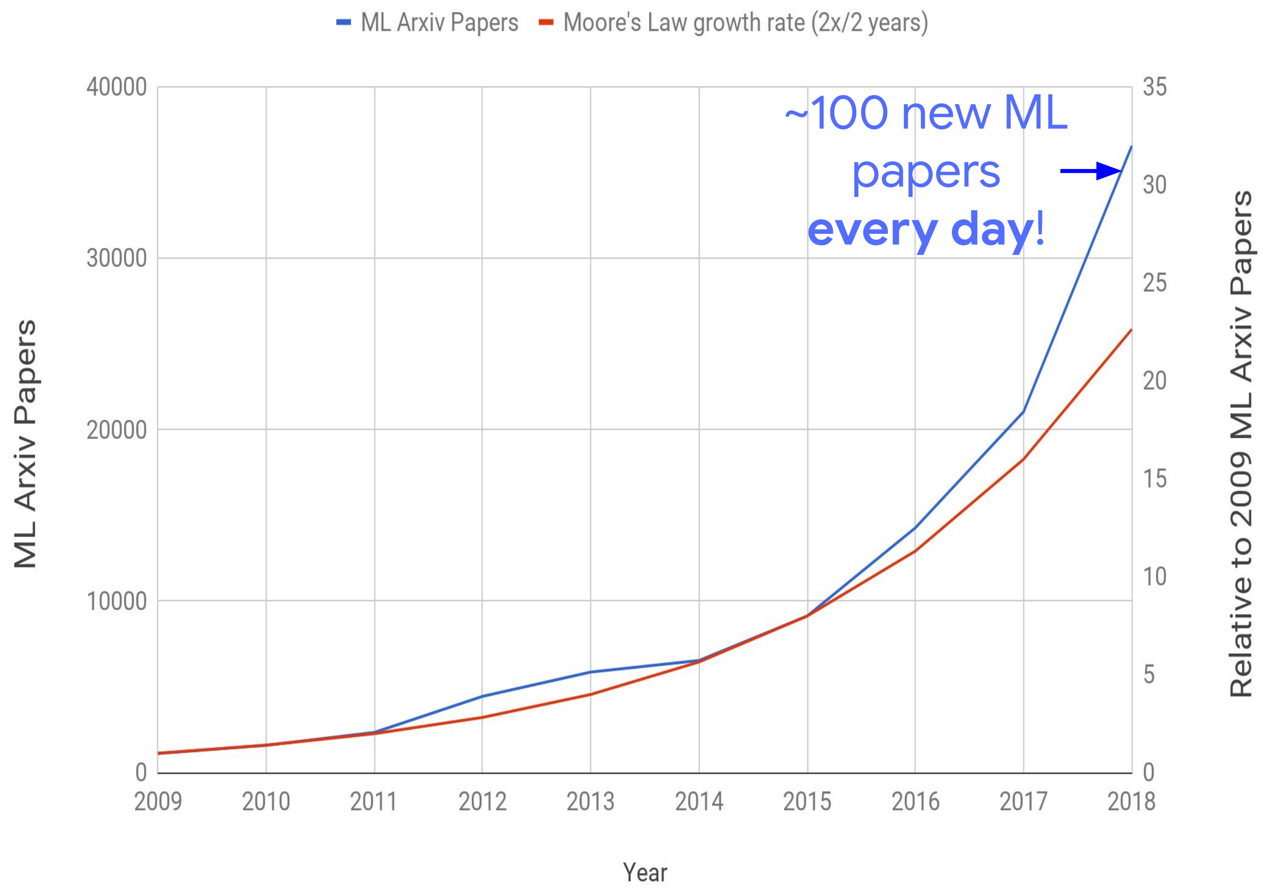
Google AI

Deep Learning to Solve Challenging Problems

Jeff Dean
Google Research
@JeffDean
ai.google/research/people/jeff

Presenting the work of **many** people at Google

Machine Learning Arxiv Papers per Year



Deep Learning

Modern Reincarnation of Artificial Neural Networks

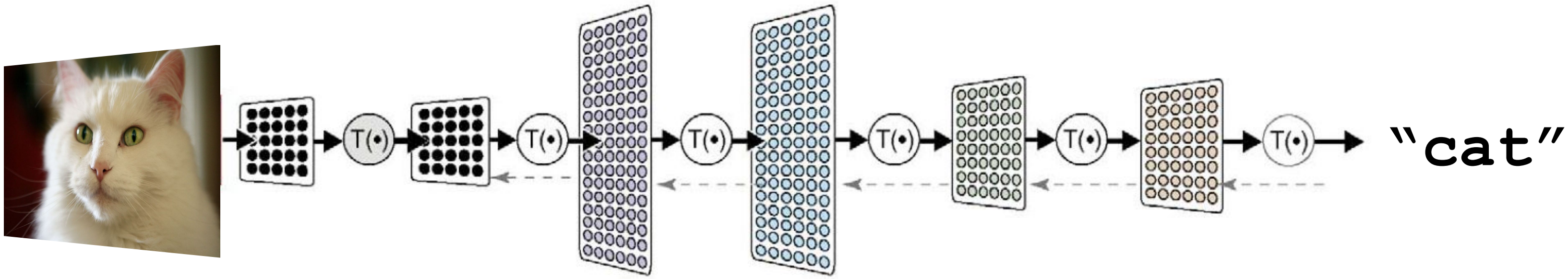
Collection of simple trainable mathematical units, organized in layers, that work together to solve complicated tasks

What's New

new network architectures,
new training math, **scale**

Key Benefit

Learns features from raw, heterogeneous, noisy data
No explicit feature engineering required

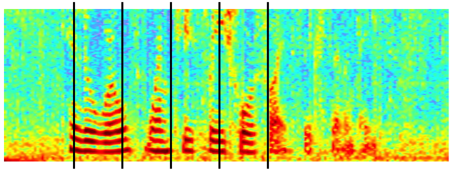


Google

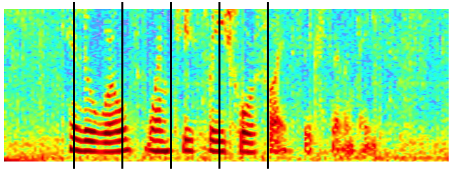
Functions a Deep Neural Network Can Learn

	input	output
Pixels:		`"leopard"`

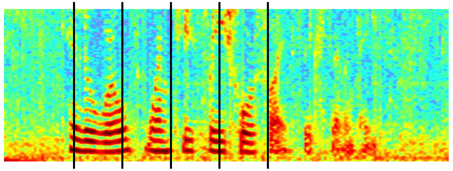
Functions a Deep Neural Network Can Learn

	input	output
Pixels:		"leopard"
Audio:		"How cold is it outside?"

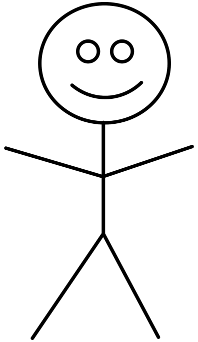
Functions a Deep Neural Network Can Learn

input		output
Pixels:		“leopard”
Audio:		“How cold is it outside?”
“Hello, how are you?”		“Bonjour, comment allez-vous?”

Functions a Deep Neural Network Can Learn

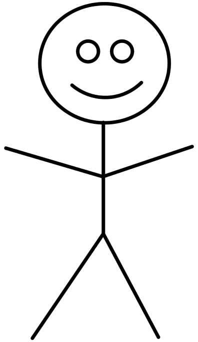
input		output
Pixels:		"leopard"
Audio:		"How cold is it outside?"
"Hello, how are you?"		"Bonjour, comment allez-vous?"
Pixels:		"A cheetah lying on top of a car"

humans



5% errors

humans



5% errors

2008: U.S. National Academy of Engineering publishes

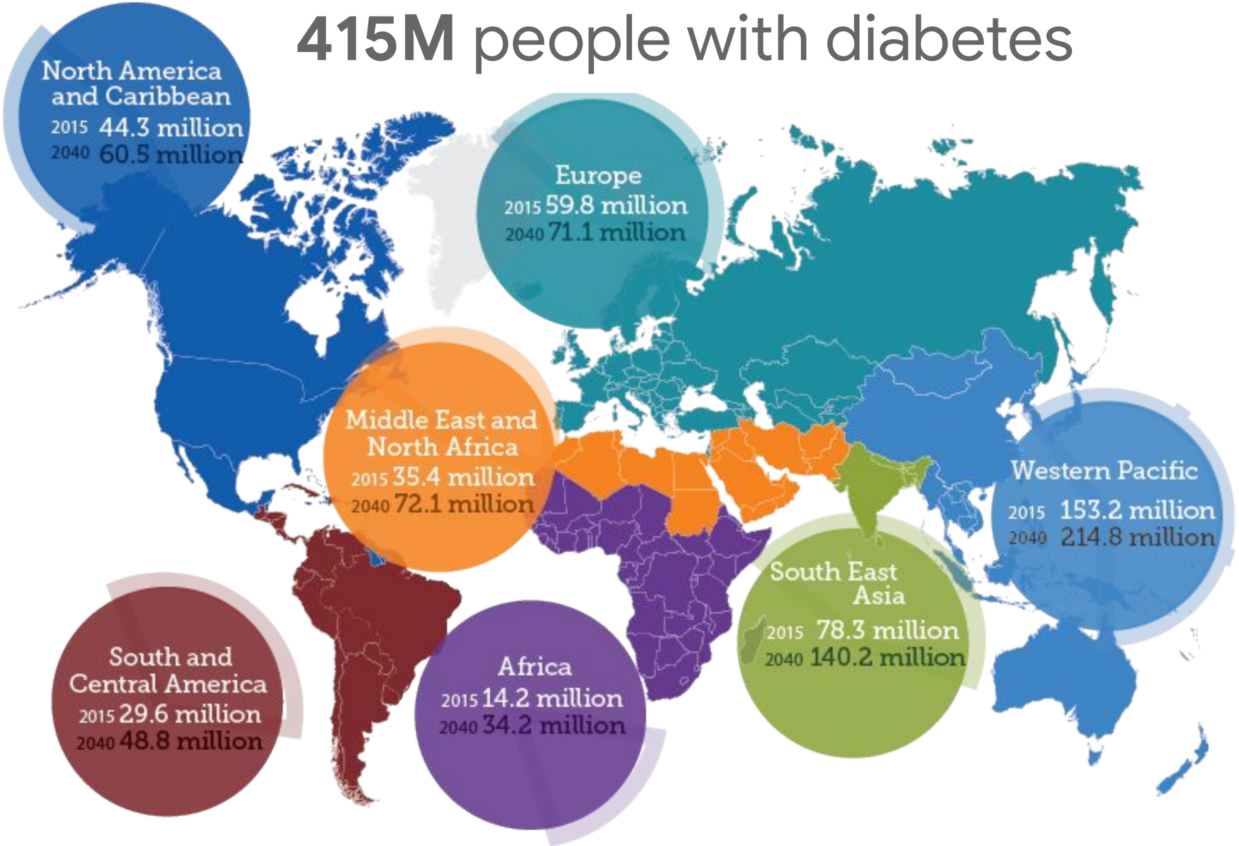
Grand Engineering Challenges for 21st Century

- Make solar energy affordable
- Provide energy from fusion
- Develop carbon sequestration methods
- Manage the nitrogen cycle
- Provide access to clean water
- Restore & improve urban infrastructure
- **Advance health informatics**
- Engineer better medicines
- Reverse-engineer the brain
- Prevent nuclear terror
- Secure cyberspace
- Enhance virtual reality
- Advance personalized learning
- **Engineer the tools for scientific discovery**

www.engineeringchallenges.org/challenges.aspx

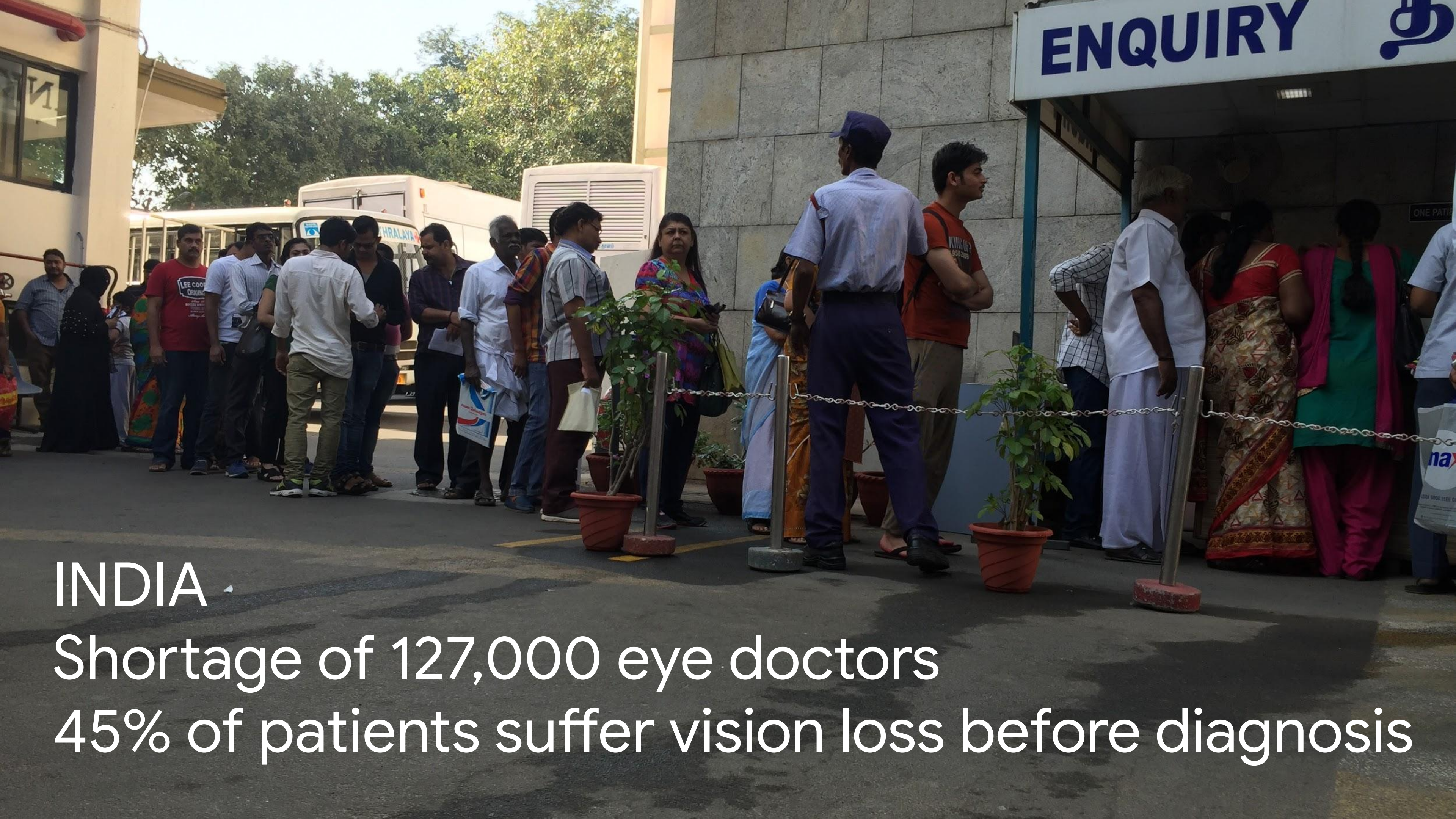
Advance health informatics

Diabetic retinopathy: fastest growing cause of blindness

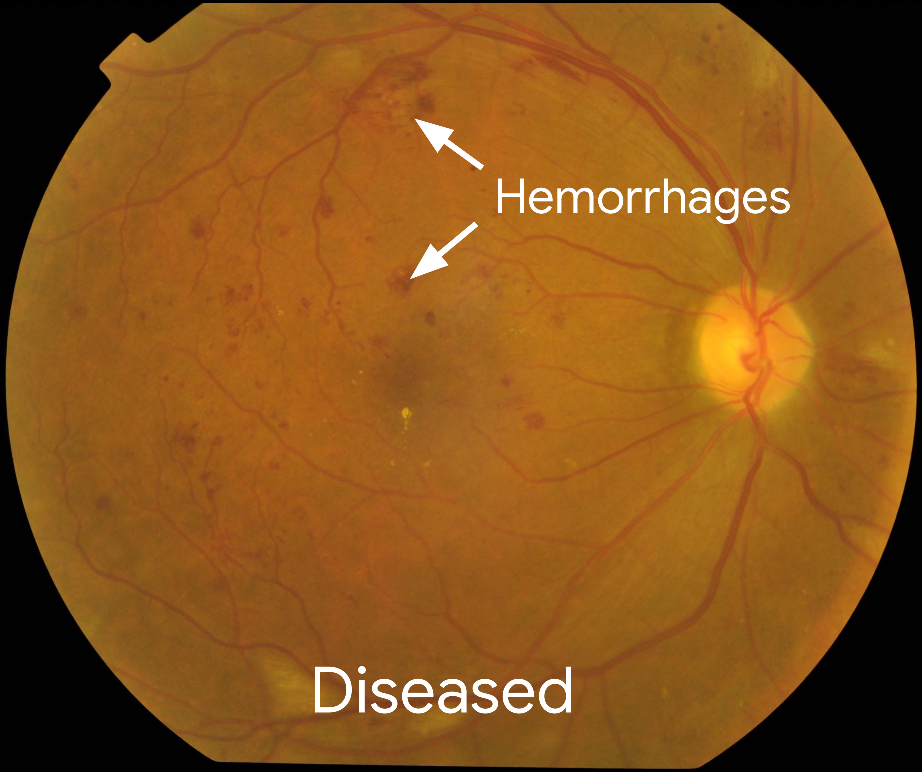


Regular screening is key to preventing blindness

=

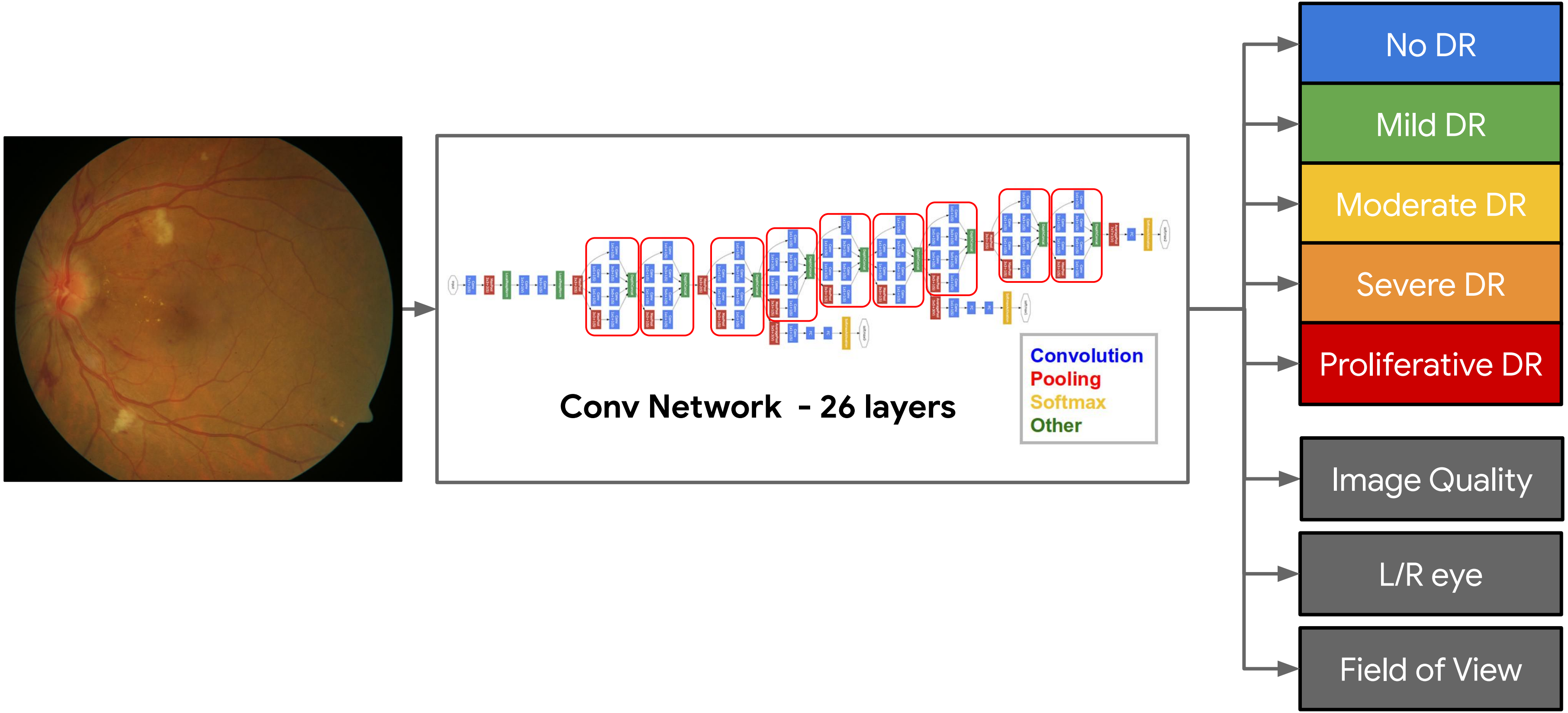

ENQUIRY
INDIA
Shortage of 127,000 eye doctors
45% of patients suffer vision loss before diagnosis

