Reducing Language Biases in Visual Question Answering with Visually-Grounded Question Encoder

Gouthaman KV and Anurag Mittal

Indian Institute of Technology Madras, India
{gkv,amittal}@cse.iitm.ac.in

Supplementary Material

Performance of the VGQE-only model

The question representation from the VGQE contains visual information too. To understand whether only such visual information in the question is enough for better performance or if we need further interactions with the image, we experiment as follows. We directly predict the answer from the VGQE output without looking at the image again. We got an overall accuracy of 44.46 on the VQA-CPv2. From this experiment, we infer that using only the visual information from the visually-grounded question may not be sufficient for questions that require specific object-level interactions from the scene, besides only the visual information. E.g., questions like, "What is the person next to the car doing?", "How many are riding bicycles?", etc. Hence, further interactions between the question and various objects from the image are required for improved performance.

Implementation details

Feature extraction: In all our experiments, we use the fixed-sized (36) object-level image features provided by [2] as V ($d_v = 2048$). These features are obtained from Faster-RCNN [12] trained on Visual Genome [7], with ResNet-101 [5] as the backbone. We did not fine-tune the image features. We use the pre-trained Glove [10] word-embeddings (trained on "Common Crawl") to extract the object-label and question word features ($d_w = 300$). If the object-label contains two words, we take the sum of the word-embeddings of the individual words as the word-embedding vector. We use a similar question pre-processing as in [4], such as lower-case transformation, removing punctuation, etc. If a question word is not there in the Glove word embedding vocabulary, we use the word embedding of its synonym or other words that give the same meaning.
Model parameters: Inside VGQE, we use a Bi-directional GRU with a hidden state dimension of 1024 as the RNN cell. We use the fine-tuned question word embedding dimension, $d = 512$. In the BLOCK fusion, we use a similar setting

as in [4], which consists of 15 chunks, each of rank 15, the dimension of the projection space is 1000, and the output dimension is 2048.

Dataset and training: We use the VQA-CPv2 dataset [1] to train and evaluate the language-bias reduction capacity of the models. We also use the VQAv2 dataset to train and evaluate the models on the standard VQA benchmark. For both VQA-CPv2 and VQAv2, we consider the most frequent 3000 answers as the answer vocabulary as in prior works [1, 11, 4] and the evaluation metric used is the VQA accuracy [3]. We also use the questions from Visual Genome [7] that matches the answer vocabulary, following the prior works [2, 6]. We train the models using the Adamw [8] optimizer with $weightdecay = 2 * 10^{-5}$ and cross-entropy loss. For the baseline model, we set the initial learning rate as 3.5×10^{-4} and linearly increase it by a factor of 0.25 till epoch 11. Then we decay the learning rate by a factor of 0.25 with a step size of 2. For the BAN baseline, we used an initial learning rate of 2×10^{-4} and followed the same scheduling algorithm as above. For the UpDn baseline, we fixed the learning rate as 2×10^{-4} and used the Binary-Cross entropy loss. We use gradient clipping with a norm threshold as 0.25. We use a batch size of 128 and *dropout* value as 0.2. For implementing all the models, we use the PyTorch [9] deep-learning framework.

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: 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. pp. 4971–4980 (2018)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
3. 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. pp. 2425–2433 (2015)
4. Cadene, R., Dancette, C., Cord, M., Parikh, D., et al.: Rubi: Reducing unimodal biases for visual question answering. In: Advances in Neural Information Processing Systems. pp. 839–850 (2019)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: Advances in Neural Information Processing Systems. pp. 1564–1574 (2018)
7. 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), 32–73 (2017)
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

9. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. https://pytorch.org (2017)
10. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
11. Ramakrishnan, S., Agrawal, A., Lee, S.: Overcoming language priors in visual question answering with adversarial regularization. In: Advances in Neural Information Processing Systems. pp. 1541–1551 (2018)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)