



Universiteit
Leiden
The Netherlands

Exploring deep learning for multimodal understanding

Lao, M.

Citation

Lao, M. (2023, November 28). *Exploring deep learning for multimodal understanding*. Retrieved from <https://hdl.handle.net/1887/3665082>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3665082>

Note: To cite this publication please use the final published version (if applicable).

Exploring Deep Learning for Multimodal Understanding

Mingrui Lao

Copyright © 2023 Mingrui Lao, All Rights Reserved

ISBN 978-94-6419-982-6

Printed by Gildeprint, The Netherlands

Cover design: Mingrui Lao

Exploring Deep Learning for Multimodal Understanding

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 28 november 2023
klokke 11.15 uur

door

Mingrui Lao

geboren te Changsha, China
in 1995

Promotores: Prof. dr. M.S. Lew
Prof. dr. A. Plaat

Promotiecommissie: Prof. dr. T.H.W. Bäck
Prof. dr. J. Batenburg
Prof. dr. K.J. Wolstencroft
Prof. dr. B.P.F. Lelieveldt (Leiden University Medical Center)
Prof. dr. C.G.M. Snoek (University of Amsterdam)



Mingrui Lao was financially supported through the China Scholarship Council (CSC) to participate in the PhD programme of Leiden University. Grant number 201903170171

The research in this thesis was performed at the LIACS MediaLab, Leiden University, The Netherlands, and we would like to thank the NVIDIA Corporation for the donation of GPU cards.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Background and Related Work	3
1.2.1	Multimodal Attention & Fusion	4
1.2.2	Language Bias Problem	5
1.2.3	Audio Visual Question Answering	6
1.2.4	Lifelong Learning	8
1.2.5	Federated Learning	9
1.3	Thesis Outline and Research Questions	10
1.4	Main Contributions	14
1.4.1	Task Contribution	15
1.4.2	Technical Contributions	18
2	Recent Advances in Robust Visual Question Answering: Challenges, Benchmarks and Strategies	21
2.1	Introduction	22
2.2	Challenges	24
2.2.1	Shortcut Bias	24
2.2.2	Multimodal inputs variations	25
2.2.3	Reasoning Consistency	26
2.2.4	Domain Robustness	28
2.3	Benchmarks	28
2.3.1	Distribution-Shift driven Benchmarks	29
2.3.2	Sub-Question driven Benchmarks	32
2.3.3	Visual perturbations driven Benchmark	35
2.3.4	Multi-Task Benchmarks	35
2.3.5	Human-Adversarial Benchmarks	36
2.4	Bias Mitigation Techniques	37
2.4.1	Anti-Bias Model Architectures	38
2.4.2	Instance-Level Gradient Adjustment	39
2.4.3	Debiased-Regularization based Techniques	42
2.4.4	Distribution-Rebalance based Approaches	44
2.5	Conclusions	46

3	Multi-Stage Hybrid Embedding Fusion Network for Visual Question Answering	47
3.1	Introduction	48
3.2	Related Work	50
3.2.1	Visual Question Answering	50
3.2.2	Multimodal Fusion	50
3.2.3	Dual Embedding	51
3.2.4	Multi-stage Learning	51
3.3	Proposed Model	52
3.3.1	Hybrid Embedding Fusion	52
3.3.2	Multi-stage Fusion Structure (MFS)	53
3.3.3	Multi-stage Hybrid Embedding Fusion Network (MHEF)	55
3.3.4	VQA framework with MHEF	56
3.4	Experiments	58
3.4.1	Datasets and Evaluation Metric	58
3.4.2	Implementation Details	58
3.4.3	Ablation Study	59
3.4.4	Fusion Scheme Comparison	61
3.4.5	Qualitative analysis	62
3.4.6	Comparison with the State-of-the-Art	63
3.5	Conclusion	65
4	A Language Prior Based Focal Loss for Visual Question Answering	67
4.1	Introduction	68
4.2	Methodology	70
4.2.1	Preliminary	70
4.2.2	LP-Focal Loss	70
4.3	Experiments	73
4.3.1	Datasets	73
4.3.2	Implementation details	73
4.3.3	Ablation studies	73
4.3.4	Compared with the standard focal loss	75
4.3.5	Comparison with the state-of-the-art	75
4.3.6	Case study	76
4.4	Conclusion	77
5	From Superficial to Deep: Language Bias driven Curriculum Learning for Visual Question Answering	79
5.1	Introduction	80
5.2	Related work	82
5.3	Language Bias driven Curriculum Learning	83
5.3.1	Difficulty Metric	85
5.3.2	Curriculum Selection Function	86

