AI, Solarpunk, and an Uncertain Future in Computing

rolltime (*she/her*)

July 12, 2024

AI, Solarpunk, and an Uncertain Future in Computing

rolltime (she/her)
July 12, 2024
Licensed under CC-BY-SA

Thank people, now that they're placated, let's talk about me!

Introduction

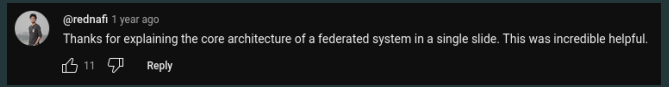
- I'm rolltime!
- Hacker, Software Engineer, Optimist

• I'm rolltime!
• Hacker, Software Engineer, Optimist

Recently finished my BA, wrote my thesis on Solarpunk and AI, I'm the lead engineer at a startup doing neither of those things.

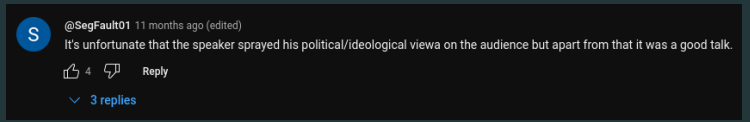
Now I want to bring my findings here. It's my hope (ha) that you all leave this talk with newfound understanding and passion. This is my second HOPE talk.

I spoke about ActivityPub, the federated social networking protocol which was thrust into the spotlight following Elon Musk's acquisition of Twitter. I spoke about my vision for a better kind of social media and how I felt we could get there. The talk was fairly popular and got a lot of feedback. I really value feedback so I'd like to take the time to review some of that feedback now.



@rednafi: Thanks for explaining the core architecture of a federated system in a single siide. This was incredible helpful.

@forrest_wilkins: Loved this, great talk!



Segfault01: It's unfortunate that the speaker sprayed his political/ideological viewa on the audience but apart from that it was a good talk.

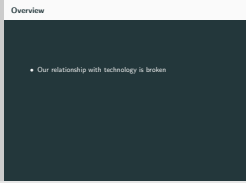
RADICAL SOCIAL MOVEMENTS

Radical social movements, ending capitalism, and curbing climate change. Segfault01, this one is on you.

@SegFault01

Overview

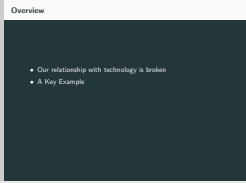
- Our relationship with technology is broken



Not just technology as it exists, but how we develop it

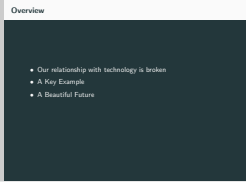
Overview

- Our relationship with technology is broken
- A Key Example



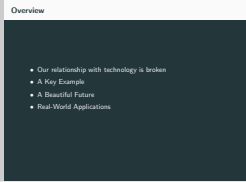
Overview

- Our relationship with technology is broken
- A Key Example
- A Beautiful Future



Overview

- Our relationship with technology is broken
- A Key Example
- A Beautiful Future
- Real-World Applications



Overview

- Our relationship with technology is broken
- A Key Example
- A Beautiful Future
- Real-World Applications
- Challenging the AI Mainstream

Overview
- Our relationship with technology is broken
- A Key Example
- A Beautiful Future
- Real-World Applications
- Challenging the AI Mainstream

AI: The Neverending Bubble

A Very Brief History

A Very Brief History

A Very Brief History

- 2016: RNNs, CNNs, GANs

A Very Brief History
• 2016: RNNs, CNNs, GANs

A Very Brief History

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover

A Very Brief History

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit

A Very Brief History
- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit

A Very Brief History

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2

A Very Brief History
- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2

A Very Brief History

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2
- 2022: Stable Diffusion

A Very Brief History
- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2
- 2022: Stable Diffusion

A Very Brief History

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2
- 2022: Stable Diffusion
- 2023: GPT-4, Claude 2, LLaMA

A Very Brief History
- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2
- 2022: Stable Diffusion
- 2023: GPT-4, Claude 2, LLaMA

GPT-4 was the first robot to pass the bar exam, unless you count Alan Dershowitz.

