World Models

Ondřej Podsztavek[1]

[1]Houdkovice, Czechia

April 18, 2018

World Models

David Ha [1] **Jürgen Schmidhuber** [2 3]

Abstract

We explore building generative neural network models of popular reinforcement learning environments. Our *world model* can be trained quickly in an unsupervised manner to learn a compressed spatial and temporal representation of the environment. By using features extracted from the world model as inputs to an agent, we can train a very compact and simple policy that can solve the required task. We can even train our agent entirely inside of its own hallucinated dream generated by its world model, and transfer this policy back into the actual environment.

An interactive version of this paper is available at https://worldmodels.github.io

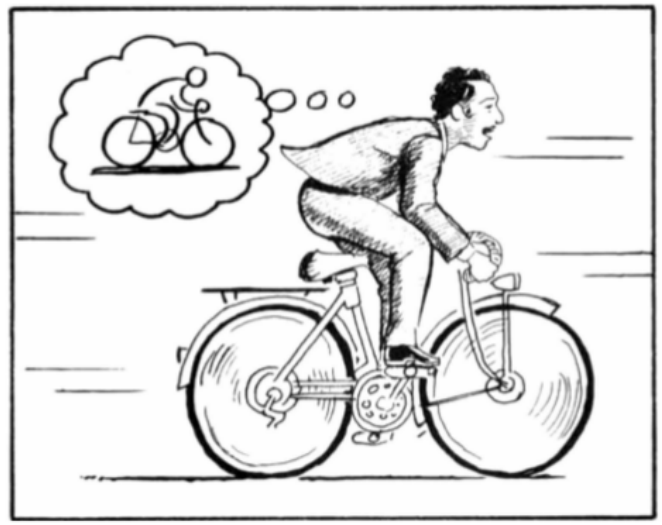


Figure 1. A World Model, from Scott McCloud's *Understanding Comics.* (McCloud, 1993; E, 2012)

Figure: arXiv version of World Models paper.[1]

[1] D. Ha and J. Schmidhuber. "World Models". In: (2018). DOI: 10.5281/zenodo.1207631. eprint: arXiv:1803.10122. URL: https://worldmodels.github.io.

Outline

Outline

Mental Model

"The mental images in one's head about one's surroundings are models. (...) One uses selected concepts and relationships to represent real systems."[2]

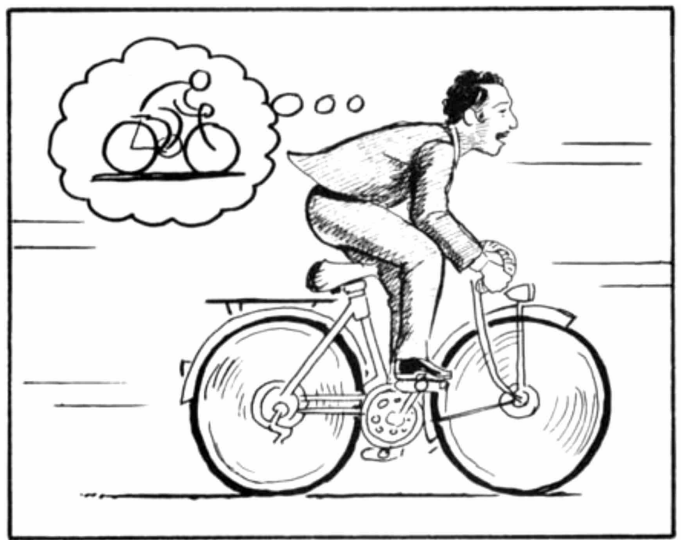


Figure: A World Model from Scott McCloud's *Understanding Comics*.

[2] Jay W. Forrester. "Counterintuitive behavior of social systems". In: *Technological Forecasting and Social Change* 3 (1971), pp. 1–22. ISSN: 0040-1625. DOI: 10.1016/S0040-1625(71)80001-X.

