The Impact of SqueezeNet

Bichen Wu and Kurt Keutzer

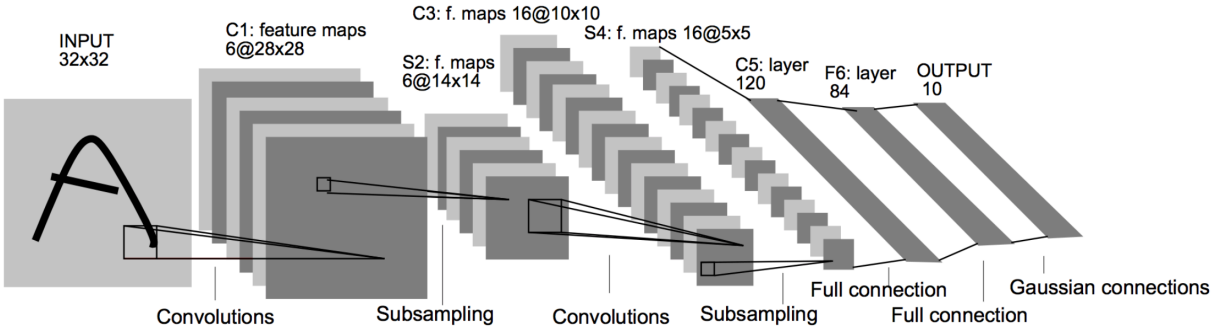
& the **PALLAS** group at UC Berkeley
Amir Gholami, Peter Jin, Alvin Wan,
Bichen Wu, Xiangyu Yue, Yang You, and Sicheng Zhao

as well as recent grads at
DeepScale
Forrest Iandola, Matthew Moskewicz, Anting Shen
and Sammy Sidhu

and visitors
Ali Jannesari (TU Darmstadt) and Kiseok Kwon (Samsung)

keutzer@berkeley.edu

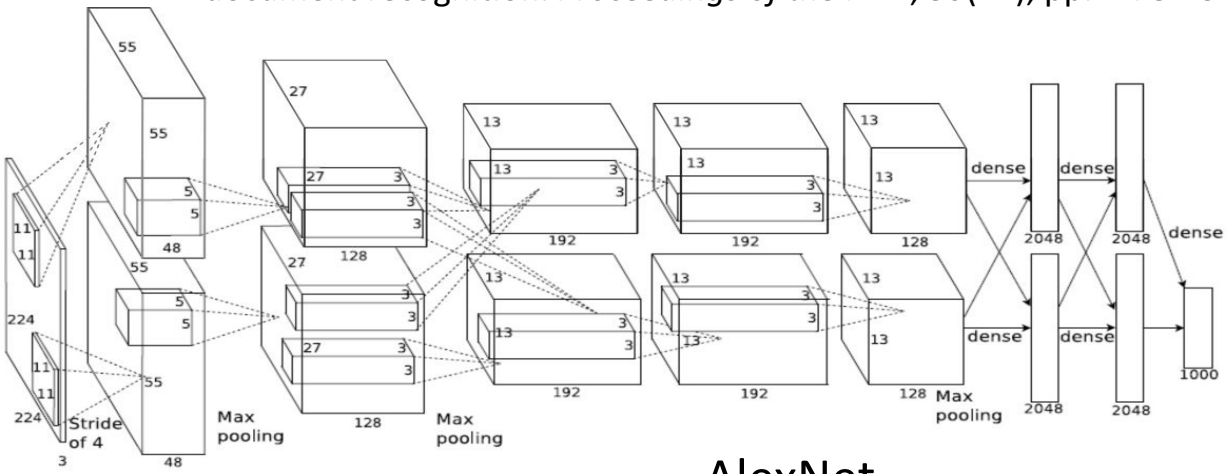
Evolution of CNNs: LeNet vs AlexNet (140x)



LeNet 5

7 layers
431K parameters
4.6 MFLOPs/Inference
0.8 TFLOPs/Epoch

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), pp.2278-2324.

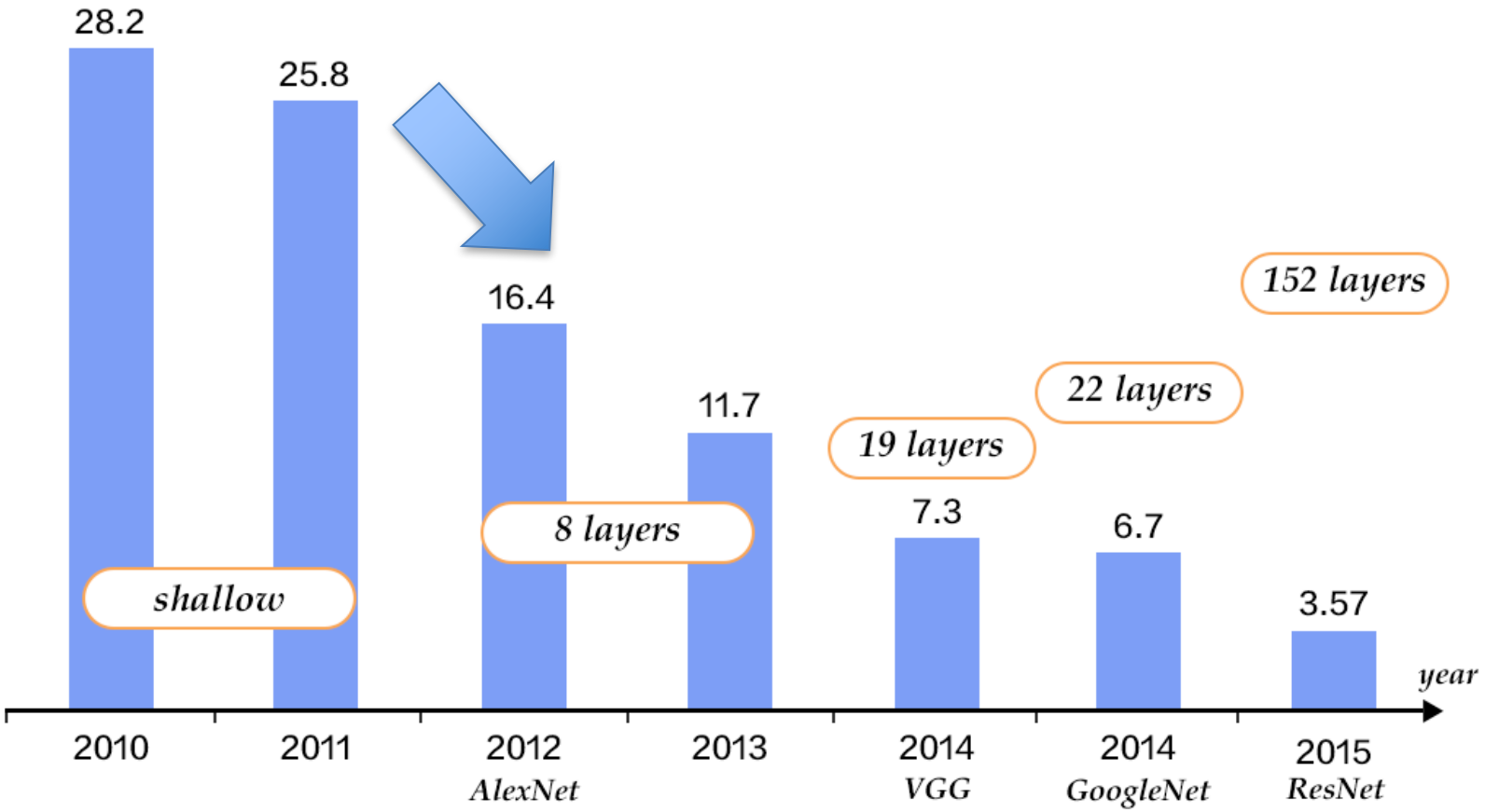


AlexNet

8 layers
61M parameters
1.5 GFLOPs/Inference
5.4 PFLOPs/Epoch

Krizhevsky, A., Sutskever, I. and Hinton, G. E., ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012: Neural Information Processing Systems, Lake Tahoe, Nevada

Accuracy Improvement after AlexNet



Source: http://paddlepaddle.org/docs/develop/book/03.image_classification/index.html

Mainstream Computer Vision Research

- Focused on fundamental issues such as how to create vision systems that equal or surpass humans in their ability to comprehend their environment
- Leads to a preeminent concern on accuracy on whatever is the latest thing – e. g. image captioning
- "only a small subset of papers discuss running time in any detail"
  - J. Huang, *Speed/Accuracy Trade-offs for Modern Convolutional Object Detectors,* 2016.

