Resources

Open questions

- By 2024, will there have been a 2-year interval in which the AI-compute trend did not double?
- By 2024, will there have been a 2-year interval in which the AI-compute trend did not grow 8x?
- By 2024, will there have been a 2-year interval in which the AI-compute trend did not grow 32x?
- Maximum training compute used by systems in 2019 and 2020

Models

- Compute growth calculator, which allow you to enter a target compute/cost and find out how long until it's reached, according to the AI Compute doubling trend (to find the calculators press the little calculator symbol next to “redo” in the menu bar).
- Spreadsheet with data on the AI-compute trend

Invite link to join discord server

Agenda

- Structure of workshop
    - *3 min, heads-down:* List down technical uncertainties regarding resolution etc.
    - *5 min, group discussion:* Going through uncertainties together
    - *2 min*, *heads-down:* Give your lower and upper bounds/or current probability on the question
    - *3 min, group discussion*: Sharing initial estimates
    - *30 min, heads-down:* Collaborate in the Google doc, and use discord for conversations.
        - Try to stick to the headings:
            - Outside views
            - Factorizations
            - Uncertainties
            - Scenarios & inside views
            - What would change your mind
            - Meta
        - Use tags! For example:
            - [Important] Next to your own points that you’d like people to answer
            - [+1] Next to others points that seem important
            - [Info request]
    - *15 min, group discussion*: Most highly upvoted points
    - *5 min:* updating and sharing our final estimates. Time for predicting on ai.metaculus.com and writing up brief comments.

Resolution uncertainties

Is anything ambiguous about the question formulation? If we gave this question do a clairvoyant who would see the future, would she have to ask "what did you mean"?

- [answered] JC: Is that date of the experiment the a) date the experiment finishes b) publication (on arxiv? blog? accepted in journal or conference?) c) something else - this is especially relevant in the case where the training takes a very long time (eg. OA5 taking 10 months or whatever it was)
    - First public announcement
- [answered] What is meant by "AI-Compute"? The amount of money spent? The amount of operations applied toward AI experiments?
    - Amount of operations, measured as petaflops/s-days used in training (not architecture search or hyperparameter tuning). See the question background here, as well as the original paper deriving the methodology here.
- [answered] What numerical accuracy? 8bit int vs bfloat13 are different. Accumulate? They don't say.
    - Can get a lot more operations per second given different size of the registers being used.
    - the original post uses 2 flops per add-multiply, ignoring precision
        - "We count adds and multiplies as separate operations, we count any add or multiply as a single operation regardless of numerical precision (making "FLOP" a slight misnomer), and we ignore ensemble models. "
    - [info-request] Is there a general method for normalizing between different "register sizes" to calculate performance? My guess is that standard benchmark tests would account for this.
- [answered] DF: How will the amount of compute used in an experiment be quantified? E.g. if in a 2-year period experiment A happens and uses 1 flop, B happens and uses 1.01 flop, and C happens and looks like it might have used 1000 flop but the authors don't say exactly, how does this resolve?
    - Sometimes order of magnitude estimates will work fine for resolution, other times resolution might be ambiguous, other times we might have to wait until sufficient information to make an estimate is released
- [answered] ~~Are we talking about price performance, or flops per largest experiment?~~
    - *See question 2 above.*
- [answered] What counts as one experiment? One training run, or the whole project to achieve a goal (e.g. Go)
    - Training for one particular system.

Lower and upper bounds

JC: 20% - 60% - 85% (point estimates for 2x, 8x, 32x) - Would need 32x happen 2.5 times (if you staggered announcements optimally), but still the end thing would be hugely costly
DF: 32x: bid 0.7, at 0.92. 8x: bid 0.6, at 0.85. 2x: bid 0.3, at 0.65
datscilly: double: 25%; 8x: 40%, resolves negatively if there are large 8x projects in 2020 and 2022; 32x: 60%
Michael: 30%, 50%, 80%
BG: 20% 40% - 60%? (highly uncertain about these point estimates)
BB: 10% 50% 90%
Jotto: 50% (8x)

Main considerations

Outside views

Are there some reference classes of previous events that give evidence about this?

If you look at 2-year periods between 2012 and 2018, of which there are 5:

- 4 (80%) have a 2x on the max compute used in that period
- 3 (60%) have a 10x
- Either 1 (20%) or 2 (40%) have a 32x (a bit hard to tell, I'm leaning towards 1)

JC: Is the relevant question something like "what fraction of 5 year periods contain a 2-year period which does not contain a jump of X"?

