RUBi: Reducing Unimodal Biases in Visual Question Answering

Remi Cadene [1*] **Corentin Dancette** [1*] **Hedi Ben-younes** [1]
Matthieu Cord [1] **Devi Parikh** [2,3]
[1] Sorbonne University
[2] Facebook AI Research
[3] Georgia Institute of Technology

Abstract

Visual Question Answering (VQA) is the task of answering questions about an image. Some VQA models often exploit unimodal biases to provide the correct answer without using the image information. As a result, they suffer from a huge drop in performance when evaluated on data outside their training set distribution. This critical issue makes them unsuitable for real-world settings.

We propose RUBi, a new learning strategy to reduce biases in any VQA model. It reduces the importance of the most biased examples, i.e. examples that can be correctly classified without looking at the image. It implicitly forces the VQA model to use the two input modalities instead of relying on statistical regularities between the question and the answer. We leverage a question-only model that captures the language biases by identifying when these unwanted regularities are used. It prevents the base VQA model from learning them by influencing its predictions. This leads to dynamically adjusting the loss in order to compensate for biases. We validate our contributions by surpassing the current state-of-the-art results on VQA-CP v2. This dataset is specifically designed to assess the robustness of VQA models when exposed to different question biases at test time than what was seen during training.

Our code is available: github.com/cdancette/rubi.bootstrap.pytorch

1 Introduction

The recent Deep Learning success in computer vision [1] and natural language understanding [2] allowed researchers to tackle multimodal tasks that combine visual and textual modalities [3–7]. Among these tasks, Visual Question Answering (VQA) attracts increasing attention. The goal of the VQA task is to answer a question about an image. It requires a high-level understanding of the visual scene and the question, but also to ground the textual concepts in the image and to use both modalities adequately. Solving the VQA task could have tremendous impacts on real-world applications such as aiding visually impaired users in understanding their physical and online surroundings, searching through large quantities of visual data via natural language interfaces, or even communicating with robots using more efficient and intuitive interfaces.

Several large real image VQA datasets have recently emerged [8–14]. Each one of them targets specific abilities that a VQA model would need to be used in real-world settings such as fine-grained recognition, object detection, counting, activity recognition, commonsense reasoning, etc. Current end-to-end VQA models [15–22] achieve impressive results on most of these benchmarks and are even able to surpass the human accuracy on a specific benchmark accounting for compositional

*Equal contribution

Preprint. Under review.

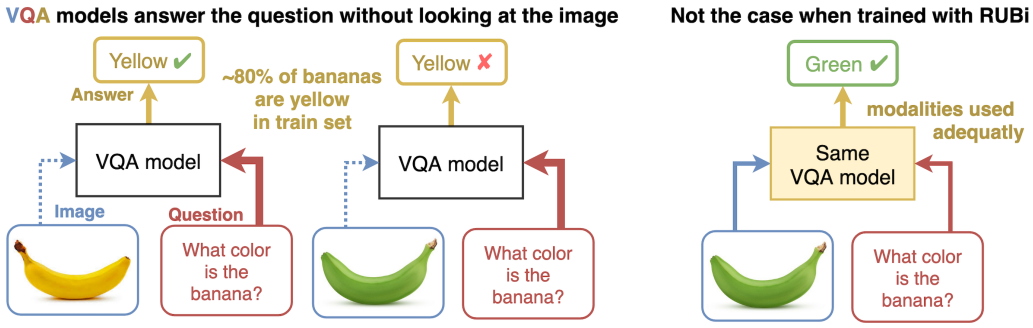


Figure 1: Our RUBi approach aims at reducing the amount of unimodal biases learned by a VQA model during training. As depicted, current VQA models often rely on unwanted statistical correlations between the question and the answer instead of using both modalities.

reasoning [23]. However, it has been shown that they tend to exploit statistical regularities between answer occurrences and certain patterns in the question [24, 10, 25, 23, 13]. While they are designed to merge information from both modalities, in practice they often answer without considering the image modality. When most of the bananas are *yellow*, a model does not need to learn the correct behavior to reach a high accuracy for questions asking about the color of bananas. Instead of looking at the image, detecting a banana and assessing its color, it is much easier to learn from the statistical shortcut linking the words *what*, *color* and *bananas* with the most occurring answer *yellow*.