How DR is Diagnosed: Retinal Fundus Images



No DR	Mild DR	Moderate DR	Severe DR	Proliferative DR

Adapt deep neural network to read fundus images

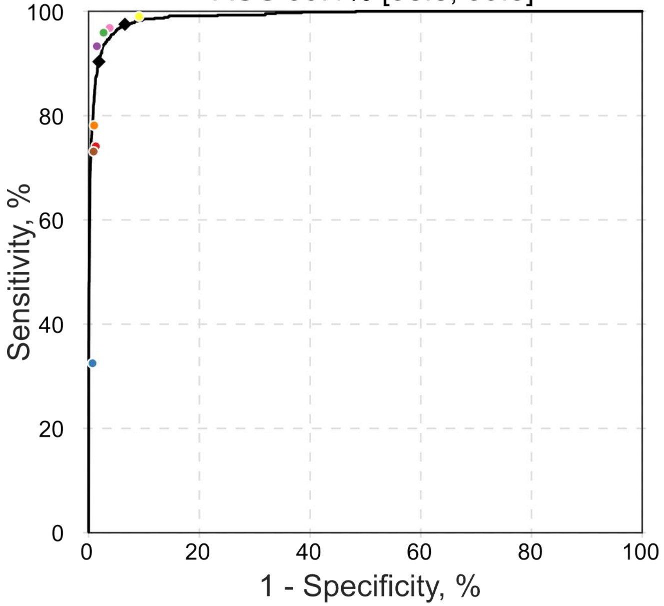


JAMA The Journal of the American Medical Association

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

n = 9,963 images
AUC 99.1% [98.8, 99.3]



F-score

0.95	0.91
Algorithm	Ophthalmologist (median)

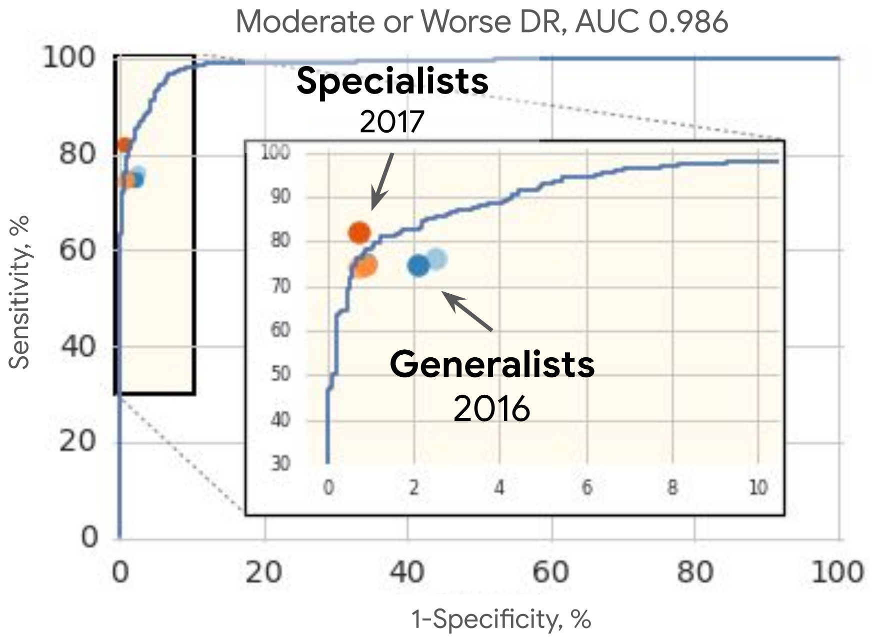
"The study by Gulshan and colleagues **truly represents the brave new world in medicine.**"

Dr. Andrew Beam, Dr. Isaac Kohane
Harvard Medical School

"Google just published this paper in JAMA (impact factor 37) [...] **It actually lives up to the hype.**"

Dr. Luke Oakden-Rayner
University of Adelaide

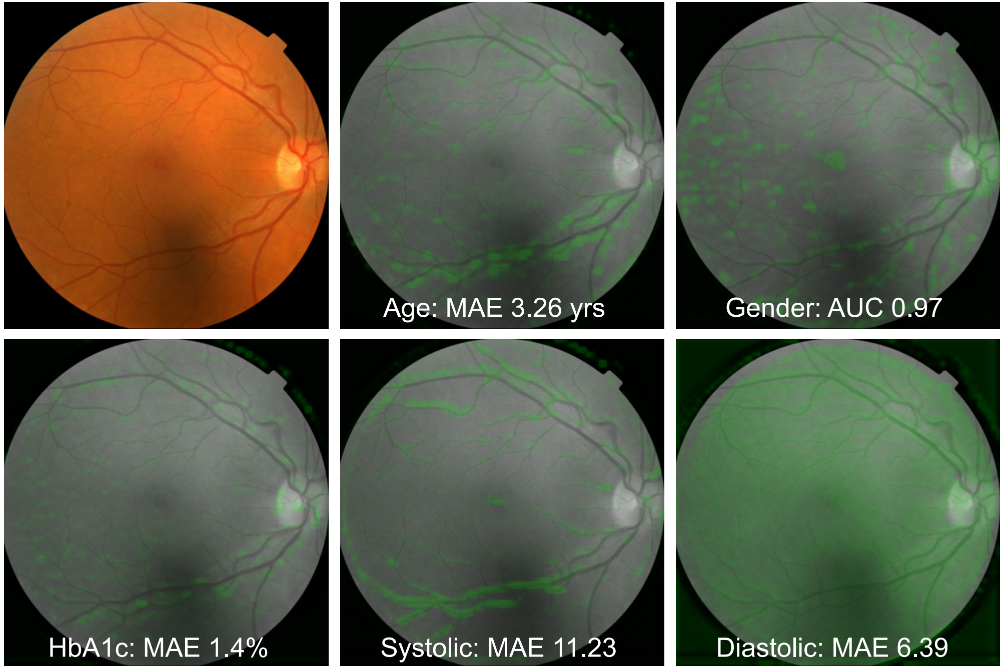
2016 - On Par with General Ophthalmologists
2017 - On Par with Retinal Specialist Ophthalmologists



	Weighted Kappa
Ophthalmologists Individual	0.80-0.84
Algorithm	0.84
Retinal Specialists Individual	0.82-0.91

Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. J. Krause, *et al.*, *Ophthalmology*, doi.org/10.1016/j.ophtha.2018.01.034

Completely new, novel scientific discoveries



Predicting things that doctors can't predict from imaging

—

Potential as a new biomarker

Preliminary 5-yr MACE AUC: 0.7

—

Can we predict cardiovascular risk? If so, this is a very nice non-invasive way of doing so

Can we also predict treatment response?

R. Poplin, A. Varadarajan *et al.* Predicting Cardiovascular Risk Factors from Retinal Fundus Photographs using Deep Learning. *Nature Biomedical Engineering, 2018.*

Pathology

Detecting Cancer Metastases on Gigapixel Pathology Images

Yun Liu[1*], Krishna Gadepalli[1], Mohammad Norouzi[1], George E. Dahl[1], Timo Kohlberger[1], Aleksey Boyko[1], Subhashini Venugopalan[2**], Aleksei Timofeev[2], Philip Q. Nelson[2], Greg S. Corrado[1], Jason D. Hipp[3] Lily Peng[1], and Martin C. Stumpe[1]

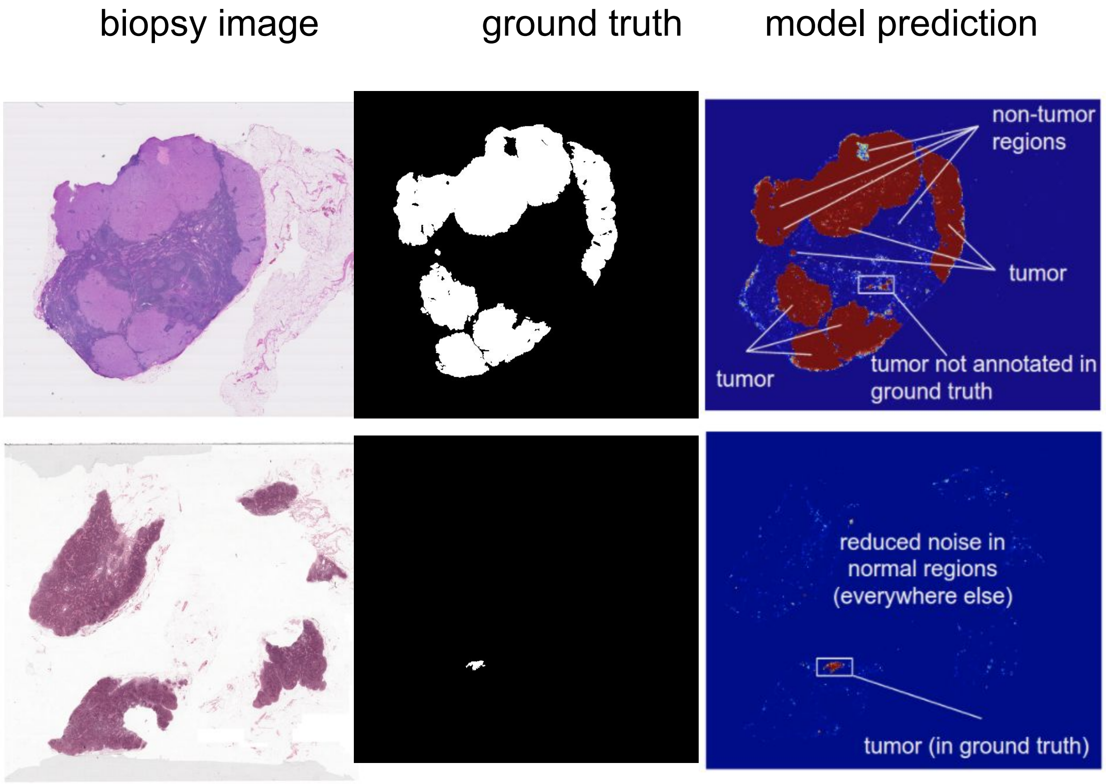
{liuyun,mnorouzi,gdahl,lhpeng,mstumpe}@google.com

[1]Google Brain, [2]Google Inc, [3]Verily Life Sciences,
Mountain View, CA, USA

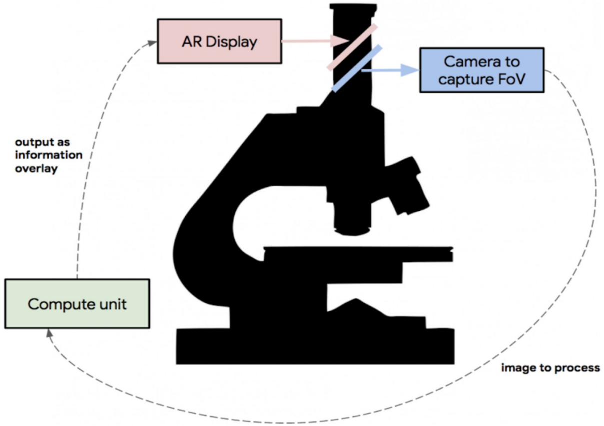
Tumor localization score (FROC):

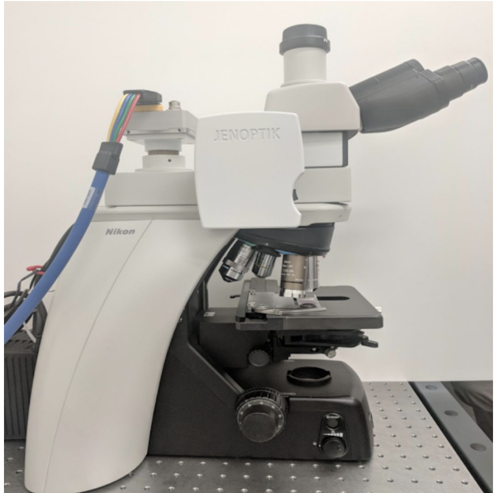
model:	**0.89**
pathologist:	0.73

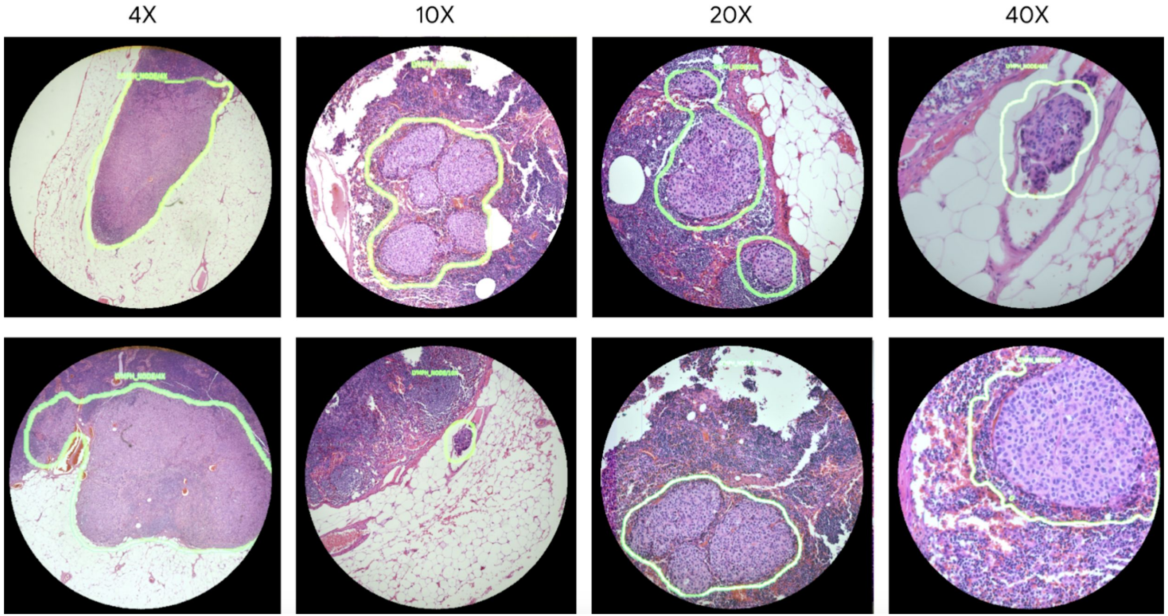
arxiv.org/abs/1703.02442



Augmented Reality Microscope







research.googleblog.com/2018/04/an-augmented-reality-microscope.html

Predictive tasks for healthcare

Given a patient's electronic medical record data, **can we predict the future**?

Deep learning methods for sequential prediction are becoming extremely good

e.g. recent improvements in Google Translation

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Published in NIPS, Dec. 2014, https://arxiv.org/abs/1409.3215

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Published in NIPS, Dec. 2014, https://arxiv.org/abs/1409.3215

GMail Smart Reply
Now ~12% of all mobile responses

Smart Reply: Automated Response Suggestion for Email,
Kannan *et al.*, KDD 2016:
https://arxiv.org/abs/1606.04870

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Sep. 2016, https://arxiv.org/abs/1609.08144

...

Predictive tasks for healthcare

Given a large corpus of training data of de-identified medical records, can we predict interesting aspects of the future for a patient not in the training set?

- *will patient be readmitted to hospital in next N days?*
- *what is the likely length of hospital stay for patient checking in?*
- *what are the most likely diagnoses for the patient right now?*
- *what medications should a doctor consider prescribing?*
- *what tests should be considered for this patient?*
- *which patients are at highest risk for X in next month?*

and why*?*

Collaborating with several healthcare organizations, including UCSF, Stanford, and Univ. of Chicago.

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

THE UNIVERSITY OF CHICAGO MEDICINE

UCSF University of California San Francisco

Stanford HEALTH CARE

Google

Mortality Risk Prediction Accuracy

https://arxiv.org/abs/1801.07860 and https://www.nature.com/articles/s41746-018-0029-1

Many Advances Depend on Being Able to Understand Text

Many Advances Depend on Being Able to Understand Text

Recent Encouraging Improvements in Language Understanding

2017: Transformer Model

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Attention Is All You Need,
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin,
June, 2017, [https://arxiv.org/abs/1706.03762,](https://arxiv.org/abs/1706.03762) appeared in NeurIPS, Dec. 2017

2017: Transformer Model

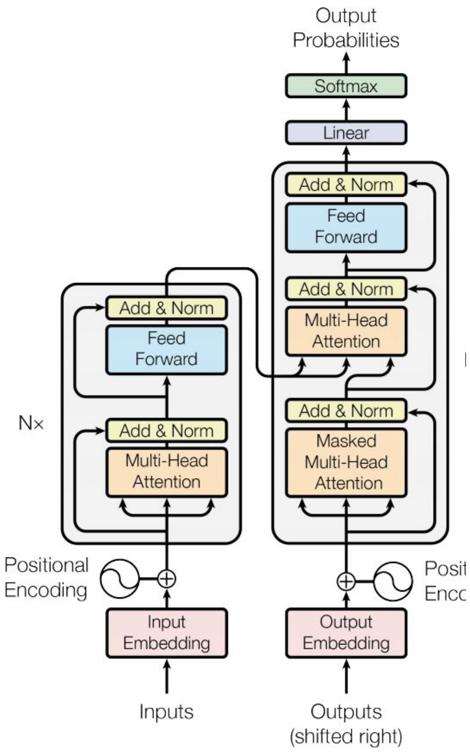


Figure 1: The Transformer - model architecture.

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

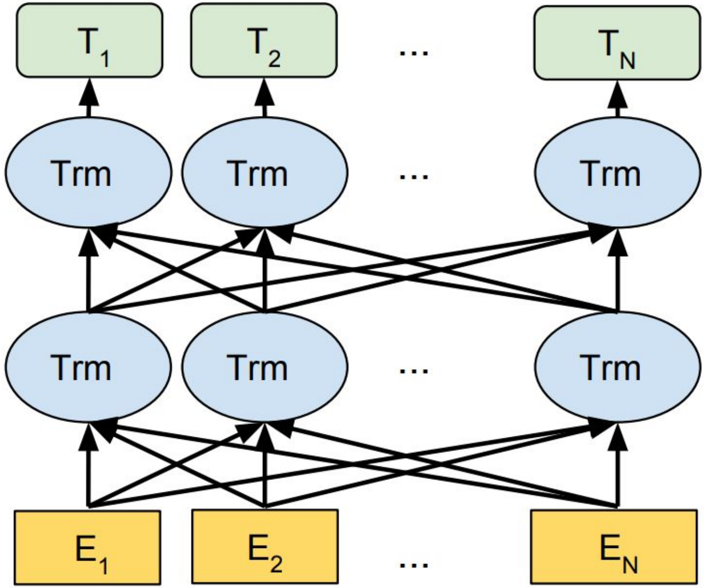
higher accuracy w/ 10X-100X less compute!

