HOCHSCHULE OSNABRÜCK
UNIVERSITY OF APPLIED SCIENCES

SCIENTIFIC PAPER PRESENTATION

DEEP REINFORCEMENT LEARNING AGENT TUNING FRAMEWORK

Fachseminar (SoSe 2020)

Aleksei Ivanov

OUTLINE

- Introducton
- Concepts and methods
  - What is reinforcement learning?
  - Current advancements (state-of-the-art)
  - Problem area
  - Methodology
- Implementation
- Experiments and results
- Summary & outlook

INTRODUCTION

Applying machine learning algorithms from scratch is not straight-forward. Neural networks have many ***hyperparameters:***

- Learning rate (fixed by Adagrad)
- Types of layers
- Number of layers
- Each layer has their own parameters:
  - Fully connected: *number of neurons and initial values*
  - Convolutional: *kernel size, strides, etc.*

INTRODUCTION

Deep reinforcement learning builds on top of neural networks

More paramaters!

- Agent learning rate (different from neural network)
- Discount factor
- Reward function
- Replay memory capacity
- Memory sample size

INTRODUCTION

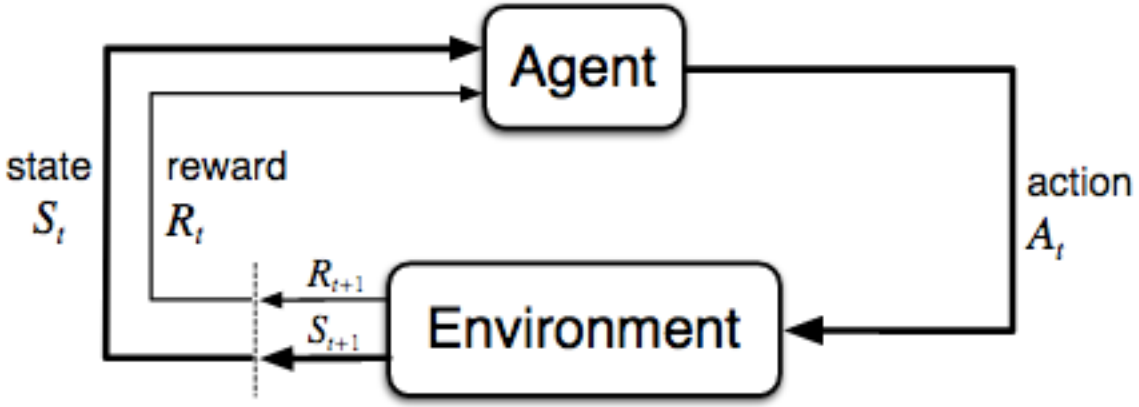
The aim of this work is to try and find systematic approach to parameter tuning using DQN in discrete state- and action-spaces

Custom grid-world environment is presented with state space in the form of vision around the agent instead of absolute position

Developed solution allows to run **multiple passes** of the same agent configuration to get better statistical stability of the results

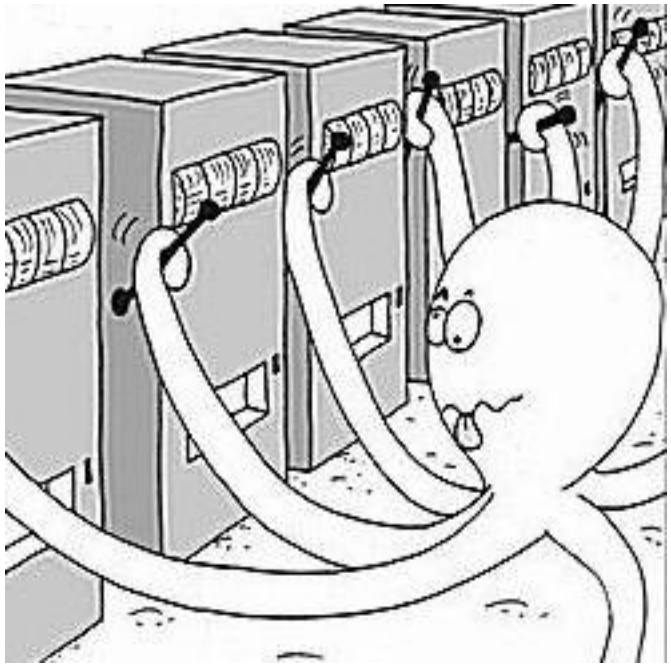
WHAT IS REINFORCEMENT LEARNING?

