Research Proposal

In neuroscience, the cognitive map theory claims that our brain represents everything in this world as a relative relationship, whether from the temporal, spatial, or abstract perspective. Correspondingly, contrastive learning trains a network with data pairs. Here, I want to build an **embodied agent** without supervision from scratch based on contrastive learning, joint-embedding predictive architecture (JEPA), and the "World Model" proposed by LeCun from a temporal perspective. The agent only depends on RGBD perception and proprioception. The agent will "grow up" like humans and learn everything gradually, from self-body control, locomotion, and semantic segmentation to language. In "World Model", LeCun illustrated how an intelligence system should be. Here, I want to instantiate his idea to a more detailed framework.

One critical function of our cognitive system is predicting the future. We must estimate how our decisions and actions will change the current state. We need to predict the actions and inner states of others in cooperation or competition. As an essential dimension of this world, most physical rules include time as an explicit or implicit parameter. Based on these observations, I want to instantiate an intelligence system from scratch under the temporal JEPA framework without supervision. The model depends on the temporal dynamic of both the outer world and inner states. LeCun described JEPA as $pred(emb(x), z) = emb(y)$. When I talk about temporal JEPA, I mean $pred(emb(s_t), action) = emb(s_{t+1})$.

Why do we need an embodied environment? Memory can be divided into declarative memory and non-declarative memory. We can't learn how to ride a bicycle by reading books. We learn to control our bodies and the bike by trial and error. We learn from mistakes (reward prediction error), and embodied simulators provide a low-cost playground for making mistakes. The agent can learn forward and inverse kinetics between body states and motion commands in simulators, ex: $FK(s_t^{inner}, cmd) = s_{t+1}^{inner}$. This form means self-body control can be learned by temporal JEPA. Another advantage of the embodied environment is that the agent can get multimodality inputs. With an embodied environment, multimodality information like vision, audio, and touch is fused naturally. Traditional approaches like ImageBind and CLIP learn the relative relationship between multimodality data. However, embodied simulators like ThreeDWorld and SoundSpaces naturally bind multimodality data into specific objects. Then, we can use a method like multisensory NeRF to build a unified implicit representation of the environment (subjective embedding of perception). Local features will be retrieved only if we observe them. The location parameter in NeRF satisfies the setting up of active perception.

Why do we have a concept of "myself" in our minds? Allocentric representation is an extension of egocentric recognition. We can control our bodies with a single thought. But we can't control the environment without actions. Our body is deterministic, and the environment is highly dynamic. As humans, we learn how to control our bodies as infants. Then, we learn how to influence the environment with actions. Although foundation models like RT-2 combined perception with robot controls, I believe we need to separate these two systems since they are heterogeneous in principle. This viewpoint is also mentioned in LeCun's paper: the world model and the action

module are separated. Another divergence between allocentric and egocentric cognition is that our body is not detachable. Our proprioception is constant, and we can always feel our limbs. But for external objects, our perception is invoked only if they somehow connect to us. “Use a sword like an arm”. It may take years of practice if an external object wants to merge into our inner control. The agent must have a clear boundary between predictable behaviors and the unreliable environment to get the desired outcome of actions.

How do we learn physical rules and common sense? Before we can understand the foundation rules of this world, we learn to segment objects. An object is an indivisible entity. Different properties distinguish various things. When people use their hands to push an object, it will move with time in our visual perception. The object can be segmented by its dynamic property (interactive perception). Visual tasks like segmentation, detection, and tracing can be learned this way. With object definition, we can learn physical rules and common sense from constant events. “Balls always fail on the ground” implies gravity. Now, our predicting system can extend from inner states to world states, ex: $WorldModel(s_t^{outer}, force) = s_{t+1}^{outer}$. We can predict what will happen in the next time step. We can infer how to grape and manipulate external objects by using this world model as a reward or critic function in RL search.

What about generative models? In LeCun’s paper, generative models can be interpreted under the JEPA framework, which predicts the embedding of y with latent variable z. In the second paragraph of this research proposal, actions are substituted into JEPA as a latent variable. The fourth paragraph claimed that "allocentric representation is an extension of egocentric recognition.” If we extend our egocentric action into allocentric force or other latent variables, we can extend the self-body control model into a world state prediction model. Under this assumption, generative models like GAN, VAE, or GPT are equivalent to imagination in our mind. These models have no essential difference from the self-control model mentioned above. Each sample generated by these models is a possible future. Recent work UniSim (which uses a generative model as a universal embodied simulator) is a brilliant implementation of this viewpoint. However, using videos as a universal proxy for RL and imitation learning is not a new idea. In MineDoJo, video-language pairs are used to train a game agent in Minecraft. In VPT, an inverse dynamic model (IDM) is trained to infer the low-level actions given a video. In UniPi, video-as-policy is proposed and IDM is combined with diffusion models to achieve text-guided action planning.

How do we develop language? Language is an abstraction of embodied experience and externalization of the world model. The world model is virtual in our minds, and everyone has his or her version of the world model. The language was first developed as a bridge to connect everyone’s world models (just like communication protocols). It is a pure symbol game if language is not grounded with specific entities in the real world. There are hundreds of languages, but they all describe one shared real world. That is the foundation of the cross-language translation. For an embodied agent, we should consider how to bind its inner representation of the environment with human language. Object-based scene graphs and word embedding (token) maps may help, ex: $similarity(edge(obj1, obj2), edge(token1, token2))$.

Another question is how to build a foundation model with a close-loop data pipeline. Existing foundation models are used top-down and LLMs are applied in various downstream tasks. But I

want to build foundation models bottom-up. The training data we use in deep learning is a concrete description of embodied experience. All agents discussed above are individuals, and their thoughts are not connected. Foundation models can bridge all individuals to share knowledge (ex: LLM-based multi-agent cooperation). Note that all agents mentioned in this proposal can grow up in their environment without supervision. This provides a large amount of annotated data without human effort. Data collected by each individual can be used to train foundation models. This approach corresponds to how we train LLM today. The LLM is trained with large-scale data from the Internet. And we created this data in our daily life. Each agent is an expert in their environment and tasks. But foundation models trained with everyone's experience can be loaded as a pre-trained model for all scenarios. Recent works like KOSMOS and 3D LLM, which treat everything (ex: image, location) as a token, provided a unified framework for data in any form.

Zheng, Zishuo
2023.10.10