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

C. Kervadec[1,2]

Joint work with: T. Jaunet[2,3] G. Antipov[1] M. Baccouche[1] R. Vuillemot[2,4] C. Wolf[2,3]

[1]Orange Innovation, France. [2]LIRIS, [3]INSA, [4]ECL, Lyon, France.

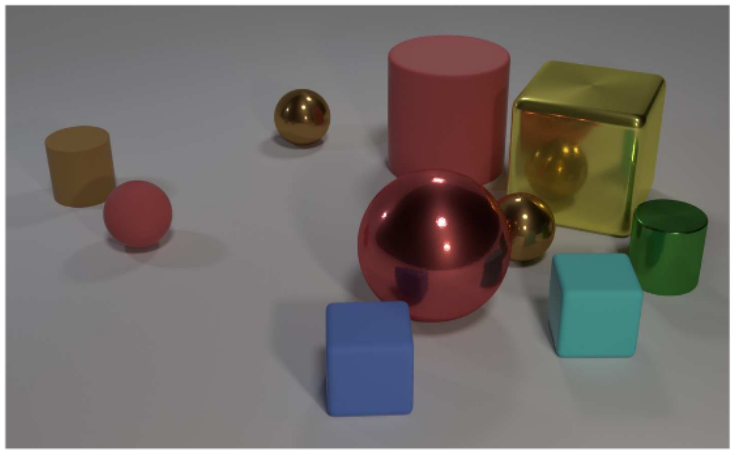
Biases and Reasoning

Visual reasoning

Algebraically manipulating words and visual objects to answer a new question [Bottou, 2014]

A1. Is the **tray** on top of the **table** black or light brown? light brown
A2. Are the **napkin** and the **cup** the same color? yes
A3. Is the small **table** both oval and wooden? yes
A4. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
A5. Are there any **cups** to the left of the **tray** on top of the **table**? no
B1. What is the brown **animal** sitting inside of? **box**
B2. What is the large **container** made of? cardboard
B3. What **animal** is in the **box**? **bear**
B4. Is there a **bag** to the right of the green **door**? no
B5. Is there a **box** inside the plastic **bag**? no

(a) GQA [Hudson and Manning, 2019]



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

(b) CLEVR [Johnson et al., 2017]

Figure: Using Visual Question Answering (**VQA**) to evaluate reasoning skills.

Biases and Reasoning

Reasoning *vs.* **shortcut learning**

"decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions" [Geirhos et al., 2020]

What is the person holding?

Answer: Paper
Pred: Banana.

Also known as: biases, educated guesses, *etc*...

Biases and Reasoning

Reasoning *vs.* **shortcut learning**

"decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions" [Geirhos et al., 2020]

What is the person holding?

Answer: Paper
Pred: Banana.

Also known as: biases, educated guesses, *etc*...

GQA-OOD: a VQA benchmark for OOD settings

Roses are Red, Violets are Blue...
But Should VQA expect Them To? (CVPR'21)

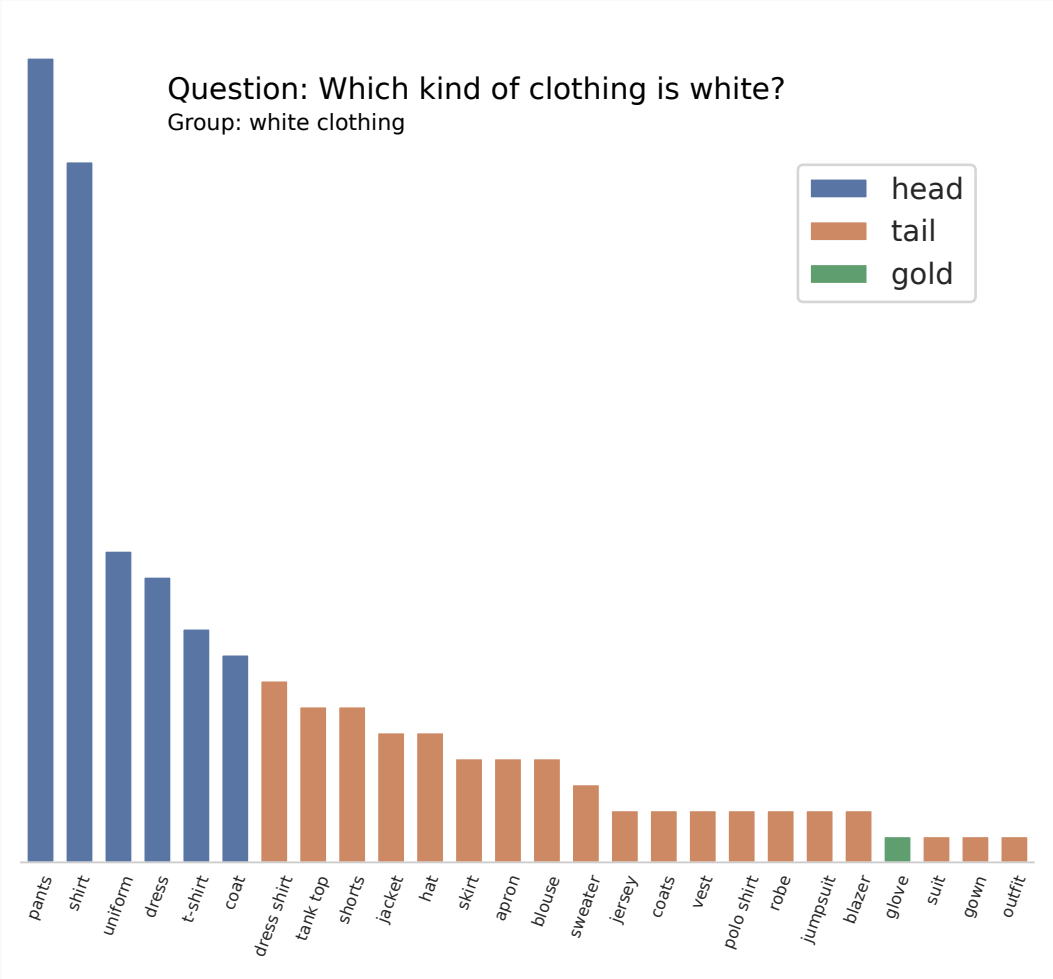
Corentin Kervadec, Grigory Antipov, Moez Baccouche and Christian Wolf.




GQA-OOD: Measuring biases in VQA [Kervadec et al., 2021]

In VQA, questions and concepts are naturally unbalanced.

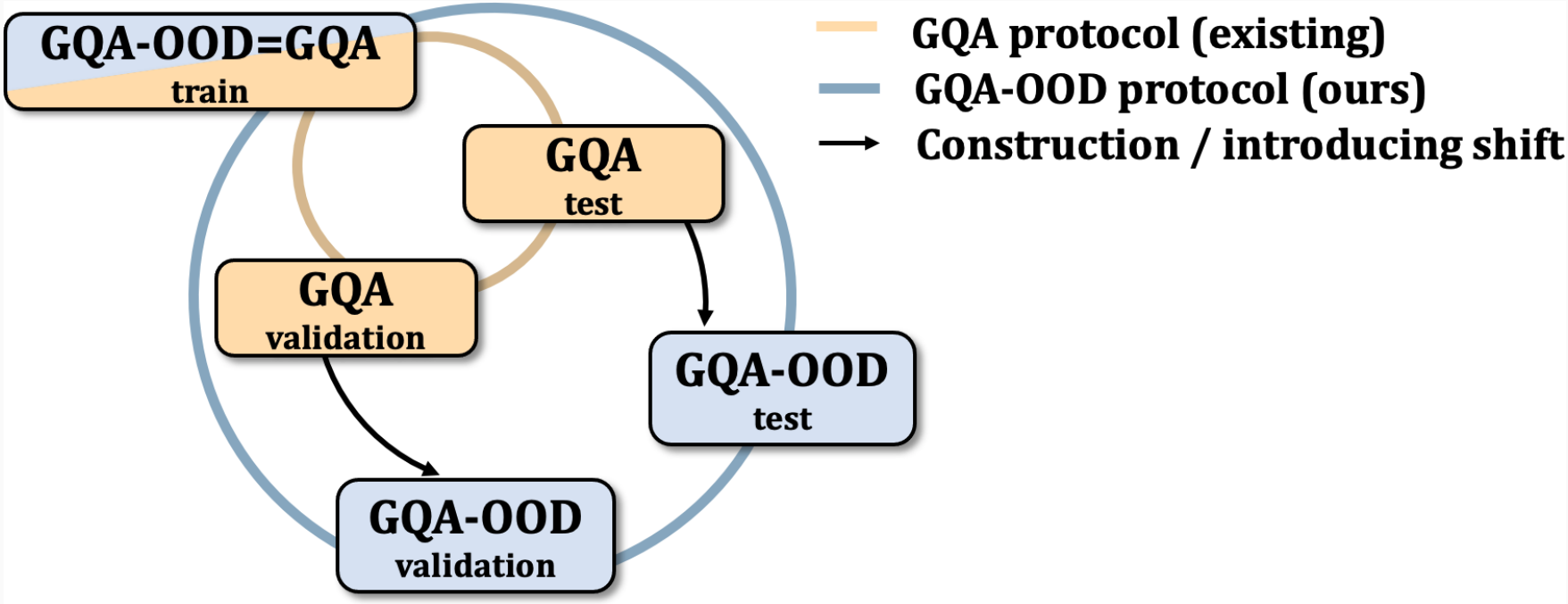
$\rightarrow$ many biases



GQA-OOD: a VQA benchmark for OOD settings

GQA-OOD (Out-Of-Distribution)

We measure and compare accuracy over both rare and frequent question-answer pairs



GQA-OOD: a VQA benchmark for OOD settings

"What is on the wall?"

Group: objects on walls

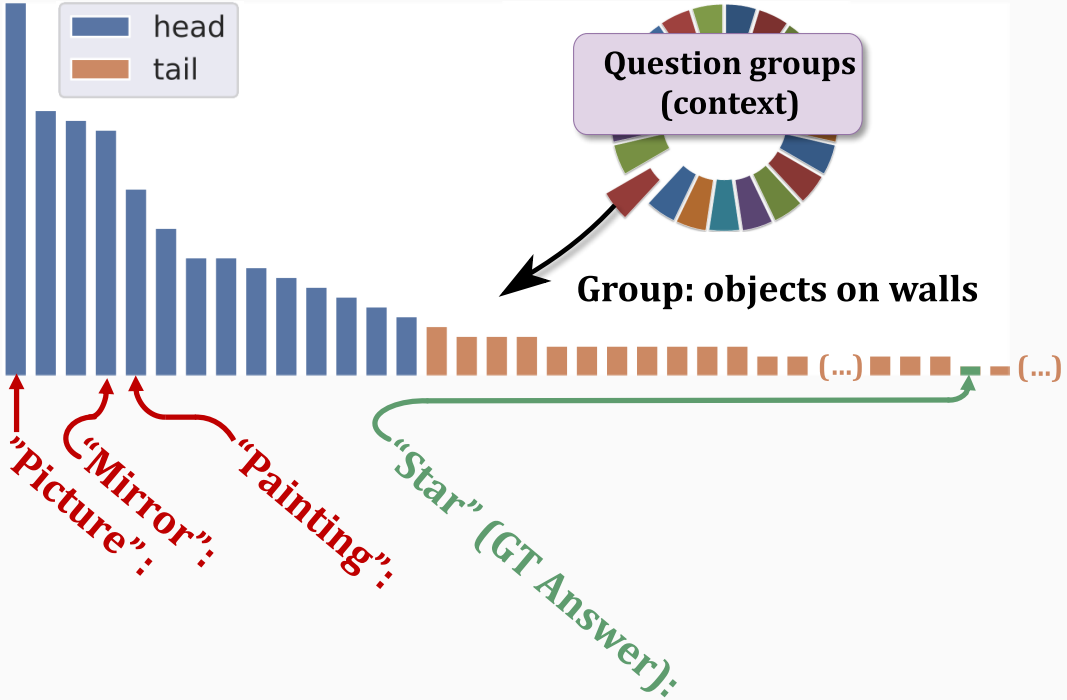
We come up with three metrics for test and validation:

- **acc-all**: all samples
- **acc-tail**: samples with **rare** answer given the question's group
- **acc-head**: samples with **frequent** answer given the question's group

In the main paper, we evaluated the validity of these metrics.

GQA-OOD: a VQA benchmark for OOD settings

“What is on the wall?”

