RMLVQA: A Margin Loss Approach For Visual Question Answering with Language Biases

Abhipsa Basu
abhipsabasu@iisc.ac.in

Sravanti Addepalli
sravantia@iisc.ac.in

R. Venkatesh Babu
venky@iisc.ac.in

Vision and AI Lab, Indian Institute of Science, Bangalore

Abstract

Visual Question Answering models have been shown to suffer from language biases, where the model learns a correlation between the question and the answer, ignoring the image. While early works attempted to use question-only models or data augmentations to reduce this bias, we propose an adaptive margin loss approach having two components. The first component considers the frequency of answers within a question type in the training data, which addresses the concern of the class-imbalance causing the language biases. However, it does not take into account the answering difficulty of the samples, which impacts their learning. We address this through the second component, where instance-specific margins are learnt, allowing the model to distinguish between samples of varying complexity. We introduce a bias-injecting component to our model, and compute the instance-specific margins from the confidence of this component. We combine these with the estimated margins to consider both answer-frequency and task-complexity in the training loss. We show that, while the margin loss is effective for out-of-distribution (ood) data, the bias-injecting component is essential for generalising to in-distribution (id) data. Our proposed approach, Robust Margin Loss for Visual Question Answering (RMLVQA) [1] *improves upon the existing state-of-the-art results when compared to augmentation-free methods on benchmark VQA datasets suffering from language biases, while maintaining competitive performance on id data, making our method the most robust one among all comparable methods.*

1. Introduction

Visual question answering (VQA) lies at the intersection of computer vision and natural language processing. It is the task of answering a question based on a given image. VQA networks need to combine knowledge from both visual scene and the question to predict the answer. These systems have numerous applications such as aiding the visually impaired in understanding their surroundings, image retrieval systems in e-commerce, and robotics.

With the success of deep learning, research in VQA has made great strides in recent years [3, 5, 16]. However, studies have shown that deep networks may learn correlations between the question and answer alone, ignoring the image modality [3, 21, 36]. They fail to do multimodal reasoning, specifically, if there is a class imbalance in the answer distributions of the training and test sets. For example, if most of the questions starting with "What color..?" are paired with the answer "red" in the training data, the model memorizes this trend to answer "red" for all color based questions in the test set, irrespective of the image.

One solution to this problem is to perform data augmentations in various ways [1, 4, 11, 15, 27, 36, 42, 43, 46]. These methods outperform most other debiasing techniques in the literature. However, augmentation strategies are dependent on the dataset in question and the type of biases observed, which makes the process manual and tedious. Another extensively explored solution is to learn the bias in the dataset separately using a question-only branch [9, 12, 18] and explicitly removing the learnt bias from the base model.

Margin losses have been widely used for a number of tasks. Cao et al. [10] address the problem of long tailed recognition through an *adaptive* margin loss, ensuring higher cosine margin penalty for the tail classes and lower penalty for other classes. A similar adaptive cosine margin penalty has been applied to mitigate the language bias problem in VQA by a previous work [17]. Margin losses have been widely used in deep face recognition as well, where it is desired that the features of two images of the same person should be as similar as possible, whereas those of two different people should be far apart. Margin losses

[1] Code available at https://github.com/val-iisc/RMLVQA

are used in this regard for facial feature discrimination to maximize the decision margin. Rather than the Euclidean space, both these margin losses project their feature spaces onto a fixed radius hypersphere. While the well known *CosFace* loss [41] uses cosine margin penalties like Cao et al. [10], *ArcFace* [14], uses angular margins to maximize the decision margins among different classes in this feature space. Angles correspond to geodesic distance on the projected hypersphere which stabilizes the training process and is shown to improve the discriminative power of face recognition models as compared to cosine margin based methods.

We believe that the language bias problem of VQA can be addressed by learning discriminative features for the biased and unbiased samples in the dataset. To this end, we implement an *adaptive* angular margin loss, inspired by ArcFace, as margins allow models to distinguish between different kinds of samples. However, the key question here would be how to set the adaptive margins to cater to the specific problem of language biases. Traditional models over-trust the questions over the images due to a class imbalance in the training and test set answer distributions. Hence, one way to set the margins would be to ensure that the training samples with frequent answers are given a smaller penalty due to the abundance of those answers in the dataset, whereas those with rare answers are given a higher margin value. We refer to the resultant margin values as the frequency-based margins. However, one factor that is ignored in this aspect is the answering difficulty of each sample. We argue that more margin should be given to hard samples compared to the easier ones even if the corresponding answers are frequent. In this regard, we propose the learning of instance-specific margins during training, so as to allow the model to distinguish between hard and easy samples alongside frequent and rare samples. To the best of our knowledge, this is one of the first attempts in learning margins automatically and parallely during training. We combine these learnable margins with the frequency based margins used in prior works [17], so as to allow both frequency and complexity of data samples to control the training dynamics. To compute these learnable margins, we introduce a bias injecting module to our model that is trained using Cross-Entropy (*CE*) loss. We show that the *CE* loss additionally clusters the training samples based on the dataset bias itself, thereby aiding the margin loss further. We further introduce a supervised contrastive loss [23] to pull the features of samples having the same answer together, while pushing others apart.

