applied sciences

MDPI

Article

Language Bias-Driven Self-Knowledge Distillation with Generalization Uncertainty for Reducing Language Bias in Visual Question Answering

Desen Yuan †, Lei Wang †, Qingbo Wu *, Fanman Meng, King Ngi Ngan and Linfeng Xu

School of Information and Communication Engineering, University of Electronic Science and Technology of China, Xiyuan West Road 2006, Chengdu 611731, China; desenyuan97@163.com (D.Y.); wangleiuestc@outlook.com (L.W.); fmmeng@uestc.edu.cn (F.M.); knngan@uestc.edu.cn (K.N.N.); lfxu@uestc.edu.cn (L.X.)

* Correspondence: qbwu@uestc.edu.cn

† These authors contributed equally to this work.

Abstract: To answer questions, visual question answering systems (VQA) rely on language bias but ignore the information of the images, which has negative information on its generalization. The mainstream debiased methods focus on removing language prior to inferring. However, the image samples are distributed unevenly in the dataset, so the feature sets acquired by the model often cannot cover the features (views) of the tail samples. Therefore, language bias occurs. This paper proposes a language bias-driven self-knowledge distillation framework to implicitly learn the feature sets of multi-views so as to reduce language bias. Moreover, to measure the performance of student models, the authors of this paper use a generalization uncertainty index to help student models learn unbiased visual knowledge and force them to focus more on the questions that cannot be answered based on language bias alone. In addition, the authors of this paper analyze the theory of the proposed method and verify the positive correlation between generalization uncertainty and expected test error. The authors of this paper validate the method's effectiveness on the VQA-CP v2, VQA-CP v1 and VQA v2 datasets through extensive ablation experiments.

Keywords: visual question answering; self-knowledge distillation; language bias; generalized uncertainty

Citation: Yuan, D.; Wang, L.; Wu, Q.; Meng, F.; Ngan, K.N.; Xu, L. Language Bias-Driven Self-Knowledge Distillation with Generalization Uncertainty for Reducing Language Bias in Visual Question Answering. *Appl. Sci.* **2022**, *12*, 7588. https://doi.org/10.3390/app12157588

Academic Editor: João M. F. Rodrigues

Received: 18 June 2022
Accepted: 25 July 2022
Published: 28 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Visual Question Answering (VQA) [1,2] is a cross-domain task of computer vision and natural language processing, and it has become increasingly important in the research and application of multimodal machine learning. In the past few decades, significant advances have been made in computer vision and natural language processing, with an explosion of visual and textual data to acquire and process. The most common VQA consists of an image and a question to be answered by the machine. Compared with other computer vision tasks, this model answers in real-time, not in advance. Moreover, the VQA model is required to comprehend the multimodal information of images and texts in a more artificially intelligent [3] way, leading to an in-depth understanding of vision and language.

VQA remains a challenging and open research topic. Recent research has focused on how to solve language bias. Language bias [4–8] threatens the implementation of VQA, which indicates that the current VQA model has an inadequate understanding of multimodal information. Language bias seems to be caused by the uneven distribution of datasets, a common problem in the real world. For example, if 90 percent of the bananas in the training set are yellow, the model would ask, "what color the banana is" and answer, "yellow" all the time, based on language bias. As shown in Figure 1, many VQA models tend to answer "yes" or "no" directly. Take another typical example. For the question "what color is the banana in the image?", although the banana is green, the model still tends to predict "yellow".

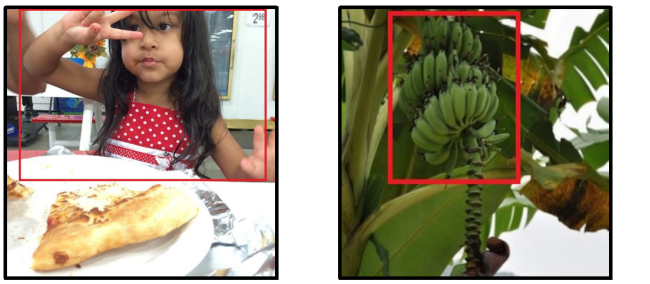

Figure 1. Example of the language bias in VQA. The output of the model is directly affected by the question. For example, the model answers "yellow" to all the questions regarding the color of the banana. If it is a yes-or-no question type, the model tends to simply answer "yes".

With language bias, the model overly relies on the correlation between the question and the answer while it ignores the information in the image. In essence, language bias arises from data imbalance, which leads to over-fitting of the model; that is, the model fits the head samples in the dataset [8]. The over-fitting of the model is an inherent problem of the model itself. The deep neural network has a variance in the case of an individual model, and the variance can be reduced by an ensemble or knowledge distillation [9–12] (Over-fitting to the label imbalance may lead to some models not training very well). At the feature level, the variance is caused by the incompleteness of the feature subgraph [12], so it is easy to produce over-fitting. Recently, knowledge [13,14] distillation and self-knowledge distillation [15,16] have been proven to be able to learn multi-view features and reduce over-fitting [12,17,18].

Neural network analysis of information is multi-view; for the same object, its different views have semantic consistency [19–21], and the multi-view structure can be ubiquitous in the dataset and feature level [12]. Therefore, the model can give the prediction based on a learned subgraph. However, if the subgraph is not comprehensive, the prediction can be biased. As shown in Figure 2, for the same question, "what color is the banana?", the model learns the feature of yellow bananas while it ignores the feature of green bananas, which are less frequent in the training set. In other words, the model ignores the view of green bananas, causing visual bias, which further leads to language bias. Therefore, the model needs to focus on the feature of the less distributed samples in the training set and learn a comprehensive set of multi-view features so as to overcome VQA language bias and over-fitting.

The paper discusses how to reduce the language bias of the VQA model via self-knowledge distillation and proposes a new online learning framework, "language bias-driven self-knowledge distillation (LBSD)", for implicit learning of multi-view visual features. Self-knowledge distillation enables the model to acquire more dark knowledge and improves its generalization ability. In short, with self-knowledge distillation, the model can have a more comprehensive understanding of view features. Online knowledge distillation no longer uses teacher models but allows student models to learn from each other by using KL divergence to uniformly constrain the output. It is worth mentioning that the student network is actually equal to the teacher network; the two networks are the same. However, the learning degree of student models cannot be described by using KL divergence alone [22,23]. Therefore, the authors of this paper put forward the concept of generalization uncertainty to help the model learn unbiased knowledge.

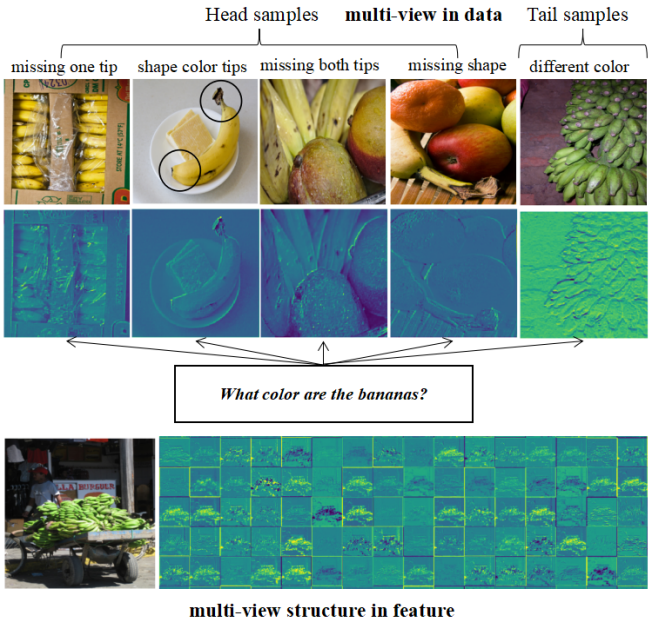

Figure 2. The VQA-CP v2 dataset contains images with multi-view (features). The authors of this paper visualize the features at the same layer in the neural network. For the same question, although the images, views and features are different, the semantic information is the same. Broadly speaking, this "multi-view" structure [12] exists both in the original data and the feature sets extracted from the middle layer.

LBSD enables two debiased models to distill knowledge from each other to learn more complete visual features. It distinguishes between debiased students and biased students by calculating the generalization uncertainty of the prediction of student models and reinforces the mutual learning of the two models about unbiased knowledge. The paper also finds that heterogeneous student models can be used to reduce language bias. LBSD enables the model to learn a more complete set of visual features and to focus on the features of the less distributed samples in the training set by utilizing generalization uncertainty, thus reducing the language bias of the model and improving the robustness of the VQA model.

Contribution. In summary, the contributions of this paper are as follows:

(1) The authors of this paper propose a training framework (LBSD) based on online self-knowledge distillation, which can considerably reduce the VQA language bias. Moreover, the authors of this paper explore the different cases of student models (heterogeneous networks). The authors of this paper verify the effectiveness of the LBSD method and analyze the theory behind it.

(2) The authors of this paper propose a method to measure generalization uncertainty based on Top-k information entropy, and use it to distinguish between debiased students and biased students, so as to force the model to focus on the samples that cannot be directly answered by language bias in the VQA datasets. The authors of this paper also prove the proportional relationship between the generalized uncertainty and the expected test error.

