首都师范大学

為學為師求實求新

深度学习应用与工程实践
1.课程介绍
1. Introduction

李冰
副研究员，硕士生导师
交叉科学研究院
首都师范大学

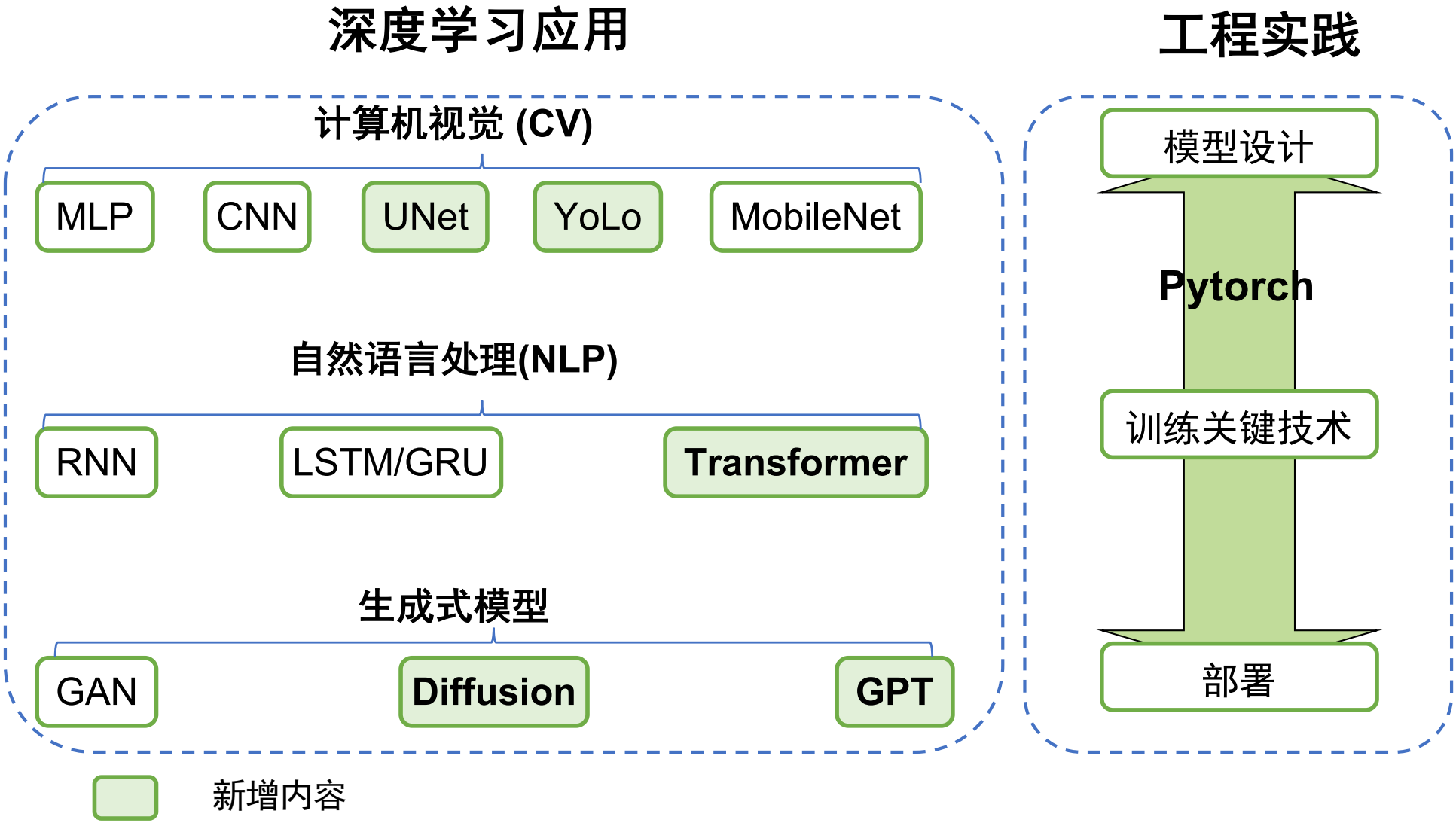
自我介绍

老师:	李冰 校本部教二楼321研究室 bing.li@cnu.edu.cn
上课时间:	周一 13:30-15:50 教二楼612
答疑:	邮件 企业微信
课程联系人	王炳猛

