Carnegie Mellon University
Language Technologies Institute

Lecture 1: Course Overview

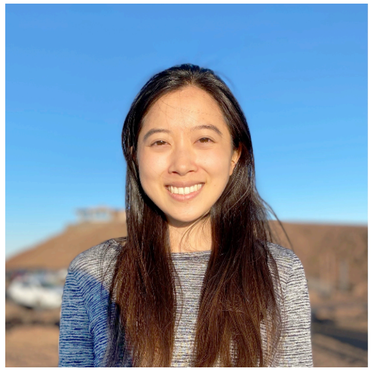
Yonatan Bisk & Emma Strubell

Welcome to On-Device Machine Learning!

Yonatan Bisk
he/him
Assistant Professor
Language Technologies Institute

Emma Strubell
she/her
Assistant Professor
Language Technologies Institute

Jared Fernandez
he/him



Clara Na
she/her

Models are growing…

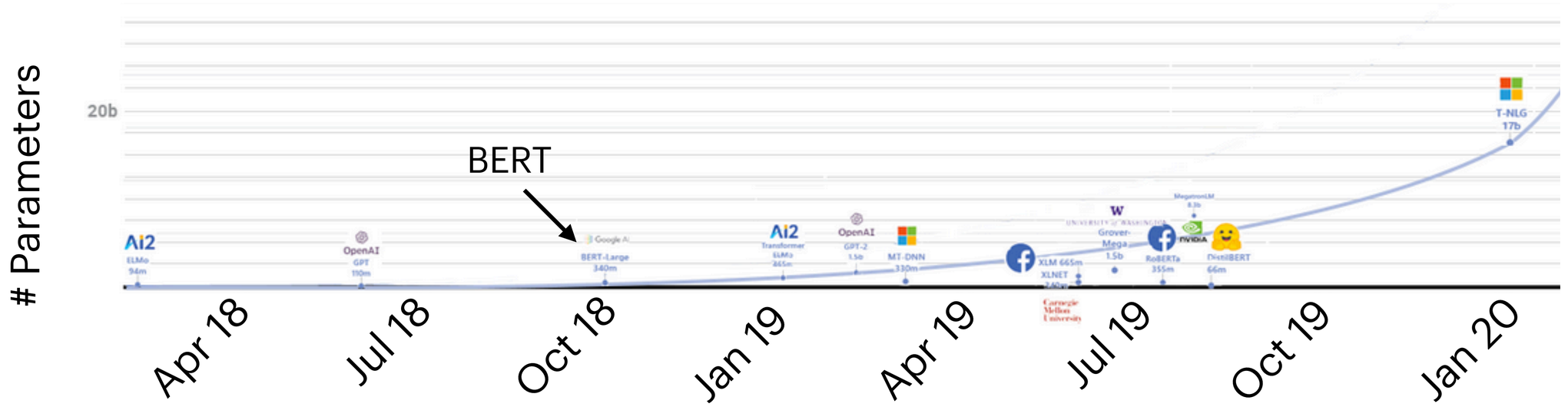
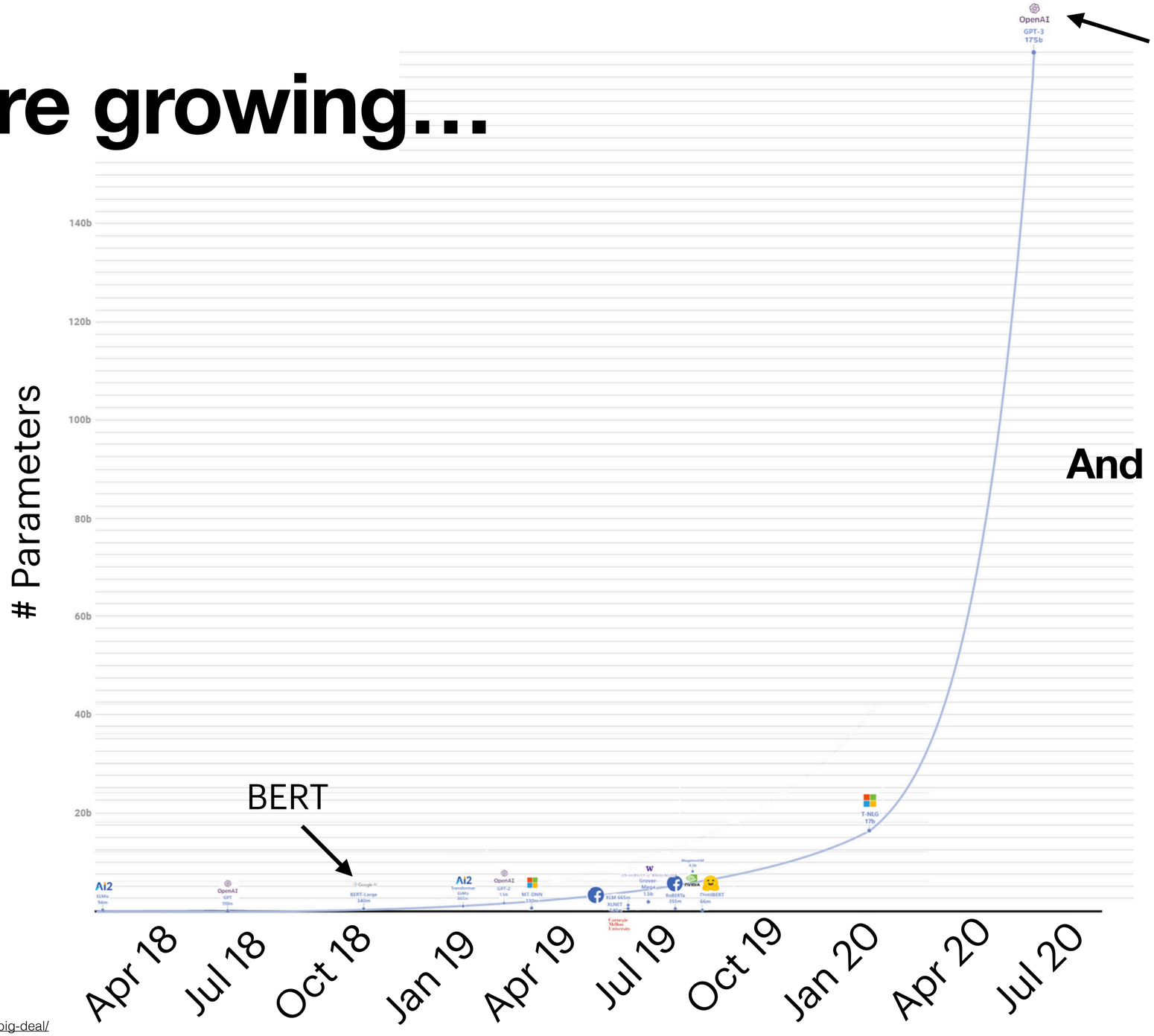


Figure: https://raohacker.com/why-the-new-ai-nlp-language-model-gpt-3-is-a-big-deal/, https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Models are growing…



And GPT-4 is 10x?!

Figure: https://raohacker.com/why-the-new-ai-nlp-language-model-gpt-3-is-a-big-deal/

ML models are growing…

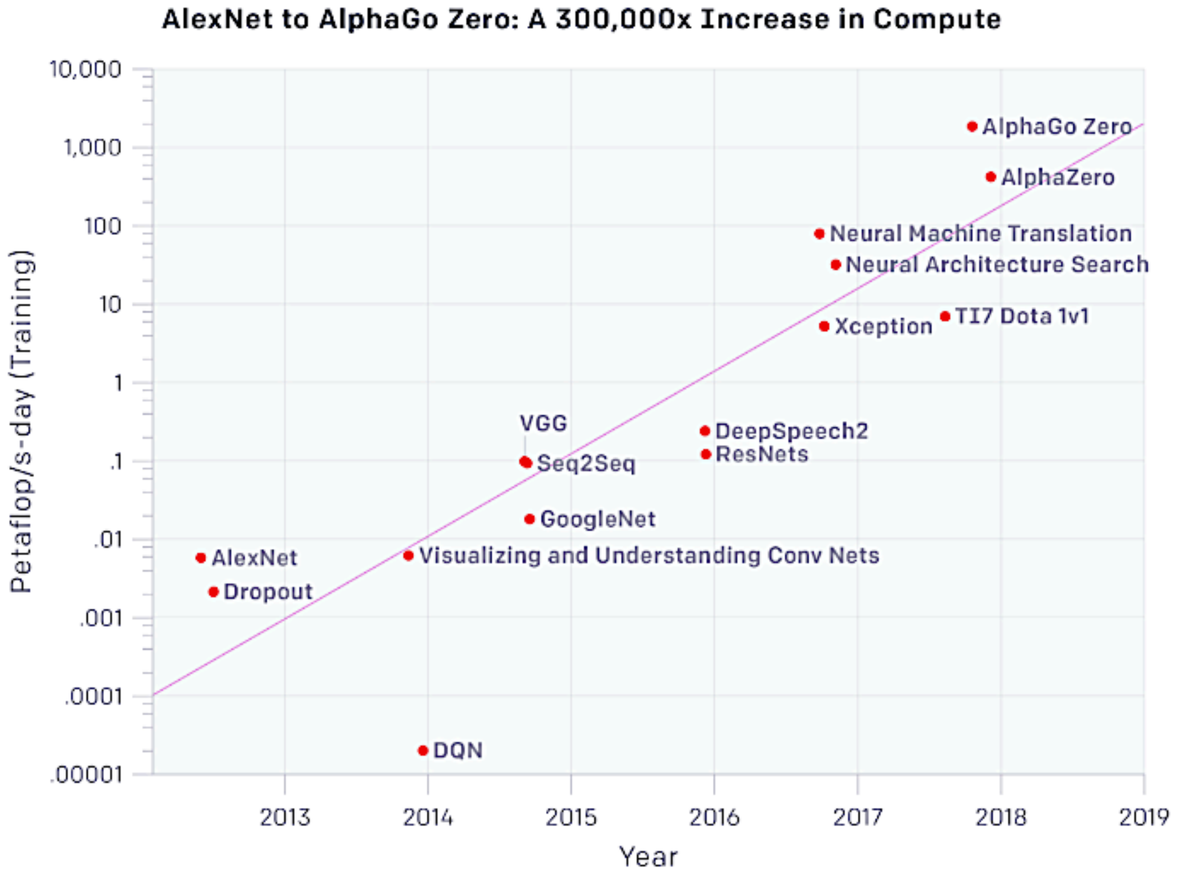


Figure: Amodei and Hernandez. AI and Compute. OpenAI Blog, 2018. https://openai.com/blog/ai-and-compute