Attention Is All You Need,
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin,
June, 2017, https://arxiv.org/abs/1706.03762, appeared in NeurIPS, Dec. 2017

2018: Bidirectional Encoder Representations from Transformers (BERT)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova, Oct. 2018, https://arxiv.org/abs/1810.04805
appeared in NAACL 2019 (Best Paper Award)

2018: Bidirectional Encoder Representations from Transformers (BERT)

Original words

Obama was born in 1961 in Honolulu , Hawaii , two years after the territory was admitted to the Union as the 50th state . Raised largely in Hawaii , he also spent one year of his childhood in Washington state and four years in Indonesia.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova, Oct. 2018, https://arxiv.org/abs/1810.04805
appeared in NAACL 2019 (Best Paper Award)

2018: Bidirectional Encoder Representations from Transformers (BERT)

Masked words

Obama was ____ in 1961 in Honolulu , Hawaii , ____ ____ after the territory was admitted to the ___ as the 50th ____ . Raised largely in ____ , he also spent one year of his ____ in Washington state and four years in Indonesia.

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](https://arxiv.org/abs/1810.04805)
Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova, Oct. 2018, https://arxiv.org/abs/1810.04805
appeared in NAACL 2019 (Best Paper Award)

2018: Bidirectional Encoder Representations from Transformers (BERT)

Masked words

Obama was ____ in 1961 in Honolulu , Hawaii , ____ ____ after the territory was admitted to the ___ as the 50th ____ . Raised largely in ____ , he also spent one year of his ____ in Washington state and four years in Indonesia.

Masks
(3, 11, 12, 20, 24, 29, 33)

Original words

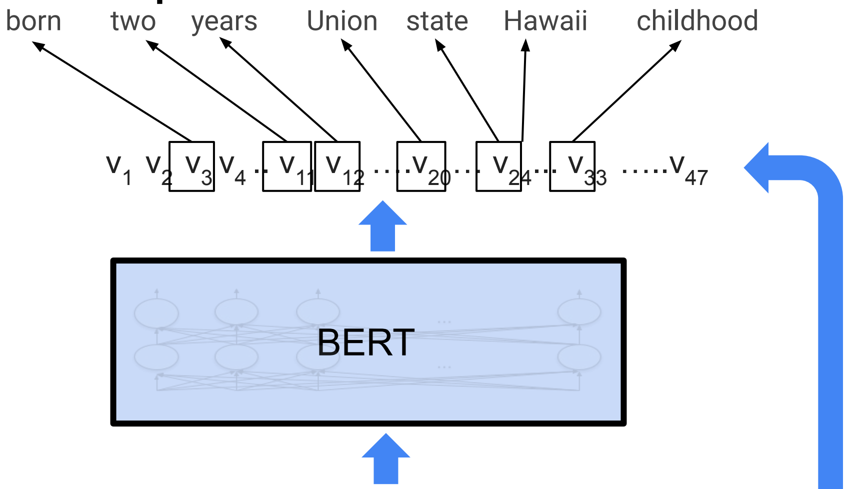
Obama was born in 1961 in Honolulu , Hawaii , two years after the territory was admitted to the Union as the 50th state . Raised largely in Hawaii , he also spent one year of his childhood in Washington state and four years in Indonesia.

[*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*](https://arxiv.org/abs/1810.04805)
Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova, Oct. 2018, https://arxiv.org/abs/1810.04805
appeared in NAACL 2019 (Best Paper Award)

2018: Bidirectional Encoder Representations from Transformers (BERT)

Predict masked words

Encode



Masked words

Obama was ____ in 1961 in Honolulu , Hawaii , ____ ____ after the territory was admitted to the ___ as the 50th ____ . Raised largely in ____ , he also spent one year of his ____ in Washington state and four years in Indonesia.

Masks
(3, 11, 12, 20, 24, 29, 33)

Original words

Obama was born in 1961 in Honolulu , Hawaii , two years after the territory was admitted to the Union as the 50th state . Raised largely in Hawaii , he also spent one year of his childhood in Washington state and four years in Indonesia.

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](https://arxiv.org/abs/1810.04805)
Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova, Oct. 2018, https://arxiv.org/abs/1810.04805
appeared in NAACL 2019 (Best Paper Award)

2018: Bidirectional Encoder Representations from Transformers (BERT)

Key thing that works extremely well:

Step 1: pre-train a model on this “fill in the blanks” task using large-amounts of self-supervised text

Step 2: fine-tune this model on individual language tasks with small amounts of data

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova, Oct. 2018, https://arxiv.org/abs/1810.04805
appeared in NAACL 2019 (Best Paper Award)

GLUE results (General Language Understanding Evaluation), gluebenchmark.com

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server.

large improvements over state of the art (SOTA) on wide variety of language tasks

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova, Oct. 2018, https://arxiv.org/abs/1810.04805
appeared in NAACL 2019 (Best Paper Award)

Natural Questions dataset

https://ai.google.com/research/NaturalQuestions
natural-questions@google.com

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins

Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang

Andrew M. Dai, Jakob Uszkoreit, Quoc Le, Slav Petrov

Data statistics:

- 307,373 training examples
- 7,830 five way annotated examples for development
- 7,842 five way annotated examples for test
- 49% of examples have a long answer
- 35% of examples have a short answer span
- 1% of examples have a yes/no answer

Natural Questions --- Data

We like question answering as a testbed because

- Questions can be arbitrarily complex
  - require world knowledge
  - require reasoning about events
- Task is relatively easy to evaluate

This example requires us to know that disabling telephony implies that you cannot make a call.

Question: *Can you make and receive calls in airplane mode?*

Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, ***suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony****, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.*

Answer: *No*

Natural Questions

https://ai.google.com/research/NaturalQuestions

Google AI

Leaderboards

Details

LONG ANSWER

Rank	Model	Participant	Affiliation	F1
1	bert_dm	dancingsoul	individual	0.7196
2	bert-dm	dancingsoul	individual	0.70248
3	BERT-syn	anon_83692	Anon	0.66774
4	Insight-baseline	L.Xiao_R.Ren	PAII Insight Team	0.66458
5	BERT	Chris-A	Google	0.66157
6	BERT-mnlp	BANQ	IBM Research AI	0.64587
7	Insight-BERT-single	L.Xiao_R.Ren	PAII Insight Team	0.63949

Engineer the Tools of Scientific Discovery

TensorFlow

http://tensorflow.org/

and

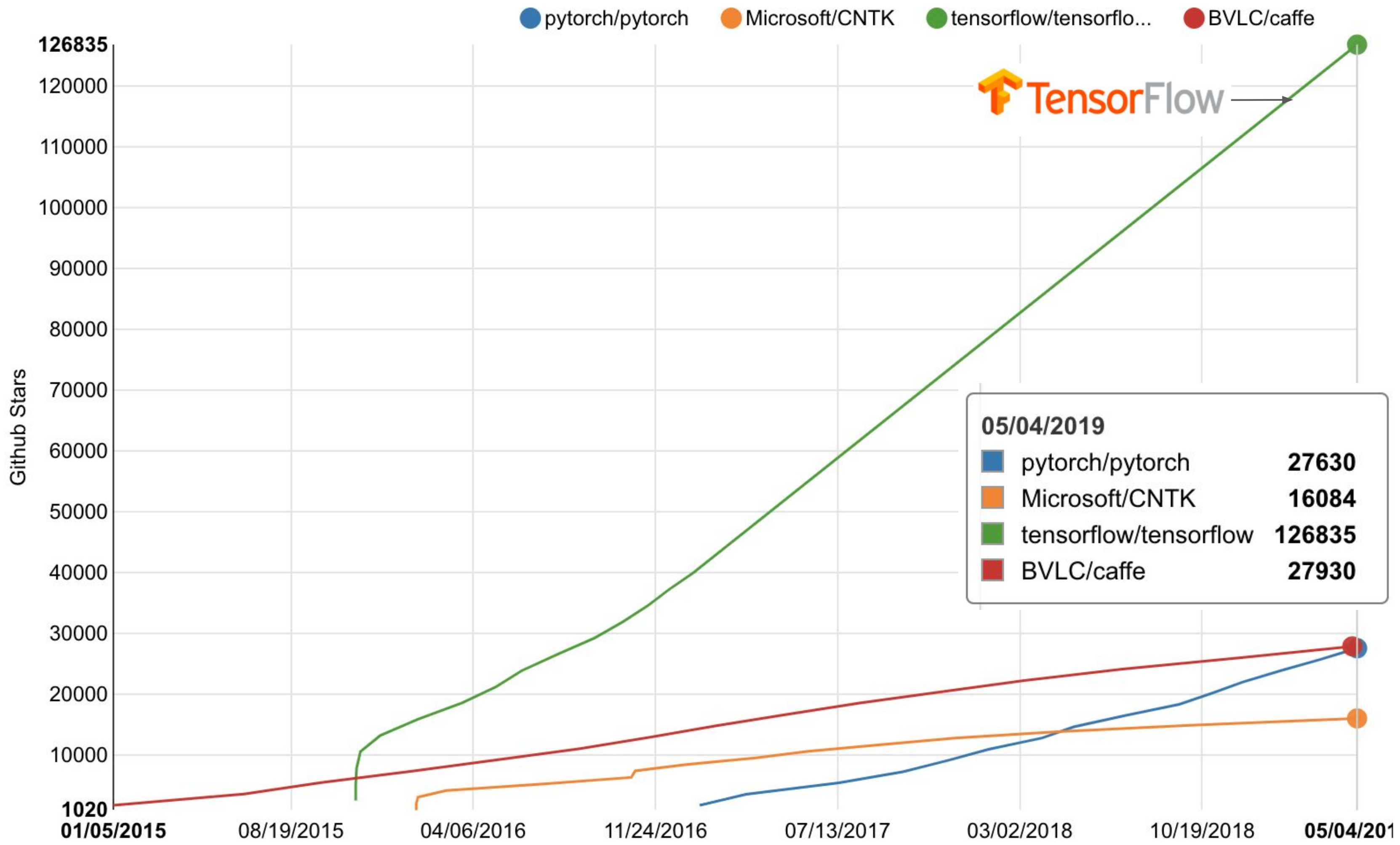
https://github.com/tensorflow/tensorflow

Open, standard software for general machine learning

Great for Deep Learning in particular

First released Nov 2015

Apache 2.0 license



Source: star-history.t9t.io/

A vibrant **Open-Source Community**

Positive Reviews

125,000+
GitHub Stars

58,000+
GitHub repositories with ‘TensorFlow’ in the title

Rapid Development

1,900+
Contributors

55,000+
Commits in <4 years

Direct Engagement

20,000+
Stack Overflow questions answered

100+
Community-submitted GitHub issues responded to weekly

50,000,000+
Downloads

https://www.blog.google/topics/machine-learning/using-tensorflow-keep-farmers-happy-and-cows-healthy/

frontiers in Plant Science

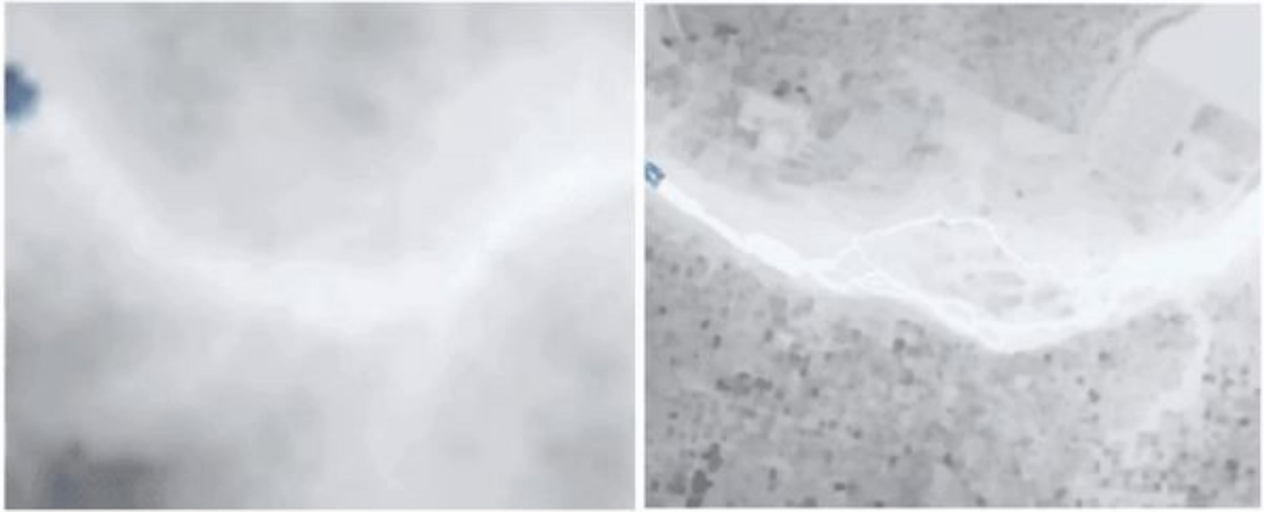
Deep Learning for Image-Based Cassava Disease Detection

Amanda Ramcharan,[1] Kelsee Baranowski,[1] Peter McCloskey,[2] Babuali Ahmed,[3] James Legg,[3] and David P. Hughes[1,4,5,*]

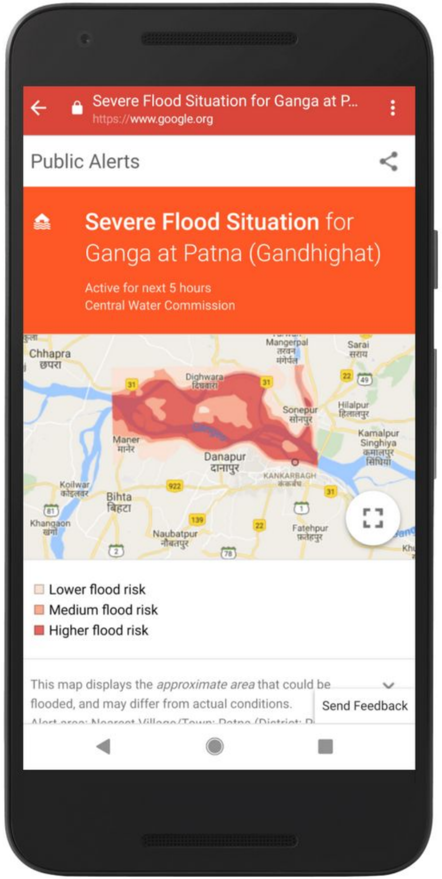
Penn State and International Institute of Tropical Agriculture

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5663696/

Better models for flood forecasting



A flood simulation of a river in Hyderabad, India. The left side uses publicly available data while the right side uses additional data and more sophisticated machine learning models. Our models contain higher resolution, accuracy, and up-to-date information.



Flood alert shown to users in Patma region

https://www.blog.google/products/search/helping-keep-people-safe-ai-enabled-flood-forecasting/

Some pieces of work and how they fit together

Google

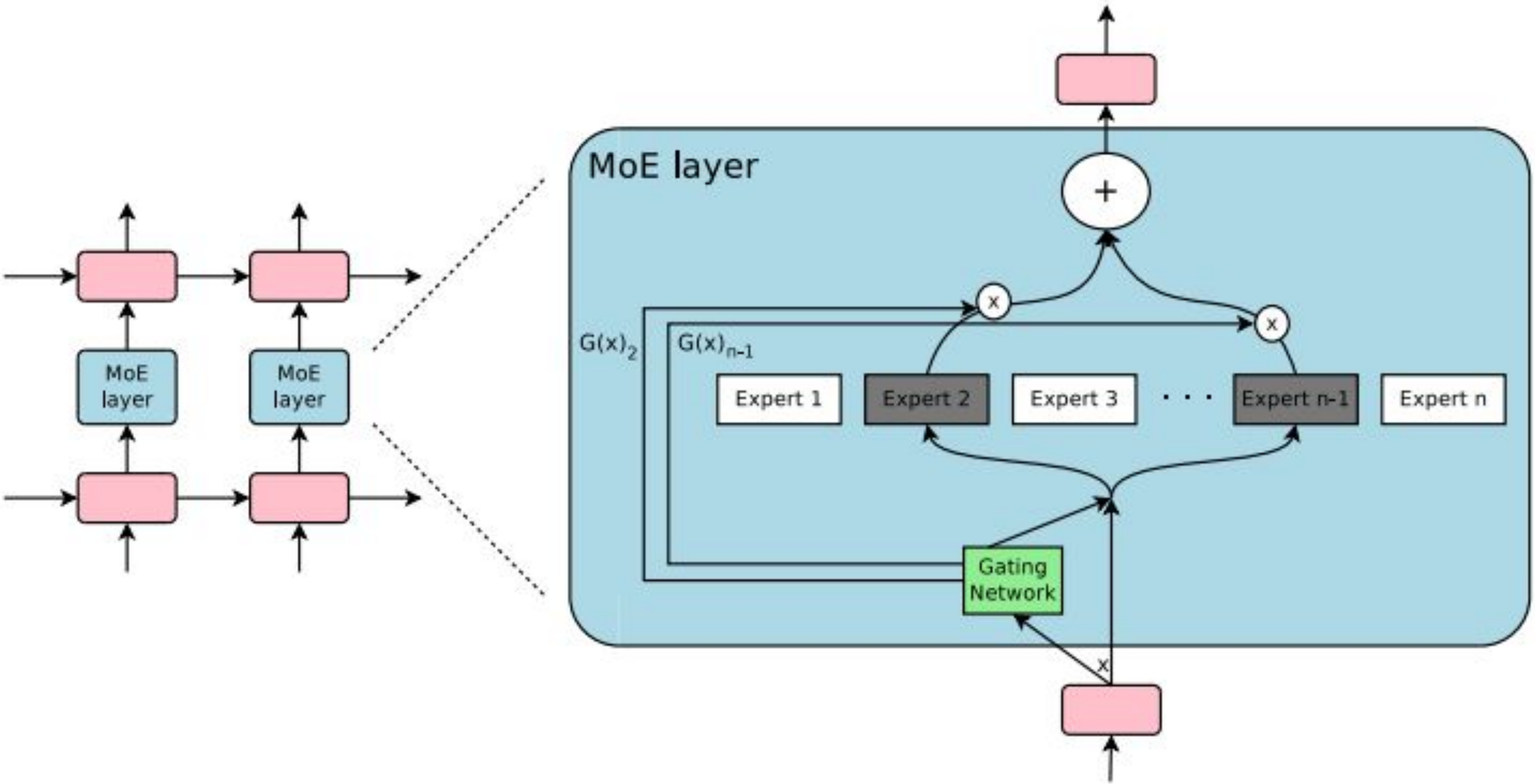
Bigger models, but sparsely activated

Bigger models, but sparsely activated

Motivation:

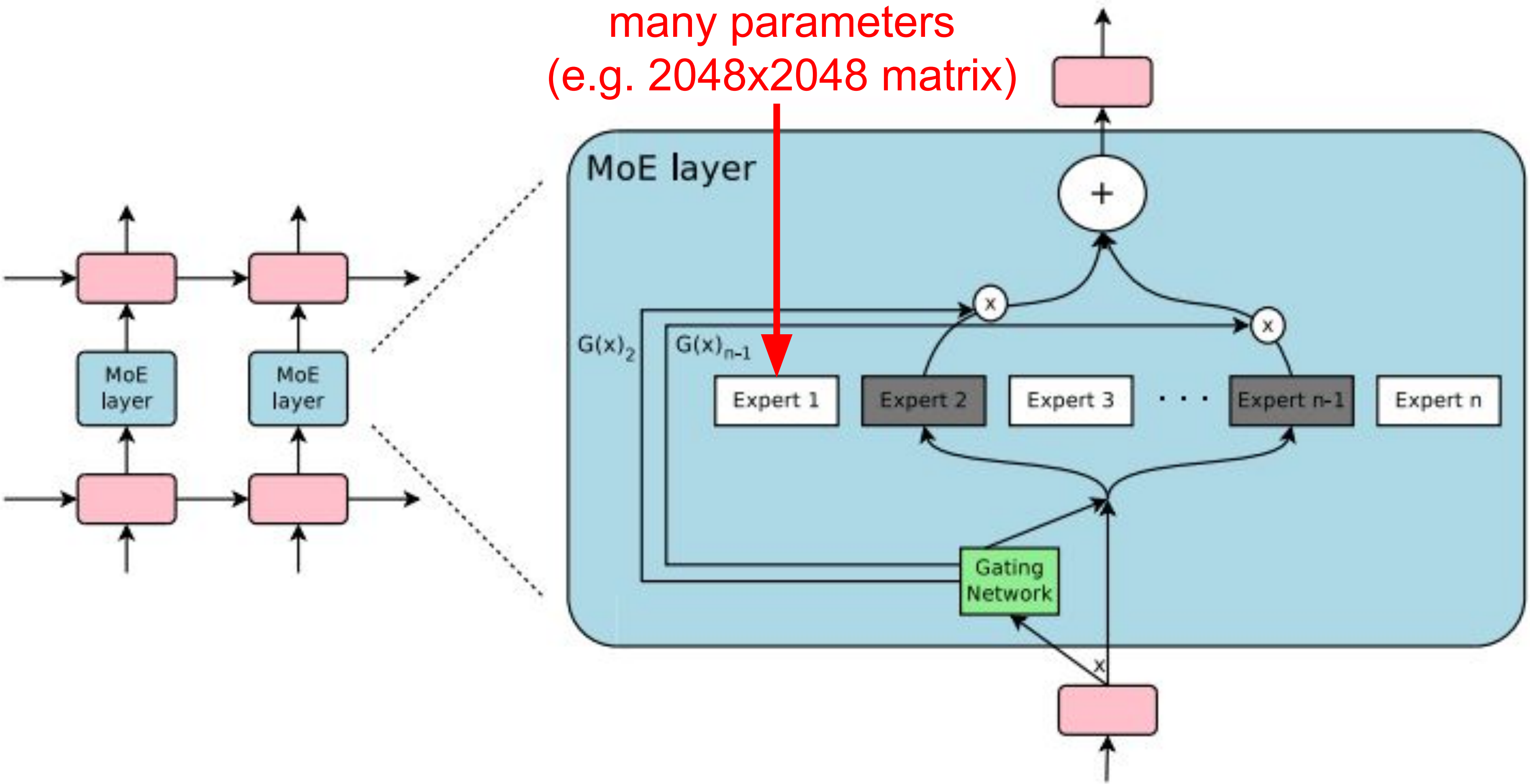
Want **huge model capacity** for large datasets, but want individual example to **only activate tiny fraction** of large model

Per-Example Routing

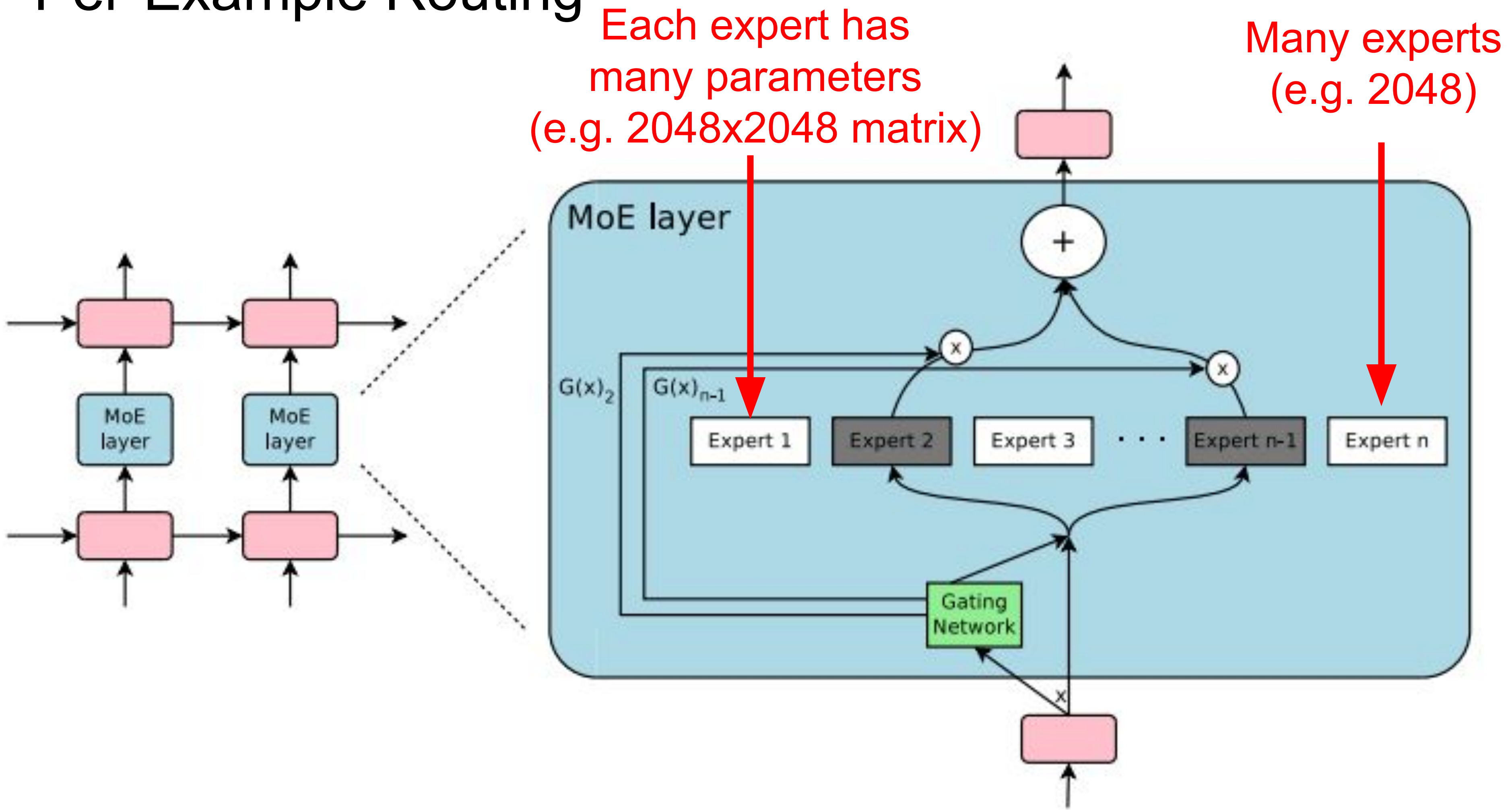


Per-Example Routing

Each expert has
many parameters
(e.g. 2048x2048 matrix)



Per-Example Routing



Expert 381	Expert 752	Expert 2004
... with **researchers** , ...	... plays **a** core ...	... with **rapidly** growing ...
... to **innovation** .	... plays **a** critical ...	... under **static** conditions ...
... tics **researchers** .	... provides **a** legislative ...	... to **swift** ly ...
... the **generation** of ...	... play **a** leading ...	... to **dras** tically ...
... technology **innovations** is ...	... assume **a** leadership ...	... the **rapid** and ...
... technological **innovations** , ...	... plays **a** central ...	... the **fast** est ...
... support **innovation** throughout ...	... taken **a** leading ...	... the **Quick** Method ...
... role **innovation** will ...	... established **a** reconciliation ...	... rec **urrent**) ...
... research **scienti** st ...	... played **a** vital ...	... provides **quick** access ...
... promoting **innovation** where ...	... have **a** central ...	... of **volatile** organic ...
...	...	...

Per-Example Routing

+1 BLEU point
1/10 the training cost
1/2 the inference cost

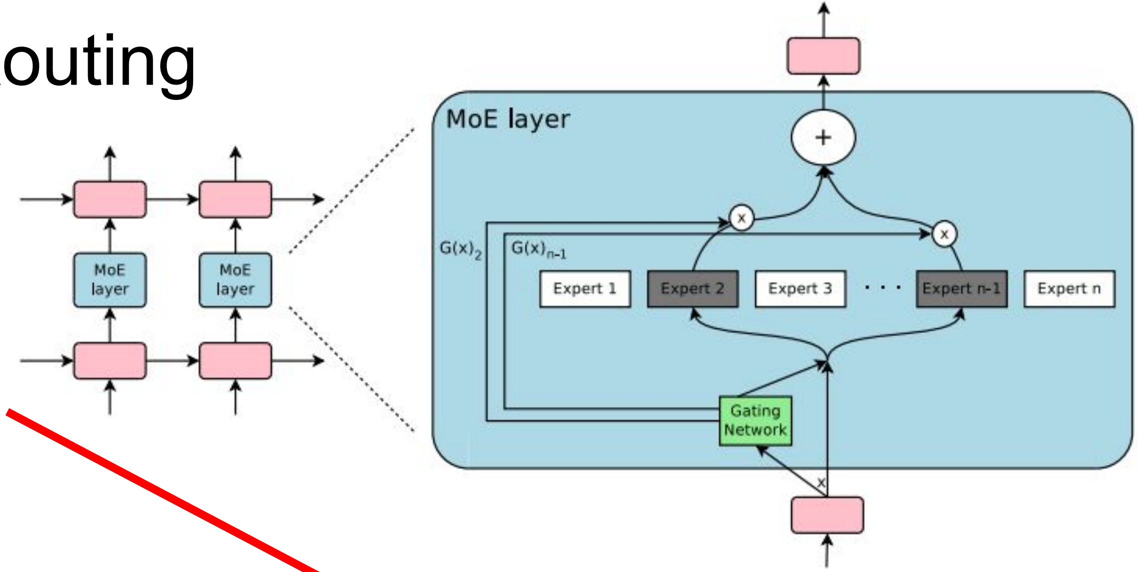


Table 7: Perplexity and BLEU comparison of our method against previous state-of-art methods on the Google Production En→ Fr dataset.

Model	Eval Perplexity	Eval BLEU	Test Perplexity	Test BLEU	Computation per Word	Total #Parameters	Training Time
MoE with 2048 Experts	2.60	37.27	2.69	36.57	100.8M	8.690B	1 day/64 k40s
GNMT (Wu et al., 2016)	2.78	35.80	2.87	35.56	214.2M	246.9M	6 days/96 k80s

Outrageously Large Neural Networks: The Sparsely-gated Mixture-of-Experts Layer,
Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le & Jeff Dean
Appeared in ICLR 2017, https://openreview.net/pdf?id=B1ckMDqlg

AutoML: Automated machine learning ("learning to learn")

Current:

Solution = ML expertise + data + computation

Current:

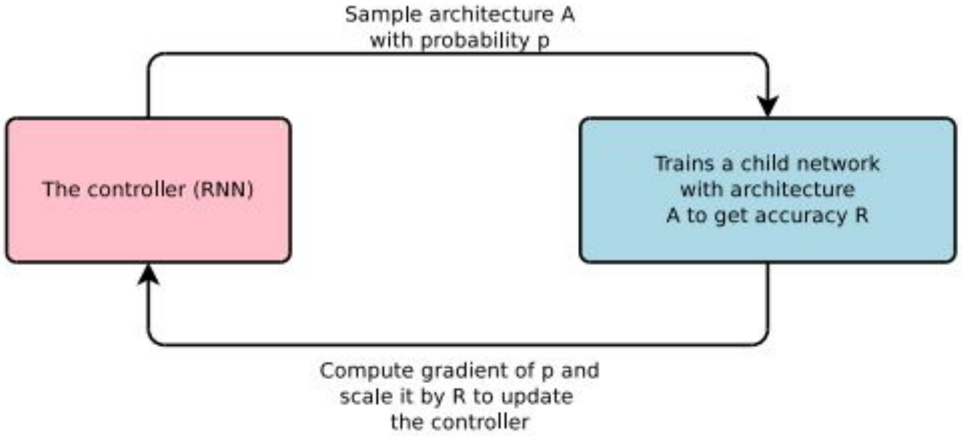
Solution = ML expertise + data + computation

Can we turn this into:

Solution = data + computation

???

Neural Architecture Search

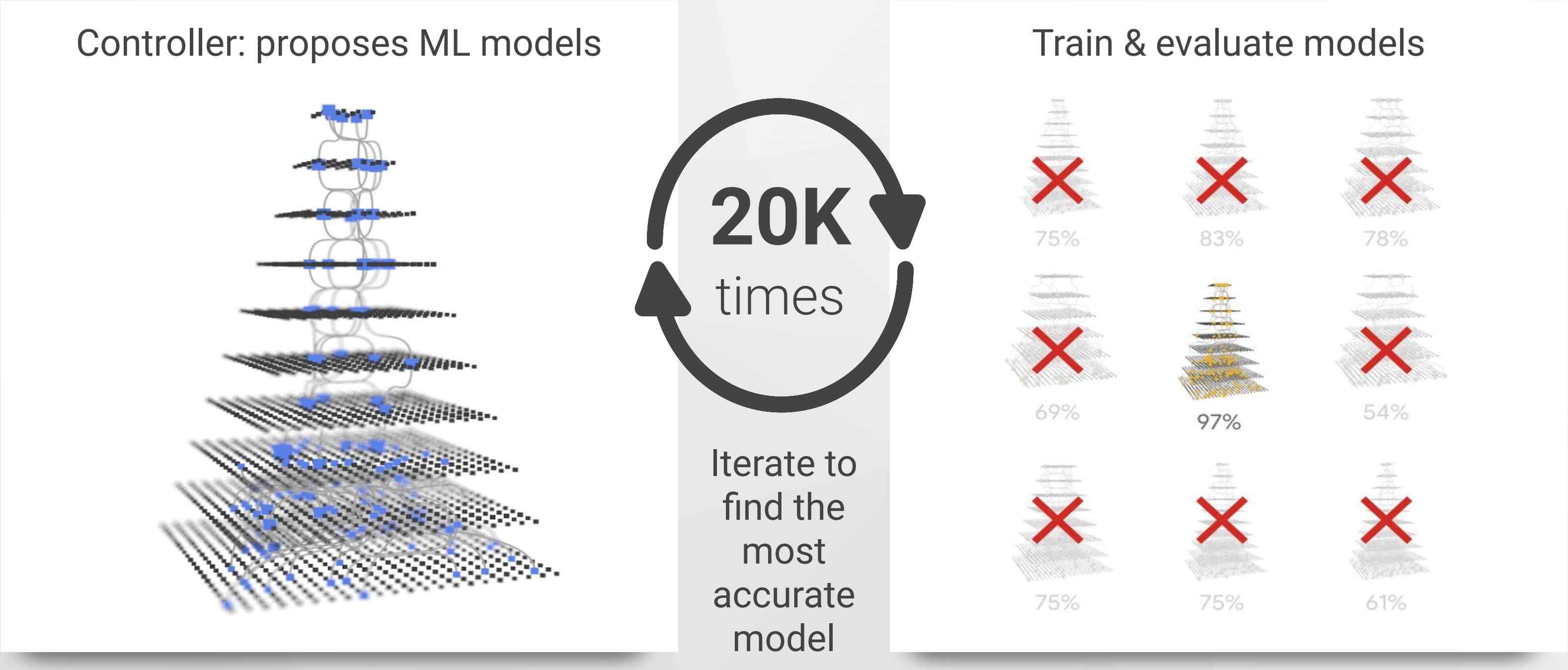


Idea: model-generating model trained via reinforcement learning

(1) Generate ten models
(2) Train them for a few hours
(3) Use loss of the generated models as reinforcement learning signal

Neural Architecture Search with Reinforcement Learning, Zoph & Le, ICLR 2016
arxiv.org/abs/1611.01578

Neural Architecture Search to find a model



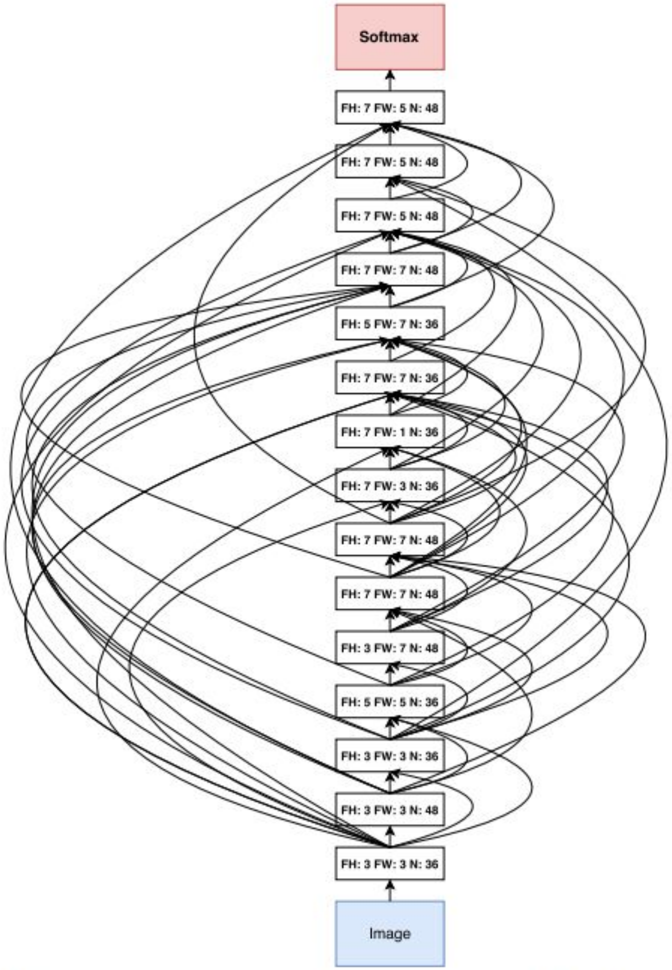
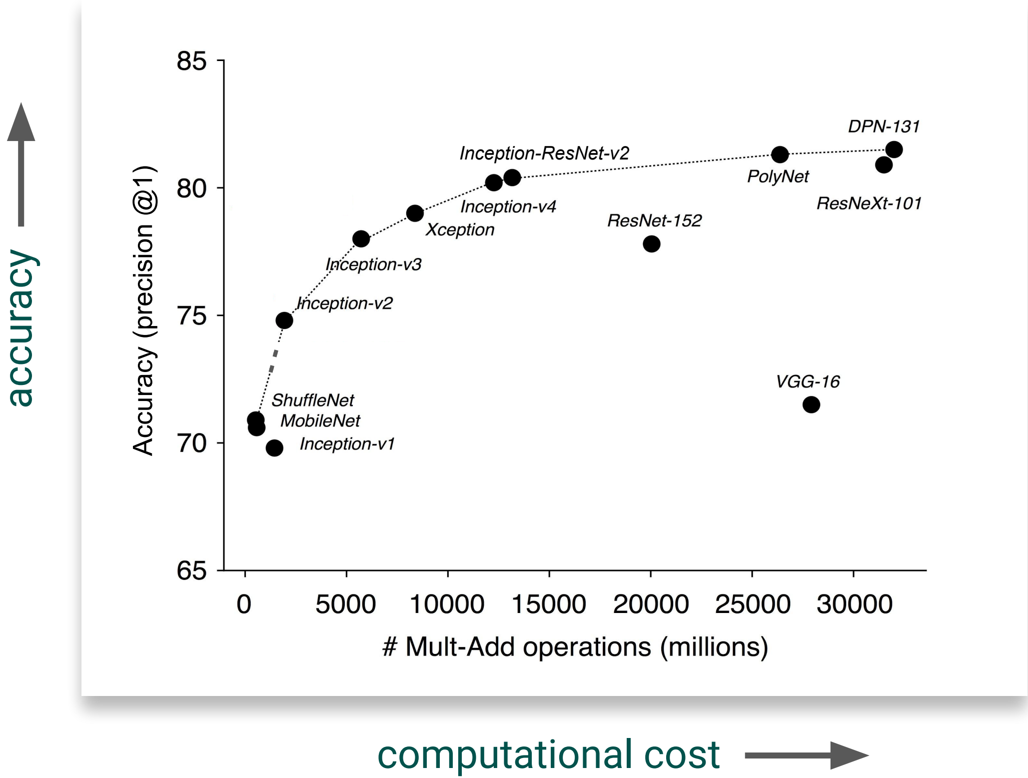


