Learning rollout policy for internal planning in deep reinforcement learning agent

Metody planowania w głębokim uczeniu ze wzmocnieniem

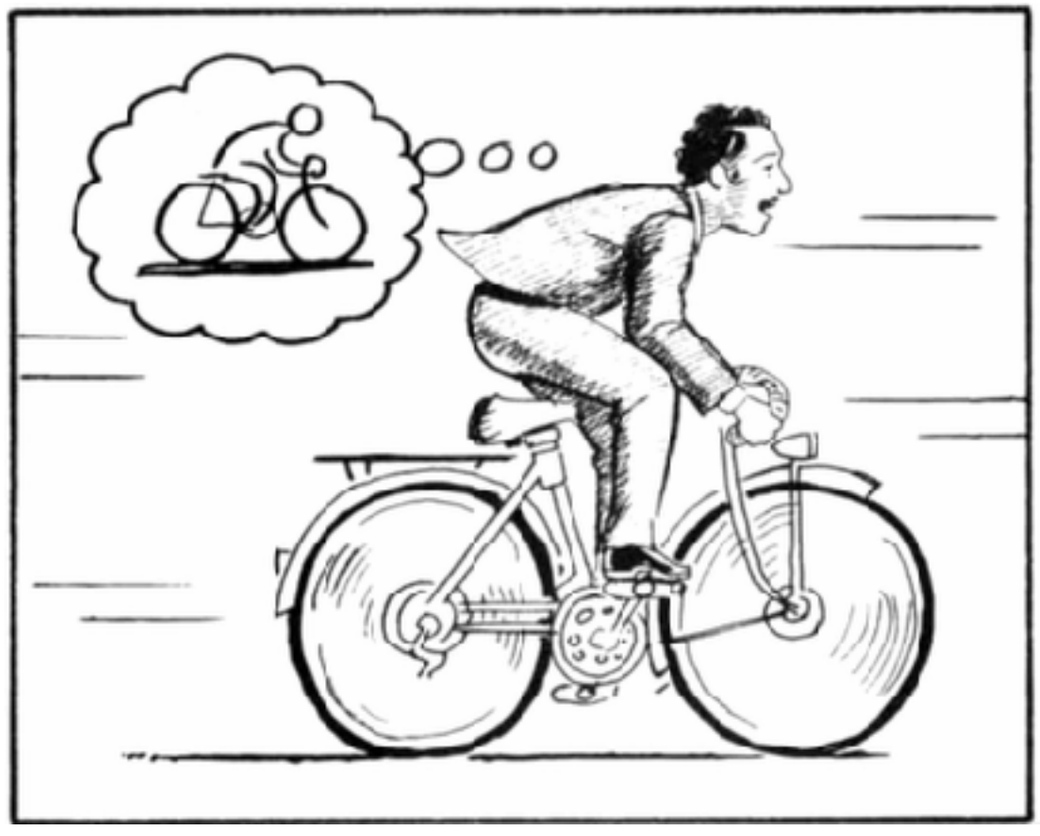
Piotr Januszewski

Motivation

Mental model

The image of the world around us, which we carry in our head, is just a model. Nobody in his head imagines all the world, government or country. He has only selected concepts, and relationships between them, and uses those to represent the real system

~ Jay Wright Forrester, 1971



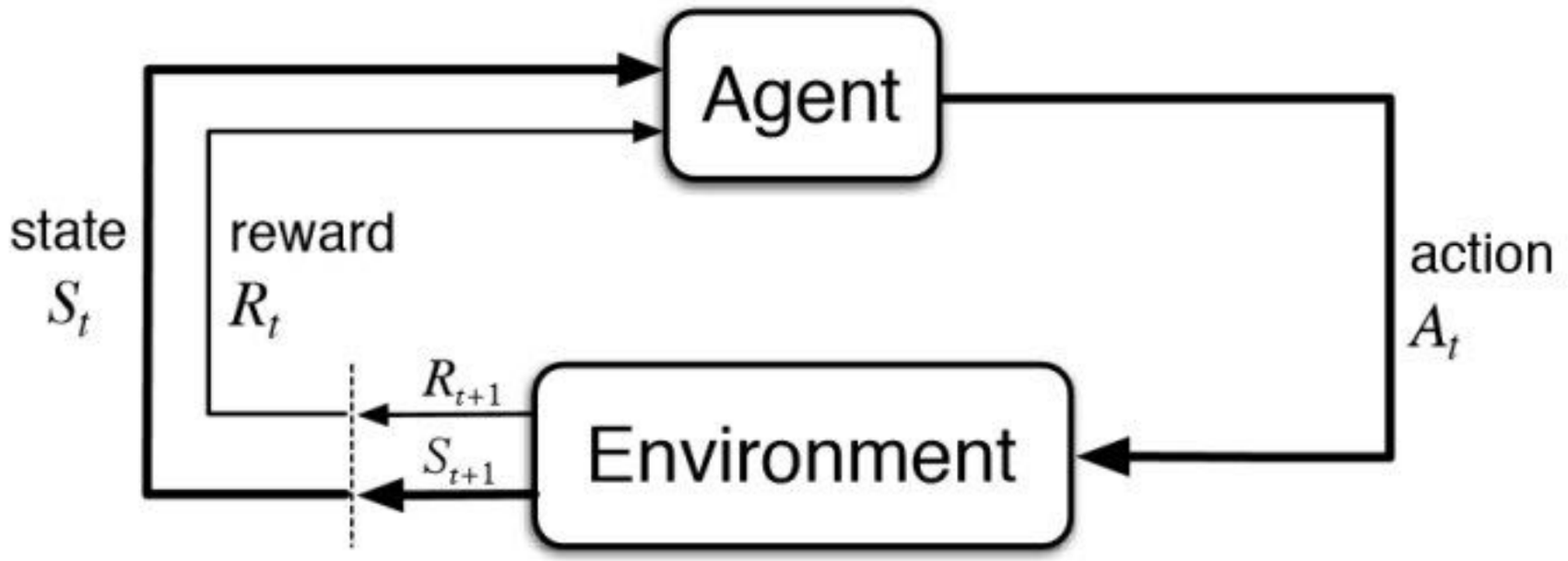
A World Model, from Scott McCloud's Understanding Comics.

