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

Supplementary Material

Corentin Kervadec[1,2] Grigory Antipov[1] Moez Baccouche[1] Christian Wolf[2]
[1]Orange, Cesson-Sévigné, France. [2]LIRIS, INSA-Lyon, UMR CNRS 5205, France
corentinkervadec.github.io {grigory.antipov, moez.baccouche}@orange.com christian.wolf@insa-lyon.fr

A. Additional examples from the GQA-OOD validation split

In order to give a better insight about the benchmark's goals and possibilities, we provide additional samples extracted from the GQA-OOD validation split. In Figure 6 and 7, we show two question-answer pairs belonging to the tail. The histogram represents the answer frequency measured over the set of all questions belonging to the group of the given question. We colored the answers according to their label, head or tail. First, we can observe that the histogram is very imbalanced, which motivates the GQA-OOD approach. Second, in the caption we provide the predicted answer for each one of the evaluated model. One can notice that the predictions are diverse, showing various degree of bias dependency. However, all models are mostly relying on context biases, as shown in Figure 8. Finally, in Figure 9, we show a question-answer pair labelled as head, where all models (excepted the blind LSTM) are correct.

B. Dataset statistics

We provide some analysis and statistics to assess the reliability of the proposed benchmark. In particular, we analyse the nature and the distribution of the questions involved in *GQA-OOD* and demonstrate that it preserves the original question diversity of GQA [4].

Question diversity — Figure 3 and Figure 4 show the distribution of question structure type as defined in GQA [4] on the validation split. As one can observe, the process implemented to construct *GQA-OOD* does not alter the question diversity of the original split. However, the proportion of open questions – 'query' in Figure 1 and Figure 2 – has increased in *GQA-OOD*. Indeed, open questions – such as color questions – generally accept a wider diversity of answer, therefore it is prone to be more imbalanced. At contrary, other types such as 'choose', 'verify' or 'compare' usually accept only two possible answers and are easier to balance. Figure 1 and Figure 2 details the distribution of the structure types in the validation in *GQA-OOD* compared to GQA.

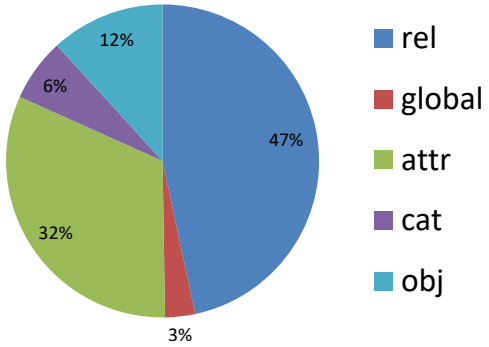

Figure 1. Distribution of the semantic types in GQA.

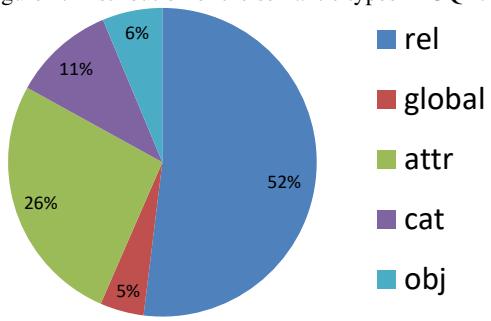

Figure 2. Distribution of the semantic types in *GQA-OOD* (tail).

C. Training details

Training hyper-parameters — All models evaluated on GQA and *GQA-OOD* have been trained on the balanced training set of GQA, and validated on the validation split. For MCAN and BUTD we use publicly available implementations at https://github.com/MILVLG/openvqa. LSTM, BUTD [1], RUBi [2], BP [3] and LM [3] are trained during 20 epochs with a batch size equals to 512 and Adam [6] optimizer. At the beginning of the training we linearly increase the learning rate from $2e^{-3}$ to $2e^{-1}$ during 3 epochs, followed by a decay by a factor

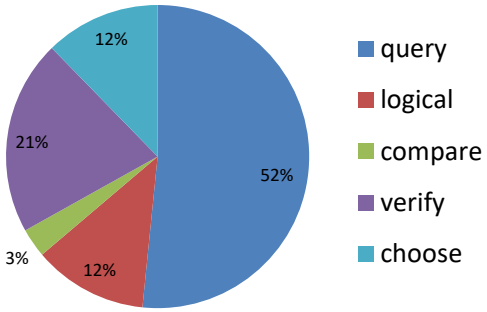

Figure 3. Distribution of the structural types in GQA.

