Generative Bias for Visual Question Answering

Jae Won Cho1  Dong-Jin Kim2  Hyeonggon Ryu1  In So Kweon1
1KAIST, South Korea.  2UC Berkeley, CA.
1{chojw,gonhy.ryu,iskweon77}@kaist.ac.kr  2djkim93@berkeley.edu

Abstract

The task of Visual Question Answering (VQA) is known to be plagued by the issue of VQA models exploiting biases within the dataset to make its final prediction. Many previous ensemble based debiasing methods have been proposed where an additional model is purposefully trained to be biased in order to aid in training a robust target model. However, these methods compute the bias for a model from the label statistics of the training data or directly from single modal branches. In contrast, in this work, in order to better learn the bias a target VQA model suffers from, we propose a generative method to train the bias model directly from the target model, called GenB. In particular, GenB employs a generative network to learn the bias through a combination of the adversarial objective and knowledge distillation. We then debias our target model with GenB as a bias model, and show through extensive experiments the effects of our method on various VQA bias datasets including VQA-CP2, VQA-CP1, GQA-OOD, and VQA-CE.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015) is a challenging task that requires a model to correctly understand and predict an answer given a input pair of image and question. Various studies have shown that VQA is prone to biases within the dataset and tend to rely heavily on language biases present in the dataset (Agrawal et al., 2016; Goyal et al., 2017; Zhang et al., 2016), where VQA models tend to predict similar answers only depending on the question regardless of the image. In response to this, recent works have developed various bias reduction techniques and recent methods have exploited ensemble based methods (Cadene et al., 2019; Clark et al., 2019; Han et al., 2021; Ramakrishnan et al., 2018) extensively.

Among ensemble based methods, additional models are introduced to concurrently learn biases that might exist within each modality or dataset. For example, in works such as (Cadene et al., 2019; Han et al., 2021), the Question-Answer model is utilized to determine the language prior biases that exist when a model is asked to give an answer based solely off of the question. This model is then utilized to train a robust “target” model, which is used for inference. The key purpose of an ensemble “bias” model is to capture the biases that are formed with its given inputs (i.e., language prior biases from the Question-Answer model). In doing so, if this model is able to represent the bias well, this bias model can be used to teach the target model to avoid such biased answers. In other words, the better the bias model can learn the biases, the better the target model can avoid such biases.

Existing ensemble based methods either use pre-computed label statistics of training data (GGED (Han et al., 2021), LMH (Clark et al., 2019)), or trained by directly computing the answer distribution from either the question or image (Cadene et al., 2019; Clark et al., 2019; Han et al., 2021; Niu et al., 2021). However, we propose that there is a limit to the bias representation that can be obtained...
from such methods, as the model’s representative capacity is limited due to its input. In addition, pre-computed label statistics represents part of the bias (Han et al., 2021). We show in Fig. 1, that given a question type, the pre-computed label statistics (or known dataset bias) are actually different to the predictions of a model trained with the question or with the image and question. This discrepancy means that there is a part of the bias that we cannot fully model simply with the previous methods. In light of this, we propose to utilize a novel ensemble based bias model that can have stochastic representations and also capture the biases that the target model inhibits.

More specifically, to capture bias by mimicking the target model’s answer representation given the same question input, we model the bias model as a Generative Adversarial Network (GAN) (Goodfellow et al., 2014) to generate stochastic bias representation given a single modality input by introducing an additional random noise vector. As seen through literature, most biases are held within the question (Agrawal et al., 2016), so we use questions as the main bias modality. In addition, we utilize knowledge distillation on top of adversarial training so that the target model learns from harder negative supervision from the bias model. Finally, with our generative bias model, we then use our modified debiasing loss function to train our target model. Our final bias model is able to train the target model that outperforms previous uni-modal and multi-modal ensemble based methods by a large margin. To the best of our knowledge, we are the first to train the bias model by directly leveraging the behavior of the target model using a generative model for the task of VQA.

