NeurIPS 2019 recap

René Donner, contextflow

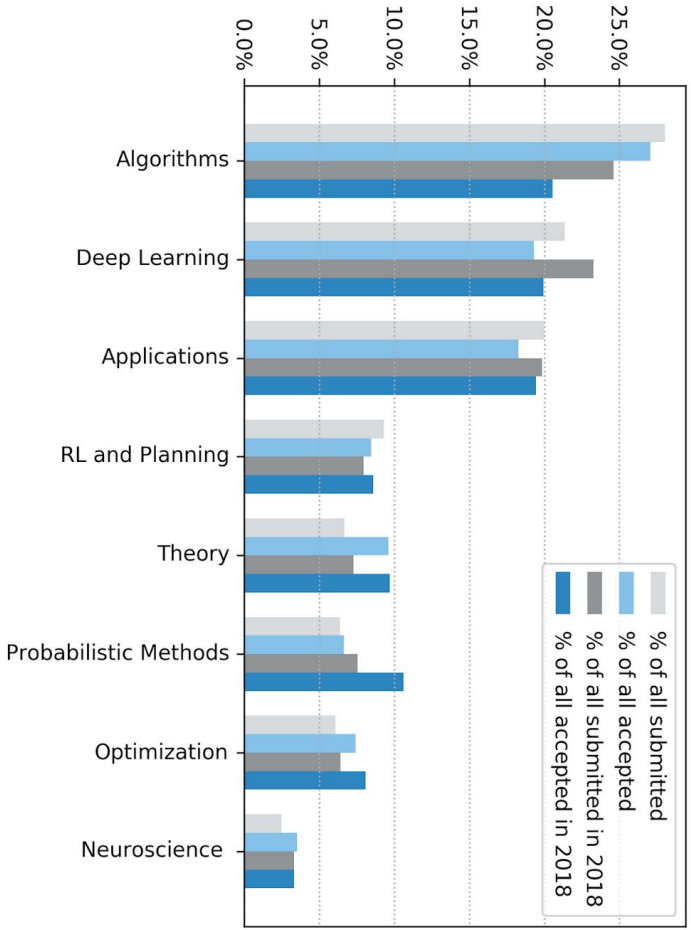

Overview
Tutorials
Papers
Workshops & Hardware

Overview

13.500 attendees - ticket lottery
1000 papers at main conference
430 papers at workshops

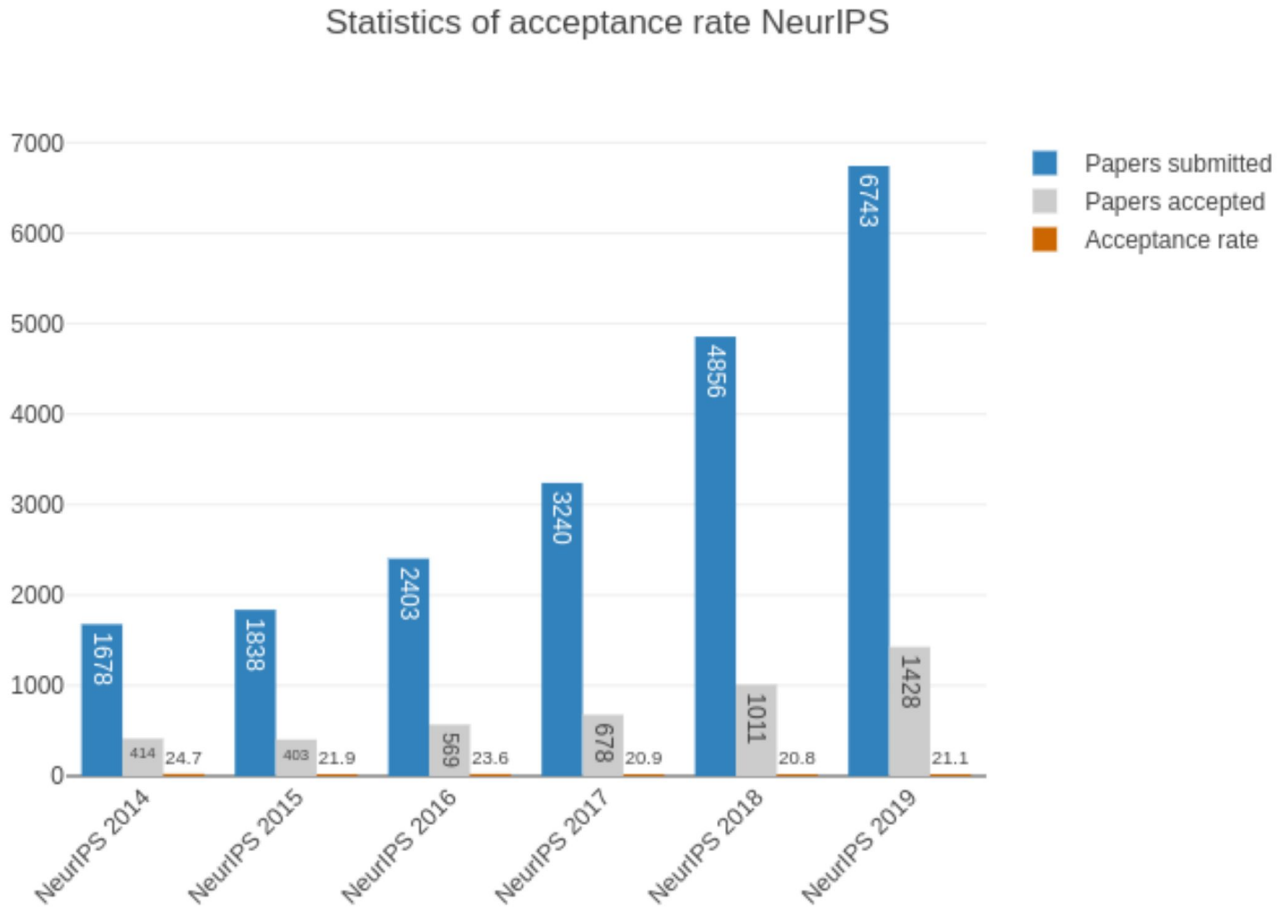
1 day of industrial talks / expo
1 day tutorials
3 day orals + posters
2 days workshops

Overview



https://medium.com/@NeurIPSConf/what-we-learned-from-neurips-2019-data-111ab996462c

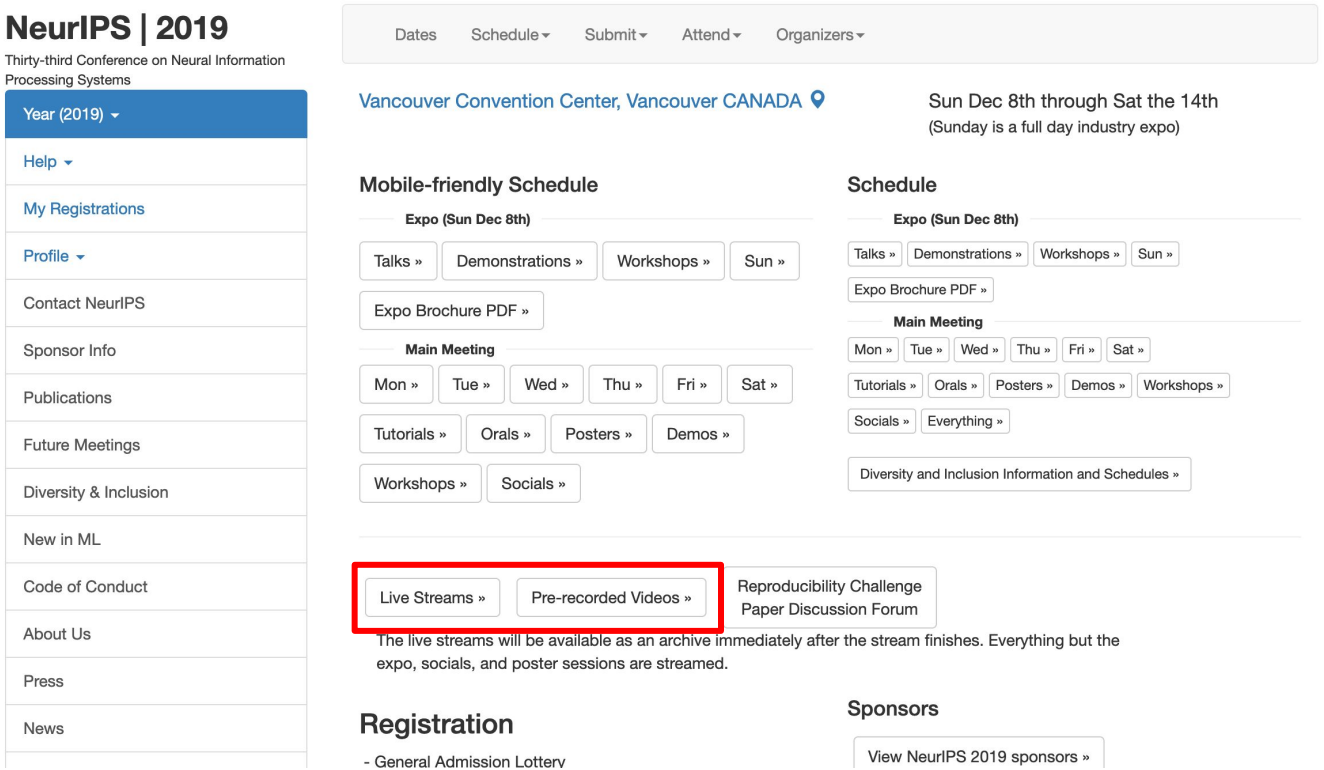
Overview



https://medium.com/@dcharrezt/neurips-2019-stats-c91346d31c8f


Welcome to NeurIPS
Poster Session is Full
Please try again later
NeurIPS 2019

Videos


NeurIPS | 2019
Thirty-third Conference on Neural Information Processing Systems
Year (2019)
Help
My Registrations
Profile
Contact NeurIPS
Sponsor Info
Publications
Future Meetings
Diversity & Inclusion
New in ML
Code of Conduct
About Us
Press
News
Dates
Schedule
Submit
Attend
Organizers
Vancouver Convention Center, Vancouver CANADA
Sun Dec 8th through Sat the 14th
(Sunday is a full day industry expo)
Mobile-friendly Schedule
Expo (Sun Dec 8th)
Talks »
Demonstrations »
Workshops »
Sun »
Expo Brochure PDF »
Main Meeting
Mon »
Tue »
Wed »
Thu »
Fri »
Sat »
Tutorials »
Orals »
Posters »
Demos »
Workshops »
Socials »
Schedule
Expo (Sun Dec 8th)
Talks »
Demonstrations »
Workshops »
Sun »
Expo Brochure PDF »
Main Meeting
Mon »
Tue »
Wed »
Thu »
Fri »
Sat »
Tutorials »
Orals »
Posters »
Demos »
Workshops »
Socials »
Everything »
Diversity and Inclusion Information and Schedules »
Live Streams »
Pre-recorded Videos »
Reproducibility Challenge
Paper Discussion Forum
The live streams will be available as an archive immediately after the stream finishes. Everything but the expo, socials, and poster sessions are streamed.
Registration
- General Admission Lotterv
Sponsors
View NeurIPS 2019 sponsors »

Schedule

Thu Dec 12th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #84 Poster

PC-Fairness: A Unified Framework for Measuring Causality-based Fairness

Yongkai Wu · Lu Zhang · Xintao Wu · Hanghang Tong In Applications -- Fairness, Accountability, and Transparency

Paper » Poster » Slides »

Thu Dec 12th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #85 Poster

This Looks Like That: Deep Learning for Interpretable Image Recognition

Chaofan Chen · Oscar Li · Daniel Tao · Alina Barnett · Cynthia Rudin · Jonathan K Su In Applications -- Fairness, Accountability, and Transparency

Paper » 3 min Video »

Thu Dec 12th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #86 Poster

Towards Automatic Concept-based Explanations

Amirata Ghorbani · James Wexler · James Zou · Been Kim In Applications -- Fairness, Accountability, and Transparency

Paper » Slides »

Thu Dec 12th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #87 Poster

Adversarial Training and Robustness for Multiple Perturbations

Florian Tramer · Dan Boneh In Applications -- Privacy, Anonymity, and Security

Paper » Slides » 3 min Video »

Thu Dec 12th 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #88 Poster


Overview
Tutorials
Papers
Workshops & Hardware

9 Tutorials

Mon Dec 9th 11:15 AM -- 01:15 PM @ West Exhibition Hall C + B3 Tutorial

Efficient Processing of Deep Neural Network: from Algorithms to Hardware Architectures