Goal

The aim of this work is to derive from previous work on model learning in complex high-dimensional decision making problems and apply them to planning in complex tasks.
Those methods proved to train accurate models, at least in short horizon, and should open a path for application of simulation-based search algorithms with the learned world model to i.e. Atari 2600 games, a platform used for evaluation of general competency in artificial intelligence.

The goal is to improve data efficiency without loss in performance compared to model-free methods. This work focuses on three benchmarks: an arcade game with dense rewards Boxing, a challenging environment with sparse rewards Freeway and a complex puzzle game Sokoban.

Reinforcement Learning



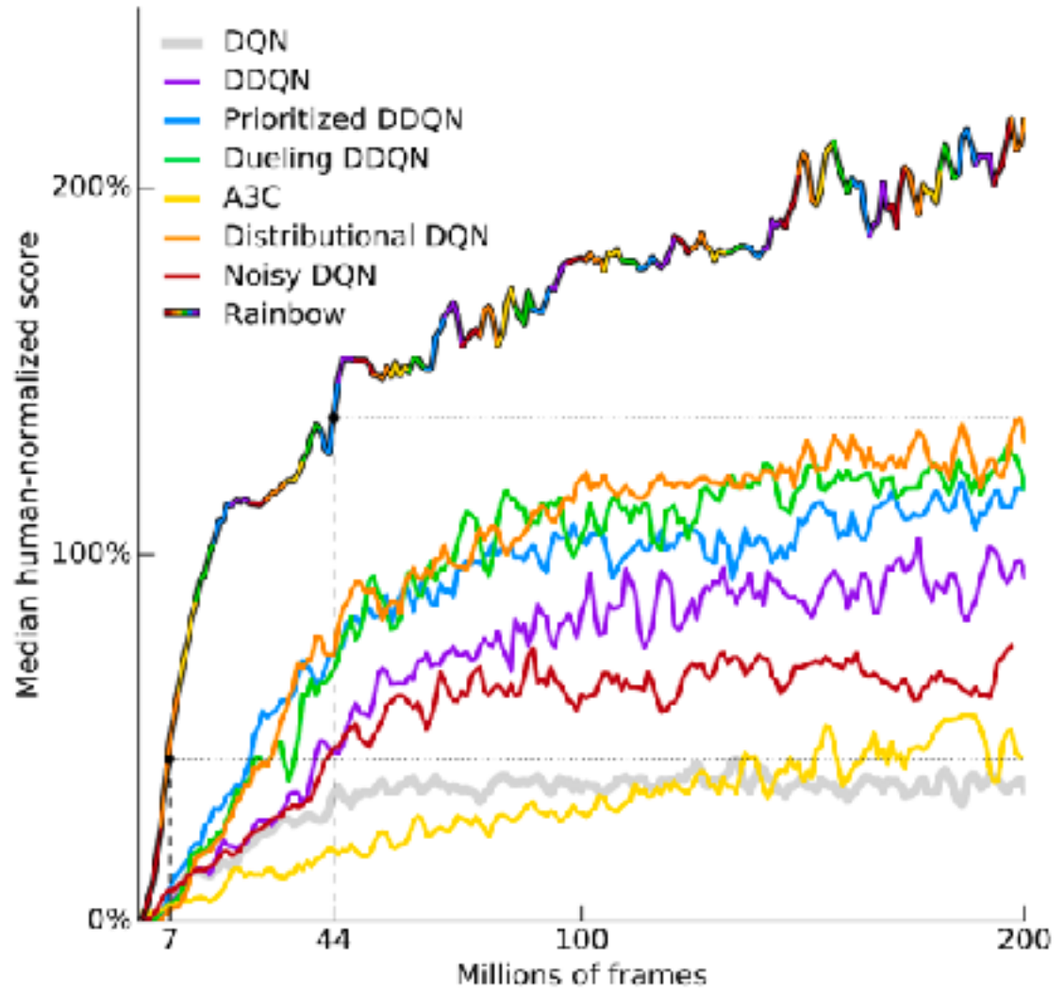
Reinforcement learning (RL) is learning what to do, how to map situations to actions, so as to maximise an expected return ~ Reinforcement Learning: An Introduction [2nd edition]