To show the efficacy of our methods, we perform extensive experiments on commonly used robustness testing VQA datasets: VQA-CP2 and VQA-CP1 (Agrawal et al., 2018) as well as the most recent and challenging VQA setups such as GQA-OOD (Kervadec et al., 2021) and the new evaluation protocol, VQA-CE (Dancette et al., 2021). Our method show the state-of-the-art results on all settings without the use of external human annotations and dataset reshuffling methods.

Our contributions are as follows:

- We propose a novel bias model for ensemble based debiasing for VQA by directly leveraging the target model that we name GenB.

- In order to effectively train GenB, we employ a Generative Adversarial Network and knowledge distillation loss to capture both the dataset distribution bias and the bias from the target model.

- We achieve the state-of-the-art performance on VQA-CP2, VQA-CP1 as well as the more challenging GQA-OOD datasets and VQA-CE using the simple UpDn baseline without extra annotations or dataset reshuffling.

2 Related Work

VQA (Antol et al., 2015; Gurari et al., 2018; Johnson et al., 2017) has been actively studied in recent years with performance reaching close to human performance (Kim et al., 2018; Tan and Bansal, 2019; Su et al., 2019; Lu et al., 2019; Chen et al., 2020b) in the most recent works. Even still, the VQA dataset has been notorious for its reliance on language biases as shown by (Agrawal et al., 2018). To this date, many VQA datasets and evaluation protocols have been derived from the original VQA dataset and have been released to the public as a means to test and understand the biases and robustness of VQA models such as VQA-CP2 and CP1 (Agrawal et al., 2018), GQA-OOD (Kervadec et al., 2021), and VQA-CE (Dancette et al., 2021).

In light of this, many methods have tackled this issue in various ways. One line of work strives to improve visual attention through visual grounding by including additional human annotations or explanations (Selvaraju et al., 2019; Wu and Mooney, 2019), which show limited improvements. Another line of work changes the distribution of the training set by either randomly replacing the images (Teney et al., 2020; Zhu et al., 2021; Wen et al., 2021) or by augmenting the dataset through counterfactual samples (Chen et al., 2020a) or through the use of an external inpainting network (Gokhale et al., 2020). Although powerful, according to (Niu et al., 2021), changing the distribution of the training set does not agree with the philosophy of the creation of this dataset. Therefore, we do not directly compare experimentally with this line of works. Most recently, another line of work has been released where an additional objective of Visual Entailment is added to further boost performance (Si et al., 2021), which we do not compare for fairness.

Excluding those works, the most effective line of work so far has been ensemble based methods (Ab-
basnejad et al., 2020; Cadene et al., 2019; Niu et al., 2021; Han et al., 2021); thus, we place our work in line with these ensemble based methods. Within the ensemble based methods, AReg (Abbasi et al., 2020), RUBi (Cadene et al., 2019), LMH (Clark et al., 2019), GGE (Han et al., 2021) tackle the language prior directly and only use a question-only branch. Unlike these single branch model based methods, GGE (Han et al., 2021) shows the possibility of using the same model as the target model to learn the biases but to a limited extent. On the other hand, CF-VQA (Niu et al., 2021) uses both question and image, but uses the two modalities individually without combining them. Our work is distinct from all previous ensemble based methods as we use a generative network with a noise input to aid the bias model in learning the bias directly from the target model.

3 Methodology
In this section, we explain VQA briefly and describe in detail our bias model, GenB and how we train it and how we debias with it.