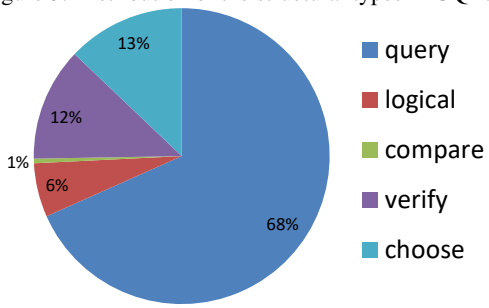

Figure 4. Distribution of the semantic types in *GQA-OOD* (tail).

of 0.2 at epochs 10 and 12. MCAN [10] is trained during 11 epoch with a batch size equals to 64 and Adamax [6] optimizer. At the beginning of the training we linearly increase the learning rate from $1e^{-4}$ to $2e^{-1}$ during 3 epochs, followed by a decay by a factor of 0.2 at epochs 10 and 12. For MMN [**?**], we use the author's implementation and trained model available at https://github.com/wenhuchen/Meta-Module-Network.

LXMERT pre-training — LXMERT [9] is pre-trained on a corpus gathering images and sentences from MSCOCO [8] and VisualGenome [7]. As the GQA dataset is built upon VisualGenome, the original LXMERT pre-training dataset contains samples from the GQA validation split. Hence, we remove those samples before pre-training in order to correctly evaluate on the GQA and GQA-OOD validation split.

Visual Oracle — The VIS-ORACLE model is based on a tiny version of the LXMERT architecture [9], where we set the hidden size to 128 and the number of per-layer heads to 4. This perfect-sighted model is taken as input objects extracted from the ground-truth GQA annotation [4]. Each object is constructed using one hot vectors encoding its class, its attributes and its in and out scenegraph relationships.

LM hyper-parameters — Figure 5 details the hyper-parameter search for the entropy penalty weight in LM [3]. We found that the entropy penalty was degrading the *GQA-OOD* accuracy when training on GQA. In particular, the flattening of the right side of the curve (most frequent samples) is even more present for higher penalty weight.

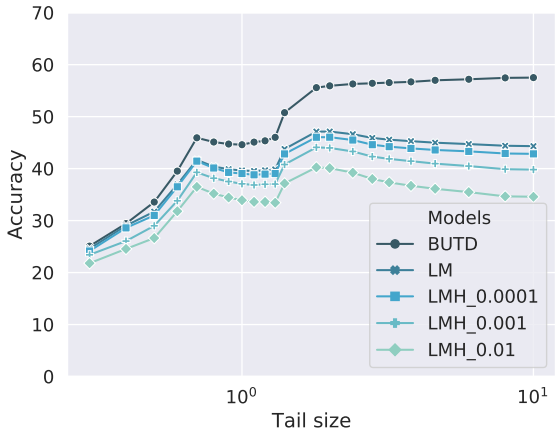

Figure 5. Influence of the LM entropy penalty weight on the prediction error distribution.

D. Measuring head/tail confusion

In the paper, we measure the head/tail confusion to get an insight on what extent the prediction errors are related to a context bias dependency. For the sake of clarity, we omit the detailed description of this procedure in the main paper. Nevertheless, the reader can find the exact methodology in the following paragraph.

The confusion corresponds to the proportion of questions where the model predicts a *head* answer with a *tail* GT answer. When plotting the confusion versus α, we decrease the size of the tail set (*i.e* we keep only the rarest question-answer pairs) while keeping the head set unchanged. Then we observe that for the majority of models, the rarest the GT answer is, the more probable the prediction belongs to the head.