Adaptive margin losses, with margins calculated by using frequency of answers in the dataset, perform well in the *ood* setting. However, we show that they cause a drop in the in-domain performance, which raises questions on their robustness. As pointed by [40], it is crucial for VQA models to perform well on both in-domain and *ood* data since the test set distibution is unknown apriori. We mitigate this issue by ensembling the outputs of the bias injecting component and the proposed learnable margin-loss trained classification head during inference. This makes the model robust to the difference in answer distributions of the training and test sets. Our overall method is called ***RMLVQA*** - Robust Margin Loss for Visual Question Answering with language biases. The key contributions of this work can be summarized as follows:

- We propose to mitigate the well known problem of language bias in VQA models by introducing an instance-specific adaptive margin loss, to allow the use of different margins for the learning of samples with varying complexities, in addition to the use of frequency-based margins. To achieve this, we introduce a bias-injecting component and allow the margins to be computed based on prediction probabilities of this branch. We show that this clusters samples in the feature space based on the bias present in the dataset.
- We propose to overcome the *id-ood* trade-off in margin-based losses, by ensembling the outputs of the bias-injecting component and the main model.
- We further introduce a supervised contrastive loss that pulls features of training samples having the same ground truth answers together, while pushing apart others. This aids the margin loss further.
- Through extensive experiments and ablations, we show how the proposed approach achieves state-of-the art results when compared to augmentation-free methods on the *ood* VQA-CP v2 dataset, while maintaining competitive performance on the *id* VQA v2 dataset. This makes our model the most robust one among all non-augmentation based methods.

The code, hyperparameter analyses, and results on multiple datasets are shared as supplementary material.

2. Related Work

VQA and the language bias problem. VQA [5, 16] is the task of answering natural language questions given an image. VQA v1 & v2 [5, 16] are widely used datasets that are used to benchmark different algorithms. However, prior works have shown that these datasets have biases which can be amplified in the trained models [2, 3, 19, 36, 37]. Specifically, in case of language bias, the model learns a correlation between the question and the answer directly. For example, if the model always predicts "tennis" for the sport-related questions, it can achieve around 40% accuracy on the VQA v1 dataset, as the images of the VQA datasets are from MS-COCO [28], which has many "tennis" related images. To effectively quantify the extent of bias in the trained models, a new dataset VQA-CP v2 [3] was created from the

VQA v2 dataset. In this work, we aim to solve the language bias problem of VQA, where we evaluate our models on the *ood* benchmark dataset VQA-CP v2 and the *id* VQA v2.

Bias mitigation techniques. The language bias problem reduces the generalizability of the VQA models, resulting in a drop in performance when the training set answer distribution is different from that of the test set. There are three popular categories of solutions in the literature tackling this problem, which we discuss below.

Visual Grounding based methods [34, 45] used human annotations as supervision to attention maps [13, 33], whereas more recent works use Grad-CAM for the same [36, 43].

Ensemble based methods use a separate question-only model to capture the bias, and further remove the bias explicitly from the base VQA model [9, 12, 35]. Apart from these techniques, Niu et al. [31] introduce a counterfactual inference framework to mitigate the biases. KV and Mittal [25] use a visually grounded question encoder such that the visual information is embedded in the question representation itself. Niu et al. [32] improve both *ood* generalizability and *id* performance through introspective knowledge distillation.

Augmentation based methods remove bias by augmenting the dataset to make it more balanced. Some works generate counterfactual data by masking or transforming the critical objects and words, and further generating appropriate ground truth answers [11, 15, 27]. Other works introduce negative samples by shuffling the inputs (images and questions) in the minibatch to reduce bias without requiring any annotation [40, 42, 46]. Wen et al. [42] use two kinds of negative samples for balancing the dataset, and construct both question-only and vision-only models to remove the biases.

Margin-based methods. Max-margin classifiers are used in popular machine learning algorithms like SVMs [38]. Margin losses are commonly used in many problems such as deep face recognition [7, 14, 29, 41], long tailed or class imbalanced learning [10, 22, 30] and few-shot learning [26]. Guo et al. [17] apply an adaptive cosine margin loss to discriminate the frequent and rare answers for a given question type and implement this by transforming the multimodal feature space to a fixed radius hypersphere. The adaptive margins are estimated from the training data based on the frequency of occurrence of an answer in questions of a given type, ensuring that rare answers are given more margin penalty, whereas frequent classes are given less penalty.

3. Proposed approach

3.1. Preliminaries

We consider the problem of classification based Visual Question Answering (VQA). It is the task of answering questions based on an image. Given a dataset $\mathcal{D}$ having n samples, with image $v \in \mathcal{V}$, question $q \in \mathcal{Q}$ and answer (class label) $a \in \mathcal{A}$, the goal is to train a model to optimize a mapping function $f : \mathcal{V} \times \mathcal{Q} \rightarrow \mathbb{R}^{|\mathcal{A}|}$ to generate predictions for the given question and image. The VQA model consists of four parts: (1) image feature extractor e_v, (2) question feature extractor e_q, (3) the VQA model m_f which fuses the image and question features to generate the joint multimodal features $\mathbf{x}$, (4) classifier c having weights $\mathbf{W}$, that generates the logits f. Most of the classical VQA models [4, 6, 8, 24, 44] follow this paradigm and formulate the problem as follows:

$$f(v, q) = c(m_f(e_v(v), e_q(q))) \tag{1}$$

Standard Cross-Entropy loss. VQA models can be trained by optimizing the Cross-Entropy (*CE*) loss[2] shown below:

$$L_s = \sum_{i=1}^{|\mathcal{A}|} -a_i \log \frac{\exp(f_i)}{\sum_{j=1}^{|\mathcal{A}|} \exp(f_j)} \tag{2}$$

To solve the problem of language bias in VQA, the model should learn discriminative features for different answers for a given question type so that it can distinguish frequently occurring answers from rarely occurring answers, as well as those with varying complexity. However, standard *CE* loss only favours the samples with frequently occurring answers.

