

Deep Learning Meets Heterogeneous Computing

Dr. Ren Wu

Distinguished Scientist, IDL, Baidu

wuren@baidu.com



Everyday

5b+ queries
500m+ users
100m+ mobile users
100m+ photos
...

As of January 2014, Baidu's market capitalization stood at roughly \$55 billion.^[8] It is considered one of the most successful internet startups of all time by market capitalization, revenue, growth and cultural impact.^[9]



WIKIPEDIA
The Free Encyclopedia

Big Data



Storage

- >2000PB

Processing

- 10-100PB/day

Webpages

- 100b-1000b

Index

- 100b-1000b

Update

- 1b-10b/day

Log

- 100TB~1PB/day

Will Baidu's Data Center Be the World's Largest?



by Mark Hachman | September 6, 2012

Chinese Web giant says that it's spending \$1.6 billion on a cloud data center, a massive investment in its infrastructure.



Infrastructure



ARM全球首个Server端规模应用
存储密度提升70%，TCO降低25%

ARM Servers

- Higher density



400倍

GPU并行架构引入语音产品线，性能大幅提升

GPU服务器

- Much better performance



Data center containers

- Faster deployment

整机柜定制

TCO	峰值交付能力
降低10%	提升10倍

国内首个成规模整机柜服务器项目

Self-design switches

- Much lower cost

Switch

DAVICU

PPC Intel

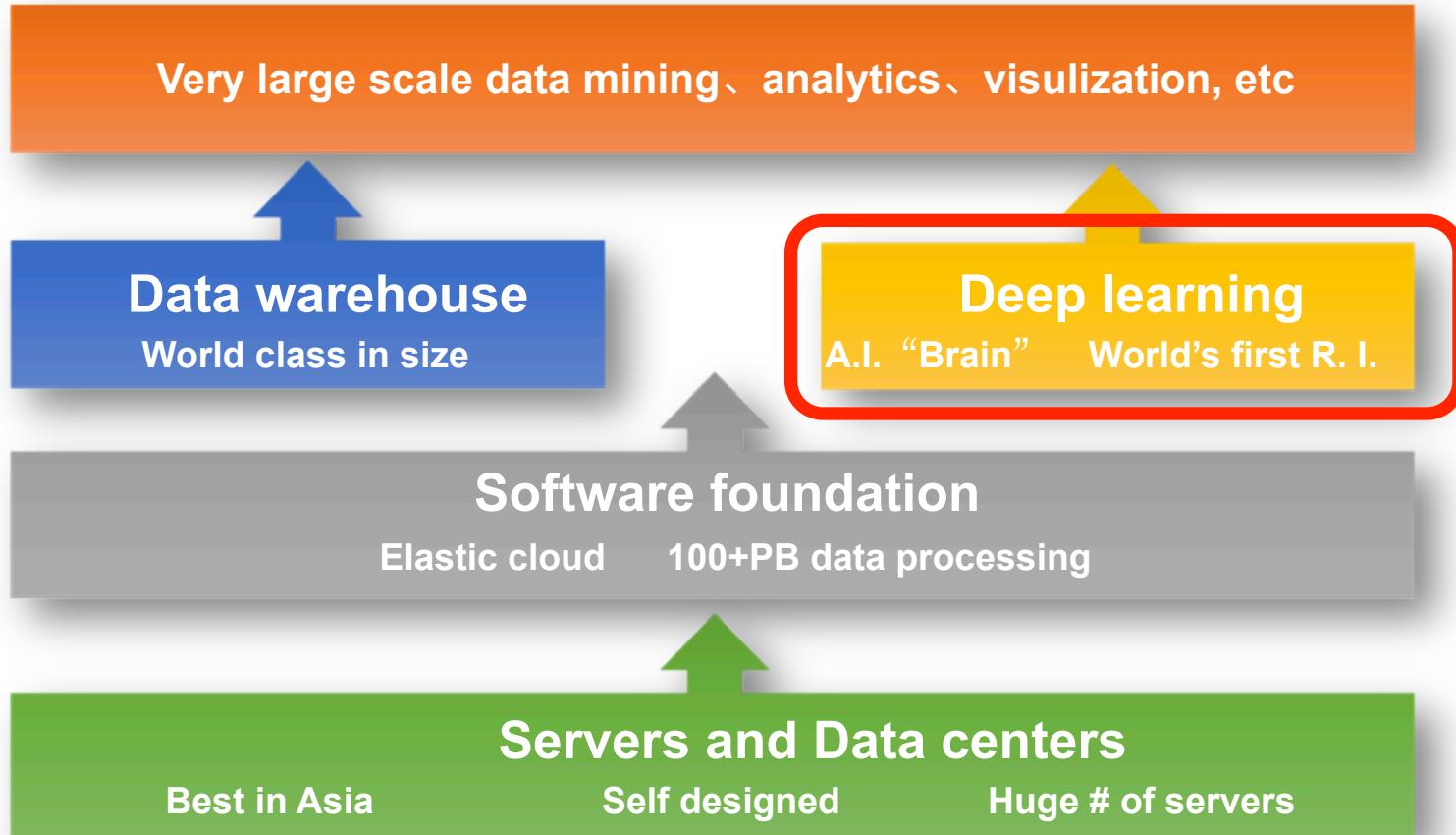
X86 Marvell

ARM Broadcom

DAC万兆部署
基于开源自研OS
HPC上线3000节点

行业首个规模部署

Big Data @Baidu



Nine Technology Challenges



On Aug 13, 2012, CEO Robin Li gave a keynote speech at ACM KDD, and proposed nine major technological challenges to the academic research community. The top three are:

1. OCR in natural images
2. Speech recognition and understanding
3. Content-based image retrieval (visual search)