2. Related Work

2.1. Language Bias in VQA

The language bias [8] in VQA has a negative impact on the general application of the model in real-world scenarios. The reason behind it is that there is often a strong correlation between questions and answers. Moreover, the questions tend to concern conspicuous objects in the image. In VQA v1 [1] and v2 [7], a positive answer or a question-related answer tends to have higher accuracy. When the questions and answers in the training set and the test set are distributed inconsistently, this language bias is obvious. Therefore, the VQA-CP v2 dataset was recently proposed to evaluate the language bias.

Train and test splits of the VQA-CP v2 have different question-answer distributions. The current approach to language bias can be divided into (1) Strengthening visual information: AttAlign [24], HINT [24], SCR [25], ReGAT [26], ESR [27], VGQE [28] and so on; (2) Weakening language priors: AdvReg [29], GRL [30], RUBi [31], LM [32], LMH [32], Bias-Product (POE) [32], RMFE [33], CF-VQA [34] and GGE-DQ [35]; (3) Using various data enhancement: CSS [36], CL-VQA [37], GradSup [38], Loss-Rescaling [39], Mutant [40], RandImg [41], Unshuffling [42], ADA-VQA [43] and X-GGM [44].

2.2. Knowledge Distillation

In recent years, knowledge distillation [45–48] has been widely used in deep learning to transfer knowledge between different models. Hinton et al. [13] used knowledge distillation for model compression; that is, moving knowledge from powerful but complex models (teacher models) to simple models (student models). By minimizing the Kullback–Leibler (KL) divergence loss of the categorical output probability, the student can imitate the output of the teacher model. In addition, some new knowledge transfer goals have been proposed, such as intermediate feature maps [49], attention maps [50], second-order statistics [46], contrastive features [51,52] or structured knowledge [53–55].

However, these methods require a distinction between the roles of the teacher and the student and are typically distilled offline. Online knowledge distillation is a knowledge distillation based on a series of student (generally two) models by eliminating cumbersome teacher models. Based on the Kullback–Leibler divergence, Zhang et al. [16] proposed a technique for deep mutual learning (DML) in which pair-wise students learn from each other using a mimicry loss. By adding distillation loss after updating enough steps, co-distillation [15] (similar to DML) enables student networks to sustain their diversity for a longer time. However, KL divergence alone cannot capture the learning degree of student models. The authors of this paper put forward the notion of generalization uncertainty as a way for the model to learn unbiased knowledge.

3. Methods

In order to reduce VQA language bias, the authors of this paper consider making the model focus on the less distributed samples in the training set to learn a more complete set of multi-view features. To this end, the authors of this paper propose a new online self-knowledge distillation learning framework (LBSD) for implicit learning of multi-view visual feature sets to alleviate language bias. The methods are divided into: (1) language bias-driven self-knowledge distillation and (2) using generalization uncertainty to help student models learn unbiased visual knowledge. In the following sections, the authors of this paper explain the workflow of LBSD and analyze the theory behind it. The block diagram of the method presented in this paper is shown in Figure 3 and Algorithm 1.

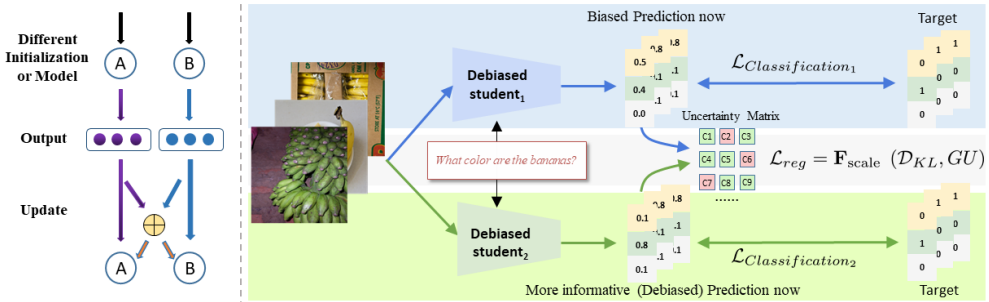

Figure 3. The flowcharts of the language bias-driven self-distillation framework, including: (1) language bias-driven self-knowledge distillation and (2) using generalization uncertainty to help student models learn unbiased visual knowledge.

Algorithm 1: Language Bias-Driven Self-Distillation

Input: Training set $\mathcal{I}$,$\mathcal{Q}$ ($\mathcal{X}$), label set $\mathcal{A}$ ($\mathcal{Y}$), learning rate $\gamma_{1,t}$ and $\gamma_{2,t}$.
Initialize: Debiased Models N_1 and N_2 (different initial conditions or models).
Repeat :
$t = t + 1$
Randomly sample data I_i, Q_i from $\mathcal{I}$,$\mathcal{Q}$.
1: Update the predictions $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ of I_i, Q_i for the current mini-batch
2: Compute the stochastic gradient and update N_1 by equation (13) :

$$N_1 \leftarrow N_1 + \gamma_{1,t} \frac{\partial L_{N_1}}{\partial N_1}$$

3: Update the predictions $\boldsymbol{p}_1$ of I_i, Q_i.
4: Compute the stochastic gradient and update Θ_2 :

$$N_2 \leftarrow N_2 + \gamma_{2,t} \frac{\partial L_{N_2}}{\partial N_2}$$

5: Update the predictions $\boldsymbol{p}_2$ of I_i, Q_i.
Until : convergence

3.1. Preliminaries

To tackle the multi-class classification problem in VQA field, the general form of VQA is: A dataset is given $\mathcal{D} = \{I_i, Q_i, a_i\}^N$ containing N triplets of images $I_i \in \mathcal{I}$, questions $Q_i \in \mathcal{Q}$ and answers $a_i \in \mathcal{A}$.

The aim of the VQA task is to learn a mapping function f_{vqa}:$I \times Q \rightarrow [0,1]^{|\mathcal{A}|}$, which generates the answer distributions for any given image-question pair. The authors of this paper omit subscript i in the following.

For each question Q and image I, the Bottom-Up Top-Down (UpDn) [56] model uses a question encoder e_q and an object detector separately i_q to extract a set of word embeddings $\boldsymbol{Q}$ and a set of visual object embeddings $\boldsymbol{V}$. The model is fed both $\boldsymbol{V}$ and $\boldsymbol{Q}$ to get the joint feature $mm(V, Q)$. Then, the joint features are fed into the classifier $\boldsymbol{C}$ to get the final predictions.

$$P_{vqa}(a|I, Q) = f_{vqa}(\boldsymbol{V}, \boldsymbol{Q}) = \boldsymbol{C}(mm(V, Q)) \tag{1}$$

For fair comparisons, the authors of this paper use the Bottom-Up Top-Down (UpDn) model [56], which is mainly used by many researchers as the backbone network.

3.2. Language Bias-Driven Self-Distillation

The method aims to learn unbiased visual knowledge via the mutual learning of two debiased models so as to reduce VQA language bias. The training strategy, which can be integrated with the current debiased methods, consists of the mutual learning of two debiased models. A dataset is given $\mathcal{D} = \{I_i, Q_i, a_i\}^N$ containing N triplets of images $I_i \in \mathcal{I}$, questions $Q_i \in \mathcal{Q}$ and answers $a_i \in \mathcal{A}$, it can be input into two identical models with different random initializations, $N1$ and $N2$, and two probability vectors p can be predicted by the model, z means Softmax output.

$$p_1^k(I_i, Q_i) = \frac{\exp\left(z_1^k\right)}{\sum_{k=1}^{K} \exp\left(z_1^k\right)} \tag{2}$$

where k represents the number of outputs or classes of the neural network.

At the same time, the VQA model is generally defined as multi-type. Therefore, for multiple types, the objective function of the training network $N1$ is defined as the cross-

entropy error between the prediction and the correct label, as shown as follows, K means samples, M means classes and L_C means the cross entropy error:

$$L_C = -\sum_{i=1}^{K}\sum_{m=1}^{M} I(a_i, m)\log(p_1^m(I_i, Q_i)) \tag{3}$$

In order to allow the two student models to learn unbiased visual features from each other (similar to self-knowledge distillation), the authors of this paper use KL divergence to constrain all the predictions, thus distilling the unbiased knowledge of the two models. The formula of KL divergence between $N1$ and $N2$ is shown as follows:

$$D_{KL}(\boldsymbol{p}_2 \| \boldsymbol{p}_1) = \sum_{i=1}^{K}\sum_{m=1}^{M} p_2^m(I_i, Q_i)\log\frac{p_2^m(I_i, Q_i)}{p_1^m(I_i, Q_i)} \tag{4}$$

The two student models simultaneously start parameter optimization, and the optimization loss is shown as follows. The consistency constraint of the predictions of the two models can realize the mutual learning of unbiased knowledge between the two models.

$$\begin{aligned} L_{N_1} &= L_{C_1} + D_{KL}(\boldsymbol{p}_2 \| \boldsymbol{p}_1) \\ L_{N_2} &= L_{C_2} + D_{KL}(\boldsymbol{p}_1 \| \boldsymbol{p}_2) \end{aligned} \tag{5}$$

Since KL divergence is asymmetric, it can be replaced by Jensen–Shannon (JS) divergence (a variation of KL divergence) to ensure the consistency constraint between the two student models. Such replacement will not affect the final precision of the model.

$$L_{JS(\boldsymbol{p}_1 \| \boldsymbol{p}_2)} = \frac{1}{2}KL\left(\boldsymbol{p}_1 \| \frac{\boldsymbol{p}_1 + \boldsymbol{p}_2}{2}\right) + \frac{1}{2}KL\left(\boldsymbol{p}_2 \| \frac{\boldsymbol{p}_1 + \boldsymbol{p}_2}{2}\right) \tag{6}$$

Moreover, all the current self-knowledge distillation models use student models with different random initializations. The strategy is effective because the model learns more complete sets of multi-view features. The authors of this paper also explore the case where two heterogeneous student networks serve as the student models. The heterogeneous networks have the same feature extraction structure, but they have different loss functions and network branches.

