CMSC 5743

Efficient Computing of Deep Neural Networks

Lecture 01: Introduction

Bei Yu
CSE Department, CUHK
byu@cse.cuhk.edu.hk

(Latest update: September 2, 2024)

2024 Fall

What We Focus on?

What you expect to Learn?

How About the Workload?

Grading System?

Overview

1. CNN Architecture Overview
2. CNN Energy Efficiency
3. CNN on Embedded Platform

Object Detection: PASCAL VOC mean Average Precision (mAP)

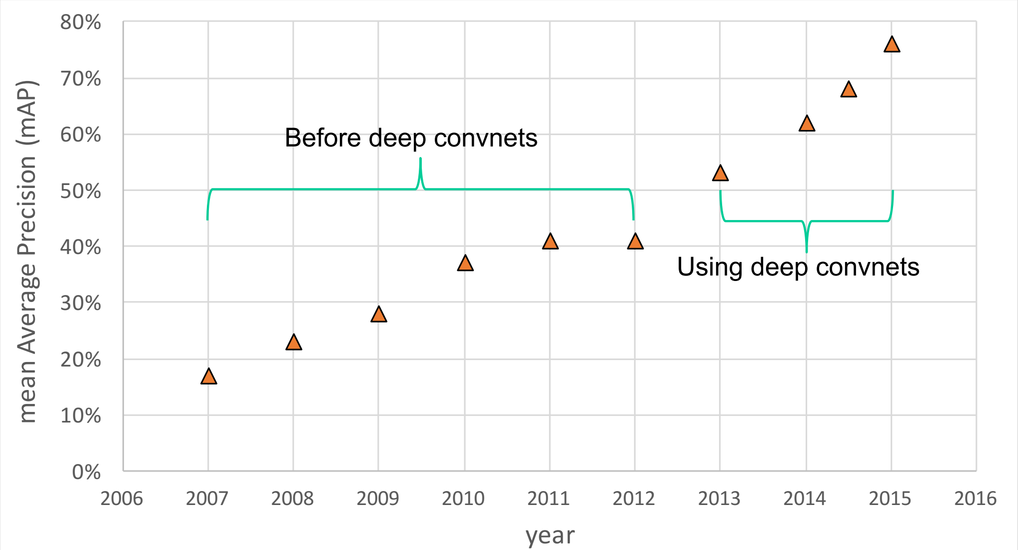
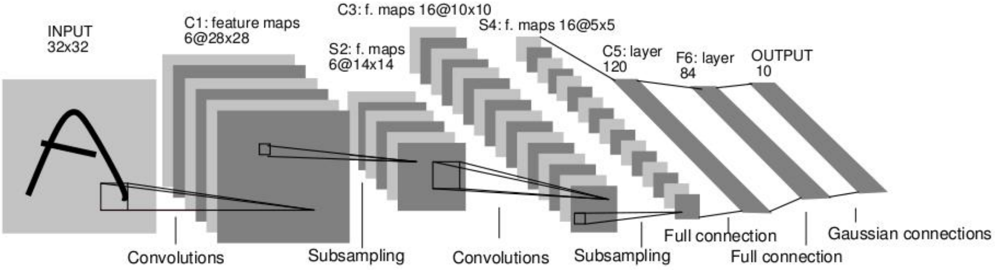


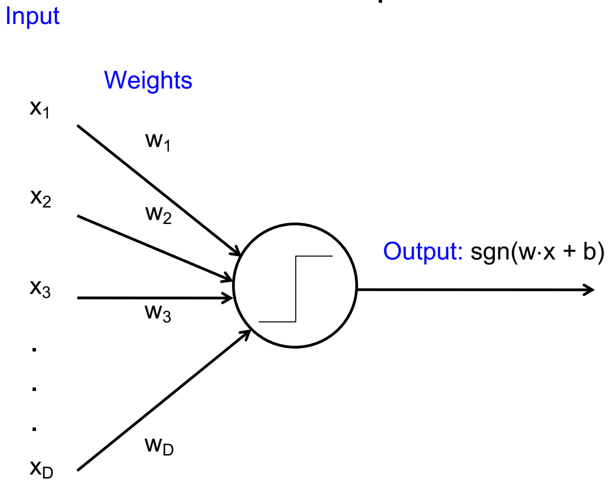
Figure source: Ross Girshick

LeNet 5

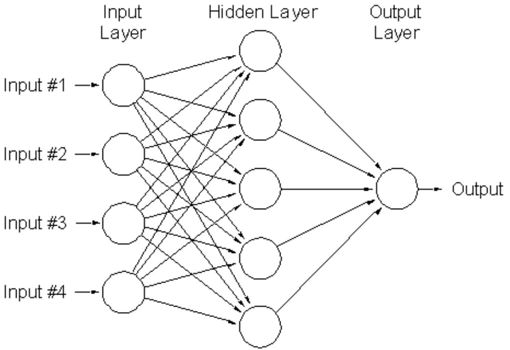


Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86(11): 2278–2324, 1998.

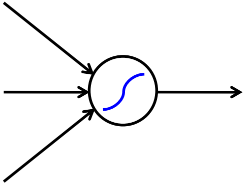
The Perceptron



Rosenblatt, Frank (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, Psychological Review, v65, No. 6, pp. 386–408.

