The meaning of “most” for visual question answering models

Alexander Kuhnle   Ann Copestake

Department of Computer Science and Technology
University of Cambridge
{aok25,aac10}@cam.ac.uk

Pietroski et al. (2009): two interpretation strategies for “most”

Cardinality-based strategy

$$\text{most}(A, B) \Leftrightarrow |S_{A \wedge B}| > 1/2 \cdot |A|$$
$$\Leftrightarrow |S_{A \wedge B}| > |S_{A \wedge \neg B}|$$

1. Estimate the number of entities satisfying both predicates (*“red squares”*) and the number satisfying one predicate but not the other (*“non-red squares”*).
2. Compare these number estimates and check whether the former is greater than the latter.

Pairing-based strategy

$$A \leftrightarrow B :\Leftrightarrow \forall x : A(x) \Leftrightarrow B(x) \;\; \Leftrightarrow |S_A| = |S_B|$$
$$\text{most}(A, B) \;\; \Leftrightarrow \exists S \subsetneq S_{A \wedge B} : S \leftrightarrow S_{A \wedge \neg B}$$

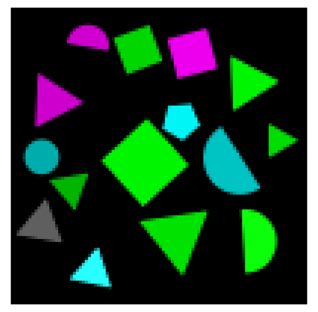
1. Successively match entities satisfying both predicates (*“red squares”*) uniquely with entities satisfying one predicate but not the other (*“non-red squares”*).
2. The remaining entities are all of one type, so pick one and check whether it is of the first type (*“red square”*).

x: entity, A and B: predicates (e.g., *“square”* and *“red”*), $A(x)$ true iff x satisfies A, and $S_A = \{x : A(x)\}$: set of entities satisfying A.

Examples

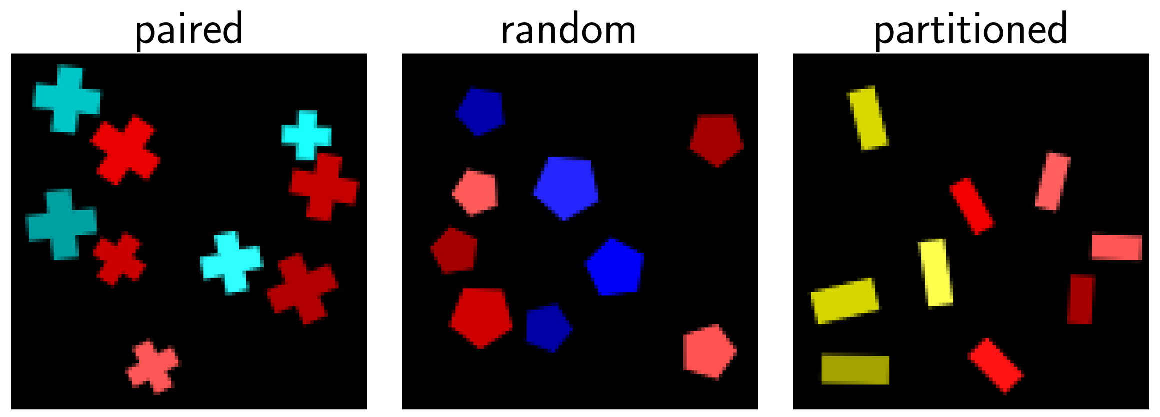
Training examples

- Exactly two squares are yellow.
- Exactly no square is red.
- More than half the red shapes are squares.
- More than a third of the shapes are cyan.



