Exploring deep learning for multimodal understanding

Lao, M.

Chapter 4

A Language Prior Based Focal Loss for Visual Question Answering

In the previous chapter, we have studied the VQA research on the architecture side to improve to model performance. In this chapter, we move to concentrate on the language bias issue, which severely hinder the interpretability and robustness of current VQA algorithms. Specifically, we attempt to solve the **RQ2** by proposing a simple yet effective loss scaling function according to the bias-only model.

According to current research, one of the major challenges in Visual Question Answering (VQA) models is the overdependence on language priors (and neglect of the visual modality). VQA models tend to predict answers only based on superficial correlations between the first few words in question and frequency of related answer candidates. To address this issue, we propose a novel Language Prior based Focal Loss (LP-Focal Loss) by rescaling the standard cross entropy loss. Specifically, we employ a question-only branch to capture the language biases for each answer candidate based on the corresponding question input. Then, the LP-Focal Loss dynamically assigns lower weights to biased answers when computing the training loss, thereby reducing the contribution of more-biased instances in the train split. Extensive experiments show that the LP-Focal Loss can be generally applied to common baseline VQA models, and achieves significantly better performance on the VQA-CP v2 dataset, with an overall 18% accuracy boost over benchmark models.

This chapter is based on the published conference paper:

- **Lao, M.**, Guo, Y., Liu, Y. and Lew, M. S. "A Language Prior based Focal Loss for Visual Question Answering." IEEE International Conference on Multimedia and Expo, 2021.

4.1 Introduction

Visual Question Answering (VQA) [30] is an attractive multi-modal deep learning task, which aims to automatically answer natural language questions according to visual scenes. It is useful in diverse applications from medical assistance to robot tutors. With the application of various attention mechanisms [62] and multi-modal fusion strategies [267], many VQA models [68, 74, 191] achieve promising performance in current benchmark datasets [30, 48]. However, many recent studies [48, 85] point out that current VQA models are heavily suffering from the problem of language priors. Specifically, VQA models tend to rely on the superficial correlations between the patterns of questions (first few words) and answer candidates (the frequency of each answer). As a result, these models often blindly select answer without considering the visual content. For instance, as shown in the Fig. 4.1, the question "what color is the sky?" is mostly answered with "blue" in the train split. It causes the VQA models to overwhelmingly answer "blue" for this question in the test set, and neglect the analysis of visual information. This undesirable behavior restricts the generalization of existing VQA models, and limits their applicability in practical scenarios.

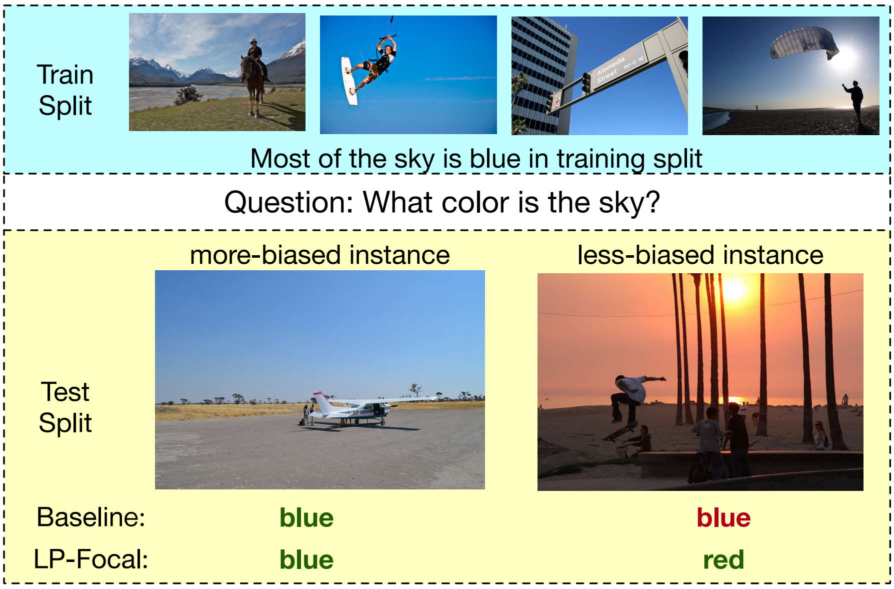

Figure 4.1: The illustration of the language priors problem in the VQA task. Existing VQA models tend to predict answers relying on spurious correlations between questions ("what color is the sky?") and prior answers ("blue") in train split. Consequently, they suffer a serious performance degradation when testing less-biased instances whose answers ('red') are not amongst the majority answers in the train split. Our LP-Focal Loss can benefit VQA models in alleviating this problem.

