Mapping Invasive Plant Species using Machine Learning

The Real Costs of Deep Learning: Green AI Initiative

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Surbhi Singh

October 31, 2019

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.

Signed: Surbhi Singh Date: 12/5/19

Approved: Catherine D. Baritaud Date: Dec. 9, 2019
Catherine D. Baritaud, STS Division, Department of Engineering and Society

Approved: Madhav Marathe Date: 12/5/19
Madhav Marathe, Department Computer Science

According to the USDA's National Institute of Food and Agriculture (NIFA), the spread of invasive plant species is currently one of the greatest epidemics facing the agricultural industry. The NIFA cites that invasive plant species are estimated to produce a loss of $137 billion per year in the United States alone, and that they affect nearly every type of ecosystem in the United States (NIFA, n.d., para 1). Due to the growth of human factors such as global travel and foreign imports, the prevalence of invasive species has risen substantially. This phenomenon is not by any means isolated in the United States, but rather is a global pandemic crisis. Understanding and combating the significant contribution of natural and human processes in the spread of invasive species must be addressed quickly.

The research proposed here aims to map invasive plants using remote sensing satellite imagery and machine learning algorithms. Specifically, the type of satellite images used to map species distributions will be investigated. Machine learning (ML) uses patterns from data to predict future occurrences, which can be applied to complex and non-linear data. Invasive plants species are a global epidemic, but they are especially prevalent in biodiversity hotspots, such as the Chitwan-Annapurna Landscape (CHAL) of Nepal. CHAL is the region of interest in this research for that very reason. A biodiversity hotspot is a region which is both rich in plant and animal species, but a region that is also threatened with destruction (Chepkemoi, 2017). With recent developments in high performance computing and machine learning, satellite imagery has become a viable tool in mapping plant species distributions.

While Machine Learning and deep learning models have been proven to be effective, they come at a cost. The Science Technology and Society (STS) part of the prospectus aims to analyze the environmental effects of training complex Machine Learning models by measuring

the computational complexity of the algorithm. Complex models which involve iterative training and parameters often require large data centers to run on. One model can emit more than 626,000 pounds of carbon dioxide (Hao, 2019). The STS portion of the prospectus is loosely coupled with the technical portion of the prospectus as it is analyzing the direct impact of the machine learning algorithms being used in the technical research. Pacey's Triangle will be used to understand the cultural, technical, and organizational barriers to the adoption of environmentally-friendly ML algorithms. The STS research could potentially impact the methodology used in the research to reduce the carbon footprint of training large ML algorithms. This research will be performed under the United States Agency for International Development (USAID) project for climate change and diversity. The work will be performed for one semester at the Biocomplexity Institute and Initiative under supervision of computer science faculty member Madhav Marathe.

IMPACT OF INVASIVE PLANT SPECIES

In addition to agriculture, invasive species impede several sustainability efforts put forth by the United Nations, including efforts to minimize animal extinction and alleviate poverty in low income regions. Invasive species disproportionately affect poor areas, in which people rely on healthy and natural ecosystems for survival (Kobilinsky, 2016). The CHAL has an abundance of plants and animals, with over 3,430 plant species. With the combination of climate change and human activities, invasive species rapidly spread and left several plant species endangered. The research focuses on the distribution of *Lantana Camara*, one of the most threatening invasive plant species. With its allopathic properties, *Lantana* poses a large threat to agriculture and natural ecosystems (Stegelmeier et al., 2013).

METHODS

Many types of satellite imagery have been used to develop species distribution maps. He et al. explains that remote sensing using satellite imagery is the most viable option for tracking plant species with unique phenology or growth form over time (2015). The use of multispectral remote sensing allows for more accurate identification because each layer of the image provides unique information.