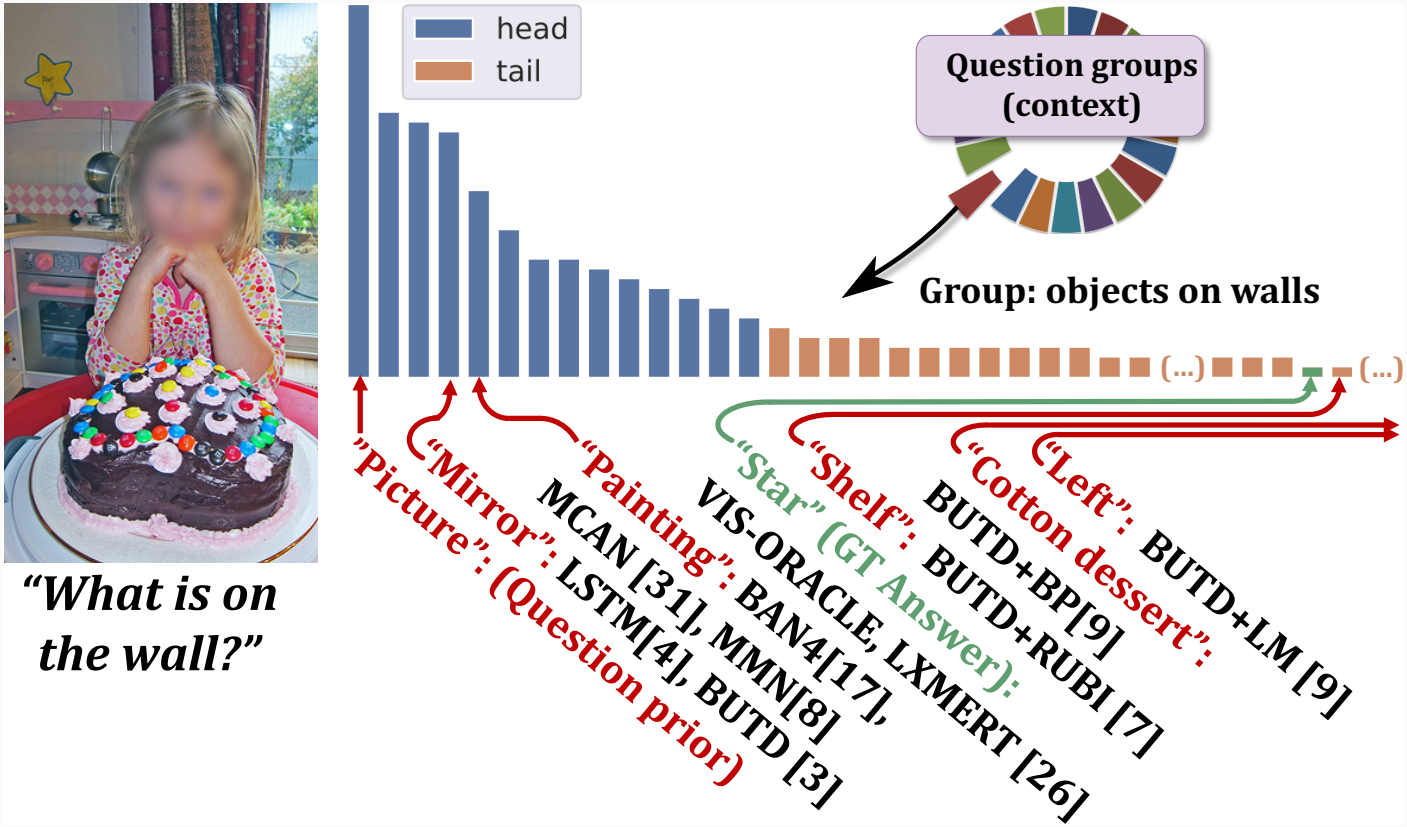


We come up with three metrics for test and validation:

- **acc-all**: all samples
- **acc-tail**: samples with **rare** answer given the question’s group
- **acc-head**: samples with **frequent** answer given the question’s group

In the main paper, we evaluated the validity of these metrics.

GQA-OOD: a VQA benchmark for OOD settings

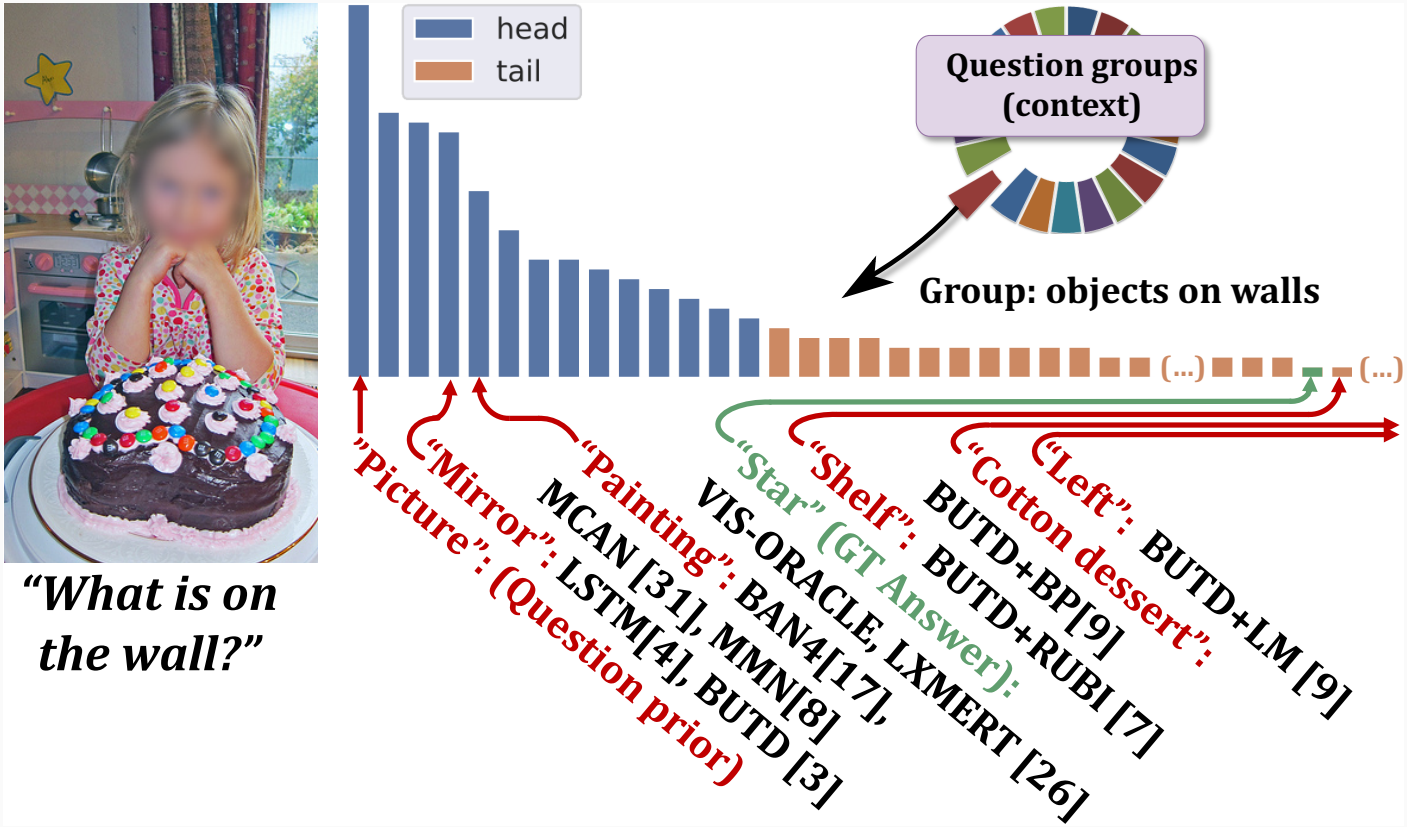


We come up with three metrics for test and validation:

- **acc-all**: all samples
- **acc-tail**: samples with **rare** answer given the question's group
- **acc-head**: samples with **frequent** answer given the question's group

In the main paper, we evaluated the validity of these metrics.

GQA-OOD: a VQA benchmark for OOD settings



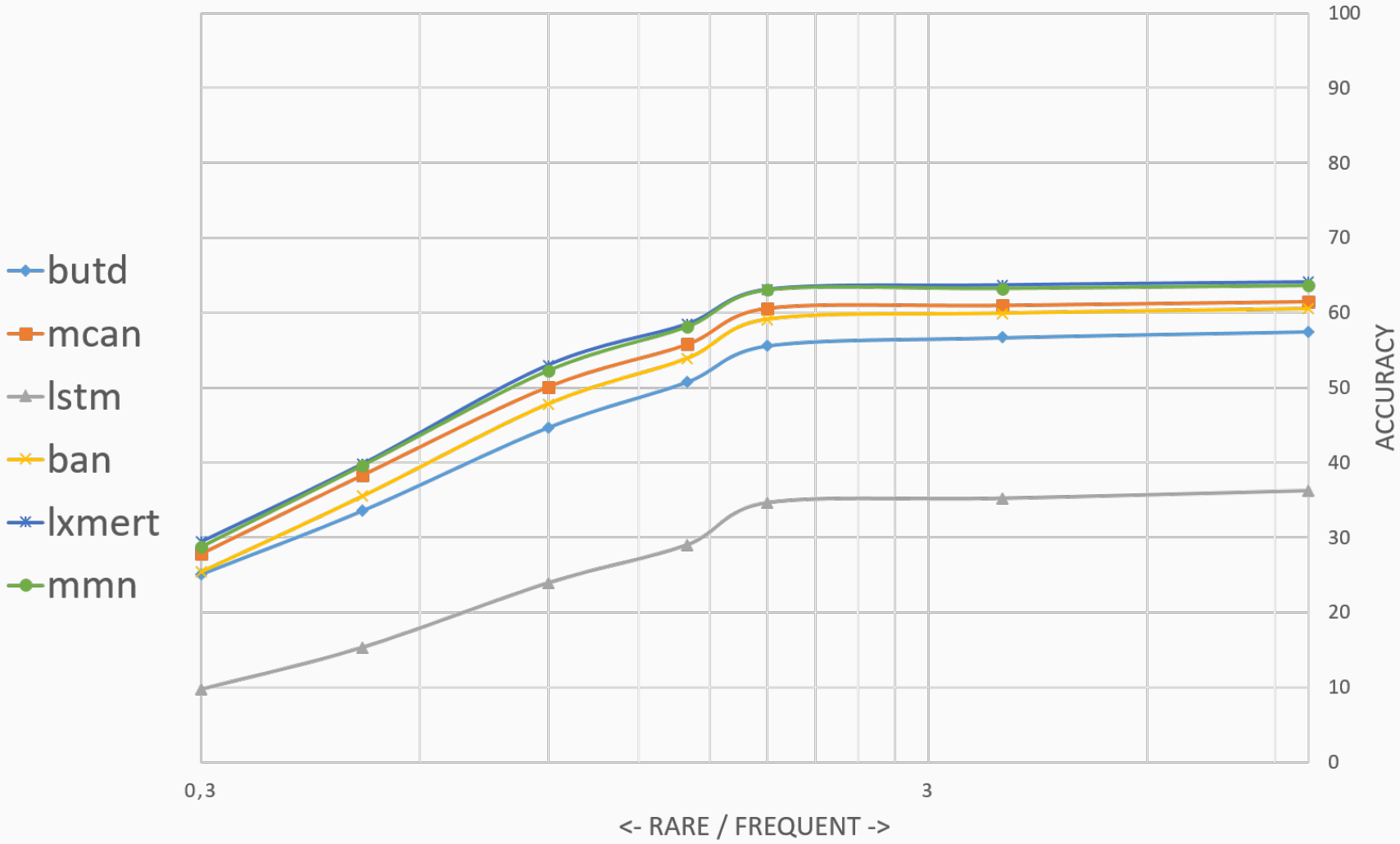
We come up with three metrics for test and validation:

- **acc-all**: all samples
- **acc-tail**: samples with **rare** answer given the question's group
- **acc-head**: samples with **frequent** answer given the question's group

In the main paper, we evaluated the validity of these metrics.

