Overcoming Language Priors with Self-supervised Learning for Visual Question Answering

Xi Zhu[1,2], **Zhendong Mao**[3*], **Chunxiao Liu**[1,2], **Peng Zhang**[1], **Bin Wang**[4] and **Yongdong Zhang**[3]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]University of Science and Technology of China, Hefei, China
[4]Xiaomi AI Lab, Xiaomi Inc., Beijing, China
{zhuxi, liuchunxiao, pengzhang, wangbin}@iie.ac.cn {zdmao,zhyd73}@ustc.edu.cn

Abstract

Most Visual Question Answering (VQA) models suffer from the language prior problem, which is caused by inherent data biases. Specifically, VQA models tend to answer questions (*e.g.*, what color is the banana?) based on the high-frequency answers (*e.g.*, yellow) ignoring image contents. Existing approaches tackle this problem by creating delicate models or introducing additional visual annotations to reduce question dependency and strengthen image dependency. However, they are still subject to the language prior problem since the data biases have not been fundamentally addressed. In this paper, we introduce a self-supervised learning framework to solve this problem. Concretely, we first automatically generate labeled data to balance the biased data, and then propose a self-supervised auxiliary task to utilize the balanced data to assist the VQA model to overcome language priors. Our method can compensate for the data biases by generating balanced data without introducing external annotations. Experimental results show that our method achieves state-of-the-art performance, improving the overall accuracy from 49.50% to 57.59% on the most commonly used benchmark VQA-CP v2. In other words, we can increase the performance of annotation-based methods by 16% without using external annotations. Our code is available on GitHub[1].

1 Introduction

Visual Question Answering (VQA) has attracted increasing attention as an AI-complete task, whose goal is to automatically answer natural language questions according to images. The paradigm of VQA [Antol *et al.*, 2015; Yang *et al.*, 2016; Anderson *et al.*, 2018; Kim *et al.*, 2018] is to first project the image and the question into a common feature space, and then fuse them as a joint vector to make prediction. Recently, some researchers [Agrawal *et al.*, 2018; Goyal *et al.*, 2017] have demonstrated that most existing VQA models suffer from the language prior problem and tend to ignore the image contents. For example, the question "*what color is the grass?*" can be answered by "*green*" generally, no matter what images are given, since most corresponding answers are "*green*" in the dataset. As a result, the VQA model memorizing the language priors will perform poorly on out-of-domain datasets.

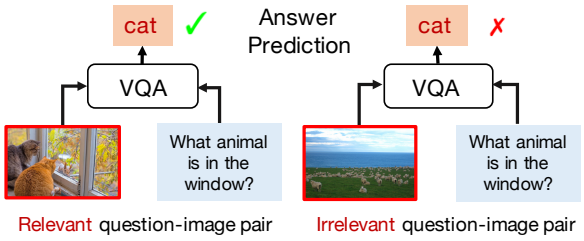

Figure 1: A question can only be answered based on relevant images.

To alleviate the influence of language priors, existing approaches focus on reducing question dependency and increasing image dependency, and they can be roughly categorized as non-annotation-based methods and annotation-based methods. Non-annotation-based methods often involve delicate models and complex learning strategies. Ramakrishnan *et al.*[2018] proposed an adversarial learning method to overcome the language priors by minimizing the performance of the question-only branch. Cadene *et al.*[2019] reduced the influence of the most-biased instances and increased the impact of the less-biased instances by dynamically adjusting their weights. Different from non-annotation-based methods, annotation-based methods try to increase image dependency directly by introducing external visual supervision. Selvaraju *et al.*[2019] used human-attention maps to ensure the alignment between model-attention and human-attention. Wu and Mooney[2019] maintained the consistency of correct answers and influential objects annotated by human explanations. Typically, annotation-based methods can achieve better performance than non-annotation-based methods, since they can better understand image contents with the guidance of visual supervision. Nonetheless, these methods require large-scale visual annotations, which are not easily accessible.

