Green AI. Constraint Based Tools used for Sustainability and for Achieving the SDG Objectives.

Project Report for LAAI

By

Enab Muneer (Matricola. 0001057723)

Submitted to

Prof. Maurizio Gabbrielli

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Contents

1. Introduction

We have heard about the words like green environment, green economy, green earth but the term green AI may sound strange when one hears it for the first time. We all know that Artificial Intelligence simply referred to as AI is doing wonders in all the fields of society but very few people are aware of the fact that AI can actually contribute to the ill health of our environment, like any other industry which generates pollution can do. To better understand the Green AI, one should first understand what Red AI is and what disadvantages are associated with it. Only then will we be able to find the solutions to shift from Red AI to Green AI.

2. What is Red AI?

Since the year 2012, the field of AI has shown remarkable progress in a broad range of capabilities including object recognition, game playing, speech recognition and machine translation. Much of this progress has been achieved by increasingly large and computationally intensive deep learning models.

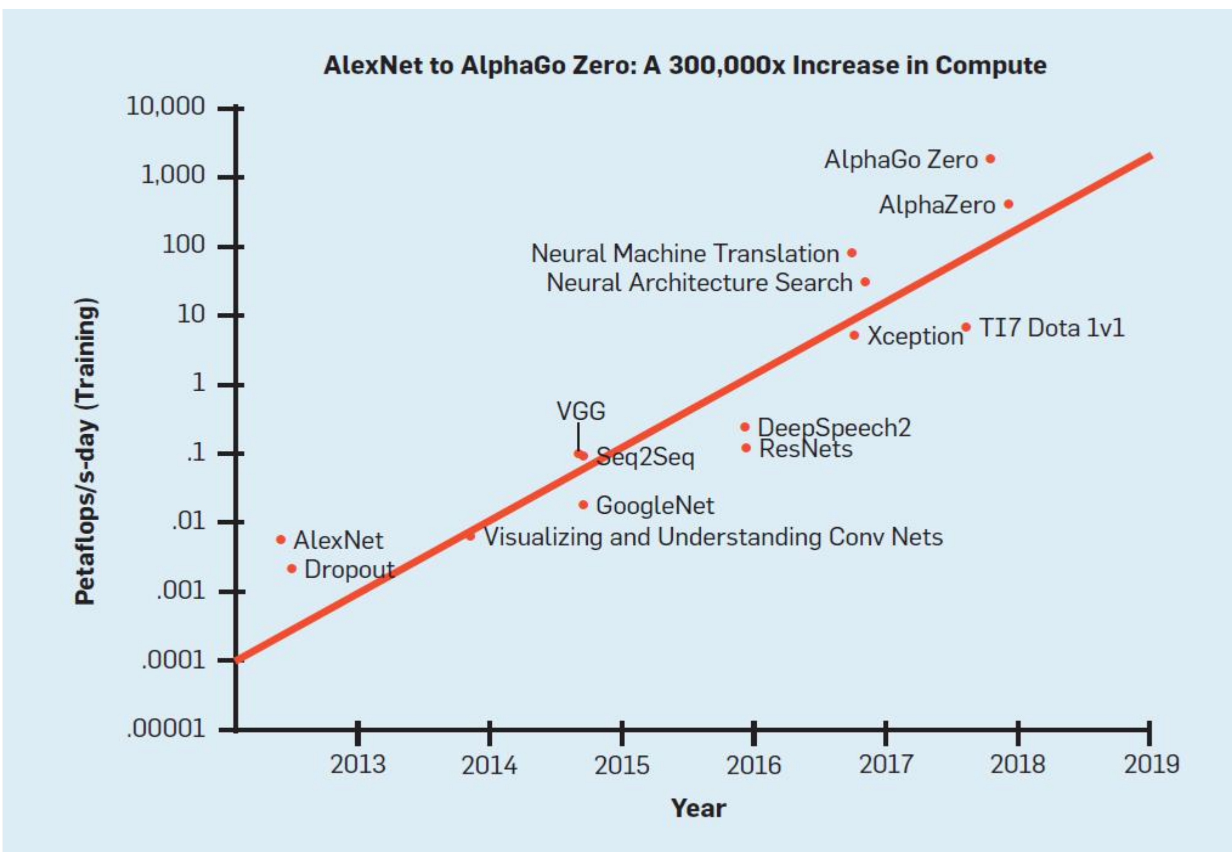


Figure 2.1

Figure 2.1 shows the plot of training cost of the various AI architectures over time. The petaflops are the units of computing the speed equal to one thousand million million (10^{15}) floating point operations per second. From the plot above, we see that the training cost of the architectures have increased drastically upto 300,000 times with training cost doubling every few months starting from AlexNet to AlphaZero. An even sharper trend can be observed in NLP word-embedding approaches by looking at ELMo followed by BERT, and GPT-3. An important paper[1] has estimated the carbon footprint of several NLP models and argued this trend is both environmentally unfriendly and prohibitively expensive, raising barriers to participation in NLP research. Such work is referred to as Red AI.

The AI community has paid relatively less attention towards computational efficiency. The trend can be seen in figure 2.2 below.

