Biases and Reasoning Patterns in Visual Question Answering (VQA)

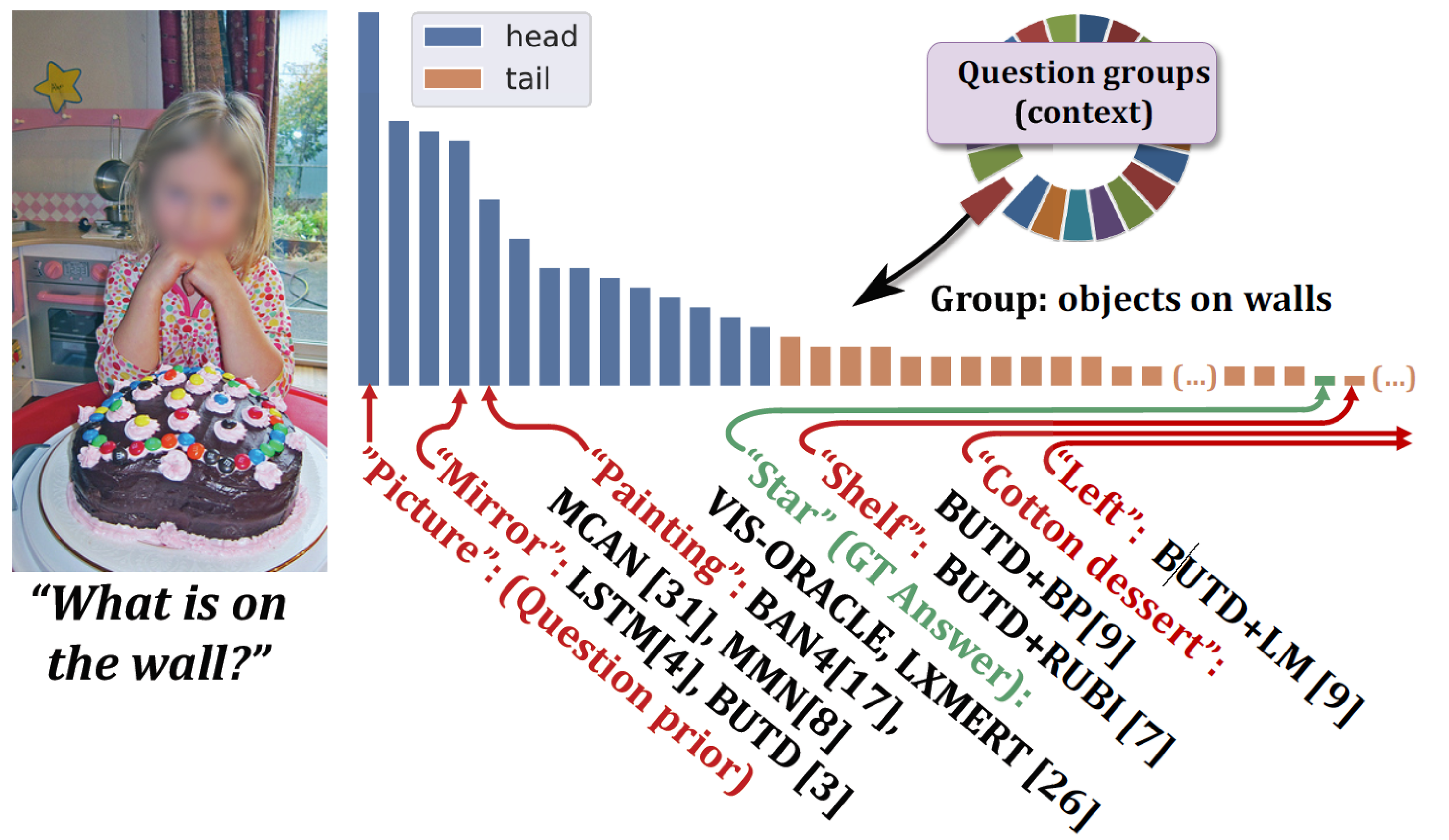
CVPR VIRTUAL JUNE 19-25

C. Kervadec[1,2] T. Jaunet[2] G. Antipov[1] M. Baccouche[1] R. Vuillemot[3] C. Wolf[2]

[1] Orange Innovation [2] LIRIS, INSA Lyon [3] LIRIS, EC Lyon

Roses are red, violets are blue… But should VQA expect them to?

