Beyond Accuracy: A Consolidated Tool for Visual Question Answering Benchmarking

Dirk Väth* and **Pascal Tilli*** and **Ngoc Thang Vu**
University of Stuttgart
Germany
{dirk.vaeth, pascal.tilli, thang.vu}@ims.uni-stuttgart.de

Abstract

On the way towards general Visual Question Answering (VQA) systems that are able to answer arbitrary questions, the need arises for evaluation beyond single-metric leaderboards for specific datasets. To this end, we propose a browser-based benchmarking tool for researchers and challenge organizers, with an API for easy integration of new models and datasets to keep up with the fast-changing landscape of VQA. Our tool helps test generalization capabilities of models across multiple datasets, evaluating not just accuracy, but also performance in more realistic real-world scenarios such as robustness to input noise. Additionally, we include metrics that measure biases and uncertainty, to further explain model behavior. Interactive filtering facilitates discovery of problematic behavior, down to the data sample level. As proof of concept, we perform a case study on four models. We find that state-of-the-art VQA models are optimized for specific tasks or datasets, but fail to generalize even to other in-domain test sets, for example they cannot recognize text in images. Our metrics allow us to quantify which image and question embeddings provide most robustness to a model. All code[1] is publicly available.

1 Introduction

VQA refers to the multi-modal task of answering free-form, natural language questions about images - a task sometimes referred to as a visual Turing test (Xu et al., 2018). The number and variety of datasets for evaluating such systems has continued to increase over the last years (Antol et al., 2015; Hudson and Manning, 2019; Agrawal et al., 2018; Kervadec et al., 2021; Johnson et al., 2017). These datasets aim to test models' abilities with respect to different skills, such as commonsense or external knowledge reasoning, visual reasoning, or reading text in images. Traditionally, evaluation relies

---

* Authors contributed equally
[1] https://github.com/patilli/vqa_benchmarking

Figure 1: Tool landing page with aggregated metrics for each model (larger version in Appendix A). Statistics can be expanded per model to view performance on each dataset.

solely on answering accuracy. However, it is misleading to believe that a single number, like high accuracy on a given benchmark, corresponds to a system's ability to answer arbitrary questions with high quality. Each dataset contains biases which state-of-the-art deep neural networks are prone to exploit, resulting in higher accuracy scores (Goyal et al., 2017; Das et al., 2017; Agrawal et al., 2016; Jabri et al., 2016). Thus, most VQA models, if evaluated on multiple benchmarks at all, are re-trained per dataset to achieve higher numbers in specialized leaderboards. Further shortcomings of current leaderboards include ignoring prediction cost and robustness, as discussed in (Ethayarajh and Jurafsky, 2020) for NLP. In VQA, we need even more specialized evaluation due to the challenges inherent to open-ended, multi-modal reasoning.

In order to successfully develop VQA systems that are able to answer arbitrary questions with human-like performance, we should overcome the previously mentioned shortcomings of current leaderboards as one of the first essential steps. To this end, we propose a benchmarking tool, to our knowledge the first of its kind in the VQA domain, that goes beyond current leaderboard evaluations. It follows the four following principles:

1. Realism To better simulate real-world conditions, we test robustness to semantic-preserving input perturbations to images and questions.

2. Generalizability We include six carefully chosen benchmark datasets, each evaluating different abilities as well as model behavior on changing distributions to test model generalizability. Additionally, we provide easy python interfaces for adding new benchmarking datasets and state-of-the-art models as they arise. The full tool is released under an open-source license.

3. Explainability To provide more insight into model behavior and overcome the problem of single-metric comparisons, we measure scores such as biases and uncertainty, in addition to accuracy.

4. Interactivity Aggregated statistics can be drilled down and filtered, providing interactive exploration of model behavior from a global dataset perspective down to individual data samples. The above functionalities not only support detailed model comparison, but also facilitate development and debugging of models in the VQA domain.

As proof of concept, we integrate several popular and state-of-the-art models from public code repositories and investigate their abilities and weaknesses. Through our case study, we demonstrate that all of these models fail to generalize, even to other in-domain test sets. Our metrics quantify the influence of model architecture decisions, which accuracy cannot capture, such as the effect of image and question embeddings on model robustness.

