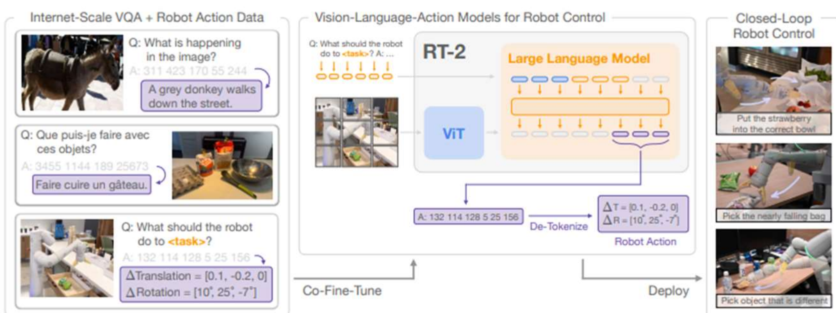
RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

Paper Review

Review by:
Tyler Kim (tkj9ep)
tkj9ep@virginia.edu

Summary

Brohan et al's, from Google DeepMind, paper *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotics Control* explore using pretrained Vision-Language Models (VLM) to train robots to perform specific tasks by **tokenizing robotic actions into text tokens** and creating "multimodal sentences". They name these models as **Vision-Language-Action Models (VLA)** and create their own instance of this **family of models** called **RT-2**. Previous attempts to use LLMs and VLMs for training robots encounter multiple obstacles such as **translating outputs** to low-level robotic actions, **rebuilding and retraining** the model, and sometimes the **lack of data**. Although previous attempts were relatively successful, Brohan et al argue that greater benefits for robots come from using Internet-scale models. Hence, the development of VLA models and RT-2. The main contribution of the paper is **RT-2** which is a family of fine-tuned VLM that uses **web-scale data to provide better generalization abilities**.



The figure above was taken from Figure 1 of the paper which details the overview of RT-2.

Brohan et al's *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotics Control* use PaLI-X from Chen et al and PaLM-E from Driess et al as the VLMs that will act as VLA models. For the VLMs to output actions, Brohan's team **encode actions on the discretization** proposed by Brohan et al for RT-1 model where the action space has 6-DoF positional and rotational displacement of the robot end-effector, gripper level extensions, and a terminate command. The continuous dimensions are uniformly discretized into 256 bins where an action is represented using **8 integer numbers**. Action tokens are chained with space characters to create an action vector. Brohan's team *co-fine-tunes* the robotics data using the action vector to convert the robot data into VQA format for the input, a **string of tokens to represent an action**, and **balancing** the number of **robot vs web data** in each training batch.

$$\text{"terminate } \Delta\text{pos}_x \ \ \Delta\text{pos}_y \ \ \Delta\text{pos}_z \ \ \Delta\text{rot}_x \ \ \Delta\text{rot}_y \ \ \Delta\text{rot}_z \ \ \text{gripper\_extension".}$$

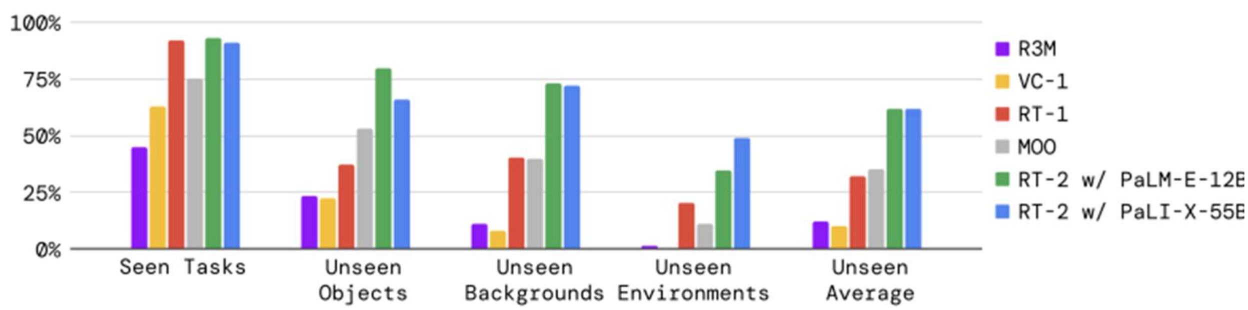
The figure above was taken from the paper describing action vectors.

