Roses are Red, Violets are Blue... But Should VQA expect Them To?

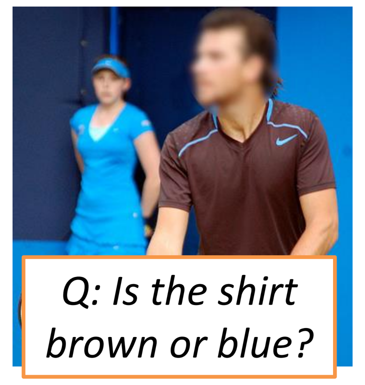
Corentin Kervadec[1,2] Grigory Antipov[1] Moez Baccouche[1] Christian Wolf[2] [1] Orange Innovation [2] LIRIS, INSA Lyon

GQA-OOD: a benchmark targeting biases in VQA.

https://github.com/gqa-ood/GQA-OOD

Visual Question Answering

- **Answer questions** posed over **images**
- Evaluate high-level **reasoning**
- Datasets are very **imbalanced**
- Models overly rely on **biases**



Biases in VQA

In-domain evaluation (*overall accuracy*) is misleading:
➢ favour models exploiting subtle training set statistics.

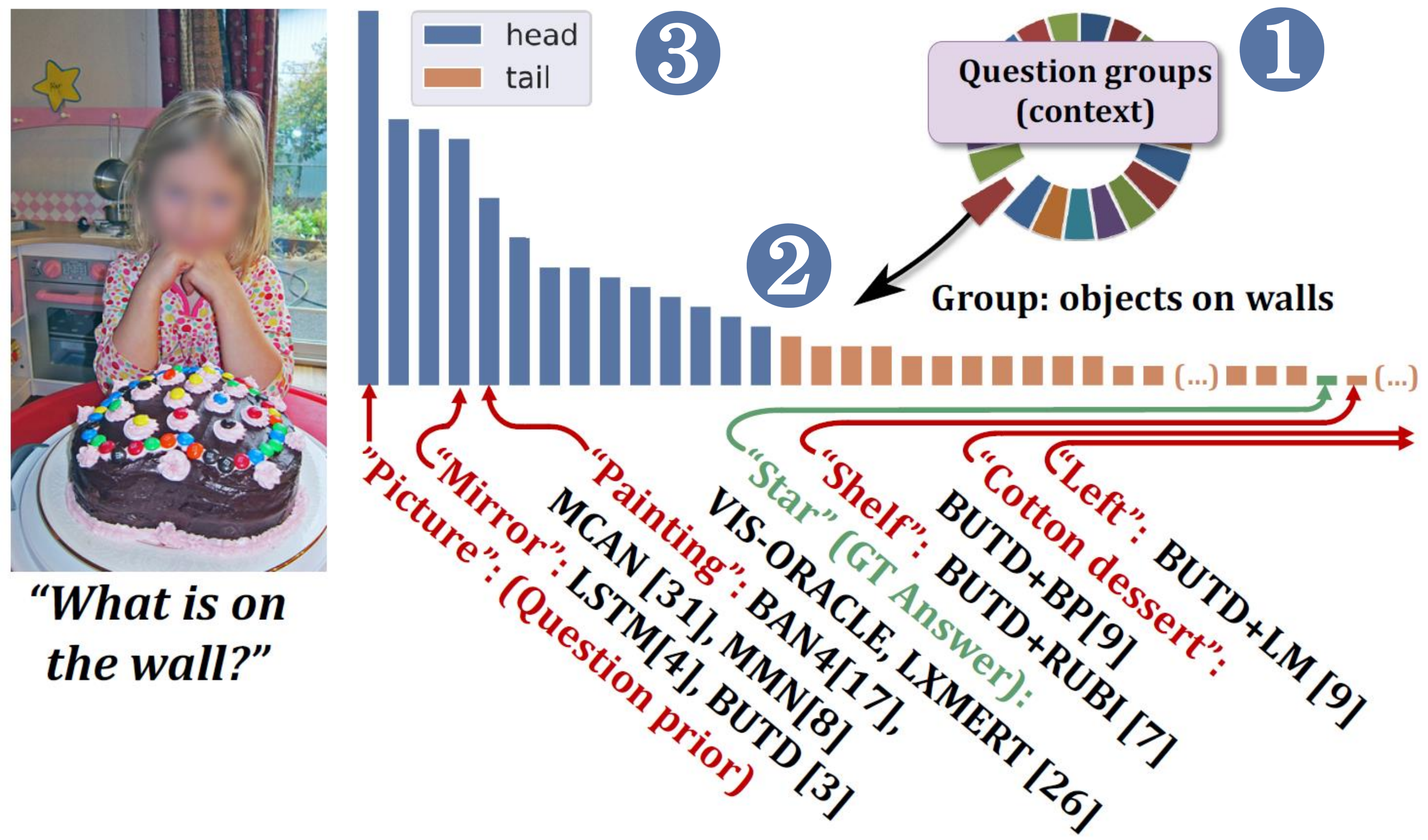
Alternatively, naively evaluating generalization by introducing **artificial distribution shift** between train and test splits is also not completely satisfying [1].

