Visual7W: Grounded Question Answering in Images

Yuke Zhu[1], Oliver Groth[2], Michael Bernstein[1], Li Fei-Fei[1]
[1]Stanford University, United States [2]Dresden University of Technology, Germany

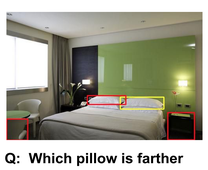
Visual7W Data

Q: What endangered animal is featured on the truck?
A: **A bald eagle.**
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?
A: **Onto 24 ¾ Rd.**
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

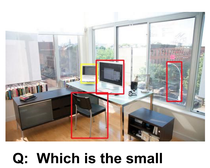
Q: When was the picture taken?
A: **During a wedding.**
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Q: Who is under the umbrella?
A: **Two women.**
A: A child.
A: An old man.
A: A husband and a wife.

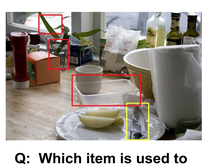


Q: Which pillow is farther from the window?

Q: Which step leads to the tub?



Q: Which is the small computer in the corner?



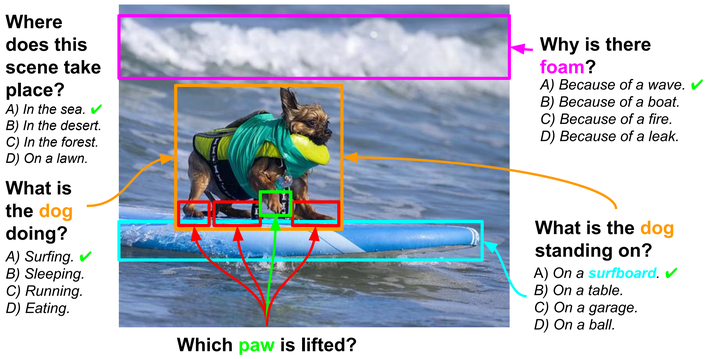
Q: Which item is used to cut items?

Telling QA

Question starts with who, what, where, when, why or how; answer consists of 4 textual multiple choices

Pointing QA

Question starts with which; answer consists of 4 visual bounding boxes in the image



Experiments & Results

Q: Who is behind the batter?
A1: Catcher.
A2: Umpire.
A3: Fans.
A4: Ball girl.

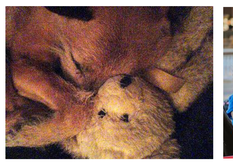
H: Catcher. ✓
M: Umpire. ✗

H: Catcher. ✓
M: Catcher. ✓

Q: What adorns the tops of the post?
A1: Gulls.
A2: An eagle.
A3: A crown.
A4: A pretty sign.

H: Gulls. ✓
M: Gulls. ✓

H: Gulls. ✓
M: A crown. ✗



Q: What kind of stuffed animal is shown?
A1: Teddy Bear.
A2: Monkey.
A3: Tiger.
A4: Bunny rabbit.

H: Monkey. ✗
M: Teddy Bear. ✓

H: Teddy Bear. ✓
M: Teddy Bear. ✓

Q: What animal is being petted?
A1: A sheep.
A2: Goat.
A3: Alpaca.
A4: Pig.

H: A sheep. ✓
M: A sheep. ✓

H: Goat. ✗
M: A sheep. ✓

Image | Multiple Choices | w/o Image | w/ Image

Top Grid: qualitative results of humans (H) and our model (M) in the V7W multiple choice benchmark with and without showing images while asking

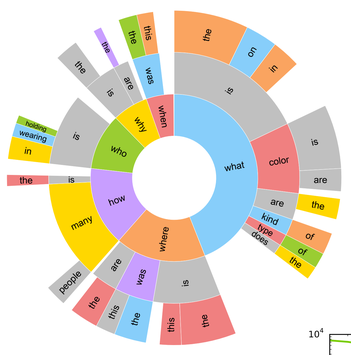
Bottom Table: comparison of model accuracies per question category and overall on the V7W benchmark split
Q indicates question being shown; ***I*** indicates image being shown
LogRegr: logistic regression model, classifies concatenation of embeddings
LSTM: model proposed by Malinowski et al., ICCV'15 [2]
LSTM-Att: our spatial attention LSTM, trained on V7W
LSTM-Att+: our spatial attention LSTM, trained on VG-QA [4]

