Generative Bias for Robust Visual Question Answering

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

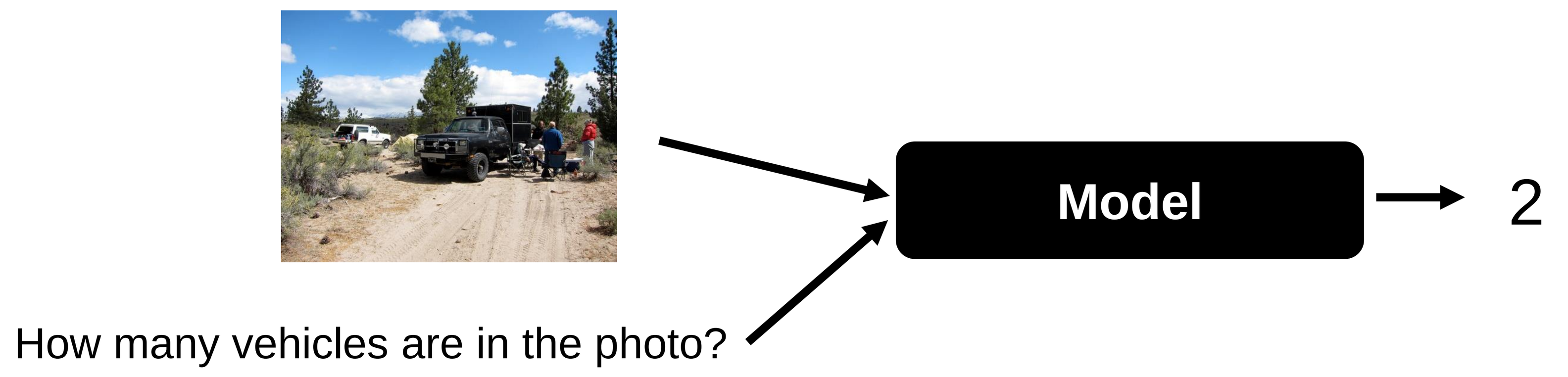
[1] KAIST [2] Hanyang University

https://github.com/chojw/genb

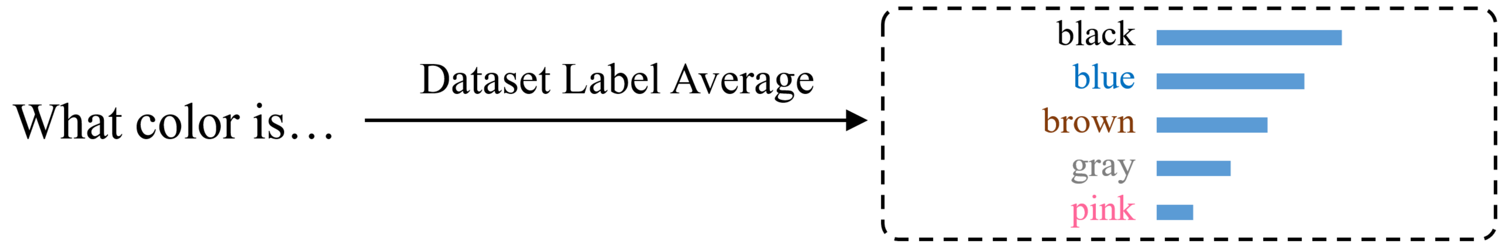
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

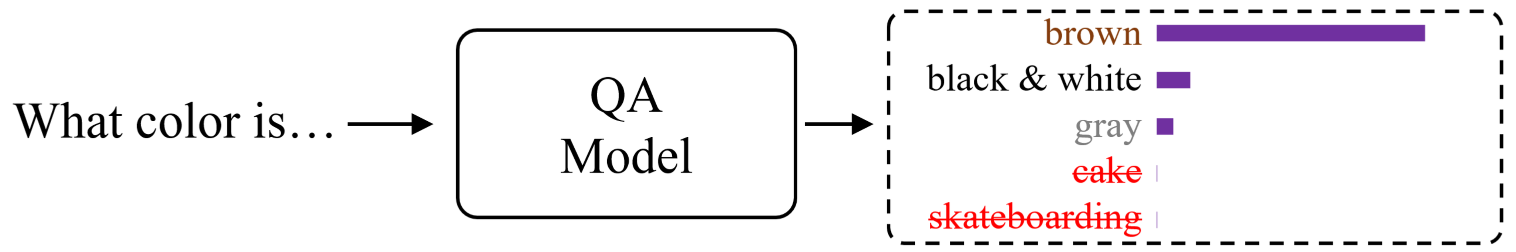
Task: Visual Question Answering (VQA)