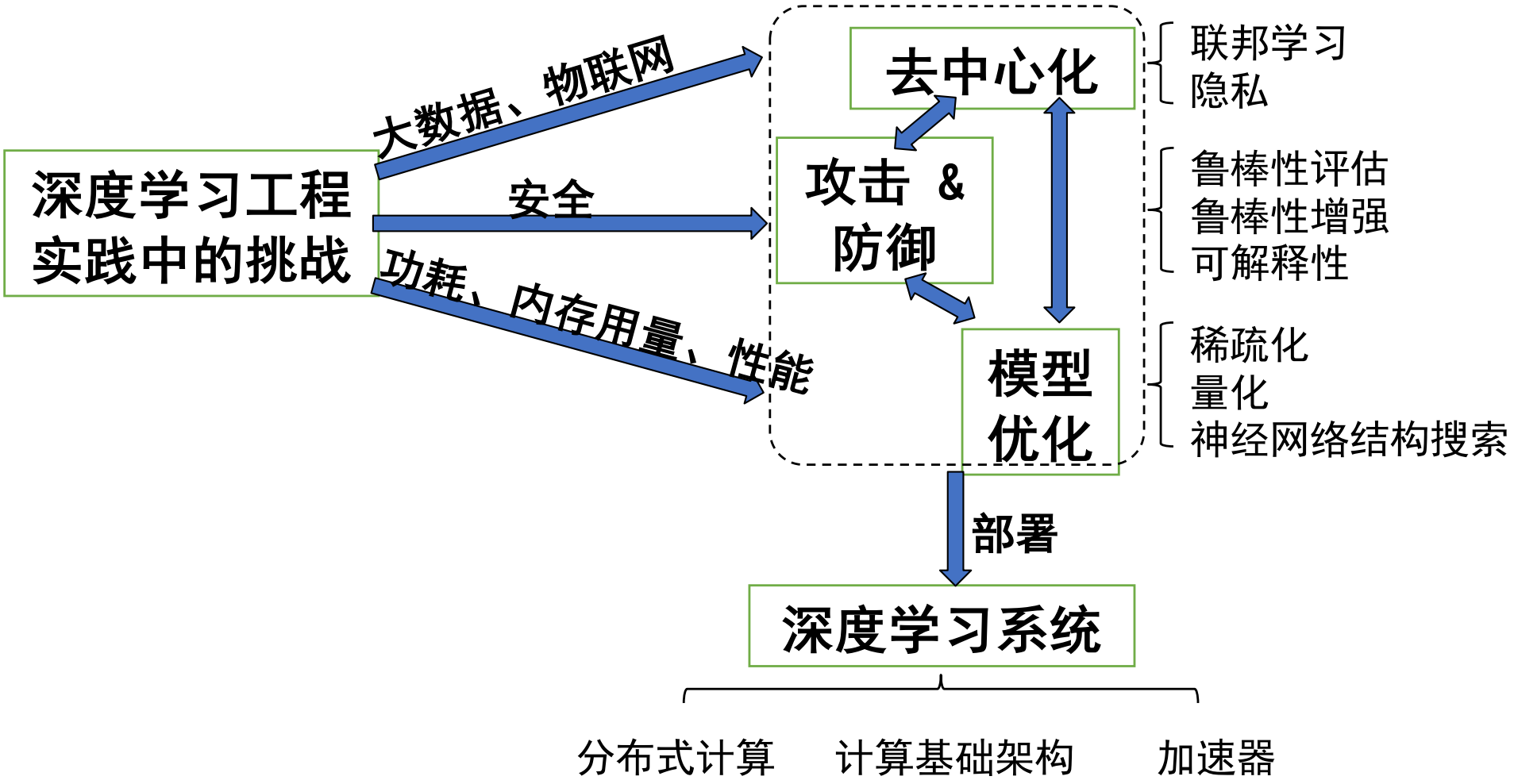
主要内容

- 课程介绍
- 深度学习
  - 应用
  - 类别
  - 重要指标
  - 软件框架
  - 硬件平台

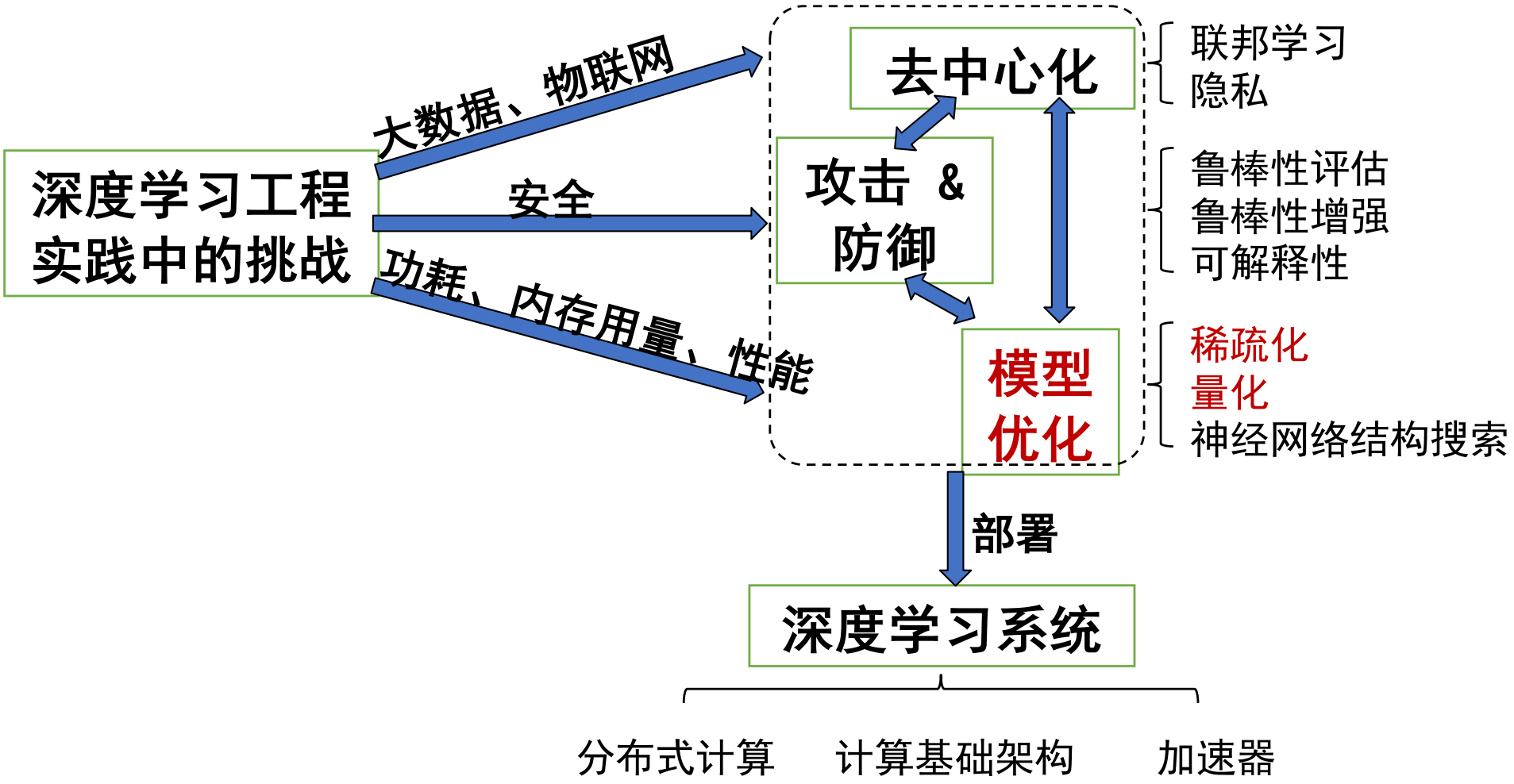
本门课的内容


深度学习应用
工程实践
计算机视觉 (CV)
MLP
CNN
UNet
YoLo
MobileNet
自然语言处理(NLP)
RNN
LSTM/GRU
Transformer
生成式模型
GAN
Diffusion
GPT
模型设计
Pytorch
训练关键技术
部署
新增内容

