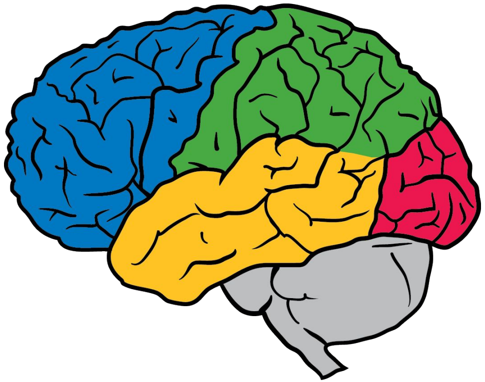


Large-Scale Deep Learning for Intelligent Computer Systems

Jeff Dean

In collaboration with **many** other people at Google

“Web Search and Data Mining”

“Web Search and Data Mining”

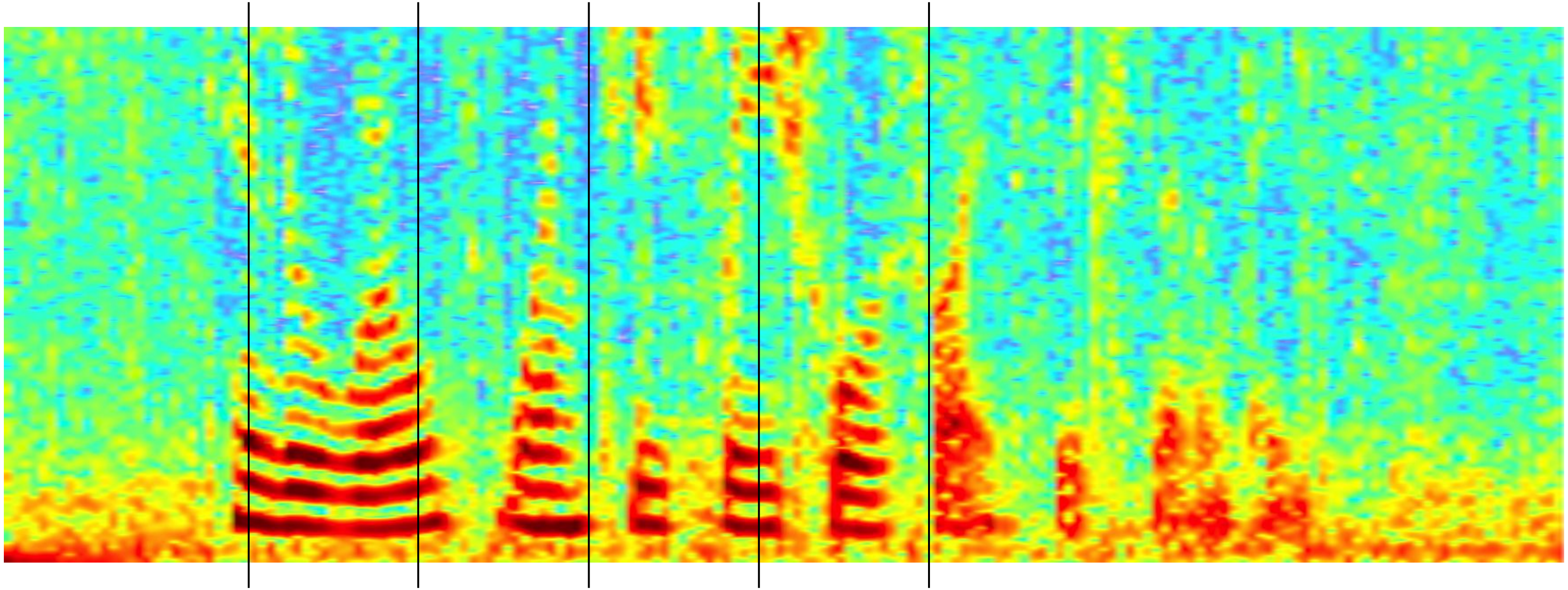
Really hard without **understanding**

Not there yet, but making significant progress

What do I mean by understanding?

What do I mean by understanding?

What do I mean by understanding?



What do I mean by understanding?

Query

[car parts for sale]

What do I mean by understanding?

Query

[car parts for sale]

Document 1

… car parking available for a small fee.
… parts of our floor model inventory for sale.

Document 2

Selling all kinds of automobile and pickup truck parts, engines, and transmissions.

Outline

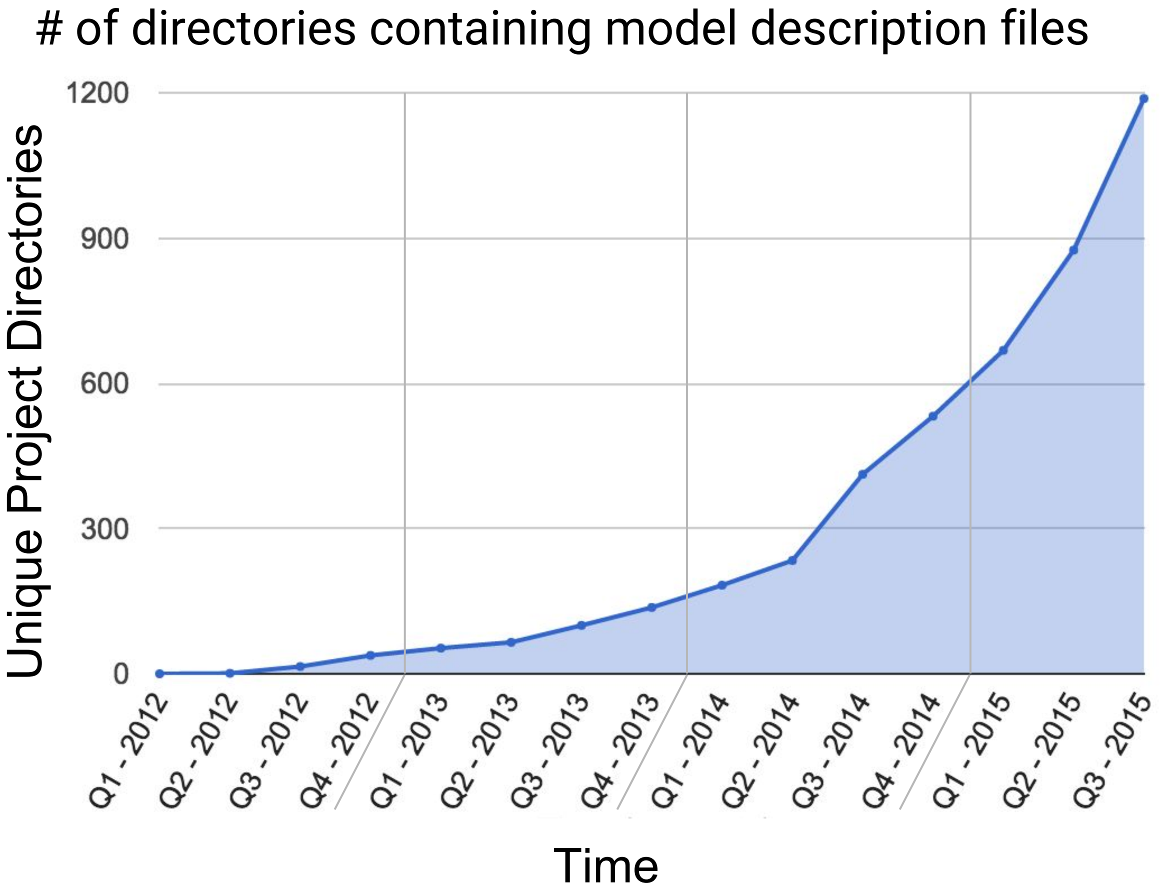
- Why deep neural networks?
- Perception
- Language understanding
- TensorFlow: software infrastructure for our work (and yours!)

Google Brain project started in 2011, with a focus on pushing state-of-the-art in neural networks. Initial emphasis:

- use large datasets, and
- large amounts of computation

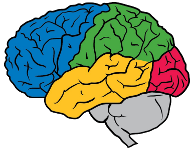
to push boundaries of what is possible in perception and language understanding

Growing Use of Deep Learning at Google



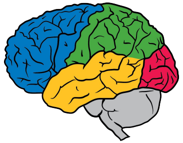
Across many products/areas:

Android
Apps
drug discovery
Gmail
Image understanding
Maps
Natural language understanding
Photos
Robotics research
Speech
Translation
YouTube
… many others ...



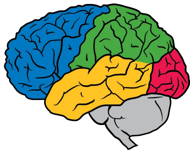
The promise (or wishful dream) of Deep Learning

Speech
Text
Search Queries
Images
Videos
Labels
Entities
Words
Audio
Features



Simple,
Reconfigurable,
High Capacity,
Trainable end-to-end
Building Blocks

Speech
Text
Search Queries
Images
Videos
Labels
Entities
Words
Audio
Features



The promise (or wishful dream) of Deep Learning

Common representations across domains.

Replacing piles of code with **data and learning**.

Would merely be an interesting academic exercise…

…**if it didn't work so well!**

In Research and Industry

Speech Recognition

Speech Recognition with Deep Recurrent Neural Networks
Alex Graves, Abdel-rahman Mohamed, Geoffrey Hinton

Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks
Tara N. Sainath, Oriol Vinyals, Andrew Senior, Hasim Sak

Object Recognition and Detection

Going Deeper with Convolutions
Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed,
Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich

Scalable Object Detection using Deep Neural Networks
Dumitru Erhan, Christian Szegedy, Alexander Toshev, Dragomir Anguelov

In Research and Industry

Machine Translation

Sequence to Sequence Learning with Neural Networks
Ilya Sutskever, Oriol Vinyals, Quoc V. Le

Neural Machine Translation by Jointly Learning to Align and Translate
Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio

Language Modeling

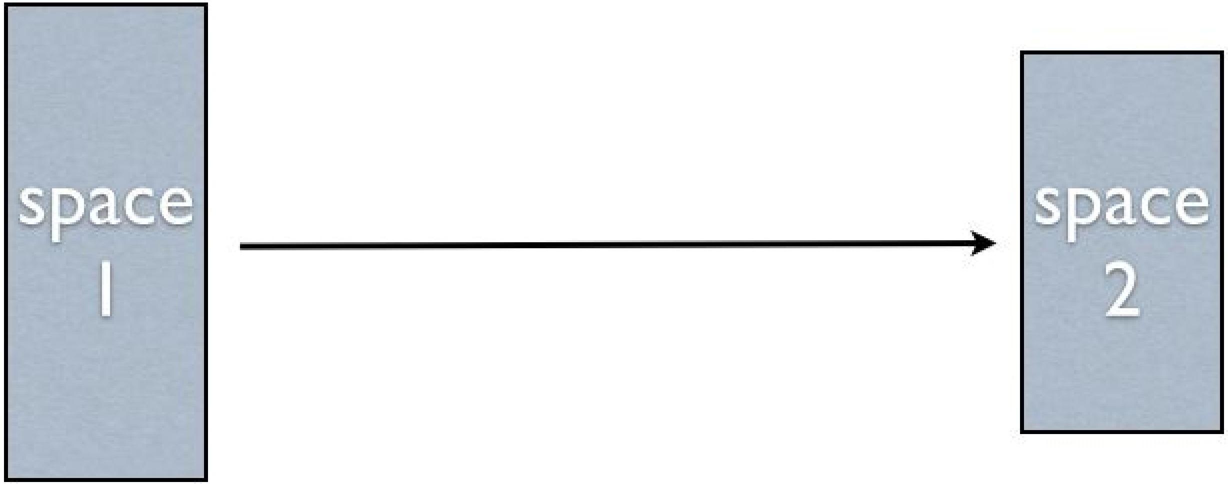
One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling
Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, Tony Robinson

Parsing

Grammar as a Foreign Language
Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, Geoffrey Hinton

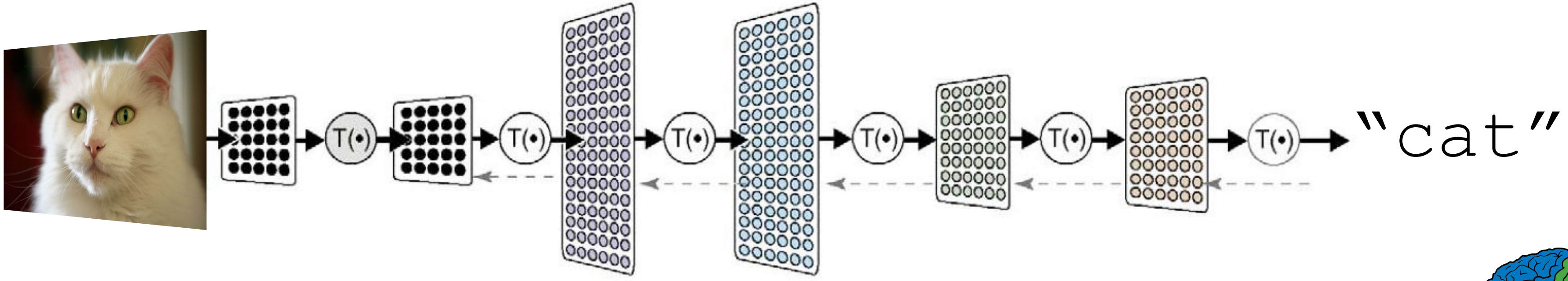
Neural Networks

- Learn a complicated function from data



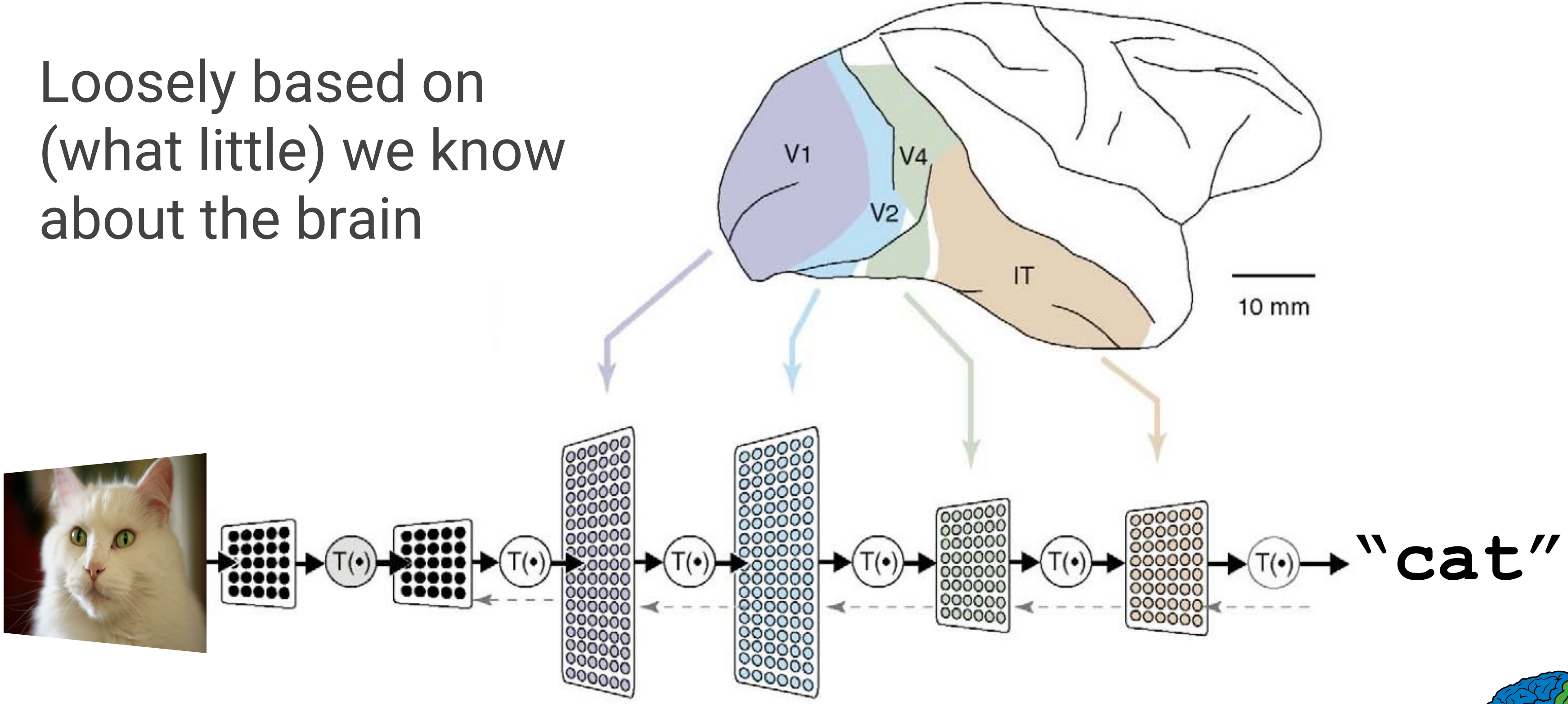
What is Deep Learning?

- A powerful class of machine learning model
- Modern reincarnation of artificial neural networks
- Collection of simple, trainable mathematical functions
- Compatible with many variants of machine learning



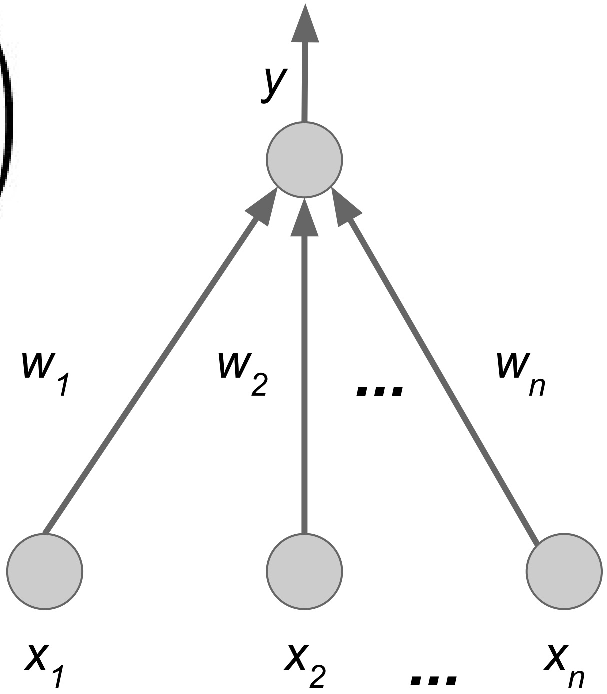
What is Deep Learning?

