

人工智能芯片产业发展之路：新阶段应对新思维

姚颂

创始人 & CEO
DeePhi Tech 深鉴科技
song.yao@deephi.tech

www.deephi.com

© 2018 DeePhi Tech. All Rights Reserved.

Part 01

关于深鉴科技

Discover the Philosophy behind Deep Learning Computing

以场景切入
深鉴核心平台



注：DNNDK 即 Deep Neural Network Development Kit，深度学习SDK，DPU 即 Deep-learning Processing Unit，深度学习专用处理器

汪玉 联合创始人

清华大学电子系长聘副教授，清华大学电子系党委副书记
国家自然科学基金“优青”获得者，ACM FPGA技术委员会亚太唯一成员
IEEE/ACM Senior Member

单羿 合伙人&CTO

清华大学电子系博士，伦敦帝国理工联合培养
IBM PhD Fellowship，前百度IDL异构计算方向创始成员
前地平线机器人FPGA技术负责人



韩松 联合创始人

MIT EECS助理教授，斯坦福大学博士
Deep Compression技术提出者，ICLR 2016 最佳论文
FPGA 2017最佳论文

姚颂 联合创始人&CEO

清华大学电子系毕业
斯坦福大学电子系访问学者
福布斯亚洲30 Under 30，MIT TR35 China



NIPS 2015: 神经计算领域世界最顶级学术会议
FPGA 2016 & 2017: FPGA领域世界最顶级学术会议
ICLR 2016: 深度学习领域世界最顶级学术会议
ISCA 2016: 计算机体系结构领域世界最顶级学术会议
Hot Chips 2016: 半导体工业界最顶级会议
中国计算机学会科学技术一等奖

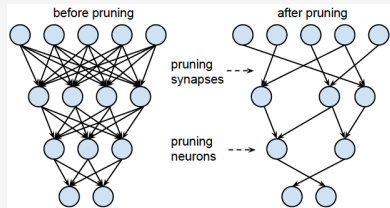
系列原创成果在国际顶尖期刊和会议上集中发表
知识产权中美同时布局，注册数百个发明专利

全球范围内率先提出神经网络压缩、稀疏神经网络概念与方法
“Learning both Weights and Connections for Efficient Neural Networks”, NIPS 2015
“Deep Compression”, ICLR 2016 Best Paper

全球范围内率先研究与发表稀疏神经网络处理架构
“EIE: Efficient Inference Engine on Compressed Deep Neural Network”, ISCA 2016

全球范围内率先将稀疏神经网络处理器用于实用
Collaboration with 搜狗, partly revealed in:
“ESE: Efficient Speech Recognition Engine with Compressed LSTM on FPGA”, **FPGA 2017 Best Paper**