State of the art

Model-free:

- Rainbow - a compilation of several independent improvements to the DQN algorithm made by the deep reinforcement learning community.

- PPO - the new family of policy gradient methods for reinforcement learning, which alternate between sampling data through interaction with the environment, and optimising a surrogate objective function using stochastic gradient ascent.



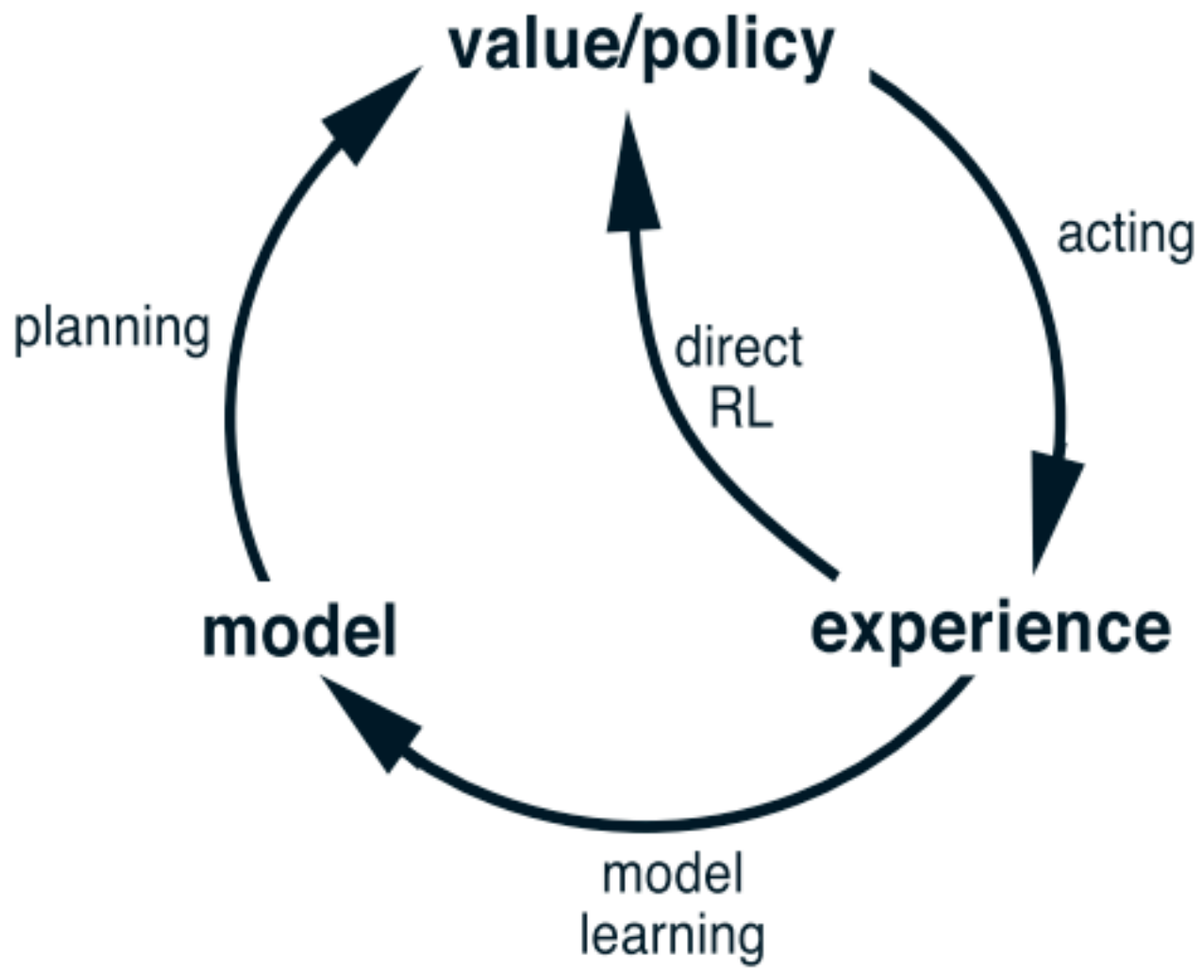
State of the art

Model-based:

- World Models - a simple model inspired by human cognition.

