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 21, 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 21, 2020

Abstract

Acknowledgements

Contents

List of Figures

List of Tables

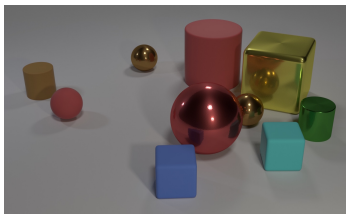
Chapter 1

Introduction

We live in an exciting technological era. In the last decade, we have witnessed the emergence of smart-home devices that harness the power of complex natural language understanding, text-to-speech and speech-to-text models to assist millions of people every day. The collaboration of hardware and software engineers have fueled the rapid evolution of the deep learning field, leading us into a world where semi-autonomous vehicles are becoming commonplace and face detection models are being used to unlock mobile devices. Given the wide adoption of deep learning (DL) techniques in industry for natural language processing (NLP) and computer vision (CV) tasks, we have seen a recent shift in the research community towards audio-visual and visio-linguistic tasks that require the learning of complex interactions between multiple input data modes. As a research community, we are excited by the technical challenge that these new tasks present, in addition to the extensive practical applications of multimodal reasoning tasks; for example, well-designed image captioning and visual question answering models could aid the visually impaired in consuming and understanding visual information in a natural manner when combined with existing speech-to-text and text-to-speech systems. Other multimodal tasks like text-to-image generation could be leveraged by law enforcement teams to build realistic composites of people or places relevant to an investigation, or by creatives as an external source of visual inspiration.

In this dissertation, I focus solely on visual question answering (VQA). More specifically, I delve into how we can leverage graphical representations of both visual and textual data to aid reasoning models in their decision-making processes. At its core, the VQA problem takes two inputs, an image and a question pertaining to one or more objects, relationships and/or concepts presented in the image. Given these inputs, we aim to answer the question, as illustrated in Figure 1.1. As humans, we implicitly solve the VQA problem every day when making decisions grounded on visual inputs. When crossing the road, we ask the implicit question *'Is it safe to cross the road?'*, formulate an answer based on visual signals, and then act upon our decision. The holy grail of VQA is to be able to perform such decision-making for all feasible combinations of visual inputs and questions. Naturally, the true distribution of these input combinations is unknown, motivating the first major

(a) Is there something to cut the vegetables with? *no*



(b) How many objects are either small cylinders or red things? *5*

(c) What kind of furniture is to the right of the chair? *sofa*

Figure 1.1: Example image, question and answer triples from the VQA 2.0 [1], CLEVR [2] and GQA [3], figures 1.1a, 1.1b and 1.1c respectively.

hurdle in VQA research: How do we design a dataset that effectively emulates subtle relationships between questions and images as presented in real-world situations where VQA would prove useful? As evident in Figure 1.1, this question is still open to interpretation; the VQA 2.0 dataset contains real-world images and free-form questions often requiring conceptual reasoning, whilst CLEVR leverages generated images alongside questions requiring more compositional reasoning, targeting logical reasoning operations like counting and boolean arithmetic. I will elaborate on these ideas in Section 2.1.

Assuming we do have an ideal dataset that effectively models real-world VQA problems, we need some way of combining both visual and textual data in a way that enables a model to perform the complex reasoning required for answer formulation.

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

In this section, I address the first major hurdle for the VQA task as introduced in Chapter 1: How do we design a dataset that effectively emulates subtle relationships between questions and images as presented in real-world situations where VQA would prove useful? More formally, we wish to develop a dataset $\mathcal{X}$ such that $x \sim \mathcal{D}\ \forall x \in \mathcal{X}$ for some fixed, underlying distribution $\mathcal{D}$ [4]. For a dataset to be *useful*, we require that $\mathcal{D}$ captures the distribution of real-world VQA problems that we wish to solve.

