Environmental Consequence of Deep Learning

Damian Borowiec, Richard R. Harper, and Peter Garahan discuss Sustainable Deep Learning

Since 2010s, Artificial Intelligence (AI) and Deep Learning (DL) have been accelerating the digital transformation of technologies, industries, and society at an unprecedented pace, with projected market size in 2028 reaching ~£116b globally and the rate of growth (CAGR) increasing from 20% (2019) to 40%. This is due to the opportunities for R&D enabled by DL across automotive, healthcare, defense, and other industries. With the backdrop of climate change, greenhouse gas emissions and scarcity of fossil fuels looming over societies, more action can be observed in academia, industries, and individuals in attempt to address these problems, with AI/DL often positioned as the panacea. Many commentaries paint ambitious futures for AI/DL as eventually net-zero tools for net-zeroing other industries. However, we are now facing a chicken-and-egg problem where the accelerated growth is increasing environmental costs of AI, with a single DL Transformer model involving ~300t of CO2 being released into the atmosphere during its training (Emma Strubell, Ganesh, & McCallum, 2019). With thousands of new models developed every year and their numerous instances operating simultaneously throughout industries, their impact becomes globally significant.

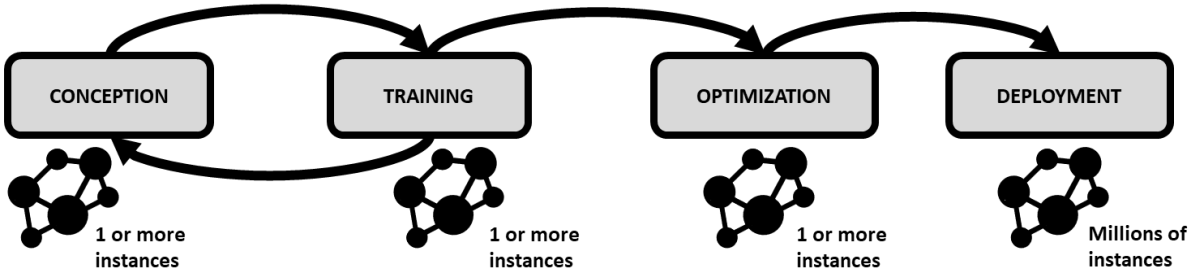
Popular commentaries often omit current practicalities of computing systems that make DL possible, in favour of imaginative future scenarios, lacking in depth analysis of the interplay between DL's ability to provide intellectual value and the associated costs. Such black box view of DL obscures the understanding of the positive (and measurable) affect AI could have in the future, in terms of creative, economic, and social well-being, yet importantly, the inevitable environmental costs of DL itself. Hence, we need concentrated efforts to combine understanding of the bleeding edge of DL's technical operation with the drivers such as research agendas, policies, and the economy.

Deep Neural Networks and computation:

DL is underpinned by Deep Neural Networks (DNNs) - virtual constructs responsible for the surge in AI developments, capable of big data (images, text, numbers) processing to extract their progressively abstract features, learning their approximate representation. At a fundamental level, DNNs are series of instructions, akin to any other software, and are thus governed by similar assumptions about performance or costs – the more abstract the features, the more computation is required to utilize them for DL inference. For example, recognizing objects in a digitized painting is more computationally taxing than recognizing letters in hand-written text. At a functional level, DNNs mimic brain's operation by representing connections between artificial neurons using *weights* – N-dimensional structures of floating-point numbers residing in RAM. Artificial neurons combine their inputs, the *weights,* and an activation function to determine their *firing*, and pass on the result to subsequent network neurons – out comes the answer at the end. During training, *weights* are iteratively updated to progressively better represent the discovered data features. Such DNN functionality leverages calculus, algebra and probability theory typically expressed as matrix operations (transposition, dot-product), enabling modern CPUs and accelerators (GPUs, TPUs, NPUs) to process them – with the latter often providing dedicated matrix-oriented circuitry for improved performance. In the end, a high-level structure of a DNN is boiled down to a small, repeating set of computer instructions.

