Loss-rescaling VQA: Revisiting Language Prior Problem from a Class-imbalance View

Yangyang Guo, Liqiang Nie *Senior Member, IEEE*, Zhiyong Cheng, and Qi Tian *Fellow, IEEE*

***Abstract*—Recent studies have pointed out that many well-developed Visual Question Answering (VQA) models are heavily affected by the language prior problem, which refers to making predictions based on the co-occurrence pattern between textual questions and answers instead of reasoning visual contents. To tackle it, most existing methods focus on enhancing visual feature learning to reduce this superficial textual shortcut influence on VQA model decisions. However, limited effort has been devoted to providing an explicit interpretation for its inherent cause. It thus lacks a good guidance for the research community to move forward in a purposeful way, resulting in model construction perplexity in overcoming this non-trivial problem. In this paper, we propose to interpret the language prior problem in VQA from a class-imbalance view. Concretely, we design a novel interpretation scheme whereby the loss of mis-predicted frequent and sparse answers of the same question type is distinctly exhibited during the late training phase. It explicitly reveals why the VQA model tends to produce a frequent yet obviously wrong answer, to a given question whose right answer is sparse in the training set. Based upon this observation, we further develop a novel loss re-scaling approach to assign different weights to each answer based on the training data statistics for computing the final loss. We apply our approach into three baselines and the experimental results on two VQA-CP benchmark datasets evidently demonstrate its effectiveness. In addition, we also justify the validity of the class imbalance interpretation scheme on other computer vision tasks, such as face recognition and image classification.**

***Index Terms*—Visual Question Answering, Language Prior Problem, Class Imbalance, Loss Re-scaling**

I. Introduction

VIsion and language are two ubiquitous elements in human cognition. One key artificial intelligence effort in bridging these two is to answer natural language questions about a visual scene, the *de facto* Visual Question Answering (VQA) task. The past few years have witnessed the great progress of VQA owing to the considerable development of computer vision and natural language processing. Conventional VQA methods generally cast it as a **classification problem**, where the image and question are handled via Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), respectively, and the fused multi-modal features are classified into one or a few candidate correct answers [1], [2].

Though a lot of VQA approaches have pushed the advancement over various benchmarks [1], [3], one notorious issue is that they are often hindered by the widely known language prior problem, namely, many VQA models answer questions without comprehensively exploiting the visual contents. In fact, existing VQA models are often brittle and easy to be fooled by some statistical shortcuts owing to the linguistically superficial correlation between questions and answers. For example, the questions under the question type *how many*[1] can mostly be answered with *2* in the training set [1]. This actually leads the VQA models to overwhelmingly reply to *how many* questions with *2* yet not truly reasoning the given image (See the training answer distribution and answer prediction in Figure 1 for example). As a result, it is problematic to assert that the previous benchmark progress in VQA drives by the real visual understanding rather than merely learning the dataset bias. To tackle this, a diagnostic benchmark, i.e., VQA-CP (VQA under Changing Priors) [4] has recently been proposed. It re-splits the VQA v1 [1] and VQA v2 [3] datasets in a way that the distribution of answers per question type is significantly distinct between the training and testing sets. As can be expected, the performance of many prior state-of-the-art VQA models [2], [5], [6] largely degrades on the VQA-CP datasets.

It is worth mentioning that great success has been achieved by prior efforts dedicated to the language prior problem in VQA. These methods can be roughly grouped into two categories: *single-branch* and *two-branch*. Specifically, approaches in the first category directly strengthen the visual feature learning for weakening the textual shortcut influence on model decision [7]–[10]. For instance, HINT [7] and SCR [10] leverage additional visual annotations to align the image region-level importance[2] with human attention map. Though these methods show competitive capability over their rivals, nonetheless, collecting such human annotations is expensive and burdensome, limiting the practicality in method adaptation across datasets. By contrast, the two-branch methods deliver the dominating performance due to their explicit counterwork over the language priors [12]–[15]. A typical design is to append an additional question-only model training branch to the orthodox question-image one, which intentionally captures the language prior and is further suppressed by the model branch with both image and question. The computational cost of this kind of methods is thereby increased since there is a new question-only branch being introduced compared with the traditional question-image branch methods.

Yangyang Guo and Liqiang Nie are with Shandong University, China. E-mail: {guoyang.eric, nieliqiang}@gmail.com.

Zhiyong Cheng is with Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), China. E-mail: jason.zy.cheng@gmail.com.

Qi Tian is with Huawei Noah's Ark Laboratory, Huawei, China. Email: tian.qi1@huawei.com

[1] We refer *question type* to the first few words in the given question.

[2] The image region importance is expressed via the model gradient magnitude of the ground-truth answers, i.e., Grad-CAM [11].

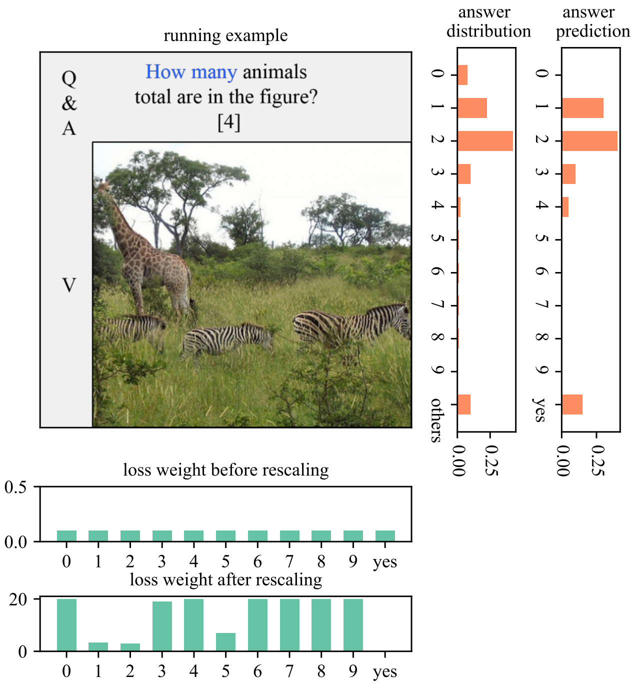


Fig. 1. Illustration of how the language prior problem arises and our solution. On the right orange bars, 1) the answer distribution in the training set for the current question type *how many* is biased, which leads the answer prediction to drift towards the frequent answers including an out-of-type answer *yes*. On the bottom green bars, 2) the original loss weight is evenly distributed over each candidate answer, while our solution is to re-scale it for the model decision balancing.

In general, the pipeline of the existing studies can be summarized into two consecutive steps: 1) describing the background and consequence of the language prior problem in VQA, and 2) developing effective methods to mitigate it and demonstrating the superiority over their contenders. However, little attention has been paid to interpreting or analyzing why the statistical shortcuts lead the model to blindly answer questions even though the corresponding image shows very intuitive cues for the right answer. This is tricky since without an appropriate interpretation for why the language prior problem occurs, the SOTA results on the VQA-CP benchmark mostly manifest to be skeptical. In other words, it is ambiguous that the performance improvement is from the alleviation of the language prior, or the over-fitting on the new curated dataset. Moreover, another imperative issue of the existing approaches overcoming the language prior problem in VQA is that they mostly involve many parameters to the baseline model, resulting in much heavier computation. Take the more advanced two-branch approaches for example, the model at least increases the parameters for the answer prediction of the question-only branch (i.e., proportional to the product of the size of predicted answers and the number of question features), which is burdensome and demands further optimization.