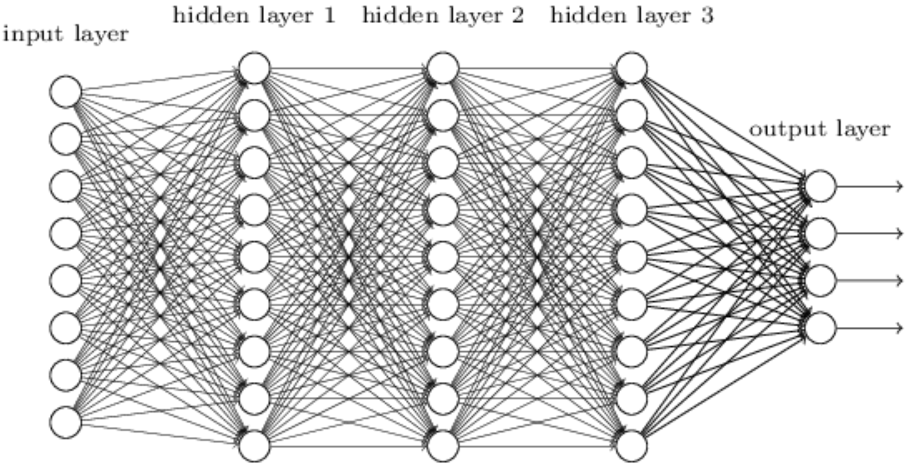


- Can learn nonlinear functions provided each perceptron has a differentiable nonlinearity



Sigmoid: $g(t) = \dfrac{1}{1+e^{-t}}$

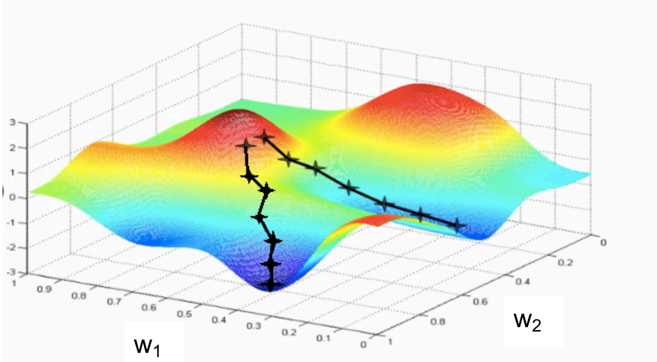
Multi-layer neural network



- Find network weights to minimize the *training error* between true and estimated labels of training examples, e.g.:

$$E(\mathbf{w}) = \sum_{i=1}^{N} \left(y_i - f_{\mathbf{w}}(\mathbf{x}_i) \right)^2$$

- Update weights by **gradient descent:** $\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial E}{\partial \mathbf{w}}$



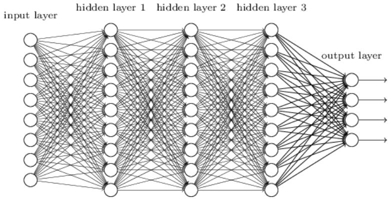
- Find network weights to minimize the *training error* between true and estimated labels of training examples, e.g.:

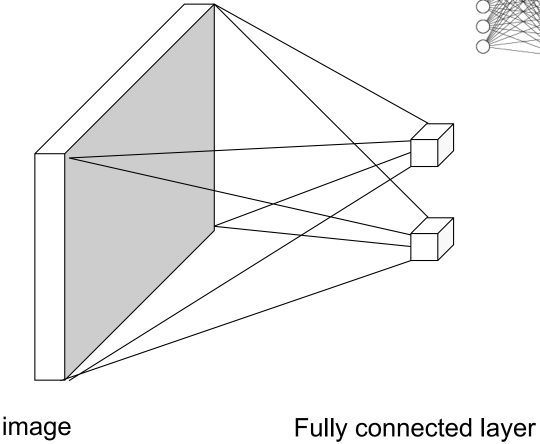
$$E(\mathbf{w}) = \sum_{i=1}^{N} \left(y_i - f_{\mathbf{w}}(\mathbf{x}_i) \right)^2$$

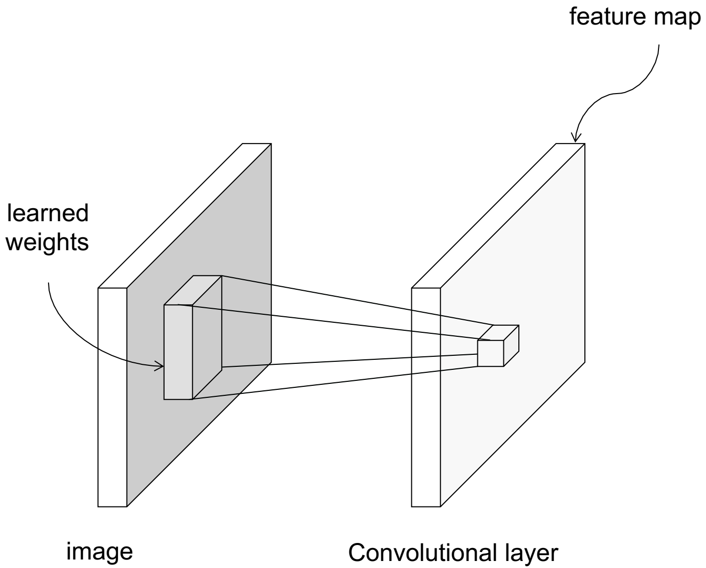
- Update weights by **gradient descent:** $\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial E}{\partial \mathbf{w}}$

- **Back-propagation:** gradients are computed in the direction from output to input layers and combined using chain rule
- **Stochastic gradient descent:** compute the weight update w.r.t. one training example (or a small batch of examples) at a time, cycle through training examples in random order in multiple epochs