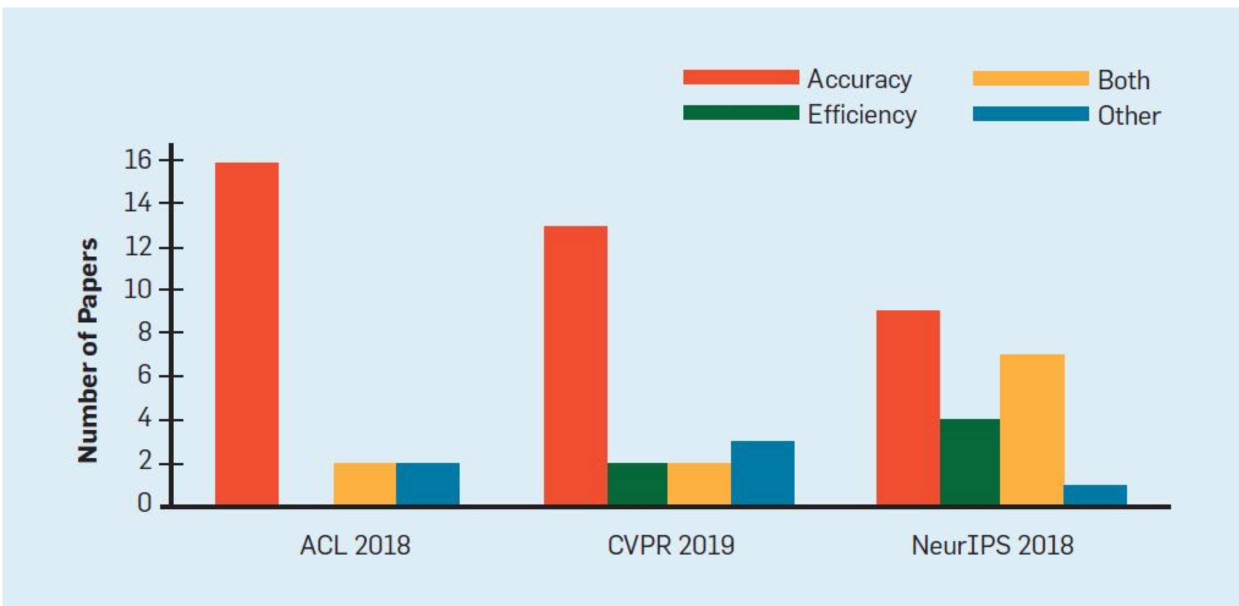


Figure 2.2

Figure 2.2 is achieved by taking a random sample of 60 papers from top AI conferences (ACL, NeurIPS, and CVPR). For each paper the note is taken whether the authors have claimed their main contribution to be (a) an improvement to accuracy or some related measure, (b) an improvement to efficiency, (c) both, or (d) other. A large majority of the papers target accuracy (90% of ACL papers, 80% of NeurIPS papers and 75% of CVPR papers). Moreover, for both empirical AI conferences (ACL and CVPR) only a small portion (10% and 20% respectively) argue for a new efficiency result. This highlights the focus of the AI community on measures of performance such as accuracy, at the expense of measures of efficiency such as speed or model size.

3. What contributes to Red AI

To better understand the computational cost of the AI architectures, we consider the three dimensions. The cost of executing the model on a single (E)xample (either during training or at inference time); the size of the training (D)ataset, which controls the number of times the model is executed during training, and the number of (H)yperparameter experiments, which controls how many times the model is trained during model development. The total cost of producing a (R)esult in machine learning increases linearly with each of these quantities. This cost can be estimated as follows:

$$\text{Cost}(R) \propto E \,.\, D \,.\, H$$

Expensive processing of one example: If we consider the neural models, where it is common for each training step to require inference, we consider training and inference cost together as "processing". Some works have used increasingly large models in terms of, for example, model parameters, and as a result, in these models, performing inference can require a lot of computation, and training even more so. For instance, Google's BERT-large contains roughly 350 million parameters. OpenAI's openGPT2-XL model contains 1.5 billion parameters. AI2, our home organization, released Grover, also containing 1.5 billion parameters. NVIDIA released Megatron-LM, containing over 8 billion parameters. Google's T5-11B contains 11 billion parameters. Most recently, openAI released openGPT-3, containing 175 billion parameters. In the computer vision community, a similar trend is observed as seen in Figure 2.1.

Such large models have high costs for processing each example, which leads to large training costs. BERT-large was trained on 64 TPU chips for four days at an estimated cost of $7,000. Grover was trained on 256 TPU chips for two weeks, at an estimated cost of $25,000. XLNet had a similar architecture to BERT-large, but used a more expensive objective function (in addition to an order of magnitude of more data), and was trained on 512 TPU chips for 2.5 days, costing more than $60,000. Specialized models can have even more extreme costs, such as AlphaGo, the best version of which required 1,920 CPUs and 280 GPUs to play a single game of Go, with an estimated cost to reproduce this experiment of $35,000,000.

We see that larger models can have stronger performance, which is a valuable scientific contribution. However, this implies the financial and environmental cost of increasingly large AI models will not decrease soon, as the pace of model growth far exceeds the resulting increase in model performance. As a result, more and more resources are going to be required to keep improving AI models by simply making them larger.

Processing many examples: Processing many examples leads to Increased amounts of training data that have also contributed to progress in state-of-the-art performance in AI. BERT-large had top performance in 2018 across many NLP tasks after training on three billion word-pieces. XLNet outperformed BERT after training on 32 billion word-pieces, openGPT-2-XL trained on 40 billion words, FAIR's RoBERTa was trained on 160GB of text, roughly 40 billion word-pieces, requiring around 25,000 GPU hours to train. T5-11B was trained on 1 trillion tokens, 300 times more than BERT-large. In computer vision, researchers from Facebook pretrained an image classification model on 3.5 billion images from Instagram, three orders of magnitude larger than existing labeled image datasets such as Open Images. Storing the data is also expensive. Finally, as in the case of model size, relying on more data to improve performance is notoriously expensive because of the diminishing returns of adding more data. For instance, Figure 3.1 shows a logarithmic relation between the object recognition top-1 accuracy and the number of training examples.