5.3.3	Knowledge Distillation	88
5.4	Experiments	88
5.4.1	Datasets and Baselines	88
5.4.2	Implementation Details	90
5.4.3	Ablation Study	90
5.4.4	State-of-the-art comparison	93
5.4.5	Experiments on small-scale datasets	94
5.4.6	Qualitative Results	95
5.5	Case Study	95
5.6	Conclusion	96
6	COCA: COLlaborative CAusal Regularization for Audio-Visual Ques- tion Answering	99
6.1	Introduction	100
6.2	A Causal View with Bias Revelation on AVQA	102
6.2.1	Causal Graph of AVQA	102
6.2.2	Potential Shortcut Biases in AVQA	103
6.3	Methodology	104
6.3.1	Bias-Centered Causal Regularization	104
6.3.2	Multi-Shortcut Collaborative Debiasing	106
6.4	Experiments	108
6.4.1	State-Of-The-Art Comparisons	109
6.4.2	Ablation Studies	110
6.5	Related Works	111
6.6	Conclusion	113
7	Multi-Domain Lifelong Visual Question Answering via Self-Critical Distillation	115
7.1	Introduction	116
7.2	Related Works	118
7.2.1	Multi-Domain Learning in Visual Question Answering	118
7.2.2	Lifelong Learning in Vision-Language Tasks	118
7.2.3	Knowledge Distillation for Overcoming Forgetting	119
7.3	Multi-Domain Lifelong Visual Question Answering	120
7.3.1	Problem Definition	120
7.3.2	Analyses of Multi-Domain Lifelong VQA Benchmark	120
7.3.3	Baseline Approach	121
7.3.4	Limitations	123
7.4	Self-Critical Distillation	124
7.4.1	Logits-level SCD	124
7.4.2	Feature-level SCD	126
7.4.3	Optimization	128

7.5	Experiments	128
7.5.1	Implementation Details	128
7.5.2	Evaluation Metrics	129
7.5.3	Datasets	130
7.5.4	Performance Evaluation	130
7.5.5	Ablation Study	132
7.5.6	Qualitative Results	135
7.6	Conclusion	136
8	FedVQA: Personalized Federated Visual Question Answering over Heterogeneous Scenes	139
8.1	Introduction	140
8.2	Related Work	142
8.2.1	Visual Question Answering	142
8.2.2	Personalized Federated Learning	142
8.2.3	Forgetting Issue in Personalized Learning	144
8.3	Methodology	145
8.3.1	Benchmark Formulation	145
8.3.2	Training Pipeline	146
8.3.3	FedP ³ : Pairwise Preference Preserving	147
8.4	Experiments	152
8.4.1	Datasets	152
8.4.2	Implementation Details	152
8.4.3	Comparative Approaches	152
8.4.4	State-of-the-art Comparisons	153
8.4.5	Ablation Study	154
8.4.6	Case Study	156
8.5	Conclusion	157
9	Conclusions	159
9.1	Main Findings	159
9.2	Limitations	162
9.2.1	Algorithmic perspective	162
9.2.2	Practical perspective	163
9.3	Future Research Directions	164
9.3.1	VQA Benchmark beyond Accuracy	164
9.3.2	Bias Mitigation in the Open-Set Learning Settings	164
9.3.3	Federated Lifelong Visual Question Answering	165
	Bibliography	167
	List of Abbreviations	187
	English Summary	189

Nederlandse Samenvatting	193
Curriculum Vitae	199
Publication List	201