"Reinforcement learning problems involve learning what to do in context of how to map situations to actions so as to maximize a numerical reward signal" [Sut18]



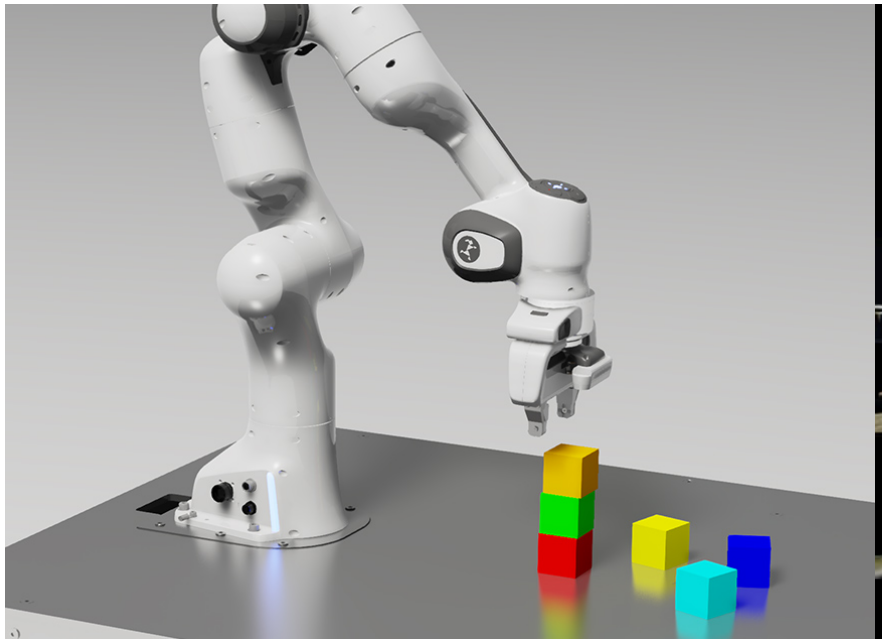
Typical agent–environment interaction in reinforcement learning [Sut18]

DISCRETE ACTION SPACE (K-ARMED BANDITS)

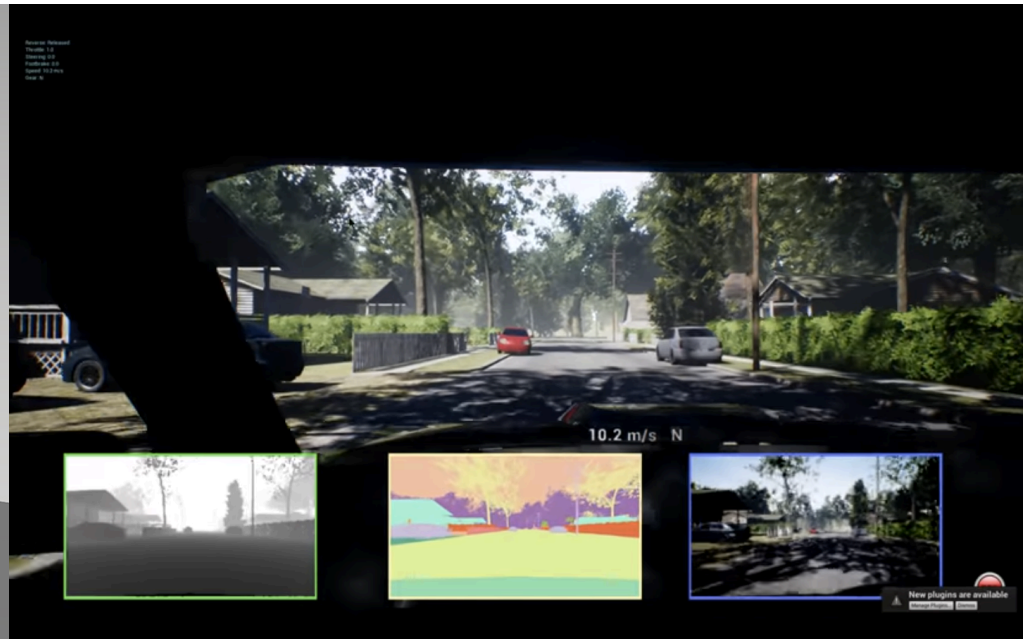


Source: Microsoft Research

CONTINUOUS ACTION SPACE (ACTOR–CRITIC)