To tackle the above two issues, in this paper, we propose to bridge the connection between the language prior problem in VQA and the class-imbalance issue exploration, and make contributions from the following two aspects. Firstly, to analyze the inherent reasons that cause the language prior problem in VQA models, we offer to revisit the problem from a class-imbalance view and contribute a novel viewpoint to interpret it. Recall that each candidate answer is taken as a single class and the answer distribution for each question type is highly biased. This observation is in consistent with that in the class-imbalance classification problem when restricting the same question type for each <question, answer> pair. In the light of this, we argue that the language prior problem is essentially aroused by the imbalanced answer class distribution per question type. Accordingly, the VQA models are severely affected by the direct correlation between the question type and its frequent answers, leaving the visual features overlooked for answer prediction. To validate this point, we make two assumptions: 1) the loss for the mis-predicted answers with different numbers under each question type in the training set is significantly distinct from the well-trained model; and 2) the gradient norm of the parameters from the question encoding layer is much larger than that of the classification layer. The former assumption interprets why the model cannot fit the tail classes (sparse answers per question type) even with more training iterations. And the latter one points out that the parameters from the question encoding layer should be the blame for the language prior learning, which exactly matches the practice of the previous two-branch methods [12], [14] that overcomes the language prior problem through tuning the question encoding parameters, and yields strong support for these methods. After validating both assumptions, secondly, we further propose a loss re-scaling approach, which is motivated by the idea from models tackling the class-imbalance issue. The key to our loss re-scaling approach is to design distinct weights over each answer for computing the loss based on the corresponding question type. Specifically, smaller weights are assigned to the head answers while larger weights are attached to the tail ones under a question type. Notably, the proposed solution is model-agnostic, which means it can be plugged into any existing methods to overcome the language prior problem. Besides, the loss re-scaling module introduces **no** training parameters, i.e., **zero** incremental inference time to the baseline model. The increased computational cost is thus negligible compared to previous methods.

To testify the effectiveness of the proposed loss re-scaling approach, we conduct detailed ablation study with three baseline methods on the VQA-CP v2 and VQA-CP v1 datasets. The experimental results demonstrate the enhanced performance when applying our method to the baseline models under different settings. Moreover, we also test the viability of the class-imbalance interpretation scheme on other Computer Vision (CV) tasks which are also hindered by the class-imbalance problem.

In summary, the contributions of this paper are three-fold:

- We revisit the language prior problem in VQA from a class-imbalance view. The proposed interpretation scheme is functionary and can be extended to other relevant CV fields. To the best of knowledge, this is the first work to provide interpretations for the cause of the language prior problem in VQA.
- Based on the class-imbalance view, we design a simple yet effective loss re-scaling method to alleviate the lan-

guage prior effect in VQA.

- Extensive experiments to investigate the effectiveness of the proposed loss re-scaling method are performed on several benchmark datasets and promising results are observed. To benefit other research along this direction, the code has been made available[3].

The rest of the paper is structured as follows. We briefly review the related literature in Section II, and then detail the interpretation scheme in Section III, followed by the proposed loss re-scaling method in Section IV. Experimental setup and result analyses are presented in Section V and VI, respectively. We finally conclude our work and discuss the future directions in Section VII.

II. Related Work

In this section, we discuss three directions of related literature which are highly relevant to this work: Language Prior Problem in VQA, Data Imbalance Issue, and Visual Question Answering.

A. Language Prior Problem in VQA

Despite VQA has gained much attention from researchers and practitioners, a large body of studies have pointed out that VQA is greatly affected by the language prior problem [4], [12], [16]–[18]. In fact, it is undesirable that most VQA models perform well with the linguistic shortcut pattern between questions and answers instead of performing visual reasoning. To address this problem, researchers devoted their efforts into the following two directions.

Developing less-biased datasets. The dataset bias is almost unavoidable when crowdsourcing with human annotators. In VQA v1 [1], Antol et al. collected the first large-scale VQA dataset, which manifests a lot of 'visual priming bias'. For example, when people ask the question 'Is there a clock tower in the picture?', there actually contains clock towers in the given image. This is counterfactual since a balanced dataset should at least involve images with no clock towers for this kind of questions. To amend this problem, [3], [19] later presented the VQA v2 dataset which adds complementary samples in the way that each question is associated with a pair of images which results in two different answers. The role of image understanding is thereby elevated when deploying VQA models on this dataset. However, [4], [16] indicate that the VQA v2 dataset is still limited by the bias problem, which can potentially drive VQA models to learn the language prior. In view of this, VQA-CP [4] was created by re-spliting the VQA v1 and v2 datasets following the rule that the answer distribution for each question type is different for the training and testing sets. The performance of many VQA models drops dramatically on the VQA-CP datasets. Different from [4] re-organizing the existing biased datasets, Johnson et al. [20] curated a new diagnostic 3D shape dataset called Clevr, to control the question-conditional bias via rejection sampling within families of related questions. The answer distribution is therefore balanced for each question type.

From the model side. In contrast to the first kind of methods, studies in this group make efforts to directly designing mechanisms to reduce the language prior effect, which can be summarized into two types of approaches: single-branch and two-branch models. Specifically, single-branch methods are proposed to directly enhance the visual feature learning in VQA. For example, HINT [7] and SCR [10] align the image region-level importance with the additional human attention map. VGQE [21] utilizes the visual and textual modalities equally when encoding the question, where the question features include not only the linguistic information from the question but also the visual information from the image. Zhu et al. [8] proposed a self-supervised framework to generate balanced question-image pairs, where another irrelevant image with an incorrect answer is appended for training to alleviate the effects of dataset bias. Differently, two-branch methods intend to explicitly counter the language prior learned by the model. A common processing is to introduce a question-only branch for deliberately capturing the language prior of the baseline model, followed by the question-image branch training to restrain the bias learning. This kind of approach is more popular due to its predominant effectiveness. For instance, Q-Adv [12] trains the above two models in an adversarial way, which minimizes the loss of the question-image model while maximizing the loss of the question-only one. More recent fusion-based methods [13]–[15] use the late fusion strategy to combine the two answer predictions and guide the model to focus more on the answers which cannot be correctly answered by the question-only branch.

Though the language prior problem in VQA has been extensively studied in recent years, nevertheless, few work contributes to analyzing the cause of this problem., i.e., what arouses the language prior problem is largely untapped in literature. Hence, in this work, we tentatively proposed to interpret the problem from a class-imbalance view, and developed a simple yet effective approach to tackle it.

B. Data Imbalance Issue

The data imbalance issue has been long seen as a troublesome problem in many research fields. Recent efforts mainly pursue solutions along the following three directions [22]: 1) *Data re-weighting* methods attach different weights to different classes for computing the final loss. A common solution is to assign large weights for tail classes (i.e., less frequent ones) while small weights for head classes (i.e., more frequent ones). For example, Focal loss [23] is proposed to automatically put more focus on hard and misclassified samples. 2) *Data re-sampling* approaches target at re-sampling the highly biased data to attain a more balanced data distribution. In particular, over-sampling methods [24] simply repeat the minority classes several times to alleviate the training bias, while under-sampling ones [25] randomly choose a subset of the majority classes. Nevertheless, the two sampling methods all have some drawbacks: over-sampling is easy to over-fitting and under-sampling undermines the model generalization capability. And 3) *Transfer learning*. Different from the above two lines of efforts, methods along this line transfer the knowledge learned

[3] https://github.com/guoyang9/class-imbalance-VQA.

TABLE I
TOY EXAMPLE OF WHY THE LANGUAGE PRIOR PROBLEM OCCURS UNDER THE QUESTION TYPE *how many*.

ground truth	proportion in training set	prediction	loss value	endowed name
2	80%	4	large	hard mistake
4	4%	2	small	easy mistake

from abundant head classes to the tail classes [26]. The knowledge can be intra-class variance [27] or deep semantic features [28].

In this work, we take the re-weighting solution as an exemplar to deal with the language prior problem in VQA. Note that the other two directions of methods can also be explored, and we leave this as a future work.

C. Visual Question Answering

Visual Question Answering is a typical vision-and-language task [29]–[31], which involves the input of an image and a natural language question. Almost all the existing VQA models follow such a paradigm: the image and the question are encoded with CNNs and RNNs, respectively, followed by a multi-modal fusion module and a classifier to predict the right answer(s). Based on this paradigm, existing models can be roughly classified into four categories [32]: *Joint Embedding*, *Attention Mechanism-based*, *Compositional Models*, and *Knowledge-enhanced*. The *Joint Embedding* methods [1], [33] encode the image and question into a common latent space, wherein the visual reasoning is operated in this feature space. Later on, the *attention mechanism* is introduced to focus on salient image regions [6], [34] or with question words [5], [35] to achieve finer-grained feature learning. The modular structure of questions is also exploited for more explicit reasoning of *Compositional VQA models* [36]. It is natural that some questions cannot be correctly answered from the image information, which require to access external knowledge for better reasoning. *Knowledge Enhanced* approaches [37], [38] thus integrate external knowledge base such as DBpedia [39] or ConceptNet [40] into VQA models.

