Lec22_rawScript

Today we are going to talk about foundational models for decision making, which I think is one of the most exciting areas of research in robotics right now. Foundational models have been a buzzword for a while since ChatGPT, and many of you have tried ChatGPT, so you already have some knowledge about what we are going to discuss today. However, today's talk will extend beyond just ChatGPT, and you will see this by the end of the lecture. I have prepared a lot of slides, and hopefully, we can cover them all. Our agenda includes three major topics: large language models and their application in embodied AI or game-playing agents, the challenges and opportunities in foundational models for decision making, and Vision Foundation models, which are lesser-known.

In the first section on task planning with large language models, subtitled "Connecting Unstructured Words with Structured Algorithms," we'll explore how this applies in everyday life. Imagine it's the 22nd century, and you have a general-purpose robot. You might want to tell this robot to help clean up a drink you've spilled or bring you a snack after a workout, expecting it to understand and act on these free-form instructions just like a human would. However, we're still in the 21st century, and current robot algorithms are quite different from what we might expect in the future. Currently, you need to specify many things manually, such as what to segment in a scene for a robot to pick up an object, or plan a grasp using calculated surface normals. These examples show how structured current algorithms are compared to the unstructured nature of human instructions.

Large language models, like ChatGPT, offer a way to bridge this gap by providing structured outputs from unstructured inputs. They can suggest a series of actions, like setting a can upright or picking up a napkin, based on the input command to clean up a spilled Coke. This capability is possible because humans use language as task abstractions and plans. The internet, filled with text-based human knowledge, provides a vast resource that large language models can learn from to assist in task planning.

For those unfamiliar with how large language models work, imagine a model predicting the next word in a sentence based on the context it has seen, such as predicting the word "books" after "The students opened their." These models provide a probability distribution over possible next words, which can be used to generate text or evaluate the likelihood of given words fitting into the context. This process illustrates the core functionality of language models.

Language models can be particularly powerful in robotics planning. For example, asking a model to list what's needed to make a coffee can result in it listing relevant items like coffee beans and water, which are useful for planning structured tasks in robotics. Furthermore, these models can output highly detailed instructions for tasks, demonstrating their ability to capture and utilize extensive human knowledge.

However, integrating these capabilities with actual robotic systems involves aligning unstructured language model outputs with structured robotic algorithms. This can be done by associating text descriptions with executable robot skills, selecting from these skills based on the model's predictions, and structuring the model's outputs to fit predefined templates that facilitate easier parsing and execution by robotic systems.

One real-world application we'll examine involves a Google paper where language models were combined with robotic affordances to better align model predictions with actual robot capabilities. For instance, if a robot only sees an apple but the model suggests picking up various fruits, the robot's action choices need to consider what's physically present and feasible.

We'll also explore how my work and others have extended these concepts by integrating visual data with language models to enhance robot understanding and interaction with their environment. This involves using techniques like unsupervised exploration and object recognition to improve the flexibility and applicability of robotic systems in real-world environments.

These advancements point to a future where robots can understand and execute tasks with a level of flexibility and intelligence closer to that of humans, significantly powered by the evolving capabilities of large language models.