One way to quantify the amount of statistical shortcuts from each modality is to train unimodal models. For instance, a question-only model trained on the widely used VQA v2 dataset [9] predicts the correct answer approximately 44% of the time over the test set. VQA models are not discouraged to exploit these statistical shortcuts from the question modality, because their training set often follows the same distribution as their testing set. However, when evaluated on a test set that displays different statistical regularities, they usually suffer from a significant drop in accuracy [10, 25]. Unfortunately, these statistical regularities are hard to avoid when collecting real datasets. As illustrated in Figure 1, there is a crucial need to develop new strategies to reduce the amount of biases coming from the question modality in order to learn better behaviors.

We propose RUBi, a training strategy to reduce the amount of biases learned by VQA models. Our strategy reduces the importance of the most biased examples, i.e. examples that can be correctly classified without looking at the image modality. It implicitly forces the VQA model to use the two input modalities instead of relying on statistical regularities between the question and the answer. We take advantage of the fact that question-only models are by design biased towards the question modality. We add a question-only branch on top of a base VQA model during training only. This branch influences the VQA model, dynamically adjusting the loss to compensate for biases. As a result, the gradients backpropagated through the VQA model are reduced for the most biased examples and increased for the less biased. At the end of the training, we simply remove the question-only branch.

We run extensive experiments on VQA-CP v2 [10] and demonstrate the ability of RUBi to surpass current state-of-the-art results from a significant margin. This dataset has been specifically designed to assess the capacity of VQA models to be robust to biases from the question modality. We show that our RUBi learning framework provides gains when applied on several VQA architectures such as Stacked Attention Networks [26] and Top-Down Bottom-Up Attention [15]. We also show that RUBi is competitive on the standard VQA v2 dataset [9] when compared to approaches that reduce unimodal biases.

2 Related work

Real-world datasets display some form of inherent biases due to their collection process [27–29]. As a result, machine learning models tend to reflect these biases because they capture often undesirable correlations between the inputs and the ground truth annotations [30–32]. Procedures exist to identify certain kinds of biases and to reduce them. For instance, some methods are focused on gender biases

[33, 34], some others on the human reporting biases [35], and also on the shift in distribution between lab-curated data and real-world data [36]. In the following, we discuss about related works that assess and reduce unimodal biases learned by VQA models.

Assessing unimodal biases in datasets and models Despite being designed to merge the two input modalities, it has been found that VQA models often rely on superficial correlations between inputs from one modality and the answers without considering the other modality [37, 32]. An interesting way to quantify the amount of unimodal biases that can potentially be learned by a VQA model consists in training models using only one of the two modalities [8, 9]. The question-only model is a particularly strong baseline because of the large amount of statistical regularities that can be leveraged from the question modality. With the RUBi learning strategy, we take advantage of this baseline model to prevent VQA models from learning question biases.

Unfortunately, biased models that exploit statistical shortcuts from one modality usually reach impressive accuracy on most of the current benchmarks. VQA-CP v2 and VQA-CP v1 [10] were recently introduced as diagnostic datasets containing different answer distributions for each question-type between train and test splits. Consequentially, models biased towards the question modality fail on these benchmarks. We use the more challenging VQA-CP v2 dataset extensively in order to show the ability of our approach to reduce the learning of biases coming from the question modality.

Balancing datasets to avoid unimodal biases Once the unimodal biases have been identified, one method to overcome these biases is to create more balanced datasets. For instance, the synthetic datasets for VQA [23, 13] minimize question-conditional biases via rejection sampling within families of related questions to avoid simple shortcuts to the correct answer.

Doing rejection sampling in real VQA datasets is usually not possible due to the cost of annotations. Another solution is to collect complementary examples to increase the difficulty of the task. For instance, VQA v2 [9] has been introduced to weaken language priors in the VQA v1 dataset [8] by identifying complementary images. For a given VQA v1 question, VQA v2 also contains a similar image with a different answer to the same question. However, even with this additional balancing, statistical biases from the question remain and can be leveraged [10]. That is why we propose an approach to reduce unimodal biases during training. It is designed to learn unbiased models from biased datasets. Our learning strategy dynamically modifies the loss values to reduce biases from the question. By doing so, we reduce the importance of certain examples, similarly to the rejection sampling approach, while increasing the importance of complementary examples which are already in the training set.

Architectures and learning strategies to reduce unimodal biases In parallel of these previous works on balancing datasets, an important effort has been carried out to design VQA models to overcome biases from datasets. [10] proposed a hand-designed architecture called Grounded VQA model (GVQA). It breaks the task of VQA down into a first step of locating and recognizing the visual regions needed to answer the question, and a second step of identifying the space of plausible answers based on a question-only branch. This approach requires training multiple sub-models separately. In contrast, our learning strategy is end-to-end. Their complex design is not straightforward to apply on different architectures while our approach is model-agnostic. While we rely on a question-only branch, we remove it at the end of the training.

