Exploring deep learning for multimodal understanding

Lao, M.

Chapter 5

From Superficial to Deep: Language Bias driven Curriculum Learning for Visual Question Answering

In the previous chapter, we proposed a loss-objective based approaches to alleviate language bias. However, under the severe out-of-distribution settings where train and test distributions are entirely different, the loss objective approach is not sufficient to obtain significant anti-bias performance. Motivated by the human cognition process, we propose a novel curriculum-learning based framework to alleviate bias and improve out-of-distribution performance, which corresponds to **RQ3**.

In this chapter, we overcome the language prior problem by proposing a novel Language Bias driven Curriculum Learning (LBCL) approach, which employs an easy-to-hard learning strategy with a novel difficulty metric Visual Sensitive Coefficient (VSC). Specifically, in the initial training stage, the VQA model mainly learns the superficial textual correlations between questions and answers (easy concept) from more-biased examples, and then progressively focuses on learning the multimodal reasoning (hard concept) from less-biased examples in the following stages. The curriculum selection of examples on different stages is according to our proposed difficulty metric VSC, which is to evaluate the difficulty driven by the language bias of each VQA sample. Furthermore, to avoid the catastrophic forgetting of the learned concept during the multi-stage learning procedure, we propose to integrate knowledge distillation into the curriculum learning framework. Extensive experiments show that our LBCL achieves remarkably better performance on the VQA-CP v1 and v2 datasets, with an overall 20% accuracy boost over baselines.

This chapter is based on the following publication:

- **Lao, M.**, Guo, Y., Liu, Y., Chen, W., Pu, N., and Lew, M. S. "From Superficial to Deep: Language Bias driven Curriculum Learning for Visual Question Answering." ACM International Conference on Multimedia, 2021.

5.1 Introduction

Vision and language are two fundamental modalities for human beings to explore and understand the real world. Benefiting from the tremendous progress of deep learning [270], Visual Question Answering (VQA) [30] at the intersection of computer vision and natural language processing has become an attractive research over the past few years. VQA is considered as an 'AI-complete' task that enables machines to answer a natural language question about a given image. With the significant development of techniques such as attention mechanisms [62] and multimodal fusion strategies [267], current VQA models have shown rapid performance boost on common benchmark datasets [30, 48], especially the datasets where the train and test splits have similar answer distributions.

Despite the impressive performance improvement, recent works [85, 271] have pointed out that, a majority of VQA models are suffering heavily from the **language bias** (prior) problem caused by the dataset itself. Concretely, these models can easily predict correct answers relying on the statistical co-occurrence patterns between the given question and the prior answer candidate, instead of combining the image to make an inference, such as overwhelmingly answering '*what sport is*' as '*tennis*' or '*what color is the sky*' as '*blue*'. Initially, VQA models are designed to achieve high-level understanding of both visual scene and textual question, and predict correct answer by exploiting multimodal information jointly and comprehensively. However, the overdependence on the textual modality for question answering is not consistent with the intention of VQA task, which severely limits the generalization and applicability of VQA models.

The language bias problem is mainly derived from the label imbalance for a question type in the train split, which also inherently exists in the real world. For instance, if 80% of the bananas are yellow in the train set, VQA models would easily achieve high training accuracy by overwhelmingly selecting the prior answer '*yellow*' for the question '*what color is the banana?*'. Hence, the VQA example about '*yellow banana*' is considered as a more-biased instance, which heavily prevents the model from learning new and adequate knowledge from the visual data. On the contrary, the training for less-biased instance like '*green banana*' can hardly obtain benefits from language bias, and they require more visual reasoning oriented by textual question to achieve better performance for selecting the correct answer. Therefore, it is necessary to develop a learning strategy to leverage various training instances with different levels of language prior, and further overcome the inherent data biases.

In this chapter, we explore the idea of Curriculum Learning (CL) [272] for unbiased VQA models, and propose a novel **Language Bias driven Curriculum Learning** (LBCL) method. CL aims to embody the cognitive process of human being on the training of machine learning models. Its core idea is to initially train with the easier examples, and then progressively focus on the hard examples in accordance with

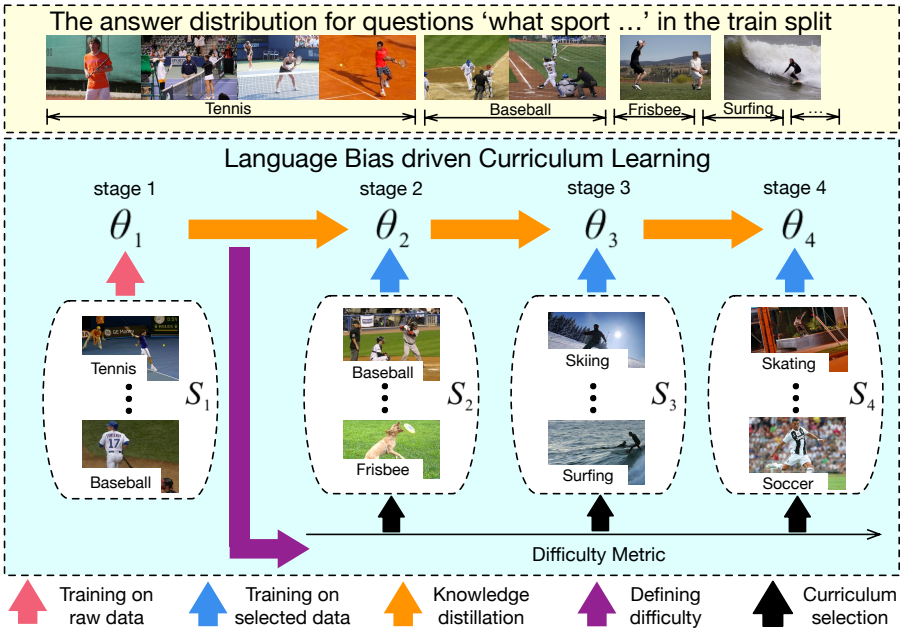

Figure 5.1: Conceptual illustration of our Language Bias driven Curriculum Leaning (LBCL) for VQA. The yellow box depicts the imbalanced answer distribution for the training of VQA model. We update the parameters θ of VQA model in accordance with an easy-to-hard learning strategy on the selected dataset S_i with a desired curriculum difficulty.

a pre-defined training criteria. In our LBCL, on the one hand, the easy concept refers to the **superficial textual correlations** between questions and answers, which could be easily captured from more-biased training instances (e.g. answer '*the color of banana*' with '*yellow*'). Although these spurious correlations play negative roles on the joint reasoning of multimodal information, they still contain some basic elements for answer prediction (e.g. the intention of question and related answer candidates for a question type), which would be beneficial for VQA models to narrow the answer space in their initial learning phase. On the other hand, the hard concept is the **deep multimodal reasoning** that adequately exploits both visual and textual modalities. We declare that the hard concept is mainly acquired from the training of less-biased examples (e.g. '*green banana*'), as they could not regularize the training process with lower errors based on textual correlations, and require more visual dependence for question answering.

