World Models[1]

[1] Ha, D., & Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*. ISO 690

Introduction

Why?

- Exciting research in the RL domain
  - Exploration of models that can model spatial and temporal abstractions
  - Ideas introduced may be inspiring for our research

- Computationally the brain probably uses mechanisms similar to RL [2]

[2] Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139-154.

Presenter: Eloy Geenjaar

Reinforcement Learning

A quick recap/introduction

Some of the content on these slides has been adapted from:

- Sergey Plis' CS8550 class on Advanced Machine Learning
- Hado van Hasselt's Introduction to Reinforcement Learning https://hadovanhasselt.files.wordpress.com/2016/01/intro_rl_20161.pdf

What is RL?

- Humans and other intelligent beings learn by interacting
- RL: learning to make decisions from interactions with the environment.
- E.g. Babies learn about gravity and causality by letting blocks fall on the ground, and stacking them on top of each other

LeCun's cake analogy

- Reinforcement Learning:
  - The cherry
  - Reward: only a few bits per sample
- Supervised learning:
  - The icing
  - 10-10k bits per sample
  - E.g. class label
- Self-supervised/unsupervised Learning:
  - The inside of the cake
  - 1-100M bits per sample

Reward signal

- The reward signal may be delayed
  - E.g.: Multiple ways to pick up a cup

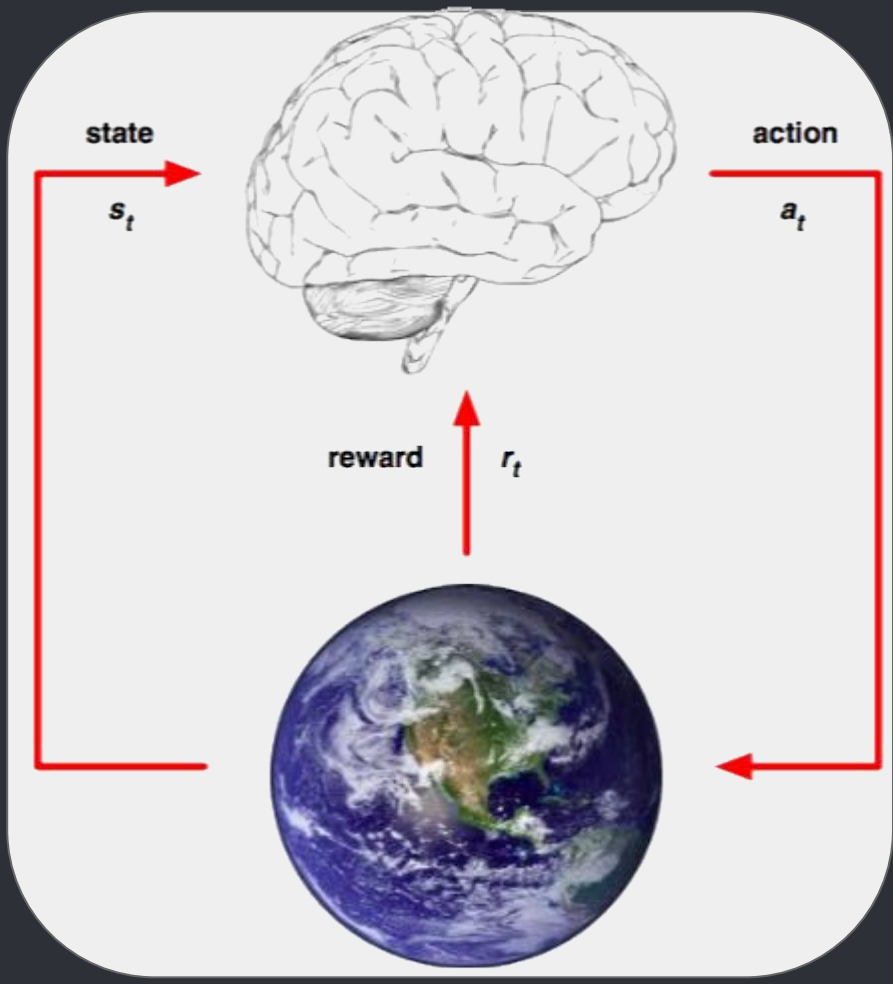
- Sequential, temporality is thus important
  - Earlier actions affect **decisions** later

Elements

- The environment
  - State (s_t)
  - Action (a_t)
  - Reward (r_t)

- The agent
  - State (s_t)
  - Action (a_t)

- Observation of environment's state (o_t)
- Maximize cumulative reward by selecting actions