The work most related to ours in terms of approach is [25]. The authors propose a learning strategy to overcome language priors in VQA models. They first introduce an adversary question-only branch. It takes as input the question encoding from the VQA model and produces a question-only loss. They use a gradient negation of this loss to discourage the question encoder to capture unwanted biases that could be exploited by the VQA model. They also propose a loss based on the difference of entropies between the VQA model and the question-only branch output distributions. These two losses are only backpropagated to the question encoder. In contrast, our learning strategy targets the full VQA model parameters to reduce the impact of unwanted biases more effectively. Instead of relying on these two additional losses, we use the question-only branch to dynamically adapt the value of the classification loss in order to reduce the learning of biases in the VQA model.

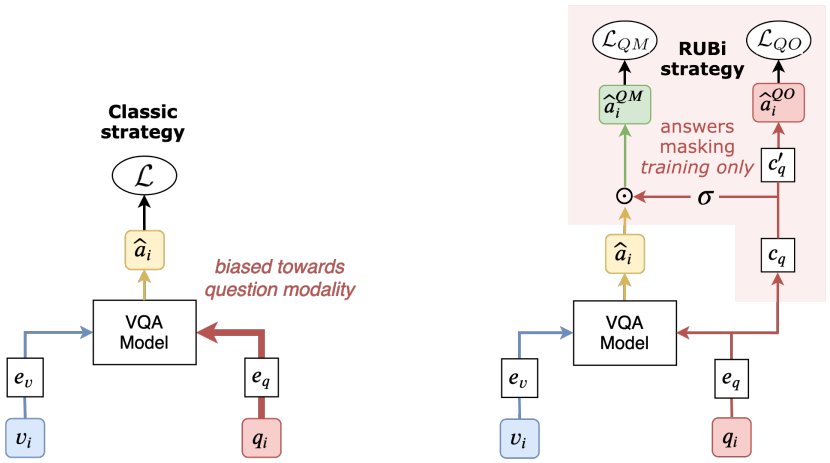


Figure 2: Visual comparison between the classical learning strategy of a VQA model and our RUBi learning strategy. The red highlighted modules are removed at the end of the training. The ouput $\hat{a}_i$ is used as the final prediction.

3 Reducing Unimodal Biases Approach

We consider the common formulation of the Visual Question Answering (VQA) task as a multi-class classification problem. Given a dataset $\mathcal{D}$ consisting of n triplets $(v_i, q_i, a_i)_{i\in[1,n]}$ with $v_i \in \mathcal{V}$ an image, $q_i \in \mathcal{Q}$ a question in natural language and $a_i \in \mathcal{A}$ an answer, one must optimize the parameters θ of the function $f : \mathcal{V} \times \mathcal{Q} \rightarrow \mathbb{R}^{|\mathcal{A}|}$ to produce accurate predictions. For a single example, VQA models use an image encoder $e_v : \mathcal{V} \rightarrow \mathbb{R}^{n_v \times d_v}$ to output a set of n_v vectors of dimension d_v, a question encoder $e_q : \mathcal{Q} \rightarrow \mathbb{R}^{n_q \times d_q}$ to output a set of n_q vectors of dimension d_q, a multimodal fusion $m : \mathbb{R}^{n_v \times d_v} \times \mathbb{R}^{n_q \times d_q} \rightarrow \mathbb{R}^{d_m}$, and a classifier $c : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^{|\mathcal{A}|}$. These functions are composed as follows:

$$f(v_i, q_i) = c(m(e_v(v_i), e_q(q_i))) \tag{1}$$

Each one of them can be defined to instantiate most of the state of the art models, such as [26, 38, 19, 39, 17, 40, 16] to cite a few.

Classical learning strategy and pitfall The classical learning strategy of VQA models, depicted in Figure 2, consists in minimizing the standard cross-entropy criterion over a dataset of size n.