Costs and efficiencies of DL operation:

Whether operating on server-grade CPUs, accelerators or mobile devices, DL workloads rely upon transistors, which represent logical 1s and 0s by switching electrical current from high to low. Simplistic in their isolated behaviour, when grouped, transistors construct adders, multipliers, and control units – the building blocks of modern hardware – which consist of as many as ~54b transistors. The switching activity of transistors dissipates (consumes) energy, with the extent dependent on the number of engaged transistors and the switching speed, dictated by the clock. To process more complex DL workloads, representing many abstract data features, either more engaged transistors or more time are required, with both increasing the total energy consumption of the circuit proportionally. DL workload efficiency can often be improved via parallel computing, due to the multidimensional nature of DNNs, packing more computation in the same time quantum. However, whilst improving DL efficiency locally, parallelism and other solutions are not immune to the economies of scale that continue to impact DL's global computational requirements and costs, which have grown 300,000-fold from 2012 to 2018 and keep doubling every 3-4 months.



Adulthood of DL models:

DL is not just training – models have a complex life cycle, which begins with experts determining the purpose of the model, training data, network structure and the hyperparameters deciding how training should progress. Once trained, DL models are deployed in various applications, operate on different devices, and distil insights from datapoints similar-to but not identical to the ones they trained on. As such, DL models could be thought of as replicas of the same brain, living in various environments where they are exposed to different information reaching their sensory inputs. Each such instance requires computation and thus consumes energy. While environmental impact of training the *instance-zero* can be equivalent to driving 5 mid-sized vehicles for a year (Transformer model), the in-field impact of its deployment for inference has orders of magnitude larger implications, with potentially millions of replicas operating around the world for unpredictably long time, consuming vast amounts of energy. Whilst the development of more efficient compute infrastructure for DL is continuing, there are more barriers to achieving DL sustainability and enabling DL to have positive impact on other industries, that are societal and economically systemic in nature, with growth and demand both cross cutting the problem.

The Rat Race:

Higher DL model complexity enables inference of more abstract data features, whilst longer training improves accuracy. Thus, the default method of developing a "better" model is to throw computational resources at it, moving up the DL leader board hierarchy. Such mentality leads to DL models outperforming the needs of their intended application areas, with development costs

seldom reported, as measuring them is difficult. Globally, DL growth may seem uncontrollable and without concern for cost, however, we posit that it occurs not due to “lavishness” or competitiveness of experts but the lack of clearly defined concepts of DL’s intellectual value. Elementally, the simplest way to compete is to be stronger or faster, given no other assumptions about the competition. We believe growth and global environmental cost of DL are exactly the assumptions to be had.

Hidden Costs of Deep Learning:

Given two vastly different devices (Cloud server and mobile phone) executing the same DL model, their energetic footprint can be quite different given their chip design and operating assumptions (e.g., battery-powered vs. constant access to electricity). This is exacerbated by the fact that even insignificant changes to the model design during initial stages, can dramatically impact costs of training and importantly, in-field operation. I.e., changing the layout of the model data after training, changes the cache access patterns and (co)processor scheduling decisions during inference, potentially affecting operating cost. Examining these relationships is difficult from the perspective of DL designers as well as engineers, both due to gaps in common knowledge and the black box nature DNNs, also preventing management from corelating organization-level issues such as environmental costs to the practicalities of product operation. DL model explainability and adequate (non-gratuitous) intellectual capabilities should be the paramount driving forces to enable more sustainable AI.

Budgeting Cognition:

DL models can be thought of as distilleries, digesting data and discovering abstract features within them, enabling inference. The value provided by an oil distillery is the function of its ability to decompose crude oil into petrol and other substances, quantifiable by the crude oil volume and processing rate - a clear value added. DL model operation can also be quantified by its ability to distil insights from data - the inference speed, throughput, and related computation costs – a clear relationship between insights and energy cost. This however does not determine its intellectual value. For example, an DL CCTV system may process 5m HD images/day at a busy car park, identifying vehicles at 99% accuracy, yet these inferences are only used during parking policy violations (0.2% of observed time) or serve as statistical summary of vehicle count. The value added of such system is not easily quantifiable and must be determined as a complex function of the system in the context of the business, community, actors, and economies. In other words, we must acknowledge the inherent complexity of AI technologies and re-examine their intellectual value added, given their monetary and environmental costs.

With clear definitions, many of the inefficiencies or unaccounted-for costs could be alleviated or reduced, and the role of economies of scale could be reversed to drastically improve DL’s environmental impact. Overall, more attention towards the complex DL life cycle and the apparent and hidden costs, could help to achieve sustainable DL in a robust and long-term way.

Bibliography

Emma Strubell, Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645-3650). Florence, Italy: Association of Computational Linguistics.