Predictive Model in Our Brain

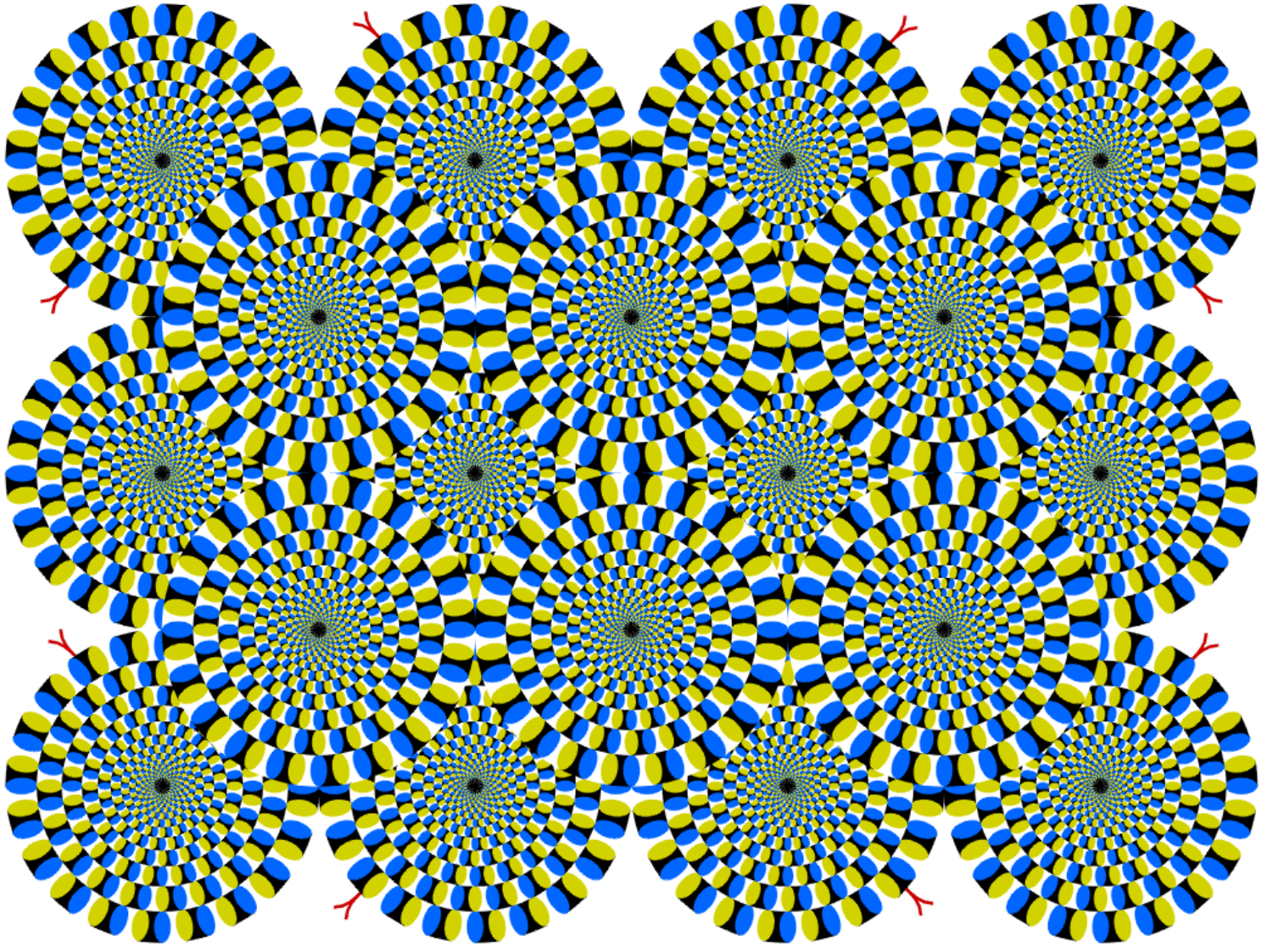


Figure: What we see is based on our brain's future prediction.

World Models in Reinforcement Learning (RL)

- A RL algorithm is bottlenecked by the credit assignment problem.
- A large world model and a small controller model.
- Small search space lets the algorithm focus on the credit assignment.