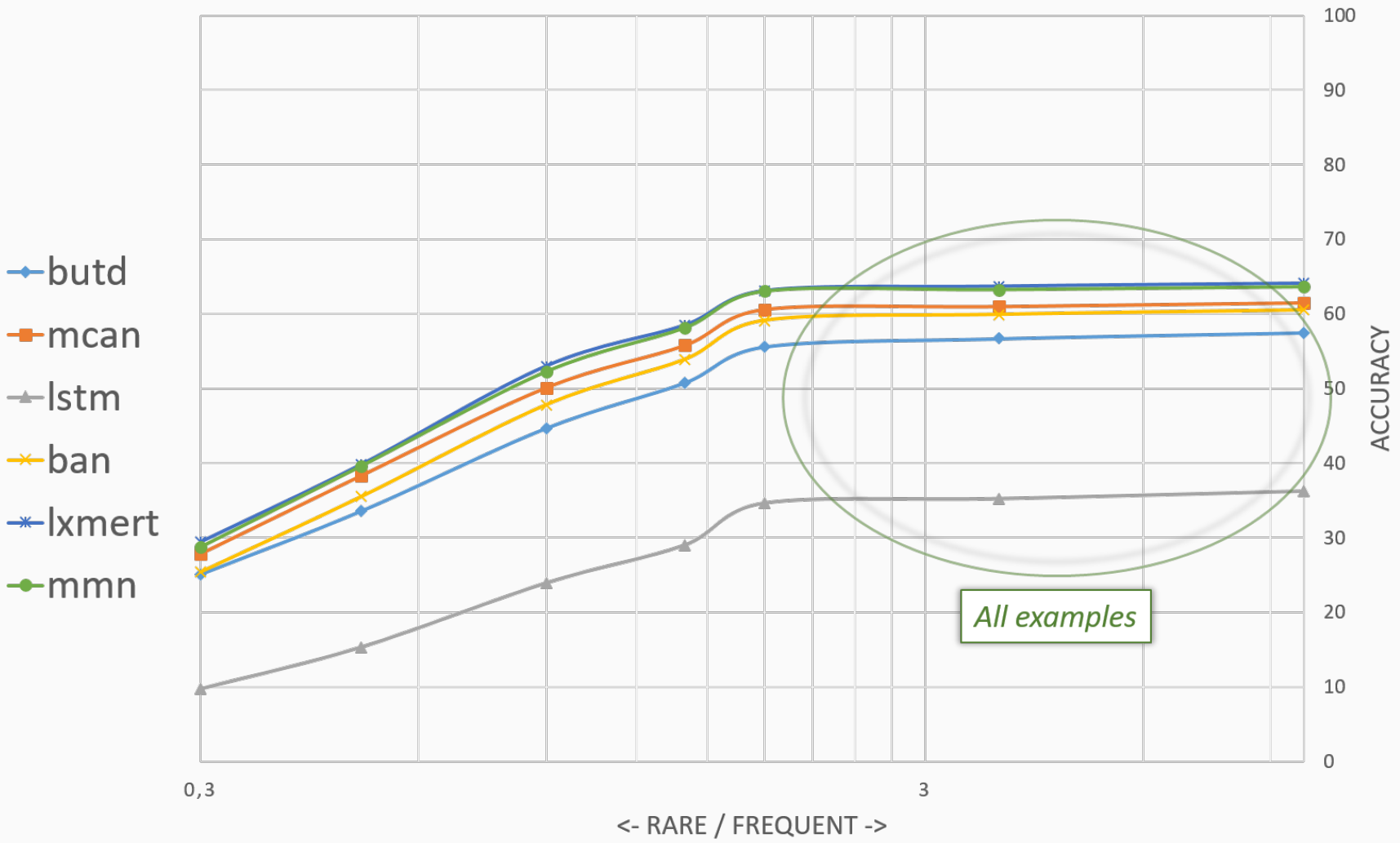
Experiments: SOTA VQA models

We plot accuracy (*y-axis*) versus the question-answer pairs rareness (*x-axis*):



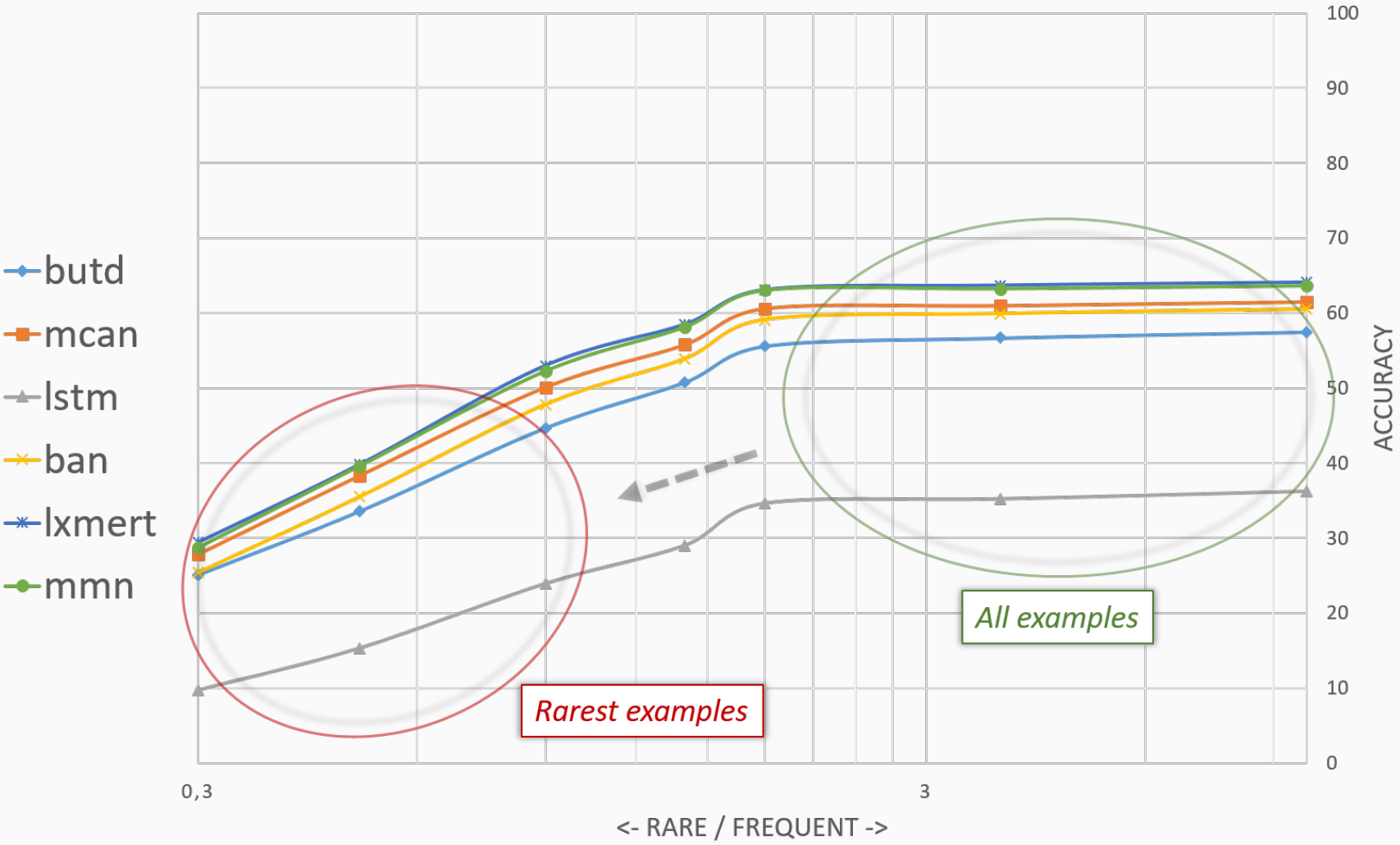
Experiments: SOTA VQA models

We plot accuracy (*y-axis*) versus the question-answer pairs rareness (*x-axis*):



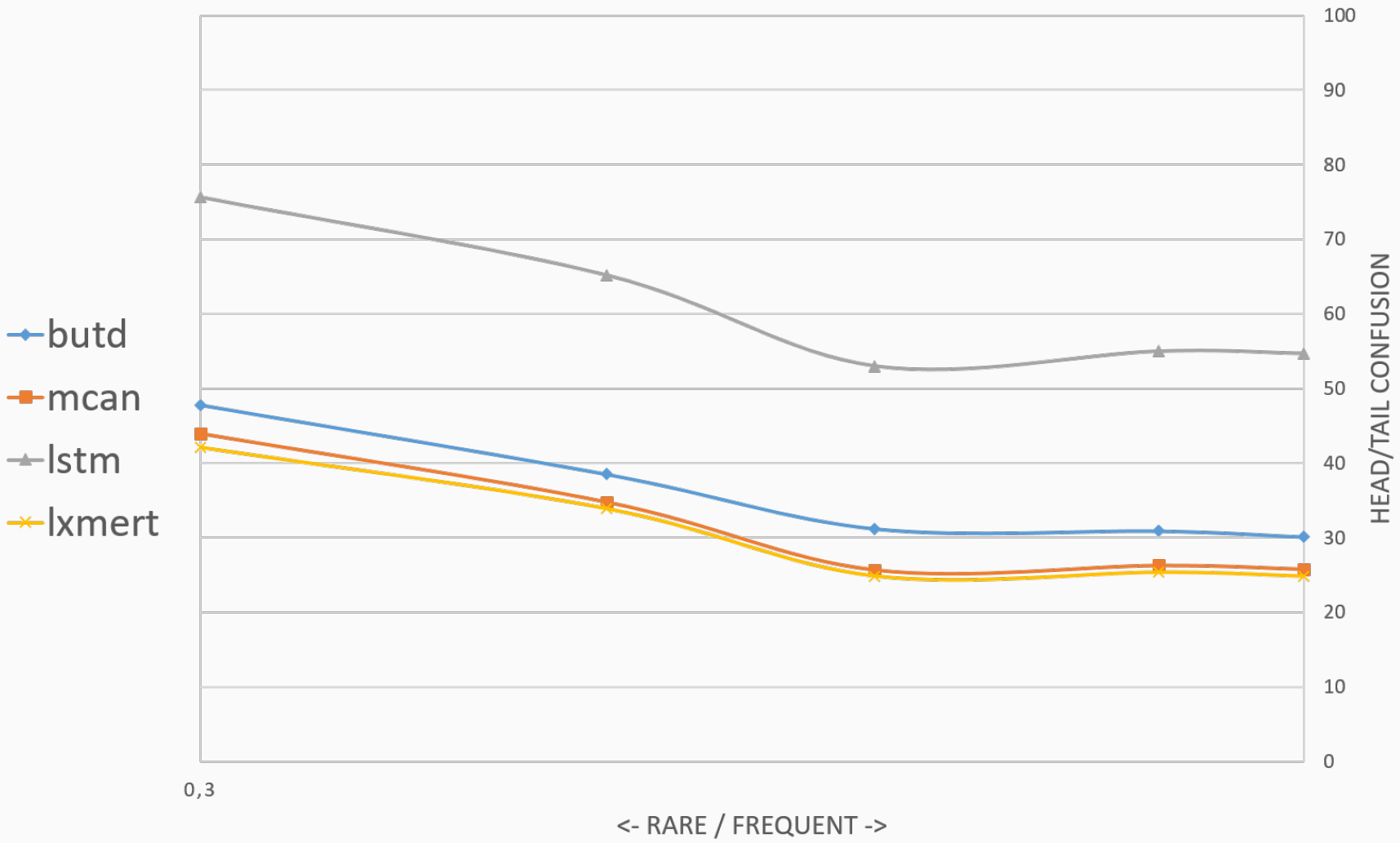
Experiments: SOTA VQA models

We plot accuracy (*y-axis*) versus the question-answer pairs rareness (*x-axis*):