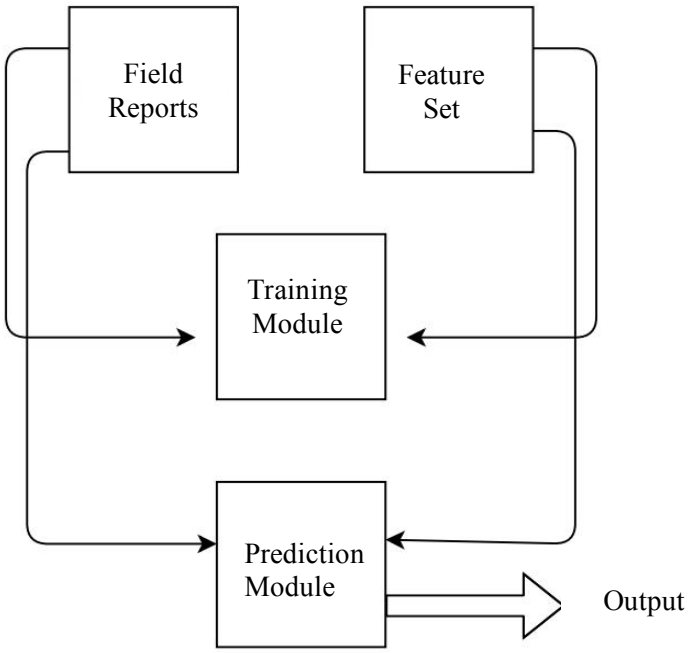
This study uses two different types of satellite imagery data to visualize the spread of invasive plant species over time. DigitalGlobe's Worldview-2 commercial satellite was purchased for high resolution images. These images contained eight bands of multispectral imagery which span a 1.84m area. Because DigitalGlobe is very expensive, only five select areas of the CHAL region with varying plants and elevations were acquired as shown in Figure 1. The

Figure 1: The CHAL region in Nepal with 6 specified regions for which DigitalGlobe imagery was acquired (Adapted from 'Mapping Invasive Plants in a Biodiversity Hotspot' by A. Adiga, 2019)

second type of imagery, Landsat8, is free but contains a much lower resolution. Landsat8 consists of 11 spectral bands which span a 30m region. Imagery for the whole CHAL region was acquired from Landsat8. The objective of the current research is to determine the effect of type of imagery used on the quality of data predictions produced by the Machine Learning algorithms. If the Landsat8 data produces comparable to the

DigitalGlobe data, the need to purchase expensive images is eliminated or reduced. Several different models will be trained on the two different types of imagery, both individually and in combination, to determine the best classification accuracy (Wang, 2008).

The research entails using Convolutional Neural Networks (CNNs) to model the distribution of invasive species using satellite imagery. This type of neural network is a deep learning algorithm which takes an image as input and determines the importance of certain features (Saha, 2018). They are popularly used for mapping invasive species because of the ability of the algorithm to capture the spatial and temporal dependencies of the image. CNNs pinpoint which features are important in the image, and can be used to reduce the presence of invasive species by analyzing and identifying the environmental predictors. Field Reports were generated by a team in Nepal to by gathering recording the presence of different invasive species. The feature set will be generated using the acquired Landsat8 imagery and metadata such as temperature and humidity. As seen in Figure 2, both the field report and features set are inputted to the training module and prediction modules to develop output a binary classifier. This classifier will be used to predict the



Singh, S. (2019) *Architecture to Develop Convolutional Network using Field Reports and Feature Set Data.* [Figure 2] *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

presence or absence of particular invasive species given a multispectral image.

The research is being performed at the Biocomplexity Institute and Initiative for one semester. The work is directly supervised by Abhijin Adiga, research assistant professor in network systems science and advanced computing. In addition, the work is closely guided by postdoctoral research associate, Aniruddha Adiga. The intended outcome of the end of the research project is a published paper.

ENVIRONMENTAL COSTS OF RUNNING MACHINE LEARNING ALGORITHMS: GREEN AI AS A SOLUTION

Machine Learning is a subset of Artificial Intelligence (AI), which is the broader concept of intelligent machines. While AI has become a very popular tool in solving today's problems, most people fail to acknowledge the environmental impacts. In the past few years, these algorithms have made breakthroughs 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. These financial and environmental costs are much higher in research which requires retraining of model architecture and parameters. The amount of computations ran by deep learning research was estimated to have increased by 300,000 times from 2012 to 1028(Strubell, 2019).

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 a STS framework to analyze this problem by identifying the different groups involved and their relationships (1983, p. 6). As seen in Figure 3, the triangle is organized into 3 parts which are organizational, technical, and cultural. The following 3 sections will be used to explain how each part plays a role in the lack of the adoption of environmentally friendly AI algorithms.