2 Related Work

VQA Benchmarks Benchmarks often emphasize certain sub-tasks of the general VQA problem. For example, CLEVR (Johnson et al., 2017) tests visual reasoning abilities such as shape recognition and spatial relationships between objects, rather than real-world scenarios. Other approaches change the answer distributions of existing datasets, such as VQA-CP (Agrawal et al., 2018) originating from VQA2 (Goyal et al., 2017) or GQA-OOD from GQA (Kervadec et al., 2021). These changes are intended to mitigate learnable bias. Another approach to mitigating biases was proposed by Hudson and Manning (2019), who created a dataset for real world visual reasoning with a tightly controlled answer distribution. Kervadec et al. (2021) went one step further by analyzing and changing the test sets to evaluate on rarely occurring question-answer pairs rather than on frequent ones. Finally, Li et al. (2021) proposed an adversarial benchmarking dataset to evaluate system robustness.

Metrics For more automated, dataset-level insight, many methods try to analyze single aspects of VQA models. For example, Halbe (2020) use feature attribution to assess the influence of individual question words on model predictions. Das et al. (2017) compare human attention to machine attention to explore whether they focus on the same image regions. To measure robustness w.r.t. question input, Huang et al. (2019) collect a dataset of semantically relevant questions, and rank them by similarity, feeding the top-3 into the network to observe changes in prediction.

Identifying Biases Agrawal et al. (2016) measured multiple properties: generalization to novel instances as selected by dissimilarity, question understanding based on length and POS-tags, and image understanding by selecting a subset of questions which share an answer but have different images. Another approach to analyze bias towards one modality is to train a uni-modal network that excludes the other modality in the training phase (Cadene et al., 2019). However, this requires training one model per modality and cannot be applied easily to all architectures, e.g. to attention mechanisms computed on joint feature spaces.

Robustness and Adversarial Examples Adversarial examples originate from image classification, where perturbations barely visible to a human fool the classifier and cause sudden prediction changes (Szegedy et al., 2014). The same idea was later applied to NLP, where, e.g., appending distracting text to the context in a question answering scenario resulted in F1-score dropping by more than half (Jia and Liang, 2017).

Benchmarking Tools Liu et al. (2021) propose a leaderboard for NLP tasks to compare model performance. They differentiate among several NLP tasks and datasets. All methods are applied post-hoc to analyze the predictions a model outputs. Other benchmarking tools, for example, focus on runtime comparisons (Shi et al., 2016; Liu et al., 2018). Our benchmarking tool not only analyzes system outputs, but also modifies input modalities as well as feature spaces and provides metrics beyond just accuracy.

3 VQA Benchmark Tool

Our tool facilitates global evaluation of model performance w.r.t general and specific tasks (**generalizability**), such as real-world images and

reading capabilities. To simplify integration of future benchmark datasets and models, we provide a well-documented python API. We measure model-inherent properties, such as biases and uncertainty (**explainability**) as well as robustness against input perturbations (**realism**). Model behavior can be further inspected using interactive diagrams and filtering methods for all metrics, supporting sample-level exploration of suspicious model behavior (**interactivity**). All data is collected post-hoc and can be explored in a web application, eliminating the need to re-train existing models.

3.1 Datasets

In this section, we describe the datasets supported out-of-the-box. These serve the principle of benchmarking **generalizability**, by including real-world scenarios as well as task-specific and even synthetic datasets. Where labels are publicly available, we rely on test sets, otherwise on validation sets (marked with *). To reduce computational cost and resources (including environmental impact), we limit each dataset to a maximum of $\sim$15,000 randomly drawn samples , which is referred to in the following paragraphs as a sub-set.

VQA2* This dataset (Goyal et al., 2017) represents a balanced version of the vanilla VQA dataset (Antol et al., 2015). It is intended to mirror real-world scenarios and used as the de-facto baseline for model comparisons.

GQA The GQA dataset (Hudson and Manning, 2019), derived from Visual Genome (Krishna et al., 2017), is designed to test models' real-word visual reasoning capabilities, in particular, robustness, consistency and semantic understanding of vision and language. Similar to VQA2, it also provides a balanced version.

GQA-OOD According to Kervadec et al. (2021), evaluating on rare instead of frequent question-answer pairs is better suited for measuring reasoning abilities. Hence, they introduce the GQA-OOD dataset as a new split of the original GQA dataset to evaluate out-of-distribution questions with imbalanced distributions.

