# Overcoming Language Priors in VQA via Decomposed Linguistic Representations

**Chenchen Jing,**[1] **Yuwei Wu,**[1*] **Xiaoxun Zhang,**[2] **Yunde Jia,**[1] **Qi Wu**[3]
[1]Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology, China
[2]Alibaba Group
[3]Australian Centre for Robotic Vision, University of Adelaide, Australia
{chenchen.jing,wuyuwei,jiayunde}@bit.edu.cn, xiaoxun.zhang@alibaba-inc.com, qi.wu01@adelaide.edu.au

## Abstract

Most existing Visual Question Answering (VQA) models overly rely on language priors between questions and answers. In this paper, we present a novel method of language attention-based VQA that learns decomposed linguistic representations of questions and utilizes the representations to infer answers for overcoming language priors. We introduce a modular language attention mechanism to parse a question into three phrase representations: type representation, object representation, and concept representation. We use the type representation to identify the question type and the possible answer set (yes/no or specific concepts such as colors or numbers), and the object representation to focus on the relevant region of an image. The concept representation is verified with the attended region to infer the final answer. The proposed method decouples the language-based concept discovery and vision-based concept verification in the process of answer inference to prevent language priors from dominating the answering process. Experiments on the VQA-CP dataset demonstrate the effectiveness of our method.

## Introduction

Recent studies (Kafle and Kanan 2017; Agrawal et al. 2018; Selvaraju et al. 2019) demonstrate that most existing Visual Question Answering (VQA) models overly rely on superficial correlations between questions and answers, *i.e.*, language priors, and ignore image information. For example, they may frequently answer "white" for questions about color, "tennis" for questions about sports, and "yes" for questions beginning with "is there a", no matter what images are given with the questions.

The main reason why these models are vulnerable to language priors is that different kinds of information of questions are entangled in the answer inference process. Most VQA models (Fukui et al. 2016; Yang et al. 2016; Anderson et al. 2018) consist of three parts: extracting informative representations for both images and questions, fusing these representations to obtain joint embeddings of images and questions, and predicting final answers with the joint embeddings. However, these models do not explicitly distinguish and utilize different information in questions, and thus inevitably use the co-occurrences of the answers and interrogative words to infer answers. Although some VQA models (Lu et al. 2016; Ma et al. 2018) adopt the question attention to focus on relevant words with image representations as guidance, they do not eliminate the effect of interrogative words during answer inference and thus are also susceptible to language priors.

To overcome language priors, Agrawal *et al.* (2018) proposed a grounded visual question answering model that exploits different information in questions by using multiple hand-designed modules. They devised a question classifier to classify questions into yes/no questions or non-yes/no questions, a Part-of-Speech-based concept extractor to extract concepts in yes/no questions, and an answer cluster predictor to identify the answer type of non-yes/no questions. In this paper, we propose to learn and exploit decomposed linguistic representations of different kinds of information in questions for overcoming language priors.

A question-answer pair usually contains three kinds of information: question type, referring object, and expected concept. Note that the expected concept is included in the question for yes/no questions, and in the answer for non-yes/no questions. Humans can effortlessly identify and utilize different information in questions to infer answers and are undoubtedly insusceptible to language priors. For instance, if one is asked to answer a question, "Is the man's shirt white?", at a glance of "is", he/she knows that it's a verification question whose possible answers are "yes" and "no". Then he/she localizes the man's shirt in the image via the phrase "man's shirt" and verifies the visual presence/absence of "white", the excepted concept in the question, based on the shirt. This process is also applicable to non-yes/no questions such as "What color is the man's shirt?". The only difference is that apart from knowing that the question is about color via "What color", several concepts (*e.g.*,"white", "black" and "blue") that are possible to be the answer may arise in his/her mind for further verification. It is thus desirable that a VQA model capable of flexibly learning and utilizing the decomposed representations of different information in questions should be established for alleviating the

