

Being Negative but Constructively:

Lessons Learnt from Creating Better Visual Question Answering Datasets

Wei-Lun (Harry) Chao*, Hexiang (Frank) Hu*, and Fei Sha

U. of Southern California

This work is partially supported by USC Graduate Fellowship, NSF IIS-1065243, 1451412, 1513966, 1208500, CCF-1139148, a Google Research Award, an Alfred. P. Sloan Research Fellowship and ARO\# W911NF-12-1-0241 and W911NF-15-1-0484.

Highlights

- Goal: How to design good visual QA dataset?
- Observation: On existing multiple-choice (MC) datasets, models can ignore information while still doing well.
- **Insight:** The design of *negative answers (decoys)* significantly affects the learning behavior.
- **Contributions:** Propose *principles and automatic* procedures to generate decoys, remedying two popular datasets (VQA, Visual7W) as well as creating a new one based on the Visual Genome (VG) projects.
- Link: http://www.teds.usc.edu/website_vqa/

Introduction

- Multiple-choice Visual Question Answering (QA): Given an *image (I)*, a *question (Q)*, and a *candidate* answers set (A)—a target (T)+ K decoys (D)—a machine needs to select the correct one.
- **Goal**: comprehend and reason with visual + language info.



Q: What vehicle is pictured?

A:	A car.	(0.21)
	A bus.	(0.62)
	A cab.	(0.50)
	A train.	(0.73)

How to design decoys is rarely discussed: random, high frequency, or <u>human generated ones by looking at Q and T</u>

Analysis

- Dataset: Visual7W (each IQA triplet has 4 candidates (C))
- VQA model:
- MLP to predict the score of each IQC triplet
- Features: CNN for I, WORD2VEC for Q and C, by concatenation

		,
Info.	Machine	Human
Random	25.0	25.0
Α	52.9	-
I+A	62.4	75.3
Q+A	58.2	36.4
I+Q+A	65.7	88.4
	•	-

Machines do well with partial info.

Diagnosis

- Decoys: not visually grounded (I+A: object/concept detection)
- > Targets: less used as decoys
- ➤ The following rule gives 48.73%

P(correct|C) = $\frac{\pi \cos x}{\# C \text{ as T+(\# C as D)/K}}$

Principles and automatic procedures

Principles

- > Neutrality (remove incidental statistics)
- > QoU (question only unresolvable)
- > IoU (image only unresolvable)

Automatic procedures

- Requirements: (1) IQT triplets are provided. (2) I with multiple QT pairs
- QoU-decoys: targets of similar Q'
- > IoU-decoys: targets of Q' of the same I
- Resolve ambiguity: (1) string matching (2) Wu-Palmer scores

IoU decoys

Overcast.	(0.55)
Daytime.	(0.49)
A building.	(0.48)
A train.	(0.54)

QoU decoys

A bicycle.	(0.28)
A truck.	(0.54)
A boat.	(0.46)
A train.	(0.51)

Experiments

Dataset	# images	# triplets	# Orig. D
VQA	205K	614K	17
Visual7W	27K	140K	3
VG	97K	1,445K	-

- ➤ All use MSCOCO images
- We create for each IQT triplet 3 IoU-decoys & 3 QoU-decoys
- User studies on 1000 triplets per dataset via AMT

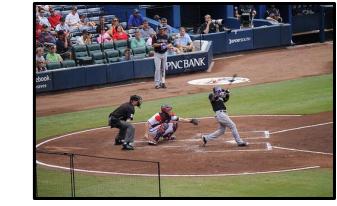
Visual7W					VQA	VG	
Method	Orig.	loU	QoU	IoU +QoU	Orig.	IoU +QoU	IoU +QoU
ЛLP-A	52.9	27.0	34.1	17.7	31.2	31.2	19.5
∕ILP-IA	62.4	27.3	55.0	23.6	42.0	34.1	25.2
MLP-QA	58.2	84.1	40.7	37.8	58.0	54.4	43.9
MLP-IQA	65.7	84.1	57.6	52.0	64.6	63.7	58.5
luman	88.4	1	-	84.1	88.5	89.0	82.5
Random	25.0	25.0	25.0	14.3	5.6	14.3	14.3

Machines need to use all three information (i.e., I, Q, A) to perform well.



What is the man wearing?

- A. Black. B. Mountains.
- C. The beach.
- D. Board shorts.
- E. He wears white shoes.
- F. A white button down shirt and a black tie.
- G. Wetsuit.



What is the right man on the right holding?

- A. Brown. B. The man on the right.
- C. Four. D. A bottle.
- E. A surfboard. F. Cellphone
- G. A bat.



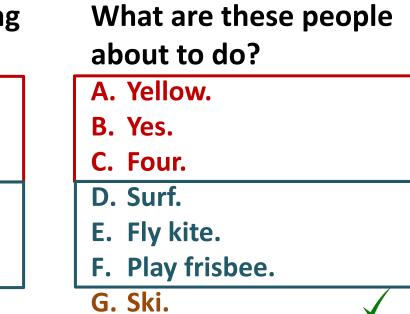
What is the train traveling over?



D. Train. E. South.

F. Forward

G.	Bridge.	



- [1] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In CVPR, 2016.
- [2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Zitnick, and D. Parikh. Vqa: Visual question answering. In ICCV, 2015.
- [3] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L-J Li, D. Shamma, M. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV, 2017.