Beyond Question-Based Biases: Assessing Multimodal Shortcut Learning in Visual Question Answering Supplementary Material

1. Supplementary material

1.1. Additional statistics about shortcuts

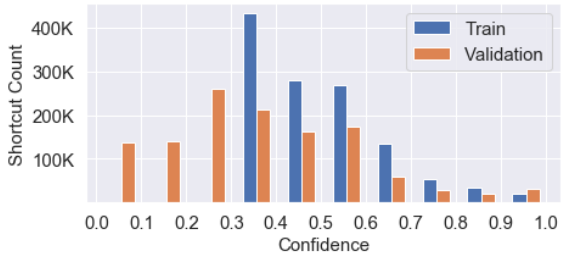

Figure 1. Histogram of shortcuts binned per confidence on the VQA v2 training and validation sets. Our shortcuts are detected on the training set and selected to have a confidence above 30%. Even though their confidence could be expected to be lower on the validation set, it still is above 30% for a large number of them, indicating that the selection transfers well to the validation set.

Confidence distribution on training and unseen data Here we show that shortcuts detected on the VQA v2 training set transfer to the validation set. In Figure 1, we display the confidence distribution of these shortcuts. As told earlier, we only consider shortcuts that reach a confidence greater than 0.3 on the training set. The number of shortcuts decreases when the confidence increases. It is expected to find fewer shortcuts with higher levels of confidence due to the collection procedure of VQA v2 which focused on reducing the amount of data biases and shortcuts. We evaluate on the validation set the same shortcuts detected on the training set and also display the confidence distribution. We show that our shortcuts are predictive on both training data, and unseen data that follows the training set distribution. The number of shortcuts that reach a confidence between 0.9 and 1.0 is even higher on the validation set than on the training set. The confidences are overall slightly lower on the validation set, but a large number of them are still above 0.3, indicating that they generalize to new examples from the same distribution. The great majority of shortcuts, which obtain a confidence lower than 1.0, allows finding examples that contradict them by leading to the wrong answers. We manually verified by looking at these examples that only a minority are wrongly annotated or ambiguous, most of them are counterexamples. These counterexamples are the core of our approach to assess the VQA model's reliance on shortcuts.

Distribution of examples per question-type In Figure 2, we display the distribution of examples per question type, and their split between the Easy and the Counterexamples split. We show that examples of a question-type that can be answered by *yes* or *no*, such as *is*, *are*, *does*, *do*, mostly belong to the Easy subset. Examples of a question-type beginning by *what*, *where* or *why* mostly belong to the Counterexamples subset. These examples need to be answered using a richer vocabulary than *yes* or *no*. Examples of a question-type beginning by *how* belong to both subset.

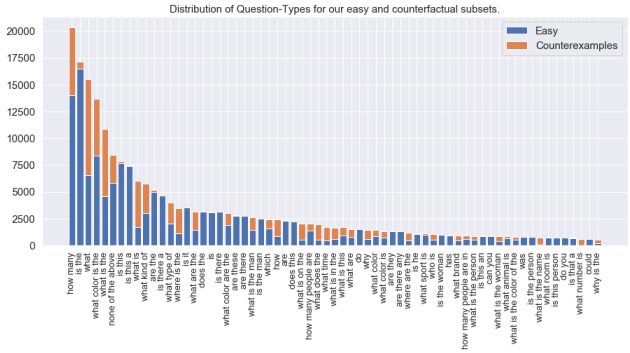

Figure 2. Distribution of the number of examples per question type. Examples associated to our Counterexamples subset are matched by some shortcuts, but no shortcut leads to the correct answer. Examples associated to our Easy subset are matched by at least one shortcut that leads to the correct answer.

Distribution of examples per answer type In Figure 3, we display the distribution of examples in our two subsets per answer type. We see that most yes-no questions are going in the Easy subset, as they are correctly predicted by some rules. On the contrary, for the two other answer types,

examples are more evenly distributed between the Easy and Counterexamples subsets.

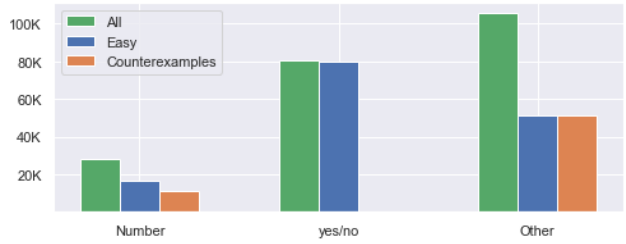

Figure 3. Number of examples per answer type. "All" corresponds to all the examples from the VQA v2 validation set. Among them, examples associated to our "Counterexamples" subset are matched by some shortcuts, but none of these shortcuts leads to the correct answer. Inversely, examples associated to our Easy subset are matched by at least one shortcut that leads to the correct answer.

