Proposal: Hierarchical Predictive Coding Framework with Diffusion-Based Environment Generation for 3D Scene Understanding

Ramakrishna Kompella

1 Conceptual Overview

1.1 Hierarchical Latent Representations

- **Sensory Layer (Bottom)**: Encodes the agent's immediate visual input (e.g., the current camera frame). This is typically a high-resolution but localized representation.
- **Environment Layer (Top)**: Encodes a latent representation of the entire environment (the "room" or scene) that the agent is navigating. As the agent moves, the environment latent is updated/inferred by fusing new sensor data and prior beliefs about the environment.

1.2 Active Predictive Coding (APC)

- Each layer of the hierarchy predicts the representation at the layer below and updates itself to minimize prediction error.
- The agent (or model) is "active,". It moves and orients its sensors to explore the environment in an effort to reduce uncertainty

1.3 Generative Model (Diffusion Model) for the Environment

- Once the environment latent representation (upper layer) is learned for each environment, a separate diffusion-based generative model is trained to reconstruct or synthesize entire 3D environments or environment-level features.
- The generative capability allows us to generate novel environments.

1.4 Consistent 3D Generation

- By combining the environment-level diffusion model with the active state-prediction mechanism of APC, the system aims to produce consistent and realistic 3D reconstructions (or novel generations) of the agent's surroundings.

2 Technical Formalization

Let:

- $\mathbf{x}_t$ be the raw sensory input at time t (e.g., RGB or RGB-D images).
- $\mathbf{z}_t^s$ be the latent representation in the **Sensory Layer** at time t.
- $\mathbf{z}_t^e$ be the latent representation in the **Environment Layer** (upper layer) at time t.

2.1 APC Forward Model

2.1.1 Sensory Layer

$$\mathbf{z}_t^s = f_s(\mathbf{x}_t; \theta_s) \quad \text{(Encoder)} \tag{1}$$

$$\hat{\mathbf{x}}_t = g_s(\mathbf{z}_t^s, \mathbf{z}_t^e; \phi_s) \quad \text{(Decoder)} \tag{2}$$

The reconstruction or predictive coding mechanism minimizes the prediction error:

$$\mathcal{L}_{\text{sensory}} = \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2 + \text{regularization terms.} \tag{3}$$

2.1.2 Environment Layer

$$\mathbf{z}_t^e = f_e(\mathbf{z}_{t-1}^e, \mathbf{z}_t^s; \theta_e) \tag{4}$$

$$\hat{\mathbf{z}}_t^s = g_e(\mathbf{z}_t^e; \phi_e) \tag{5}$$

The environment layer integrates information over time and minimizes:

$$\mathcal{L}_{\text{env}} = \|\mathbf{z}_t^s - \hat{\mathbf{z}}_t^s\|^2 + \text{regularization terms.} \tag{6}$$

2.1.3 Overall Objective

$$\mathcal{L}_{\text{APC}} = \sum_t \Big(\mathcal{L}_{\text{sensory}} + \mathcal{L}_{\text{env}}\Big). \tag{7}$$

2.2 Diffusion Model for Environment Generation

We introduce a diffusion model D_α (with parameters α) over the **environment latent space** $\mathbf{z}^e$. Training includes environment-level latents $\mathbf{z}^e$ extracted from large-scale 3D datasets. The diffusion model learns:

$$p(\mathbf{z}^e) \approx D_\alpha(\mathbf{z}^e). \tag{8}$$

At inference time, the model can sample from the learned distribution to produce plausible completions or novel environment latents.

3 Choice of Models and Techniques

- **Sensory Encoder/Decoder**: CNNs, Vision Transformers, or VAEs for compact latent representations $\mathbf{z}^s$.
- **Environment Aggregator**: GRU/LSTM/Transformers for temporal aggregation or NeRF-like scene representation.
- **Diffusion Model**: Latent Diffusion for scalable 3D latent space modeling.
- **Active Policy**: Reinforcement Learning or Active Inference to minimize predictive error.

4 Datasets and Training Procedure

4.1 Datasets

- **Matterport3D**: Real indoor multi-room scans.
- **Replica**: High-quality reconstructions of indoor spaces.
- **RealEstate10K**: Real-world property tours.

4.2 Training Phases

1. **Phase A: APC End-to-End Pre-Training**
2. **Phase B: Diffusion Model Training**
3. **Phase C: Integration and Fine-Tuning**

5 Relation to Prior Work

- Predictive Coding: Lotter et al. (2016), Friston et al. (2009).
- World Models: Ha and Schmidhuber (2018), Hafner et al. (2019).
- Scene Representation: Eslami et al. (2018), Mildenhall et al. (2020).
- Diffusion Models: Ho et al. (2020), Rombach et al. (2022).