Make Up Your Mind: Towards Consistent Answer Predictions in VQA Models

Arijit Ray, Giedrius T. Burachas, Karan Sikka, Anirban Roy, Avi Ziskind, Yi Yao, Ajay Divakaran

SRI International, Princeton, NJ, USA

1 Introduction

"*A skeptic, I would ask for consistency first of all.*" - Sylvia Plath

Visual Question Answering (VQA) involves answering natural language questions on images [1]. While state-of-the-art models can answer such questions satisfactorily well on a standard VQA dataset [1], we observe that it still often makes blatant mistakes while answering questions from slightly different perspectives. This reduces their perceived trust and raises concerns as to whether they can represent the semantics of concepts in questions and images accurately.

We introduce the task of belief consistency in VQA models, i.e. the task of answering different semantically-grounded questions about a certain concept consistently. For example, if the answer to "is it a vegetarian pizza?" is "yes", the answer to "is there pepperoni on the pizza?" should be "no". Current VQA datasets do not have such multiple consistent questions about a single concept to test the consistency of a model. We propose a simple approach to auto-generate a consistent VQA dataset (**ConVQA**) on top of Visual Genome [2] that comprises such consistent QA sets. We propose a new performance metric and benchmark a current VQA model on ConVQA. We also release a fine-tuned baseline model that slightly improves both ConVQA and VQA2.0 performance.

2 Related Work

Consistency checking can be thought of as an interrogative Turing Test [3]. Consistently being able to answer questions from different angles is a necessity to suggest that a machine truly understands the premise in a question for a given image [4] [5] and the true meaning of the semantic concepts rather than relying on dataset biases [6] [7]. Consistent QA pairs to a given question can also be used as a natural language explanation [8] of the model's beliefs, thus helping the model appear human-like when asked "why?" or for further related dialog [9]. Training for consistency is also related to adversarial training [10], however, the consistent inputs are not tailored to intentionally disrupt the system, but rather to enforce consistency.

3 Dataset

We use scene graph annotations from the Visual Genome Dataset and template-based NLP techniques to generate a dataset of consistent QA sets (**ConVQA**)

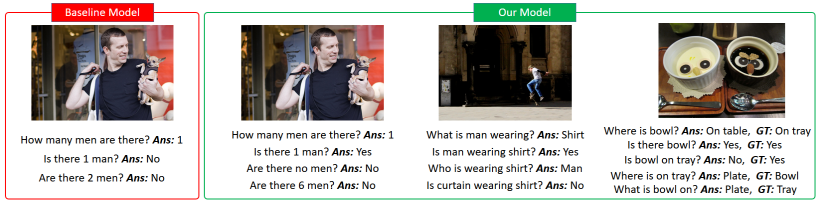

Fig. 1. After fine-tuning the baseline model on our ConVQA Train set, our model predicts answers more consistently on questions phrased differently.

Table 1. Accuracies for models on our ConVQA and VQA2.0 Dataset

	ConVQA			VQA 2.0			
	Perfect-Con	**Avg-Con**	**Top1**	**VQA-Val**	**Count**	**Yes/No**	**Other**
Baseline	2.67%	44.29%	45.21%	62.17%	42.89%	75.36%	**55.87%**
Fine-Tuned	**10.94%**	**63.57%**	**64.97%**	**62.27%**	**43.44%**	**75.96%**	55.39%

(more details in appendix). Currently, we only focus on counting, attribute, existential and relational consistency. We generate groups of questions phrased differently about a certain concept to make consistent QA sets. For example, for the attribute "white" of object "cup" in the Visual Genome dataset, we generate "is the cup white? Yes", "Is the cup black? No" and "What color is cup? White".

4 Baseline Benchmark Results

We introduce two metrics - we say a model is perfectly consistent when it answers all questions in a consistent set correctly (**Perfect-Con**). We also measure the average consistency percentage (fraction of questions answered correctly in a consistent set) across all consistent sets (**Avg-Con**).

As the baseline, we use a VQA model similar to [11] and [12]. It takes both Region Proposal features [13] and ResNet152 [14] features as input and gets 62.17% on 10,000 image-question pairs on the VQA2.0 Val Dataset. We fine-tune jointly on our ConVQA and VQA2.0 Train splits (more details in appendix). We test on 17,000 consistent sets (3,000 images) on our ConVQA Test split. Joint ConVQA+VQA fine-tuning significantly improves ConVQA performance while still maintaining (and slightly improving) the accuracy on VQA2.0 Val set. The results are outlined in Table 1 and qualitative examples are shown in Fig 1.