Recently, many approaches are proposed to alleviate this problem. Apart from developing a more balanced dataset [48], related works can be roughly divided into two categories: visual grounding based approaches and language prior based approaches. The first category [92] [105] exploits external information such as human explanations to strength visual grounding in VQA models. The second category [89, 101, 106, 107, 243, 268] aims to design efficient models and learning strategies to capture and reduce language priors in a dataset. As the external knowledge and data are too expensive to collect, compared with visual grounding approaches, language prior based methods have become more attractive and accessible.

One mainstream solution in the language prior based methods is to build an additional neural network branch to predict answers only based on the textual modality. In this manner, the VQA model can explicitly exploit the language priors, and thus address this problem. One challenge in this direction is how to design a learning strategy to exclude language priors captured from the question-only branch. Ramakrishnan et al. [89] proposed an adversarial learning method by minimizing the performance of the question-only branch, while maximizing that of the image-question branch. This adversarial learning strategy discourages the VQA model from capturing language biases in its question encoding, and promotes the contribution of the visual modality in question answering. Another effective learning strategy is fusion-based methods (e.g. RUBI [106] and LM [107]). In this strategy, two outputs (predicted answer distributions) of the VQA model and the question-only branch are merged together, and then the fused output acts as the final prediction of VQA model in training. It effectively prevents VQA models from making use of bias for answer prediction.

In this chapter, we introduce a novel Language Prior based Focal Loss (LP-Focal Loss) to overcome language priors captured from the question-only branch. The loss function is a dynamic reweighting cross entropy loss, which distributes distinct weights over each answer candidate for computing the loss. As for the reweighting factor, we intuitively employ the predicted results from the question-only branch as the evidence to down-weight the contributions of more-biased answer candidates, and further concentrate on the answers with fewer language priors. Additionally, similar to the focal loss [269] in object detection, we add a tunable focusing parameter to flexibly adjust the contribution gap between more-biased and less-biased answers. Importantly, the LP-Focal Loss is a generic approach which can be adapted to different datasets with various levels of biases. In Fig. 4.1, "blue" is a biased answer candidate for question "what color is the sky?", which has higher expectation predicted by the question-only branch. Consequently, the aforementioned question with an image about blue sky is considered as a more-biased training instance (VQA model can easily benefit from textual modality). For the standard cross entropy loss, all answers are treated equally to compute the training loss, which leads to a performance decline for testing less-based instances (pictures with "red" sky). However, when applying the LP-Focal Loss, the reweighting factor can significantly reduce the loss weight for prior answer "blue", and further reduce the backpropagated gradients for the more-biased instances.

To validate the effectiveness and generalizability of the LP-Focal Loss, we implement extensive ablation studies on three widely-used and well-performed VQA baseline models (UpDn [191], BAN [68] and MCAN [74]). From experimental results, the LP-Focal Loss outperforms baseline models with a huge gain of 18%, and achieves superior performance on smaller training sets. Finally, we compare the LP-Focal Loss with competitive approaches for reducing language priors, and our method achieves state-of-the-art performance (58.45%) on the VQA-CP v2 dataset.

4.2 Methodology

4.2.1 Preliminary

The purpose of VQA models is to accurately answer textual questions based on given images. We represent a VQA dataset with N training instances as $S=\{v_i,q_i,a_i\}_{i=1}^{N}$, where $v_i\in V$ and $q_i\in Q$ are the visual and question input in the i^{th} instance, while $a_i\in A$ is the corresponding ground truth answer in the answer dictionary. The VQA model aims to leverage image-question inputs and learn a fusion function $f:Q\times V\rightarrow R^A$ for generating a predicted distribution over the answer label space. We format the equation of it as:

$$P\left(A\mid v_i,q_i\right)=\sigma\left(f\left(v_i,q_i;\theta\right)\right), \tag{4.1}$$

where $\sigma(\cdot)$ implies the softmax activation function and θ is the learning parameter in VQA model. Then the model could be trained with standard cross-entropy, optimizing parameters to minimize the Equ. (4.2) over the correct answer labels.

