Measure energy consumption

Highlights

•

Literature review of energy estimation methods from computer architecture for machine learning applications.

•

State-of-the-art approaches to estimate energy consumption in machine learning.

•

Software tools from the power and performance monitoring field and their applicability to machine learning.

•

Two use cases to estimate energy consumption for deep learning and data mining.

Abstract

Energy consumption has been widely studied in the computer architecture field for decades. While the adoption of energy as a metric in machine learning is emerging, the majority of research is still primarily focused on obtaining high levels of accuracy without any computational constraint. We believe that one of the reasons for this lack of interest is due to their lack of familiarity with approaches to evaluate energy consumption. To address this challenge, we present a review of the different approaches to estimate energy consumption in general and machine learning applications in particular. Our goal is to provide useful guidelines to the machine learning community giving them the fundamental knowledge to use and build specific energy estimation methods for machine learning algorithms. We also present the latest software tools that give energy estimation values, together with two use cases that enhance the study of energy consumption in machine learning.

Introduction

Computer architecture researchers have been investigating energy consumption for decades, especially to be able to deliver state-of-the-art energy efficient processors. Machine learning researchers, on the other hand, have been mainly focused on producing high accurate models without considering energy consumption as an important factor [18]. This is the case for deep learning, where the goal has been to produce deeper and more accurate model without any constraints in terms of computation. These models have grown in computation (typically in the GigaFlops) and memory requirements (typically in the millions of parameters or weights). These algorithms require high levels of computing power during training as they have to be trained on large amounts of the data while during deployment they may be used multiple times. Some awareness in energy consumption is starting to arise, originating from a few machine learning research groups [12], [14], [47], [61] and challenges such as The Low Power Image Recognition Challenge (LPIRC) [26]. Thus, we believe that efforts towards estimating energy consumption and developing tools for researchers to advance their research in energy consumption are necessary for a more scalable and sustainable future.

We believe that the reasons why the machine learning community has not shown more interest in energy consumption is because of their lack of familiarity with the current approaches to estimate energy and the lack of power models in existing machine learning frameworks, for example, in Tensorflow [1], Caffe2 [43], PyTorch [56] and others to support energy evaluations. This study addresses this challenge by making the following contributions:

i.

We present a literature review of different energy estimation approaches from the computer architecture community (Section 4). We synthesize and classify the papers into high-level taxonomy categories (Section 3) and modeling techniques to enable a user from the machine learning or computer architecture community to decide which estimation model could be used or built for a given scenario. We also present the advantages and disadvantages for each category.

ii.

We present the current state-of-the-art approaches to estimate energy consumption in machine learning (Section 6).

iii.

We present the currently available software tools and present their characteristics to the user to facilitate building energy consumption models (Section 5). We categorize the tools based on the granularity of the energy estimations, software that is supported, precision etc.

Background

This section explains the main concepts and terminology used throughout the rest of the paper. Energy, measured in joules (J), is the total power consumed during an interval of time. Power, i.e., the rate at which energy is consumed, is the sum of static and dynamic power. Static power, also known as leakage power, is the power consumed when there is no circuit activity.

Taxonomy of power estimation models

Power models are built to design better hardware, design better algorithms or design better software to map these algorithms onto hardware. Following the abstraction levels at the hardware–software stack, there already exists a taxonomy of power models ranging from low-level transistors to high-level system description of hardware components, such as processor, cache, bus and others [73]. In this paper we propose a more detailed taxonomy that better aligns with our goal of mapping energy estimation techniques to machine learning applications. We focus on power estimation models that can be built at the system-level to understand the energy consumption at the application level. We propose the following taxonomy of power estimation models at the system-level:

•

Software level: The developers of the model at this level are interested in the energy consumption of the application or software implementation and explore optimization techniques that include designing efficient algorithms or better software implementation of the algorithm.

–

Application-level: At the topmost abstraction level, a power consumption model can be built by relating algorithmic properties of the application directly to the power estimation. Here, the developer of the power models extracts characteristics of the application, for example, kernel sizes in a neural network, and relate it to the energy profile of the application.

–

Instruction-level: At the next level, a power consumption model can be built to understand which specific instructions in the program contribute to the energy consumption. The instruction traces can be extracted using an instruction-set simulator or performance counter profiling, and the cost for each instruction can be added either by known or relative the cost of the instruction or experimental data. This can be applied to estimate the energy consumption of the different functions of machine learning algorithms. This is useful for understanding which parts of the algorithm are consuming most of the energy, to focus the efforts on reducing the energy consumption of such parts .

Approaches to estimate energy consumption

The goal of this section is to introduce key approaches to estimate energy to the machine learning expert. First, we give a general overview of the energy estimation field, providing the reader with the basic knowledge. Second, we explain in detail how the energy estimation models are built, to give the machine learning expert or computer architecture researcher the starting point to build or use more specific machine learning energy models.

The papers reviewed in this section are chosen from a more general survey [30] on power measurement, estimation, and management. We also include papers that we discovered from researching the field on ways to estimate the energy consumption that could be applied in machine learning scenarios. This was achieved by searching on online databases (e.g. Google Scholar) and by looking at the references from some key papers in the area.

The surveyed papers can be clustered into four groups: (i) papers that obtain the activity factors with performance counters and use regression or correlation techniques to obtain the power or energy; (ii) papers that use simulation data to obtain the activity factors; (iii) papers providing architecture or instruction level information; (iv) papers that provide real-time power or energy estimation

Conclusions

Machine learning algorithms consume significant amounts of energy. However, the lack of evaluations based on energy consumption of these algorithms can be attributed to the lack of appropriate tools to measure and build power models in existing machine learning suites, and because estimating energy consumption is a challenging task.

This paper addresses that challenge by presenting a review of the key approaches to estimate energy consumption from the computer architecture field, mapped to machine learning applications. We also describe the state-of-the-art methods to estimate energy consumption in particular for data mining and convolutional neural networks. Our synthesis of the surveyed papers provides the necessary guidelines to expose energy consumption methods to machine learning audiences interested in incorporating energy as metric in the design of machine learning systems. To demonstrate the usefulness of the synthesis, we present two use cases, which show, from the data mining and neural networks perspectives, how to apply the different estimation approaches. We show that the benefits of further research in energy estimations can help machine learning researchers gain significant insights when building machine learning systems.

Our survey also reveals the current state of energy estimations in machine learning. In particular, there are several works emerging to enable energy evaluations in machine learning either through energy prediction modeling as seen in NeuralPower [12] or by direct integrating power monitoring tools to existing machine learning suites as seen in SyNERGY [61]. However, current modeling approaches face another challenge; which is the rapid changes in neural network designs, implementations and hardware. Moreover, there is a fragmentation of the machine learning software ecosystem with multiple software suites and no common benchmarking suites. These modeling approaches will have to be adaptable to these changes and also be comprehensive. This is because most works target a few computational intensive layers with majority of the work confined to just convolutional neural networks.

Another area in which energy estimation is lacking are GPUs that are extensively used by machine learning researchers to train machine learning models. Few works such as NeuralPower [12] build desktop GPU-based energy estimation models while others target mobile CPUs and GPUs [61]. Future work will require further research to integrate the literature on GPU-based estimation methods applied for machine learning scenarios. Finally, there is a recent upsurge to build application specific hardware for machine learning [45]. To keep pace with both advancements in neural network designs and advances in the hardware community will require energy estimation models for these new emerging architectures either gathered on real systems or through neural network hardware simulators such as ScaleSIM [63].