3.1 Visual Question Answering Baseline

With an image and question as a pair of inputs, a VQA model learns to correctly predict an answer from the whole answer set $A$. A typical VQA model $F(\cdot, \cdot)$ takes both a visual representation $v \in \mathbb{R}^{n \times d_v}$ (a set of feature vectors computed from a Convolutional Neural Network given an image where $n$ is the number of number of objects in the image and $d_v$ being the vector dimension) and a question representation $q \in \mathbb{R}^{d_q}$ (a single vector computed from a Glove (Pennington et al., 2014) word embedding followed by a Recurrent Neural Network given a question) as input. Then, an attention module followed by a multi-layer perceptron classifier $F : \mathbb{R}^{n \times d_v} \times \mathbb{R}^{d_q} \rightarrow \mathbb{R}^{|A|}$ which generates an answer logit vector $y \in \mathbb{R}^{|A|}$ (i.e., $y = F(V, Q)$). Then, after applying a sigmoid function $\sigma(\cdot)$, our goal is to make an answer probability prediction $\sigma(y) \in [0, 1]^{|A|}$ close to the ground truth answer probability $y_{gt} \in [0, 1]^{|A|}$. In this work, we adopt one of the popular state-of-the-art architectures UpDn (Anderson et al., 2018) widely used in VQA research.

3.2 Ensembling with Bias Models

In this work, our scope is bias mitigation through ensembling bias model similar to previous works (Cadene et al., 2019; Clark et al., 2019; Han et al., 2021). In ensemble based methods, there exist a “bias” model that generates $y_b \in \mathbb{R}^{|A|}$ which we define as $F_b(\cdot, \cdot)$ and a “target” model, defined as $F(\cdot, \cdot)$. Note that, we discard $F_b(\cdot, \cdot)$ during testing and only use $F(\cdot, \cdot)$. As previously mentioned, the goal of the existing bias models is to overfit to the bias as much as possible. Then, given the overfitted bias model, the target model is trained with a debiasing loss function (Cadene et al., 2019; Clark et al., 2019; Han et al., 2021) to improve the robustness of the target model. Ultimately, the target model learns to predict an unbiased answer by avoiding the biased answer from the bias model. The bias model $F_b(\cdot, \cdot)$ can either be the same or different from the original $F(\cdot, \cdot)$ and there could be multiple models as well (Niu et al., 2021). Although previous works try to leverage the bias from the individual modalities, we propose that this limits the ability of the model to represent biases. Hence, in order to try to represent the biases similar to the target model, we set the architecture of $F_b(\cdot, \cdot)$ to be the same as $F(\cdot, \cdot)$ and we use the UpDn (Anderson et al., 2018) model.

3.3 Generative Bias

As mentioned in the Sec. 1, as our goal is to create a bias model that can generate stochastic bias representations, we use a random noise vector in conjunction with a given modality to learn both the dataset bias and the bias that the target model could exhibit. As the question is known to be prone to bias, we keep the question modality and use it as the input to our bias model $F_b(\cdot, \cdot)$. But instead of using the image features, we introduce a random noise vector $z \in \mathbb{R}^{n \times 128}$ in addition to a generator network $G : \mathbb{R}^{n \times 128} \rightarrow \mathbb{R}^{n \times d_v}$ to generate the corresponding input to the bias model $F_b(\cdot, \cdot)$. Formally, given a random Gaussian noise vector $z \sim \mathcal{N}(0, 1)$, a generator network $G(\cdot)$ synthesizes a vector that has the same dimension as the image feature representation, i.e., $\tilde{v} = G(z) \in \mathbb{R}^{n \times d_v}$. Ultimately, our model takes in the question $q$ and $G(z)$ as its input and generates the bias logit $y_b$ in the form $F_b(G(z), q) = y_b$. Note, this can be done on either modality, such as in the form $F_b(G(z), v) = y_b$, but we found this unhelpful. For simplicity, we consider generator and bias model as one network and rewrite $F_b(G(z), q)$ in the form $F_{b,G}(z, q)$ and call our “Generative Bias” method GenB.
3.4 Training the Bias Model

In order for our bias model GenB to learn the biases given the question, we use the traditional VQA loss, the Binary Cross Entropy Loss:

$$L_{GT}(F_{b,G}) = L_{BCE}(\sigma(F_{b,G}(z, q)), y_{gt})$$ \hspace{1cm} (1)

