May, 2020
Ekkono Solutions AB

Energy Efficient Algorithms

While more and more devices are being connected – we are talking billions of units – imagine the amount of energy they consume. What happens when you also run machine learning (ML) on these devices? How can we design ML models that are as environmentally friendly as possible from the start? And better – how can we design algorithms that also train devices to be better and smarter and therefore minimize their CO_2 footprint?

We are currently in the middle of a world-wide energy transformation, where we reduce the dependency on non-renewable energy sources and scale up renewable energy solutions, such as wind or solar energy. We know that the resource will be scarce, which requires us to use it wisely. At Ekkono we have a sustainable mindset from start to finish. Early on, we decided to assess each customer case according to the UN Agenda 2030, to be sure that our technology does not violate any of the sustainability goals. Further, we strive to prioritize those cases where our technique contributes to a more sustainable future. Another aspect of this is to make sure that our technique in itself, that is, our set of algorithms and software library, does not consume more energy than necessary. This is why we focus on making our machine learning algorithms as energy efficient as possible.

But it is more to it than sustainability. Energy efficient ML algorithms take less processing power, run fewer CPU cycles and consume less memory, this enables us to run our algorithms on tinier devices that never before had the capacity to run machine learning on the edge. It also gives the opportunity to retrofit Ekkono to an already installed base, even though the processor was not originally built for Edge Machine Learning[*]. Imagine the possibilities that come with OTA (over the air) updates of new functionalities, without the need of a new more powerful processor. Our arguments for why energy efficient machine learning is important are endless, but let us start with a deep dive into what it actually is.

[*] Edge machine learning means running machine learning (ML) at the edge of the network – onboard the connected device. Ekkono develops an edge machine learning software. In Ekkono's case, it is possible to do incremental learning at the edge, which means that the ML model continuously gets better but also that it gets personalized as it is fed with sensor data while in production.

1. Green Machine Learning

Did you know that the bitcoin network alone consumes around 22.9 million metric tons of CO_2? That equals the same amount of energy consumption as the entire country of Jordan and 0.2% of the total energy consumption in the world. Consider all 50 billion connected units we have in the world right now – do we have enough energy available to run ML on them? And in addition to energy, do we even have that amount of processing power and memory available right now? One solution is to develop better and more powerful processors. Another is to optimize the techniques used for training the devices to become smart and use as little energy as possible in action.

At Ekkono we know that if we want to save the world with new technology, we need to start looking at solutions that are possible. One of our main focuses is to develop energy efficient algorithms – ML algorithms optimized to run on low processing power to decrease energy consumption. We call it green machine learning and it aims at designing algorithms and models that are optimized for low computational and power requirements.

2. Optimized and Customized Algorithms

An ML algorithm is a set of instructions to be executed in a computer. Those instructions can be computations or memory accesses. When those instructions are carried out, the processor and the memory dissipate power. The more instructions, the more computations and memory accesses, thus the more energy consumption. Therefore, depending on what the algorithm is doing, it will have a different energy consumption pattern. We optimize which set of algorithms are most suitable for each type of problem.

When one designs ML algorithms, we need to start with an evaluation of its energy consumption and from there design algorithms that are energy efficient without affecting predictive accuracy results. For every algorithm that we develop at Ekkono, we ensure that they are as energy efficient as possible. We constantly benchmark our solutions to thrive for efficiency. Every time we decide which algorithm best suits a specific task, we analyze its energy footprint, to guarantee that we pick the best solution in terms of performance and sustainability. This means the ML algorithm will run as energy efficient as possible without affecting predictive accuracy.

3. Technical Details

Different ML algorithms consume energy in different ways. Two very distinct examples are deep learning algorithms and Hoeffding Trees.

Hoeffding Trees are a set of online decision trees that analyze data in real time while incurring in low computational and power requirements. These algorithms are designed to run on small platforms, thus designing them even more energy efficiently has a great impact on the performance of these platforms. To build energy efficient ML algorithms we propose three steps: i) analyze the energy hotspots of the algorithms – see where energy is being consumed, ii) optimize the algorithms in particular on those hotspots to, iii) reduce their overall energy consumption. When we analyzed the energy consumption of Hoeffding Trees, we discovered that there were many repetitive tasks that could be avoided, and that they were using static hyper parameters for data of a dynamic nature. We proposed a solution that set the hyper parameters of the algorithms adaptively, based on the characteristics of the incoming data. On top of that, we proposed a splitting criteria for these types of trees that is energy efficient. The idea was to have an energy budget to be used by the different types of splitting nodes. The nodes that were most responsible for an accuracy increase utilize more of the energy budget, while the rest of the nodes ran in an ECO mode utilizing less part of the overall energy budget. These solutions allowed us to reduce the energy consumption by 20% to 40% for different test cases.