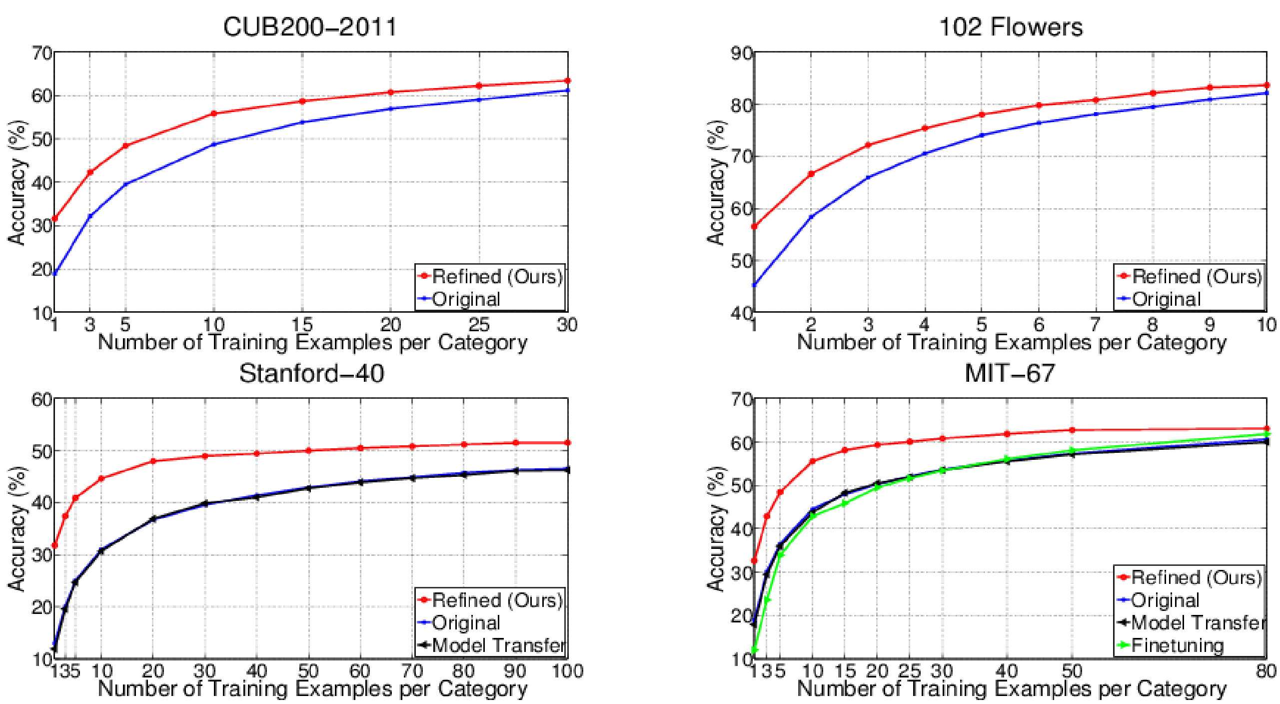


Figure 3.1

Massive number of experiments: Some projects have poured large amounts of computation into tuning hyperparameters or searching over neural architectures, well beyond the reach of most researchers. For instance, researchers from Google trained over 12,800 neural networks in their neural architecture search to improve performance on object detection and language modeling. The number of experiments performed during model construction is often underreported. Nonetheless, evidence for a logarithmic relation exists here as well.

The increasing costs of AI experiments offer a natural economic motivation for developing more efficient AI methods. It might be the case that at a certain point prices will be too high, forcing even researchers with large budgets to develop more efficient methods. Analysis in Figure 2.2 shows that currently most effort is still being dedicated to accuracy rather than efficiency. At the same time, AI technology is already very expensive to train or execute, which limits the ability of many researchers to study it, and of practitioners to adopt it. Combined with the environmental price tag of AI, we believe more effort should be devoted toward efficient AI solutions.

Red AI work is extremely valuable, but there is value in pushing the limits of model size, dataset size, and the hyperparameter search budget. That is why we should take measures to make AI more green.

4. What is Green AI?

The term Green AI refers to AI research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent. Green AI promotes approaches that have favorable performance/efficiency trade-offs. If measures of efficiency are widely accepted as important evaluation metrics for research alongside accuracy, then researchers will have the option of focusing on the efficiency of their models with positive impact on both inclusiveness and the environment.

5. How to achieve Green AI?

To measure the efficiency of an AI system, one should report and consider the amount of work required to generate the result. This amount of work should include the work required to train a model, the aggregated amount of work required for all hyperparameter tuning experiments, the size of the dataset, and the number of experiments. Reducing the amount of work in each of these steps will result in AI that is more green. The parameters that can be considered to measure the efficiency of an AI system are as follows:

1. Carbon Emission: Carbon emission is the quantity that we want to directly minimize. But, it is difficult to measure the amount of carbon released by training or executing a model, as this amount depends highly on the local electricity infrastructure. Therefore, it is not comparable between researchers in different locations or even the same location at different times.

2. Electricity usage. Electricity usage is correlated with carbon emission. Moreover, GPUs often report the amount of electricity each of their cores consume at each time point, which facilitates the estimation of the total amount of electricity consumed by generating an AI result. Without a doubt, this measure is hardware dependent, and hence, does not allow for a fair comparison between different models developed on different machines.

3. Number of parameters. Another common measure of efficiency is the number of parameters used by the model. As with runtime, this measure is correlated with the amount of work. Unlike the other measures, it does not depend on the underlying hardware. Moreover, this measure also highly correlates with the amount of memory consumed by the model. Nonetheless, different algorithms make different use of their parameters, for instance by making the model deeper vs. wider. As a result, different models with a similar number of parameters often perform different amounts of work.