However, unlike existing works, we want the bias model to also capture the biases in the target model. Hence, in order to mimic the bias of the target model as a dynamic distribution of the answer, we propose the adversarial training (Goodfellow et al., 2014) to train our bias model. In particular, we introduce a discriminator to try to distinguish the answers from the target model and the bias model as “real” and “fake” answers, respectively. The discriminator is formulated as $D(F(v, q))$ and $D(F_{b,G}(z, q))$ or rewritten as $D(y)$ and $D(y_b)$. The objective of our generative adversarial network with generator $F_{b,G}(\cdot, \cdot)$ and $D(\cdot)$ can be expressed as follows:

$$\min_{F_{b,G}} \max_D L_{GAN}(F_{b,G}, D), \quad \text{where} \quad L_{GAN}(F_{b,G}, D) = \mathbb{E}_{v,q} \left[ \log \left( D(F(v, q)) \right) \right] + \mathbb{E}_{q,z} \left[ \log \left( 1 - D(F_{b,G}(z, q)) \right) \right] = \mathbb{E}_y \left[ \log \left( D(y) \right) \right] + \mathbb{E}_{y_b} \left[ \log \left( 1 - D(y_b) \right) \right].$$ \hspace{1cm} (2)

The generator ($F_{b,G}$) tries to minimize the objective against an adversarial discriminator ($D$) that tries to maximize it. Through alternative training of $D$ and $F_{b,G}$, the distribution of the answer vector from the bias model ($y_b$) should be close to that from the target model ($y$).

In addition, in order to further aid in the bias model’s ability to capture the intricate biases present in the target model, we add an additional objective that encourages the bias model to directly follow the behavior of the target model with only the $q$ given to it. We empirically find that it is beneficial to include a sample-wise distance based metric such as KL divergence. This phenomenon is well known in the image to image translation task (Isola et al., 2017). Then, the goal of the generator is not only to fool the discriminator but also to try to imitate the answer output of the target model in order to give the target model more challenging supervision in the form of hard negative sample synthesis. We add another objective to our adversarial training for $F_{b,G}(\cdot, \cdot)$ as follows:

$$L_{distill}(F_{b,G}) = \mathbb{E}_{v,q,z} \left[ D_{KL}(F(v, q)\|F_{b,G}(z, q)) \right].$$ \hspace{1cm} (3)

Ultimately, the final training loss for the bias model, or GenB, is as follows:

$$\min_{F_{b,G}} \max_D L_{GenB}(F_{b,G}, D), \quad \text{where} \quad L_{GenB}(F_{b,G}, D) = L_{GAN}(F_{b,G}, D) + \lambda_1 L_{distill}(F_{b,G}) + \lambda_2 L_{GT}(F_{b,G}),$$ \hspace{1cm} (4)

where $\lambda_1$ and $\lambda_2$ are the loss weight hyperparameters.

3.5 Debiasing the Target Model

Given a generated biased answer $y_b$, there are several debiasing loss functions that we can use such as (Cadene et al., 2019; Clark et al., 2019; Han et al., 2021), and we show the effects of each one in Table 5. The GGE (Han et al., 2021) loss is one of the best performing without the use of label distribution. The GGE loss takes the bias predictions/distributions and generates a gradient in the opposite direction to train the target model. With this starting point, we modify this equation with
the ensemble of the biased model in this work as follows:

\[ L_{\text{target}}(F) = L_{\text{BCE}}(y, y_{DL}), \]  

(5)

where the \( i \)-th element of the pseudo-label \( y_{DL} \) is defined as follows:

\[ y_{DL}^i = \min \left( 2 \cdot y_{gt}^i \cdot \sigma(-2 \cdot y_{gt}^i \cdot y_b^i), 1 \right), \]  