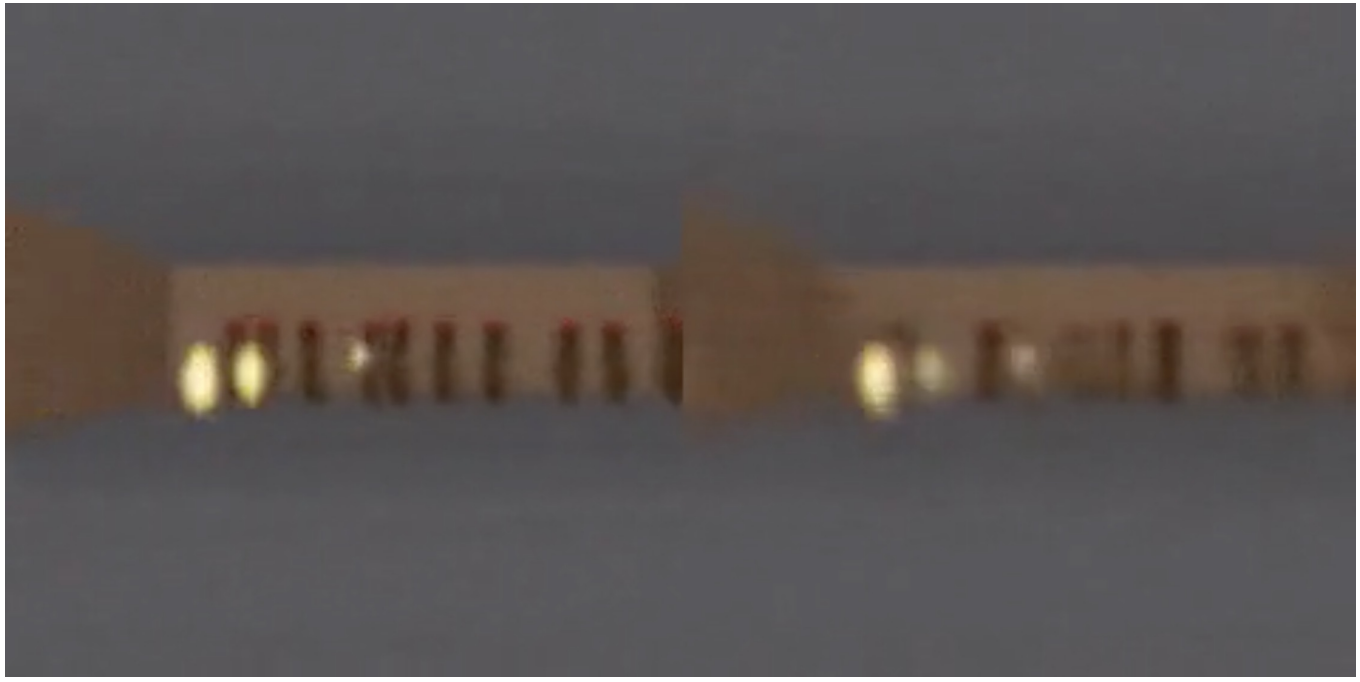


Figure: Despite losing details during this lossy compression process, latent vector captures the essence of each image frame.

Outline

Agent Model

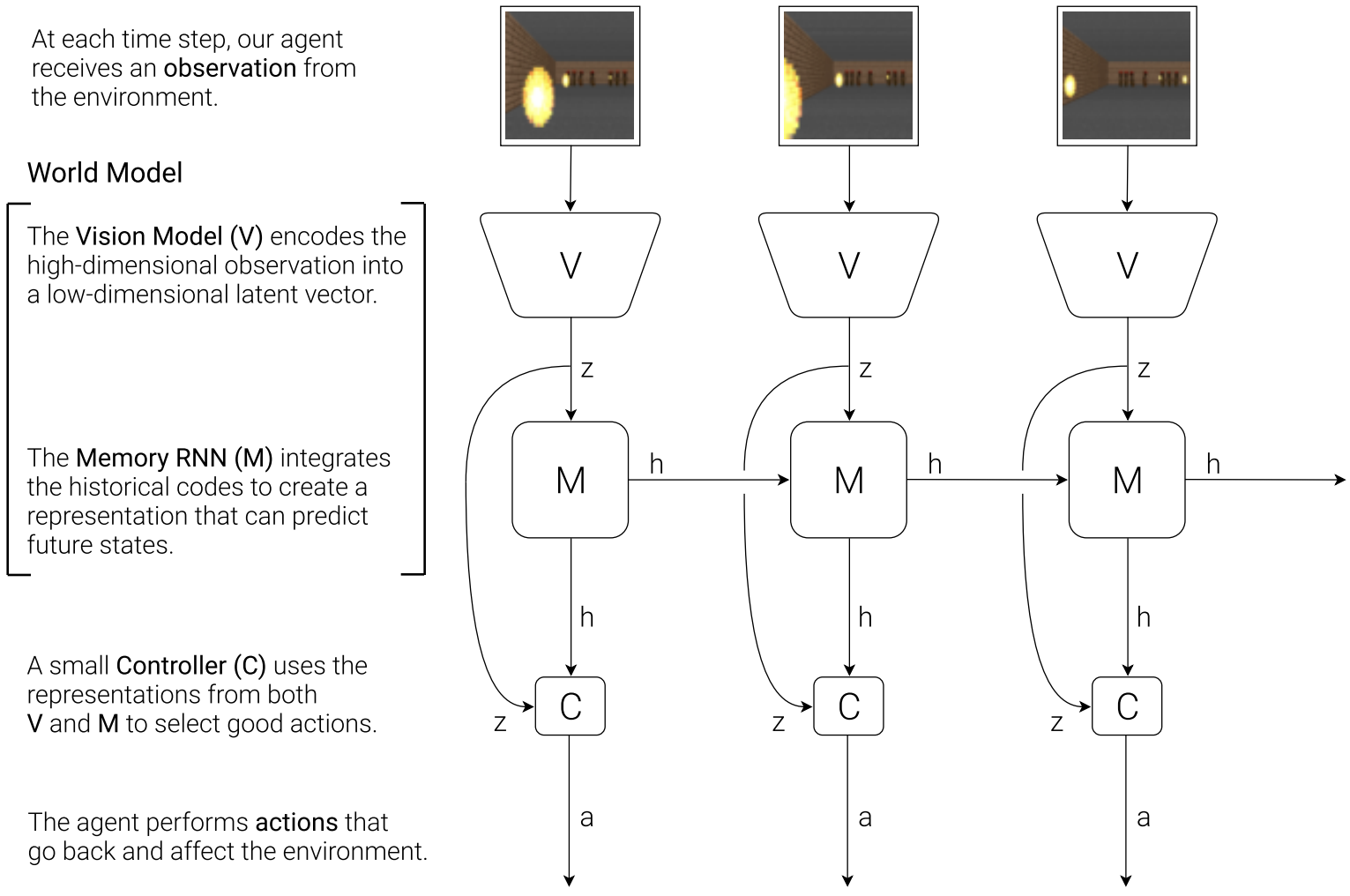


Figure: A simple model inspired by our cognitive system consists of three components: **Vision (V)**, **Memory (M)** and **Controller (C)**.