Dataset	Year	Image Count	Question Count	Image Source	Question Source	Answer Type	Additional Data	Evaluation Metrics
DAQUAR [5]	2014	1K	12K	NYU-Depth V2 [6]	Both	Multi-label	-	Accuracy, WUPS
Visual Madlibs [7]	2015	10K	360K	COCO [8]	Human	Fill in the blank open-ended & multi-choice	-	Accuracy, BLEU
COCO-QA	-	-	-	COCO	-	-	-	Accuracy, BLEU
VQAv1 [9]	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) [10]	2018	≈204K	≈370K	COCO	Human	Open-ended	See VQAv1	Accuracy[1]
Compositional VQA (C-VQA) [11]	2017	204K	369K	COCO	Human	Open-ended	See VQAv1	Accuracy[1]

VQAv2 [1]	2017	204K	1.1M	COCO	Human	Open-ended	COCO image captions, Complementary image pairs	Accuracy[1]
Balanced Binary Abstract Scenes [12]	2016-17	31K	33K	Clip art, 2D	Human	Multiple choice	Image captions	Accuracy[1]
Changing Priors (CP) [10]	2018	≈219K	≈658K	COCO	Human	Open-ended	See VQAv2	Accuracy[1]
Visual Genome [13]	2016	108K	1.7M	COCO, YFCC100M [14]	Human	Open-ended	COCO annotations, Region descriptions, Scene graphs	Accuracy
Visual7W [15]	2016	47K	327K	COCO	Human	-	-	-

TDIUC [16]	2017	167K	1.6M	COCO, Visual Genome	Both	Open-ended	-	Per-question-type accuracy, regular & normalised arithmetic & harmonic mean accuracy
CLEVR [2]	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 [3]	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.

2.1.1 Early Work

Given the recent rise in popularity of Visual Question Answering tasks, Antol et al. recognised a need for a new, sufficiently large dataset that contained free-form, open-ended questions that still allowed results to be quantified; this need was realised by two datasets: an open-ended dataset derived from COCO [8], as well as a smaller synthetic dataset with similarly created questions.[9] The key contributions of these datasets were two-fold:

1. The real-world image dataset provided the first feasible dataset for benchmarking VQA models, thanks to its large size compared to existing VQA datasets like DAQUAR [5] and COCO-QA [17]. Whilst smaller, the synthetic dataset allowed researchers to investigate the underlying theory behind visual reasoning without focusing on the difficult task of extracting features from images.
2. Both datasets employed a new open-ended accuracy metric that rewarded models for feasible answers based on answer distributions generated from human answers to questions; an answer is considered "100% accurate if at least 3 workers provided that exact answer." Antol et al. Whilst not a perfect metric due to the coarseness of the collected human answer distributions, the authors justify that metrics such as BLEU and ROUGE from other NLP tasks are not suited to the dataset, and thus promote the importance of further research into new metrics for VQA tasks.

2.1.2 Mitigating the Exploitation of Language Priors

Despite VQA 1.0's strengths, the authors identify the first main hurdle for VQA datasets, highlighting that if left to their own devices, VQA models will exploit statistical priors present in textual information. They describe how language-only methods perform surprisingly well compared to models that harness both question and image features, noting that a simple "per-question type prior" model achieves 71.03% and 35.77% on binary and numeric questions, and an LSTM with only question word embeddings performs similarly at 78.20% and 35.68% [9]. Further studies [1, 12] reinforce these observations, highlighting the skewed answer distributions of both the real-world and synthetic VQA datasets, demonstrating that models could leverage language priors to effectively answer some types of questions without requiring any knowledge of associated images. Zhang et al. focus on the imbalance in the answer distribution of binary questions in the VQA dataset, pointing out that 'yes' is the answer to 68% of all binary questions in the abstract scenes dataset. Furthermore, they propose a potential solution that allows the answer distribution of binary questions in the existing abstract scenes dataset to be balanced: for each (image, question) pair, humans were tasked with the creation of a new scene using existing clip-art objects [18] that was similar to the existing image, but had a negated answer on the same question, as shown in Figure 2.1 below.

[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.

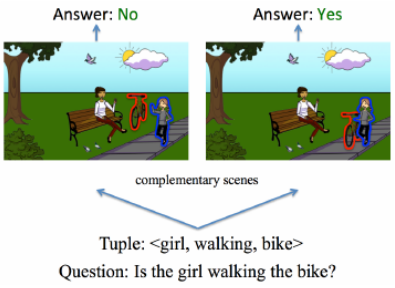


Figure 2.1: Zhang et al. demonstrate their answer distribution balancing technique. Given the scene on the left and the question "Is the girl walking the bike", workers were tasked with creating a scene that differs from the scene on the left but has the opposite answer to the question, as illustrated by the scene and answer on the right.

