Octo

An Open-Source Generalist Robot Policy

Octo Model Team · RSS 2024 · arXiv:2405.12213

The problem

- Earlier policies often fix inputs, e.g. one camera
- Hard to finetune to new sensors or actions
- The largest models are not public

so

The model

Tokenize: text, goal image, cameras

↓

Transformer, block-wise masked

↓

Readout tokens (like BERT's [CLS])

↓

Diffusion head → action chunk

800k

trajectories from 25 OXE datasets

93M

Octo-Base params (Small: 27M)

9

robot setups at 4 institutions

~5 h

to finetune Octo-Base on one 24 GB GPU

Out of the box

- WidowX, URS, RT-1 Robot: in the mix
- Above RT-1-X on 3 robots ("29%")
- Similar to SSB RT-2-X on 2 robots
- WidowX: goals beat language ("25%")

then

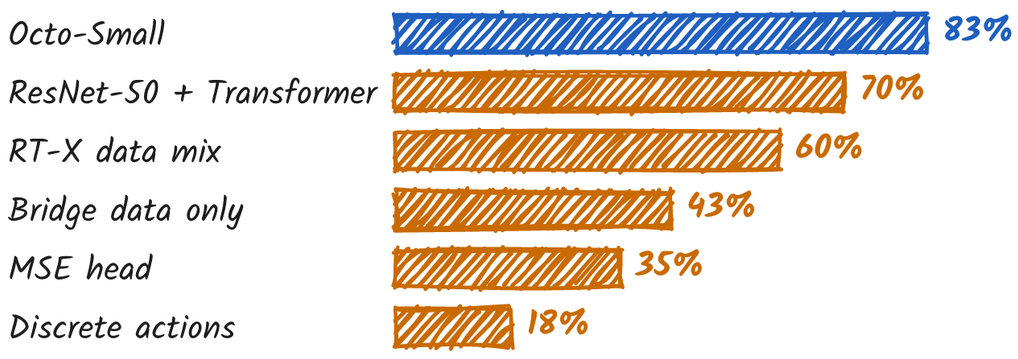
Finetuning

~100 demos of a new setup → Add an encoder or a head → Full model, 50k steps →

72%

finetuned success, 6 setups (best baseline 20%)

Ablations on WidowX (Table II)



Takeaway

- Flexible tokens + diffusion head + wide data
- A good start for finetuning on ~100 demos
- Weak: wrist camera, language; zero-shot tested only on seen robots

bushuyeu.com/read-along-octo · paper CC BY 4.0 · sketchnote by Pavel Bushuyeu

Active listening · Octo

Write while you listen. One line per box is enough.

Octo Model Team · RSS 2024 · arXiv:2405.12213

Date _____ Presenter _____ Role _____

Their main claim, in my words

The evidence they show

What I would push on

My question for the room

How it connects to my work

Listen for

- ☐ Zero-shot: Fig. 5 seen tasks, Table VII broader
- ☐ How readout tokens let finetuning add inputs and heads
- ☐ What is borrowed: diffusion head, chunks, ViT recipe
- ☐ The finetuning baselines, and the head each one uses
- ☐ Limits they admit: wrist camera, language, data mix

Ask the room

1. VC-1 has an MSE head: is the 72 vs 15 gap pretraining?
2. Does many-robot pretraining beat one robot for a new arm?
3. Wrist views often hurt finetuning: more data or new tokenizer?

Page, video and FAQ:
bushuyeu.com/read-along-octo