Minmaxing the energy efficiency of biological computing

A study on the energy efficiency of Ph.D. students for neural network inference and other computing shenanigans

Hugo Waltsburger[†‡]

[†]SONDRA, CentraleSupélec, Université Paris Saclay, Gif-sur-Yvette
Hugo.Waltsburger@centralesupelec.fr
[‡]Université Paris-Saclay, CentraleSupélec, CNRS, Laboratoire de Génie Electrique et Electronique de Paris, 91192, Gif-sur-Yvette, France.
Sorbonne Université, CNRS, Laboratoire de Génie Electrique et Electronique de Paris, 75252, Paris, France
Hugo.Waltsburger@geeps.centralesupelec.fr

Yes Anthony, the Ph.D. manuscript is coming along.
No, I am not getting sidetracked haha, I would never.
No, you need not prepare the crucifixion nails.

-The author to one of his (much beloved) Ph.D. supervisors, probably

***Abstract*—Who are we? What is our purpose in life? This paper will not answer these questions, but will nonetheless hint at the possibility that, if humans' evolutionary purpose is neural network inference, then we are awfully badly designed. We will also answer some fundamental philosophical questions, such as "How many human computers do we need to play Doom?".**

I. Introduction

A. Context

The artificial neurons that make up modern neural networks have barely changed ever since their initial appearance in 1943 [1]. They went through two successive cycles of hype and disappointment (the so-called *AI Winters*), before beginning their third hype cycle in the early 2010s decade, a hype that lasts to this day. Will it last? It's hard to make predictions at this point. Large language models (LLMs) have had some spectacular results. But with modern LLMs having been trained on the almost complete dataset of data available on the Internet using tens of thousands of GPUs for months, the path of least resilience that was "throw more layers, data, and GPUs at the problem until there's no more problem" is bound to break down sooner or later.

With this matter comes an efficiency problem. The energy footprint of high-performance computing is significant. The creators of the open-source LLaMa II 70B LLM have estimated its training-related greenhouse-gas (GHG) emissions to 500t of carbon dioxide equivalent (CO_2e) [2]. This amounts to 100 times the target yearly emissions of an average European citizen in 2030, as per the Paris Agreement. This figure does not count the energy consumption of inference. Some modern state of the art neural networks are so expensive to train that it is doubtful that anyone will try to retrain them from scratch before they are obsolete - which at the current speed, may be a matter of months, a couple of years at most. Modern neural networks become more susceptible of consuming more energy during inference than during training. There is a growing importance of measuring the environmental impact of artificial intelligence, to create more energy-efficient networks. A first (modest) attempt was made by yours truly in [3].

B. Getting efficiency out of the way

Now that the mandatory serious and forward-thinking part of this paper is out of the way, we can think backward: what would be the most efficient inefficient way of performing the tasks that neural networks do? Our initial idea started with the procurement of a monkey, a typewriter, and a couple trillion years (along with lots of ink ribbons and a supertanker worth of paper). Unfortunately, monkeys are not considered pets in France (thus demanding special authorizations, which I don't have, to take care of one), and typewriters have become vintage - thus expensive. So I settled on the thing closest to a monkey with a typewriter that was readily and cheaply available: a Ph.D. student (myself) with a computer[1]. In this paper, we will explore how a human subject can optimize its neural network inference performance.

[1]The quality of the resulting thesis should be about equivalent

II. Historical Perspective

In the early days of computing, a lot of calculations were performed by hand, filling notebooks worth of computations. This led to the creation of a profession, "Computer", whose job would be to carry out much of the lengthy calculations that mathematicians did not have the time or want to perform themselves. The First World War, a period fraught with ballistics calculations, led to the employment of many computers. When mechanical and electronic computers appeared, these human computers were also tasked with programming them - a tedious task involving wires and punched cards. Oftentimes, these venues would figure among the few opportunities for college-educated women or minorities to enter an ""academic"" career. The early days of computer science saw many such people, whose opportunity came from the fact that computer science was generally considered a "less noble" field than theoretical science. The reader interested in a deeper exploration of human computers may refer to [4] or [5]. In the brief blip of time we are currently experiencing where computers are silicon-based rather than organic (between these past-but-not-so-remote times, and the mentats of the post-Butlerian-Jihad era – it's coming and you know it), it may be hard to evaluate how easy we have it with calculations. This paper will remind the modern reader of this fact, and hopefully cement my position as an early prophet when the Butlerian Jihad comes and AI gets relabeled as either "Abominable Intelligence", or SALAMI (Systematic Approaches to Learning Algorithms and Machine Inferences).