- PlaNet - a purely model-based agent that learns the environment dynamics from images and chooses actions through fast online planning in latent space.

- SimPLe - the authors explore how video prediction models can enable RL agents to solve Atari games with orders of magnitude fewer interactions than model-free methods.



Planning and learning

The experience can improve value functions and policies either directly or indirectly via the transition model.

Latent state-space models

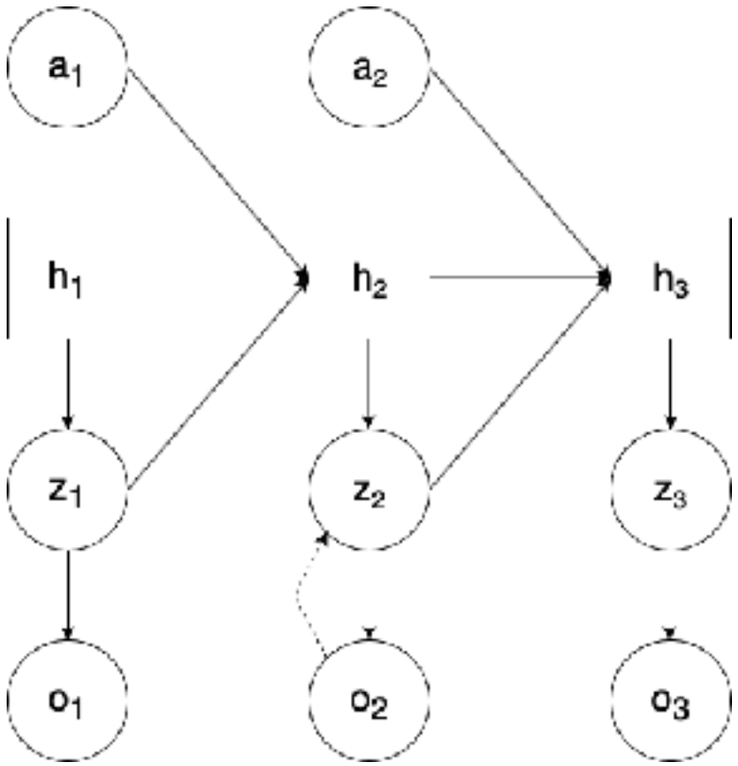
A key challenge in model-based reinforcement learning is learning accurate, computationally efficient models of complex domains and using them to solve RL problems. Recent solution to this problem are computationally efficient state-space environment models that make predictions at a higher level of abstraction, both spatially and temporally, than at the level of raw pixel observations. Such models substantially reduce the amount of computation required to make predictions, as future states can be represented much more compactly. Moreover, in order to increase model accuracy, they exploit the benefits of explicitly modeling uncertainty in state transitions.

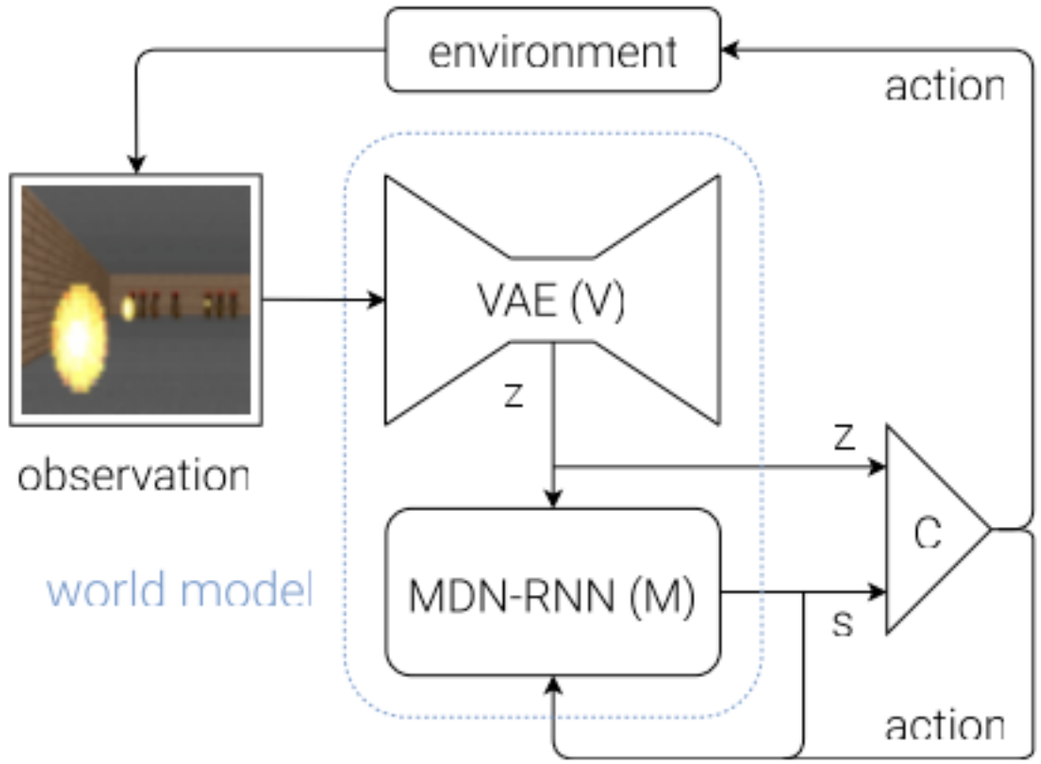
Simulation-based search

- Planning carries the promise of increasing performance just by increasing the computational budget for searching for actions.