Vivienne Sze

Mon Dec 9th 02:45 -- 04:45 PM @ West Exhibition Hall A Tutorial

Representation Learning and Fairness

Moustapha Cisse · Sanmi Koyejo

Mon Dec 9th 05:45 -- 06:35 PM @ West Exhibition Hall C + B3 Invited Talk

How to Know

Celeste Kidd

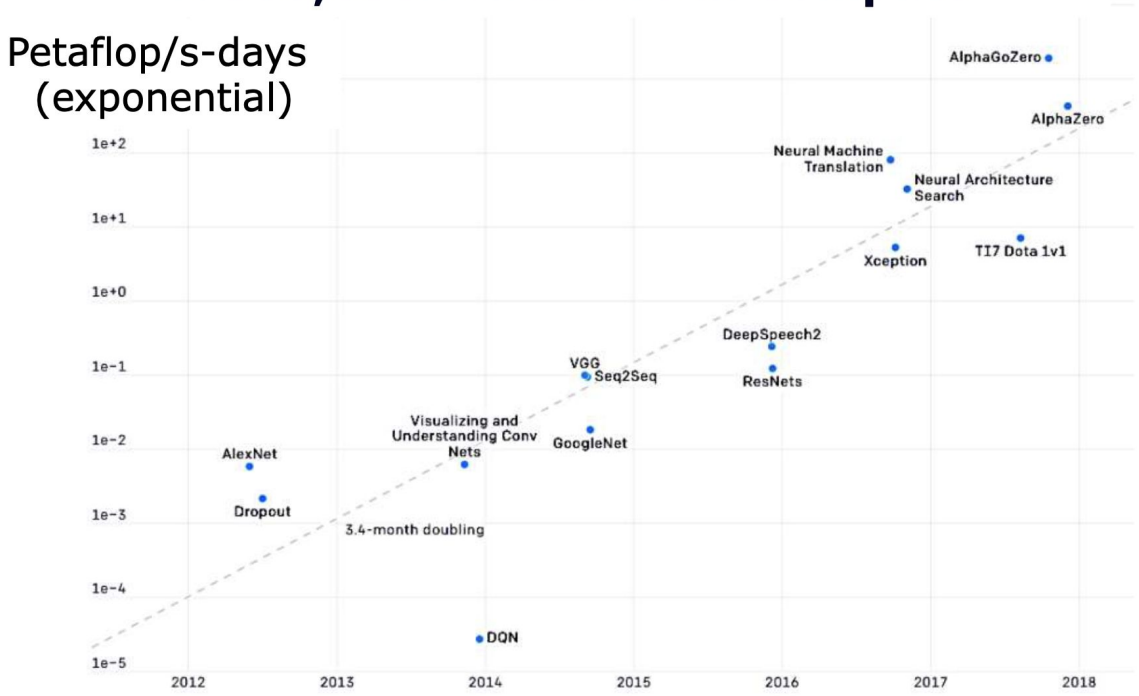
PDF »

Mon Dec 9th 11:15 AM -- 01:15 PM @ West Exhibition Hall C + B3 Tutorial

Efficient Processing of Deep Neural Network: from Algorithms to Hardware Architectures

Vivienne Sze

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute

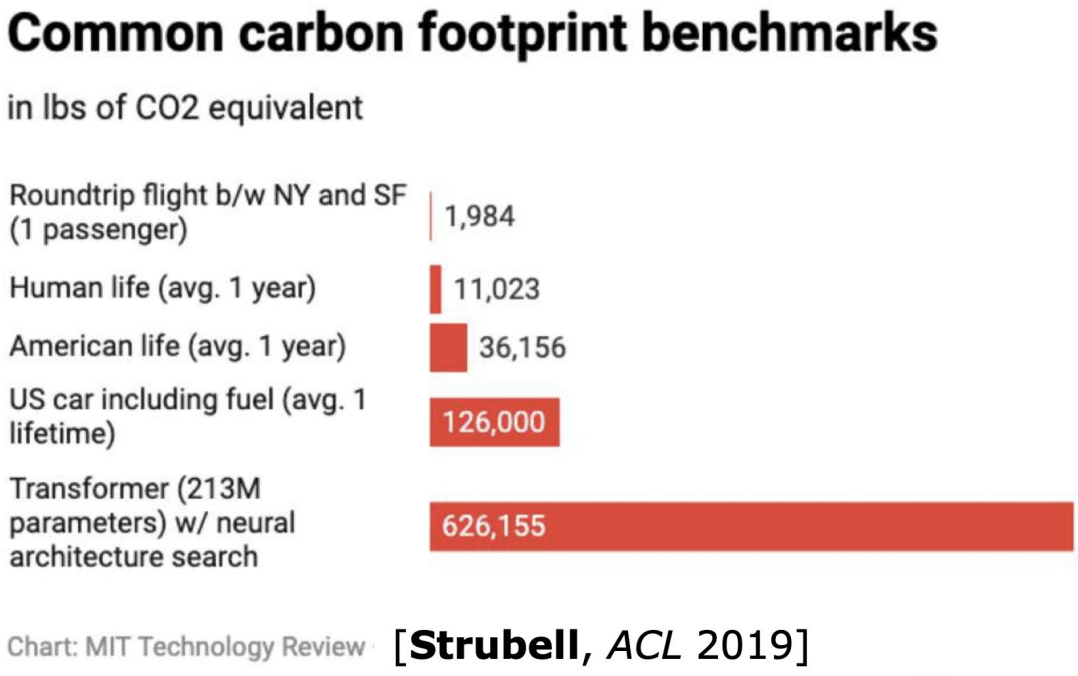


http://eyeriss.mit.edu/2019_neurips_tutorial.pdf

Mon Dec 9th 11:15 AM -- 01:15 PM @ West Exhibition Hall C + B3 Tutorial

Efficient Processing of Deep Neural Network: from Algorithms to Hardware Architectures

Vivienne Sze



[**Strubell**, *ACL* 2019]

http://eyeriss.mit.edu/2019_neurips_tutorial.pdf

Mon Dec 9th 02:45 -- 04:45 PM @ West Exhibition Hall A

Tutorial

Representation Learning and Fairness

Moustapha Cisse · Sanmi Koyejo

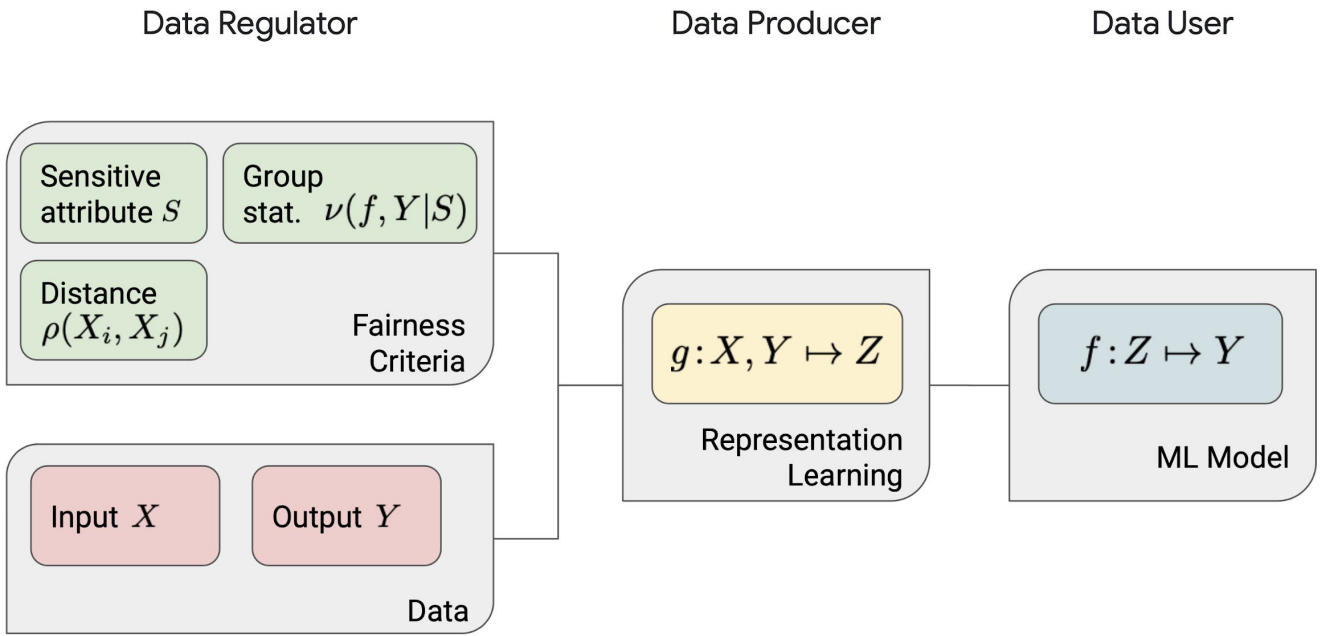
http://sanmi.cs.illinois.edu/documents/Representation_Learning_Fairness_NeurIPS19_Tutorial.pdf

Mon Dec 9th 02:45 -- 04:45 PM @ West Exhibition Hall A

Tutorial

Representation Learning and Fairness

Moustapha Cisse · Sanmi Koyejo



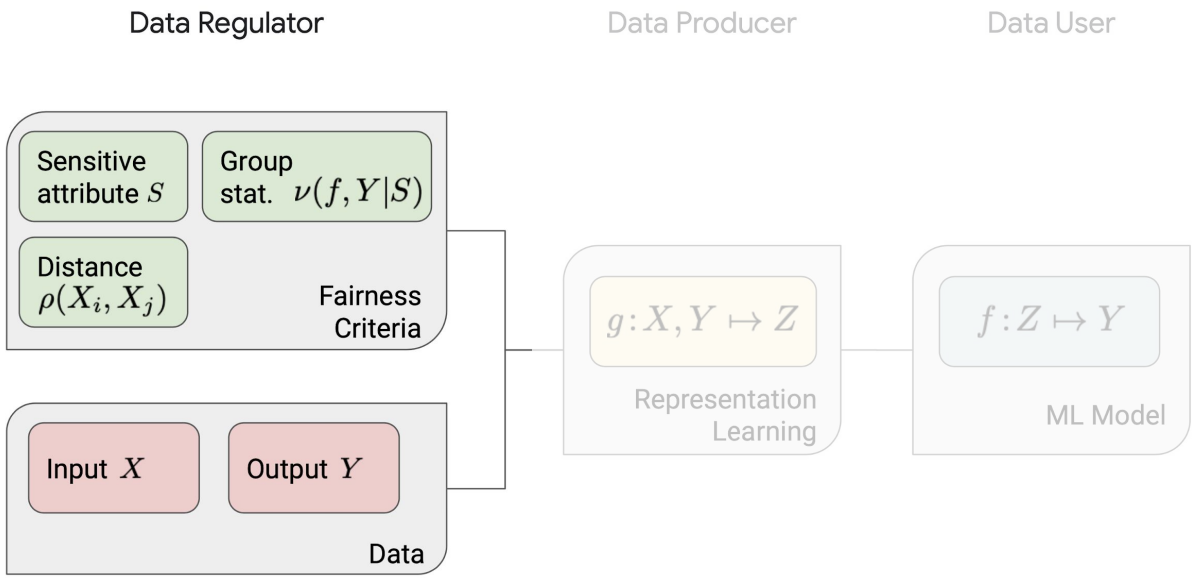
http://sanmi.cs.illinois.edu/documents/Representation_Learning_Fairness_NeurIPS19_Tutorial.pdf

Mon Dec 9th 02:45 -- 04:45 PM @ West Exhibition Hall A

Tutorial

Representation Learning and Fairness

Moustapha Cisse · Sanmi Koyejo



Data Regulator

Determines fairness criteria, determines data source(s), audits results

- INPUT: Data
- OUTPUT: Fairness criteria

AUDITING

- INPUT: Models
- OUTPUT: Satisfactory?

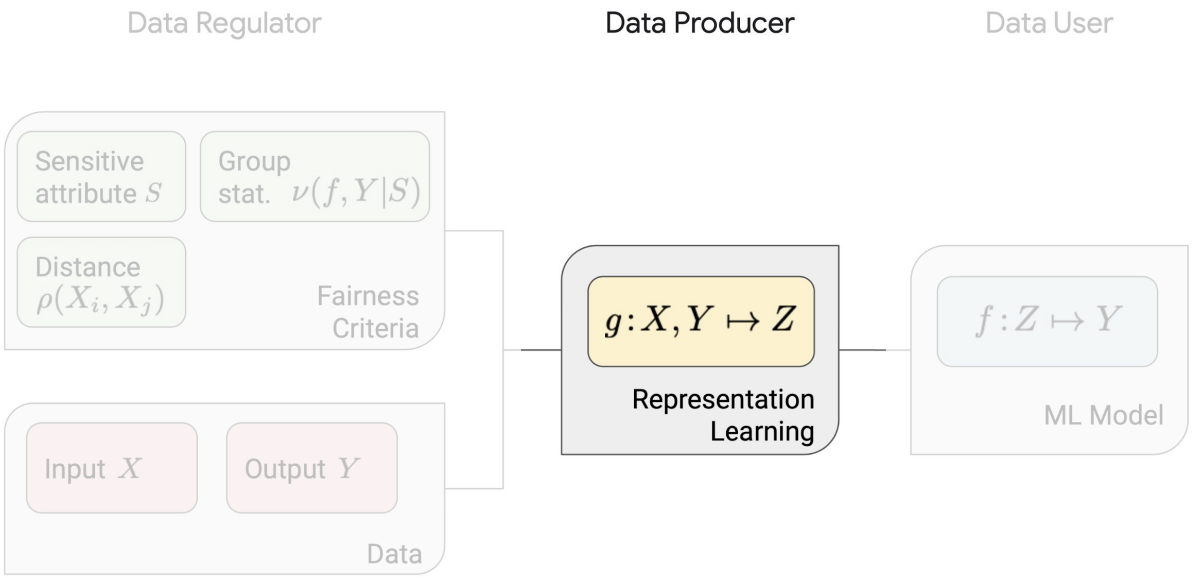
http://sanmi.cs.illinois.edu/documents/Representation_Learning_Fairness_NeurIPS19_Tutorial.pdf