From fully connected to convolutional networks

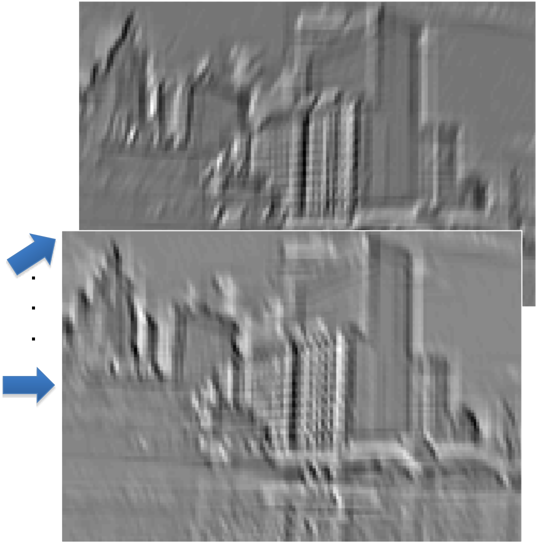





feature map
learned
weights
image
Convolutional layer

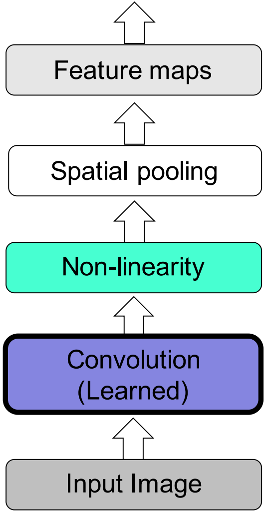
Convolution as feature extraction

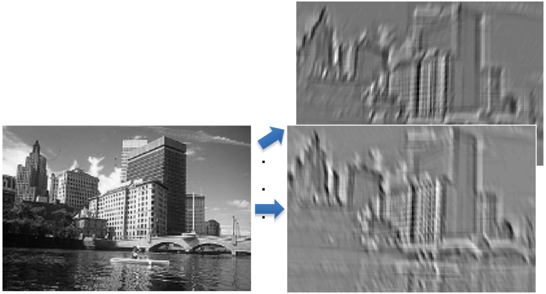
Input



Feature Map

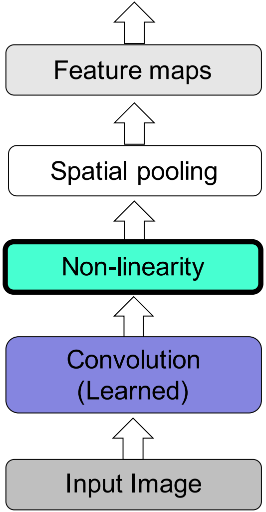
Key operations



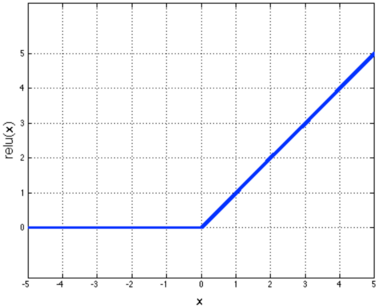


Input Feature Map

Source: R. Fergus, Y. LeCun

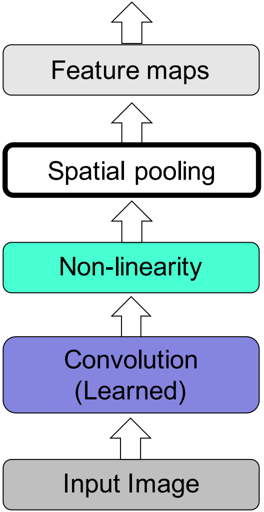


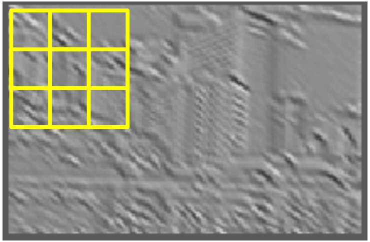
Rectified Linear Unit (ReLU)



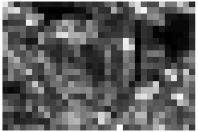
Source: R. Fergus, Y. LeCun

Key operations





Max



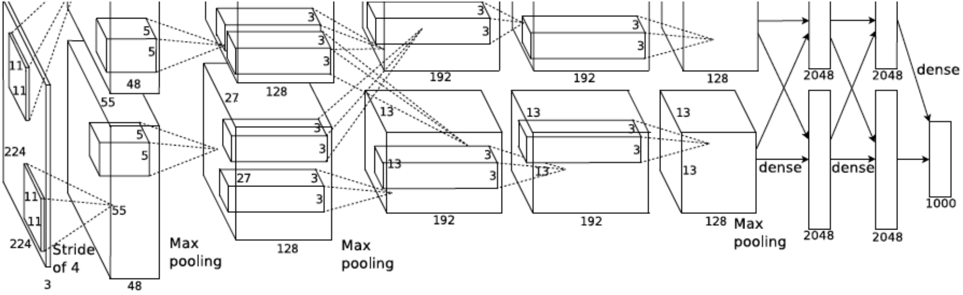
Source: R. Fergus, Y. LeCun

Fast forward to the arrival of big visual data

- ~14 million labeled images, 20k classes
- Images gathered from Internet
- Human labels via Amazon MTurk
- ImageNet Large-Scale Visual Recognition Challenge (ILSVRC):
  1.2 million training images, 1000 classes