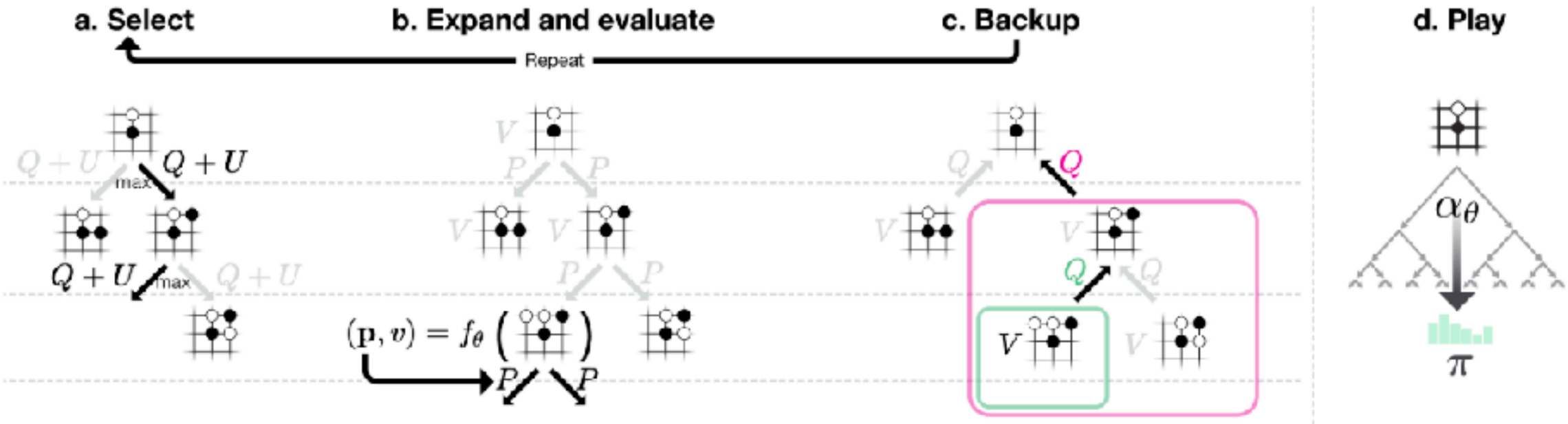
- Search exploits temporality, in other words focuses on the current situation, by learning from this specific distribution of future experience, rather than learning from the distribution of all possible experience.