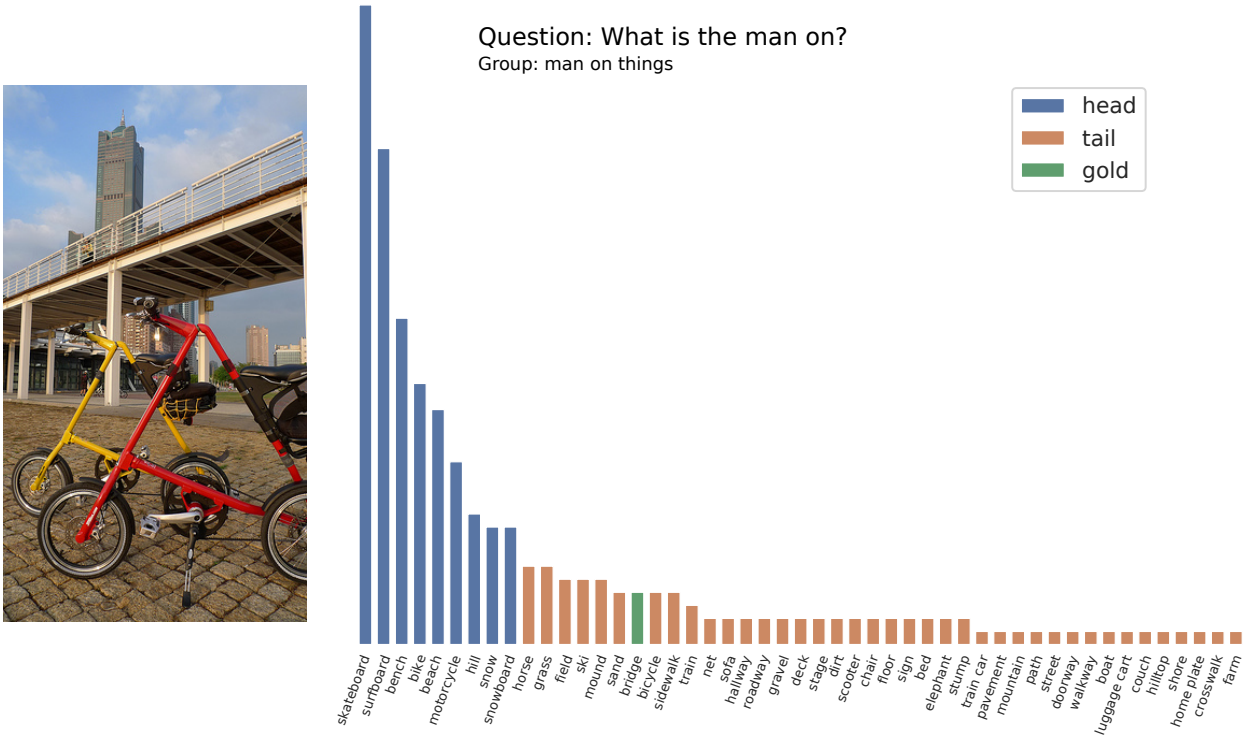

Figure 6. Tail sample from the GQA-OOD validation split. Question:*What is the man on?*. Answer:*bridge*. The evaluated models have predicted: LSTM=*skateboard*; BUTD [1], MCAN [10] = *bike*; BAN [5], BUTD+LM [3], MMN [**?**], BUTD+RUBI [2], BUTD+BP [3] = *bicycle*; LXMERT [9], ORACLE-VIS = *bridge*. The histogram represents the answer frequency measured over the set of all questions belonging to the question group.

