Generative Bias for Robust Visual Question Answering*

Jae Won Cho[1], Dong-Jin Kim[2], Hyeonggon Ryu[1], and In So Kweon[1]

[1] KAIST, Daejeon, South Korea
[2] Hanyang University, Seoul, South Korea
chojw@kaist.ac.kr & djdkim@hanyang.ac.kr & gonhy.ryu@kaist.ac.kr & iskweon77@kaist.ac.kr

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. Various 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 simply from the label statistics of the training data or from single modal branches. 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 in the target model 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 and VQA-CP1.

Keywords: Visual Question Answering · Vision & language · Robustness.

1 Introduction

Visual Question Answering (VQA) [4] 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 [1,8,17], 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 debiasing methods [5,6,9,15] 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 [5,9], the Question-Answer (QA) model is utilized to

* This paper is a short version of CVPR'23 Submission and 28th Samsung HumanTech Bronze award winner

determine the language prior biases that exist when a model is asked to give an answer based solely off of the question. This QA 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 QA 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 (GGE-D [9] and LMH [6]), or single modal branches that compute the answer from either the question or image [5,6,9,12]. However, we conjecture 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 only part of the bias [9]. As shown in Fig. 1, given a question type, the pre-computed label statistics (or known dataset bias) are noticeably different to the predictions of a model trained with the question or with the image and question. This discrepancy signifies that there is a part of the bias that we cannot fully model simply with the previous methods. Therefore, we propose a novel stochastic bias model that learns the bias *directly from the target model*.

More specifically, to directly learn the bias distribution of the *target model*, we model the bias model as a Generative Adversarial Network (GAN) [7] to stochastically mimic the target model's answer distribution given the same question input by introducing a random noise vector. As seen through literature, most biases are held within the question [1], so we use questions as the main bias modality. In addition, we utilize knowledge distillation [10] on top of adversarial training to force the bias model to be as close as possible to the target model, 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 debiasing 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.

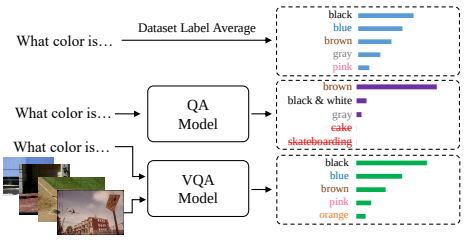

Fig. 1. Given a Question Type ("What color is..."), we show all of the averaged answers within the training dataset. The answer computed from the entire training dataset is the known dataset bias as in [6,9]. We see that the averaged model predictions of the Question-Answer Model and Visual-Question-Answer Model are significantly different.

To show the efficacy and robustness of our method, we perform extensive experiments on commonly used robustness testing VQA datasets and various

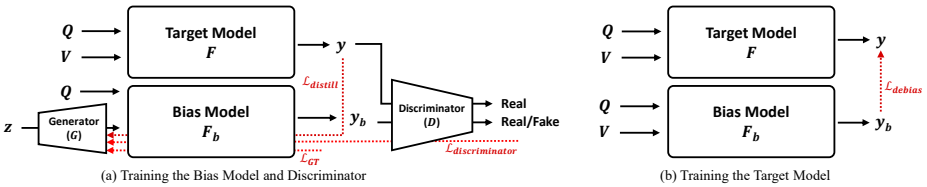

Fig. 2. (a) shows how we train our Bias Model and Discriminator. The Bias Model is trained with 3 different losses including the ground truth BCE (Eq. (1)), knowledge distillation (Eq. (3)), and adversarial Eq. (2)) losses. (b) shows how our Target Model is trained with the Bias model with debiasing loss functions (refer existing works). Note, steps (a) and (b) happen concurrently. Note that we only use the Target Model during inference.

different VQA architectures. 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: (1) We propose a novel bias model for ensemble based debiasing for VQA by directly leveraging the target model that we name *GenB*. (2) 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. (3) We achieve state-of-the-art performance on VQA-CP2 and VQA-CP1 using the simple UpDn baseline without extra annotations or dataset reshuffling.

2 Methodology

In this section, we explain VQA briefly and describe in detail our method GenB, how we train and debias with it, and a short analogous discussion.