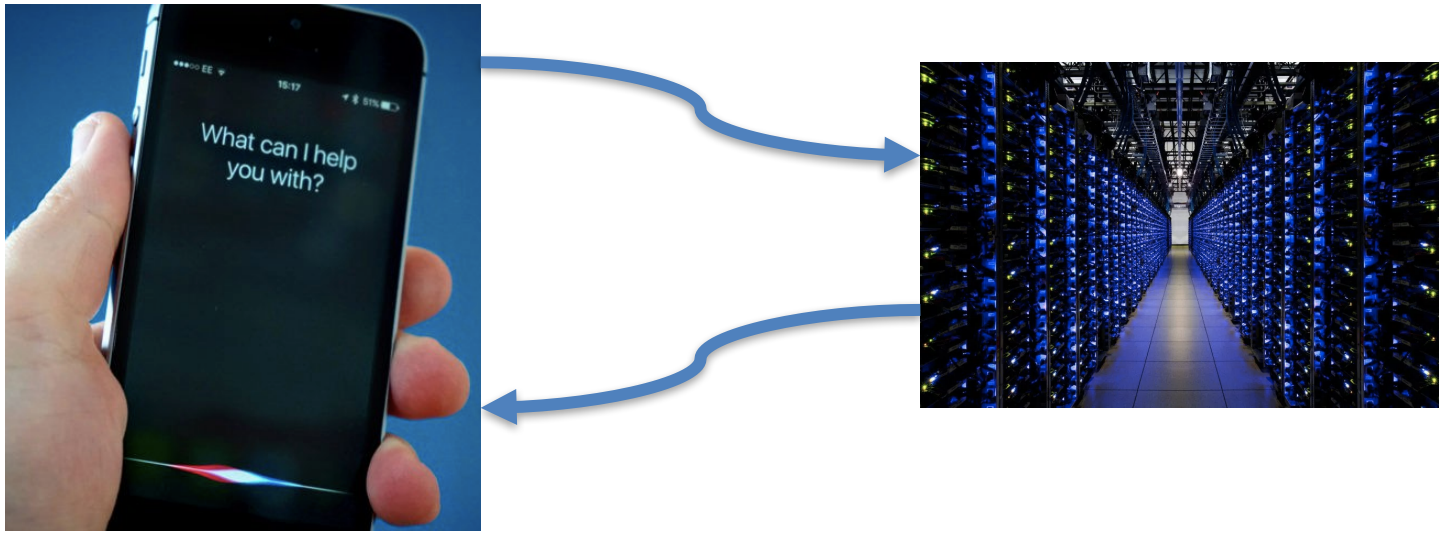
8 x P100

PC for Computer Vision Researchers
~ 80 TeraFlops

DGX-1

Google Data Center

2015: Most Apps Using DNNs Run on Clusters or the Cloud



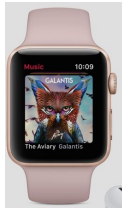
- Economics and technology of client-cloud interactions is complicated – another problem we have been working on for a decade
- We want client-centric apps because:
  - Privacy
  - Low latency – as a requirement or better user experience
  - "Always on" reliability – even there's no network connection available
  - Transmission cost for Internet-of-things (IOT) applications

We want the accuracy of CNNs/DNNs but within embedded constraints

DGX-1
130-170 TFLOPS
3200 Watts
128 GB

TitanX
11 TFLOPS
223 Watts
12GB

Smartphones
800 MFLOPs
3 Watts
2-4GB



IOT Devices
100's MHz
<1Watt
<1GB

Experimental
Level 5
Urban Taxi
KiloWatts

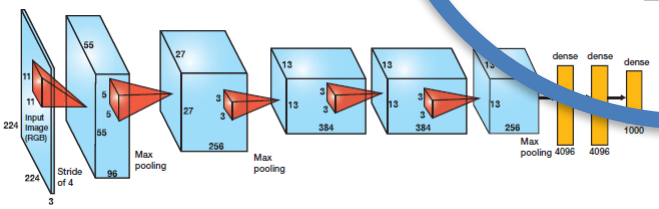
Level 4-5
Urban Taxi
100's Watts

Level 1-3
Passenger
10's of watts

Individual Sensors
500mW – 5W

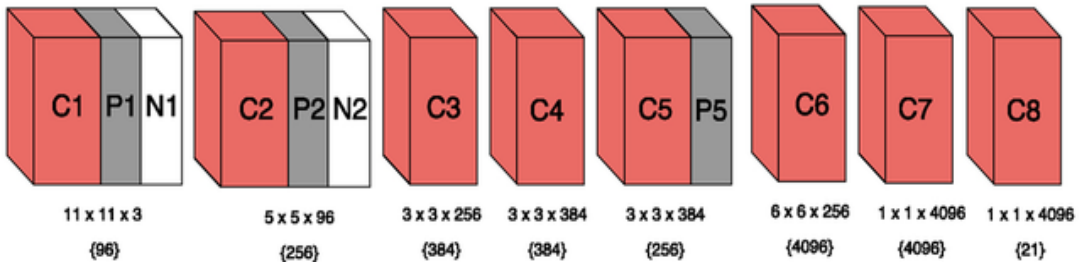
- What can we do with 1000x less speed and 100x less power?
  - 11 TFLOPS → 800 MFLOPS – 223 Watts to 3 Watts

Maybe It's Time to Re-evaluate More Complex Nets

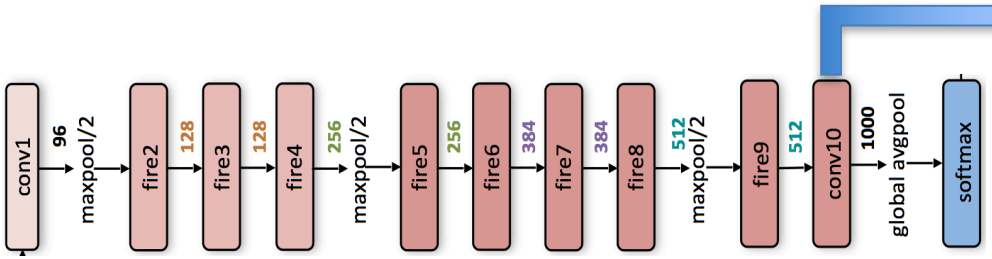


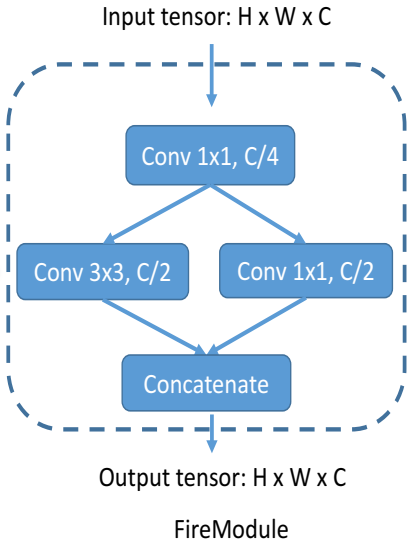
Equivalent Accuracy 50x Smaller SqueezeNet

AlexNet [1]



SqueezeNet [2]





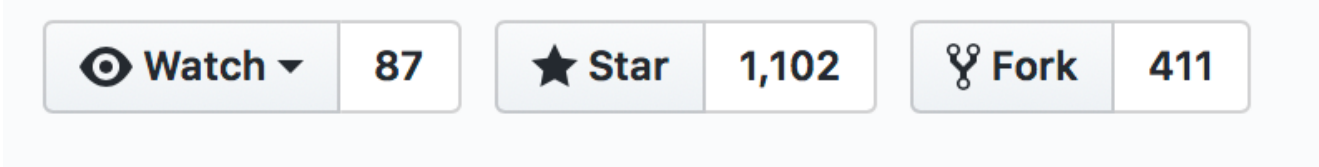
CNN	Top-5 Accuracy ImageNet	Model Parameters	Model Size	After Deep Compression
AlexNet[1]	80.3%	60M	243MB	6.9MB
SqueezeNet[2]	80.3%	1.2M	4.8MB	0.47MB

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012. APA

[2] Iandola, Forrest N., et al. "**SqueezeNet**: AlexNet-level accuracy with 50x fewer parameters and< 1MB model size." arXiv preprint arXiv: 1602.07360 (2016). (February 2016)

Impact of SqueezeNet

Github stars:



Paper citations:

SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size

FN Iandola, S Han, MW Moskewicz, K Ashraf… - arXiv preprint arXiv …, 2016 - arxiv.org

Abstract: Recent research on deep neural networks has focused primarily on improving accuracy. For a given accuracy level, it is typically possible to identify multiple DNN architectures that achieve that accuracy level. With equivalent accuracy, smaller DNN

☆ ❞ Cited by 226 Related articles All 12 versions

Impact of SqueezeNet

- SqueezeNet in deep learning frameworks
- SqueezeNet showcased on embedded processors
- SqueezeNet in mobile software development kits
- SqueezeNet-based mobile applications
- Squeezing becomes a meme for mobile applications
- SqueezeNet in education
- Life after SqueezeNet: SqueezeNext, ShiftNet