[*]corresponding author

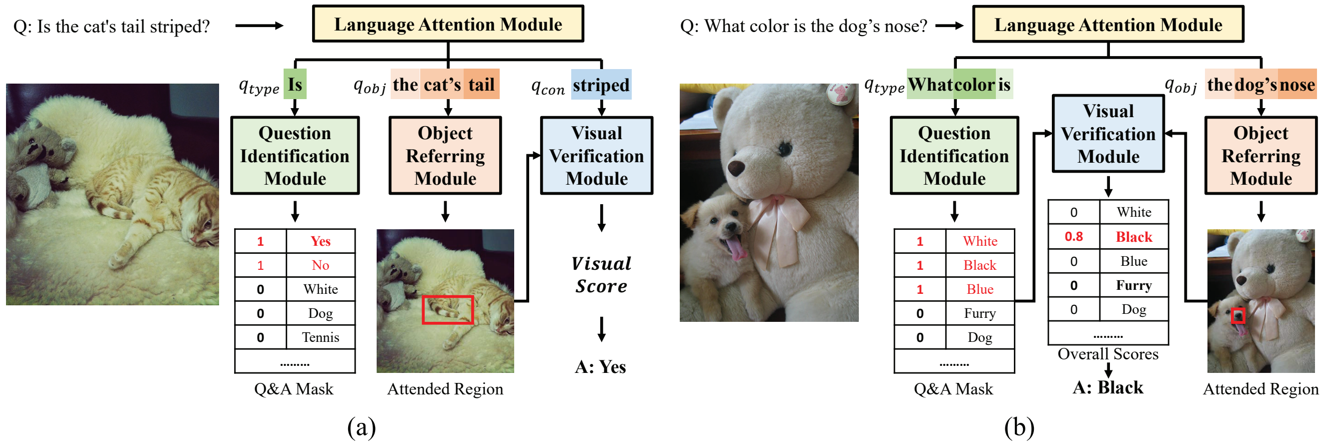

Figure 1: The framework of the proposed method. It decomposes a question into three phrase representations via a language attention module and further utilizes the phrase representations to infer answers via three specific modules: question identification module, object referring module, and visual verification module. (a) is the answering process of the proposed method for yes/no questions, and (b) is the answering process for non-yes/no questions.

influence of language priors.

To this end, we present a novel method of language attention-based VQA that includes a language attention module, a question identification module, an object referring module and a visual verification module, as shown in Figure 1. *The language attention module* parses a question into three phrase representations: type representation, object representation, and concept representation. The decomposed linguistic representations are then fed into the following modules, respectively. By combining the hard attention mechanism and soft attention mechanism, *the language attention module* eliminates the influence of interrogative words when learning concept representations. *The question identification module* uses the type representation to identify the question type and the possible answer set (yes/no or specific concepts such as colors or numbers). A question-answer mask (Q&A mask), denoting whether candidate answers are possible to be the correct answer, is generated by measuring the relevance between the type representation and candidate answers. *The object referring module* adopts the top-down attention mechanism (Anderson et al. 2018) to attend to the relevant region of the image with the object representation as guidance. *The visual verification module* measures the visual score between the attended region and the concept representation to infer answers via a threshold comparison for yes/no questions. For non-yes/no question, possible answers discovered by *the question identification module* serve as the concepts to be verified. The visual scores between the attended region and candidate answers are measured and then fused with the Q&A mask to obtain the final answer.

By identifying and utilizing different information in questions, the proposed method decouples the language-based concept discovery and vision-based concept verification from the answer inference process. Thus the superficial correlations between questions and answers do not dominate the answer inference process, and the model must exploit the image content to infer the final answer from the possible answer set. Moreover, our method achieves a transparent answering process benefiting from the modular design. The intermediate results of the four modules (decomposed phrases, Q&A masks, attended regions and visual scores) actually form explanations of why and how a specific answer is inferred.

The contributions of this paper are summarized as follows: (1) We learn decomposed linguistic representations of questions and decouple the language-based concept discovery and vision-based concept verification to overcome language priors. (2) We use a language attention module combining both the hard attention mechanism and soft attention mechanism to flexibly identify different information in questions while separating concept representations from type representations.

## Related work

**Visual Question Answering.** VQA methods can be divided into two categories: holistic and modular. The holistic methods (Yang et al. 2016; Anderson et al. 2018; Hudson and Manning 2018; Gao et al. 2019; Cadene et al. 2019; Ben-Younes et al. 2019) use a single model for different question-image pairs and are widely used in real-world image VQA datasets such as (Antol et al. 2015; Goyal et al. 2017) where images and questions are diverse. The modular methods (Andreas et al. 2016; Johnson et al. 2017b; Hu et al. 2018; Shi, Zhang, and Li 2019), which focus on compositional reasoning, devise different modules for different sub-tasks and perform better in the synthetic VQA datasets such as (Johnson et al. 2017a). Modular methods first parse a question into a module layout and then execute the modules to infer the answer.

Although our method adopts modular design and parses the questions, our method belongs to the holistic methods because it still uses only one model for different questions. Besides, we parse the questions to obtain phrase representations, which serve as the input of the following modules, rather than to predict the module layout.

**Overcoming language priors for VQA.** To diagnose to what extent VQA models are influenced by language priors, Agrawal *et al.* (2018) curated the VQA-CP dataset, a new split of original VQA dataset (Antol et al. 2015; Goyal et al. 2017). In this dataset, for each question category, the answer distributions of both the train split and test split are different such that models overly rely on language priors perform poorly. Along with the VQA-CP dataset, they proposed a grounded visual question answering (GVQA) model that disentangles the visual concept recognition from the answer space prediction to overcome language priors. They predefined 2000 visual concepts belonging to 50 clusters, and devised visual concept classifiers and an answer cluster classifier to identify the visual concepts of an image and plausible answers of a question, respectively.