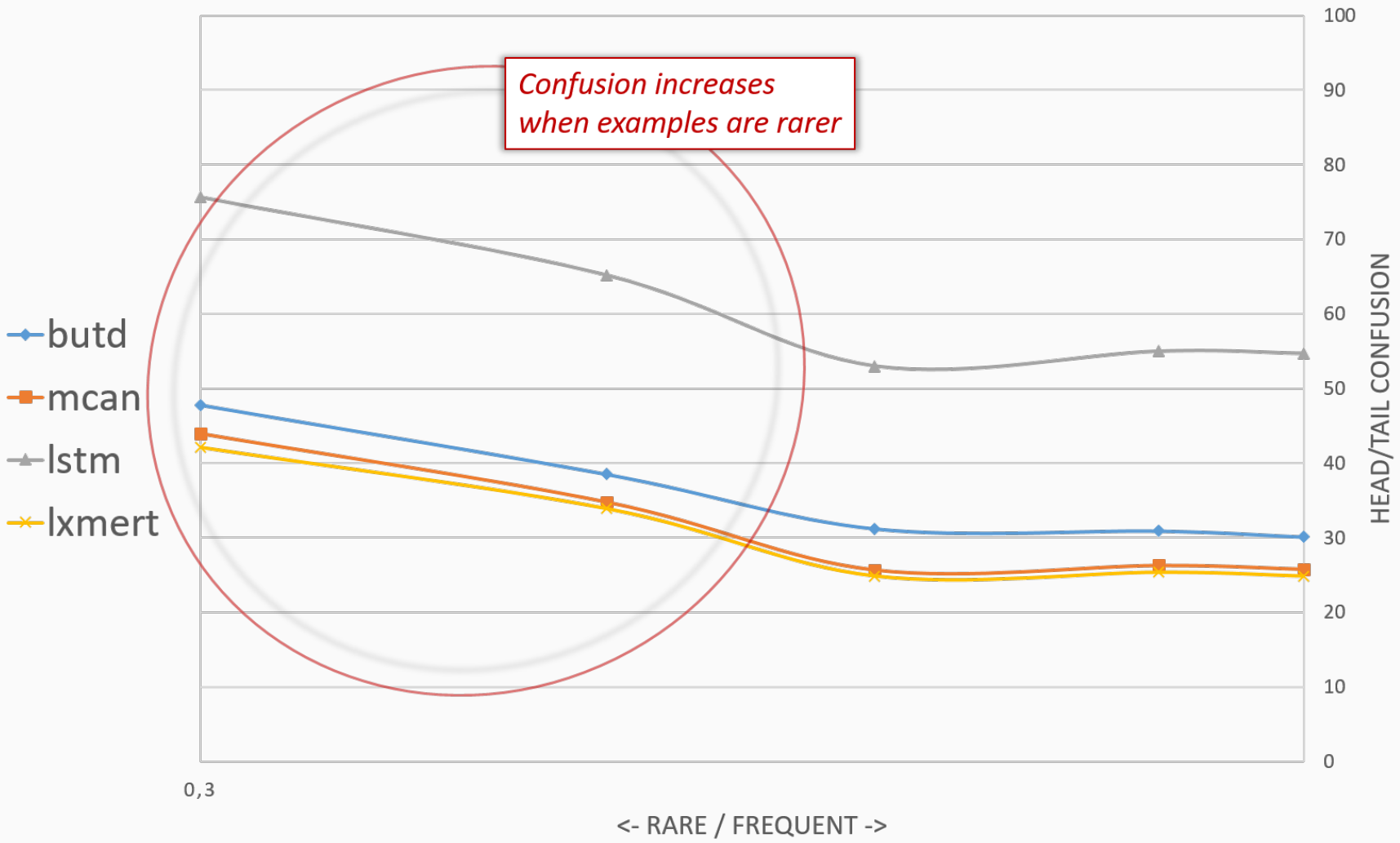
Experiments: SOTA VQA models

head/tail confusion (when the model predicts a frequent instead of rare answer):



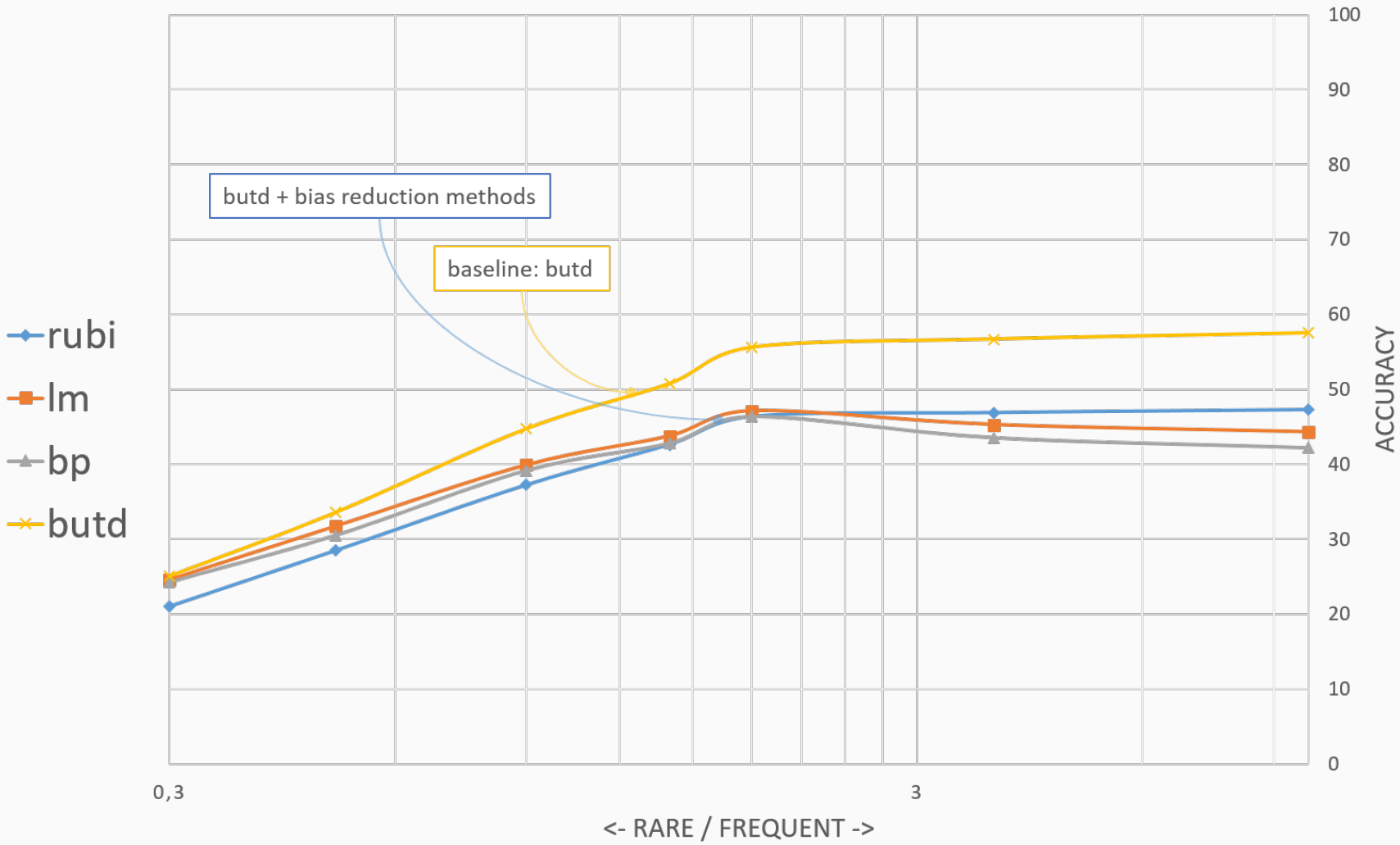
Experiments: SOTA VQA models

head/tail confusion (when the model predicts a frequent instead of rare answer):



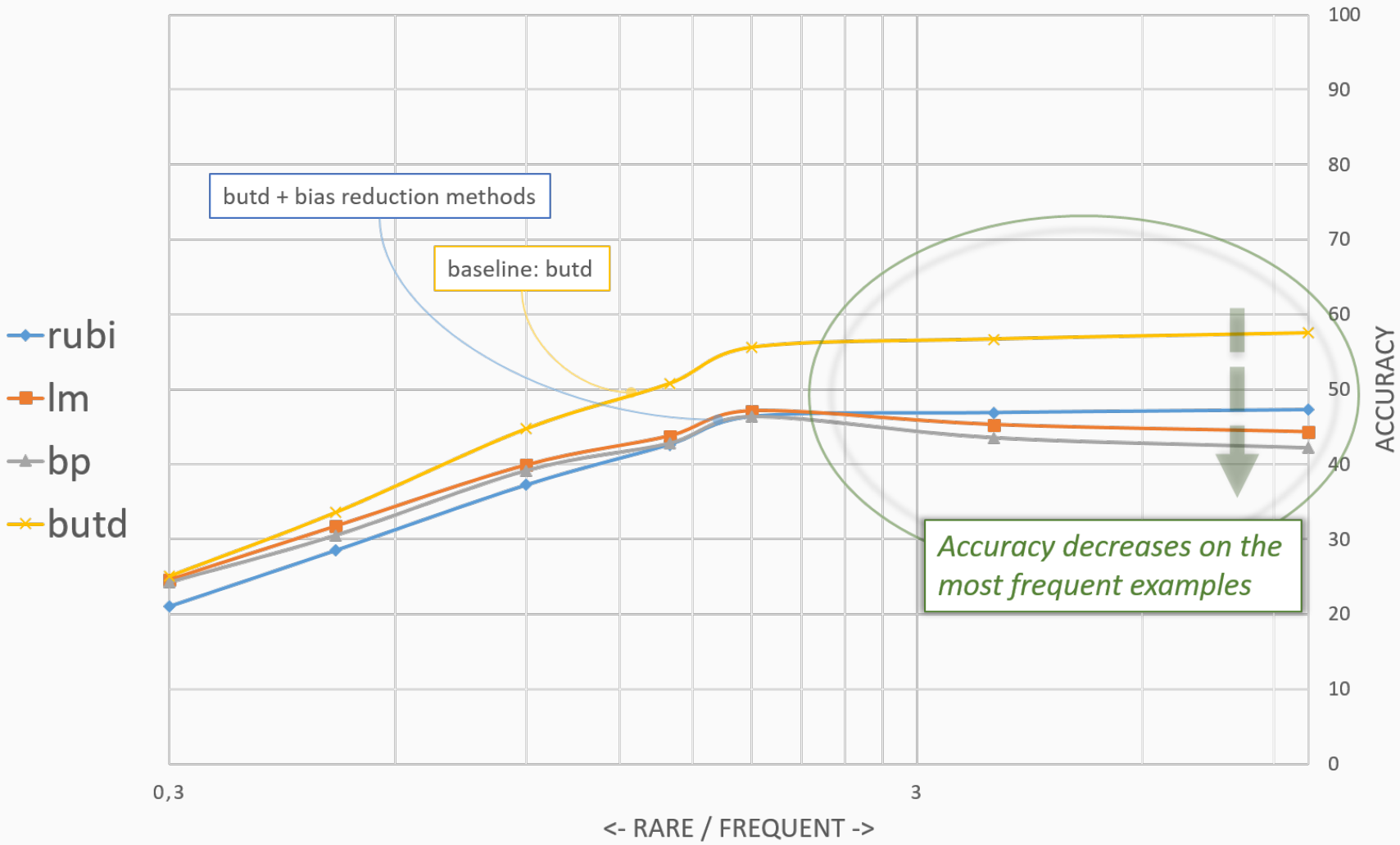
Experiments: bias reduction techniques

Bias-reduction methods also fail in this setup:



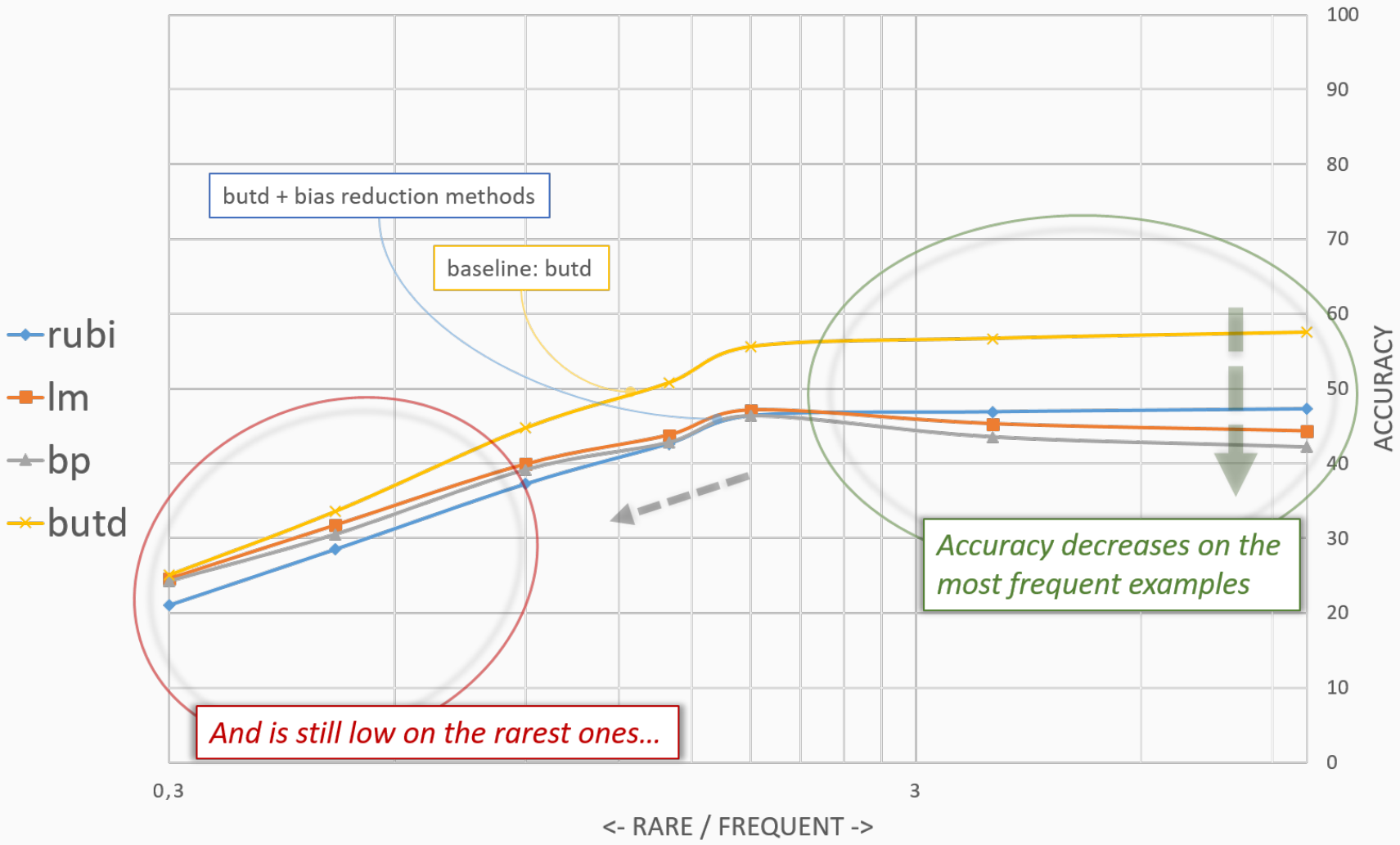
Experiments: bias reduction techniques

Bias-reduction methods also fail in this setup:



Experiments: bias reduction techniques

Bias-reduction methods also fail in this setup:



How Transferable are Reasoning Patterns in VQA? (CVPR'21)

C. Kervadec, T. Jaunet, G. Antipov, M. Baccouche, R. Vuillemot and C. Wolf






Motivation and related works

Reasoning patterns in Transformers

Analysing self-attention mechanisms

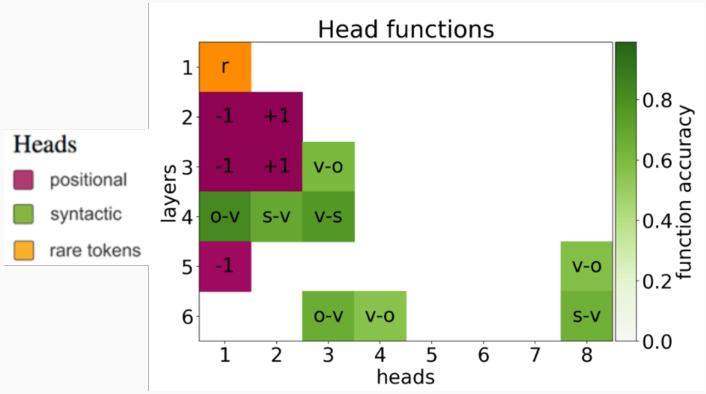


Figure: [Voita et al., 2019]

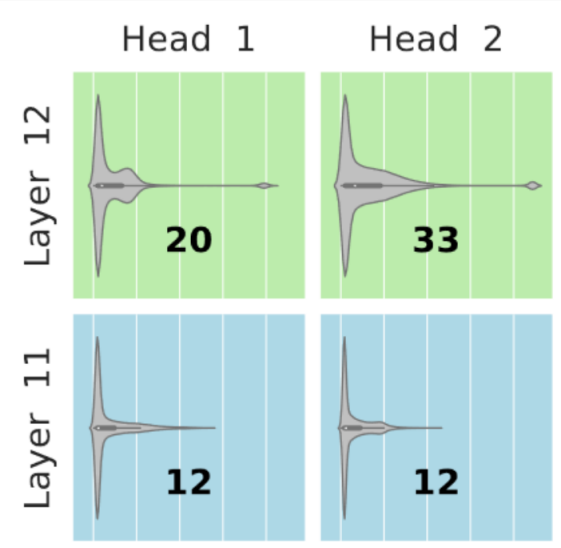
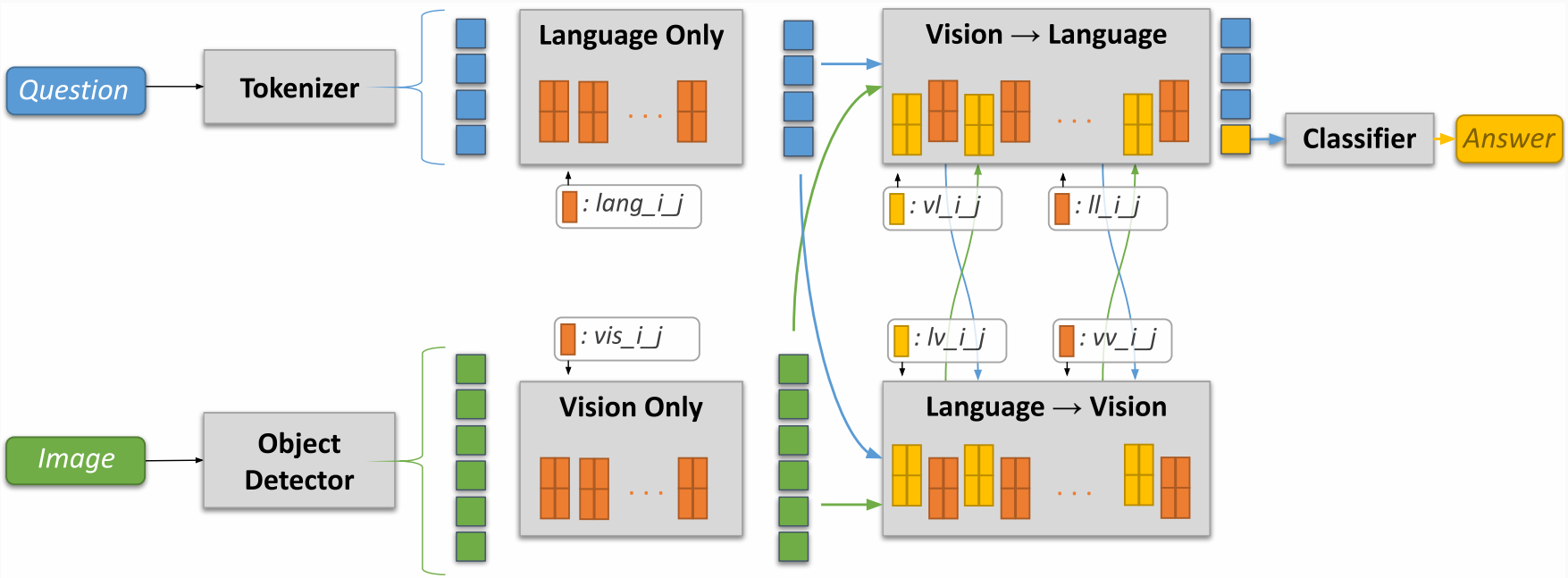


Figure: [Ramsauer et al., 2020]

VL-Transformer

Vision-Langage (VL)-Transformer [Tan and Bansal, 2019]

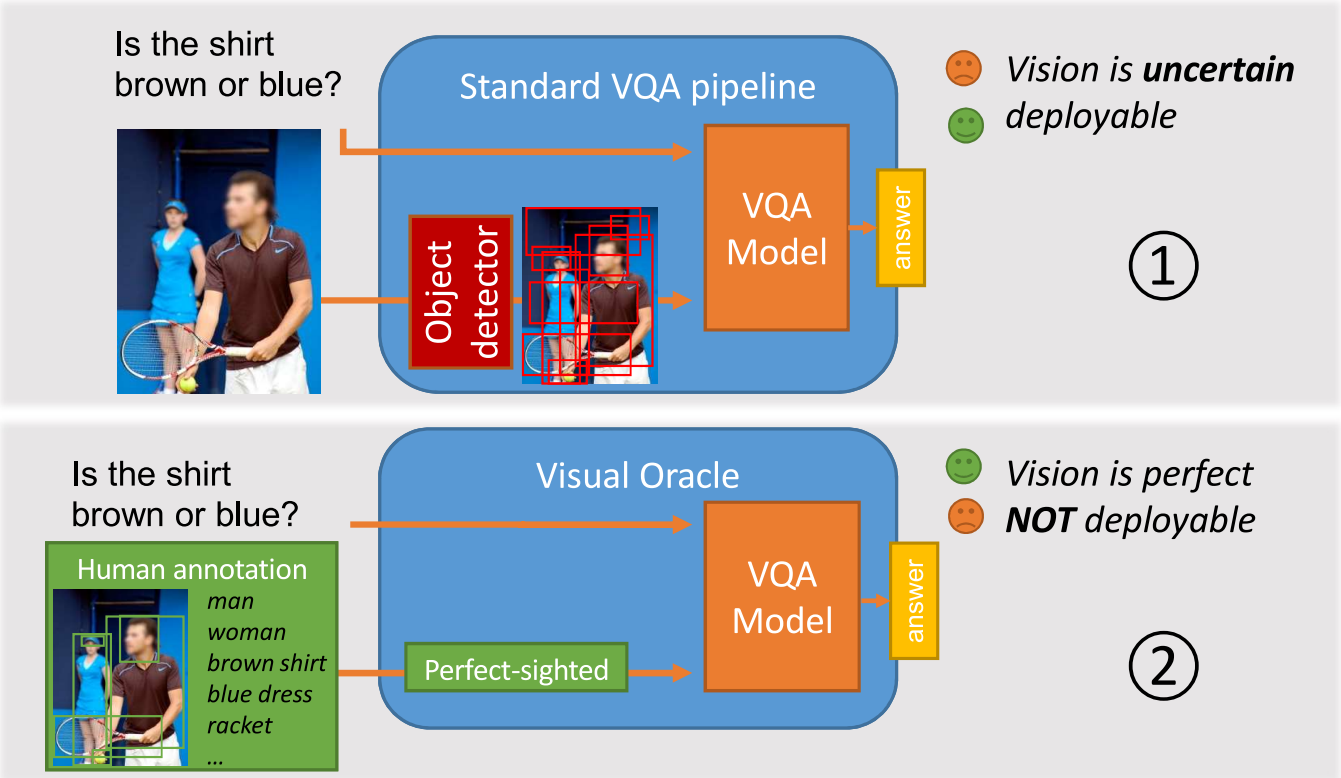
- **input:** visual objects and question words
- **output:** answer prediction
- Use **uni-modal** and **cross-modal** Transformer layers



Hypothesis

Visual bottleneck

Shortcut learning is in part caused by the visual uncertainty



① Standard VQA model with imperfect vision
② Oracle model with perfect sight.