- VQA is known to have **bias issues** where models rely on **language priors**
- **Ensemble-based** method use a **biased** model to *debias* a **target** ***"robust"*** model
- Previous ensemble-based methods primarily utilize **two label statistics**

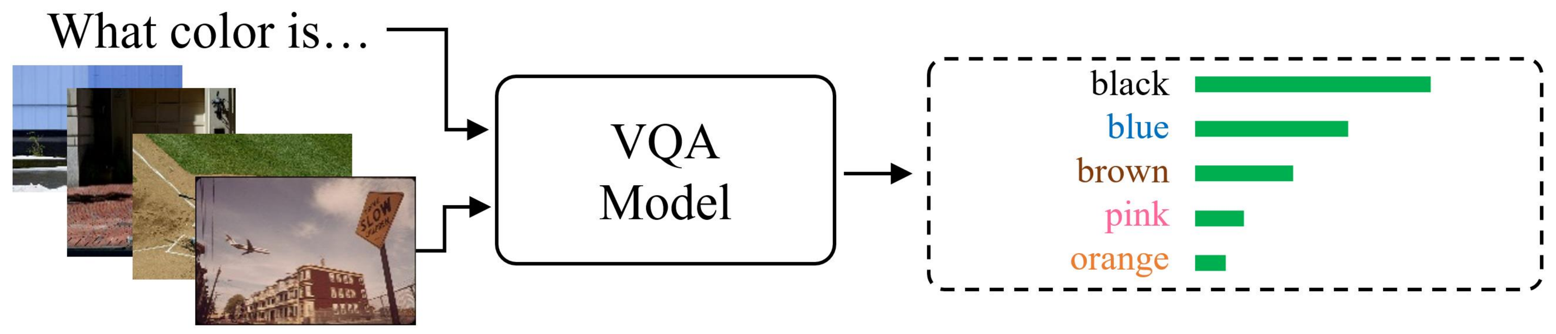


Dataset label average statistics



Single-modal model's average output statistics

- The bias experienced by an *actual VQA model*

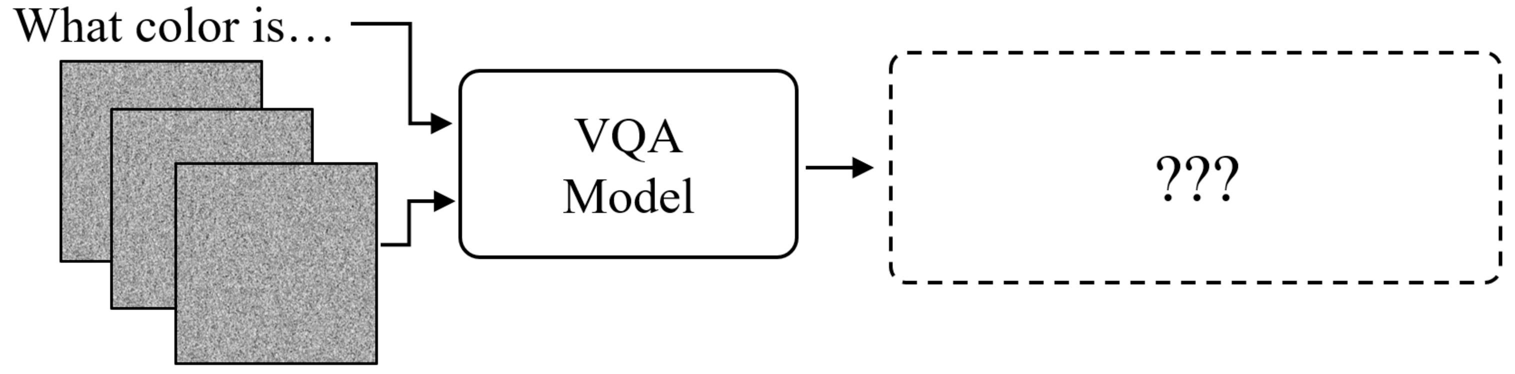


Motivation:

The better we can capture the bias, the better we can debias

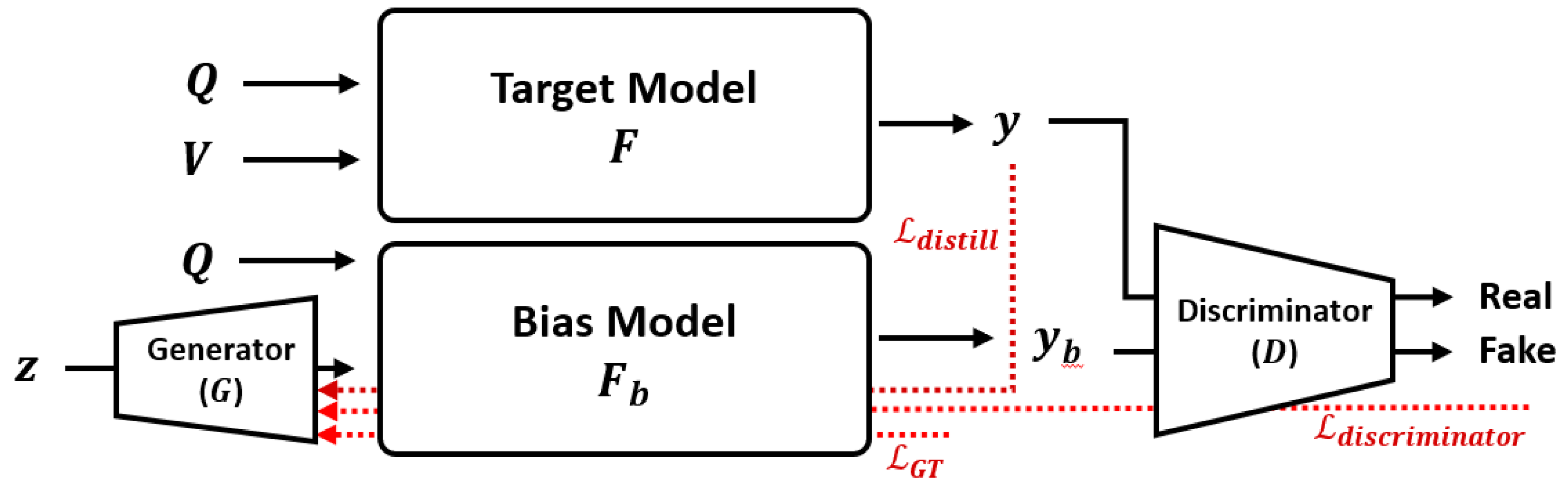
Method

- Bias representations are limited by the *static inputs* of images or questions
- We replace image input with a **generator model** to *capture* the bias



Training the Bias model

The bias model captures
- the **distribution bias** through *GT Loss*
- the **model bias** through *discriminator loss* + *distillation loss*



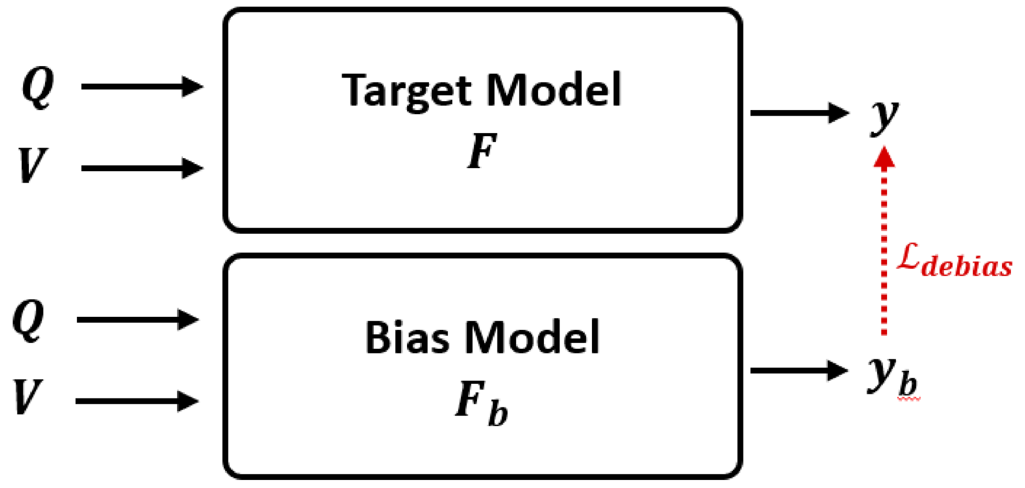
Training the Target Model

- Gradient based debiasing loss for target model

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

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

- Raw unbounded output + clamping allows the loss to take into consideration the **intensity** of bias



