人工智能手机

石恩名

广州优亿信息科技有限公司

01 AI智能手机

人工智能手机是一种由AI技术强化加持的智能通讯设备，从**硬件芯片层**到**操作系统层**和**应用交互层**均整合了人工智能技术的移动通讯设备，从而帮助用户可以更轻松便捷地下达指令，完成用户要做的事，降低用户的认知负担。

AI智能手机

- 人工智能和5G技术的成熟重新定义智能手机，新的智能手机将拥有语音界面，它既可以与触摸屏一起运行，也可以单独运行，手机还将搭载与当前不同的应用程序。
- 人工智能手机是一种由AI技术强化加持的智能通讯设备，从**硬件芯片层**到**操作系统层**和**应用交互层**均整合了人工智能技术的移动通讯设备，从而帮助用户可以更轻松便捷地下达指令，完成用户要做的事，降低用户的**认知负担**。

背景

人工智能手机具备能力：

- 知觉: 视觉， 听觉， 触觉， 嗅觉， 味觉， (以及其他传感器带来的超越人类的感觉， 例如加速度， 磁场， 气压， 脑电波等等)
- 学习: 通过经验不断获得正确知识的能力.
- 知识表示: 如何将万物及其属性之间建立各种联系.
- 演绎、推理和解决问题
- 自然语言处理: 能听懂人的语言， 能说人的语言.
- 规划: 根据现有信息， 制订最有利的行动事项表.
- 运动和控制: 能移动并控制一些工具达成行动.
- 社交: 能感知他人的情绪， 并做出 不导致负面情绪 的反应

AI手机

降低用户的认知负担	通过机器深度学习，实现智能感知。它不仅能够智能学习用户的应用使用习惯，让手机越用越懂用户，还能准确预测用户行为，提前保障，降低应用冷启动的概率，能够实现85%的用户行为预测，使应用启动时间提升20%以上。
资源分配	按需动态调度CPU、GPU、内存等手机资源，彻底改善“公平资源调度”策略，而不是放任APP之间互相争夺，真正实现基于应用的优先级和体验需求来分配硬件资源，使手机持久使用后依然能够保持流畅。
性能优化	基于人工智能的自学习系统，软硬结合的精细化资源调度和安卓系统组件深入优化，解决了Android手机久用卡顿问题。在此基础上，EMUI在手机内存管控上继续发挥技术优势，通过更高效稳定的内存回收，更高压缩比的后台内存压缩，更智能的内存碎片整理策略，让手机在相同RAM的硬件基础上用户体验更流畅。
	基于手机的人工智能学习技术，实现基于用户使用习惯的智能排序算法，优先压缩后台不常用的应用和文件，提升压缩比例
提高用户体验	能够预测触摸滑动的轨迹，使手机的响应速度提升10%。它使用智能触控位置预测算法，预测用户的滑动轨迹，提前进行界面显示刷新，触控操作响应速度提升10%。

AI手机

手机更懂你

- 手机将变成用户的“人格扩展”，未来我们的手机将会识别并预测用户的行为。他们会清楚用户的身份，用户想要做什么，用户什么时候想要做以及用户想如何做这件事。智能手机会一直追踪你来进行学习，计划并解决你的问题。它会调用自己所有的传感器和数据来完成这个任务。

用户验证

- 传统的密码验证系统已经渐渐变得复杂低效，带来的结果是安全漏洞和糟糕的用户体验。机器学习、生物识别以及用户行为将会促进安全技术的可用性和自助服务的能力。

AI手机

情绪识别

- 情绪感应系统和情感计算能够让手机探测、分析、处理并对用户的情感和情绪做出相应的反应。激增的智能助理和AI激发了基础的对话系统对“情感智能”的需求，以便能够为用户提供更好的语境和使用体验。

物体识别

- 智能手机能够帮助用于对事物进行认知，只需要用手机对事物进行拍照，及可以获取事物的基本信息。

音频分析

- AI能够分辨出这些声音是什么，通知用户或者触发警报。例如，手机能够检测到用户在打鼾，进而触发手环震动，促使用户调整睡眠姿势。

AI手机

AI智能手机主要核心架构可以分为四层：

（1）硬件层。由智能芯片为核心组成的一套具备多元感知和强大的处理能力的硬件配套组合。

（2）深度学习模型库。各种深度学习算法模型库，为上层应用实现提供技术保障。

（3）SDK软件开发工具包，提供了各种人工智能所需要的软件工具开发包，包括语音识别、图像识别、人脸识别和自然语言处理等。

（4）应用层。基于SDK和深度模型库实现的各种交互应用。

02 智能芯片介绍

人工智能手机是一种由AI技术强化加持的智能通讯设备，从**硬件芯片层**到**操作系统层**和**应用交互层**均整合了人工智能技术的移动通讯设备，从而帮助用户可以更轻松便捷地下达指令，完成用户要做的事，降低用户的认知负担。

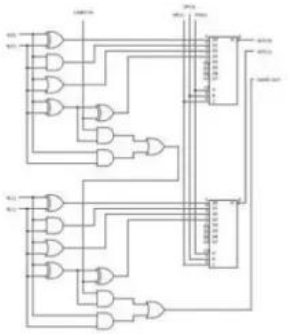
硬件层

人工智能手机在硬件层首先具有一颗可以足够强劲的AI芯片，同时为了支持视觉，听觉， 触觉等方面的能力需要其具备各类传感器，如温度传感器、压力传感器、重力传感器、加速度传感器、磁场传感器、陀螺仪等。