A Very Brief History

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2
- 2022: Stable Diffusion
- 2023: GPT-4, Claude 2, LLaMA
- 2024: GPT-4o, Claude 3.5, AlphaFold 3

A Very Brief History
- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2
- 2022: Stable Diffusion
- 2023: GPT-4, Claude 2, LLaMA
- 2024: GPT-4o, Claude 3.5, AlphaFold 3

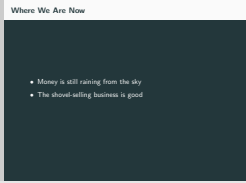
Where We Are Now

- Money is still raining from the sky

AI Ventures Attracted 96 Billion Dollars of Investment in 2023. Crypto: 30 Billion in 2021. AI: 42 Billion in 2021.

Where We Are Now

- Money is still raining from the sky
- The shovel-selling business is good



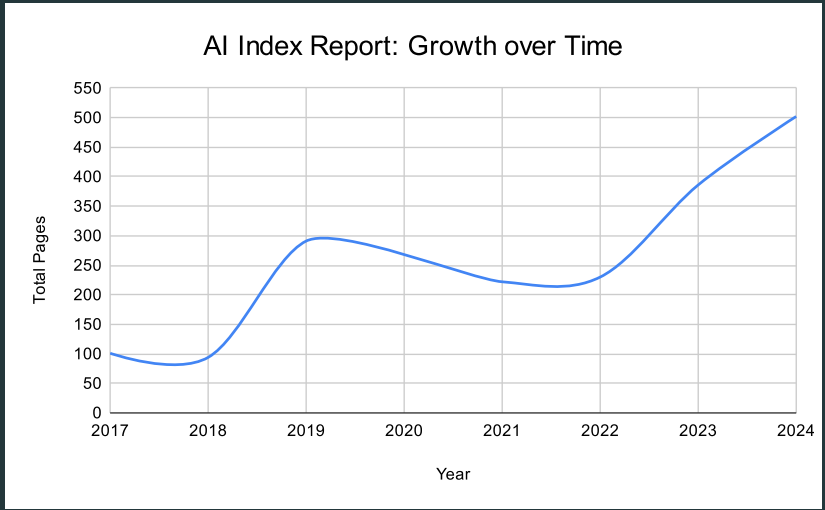
nvidia briefly became the world's most valuable company

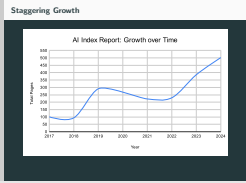
Where We Are Now

- Money is still raining from the sky
- The shovel-selling business is good
- Around 30% of people use ChatGPT weekly

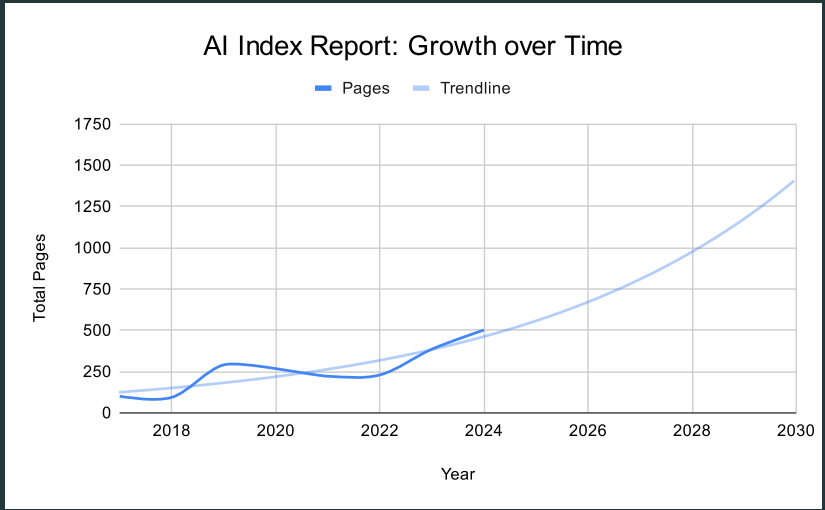
ChatGPT had the second fastest user growth, behind threads. This number is probably higher in the US.