Source: NVIDIA



Source: Microsoft Research

CURRENT ADVANCEMENTS

A number of modifications have been proposed:

- Multi-agent learning [Doy02], [Foe16]
- Reward shaping
- Policy search [Yng03]
- Learning without exploration [Fuj19]
- Self-supervision [Eva17]

CURRENT PROBLEMS

Lack of research in how memory capacity and sample size affect learning performance, along with other hyperparameters

No systematic approach – modern algorithms "just work", but before that they require manual configuration

Recently there were attempts to stand back and **understand why simple algorithms work** [Jin19]

METHODOLOGY

Several experiments are conducted and the averages of loss and reward over five runs are presented

The grid world is used instead of computer vision task to **reduce complexity**

Algorithm performance evaluation allows to build a **systematic approach** to agent hyperparameter choice *at least a given environment*

IMPLEMENTATION

1. Environment
2. Learned policy
3. Agent vision
4. Legend

EXPERIMENTS

Focus on different parameters:

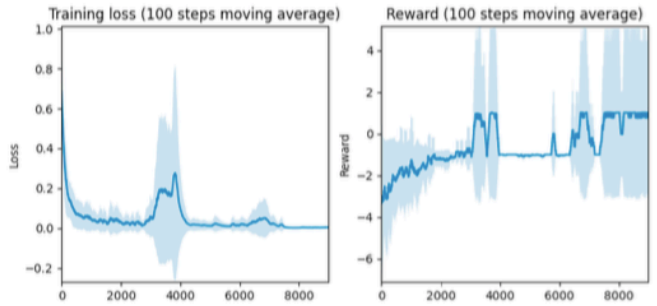
1. Memory sample size
2. Discount factor
3. Memory capacity
4. Agent vision range

EXPERIMENT 1 (MEMORY SAMPLE SIZE)

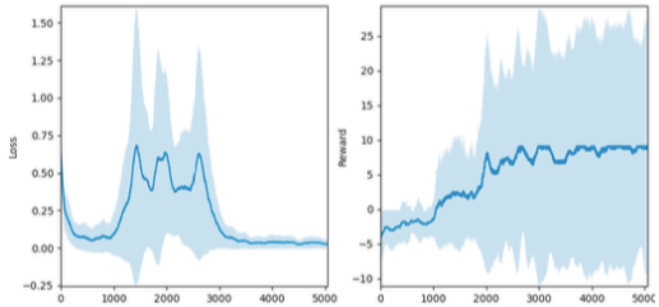
Visible change in learning performance

Impractical to pick arbitrarily large sample size because of computation costs

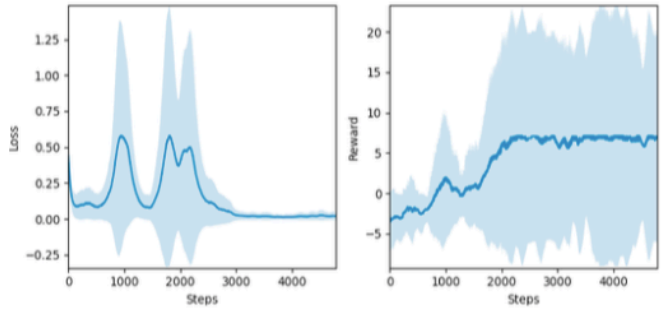
Not trivial to pick correct sample size – 70/30 from supervised learning does not work



