MUREL: Multimodal Relational Reasoning for Visual Question Answering Supplementary Materials

Remi Cadene [1*] Hedi Ben-younes [1,2*] Matthieu Cord [1] Nicolas Thome [3]
[1] Sorbonne Université, CNRS, LIP6, 4 place Jussieu, 75005 Paris
[2] Heuritech, 110 avenue de la République, 75011 Paris
[3] Conservatoire National des Arts et Métiers, 75003 Paris
remi.cadene@lip6.fr, hedi.ben-younes@lip6.fr, matthieu.cord@lip6.fr, nicolas.thome@cnam.fr

Experimental setup

Image We use the pretrained Faster-RCNN [10] by [1] on Visual Genome [9] to extract objects features from each image. Two setups have been proposed in the literature. A first one that extract 36 regions per image and a second one that extract 10 to 100 regions depending on a threshold. For the sake of simplicity, we choose the first setup in order to always represent our image as $\mathbb{R}^{36\times2048}$. We do not fine tune any pf the Faster-RCNN parameters.

Question We use the same preprocessing as [4], which apply a lower case transformation and remove all the punctuation. We only consider the questions that are associated to the 3000 most occurring answers (1480 for the TDIUC dataset) while containing less than 26 words. We use a pretrained Skip-thought encoder by [8] with a two glimpses self attention mechanism [11] to represent our question in a 4800-dimensional space. We fine tune every parameters of the Skip-thoughts including the embedding layer.

Optimization process We use the Adam optimizer [7] with a learning rate of $5 * 10^{-5}$ and a batch size of 256. During the first 7 epochs, we linearly increase the learning rate to $2 * 10^{-4}$. After the epoch 14, we decrease it by a factor 0.25 every two epochs until convergence. We also apply a gradient clipping of 0.25. We use early stopping based on the validation accuracy. This process is inspired from [12, 5].

Loss function We use the standard cross-entropy loss function for multi-class classification problems.

Comparison with classic attention

MuRel leverages the bilinear strategy in a different way than the classical VQA models [2, 3, 4, 6]. Instead of scalar question-guided visual attention maps, the fusion between question and each region is represented as a vector. This more expressive multidimensional representation allows MuRel to focus on specific features of a particular region given a textual context.

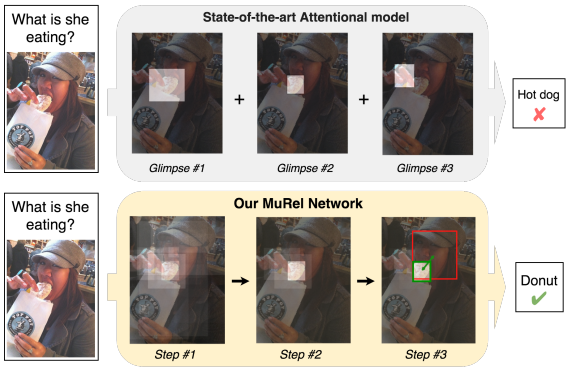

Figure 1. Comparing MuRel visualization w.r.t. classic attention. As MuRel uses pairwise representations between regions, we can infer the most important interaction, here between the woman's face and the donut, enabling correct answer prediction.

An other important aspect of MuRel lies in its pairwise module which models the relations between regions over multiple steps. Besides bringing more capacity, this pairwise modeling also allows to visualize the strongest region interactions, as we show in Figure 1, which is not possible with a classic attention model.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. 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. 1

[2] H. Ben-Younes, R. Cadène, N. Thome, and M. Cord. Mutan: Multimodal tucker fusion for visual question answering. *ICCV*, 2017. 1

*Equal contribution

[3] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the 33st Conference on Artificial Intelligence (AAAI)*, 2019. 1

[4] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*. The Association for Computational Linguistics, 2016. 1

[5] J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear Attention Networks. *arXiv preprint arXiv:1805.07932*, 2018. 1

[6] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017. 1

[7] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[8] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3294–3302, Cambridge, MA, USA, 2015. MIT Press. 1

[9] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1

[10] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 1

[11] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *IEEE International Conference on Computer Vision (ICCV)*, pages 1839–1848, 2017. 1

[12] Yu Jiang*, Vivek Natarajan*, Xinlei Chen*, M. Rohrbach, D. Batra, and D. Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 1