在芯片上，拥有专用的智能芯片比通用硬件如CPU、GPU要更具备高性能、低功耗的特点。

专用硬件为什快？

- 没有取指，译码等过程
- 高度并行的计算单元
- 流水线，硬件利用率高
- 片内带宽很高，没有传输瓶颈

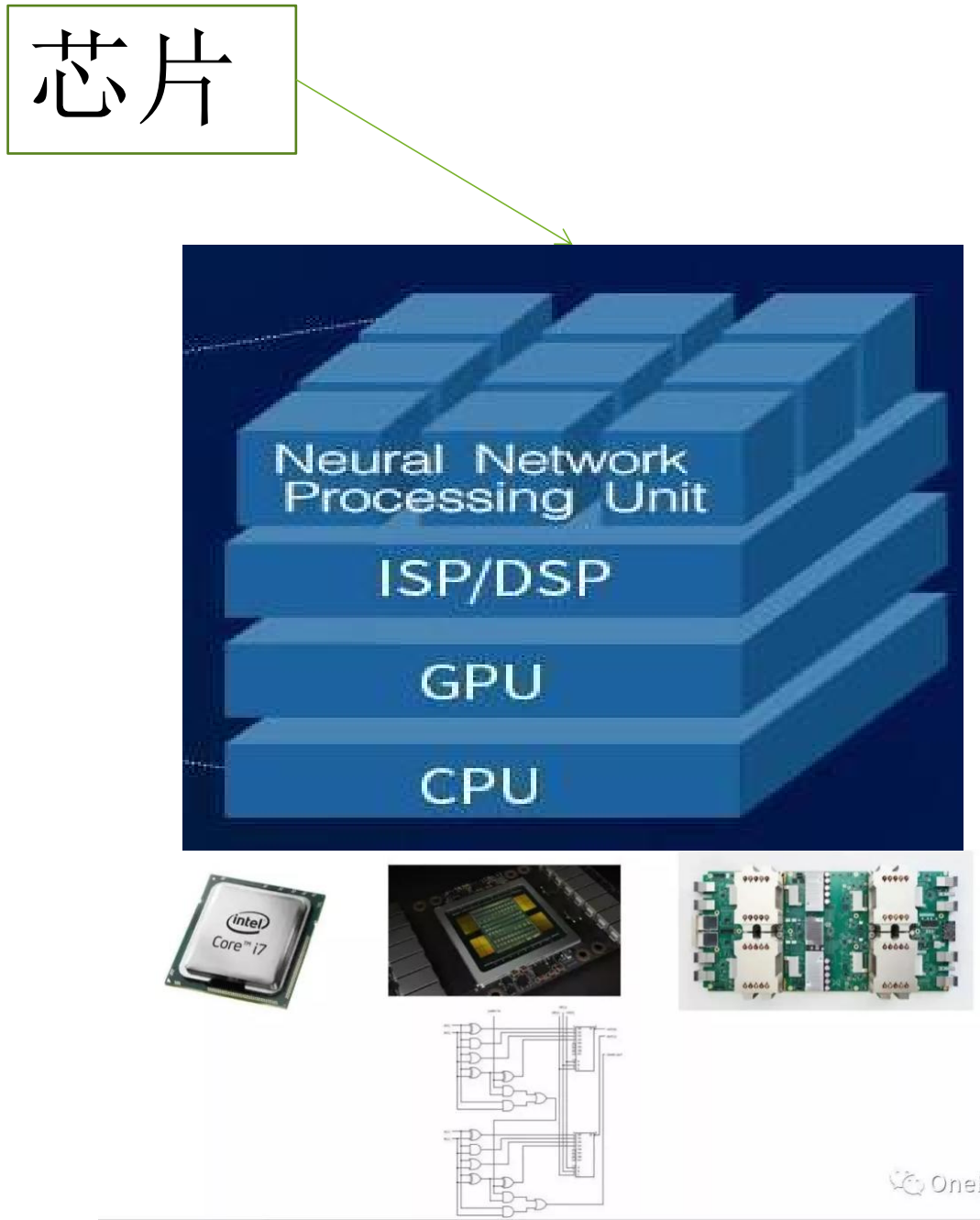


人工智能芯片

通用芯片并不能很好地适应深度学习算法的要求，效率低，功耗大，成本高。各种神经网络算法需要专用芯片来保证其运行效率。

云端要求 AI 芯片适应多种神经网络架构，同时能进行高精度浮点运算，峰值性能至少要达到Tflops级别，对功耗没有严苛要求；支持阵列式结构以进一步提高性能。

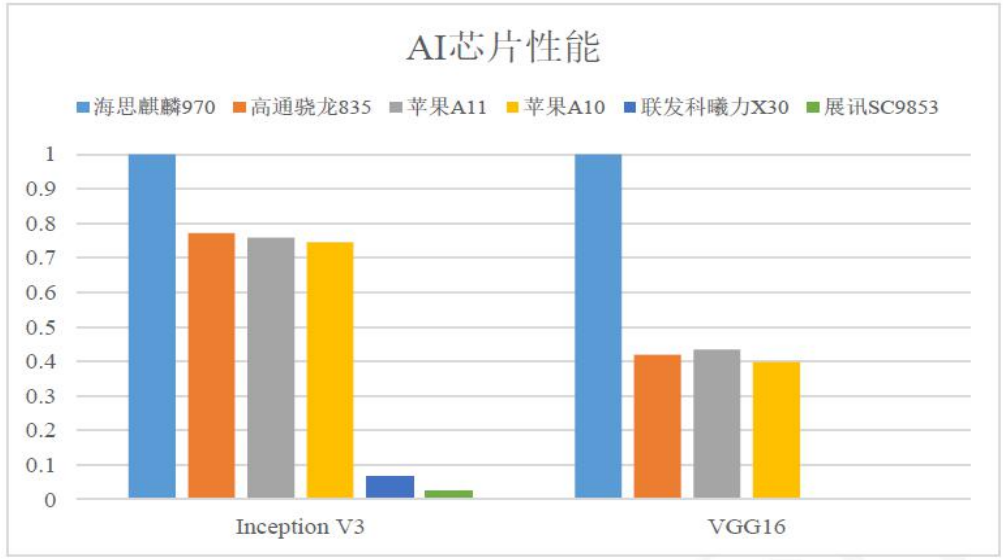
移动端 AI 芯片对设计的要求截然不同。一个根本的要求是控制功耗，这就需要使用一些办法（如网络压缩）来提升计算能效，同时尽可能少地降低计算性能和计算精度的损失。

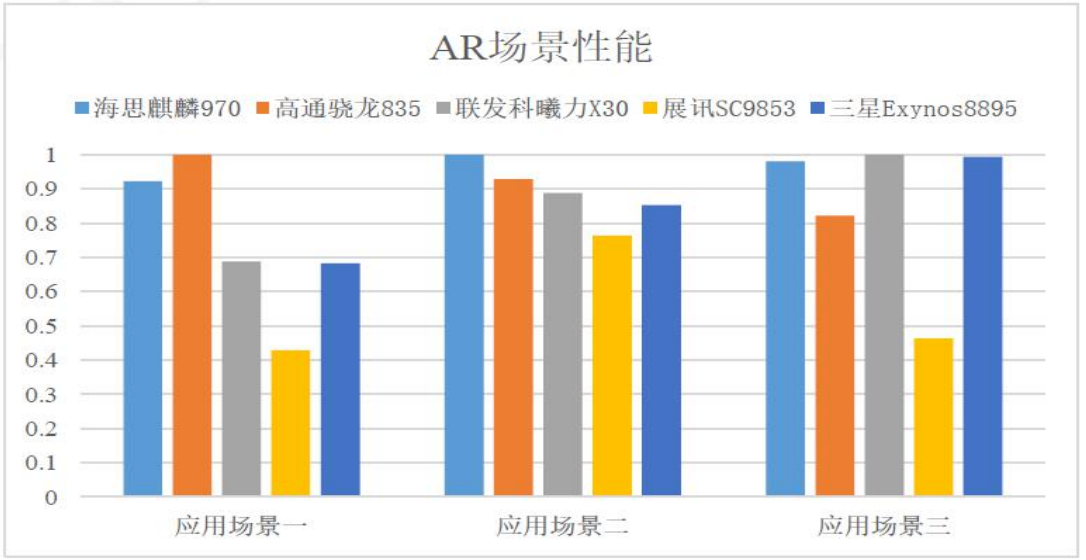


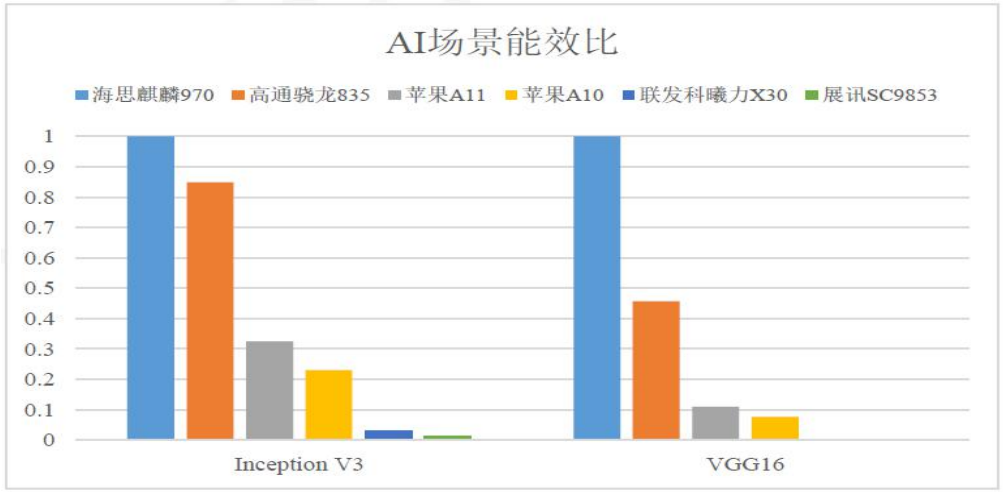
移动端芯片

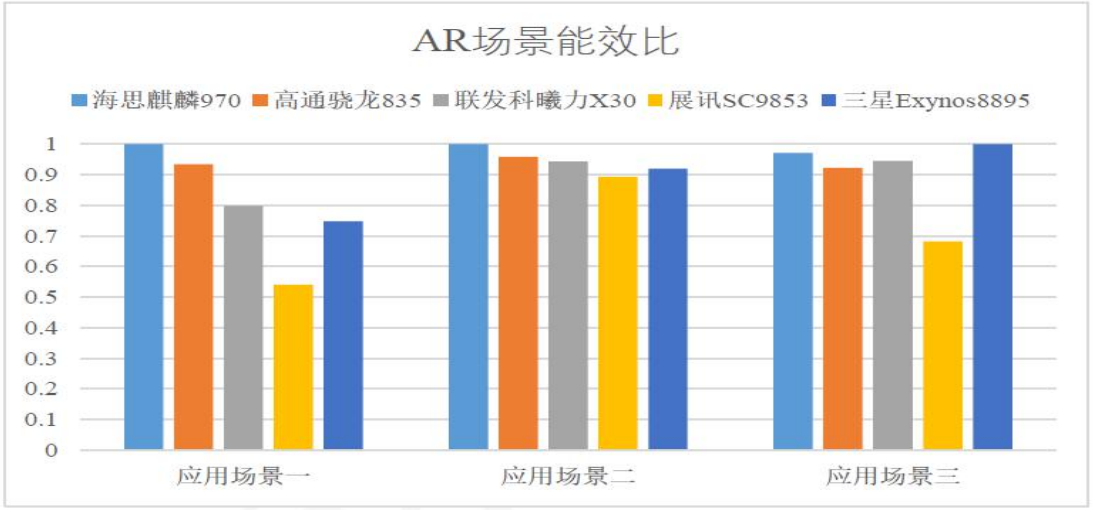
公司	芯片	CPU	GPU	NPU	浮点计算能力(16位)	最高主频	工艺
华为	麟970芯片	4xA73(2.4Ghz)+4xA53(1.8Ghz)	Mali-G72 MP12(12核心)	寒武纪A1	346.8 Gflops	2.4GHz	TSMC 10nm工艺
高通	高通骁龙845	4x Kryo 385 Gold 4x kryo 385 Silver	Adreno 630	Hexagon685	-	2.8Hz	10nm LPP
三星	Exynos9810	4x Exynos M3 4x Cortex A55	Mali G72MP18	否	334 Gflops	2.9GHz	10nm LPP