Beyond the simple reasoning on images, researchers have also explored other auxiliary information and built more challenging benchmarks. For instance, Visual Commonsense Reasoning (VCR) requires a VQA model to provide an accurate answer as well a rationale (expressed via text) explaining why the chosen answer is true. TextVQA asks models to read and reason the scene text in images. And video question answering [41]–[43] extends the question answering procedure from images to videos.

III. REVISITING LANGUAGE PRIOR PROBLEM

The language prior problem in VQA has been extensively studied in recent years, however, few work tries to interpret the reasons from the angle of model learning, which makes the advancement on benchmarks unreliable to some extent. In the light of this, we intend to fill this gap by revisiting it from a class-imbalance view since current VQA models treat VQA as a classification problem and the answer distribution of each question type is highly imbalanced. In the following, we start with two assumptions and then empirically prove them. We further extend this interpretation scheme to other two CV tasks.

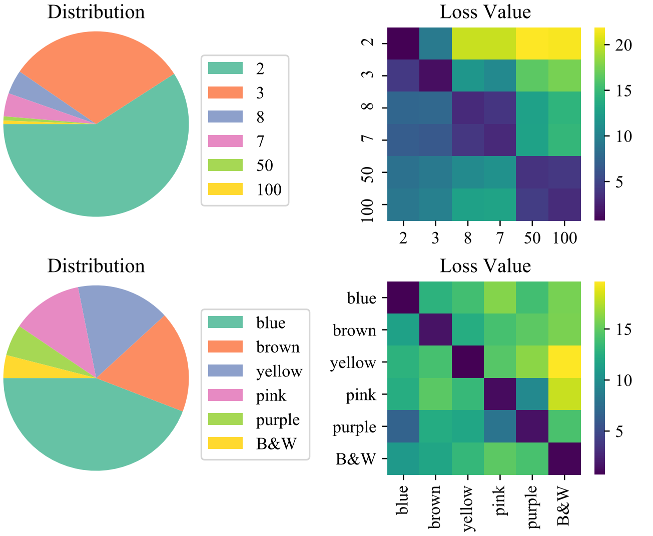


Fig. 2. Answer distribution and loss value of UpDn from two highly biased question types *How many* on the top and *What color* on the bottom. For the confusion matrix, the x-axis denotes the ground-truth answers from frequent to sparse, y-axis represents the predicted answers with the same order, and value signifies the loss. Note that the numbers of each case are kept identical for fair analysis.

A. New Viewpoint

We firstly present a new viewpoint for the class-imbalance issue based on the loss computation.

Assumption 1. *The loss value of hard mistake samples is larger than that of easy mistake ones during the late training phase (when the learning converges).*

We use a toy example in Table I to illustrate this assumption. Note that we only consider those samples which are incorrectly predicted, since the loss from the correctly predicted ones would be zero and back-propagate no gradients. As shown in Table I, *2* takes 80% of the answers to questions under the question type *how many*, while *4* accounts for only 4% in the training set. After the model is well-trained and the loss goes convergence, we assume that the loss value of these samples which are incorrectly predicted to *4* from *2* (hard mistake) is much larger than that of the ones predicted to *2* from *4* (easy mistake). In fact, the two endowed names in Table I are intuitive since the dataset bias would guide the model to more easily and wrongly predict an answer *4* to *2*. During the early training phase, the loss of both easy and hard mistakes should be numerically similar. And the easy mistakes induce the model parameters to fit their pace by drastically updating them. In contrast, the hard mistakes try to pull the parameters back to normal learning, yet short in quantity, leading to insignificant counterwork. This is straightforward to understand. However, when it comes to the late training phase or when the learning converges, the study on how the class imbalance issue affects the model prediction is particularly sparse in literature to our knowledge. It thus demands an insightful analysis on whether the inherent mechanism of the class-imbalance learning is consistent or not between the early and late training phases. Go a step further, if the model learning is special during the late training phase, it can provide us with insight on why the class-imbalance issue affects the model prediction when testing, and motivate us to design specific methods to tackle it. In view of this, we deem that during the late training phase, the loss value is significantly different between easy and hard mistakes. That is, for hard mistakes, the loss is relatively large, avoiding to 'make this mistake' next time. Nevertheless, for easy mistakes, the loss becomes negligible, receiving much 'tolerance' from the model.

Towards this end, we would like to empirically prove the assumption with a well-designed experiment. Specifically, we leverage a well-trained model, e.g., UpDn [2], to observe the loss value with respect to different mistakes. To begin with, we fix the model parameters when the loss goes convergence. And then, we reuse the samples in the training set and save the resultant prediction for further processing. Thereafter, for some highly biased question types, we use the confusion matrix to analyze the model prediction in Figure 2. It can be observed that the loss values of the confusion matrix in the upper triangle are much larger than that of the lower triangle. Especially for the *How many* questions, the loss values predicted from other sparse answers (i.e., *50*, *100*) to the most frequent answer *2* are the highest among others, which denotes that the model penalizes these hard mistakes to a great extent. On the other hand, the loss values of easy mistakes (corresponding to the lower triangle values) tend to be much smaller, signifying that the model seems to overlook these easy mistakes.

Based on the first Assumption, we further argues:

Assumption 2. *The gradient norm from the parameters of the question encoding layer is larger than that of the prediction layer.*

It is widely known that larger gradient norm of parameters often goes with more updating when performing back-propagating. This assumption denotes that the question encoding layer affects more to the language prior learning than the prediction layer. To validate it, we plot the gradient norm change of two groups of parameters with respect to the training steps in Figure 3. It can be seen that the gradient norm of the parameters from the question encoding layers continues to increase, while the parameters from the prediction layer decreases a lot with more training steps. This demonstrates that the parameters from the question encoding layer affect the model decision more drastically compared to the prediction layer. Note that the VQA models are affected more by the language prior problem in the late training phase. As a result, the parameters fluctuate larger should contribute more to the learning of the language prior with deeper training. As shown in Figure 3, in this case, we can conclude that the question encoding layer should be the blame for the language prior problem instead of the prediction one. This observation is in consistent with many of the two-branch studies [12], [14] that aim to modify the question encoding parameters to reduce the language prior. And this is the first time to evidently interpret why we should follow such a paradigm to alleviate the language prior problem. Arguably, our interpretation scheme produces a strong experimental support for the existing two-branch methods.

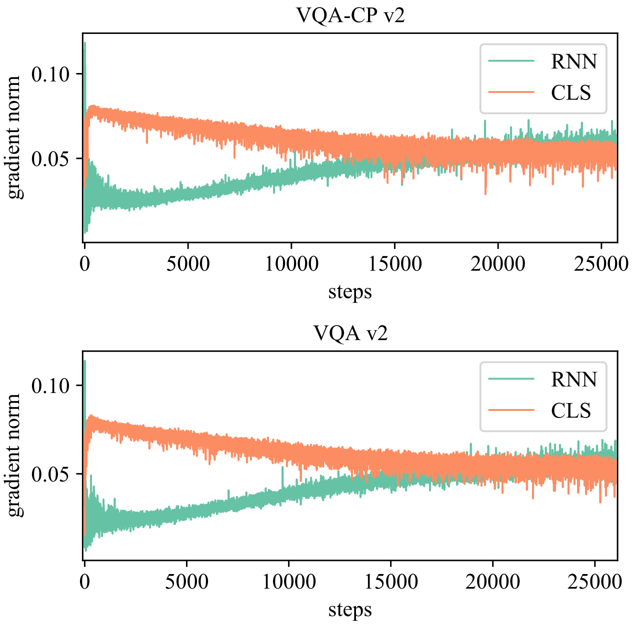


Fig. 3. Gradient norm of two groups of parameters with respect to training steps on two benchmark datasets. RNN represents the parameters from the question encoding layers, while CLS denotes the prediction layer.

