GATO: A Generalist Agent

CECS 550 : Pattern Recognition
Spring 2023

Group :- 7

Anthony Martinez
Diksha Patil
Pratik Jadhav
Pavan More
Sudarshan Powar

Instructor: Prof Mahshid Fardadi
Teaching Assistant: Rahul Deo Vishwakarma

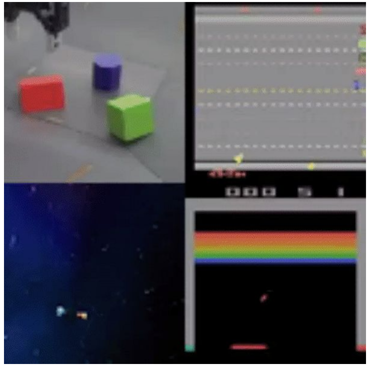
CALIFORNIA STATE UNIVERSITY
LONG BEACH

Agenda

- Introduction
- Datasets
- Data Preparation
- Training
- Performance
- Limitations
- Key Findings

Introduction

- What is Artificial General Intelligence?
- Narrow AI vs General AI
- Benefits of General AI
    - No need create domain-specific models
    - A single network would have a lot of diverse data to train on.



Datasets

- Vision and language
- Atari games
- Robot arm
- Text
- Question and Answers

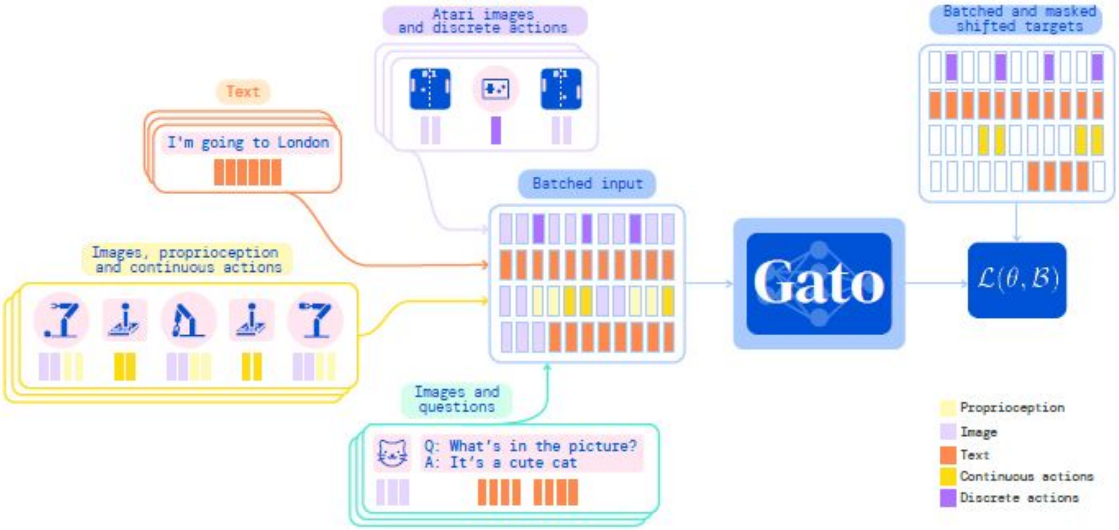
Table 1: **Datasets.** Left: Control datasets used to train Gato. Right: Vision & language datasets. Sample weight means the proportion of each dataset, on average, in the training sequence batches.

Control environment	Tasks	Episodes	Approx. Tokens	Sample Weight
DM Lab	254	16.4M	194B	9.35%
ALE Atari	51	63.4K	1.26B	9.5%
ALE Atari Extended	28	28.4K	565M	10.0%
Sokoban	1	27.2K	298M	1.33%
BabyAI	46	4.61M	22.8B	9.06%
DM Control Suite	30	395K	22.5B	4.62%
DM Control Suite Pixels	28	485K	35.5B	7.07%
DM Control Suite Random Small	26	10.6M	313B	3.04%
DM Control Suite Random Large	26	26.1M	791B	3.04%
Meta-World	45	94.6K	3.39B	8.96%
Procgen Benchmark	16	1.6M	4.46B	5.34%
RGB Stacking simulator	1	387K	24.4B	1.33%
RGB Stacking real robot	1	15.7K	980M	1.33%
Modular RL	38	843K	69.6B	8.23%
DM Manipulation Playground	4	286K	6.58B	1.68%
Playroom	1	829K	118B	1.33%
Total	596	63M	1.5T	85.3%