VAE (V) Model

Learn an abstract, compressed representation of an observed frame.

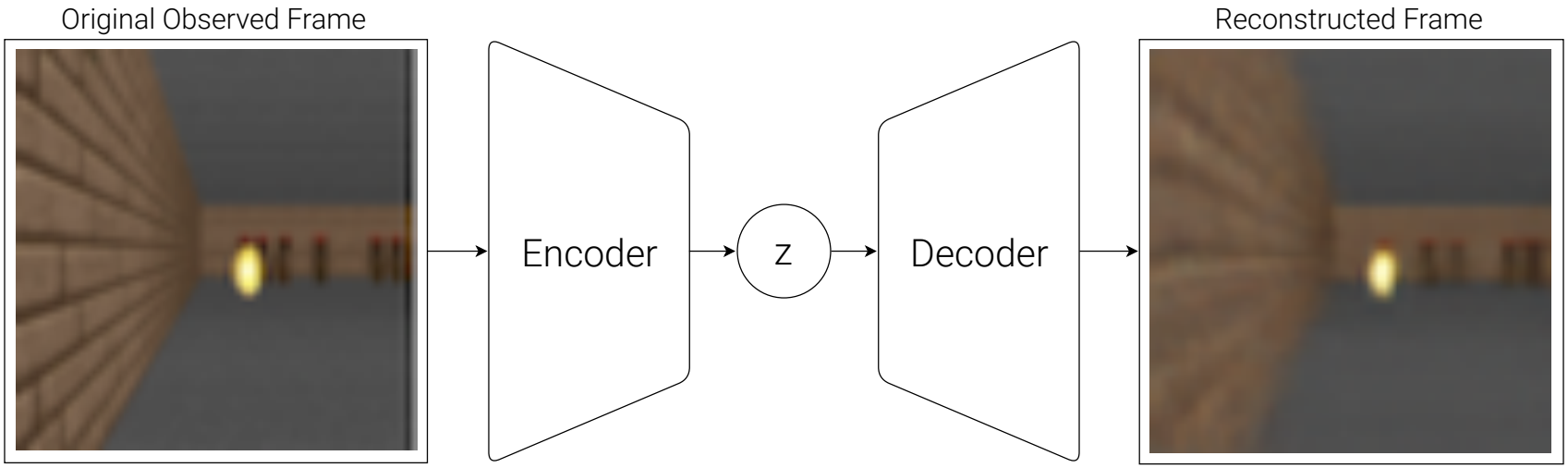


Figure: Flow diagram of a Variational Autoencoder.[3]

[3] D. P Kingma and M. Welling. "Auto-Encoding Variational Bayes". In: *ArXiv e-prints* (Dec. 2013). arXiv: 1312.6114 [stat.ML].

MDN-RNN (M) Model[4]

Compress what happens over time, predict future.