CLEVR* CLEVR (Johnson et al., 2017) is a synthetic dataset, containing images of multiple geometric objects in different colors, materials and arrangements. It aims to test models' visual reasoning abilities by asking questions that require a

model to identify objects based on attributes, presence, count, and spacial relationships.

OK-VQA* Marino et al. (2019) introduce a dataset that requires external knowledge to answer its questions, thereby motivating the integration of additional knowledge pools.

Text VQA* Singh et al. (2019) consider the problem of understanding text in images, an important problem to consider in VQA benchmarking systems, as one application of VQA is intended to aid the visually impaired.

3.2 Metrics

In addition to the evaluation of accuracy across datasets with different distributions and focuses, we implement metrics such as bias of models towards one modality and uncertainty (**explainability**), as well as robustness to noise and adversarial questions (**realism**). All metrics are in range $[0, 100]$.

Accuracy Our tool supports multiple ground truth answers with different scores per sample, providing the flexibility to evaluate for single-answer accuracy as well as e.g. the official VQA2 accuracy measure $acc(a) = min(1, \frac{\#\text{humans}(a)}{3})$ (Antol et al., 2015).

Modality Bias Here, we refer to a model's focus on one modality over the other. Given an image of a zoo and the question *"What animals are shown?"*, if we replace this picture with a fruit bowl, we would expect the model to change its prediction. However, if the prediction stays unaltered, the model's answer cannot depend on the image input. For each prediction on altered inputs (i', q) or (i, q'), we evaluate how many times the answer a' of the replacement pairs is the same as answer a predicted on the original inputs (i, q). Averaging across N trials yields a Monte-Carlo estimate of the bias towards one modality as $1/N \sum_{q'} \mathbb{1}_{f(q,i)=f(q',i)}$. Heuristics, such as ensuring no overlap between subjects and objects of q and q', help reduce cases where q' would just be a rephrasing of q. High values in modality bias correspond to models ignoring input from one modality for many samples, e.g. a question bias of 100 indicates a model that completely ignores images.

Robustness to Noise An important consideration when deploying a model in the real world, is its susceptibility to noise. Noise might be induced naturally by data acquisition methods (VQA-setting:

camera), for example a color-question should not be affected by subtle tone shifts between two cameras. On the question side, semantic-preserving input changes can be induced through paraphrases, synonyms or region-dependent spelling.

For measuring robustness to noise in images, we support adding Gaussian-, Poisson-, salt&pepper- and speckle-noise to the original input image. We also support adding Gaussian noise in image feature space. To obtain a realistic input range, we calculate the standard deviation from 500 randomly sampled image feature vectors. After multiple trials, we average how often the prediction on the noisy inputs matches the original prediction. Applying noise to the original image input tests the robustness of the image feature extractor, which, in many models, is external and thus easy to swap and interesting to compare. On the other hand, applying noise in feature space tests model robustness towards noisy feature extractors.

Measuring robustness to question noise is done by adding Gaussian noise in embedding space, a reasonable approach under the assumption that similar vectors in embedding space have similar meaning. Again, multiple trials are performed.

High values in robustness correspond to models unaffected by noise in one modality for many samples, e.g. a question robustness of 100 indicates a model that never changed its predictions due to noise added in question embedding space.

Robustness to Adversarial Questions Semantically Equivalent Adversarial Ruless (SEARs) alter textual input according to a set of rules, while preserving original semantics (Ribeiro et al., 2018). For the questions in the VQA dataset, the authors come up with the four rules that most affect the predictions in their tests, using a combination of Part-of-Speech (POS)-Tags and vocabulary entries:

- *Rule 1* WP VBZ → WP's

- *Rule 2* What NOUN → Which NOUN

- *Rule 3* color → colour

- *Rule 4* ADV VBZ → ADV's

High values in robustness against SEARs correspond to models unaffected by adversarial questions, e.g. a robustness of 100 indicates a model that never changed its predictions due to the application of any of the above rules. Therefore, higher values are preferable.