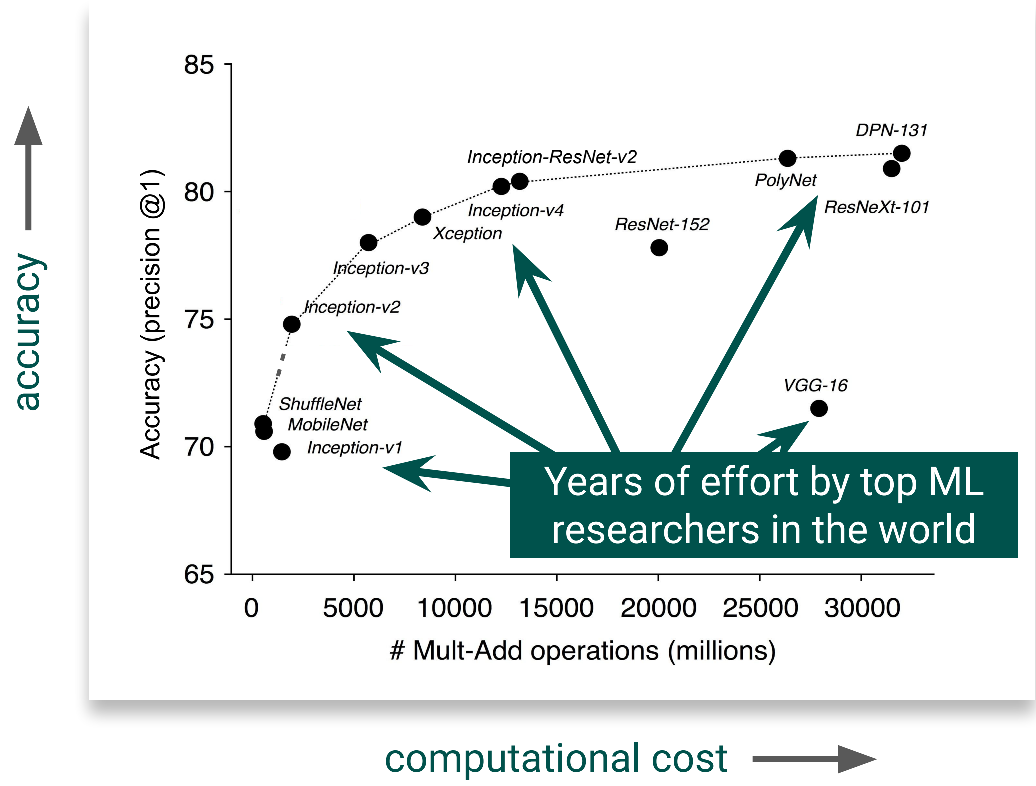
Figure 7: Convolutional architecture discovered by our method, when the search space does not have strides or pooling layers. FH is filter height, FW is filter width and N is number of filters.

AutoML outperforms handcrafted models



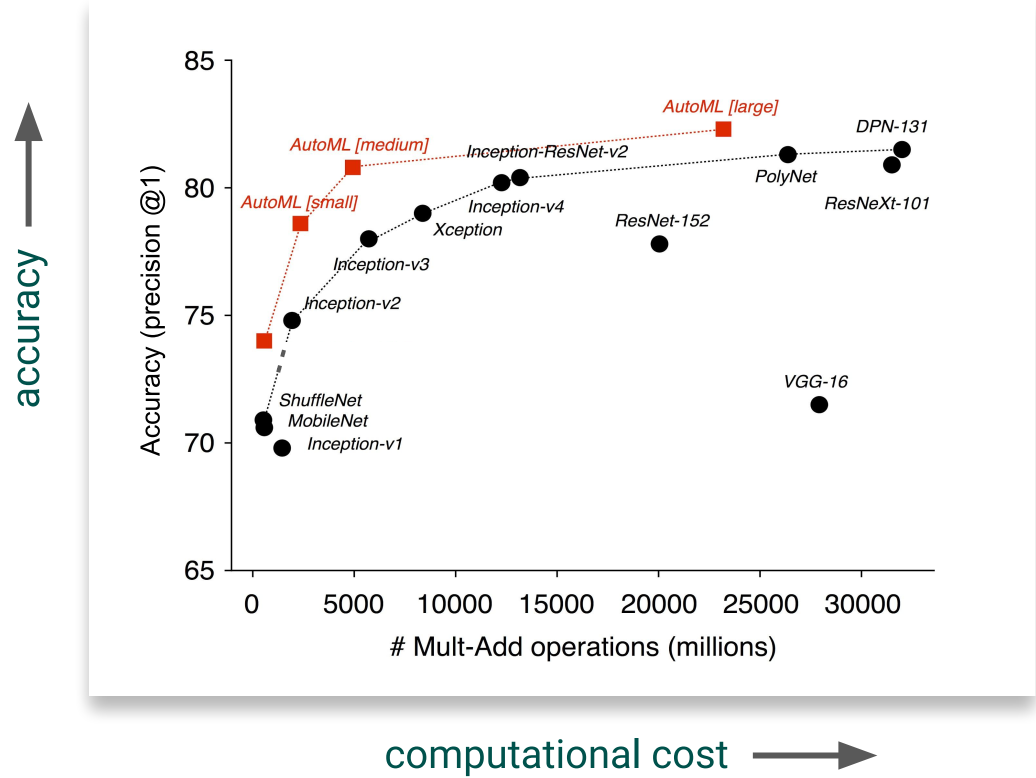
Learning Transferable Architectures for Scalable Image Recognition, Zoph et al. 2017,
https://arxiv.org/abs/1707.07012

AutoML outperforms handcrafted models



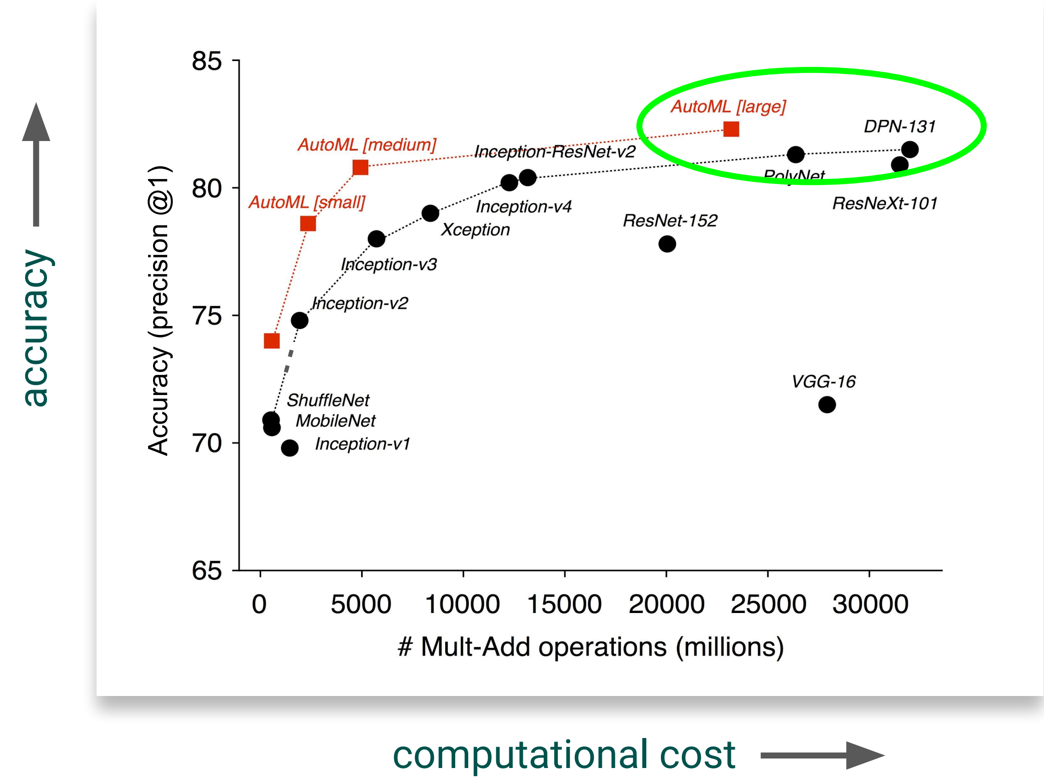
Learning Transferable Architectures for Scalable Image Recognition, Zoph et al. 2017, https://arxiv.org/abs/1707.07012

AutoML outperforms handcrafted models



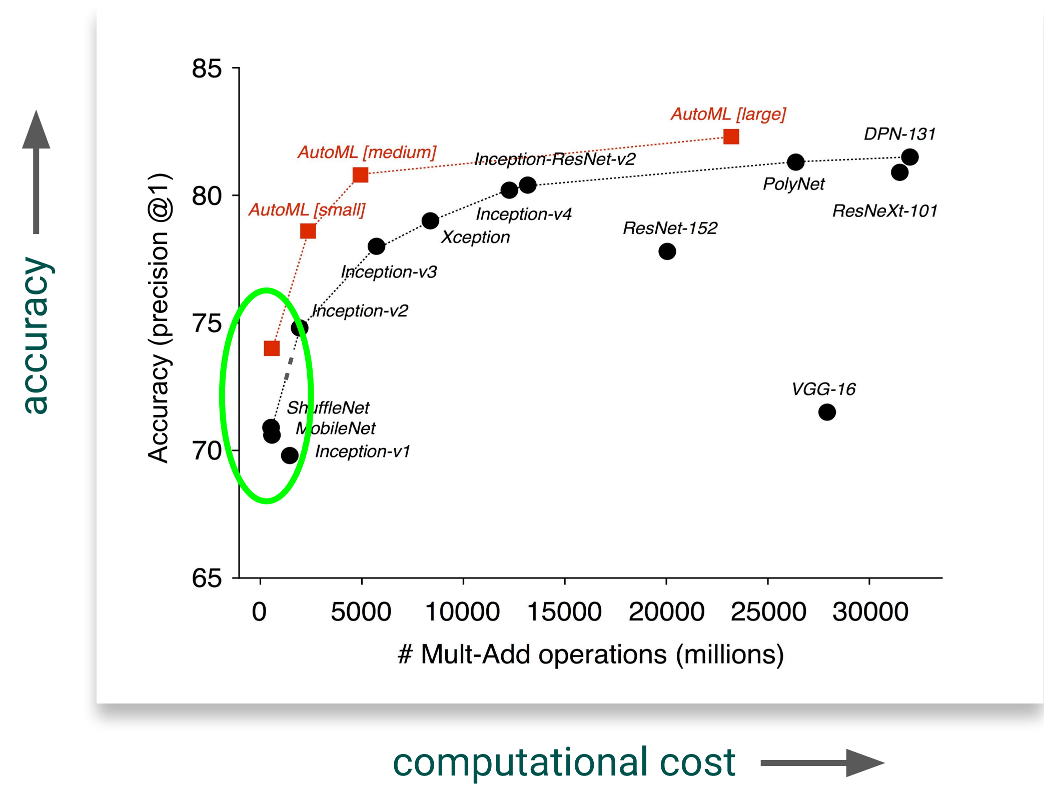
Learning Transferable Architectures for Scalable Image Recognition, Zoph et al. 2017, https://arxiv.org/abs/1707.07012

AutoML outperforms handcrafted models



Learning Transferable Architectures for Scalable Image Recognition, Zoph et al. 2017, https://arxiv.org/abs/1707.07012

AutoML outperforms handcrafted models



Learning Transferable Architectures for Scalable Image Recognition, Zoph et al. 2017, https://arxiv.org/abs/1707.07012

AI & Machine Learning Products

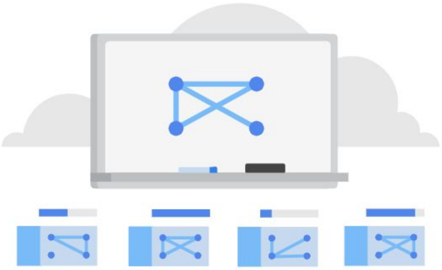
Cloud AutoML BETA

Train high-quality custom machine learning models with minimal effort and machine learning expertise.

TRY AUTOML VIEW DOCUMENTATION

Train custom machine learning models

Cloud AutoML is a suite of machine learning products that enables developers with limited machine learning expertise to train high-quality models specific to their business needs. It relies on Google's state-of-the-art transfer learning and neural architecture search technology.



https://cloud.google.com/automl/

AI & Machine Learning Products

AutoML products

Create your own custom machine learning models with an easy-to-use graphical interface.

Sight

AutoML Vision

Derive insights from images in the cloud or at the edge.

LEARN MORE

AutoML Video Intelligence

Enable powerful content discovery and engaging video experiences.

LEARN MORE

Language

AutoML Natural Language

Reveal the structure and meaning of text through machine learning.

LEARN MORE

AutoML Translation

Dynamically detect and translate between languages.

LEARN MORE

Structured data

AutoML Tables

Automatically build and deploy state-of-the-art machine learning models on structured data.

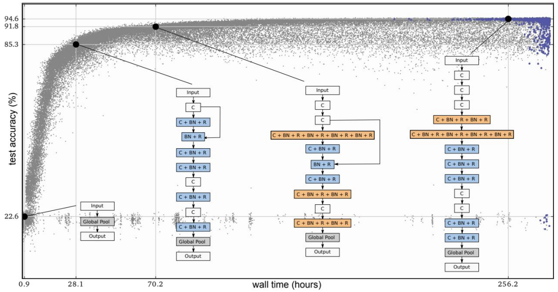
LEARN MORE

https://cloud.google.com/automl/

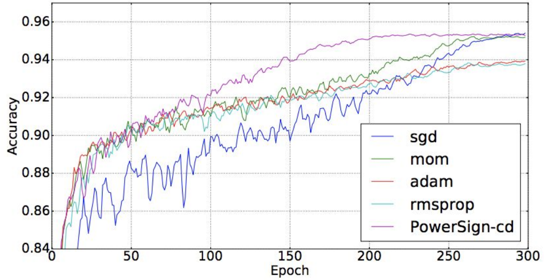
Additional Work in AutoML

Evolution for search rather than reinforcement learning:
Regularized Evolution for Image Classifier Architecture Search,
Esteban Real, Alok Aggarwal, Yanping Huang, Quoc V Le,
https://arxiv.org/abs/1802.01548

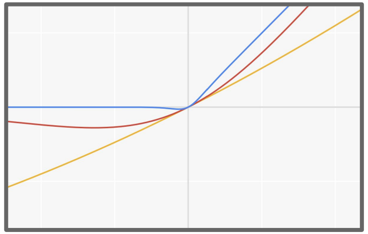
Large-Scale Evolution of Image Classifiers,
Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc Le, Alex Kurakin
https://arxiv.org/abs/1703.01041



Learn the optimization update rule:
Neural Optimizer Search with Reinforcement Learning,
Irwan Bello, Barret Zoph, Vijay Vasudevan, Quoc V. Le,
https://arxiv.org/abs/1709.07417

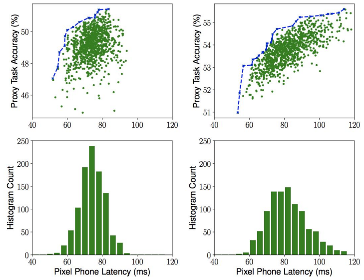


Learn the non-linearity to use as an activation function:
Searching for Activation Functions,
Prajit Ramachandran, Barret Zoph, Quoc V. Le,
https://arxiv.org/abs/1710.05941



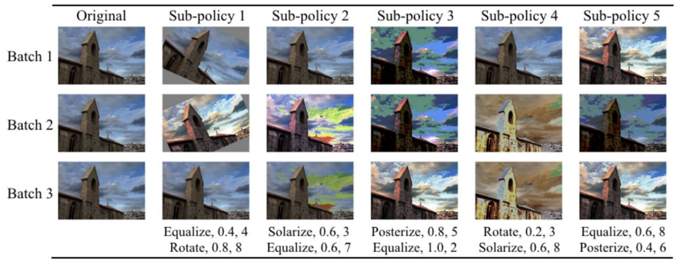
Additional Work in AutoML (cont)

Incorporate inference latency & accuracy into reward:
MnasNet: Platform-Aware Neural Architecture Search for Mobile,
Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan,
Quoc V. Le, https://arxiv.org/abs/1807.11626

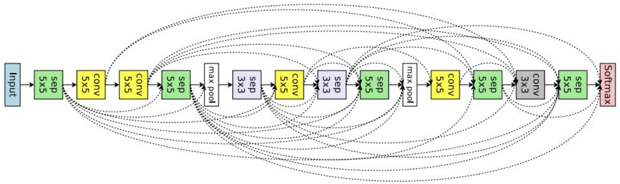


(a) $\alpha = 0, \beta = -1$ (b) $\alpha = \beta = -0.07$

Learn data augmentation policies:
AutoAugment: Learning Augmentation Policies from Data,
Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan,
Quoc V. Le, https://arxiv.org/abs/1805.09501



Explore many architectures simultaneously w/ parameter sharing:
Efficient Neural Architecture Search via Parameters Sharing In Deep Learning,
Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, Jeff Dean
https://arxiv.org/abs/1802.03268



More computational power needed

Deep learning is transforming how we design computers

Special computation properties

reduced precision ok

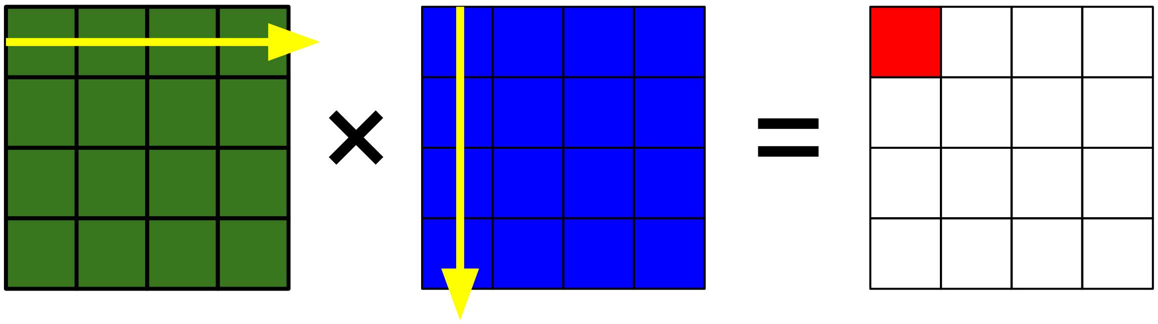
about 1.2		~~1.21042~~
× about 0.6	**NOT**	~~× 0.61127~~
about 0.7		~~0.73989343~~

Special computation properties

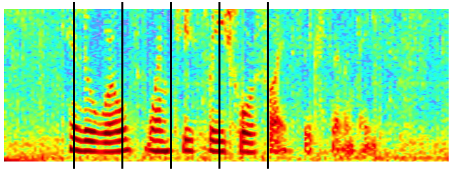
reduced precision ok

about 1.2 × about 0.6 about 0.7	**NOT**	~~1.21042~~ ~~× 0.61127~~ ~~0.73989343~~

handful of specific operations



~2012:
Great initial success with deep neural nets for speech recognition and image recognition

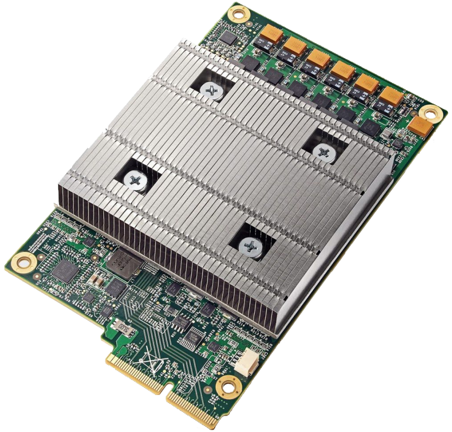


2012 thought exercise:
What if 100M of our users started talking to their phones for three minutes per day?