Vision / language dataset	Sample Weight
MassiveText	6.7%
M3W	4%
ALIGN	0.67%
MS-COCO Captions	0.67%
Conceptual Captions	0.67%
LTIP	0.67%
OKVQA	0.67%
VQAV2	0.67%
Total	14.7%

Data Preparation

- Tokenization
- Sequence Ordering
- Embedding



Training

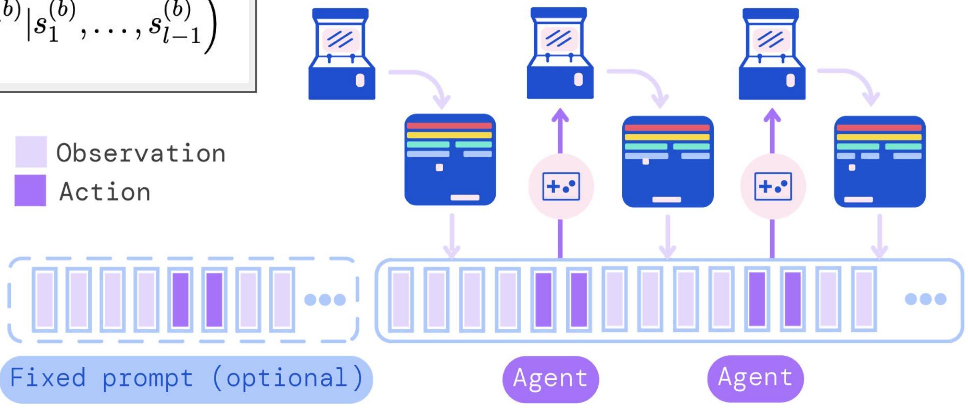
- Loss Function

$$\log p_\theta(s_1, \ldots, s_L) = \sum_{l=1}^{L} \log p_\theta(s_l | s_1, \ldots, s_{l-1}),$$

$$\mathcal{L}(\theta, \mathcal{B}) = -\sum_{b=1}^{|\mathcal{B}|} \sum_{l=1}^{L} m(b,l) \log p_\theta \left(s_l^{(b)} | s_1^{(b)}, \ldots, s_{l-1}^{(b)} \right)$$

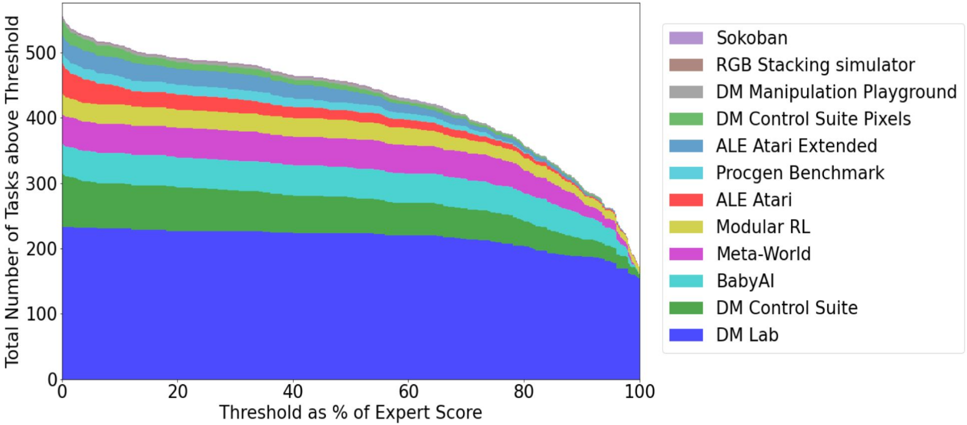
Masking function

$$m(b,l) = \begin{cases} 1 & \text{Text or Action} \\ 0 & \text{Image or Observation} \end{cases}$$



Performance of model

- 450 out of 604 tasks at over 50% expert score.
- Ale Atari : 23 >= average human score
- BabyAI: 80% of expert score for almost all levels.
- Rudimentary dialogue and image captioning.



A bearded man is holding a plate of food.

