Research Proposal

Abstract

In neuroscience, the cognitive map theory believes that our memory represents everything as a relative relationship. In the Hopfield network, memory retrieval is finished by pattern completion. Correspondingly, contrastive learning trains a network with data pairs. Autoregression achieves pattern completion in the temporal dimension. Here, I want to build an **embodied agent** without any supervision from scratch based on contrastive learning in the temporal dimension. This model is developed with the joint-embedding predictive architecture (JEPA) and the "World Model" mentioned by LeCun. The agent only depends on RGBD perception and proprioception. The agent will "grow up" like humans and learn everything gradually, including self-body control, imagination, semantic segmentation, and language. In "A Path Towards Autonomous Machine Intelligence," LeCun illustrated how an intelligence system should be. Here, I want to instantiate his idea into a detailed framework.

Model Selection

One essential ability of the 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 use time as an explicit or implicit parameter. Based on these observations, the model should depend on the temporal dynamic of the outer world and inner states. LeCun described JEPA as $pred(emb(x), z) = emb(y)$. But this article emphasis on the temporal perspective: $pred(emb(s_t), action) = emb(s_{t+1})$.

In recent years, Transformer become the dominant model in various tasks, including NLP, CV, and decision making. If we treat the embedding variables in JEPA as tokens and temporal prediction as auto-regression, Transformer is a perfect implementation of the JEPA framework. Another advantage of Transformer is multimodal perception. Evolution helps humans choose the most useful modality for survival. We integrate various modalities to reveal the underlying structure of this physical world. CNN and RNN are modality-specific and successful in small datasets. But Transformer has less inductive bias and success in large datasets. This provides a unified interface for multimodality data. KOSMOS, RT-2, and 3D-LLM treat everything as tokens (ex, text, image, location, cloud point, and action). Q-Former in BLIP-2 and CLIP align multimodal feature spaces. Moreover, the attention mechanism supports the viewpoint that "everything is a relationship" in the abovementioned cognitive map theory.

Embodied Simulator

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). Embodied simulators provide a low-cost playground for making mistakes. Another advantage of the embodied environment is multimodal perception. 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, MineDoJo, and SoundSpaces naturally bind multimodality data with specific objects.

Texts and APIs can be viewed as another kind of abstract simulator. The success of RHLF proves the importance of environmental feedback. This scenario has no embodied experience, but human thoughts constrain the model. "Embodied" is not the key but the action feedback. Agents must be constrained by rules, whether physical (embodied simulators), logical (APIs), or semantic (language).

Egocentric Control

Why do we have a concept of "myself" in our minds? We can control our bodies with a single thought. But we can't control the environment without actions. Our body is deterministic and fully observed, but the environment is highly dynamic and partially observed. 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. The agent must have a clear boundary between predictable behaviors and unreliable environments to get the desired outcome of actions. As humans, we learn how to control our bodies as infants. Then, we know how to influence the environment with actions. The agent can learn forward and inverse kinetics between body states and motion commands in simulators without supervision, ex: $FK(s_t^{inner}, cmd) = s_{t+1}^{inner}$.

Allocentric Representation

How do we learn physical rules and common sense? First, we need to extend egocentric cognition into allocentric representation. The forward dynamic model mentioned above is egocentric. But in our minds, we can image body motion without actually moving. The control rules of body joints are the same but extended from a first-person view into a third-person view. With imagination, we can reason and predict other one's moving trajectories.

This kind of prediction is not restricted to persons. Objects are also driven by forces. But how do we 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 without supervision. With object definition, we can learn physical rules and common sense from constant events. "Balls always fail on the ground" implies gravity.

Under this assumption, our predicting system can extend from inner states to world states by contrastive learning, 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.

Imagination as Generative Models

"What I cannot create, I do not understand." We learn this world by building a copy in our mind, and this copy is called “world model.” This model can be constructed by generative models. If we view actions as a latent variable z, generative models like GAN or VAE are equivalent to the imagination in our mind. Each generated sample 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 introduced 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.