人工智能芯片









优亿科技
USEEASE

人工智能芯片

公司	芯片	说明
高通	骁龙	发布骁龙神经处理引擎软件开发工具包挖掘骁龙SoC AI计算能力；与Facebook AI研究所合作研制AI芯片；收购NXP致力于发展智能驾驶芯片
谷歌	TPU (TensorFlow Processing Unit)	专为其深度学习算法Tensor Flow设计，也用在AlphaGo系统、StreetView和机器学习系统RankBrain中，第二代Cloud TPU理论算力达到了180T Flops，能够对机器学习模型的训练和运行带来显著的加速效果。
英伟达	GPU	适合并行算法，占目前AI芯片市场最大份额，应用领域涵盖视频游戏、电影制作、产品设计、医疗诊断等各个门类。
AMD	GPU	GPU第二大市场。
英特尔	FPGA	来自167亿美元收购的Altera，峰值性能逊色于GPU，指令可编程，且功耗也要小得多，适用于工业制造、汽车电子系统等，可与至强整合。
	Xeon Phi Knights Mill	适用于包括深度学习在内的高性能计算，能充当主处理器，（据称）可以在不配备其它加速器或协处理器高效处理深度学习应用。
微软	FPGA	自主研发，已被用于Bing搜索，能支持微软的云服务Azure，速度比传统芯片快得多。
微软	FPGA	自主研发，已被用于Bing搜索，能支持微软的云服务Azure，速度比传统芯片快得多。
Xilinx	FPGA	世界最大的FPGA制造厂商，2016年底推出支持深度学习的reVision堆栈。
IBM	TrueNorth 真北类脑芯片	是一种基于神经形态工程，2011年和2014年分别发布了“TrueNorth”第一代和第二代类脑芯片，二代神经元增加到100万个，可编程数量增加976倍，每秒可执行460亿次突破计算。
苹果	专用芯片 Apple Neural Engine	该芯片定位于本地设备AI任务处理，把面部识别、语音识别等AI相关任务集中到AI模块上，提升AI算法效率，未来可能嵌入苹果的终端设备中。
Mobileye	EyeQ5	用于汽车辅助驾驶系统
Movidius	Myriad 图形处理器	Myriad系列图形处理器已经被联想用来开发下一代虚拟现实产品；谷歌Project Tango 3D传感器技术背后的功臣。
地平线机器人	BPU	推出BPU架构，第一代BPU在FPGA和ARM架构上实现
深鉴科技	DPU	基于Xilinx FPGA推出DPU
华为	麒麟970	嵌入式神经网络处理器（NPU）芯片，解决端侧AI挑战；单元架构能够对深度学习的神经网络架构实现通用性的支撑，以超高的性能功耗比实现AI训练及应用在移动端的落地。
中星微电子	星光智能一号	嵌入式神经网络处理器（4个NPU核）芯片，加速人工神经网络模型。
寒武纪	寒武纪一号 DianNao	嵌入式神经网络处理器（NPU）芯片，加速人工神经网络模型。

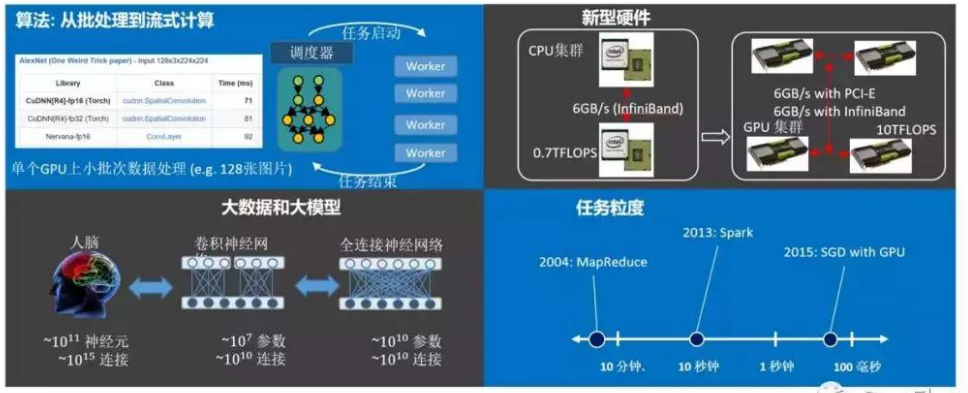
03 移动端模型

人工智能手机是一种由AI技术强化加持的智能通讯设备，从**硬件芯片层**到**操作系统层**和**应用交互层**均整合了人工智能技术的移动通讯设备，从而帮助用户可以更轻松便捷地下达指令，完成用户要做的事，降低用户的认知负担。

软件层模型优化

对于终端来说，需要将软件应用层做到可以实用，最重要的是需要将原来的深度学习模型进行优化处理。传统的深度学习网络一般都是具有上亿参数，而对于需要考虑功耗和散热的移动终端来说，AI芯片无法具备和PC芯片一样的性能，其巨大的存储和计算代价也使得其实用性特别是在移动设备上的应用受到了很大限制，因此需要对原来的软件和模型进行优化和适应。

硬件越快，软件越难



模型优化

模型压缩有很多分支包括：

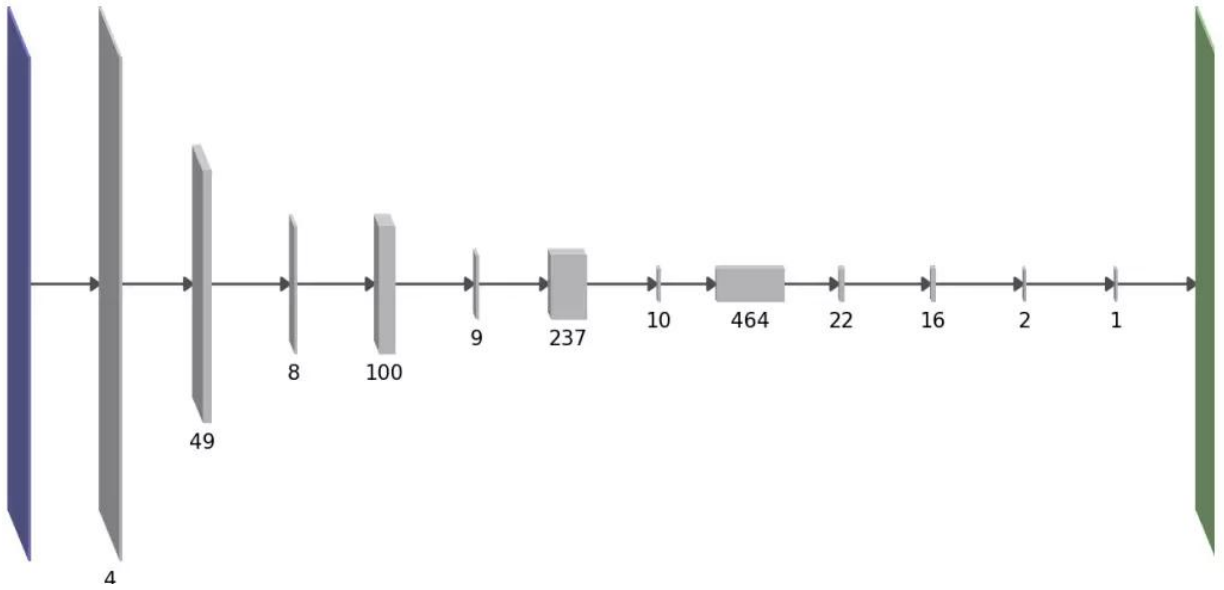
（1）剪枝（Pruning）；

（2）基于稀疏表示（Sparse Representation）；

（3）量化（Quantization）；

（4）矩阵分解（Matrix decomposition）；

（5）学习小网络（teacher-student learning paradigm）。



模型优化

- **剪枝与稀疏**

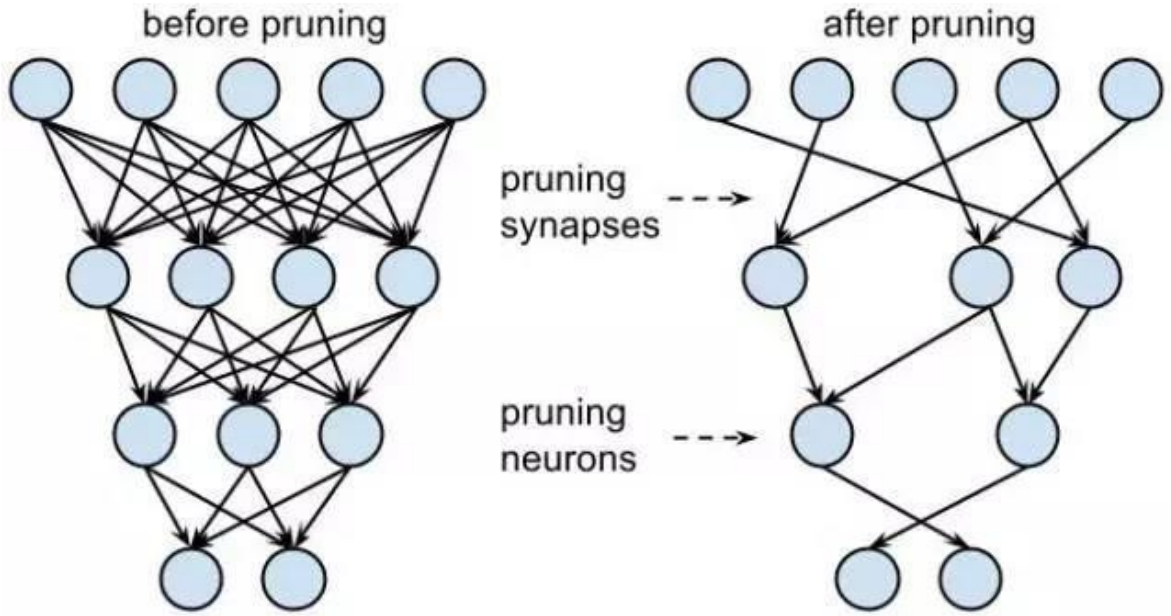
  剪枝就是移除掉一些不重要的权重连接，在尽量保持相同性能的前提下降低计算成本，删除那些在深度网络结构中不真正使用的特征可以加速推断和训练的过程。先在原始的网络结构下训练网络，然后通过设置一个阈值，把一些权值较小的连接进行剪枝，重新训练权重，对训练好的模型再剪枝，再重新训练，直到满足设定条件为止。