III. Motivation

At this point, I would honestly do just about anything to avoid writing my Ph.D. manuscript[2].

IV. Experimental setup

Ph.D. students are typically human, and adult[citation needed]. In this specific setup, the author is a human male, aged 28 solar years. The average human male adult needs to consume, on average, about 10.6MJ (2,500kcal) of energy each day [6], [7], in the form of food. The use of a most-likely-not-very-scientific online calorie calculator (taking age, height, weight, and physical activity into account) places the author at a remarkably average energy consumption of 10.2MJ/day (2,437kcal/day). In the following, we will assimilate the author with the average male Ph.D. student. This yields an average thermal dispersion power (TDP) of around 120W. The study of the influence of intellectual effort on the metabolism of the body and the brain is a whole can of worms of contradicting and/or inconclusive results [8] that I am not willing to touch with a 10-foot pole. Thus, for the sake of this experiment, we will simply assume that, for a given experimental setup, our computer's TDP remains constant over time.

Now, how can we perform neural network inference?

[2]In case my supervisors ever come across this paper: it was *mostly* written in my free time. Its characterization as an exercise of procrastination is left entirely to the reader's evaluation

Neural networks' workloads are mostly made up of multiply-accumulates (MACs) – essentially matrix multiplications – and composition by nonlinear functions of variable complexity. Computers typically store numbers in the form of 32-bit floating point numbers. This number format comprises three parts: the sign bit, the exponent, and the significand (sometimes called the mantissa).

We won't go into too much detail, but people interested in the eldritch horror that is the IEEE-754 floating point standard may refer to my Ph.D. thesis for more details. It's gonna be great. I swear.

The way floating point additions and multiplications are computed can be visualized in figures 1 and 2.

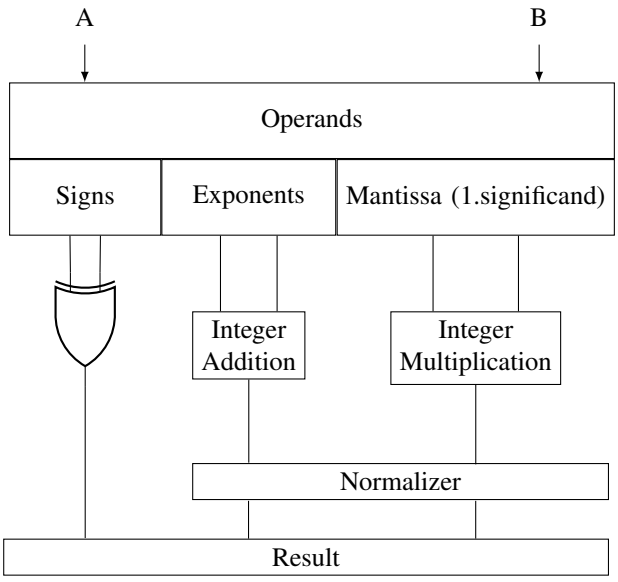

Figure 1: A float multiplier[3]

Look at those drawings, look at them well. Think of the obscene amount of time it took me to make them. Did you give them enough appreciation? Yes? Are you sure? Then we can move on.

I initially set out to try and measure my throughput in terms of manual floating-point multiply-accumulates, equipped with just my brain, a pencil, and some paper. After much effort, reading of low-level logic, and a lot of swearing, I concluded that performing floating-point operations manually required a level of masochism that not even I have. And that is saying something.

Fortunately, machine-learning researchers have come up with ways of simplifying the compute load of neural networks by changing the number representation of a neural network's weights, a process that is called quantization.

[3]The diagrams are on my GitHub so that nobody ever needs to import grainy JPGs, or suffer like I did to redraw them: https://github.com/frost-is/TikZ-Diagrams

