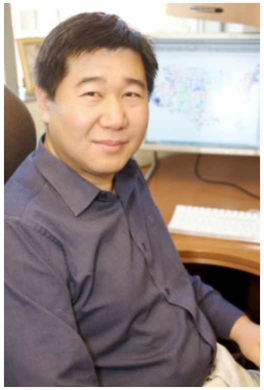
Being Negative but Constructively:

Lessons Learnt from Creating Better Visual Question Answering Datasets

Hexiang (Frank) Hu*

Wei-Lun (Harry) Chao*

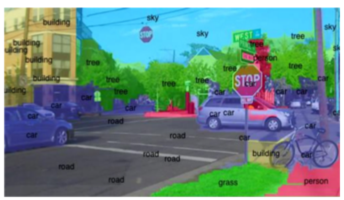


Fei Sha

USC

Vision and language

Computer Vision



Visual Captioning

Natural Language Image Retrieval

Visual Question Answering (Visual QA)

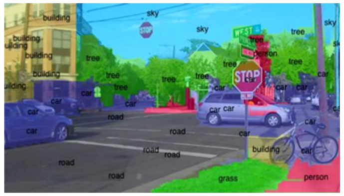
Natural Language Processing

Algorithm ⇐ **Dataset** ⇒ **Evaluation**

(learning signal) (test environment)

Vision and language

Computer Vision



Visual Captioning

Natural Language Image Retrieval

Visual Question Answering (Visual QA)

Natural Language Processing

How to design good datasets?

Outline

- Introduction on Visual QA

- Issues on existing datasets

Machines can do well while ignoring either visual or language information!

- **Our contributions:**
  - ➢Diagnosis of the issues
  - ➢Automatic procedures to remedy existing datasets
  - ➢Comprehensive evaluation on five existing datasets

Outline

- Introduction on Visual QA

- Issues on existing datasets

Machines can do well while ignoring either visual or language information!

- Our contributions:
  - ➢Diagnosis of the issues
  - ➢Automatic procedures to remedy existing datasets
  - ➢Comprehensive evaluation on five existing datasets

Visual question answering (Visual QA)

Intelligent system

Answer
Motorcycle cop.

Question
Who leads the parade?

comprehend and reason with both visual and language information

Evaluation

Difficult to evaluate

- **Open-world:** Prediction: <u>policeman</u> vs. GT: <u>motorcycle cop</u>
- **Multiple-choice:**

Candidate Answer Set (A):

The mayor.
The governor. *Decoy (D)*
The clowns.
Motorcycle cop. *Target (T)*

Candidate (C)

Selection accuracy as metric!

- **Goal:**

comprehend and reason with both visual and language information

Outline

- Introduction on Visual QA
- Issues on existing datasets

Machines can do well when they ignores either visual or language information!

- Our contributions:
  - Diagnosis of the issues
  - Automatic procedures to remedy existing datasets
  - Comprehensive evaluation on five existing datasets

How Visual QA datasets are created?