B. Extension to other CV Tasks

To explore whether the proposed class-imbalance interpretation scheme works under highly imbalanced scenarios in other CV tasks, we tested the face recognition and image classification tasks and showed the results in Figure 4. The method and dataset we used for face recognition are Center Loss [44] the LFW dataset (Labeled Face in the Wild) [45], where we sampled seven persons with distinctive numbers in the training set. While for the image classification, we leveraged the imbalanced CIFAR-10 dataset [46] and the Resnet [47] method for evaluation. From Figure 4, we can observe that for the both tasks, the loss values show a similar manifestation with the samples in Figure 2. For example, when incorrectly predicting from the sparse *Chavez* class to other more frequent classes for face recognition, i.e., hard mistakes, the loss values become much larger than others. This can draw the parameters to avoid these mistakes and predict to other classes even though the model could make mistakes. Similar observation can also be found for the image classification task. Another interesting point from the imbalanced image classification one is that the model tries to ignore the mistakes from more frequent classes while penalizes the ones from the sparser classes.

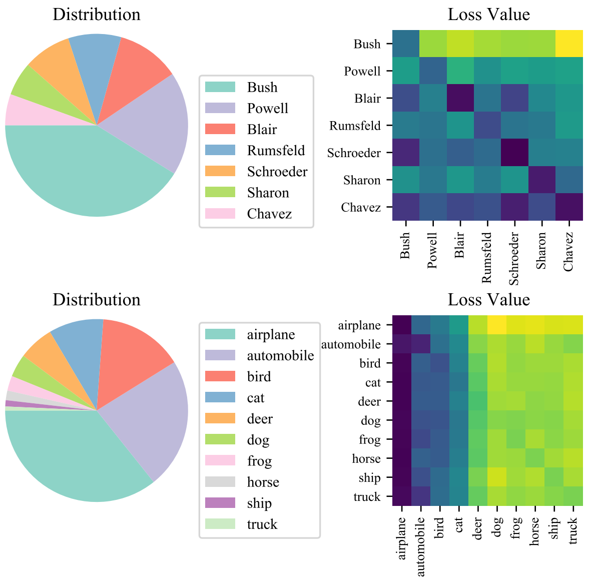


Fig. 4. Face and image class distribution and loss value prediction in the distinctive training set. The top one corresponds to seven persons from the LFW dataset. And the bottom one denotes the ten classes from the imbalanced CIFAR-10 dataset.

IV. Proposed Method

In this section, we present our method to overcome the language prior problem in VQA. Our method borrows the idea from the approaches attacking the class-imbalance issue. As mentioned above (in Section II-B), there are three directions of approaches to alleviate the class-imbalance problem, i.e., data re-weighting, data re-sampling and transfer learning. In this work, we adopt the data re-weighting one to deal with the language prior problem here. In particular, we design a novel mechanism to assign different weights over different answers (classes) under each question type to compute the final loss. Before delving into the details of the loss re-scaling method, we firstly quickly recap the basic learning functions in VQA.

A. Preliminary

The goal of VQA is to provide an accurate answer $\hat{a}$ to a given question Q upon an image I, which can be achieved by,

$$\hat{a} = \arg\max_{a \in A} p(a|Q, I; \Theta), \tag{1}$$

where Θ denotes the model parameters and A represents all the candidate answers. And the existing methods mostly tackle the VQA as a classification task after the multi-modal fusion module of image and question producing the fused feature $\boldsymbol{h}_f$. There are actually two possible solutions to predict the correct answer(s): Sigmoid activation function followed by Binary Cross Entropy loss and Softmax activation function followed by the Cross Entropy loss. We use Sigm-BCE and Soft-CE to denote the two solutions, respectively. Note that the ground-truth for each image-question pair is not necessarily unique, that is, one question is answered by ten annotators and then the label is softened by ten. In this way, for Sigm-CE, which treats each candidate answer as a single label for the binary classification problem[4], is given as,

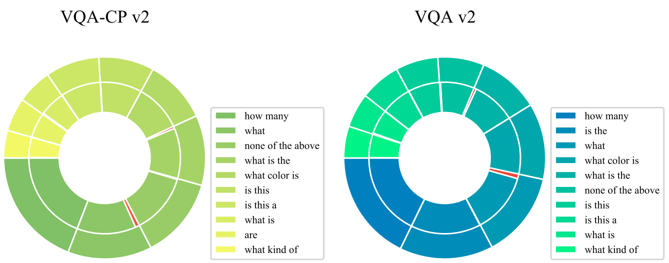


Fig. 5. Answer distribution in the training set (outer circle) and the validation set (inner circle) of the most frequent question types. The red samples represent the portion of answers from the validation set which are not in the answer set of the corresponding question type in the training set.

$$\begin{aligned} \boldsymbol{p} &= \boldsymbol{W}_p \boldsymbol{h}_f + \boldsymbol{b}_p, \\ \hat{\boldsymbol{p}} &= \sigma(\boldsymbol{p}), \\ \mathcal{L}_{bce} &= -\sum_{i=1}^{|A|} a_i \log \hat{p}_i + (1 - a_i) \log(1 - \hat{p}_i), \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ represents the Sigmoid activation function, a_i denotes the ground-truth answer label, $\boldsymbol{W}_p$ and $\boldsymbol{b}_p$ represent the learnable matrix and bias, respectively. Note that the current mainstream VQA models all adopt the Sigm-BCE solution, such as UpDn [2] and LMH [15], due to the insightful finding in [2], [48] that sigmoid output allow optimization for multiple correct answers per question than the single-label softmax one.

In contrast, Soft-CE tackles the multi-label classification in VQA via,

$$\begin{aligned} \boldsymbol{p} &= \boldsymbol{W}_p \boldsymbol{h}_f + \boldsymbol{b}_p, \\ \hat{\boldsymbol{p}} &= \text{Softmax}(\boldsymbol{p}), \\ \mathcal{L}_{ce} &= \sum_{i=1}^{|A|} a_i \log \hat{p}_i, \end{aligned} \tag{3}$$

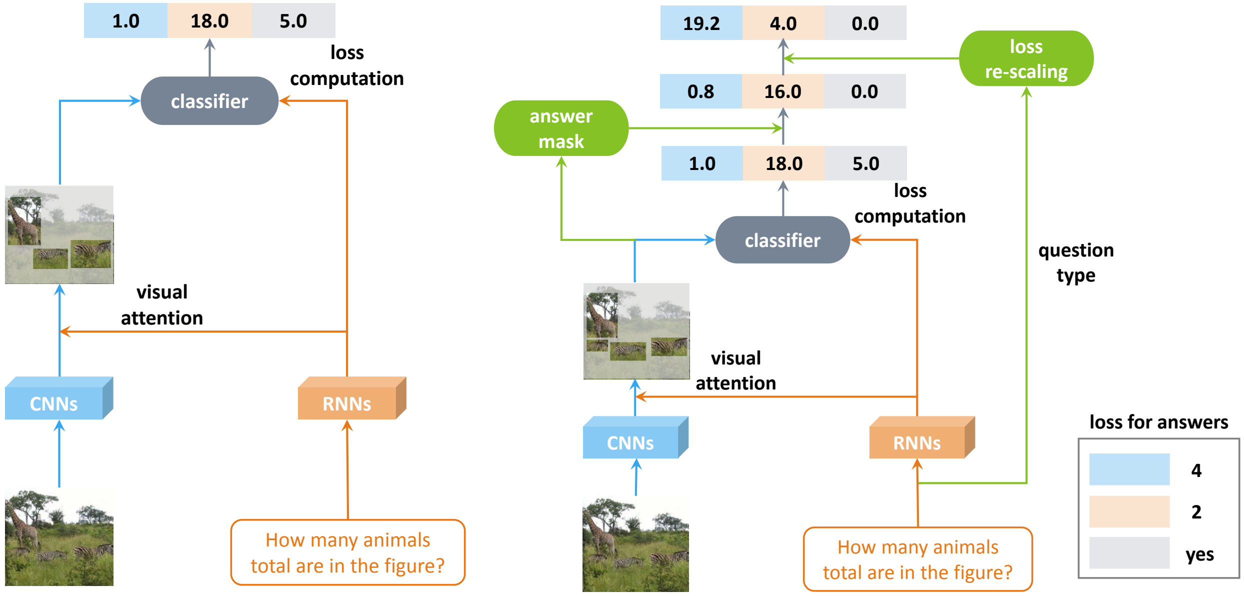
where Softmax$(\cdot)$ implies the Softmax activation function.