However, the inherent data bias problem has not been fun-

*Corresponding Author
[1]https://github.com/CrossmodalGroup/SSL-VQA

damentally addressed, and the above methods just weaken the adverse effect to some extent and hence yield unsatisfactory performance. The inherent data biases will inevitably force the VQA model to select the high-frequency answers and eventually arouse the language prior problem. Therefore, it is of crucial importance to alleviate the inherent data biases, i.e., transforming biased data to balanced data without introducing external annotations.

To this end, we propose a self-supervised learning framework for VQA to automatically balance the biased data to overcome language prior problem. Our method is motivated by an interesting and intuitive finding. As shown in Figure 1, a question can only be answered when the given image containing the key information for answering the question. Such a question-image pair can be defined as a relevant pair. Based on this observation, it is necessary to estimate whether the given question and image are relevant or not before answering the question. For that purpose, we introduce an auxiliary task named *question-image correlation estimation* to estimate the relevance between questions and images. Specifically, we first automatically generate a set of balanced question-image pairs with binary labels (relevant and irrelevant), which are then consumed by the self-supervised auxiliary task to assist the VQA model to overcome language priors. We incorporate the auxiliary task into the VQA model by feeding the relevant and irrelevant pairs. When fed a relevant question-image pair, the model is encouraged to predict the correct answer with a high confidence score, where the confidence score is the probability of the question-image pair being relevant. On the contrary, the model is pushed to predict the correct answer with a low confidence score when the input pair is irrelevant. By optimizing these two objectives simultaneously, we can achieve a balance between answering questions and overcoming language priors. Therefore, our method can also be interpreted as an underlying multi-task learning framework.

To summarize, our contributions are as follows:

- We introduce a self-supervised framework by transforming the inherently biased data into balanced data automatically, and propose an auxiliary task to exploit such balanced data to overcome language priors fundamentally. To the best of our knowledge, this is the first work to use self-supervised learning in this task.
- Extensive experiments are conducted on the popular benchmark VQA-CP v2. Experimental results show that our approach without using external annotations can significantly outperform the state-of-the-art methods, including the models using human supervision. We increase the overall accuracy from 49.50% to 57.59%.

2 Related Works

2.1 Visual Question Answering

Visual Question Answering (VQA) aims to answer questions according to images, which involves technologies from both natural language processing and computer vision communities [Liu *et al.*, 2016; Parkhi *et al.*, 2015; Conneau *et al.*, 2016; Liu *et al.*, 2018]. Existing VQA approaches can be coarsely classified into four categories: 1) Joint Embedding approaches [Antol *et al.*, 2015] first project images and questions into a common feature space, and then combine them to predict answers by a classifier. 2) Attention-based methods [Anderson *et al.*, 2018] mainly focus on learning the interactions between the question words and image regions, making the answering process to be more interpretable. 3) Compositional models [Andreas *et al.*, 2016] leverage the compositional structure of questions to assembling modules that operate in the space of attention. 4) Knowledge-based approaches [Wu *et al.*, 2016] are proposed to answer common sense questions by exploiting external knowledge.

However, existing models tend to memorize the language priors during training without considering image information. Such models may achieve impressive results on the test set sharing the same distribution with the training set, but often performs poorly on out-of-domain test sets.

2.2 Overcoming Language Priors in VQA

Existing approaches in overcoming language priors can be roughly categorized as non-annotation-based methods and annotation-based methods. Non-annotation-based methods focus on creating delicate models to directly reduce the question dependency, while the annotation-based methods concentrate on strengthening the visual grounding by introducing additional human visual supervision.

For the non-annotation-based methods, Agrawal *et al.* [2018] proposed a hand-designed VQA framework, which explicitly disentangled the visual recognition from answer space prediction for different question types. Similarly, Jing *et al.*[2020] also decoupled the concept discovery and the question answering. Apart from shrinking the answer space, Ramakrishnan *et al.*[2018] proposed an adversarial learning strategy by minimizing the performance of the question-only branch. Guo *et al.*[2019] adopted a pair-wise ranking schema, forcing the question-only branch to make worse predictions than the base model did. Cadene *et al.*[2019] dynamically adjusted the weights of training instances via their prior masks learned by a question-only branch, reducing the influence of the most-biased instances and increasing the impact of the less-biased instances. Yi *et al.*[2018] proposed a neural-symbolic model incorporating the symbolic program executor into DNN for visual reasoning, which is distinct from the above models and can also solve the bias problem. [Mao *et al.*, 2019] combined the neural-symbolic model with curriculum concept learning, making it more generalizable.