Staggering Growth



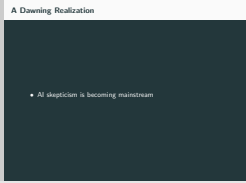


Staggering Growth



A Dawning Realization

- AI skepticism is becoming mainstream



More people are now saying what artists have been saying for months.
I will fucking piledrive you if you mention AI again - Ludicity
I'm going to ask ChatGPT how to prepare a garotte and then I am going to strangle you with it, and you will simply have to pray that I roll the 10% chance that it freaks out and tells me that a garotte should consist entirely of paper mache and malice.

A Dawning Realization

- AI skepticism is becoming mainstream
- Even Sequoia says it's unsustainable!

AI will need to generate an additional $600bn of revenue to become profitable

Energy Demand: Just How Bad Is It?

- Bad.

Billions of dollars into waste heat. One person flying from New York to LA results in a staggering two metric tons of CO2 emissions. Training GPT-4 cost OpenAI $100m in hardware and electricity and produced the equivalent of 2500 of these flights, or 156 person-years of CO2 for the average American.

Energy Demand: Just How Bad Is It?

- Bad.
- Training GPT-4: 2500 Flights, 100 cars

Training is energy-intense, but depending on model lifecycle, inference can be an even bigger problem. ChatGPT consumed an estimated 564 MWh per day in early 2023 to serve 195 million requests. Accounting for user growth and larger models, a conservative estimate has ChatGPT producing 100 flights per day of CO2.

Energy Demand: Just How Bad Is It?

- Bad.
- Training GPT-4: 2500 Flights
- ChatGPT: >100 Flights Per Day

Energy Demand: Just How Bad Is It?
• Bad.
• Training GPT-4: 2500 Flights
• ChatGPT: >100 Flights Per Day

Energy Demand: Just How Bad Is It?

- Bad.
- Training GPT-4: 2500 Flights
- ChatGPT: >100 Flights Per Day
- Google's carbon emissions up 48%

Google's carbon emissions are up 48% since 2019, mostly due to new datacenter demand. They have directly stated that AI is at least partially responsible; "reducing emissions may be challenging due to increasing energy demands from the greater intensity of AI compute".

They've also claimed that their AI-powered Search is ten times more expensive than the standard version. Based on that higher cost, we can estimate that if Google were to use AI for every search, the system's power usage might approach 30 TWh per year.

Energy Demand: Just How Bad Is It?

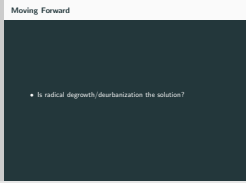
- Bad.
- Training GPT-4: 2500 Flights
- ChatGPT: >100 Flights Per Day
- Google's carbon emissions up 48%
- AI-Powered Google Search: 30 TWh

I've often heard it claimed that AI is using "as much power as a small country". That's actually false: it's as much as a large one. 30 TWh per year has Google alone tied with Bulgaria as the 66th largest consumer of electricity in the world, producing 1.8 megatons of co2 per year, or almost 2500 flights per day.

These numbers show a real, meaningful impact on the climate. That's scary, especially at a time when reducing carbon emissions should be humanity's number one priority. That raises the question: what do we do about it?

Moving Forward

- Is radical degrowth/deurbanization the solution?