www.image-net.org/challenges/LSVRC/

AlexNet: ILSVRC 2012 winner



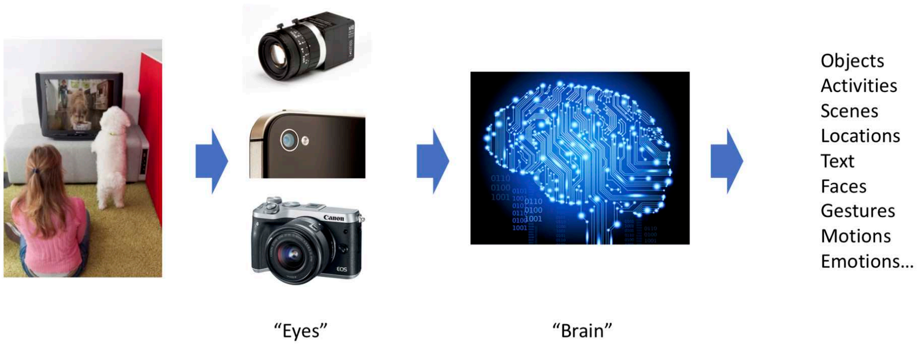
- Similar framework to LeNet but:
  - Max pooling, ReLU nonlinearity
  - More data and bigger model (7 hidden layers, 650K units, 60M params)
  - GPU implementation (50x speedup over CPU)
    - Trained on two GPUs for a week
  - Dropout regularization

A. Krizhevsky, I. Sutskever, and G. Hinton, [ImageNet Classification with Deep Convolutional Neural Networks](), NIPS 2012

Overview

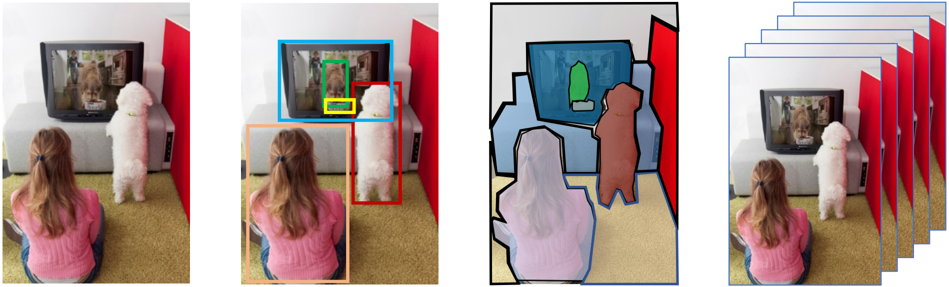
Computer Vision

- Humans use their eyes and their brains to visually sense the world.
- Computers user their cameras and computation to visually sense the world



Jian Sun, "Introduction to Computer Vision and Deep Learning".

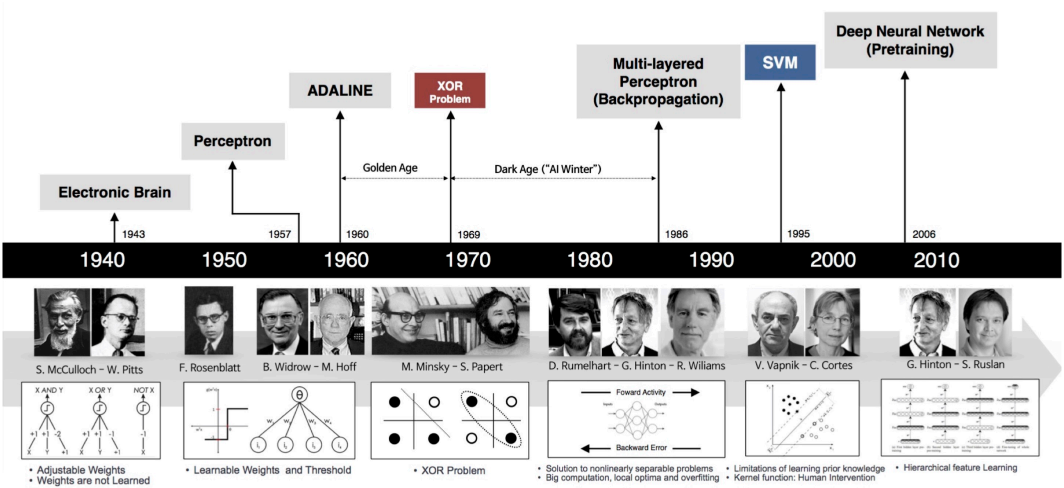
Few More Core Problems



Classification | Detection | Segmentation | Sequence

Image → Region → Pixel | Video

A Bit of History



Jian Sun, "Introduction to Computer Vision and Deep Learning".

Winter of Neural Networks (mid 90′ – 2006)

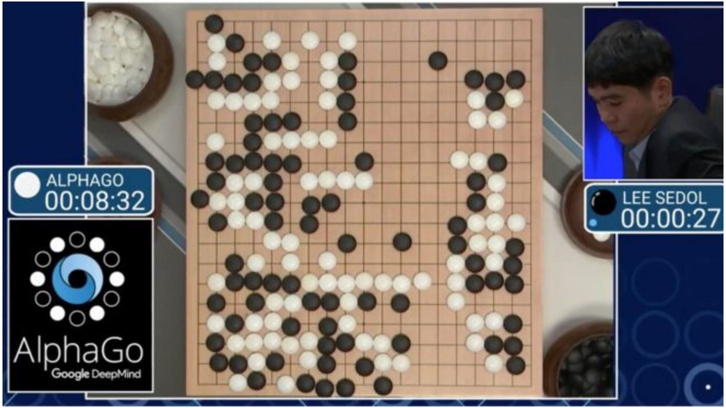
- The rises of SVM, Random forest
- No theory to play
- Lack of training data
- Benchmark is insensitive
- Difficulties in optimization
- Hard to reproduce results

Curse

"Deep neural networks are no good and could never be trained."

