JISEA Joint Institute for Strategic Energy Analysis

Green AI: Insights Into Deep Learning's Looming Energy Efficiency Crisis

Charles Tripp, Erik Bensen, Lucas Hayne, Jordan Perr-Sauer, Monte Lunacek

EPRI AI and Electric Power Summit

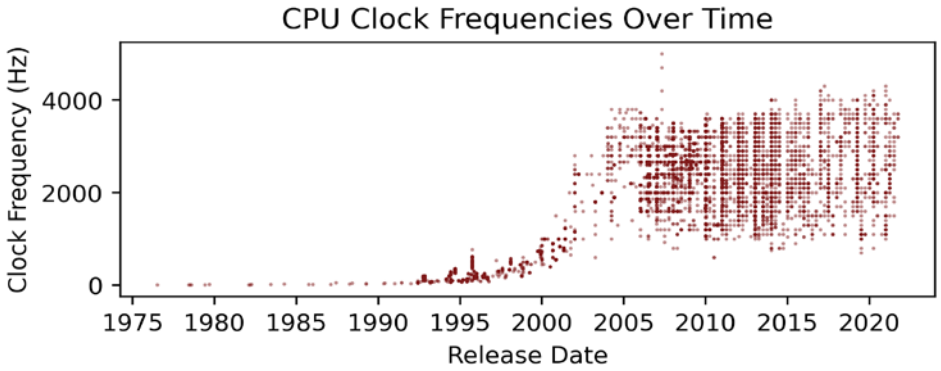
10/06/2022

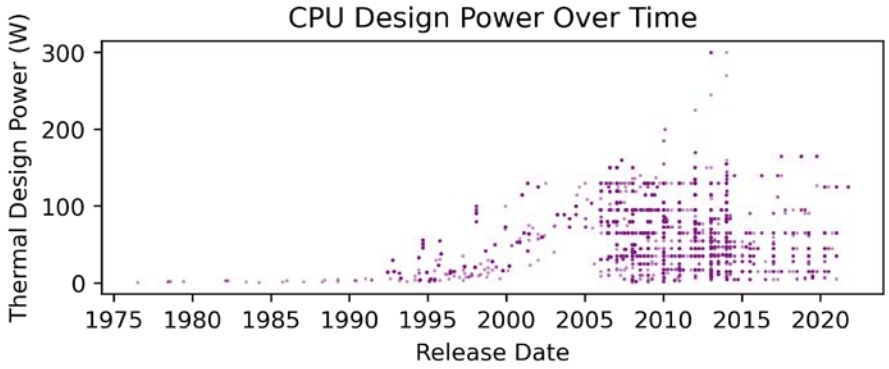
NREL Transforming ENERGY
COLORADOSCHOOLOFMINES EARTH • ENERGY • ENVIRONMENT
COLORADO STATE UNIVERSITY
Massachusetts Institute of Technology
Stanford University
University of Colorado Boulder

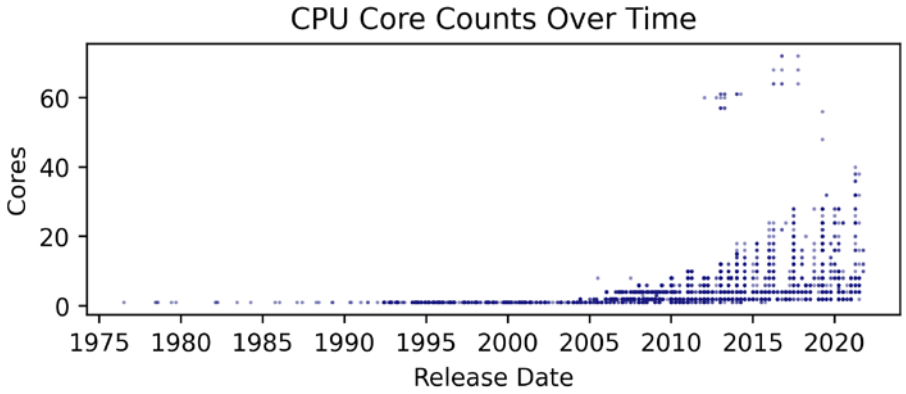
Agenda

- Background
  - A Brief History of Energy & Computing
  - The Growing Energy Costs of Computing
- An Empirical Study of Deep Learning's Energetic Dynamics
- How to Reduce Training Energy By 98%+ Today
- A Path Toward Increased Deep Learning Energy Efficiency By Orders of Magnitude