4. FPO: It is the measure of the total number of floating-point operations (FPO) required to generate a result. FPO provides an estimate of the amount of work performed by a computational process. It is computed analytically by defining a cost to two base operations, ADD and MUL. Based on these operations, the FPO cost of any machine learning abstract operation can be computed as a recursive function of these two operations. FPO has several appealing properties. First, it directly computes the amount of work done by the running machine when executing a specific instance of a model and is thus tied to the amount of energy consumed. Second, FPO is agnostic to the hardware on which the model is run. This facilitates fair comparisons between different approaches, unlike the measures described in the above points. Third, FPO is often correlated with the running time of the model. FPO also considers the amount of work done at each time step.

FPO seems to be the best way to calculate the cost of an AI model, but it also has some limitations. The energy consumption of a model is not only influenced by the amount of work, but also from other factors such as the communication between the different components, which is not captured by FPO. As a result, FPO doesn't always correlate with other measures such as runtime and energy consumption. Second, FPO targets the number of operations performed by a model, while ignoring other potential limiting factors for researchers such as the memory used by the model, which can often lead to additional energy and monetary costs. Finally, the amount of work done by a model largely depends on the model implementation, as two different implementations of the same model could result in very different amounts of processing work. Due to the focus on the modeling contribution and accuracy, the AI community has traditionally ignored the quality or efficiency of models' implementation. This norm should be reversed and exceptionally good implementations that lead to efficient models should be credited by the AI community.

Additional ways to promote Green AI:
In addition to reporting the FPO cost, the researchers should report the budget/performance curves wherever possible.

6. Sustainable Development Goals(SDG):

The 17 Sustainable Development Goals (SDGs) are humanity's most ambitious plan for a better world.In September 2015, 193 countries agreed on these 17 goals and 169 sub-goals at the United Nations General Assembly. Since the goals are to be achieved by 2030, they are also called the 2030 Agenda. The SDGs provide worldwide guidance for addressing the global challenges facing the international community. It is about better protecting the natural foundations of life and our planet everywhere and for everyone, and preserving people's opportunities to live in dignity and prosperity across generations.

The 17 goals cover all three areas of sustainable development: ecological, economic and social. These goals are as follows:

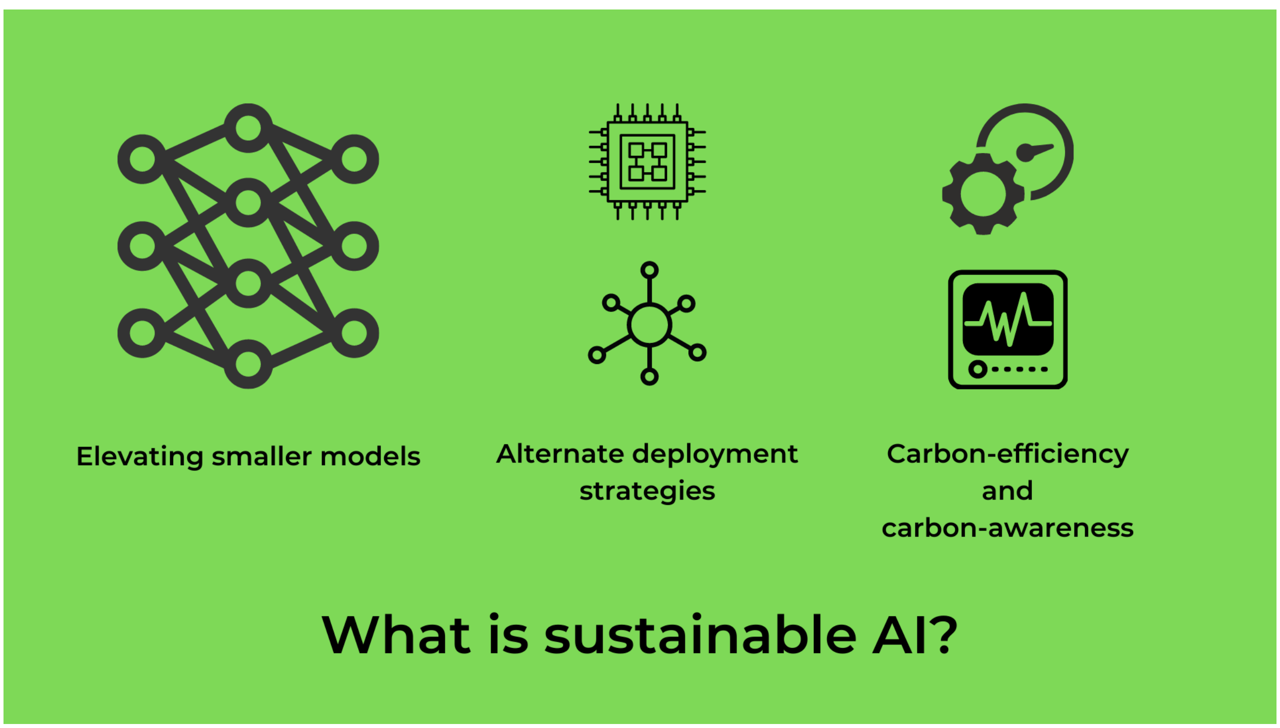
1. No Poverty
2. Zero hunger

3. Good health and well being
4. Quality Education
5. Gender Equality
6. Clean water and sanitation
7. Affordable and clean energy
8. Decent work and economic growth
9. Industry, innovation and infrastructure
10. Reduced inequality
11. Sustainable cities and communities
12. Responsible consumption and production
13. Climate action
14. Life below water
15. Life on hand
16. Peace and justice strong institutions
17. Partnership to achieve the goal

The above 17 SDG goals are built on the principle of "Leaving No One Behind". This agenda emphasizes on the approach of achieving sustainable development for all. SDG ensures access to affordable, reliable, sustainable and modern forms of energy for all, which in turn could reduce poverty, improve health and wellbeing, and mitigate climate change.