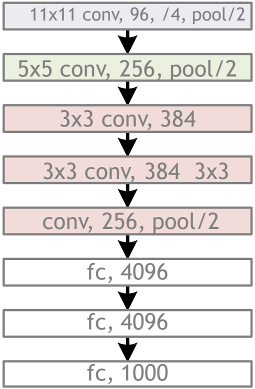
Renaissance of Deep Learning (2006 –)

- A fast learning algorithm for deep belief nets. [Hinton et.al 1996]
- Data + Computing + Industry Competition
- NVidia's GPU, Google Brain (16,000 CPUs)
- Speech: Microsoft [2010], Google [2011], IBM
- Image: AlexNet, 8 layers [Krizhevsky et.al 2012] (26.2% -> 15.3%)



Revolution of Depth

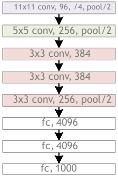
AlexNet, 8 layers (ILSVRC 2012)



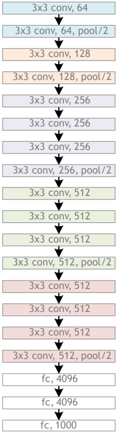
Slide Credit: He et al. (MSRA)

Revolution of Depth

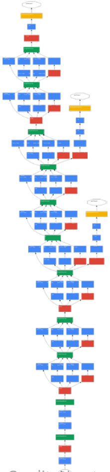
AlexNet, 8 layers (ILSVRC 2012)



VGG, 19 layers (ILSVRC 2014)



GoogleNet, 22 layers (ILSVRC 2014)



Slide Credit: He et al. (MSRA)

Revolution of Depth

AlexNet, 8 layers (ILSVRC 2012)

VGG, 19 layers (ILSVRC 2014)

ResNet, 152 layers (ILSVRC 2015)

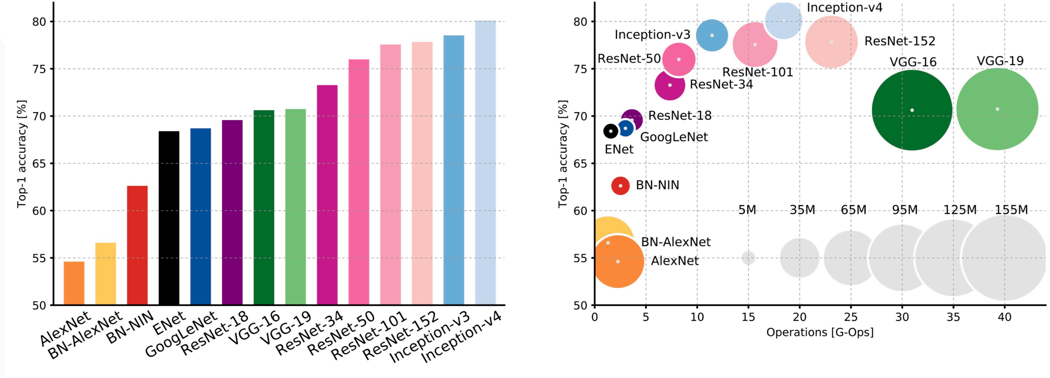
Slide Credit: He et al. (MSRA)

Some Recent Classification Architectures

- AlexNet (Krizhevsky, Sutskever, and E. Hinton 2012) 233MB
- Network in Network (Lin, Chen, and Yan 2013) 29MB
- VGG (Simonyan and Zisserman 2015) 549MB
- GoogleNet (Szegedy, Liu, et al. 2015) 51MB
- ResNet (He et al. 2016) 215MB
- Inception-ResNet (Szegedy, Vanhoucke, et al. 2016)
- DenseNet (Huang et al. 2017)
- Xception (Chollet 2017)
- MobileNetV2 (Sandler et al. 2018)
- ShuffleNet (Zhang et al. 2018)

Some Recent Classification Architectures

- AlexNet (Krizhevsky, Sutskever, and E. Hinton 2012) 233MB
- Network in Network (Lin, Chen, and Yan 2013) 29MB
- VGG (Simonyan and Zisserman 2015) 549MB
- GoogleNet (Szegedy, Liu, et al. 2015) 51MB
- ResNet (He et al. 2016) 215MB
- Inception-ResNet (Szegedy, Vanhoucke, et al. 2016) 23MB
- DenseNet (Huang et al. 2017) 80MB
- Xception (Chollet 2017) 22MB
- MobileNetV2 (Sandler et al. 2018) 14MB
- ShuffleNet (Zhang et al. 2018) 22MB



[1]

[1] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello (2017). "An analysis of deep neural network models for practical applications". In: *arXiv preprint*.