Beyond that, annotation-based methods are shown to be effective by highlighting the important visual regions under the guidance of external visual supervision. HINT [Selvaraju *et al.*, 2019] increased the image dependency by optimizing the alignment between human-attention maps and gradient-based visual importance. SCR [Wu and Mooney, 2019] also emphasized the correspondences between correct answers and influential objects annotated by human textual explanations. However, these models are heavily dependent on human supervision, which is not always accessible.

Different from these methods, our self-supervised approach does not need to construct complex architectures or introduce external supervision. We first balance the original biased data automatically, and then overcome language priors

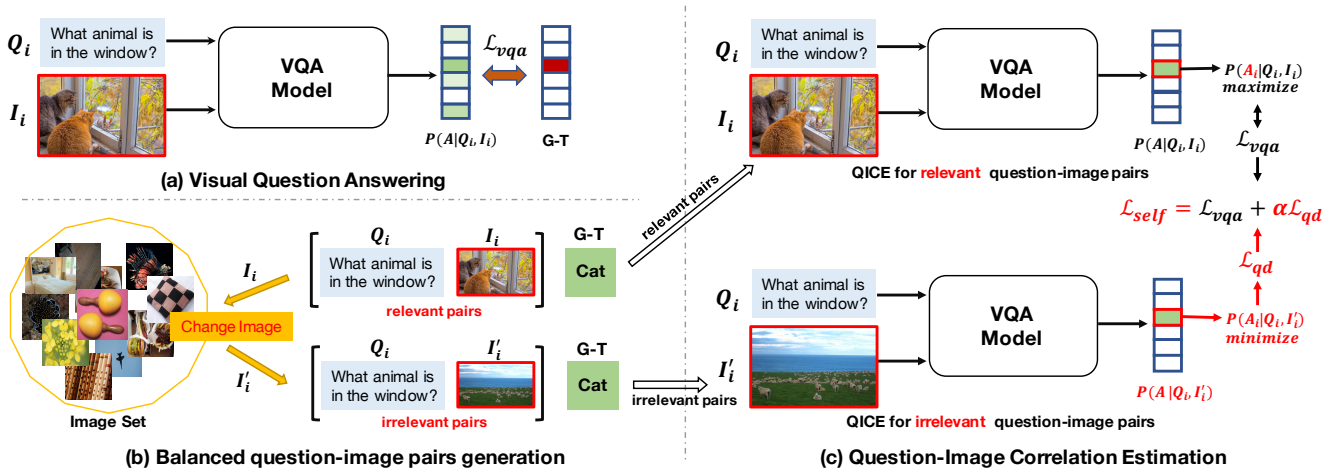

Figure 2: The framework of our self-supervised approach. The base VQA model is depicted in part (a), which aims to answer a question according to an image. Part (b) displays how we automatically generate balanced question-image pairs. To be more clear, part (c) shows how the question-image correlation estimation works for relevant and irrelevant pairs separately. G-T denotes the ground-truth.

based on these balanced data in a self-supervised manner.

2.3 Self-supervised Learning

Self-supervised learning constructs some supervisory signals automatically computed from the input data, and efficiently exploits the input itself to learn the high-level representation for some down-stream tasks. For example, Gidaris *et al.*[2018] proposed to randomly rotate an image by one of four possible angles and let the model predict that rotation. Apart from trying to predict the rotation, one can also try to recover part of the data, such as image completion [Pathak *et al.*, 2016]. In this paper, we utilize self-supervised learning for question-image correlation estimation as an auxiliary task to assist the VQA model to overcome language priors. We randomly change the image in the original relevant question-image pair, and then let the model predict its relevance.

3 Method

The framework of our approach is illustrated in Figure 2. Next, we will make detailed description of how it works.

