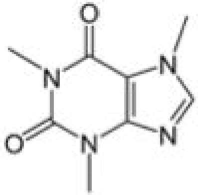
A Brief Introduction to Deep Learning and Caffe

Maximally accurate	Maximally specific
espresso	2.23192
coffee	2.19914
beverage	1.93214
liquid	1.89367
fluid	1.85519



caffe.berkeleyvision.org

github.com/BVLC/caffe

embedded VISION ALLIANCE

Evan Shelhamer, Jeff Donahue, Jon Long

Empowering Product Creators to Harness Embedded Vision

The Embedded Vision Alliance (www.Embedded-Vision.com) is a partnership of 50+ leading embedded vision technology and services suppliers

Mission: Inspire and empower product creators to incorporate visual intelligence into their products

The Alliance provides high-quality, practical technical educational resources for engineers

- Alliance website offers tutorial articles, video "chalk talks," forums
- *Embedded Vision Insights* newsletter delivers news and updates

Register for updates at www.Embedded-Vision.com

Alliance Member Companies

embedded VISION ALLIANCE

- 8tree – revealing more dimensions
- AdasWorks
- Allied Vision
- Almalence – Innovative Imaging Technologies
- AMD
- Analog Devices
- apical
- Argus Insights
- ARM
- Auviz Systems
- Basler
- BDTi
- cadence
- CEVA – The DSP Powerhouse
- Cogent Embedded
- dji
- DMP
- eUtecUs
- FotoNation
- Huawei
- Imagination
- infineon
- intel
- Irida Labs
- itSeez3D
- JPR – Jon Peddie Research
- Luxoft
- MathWorks
- Mentor Graphics
- Mitek
- Morpho
- Movidius
- nVidia
- NXP
- ON Semiconductor
- PathPartner
- PreeMadonna
- PTC
- Qualcomm
- Quest – Born to Engineer
- QuickPlay
- Rambus
- Renesas
- Sighthound
- Sony
- StradVision
- Synopsys – Silicon to Software
- tend
- TeraDeep
- Texas Instruments
- ThunderSoft
- VectorBlox – embedded supercomputing
- VeriSilicon
- videantis – passion for video
- Vision Ventures
- Vivante
- WINDs – CMOS Imager
- Xilinx – All Programmable.
- zmicro

Hands-on Tutorial on Deep Learning and Caffe

Want to get a jump start in using convolutional neural networks (CNNs) for vision applications?

Sign up for a day-long tutorial on CNNs for deep learning with hands-on lab training on the Caffe software framework.

- *How CNNs work, and how to use them for vision*
- *How to use Caffe to design, train, and deploy CNNs*

September 22nd, 9 am to 5 pm, in **Cambridge, Massachusetts**

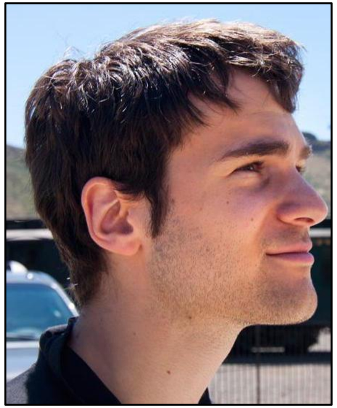
Register at [*http://www.embedded-vision.com/caffe-tutorial*](http://www.embedded-vision.com/caffe-tutorial)

- **Use promo code "CNN16-0824" for a 10% discount**

Speakers (and Caffe developers)

Evan Shelhamer

Jeff Donahue



Jon Long

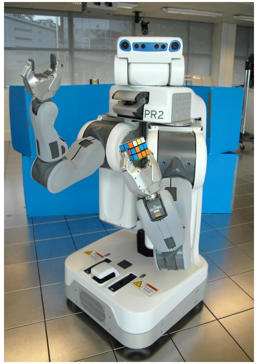
Why Deep Learning?

End-to-End Learning for Many Tasks

vision

speech

text



control

Visual Recognition Tasks

Classification

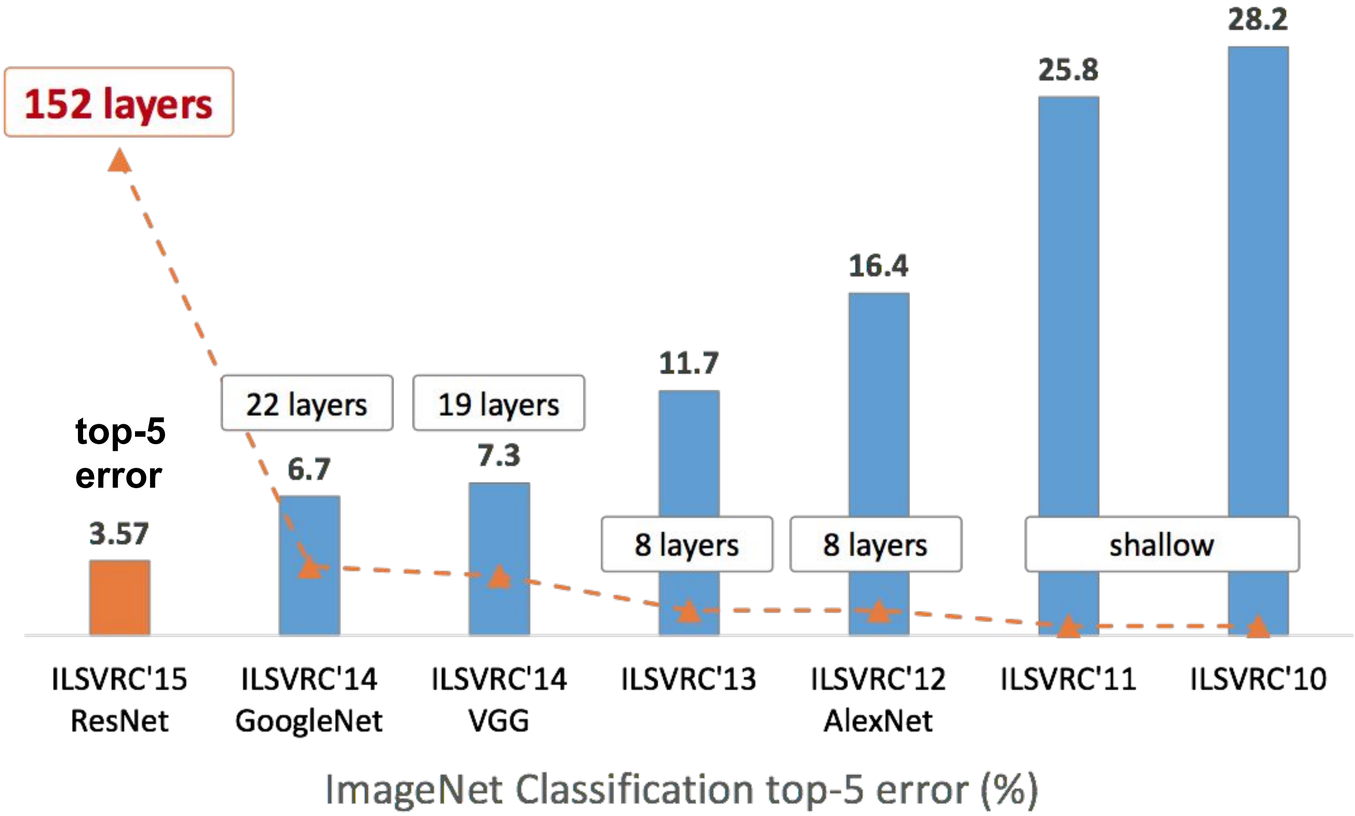
- what kind of image?
- which kind(s) of objects?

