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

Corentin Kervadec, Grigory Antipov, Moez Baccouche, Christian Wolf

1Orange, Cesson-Sévigné, France. 2LIRIS, INSA-Lyon, UMR CNRS 5205, France

corentinkervadec.github.io {grigory.antipov, moez.baccouche}@orange.com christian.wolf@insa-lyon.fr

Abstract

Models for Visual Question Answering (VQA) are notorious for their tendency to rely on dataset biases, as the large and unbalanced diversity of questions and concepts involved and tends to prevent models from learning to "reason", leading them to perform "educated guesses" instead. In this paper, we claim that the standard evaluation metric, which consists in measuring the overall in-domain accuracy, is misleading. Since questions and concepts are unbalanced, this tends to favor models which exploit subtle training set statistics. Alternatively, naively introducing artificial distribution shifts between train and test splits is also not completely satisfying. First, the shifts do not reflect real-world tendencies, resulting in unsuitable models; second, since the shifts are handcrafted, trained models are specifically designed for this particular setting, and do not generalize to other configurations. We propose the GQA-OOD benchmark designed to overcome these concerns: we measure and compare accuracy over both rare and frequent question-answer pairs, and argue that the former is better suited to the evaluation of reasoning abilities, which we experimentally validate with models trained to more or less exploit biases. In a large-scale study involving 7 VQA models and 3 bias reduction techniques, we also experimentally demonstrate that these models fail to address questions involving infrequent concepts and provide recommendations for future directions of research.

1. Introduction

Visual Question Answering (VQA), i.e. the task of answering a question posed over an image, is often seen as a testbed for the capability of learning-based systems to perform high-level reasoning. This multimodal problem is notorious for its diversity, meaning that VQA models are required to learn various high-level general representations of concepts of the physical world as well as their interactions.

Efforts to learn the necessary high-level reasoning from large-scale datasets depend on the absence of harmful biases in the data, which could provide unwanted shortcuts to learning in the form of "Clever Hans" effects. Unfortunately, and in spite of recent efforts [11, 15], most VQA datasets remain very imbalanced. Common concepts are significantly more frequent, e.g. the presence of a "red rose", compared to out of context concepts like the presence of a "zebra in a city". This causes the tendency of models to overly rely on biases, hindering generalisation [7, 9]. Despite a general consensus on this diagnostic, systemic evaluations of error distributions are rare. In particular, overall accuracy is still the major, and often unique, metric used to evaluate models and methods, although it is clearly insufficient. Several questions remain open. How is error distributed? Are true positives due to reasoning or to exploitation of bias? What is the prediction accuracy on infrequent vs. frequent concepts? How can we validate models in Out Of Distribution (OOD)-settings?

In this work we propose a new benchmark and a study of State-Of-The-Art (SOTA) VQA models, which allows to precisely answer these questions. The proposed new evaluation protocol is complementary to existing ones, but al-
allows a better diagnostic of current VQA performance. In particular, our benchmark can be viewed as an alternative to the VQA-CP [2] dataset, which has lead to mixed results (see [25] for a recent comprehensive study). Our benchmark comprises (i) a new fine-grained reorganization of the GQA dataset [15] introducing distribution shifts in both validation and test sets (see Figure 1-a); (ii) a set of evaluation metrics; (iii) new evaluation plots illustrating the generalisation behavior of VQA models on different operating points. The choice of the GQA dataset is motivated by its useful structuring into question groups, which allows to capture biases precisely, to select groups with strong biases and to create distribution shifts tailored to the exact nature of each question (see Figure 1-b). It also makes it possible to analyze how errors are distributed over different associations of concepts according to their frequency in the dataset.

Our contributions — (i) We propose and make public\(^1\) a new fine-grained re-organization of the GQA dataset and a set of the respective evaluation metrics allowing to precisely evaluate the reasoning behavior of VQA models and to characterise and visualise their generalisation behaviour on different operating points w.r.t distribution shifts. (ii) Compared to competing benchmarks, our dataset features distribution shifts for both, validation and test, allowing to validate models under OOD conditions. (iii) We experimentally evaluate the usefulness of the proposed metric showing its behavior on models trained to, more or less, exploit biased regions [10]. They tend to overly rely on dataset biases [12], and are not able to generalize to unseen distributions [2]. Methods to solve this include e.g setting up an adversarial game against a question-only adversary to regularize training [21], and a question-only branch in addition to a base model during training to prevent it from learning textual biases [7] (RUBi). At test time, this branch is omitted. [9] regularize model predictions using question type statistics from the training set. Other attempts force VQA models to attend to the most significant visual regions from humans’ perspective [28, 22]. While they show promising results when evaluated on unseen distributions [2], they generally slightly degrade performances on standard benchmarks, which tend to favor models relying on dataset biases.