Our contributions

We propose the GQA-OOD benchmark:
➢ fine-grained reorganization of GQA dataset [2]
A two-in-one evaluation:
➢ measure accuracy over both rare and frequent QA
➢ compare in- *vs.* out-of-distribution accuracy

SOTA VQA models, *including bias reduction methods*, fail to address questions involving infrequent concepts.

GQA-OOD: a benchmark for OOD settings



1. Split data into question groups
2. Construct answer histogram of each group
3. Identify **head** (frequent) and **tail** (rare) questions *in the group*

We obtain 3 metrics: *acc-all* (**all** samples) • *acc-tail* (**rare**) • *acc-head* (**frequent**)

Models fail on rare question-answer pairs

Model	Uses image	acc-all	acc-tail	acc-head	Δ
Quest. Prior	✗	21.6	17.8	24.1	35.4
LSTM [4]	✗	30.7	24.0	34.8	45.0
BUTD [3]	✓	46.4±1.1	42.1±0.9	49.1±1.1	16.6
MCAN [29]	✓	50.8±0.4	46.5±0.5	53.4±0.6	14.8
BAN4 [18]	✓	50.2±0.7	47.2±0.5	51.9±1.0	**9.9**
MMN [8]	✓	52.7	48.0	55.5	15.6
LXMERT [24]	✓	**54.6**	**49.8**	**57.7**	15.9

Technique	acc-all	acc-tail	acc-head	Δ
BUTD [3]	46.4±1.1	**42.1**±0.9	49.1±1.1	16.6
+RUBi+QB	**46.7**±1.3	**42.1**±1.0	**49.4**±1.5	17.3
+RUBi [7]	38.8±2.4	35.7±2.3	40.8±2.7	14.3
+LM [9]	34.5±0.7	32.2±1.2	35.9±1.2	**11.5**
+BP [9]	33.1±0.4	30.8±1.0	34.5±0.5	12.0