The conceptual illustration of our LBCL is shown in Fig. 5.1. We take VQA instances related to the question type about 'what sport' for example. As depicted in the yellow box about answer distribution, 'tennis' is the most prior answer for given question, while the proportions of some answers (e.g. '*frisbee*' and '*surfing*') are not sufficient in the train split. In the first stage of the LBCL, we train the VQA model on the original training data. It is considered as the easy learning stage, because the training is dominated by the prior ground truth answers, and the VQA model tends to learn more about the textual correlation between the question and

dominating answers (the question '*what sport*' can be probably answered by '*tennis*' or '*baseball*' as well). Meanwhile, on the basis of the prediction in the first stage, we can define the difficulty metric for curriculum learning based on a novel **Visual Sensitive Coefficient** (VSC), which is proposed to evaluate a VQA sample about how much the benefit obtained from visual modality under severe language bias. In the following stages, we achieve the easy-to-hard transition by a curriculum selection function to select examples with a desired difficulty. As the training stage proceeds, the learning gradient for more-biased examples would be excluded, and VQA model turns to concentrate on less-biased examples ('*skating*' and '*soccer*'). Furthermore, we propose to incorporate knowledge distillation with the curriculum learning for unbiased VQA, so as to alleviate the catastrophic forgetting problem through the easy-to-hard learning.

To the best of our knowledge, this is the first attempt to introduce the curriculum learning for overcoming language bias in VQA task. Our LBCL module is generic and model-agnostic, which can be applied to various VQA models. In this paper, we demonstrate the capacity of LBCL to alleviate the language prior on three recent top-performing baseline models, i.e. UpDn [191], BAN [68] , and MCAN [74]. To prove the generalizability and effectiveness of our LBCL method, we carry out extensive ablation studies on VQA-CP v2 and v1 datasets, which are specially built to assess the robustness of VQA models under severe language prior. Experimental results show that our approach surpasses the baseline models by a significant gain of 20%, and also obtains superior performance on the small-scale training datasets of VQA-CP v1 and v2 described in Appendix. Finally, we compare LBCL with state-of-the-art debiasing strategies, and our method achieves remakably better performance on the VQA-CP v2 (60.74%) and v1 (61.57%) datasets.

5.2 Related work

Visual Question Answering: In most VQA models, the attention mechanism and multi-modal fusion are two essential techniques to boost the performance. The attention mechanism aims to measure relevant image regions or objects with different importances based on the give question. It builds a crucial bridge for joint reasoning between multimodal features. Some remarkable approaches (e.g. SAN [66], UpDn [191], BAN [68], and MCAN [74]) significantly enhance accuracy and interpretability of VQA models. Multimodal fusion is to achieve high-level and complex interactions between visual and textual features for answer prediction. Most state-of-the-art VQA models employ the advanced bilinear pooling approaches to fulfill effective second-order interactions with less resources, such as MLB [81], MCB [273], MUTAN [83], BLOCK [274], and MHEF [275].

Language Bias in VQA: In order to better evaluate the language bias in current VQA models, Agrawal et al. [85] propose a new dataset VQA-CP for unbiased training, where the distributions of answers per question between train and test

splits are quite different. There are also many great efforts to overcome this problem, and we roughly divided these debiasing strategies into four categories: 1) The annotation-based methods [92, 105] assume that the image feature is not sufficient to tackle the VQA task, and attempt to utilize extra visual and textual annotations to strength the visual grounding for VQA. 2) The fusion-based approaches [106, 107] are to train VQA model with a question-only model, and combine the outputs of two models as the final prediction for training. It effectively removes the excessive dependence of language prior for VQA models. 3) The data rebalance-based methods [94, 110, 232, 236, 243] attempt to propose some data augmentation strategies to generate counterfactual training instances automatically, thereby balancing the answer distribution of training data. 4) The other methods: there are also many impressive works to overcome language bias through adversarial learning [89], modifying language module [101, 268], and casual inference [88]. Among most of these debiasing methods [88, 89, 106, 107], the question-only (unimodal) branch is crucial for capturing spurious relationships between questions and answer candidates. In our LBCL method, it is also an essential component for quantifying the difficulty metric Visual Sensitive Coefficient (VSC) of each training instance in curriculum learning.

Curriculum Leaning: Bengio et al. [272] first propose the concept of curriculum learning (CL), which formalizes an easy-to-hard strategy for training machine learning models inspired by human brains. One character in CL is that the definition of the difficulty metrics is task-specific. CL has been extensively exploited to tackle various computer vision [276, 277] and natural language processing tasks [278, 279, 280]. CL can also be combined with other machine learning methods to solve the corresponding problems better, such as meta learning [281], reinforcement learning [282] and active learning [283]. Based on the difficulty metric for CL, the common strategies to select sub training dataset with a proper difficulty level are sampling [284], weighting [285, 286] and batching [287]. In our LBCL, we employ weighting method to dynamically distribute binary weights on training instances for adjusting the training difficulty.

5.3 Language Bias driven Curriculum Learning

In this section, we first introduce the preliminary of VQA, and then describe the overview learning framework of our Language Bias driven Curriculum Learning (LBCL) method. In the following three subsections, we elaborate the central concepts in the LBCL, including difficulty metric, curriculum selection function and knowledge distillation. Finally, we conclude the algorithmic pipeline of LBCL.