World Models

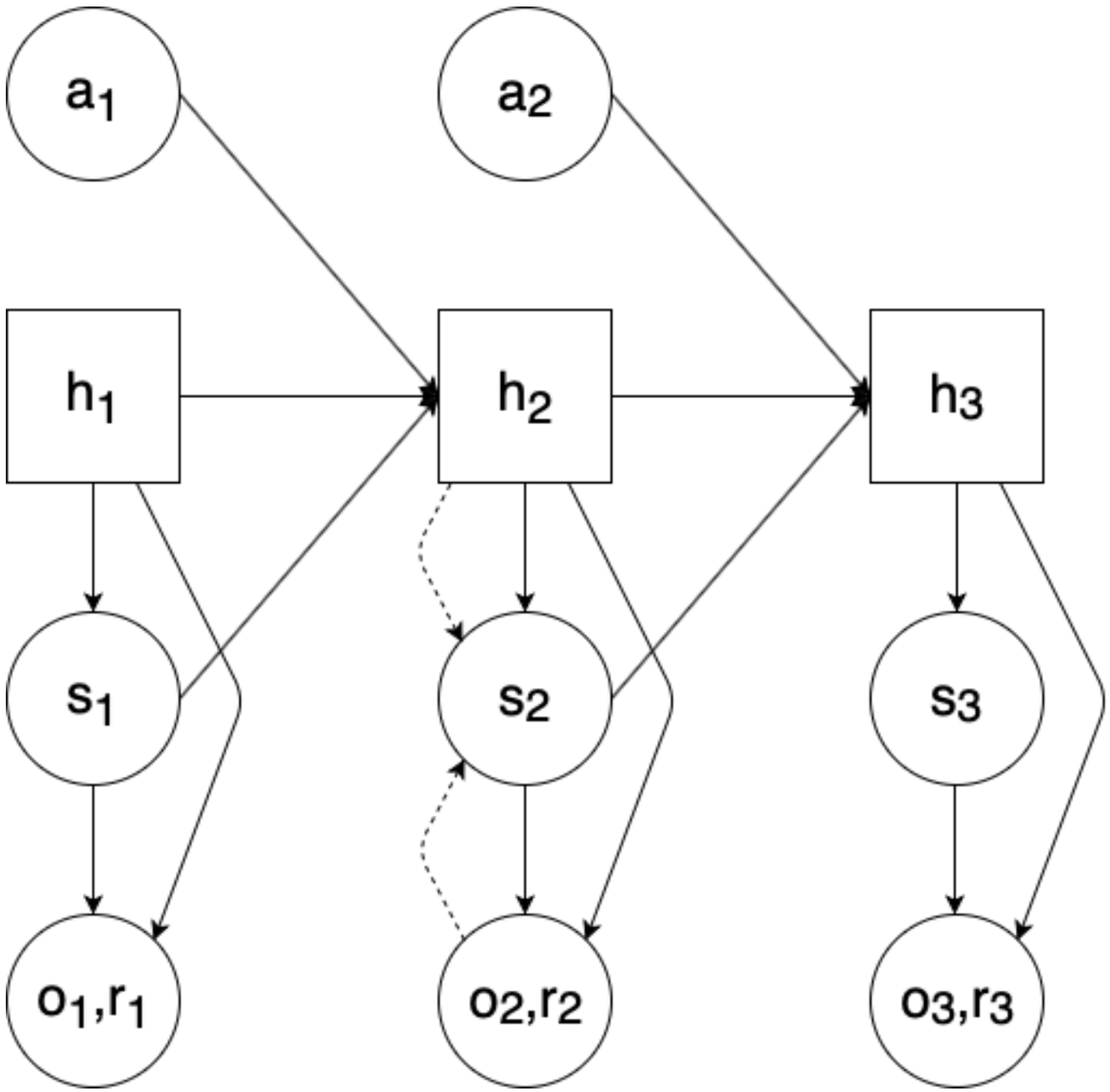




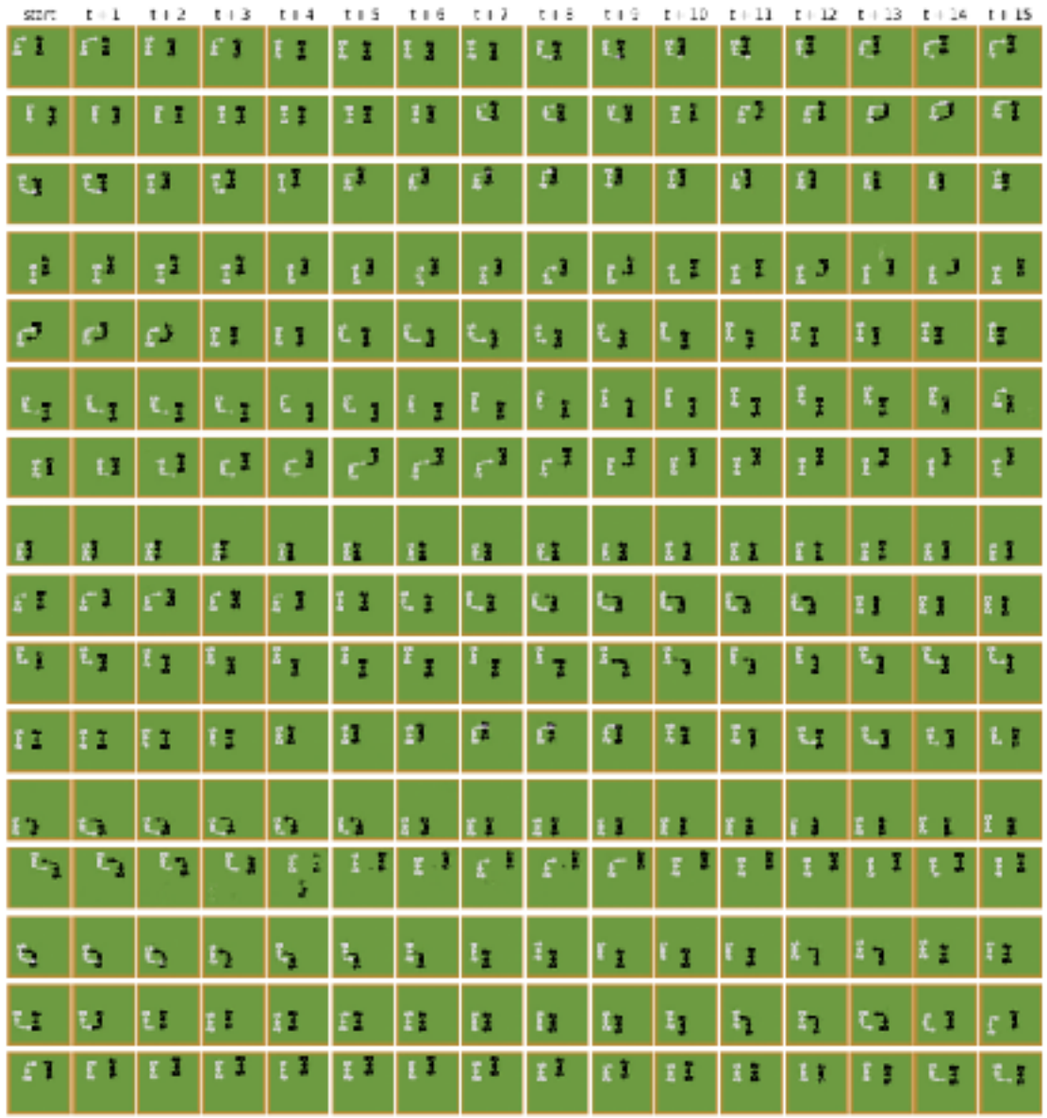
AlphaZero


a. Select
b. Expand and evaluate
c. Backup
d. Play
Repeat
$Q + U$
max
V
P
Q
$(\mathbf{p}, v) = f_\theta\left(\right)$
α_θ
π