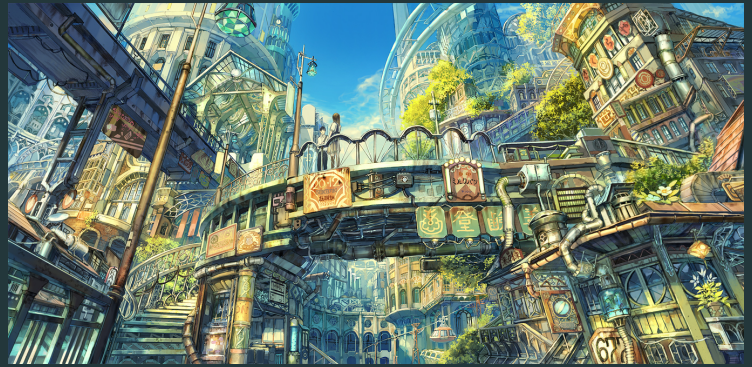


Why not just ban AI?

Neo-Pastoral Anarcho-Primitivist Cottagecore Bullshit

- Unrealistic; if we all move to the country it won't be the country anymore.
- Density isn't inherently bad for the environment
- Long-time human urge for a "simpler, greener" time before cities. Cottagecore
- Impossible to disentangle that urge from racism. White flight. Debt spiral.
- Not the first time; look at Andrew Jackson.
- We need a different dream. We need to believe in a world where the practical merges with the beautiful. Where dense urban environments thrive with lush greenery and sustainable food systems. Where optimism isn't a fallacy, but the path to a brighter future. We need Solarpunk.

Solarpunk



Munashichi, Future Economic View of Innocence, 2015

Solarpunk

Munashichi, Future Economic View of Innocence, 2015

Solarpunk is what happens when we dare to ask the question: What if we could do better?

Solarpunk: An Introduction

- Response to Cyberpunk

Cyberpunk is dystopian, solarpunk is hopeful

Solarpunk: An Introduction

- Response to Cyberpunk
- Beautiful solutions are better

Science fiction is a form of activism. Out of the box thinking

Solarpunk: An Introduction

- Response to Cyberpunk
- Beautiful solutions are better
- Post-capitalist

Solarpunk is counterculture; to progress is to move past competition and towards collaboration.

What do you get when you combine radical optimism, out of the box thinking, and counterculture? What do you get when you form communities centered around bold new ideas and inspired problem solving? You get hackers! Solarpunk is how we hack our way towards a better world. So let's do it!

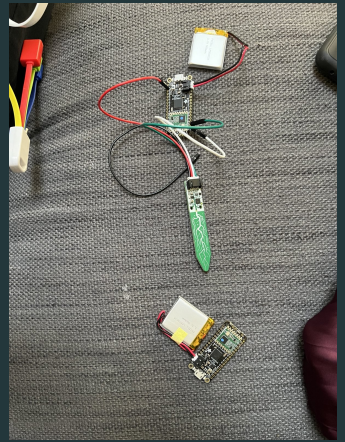
Solarpunks Make Beautiful Things

Stela Xhiku, Heliostat #1, 2023

Solarpunks Make Beautiful Things

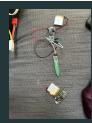
Stela Xhiku, Heliostat #1, 2023

Solarpunks Make Beautiful Things



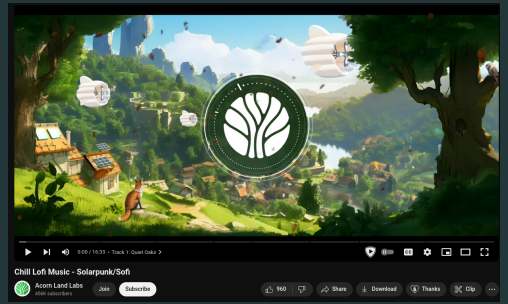
Me, Farming Web of Trust, 2022

Solarpunks Make Beautiful Things



Me, Farming Web of Trust, 2022

Solarpunks Make Beautiful Things



Sofi, Acorn Land Labs, 2023

Solarpunks Make Beautiful Things
Sofi, Acorn Land Labs, 2023

Let's look at a cool technological application of Solarpunk ideas.

