Quantifying and Alleviating the Language Prior Problem in Visual Question Answering

Yangyang Guo[†], Zhiyong Cheng[§], Liqiang Nie[†], Yibing Liu[†], Yinglong Wang[§], Mohan Kankanhalli[‡]

[†]School of Computer Science and Technology, Shandong University
[§]Shandong Computer Science Center (National Supercomputer Center in Jinan),
Qilu University of Technology (Shandong Academy of Sciences)
[‡]School of Computing, National University of Singapore
{guoyang.eric,jason.zy.cheng,nieliqiang,lyibing112}@gmail.com,wangyl@sdas.org,mohan@comp.nus.edu.sg

ABSTRACT

Benefiting from the advancement of computer vision, natural language processing and information retrieval techniques, visual question answering (VQA), which aims to answer questions about an image or a video, has received lots of attentions over the past few years. Although some progress has been achieved so far, several studies have pointed out that current VQA models are heavily affected by the *language prior problem*, which means they tend to answer questions based on the co-occurrence patterns of question keywords (e.g., *how many*) and answers (e.g., *2*) instead of understanding images and questions. Existing methods attempt to solve this problem by either *balancing the biased datasets* or *forcing models to better understand images*. However, only marginal effects and even performance deterioration are observed for the first and second solution, respectively. In addition, another important issue is the inability to quantitatively measure the extent of the language prior effect, which severely hinders the advancement of related techniques.

In this paper, we make contributions towards solving the above problems from two perspectives. Firstly, we design a metric to quantitatively measure the language prior effect on VQA models. The proposed metric has been demonstrated to be effective in our empirical studies. Secondly, we propose a regularization method (i.e., score regularization module) to enhance current VQA models by alleviating the language prior problem as well as boosting the backbone model performance. The proposed score regularization module adopts a pair-wise learning strategy, which makes the VQA models answer the question based on the reasoning on the image (upon this question) instead of basing on question-answer patterns observed in the biased training set. The score regularization module is versatile to be integrated into various VQA models. We conducted extensive experiments over two popular VQA datasets (i.e., VQA 1.0 and VQA 2.0) and integrated the score regularization module into three state-of-the-art VQA models. Experimental results show that the score regularization module can not only effectively reduce the language prior problem of these VQA models but also consistently improve their question answering accuracy.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Natural language processing**; **Computer vision**;

KEYWORDS

Visual Question Answering, Language Prior Problem, Evaluation Metric

ACM Reference Format:

Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang and Mohan Kankanhalli. 2019. Quantifying and Alleviating the Language Prior Problem in Visual Question Answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331186

1 INTRODUCTION

Question Answering (QA) has been long recognized as a challenging information retrieval task. At the beginning, it focused only on the text domain. With the great progress of natural language processing (NLP), computer vision (CV) and information retrieval (IR), a new 'AI-complete' task, namely visual question answering (VQA), has become an emerging interdisciplinary research field over the past few years. VQA aims to accurately answer natural language questions about a given image or a video, bringing bright prospects in various applications including medical assistance and human-machine interaction. Recently, several benchmark datasets have been constructed to facilitate this task [5, 21, 25, 45], followed by a number of devised deep models [4, 5, 24–26, 40].

Although these methods have achieved state-of-the-art performance over their contemporary baselines, many studies point out that today's VQA models are still heavily driven by superficial correlations between questions and answers in the training data and lack sufficient visual understanding [2, 18, 32]. As a consequence,

Corresponding Author: Zhiyong Cheng and Liqiang Nie.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331186

it turns out that the carefully designed VQA models actually provide answers based upon the first few words in questions and can frequently yield not bad performance. Taking the VQA 1.0 training set [5] as an example, *2* accounts for 31% of the answers to questions initiating with *how many*. This leads to VQA models overwhelmingly replying to 'how many ...' questions with *2* without truly understanding the given images when testing. The problem that *the predicted answers are strongly driven by the answer set from the same question type*[1] *in the training set* is the so called *language prior problem* [2, 13, 37] that many VQA models confront with.

It is not hard to understand the reason of the language prior problem, however this problem is non-trivial to deal with. One reason for this unsatisfactory behavior is the fundamentally problematic nature of independent and identically distributed (i.e., IID) train-test splits in the presence of strong priors. Accordingly, it is hard to distinguish in a well-performing model between making progress towards the goal of understanding images correctly and only exploiting language priors to achieve high accuracy [2]. Moreover, tackling with the language prior problem without deteriorating the model performance poses another challenge.

With the realization of the language prior problem in VQA, researchers have devoted great efforts towards solving or somehow alleviating the problem and developed a set of approaches. Existing approaches can be broadly classified into two directions: 1) making the datasets less biased; and 2) making the model answer questions by analyzing the image contents. In the first direction, researchers in [13, 42] tried to balance the existing VQA 1.0 dataset by adding complementary entries and built the VQA 2.0 dataset [13]. More concretely, for each <image, question, answer> triplet, another triplet with a visually similar image but a different answer is collected to elevate the role of images in VQA. However, even with this balance, there still exists significant bias in the augmented VQA 2.0 dataset. For instance, *2* still accounts for 27% of the question type *how many* in the training set of the VQA 2.0 dataset. Instead of amending the datasets, Johnson et al. [17] designed a diagnostic 3D shape dataset to control the question-conditional bias via rejection sampling within families of related questions. Since they dealt with the problem from the perspective of datasets and attempted to circumvent the inherent deficiency in traditional biased datasets, the language prior problem of previous methods is still not settled.

In contrast, researchers in the second direction make efforts to design mechanisms to make the VQA models avoid the language prior problem. Approaches in this direction can be directly used in the biased datasets and thus are more generalizable. For example, the method in [2] explicitly disentangles the recognition of visual concepts present in the image from the answer prediction for a given question. And more recently, Ramakrishnan et al. [32] treated the training as an adversarial game between the VQA model and the QA model (eliminating images from the current triplet) to reduce the impact of language biases. Both methods are built upon the widely used VQA model Stacked Attention Networks (SAN) [40]. Nevertheless, performance deterioration is observed for both methods as compared to the backbone model SAN. We argue that a better regularization can not only alleviate the language prior problem but also improve the model performance.

[1]Questions initiate with the same words.

