ERI 2023 - Embodied AI Reference Implementation

Rene Schulte
Reply DE
r.schulte@reply.com

Maccagni Giacomo
Machine Learning Reply
g.maccagni@reply.com

Federico Minutoli
Machine Learning Reply
f.minutoli@reply.com

Riccardo Zuppetti
Machine Learning Reply
r.zuppetti@reply.com

Simone Voto
Concept Engineering Reply
s.voto@reply.com

Samuele Giannetto
Concept Engineering Reply
s.giannetto@reply.com

Figure 1: Reply Xchange Milan 2023 - Demo Area

Abstract

This work is related to the implementation of the most recent advances in embodied AI from Google and Meta on Boston Dynamics Spot Robot. The ultimate goal of the work is to develop a long-horizon planning coordinator who coordinates Spot robot to complete long-horizon tasks based on their short-horizon skills. The context will be that of navigation and mobile manipulation, i.e., the robots will be asked to fetch objects and move them to other points in a pre-set area. All this will pass through the short-term skills built-in in the Spot SDK, Meta's recent model for artificial visual cortex (VC-1), and Google's vision-language (VLM) model for visual-language navigation for give long-horizon tasks to robots through natural language prompts in the area. Find the work here `https://marino-multipla.github.io/cop/portfolio/eri2023.html`

Keywords: EmbodiedAI Agent, GroundingDINO, Visual Language Map, Quadruped Robot, Natural Language Control

1 Introduction

At Reply, we harness visual representations to enable the Spot robot to understand the environment and perform complex tasks like navigation and object manipulation with minimal training, enhancing human-robot interaction. This enables the control of the AI agents using natural language and voice commands, eliminating the need for complex model management. Spot's interaction begins with converting human commands spoken in natural language and voice into text through the Speech-to-Text phase, a crucial step for enabling seamless communication. The natural language text is then subjected to Task Processing, where subtasks are extracted, enabling Spot to gain a more comprehensive understanding of the user's intent. Spot's capabilities extend to Navigation Tasks, facilitated by the use of Vision Language Maps (VLMaps) from Google. These maps provide Spot with a semantic understanding of its environment, assisting in tasks such as autonomous exploration and mapping. In Manipulation Tasks, Spot employs two distinct AI models: Grounding DINO for object detection and Visual Cortex 1 for effective manipulation. DINO plays a pivotal role in accurately detecting and locating objects within Spot's surroundings, instead, Visual Cortex 1 enhances Spot's ability to interact with objects, ensuring precision and effectiveness, particularly in tasks like pick-and-place operations.

2 Related Work

Below are the works that inspired our implementation.

2.1 CortexBench

This study presents CortexBench, an extensive evaluation of pre-trained visual representations (PVRs) for Embodied AI. It includes 17 tasks across locomotion, navigation, dexterity, and mobile manipulation. Despite no universally dominant PVR, the research explores the impact of pre-training data scale and diversity using over 4,000 hours of egocentric videos. Interestingly, scaling dataset size and diversity doesn't universally enhance performance. The largest model, VC-1, surpasses prior PVRs on average but lacks universal dominance. However, when adapted for task-specific domains, VC-1 demonstrates substantial performance gains, outperforming known benchmarks on all CortexBench tasks. VC-1 models, requiring over 10,000 GPU-hours to train, are accessible on the researchers' website for the benefit of the research community.

`https://eai-vc.github.io/`

`https://arxiv.org/pdf/2303.18240.pdf`

2.2 Visual Language Maps for Robot Navigation

This research proposes VLMaps, a novel spatial mapping approach that combines pretrained visual-language features with 3D reconstructions of the physical environment. In contrast to off-the-shelf visual-language models, VLMaps autonomously build maps from robot video feeds, allowing natural language indexing without additional labeled data. When integrated with large language models (LLMs), VLMaps enable the translation of natural language commands into spatial navigation goals with detailed instructions, such as positioning between furniture or specifying distances. Moreover, VLMaps can be shared among diverse robots, generating obstacle maps on-the-fly. Experiments in both simulated and real-world settings demonstrate that VLMaps significantly improve navigation based on complex language instructions compared to existing methods.
`https://vlmaps.github.io/`

`https://arxiv.org/pdf/2210.05714.pdf`

2.3 ASC: Adaptive Skill Coordination for Robotic Mobile Manipulation

The paper introduces Adaptive Skill Coordination (ASC) for long-horizon tasks like mobile pick-and-place. ASC has three components: a library of basic visuomotor skills, a skill coordination policy choosing when to use each skill, and a corrective policy adapting skills in out-of-distribution states. ASC relies on onboard sensing, eliminating the need for detailed maps or precise object locations, facilitating real-world deployment. Trained in simulated indoor environments, ASC is deployed zero-shot on the Boston Dynamics Spot robot in real-world settings. Perturbation experiments show ASC's robustness to errors, layout changes, dynamic obstacles, and disturbances.
`https://adaptiveskillcoordination.github.io/`

`https://arxiv.org/pdf/2304.00410.pdf`

2.4 R3M: A Universal Visual Representation for Robot Manipulation

This research explores the efficacy of visual representations pre-trained on diverse human video data for enhancing data-efficient learning in robotic manipulation tasks. Using the Ego4D human video dataset, the study employs time-contrastive learning, video-language alignment, and an L1 penalty to create a visual representation called R3M. R3M, when utilized as a frozen perception module for downstream policy learning, demonstrates notable improvements in task success. Across 12 simulated robot manipulation tasks, R3M outperforms training from scratch and surpasses state-of-the-art representations like CLIP and MoCo by over 10 per cent. In real-world experiments, R3M enables a Franka Emika Panda arm to learn various manipulation tasks with just 20 demonstrations in a cluttered apartment setting.
`https://tinyurl.com/robotr3m`