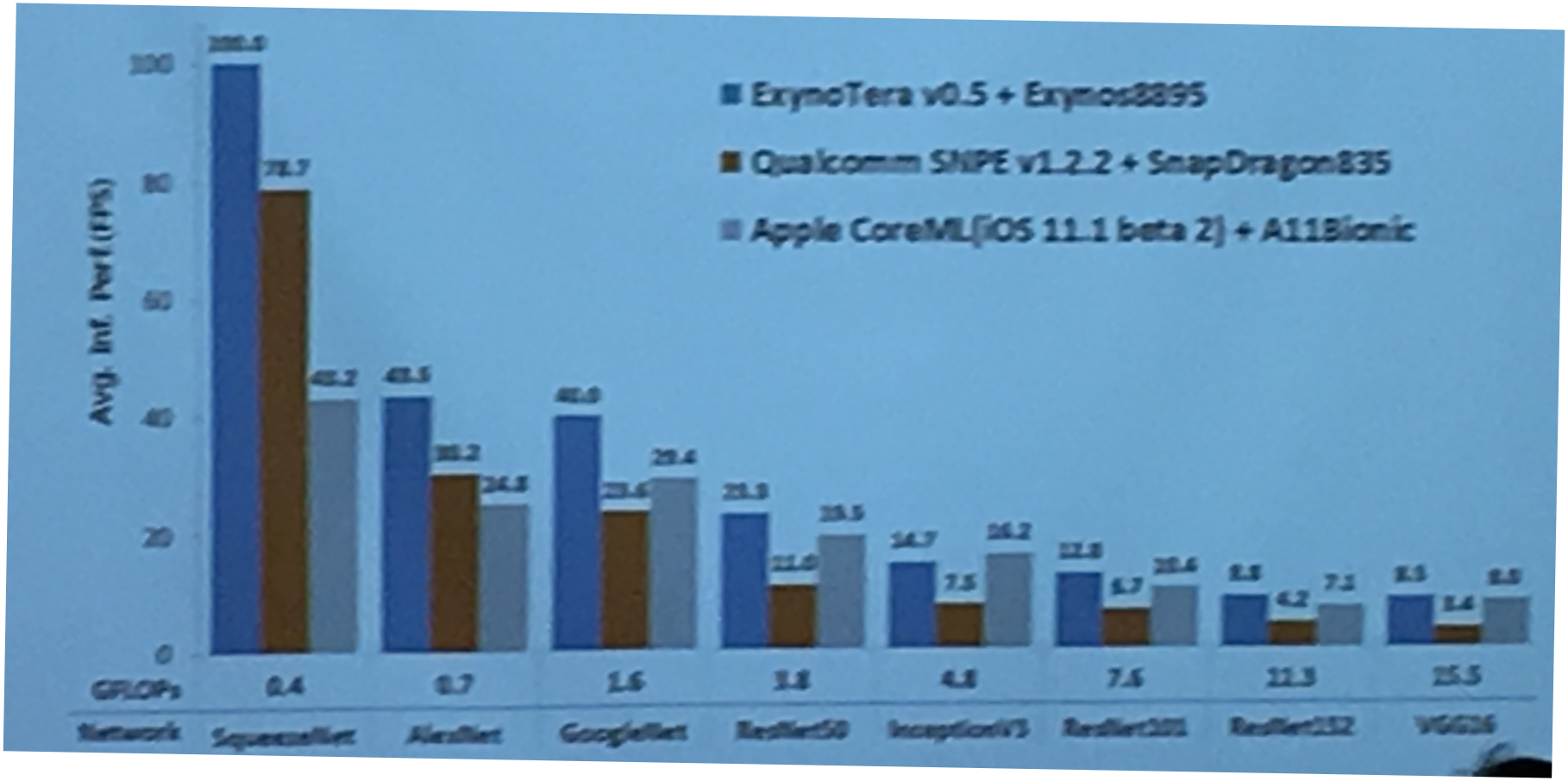
SqueezeNet ported to DL frameworks

Caffe

Impact of SqueezeNet

- SqueezeNet in deep learning frameworks
- SqueezeNet showcased on embedded processors
- SqueezeNet in mobile software development kits
- SqueezeNet-based mobile applications
- Squeezing becomes a meme for mobile applications
- SqueezeNet in education
- Life after SqueezeNet: SqueezeNext, ShiftNet