Another important issue is the lack of proper evaluation metrics to measure the extent of language prior effect of VQA models. Although the language prior problem has been pointed out by various previous studies [1, 13, 17, 18, 42] and many approaches have been proposed to deal with this problem [2, 32], few efforts have been devoted into how to numerically quantify the language prior effect. As discussed, it is hard to distinguish whether the model really understands the question and image contents before answering the question or it just simply discovers some patterns between question words and answers. Besides, it is also difficult to evaluate how well a newly designed model solves the language prior problem.

In order to tackle the aforementioned limitations of the previous approaches and the lack of language prior measurement, in this paper, we establish a formal quantitative metric to measure the extent of language prior effect (called LP score) and design a generalized regularization method to alleviate the language prior problem in VQA. On the one hand, the proposed LP score evaluates the language prior effect by taking into account both the training dataset bias and model deficiency. In this way, the LP score can measure the language prior effect quantitatively and guide further studies on alleviating the language prior problem. On the other hand, our proposed regularization method leverages a *score regularization module* to force backbone models to better reason with the image contents before predicting answers. More specifically, the score regularization module is added to the backbone models before the final answer prediction layer. This is to guarantee that the VQA model answers questions by understanding questions and corresponding image contents instead of simply analyzing the co-occurrence patterns of question key words (e.g., *how many*) and the answer (e.g., *2*). To achieve the goal, the inputs to the score regularization module are from two streams: *fused question-image feature with the embedding feature of the true answer* and *question feature with the embedding feature of the true answer*; and then the score regularization module computes two scores and employs a pair-wise learning scheme for training. Different from the multi-step learning as adopted in [2, 32], we train the proposed regularizer with the backbone model in an end-to-end multi-task learning scheme. Moreover, our proposed regularization method can be applied to most of the existing VQA models on the biased datasets.

To verify the effectiveness of our proposed regularization method, we conducted extensive experiments on two most popular datasets VQA 1.0 [5] and VQA 2.0 [13]. Moreover, we added the proposed regularization module to three state-of-the-art models. Experimental results demonstrate that our proposed methods can yield better performance as compared to the corresponding backbone models, and thus achieve state-of-the-art performance.

In summary, our main contributions in this paper are threefold:

- To the best of our knowledge, we are the first to study the lack of language prior measurement and emphasize its importance on facilitating the advancement of related techniques. We further design an evaluation metric to quantitatively measure the extent of language prior effect of VQA models.
- We propose a regularization method on VQA models by forcing the backbone model to better reason visual contents before the final answer prediction. The proposed regularizer can help

reduce the language prior effect as well as boost the model performance. It is expected that our method can be extended to other visual-language reasoning tasks which also suffer from the language prior problem, e.g., image captioning.

- We conducted extensive comparative experiments on two publicly available datasets to validate the effectiveness of the proposed regularization method and the feasibility of the proposed evaluation metric. Moreover, we have released the code and setting to facilitate future research in this direction[2].

2 LANGUAGE PRIOR QUANTIFICATION

2.1 Observations

Before elaborating the conception of our language prior measurement, let us go through some examples to show the language prior problem intuitively. Figure 1 shows the answer distributions in the VQA 1.0 dataset [5] of two question types: *how many* and *what color*[3]. For both sub figures, the leftmost bar represents the ground truth answer distribution (i.e., GT-train) in the training set. For example, the 31% answers to the question type *how many* are *2*. And the right ones are the distributions of the predicted wrong answers of three baselines in the validation set. The Question-only [5, 13] is the model trained without reasoning the image, and it is for sure that in this baseline would arise the language prior problem. The other two HieCoAttn [24] and Strong-baseline [20] are the state-of-the-art models on the VQA 1.0 dataset. For example, if the true answer for a given question of the question type *how many* is *6*, and the predicted answer is *2*, then it is counted as a predicted wrong answer of *2* under the question type *how many*. Without dataset bias or language prior, the predicted wrong answers from VQA models should be more diverse (*5*, *apple*, *on the left*, etc) or roughly follow uniform distribution over all answers instead of being proportional to the answer distribution in the training set. For example, a large portion of answers from *how many* questions are mispredicted *2*, *1* and *3*, which are also the most frequent true answers in the training dataset (as shown in Figure 1). This indicates that VQA models tend to provide answers according to the patterns observed between question types and answers in the training set rather than reasoning about images for the current question.

In the recent work, Goyal et al. [13] attempted to deal with the language prior problem of the popular VQA 1.0 dataset by collecting <image, question, answer> triplets to construct the VQA 2.0 dataset. Instead of associating each question with just one image as in the VQA 1.0 dataset, the VQA 2.0 dataset assigns a pair of similar images with different answers to the same question. However, as shown in Figure 2, the language prior problem still exists. The predicted wrong answer distributions from the state-of-the-art models Up-Down [3] and Counter [43] are still highly biased.

The phenomenon arises from the two aspects: 1) training dataset bias. It is common that some answers are more frequently relevant to a certain question type. For example, *red*, *white*, *blue* and *black* are the most frequent colors in daily life, constituting the most frequent answers to the question type *what color*. Moreover, people only ask the question 'Is there a clock tower in the picture?' on images actually containing a clock tower. As experimented by [13], blindly

[2]https://github.com/guoyang9/vqa-prior.

[3]These two types of questions take about 20% of the questions in the whole dataset.

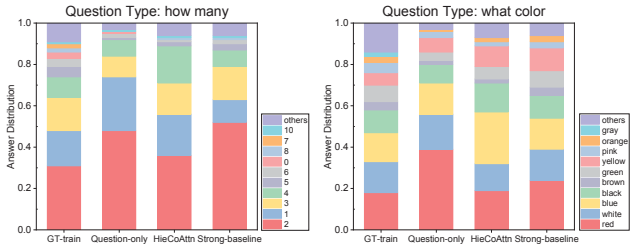

Figure 1: Answer distribution of two question types in the VQA 1.0 dataset.

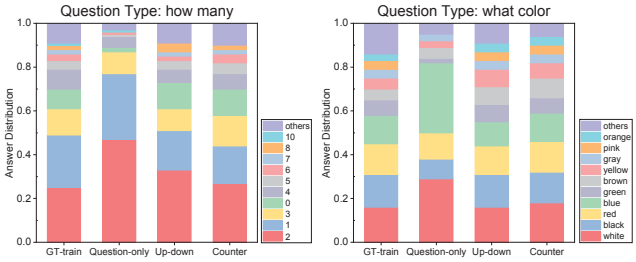

Figure 2: Answer distribution of two question types in the VQA 2.0 dataset.

answering *yes* for the question type *Do you see a* without reading the rest of the question or looking at the associated image results in a VQA accuracy of 87%. And 2) model deficiency. The predictive capability of the language over images from today's VQA models have been corroborated by ablation studies in [1]. Figure 1 and 2 also validate that these VQA methods suffer from the language prior problem.