Mon Dec 9th 02:45 -- 04:45 PM @ West Exhibition Hall A

Tutorial

Representation Learning and Fairness

Moustapha Cisse · Sanmi Koyejo



Data Producer

Computes the fair representation given data regulator criteria

- INPUT: Fairness criteria
- OUTPUT: Representation

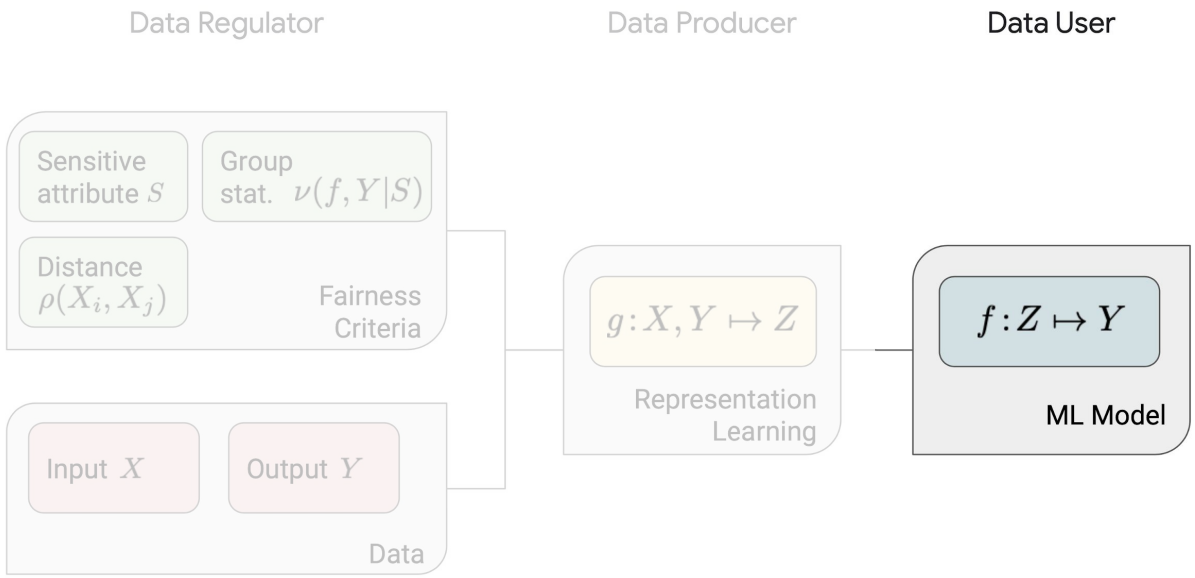
http://sanmi.cs.illinois.edu/documents/Representation_Learning_Fairness_NeurIPS19_Tutorial.pdf

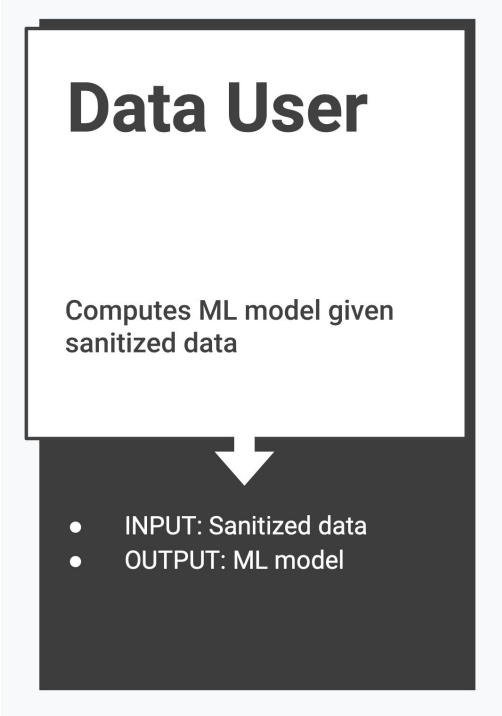
Mon Dec 9th 02:45 -- 04:45 PM @ West Exhibition Hall A

Tutorial

Representation Learning and Fairness

Moustapha Cisse · Sanmi Koyejo





http://sanmi.cs.illinois.edu/documents/Representation_Learning_Fairness_NeurIPS19_Tutorial.pdf

Mon Dec 9th 05:45 -- 06:35 PM @ West Exhibition Hall C + B3 Invited Talk

How to Know

Celeste Kidd

PDF »

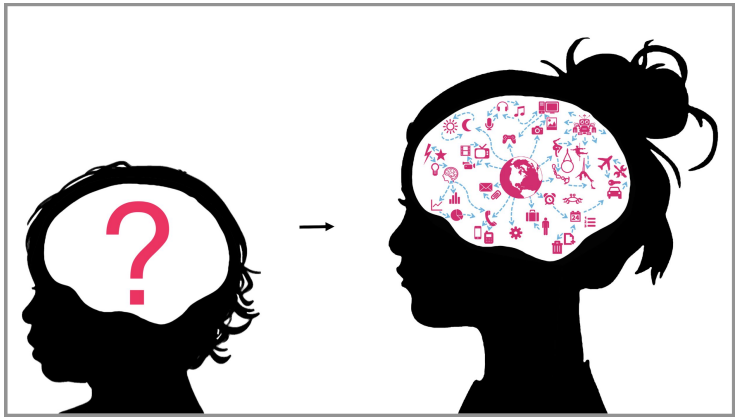
Berkeley
UNIVERSITY OF CALIFORNIA

How to know

Celeste Kidd (celestekidd@berkeley.edu)
Psychology, University of California, Berkeley

@celestekidd

NeurIPS, Vancouver, 9 Dec 2019




Overview
Tutorials
Papers
Workshops & Hardware

SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems

Alex Wang* New York University

Yada Pruksachatkun* New York University

Nikita Nangia* New York University

Amanpreet Singh* Facebook AI Research

Julian Michael University of Washington

Felix Hill DeepMind

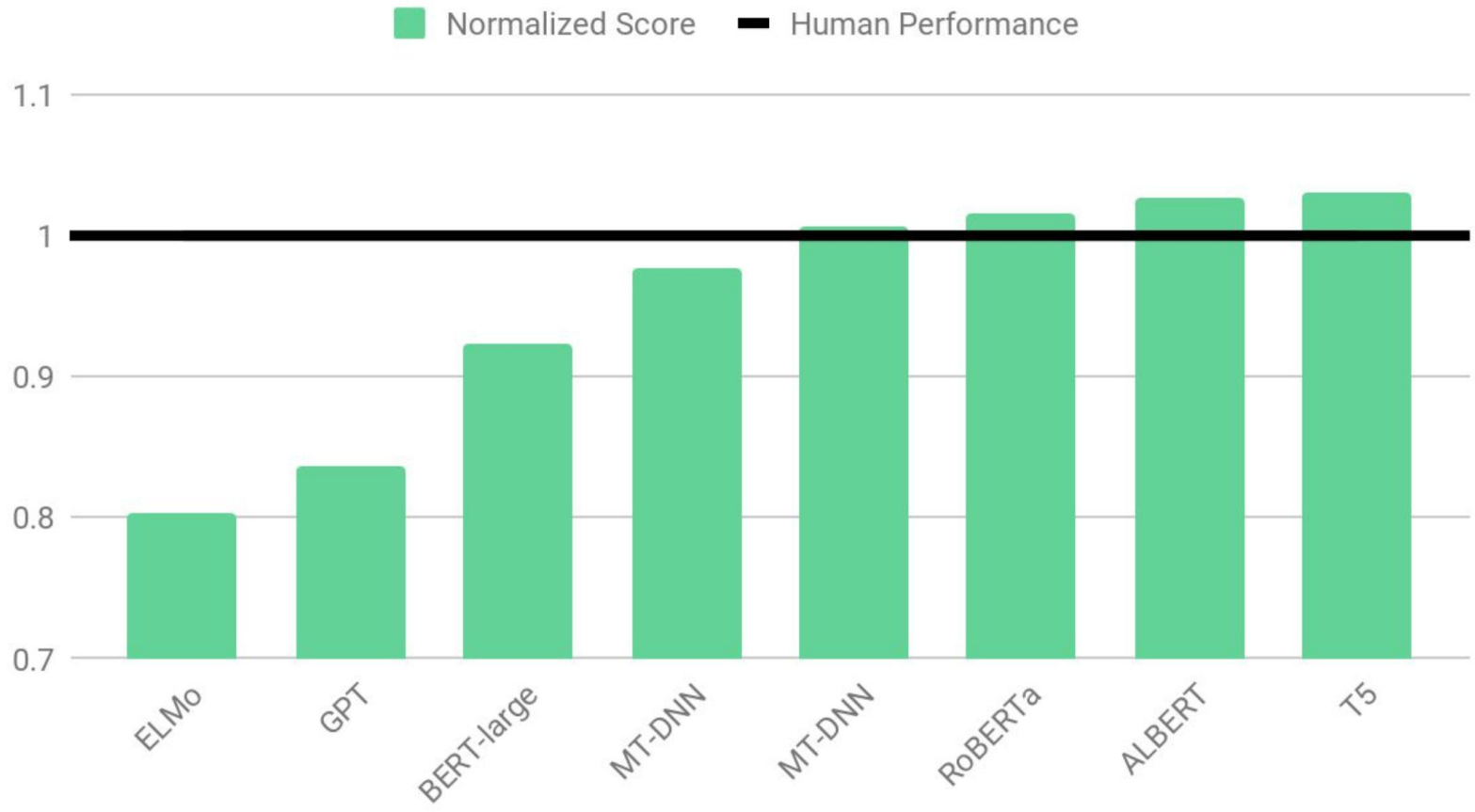
Omer Levy Facebook AI Research

Samuel R. Bowman New York University

First Attempt: GLUE

- Benchmark of 9 sentence- and sentence-pair classification tasks
  - Different tasks (sentiment analysis, paraphrase detection, etc.), genre, amount of data
  - Evaluate system on all nine tasks; overall score is average across tasks
- Released May 2018

https://neurips.cc/media/Slides/nips/2019/westballa+b(11-15-50)-11-16-55-15712-superglue_a_st.pdf



https://neurips.cc/media/Slides/nips/2019/westballa+b(11-15-50)-11-16-55-15712-superglue_a_st.pdf

SuperGLUE

- New benchmark of 8 NLU tasks
- Also:
  - Additional diagnostics
  - Rules updates
  - Starter code
- Tasks were selected from an open call to the NLP community
  - Screen each proposed task to be easy for humans, hard for machines
  - Emphasized tasks with little training data
  - More diverse set of task formats, e.g. QA, coreference
- Released May 2019

https://neurips.cc/media/Slides/nips/2019/westballa+b(11-15-50)-11-16-55-15712-superglue_a_st.pdf

BoolQ

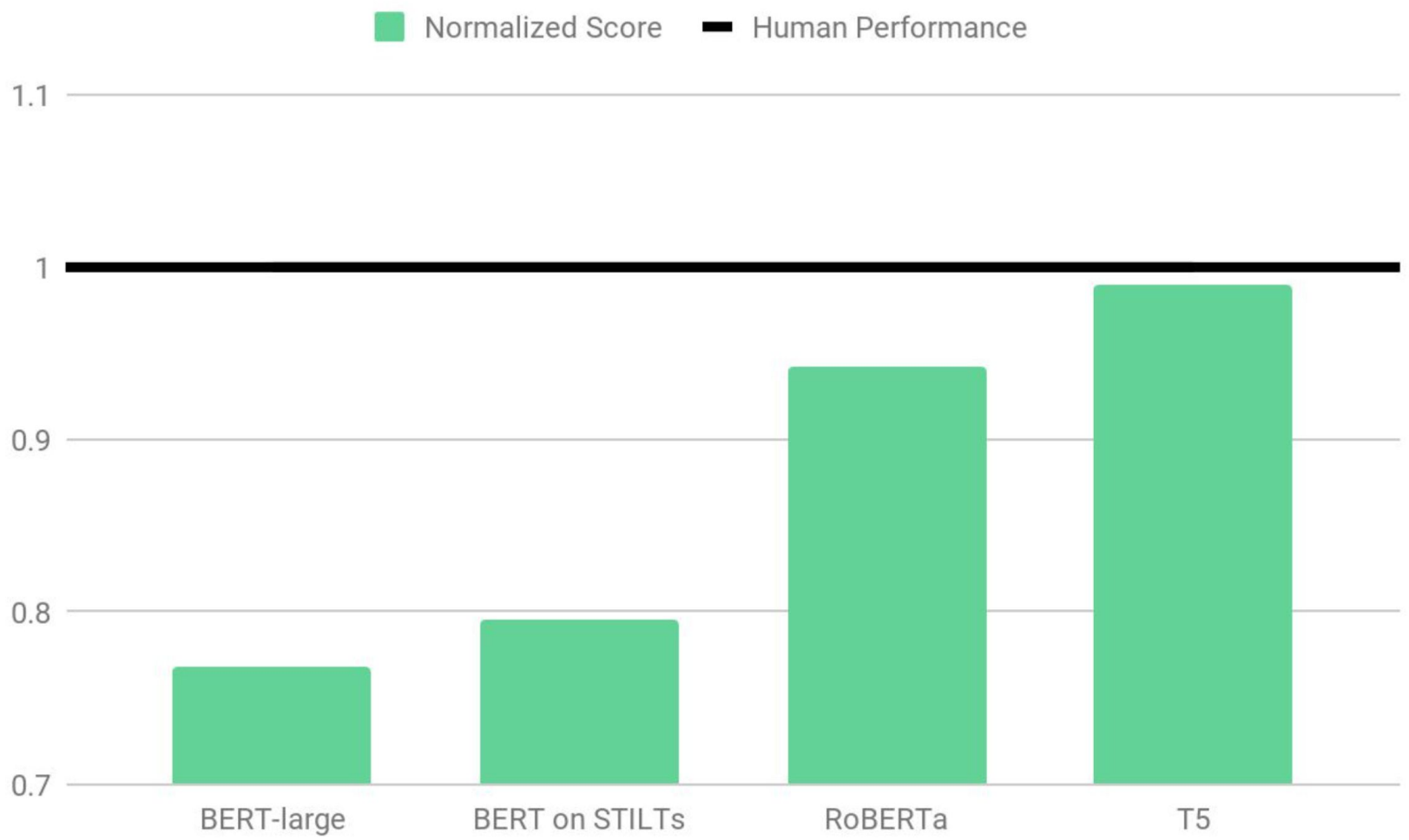
Passage: *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.*
Question: *is barq's root beer a pepsi product* **Answer:** `No`

COPA

Premise: *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?*
Alternative 1: *The sun was rising.* **Alternative 2:** *The grass was cut.*
Correct Alternative: `1`

MultiRC

Paragraph: *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*
Question: *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (`T`), *No* (`F`), *Yes* (`T`), *No, she didn't recover* (`F`), *Yes, she was at Susan's party* (`T`)



https://neurips.cc/media/Slides/nips/2019/westballa+b(11-15-50)-11-16-55-15712-superglue_a_st.pdf

CondConv: Conditionally Parameterized Convolutions for Efficient Inference

Brandon Yang*
Google Brain
bcyang@google.com

Gabriel Bender
Google Brain
gbender@google.com

Quoc V. Le
Google Brain
qvl@google.com

Jiquan Ngiam
Google Brain
jngiam@google.com

CondConv

Fundamental assumption:

Convolutional kernels should be shared for all examples in a dataset

Conditionally parameterized convolutions (CondConv)

each individual example is processed with different weights!