The experiments test the RT-2 model based on **PaLI-X (RT-2-PaLI-X)** and RT-2 model based on **PaLM-E (RT-2-PaLM-E)** using around 6000 trajectories, original web scale data from Chen et al and Driess et al for training, and aims to answer four main questions as quoted from the paper:

1. "How does RT-2 perform on seen task and generalize over new objects, backgrounds, and environments?"
2. "Can we observe and measure any emergent capabilities of RT-2?"
3. "How does the generalization vary with parameter count and other design decisions?"
4. "Can RT-2 exhibit signs of chain-of-thought reasoning similarly to vision-language models?"

The paper used **RT-1** from Brohan et al, **VC-1** from Majumdar et al, **R3M** from Nair et al, and **MOO** from Stone et al as **baselines**.

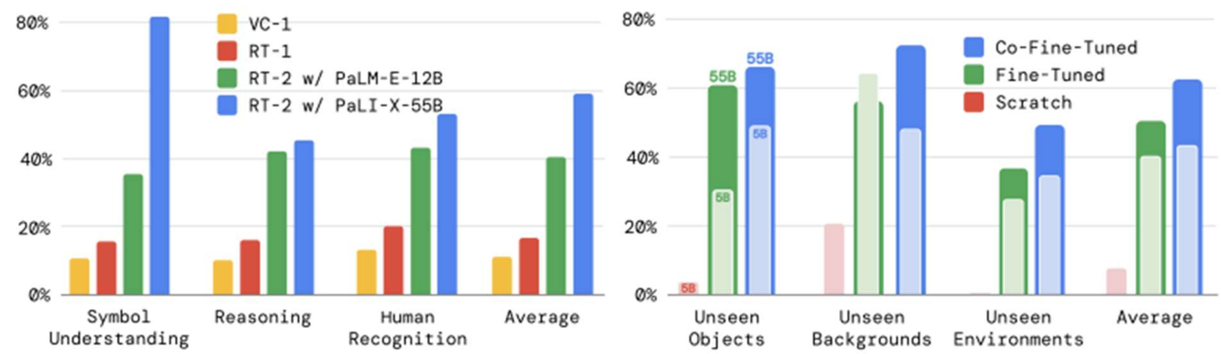
For the experiment for the first question, the paper compared the two RT-2 models and the baselines with tasks that have both *seen* and *unseen* categories such as objects, backgrounds, and environments and split the tasks by difficulty. The paper reports that both RT-2 models performed similarly with each other but performed **about 2x better than RT-1 and MOO** and **about 6x better than other baselines** which indicates that RT-2 generalized better than other approaches.



The figure above taken from Figure 4 of the paper shows the results of the first experiment of the paper.

For the experiment for the second question, Brohan's team aims to test whether RT-2 has inherited or *emergent* capabilities, new capabilities that emerge from transferring Internet-scale pretraining. They test all six models with **nuanced tasks**, such as "put strawberry into correct bowl" or "pick up the bag about to fall off the table" and split the capabilities into three categories: symbol understanding, reasoning, and human recognition. The paper reports that both VLA models **performed significantly better than all the baselines** which RT-2-PaLI-X achieved better in symbol understanding while RT-2-PaLM-E performed better in math reasoning. The paper argues that the RT-2 **does indeed have emergent capabilities** and can be measured.