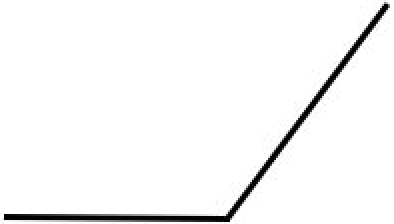
- Loosely based on (what little) we know about the brain



The Neuron

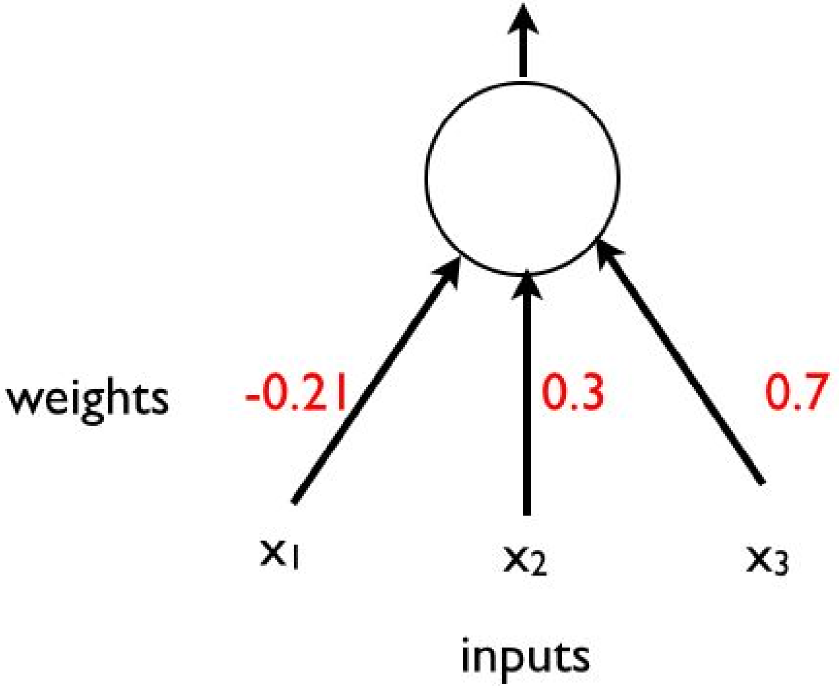
$$y = F\left(\sum_i w_i x_i\right)$$

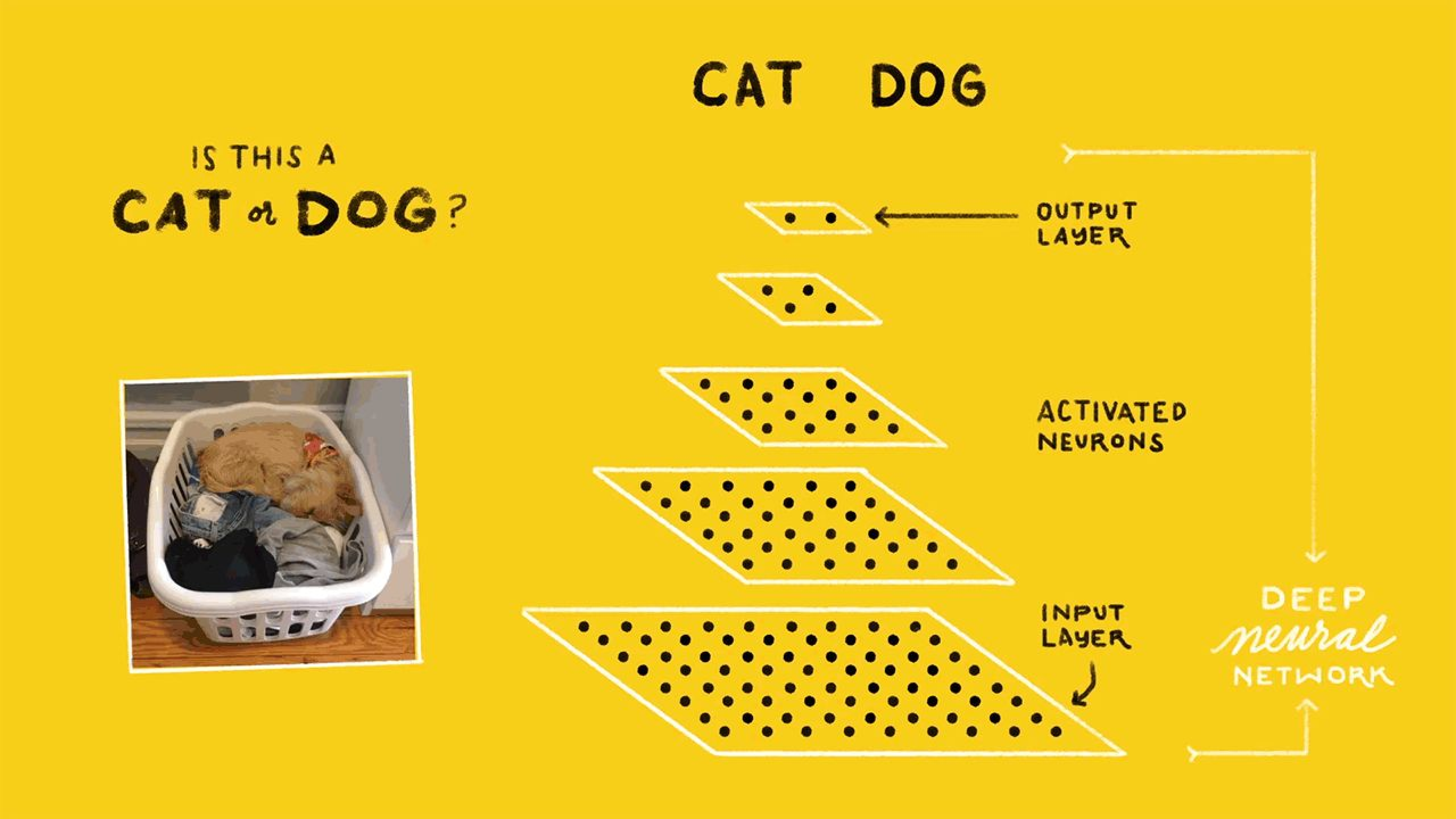




$$F(x) = \max(0, x)$$

$y = \max(0, -0.21 * x_1 + 0.3 * x_2 + 0.7 * x_3)$




IS THIS A
CAT or DOG?
CAT DOG
OUTPUT LAYER
ACTIVATED NEURONS
INPUT LAYER
DEEP neural NETWORK

Learning algorithm

While not done:

Pick a random training example "(input, label)"

Run neural network on "input"

Adjust weights on edges to make output closer to "label"

Learning algorithm

While not done:

Pick a random training example “(input, label)”

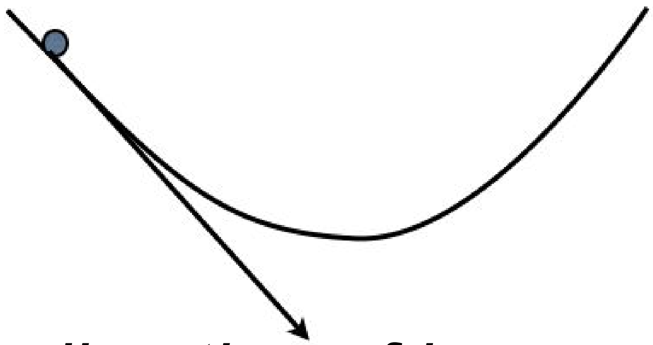
Run neural network on “input”

Adjust weights on edges to make output closer to “label”

Backpropagation

Use partial derivatives along the paths in the neural net

Follow the gradient of the error w.r.t. the connections

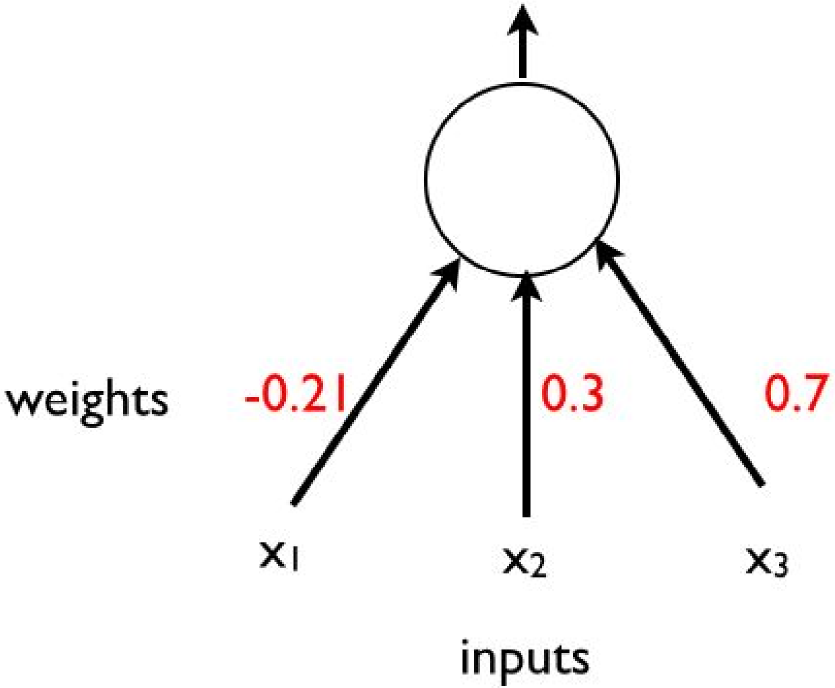


Gradient points in direction of improvement

Good description: **"Calculus on Computational Graphs: Backpropagation"**

[*http://colah.github.io/posts/2015-08-Backprop/*](http://colah.github.io/posts/2015-08-Backprop/)

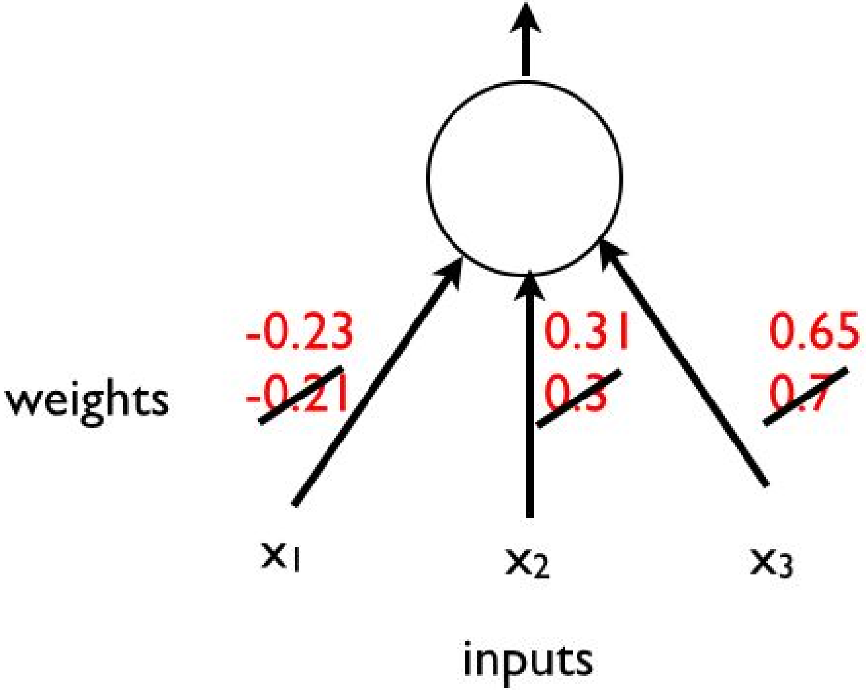
$$y = \max(0, -0.21 * x_1 + 0.3 * x_2 + 0.7 * x_3)$$

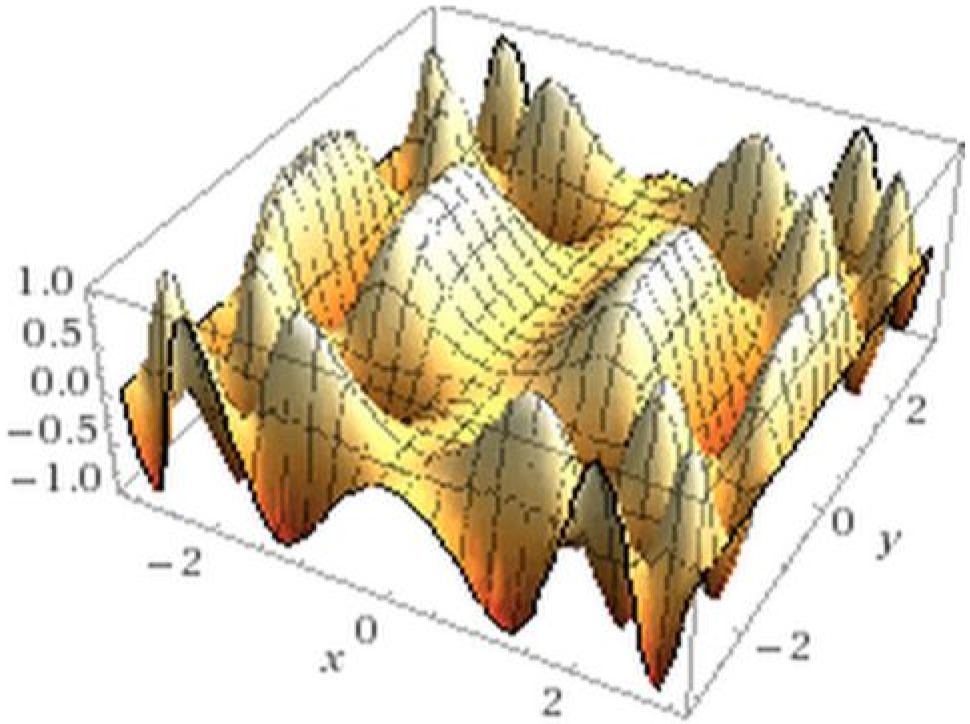


next time:

$\text{output} = \max(0, -0.23 * x_1 + 0.31 * x_2 + 0.65 * x_3)$

$\cancel{\text{output} = \max(0, -0.21 * x_1 + 0.3 * x_2 + 0.7 * x_3)}$





This shows a function of 2 variables: real neural nets are functions of hundreds of millions of variables!

Plenty of raw data

- **Text**: trillions of words of English + other languages
- **Visual data**: billions of images and videos
- **Audio**: tens of thousands of hours of speech per day
- **User activity**: queries, marking messages spam, etc.
- **Knowledge graph**: billions of labelled relation triples
- ...

How can we build systems that truly understand this data?

Important Property of Neural Networks

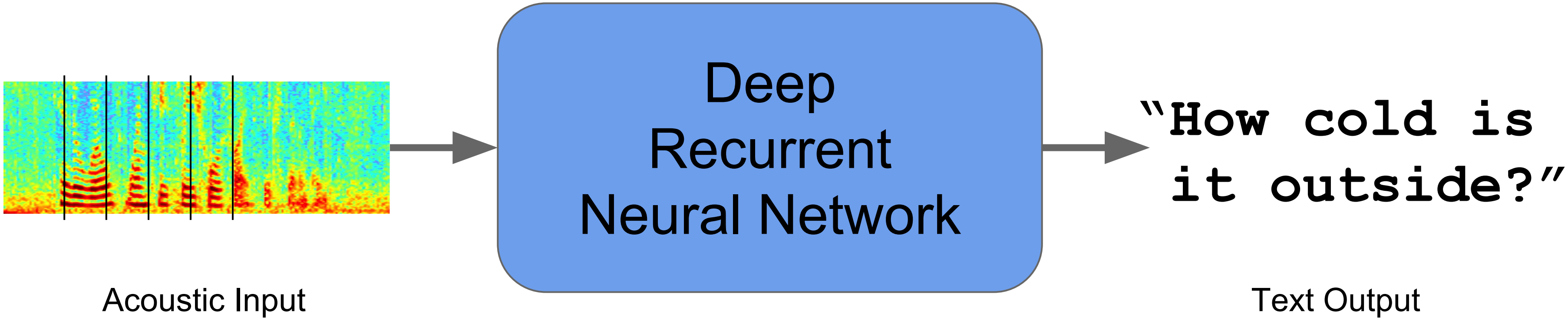
Results get better with

more data +
bigger models +
more computation

(Better algorithms, new insights and improved techniques always help, too!)

What are some ways that
deep learning is having
a significant impact at Google?