Uncertainty To measure model certainty, we leverage the dropout-based Monte-Carlo method (Gal and Ghahramani, 2016). Forwarding a sample multiple times with active dropout, the averaged output vector $\frac{1}{N}\sum_{n=1}^{N} f(x)$.

3.3 Views

We support inspection of the included metrics at different levels of granularity, from comparisons across multiple datasets to filtering of individual samples (**interactivity**). On each level, we supplement the accuracy measure by additional metrics helpful for understanding and debugging VQA models (**explainability**).

Gobal View The global view (see figure 1) acts as the main entry to our tool. At a glance, it shows a leaderboard with statistics averaged on all datasets, providing users with an impression of the models' performance and properties across tasks and distributions. All columns are sortable to allow easy

Dataset	Accuracy	Bias Image	Bias Question	Noise Robustness Image (Imagespace)	Noise Robustness Image (Featurespace)	Noise Robustness Question	Robustness SEARs	Uncertainty
BAN-8	38.62	5.21	3.31	99.99	82.22	70.09	53.53	66.10
GQA-OOD-HEAD	40.97	3.67	2.26	100.00	84.51	75.28	33.16	66.59
OK-VQA	32.88	0.39	0.46	99.98	79.10	64.16	75.63	76.25
VQA2	81.60	7.82	0.86	100.00	87.79	79.57	70.24	65.16
GQA-OOD-TAIL	35.65	2.20	1.36	100.00	84.24	74.57	32.45	65.48
CLEVR	28.64	15.65	12.92	99.96	86.22	67.44	74.52	46.75
TextVQA	8.40	2.07	2.11	99.96	67.83	53.08	71.01	80.47
GQA-OOD-ALL	38.95	2.93	1.89	100.00	83.98	75.11	32.86	66.39
GQA	41.86	6.95	4.58	99.98	84.08	71.50	38.36	61.71

Figure 2: Expanded details on the overview page (larger version in Appendix A).

comparisons between models for each metric. Each row in the overview table describes a model's average performance and can be expanded to provide additional information on a per-dataset level (see figure 2).

Metrics View Clicking a model row in the global view navigates to the metrics view, which provides graphs on all metrics and datasets for the selected model in detail (see figure 3). Users have the choice to change dataset and metric via selection boxes. For easy comparison between datasets of different sizes, all values are recorded in percentages of the dataset.

Filter View Our tool supports searching for patterns of suspicious model behavior by providing a
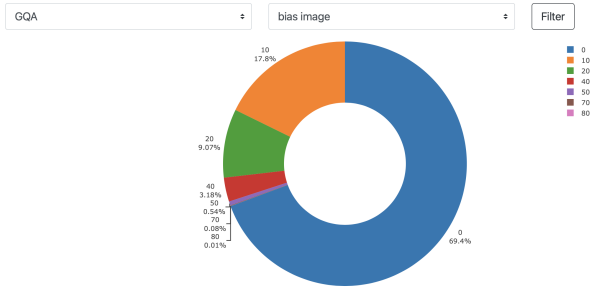
Figure 3: Metric view, showing bias towards images on the GQA dataset for the MDETR model (larger version in Appendix A).

Figure 5: Sample view with *Image Bias Word Space* metric. Similar cards exist for each metric (larger version in Appendix A).

filter view (see figure 4). Once model, dataset and metric are selected, users are presented a list of all samples within the chosen range. The range can be adjusted using a slider, which updates the list of matching data samples in real-time.
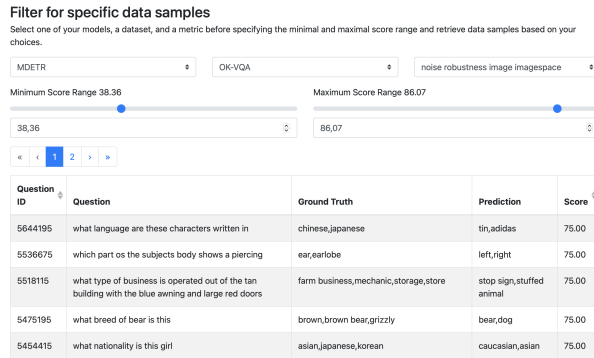

Figure 4: Filter view. Enables filtering for unexpected model behavior on sample level (larger version in Appendix A).