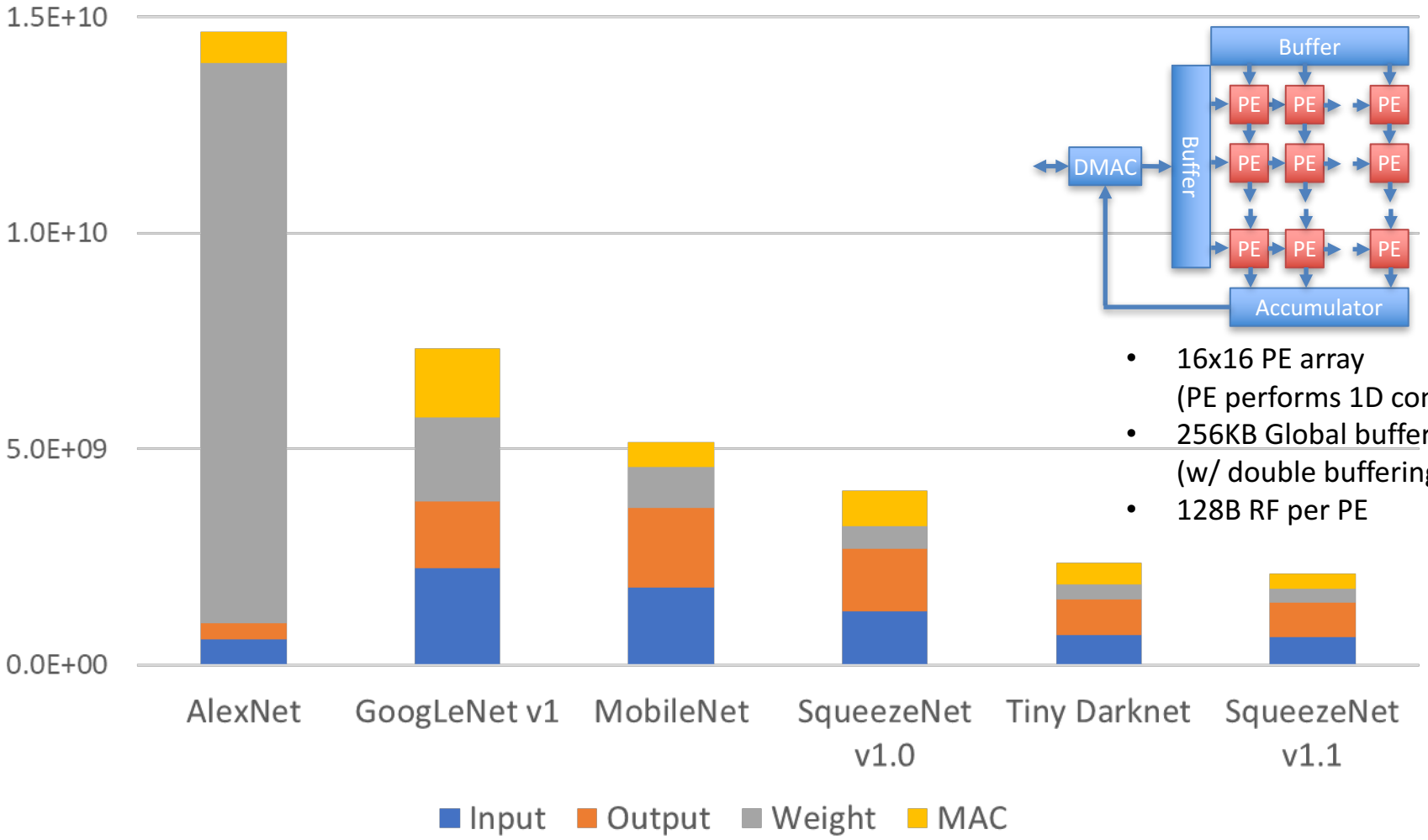
SqueezeNet 100fps on ExynoTera



SqueezeNet is 50x smaller than AlexNet

SqueezeNet is 12x Faster than ResNet 152

Normalized Energy Consumption

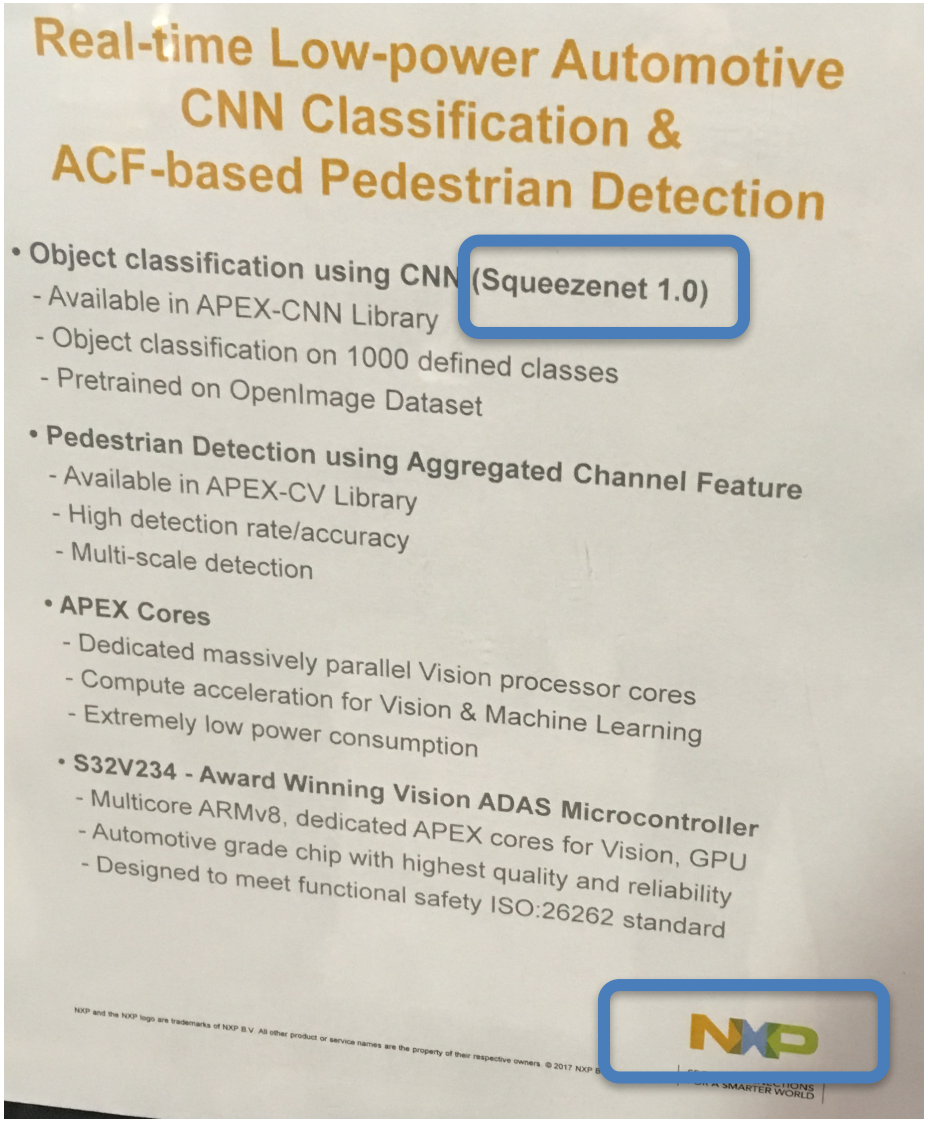


SqueezeNet 2 – 10X more energy efficient than other popular nets

Kiseok Kwon: Samsung Digital Media City

NXP Demos at Embedded Vision Summit

Introduced SqueezeNet for object classifcation to sponsors at BDD opening 3/20/2016

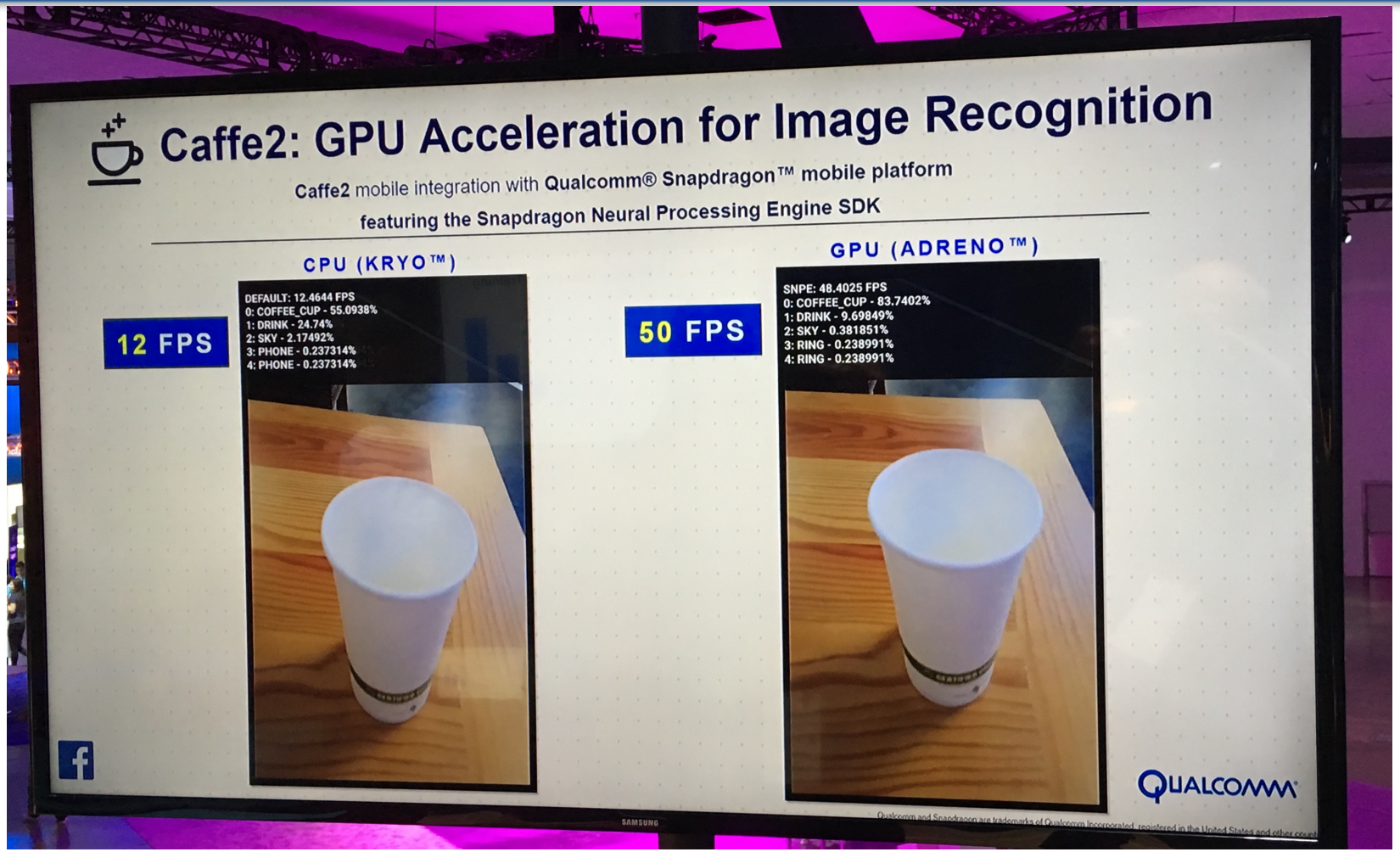


NXP at Embedded Vision Summit

Speed: 20 FPS
Power: 500 mW
For more NXP demo: https://www.nxp.com

NXP demo

Qualcomm demo at F8



- QualComm and FB collaborate to show speedups on SqueezeNet at F8

ARM

Enabling Embedded Inference Engine with the ARM Compute Library: A Case Study

Dawei Sun, Shaoshan Liu*, and Jean-Luc Gaudiot

If you need to enable deep learning on low-cost embedded SoCs, should you port an existing deep learning framework or should you build one from scratch? In this paper, we seek to answer this question by sharing our practical experience of building an embedded inference engine using the ARM Compute Library (ACL). The results show that, contradictory to conventional wisdom, for simple models, it takes much less development time to build an inference engine from scratch as opposed to porting existing frameworks. In addition, by utilizing ACL, we managed to build an inference engine that outperforms TensorFlow by 25%. Our conclusion is that, with embedded devices, we must

Paper: Sun, Dawei, Shaoshan Liu, and Jean-Luc Gaudiot. "Enabling Embedded Inference Engine with ARM Compute Library: A Case Study." *arXiv preprint arXiv:1704.03751* (2017).

Code: https://github.com/ARM-software/ComputeLibrary

FPGA

Accelerating SqueezeNet on FPGA

by Megha Arora and Samyukta Lanka.

Final | Initial Proposal

Project Summary

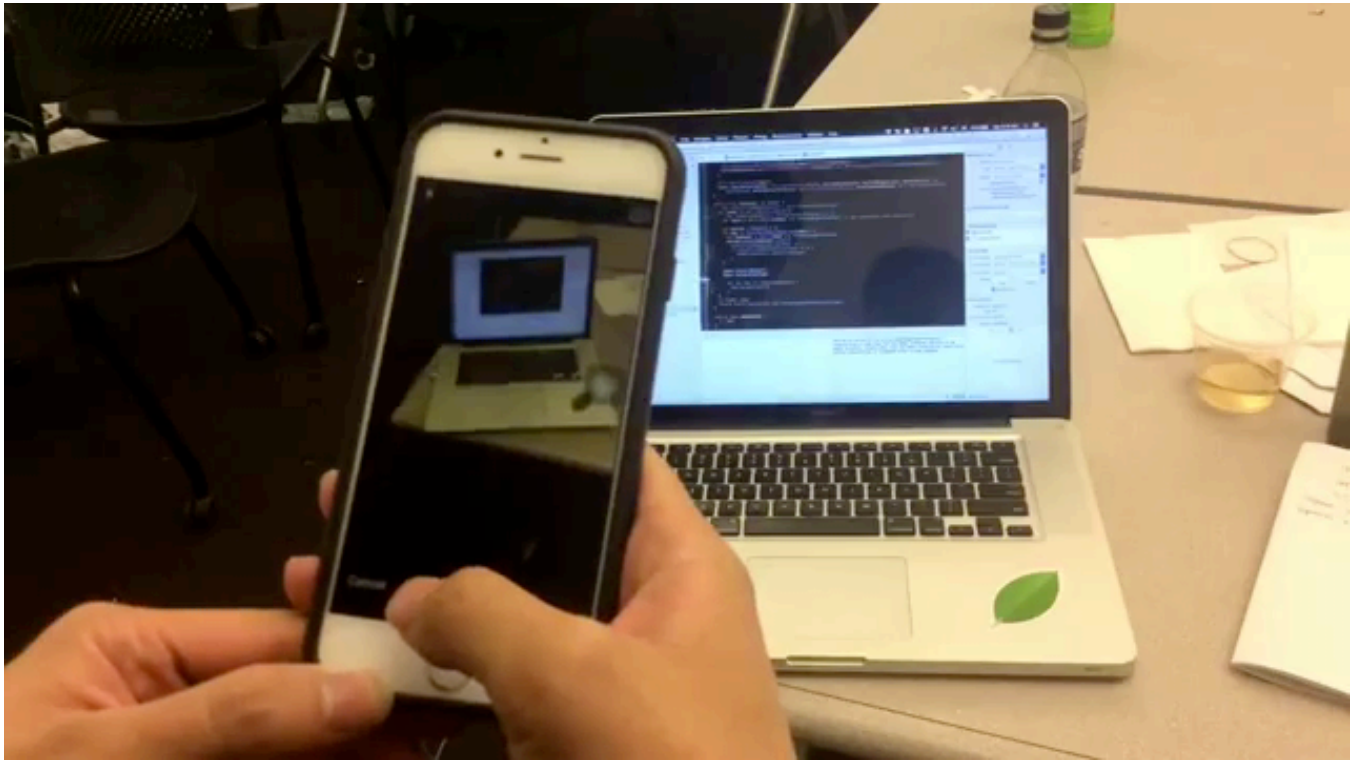
We have successfully been able to accelerate the SqueezeNet on Zybo Zynq-7020 FPGA. Our implementation (when evaluated using the ILSVRC2012 ImageNet data) is faster and more energy efficient as compared to our baseline!

