Learning Reasoning Mechanisms for Unbiased Question-based Counting

Corentin Dancette, Remi Cadene, Xinlei Chen, Matthieu Cord

Learning Reasoning Mechanisms for Unbiased Question-based Counting

Corentin Dancette[1*] Remi Cadene[1*] Xinlei Chen[2] Matthieu Cord[1,3]
[1]Sorbonne Université, 4 place Jussieu, 75005, Paris [2]Facebook AI Research [3]Valeo.ai
[1]first.last@sorbonne-universite.fr [2]xinleic@fb.com

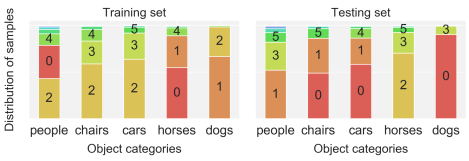
1. Introduction

Biases allow deep learning models to reach impressive performances on regular datasets but can be harmful if exploited in the real-world [10, 17]. It is critical to create benchmarks that reflect these failures [2, 8]. We tackle this issue for the visual counting task [1, 7, 18], a subset of VQA [4, 11] which also requires high level reasoning abilities and displays similar biases. The appearance of certain words in the question or objects in the image is predictive of the count label. An example of bias from the TallyQA counting dataset [1] is displayed in Figure 1.

First, we introduce two counting datasets meant to evaluate a model's ability to avoid biases. Both datasets are built on with the idea of *changing distributions*, meaning the training and testing distributions are different, similarly to VQA-CP. Second, we introduce the Spatial Counting Network (SCN), a model that better avoids learning the biases and instead relies on more suited mechanisms for counting, and we show its superior performance on our evaluation procedure.

2. Novel out-of-distribution datasets

TallyQA-CP Inspired by VQA-CP [2], we build TallyQA-CP to penalize models that over-rely on the question-related biases. We construct a new training set and testing set by extracting the main concept to be counted from each question (e.g. in "how many tables are green", the concept will be "tables"), and use it to conditions the answer distribution differently between the training and the testing set. We display on the left the shift in distributions for the five most common concepts.



TallyQA-Odd-Even A characteristic of our proposed TallyQA-CP is that it mostly penalizes the use of question-related biases. Instead, we introduce the Odd-Even version that penalizes, **by construction**, the use of biases from both question and image. We generate the TallyQA-Odd-Even dataset by removing 90% of the samples associated to an even count label from the TallyQA training set and 90% of the samples associated to an odd label from the testing set. Models that are trained on odd numbers should generalize to even numbers if they do not rely on biases.

To avoid adaptive over-fitting [9, 16], we hold out 10% of the training set as a validation set for early-stopping on both our datasets.

*Equal contribution

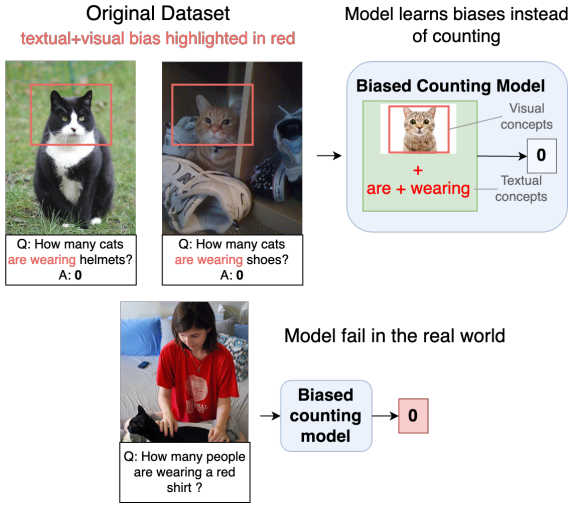


Figure 1. Matching simple patterns from the training set can be enough to answer a large number of counting questions and obtain higher accuracy on the OOD testing set. In the real-world, biased models that rely on such a pattern would fail to provide the correct answer.

3. Spatial Counting Network

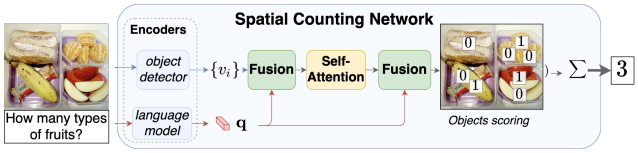


Figure 2. **Spatial Counting Network**.

Our SCN model contains inductive biases to encourage the learning of the counting mechanism, and avoid learning biases. As shown in Figure 2, for the image v, a we use a pre-trained object detector [3] to transforms the raw pixels to region features. For the question, we use a pretrained skip-thought encoder [13]. We then merge question and image vectors using a bilinear **multi-modal fusion** module [12]. To allow box deduplication and spatial reasoning, we add a **self-attention** [19] layer. The resulting vectors are then again fused with the question representation and produce a counting score s_i for each region via sigmoid activation. Finally, the global count output $\hat{c} = \sum_i s_i$ is a simple summation of all the individual counting scores.

Loss: Unlike many state-of-the-art counting or VQA models [20, 1, 14, 15] that treat count numbers as classification labels, we interpret them as numbers and we train the model with a MSE regression loss, $\mathcal{L}_{\text{MSE}}$. Additionally, to encourages each sigmoid output s_i to be close to 0 or 1, we add a **binary entropy regularization** term per-region: $\mathcal{L}_H = -\frac{1}{n}\sum_{\mathcal{D}} \frac{1}{n_v} \sum_{i=1}^{n_v} H(s_i)$. We show its effect in Figure 3. Our final training loss is $\mathcal{L} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_H$.

