ENVIRONMENTAL COSTS OF MACHINE LEARNING ALGORITHMS: REDUCING UNNECESSARY COMPUTATIONS

A Research Paper submitted to the Department of Engineering and Society
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Surbhi Singh

March 27, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Catherine D. Baritaud, Department of Engineering and Society

Machine Learning is a subset of Artificial Intelligence (AI), which is the broader concept of intelligent machines. With the recent rise in Artificial Intelligence research, breakthroughs have been made in natural language processing, image recognition, gaming, and much more. However, with all larger models, more data is required for training and thus creates more computationally expensive algorithms. The technical project aims to map invasive plant species using machine learning. The STS research is loosely coupled with the technical portion as it analyzes the environmental impacts of training the types of complex machine learning algorithms being used in the research. The STS research could potentially impact the methodology used in the research to reduce the carbon footprint of training large ML algorithms. Pacey's Triangle was used to investigate the barriers to developing more efficient and environmentally friendly algorithms which aim to reduce unnecessary computations.

Although AI has become a very popular tool in solving today's problems, most people fail to acknowledge the environmental impacts. The research here aims to analyze what factors are preventing the wide adoption of environmentally friendly algorithms.

RISING COSTS OF ARTIFICIAL INTELLIGENCE RESEARCH

The rapid pace of artificial intelligence growth has created significant progress in many areas, but also increased costs in many areas Researchers at the University of Massachusetts found that several common AI models can emit CO_2 in ranges from 39-626,155 lbs. For reference, an average car emits 126,000 lbs. in its lifetime (Strubell, 2019). These financial and environmental costs are much higher in research which requires retraining of model architecture and parameters. AI models are usually trained using data centers, which are large contributors to carbon emissions. Some argue that computationally expensive algorithms are not a pressing issue at the moment (Biewald, 2019). While it is true that these algorithms currently comprise a very

small percentage of current data center power usage, the amount of computations is increasing exponentially and will likely become an issue in the near future. The amount of computations run by deep learning research was estimated to have increased by 300,000 times from 2012 to 2018 (Amodei and Hernandez , 2018). If the growth continues to grow exponentially, action must be taken soon before the environmental impacts become detrimental.

STS FRAMEWORK: PACEY'S TRIANGLE

This study investigates the question of how to incentivize researchers to develop more efficient and environmentally friendly algorithms. Efficient algorithms are both cheaper and have a lower carbon footprint, yet efficiency is not prioritized in current research. Pacey's Triangle can be used to understand what cultural, social, and technological factors are inhibiting the adoption of environmentally friendly AI algorithms (1983, p. 6). This framework is chosen because it can highlight the different groups involved in the problem and their relationships.

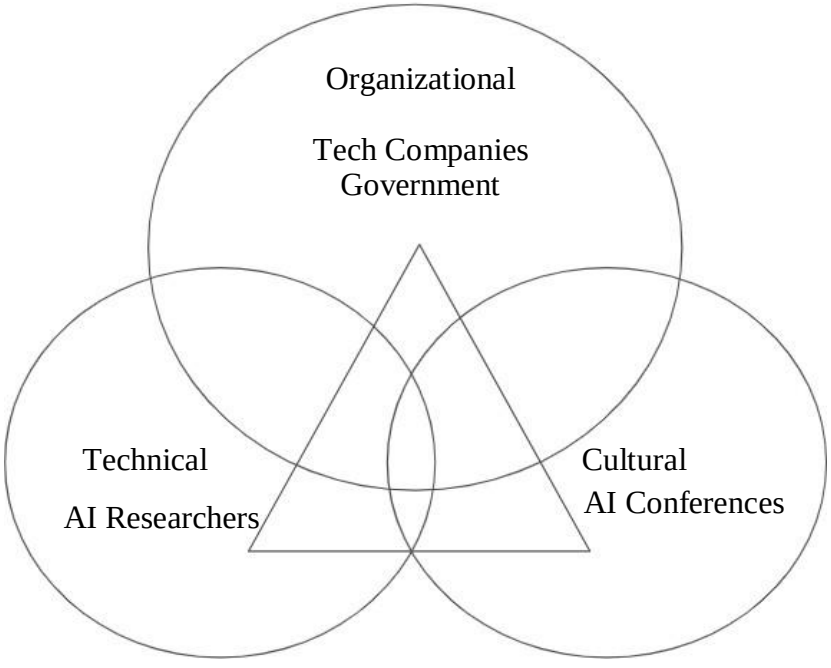

Figure 1 to the left shows Pacey's Triangle for the actors involved in AI research. The organizational aspect includes big tech companies and the government who decide what types of projects to fund. The technical portion are the machine learning researchers as they are the ones who chose what methodologies should be used. The cultural aspect is what measurements

Figure 1: Pacey's Triangle for the creation of computationally expensive ML algorithms: An illustration of the three aspects from Pacey's (adapted by Singh from Pacey, 1983)

are deemed important by top AI conferences and publishers. This heavily impacts what researchers focus on during development. The three aspects will be expanded upon in the following sections.