Normalized *CE* loss. Before defining the margin loss, we define a reformulation of the *CE* loss as a cosine loss [14, 41], by $L2$-normalizing the classifier weight vectors $\mathbf{W_i} \in c$ for each $a_i \in \mathcal{A}$, and feature $\mathbf{x} = m_f(e_v(v_i), e_q(q_i))$. Therefore, we define $\hat{\mathbf{W}}_\mathbf{i} = \frac{\mathbf{W_i}}{\|\mathbf{W_i}\|}$ and $\hat{\mathbf{x}} = s\frac{\mathbf{x}}{\|\mathbf{x}\|}$, where s is a scaling parameter. Let θ_i be the angle between $\mathbf{x}$ and $\mathbf{W_i}$. Therefore the logit for each a_i is transformed as (keeping the bias term as 0 for simplicity):

$$f_i = \hat{\mathbf{W}}_\mathbf{i}^{\top} \hat{\mathbf{x}} = \|\hat{\mathbf{W}}_\mathbf{i}\| \|\hat{\mathbf{x}}\| \cos\theta_i = s \ \cos\theta_i \tag{3}$$

The joint features $\hat{\mathbf{x}}$ are thus distributed on a hypersphere with a radius s. This makes the normalized *CE* loss as:

$$L_{ns} = \sum_{i=1}^{|\mathcal{A}|} -a_i \log \frac{\exp(s \ \cos\theta_i)}{\sum_{j=1}^{|\mathcal{A}|} \exp(s \ \cos\theta_j)} \tag{4}$$

3.2. RMLVQA

In this subsection, we define our method, which is a Robust Margin Loss based approach for solving the problem of language biases in VQA.

Adaptive Angular Margin Loss. We introduce an adaptive angular margin loss that ensures decision margin maximization through angular margins in an adaptive way. The normalized *CE* loss transforms the feature space to an angular space of radius s. The motivation behind an adaptive

[2] In this paper, all losses will be expressed as individual sample losses

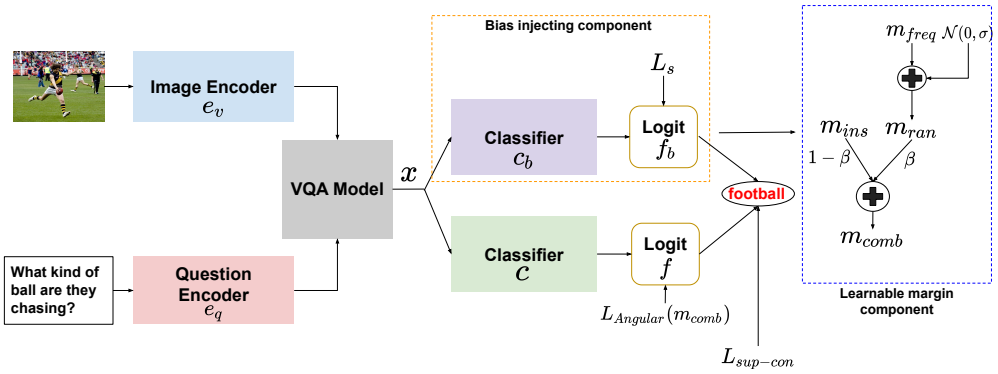

Figure 1. **Overview of our model**. The orange dashed region shows the proposed bias-injecting module, trained using the *CE* loss. The blue dashed region shows how the instance-based margins are generated and then combined with the randomized frequency-based margins to compute the final margins. The final prediction is obtained by combining the two logit heads.

margin instead of a constant margin is based on the root of the language bias problem in VQA. These biases occur when similar questions in the dataset are paired with the same answers most of the time, fooling the model to memorize these questions and the answers by ignoring the image completely. Thus, while some samples are overlooked because of the rarity of their corresponding answers, others are overly paid attention to. Adaptive margins allow the model to set different penalties for the varying samples, thus leading to highly discriminative features for the different training instances.

One aspect of the adaptive margins is to ensure that the instances with frequent answers are allowed a smaller margin than those with the rare ones. We call these margins m_{freq}. Similar to [17], these margins are calculated based on the question type as shown below:

$$\bar{m}^k_{freq}[i] = \frac{n^k_i + \epsilon}{\sum_{j=1}^{|\mathcal{A}|} n^k_j + \epsilon} \tag{5}$$

$$m^k_{freq}[i] = 1 - \bar{m}^k_{freq}[i] \tag{6}$$

where $\bar{m}^k_{freq}[i]$ measures the probability of occurrence of answer a_i in the training data for a given question type qt_k. n^k_i is the frequency of answer a_i in the training set calculated for the given qt_k, and ϵ is a hyperparameter for avoiding computational overflow. $m^k_{freq}[i]$ is the adaptive margin for answer $a_i \in \mathcal{A}$ corresponding to qt_k.