- **深度压缩**

  深度压缩是指通过对权重进行量化，使得许多连接共享同一权重，并且只需要存储码本(有效的权重)和索引的方法。深度压缩包括权值量化和哈希编码、定值化等。神经网络中的权值一般用单精度的浮点数表示，需要占用大量的存储空间。而用码书对权值进行量化可以共享权值，从而减轻存储负担。哈希编码是指利用哈希对权重进行编码，从而达到压缩权重的目的。

剪枝与稀疏存储

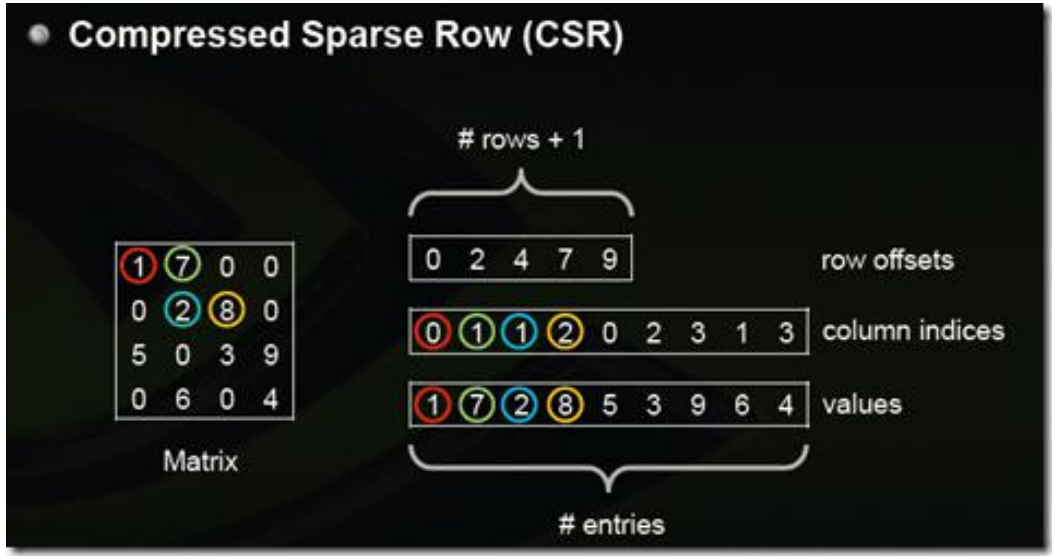
- **“剪枝”的步骤主要可以分为三步：**
- （1） 进行正常的网络训练；
- （2） 删除所有权重小于一定阈值的连接；
- （3） 对上面得到的稀疏连接网络再训练；



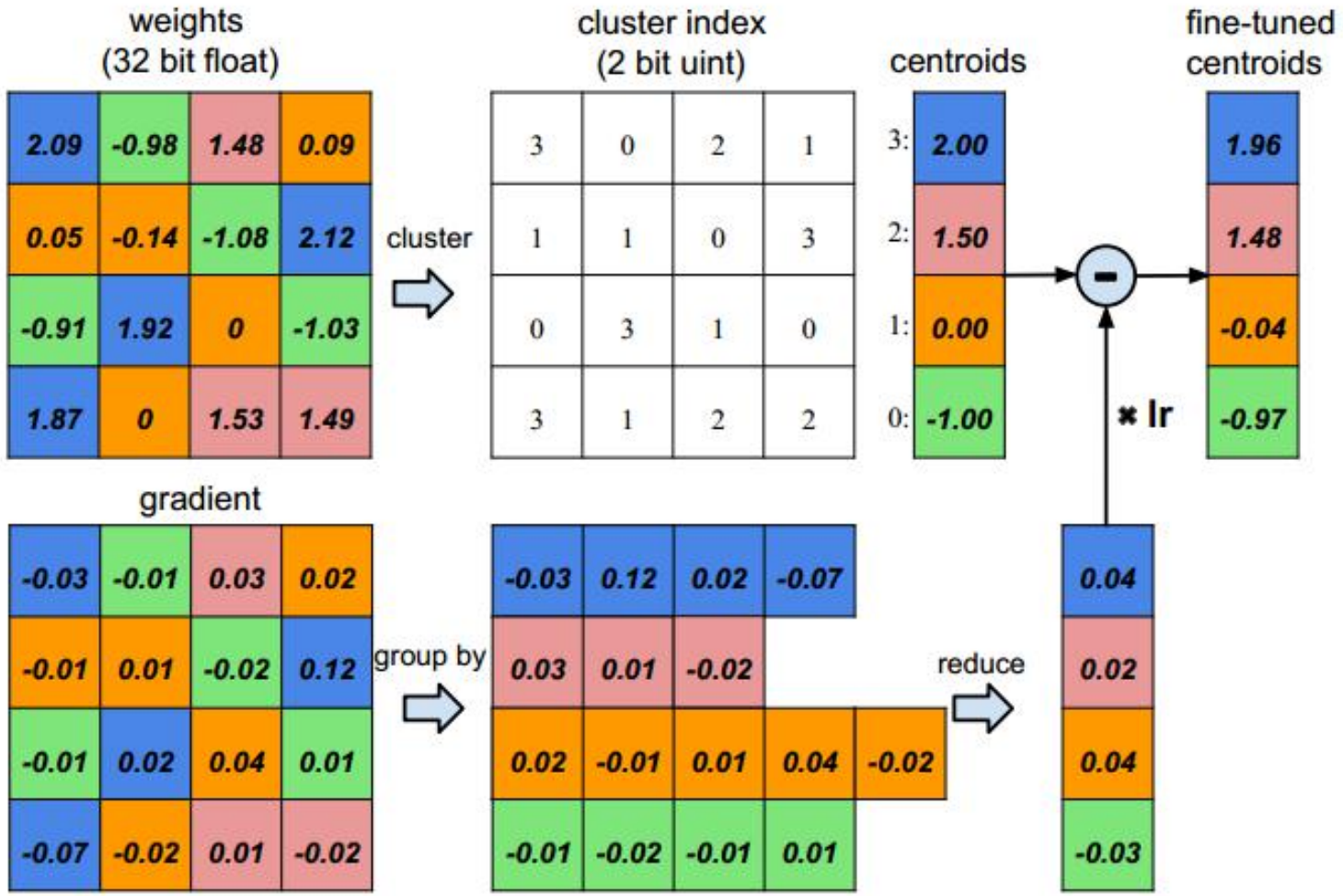
剪枝与稀疏存储

- **存储稀疏**

存储稀疏结构时采用的是稀疏压缩行CSR或者稀疏压缩列CSC，假设非0元素个数为a,行或者列数为n,那么我们需要存储的数据量仅需要 2a+n+1。 以CSR为例，我们存储时采用的是3元组结构，即：行优先存储a个非零数，记为A；a个非零数所在列的列号；每一行第一个元素在A中的位置+非零数个数。