2.1 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 $\mathcal{A}$. A typical VQA model $F(\cdot,\cdot)$ takes both a visual representation $\mathbf{v} \in \mathbb{R}^{n \times d_v}$ (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 $\mathbf{q} \in \mathbb{R}^{d_q}$ (a single vector computed from a Glove [14] word embedding followed by a Recurrent Neural Network given a question) as input. Then, an attention module followed by a multi-layer perceptron classifier $F : \mathbb{R}^{n \times d_v} \times \mathbb{R}^{d_q} \rightarrow \mathbb{R}^{|\mathcal{A}|}$ which generates an answer logit vector $\mathbf{y} \in \mathbb{R}^{|\mathcal{A}|}$ (*i.e.*, $\mathbf{y} = F(V, Q)$). Then, after applying a sigmoid function $\sigma(\cdot)$, our goal is to make an answer probability prediction $\sigma(\mathbf{y}) \in [0, 1]^{|\mathcal{A}|}$ close to the ground truth answer probability $\mathbf{y}_{gt} \in [0, 1]^{|\mathcal{A}|}$. In this work, we adopt one of the popular state-of-the-art architectures UpDn [3] widely used in VQA research.

2.2 Ensembling with Bias Models

In this work, our scope is bias mitigation through ensembling bias model similar to previous works [5,6,9]. In ensemble based methods, there exist a "bias" model that generates $\mathbf{y}_b \in \mathbb{R}^{|\mathcal{A}|}$ 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 [5,6,9] 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 [12]. Although previous works try to leverage the bias from the individual modalities [5,6,9,12], we propose that this limits the ability of the model to represent biases. Hence, in order 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 [3] model.

2.3 Generative Bias

As mentioned in the Sec. 1, as our goal is to train 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 $\mathbf{z} \in \mathbb{R}^{n\times 128}$ 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 $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{1})$, a generator network $G(\cdot)$ synthesizes a vector that has the same dimension as the image feature representation, *i.e.*, $\hat{\mathbf{v}} = G(\mathbf{z}) \in \mathbb{R}^{n\times d_v}$. Ultimately, our model takes in the question $\mathbf{q}$ and $G(\mathbf{z})$ as its input and generates the bias logit $\mathbf{y_b}$ in the form $F_b(G(\mathbf{z}),\mathbf{q}) = \mathbf{y_b}$. Note, this can be done on another modality, (*i.e.*, $F_b(G(\mathbf{z}),\mathbf{v}) = \mathbf{y_b}$), but we found this is unhelpful. For simplicity, we consider generator and bias model as one network and rewrite $F_b(G(\mathbf{z}),\mathbf{q})$ in the form $F_{b,G}(\mathbf{z},\mathbf{q})$ and call our "Generative Bias" method **GenB**.

2.4 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:

$$\mathcal{L}_{GT}(F_{b,G}) = \mathcal{L}_{BCE}(\sigma(F_{b,G}(\mathbf{z},\mathbf{q})),\mathbf{y}_{gt}). \tag{1}$$

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 random distribution of the answer, we propose adversarial training [7] to train our bias model. In particular, we introduce a discriminator that tries to distinguish the answers from the target model and the bias model as "real" and "fake" answers, respectively. The discriminator is formulated as $D(F(\mathbf{v},\mathbf{q}))$ and $D(F_{b,G}(\mathbf{z},\mathbf{q}))$ or rewritten as $D(\mathbf{y})$ and $D(\mathbf{y}_b)$. The objective of our generative adversarial network with generator $F_{b,G}(\cdot,\cdot)$ and $D(\cdot)$ can be expressed as:

$$\begin{aligned}
&\min_{F_{b,G}}\max_{D}\mathcal{L}_{GAN}(F_{b,G},D), \text{where}\\
&\mathcal{L}_{GAN}(F_{b,G},D)=\mathop{\mathbb{E}}_{\mathbf{v},\mathbf{q}}\Big[\log\Big(D\big(F(\mathbf{v},\mathbf{q})\big)\Big)\Big]+\mathop{\mathbb{E}}_{\mathbf{q},\mathbf{z}}\Big[\log\Big(1-D\big(F_{b,G}(\mathbf{z},\mathbf{q})\big)\Big)\Big]\\
=&\mathbb{E}_{\mathbf{y}}\Big[\log\big(D(\mathbf{y})\big)\Big]+\mathbb{E}_{\mathbf{y}_b}\Big[\log\big(1-D(\mathbf{y}_b)\big)\Big].
\end{aligned}\tag{2}$$

The generator ($F_{b,G}$) tries to minimize the objective ($\mathcal{L}_{GAN}$) 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 ($\mathbf{y}_b$) should be close to that from the target model ($\mathbf{y}$).