Scaling Vision Transformers

Xiaohua Zhai*, Alexander Kolesnikov*, Neil Houlsby, Lucas Beyer*

Google Research, Brain Team, Zürich

{xzhai,akolesnikov,neilhoulsby,lbeyer}@google.com

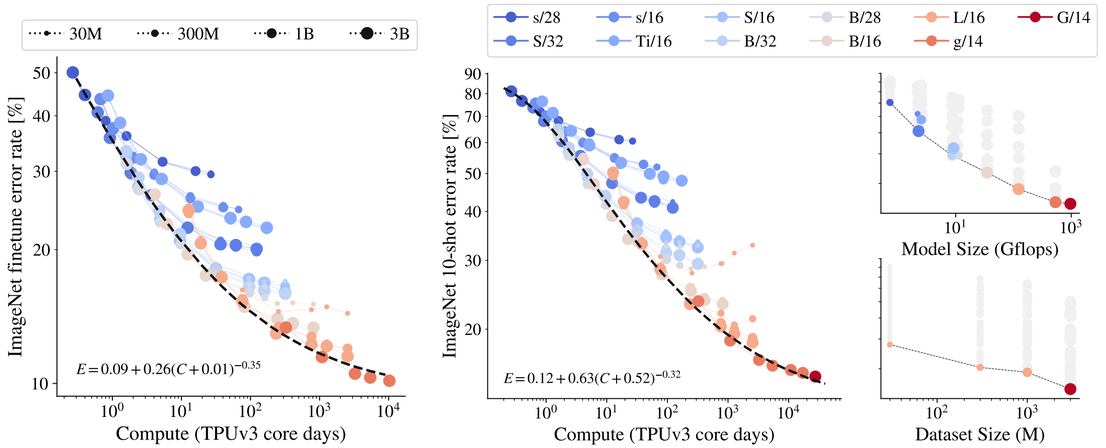
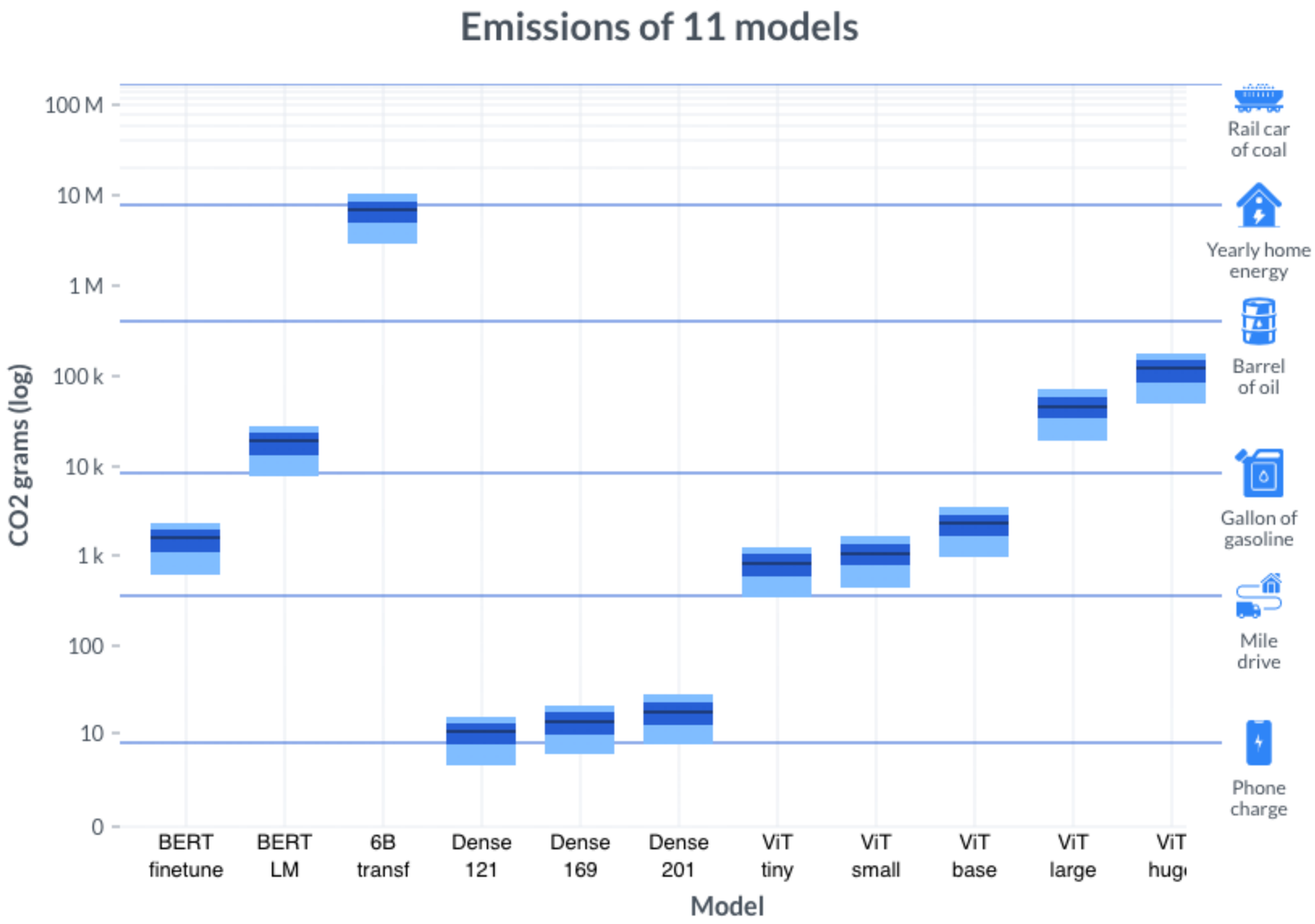


Figure 1: **Left/Center**: Representation quality, measured as ImageNet finetune and linear 10-shot error rate, as a function of total training compute. A saturating power-law approximates the Pareto frontier fairly accurately. Note that smaller models (blue shading), or models trained on fewer images (smaller markers), saturate and fall off the frontier when trained for longer. **Top right**: Representation quality when bottlenecked by model size. For each model size, a large dataset and amount of compute is used, so model capacity is the main bottleneck. Faintly-shaded markers depict sub-optimal runs of each model. **Bottom Right**: Representation quality by datasets size. For each dataset size, the model with an optimal size and amount of compute is highlighted, so dataset size is the main bottleneck.

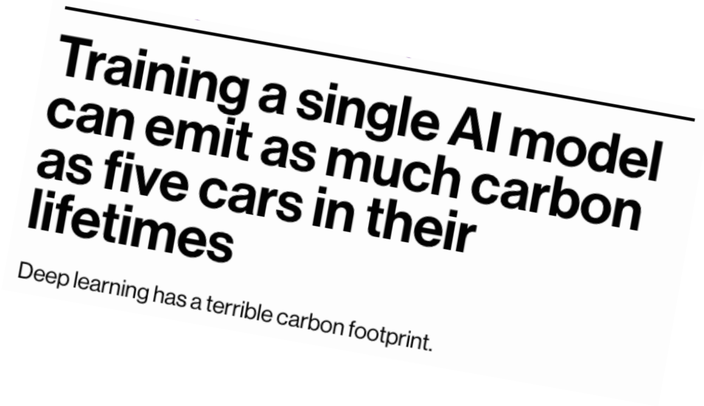
NLP models are growing…
so are their energy demands



	CO_2e (pounds)
enger, NYC↔SFO	1984
ge, 1 year	11,023
erage, 1 year	36,156
time w/ fuel	126,000

del (GPU)	
t-of-speech,	20
eter tuning &	39,234
	96
tecture search	313,078

Figure: J. Dodge, T. Prewitt, R. Tachet des Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, W. Buchanan. Measuring the Carbon Intensity of AI in Cloud Instances. ACM FAccT, 2022.



Deep Learning's Carbon Emissions Problem

AI Can Do Great Things—if It Doesn't Burn the Planet

The computing power required for AI landmarks, such as recognizing images and defeating humans at Go, increased 300,000-fold from 2012 to 2018.

Is large ML training destroying the planet?

*nobody is really doing the accounting, so we don't know, especially when it comes to inference.

- *Probably not* due to direct GHG emissions.*
- Datacenters responsible for < 1% global electricity use, small subset* used for AI/ML.

Emissions due to hardware supply chain

Chasing Carbon: The Elusive Environmental Footprint of Computing

Udit Gupta[1,2], Young Geun Kim[3], Sylvia Lee[2], Jordan Tse[2],
Hsien-Hsin S. Lee[2], Gu-Yeon Wei[1], David Brooks[1], Carole-Jean Wu[2]

[1]Harvard University, [2]Facebook Inc., [3]Arizona State University

ugupta@g.harvard.edu carolejeanwu@fb.com

***Abstract*—Given recent algorithm, software, and hardware innovation, computing has enabled a plethora of new applications. As computing becomes increasingly ubiquitous, however, so does its environmental impact. This paper brings the issue to the attention of computer-systems researchers. Our analysis, built on industry-reported characterization, quantifies the environmental effects of computing in terms of carbon emissions. Broadly, carbon emissions have two sources: operational energy consumption, and hardware manufacturing and infrastructure. Although carbon emissions from the former are decreasing thanks to algorithmic, software, and hardware innovations that boost performance and power efficiency, the overall carbon footprint of computer systems continues to grow. This work quantifies the carbon output of computer systems to show that most emissions related to modern mobile and data-center equipment come from hardware manufacturing and infrastructure. We therefore outline future directions for minimizing the environmental impact of computing systems.**

***Index Terms*—Data center, mobile, energy, carbon footprint**

I. Introduction

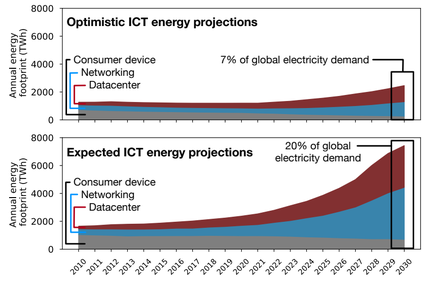


Fig. 1. Projected growth of global energy consumption by information and computing technology (ICT). On the basis of optimistic (top) and expected (bottom) estimates, ICT will by 2030 account for 7% and 20% of global demand, respectively [1].