Enables to increase the size and capacity of a network

while maintaining efficient inference

CondConv

Fundamental assumption:

Convolutional kernels should be shared for all examples in a dataset

Conditionally parameterized convolutions (CondConv)

each individual example is processed with different weights!

Enables to increase the size and capacity of a network

while maintaining efficient inference

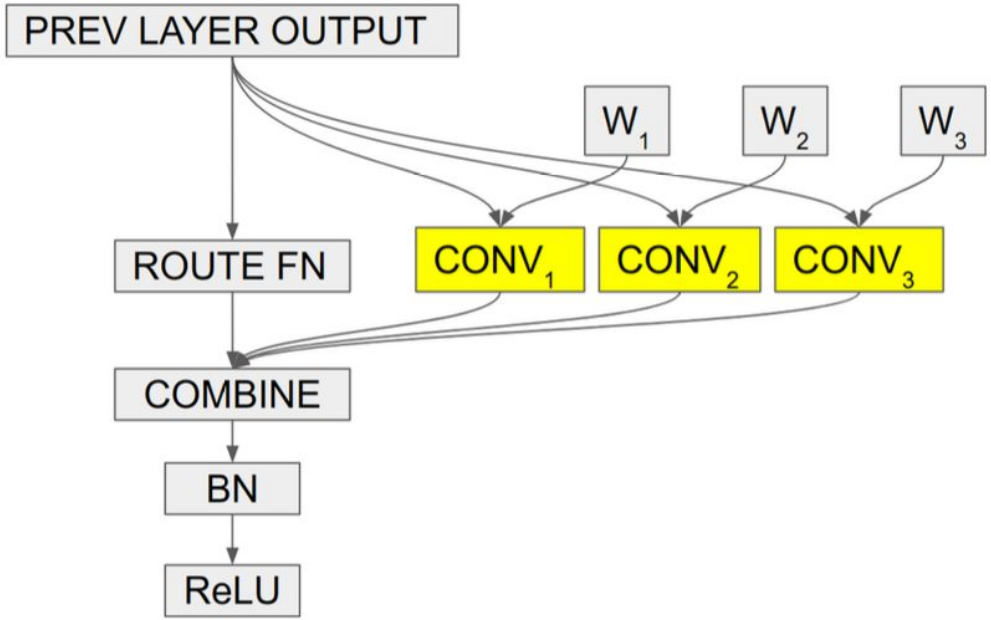
CondConv

Fundamental assumption:
    Convolutional kernels should be shared for all examples in a dataset

Conditionally parameterized convolutions (CondConv)
    each individual example is processed with different weights!

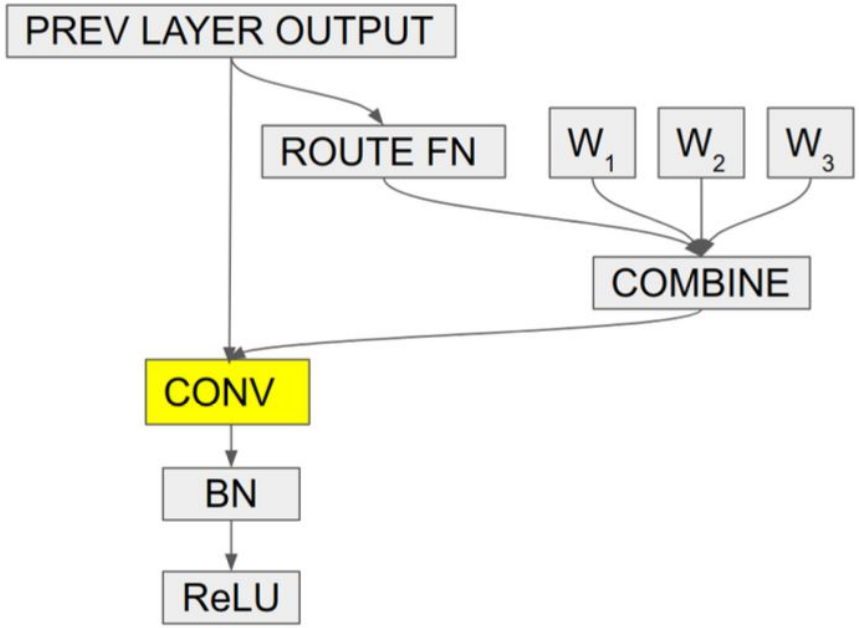
Enables to increase the size and capacity of a network
    while maintaining efficient inference

CondConv



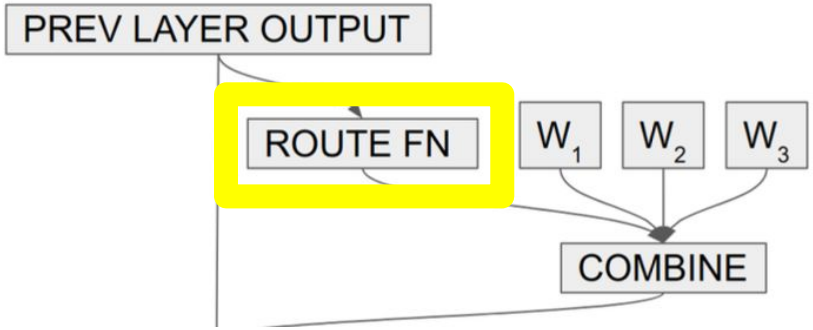
(b) Mixture of Experts: $\alpha_1(W_1 * x) + \ldots + \alpha_n(W_n * x)$

CondConv



(a) CondConv: $(\alpha_1 W_1 + \ldots + \alpha_n W_n) * x$

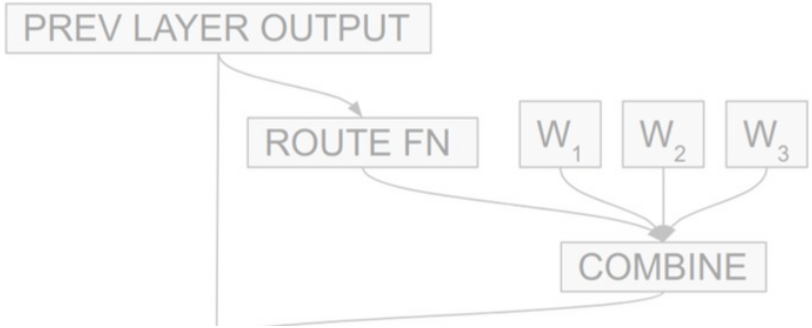
CondConv



We wish to design a per-example routing function that is computationally efficient, able to meaningfully differentiate between input examples, and is easily interpretable. We compute the example-dependent routing weights $\alpha_i = r_i(x)$ from the layer input in three steps: global average pooling, fully-connected layer, Sigmoid activation.

$$r(x) = \text{Sigmoid}(\text{GlobalAveragePool}(x)\ R)$$

CondConv

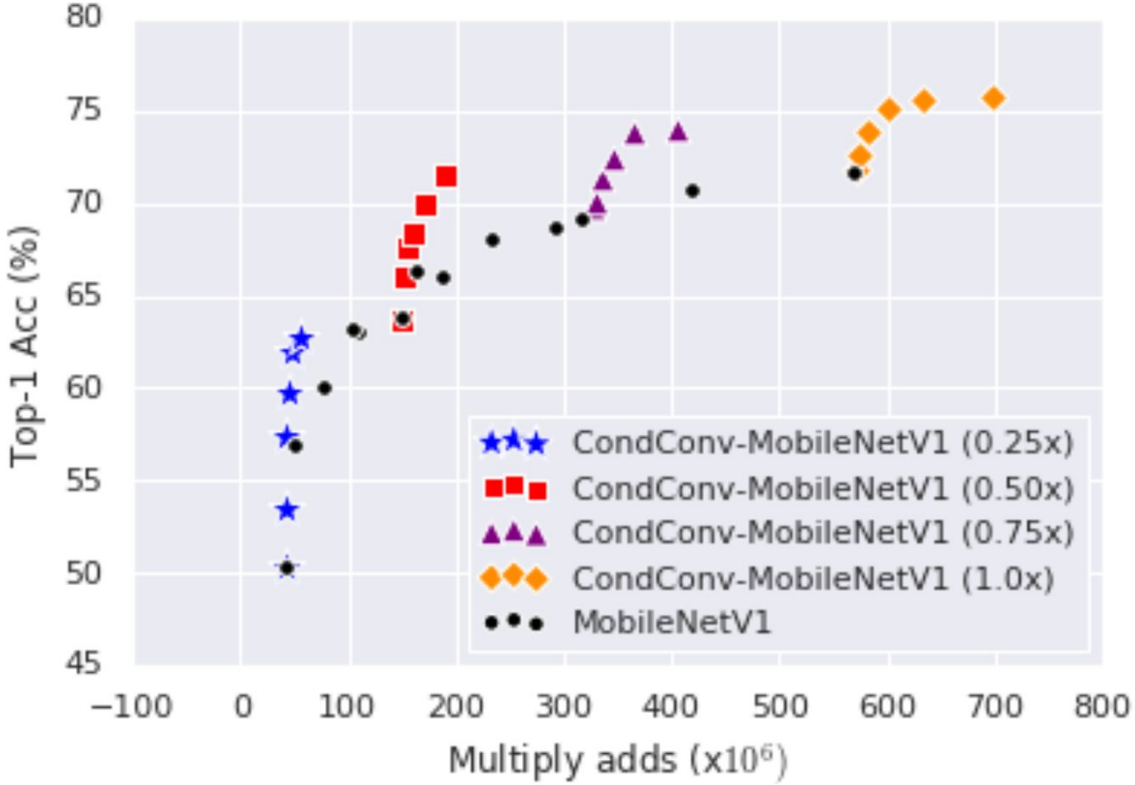


We wish to design a per-example routing function that is computationally efficient, able to meaningfully differentiate between input examples, and is easily interpretable. We compute the example-dependent routing weights $\alpha_i = r_i(x)$ from the layer input in three steps: global average pooling, fully-connected layer, Sigmoid activation.

$$r(x) = \text{Sigmoid}(\text{GlobalAveragePool}(x)\ R)$$

Finally, we experiment with the *Softmax* activation function to compute routing weights. The baseline's *Sigmoid* significantly outperforms *Softmax*, which suggests that multiple experts are often useful for a single example.

CondConv


80
75
70
65
60
55
50
45
Top-1 Acc (%)
−100
0
100
200
300
400
500
600
700
800
Multiply adds (x10^6)
CondConv-MobileNetV1 (0.25x)
CondConv-MobileNetV1 (0.50x)
CondConv-MobileNetV1 (0.75x)
CondConv-MobileNetV1 (1.0x)
MobileNetV1

This Looks Like *That*: Deep Learning for Interpretable Image Recognition

Chaofan Chen*
Duke University
cfchen@cs.duke.edu

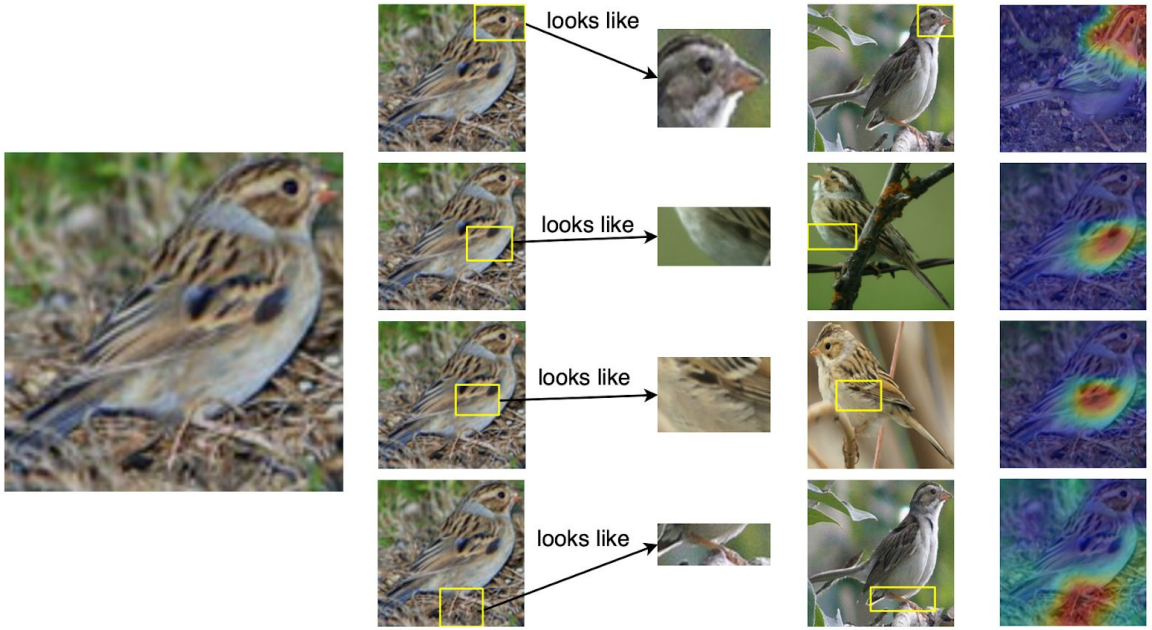
Oscar Li*
Duke University
oscarli@alumni.duke.edu