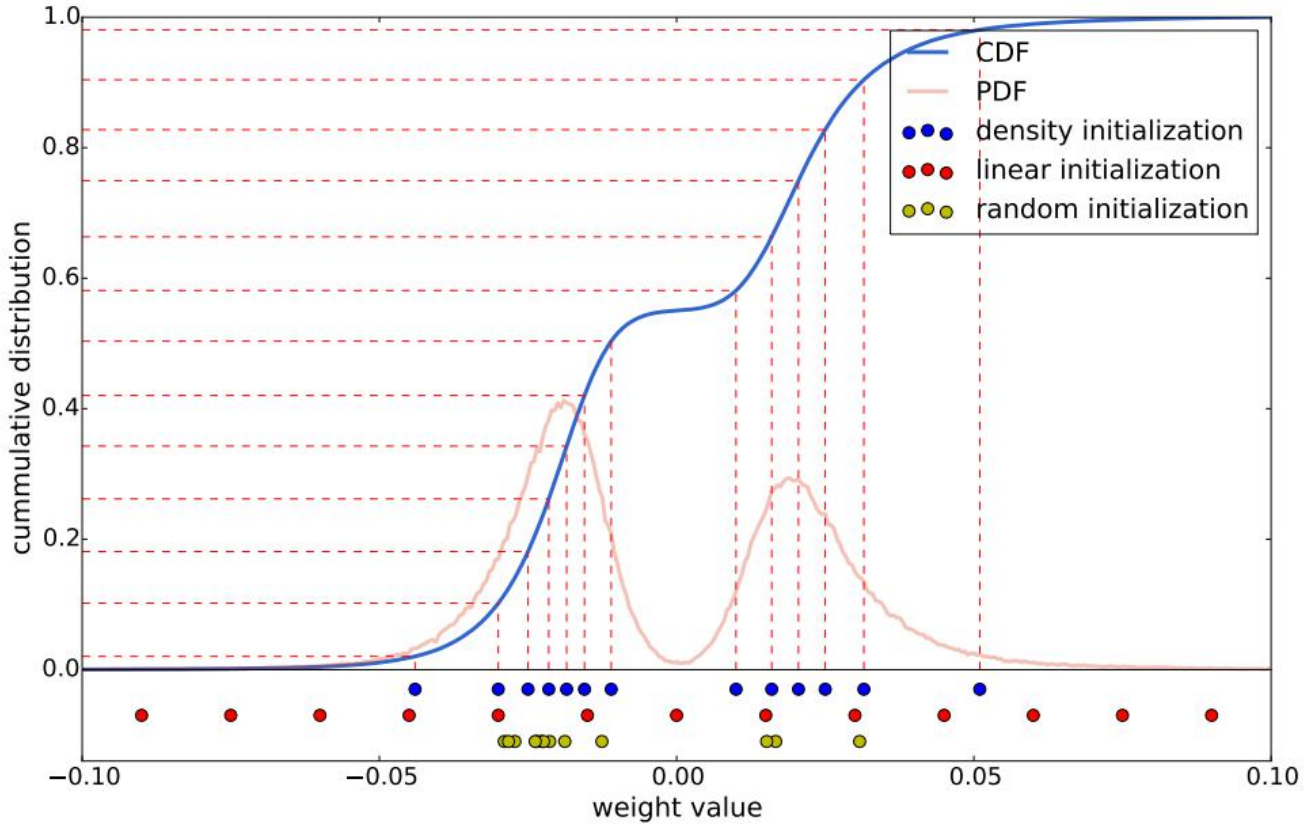


训练量化和权重共享



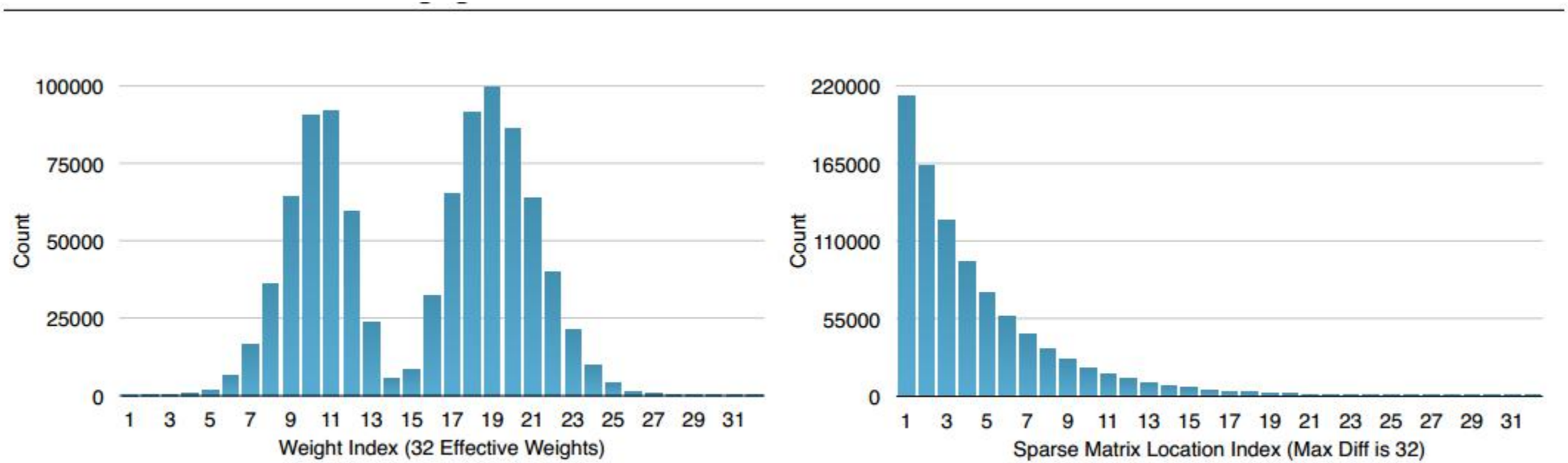
训练量化和权重共享

- 码字的确定



哈夫曼编码

Huffman编码是一种可变字长编码(VLC)，该方法完全依据字符出现概率来构造异字头的平均长度最短的码字。Huffman编码作为优化主要用在最后一个全连接层，由于其非均匀分布，利用Huffman编码来对其进行处理，最终可以使的网络的存储减少20%~30%。



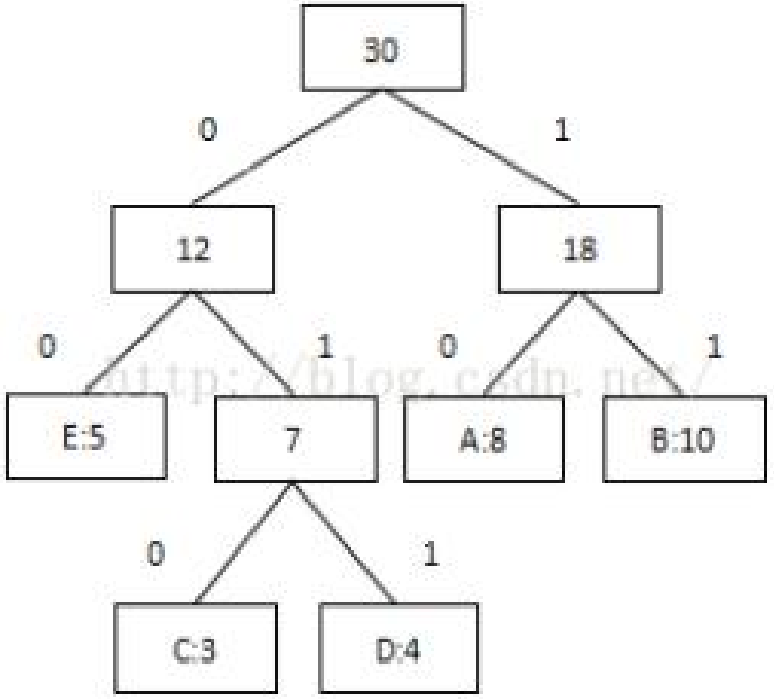
Distribution for weight (Left) and index (Right). The distribution is biased.

哈夫曼编码

Huffman编码

计算出每个字符出现的次数，把出现次数（概率）最小的两个相加，并作为左右子树，出现次数（概率）越多的会越在上层，编码也越短，出现频率越少的就越在下层，编码也越长。

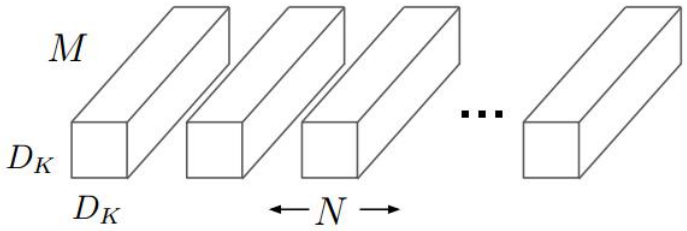
字符	次数	编码
B	10	11
A	8	10
C	3	010
D	4	011
E	5	00



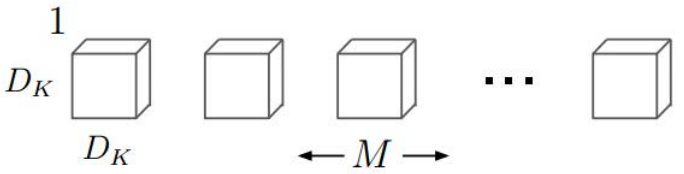
几种常用的模型压缩

- **MobileNet**

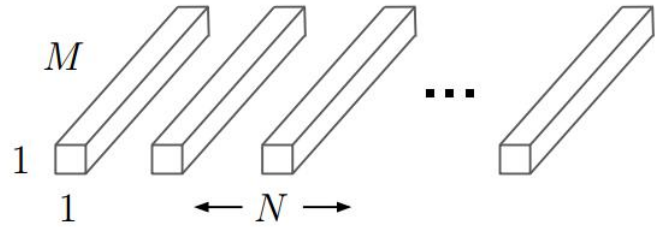
  MobileNet是基于一个流线型的架构，它使用深度可分离的卷积来构建轻量级的深层神经网络。MobileNets模型**基于深度可分解的卷积**，它可以将标准卷积分解成一个深度卷积和一个点卷积（1 × 1卷积核）。深度卷积的特点是一个卷积核负责一部分 feature map，每个 feature map 只被一个卷积核卷积。



