2024-2025 SPRING SEMESTER

Introduction to Robotics

-EE422 / CS421-

VISION LANGUAGE MODELS FOR ROBOTIC PERCEPTION

Week-1

AGAH KUTAY ORUC

EMIN HAN CELIK

TIMUR GULMEZ

MELIH EGREK

INTRODUCTION

Vision-language-action models (VLAs) are a class of multimodal models within the field of embodied AI, designed to process information from vision, language, and action modalities. Embodied AI requires controlling physical embodiments that interact with the environment. The policy must possess the capability to understand language instructions, visually perceive the environment, and generate appropriate actions. They are well-suited not only for controlled settings like factories but also for everyday tasks in household environments. Earlier AI models were simpler, focusing only on one type of data like images or text. For example, AlexNet focused on images, while GPT focused on text. Reinforcement learning models, like the Deep Q-network, learned from rewards. Modern VLA models combine these abilities to perform tasks involving vision, language, and actions.

1.Background

1.1 Unimodal Models

Before VLAs, most AI models focused only on one type of data (unimodal), such as images, text, or actions. These were the starting points for today's multimodal (multiple data types) VLA models.

1.2 Vision-Language Models(VLMSs)

VLMs combine vision and language abilities, such as describing an image or answering questions about it. Early examples include models like ViLBERT and CLIP. Recent models like Flamingo and LLaVA show how powerful these combinations can be.

1.3 Embodied AI & Robot Learning

Embodied AI interacts with the real world, unlike conversational AI, which only deals with text. Robot learning usually involves reinforcement learning (learning from trial and error). Sometimes, imitation learning is used, where robots learn directly by copying human actions.

2.Components of VLA

Vision-Language-Action (VLA) models combine multiple technologies to help robots understand their environment, better interact with humans, and succeed at tasks. Reinforcement learning (RL) and human feedback accelerate robots' learning from experience, while pre-trained visual representations (PVRs) and video representations help them perceive their environment more clearly. Dynamic learning allows them to predict the consequences of actions, while world models enable robots to create imaginary environments and generate risk-free solutions. Large language models (LLMs) facilitate task planning by using common sense and language-based reasoning in a human context. Visual world models and reasoning mechanisms such as chained thinking enable robots to solve more complex situations more effectively. All of these components aim to make robots smarter, more agile, and more compatible with humans.

3. Low-level Control Policies

Low-level control policies help robots follow instructions by combining modules like action decoders (deciding actions), vision encoders (understanding images), and language encoders (understanding text instructions). These parts form Vision-Language-Action (VLA) models, helping robots perform actions based on given instructions. Here are some various approaches to designing low-level control policies; Non-Transformer Control Policies,Non-Transformer Control Policies,Control Policies for Multimodal Instructions,Control Policies with 3D Vision,Diffusion-based Control Policies,Diffusion-based Control Policies with 3D Vision,Control Policies for Motion Planning,Control Policies with Point-based Action,Large VLA Pros: Different architectures are used in VLA models to combine vision and language information. Cross-attention works well with small models but is complex. FiLM, used by RT-1, effectively conditions actions based on instructions. Concatenation is simpler and works effectively with larger models.

4. Task Planners

In robotic systems, different planning methods are applied to ensure a robot can successfully complete complex tasks. These methods are optimized based on specific applications and system requirements. Below, the planning models are ranked in order of most to least frequently used.

4.1 End-to-End Planning

This method directly makes decisions using sensor data and deep learning models. It is the most widely used approach, particularly in applications like autonomous driving systems.

4.2 Modular Planning

Tasks are broken down into smaller, manageable subtasks, making execution more efficient. This planning method is commonly used in warehouse automation and industrial robotics.

4.3 D Visual Planning

Robots utilize visual perception to map their surroundings in three dimensions, making this approach widely used in logistics and navigation robotics.

4.4 Language-Based Planning

This method interprets natural language commands from users and translates them into appropriate robotic actions. It is frequently applied in smart home systems and human-robot interactions.

4.5 Code-Based Planning

Robots generate their own motion scripts through automated coding. This technique is primarily used in autonomous software systems and AI-driven robotic development.

4.6 Grounded Planning

Robots make real-time decisions based on direct environmental feedback. This method is mainly implemented in specialized applications, such as Mars rovers and search-and-rescue robots.

4.7 Monolithic Planning

Based on predefined rules, this method offers limited flexibility. It is primarily found in older generations of industrial robots and legacy automation systems.

These planning methods can be adapted for various applications and are often combined to enhance efficiency and flexibility.