Based on the aforementioned two VQA solutions, we propose to assign distinctive weights for each answer when computing the final loss. To this end, in the following, we firstly design an *Answer Mask* module to constrain the predicted answers fall into the correct answer set for each question type. We then employ a novel strategy to compute the answer weights under the current question type for the final loss.

B. Answer Mask Modeling

Remind the example in Figure 1 that for the question type *how many*, there even exists an out-of-type answer *yes* which does not belong to the answer set for the current question type in the training set. This perplexes a lot as very rare *how many* questions should be answered with binary answers *yes* or *no*. One possible reason is that the *yes* answer takes large parts of all the answers when training, which results the model into

[4] Notice that we omit the mini-batch summation in all loss functions for simplicity.



(a) Baseline learning pipeline

(b) Our proposed learning pipeline

Fig. 6. Visual comparison of loss computation between the baseline learning pipeline and ours. We use three answers, i.e., *4*, *2* and *yes*, to intuitively illustrate how the proposed two modules work.

unexpectedly drifting to this frequent answer. To tackle this problem and ensure the predicted answers lie into the specific answer set of the given question type, we introduce an answer mask which makes sure that the predicted answers will not escape from the corresponding answer set. The motivation is supported by the findings in Figure 5, where only a very small portion of samples in the validation set are split to the out-of-type answers from the corresponding answer set in the training set. Concretely, the answer mask module takes the question feature and the image feature as inputs, and then employs a fully-connected layer equipped with non-linear activation functions,

$$\boldsymbol{m}_p = f(\boldsymbol{W}_q\boldsymbol{q} + \boldsymbol{W}_v\boldsymbol{v}), \tag{4}$$

where the size of $\boldsymbol{m}_p \in \mathbb{R}^{|A|}$ is the same with the candidate answer size, $f(x)$ denotes the activation function, and $\boldsymbol{W}_q$ and $\boldsymbol{W}_v$ are the learnable weights. To enable the answer mask modeling, we need to collect the true mask label $\boldsymbol{m}_a \in \mathbb{R}^{|A|}$ where the answer positions belonging to the given question type are set to be 1 and others are 0. Finally, the BCE loss is used to train this module,

$$\mathcal{L}_{mask} = \sum_{i=1}^{|A|} m_a^i \log m_p^i + (1 - m_a^i)\log(1 - m_p^i), \tag{5}$$

where m_a^i and m_p^i represent the i-th element in $\boldsymbol{m}_a$ and $\boldsymbol{m}_p$, respectively.

In fact, a straightforward design of $f(x)$ is to adopt the sigmoid activation function, which can map the features to a mask value within $(0, 1)$. However, we favor the output to be larger than 0.5 in most times so that the final predicted answer would be grounded more by this mask. Note that deploying all-ones mask will lead to over-fitting and is not flexible. We then devise a generalized softplus function in Equation 6,

$$f(x) = \max(1, \frac{1}{\alpha}\log(1 + \exp \alpha x)), \tag{6}$$

where α is a hyper-parameter which is fixed to be 1 in our experiment.

C. *Loss Re-scaling*

In this subsection, we present our proposed loss re-scaling module in detail. To begin with, we elaborate the computation of the re-scaling weights. Thereafter, we provide the re-scaled loss function and partial derivatives.

Re-scaling Weight Computation. We intuitively introduce a fixed weight to each answer in the loss function, where it can prevent the hard mistakes to update the parameters harshly while guide the easy mistakes for more influential model learning. In particular, we name the weight with μ_i, which is obtained via,

$$\mu_i = \frac{\sum_k^{|A|} n_k^j - n_i^j}{n_i^j}, \tag{7}$$

where n_i^j represents the number of answer a_i under the question type qt_j (for other answers not in the answer set of the current question type, μ_i is equal to 1). Under this situation, recollect the previous example in Table I, the loss weights for answer '2' and '4' are $(1 - 0.8)/0.8 = 0.25$, and $(1 - 0.04)/0.04 = 24$, respectively. In this way, the easy mistakes become relatively hard to propagate information for model parameters, while hard mistakes gain more chance to update parameters, especially for earlier layers. In implementation, we found that when the values are too large, the training becomes more unstable, we therefore employ a smoothing function and truncate too large values to be 100 for more stable training,

$$\mu_i = \max(100, \log(1 + \exp \mu_i)). \tag{8}$$

This scheme is specially designed for avoiding VQA models to learn too much language prior effect when training. In the testing phase, the loss re-scaling is deactivated and only the baseline model remains. This assures that there is **no incremental computation** at all when evaluating compared with the baseline model[5].

Losses and Partial Derivatives. Based on the re-scaling weight, we further derive the losses and its partial derivatives over the prediction for both the Soft-CE and Sigm-BCE below. With this, the gradient of other learnable parameters can be easily computed.

1) For Sigm-BCE, the loss function becomes,

$$\mathcal{L}^{\mu}_{bce} = -\sum_{i=1}^{|A|} \mu_i (a_i \log \sigma(p_i) + (1 - a_i) \log(1 - \sigma(p_i))), \tag{9}$$

where the partial derivative can be obtained by,

$$\begin{aligned}\frac{\partial \mathcal{L}^{\mu}_{bce}}{\partial p_i} &= -\mu_i (a_i \frac{1}{\sigma(p_i)} (\sigma(p_i)(1 - \sigma(p_i))) \\ &+ (1 - a_i) \frac{1}{1 - \sigma(p_i)} (-\sigma(p_i)(1 - \sigma(p_i)))) \\ &= \mu_i ((1 - a_i)\sigma(p_i) - a_i (1 - \sigma(p_i))).\end{aligned} \tag{10}$$

2) For the Soft-CE, the loss function is given as,

$$\mathcal{L}^{\mu}_{ce} = -\sum_{i=1}^{|A|} \mu_i a_i \log \frac{\exp p_i}{\sum_{j=1}^{A} \exp p_j}. \tag{11}$$

To compute the partial derivative p_i of from $\mathcal{L}_{ce}$, we firstly provide the partial derivative of the Softmax function,

$$\frac{\partial \hat{p}_i}{\partial p_i} = \begin{cases} \hat{p}_i (1 - \hat{p}_j), & \text{if } i == j \\ -\hat{p}_i \hat{p}_j, & \text{otherwise.} \end{cases} \tag{12}$$

And then the partial derivative is given by,

$$\begin{aligned}\frac{\partial \mathcal{L}^{\mu}_{ce}}{\partial p_i} &= -\sum_{i=1}^{|A|} \mu_i a_i \frac{\partial \log \hat{p}_i}{\partial p_i} \\ &= -\sum_{i=1}^{|A|} \mu_i a_i \frac{1}{\hat{p}_i} \frac{\partial \hat{p}_i}{\partial p_i} \\ &= \mu_i a_i (\hat{p}_i - 1) + \sum_{k \neq i}^{|A|} \mu_k a_k \hat{p}_i \\ &= (\mu_i a_i + \sum_{k \neq i}^{|A|} \mu_k a_k) \hat{p}_i - \mu_i a_i \\ &= \sum_{k=1}^{|A|} \mu_k a_k \hat{p}_i - \mu_i a_i,\end{aligned} \tag{14}$$

where line 3 is derived from Equation 12. In this way, the final loss becomes,

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{mask}, \tag{15}$$

where the $\mathcal{L}_{cls}$ can be either $\mathcal{L}^{\mu}_{ce}$ or $\mathcal{L}^{\mu}_{bce}$.