paper link



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

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

Model	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

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

Poster at CVPR'21: SESSION TWO

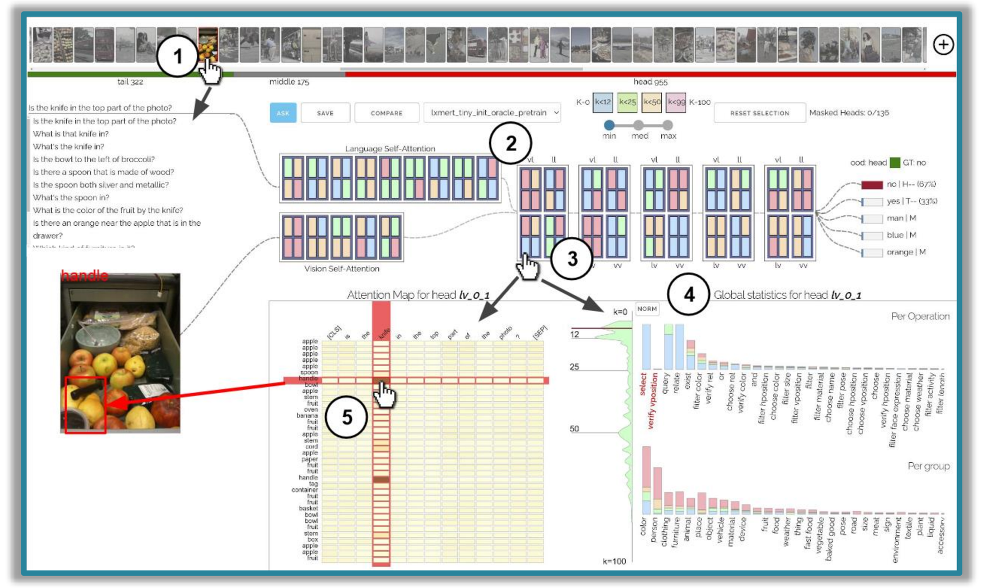
VisQA

paper link arxiv/submitted

Interactive tool **visualizing** attention heads of **VL-Transformers** for **VQA**.

- Fine grained visualization of the interactions at work in the attention layers.
- Instance based, you can ask your question

Explore the question of **reasoning** vs. **bias** exploitation. Check out the demo!

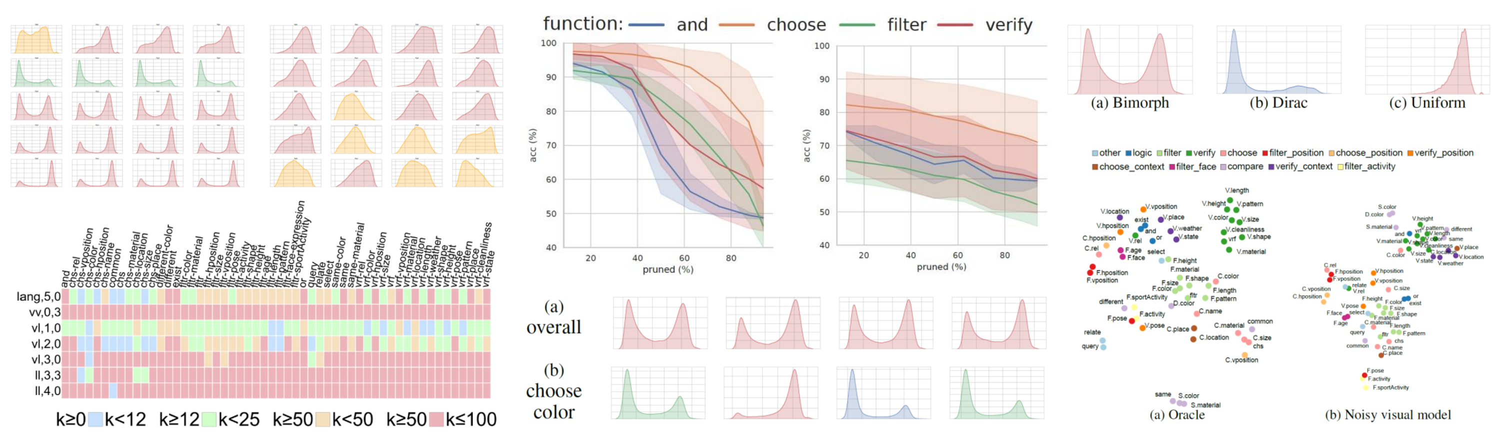


https://visqa.liris.cnrs.fr/

Reasoning patterns in VQA

In-depth analysis of **reasoning patterns** at work in VQA

- Analysing attention mechanisms at work in a **VL-Transformer**
- Comparing models with **perfect-sight** *vs.* **noisy visual inputs**



*See the paper to get more details on these figures

We observe significant differences between **Oracle** (perfect-sight) and **Standard** (noisy vision): we highlighted the **Oracle** ability of *adapting reasoning to the task at hand*.