Recently, Ramakrishnan *et al.* (2018) proposed an adversarial regularization scheme for VQA models to mitigate the effect of language priors. They introduced a question-only adversary model and optimized question representations to minimize the accuracy of the question-only model while maintaining base VQA model's performance. Aiming to emphasize the significance of visual information, they weakened unwanted correlations between questions and answers while we appropriately use information in questions to guide the vision-based concept verification. Selvaraju *et al.* (2019) proposed a human importance-aware network tuning method that uses human supervision to improve visual grounding. They forced the model to focus on the right region by optimizing the alignment between human attention maps and gradient-based network importance. By contrast, we decouple the concept discovery and concept verification to guarantee that the model exploits image content to infer answers.

Our method is similar to the work of Agrawal *et al.* (2018) in that we exploit different information in questions to decouple the concept discovery and concept verification. However, our method differs from theirs in two aspects. First, we use the language attention mechanism to flexibly learn decomposed representations of questions instead of using Part-of-Speech-based extractors to extract phrases from questions. Second, we regard candidate answers as visual concepts and learn their relevance with questions and images in an end-to-end manner, while they predefined various visual concepts and identified the concepts in images with pre-trained classifiers. In summary, our method guarantees that different information in questions can be flexibly identified and appropriately utilized in a unified framework by learning decomposed linguistic representations.

## Method

As shown in Figure 1, the proposed method includes four modules: (a) a language attention module parses a question into the type representation, the object representation, and the concept representation; (b) a question identification module uses the type representation to identify the question type and possible answers; (c) an object referring module uses the object representation to attend to the relevant region of an image; (d) a visual verification module measures the relevance between the attended region and the concept representation to infer the answer.

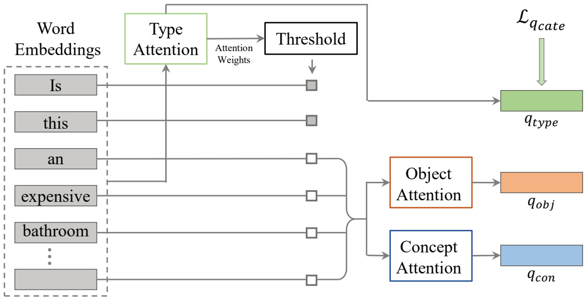

Figure 2: The architecture of the language attention module.

The input of our method includes an image $I \in \mathcal{I}$ represented by $K$ local features $\{\boldsymbol{v}_k\}_{k=1}^{K}$, a question $Q \in \mathcal{Q}$, and the candidate answer set $\mathcal{A}$. For the question and answers, the pre-trained GloVe (Pennington, Socher, and Manning 2014) is utilized to initialize the embedding of each word.

### Language Attention Module

To obtain phrase representations of questions, a possible solution is using off-the-shelf parsers (Socher et al. 2013) to parse a question into different grammatical components, such as a triplet $\langle subject, predicate, object \rangle$. However, the phrases generated by the parsers are not always satisfactory because the parsers do not allow a word to appear in more than one grammatical component. For a simple question like "Is there a dog?", compared with utilizing the word "there" to focus on the whole image and verify it with the concept "dog", a better solution is to find the region that contains "dog" and verify the region also with "dog". In this case, the word "dog" should appear in both the phrase about the referring object and the phrase about the expected concept .

Recently, Hu *et al.* (2017) and Yu *et al.* (2018) utilized the modular soft attention mechanism, where the embeddings of all words can be adaptively re-weighted and then aggregated to be phrase representations, to automatically parse referring expressions into different phrases. However, their method is still not applicable to our task. Due to the existence of language priors, directly using modular soft attention to learn different phrase representations without any constraints will lead to that the model may exploit the co-occurrence of the interrogative words and answers when learning concept representations. In other words, superficial correlations will dominate the answering process. Thus it is critical to flexibly discover different information in questions while alleviating the effect of language priors.

Inspired by the hard attention mechanism (Malinowski et al. 2018), where only a subset of information is selected for further processing, we propose a *language attention module* to obtain decomposed linguistic representations. The language attention module combines the hard attention mechanism and soft attention mechanism to separate the concept representations of yes/no questions from the type representations. As shown in Figure 2, the language attention module uses three kinds of attention, which are type attention, object attention and concept attention, to learn the three decomposed representations, respectively. Specifically, we adopt a question category identification loss to guarantee the type attention attends to the interrogative words and a threshold to filter the interrogative words out when learning the object representations and concept representations. Here we also exclude the interrogative words for the object representations of yes/no questions, because the interrogative words are useless in object referring.