A Brief History of Computing



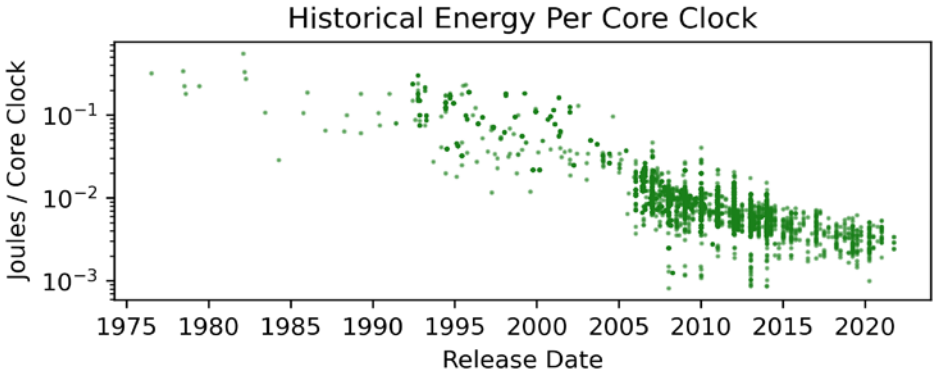


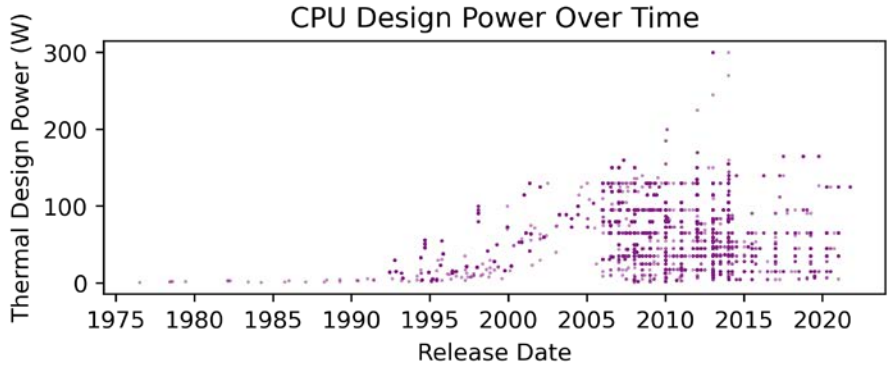


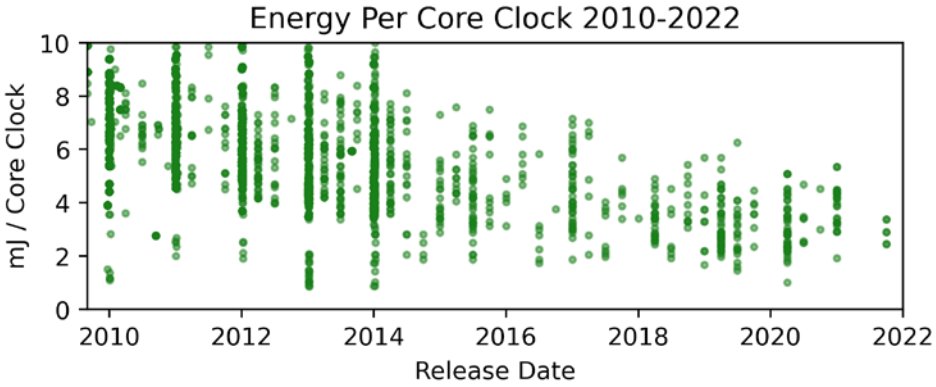
- Exponential Dennard scaling of clock frequencies ended ~2005
- Core counts have been linearly increasing since then
- CPU Design Power increased exponentially during Dennard scaling, and is now leveling-off

Figure data from: ark.intel.com, github.com/toUpperCase78/intel-processors, and Danowitz, A., Kelley, K., Mao, J., Stevenson, J. P., & Horowitz, M. (2012). CPU DB: recording microprocessor history. *Communications of the ACM*, *55*(4), 55-63.

A Brief History of The Energy Efficiency of Computing







- Energy per core clock cycle is a rough approximation of the energy cost per operation
- Energy cost per clock dropped exponentially during and beyond Dennard scaling and could be approaching a non-zero asymptote.

Figure data from: ark.intel.com, github.com/toUpperCase78/intel-processors, and
[1] Danowitz, A., Kelley, K., Mao, J., Stevenson, J. P., & Horowitz, M. (2012). CPU DB: recording microprocessor history. *Communications of the ACM*, *55*(4), 55-63.