Link: https://lankas.github.io/15-618Project/
Code: https://github.com/lankas/SqueezeNet

Impact of SqueezeNet

- SqueezeNet in deep learning frameworks
- SqueezeNet showcased on embedded processors
- SqueezeNet in mobile software development kits
- SqueezeNet-based mobile applications
- Squeezing becomes a meme for mobile applications
- SqueezeNet in education
- Life after SqueezeNet: SqueezeNext, ShiftNet

Espresso: a mobile SDK for iPhone6



iPhone 6: CMU 5/2016: http://codinfox.github.io/espresso/

- Zhihao Li and Zhenrui Zhang
- 1st prize in 2016 CMU Annual Parallel Competition

SqueezeNet in Apple's CoreML

```
        AVCaptureConnection) {
42 //      print("Camera was able to capture a frame:", Date())
43
44      guard let pixelBuffer: CVPixelBuffer =
            CMSampleBufferGetImageBuffer(sampleBuffer) else
            { return }
45
46      guard let model = try? VNCoreMLModel(for:
            SqueezeNet().model) else { return }
47      let request = VNCoreMLRequest(model: model)
            { (finishedReq, err) in
48
49          //perhaps check the err
50
51 //           print(finishedReq.results)
52
53          guard let results = finishedReq.results as?
                [VNClassificationObservation] else { return }
54
55          guard let firstObservation = results.first else
                { return }
56
57          print(firstObservation.identifier,
                firstObservation.confidence)
```

SqueezeNet in Baidu's MDL (mobile deep learning)

baidu / mobile-deep-learning

Watch 207 | Star 3,274 | Fork 534

Code | Issues 18 | Pull requests 0 | Projects 0 | Wiki | Insights

This research aims at simply deploying CNN(Convolutional Neural Network) on mobile devices, with low complexity and high speed.

mobile deep-learning neon cnn neural-network arm ios android googlenet mobilenet squeezenet

Mobile-deep-learning （MDL）

license MIT License build passing

Free and open source mobile deep learning framework, deploying by Baidu.

This research aims at simply deploying CNN on mobile devices, with low complexity and high speed. It supports calculation on iOS GPU, and is already adopted by Baidu APP.

- Size: 340k+ (on arm v7)
- Speed: 40ms (for iOS Metal GPU Mobilenet) or 30 ms (for Squeezenet)

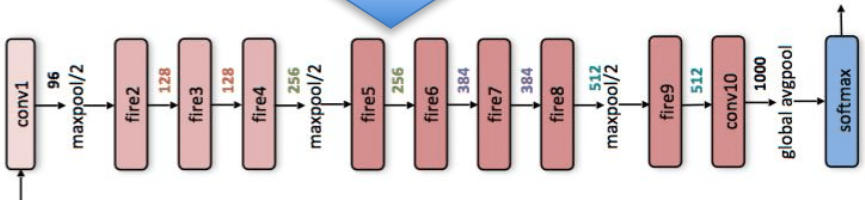
Impact of SqueezeNet

- SqueezeNet in deep learning frameworks
- SqueezeNet showcased on embedded processors
- SqueezeNet in mobile software development kits
- SqueezeNet-based mobile applications
- Squeezing becomes a meme for mobile applications
- SqueezeNet in education
- Life after SqueezeNet: SqueezeNext, ShiftNet

Style Transfer using SqueezeNet

Original model[1] based on VGG 19: **575MB**

>100X reduction



Efficient model[2] based on SqueezeNet: **4.8MB**

Now you can run it locally:

- Interactively
- Without cloud access

(if you're a teenager)

[1] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "A neural algorithm of artistic style." *arXiv preprint arXiv:1508.06576*(2015).
[2] https://github.com/lizeng614/SqueezeNet-Neural-Style-Pytorch

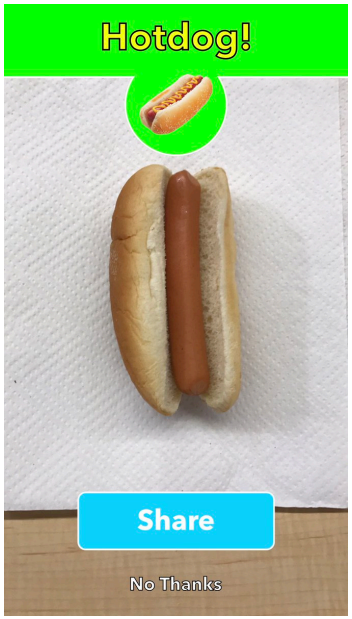
America's Favorite Mobile App: *Not Hotdog*

Tim Anglade Follow

Startup guy working on the TV show Silicon Valley—timanglade@gmail.com

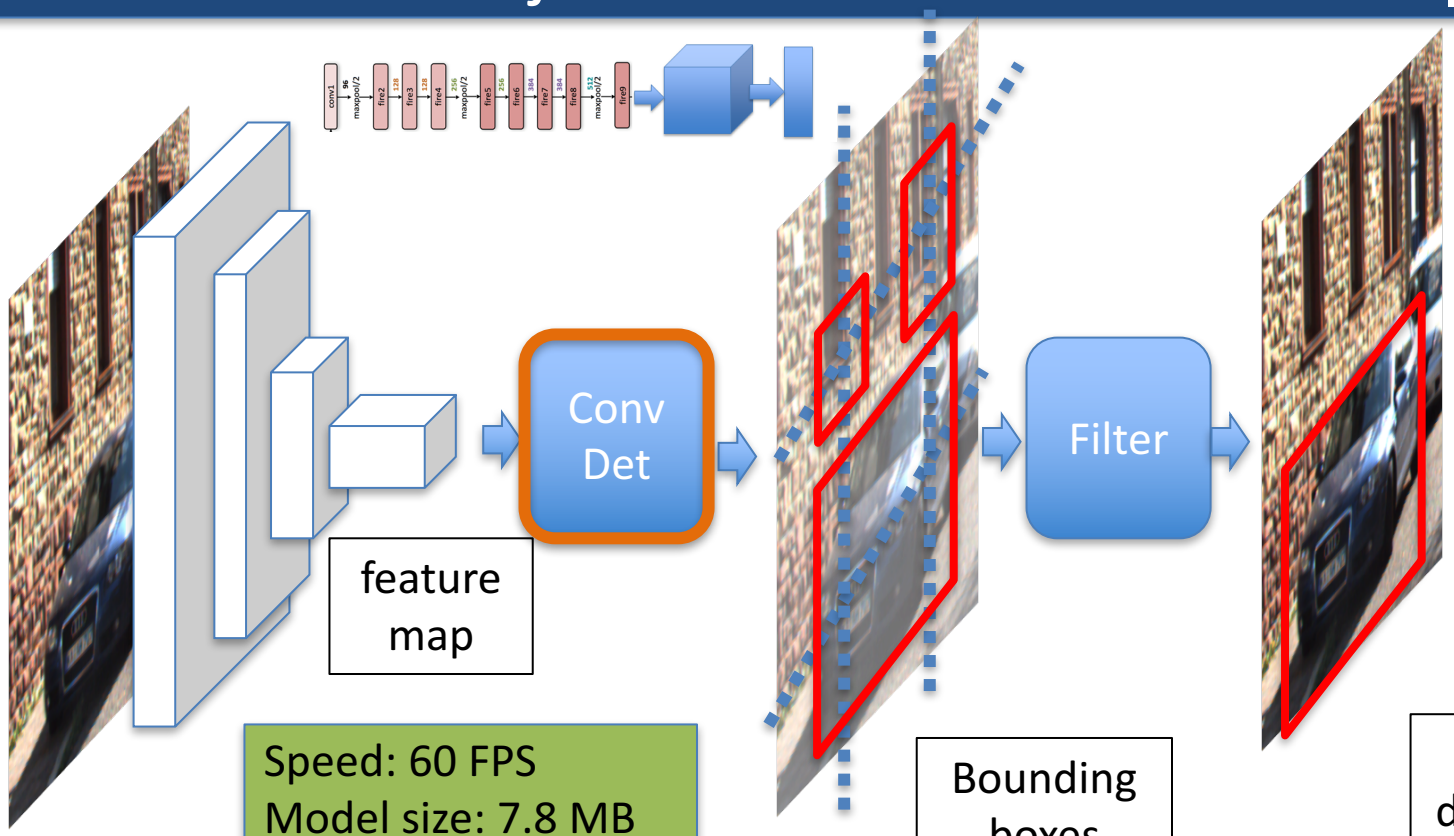
Jun 26 · 23 min read

How HBO's Silicon Valley built "Not Hotdog" with mobile TensorFlow, Keras & React Native

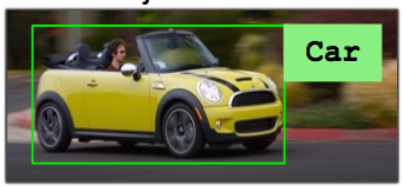


- SqueezeNet powers Version 2 of the *Not Hotdog* app from the Silicon Valley TV show.
- A variant of MobileNets powers Version 3.