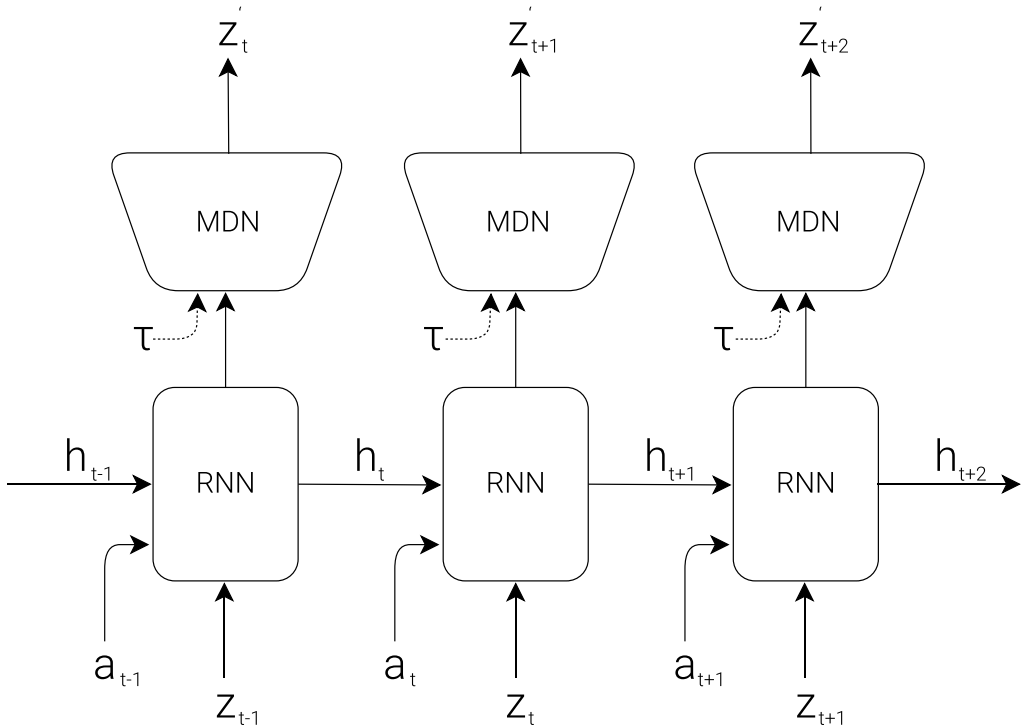


Figure: Long Short-Term Memory (LSTM) with a Mixture Density Network (MDN) models $\mathrm{P}(z_{t+1}|a_t, z_t, h_t)$ with Gaussian mixture.

---

[4] Alex Graves. "Generating Sequences With Recurrent Neural Networks". In: *CoRR* abs/1308.0850 (2013). arXiv: 1308.0850. URL: http://arxiv.org/abs/1308.0850.

Controller (C) Model

Simple and small so that most complexity is in the *world model*.

$$a_t = W_c[z_t, h_t] + b_c$$

Trained with Covariance-Matrix Adaption Evolution Strategy.[5]

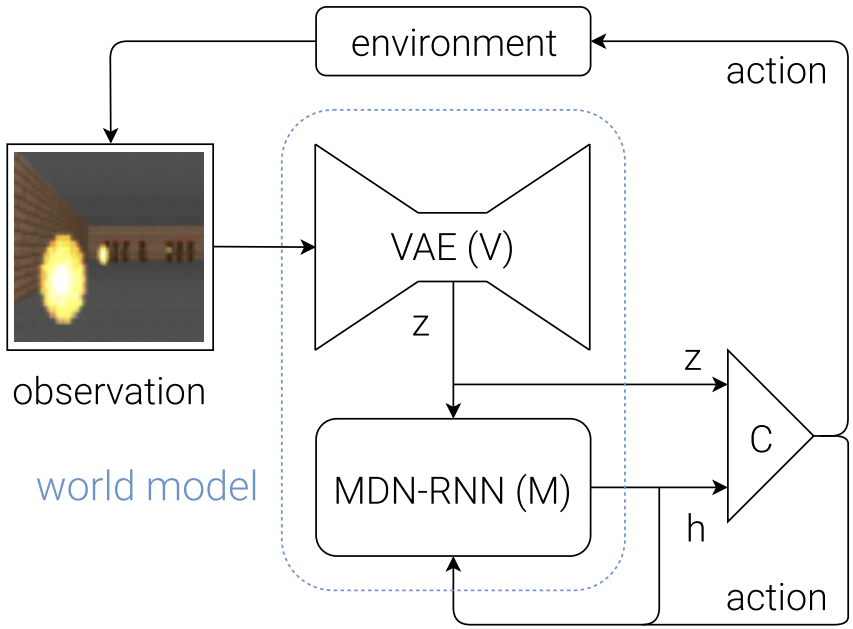


Figure: Flow diagram of the agent model.

[5] Nikolaus Hansen. "The CMA Evolution Strategy: A Tutorial". In: *CoRR* abs/1604.00772 (2016). arXiv: 1604.00772. URL: http://arxiv.org/abs/1604.00772.