Let's hack the web! (the other way)

Let's hack the web! (the other way)

The Solar Protocol

- Energy-Centered

The Solar Protocol

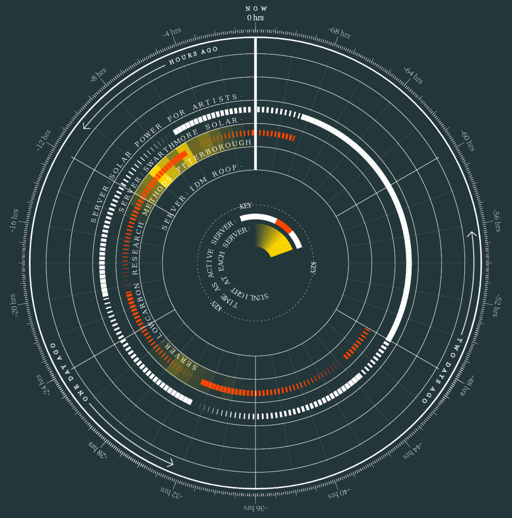
- Energy-Centered
- Planet-Scale

The Solar Protocol

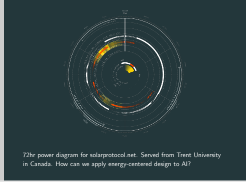
- Energy-Centered
- Planet-Scale
- Naturally Intelligent

The Solar Protocol
• Energy-Centered
• Planet-Scale
• Naturally Intelligent

permacomputing, energy-centered design



72hr power diagram for solarprotocol.net. Served from Trent University in Canada. How can we apply energy-centered design to AI?



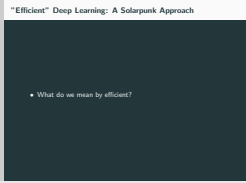
Servers in Canada, the US, China, Chile, Kenya, and more.

Approximate Multipliers

Summary of my BA thesis

"Efficient" Deep Learning: A Solarpunk Approach

- What do we mean by efficient?



Reducing memory- reduces hardware cost. Reducing compute (FLOPs)- increases speed and reduces power requirements. As Solarpunks, we're concerned with power. Let's look at some techniques from the AI mainstream.

Some Other Ideas

	Quant.	Pruning
Reduce memory	✓	
Reduce FLOPs		✓
Reduce power	✓	✓
Used in training		
Zero-shot		✓
Equal Performance	✓	✓

Some Other Ideas

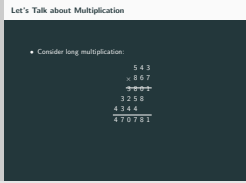
	Quant.	Pruning
Reduce memory	✓	
Reduce FLOPs		✓
Reduce power	✓	✓
Used in training		
Zero-shot		✓
Equal Performance	✓	✓

Quantization: Reduces bit width. Saves a lot of memory but for large models inability to represent small gradients is a problem. Often used after training. Pruning: Doesn't affect memory, simply sets some weights to zero to save compute. Sometimes retrained. What all of these have in common is they attempt to reduce the amount of multiplication done.

Let's Talk about Multiplication

- Consider long multiplication:

$$
\begin{array}{r}
5\,4\,3 \\
\times\, 8\,6\,7 \\
\hline
\cancel{3\,8\,0\,1} \\
3\,2\,5\,8 \\
4\,3\,4\,4 \\
\hline
4\,7\,0\,7\,8\,1
\end{array}
$$

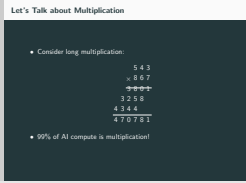


Let's Talk about Multiplication

- Consider long multiplication:

$$
\begin{array}{r}
5\,4\,3 \\
\times\, 8\,6\,7 \\
\hline
\cancel{3\,8\,0\,1} \\
3\,2\,5\,8 \\
4\,3\,4\,4 \\
\hline
4\,7\,0\,7\,8\,1
\end{array}
$$

- 99% of AI compute is multiplication!



