General-Purpose In-Context Learning by Meta-Learning Transformers

Louis Kirsch[1,2], James Harrison[1], Jascha Sohl-Dickstein[1], Luke Metz[1]

[1]Google Research, Brain Team [2]The Swiss AI Lab IDSIA, USI, SUPSI

@LouisKirschAI

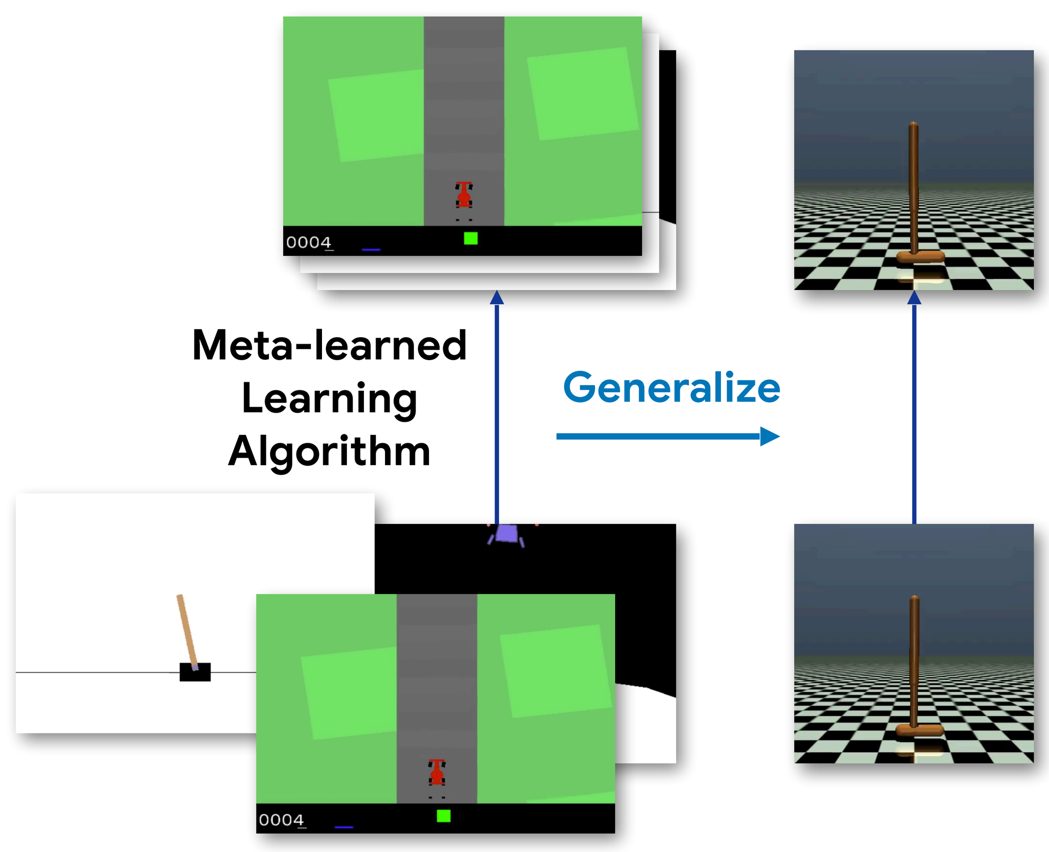
Motivation

General Purpose Meta Learning

Drive **advancements** in Machine Learning via Meta Learning

Enable **reusability** across a wide range of tasks

Here: Focus on memory-based / **in-context** learning



Conclusion

- Transformers and other black-box models can be meta-trained to act as **general-purpose in-context learners**
- There are **phase transitions** between algorithms that **generalize**, algorithms that **memorize**, and algorithms that fail to meta-train at all, induced by changes in model size, number of tasks, and meta-optimization
- The capabilities of meta-trained learning algorithms are **bottlenecked by the accessible state size (memory)** unlike standard models which are thought to be bottlenecked by parameter count

