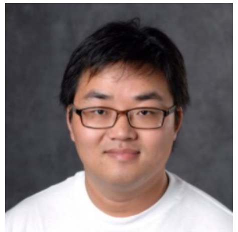
Introduction



Jiliang Tang

Privacy

Xiaorui Liu

Safety & Robustness

Yaxin Li

Explainability

Wenqi Fan

Non-discrimination & Fairness

Environmental Well- being

Haochen Liu

Accountability & Auditability

Dimension Interactions

Future Directions

Yiqi Wang

Energy Consumption of Large NLP models

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Strubell et al. "Energy and Policy Considerations for Deep Learning in NLP." 2019.

Environmental Well-being

- ❑ Sustainable
- ❑ Environmentally friendly

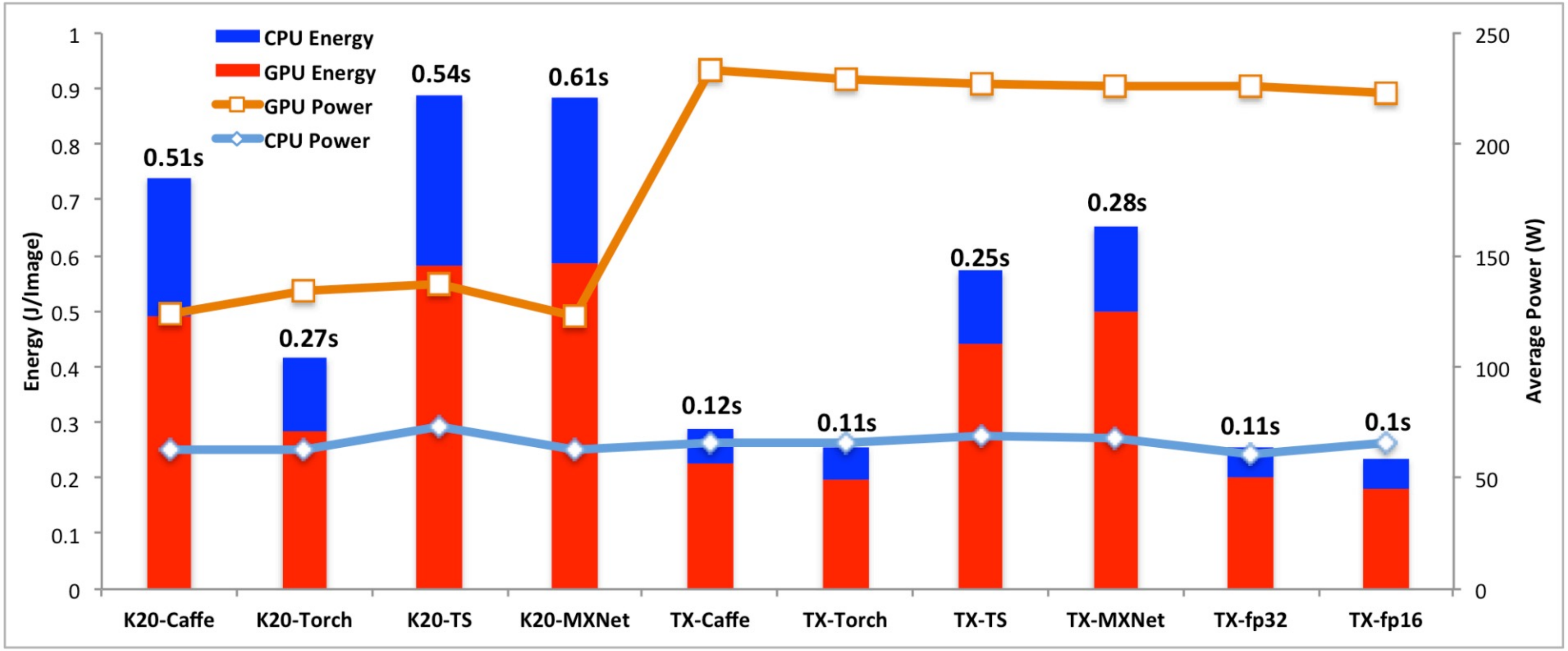
Energy Consumption Estimation → Energy Consumption Reduction