7. Sustainable AI

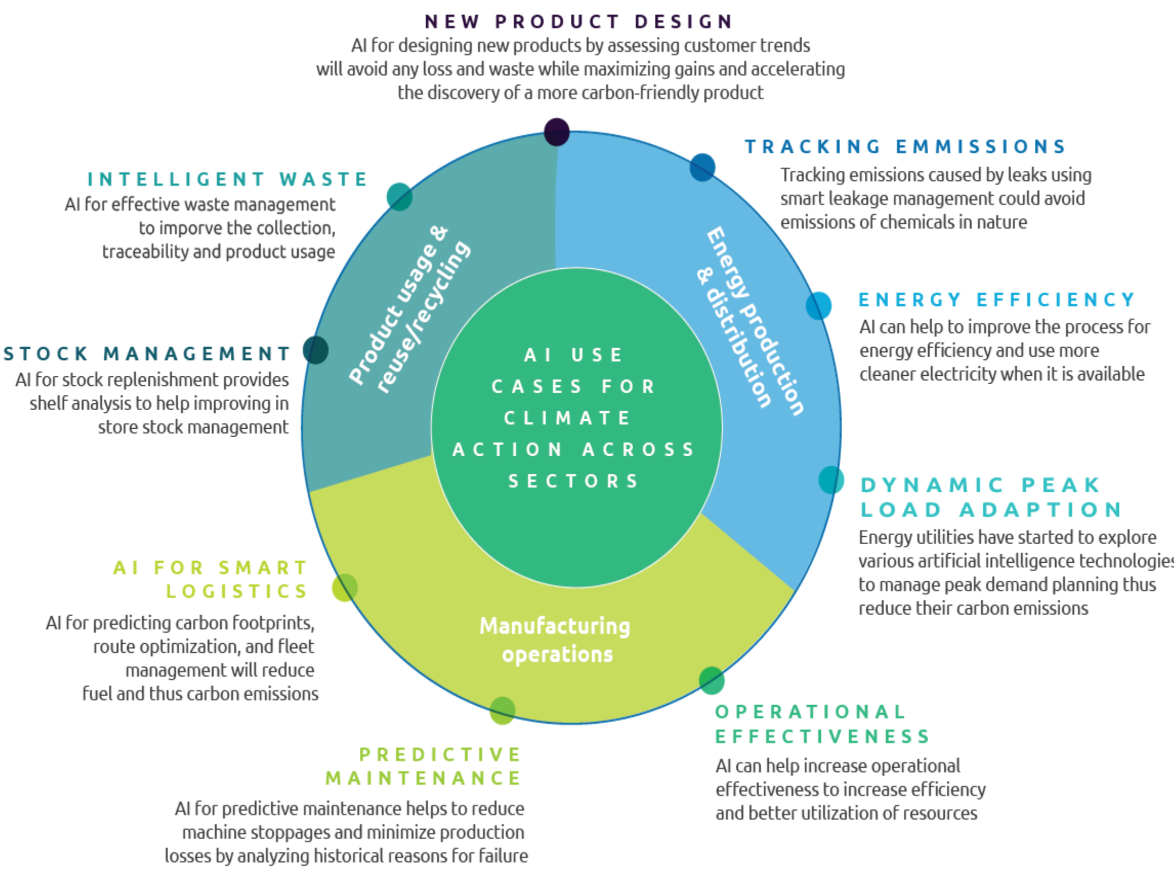
Sustainable AI is referred to as the AI used to solve environmental issues. Sustainable AI is a movement to foster change in the entire lifecycle of AI products (i.e. idea generation, training, re-tuning, implementation, governance) towards greater ecological integrity and social justice. Sustainable AI is focused on the whole sociotechnical system of AI, that is, to develop AI that is compatible with sustaining environmental resources for current and future generations.



8. How AI can contribute to SDG/ AI and sustainability

In 2017 the UN started its 'AI for Good' global summit series hosted by the International Telecommunication Union (ITU) in cooperation with other UN organizations envisaged as the leading dialogue platform on AI contributing to the SDGs. Artificial Intelligence (AI) is the ally that sustainable development needs to design, execute, advise and to plan the future of our planet and its sustainability more effectively. Technology like AI will help us build more efficiently, use resources sustainably and reduce and manage the waste we generate more effectively, among many other matters. According to a study, AI could help achieve 79 % of the Sustainable Development Goals (SDGs).

A clear example of AI's contribution to sustainability is traffic management. Applying Artificial Intelligence in urban mobility allows traffic jams to be predicted and alternative routes to be suggested. With shared mobility, this technology predicts vehicle demand by zone and time. This means that companies can organize the availability of vehicles for citizens based on their needs. This solution not only facilitates mobility, but also minimizes its environmental impact.



AI can also help enhance the efficiency of renewable energies. Companies are already using this technology to find out the daily availability of energy-generating facilities (wind turbines, hydraulic plants, biomass plants etc.), in order to predict the energy production required to be produced in the coming days and, ultimately, to prevent and diagnose breakdowns.

Beyond the energy sector, there are many industries and businesses that can improve by implementing AI technologies, all while helping the planet. In agriculture, for example, it is used to make irrigation and fertilization more efficient. Thanks to humidity, temperature and fertilization sensors, Artificial Intelligence is able to predict crops' needs. The most innovative solutions within agricultural sustainability are drones that help farmers with surveillance, in addition to hyperspectral image analysis for comprehensive pest control.