`https://arxiv.org/pdf/2203.12601.pdf`

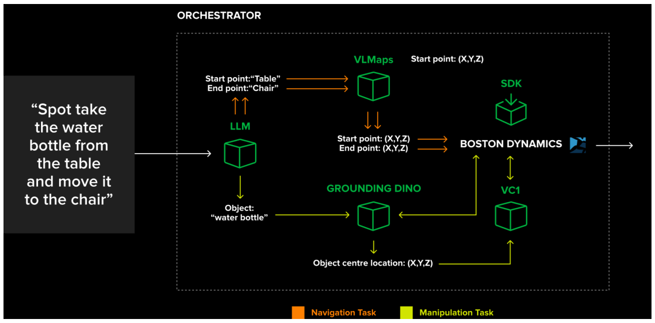

Figure 2: Architecture

Figure 3: Area42 Laboratory

2.5 Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection

This paper introduces Grounding DINO, an open-set object detector that combines Transformer-based detector DINO with grounded pre-training. It incorporates language into a closed-set detector for open-set concept generalization, utilizing a feature enhancer, language-guided query selection, and a cross-modality decoder for fusion. Grounding DINO performs well across benchmarks, achieving a 52.5 AP on COCO's zero-shot transfer and setting a record on ODinW with a mean of 63.0 AP after fine-tuning with COCO data.
`https://github.com/IDEA-Research/GroundingDINO`

`https://arxiv.org/pdf/2303.05499.pdf`

3 Implementation

3.1 Architecture

The implementation is based on several integrations. The user sends the command to the robot using their voice. The

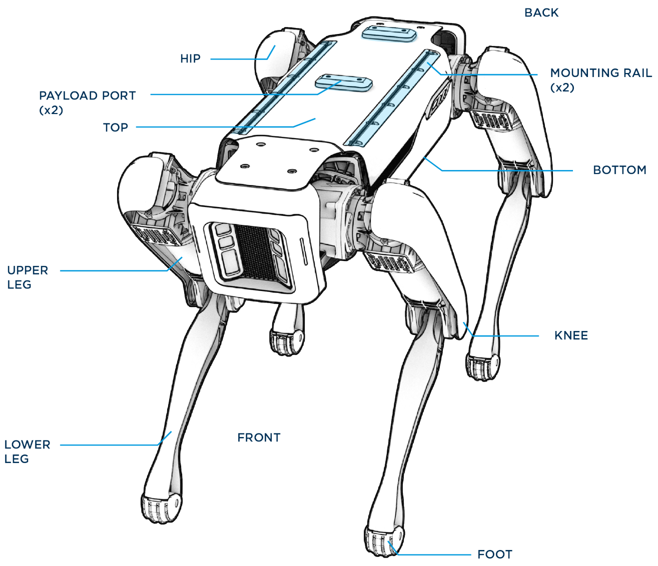

Figure 4: Boston Dynamics SPOT Anatomy

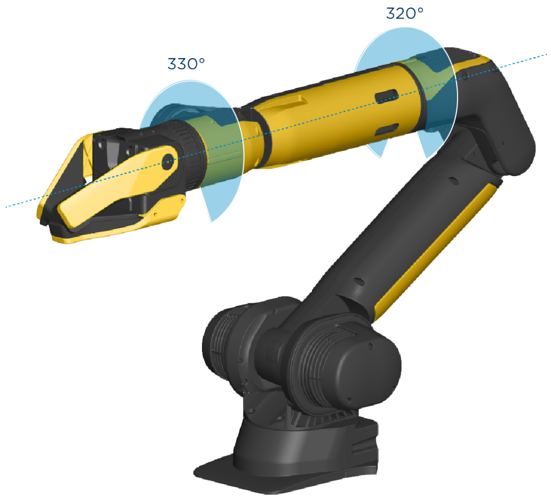

Figure 5: SPOT Arm

voice is decoded into text using a Speech to Text model, in this case OpenAI's Whisper. After which the text is processed by an LLM with gpt-3.5-turbo model to understand the natural language and temporal correlations of the commands to derive the low-level tasks that the robotic agent will have to perform. The main tasks are: Navigation, grasping, placing. Navigation tasks will be managed both by Bosotn Dynamics' GraphNavMap and by the VLMaps model. The manipulation tasks will be handled by both the Boston Dynamics SDK and the model in Visual Representation R3M. You find here 2 all the main components involved in the implementation.

3.2 Agent

We used the most advanced quadruped robot on the market, also equipped with a mechanical arm to be able to carry out object manipulation tasks. You find here 4 the main anatomy of the robot.

3.3 Laboratory

The development of the solution took place in Area42, which is our laboratory located in Turin. In the lab there are more than 9 workbenches with variable settings of both surrounding objects and light and passage conditions.

4 Future developments

This reference implementation lays the foundations for the development of a silent-agent orchestrator, with robotic agents of different nature. Another line of research is that relating to the full use of navigation and object manipulation models, completely bypassing the factory libraries provided by Boston Dynamics