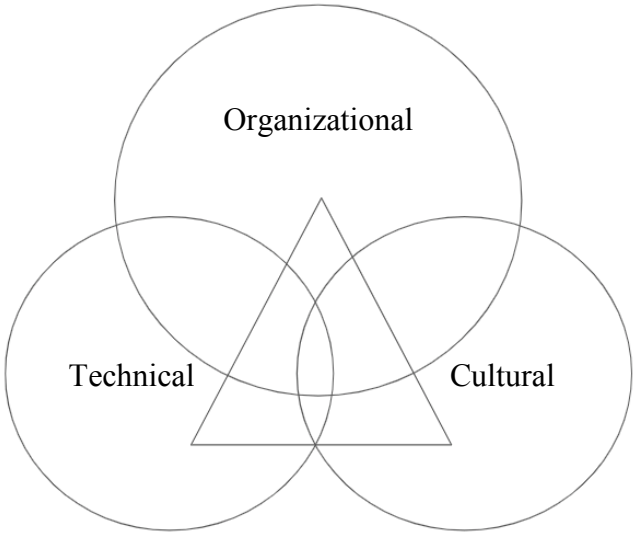


Figure 3: Pacey's triangle for the creation of computationally expensive ML algorithms: An illustration of the three aspects from Pacey's (adapted by Singh from Pacey, 1983, p. 6).

Cultural Pressure

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 in research purposes. Figure 4 below shows the ratio of papers from top AI conferences that report accuracy versus efficiency.

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 research which is environmentally friendly and efficient. Red AI is computationally expensive and often sacrifices large amounts of efficiency for small accuracy gains.

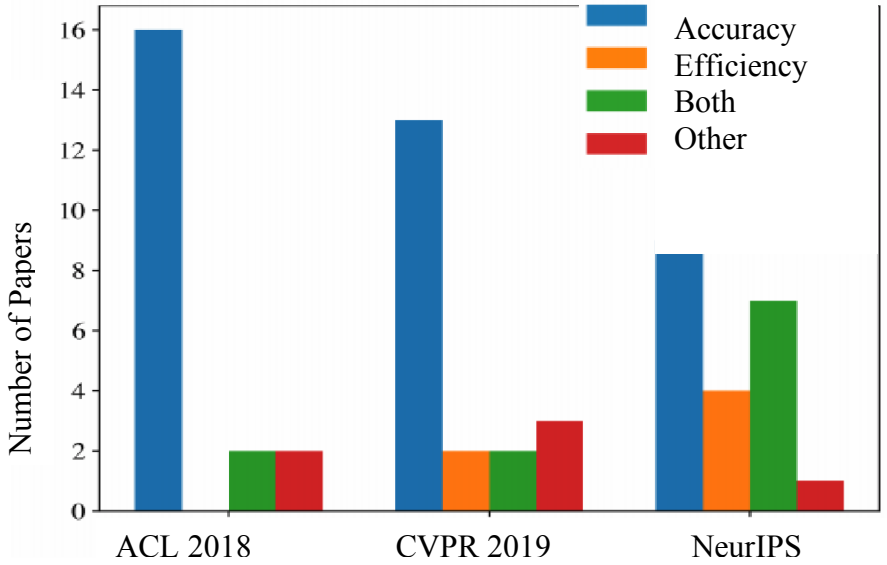


Figure 4: Proportion of AI papers that reported accuracy and efficiency from a sample of 60 papers at renowned AI conferences. (Adapted from 'Green AI' by R. Schwartz, J Dodge, N.Smith, & O. Etzioni 2019)

The lack of adoption of Green AI by the large majority of AI researchers could be due to several reasons. Cultural factors play a large role in the dominance of Red Ai. The pressure to get a paper accepted into a top conference or publish could drive researchers to excessively train models, disregarding the computational costs. For example, the ImageNet Challenge evaluates algorithms on their ability to correctly identify a set of images/videos. The only criteria for the challenge is the accuracy of the model, disregarding the number of training parameters or amount of computations. Figure 5 to the right exemplifies how large jumps in computational costs do not always equate to significant accuracy gains. The computational costs are represented as the total number of floating point