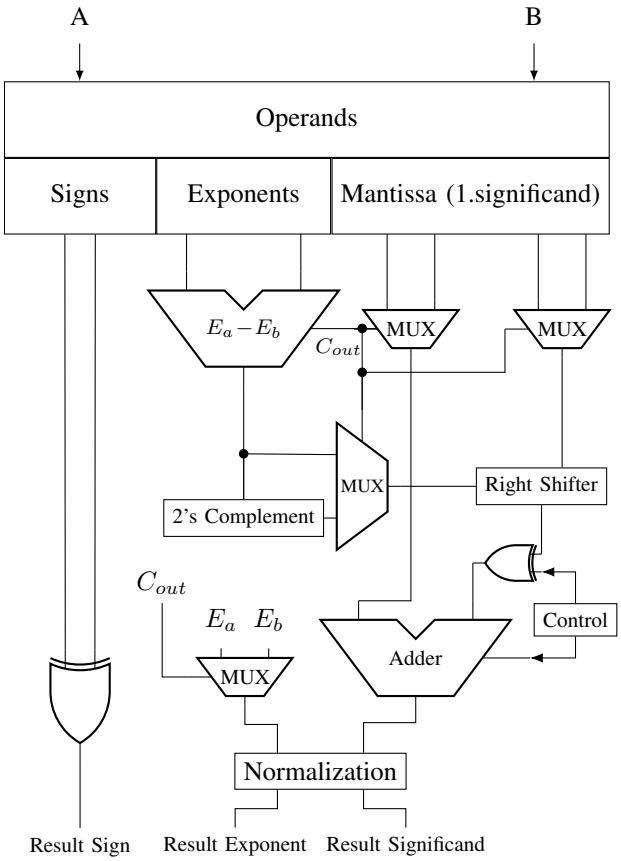

Figure 2: Seriously, who came up with this monstrosity? It took me ages to do a proper Tikz diagram[3]. This is a float adder, slightly simplified, but not by much.

The way it is usually done is by using the INT8 or INT16, signed integer formats. Here, for simplification purposes, we will consider the INT8 format, yielding a number range in $[\![-128; 127]\!]$. Using a typical pen and paper, averaged over a couple of tens of randomly generated operations and counting conservatively, the author of this paper needs slightly under 30 seconds to perform an INT8 multiplication and about 10 seconds for an addition. For the activation functions, we should be able to refer to computation tables - since INT8 can only take 256 different values, these tables could fit on a page each. The bulk of the compute load should be made out of MACs, since the amounts of MACS we need to perform in layer k is *number of neurons in layer* $(k-1) \times$ *number of neurons in layer* k, when the number of activations to compute is only *number of neurons in layer* k. Taking this into account, plus the relative simplicity of computing activations using tables, we will consider that the compute load of neural inference is solely made up of multiplications and additions, each equally represented.

If we want to try our hand with a rather lightweight neural network, we can try MobileNetV3 Small[9]. It comprises 44 million MACs. Following the logic we developed above, it should take about 28 years of continuous computing to perform a single inference, for a total inference cost of 105GJ of energy. To give an order of magnitude, this represents about 3 minutes' worth of the full power of a nuclear reactor. So *a lot* of energy. This also means that the author of this paper, working 35-hour workweeks with 6 weeks of paid leave *per annum*, should take about 150 years to compute a single inference[4].

V. This is bad. Can we do better ?

A. Slide Rules

Slide Rules allow us to change multiplications into additions through the power of logarithms. We therefore reach a throughput of one operation per 10s for both additions and multiplications – a twofold improvement. We will, however, consider that there are not enough of them left around to use at a truly large scale.

Figure 3: The slide rule. An elegant weapon for a more civilized age. This one was my late Grandma's

B. Sex

According to the French medical database Vidal [7] and the British National Health Service (NHS) [6], adult men need on average about 10,500kJ of food each day. Women, however, only need about 8400kJ. Thus, female Ph.D. students are on average 20% more energetically efficient than male Ph.D. students. To maximize efficiency, hiring female Ph.D. students for our inference setup seems the way to go, bringing us down to 85GJ per inference, albeit with identical inference time.

[4] Just in time for my retirement

C. Physical activity

The energy-consumption figures cited above are for average humans, who are moderately physically active. However, if we take humans and have them live where they work and not do anything besides working and sleeping, then their energy consumption is only their basal metabolic rate (BMR).

The topic of BMR is a surprisingly complex one, as BMR varies based on age, sex, ethnicity, muscles, etc. We will not delve too deep into it, but a good intro can be found in [10]. According to this data, the BMR of men and women between 18 and 30 years old can be approximated using the following formulas: $0.0546 \times \text{weight} + 2.33$ for women, and $0.0669 \times \text{weight} + 2.43$ for men (in $MJ.day^{-1}$). Now, this estimation requires data on the average weight of men and women. This statistic is highly variable depending on the country of origin, so here we will assume that our subjects are French. We will also neglect the effects of socioeconomic status on weight.