本门课的内容


深度学习工程
实践中的挑战
大数据、物联网
安全
功耗、内存用量、性能
去中心化
攻击 &
防御
模型
优化
联邦学习
隐私
鲁棒性评估
鲁棒性增强
可解释性
稀疏化
量化
神经网络结构搜索
部署
深度学习系统
分布式计算
计算基础架构
加速器

本门课的内容


深度学习工程
实践中的挑战
大数据、物联网
安全
功耗、内存用量、性能
去中心化
攻击 &
防御
模型
优化
联邦学习
隐私
鲁棒性评估
鲁棒性增强
可解释性
稀疏化
量化
神经网络结构搜索
部署
深度学习系统
分布式计算
计算基础架构
加速器

你们将会学到

- DNN 基础
  - 不同应用领域的经典网络
    - 卷积神经网络(CNN)、循环神经网络 (RNN)等
  - 网络训练方法
    - 数据集准备；优化方法；误差设计
- DNN 优化技术
  - 轻量化网络, 模型压缩, 剪枝, 量化, 稀疏化, …
- DNN工程实践能力
  - 搭建工程环境
  - 查找资源
  - Python (Numpy), Pytorch 框架的使用
- 高阶话题
  - 分布式计算, 网络结构搜索(NAS), 去中心化训练和隐私保护，深度学习安全

考核方式

- 考勤 （20%）
  - 不点名
- 作业 (50%)
- 期末 (40%)：
    - 论文研读
    - ppt报告

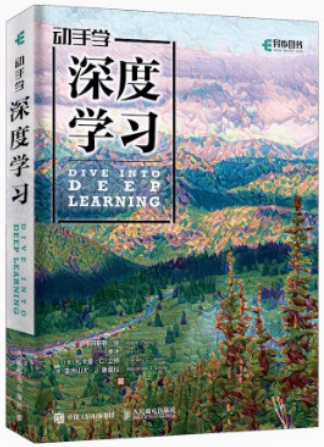
教材和软件框架

1.动手学深度学习pytorch

作者: 阿斯顿·张（Aston Zhang） / 李沐（Mu Li） / [美] 扎卡里·C. 立顿（Zachary C. Lipton） / [德] 亚历山大·J. 斯莫拉（Alexander J. Smola）

- http://zh.d2l.ai/

《动手学深度学习》



《动手学深度学习》

第二版

跳转第一版

面向中文读者的能运行、可讨论的深度学习教科书

含 PyTorch、NumPy/MXNet、TensorFlow 和 PaddlePaddle 实现

被全球 60 多个国家 400 多所大学用于教学

☆ Star 38,930

教材和软件框架

1.动手学深度学习pytorch

作者: 阿斯顿·张（Aston Zhang） / 李沐（Mu Li） / [美] 扎卡里·C. 立顿（Zachary C. Lipton） / [德] 亚历山大·J. 斯莫拉（Alexander J. Smola）

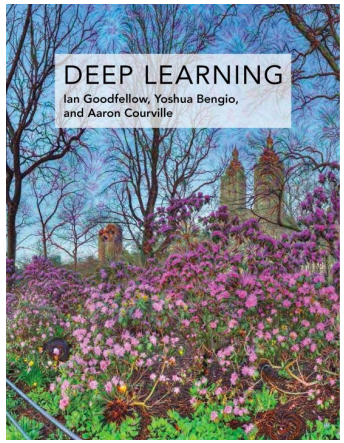
2.深度学习 (2016),

作者: [美] 伊恩·古德费洛 / [加] 约书亚·本吉奥 / [加] 亚伦·库维尔

3. 机器学习

作者：周志华，清华大学出版社，2016.

