International Journal of Advance Research in Science and Engineering
Volume No. 10, Issue No. 08, August 2021
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Impact of AI on Environment

[1]Lokesh Singh, [2]Harshvardhan Dwivedi, [3]Bhovnit Singh,
[4]Harshvardhan Saxena
[1](lokeshsingh899@gmail.com), [2](vardhanharsh0401@gmail.com)

Abstract
Nowadays we see and use various AI(artificial intelligence) solutions in our daily life for solving a myriad of day-to-day problems. Various fields are being found and researched on where artificial intelligence can be used for making systems more efficient but do we know about the cost our environment pays for our use of these solutions implemented using AI and machine learning. The computational cost of the AI solutions should be considered too. We introduce you to these costs incurred in the environment and some solutions to minimize them.

1 Introduction

Artificial Intelligence is claimed to be the simulation of human intelligence within the machines which are programmed to think like humans and may mimic their actions. This term can also be applied to a machine that can exhibit traits such as problem solving and learning new things.

Artificial Intelligence has various applications in today's society, many industries such as healthcare, education, finance, entertainment have implemented AI for solving different complex problems. As it can solve various complex problems efficiently it has made our daily life easy and comfortable.

But while we use artificial intelligence to solve these problems, have we ever considered what cost must the environment pay for implementation of such AI based solutions. There are many areas and fields related to artificial intelligence that we are expanding our research to but have we thought about the carbon footprint it leaves. One might not think that it could cost much to our environment but have we ever tallied the costs.

In order to achieve accuracy in the machine learning models, they are trained repeatedly without much consideration of the computational costs as there are resources available right now.

How much does it cost?

For quantification of computational cost, a life cycle assessment of many AI models was performed at researchers at Massachusetts. This computational cost in terms of CO2 emissions came out to be five times the average lifetime CO2 emissions of an American car.

AI Applications in Environmental Sector

Many MNCs such as *Microsoft, Google and Tesla*, whilst pushing the limits for man's innovations, have made considerable efforts in developing 'Eco Friendly' AI systems. For instance, Google's very own *DeepMind* AI has helped the organization to curb their data center energy consumption by 40 percent making them more energy efficient and decreasing overall GHG emissions. As data centers alone consume 3% of global energy every year, development of such AI's not only improve the energy efficiency but also assist in providing energy access to remote communities, setting up microgrids and integrating renewable energy resources.

Installation of smart grids in cities can utilize artificial intelligence techniques to regulate and control parts of neighborhood power grid to deliver exactly the amount of electricity needed, or requested from its dependents, against the use of conventional power grids that can be wasteful due to unplanned power distribution.

2 Methods

From what we know to quantify the computational and environmental cost of coaching deep neural network models for NLP, we perform an analysis of the energy required to coach a spread of popular off the-shelf NLP models, as well as a case study of the complete sum of resources required to develop LISA, a state-of-the-art NLP model from EMNLP 2018, including all tuning and experimentation. We measure energy use as follows. We train the models described in fig.2.1 using the default settings provided, and sample GPU and CPU power consumption during training. Each model was trained for a maximum of 1 day. We train all models on a single NVIDIA Titan X GPU, with the exception of ELMo which was trained on 3 NVIDIA GTX 1080 Ti GPUs. While training, we repeatedly query the NVIDIA System Management Interface2 to sample the GPU power consumption and report the average over all samples. To sample CPU power consumption, we use Intel's Running Average Power Limit interface.

International Journal of Advance Research in Science and Engineering
Volume No. 10, Issue No. 08, August 2021
www.ijarse.com
IJARSE
ISSN 2319 - 8354

2.1 The Models

We analyze four models, the computational requirements of which we describe below. All models have code freely available online, which we used out-of-the-box. For more details on the models themselves, please refer to the original papers. Transformer. The Transformer model is an encoder-decoder architecture primarily recognized for efficient and accurate machine translation. The encoder and decoder each consist of 6 stacked layers of multi-head self attention. Transformer base model (65M parameters) was trained on 8 NVIDIA P100 GPUs for 12 hours, and the Transformer big model (213M parameters) was trained for 3.5 days (84 hours; 300k steps). This model is also the basis for recent work on neural architecture search (NAS) for machine translation and language modeling , and the NLP pipeline that we study in more detail. So their full architecture search ran for a total of 979M training steps, and that their base model requires 10 hours to train for 300k steps on one TPUv2 core. This equates to 32,623 hours of TPU or 274,120 hours on 8 P100 GPUs.

ELMo. The ELMomodel is based on stacked LSTMs and provides rich word representations in context by pre-training on a large amount of data using a language modeling objective. Replacing context-independent pre-trained word embeddings with ELMo has been shown to extend performance on downstream tasks like as named entity recognition, participant role labeling, and coreference. ELMo was trained on 3 NVIDIA GTX 1080 GPUs for 2 weeks (336 hours) .

BERT. The BERT model provides a Transformer-based architecture for building contextual representations similar to ELMo, but trained with a different language modeling objective. BERT substantially improves accuracy on tasks requiring sentence-level representations such as question answering and natural language inference. BERT base model (110M parameters) was trained on 16 TPU chips for 4 days (96 hours). NVIDIA reports that they can train a BERT model in 3.3 days (79.2 hours) using 4 DGX-2H servers, totaling 64 Tesla V100 GPUs.

GPT-2. This model is the latest edition of OpenAI's GPT general-purpose token encoder, also based on Transformer-style self-attention and trained with a language modeling objective. By training a really large model on massive data, Radford et al. (2019) show high zero-shot performance on question answering and language modeling benchmarks. The large model described in Radford et al. (2019) has 1542M parameters and is reported to need 1 week (168 hours) of training on 32 TPUv3 chips.