D. Method Application

Most state-of-the-art VQA architectures are compatible with our method. To show how the above two modules work, we leverage the UpDn [2] and illustrate its learning pipeline and ours in Figure 6. Specifically, for each input image, UpDn utilizes the pre-trained object detection networks (i.e., Faster R-CNN [49]) to obtain the most salient objects; meanwhile, the question words are sequentially processed via GRU (i.e., Gated Recurrent Network [50]) to capture the semantic information. A visual attention is then appended to refine the visual features by employing the question features to attend on each object features in the current image, which is followed by the multi-modal fusion module and the classifier for predicting the answers. As can be seen in Figure 6a, the BCE losses for answers *4*, *2* and *yes* are 1.0, 18.0 and 5.0, respectively. With our answer mask module in Figure 6b, the *yes* answer no longer takes effects since it is an out-of-type outlier for the question type *how many*. Our final loss re-scaling module then assigns distinct loss weights to each candidate answer according to the statistics in Table I. One can see that the loss for the sparse answer *4* is augmented while for the frequent answer *2* is diminished.

V. Experimental Setup

We conducted extensive experiments on three benchmark datasets to validate the effectiveness of the loss re-scaling method. In particular, the experiments are mainly leveraged to answer the following research questions:

- **RQ1**: Can the proposed loss re-scaling method outperform the state-of-the-art VQA models?
- **RQ2**: How do the answer mask and loss re-scaling modules perform on the two learning strategies?
- **RQ3**: Is the proposed loss re-scaling method superior than other loss re-weighting approaches?
- **RQ4**: Why does the proposed method perform better than the baseline methods?

In the following, we first provide the basic information of the evaluated benchmark datasets. We then detail the standard evaluation metric, followed by the key baselines used in this work, and we end this section with some implementation details.

A. Datasets

We tested our proposed method mainly on the two VQA-CP datastes: VQA-CP v2 and VQA-CP v1 [4], which are two well-known public benchmarks for evaluating the models' capability to overcome the language prior problem in VQA and are actually the new splits of the traditional VQA v2 [3] and VQA v1 [1] datasets, respectively. The VQA-CP v2 and VQA-CP v1 datasets consist of ∼122K images,

[5] By zero incremental inference time, we explicitly refer to the loss re-scaling module excluding the answer mask one since the loss re-scaling module produces a primary effect as shown in Table IV.

TABLE II

ACCURACY COMPARISON BETWEEN THE PROPOSED METHOD AND BASELINES OVER THE VQA-CP V2 TEST AND VQA V2 VAL SETS. GAP△ IMPLIES THE PERFORMANCE GAP OF VQA V2 VAL AND VQP-CP V2 TEST OF THE *All* ANSWER CATEGORY. REGARDING THE METHOD CATEGORY, 'PLAIN' REPRESENTS TRADITIONAL VQA METHODS WITHOUT SPECIALIZATION FOR OVERCOMING THE LANGUAGE PRIOR PROBLEM. 'TWO-BRANCH' AND 'SINGLE-BRANCH' DISTINGUISH WHETHER THERE IS ANOTHER QUESTION-ONLY TRAINING BRANCH. '−' AND '†' DENOTE THE NUMBER IS NOT AVAILABLE AND OUR IMPLEMENTATION, RESPECTIVELY. THE BEST PERFORMANCE IN CURRENT SPLITS IS HIGHLIGHTED IN BOLD.

Method Category	Method	VQA-CP v2 test				VQA v2 val				Gap△ ↓
		Y/N	Num.	Other	All	Y/N	Num.	Other	All	
Plain	Question-only [4]	35.09	11.63	07.11	15.95	67.95	30.97	27.20	43.01	27.06
	SAN [6]	38.35	11.14	21.74	24.96	70.06	39.28	47.84	52.41	27.45
	NMN [36]	38.94	11.92	25.72	27.47	73.38	33.23	39.93	51.62	24.15
	MCB [51]	41.01	11.96	40.57	36.33	77.91	37.47	51.76	59.71	23.38
	HAN [52]	52.25	13.79	20.33	28.65	-	-	-	-	-
	UpDn [2]	42.27	11.93	46.05	39.74	81.18	42.14	55.66	63.48	23.74
	UpDn [2]†	42.31	11.84	44.35	38.80	80.59	40.51	53.69	62.07	23.27
Two-branch	AdvReg [12]	65.49	15.48	35.48	41.17	79.84	42.35	55.16	62.75	21.58
	Rubi [14]	68.65	20.28	43.18	47.11	-	-	-	61.16	14.05
	LM [15]†	73.65	41.44	42.82	51.68	80.37	40.62	53.53	61.92	10.24
	LMH [15]†	70.29	44.10	44.86	52.15	74.18	38.43	53.05	59.07	6.92
Single-branch	GVQA [4]	57.99	13.68	22.14	31.30	72.03	-	-	48.24	16.94
	HINT [7]	67.27	10.61	45.88	46.73	81.18	42.99	55.56	63.38	16.65
	SCR [10]	72.36	10.93	**48.02**	49.45	78.8	41.6	54.5	62.2	12.75
	VGQE [21]	66.35	27.08	46.77	50.11	-	-	-	63.18	13.07
-	UpDn+Ours	68.42	21.71	42.88	47.09	64.22	39.61	53.09	55.50	8.41
	LM+Ours	**73.74**	45.14	44.59	53.17	76.28	36.61	52.71	59.45	6.28
	LMH+Ours	72.82	**48.00**	44.46	**53.26**	68.21	36.37	52.29	56.81	**3.55**

∼658K questions and ∼6.6M answers, and ∼122K images, ∼370K questions and ∼3.7M answers, respectively. And each question is answered by ten annotators. Moreover, the answer distribution of each question type is significantly different between the training and testing sets in the VQA-CP datasets. For example, in VQA-CP v1, *1* and *3* take about 50% of the answers under the question type *how many* in the training set, while *2* and *4* hold around 75% in the testing set. Other than VQA-CP, we also reported results on the VQA v2 dataset, a more biased benchmark, for completeness. For all the three evaluated dataset, the answers are divided into three categories: *yes/no*, *number* and *other*.

B. Evaluation Metric

We adopted the standard metric in VQA for evaluation [1]. For each predicted answer a, the accuracy is computed as,

$$Acc_a = \min(1, \frac{\#\text{humans that provide answer } a}{3}). \tag{16}$$

Note that each question is answered by ten annotators, and this metric takes the disagreement in human answers into consideration [1], [3].

C. Tested Baselines

We mainly tested our method over the following three baselines. Thereinto, UpDn is the most popular VQA baseline, LM and LMH are the current state-of-the-art models on the VQA-CP v2 dataset. Note that all the three baselines are based on the Sigm-BCE strategy.

UpDn [2] firstly leverages the pre-trained object detection networks to obtain salient object features for VQA. It employs a simple attention network to focus on the most important objects which are highly related with the given question.

LM [15] is a typical two-branch method which takes UpDn as baseline and combine the answer prediction from the two branches (i.e., question-image and question-only) in an ensemble mode.

LMH [15] extends LM by introducing another entropy loss to encourage the question-only branch prediction to be non-uniform, posing a greater impact on the ensemble.

D. Implementation Details

We adopted the publicly available implementation[6] for UpDn and implementation[7] for LM and LMH in our experiments.

For the answer mask module, we append a dropout layer with the dropout rate tuning from 0.0 to 1.0 with a step size of 0.1. And the answer mask loss weight is fixed to 1.0. For the loss re-scaling approach, we first train the answer mask module until converging, and then fine-tune the model with our appended re-scaling weights for another ten epoches. Note that we kept all the other settings being unchanged, such as learning rate, optimizer, mini-batch size.

VI. EXPERIMENTAL RESULTS

A. Overall Performance Comparison (RQ1)

Table II and III illustrate the performance results on VQA-CP v2 testing, VQA v2 validation and VQA-CP v1 testing set. The key observations from these two tables are as follows:

- Our methods achieve the best over all answer categories on the VQA-CP v2 and VQA-CP v1 datasets except for the *other* category on VQA-CP v2. Since SCR leverages additional annotations to recognize important objects with

[6] https://github.com/hengyuan-hu/bottom-up-attention-vqa.

[7] https://github.com/chrisc36/bottom-up-attention-vqa.