The Rising Energetic Cost of Computing

- Computing systems are no longer exponentially decreasing the energy cost of computing towards zero.
- Total computing energy usage has been growing exponentially for more than a decade.
- If this trend continues, computing is forecasted to consume >8 PWh, or **~20% of generated electricity in 2030** [2].

- To mitigate the looming computing energy crisis:

  It is time to invest in efficient algorithms, software, and operating strategies.

[2] Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. *Nature*, *561*(7722), 163-167.

Deep Learning's Energy Footprint

- Computing is an increasing consumer of energy resources worldwide.
- AI is a popular consumer of computing resources.
- Within AI, Deep Learning is both highly prevalent and computationally expensive.
- Training and applying popular neural network models are consuming vast quantities of energy and incurring large carbon footprints.

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer_{big}	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT_{base}	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Table from [3] Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645-3650).

Quantifying The Energy Cost of Deep Learning

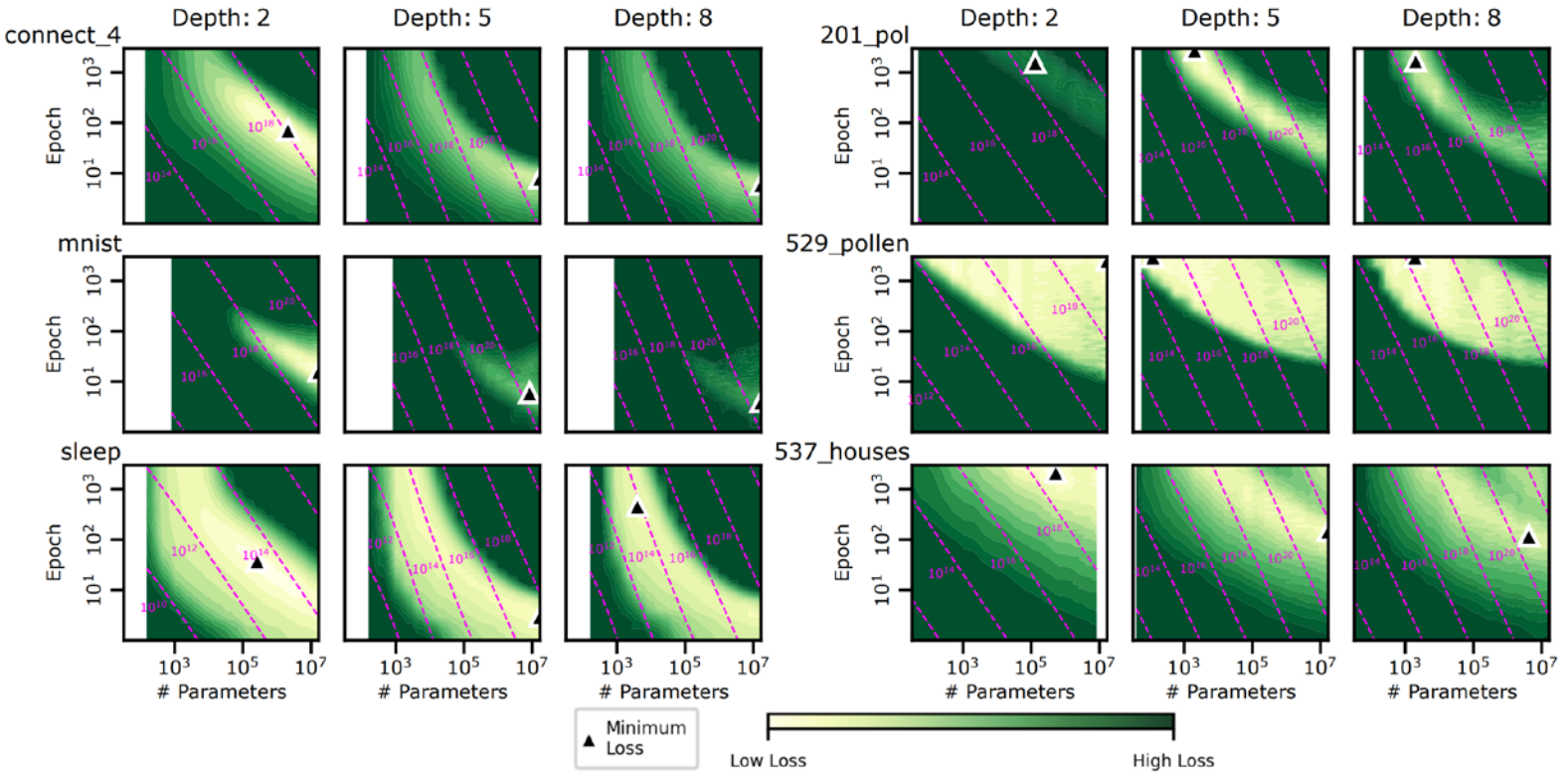
- We are empirically measuring the energy consumption of deep learning algorithms.
  - To connect AI researchers to the concrete energetic costs of their work
  - To understand the energy-performance trade-offs of network configuration design decisions
  - To propose and evaluate more energy-efficient approaches to deep learning

