
Intelligent Resource Management with Deep
Reinforcement Learning in Device-to-Device
Communication

by

David Cotton

Thesis submitted in fulfilment of the requirements for the degree of

Master of Analytics (Research)

Under the supervision of

Zenon Chaczko, Doan Hoang & Massimo Piccardi

at

School Electrical and Data Engineering
Faculty Engineering and Information Technology
University of Technology Sydney
NSW, 2007, Australia

March 2022

Certificate of original authorship

I, *David Cotton* declare that this thesis, is submitted in fulfilment of the requirements for the award of *Master of Analytics (Research)*, in the *School of Electrical and Data Engineering, Faculty Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: _____

[David Cotton]

DATE: 1st March, 2022

PLACE: Sydney, Australia

Acknowledgements

I would like to thank the many people who have helped me complete this thesis. Thanks to my supervisors Dr. Zenon Chackzo, Prof. Doan Hoang and Prof. Massimo Piccardi for all the support and insightful comments that have immeasurably improved this work. Zenon, thank you for all your help and friendship. You helped me find my academic feet when I was lost and build my confidence to publish our research. Doan, thank you for your challenging questions and stimulating discussions. Massimo, thank you for your fantastic advice and support. Whenever there was a challenge, you were always two-steps ahead with timely, clear feedback and a path forward. I would also like to acknowledge Prof. Richard Xu for providing me the opportunity to undertake this degree, sharing your deep machine learning knowledge and pushing my work to a higher level. Thank you Dr. Jason Traish for your mentorship and supporting the development of my ideas. You were always extremely generous with your time and knowledge.

I would also like to thank my friends in academia Sam Hartridge and Prof. Brendan Mulhern. Even though you specialise in fields very different from mine, listening to me whinge and providing me your helpful insight has helped me untangle byzantine academic process. Also, to friends outside academia with no interest in machine learning that allowed for an escape from this thesis when needed.

I would also like to thank my kids, James and Ellie. I am sorry that completing this degree has taken so much of our time together and I promise to spend more time playing with you from now on. Lastly, and most importantly, my beautiful wife Clare. This thesis deserves to have your name on the cover for all the support you have provided. When your friends and family said I was crazy for leaving my job for the destitute student world, you backed me. Through the highs and many lows of this degree you have helped me vent, unpack and calculate the best response. Despite having no interest in maths or computers, you have probably proofread more than twenty drafts of this thesis and the research papers it contains, of impossibly dense and incredibly boring academic writing. For all this and everything I've missed, I am truly, truly grateful of your love and support.

List of Publications

Conference :

1. D. Cotton, J. Traish and Z. Chaczko, “Coevolutionary Deep Reinforcement Learning”, *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Canberra, Australia, 1–4 December 2020.
2. D. Cotton and Z. Chaczko, “GymD2D: A Device-to-Device Underlay Cellular Offload Evaluation Platform”, *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, Nanjing, China, 29 March–1 April 2021.

Table of Contents

List of Publications	iii
List of Figures	viii
List of Tables	ix
Acronyms	x
1 Introduction	1
1.1 Background	1
1.1.1 Research Problems	3
1.1.2 Research Questions	4
1.2 Aim and Objectives	4
1.2.1 Aim	4
1.2.2 Objectives	4
1.3 Methodology	5
1.3.1 Data Management Plan	5
1.4 Results	5
1.5 Organisation	6
2 Literature Review	8
2.1 Cellular Networks	8
2.1.1 Propagation and Path Loss	9
2.1.2 Interference	9
2.1.3 Spectral Efficiency	10
2.1.4 Traffic Offloading	10
2.1.5 Opportunistic Spectrum Access	11
2.2 D2D Communications	12

