho \leq \eta \end{cases}, \tag{7}$$

where ρ and η are respectively the length of generated question and effective annotation length. We respectively set the score coefficients of strong-positive samples, weak-positive samples and negative samples to be 1.0, 0.5 and -0.5. We use a equal weights for up to 4-grams.

As a supplement, we also use BLEU [36] and METEOR [24] to evaluate the textual similarity between generated questions and positive annotations in the test set.

Table 2: Experiment results on full VDQG dataset.

Model	ΔBLEU	BLEU	METEOR
Human$_{top}$	69.2	85.5	57.5
Human$_{random}$	62.9	82.4	54.9
Retrieval	24.3	42.5	29.1
CNN-LSTM	33.4	56.2	37.3
ACQG$_{ac}$	29.4	52.9	35.3
ACQG$_{ac+qs}$	40.1	59.1	39.6
ACQG$_{full}$	**40.6**	**59.4**	**39.7**

Table 3: Experiment results on VDQG hard subset.

Model	ΔBLEU	BLEU	METEOR
Human$_{top}$	62.3	79.2	52.2
Human$_{random}$	53.7	74.9	48.9
Retrieval	13.4	36.9	25.9
CNN-LSTM	20.3	47.8	32.7
ACQG$_{ac}$	13.5	44.3	30.4
ACQG$_{ac+qs}$	32.6	53.2	36.1
ACQG$_{full}$	**33.5**	**53.6**	**36.4**

5.1. Results

We conducted two experiments based on the VDQG dataset. The first experiment was conducted on the full samples. The second experiment was performed by using only a hard subset of VDQG. We constructed the hard subset by selecting 50% samples with a lower ratio of positive annotations within each object category.

Table 2 summarizes the results on the full VDQG dataset. The proposed method outperforms baseline methods according to all metrics. We also performed ablation study by gradually dropping the similarity terms in Eq. (1) out of our full model. The results suggest that question similarity dominates the performance improvement while other terms also play an essential role. It is noted that ACQG$_{ac}$ yields poor results in comparison to the baseline CNN-LSTM. Based on our conjecture, the attribute score contrast term may be too simple therefore overwhelmed by the noisy prediction scores of attributes. Experimental results on the hard subset are shown in Table 3. Compared with the results in Table 2, the performance gap between ACQG$_{full}$ and non-attribute-guided models increases in hard cases, which shows the significance of discriminative attributes in the task of VDQG.

We also performed an interesting experiment based on the collected question annotations in VDQG dataset. Specifically, 'Human$_{top}$' indicates the first-annotated positive question of each sample, while 'Human$_{random}$' indicates a random positive annotation among all the human annotations of each sample. It is reasonable to assume that the first-written questions are likely to ask the most distinguishing differences between two images. From both Tables 2 and 3, we observe that 'Human$_{top}$' consistently outperforms 'Human$_{random}$. The results suggest the effective-

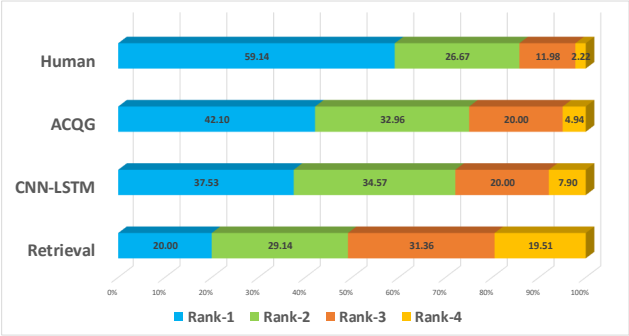


Figure 6: User study on VDQG full.

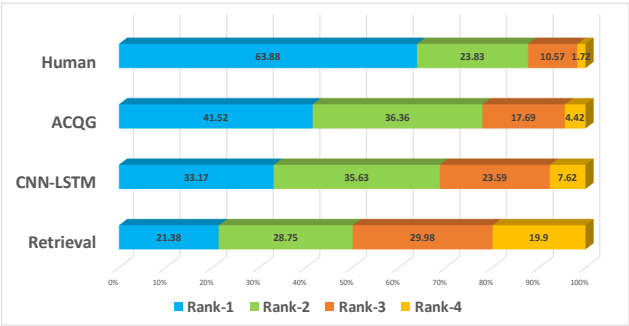


Figure 7: User study on VDQG hard subset.

ness of the proposed VDQG dataset and metric settings for VDQG evaluation.

5.2. User Study

We gathered a total of 27 participants to join our user study. Each time we showed the participant an image pair and four questions generated respectively by a human annotator (the groundtruth), the proposed ACQG$_{full}$, CNN-LSTM, and Retrieval. Then the participant was asked to rank these questions according to their capability of distinguishing the given image pair. Figure 6 shows the results of user study. We also separately analyze the hard samples, and show the results in Fig. 7. The proposed ACQG$_{full}$ outperforms other baseline models in the user study. It is observed that the performance gap becomes more significant on hard samples.