Experiments

Comparison with state-of-the-art

Method	Base	VQA-CP2 test			
		All	Yes/No	Num	Other
SAN [50]	-	24.96	38.35	11.14	21.74
GVQA [3]	-	31.30	57.99	13.68	22.14
S-MRL [7]	-	38.46	42.85	12.81	43.20
UpDn [4]	-	39.94	42.46	11.93	45.09
Methods based on modifying language modules					
DLR [24]	UpDn	48.87	70.99	18.72	45.57
VGQE [35]	UpDn	48.75	–	–	–
VGQE [35]	S-MRL	50.11	66.35	27.08	46.77
Methods based on strengthening visual attention					
HINT [42]	UpDn	46.73	67.27	10.61	45.88
SCR [48]	UpDn	49.45	72.36	10.93	48.02
Methods based on ensemble models					
AReg [41]	UpDn	41.17	65.49	15.48	35.48
RUBi [7]	UpDn	44.23	67.05	17.48	39.61
LMH [13]	UpDn	52.45	69.81	**44.46**	45.54
CF-VQA(SUM) [37]	UpDn	53.55	**91.15**	13.03	44.97
CF-VQA(SUM) [37]	S-MRL	55.05	90.61	21.50	45.61
CF-VQA(SUM) [37] + IntroD [38]	S-MRL	55.17	**90.79**	17.92	46.73
GGE [19]	UpDn	**57.32**	87.04	27.75	**49.59**
GenB (Ours)	UpDn	**59.15**	88.03	**40.05**	**49.25**
Methods based on balancing training data					
CVL [1]	UpDn	42.12	45.72	12.45	48.34
RandImg [46]	UpDn	55.37	83.89	41.60	44.20
SSL [52]	UpDn	57.59	86.53	29.87	50.03
CSS [9]	UpDn	58.95	84.37	49.42	48.21
CSS [9] + IntroD [38]	UpDn	60.17	89.17	46.91	48.62
MUTANT [15]	UpDn	61.72	88.90	49.68	50.78
D-VQA [47]	UpDn	61.91	88.93	52.32	50.39
KDDAug [10]	UpDn	60.24	86.13	55.08	48.08

VQA-CP2

Loss Component Ablation

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

Bias Model Ablation

Bias Model	VQA-CP2 test			
	All	Yes/No	Num	Other
UpDn	39.94	42.46	11.93	45.09
UpDn	52.47	88.20	30.09	40.38
Visual-Answer	41.03	42.69	12.66	47.93
Question-Answer	56.68	**89.30**	20.78	**49.43**
GenB Visual	49.54	72.05	12.58	47.89
GenB Question (Ours)	**59.15**	88.03	**40.05**	49.25

Architecture Ablation

Architecture	VQA-CP2 test				Δ Gap
	All	Yes/No	Num	Other	
UpDn [4]	39.94	42.46	11.93	45.09	**+19.21**
UpDn [4] + GenB	**59.15**	**88.03**	**40.05**	**49.25**	
BAN† [34]	37.35	41.96	12.08	41.71	**+20.02**
BAN† [34] + GenB	**57.37**	**89.11**	**29.52**	**48.37**	
SAN† [50]	38.65	40.59	12.98	44.67	**+18.07**
SAN† [50] + GenB	**56.72**	**88.84**	**19.04**	**50.22**	
LXMERT [45]	46.23	42.84	18.91	55.51	**+24.93**
LXMERT [45] + GenB (**Ours Best**)	**71.16**	**92.24**	**64.71**	**61.89**	
Reported LXMERT Performance					
LXMERT [45] + MUTANT [15]	69.52	93.15	67.17	57.78	
LXMERT [45] + D-VQA [47]	69.75	80.43	58.57	67.23	
LXMERT [45] + SAR [43]	62.12	85.14	41.63	55.68	

Qualitative Visualizations