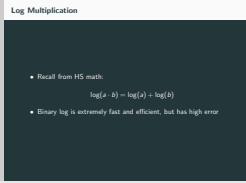
repeated addition; quadratic complexity. Quantization and pruning reduce the amount of multiplication done. What if we could use hardware that made multiplication more efficient?

Log Multiplication

- Recall from HS math:

$$\log(a \cdot b) = \log(a) + \log(b)$$

- Binary log is extremely fast and efficient, but has high error



Logarithm product rule. This idea gives rise to Approximate Multipliers. AI workloads have been shown to be tolerant to small errors in multiplication. If we replace regular multiplication with log multiplication, we can reduce the power significantly.

Some Other Ideas

	Quant.	Pruning	AMs
Reduce memory	✓		
Reduce FLOPs		✓	
Reduce power	✓	✓	✓
Used in training			✓
Zero-shot	*	✓	✓
Equal Performance	*	✓	✓

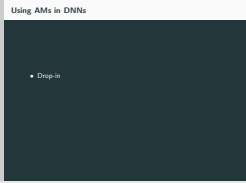
Some Other Ideas

	Quant.	Pruning	AMs
Reduce memory	✓		
Reduce FLOPs		✓	
Reduce power	✓	✓	✓
Used in training			✓
Zero-shot	*	✓	✓
Equal Performance	*	✓	✓

AMS affect the underlying framework. They reduce the computational cost of multiplication. They can be used in concert with all the other techniques here. Equal or better performance. Energy-centered design

Using AMs in DNNs

- Drop-in



No change in architecture, no retraining90+% energy savings.

Using AMs in DNNs

- Drop-in
- Minimal Performance Degradation

Stochastic regularization90+% energy savings.

Using AMs in DNNs

- Drop-in
- Minimal Performance Degradation
- Modular

Usable alongside everything else we talked about90+% energy savings.

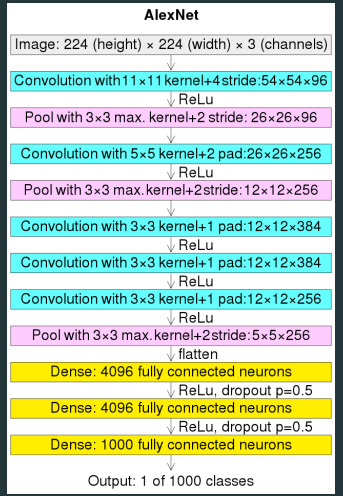
Using AMs in DNNs

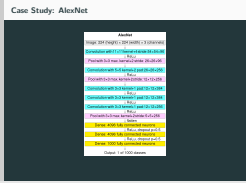
- Drop-in
- Minimal Performance Degradation
- Modular
- Massive Energy Savings

Using AMs in DNNs
• Drop-in
• Minimal Performance Degradation
• Modular
• Massive Energy Savings

90+% energy savings.