"efficiently amortizing the manufacturing carbon footprint of a Google Pixel 3 smartphone re… image-classifi… — beyond the…

https://www.nytimes.com/interactive/2022/08/02/world/americas/brazil-airstrips-illegal-mining.html

Wide array of environmental and social impacts fueling our hardware supply chain:

- Water
- Rare earth minerals! (Martin & Iles, 2020)

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

Deep Learning's Carbon Emissions Problem

AI Can Do Great Things—if It Doesn't Burn the Planet

The computing power required for AI landmarks, such as recognizing images and defeating humans at Go, increased 300,000-fold from 2012 to 2018.

Is large LM training destroying the planet?

*nobody is really doing the accounting, so we don't know, especially when it comes to inference.

- *Probably not* due to direct GHG emissions.*
- Datacenters responsible for < 1% global electricity use, small subset* used for AI/ML.

Should we still be concerned about the unprecedented computational requirements of large LMs?

- **Yes! Many practical reasons** (e.g. ML on-device, w/ no internet).

Everything Uses Power

How much power do I lose from AC? Heat?

How much from accelerating too hard? (Model pred error)

How much from each additional camera? Lidar?

How much from the self-driving processing?

Apple M2	NVIDIA A6000	NVIDIA Drive	Tesla FSD
20 Watts	300 Watts	500 Watts	75 Watts

Most ML will be on the edge

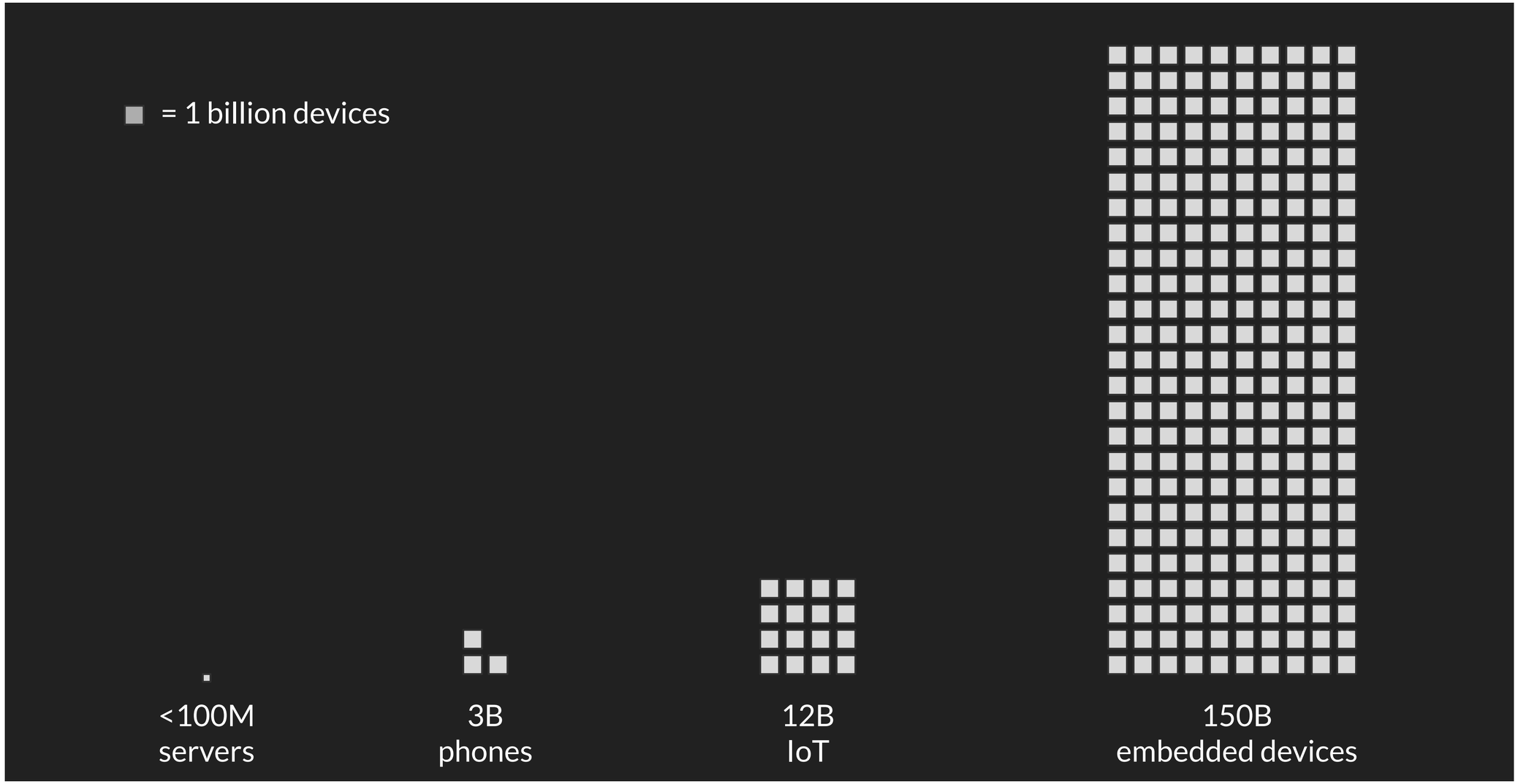


Figure: Jameson Toole. Smaller, faster mobile models. O'Reilly AI London, 2019.

ML accelerators beyond GPU/TPU

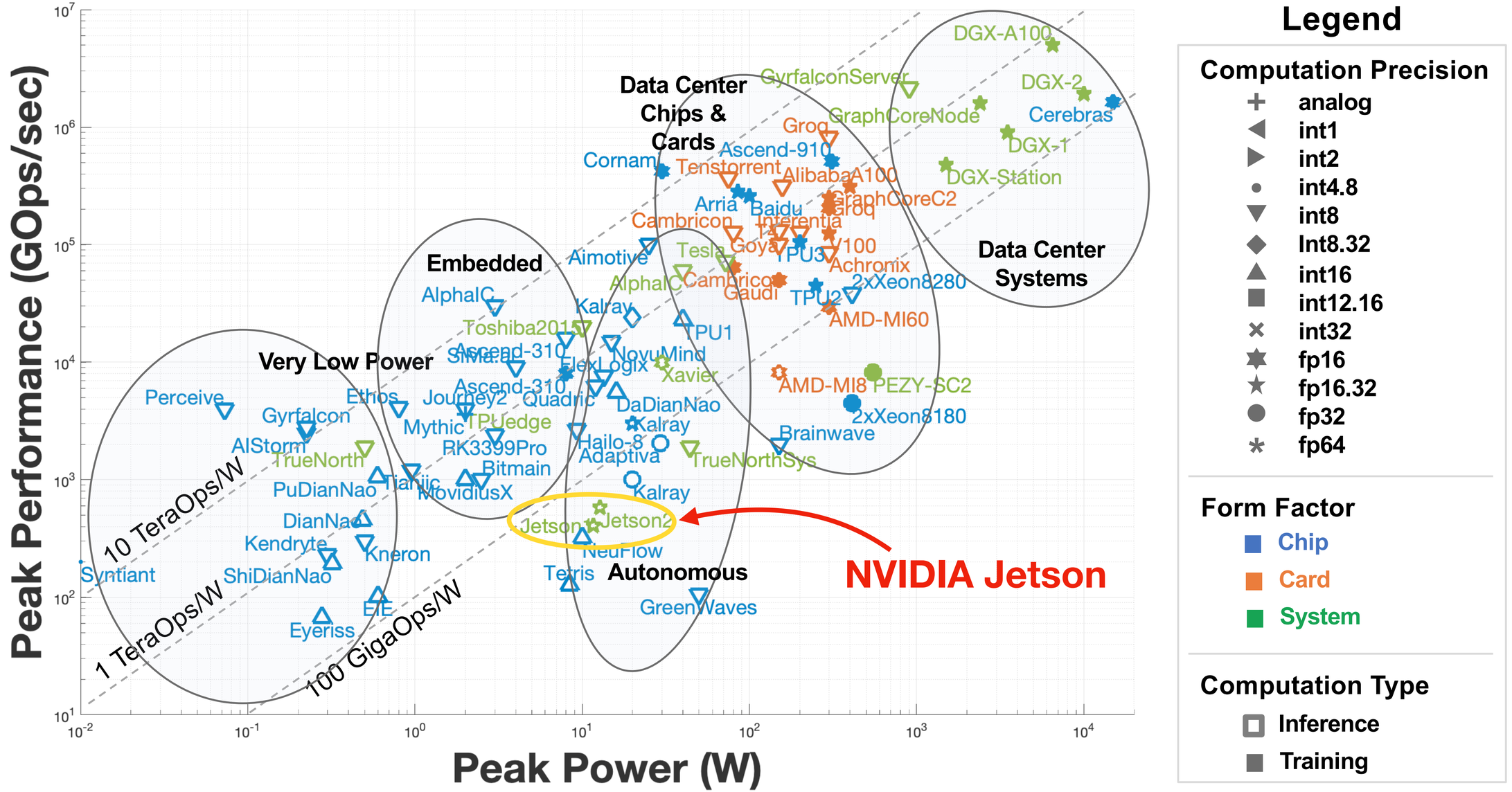
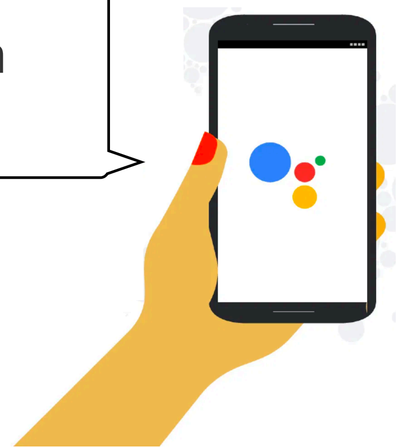


Figure: A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner. Survey of Machine Learning Accelerators. IEEE-HPEC, 2020.