3.1 The Paradigm of VQA

The purpose of VQA is to automatically answer textual questions according to images. Concretely, given a VQA dataset $\mathcal{D} = \{I_i, Q_i, A_i\}_{i=1}^{N}$ with N instances, where $I_i \in \mathcal{I}$, $Q_i \in \mathcal{Q}$ are the image and question for the i^{th} instance while $A_i \in \mathcal{A}$ is the corresponding annotation, the VQA model aims to learn a mapping function $\mathcal{F} : \mathcal{I} \times \mathcal{Q} \rightarrow \mathcal{R}^{\mathcal{A}}$ to produce an accurate distribution over the answer space $\mathcal{A}$. $\mathcal{F}$ typically consists of three parts: extracting features for both image and question, fusing them to obtain a joint multi-modal representation, and predicting a distribution over the answer space. We can write the answer prediction for the i^{th} image and question as $\mathcal{F}(\mathcal{A}|I_i, Q_i)$. Note that almost all the existing VQA models [Yang *et al.*, 2016; Kim *et al.*, 2018; Anderson *et al.*, 2018] follow this paradigm and their parameters are optimized by minimizing the cross-entropy loss $\mathcal{L}_{vqa\_ce}$ in Equation (2) or multi-label soft loss $\mathcal{L}_{vqa\_ml}$ in Equation (3):

$$P(\mathcal{A}|I_i, Q_i) = softmax(\mathcal{F}(I_i, Q_i)) \tag{1}$$

$$\mathcal{L}_{vqa\_ce} = -\frac{1}{N}\sum_{i}^{N} logP(\mathcal{A}|I_i, Q_i)[A_i] \tag{2}$$

$$\begin{aligned}\mathcal{L}_{vqa\_ml} = -\frac{1}{N}\sum_{i}^{N} [t_i log(\delta(\mathcal{F}(\mathcal{A}|I_i, Q_i))) \\ + (1 - t_i) log(1 - \delta(\mathcal{F}(\mathcal{A}|I_i, Q_i)))]\end{aligned} \tag{3}$$

where $\delta(\cdot)$ denotes the sigmoid function, t_i is the soft target score of each answer for the i^{th} instance, denoted as $t_i = \frac{number\ of\ votes}{n}$, where n is the number of valid answers for the i^{th} question, and $number\ of\ votes$ is the number of each answer that human annotated for this question.

3.2 Question-Image Correlation Estimation

A VQA model memorizing the language priors tends to make predictions only based on the questions while ignoring the images. Ideally, a question can only be answered when the given image containing the information related to it. Therefore, it is of crucial importance to require the VQA model to judge whether the given image can be used as the reference or not before answering a specific question. Unfortunately, this requirement has been neglected by all the previous works since the question-image pairs have been matched correctly in existing benchmarks. We illustrate that such validation is necessary to alleviate language priors in VQA, because it can force the model to refer image contents rather than answer blindly. To this end, we propose an auxiliary task called Question-Image Correlation Estimation (QICE), a binary classification task, to predict whether the question-image pair is relevant before answering a question. In this paper, we define the relevant question-image pair as the image can be used to answer the question with a specific answer.

Balanced question-image pairs generation. As shown in Figure 2 (b), we first automatically generate a set of labeled question-image pairs from the the original dataset without introducing extra human annotations for the auxiliary task. Specifically, each question-image pair (Q, I) in the training set is treated as a relevant pair with label $c = 1$, because there is an answer A for this pair in the dataset. And then for each relevant pair (Q, I), we replace the original image I by a randomly selected image from the image set $\mathcal{I}$, which is denoted as $I' = Sample(\mathcal{I} \backslash \mathrm{I})$. In this way, we can get another question-image pair (Q, I'). Obviously, the probability of (Q, I') being a relevant pair is very low considering the huge size of $\mathcal{I}$, thus we assign an irrelevant label $c = 0$ to each generated pair. As a result, we can obtain a balanced question-image pair matching dataset where the number of relevant pairs is equal to that of the irrelevant pairs. Note that the construction of the balanced question-image pairs does not need any human annotation.