- We have produced an empirical deep learning dataset, "BUTTER", [4] [5] recording the impact of network topology and hyperparameter choice on network training and test performance.
  - 13 datasets, 8 shapes, 14 depths, 23 sizes, 4 learning rates, 6 batch sizes, 18 regularization levels, and 5 optimizers
  - 483k distinct experiments, 11M training runs, 40B training epochs, 20 statistics per epoch

- We are now augmenting this dataset with both energy usage estimators and actual energy usage measurements.
- We hope that analysis of this dataset will lead to greater awareness of deep learning's energy cost and aid in developing more energy-efficient practices.

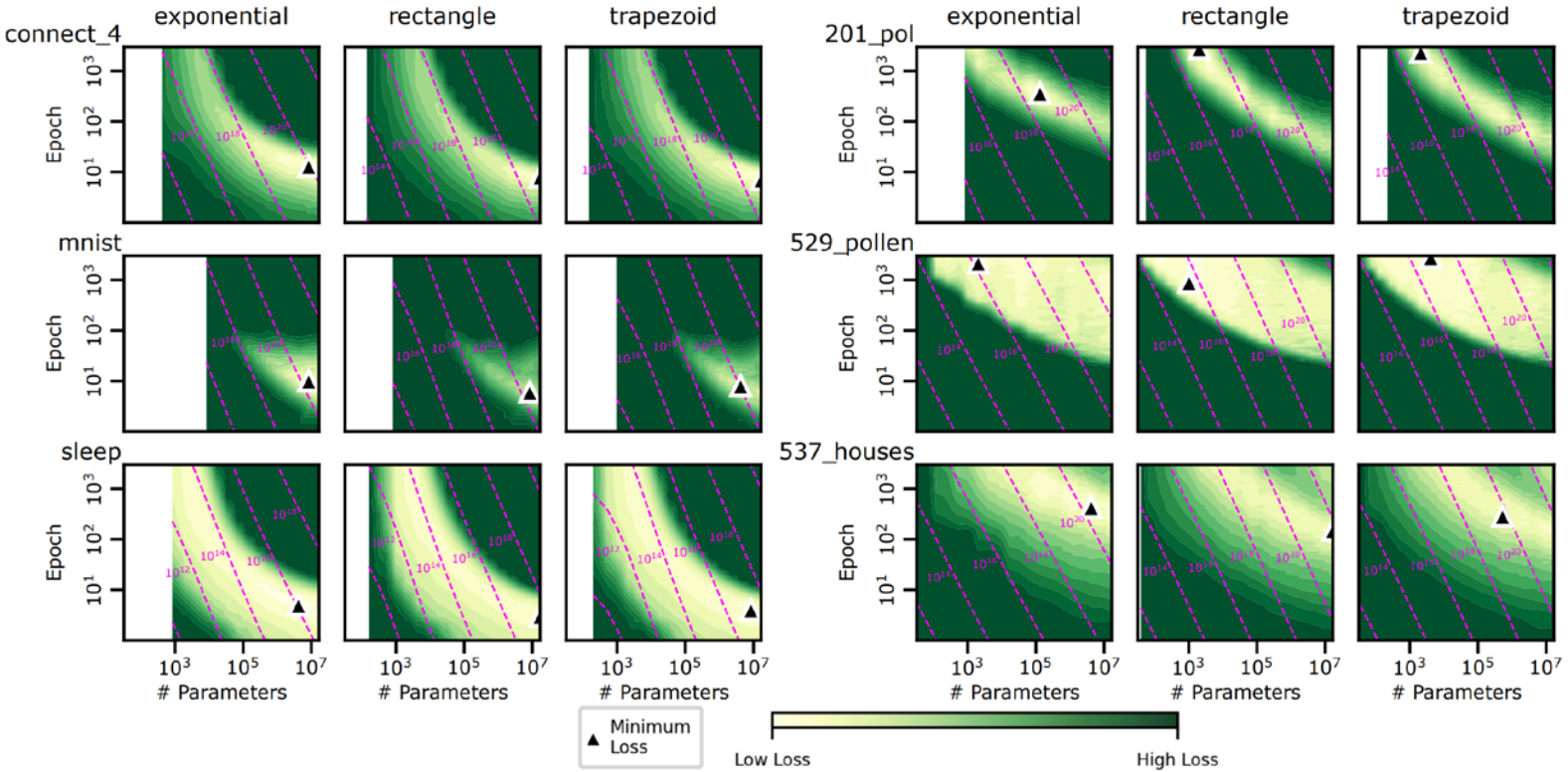
[4] Tripp, Charles, Perr-Sauer, Jordan, Hayne, Lucas, & Lunacek, Monte. *BUTTER - Empirical Deep Learning Dataset*. United States. https://dx.doi.org/10.25984/1872441
[5] Tripp, C. E., Perr-Sauer, J., Hayne, L., & Lunacek, M. (2022). An Empirical Deep Dive into Deep Learning's Driving Dynamics. *arXiv preprint arXiv:2207.12547*.