The adaptive angular margin loss adds a margin penalty to the angle between the features $\mathbf{x}$ and the classifier weights $\mathbf{W_i}$ for the i^{th} class as shown below. Since the margin is placed on the angle, it maps exactly to the "geodesic" distance on the hypersphere [14]. While Deng et al. [14] set a constant value for the margin, i.e. 0.5, our margins are adaptive in nature.

$$L^k_{Angular} = \sum_{i=1}^{|\mathcal{A}|} -a_i \log \frac{\exp(s\ \cos(\theta_i + m^k_{freq}[i]))}{\sum_{j=1}^{|\mathcal{A}|} \exp(s\ \cos(\theta_j + m^k_{freq}[j]))}$$

For notational simplicity, in the coming sections, we omit k from the margin and loss computations.

Randomization of the estimated margins. Boutros et al. [7] suggest that in typical margin losses, setting constant margins can limit the generalizability and discriminative power of a model. We note that $m_{freq}[i]$ is constant for each $a_i \in \mathcal{A}$ for a given question type qt_k over all samples in the training set. When qt_k is "how many", the margin for the answer "2" is same for all the training instances belonging to this question type. We believe that setting the same high margin value for a specific rare answer in a given question type and similarly lower value for a frequently appearing answer can lead to overcorrection of the language bias, by forcing the model to focus on the rare answers more than the frequent ones. Therefore, under a given question type qt_k, for every answer a_i, we use a randomized version of $\bar{m}_{freq}[i]$, called $\bar{m}_{ran}[i]$ in the following manner:

$$\bar{m}_{ran}[i] = \mathcal{N}(\bar{m}_{freq}[i], \sigma) \tag{7}$$

where $\mathcal{N}$ is the Gaussian distribution, and σ is the standard deviation, which is a hyperparameter. This impedes the model from overcorrecting with respect to the rare answers, thus increasing its generalizability. Finally, we obtain the randomized margin $m_{ran}[i] = 1 - \bar{m}_{ran}[i]$ for each $a_i \in \mathcal{A}$ in each qt_k.

Instance-based margins. The frequency-based margins estimated from the training data are constant for a given question type. So, for two different questions starting with "how many", the margins are same for each answer, irrespective of the model's difficulty in answering these instances. Ideally if one of the samples is hard to answer, more margin should be given to its ground truth class compared to the easier one. Randomizing m_{freq} as described above does not take the complexity of each sample into account. To this end, our model learns instance-level margins during training. We augment an auxiliary classifica-

tion head to the feature $\mathbf{x}$ (the orange dashed box in Fig. 1) which is architecturally same as the original model classifier c. We call this the bias-injecting component. Following Eq. 1, the formulation of this component becomes:

$$f_b(v, q) = c_b(\mathbf{x}) \tag{8}$$

where, f_b refers to the logits generated by this component and c_b refers to the classification head. It is to be noted that while $c_b \neq c$, both are trained to predict the answers given the questions and the images. Finally, we denote $\bar{m}_{ins} = softmax(f_b/\tau)$ as the confidence, (i.e. prediction probability) of the bias-injecting component in answering a question. Here τ stands for temperature which is a hyperparameter to the network. We define $m_{ins} = 1 - \bar{m}_{ins}$ as the instance-level learnable margin, obtained from the model during training. These learned margins are similar with the frequency-based margins when samples are far away from the decision boundary. However, for training instances that are close, m_{ins} increases the margins if the model confidence is low for the ground truth class, and decreases the same in case the model has high confidence in predicting the ground truth answer. We combine this instance-level confidence for the ground truth class with the previously calculated m_{ran} for each sample (v, q) to obtain the final margins m_{comb}, as shown in the blue dashed box in Fig. 1.

$$m_{comb}[gt] = \beta m_{ran}[gt] + (1 - \beta) m_{ins}[gt] \tag{9}$$

gt refers to the index of the ground truth class of the sample, and β is a hyperparameter used for combining the two margins. For all indices $l \neq gt$, $m_{comb}[l] = m_{ran}[l]$. In the first few training epochs, we do not compute m_{ins} as the model is in the initial stage of its learning, i.e. we set $\beta = 1$. As training progresses, β is decreased in a step-wise manner to increase the contributions from the learnt margins.

The bias-injecting component. The bias-injecting component is a classifier appended to the features $\mathbf{x}$ trained using the standard *CE* loss. As the *CE* loss is known to favour the frequently appearing answers in the training data, it automatically learns the bias in the dataset, and by backpropagating this bias into the network, it clusters samples in the feature space based on the source of the bias. This aids the margin loss as it can now separate the (frequent, rare), and the (easy, difficult) samples inside the individual clusters. Moreover, the advantage of using it to generate the learnable margins is that the bias captured is now reflected directly in $\bar{m}_{ins}$, which ensures that it is close to $\bar{m}_{freq}$, but still provides information on the answering difficulty of a samples lying close to the decision boundary. Another advantage of this component is that by learning the data bias, this module is capable of generalizing to any *in-distribution* data, whereas the predictions from the primary classifier c favour the *ood* data.