Despite this method's success in reducing the proportion of unbalanced (image, question) pairs by almost 72%, 20.48% (image, question) pairs in the balanced dataset did not have a complementary (image, question) pair; 5.93% because a complementary scene could not be created due to limitations of the abstract scenes clip-art library, and 14.55% due to disagreement between the answers collected by workers for a given scene. A similar approach was used for non-binary questions in the creation of the VQA 2.0 dataset, as summarised in Table 2.1 above. Hudson and Manning approach the task of dataset balancing from a different angle, aiming to mitigate biases in the answer distributions of the GQA dataset whilst maintaining some degree of representation of real-world priors.

Whilst it is certainly important that dataset creators mitigate biases in their work, it is also the responsibility of VQA model authors to ensure that the ways in which image and question data is fed to their models is also able to be justified; many models, incorporate BiLSTMs [19] or LSTMs [20] to obtain question-level representations, however by processing the input data in this manner, it becomes difficult to determine the extent to which the reasoning part of the model is exploiting probabilistic priors in the dataset. By embedding and processing the question as a graph, interpretability studies can be performed on the model to ensure the reasoning part of the model performs as intended.

Approaches and Architectures

It has been observed that the architectures that leverage additional information such as scene graphs rather than just raw image features tend to perform better. This is clearly illustrated by Hudson and Manning in Figure 2.2 below:

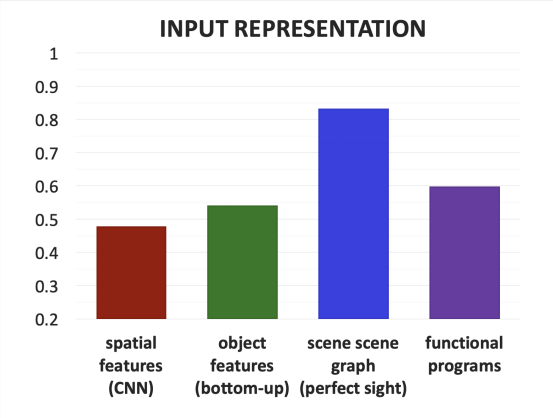


Figure 2.2: Hudson and Manning illustrate the effect of various forms of input embeddings, noting that accuracy of the trained MAC network [19] increases drastically with the semantic richness of the embedding.

Hudson and Manning rightly point out that scene graphs cannot always be relied on as a supervisory data source, however with the rise of new scene graph generation models, scene graphs can be generated automatically from images. Since these graphs no longer have to be created through labour-intensive means as seen in the Visual Genome dataset, the use of scene graphs as an image embedding is more feasible, as they can be created quickly for any input image.

Despite this, very few (if any) models are effectively using similar graph embeddings for question data. By embedding this data similarly to the image, it becomes possible to perform reasoning steps about the image based on the question and vice versa. Commonly used image and quesion embeddings include:

Image Representations

- CNN features
- Image object features and annotations
- Scene Graphs

Question Representations

- Traditional word embeddings e.g. word2vec, GloVe
- Recurrent neural network embeddings, e.g. LSTM, GRU and their bidirectional counterparts.
- Dependency-based word embeddings [**levy2014dependency**]
- Functional Programs

Combined Question and Image Representations

- Question-Image co-attention [**lu2016hierarchical**]
- Bilinear models e.g. BLOCK Fusion [**ben2019block**]

Module Networks

Whilst deep neural networks have proved useful for many classification tasks due to their inherent ability to extract correlations between input and output, they are difficult to interpret and struggle to perform complex reasoning tasks. In order to increase interpretability and coerce deep learning models to perform more structured reasoning steps, Andreas et al. proposed the creation of multiple network 'modules', each capable of performing a specific reasoning step, similar to each component of the functional programs found in the CLEVR [2] and GQA [3] datasets.

Whilst attaining a state-of-the-art result in 2016 on the VQA 1.0 test dataset [9] of 58.7%, the per-question-type statistics report only a 2.5% and 1.4% performance improvement over a LSTM language-only baseline on binary and numerical questions despite the 16% improvement on 'other'-type questions. This suggests that the success of the model is likely clouded by statistical biases in the dataset, evidenced by the use of a text-only model in the ensemble shown in Figure 2.3. As discussed by Hudson and Manning, the model depends on unreliable parsers and their "hand-crafted" design lacks extensibility. Moreover, attention maps on raw CNN features are used as the primary form of image data input. Whilst supplementary information like scene graphs and object annotations were not available at the time of writing, these data formats would allow for additional interactions and data transfer between modules, likely improving the network's ability to perform logical reasoning processes.