CULTURAL INFLUENCES: CHANGING RESEARCH PRIORITIES

Currently, the pressure to get a paper accepted into a top conference or publish is the cultural barrier to the wide adoption of efficient AI. The field of Artificial Intelligence has been designed after the human brain, which is very efficient. However, most AI algorithms are quite inefficient, performing many more computations than required. To reduce the exponentially increasing carbon emissions that have risen with increasingly complex models, several theories have been proposed. Along with the accuracy and cost of AI models, the efficiency of the algorithm should be strongly considered for research purposes. The pressure to get a paper accepted to top conferences causes researchers to excessively train models for improved accuracy, disregarding computational costs. Figure 2 to the right shows the ratio of papers from top AI conferences that report accuracy rather than efficiency. Some top conferences such as NeurIPS 2019 and EMNLP 2020 have taken steps in the right direction to require a computational budget in paper submissions.

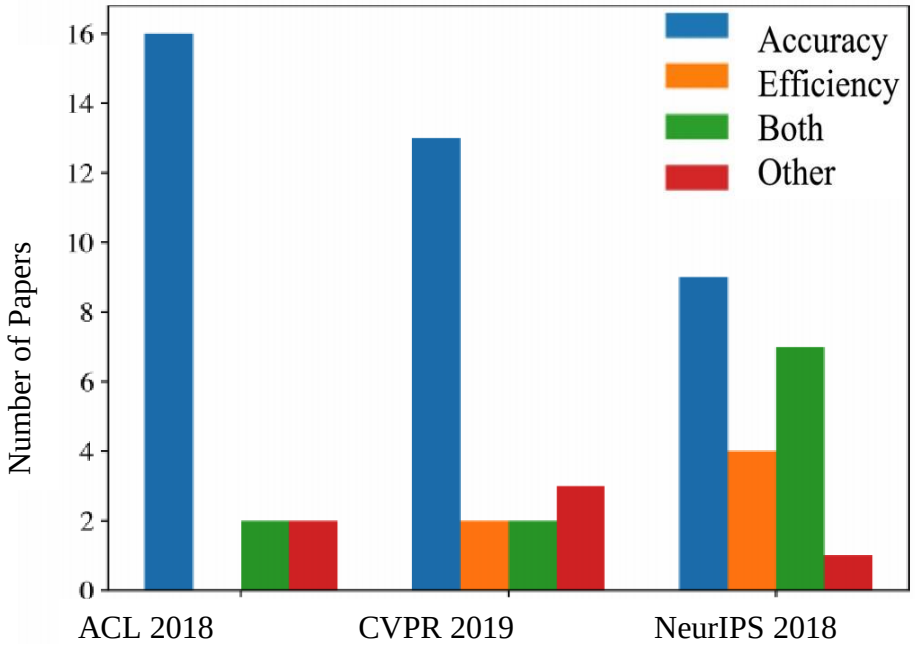

Figure 2: Proportion of AI papers that reported accuracy and efficiency from a sample of 60 papers at renowned AI conferences. (Adapted from 'Green AI' by R. Schwartz, J Dodge, N. Smith, & O. Etzioni, 2019)

Green AI

Researchers at the Allen Institute for Artificial Intelligence Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni recently proposed a novel idea known as Green AI. The article classifies algorithms as 'Red AI' and 'Green AI' depending on the amount of computations required. Green AI includes novel research which is environmentally friendly and is considerate of the amount of resources required. This is often achieved by prioritizing the efficiency of algorithms rather than just the accuracy. Red AI is the opposite; computationally expensive and often sacrifices large amounts of efficiency for small accuracy gains. The purpose of this research is not to diminish the importance of Red AI, but to understand why it is so prevalent and how to shift towards more Green AI.

ORGANIZATIONAL POWER: FUNDING THE RIGHT PROJECTS

The government invests a large amount of money on AI research every year. To maintain the nation's technological edge, the Fiscal 2020 budget for AI and ML research is about $1 billion (Cornille, 2019). Currently, the government tends to choose to fund projects which are producing the best results in terms of performance and accuracy. By implementing new priorities including imposing restrictions on total computations performed or requiring some standard of efficiency for government funded projects Green AI, researchers would be encouraged to adapt better methodologies.