Reasoning Patterns

Visual oracle is less prone to learn shortcuts:

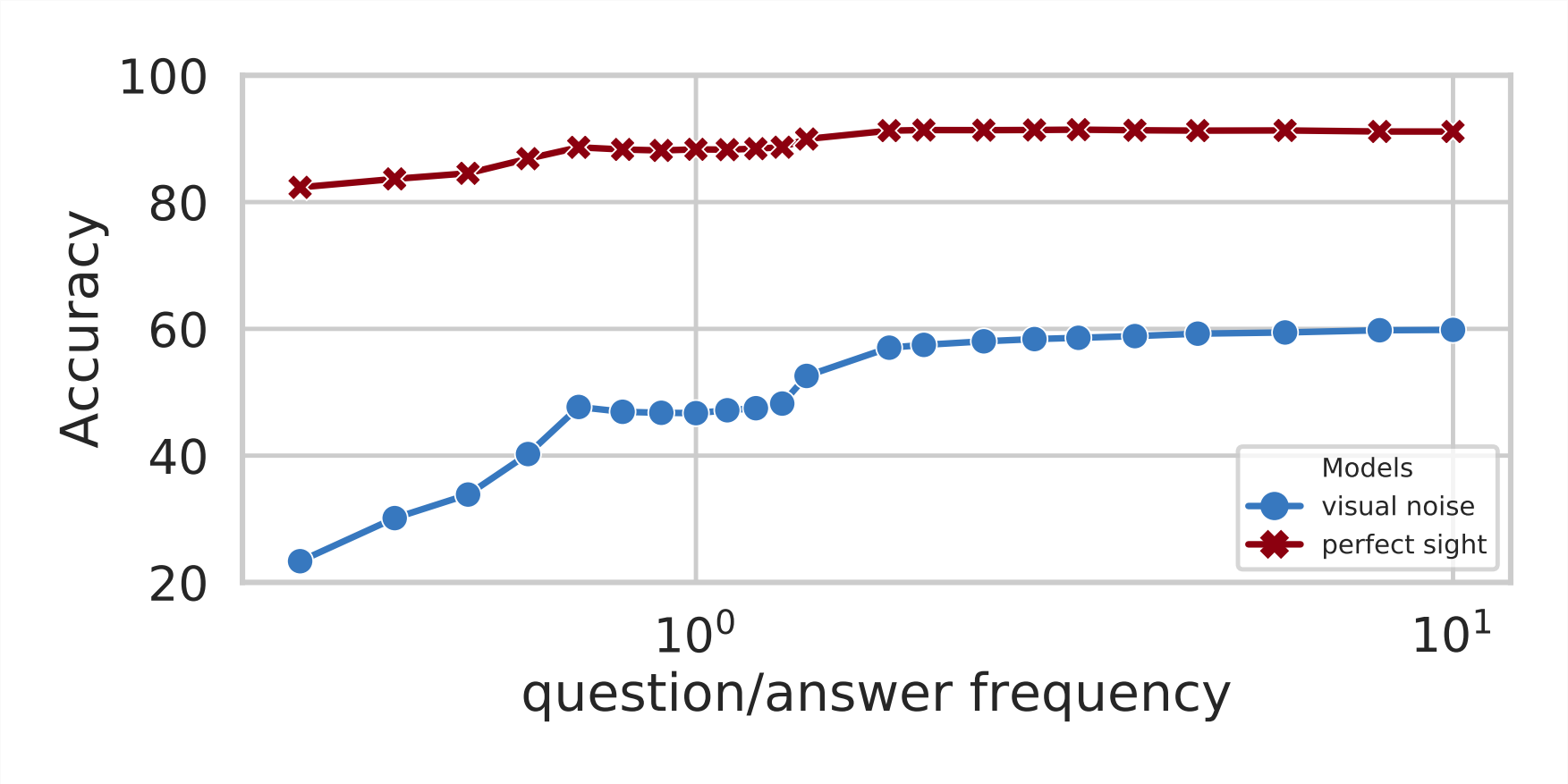
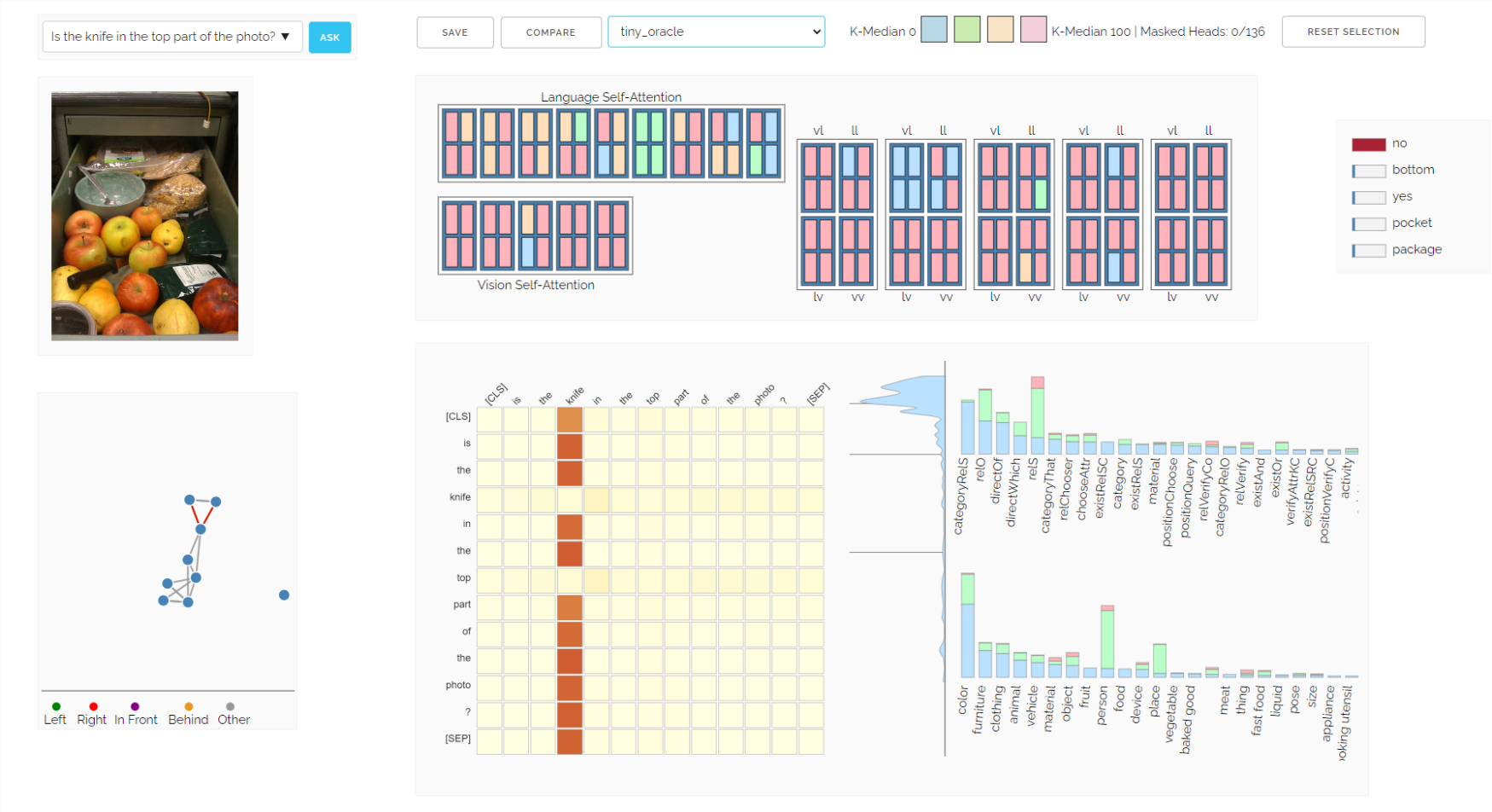


Figure: Comparison of the out-of-distribution generalization: a perfectly-sighted oracle model vs. a standard noisy vision based model (GQA-OOD benchmark [Kervadec et al., 2021]).

Reasoning Patterns

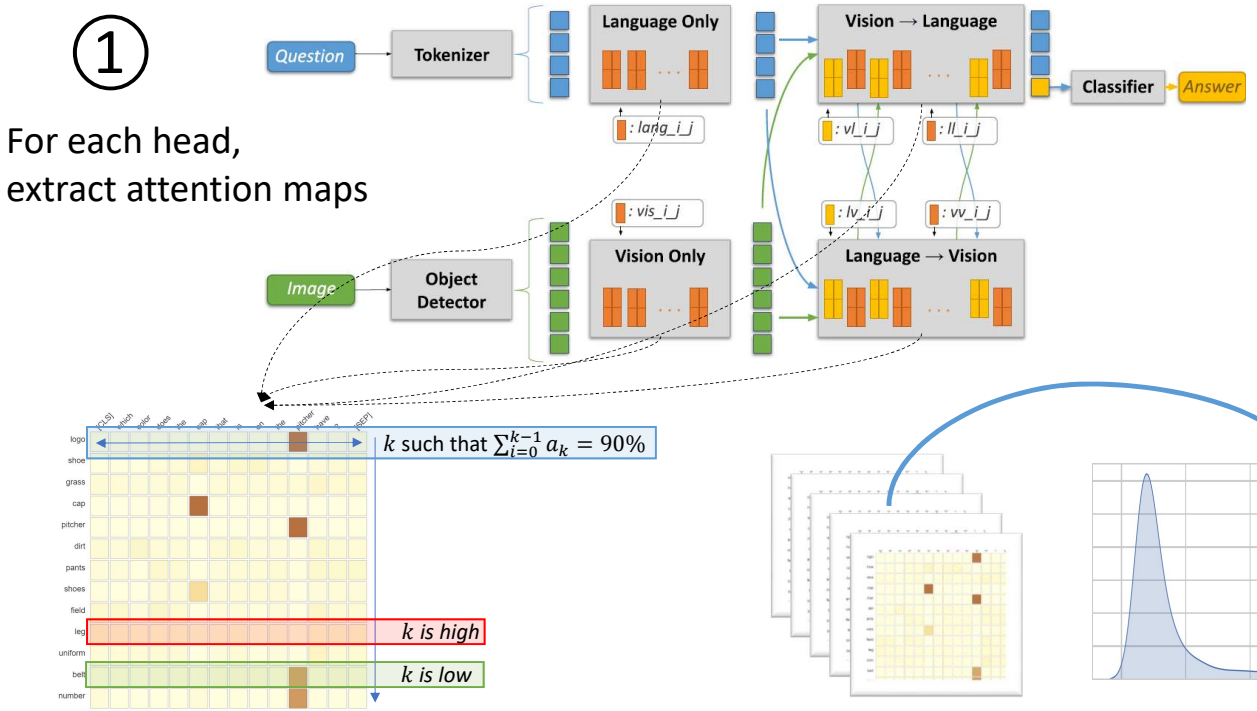
Interactive tool

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



Analysis of Reasoning Patterns in Attention Heads

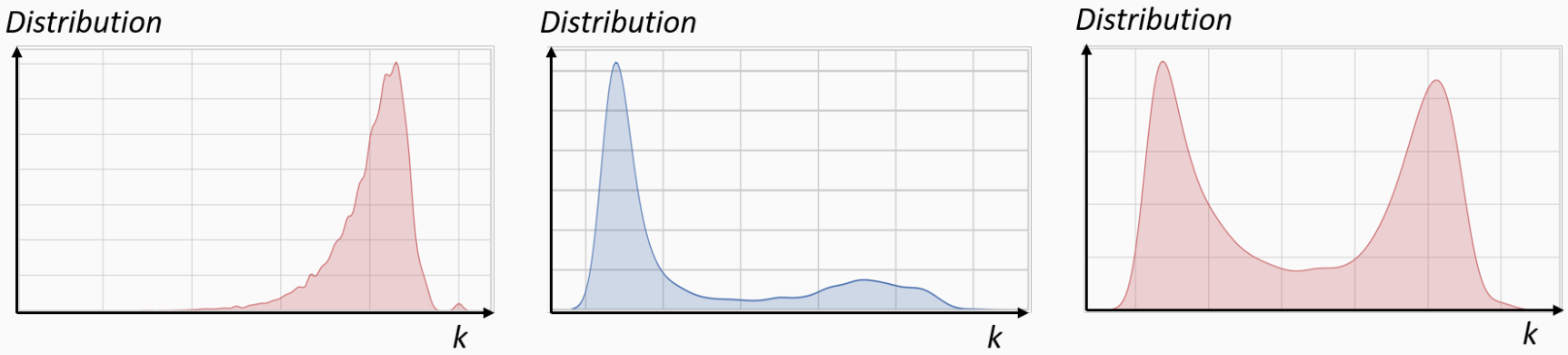
Measuring attention modes in attention heads



Analysis of Reasoning Patterns in Attention Heads

Measuring attention modes

We identify three attention modes learned by the Oracle: *bimorph*, *dirac* and *uniform*