Inference. While margin losses are effective on an *ood* test data, they result in degraded performance on *id* test set. To mitigate this problem, during inference time, we take advantage of both the bias-injecting component f_b and the primary logit head f, as the former is useful in making predictions for an *id* test data, while the latter is useful in an *ood* setting. This makes our model robust to different data distributions. We donote by $\hat{p_{comb}}$ the combined predicted answer probabilities. This is obtained by combining the predictions from c and c_b as defined below:

$$\hat{p_{comb}} = \alpha.\hat{p} + (1 - \alpha).\hat{p_b} \tag{10}$$

where α is a hyperparameter controlling the weight of each logit head during inference, $\hat{p} = \text{softmax}(f)$ are the predictions from classifier c, $\hat{p_b} = \text{softmax}(f_b/\tau)$ are the predictions from the bias-injecting component c_b. The answer predicted is therefore $\hat{a} = argmax(\hat{p_{comb}})$. The temperature τ is same as that defined previously for the instance-based margins.

Supervised Contrastive (SupCon) Loss. We further separate samples in the feature space based on the SupCon loss. In addition to creating discriminative features for (frequent, rare) and (easy, difficult) samples, the SupCon loss leads to a degree of discrimination among samples with different ground truth answers in the feature space. We consider a mini-batch of size $\mathcal{B}$ of multimodal features denoted as $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_\mathcal{B}}\}$ and corresponding answers as $\{a_1, a_2, \ldots, a_\mathcal{B}\}$. If we consider the current sample with index j, the set of positive examples from the mini-batch is denoted by $P_j : \{i \in \mathcal{B} \quad s.t. \quad a_i = a_j\}$. Similarly, the set of negative examples is denoted by $N_j : \{i \in \mathcal{B} \quad s.t. \quad a_i \neq a_j\}$. The SupCon loss [23] is defined as,

$$L_{sup-con} = \sum_{j \in \mathcal{B}} \frac{-1}{|P_j|} \sum_{p \in P_j} \log \frac{\exp(\mathbf{x_j}^T \mathbf{x_p}/\tau)}{\sum_{n \in N_j} \exp(\mathbf{x_j}^T \mathbf{x_n}/\tau)}, \tag{11}$$

We set the temperature $\tau = 1$ for all experiments as we do not find any significant improvements by changing its value.

Total Loss: The final loss function for each sample is:

$$\mathcal{L} = L_{Angular}(m_{comb}) + L_s + L_{sup-con} \tag{12}$$

where $L_{Angular}$ is the angular margin loss, L_s is the *CE* loss training the bias injecting component, $L_{sup-con}$ is the SupCon loss. Our final model can be seen in Fig. 1. We name our overall method ***RMLVQA***.

4. Experiments

Dataset details. We evaluate our model on the VQA-CP v2 dataset (Visual Question Answering Under Changing Priors) [3], which is an *ood* benchmark, based on the standard evaluation metric shown by Antol et al. [5]. It was created by reorganizing the training and validation sets of the VQA v2 [16] dataset, ensuring that the distribution

of answers for each question type in the training set and the test set are different. The training set of VQA-CP v2 contains approximately 121k images and 438k questions, the test set contains approximately 98k images and 220k questions. Following previous works, we also evaluate our model on the VQA v2 validation set.

4.1. Results

Quantitative analysis. In Table 1, we show a comparative analysis of the performance of *RMLVQA* on both VQA-CP v2 test and VQA v2 validation sets. We report the overall accuracies along with those on "yes/no", "number" and "other" type questions. We observe that our method gains around 6.39% in the overall accuracy for VQA-CP in comparison with *AdaVQA* [17], which is the cosine margin loss trained model. We observe that *AdaVQA*, while performing well on VQA-CP v2, has considerably low scores for the in-domain VQA v2. With *RMLVQA*, the validation accuracy of VQA v2 is 59.99%, which is 13.01% higher than *AdaVQA*. This shows that our method is the first one trained by adaptive margins that is robust to both *id* and *ood* data. With a score of 60.41% for VQA-CP v2, the model improves upon the current state-of-the-art augmentation-free models. In addition to reporting the performance of the various models in literature on the two datasets, we show the relative difference in accuracies between VQA-CP v2 test data and VQA v2 validation data. *RMLVQA* outperforms all other non-augmentation based models in terms of robustness, with the lowest difference in the two accuracies (see Table 1), indicating that our method is the most robust compared to all other methods that do not use any augmentation, and perform equally well on both VQA-CP v2 and VQA v2. For all experiments, Updn [4] is our base network.

Qualitative analysis. We further demonstrate the effectiveness of our method in Fig. 2 which provides qualitative results on the VQA-CP v2 test set for ***RMLVQA*** and the base network UpDn. In the first row, we show examples for question type "what color". The base network UpDn [4] focuses on the incorrect region in the image, given by attention map scores, and outputs "blue", the most frequent answer in the training set for the given question type. The best scoring region from the attention maps of our model is localized at the correct region in the image, and it correctly outputs "red", which is not a frequent answer in "what color". In the second example, belonging to question type "what is the person", we see that while the baseline outputs "phone", which is a frequent class under that question type, ***RMLVQA*** predicts the correct answer "hat", which is a rare class, thus showing the effectiveness of our method. While the baseline outputs incorrect answers, it still focuses on the correct region in the image. This crucial observation indicates that the bias in the model weights do not allow this visual information to propagate forward. Thus it still picks from the frequent answers for the given question type.