Empirical Deep Learning Efficiency Study



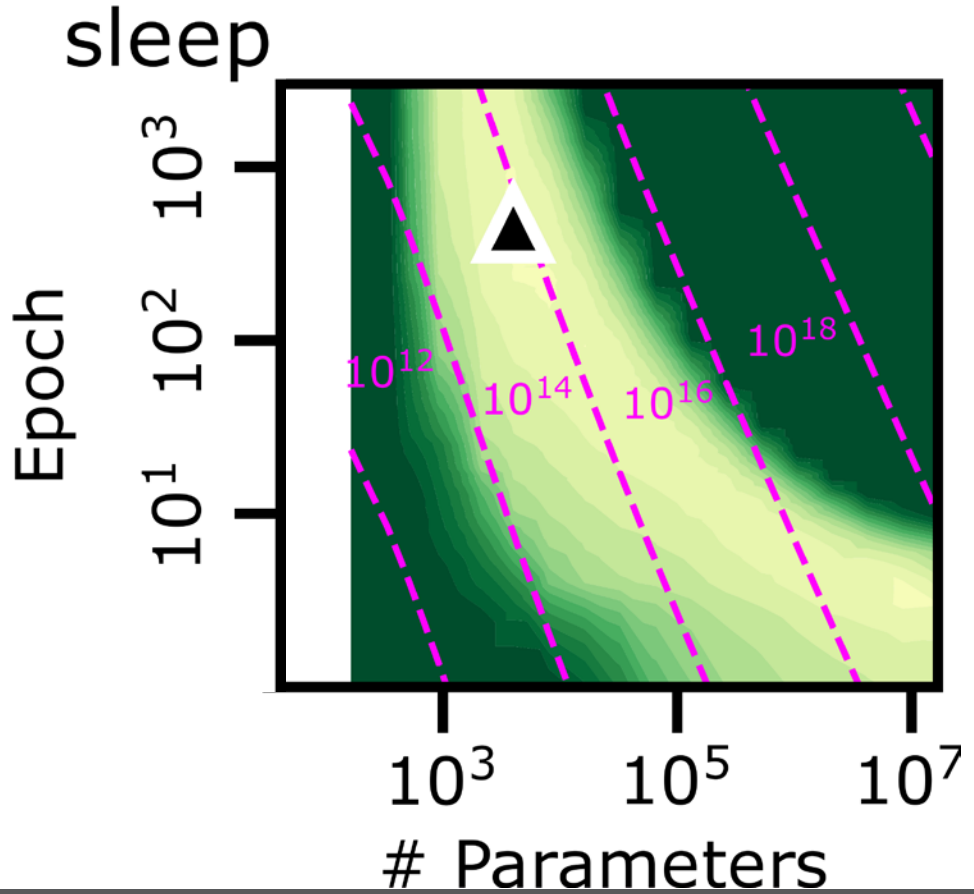
Larger models can learn in fewer epochs, and reach higher performance levels in fewer iterations, with exponentially diminishing returns. Each task has an optimum depth.

Empirical Deep Learning Efficiency Study



Shape has only a modest impact on performance, as long as the topology is not sparse and does not contain any bottlenecks.