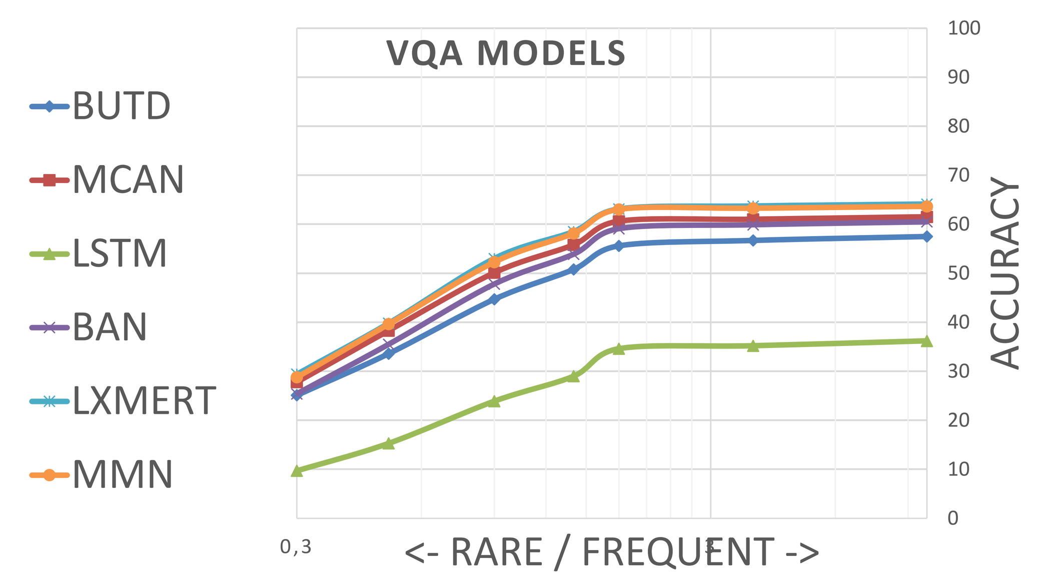
Left: VQA models. Up: bias reduction methods*

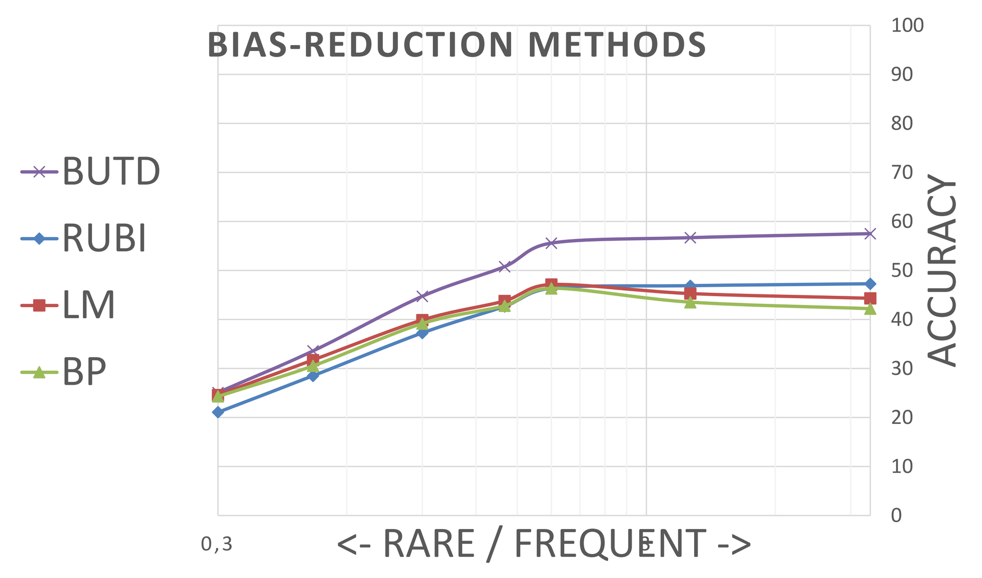
All models are trained on the GQA train split. VQA models (including bias-reduction methods!) fail to generalize on infrequent association of concepts.

* References are in the paper

Visualising the **generalisation** behaviour

We plot *acc-tail* (*y axis*) while varying the degree of rarity (*x-axis*):