Speech Recognition



Reduced word errors by more than 30%

Google Research Blog - August 2012, August 2015

ImageNet Challenge

Given an image, predict one of 1000 different classes

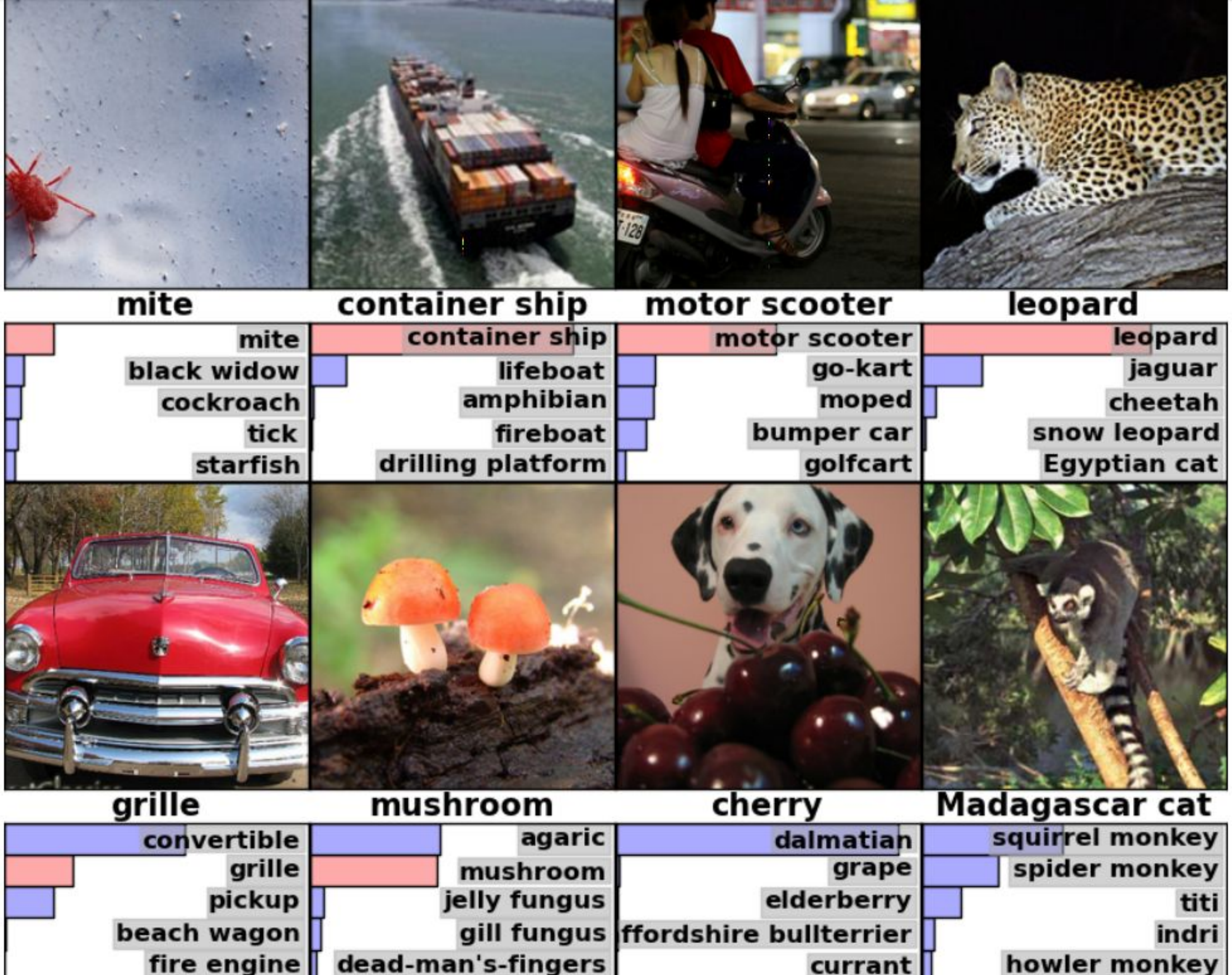
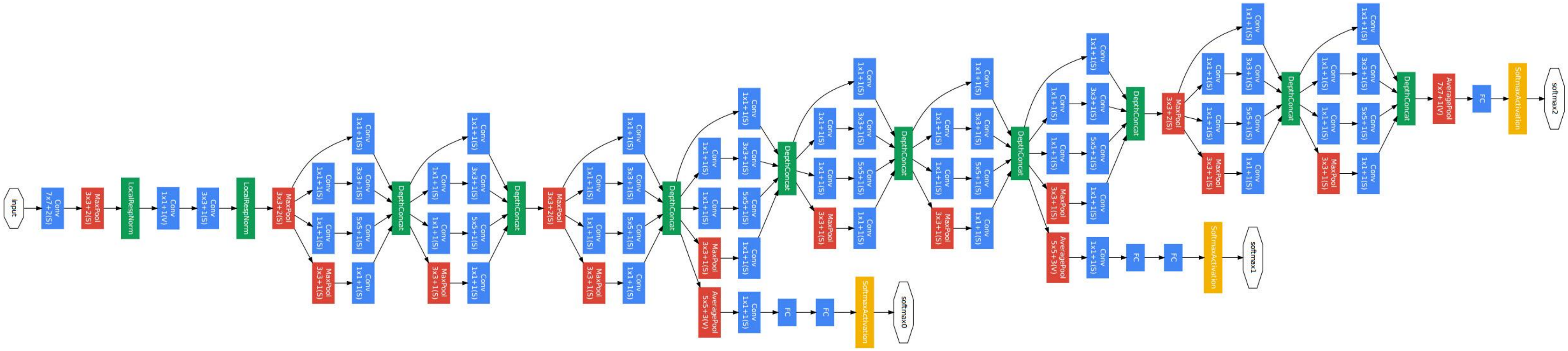


Image credit: www.cs.toronto.edu/~fritz/absps/imagenet.pdf

The Inception Architecture (GoogLeNet, 2014)



Going Deeper with Convolutions

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich

ArXiv 2014, CVPR 2015

Neural Nets: Rapid Progress in Image Recognition

Team	Year	Place	Error (top-5)
XRCE (pre-neural-net explosion)	2011	1st	25.8%
Supervision (AlexNet)	2012	1st	16.4%
Clarifai	2013	1st	11.7%
GoogLeNet (Inception)	2014	1st	6.66%
Andrej Karpathy (human)	2014	N/A	5.1%
BN-Inception (Arxiv)	2015	N/A	4.9%
Inception-v3 (Arxiv)	2015	N/A	3.46%

ImageNet challenge classification task

Good Fine-Grained Classification

"hibiscus"

"dahlia"

Good Generalization

Both recognized as “meal”

Sensible Errors

"snake"

"dog"

Google Photos Search

Your Photo

Deep Convolutional Neural Network

"ocean"

Automatic Tag

Search personal photos without tags.

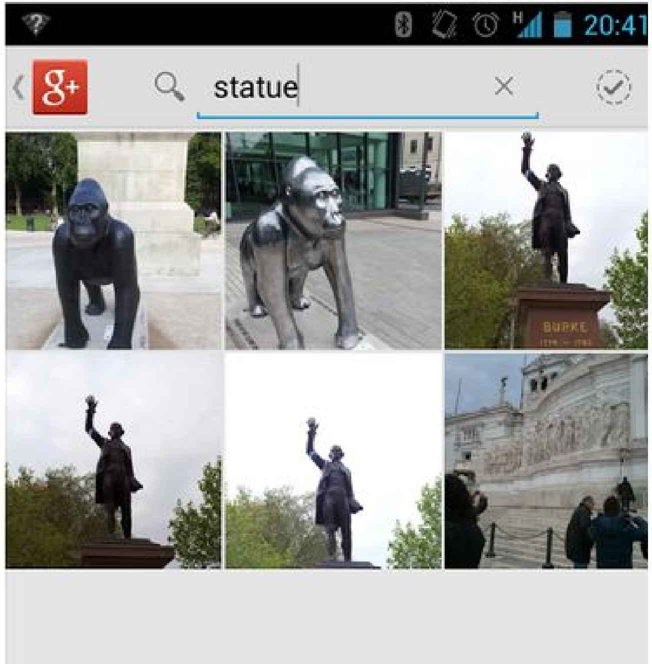
Google Research Blog - June 2013

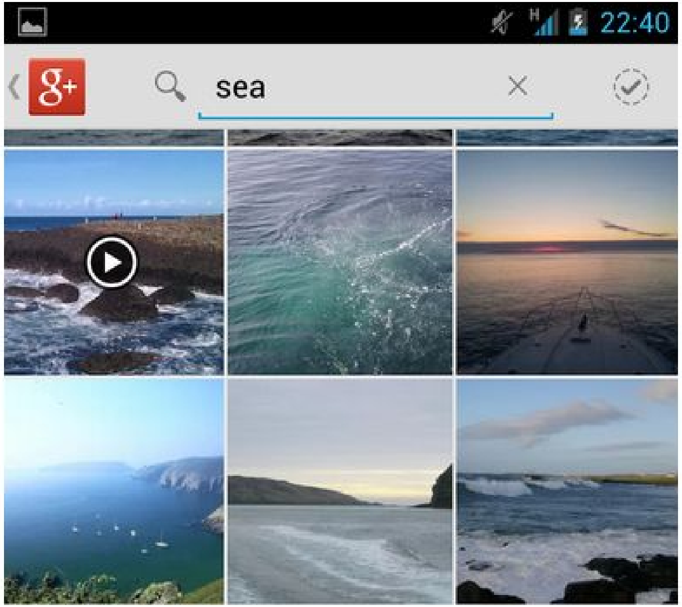
Google Photos Search

Wow.

The new Google plus photo search is a bit insane.

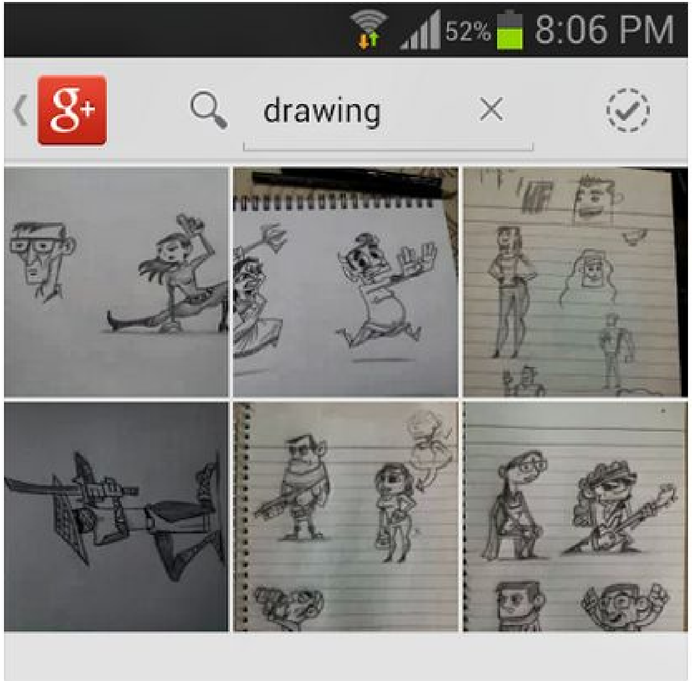
I didn't tag those... :)

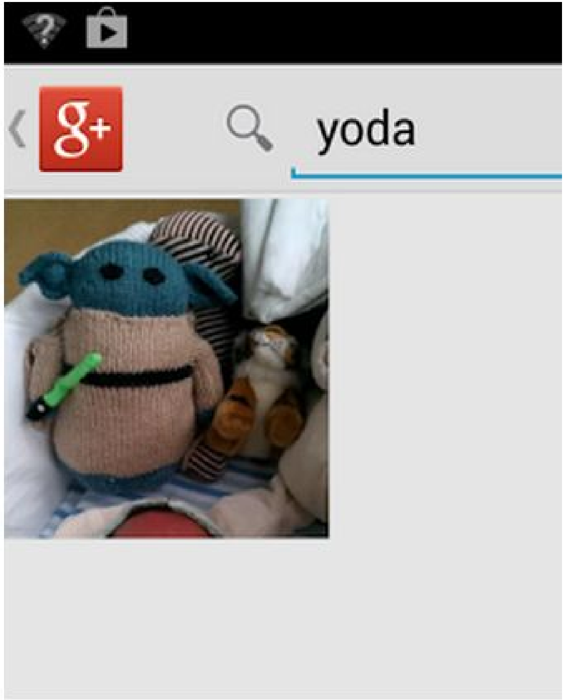




Google Photos Search

Google Plus photo search is awesome. Searched with keyword 'Drawing' to find all my scribbles at once :D






ASIAWIDE TRAVEL
(02) 9745 3355 1st Floor, 240 BURWOOD RD
aria's Bakery Inn 超羣餅屋
Maria's Bakery Inn 超羣餅屋


Corner
Cubbyhouse
CIANO MOTOR ENGINEERS
MECHANICAL REPAIRS TO ALL MAKES AND MODELS
Specialising In BMW, MINI & TOYOTA
8 REGATTA ROAD FIVE DOCK 9745 3173
88
LATEST DIAGNOSTIC EQUIPMENT • REGO INSPECTIONS
STEERING • SUSPENSION • TYRES • WHEEL ALIGNMENTS
FUEL INJECTION SERVICING • BATTERIES • AUTO ELECTRICAL
Factory Trained Technicians

Language Understanding

Query

[car parts for sale]

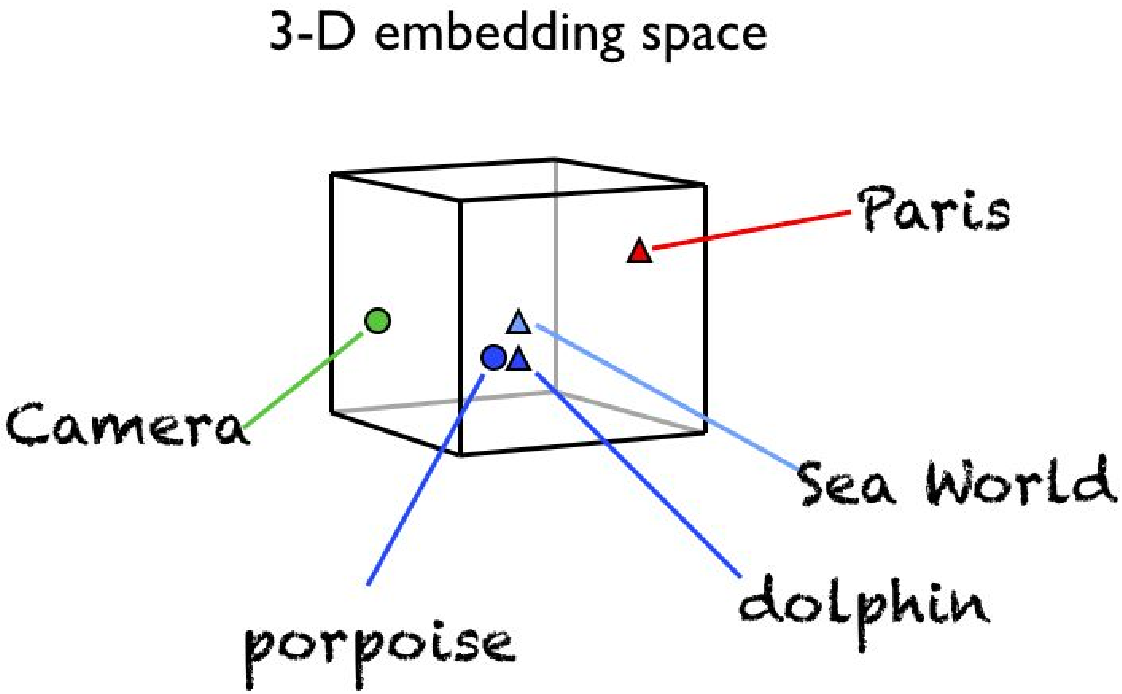
Document 1

… car parking available for a small fee.
… parts of our floor model inventory for sale.

Document 2

Selling all kinds of automobile and pickup truck parts, engines, and transmissions.

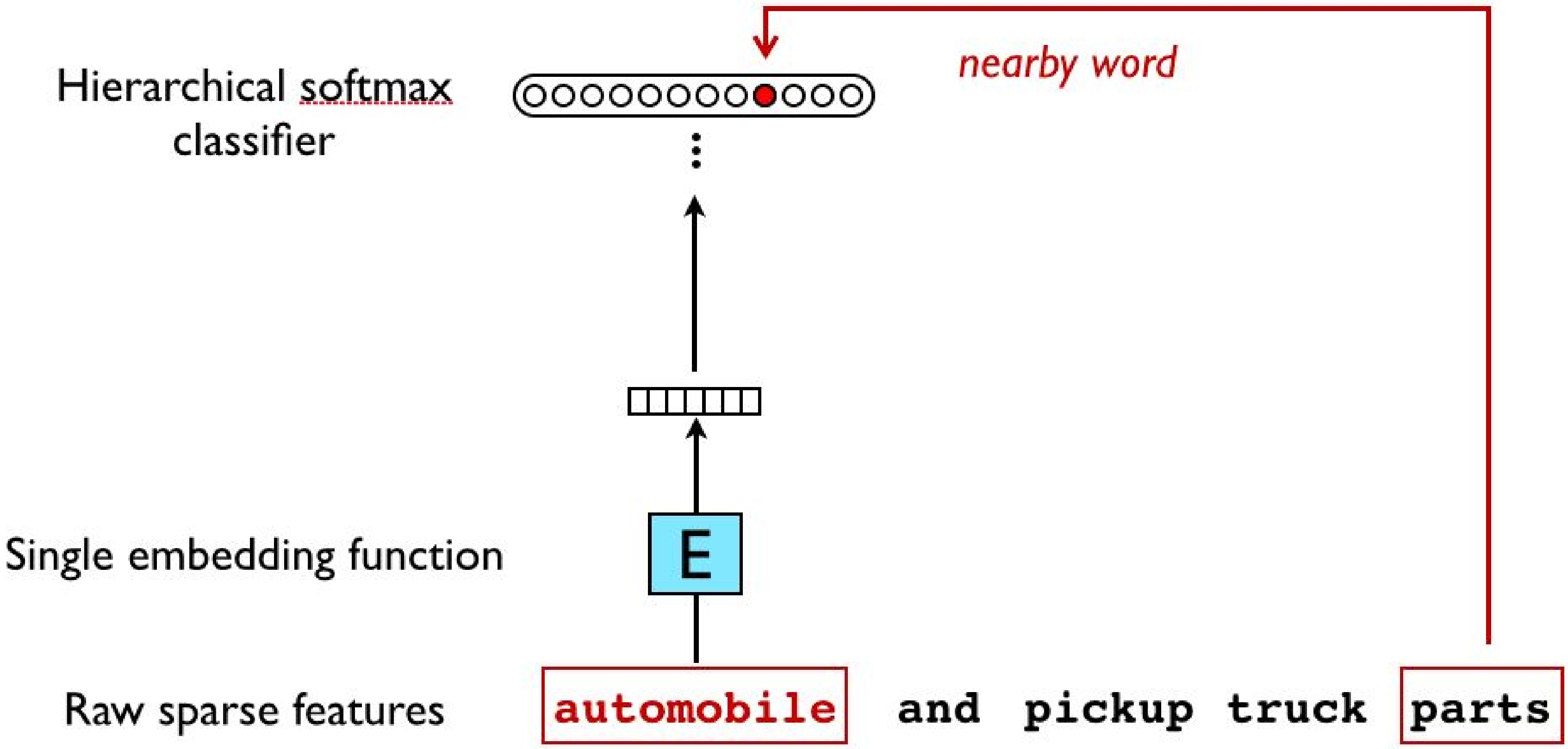
How to deal with Sparse Data?