Challenges

- appearance varies by lighting, pose, context, ...
- clutter
- fine-grained categorization (horse or exact species)

- [ ] dog
- [x] car
- [x] horse
- [ ] bike
- [ ] cat
- [ ] bottle
- [x] person

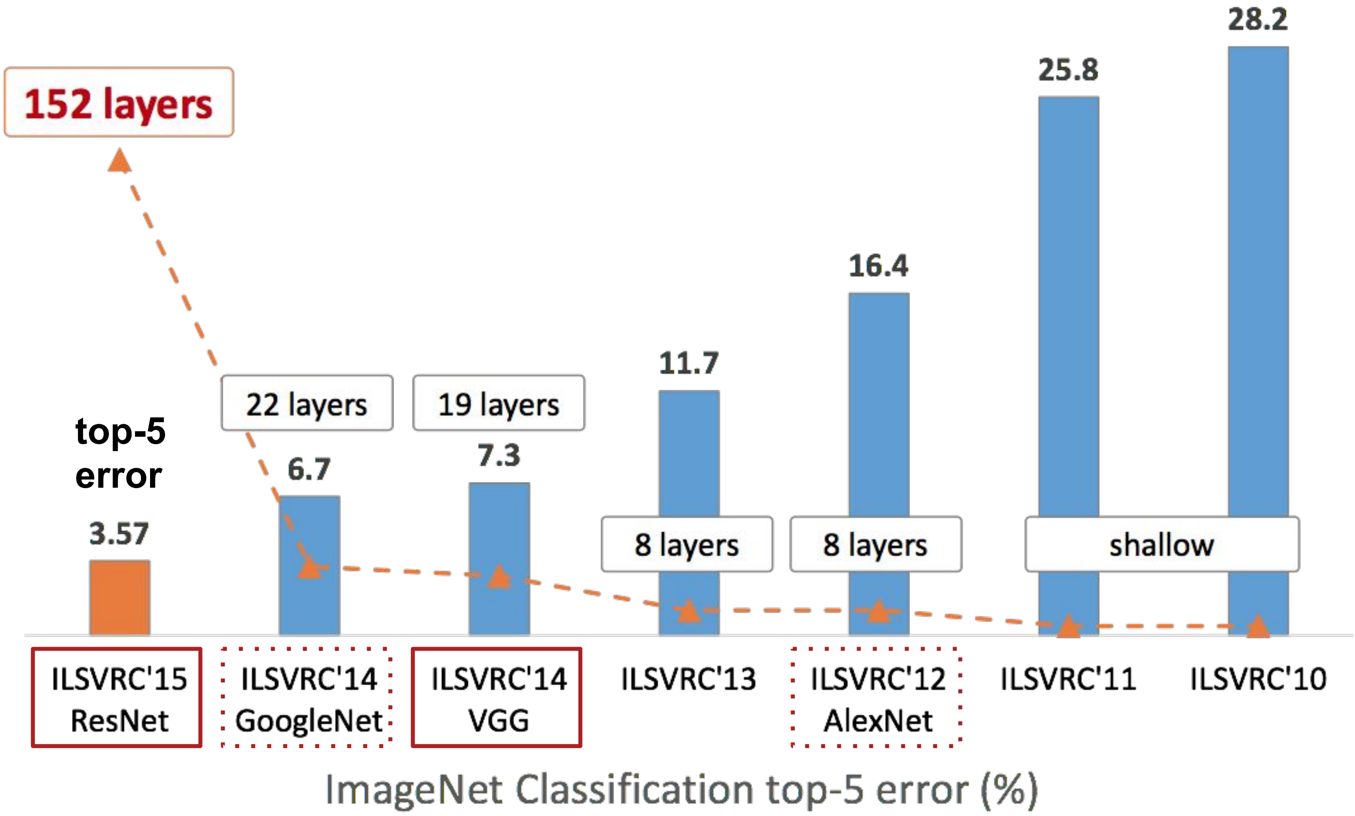
Image Classification: ILSVRC 2010-2015



- ❑ dog
- ☑ car
- ☑ horse
- ❑ bike
- ❑ cat
- ❑ bottle
- ☑ person

[graph credit K. He]

Image Classification: ILSVRC 2010-2015



- ❑ dog
- ☑ car
- ☑ horse
- ❑ bike
- ❑ cat
- ❑ bottle
- ☑ person

[graph credit K. He]

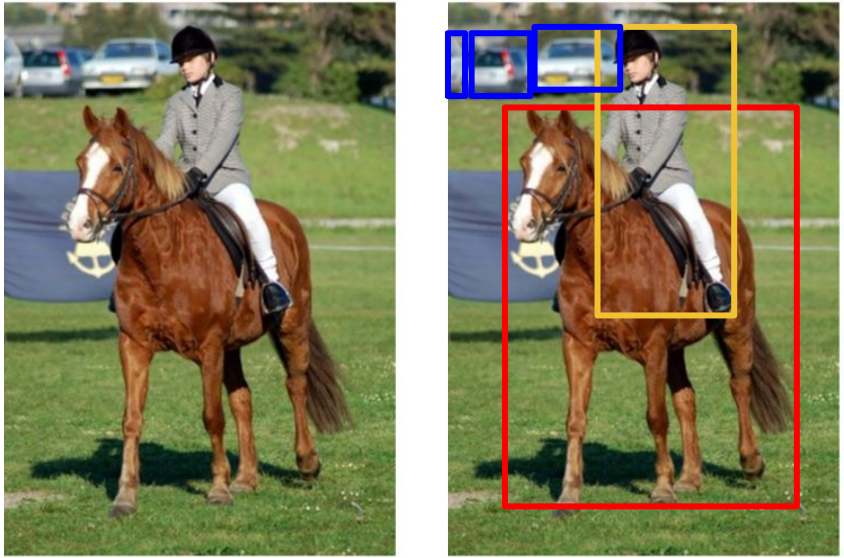
Visual Recognition Tasks

Detection

- what objects are there?
- where are the objects?

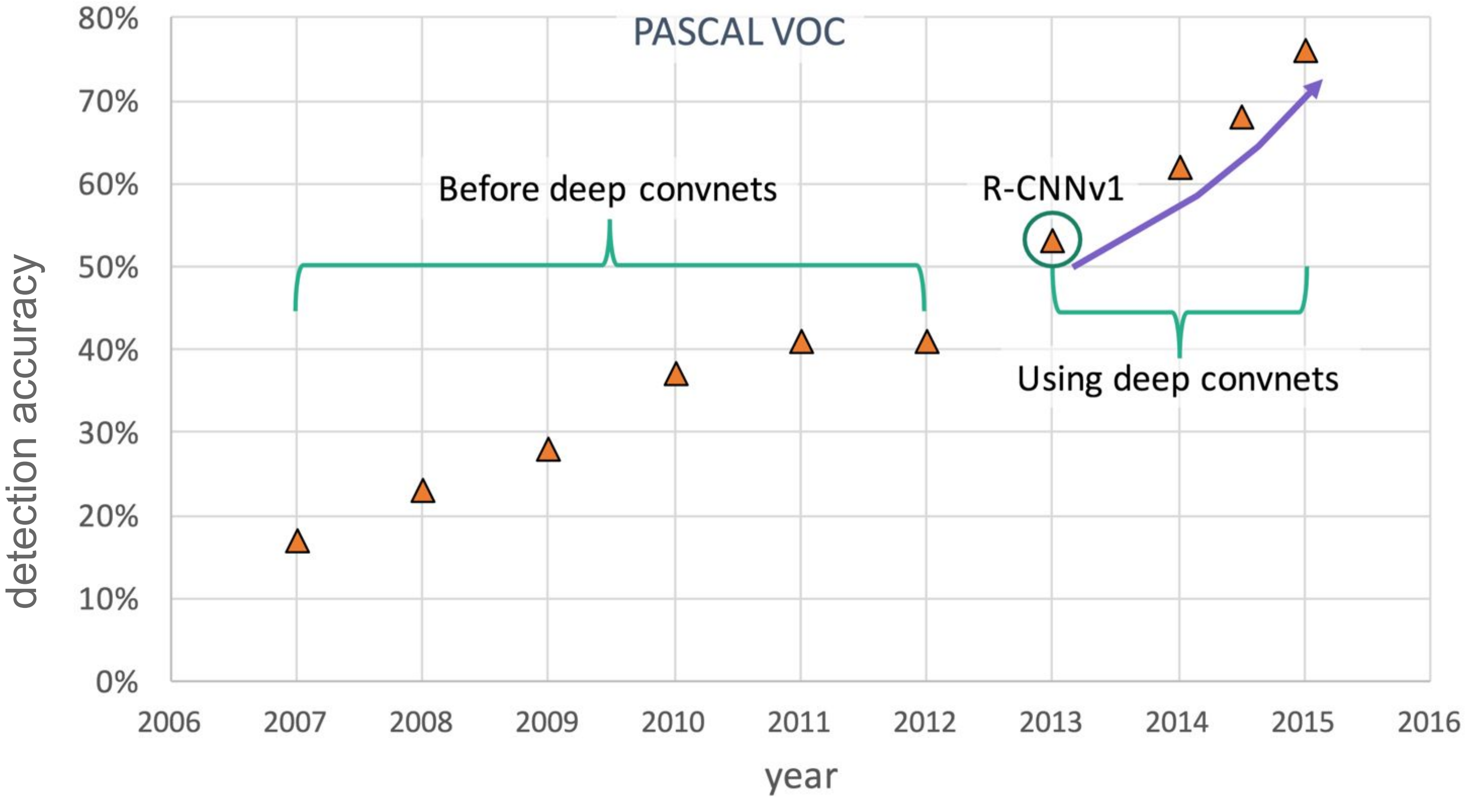
Challenges

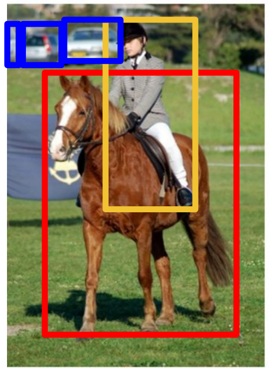
- localization
- multiple instances
- small objects



car person horse

Detection: PASCAL VOC





R-CNN:
regions +
convnets

state-of-the-art,
in Caffe

[graph credit R. Girshick]