Reinventing VQA evaluation — Propositions of new evaluation methods for VQA went hand in hand with the design of new models. Early work [19] proposed a soft evaluation score based on a lexical database, then replaced by a hard classification score, more prone to favor biased predictions, but easier to use in practice. The authors of GQA [15] proposed several additional metrics, namely: consistency, plausibility, validity, distribution, etc. They provide insights on VQA performance, but do not evaluate its ability to predict correct answers in OOD setting, especially when applied to the original balanced GQA dataset. To evaluate the generalization capability of Neural State Machines, [14] proposed interesting splits of GQA, reorganizing the train and validation splits to distinctly evaluate how well the models generalize on visual content and on the linguistic structure of the question. The authors of [5] analyzed generalization to unseen associations of known linguistic structures using their CLOSURE benchmark. They demonstrated that SOTA models (including those which were explicitly designed to overcome this issue) fail to generalize in these conditions. This confirms the need of carefully designed benchmarks to evaluate the true capabilities of VQA models. In this context, [2] introduced a reorganisation of the VQA2 [11] dataset splits, namely VQA-CP2, where the training distribution is made explicitly different from the one in the test. We discuss the differences between VQA-CP2 and our benchmark in detail in Section 3.

---

\(^1\)https://github.com/gqa-ood/GQA-OOD
Shortcomings of Out-of-distribution testing Recent works have raised some criticizes about OOD evaluation protocols. In particular, [25] points out several pitfalls observed when evaluating VQA in OOD setting using VQA-CP [2]: (1) this protocol does not allow to test a model in both out- and in-distribution settings without having to retrain it on a different set of data; (2) several works rely of known construction procedures of the OOD test split, e.g. the test answers’ distribution is the inverse of the train one, coming up with methods specialized to a very particular setup; and (3) as no validation split is provided, the hyper-parameters are usually selected on the test split. Although these wrong practices are in part to the responsibility of the model designers, they show us the need to rethink the evaluation protocol. In parallel, [23] come up with an interesting negative result while analysing visual grounding bias-reduction methods designed on top of the VQA-CP dataset. These methods [28, 22], consisting in supervising a VQA model to attend to visual regions which are relevant to a human considering the question, are very efficient on VQA-CP. Surprisingly, [23] found that simply enforcing the model to attend to random visual regions was at least as much efficient on out- and in-distribution settings. Why was this negative result not observed before? We think that a more profound empirical evaluation of models’ behavior would help to better judge and compare the efficiency of new VQA methods.

These works unveil the need of a better evaluation protocol, allowing to test in both out- and in-distribution settings without falling into ood testing pitfalls. In this paper, we try to provide this evaluation protocol. We complement it with a large experimental study of the generalisation abilities of VQA architectures and de-bias methods in order to better understand theirs pros and cons.

3. GQA-OOD: a benchmark for OOD settings

We introduce a new VQA benchmark named GQA-OOD designed to evaluate models and algorithms in OOD configurations. We here define OOD samples as rare events, in particular measured w.r.t. to a base distribution, e.g. a training distribution. These rare events might involve concepts which are also present in the training set. Let’s for instance consider the question: ‘What color is this rose?’

If the image represents a rose, then red would be a common color, but in an OOD setting, infrequent (correct) test answers would be, for instance, blue, requiring models to reason to provide the correct answer. We design a benchmark where this shift is not global but depends on the context. If the context changes, and the flower type is a violet, then a (correct) OOD answer would now be red instead of blue.

The GQA-OOD benchmark consists of a dataset and new evaluation metrics. The dataset itself is based on the existing GQA [15] dataset, which provides more fine-grained annotations compared to competing VQA2 [11] (the questions in GQA have been automatically generated from scene graphs, which allows better control of the context). Figure 2 shows how the proposed protocol compares to the existing GQA protocol: the two share the same (existing) training set, but we introduce fine-grained shifts into both the validation and the test sets applying the process further described below. The shifted subsets have been constructed in 3 steps: (i) dividing questions into groups according to their contexts; (ii) extracting the most imbalanced question groups considering their answer distributions; (iii) selecting OOD samples among the remaining questions.

**Question groups** — To structure the process introducing distribution shifts, we use the notion of local groups provided in the GQA annotation. They allow to precisely define the type of a question, e.g. ‘What color ...?’, ‘Where is ...?’, etc. They also depend on the concepts related to the question, e.g. ‘zebra’, ‘violet’, etc. There is a total of ≈ 37K local groups related to ≈ 132K questions in the GQA validation split. We use the balanced version of GQA, whose question distribution has been smoothed in order to obtain a more uniform answer distribution. However, this does not impact the imbalanced nature of the dataset, which is often due to real world tendencies, e.g. that ‘roses are red’.