a) Memory sample size: 32



b) Memory sample size: 64



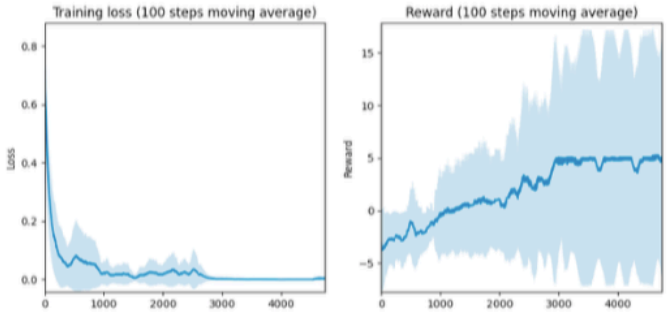
c) Memory sample size: 128

EXPERIMENT 2 (DISCOUNT RATE)

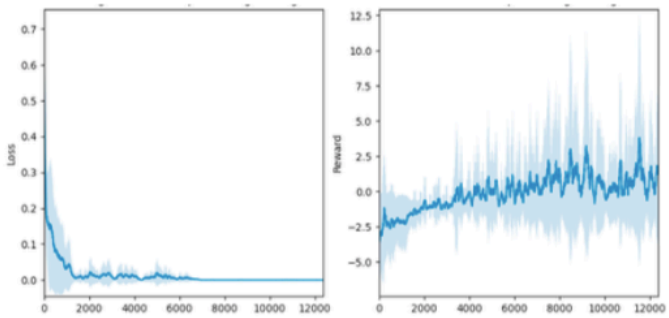
Optimal memory sample size is used as a **baseline**

Smaller discount rate – **harder to remember** what is "good"

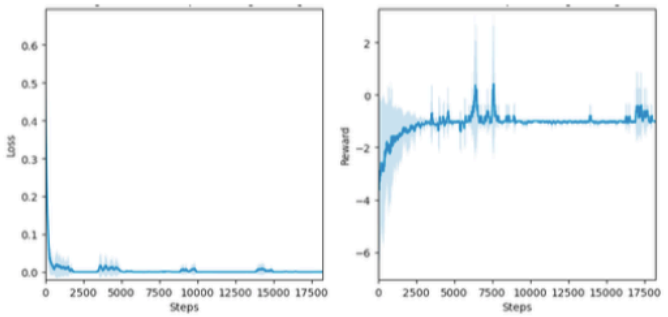
More noise in agend actions, however **faster convergence** of loss



a) Discount factor: 0.50

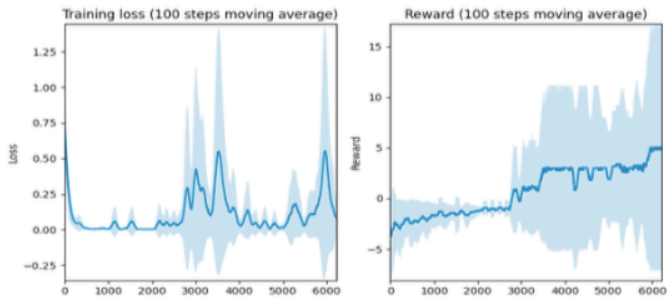


b) Discount factor: 0.25

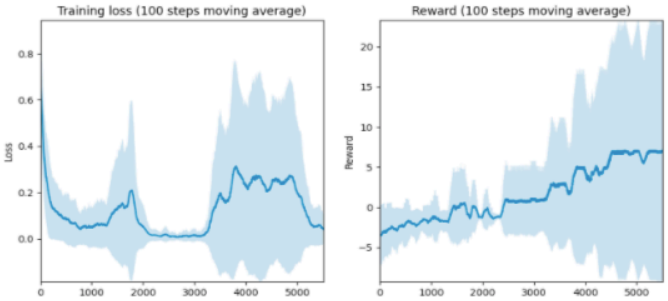


c) Discount factor: 0.10

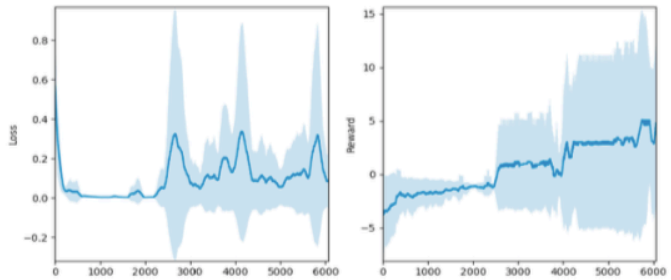
EXPERIMENT 3 (MEMORY CAPACITY)