Though many studies mention the language prior problem, nevertheless, little attention has been paid to develop an evaluation metric to quantitatively measure the extent of language prior effect numerically. Therefore, in this paper, we propose a metric to effectively measure the VQA models' language prior degree and provide full validation of this metric in Section 5. We hope this metric can facilitate the advancement of VQA models and other domains which also suffer from the language prior problem.

2.2 Definition

In this subsection, we will give a detailed definition and explanation of the proposed metric - language prior score (dubbed as LP score). We first list the main notations used in the metric.

Notations. Let $\overline{A}$ denote all the answer multiset[4] in the training set, and QT be the question type set. For a question type qt_j, $\overline{A}_j$ indicates the corresponding answer multiset, which is a subset of $\overline{A}$; A_j indicates the corresponding answer set, which contains the non-redundant elements in $\overline{A}_j$. And n_j^i is the number of answer a^i in $\overline{A}_j$. For example, let us assume there is only one question type *how many* - qt_j, and $\overline{A}$ is {0, 0, 1, 2, 2, 2, 3, 4, 4, 4}. Now $\overline{A}_j$ should be the same as $\overline{A}$, then A_j is {0, 1, 2, 3, 4}. If a^i is 4, then n_j^i should be 3.

[4]Allow for duplicate elements.

Answer Precision per Question Type. After evaluating the model in the validation set, we can compute the answer precision for each question type. We ignore the case that a predicated answer a^i has not been included in the current answer multiset A_j (i.e., $a^i \notin A_j$)[5]. Otherwise we compute P_j^i, which is the precision of the predicted answer a^i under the question type qt_j, is computed as:

$$P_j^i = \frac{TP_j^i}{TP_j^i + FP_j^i}, \tag{1}$$

where TP_j^i denotes the number of true positive answers, i.e., the predicted answer a_i is the same as the ground truth answer under the question type qt_j. And FP_j^i denotes the number of false positive answers, i.e., the predicted answer a_i is not consistent with the groundtruth answer under the question type qt_j. For example, if a testing question belongs to the question type qt_j and the predicted answer is a_i, and then $TP_j^i + 1$ if the groudtruth answer is a_i, otherwise $FP_j^i + 1$. Apparently, a larger P_j^i indicates that more questions of this type are correctly answered, and vice versa.

Language Prior Score. Let LP_j^i denote the LP score for the predict answer a^i under the question type qt_j. Formally, it is defined as:

$$LP_j^i = (1 - P_j^i) * \sigma(\frac{n_j^i}{|\overline{A}_j|}), \tag{2}$$

where $\sigma(\cdot)$ refers to a non-linear function (here the sigmoid function is adopted) and $|\overline{A}_j|$ is the size of multiset $\overline{A}_j$. $(1 - P_j^i)$ of Equation 2 represents the model deficiency when testing. In extreme cases, if a model performs best as oracle, the P_j^i should be near to 1. And accordingly, $(1 - P_j^i)$ should be near to 0. $\sigma(\frac{n_j^i}{|\overline{A}_j|})$ represents the proportion of the true answer a^i of a certain question type qt_j in the whole training set. The reason why we use $\sigma(\cdot)$ for smoothing this part is the proportion of different answers varies largely and we hope sparse answers can also contribute to this metric. We can see that a larger LP_j^i is obtained only when 1) the answers of more questions in the validation set (or testing set) are incorrectly predicted to be a^i; and 2) the true answers of more questions are a^i in the training set. In other words, if more answers of a question type in the training data are biased towards a^i, and more questions of this type are wrongly answered to a^i, (i.e., the language prior problem), the larger LP score will be obtained. Therefore, the measurement considers both the training dataset bias and the model deficiency - the two factors that cause the language prior problem as discussed. Finally, the LP score over the whole validation set can be computed as,

$$LP = \frac{1}{|\boldsymbol{QT}|} \sum_{j \in \boldsymbol{QT}} \frac{1}{|A_j|} \sum_{i \in \boldsymbol{A}_j} LP_j^i, \tag{3}$$

where $|\boldsymbol{QT}|$ is the size of the whole question type set, and $|\boldsymbol{A}_j|$ is the size of the answer set under the question type qt_j. We can easily conclude that $LP \in [0, 1]$ and the larger the LP score is, the more language prior is resulted in by the model.

[5]If most of the predicted answers do not belong to the answer set of the current question types, it is obviously that a very low accuracy will be obtained. In the experiments, only around 0.1% answers are ignored for all the baselines. Therefore, ignoring those answers ($a^i \notin A_j$) has negligible effects on the language prior measurement.

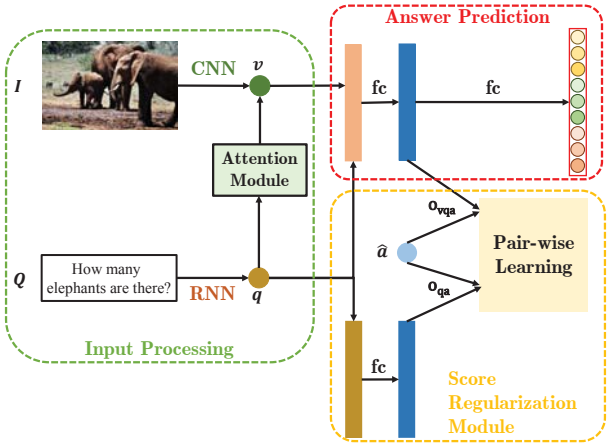

Figure 3: Demonstration of the proposed method for alleviating the language prior problem.

3 PROPOSED REGULARIZATION METHOD

3.1 Problem Formulation

The goal of VQA is to provide an accurate answer $\hat{a}$ to a given textual question Q upon an image or a video I. A general approach is to regard the VQA problem as a classification task:

$$\hat{a} = \arg\max_{a \in \Omega} p(a|Q, I; \Theta), \tag{4}$$

where Ω denotes the candidate answer set and Θ denotes the model parameters.