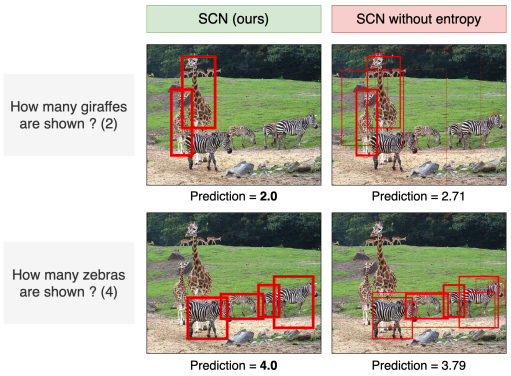


Figure 3. Comparison of our model with and without entropy loss. Boxes are in bold when their count value $c_i \approx 1$. The model trained with entropy selects the correct two regions, while the model without it associates fractional values to multiple regions and fails to distinguish duplicates.

4. Results

4.1. State-of-the-art models are biased

As shown in Table 1, all models suffer from a large drop in accuracy, compared to their scores on the original version of TallyQA [1]. For example, RCN had an overall accuracy of 65.49% on TallyQA. However, it only gets 2% accuracy on TallyQA-CP testing set, and 28.4% on TallyQA-Odd-Even. The bias-reduction methods (uniform sampling and RUBi [6]) have a positive impact on TallyQA-CP, which is expected, especially for RUBi, since it targets specifically question-related biases. Finally, we can notice that **most of the models**, in both benchmarks, reach lower performance than **Random** $\mathcal{D}_{\text{test}}$ that follows the testing set distribution.

	TallyQA-CP		Odd-Even		TallyQA [1]	
	Acc. ↑	RMSE ↓	Acc.	RMSE	Acc.	RMSE
Random $\mathcal{D}_{\text{train}}$	19.53	2.84	10.26	2.81	20.18	2.92
Random $\mathcal{D}_{\text{test}}$	20.40	2.89	30.68	2.61	31.80	2.20
Q-Only [1]	0.63	2.23	16.92	1.91	42.38	1.74
I-Only [1]	21.55	2.24	9.80	2.06	38.14	1.70
Q+I [1]	1.68	1.97	20.86	1.80	52.32	1.49
MUTAN [5]	1.91	1.96	24.99	1.67	53.51	1.54
Counter [20]	0.64	2.08	19.89	1.83	62.58	1.34
RCN [1]	2.00	1.76	28.40	1.61	**65.49**	**1.26**
RCN + Sampling	5.58	1.82	27.10	1.63	53.78	1.58
RCN + RUBi [6]	31.04	1.56	25.35	1.71	59.83	1.35
RCN + $\mathcal{L}_{\text{MSE}}$	14.99	1.60	31.44	1.51	60.35	1.2
SCN (ours)	**34.79**	**1.46**	**40.87**	1.50	57.39	1.24
SCN without $\mathcal{L}_{\mathcal{H}}$	26.88	1.47	39.54	**1.48**	57.07	1.24

Table 1. Benchmark of question-based visual counting models on our TallyQA-CP and TallyQA-Odd-Even datasets. We report the accuracy and the RMSE scores on the testing and validation sets. "Sampling" stands for uniform sampling strategy.

Impact of regression loss: We report a gain of +12.99 on TallyQA-CP and +3.04 on TallyQA-Odd-Even points for **RCN with** $\mathcal{L}_{\text{MSE}}$ over RCN. These good performances suggest that regression is a better design choice to avoid learning biases.

4.2. Spatial Counting Network

On TallyQA-CP, we report the best accuracy of 34.79% for our SCN on TallyQA-CP (+32.79 over RCN). On TallyQA-Odd-Even, SCN reaches the best accuracy of 40.87% (+12.47 over RCN) We also perform an ablation of SCN, without the **entropy loss** (**SCN without** $\mathcal{L}_{\mathcal{H}}$). We report an important effect on TallyQA-CP, with -7.91 points, which confirms the effect seen in Figure 3.

Accuracy per count label: In Figure 4, we display a comparison between our model and RCN on the count labels. Interestingly, we report a higher accuracy on even count labels, less represented in the training set and a lower accuracy on odd count labels, more represented. We also report a smaller differences between adjacent count labels, compared with RCN. These results suggest that our design choices help to generalize to a different distribution of count labels, by learning mechanisms more suited for counting.

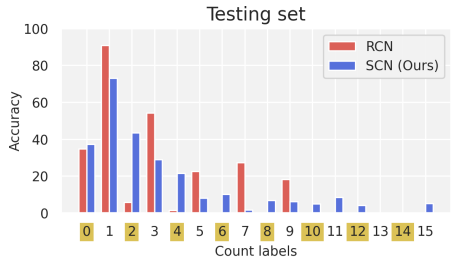


Figure 4. Accuracy per count labels of our model and RCN [1] on TallyQA-Odd-Even.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 1, 2

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 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 (CVPR)*, 2018. 1

[3] 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. 2

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[5] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[6] 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. 2

[7] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[8] Corentin Dancette, Remi Cadene, Xinlei Chen, and Matthieu Cord. Overcoming statistical shortcuts for open-ended visual counting. *arXiv preprint arXiv:2006.10079*, 2020. 1

[9] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015. 1

[10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020. 1

[11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[12] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2

[13] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[14] 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. 2

[15] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 2

[16] 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. 1

[17] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[18] Alexander Trott, Caiming Xiong, and Richard Socher. Interpretable counting for visual question answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2

[20] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 2