**Measuring group imbalance** — We extract a subset of the most imbalanced question groups, as we are interested in evaluating the prediction error specifically in the context, where shifts in distribution are meaningful and strong. We measure balance through Shannon entropy given as $e(x) = -\sum_{i=0}^{d} p(x_i) \log p(x_i)$, where $p(x_i)$ is the estimated probability of the class $i$. As entropy depends on the number of answer classes, which is highly variable between different question groups, we normalize entropy w.r.t. the number $d$ of possible answers in the group: $\tilde{e}(x) = \frac{e(x)}{\log(d)}$, where $\log(d)$ is equal to the entropy of a uniform distribution of size $d$. Normalized entropy $\tilde{e}(x)$ thus measures how

<table>
<thead>
<tr>
<th>Dataset Split</th>
<th>#Quest.</th>
<th>#Groups</th>
<th>#Imgs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GQA-OOD val testdev</td>
<td>51,045</td>
<td>3,849</td>
<td>12,062</td>
</tr>
<tr>
<td>GQA testdev</td>
<td>12,578</td>
<td>7,803</td>
<td>906</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset Split</th>
<th>Subset</th>
<th>#Quest.</th>
<th>#Groups</th>
<th>#Imgs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GQA-OOD val head</td>
<td>33,882</td>
<td>3,849</td>
<td>8,664</td>
<td></td>
</tr>
<tr>
<td>tail</td>
<td>17,163</td>
<td>3,849</td>
<td>6,632</td>
<td></td>
</tr>
<tr>
<td>GQA-OOD testdev head</td>
<td>1,733</td>
<td>471</td>
<td>365</td>
<td></td>
</tr>
<tr>
<td>tail</td>
<td>1,063</td>
<td>471</td>
<td>330</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Data statistics: (a) GQA-OOD vs. GQA; (b) head vs. tail
close the distribution \( p(x) \) is to a uniform distribution of the same dimension. Finally, we keep groups with a normalised entropy smaller than a threshold empirically set to \( T=0.9 \). This selects all benchmark’s questions, but further work is done in order to select specific answer classes for each group.

**OOD setting and metrics** — We introduce a shift in distribution by selecting a subset of answer classes for each question group according to their frequencies, and introduce three different metrics according to which classes are used for evaluation. All these metrics are defined over the aforementioned imbalanced local groups. Figure 1 illustrates how the subsets are selected using the concrete example mentioned imbalanced local groups. Figure 1 illustrates how the subsets are selected using the concrete example answer histogram of question group objects on walls.

- **Acc-tail**: the accuracy on OOD samples, which are the samples of the tail of the answer class distribution, i.e. the rarest answers given the context. We define the tail classes as classes \( i \) with \( |a_i| \leq \alpha \mu(a) \), where \( |a_i| \) is the number of samples belonging to the class \( i \) and \( \mu(a) \) is the average sample count for the group. We empirically set the parameter \( \alpha=1.2 \), and in Section 4.2 we analyze and illustrate the impact of the choice of \( \alpha \) on Acc-tail. Figure 1 provides an example of such a tail question — we can see that the answer Star is rare in this group, therefore it belongs to the tail set like the other answers shown in orange.

- **Acc-head**: the accuracy on the distribution head for each local group, given as difference between the whole group and its tail (blue answers in Figure 1).

- **Acc-all**: the overall (classical) accuracy over all GQA-OOD samples, i.e. the in-domain accuracy. In Figure 1, this corresponds to the blue and orange answers.

Table 1 provides statistics of the proposed benchmark. We also analyzed the nature, distribution and diversity of the questions w.r.t to GQA [15], c.f. Appendix.

**Difference with VQA-CP2** — The VQA-CP2 [2] dataset was a first of its kind and paved the way for follow-up work on bias reduction methods in VQA. However, its construction is conceptually different from our work, partially due to the restrictions of the base dataset VQA2 w.r.t. to GQA, but also due to key design choices. Lacking annotations on group structure in the base dataset, questions are grouped according to their first words and the ground-truth answer. The shift is created by splitting according to types. In contrast, our proposed GQA-OOD dataset allows fine-grained analysis of the generalization behavior of a VQA model by (i) question group, and via (ii) different metrics corresponding to different amounts of shifts (acc-tail vs. acc-head) in out- and in-distribution settings, and (iii) even through the possibility of continuous evaluation along different operating points (see Figure 3).

