Distilling Symbols from Perceptual World Models

CSE 598: Perception In Robotics

Lucas Saldyt, Maxime Zand

May 18, 2024

Arizona State University

Perceptual World Models and Symbols (Problem)

World models **predict future states from image and action input**.
E.x. if an agent moves forward, what do they see next?

An agent that predicts minecraft dynamics should implicitly understand symbols, such as types of blocks

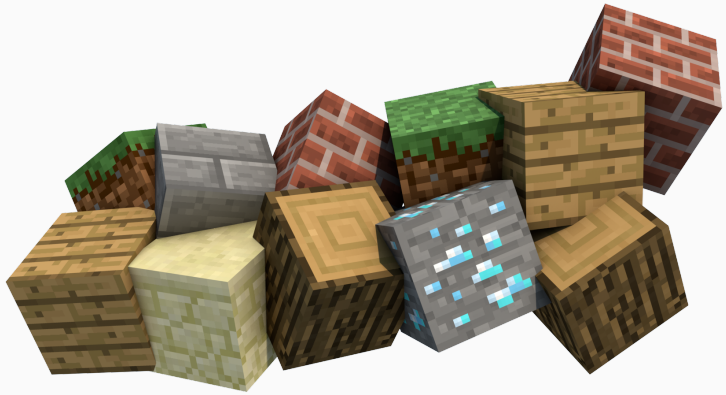

Figure 1: Minecraft blocks e.g. grass, wood, stone

We study the crafter domain (2D minecraft)
Our project goal is to distill these symbols without supervision

Dreamer (Related Work)

The Dreamer world model predicts future states given actions, observations, and memory. Memory is stored in a real vector $h_t \in \mathbb{R}^n$

h_t is not symbolic, but could encode implicit symbols.

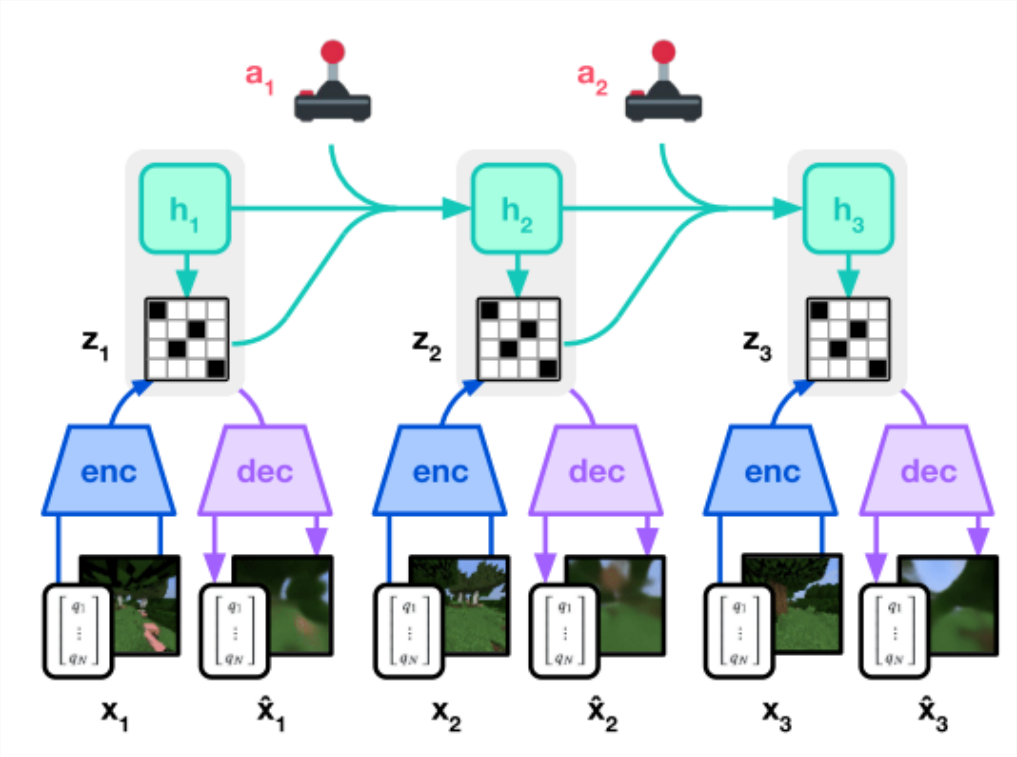

Figure 2: Dreamer Architecture

Methods

Sparse Autoencoding

We propose leaving Dreamer's z_t and h_t unaltered, and introducing a new learned latent code $\hat{c}_t$, produced by an autoencoder on h_t, which is regularized to be sparse using an L1 norm:

$$\underset{\text{reconstruction}}{\mathcal{L}}(\phi) = \underset{\text{reconstruction}}{\beta} \left\| h_t - \hat{h}_t \right\|_k \tag{1}$$

$$\underset{\text{sparsity}}{\mathcal{L}}(\phi) = \underset{\text{sparsity}}{\beta} \|\hat{c}_t\|_k \tag{2}$$