Artificial Intelligence can be used in error prediction. In manufacturing and distribution systems, the AI can be used to identify errors in assembly lines that are invisible to the human eye.

9. Constraint based problem

Constraint based problem is a search process for a specific problem according to special constraints/conditions of that problem. The goal of the tool is to find the best feasible solution with the consideration of the problem constraints. Constraint based tools are nowadays used for energy resources sustainability, environmental sustainability, economic sustainability, and social sustainability. Following are some of the work that has been done in using the constraint based algorithms to achieve SDG objectives.

9.1. Constraint based tools for sustainability and for achieving SDG objectives

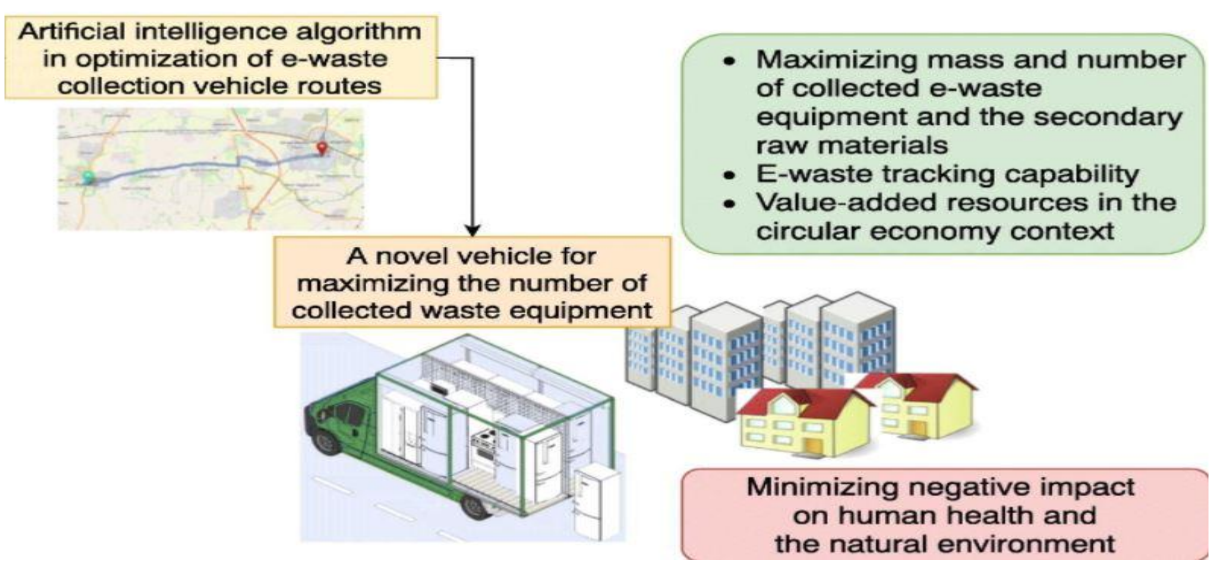
Here are some of the constraint based tools examples that have been implemented in the real world to achieve the SDG objectives:

9.2. Solid Waste Collection

Urbanization is rapidly increasing in developing countries. The increasing trend towards urbanization and population growth combined with the growing concern about the negative environmental impacts have created a critical situation for the management of household solid waste. Ineffective waste management can lead to degradation of valuable land resources.

The optimization objectives in solid waste collection (SWC) are route optimization which is performed to improve the collection and transportation of the SWC system. This is achieved by reducing the route length. Time optimisation, Allocation/relocation of bins, waste content, transfer/collection station, cost optimization and environmental impacts. To optimize these parameters in a more realistic environment, the various constraints have been considered. Since waste is collected by a waste collecting vehicle, the most common constraint in different studies is the vehicle capacity constraints, time constraints, environmental constraints.

SWC operation can vary among countries due to various socio-economic factors including population size, financial status, road congestion, air pollution, human labor availability, amount of collection vehicle, etc. The waste collection process could also differ due to type building such as residential and commercial. Besides, the feasibility of various technologies are considered as main barriers to execute the SWC model operations. The safety and waste disposal are reported as some other constraints of SWC. The implementation of an optimal SWC system can contribute to multiple SDGs and their targets related to improving the sustainable smart city/smart village through environmental protection, public health protection and poverty reduction.



The solution approaches of SWC:

The conventional approaches cannot deliver satisfactory solutions while optimizing multiple objectives and handling uncertain parameters. The heuristic approaches are dominant to conventional approaches in terms of computational time and problem complexity. The cluster-first-route-second heuristic method was used to solve the capacitated arc routing problem (CARP) by the Vertex-one Center algorithm. In this method, at first, the groups are formed by clustering the customers and then allocated to waste vehicles (phase I), after which, efficient routes are identified for each group (phase II). The results demonstrate that the routing cost drops by 25% .

Another way of optimization is using the meta-heuristic approaches. The particle swarm optimization (PSO) algorithm is a popular meta-heuristic population-based optimization approach which uses the behaviors and movement of an organism such as bird flocking and fish schooling in searching for food. A capacitated vehicle-routing problem CVRP model is designed based on a modified particle swarm optimization (PSO) algorithm. The results indicate that the proposed model achieves the best waste collection and route optimization with regard to fuel consumption, travel length, cost and efficiency. Another meta heuristic approach is the Ant colony optimisation ACO algorithm. It is based on the behavior of an ant in a path that is designed using the deposition of pheromones by the previous ants. A clustering-based improved multiple ACO is proposed to minimize the route distance of waste collection. The results indicate a decrease in route length by 31.1% . Another meta-heuristic approach is the Genetic Algorithm(GA). The GA uses three operational stages including, reproduction, crossover, and mutation to achieve the final best solution. A VRP for SWC is modeled through GA to find the optimal routes leading to the minimization of the traveled path and operational costs. The results indicate a reduction in travel distance by 66%.