Chaofan Tao
Duke University
chaofan.tao@duke.edu

Alina Jade Barnett
Duke University
abarnett@cs.duke.edu

Jonathan Su
MIT Lincoln Laboratory†
su@ll.mit.edu

Cynthia Rudin
Duke University
cynthia@cs.duke.edu



Leftmost: a test image of a clay-colored sparrow

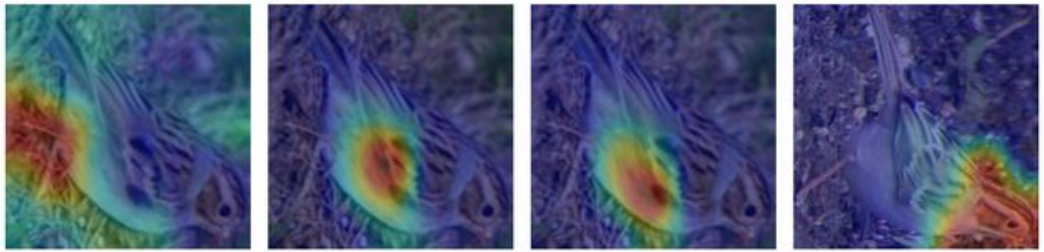
Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm

Third column: prototypical parts learned by our algorithm

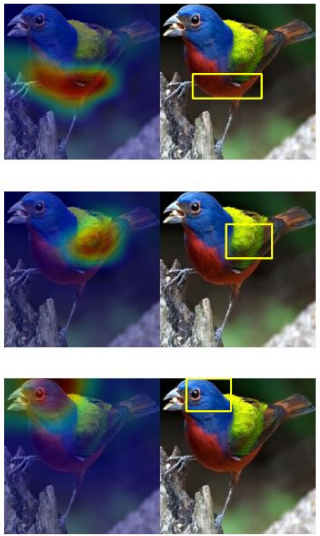
Fourth column: source images of the prototypical parts in the third column

Rightmost column: activation maps indicating how similar each prototypical part resembles part of the test bird

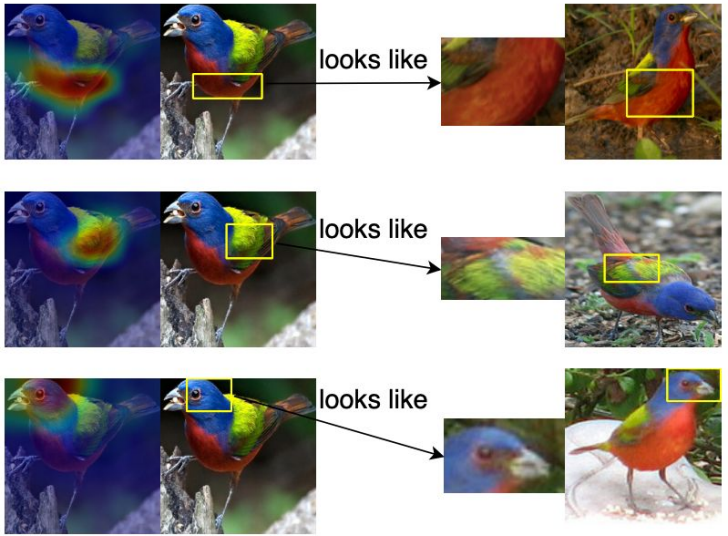
Our work relates to (but contrasts with) those that perform *posthoc* interpretability analysis for a trained convolutional neural network (CNN). In posthoc analysis, one interprets a trained CNN by fitting explanations to how it performs classification. Examples of posthoc analysis techniques include activation maximization [6, 13, 23, 43, 31, 37, 47], deconvolution [48], and saliency visualization [37, 41, 40, 35]. All of these posthoc visualization methods do not explain the reasoning process of how a network *actually* makes its decisions. In contrast, our network has a built-in case-based reasoning process, and the explanations generated by our network are actually used during classification and are not created posthoc.



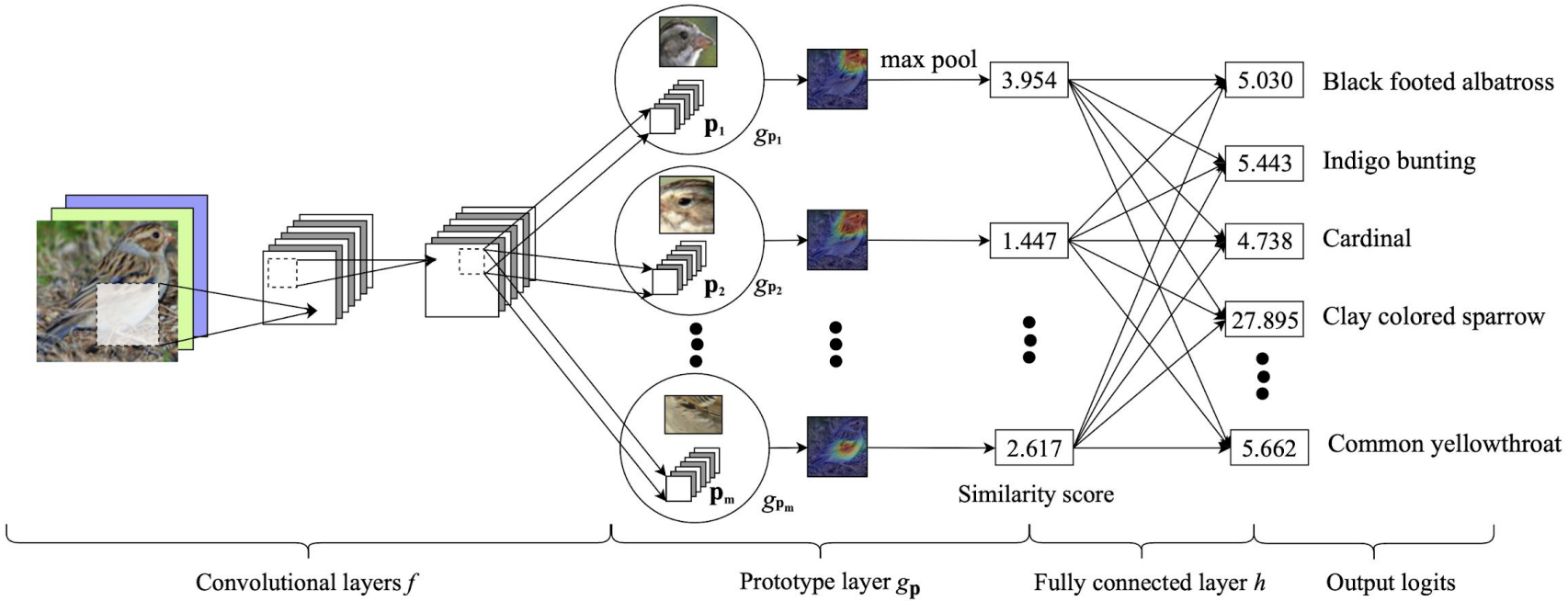
(a) Object attention
(class activation map)

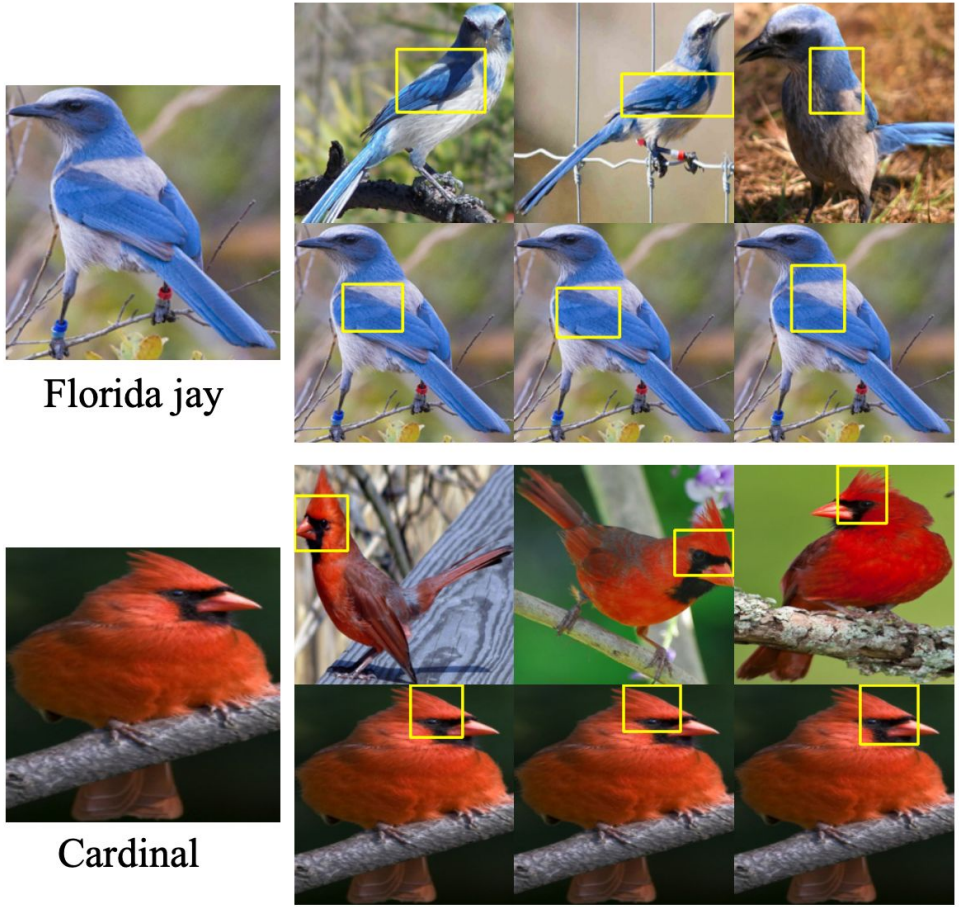


(b) Part attention
(attention-based models)

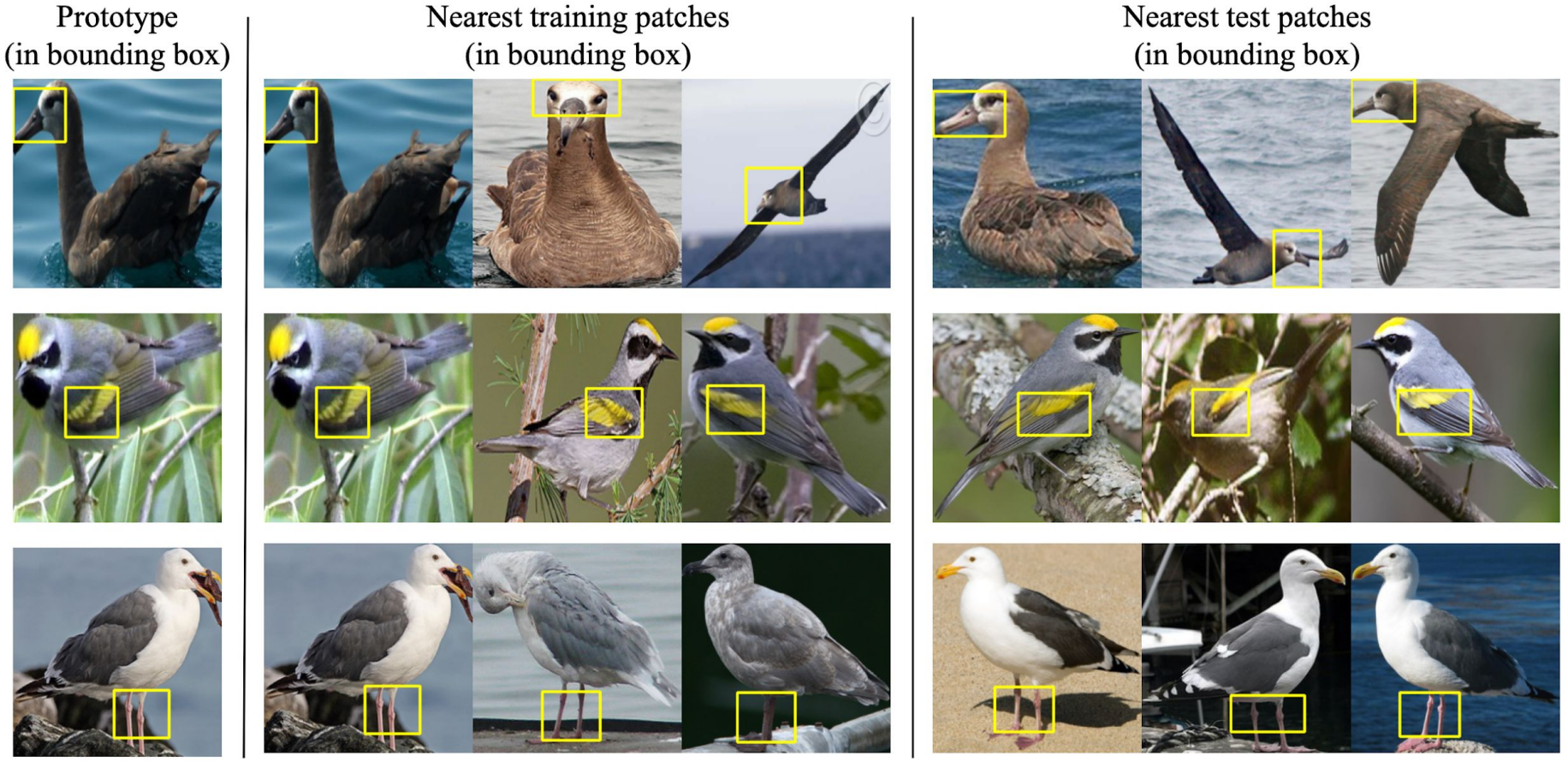


(c) Part attention + comparison with learned
prototypical parts (our model)


max pool
3.954
1.447
2.617
Similarity score
5.030
Black footed albatross
5.443
Indigo bunting
4.738
Cardinal
27.895
Clay colored sparrow
5.662
Common yellowthroat
$\mathbf{p}_1$
$\mathbf{p}_2$
$\mathbf{p}_m$
$g_{\mathbf{p}_1}$
$g_{\mathbf{p}_2}$
$g_{\mathbf{p}_m}$
Convolutional layers f
Prototype layer $g_{\mathbf{p}}$
Fully connected layer h
Output logits