3.2 Background of VQA Models

As shown in Figure 3, the main framework is composed of three components: **Input Processing**, **Answer Prediction** and **Score Regularization**. The core of our proposed regularization method lies in the **Score Regularization** part, which will be elaborated in Section 3.3.

3.2.1 Input Processing. There are mainly three parts in the **Input Processing** component: Image Processing, Question Processing and Attention Module.

Image Processing. The predominant VQA models leverage pre-trained Convolutional Neural Network (CNN) frameworks (e.g., VGG [35] or ResNet [15]) to extract image features $\boldsymbol{v}$,

$$\boldsymbol{v} = \text{CNN}(I). \tag{5}$$

As most of the state-of-the-art VQA models use the attention mechanism, it is worth mentioning that there are two kinds of image feature extraction. The first one is splitting the image into equal-size regions (e.g., 14×14) and then extracting image features from each region. This will result in a tensor size of $k \times 14 \times 14$ (k is the feature size of each equal-sized image region). The other one is using the region proposal techniques (e.g., Faster R-CNN [33]) to extract image features for salient image regions, leading to a tensor size of $k \times n$ (k is the feature size of each image region and n is the proposed salient image region number).

Question Processing. Recurrent Neural Network (RNN, e.g., LSTM [16]) is often used in VQA models to extract question features

$\boldsymbol{q}$,

$$\boldsymbol{q} = \text{RNN}(Q). \tag{6}$$

More concretely, for a question sentence consisting of T words, its words are fed into the RNN network one by one to obtain their hidden features $\boldsymbol{h}$. Usually the last hidden feature $\boldsymbol{h}_T$ or all the hidden features (when the attention mechanism is used on each question word) are used to represent this question.

Attention Module. After the processing of images and questions, a series of image region features $\{\boldsymbol{v}_0, \boldsymbol{v}_1, ...\}$ (14×14 or n) and one question feature $\boldsymbol{h}_T$ are obtained. Then the VQA models use the question feature to attend on each image region through a multi-layer perceptron (MLP) network or CNN to obtain the attention weights for each image region $\boldsymbol{v}_j$:

$$g_j = \text{ATT}(\boldsymbol{h}_T, \boldsymbol{v}_j), \tag{7}$$

where g_j is normalized through a softmax function. Finally the attended image feature is given by:

$$\widetilde{\boldsymbol{v}} = \sum \boldsymbol{v}_j * g_j. \tag{8}$$

3.2.2 Answer Prediction. With the attended image feature $\widetilde{\boldsymbol{v}}$ and question feature $\boldsymbol{h}$, typically, a fusion function (e.g., element-wise addition, element-wise multiplication or concatenation) can be adopted to fuse the question and attended image features. After merging the question and the image features, the VQA models frequently use several linear layers with non-linear activation functions (e.g., ReLU) to make full interactions. Finally the models predict a normalized fixed-length vector and each dimension corresponds to one fixed answer,

$$p_{answer} = \text{softmax}(\widetilde{\boldsymbol{v}}, \boldsymbol{h}). \tag{9}$$

The models can be trained by minimizing the log-likelihood loss function, such as:

$$\mathcal{L}_{answer} = -a_{gt} * \log p_{answer}, \tag{10}$$

where a_{gt} is the distribution of the ground truth answers.

3.3 Proposed Regularization Method

As shown in Figure 1 and 2, there are some frequent patterns between question types and answers. And these patterns are easily captured by the VQA models. As a result, the model will directly give answers based only upon the text questions without referring to the image contents. The VQA then degenerates to a QA problem to some extent.

Based on the above discussion, we would like the VQA model to better reason the image contents upon the corresponding questions before predicting the answers, instead of relying on the discovered question-answer patterns to make prediction. To achieve this, we design a score regularization module, which adopts a pair-wise learning scheme to make the predicted score obtained from the <image, question, answer> higher than the predicted score obtained from the <question, answer>.

As shown in Figure 3, there are two stream inputs to the score regularization module: $\boldsymbol{o}_{vqa}$ and $\boldsymbol{o}_{qa}$. The former one represents the integration representation of image, question and answer, while the latter one denotes the integration of the question and answer. $\widehat{\boldsymbol{a}}$ is the pre-trained word embedding of true answers and it can be fused with other elements (e.g., <image, question> feature or only question feature) to obtain $\boldsymbol{o}_{vqa}$ and $\boldsymbol{o}_{qa}$. The fusion method includes element-wise addition, multiplication and concatenation. More analysis can be found in Section 5. After this step, the fused features of <image, question, answer> and <question, answer> are used to predict s_{vqa} and s_{qa},

$$s_{vqa} = \text{MLP}(\boldsymbol{o}_{vqa}), \tag{11}$$

$$s_{qa} = \text{MLP}(\boldsymbol{o}_{qa}), \tag{12}$$

where the MLP is leveraged to implement our score regularization module. In order to achieve that questions with images are better than merely questions for answer prediction, a pair-wise learning method is adopted,

$$\mathcal{L}_{score} = \max(0, s_{vqa} - s_{qa} + \gamma), \tag{13}$$

where γ is a relatively small margin. In this way, the backbone models are forced to consider image content for answering questions, instead of only basing on the frequent patterns between question words and answers.

With the proposed regularization method, the final loss function of the backbone VQA model is a combination of both the answer prediction loss and the score restriction loss,

$$\mathcal{L} = \mathcal{L}_{answer} + \beta * \mathcal{L}_{score}, \tag{14}$$

where β is a hyper-parameter balancing these two loss functions. This enables us to train the backbone model with our regularization method in an end-to-end multi-task learning scheme. The default optimization method of the backbone models is kept unchanged to optimize the final loss function.

In Section 5, we will show that different from the methods in [2, 32] which deteriorate the backbone models' performance, our proposed regularization method can boost the backbone models' performance as well as alleviate the language prior problem.

4 EXPERIMENTAL SETUP

We conducted extensive experiments on two datasets to thoroughly justify the effectiveness of our proposed regularization method as well as the feasibility of our proposed evaluation metric. In particular, our experiments mainly answer the following research questions:

- **RQ1**: Can our proposed regularization method outperform the state-of-the-art VQA methods?
- **RQ2**: Is the proposed evaluation metric (i.e., LP score) feasible for measuring the extent language prior effect?
- **RQ3**: Is the proposed regularization helpful for boosting the answering accuracy as well as alleviating the language prior problem?
- **RQ4**: Can backbone models with our proposed regularization method better understand images than those without it?