materials are identical for all configurations. The blue bars in Fig. 1 summarize the measured SHG signals. For excitation of the *LC* resonance in Fig. 1A (horizontal incident polarization), we find an SHG signal that is 500 times above the noise level. As expected for SHG, this signal closely scales with the square of the incident power (Fig. 2A). The polarization of the SHG emission is nearly vertical (Fig. 2B). The small angle with respect to the vertical is due to deviations from perfect mirror symmetry of the SRRs (see electron micrographs in Fig. 1). Small detuning of the *LC* resonance toward smaller wavelength (i.e., to 1.3- μm wavelength) reduces the SHG signal strength from 100% to 20%. For excitation of the Mie resonance with vertical incident polarization in Fig. 1D, we find a small signal just above the noise level. For excitation of the Mie resonance with horizontal incident polarization in Fig. 1C, a small but significant SHG emission is found, which is again po-

Reducing the Dimensionality of Data with Neural Networks

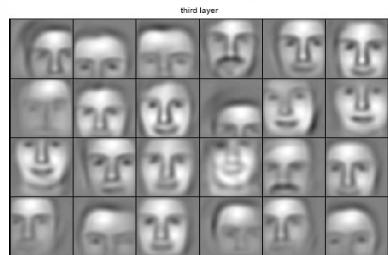
G. E. Hinton* and R. R. Salakhutdinov

High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

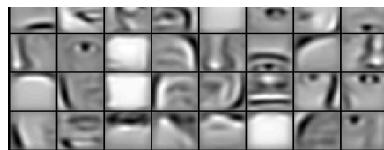
Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which

finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer "encoder" network

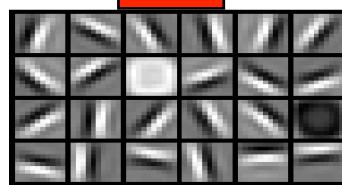
Deep Learning vs. Human Brain



object models



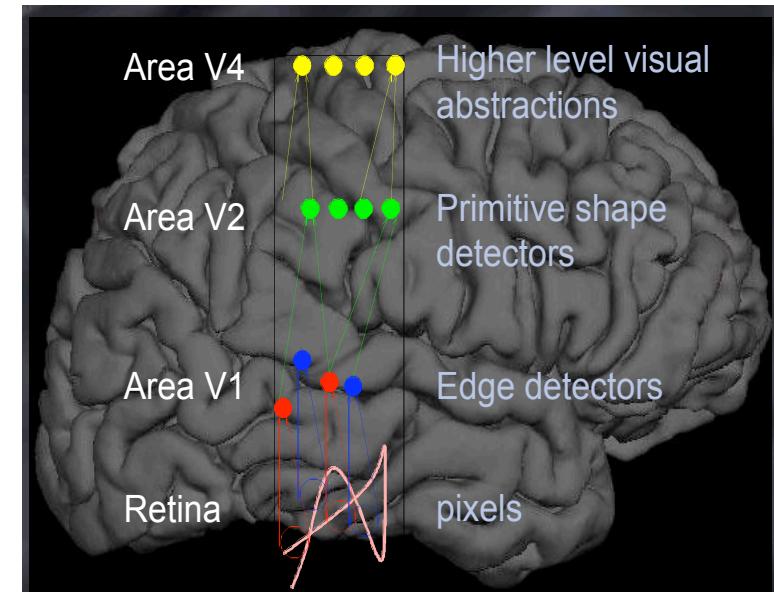
object parts
(combination
of edges)