Autonomous drive

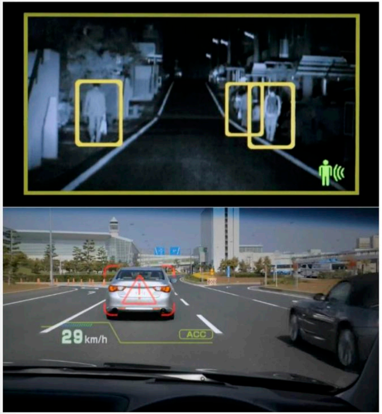
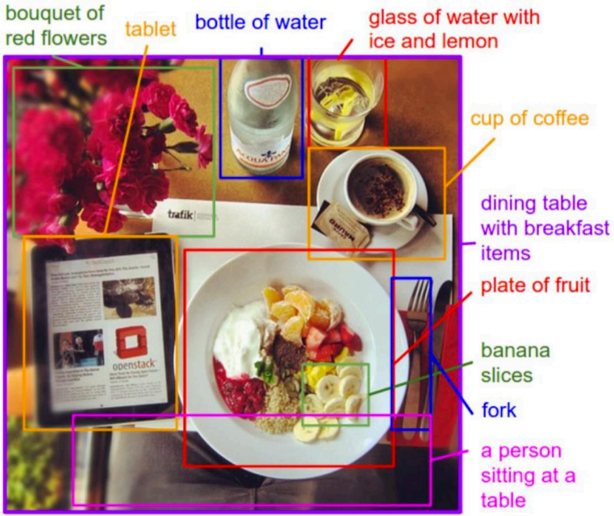
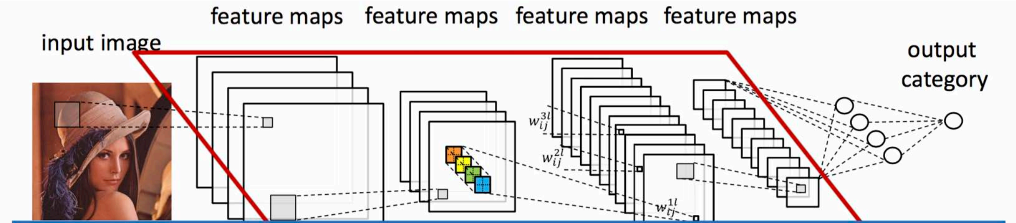


Image recognition

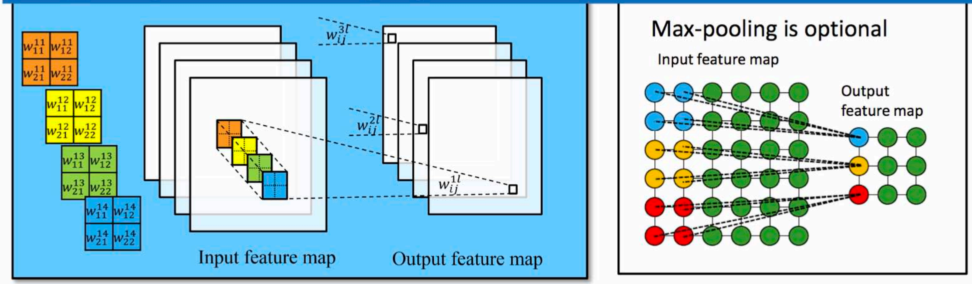


Convolutional Neural Network (CNN)



Convolutional layers account for over 90% computation

[1] A. Krizhevsky, etc. Imagenet classification with deep convolutional neural networks. NIPS 2012.
[2] J. Cong and B. Xiao. Minimizing computation in convolutional neural networks. ICANN 2014



Embedded CV

Example: Hisense ADAS

Start Video

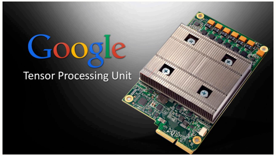
When Machine Learning Meets Hardware

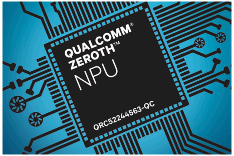
Convolution layer is one of the most expensive layers