TABLE III
ACCURACY COMPARISON BETWEEN THE PROPOSED METHOD AND BASELINES OVER THE VQA-CP V1 DATASET. '—' AND '†' DENOTE THE NUMBER IS NOT AVAILABLE AND OUR IMPLEMENTATION, RESPECTIVELY. THE BEST PERFORMANCE IN CURRENT SPLITS IS HIGHLIGHTED IN BOLD.

Method	VQA-CP v1 test			
	Y/N	Num.	Other	All
Question-only [4]	35.72	11.07	08.34	20.16
SAN [6]	35.34	11.34	24.70	26.88
NMN [36]	38.85	11.23	27.88	29.64
MCB [51]	37.96	11.80	39.90	34.39
UpDn [2]†	43.76	12.49	42.57	38.02
AdvReg [12]	74.16	12.44	25.32	43.43
LM [15]†	75.01	28.86	42.22	53.59
LMH [15]†	76.61	29.05	43.38	54.76
GVQA [4]	64.72	11.87	24.86	39.23
UpDn+Ours	44.25	18.04	43.03	39.34
LM+Ours	75.51	26.93	**43.67**	54.07
LMH+Ours	**78.26**	**29.80**	42.76	**55.32**

more supervision, it thus achieves better performance on the *other* category wherein most answers belong to objects. In general, as these two datasets are deemed as an effective protocol for testing whether the evaluated method overcomes the language prior problem, we can conclude that the proposed method is capable of dealing with this issue and superior over the existing methods.

- For all the three baselines, with our answer mask and loss re-scaling modules, the baseline models can obtain a significant performance improvement. For example, in the VQA-CP v2 dataset, the overall accuracy of UpDn boosts from 38.80% to 47.09%, a 8.29% absolute improvement.
- The performance gap between VQA v2 val and VQA-CP v2 test in Table II denotes the model over-fitting level for the language prior since the VQA v2 is a more biased dataset, i.e., the smaller, the better. It can be seen that the methods specially designed for tackling the language prior problem outperform the traditional methods, and our approaches even surpass all the existing methods, which indicates our approaches learn less language prior than other methods.
- For the comparison of two-branch and single-branch methods, it can be seen that two-branch ones often outperform their contemporary rivals, such as AdvReg vs GVQA, Rubi vs HINT, and LMH even surpasses the succeeding method VGQE. One noticeable reason for this is that the question-only model in the two-branch methods explicitly captures the language prior and is then suppressed by the question-image model.

B. Ablation Study (RQ2)

We conducted detailed experiments to validate the effectiveness of each module and reported the results in Table IV and Figure 7. Note that the LM and LMH are particularly developed based on BCE loss function which cannot be adapted with the Soft-CE solution. And we extended the UpDn with the Soft-CE solution ourselves. We can find that:

- The loss re-scaling scheme is effective under a series of settings. When directly applying it to the baseline models, the accuracy is enhanced with a large margin. For instance, the UpDn and LMH achieve a 2.04% and 0.53% absolute improvements of the BCE loss, respectively. With the pre-training of the answer mask module, the model can further achieve a much higher accuracy and obtain the state-of-the-art performance.
- For the answer mask module, the softplus-G activation function does surpass the plain sigmoid one. The reason is that the mask is favored to be larger than 0.5 with a high probability.
- We further evaluated the dropout rate effect in the answer mask module and illustrated the results in Figure 7. It can be observed that a smaller dropout rate is better for UpDn, while LM and LMH prefer larger dropout rates.
- Surprisingly, our methods can achieve a substantially better results over the baseline method when using the CE

TABLE IV
EFFECTIVENESS VALIDATION OF THE ANSWER MASK AND LOSS RE-SCALING MODULES. '—' REPRESENTS THE NUMBER IS NOT AVAILABLE, SINCE LM AND LMH DIRECTLY PERFORMS ON THE BCE LOSS. AND SOFTPLUS-G IMPLIES THE ACTIVATION FUNCTION IN EQUATION 6.

Loss	Module	UpDn	LM	LMH
BCE	baseline	38.80	51.68	52.15
	+answer mask (sigmoid)	40.94	51.71	51.56
	+answer mask (softplus-G)	41.69	52.76	52.70
	+loss re-scaling	40.84	51.96	52.68
	+both	47.09	53.17	53.26
CE	baseline	40.49	-	-
	+answer mask (sigmoid)	39.32	-	-
	+answer mask (softplus-G)	40.30	-	-
	+loss re-scaling	55.93	-	-
	+both	**57.02**	-	-

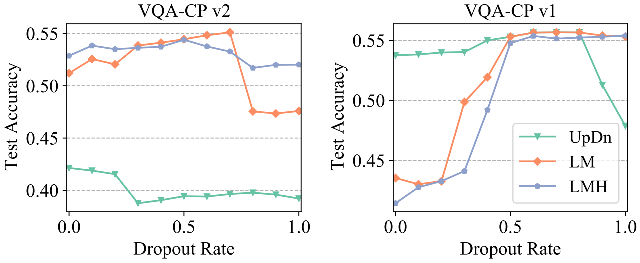


Fig. 7. Testing set accuracy of the three evaluated methods over the two datasets with respect to dropout rate.

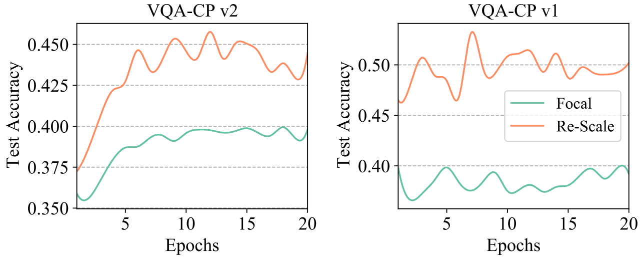


Fig. 8. Testing set accuracy of loss re-weighing methods on the UpDn baseline over the two datasets with respect to training epochs.

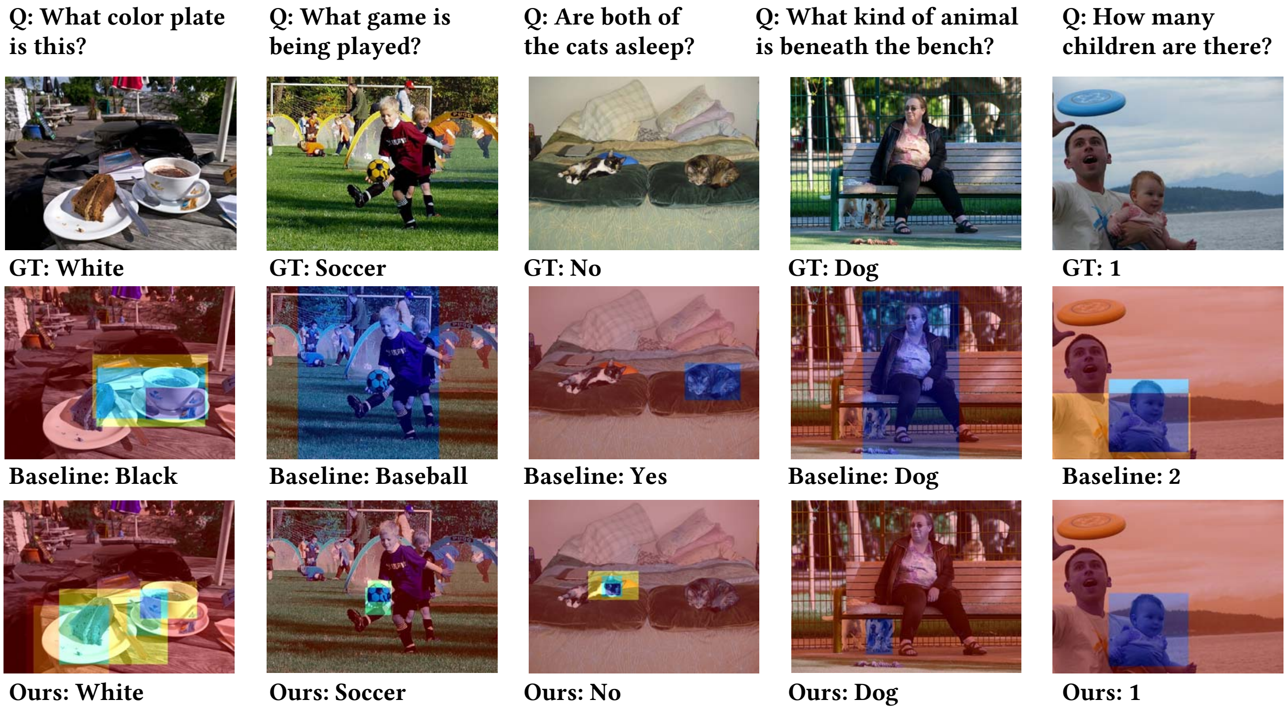

Q: What color plate is this?
Q: What game is being played?
Q: Are both of the cats asleep?
Q: What kind of animal is beneath the bench?
Q: How many children are there?
GT: White
GT: Soccer
GT: No
GT: Dog
GT: 1
Baseline: Black
Baseline: Baseball
Baseline: Yes
Baseline: Dog
Baseline: 2
Ours: White
Ours: Soccer
Ours: No
Ours: Dog
Ours: 1