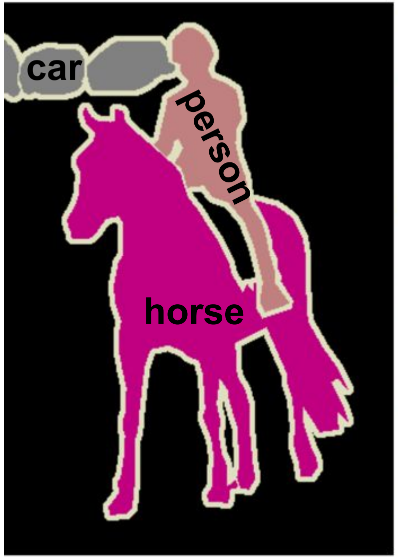
Visual Recognition Tasks

Semantic Segmentation

- what kind of thing
  is each pixel part of?
- what kind of stuff
  is each pixel?

Challenges

- tension between
  recognition and localization
- amount of computation



Segmentation: PASCAL VOC

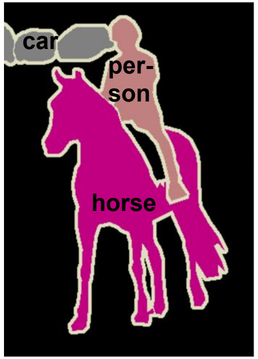
Leaderboard

MSRA_BoxSup [?]	75.2
Oxford_TVG_CRF_RNN_COCO [?]	74.7
DeepLab-MSc-CRF-LargeFOV-COCO-CrossJoint [?]	73.9
Adelaide_Context_CNN_CRF_VOC [?]	72.9
DeepLab-CRF-COCO-LargeFOV [?]	72.7
POSTECH_EDeconvNet_CRF_VOC [?]	72.5
Oxford_TVG_CRF_RNN_VOC [?]	72.0
DeepLab-MSc-CRF-LargeFOV [?]	71.6
MSRA_BoxSup [?]	71.0
DeepLab-CRF-COCO-Strong [?]	70.4
DeepLab-CRF-LargeFOV [?]	70.3
TTI_zoomout_v2 [?]	69.6
DeepLab-CRF-MSc [?]	67.1
DeepLab-CRF [?]	66.4
CRF_RNN [?]	65.2
TTI_zoomout_16 [?]	64.4
Hypercolumn [?]	62.6
FCN-8s [?]	62.2
MSRA_CFM [?]	61.8
TTI_zoomout [?]	58.4
SDS [?]	51.6
NUS_UDS [?]	50.0
TTIC-divmbest-rerank [?]	48.1
BONN_O2PCPMC_FGT_SEGM [?]	47.8
BONN_O2PCPMC_FGT_SEGM [?]	47.5
BONNGC_O2P_CPMC_CSI [?]	46.8
BONN_CMBR_O2P_CPMC_LIN [?]	46.7

deep learning with Caffe

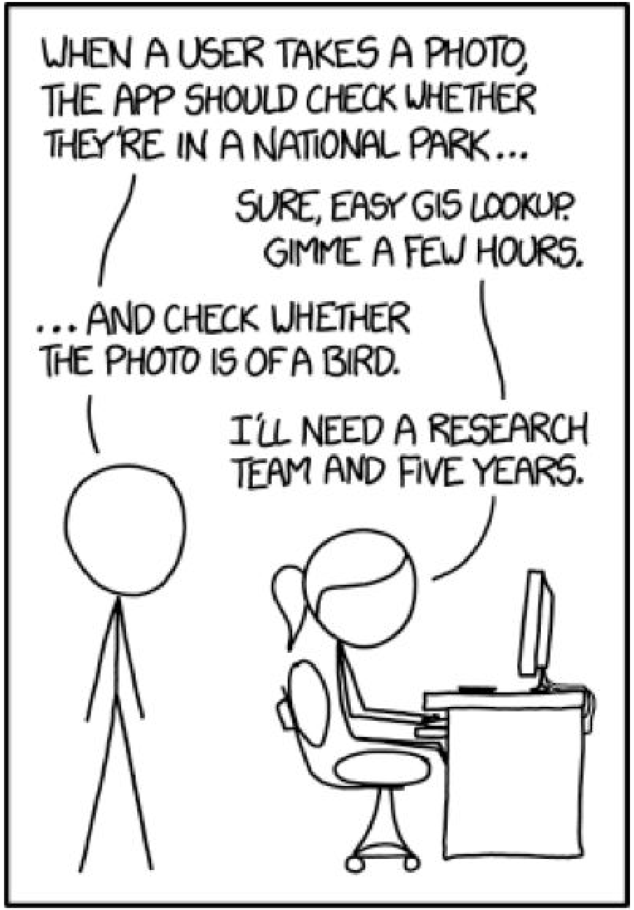
end-to-end networks lead to
30 points absolute or 50% relative improvement
and >100x speedup in 1 year!

(papers published for +1 or +2 points)



FCN:
pixelwise
convnet

state-of-the-art,
in Caffe



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

xkcd: Tasks

"The Virtually Impossible"

PARK or BIRD

Want to know if your photo is from a U.S. national park? Want to know if it contains a bird? Just drag it into the box to the left, and we'll tell you. We'll use the GPS embedded in your photo (if it's there) to see whether it's from a park, and we'll use our super-cool computer vision skills to try to see whether it's a bird (which is a hard problem, but we do a pretty good job at it).

To try it out, just drag any photo from your desktop into the upload box, or try dragging any of our example images. We'll give you your answers below!

Want to know more about PARK or BIRD, including why the heck we did this? Just click here for more info → i

EXAMPLE PHOTOS

Photo credits

PARK?

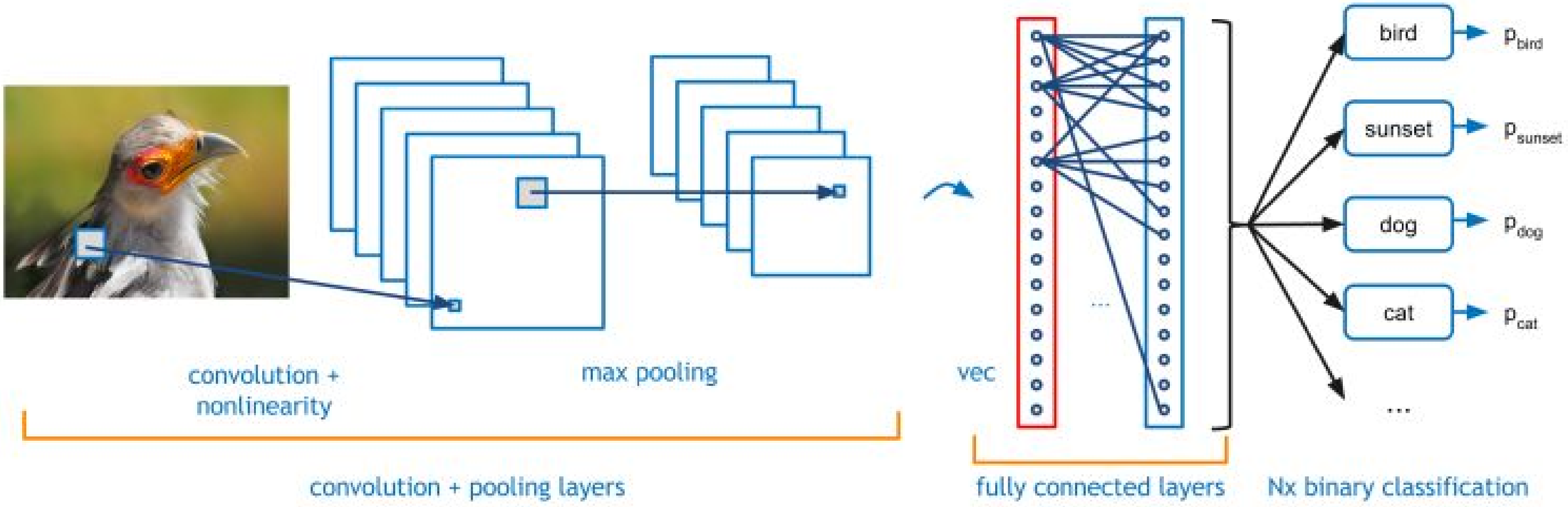
YES

Ah yes, Everglades is truly beautiful.