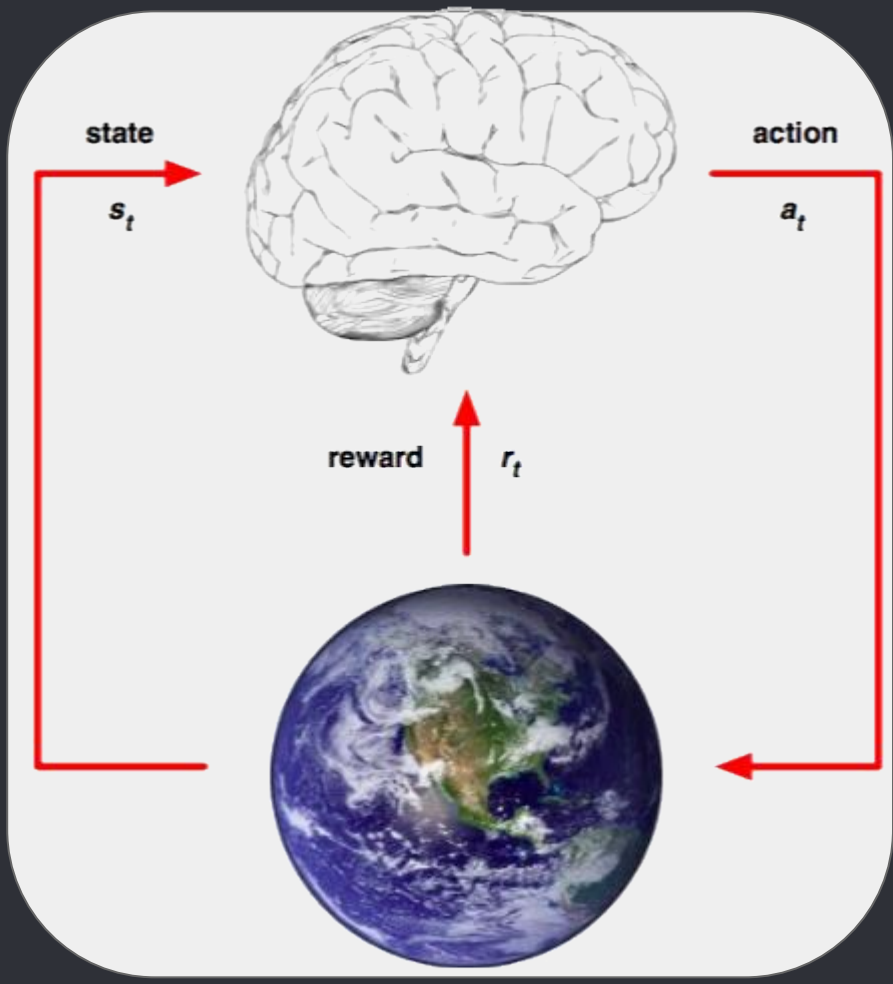
Agent

- State (s_t)
- Action (a_t)

- Policy (π)
  - Maps state to actions

- Value function (v_π)
  - Expected future reward)

- Model (m)



Q-Learning

- Maximum attainable reward that can be reached in the future, given the next action is taken into account when deciding an action

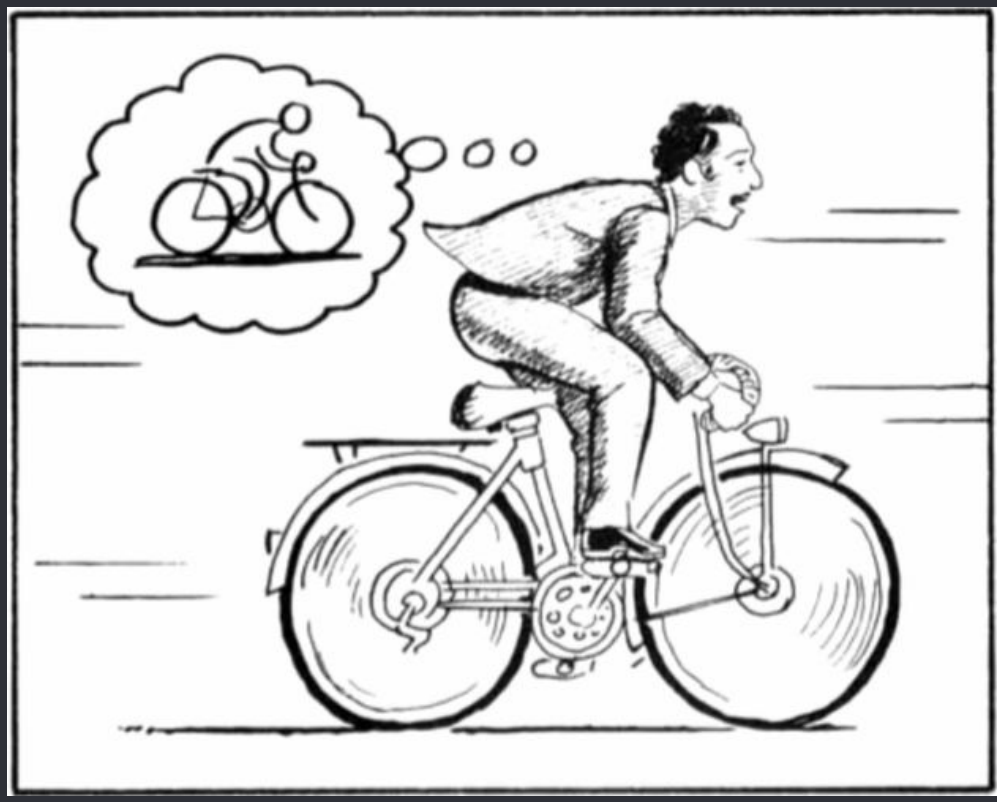
- Deep Q-Learning: Use a DNN (probably CNN)
  - Unstable: Observations are highly correlated
  - Unstable: Actions have a large impact on future observations
  - E.g.: Emotionally deciding something
  - Experience replay: random sample of previous actions