The Obepi-Roche study indicates that French men and women have an average weight of 81kg and 67kg, respectively [11], [12]. This means that the average French Ph.D. student's BMR is about 6MJ/day (1,435kcal/day) for women and 7.85MJ/day (1,880kcal/day) for men. The logical conclusion that follows is that, if we seal our Ph.D. students in capsules that prevent them from moving, we consume 25 to 30% less energy. This also means that we can make them work 16 hours a day 365 days a year without compromising their sleep too much. This won't change the amount of energy used, but will nonetheless reduce our inference time, down to just a little under 42 years per inference.

D. Age

One way of explaining the efficiency of GPUs for neural network inference and training – and video games – (or, at least, the way I explain it to those among my students who have little background in computer science or electronics) is the following:

"You can compare a CPU to a mathematician and a GPU to a bunch of middle-schoolers. The mathematician can perform much more complicated operations. However, it takes time and resources to raise someone into a mathematician - and to have them perform, since they consume more food. This means that, for the same cost, you can have more middle-schoolers than mathematicians. Mathematicians can compute operations that middle-schoolers are incapable of. However, as far as simple operations - such as additions and multiplications - are concerned, they are only marginally better than middle schoolers. On the opposite end of the spectrum, if your computations are complex and data dependent, with a lot of logic and branching, then middle schoolers will be slower – they will take a lot of time to break the task into many smaller tasks, often they will have to wait for one of them to give the output of one operation before being able to advance on the task, and this will typically make them slower. This translates to CPUs being specially adapted for serialized, data-dependent operations with branching and logic, while GPUs are a better fit for parallelized, data-independent operations including little logic.

Let us assume mathematicians can do 3 times as many simple operations as middle-schoolers per unit of time. If the cost of a middle-schooler is less than a third of that of a mathematician, then buying middle-schoolers[5] *will be more efficient than buying mathematicians*[6]*. There are caveats. This supposes that the task can be split into simple chunks that can be processed independently (easily parallelizable). This is the case for machine learning and videogames/rendering: they are composed of simple, repetitive, data-independent computations with little logic - which mainly comprise additions, multiplications, and compositions by some basic functions. Thus, an army of middle-schoolers will tend to perform them more quickly than a handful of mathematicians bought for the same price. Although you will usually need mathematicians to split the task into chunks that middle-schoolers can process.*

As an example: the latest Nvidia H100 (PCIe version) comprises 14,592 CUDA cores, each capable of a single floating multiply-accumulate per instruction. The largest modern CPU, the EPYC 9754, made by AMD, reaches 128 cores [7]*. Even using vector instructions such as AVX512, which can process 8 32-bit multiply-accumulates per clock cycle, we only reach 1024 multiply-accumulates per instruction at full load – under ideal conditions. That is 14 times fewer than the H100. Both of these processing units have equivalent silicon areas and transistor counts – Nvidia's H100 boasts an 814mm^2 monolithic die comprising 80 billion transistors, and AMD's EPYC 9754 implements 8 chiplets of 73mm^2 die area each (584mm^2 in total), plus a 391mm^2 I/O die, for a total of 981mm^2 comprising 71 billion transistors. Their respective thermal dispersion powers are also equivalent, 360W for the EPYC and 350W for the H100.*

Now, this explanation is very much simplified and would warrant a deep dive into the respective capabilities and tradeoffs of CPUs and GPUs – GPUs can't run operating systems, for example. But that's the intuition."

So, what if we use middle schoolers instead of Ph.D. students for our computations? Our best review of the literature did not find a satisfactory answer. One reason is that this axis of research is supremely inefficient and profoundly unethical.

[5]The author speaks from a purely theoretical point of view and certainly does not condone buying middle-schoolers

[6]The author does not condone buying mathematicians either

[7]We won't account for AMD's simultaneous multi-threading (SMT) or Intel's Hyperthreading, since the "operations per instruction" metric should, theoretically, not benefit from it. Almost all modern CPUs implement 2 threads per physical core, but it is not equivalent to having 2 physical cores

Another is that it's extremely difficult to find statistics for children, for understandable reasons. In any case, we were not able to find data for this age bracket.

Nonetheless, the World Health Organization (WHO) provides weight-for-age charts for children of 5 and below [13]. These charts are meant to help pediatricians determine the extent to which a child's physiological needs for growth and development are satisfied. These statistics go up to 6 years old for boys and 5 years old for girls. To keep the comparison fair, we will compare 5-year-old children. Their median weight is 18kg. The equations in [10], taken for children between 3 and 10, give us a median daily BMR of 3.60MJ/day (870kcal/day) for girls and 3.80MJ/day (920kcal/day) for boys, which is a significative gain in efficiency compared to Ph.D. students and would amount to 55GJ (resp. 58GJ) per inference.