Data privacy for users

VizWiz

Microwave on high for 4 minutes, stir, then an additional 3 to 3.5 minutes.



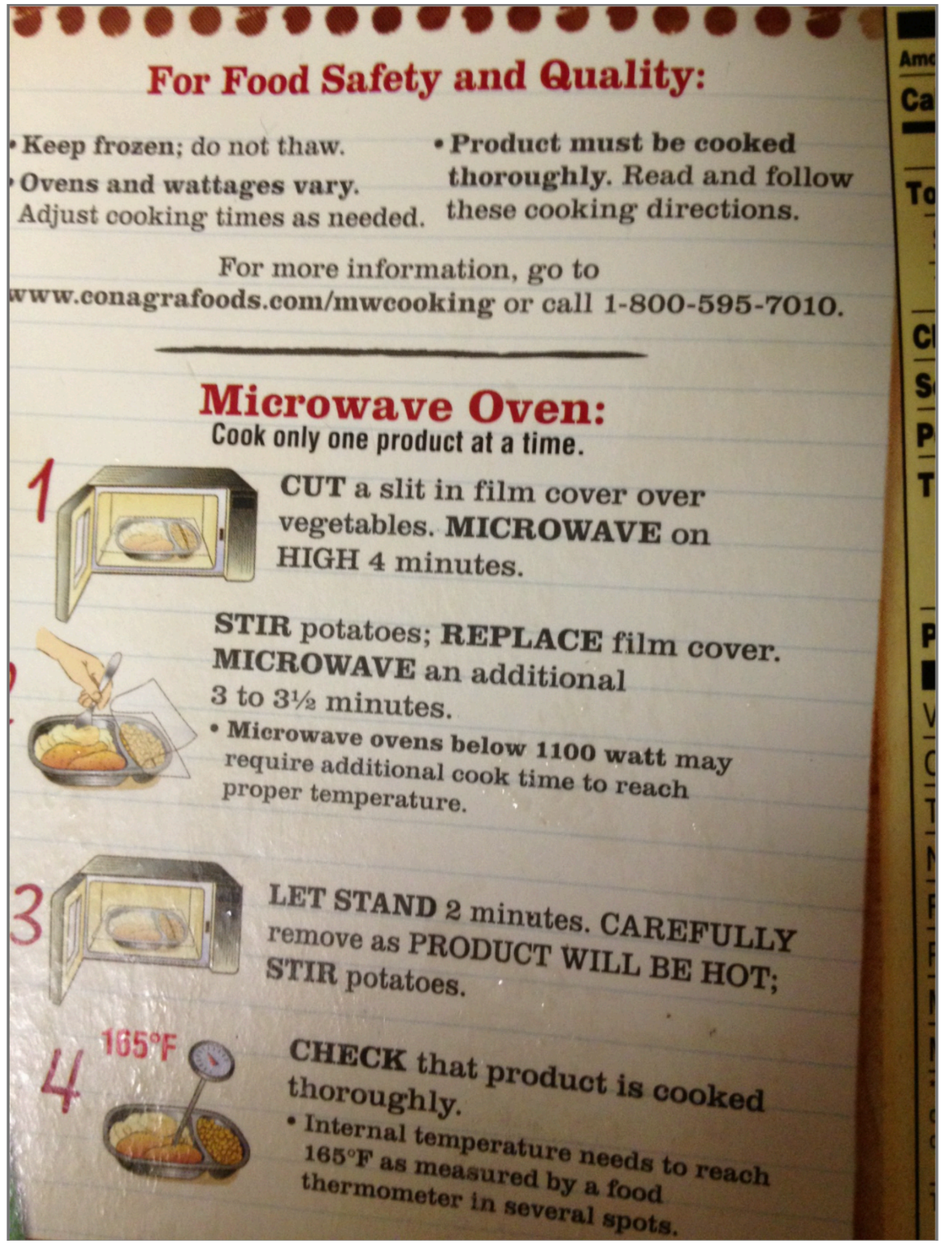


Image: https://www.wikihow.com/Interact-with-the-Blind

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

Deep Learning's Carbon Emissions Problem

AI Can Do Great Things—if It Doesn't Burn the Planet

The computing power required for AI landmarks, such as recognizing images and defeating humans at Go, increased 300,000-fold from 2012 to 2018.

Is large LM training destroying the planet?

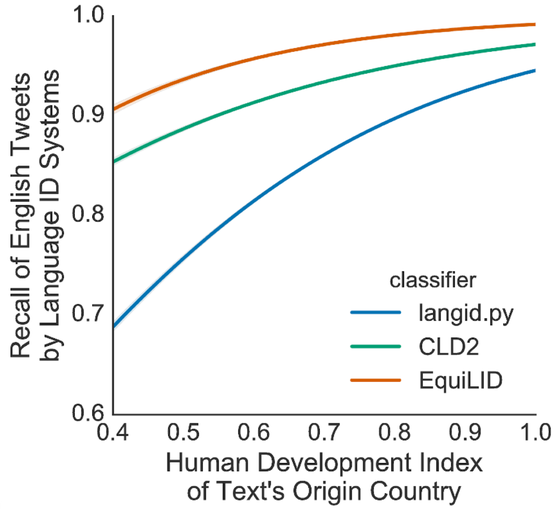
*nobody is really doing the accounting, so we don't know, especially when it comes to inference.

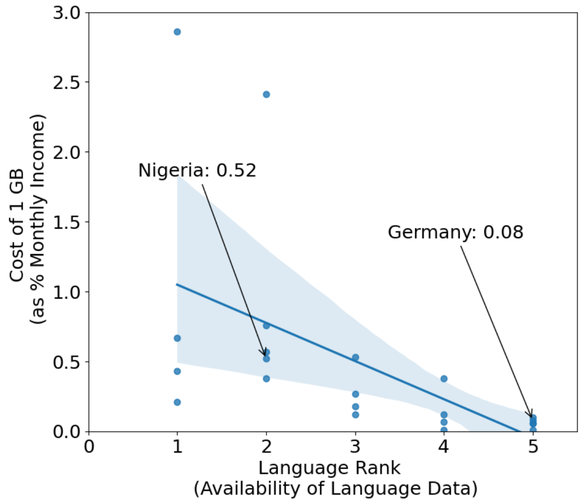
- *Probably not* due to direct GHG emissions.*
- Datacenters responsible for < 1% global electricity use, small subset* used for AI/ML.

Should we still be concerned about the unprecedented computational requirements of large LMs?

- **Yes! Many practical reasons** (e.g. ML on-device, w/ no internet).
- **Ethical issue: equity of access.** Increasingly few individuals/organizations have access to the resources necessary to use, develop, and shape this powerful technology.

Equity of access and aligning values

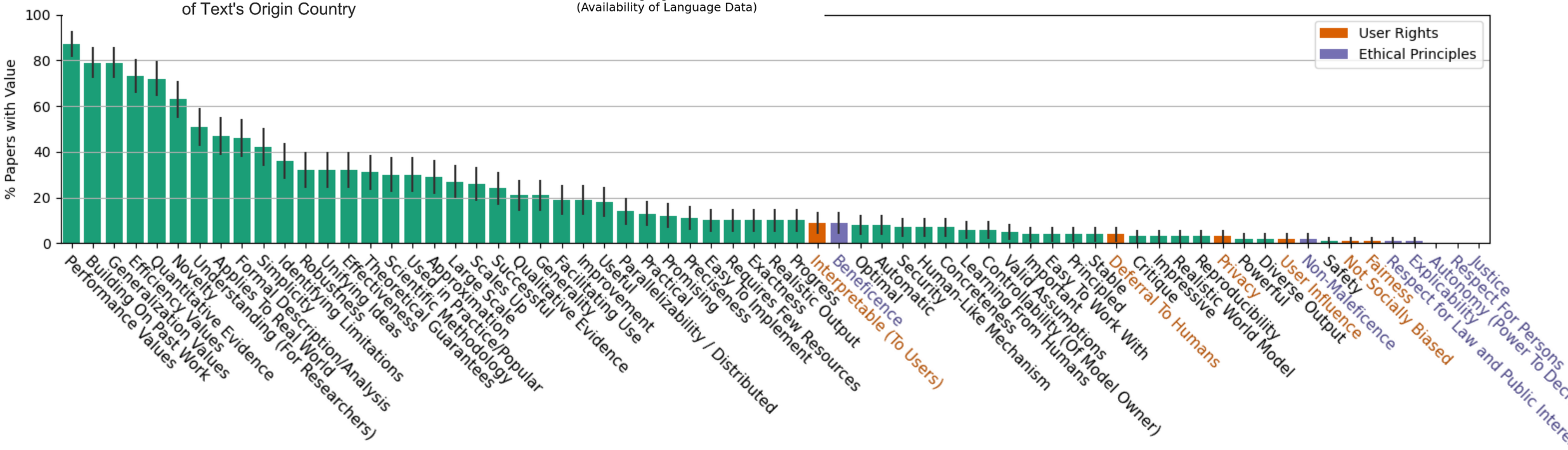




A. Birhane, et al. The Values Encoded in Machine Learning Research. 2020.

O. Ahia, et al. The Low-Resource Double-Bind: An Empirical Study of Pruning for Low-Resource Machine Translation. EMNLP Findings 2021.

D. Jurgens, et al. Incorporating Dialectal Variability for Socially Equitable Language Identification. ACL 2017.



Ok, so what do we do?