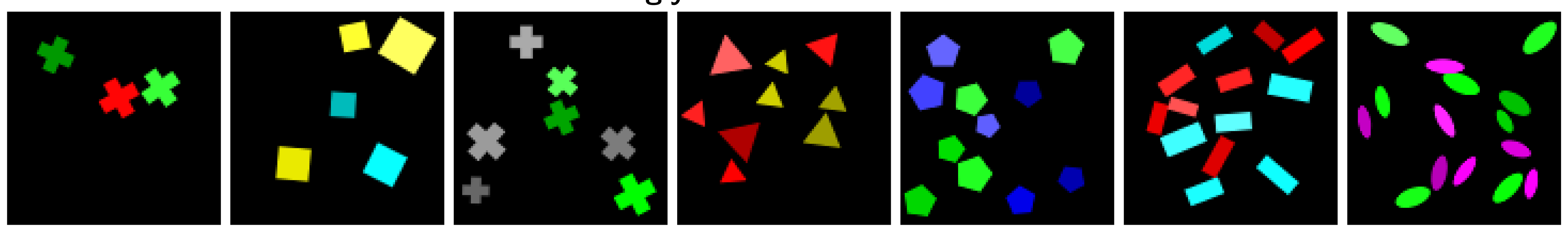
- Less than half the shapes are green.
- Exactly all magenta shapes are squares.
- At most five shapes are magenta.
- At least one triangle is gray.

Three types of spatial arrangements



“More than half the shapes are red shapes?”

Increasingly balanced attribute ratios



GitHub projects & PDF versions

ShapeWorld: `https://github.com/AlexKuhnle/ShapeWorld`

FiLM for ShapeWorld: `https://github.com/AlexKuhnle/film`

Paper & poster PDF, plus related papers: `https://www.cl.cam.ac.uk/~aok25/`

Experimental setup: task, model, data, etc

Task: image caption agreement

Model: FiLM (Perez et al. 2018)

Variants: -pre indicates pretrained CNN module, -coll indicates shape collisions allowed

Training data: 100k images with 5 captions per image

Training: 100k iterations with batch size 64 ($\sim$ 13 epochs)

Validation data: 20k instances

Test data: 48 configurations with 1024 instances each

Numbers: *“zero”* to *“five”*

Quantifiers: *“no”, “a/three quarter(s)”, “a/two third(s)”, “all”*

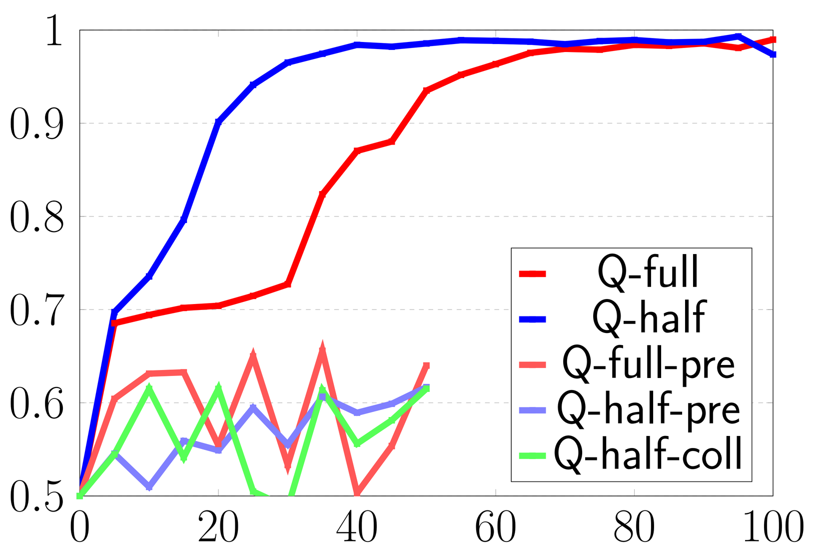
Modifiers: *“less than”, “at most”, “exactly”, “at least”, “more than”, “not”*

Training datasets: Q-full contains all quantifiers, Q-half contains only *“more than half”* and *“less than half”*

Test datasets: One contrasting attribute, close-to-balanced contrast attribute ratios, area- vs size-controlled, random/paired/partitioned positioning