edges



pixels



Slide credit: Andrew Ng

Top breakthrough technology 2013



HOME ▾ MENU ▾ CONNECT THE LATEST POPULAR MOST SHARED

Introduction The 10 Technologies Past Years

MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013

Deep Learning
With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

Temporary Social Media
Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous.

Prenatal DNA Sequencing
Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child?

Additive Manufacturing
Skeptical about 3-D printing? GE, the world's largest manufacturer, is on the verge of using the technology to make jet parts.

Baxter: The Blue-Collar Robot
Rodney Brooks's newest creation is easy to interact with, but the complex innovations behind the robot show just how hard it is to get along with people.

Memory Implants

Smart Watches

Ultra-Efficient Solar Power

Big Data from Cheap Phones

Supergrids

MIT Technology Review, April 23rd, 2013

- Announced its first research arm in Jan. 2013
- Institute of Deep Learning (IDL)
- The focus is Artificial Intelligence
- Two locations: Beijing and Silicon Valley



Progress of Deep Learning at Baidu



- Big improvement on speech & image recognition (2013)
 - Speech: error rate reduced by **25%**
 - OCR: error rate reduced by **30%**
 - Face: LFW benchmark, **94%** correct
- **DNN CTR** for search ads was launched on May 20th 2013, serving billions of search queries everyday – **substantial improvement**

Baidu – Visual Search



本地上传

请粘贴图片网址或拖拽图片至此 | →

相似美图

猜猜TA是谁？

高清素材

图片来源

人脸识别

识图插件下载

即使相隔汹涌的大海，
它依旧会
照亮你的方向

点击发现更多 >

Visual Search: Faces



人脸搜索 搜索结果 支持图片

A grid of 15 portrait photographs of a woman, likely Liu Yifei, used for visual search results. The images are arranged in three rows: the first row has five images, the second row has five images, and the third row has five images. The images show the woman from different angles and in different settings. A small red circular icon with the number '401' is visible in the middle of the second row. A watermark '您上传的图片' (Your uploaded image) is at the bottom left of the first image.

Visual Search Example



本地上传

请粘贴图片网址或拖拽图片至此



您的图片可能是：

邓丽君

成龙

陈晓旭

670x753

邓丽君

· 来自百度百科

邓丽君（Teresa Teng，1953年1月29日—1995年5月8日），是一位在亚洲地区和全球华人社会极具影响力的台湾歌唱家，亦是20世纪后半叶最富盛名的日本歌坛巨星之一。她的歌曲在华人社会广泛的知名度和经久不衰的传唱度为其赢得了“十亿个掌声”的美誉，被日本艺能界尊为“亚洲歌唱女王”。其生前演艺

[人物概况](#) [早年经历](#) [演艺经历](#) [个人生活](#) [主要作品](#)

相似图片

· 全部相似图片



相关网页

· 更多来源

揭秘: 邓丽君 ,梅艳芳,陈晓旭的



她的歌声甜蜜依旧 她的笑容
甜美动人 她的情路一波三折
她的柔情似水东流 ...

1/2/2013 2:28:11 PM

http://ent.rednet.cn/c/2009/04/10/1741997_2.htm

"聆听永恒的经典"纪念 邓丽君 20



旷世美女 邓丽君 的终身遗憾



邓丽君 成龙恋情昙花一现 盘点



芝兰之室(a) - 邓丽君 的绝版私



人脸搜索

· 更多人脸搜索



Visually similar images



The competition

Baidu

Another Example



Image uploaded



Baidu

Baidu 图片

粘贴图片网址 本地上传

搜索一下

提示：您也可以把图片拖到这里

全部 相册照片

The screenshot shows the Baidu Image Search interface. At the top, there's a search bar with the placeholder "粘贴图片网址 本地上传" (Paste image URL/Local upload) and a "搜索一下" (Search) button. Below the search bar is a note: "提示：您也可以把图片拖到这里" (Tip: You can also drag the image here). Underneath, there are two tabs: "全部" (All) and "相册照片" (Album photos). The main area displays a grid of search results, all showing various scenes of buildings by water under a blue sky, similar to the uploaded image.