World models

Here we go

A world model

- Compressed spatial and temporal representation of the environment



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

Brain predictions

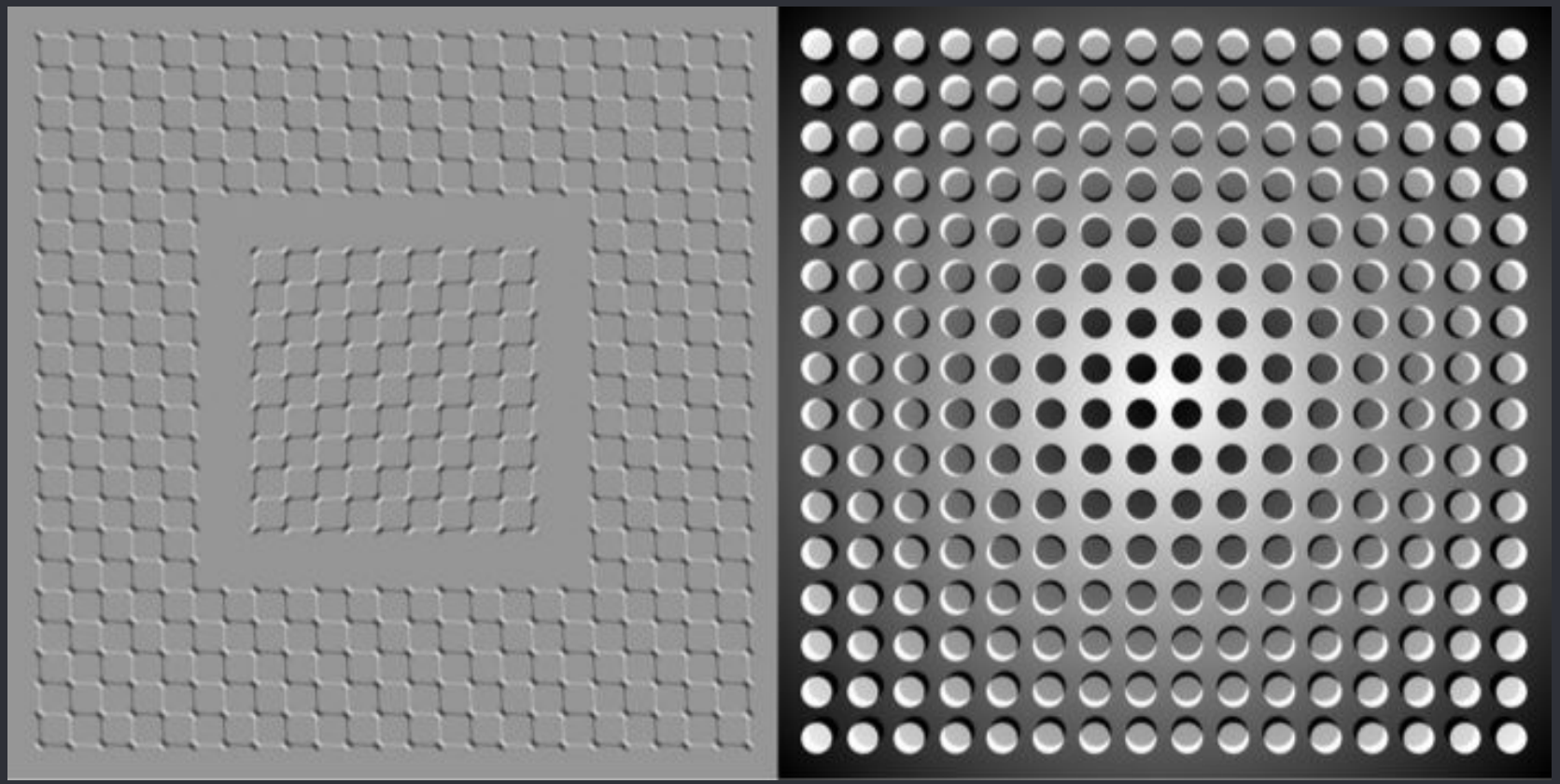


Figure 2. What we see is based on our brain's prediction of the future (Kitaoka, 2002; Watanabe et al., 2018).