Choice of hyperparameter α for inference time combination of prediction probabilities. Choice of α is crucial as it decides how much weight to put on the prediction probabilities of the bias-injecting component c_b and the primary classifier c. As VQA-CP v2 does not have a validation set, we choose $\alpha = 0.5$ for fair evaluation. However, we report the accuracies of VQA-CP v2 test and VQA v2 validation sets for different values of α in Table 2. We observe that $\alpha = 0.5$ leads to the most robust performance of our method on both *id* and *ood* data, where the relative gap of accuracies is just 0.42%. While we already choose the optimal value of α for reporting model performance, our experiments demonstrate that one can choose an appropriate value to control the tradeoff between *id* and *ood* performances.

4.2. Ablation studies

In this subsection, we discuss the role of each component of ***RMLVQA***. We evaluate them on the VQA-CP v2 test and VQA v2 validation data and show the results in Table 3.

Effectiveness of angular margins. We call the base adaptive angular margin loss trained model RMLVQA-Base, which only considers the frequency based margins. Compared to AdaVQA, it leads to a rise of 3.22% in the VQA-CP test accuracy and an increase of 2.74% in the accuracy of the VQA-v2 validation set. This shows the effectiveness of the angular margin loss over the cosine margins both in *id* and *ood* data, validating the choice of the angular margin loss for addressing the language bias problem of VQA.

Effect of randomization of the margins. Randomizing the margins of each answer leads to an improvement in the accuracy values of VQA-CP v2 test, as shown in Table 3. Further, we also notice that it improves the in-domain accuracy of VQA v2 by a large margin (7.12%), thus showing its effectiveness in generalizability.

Role of the learnable margins. As explained earlier, m_{ins} represents the confidence of the bias-injecting component in answering a question. We show the positive effect of adding these margins to m_{ran} in Table 3 for VQA-CP v2, with a slight loss in VQA v2. In Fig. 4 we show the Spearman's rank correlation coefficient ρ between m_{freq} and m_{ins} for the question type 'how many' for our method (for the angular and cosine margin losses defined in Sec 3 and AdaVQA [17] respectively), averaged over the relevant samples in each training epoch. We note the following: a) ρ is high throughout, indicating that the order of margin values in m_{ins} for a certain sample remains similar to that of the margin values in m_{freq} for the question type to which the sample belongs. b) ρ decreases over epochs, indicating the excess information being carried by m_{ins}, aiding in the discrimination of samples whose answer frequencies are similar, but have different complexities.

Role of the bias-injecting component. In Table 3, we show that the bias-injecting component increases the model's overall accuracy by 1.47% for VQA-CP v2 (3.65%

Table 1. **Accuracy comparisons with other methods** on the VQA-CP v2 and VQA v2 datasets. The methods have been grouped into different categories (separated by lines in the table): base approaches, methods using human annotations, those that modify the language encoder, models weakening the language bias in different ways, our approach ***RMLVQA*** trained with UpDn as backbone, and finally the augmentation based methods. The best performance in each column is highlighted in bold. We highlight the overall accuracy of RMLVQA on VQA-CP with an underline as it achieves state-of-the art on augmentation-free methods. $*$ denotes numbers shown in [17]. We report the average accuracy of our model over 5 random seeds (along the with standard deviations in the superscript). Columns ending with "-CP" denote accuracies for VQA-CP v2, others denote accuracies for VQA v2. † indicates our implementation.

Model	Y/N-CP	Num-CP	Others-CP	Overall-CP	Y/N	Num	Others	Overall	Diff
SAN [44]	38.35	11.14	21.74	24.96	70.06	39.28	47.84	52.41	27.45
GVQA [3]	57.99	13.68	22.14	31.30	72.03	31.17	34.65	48.24	16.94
UpDn [4]	42.27	11.93	46.05	39.74	81.18	42.14	55.66	63.48	23.74
S-MRL [9]	42.85	12.81	43.2	38.46	41.96	12.54	41.35	37.13	1.33
AttAlign [36]	43.02	11.89	45.00	39.37	80.99	42.55	55.22	63.24	23.87
HINT [36]	67.27	10.61	45.88	46.73	81.18	42.99	55.56	63.38	16.65
SCR [43]	70.41	10.42	47.29	48.47	78.8	41.6	54.5	62.2	13.73
VGQE [25]	66.35	27.08	46.77	50.11	-	-	-	64.04	13.93
DLR [20]	70.99	18.72	45.57	48.87	76.82	39.33	48.54	57.96	9.09
AdvReg [35]	65.49	15.48	35.48	41.17	79.84	42.35	55.16	62.75	21.58
RUBi [9]	68.65	20.28	43.18	47.11	-	-	-	-	-
LMH* [12]	70.29	44.10	44.86	52.15	65.06	37.63	54.69	56.35	4.2
CF-VQA [31]	90.61	21.50	45.61	55.05	81.13	43.86	50.11	60.94	5.89
IntroD [32]	**90.79**	17.92	46.73	55.17	**82.48**	**46.60**	54.05	63.40	8.23
GGE-DQ [18]	87.04	27.75	49.59	57.32	73.27	39.99	54.39	59.11	1.79
AdaVQA † [17]	70.83	49.00	46.29	54.02	47.78	34.13	51.14	46.98	7.04
RMLVQA	$89.98^{\pm 0.46}$	$45.96^{\pm 0.57}$	$48.74^{\pm 0.13}$	$\underline{\mathbf{60.41}}^{\pm \mathbf{0.32}}$	$76.68^{\pm 0.37}$	$37.54^{\pm 0.49}$	$53.26^{\pm 0.12}$	$59.99^{0.06}$	**0.42**
CVL [1]	45.72	12.45	48.34	42.12	-	-	-	-	-
Unshuffling [39]	47.72	14.43	47.24	42.39	78.32	42.16	52.81	61.08	18.69
RandImg [40]	83.89	41.60	44.20	55.37	76.53	33.87	48.57	57.24	1.87
SSL [46]	86.53	29.87	50.03	57.59	-	-	-	63.73	6.14
CSS [11]	84.37	49.42	48.21	58.95	73.25	39.77	55.11	59.91	0.96
CSS + CL [27]	86.99	49.89	47.16	59.18	67.27	38.40	54.71	57.29	1.89
Mutant [15]	88.90	49.68	**50.78**	61.72	82.07	42.52	53.28	62.56	0.84
D-VQA [42]	88.93	**52.32**	50.39	**61.91**	82.18	44.01	**57.54**	**64.96**	3.05