There is, however, a caveat: to what extent can one teach 5-year-olds how to add and multiply, and how efficiently do they accomplish this task? This warranted inquiry. I went to my mother, a Kindergarten teacher, and asked her if I could borrow a couple of her kindergarteners to run an experiment. Her enthusiasm was rather lukewarm ("What?! Why? No!"). None of my friends have children in the right age bracket, and since I would do anything for Science, so long as it does not involve talking to strangers, this was the end of this way of exploration. Feel free to contact me if you have subjects available for this ethically undefendable but scientifically glorious experiment.

VI. Optimizing against carbon emissions

Optimizing biological neurons against energy is not necessarily relevant, as the biologically exploitable energy of food does not reflect the energy that was expended to produce it. This means that creating an accurate reflection of the energy efficiency of humans would need to take into account the entire supply chain, which would represent a tremendous work. Here, we argue that optimizing our workflow with regards to carbon emissions is more relevant, as carbon emissions a) are positively correlated with energy emissions, b) have direct consequences on our biome, where energy expenditure does not – or to a smaller extent –, and c) I'm starting to feel guilty of over procrastinating so I need to wrap this up.

Therefore, our problem of minimizing energy consumption can be simplified by trying to minimize carbon emissions. So, how can we do it? The origin of the food we feed our computers seems crucial here.

From a purely energy-efficient standpoint, vegetable oil is about as good as can be for humans, with energetic values around 37MJ/kg (8850kcal/kg)[8]. How does it translate in terms of carbon emissions? There are discrepancies when it comes to the type of seed the oil comes from. Rapeseed oil seems to be the most carbon-efficient crop for oil production [14], with median emissions of 2.49kgCO_2e per kg. This translates to about 14.85MJ/kgCO_2e (3,550kcal/kgCO_2e). But if we try to tackle the problem from a carbon emission standpoint, then we can further optimize the MJ/kgCO_2e metric.

Supposing that we have the logistics available to plant and process crops on site (a good way of mitigating, if not negating, carbon emissions related to transport, packaging, retail, and waste), then using the data from [15] and only accounting for the emissions related to farming, animal feed (if applicable) and processing, then we obtain table I.

Food	MJ/kg	kgCO_2e/kg	MJ/CO_2e
Rapeseed oil [14]	37 (885kcal)	2.49	14.9 (3,550kcal)
Potatoes [16]	3.6 (87kcal)	0.19	18.9 (4,578kcal)
Peanuts [17]	23.7 (567kcal)	0.6	39.5 (9,450kcal)

Table I: Some of the least greenhouse-gas emitting foods

So it seems peanuts are the most efficient crop as far as energy per kgCO_2e is concerned. Moreover, peanuts are edible raw, and since cooking leads to more GHG emissions, this further reduces our greenhouse gas footprint. For confirmation purposes, the author ate an entire potato raw to determine whether raw potatoes could reasonably be the sole component of a person's alimentation without inducing lifelong trauma, and the conclusion was a resounding "No".

Unfortunately, cal/kgCO_2e is not the only metric we need to account for. Humans are pesky creatures that, unlike computers, require more than just energy to function properly: they require various nutrients - among them proteins, fat, carbohydrates, and miscellaneous vitamins and minerals. Our best review of the literature did not come up with any metric measuring how much of a human's physiological needs are met by any given food. However, peanuts do contain a good amount of proteins, fibers, carbohydrates, and fat. Trying to come up with a mix of what could be added to peanuts to be the sole basis for humans would be way too involved for this paper, which already went way too far for something that was initially supposed to be a joke. However, given the fact that the "Plumpy'Nut" ready-to-use therapeutic food (RUTF), which was created for the treatment of severe acute malnutrition, uses peanut paste for its basis (it contains peanut paste, vegetable oil, powdered milk, powdered sugar, vitamins, and minerals), we feel confident that peanuts are, if not the best choice, among the better ones.

So, according to these metrics, a farm of French Ph.D. students doing nothing but computing 16 hours a day and sleeping the rest of the time would need to feed its computers 250g (women) to 330g (men) of locally produced peanut slurry per day. We will assume that men and women are equally represented in our setting, thus yielding an average of 290g of daily peanut requirements per person or about 106kg per year. This translates to 175g of CO_2e emissions

[8] This is about 1% of the energy density of reactor-grade uranium and 0.02% of that of pure Uranium 235. Unfortunately, humans tend to have a hard time digesting radioactive isotopes, which is a shame

per day and 65kg per year. The best lands give a yield above 4t of peanuts per ha [17], with harvest taking place once a year. Given that feeding the entire population of France using peanut slurry should need around 6.9Mt of peanuts per year, locally producing them would require dedicating 17,200km² of arable land to peanuts, or $\approx 15.6\%$ of the total land area used for agriculture in France in 2020[18].