Experimental results

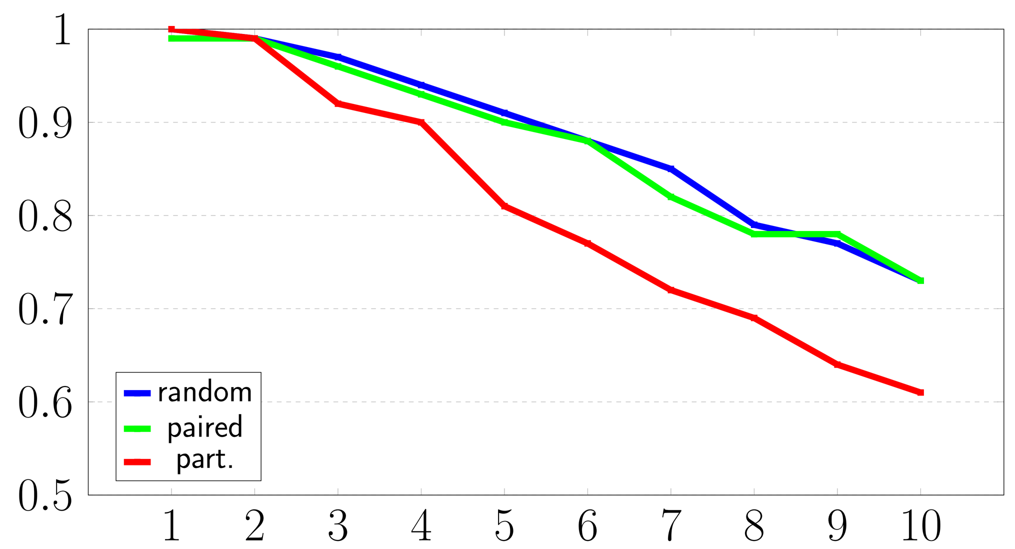
Training performance



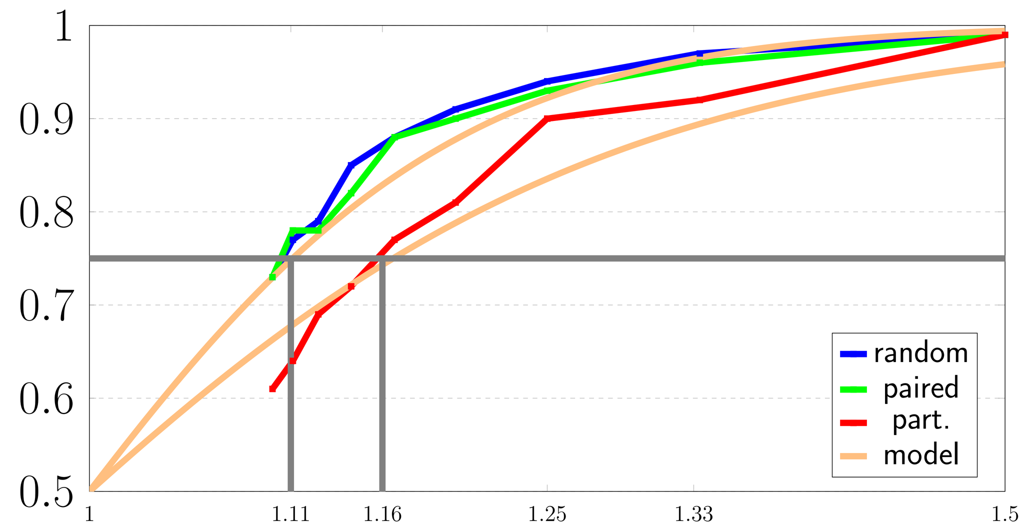
Evaluation performance

train	mode	size-controlled all	1:2	2:3	3:4	4:5	5:6	6:7	7:8	area-controlled all	1:2	2:3	3:4	4:5	5:6	6:7	7:8
Q-full	random	92	100	99	97	94	91	88	85	93	100	99	97	93	91	86	82
	paired	93	99	99	96	93	90	88	82	93	99	99	96	91	87	84	80
	part.	89	100	99	92	90	81	77	72	89	99	98	92	88	82	78	72
Q-half	random	92	100	100	98	93	88	88	87	93	100	100	97	92	86	85	82
	paired	92	100	100	96	90	86	84	79	92	100	99	96	87	84	79	76
	part.	91	100	99	96	86	83	83	80	91	100	99	94	89	83	83	80

Weber fraction: performance for increasingly balanced ratios



Q-full model performance for increasingly balanced ratios (x-axis indicates ratio via n:n+1)



Performance as a function of the actual ratio fraction (n+1)/n, with Weber fraction (75%) highlighted