- Computation pattern
- Emerging challenges

More and more end-point devices with limited memory

- Cameras
- Smartphone
- Autonomous driving

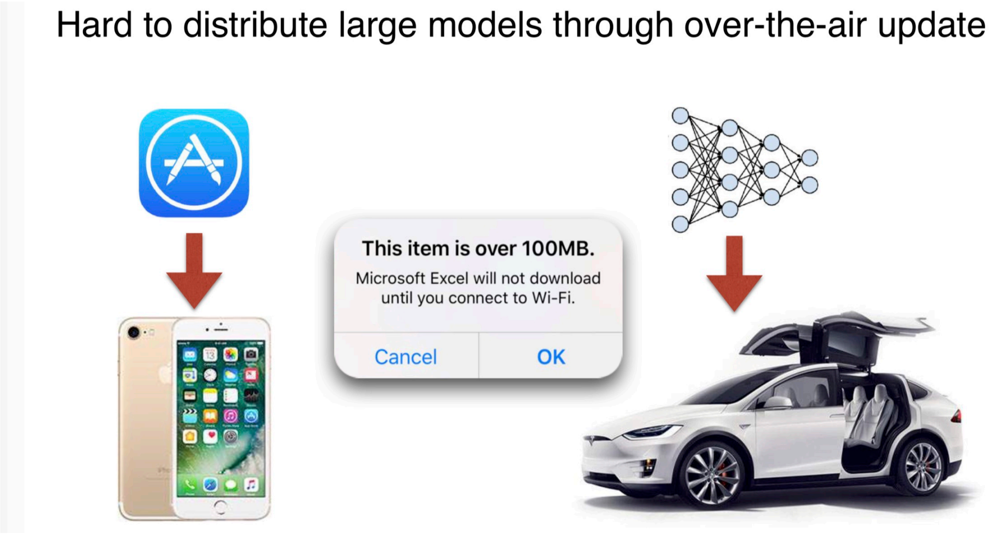




XILINX

MOBILEYE®
An Intel Company

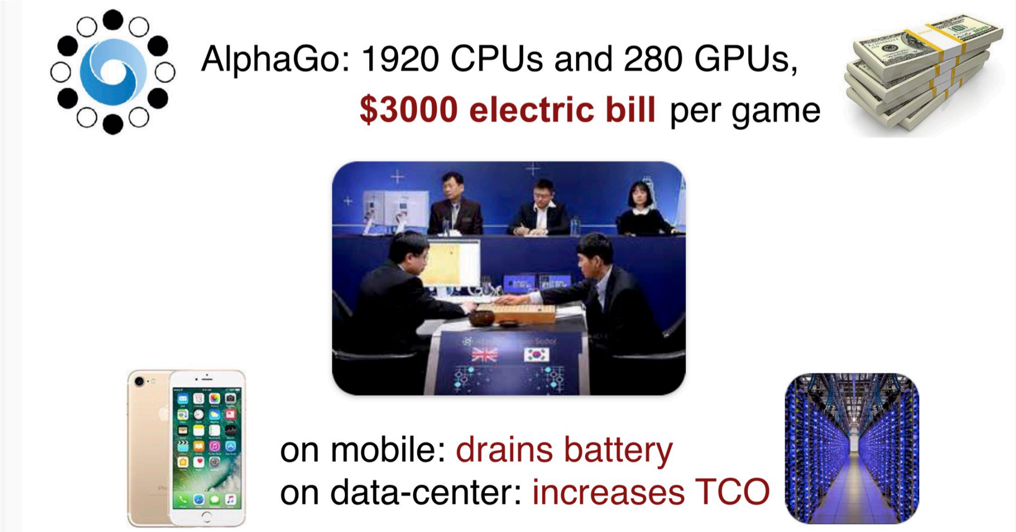
1st Challenge: Model Size



[2]

[2]Song Han and William J. Dally (2018). "Bandwidth-efficient Deep Learning". In: *Proc. DAC*, 147:1–147:6.

2nd Challenge: Energy Efficiency



[3]

[3]Song Han and William J. Dally (2018). “Bandwidth-efficient Deep Learning”. In: *Proc. DAC*, 147:1–147:6.

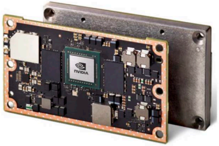
Application Category

Both	Datacenter	Edge
Intel, Nvidia, IBM, Xilinx, HiSilicon, Google, Baidu, Alibaba Group, Cambricon, DeePhi, Bitmain, Wave Computing	AMD, Microsoft, Apple, Tencent Cloud,Aliyun, Baidu Cloud, HUAWEI Cloud, Fujitsu, Nokia, Facebook, HPE, Thinkforce, Cerebras, Graphcore, Groq, SambaNova Systems, Adapteva, PEZY	Qualcomm, Samsung, STMicroelectronics, NXP, MediaTek, Rockchip, Amazon_AWS, ARM, Synopsys, Imagination, CEVA, Cadence, VeriSilicon, Videantis, Horizon Robotics, Chipintelli, Unisound, AISpeech, Rokid, KnuEdge, Tenstorrent, ThinCI, Koniku, Knowm, Mythic, Kalray, BrainChip, AImotive, DeepScale, Leepmind, Krtkl, NovuMind, REM, TERADEEP, DEEP VISION, KAIST DNPU, Kneron, Esperanto Technologies, Gyrfalcon Technology, GreenWaves Technology, Lightelligence, Lightmatter, ThinkSilicon, Innogrit, Kortiq, Hailo,Tachyum

Source: https://basicmi.github.io/Deep-Learning-Processor-List/

Flexibility vs. Efficiency

CPU
(Raspberry Pi3)



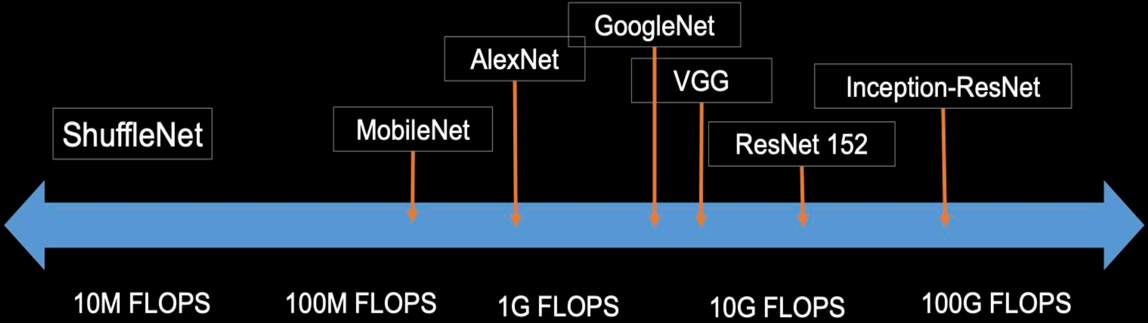
GPU
(Jetson TX2)

FPGA
(UltraZed)

ASIC
(Movidius)

Flexibility ⟷ Power/Performance Efficiency

Computing Spectrum



ASIC Example: TPU

In-Datacenter Performance Analysis of a Tensor Processing Unit™

Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon

Google, Inc., Mountain View, CA USA

Email: {jouppi, cliffy, nishantpatil, davidpatterson} @google.com

To appear at the 44th International Symposium on Computer Architecture (ISCA), Toronto, Canada, June 26, 2017.

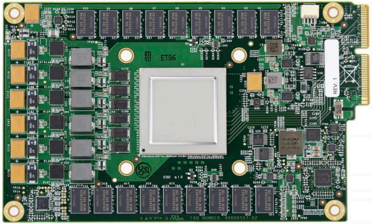


Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

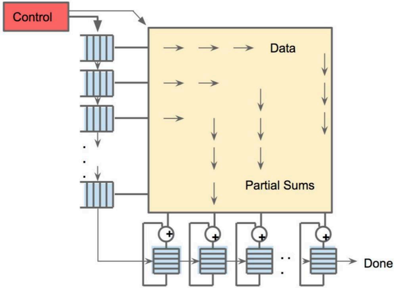
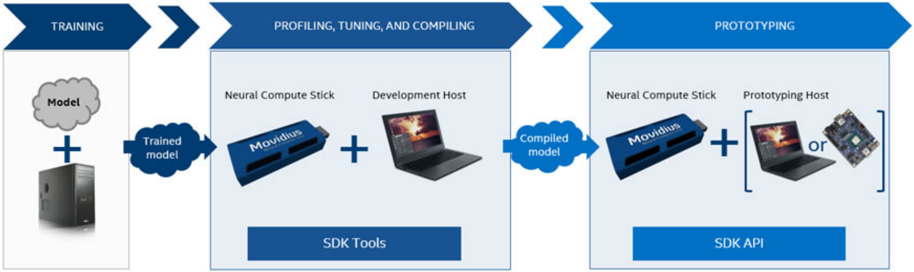