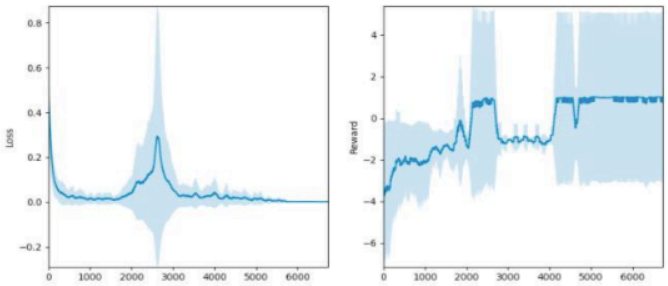
a) Memory capacity: 128



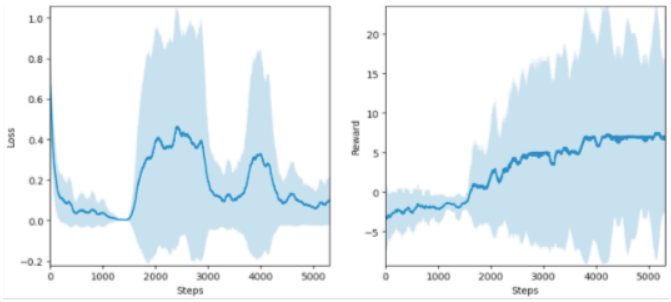
d) Memory capacity: 1024



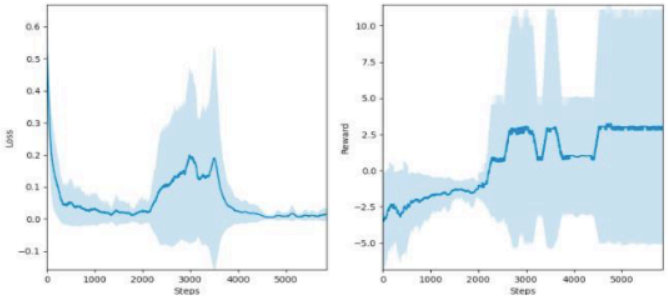
b) Memory capacity: 256



e) Memory capacity: 2048



c) Memory capacity: 512



f) Memory capacity: 4096

EXPERIMENT 3 (MEMORY CAPACITY)

More memory capacity – **longer history is remembered**, however agent might not notice any improvements that it has learned

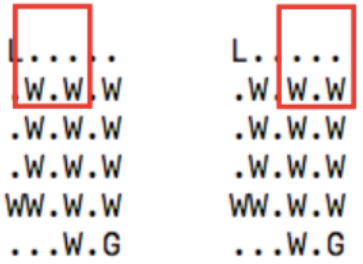
Learning performance is not in a linear relation to memory capacity!

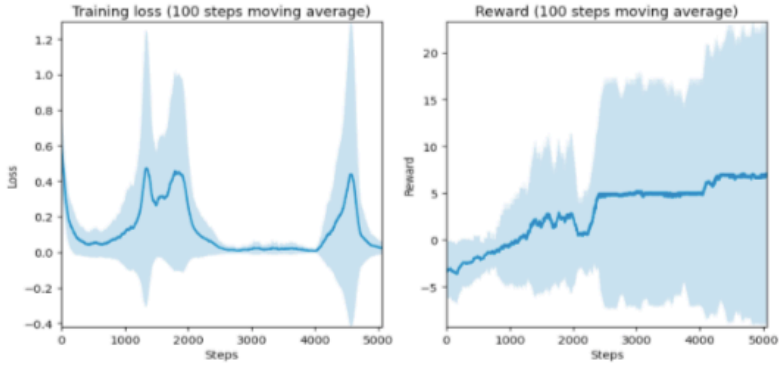
Run No.	*Memory size*	*Loss first convergence (in steps)*	*Average reward at convergence point*	*Final approx. average reward*
3a	128	300	-2.50	2.50
3b	256	500	-2.00	3.00
3c	512	1500	-2.00	6.00
3d	1024	**2500**	**0.00**	**7.00**
3e	2048	1500	-1.00	1.00
3f	4096	2000	-1.25	2.50

