Hierarchical Relational Inference

Aleksandar **Stanić**
aleksandar@idsia.ch

Sjoerd van **Steenkiste**
sjoerd@idsia.ch

Jürgen **Schmidhuber**
juergen@idsia.ch

IDSIA — Scuola universitaria professionale della Svizzera italiana SUPSI — Università della Svizzera italiana

GitHub/astanic/hri

Summary

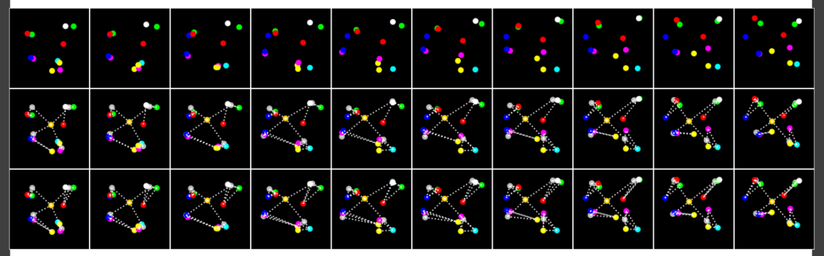
- Common-sense physical reasoning, requires learning about objects, their interactions and dynamics directly from raw visual input
- Real world objects vary greatly in terms of their properties, which complicates modelling their dynamics
- Many objects can be viewed as a hierarchy of parts that locally behave independently, but also act more globally as a single whole
- We learn a hierarchical world model (HRI) that explicitly distinguishes objects at multiple levels of abstraction and relations between them
- Use cases: interacting with complex objects (e.g. humanoid robot), or learning a hierarchically structured predictive model for MBRL

Experimental Questions

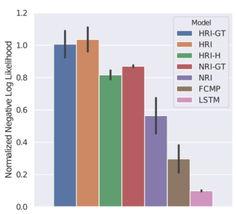
- Can parts, objects and relations be inferred from visual data?
- What is the effect of relational inference?
- Does hierarchical graph inference and reasoning help?
- Can HRI be scaled to real world visual scenarios?

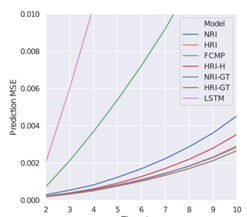
State Springs

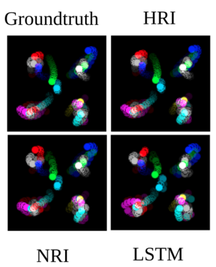
- Synthetic dataset of state trajectories of objects and parts connected via finite-length springs in a hierarchical structure
- HRI makes accurate predictions about the future state of the environment and correctly infers the underlying interaction graph



- Comparisons to baselines show benefits of relational inference (HRI vs LSTM) and hierarchical message passing (HRI vs NRI)



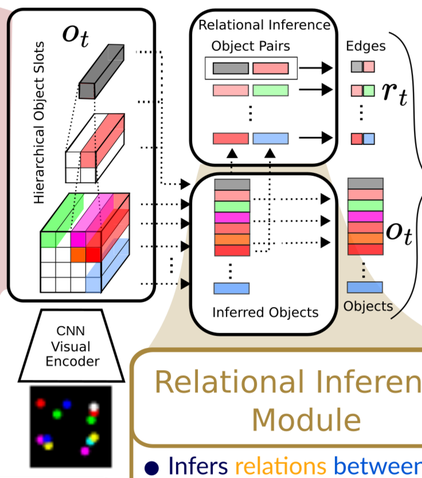




HRI model

Visual Encoder

- Infers representations of objects and parts from the input image
- Object slots: partition CNN's feature maps
- Spatial bias ensures part-object grouping
- Weight sharing ensures common format



Visual Decoder

- Renders predicted object states back to image space
- Allows to train HRI like a VAE
- Objects (slots) are decoded individually, and then combined in pixel space [SlotDec]
- ParDec: decode all objects together (weaker inductive bias)