The Paradigm of VQA Model: Given an image-question input, the classification-based VQA model aims to generate a predicting distribution for answer dictionary. A VQA dataset with N training instances is represented as $S = \{I_i, Q_i, a_i\}_{i=1}^{N}$, where $I_i \in \mathcal{I}$ and $Q_i \in \mathcal{Q}$ are the image and question input of the i^{th} instance, while

$a_i \in \mathcal{A}$ is the correct answer in the answer dictionary. The VQA model is to utilize multimodal inputs to learn a fusion function $f : \mathcal{Q} \times \mathcal{I} \rightarrow [0,1]^{|\mathcal{A}|}$ for producing a distribution over the answer space $\mathcal{A}$. The function is denoted as:

$$P\left(\mathcal{A} \mid I_i, Q_i\right) = softmax\left(f\left(I_i, Q_i; \theta\right)\right), \tag{5.1}$$

where θ implies the learning parameters in VQA model. We can train the VQA model with standard cross-entropy loss, and optimize the network parameters θ to minimize the formulation below,

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i}^{N} \log\left(P\left(\mathcal{A} \mid I_i, Q_i\right)\right)\left[a_i\right]. \tag{5.2}$$

It is worth noting that, for some commonly-used VQA datasets [30, 48], an image-question input is corresponding to many correct answers. Consequently, we can employ the soft cross-entropy loss for multi-label classification to solve this task, which is given as:

$$\mathcal{L}_{sce} = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{|\mathcal{A}|} a_{ij}^{*} \log\left(P\left(a_{ij} \mid I_i, Q_i\right)\right), \tag{5.3}$$

where a_{ij} denotes the j^{th} answer candidate in dictionary for the i^{th} training instance, and a_{ij}^{*} is its value of the ground truth label.

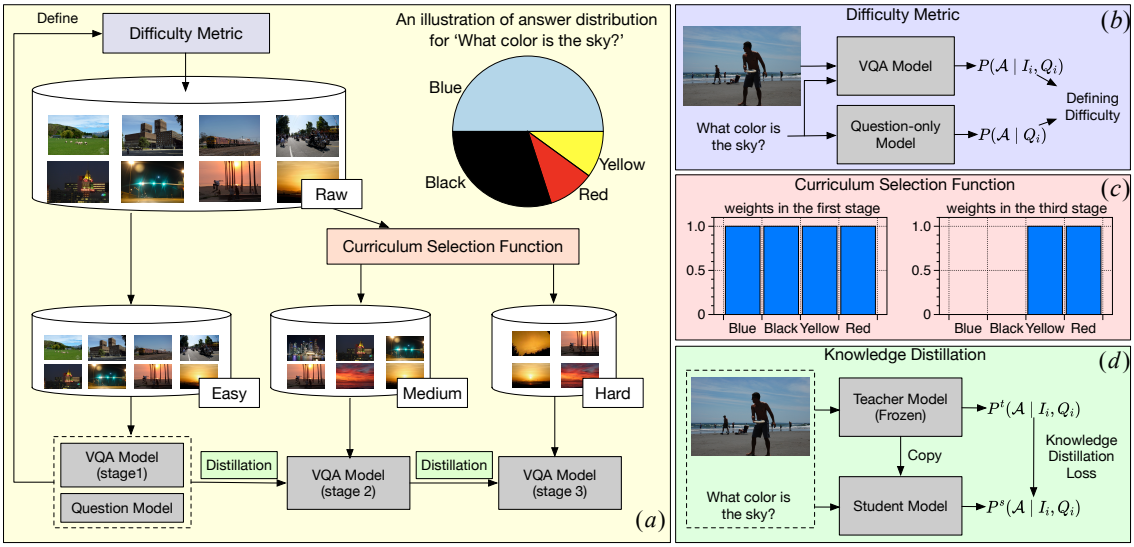

Figure 5.2: (a) Overview of our proposed LBCL method with training examples related to 'what color is the sky?' . (b) The model architecture for defining difficulty metric in the first learning stage. (c) The loss weights of answer candidates in different training stages. (d) The framework of knowledge distillation approach.

Overview Framework: LBCL aims to train the VQA model with an easy-to-hard strategy for overcoming language bias problem. The overview of LBCL architecture

is shown in Fig. 5.2(a), and we take VQA training instances related to '*what color is the sky?*' as an example to analyze our method. From the pie chart illustrated in Fig. 5.2(a), the ground truth answers for the aforementioned question are dominated by '*blue*', whereas '*red*' and '*yellow*' are the rare labels. In the first learning stage, the VQA model with an extra question-only model are trained on the raw biased dataset. During this period, VQA model is prone to learning the easier VQA concept ('*blue*' is usually used to describe the color of sky) due to the imbalanced training answer distribution. Additionally, we exploit the predictions in this stage as the prior knowledge, and further define the difficulty metric Visual Sensitive Coefficient (VSC) as the training criteria for the next multi-stage easy-to-hard training stages (the purple box in Fig. 5.2(b)).

The following training stages (from 2 to n) keep a sequence of training criteria based on the difficulty metric VSC defined in the first stage. The easy-to-hard transition is achieved by a curriculum selection function (the red box in Fig. 5.2(c)) that dynamically adjusts the loss weights of training instances with a desired difficulty. As the training proceeds, the backpropagated gradients from more-biased instances (the '*blue*' and '*black*' sky) would be removed, and the VQA model turns to learn 'hard' concepts by focusing on the less-biased instances progressively (the '*red*' and '*yellow*' sky). Furthermore, to avoid the catastrophic forgetting problem through multi-stage learning, we exploit the knowledge distillation (the green box in Fig. 5.2(d)) to transfer knowledge from the previous learning stage to next stage.

5.3.1 Difficulty Metric

Difficulty Metric is an indispensable component for curriculum learning, which acts as an evaluation indicator that represents the task-specific difficulty of a VQA example. For the curriculum learning of overcoming language bias in VQA task, we first introduce an intuitive difficulty metric based on the Training Error (TE), and further propose a novel metric Visual Sensitive Coefficient (VSC) to evaluate a VQA training instance about the visual dependence for question answering under language prior.

Training Error (TE): TE-based difficulty is derived from the intuition that more-biased examples tend to have less training error, as they can easily achieve higher performance only according to the language prior. On the contrary, less-biased samples are rare in the imbalanced VQA dataset, and are usually trained insufficiently for VQA model. Therefore, we propose a difficulty metric TE, and the difficulty TE_{ij} for the j^{th} answer candidate of the i^{th} training instance a_{ij} is formulated as:

$$TE_{ij} = \log \frac{1}{P(a_{ij} \mid V_i, Q_i)}. \tag{5.4}$$

However, the TE metric fails to reflect and measure the bias level and visual dependence of each VQA training sample.