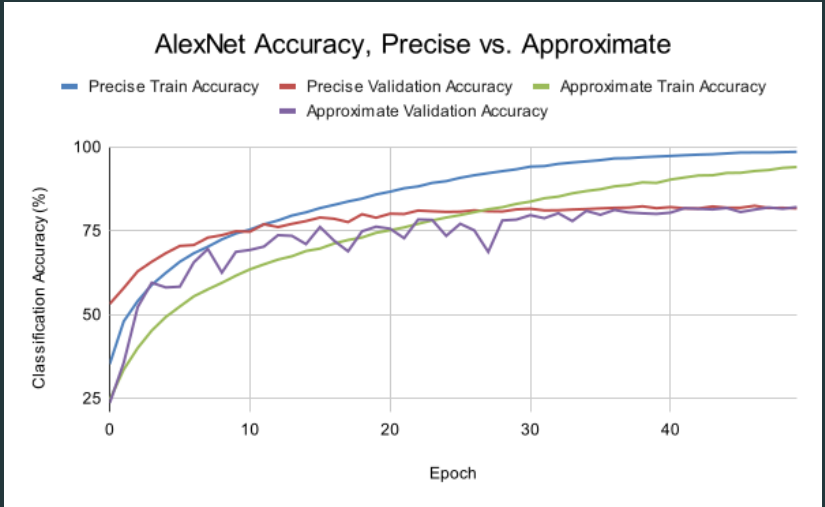
Case Study: AlexNet

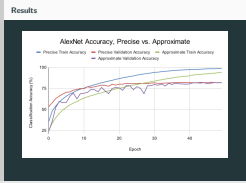




This is AlexNet, a well-known CNN which is complex enough to be a useful example but small enough to test easily. I implemented both precise and approximate versions and trained them on the CIFAR-10 dataset for 50 epochs.

Results





Precise: Final validation 81.8, training 98.5

Approx: Final validation 81.7, training 93.4

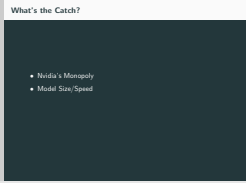
Rates of learning are similar, AMs get off to a slower start.

What's the Catch?

- Nvidia's Monopoly

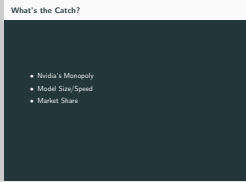
What's the Catch?

- Nvidia's Monopoly
- Model Size/Speed



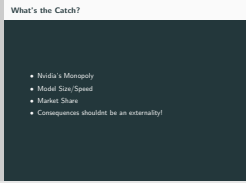
What's the Catch?

- Nvidia's Monopoly
- Model Size/Speed
- Market Share



What's the Catch?

- Nvidia's Monopoly
- Model Size/Speed
- Market Share
- Consequences shouldnt be an externality!



Conclusion

Conclusion

Why are we here? Because of the name of the conference: Hope.
Solarpunk isn't about saving the world, but it is about trying our hardest to.
If the rest of this means nothing to you, look at me now and remember this: you are the future. Everyone in the room with me, everyone watching live, everyone viewing the recording; you are the future. The freaks and the hackers and the weirdos. We all have a role to play in saving the world. The more people believe that it's possible to change the world, the more likely it is to happen. So get out there and make it happen.

Thank You

Thank You

Links and Such

@rolltime@freeradical.zone

rollti.me/hope2024

Links and Such
@rolltime@freeradical.zone
rollti.me/hope2024

Mitchell's Multiplier Example

- Let's take the binary log of 18. $18 = 00010010$
  Find the most significant one bit, here 2^4.
  Take the position (4) as the significand, and the rest as the mantissa; $b0.0010 = 0.125$. Thus:

$$\log_2 18 \approx 4.125$$

  In fact it's about 4.17, so we got pretty close!

Mitchell's Multiplier Example
• Let's take the binary log of 18. 18 = 00010010
Find the most significant one bit, here 2^4.
Take the position (4) as the significand, and the rest as the mantissa; b0.0010 = 0.125. Thus:
$\log_2 18 \approx 4.125$
In fact it's about 4.17, so we got pretty close!

Computers- shift to left, counter, etc.

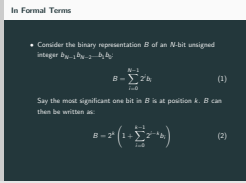
In Formal Terms

- Consider the binary representation B of an N-bit unsigned integer $b_{N-1}b_{N-2}...b_1b_0$:

$$B = \sum_{i=0}^{N-1} 2^i b_i \tag{1}$$

Say the most significant one bit in B is at position k. B can then be written as:

$$B = 2^k \left(1 + \sum_{i=0}^{k-1} 2^{i-k} b_i\right) \tag{2}$$



In Formal Terms, Cont.