Embedding Function: A look-up-table that maps sparse features into dense floating point vectors.

Usually use many more than 3 dimensions (e.g. 100D, 1000D)

Embeddings Can be Trained With Backpropagation



Mikolov, Sutskever, Chen, Corrado and Dean. *Distributed Representations of Words and Phrases and Their Compositionality*, NIPS 2013.

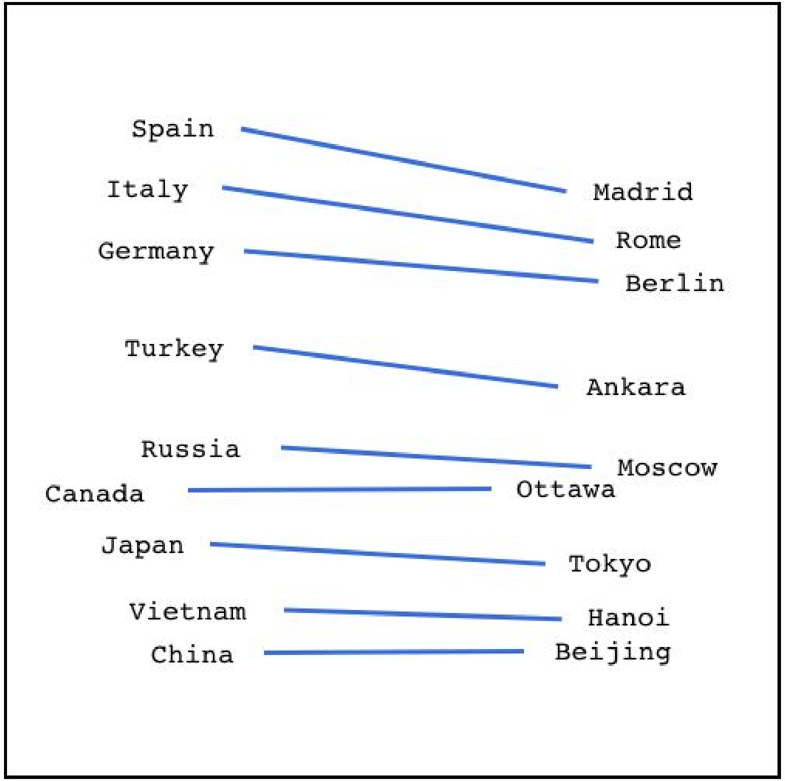
Nearest Neighbors are Closely Related Semantically

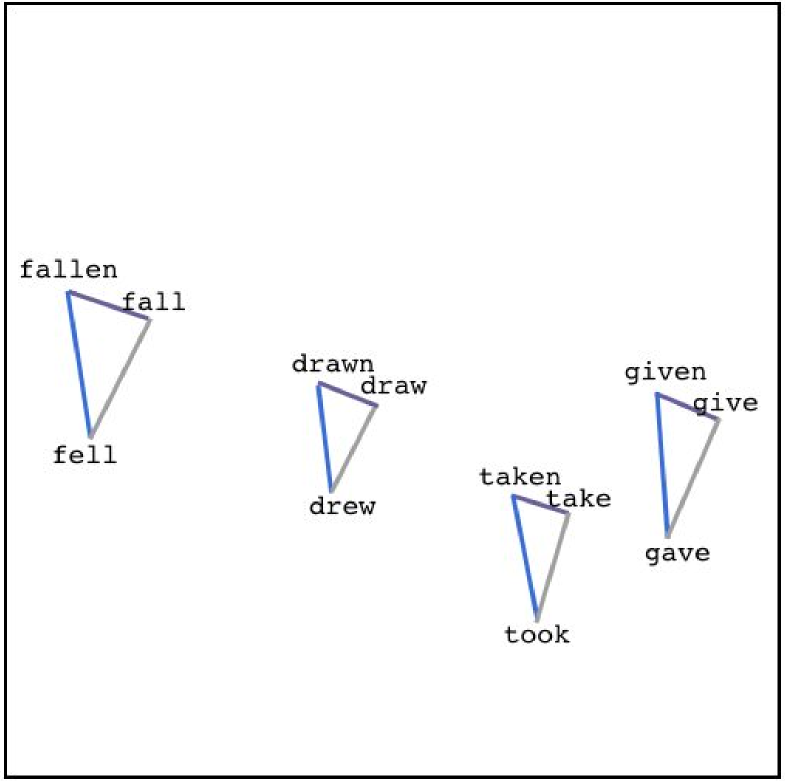
Trained language model on Wikipedia

tiger shark	**car**	**new york**
bull shark	cars	new york city
blacktip shark	muscle car	brooklyn
shark	sports car	long island
oceanic whitetip shark	compact car	syracuse
sandbar shark	autocar	manhattan
dusky shark	automobile	washington
blue shark	pickup truck	bronx
requiem shark	racing car	yonkers
great white shark	passenger car	poughkeepsie
lemon shark	dealership	new york state

* 5.7M docs, 5.4B terms, 155K unique terms, 500-D embeddings

Directions are Meaningful





Solve analogies with vector arithmetic!

$$V(\mathtt{queen}) - V(\mathtt{king}) \approx V(\mathtt{woman}) - V(\mathtt{man})$$

$$V(\mathtt{queen}) \approx V(\text{king}) + (V(\mathtt{woman}) - V(\mathtt{man}))$$

RankBrain in Google Search Ranking

`Query: "car parts for sale",`
`Doc: "Rebuilt transmissions …"`

Query & document features

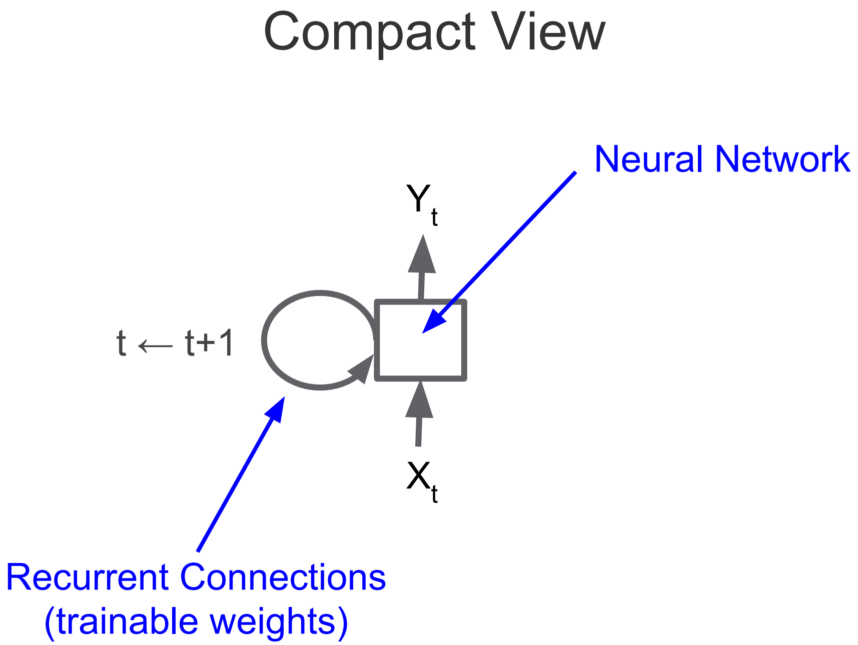
→ Deep Neural Network →

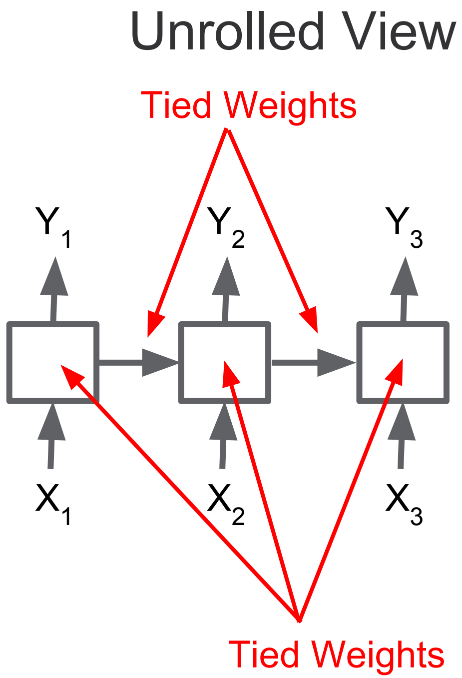
`Score for doc,query pair`

Launched in 2015
Third most important search ranking signal (of 100s)

Bloomberg, Oct 2015: "*Google Turning Its Lucrative Web Search Over to AI Machines*"

Recurrent Neural Networks





Recurrent Neural Networks

RNNs very difficult to train for more than a few timesteps: numerically unstable gradients (vanishing / exploding).

Thankfully, **LSTMs**… ["*Long Short-Term Memory*", Hochreiter & Schmidhuber, 1997]

LSTMs: Long Short-Term Memory Networks

‘RNNs done right’:

- Very effective at modeling long-term dependencies.
- Very sound theoretical and practical justifications.
- A central inspiration behind lots of recent work on using deep learning to learn complex programs:
  Memory Networks, *Neural Turing Machines*.