For a question $Q$ that is a sequence of $T$ words $\{w_t\}_{t=1}^{T}$ with its embeddings $\{\boldsymbol{e}_t\}_{t=1}^{T}$, we adopt a trainable vector $\boldsymbol{w}_{a,t}$ to compute the attention weights and obtain the type representation $\boldsymbol{q}_{type}$ by

$$\boldsymbol{q}_{type} = \sum_{i=1}^{T} \alpha_t^{type} \boldsymbol{e}_i, \quad \alpha_i^{type} = \frac{exp(\boldsymbol{w}_{a,t}^T \boldsymbol{e}_i)}{\sum_{j=1}^{T} exp(\boldsymbol{w}_{a,t}^T \boldsymbol{e}_j)}. \quad (1)$$

To guarantee that the type attention focuses on interrogative words, we use the interrogative word-based question category provided in the original VQA dataset (Antol et al. 2015; Goyal et al. 2017) as supervision information to guide the learning of type representations. Note that the question category, that includes "is there", "are there", etc., is different from the aforementioned question type that only contains two types: yes/no questions and non-yes/no questions. Thus we use "question category" here to avoid confusion. We devise a question category identification loss given by

$$\mathcal{L}_{q_{cate}} = -\frac{exp(\boldsymbol{W}_{y_{cate}}^T \boldsymbol{q}_{type})}{\sum_{j=1}^{C} exp(\boldsymbol{W}_j^T \boldsymbol{q}_{type})}, \quad (2)$$

where $C$ is the number of question categories, $\boldsymbol{W}$ is a trainable weight matrix, and $y_{cate} \in \{1, 2, 3, ....C\}$ denotes the ground truth question category of the question $Q$.

We further introduce a scalar $\beta$ as the threshold to filter out the words most related to the question type, to eliminate the influence of interrogative words for the concept attention. By comparing the type attention weight of each word with $\beta$, we can obtain a set of words $\{w_p\}_{p=1}^{p}$, where the type attention weight of each word is lower than $\beta$. These words are supposed to be relevant to the referring object and the expected concept. Then another two trainable vectors, $\boldsymbol{w}_{a,o}$ and $\boldsymbol{w}_{a,c}$, are introduced to compute the attention weights for the object attention and the concept attention, respectively. The final phrase representations are obtained by

$$\begin{aligned} \boldsymbol{q}_{obj} &= \sum_{p=1}^{P} \alpha_p^{obj} \boldsymbol{e}_p, \quad \alpha_p^{obj} = \frac{exp(\boldsymbol{w}_{a,o}^T \boldsymbol{e}_p)}{\sum_{j=1}^{P} exp(\boldsymbol{w}_{a,o}^T \boldsymbol{e}_j)}, \\ \boldsymbol{q}_{con} &= \sum_{p=1}^{P} \alpha_p^{con} \boldsymbol{e}_p, \quad \alpha_p^{con} = \frac{exp(\boldsymbol{w}_{a,c}^T \boldsymbol{e}_p)}{\sum_{j=1}^{P} exp(\boldsymbol{w}_{a,c}^T \boldsymbol{e}_j)}. \end{aligned} \quad (3)$$

In this manner, the object attention and the concept attention function as the hard attention as only a subset of words are taken into consideration.

For the answer $A$ that is a sequence of $S$ words $\{w_s\}_{s=1}^{S}$ with its embeddings $\{\boldsymbol{e}_s\}_{s=1}^{S}$, the Gated Recurrent Unit (GRU) (Cho et al. 2014) is adopted to obtain the sentence-level answer representation $\boldsymbol{a}$.

## Question Identification Module

Given the type representation $\boldsymbol{q}_{type}$, we first identify the question type via a question identification loss,

$$\mathcal{L}_{q_{type}} = CE(y_{type}, \boldsymbol{w}_{type}^T \boldsymbol{q}_{type}), \quad (4)$$

where $CE(a, b) = -a\log(b) - (1-a)log(1-b)$ denotes the cross entropy function, $\boldsymbol{w}_{type}$ is a trainable vector, and $y_{type}$ denotes the ground truth type of question $Q$, which is 1 for yes/no questions and 0 for non-yes/no questions.

The possible answers of yes/no questions are naturally "yes" and "no", while the possible answers of non-yes/no questions are identified by measuring the relevance between the question and the candidate answers. For each non-yes/no question, a Q&A mask $\boldsymbol{m}_q \in (0,1)^{|\mathcal{A}| \times 1}$ is generated, where each element denotes the possibility for a candidate answer to be the correct answer, and $|\mathcal{A}|$ denotes the number of candidate answers. We first compute the relevance scores between the question $Q$ and all the candidate answers as $s_{qa}(Q, A_j) = \boldsymbol{q}_{type} \cdot \boldsymbol{a}_j$ to obtain the relevance scores $\boldsymbol{s}_q \in \mathbb{R}^{|\mathcal{A}| \times 1}$, where $\cdot$ denotes the dot product. Then the sigmoid function is utilized to project the relevance scores to the range of $(0, 1)$ and the mask is thus generated by $\boldsymbol{m}_q = sigmoid(\boldsymbol{s}_q)$.

