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...

Author: Josep Brugués i Pujolràs, bruguesjosep@protonmail.com
Advisor 1: Lluís Gómez, 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

II. State of the art

In this section, ...

[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27]

A. Visual Question Answering

1) Methods:

2) Datasets:

B. Optical Character Recognition

1) Single language:

2) Multiple language:

C. Multilingual embeddings

1) Word embeddings:

2) Sentence embeddings:

D. Summary

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...

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, Y. Natarajan, Vivand ek 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.
[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805
[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," 2016. [Online]. Available: https://arxiv.org/abs/1602.07332
[18] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
[19] P. Jawanpuria, A. Balgovind, A. Kunchukuttan, and B. Mishra, "Learning multilingual word embeddings in latent metric space: a geometric approach," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 107–120, 2019.
[20] M. Artetxe and H. Schwenk, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 09 2019. [Online]. Available: https://doi.org/10.1162/tacl_a_00288
[21] X. Chen and C. Cardie, "Unsupervised multilingual word embeddings," *arXiv preprint arXiv:1808.08933*, 2018.
[22] J. Tiedemann, "Parallel data, tools and interfaces in opus," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.
[23] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.
[24] R. Speer and J. Lowry-Duda, "Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge," *arXiv preprint arXiv:1704.03560*, 2017.
[25] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016.
[26] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," 2017.
[27] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," 2013.