The competition

This screenshot shows a search results page from another search engine. At the top, it has a toolbar with tabs for "网页" (Web pages), "图片" (Images), "地图" (Maps), "更多" (More), and "搜索工具" (Search tools). Below the toolbar is a filter section with options for "大小" (Size), "颜色" (Color), "类型" (Type), "时间" (Time), "相关新闻" (Related news), and "更多工具" (More tools). The main content area is a grid of image thumbnails, which are mostly architectural or landscape images, including a large dome, a bridge over water, and a city skyline, which are significantly different from the image uploaded to Baidu.

Visually Similar Images - Comparison



Image uploaded



Baidu

Baidu Images

识图

输入图片网址 | 从本地上传

识图一下

提示：也可以把图片拖到这里

找乐购西，赢礼品



The competition



CBIR – The Competition

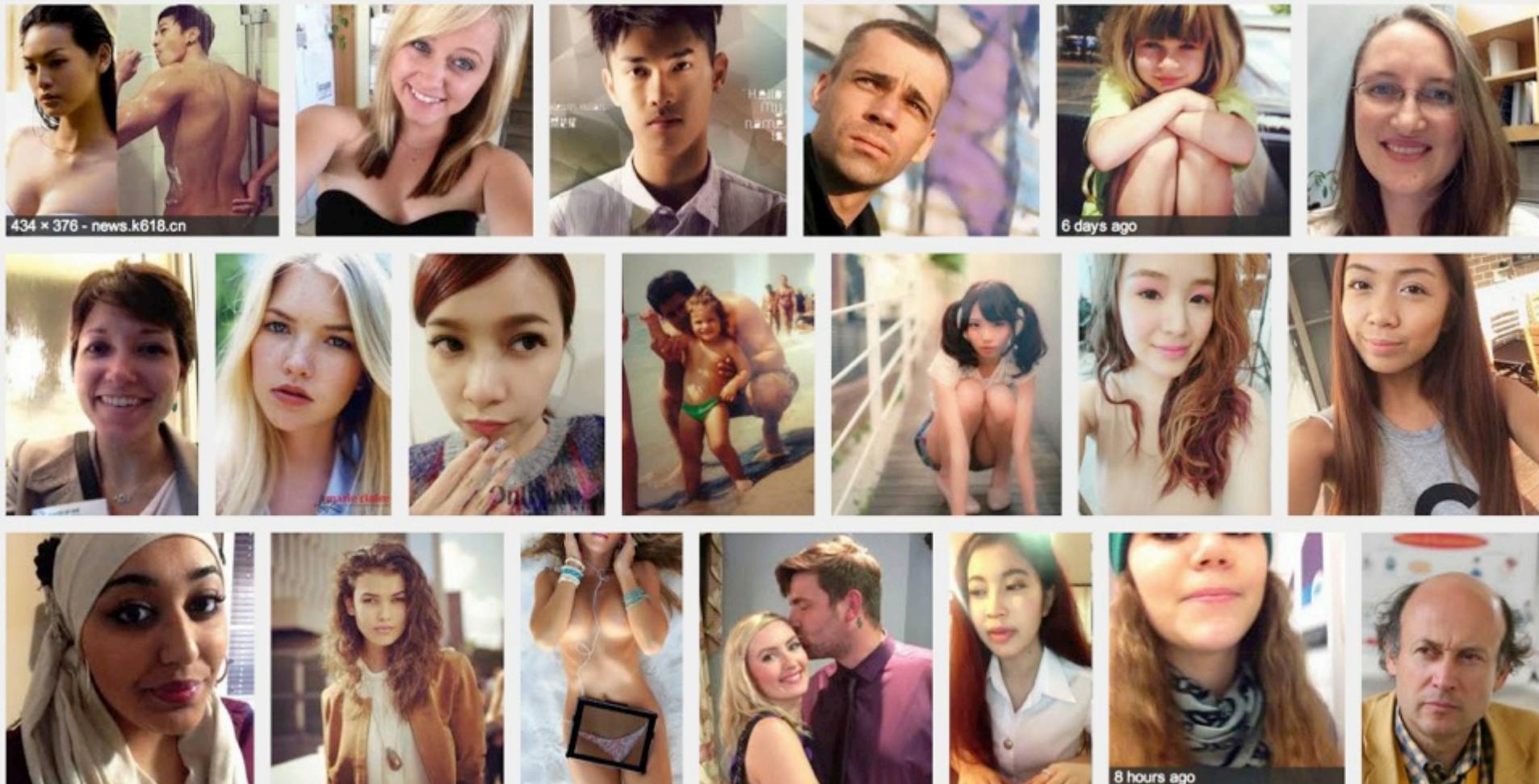


Image Recognition - Flowers



原图尺寸:750x502

我猜测这种花是:

荷包牡丹

99%的可能性

荷包牡丹原产中国、西
茎。叶对生，有长柄，
萼片两枚，早落。花瓣
荷...

[植物简介 - 科属分类 -](#)



全部 相似图片

默认排序 花类别排序

荷包牡丹



莲



百度魔图：PK大咖



百度魔图
相似度: 84.47%
偶霸，明星style!

我的照片

快来下载百度魔图
看看你最像哪位明星吧!

百度魔图
相似度: 74.41%
布死痕象啊~绳命作弄人~(>_<)~

我的照片

丹尼尔·克雷格

快来下载百度魔图
看看你最像哪位明星吧!

无 SIM 卡 上午 12:00 68%

排行榜 分类 免费排行

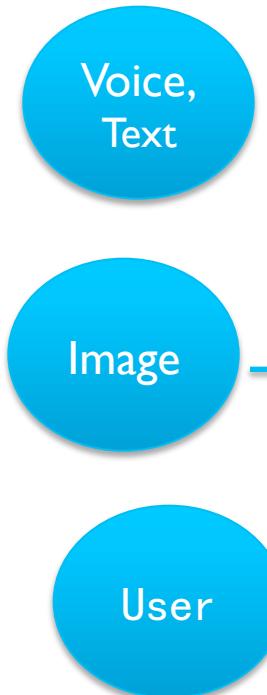
名次	应用图标	应用名称	类别	评分	操作
1		百度魔图	摄影与录像	★★★★★ (52)	打开
2		百度魔拍	摄影与录像	★★★★★ (12)	打开
3		君王2 HD	游戏	★★★★★ (1.2M)	免费
4		我叫MT Online	游戏	★★★★★ (1.1M)	免费

精品推荐 排行榜 Genius 搜索 更新

Peak uploadin

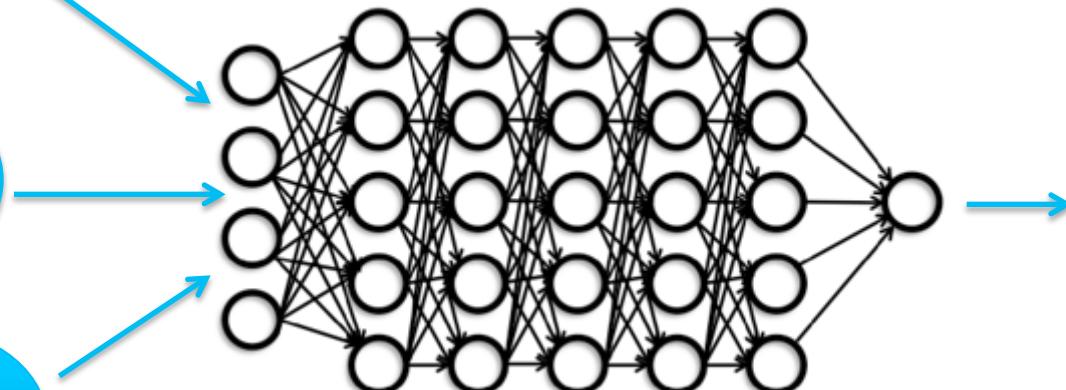
DOS APP #1 for 3 weeks

Deep Learning



DNN for Speech

10k hours of voice data
10b training samples
Months on a GPU cluster



Typical scale of training data



Datasets

- Image recognition: 100 millions
- OCR: 100 millions
- Speech: 10 billions
- CTR: 100 billions

Training time:

Weeks to Months
on GPU clusters

**Big data + Deep learning + HPC
= Success**

Projected training data to
grow 10x each year



Mobile Applications of DNN



“手机百度 随时知道”