Outline

Vision (V) Model Only

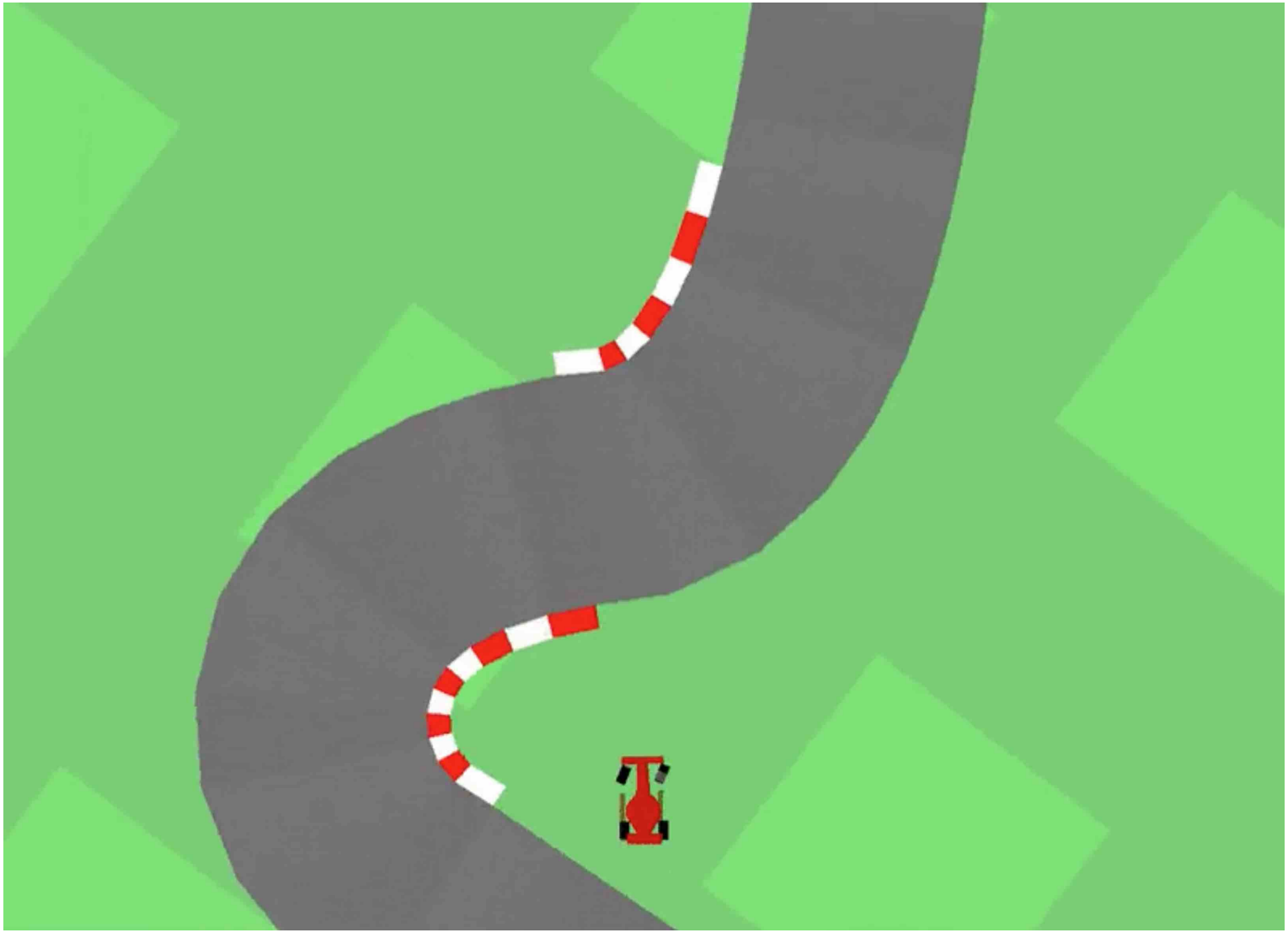


Figure: Limiting the controller to see only z_t.

Full World Model (V and M)