(a) Standard Convolution Filters



(b) Depthwise Convolutional Filters

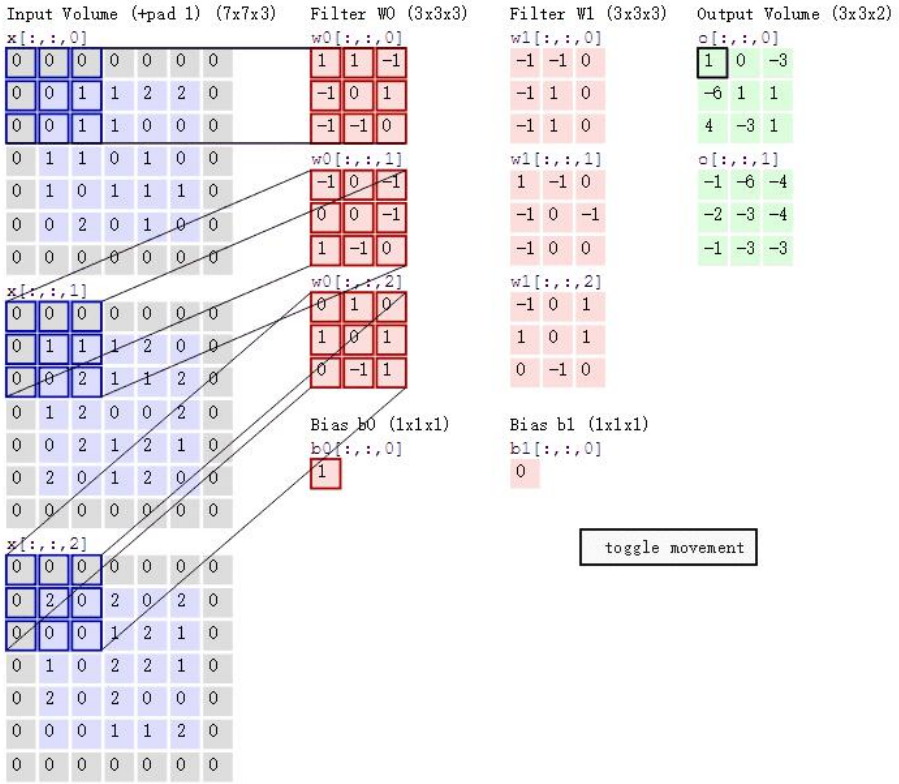


(c) 1 × 1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

几种常用的模型压缩

- **深度可分解卷积**

输入图片维度是11 × 11 × 3，标准卷积为3 × 3 × 3 ×16，那么可以得到输出为6 × 6 × 16的输出结果。现在输入图片不变，先通过一个维度是3 × 3 × 1 × 3的深度卷积，得到6 × 6 × 3的中间输出，然后再通过一个维度是1 × 1 × 3 ×16的1 ×1卷积，同样得到输出为6 × 6 × 16。但对于总共的参数为：11 × 11 × 3 ×3 × 3 ×16=52272个参数；第二种的参数为11 × 11 × 3×3 × 3 × 3+6 × 6 × 3× 1× 1×16=11529个参数，减少了近5倍的参数。



几种常用的模型压缩

- **MobileNet实验对比效果**

Table 8. MobileNet Comparison to Popular Models

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
GoogleNet	69.8%	1550	6.8
VGG 16	71.5%	15300	138

几种常用的模型压缩

- **ShuffleNet**

  ShuffleNet在MobileNet提出的深度可分解卷积的基础上主要采用通道重排（channel shuffle）和分组逐点卷积（group pointwise convolution）对模型的性能进行提升。

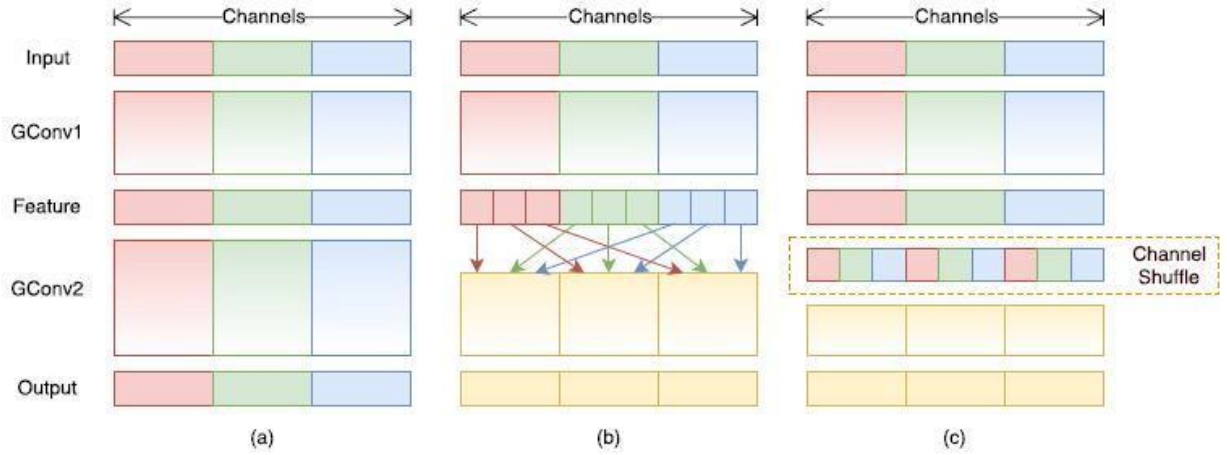


Figure 1: Channel shuffle with two stacked group convolutions. GConv stands for group convolution. a) two stacked convolution layers with the same number of groups. Each output channel only relates to the input channels within the group. No cross talk; b) input and output channels are fully related when GConv2 takes data from different groups after GConv1; c) an equivalent implementation to b) using channel shuffle.

几种常用的模型压缩

- **ShuffleNet实验效果对比**

Table 4: Classification error vs. various structures (%, *smaller number represents better performance*)

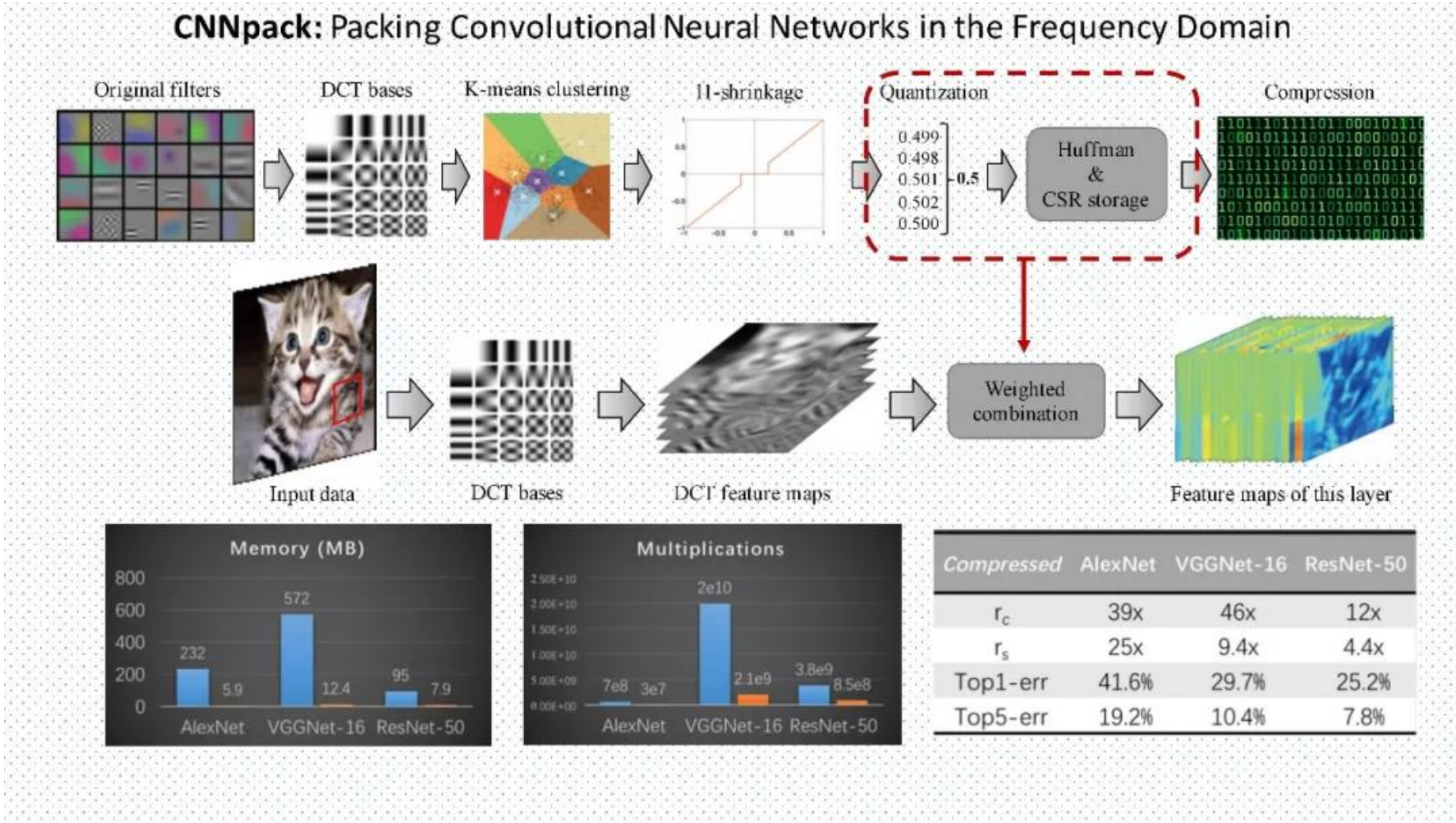
Complexity (MFLOPs)	VGG-like[4]	ResNet	Xception-like	ResNeXt	ShuffleNet (ours)
140	56.0	38.7	35.1	34.3	**34.1** ($1\times$, $g = 3$)
38	-	48.9	46.1	46.3	**43.7** ($0.5\times$, $g = 4$)
13	-	61.6	56.7	59.2	**53.7** ($0.25\times$, $g = 8$)
40 (arch2)	-	48.5	45.7	47.2	**42.7** ($0.5\times$, $g = 8$)
13 (arch2)	-	61.3	56.5	61.0	**53.3** ($0.25\times$, $g = 8$)

Table 5: ShuffleNet vs. MobileNet [12] on ImageNet Classification

Model	Complexity (MFLOPs)	Cls err. (%)	Δ err. (%)
1.0 MobileNet-224	569	29.4	-
ShuffleNet $2\times$ ($g = 3$)	524	**29.1**	0.3
0.75 MobileNet-224	325	31.6	-
ShuffleNet $1.5\times$ ($g = 3$)	292	**31.0**	0.6
0.5 MobileNet-224	149	36.3	-
ShuffleNet $1\times$ ($g = 3$)	140	**34.1**	2.2
0.25 MobileNet-224	41	49.4	-
ShuffleNet $0.5\times$ (arch2, $g = 8$)	40	**42.7**	6.7
ShuffleNet $0.5\times$ (shallow, $g = 3$)	40	45.2	4.2

