Visual Question Answering

Alexander Mirrington

This thesis is presented as part of the requirements for the conferral of the degree:

Bachelor of Information Technology (Honours)

Supervisors:
Dr. Caren Han
Dr. Josiah Poon

The University of Sydney
School of Computer Science

September 11, 2020

Declaration

I, *Alexander Mirrington*, declare that this thesis is submitted in partial fulfilment of the requirements for the conferral of the degree *Bachelor of Information Technology (Honours)*, from the University of Sydney, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Alexander Mirrington

September 11, 2020

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce volutpat lobortis ipsum. Pellentesque et rhoncus turpis. Nam hendrerit ligula eu justo hendrerit eleifend a eu nunc. Suspendisse felis risus, pellentesque at pretium a, imperdiet quis mauris. Pellentesque rutrum, mi sit amet faucibus posuere, neque velit mattis ligula, ac dapibus lectus enim vitae arcu. Aenean cursus, mi quis aliquam dapibus, justo neque posuere ipsum, sit amet pharetra sem nulla non nisi. Phasellus laoreet faucibus metus id convallis. Integer augue sapien, tempus eu aliquet vehicula, ornare ut ligula. Sed ex magna, tempus ut nulla a, pharetra cursus dui. Aenean dapibus dui commodo, pretium nunc gravida, dapibus elit. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae;

Acknowledgements

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Contributions

Summary of main contributions to the field.

1.2 Outline

Overall thesis outline.

Chapter 2

Literature Review

2.1 Visual Question Answering Datasets

Dataset	Year	Image Count	Question Count	Image Source	Question Source	Answer Type	Additional Data	Evaluation Metrics
DAQUAR [1]	2014	1K	12K	NYU-Depth V2 [2]	Both	Multi-label	-	Accuracy, WUPS
Visual Madlibs [3]	2015	10K	360K	COCO [4]	Human	Fill in the blank open-ended & multi-choice	-	Accuracy, BLEU
COCO-QA	-	-	-	COCO	-	-	-	Accuracy, BLEU
VQAv1 [5]	2015	204K	614K	COCO	Human	Open-ended, Multi-choice	COCO image captions	Accuracy[1]
Abstract Scenes	2015	50K	150K	Clip art, 2D	Human	Open-ended, Multi-choice	Image captions	Accuracy[1]
Changing Priors (CP) [6]	2018	≈204K	≈370K	COCO	Human	Open-ended	See VQAv1	Accuracy[1]

Compositional VQA (C-VQA) [7]	2017	204K	369K	COCO	Human	Open-ended	See VQAv1	Accuracy[1]
VQAv2 [8]	2017	204K	1.1M	COCO	Human	Open-ended	COCO image captions, Complementary image pairs	Accuracy[1]
Balanced Binary Abstract Scenes [9]	2016-17	31K	33K	Clip art, 2D	Human	Multiple choice	Image captions	Accuracy[1]
Changing Priors (CP) [6]	2018	≈219K	≈658K	COCO	Human	Open-ended	See VQAv2	Accuracy[1]
Visual Genome [10]	2016	108K	1.7M	COCO, YFCC100M [11]	Human	Open-ended	COCO annotations, Region descriptions, Scene graphs	Accuracy
Visual7W [12]	2016	47K	327K	COCO	Human	-	-	-

TDIUC [13]	2017	167K	1.6M	COCO, Visual Genome	Both	Open-ended	-	Per-question-type accuracy, regular & normalised arithmetic & harmonic mean accuracy
CLEVR [14]	2017	100K	999K	Computer-generated, 3D	Generated	Open-ended	Functional programs, Scene graphs	Accuracy
CoGenT-A & B	2017	100K	999K	Computer-generated, 3D	Generated	Open-ended	Functional programs, Scene graphs	Accuracy
Humans	2017	-	32K	CLEVR	Human	Open-ended	See CLEVR	Accuracy
GQA [15]	2019	113K	22.6M	-	Both	Open-ended	Scene graphs, Functional programs, Full-sentence answers	Accuracy, Consistency, Validity, Plausibility, Distribution, Grounding

Table 2.1: A comparison of relevant features of the most popular VQA datasets. Dataset variations are listed in regular font below their bolded counterparts.

2.2 Question Embedding in Visual Question Answering

2.3 Image Embedding in Visual Question Answering

2.4 Multi-modal Fusion in Visual Question Answering

[1]For open-ended answers, an answer is considered 'correct' if it matches at least three of the ten human-provided answers. For multiple choice answers, a traditional accuracy metric is used.

Chapter 3

Results

3.1 Performance Evaluation

3.2 Ablation Studies

3.3 Hyperparameter Optimisation

Chapter 4

Conclusion

4.1 Future Work

- Scene graph generation

Appendix A

Your first appendix

A.1 The title of the first section

The appendices work exactly the same way as chapters, they are numbered with letters rather than numbers though.

Bibliography

(1) M. Malinowski and M. Fritz, in *Advances in Neural Information Processing Systems 27*, ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, Curran Associates, Inc., 2014, pp. 1682–1690.

(2) N. Silberman, D. Hoiem, P. Kohli and R. Fergus, European conference on computer vision, 2012, pp. 746–760.

(3) L. Yu, E. Park, A. C. Berg and T. L. Berg, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2461–2469.

(4) T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, European conference on computer vision, 2014, pp. 740–755.

(5) S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick and D. Parikh, The IEEE International Conference on Computer Vision (ICCV), 2015.

(6) A. Agrawal, D. Batra, D. Parikh and A. Kembhavi, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4971–4980.

(7) A. Agrawal, A. Kembhavi, D. Batra and D. Parikh, "C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset", *arXiv preprint arXiv:1704.08243*, 2017.

(8) Y. Goyal, T. Khot, D. Summers-Stay, D. Batra and D. Parikh, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6904–6913.

(9) P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra and D. Parikh, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5014–5022.

(10) 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.(Article)", *International Journal of Computer Vision*, 2017, **123**, 32–73.

(11) B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth and L.-J. Li, "YFCC100M: The New Data in Multimedia Research", *Commun. ACM*, 2016, **59**, 64–73.

(12) Y. Zhu, O. Groth, M. Bernstein and L. Fei-Fei, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4995–5004.

(13) K. Kafle and C. Kanan, The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1965–1973.

(14) J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick and R. Girshick, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2901–2910.

(15) D. A. Hudson and C. D. Manning, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6700–6709.