paper link

- **Uncertainty** in vision **prevents from** successful learning of **reasoning**

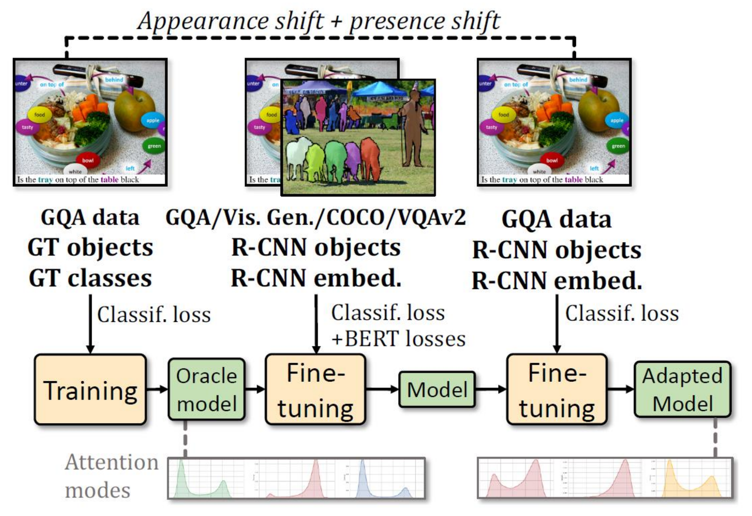
Poster at CVPR'21: SESSION THREE

Oracle transfer

paper link

We propose to **transfer** learned **reasoning patterns** form **Oracle** to **Standard**:

1. Train the oracle on perfect vision
2. Optionally, BERT-like pretraining
3. Finetune with standard (noisy) vision



Oracle transfer improves accuracy in **in-** and **out-of distribution** settings!

Model	Pretraining: Oracle	Pretraining: LXMERT/BERT	GQA-OOD [22] acc-tail	GQA-OOD [22] acc-head	GQA [19] overall	VQAv2 [17] overall
(a) Baseline			42.9	49.5	52.4	-
(b) Ours	✓		**48.5**	**55.5**	**56.8**	-
(c) Baseline (+LXMERT/BERT)		✓	47.5	54.7	56.8	69.7
(d) Ours (+LXMERT/BERT)	✓	✓	**48.3**	**55.2**	**57.8**	**70.2**

Do VQA models reason?

Evaluate — Analyse — Improve

VQA models are notorious for their tendency to rely on dataset **biases**.

The large and unbalanced diversity of concepts involved in VQA and the lack of well-annotated data tend to prevent deep learning models from learning to **reason**. Instead, it leads them to perform **shortcuts**[1], relying on specific training set statistics, which is not helpful for generalizing to real-world scenarios.

*We propose to **evaluate**, **analyse** and **improve** Visual Question Answering (VQA) models through the lens of **biases and reasoning**.*

Supervising reasoning transfer

paper link arxiv /submitted

Supervising the model to predict reasoning operations:

- A catalyst for transferring reasoning patterns

Theoretical analysis (based on PAC learning):

- Show benefice of supervising program prediction in VQA deriving bounds on sample complexity.

✓ Enhances the oracle transfer performance.
✓ Achieves SOTA accuracy while using a smaller model and less training data.

Method	Visual feats.	Additional supervision	Training data (M) Img	Training data (M) Sent	GQA-OOD acc-tail	GQA-OOD acc-head	GQA bin.	GQA open	GQA all
BAN4 [Kim et al., 2018]	RCNN	-	≈ 0.1	≈1	47.2	51.9	76.0	40.4	57.1
MCAN [Yu et al., 2019]	RCNN	-	≈ 0.1	≈1	46.5	53.4	75.9	42.2	58.0
Oracle transfer (ours)	RCNN	-	≈0.18	≈1	48.3	55.5	75.2	44.1	58.7
MMN [Chen et al., 2021]	RCNN	Program	≈0.1	≈15	48.0	55.5	78.9	44.9	60.8
LXMERT [Tan and Bansal, 2019]	RCNN	-	≈0.18	≈9	**49.8**	57.7	77.8	45.0	60.3
Supervised transfer (ours)	VinVL	Program	≈0.1	≈15	49.1	59.7	80.1	48.0	63.0
NSM [Hudson and Manning, 2019]	SG	Scene graph	≈0.1	≈1	-	-	78.9	**49.3**	63.2
OSCAR+VinVL [Zhang et al.,]	VinVL	-	≈5.7	≈9	-	-	**82.3**	48.8	**64.7**

[1] R. Geirhos, et al. Shortcut learning in deep neural networks. In Nature Machine Intelligence 2020
[2] D. Hudson, et al. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR 2019