Visual Sensitive Coefficient (VSC): To better evaluate the difficulty of examples from a view of language bias, we propose a novel difficulty metric Visual Sensitive Coefficient (VSC) for the LBCL method. In the VQA task under severe language bias, the easy concept is textual correlations between questions and answers, whereas the hard is the visual reasoning for answer prediction. Hence, the VSC is introduced to evaluate how much the VQA model predicting answers rely on the visual modality under strong language bias. Particularly, the difficulty metric VSC is defined based on the pair-wise mutual information (pmi) [288], which measures the mutual dependence between two discrete random variables. The mutual information of variables x and y is as follows:

$$pmi(x;y) \equiv \log \frac{P(x,y)}{P(x)P(y)} = \log \frac{P(y \mid x)}{P(y)}. \tag{5.5}$$

On the basis of the pmi, we further define the difficulty metric VSC based on the difference between the mutual correlation of $(\mathcal{A};(V_i,Q_i))$ and $(\mathcal{A};Q_i)$. Specifically, the VSC_{ij} for the a_{ij} is:

$$\begin{aligned} VSC_{ij} &= pmi(a_{ij};(V_i,Q_i)) - pmi(a_{ij};Q_i) \\ &= \log \frac{P(a_{ij} \mid V_i,Q_i)}{P(a_{ij})} - \log \frac{P(a_{ij} \mid Q_i)}{P(a_{ij})} = \log \frac{P(a_{ij} \mid V_i,Q_i)}{P(a_{ij} \mid Q_i)}. \end{aligned} \tag{5.6}$$

If a_{ij} is the ground truth answer for the i^{th} example, the higher VSC_{ij} value indicates that VQA model acquires more visual reasoning for training this example when predicting answers. Conversely, $VSC_{ij} < 0$ illustrates the example excessively depends on the question, and adding the image feature would impair its performance.

Discussion: The framework for defining difficulty metric in LBCL is depicted in Fig. 5.2(b). Apart from the VQA model (Equ. 5.1), similar to other debiasing strategies [88, 89, 106, 107], we additionally employ a question-only model. It is a single-branch network to obtain the biased prediction only on the basis of question features:

$$P\left(\mathcal{A} \mid Q_i\right) = softmax\left(g\left(Q_i;\phi\right)\right), \tag{5.7}$$

where $g(\cdot)$ and ϕ imply the mapping function and parameters of the question model. For multi-label classification VQA (Equ. 5.3), to encourage the biased prediction to be more non-uniform, we use the answer with largest label value as the sole ground truth a_i, and train the question-only model with the cross-entropy loss (Equ. 5.2).

5.3.2 Curriculum Selection Function

Curriculum selection function is to determine which examples should be utilized for training at the current stage. In our LBCL, we adopt weighting method to

dynamically select training examples based on the difficulty metric (VSC or TE) defined in the first stage.

The core in the selection function is to determine the real-time model competence (the difficulty threshold in the current stage) and the number of total training stages. Given the pre-defined maximum model competence C_{max} (the desired difficulty threshold in the final training stage), a difficulty increment parameter γ and the initial difficult value d_2 (first used in the second stage in LBCL), the real-time model competence $C(t)$ in the t^{th} training stage and the total training stage T are formatted as:

$$C(t) = d_2 + (t-2)\gamma, \tag{5.8}$$

$$T = \lceil (C_{max} - d_2)/\gamma \rceil + 2. \tag{5.9}$$

Then, we dynamically select training examples by assigning a binary weight $w_i(t)$ for the i^{th} example on the cross-entropy loss in the t^{th} learning stage:

$$\mathcal{L}_{ce}(t) = -\frac{1}{N}\sum_{i}^{N} w_i(t) \log\left(P\left(\mathcal{A} \mid I_i, Q_i\right)\right)[a_i]. \tag{5.10}$$

Specifically, the binary weight $w_i(t)$ is determined by the comparison between the pre-defined difficulty (VSC) and the real-time model competence $C(t)$ in the t^{th} learning stage:

$$w_{ij}(t) = \begin{cases} 0, & VSC_{ij} < C(t) \\ 1, & VSC_{ij} \geq C(t) \end{cases}. \tag{5.11}$$

For multi-label classification VQA task, the binary weight $w_{ij}(t)$ is distributed to the j^{th} answer candidates of the i^{th} example:

$$\mathcal{L}_{sce}(t) = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{|\mathcal{A}|} w_{ij}(t) a_{ij}^{*} \log\left(P\left(a_{ij} \mid I_i, Q_i\right)\right), \tag{5.12}$$

where $w_{ij}(t)$ is similarly computed by the Equ. (5.11) based on VSC_{ij} and $C(t)$. Fig. 5.2(c) illustrates the change of loss weight from easy stage to hard stage. The loss weight in the first stage is evenly distributed over each candidate answer, while in the following stages the weights for answers of more-biased examples ('*blue*' and '*black*') are zeros, thereby benefiting VQA model from focusing on the hard examples ('*red*' and '*yellow*').

5.3.3 Knowledge Distillation

The LBCL method is a multi-stage strategy learning the knowledge from easy to hard. It is inevitable for VQA model to forget previous knowledge during the progressive training. To tackle this problem, we propose to combine the knowledge distillation [289] approach with the curriculum learning to overcome the forgetting issue.