3.3. Debiased Mutual Students

As mentioned above, the language bias of datasets is, in essence, the distribution bias of image samples. For the same input image/text sample pair, the two student networks may have different outputs because of an inconsistent random seed, order of data reading or even network structure.

As shown in Figure 4, for more-distributed image/text sample pairs in the dataset, the model can simply answer the question through language bias, and the confidence of the answer is very high. The different student models tend to have the same answer. For the gradient update of neural networks, the cross-entropy loss and KL divergence loss of image/text samples that can answer the question by language bias are minimal. However, for the less-distributed samples, the model is more likely to have different answers. Therefore, the different answers can be measured and analyzed to help the model focus more on the samples that cannot be directly answered by language bias so as to reduce language bias.

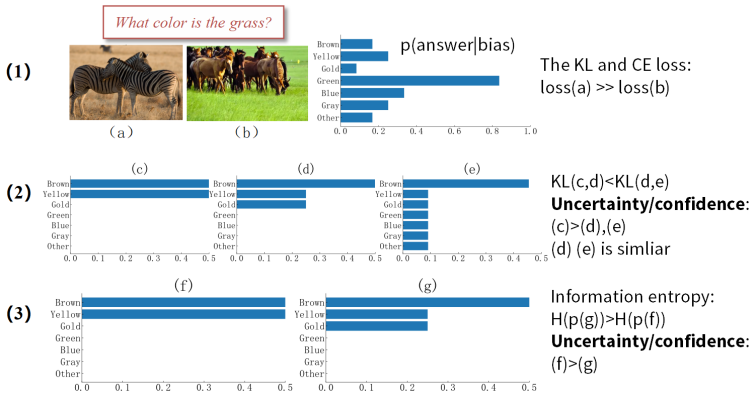

Figure 4. Examples used to show the difference between KL divergence and uncertainty.

In general, the current self-distillation methods only use KL divergence for the mutual distillation of knowledge. As KL divergence is not commutative, it cannot be understood as "distance", which measures the information loss between two distributions. Simply constraining the KL divergence of the two student models cannot figure out the difference between the output and help the two models learn from each other with more precision. As shown in Figure 4, KL divergence for different distributions and consistency constraints is not always consistent with our expectations. For this reason, the authors of this paper consider using information entropy to evaluate the output uncertainty of the two models and evaluate the output difference based on the uncertainty.

As shown in Figure 4, although information entropy H is a common method to measure information uncertainty, the output is not always consistent with our understanding. For p_1 = [0.5, 0.25, 0.25] and p_2 = [0.5, 0.5, 0], the formula leads to $H(p_a) > H(p_b)$. For general classification scenarios, it is clear that p_b is less certain than p_a, and the confidence of predictions is extremely low. Therefore, in order to describe the prediction uncertainty, the authors of this paper adopt a simple and improved version: Top-k information entropy.

Suppose that $p_1, p_2, \ldots, p_k$ are k values with the highest probability, the following formula can be obtained:

$$H_{noraml}(p) = -\sum_{i=1}^{m} p_i \log p_i \tag{7}$$

$$H_{\text{top}-k}(p) = -\sum_{i=1}^{k} \tilde{p}_i \log \tilde{p}_i \tag{8}$$

$$\tilde{p}_i = p_i / \sum_{i=1}^{k} p_i \tag{9}$$

$$C = H_{\text{top-}k}(p) / \log k \tag{10}$$

By using the above formula, the authors of this paper can get a result in the range of 0 to 1 and take C as the final uncertainty measure.

In order to measure the output difference between the two student models, for the uncertainty $C1$ and $C2$, the output difference can be defined as $|C1 - C2|$. In order to enhance the mutual learning of the two student models in the case of output difference (questions that cannot be answered directly with language bias), the authors of this paper define a generalization uncertainty index GU to represent the intensity. The formula is $GU = e^{|C1-C2|}$, and the final loss function of the generalization uncertainty index can be obtained. The formula is as follows:

$$L_{reg} = \mathbf{F}_{\text{scale}}(GU, D_{KL}) = e^{|C1-C2|} D_{KL}(\boldsymbol{p}_2 \| \boldsymbol{p}_1) \tag{11}$$

$$L_{N_1} = L_{C_1} + L_{reg} \tag{12}$$

In the next section, the authors of this paper will prove that the generalization uncertainty index *GU* of the two student models can be used to estimate the test error of the model.

3.4. Theoretical Analysis

3.4.1. Theoretical Analysis of Generalization Uncertainty (GU)

In this section, the authors of this paper demonstrate that the generalized uncertainty index between two student models can be used to estimate the model test error on image-text sample pairs. Thus, generalized uncertainty is used in the training process to help students learn unbiased knowledge. Following the research of Nakkiran and Bansal [57], Jiang et al. [58] and others [59–61], the authors of this paper use class-segregated calibration (or class-wise calibration) [58,62–66] to prove the proportional relationship between the generalized uncertainty and the test error.

Notation 1. *The authors in this paper define two neural networks trained from different random seeds as n, n'. The data of this model include K categories with input I_i, Q_i and label a_i (Y). The model is parameterized by stochastic learning. The probability expression of the predicted output of the model is $\mathrm{p}(\hat{a} \mid I, Q)$. D_{vqa} is the distribution map from $I_i, Q_i \times [K]$, and $\mathrm{p}(\omega)$ is the sample estimate for the different parameter distribution of models. The parameters of the model can be defined as Ω. The $1[\ldots]$ function is the indicator function, which means the prediction is true or otherwise.*

Definition 1. *The model N ($\mathrm{p}(\hat{a}, \omega \mid I, Q)$) satisfies the generalization uncertainty proportional (GUP) on the distribution D_{vqa} if:*

$$\begin{gathered} TestErr_{D_{vqa}}(n) E_{D_{vqa}}[1|n(I,Q) \neq a] \propto \\ GUErr_{D_{vqa}}(n,n')(TopK) E_{D_{vqa}}[1|n(I,Q) \neq n'(I,Q)] \end{gathered} \tag{13}$$

Definition 2. *The self-knowledge distillation model N (n, n') satisfies class-wise calibration (or class-segregated calibration) on D_{vqa} if any kind of confidence value q falls in $[0, 1]$ and for any class k falls in $[K]$,*

$$p(a = k \mid \tilde{n}_k(I,Q) = q) = q \tag{14}$$

$$\frac{\sum_{k=0}^{K-1} p(a = k, \tilde{n}_k(I,Q) = q)}{\sum_{k=0}^{K-1} p(\tilde{n}_k(I,Q) = q)} = q \tag{15}$$

Theorem 1. *If the self-knowledge distillation model N (n, n') satisfies class-wise calibration (or class-segregated calibration) on D_{vqa}, then N satisfies the generalization uncertainty proportional (GUP) on D_{vqa}.*

$$\mathbb{E}_{\Omega,\Omega'}\left[\mathrm{GUErr}_{D_{vqa}}(n,n')\right] \propto \mathbb{E}_{\Omega}\left[TestErr_{D_{vqa}}(n)\right] \tag{16}$$

Proof. The authors in this paper define the Expected Test error as TE (Test error). The TOP-K error of the two-student model with generalized uncertainty is fixed at GUE. By simplifying the two errors, the following results can be obtained, and the proportional relationship (GUP) between them can be obtained. Since *K* previously represented the number of categories, for this reason, the authors of this paper represent *J* as the *K* term in TOP-K, *i* as the corresponding prediction at the sample of *J*-th value.

$$\mathrm{TE} = \int_{q \in [0,1]} q(1-q) \sum_{k=0}^{K-1} p(\tilde{n}_k(I,Q) = q) dq \tag{17}$$

$$\begin{aligned} \mathrm{GUE} &= \sum_{j=0}^{J} 1/p_i \underbrace{\sum_{k=0}^{K-1} \int_{q\in[0,1]}}_{\text{swap}} p(\tilde{n}_k(I,Q)=q)q(1-q)dq \\ &= \sum_{j=0}^{J} 1/p_i \int_{q\in[0,1]} q(1-q) \sum_{k=0}^{K-1} p(\tilde{n}_k(I,Q)=q)dq \propto TE. \end{aligned} \tag{18}$$

The detailed proof of generalization uncertainty can be found in Appendix A. □

3.4.2. Theoretical Analysis of Debiased Self-Distillation

In this section, the authors of this paper demonstrate that self-knowledge distillation and generalized uncertainty can enable models to learn more complete multi-view feature sets and reduce language bias in VQA. The authors of this paper followed the research of Allen Zhu and Zhiyuan Li [12,19].