$\underset{\text{sparsity}}{\mathcal{L}}$ and $\underset{\text{reconstruction}}{\mathcal{L}}$ are added to the overall loss $\mathcal{L}(\phi)$

Mutual Info Regularization

Use methods from InfoGAN to distill symbols in Dreamer

See future work section

Results (Aggregate)

A sparse autoencoder finds more symbol correlations more quickly. Correlation are measured against block class labels

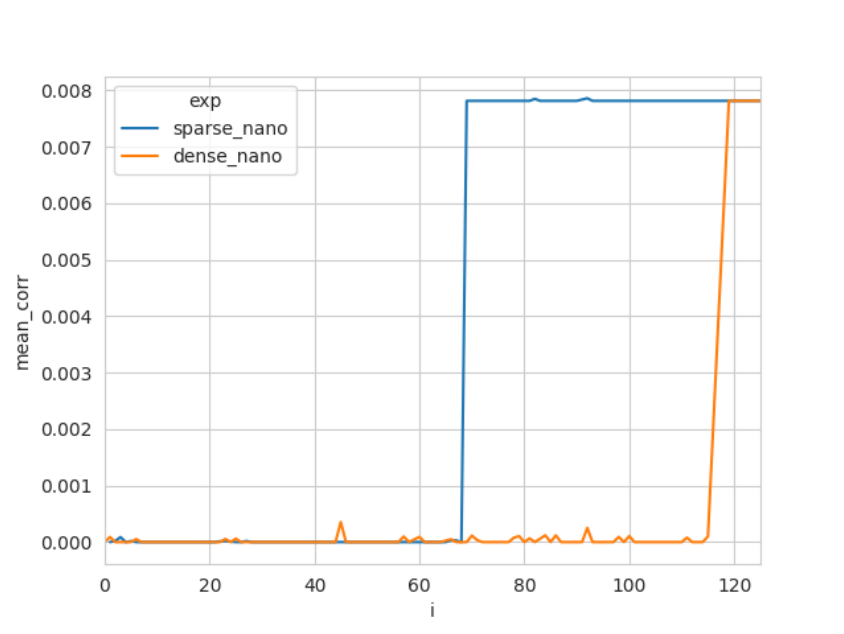

Figure 3: Mean Correlation

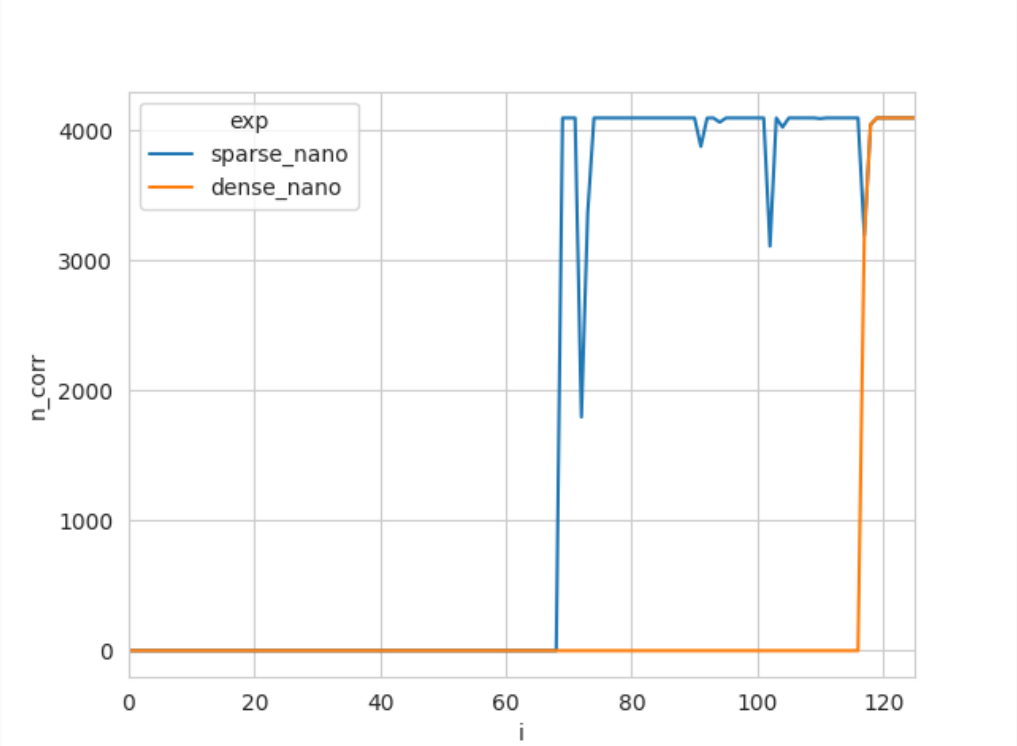

Figure 4: Number of correlations > threshold

X-axis is episodes, training from scratch.