The knowledge distillation framework is a teacher-student network [290] structure, which is described in Fig. 5.2(d). Specifically, the teacher model is a frozen structure pretrained from the previous stage. Before the training of current stage, the student model is copied from the teacher model, whose parameters and network structure are the same as its teacher. Then, we train the student model with the weighted VQA instances based on the curriculum selection function in corresponding stage. Meanwhile, we also feed the same training data into the detached teacher model, and further exploit its predicting results as the supervision for student model. Practically, we adopt the common Kullback-Leibler divergence constraint as distillation loss. Its formulation is defined as:

$$\begin{aligned}\mathcal{L}_{kd} &= \frac{1}{M}\sum_{i}^{M} KL\left(P^t\left(\mathcal{A} \mid I_i, Q_i\right) \| P^s\left(\mathcal{A} \mid I_i, Q_i\right)\right) \\ &= -\frac{1}{M}\sum_{i}^{M} P^t\left(\mathcal{A} \mid I_i, Q_i\right) \log \frac{P^s\left(\mathcal{A} \mid I_i, Q_i\right)}{P^t\left(\mathcal{A} \mid I_i, Q_i\right)},\end{aligned} \tag{5.13}$$

where $\mathcal{L}_{kd}$ denotes the KL distance from the prediction of teacher model $P^t\left(\mathcal{A} \mid I_i, Q_i\right)$ to the prediction of student model $P^s\left(\mathcal{A} \mid I_i, Q_i\right)$, and M is the number of samples in a mini-batch. Consequently, the total loss $\mathcal{L}_{all}(t)$ in the t^{th} training stage ($t \geq 2$) is:

$$\mathcal{L}_{all}(t) = \mathcal{L}_{sce}(t) + \lambda_d \mathcal{L}_{kd}, \tag{5.14}$$

where the λ_d is a trade-off factor applied to adjust the contributions of the loss terms between VQA loss and knowledge distillation loss.

Algorithmic Pipeline: Based on the aforementioned crucial components in LBCL, the detailed descriptions about how our method works are summarized in Algorithm 1. The testing phase is performed only once by using the final trained model. It is worth noting that, LBCL is model-agnostic and can be applied to any classification-based VQA models for alleviating language bias.

5.4 Experiments

5.4.1 Datasets and Baselines

Datasets: In this chapter, we train and evaluate the LBCL approach on the VQA-CP v1 and v2 datasets [85]. They are the two most commonly used benchmark datasets proposed to test the robustness of VQA models under severe language

Algorithm 1: Language Bias driven Curriculum Learning

Input: Training set $S = \{I_i, Q_i, a_i\}_{i=1}^{N}$, difficulty metric $d(\cdot)$,curriculum selection function $select(\cdot)$,VQA model f_θ,question-only model f_ϕ ,total training stage T, and maximum model competence C_{max}

Output: Trained VQA model

initialization;

for *t = 1, 2, ..., T* **do**

  if *t=1* **then**

    $\theta \leftarrow \text{train}\,(\theta, S)$;

    $\phi \leftarrow \text{train}\,(\phi, S)$;

    compute the difficulty d_i for each example $\in S$ **(section 3.1)**;

  else

    compute the real-time model competence $C(t)$

    $S^* \leftarrow \text{select}(S, C(t))$ **(section 3.2)**;

    $\theta \leftarrow \text{train}\,(\theta, S^*)$ with knowledge distillation **(section 3.3)**;

Table 5.1: Performance on three benchmark models and the models applied with our LBCL approach.

Model	VQA-CP v2				VQA-CP v1
	Overall	Yes/No	Number	Other	Overall
UpDn	40.75	42.10	12.77	47.74	38.36
+Ours	60.74	88.28	45.77	50.14	61.57
BAN	40.69	43.49	13.66	46.64	38.91
+Ours	60.62	87.72	49.39	49.50	61.08
MCAN	42.48	48.27	14.70	47.06	39.20
+Ours	61.84	88.17	53.60	50.30	61.40

prior, which are constructed by re-organizing the train and val splits of VQA v1 [30] and v2 [48] datasets respectively. The train and test sets of VQA-CP dataset have entirely different answer distributions. Consequently, a model strongly suffering from the language bias in the train set will perform poorly on the test set. Specifically, it consists of the overall accuracy on the whole dataset, and the performance for VQA samples related to three different question categories (Yes/No, Number, Other).

Baselines: To demonstrate that our LBCL is model-agnostic, we test it on top of three VQA models: UpDn [191], BAN [68] and MCAN [74]. The UpDn model is a baseline model that uses a bottom-up visual attention to assign weights for different visual objects. The BAN model is proposed to achieve multi-hop reasoning by stacking multiple bilinear attention layers with residual connections. MCAN is a transformer-based VQA model that jointly achieves pairwise interactions for inter-modal and intra-modal features.

Table 5.2: Comparisons between difficulty metrics Training Error (TE) and Visual Sentisive Coefficient (VSC) under two-stage LBCL with different settings of maximum model competence C_{max} (best results are in bold).

C_{max}	TE					VSC				
	VQA-CP v2				VQA-CP v1	VQA-CP v2				VQA-CP v1
	Overall	Yes/No	Number	Other	Overall	Overall	Yes/No	Number	Other	Overall
0	41.15	43.13	12.33	48.02	38.47	55.34	86.87	13.76	**50.21**	56.58
0.25	52.91	79.42	18.57	**48.44**	51.00	56.64	**89.55**	18.40	49.89	58.62
0.5	55.34	86.17	25.60	47.35	55.91	58.26	88.84	30.25	49.92	59.15
0.75	55.09	86.10	29.41	45.88	56.11	58.06	85.20	37.20	49.57	59.75
1	**55.45**	**86.34**	29.42	44.76	57.48	**58.97**	82.81	46.69	49.90	**60.27**
1.25	54.28	86.20	33.35	43.29	58.21	58.87	80.38	51.87	49.52	59.87
1.5	54.47	84.83	41.23	42.20	**58.35**	58.59	79.72	54.22	48.86	58.44
1.75	54.16	81.91	49.92	40.79	57.90	57.56	76.81	**55.32**	48.09	57.13
2	53.29	79.92	**51.80**	39.75	57.20	56.78	76.08	54.40	47.32	55.58

5.4.2 Implementation Details

Network Architecture: For visual representation, we use the pre-trained Faster R-CNN [246] to obtain object-level image features with no more than 100 proposals with their 2048-d features. For question features, we adopt a Glove [247] to encode the question as a word-level vector, and set the max length of words in question to 14. As for the architecture of the question-only model, we first exploit a LSTM [60] to encode the word-level textual features, and then feed the output of LSTM into a classifier including three fully connected layers to predict the correct answers.