To effectively guide the mask generation, we search all possible answers for each question category in the dataset to obtain a ground truth Q&A mask $\boldsymbol{M} \in \mathbb{R}^{C \times |\mathcal{A}|}$ as supervision information. For each question category, possible answers are marked as 1 and otherwise 0. The KL-divergence is utilized to measure the distance between the generated mask $\boldsymbol{m}_q$ of a question and the ground truth mask $\boldsymbol{M}_{y_{cate}}$ determined by its question category. The final mask generation loss is given by

$$\mathcal{L}_{mask} = \sum_{j=1}^{|\mathcal{A}|} \boldsymbol{M}_{y_{cate},j} log \frac{\boldsymbol{M}_{y_{cate},j}}{\boldsymbol{m}_{q,j}}. \quad (5)$$

Note that $M$ only provides weak supervision information because the provided question categories are relatively simple. For example, the category "what is" contains many sub-categories such as "what is the color", "what is the animal", "what is the number". Each sub-category has specific possible answer set, but $M$ does not explicitly differentiate these sub-categories. Thus to better characterize the relevance between questions and answers, in testing, we round the elements in $\boldsymbol{m}_q$ to the nearest tenth, that is $\{0, 0.1, 0.2, ..., 0.9, 1\}$, instead of generating a binary mask via a threshold comparison.

## Object Referring Module

The object referring module uses the object representation $\boldsymbol{q}_{obj}$ to attend to the region relevant to the question in the image. Given a set of local features of the image $\{\boldsymbol{v}_k\}_{k=1}^{K}$ and the object representation $\boldsymbol{q}_{obj}$, the top-down attention mechanism (Anderson et al. 2018) is adopted to weight the local features by their relevance with the question and further obtain the final visual representation $\boldsymbol{v}$ as

$$\boldsymbol{v} = \sum_{k=1}^{K} \alpha_k^v \boldsymbol{v}_k,\ \alpha_k^v = \frac{exp(\boldsymbol{W}_v^T \boldsymbol{v}_k \cdot \boldsymbol{W}_q^T \boldsymbol{q}_{obj})}{\sum_{j=1}^{K} exp(\boldsymbol{W}_v^T \boldsymbol{v}_j \cdot \boldsymbol{W}_q^T \boldsymbol{q}_{obj})}, \quad (6)$$

where $\boldsymbol{W}_v$ and $\boldsymbol{W}_q$ are trainable weight matrices.

## Visual Verification Module

The visual verification module verifies the visual presence/absence of concepts based on the attended region to infer the final answer. For yes/no questions, given the concept representation $\boldsymbol{q}_{con}$ and the visual representation $\boldsymbol{v}$ for the attended region, the visual score is computed as $S_{qv}(I,Q) = \boldsymbol{q}_{con} \cdot \boldsymbol{v}$. The cross entropy loss is adopted as objective as shown in Eq. (8).

For non-yes/no questions, we first compute the visual scores between the attended region and all candidate answers by $S_{va}(A_j, I) = \boldsymbol{a}_j \cdot \boldsymbol{v}$ and obtain a visual score vector $\boldsymbol{s}_v \in \mathbb{R}^{|\mathcal{A}|\times 1}$ . Then, we fuse the visual score vector $\boldsymbol{s}_v$, which represents the relevance between the image and candidate answers, and the Q&A mask $\boldsymbol{m}_q$, which represents the relevance between the question and candidate answers, to obtain the overall scores $\boldsymbol{s}_{vqa} = \boldsymbol{m}_q \circ \boldsymbol{s}_v$, where $\circ$ denotes the element-wise product. Thus given an image $I$ and a question $Q$, the probability for a candidate answer $A_j$ to be correct is

$$P(A_j|I,Q) = \frac{exp(\boldsymbol{s}_{vqa,j})}{\sum_{l=1}^{|\mathcal{A}|} exp(\boldsymbol{s}_{vqa,l})}. \quad (7)$$

In practice, each question-image pair is assigned with one or several similar correct answers provided by different annotators. Thus the answers for an question-image pair $\langle I, Q\rangle$ can be regarded as a distribution vector $\boldsymbol{y} \in (0,1)^{|\mathcal{A}|\times 1}$, where $\boldsymbol{y}_j$ indicates the occurrence probability of the answer $A_j$ across human labeled answers. Thus we adopt the KL-divergence as the distance metric and the overall verification loss is given by

$$\mathcal{L}_{veri} = \begin{cases} \sum\limits_{j=1}^{|\mathcal{A}|} \boldsymbol{y}_j \log \frac{\boldsymbol{y}_j}{P(A_j|I,Q)}, & y_{type} = 0, \\ CE(b, \sigma(S_{qv(I,Q)})), & y_{type} = 1, \end{cases} \quad (8)$$

where $\sigma(\cdot)$ is the sigmoid function and $b$ is the ground truth label for yes/no questions.

In model learning, we sum up all the aforementioned losses as the overall objective of the proposed method. In testing, we first identify the question type via Eq. (4). Then for yes/no question, the visual score is computed and compared with 0.5 to obtain the final answer. For non-yes/no question, the model takes as input all the candidate answers and the answer with the highest overall score are selected to be the answer.

