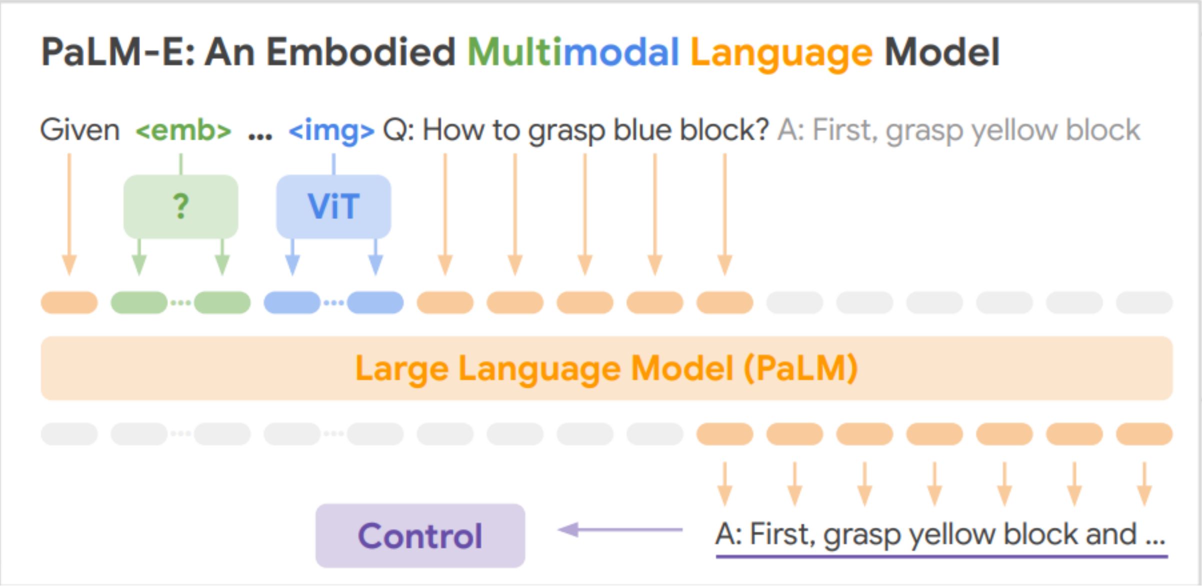
PaLM-E: An Embodied Multimodal Language Model

Figure 1
Abstract
Introduction
PaLM-E: An Embodied Multimodal Language Model
Conclusion

Figure 1


PaLM-E: An Embodied Multimodal Language Model
Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block
?
ViT
Large Language Model (PaLM)
Control
A: First, grasp yellow block and ...

PaLM-E is a single general-purpose multimodal language model for embodied reasoning tasks, visual-language tasks, and language tasks. PaLM-E transfers knowledge from visual-language domains into embodied reasoning – from robot planning in environments with complex dynamics and physical constraints, to answering questions about the observable world. PaLM-E operates on multimodal sentences, i.e. sequences of tokens where inputs from arbitrary modalities (e.g. images, neural 3D representations, or states, in green and blue) are inserted alongside text tokens (in orange) as input to an LLM, trained end-to-end.

Abstract

Large language models have been demonstrated to perform complex tasks. However, enabling general inference in the real world, e.g. for robotics problems, raises the challenge of grounding. We propose embodied language models to directly incorporate real-world continuous sensor modalities into language models and thereby establish the link between words and percepts. Input to our embodied language model are multi-modal sentences that interleave visual, continuous state estimation, and textual input encodings. We train these encodings end-to-end, in conjunction with a pre-trained large language model, for multiple embodied tasks including sequential robotic manipulation planning, visual question answering, and captioning. Our evaluations show that PaLM-E, a single large embodied multimodal model, can address a variety of embodied reasoning tasks, from a variety of observation modalities, on multiple embodiments, and further, exhibits positive transfer: the model benefits from diverse joint training across internet-scale language, vision, and visual-language domains. Our largest model, PaLM-E-562B with 562B parameters, in addition to being trained on robotics tasks, is a visual-language generalist with state-of-the-art performance on OK-VQA, and retains generalist language capabilities with increasing scale.

Introduction

Large language models (LLMs) demonstrate strong reasoning capabilities across various domains, including dialogue (Glaese et al., 2022; Thoppilan et al., 2022), step-by-step reasoning (Wei et al., 2022; Kojima et al., 2022), math problem solving (Lewkowycz et al., 2022; Polu et al., 2022), and code writing (Chen et al., 2021a). However, a limitation of such models for inference in the real world is the issue of grounding: while training LLMs on massive textual data may lead to representations that relate to our physical world, connecting those representations to real-world visual and physical sensor modalities is essential to solving a wider range of grounded real-world problems in computer vision and robotics (Tellex et al., 2020). Previous work (Ahn et al., 2022) interfaces the output of LLMs with learned robotic policies and affordance functions to make decisions, but is limited in that the LLM itself is only provided with textual input, which is insufficient for many tasks where the geometric configuration of the scene is important. Further, in our experiments we show that current state-of-the-art visual- language models trained on typical vision-language tasks such as visual-question-answering (VQA) cannot directly solve robotic reasoning tasks.