深度压缩：让算法更小、更快、更省功耗



算法特点

压缩效率

精确程度

方便易用

1/3

权重数量

1/10

带宽负载

1/10

模型大小

3 X

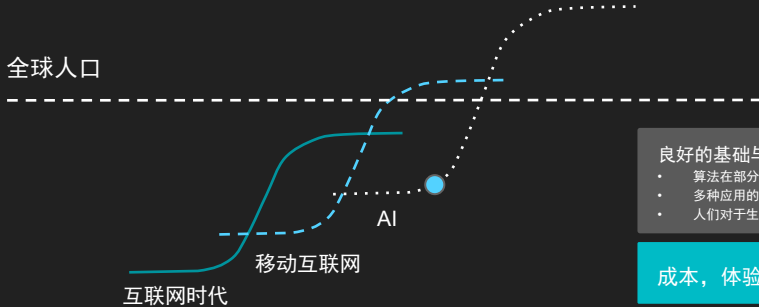
性能

深度压缩工具：可在CNN和RNN上获得**显著的压缩比**

在SSD物体检测框架下，可对算法进行**7倍压缩而不影响精确度**

深鉴DNNDK软件开发简单，仅需**50行代码**即可运行ResNet-50网络

全球人口



良好的基础与条件

- 算法在部分领域已经超过人的识别精度
- 多种应用的算法类似，具备大规模可复制
- 人们对于生活品质的追求正在上升

成本，体验，应用门槛，稳定性？

2012年，于清华大学与斯坦福大学开始学术研究

2015年中，取得阶段性研究成果，开始学术发表

2016年3月，深鉴科技公司于北京正式成立

2016年4月，获得金沙江创投与高榕资本天使轮融资

2016年6月，团队与DeepMind共同获得ICLR'16最佳论文奖项

2016年8月，与Intel、Nvidia、ARM共同站上Hot Chips

2017年1月，签订Xilinx、联发科等领投的重量级A轮融资

2017年2月，在FPGA2017会议独揽最佳论文奖项

2017年8月，获得蚂蚁金服、三星风投等领投的重量级A+轮融资

2017年10月，公司第一次发布会，发布6款基于DPU平台的产品

2018年3月，语音识别加速系统上线AWS，开启全球用户云服务模式



人脸检测识别模组

终端解决方案

云端解决方案



人脸分析解决方案



视频结构化解决方案



语音识别加速方案

深度学习开发SDK - DNNDK
(Deep Neural Network Development Kit)

DECENT 神经网络压缩工具

DNNC 神经网络编译器

DPU架构、指令集及芯片
(Deep-learning Processing Unit)



Aristotle
架构



Descartes
架构

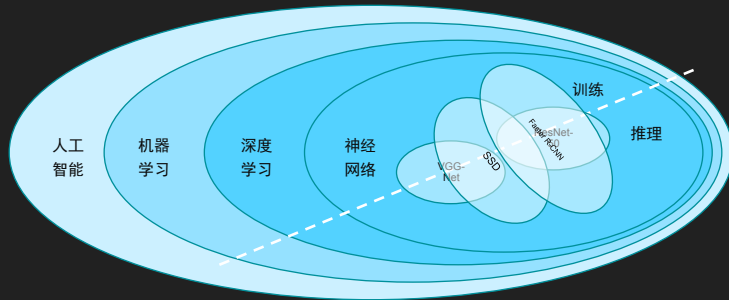


听涛/观海
SoC芯片

Part 02

技术新思维：软硬件协同

Discover the Philosophy behind Deep Learning Computing



今天，我们谈到AI时，更多地指的是机器学习和深度学习

通用芯片

机器学习

机器学习
(OpenCV + DL)

深度学习
(推理阶段)

