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

C. Kervadec\textsuperscript{1,2}  
Joint work with: T. Jaunet\textsuperscript{2,3}  G. Antipov\textsuperscript{1}  M. Baccouche\textsuperscript{1}  R. Vuillemot\textsuperscript{2,4}  C. Wolf\textsuperscript{2,3}

\textsuperscript{1}Orange Innovation, France.  \textsuperscript{2,LIRIS,}  \textsuperscript{3,INSA,}  \textsuperscript{4,ECL,} Lyon, France.
Biases and Reasoning

Visual reasoning

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

(a) GQA [Hudson and Manning, 2019]
(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...
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...
Roses are Red, Violets are Blue...
But Should VQA expect Them To? (CVPR’21)

Corentin Kervadec, Grigory Antipov, Moez Baccouche and Christian Wolf.
In VQA, questions and concepts are naturally unbalanced. → 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
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.
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.
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.
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.
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**):

![Graph showing accuracy vs. question-answer pairs rareness](image-url)
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):

Confusion increases when examples are rarer
Experiments: bias reduction techniques

Bias-reduction methods also fail in this setup:
Experiments: bias reduction techniques

Bias-reduction methods also fail in this setup:

Accuracy decreases on the most frequent examples
Experiments: bias reduction techniques

Bias-reduction methods also fail in this setup:

- butd + bias reduction methods
- baseline: butd

Accuracy decreases on the most frequent examples

And is still low on the rarest ones...
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.
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

1. For each head, extract attention maps

2. Measure attention energy

3. Plot the per-head energy distribution over the dataset
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
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
Comparison: Impact of the function `choose color` on attention modes.

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

![Graphs showing attention modes](image_url)

**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](image1)

![Figure: Standard VQA model](image2)
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

![Diagram showing the impact of head pruning on accuracy for different functions.](image-url)
**Oracle transfer**

Using **oracle** and **standard data**:

1. train the visual oracle;
2. optionally, BERT-like pretraining;
3. finetune on target dataset.
Oracle transfer

Using **oracle** and **standard data**:

1. train the visual oracle;
2. optionally, BERT-like pretraining;
3. finetune on target dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Pretraining</th>
<th>GQA-OOD</th>
<th>GQA overall</th>
<th>VQA overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>acc-tail</td>
<td>acc-head</td>
<td></td>
</tr>
<tr>
<td>(a) Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) Ours</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c) Baseline (+BERT)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d) Ours (+BERT)</td>
<td>✓ ✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Classification loss + BERT losses**

**Attention modes**

**Oracle transfer**

```
Appearance shift + presence shift
```

**Oracle**

```
Oracle model
```

**Fine-tuning**

```
Fine-tuning
```

**Model**

```
Model
```

**Adapted Model**

```
Adapted Model
```
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!
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


**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*.

**Hopfield networks is all you need.**

**Lxmert: Learning cross-modality encoder representations from transformers.**

**Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned.**