TECHNICAL OPERATIONS: MEASURING COMPUTATIONS

Many options were explored to determine a way to quantify computational cost including GPU usage time, carbon emissions, and number of parameters used to train a model. However, these measures do not allow for a fair comparison among models as they are largely dependent on the type of infrastructure and underlying hardware being used. Efficiency analysis can use the total

number of floating point operations (FPO) to accurately compare across different models. FPO allows for comparison amongst different researchers because it does not involve the hardware or electricity usage. Any abstract machine learning operations can be represented by ADD or MUL operations at the hardware level, and will be the same on any type of hardware (Schwarz, 2020). In addition, FPO is strongly associated with run time. However, FPO does not take memory consumption into account, which can affect the amount of energy used. While imperfect, FPO is the fairest measurement among different models.

Researchers use many different methods to increase a model's accuracy including using more complex models, training models on more data, and running more experiments. This is often referred to as hyperparameter tuning. Each of these ways increases the energy and financial

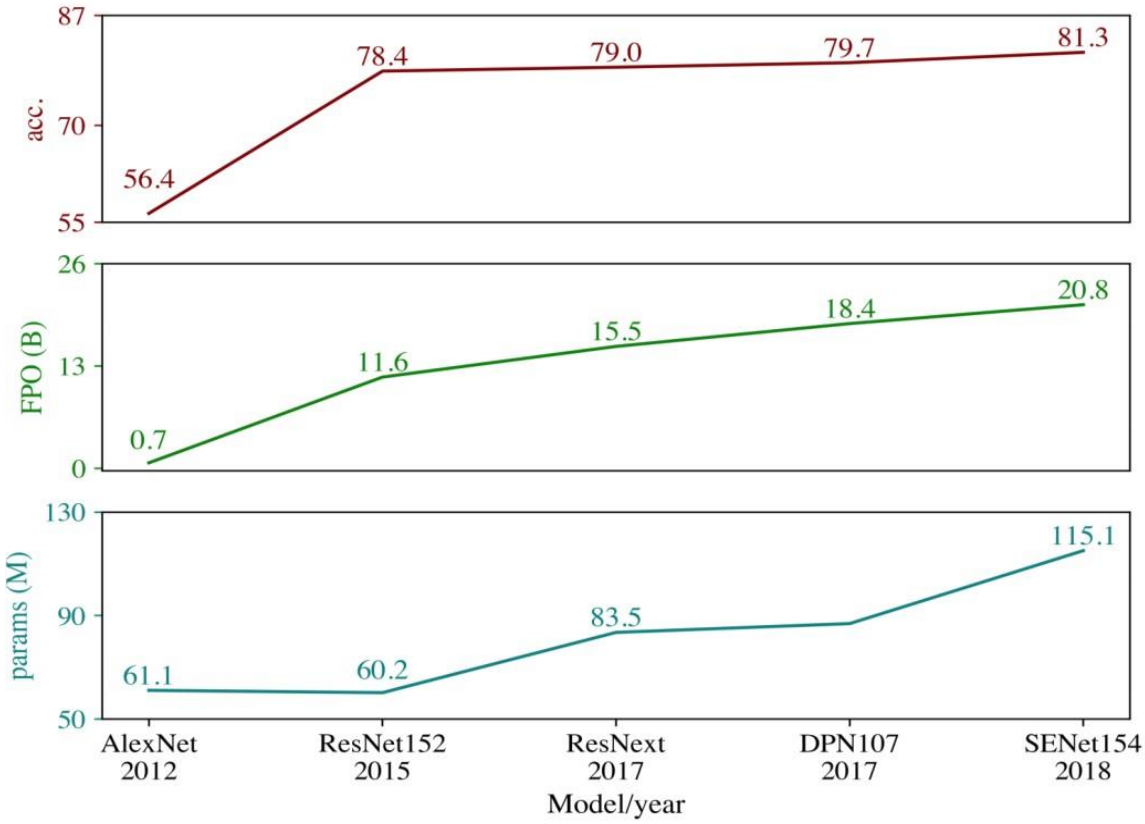

Figure 3: Floating Point Operation(FPO) in billions for different top ImageNet models. This shows how the rate accuracy increases much less steeply than the rate of the FPO increase. (Adapted from 'Green AI' by R. Schwartz, J Dodge, N.Smith, & O. Etzioni. 2019)

costs, often resulting in Red AI. The problem here is to understand at what point is the accuracy gain no longer worth the increased costs. Usually, the accuracy gains level out and minimal improvement are seen as number of computations increases. As shown in Figure 3 below, a 33.6% increase in FPO resulted in just a 0.76% increase in accuracy (Schwarz, 2020). Thus, holistically, ResNet can arguably be seen as a superior algorithm to ResNext, despite its lower accuracy.