- 推荐下载使用Pytorch (https://pytorch.org/)



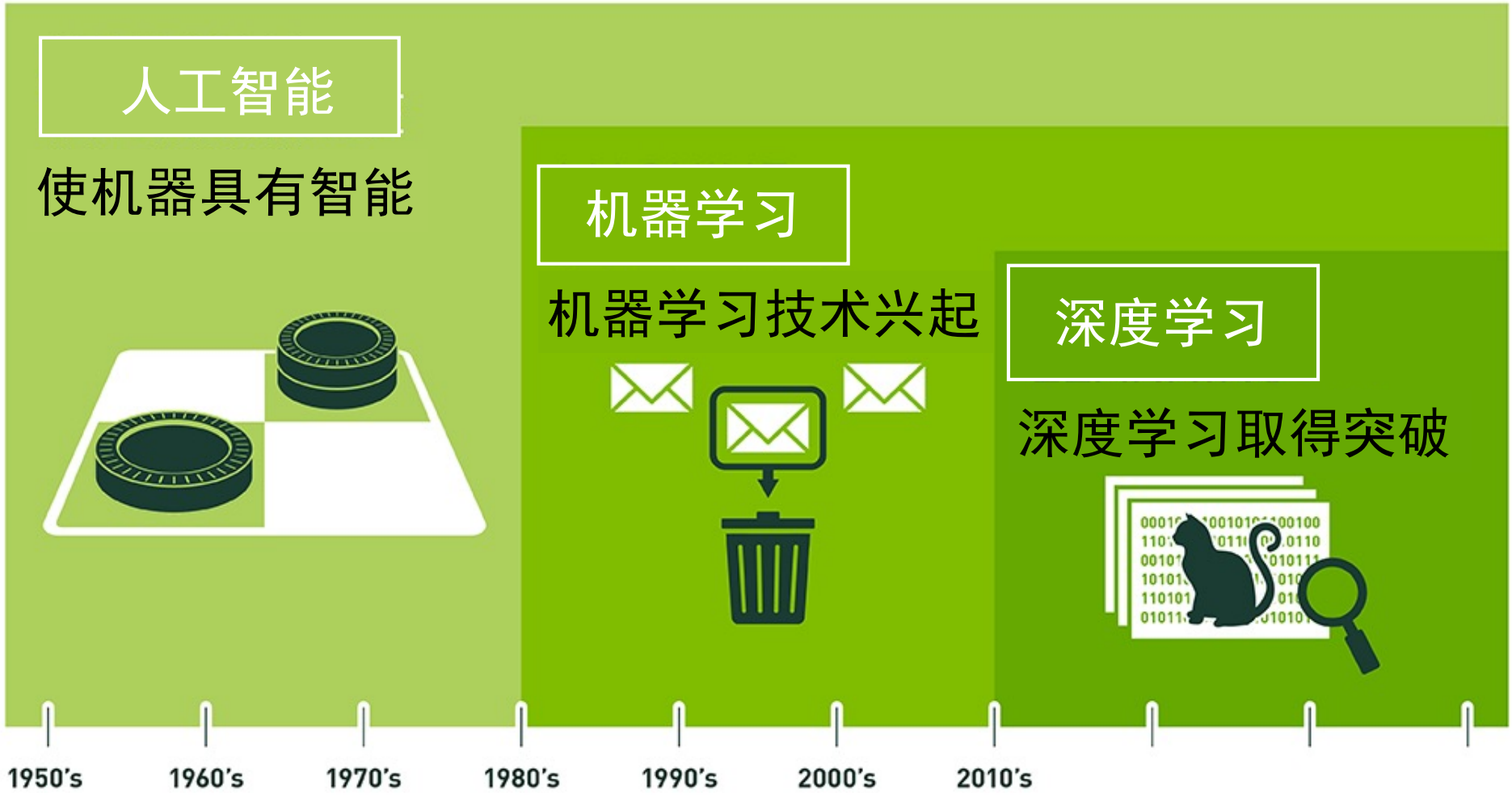
课程目标

- 了解深度神经网络
- 了解深度学习训练常用技术
- 学会这些技术进行实验
  - 学会分析实验结果并总结
  - 通过图像、图例和“看图说话”来呈现实验结果和数据
- 学会查找资料