(6)

where \( y_{gt}^i \) and \( y_b^i \) are the \( i \)-th element of the ground truth and the output of the biased model, respectively. The key point of difference is that unlike (Han et al., 2021) that suppresses the output of the biased model with the sigmoid function, we use \( y_b \) without using the sigmoid function. In this case, as the value of \( y_{DL} \) can exceed 1, we additionally clip the value so that the value of \( y_{DL} \) is bounded in [0, 1]. We empirically find these simple modifications on the loss function significantly improves the performance. We conjecture the unsuppressed biased output \( y_b \) allows our target model to better consider the intensity of the bias, leading to a more robust target model. In addition, when we train the target model, we do not use the noise inputs as in \( F_{b,G}(z, q) \), rather we use the real images as such \( F_b(v, q) \), and use this output to train our target model. When the bias model is trained, it is trained with a noise vector to hallucinate the possible biases when only given the question, then, to fully utilize the biases that the bias model captures, we give it the real images.

4 Experiments

In this section, we explain our experimental setting and show our quantitative and qualitative experimental results.

4.1 Setup

Dataset and evaluation metric. We conduct our experiments within the VQA datasets that are commonly used for diagnosing bias in VQA models. In particular, we test on the the VQA-CP2 and VQA-CP1 datasets (Agrawal et al., 2018) and GQA-OOD dataset (Kervadec et al., 2021). For evaluation on all datasets, we take the standard VQA evaluation metric (Antol et al., 2015). In addition to this, we also evaluate our method on the VQA-CE (Dancette et al., 2021), which is based on the VQA-v2 dataset, and is a newly introduced VQA evaluation protocol for diagnosing how reliant VQA models are on shortcuts.

Baseline architecture. Similar to recent works, we adopt a popular VQA baseline architecture UpDn (Anderson et al., 2018) as both our ensemble bias model \( F_b \) and our target model \( F \). We list the details of the generator and discriminator in the supplementary material. During training, we train both the bias model and target model together, then we use the target model only for inference.

4.2 Results on VQA-CP2 and VQA-CP1

We compare how our method GenB performs in relation to the recent state-of-the-art methods that focus on bias reduction as shown in Table 1. For VQA-CP2, we first list the baseline architectures in the first section. Then, we compare our method to the methods that modify language modules (DLR (Jing et al., 2020), VGQE (KV and Mittal, 2020)), strengthen visual attention (HINT (Selvaraju et al., 2019), SCR (Wu and Mooney, 2019)), ensemble based methods, (AReg (Ramakrishnan et al., 2018), RUBi (Cadene et al., 2019), LMH (Clark et al., 2019), CF-VQA (Niu et al., 2021), and GGE (Han et al., 2021)) and balance training data by changing the training distribution (CVL (Abbasnejad et al., 2020), RandImg (Teney et al., 2020), SSL (Zhu et al., 2021), CSS (Chen et al., 2020a), Mutant (Gokhale et al., 2020), and D-VQA (Wen et al., 2021)) in the respective sections.

Among the balancing training data methods, while some methods swap the image and questions from the supposed pairs (Teney et al., 2020; Zhu et al., 2021; Wen et al., 2021), counterfactuals based methods generate counterfactual samples masking critical words or objects (Chen et al., 2020a) or by using an external inpainting network to create a new subset of data (Gokhale et al., 2020). Our model (GenB) is in line with the ensemble models listed. Following previous works (Niu et al., 2021), we do not compare to methods that change training distribution as these methods conflict with the original intent of VQA-CP (which is to test whether VQA models rely on prior training data (Agrawal et al., 2018)) and are listed in the bottom of Table 1.

In Table 1, our method achieves the state-of-the-art performance on the VQA-CP2 dataset, surpassing the second best (GGE (Han et al., 2021)) by 1.43%. The performance of our model on all three categories (“Yes/No,” “Num,” and “Other”) are within the top-3 consistently. Our method also performs highly favorably in the “Other” metric.