Training strategy: As VQA examples tend to have several ground truth answers in VQA-CP v1 and v2 datasets, we adopt soft cross-entropy loss (Equ. (5.3)) to train VQA model as multi-label classification. In the first learning stage, we train VQA model with a question-only branch for 12 epochs with the same setting in original papers. After defining the difficulty metric based on the predictions in the first stage, we progressively train the model with weighted training instances based on the curriculum selection function. Specifically, for the optimal setting of the curriculum selection function in LBCL, we set the initial difficulty value $d_2 = 0$, the difficulty increment parameter $\gamma = 0.25$, the maximum model competence $C_{max} = 1.25$, and the number of total training stage $T = 7$, which can be computed with Equ. (5.9). After the first stage, the number of epoch for each training stage is 2. In the final stage, we additionally train VQA model for two epochs, where the learning rate is decayed by 1/5. 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.

5.4.3 Ablation Study

In this subsection, we carry out extensive ablation studies for the LBCL on VQA-CP v1 and v2 datasets. For simplicity, we only show the overall accuracy on the VQA-CP v1 test split. To be specific, we first make a general comparison between

three baseline models and the models applied with our LBCL approach. Then, we progressively validate the contributions of different components in the LBCL.

Curriculum Learning vs Non-Curriculum Learning: As depicted in Tab. 5.1, incorporating LBCL and VQA models can significantly improve the 'overall' performance, with average 22% and 20% accuracy boosts over three benchmark models on VQA-CP v1 and v2 datasets, respectively. For different question types, LBCL achieves remarkable improvements on biased types ('Yes/No' and 'Number'), and is also beneficial to the less-biased 'other' type with a gain of 2.5%. All these results indicate that LBCL can effectively overcome language bias problem through the easy-to-hard training strategy. The following content in this subsection validates the effectiveness of several crucial components in our LBCL framework, which is mainly implemented on the baseline UpDn model.

VSC vs TE: We compare proposed difficulty metrics Training Error (TE) and Visual Sensitive Coefficient (VSC) under the two-stage LBCL. Specifically, after the first-stage training, the real-time model competence is equal to the predefined maximum model competence ($C(2) = C_{max}$) in the second stage. The C_{max} is set from 0 to 2 for experimental analysis. As illustrated in Tab. 5.2, with the C_{max} rising from 0 to 1, the 'overall' accuracies of two difficulty metrics yield better results. However, its higher values ($C_{max} > 1.25$) may fail to fulfill further improvements. For different question types, we can notice that the optimal maximum model competence C_{max} are different. Specifically, the appropriate C_{max} for questions related 'Yes/No' and 'Other' are 0 and 0.25 evaluated by the VSC, while that for 'Number' type is 1.5. It verifies that the levels of language bias for different question types are different. In contrast to other types, the questions related to 'Number' are more seriously suffered from the language bias problem. For the comparisons between two difficulty metrics, VSC is remarkably superior to TE, as VSC achieves better 'overall' results, and also keeps more stable performance on 'Other' type with the change of C_{max}. It can be explained that, in contrast to TE, VSC have better capacity to evaluate the language bias driven difficulty of each VQA sample by measuring visual dependence. Thereby, we adopt VSC as the difficulty metric to implement the following experiments.

The contribution of easy concept: In Tab. 5.3, we demonstrate the importance of easy concept (textual correlation) under two-stage LBCL as well (the same as that in 'TE vs VSC'). Since the easy concept is mainly learned from the initial learning stage, for VQA models without the easy concept, we directly train them with the 'hard' VQA samples from scratch, instead of training them from the first stage to second stage progressively. From the results with different maximum model competence C_{max}, training VQA model without the easy concept would severely impair the model performance, especially for the accuracy of 'Other'. This is due to the fact that, compared with other question categories, the textual correlations between questions and answers related to 'Other' are more complex. Hence, the

Table 5.3: The effect of easy concept in the first stage under two-stage curriculum learning.

C_{max}	Easy Concept	VQA-CP v2			
		Overall	Yes/No	Number	Other
0.5	✓	58.26	88.84	30.25	49.92
	×	52.88	88.13	20.87	43.19
0.75	✓	58.06	85.17	37.20	49.57
	×	53.82	86.58	31.62	42.75
1	✓	58.97	82.81	46.69	49.90
	×	54.36	85.91	43.80	40.74
1.25	✓	58.87	80.83	51.87	49.52
	×	52.25	81.74	43.68	39.16
1.5	✓	58.59	79.22	54.22	48.86
	×	50.69	79.53	43.19	37.63

easy concept (textual correlations) plays a positive role on the unbiased learning, and benefits VQA models to narrow the answer space in their initial learning stage.

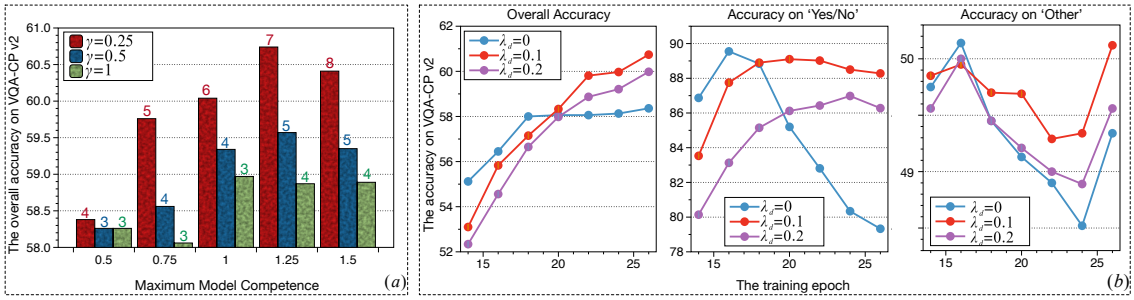

Figure 5.3: (a) Results of LBCL with different settings of difficulty increment parameter γ, where the number above each bar represents the total training stage T for LBCL with the corresponding setting. (b) Results ('Overall', 'Yes/No' and 'Other' accuracies) of LBCL with different trade-off factor λ_d in knowledge distillation.

The variant of difficulty increment parameter: Fig. 5.3(a) depicts the comparisons for LBCL with different settings of difficulty increment parameters γ, which implies the increase of model competences $C(t)$ from the t^{th} stage to the $(t+1)^{th}$ stage. We can see that, the smaller setting of γ (0.25) obtains better results in contrast to the larger value (0.5 or 1), and reaches its best performance at 60.74, when the maximum model competence C_{max} is 1.25. This is because, based on a fixed C_{max}, the relative low difficulty increment divides the whole learning process into more stages with different difficulty levels. For instance, when the $C_{max} = 1.25$, LBCL with $\gamma = 0.25$ requires 7 training stages, whereas that with $\gamma = 1$ only keeps 4 stages. This fine-grained easy-to-hard learning is more conducive for VQA model to alleviate language bias progressively.