The big question now becomes how to determine when to reduce or stop training these models if such a plateau in accuracy exists. If the new model is open source and being reused by several other researchers, it may be worth investing lots of resources to train it to its highest accuracy potential. However, researchers should be aware of the costs of Red AI and be mindful of unnecessary computations.

CASE STUDY: OPEN SOURCED AI MODELS

While some AI models are publicly available for reuse, retraining of unavailable models is one of the big costs in AI training. Open sourced code is software that is available to anyone to use and modify freely. There are several reasons why some researchers choose not to make their models open-sourced. This case study will focus on one complicated model developed throughout 2019 by Open AI to analyze the benefits and dangers of releasing open source code. The second Generative Pre-Training model (GPT-2) contains over 1.5 billion parameters, and is the most advanced text generation and language generating technology we currently have (Vig, 2019). In order for a model to learn something as complicated as language generation, it must be extremely large and requires training of specific tasks. Fine-tuning the tasks is particularly energy consuming as it involves a lot of trial and error.

Researchers at Open AI warn about several malicious uses of a powerful AI text generator. The GPT-2 could be used to generate false news articles, impersonate fake user accounts, automate phishing/spamming, and other deceptive language models on a large scale (Radford et. al., 2019) While the OpenAI engineers claim that the model was not released for several months because of these potential dangers, this does not make sense given the fact that the model training could be replicated given all the information was publicly available. Radford et al. stated that "We are aware that some researchers have the technical capacity to reproduce and open- source our results. We believe our release strategy limits the initial set of organizations who may choose to do this, and gives the AI community more time to have a discussion about the implications of such systems" (Radford et. al., 2019, para. 15). Therefore, this simply limits access to those who don't have the resources and money to train such a large model.

Initially, OpenAI did not release any parts of the pre-trained model. Then, over the course of 6 months, a small, medium, and large version of the model was released, but never the full thing. This strategy was used to assess the implications of the new technology on society and evaluate any misuse. No concerning amounts of misuse were detected with the slow release. The GPT-2 could be used to set the precedent for establishing policy for the release of any open source code by creating sharing safety and research standards. The government should also take action to help monitor the diffusion of AI technologies into society to measure progress and identify potential problems. Big corporations and researchers should be encouraged to make computationally expensive models open source, especially if others are likely to retrain to get the same results. Retraining a model with 1 billion parameters is simply a waste of time and resources. The GPT-2 requires 168 yours of training on 32 TPUv3 chips, which amount to a cost between $12,902 and $14,003 of cloud compute in data centers (Strubell, 2019).

While the carbon footprint of TPUs are unknown for this model, the amount of CO_2 emitted in pounds per kilowatt-hour are shown in Figure 4 below for several other popular models.

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
Transformer$_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer$_{big}$	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT$_{base}$	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT$_{base}$	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Figure 4: Estimated costs of training popular AI models in terms of CO_2 emissions and cloud compute cost (USD). (Adapted from 'Energy and Policy Considerations for Deep Learning in NLP by E. Strubell, A. Ganesh, & A. McCallum, 2019)

While these powerful AI models have the potential to do good, especially if available to anyone, we must also remain cautious and anticipate ethical and social repercussions that accompany any technology, especially such powerful ones.

IMPACT

Changing the culture of AI conferences, imposing restrictions by the government, and adopting other Green practices will not only reduce the carbon footprint, but they will also reduce the financial costs of training complex AI models. This would allow a larger group to participate in the development of AI algorithms and promote inclusivity. Rather than requiring high computing power resources, anyone with a laptop should be able to run machine learning algorithms that could be presented at top conferences. The goal for this paper is to raise awareness in the AI community about the environmental and financial costs of large AI models in hopes that researchers will take

efficiency into consideration while developing new models. This tradeoff between prediction accuracy and prediction cost may add another layer of analysis to both old and new algorithms. In addition, this research aims to identify methods of quantifying and reducing the overall computing costs of machine learning algorithms. Current Machine Learning researchers are blindly wasting a lot of computing resources on redundant training, with little to no accuracy gains. While Red AI has made significant strides in research, it is overly prevalent. This research does not dismiss the importance of computing intensive algorithms, but instead urges researchers to take a holistic view and not sacrifice some areas for minor improvements in others. Machine Learning has endless potential to solve complex problems, but unless new methodology is adopted by researchers, the environmental impacts could soon be disastrous.