Consumer	**Renew.**	**Gas**	**Coal**	**Nuc.**
China	22%	3%	65%	4%

Germany	40%	7%	38%	13%
United States	17%	35%	27%	19%
Amazon-AWS	17%	24%	30%	26%
Google	56%	14%	15%	10%
Microsoft	32%	23%	31%	10%

Table 1: Percent energy sourced from: Renewable (e.g. hydro, solar, wind), natural gas, coal and nuclear for the top 3 cloud compute providers, compared to the United States,4 China 5 and Germany.

3. Experimental results

3.1 Cost of training

Model	Hardware	Power (W)	Hours	kWh·PUE	CO2e	Cloud compute cost
Transformer base	P100x8	1415.78	12	27	26	€34–€116
Transformer big	P100x8	1515.43	84	201	192	€239–€814
ELMo	P100x3	517.66	336	275	26	€359–€1221
BERT base	V100x64	12,041.51	79	1507	143 8	€3113–€10433
BERT base	TPUv2x16	—	96	—		€1721–€5736
NAS	P100x8	1515.43	274,120	656,347	626,155	€782611–€2657240
NAS	TPUv2x1	—	32,623	—	—	€36563–€121875
GPT-2	TPUv3x32	—	168	—	—	€10707–€35694

Table 3: lists CO2 emissions and estimated cost of training the models described in 2.1. Of note is that TPUs are more cost-efficient than GPUs on workloads that make sense for that hardware (e.g. BERT). We also see that models emit substantial carbon emissions; training BERT on GPU is roughly equivalent to a trans-American flight. So report that NAS achieves a new state of-the-art BLEU score of 29.7 for English to German machine translation, an increase of just 0.1 BLEU at the cost of at least $150k in on-demand compute time and non-trivial carbon emissions.

3.2 Cost of development: Case study

To quantify the computational requirements of R&D for a new model we study the logs of all training required to develop LinguisticallyInformed Self-Attention, a multi-task model that performs part-of-speech tagging, labeled dependency parsing, predicate detection and semantic role labeling. This model makes for an interesting case study as a representative NLP pipeline and as a Best Long Paper at EMNLP. Model training related to the project spanned a period of approx 6 months. During that point 123 small hyperparameter grid searches were performed, leading to 4789 jobs in total. Jobs varied long starting from a minimum of three minutes, indicating a crash, to a maximum of 9 days, with a mean job length of 52 hours. All training was done on a combination of NVIDIA Titan X (72%) and M40 (28%) GPUs.8 The sum GPU time required for the project totaled 9998 days (27 years). This averages to about 60 GPUs running constantly throughout the duration of the project. Table 4 lists upper and lower bounds of the estimated cost in terms of Google Cloud compute and raw electricity required to develop and deploy this model.9 We see that while training one model is comparatively inexpensive, the value of tuning a model for a replacement dataset, which we estimate here to need 24 jobs, or performing the full R&D required to develop this model, quickly becomes extremely costly.

Estimated cost (USD)

Models	Hours	Cloud compute	Electricity
1	120	\$52–\$175	\$5
24	2880	\$1238–\$4205	\$118
4789	239,942	\$103k–\$350k	\$9870

Table 2: Estimated cost in terms of cloud compute and electricity for training: (1) a single model (2) a single tune and (3) all models trained during R&D

.

Conclusion

With the deteriorating environment we would not want AI and machine learning to add more to it. Thus we must find solutions for minimizing the computational cost of these models. further measures that can be adopted to do so can be as follows

Think before you train, we must always think first before training a model if the training of the model will achieve desired results, training models repeatedly to achieve accuracy by a very small measure may incur very high costs thus we must always consider the feasibility keeping in mind the computational cost.

Better optimizing of datasets, we must find ways for better optimising of datasets which could further decrease the computational cost of the model.

Tallying the costs, we must calculate the costs associated with the models for better determination if the feasibility of the model training.

Promotion of usage of alternative and renewable resources, the more renewable resources are used instead of the non-renewable resources the more we can bring down the harmful impact of AI and ML computational costs on the environment.

Tiny AI

The next advancement in AI and ML is tiny AI, more research should be promoted on tiny AI as this solution is made to dramatically reduce the size of the algorithms which can be built on small hardware or on devices with low power consumption thus ultimately reducing the threat to nature.

References:

1. https://en.m.wikipedia.org/wiki/Artificial_intelligence
2. https://www.forbes.com/sites/forbestechcouncil/2020/08/17/why-we-should-care-about-the-environmental-impact-of-ai/amp/
3. Artificial Intelligence Tutorial - Tutorialspoint
4. https://paperswithcode.com/paper/linguistically-informed-self-attention-for
5. https://turcomat.org/index.php/turkbilmat/article/download/4547/3844
6. https://arxiv.org/pdf/1906.02243
7. https://www.semanticscholar.org/paper/AI-Tax%3A-The-Hidden-Cost-of-AI-Data-Center-Richins-Doshi/4f5e12bee7ea0357dac0b05b1cf3b7ba6439795d
8. https://scholar.harvard.edu/files/pakes/files/ai_v19.pdf
9. http://scholar.google.com/citations?view_op=search_authors&hl=en&mauthors=label:artificial_intelligence
10. https://link.springer.com/article/10.1007/s43681-020-00007-2
11. https://www.nber.org/system/files/working_papers/w24449/w24449.pdf