Correlation estimation. With the generated balanced data, we can train a QICE model to predict the label of each question-image pair by optimizing the cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{self} = -\frac{1}{2N}\sum_{i}^{2N} c_i log\, QICE(Q_i, I_i) \\ + (1 - c_i) log(1 - QICE(Q_i, I_i))]\end{aligned} \tag{4}$$

where $\mathcal{L}_{self}$ can be interpreted as a self-supervised training loss since it only leverages the label supervision c from our generated data. The objective function guarantees the QICE model to understand the question as well as image contents because each Q corresponds to balanced relevant and irrelevant instances and no language priors can be depended on. In the next subsection, we will discuss how to leverage our auxiliary task QICE with the balanced data to assist the VQA model to eliminate language biases in a unified framework.

3.3 Unified Self-supervised Framework

In this section, we present our unified VQA framework that can simultaneously answer questions and estimate question-image relevance. Obviously, the QICE task defined above can share the same network structure with VQA because they have the completely same inputs and similar outputs: they all take question-image pair (I, Q) as input, and VQA predicts a distribution over answer space $\mathcal{A}$ while QICE produces a binary label on a specific answer A. Such property motivates us to settle these two tasks concurrently in a unified VQA framework as shown in Figure 2.

For the VQA model depicted in Figure 2 (a), it takes a relevant question-image pair (Q, I) as input, and predict a distribution $\mathcal{F}(\mathcal{A}|Q, I)$ over answer space $\mathcal{A}$, which can be optimized by minimizing VQA loss $\mathcal{L}_{vqa\_ce}$ or $\mathcal{L}_{vqa\_ml}$. This objective function teaches the model to learn the capability of answering questions. For the QICE task displayed in Figure 2 (c), given a question-image pair (I, Q) corresponding to a specific answer A, the prediction probability $P(A|Q, I)$ of the VQA model can be regarded as the confidence of (I, Q) being a relevant pair. The larger the probability, the higher the matching degree. Therefore, $\mathcal{L}_{self}$ can be rewritten as:

$$\begin{aligned}\mathcal{L}_{self} = -\frac{1}{2N}\sum_{i}^{2N}[c_i log P(A_i|Q_i, I_i) \\ + (1 - c_i) log(1 - P(A_i|Q_i, I_i))]\end{aligned} \tag{5}$$

The model is required to make the right binary predictions for the QICE task, which can enforce the model to better understand images since each question is paired with equal amounts of relevant and irrelevant images. More specifically, the first term of $\mathcal{L}_{self}$ aims to maximize the confidence of a question-image pair to be relevant, which is consistent with the objective of the VQA task that makes a prediction on the ground-truth A with high confidence. What's more important, the second term of $\mathcal{L}_{self}$ is designed to minimize the confidence of a pair to be relevant, which can exactly meet with the requirement of language prior reduction. Intuitively, the question dependency of a VQA model can be measured by the confidence of a question being answered correctly even with irrelevant images. The larger the confidence, the stronger the dependency. Minimizing the confidence of irrelevant pairs being relevant can explicitly prevent the VQA model from being overly driven by the language priors, and here we name it as question dependency loss $\mathcal{L}_{qd}$:

$$\mathcal{L}_{qd} = -\frac{1}{N}\sum_{i}^{N} log(1 - P(A_i|Q_i, I_i')) \tag{6}$$

We omit c_i in Equation (6) since $\mathcal{L}_{qd}$ is only valid for irrelevant question-image pairs (Q, I'). Mathematically, minimizing $-log(1 - P(A|Q, I'))$ in $\mathcal{L}_{qd}$ is equivalent to minimizing $P(A|Q, I')$. Experimentally, minimizing $P(A|Q, I')$ is more stable than minimizing $-log(1-P(A|Q, I'))$ during training, which is because the gradient of $P(A|Q, I')$ is more stable than that of $-log(1 - P(A|Q, I'))$. Therefore, we propose to minimize $P(A|Q, I')$ directly, and the updated question dependency loss $\mathcal{L}_{qd}$ can be defined as:

$$\mathcal{L}_{qd} = \frac{1}{N}\sum_{i}^{N} P(A_i|Q_i, I_i') \tag{7}$$

Consequently, QICE can be naturally regarded as an underlying multi-task learning task, containing two parts: visual question answering and language priors reduction. We can reformulate $\mathcal{L}_{self}$ as:

$$\mathcal{L}_{self} = \mathcal{L}_{vqa} + \alpha\mathcal{L}_{qd} \tag{8}$$

where $\mathcal{L}_{vqa}$ can be any VQA loss ($\mathcal{L}_{vqa\_ce}$ or $\mathcal{L}_{vqa\_ml}$), and α is a hyper-parameter. Obviously, $\mathcal{L}_{self}$ can be seen as a generalized VQA loss, as it degenerates to $\mathcal{L}_{vqa}$ when $\alpha = 0$. That means the question dependency loss $\mathcal{L}_{qd}$ actually acts as a regularizer, preventing the VQA model from memorizing the language priors and forcing it to better understand images. As a result, $\mathcal{L}_{self}$ offers flexibility in controlling the balance between answering questions and reducing language priors. Moreover, we do not need to explicitly optimize the model to be expert in estimating the correlations of question-image pairs, and we just use its balanced supervision to compensate for the data biases with our self-supervised loss. Following this, our method can alleviate language priors in a self-supervised manner without using external supervision.

4 Experiments

4.1 Datasets and Baselines

Datasets. Our approach is evaluated on the most commonly used benchmark VQA-CP v2 [Agrawal *et al.*, 2018] with the standard evaluation metric [Antol *et al.*, 2015]. The VQA-CP v2 dataset is derived from VQA v2 [Goyal *et al.*, 2017] by reorganizing the train and validation splits, and the Q-A pairs in the training set and test set have different distributions. Therefore, it is suitable for evaluating the model's generalizability. We also evaluate our model on the VQA v2 dataset containing strong biases and report the results on its validation split.

Baselines. We compare our approach against the following baseline methods: (1) non-annotation-based methods: UpDn [Anderson *et al.*, 2018], AdvReg [Ramakrishnan *et al.*, 2018], Rubi [Anderson *et al.*, 2018] and DLR [Jing *et al.*, 2020]; (2) annotation-based methods: HINT [Selvaraju *et al.*, 2019] and SCR (best-performing method) [Wu and Mooney, 2019].

4.2 Implementation Details

Our approach is model agnostic and can be applied to different VQA models. In this paper, we mainly evaluate our method based on UpDn [Anderson *et al.*, 2018], and we add one Batch Normalization layer before the classifier. Following previous work, we use the pre-trained Faster R-CNN to extract image features. For each image, it is encoded as a set of 36 objects with corresponding 2048-dimensional feature vectors. All the questions are trimmed to the same length 14. For each question, the words are initialized by the 300-dimensional Glove embeddings and then feed into GRU to get a sentence-level representation with the dimension of 1280.

We pre-train the model with the VQA loss for 12 epochs and fine-tune it with the self-supervised loss for 20 epochs. The batch size is 256, and the irrelevant images are randomly selected from mini-batches. The Adam optimizer is adopted with the initial learning rate of 0.001 which is halved every 5 epochs after 10 epochs. We evaluate our approach with different VQA losses in our main experiment, setting $\alpha = 3$ for multi-label VQA loss and $\alpha = 1.2$ for cross-entropy VQA loss. All the other experiments in this paper are based on multi-label VQA loss with $\alpha = 3$. The hyper-parameter α setting is also investigated in the next subsection.

4.3 Experimental Results and Analysis

Comparison with state-of-the-art. Our approach is tested based on two VQA losses (cross-entropy loss and multi-label loss) separately. To eliminate the stochasticity from the random sampling strategy, we report an average score of 10 experiments on the test set. From the results shown in Table 1, we can observe that: (1) Our approach can not only improve the overall performance of the baseline UpDn (**+14.35%** for cross-entropy loss and **+16.06%** for multi-label loss), but also significantly outperform the best-performing method SCR (**+3.13%** for cross-entropy loss and **+8.09%** for multi-label loss). (2) The improvements based on both VQA losses are all remarkable. Typically, using multi-label loss can achieve better performance since it is consistent with the evaluation metric and considers multiple feasible answers, which is shown to be more generalizable. (3) No matter which VQA loss is used, our approach can achieve extremely high accuracy (**87.75%** and **86.53%**) on the "Yes/No" question type, which indicates that our strategy is indeed to be effective in overcoming the language priors since biases are more likely to exist in these simple questions. (4) For the hardest "Num" questions, we can also get surprising improvements, which strongly illustrates that our approach can jointly understand images and questions, and reason them efficiently.