Credit assignment problem

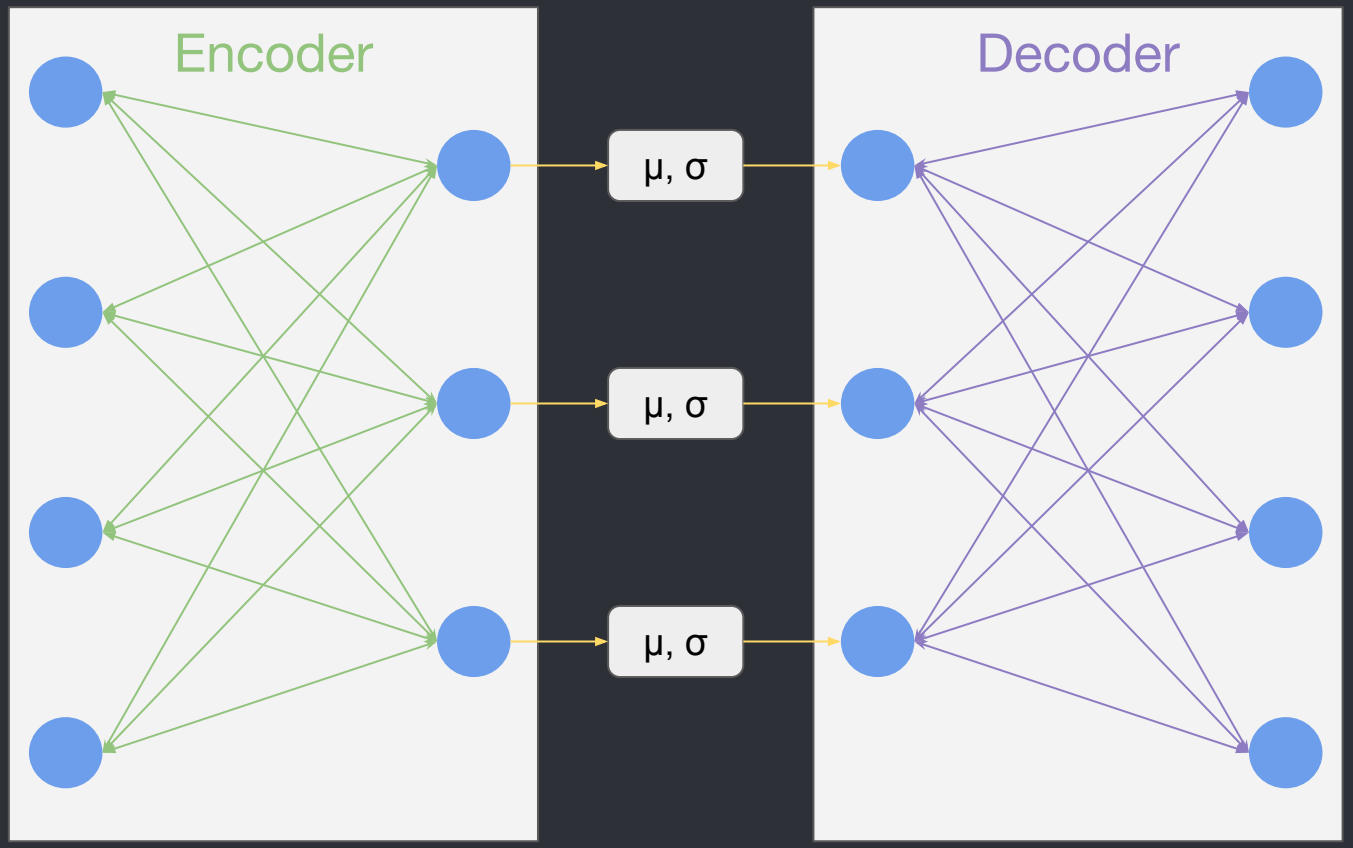
- RL algorithms are bottlenecked by the credit assignment problem
- Which value is assigned to each artificial unit in a neural network

Solution:

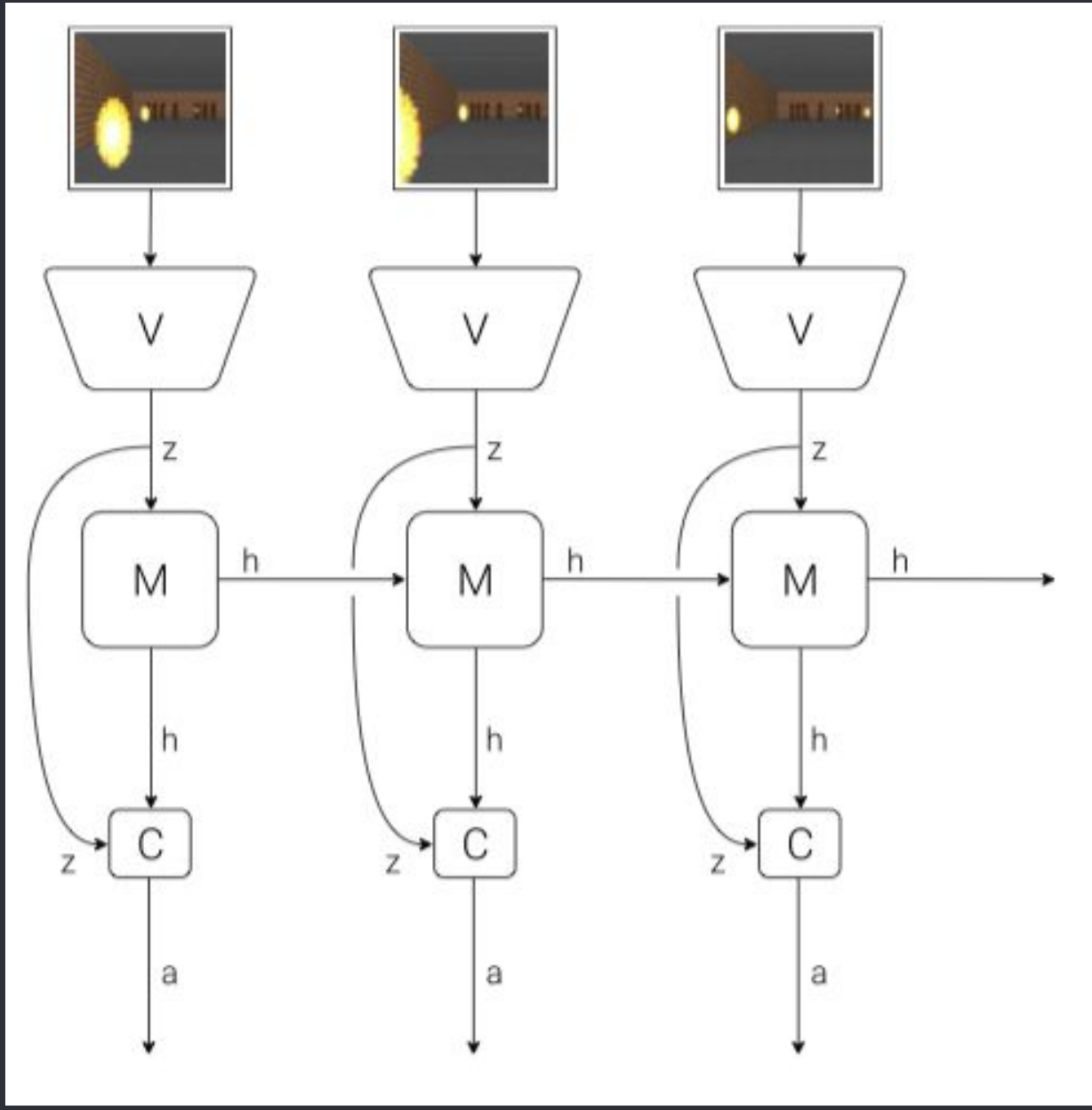
1. Train large neural network to learn model of the world
2. Train smaller controller model that uses embedding space of the larger neural network to select next action