Notation 2. *Let us set up a model whose dataset contains K categories, p-input patch s and the ReLU function. The model input is I_i, Q_i and the label is a_i. To simplify the problem, the authors of this paper assume that each category contains related features that are orthogonal to each other. The authors of this paper define these features as vectors of $vqa_{j,1}$, $vqa_{j,2}$.*

Following the settings of Zhu et al.'s research. The authors of this paper get the definitions as follows. The set of all features:

$$X_{vqa} \stackrel{\text{def}}{=} \left\{vqa_{j,1}, vqa_{j,2}\right\}_{j\in[k]} \tag{19}$$

$$vqa_{j,\ell} \perp vqa_{j',\ell'} when (j,\ell) \neq (j',\ell') \tag{20}$$

Definition 3. *(Data distribution) The authors of this paper define the multi-view and single-view distribution D_{vqam} and D_{vqas}, $D \in D_{vqam/s}$, and $(I_i, Q_i, a_i) \sim D$. Sample features with probability s/k, $s \in \left[1, k^{0.2}\right]$. The coefficients z_p, $n_{p,vqa'}$ is the feature noise, ξ_p is the random Gaussian noise. For each $p \in [P] \backslash \mathcal{P}(I_i, Q_i)$, the authors of this paper set:*

$$x_p = z_p vqa + \sum_{vqa' \in X_{vqa}} n_{p,vqa'} vqa' + \xi_p \tag{21}$$

Definition 4. *(The final data distribution D and the training dataset S_d. Suppose D contains $1-\mu$ D_{vqam} and μ D_{vqas}. For N samples in D, the training dataset $S_d = S_{dm} \cup S_{ds}$. (I_i, Q_i, a_i) random sampling from the set S_d. $\mu = \frac{1}{\mathrm{poly}(k)}$, and $N = k^{1.2}/\mu$.*

Definition 5. *The authors of this paper define a network $VQA(I_i, Q_i)$ with a cross-entropy loss function using a stochastic learning algorithm as follows:*

$$L(VQA) = \mathbb{E}_{(I_i,Q_i,a_i)\sim S_d}[L(VQA; I_i, Q_i, a_i)] \tag{22}$$

The logits function of the single model ($\eta \leq \frac{1}{\mathrm{poly}(k)}$, $T = \frac{\mathrm{poly}(k)}{\eta}$) can be defined as

$$\mathrm{logit}_i(VQA, I, Q) \stackrel{def}{=} \frac{e^{VQA_i(I,Q)}}{\sum_{j\in[k]} e^{VQA_j(I,Q)}} \tag{23}$$

The logits function of the model using knowledge distillation can be defined as

$$\mathrm{logit}_i^{\tau}(VQA, I, Q) = \frac{e^{\min\left\{\tau^2 VQA_i(I,Q),1\right\}/\tau}}{\sum_{j\in[k]} e^{\min\left\{\tau^2 VQA_j(I,Q),1\right\}/\tau}} \tag{24}$$

Theorem 2. *For the single model, the authors of this paper use the prediction error as follows,* $\exists i \in [k] \backslash \{a\}$*:*

$$\Pr_{(I,Q,a)\sim D}\left[VQA_a^{(T)}(I,Q) < VQA_i^{(T)}(I,Q)\right] \in (0.5 \pm 0.01)\mu \tag{25}$$

Theorem 3. *For self-knowledge distillation with the generalization uncertainty model,* λ *(*$\lambda > 1$*) is the gain from generalized uncertainty. The authors of this paper use the prediction error as follows,* $\exists i \in [k] \backslash \{y\}$*:*

$$\Pr_{(I,Q,a)\sim D}\left[VQA_a^{(T+T')}(I,Q) < VQA_i^{(T+T')}(I,Q)\right] \leq 0.26\mu/\lambda \tag{26}$$

The authors of this paper find that when comparing Theorems 2 and 3, the prediction error of the model decreased. That means the LBSD method can reduce language bias. The detailed proof can be found in Appendix A.

4. Settings, Results and Discussion

In this section, the authors of this paper evaluate the effectiveness of all the LBSD methods in the three mainstream datasets (VQA-CP v2, VQA-CP v1 and VQA v2), carry out an ablation experiment with the typical debiased method and compare the performance of the LBSD methods and that of the latest method. Table 1 shows the statistics of all the datasets.

Table 1. Statistics of VQA-CP v2, VQA v2 and VQA-CP v1.

Dataset	VQA-CP v2 [67]			VQA-CP v1 [67]			VQA v2 [7]		
	Train	Test	Total	Train	Test	Total	Train	Test	Total
Images	121 K	98 K	219 K	118 K	87 K	205 K	440 K	214 K	654 K
Questions	438 K	220 K	658 K	245 K	125 K	370 K	83 K	41 K	124 K
Answers	4.4 M	2.2 M	6.6 M	2.5 M	1.3 M	3.8 M	4.4 M	2.1 M	6.5 M

4.1. Experimental Settings Setup

4.1.1. Datasets and Backbone

The paper uses the standard VQA evaluation metric [1] to evaluate the performance of the model on the VQA-CP v2 [67], VQA-CP v1 [67] and VQA v2 [7] datasets. For fair comparisons, all the methods are based on the UpDn model, and their best-recorded performance is compared. The experiment trains and tests the models on two Titan Xp GPUs.

Currently, for the VQA language bias issue, researchers evaluate the performance of the proposed models on the VQA-CP v2 dataset and conduct auxiliary verification on the VQA v2 dataset. Most findings test the models on VQA-CP v2 and VQA v2 and calculate the gap index [36] as an auxiliary index to verify the robustness of the model.

VQA-CP v2. The researchers propose the VQA-CP v2 dataset, which is derived from the re-classification of the samples in the VQA v2 dataset, to measure language bias. The VQA-CP v2 and VQA-CP v1 datasets are the only open-source datasets for language bias evaluation. The questions and answers in the training and testing sets are distributed in considerably different ways. In other words, for the same type of questions, the answers in the training set and testing set are distributed very differently. Therefore, the VQA-CP v2 dataset is suitable for measuring the language bias of the models. The training set consists of 121 K images, 438 K questions and 4.4 million answers, and the testing set consists of 98 K images, 220 K questions and 2.2 million answers.

VQA-CP v1. The VQA-CP v1 dataset, the first version of the VQA-CP dataset, is the first-ever dataset for language bias evaluation. It is derived from the re-classification of the VQA v1 [1] dataset. The VQA-CP v1 training set consists of 118 K images, 245 K questions

and 2.5 million answers, and the VQA-CP v1 testing set consists of 87 K images, 125 K questions and 1.3 million answers.

VQA v2. The VQA v2 dataset is the second version of the VQA dataset. The training set consists of 82,783 images, 443,757 questions and 4,437,570 answers. The testing set consists of 40,504 images, 214,354 questions and 2,143,540 answers. The VQA v2 dataset is double the VQA v1 dataset in size.

4.1.2. Experimental Details

For LBSD, the k in generalization uncertainty is set at 3, and the KL divergence coefficient is set at 2 or 3. The basic VQA network UpDn uses a pre-trained Faster-RCNN to extract image features, a pretrained model GloVe (300 dimensions) to extract text features and a single-layer GRU to obtain question-embedded vectors (512 dimensions). Finally, the joint embedding is 2048 dimensions. In addition, the batch size is set at 512 and trained and tested on two Titan Xp GPUs. Because VQA-CP v1 and v2 lack validation datasets, VQA v2 datasets generally display results on validation datasets. In order to select the parameters of the model, the authors of this paper divide 10% of the samples from the test datasets or the validation datasets to act as the validation dataset, select the parameters of the model on the validation datasets and then test the precision of the model on the test datasets. The results of our experiments are based on the results of the original published papers, and for experiments that were not performed in the original papers, we reproduced them using the official code, and for the results that we reproduced, we put an asterisk in the upper right-hand corner. With regard to run time, the proposed method runs 30 epochs for 15 h in a 256 GB memory and two Titan XP GPUs environment. For the statistical analysis, we performed multiple experiments with a confidence of 95% for the experimental results, and for the purpose of fillability of the experimental results, we selected the median of the results of multiple experiments as the final result, the final precision and the precision of each index are filled in the table.

4.2. Ablation Studies

To verify the effectiveness of LBSD, the authors of this paper conduct an ablation experiment on every aspect. For fair comparisons, the authors of this paper select the mainstream VQA network UpDn as the skeleton and carry out ablation experiments on typical debiased methods such as Bias product, Reweight and LMH. In these tables, * indicates the results of our reimplementation from the official code.

4.2.1. Architecture Agnostic

Since LBSD is irrelevant to the model, it can be integrated into various VQA networks. To evaluate the performance of LBSD on debiased methods, the authors of this paper combine it with other typical methods and baseline, including UpDn, Bias product (Product of Experts), reweight and LMH. Reweight, a non-ensemble method, encourages the model to focus on the samples that are predicted erroneously by the language bias model. While Bias product and LMH are ensemble models. Compared with these, the LBSD-integrated models have higher precision.