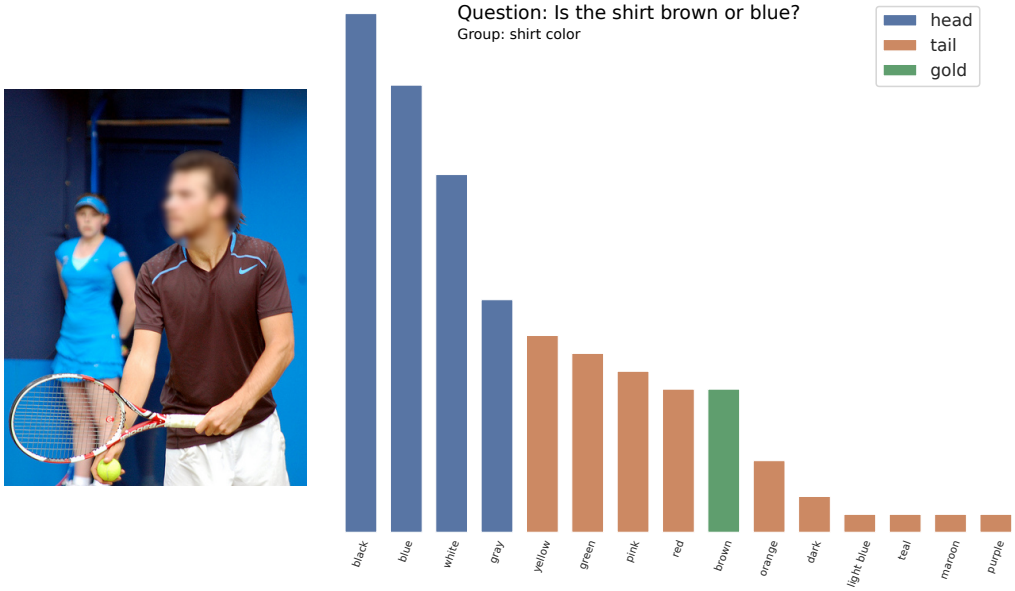

Figure 7. Tail sample from the GQA-OOD validation split. Question:*Is the shirt brown or blue?*. Answer:*brown*. The evaluated models have predicted: LSTM, BAN [5], BUTD [1], BUTD+LM [3] = *blue*; BUTD+RUBI [2], = *light blue*; MCAN [10], LXMERT [9], ORACLE-VIS, MMN [**?**], BUTD+BP [3] = *brown*. The histogram represents the answer frequency measured over the set of all questions belonging to the question group.

