sustAInability

Jon Crowcroft

15/7/24

Basic cs resources

- If you want to process more data, in the same time,
  - need to clock processors faster…which is no longer happening (Moore's out)
  - Batch/pipeline
  - or parallelize (lots of cores, either cpu or gpu)
- Some data processing doesn't parallelize much easily either
  - E.g. graph neural nets – typically require vast RAM
  - Some does (e.g. LLMs, even with multi-head attn)
- Energy cost scales between n and n^2 (sometimes worse)

Scale but at what cost?

- Hotchip 2023 Amin Vahdat and Jeff Dean – summary here
  https://www.nextplatform.com/2023/08/29/the-next-100x-for-ai-hardware-performance-will-be-harder/
- But see earlier warning/corrective from systems people:
  https://www.usenix.org/system/files/conference/hotos15/hotos15-paper-mcsherry.pdf
- Sustainability has a number of dimensions in AI
  - People
  - Data
  - Compute
  - Energy
- Not optional:- new EU reporting requirement
  - https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en

Training v. inference cost – LLMs and the rest

- Where will new *training* data come from?
  - After Common Crawl (some of which is private or copyrighted)
  - There is only 1 Interweb, and it took 32 years to get to here.
  - Growing from synthetic data would lead to model collapse
- Adding alignment/guardrails during *inference* also costs
- Why is that any different from all pre-LLM AI?
  - Labelling data isn't free (Moorfield's versus Deepmind - who contributed most?)
  - How many people comprehend what they are using? (skills shortage)
- Many things aren't "language" – other structures
  - Time series, causal graphs, make life simpler
- Many cheaper AI techniques tackle things without needing a transformer – have some prior model knowledge, and inherent explainability
  - E.g. Phi-ML
  - RL
  - Bayes
  - Relates to Synthetic Data (and GANs)

Model attacks

- Model inversion attack
  - & Set membership inference
- Explainers,interpreters,uncertainty/confidence
  - Cost (Shapley, Integrated Gradients etc)
  - Model collapse risk
  - Not forgetting, forgetting
- Who owns the model, anyhow?
  - Derived work , still legal “grey” area, but
  - probably data lake owner may assert majority ownership…

sustainability

- 2x or even 2000x reduction in electricity/water cost still too bad
  - Inclusion, global south, even just plain business model
- Alternatives abound, so why be lazy?
  - Carbon Emissions and Large Neural Network Training https://arxiv.org/abs/2104.10350
- Controlling ownership, Maintaining model
  - What is the long term biz to fix (vulnerabilities, unlearn etc)
- Access control (AI as service needs protecting)
  - TEE/Enclave/DRM – e.g. h/w locker https://arxiv.org/pdf/2405.20990
  - Or s/w equiv (FHE, zero knowledge systems etc)
  - All increase inference time cost/decrease sustainability

What next?

- Winter…

  https://www.telegraph.co.uk/business/2024/07/15/ai-obsessed-bosses-are-about-to-get-a-rude-awakening/

- Irritating, since (as mentioned) there are many useful AI tools
- And much private useful data too.
- FL/Privacy Preserving learning also gaining traction
- Other challenges include forgetting/unlearning
- TBD: UK AI Regulation (tomorrow☺)
- See also

  https://fishcalledbush.blogspot.com/2024/07/unsustainable-inconceivable.html