The authors of this paper conduct ablation experiments on the VQA-CP v2 and VQA-CP v1 datasets. As shown in Table 2, for typical debiased methods, including ensemble and non-ensemble methods, LBSD improves the precision of the model on the VQA-CP v2 dataset. For example, the performance of reweighting (non-ensemble) and LMH (ensemble) improves by 1.26% and 2.22%, respectively. Even for UpDn without debiased methods, LBSD improves the precision by 0.25%, which demonstrates that LBSD reduced the language bias from the perspective of feature learning. As shown in Table 3, for reweight (non-ensemble) and bias product (ensemble), LBSD improves the performance by 2.2% and 0.58% ("NUM" index has been improved by 7.51%), respectively, on the VQA-CP v1 dataset.

Table 2. VQA-CP v2: Ablation experiments of the LBSD method on the VQA-CP v2 dataset. * indicates the results from our reimplementation using officially released codes.

Model	Overall	Yes/No	Num	Other
UpDn [56]	39.74	42.27	11.93	46.05
+**LBSD**	**39.99**	42.76	12.36	46.12
Bias Product	39.93	–	–	–
Bias Product *	39.86	41.96	12.59	46.25
+**LBSD**	**40.47**	44.28	12.28	46.21
Reweight	40.06	–	–	–
Reweight *	40.02	45.09	12.30	44.96
+**LBSD**	**41.28**	47.07	12.30	46.20
LMH	52.05	69.81	44.46	45.54
+**LBSD**	**54.27**	75.49	44.02	45.96

Table 3. VQA-CP v1: Ablation experiments of the LBSD method on the VQA-CP v1 dataset. * indicates the results from our reimplementation using officially released codes.

Model	Overall	Yes/No	Num	Other
UpDn [56]	37.87	42.58	14.16	42.71
+**LBSD**	**38.55**	**43.29**	12.90	**44.13**
Bias Product *	38.81	42.96	13.34	44.91
+**LBSD**	**39.39**	**45.07**	13.08	44.32
Reweight *	41.46	61.52	13.02	32.94
+**LBSD**	**43.66**	**66.63**	12.23	**33.45**
LMH	55.27	76.47	26.66	45.68
+**LBSD**	**55.93**	75.43	**34.17**	45.28

4.2.2. Effectiveness of GU

To verify the effectiveness of generalization uncertainty in the reduction of language bias, the authors of this paper conduct ablation experiments on VQA-CP v2. Two debiased methods, including Reweight (non-ensemble) and LMH (ensemble), are selected for verification. As shown in Table 4, the results show that, compared with LBSD without the generalization uncertainty constraint, LBSD with the generalization uncertainty constraint improves the performance by 0.27% and 0.63%, respectively, on Reweight and LMH. For the question types "YES/NO" and "Other" that are highly dependent on language bias, the generalization uncertainty constraint can be added to reduce the language bias of these question types.

Table 4. VQA-CP v2: Ablation experiments of the generalization uncertainty method on the VQA-CP v2 dataset. * indicates the results from our reimplementation using officially released codes.

Model	Overall	Yes/No	Num	Other
Reweight	40.06	–	–	–
Reweight *	40.02	45.09	12.30	44.96
+**LBSD** without GU	**40.99**	**46.34**	12.55	**45.98**
+**LBSD** with GU	**41.28**	**47.07**	12.30	**46.20**
LMH	52.05	69.81	44.46	45.54
+**LBSD** without GU	**53.64**	**75.44**	40.19	**45.91**
+**LBSD** with GU	**54.27**	**75.49**	44.02	**45.96**

4.2.3. Heterogeneous Student Networks

Generally, the student models of self-knowledge distillation have identical network structures. The authors of this paper also explore heterogeneous student networks, where the two student models are not identical. The authors of this paper select two debiased methods based on the UpDn model to verify the effectiveness of heterogeneous student networks. As shown in Table 5, heterogeneous student networks can have similar effects to homogeneous student networks. Moreover, the precision of the two heterogeneous student models is improved.

Table 5. VQA-CP v2: Heterogeneous student networks. * indicates the results from our reimplementation using officially released codes.

Model	Overall	Yes/No	Num	Other
Bias Product *	39.86	41.96	12.59	46.25
+**LBSD**-with LMH	40.06	43.31	12.41	45.94
+**LBSD**-same stu	**40.47**	44.28	12.28	46.21
Reweight *	40.02	45.09	12.30	44.96
+**LBSD**-with LMH	40.60	44.92	12.56	46.03
+**LBSD**-same stu	**41.28**	47.07	12.30	46.20
LMH	52.05	69.81	44.46	45.54
+**LBSD**-with Bias product	53.89	75.25	43.58	45.52
+**LBSD**-with Reweight	53.82	75.56	41.79	45.73
+**LBSD**-same stu	**54.27**	75.49	44.02	45.96

4.3. Comparisons with State-of-the-Arts

To evaluate the performance of LBSD, the authors of this paper carry out an experiment on VQA-CP v2, VQA-CP v1, and VQA v2 and compare it with the state-of-the-art method. In these tables, * indicates the results of our reimplementation from the official code.

4.3.1. Performance on VQA-CP v2

Setting. The authors of this paper combine LBSD with LMH and name it LBSD-LMH. For fair comparisons, the authors of this paper choose the debiased method based on UpDn. According to the principles of reducing language bias, the authors of this paper divide the methods into groups: (1) Strengthening visual information [24,25]. (2) Weakening language priors [29,31,32]. (3) Using various data enhancement and data balance [36,68].

Since LBSD improves the performance by enabling the model to focus more on visual information and difficult samples (the model cannot answer based on language bias), the authors of this paper compare other methods with those in the first and second groups. Moreover, according to the experiment settings of CSS [36], the authors of this paper test and calculate the gap index as an auxiliary index on VQA v2 to verify the robustness of the model.

Results. Comparisons are reported in Table 6. As shown in Table 6, compared with other methods with UpDn as the standard VQA model, LBSD improves the performance on VQA-CP v2. The gap index has also been improved ("All" and "Other"). The results show that the proposed LBSD can reduce language bias in VQA. For individual items, such as Num, yes/no and others, CFVQA is slightly higher than our method in num index; it is an ensemble method based on causal inference. Similar to boosting, CFVQA uses more

ensemble networks as additional information than ours. Therefore, it is unfair to compare directly on small indices.

Table 6. Performance comparison (Accuracies (%)) on the VQA-CP v2 test set and the VQA v2 val set of state-of-the-art models. The gap means the performance difference between VQA v2 and VQA-CP v2.

Model	Venue	VQA-CP v2 Test ↑				VQA v2 val ↑				Gap Δ↓	
		All	Yes/No	Num	Other	All	Yes/No	Num	Other	All	Other
GVQA [67]	*CVPR'18*	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65	16.94	12.51
UpDn [56]	*CVPR'18*	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66	23.74	9.61
methods based on strengthening visual information											
AttAlign [24]	*ICCV'19*	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22	23.87	10.22
HINT [24]	*ICCV'19*	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56	16.55	9.68
ReGAT [26]	*ICCV'19*	40.42	–	–	–	67.18	–	–	–	26.76	–
VGQE [28]	*ECCV'20*	48.75	–	–	–	64.04	–	–	–	15.29	–
ESR [27]	*ACL'20*	48.9	69.8	11.3	47.8	62.6	–	–	–	13.70	–
KAN [69]	*TNNLS'20*	42.60	42.12	15.52	50.28	–	–	–	–	–	–
methods based on weakening language priors:											
AReg [29]	*NeurIPS'18*	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16	21.58	19.68
GRL [30]	*ACL'19*	42.33	59.74	14.78	40.76	51.92	–	–	–	9.59	–
RUBi [31]	*NeurIPS'19*	45.23	64.85	11.83	44.11	50.56	49.45	41.02	53.95	5.33	9.84
LM [32]	*EMNLP'19*	48.78	72.78	14.61	45.58	63.26	81.16	42.22	55.22	14.48	9.64
LMH [32]	*EMNLP'19*	52.05	69.81	44.46	45.54	61.64	77.85	40.03	55.04	9.59	9.50
CF-VQA [34]	*CVPR'21*	53.55	91.15	13.03	44.97	63.54	82.51	43.96	54.30	9.99	9.33
LBSD-LMH	*Ours*	**54.27**	75.49	44.02	**45.96**	57.95	69.22	38.19	54.63	**3.68**	**8.67**
methods based on data argumentation											
RandImg [70]	*NeurIPS'20*	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57	1.87	4.37
CVL [68]	*CVPR'20*	42.12	45.72	12.45	48.34	–	–	–	–	–	–
LMH+CSS [36]	*CVPR'20*	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11	0.96	6.90
LMH+CSS+CL [37]	*EMNLP'20*	59.18	86.99	49.89	47.16	57.29	67.27	38.40	54.71	1.89	7.55
Unshuffling [42]	*ICCV'21*	42.39	47.72	14.43	47.24	61.08	78.32	42.16	52.81	18.69	5.57
X-GGM [44]	*ACM MM'21*	45.71	43.48	27.65	52.34	–	–	–	–	–	–

4.3.2. Performance on VQA-CP v1

Settings. The authors of this paper compare the state-of-the-art methods to LBSD-LMH and VQA-CP v1. According to the principle and method of reducing language bias, the authors of this paper divide them into groups: (1) Strengthening visual information [24,25]. (2) Weakening language priors [29,31,32]. (3) Using various data enhancement and data balance [36,68]. Moreover, the authors of this paper conducted another experiment based on the official codes of the methods, as the results of some methods on VQA-CP v1 were not shown.

