DL final presentation

Team 14
Yizheng Li, Yubo Wang, Zichu Wang

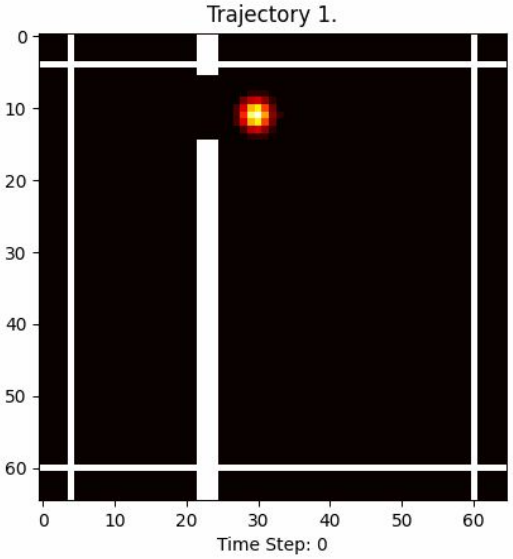
Task

We want

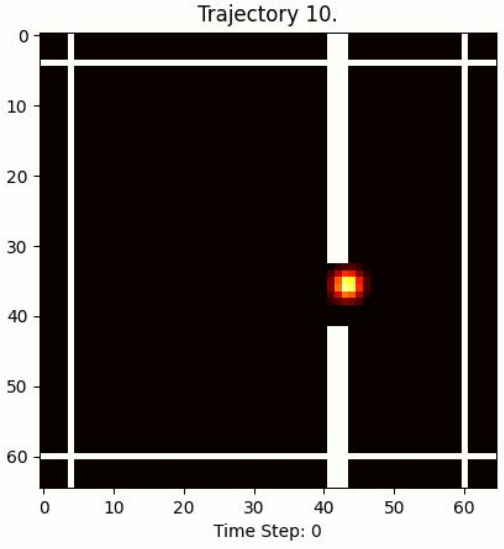
1) the **ball** to correctly **follow the action sequence**, and

2) the **model** to **stop the ball** from achieving the action when the **wall** hinders it.

Example of a free trajectory:



Example of a stopped trajectory:



Main Challenges

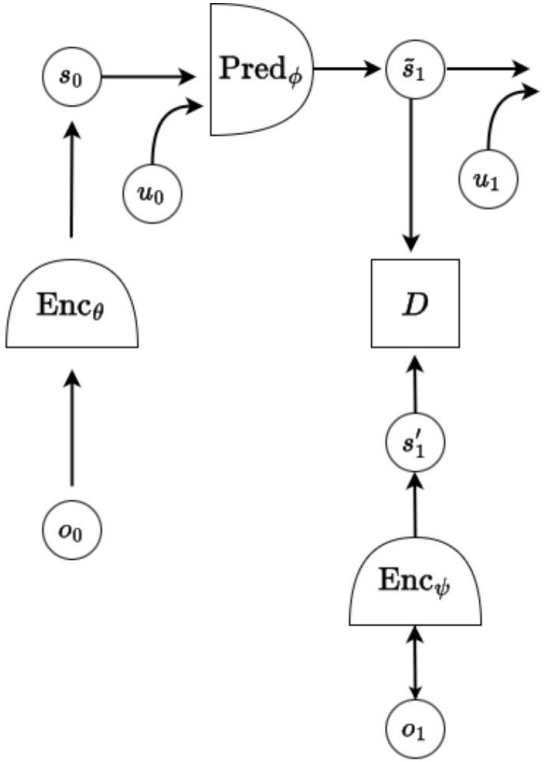
Joint-embedding models risk **representation collapse**, where all inputs map to nearly identical vectors. There are two usual directions from an energy model perspective explanation:

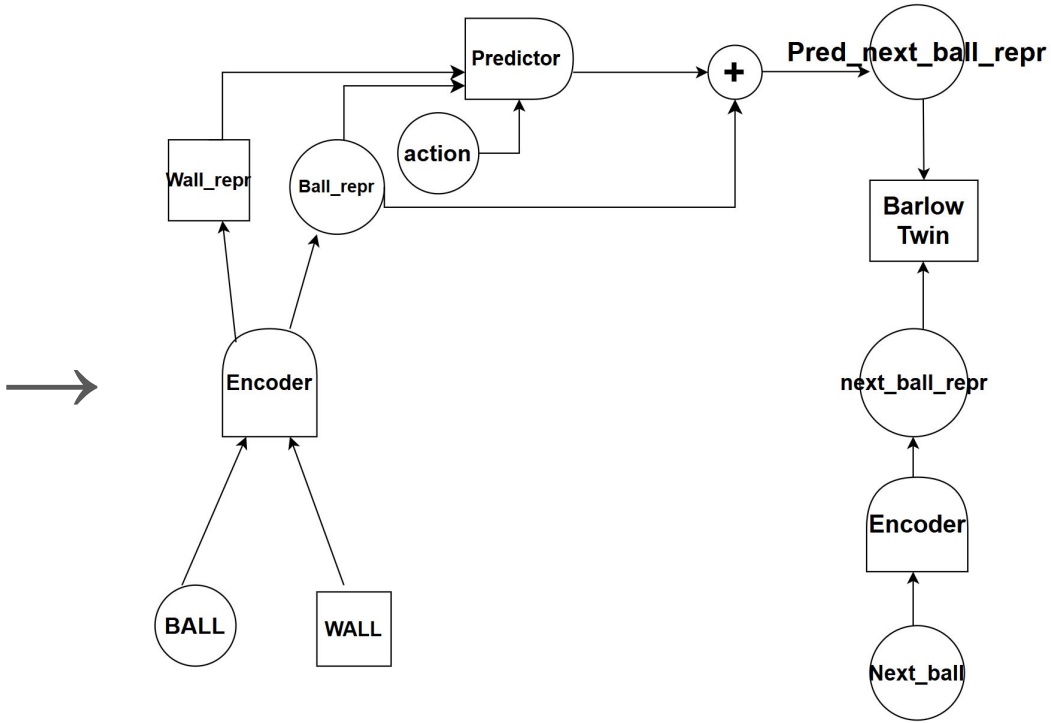
- **Contrastive Methods** (e.g., SimCLR, MoCo) push down the energy(incompatibility score) of the positive sample and pull up the energy of the negative sample.

- **Regularized Methods** (e.g., Barlow Twins, VICReg) push down on the energy of training samples and use a regularizer term that minimizes the volume of low-energy regions.

We adopt the second approach, trying both Barlow Twins and VICReg to produce decorrelated, information-rich embeddings without contrastive learning.

