OPEN SCIENCE WITH RESPECT TO ARTIFICIAL INTELLIGENCE

Sagnik Mazumder
Computer Science and Engineering
Government College of Engineering and Textile Technology, Berhampore
sagnik.mazumder@gcettb.ac.in

July 20, 2021

ABSTRACT

Artificial Intelligence is one of those fields in computer science that is currently being extensively studied. In this paper, the author attempts to summarise the current state of research in the field with respect to openness to the general community, and has found a profound lack of opportunity to contribute to the field as a novice, and a near monopoly of effective research by large industries while production environments continue to largely remain safe from such influences.

1 Introduction

Artificial Intelligence (AI) is a prominent field in Computer Science and for the past few decades has seen increasing prospects both in research and application. However, considering the wide range of application it provides, it is necessary to constantly verify the interpretability and accessibility of the field in order to prevent a future in which the field becomes approachable to only a handful of people. As such, the field has been to some level biased in its evolution and requires constant regulation it to avoid potential misuse of technology

2 Availability of Code

Research code implementations are important when accountability and reproducibility are considered, especially since many researchers often resolve to overuse of technical jargon to make their paper seem more authentic and thereby increase their chances of getting published.

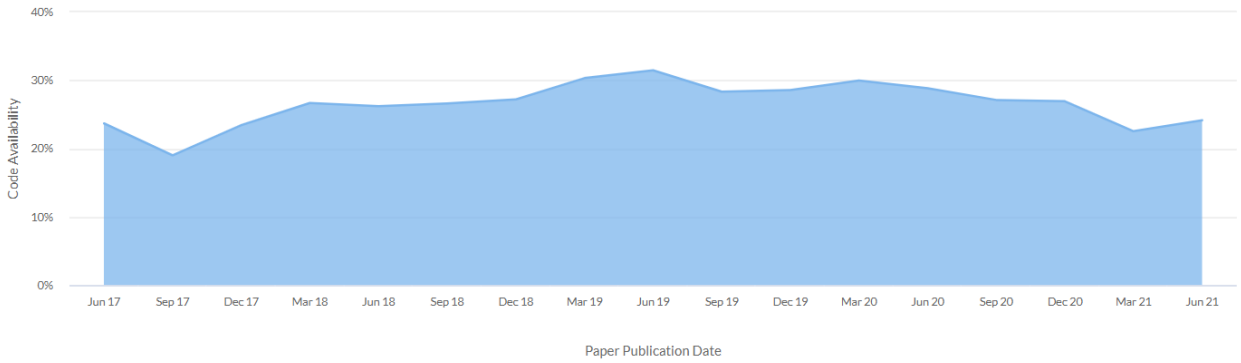

Figure 1: Code availability vs Paper publications by date [1]

However as apparent from Figure 1, a mere 24% as of June 2021 chose to make their code public, a majority of which are often academic groups. Industry groups especially bigger tech companies tend to not publish their code as it is often

intertwined with proprietary scaling infrastructure that cannot be made public. This suggests a centralisation of talent and compute otherwise unmentioned in the industry and unless proper regulation is put into effect, such practices will continue to drive research.

3 Computational Power, Dataset Sizes, and Costs

Since genuine ideas are not easy to come by and everyone wants to publish a paper, a majority of the research is reuse of previously made models with minor changes which may or may not affect performance and scaling the models to larger sizes for a slight increase in performance. However, training such models of billions of parameters often require specialised hardware and substantial compute time which are almost always unavailable to more general researchers.

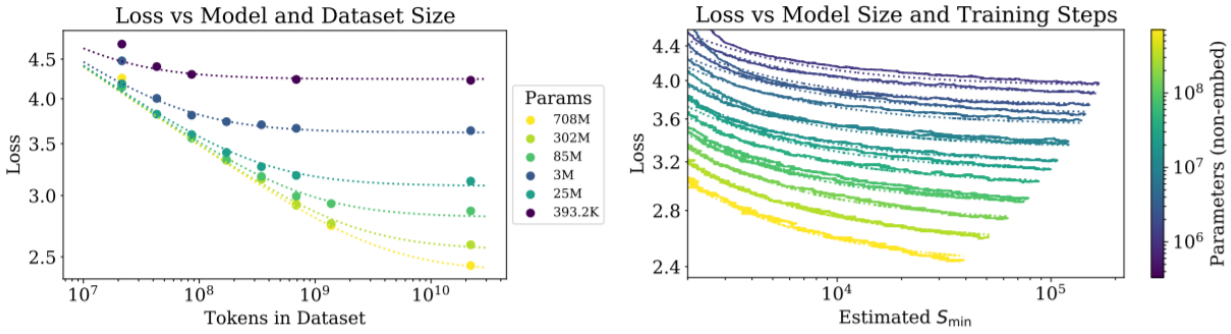

Figure 2: Effect of performance with respect to model size, dataset size, and training steps [2]

On the other hand, smaller models tend to require larger datasets than their larger counterparts to provide similar levels of performance. But except for a few general-purpose datasets, larger datasets are almost exclusively available to large corporations which bar academic researchers access to them. For some use cases however, smaller and data efficient models can outperform larger models with relatively smaller datasets. Also with more modern architectures, deep learning continues to grow more efficient regarding compute even though its data requirements keep increasing.

Along with compute and data requirements, it is important to note, while the cost of floating-point operations (FLOPs) has been decreasing with time due to advanced hardware, usage of larger models means more parameters and combined with larger datasets mean considerably more FLOPs compared to earlier models hence overall cost of training of models keep increasing.

