Multilingual Scene Text VQA

Josep Brugués i Pujolràs

Abstract

Short work summary (around 100 words).

Index Terms

Visual Question Answering, ST-VQA, Optical Character Recognition, Multilingual embeddings

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

As the introduction has pointed out, the main objective of this project is to adapt current Visual Question Answering (VQA) models to work on multiple languages and not just English. To achieve the goals of this project, a series of VQA methods have been used, alongside Object Character Recognition (OCR) techniques and Multilingual embedding methods to encode the different languages into a single, common space. In this section, we present a brief history on all the topics related to the project, from their beginnings to the latest and most important developments.

In view of that, Section II-A focuses on VQA: Section II-A1 explains some of the methodologies to solve this computer vision task while Section II-A2 focuses on the most important datsets that have been released. Next, II-B spotlights on Optical Character Recognition: Section II-B1 on single language OCR and Section II-B2 on multiple language OCR. Following, Section II-C explains the different approaches to obtain Multilingual embeddings: II-C1 explains several techniques to obtain word embeddings while II-C2 explains the approaches to obtain sentence embeddings. To finish up, Section II-D wraps up the most important state-of-the-art research works seen in the following lines.

A. Visual Question Answering

VQA is a computer vision task where a system is given a text-based question about an image, and the algorithm must infer the answer based on information extracted on the image. Questions can be arbitrary and they encompass many sub-problems in computer vision, such as:

- Object recognition: “What is in the image?”
- Object detection: “Are there any cats in the image?”
- Attribute classification: “What color is the cat?”
- Scene classification: “Is it sunny?”
- Counting: “How many cats are in the image?”

1) VQA Methods: Historically, VQA methods have focused on using text found on the image to answer the given question. [1], [2], [3], [4], [5], [6], [7]

[8]

2) VQA Datasets: [9], [10], [11], [12], [13], [4]

B. Optical Character Recognition

OCR is a computer vision task used to extract text –printed or handwritten— from images. As we have seen from Section II-A1, it is one of the backbones on almost all VQA approaches.

[14], [15]

1) Single language OCR: []

2) Multiple language OCR: [16]

C. Multilingual embeddings

Multilingual embeddings are used to represent words or sentences from multiple languages in a single vector space. Unsupervised methods acquire the embeddings without the need of a cross-lingual supervision, which is a significant advantage over traditional, supervised methods that require some sort of supervision between languages [17].

[18], [19], [20], [21]

1) Word embeddings: [22], [23], [24], [25], [26], [27], [26], [28], [29], [30], [31], [32], [33], [17], [34]

2) Sentence embeddings: [35], [36]

D. Summary

Visual Genome [X]	COCOText [X]	Clevr [X]	VizWiz [X]	TextVQA [X]	ST-VQA [X]
Year x Images x Questions	Year x Images x Questions	Year x Images x Questions	Year x Images x Questions	Year x Images x Questions	Year x Images x Questions

TABLE I
Summary of the available VQA datasets and their characteristics

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

[2] 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

[3] 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.

[4] 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.

[5] 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.

[6] 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

[7] 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.

[8] 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.

[9] 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

[10] 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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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

[16] 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.

[17] X. Chen and C. Cardie, "Unsupervised multilingual word embeddings," *arXiv preprint arXiv:1808.08933*, 2018.

[18] 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.

[19] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," 2013.

[20] D. C. Ferreira, A. F. Martins, and M. S. Almeida, "Jointly learning to embed and predict with multiple languages," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2019–2028.

[21] 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

[22] A. Klementiev, I. Titov, and B. Bhattarai, "Inducing crosslingual distributed representations of words," in *Proceedings of COLING 2012*, 2012, pp. 1459–1474.

[23] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 462–471.

[24] K. M. Hermann and P. Blunsom, "Multilingual models for compositional distributed semantics," *arXiv preprint arXiv:1404.4641*, 2014.

[25] S. Lauly, H. Larochelle, M. M. Khapra, B. Ravindran, V. Raykar, A. Saha *et al.*, "An autoencoder approach to learning bilingual word representations," *arXiv preprint arXiv:1402.1454*, 2014.

[26] S. Gouws, Y. Bengio, and G. Corrado, "Bilbowa: Fast bilingual distributed representations without word alignments," in *International Conference on Machine Learning*. PMLR, 2015, pp. 748–756.

[27] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015.

[28] M.-T. Luong, H. Pham, and C. D. Manning, "Bilingual word representations with monolingual quality in mind," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015, pp. 151–159.

[29] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016.

[30] 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

[31] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.

[32] 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.

[33] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," 2017.

[34] 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.

[35] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[36] 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