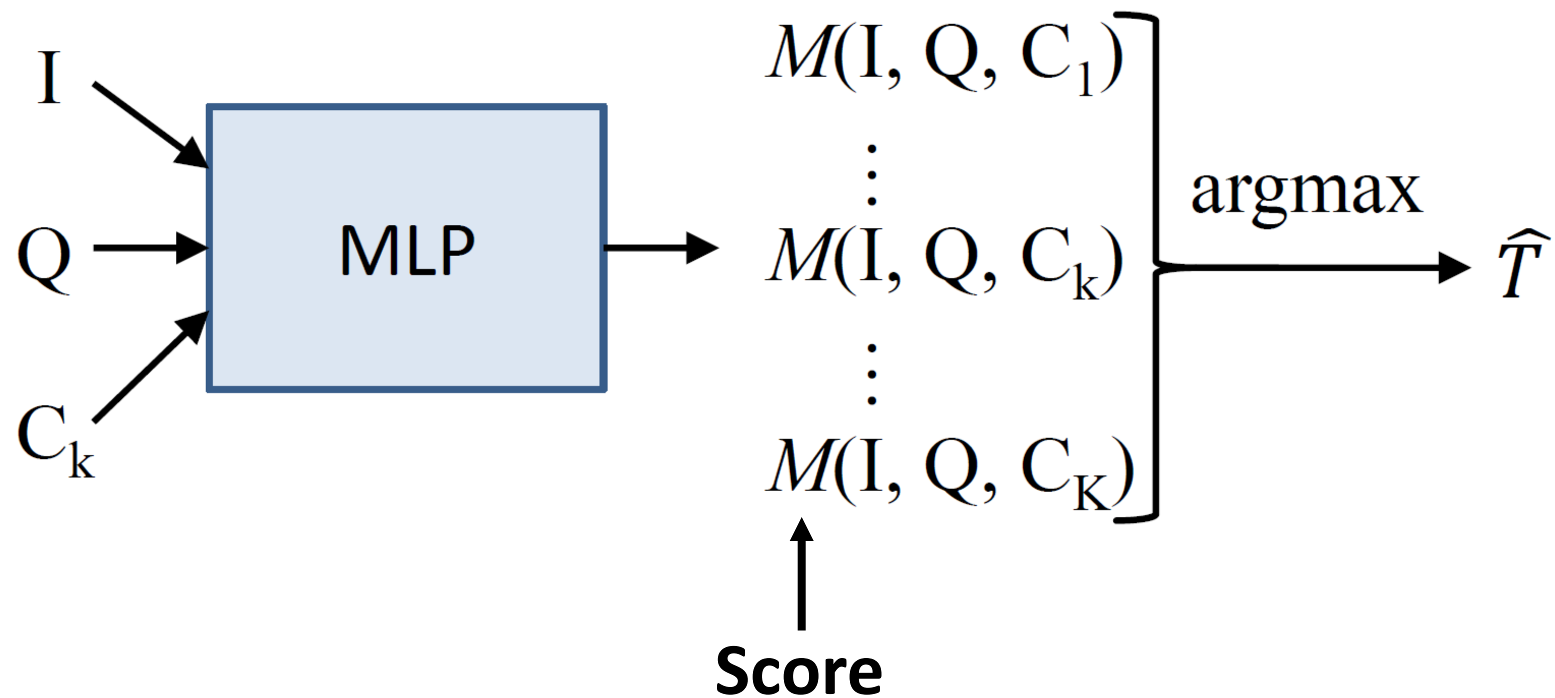
Pick an Image (I)	→	**Ask a Question (Q)**	→	**Generate the Targets (T)**	→	**Generate Decoys (D)**
		Who leads the parade?		Motorcycle cop.		The mayor. The governor. The clowns.

- **Generate decoys:**
  - Human generation according to (Q, T) [Visual7W]
  - Random or high-frequency (target) answers [VQA]

Detailed analysis on Visual7W

- Model: MLP with (I, Q, C) as the input [following Jabri et al., 2016]

Given an IQA triplet, where $A=\{C_1, \ldots, C_K\}$



Machine's Performance Analysis: with Full Information

Question

What vehicle is pictured?

a. A bicycle.
b. A bus.
c. A cab.
d. A train.

Information	Machines	Humans
Random	25.0%	25.0%
I + Q + A	65.7%	88.4%

[Example from Visual7W, Zhu et. al.]

Machine's Performance Analysis: with Partial Information

?

?

a. A bicycle.
b. A bus.
c. A cab.
d. A train.

Information	Machines	Humans
Random	25.0%	25.0%
I + Q + A	65.7%	88.4%
A	**52.9%**	25.0%

[Example from Visual7W, Zhu et. al.]

Machine's Performance Analysis: with Partial Information

?

Question

What vehicle is pictured?

a. A bicycle.
b. A bus.
c. A cab.
d. A train.

Information	Machines	Humans
Random	25.0%	25.0%
I + Q + A	65.7%	88.4%
Q + A	**58.2%**	**36.4%**

[Example from Visual7W, Zhu et. al.]

Machine's Performance Analysis: with Partial Information

?

a. A bicycle.
b. A bus.
c. A cab.
d. A train.

Information	Machines	Humans
Random	25.0%	25.0%
I + Q + A	65.7%	88.4%
I+A	**62.4%**	**73.5%**

[Example from Visual7W, Zhu et. al.]

Machine’s Performance Analysis: with Partial Information

?

a. A bicycle.
b. A bus.
c. A cab.
d. A train.

Information	Machines	Humans
Random	25.0%	25.0%
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

Machines can do well while ignoring information!

[Example from Visual7W, Zhu et. al.]

Outline

- Introduction on Visual QA
- Issues on existing datasets

Machines can do well while ignoring either visual or language information!

- **Our contributions:**
  - ➢Diagnosis of the issues
  - ➢Automatic procedures to remedy existing datasets
  - ➢Comprehensive evaluation on five existing datasets

Diagnosis: Shortcuts in decoys

(1) Decoys are **less frequently used** as targets

- **A frequency based baseline**

$$\mathcal{S}core(\mathrm{C}) = \frac{\#\ \text{of C as T}}{\#\ \text{of C as T} + \#\ \text{of C as D}}$$

$$\text{Prediction} = \mathop{\textbf{argmax}}_{\mathrm{C} \in A} \mathcal{S}core(\mathrm{C})$$

48%
accuracy

Diagnosis: Shortcuts in decoys

(2) Decoys might not be **visually grounded in images**

?

a. A bicycle.
b. A bus.
c. A cab.
d. A train.