This means that, thus far, our grand plan of turning the entire French land and population into a gigantic computer farm is on track, and would yield about 3.2 INT8 MOPS of computing capabilities 16 hours a day, or 2.2 INT8 MOPs when taking shifts. Nice[9].

Now, there are some caveats:

1) As the scientifically shaky but visually grand experiment known as *Christopher Nolan's Interstellar* has exemplified, monoculture is typically not a great idea in the long run, for a variety of reasons (mainly vulnerability to pests and diseases and lack of genetic variability). This is not a problem because, like all cartoonishly evil geniuses, I will be thwarted at the last second by a hero in shining armor straight from the old stories. This protagonist will either make me realize the error of my ways via the power of love and friendship (I certainly hope not), or defeat me in a battle of wits, whereupon I shall deliver my *Evil Monologue*™ (which I may or may not have already written while procrastinating on my Ph.D. dissertation), ending with "Someday you will realize that I was right, Iggy MacRainbows", before falling down the shaft of my soon-to-explode ~~Death Star~~ server farm, cackling in laughter. Of course, my death will be unconfirmed depending on whether box-office results determine that my story deserves a sequel. I give the aforementioned sequel 50/50[10] chances of either butchering my character or giving me an awesome redemption arc. I honestly can't wait.
2) This study does not include the use of water. Humans require water. However, this would send us on the tangent of the tangent, and I'm already way late on my Ph.D. redaction so I ask for your understanding.
3) Nourishment by peanut slurry only would lack some essential nutrients. If we seek to include some vitamins and minerals in our slurry, then the emissions should be higher than announced here. This one will the tackled in the following section

So, to try and preserve our human computers for as long as possible, we will need to introduce vitamins and minerals in our peanut slurry, which will induce further GHG emissions. However, we argue that these elements are effective in such small quantities that artificial synthesis amounts to a negligible amount of marginally added carbon emissions. In [19], it was determined that adding artificial D vitamin to bread, milk, and oil, only increases the food's carbon footprint by $\approx 0.001\%$. Thus, we will assume that the enrichment of our peanut slurry is negligible in terms of added carbon emissions.

Moreover, human bodies are reasonably nutritious: according to [20], a male human body of 55.26kg contains 32,376kcal of skeletal flesh. Should we eat the entire body (marrow, skin, adipose tissue, etc.), then we reach 143,771kcal per body[11]. The study does not provide figures for female bodies as no chemical composition of the female body was available in the literature at the time of writing. If we do a back-of-the-envelope calculation assuming that these figures scale linearly and identically with weight for both sexes (they don't but we'll soon show that this ends up amounting to a rounding error), we can infer that the average French person's body contains about 190,000kcal, brain and rachis excluded (172,000 and 208,000kcal for women and men, respectively). If we liquefy the corpses of our dead computers via a comically large meat grinder to feed them to living computers, then this positively influences our carbon efficiency. Let us work out how much.

French data shows that 638,266 French people died in 2023 and that the French population at the time of writing is 64,842,629. The death of about 1% of the population each year for a country whose population is growing seems like a reasonable assumption in general. Assuming men and women are equivalently represented in the French population, dead bodies would average in at about 5kcal of food per living French person per day, about 0.3% of the average French person's BMR. Thus, the bulk of our computers' emissions will come from peanut culture, and we will simply assume that the carbon impact of the additives in the peanut slurry is offset by the corpse slurry.

Incidentally, we have just demonstrated that fundamental assumptions underlying the plots of *Soylent Green* and *The Matrix* cannot hold in real life.

VII. Okay but can we do better?

If we go out of the realm of what is currently achievable with 2024 technology, then we can further optimize: the human brain consumes about 20% of the entire body's calories [22]. If we remove people's brains and have them live in small vats filled with nutritive substances[12], then our apparatus only consumes about 20W per human. Estimating the nutritional needs of a brain in a vat is hell. Considering a brain cannot digest anything, nutrients must be supplied in a heavily processed

[9] The astute reader may remark that H100 GPUs boast 3,026 INT8 TOPS. They are, however, much less visually entertaining

[10] 100/0 if Disney buys the adaptation rights

[11] Although we advise against eating the brain, as eating the brain of an individual of one's species leads to higher rates of prion diseases, typically spongiform encephalopathy (the so-called "mad cow disease") [21]. Discarding the brain, rachis, and spinal cord would reduce the caloric value of the male human body by 2706kcal.

