# Multilingual Scene Text VQA

Josep Brugués i Pujolràs

**Abstract**

Short work summary (around 100 words).

**Index Terms**

Your keywords here,...

## I. Introduction

PROBLEM definition and working hypothesis

### A. Subsection Heading Here

Subsection text here, if needed...

## II. State of the art

In this section, ...
[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]

### A. VQA methods

### B. VQA datasets

### C. Optical Character Recognition

### D. Multilingual word embeddings

## III. Method

Computational approach used to solve the problem

## IV. Experiments

All the details about the experiments design and process

## V. Results

Explanation about the performance evaluation procedure and results analysis.

## VI. Conclusions

Summary about the degree of achievement according to the given problem and the adopted hypothesis; and outline about open research lines...

## Appendix A
## Appendix Title

Appendix one text goes here.

## Acknowledgment

The authors would like to thank...

Author: Josep Brugués i Pujolràs, bruguesjosep@protonmail.com
Advisor 1: Lluis Gomze, Document Analysis Group, Centre de Visió per Computador
Advisor 2: Dimosthenis Karatzas, Document Analysis Group, Centre de Visió per Computador
Thesis dissertation submitted: September 2021

## REFERENCES


[1] L. Gómez, A. F. Biten, R. Tito, A. Mafla, M. Rusiñol, E. Valveny, and D. Karatzas, "Multimodal grid features and cell pointers for scene text visual question answering," *CoRR*, vol. abs/2006.00923, 2020. [Online]. Available: https://arxiv.org/abs/2006.00923

[2] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," 2017.

[4] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," *CoRR*, vol. abs/1702.03859, 2017. [Online]. Available: http://arxiv.org/abs/1702.03859

[5] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. Jawahar, and D. Karatzas, "Scene text visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[6] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[7] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 71–79.

[9] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, "Ocr-vqa: Visual question answering by reading text in images," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 947–952.

[10] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh, "Pythia v0.1: the winning entry to the VQA challenge 2018," *CoRR*, vol. abs/1807.09956, 2018. [Online]. Available: http://arxiv.org/abs/1807.09956

[11] A. Singh, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, and D. Parikh, "Pythia-a platform for vision & language research," in *SysML Workshop, NeurIPS*, vol. 2018, 2018.

[12] K. Wang and S. Belongie, "Word spotting in the wild," in *European conference on computer vision*. Springer, 2010, pp. 591–604.

[13] A. Veit, T. Matera, L. Neumann, J. Matas, and S. J. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *CoRR*, vol. abs/1601.07140, 2016. [Online]. Available: http://arxiv.org/abs/1601.07140

[14] A. Mishra, K. Alahari, and C. Jawahar, "Scene Text Recognition using Higher Order Language Priors," in *BMVC - British Machine Vision Conference*. Surrey, United Kingdom: BMVA, Sep. 2012. [Online]. Available: https://hal.inria.fr/hal-00818183

[15] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.