Making the V in VQA Matter (aka VQA v2.0): Elevating the Role of Image Understanding in Visual Question Answering

Yash Goyal*[1] Tejas Khot*[1] Douglas Summers-Stay[2] Dhruv Batra[3] Devi Parikh[3]

[1]Virginia Tech [2]Army Research Laboratory [3]Georgia Tech *=equal contribution

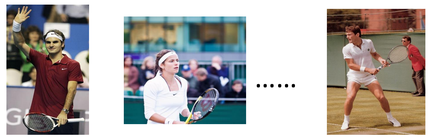
Dataset and challenge details at:
www.visualqa.org

Contributions

- Introduce a larger (1.8x) and more balanced VQA v2.0 dataset
  - Each question is associated with a pair of similar images that result in two different answers to the question
- Reduce language priors from VQA dataset (Antol et al., ICCV 15)
- Develop a model for counter-example explanations!
  - What would the image look like if the answer was different?

Motivation

Q: What sport is … ?
A: "tennis" - 41%



Q: How many … ?
A: "2" - 39%

Q: Is there a clock … ?
A: "yes" - 98%

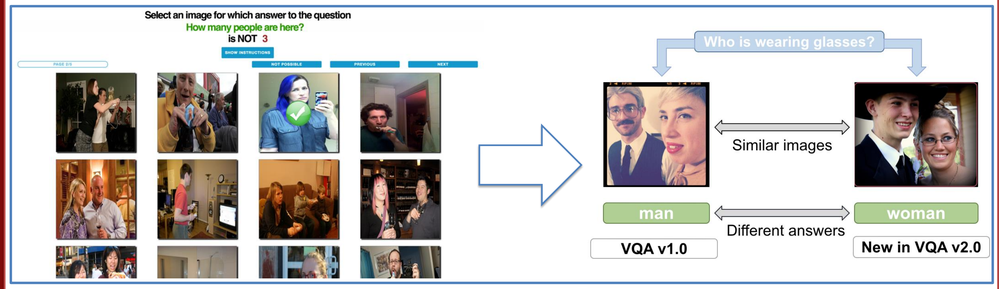
Q: Is the man standing… ?
A: "no" - 69%

Datasets have language biases!
These biases make it difficult to measure progress in image understanding.

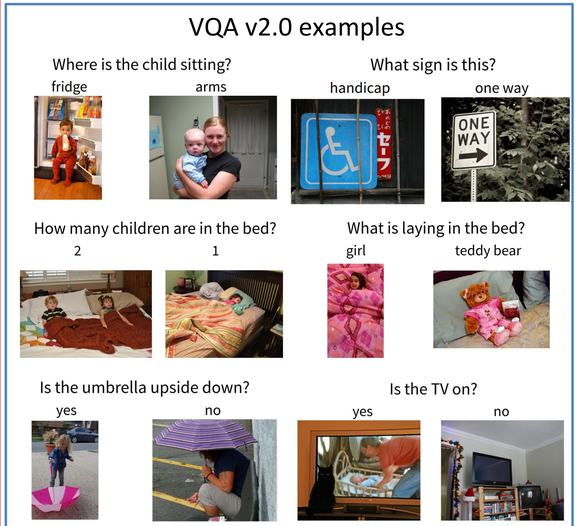
Solution: Balancing the VQA dataset

We balance each question = Collect an image with different answer

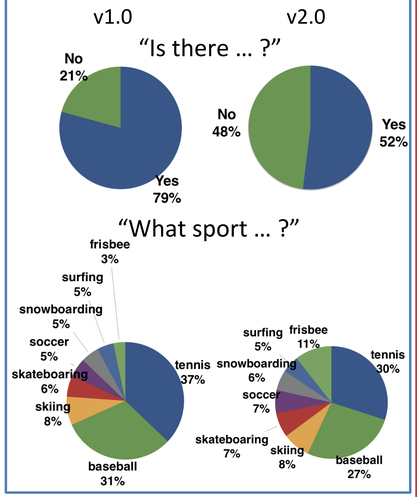
Humans are shown the *question*, the *answer* and 24 *images* similar to the original image, and asked to pick an *image* such that the *question* makes sense and the *answer* is not correct for the *question*.



VQA v2.0 examples



Answer entropy increases by 56%!



VQA v2.0 dataset: 443K train, 214K val and 453K test questions

Benchmarking VQA Models

Train/Test	v1.0/v1.0	v1.0/v2.0	v2.0*/v2.0
Prior	27.38	24.04	24.04
Language only	48.21	41.40	41.47
LSTM + CNN (Antol et al., ICCV 15)	54.40	47.56	49.23
HieCoAtt (Lu et al., NIPS 16)	57.09	50.31	51.88
MCB (Fukui et al., EMNLP 16)	60.36	54.22	56.08

1. Drop in performance by 6-7% when evaluated on VQA v2.0 dataset
2. Gain 1-2% back when re-trained on VQA v2.0 dataset

Yes/No Type Questions	v1.0/v1.0	v1.0/v2.0	v2.0*/v2.0
HieCoAtt (Lu et al., NIPS 16)	79.99	67.62	70.93
MCB (Fukui et al., EMNLP 16)	81.20	70.40	74.89

Biggest drop in performance (11-12%) in yes/no type questions
- Important because SOTA VQA models achieve similar and high accuracy on yes/no type questions in VQA v1.0

* models are trained on half of v2.0 train set to be comparable to v1.0 train set

Counter-Example Explanations

New explanation modality: when asked a question about image, model gives an answer along with a set of ranked images it believes are counter-examples for the QA pair

Q: Which way is its head turned?
A: left

Q: What color is the plate?
A: blue

Counter-example images