BIRD?

YES

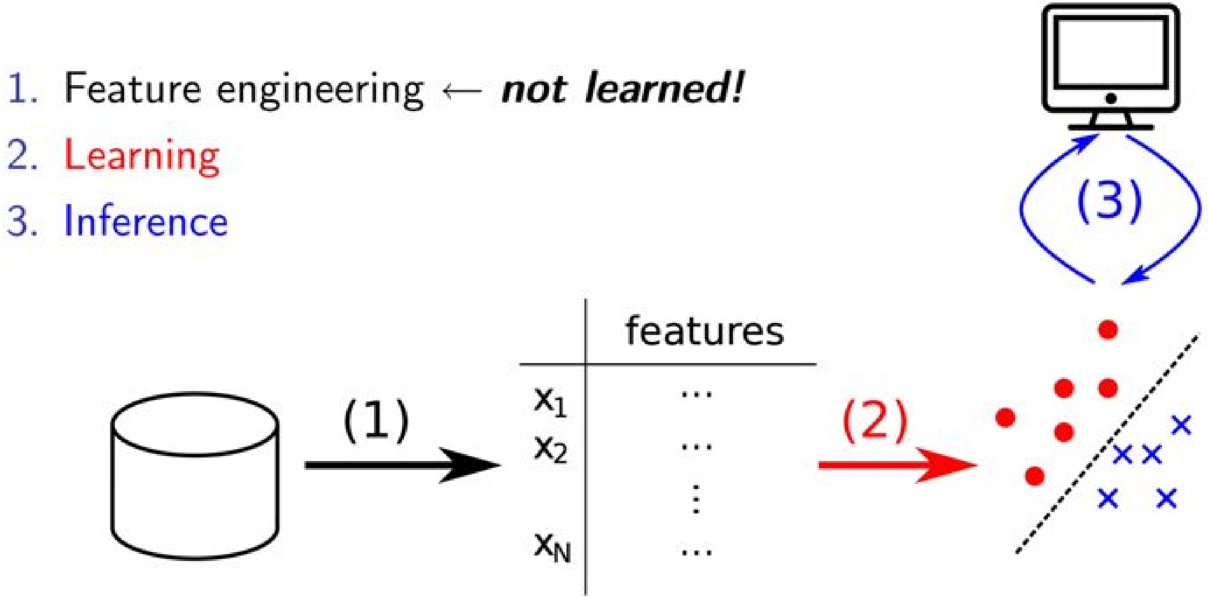
Dude, that is such a bird.



All in a day’s work with Caffe

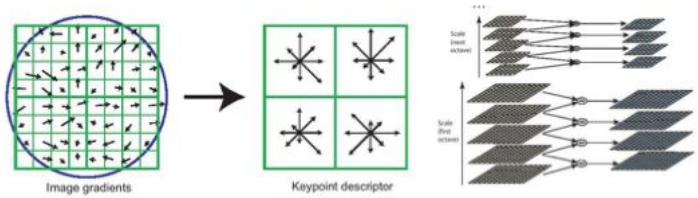
http://code.flickr.net/2014/10/20/introducing-flickr-park-or-bird/

Shallow Learning

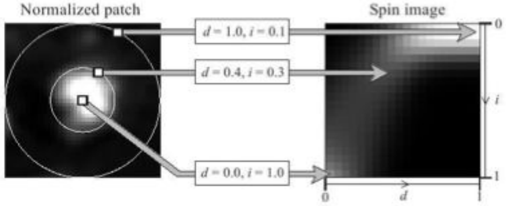


Separation of hand engineering and machine learning

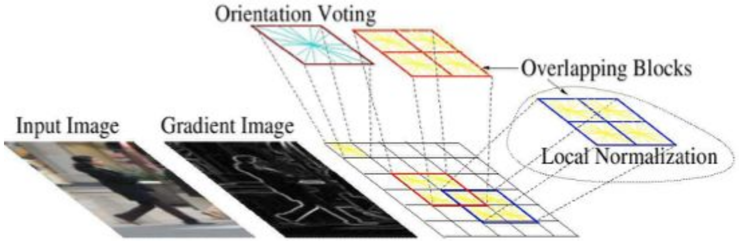
Hand-Engineered Features



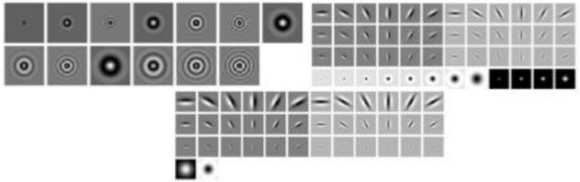
SIFT



Spin image



HoG

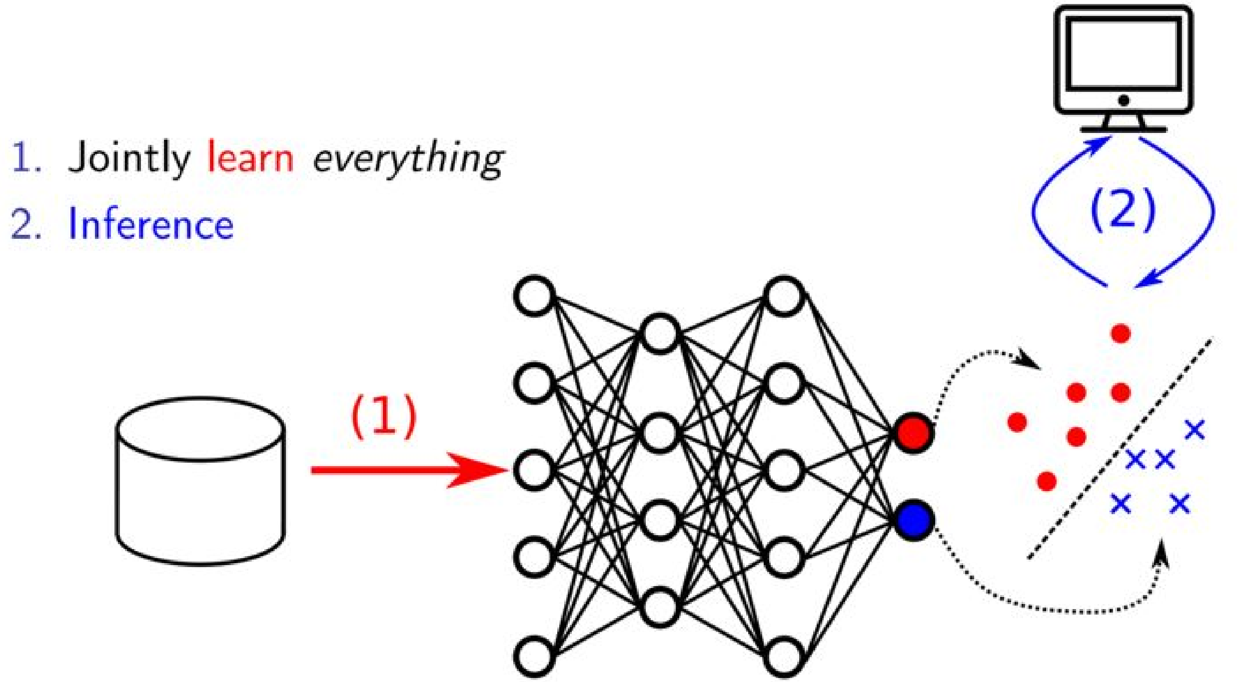


Textons

[figure credit R. Fergus]

Features from years of vision expertise by the whole community are now surpassed by *learned* representations and these *transfer across tasks*