主要内容

- 课程介绍
- 深度神经网络
  - 发展历程
  - 应用
  - 类别
  - 重要指标
  - 软件框架
  - 硬件平台

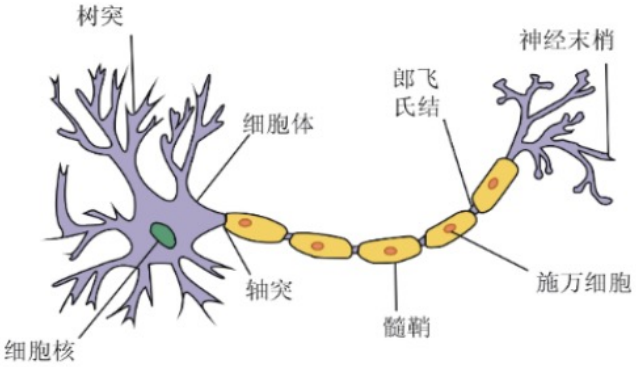
人工智能 ⇔ 机器学习 ⇔ 深度学习



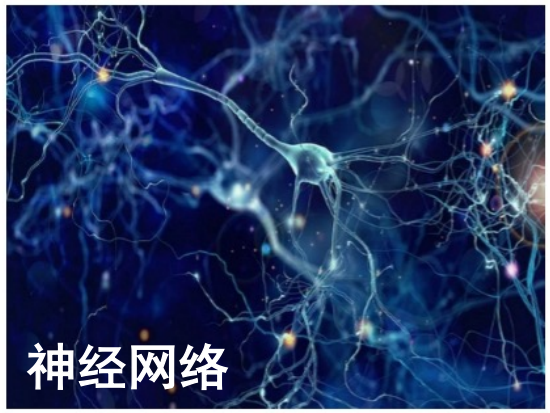
深度学习使人工智能时代更快到来

Source of image: https://blogs.nvidia.com.tw/2016/07/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

ANN：神经元与神经网络



神经元



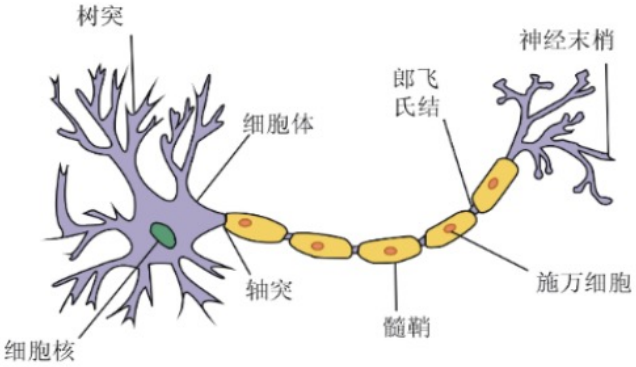
由1000多亿神经细胞（神经元）

深度学习的基础理论源于人工神经网络（Artificial Neural Network, ANN）。

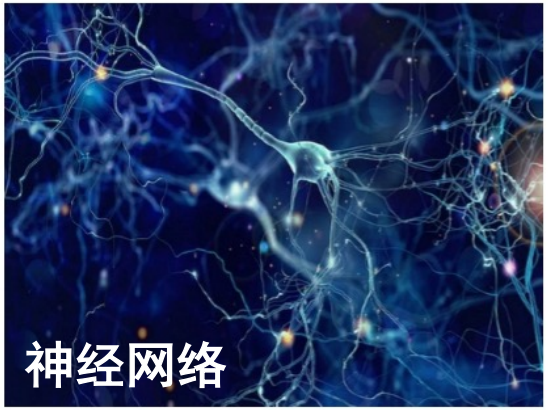
“早期ANN：构建类似人脑神经系统的神经网络结构，运用**大量简单处理单元**经广泛连接来组成**人工神经网络**，从而达到**模拟人脑学习**的目的。”---- **类脑**

“深度学习通过逐层提取低层特征形成更加抽象的高层表示，来进行事物类别或属性特征的自动化学习，以发现数据的分布式特征表示。”