典型
神经网络

人工智能 > 机器学习 > 深度学习 > 神经网络

定制化应用

传统芯片设计：接口/协议明确，只优化某一方面

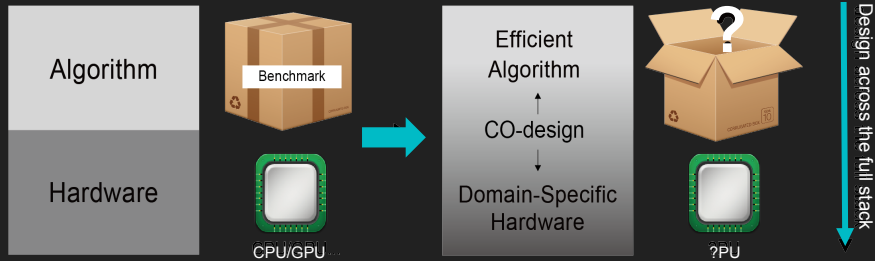


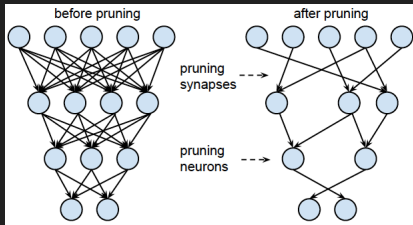
传统芯片设计：接口/协议明确，只优化某一方面



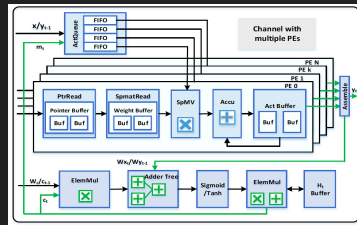
<i>Application</i>	<i>MLP0</i>	<i>MLP1</i>	<i>LSTM0</i>	<i>LSTM1</i>	<i>CNN0</i>	<i>CNN1</i>	<i>Mean</i>	<i>Row</i>
Array active cycles	12.7%	10.6%	8.2%	10.5%	78.2%	46.2%	28%	1
Useful MACs in 64K matrix (% peak)	12.5%	9.4%	8.2%	6.3%	78.2%	22.5%	23%	2
Unused MACs	0.3%	1.2%	0.0%	4.2%	0.0%	23.7%	5%	3
Weight stall cycles	53.9%	44.2%	58.1%	62.1%	0.0%	28.1%	43%	4
Weight shift cycles	15.9%	13.4%	15.8%	17.1%	0.0%	7.0%	12%	5
Non-matrix cycles	17.5%	31.9%	17.9%	10.3%	21.8%	18.7%	20%	6
RAW stalls	3.3%	8.4%	14.6%	10.6%	3.5%	22.8%	11%	7
Input data stalls	6.1%	8.8%	5.1%	2.4%	3.4%	0.6%	4%	8
TeraOps/sec (92 Peak)	12.3	9.7	3.7	2.8	86.0	14.1	21.4	9

Source: Google TPU paper, ISCA 2017





算法：深度压缩



硬件：为压缩后算法优化

Part 03

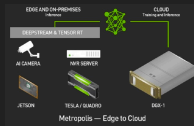
产品新思维：系统解决方案

Discover the Philosophy behind Deep Learning Computing





英伟达：最核心竞争力来自于丰富的软件、算法堆栈



DeepStream in
Security Industry

DRIVEWORKS SDK	DETECTION	LOCALIZATION	DRIVING	VISUALIZATION
	Detection/Classification	Map Localization	Vehicle Control	Streaming to cluster
	Sensor Fusion	HD-Map Interfacing	Scene understanding	ADAS rendering
	Segmentation	Egomotion (SFM, Visual Odometry)	Path Planning solvers	Debug Rendering

DriveWorks in Automotive Industry



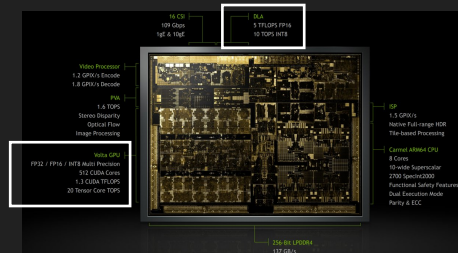
CUDA

英伟达已转型为系统与平台厂商，而不是GPU芯片厂商



英伟达商务副总裁Jeff Herbst

1.3 CUDA TFLOPS
20 Tensor Core TOPS
5 DLA TFLOPS



英伟达在CES2018发布的 Drive Xavier SoC

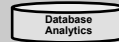
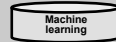
应用需求定义系统架构，系统架构定义芯片架构

Xilinx正着力建立完整的软硬件堆栈

Accelerated
Open Frameworks
(Xilinx and partners)



Accelerated
Libraries
(Xilinx and partners)



Development
Environment
(Xilinx)



Boards,
runtime software



Development Stack

Platform

Xilinx新任CEO Victor Peng, 此前负责Xilinx软件与解决方案



Victor led the FPGA development software and IP solutions team previously

Part 04

商业新思维：把握应用，换道超车

Discover the Philosophy behind Deep Learning Computing



定损



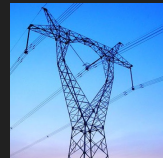
无人机



辅助驾驶



遥感



电网



医疗



官方微信



www.deephi.com

THANK YOU!

2018年5月25日星期五

Discover the Philosophy behind Deep Learning Computing