WORKS CITED

Amodei, D., & Hernandez, D. (2019, December 12). AI and Compute. [Web log post]. Retrieved from https://openai.com/blog/ai-and-compute/

Cai, F. (2019). Greening AI: New AI2 initiative promotes model efficiency. Retrieved from https://syncedreview.com/2019/07/31/greening-ai-new-ai2-initiative-promotes-model-efficiency/

Cornillie, C. (2019). Finding artificial intelligence money in the fiscal 2020 budget: BGOV. Retrieved from https://about.bgov.com/news/finding-artificial-intelligence-money-fiscal- 2020-budget/

Hao, K. (2019). Training a single AI model can emit as much carbon as five cars in their lifetimes. Technology Review, Retrieved from https://www.technologyreview.com

Devulapalli, H. (2017). The ethics of open data in closed societies. Retrieved from https://medium.com/open-and-shut/conference-notes-ethics-of-open-data-in-closed-societies-4d5bbfac8a7d

Ferri, B. (2019). The democratization of artificial intelligence (AI) for Data Science. Retrieved from https://www.iotforall.com/democratization-of-artificial-intelligence/

Horev, R. (2018). BERT Explained: State of the art language model for NLP. Retrieved from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Pacey, A. (1983). The Culture of technology. Cambridge, MA: MIT Press.

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019, February 14). Better language models and their implications. [Web log post].Retrieved from https://openai.com/blog/better-language-models/

Schwartz, R. (2020, February 17). Green AI. [Web log post]. Retrieved from https://medium.com/ai2-blog/green-ai-db24a414a7a4

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2019). Green ai. arXiv preprint, Retrieved from https://arxiv.org/abs/1907.10597

Strubell, E., Ganesh, A., McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint, Retrieved from https://arxiv.org/abs/1906.02243

Vig, J. (2019). OpenAI GPT-2: Understanding language generation through visualization. Retrieved from https://towardsdatascience.com/openai-gpt-2-understanding-language-generation-through-visualization-8252f683b2f8

BIBLIOGRAPHY

Biewald, L. (2019). Deep Learning and carbon emissions. *Towards Data Science,* Retrieved from https://towardsdatascience.com/deep-learning-and-carbon-emissions

Bradley, B. A. (2013). Remote detection of invasive plants: A review of spectral, textural and phenological approaches. *Biological Invasions, 16*(7), 1411–1425. doi: 10.1007/s10530-013-0578-9

Cai, F. (2019). Greening AI: New AI2 initiative promotes model efficiency. Retrieved from https://syncedreview.com/2019/07/31/greening-ai-new-ai2-initiative- promotes-model-efficiency/

Chepkemoi, J. (2017). What is a biodiversity hotspot? Retrieved from https://www.worldatlas.com/articles/what-is-a-biodiversity-hotspot.html.

Devulapalli, H. (2017). The Ethics of open data in closed societies. Retrieved from https://medium.com/open-and-shut/conference-notes-ethics-of-open-data-in-closed-societies-4d5bbfac8a7d

Ferri, B. (2019). The Democratization of Artificial Intelligence (AI) for Data Science. Retrieved from https://www.iotforall.com/democratization-of-artificial-intelligence/

Hao, K. (2019). Training a single AI model can emit as much carbon as five cars in their lifetimes. *Technology Review,* Retrieved from https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their lifetimes

Horev, R. (2018). BERT Explained: State of the art language model for NLP. Retrieved from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Kobilinsky, D. (2016). Invasive species bigger threat in developing countries. *The Wildlife Society,* Retrieved from https://wildlife.org/invasive-species-bigger-threat-in-developing-countries/

National Institute of Food and Agriculture. (n.d.). *USDA*, Retrieved from https://nifa.usda.gov/topic/invasive-pests-and-diseases

Pacey, A. (1983). The Culture of technology. Cambridge, MA: MIT Press.

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019, February 14). Better language models and their implications. [Web log post].Retrieved from https://openai.com/blog/better-language-models/

Saha, S. (2018). A comprehensive guide to convolutional neural networks - the ELI5 Way. *Medium*, Retrieved from https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2019). Green ai. *arXiv preprint,* Retrieved from https://arxiv.org/abs/1907.10597

Strubell, E., Ganesh, A., McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint,* Retrieved from https://arxiv.org/abs/1906.02243

Wang, L. (2008). Invasive species spread mapping using multi-resolution remote sensing data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 37*

Vig, J. (2019). OpenAI GPT-2: Understanding language generation through visualization. Retrieved from https://towardsdatascience.com/openai-gpt-2-understanding-language-generation-through-visualization-8252f683b2f8