VQA-CP2 comprises two splits only (train-test), lacking the possibility of validating model hyper-parameters. Most techniques therefore seem to optimize their hyper-parameters on the test split [7, 9, 21, 28, 22], which should be frowned upon, or, alternatively, validate on a subset of train which does not include a shift [26]. Our GQA-OOD dataset contains a validation set with a shift w.r.t. to the train set, which allows to validate hyper-parameters in OOD settings. Finally, unlike VQA-CP, our proposed dataset requires models to be trained on the existing GQA train split. This forces models to reduce bias in their test results while being exposed to natural tendencies and biases captured in the training corpus, favoring work on bias reduction through methodology instead of through cleaning of training data.

**Limitations** — The proposed benchmark is built on the GQA dataset, whose questions have been automatically generated, resulting in a limited vocabulary (GQA covers 70% of VQA2 answers [15]) and a synthetic syntax. While the images are natural and real, one might argue, that the questions are not “in the wild”. However, the benefits of the synthetic nature of the questions largely out-weight its limitations. In particular, this offers a better control on the data and excludes unmodelled external knowledge, which leads to a better evaluation of reasoning abilities. We made the source code publicly available, and we encourage the field to use it to study robustness in OOD settings.

4. Experiments

In our experiments we used several SOTA VQA models and we compared the proposed GQA-OOD benchmark to the standard benchmarks VQA2 [11], GQA [15] and VQA-CP2 [2]. The line-up includes strong models with object-level attention and two Transformer [27]-based model, as well as two blind baseline models and a visual oracle.

**Question Prior** — this blind baseline returns the most probable answer estimated from training set statistics. Following [2, 11], we use the question types priors when evaluating on VQA-CP and VQA2. For GQA, we use the training global group priors. **LSTM** [4] — this blind baseline takes GloVe embeddings [20] and encodes them using an LSTM [13] followed by a feed-forward network. Neither of the two blind baselines use the input image. **BUTD** [3] — a strong VQA baseline based on object-level attention, in particular, bounding boxes with dense visual feature vectors extracted from the image using an object detector. **BAN4** [18] — another object-level attention based model using a bilinear fusion mechanism to models relations between words and objects with multi-hop reasoning. We use the 4-layers version of BAN. **MCAN** [29] — a Transformer [27] architecture designed to model both intra-modality interactions and the inter-modality ones. It allows complex multi-hop reasoning through several stacked self-attention blocks. In our experiments, we use the 6-layers version of MCAN. **LXMERT** [24] — a SOTA approach combining a Transformer architecture (close to the
one used in MCAN) with a large scale pre-training procedure. LXMERT is pretrained on roughly 9 millions question-image pairs with semi-supervised tasks such as reconstructing masked words and visual objects. It is then fine-tuned on the GQA dataset. MMN [8] — a Meta Module Network for compositional visual reasoning. It is based on a hybrid approach combining neural module networks (NMN) and monolithic architectures (such as MCAN). The former, NMN, is based on hand-crafted neural network program blocks and is supposed to allow a better compositionality and interpretability. The latter, monolithic architecture, performs its operations in a latent space and has been shown to be experimentally more efficient. NMN tries to combine the best of both worlds. VIS-ORACLE — a model with a perfect sight, i.e. taking as input the question and a set of ground truth objects directly taken from the annotation of GQA\(^3\). It allows to evaluate the performance of a model without the imperfection of the visual extractor. Is is based on a Transformer architecture similar to the one used for LXMERT with smaller capacity.

For training details of all models we refer to the Appendix. Models evaluated on GQA-OOD were trained on the training set of GQA (balanced) and validated on the validation split of GQA-OOD, unless otherwise stated. When available, we provide the standard deviation computed over at least four different seeds.

### 4.1. Evaluation of the proposed metric

We believe that a good evaluation metric satisfies at least two properties: it is easy to interpret, and it provides an estimate for the quality targeted by the evaluation. We argued above on the merits of our proposed tail accuracy (Acc-tail) as a way of estimating VQA performance less influenced by bias. In what follows, we complete this by an experimental validation of the metric. To this end, we compared two different VQA models, one of which has been trained in a way known to be biased (BUTD [3]+all), and evaluate the proposed metric’s capacity to detect this bias. All scores in % on the GQA-OOD val split.

![Figure 3. Performance (higher is better) (a) and head/tail confusion (lower is better) (b) for different definitions of the tail distribution (\(\alpha\) parameter values) on the GQA-OOD benchmark. We compare several VQA models. The x-axis is in log-scale.](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>Baseline benchmark (Tot. Acc.)</th>
<th>Proposed benchmark (Acc-tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUTD [3]++bal</td>
<td>60.7±0.4</td>
<td>45.4±0.3</td>
</tr>
<tr>
<td>BUTD [3]+all</td>
<td>59.8±0.1</td>
<td>41.2±0.1</td>
</tr>
<tr>
<td>(\Delta) (relative)</td>
<td>-1.4%</td>
<td>-7.7%</td>
</tr>
</tbody>
</table>

\(\Delta\) Table 2. We compare two different VQA models based on BUTD [3], one of which has been trained on a split known to be biased (BUTD [3]+all), and evaluate the proposed metric’s capacity to detect this bias. All scores in % on the GQA-OOD val split.