In this paper we propose embodied language models, which directly incorporate continuous inputs from sensor modalities of an embodied agent and thereby enable the language model itself to make more grounded inferences for sequential decision making in the real world. Inputs such as images and state estimates are embedded into the same latent embedding as language tokens and processed by the self-attention layers of a Transformer-based LLM in the same way as text. We start from a pre-trained LLM in which we inject the continuous inputs through an encoder. These encoders are trained end-to-end to output sequential decisions in terms of natural text that can be interpreted by the embodied agent by conditioning low-level policies or give an answer to an embodied question. We evaluate the approach in a variety of settings, comparing different input representations (e.g. standard vs. object-centric ViT encodings for visual input), freezing vs. finetuning the language model while training the encoders, and investigating whether co-training on multiple tasks enables transfer.

To investigate the approach's breadth, we evaluate on three robotic manipulation domains (two of which are closed- loop in the real-world), standard visual-language tasks such as VQA and image captioning, as well as language tasks. Our results indicate that multi-task training improves performance compared to training models on individual tasks. We show that this transfer across tasks can lead to high data-efficiency for robotics tasks, e.g. significantly increasing learning success from handfuls of training examples, and even demonstrating one-shot or zero-shot generalization to novel combinations of objects or unseen objects.

We scale PaLM-E up to 562B parameters, integrating the 540B PaLM (Chowdhery et al., 2022) LLM and the 22B Vision Transformer (ViT) (Dehghani et al., 2023) into, to our knowledge, the largest vision-language model currently reported. PaLM-E-562B achieves state-of-the-art performance on the OK-VQA (Marino et al., 2019) benchmark, without relying on task-specific finetuning. Although not the focus of our experimentation, we also find (Fig. 2) that PaLM-E-562B exhibits a wide array of capabilities including zero-shot multimodal chain-of-thought (CoT) reasoning, few-shot prompting, OCR-free math reasoning, and multi-image reasoning, despite being trained on only single-image examples. Zero-shot CoT (Kojima et al., 2022), originally a language-only concept, has been shown on multimodal data with task-specific programs (Zeng et al., 2022) but to our knowledge, not via an end-to-end model.

To summarize our main contributions, we (1) propose and demonstrate that a generalist, transfer-learned, multi-embodiment decision-making agent can be trained via mixing in embodied data into the training of a multimodal large language model. We show that, (2) while current state-of-the-art general-purpose visual-language models out-of-the box (zero-shot) do not well address embodied reasoning problems, it is possible to train a competent general-purpose visual-language model that is also an efficient embodied reasoner. In studying how to best train such models, we (3) introduce novel architectural ideas such as neural scene representations and entity-labeling multimodal tokens. Finally, in addition to our focus on PaLM-E as an embodied reasoner we (4) show that PaLM-E is also a quantitatively competent vision and language generalist, and (5) demonstrate that scaling the language model size enables multimodal finetuning with less catastrophic forgetting.

PaLM-E: An Embodied Multimodal Language Model

The main architectural idea of PaLM-E is to inject continuous, embodied observations such as images, state estimates, or other sensor modalities into the language embedding space of a pre-trained language model. This is realized by encoding the continuous observations into a sequence of vectors with the same dimension as the embedding space of the language tokens. The continuous information is hence injected into the language model in an analogous way to language tokens. PaLM-E is a decoder-only LLM that generates textual completions auto-regressively given a prefix or prompt. We call our model PaLM-E, since we use **PaLM** (Chowdhery et al., 2022) as the pre-trained language model, and make it **Embodied**.

The inputs to PaLM-E consist of text and (multiple) continuous observations. The multimodal tokens corresponding to these observations are interleaved with the text to form multi-modal sentences. An example of such a multi-modal sentence is `Q: What happened between <img_1> and <img_2>?` where represents an embedding of an image. The output of PaLM-E is text generated auto-regressively by the model, which could be an answer to a question, or a sequence of decisions produced by PaLM-E in textual form that should be executed by a robot. When PaLM-E is tasked with producing decisions or plans, we assume that there exists a low-level policy or planner that can translate these decisions into low-level actions. Prior work has discussed a variety of ways to train such low-level policies (Lynch & Sermanet, 2020; Brohan et al., 2022), and we use these prior methods directly without modification. In the following, we describe our approach more formally.

Conclusion

We proposed to build an embodied language model by injecting multi-modal information such as images into the embedding space of a pre-trained LLM. Experiments showed that off-the-shelf state-of-the-art vision-language models trained on general VQA and captioning tasks are not sufficient for embodied reasoning tasks, as well as limitations of a recent proposal for grounding language models through affordances. To overcome these limitations, we proposed PaLM-E, a single model that is able to control different robots in simulation and in the real world, while at the same time being quantitatively competent at general VQA and captioning tasks. In particular the novel architectural idea of ingesting neural scene representations (i.e., OSRT) into the model is particularly effective, even without large-scale data. PaLM-E is trained on a mixture of diverse tasks across multiple robot embodiments as well as general vision-language tasks. Importantly, we have demonstrated that this diverse training leads to several avenues of transfer from the vision language domains into embodied decision making, enabling robot planning tasks to be achieved data efficiently. While our results indicate that frozen language models are a viable
path towards general-purpose embodied multimodal models that fully retain their language capabilities, we have also surfaced an alternative route with unfrozen models: scaling up the language model size leads to significantly less catastrophic forgetting while becoming an embodied agent. Our largest model, PaLM-E-562B, showcases emergent capabilities like multimodal chain of thought reasoning, and the ability to reason over multiple images, despite being trained on only single-image prompts.