Uh oh:
Running speech models on CPUs, we'd need to double the number of computers in Google datacenters

TPUv1: Google's first Tensor Processing Unit (TPU)

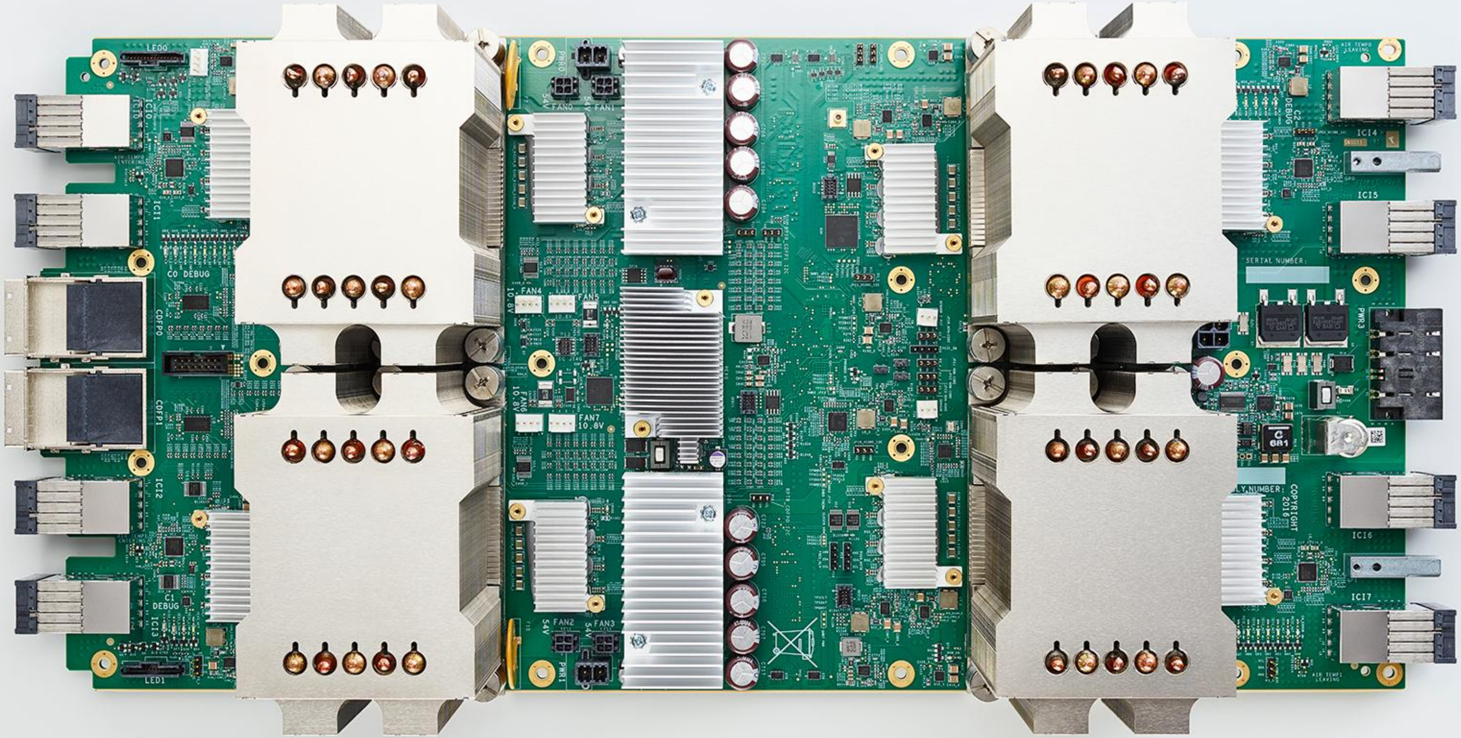
Google-designed chip for neural net **inference**



In production use for ~4 years: used on search queries, for neural machine translation, for speech, for image recognition, for AlphaGo match, …

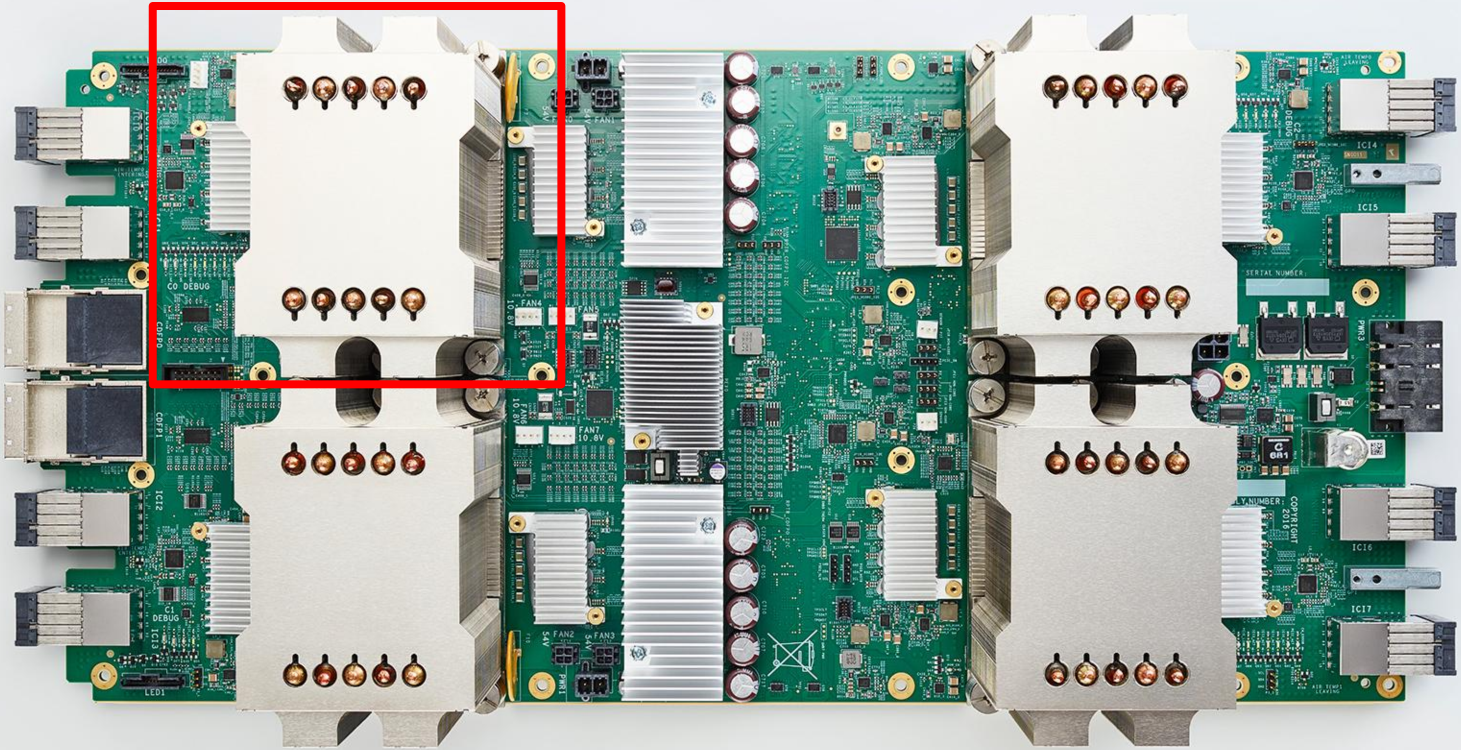
In-Datacenter Performance Analysis of a Tensor Processing Unit, Jouppi, Young, Patil, Patterson *et al.*, ISCA 2017, arxiv.org/abs/1704.04760

TPUv2: for training and inference (available as Cloud TPUv2)



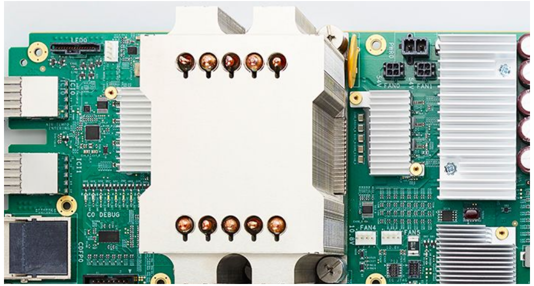
g.co/cloudtpu

TPUv2: for training and inference (available as Cloud TPUv2)

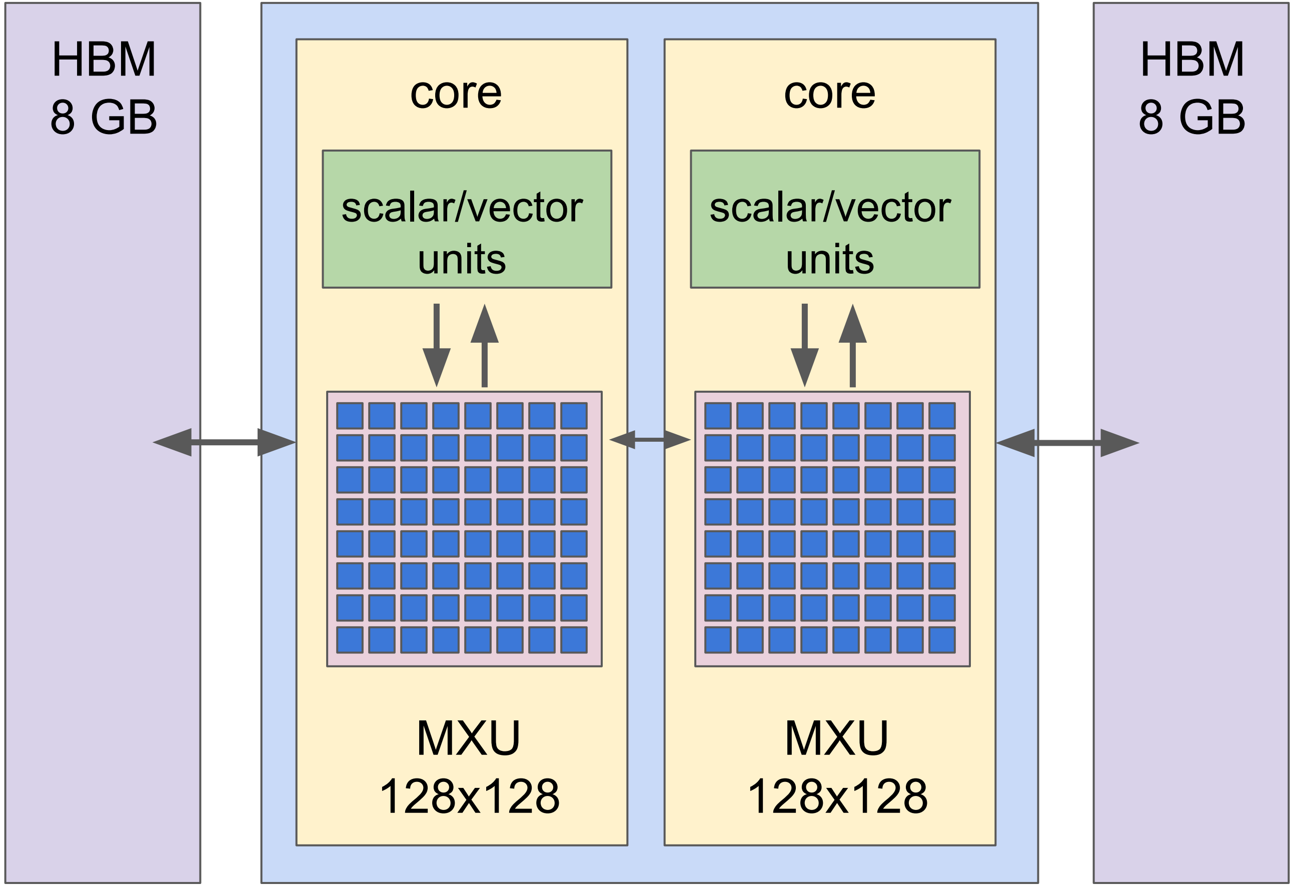


g.co/cloudtpu

TPUv2 Chip



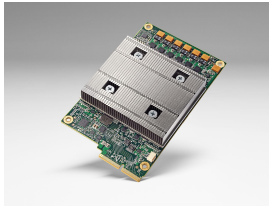
- 16 GB of HBM
- 600 GB/s mem BW
- Scalar/vector units: 32b float
- MXU: 32b float accumulation but reduced precision for multipliers
- 45 TFLOPS



Rapid progress

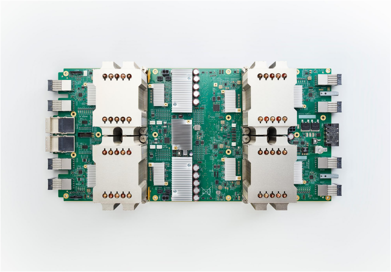
g.co/cloudtpu

TPU v1
(2015)



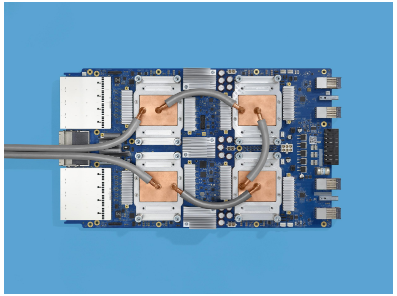
92 teraops
Inference only

Cloud TPU v2
(2017)



180 teraflops
64 GB HBM
Training and inference
Generally available (GA)

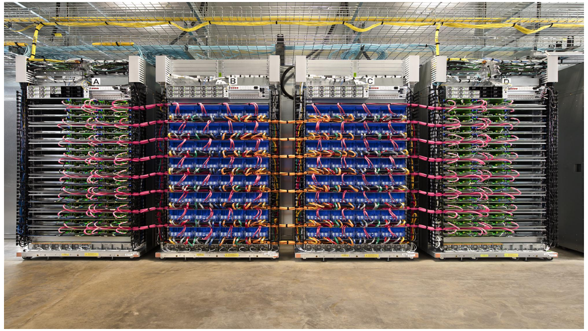
Cloud TPU v3
(2018)



420 teraflops
128 GB HBM
Training and inference
Generally available (GA)

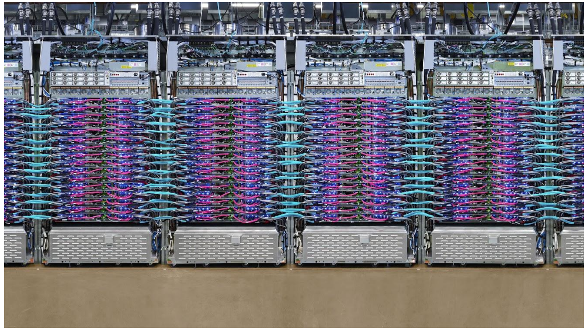
Rapid progress

g.co/cloudtpu



Cloud TPU v2 Pod (2017)

11.5 petaflops
4 TB HBM
2-D toroidal mesh network
Training and inference
Beta

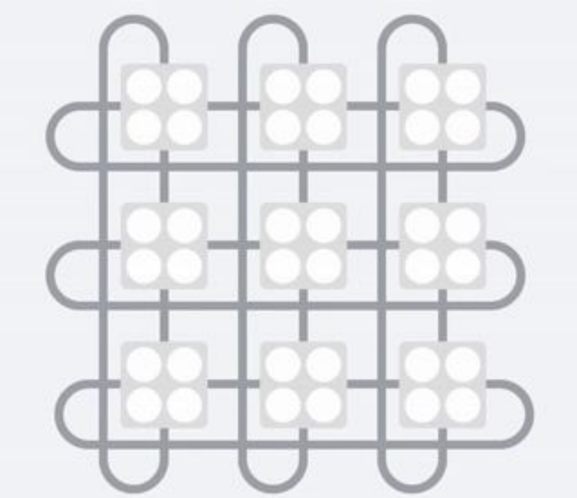


Cloud TPU v3 Pod (2018)

> 100 petaflops!
32 TB HBM
Liquid cooled
New chip architecture + larger-scale system
Beta
Now available to the public for the first time

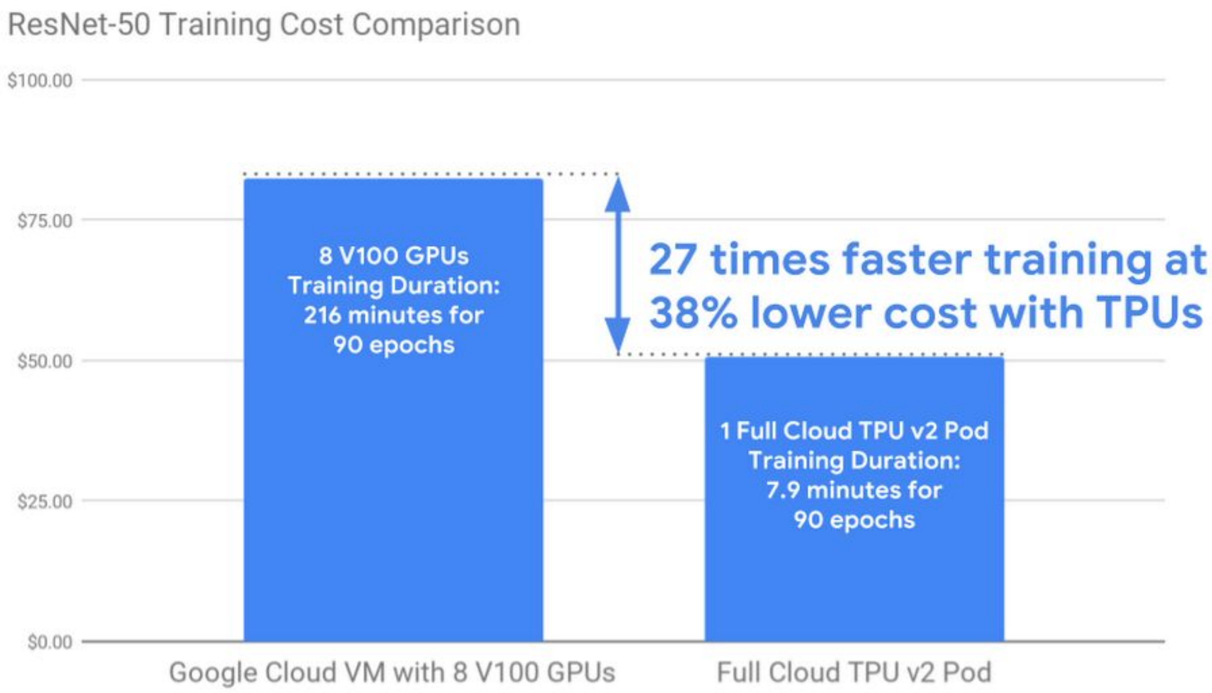
Key to performance of pods:
High-speed 2-D toroidal mesh interconnect => "AI Supercomputers"