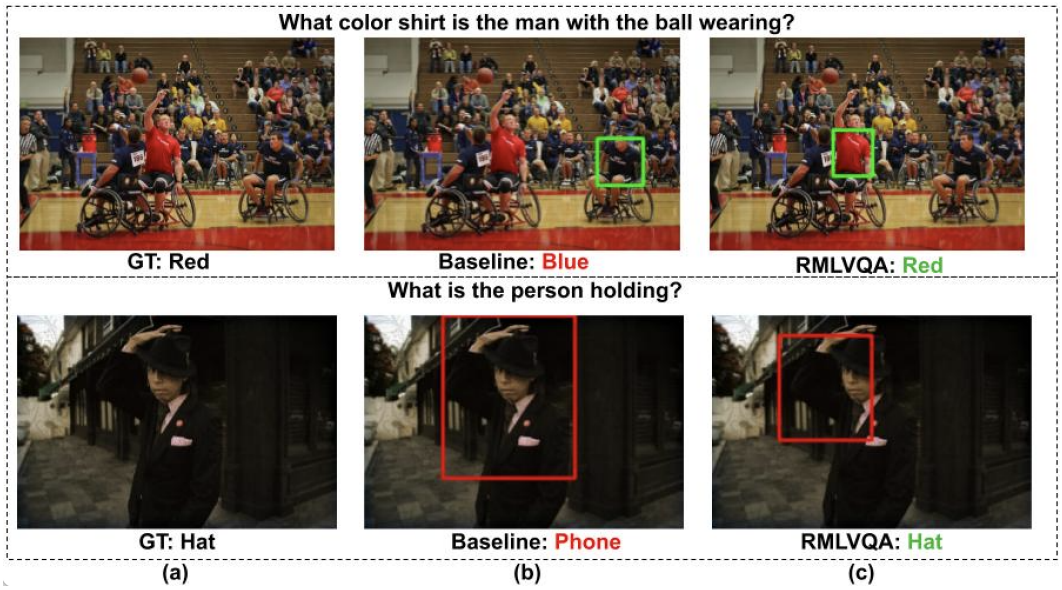

Figure 2. **Qualitative analysis of *RMLVQA***. We show two images from two question types in the rows. Column (a) represents the ground truth answers, column (b) represents the answers predicted by the baseline UpDn model. Column (c) represents the answer predicted by our method. The bounding boxes show the highest scored region in the attention map of the individual models.

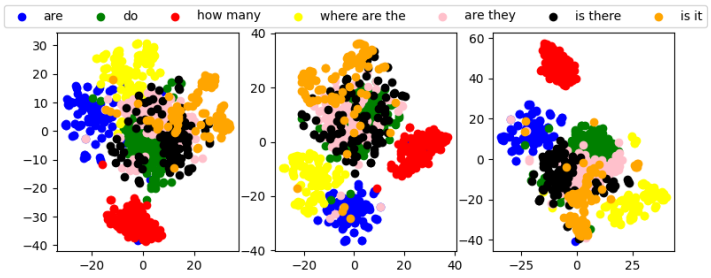

Figure 3. TSNE visualization of the feature space with respect to different question types for the model components, evaluated incrementally: a) *RMLVQA-Base*, b) Gaussian randomization, c) Bias injection

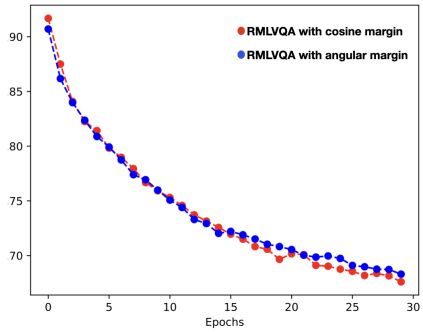

Figure 4. **Spearman's rank correlation coefficent** between $\mathbf{m_{freq}}$ and $\mathbf{m_{ins}}$ over epochs for question type "how many" on VQA-CP v2 test set

Table 2. **Role of α in the inference stage of *RMLVQA*.**

α	VQA-CP v2 Test	VQA-v2 Val	Gap
1.0	60.54	58.16	2.38
0.8	60.50	58.83	1.67
0.6	60.33	59.57	0.76
0.5	60.41	59.99	**0.42**
0.4	59.87	60.46	0.59
0.2	57.46	61.26	3.8
0.0	39.48	61.54	22.06

Table 3. **Ablations.** In this table we show the ablations of ***RMLVQA***, evaluated on VQA-CP v2 test & VQA v2 val.

