

Good, Better, Best: Textual Distractors Generation for Multiple-Choice Visual Question Answering via Reinforcement Learning

Jiaying Lu, Xin Ye, Yi Ren, Yezhou Yang

Introduction

- DG-VQA: A novel task called textual Distractors Generation for VQA is proposed, as the demand of automated multiple-choice VQA grows.
- GOBBET: A Reinforcement Learning Framework for DG-VQA is developed, which does not require training samples.
- Experiments: Distractors generated by GOBBET can fool existing VQA models; these distractors can also help build robust VQA models.
- Case Study: The quality of generated distractors is further examined.



Q: What can be seen from the windows?

(a) Input Image and Question

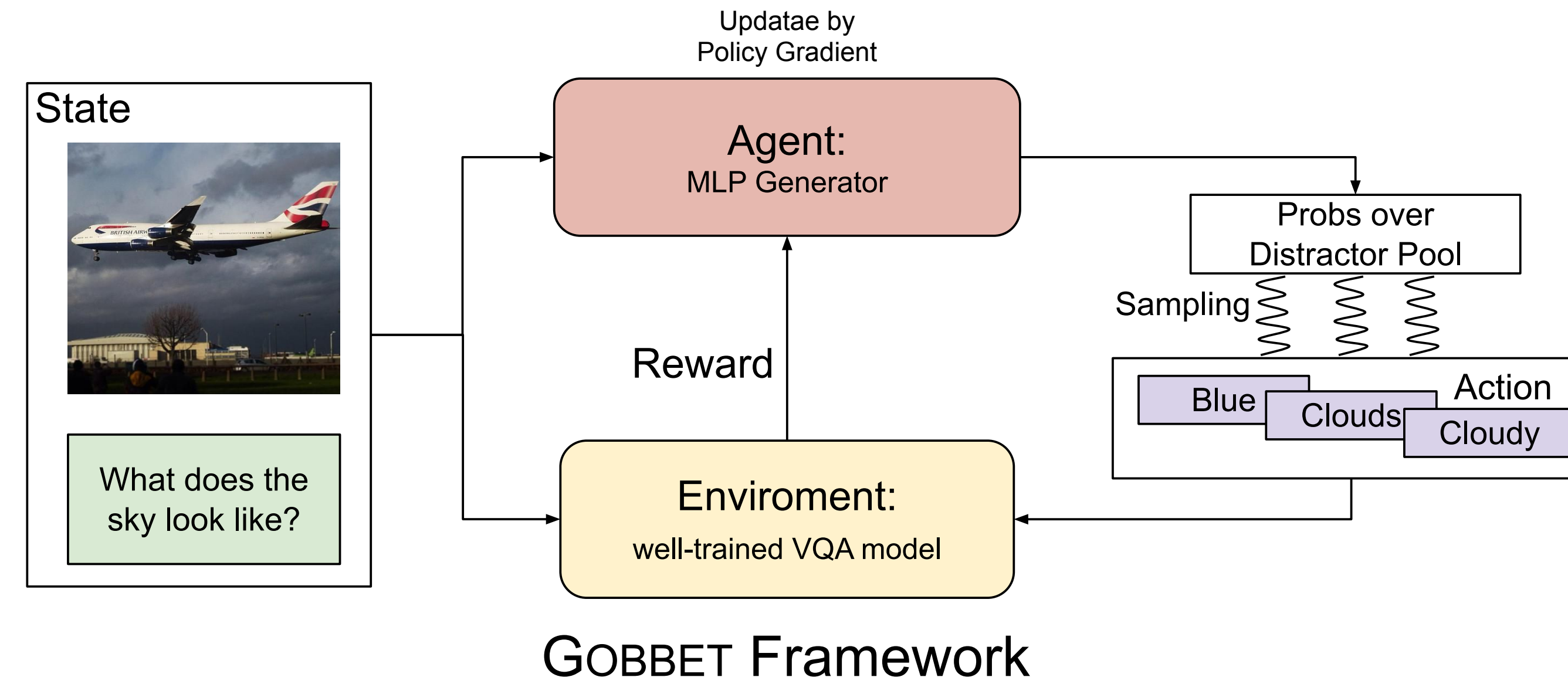
A: **More windows** ✓
A: The passing scenery
A: Snow falling
A: The backyard patio

(b) Original Answer Choices

A: More windows
A: A mirror
A: **Lights** ✗
A: A laptop

(c) Generated Distractors

Methods



- Agent: a distractor generator produces distractors according to the input image, question and answer
- Environment: a pre-trained VQA model serves as alternative knowledge source
- Reward: the performance degradations of pre-trained VQA models
- RL algorithm: a policy gradient-based method to maximize the "expected" rewards

Main Experiment

Table 1 in the paper. ΔAcc denotes the performance degradation.