- Let x be:

$$x = \sum_{i=0}^{k-1} 2^{i-k} b_i \tag{3}$$

where $0 \leq x < 1$. We can write:

$$B = 2^k(1 + x) \tag{4}$$

By log rules, the accurate binary log of B is:

$$\log_2 B = \log_2 \left(2^k (1 + x)\right) \tag{5}$$

$$= k + \log_2(1 + x) \tag{6}$$

In Formal Terms, Cont.
• Let x be:
$$x = \sum_{i=0}^{k-1} 2^{i-k} b_i \quad (3)$$
where $0 \leq x < 1$. We can write:
$$B = 2^k(1 + x) \quad (4)$$
By log rules, the accurate binary log of B is:
$$\log_2 B = \log_2 \left(2^k (1 + x)\right) \quad (5)$$
$$= k + \log_2(1 + x) \quad (6)$$

In Formal Terms, Cont.

- To complete the calculation, we approximate.
  The mantissa $log_2(1 + x)$ is approx. x, since $0 \leq x < 1$.
  Thus:

$$\log_2 B \approx k + x \tag{7}$$

  To calculate the approximate product, we sum two approximate log values and calculate the approximate antilog using a similar method.

In Formal Terms, Cont.
• To complete the calculation, we approximate.
The mantissa $log_2(1 + x)$ is approx. x, since $0 \leq x < 1$.
Thus:
$\log_2 B \approx k + x$ (7)
To calculate the approximate product, we sum two approximate log values and calculate the approximate antilog using a similar method.

Quantifying Area, Power, and Error

- For int16, MA is 70% smaller and uses 78% less power than accurate multipliers
- However, error is both large and unpredictable

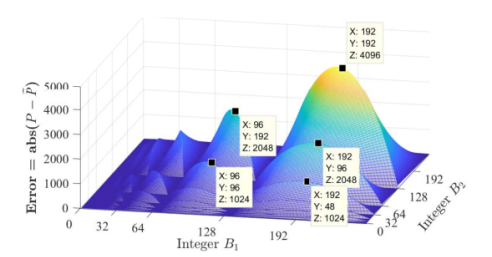
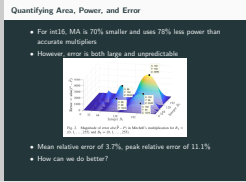


Fig. 2. Magnitude of error $abs(\tilde{P} - P)$ in Mitchell's multiplication for $B_1 = \{0, 1, \ldots, 255\}$ and $B_2 = \{0, 1, \ldots, 255\}$.

- Mean relative error of 3.7%, peak relative error of 11.1%
- How can we do better?

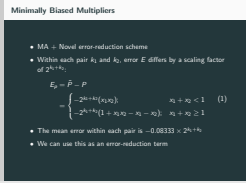


Minimally Biased Multipliers

- MA + Novel error-reduction scheme
- Within each pair k_1 and k_2, error E differs by a scaling factor of $2^{k_1+k_2}$:

$$E_p = \tilde{P} - P$$
$$= \begin{cases} -2^{k_1+k_2}(x_1 x_2); & x_1 + x_2 < 1 \\ -2^{k_1+k_2}(1 + x_1 x_2 - x_1 - x_2); & x_1 + x_2 \geq 1 \end{cases} \quad (1)$$

- The mean error within each pair is $-0.08333 \times 2^{k_1+k_2}$
- We can use this as an error-reduction term



Quantifying MBM vs. MA

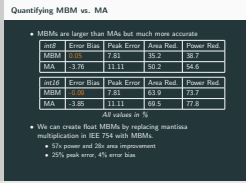
- MBMs are larger than MAs but much more accurate

int8	Error Bias	Peak Error	Area Red.	Power Red.
MBM	0.05	7.81	35.2	38.7
MA	-3.76	11.11	50.2	54.6

int16	Error Bias	Peak Error	Area Red.	Power Red.
MBM	-0.09	7.81	63.9	73.7
MA	-3.85	11.11	69.5	77.8

All values in %

- We can create float MBMs by replacing mantissa multiplication in IEE 754 with MBMs.
  - 57x power and 28x area improvement
  - 25% peak error, 4% error bias



Mention error bias, all errors essentially average to near zero