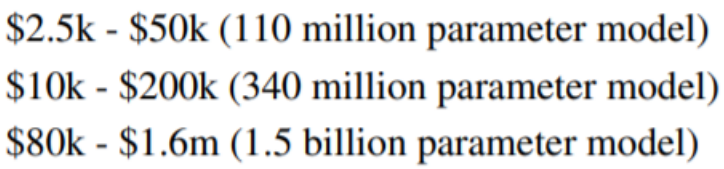

Figure 3: Cost estimate with respect to numbers of parameters [3]

Figure 3 provides an estimate for a single run of Natural Language Processing (NLP) models, the exact figures are proprietary information. This suggests OpenAI's GPT-3 [4] would cost at least $1.9m. Though the cost for computer vision models is relatively less thanks to the less number of parameters required, the reasons for which we will not discuss here, large models still require huge costs to train as estimated in Figure 4. Such costs are obviously unaffordable for many companies and especially academic research teams and startups.

The effects of such requirements are also concerning both economically and environmentally. Without major breakthroughs, and a continuation of the current methodology of often spending millions of dollars for marginal improvements in performance, billions of dollars would be required for any decent improvements which will considerably effect the economy, while running such large computations would also cause significant harm to the environment. An estimate for such costs was provided in a 2019 paper [5], a part of which has been shown in Figure 4.

Benchmark	Error rate	Polynomial			Exponential		
		Computation Required (Gflops)	Environmental Cost (CO_2)	Economic Cost ($)	Computation Required (Gflops)	Environmental Cost (CO_2)	Economic Cost ($)
ImageNet	Today: 11.5%	10^{14}	10^{6}	10^{6}	10^{14}	10^{6}	10^{6}
	Target 1: 5%	10^{19}	10^{10}	10^{11}	10^{27}	10^{19}	10^{19}
	Target 2: 1%	10^{28}	10^{20}	10^{20}	10^{120}	10^{112}	10^{112}
MS COCO	Today: 46.7%	10^{14}	10^{6}	10^{6}	10^{15}	10^{7}	10^{7}
	Target 1: 30%	10^{23}	10^{14}	10^{15}	10^{29}	10^{21}	10^{21}
	Target 2: 10%	10^{44}	10^{36}	10^{36}	10^{107}	10^{99}	10^{99}
SQuAD 1.1	Today: 4.621%	10^{13}	10^{4}	10^{5}	10^{13}	10^{5}	10^{5}
	Target 1: 2%	10^{15}	10^{7}	10^{7}	10^{23}	10^{15}	10^{15}
	Target 2: 1%	10^{18}	10^{10}	10^{10}	10^{40}	10^{32}	10^{32}
CoLLN 2003	Today: 6.5%	10^{13}	10^{5}	10^{5}	10^{13}	10^{5}	10^{5}
	Target 1: 2%	10^{43}	10^{35}	10^{35}	10^{82}	10^{73}	10^{74}
	Target 2: 1%	10^{61}	10^{53}	10^{53}	10^{181}	10^{173}	10^{173}
WMT 2014 (EN-FR)	Today: 54.4%	10^{12}	10^{4}	10^{4}	10^{12}	10^{4}	10^{4}
	Target 1: 30%	10^{23}	10^{15}	10^{15}	10^{30}	10^{22}	10^{22}
	Target 2: 10%	10^{43}	10^{35}	10^{35}	10^{107}	10^{99}	10^{100}

Figure 4: Implications of achieving performance benchmarks on the computation, carbon emissions, and economic costs from deep learning based on projections from polynomial and exponential models. [6][5]

4 Tools of the Trade

Unlike the previously mentioned resources, there is an open availability of tools required for research and application purposes. A majority of the libraries used are open sourced, and while there is a low rate of publication of code, models are often made available enabling one hands on usage of models. The computation power, dataset size and cost problems can be somewhat mitigated, mostly for application purposes, by using transfer learning. PyTorch and TensorFlow provide extensive collections of pretrained models which can be deployed directly, fine tuned to a use case, or re-purposed for a completely new objective. Hence while research remains difficult to access, building applications is easy with the freely available tools.

A somewhat odd problem remains regarding the interpretability of models. As focus has shifted through the years from purely statistical methods to more black-box methods, models are often found to be unreliable. Misinterpretation can mean models learning wrong information from given data, which may give benchmark results in development but are not safe to use in production environments, this is an architectural problem. A dataset problem in this regard is finding or making the correct dataset to maximize learning with smallest possible data points, while cross domain datasets sometimes improve learning as observed in transfer learning techniques, reliability of such models need to be thoroughly ensured. It is often found models are learning by reference instead of reasoning.

5 Conclusion

While the number of research papers generated in AI remain high, a large number of them continue to remain incomprehensible and redundant, their implementations unreleased, and thereby results often irreproducible. With the current techniques, improving benchmark performances remains difficult without a huge economic and environmental cost. But since tools to research modern techniques continue to remain in the open source, the current situation in this regard is not irreparable, AI also continues to remain open as ever in regard to applications and with the increasing interest in improving ethics and interpretability, this trend promises an positive curve at least in the near future.

References

[1] Papers with code : Trends | papers with code, 2021.

[2] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[3] Or Sharir, Barak Peleg, and Yoav Shoham. The cost of training nlp models: A concise overview, 2020.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp, 2019.

[6] Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. The computational limits of deep learning, 2020.