Method	Yes/No	Num	Other	Overall
UpDn [2018]	42.27	11.93	46.05	39.74
AdvReg [2018]	65.49	15.48	35.48	41.17
Rubi [2019]	68.65	20.28	43.18	47.11
DLR [2020]	70.99	18.72	45.57	48.87
HINT [2019]	70.04	10.68	46.31	47.70
SCR [2019]	71.60	11.30	48.40	49.50
UpDn†-$\mathcal{L}_{ce}$	47.27	13.67	40.32	38.28
UpDn†-$\mathcal{L}_{ml}$	43.45	13.64	48.18	41.53
UpDn+Ours-$\mathcal{L}_{ce}$	**87.75**	26.40	41.42	52.63
UpDn+Ours-$\mathcal{L}_{ml}$	86.53	**29.87**	**50.03**	**57.59**

Table 1: Performance on VQA-CP v2 test split. The first row shows the performance of non-annotation-based models, while the second row displays that of annotation-based methods. Our method significantly outperforms all these methods (including the best-performing method) no matter which VQA loss is used. † denotes the reimplementation of our baseline. $\mathcal{L}_{ce}$ is cross-entropy VQA loss and $\mathcal{L}_{ml}$ is multi-label VQA loss. Accuracies in percentage (%) are reported.

Model	Proportion of Training Set				
	20%	40%	60%	80%	100%
UpDn†[2018]	36.22	38.90	39.40	40.61	41.53
SCR [2019]	-	-	-	-	49.50
UpDn+Ours	**52.71**	**54.42**	**56.83**	**57.31**	**57.59**

Table 2: Performance on the VQA-CP v2 test set with different amounts of training data. Our approach outperforms UpDn with an average improvement of +16.44%. † is the reimplementation of the baseline. Overall accuracies in percentage (%) are reported.

Performance on smaller training sets. To further demonstrate the advantage of our approach, we randomly sample different amounts of training data from the original training set and conduct a series of experiments. All the experiments are tested on the standard test set and results are shown in Table 2. We find that our method gets an average accuracy improvement of **+16.44%** over baseline UpDn. What's more important, even with 20% of the training data, our approach can also significantly surpass the best-performing method SCR trained with external supervision on the full training set. We believe this is because our approach can effectively leverage the balanced data with the assistance of our regularizer, which is more likely to exhibit great generalizability.

Performance based on different baselines. We also conduct experiments based on two additional VQA models: SAN [Yang *et al.*, 2016] and BAN [Kim *et al.*, 2018]. From the

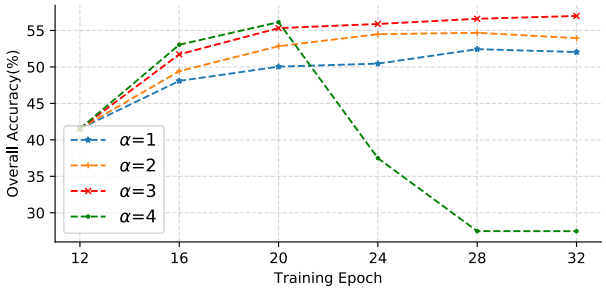

Figure 3: Comparison of overall accuracies with different α settings. Our method achieves better performance when $\alpha = 3$.

Method	Overall	Gap$\Delta\uparrow$
SAN [Yang *et al.*, 2016]	24.96	**+12.68**
SAN+Ours	**37.64**	
BAN [Selvaraju *et al.*, 2019]	41.48	**+13.48**
BAN+Ours	**54.96**	

Table 3: Performance on the VQA-CP v2 test set based on different baselines. Overall accuracies in percentage (%) are reported.

results depicted in Table 3, we can observe that the improvements for different baselines are all remarkable and consistent, which demonstrates that our method is model agnostic.

Performance on biased VQA dataset. We also evaluate our approach on the VQA v2 dataset containing strong language biases. We pre-train the model with VQA loss for 6 epochs and then fine-tune it for 10 epochs. As shown in Table 4. Our approach gets an improvement on VQA v2 val, while other baseline methods result in performance drops. The reason is that our self-supervised loss can achieve a balance between answering questions and eliminating language priors.