(a) nearest prototypes of two test images



(b) nearest image patches to prototypes

Table 1: Top: Accuracy comparison on cropped bird images of CUB-200-2011
Bottom: Comparison of our model with other deep models

Base	ProtoPNet	Baseline	Base	ProtoPNet	Baseline
VGG16	76.1 ± 0.2	74.6 ± 0.2	VGG19	78.0 ± 0.2	75.1 ± 0.4
Res34	79.2 ± 0.1	82.3 ± 0.3	Res152	78.0 ± 0.3	81.5 ± 0.4
Dense121	80.2 ± 0.2	80.5 ± 0.1	Dense161	80.1 ± 0.3	82.2 ± 0.2


Overview
Tutorials
Papers
Workshops & Hardware

Black in AI (BAI) Affinity Workshop
Women in Machine Learning (WiML) Affinity Workshop
LatinX in AI (LAI) Affinity Workshop
New In Machine Learning
Queer in AI (QAI) Affinity Workshop
CiML 2019: Machine Learning Competitions for All
Solving inverse problems with deep networks: New architectures, theoretical foundations, and applications
Retrospectives: A Venue for Self-Reflection in ML Research
Biological and Artificial Reinforcement Learning
Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness, and Privacy
Bayesian Deep Learning
MLSys: Workshop on Systems for ML
Information Theory and Machine Learning
KR2ML - Knowledge Representation and Reasoning Meets Machine Learning
Machine Learning for Health (ML4H): What makes machine learning in medicine different?
Competition Track Day 1
Workshop on Human-Centric Machine Learning
Safety and Robustness in Decision-making
Visually Grounded Interaction and Language
Machine Learning for the Developing World (ML4D): Challenges and Risks
Minding the Gap: Between Fairness and Ethics
Shared Visual Representations in Human and Machine Intelligence
Graph Representation Learning
Beyond first order methods in machine learning systems
Perception as generative reasoning: structure, causality, probability
EMC2: Energy Efficient Machine Learning and Cognitive Computing (5th edition)
Optimal Transport for Machine Learning
AI for Humanitarian Assistance and Disaster Response
Workshop on Federated Learning for Data Privacy and Confidentiality
Learning Meaningful Representations of Life
Learning with Rich Experience: Integration of Learning Paradigms
Meta-Learning
Real Neurons & Hidden Units: future directions at the intersection of neuroscience and AI
Fair ML in Healthcare
Document Intelligence
Deep Reinforcement Learning
Privacy in Machine Learning (PriML)
Robot Learning: Control and Interaction in the Real World
Tackling Climate Change with ML
Medical Imaging meets NeurIPS
Learning Transferable Skills
Machine Learning with Guarantees
Machine Learning and the Physical Sciences
Emergent Communication: Towards Natural Language
Context and Compositionality in Biological and Artificial Neural Systems
Sets and Partitions
The third Conversational AI workshop – today's practice and tomorrow's potential
Program Transformations for ML
ML For Systems
NeurIPS Workshop on Machine Learning for Creativity and Design 3.0
Learning with Temporal Point Processes
Machine Learning for Autonomous Driving
Joint Workshop on AI for Social Good
“Do the right thing”: machine learning and causal inference for improved decision making
The Optimization Foundations of Reinforcement Learning
Science meets Engineering of Deep Learning
Competition Track Day 2
Bridging Game Theory and Deep Learning

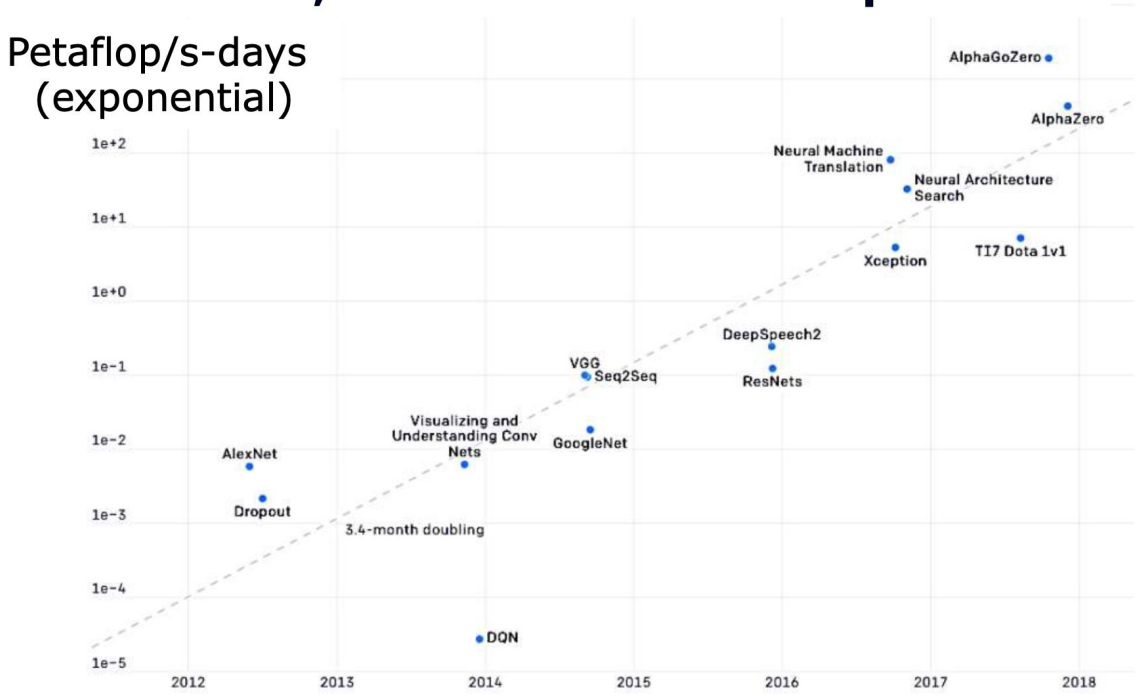

Overview
Tutorials
Papers
Workshops & Hardware

Mon Dec 9th 11:15 AM -- 01:15 PM @ West Exhibition Hall C + B3 Tutorial

Efficient Processing of Deep Neural Network: from Algorithms to Hardware Architectures

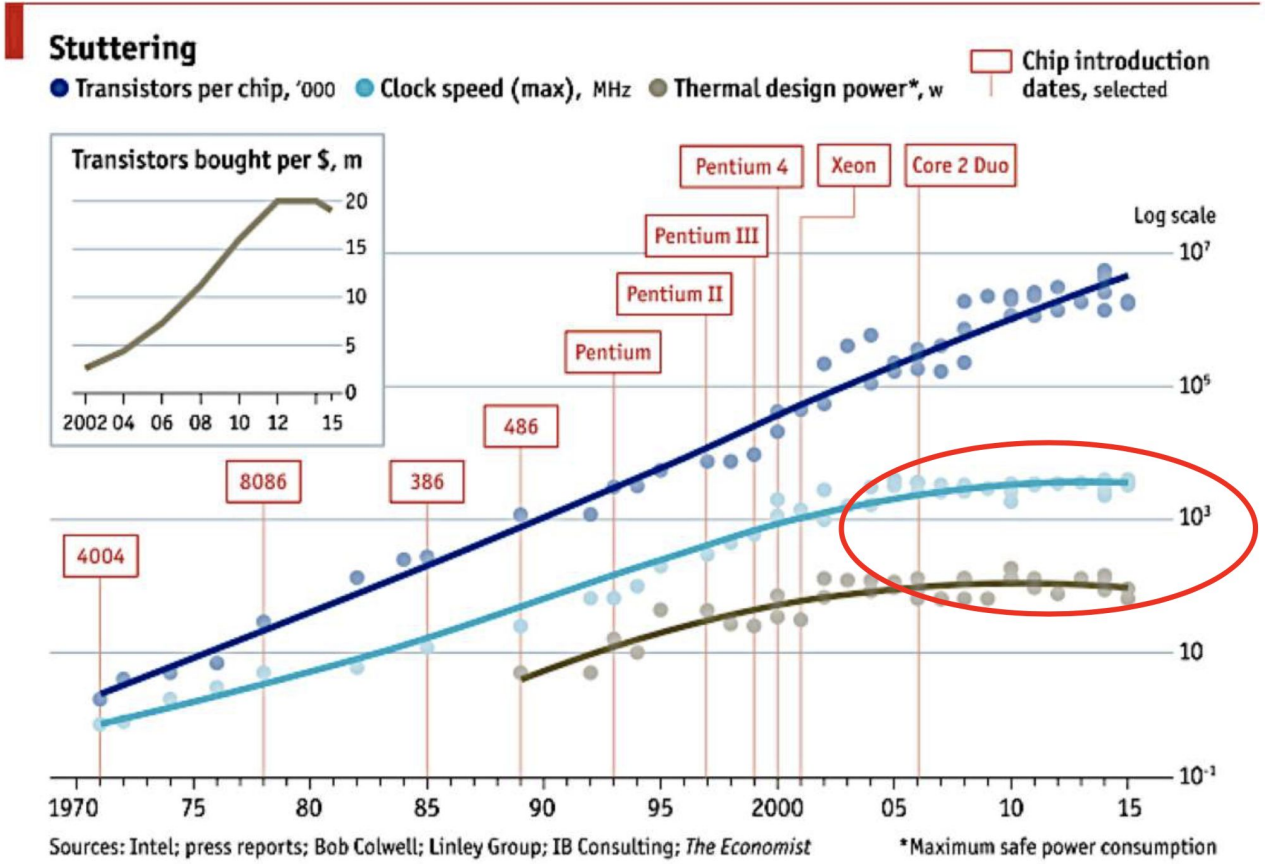
Vivienne Sze

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



http://eyeriss.mit.edu/2019_neurips_tutorial.pdf

Future Hardware



Future Hardware

Intel AI: *Intel® Nervana™ NNP: domain specific architectures for inference & training*

Graphcore: *Innovative approaches in training large scale language models*

Cerebras Systems: *Accelerating deep learning at wafer scale*

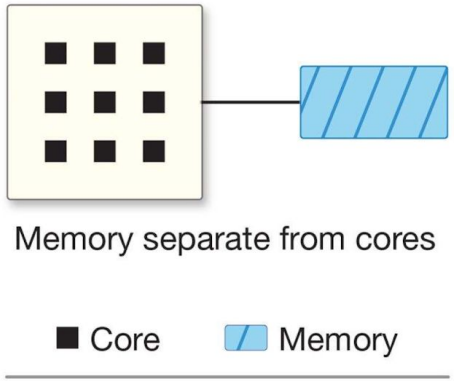
Habana Labs: *Hands-on Workshop: Implementing High-Performance AI Workloads with Habana AI Processors*

Traditional Memory Architectures *not* Optimized for DL

In neural networks, weights and activations are local, with low reuse

Traditional memory designs are punished

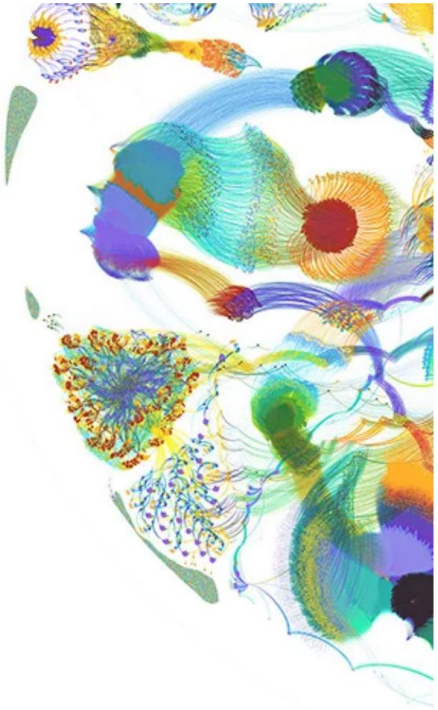
- Central shared memory is slow & far away
- Requires high data reuse (caching)
- Fundamental operation (matrix*vector) has low data reuse



cerebras

https://www.hotchips.org/hc31/HC31_1.13_Cerebras.SeanLie.v02.pdf

Graphcore



Graphcore

IPU-Tiles™

1216 IPU-Tiles™ each with an independent IPU-Core™ and tightly coupled In-Processor-Memory™

IPU-Core™

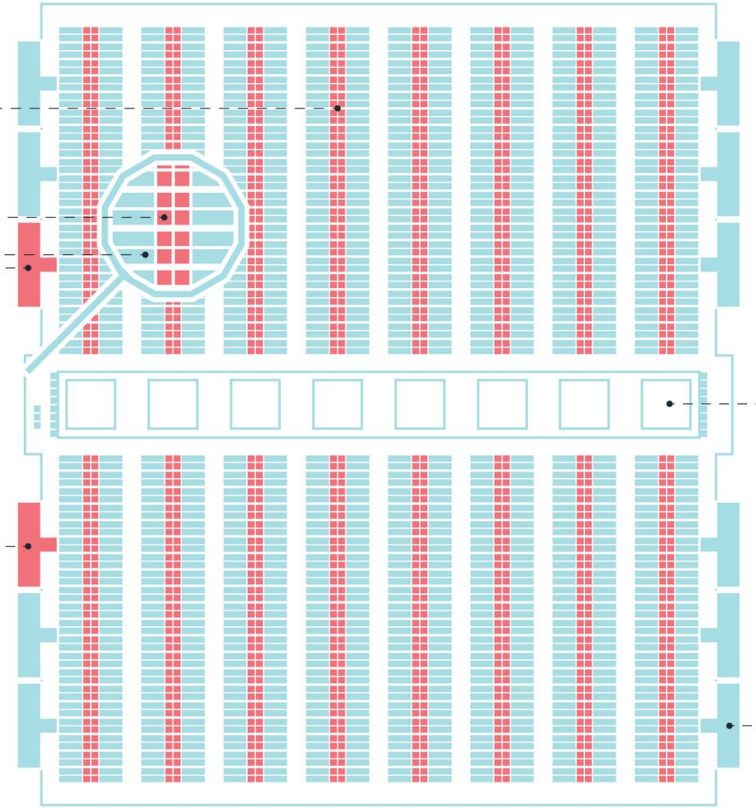
1216 IPU-Cores™ with 7296 programs executing in parallel