Deep Learning

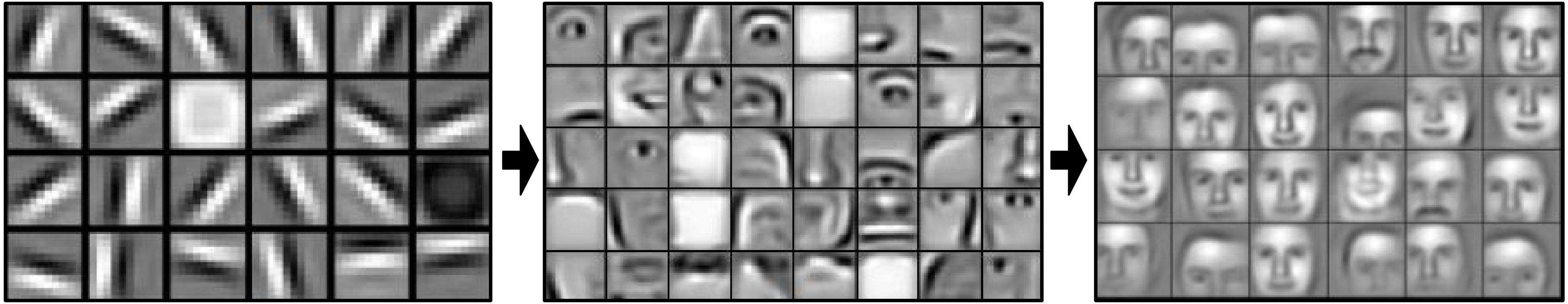


The data decides –Yoshua Bengio

[slide credit K. Cho]

End-to-End Learning Representations

The visual world is too vast and varied
to fully describe by hand



local appearance parts and texture objects and semantics

Learn the representation from data

[figure credit H. Lee]

End-to-End Learning Tasks

The visual world is too vast and varied to fully describe by hand

Learn the task from data

Designing for Sight

Convolutional Networks or convnets are nets for vision

- functional fit for the visual world
  by compositionality and feature sharing

- learned end-to-end to handle visual detail
  for more accuracy and less engineering

Convnets are the dominant architectures for visual tasks

Visual Structure

Local Processing: pixels close together go together

receptive fields capture local detail

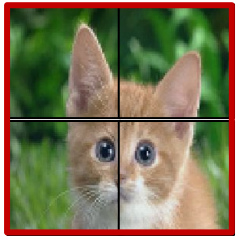
Across Space: the same what, no matter where

recognize the same input in different places

Visual Structure

Local Processing: pixels close together go together

receptive fields capture local detail



Can rely on spatial coherence

This is not a cat

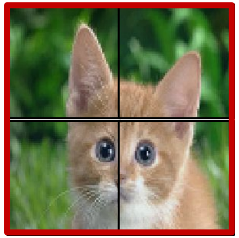
Across Space: the same what, no matter where

recognize the same input in different places

Visual Structure

Local Processing: pixels close together go together

receptive fields capture local detail



Can rely on spatial coherence

This is not a cat

Across Space: the same what, no matter where

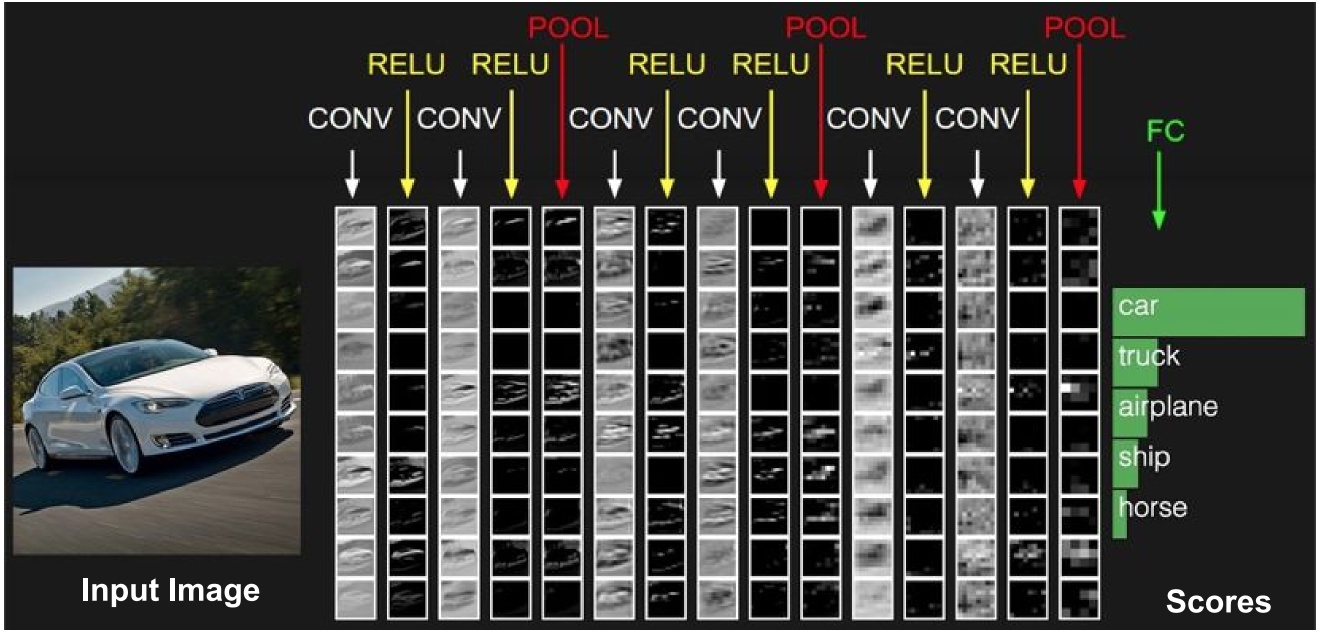
recognize the same input in different places

All of these are cats

Convnet Architecture

Stack convolution, non-linearity, and pooling until global FC layer classifier

Conv 3x3s1, 10 / ReLU **Type**: Conv **Kernel Size**: 3x3 **Stride**: 1 **Channels**:10 **Activation**: ReLU



FC 10

Max Pool 3x3s1

Conv 3x3s1, 10 / ReLU

Conv 3x3s1, 10 / ReLU

Max Pool 3x3s1

Conv 3x3s1, 10 / ReLU

Conv 3x3s1, 10 / ReLU

Max Pool 3x3s1

Conv 3x3s1, 10 / ReLU

Conv 3x3s1, 10 / ReLU

[figure credit A. Karpathy]

Why Now?

1. Data
   ImageNet et al.: millions of *labeled* (crowdsourced) images

2. Compute
   GPUs: terabytes/s memory bandwidth, teraflops compute

3. Technique
   new optimization know-how,
   new variants on old architectures,
   new tools for rapid experimentation

Why Now? Data

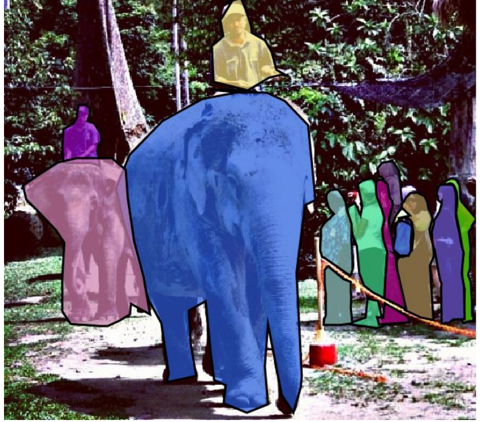
For example:

IMAGENET

>10 million labeled images
>1 million with *bounding boxes*

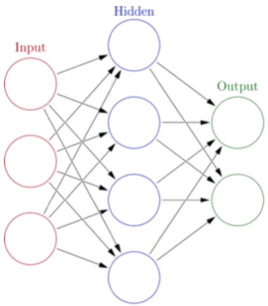
Microsoft
Common Objects in Context

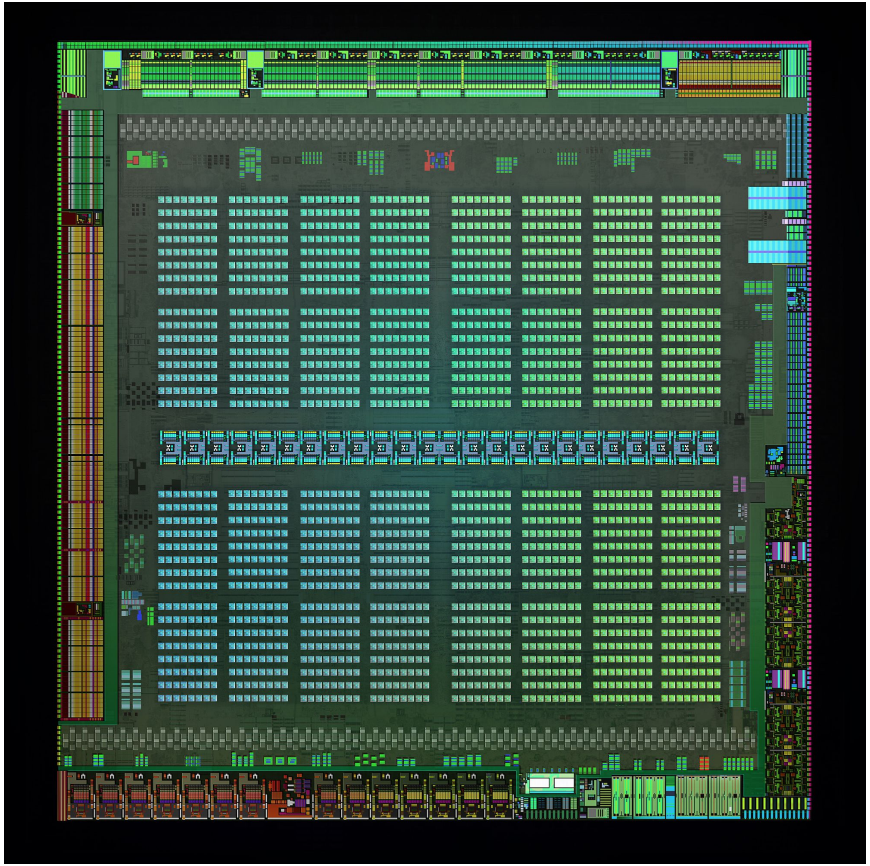
>300,000 images with *labeled and segmented* objects



Why Now? GPUs

Parallel processors
for parallel models:





Inherent Parallelism

same op, different data

Bandwidth

lots of data in and out

Tuned Primitives

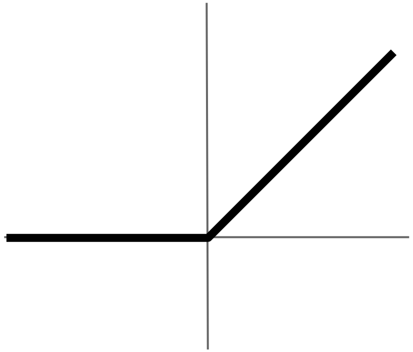
cuDNN and cuBLAS

for deep nets for matrices

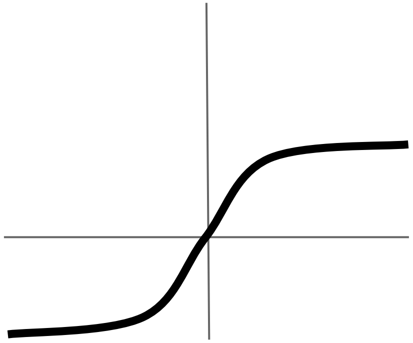
Why Now? Technique

Non-convex and high-dimensional learning is okay with the right design choices

e.g. non-saturating non-linearities



instead of



Learning by Stochastic Gradient Descent (SGD) with momentum and other variants

Why Now? Deep Learning Frameworks

frontend:
a language for any network, any task

network internal representation

tools:
visualization, profiling, debugging, etc.

backend:
dispatch compute for learning and inference

layer library:
fast implementations of common functions and gradients

framework

Deep Learning Frameworks

Caffe
Berkeley / BVLC
C++ / CUDA,
Python, MATLAB

Torch
Facebook + NYU
Lua (C++)

Theano
U. Montreal
Python

TensorFlow
Google
Python (C++)

all open source
we like to brew our networks with **Caffe**

What is Caffe?

Open framework, models, and worked examples for deep learning

- 2 years old
- 2,000+ citations, 200+ contributors, 10,000+ stars
- 7,000+ forks, >1 pull request / day average
- focus has been vision, but branching out: sequences, reinforcement learning, speech + text

Prototype

Train

Deploy

What is Caffe?

Open framework, models, and worked examples for deep learning

- Pure C++ / CUDA architecture for deep learning
- Command line, Python, MATLAB interfaces
- Fast, well-tested code
- Tools, reference models, demos, and recipes
- Seamless switch between CPU and GPU


Prototype

Train


Deploy

Caffe is a Community

project pulse

BVLC / caffe

Unwatch 1,205 | Unstar 8,498 | Fork 4,821

January 19, 2016 – February 19, 2016

Period: **1 month**

Overview

45 Active Pull Requests

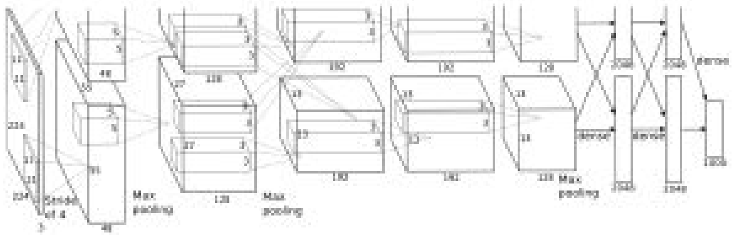
90 Active Issues

22	**23**	**52**	**38**
Merged Pull Requests	Proposed Pull Requests	Closed Issues	New Issues

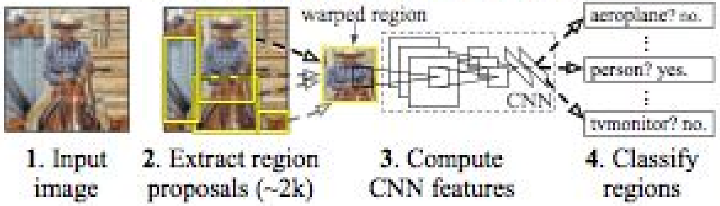
Excluding merges, **20 authors** have pushed **19 commits** to master and **53 commits** to all branches. On master, **44 files** have changed and there have been **2,268 additions** and **162 deletions**.

Reference Models

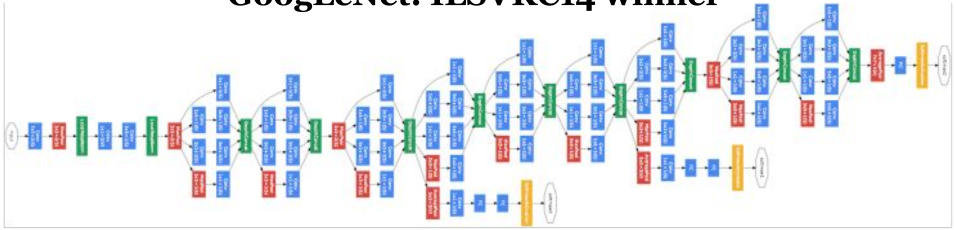
AlexNet: ImageNet Classification



R-CNN: *Regions with CNN features*



GoogLeNet: ILSVRC14 winner



Caffe offers the

- model definitions
- optimization settings
- pre-trained weights

so you can start right away

The BVLC models are licensed for unrestricted use

The community shares models in our [Model Zoo]

Embedded Caffe

Caffe runs on embedded CUDA hardware and mobile devices

- same model weights,
  same framework interface

- out-of-the-box on
  CUDA platforms

- OpenCL port
  thanks Fabian Tschopp!
  + AMD, Intel, and the community
- community Android port
  thanks sh1r0!

CUDA Jetson TX1, TK1

OpenCL branch

Android lib, demo

Industrial and Applied Caffe

… startups, big companies, more ...

Caffe at Facebook

- in production for vision at scale: uploaded photos run through Caffe
- Automatic Alt Text for the blind
- On This Day for surfacing memories
- objectionable content detection
- contributing back to the community: inference tuning, tools, code review include fb-caffe-exts thanks Andrew!