$$\text{Loss}_{\ ce}=-\frac{1}{N}\sum_{i}^{N}\log\left(P\left(A\mid v_i,q_i\right)\right)\left[a_i\right]. \tag{4.2}$$

However, in many widely-used VQA datasets [30], the ground truth answer for each question-image input is not unique. Consequently, the cross entropy loss for multi-label classification can be rewritten as the Equ. (4.3):

$$\text{Loss}_{ce}=-\frac{1}{N}\sum_{i}^{N}\sum_{j}^{|A|}\left[a_{ij}^{*}\log\left(P\left(a_{ij}\mid v_i,q_i\right)\right)\right], \tag{4.3}$$

where a_{ij} implies the j^{th} answer candidate for the i^{th} VQA example, and a_{ij}^{*} is its value of the label.

4.2.2 LP-Focal Loss

Building a question-only branch to capture language priors is the prerequisite to employ the LP-Focal loss to reduce priors in VQA. As shown in Fig. 4.2(a), similar

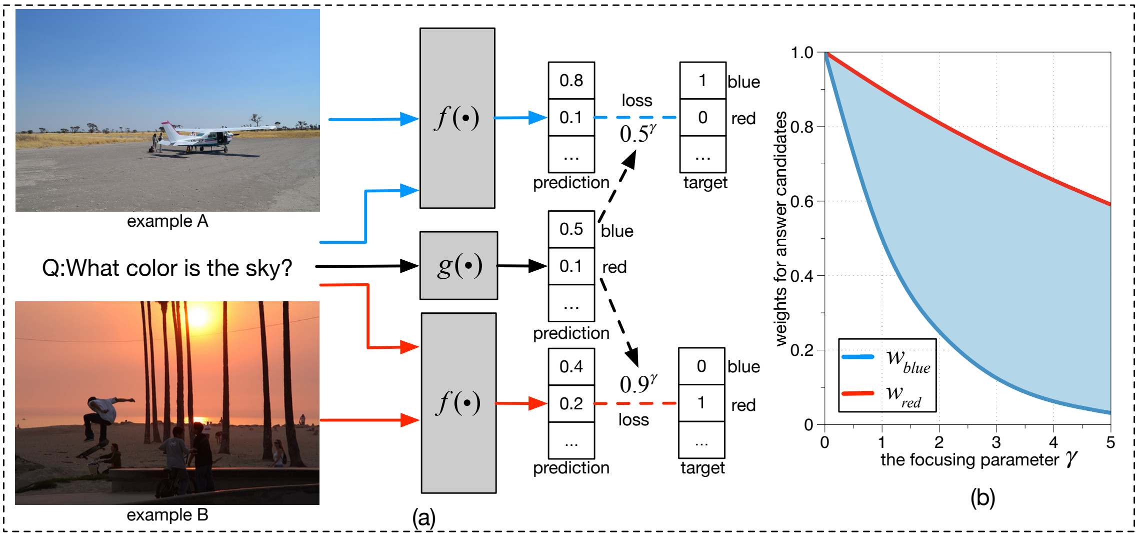

Figure 4.2: (a) Detailed illustration of the LP-Focal Loss for reducing language priors in the VQA model. Given two training examples with the same question, LP-Focal Loss effectively adjusts weights between more-biased and less-biased examples. (b) depicts loss weights for two answer candidates with variant settings of γ, where the region in light blue reveals the gap of weights between two answers.

to other question-only branch based approaches [106, 107], we additionally use a single-branch neural network to obtain the biased prediction only on the basis of question features:

$$P\left(A \mid q_i\right)=\sigma\left(g\left(q_i ; \phi\right)\right), \tag{4.4}$$

where $g(\cdot)$ and ϕ denote the mapping function and parameters of the question-only branch.