4.1 Datasets

We tested our proposed method on VQA 1.0 [5] and VQA 2.0 [13] datasets. Both datasets consist of real images from MSCOCO [22] and abstract cartoon scenes. For each image, three different questions are given by Amazon Mechanical Turk (AMT) workers, with ten answers per question. The answers are divided into three categories: *yes/no*, *number* and *other*. Besides, both datasets are split

into training, validation and testing (or test-std) splits. The ground truth answers are only available for the first two splits.

4.2 Evaluation Metric

Accuracy. We adopt the standard accuracy metric for evaluation [5, 13]. Given an image and a corresponding question, for a predicted answer a, the accuracy is computed as:

$$Acc_a = \min(1, \frac{\#\text{humans that provide that answer } a}{3}). \quad (15)$$

Note that each question is answered by ten participants, this metric takes the disagreement in human answers into consideration. The reported results are the averaged accuracy over all questions.

LP Score. As the ground truth answers are not published for the testing set data, we only compute the LP score on the validation set. The computation of LP score is elaborated in Equation 3.

4.3 Compared Baselines

We added our regularization method into the following three state-of-the-art baselines. The first one is from the VQA 1.0 dataset, while the last two are from the VQA 2.0 one.

- **Strong-baseline** [20] leverages two stacked ConvNets to obtain the final attention weights for each equal-sized image region. After that it fuses the attentive image feature with the question feature through the vector addition approach.
- **Up-Down** [3] utilizes a *top-down* mechanism to determine attention weights from *bottom-up* image features (object level and other salient image regions).
- **Counter** [43] is an upgraded version of the Strong-baseline, introducing a counting module to enable robust counting from object proposals.

4.4 Implementation Details

We kept most of the setting of backbone models unchanged, including batch size, optimization method, number of non-linear layers. For all the three backbone models, the trade-off parameter β was tuned in the range [0.001, 0.01, 0.1, 1, 10, 100]; the margin γ was tuned in [0.0, 1.0] with a step size 0.1; and the number of MLP in our score regularization module is fixed to 2; a dropout layer is added between the two layers with a dropout rate 0.5.

5 EXPERIMENTAL RESULTS

5.1 Performance of Accuracy Comparison (RQ1)

Table 1 and Table 2 summarize the accuracy comparison results between our proposed score regularization method with baselines from two groups: traditional VQA models (e.g., Question-only [5], NMN [4], DCN [27]) and VQA models designed to alleviate the language prior problem (i.e., SAN-GVQA [2] and SAN+Q-Adv+DoE [32]). The answers are divided into three categories: *Y/N*, *Num.* and *Other*. And the split *All* represents the overall accuracy. Besides, Strong-baseline-SR, Up-down-SR and Counter-SR[6] are backbone models Strong-baseline, Up-down and Counter with our regularization method, respectively.

[6]The fusion method between <question, image> and answer features is element-wise multiplication. More analysis can be found in Section 5.3.

For VQA models from the second group, the observation from Table 1 and Table 2 is that these VQA models all deteriorate the corresponding backbone models's performance. For example, the overall accuracy deterioration on the validation set of SAN-GVQA over backbone model SAN is 5.88% on Table 1, and Up-down+Q-Adv+DoE over backbone model Up-down is 0.45% on Table 2. Compared with the models in this group, the final models with our score regularization module (i.e., Strong-baseline-SR, Up-down-SR and Counter-SR) can outperform these baselines with a large margin on both the VQA 1.0 and 2.0 datasets. For example, on the VQA 1.0 dataset, the absolute improvement of Strong-baseline-SR over SAN+Q-Adv+DoE on Validation All is 9.15%; on the VQA 2.0 dataset, Counter-SR over Up-down+Q-Adv+DoE is 2.74%.

Note that the methods in the second group are carefully designed to alleviate the language prior problem, however, there is no evidence in their reports to validate their effects. That means although those methods can indeed alleviate the language prior problem, we still do not know to what extent they can achieve this[7]. Next, we analyze the feasibility of our proposed metric LP score and use it to measure the extent of language prior effect of the model with and without our regularization model.

5.2 Feasibility of the Proposed Metric (RQ2)

5.2.1 Case Analysis. We chose two question types *how many* and *what animal* to analyze the feasibility of the proposed LP score metric. The answer distribution in the training set of question type *what animal* is much more uniform than that of *how many*. Note that the *Question-only* method answers the questions merely based on the question features without reasoning images which will certainly result in the language prior problem. From Table 3 we could see that the LP scores of the state-of-the-art approaches are lower than that of the *Question-only* ones, which is consistent with that the language prior problem affects smaller on the former ones than the latter ones. Moreover, for question type *how many*, the LP scores of state-of-the-art methods and the regularized methods[8] on both the VQA 1.0 and VQA 2.0 datasets are just slightly better than the *Question-only* one, respectively. In contrast, for the more uniform answer distribution of question type *what animal*, there is a large margin between the state-of-the-art models and the Question-only model. This indicates that for the state-of-the-art VQA models, the language prior effect of these question types with less uniform answer distributions is higher than those with more uniform answer distributions. Based on the analysis, we can deduce that our proposed metric is capable of measuring the language prior effect.

5.2.2 Overall Analysis. Figure 4 shows how the LP score and accuracy changes with the increase of training epochs. The red line shows the accuracy of one typical baseline model, while the other three on each sub-figure show the LP scores of three baselines. At the very beginning, the VQA models answer questions mainly based on the learned language prior, which results in a higher LP score in the first few training epochs. With more iterations on the

[7]The codes of SAN-GVQA [2], SAN+Q-Adv+DoE and Up-down+Q-Adv+DoE [32] are not available, and it is hard for us to replicate their results due to the complicate parameter tuning in the SAN model, therefore, we cannot get the LP scores for those models.

[8]The fusion method is element-wise multiplication.

Table 1: Performance of accuracy comparisons between the proposed method and baselines over the VQA 1.0 dataset. The best performance in current splits is highlighted in bold.