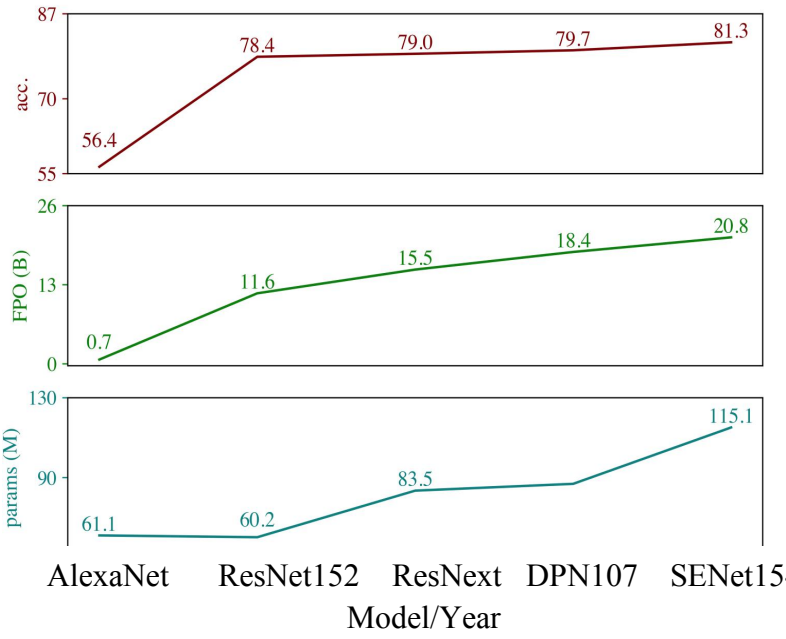


Figure 5: Floating Point Operation(FPO) in billions for different top ImageNet models. Shows how the rate accuracy increase much less steeply than the rate of the FPO increase. (Adapted from 'Green AI' by Schwartz et al., 2019)

operations (FPO) because it is the most objective comparison among algorithms. As seen in Figure 5, a 33.6% increase in FPO resulted in just a 0.76% increase in accuracy. Thus, holistically, ResNet can arguably be seen as a superior algorithm to ResNext. FPO allows for comparison amongst different researchers because it does not involve the hardware or electricity usage.

However, it is not perfect. FPO does not take memory consumption into account, which can affect the amount of energy used. Another limitation of FPO is that it does not account for the way the model is implemented.

Or

Organizational Power

The lack of Green AI adoption is heavily impeded by the Organizational structure. Most AI projects are funded by big tech companies and the government. Currently, tech companies just care about delivering results fast and with highest accuracy. If these organizations were to impose restrictions on the amount of computations certain algorithms should perform, the carbon emissions could be more regulated. In addition, they could prioritize the funding of projects which follow more Green AI principles.

Technical Accessibility

Training AI algorithms efficiently has several benefits beyond lowering carbon emissions. Less computations 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 should be able to run machine learning algorithms with a laptop that could be presented at top conferences. Some argue that computationally expensive algorithms are not a pressing issue at the moment (Biewald, 2019). While these algorithms 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 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. 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 these Red AI algorithms have 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

Biewald, L. (2019). Deep Learning and carbon emissions. *Towards Data Science,* Retrieved from https://towardsdatascience.com/deep-learning-and-carbon-emissions-79723d5bc86e

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-initi ative-promotes-model-efficiency/

Chepkemoi, J. (2017, March 28). What is a biodiversity hotspot? Retrieved from https://www.worldatlas.com/articles/what-is-a-biodiversity-hotspot.html.

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-theirlifetimes/?fbclid=IwAR391STSJzUqg7QQMphxxuELbxQ_HiXICTJ2K0Y0it-6kD8_2HwX9U-jLmY

He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M. N., Schmidtlein, S., ... & Pettorelli, N. (2015). Will remote sensing shape the next generation of species distribution models? *Remote Sensing in Ecology and Conservation*, *1*(1), 4-1

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

Saha, Sumit.(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

Social Construction of Technology. Retrieved October 31, 2019 from: https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/social-construction-technology

Stegelmeier, B. L., Field, R., Panter, K. E., Hall, J. O., Welch, K. D., Pfister, J. A., ... & Green, B. T. (2013). Selected poisonous plants affecting animal and human health. *Haschek and Rousseaux's Handbook of Toxicologic Pathology* (pp. 1259-1314). https://doi.org/10.1016/B978-0-12-415759-0.00040-6

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*, 135-142.