The standard cross entropy loss in the VQA model equally computes loss for each training instance. In this paper, we endow the cross entropy loss with the capability to adjust weights for examples with diverse levels of language priors. Given an image-question pair, we intuitively introduce a reweighting factor for each answer candidate. It could effectively decrease the influence of more-biased examples, and further prevent learning parameters updating sharply from language priors. The reweighting factor $\boldsymbol{w}_{ij}$ for the i^{th} example in the j^{th} answer candidate is obtained as via:

$$w_{ij}=\left(1-P\left(a_{ij} \mid q_i\right)\right)^\gamma, \tag{4.5}$$

where γ is the focusing parameter. Similar to the standard focal loss [269], this parameter smoothly adapts the rate at which the prior answer candidates are down weighted. Note that if the focusing parameter $\gamma=0$, the loss function would degenerate to vanilla cross entropy loss. When a training instance strongly suffers from language priors, the $P\left(a_{ij} \mid q_i\right)$ for its ground truth answer is large, and w_{ij} is close to 0. Hence, the reweighting factor w_{ij} can assign less weights to the more-biased instances, and promote less-biased instances.

Then we propose the LP-Focal loss by combining the reweighting factor with the cross entropy loss for multi-label classification, which is defined as:

$$\text{Loss}_{lpf} = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{N_d}\left[w_{ij}a_{ij}\log\left(P\left(a_{ij} \mid v_i, q_i\right)\right)\right] \tag{4.6}$$

For the model architecture to implement the LP-Focal Loss in the Fig. 4.2(a), as our method is model agnostic and could be applied to variants VQA models, the framework to achieve the function $f(\cdot)$ can be any classification-based VQA models. The question-only branch $g(\cdot)$ consists of a LSTM with a single-layer fully connected layer to process the summation of word-level question features. Then it is followed by a two-layer classifier network for answer prediction, which is similar as that in the UpDn [191] model.

Fig. 4.2 depicts how the LP-Focal Loss adjusts loss weights for different examples. The example A is a more-biased example whose correct answer "blue" obtains a predicted result of 0.5 from the question-only branch, while example B is the less-biased example (0.1 for "red"). On the basis of the prediction from the question-only branch $P(A \mid q_i)$, the LP-Focal Loss distributes weights 0.9^{γ} and 0.5^{γ} for answer candidates "blue" and "red" respectively. From the Equ. (4.6), if the correct answer is unique for aforementioned two examples, the weights for two answer candidates can also represent the loss weights for two training examples. As shown in the Fig. 4.2(b), with the increase of γ, the gap of loss weights (the region in light blue) between less-biased and more biased example is significantly enlarged, which benefits VQA model in alleviating the influence of training instances with strong biases, and focusing on less-based instances.

LP-Focal Loss vs Focal Loss: The difference between standard Focal Loss and LP-Focal loss is the reweighting factor. When employed in VQA models, the reweighting factor of the Focal Loss w'_{ij} is defined as:

$$w'_{ij} = \left(1 - P\left(a_{ij} \mid v_i, q_i\right)\right)^{\gamma} \tag{4.7}$$

Since the Focal Loss aims to focus on hard examples, w'_{ij} is used to assign weights based on the difficulty of image-question inputs $P(a_{ij} \mid v_i, q_i)$. On the contrary, the LP-Focal Loss focuses on the less-biased examples, and its reweighting factor $\boldsymbol{w}_{ij}$ is on the basis of the language priors captured from the question-only branch $P(a_{ij} \mid q_i)$.

Table 4.1: The effect of the focusing parameter γ on the VQA-CP v2 test split.

UpDn	Overall	Other	Number	Yes/No
$\gamma = 0$	40.56	47.70	12.30	41.75
$\gamma = 1$	55.89	49.43	17.54	88.28
$\gamma = 2$	56.81	**49.97**	20.79	**88.73**
$\gamma = 3$	57.90	49.96	29.24	88.07
$\gamma = 4$	**58.45**	49.32	**34.67**	88.34
$\gamma = 5$	57.14	47.45	33.42	88.04

BAN	Overall	Other	Number	Yes/No
$\gamma = 0$	40.69	46.64	13.66	43.49
$\gamma = 1$	56.79	48.92	22.83	89.57
$\gamma = 2$	57.81	**49.45**	28.77	**88.97**
$\gamma = 3$	58.04	49.20	31.90	88.59
$\gamma = 4$	**58.39**	48.64	**37.41**	87.99
$\gamma = 5$	57.88	48.42	36.54	87.11