In addition, to further aid in the bias model's ability to capture the intricate biases present in the target model, we add an additional knowledge distillation objective [10] that encourages the bias model to directly follow the behavior of the target model with only the $\mathbf{q}$ given to it. We empirically find that it is beneficial to include a sample-wise distance based metric such as KL divergence. This method is similar to the approaches in the image to image translation task [11]. 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:

$$\mathcal{L}_{distill}(F_{b,G})=\mathop{\mathbb{E}}_{\mathbf{v},\mathbf{q},\mathbf{z}}\Big[D_{KL}\big(F(\mathbf{v},\mathbf{q})\|F_{b,G}(\mathbf{z},\mathbf{q})\big)\Big].\tag{3}$$

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

$$\begin{aligned}
&\min_{F_{b,G}}\max_{D}\mathcal{L}_{GenB}(F_{b,G},D), \text{where}\\
&\mathcal{L}_{GenB}(F_{b,G},D)=\mathcal{L}_{GAN}(F_{b,G},D)+\lambda_1\mathcal{L}_{distill}(F_{b,G})+\lambda_2\mathcal{L}_{GT}(F_{b,G}),
\end{aligned}\tag{4}$$

where λ_1 and λ_2 are the loss weight hyper-parameters.

2.5 Debiasing the Target Model

Given a generated biased answer $\mathbf{y}_b$, there are several debiasing loss functions that we can use such as [5,6,9]. The GGE [9] loss is one of the best performing losses without the use of label distribution. The GGE loss takes the bias predictions/distributions and generates a gradient in the opposite direction to train

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. Among the compared baselines, our method GenB shows the best performance by a noticeable margin.

Method	Base	VQA-CP2 test				VQA-CP1 test			
		All	Yes/No	Num	Other	All	Yes/No	Num	Other
SAN [16]	-	24.96	38.35	11.14	21.74	32.50	36.86	12.47	36.22
GVQA [2]	-	31.30	57.99	13.68	22.14	39.23	64.72	11.87	24.86
S-MRL [5]	-	38.46	42.85	12.81	43.20	36.38	42.72	12.59	40.35
UpDn [3]	-	39.94	42.46	11.93	45.09	36.38	42.72	42.14	40.35
Methods based on ensemble models									
AReg [15]	UpDn	41.17	65.49	15.48	35.48	43.43	74.16	12.44	25.32
RUBi [5]	UpDn	44.23	67.05	17.48	39.61	50.90	80.83	13.84	36.02
LMH [6]	UpDn	52.45	69.81	**44.46**	45.54	55.27	76.47	**26.66**	**45.68**
CF-VQA(SUM) [12]	UpDn	53.55	**91.15**	13.03	44.97	57.03	**89.02**	17.08	41.27
CF-VQA(SUM) [12]	S-MRL	55.05	90.61	21.50	45.61	**57.39**	**88.46**	14.80	43.61
CF-VQA(SUM) [12] + IntroD [13]	S-MRL	55.17	**90.79**	17.92	46.73	–	–	–	–
GGE [9]	UpDn	**57.32**	87.04	27.75	**49.59**	–	–	–	–
GenB (Ours)	UpDn	**59.15**	88.03	**40.05**	**49.25**	**62.74**	86.18	**43.85**	**47.03**

the target model. With this starting point, we modify this equation with the ensemble of the biased model in this work as follows:

$$\mathcal{L}_{target}(F) = \mathcal{L}_{BCE}(\mathbf{y}, \mathbf{y}_{DL}), \tag{5}$$

where the i-th element of the pseudo-label $\mathbf{y}_{DL}$ is defined as follows:

$$\mathbf{y}^i_{DL} = \min\left(1\ ,\ 2 \cdot \mathbf{y}^i_{gt} \cdot \sigma(-2 \cdot \mathbf{y}^i_{gt} \cdot \mathbf{y}^i_b)\right), \tag{6}$$