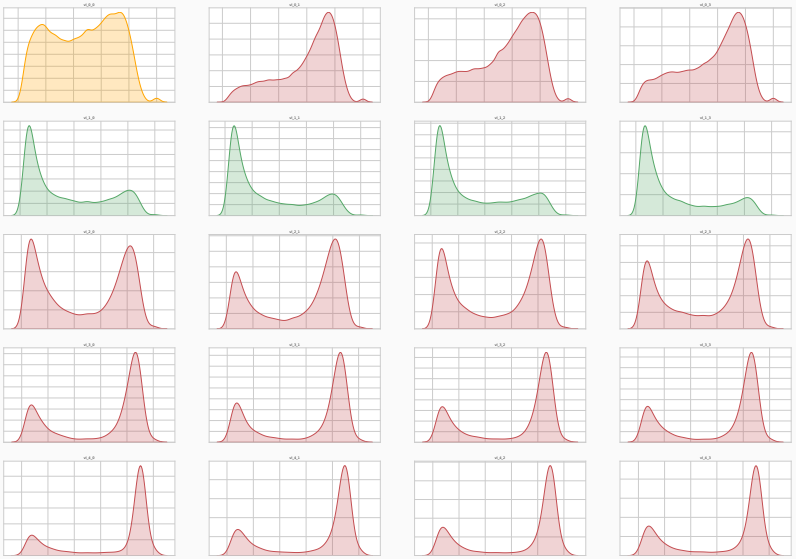
(a) Uniform (b) Dirac (c) Bimorph

Analysis of Reasoning Patterns in Attention Heads

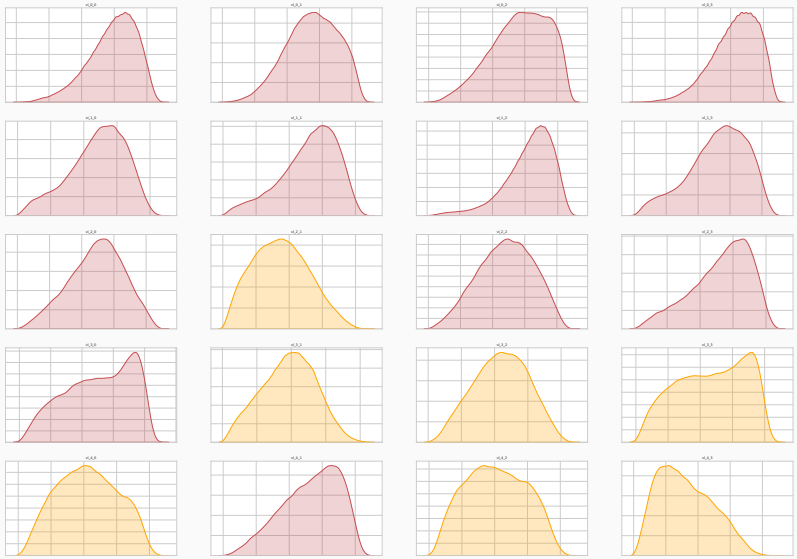
Attention modes: oracle vs. standard VQA

Measuring attention modes of vision-to-language attention heads:

- Higher *diversity* in visual oracle



(a) Oracle

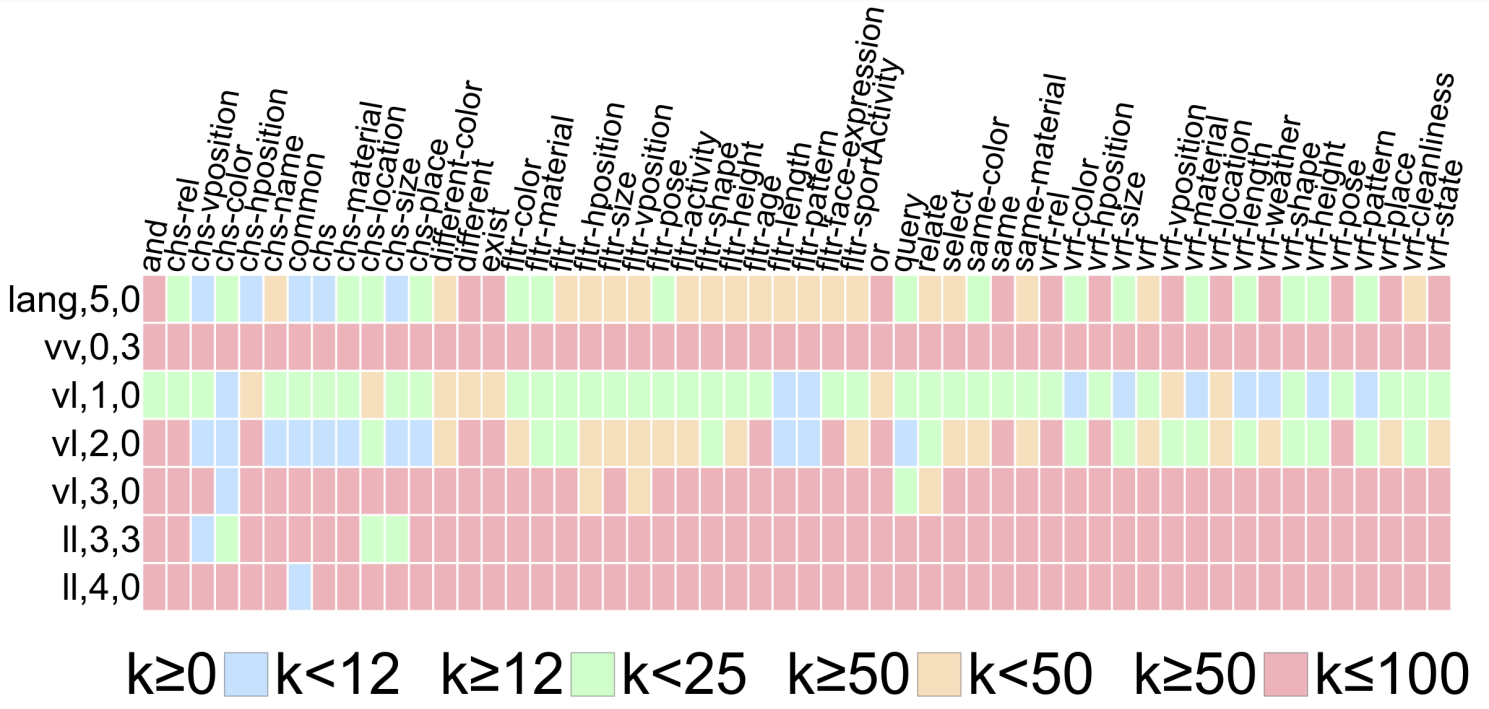


(b) Standard VQA

Analysis of Reasoning Patterns in Attention Heads

Attention modes vs. task functions

ex: `filter size, choose color, query name, relate, verify material`, *etc*...



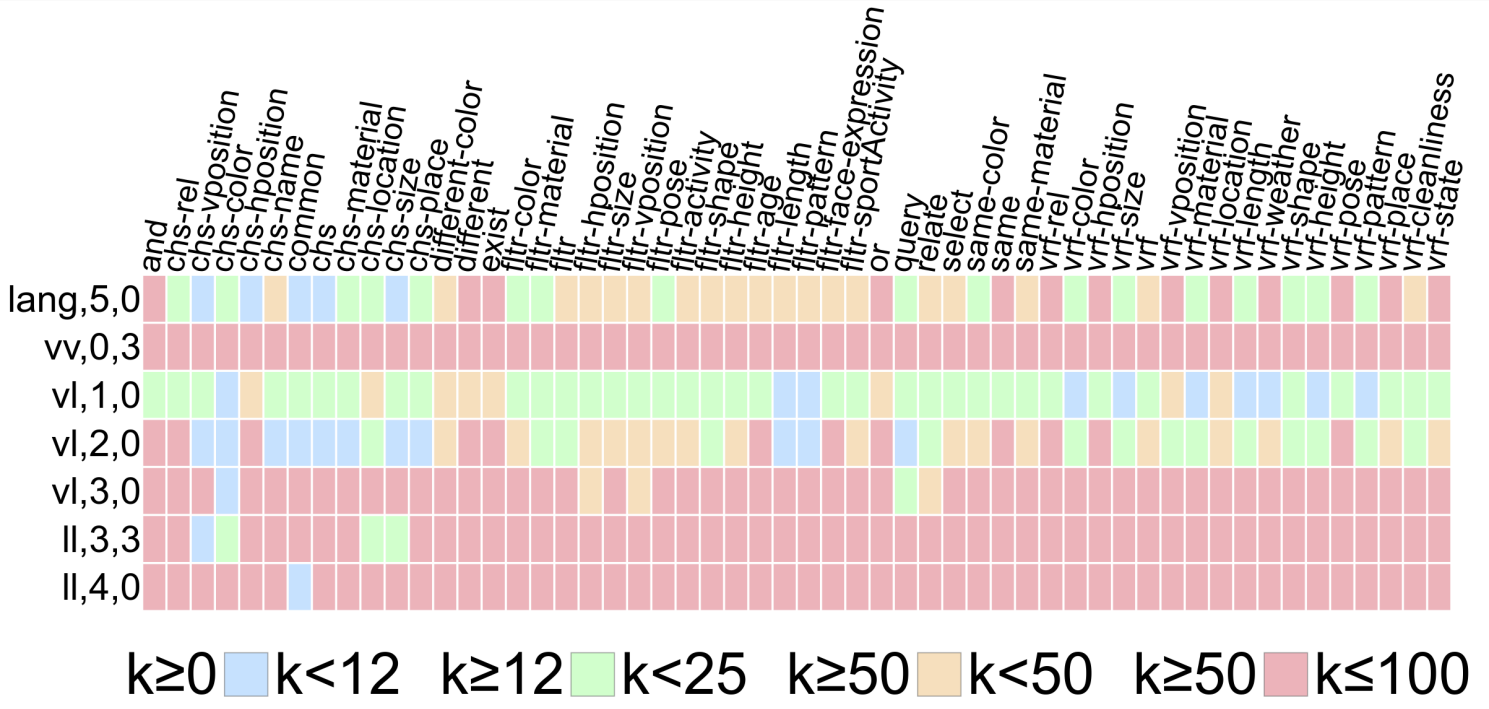
Oracle

- Attention heads behave differently depending on the function
- A given function causes different attention modes for different heads

Analysis of Reasoning Patterns in Attention Heads

Attention modes vs. task functions

ex: `filter size, choose color, query name, relate, verify material`, *etc*...



Oracle

- Attention heads behave differently depending on the function
- A given function causes different attention modes for different heads

Analysis of Reasoning Patterns in Attention Heads

Comparison: Impact of the function `choose color` on attention modes.

- Standard model: no clear relationships between attention modes and functions

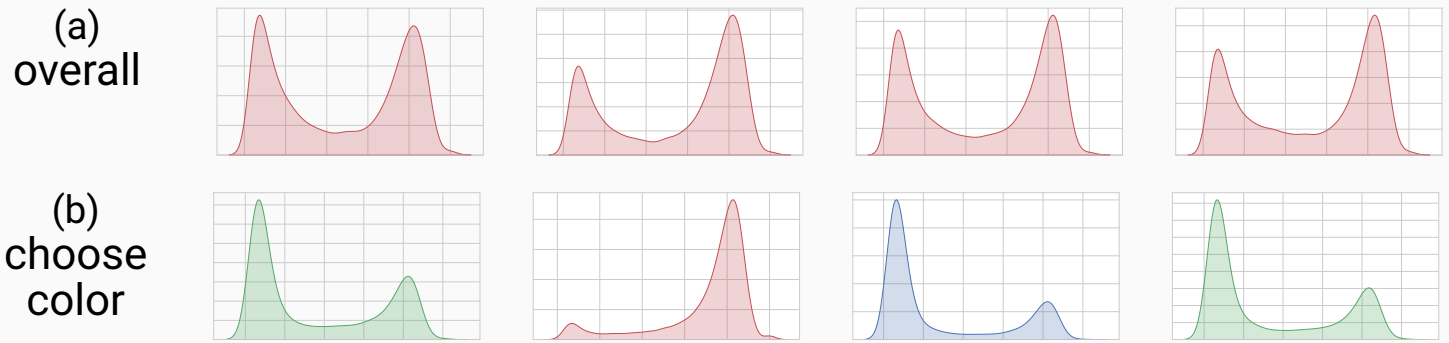


Figure: Oracle model

Analysis of Reasoning Patterns in Attention Heads

Comparison: Impact of the function `choose color` on attention modes.