几种常用的模型压缩

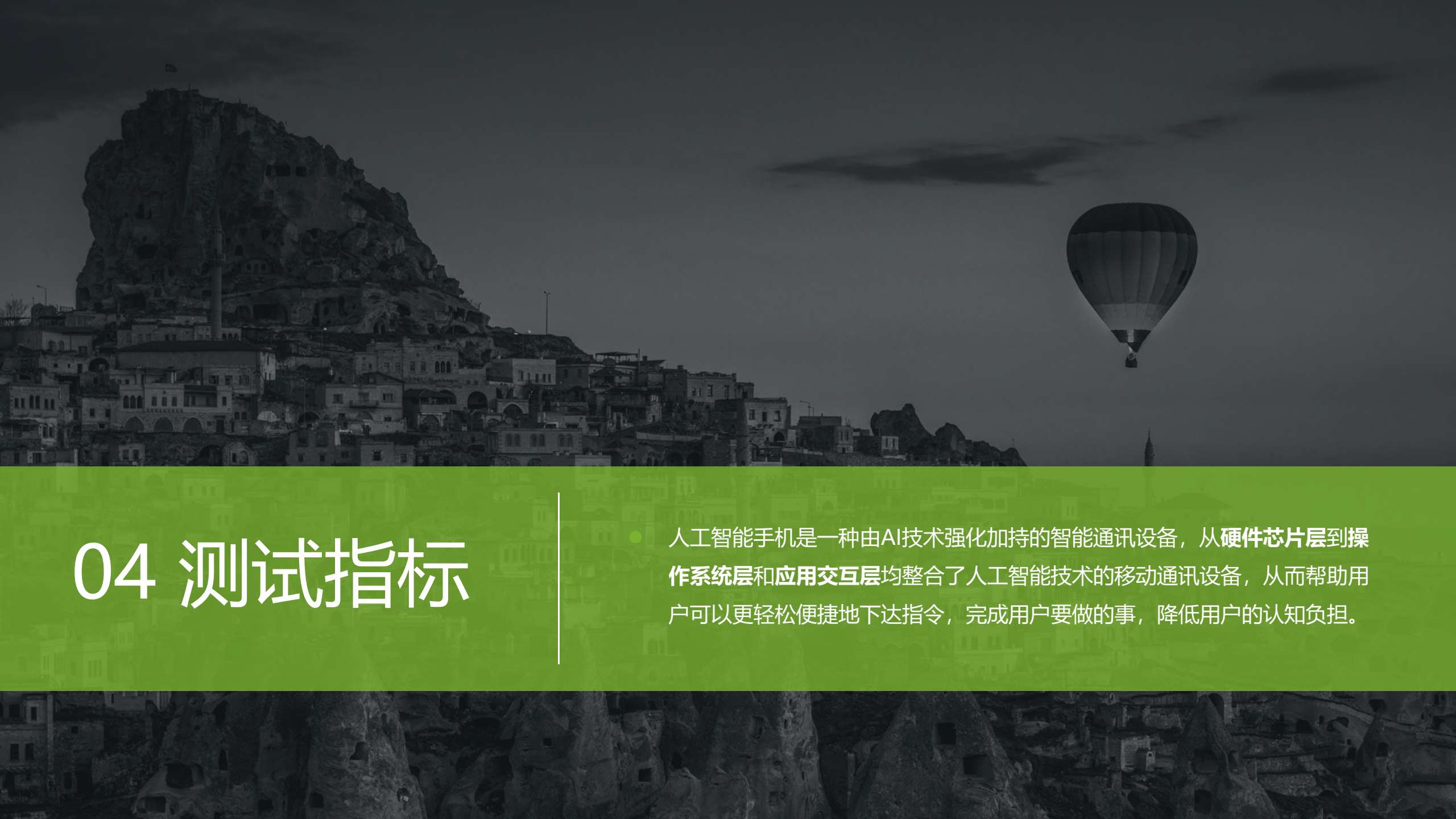
- **CNNpack**



Compressed	AlexNet	VGGNet-16	ResNet-50
r_c	39x	46x	12x
r_s	25x	9.4x	4.4x
Top1-err	41.6%	29.7%	25.2%
Top5-err	19.2%	10.4%	7.8%

模型优化

- 由于不同的神经网络模型的优化可以产生很大的区别，因此，我们在制定测试指标的时候并不能只考虑硬件和芯片的性能，还要直接测试其基于软件层面（模型）所表现出来的性能指标。



04 测试指标

人工智能手机是一种由AI技术强化加持的智能通讯设备，从**硬件芯片层**到**操作系统层**和**应用交互层**均整合了人工智能技术的移动通讯设备，从而帮助用户可以更轻松便捷地下达指令，完成用户要做的事，降低用户的认知负担。

智能手机的测试

- 智能手机的测试分为**硬件层和应用层**两个方面，其中对于硬件层来说主要针对的是芯片和硬件能力的测试；对于应用层来说，其主要反映的是一种综合能力的对比和测试。
- 对于硬件层的测试主要集中在对芯片的性能、功耗等指标。
- 对于应用层的测试是对一台手机智能化程度综合水平的测试，包括对手机智能识别相关的API的响应速度、耗电量、网络峰值等指标。

05 芯片测试

人工智能手机是一种由AI技术强化加持的智能通讯设备，从**硬件芯片层**到**操作系统层**和**应用交互层**均整合了人工智能技术的移动通讯设备，从而帮助用户可以更轻松便捷地下达指令，完成用户要做的事，降低用户的认知负担。

深度学习专用芯片基准测试工具

- 基准测试（Benchmarking）：是指通过设计科学的测试方法、测试工具和测试系统，实现对一类测试对象的某项性能指标进行定量的和可对比的测试。基准测试通常用于评估计算机硬件的性能特征，例如CPU的浮点运算性能，GPU的图像处理能力，存储系统访问的速度等等，有时也用于软件或者编译工具。总的来说，基准测试提供了一种比较不同软硬件系统性能的方法。

智能芯片基准测试工具

- 2017年10月23日，中国人工智能产业发展联盟（AIIA）总体组组织了人工智能芯片Benchmark研讨会，正式启动人工智能芯片Benchmark项目。现在市面上主要的智能芯片测试工具有：

  （1）斯坦福大学的DAWNBench；

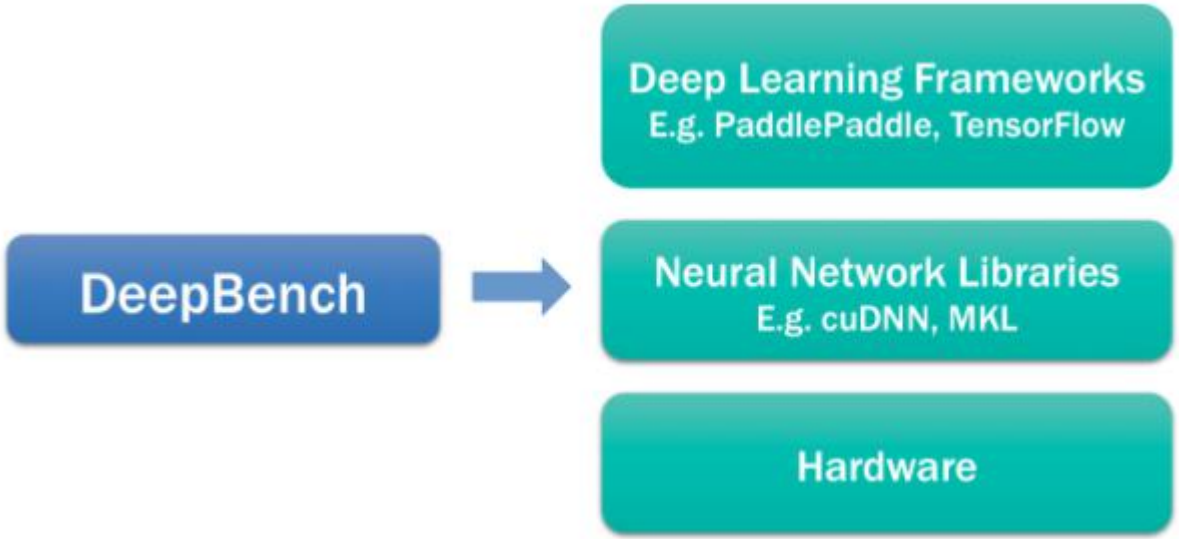
  （2）MIT Eyeriss团队的DNN Processor Benchmarking Metrics；

  （3）百度的DeepBench；

  （4）鲁大师的AI芯片测试。

深度学习专用芯片基准测试工具

- **DeepBench** 测试的是深度学习模型训练中的基础运算，主要包括一系列基础操作（稠密矩阵相乘、卷积和通信）以及一些循环层类型，推理工作负载（inference workloads）和更低精度的算法。DeepBench的测试包括了深度学习框架、神经网络库和硬件。