Method	Validation				Test-dev				Test-std			
	Y/N	Num.	Other	All	Y/N	Num.	Other	All	Y/N	Num.	Other	All
Question-only [5]	77.86	30.24	27.61	46.75	78.20	35.68	26.59	48.76	78.12	34.94	26.99	48.89
HieCoAttn [24]	79.6	35.0	45.7	57.0	79.7	38.7	51.7	61.8	-	-	-	62.1
SAN [40]	78.6	41.8	46.4	57.6	79.30	36.60	46.10	58.70	79.11	36.41	46.42	58.85
NMN [4]	80.44	34.03	40.66	54.72	81.2	38.0	44.0	58.6	-	-	-	58.7
Strong-baseline [20]	82.31	35.77	51.67	61.10	82.2	**39.1**	55.2	64.5	82.0	39.1	**55.2**	64.6
Ask-me-anything [38]	-	-	-	55.96	81.01	38.42	45.23	59.17	81.07	37.12	45.83	59.44
SMem [39]	-	-	-	-	80.87	37.32	43.12	57.99	80.80	37.53	43.48	58.24
SAN-GVQA [2]	76.90	-	-	51.12	-	-	-	-	-	-	-	-
SAN+Q-Adv+DoE [32]	71.06	32.59	42.91	52.15	-	-	-	-	-	-	-	-
Ours (Strong-baseline-SR)	**82.51**	**35.80**	**51.68**	**61.30**	**83.10**	39.05	**55.9**	**65.15**	**83.2**	**39.14**	55.12	**65.28**

Table 2: Performance of accuracy comparisons between the proposed method and baselines over the VQA 2.0 dataset. The best performance in current splits is highlighted in bold.

Method	Validation				Test-dev				Test-std			
	Y/N	Num.	Other	All	Y/N	Num.	Other	All	Y/N	Num.	Other	All
Question-only [5, 13]	67.90	30.48	26.49	42.57	67.17	31.41	27.36	44.22	67.01	31.55	27.37	44.26
Up-Down [3]	80.3	42.8	55.8	63.2	81.82	44.21	56.05	65.32	82.20	43.90	56.26	65.67
DCN [27]	-	-	-	62.94	83.51	46.61	57.26	66.87	83.85	47.19	56.95	67.04
DA-NTN [6]	**83.09**	44.88	55.71	64.58	**84.29**	47.14	57.92	67.56	**84.60**	47.13	58.20	67.94
Counter [43]	81.81	**49.22**	**56.96**	65.28	83.14	51.62	**58.97**	68.09	83.56	51.39	59.11	68.41
SAN-GVQA [2]	72.03	-	-	48.24	-	-	-	-	-	-	-	-
SAN+Q-Adv+DoE [32]	69.98	39.33	47.63	52.31	-	-	-	-	-	-	-	-
Up-down+Q-Adv+DoE [32]	79.84	42.35	55.16	62.75	-	-	-	-	-	-	-	-
Ours (Up-down-SR)	80.91	43.2	55.03	63.68	81.86	44.12	56.20	66.35	82.98	43.97	56.96	66.58
Ours (Counter-SR)	82.48	49.02	56.88	**65.29**	83.67	**51.63**	58.57	**68.12**	83.87	**51.60**	**59.16**	**68.43**

Methods	How many		What animal	
	VQA 1.0	VQA 2.0	VQA 1.0	VQA 2.0
Question-only	50.37	49.80	54.49	53.55
Strong-baseline	49.89	-	33.84	-
Strong-baseline-SR	49.81	-	33.85	-
Up-down	-	48.01	-	33.81
Up-down-SR	-	47.90	-	33.69
Counter	-	46.30	-	31.09
Counter-SR	-	46.26	-	31.04

Table 3: LP scores of four baselines and three regularized methods on two typical question types.

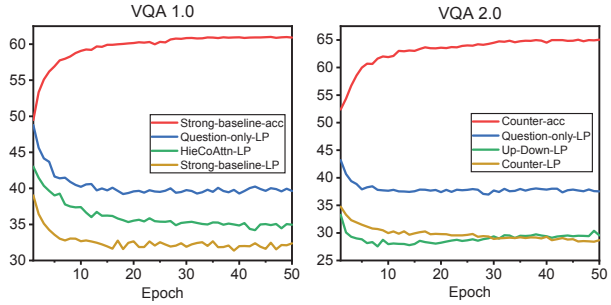

Figure 4: The convergence illustration of LP scores and accuracy over several baselines.

training set, the LP scores begins to drop and the accuracy begins to rise. This denotes that the current VQA models learn to weaken the influence of language prior problem so that the overall accuracy can obtain improvement. If more language prior can be alleviated or overcome, there should be accuracy improvement instead of accuracy degradation. Therefore, it is promising to study and alleviate the language prior problem.

5.3 Effect of the Proposed Method (RQ3)

Table 4 and Table 5 show the influence of our score regularization module over three baselines on the VQA 1.0 and VQA 2.0 datasets, respectively. The second and the last column report the accuracy metric, while the third column reports the LP score metric. To be more specific, *mul* represents that the element-wise multiplication is used for the feature fusion of <image, question, answer> or <question, answer> in score regularization module Equation 11, *add* denotes element-wise addition, while *con* represents concatenation.

Table 4: Influence of the score regularization module on the VQA 1.0 dataset.

Method	Valid-All	Valid-LP	Test-dev-All
Strong-baseline	61.10	31.54	64.5
Strong-baseline-SR (mul)	**61.30**	31.36	65.11
Strong-baseline-SR (add)	61.19	**31.33**	64.88
Strong-baseline-SR (con)	61.13	31.38	**65.15**

Table 5: Influence of the score regularization module on the VQA 2.0 dataset.

Method	Valid-All	Valid-LP	Test-dev-All
Up-down	63.20	29.71	65.32
Up-down-SR (mul)	**63.68**	29.44	66.35
Up-down-SR (add)	63.53	**29.43**	66.25
Up-down-SR (con)	63.55	29.50	**66.46**
Counter	65.28	29.74	68.09
Counter-SR (mul)	**65.29**	**29.67**	**68.12**
Counter-SR (add)	65.03	29.84	67.88
Counter-SR (con)	65.01	29.88	67.86

We could observe that different from SAN-GVQA [2] and SAN+Q-Adv+DoE [32] deteriorating the backbone models, our proposed regularization method can achieve comparative performance or boost the backbone accuracy performance (e.g., Up-down-SR (mul) over Up-down is 0.48%). Moreover, the LP score of the proposed regularization method can also outperform the corresponding backbone models. This demonstrates the advantage of our regularization method over the existing ones that we can alleviate the language prior problem as well as boost the backbone models' performance.