9.3. Biomass Plant Placement with Energy-Effective Supply

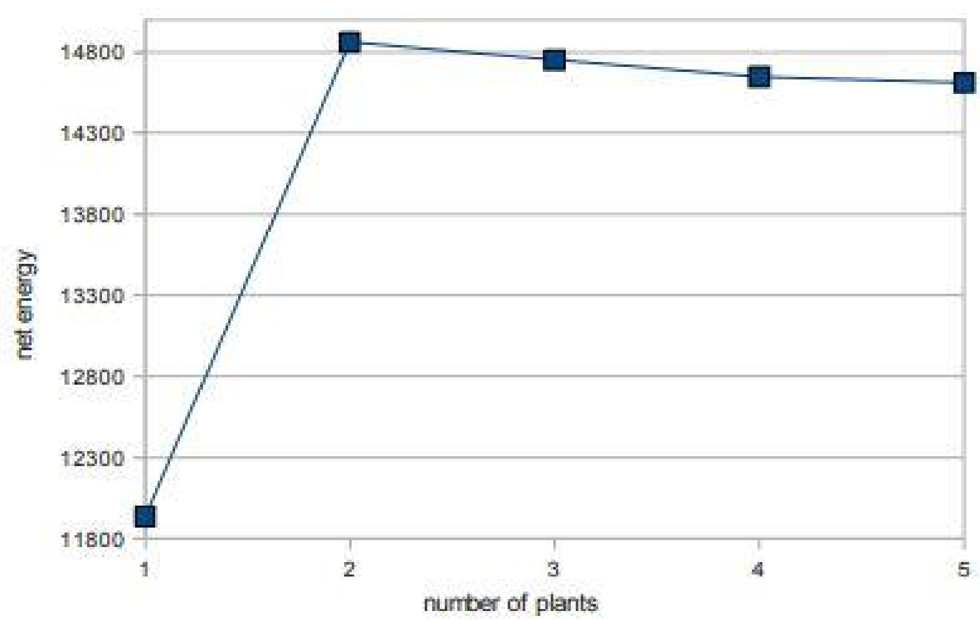
The end of fossil fuels seems postponed to a midterm future, the search for new energy sources is recently having a new boost. After decades of warnings from the environmental experts, citizens and governments start to realize that we cannot continue polluting our planet indefinitely. Carbon oxides are known to increase the greenhouse effect, widely recognized as the main reason for climate change. In the search for reducing carbon oxides emissions, biomass-powered power plants are very promising, because they provide energy with an almost carbon neutral process since biomass mostly comes from trees and vegetables that during their life converted carbon dioxide to oxygen. Biomass power plants are aimed at obtaining energy from a variety of different fuels: from garbage, to forest/sawmill residues, to manure. Also, for countries that mainly rely on imported energy, biomass power may mean an economy less dependent on the price of oil. On the other hand, building a biomass plant does not necessarily imply any improved sustainability, as the plant is inserted into an environment, with complex interrelations, including production of fumes, the need for refrigeration, transport of the biomass from the production sites to the power plant, and so on.

The CLP model is used to address the given problem. The problem is modeled into a graph that connects possible locations of the biomass power plants to the location of the forests. Each arc is labeled with the shortest path length between the two nodes. The (shortest path) distance between two nodes i and j is called d_{ij}. Array P represents the possible locations of the plants. The element p_i can be 0 or 1. It is 1 iff we build a power plant in node i. The constraints defined are:

1. The total quantity of wood provided by a forest node does not exceed its carry capacity, that is the production capacity that a forest has, cannot be exceeded.
2. Minimal quantity of wood that a power plant shall receive in order to keep it running.
3. Maximum quantity of wood that a power plant can accept.
4. Another constraint can be that one cannot receive wood if the power plant has not been installed.

From a sustainability viewpoint, the aim is that the whole system generated by the new plant produces renewable energy. The system contains the plant, as well as the forest owners collecting wood and transporting it to the plant, so we must ensure that the energy spent for the transport and for building the plant does not overpass that produced by the wood in the plant.

The CLP model was applied to the two subregions of the Emilia-Romagna region.The two areas were selected with different characteristics. The western area with large availability of forests and the eastern area with less forest nodes. The goal was to place a number of plants not greater than 5 and whose production had to be more than 0.2 MW but less than 1 MW. The total net energy produced by the plants in the two optimal configurations was respectively 88500 GJ and and 16800 GJ for the western and eastern areas, while the total energy necessary for the transportation of the biomass to the power plant is, respectively 96133 MJ and 4707 MJ. In both cases, the net produced energy is positive, which shows that the optimal placement reached its objective.