TABLE OF CONTENTS

2.2.1	D2D Advantages	12
2.2.2	D2D Use Cases	13
2.2.3	D2D Taxonomy	15
2.2.4	D2D Challenges	17
2.3	D2D Resource Management	19
2.3.1	Optimisation Classification	19
2.3.2	Simulation Models	22
2.3.3	RL-based D2D Resource Management	23
2.4	Reinforcement Learning	24
2.4.1	Markov Decision Processes	24
2.4.2	Policies and Value Functions	25
2.4.3	Value-Based Methods	26
2.4.4	Policy Gradient Methods	26
2.4.5	Actor-Critic Methods	27
2.5	Deep Reinforcement Learning	28
2.5.1	Deep Q-Networks	28
2.5.2	Double DQN	29
2.5.3	Dueling DQN	29
2.5.4	Prioritised Replay	30
2.5.5	Multi-Step Learning	30
2.5.6	NoisyNets	31
2.5.7	Categorical DQN	31
2.5.8	Rainbow DQN	32
2.5.9	Asynchronous Advantage Actor-Critic	33
3	Coevolutionary Deep Reinforcement Learning	34
3.1	Overview	34
3.2	Preliminaries	34
3.2.1	Multi-Agent Reinforcement Learning	34
3.2.2	Competitive Training Methods	35
3.3	Background	37
3.3.1	Self-Play	37
3.3.2	Coevolutionary Algorithms	39
3.3.3	Evolutionary Reinforcement Learning	40
3.4	Coevolutionary Reinforcement Learning	40
3.4.1	Training Procedure	41

TABLE OF CONTENTS

3.4.2	Survivor Selection	42
3.4.3	Related Work	43
3.5	Method	43
3.5.1	Evaluation Environment	44
3.5.2	Observation Space	45
3.5.3	Action Space	46
3.5.4	Reward Function	47
3.5.5	Neural Network Architecture	47
3.5.6	Agent Configuration	48
3.6	Evaluation	49
3.6.1	Evaluated Algorithms	49
3.6.2	Methodology	50
3.6.3	Results	51
3.6.4	Ablation Study	51
3.7	Conclusion	54
4	GymD2D: A Device-to-Device Underlay Cellular Offload	
	Evaluation Platform	56
4.1	Overview	56
4.2	Preliminaries	57
4.2.1	Device-to-Device Communication	57
4.2.2	OpenAI Gym	59
4.2.3	Network Simulation	59
4.3	GymD2D	60
4.3.1	Design Principles	60
4.3.2	Architecture	61
4.3.3	System Model	61
4.3.4	Path Loss Models	63
4.3.5	Network Simulator	64
4.3.6	Gym Environment	65
4.3.7	Capabilities and Limitations	65
4.4	Evaluation Methods	68
4.4.1	Agent Architecture	68
4.4.2	Observation Space	68
4.4.3	Action Space	69
4.4.4	Reward Function	69

TABLE OF CONTENTS

4.4.5	Neural Network Architecture	70
4.4.6	Agent Configuration	70
4.5	Evaluation	72
4.5.1	Methodology	72
4.5.2	Results	72
4.5.3	Discussion	73
4.6	Conclusion	76
5	Conclusion	78
5.1	Summary of Results	78
5.2	Future Work	79
	Bibliography	81

List of Figures

1.1	D2D Cellular Offload	2
2.1	D2D Use Cases	14
2.2	The Reinforcement Learning Cycle	24
3.1	Coevolutionary RL Training Cycle	41
3.2	Connect Four Observation Representation	46
3.3	Connect Four Action Mask	47
3.4	Connect Four Action Space	48
3.5	Evaluated Algorithms	50
3.6	Evaluation Win Percentage	52
3.7	Evaluation Neural Network Loss	52
3.8	Evaluation Temporal-Difference Error	53
3.9	Evaluation Maximum Q-Values	53
3.10	CLaRE Ablations	54
4.1	GymD2D Architecture Diagram	61
4.2	Network Simulator Architecture	64
4.3	Agent Architecture	69
4.4	Total System Capacity of All Agents	74
4.5	Total System Capacity of DRL Agents	74
4.6	Total DUE Capacity	75
4.7	Mean DUE Transmit Power	75

List of Tables

3.1	Environment Configuration	45
3.2	Rainbow DQN Hyperparameters	49
4.1	GymD2D BS Configuration Parameters	63
4.2	GymD2D UE Configuration Parameters	63
4.3	GymD2D Environment Configuration Parameters	66
4.4	Rainbow DQN Hyperparameters	71
4.5	SAC Hyperparameters	71
4.6	A2C Hyperparameters	71
4.7	Simulation Parameters	73