Deep learning algorithms are a wide variety of neural networks meant to address complex problems such as object recognition and natural language processing. While these networks are very powerful, they usually consume significant amounts of energy. That is why researchers in this area have proposed different solutions to make smaller, more energy efficient networks. Two of the many solutions are quantization and pruning. Quantization helps to reduce the memory space of the weights of the networks, by representing such numbers with a lower precision. This leads to both easier computations and less memory requirements. Pruning is used to remove redundant connections between some of the neurons, to reduce both memory accesses (since less nodes need to be accessed) and computations. These advancements have led to very efficient deep learning networks that are able to run on mobile and embedded devices.

4. Use Case

Moving machine learning towards a greener future, with more energy efficient and sustainable solutions is the ultimate goal. The main challenge nowadays arises as a sustainable mindset is hardly ever encountered for in many of the state-of-the-art ML solutions. A slightly more energy efficient ML algorithm saves battery, CPU cycles, etc., which means we can run on smaller things. While it saves energy it also contributes to a development of more sustainable devices – and less CO_2 emissions.

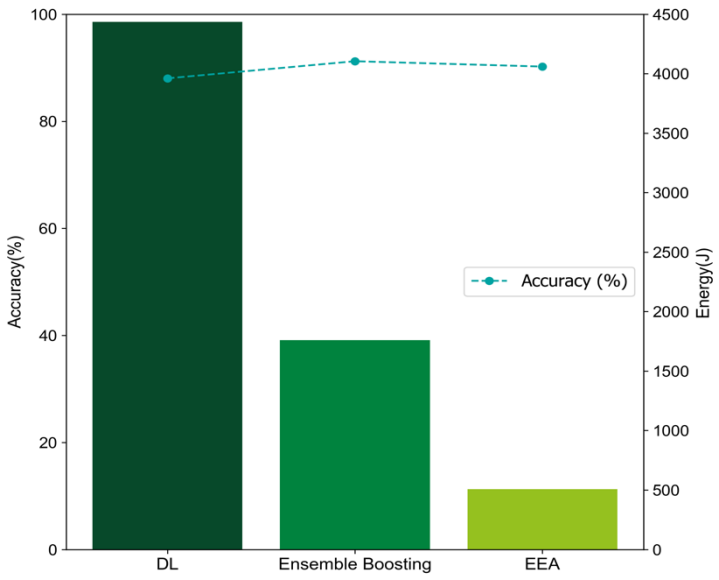

Figure 1. Accuracy (%) and Energy Consumption (J) comparison between three algorithms: a deep learning (DL) algorithm, ensemble boosting algorithm (Ensemble Boosting) and the extended Hoeffding Tree algorithm (EEA).

Figure 1 shows the comparison between three different algorithms, a deep neural network (DL), an online ensemble boosting algorithm (Ensemble Boosting), and an energy efficient extension of the Hoeffding Tree algorithm (EEA). We compared these algorithms on a dataset that has different forest cover types, where the goal is to classify such types correctly.

The data shows that a regular deep learning (DL) algorithm consumes eight (8) times more energy than an optimized and sustainable one, while achieving 2% less accuracy. The energy efficient Hoeffding Tree algorithm (EEA) consumes almost four (4) times less energy than the ensemble boosting algorithm but achieves the same accuracy. A simple analysis like this enlightens new and interesting advancements, since it illustrates that the highest performing algorithms can be the most energy efficient ones, contrary to public belief.

5. Conclusions

Considering that ML is used on more and more connected things, it is vital that energy efficiency becomes a standard performance metric when evaluating ML techniques. Energy efficient machine learning algorithms will help build a more sustainable future, saving energy

by running more efficiently onboard connected units, but also by training devices to run optimally and energy efficiently.

Apart from sustainability, energy efficient ML has a number of benefits for practical reasons. An energy efficient algorithm runs fewer CPU cycles and consumes less energy, which enables Ekkono to run on smaller devices than ever. Based on the fact that we consume a minimum of CPU power and memory, it also enables Ekkono to be retrofitted into an already installed base; i.e. the ability to update functionalities on devices that are already running in for example production. This is an interesting idea, considering the fact that it historically has been about producing and selling new products in order to launch new features. An efficient way of retrofitting into the installed base gives the opportunity to sell updates of functionalities, without the costs and resource consumption of producing new products. This leads to a business model that is essentially about producing fewer products, with the ability to earn money on the entire product lifecycle – which will have a positive impact on sustainability as well.

***

Ekkono #openfika is a short open, online fika[†] session, hosted by Ekkono, on hot, contemporary and relevant topics, where a 15 minutes presentation is followed by discussion and Q&A. Keep an eye on www.ekkono.ai and LinkedIn for the next #openfika session.

Ekkono Solutions AB is a software company that develops Edge Machine Learning. Our product is the result of seven years of research at the University of Borås, Sweden, and assists product OEMs in different industries to rapidly develop smart features onboard their products, using machine learning to make them self-learning and predictive. For more information, visit www.ekkono.ai.

[†] fika (wikipedia.org); Swedes have fika (pronounced [ˈ fiːka]), meaning "coffee break". The tradition has spread throughout Swedish businesses around the world. Fika is a social institution in Sweden and a common practice at workplaces in Sweden. Fika may also function partially as an informal meeting between co-workers and management people, and it may even be considered impolite not to join in.