Model	Y/N	Num	Others	Overall	VQA v2 Val
RMLVQA-Base	79.78	49.62	48.49	57.24	49.72
+Randomization	83.72	49.47	48.31	57.97	56.84
+Bias Injection	88.27	44.13	48.55	59.44	60.49
+Learnable Margin	88.06	46.63	48.74	59.87	59.60
+SupCon Loss	89.98	45.96	48.74	60.41	59.99
-Backprop	89.36	40.56	47.80	58.80	59.67

for VQA v2). The effectiveness of this component can be understood from Fig. 3, where we show the TSNE visualizations of how answers from different question types are placed in the feature space for the components of *RMLVQA*. Fig. 3(a) shows that while for some question types like 'how many', 'are', answers are well separated, some others like 'do', 'are they', 'is there' are close, out of which some are entangled. Although the randomization increases flexibility of the adaptive margins, it does not disentangle these question types, as seen in Fig. 3b. Finally, in Fig. 3(c), we see that the bias-injecting component aids in separating the answers according to their question types. This is because the bias sources of VQA-CP is its question types, and hence this component clusters samples in the training set based on the same. This is crucial, as with more entanglement, the model tends to answer questions of one question type with the answers of another type. In the last row of Table 3, we also observe the effect of removing the backpropagation of the *CE* loss (used to train the bias-injecting component) into the main network for VQA-CP v2. The classifier weights in the bias-injecting component learn the language bias in the dataset and backpropagate the same into the network, thus helping the model separate the answers across different question types in the feature space. We utilize this knowledge in the inference stage, where we combine the two logit heads, thus helping the model to generalize to both *id* and *ood* test data, as is evident from Tables 1 and 2.

Role of the SupCon Loss. The addition of the SupCon loss is shown to improve the performance of our model empirically for both VQA-CP and VQA v2 as shown in Table 3.

5. Conclusion

Although margin-based losses are useful in mitigating language biases in VQA, they may not distinguish between answers of different task-complexity levels in the feature space. To distinguish between samples having similar answer frequency, but different complexity, we learn instance-based margins for each sample from the model during training. We introduce a bias-injecting component in our model and utilize its confidence in answer prediction as the instance-level margins. Being trained by the *CE* loss, this component further clusters the training samples based on the source of the data bias, guiding the margin loss to distinguish frequent and rare answers inside each cluster. We ensemble the outputs of the bias-injecting component with those of the main model during inference to achieve state-of-the-art results among existing augmentation-free methods on the *ood* VQA-CP v2 dataset, while being the most robust with respect to both *id* and *ood* test sets.

6. Acknowledgments

This work was supported by a research grant (CRG/2021/005925) from SERB, DST, Govt. of India. Abhipsa Basu is supported by the PMRF fellowship.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10044–10054, 2020. 1, 7

[2] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016. 2

[3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018. 1, 2, 5, 7

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1, 3, 6, 7

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2, 5

[6] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8102–8109, 2019. 3

[7] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. *arXiv preprint arXiv:2109.09416*, 2021. 3, 4

[8] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1989–1998, 2019. 3

[9] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019. 1, 3, 7

[10] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 1, 2, 3

[11] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. 1, 3, 7

[12] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019. 1, 3, 7

[13] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 3

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 3, 4

[15] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. *arXiv preprint arXiv:2009.08566*, 2020. 1, 3, 7

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1, 2, 5

[17] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. Adavqa: Overcoming language priors with adapted margin cosine loss. *arXiv preprint arXiv:2105.01993*, 2021. 1, 2, 3, 4, 6, 7

[18] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1584–1593, 2021. 1, 7

[19] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016. 2

[20] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11181–11188, 2020. 7

[21] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973, 2017. 1

[22] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2019. 3

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2, 5

[24] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[25] Gouthaman KV and Anurag Mittal. Reducing language biases in visual question answering with visually-grounded question encoder. In *European Conference on Computer Vision*, pages 18–34. Springer, 2020. 3, 7

[26] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[27] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3285–3292, 2020. 1, 3, 7

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[29] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 3

[30] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 3

[31] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021. 3, 7

[32] Yulei Niu and Hanwang Zhang. Introspective distillation for robust question answering. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 7

[33] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. 3

[34] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[35] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *Advances in Neural Information Processing Systems*, 31, 2018. 3, 7

[36] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600, 2019. 1, 2, 3, 7

[37] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019. 2

[38] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999. 3

[39] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*, 2020. 7

[40] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *Advances in Neural Information Processing Systems*, 33:407–417, 2020. 2, 3, 7

[41] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 2, 3

[42] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. Debiased visual question answering from feature and sample perspectives. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 3, 7

[43] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3, 7

[44] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 3, 7

[45] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 ieee winter conference on applications of computer vision (wacv)*, pages 349–357. IEEE, 2019. 3

[46] Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. Overcoming language priors with self-supervised learning for visual question answering. *arXiv preprint arXiv:2012.11528*, 2020. 1, 3, 7