Training Objective

- HRI is tranied like a VAE: latent variables are the inferred objects and relations

$$\mathcal{L} = \mathbb{E}_{q_\phi(r|x)}[\log p_\theta(x|r,o)] - D_{KL}[q_{\phi_o}(o|x)||p_{\theta_o}(o)] - D_{KL}[q_{\phi_r}(r|x,o)||p_{\theta_r}(r)]$$

Relational Inference Module

- Infers relations between objects and parts, to form the (dynamic) interaction graph
- Edges computed via message-passing in a GNN

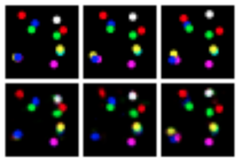
$$e_{i,j} = f_e([o_i, o_j, r_{i,j}]), o'_j = f_o([\sum_{i \in \mathcal{N}_{o_j}} e_{i,j}, o_j])$$

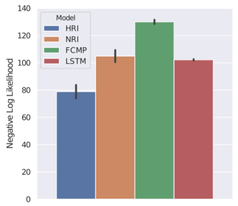
Dynamics Predictor

- Predicts future object states based on inferred interaction graph and current objects via Hierarchical Message-Passing (HMP)
- HMP distinguishes 3 separate phases for efficiency: bottom-up, within-siblings and top-down phase

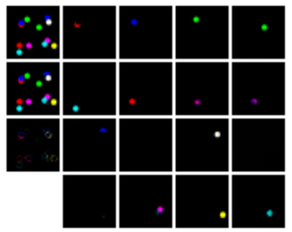
Visual Springs

- Rendered version of the springs dataset only a part of the hierarchy is visible
- HRI outperforms both NRI and LSTM in terms of NLL for physical reasoning
- Spatial slots learned by the visual encoder capture individual objects and yield object representations
- HRI outperforms NRI at inferring the interaction graph in the visual domain

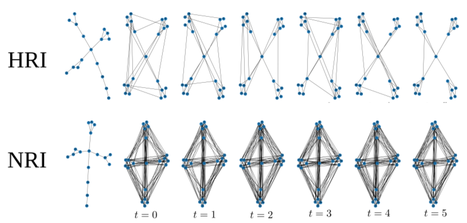




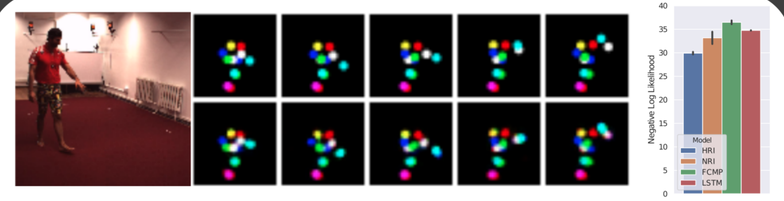
Object Slots Introspection



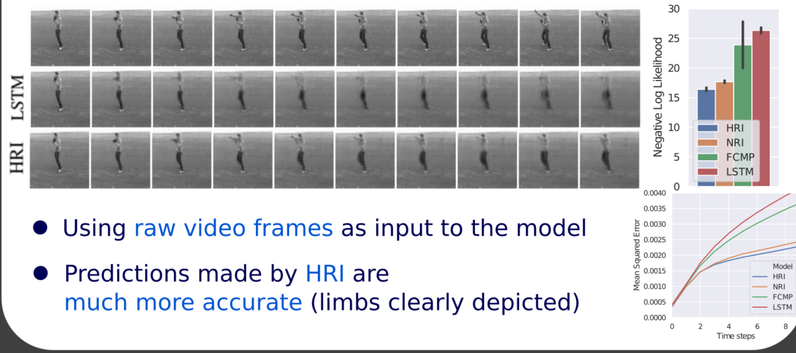
Inferred Graphs



Human 3.6M and KTH



- Render joints - a visual representation based on keypoints
- Non-stationary dynamics and less hierarchical interactions
- Still able to observe improvements using HRI



- Using raw video frames as input to the model
- Predictions made by HRI are much more accurate (limbs clearly depicted)