Information	Machines
Random	25.0%
I + Q + A	65.7%
I+A	**62.4%**

➢ machines can perform attribute/object detection

(3) Decoys might not be **grounded in questions**

[Example from Visual7W, Zhu et. al.]

Outline

- Introduction on Visual QA
- Issues on existing datasets

Machines can do well while ignoring either visual or language information!

- **Our contributions:**
  - Diagnosis of the issues
  - Automatic procedures to remedy existing datasets
  - Comprehensive evaluation on five existing datasets

Principles for decoys

- **Neutrality:**

Equal likely used as the target

- **IoU (Image-only-Unresolvable):**

Plausible to the image

- **QoU (Question-only-Unresolvable):**

Plausible to the question

Automatic procedures

- Assumptions:
  - ➢A dataset with (I, Q, T) triplets is given.
  - ➢An image is associated with multiple (Q, T).

- For a (I, Q, T) triplet:

IoU-decoys: from T' of triplets with the same I

Q: What vehicle is pictured?
T: **A train.**

Q': When is the picture taken?
T': **Daytime.**

Automatic procedures

- Assumptions:
  - A dataset with (I, Q, T) triplets is given.
  - An image is associated with multiple (Q, T).

- For a (I, Q, T) triplet:

IoU-decoys: from T’ of triplets with the same I

QoU-decoys: from T’ of triplets with similar Q’

Q: What vehicle is pictured?
T: **A train.**

Q’: What is the vehicle?
T’: **A truck.**

Automatic procedures

- Assumptions:
  - A dataset with (I, Q, T) triplets is given.
  - An image is associated with multiple (Q, T).

- For a (I, Q, T) triplet:

IoU-decoys: from T' of triplets **with the same I**

QoU-decoys: from T' of triplets **with similar Q'**

Neutrality follows naturally

Illustration

Question:

What vehicle is pictured?

Candidate Answers:

Original	
a. A car.	(0.2083)
b. A bus.	(0.6151)
c. A cab.	(0.5000)
d. A train. ✓	(0.7328)

Freq-Baseline: 48%

Image only Unresolvable (IoU)	
a. Overcast. ✗	(0.5455)
b. Daytime.	(0.4941)
c. A building.	(0.4829)
d. A train.	(0.5363)

Freq-Baseline: 26%

Question only Unresolvable (QoU)	
a. A bicycle.	(0.2813)
b. A truck. ✗	(0.5364)
c. A boat.	(0.4631)
d. A train.	(0.5079)

Freq-Baseline: 30%

[Numbers are Score(C); accuracy are based on each set of decoys.]

Outline

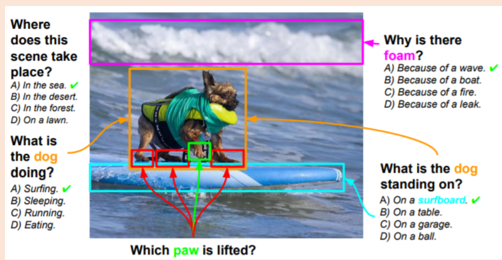
- Introduction on Visual QA
- Issues on existing datasets

Machines can do well while ignoring either visual or language information!

- **Our contributions:**
  - ➢ Diagnosis of the issues
  - ➢ Automatic procedures to remedy existing datasets
  - ➢ Comprehensive evaluation on five existing datasets

Experimental setup

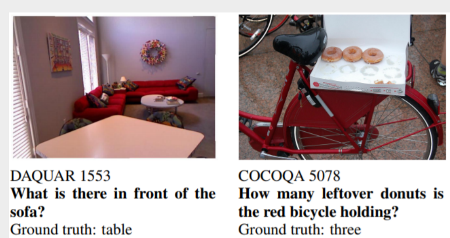
- Five datasets



Visual7W [CVPR 2016]

VQA [ICCV 2015]

Visual Genome (VG) [IJCV 2017]



COCOQA [NIPS 2015]

VQA2 [CVPR 2017]

Experimental setup

- Five datasets: all with images from MSCOCO

Dataset	# Training/Test triplets	Original decoys
Visual7W [CVPR 2016]	69K/42K	**3 (4 choose 1)**
VQA [ICCV 2015]	248K/121K	**17 (18 choose 1)**
Visual genome [IJCV 2017]	727K/433K	None
VQA2 [CVPR 2017]	444K/214K	None
COCOQA [NIPS 2015]	79K/39K	None