Factorisations

How might you break this question into sub-questions?

(For example, US demand for cars can be broken down into the number of drivers, how often they change cars, and their average salary.)

- Things driving the trend:
    - Economics
        - Supply of compute
        - Demand for compute
    - Parallelizability
        - Algorithmic
        - Data
- Price-performance
    - moore's law
        - Modulo people not actually having things to try, this should give >2x compute doubling/year
    - Difficulties in continuing to improve standard CMOS electronics:
        - There are now only 3 major chip manufacturers aiming at 7nm and below: Intel, TSMC, and Samsung. Global Foundries dropped out at 12nm. It costs $10B to build new fabs.
    - specialised hardware
        - how simple is the training task? are we just doing lots of matrix multiplies or are we doing search over architectures or something that needs a more general-purpose chip?
            - currently seems like a large part of the cost of experiments is spent on CPUs (collecting experience in the environment) even though most of the flops are performed on GPUs or TPUs (gradient updates)
                - Re this, c.f. these guesstimates showing the CPU flops *and cost* to just be a rounding error on TPUs for AlphaStar
                - definitely not the case for OAI5 (dota) - think CPU cost was 10x GPU cost. shadowhand similar. But I think DOTA is unusually bad for this.
    - new paradigms e.g. optical?
        - my impression is that 6 years is too soon here, but I'm not sure
            - There are multiple optical computing companies that will have 'something' within that time. (optalysys, lightmatter, lightelligence, fathom)

                - will it be significantly better on price-performance than ML chips? (I'm probably biased, but at least one of them will)
                - what sort of factor? 10-1000x (in the 2.5-3.5yr timeframe)
                - than the TPUs at the time they're ready? Woah
            - I update downwards (less likely that there wouldn't be a massive compute increase)
        - competition and pricing in GPU market (seems like ML GPUs are fairly overpriced compared to actual production cost)
- willingness to spend
    - benefits of more expensively trained AI
        - On the one hand, compute is surprisingly useful (Rich Sutton's "bitter lesson")
        - On the other hand, quantitatively, compute has surprisingly small returns (this is the correct update if we were surprised by how fast this trend is)
        - Some products seem to get a lot more valuable if they are best in class. So a company that spends $1B on the best object recognition could win the market by being a few percent more accurate.
    - awareness of the benefits/is progress incremental or do we have to spend $1bn on an experiment, wait a year, and only then see if it worked?
    - total cost vs cost of biggest experiment
        - an actor might be spending 10bn on an AI research project if they're spending 1bn on a single training run
- speed of ramping up larger experiments
    - supply of chips
    - time to buy, assemble, connect up and program data-centres/supercomputers
        - are chips used for other things or just for this one experiment
- things that affect counting:
    - precision (if we're counting flops regardless of precision, then a move to lower precision would lead to higher flops)
        - Indeed, there's some literature on using 8-bit floats instead of 32- or 64-bit ones for NN training, showing that it's just as good. Maybe people gradually reduce the number of bits that you need? If that were to go down by 1 bit a year, that's the sort of trend that could continue for 5 years.
    -

Uncertainties

What are the sources of variation in your forecast?

Remember to use the [Important] tag for things that would be useful to reduce uncertainty on!

- How many new organizations will have the budget to run large compute experiments that could fulfill this growth req?
    - Like James mentioned, and as Ryan Carey mentions in the post, these experiments will cost billions in a few years if trend continues. So I'd surprised if anyone but current tech monopolies, governments, and maybe OpenAI could do this
    - To attempt to enumerate the organisations that could/would plausibly spend >$20m on a single experiment [+1][important]
        - AGI labs
            - OpenAI → likely
            - Deepmind → highly likely
            - Google Brain (now called: "Google AI") (or non-Deepmind Google more generally)
        - Tech industry
            - US
                - Facebook
                - Microsoft
                - Amazon
                - Apple
                - Nvidia
                - Tesla
                - Intel
                - uber
            - Chinese
                - Baidu
                - Tencent
                - Alibaba
            - Other
            - Canada?
        - Governments
            - US
                - Would any US government organization publish their results from an ML experiment? If so which ones?
                    - NSF
                    - -ARPAs
            - Chinese