Table 1: Number of steps until model error first convergence and respective average reward

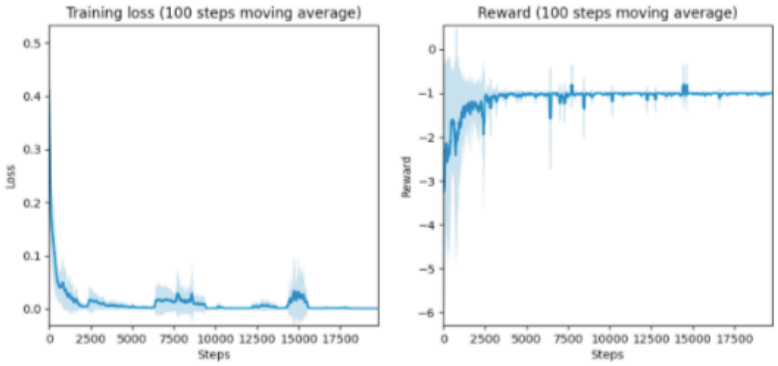
EXPERIMENT 4 (VISION RANGE)

Agent is unable to differentiate between different states in the environment
Relatively to the agent those states are *identical*





a) Vision range: 3 units



b) Vision range: 1 unit

RESULTS

Experiment No.	*Vision range*	*Discount factor*	*Memory size*	*Memory sample size*	*Final approx. average reward*
1a	2	0.99	1024	32	0.75
1b				**64**	**7.00**
1c				128	6.50
2a		**0.50**		64	**5.00**
2b		0.25			0.50
2c		0.10			-1.00
3a			128		2.50
3b			256		3.00
3c			512		6.00
3d			**1024**		**7.00**
3e			2048		1.00
3f			4096		2.50
4a	1		1024		-1.00
4b	**3**				**6.00**

SUMMARY

- Experiments have shown that it is **possible to find optimal** parameter values by hand by picking an arbitrary value first
- Memory capacity and sample size change had the **most significant** effect and non-linear relation
- Application of training set processing from supervised learning (even distribution to **avoid bias** – balancing training data set) might improve learning performance

OUTLOOK

- This opens a new direction in developing **new adaptive methods** (similar to Adagrad [Duc11]) for memory parameters adaptation
- Will probably require building **new data structure** (currently deque-like structures are used)
- Could general **adaptation framework** be the way to general artificial intelligence?

THANK YOU

FOR YOUR ATTENTION

REFERENCES

[Doy02] Doya K.; Samejima K.; Katagiri K. et al. (2002): Multiple model-based reinforcement learning. Neural Computation, 14, 6, 1347-1369.

[Duc11] Duchi J.; Hazan E. und Singer Y. (2011): Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. In: Journal of Machine Learning Research 12.. S. 2121-2159.

[Eva17] Evan S.; Mahmoudieh P.; Argus M. et al. (2017): Loss is its own Reward: Self-Supervision for Reinforcement Learning. arXiv preprint arXiv:1612.07307.

[Foe16] Foerster J.N.; Assael Y.M.; de Freitas N. et al. (2016): Learning to Communicate withDeep Multi-Agent Reinforcement Learning. In: 30th Conference on Neural Information Processing Systems (NIPS 2016). Barcelona, Spain.

[Fuj19] Fujimoto S.; Meger D. und Precup D. (2019): Off-Policy Deep Reinforcement Learning without Exploration. In: Proceedings of the 36th International Conference on Machine Learning. California.

[Jin19] Jin C. (2019): Machine Learning: Why Do Simple Algorithms Work So Well? (PhD Thesis) EECS Department, University of California, Berkeley.

[Sut18] Sutton R.S. und Barto A.G. (2018): Reinforcement Learning: An Introduction. Cambridge: MIT Press.

[YNg03] Y. Ng A. (2003): Shaping and Policy Search in Reinforcement Learning. (PhD Thesis) University of California, Berkeley.