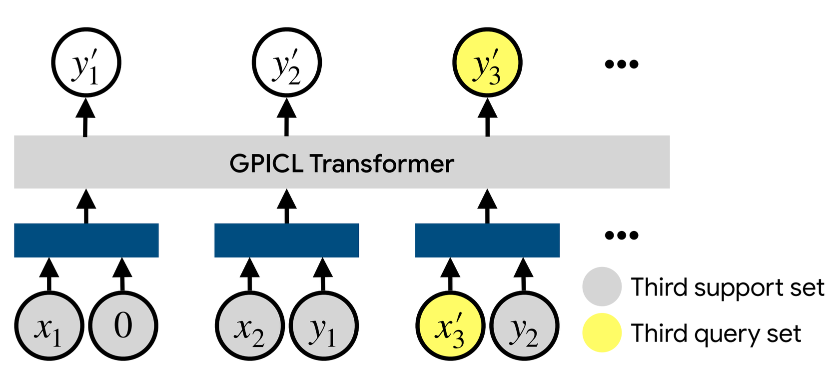
General-Purpose In-Context Learning (GPICL)

What is an In-Context Learning Algorithm?

In supervised learning $\left(\{x_i, y_i\}_{i=1}^{N_D}, x'\right) \mapsto y'$

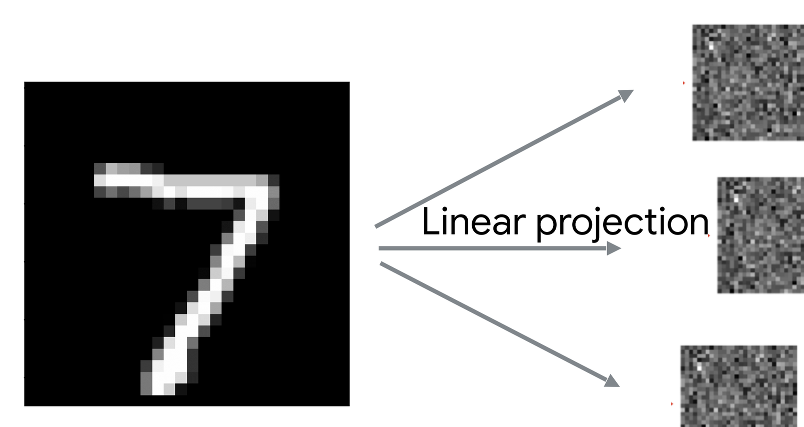
Learning = Improving predictions y' with larger $D = \{x_i, y_i\}_{i=1}^{N_D}$

With black-box models such as LSTMs or **Transformers**



Hypothesis: Many diverse tasks → **G**eneral-**P**urpose **I**n-**C**ontext **L**earning-to-learn

Generating Tasks for Learning-To-Learn



- Meta-train multi-task across n tasks
- Only a single prediction head

$$J(\theta) = \mathbb{E}_{D \sim p(D)}\left[\sum_{j=1}^{N_D} l(f_\theta(D_{1:j-1}, x_j), y_j)\right]$$

Base dataset eg MNIST dataset

Create n tasks

$\bar{D} = \{\bar{x}_i, \bar{y}_i\}$ $\quad D = \{A\bar{x}_i, \rho(\bar{y}_i)\}$ $\quad A_{ij} \sim \mathcal{N}\left(0, \frac{1}{N_x}\right)$