The contribution of knowledge distillation: Fig. 5.3(b) shows the advantage of knowledge distillation in our approach, and all experiments are implemented under LBCL with $C_{max} = 1.25$ and $\gamma = 0.25$ (the best result in Fig 5.3(a)). With the training epoch increasing from 18 to 26, the LBCL without distillation ($\lambda_d = 0$)

Table 5.4: Comparisons with the state-of-the-art based on the UpDn model on VQA-CP v2 dataset. Best and second best numbers are in bold and underlined, respectively.

	Method	VQA-CP v2			
		Overall	Yes/No	Number	Other
Baseline	UpDn [191]	39.74	42.27	11.93	46.05
Annotation Based	AttAlign [92]	39.37	43.02	11.89	45.00
	HINT [92]	46.73	67.27	10.61	45.88
	SCR [105]	49.45	72.36	10.93	48.02
Data Rebalance Based	SSL [243]	57.59	86.53	29.87	50.03
	CSS [94]	58.95	84.37	49.42	48.21
	CSS+CL [236]	59.18	86.99	**49.89**	47.16
	Mutant [110]	**61.72**	88.90	49.68	**50.78**
Reducing Language Prior Based	AdvReg [89]	41.17	65.49	15.48	35.48
	RUBi [106]	47.11	68.65	20.28	43.18
	DLR [268]	48.87	70.99	18.72	45.57
	VGQE [101]	50.11	66.35	27.08	46.77
	LM+H [107]	52.01	72.58	31.12	46.97
	RMFE [99]	54.55	74.03	49.16	45.82
	CF-VQA [88]	55.05	**90.61**	21.50	45.61
	LBCL (Ours)	60.74	88.28	45.77	50.14

could hardly achieve further improvements on the overall accuracy, since the VQA model encounters the catastrophic forgetting problem on the VQA examples related to 'Yes/No' and 'Other'. By integrating knowledge distillation into LBCL, the VQA model ($\lambda_d = 0.1$) can effectively avoid the performance degradation caused by forgetting problem for aforementioned two question types, and obtain superior overall results, with an accuracy gain of 2.5%. The larger setting of the factor λ_d ($\lambda_d = 0.2$) fails to fulfill further improvements.

5.4.4 State-of-the-art comparison

Performance on VQA-CP v2: We first compare the LBCL with state-of-the-art methods proposed to alleviate language prior on the VQA-CP v2 dataset. We roughly divide these debiasing strategies into three categories: annotation based, data rebalance based and reducing language prior based approaches. Annotation based methods attempt to use human explanations to enhance the visual grounding for VQA models. The data rebalance based approaches tend to generate counterfactual VQA examples, and further rebalance the training answer distribution. We consider other debiasing approaches as the reducing language prior based methods. As the UpDn model is the widely-used benchmark model for evaluating language bias, we list the performance of state-of-the-art approaches implemented on the UpDn model in the Tab. 5.4.

Overall, the data rebalance based methods tend to achieve better results than approaches belonging to other categories. However, this kind of methods aim to create

Table 5.5: Comparisons with the state-of-the-art based on UpDn model on VQA-CP v1 dataset (best results are in bold).

Method	VQA-CP v1			
	Overall	Yes/No	Number	Other
UpDn [191]	37.87	42.58	14.16	42.71
AdvReg [89]	45.69	77.64	13.21	26.97
GRL [291]	44.09	75.01	13.40	42.67
RUBi [106]	44.81	69.65	14.91	32.13
LM+H [107]	55.27	76.47	26.66	45.68
CF-VQA [88]	56.80	87.76	13.89	43.25
CSS [94]	59.63	86.62	28.93	45.12
CSS+GS [292]	58.05	78.50	37.24	46.08
CSS+CL [236]	61.27	**88.14**	34.43	46.08
LBCL (Ours)	**61.57**	84.48	**42.84**	**46.32**

more examples to change the training bias, instead of achieving jointly multimodal reasoning under language prior. This behavior is not consistent to the intention of VQA-CP dataset. Nevertheless, our method still outperforms most Data Rebalance Based methods (SSL, CSS and CL), and obtains competitive results compared with the state-of-the-art method Mutant [110]. In comparison to the baseline model, our LBCL significantly enhances performance (+21% for 'Overall' accuracy), and is superior to other reducing language prior based approaches. These results powerfully support that our LBCL effectively alleviates language bias with curriculum learning, and further benefits VQA models to achieve unbiased reasoning for multimodal information.

Performance on VQA-CP v1: Tab. 5.5 illustrates performance comparisons with the existing competitive models on the VQA-CP v1 dataset. We achieve a new state-of-the-art result on this dataset, with a significant accuracy boost (from 37.87% to 61.57%) over the UpDn model. In particular, the LBCL module obtains highest accuracy of 42.84% on the hardest 'Number' question type . In addition, the performance of our method also outperforms those impressive debiasing strategies (such as CF-VQA [88], CSS[94] and CSS+CL [236]), which is consistent to the performance on the VQA-CP v2 dataset. All these results further verify the effectiveness of the LBCL for alleviating language bias in VQA task.

5.4.5 Experiments on small-scale datasets

To further demonstrate the generalizability of our method, we randomly sample different proportions of the train split from VQA-CP v1 and v2 datasets, and carry out a series of experiments for verifying our method. In general, LBCL achieves remarkable and stable improvements under different data scales, with overall 23% and 20% performance boosts on VQA-CP v1 and v2 datasets. Particularly, even trained with limited 20% training data, our method implemented on UpDn model

Table 5.6: Results of our LBCL on the VQA-CP v1 and v2 datasets with different percentages of training split.