Fig. 9. Visualization of results from the UpDn and our method on the VQA-CP v2 dataset. The first line of text is the given question, followed by the image in the second line and the ground-truth answer in the third line. The attention maps and predicted answers of UpDn and our method are sequentially provided.

loss function. For example, the final absolute performance improvement of our method over the baseline is over 16.53%! It is even better than the recent strong baseline LM and LMH. This highlights the effectiveness of the loss re-scaling scheme and the potential of the class-imbalance view on alleviating the language prior problem in VQA.

C. Other Loss Re-weighting Method (RQ3)

To examine whether the loss re-scaling scheme is superior than other commonly used ones in other tasks, we adopted a well-developed Focal Loss [23], [53] and tested its performance in VQA. Focal Loss is designed to put more focus on hard, misclassified examples, where the loss weights are auto-learned from the training data, which is given as,

$$FL(p_t) = -\alpha(1-p_t)^{\gamma} log(p_t), \tag{17}$$

where α and γ are hyper-parameters and p_t denotes the prediction for binary classification. We further extended this loss function to the VQA multi-class case and plotted the results in Figure 8. We can observe that there exists a quite large performance margin between our method and the Focal Loss. One possible reason is that our loss weights is particularly computed under each question type, which provides more explicit guidance than the weight auto-learning strategy in Focal Loss.

D. Case Study (RQ4)

We visualized the attention map and predicted answers of UpDn and our final method in Figure 9. Specifically, for the first, second and the last samples, the baseline all provides the incorrect answers due to the capturing of the language prior. The reason is that the answers *black*, *baseball* and *2* are the dominated answers under the question types *what color*, *what* and *how many*, respectively. For the last sample, the attention map from the baseline focuses on both the man and the child, which incorrectly leads the model to predict that there are *2* children. In contrast, our method attends solely on the *child* object and yields the correct answer. For the third sample, the baseline only looks an asleep *cat* and concludes that both cats are asleep, while our method notices the awake *cat*. And for the fourth sample, though the baseline also predicts the right answer, nevertheless, its attention is focused on both the *women* and the *dog*. Our method puts more attention on the animal beneath the bench - *dog*.

VII. Conclusion and Future Work

There is few work analyzing the inherent cause of the language prior problem in VQA. In this work, we propose to fill this gap via interpreting it from a class-imbalance view. To begin with, we present two assumptions and prove their viability. We then develop a simple yet effective loss re-scaling approach to attach distinct weights to the answers under the given question type. Extensive experiments on three publicly available datasets validate the effectiveness of the proposed loss re-scaling method. Moreover, we also extend the interpretation scheme to other two CV tasks. In the future, we would like to exploit other typical methods overcoming the class-balance issue, i.e. data re-balancing or transfer learning, for alleviating the language prior problem. Moreover, the class-imbalance view will be explored over other vision-and-language problems which are also hindered by the language bias issue, e.g., image captioning.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[3] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017.

[4] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *CVPR*, 2018.

[5] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NeurIPS*, 2016.

[6] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016.

[7] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh, "Taking a hint: Leveraging explanations to make vision and language models more grounded," in *ICCV*, 2019.

[8] X. Zhu, Z. Mao, C. Liu, P. Zhang, B. Wang, and Y. Zhang, "Overcoming language priors with self-supervised learning for visual question answering," in *IJCAI*, 2020.

[9] G. KV and A. Mittal, "Reducing language biases in visual question answering with visually-grounded question encoder," in *arXiv preprint arXiv:2007.06198*, 2020.

[10] J. Wu and R. Mooney, "Self-critical reasoning for robust visual question answering," in *NeurIPS*, 2019.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[12] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *NeurIPS*, 2018.

[13] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *CVPR*, 2020.

[14] R. Cadene, C. Dancette, M. Cord, D. Parikh *et al.*, "Rubi: Reducing unimodal biases for visual question answering," in *NeurIPS*, 2019.

[15] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't take the easy way out: ensemble based methods for avoiding known dataset biases," in *EMNLP*, 2019.

[16] Y. Guo, Z. Cheng, L. Nie, Y. Liu, Y. Wang, and M. Kankanhalli, "Quantifying and alleviating the language prior problem in visual question answering," in *SIGIR*, 2019.

[17] C. Jing, Y. Wu, X. Zhang, Y. Jia, and Q. Wu, "Overcoming language priors in vqa via decomposed linguistic representations," in *AAAI*, 2020.

[18] D. Teney, E. Abbasnejad, and A. v. d. Hengel, "Unshuffling data for improved generalization," in *arXiv preprint arXiv:2002.11894*, 2020.

[19] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and yang: Balancing and answering binary visual questions," in *CVPR*, 2016.

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

[21] G. KV and A. Mittal, "Reducing language biases in visual question answering with visually-grounded question encoder," in *arXiv preprint arXiv:2007.06198*, 2020.

[22] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *ICLR*, 2020.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *JAIR*, 2002.

[25] C. Drummond, R. C. Holte *et al.*, "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on learning from imbalanced datasets II*, 2003.

[26] X.-S. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *TIP*, 2019.

[27] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *CVPR*, 2019.

[28] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *CVPR*, 2019.

[29] Z. Zhao, Z. Zhang, X. Jiang, and D. Cai, "Multi-turn video question answering via hierarchical attention context reinforced networks," *TIP*, 2019.

[30] J. Ji, C. Xu, X. Zhang, B. Wang, and X. Song, "Spatio-temporal memory attention for image captioning," *TIP*, 2020.

[31] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, and T.-S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," *TIP*, 2018.

[32] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *CVIU*, 2017.

[33] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *ICCV*, 2015.

[34] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," in *arXiv preprint arXiv:1704.03162*, 2017.

[35] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *ICCV*, 2017.

[36] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *CVPR*, 2016.

[37] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *CVPR*, 2016.

[38] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "Fvqa: Fact-based visual question answering," *TPAMI*, 2018.

[39] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*, 2007.

[40] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, 2004.

[41] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *IJCV*, 2017.

[42] J. Lei, L. Yu, M. Bansal, and T. Berg, "Tvqa: localized, compositional video question answering," in *EMNLP*, 2018.

[43] Z. Zhao, Q. Yang, D. Cai, X. He, Y. Zhuang, Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang, "Video question answering via hierarchical spatio-temporal attention networks." in *IJCAI*, 2017.

[44] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016.

[45] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," 2008.

[46] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[48] D. Teney, P. Anderson, X. He, and A. Van Den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *CVPR*, 2018.

[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.

[50] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *EMNLP*, 2014.

[51] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *EMNLP*, 2016.

[52] M. Malinowski, C. Doersch, A. Santoro, and P. Battaglia, "Learning visual question answering by bootstrapping hard attention," in *ECCV*, 2018.

[53] Q. Zhou, B. Zhong, X. Lan, G. Sun, Y. Zhang, B. Zhang, and R. Ji, "Fine-grained spatial alignment model for person re-identification with focal triplet loss," *TIP*, 2020.