Results. As shown in Table 7, compared with the methods in group 1 and group 2, LBSD realizes the best performance on VQA-CP v1. In particular, LBSD improves the performance of LMH and Reweight by 0.66% and 2.2%, respectively. The results show that the proposed method is effective for different datasets and is effective for different types of debiased methods. The results verify the effectiveness of LBSD.

Table 7. Performance comparison (Accuracies %) with the state-of-the-art model's accuracy on the VQA-CP v1 test. * indicates the results from our reimplementation using officially released codes.

Model	All	Yes/No	Num	Other
GVQA [67]	39.23	64.72	11.87	24.86
UpDn [56]	37.87	42.58	14.16	42.71
Group 1				
Reweight * [32]	41.46	61.52	13.02	32.94
LBSD-Reweight	**43.66**	66.63	12.23	33.45
Group 2				
AReg [29]	41.17	65.49	15.48	35.48
RUBi [31]	44.81	69.65	14.91	32.13
LMH [32]	55.27	76.47	26.66	45.68
LBSD-LMH	**55.93**	75.43	34.17	45.28
Group 3				
CSS [36]	60.95	85.60	40.57	44.62
CSS+GS [38]	58.05	78.50	37.24	46.08

4.4. Qualitative Examples

In order to better show the results, the authors of this paper conduct a visualization analysis of some representative findings of the model from the perspective of qualitative analysis and compare it with other methods. Figure 5 shows that our method is superior to the baseline method.

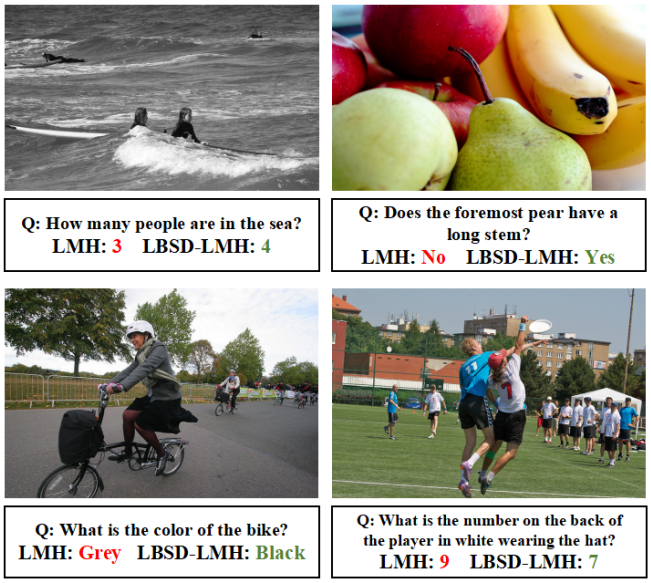

Figure 5. Qualitative examples of VQA-CP v2 (test set). The wrong and right answers are highlighted in red and green.

5. Conclusions

This paper discusses how to reduce the language bias of the VQA model via self-knowledge distillation and proposes a new online learning framework, "language bias-driven self-knowledge distillation (LBSD)", for implicit learning of multi-view visual features. Moreover, in order to help student models learn unbiased visual knowledge, the authors of this paper propose generalization uncertainty to measure the learning results of student models and use KL divergence to reinforce the debiased mutual learning of student models. In this way, the student model can learn unbiased knowledge from

each other through the output of Top-K information entropy. In addition, the paper also discusses the effect of the heterogenous student models on the reduction of language bias. The experiment proves that even the heterogeneous student model can improve the unbiased learning ability through the LBSD method. Extensive experiments and ablation experiments on the VQA-CP v2, VQA-CP v1 and VQA v2 datasets verify the effectiveness of the proposed method. In the future, we will continue to explore how to better define the concept of unbiased knowledge, such as using multimodal knowledge graphs to help the model understand the type of knowledge in the dataset and how to optimize the loss function to enable the model to distinguish biased and unbiased knowledge, so as to reduce the experimental bias against language.

Author Contributions: Conceptualization, D.Y., L.W., Q.W., F.M., K.N.N. and L.X.; methodology, visualization, experiments and writing, D.Y. and L.W.; writing, review and editing, Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China; grant numbers 61831005 and 61971095.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Detailed Proof

Appendix A.1.1. Proof of Generalization Uncertainty (GU)

Proof. Predicting generalization and Calibration. It is found that the distribution over predicted classes and ground truth labels match each other within a set of confidence levels; a measure of disagreement between an ensemble and the ensemble itself boils down to measuring disagreement against the ground truth.

Recall Theorem 1. The authors of this paper can express the expected disagreement rate between two debiased students as an integral over the confidence values.

In order to simplify the expected TOP-K disagreement rate (GU), the authors of this paper will first simplify the expected test error (TE) as follows.

$$\begin{aligned} \mathrm{TE} &\triangleq \mathbb{E}_{n\sim H_A}[p(n(I,Q)\neq a \mid n)] \\ &= \mathbb{E}_{n\sim H_A}\left[\mathbb{E}_{(I,Q,a)\sim D}[\mathbb{1}[n(I,Q)\neq a]]\right] \\ &= \mathbb{E}_{(I,Q,a)\sim D}\left[\mathbb{E}_{H_A}[\mathbb{1}[n(I,Q)\neq a]]\right] \\ &= \mathbb{E}_{(I,Q,a)\sim D}[1-\tilde{n}_a(I,Q)]. \end{aligned} \tag{A1}$$

Deal with integrals, and define $\tilde{n}_k(I,Q)$ as q_k, $(1-\tilde{n}_k(I_i,Q_i))$ as f_k, the authors of this paper can get:

$$\mathrm{TE} = \sum_{k=0}^{K-1} \int_{I_i,Q_i} f_k p(I, Q=I_i, Q_i, a=a_k) d(I_i, Q_i) \tag{A2}$$

$$= \int_{\boldsymbol{q}\in\Delta^K} \sum_{k=0}^{K-1} \int_{I_i,Q_i} f_k p(I, Q=I_i, Q_i, a=a_k, \tilde{n}(I,Q)=\boldsymbol{q}) d(I_i,Q_i) d\boldsymbol{q} \tag{A3}$$

$$= \underbrace{\int_{\boldsymbol{q}\in\Delta^K} \sum_{k=0}^{K-1}}_{\text{swap}} p(a=a_k, \tilde{n}(I,Q)=\boldsymbol{q})(1-q_k) d\boldsymbol{q}. \tag{A4}$$

$$= \sum_{k=0}^{K-1} \int_{\boldsymbol{q} \in \Delta^K} p(a = a_k, \tilde{n}(I,Q) = \boldsymbol{q})(1 - q_k) d\boldsymbol{q} \tag{A5}$$

$$= \sum_{k=0}^{K-1} \int_{q_k} p(a = a_k, \tilde{n}_k(I,Q) = q_k)(1 - q_k) dq_k \tag{A6}$$

$$= \int_{q \in [0,1]} \sum_{k=0}^{K-1} p(a = a_k, \tilde{n}_k(I,Q) = q)(1 - q) dq \tag{A7}$$

Using the calibration in aggregate assumption, the authors of this paper can get:

$$\mathrm{TE} = \int_{q \in [0,1]} q(1-q) \sum_{k=0}^{K-1} p(\tilde{n}_k(I,Q) = q) dq \tag{A8}$$

As defined in Section 3.4.2, the authors of this paper need to prove a direct relationship between GU error and TE. Different from the method proposed by Jiang et al., the GU can provide more soft information than the hard target. Similar to the definition of TE, the authors of this paper can get:

$$\begin{aligned} \mathrm{GUerror} &\triangleq Topk\mathbb{E}_{n,n' \sim H_{\mathcal{A}}} \left[p\left(n(I,Q) \neq n'(I,Q) \mid n, n' \right) \right] \\ &= Topk\mathbb{E}_{n,n' \sim H_{\mathcal{A}}} \left[\mathbb{E}_{(I,Q,a) \sim D} \left[\mathbb{1}\left[n(I,Q) \neq h'(I,Q) \right] \right] \right] \\ &= Topk\mathbb{E}_{(I,Q,a) \sim D} \left[\mathbb{E}_{n,n' \sim H_{\mathcal{A}}} \left[\mathbb{1}\left[n(I,Q) \neq n'(I.Q) \right] \right] \right] \end{aligned} \tag{A9}$$

Similar to the simplified proof of TE, the authors of this paper can get:

$$\begin{aligned} \mathrm{GUerror} &= \sum_{j=0}^{J} 1/p_i \underbrace{\sum_{k=0}^{K-1} \int_{q \in [0,1]}}_{\text{swap}} p(\tilde{n}_k(I,Q) = q) q(1-q) dq \\ &= \sum_{j=0}^{J} 1/p_i \int_{q \in [0,1]} q(1-q) \sum_{k=0}^{K-1} p(\tilde{n}_k(I,Q) = q) dq \propto TE. \end{aligned} \tag{A10}$$

Proof finished. □

Appendix A.1.2. Proof of Debiased Self-Distillation

Proof. Referring to the research work of Allen Zhu, the authors of this paper use the same lottery winning theory and other lemmas to prove it. Refer to Allen Zhu's research for the details of the lemma. The authors of this paper expand the research to VQA with GU.