5.4 Visualization of Attention Kernels (RQ4)

As the attention module becomes an indispensable part of current VQA models, we visualized some examples of the attention kernels from these backbone models with and without our score regularization module. We mainly listed the questions which belong to the question type *how many* and *what color*, other more uniform question type samples are also analyzed in Figure 5, e.g., *what is*. There are three rows of six examples, each row illustrates two samples from one backbone model with and without regularization, where the backbone models without regularization predicted incorrectly and backbone models with regularization predicted correctly. From the figure, we can see the failure cases of VQA methods without regularization can be grouped into two categories: 1) attending to wrong regions and predicting answers incorrectly and 2) attending to correct regions but predicting answers wrongly.

Both samples from the second row belong to the first category. For instance, the first example is about the color of the countertop tiles, and the backbone model Up-down without regularization focuses on the closet and the white tile while the true region should be the tile on top of the countertop. As illustrated in Figure 1, the number of the wrong answer of *white* is much larger than that of the true answer *blue* in the VQA 1.0 training set, which leads to the language prior problem here. The second example from backbone model Up-down shares the same problem. In contrast, examples falling into the second category attended to correct image regions but predicting answers incorrectly. For instance, the second example from Strong-baseline is similar to the image classification task, where the true answer should be a bird instead of a cat. And the second example of backbone model Counter belongs to *how many* question type. Since the number of the answer *1* is much more than that of the answer *0* under this question type in the VQA 2.0 dataset, this example also testified that the language prior learned by the model Counter causes a wrong prediction and it can be corrected by our regularization method (as shown in the result of Counter-SR).

6 RELATED WORK

6.1 Visual Question Answering

Traditional text-based QA [8, 9, 30] has been long recognized as a challenging information retrieval task. Derived from it, other QA systems like community QA (CQA) [29], multimedia QA [28]and visual QA (VQA) [5, 13, 41] have attracted researchers' interest in recent years. We mainly recap the related studies of VQA in this subsection.

VQA has witnessed a renewed excitement in multi-discipline AI research problems due to the development of CV, NLP and IR. Generally speaking, the existing VQA methods can be classified into four categories [37]: *Joint Embedding*, *Attention Mechanism-based*, *Compositional* and *Knowledge Base-enhanced*. However, the language prior problem is observed across the existing VQA models [5, 38, 40]. It is impossible to distinguish an answer arising because of image reasoning and one selected because it occurs frequently in the training set. In the view of amending biased datasets, researcher in [42] added <image, question, answer> triplets by compositing another visually similar image but with an opposite answer to a binary question for VQA 1.0 abstract scenes. Similarly, authors in [13] added triplets based on all varieties of questions for VQA 1.0 real images. Instead of supplementing biased datasets, authors in [17] designed a diagnostic 3D shape to balance answer distribution for each question type from scratch. Different from the above ones, methods in [2, 32] aim to force VQA models to better understand the images. The authors built their models on SAN [40] with restrictions to prevent the model from exploiting language correlations in the training data.

It is worth emphasizing that the previous methods solve the language prior problem either by introducing other bias into existing datasets [13, 42] or degrading performance over the backbone models [2, 32]. In this paper, we proposed a regularization method for several publicly released state-of-the-art attentional VQA models. In addition to the capability of alleviating the language prior problem, better accuracy is observed for the VQA models with our regularization module than the corresponding ones without.

6.2 Deep Multimodal Fusion

In this work, we fuse the multi-modality features - image, question, and answer in the proposed regularization module. Here we briefly review related works in this direction. There is a large amount of studies on integrating multimodal data sources in deep neural networks, including recommendation [7, 11, 44], multimodal retrieval [14, 23], and user profiling [12], image captioning [10, 19]. The flexibility of deep architecture advances the implementation of

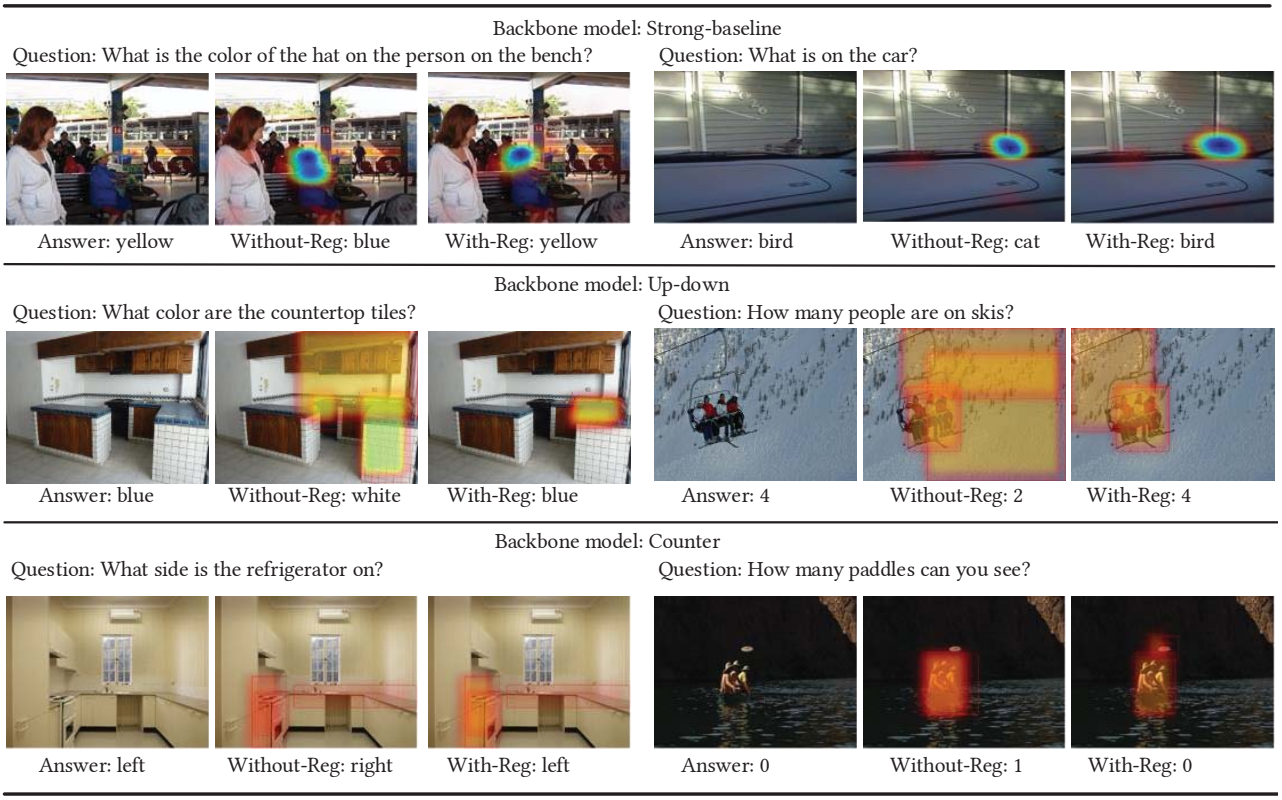

Figure 5: Visualization of three backbone models with and without the proposed regularization method.

multimodal fusion either as feature-level fusion or decision-level fusion [31].