4.3 Experiments

4.3.1 Datasets

We evaluate our LP-Focal Loss mainly on the VQA-CP v2 dataset [85]. It is the most commonly used benchmark for validating the robustness of VQA models in reducing language priors, as the distribution of answers per question type varies significantly between the train and test splits. We also utilize the VQA-v2 dataset to train and evaluate our method on the benchmark VQA model for completeness. To validate the generalizability of our proposed loss function, we implement the LP-Focal Loss on three well-performed benchmark VQA models: UpDn [191], BAN [68], and MCAN [74].

4.3.2 Implementation details

As for network details, we use the pre-trained Faster R-CNN to extract object-level image features with no more than 100 proposals with their 2048-d features. We use a Glove to encode the question as a word-level vector, and set the max length of words in question is 14. In general, for the compared VQA models, we used the same setting as in the original papers. For training details, we pretrain the VQA model and the question-only branch for 10 epochs with the same setting in original papers, and then we train the model with our proposed LP-Focal Loss. After 10 epochs, the learning rate is decayed by 1/2 for every 5 epoch up to 20 epochs and stop training. The initial learning rate is 1e-3 for UpDn and BAN models and 1e-4 for the MCAN model. We set the mini-batch to 256 for UpDn and BAN, and 64 for MCAN model.

4.3.3 Ablation studies

In this subsection, we carry out the ablation studies on three benchmark models (UpDn, BAN and MCAN) to examine contributions of focusing parameter γ and question-only branch in the LP-Focal loss. In addition, to further demonstrate the superiority and generalizability of the our method, we implement the LP-Focal loss on different proportions of the training split. We train all models on the VQA-CP v2 train split, and evaluate on its test split.

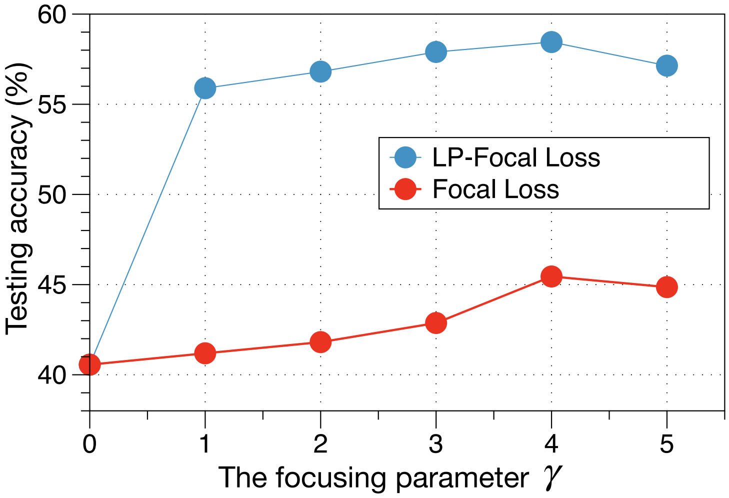

Figure 4.3: The comparison of the standard Focal Loss and our LP-Focal loss under different settings of focusing parameters γ on the VQA-CP v2 test split.

Effect of the focusing parameter: As depicted in the Tab. 4.1, with different settings of the focusing parameter ($\gamma > 0$), our LP-Focal loss achieves a significant accuracy boost over the benchmark cross entropy loss ($\gamma = 0$). To be specific, with the γ rising from 1 to 4, all three tested models yield better results. Notably, the accuracy of questions in "number" type increases dramatically. This is due to the fact that, compared with question types "yes/no" and "other", the gap of language priors between more-biased instances and less-biased instances is not too much. Consequently, adding the value of γ could effectively benefit models to further down-weight biased examples, and focus on less-biased examples. In general, the optimal setting for focusing parameter is $\gamma = 4$ in this dataset, as larger value (e.g. $\gamma = 5$) may fail to fulfill further improvements.