6. Comparison with Referring Expression

A referring expression is a kind of unambiguous description that refers to a particular object within an image. Despite the linguistic form differences between the discriminative question and the referring expression, they have the common objective of disambiguation. In this section, we compared discriminative question with referring expression by conducting a user study with 14 participants. Specifically, each time we showed the participant an image with two ambiguous objects marked with their respective bounding boxes. Meanwhile, we showed the participant a referring expression[2] or a discriminative question with its conditioning attribute that refers to one of the objects. Then the participant was asked to retrieve the referred object by the given information. We compute the mean retrieval accuracy to measure the disambiguation capability of the given textual information.

Figure 8: Visualization of the Discriminative Question (DQ) and Referring Expression (RE) generated from the ambiguous objects in images. Referred objects and distractors are marked with green and red bounding boxes respectively. The second row shows some failure cases.

The results are interesting – showing referring expressions results in a mean retrieval accuracy of 65.14%, while showing discriminative question+attribute achieves a competitive result of 69.51%. In Fig. 8 we show some of the generated referring expressions and discriminative questions on ambiguous objects within images. It is interesting to notice that referring expressions and discriminative questions fail in different cases, which indicates that they could be further studied as complementary approaches to visual disambiguation.

7. Conclusion

We have presented a novel problem of generating discriminative questions to help disambiguate visual instances. We built a large-scale dataset to facilitate the evaluation of this task. Besides, we proposed a question generation model that is conditioned on discriminative attributes. The method can be trained by using weak supervisions extracted from existing VQA dataset (single image + single question), without using full supervision that consists of paired-image samples annotated with discriminative questions.

Acknowledgement: This work is supported by SenseTime Group Limited and the General Research Fund sponsored by the Research Grants Council of the Hong Kong SAR (CUHK 416713, 14241716, 14224316. 14209217).

[2]We generate referring expressions using the state-of-the-art REG model [28] trained on RefCOCO+ dataset [50]. The images used in the user study are selected from the validation set of RefCOCO+. In particular, we select the images containing two ambiguous objects.

References

[1] H. Ali, Y. Chali, and S. A. Hasan. Automation of question generation from sentences. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 58–67, 2010. 3

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *CVPR*, pages 39–48, 2016. 2

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 2, 3

[4] W. Chen. Aist, g., mostow, j.: Generating questions automatically from informational text. In *Proceedings of the Second Workshop on Question Generation, held at the Conference on AI in, Education*, pages 17–24, 2009. 3

[5] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual Dialog. In *CVPR*, 2017. 2

[6] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017. 3

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015. 2, 6

[8] H. J. Escalante, C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger. The segmented and annotated iapr tc-12 benchmark. *CVPR*, 114(4):419–428, 2010. 2

[9] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015. 2, 4

[10] M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, and B. Dolan. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv:1506.06863*, 2015. 7

[11] D. Golland, P. Liang, and D. Klein. A game-theoretic approach to generating spatial descriptions. In *EMNLP*, pages 410–419, 2010. 2

[12] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017. 2, 3

[13] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. LSTM: A search space odyssey. *arXiv preprint*, arXiv:1503.04069, 2015. 2

[14] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *ECCV*, pages 3–19, 2016. 1

[15] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016. 3, 4

[16] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars. Guiding the long-short term memory model for image caption generation. In *ICCV*, pages 2407–2415, 2015. 2

[17] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 2, 3

[18] S. Kalady, A. Elikkottil, and R. Das. Natural language question generation using syntax and keywords. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 1–10, 2010. 3

[19] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 2

[20] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 2

[21] E. Krahmer and K. Van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012. 2

[22] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, 2017. 2

[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 2, 3, 5, 6

[24] M. D. A. Lavie. Meteor universal: Language specific translation evaluation for any target language. *ACL*, page 376, 2014. 7

[25] T.-Y. Lin, M. Maire, S. 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. 3

[26] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297, 2016. 2

[27] 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. 2

[28] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 2, 3, 4, 8

[29] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv:1412.6632*, 2014. 2

[30] I. Masuda-Mora, S. Pascual-deLaPuente, and X. Giró-i Nieto. Towards automatic generation of question answer pairs from images. In *CVPRW*, 2016. 1, 2, 3

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013. 6

[32] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. P. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv:1701.08251*, 2017. 2, 3

[33] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. In *ACL*, pages 1802–1813, 2016. 1, 2, 3, 7

[34] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 3

[35] H. Noh, P. Hongsuck Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, pages 30–38, 2016. 2

[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. Association for Computational Linguistics, 2002. 7

[37] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, pages 2953–2961, 2015. 2, 3

[38] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 2

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 4

[40] I. V. Serban, A. García-Durán, C. Gulcehre, S. Ahn, S. Chandar, A. Courville, and Y. Bengio. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv:1603.06807*, 2016. 3

[41] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, pages 4613–4621, 2016. 2

[42] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv:1602.07261*, 2016. 3, 4, 7

[43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 2

[44] T. Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972. 2

[45] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *PAMI*, 2017. 2, 4

[46] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, pages 451–466, 2016. 2

[47] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015. 2

[48] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016. 2

[49] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*, pages 2461–2469, 2015. 2

[50] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 2, 3, 8

[51] L. Yu, H. Tan, M. Bansal, and T. L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, 2017. 3

[52] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. Yin and yang: Balancing and answering binary visual questions. In *CVPR*, pages 5014–5022, 2016. 2

[53] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pages 4995–5004, 2016. 2, 3