Methods in the feature-level fusion group transform the raw inputs from multiple paths into separate intermediate representations, followed by a shared representation layer to merge them. For instance, Chen et al. [7] proposed a two-level attention mechanism to fuse the features from component-level and item-level for multimedia recommendation. Farnadi et al. [12] utilized a shared representation between different modalities to arrive at more accurate user profiles. Some efforts [10, 19] in image captioning merged previous word representations and image features to produce the next word.

By contrast, decision-level fusion refers to the aggregation of decisions from multiple classifiers, each trained on separate modalities. These fusion rules could be max-fusion, averaged-fusion, Bayes' rule based, or even learned using a meta-classifier [31]. For example, the work in [34] presents a two-stream CNN (i.e., Spatial stream ConvNet and Temporal stream Convnet) and then combines them with a *class score fusion* approach for action recognition in videos. In order to achieve simultaneous gesture segmentation and recognition, authors in [36] integrated the emission probabilities estimated from different inputs (i.e, skeleton joint information, depth and RGB images) as a simple linear combination.

7 CONCLUSION

The language prior problem severely hinders the advancement of VQA. In this paper, we target this problem and make contributions from two perspectives. Firstly, we propose an evaluation metric called LP score to measure the extent of language prior effect. The evaluation metric can quantitatively measure the extent of language prior effect of different VQA models and thus can facilitate the development of related techniques. Secondly, we design a score regularization module, which is versatile to be integrated into various current VQA models. The proposed regularizer can effectively make the VQA models better reason images upon questions before result prediction, and thus can alleviate the language prior problem as well as improve the answer accuracy. Extensive experiments have been conducted on two widely used VQA datasets to validate the feasibility of the proposed metric and the effectiveness of the designed regularization method. We hope this metric can be used to compare the VQA models on alleviating the language prior problem in the future. Besides, we would like to further extend our regularizer based on the non-uniform granularity of different question types and explore its effectiveness on more diversified VQA models.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.: 61772310, No.:61702300, No.:61702302, No.: 61802231, and No.: U1836216; the Project of Thousand Youth Talents 2016; the Tencent AI Lab Rhino-Bird Joint Research Program, No.:JR201805; and the National Research Foundation, Prime Minister's Office, Singapore under its Strategic Capability Research Centres Funding Initiative.

REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *EMNLP*. ACL, 1955–1960.

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*. IEEE, 4971–4980.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*. IEEE, 6077–6086.

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *CVPR*. IEEE, 39–48.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*. IEEE, 2425–2433.

[6] Yalong Bai, Jianlong Fu, Tiejun Zhao, and Tao Mei. 2018. Deep attention neural tensor network for visual question answering. In *ECCV*. Springer, 21–37.

[7] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*. ACM, 335–344.

[8] Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. 2019. Driven answer generation for product-related questions in e-commerce. In *WSDM*. ACM, 411–419.

[9] Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. 2019. Driven answer generation for product-related questions in E-commerce. In *WSDM*. ACM, 411–419.

[10] Yong Cheng, Fei Huang, Lian Zhou, Cheng Jin, Yuejie Zhang, and Tao Zhang. 2017. A hierarchical multimodal attention-based neural network for image captioning. In *SIGIR*. ACM, 889–892.

[11] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. *TOIS* 37, 2 (2019), 16.

[12] Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User profiling through deep multimodal fusion. In *WSDM*. ACM, 171–179.

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*. IEEE, 6325–6334.

[14] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Xin-Shun Xu, and Mohan Kankanhalli. 2018. Multi-modal preference modeling for product search. In *MM*. ACM, 1865–1873.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. IEEE, 770–778.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*. IEEE, 1988–1997.

[18] Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *CVIU* 163 (2017), 3–20.

[19] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*. IEEE, 3128–3137.

[20] Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. In *arXiv preprint arXiv:1704.03162*.

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata andJoshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* 123, 1 (2017), 32–73.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer, 740–755.

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Philip S Yu. 2016. Composite correlation quantization for efficient multimodal retrieval. In *SIGIR*. ACM, 579–588.

[24] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*. 289–297.

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

[26] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*. IEEE, 1–9.

[27] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*. IEEE, 6087–6096.

[28] Liqiang Nie, Meng Wang, Zhengjun Zha, Guangda Li, and Tat-Seng Chua. 2011. Multimedia answering: enriching text QA with media information. In *SIGIR*. ACM, 695–704.

[29] Adi Omari, David Carmel, Oleg Rokhlenko, and Idan Szpektor. 2016. Novelty based ranking of human answers for community questions. In *SIGIR*. ACM, 215–224.

[30] Marius A Pasca and Sandra M Harabagiu. 2001. High performance question/answering. In *SIGIR*. ACM, 366–374.

[31] Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *Signal Processing* 34, 6 (2017), 96–108.

[32] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *NIPS*. 1546–1556.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*. 91–99.

[34] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*. 568–576.

[35] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

[36] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. 2016. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *TPAMI* 38, 8 (2016), 1583–1597.

[37] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *CVIU* 163 (2017), 21–40.

[38] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*. IEEE, 4622–4630.

[39] Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*. Springer, 451–466.

[40] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *CVPR*. IEEE, 21–29.

[41] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*. IEEE, 1839–1848.

[42] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *CVPR*. IEEE, 5014–5022.

[43] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. 2018. Learning to count objects in natural images for visual question answering. In *ICLR*.

[44] Zhou Zhao, Qifan Yang, Hanqing Lu, Min Yang, Jun Xiao, Fei Wu, and Yueting Zhuang. 2017. Learning max-margin geoSocial multimedia network representations for point-of-interest suggestion. In *SIGIR*. ACM, 833–836.

[45] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *CVPR*. IEEE, 4995–5004.