Vision model: VAE

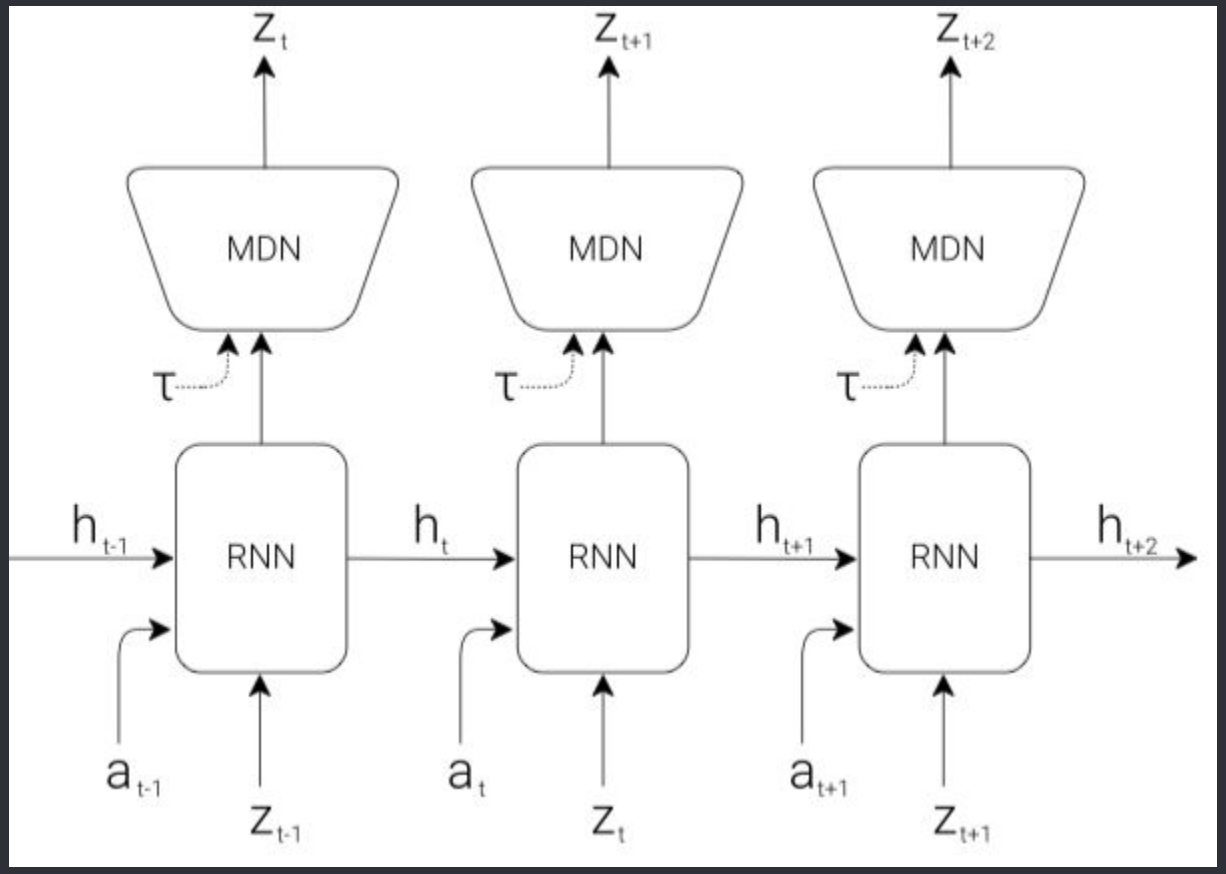
Input layer | Variational autoencoder | Output layer