We also show how our method performs on the
| Method                        | Base | VQA-CP2 test | VQA-CP1 test |
|------------------------------|------|--------------|--------------|
|                              |      | All          | Yes/No       | Num          | Other         |
|                              |      | All          | Yes/No       | Num          |
| SAN (Yang et al., 2016)      | -    | 24.96        | 38.35        | 11.14        | 21.74         |
| GVQA (Agrawal et al., 2018)  | -    | 31.30        | 57.99        | 13.68        | 22.14         |
| S-MRL (Cadene et al., 2019)  | -    | 38.46        | 42.85        | 12.81        | 43.20         |
| UpDn (Anderson et al., 2018) | -    | 39.94        | 42.46        | 11.93        | 45.09         |
|                              |      | 32.50        | 36.86        | 12.47        | 36.22         |
|                              |      | 36.38        | 42.72        | 12.59        | 40.35         |
|                              |      | 36.38        | 42.72        | 12.44        | 40.35         |
| **Methods based on modifying language modules** | |  |  |  |
| DLR (Jing et al., 2020)      | UpDn | 48.87        | 70.99        | 18.72        | 45.57         |
| VQGE (KV and Mittal, 2020)   | UpDn | 48.75        | –            | –            | –             |
| VQGE (KV and Mittal, 2020)   | S-MRL | 50.11        | 66.35        | 27.08        | 46.77         |
|                              |      | 50.90        | 80.83        | 13.84        | 36.02         |
| **Methods based on strengthening visual attention** | |  |  |  |
| HINT (Selvaraju et al., 2019) | UpDn | 46.73        | 67.27        | 10.61        | 45.88         |
| SCR (Wu and Mooney, 2019)    | UpDn | 49.45        | 65.49        | 27.08        | 46.77         |
|                              |      | 50.90        | 80.83        | 13.84        | 36.02         |
| **Methods based ensemble models** | |  |  |  |
| AReg (Ramakrishnan et al., 2018) | UpDn | 41.17        | 65.49        | 15.48        | 45.54         |
| RUBi (Cadene et al., 2019)   | UpDn | 44.23        | 67.05        | 17.48        | 46.77         |
| LMH (Clark et al., 2019)     | UpDn | 52.45        | 69.81        | 44.46        | 45.54         |
| CF-VQA(SUM) (Niu et al., 2021)| UpDn | 53.55        | 69.41        | 43.15        | 46.77         |
| CF-VQA(SUM) (Niu et al., 2021)| S-MRL | 55.05        | 70.61        | 21.50        | 46.77         |
| GGE (Han et al., 2021)       | UpDn | 57.32        | 87.04        | 27.75        | 49.59         |
| CFB (Ours)                   | UpDn | 58.75        | 88.91        | 35.37        | 49.37         |
|                              |      | 62.74        | 86.18        | 43.85        | 47.03         |
| **Methods based on balancing training data** | |  |  |  |
| CVL (Abbassinejad et al., 2020)| UpDn | 42.12        | 45.72        | 12.45        | 48.34         |
| Randling (Teney et al., 2020)| UpDn | 55.37        | 83.89        | 41.60        | 44.20         |
| SSL (Zhu et al., 2021)       | UpDn | 57.59        | 86.53        | 29.87        | 50.03         |
| CSS (Chen et al., 2020a)     | UpDn | 58.95        | 84.37        | 49.42        | 48.21         |
| MUTANT (Gokhale et al., 2020)| UpDn | 61.72        | 88.90        | 49.68        | 50.78         |
| D-VQA (Wen et al., 2021)     | UpDn | 61.91        | 88.93        | 52.32        | 50.39         |

Table 1: Experimental results of our method on the VQA-CP2 test set and VQA-CP1 test set. Best and second best results are styled in this manner within the column. We do not directly compare to methods that change the distribution of the training data as it does not go with the philosophy of VQA-CP (Niu et al., 2021). Among the compared baselines, our method GenB shows the best performance by a noticeable margin.