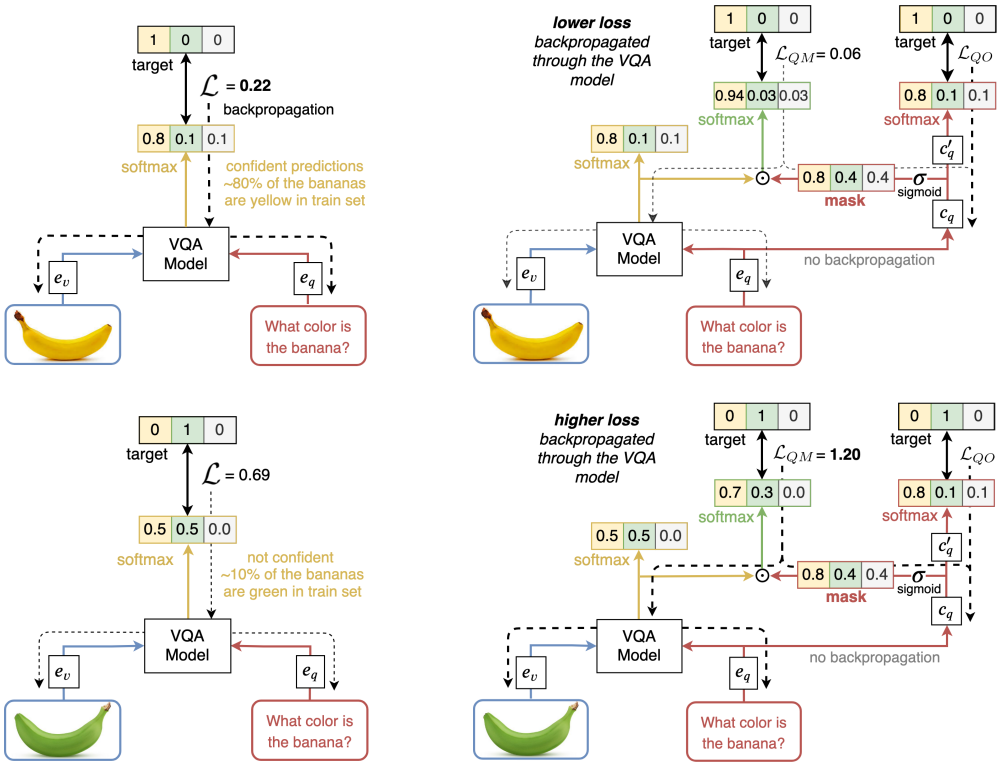
$$\mathcal{L}(\theta; \mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} \log(\text{softmax}(f(v_i, q_i)))[a_i] \tag{2}$$

VQA models are inclined to learn unimodal biases from the datasets [10]. This can be shown by evaluating models on datasets that have different distributions of answers for the test set, such as VQA-CP v2. In other words, they rely on statistical regularities from one modality to provide accurate predictions without having to consider the other modality. As an extreme example, strongly biased models towards the question modality always output *yellow* to the question *what color is the banana*. They do not learn to use the image information because there are too few examples in the dataset where the banana is not yellow. Once trained, their inability to use the two modalities adequately makes them inoperable on data coming from different distributions such as real-world data. Our contribution consists in modifying this cost function to avoid the learning of these biases.

3.1 RUBi learning strategy

Learning biases with a question-only branch One way to measure the unimodal biases in VQA datasets is to train an unimodal model which takes only one of the two modalities as input. A question-only model can be formalized as a function $f_Q : \mathcal{Q} \rightarrow \mathbb{R}^{|\mathcal{A}|}$ parameterized by θ_Q, and composed of a question encoder e_q and a classifier c_q. All parameters are learned with the standard cross-entropy criterion from Equation (2).

$$f_Q(q_i) = c_q(e_q(q_i)) \tag{3}$$



(a) Classical learning strategy

(b) RUBi learning strategy

Figure 3: Detailed illustration of the RUBi impact on the learning. In the first row, we illustrate how RUBi reduces the loss for examples that can be correctly answered without looking at the image. In the second row, we illustrate how RUBi increases the loss for examples that cannot be answered without using both modalities.

The key idea of our approach, depicted in Figure 2, is to use the question-only model to capture the question biases, allowing the model to focus on the examples that cannot be answered correctly using the question modality only. For any VQA model of the form presented in Equation (1), we add a question-only branch, i.e. a classifier using the encoder e_q from the VQA model. During training, the branch acts as a proxy preventing the VQA model from learning biases. At the end of the training, we simply remove the branch and use the predictions from the base VQA model.