Linear projection **Label Permutation**

Label ↦ one-hot index

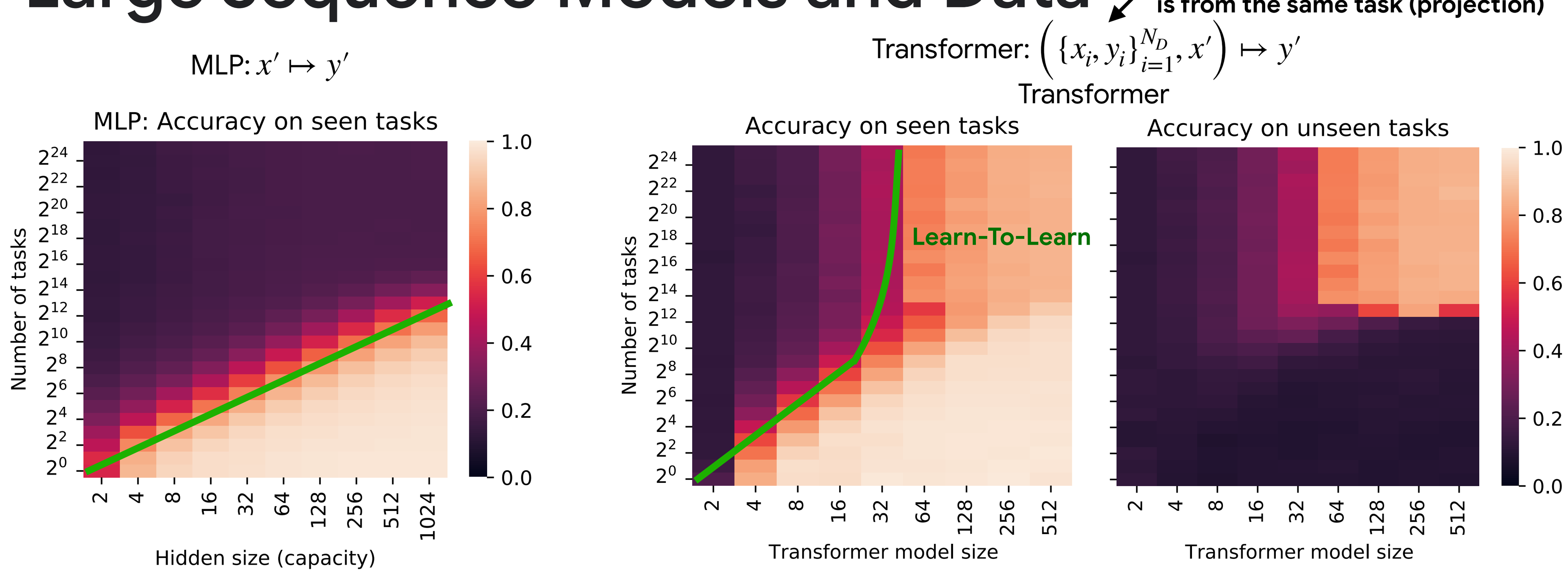
Results

Large Sequence Models and Data

Each element in the sequence is from the same task (projection)