[12] Any resemblance with parts of Metal Gear Rising Revengeance's plot will result from pure chance

form, and we can throw our entire previous calculations out the window. So we won't explore this venue any further.

VIII. What tasks can we perform?

A. Neural Network Inference

Some of the most popular tasks performed by modern-day neural networks are computer vision and text generation. We have exemplified a single neural network inference of MobileNetV3 Small costing 42 years of computer inference, or 95MJ of energy on average. Compared to that, using an Nvidia A100 costs about 3.6mJ per inference [3], or 26×10^{12} less energy.

Now, how does that translate in terms of carbon emissions? Energy-related carbon emissions are a complicated endeavor: when adding load to the energy grid, pilotable energy sources (gas, oil, coal) will ramp up to keep the voltage and frequency at the desired level. So estimating the marginal impact of adding a load to the grid in terms of carbon emissions as *added load* $\times$ *average emissions per unit of energy* is a bit of an oversimplification. We will do this oversimplification, as we assume that the added load that is our inference server is small enough (<1kW) to blend in the base power mix, but this assumption does not hold at larger scales. In 2023, the average energy-related carbon emissions in France were $39gCO_2e/KWh$ [23]. Thus, our A100 would emit $3.5\mu gCO_2e$ per inference. Our human computers, on the other hand, would emit about $2.7tCO_2e$ to perform the same task, around a trillion times worse.

Now, humans can have a very good level of performance provided sufficient training. Russakovsky et al. [24] note that the best human classifiers on the ImageNet dataset only had a 5.1% error rate, much better than MobileNet S's 12.7% error rate.

It seems a reasonable approximation to assume that a good human classifier should take about 10s for a single inference, yielding an energy efficiency of 1kJ per inference or $3mgCO_2e$ per inference, with much better accuracy and throughput than our server farm. So it seems that all of the pages above were but a fruitless endeavor. But a fun one to be sure![13]

B. Playing Doom

How many human computers do we need to play Doom? Well, a single one. With abysmally low framerates, of course. Can we do better? Surely!

Going through the manual of Doom (1993), we find requirements comprising a 386 CPU (or better), 4Mb of RAM, and a VGA card. Doom ran solely on the CPU, and the i386 is capable of around 5 MIPS. It is excruciatingly hard to assume the compute performance of a CPU based on its MIPS metric alone. However, we can safely assume that if we parallelize the workload over the entire population of France, then we have enough computers to dispatch the entire workload efficiently, if only since there are more French people than transistors on an i386 die (275,000 or 855,000 depending on the version[25]). The problem we encounter here is that Amdahl's law only goes so far - parallelizing over so many humans, we become latency-bound, not compute-bound. Taking into account our previous results, we conclude that we could probably play Doom at a maximum efficiency of about 0.1 frames per second, at best.

IX. Lifecycle analysis

At this point, it would probably be less expensive (and more interesting) to use the entire population of France to recreate an entire computer industry, software plus hardware, and reimplement games and neural networks entirely from scratch, than to use those same people as computers[14].

X. Conclusion and future works

In *The Matrix*, the machines went about as far as can be to exploit the brain power of humans efficiently while (relatively) preserving their complete bodily functions. It's just a shame that human brains are not very good at doing what processors do in the first place.

Also, energy-wise, that bit about liquefying the dead so that they can be fed intravenously to the living doesn't amount to all that much in the grand scheme of things. So Morpheus is being a tad dramatic here. However, it adds to the grimdark ambiance, so he gets a pass.

Finally, this paper allows us to remind people at family diners[15] that AI is mainly just linear algebra: using a pen and paper and *some* patience, anyone capable of performing high school level Mathematics can perform the same tasks as the latest AI model. Also it's neither intelligent, nor artificial, and you should call it SALAMI, at least you'd feel rightfully embarrassed when asking me whether linear algebra should be given the same rights as humans. I swear when the Butlerian Jihad comes you people are the first to go[16].

While this paper achieves little in the way of practical science, maybe the true science was the procrastination we made along the way. Now, there is a Ph.D. thesis to be finished.

XI. Ethics Statement

Creating labor camps, or worse, reproducing *The Matrix*™, is bad, and you should not do it.

XII. Special Thanks

I would like to thank the entire Sigbovik team for having me research way more biology than I would have ever liked. I would like to thank my Ph.D. supervisors for not immediately assassinating me, should they ever come across this paper.