Preventing biases by masking predictions Before passing the predictions of our base VQA model to the loss function defined in Equation (2), we merge them with the output of the question-only branch. This output is a mask containing a scalar value between 0 and 1 for each answer. It is obtained by passing the output of c_q through a sigmoid function σ. The goal of this mask is to dynamically alter the loss by modifying the predictions of the VQA model. To obtain the new predictions, we simply compute an element-wise product $\odot$ between the mask and the original predictions as defined in the following equation. We use $\mathcal{L}_{QM}$ to refer to the cross-entropy loss associated with these predictions. We use θ_{QM} to refer to the union of the parameters of the base model and the question classifier c_q.

$$f_{QM}(v_i, q_i) = f(v_i, q_i) \odot \sigma(f_Q(q_i))) \tag{4}$$

We modify the predictions in this specific way to prevent the VQA model to learn biases from the question. To better understand the impact of our approach on the learning, we examine two scenarios. First, we reduce the importance of the most biased examples, i.e. examples that can be correctly classified without using the image modality. To do so, the question-only branch outputs a mask to increase the score of the correct answer while decreasing the scores of the others. As a result, the loss is much lower for these biased examples. In other words, the gradients backpropagated through the VQA model are smaller, thereby reducing the importance of these examples in the learning. As illustrated in the first row of Figure 3, given the question *what color is the banana*,

the mask takes a high value of 0.8 for the answer *yellow* which is the most likely answer for this question in the training set. On the other hand, the value for the other answers *green* and *white* are smaller. We see that the mask influences the VQA model to produce new predictions where the score associated with the answer *yellow* increases from 0.8 to 0.94. Compared to the classical learning approach, the loss is smaller with RUBi and decreases from 0.22 to 0.06. Secondly, we increase the importance of examples that cannot be answered without using both modalities. For these examples, the question-only branch outputs a mask that increases the score of the wrong answer. As a result, the loss is much higher and the VQA model is encouraged to learn from these examples. We illustrate this behavior in the second row of Figure 3 for the same question about the color of the banana. When the image contains a green banana, RUBi increases the loss from 0.69 to 1.20.

We also add a classifier c'_q with a question-only cross-entropy loss $\mathcal{L}_{QO}$ on top of the question-only branch. By doing so, we further improve the unimodal branch ability to capture biases. It reduces the amount of unimodal biases learned by the VQA model. With this loss, we optimize θ_{QO} containing the parameters of the question-only branch modules, c'_q and c_q. Note that we don't backpropagate this loss to the question encoder to prevent it from directly learning question biases.

$$f_{QO}(q_i) = c'_q(f_Q(q_i))) \tag{5}$$

We obtain our final loss $\mathcal{L}_{\text{RUBi}}$ by summing the two losses together in the following equation:

$$\mathcal{L}_{\text{RUBi}}(\theta_{QM}, \theta_{QO}; \mathcal{D}) = \mathcal{L}_{QM}(\theta_{QM}; \mathcal{D}) + \mathcal{L}_{QO}(\theta_{QO}; \mathcal{D}) \tag{6}$$

3.2 Baseline architecture

Most VQA architectures from the state of the art are compatible with our RUBi learning strategy. To test our strategy, we design a fast and simple architecture inspired from [16]. This baseline architecture is detailed in the supplementary material. As common in the state of the art, our baseline architecture encodes the image as a bag of n_v visual features $\mathbf{v}_i \in \mathbb{R}^{d_v}$ using the pretrained Faster R-CNN by [15], and encodes the question as a vector $\mathbf{q} \in \mathbb{R}^{d_q}$ using a GRU, pretrained on the skipthought task [3]. The VQA model consists of a Bilinear BLOCK fusion [17] which merge the question representation $\mathbf{q}$ with the features $\mathbf{v}_i$ of each region of the image. The output is aggregated using a max pooling on the n_v regions. The resulting vector is then fed into a MLP classifier which outputs the final predictions. While most of our experiments are done with this fast and simple baseline architecture, we experimentally demonstrate that the RUBi learning strategy is effective on other VQA architectures.

4 Experiments

Experimental setup We train and evaluate our models on VQA-CP v2 [10]. This dataset was developed to evaluate the models robustness to question biases. We follow the same training and evaluation protocol as [25], who also propose a learning strategy to reduce biases. For each model, we report the standard VQA evaluation metric [8]. We also evaluate our models on the standard VQA v2 [9]. Further implementation details are included in the supplementary materials.

4.1 Results

State-of-the-art comparison In Table 1, we compare our approach consisting of our baseline architecture trained with RUBi on VQA-CP v2 against the state of the art. To be fair, we only report approaches that use the strong visual features from [15]. We compute the average accuracy over 5 experiments with different random seeds. Our RUBi approach reaches an average overall accuracy of 47.11% with a low standard deviation of $\pm$0.51. This accuracy corresponds to a gain of +5.94 percentage points over the current state-of-the-art UpDn + Q-Adv + DoE. It also corresponds to a gain of +15.88 over GVQA [10], which is a specific architecture designed for VQA-CP. RUBi reaches a +8.65 improvement over our baseline model trained with the classical cross-entropy. In comparison, the second best approach UpDn + Q-Adv + DoE only achieves a +1.43 gain in overall accuracy over their baseline UpDn. In addition, our approach does not significantly reduce the accuracy over our baseline for the answer type *Other*, while the second best approach reduces it by 10.57 point.