Full model



Memory model

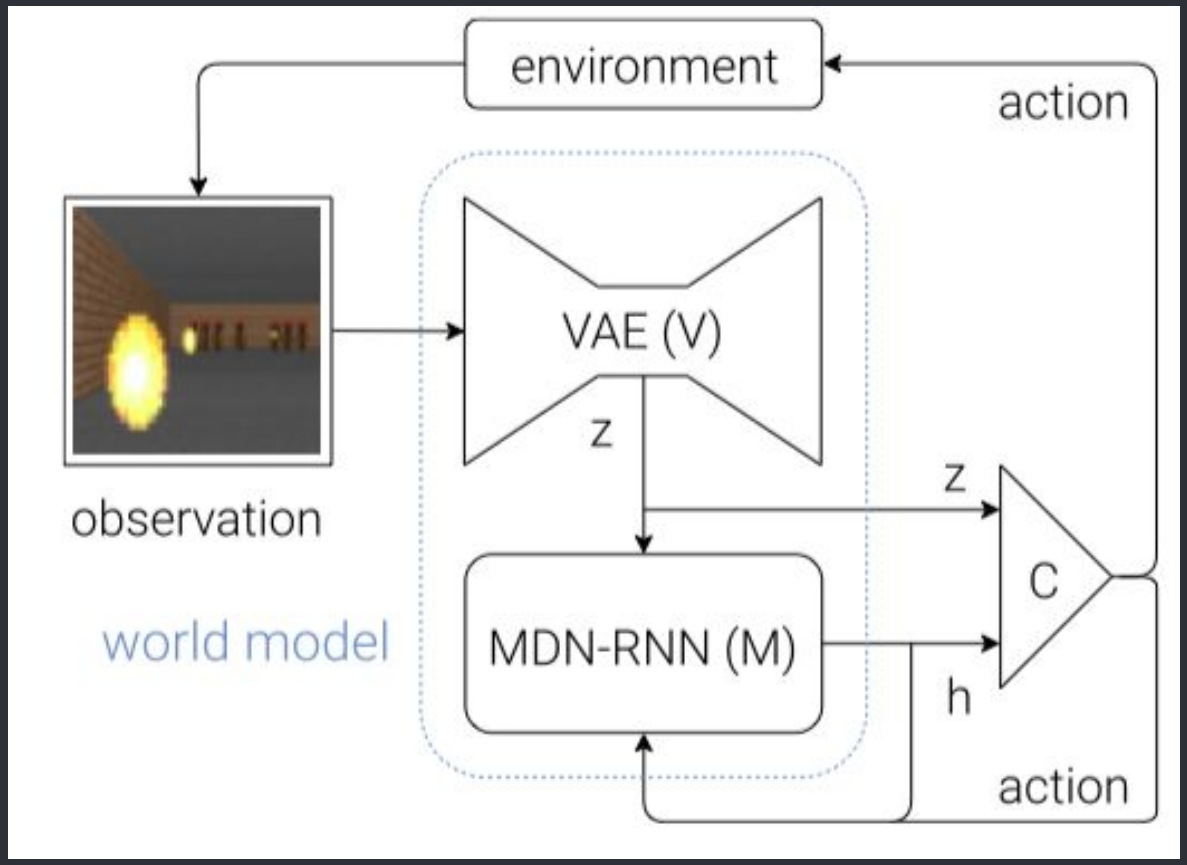


Controller model

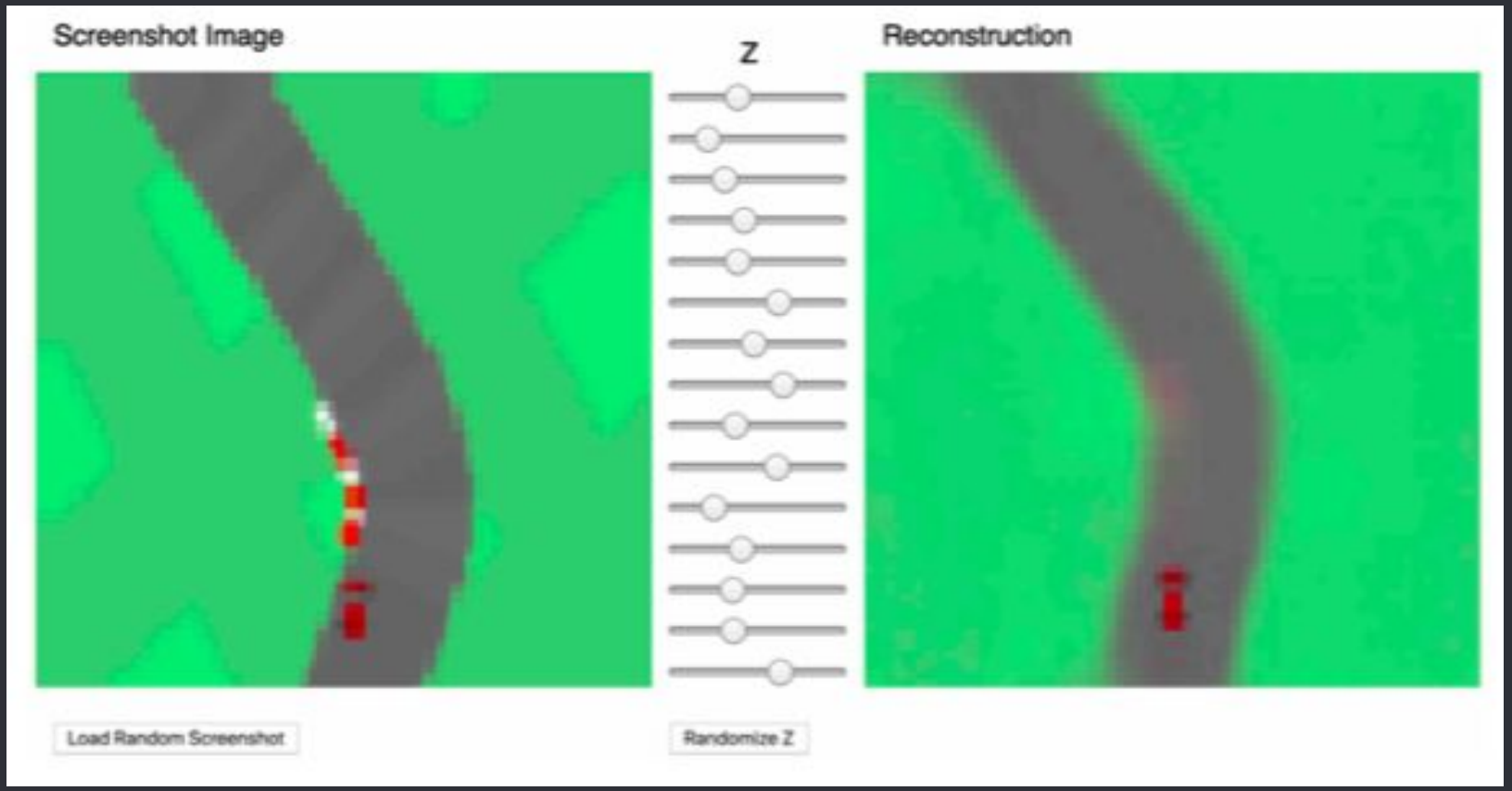
$$a = W_c * [z_t\ h_t] + b_c$$

- Trained using a genetical algorithm

Full model: overview



Example of reconstruction using VAE



Results: CarRacing

METHOD	AVG. SCORE
DQN (PRIEUR, 2017)	343 ± 18
A3C (CONTINUOUS) (JANG ET AL., 2017)	591 ± 45
A3C (DISCRETE) (KHAN & ELIBOL, 2016)	652 ± 10
CEOBILLIONAIRE (GYM LEADERBOARD)	838 ± 11
V MODEL	632 ± 251
V MODEL WITH HIDDEN LAYER	788 ± 141
FULL WORLD MODEL	$\mathbf{906 \pm 21}$

Driving inside a dream

- The VAE learns a low dimensional space
  - The RNN-MDN learns to predict the next location in that space
- The RNN-MDN can thus be used to drive around in that space, without needing any observations

Learning inside a dream: VizDoom

Learning inside a dream: continued

- Temperature for the RNN-MDN can be increased to make the environment harder (more uncertain)
- Fireballs may behave in a less predictable path compared to the game
- Agents that perform well in higher temperature settings generally perform better in the normal setting

Learning inside a dream: adversarials