\(\Delta\) As GT annotations (scene-graphs) are only available for the train and validation split, we do not evaluate VIS-ORACLE on the testdev split.
the metric focuses on the rarer and rarer question-answer pairs, providing valuable evidence that acc-tail is indeed well suited for measuring VQA performance undisturbed by bias dependencies.

4.2. Analysis of VQA model error distributions

The GQA-OOD benchmark allows us to perform an analysis of the prediction error distributions for various VQA models as shown in Table 3. We provide the three metrics introduced in Section 3: acc-tail, acc-head and acc-all. We also measure the difference $\Delta(\text{tail,head}) = \frac{\text{acc-head} - \text{acc-tail}}{\text{acc-head}}$ to illustrate how much is the error prediction imbalanced between frequent and rare answers.

**Models fail on rare question-answer pairs** — We can see that VQA models (dramatically) fail to generalize on infrequent association of concepts. The two blind models (Question Prior and LSTM in Table 3-a) obtain the highest gap between acc-tail and acc-head, explained by the fact that they uniquely rely on question biases. The $\Delta$ score indicates that BUTD, MMN, MCAN, BAN4 and LXMERT also struggle (in a lesser extent) to generalize on the less frequent question-answer pairs. Nevertheless, we observe that the Transformer-based architecture combined with large-scale BERT training, LXMERT, outperforms all models on the acc-tail metric, confirming its superiority. This is corroborated by [12], who show that pretrained Transformers improve OOD robustness in NLP.

In contrast to our proposed acc-tail metric, the metric acc-all, close to the standard VQA metric, does not reflect the true model’s performances, since it is mechanically increased by the high scores obtained on the most frequent question-answers. This confirms the need of a two-in-one evaluation: measuring the out- and in-distribution performance scores, as we propose.

**Visualising the generalisation behavior** — The definition of what constitutes a “rare” answer, i.e. the size of the tail, depends on the parameter $\alpha$. In Figure 3-a, we analyze how VQA model prediction errors (acc-tail) depend on this definition, i.e. how models behave w.r.t. to questions whose answers are more and more rare. Increasing $\alpha$ increases the tail — in the extreme case it is equal to the whole distribution (right side of the plot). With small $\alpha$, only the most infrequent question-answer pairs are evaluated (left side of the plot). All models follow the same dynamic: starting from a tail size which represents roughly half of the question-answer pairs, tail accuracy starts to linearly decrease until reaching a dramatically low score (about 30 pts lower than the overall accuracy). An exception is VIS-ORACLE: its dynamics is nearly flat, prediction error is almost decorrelated from answer rareness. This provides evidence that a model using perfect visual input is able to learn reasoning with significantly decreased dependency on dataset biases.

We complement this analysis by measuring the confusion between head and tail as a function of $\alpha$, shown in Figure 3-b, which provides insights on the causes of the generalisation failure observed in Figure 3-a. This confusion corresponds to the proportion of questions where the model predicts a head answer with a tail GT answer$^4$. For $\alpha=1.2$, LXMERT confuses answers for 25% of questions, which increases up to 42% for $\alpha=0.3$. Similar behavior is observed for the other models, but interestingly not for VIS-ORACLE, where curve is nearly flat, again providing evidence for a low dependency on statistical biases in the training set. As a side note, we found in [17] that initialising LXMERT weights with VIS-ORACLE allows to boost the accuracy on acc-tail.

**Exploiting biases vs. “reasoning”** — It is difficult to assess, whether a model reasons or not, in particular since the term “reasoning” has various different definitions. Referring to [6], “a plausible definition of ‘reasoning’ could be ‘algebraically manipulating previously acquired knowledge in order to answer a new question’”. In the context of VQA, we could interpret this as “algebraically manipulating words and visual objects to answer a new question”. With this definition, using statistics biases cannot be considered reasoning, but should rather be denoted as “educated guesses” [15] or biased answers. Using the proposed GQA-OOD benchmark, we explore the estimation of three reasoning labels qualifying the mode of operation a model uses for a given input: bias, reason and other/unknown. In absence of GT information, we propose to estimate these labels from proxy rules: a VQA model is estimated to reason, when it correctly predicts an answer, which is rare in GT and rare in prediction; it is considered biased, when it wrongly predicts an answer, which is frequent in its prediction and rare in GT.