Architecture agnostic RUBi can be used on existing VQA models without changing the underlying architecture. In Table 2, we experimentally demonstrate the generality and effectiveness of our

Table 1: State-of-the-art results on VQA-CP v2 `test`. All reported models use the same features from [15]. Models with * have been trained by [25]. Models with ** have been trained by [41].

Model	Overall	Answer type		
		Yes/No	Number	Other
Question-Only [10]	15.95	35.09	11.63	7.11
UpDn [15] **	38.01	.	.	.
RAMEN [41]	39.21	.	.	.
BAN [19] **	39.31	.	.	.
MuRel [16]	39.54	42.85	13.17	45.04
UpDn [15] *	39.74	42.27	11.93	**46.05**
UpDn + Q-Adv + DoE [25]	41.17	65.49	15.48	35.48
Baseline architecture (ours)	38.46 ± 0.07	42.85 ± 0.18	12.81 ± 0.20	43.20 ± 0.15
RUBi (ours)	**47.11** ± 0.51	**68.65** ± 1.16	**20.28** ± 0.90	43.18 ± 0.43

Table 2: Effectiveness of the RUBi learning strategy when used on different architectures on VQA-CP v2 `test`.

SAN	Overall	UpDn	Overall
Baseline [26]	24.96	Baseline [15]	39.74
+ Q-Adv + DoE [25]	33.29	+ Q-Adv + DoE [25]	41.17
+ RUBi (ours)	**36.69**	+ RUBi (ours)	**44.23**

Table 3: Overall accuracy of the RUBi learning strategy on VQA v2 `val` and `test-dev` splits.

Model	`val`	`test-dev`
Baseline (ours)	**63.10**	**64.75**
RUBi (ours)	61.16	63.18

learning scheme by showing results on two additional architectures, Stacked Attention Networks (SAN) [26] and Bottom-Up and Top-Down Attention (UpDn) [15]. First, we show that applying RUBi on these architectures leads to important gains over the baselines trained with their original learning strategy. We report a gain of +11.73 accuracy point for SAN and +4.5 for UpDn. This lower gap in accuracy may show that UpDn is less driven by biases than SAN. This is consistent with results from [25]. Secondly, we show that these architectures trained with RUBi obtain better accuracy than with the state-of-the-art strategy from [25]. We report a gain of +3.4 with SAN + RUBi over SAN + Q-Adv + DoE, and +3.06 with UpDn + RUBi over UpDn + Q-Adv + DoE.

Impact on VQA v2 We report the impact of our method on the standard VQA v2 dataset in Table 3. It uses the same data as VQA-CP v2, but includes the same statistical regularities in its train, val and test sets. In this context, we usually observe a drop in accuracy using approaches focused on reducing biases. This is due to the fact that exploiting unwanted correlations from the VQA v2 train set is not discouraged and often leads to a higher accuracy on the test set. Nevertheless, our RUBi approach leads to a comparable drop to what can be seen in the state-of-the-art. We report a drop of 1.94 percentage points with respect to our baseline, while [10] report a drop of 3.78 between GVQA and their SAN baseline. [25] report drops of 0.05, 0.73 and 2.95 for their three learning strategies with the UpDn architecture which uses the same visual features as RUBi. As shown in this section, RUBi improves the accuracy on VQA-CP v2 from a large margin, while maintaining competitive performance on the standard VQA v2 dataset compared to similar approaches.

4.2 Qualitative analysis

To better understand the impact of our RUBi approach, we compare in Figure 4 the answer distribution on VQA-CP v2 for some specific question patterns. We also display interesting behaviors on some examples using attention maps extracted as in [16]. In the first row, we show the ability of RUBi to reduce biases for the *is this person skiing* question pattern. Most examples in the train set have the answer *yes*, while in the test set, they have the answer *no*. Nevertheless, RUBi outputs 80% of *no*, while the baseline almost always outputs *yes*. Interestingly, the best scoring region from the attention map of both models is localized on the shoes. To get the answer right, RUBi seems to reason about