SqueezeDet for Object Detection



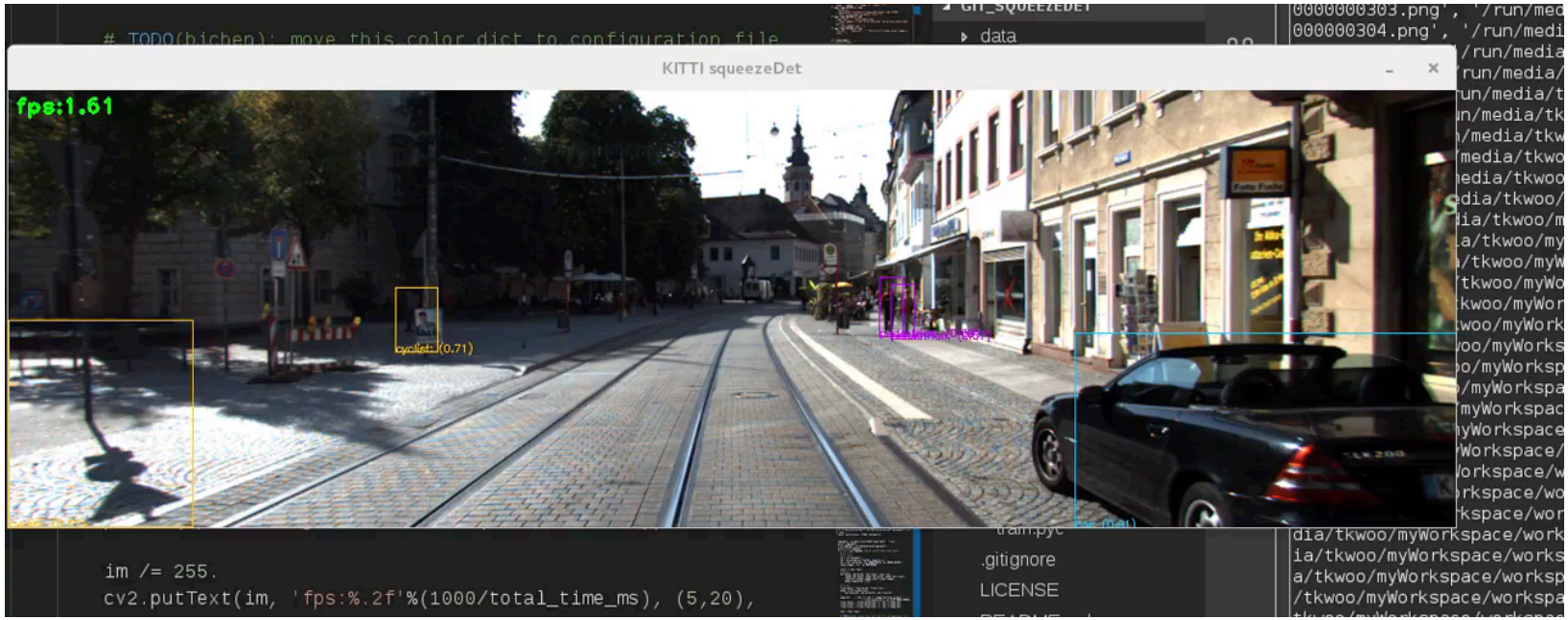
Object *detection*



Best Paper Award: Bichen Wu, Forrest Iandola, Peter H. Jin, and Kurt Keutzer. 2017. SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In Proceedings, CVPR Embedded Computer Vision Workshop, July 2017.

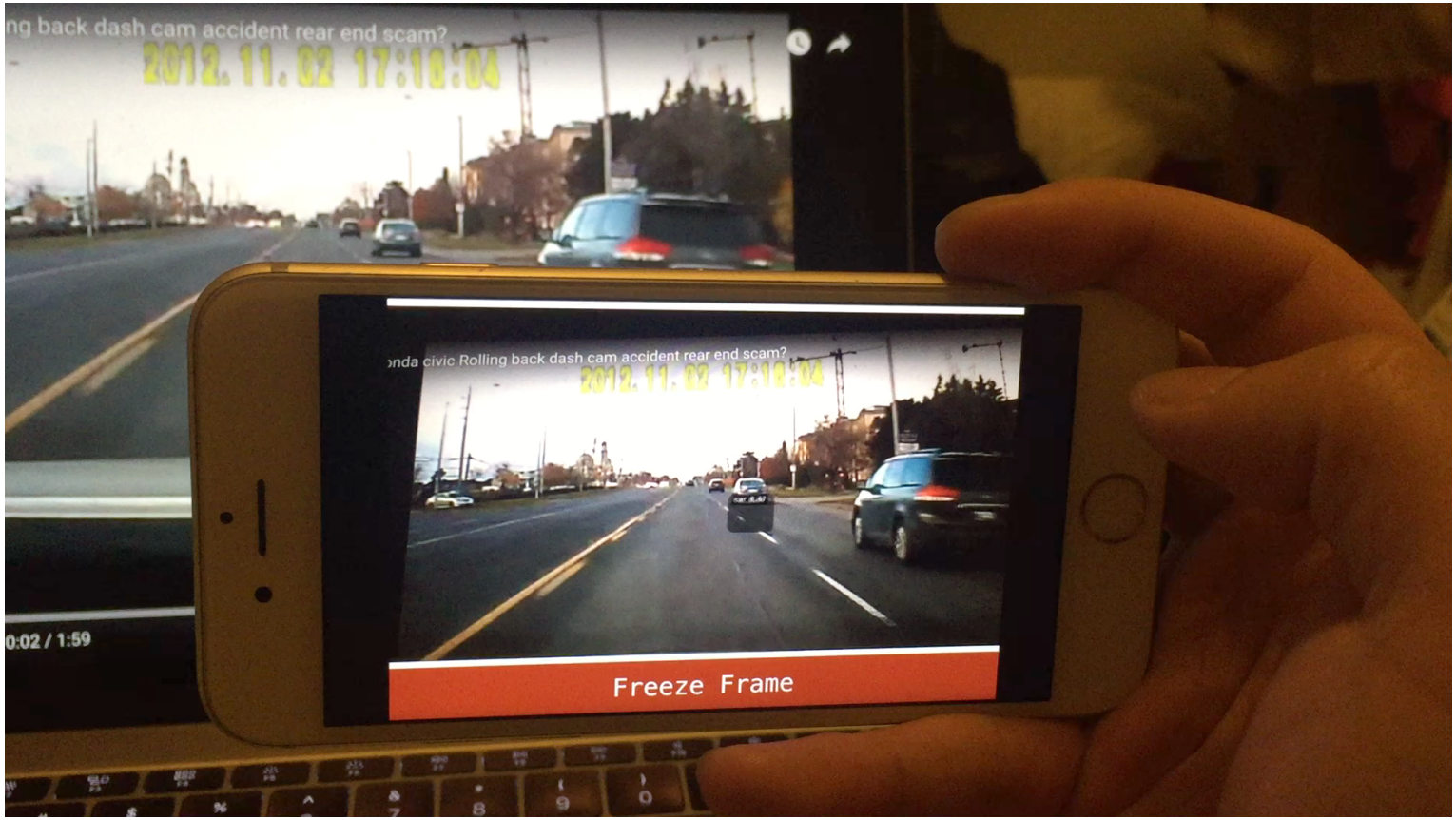
SqueezeDet demo

- Created by Youtuber TK Woo
  - Link: https://youtu.be/O5RcHs9uqVA
  - Search on Youtube: **SqueezeDet Demo**