Model	TellingVQA [49]		RevisitedVQA [19]		MCB [7]	
	Acc	ΔAcc	Acc	ΔAcc	Acc	ΔAcc
Original distractors	55.6%	-	64.8%	-	62.2%	-
Baselines						
<i>Q-type prior</i>	57.3%	-1.7%	68.7%	-3.9%	85.7%	-23.5%
<i>Adversarial Matching</i> [47]	54.7%	0.9%	71.7%	-6.9%	51.3%	10.9%
<i>LSTM Q+I</i> [1]	41.7%	13.9%	68.9%	-4.1%	85.7%	-23.5%
Proposed Methods						
Reward from RevisitedVQA						
- GOBBET-base	86.5%	-30.9%	0.01%	64.7%	26.5%	35.7%
- GOBBET-warmup	33.7%	21.9%	49.1%	15.8%	37.5%	24.7%

Case Study

Text in boldface denotes the answer option chosen by pre-trained VQA model.



Q: What does the sky look like?

Original Choices

A: **Stormy** ✓
A: Hazy
A: Windy

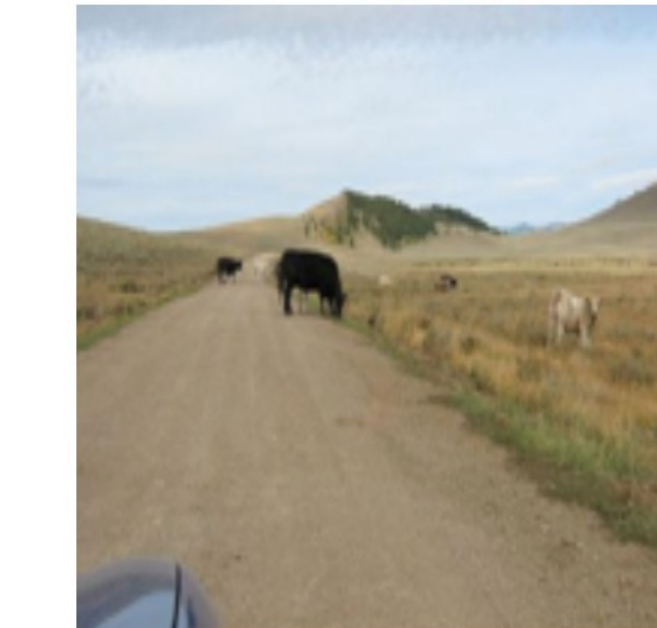
A: Sunny

Distractors by *Adversarial Matching*

A: Stormy
A: **Sky** ✗
A: Blue
A: Cloudy
A: **0** ✗

Distractors by GOBBET-warmup

A: Stormy
A: **Cloudy** ✗
A: Blue
A: Clouds
A: **One** ✗



Q: How many black cows are there?

A: **3** ✓

A: 9
A: 8

A: 7

A: 3

A: Zero

A: 5

A: **0** ✗

A: 3

A: Two

A: Four

A: **One** ✗



Q: What sport are they playing?

A: **Golf**

A: **Baseball** ✗
A: Hockey

A: Basketball

A: **Golf** ✓

A: Volleyball

A: Playing soccer

A: Soccer

A: Golf

A: **Baseball** ✗

A: Soccer

A: Tennis

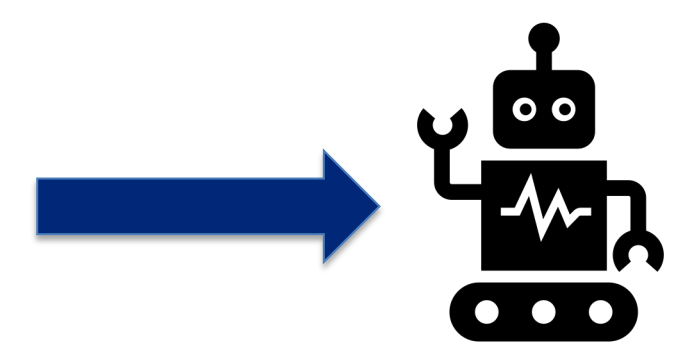
Problem Definition



I:

Q: How many black cows are there?

A: 3



D_1 : Two

D_2 : Four

D_3 : One

References

- [1] Antol, et al. "Vqa: Visual question answering." ICCV. 2015.
- [7] Fukui, et al. "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding." EMNLP. 2016.
- [19] Jabri, et al. "Revisiting visual question answering baselines." ECCV, 2016.
- [47] Zellers et al. "From recognition to cognition: Visual commonsense reasoning." CVPR. 2019.
- [49] Zhu, et al. "Visual7w: Grounded question answering in images." CVPR. 2016.