Acronyms

- 3GPP** 3rd Generation Partnership Project. 12
- 5G** fifth generation. 3, 12
- A2C** Advantage Actor-Critic. ix, 68, 71
- A3C** Asynchronous Advantage Actor-Critic. 33, 71
- AI** artificial intelligence. 37, 40
- API** application programming interface. 59, 65
- AWGN** additive white Gaussian noise. 63
- BS** base station. ix, 3, 8, 15, 16, 59, 62–65, 67–69, 72
- CCI** co-channel interference. 9, 10
- CERL** collaborative evolutionary reinforcement learning. 40, 43
- CLaRE** Coevolutionary Learning and REinforcement. 5, 6, 40, 43, 49, 51, 79
- CNN** convolutional neural networks. 45, 47, 65
- CoEA** coevolutionary algorithms. 39
- CSI** channel state information. 21
- CUE** cellular user equipment. 15–17, 20, 22, 58, 59, 62–66, 68, 69, 72–74, 76, 77
- D2D** device-to-device. v, 1–8, 11–23, 56–63, 68, 69, 72–79

- DNN** deep neural network. 3, 8, 23, 28
- DQN** Deep Q-Networks. ix, 23, 28–32, 43, 47–49, 68, 70, 71, 76
- DRL** deep reinforcement learning. viii, 1, 3–8, 22, 23, 28, 40, 43, 54, 58, 60, 65, 68–70, 72–79
- DUE** D2D user equipment. viii, 13, 15–18, 20, 22, 23, 58, 62–66, 68–70, 72–76
- EA** evolutionary algorithm. 39, 40
- EIRP** effective isotropic radiated power. 62
- ERL** evolutionary reinforcement learning. 40, 43
- FSP** fictitious self-play. 36
- FSPL** free space path loss. 63, 64
- GAE** Generalised Advantage Estimator. 71
- HetNet** heterogeneous cellular network. 10, 20, 22
- IoT** Internet of things. 14
- LTE** Long Term Evolution. 3, 65, 72
- LTE Advanced** Long Term Evolution Advanced. 12
- M2M** machine-to-machine. 14, 19, 67
- MARL** multi-agent reinforcement learning. 34, 35
- MBS** macro base station. 10, 22, 58, 62
- MCTS** Monte Carlo tree search. 44, 45, 49–52
- MDP** Markov decision process. 24–26, 36
- MEC** mobile edge caching. 4, 13
- NR** New Radio. 3, 65, 72

ACRONYMS

- OFDMA** orthogonal frequency division multiple access. 22, 58, 62
- PLE** path loss exponent. 63, 72
- PU** primary user. 11, 15
- QoS** quality of service. 23, 67
- RB** resource block. 20, 22, 23, 58, 62–66, 69, 71, 72, 76
- ReLU** Rectified Linear Unit. 48, 70
- RF** radio frequency. 1, 13, 18
- RL** reinforcement learning. v, viii, 2, 3, 5–8, 22–26, 28, 31, 34–38, 40, 43, 45, 54, 56, 58, 59, 78
- RRM** radio resource management. 2, 6, 7, 18, 19, 21, 56–58, 61, 64, 65, 72–79
- SAC** soft actor critic. ix, 68, 70, 71
- SBS** small base station. 10, 21, 22
- SINR** signal-to-interference-plus-noise ratio. 9, 10, 18, 20, 57, 59, 63, 65, 69
- SNR** signal-to-noise ratio. 69
- SU** secondary user. 11
- UE** user equipment. ix, 3, 8, 10, 12, 14, 17, 18, 20, 63–65, 67, 76, 79
- V2V** vehicle-to-vehicle. 14, 19, 67
- Wi-Fi** wireless fidelity. 2, 10

Abstract

Radio resource management in device-to-device cellular offload can be optimised to increase network capacity, quality of service, energy efficiency, lower latency and provide more resilient networks. However, this resource optimisation problem is both NP-Hard and required to operate at a millisecond timescale, limiting feasible solutions.

In this thesis, we investigate how deep reinforcement learning can be applied to improve resource allocation. To empirically demonstrate our approach, we develop a network simulator for device-to-device cellular offload research. We also introduce an improved self-play algorithm for training reinforcement learning without expert guidance.

We apply our self-play training algorithm to the game Connect Four. Leveraging the competitive pressures of coevolution, we improve the performance of agents trained with our method, achieving a 15% higher win rate. Furthermore, agents exhibit more stable training dynamics and suffer fewer performance regressions.

We evaluate our network simulator and demonstrate deep reinforcement learning can significantly increase network capacity. Our network simulator reduces research friction and provides an evaluation platform to compare, share and build upon results. Our toolkit is provided to other researchers as open-source software.

