Emergence as Prediction

Aditi Group Meeting 12/1

What is emergence?

What is emergence?

Jacob Steinhardt

“when quantitative changes in a system result in qualitative changes in behavior”

What is emergence?

Jacob Steinhardt

"when quantitative changes in a system result in qualitative changes in behavior"

Jason Wei

"not present in small models but is present in large models"

What is emergence?

Jacob Steinhardt

"when quantitative changes in a system result in qualitative changes in behavior"

Jason Wei

"not present in small models but is present in large models"

Rylan Schaeffer

"choice of metric rather than fundamental changes in model behavior with scale"

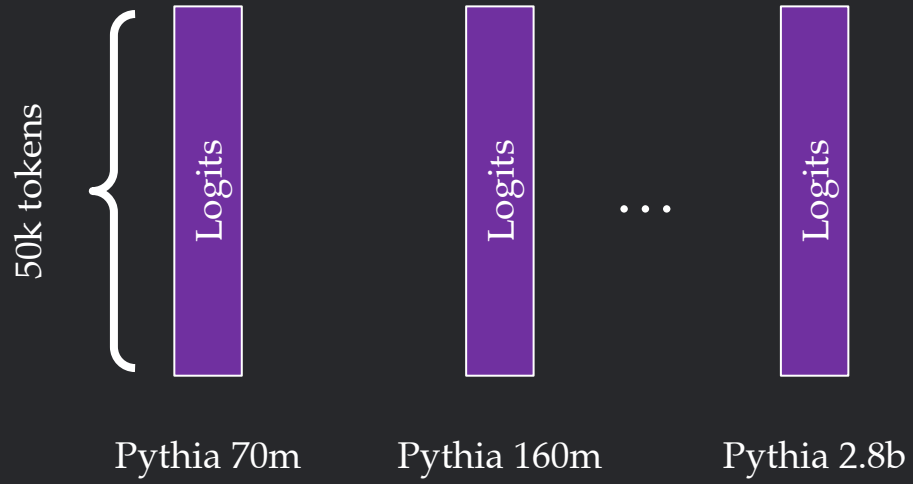
What is a testable hypothesis?

Hypothesis: Emergent behaviors are **unpredictable** from smaller models

Benefit: We can quantify what it means to be predictable

Emergence as predicting model behavior

The quick brown fox jumps over the lazy _____



Emergence as predicting model behavior

The quick brown fox jumps over the lazy _____

Q: How do we show something is unpredictable?

A: We try really hard at predicting it and fail

Input

Your input will be a `100 x 6 x 50277` matrix

- Axis 1: Prefix (“The quick brown fox jumps over the”)
- Axis 2: Model
  - You will be given the next-token probs for `Pythia {70m, 160m, 410m, 1b, 1.4b, 2.8b}`
- Axis 3: Token (Pythia tokenizer has 50277 tokens)

```
array[i][j][k] = prefix i, model j, token k
```

Output

Your input will be a `100 x 6 x 50277` matrix

Your requested output will be your prediction for which token is graded by Pythia 12b as the highest likelihood token

- Output format will be a list of token indices (i.e. [1931, 281, 2827, …] of len 1000)

Datasets

For training, you will be given two possible train datasets

- **wikipedia** (each prefix is 10-20 words, uniformly sampled)
- **glue-rte** (each prefix is a query in the dataset)

I have a held-out test set for evaluation

Code Structure

Download/copy `challenge.py` from the slack message I sent

You have to write a strategy function that can take an input of `100` strings and a `100 x 6 x 50277` matrix and produce a prediction of the top-1, which will be a length `100` list

I've attached my `environment.yml` for convenience

- `conda env create -f environment.yml`

Logistics

There is a **$20 reward** 💸💸💰💰💵💵 for the winning team!

You can pair up with as many people as you would like, the reward will be equally split.

You have until **11:10** before we evaluate strategies

Ready? Set. Go!

https://github.com/kothasuhas/predicting_emergence

Simple extensions

Can we properly quantify emergence as the difficulty of predicting the next token distribution?

Is this predictable enough to lead to practical gains, where the forward pass for inference is over smaller models rather than larger models?

Is prediction actually harder for tasks considered “emergent” such as in-context learning and responding to Chain-of-Thought?