→ Surveys and Tools → Future Directions

Energy Consumption Estimation → Energy Consumption Reduction

→ Surveys and Tools → Future Directions

Energy consumption estimation in computer vision

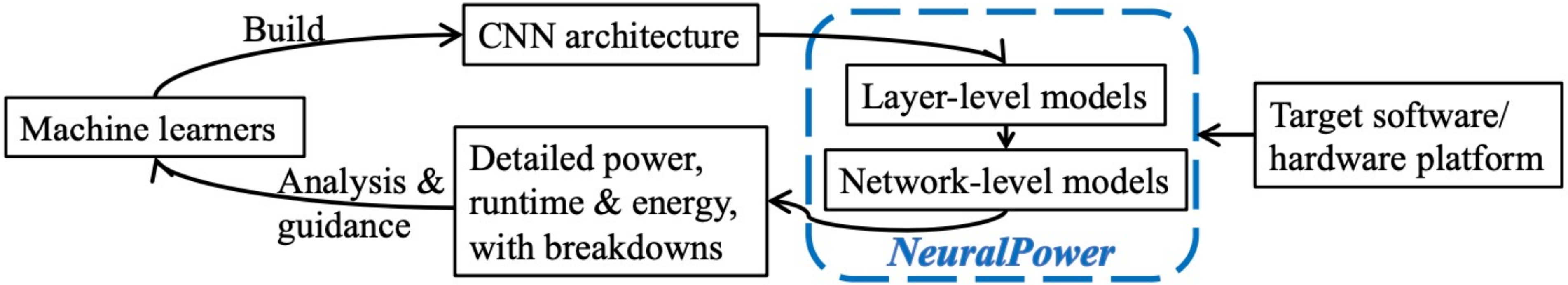


Energy consumption comparison among different CNN frameworks

Li, Da, et al. "Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs." 2016.

NeuralPower

NeuralPower: a predictive framework for power, runtime, and energy of CNNs during the testing phase (Cai et al., 2017)



Cai, Ermao, et al. "Neuralpower: Predict and deploy energy-efficient convolutional neural networks." 2017.

Energy consumption estimation in NLP

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

Carbon emissions of training popular NLP models on different types of hardware

Strubell et al. “Energy and Policy Considerations for Deep Learning in NLP.” 2019.

Energy Consumption Estimation → Energy Consumption Reduction

→ Surveys and Tools → Future Directions

Reducing energy consumption

- Model Compression
  - The size of a deep model is reduced via model compression techniques.
- Adaptive Design
  - The architecture of a model is adaptively designed to optimize its energy efficiency.
- Hardware
  - Energy-efficient computing devices or platforms are designed for specific AI applications.

Model compression

Category Name	Description
Parameter pruning and quantization	Reducing redundant parameters which are not sensitive to the performance
Low-rank factorization	Using matrix/tensor decomposition to estimate the informative parameters
Transferred/compact convolutional filters	Designing special structural convolutional filters to save parameters
Knowledge distillation	Training a compact neural network with distilled knowledge of a large model

Cheng et al. “A Survey of Model Compression and Acceleration for Deep Neural Networks.” 2019.

Adaptive Design

- Pruning Approach (Yang et al., 2017)
  - The CNN layers which consume much energy are pruned.
- Hyperparameter Optimization (Stamoulis et al., 2018)
  - The design of a CNN architecture is formulated as a hyperparameter optimization problem under energy consumption restrictions.

Yang, Tien-Ju, Yu-Hsin Chen, and Vivienne Sze. "Designing energy-efficient convolutional neural networks using energy-aware pruning." 2017.
Stamoulis, Dimitrios, et al. "Designing adaptive neural networks for energy-constrained image classification." 2018.

Hardware

- Neural processing unit (NPU) (Esmaeilzadeh et al., 2012)
  - NPU executes some fixed computations in neural networks such as multiplication, accumulation, and sigmoid, on chips.
- RENO (Liu et al., 2015)
  - A more advanced on-chip architecture is proposed for neural network acceleration.
- ReGAN (Chen et al., 2018)
  - It is specially designed for accelerating generative adversarial networks.

Esmaeilzadeh, Hadi, et al. "Neural acceleration for general-purpose approximate programs." 2012.
Liu, Xiaoxiao, et al. "RENO: A high-efficient reconfigurable neuromorphic computing accelerator design." 2015.
Chen, Fan, Linghao Song, and Yiran Chen. "Regan: A pipelined reram-based accelerator for generative adversarial networks." 2018.

Energy Consumption Estimation → Energy Consumption Reduction

→ **Surveys and Tools** → Future Directions

Surveys

- García-Martín et al. “Estimation of energy consumption in machine learning.” 2019.
- Cheng et al. “A survey of model compression and acceleration for deep neural networks.” 2017.
- Wang et al. “Benchmarking the performance and energy efficiency of ai accelerators for ai training.” 2020.
- Mittal et al. “A survey of methods for analyzing and improving GPU energy efficiency.” 2014.
- Chen et al. “A survey of accelerator architectures for deep neural networks.” 2020.

Tools

- *SyNERGY* (Rodrigues et al., 2018).
- *Machine Learning Emissions Calculator* (Lacoste et al., 2019).
- *Accelergy* (Wu et al., 2019).
- *Timeloop* (Parashar et al., 2019).

García-Martín, Eva, et al. "Estimation of energy consumption in machine learning." 2019.
Lacoste, Alexandre, et al. "Quantifying the carbon emissions of machine learning." 2019.
Wu et al. "Accelergy: An architecture-level energy estimation methodology for accelerator designs." 2019.
Parashar et al. "Timeloop: A systematic approach to dnn accelerator evaluation." 2019.

Energy Consumption Estimation → Energy Consumption Reduction

→ Surveys and Tools → Future Directions

Future Directions

- Algorithm level
  - AutoML has the potential to design energy-saving models.

- Hardware level
  - Designing efficient devices to facilitate model training needs more attention.