where $\mathbf{y}^i_{gt}$ and $\mathbf{y}^i_b$ 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 [9] that suppresses the output of the biased model with the sigmoid function, we use $\mathbf{y}_b$ without using the sigmoid function. In this case, as the value of $\mathbf{y}_{DL}$ can exceed 1, we additionally clip the value so that the value of $\mathbf{y}_{DL}$ 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 $\mathbf{y}_b$ 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 $F_{b,G}(\mathbf{z}, \mathbf{q})$, rather we use the real images as such $F_b(\mathbf{v}, \mathbf{q})$, 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.

3 Experiments

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 [2]. For evaluation on all datasets, we take the standard VQA evaluation metric [4].

Baseline architecture. We adopt a popular VQA baseline architecture UpDn [3] as both our ensemble bias model F_b and our target model F. During training,

Table 2. Loss ablation for GenB. All inferences scores are based on the target model except the first row. Although the DSC and Distill losses independently do not show large improvement, our final model with all losses show a large margin of improvement.

Training Loss	Bias Model	VQA-CP2 test			
		All	Yes/No	Num	Other
BCE	UpDn	39.94	42.46	11.93	45.09
BCE	GenB	56.98	88.82	19.39	**49.86**
BCE + DSC	GenB	56.54	**89.06**	21.29	49.79
BCE + Distill	GenB	57.06	88.91	23.24	49.65
BCE + DSC + Distill	GenB	**59.15**	88.03	**40.05**	49.25

we train both the bias model and target model together, then we use the target model only for inference.

Results on VQA-CP2 and VQA-CP1. We compare GenB in relation to the recent state-of-the-art ensembling methods that focus on bias reduction as shown in Table 1. For VQA-CP2, we first list the *baseline architectures* and then compare only with in this paper due to lack of space. The *ensemble based methods* listed are, (AReg [15], RUBi [5], LMH [6], CF-VQA [12], and GGE [9]).

In Table 1, our method achieves state-of-the-art performance on VQA-CP2, surpassing the second best (GGE [9]) by 1.83%. The performance of our model on all three categories ("Yes/No," "Num," "Other") are within the top-3 consistently for the same backbone architecture. Our method also performs highly favorably in the "Other" metric. We also show how our method performs on the VQA-CP1 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) [12] and our method improves the overall performance by 5.60% while also having the best performance in both "Num" and "Other" category, by 3.28% and 2.41% performance improvements, respectively.

3.1 Ablation Studies

Our method (GenB) includes several different components as shown from Sec. 2.3 to Sec. 2.5. 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. To further understand whether GenB can be applied to other networks, we further include an ablation study of GenB on other VQA architectures.

4 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 both the distribution and 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 and we believe that this work can be extended to other multi-modal and uni-modal research in understanding and mitigating bias.

References

1. Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. In: EMNLP (2016) 1, 2
2. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: CVPR (2018) 6
3. 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: CVPR (2018) 3, 4, 6
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: ICCV (2015) 1, 6
5. Cadene, R., Dancette, C., Ben-younes, H., Cord, M., Parikh, D.: Rubi: Reducing unimodal biases in visual question answering. In: NeurIPS (2019) 1, 2, 4, 5, 6, 7
6. Clark, C., Yatskar, M., Zettlemoyer, L.: Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In: EMNLP (2019) 1, 2, 4, 5, 6, 7
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014) 2, 5
8. 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: CVPR (2017) 1
9. Han, X., Wang, S., Su, C., Huang, Q., Tian, Q.: Greedy gradient ensemble for robust visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1584–1593 (2021) 1, 2, 4, 5, 6, 7
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 2, 5
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017) 5
12. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: CVPR (2021) 2, 4, 6, 7
13. Niu, Y., Zhang, H.: Introspective distillation for robust question answering. Advances in Neural Information Processing Systems **34**, 16292–16304 (2021) 6
14. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: EMNLP (2014) 3
15. Ramakrishnan, S., Agrawal, A., Lee, S.: Overcoming language priors in visual question answering with adversarial regularization. In: NeurIPS (2018) 1, 6, 7
16. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR (2016) 6
17. 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. pp. 5014–5022 (2016) 1