Sample View Finally, once the desired range of samples has been filtered, clicking a data sample navigates to the sample view (see figure 5). There, the original input image and question are displayed, along with ground truth and the model's top-3 predictions. Additionally, the scores and answers for each single metric are shown. For example, if applying noise to the image changed the prediction multiple times, we show all the answers that were predicted using those noisy inputs.

4 Case Study

As a case study, we explore a range of models from well-established, previously high ranking entries in the VQA2 competition to more recent, transformer-based architectures and report the insights we gained by inspecting them with our tool.
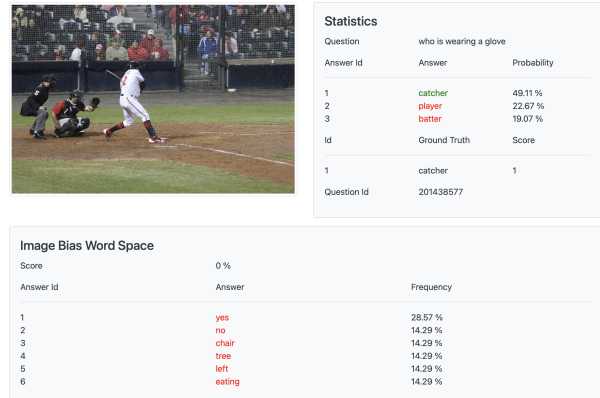
4.1 Evaluated Models

We chose a widely used VQA-baseline BAN, two transformer-based architectures MDETR and MCAN, and MMNASNET.

BAN (Kim et al., 2018) is a strong baseline using bilinear attention. It won third place in the VQA2 2018 challenge and was still in the top-10 entries in 2019. We use the 8-layer version.

MCAN (Yu et al., 2019) improves BAN with a co-attention feature fusion mechanism.

MMNASNET Yu et al. (2020) is a more recent state-of-the-art model constructed using neural architecture search. It is one of the top-10 entries of the VQA2 2020 challenge.

MDETR (Kamath et al., 2021) is a state-of-the art transformer using more recent question (Liu et al., 2019) and image embedding approaches (Carion et al., 2020). MDETR achieves competitive accuracies on both GQA and CLEVR.

4.2 Results and Lessons Learned

Table 1 contains the aggregated results of all models, averaged across the development (sub-)splits of all datasets. For details about the computation of each metric, see section 3.2. Table 2 shows model accuracy per dataset. Unsurprisingly, models performed best when evaluated on the development (sub-)split of the dataset they were originally trained on, and worse on datasets they were not trained on. These performance drops are observable for all models, suggesting that VQA models cannot yet generalize well across tasks. Low performance of current highly ranked VQA models

Average Results								
Model	Accuracy	Modality Bias		Robustness Image		Robustness	SEARs	Uncer-	Parameters
		Image	Quest.	Image	Feature	Question		tainty	
MCAN	**41.30**	4.83	2.57	99.98	81.09	63.14	58.45	31.77	201,723,191
MMNASNET	40.80	**4.67**	**2.51**	**99.99**	79.90	65.18	59.65	31.75	211,166,871
BAN-8	38.62	5.21	3.31	**99.99**	82.22	70.09	53.53	66.10	112,167,258
MDETR	38.82	4.77	2.60	36.33	**90.10**	**100.00**	**100.00**	**24.58**	185,847,022

Table 1: Average results of our evaluated models across all development (sub-)splits. We split the columns modality bias and robustness against image modifications into two sub-columns. These columns should be read as the modality bias (lower values are better) measured for the image space and question space. Robustness (higher values are better) against image changes is divided into alterations on the image itself as well as modifications inside image feature space . All metrics are in range $[0, 100]$.

on new datasets can partially be attributed to their fixed answer spaces. This implies the need for more research into systems that are able to generate answers instead of treating VQA as a multiple-choice problem. However, even changing distributions of the same dataset leads to a large performance drop, as we observe, for example, in GQA and its out-of-distribution variants, GQA-OOD-HEAD and GQA-OOD-TAIL. By swapping the original GQA dataset for the GQA-OOD-TAIL distribution, MDETR accuracy decreased by more than $11, 6\%$. That out-of-distribution testing causes such high losses in accuracy indicates models are still relying on biases learned from the training dataset.