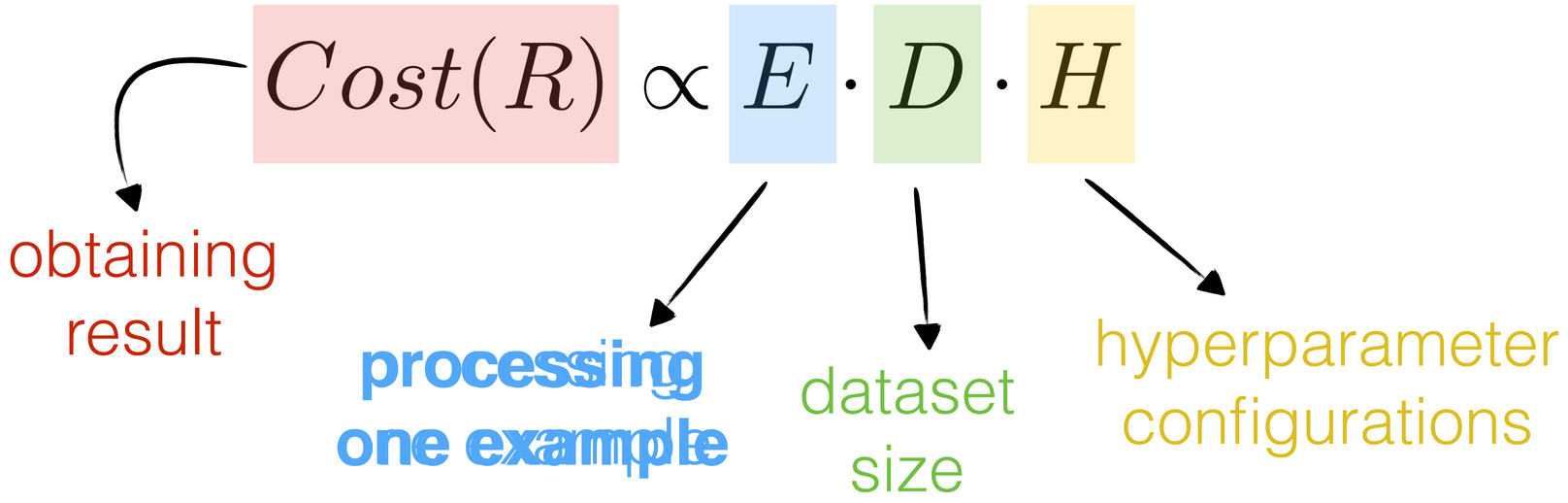
Model development

- **How do we make large models smaller? Avoid them altogether?**
- Reduce wasteful retraining, reimplementation, and execution.
- **Standardize reporting of ML software energy requirements.**

Compute infrastructure

- Reduce emissions across the supply chain and hardware lifecycle.
- Maximize data center energy efficiency.
- Use low-carbon electricity.
- Hardware innovation.

How?



Equation from: Schwartz, Dodge, Smith, Etzioni. Green AI. Communications of the ACM, December 2020.

Challenges

Trading off accuracy and efficiency

- **Pareto efficiency** or **Pareto optimality** is a situation where no individual or preference criterion can be better off without making at least one individual or preference criterion worse off or without any loss thereof.

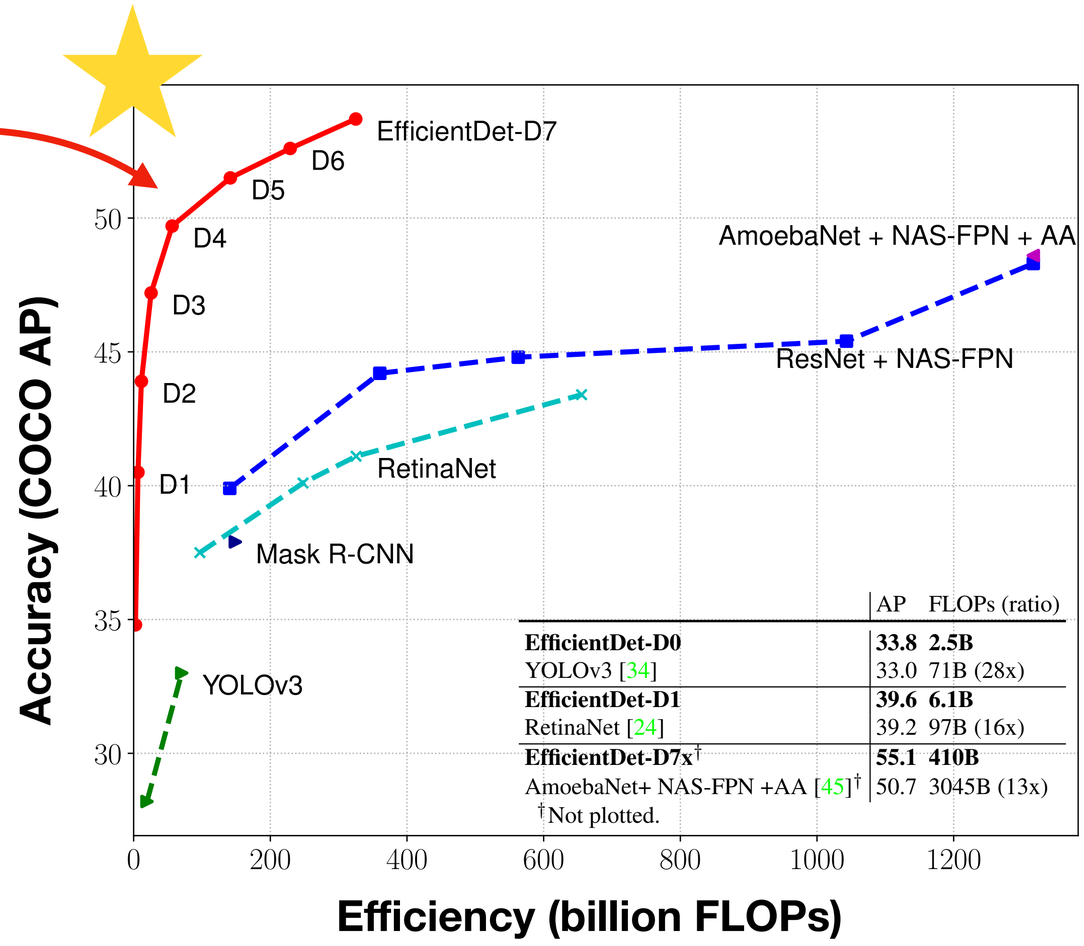


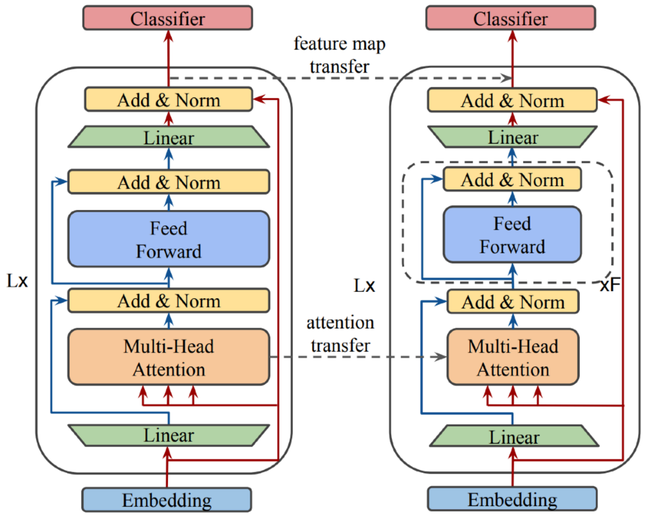
Figure: Mingxing Tan, Ruoming Pang, Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. CVPR 2020.

Solutions: Model compression

- **Premise:** models are over-parameterized
  (but this aids optimization; e.g. lottery ticket hypothesis, Frankle & Carbin 2019)

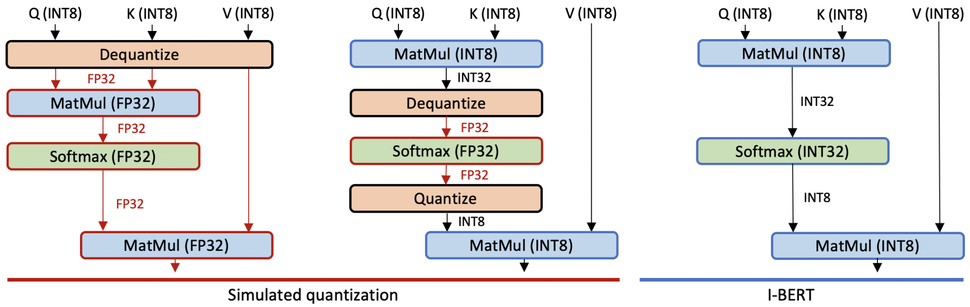
Knowledge distillation:

- Train smaller model to have same activations as larger model (Hinton et al. 2015; Sanh et al. 2020; Sun et al. 2020).



Quantization:

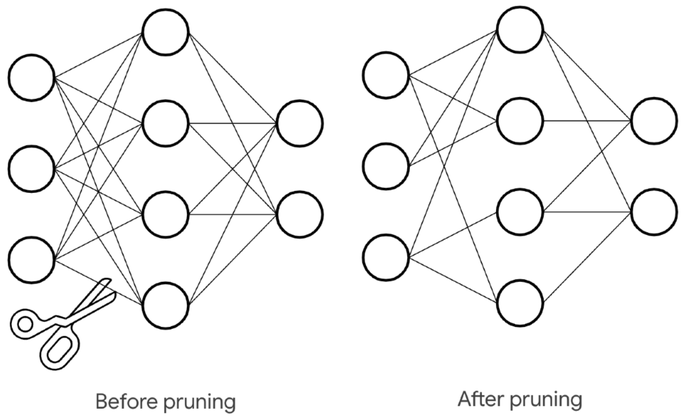
- Use far fewer bits to represent each model parameter (Zafrir et al. 2019; Shen et al 2020, Kim et al. 2021).