JEPA architecture

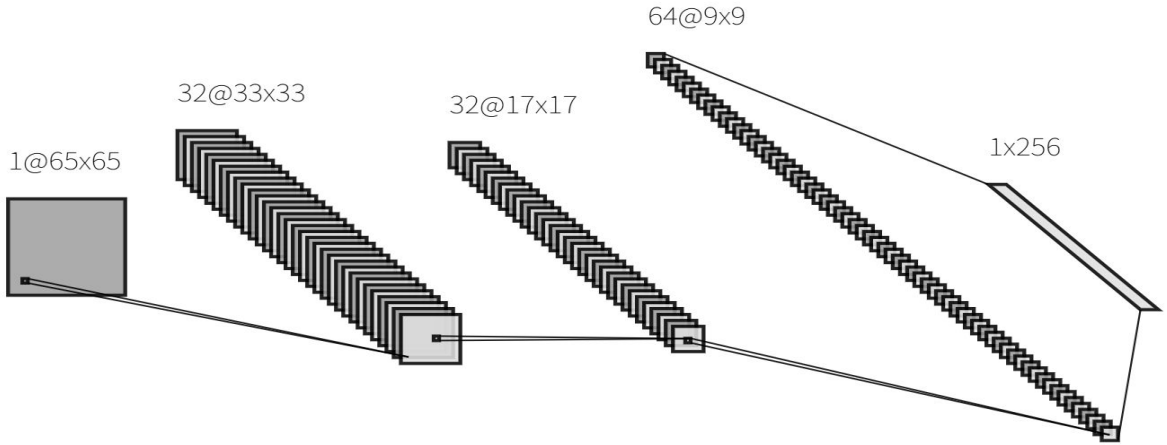




Architecture

Encoder: 3 layer CNN with **BN+ReLU**, output was flattened and fed to a single layer FC networks with **tanh** activation, output is **256d** representation of ball or wall. We **separate the wall and ball channel** and encode them independently.

The structure:

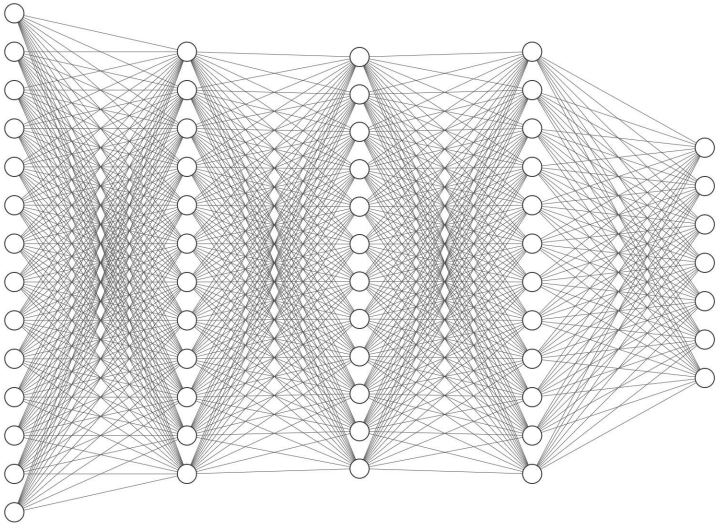


Architecture

Predictor: 4 layer **FC** networks with 3 **ReLU** activations between the middle layers and **tanh** activation at the end. 256d ball_repr, 256d wall_repr and 2d action are fed into the the predictor, output is 256d next_ball_repr.

Input: [ball_repr, wall_repr, action], 514 dim

Output: [next_ball_repr], 256 dim



Training

1st stage: Use BT to train the whole model.

- We hope to learn a non-collapsing representation, Barlow Twins is effective
- However the descent was limited. after 10 epochs, Normal Loss ~ 3, Wall Loss ~ 7

2nd stage: switch to MSE for further training

- The final evaluation uses MSE to measure the distance between the predicted and real coordinates of the ball, so we try to align our loss with it, and found it worked.
- After change to MSE, we find that it's **OK to compromise the representation**, so long as it doesn't completely collapse (i.e. Barlow Twin loss doesn't explode).
- The descent continued, but the wall loss decreased very slowly and got stuck at a point.

after ~35 epochs, Normal Loss ~ 0.5, Wall Loss ~ 1.7

3rd stage: stop training Ball Encoder

- We then freeze ball encoder and focus on the wall encoder and the predictor with MSE, and we find that the wall loss started to decrease again.

after ~15 epochs, Normal Loss ~ 0.35, Wall Loss ~ 0.75

Techniques explored

Architecture:

Complex encoders and predictors may not be optimal (we tried ViT, multi-head attention, etc.)

Activation: **Tanh** works well as the final activation function (all coordinates are limited in [-1,1]), as it may provides a soft boundary at the tails, which could meet the geometry of the task

Regularization: **Barlow Twins**, VICReg, weighted combinations of some terms

weight sharing for the input and target encoders.

Training pipelines are very important (separating stages).

Augmentation:

Random cuts and lengthening on trajectories

Flip states and actions

Add Gaussian noise to the pictures.

Results

Normal Loss: 0.32467275857925415

Wall Loss: 0.741540789604187

Wall Other Loss: 2.796499729156494

Expert Loss: 1.667962074279785 2

Total loss around 5.52

Number of parameters: 3631680