Ultra-fast all-reduce using custom hardware



As easy to program as a single node

Cloud TPU v2 Pod (512 cores) vs. NVIDIA V100 (8 GPUs): 27X faster training at 38% lower cost



As of Dec 12, 2018 using TF 1.12

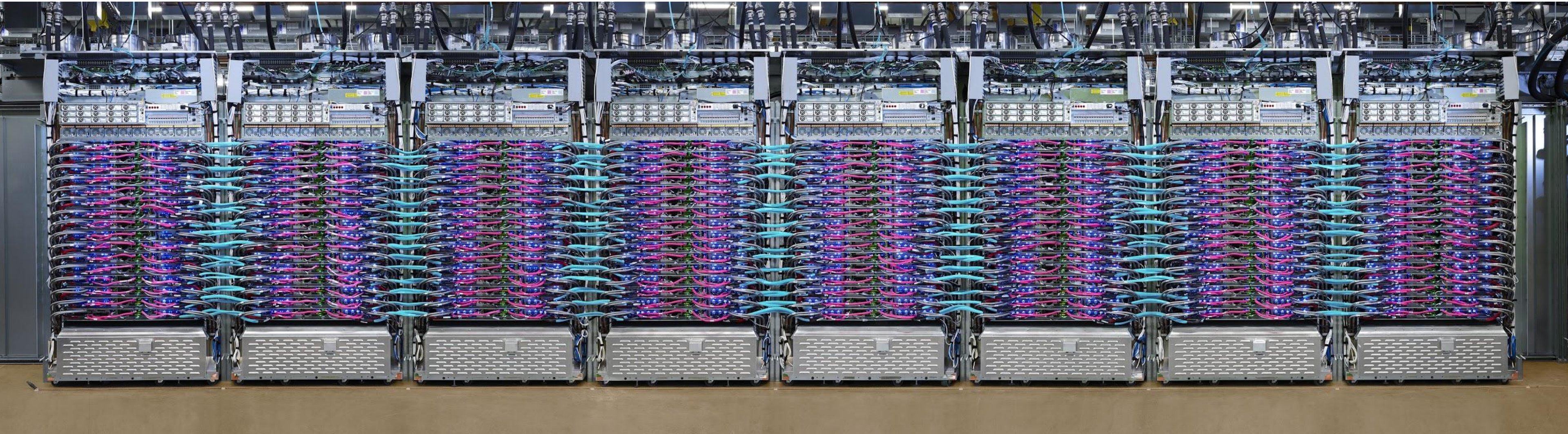
Cloud TPU v3 Pod Performance: Two Examples

Train **ResNet-50 ImageNet image classification** model from scratch in **<2 minutes** on full v3 Pod

Process **more than 1.05M images / second** along the way! (~1 epoch per second)

Train **BERT language representation** from scratch in **just 76 minutes** on full v3 Pod

Training BERT takes days on smaller systems



Many ready-to-use open source models for Cloud TPUs

Image recognition, segmentation, & more

Image Recognition:
AmoebaNet-D
ResNet-50/101/152/200
Inception v2/v3/v4

Object Detection:
RetinaNet, Mask R-CNN

Image Segmentation:
Mask R-CNN, DeepLab, and RetinaNet

Low-Resource Models:
MnasNet, MobileNet, SqueezeNet

Machine translation and language modeling

Machine translation
Language modeling
Sentiment analysis
(all Transformer-based)

Question-answering (QANet)

BERT:
State-of-the-art results across 10+ natural language tasks

Speech recognition

ASR Transformer
(LibriSpeech)

Image generation

Image Transformer
DCGAN
BigGAN
Compare GAN library

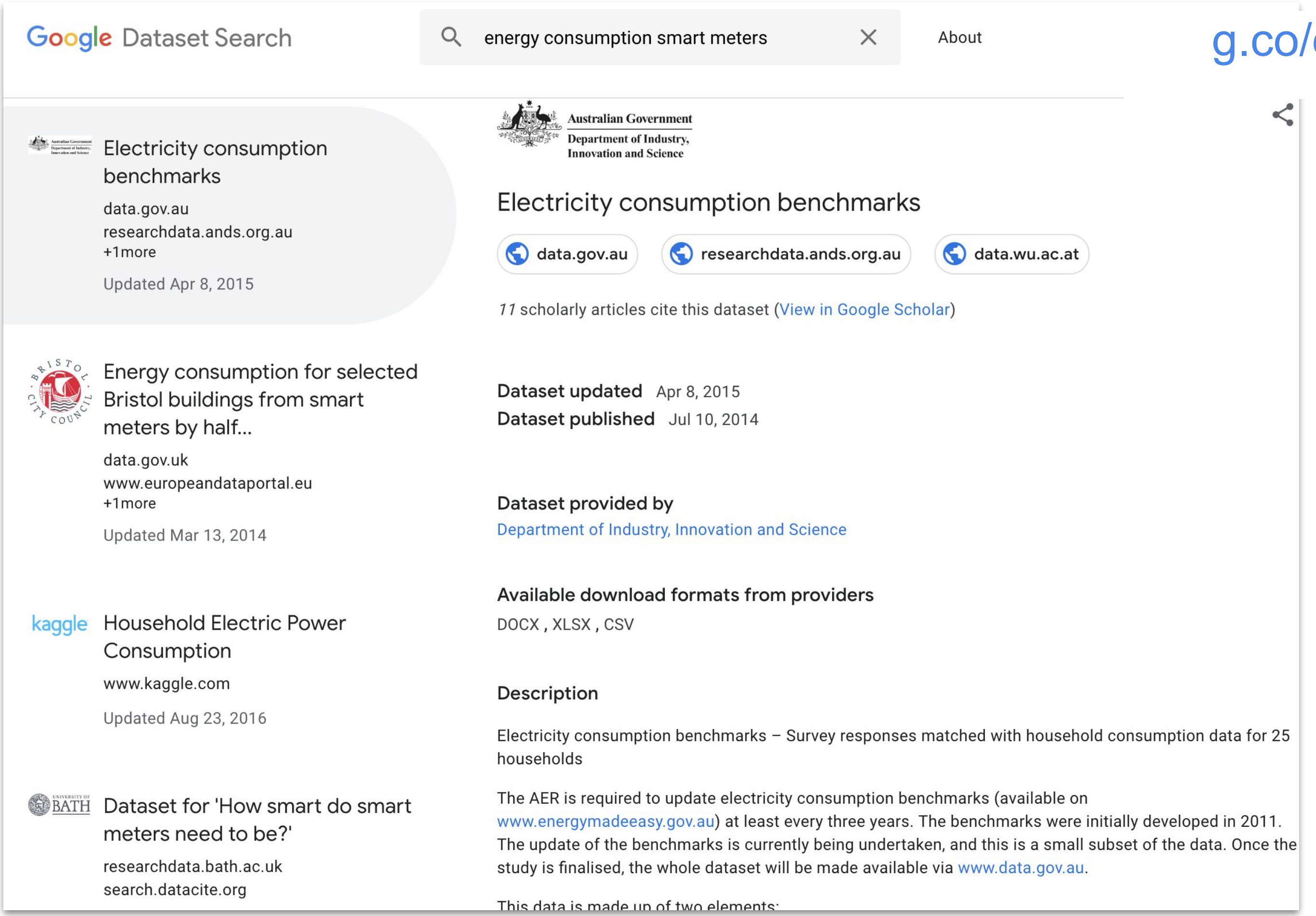
g.co/cloudtpu

Engineer the Tools of Scientific Discovery

Google Dataset Search

g.co/datasetsearch

g.co/datasetsearch



Searching metadata from 1000s of providers totalling more than 10M datasets

How do these fit together?

Combine many of these ideas:

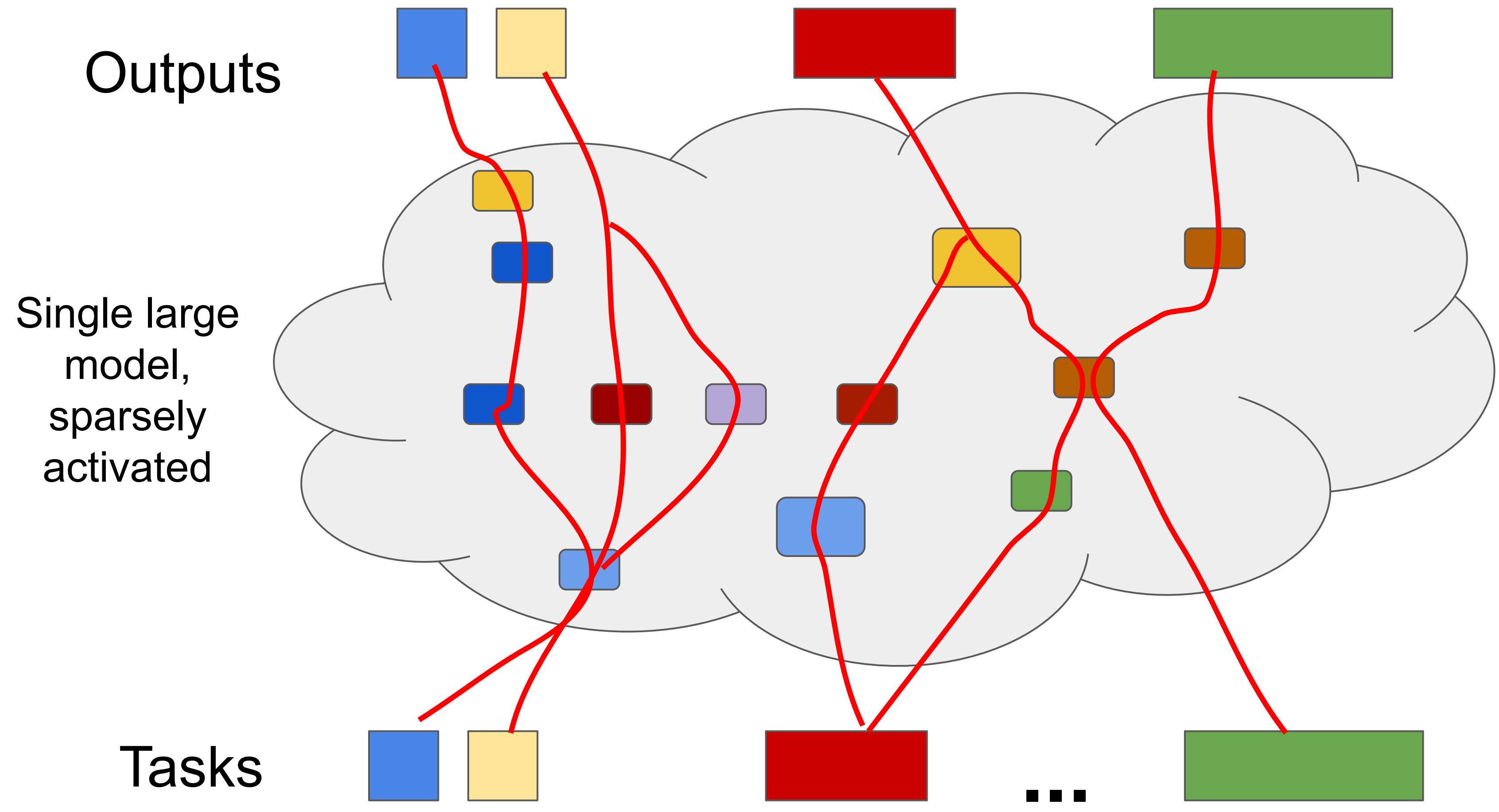
Large model, but **sparsely activated**

Single model to **solve many tasks** (100s to 1Ms)

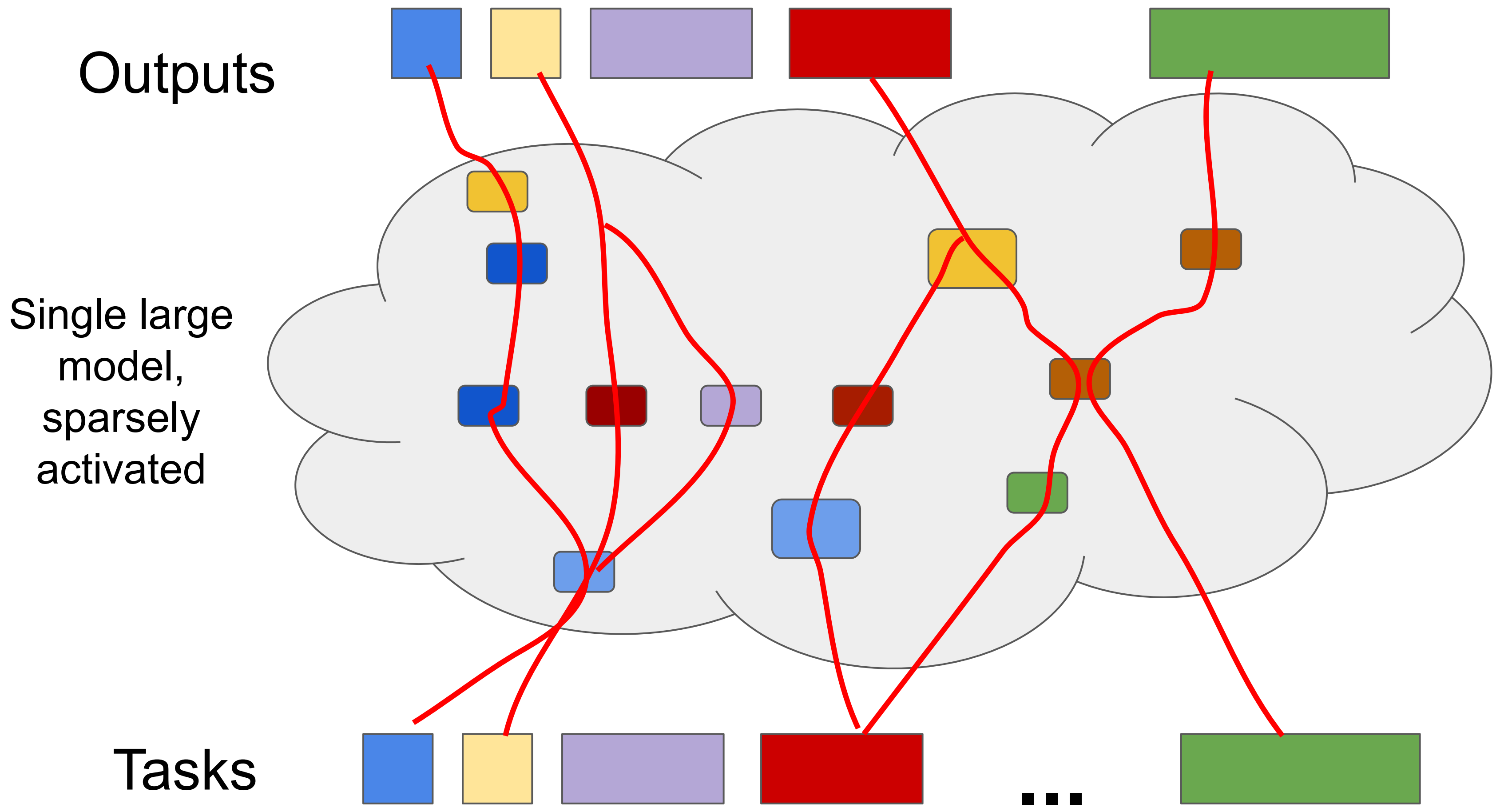
Dynamically learn and **grow pathways** through large model

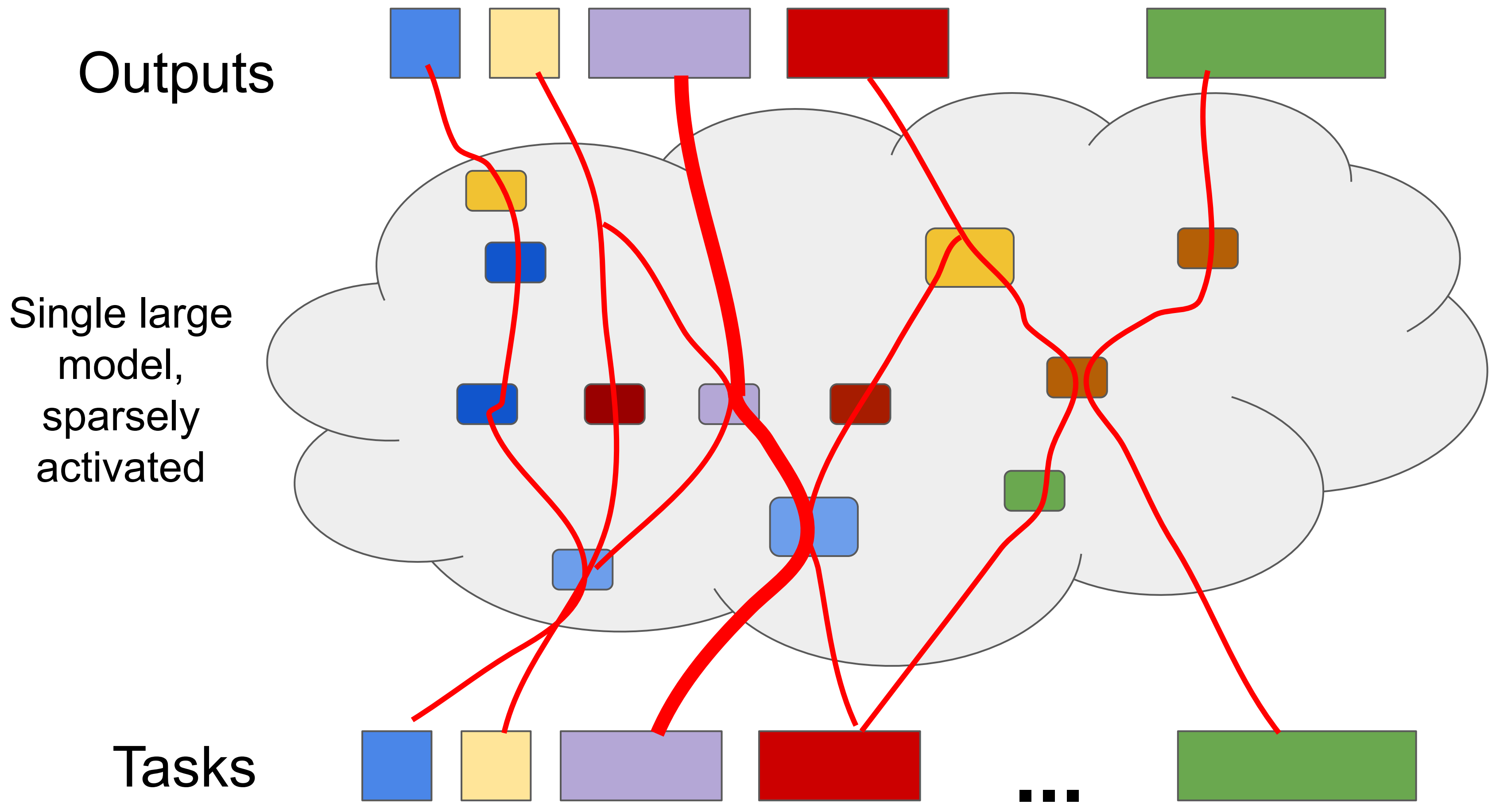
Hardware **specialized for ML supercomputing**