1.2. Examples that are not matched by any rule

In Figure 4, we display some representative examples that are neither in the Easy subset nor in the Counterexamples subset. These examples are not matched by any antecedent of our rules. Their input might be unusual. We do not add these examples to our Counterexamples subset, as they do not contradict the shortcuts we found. We discard them entirely from our analysis. There consists in about 3K of examples.

Figure 4. Representative instances of image-question-answer examples that are not matched by any of our shortcuts. These examples have unusual questions, images or answers.

Results with ground-truth visual labels We report in Table 1 the results of our analysis with ground-truth visual labels from the COCO [7] dataset, instead of labels detected with Faster R-CNN. We make similar observations to the main experiments of the paper: bias-reductions methods often degrade performances, on both easy and counterexamples split. A few methods slightly improve the counterexamples score, but much less than on VQA-CP. The only method which improves both overall and counterexamples scores is LfF [9]. We observed similar results on the dataset with detected labels, reported in Table 1 of the main paper.

Results on VQA v1 We report in Table 2 the results of our analysis on the VQA v1 dataset. We observe similar results as in Table 1 from the main paper. Most bias-reduction methods degrade performances on the counterexamples split, and only LfF [9] improves performances on all three splits.

	Approaches	Overall	**Counterexamples (ours)**	**Easy (ours)**
	Number of examples	*214,354*	*63,925*	*135,324*
Baselines	Shortcuts	42.14	0.43	65.95
	Image-Only	23.70	2.92	35.39
	Question-Only	44.12	13.98	60.88
VQA models	SAN [12] – *grid features*	55.61	28.99	70.04
	UpDown [1]	63.52 (+0.00)	37.77 (+0.00)	77.52 (+0.00)
	BLOCK [2]	63.89	37.06	78.52
	VilBERT [8] – *pretrained*†	67.77	43.32	81.27
Bias-reduction methods	*UpDown [1] is used as a base architecture for bias-reduction methods*			
	RUBi [3]	61.88 (-1.64)	36.05 (-1.72)	75.84 (-1.68)
	LMH + RMFE [6]	60.12 (-3.40)	34.97 (-2.80)	73.80 (-3.72)
	ESR [10]	62.96 (-0.56)	37.22 (-0.55)	76.98 (-0.54)
	LMH [5]	61.15 (-2.37)	37.82 (+0.05)	73.91 (-3.61)
	LfF [9]	63.57 (+0.05)	38.18 (+0.41)	77.44 (-0.08)
	LMH+CSS [4]	53.55 (-9.97)	37.27 (-0.50)	62.30 (-15.22)
	RandImg [11]	63.34 (-0.18)	38.13 (+0.36)	77.05 (-0.47)

Table 1. Results of our VQA-CE evaluation protocol with **ground-truth visual labels**. We report accuracies on VQA v2 full validation set and on our two subsets: **Counterexamples** and **Easy** examples. We re-implemented all models and bias-reduction methods. †VilBERT is pretrained on Conceptual Caption and fine-tuned on VQA v2 training set. Scores in (green) and (red) are relative to UpDown [1].

	Approaches	Overall	**Counterexamples (ours)**	**Easy (ours)**
	Number of examples	*121,512*	*40,052*	*80,539*
Baselines	Shortcuts	44.71	0.05	67.35
	Image-Only	24.39	1.75	35.83
	Question-Only	49.20	13.48	67.27
	SAN [12] – *grid features*	58.35	26.09	74.58
	UpDown [1]	62.83 (+0.00)	31.71 (+0.00)	78.49 (+0.00)
Bias-reduction methods	*UpDown [1] is used as a base architecture for bias-reduction methods*			
	RUBi [3]	55.82 (-7.01)	23.87 (-7.84)	71.90 (-6.59)
	LMH + RMFE [6]	62.97 (+0.14)	31.09 (-0.62)	79.02 (+0.53)
	ESR [10]	63.03 (+0.20)	31.50 (-0.21)	78.91 (+0.42)
	LMH [5]	59.74 (-3.09)	32.80 (+1.09)	73.30 (-5.19)
	LfF [9]	63.26 (+0.43)	32.05 (+0.34)	78.97 (+0.48)
	RandImg [11]	62.87 (+0.04)	31.09 (-0.62)	78.87 (+0.38)

Table 2. Results of our VQA-CE evaluation protocol on **VQA v1** full validation set and on our two subsets: **Counterexamples** and **Easy** examples. We re-implemented all models and bias-reduction methods. Scores in (green) and (red) are relative to UpDown [1].

1.3. Rules with supporting examples and counterexamples

In Figure 5, we display some counterexamples to some rules displayed in Table 2 of the main paper. Some of those examples are "true" counterexamples, where the input does match the rule's antecedent, but the answer is different. For instance, in the first example of the first rule, the question is actually about the clothes and not the sport, and the man is dressed in a basketball outfit. On the contrary, some examples are there due to an incorrect object detection: in the second example of the first rule, the object detection module detected a skateboard instead of a scooter. Thus, the example is incorrectly matched.