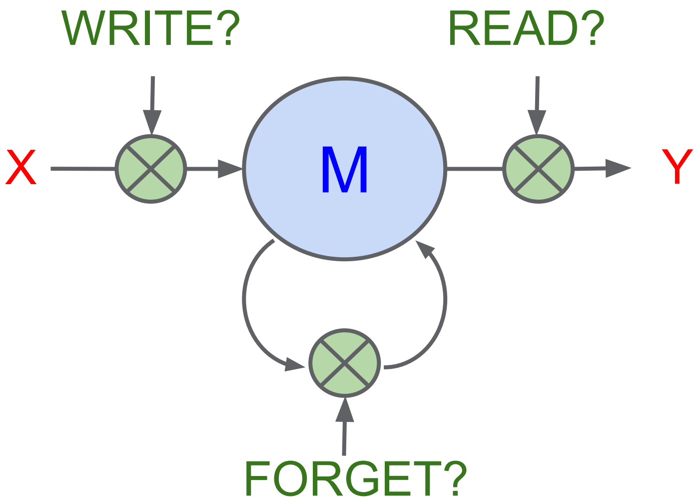
A Simple Model of Memory

Instruction
Input
Output

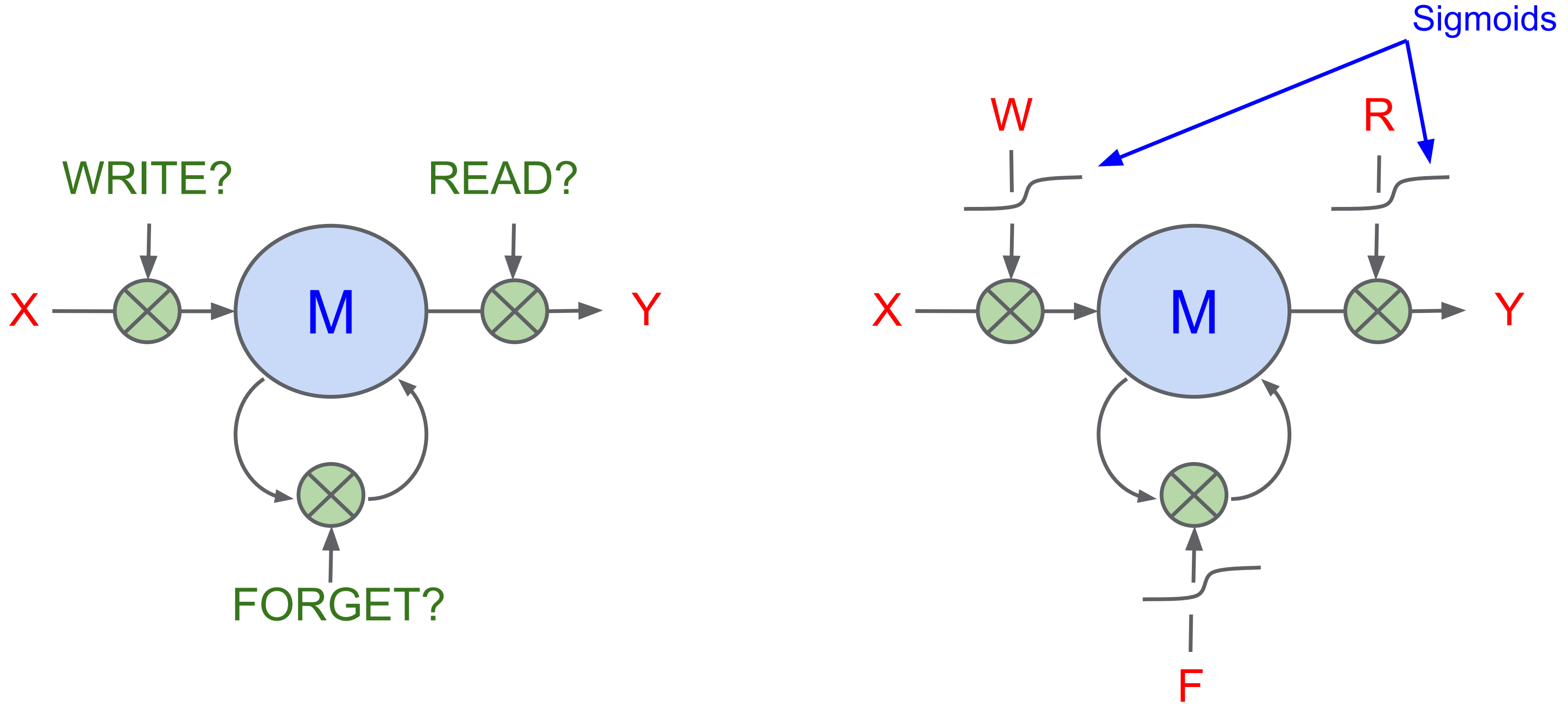
WRITE X, M

READ M, Y

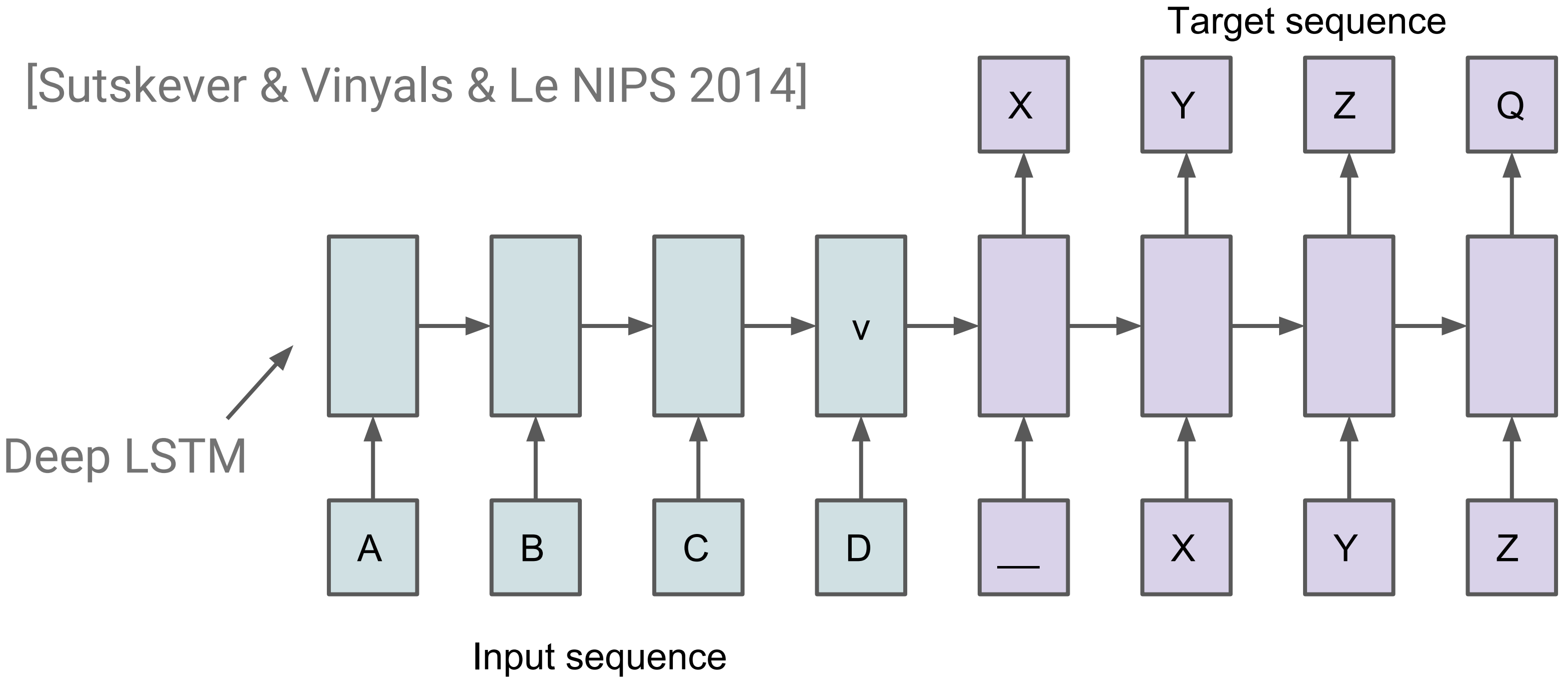
FORGET M



Key Idea: Make Your Program Differentiable



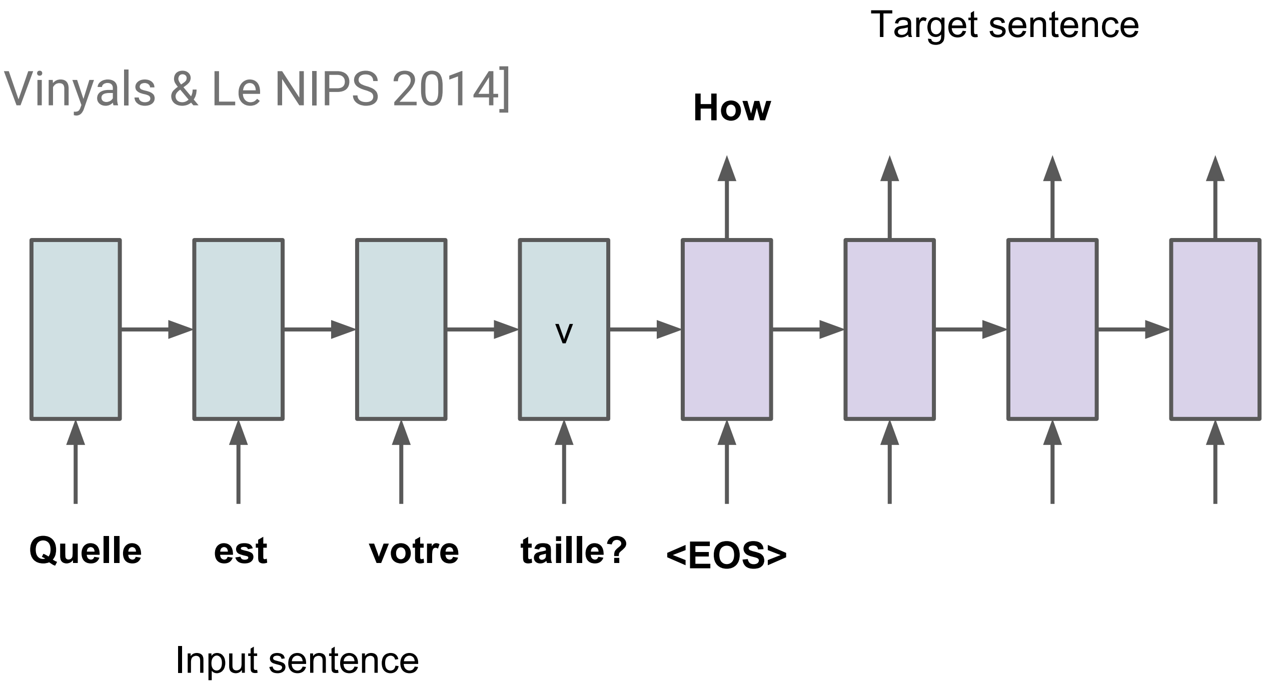
Sequence-to-Sequence Model



$$P(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$

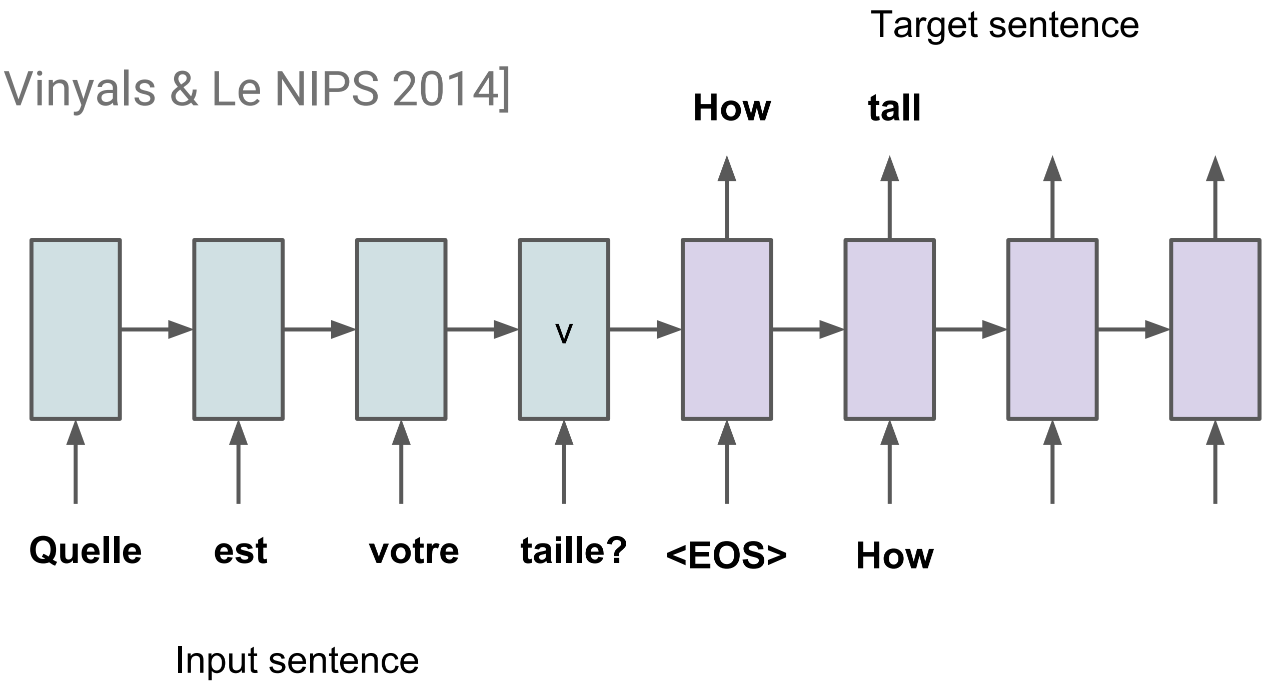
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



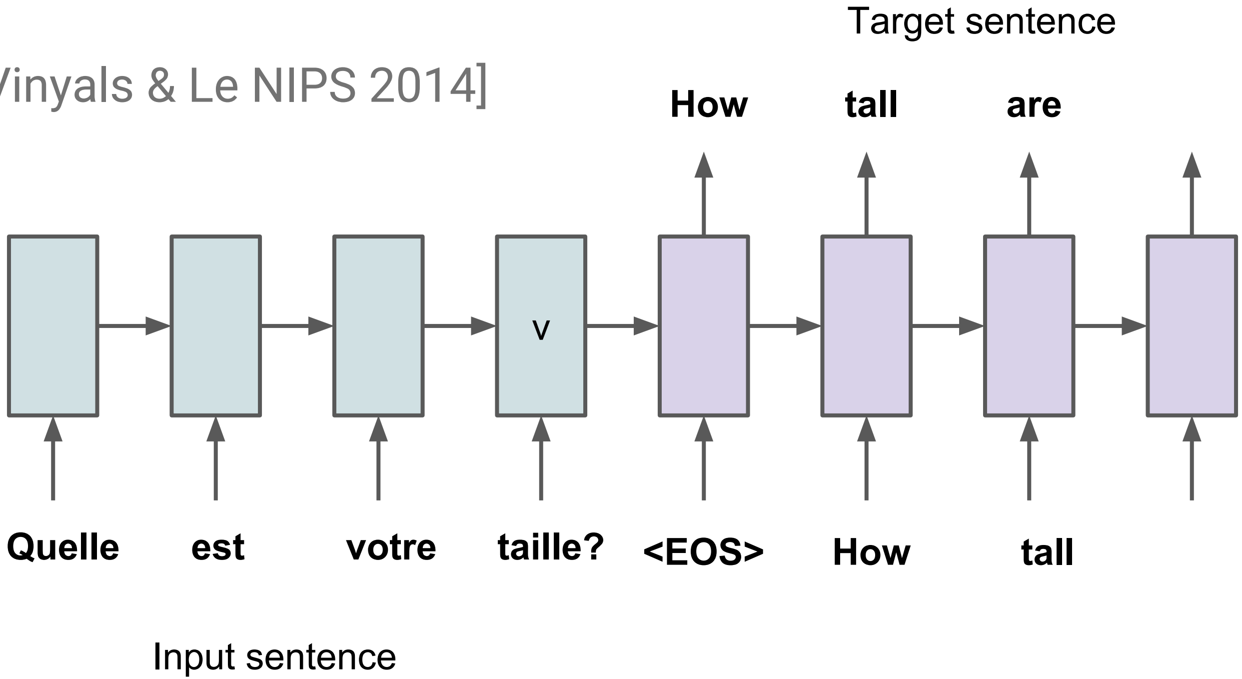
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



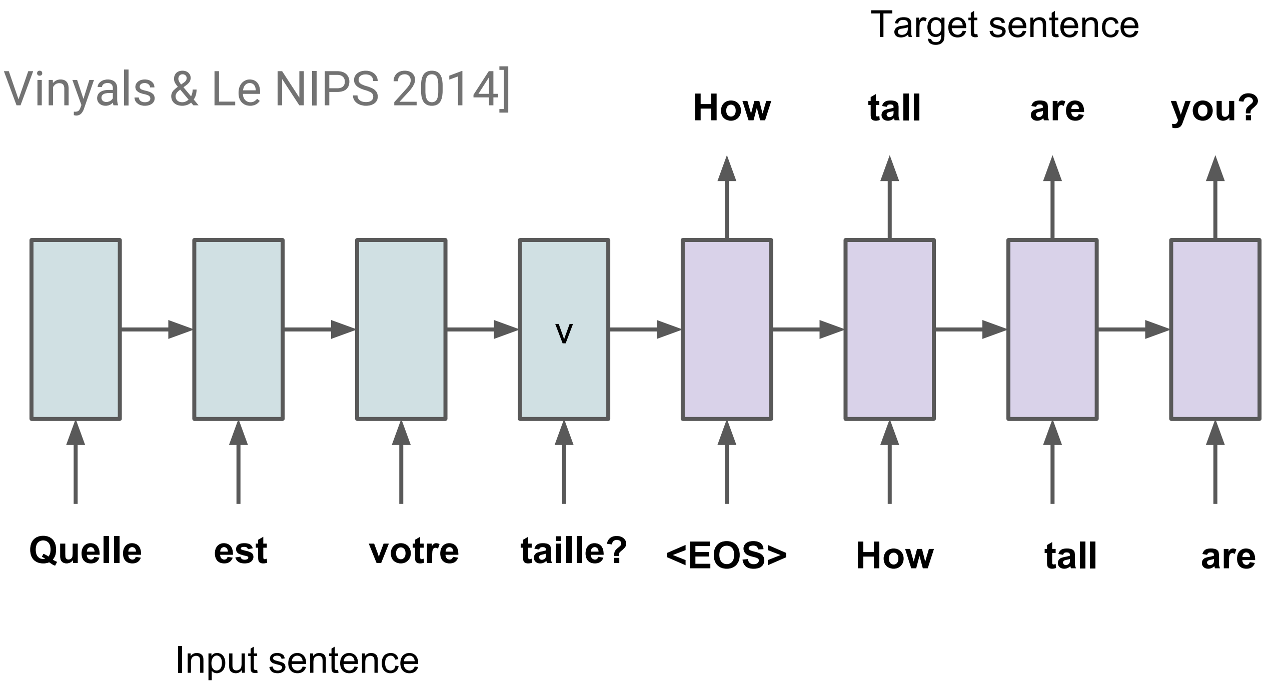
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



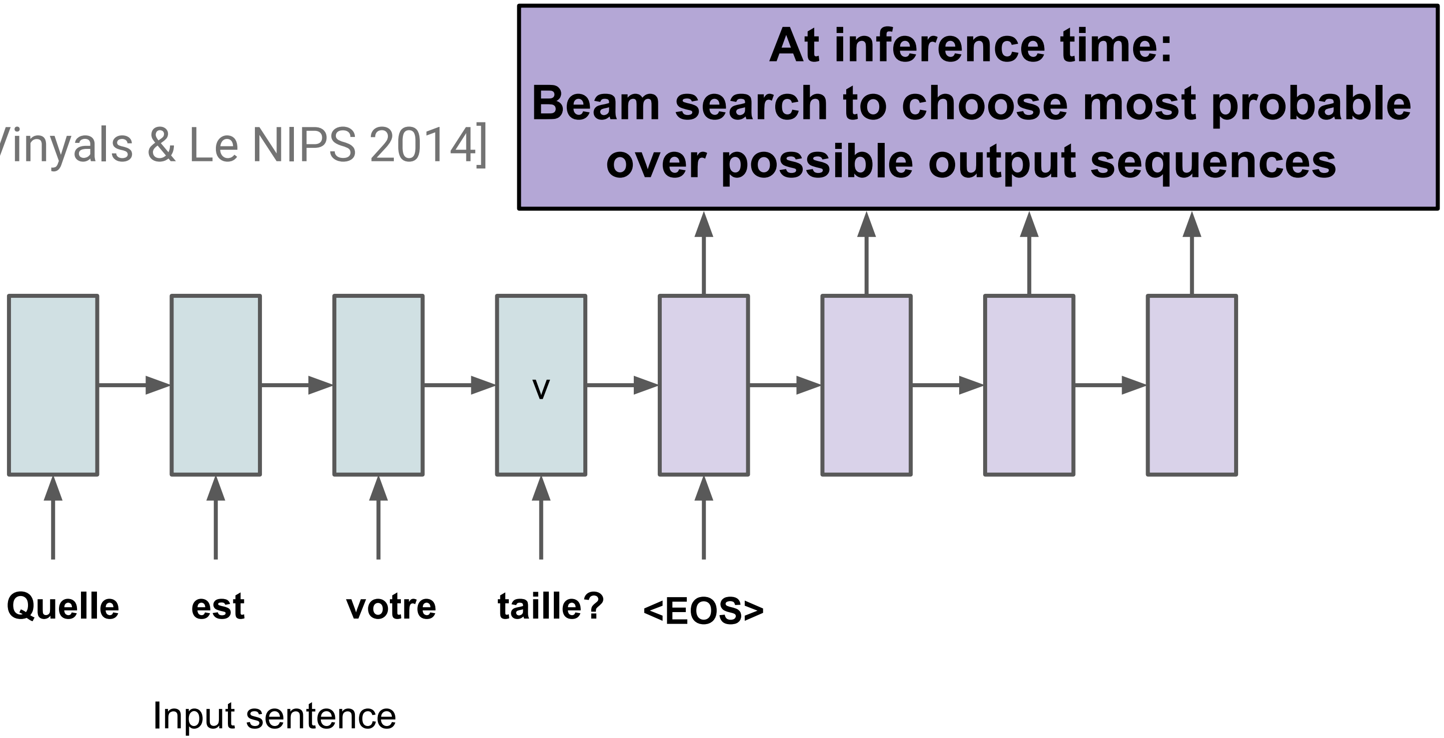
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



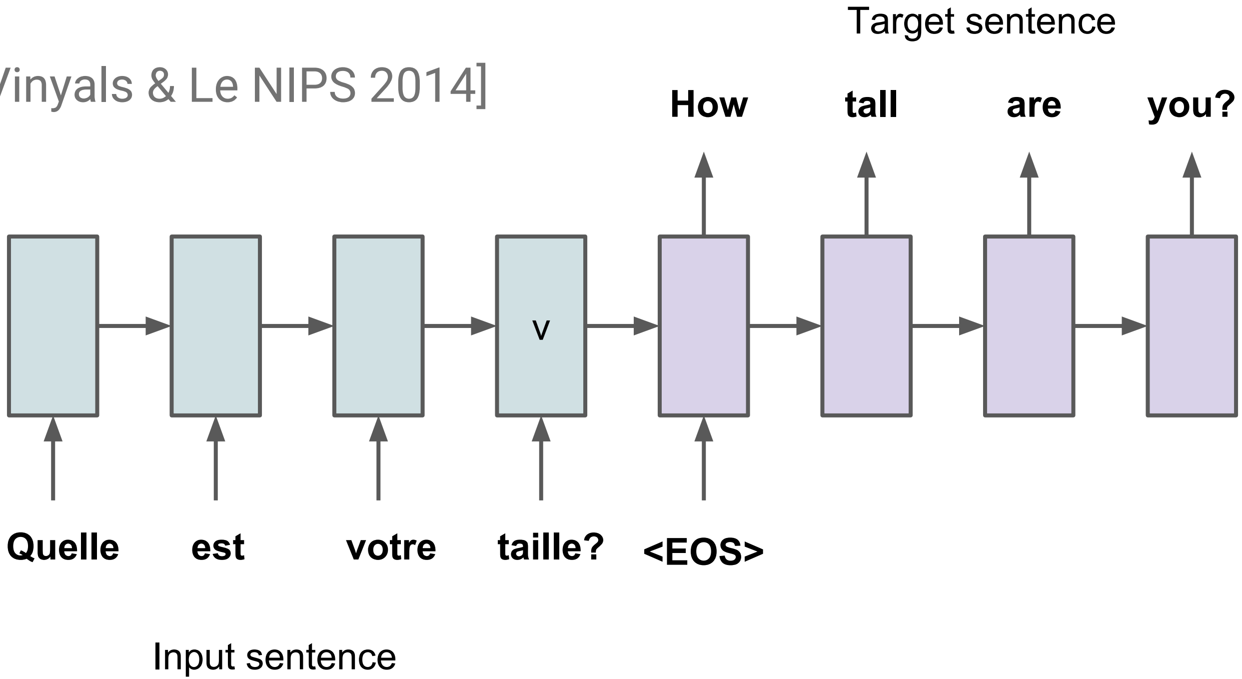
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



Sequence-to-Sequence

- Active area of research
- Many groups actively pursuing RNN/LSTM
    - Montreal
    - Stanford
    - U of Toronto
    - Berkeley
    - Google
    - ...
- Further Improvements
    - Attention
    - NTM / Memory Nets
    - ...

