Abstract

- (Vision language action model) Express actions as text tokens

Introduction

- Previous approaches Generally addresses higher level aspect of planning
- Previous approaches use vision model or language model pre-training but not vision language model pre-training
- This work does not restrict action space to 2D space and does not require a calibrated camera

Goal

- Boost generalization and emergent semantic reasoning in end to end robot learning using vision language action (vlm) models

New terms

- Chain of thought reasoning (multistage semantic reasoning)
- Multimodal sentences
- Clip model(representation learning that learns both modalities)
- Visual language model (vision, texts to free form text)
- Symbol running
- Vc-1
- R3M
- Moo

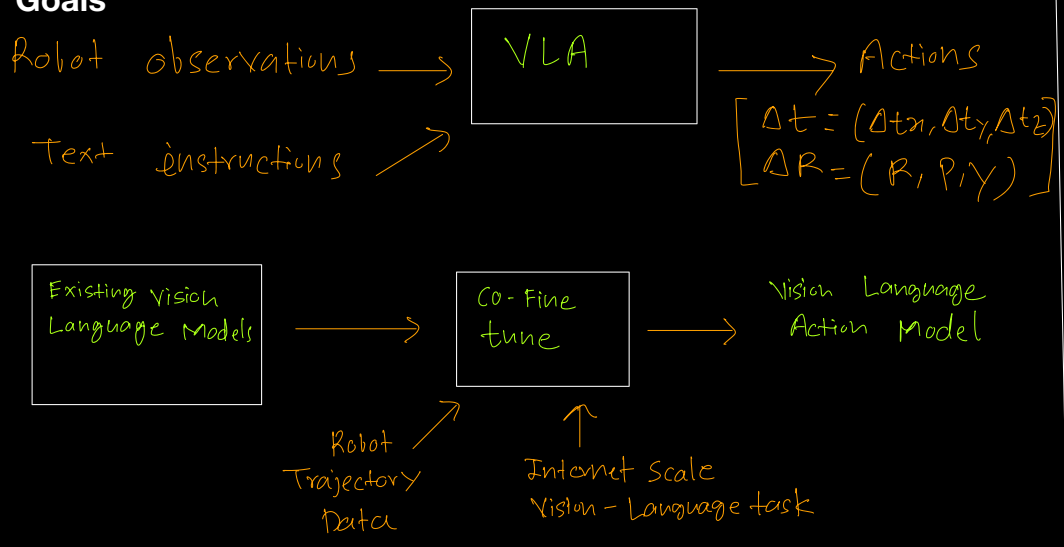
Technology used

- Pali -x
- Palm-E
- Open-source language table simulation environment by lynch

Questions

- **Is more data helping here or the architecture of vision language action model is the reason for better performance than rt-1**

Goals



Hypothesis

This Internet - scale training would improve generalization and enable emergent semantic reasoning

In previous approaches the low level controller did not benefit from the rich semantic knowledge of internet scale models during training