Figure 5-a shows the calculation of these labels based on the distribution of the head and tail labels of each answer in the predictions (rows) and GT (columns) for LXMERT [24] on the validation split of GQA-OOD. We add a borderline label representing the fuzzy frontier between reasoning and bias exploitation$^5$. In Figure 5-b, we show the distribution $\text{bias} < \alpha < 1.2$, $\text{head} > 1.2$, $\text{borderline} = 0.7 < \alpha < 1.2$, $\text{tail} < 0.7$.

Figure 4. Acc-tail performance and head/tail confusion, as in Fig. 3, but for different bias-reduction methods on top of BUTD [3].

Figure 5. We estimate “reasoning labels”: the model is estimated to reason, when it correctly predicts an answer rare in GT and rare in prediction; it is considered biased, when it wrongly predicts an answer, which is frequent in its prediction and rare in GT. All values are computed over the GQA-OOD validation split. The matrix (a) shows the joint distribution of predicted and GT classes. (b): Distribution the estimated reasoning labels over the GQA [15] question types for the LXMERT [24] model. The model often predicts a biased answer on the query and compare questions while there is evidence that it may reason on verify, choose and logical questions.

Table 3. Comparison of several VQA methods on the GQA-OOD test dev split. Acc-tail: OOD settings. Acc-head: accuracy on most probable answers (given context), scores in %. (a) different models; (b) BUTD [3] model combined with different bias reduction techniques.

<table>
<thead>
<tr>
<th>Model</th>
<th>Uses image</th>
<th>acc-all</th>
<th>acc-tail</th>
<th>acc-head</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quest. Prior</td>
<td>✗</td>
<td>21.6</td>
<td>17.8</td>
<td>24.1</td>
<td>35.4</td>
</tr>
<tr>
<td>LSTM [4]</td>
<td>✗</td>
<td>30.7</td>
<td>24.0</td>
<td>34.8</td>
<td>45.0</td>
</tr>
<tr>
<td>BUTD [3]</td>
<td>✓</td>
<td>46.8 ± 1.1</td>
<td>42.1 ± 0.9</td>
<td>49.1 ± 1.1</td>
<td>16.6</td>
</tr>
<tr>
<td>MCAN [29]</td>
<td>✓</td>
<td>50.8 ± 0.4</td>
<td>46.5 ± 0.5</td>
<td>53.4 ± 0.6</td>
<td>14.8</td>
</tr>
<tr>
<td>BAN4 [18]</td>
<td>✓</td>
<td>50.2 ± 0.7</td>
<td>47.2 ± 0.5</td>
<td>51.9 ± 0.0</td>
<td><strong>9.9</strong></td>
</tr>
<tr>
<td>MMN [8]</td>
<td>✓</td>
<td>52.7</td>
<td>48.0</td>
<td>55.5</td>
<td>15.6</td>
</tr>
<tr>
<td>LXMERT [24]</td>
<td>✓</td>
<td><strong>54.6</strong></td>
<td>49.8</td>
<td><strong>57.7</strong></td>
<td>15.9</td>
</tr>
<tr>
<td>BUTD [3] + RBi+QB</td>
<td></td>
<td><strong>46.7 ± 1.3</strong></td>
<td><strong>42.1 ± 1.0</strong></td>
<td><strong>49.1 ± 1.5</strong></td>
<td>17.3</td>
</tr>
<tr>
<td>BUTD [3] + BP</td>
<td></td>
<td>38.8 ± 2.4</td>
<td>35.7 ± 2.3</td>
<td>40.8 ± 2.7</td>
<td>14.3</td>
</tr>
<tr>
<td>BUTD [3] + LM</td>
<td></td>
<td>34.5 ± 0.7</td>
<td>32.2 ± 0.2</td>
<td>35.9 ± 0.2</td>
<td><strong>11.5</strong></td>
</tr>
<tr>
<td>BUTD [3] + BP</td>
<td></td>
<td>33.1 ± 0.4</td>
<td>30.8 ± 1.0</td>
<td>34.5 ± 0.5</td>
<td>12.0</td>
</tr>
</tbody>
</table>

of these reasoning labels over the different GQA structural question types [15]: verify, choose, compare and query. We observe that LXMERT seems to “reason” on the verify, choose and logical questions, which are binary questions, while compare and query questions are the most prone to bias exploitation. From this, we conclude that future efforts on improvements of model capacities to answer open questions (e.g. typed as query) should be particular fruitful.