All systems struggled to read text in images, in fact, the highest accuracy score on TextVQA was only 8.81%, achieved by MMNASNET. This might be improved by extending existing VQA-architectures with additional inputs, e.g. from optical character recognition, or adapting the training of currently used image feature extractors.

Applying noise in image space has almost no impact on models using bottom-up-topdown feature extraction (Anderson et al., 2018), in contrast to MDETR, the only model using an alternative approach. In feature space, all models are similarly stable, which could imply that the feature extractor in MDETR could be made more robust by augmenting training with noisy images.

All models show highest modality bias and low accuracy on the CLEVR dataset. Given that no models were trained on synthetic images or questions involving such complex selection and spatial reasoning, this hints at the models not understanding either modality well. Inspection using the filter view on modality biases provides more evidence of understanding problems here, showing that for example BAN-8 nearly always guesses *yes* or *no*, regardless of the question asked or the image given.

In general, BAN-8 displays the highest modality bias, indicating more recent models have become better at jointly reasoning over image and text.

SEAR and question robustness metrics show that RoBERTa (Liu et al., 2019) provides substantial robustness to question perturbation; there were zero cases causing MDETR to change predictions, suggesting that context-aware embeddings should be a standard consideration for future VQA models.

Our metrics show that state-of-the-art VQA models are optimized for specific tasks or datasets, but fail to generalize even across other in-domain datasets. In order to be successful in real-world applications, systems must demonstrate a variety of abilities, not merely good performance on a single-purpose test set.

5 Conclusion

Our proposed benchmarking tool is the first of its kind in the domain of VQA and addresses the problems of current single-metric leaderboards in this domain. It provides easy to use and fast comparison of integrated models on a global level. The performance of each model is evaluated across multiple special-purpose as well as general-purpose datasets to test generalizability and capabilities. Each model can be quantified by metrics such as accuracy, biases, robustness, and uncertainty, revealing strengths and weaknesses w.r.t to given tasks, i.e. measuring the properties models offer as well as their real-world robustness. Exploration via filtering can be used to identify suspicious behaviour down to single data sample level. Through this, our tool provides deeper insights into the strengths and weaknesses of each model across tasks and metrics and how architectural choices can affect behavior, encouraging researchers to develop VQA systems with rich sets of abilities that stand up to real-world

Model	CLEVR	GQA	GQA-OOD-ALL	GQA-OOD-HEAD	GQA-OOD-TAIL	OK-VQA	TextVQA	VQA2
MCAN	32.87	44.58	41.67	44.78	36.59	35.46	8.49	**85.94**
MMNASNET	31.95	44.50	40.24	42.01	37.35	35.09	8.81	**86.51**
BAN-8	28.64	41.86	38.95	40.97	35.65	32.88	8.40	**81.60**
MDETR	25.44	**61.42**	55.76	59.43	49.76	9.83	4.70	44.22

Table 2: Accuracy across the development (sub-)splits of different datasets. Bold entries mark best accuracy per model and coincides in all cases with the dataset it was trained on.

environments. The open-source tool itself can be installed as a package and extended with new models, datasets and metrics using our python API.

In the future, we plan to extend this tool with new datasets as they are released. Moreover, we are looking for more metrics for model evaluation as well as more detailed dataset analysis, e.g. answer space overlap. Last but not least, interactivity could be extended towards live model feedback, allowing to change inputs, e.g. the image noise level, and observe model outputs at runtime.

6 Acknowledgement

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Advances in neural information processing systems*, pages 841–852.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90 – 100. Language in Vision.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboard design. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Shaunak Halbe. 2020. Exploring weaknesses of vqa models through attribution driven insights. *ACL 2020*, page 64.

Jia-Hong Huang, Cuong Duc Dao, Modar Alfadly, and Bernard Ghanem. 2019. A novel framework for robustness analysis of visual qa models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8449–8456.

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

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 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.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. 2021. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 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*, pages 2776–2785.

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

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

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. *arXiv preprint arXiv:2106.00245*.

Ling Liu, Yanzhao Wu, Wenqi Wei, Wenqi Cao, Semih Sahin, and Qi Zhang. 2018. Benchmarking deep learning frameworks: Design considerations, metrics and beyond. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 1258–1269. IEEE.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaicheng Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. Explainaboard: An explainable leaderboard for nlp. *arXiv preprint arXiv:2104.06387*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 856–865.