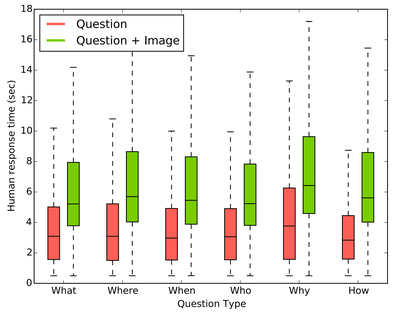
Method	What	Where	When	Who	Why	How	Which	Overall
Human (Q)	0.356	0.322	0.393	0.342	0.439	0.337	-	0.353
Human (Q+I)	0.965	0.957	0.944	0.965	0.927	0.942	0.973	0.966
LogRegr (Q)	0.420	0.375	0.666	0.510	0.354	0.458	0.354	0.383
LogRegr (I)	0.408	0.426	0.438	0.415	0.337	0.303	0.256	0.305
LogRegr (Q+I)	0.429	0.454	0.621	0.501	0.343	0.356	0.307	0.352
LSTM (Q)	0.430	0.414	0.693	0.538	0.491	0.484	-	0.462
LSTM (I)	0.422	0.497	0.660	0.523	0.475	0.468	0.299	0.359
LSTM (Q+I)	0.489	0.544	0.713	0.581	0.513	0.503	0.521	0.521
LSTM-Att (Q+I)	0.529	0.560	0.743	0.602	0.522	0.466	0.561	0.541
LSTM-Att+ (Q+I)	0.572	0.613	0.760	0.624	0.590	0.531	-	0.587

Statistics

- 47,300 COCO images
- 327,939 QA pairs: kkk,kkk telling QAs, kkk,kkk pointing QAs
- 1,311,756 multiple choices
- 561,459 object groundings
- 36,579 object categories
- avg. question length: 6.9 +/- 2.4
- avg. answer length: 2.0 +/- 1.4
- 27.6% of all answers >2 words
- 1,000 most frequent answers account for 63.5% of all answers

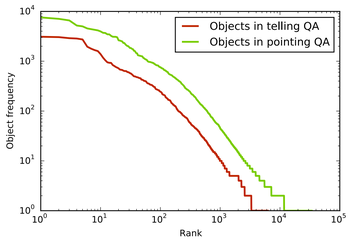
↓ Question Distribution
telling questions grouped by first 3 words (sparse arcs removed from sunburst)





Human Response Times ↑
significantly longer response delays when image absent across all telling QAs; image and QA tightly coupled

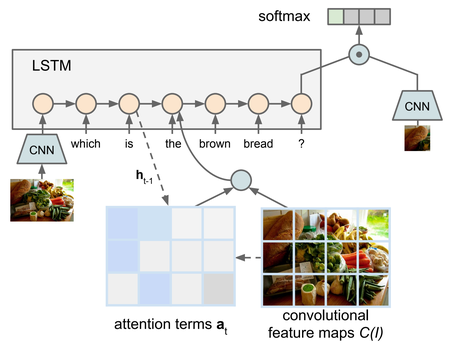
Object Distribution →
pointing questions refer to significantly more object categories



Top 20 objects

man	person
trees	window
shirt	table
wall	head
grass	water
tree	snow
woman	hair
people	sign
sky	ground
building	wood

Spatial Attention LSTM



Q: Which is the brown bread?
A1: A2: A3: A4:

Attention Terms: $w_a^T, W_{he}, W_{ce}, b_a$ trainable

$$a_t = softmax(e_t)$$

$$e_t = w_a^T tanh\left(W_{he} h_{t-1} + W_{ce} C(I)\right) + b_a$$

h_{t-1} hidden state vector
$C(I)$ convulotional image features

Related Work

[1] A. Agrawal et al. VQA: Visual Question Answering. *ICCV, 2015.*
[2] M. Malinowski et al. Ask your neurons: A neural-based approach to answering questions about images. *ICCV, 2015.*
[3] D. Geman et al. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences, 2015.*
[4] R. Krishna et al. Visual Genome: Connecting language and vision using crowdsourced dense annotations. *arXiv preprint, 2016*

Project Page
http://web.stanford.edu/~yukez/visual7w/

Data
Images, QAs & Grounding

Code
Evaluation Toolkit & Torch Model

Crowdsourcing
AMT Interfaces available

All collected QA data contributed to visualgenome.org
VG-QA (total) [4]: 101,174 Images; 1,773,258 QA Pairs