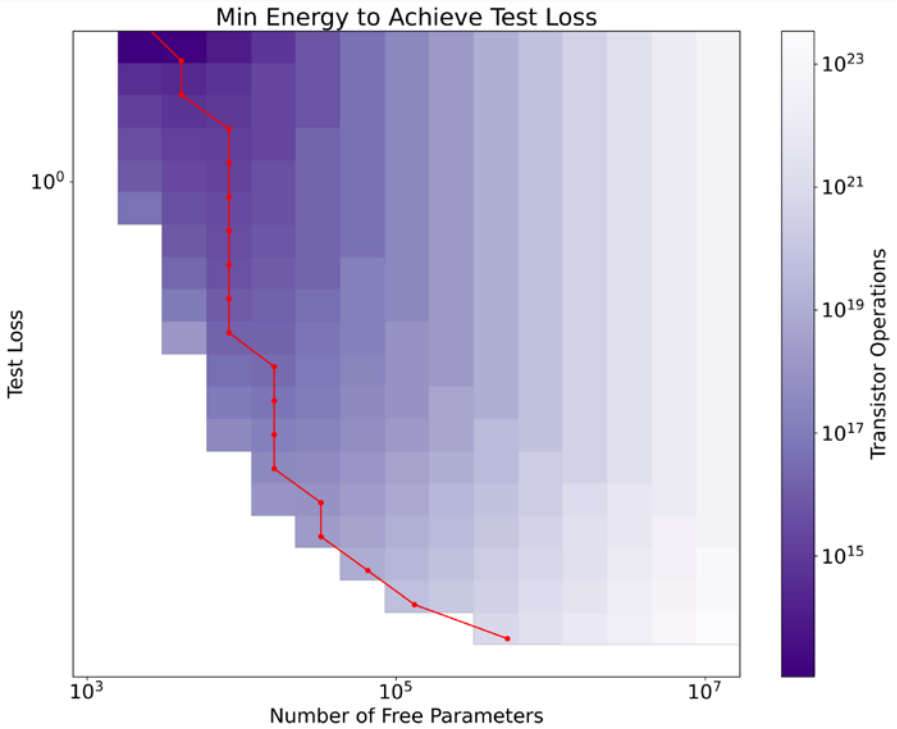
One Simple Trick For Training For <2% the Energy



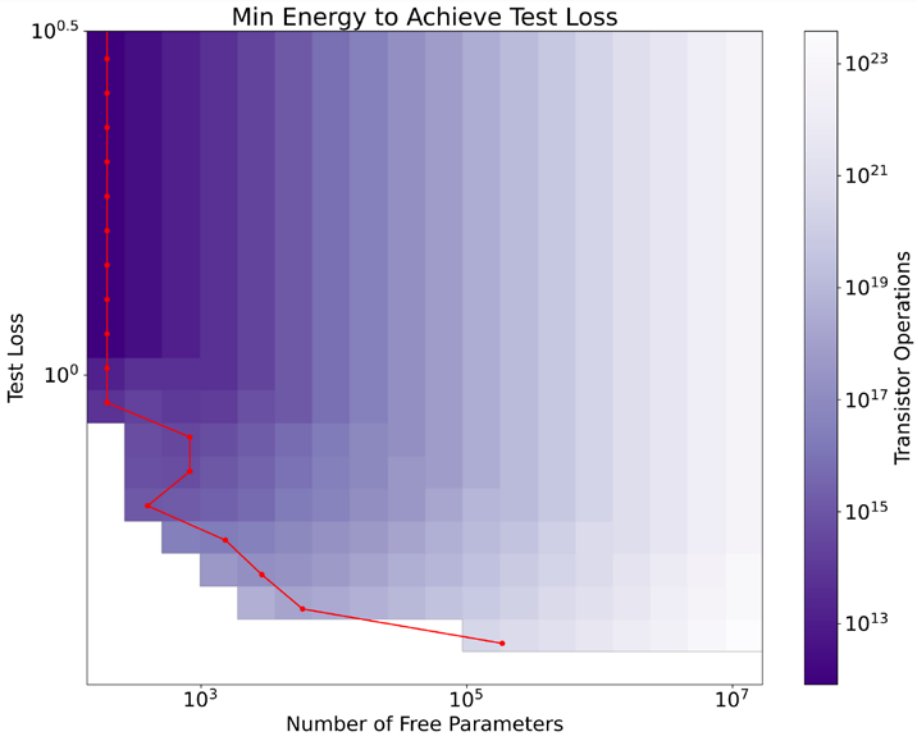
- Because larger networks can achieve better performance, a common practice is to simply train a heavily overparameterized network.
- Doing so can easily consume several orders of magnitude more energy than required to achieve a target loss level.
- A simple exponential search for a small size that can reach the performance target can reduce energy consumption to within ~2x that of the best possible network.