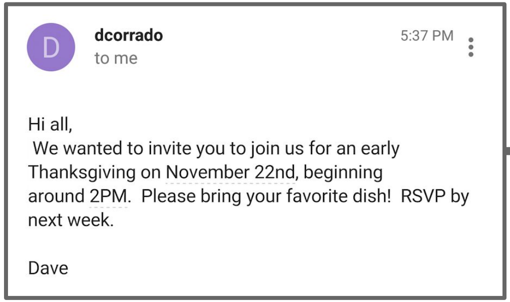
Sequence-to-Sequence

- **Translation:** [Kalchbrenner *et al.*, EMNLP 2013][Cho *et al.*, EMLP 2014][Sutskever & Vinyals & Le, NIPS 2014][Luong *et al.*, ACL 2015][Bahdanau *et al.*, ICLR 2015]
- **Image captions:** [Mao *et al.*, ICLR 2015][Vinyals *et al.*, CVPR 2015][Donahue *et al.*, CVPR 2015][Xu *et al.*, ICML 2015]
- **Speech:** [Chorowsky *et al.*, NIPS DL 2014][Chan *et al.*, arxiv 2015]
- **Language Understanding:** [Vinyals & Kaiser *et al.*, NIPS 2015][Kiros *et al.,* NIPS 2015]
- **Dialogue:** [Shang *et al.*, ACL 2015][Sordoni *et al.*, NAACL 2015][Vinyals & Le, ICML DL 2015]
- **Video Generation:** [Srivastava *et al.*, ICML 2015]
- **Algorithms:** [Zaremba & Sutskever, arxiv 2014][Vinyals & Fortunato & Jaitly, NIPS 2015][Kaiser & Sutskever, arxiv 2015][Zaremba *et al.*, arxiv 2015]

Smart Reply

Google Research Blog
- Nov 2015

Incoming Email


D
dcorrado
to me
5:37 PM

Hi all,
We wanted to invite you to join us for an early Thanksgiving on November 22nd, beginning around 2PM. Please bring your favorite dish! RSVP by next week.

Dave

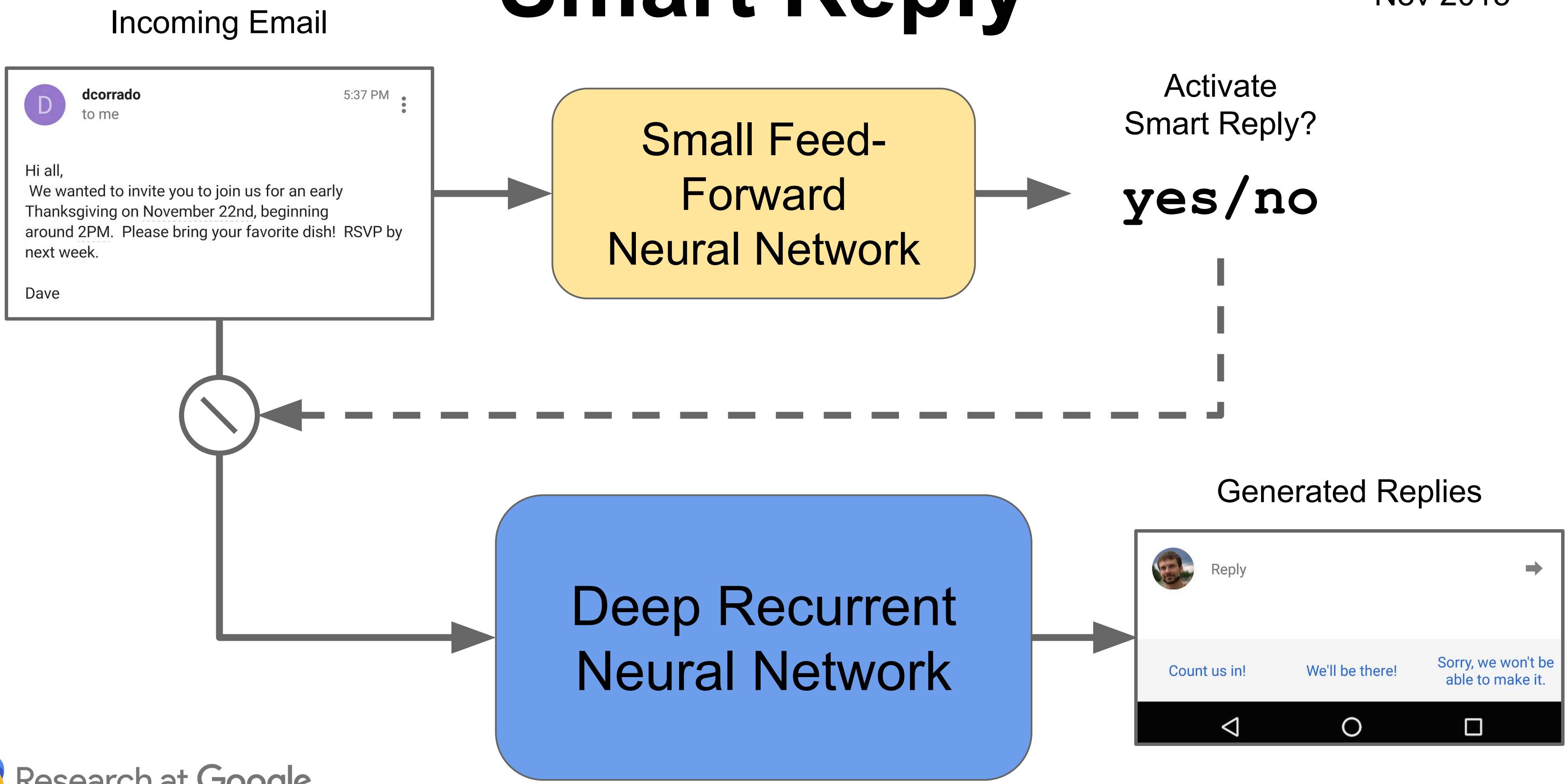
Small Feed-Forward Neural Network

Activate
Smart Reply?

`yes/no`

Smart Reply

Google Research Blog
- Nov 2015



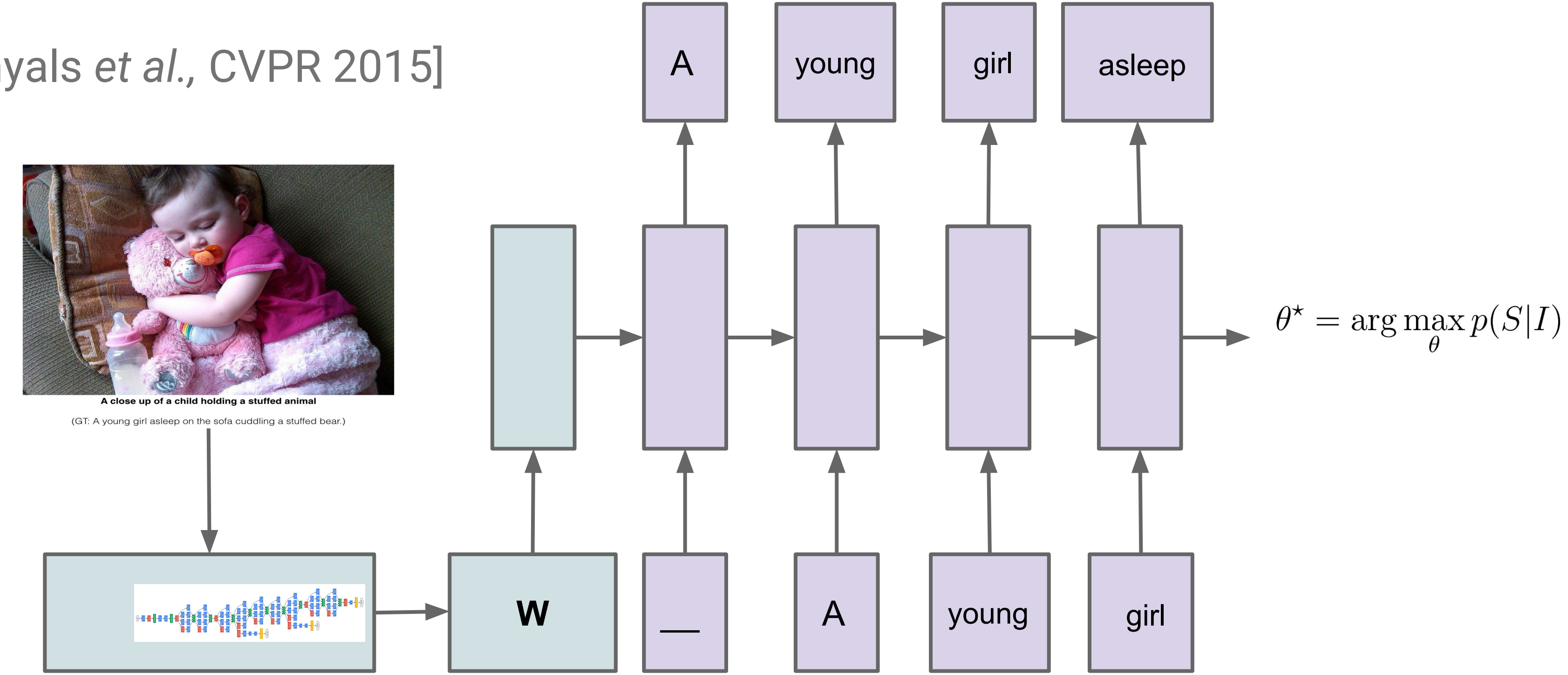
How to do Image Captions?

P(English | French)

P(English | Image)

How?

[Vinyals *et al.*, CVPR 2015]



Human: A young girl asleep on the sofa cuddling a stuffed bear.

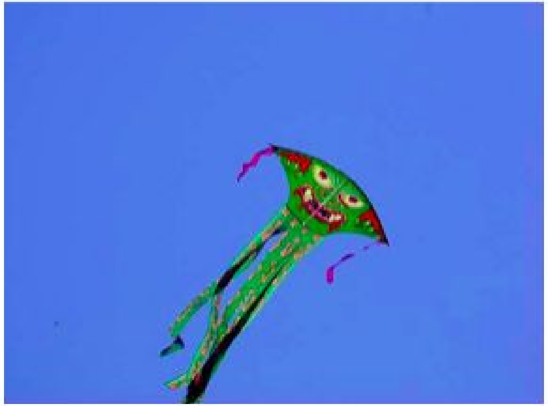
Model: A close up of a child holding a stuffed animal.

Model: A baby is asleep next to a teddy bear.

A man holding a tennis racquet on a tennis court.

Two pizzas sitting on top of a stove top oven

A group of young people playing a game of Frisbee

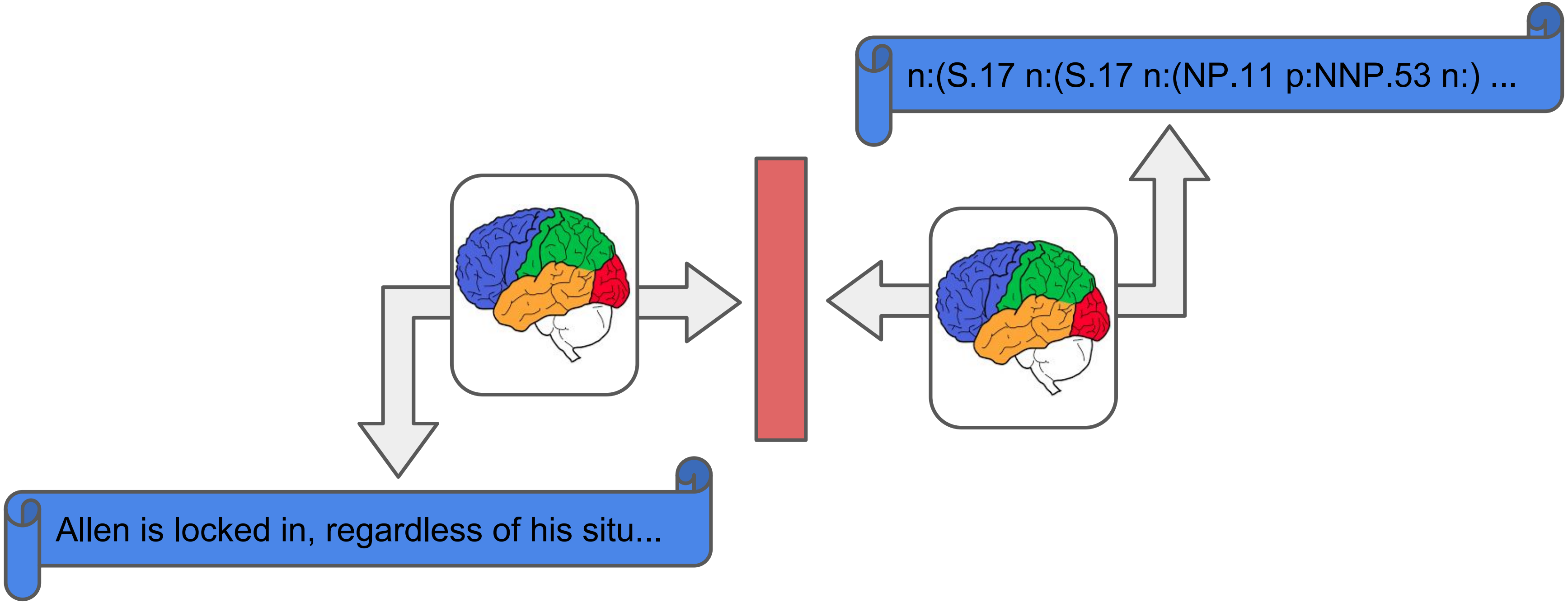


A man flying through the air while riding a snowboard

Combined Vision + Translation

Research at Google

Can also learn a grammatical parser



It works well

Completely learned parser with no parsing-specific code

State of the art results on WSJ 23 parsing task

Grammar as a Foreign Language, Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton (NIPS 2015)
http://arxiv.org/abs/1412.7449

Turnaround Time and Effect on Research

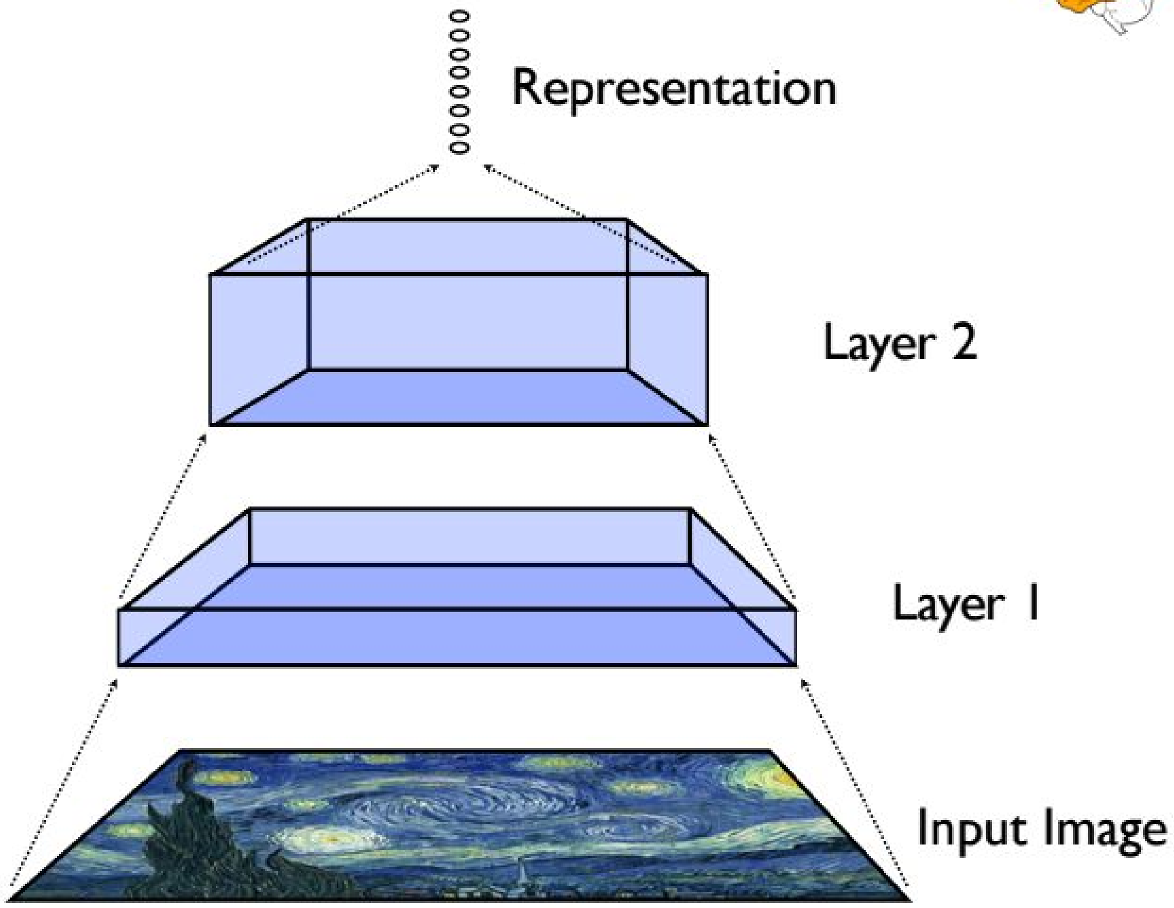
- Minutes, Hours:
  - **Interactive research! Instant gratification!**
- 1-4 days
  - Tolerable
  - Interactivity replaced by running many experiments in parallel
- 1-4 weeks:
  - High value experiments only
  - Progress stalls
- >1 month
  - Don’t even try