# Experiments

## Datasets and Experimental Settings

**Datasets.** We evaluate the effectiveness of the proposed method in the VQA-CP v2 dataset (Agrawal et al. 2018) using standard VQA evaluation metric (Antol et al. 2015). The train split and test split of VQA-CP v2 is created by re-organizing the train split and validation split of the VQA v2 (Goyal et al. 2017). In the dataset, the distribution of answers per question category such as "what number" and "are there", is different in the test split from its in the train split. Consequently, VQA models that are overly driven by language priors perform poorly in this dataset. We also report the results on the validation split of the original VQA v2 dataset for completeness.

**Implementation Detail.** We build our model on the bottom-up and top-down attention (UpDn) method (Anderson et al. 2018) as (Ramakrishnan, Agrawal, and Lee 2018) and (Selvaraju et al. 2019). The UpDn utilizes two kinds of attention mechanisms: bottom-up attention and top-down attention. The bottom-up attention generates object proposals with Faster R-CNN (Ren et al. 2015), while the top-down attention predicts an attention distribution over the proposals using the question representations as guidance. For each image, the UpDn generates no more than 100 proposals with its 2048-d feature. The questions are preprocessed to a maximum of 14 words. The extra words are discarded and the questions shorter than 14 words are padded with vectors of zeros. The pre-trained GloVe is used to initialize the word embeddings with the dimension of 300 and then the GRU is used to obtain sentence-level question embeddings with the dimension of 512.

In our implementation, we set $K$ as 36 for each image, thus the dimension of features of an image is $36 \times 2048$. For question embeddings, we replace the original GRU with our language attention module, since the proposed method does not utilize the RNN-based sentence-level embeddings. The answers are preprocessed to a maximum of 3 words. Answers appear no more than 9 times in dataset are excluded from candidate answer set $\mathcal{A}$. We use the 65 kinds of interrogative word-based question categories provided by the VQA v2 (Goyal et al. 2017) in the language attention module and the question identification module. To make sure the interrogative words are filtered out for yes/no questions, we set the threshold $\beta$ in the language attention module as 0.1, which is a little bigger than average attention weight, *i.e.*, 0.07. In the VQA-CP, we set the number of training epochs as 30 and the final model is used for evaluation without early-stopping because there is no validation set.

## Results and Analysis

**Comparison with the state-of-the-art.** The results of our method and state-of-the-art VQA models on the VQA-CP v2 dataset are listed in Table 1. It is shown that our method brings remarkably improvement for its base model, the UpDn. By learning and exploiting decomposed linguistic representations, the proposed method decouples the language-based concept discovery and vision-based concept verification from the answer inference process. For yes/no questions, the language attention module explicitly separates the concept representation from interrogative words. Therefore the model needs to verify the visual presence/absence of the concept in questions based on image content to infer answers, instead of relying on the interrogative words. For

Table 1: Results of our method and the state-of-the-art on the VQA-CP v2 and the VQA v2.

| Model | *VQA-CP v2 test* | | | | *VQA v2 val* | | | |
|---|---|---|---|---|---|---|---|---|
| | Overall | Yes/No | Numbers | Other | Overall | Yes/No | Numbers | Other |
| SAN (Yang et al. 2016) | 24.96 | 38.35 | 11.14 | 21.74 | 52.41 | 70.06 | 39.28 | 47.84 |
| UpDn (Anderson et al. 2018) | 39.49 | 45.21 | 11.96 | 42.98 | 62.85 | 80.89 | 42.78 | 54.44 |
| GVQA (SAN) (Agrawal et al. 2018) | 31.30 | 57.99 | 13.68 | 22.14 | 48.24 | 72.03 | 31.17 | 34.65 |
| AdvReg (UpDn) (Ramakrishnan et al. 2018) | 41.17 | 65.49 | 15.48 | 35.48 | 62.75 | 79.84 | 42.35 | 55.16 |
| HINT (UpDn) (Selvaraju et al. 2019) | 47.7 | 70.04 | 10.68 | **46.31** | 62.35 | 80.49 | 41.75 | 54.01 |
| Ours (SAN) | 34.83 | 57.28 | 15.11 | 28.48 | 49.27 | 66.71 | 32.47 | 40.43 |
| Ours (UpDn) | **48.87** | **70.99** | **18.72** | 45.57 | 57.96 | 76.82 | 39.33 | 48.54 |

non-yes/no questions, the possible answers identified in the question identification module are verified with the attended region. Then the answer with the highest score is selected to be the inferred answer. Thus the model needs to exploit the image content to select the most relevant answer. In summary, our method guarantees that the model must exploit the visual information of images to infer the correct answer from the possible answer set and therefore significantly alleviates the influence of language priors. Particular, we observe that our method works much better than others in the "number" subset. The main reason is that the questions of this subset are more difficult than the questions about color, sport, etc., considering the counting ability is needed to answer them. In this case, the models are more inclined to rely on language priors. As a result, our method performs better since we decouple the language-based concept discovery and vision-based concept verification to overcome the language priors.