[13] Statistics on the energy usage of LLMs will be present in my Ph.D. thesis, but I haven't had the time to process them for this paper, for which I apologize.

[14] Should politicians want to get in touch, my email is on the first page

[15] You know them

[16] Any resemblance to people or events existing or having existed would only be the result of pure chance

References

[1] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, pp. 115–133, 1943.

[2] H. Touvron, L. Martin, K. Stone, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[3] H. Waltsburger, E. Libessart, C. Ren, A. Kolar, and R. Guinvarc'h, "Neural network scoring for efficient computing," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2023, pp. 1–5.

[4] D. A. Grier, *When computers were human*. Princeton University Press, 2013.

[5] T. Haigh and P. E. Ceruzzi, *A new history of modern computing*. MIT Press, 2021.

[6] N. H. Service, "What should my daily intake of calories be?" *https://web.archive.org/web/20240220201035/https://www.nhs.uk/common-health-questions/food-and-diet/what-should-my-daily-intake-of-calories-be/*,

[7] Vidal, "Les recommandations nutritionnelles de 18 à 75 ans," *https://web.archive.org/web/20231110192642/https://www.vidal.fr/sante/nutrition/equilibre-alimentaire-adulte/recommandations-nutritionnelles-adulte.html*,

[8] T. scientific american, "Does thinking really hard burn more calories?" *https://archive.is/ExrgC*, 2012.

[9] A. Howard, M. Sandler, G. Chu, *et al.*, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.

[10] C. Henry, "Basal metabolic rate studies in humans: Measurement and development of new equations," *Public health nutrition*, vol. 8, no. 7a, pp. 1133–1152, 2005.

[11] L. C. l'Obésité (League Against Obesity), "Taille, poids et tour de taille : Photographie 2020 des français," *https://web.archive.org/web/20231208192705/https://liguecontrelobesite.org/actualite/taille-poids-et-tour-de-taille-photographie-2020-des-francais/*, 2020.

[12] A. Fontbonne, A. Currie, P. Tounian, *et al.*, "Prevalence of overweight and obesity in france: The 2020 obepiroche study by the "ligue contre l'obésité"," *Journal of Clinical Medicine*, vol. 12, no. 3, p. 925, 2023.

[13] W. H. Organization *et al.*, *WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development*. World Health Organization, 2006.

[14] T. D. Alcock, D. E. Salt, P. Wilson, and S. J. Ramsden, "More sustainable vegetable oil: Balancing productivity with carbon storage opportunities," *Science of The Total Environment*, vol. 829, p. 154 539, 2022.

[15] O. W. in Data, "Food: Greenhouse gas emissions across the supply chain," *https://ourworldindata.org/grapher/food-emissions-supply-chain?tab=table*, 2018.

[16] J. Singh and L. Kaur, *Advances in potato chemistry and technology*. Academic press, 2016.

[17] J. McCarty, S. Ramsey, and H. Sandefur, "A historical analysis of the environmental footprint of peanut production in the united states from 1980 to 2014," *Peanut Science*, vol. 43, no. 2, pp. 157–167, 2016.

[18] C. d'Agriculture, "Les chiffres 2022 de l'agriculture française," *https://web.archive.org/web/20240229192304/https://chambres-agriculture.fr/fileadmin/user_upload/National/FAL_commun/publications/National/Plaquette_chiffres_de_l_agriculture_VDEF.pdf*, 2022.

[19] M. J. Bruins and U. Létinois, "Adequate vitamin d intake cannot be achieved within carbon emission limits unless food is fortified: A simulation study," *Nutrients*, vol. 13, no. 2, p. 592, 2021.

[20] J. Cole, "Assessing the calorific significance of episodes of human cannibalism in the palaeolithic," *Scientific reports*, vol. 7, no. 1, p. 44 707, 2017.

[21] P. P. Liberski, A. Gajos, B. Sikorska, and S. Lindenbaum, "Kuru, the first human prion disease," *Viruses*, vol. 11, no. 3, p. 232, 2019.

[22] G. J. Siegel and R. W. Albers, *Basic neurochemistry: molecular, cellular, and medical aspects*. Raven Press, 1994.

[23] "French historical data - carbon emissions for energy generation," *https://www.nowtricity.com/country/france/*, 2023.

[24] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[25] "Microprocessor quick reference guide," *https://web.archive.org/web/20240316123819/https://www.intel.com/pressroom/kits/quickreffam.htm*, 2008.