Language & Abstraction

How do we develop language? Language is an abstraction of embodied experience and externalization of thoughts. The world model is virtual in our minds, and everyone has their own version of the world model. The language was 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 physical world. That is the foundation of the cross-language translation. We should consider how to bind an embodied agent's inner representation of the environment with human language. Aligning object-based scene graphs with word embeddings (tokens) space may help, ex: $similarity(distance(obj1, obj2), distance(token1, token2))$.

Bottom-Up Data Pipeline

Since super intelligence emerged in foundation models, how to generate large-scale datasets become essential. Many works leverage LLMs to label existing datasets to create new datasets (ex: “Visual Instruction Tuning”). However, language is highly abstract, with nonnegligible information loss. Training foundation models with language is not the optimal solution. We chose LLMs because of limited computing power and available datasets.

Existing foundation models are used top-down, and LLMs are applied in various downstream tasks. Here, I want to build foundation models from the bottom up. The training data we use in deep learning is a concrete description of embodied experience. Agents are individuals, and their thoughts are not connected. Foundation models can bridge individuals to share knowledge (ex, LLM-based multi-agent cooperation). Note that 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 pre-trained models for all scenarios.

Conclusion

Building human-level artificial intelligence is the ultimate goal of computer science. Although the research topics in the deep learning area have changed rapidly in recent years, the cognitive system of humans has remained the same. Each subfield in the deep learning area is trying to imitate one aspect of the cognitive system, from perception and reasoning to language. It is easy for practitioners to catch up with emerging technologies. But to make an outstanding contribution, an overall understanding of the whole framework is critical.

Zheng, Zishuo
2023.11.09

Reference

O'keefe, John, and Lynn Nadel. "Précis of O'Keefe & Nadel's The hippocampus as a cognitive map." Behavioral and Brain Sciences 2.4 (1979): 487-494.

LeCun, Yann. "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27." Open Review 62 (2022).

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Peng, Zhiliang, et al. "Kosmos-2: Grounding Multimodal Large Language Models to the World." arXiv preprint arXiv:2306.14824 (2023).

Brohan, Anthony, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." arXiv preprint arXiv:2307.15818 (2023).

Hong, Yining, et al. "3d-llm: Injecting the 3d world into large language models." arXiv preprint arXiv:2307.12981 (2023).

Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv preprint arXiv:2301.12597 (2023).

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Girdhar, Rohit, et al. "Imagebind: One embedding space to bind them all." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Gan, Chuang, et al. "Threedworld: A platform for interactive multi-modal physical simulation." arXiv preprint arXiv:2007.04954 (2020).

Fan, Linxi, et al. "Minedojo: Building open-ended embodied agents with internet-scale knowledge." Advances in Neural Information Processing Systems 35 (2022): 18343-18362.

Chen, Changan, et al. "Soundspaces 2.0: A simulation platform for visual-acoustic learning." Advances in Neural Information Processing Systems 35 (2022): 8896-8911.

Ziegler, Daniel M., et al. "Fine-tuning language models from human preferences." arXiv preprint arXiv:1909.08593 (2019).

Gadre, Samir Yitzhak, Kiana Ehsani, and Shuran Song. "Act the part: Learning interaction strategies for articulated object part discovery." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Goodfellow, Ian, et al. "Generative adversarial networks." Communications of the ACM 63.11 (2020): 139-144.

Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

Baker, Bowen, et al. "Video pretraining (vpt): Learning to act by watching unlabeled online videos." Advances in Neural Information Processing Systems 35 (2022): 24639-24654.

Dai, Yilun, et al. "Learning universal policies via text-guided video generation." arXiv preprint arXiv:2302.00111 (2023).

Sieb, Maximilian, et al. "Graph-structured visual imitation." Conference on Robot Learning. PMLR, 2020.

Liu, Haotian, et al. "Visual instruction tuning." arXiv preprint arXiv:2304.08485 (2023).

Zhang, Hongxin, et al. "Building cooperative embodied agents modularly with large language models." arXiv preprint arXiv:2307.02485 (2023).