Effect of the question-only branch: In the Tab. 4.2, we demonstrate the importance of the question-only branch in our method. Removing the training loss of the single-branch from the LP-Focal Loss ($\gamma = 4$) would seriously impair model's performance. It verifies that adding a question-only neural network can effectively capture the language biases in the dataset, which delivers crucial information for the LP-Focal Loss to reduce biases.

Smaller training splits: As shown in the Tab. 4.3, from results of variant percentages of training splits, our method obtains obviously superior performance over three benchmark models, with average improvements of 18%. In particular, with 20% training data, the LP-Focal Loss can still effectively overcome language priors captured from limited data, and further achieve overall 17% accuracy boost. These results further prove the effectiveness of the LP-Focal loss, and the potential capacity to perform well in limited datasets.

Table 4.2: The ablations of the training loss of question-only branch on the VQA-CP v2 test split. L_q is whether we implement our LFP-Loss with the loss of question-only branch.

Model	L_q	Overall	Other	Number	Yes/No
UpDn+Ours	✓	58.45	49.32	34.67	88.34
	×	41.88	48.36	13.41	44.40
BAN+Ours	✓	58.39	48.64	37.41	87.99
	×	42.12	47.50	13.43	46.86
MCAN+Ours	✓	59.33	49.22	43.10	87.11
	×	43.74	48.67	16.57	48.56

4.3.4 Compared with the standard focal loss

In this part, we make comparisons between our method with the standard Focal Loss. Theoretically, Focal Loss aims to dynamically assign more weight to the hard examples, while our LP-Focal Loss is to down-weight for more-biased examples for reducing language priors. We implement the Focal Loss on the UpDn model and the results are shown in the Fig. 4.3. Compared with the benchmark cross entropy loss ($\gamma = 0$), the Focal Loss could obtain better results. Both Focal Loss and our method show growth trends when the focusing parameter increases from 1 to 4, whereas our LP-Focal loss is significantly superior to the Focal Loss to achieve better performance. All these results indicate that focusing on hard negative examples (such as Focal Loss) has positive effect on overcoming priors, but it is is not sufficient to achieve remarkable performance like our LP-Focal Loss.

4.3.5 Comparison with the state-of-the-art

In this subsection, we compare the LP-Focal Loss with state-of-the-art debiasing methods in the VQA task. We roughly divided these approaches into two categories: visual grounding based approaches and language prior based approaches. Visual grounding based methods like HINT [92] and SCR [105] employ human explanations to increase the visual grounding for VQA. The language priors based approaches often design delicate models and learning strategies to remove language priors, including AdvReg [89], RUBi [106] and LM [107]. As the UpDn model is the mostly used baseline model for VQA, we list the performance of state-of-the-art approaches implemented on the UpDn model in the Tab. 4.4.

Compared with the baseline UpDn model on the VQA-CP v2 dataset, our LP-Focal Loss significantly improves the overall accuracy, and outperforms other question-only branch based approaches [89, 106, 107]. Furthermore, our method is also superior to the SSL [243] approach by nearly 1%. From the results of different question types, the LP-Focal loss achieves a remarkable performance (88.34%) on the “Yes/No” question type. In addition, on the hardest "Number" question type, compared with the baseline model, the our method could still fulfill a great accuracy boost from

Table 4.3: Results of the LP-Focal Loss on the VQA-CP v2 test set with different proportions of training split.

Per Model	20%	40%	60%	80%	100%
UpDn	34.91	36.69	38.02	40.00	40.56
UpDn+Ours	52.26	54.86	56.69	57.29	58.45
BAN	35.59	37.00	37.96	39.88	40.69
BAN+Ours	52.54	55.31	57.11	57.78	58.39
MCAN	36.62	37.85	39.01	40.58	41.10
MCAN+Ours	53.53	55.78	57.67	58.53	59.33

11.93% to 34.67%. These results strongly demonstrate that our method can effectively exclude language priors, and further benefit VQA models to achieve unbiased reasoning for multi-modal features. Apart from the VQA-CP v2 dataset, we also evaluate the approaches on the VQA v2 dataset, where the language priors in the train split are beneficial to the performance in the val split. From results on this dataset, most methods perform worse than the base model, and our method is also suffered from the accuracy drop to 62.45%. It can be explained that the LP-Focal loss is to reduce the language priors, which would avoid VQA models benefiting from data biases in the train split to improve the performance on the val split.