The proposed method outperforms the AdvReg (Ramakrishnan, Agrawal, and Lee 2018) and the HINT (Selvaraju et al. 2019) in the overall accuracy with the same base model. By learning question representations based on which the question-only adversary model can't infer correct answers, the AdvReg weakens the superficial correlations between questions and answers. By contrast, we learn and exploit the decomposed representations of questions to explicitly decouple the concept discovery and concept verification to prevent the language priors from dominating the answer inference. The results show the effectiveness of the decomposed linguistic representations. The HINT effectively leverages human attention maps to encourage the model to focus on the right regions and achieves state-of-the-art performance. However, the human supervision they used is not always available, which limits the generalization ability to other VQA datasets. The proposed method outperforms the HINT, without the human supervision. The ground truth of the question type, the question category, and the Q&A mask used in our method are from the original VQA dataset and can be easily obtained for any other VQA datasets.

For fair comparisons with GVQA, we particularly build our method upon the SAN (Yang et al. 2016) and obtain a model marked as "ours (SAN)" in Table 1. The main reason why this model outperforms the GVQA is that our method integrates the identification and the utilization of different information in questions into an end-to-end trained model, and thus results in better compatibility.

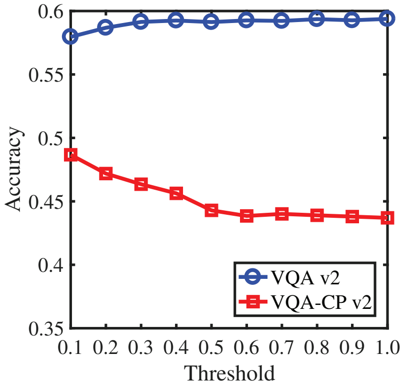

Figure 3: Results with different thresholds on two datasets.

**Results on the original VQA dataset.** The results on the VQA v2 dataset are also listed in Table 1. It is shown that all models for overcoming language priors perform worse than their base models because they avoid overfitting the dataset biases of VQA v2 dataset. Moreover, we observe our method and the GVQA suffer more performance drop on the VQA v2 dataset compared with the corresponding base model than the AdvReg and the HINT. Intuitively, by decoupling the concept discovery and concept verification, the former two methods can more effectively handle the situation that the model focuses on the right region but still predict the wrong answers via language priors. Thus they can more effectively prevent models from exploiting language priors.

**Parameter analysis.** The threshold in the language attention module is a critical hyper-parameter for the proposed method. To measure the influence of the threshold, we vary it from 0.1 to 1 to train different models on both datasets, and the results are shown in Figure 3. We observe that the threshold can control the ability of our method to overcome language priors. As the threshold increases, our method performs better on the VQA v2 but worse on the VQA-CP v2. The reason is that when the threshold is set higher, fewer interrogative words are filtered out and the model will exploit the interrogative words to infer the answer.

**Ablation studies.** To evaluate the effectiveness of several important components of our method, we re-train different versions of our model by ablating certain components. The results on the VQA-CP v2 dataset are listed in Table 2.

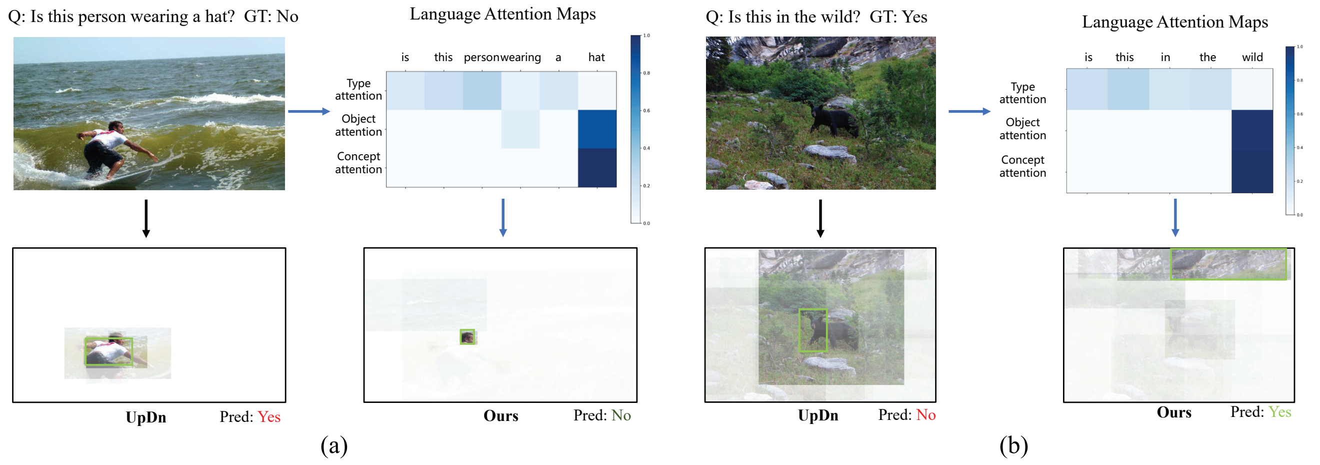

Figure 4: Qualitative comparisons between the proposed method and the UpDn. For each example, the top left shows an input question in the test split of the VQA-CP v2 dataset, along with the image and the ground-truth (GT) answer. The top right shows the language attention maps of the language attention module. The bottom row shows the visual attention maps for object referring and the predicted answers of the UpDn and our method, respectively. The region with the highest weight is marked with a green rectangle.