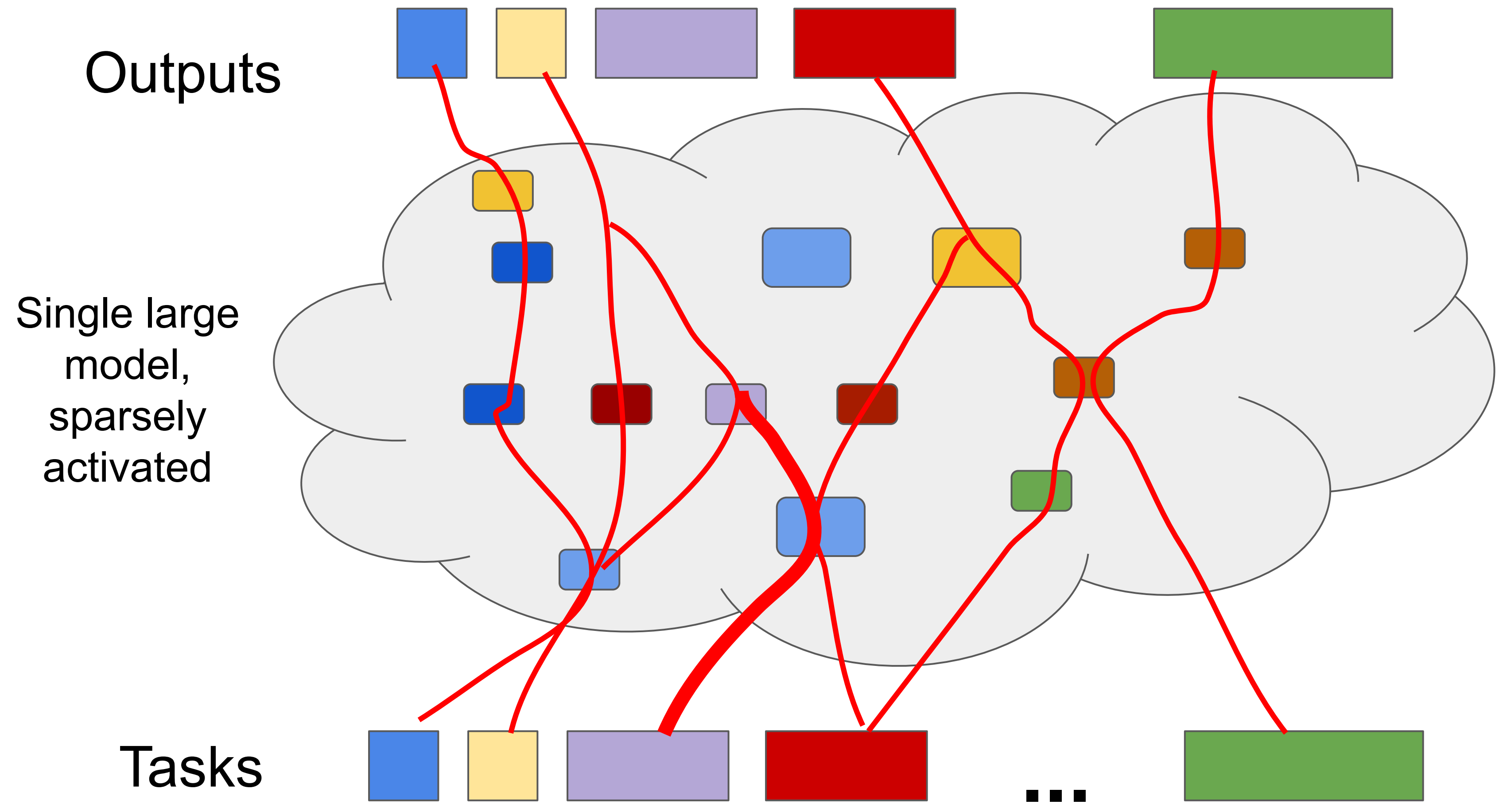
ML for efficient mapping onto this hardware

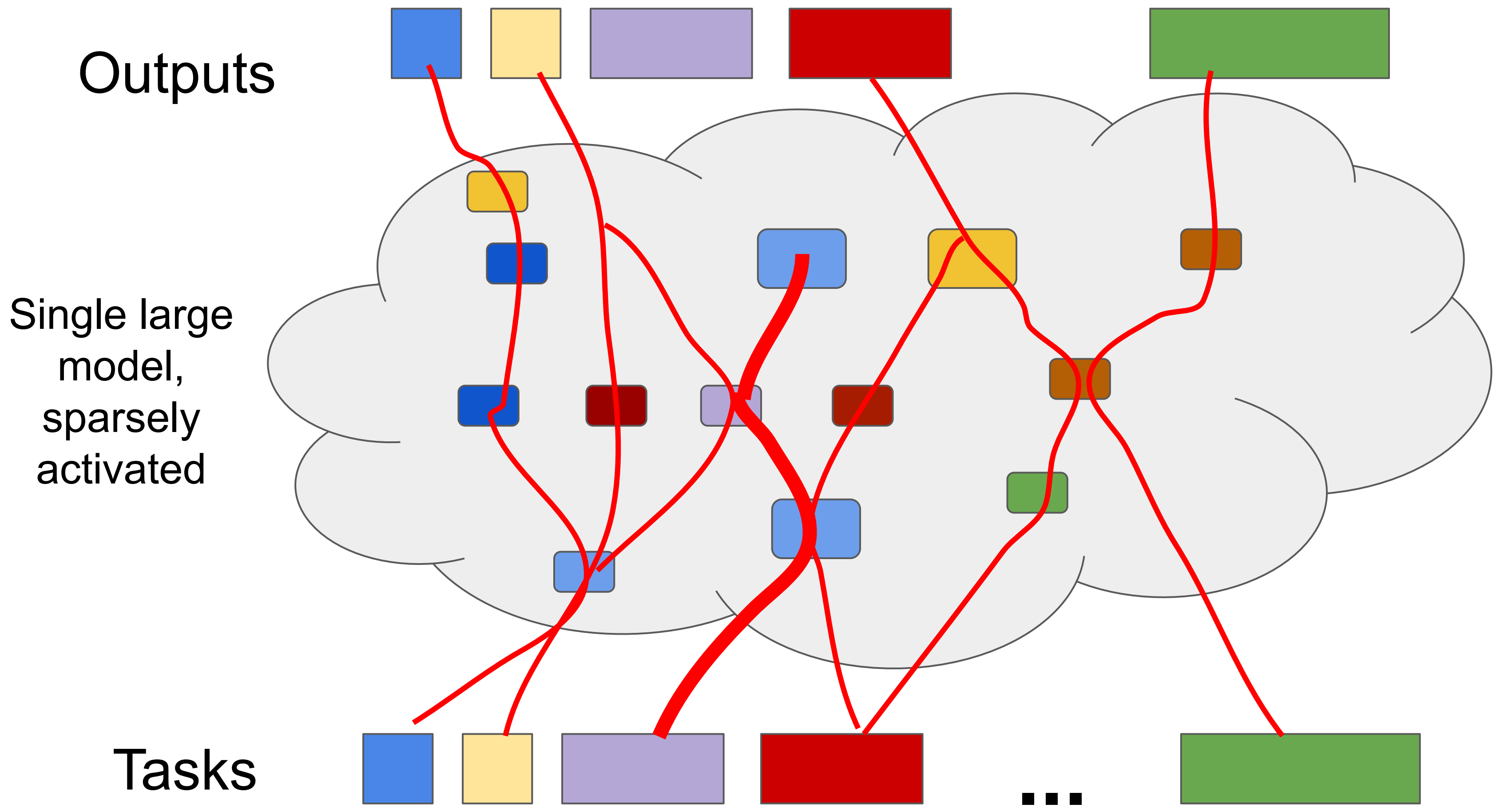


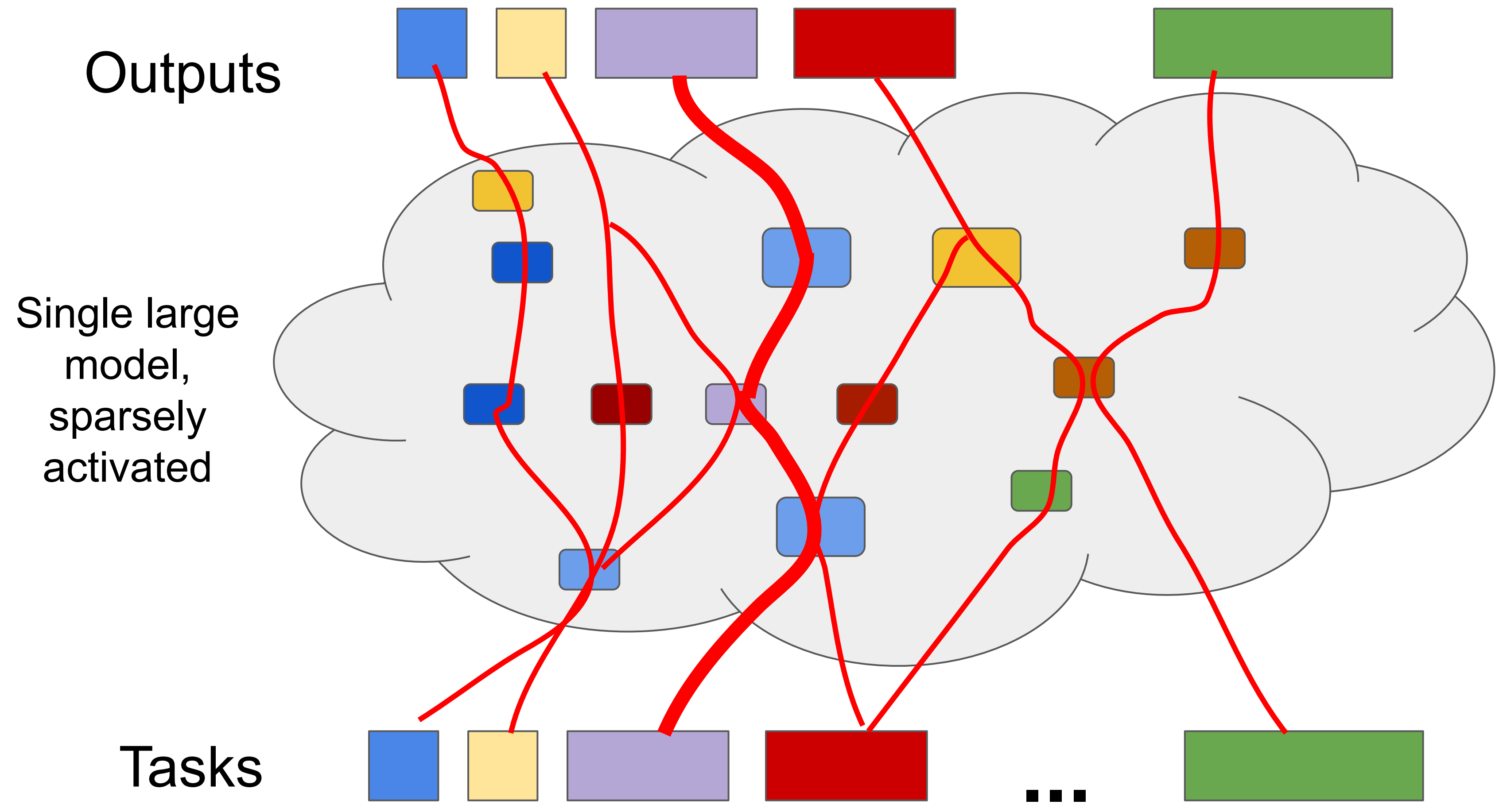
Outputs
Single large model, sparsely activated
Tasks
...

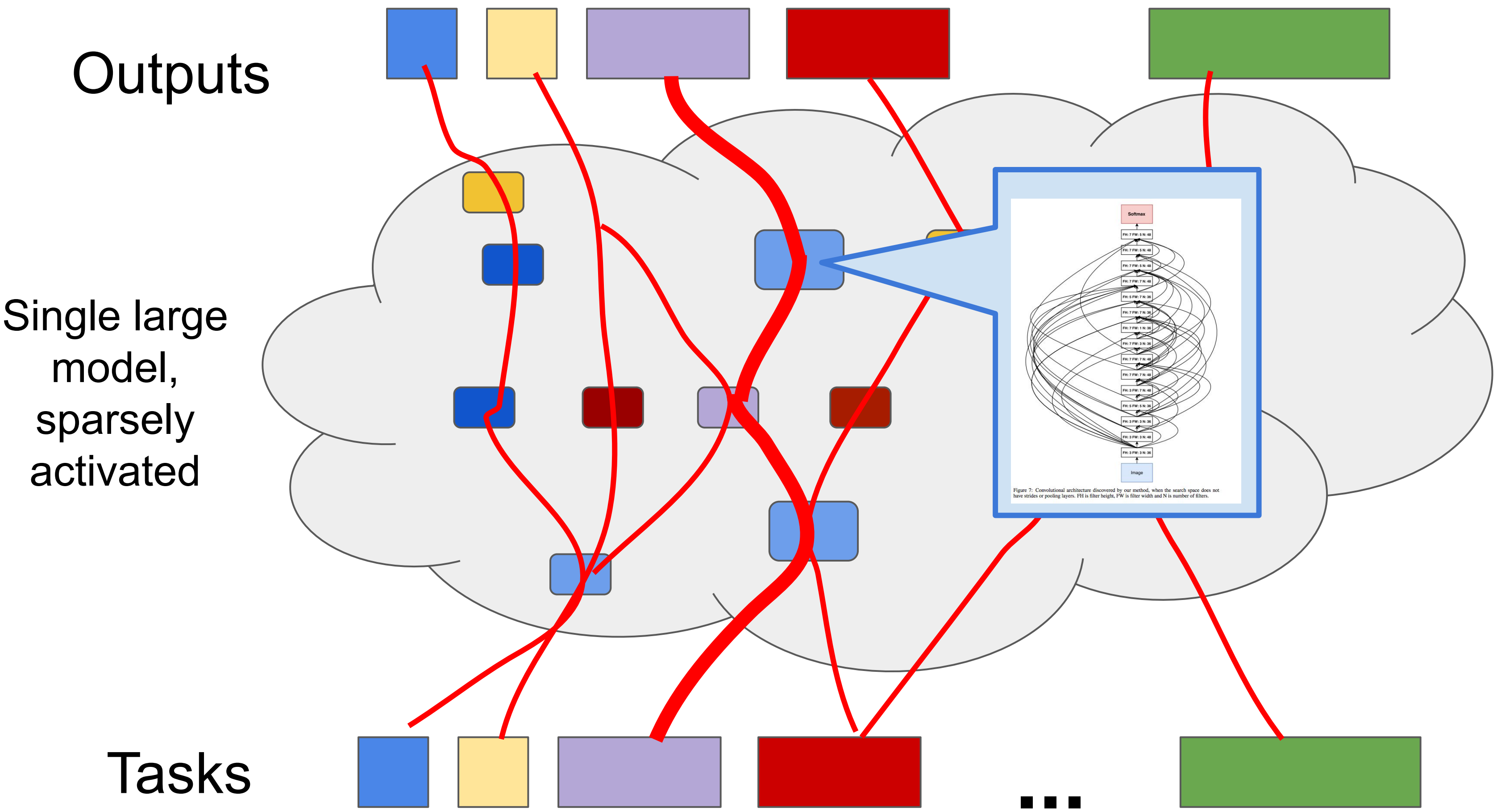

Outputs
Single large model, sparsely activated
Tasks
...


Outputs
Single large model, sparsely activated
Tasks
...


Outputs
Single large model, sparsely activated
Tasks
...


Outputs
Single large model, sparsely activated
Tasks
...


Outputs
Single large model, sparsely activated
Tasks
...


Outputs
Single large model, sparsely activated
Tasks
...
Softmax
Image
Figure 7: Convolutional architecture discovered by our method, when the search space does not have strides or pooling layers. FH is filter height, FW is filter width and N is number of filters.

Thoughtful use of AI in Society

AI at Google: our principles

Sundar Pichai
CEO

Published Jun 7,

At its heart, AI is computer programming that learns and adapts. It can't solve every problem, but its potential to improve our lives is profound. At Google, we use AI to make products more useful—from email that's spam-free and easier to compose, to a digital assistant you can speak to naturally, to photos that pop the fun stuff out for you to enjoy.

Beyond our products, we're using AI to help people tackle urgent problems. A pair of high school students are building AI-powered sensors to predict the risk of wildfires. Farmers are using it to monitor the health of their herds. Doctors are starting to use AI to help diagnose cancer and prevent blindness. These clear benefits are why Google invests heavily in AI research and development, and makes AI technologies widely available to others via our tools and open-source code.

We recognize that such powerful technology raises equally powerful questions about its use. How AI is developed and used will have a significant impact on society for many years to come. As a leader in AI, we feel a deep responsibility to get this right. So today, we're announcing seven principles to guide our work going forward. These are not theoretical concepts; they are concrete standards that will actively govern our research and product development and will impact our business decisions.

1. Be socially beneficial.
2. Avoid creating or reinforcing unfair bias.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate privacy design principles.
6. Uphold high standards of scientific excellence.
7. Be made available for uses that accord with these principles.

https://ai.google/principles

Machine Learning Fairness

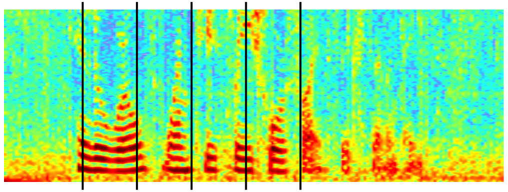
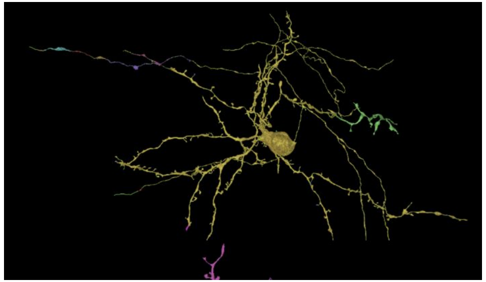
- Text Embedding Models Contain Bias. Here's Why That Matters. (Packer et al., Google 2018)
- Measuring and Mitigating Unintended Bias in Text Classification (Dixon et al., AIES 2018)
    - Exercise demonstrating Pinned AUC metric
- Mitigating Unwanted Biases with Adversarial Learning (Zhang et al., AIES 2018)
    - Exercise demonstrating Mitigating Unwanted Biases with Adversarial Learning
- Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations (Beutel et al., FAT/ML 2017)
- No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World (Shankar et al., NIPS 2017 workshop)
- Equality of Opportunity in Supervised Learning (Hardt et al., NIPS 2016)
- Satisfying Real-world Goals with Dataset Constraints (Goh et al., NIPS 2016)
- Designing Fair Auctions:
    - Fair Resource Allocation in a Volatile Marketplace (Bateni et al. EC 2016)
    - Reservation Exchange Markets for Internet Advertising (Goel et al., LIPics 2016)
- The Reel Truth: Women Aren't Seen or Heard (Geena Davis Inclusion Quotient)

https://developers.google.com/machine-learning/fairness-overview/

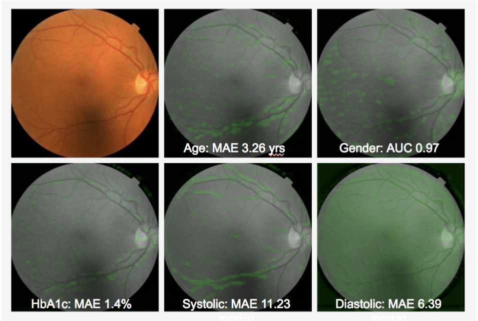
Google AI

Conclusions

Deep neural networks and machine learning are helping to make headway on some of the world's grand challenges





Thank you! More info about our work at ai.google/research

We're hiring! ai.google/research/join-us/

2018 overview: ai.googleblog.com/2019/01/looking-back-at-googles-research.html