General-Purpose In-Context Learning (GPICL)

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

Meta-learned Learning Algorithm
Generalize

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

What is an In-Context Learning Algorithm?

In supervised learning \( \{(x_i, y_i)\}_{i=1}^N \mapsto y' \)
Learning = Improving predictions \( y' \) with larger \( D = \{x_i, y_i\}_{i=1}^N \)
With black-box models such as LSTMs or Transformers

Hypothesis: Many diverse tasks \( \mapsto \) General-Purpose In-Context Learning-to-learn

Generating Tasks for Learning-To-Learn

• Meta-train multi-task across \( n \) tasks
• Only a single prediction head

Base dataset eq MNIST dataset Create \( n \) tasks
\( D = \{x_i, y_i\}_i \)
Linear projection
Label Permutation
Label \( \mapsto \) one-hot index

Large Sequence Models and Data

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

<table>
<thead>
<tr>
<th>Phase</th>
<th>Learning</th>
<th>Generalization</th>
<th>Algorithm Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No</td>
<td>No</td>
<td>Instance memorization</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>No</td>
<td>System identification / Task memorization</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>General-purpose learning algorithm</td>
</tr>
</tbody>
</table>

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