➢ Create **3 IoU & 3 QoU decoys (6 decoys in total)**

➢ Remove Yes/No triplets from VQA, VQA2 (~30%)

Original vs. New

Visual7W

Method	Original	IoU + QoU
MLP-A	52.9	17.7
MLP-IA	62.4	23.6
MLP-QA	58.2	37.8
MLP-IQA	**65.7**	**52.0**
Human	88.4	84.1
Random	25.0	14.3

VQA-
(exclude YES/NO)

Original	IoU + QoU
28.8	23.6
43.0	35.5
45.8	38.2
55.6	**53.7**
-	85.5
5.6	14.3

Original vs. New

Method	Visual7W Original	Visual7W IoU + QoU	VQA- (exclude YES/NO) Original	VQA- (exclude YES/NO) IoU + QoU
MLP-A	52.9	17.7	28.8	23.6
MLP-IA	62.4	23.6	43.0	35.5
MLP-QA	58.2	37.8	45.8	38.2
MLP-IQA	**65.7**	**52.0**	**55.6**	**53.7**
Human	88.4	84.1	-	85.5
Random	25.0	14.3	5.6	14.3

Algorithm with answer information only fails!

Original vs. New

	Visual7W		**VQA-** (exclude YES/NO)	
Method	Original	IoU + QoU	Original	IoU + QoU
MLP-A	52.9	17.7	28.8	23.6
MLP-IA	62.4	23.6	43.0	35.5
MLP-QA	58.2	37.8	45.8	38.2
MLP-IQA	**65.7**	**52.0**	**55.6**	**53.7**
Human	88.4	84.1	-	85.5
Random	25.0	14.3	5.6	14.3

Algorithm needs all information to perform well!

New multiple-choice datasets

Method	VG	VQA2-	COCOQA
MLP-A	19.5	21.3	26.6
MLP-IA	25.2	31.0	60.7
MLP-QA	43.9	37.2	51.4
MLP-IQA	**58.5**	**53.8**	**75.9**
Human	82.5	-	-
Random	14.3	14.3	14.3

Similar Results are obtained across all multiple-choice datasets!

Qualitative results

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. ✓

Where do the stairs lead?

A. A parking lot.
B. The building.
C. The windows.

D. From the canal to the bridge. ✗
E. Up.
F. To the building.

G. To the plane.

What is the color of his wetsuit?

A. When waves are bigger.
B. It is not soft and fine.
C. It is a picture of nature.

D. Green.
E. Blue.
F. Red.

G. It is black. ✓

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.

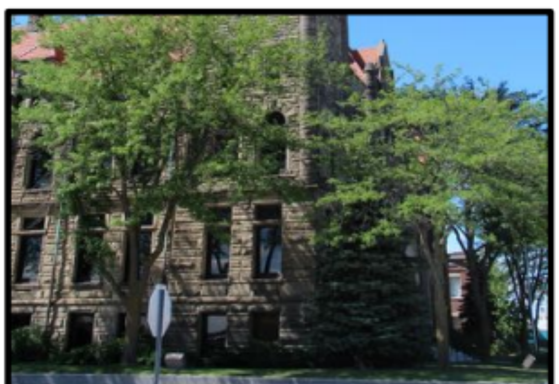
G. A bat. ✓

Failure cases

Who is wearing glasses?

A. Certificate.
B. Garland.
C. Three.
D. The man.
E. Person in chair.
F. The lady.
G. The woman.

Where are several trees?



A. Trees.
B. Clear and sunny.
C. Basement windows.
D. On both sides of road.
E. To left of truck.
F. On edge of the sidewalk.
G. In front of the building.

Conclusions

- Design good multiple-choice Visual QA datasets
- Analyze issues in existing datasets

Machines can do well while ignoring either visual or language information!

- Propose automatic procedures to remedy

IoU-decoys: from T’ of triplets with the same I

QoU-decoys: from T’ of triplets with similar Q’

- Conduct comprehensive experiments to validate

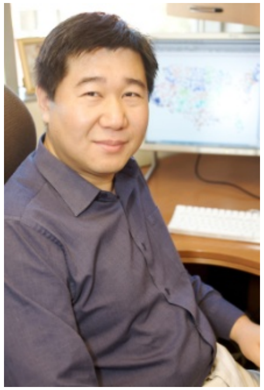
Q & A

All curated datasets available at:

http://www.teds.usc.edu/website_vqa/

Hexiang (Frank) Hu*

Wei-Lun (Harry) Chao*



Fei Sha