Learning From Online and Physically Grounded Visual Experience

Overview

In the future, applications in robotics and personal assistance require machines to learn and adapt based on individual visual experiences. However, existing machine learning algorithms have limitations in their ability to learn new concepts and skills in real-time or transfer knowledge to new situations. To overcome this challenge, we propose a research initiative aimed at creating a learning system that can learn visual knowledge and representations from physically grounded experience.

Recent advancements in self-supervised learning have demonstrated impressive visual learning capabilities without the need for human annotations. However, these methods only focus on object-centric static images from the internet. By contrast, our proposed project aims to learn visual representations from an online video stream, similar to the learning environment of real world agents like ourselves.

The proposed system consists of three research thrusts: 1) 3D motion perception to ground visual representations in the physical world; 2) detection and clustering of high-level temporal event concepts; and 3) memorization and consolidation of visual experiences. Our system will be evaluated using egocentric videos and embodied environments, and applied to real-world vision tasks.

Intellectual Merit

The proposed research aims to deepen our understanding of learning in embodied agents. Currently, large scale models can only run in large data centers since learning requires a tremendous amount of data collected from the internet. The learning process is in stark contrast to how we acquire our visual knowledge.

Our proposed subareas of 3D motion perception, temporal prediction, and visual memory consolidation are central to human vision and memory but under-explored in the current machine learning and computer vision literature. Our proposed system is poised to lay the foundation for embodied perception and learning in future AI. Our projects have the potential of addressing the following research questions: 1) What is the role of 3D and motion perception in high-level visual learning? 2) Does visual memory alleviate forgetting in an online environment? 3) How robust is online learning dealing with real world distribution shift? 4) Can real world videos make visual learning more data efficient?

Our engineering-focused approach also holds the potential to offer new insights into the principles of intelligence in human cognition. If we can develop a machine learning algorithm that can learn from personalized experiences and make use of information in embodied environments in a manner similar to humans, it is likely that the resulting system will share functional similarities to the human learning process.

Broader Impacts

The proposed research has potential applications in a variety of consumer products, such as augmented reality, home robotics, and personal assistive technologies. By overcoming the current data and computation constraints and bring learning online, our research has the potential to democratize AI and make AI more accessible to people living in a different environment than the original training data.

Furthermore, by learning from personalized visual experiences, the proposed research may help address privacy concerns associated with uploading user data to servers for training purposes. A deeper understanding of the principles of learning and memory consolidation may also inform the development of future educational technologies.