Empirical Deep Learning Efficiency Study

mnist



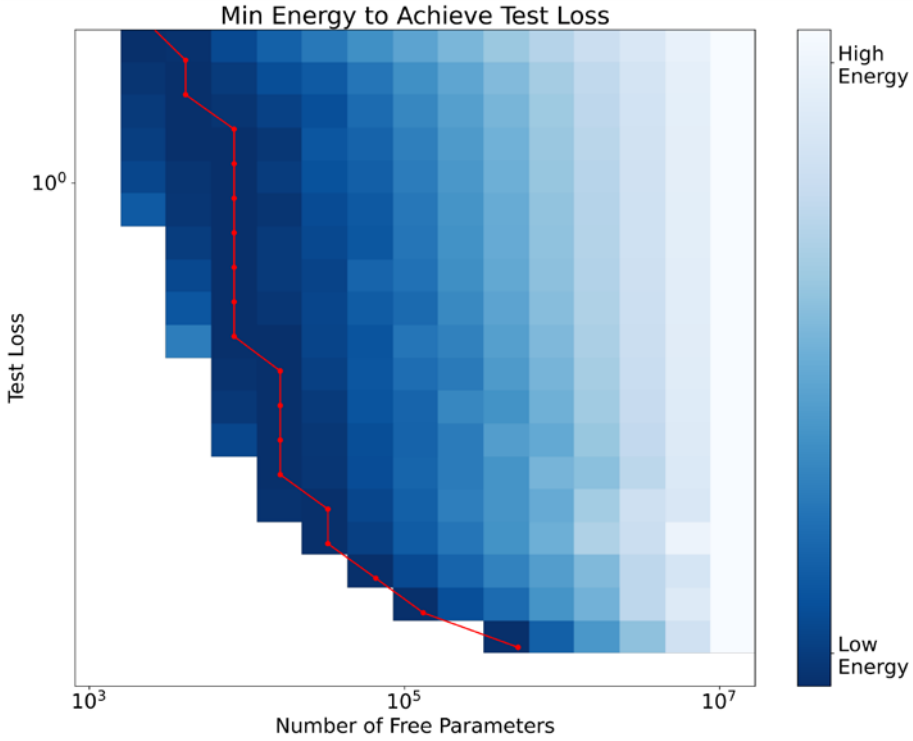
connect_4



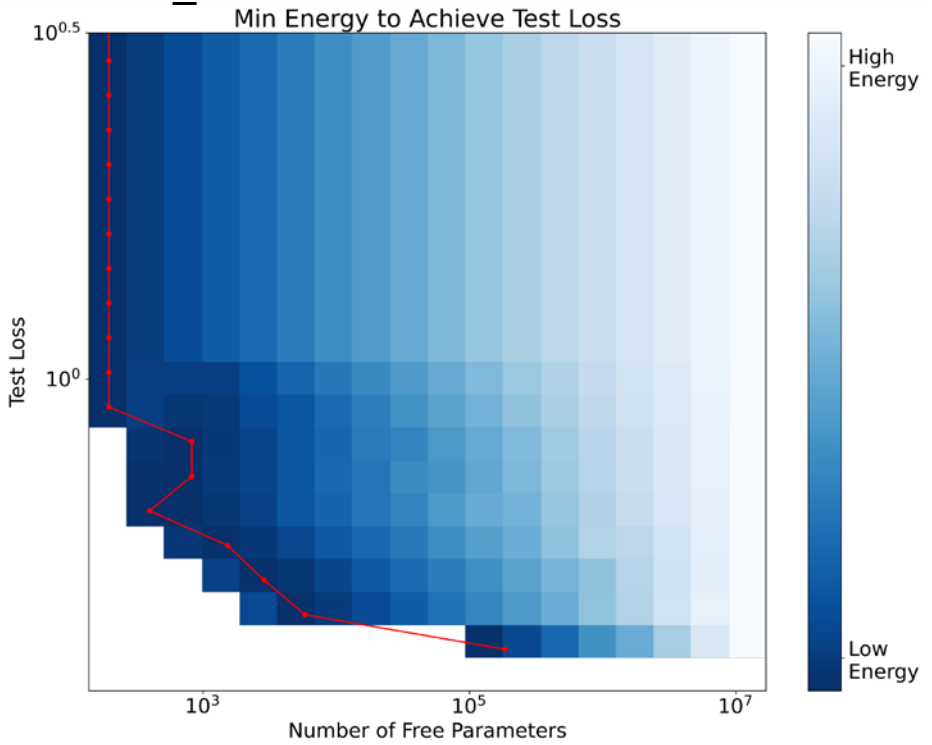
The most efficient network for any performance level is within 4x the size of the smallest network that can achieve it.