Results (Example)

Sparsity reduces entanglement
e.g. this neuron (right, 43) approximately encodes the material of blocks in the crafter environment

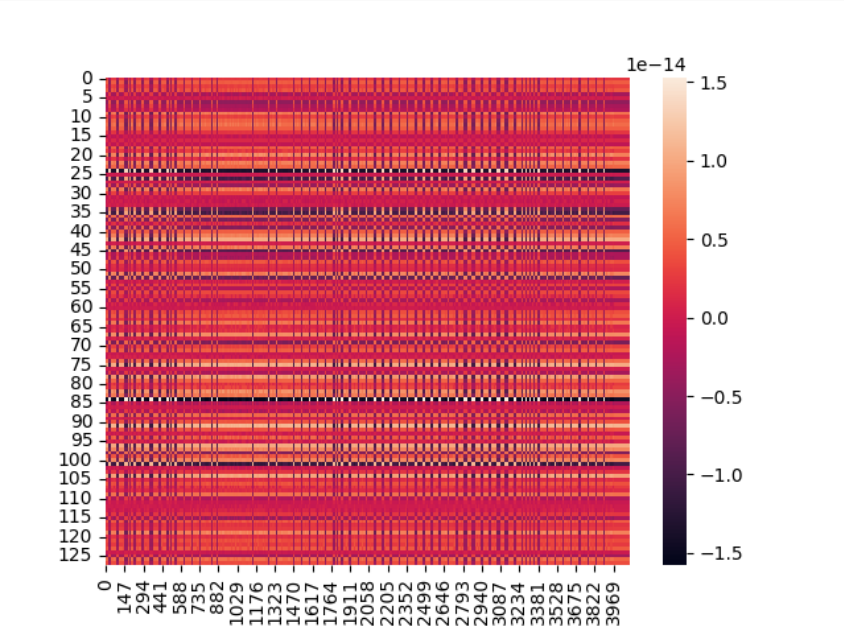

Figure 5: Dense Correlation (baseline)

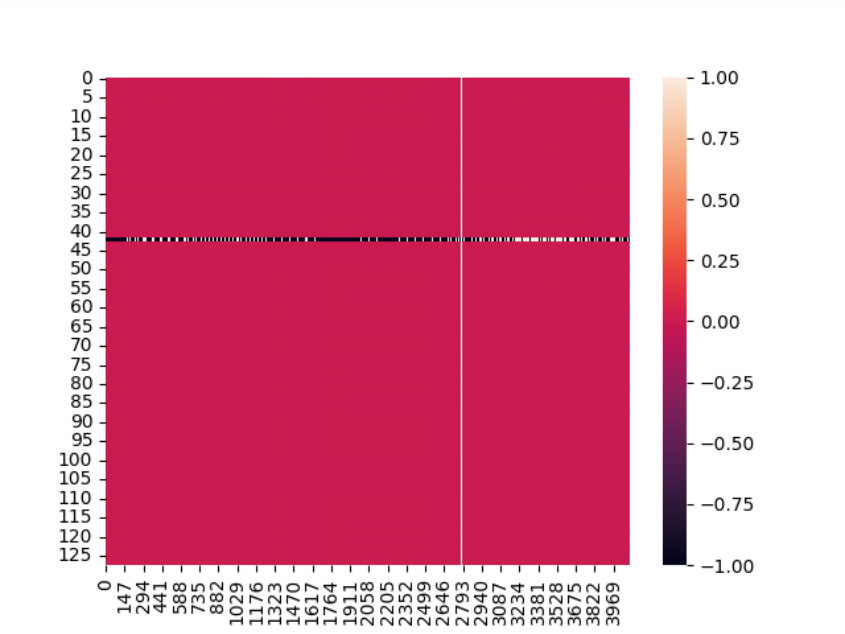

Figure 6: Sparse Correlation

The y axis is the latent vector h_t
The x axis is one-hot encoded block classes

Future Work (Backup)

Furthermore, we propose:

- Structuring c_t, e.g. concatenation of categorical and gaussian
- Predicting dynamics in terms of c_t with a separate function g:

$$f: \quad h_{t-1}, a_{t-1}, z_{t-1} \quad \mapsto \quad h_t \quad \text{Latent dynamics} \tag{3}$$

$$g: \quad c_{t-1}, a_{t-1}, z_{t-1} \quad \mapsto \quad c_t \quad \text{Symbolic dynamics} \tag{4}$$

- Repredicting a thresholded (sparse) $\dot{c}_t$ from $\hat{c}_t = p(z_t, h_t)$
- Reprediction loss is a lower bound on mutual information