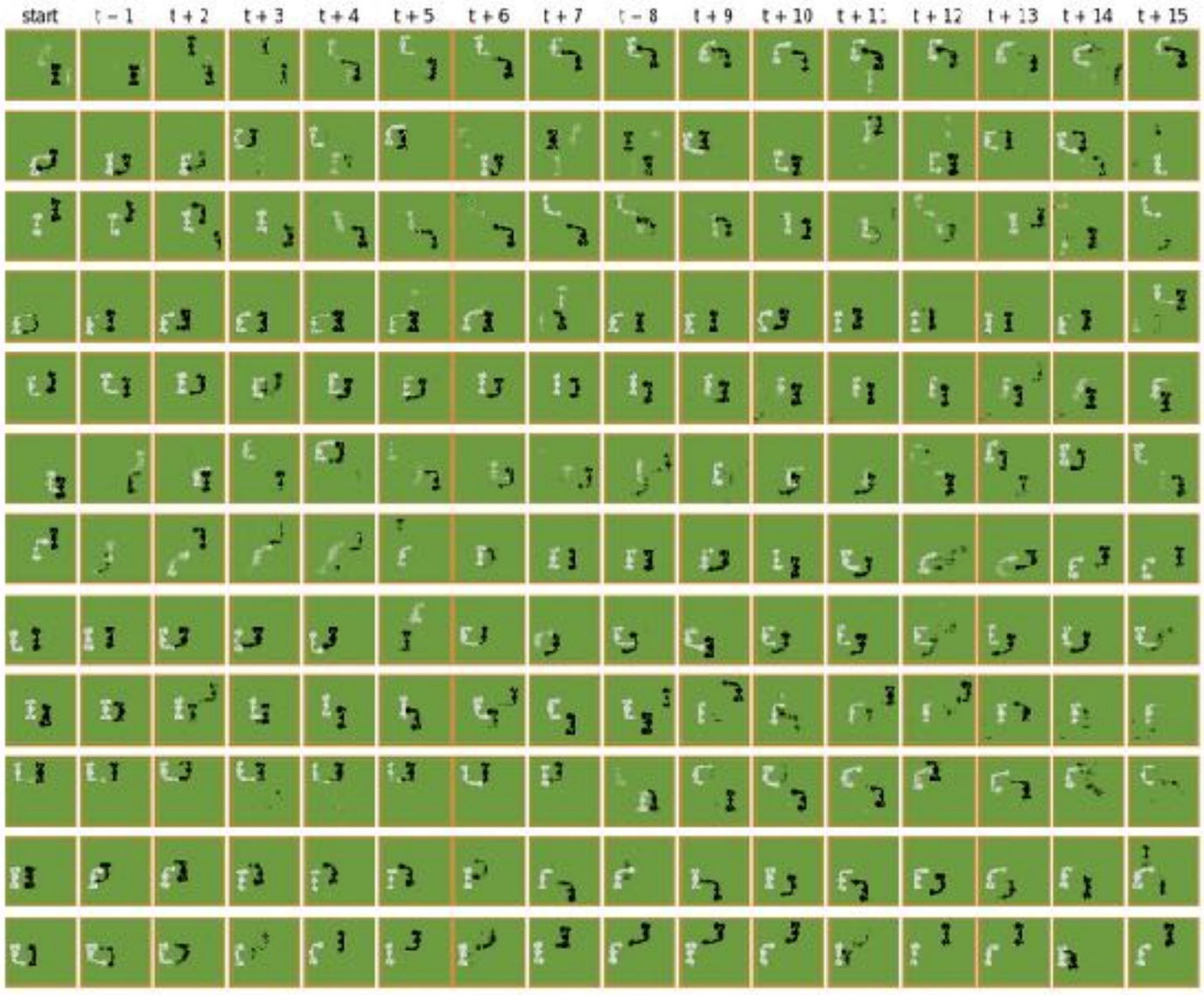
Discrete PlaNet



Experiments:
Train OWM in the Boxing environment

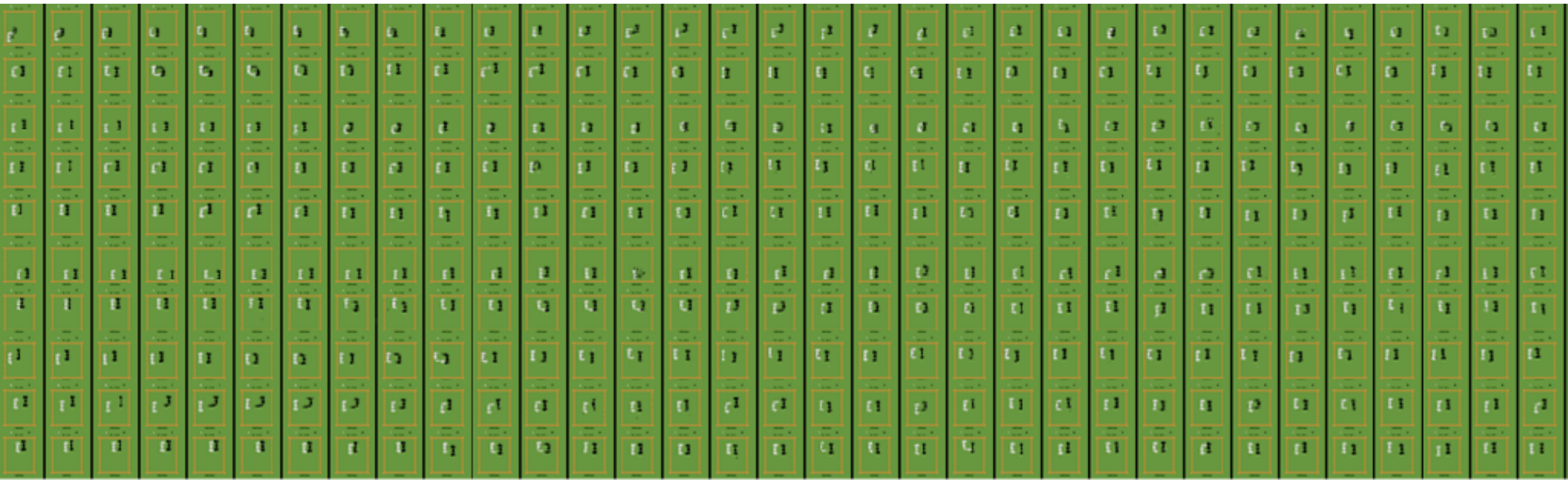


Experiments:
Train W+A in the Boxing environment



Experiments:

Train DPN in the Boxing environment



Conclusion

- Final solution in form of the „Discrete PlaNet” architecture is presented. Despite many difficulties, it finally reached a level of performance equal or higher than strong model-free and model-base baselines in low data regime of up to 1M interactions with the real environment of Atari 2600 game Boxing.

- The most challenging part of this work, underestimated at first by the author, was model learning. Current state-of-the-art model learning methods, although report promising results, were not tested for planning with them using search based algorithms, let alone planning and learning. Towards the end of the experimentation phase for this thesis, the PlaNet paper came out. It become the keystone for the final solution: the DPN architecture.

- Neither architecture was able to learn playing Sokoban. Certainly, the problem lies in model learning techniques. Sokoban dynamics, although based on simple rules of moving a character and pushing boxes, allow for incredible number of possible states and levels configurations. The models were not able to generalize well to this number of possibilities.

- The AlphaZero algorithm is very promising in a sense, that it is the general reinforcement learning algorithm which proved to solve really complex problems, like playing the game of Go \cite{Algo.AlphaGoZero}, when supplied with the perfect environment dynamics model. It should easily manage Sokoban complexity. And yet, joining it with the imperfect learned world model resulted in unstable, and at the end unsuccessful, training.

- Likewise, for the DPN architecture, sparse rewards of Freeway became an obstacle which could not be overcome.

Future work

- Future work could focus on extending this method to other challenging tasks like: sparse rewards environments, i.e. Freeway, and complex puzzle games with massive state-space sizes, i.e. Sokoban.

- DPN, unlike model-based baseline SimPLe, hold promise of increased performance with an increased computational budget for planning. This hypothesis could be put to the extensive test too.

- Furthermore, generalisation of the DPN world model to different tasks in the same or very similar environments could be explored.