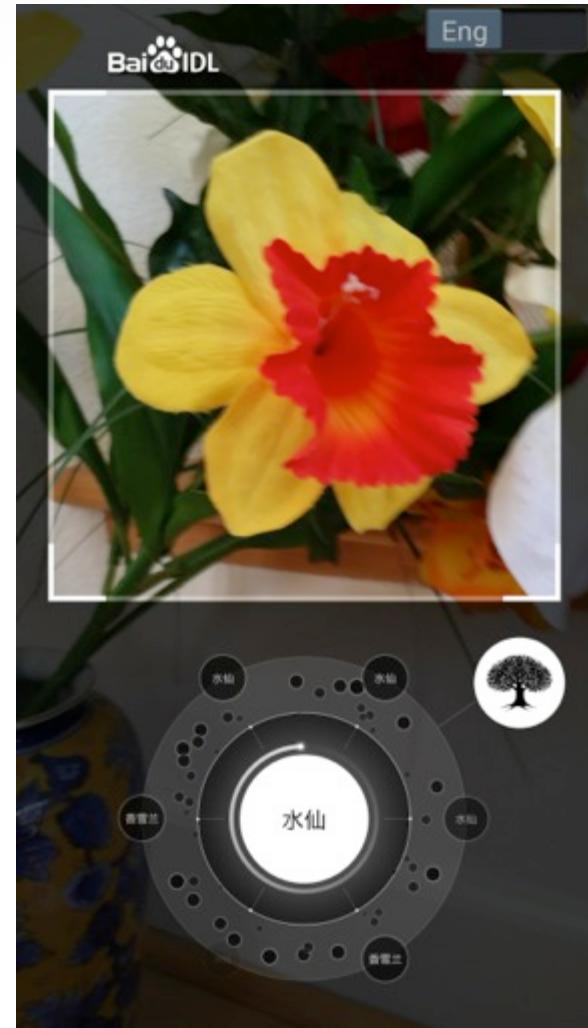
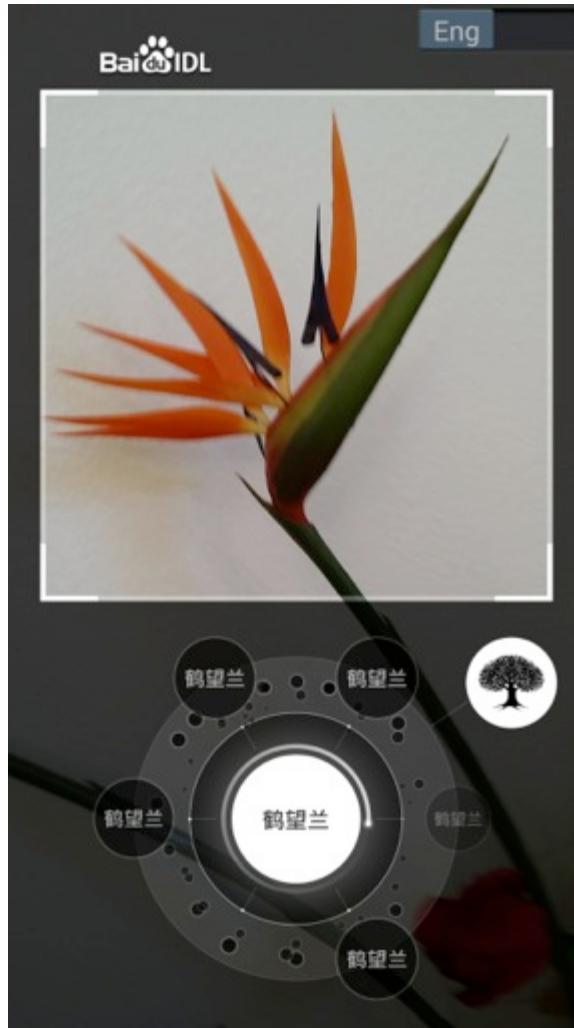
想象一下图像识别加上语音识别

DNN – Anywhere, Anytime



- DNN-based image recognition on mobile device
- No connectivity needed
- Real time, directly works on video stream
- Everything is done within the device
- What you point is what you get
- OpenCL based, highly optimized
- Large deep neural network models
- Thousands of objects, flowers, dogs, and bags etc
- Unleashed the full potential of the device hardware
- World's first in-place mobile DNN app?
- And the best!