For the single model. For every $t < T$, according to the noise lower bound and multi-view error claim, the authors of this paper can get:

$$\sum_{t=T_0}^{T} \mathbb{E}_{(I,Q,a) \sim \mathcal{S}_{dm}} \left[1 - \mathrm{logit}_a \left(VQA^{(t)}, I, Q \right) \right] \leq \widetilde{O}\left(\frac{k}{\eta} \right) \tag{A11}$$

$$\sum_{t=T_0}^{T} \mathbb{E}_{(I,Q,a) \sim \mathcal{S}_{ds}} \left(1 - \mathrm{logit}_a \left(VQA^{(t)}, I, Q \right) \right) \leq \widetilde{O}\left(\frac{N}{\eta \rho^{q-1}} \right) \tag{A12}$$

The training objective is:

$$L\left(VQA^{(t)} \right) = \mathbb{E}_{(I,Q,a) \sim S_d} \left[-\log \mathrm{logit}_a \left(VQA^{(t)}, I, Q \right) \right] \tag{A13}$$

For every data:

1. If $\mathrm{logit}_a \left(VQA^{(t)}, I, Q \right) \geq \frac{1}{2}$:

$$-\log \operatorname{logit}_a\left(VQA^{(t)}, I, Q\right) \leq O\left(1-\operatorname{logit}_a\left(VQA^{(t)}, I, Q\right)\right) \tag{A14}$$

2. If $\operatorname{logit}_a\left(VQA^{(t)}, I, Q\right) \leq \frac{1}{2}$:a naive bound,

$$-\log \operatorname{logit}_a\left(VQA^{(t)}, I, Q\right) \in [0, \widetilde{O}(1)] \tag{A15}$$

Therefore, the authors of this paper can get, that when $T \geq \operatorname{poly}(k)/\eta$:

$$\frac{1}{T}\sum_{t=T_0}^{T} \mathbb{E}_{(I,Q,a)\sim S_d}\left[-\log \operatorname{logit}_a\left(VQA^{(t)}, I, Q\right)\right] \leq \frac{1}{\operatorname{poly}(k)} \tag{A16}$$

However, the objective value does not increase monotonically, as the authors of this paper are using gradient descent and using O(1)-Lipscthiz continuous as the objective function, the authors of this paper define $\mathbb{E}_{(I,Q,a)\sim S_d}$ as $\mathbb{E}_d$, and get:

$$\mathbb{E}_d\left(1-\operatorname{logit}_a\left(VQA^{(T)}, I, Q\right)\right) \leq \mathbb{E}_d\left[-\log \operatorname{logit}_a\left(VQA^{(T)}, I, Q\right)\right] \tag{A17}$$

$$\mathbb{E}_{(I,Q,a)\sim S_d}\left[-\log \operatorname{logit}_a\left(VQA^{(T)}, I, Q\right)\right] \leq \frac{1}{\operatorname{poly}(k)} \tag{A18}$$

As a result, during training, the accuracy is perfect. The accuracy of the single-view test is as follows:

$$VQA_a^{(T)}(I, Q) \leq \max_{j\neq y} VQA_j^{(T)}(I, Q) - \frac{1}{\operatorname{poly}\log(k)} \tag{A19}$$

For the single-view D_s, the authors of this paper can get $\left|\mathcal{M}_{VQA}\right| \geq k(1-o(1))$ and the prediction error with a probability of at least $\frac{1}{2}(1-o(1)$, so the authors of this paper can get Theorem 2.

For self-knowledge distillation with the generalization uncertainty model.

The logits function of the model using knowledge distillation can be defined as

$$\operatorname{logit}_i^{\tau}(VQA, I, Q) = \frac{e^{\min\left\{\tau^2 VQA_i(I,Q),1\right\}/\tau}}{\sum_{j\in[k]} e^{\min\left\{\tau^2 VQA_j(I,Q),1\right\}/\tau}} \tag{A20}$$

Similar to the proof of knowledge distillation (from Allen Zhu) and Theorem 1. For a network with $(i, \ell) \in \mathcal{M}$, the authors of this paper can get:

$$S_{i,\ell} \stackrel{\text{def}}{=} \mathbb{E}(I, Q, a) \sim \mathcal{S}_{dm}\left[1_{a=i} \sum_{p\in P_{vqa_{i,\ell}}(I,Q)} z_p^q\right] \tag{A21}$$

$$\mathcal{M} \stackrel{\text{def}}{=} \left\{(i, \ell^*) \in [k]\times[2] \mid \Lambda_{i,\ell^*}^{(0)} \geq \Lambda_{i,3-\ell^*}^{(0)}\left(\frac{S_{i,3-\ell^*}}{S_{i,\ell^*}}\right)^{\frac{1}{q-2}}\left(1+\frac{1}{\log^2(m)}\right)\right\} \tag{A22}$$

Assume that the distribution of $\sum_{p\in\mathcal{P}_{vqa}(I,Q)} z_p^q$ for $vqa \in \{vqa_{a,1}, vqa_{a,2}\}$ are the same. The authors of this paper can get:

$$\mathcal{M}_{VQA1} \stackrel{\text{def}}{=} \left\{(i, \ell^*) \in [k]\times[2] \mid \Lambda_{i,\ell^*}^{(0)} \geq \Lambda_{i,3-\ell^*}^{(0)}\left(1+\frac{2}{\log^2(m)}\right)\right\} \tag{A23}$$

Similar to the single model, for every $(I_i, Q_i, a_i) \in \mathcal{S}_{dm}$, the authors of this paper can get:

$\forall(i,\ell) \in \mathcal{M}_{VQA1}$:
If $vqa_{i,\ell}$,is in $X_{vqa}(I,Q)$:

$$\text{logit}_i^{\tau}(VQA1, I, Q) \geq \frac{1}{s(I,Q)} - k^{-\Omega(\log k)} \tag{A24}$$

If $vqa_{i,\ell}$ is in $X_{vqa}(I,Q)$:

$$\text{logit}_i^{\tau}(VQA1, I, Q) = k^{-\Omega(\log k)} \tag{A25}$$

$\forall i \in [k]$:
If $vqa_{i,\ell}$, is in $X_{vqa}(I,Q)$:

$$\text{logit}_i^{\tau}(VQA1, I, Q) \leq \frac{1}{s'(I,Q)} + k^{-\Omega(\log k)} \tag{A26}$$

If neither $vqa_{i,1}$ or $vqa_{i,2}$ not in $X_{vqa}(I,Q)$:

$$\text{logit}_i^{\tau}(VQA1, I, Q) = k^{-\Omega(\log k)} \tag{A27}$$

Similar to the proof of the ensemble model, at the end of self-knowledge distillation, in multi-view data, the network should provide the same (near-perfect) accuracy. With the generalization uncertainty, the test accuracy has been improved by λ.

Therefore, with $\left|\mathcal{M}_{VQA1}\right| \geq k(1-o(1))$ and $\left|\mathcal{M}_{VQA2}\right| \geq k(1-o(1))$, additionally, they are totally independent random sets, and the authors of this paper can obtain that $\left|\mathcal{M}_{VQA1} \cup \mathcal{M}_{VQA2}\right| \geq \frac{3}{2}k(1-o(1))$. That means the model with self-distillation has an accuracy of $\geq \frac{3}{4}\lambda(1-o(1))$. Therefore, the authors of this paper can get:

$$\Pr_{(I,Q,a)\sim\mathcal{D}}\left[\exists i \in [k]\backslash\{a\} : VQA_a^{(T+T')}(I,Q) < VQA_i^{(T+T')}(I,Q)\right] \leq 0.26\mu/\lambda \tag{A28}$$

Proof finished. □

References

1. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference On Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
2. Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C.L.; Parikh, D.; Batra, D. Vqa: Visual question answering. *Int. J. Comput. Vis.* **2017**, *123*, 4–31. [CrossRef]
3. Teney, D.; Wu, Q.; van den Hengel, A. Visual question answering: A tutorial. *IEEE Signal Process. Mag.* **2017**, *34*, 63–75. [CrossRef]
4. Agrawal, A.; Batra, D.; Parikh, D. Analyzing the behavior of visual question answering models. *arXiv* **2016**, arXiv:1606.07356.
5. Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; Parikh, D. Yin and yang: Balancing and answering binary visual questions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
6. Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
7. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
8. Yuan, D. Language bias in Visual Question Answering: A Survey and Taxonomy. *arXiv* **2021**, arXiv:2111.08531.
9. Brown, G.; Wyatt, J.L.; Tino, P.; Bengio, Y. Managing diversity in regression ensembles. *J. Mach. Learn. Res.* **2005**, *6*, 1621–1650.
10. Mehta, P.; Bukov, M.; Wang, C.H.; Day, A.G.; Richardson, C.; Fisher, C.K.; Schwab, D.J. A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **2019**, *810*, 1–124. [CrossRef]
11. Munson, M.A.; Caruana, R. On feature selection, bias-variance, and bagging. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Geramny, 2009; pp. 144–159.
12. Allen-Zhu, Z.; Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv* **2020**, arXiv:2012.09816.
13. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
14. Yuan, L.; Tay, F.E.; Li, G.; Wang, T.; Feng, J. Revisiting Knowledge Distillation via Label Smoothing Regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3903–3911.