深度学习专用芯片基准测试工具

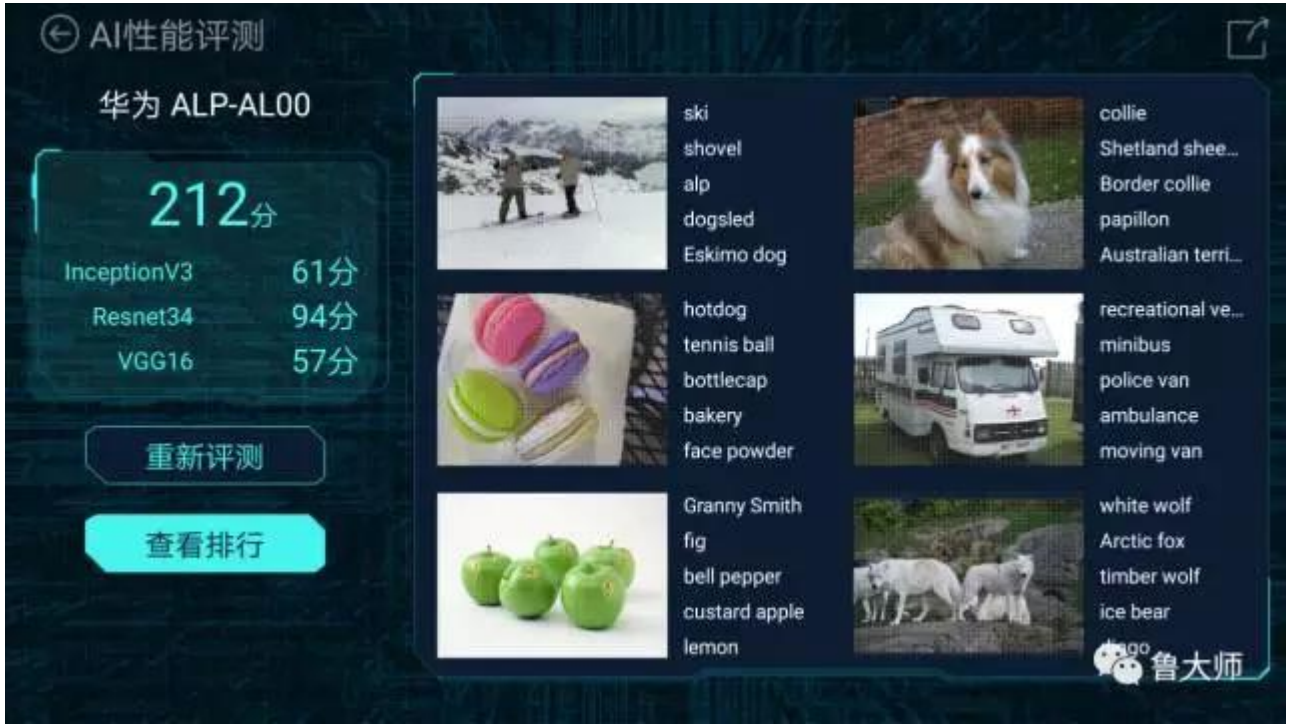
- **DAWNBench** 是一个斯坦福开源的深度学习训练基准工具，它提供了一个评测神经网络训练效率和训练时间的方法。其使用的数据集包括CIFAR10图像数据集、ImageNet图像数据集、SQuAD问答数据集。

Rank	Time to 93% Accuracy	Model	Hardware	Framework
1 Jan 2018	14:37:59	ResNet50 *DIUX* source	p3.16xlarge	tensorflow 1.5, tensorpack 0.8.1
2 Dec 2017	1 day, 20:28:27	ResNet152 *ppwwyyxx* source	8 P100 / 512 GB / 40 CPU (NVIDIA DGX-1)	tensorpack 0.8.0
3 Oct 2017	10 days, 3:59:59	ResNet152 *Stanford DAWN* source	8 K80 / 488 GB / 32 CPU (Amazon EC2 [p2.8xlarge])	MXNet 0.11.0

Rank	Cost (USD)	Model	Hardware	Framework
1 Jan 2018	$358.22	ResNet50 *DIUX* source	p3.16xlarge	tensorflow 1.5, tensorpack 0.8.1
2 Oct 2017	$1112.64	ResNet152 *Stanford DAWN* source	8 K80 / 488 GB / 32 CPU (Amazon EC2 [p2.8xlarge])	MXNet 0.11.0
3 Oct 2017	$2323.39	ResNet152 *Stanford DAWN* source	4 M60 / 488 GB / 64 CPU (Amazon EC2 [g3.16xlarge])	TensorFlow v1.3

深度学习专用芯片基准测试工具

- **鲁大师AI性能测试**使用目前较为常用的三种神经网络Inception V3、ResNet34、VGG16的特定算法，机器识别图片内容，按照概率高低输出可能的结果列表。该测试项目会提供一系列被测试的图片，如华为Mate10的麒麟970当中的AI协处理器将发挥其作用在测试中对其进行识别。



基准测试

- 通过对现有市面上测试工具的调研和分析，我们觉得基于AI芯片的性能主要从以下几种方面进行会比较合理：
- （1）稠密矩阵相乘计算测试。
- （2）卷积计算测试.
- （3）循环层计算测试.
- （4）All-Reduce 计算测试.

基准测试

- **（1）稠密矩阵相乘计算测试**：

  现在几乎所有深度学习网络都包含稠密矩阵相乘。它们被用于执行全连接层和 vanilla RNN，以及为其他类型的循环层建立基石。有时它们也被用于快速执行自定义代码缺失的新类层。

- **（2）卷积计算测试**：

  卷积构成了网络中图像和视频操作的绝大多数的浮点计算，也是语音和自然语言模型网络的重要部分。不同大小的过滤器和图像可选择不同的卷积计算方法，例如，direct approaches、基于矩阵相乘的方法、基于 FFT 的方法以及基于 Winograd 的方法。

基准测试

- **（3）循环层计算测试**：

  循环层总是由之前的运算与一元（unary）或二元（binary）运算这样的简单计算结合而成的——这些简单运算不是计算密集型的，通常只需占据总体运算时间的一小部分。然而，在循环层中，GEMM 和卷积运算相对较小，所以这些更小运算的成本变得有极大影响。

基准测试

- **(4) All-Reduce 计算测试**：

  神经网络通常在多 GPU 或多系统与 GPU 并行的情况下训练，这主要有两个技术分类：同步和异步。同步技术依赖于保持参数和所有模型实例的同步，它通常要保证在优化步骤执行前，所有模型实例有一些梯度的备份。最简单运行这些计算结果的 Message Passing Interface (MPI) 被称为 All-Reduce。

  有很多可以执行 All-Reduce 的方法，我们可以依靠数字的排列、数据的大小和网络的拓扑结构来执行。为了评估 All-Reduce，我们一般使用NVIDIA’s NCCL Ohio State University (OSU) Benchmarks作为基准库进行测试。

06 综合测试指标

人工智能手机是一种由AI技术强化加持的智能通讯设备，从**硬件芯片层**到**操作系统层**和**应用交互层**均整合了人工智能技术的移动通讯设备，从而帮助用户可以更轻松便捷地下达指令，完成用户要做的事，降低用户的认知负担。

智能手机测试

- 应用层的智能测试相对于智能芯片测试，更关注手机其作为一个整体在AI性能方面的能力，包括图像识别、自能问答和语音识别。测试类型主要包括以下几点：
- （1）采用**CIFAR数据集和ImageNet图片数据集**测试人工智能手机在图像识别和机器视觉等方面的能力，所需要测试的指标包括图像识别的错误率、图像识别的速率（千张/分钟）和所需的功耗。
- （2）采用**TIMIT语音识别数据集**测试人工智能手机在语音识别的能力，所需要测试的指标包括语言识别的准确率、语音识别的响应时间。
- （3）采用**SQuAD问答数据集**测试人工智能手机对自然语言处理的能力，所需要测试的指标包括问答的准确性、响应时间、唤醒所需要时间。

综合测试指标

	测试数据集	评估指标	性能	能耗
图像识别	CIFAR/ ImageNet	错误率	识别速率	√
目标检测	COCO	平均准确率/平均召回率	识别速率	√
人脸识别	CMU-PIE	误识率/拒识率	识别速度/注册速度	√
语音识别	TIMIT	词错误率/字错误率	识别速度/回复速度	√
同声传译	AI Challenger (英中机器同声传译)	双语互译质量评估BLEU	同步率	√
智能问答	SQuAD	精确批配/F1	响应速度	√

测试数据集介绍

- **CIFAR图片数据集**：是一个用于普通物体识别的数据集，CIFAR-10数据集包括由10个类别的事物,每个事物各有6000张彩色图像,每张图片的大小是32*32。
- **ImageNet图片数据集**：是目前深度学习图像领域应用得非常多的一个领域，关于图像分类、定位、检测等研究工作大多基于此数据集展开，ImageNet有1400多万幅图片，涵盖2万多个类别；其中有超过百万的图片有明确的类别标注和图像中物体位置的标注。
- **SQuAD问答数据集**：是一份由斯坦福整理的关于维基百科文章的问答数据集。有500多篇文章和10万多问答数据对。SQuAD是最常用的自动问答数据集。
- **TIMIT语音识别数据集**：是由德州仪器(TI)、麻省理工学院(MIT)和坦福研究院(SRI)合作构建的声学-音素连续语音语料库，数据集的语音采样频率为16kHz，一共包含6300个句子，由来自美国八个主要方言地区的630个人每人说出给定的10个句子，所有的句子都在音素级别(phone level)上进行了手动分割，标记。


Thanks for
your attention!