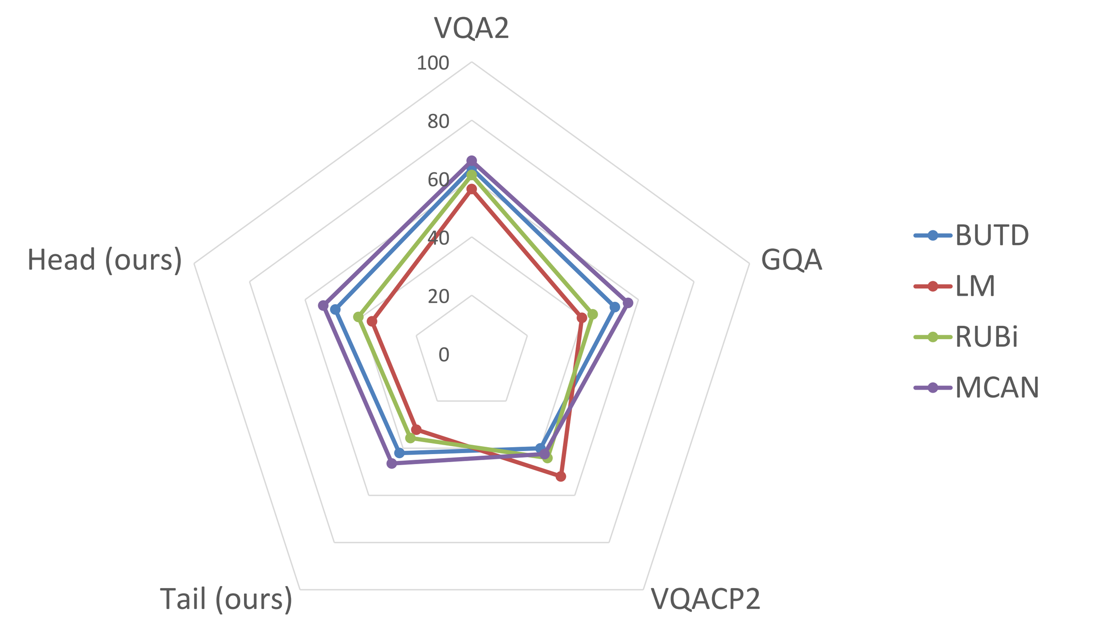




For VQA models (*left*) and bias reduction methods (*right*), accuracy dramatically decreases when QA are rare. It shows that these models exploit biases instead of reasoning.

Comparison with other benchmarks

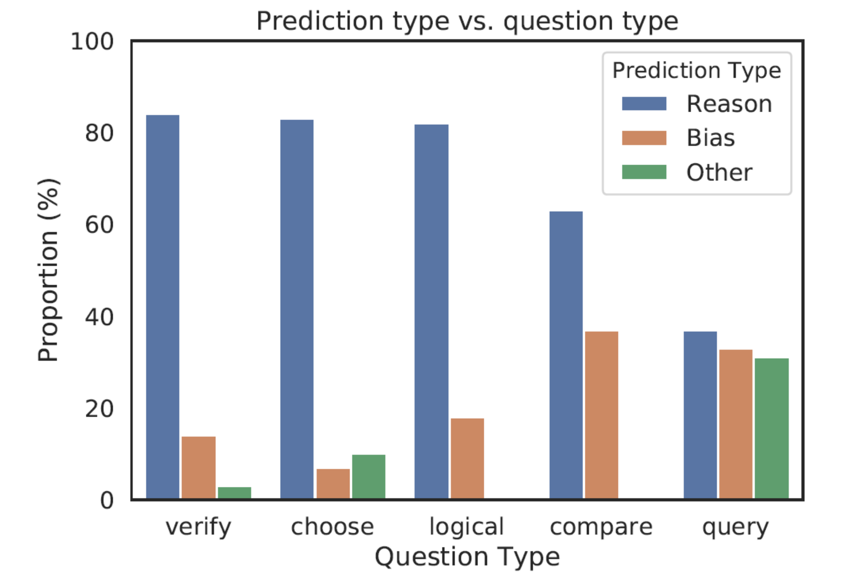
No significant impact on VQA2 or GQA ranking:



But very different from the ones on VQA CP:
➢ Bias reduction methods designed for VQA CP achieve very low performance on our benchmark.

Exploiting biases *vs.* reasoning

We use our metrics to estimate when LXMERT is reasoning or exploiting biases:



Future efforts on improvements of model capacities to answer open questions (e.g typed as query) should be particular fruitful.

[1] D. Teney, et al. On the value of out-of-distribution testing: An example of goodhart's law. In Proc. NeurIPS 2020.
[2] D. Hudson, et al. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proc CVPR 2019