5 Conclusion and Future Work

We introduced the problem of consistency in VQA and released a preliminary baseline on how it can be tackled using auto-generated datasets. While auto-generated datasets are cheaper, we do plan to systematically curate it using expert help to reduce noise and broaden our definitions of consistency. Similar to the framework of [15], we also plan to learn a way to ask these consistent questions so that machines can self-check for consistency like humans.

Appendix

5.1 Dataset Generation

Here is a summary of our consistent sets:

Relational/Existential Consistency

1. `Is <object> <relation> <subject>? Yes`. For example, is man standing near tree?, Yes
2. `Is there <object>? Yes`, For example, is there man? Yes
3. `Is there <subject>? Yes`
4. `Who/What is <relation> <subject>? <object>.` For example, Who is standing near tree? Man
5. `Is <other object> <relation> <subject>? No`, `Is <object> <relation> <other subject>? No`. We cross verify with scene graph to make sure these are "no". However, the scene graph isn't exhaustively annotated for all images and hence, these maybe noisy sometimes.

Counting Consistency

1. `How many plural(<object>) are there? <count>`
2. `Are there <count> <object>? Yes`. `<object>` is replaced by its plural if count is not 1 and "count" replaced by "no" if count is 0.
3. `Are there <other count> plural(<object>)? No`

We observe that counting questions are often noisy because of annotations not being exhaustive and non-countable objects being annotated, While annotation exhaustiveness cannot be automatically tackled, we ignore annotations of non-countable stuff ("shadow", "water" etc) to make this somewhat cleaner.

Attribute Consistency

1. `What hypernym(<attribute>) is <object>? <attribute>`. For example, "What color is cup? White". We get hypernyms using WordNet.
2. `Is <object> <attribute>? Yes`
3. `Is <object> opposite(<attribute>)? No.` We get opposite attributes using WordNet.

Some WordNet hypernyms and opposites are noisy, so we manually prune them for some cases (like ignoring opposites for other colors except "white" etc.

Dataset Statistics We generate template-based consistent sets for 108,077 images in the Visual Genome (VG) Dataset. To reduce noise and rule-out non-salient or unnatural objects, we only consider the top 100 objects and attributes that occur in VG. On average, ConVQA has 14.16 consistent sets per image and 4.43 QA pairs per consistent set. There are a total of 6,792,068 QA pairs in 1,530,979 consistent sets over 108,077 images.

5.2 Training and Evaluation Details

We generate consistent QA sets for 108,000 Visual Genome images. We split the dataset into Train (80%), Val (10%) and Test (10%). For all our fine-tuning experiments, we train for 1 epoch consisting of 1000 batches with batch size of 196. We also randomly inject batches of VQA training data 50% of the time to prevent the model from overfitting on our simpler questions. For the sake of time, we randomly choose 10 consistent sets for each image while training and also limit the number of QA pairs in each consistent set to a maximum of 10. For all evaluation purposes, we test on about 17K consistent sets spanning 3000 images from our "Test" split.

Acknowledgements

This research was supported by funding from the Defense Advanced Research Projects Agency (DARPA) under the Explainable AI (XAI) Program. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2425–2433
2. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1) (2017) 32–73
3. Radziwill, N.M., Benton, M.C.: Evaluating quality of chatbots and intelligent conversational agents. arXiv preprint arXiv:1704.04579 (2017)
4. Ray, A., Christie, G., Bansal, M., Batra, D., Parikh, D.: Question relevance in vqa: identifying non-visual and false-premise questions. arXiv preprint arXiv:1606.06622 (2016)
5. Mahendru, A., Prabhu, V., Mohapatra, A., Batra, D., Lee, S.: The promise of premise: Harnessing question premises in visual question answering. arXiv preprint arXiv:1705.00601 (2017)
6. Agrawal, A., Kembhavi, A., Batra, D., Parikh, D.: C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. arXiv preprint arXiv:1704.08243 (2017)
7. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR. Volume 1. (2017) 3
8. Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. arXiv preprint arXiv:1802.08129 (2018)
9. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2017)

10. Shi, H., Mao, J., Xiao, T., Jiang, Y., Sun, J.: Learning visually-grounded semantics from contrastive adversarial samples. arXiv preprint arXiv:1806.10348 (2018)
11. Kazemi, V., Elqursh, A.: Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162 (2017)
12. Teney, D., Anderson, P., He, X., Hengel, A.v.d.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. arXiv preprint arXiv:1708.02711 (2017)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE (2017) 2980–2988
14. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. Volume 4. (2017) 12
15. Misra, I., Girshick, R., Fergus, R., Hebert, M., Gupta, A., van der Maaten, L.: Learning by asking questions. arXiv preprint arXiv:1712.01238 (2017)