the absence of skis in this region. It seems that our baseline gets it wrong by not seeing that the skis are not locked under the ski boots. This unwanted behavior could be due to the question biases. In the second row, similar behaviors occur for the *what color are the bananas* question pattern. 80% of the answers from the train set are *yellow*, while most of them are *green* in the test set. We show that the amount of *green* and *white* answers from RUBi are much closer to the ones from the test set than with our baseline. In the example, it seems that RUBi relies on the color of the banana, while our baseline misses it. In the third row, it seems that RUBi is able to ground the textual concepts such as *top part of the fire hydrant* and *color* on the right visual region, while the baseline relies on the correlations between the fire hydrant, the yellow color of its core and the answer *yellow*. Similarly on the fourth row, RUBi grounds *color*, *star*, *fire hydrant* on the right region, while our baseline relies on correlations between *color*, *fire hydrant*, the yellow color of the top part region and the answer *yellow*. Interestingly, there is no similar question that involves the color of a star on a fire hydrant in the training set. It shows the capacity of RUBi to generalize to unseen examples by composing and grounding existing visual and textual concepts from other kinds of question patterns.

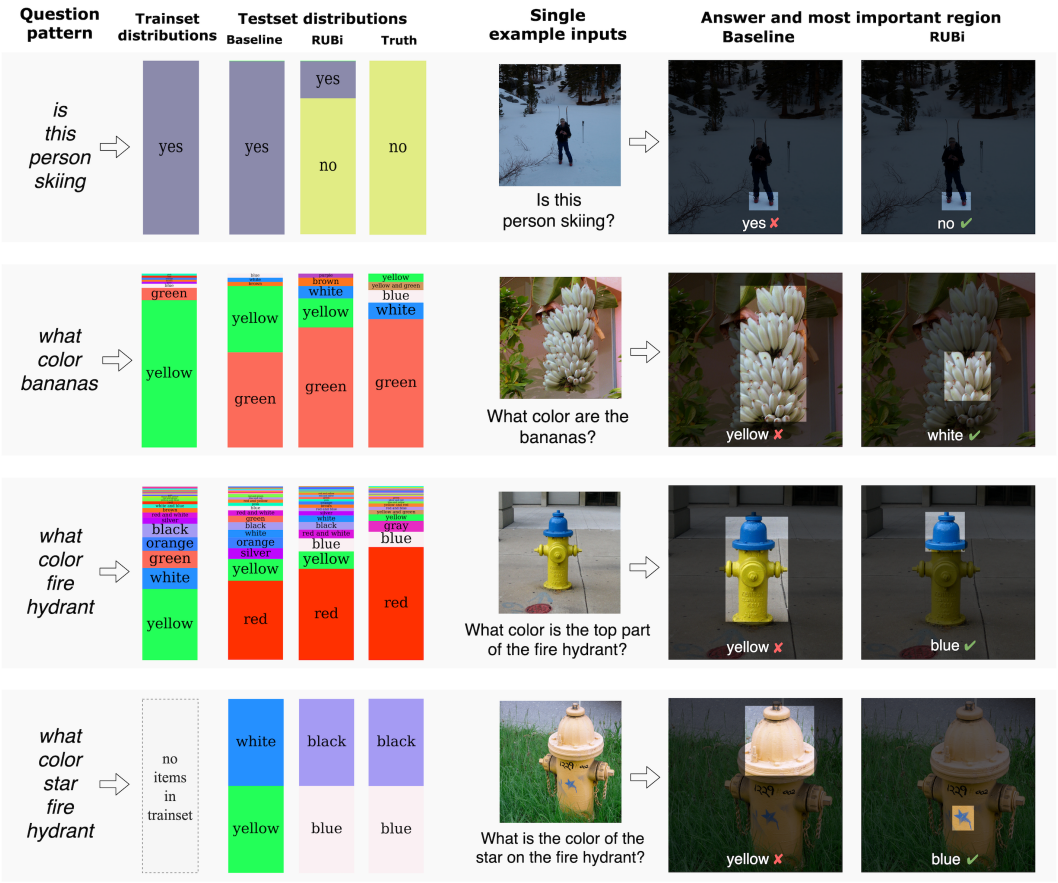


Figure 4: Qualitative comparison between the outputs of RUBi and our baseline on VQA-CP v2 `test`. On the left, we display distributions of answers for the train set, the baseline evaluated on the test set, RUBi on the test set and the ground truth answers from the test set. For each row, we filter questions in a certain way. In the first row, we keep the questions that exactly match the string *is this person skiing*. In the three other rows, we filter questions that respectively include the following words: *what color bananas*, *what color fire hydrant* and *what color star hydrant*. On the right, we display examples that contains the pattern from the left. For each example, we display the answer of our baseline and RUBi, as well as the best scoring region from their attention map.