Empirical Deep Learning Efficiency Study

mnist

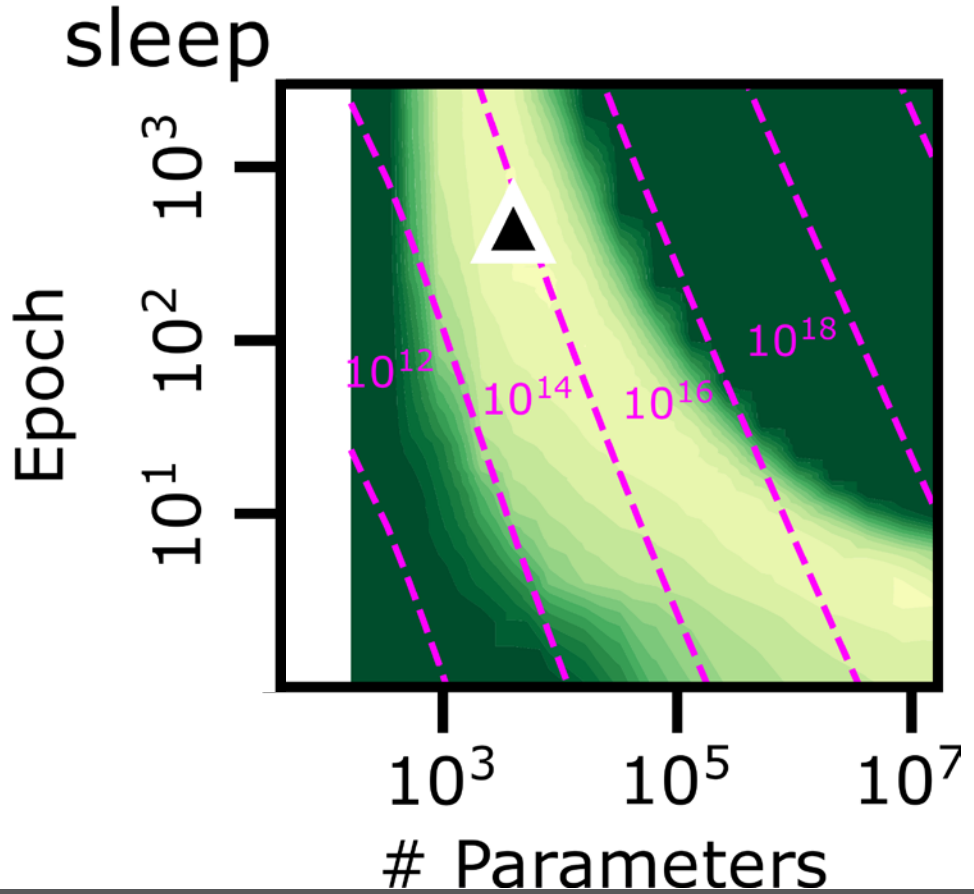


connect_4



An exponential search in network size will only cost ~2x of the minimum energy required to achieve any target loss level.

A Path Forward: Energy Efficient Deep Learning



- Simple exponential size search can reduce energy costs by 2-4 orders of magnitude.
- Interleaving training with growth may enable even greater efficiencies matching or exceeding that of using the energy-optimal size, reducing energy consumption by 50%.
- Building on the "Lottery Ticket Hypothesis", incremental pruning may reduce energy costs by 1-2 orders of magnitude.
- Regularization mimicking the effect of heavily overparameterized networks may reduce energy costs further.
- Additional refinements are possible!

Algorithmic Efficiency Can Mitigate The Coming Crisis

- For AI applications to continue growing at the current pace, energy costs must be considered and addressed.
- Targeted refinement of AI algorithms can yield meaningful reductions in the computational energy burden.
- However, we must invest in further research to achieve these reductions.

Thank you!

NREL/PR-2C00-84088

www.jisea.org

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the Joint Institute for Strategic Energy Analysis, and the National Renewable Energy Laboratory. The views expressed herein do not necessarily represent the views of the DOE, the U.S. Government, or sponsors.

JISEA Joint Institute for Strategic Energy Analysis

NREL Transforming ENERGY | COLORADO SCHOOL OF MINES EARTH • ENERGY • ENVIRONMENT | COLORADO STATE UNIVERSITY | Massachusetts Institute of Technology | Stanford University | University of Colorado Boulder