4.3. Re-evaluating bias-reduction methods

We use the proposed benchmark to re-evaluate several bias-reduction methods, which have been initially designed on the VQA-CP [2] dataset. As these methods were designed to be model-agnostic [7, 9], we use them together with the BUTD [3] architecture:

RUBi [7] — adds a question-only branch to the base model during training to prevent it from learning question biases. This branch is omitted during evaluation. To better analyze bias dependencies, we also study a modified version of RUBi, which we refer to as RUBi+QB below. In this variant, the question-only branch is kept during evaluation.

BP [9] — is similar to RUBi but differs by directly taking training set statistics to infer question type biases during training. The question type biases are fused with the base model predictions using a product of experts [9], and removed during testing. LM [9] — is an improved version of BP [9]. In this version, the question bias is dynamically weighted by the base model in order to control its influence. In the original setup, an entropy penalty is added in the loss to prevent the model to ignore the bias. Nevertheless, when training on GQA, we obtain better results without this penalty (see appendix for details).

Surprisingly, none of the three bias-reduction methods succeed to improve acc-tail (cf. Table 3-b), they even deteriorate acc-head. This is unexpected as they have been designed to overcome the dependency on question type biases. For further analysis, we evaluate RUBi while keeping the question-only branch during testing (RUBi+QB). As expected, it outperforms RUBi on acc-head, indicating it has better captured frequent patterns. However, it also outperforms RUBi on the OOD settings, demonstrating that preventing from learning frequent patterns does not necessarily increase performances on rare samples.

We provide a visualization of the generalisation behav-

---

Footnotes:

6 only 1% of the tail questions are typed as compare.

7 VQA2: biases are over question types; GQA: local groups.
bias-reduction; looking at head/tail responds the same way as BUTD, confirming the effect of same frequent/rare imbalance. We observe that RUBi+QB Figure 3-a performs worse than BUTD but conserves the ing dataset biases. As a comparison, the LSTM model in what succeeded in preventing the base model from learn-
curve, however, corresponding to rare samples, remains al-
has been reduced by bias-reduction. The left side of the 
all accuracy, dominated by frequent question-answer pairs,
the right side of the curve has flattened, indicating that over-
LM [9] and, to a lesser extent, RUBi [7], we observe that 
benchmarks. Results computed on the testdev split of
GQA-OOD
Table 4. We compare the proposed \textit{acc-tail} metric with other benchmarks. Results computed on the testdev split of \textit{GQA-OOD} and GQA [15], the test split of VQA-CP2 [2] and the VQA2 [11] validation split. Values in italic: trained and tested by ourselves.

<table>
<thead>
<tr>
<th>Model</th>
<th>VQA2 overall</th>
<th>GQA overall</th>
<th>GQA-CP2 overall</th>
<th>GQA-OOD acc-tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q. Prior</td>
<td>32.1</td>
<td>27.0</td>
<td>55.6</td>
<td>8.8</td>
</tr>
<tr>
<td>LSTM [4]</td>
<td>43.0</td>
<td>39.1</td>
<td>3.6</td>
<td>22.1</td>
</tr>
<tr>
<td>BUTD [3]</td>
<td>63.5</td>
<td>51.6±0.3</td>
<td>1.8</td>
<td>40.1</td>
</tr>
<tr>
<td>MCAN [29]</td>
<td>66.1</td>
<td>56.2±0.2</td>
<td>1.6</td>
<td>42.5</td>
</tr>
<tr>
<td>BAN4 [18]</td>
<td>65.9</td>
<td>54.7±0.4</td>
<td>1.6</td>
<td>40.7</td>
</tr>
<tr>
<td>MMN [8]</td>
<td>-</td>
<td>59.6</td>
<td>1.8</td>
<td>-</td>
</tr>
<tr>
<td>LXMERT [24]</td>
<td>69.9</td>
<td>59.6</td>
<td>1.5</td>
<td>-</td>
</tr>
<tr>
<td>BUTD [3]</td>
<td>63.5</td>
<td>51.6±0.3</td>
<td>1.8</td>
<td>40.1</td>
</tr>
<tr>
<td>+RUBi+QB</td>
<td>-</td>
<td>51.9±1.1</td>
<td>1.7</td>
<td>47.6±3.7</td>
</tr>
<tr>
<td>+RUBi [7]</td>
<td>61.2</td>
<td>43.5±2.0</td>
<td>1.9</td>
<td>44.2</td>
</tr>
<tr>
<td>+LM [9]</td>
<td>56.4</td>
<td>39.7±0.7</td>
<td>2.1</td>
<td>52.0</td>
</tr>
<tr>
<td>+BP [9]</td>
<td>63.2</td>
<td>39.6±0.3</td>
<td>2.2</td>
<td>39.9</td>
</tr>
</tbody>
</table>