Regarding the experiment for the third question, Brohan's team tests a **5B** parameter and **55B** parameter RT-2-PaLI-X model as well as **three training approaches**: training from scratch, fine-tuning pretrained model with just robot action, and co-fine-tuning. The paper reports that training from scratch results in poor performance despite size and co-fine-tuning generalized better than regular fine-tuning regardless of the size of the model. In other words, **size does not seem to make a significant difference but tuning does**.



(a) Performance comparison on various emergent skill evaluations (Figure 8) between RT-2 and two baselines.

(b) Ablations of RT-2-PaLI-X showcasing the impact of parameter count and training strategy on generalization.

The figure above is taken from Figure 6 of the paper which highlights the results of the second experiment and third experiment.

For the final experiment, the team **augments** the data to add a “Plan” step followed by actual action tokens and fine-tune a variant of RT-2 with PaLM-E. The paper reports that the robot is indeed **able to have a chain-of-thought**.

The figure was taken from Figure 7 of the paper detailing the chain-of-thought reasoning.

Finally, the paper describes limitations such as **lack** of ability for the robot to **perform new motions** and **computation costs**. Ultimately, the team believes the generalization ability of this approach leads to promising new approaches for robotics.

Strengths

One strength is that the paper was well-written. It was **easy to follow** and made sense. Another strength of the paper was that the **figures were well-designed**, and I could easily understand what the figures were displaying and their **purpose**. More specifically, all graphs for experiment results were **easy to read** and I could **pick up** on what the author was trying to **communicate quickly**. In addition, the paper provides a **new perspective** or approach to **training robots**. That is, using **pretrained models** to train the models as opposed to the traditional view **of training models from scratch**. Another value that the paper provides is a **different technique for representing actions**. In the paper, they propose **discretization** of actions for discrete and continuous actions using **bins** and using **8 integer representations to represent an action**. Finally, the paper provided a new perspective of fine-tuning a model using **co-fine-tuning** where the model is **tuned and trained simultaneously**.

In addition to the strengths of the paper, there is quite a bit I learned from it ranging from basic robotic concepts to using specific techniques for training the model. One thing I learned was a way to **transform robot actions** into something that a **LLM could read** through **tokenization**. Another interesting thing I learned was how one could use a **VLM** to **train a robot** to perform certain tasks, more generally, a **multimodal model for training**. Finally, which is the purpose of the paper, how to use a **pretrained VLM** to train a robot to do certain tasks. Other small things I learned were what an **end-effector** is and what **Degrees of Freedom (DoF) mean**.

Potential Improvements

One improvement that the paper could use to **further achieve the stated goal/contribution** is to **generalize RT-2** for a **plethora of possible actions**. One of the limitations, as stated in the paper, is that RT-2 can only do **certain types of actions**. This makes RT-2 **very specialized** which is not the end goal of the research project, that is, to make a single trained model that can take observations and web-scale data **for many actions**. Another improvement I would recommend is to reduce the computational cost of the model. This limitation was stated in the paper, but it is worth noting that a **high computation cost** can **bar** many possible applications for its use simply because there are **not enough resources**. Fixing this limitation could **allow more accessibility** for the model for robotics application. As of now, an improvement for the paper **to leverage the latest techniques or models to improve the method** would be to **try another pretrained VLM** such GPT-4o, LLaVA, and others. I think using more recent VLM could **help reduce** the **computation cost** and **generalization ability** of the model. Although these VLM models may **not have been necessarily available** at the time of the study, I do think they **may help** in **improving** the proposed method.

Extensions

There are a couple of ideas I would like to try to extend the paper. One idea is to **try newer pretrained VLMs** to improve RT-2 or make a new RT-3 family. I think using newer models may help address some of the limitations that were stated in the paper. Another idea would be to follow-up this paper with perhaps a **more computationally efficient custom pretrained VLM**. This could be making a **smaller VLM** or **trying a different architecture** or a new technique. Lastly, I think a good idea for extending this paper would be to use RT-2 for a **larger variety of tasks** such as art, solving complex math problems, and even crosswords.