On This Day
highlight content

Automatic Alt Text
recognize photo content for accessibility

[example credit Facebook]

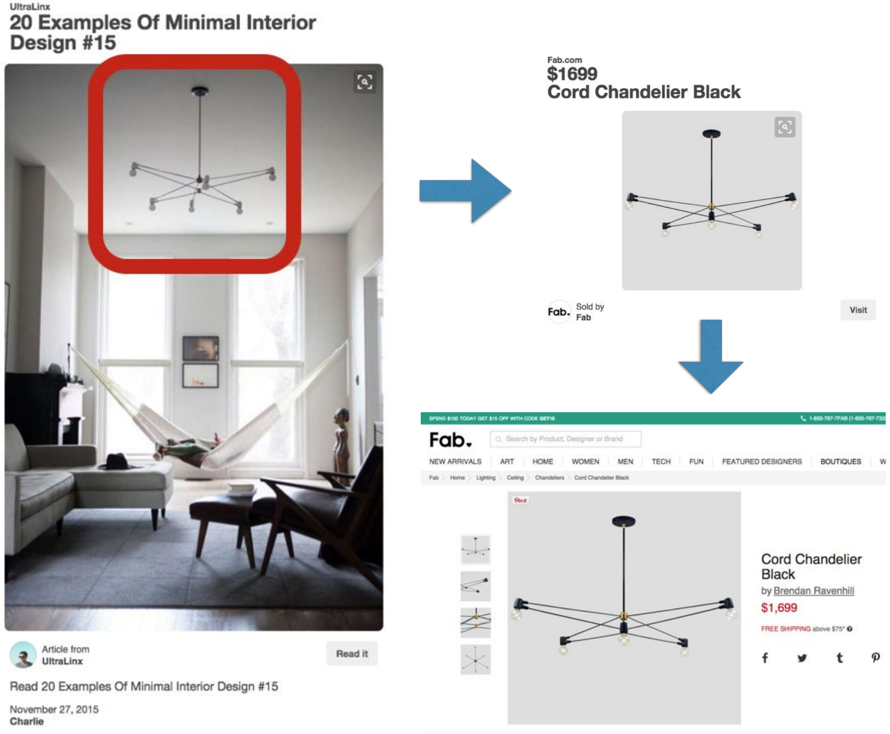
Caffe at Pinterest

- in production for vision at scale: uploaded photos run through Caffe

- deep learning for visual search: retrieval over billions of images in <250 ms

- ~4 million requests/day

- built on an open platform of Caffe, FLANN, Thrift, ...



[example credit Andrew Zhai, Pinterest]

Caffe at Yahoo! Japan

- curate news and restaurant photos for recommendation
- arrange user photo albums

News Image Recommendation
select and crop images for news

Share a Sip of Brewed Models

demo.caffe.berkeleyvision.org

demo code open-source and bundled

Maximally accurate	Maximally specific
cat	1.80727
domestic cat	1.74727
feline	1.72787
tabby	0.99133
domestic animal	0.78542

Scene Recognition http://places.csail.mit.edu/

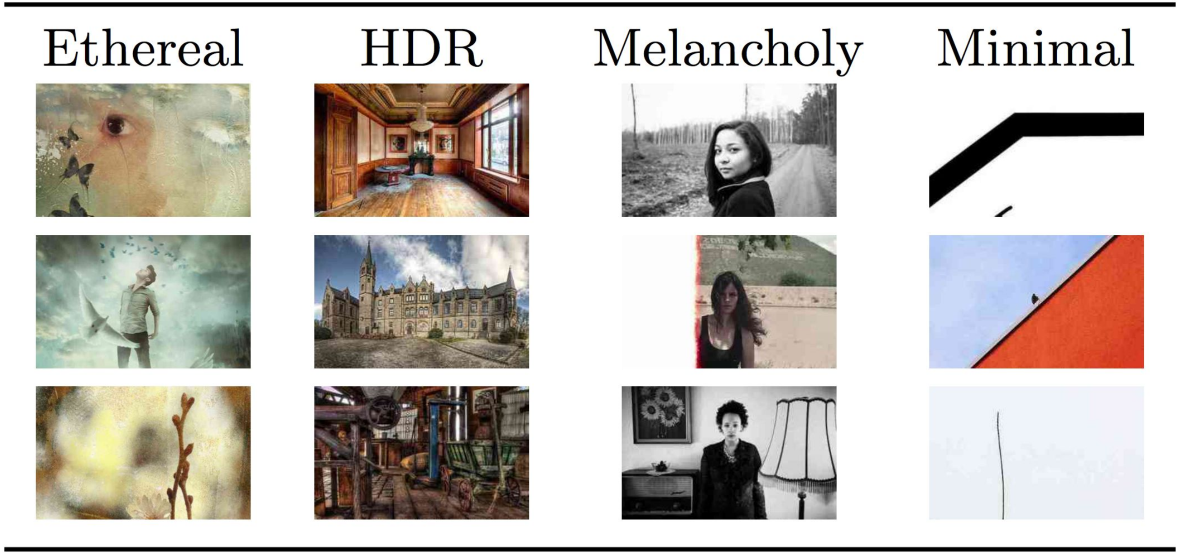
Predictions:

- **Type of environment:** outdoor
- **Semantic categories:** skyscraper:0.69, tower:0.16, office_building:0.11,
- **SUN scene attributes:** man-made, vertical components, natural light, open area, nohorizon, glossy, metal, wire, clouds, far-away horizon

B. Zhou et al. NIPS 14

Visual Style Recognition

Karayev et al. *Recognizing Image Style*. BMVC14. Caffe fine-tuning example.
Demo online at http://demo.vislab.berkeleyvision.org/ (see Results Explorer).



[Image-Style]

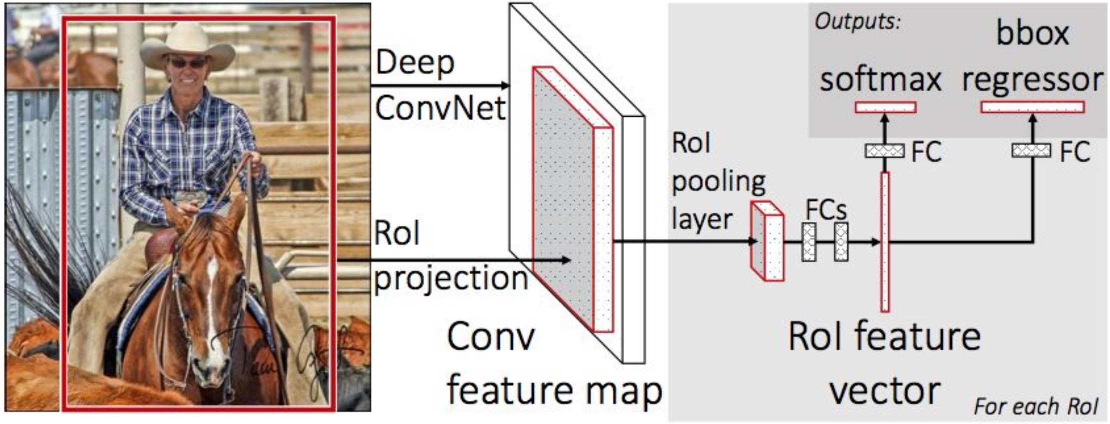
Other Styles:

[Vintage](#)
[Long Exposure](#)
[Noir](#)
[Pastel](#)
[Macro](#)
… and so on.

Object Detection

R-CNNs: Region-based Convolutional Networks

[Fast R-CNN](#)
- convnet for features
- proposals for detection



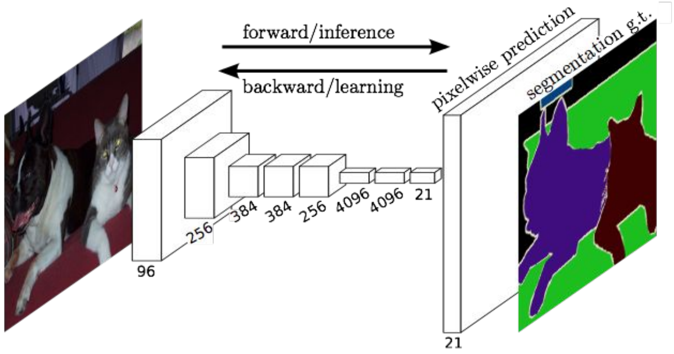
[*Faster* R-CNN](#)
- end-to-end proposals and detection
- image inference in 200 ms
- Region Proposal Net + Fast R-CNN

papers + code online

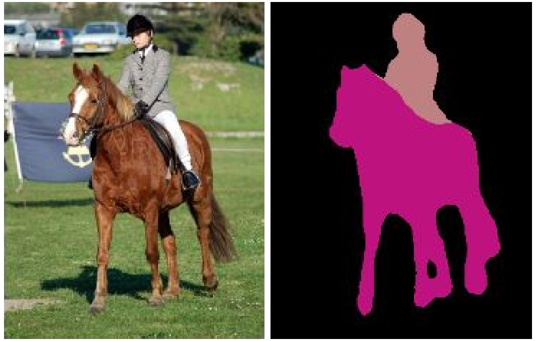
**Ross Girshick, Shaoqing Ren,
Kaiming He, Jian Sun**