Technique	Benefits	Hardware
Dynamic range quantization	4x smaller, 2x-3x speedup	CPU
Full integer quantization	4x smaller, 3x+ speedup	CPU, Edge TPU, Microcontrollers
Float16 quantization	2x smaller, GPU acceleration	CPU, GPU

Pruning / sparsity:

- Remove “unnecessary” model parameters (LeCun et al. 1989; Blalock et al. 2020, Xia et al. 2022).



Images from: https://blog.tensorflow.org/2019/05/tf-model-optimization-toolkit-pruning-API.html

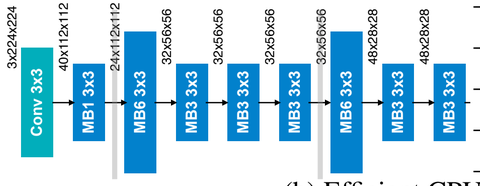
https://www.tensorflow.org/lite/performance/post_training_quantization

https://medium.com/@kaustavtamuly/compressing-and-accelerating-high-dimensional-neural-networks-6b501983c0c8

Solutions: Neural architecture search / AutoML

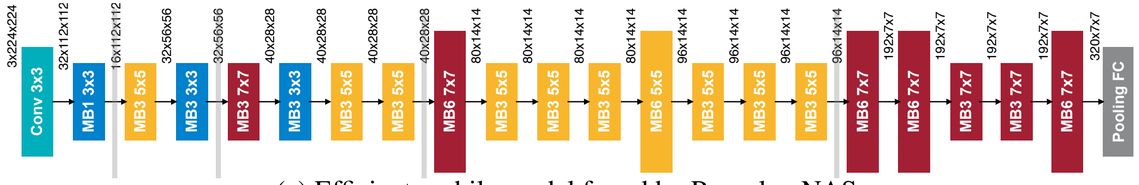
- **Key idea:** One-time expensive, combinatorial search to find a more efficient architecture

(a) Efficient GPU model found by ProxylessNAS.

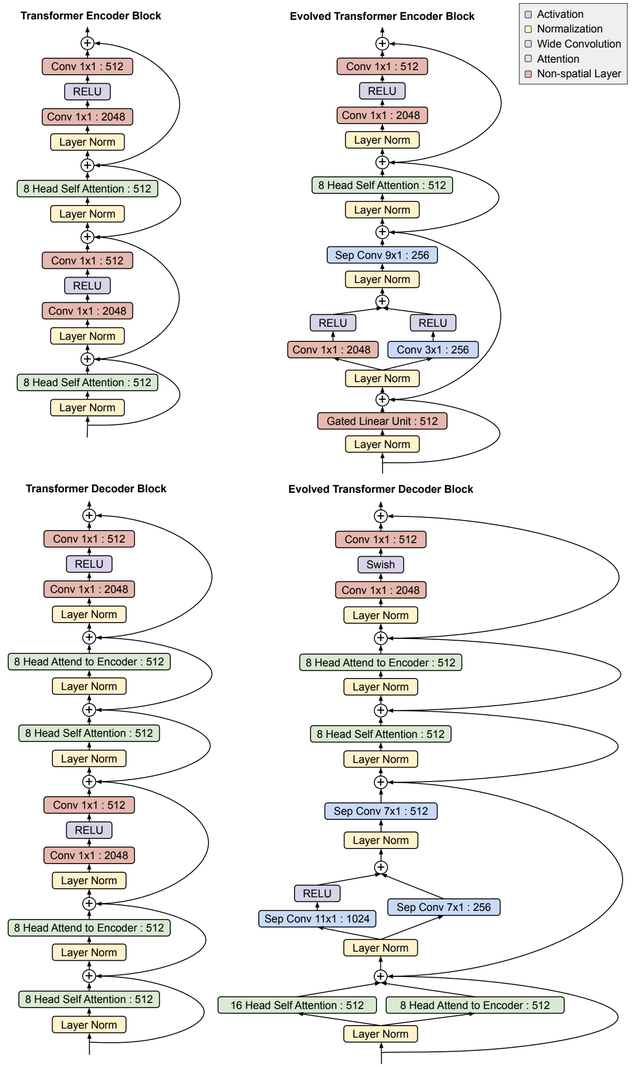


Model	Top-1 (%)	GPU latency	CPU latency	Mobile latency
Proxyless (GPU)	75.1	5.1ms	204.9ms	124ms
Proxyless (CPU)	75.3	7.4ms	138.7ms	116ms
Proxyless (mobile)	74.6	7.2ms	164.1ms	78ms

(b) Efficient CPU model found by ProxylessNAS.



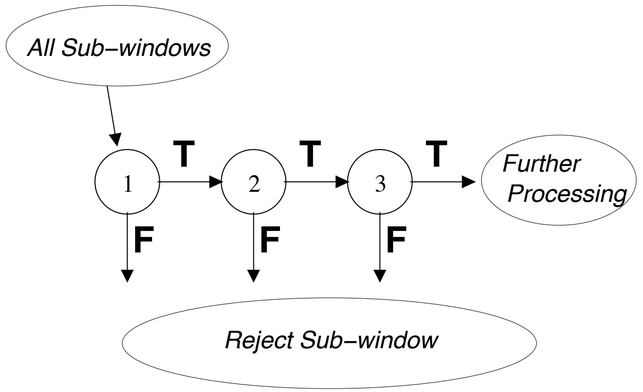
(c) Efficient mobile model found by ProxylessNAS.



- Often, inference efficiency comes at the cost of training efficiency.
- Do these architectures generalize? Is the amortized savings worth the up-front cost?

Solutions: Adaptive compute / coarse-to-fine

- **Key idea:** Use only as much computation as needed for a given example
- Old idea: Viola & Jones, CVPR 2001; Strubell et al., ACL 2015



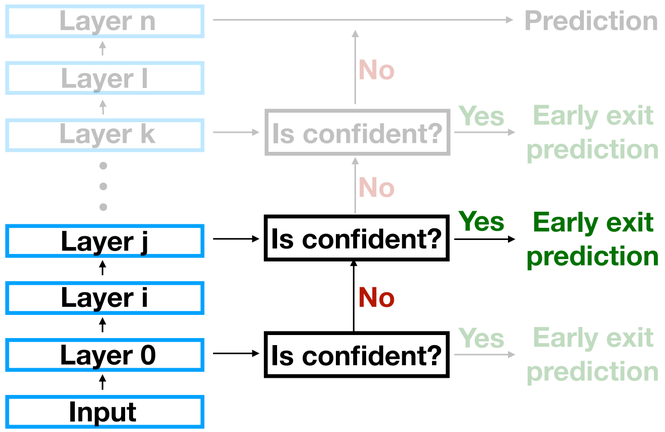
ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2001

Rapid Object Detection using a Boosted Cascade of Simple Features

Paul Viola
viola@merl.com
Mitsubishi Electric Research Labs
201 Broadway, 8th FL
Cambridge, MA 02139

Michael Jones
mjones@crl.dec.com
Compaq CRL
One Cambridge Center
Cambridge, MA 02142

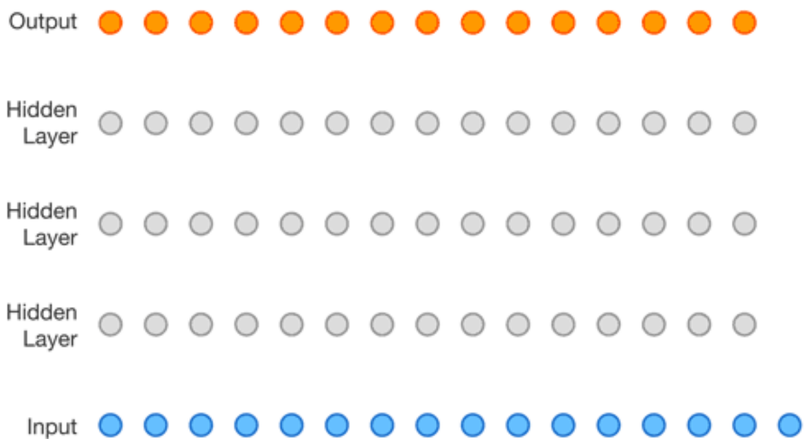
- Lots of new implementations applied to NNs
- How to efficiently decide confidence?
- Unclear how to obtain improvements on accelerators that leverage batch parallelism.



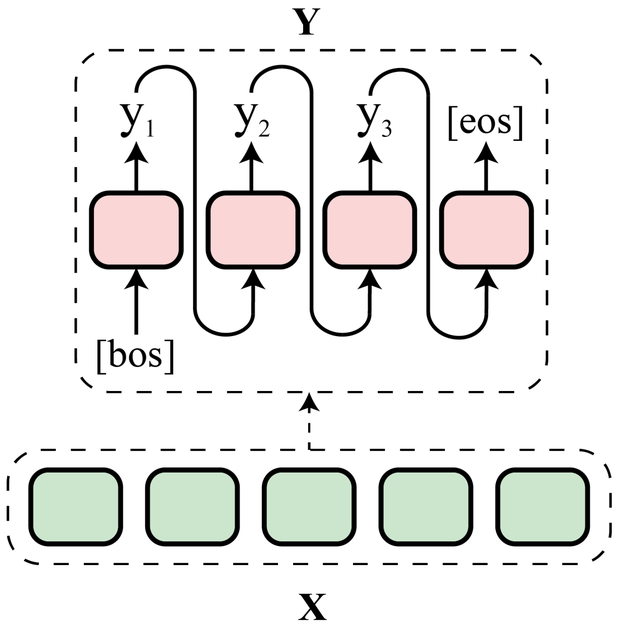
R. Schwartz, G. Stanovsky, S. Swayamdipta, J. Dodge, N. A. Smith. The Right Tool for the Job: Matching Model and Instance Complexities. ACL 2020.