We first investigate the effectiveness of the language attention module of the proposed method. To this end, we replace the language attention module with the original GRU and obtain a model marked as "Ours w/o LA", where the "w/o" represents "without". The "Ours w/o LA" obtains sentence-level question representations via the GRU and input the representations into the following modules. Apart from the "Ours w/o LA", we further replace the language attention module with the ordinary modular soft attention mechanism and obtain a model called "Ours w/o threshold". In other words, we remove the threshold and the question category identification loss of the language attention module. As shown in Table 2, the "Ours w/o LA" performs worse than the full model in all three subsets. This clearly demonstrates the effectiveness learning decomposed representations for overcoming language priors. The "Ours w/o threshold" performs better than "Ours w/o LA" but still significantly worse than the full model. This demonstrates simply using modular soft attention to learn decomposed representations without any constraints results in that the model is still vulnerable to language priors.

Then we investigate the influence of the mask generation process in the question identification module. We exclude the mask generation process and the obtained model is called "Ours w/o Mask", which selects the answer with the highest visual score as the final answer. From Table 2, we find that the Q&A mask brings substantive improvement on the "other" subset. The model without the Q&A masks performs worse than the base model in "numbers" and "other" subsets, while the full model outperforms the base model by a large margin.

**Qualitative examples.** Figure 4 depicts two qualitative examples about yes/no questions that show the effectiveness of the proposed model in question parsing and visual grounding. It can be shown that, in both cases, our method correctly identifies different kinds of information in the question and localizes the relevant region in the image more accurately than the UpDn. As a result, our method infers the right answers.

Table 2: Ablation studies on the VQA-CP v2

| Model | *VQA-CP v2 test* | | | |
|---|---|---|---|---|
| | Overall | Yes/No | Numbers | Other |
| UpDn (Anderson et al. 2018) | 39.49 | 45.21 | 11.96 | 42.98 |
| Ours w/o LA | 41.49 | 49.07 | 14.06 | 45.04 |
| Ours w/o threshold | 42.50 | 51.22 | 17.98 | 44.66 |
| Ours w/o Mask | 43.39 | 70.39 | 16.64 | 36.59 |
| Ours | **48.87** | **70.99** | **18.72** | **45.57** |

## Conclusion

In this work, we have presented a novel method of language attention-based VQA. Our method learns decomposed linguistic representations of questions to overcome language priors. Using a language attention module, we can flexibly parse a question into three phrase representations. These representations was appropriately utilized to decouple the language-based concept discovery and vision-based concept verification from the answer inference process. Thus superficial correlations between questions and answers can not dominate the answering process and the model must exploit the images to infer answers. Besides, our method can achieve a more transparent answering process with informative intermediate results. Experimental results on the VQA-CP dataset show the effectiveness of our method.

# Acknowledgments

This work was supported in part by the Beijing Municipal Natural Science Foundation under Grant No. L172027, the Natural Science Foundation of China (NSFC) under Grant No. 61702037, and Alibaba Group through Alibaba Innovative Research Program. The work of Qi Wu was supported in part by the ARC DECRA DE190100539.

# References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Dont just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4971–4980.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural module networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 39–48.

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

Ben-Younes, H.; Cadene, R.; Thome, N.; and Cord, M. 2019. Block: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection. In *The Thirty-Third AAAI Conference on Artificial Intelligence*.

Cadene, R.; Ben-Younes, H.; Cord, M.; and Thome, N. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1989–1998.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

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

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6639–6648.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 3.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1115–1124.

Hu, R.; Andreas, J.; Darrell, T.; and Saenko, K. 2018. Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 53–69.

Hudson, D. A., and Manning, C. D. 2018. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. 2017a. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1988–1997.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Hoffman, J.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017b. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2989–2998.

Kafle, K., and Kanan, C. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1983–1991.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 289–297.

Ma, C.; Shen, C.; Dick, A.; Wu, Q.; Wang, P.; van den Hengel, A.; and Reid, I. 2018. Visual question answering with memory-augmented networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6975–6984.

Malinowski, M.; Doersch, C.; Santoro, A.; and Battaglia, P. 2018. Learning visual question answering by bootstrapping hard attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–20.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Ramakrishnan, S.; Agrawal, A.; and Lee, S. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1541–1551.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 91–99.

Selvaraju, R. R.; Lee, S.; Shen, Y.; Jin, H.; Batra, D.; and Parikh, D. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Shi, J.; Zhang, H.; and Li, J. 2019. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Socher, R.; Bauer, J.; Manning, C. D.; et al. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, 455–465.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21–29.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1307–1315.