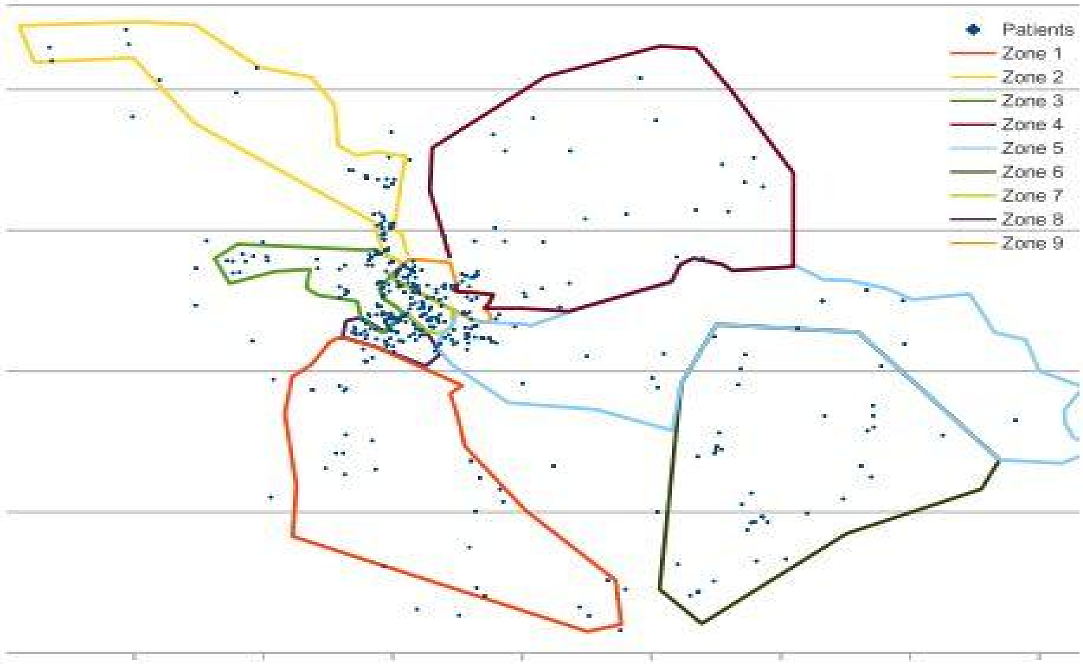
9.4. Workload-Balanced and Loyalty-Enhanced Home Health Care

The home health care (HHC) service in the city of Ferrara, Italy, is managed by the local agency of the National Health Service (NHS), namely AUSL 109. All patients who are not self sufficient and in need of medical treatment are eligible. Requests are made and a medical treatment on the specific day of the week must be delivered. Service is provided by a set of qualified nurses. Every day, each nurse who is on duty starts her job from the hospital, visits the patients in her list delivering the required treatments and traveling by car from one patient's home to the next eventually returning to the hospital. While Ferrara is a medium-size town (about 150,000), the area administered by AUSL 109 is rather large and its population is aging.

Scheduling such a service has several constraints:

1. The minimization of the travel time over the service time. i.e. the travel time during which a nurse is on duty but is not delivering any service
2. The equidistribution of the workload, which can not be guaranteed by simply equally subdividing patients, due to heterogeneous requests i.e, one patient can have several treatments and all of them are done by a single nurse.
3. A good degree of loyalty for the patient i.e, the number of different nurses who are in charge of a single patient should be minimal.

Before, nurses organized their duties themselves. In order to simplify the subdivision of the patients to the nurses, the territory pertaining to AUSL 109 was partitioned statically into 9 zones considering several factors, such as the distribution of the population, its age and historical data.



. The 9 Zones in which the area is divided, dots show the location of patients.

A CP model and search heuristic was applied aiming at finding the schedules and routing plans such that nurses workloads are balanced and patients are not visited by too many different nurses. The results achieved in CP were proved to be significantly better than the ones obtained by hand by the nurses. Also, the travel time by the nurses was reduced which means their vehicles consumed less fuel and reduced pollution impact. Also, nurses save approximately 3 hours per week of travel time, which means a significant saving of person-hours and money that can be devoted to provide even better service or to serve a higher number of patients.

10. Conclusion

In the discussion above, I have mentioned the various problems related to sustainability addressed by means of Logic and Constraint Programming. We saw that energy is the key issue under several points of view, its efficient use, its renewability and environmental friendly production. The main aspect related to sustainability is the greenhouse gas emission, whose volume in the atmosphere needs to be contained in order to avoid climate changes. We have also seen that AI can contribute more to the carbon footprint and greenhouse gasses as its processes including hardware used to power it, training emits larger carbon footprint which inturn does more harm to the environment. AI must be developed in such a way that it does no harm to the environment. Sustainable AI offers an approach that harmonizes these issues and presents a potential pathway to addressing these challenges in a holistic fashion keeping in mind our planet, organizational profits, and above all, people at the center of the design, development, and deployment phases of an AI system. We have also seen that in order to obtain sustainability, constraint programming and computational logic has effectively contributed to the solution. We can push forward on making sustainable AI and attain sustainability and achieve sustainable development goals by starting to share the ideas widely. The harmful impact of AI to our environment is not being discussed the way it should have been. Many people are still unaware of this subject. We should increase the awareness in our society, educational institutions and governments so that we can all as a society take part to curb these harms and make our planet sustainable for future generations.

References

1. Strubell, E., Ganesh, A. and McCallum, A. Energy and policy considerations for deep learning in NLP. In Proceedings of ACL, 2019.
2. Mohammed Abdallah, Manar Abu Talib, Sainab Feroz, Qassim Nasir, Hadeer, Bayan Mahfood. Artificial Intelligence applications in solid waste management. A systematic review.
3. Green AI By Roy Schwartz, Jesse Dodge, Noah A. Smith, Oren Etzioni. Communications of the CM, December 2020, Vol. 63 No. 12
4. Sustainable AI by The Gradient. https://thegradient.pub/sustainable-ai/
5. Jon Reilly. How can AI help in achieving the UN's sustainable development goals. https://www.akkio.com/post/how-can-ai-help-in-achieving-the-uns-sustainable-development-goals
6. Emma Strubell, Ananya Ganesh, Andrew McCallum. College of Information and Computer Sciences. Energy and Policy Considerations for Deep Learning in NLP.