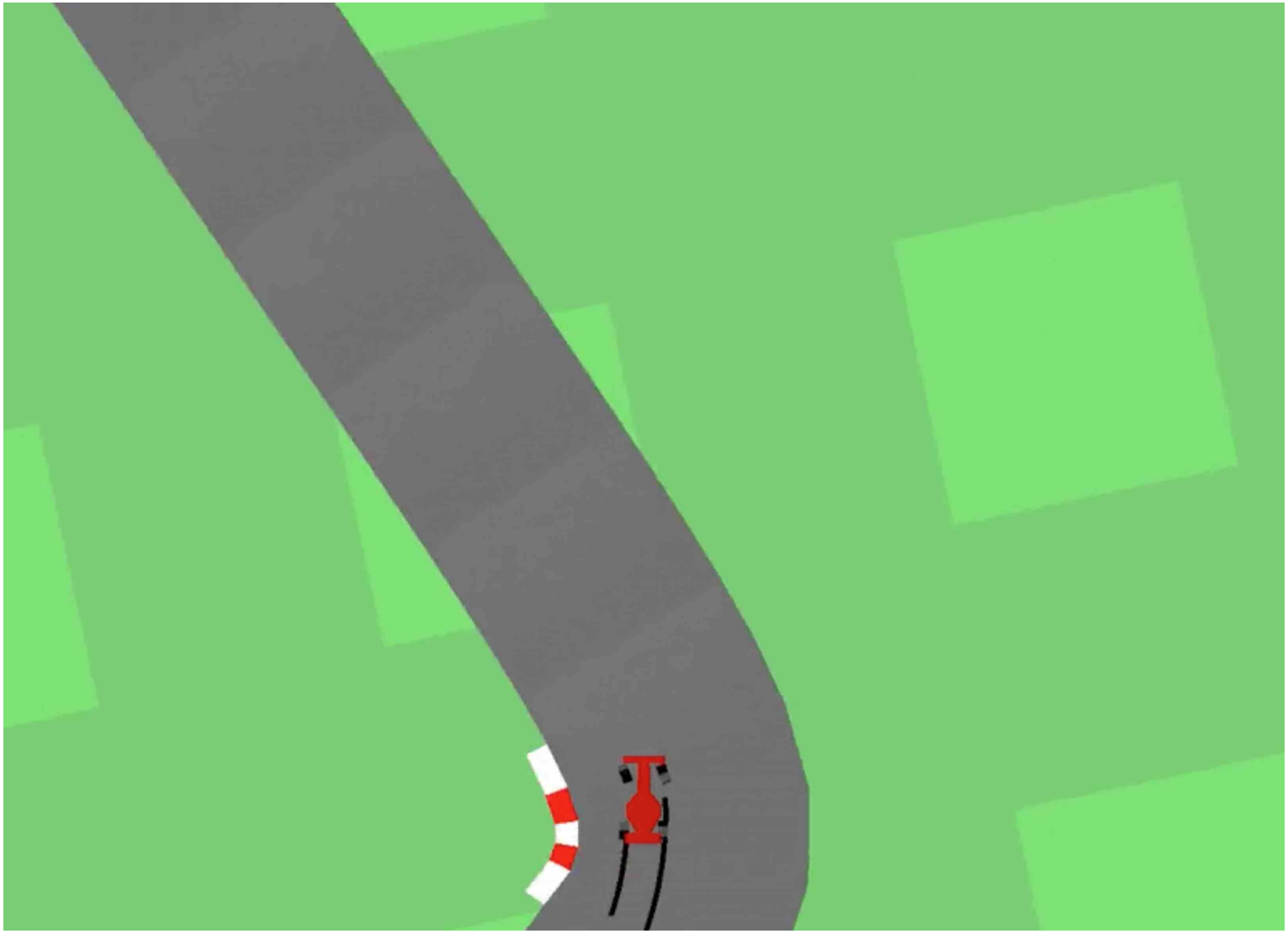


Figure: Stable driving with access to both z_t and h_t.

Car Racing Results

Method	Avg. Score
DQN	343 ± 18
A3C (continuous)	591 ± 45
A3C (discrete)	652 ± 10
ceobillionaire (Gym Leaderboard)	838 ± 11
V model	632 ± 251
V model with hidden layer	788 ± 141
Full World Model	**906 ± 21**

Table: `CarRacing-v0` scores achieved using various methods.

Outline

VizDoom Experiment

Can an agent be trained in inside its own hallucination and transfer learned policy back to the actual environment?

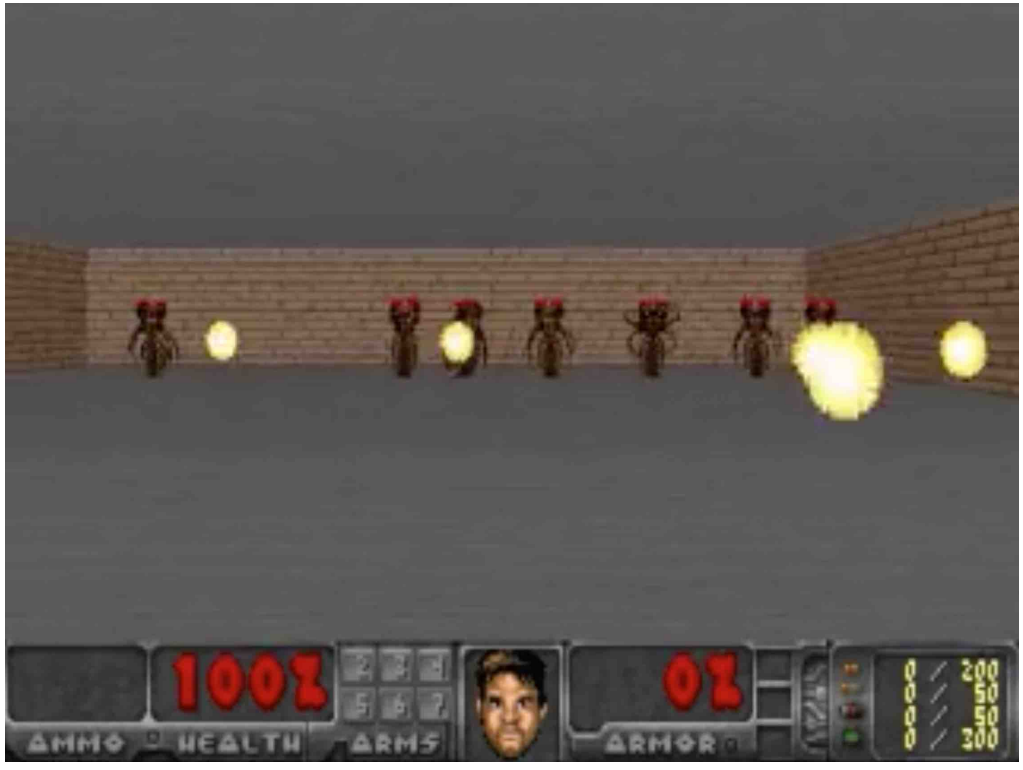


Figure: *VizDoom: Take Cover* environment.