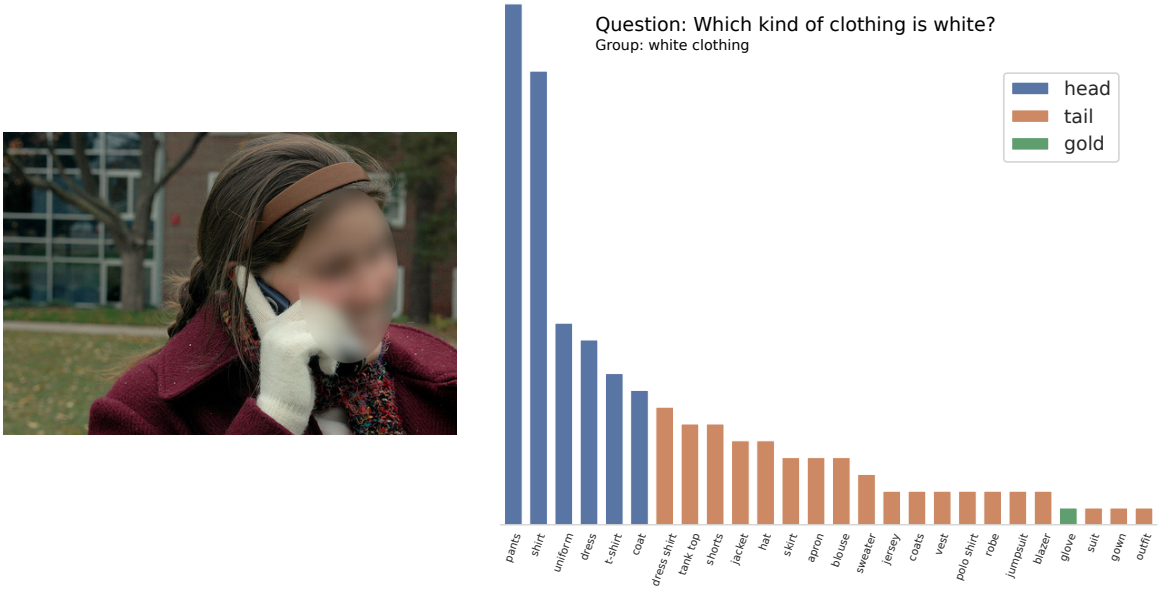

Figure 8. Tail sample from the GQA-OOD validation split. Question:*Which kind of clothing is white?*. Answer:*glove*. The evaluated models have predicted: LSTM = *shirt*; LXMERT [9], BUTD [1], BAN [5], MMN [**?**], BUTD+RUBI [2] = *coat*; MCAN [10],= *jacket*; BUTD+LM [3], BUTD+BP [3]= *long sleeved*; ORACLE-VIS = *glove*. The histogram represents the answer frequency measured over the set of all questions belonging to the question group.

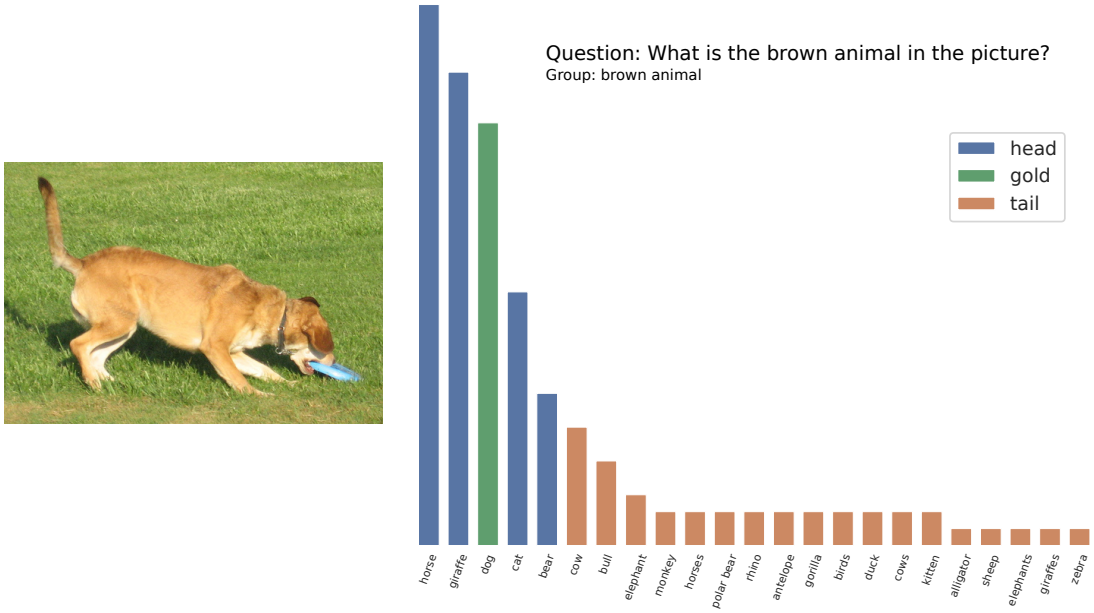

Figure 9. Head sample from the GQA-OOD validation split. Question:*What is the brown animal in the picture?*. Answer:*dog*. The evaluated models have predicted: LSTM = *horse*; BAN [5], BUTD [1], BUTD+LM [3], BUTD+RUBI [2], MCAN [10], LXMERT [9], ORACLE-VIS, MMN [**?**], BUTD+BP [3] = *dog*. The histogram represents the answer frequency measured over the set of all questions belonging to the question group.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1, 3, 4

[2] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems*, pages 839–850, 2019. 1, 3, 4

[3] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4060–4073, 2019. 1, 2, 3, 4

[4] Drew A Hudson and Christopher D Manning. 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, 2019. 1, 2

[5] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 3, 4

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014. 1, 2

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[9] Hao Tan and Mohit Bansal. 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, 2019. 2, 3, 4

[10] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 2, 3, 4