Model \ Per	VQA-CP v1					VQA-CP v2				
	20%	40%	60%	80%	100%	20%	40%	60%	80%	100%
UpDn	30.85	32.33	34.97	37.12	38.36	34.85	36.98	38.23	40.07	40.75
UpDn+Ours	54.31	56.17	58.89	60.61	61.57	52.89	55.34	58.11	60.03	60.74
BAN	29.91	31.25	35.03	37.53	38.91	35.73	37.04	37.92	39.97	40.69
BAN+Ours	55.25	56.82	58.94	60.13	61.08	51.56	54.28	58.52	59.78	60.62
MCAN	31.56	33.46	36.85	38.87	39.20	36.89	38.14	39.21	41.22	42.48
MCAN+Ours	57.34	58.13	59.34	60.89	61.40	53.49	56.37	59.18	61.02	61.84

can still obtain 52.89% overall accuracy on the VQA-CP v2 dataset, which outperforms many impressive debiasing approaches trained with the whole dataset, such as LM+H [107], VGQE [101] and DLR [268] in Tab. 5.4. These results strongly support the effectiveness of LBCL to overcome the language bias inherently existed in raw data, and tend to show its great applicability to be exploited in more real-world scenarios with data bias.

5.4.6 Qualitative Results

Fig. 5.4 reveals the qualitative results of VQA samples with different VSC (difficulty metric) related to question types of '*how many*' and '*what sport*'. To visualize the utilization of training samples in different learning stages in our method, we set the initial difficult value $d_2 = 0$, maximum model competence $C_{max} = 1$, and the difficulty increment parameter $\gamma = 0.5$ for LBCL with total four-stage learning. From the answer distributions of aforementioned question types, '*2*' and '*tennis*' are prior answers accounted for large proportions in answer candidates. Hence, the VQA samples about '*2*' and '*tennis*' are endowed with lower values of VSC, as they are severely suffered from language bias, and can hardly earn benefits from visual contents. By contrast, the VQA samples whose correct answers are rare ('*12*' and '*frisbee*') tend to acquire high VSC, which are mainly exploited in the 'Stage 4' (the stage for 'hard' concept) to explore multimodal reasoning for VQA models.

5.5 Case Study

In Fig. 5.5, we make case studies to qualitatively show the superiority of our approach, where the VQA examples are sampled from the VQA-CP v2 test split. These examples cover a broad range of question categories (e.g. recognition, verification and number), and the corresponding visual content could be color or grey images. In general, compared with the baseline model, our approach yields better results, and is inclined to obtain higher prediction certainty. In the train split of VQA-CP v2, '*Yes*' is a prior answer which has higher frequency in contrast to other answer candidates. As a result, for either relevant question '*Is there someone*

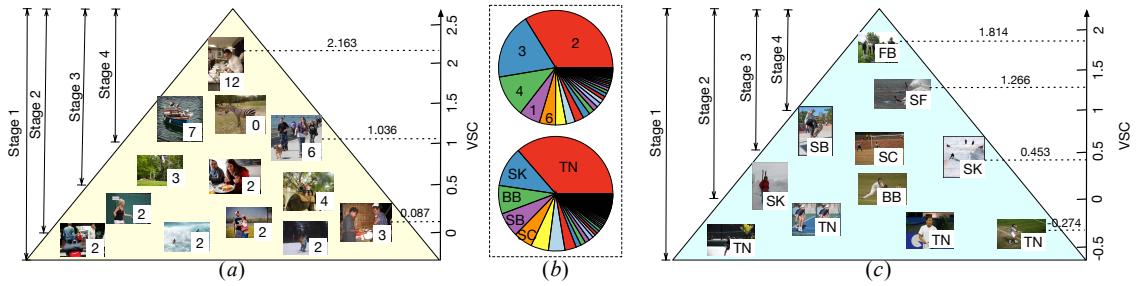

Figure 5.4: (a) The qualitative analysis for VQA samples related to 'how many'. (b) The training answer distributions for question categories 'how many' and 'what sport'. (c) The qualitative analysis for VQA samples related to 'what sport'. The abbreviations in (b) and (c) are tennis (TN), skiing (SK), baseball (BB), skateboarding (SB), soccer (SC), surfing (SF) and frisbee (FB).

clean the floor?' or irrelevant question '*Why this man be taking a break?*', the baseline model overwhelmingly selects '*Yes*' as the correct answer, instead of achieving multimodal reasoning. For the same reason, when the questions are related to 'number' and 'sport', prior answers '*tennis*' and '*2*' have higher predicted results from baseline model. On the contrary, our LBCL method can effectively avoid the aforementioned statistical prior, and predict correct answers more related to the visual modality.

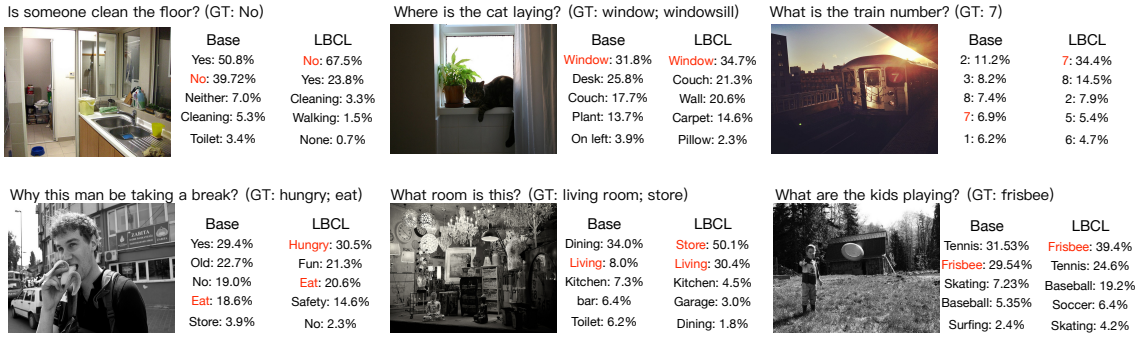

Figure 5.5: The VQA examples of case study on the VQA-CP v2 dataset

5.6 Conclusion

We presented a novel Language Bias driven Curriculum Learning (LBCL) for unbiased VQA model. Our approach enabled VQA model to overcome language prior through an easy-to-hard training strategy. In the structure of our LBCL, we defined a novel difficulty metric Visual Sensitive Coefficient (VSC) to evaluate the difficulty for each VQA instance in accordance with language prior. We also integrated knowledge distillation into LBCL to avoid catastrophic forgetting problem through progressive training. Extensive experiments showed the effectiveness of our method.

Future Works: In the future, from the theoretical aspect, we plan to integrate the meta learning into the LBCL, so as to automatically design the training difficulty and for curriculum learning. From the practical aspect, we try to put our LBCL into other multimodal scenarios, where the uni-modal bias problem is severe to impair model robustness.