Impact of different α. To investigate the impact of the hyper-parameter α, which makes a trade-off between answering questions and overcoming language priors, we conduct extensive experiments with different α settings. Due to space limitations, in this paper, we only analyze the case when using multi-label VQA loss, see Figure 3. The model yields the highest performance when $\alpha = 3$. What's more, a large α might cause model collapse after several epochs, while a small α will result in unsatisfactory performance.

Qualitative analysis. We quantitatively evaluate the effectiveness of our approach. As shown in Figure 4, our method can answer the questions correctly and focus on the right regions. For example, when answering the question "Is this a professional game?", our method can pay more attention to the characters on the man's clothes, which might be an important visual clue to judge whether the game is professional.

5 Conclusion

In this paper, we propose a novel self-supervised learning framework to overcome language priors in VQA. Based on a model-agnostic auxiliary task, our framework is able to effectively exploit the automatically generated balanced data to alleviate the influence of dataset biases. Experimental results show that our approach achieves a balance between answering questions and overcoming language priors, and leads to a better overall learning outcome, achieving a new state-of-the-art on the most commonly used benchmark VQA-CP v2. Theoretically, we believe that our work can be a meaningful step in realistic VQA and solving the language bias issue, and this self-supervision can be generalized to other tasks (e.g. image caption) that are subject to the inherent data biases.

Figure 4: Qualitative comparison between our self-supervised approach and the baseline UpDn. The bounding boxes indicate the most important regions with attention values. G-T is ground-truth.

Method	Overall
UpDn [Anderson *et al.*, 2018]	63.48
AdvReg [Ramakrishnan *et al.*, 2018]	62.75
DLR [Jing *et al.*, 2020]	57.96
HINT [Selvaraju *et al.*, 2019]	62.35
SCR [Wu and Mooney, 2019]	62.20
UpDn+Ours	**63.73**

Table 4: Overall accuracy(%) on the VQA v2 val split. Our approach does not hurt the performance of the model on the biased dataset, while other bias-reducing methods all get performance drops.

Acknowledgments

We thank all the anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China (grant No.U19A2057), the National Science Fund for Distinguished Young Scholars (grant No.61525206), the Fundamental Research Funds for the Central Universities (grant No.WK3480000008), the Strategic Priority Research Program of Chinese Academy of Sciences (grant No.XDC02040400), the National Key Research and Development Program (grant No.2016QY03D0503) and Tianjin New Generation Artificial Intelligence Major Program (grant No.19ZXZNGX00110).

References

[Agrawal *et al.*, 2018] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018.

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[Andreas *et al.*, 2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[Cadene *et al.*, 2019] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems*, pages 839–850, 2019.

[Conneau *et al.*, 2016] Alexis Conneau, Holger Schwenk, Loıc Barrault, and Yann Lecun. Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*, 2, 2016.

[Gidaris *et al.*, 2018] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[Guo *et al.*, 2019] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang, and Mohan Kankanhalli. 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*, pages 75–84, 2019.

[Jing *et al.*, 2020] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. 2020.

[Kim *et al.*, 2018] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[Liu *et al.*, 2016] An-An Liu, Yu-Ting Su, Wei-Zhi Nie, and Mohan Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):102–114, 2016.

[Liu *et al.*, 2018] Anan Liu, Ning Xu, Hanwang Zhang, Weizhi Nie, Yuting Su, and Yongdong Zhang. Multi-level policy and reward reinforcement learning for image captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 821–827, 2018.

[Mao *et al.*, 2019] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.

[Parkhi *et al.*, 2015] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *British Machine Vision Association*, page 6, 2015.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[Ramakrishnan *et al.*, 2018] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pages 1541–1551, 2018.

[Selvaraju *et al.*, 2019] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2591–2600, 2019.

[Wu and Mooney, 2019] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. In *Advances in Neural Information Processing Systems*, pages 8601–8611, 2019.

[Wu *et al.*, 2016] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4622–4630, 2016.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[Yi *et al.*, 2018] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pages 1031–1042, 2018.