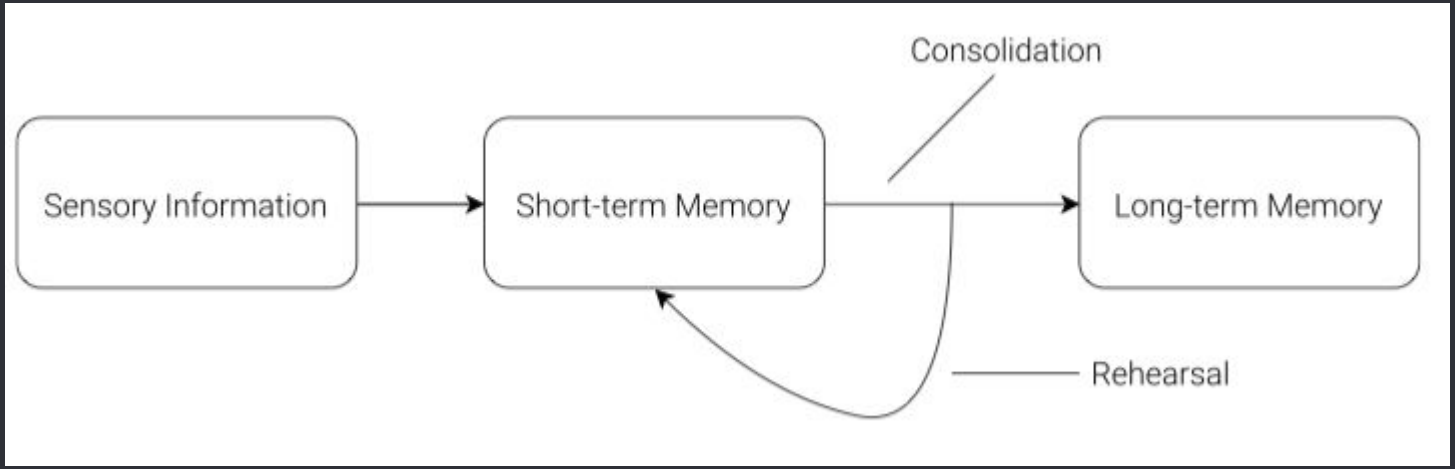
- The agent discovered an adversarial policy to move around in such a way that the monsters in this virtual environment never shoots a single fireball.
- The controller model has access to the internal state of the memory model and thus to all of the internal states and memory of the simulated game engine, instead of just the observations

Learning inside a dream: results

- SOTA at the time of publication: 820 +- 58

Temperature τ	Virtual Score	Actual Score
0.10	2086 ± 140	193 ± 58
0.50	2060 ± 277	196 ± 50
1.00	1145 ± 690	868 ± 511
1.15	918 ± 546	1092 ± 556
1.30	732 ± 269	753 ± 139
Random Policy	N/A	210 ± 108
Gym Leader	N/A	820 ± 58

Connection to neuroscience



- Hippocampal replay: replays memories when an animal rests or sleeps

Future work

- More complex models: intrinsic motivation
- Complex motor skills could also be learned (imitated) by the world model, the controller model can rely on those motor skills instead of learning them from scratch.

Thanks!

ANY QUESTIONS?

Let’s keep discussing the ideas and looking for ways to learn them deeper by applying them in unexpected ways