VQA-CP1 dataset, which is a subset of the VQA-CP2 dataset. Note that not all of the baselines are listed as we only list the scores that are made available in the respective papers. Our method also shows the state-of-the-art results on this dataset with a significant performance improvement over the second best among the methods compared, CF-VQA(SUM) (Niu et al., 2021). Even when compared to the available method CSS (Chen and Dolan, 2011) that we do not compare as it is considered unfair according to (Niu et al., 2021), our method shows a significant gain in performance. Compared to the second best method, CF-VQA(Sum) (Niu et al., 2021) on UpDn, our method improves the overall performance by 5.35% while also having the best performance in both “Num” and “Other” category, by 3.28% and 2.41% performance improvements, respectively.

4.3 Results on GQA-OOD

In light of recent revelations, we further evaluate our method on the new VQA debiasing dataset, GQA-OOD (Kervadec et al., 2021) and list our results in Table 2. We compare our method to available recent state-of-the-art methods RUBi (Cadene et al., 2019), LMH (Clark et al., 2019), and CSS (Chen et al., 2020a). Our method shows the best performance in all the metrics compared to the state-of-the-art methods by a significant margin. Even when compared to methods that show similar performance to GenB in VQA-CP2 like CSS, GenB significantly outperforms it in GQA-OOD by 5.19% in overall. Interestingly, although all of the listed previous methods outperform the base model UpDn in other datasets, they show a performance degradation on GQA-OOD. Unlike these methods, our method GenB is able to have an increase in performance, showing the robustness of GenB.

4.4 Results on VQA-CE

We also evaluate our method on the newly introduced evaluation protocol that measures how much a VQA model depends on shortcuts called VQA-CounterExamples (VQA-CE) (Dancette et al., 2021) in Table 3. This dataset is based on the VQA-v2 dataset and lists Overall, Easy, and Counter, which are the total score on VQA-v2 validation set, the subset of samples where the shortcuts of im-
Table 2: Experimental results on the GQA-OOD dataset. Our method shows the best performance in all the metrics compared to the state-of-the-art methods by a significant margin. The results show that our method robust on GQA-OOD as well.

Table 3: Evaluation on the VQA-CE protocol. Ours shows the best performance in counterexamples (listed as Counter) which is the main scope of the VQA-CE.

Table 4: We ablate the different losses we use to train the GenB bias model. All inferences scores are based on the target model except the first row. BCE loss is Eq. (1), which is the ground truth VQA loss. DSC refers to the discriminator loss Eq. (4) and Distill refers to the KL Divergence loss Eq. (3). Although the DSC and Distill losses independently do not show large improvement, our final model with all losses show a large margin of improvement.

Table 5: Ablation of ensemble debiasing loss functions and our debiasing loss function Eq. (5). Note that our loss improves the score of using the vanilla UpDn model when compared to GGE by a large margin. Note that GenB works best with our loss and shows a large performance improvement from GGE + GenB.

4.5 Ablation Studies

Our method (GenB) includes several different components as shown from Sec. 3.3 to Sec. 3.5. To understand the effects of each component, we run an ablation study on the VQA-CP2 dataset. For all experiments, the results are of the target model and as the purpose of the bias model is to capture bias instead of correctly predicting the answers, we do not consider the predictions of the bias model. For all our ablation tables, we also add the UpDn baseline in the first row, the model in which our target model and bias model is based for comparison.

Bias model training analysis. We ablate on the different losses we use to train the bias model from the target model and list our findings in Table 4 with BCE denoting the ground truth VQA loss Eq. (1), DSC denoting the discriminator loss Eq. (4), and Distill denoting the KL Divergence distillation loss Eq. (3). GenB trained on the BCE loss is already significantly helpful in debiasing the target model. Adding the DSC and Distill losses show similar performance to the GenB trained with BCE loss in a standalone manner. As the bias model and target model are trained concurrently, without a ground truth anchor, we conjecture that the bias model struggles to learn any meaningful bias. Moreover, combining all the proposed loss functions shows the best overall performance among all the baselines. Note that BCE + DSC shows the best Yes/No score while our final model shows the best score in Num score. We conjecture that the bias model trained with our adversarial training framework is good at modeling the bias for simple answers like Yes/No while the bias with more complex answers like counting is hard to learn only with the adversarial training. We conclude that combining our knowledge distillation loss to the adversarial training is essential to learn the bias with complex answers including the Num category.