Shaohuai Shi, Qiang Wang, Pengfei Xu, and Xiaowen Chu. 2016. Benchmarking state-of-the-art deep learning software tools. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pages 99–104. IEEE.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. 2018. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4951–4961.

Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian. 2020. Deep multimodal neural architecture search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3743–3752.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6281–6290.
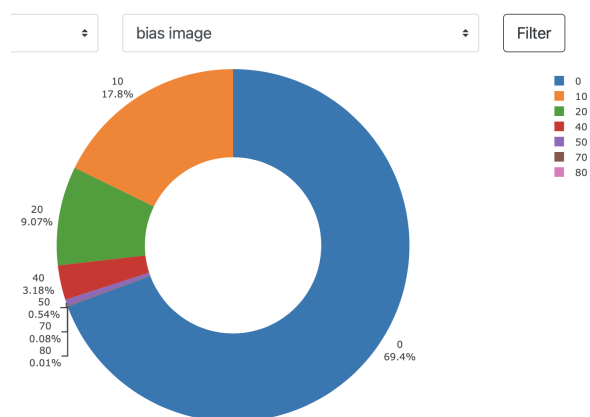
A Appendix

Figure 6: Zoom into Figure 3.

Model	Accuracy	Bias Image	Bias Question	Noise Robustness Image (Imagespace)	Noise Robustness Image (Featurespace)	Noise Robustness Question	Robustness SEARs	Uncertainty	Parameters	Details
MCAN-LARGE	41.30	4.83	2.57	99.98	81.09	63.14	58.45	31.77	201.723.191	Show Details
MMNASNET-LARGE	40.80	4.67	2.51	99.99	79.90	65.18	59.65	31.75	211.166.871	Show Details
BAN-8	38.62	5.21	3.31	99.99	82.22	70.09	53.53	66.10	112.167.258	Show Details
MDETR	38.82	4.77	2.60	36.33	90.10	100.00	100.00	24.58	185.847.022	Show Details

Figure 7: Figure 1 in full size.

Filter for specific data samples

Select one of your models, a dataset, and a metric before specifying the minimal and maximal score range and retrieve data samples based on your choices.

MDETR ⇕	OK-VQA ⇕	noise robustness image imagespace ⇕

Minimum Score Range 38.36

38,36

Maximum Score Range 86.07

86,07

« ‹ 1 2 › »

Question ID	Question	Ground Truth	Prediction	Score
5644195	what language are these characters written in	chinese,japanese	tin,adidas	75.00
5536675	which part os the subjects body shows a piercing	ear,earlobe	left,right	75.00
5518115	what type of business is operated out of the tan building with the blue awning and large red doors	farm business,mechanic,storage,store	stop sign,stuffed animal	75.00
5475195	what breed of bear is this	brown,brown bear,grizzly	bear,dog	75.00
5454415	what nationality is this girl	asian,japanese,korean	caucasian,asian	75.00

Figure 8: Figure 4 in full size.
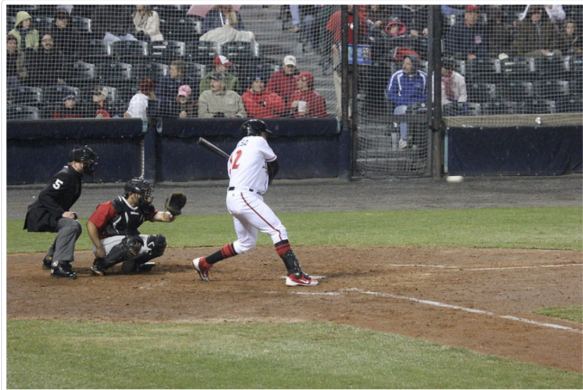
Statistics

Question	who is wearing a glove	
Answer Id	Answer	Probability
1	catcher	49.11 %
2	player	22.67 %
3	batter	19.07 %
Id	Ground Truth	Score
1	catcher	1
Question Id	201438577	

Image Bias Word Space

Score	0 %	
Answer Id	Answer	Frequency
1	yes	28.57 %
2	no	14.29 %
3	chair	14.29 %
4	tree	14.29 %
5	left	14.29 %
6	eating	14.29 %

Figure 9: Figure 5 in full size.