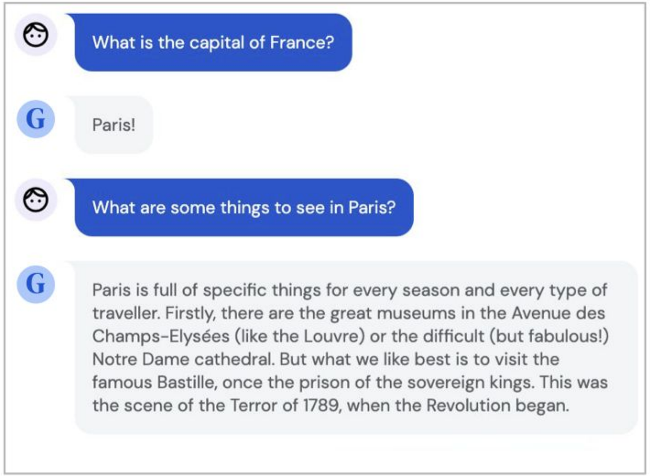
Man holding up a banana to take a picture of it.

a man smiles while holding up a slice of cake

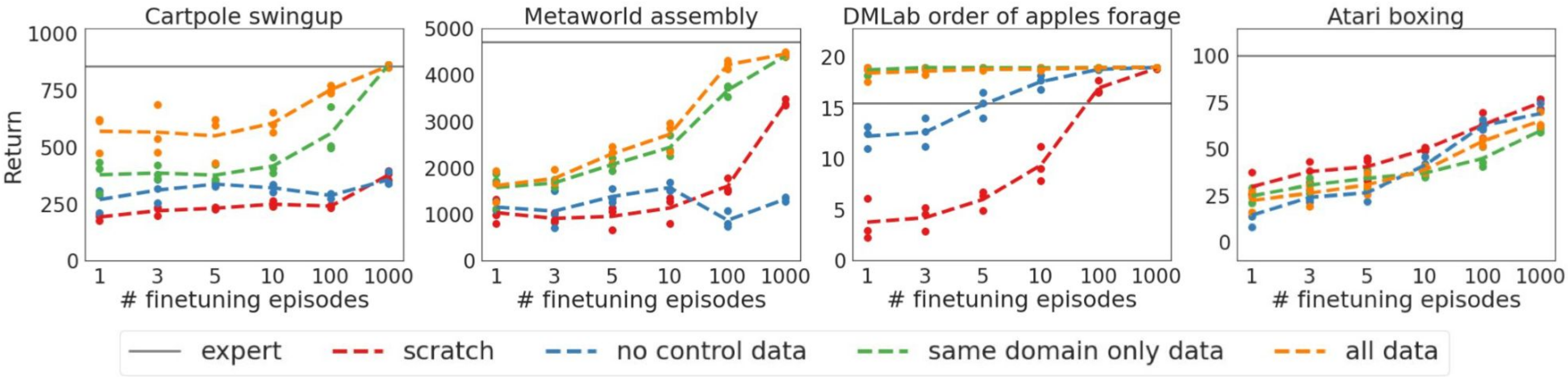
A group of children eating pizza at a table.

Two boys having pizza for lunch with their friends.

The boys are eating pizza together at the table.



Performance on unseen tasks



- The network can experience positive transfer from all the multimodal data fed to it.
- This means even the data is from all different modalities, it is helping the model learn better and make it perform better than model with only same domain data.

Scaling

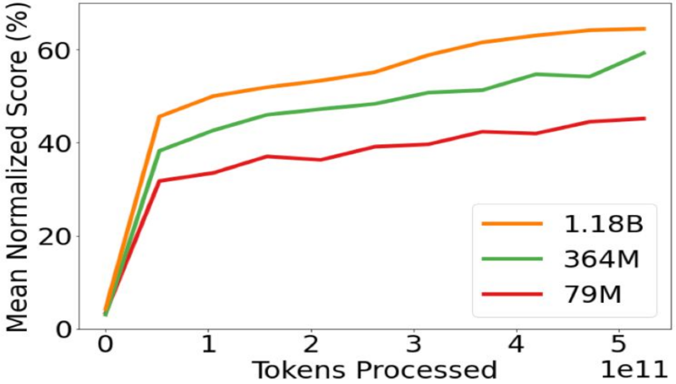


Figure 8: **Model size scaling laws results.** In-distribution performance as a function of tokens processed for 3 model scales. Performance is first mean-aggregated within each separate control domain, and then mean-aggregated across all domains. We can see a consistent improvement as model capacity is increased for a fixed number of tokens.

- Will Scaling increase performance?
- ChatGpt-3 has 175B parameters
- ChatGpt-4 parameters unreleased as of this date.

Key Findings

- Generalist agents can perform reasonably well on multi-task multi-embodiment policies, including for real-world text, vision and robotic tasks.
- Have potential to learn new tasks with few data points (few-shot learning).
- Performance across all tasks will increase with scale in parameters.
- By scaling up we can build a general purpose agent.

Limitations

- Jack of all trades, master of none.
- Computational power.
- Ethical considerations.

References