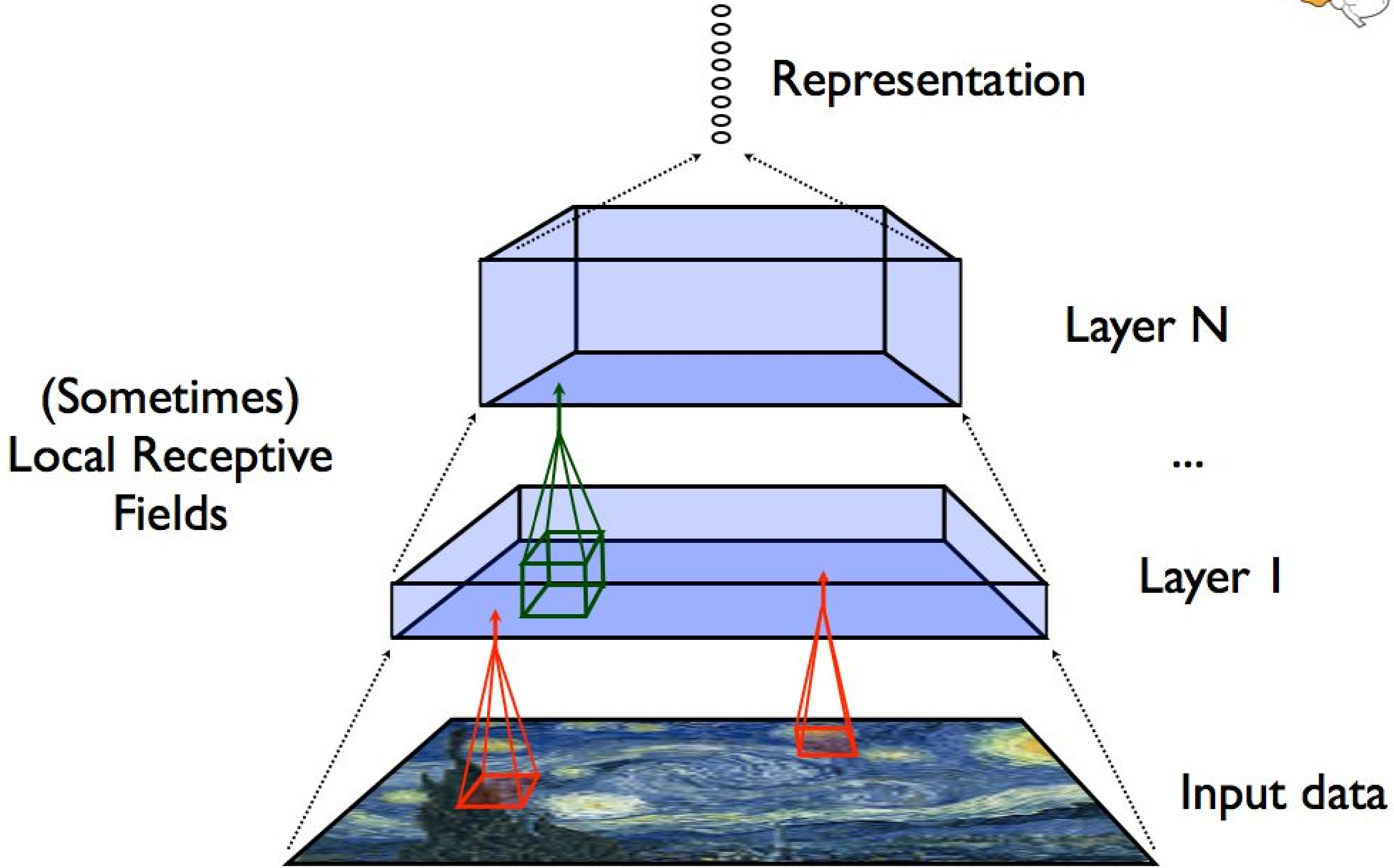
Train in a day what would take a single GPU card 6 weeks

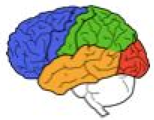
How Can We Train Large, Powerful Models Quickly?

- Exploit many kinds of parallelism
  - Model parallelism
  - Data parallelism

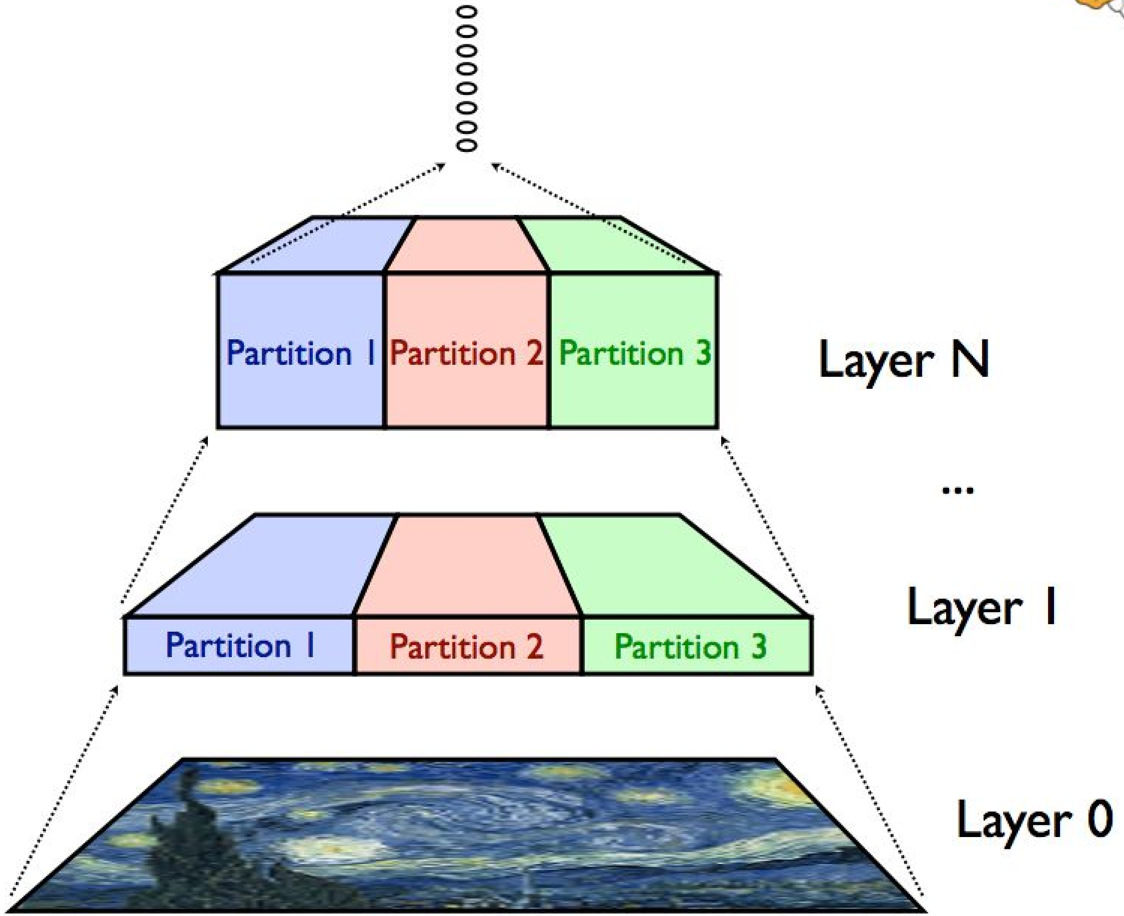
Model Parallelism



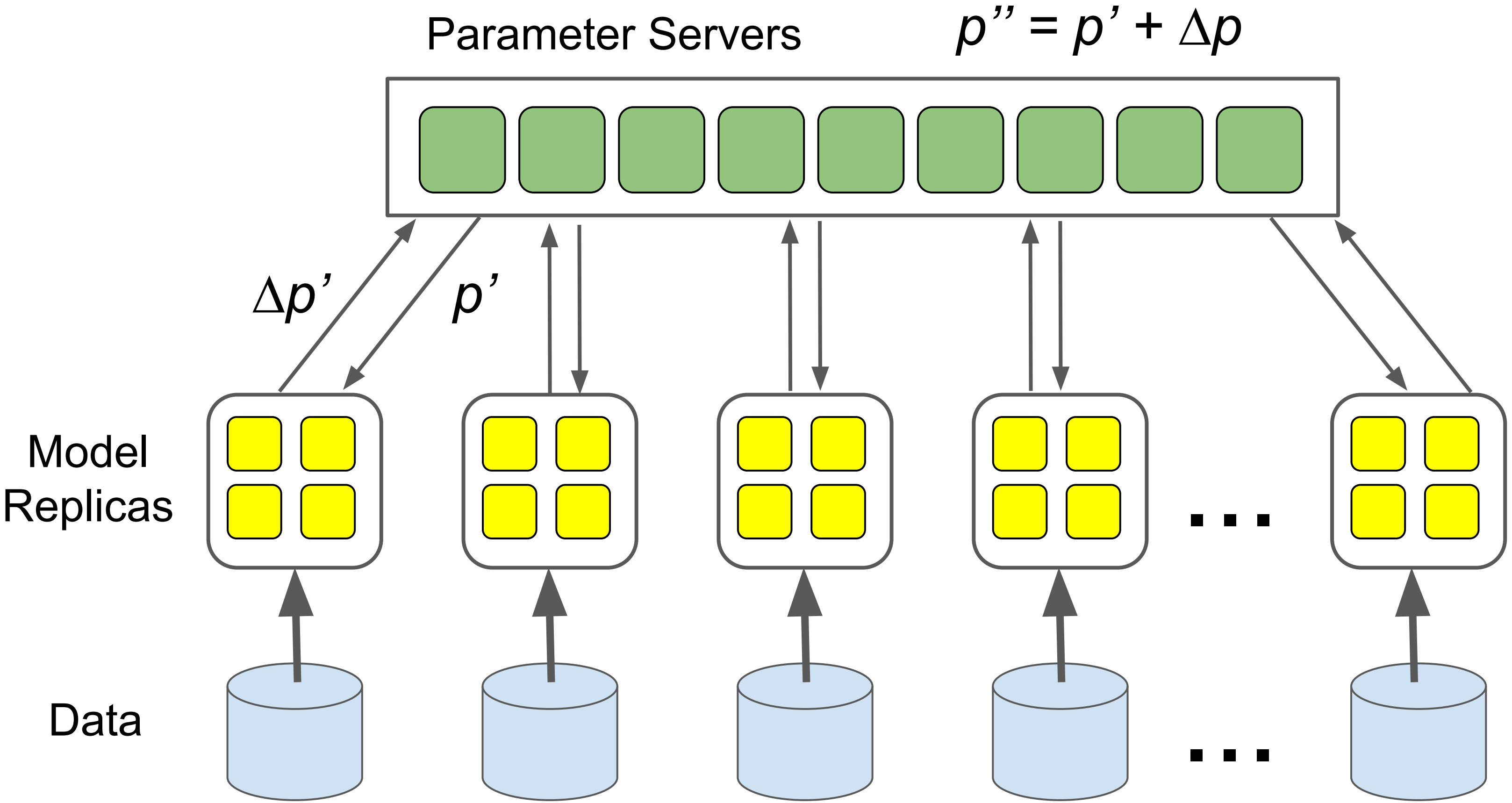

Representation
Layer N
(Sometimes)
Local Receptive
Fields
...
Layer 1
Input data



Model Parallelism: Partition model across machines



Data Parallelism



Data Parallelism Choices

Can do this **synchronously**:

- **N replicas** equivalent to an **N times larger batch size**
- Pro: No noise
- Con: Less fault tolerant (requires some recovery if any single machine fails)

Can do this **asynchronously**:

- Con: Noise in gradients
- Pro: Relatively fault tolerant (failure in model replica doesn't block other replicas)

(Or **hybrid**: M asynchronous groups of N synchronous replicas)

What do you want in a machine learning system?

- **Ease of expression**: for lots of crazy ML ideas/algorithms
- **Scalability**: can run experiments quickly
- **Portability**: can run on wide variety of platforms
- **Reproducibility**: easy to share and reproduce research
- **Production readiness**: go from research to real products

TensorFlow:
Second Generation Deep Learning System

TensorFlow

If we like it, wouldn't the rest of the world like it, too?

Open sourced single-machine TensorFlow on Monday, Nov. 9th, 2015

- Flexible Apache 2.0 open source licensing
- Updates for distributed implementation coming soon

http://tensorflow.org/

and

https://github.com/tensorflow/tensorflow

Version: master

MNIST For ML Beginners
The MNIST Data
Softmax Regressions
Implementing the Regression
Training
Evaluating Our Model
Deep MNIST for Experts
Setup
Load MNIST Data
Start TensorFlow InteractiveSession
Build a Softmax Regression Model
Placeholders
Variables
Predicted Class and Cost Function
Train the Model
Evaluate the Model
Build a Multilayer Convolutional Network
Weight Initialization
Convolution and Pooling
First Convolutional Layer
Second Convolutional Layer
Densely Connected Layer
Readout Layer
Train and Evaluate the Model
TensorFlow Mechanics 101
Tutorial Files
Prepare the Data

TensorFlow Mechanics 101

This is a technical tutorial, where we walk you through the details of using TensorFlow infrastructure to train models at scale. We use again MNIST as the example.

View Tutorial

Convolutional Neural Networks

An introduction to convolutional neural networks using the CIFAR-10 data set. Convolutional neural nets are particularly tailored to images, since they exploit translation invariance to yield more compact and effective representations of visual content.

View Tutorial

Vector Representations of Words

This tutorial motivates why it is useful to learn to represent words as vectors (called word embeddings). It introduces the word2vec model as an efficient method for learning embeddings. It also covers the high-level details behind noise-contrastive training methods (the biggest recent advance in training embeddings).

View Tutorial

Recurrent Neural Networks

An introduction to RNNs, wherein we train an LSTM network to predict the next word in an English sentence. (A task sometimes called language modeling.)

View Tutorial

Sequence-to-Sequence Models

A follow on to the RNN tutorial, where we assemble a sequence-to-sequence model for machine translation. You will learn to build your own English-to-French translator, entirely machine learned, end-to-end.

View Tutorial

TensorFlow:
Large-Scale Machine Learning on Heterogeneous Distributed Systems

(Preliminary White Paper, November 9, 2015)

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro,
Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow,
Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser,
Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray,
Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar,
Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals,
Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng

Google Research*

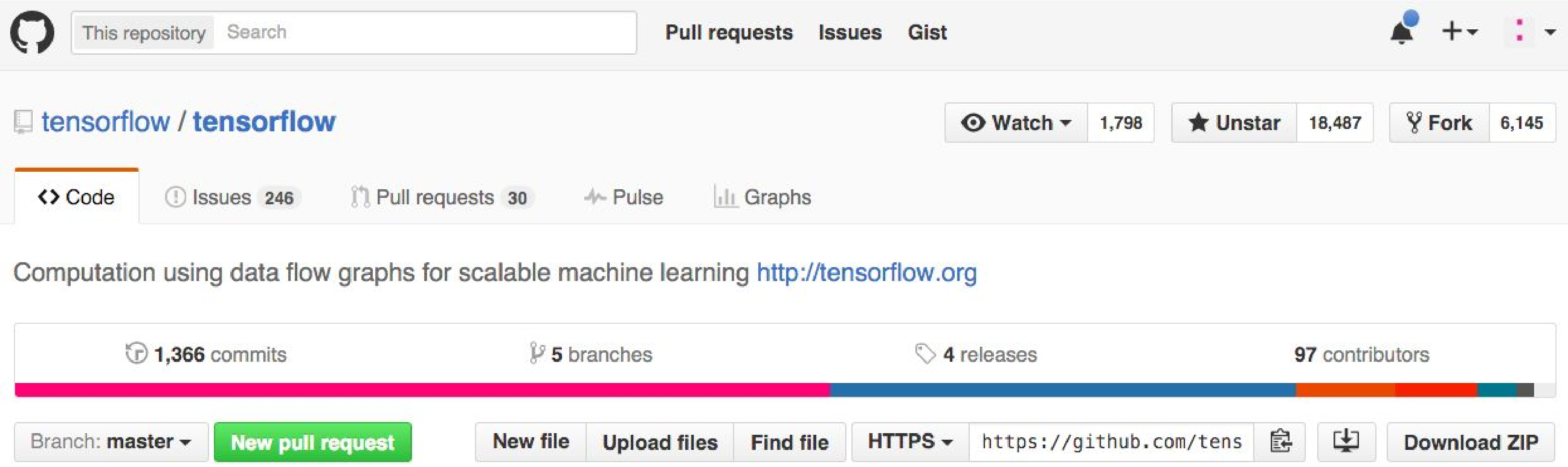
Abstract

TensorFlow [1] is an interface for expressing machine learning algorithms, and an implementation for executing such algorithms. A computation expressed using TensorFlow can be executed with little or no change on a wide variety of heterogeneous systems, ranging from mobile devices such as phones

sequence prediction [47], move selection for Go [34], pedestrian detection [2], reinforcement learning [38], and other areas [17, 5]. In addition, often in close collaboration with the Google Brain team, more than 50 teams at Google and other Alphabet companies have deployed deep neural networks using DistBelief in a wide variety

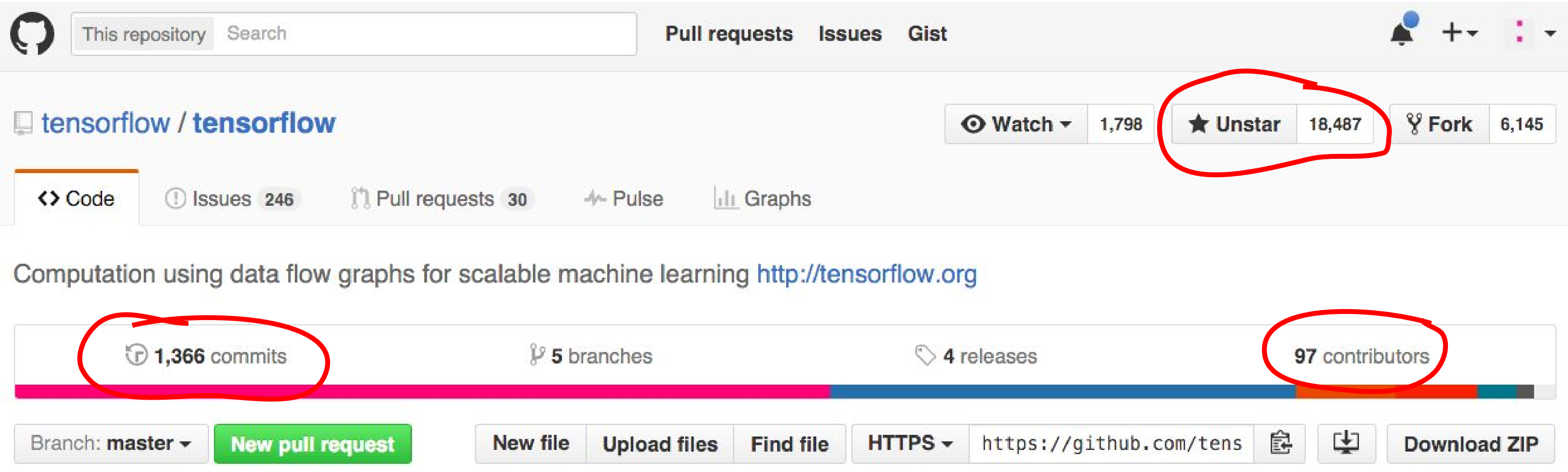
http://tensorflow.org/whitepaper2015.pdf