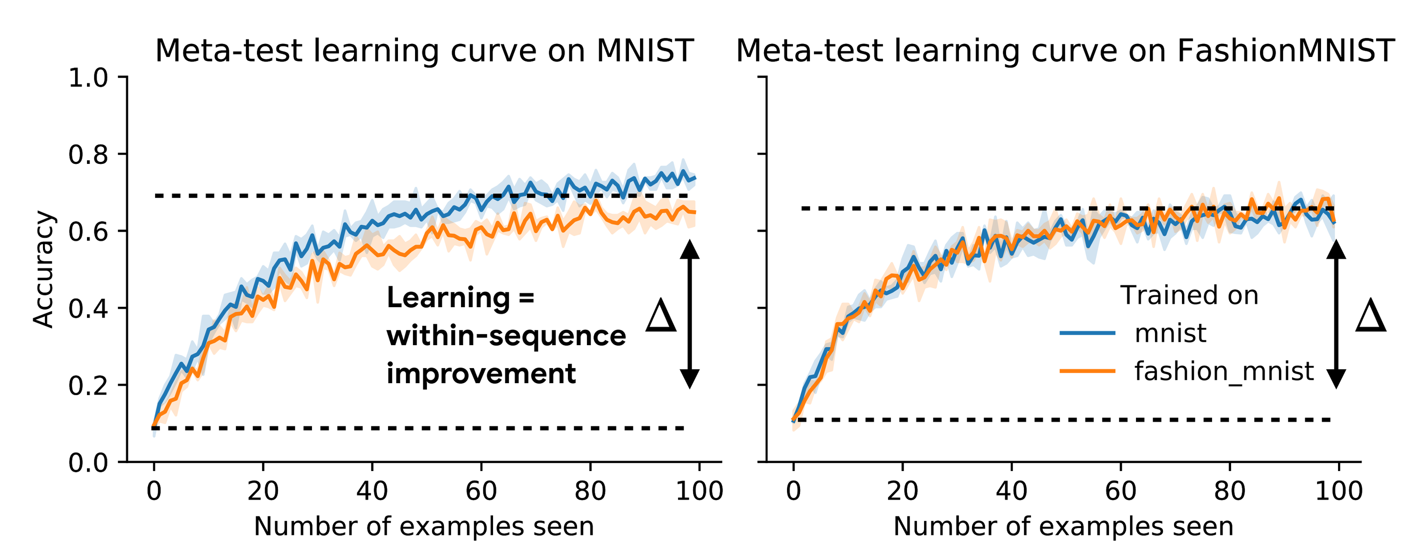
MLP: $x' \mapsto y'$

Transformer: $\left(\{x_i, y_i\}_{i=1}^{N_D}, x'\right) \mapsto y'$



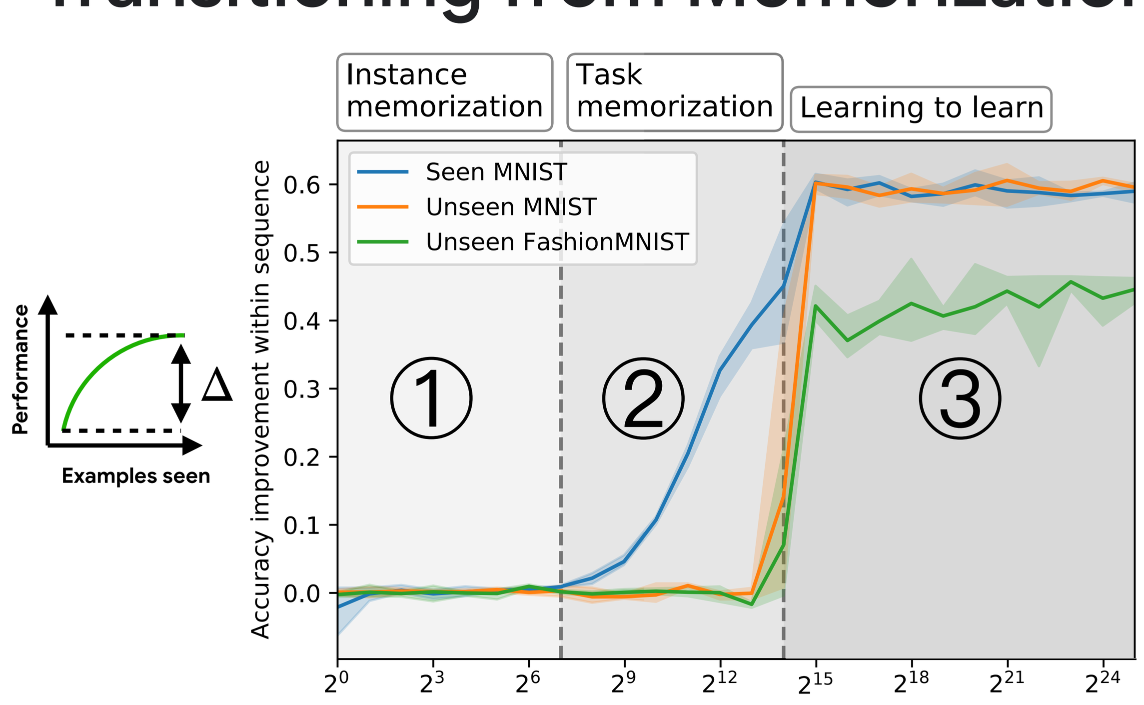
At a certain model size and number of tasks, the Transformer generalizes to a seemingly unbounded number of tasks

The Emergence of Learning-To-Learn



The meta-trained GPICL learns from examples at test time, and generalizes to unseen datasets

Transitioning from Memorization to Learning



Transformers exhibit three different phases in terms of meta-learned behavior

Phase	Learning	Generalization	Algorithm Description
①	× No	× No	Instance memorization
②	✓ Yes	× No	System identification / Task memorization
③	✓ Yes	✓ Yes	General-purpose learning algorithm

Architecture: A Large State is Crucial for Learning

The state size (accessible memory) of an architecture most strongly predicts its performance as a general-purpose learning algorithm