5 Conclusion

We propose RUBi to reduce unimodal biases learned by Visual Question Answering (VQA) models. RUBi is a simple learning strategy designed to be model agnostic. It is based on a question-only

branch that captures unwanted statistical regularities from the question modality. This branch influences the base VQA model to prevent the learning of unimodal biases from the question. We demonstrate a significant gain of +5.94 percentage point in accuracy over the state-of-the-art result on VQA-CP v2, a dataset specifically designed to account for question biases. We also show that RUBi is effective with different kinds of common VQA models. In future works, we would like to extend our approach on other multimodal tasks.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013.

[3] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[4] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[5] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[6] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2017.

[10] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

[13] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 2019.

[14] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *CVPR*, 2019.

[15] 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 *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, June 2018.

[16] Remi Cadene, Hedi Ben-Younes, Nicolas Thome, and Matthieu Cord. Murel: Multimodal Relational Reasoning for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2019.

[17] 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 33st Conference on Artificial Intelligence (AAAI)*, 2019.

[18] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *ECCV*, 2018.

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

[20] Juanzi Li Jiaxin Shi, Hanwang Zhang. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, 2019.

[21] Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. Chain of Reasoning for Visual Question Answering. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 275–285. Curran Associates, Inc., 2018.

[22] Gao Peng, Zhengkai Jiang, Haoxuan You, Zhengkai Jiang, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering. In *CVPR*, Dec 2019.

[23] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2017.

[24] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *EMNLP*, 2016.

[25] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pages 1541–1551, 2018.

[26] 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*, pages 21–29, 2016.

[27] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30. ACM, 2013.

[28] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. *NAACL*, 2018.

[29] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. *CVPR*, Jun 2011.

[30] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[31] Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini. Right for the Right Reason: Training Agnostic Networks. *Lecture Notes in Computer Science*, page 164–174, 2018.

[32] Varun Manjunatha, Nirat Saini, and Larry S. Davis. Explicit Bias Discovery in Visual Question Answering Models. In *CVPR*, Nov 2019.

[33] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[34] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[35] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939, 2016.

[36] Abhinav Gupta, Adithyavairavan Murali, Dhiraj Prakashchand Gandhi, and Lerrel Pinto. Robot learning in homes: Improving generalization and reducing dataset bias. In *Advances in Neural Information Processing Systems*, pages 9094–9104, 2018.

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

[38] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[39] Hedi Ben-Younes, Rémi Cadène, Nicolas Thome, and Matthieu Cord. Mutan: Multimodal tucker fusion for visual question answering. *ICCV*, 2017.

[40] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[41] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Answer them all! toward universal visual question answering models. *CVPR*, 2019.

6 Implementation details

Image encoder We use the pretrained Faster R-CNN by [15] to extract object features. We use the setup that extracts 36 regions for each image. We do not fine-tune the image extractor.

Question encoder We use the same preprocessing as in [16]. We apply a lower case transformation and remove the punctuation. We only consider the most frequent 3000 answers for both VQA v2 and VQA CP v2. We then use a pretrained Skip-thought encoder with a two-glimpses self attention mechanism. The final embedding is of size 4800. We fine-tune the question encoder during training.

Baseline architecture Our baseline architecture is a simplified version of the MuRel architecture [16]. First, it computes a bilinear fusion between the question vector and the visual features for each region. The bilinear fusion module is a BLOCK [17] composed of 15 chunks, each of rank 15. The dimension of the projection space is 1000, and the output dimension is 2048. The output of the bilinear fusion is aggregated using a max pooling over n_v regions. The resulting vector is then fed into a MLP classifier composed of three layers of size (2048, 2048, 3000), with ReLU activations. It outputs the predictions over the space of the 3000 answers.

Question-only branch The RUBi question-only branch feeds the question into a first MLP composed of three layers, of size (2048, 2048, 3000), with ReLU activations. First, this output vector goes through a sigmoid to compute the mask that will alter the predictions of the VQA model. Secondly, this same output vector goes through a single linear layer of size 3000. We use these question-only predictions to compute the question-only loss.

Optimization process We train all our models with the Adam optimizer. We train our baseline architecture with the learning rate scheduler of [16]. We use a learning rate of 1.5×10^{-4} and a batch size of 256. During the first 7 epochs, we linearly increase the learning rate to 6×10^{-4}. After epoch 14, we apply a learning rate decay strategy which multiplies the learning rate by 0.25 every two epochs. We train our models until convergence as we do not have a validation set for VQA-CP v2. For the UpDn and SAN architectures, we follow the optimization procedure described in [25].