Solutions: Task-specific

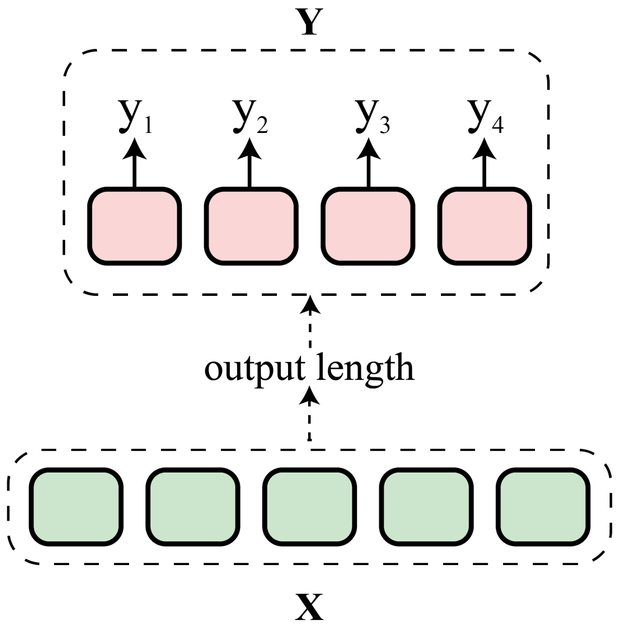
Non-autoregressive machine translation



Autoregressive



Autoregressive



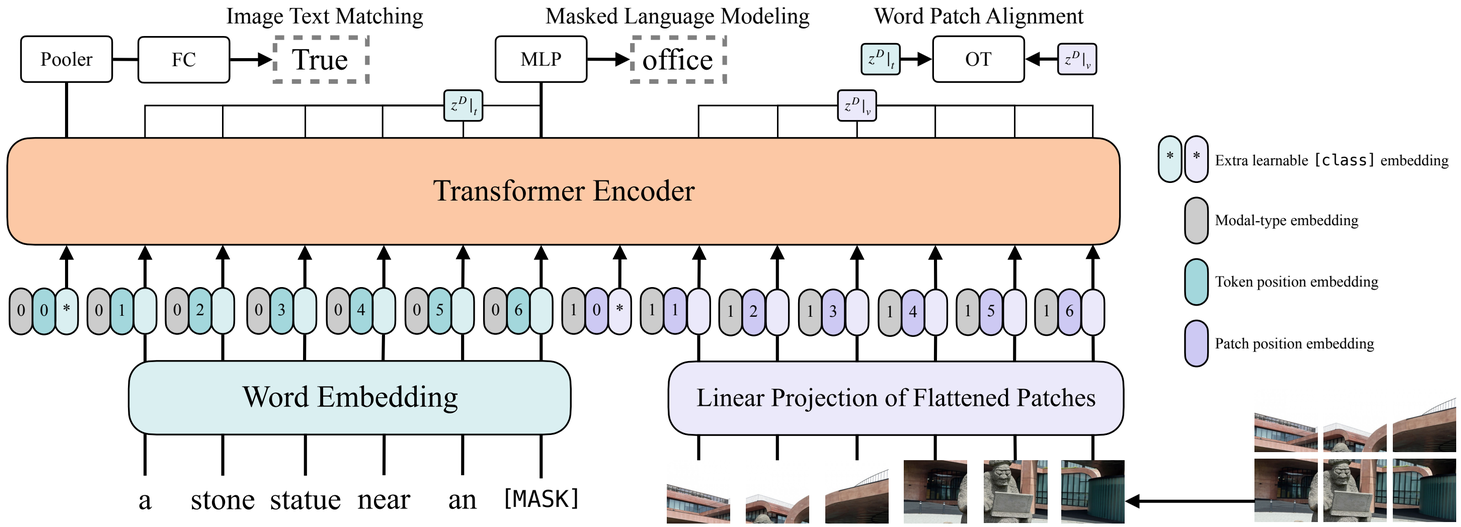
Non-autoregressive

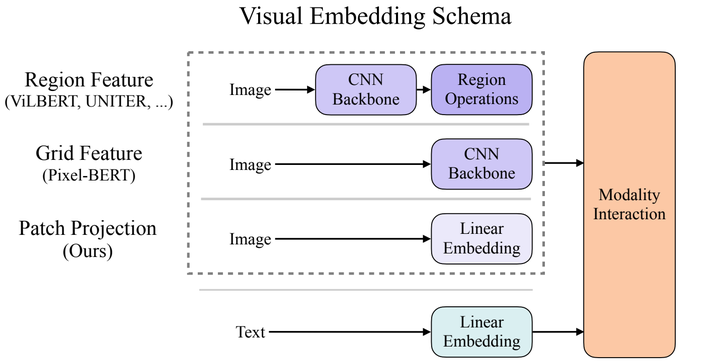
Figures from: Y. Su, D. Cai, Y. Wang, D. Vandyke, S. Baker P. Li, N. Collier. Non-Autoregressive Text Generation with Pre-trained Language Models. EACL 2021.

https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

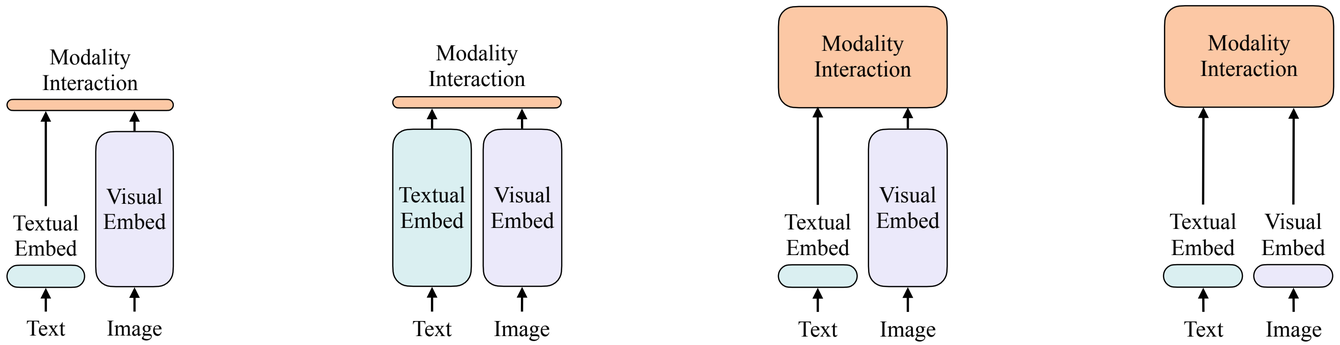
Solutions: Task-specific

Multimodal fusion

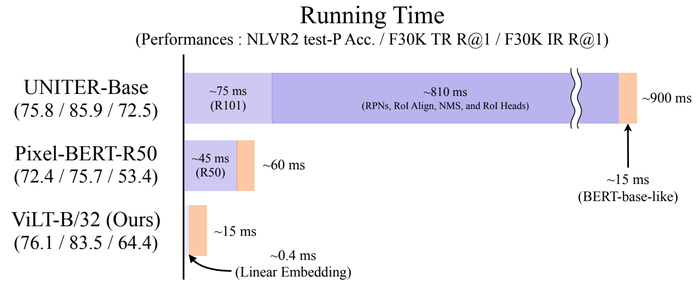




How much modality specific processing is necessary?



(a) VE > TE > MI (b) VE = TE > MI (c) VE > MI > TE (d) MI > VE = TE



ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision
https://arxiv.org/abs/2102.03334

Solutions: Other fun tricks

M. Rastegari, V. Ordonez, J. Redmon, A. Farhadi XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. ECCV 2016.
S. Zhu, L. H. K. Duong, W. Liu. XOR-Net: An Efficient Computation Pipeline for Binary Neural Network Inference on Edge Devices. ICPADS 2020.

- **X(N)OR-Nets:** Binary quantized (0-1) convolutional layers w/ activations computed using just X(N)OR and popcount

```
xor(00100110, 01100000) = 01000110
popcount(01000110) = 3
```

- Dot product of two binary matrices:

```
a = xnor(x, y) = not(xor(x, y))
b = popcount(a)
c = len(a)
dot(x, y) = 2b - c
```

remove NOT operation →

```
a = xor(x, y)
b = popcount(a)
c = len(a)
xor-dot(x, y) = c - 2b
```

- **CNNs as FFTs:** the convolution theorem

convolution ↘ Hadamard product (pointwise) ↙

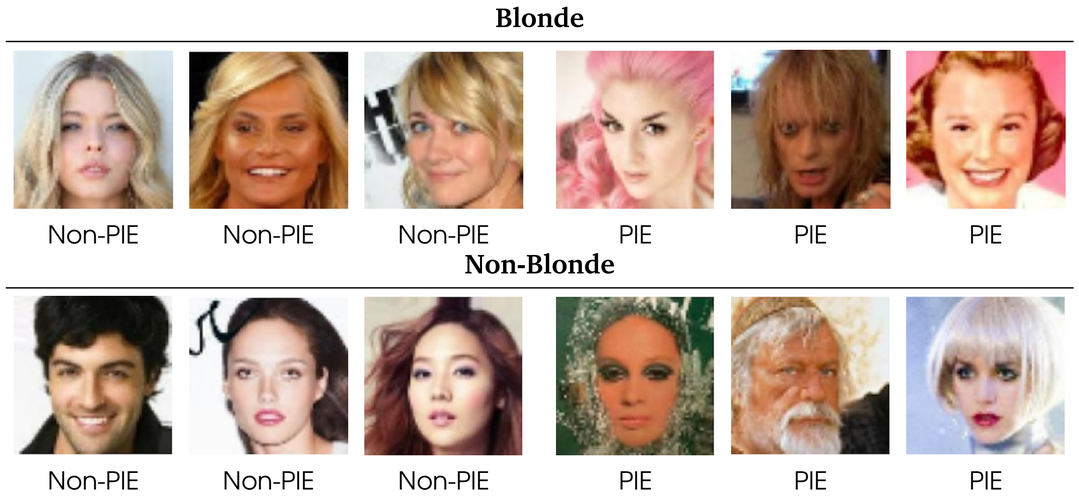
$$\mathcal{F}(\kappa * u) = \mathcal{F}(\kappa) \odot \mathcal{F}(u)$$