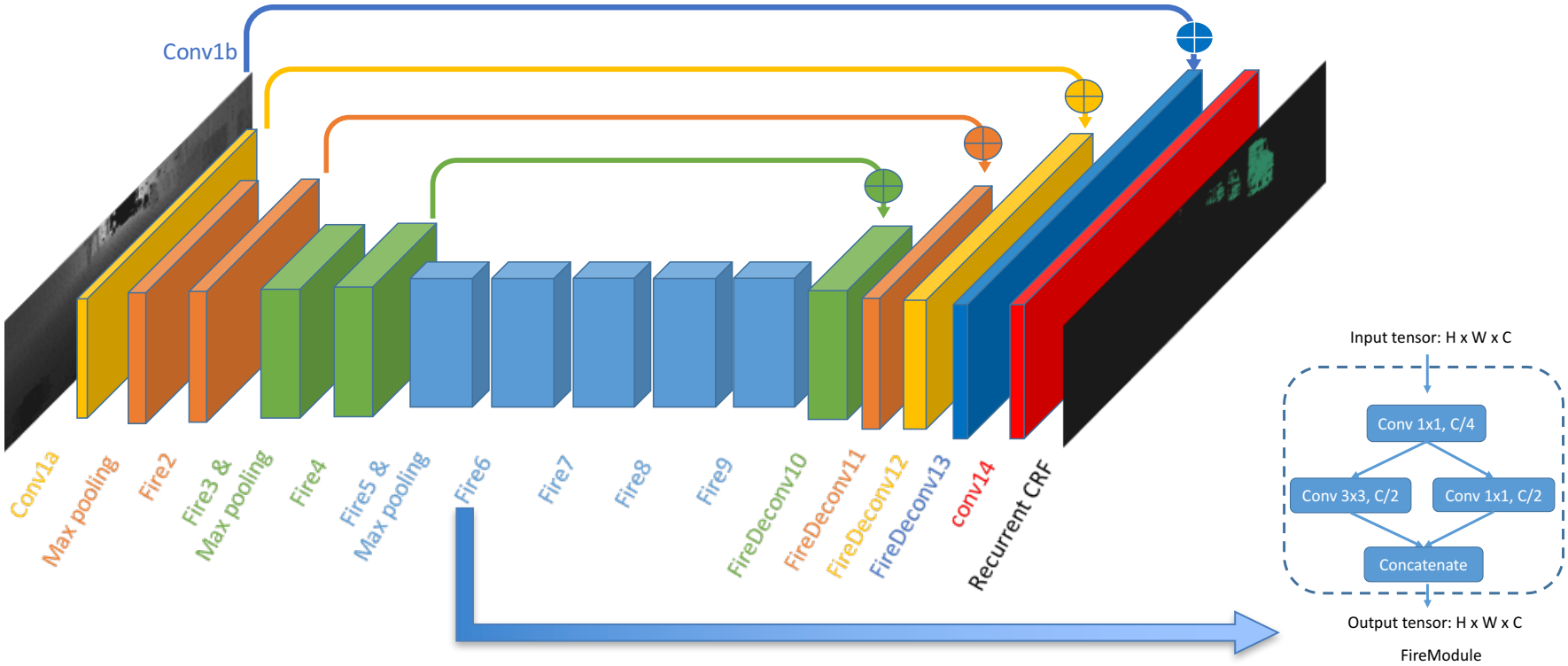
SqueezeDet deployment on mobile devices

- We used Tensorflow Mobile to deploy SqueezeDet on an iPhone 6



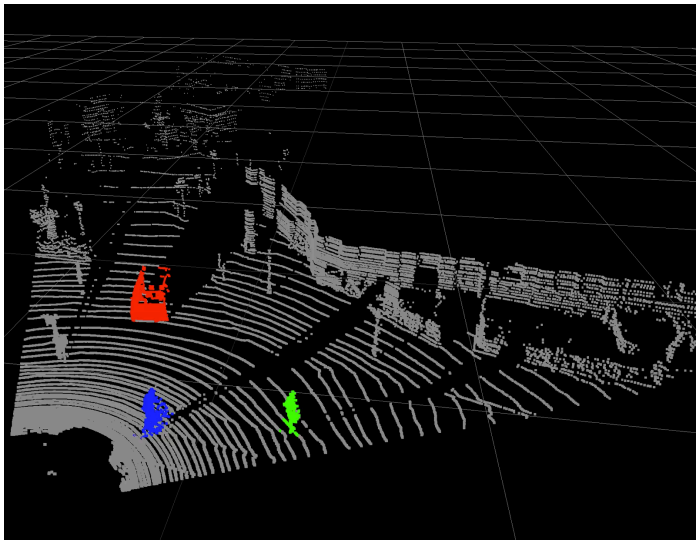
SqueezeSeg for LiDAR Point Cloud Segmentation

- Designed for LiDAR point cloud segmentation for autonomous driving
- Extremely high efficiency (on Titan X maxwell GPU):
  - 114 Frames per second
  - 3.46 MB of parameters
  - 0.7 J per frame

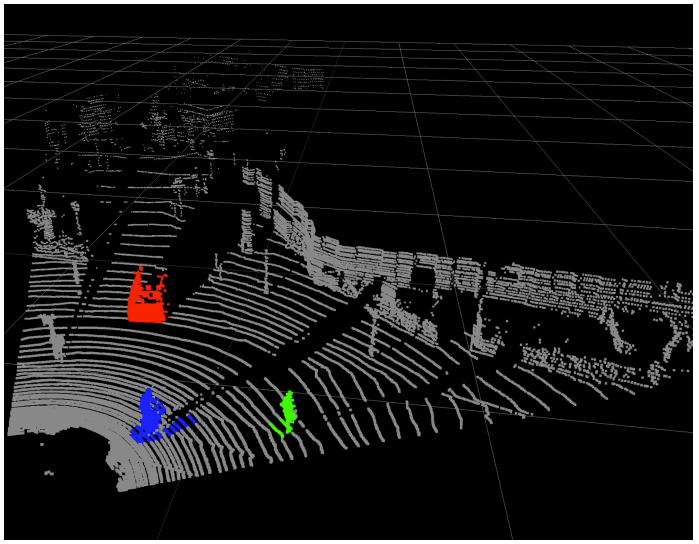


Wu, Bichen, et al. "SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud." *arXiv preprint arXiv:1710.07368*(2017).

SqueezeSeg demo



Ground truth label map



Predicted label map

Video reference

Impact of SqueezeNet

- SqueezeNet in deep learning frameworks
- SqueezeNet showcased on embedded processors
- SqueezeNet in mobile software development kits
- SqueezeNet-based mobile applications
- Squeezing becomes a meme for mobile applications
- SqueezeNet in education
- Life after SqueezeNet: SqueezeNext, ShiftNet

Squeezing AI into Mobile Becomes a Meme

"We're squeezing AI into smartphones."
- Mark Zuckerberg, Keynote, F8

Squeezing Deep Learning into mobile phones
- A Practitioners guide
Anirudh Koul

Dan DeLong/Microsoft When you're far from a cell tower and need to figure out if that bluebird is Sialia sialis or Sialia mexicana, no cloud server is going to help you. That's why companies are squeezing AI onto portable devices, and Microsoft has just taken that to a new extreme by...

Engineers are trying to squeeze outsize AI into mobile systems

Impact of SqueezeNet

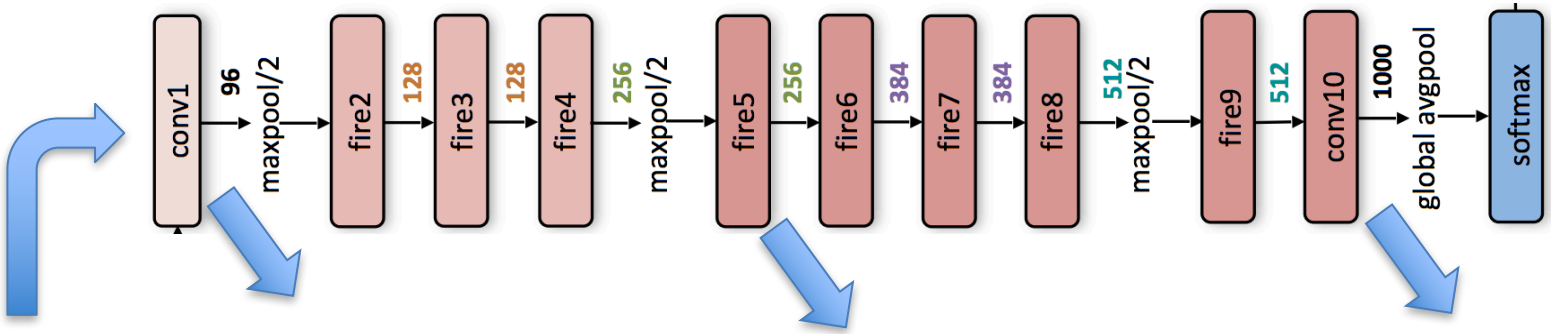
- SqueezeNet in deep learning frameworks
- SqueezeNet showcased on embedded processors
- SqueezeNet in mobile software development kits
- SqueezeNet-based mobile applications
- Squeezing becomes a meme for mobile applications
- SqueezeNet in education
- Life after SqueezeNet: SqueezeNext, ShiftNet

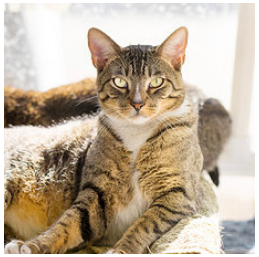
Google Brain's deeplearnjs.org folks use SqueezeNet interactively in a web browser

98ms in Javscript in a web browser!