Pixelwise Prediction

Fully convolutional networks for pixel prediction
in particular semantic segmentation
- end-to-end learning
- efficient inference and learning
  100 ms per-image prediction
- multi-modal, multi-task



Applications
- semantic segmentation
- denoising
- depth estimation
- optical flow



Jon Long & Evan Shelhamer*, Trevor Darrell. CVPR'15*

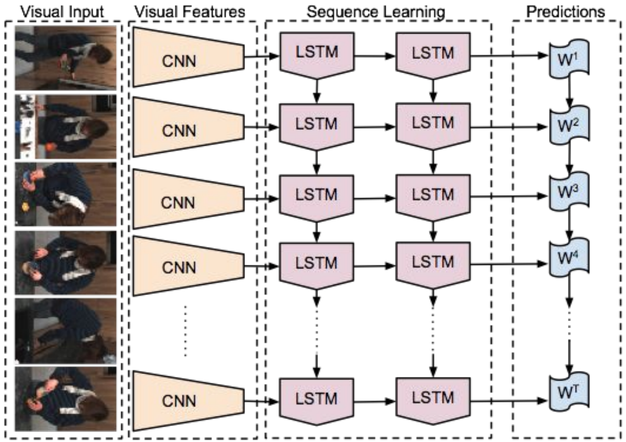
CVPR'15 paper and code + models

Recurrent Networks for Sequences

Recurrent Nets and Long Short Term Memories (LSTM)
are sequential models

- video
- language
- dynamics

learned by backpropagation through time



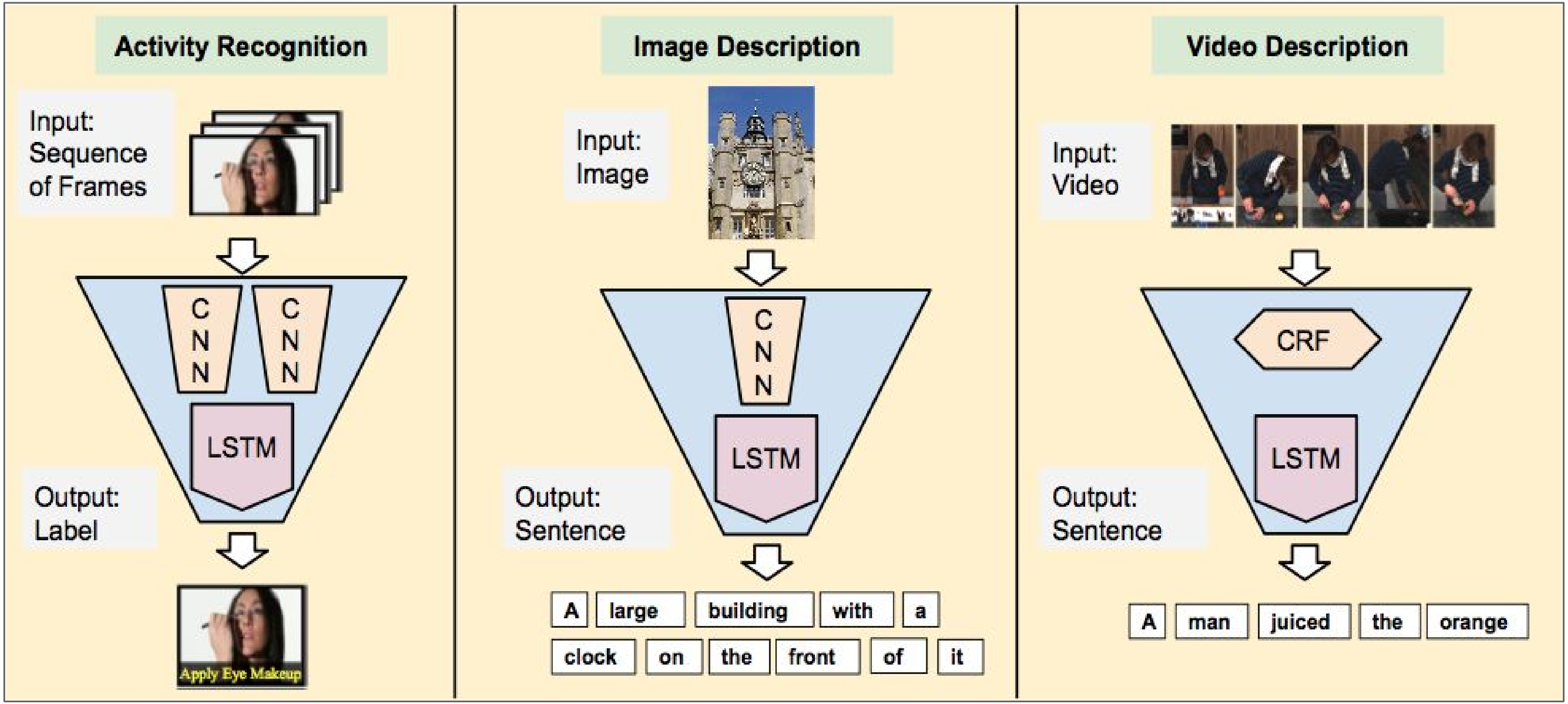
LRCN: Long-term Recurrent Convolutional Network

- activity recognition (sequence-in)
- image captioning (sequence-out)
- video captioning (sequence-to-sequence)

LRCN:
recurrent + convolutional
for visual sequences

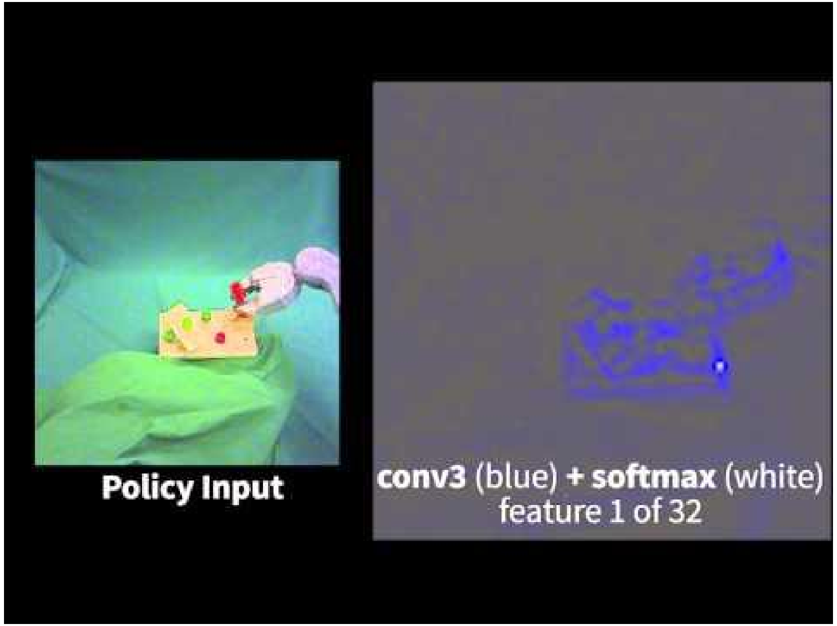
CVPR'15 paper and code + models

Visual Sequence Tasks



Jeff Donahue et al. CVPR'15

Deep Visuomotor Control



example experiments

feature visualization

Sergey Levine & Chelsea Finn*,*
Trevor Darrell, and Pieter Abbeel

Thanks to the Caffe Crew









...plus the cold-brew

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Jonathan Long,
Sergey Karayev, Ross Girshick, Sergio Guadarrama, Ronghang Hu, Trevor Darrell

and our [open source contributors]!

Acknowledgements

Thank you to the Berkeley Vision and Learning Center and its Sponsors

Thank you to NVIDIA
for GPUs, cuDNN collaboration,
and hands-on cloud instances

Thank you to A9 and AWS
for a research grant for Caffe dev
and reproducible research

Thank you to our 200+
open source contributors
and vibrant community!

Hands-on Tutorial on Deep Learning and Caffe

Want to get a jump start in using convolutional neural networks (CNNs) for vision applications?

Sign up for a day-long tutorial on CNNs for deep learning with hands-on lab training on the Caffe software framework.

- *How CNNs work, and how to use them for vision*
- *How to use Caffe to design, train, and deploy CNNs*

September 22nd, 9 am to 5 pm, in **Cambridge, Massachusetts**

Register at [*http://www.embedded-vision.com/caffe-tutorial*](http://www.embedded-vision.com/caffe-tutorial)

- **Use promo code "CNN16-0824" for a 10% discount**