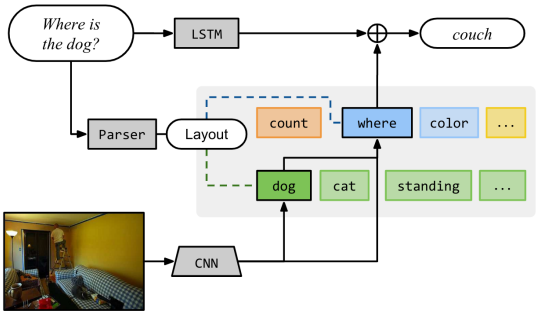


Figure 2.3: Andreas et al. propose the Neural Module Network, in which reasoning operations are strung together based on parser outputs and use combinations and transformations of attention maps on image features to predict a final answer to the question.

andreas2016learning improve upon their earlier work with a more robust method for generating compositional module layouts from questions, combining LSTM embeddings of the question with representations of candidate module layouts to select

a module layout using a reinforcement learning approach. This model sees only a slight improvement over their previous iteration, however provides a novel framework for a variety of graph-based reasoning tasks; the scene graphs and functional programs in CLEVR and GQA are strong supervisory training data for such models, and recent scene graph generation techniques [**yang2018graph**] would prove useful in generalising models to images that have not been previously annotated.

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

Chapter 3

Methodology

3.1 Architecture Overview

- Question input module
- Scene graph input module (knowledge base)
- Reasoning module
- Output module

3.2 Question Input Module

3.3 Scene Graph Input Module

3.4 Reasoning Module

- Concat + linear fusion (other fusion types?)
- Bottom-up
- MAC network

3.4.1 Bottom-up

ReLU proved to yield higher results over gated tanh when paired with GAT/GCN embeddings. In the original paper, CNN/R-CNN features are extracted in the pre-processing step, meaning there is no need for gradient propagation to the knowledge base embedding. Preliminary tests showed a need for learnable embeddings in graph convolutional models, and thus a need for end-to-end propagation of gradients.

3.4.2 MAC Network

In order to leverage the computational benefits of sparse tensor operations implemented in PyG [22], I used a PyTorch [23] re-implementation of the MAC network, which has been trained to 98.6% on the CLEVR dataset [24], just 0.3% shy of the official result reported by Hudson and Manning. Notably, this re-implementation accounts for minor details that were omitted from the original MAC network paper but enabled by default in the official MAC network repository citation required.

- Initial tests showed little performance difference between conditioning the current control state on previous control states.

3.5 Output Module

Chapter 4

Results

4.1 Performance Evaluation

4.2 Ablation Studies

4.3 Hyperparameter Optimisation

- wandb [25] implementation of bayesian optimisation, using the hyperband early stopping method

Chapter 5

Discussion

5.1 Error Analysis

Chapter 6

Conclusion

6.1 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) 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.

(2) 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.

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

(4) M. Mohri, A. Rostamizadeh and A. Talwalkar, *Foundations of machine learning*, MIT press, 2018.

(5) 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.

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

(7) 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.

(8) 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.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

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

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

(17) M. Ren, R. Kiros and R. Zemel, in *Advances in Neural Information Processing Systems 28*, ed. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, Curran Associates, Inc., 2015, pp. 2953–2961.

(18) C. L. Zitnick and D. Parikh, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

(19) D. A. Hudson and C. D. Manning, "Compositional Attention Networks for Machine Reasoning", *CoRR*, 2018, **abs/1803.03067**.

(20) J. Andreas, M. Rohrbach, T. Darrell and D. Klein, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

(21) D. A. Hudson and C. D. Manning, "GQA: a new dataset for compositional question answering over real-world images", *CoRR*, 2019, **abs/1902.09506**.

(22) M. Fey and J. E. Lenssen, ICLR Workshop on Representation Learning on Graphs and Manifolds, 2019.

(23) A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, in *Advances in Neural Information Processing Systems 32*, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, Curran Associates, Inc., 2019, pp. 8026–8037.

(24) C. Eyzaguirre and A. Soto, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12817–12825.

(25) L. Biewald, *Experiment Tracking with Weights and Biases*, Software available from wandb.com, 2020.