DNN – Anywhere, Anytime



百度酷耳



穿衣指数

DNNs Everywhere



Supercomputers



Datacenters
(cloud)

Tablets, smartphones



Wearable devices
IoTs



DNNs Everywhere



Supercomputers



Datacenters



1000s GPUs

100k-1m servers

Tablets, smartphones



Wearable devices
IoTs



Supercomputer used for training

Trained DNNs then deployed to data centers (cloud),
smartphones, and even wearables and IoTs

Heterogeneous Computing



Fastest Supercomputer In Europe
6.27 PetaFLOPS (80% Linpack Efficiency)
Piz Daint

Greenest Petascale
3110 MFLOPS/
#2: JUQUEEN: 2176 MFLOPS

4th Generation Intel® Core™ Processor Die Map
22nm Tri-Gate 3-D Transistors

Processor Graphics
Quad core die shown above
** Cache is shared across all 4 cores and processor graph

Krait 400 CPU features 28HPm process technology superior 2GHz+ performance

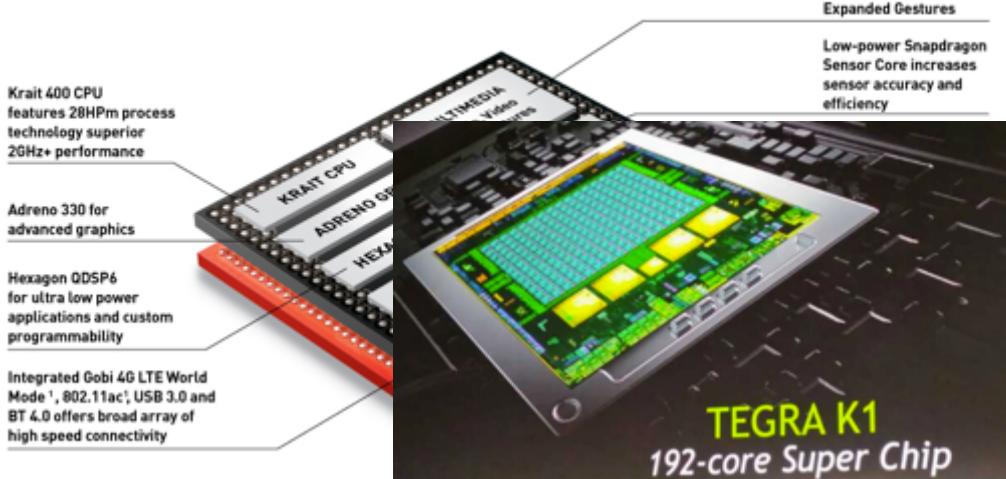
Adreno 330 for advanced graphics

Hexagon QDSP6 for ultra low power applications and custom programmability

Integrated Gobi 4G LTE World Mode, 802.11ac, USB 3.0 and BT 4.0 offers broad array of high speed connectivity

Supercomputers
Data centers (cloud)
Smart phones
Wearable devices!

Big data +
Deep learning +
HPC HC = Success



OpenCL-based Open ECO-SYSTEM



- Diverse industry participation, from cell phones to supercomputers
 - Processor vendors, system OEMs, middleware vendors, application developers.
- OpenCL is the industry standard embraced by many companies.

3DLABS
SEMICONDUCTOR

ACTIVISION | **BLIZZARD**

AMD

ARM

BROADCOM

EA

codeplay™

ERICSSON

freescale™
semiconductor

GE

HI CORP.

IBM.

intel

**Imagination
TECHNOLOGIES**



NATIONAL LABORATORY
Los Alamos

MOTOROLA

movidia

NOKIA

NVIDIA.

QNX
QNX SOFTWARE SYSTEMS

RAPID MIND

SAMSUNG

**Seaweed
SYSTEMS**

TAKUMI

**TEXAS
INSTRUMENTS**

UNIVERSITY OF
DEMA

KHRONOS
GROUP

Summary



**Big data + Deep learning + High performance computing =
Intelligence**



**Big data + Deep learning + Heterogeneous computing =
Success**

Baidu USA



And we are hiring

- Heterogeneous Computing experts
- Parallel algorithm and performance experts
- CUDA/OpenCL Experts
- FPGA experts
- Andriod/IOS experts
- Data scientist
- Infrastructure Engineer

...

usdc-jobs@baidu.com
wuren@baidu.com

<http://usa.baidu.com/>

Thank you!

Dr. Ren Wu
wuren@baidu.com
@韧在百度