AI Forecasting Research Ideas

by Jaime Sevilla, Anson Ho and Lennart Heim

Overview

- This document contains a collection of AI Forecasting research ideas, prepared by some Epoch employees in a personal capacity
- We think that these are interesting and valuable projects that research interns or students could look into, though they may vary in difficulty (depending on your background/experience)
- This is the result of a quick brainstorming and curation, rather than a thorough deliberative process. We encourage a critical outlook when reading them.
- You may also be interested in these other forecasting research ideas suggested by Jaime Sevilla
- In general, you can use Epoch's database to find notable machine learning papers with parameter, compute, and dataset sizes. On Epoch website you can also find our past research, a tool for visualising the dataset, and some other tools like this compute calculator.
- Please feel free to contact us for clarification about these questions!

Projects

Extrapolating GPT-N performance

Difficulty: Medium

Lukas Finnveden previously performed an extrapolation of GPT-N performance on several benchmark tasks, such as cloze completion and arithmetic (Finnveden, 2020). Can you expand on this methodology and apply it to more cases?

Qualitatively analysing language model / image generation improvements since ~2000

Difficulty: Easy

While we can plot graphs showing quantitative changes in language model / image generation performance over time (e.g. in terms of the perplexity), what does this actually mean in terms of model capabilities? A collection of samples from language models in the last two decades could help give a visceral sense of how much they have improved. The comparison could include a selection of the best output out of 10 prompts, a comparison of prompt completions, etc.

Do AI researchers train models using scaling laws?

Difficulty: Medium

Scaling laws have been proposed as ways to gather information about how to train large machine learning models efficiently (Kaplan *et al.*, 2020), and this has been done in practice for training LLMs like Chinchilla (Hoffmann *et al.*, 2022). But how broadly have scaling laws been used by AI researchers in general, and has there been a delay in the uptake of such an approach?

Revisiting 'Is AI Progress Impossible To Predict?'

Difficulty: Hard

Alyssa Vance argued that AI progress on a task from one model to the next was unpredictable (Vance, 2022). Can we investigate this in more detail? For instance, the authors of *Beyond the Imitation Game* (Big Bench) find that for tasks where progress is "jumpy", there are usually progress metrics that vary more smoothly (Srivastava, 2022). Can we use those metrics to predict progress?

Algorithmic breakthroughs in machine learning history

Difficulty: Medium

What were the major algorithmic innovations in machine learning over the last two decades? This could be structured as a literature review or as a survey of experts, culminating in a big list of the key algorithmic advances over the last ~20 years. Such a database helps us understand the frequency and significance of algorithmic insights.

Improvements due to "software-for-hardware"

Difficulty: Hard

Innovations in compilers and other low-level improvements have helped increase the utilisation rate of GPUs and improve training efficiency.

Make a list of such improvements and how much did they improve performance overall for tasks such as training a Neural Network.

Paradigm changes in AI

Difficulty: Medium

What were the major paradigm shifts in different domains of AI? By talking to domain experts, reading lit reviews and popular papers, discern what methods were popular at each point in time and compile a list of these domain-specific paradigm shifts. Such a list allows us to use Laplace's rule to estimate a base rate of paradigm changes in AI.

Study training run lengths

Difficulty: Easy

Epoch worked out a theoretical upper bound to training run clock length of 14-15 months. Empirically investigate trends in training run lengths, and see how it compares to this theoretical upper bound – what are the reasons for the discrepancies? This would require building a dataset of training run lengths.

AI development vignettes

Difficulty: Hard

Write down qualitative and concrete stories about AI development, exploring the possible risks and societal consequences. The emphasis here should be on detail, and you should take potential hardware, algorithmic, and data constraints into account (e.g. what happens if Moore's law ends in a few years?).

Profiler to measure compute

Difficulty: Hard

Compute is one of the key inputs in machine learning, very predictive of performance and relatively easy to measure. However, compute usage typically isn't reported even in top journal articles. Part of the reason for this is the lack of good profiling tools in GPUs and/or machine learning frameworks. The task is thus to implement an open-source solution into a framework like PyTorch. This could help shift the community's norms towards more transparent reporting, which in turn would create a lever for AI governance interventions.

Lennart Heim has an extensive draft on this issue he would be happy to share on request.

Insights-based models of AI timelines

Difficulty: Medium

The Median Group previously proposed a model of AI timelines based on key "insights" required on the way to AGI development. However, the current model is based on outdated and poorly curated data, and there are some questionable methodological choices. Collect data that is more up-to-date, and redo the model – how do your results compare to more well-known timelines models?

Brain emulation development

Difficulty: Medium

Anders Sandberg looked into a Monte Carlo model of brain emulation development (Sandberg, 2014). However, this paper is now old and has outdated estimates. Replicate the methodology of this paper – what are the new results?

Rethinking the evolutionary anchor

Difficulty: Hard

In *Forecasting TAI with biological anchors*, Ajeya Cotra proposes the “evolutionary anchor” as a hypothesis for how the compute needed to train generally intelligent systems, based on “the total FLOP performed over the course of evolution, since the first neurons” (Cotra, 2020). But there have been some concerns about whether this definition is appropriate – it does not account for the compute for simulating the environment (Sempere, 2022), and anthropic considerations might prove highly important (Erdil, 2022). Assess the significance of these concerns, and reassess the viability of the current definition of the anchor.

Investigate trends in memory bandwidth, latency and price of memory

Difficulty: Easy

Memory is a big challenge for modern deep learning, required for storing things like parameter values and intermediate gradient computations (Weng, 2021). How have memory bandwidth, latency, and the price of memory changed over time?

Investigating the parameter gap

Difficulty: Medium

In a previous investigation, Villalobos et al identified a “parameter gap” – that is, a surprising lack of notable ML models with sizes between 1e10 and 1e11 parameters. Investigate whether the proposed hypotheses in the paper are accurate or identify new hypotheses to explain the data.

Investigate possible paradigm shifts in hardware

Difficulty: Medium

Innovations in hardware might change the pace of the trend on FLOPS/$, which would have important implications for AI.
Look into a possible paradigm change (eg optical computing, 3D chip stacking, crossbar computation, etc), investigate who is working on it and what results have been demonstrated so far.
The IRDS 2020 report is a good starting point to learn about some emerging hardware paradigms.

How much can we scale up production in the compute supply chain?

Difficulty: Hard

How quickly we will be able to increase the production of hardware and the required computing infrastructure (such as data centers) as we increase the investment in compute for AI?.
To what extent could ASML, TSMC, and other suppliers across the compute supply chain scale up their production on a 1/5/10 year timescale if the prices of their products went up dramatically (eg x10)? What are the key bottlenecks for such a massive scale-up across the compute supply chain?

Will the price of compute go down if hardware performance stops improving?

Difficulty: Hard

When innovation periods for hardware accelerators are extended, amortisation happens and prices go down (Heim, 2021). Over the past few years, innovation cycles have been relatively short, thus leading to relatively high prices, but what happens when innovations stalls? Under this view, we might predict the entrance into a period of amortisation where prices are driven down, enabling an economy of scale which in turn drives prices down further still. In the event of the current trend of hardware improvements coming to a stop, how might such a dynamic affect the price of compute going forward?

What has been the share of any chip in a given year of total available compute performance?

Difficulty: Easy

New chips are continuously developed, and old chips are subsequently replaced, but not instantly. Ultimately, we want to have a better understanding of the dynamics of how chips are replaced over time. To help with this, construct a database that specifies the following: for each year between 2010 and 2020, what share of the available compute came from which chips?