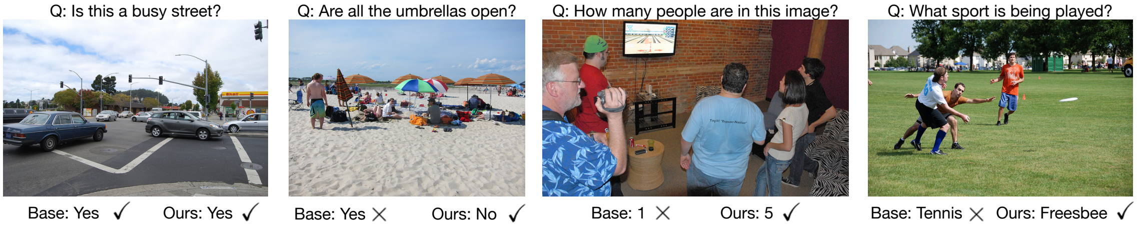
Q: Is this a busy street?
Base: Yes ✓
Ours: Yes ✓
Q: Are all the umbrellas open?
Base: Yes ×
Ours: No ✓
Q: How many people are in this image?
Base: 1 ×
Ours: 5 ✓
Q: What sport is being played?
Base: Tennis ×
Ours: Freesbee ✓

Figure 4.4: The examples of case study on the VQA-CP v2 dataset.

4.3.6 Case study

In this section, we make case studies to qualitatively demonstrate the superiority of the LP-Focal loss in the Fig. 4.4. The predicted results derived from the baseline UpDn model and the model with our method ($\gamma = 4$). These examples cover a broad range of the answer types, including Yes/No, Number and Other. For the question type Yes/No, "Yes" is the answer with relatively more priors than other answers. Consequently, in the first two examples, the baseline model overwhelmingly predicts "Yes" as the correct answer, which would lead to a decline in performance on the unbiased VQA dataset. For the same reason, in the last two cases, the baseline model directly selects "1" and "Tennis" as the predicted results, because they are the most-biased answers under questions "How many" and "What sport". On the contrary, by exploiting our LP-Focal loss, the VQA models can effectively avoid overfitting for data biases, and show better results on unbiased dataset.

Table 4.4: Comparisons with state-of-the-art approaches based on the benchmark UpDn models. Results are tested on the VQA-CP v2 test split and the VQA v2 val split.

Model	VQA-CP v2				VQA v2
	Overall	Other	Number	Yes/No	Overall
UpDn [191]	39.74	46.05	11.93	42.27	63.48
visual grounding based approaches					
AttAlign [92]	39.37	45.00	11.89	43.02	63.24
HINT [92]	46.73	45.88	10.61	67.27	63.38
SCR [105]	49.45	48.02	10.93	72.36	62.20
language prior based approaches					
AdvReg [89]	41.17	35.48	15.48	65.49	62.75
RUBI [106]	44.23	39.61	17.48	67.05	-
LangAtt [268]	48.87	45.57	18.72	70.99	57.96
VGQE [101]	48.75	-	-	-	**64.04**
LM [107]	48.78	45.58	14.61	72.78	63.26
LM+H [107]	52.01	46.97	31.12	72.58	56.35
SSL [243]	57.59	**50.03**	29.87	86.53	63.73
LP-Focal	**58.45**	49.32	**34.67**	**88.34**	62.45

4.4 Conclusion

In this chapter, we proposed a novel Language Priors based Focal Loss (LP-Focal Loss) to address the language priors problem in the VQA task. Specifically, our method exploited language priors captured by a question-only branch, and further dynamically assigned weights for different training instances. Extensive experiments verified the effectiveness and generalizability of the LP-Focal Loss, and it achieved state-of-the-art performance on the VQA-CP v2 dataset.

Future Works: we attempt to improve the LP-Focal loss from the aspect of algorithm and application. On the one hand, we will explore the different predictive uncertainty for different question types, thereby proposing an instance-level strategy to assign different focal parameters for different training samples. On the other hand, we plan on refining the LP-Focal loss and use it on other multi-modal deep learning tasks with unimodal biases, such as scene graph genration and video groudning.