Table 4. We compare the proposed \textit{acc-tail} metric with other benchmarks. Results computed on the testdev split of \textit{GQA-OOD} and GQA [15], the test split of VQA-CP2 [2] and the VQA2 [11] validation split. Values in italic: trained and tested by ourselves.

ior on bias-reduction methods in Figure 4-a. For BP [9], LM [9] and, to a lesser extent, RUBi [7], we observe that the right side of the curve has flattened, indicating that overall accuracy, dominated by frequent question-answer pairs, has been reduced by bias-reduction. The left side of the curve, however, corresponding to rare samples, remains almost unchanged, revealing that these methods have somewhat succeeded in preventing the base model from learning dataset biases. As a comparison, the LSTM model in Figure 3-a performs worse than BUTD but conserves the same frequent/rare imbalance. We observe that RUBi+QB responds the same way as BUTD, confirming the effect of bias-reduction; looking at head/tail confusion in Figure 4-b, the result is even more pronounced. In short, we demonstrate the effectiveness of bias reduction methods in preventing the base model from learning salient properties of the training set, and occasionally reducing the dependency toward dataset biases. However, this does not necessarily help the model to learn the subtle distributions, required for generalization and for learning to reason.

4.4. Comparison with other benchmarks

We compare the proposed \textit{GQA-OOD} benchmark with the following three standard VQA datasets:

\textbf{GQA} [15] (balanced version) is a dataset with \textasciitilde 1.7M question-answer pairs automatically generated from real images. Among all VQA datasets, GQA has the richest annotations allowing to evaluate models with complementary metrics: consistency, validity, plausibility, [15] etc. Here, we only discuss overall accuracy and the distribution score on the GQA testdev split as the other metrics are unrelated to the topic of our paper. The distribution score is obtained by measuring the match between the true GT answer distribution and the predicted distribution. \textbf{VQA2} [11] (265K images, \textasciitilde 3 questions each) is composed of questions created by humans, each question is annotated with 10 GT an-
swers. We compare with overall accuracy on the VQA2 validation split. \textbf{VQA-CP2} [2] has been constructed by re-organising the training and validation splits of VQA2 [11] aiming at differences in answer distributions between training and test splits. It has been designed to measure sensitivity to language bias.

\textbf{Comparison with GQA and VQA2} — In Table 4.3, we compare our \textit{acc-tail} score with the other benchmarks. We can see that overall accuracy on GQA and VQA2 is not sufficient to fully characterize the VQA performances. Our evaluation in OOD settings is the only one to reveal that even SOTA models struggle on infrequent question-answer pairs. The best-performing model LXMERT looses about 10 points in the OOD setting. Our metric also unveils that, despite performing on-par with LXMERT on GQA overall, MMN struggles more on infrequent question-answer pairs. Finally, we argue that \textit{acc-tail} is easier to interpret than the error distribution measure defined in GQA.

\textbf{Comparison with VQA-CP2} [2] — Comparing \textit{acc-tail} to VQA-CP2 overall accuracy, we observe similar scores on standard VQA models, but a completely different behavior for bias-reduction methods. While they do not improve the scores in the OOD setting (cf. Section 4.2), they achieve strong performances on VQA-CP2. The score of LM stands out, achieving the highest overall accuracy on VQA-CP2 (52.0%) but one of the lowest \textit{acc-tail} on \textit{GQA-OOD} (33%), with similar behavior for RUBi and BP. In short, while VQA-CP2 measures to what extent a VQA model struggles to generalize to a specific unseen distributions, the VQA-CP2 evaluation does not reflect the model behavior on rare question-answer pairs.

\textbf{Discussion and conclusions}

Going beyond previous attempts to reduce the influence of dataset biases in VQA evaluation, our proposed \textit{GQA-OOD} benchmark allows to \textit{both} evaluate (1) whether models have absorbed tendencies in the training data, and (2) how well they generalize to rare/unseen question-answer pairs. This was made possible by (i) a thorough choice of imbalanced question groups, (ii) a new set of metrics and finally, (iii) by allowing to control the amount of distribution shift via the hyper-parameter \(\alpha\). We have provided evidence that the benchmark and metric measure performance and dependency on dataset bias. Our experiments have also shown that neither conventional SOTA VQA models nor dedicated bias reduction methods succeed in all aspects of the proposed evaluation benchmark. We hope that this sheds light on the current shortcomings in vision and language reasoning, and we hope that \textit{GQA-OOD} will contribute to the emergence of new models, less prone to learning spurious biases and more reliable in real-world scenarios.

\textbf{Acknowledgements} — C. Wolf acknowledges support from ANR through grant “Remember” (ANR-20-CHIA-0018).
References