15. Anil, R.; Pereyra, G.; Passos, A.; Ormandi, R.; Dahl, G.E.; Hinton, G.E. Large scale distributed neural network training through online distillation. *arXiv* **2018**, arXiv:1804.03235.
16. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep Mutual Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4320–4328.
17. Lyu, S.; Zhao, Q.; Ma, Y.; Chen, L. Make Baseline Model Stronger: Embedded Knowledge Distillation in Weight-Sharing Based Ensemble Network. 2021. Available online: https://www.bmvc2021-virtualconference.com/assets/papers/0212.pdf (accessed on 17 June 2022).
18. Lukasik, M.; Bhojanapalli, S.; Menon, A.K.; Kumar, S. Teacher's pet: Understanding and mitigating biases in distillation. *arXiv* **2021**, arXiv:2106.10494.
19. Allen-Zhu, Z.; Li, Y. Backward feature correction: How deep learning performs deep learning. *arXiv* **2020**, arXiv:2001.04413.
20. Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; Liu, T.Y. R-drop: Regularized dropout for neural networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10890–10905.
21. Wen, Z.; Li, Y. Toward understanding the feature learning process of self-supervised contrastive learning. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; 2021; pp. 11112–11122.
22. Fuglede, B.; Topsoe, F. Jensen-Shannon divergence and Hilbert space embedding. In Proceedings of the International Symposium on Information Theory, 2004, ISIT 2004, Proceedings, Chicago, IL, USA, 27 June–2 July 2004; p. 31.
23. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
24. Selvaraju, R.R.; Lee, S.; Shen, Y.; Jin, H.; Batra, D.; Parikh, D. Taking a hint: Leveraging explanations to make vision and language models more grounded. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019.
25. Wu, J.; Mooney, R.J. Self-Critical Reasoning for Robust Visual Question Answering. In Proceedings of the Thirty-third Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
26. Li, L.; Gan, Z.; Cheng, Y.; Liu, J. Relation-aware graph attention network for visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 10313–10322.
27. Shrestha, R.; Kafle, K.; Kanan, C. A negative case analysis of visual grounding methods for VQA. *arXiv* **2020**, arXiv:2004.05704.
28. Kv, G.; Mittal, A. Reducing Language Biases in Visual Question Answering with Visually-Grounded Question Encoder. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020.
29. Ramakrishnan, S.; Agrawal, A.; Lee, S. Overcoming language priors in visual question answering with adversarial regularization. *Adv. Neural Inform. Process. Syst.* **2018**, *31*, 1541–1551.
30. Grand, G.; Belinkov, Y. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. *arXiv* **2019**, arXiv:1906.08430.
31. Cadene, R.; Dancette, C.; Ben-younes, H.; Cord, M.; Parikh, D. RUBi: Reducing Unimodal Biases in Visual Question Answering. In Proceedings of the Thirty-third Conference on Neural Information Processing Systems, Vancouver, DC, Canada, 8–14 December 2019.
32. Clark, C.; Yatskar, M.; Zettlemoyer, L. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. *arXiv* **2019**, arXiv:1909.03683.
33. Gat, I.; Schwartz, I.; Schwing, A.G.; Hazan, T. Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3197–3208.
34. Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.S.; Wen, J.R. Counterfactual vqa: A cause-effect look at language bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 12700–12710.
35. Han, X.; Wang, S.; Su, C.; Huang, Q.; Tian, Q. Greedy Gradient Ensemble for Robust Visual Question Answering. In Proceedings of the ICCV 2021, Virtual, 11–17 October 2021.
36. Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; Zhuang, Y. Counterfactual samples synthesizing for robust visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10800–10809.
37. Liang, Z.; Jiang, W.; Hu, H.; Zhu, J. 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), Online, 16–20 November 2020; pp. 3285–3292.
38. Teney, D.; Abbasnedjad, E.; van den Hengel, A. Learning what makes a difference from counterfactual examples and gradient supervision. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 580–599.
39. Guo, Y.; Nie, L.; Cheng, Z.; Tian, Q. Loss-rescaling VQA: Revisiting Language Prior Problem from a Class-imbalance View. *arXiv* **2020**, arXiv:2010.16010.
40. Gokhale, T.; Banerjee, P.; Baral, C.; Yang, Y. MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 878–892.
41. Teney, D.; Abbasnejad, E.; Kafle, K.; Shrestha, R.; Kanan, C.; van den Hengel, A. On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law. In Proceedings of the Advances in Neural Information Processing Systems, Virtual Event, 6–12 December 2020; Volume 33, pp. 407–417.

42. Teney, D.; Abbasnejad, E.; Hengel, A.v.d. Unshuffling Data for Improved Generalization. *arXiv* **2020**, arXiv:2002.11894.
43. Guo, Y.; Nie, L.; Cheng, Z.; Ji, F.; Zhang, J.; Del Bimbo, A. Adavqa: Overcoming language priors with adapted margin cosine loss. *arXiv* **2021**, arXiv:2105.01993.
44. Jiang, J.; Liu, Z.; Liu, Y.; Nan, Z.; Zheng, N. X-GGM: Graph Generative Modeling for Out-of-Distribution Generalization in Visual Question Answering. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 199–208.
45. Rashid, J.; Shah, S.M.A.; Irtaza, A. An efficient topic modeling approach for text mining and information retrieval through K-means clustering. *Mehran Univ. Res. J. Eng. Technol.* **2020**, *39*, 213–222. [CrossRef]
46. Yim, J.; Joo, D.; Bae, J.; Kim, J. A Gift From Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
47. Feng, Z.; Lai, J.; Xie, X. Resolution-Aware Knowledge Distillation for Efficient Inference. *IEEE Trans. Image Process.* **2021**, *30*, 6985–6996. [CrossRef] [PubMed]
48. Rashid, J.; Kim, J.; Hussain, A.; Naseem, U.; Juneja, S. A novel multiple kernel fuzzy topic modeling technique for biomedical data. *BMC Bioinform.* **2022**, *23*, 275. [CrossRef] [PubMed]
49. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
50. Komodakis, N.; Zagoruyko, S. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In Proceedings of the ICLR 2017, Toulon, France, 24–26 April 2017.
51. Tian, Y.; Krishnan, D.; Isola, P. Contrastive Representation Distillation. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
52. Xu, G.; Liu, Z.; Li, X.; Loy, C.C. Knowledge distillation meets self-supervision. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 588–604.
53. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3967–3976.
54. Liu, Y.; Cao, J.; Li, B.; Yuan, C.; Hu, W.; Li, Y.; Duan, Y. Knowledge distillation via instance relationship graph. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7096–7104.
55. Passalis, N.; Tzelepi, M.; Tefas, A. Heterogeneous knowledge distillation using information flow modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2339–2348.
56. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018.
57. Nakkiran, P.; Bansal, Y. Distributional generalization: A new kind of generalization. *arXiv* **2020**, arXiv:2009.08092.
58. Jiang, Y.; Nagarajan, V.; Baek, C.; Kolter, J.Z. Assessing generalization of sgd via disagreement. *arXiv* **2021**, arXiv:2106.13799.
59. Chuang, C.Y.; Torralba, A.; Jegelka, S. Estimating generalization under distribution shifts via domain-invariant representations. *arXiv* **2020**, arXiv:2007.03511.
60. Jiang, Y.; Krishnan, D.; Mobahi, H.; Bengio, S. Predicting the generalization gap in deep networks with margin distributions. *arXiv* **2018**, arXiv:1810.00113.
61. Jiang, Y.; Neyshabur, B.; Mobahi, H.; Krishnan, D.; Bengio, S. Fantastic generalization measures and where to find them. *arXiv* **2019**, arXiv:1912.02178.
62. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Proceedings of the NIPS 2017, Thirty-first Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
63. Dawid, A.P. The well-calibrated Bayesian. *J. Am. Stat. Assoc.* **1982**, *77*, 605–610. [CrossRef]
64. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
65. Gupta, C.; Podkopaev, A.; Ramdas, A. Distribution-free binary classification: Prediction sets, confidence intervals and calibration. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3711–3723.
66. Wu, X.; Gales, M. Should ensemble members be calibrated? *arXiv* **2021**, arXiv:2101.05397.
67. Agrawal, A.; Batra, D.; Parikh, D.; Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In Proceedings of the CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018.
68. Abbasnejad, E.; Teney, D.; Parvaneh, A.; Shi, J.; Hengel, A.v.d. Counterfactual vision and language learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10044–10054.
69. Zhang, L.; Liu, S.; Liu, D.; Zeng, P.; Li, X.; Song, J.; Gao, L. Rich Visual Knowledge-Based Augmentation Network for Visual Question Answering. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4362–4373. [CrossRef]
70. Teney, D.; Kafle, K.; Shrestha, R.; Abbasnejad, E.; Kanan, C.; Hengel, A.v.d. On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law. *arXiv* **2020**, arXiv:2005.09241.