Debiasing loss function analysis. We list in Sec. 3.5 the different losses we can use for debiasing and we show our results in Table 5. We also list our new debiasing loss function Eq. (5)
and how it performs when a vanilla UpDn model is used instead of the GenB.

We analyze how GGE’s loss function fairs when the vanilla UpDn model is used as a bias model. In the second and the third row, although the GGE loss improves the score from the baseline by introducing a bias model, our loss exceeds the performance by 5.07%. In the last two rows, we also show the debiasing losses with the full GenB model with all of the bias model training losses and show 8.81% performance gap. Given the GGE loss, the performance increase from the bias model is only 2.11% whereas our debiasing loss function is able to boost the score from the vanilla UpDn to GenB by 5.85%. We conjecture that as our loss function accounts for the amount of bias a model captures, and the generative bias model is able to stochastically model the bias, our loss function captures the bias in a much more dynamic and helpful manner.

4.6 Qualitative Results

We visualize the attention scores and predictions of our target and bias models in Fig. 3. We run the bias model three times with different noise to show the varying attention scores and biased predictions. We also input the corresponding image with the question to see where the model would attend, and we find that the attention is once again random as the image is previously unseen. The bias model’s predictions change each time as a different noise is given as expected. Even though the question type is the same (e.g., what color), the biased model’s predictions are noticeably different. Moreover, we find the target model is able to correctly attend to the salient region while predicting the correct answer by virtue of the bias model.

5 Conclusion

In this paper, we started with this intuition “the better the bias model, the better we can debias the target model. Then how can we best model the bias?” In response, we present simple, effective, and novel generative bias model that we call GenB. We use this generative model to learn the bias that may be inhibited by the target model with the aid of generative networks, adversarial training, and knowledge distillation. In addition, in conjunction with our modified loss function, our novel bias model is able to debias our target model, and our target model achieves state-of-the-art performance on various bias diagnosing datasets including VQA-CP2, VQA-CP1, GQA-OOD, and VQA-CE among methods that keep in line with the original intent of the task, and we believe that this work can be extended to other multi-modal and unimodal research in understanding and mitigating bias.
References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In CVPR.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In EMNLP.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don’t just assume; look and answer: Overcoming priors for visual question answering. In CVPR.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In CVPR.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQa: Visual question answering. In ICCV.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases in visual question answering. In NeurIPS.

David L. Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. ACL.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020a. Counterfactual samples synthesizing for robust visual question answering. In CVPR.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In ECCV.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases. In EMNLP.

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1574–1583.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In EMNLP.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In NeurIPS.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In CVPR.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In CVPR.

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

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In CVPR.

Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming language priors in VQA via decomposed linguistic representations. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 11181–11188. AAAI Press.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. Roses are red, violets are blue... but should vqa expect them to? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2776–2785.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In NeurIPS.

Gouthaman KV and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, pages 18–34. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In CVPR.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In EMNLP.
Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In NeurIPS.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In ICCV.

Qingyi Si, Zheng Lin, Ming yu Zheng, Peng Fu, and Weiping Wang. 2021. Check it again: Progressive visual question answering via visual entailment. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4101–4110.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vi-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. 2020. On the value of out-of-distribution testing: An example of goodhart’s law. arXiv preprint arXiv:2005.09241.

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

Jialin Wu and Raymond J Mooney. 2019. Self-critical reasoning for robust visual question answering. In NeurIPS.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In CVPR.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5014–5022.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2021. Overcoming language priors with self-supervised learning for visual question answering. IJCAI.