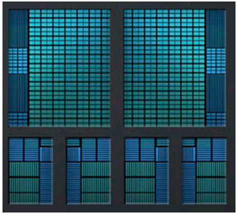
Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly

ASIC Example: Intel Movidius

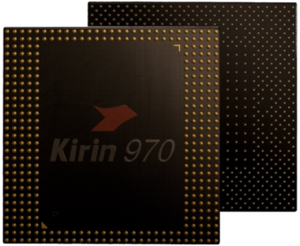


- Start Video
- Introduction Link 2

Other ASIC Examples



(a) A11



(b) Kirin970

(c) Snapdragon845

29,938 views | Aug 28, 2017, 08:01am

Microsoft: FPGA Wins Versus Google TPUs For AI

Moor Insights and Strategy Contributor

GUEST POST WRITTEN BY

Karl Freund

Karl Freund is Sr. Analyst, Machine Learning and HPC, Moor Insights & Strategy

The Microsoft Brainwave mezzanine card extends each server with an Intel Altera Stratix 10 FPGA accelerator, synthesized to act as a "Soft DNN Processing Unit," or DPU, and a fabric interconnect that enables datacenter-scale persistent neural networks. MICROSOFT

Tencent FPGA

FPGA云服务器

FPGA 云服务器（FPGA Cloud Computing）是基于FPGA（Field Programmable Gate Array）现场可编程阵列的计算服务，您只需单击几下即可在几分钟内轻松获取并部署您的FPGA计算实例。您可以在FPGA实例上编程，为您的应用程序创建自定义硬件加速。我们为您提供可重编程的环境，您可以在FPGA实例上多次编程，而无需重新设计硬件，让您能更加专注于业务发展。

立即申请

申请使用资格，将有专人为您提供服务与报价

了解更多 >>

产品介绍 产品优势 产品功能 应用场景 产品文档

Baidu FPGA


百度云
产品
解决方案
云市场
合作与生态
帮助与支持
登录
注册
备案
论坛
管理控制台
FPGA云服务器
FPGA Cloud Compute
配备现场可编程门阵列（Field Programmable Gate Array）的高性能云计算服务。同时具备开发、模拟、调试和编译硬件代码所需的各种资源，您可以基于FPGA云服务器为您的应用程序创建自定义的硬件加速能力。
申请公测
帮助文档 >
FPGA
产品概述
产品功能
产品优势
应用场景
相关产品
使用指南
产品概述

News Byte

October 12, 2017

Share this Article

Contact Intel PR

INTEL FPGAS POWER ACCELERATION-AS-A-SERVICE FOR ALIBABA CLOUD

Intel today announced that Intel® field programmable gate arrays (FPGAs) are now powering the Acceleration-as-a-Service of Alibaba Cloud*, the cloud computing arm of Alibaba Group. The acceleration service, which can be launched from the Alibaba Cloud website, enables customers to develop and deploy accelerator solutions in the cloud for Artificial Intelligence inference, video streaming analytics, database acceleration and other fields where intense computing is required.

Xilinx Selected by Alibaba Cloud for Next-Gen FPGA Cloud Acceleration

Xilinx FPGAs are accelerating machine learning and other critical compute workloads for one of the world's largest cloud providers

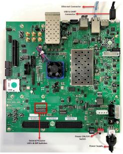
Oct 12, 2017

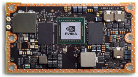
HANGZHOU, China, Oct. 12, 2017 /PRNewswire/ -- Xilinx, Inc. (NASDAQ: XLNX) today announced at the Computing Conference that Alibaba Cloud, the cloud computing arm of Alibaba Group, has chosen Xilinx for next generation FPGA acceleration in their public cloud. As the largest cloud provider in China, Alibaba Cloud offers high-performance, elastic computing power to over two million customers. Based on Xilinx® FPGAs, the new "F2" instances give Alibaba Cloud customers access to acceleration for data analytics, genomics, video processing, and machine learning workloads.

NVIDIA CEO Says "FGPA is Not the Right Answer" for Accelerating AI

https://medium.com/syncedreview/nvidia-ceo-says-fgpa-is-not-the-right-answer-for-accelerating-ai-83c810969edd

Comparisons: FPGA, ASIC, GPU[4]




	Xilinx ZCU102	Xilinx ZCU104	Huawei Atlas 200	nVIDIA Jetson TX2	Cambricon MLU 270
price	3K RMB	2K RMB	4K RMB	2.8K RMB	12K RMB
MobileNet-V1	1.14 ms	1.37 ms	1.8 ms	12.44 ms	1.85 ms
ResNet50	5.23 ms	6.81 ms	3.6 ms	24.70 ms	2.54 ms
Inception_v2	2.68 ms	3.35 ms	6.0 ms	10.81 ms	5.12 ms
Inception_v3	6.44 ms	8.53 ms	5.7 ms	32.53 ms	4.71 ms
Inception_v4	11.87 ms	17.06 ms	9.3 ms	44.37 ms	11.33 ms

[4] price is NOT accurate – reference purpose.