        - → Seem much more willing than US. C.f. this other question about how much they're spending on AI
    - Europe

- [info-request] [answered] The # of times this max compute record is broken is quite small. How often has this record of "largest experiment" been broken?
    - According to OpenAI compute trends graph, seems to be about 1 per year
    - C.f. this comment from the site:
        - "July 1st 2020 is around 15 months from now. From the chart, there seem to be around 15 noteworthy project in 7 years, or around 2.7 every 15-month period. Assuming the timing of projects follows a uniform distribution, and that each project uses more compute than the previous, we should expect the largest project before the question's resolution date to occur around 10.9 months from now (using the expected value of the highest draw from a uniform distribution formula). That gives enough time for just over 3 doublings, if the doubling time remains around the same. If we extrapolate from the estimate of Alphastar's computing of 28x10^3 PFLOP/s-days, we may expect upwards of 200x10^3 PFLOP/s-days of compute used in training."
    - Per the question, there's a lag of around 2 years between major experiments.

Scenarios and inside views

How might the future unfold for this to resolve true? And why should we expect things to turn out differently than in the past?

Nov 2017: 1.8 PFLOP/s-days (Alphago Zero)
Jan 2019: 10-25 PFLOP/s-days (AlphaStar)
Seems like the trend of 10x per year is holding up so far.

Datscilly: default scenario is hardware improving 2x per year (price/performance) and investment increasing 2x per year, for a total of 4x per year. Then 8x question should be under 50%, 32x question over 50%

- How much do other forecasters think investment ($) in large AI project to go up over time? 2x is just a first guess [Prediction request]
    - How likely is it to reach the tech-monopoly-budget-upper-bound ($20B) (for a single experiment)?
    - How likely is it to reach the government-experiment-upper-bound ($200B) (for a single experiment)?

- Scenario: There is an increase in the total amount of compute used for AI experiments, but it's spread across many different experiments, and there are no longer valuable experiments to run that "max" out on compute
- Scenario: Most valuable new experiments involve massive data instead of compute
    - I'm thinking of GPT-2 which seems to have been data not compute intensive
        - C.f. the question "how much compute did GPT-2 use" with crowd median <100 pfs-days
        - Beth: supervised models are generally *much* cheaper to train than RL agents (several OOM less expensive)

Will we know of all experiments that happen? Most are only revealed 1-3 years later.

What would change your mind?

What are some important and plausible points that would change your mind about this forecast?

DF: I think that it's unlikely that the AI-compute trend will continue, and expect it to slow down. Things that would change my mind:

- DF: AI experiments start producing significant amounts of economic value in the near future, such that it suddenly becomes very much worth it
    - In what situations would it be worth it to *retrain* that experiment often? Which is what we're measuring.
    - DF: Actually, really the thing you need is for the derivative of economic value with respect to computation used to become quite high.
- DF: Moore's law starts speeding up instead of slowing down, because of some new fancy computation technology that will predictably come on the scene and cause a discontinuity in cost-effectiveness.
    - +michael what's the pitch for how much optical computing would improve things?
        - Compute is currently limited mostly by a bottleneck of moving data between memory and logic.
    - DF: One way this could manifest is specialised chips that are very good at the operations that are used in AI, like TPUs, or perhaps some fancy software that makes good use of flops in NN training? But again, this needs to continuously happen, otherwise it affects the level and not the growth.

Meta

DF: why am I constantly hearing beeping? There's a sound that's initially a high note and then a low note immediately after that keeps on happening, and is super irritating. How can I stop that?

- I think it's people changing channels
- DF: looks like you can turn off various sounds in the notifications section of Settings

- Some type of epistemic tag/shorthand when adding answers to people's info requests.
- Better attribution of who wrote what, ex. Dropbox paper highlighting on who the side wrote something

Final estimates

Remember to post your reasoning to [the question page on Metaculus AI](#).

This helps make the AI Metaculus prediction more interpretable to AI decision-makers and researchers who might rely on it.

JC: 25% - 60% - 90%
Michael: 20%
BG:
Jacob: 14% - 40% - 80%

Future work, and next steps:

Feature work -- things to follow-up on and do next time/other times

Jotto -- better interface for question writing -- forking other's questions

Other Large AI projects:

1. OpenAI's DOTA2: https://openai.com/blog/how-to-train-your-openai-five/ used 800 petaflop/s-days and experienced about 45,000 years of Dota self-play over 10 realtime months