Training in a Hallucination

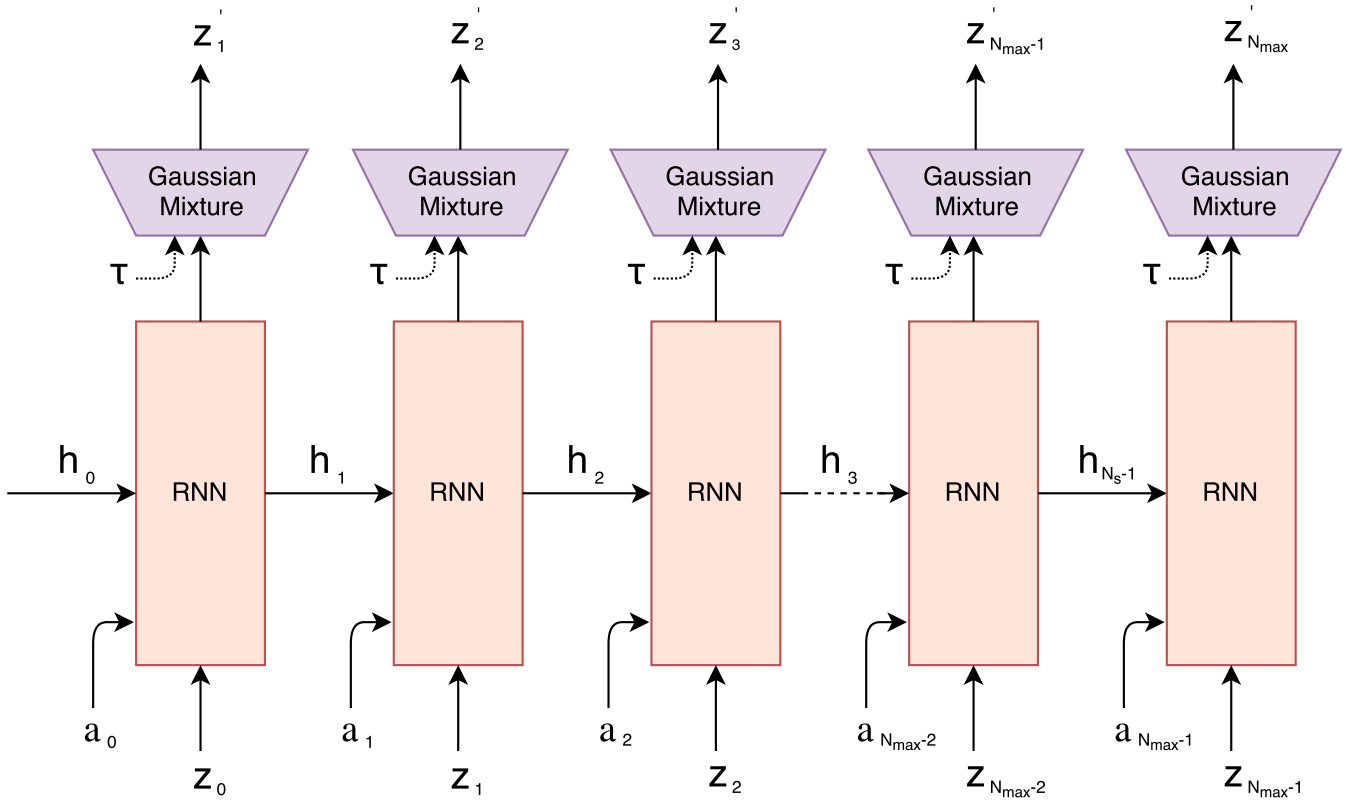
- The M model also predicts whether the agent dies in the next frame.

Training in a Hallucination

- The M model also predicts whether the agent dies in the next frame.
- Do not need the V model to encode real pixel frames.

Training in a Hallucination

- The M model also predicts whether the agent dies in the next frame.
- Do not need the V model to encode real pixel frames.
- Possible to add extra uncertainty into the virtual environment via the parameter τ.



Policy Transfer Results

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

Table: *VizDoom: Take Cover* scores at various temperature settings.

Cheating the World Model

Figure: Agent discovers an adversarial policy to automatically extinguish fireballs.

Outline

Discussion

- Do not waste cycles training agent in the actual environment.
- VAE encodes irrelevant part of observations.
- Unlike human brain[6] LSTM suffers from catastrophic forgetting.
- Only simulates future time steps without human-like hierarchical planning or abstract reasoning.[7]

---

[6] Jr Bartol Thomas M et al. "Nanoconnectomic upper bound on the variability of synaptic plasticity". In: *eLife* 4 (Nov. 2015). Ed. by Sacha B Nelson, e10778. ISSN: 2050-084X. DOI: 10.7554/eLife.10778. URL: https://doi.org/10.7554/eLife.10778.

[7] Jürgen Schmidhuber. "On Learning to Think: Algorithmic Information Theory for Novel Combinations of Reinforcement Learning Controllers and Recurrent Neural World Models". In: *CoRR* abs/1511.09249 (2015). arXiv: 1511.09249. URL: http://arxiv.org/abs/1511.09249.

Questions?