- Folks from Google Brain have created an interactive tool and code for learning deep learning
- Check out: https://deeplearnjs.org/
- Run SqueezeNet in a web browser and see visualizations of the layers:
  - https://deeplearnjs.org/demos/imagenet/

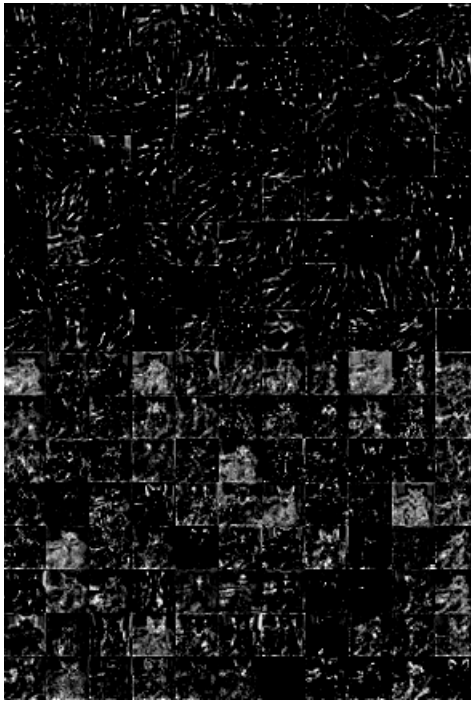
Convolutional Layers in SqueezeNet



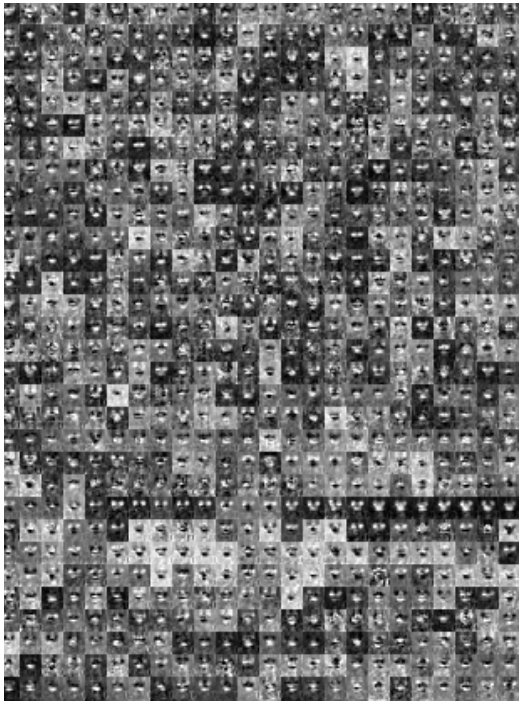


3 channels (RGB)

96 channels



256 channels



1000 channels

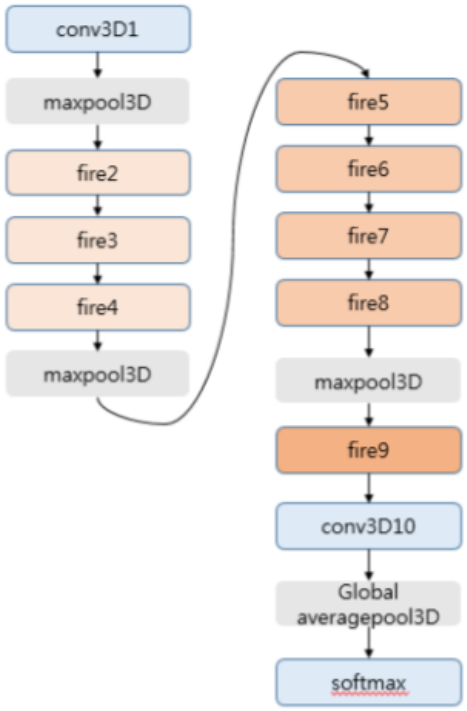
- https://deeplearnjs.org/demos/imagenet/

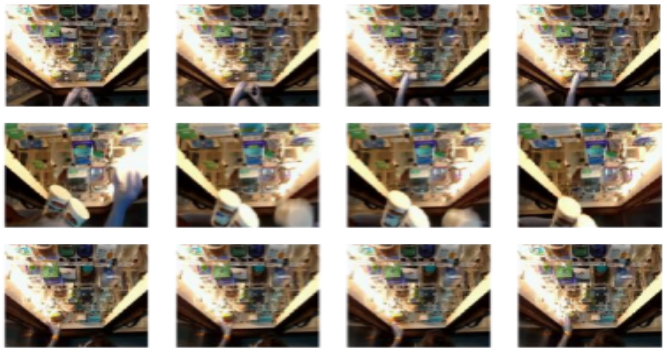
SqueezeNet in CS231 at Stanford

CS231n: Convolutional Neural Networks for Visual Recognition

Spring 2017

- SqueezeNet used in assignment and course projects:





Item removal detection:

http://cs231n.stanford.edu/reports/2017/pdfs/213.pdf

3D SqueezeNet for medical images

http://cs231n.stanford.edu/reports/2017/pdfs/23.pdf

Mice behavior analysis:

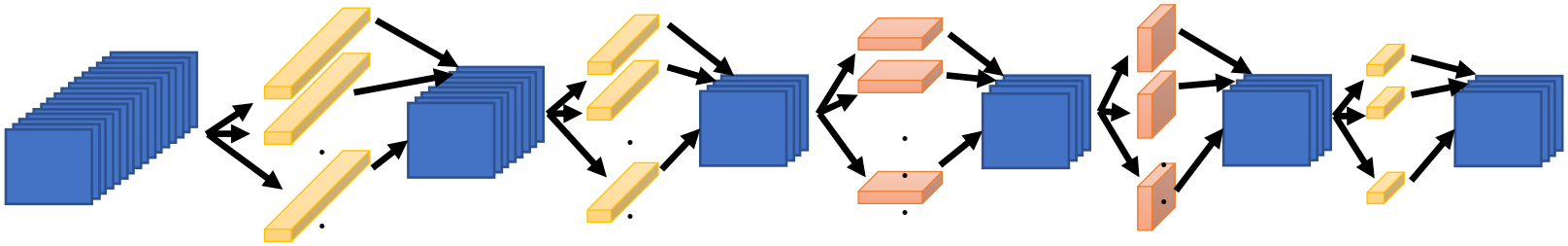
http://cs231n.stanford.edu/reports/2017/pdfs/500.pdf

Impact of SqueezeNet

- SqueezeNet in deep learning frameworks
- SqueezeNet showcased on embedded processors
- SqueezeNet in mobile software development kits
- SqueezeNet-based mobile applications
- Squeezing becomes a meme for mobile applications
- SqueezeNet in education

Life after SqueezeNet: SqueezeNext, ShiftNet

SqueezeNext

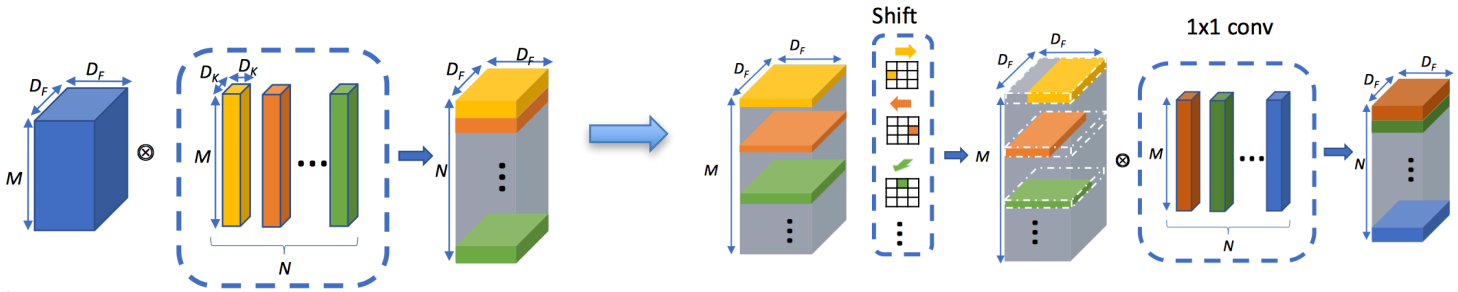


Model	Top-1	Top-5	# Params	Reduction
AlexNet	57.10	80.30	60.9M	1×
SqueezeNet	57.50	80.30	1.2M	50×
1.0-G-SqNext-23	**56.88**	**80.83**	**0.5M**	**120×**
VGG-19	**68.50**	**88.50**	**138M**	1×
2.0-SqNext-34	**68.46**	**88.78**	**3.8M**	**36×**

- Matches AlexNet with **120x** smaller parameters
- Deeper version achieves VGG accuracy with **36x** smaller model

ShiftNet

- A lesson from SqueezeNet: spatial convolution (3x3, 5x5, etc.) is expensive ...
  - Replace spatial convolutions with the "Shift" operation[1] that requires **zero-parameter, zero-FLOPs**



- Classification:

	Top-1 Acc.	Parameter size	Reduction
AlexNet	57.2	60 million	1X
SqueezeNet	57.5	1.2 million	50X
ShiftNet-C	**58.8**	**0.78 million**	**77X**

- Other tasks:
  - Face verification: 37X parameter reduction
  - Style transfer: 6X parameter reduction

[1] Wu B, Wan A, Yue X, Jin P, Zhao S, Golmant N, Gholaminejad A, Gonzalez J, Keutzer K. Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions. arXiv preprint arXiv:1711.08141. 2017 Nov 22.

Concolusion

- The increasing demand for deploying CNNs/DNNs on embedded devices requires “Squeezing” parameter size, computation and energy consumption of neural networks

- SqueezeNet very well addressed the above problem, and has been widely adopted:
  - It’s ported to other deep learning frameworks
  - It’s demonstrated in embedded processors
  - It’s included in many mobile SDKs
  - It powered many mobile applications
  - It’s used for education

- Beyond SqueezeNet:
  - We build SqueezeNext, ShiftNet to achieve better accuracy with smaller model size

Thank you!