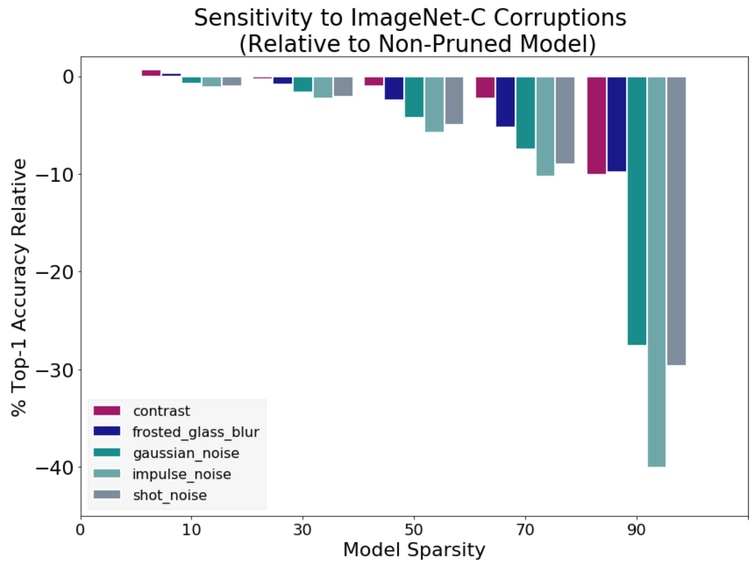
N. Vasilache, J. Johnson, M. Mathieu, S. Chintala, S. Piantino, Y. LeCun. Fast Convolutional Nets With fbfft: A GPU Performance Evaluation. ICLR 2015.
H. Pratt, B. Williams, F. Coenen, and Y. Zheng. FCNN: Fourier Convolutional Neural Networks. ECML-PKDD 2017.

Challenges

Increased bias / reduced generalization in compressed models

- Trained models fit the (undesirable) biases in their training data, such as higher error rates on attributes under-represented in the set.
- “Compression consistently **amplifies** the disparate treatment of underrepresented protected subgroups for all levels of compression that we consider” — Hooker et al. 2020





Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. Characterising and mitigating bias in compact models. WHI 2020.

Hooker, S., Courville, A., Clark, G., Dauphin, Y., & Frome, A. What Do Compressed Deep Neural Networks Forget? 2019.

Questions?

Course structure

Project-based, lab-based course

Lectures

- Introducing core concepts & papers
- Covering material necessary for labs
- Lecture highlights

Labs/Projects

- 5 Labs
- Project Presentations
- Lab reports

Goal: On-Device thinking for your own domain

Grading

- You will be graded based on a combination of group project work and discussion:

Lectures:

Lecture highlights: 20%

Labs:

Reports (5): 50%

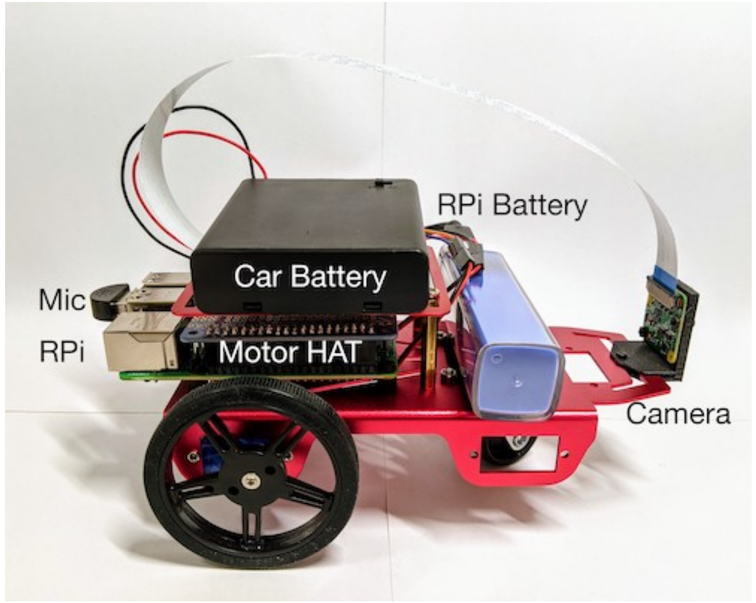
Presentation:

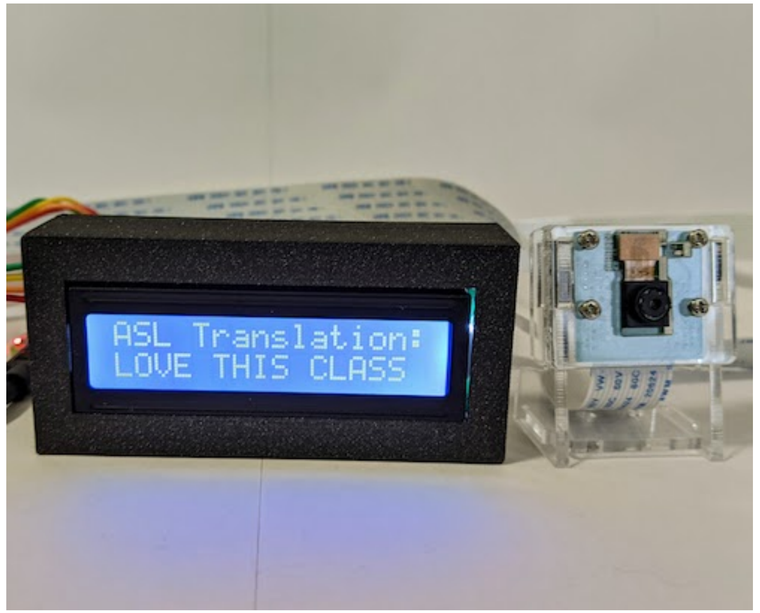
Proposal: 10%

Midterm: 10%

Final: 10%

Project/Lab





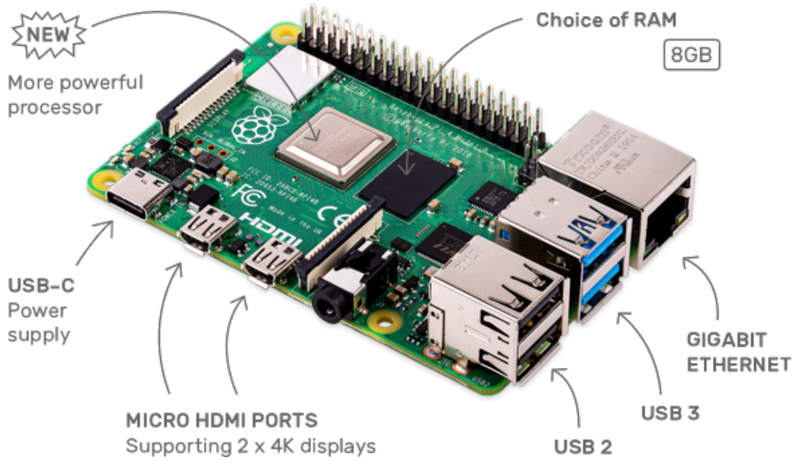
- You will work in groups of 3-4, with some classes devoted to group work on projects.
- **We will have a class for group formation, but you can start now!**

  (recommend stating your goals/interests/background to facilitate finding others)
- Project Proposal with Team due **Friday, September 29**

Yes, we can provide physical devices!

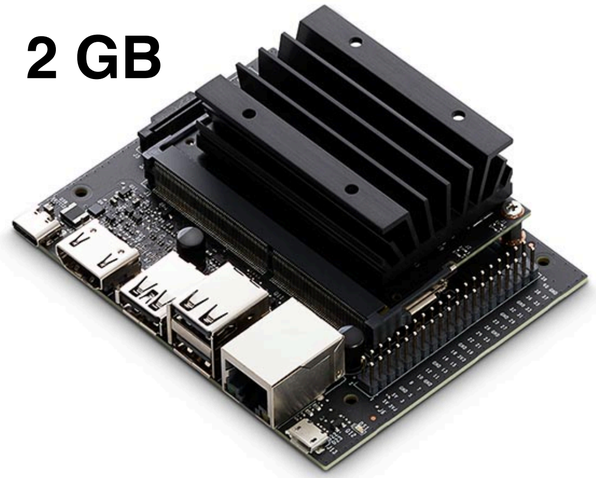
- Once we understand your goals and interests, we can decide if these are right for you

Raspberry Pi 4



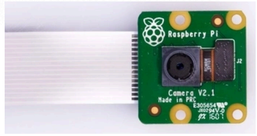
NVIDIA Jetson Nano Developer Kit

2 GB



4 GB

Plus:



- Camera
- USB Microphone
- USB WiFi dongle (for Jetson)
- 32GB SD card
- Necessary cables, etc.

- What task motivates you? What hardware does it need?
- Next week's lecture will discuss hardware in more detail.
- Start brainstorming now!

Bigger != Better

Lab reports: 50%

- Write up your Thursday in-class work as a report due the following Friday midnight.
- Each report is worth 10% of your grade.
- Lab 1 is individual, to get setup and will be done on your laptop.
- Labs 2-5 require experiments and benchmarking on the physical hardware devices.
- Labs teach you basic skills you need but also require you to think about how your task, data, hardware, domain, … is unique

Lecture Highlights: 20%

- What are the key take-aways?

  — Every summary is worth 2 points

- All highlights are submitted via Canvas (individually)

Most Lectures will contain a highlight (indicated on website and will be on canvas)

What if I get COVID?

Please do not come to class if you are sick or suspect you might be.

If you we notice you are exhibiting symptoms, we will ask you to leave.

If you become sick or are exposed awaiting test results, inform the instructors ASAP and we will make arrangements for you to join remotely.

You will still be responsible for all assignments, participation, etc. and subject to the normal late day policy.

11-767 Community Guidelines

- In this class, every individual will and must be treated with respect.
- The ways we are diverse are many and are fundamental to building and maintaining an equitable and inclusive campus community. Research shows that greater diversity across individuals leads to greater creativity in the group.
- It is the responsibility of each of us to create a safer and more inclusive environment.
- Have empathy.

https://www.youtube.com/watch?v=3LopI4YeC4I

Questions?

Course website: cmu-odml.github.io

Canvas: https://canvas.cmu.edu/courses/35445

Email: 11-767-instructors@cs.cmu.edu