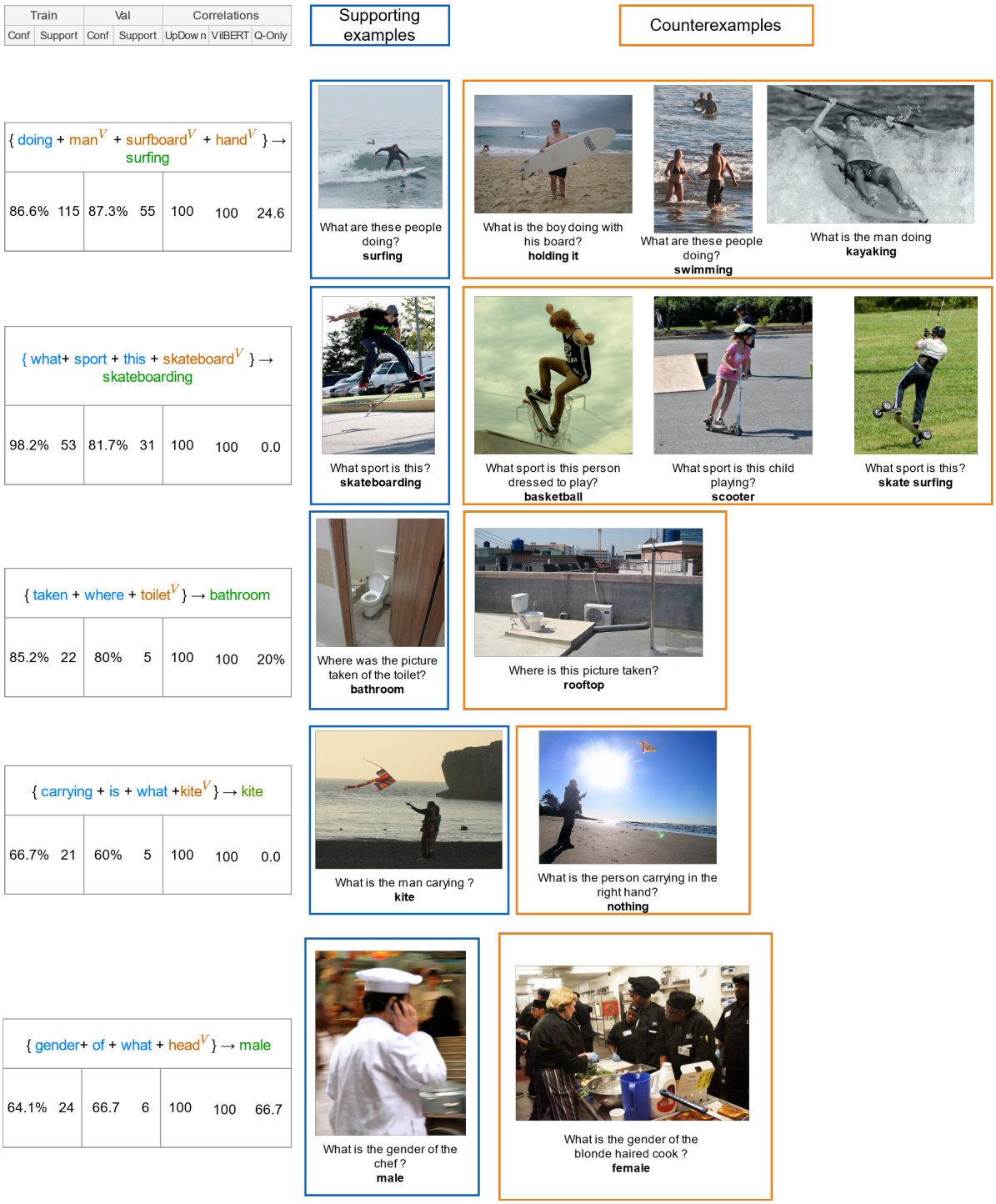

Figure 5. Instances of shortcuts that are highly correlated with VQA models' predictions. We display their antecedent made of words from the question and objects^V from the image, and their **answer**. Their support, i.e. number of examples matched by the antecedent, and confidence, i.e. percentage of correct answers among them, have been calculated on the VQA v2 training and validation sets. We report the correlation coefficients of their predictions with those of three VQA models: UpDown [3] that uses an object detector, VilBERT [31] that has been pretrained on a large dataset, and Q-only [21] that only uses the question. We also display some supporting examples, in blue, and counterexamples, in orange.

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 (CVPR)*, 2018. 3

[2] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 3

[3] Remi Cadene, Corentin Dancette, Hedi Ben-Younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing Unimodal Biases for Visual Question Answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[4] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[5] 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. 3

[6] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[7] 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 *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 2

[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[9] Junhyun Nam, Hyuntak Cha, Sung-Soo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 3

[10] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for vqa. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8172–8181, 2020. 3

[11] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *arXiv preprint arXiv:2005.09241*, 2020. 3

[12] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3