In-Processor-Memory™

300MB In-Processor-Memory™
45TB/s memory bandwidth
Whole model held on-chip

PCIe

PCIe Gen4 x16
64 GB/s bidirectional bandwidth to host

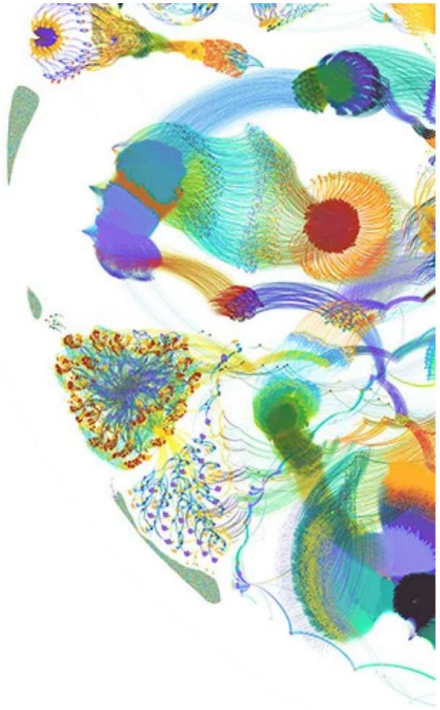


IPU-Exchange™

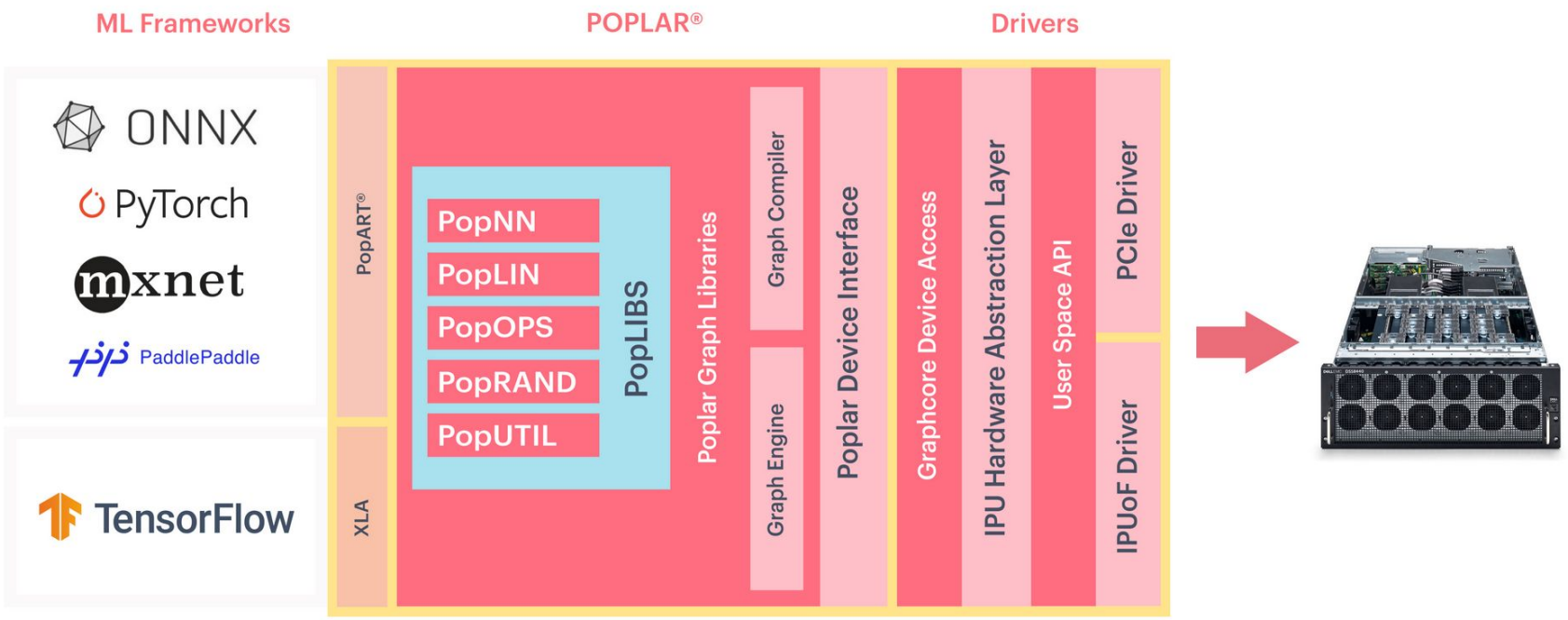
8 TB/s all to all IPU-Exchange™
Non-blocking, any communication pattern

IPU-Links™

80 IPU-Links, 320GB/s chip to chip bandwidth

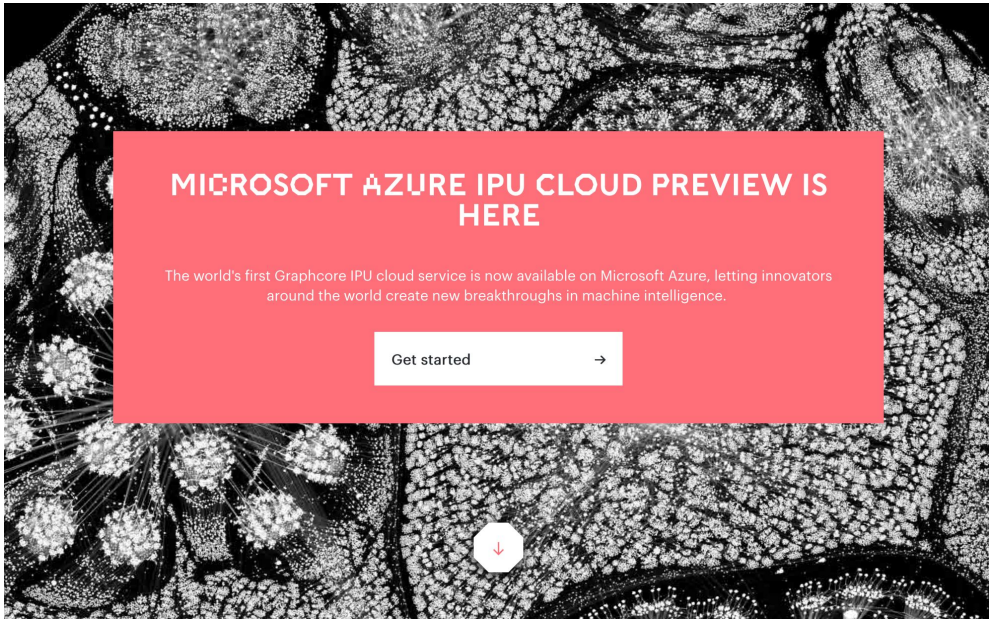


Graphcore

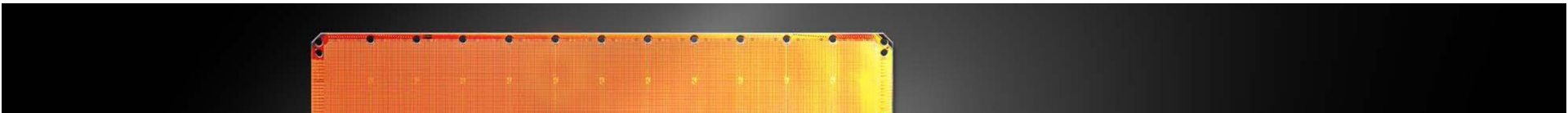

ML Frameworks
POPLAR®
Drivers
ONNX
PyTorch
mxnet
PaddlePaddle
TensorFlow
PopART®
XLA
PopNN
PopLIN
PopOPS
PopRAND
PopUTIL
PopLIBS
Poplar Graph Libraries
Graph Compiler
Graph Engine
Poplar Device Interface
Graphcore Device Access
IPU Hardware Abstraction Layer
User Space API
PCIe Driver
IPUoF Driver

Graphcore


DELL EMC DSS8440


MICROSOFT AZURE IPU CLOUD PREVIEW IS HERE
The world's first Graphcore IPU cloud service is now available on Microsoft Azure, letting innovators around the world create new breakthroughs in machine intelligence.
Get started

Cerebras



Hotchips 2019

4:45 PM	ML Training		Session Chair: Cliff Young	
4:45 PM		A Scalable unified architecture for Neural Network computing from Nano-level to high performance computing	Liao Heng	Huawei
5:15 PM		Deep Learning Training at Scale – Spring Crest Deep Learning Accelerator	Andrew Yang, Nitin Garegrat, Connie Miao & Karthik Vaidyanathan	Intel
5:45 PM		Wafer Scale Deep Learning	Sean Lie	Cerebras
6:15 PM		Habana Labs Approach to Scaling AI Training	Eitan Medina	Habana

https://www.hotchips.org/archives/2010s/hc31/

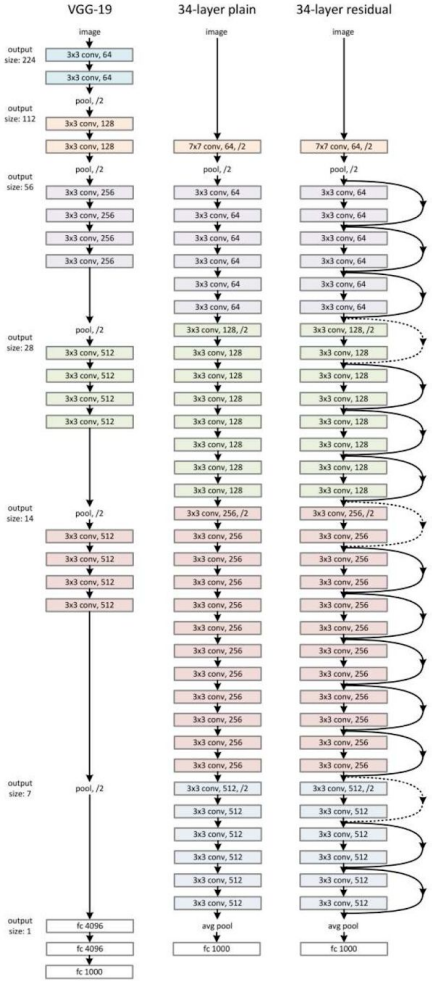
Deep Learning Training is Hard

Size:

- Billions-trillions of ops per sample
- Millions-billions of samples per training
- Peta-exa scale compute

Shape:

- **Fine-grained:** A lot of parallelism; presents opportunity to accelerate
- **Coarse-grained:** Inherently serial



cerebras

Legacy Technologies: Brute Force Parallelism

Fine-grained

- Dense vector processors (e.g. GPUs)
- Limited when compute not large uniform blocks

Coarse-grained

- Scale out clustering (e.g. PCIe, Ethernet, IB, NVLink)
- Run multiple instances of the same model (data parallel)
- Limited by inherent serial nature of problem

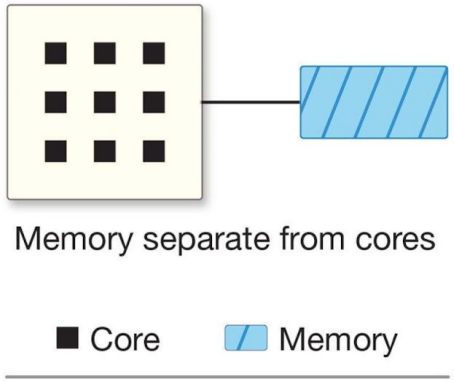
Result: scaling is limited and costly

Traditional Memory Architectures *not* Optimized for DL

In neural networks, weights and activations are local, with low reuse

Traditional memory designs are punished

- Central shared memory is slow & far away
- Requires high data reuse (caching)
- Fundamental operation (matrix*vector) has low data reuse