- Standard model: no clear relationships between attention modes and functions

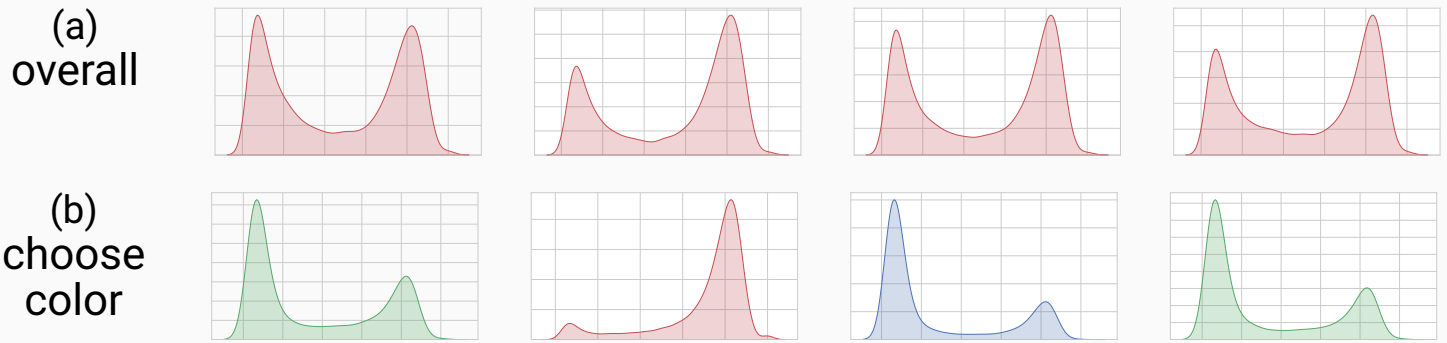


Figure: Oracle model

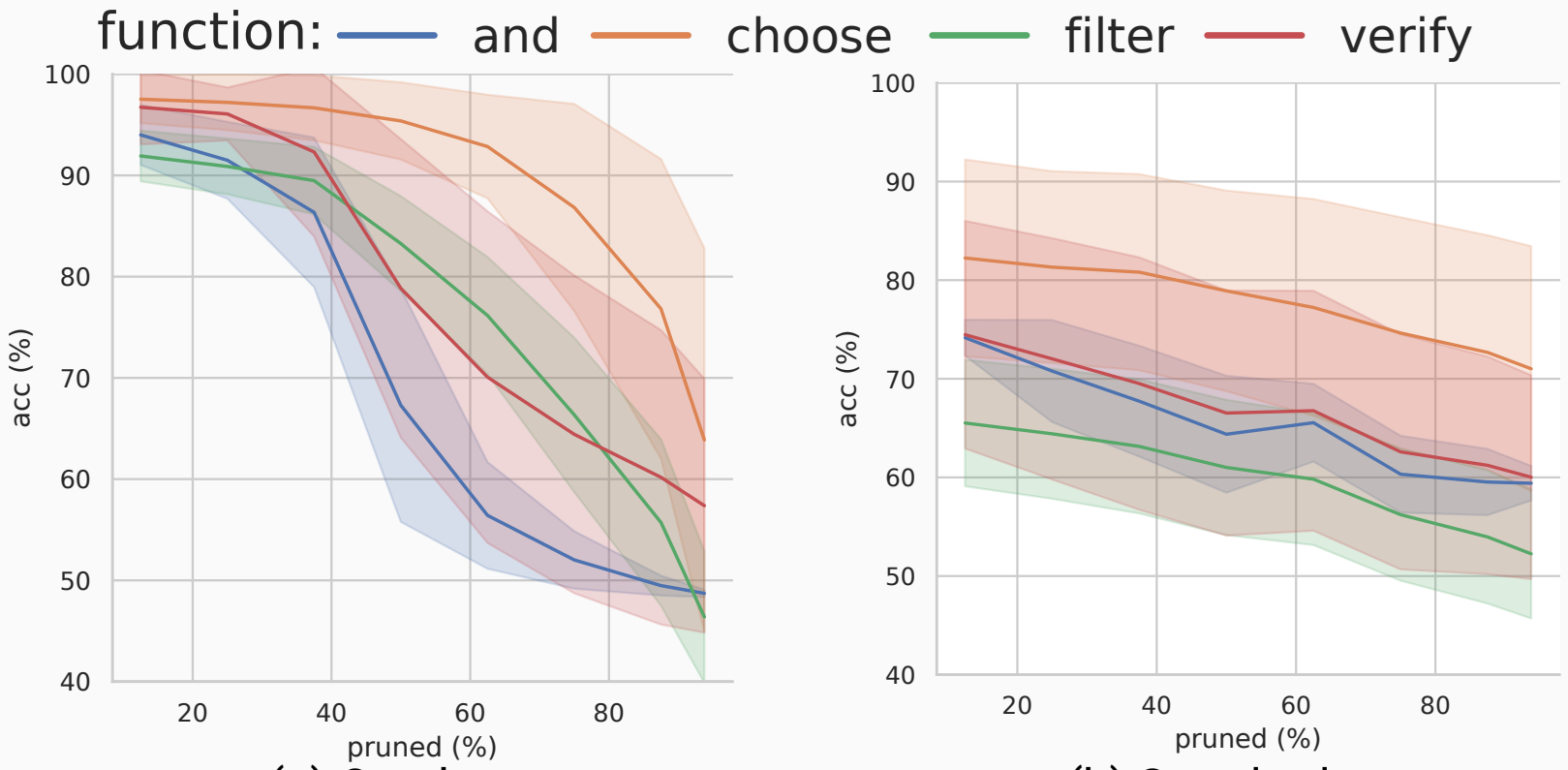
(a) overall (b) choose color

Figure: Standard VQA model

Analysis of Reasoning Patterns in Attention Heads

Head pruning: randomly removing (replace by average) cross-modal heads

- **Oracle**: impact related to the nature of the function, highlights a modular property
- **Standard**: pruning seems to be unrelated to function types



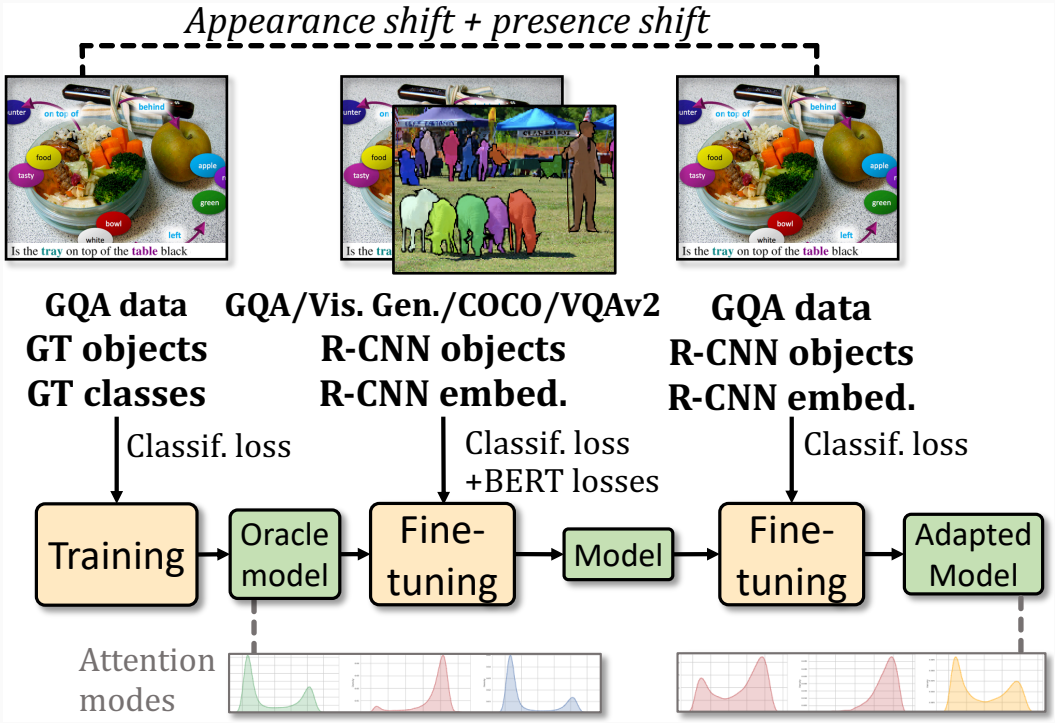
(a) Oracle

(b) Standard

Oracle transfer

Oracle transfer

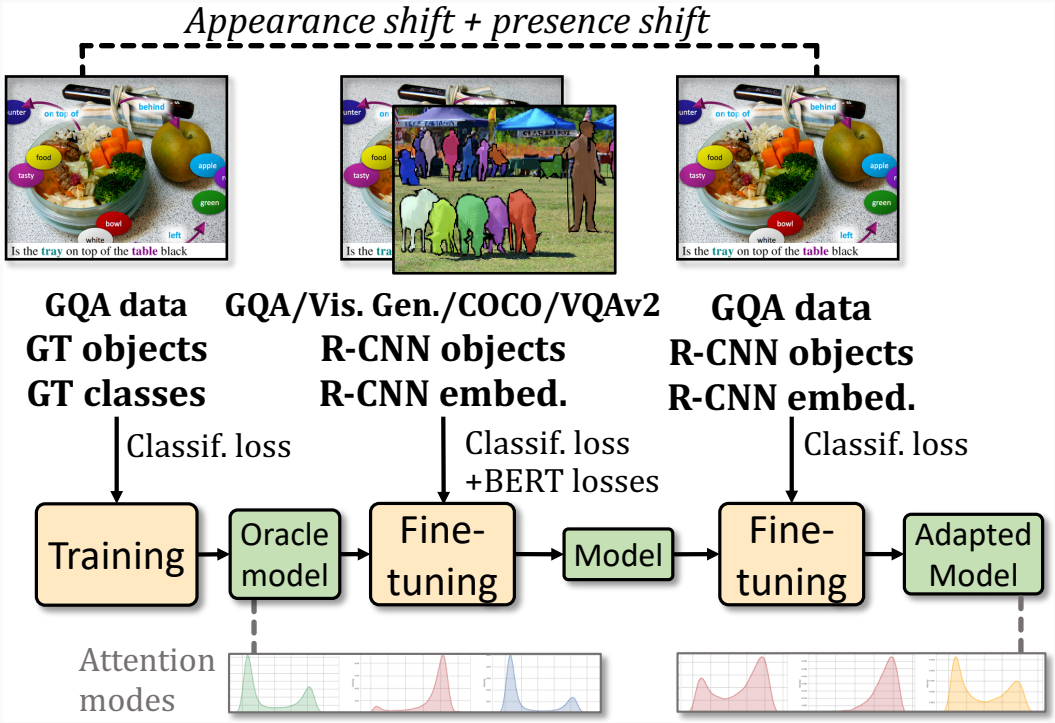
Using oracle and standard data:
(1) train the visual oracle;
(2) optionally, BERT-like pretraining;
(3) finetune on target dataset.



Oracle transfer

Oracle transfer

Using oracle and standard data:
(1) train the visual oracle;
(2) optionally, BERT-like pretraining;
(3) finetune on target dataset.



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

Conclusion & Discussion

Contributions

* *GQA-OOD*: a benchmark for better evaluating biases in VQA
* A deep analysis of several aspects of VQA models linked to reasoning
* An *oracle transfer* method to reduce biases

Limitations

* Limited to the (partially) synthetic GQA [Hudson and Manning, 2019] dataset
* The *oracle transfer* could be more efficient

Future work

* Extending OOD analysis to more natural settings
* Improving the *oracle transfer* with program prediction

Thanks

Thanks!
Any questions?

Roses are Red, Violets are Blue... But Should VQA expect Them To?
C. Kervadec, G. Antipov, M. Baccouche, C. Wolf @ *CVPR2021*

How Transferable are Reasoning Patterns in VQA?
C. Kervadec, T. Jaunet, G. Antipov, M. Baccouche, R. Vuillemot, C. Wolf @ *CVPR2021*

More at `https://corentinkervadec.github.io/`
Twitter: `https://twitter.com/CorentK`

Bibliography I

[Bottou, 2014] Bottou, L. (2014).
From machine learning to machine reasoning.
Machine learning, 94(2):133–149.

[Geirhos et al., 2020] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020).
Shortcut learning in deep neural networks.
Nature Machine Intelligence, 2(11):665–673.

[Hudson and Manning, 2019] Hudson, D. A. and Manning, C. D. (2019).
Gqa: A new dataset for real-world visual reasoning and compositional question answering.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

[Johnson et al., 2017] Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017).
Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Bibliography II

[Kervadec et al., 2021] Kervadec, C., Antipov, G., Baccouche, M., and Wolf, C. (2021).
Roses are red, violets are blue... but should vqa expect them to?
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[Ramsauer et al., 2020] Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., et al. (2020).
Hopfield networks is all you need.
arXiv preprint arXiv:2008.02217.

[Tan and Bansal, 2019] Tan, H. and Bansal, M. (2019).
Lxmert: Learning cross-modality encoder representations from transformers.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114.

[Voita et al., 2019] Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019).
Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.