Source on GitHub



https://github.com/tensorflow/tensorflow

Source on GitHub



https://github.com/tensorflow/tensorflow

Motivations

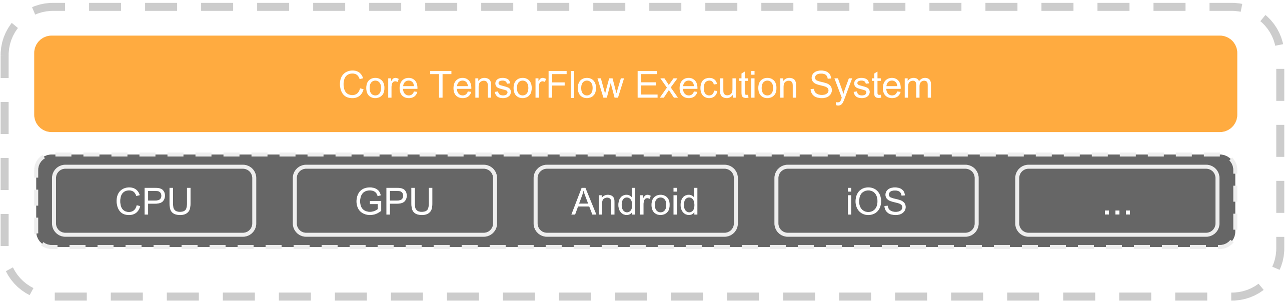
DistBelief (1st system) was great for scalability, and production training of basic kinds of models

Not as flexible as we wanted for research purposes

Better understanding of problem space allowed us to make some dramatic simplifications

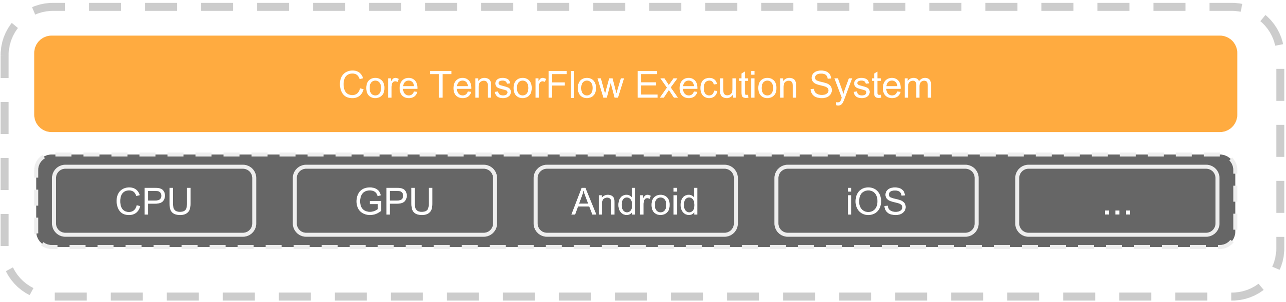
TensorFlow: Expressing High-Level ML Computations

- Core in C++
  - Very low overhead



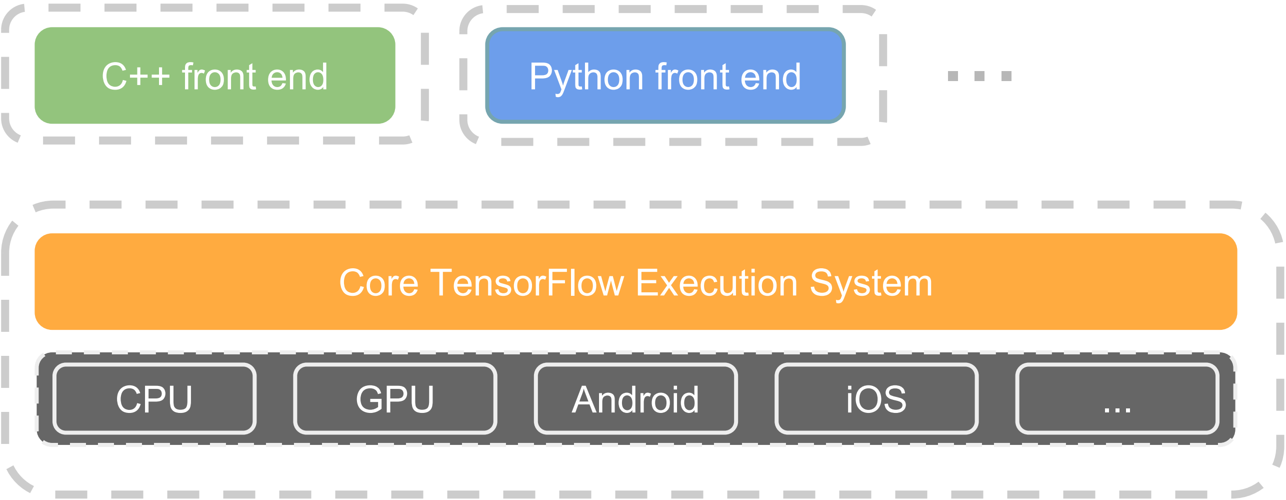
TensorFlow: Expressing High-Level ML Computations

- Core in C++
  - Very low overhead
- Different front ends for specifying/driving the computation
  - Python and C++ today, easy to add more



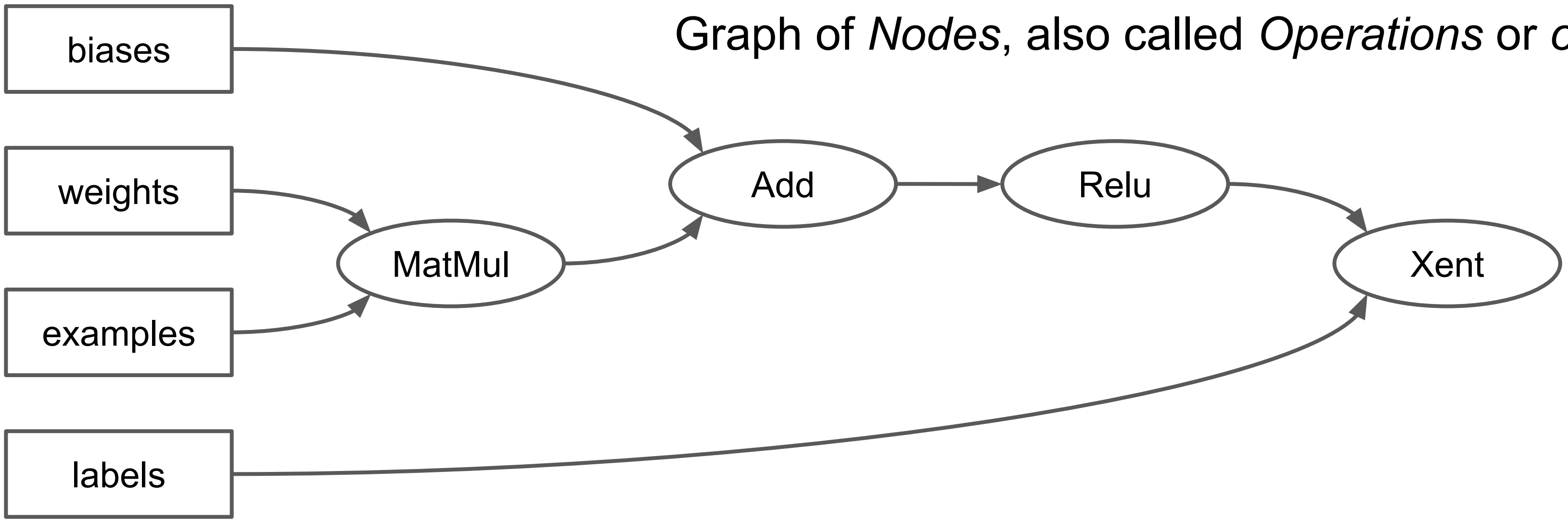
TensorFlow: Expressing High-Level ML Computations

- Core in C++
  - Very low overhead
- Different front ends for specifying/driving the computation
  - Python and C++ today, easy to add more



Computation is a dataflow graph

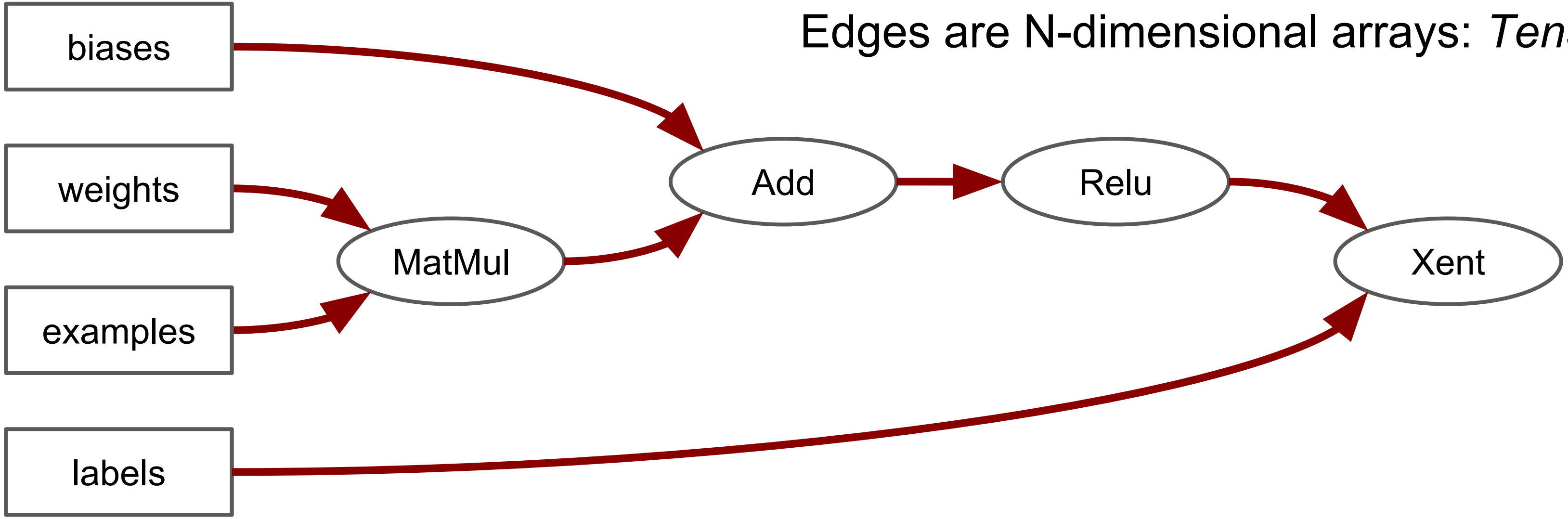
Graph of *Nodes*, also called *Operations* or *ops.*



Computation is a dataflow graph

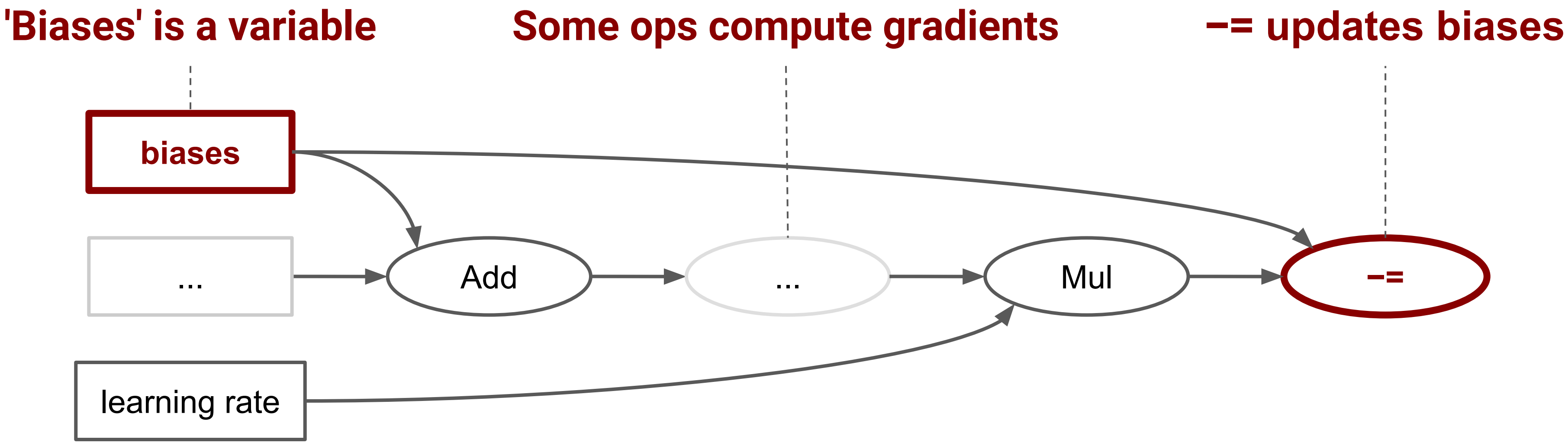
with tensors

Edges are N-dimensional arrays: *Tensors*



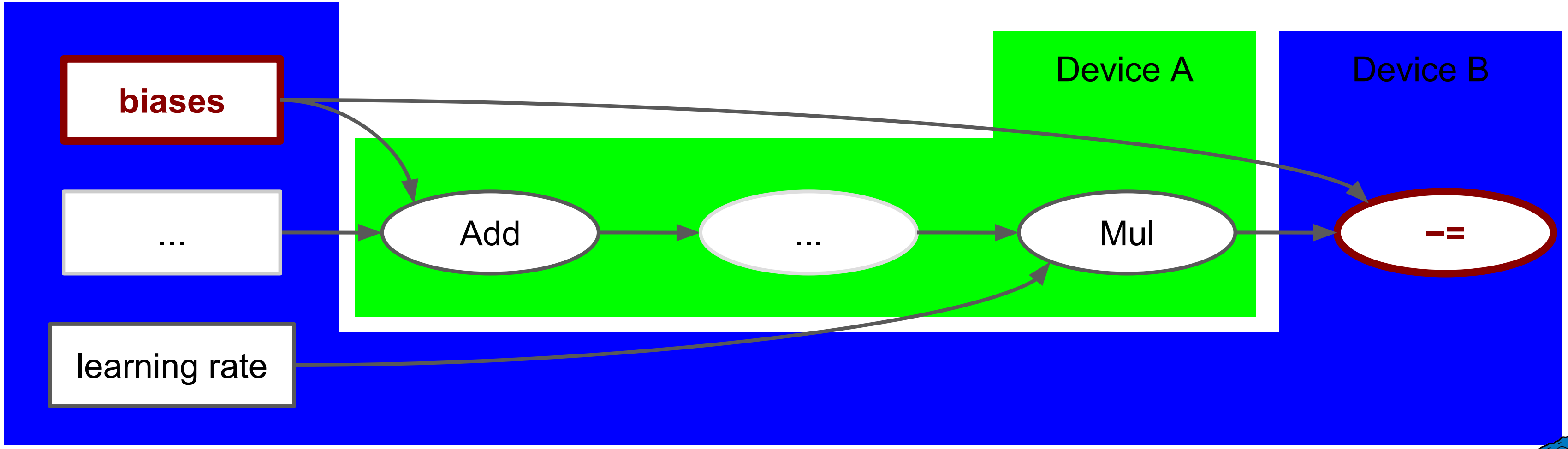
Computation is a dataflow graph

with state



Computation is a dataflow graph

distributed



Devices: Processes, Machines, GPUs, etc

TensorFlow: Expressing High-Level ML Computations

Automatically runs models on range of platforms:

from **phones** ...

to **single machines** (CPU and/or GPUs) …

to **distributed systems** of many 100s of GPU cards

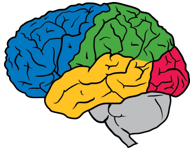
Conclusions

Deep neural networks are making significant strides in understanding:
In speech, vision, language, search, …

If you're not considering how to use deep neural nets to solve your search or understanding problems, **you almost certainly should be**

TensorFlow makes it easy for everyone to experiment with these techniques

- Highly scalable design allows faster experiments, accelerates research
- Easy to share models and to publish code to give reproducible results
- Ability to go from research to production within same system



Further Reading

- Le, Ranzato, Monga, Devin, Chen, Corrado, Dean, & Ng. *Building High-Level Features Using Large Scale Unsupervised Learning*, ICML 2012. research.google.com/archive/unsupervised_icml2012.html
- Dean, *et al.*, *Large Scale Distributed Deep Networks,* NIPS 2012, research.google.com/archive/large_deep_networks_nips2012.html.
- Mikolov, Chen, Corrado & Dean. *Efficient Estimation of Word Representations in Vector Space,* NIPS 2013, arxiv.org/abs/1301.3781.
- Le and Mikolov, *Distributed Representations of Sentences and* Documents, ICML 2014, arxiv.org/abs/1405.4053
- Sutskever, Vinyals, & Le, *Sequence to Sequence Learning with Neural Networks*, NIPS, 2014, arxiv.org/abs/1409.3215.
- Vinyals, Toshev, Bengio, & Erhan. *Show and Tell: A Neural Image Caption Generator*. CVPR 2015. arxiv.org/abs/1411.4555
- TensorFlow white paper, tensorflow.org/whitepaper2015.pdf (clickable links in bibliography)

research.google.com/people/jeff

research.google.com/pubs/MachineIntelligence.html

Questions?