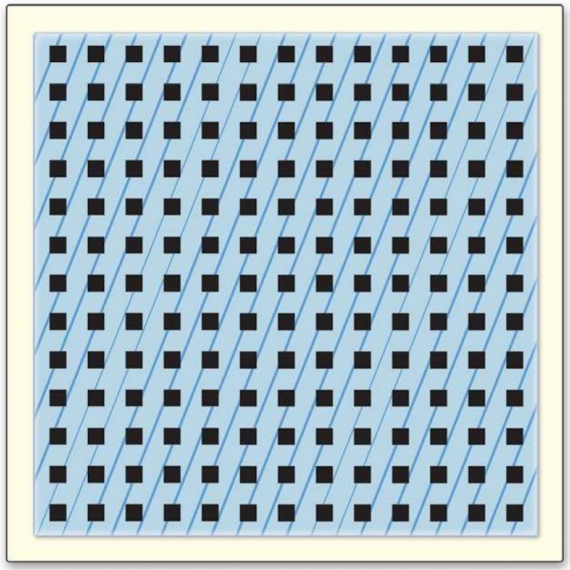


A Memory Architecture that *is* Optimized for DL

In neural networks, weights and activations are local, with low data reuse

The right answer is distributed, high performance, on-chip memory

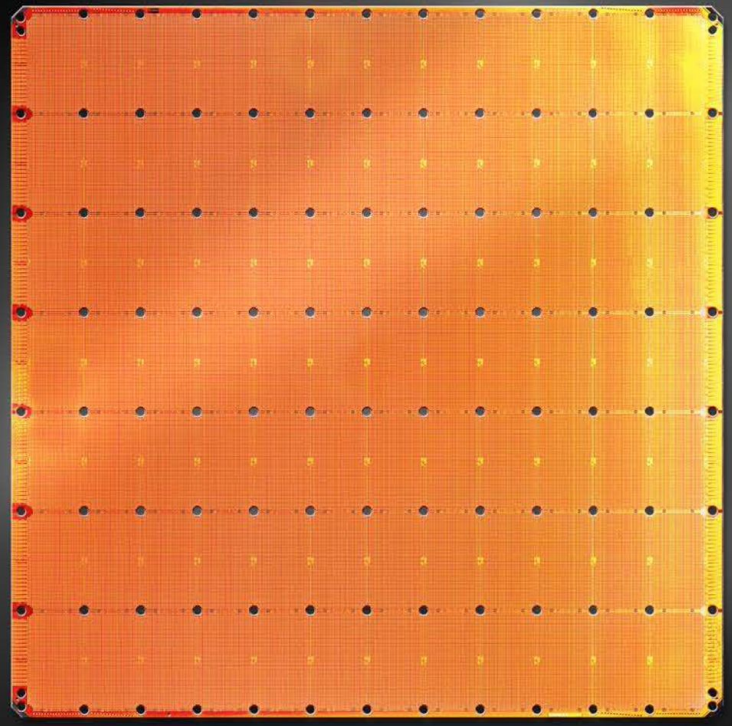
- All memory is fully distributed along with compute datapaths
- Datapath has full performance from memory



Memory uniformly distributed across cores

cerebras

Cerebras Wafer Scale Engine

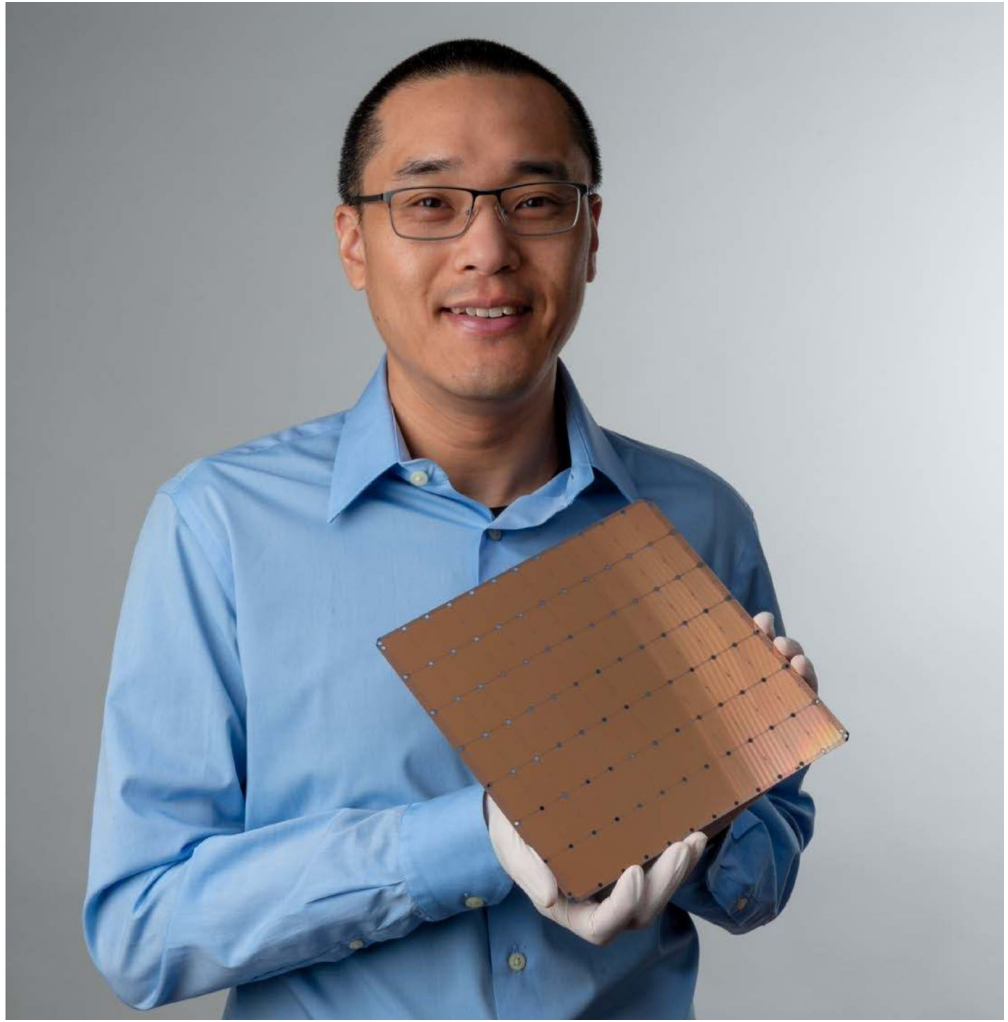


Cerebras WSE
1.2 Trillion Transistors
46,225 mm² Silicon

Largest GPU
21.1 Billion Transistors
815 mm² Silicon

Largest Chip Ever Built

- 46,225 mm^2 silicon
- 1.2 trillion transistors
- 400,000 AI optimized cores
- 18 Gigabytes of On-chip Memory
- 9 PByte/s memory bandwidth
- 100 Pbit/s fabric bandwidth
- TSMC 16nm process

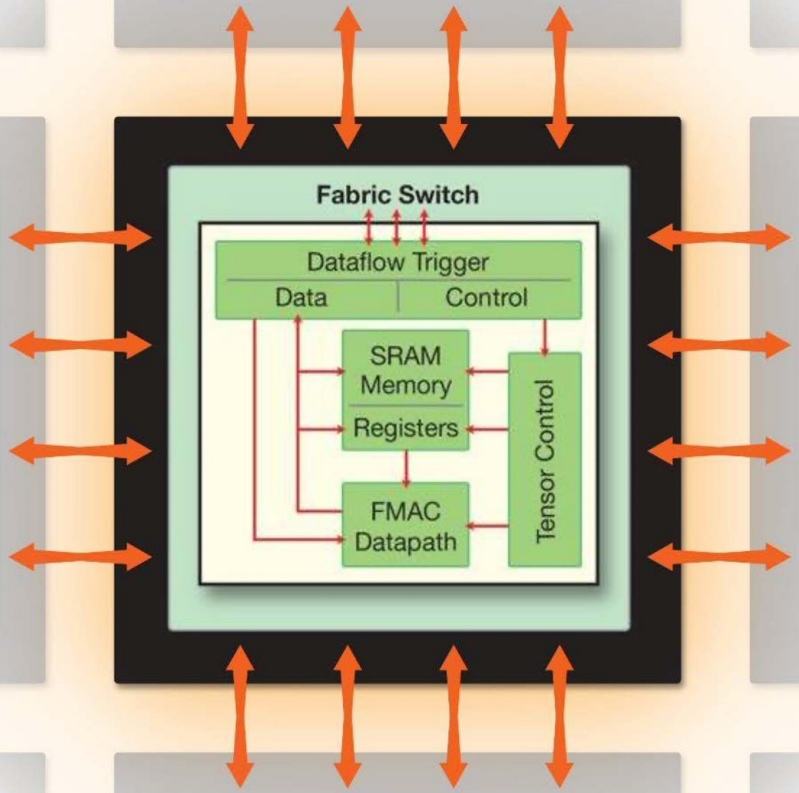


cerebras

Flexible Cores Optimized for Tensor Operations

Key to supporting rapidly evolving NN architectures

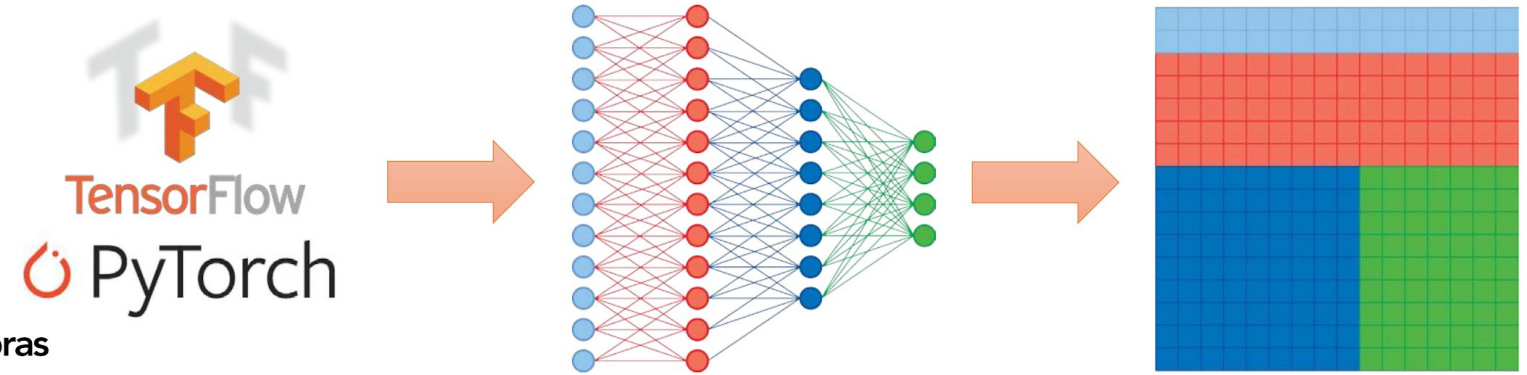
- Fully programmable compute core
- Full array of general instructions with ML extensions
- Flexible **general ops** for control processing
  - e.g. arithmetic, logical, load/store, branch
- Optimized **tensor ops** for data processing
  - Tensors as first class operands
  - e.g. `fmac [z] = [z], [w], a`
    `3D 3D 2D scalar`



cerebras

Programming the Wafer Scale Engine

- Neural network models expressed in common ML frameworks
- Cerebras interface to framework extracts the neural network
- Performs placement and routing to map neural network layers to fabric
- The entire wafer operates on the single neural network

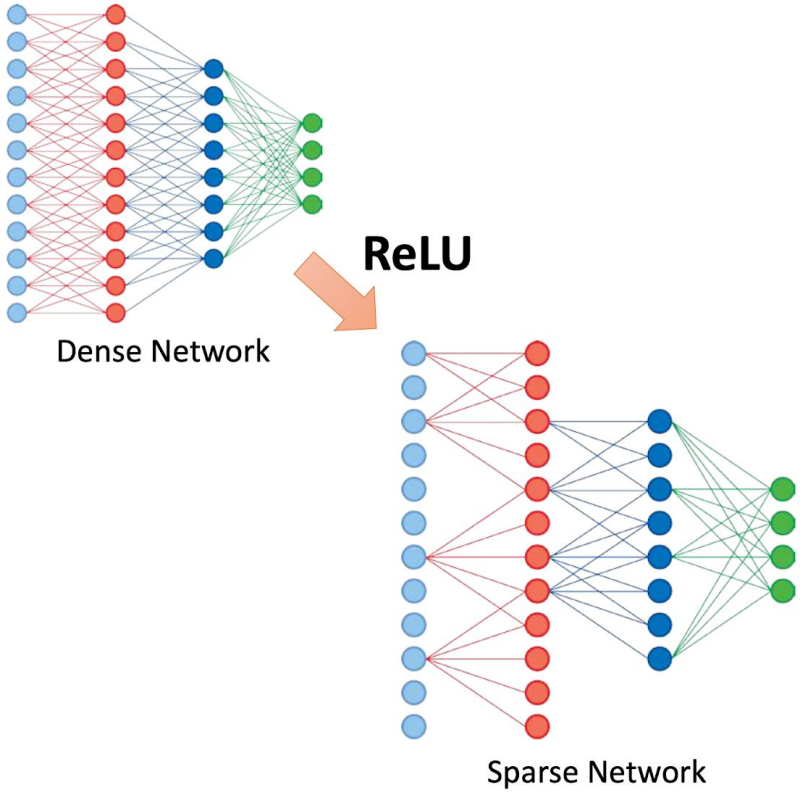


cerebras

Sparse Compute Engine for Neural Networks

NN operations like nonlinearities naturally create fine-grained sparsity

- Native, sparse processing enables higher efficiency and performance
- Dataflow scheduling in hardware
  - Triggered by data
  - Filters out sparse zero data
  - Skips unnecessary processing
- Fine-grained execution datapaths
  - Small cores with independent instructions
  - Maximizes utilization
  - Efficiently processes dynamic, non-uniform work



cerebras

Cluster-scale Deep Learning compute in a single system

15 Rack Units

Fits in a standard datacenter rack

1.2 Terabits/sec

System IO over 12x standard 100 GbE

20 kW

Maximum power draw

Explore our product


It’s working,
running customer workloads.
Stay tuned…

NeurIPS 2019 recap

René Donner, contextflow