ANN：神经元与神经网络

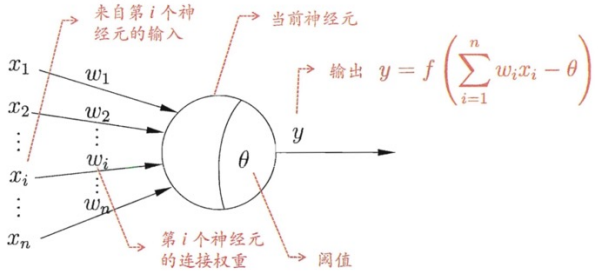


神经元

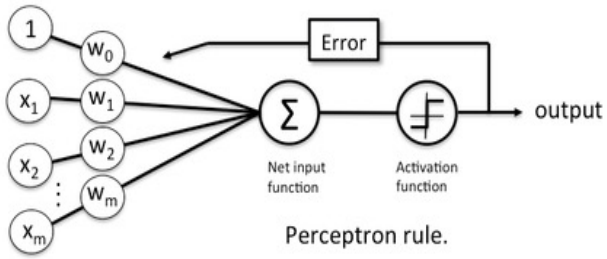


由1000多亿神经细胞（神经元）

1943年
心理学家McCulloch和数学家Pitts
第一个神经元的数学表达MP模型



1958年
Rosenblatt
第一个神经网络的原型
感知机



1969年
AI先驱Marvin Minsky

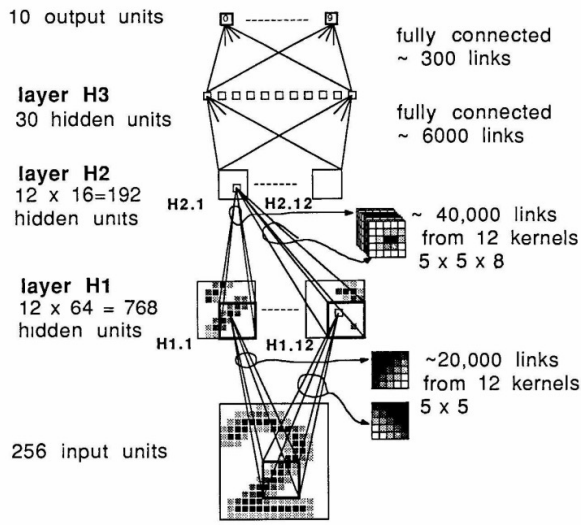
- **感知机的局限**

无法解决简单的异或（XOR）等线性不可分问题

- **专家系统兴起**

发展历程 --- 总览

卷积神经网络
(80年代)



字长：32位;
晶体管数目：275K
峰值主频： 40 MHz

杰弗里·辛顿
“深度学习之父”

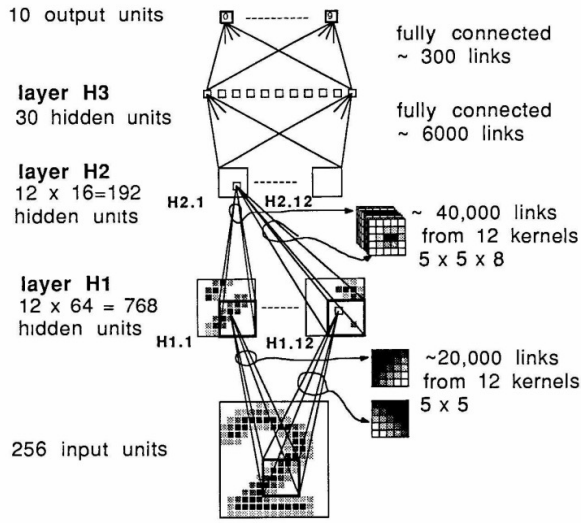
Y. Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition.1989.

J. Schmidhuber. Deep Learning in Neural Networks: An Overview. arxiv, 2014.

G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. Science, 2006.

发展历程 --- 总览

卷积神经网络 (80年代)



字长：32位;
晶体管数目：275K
峰值主频： 40 MHz

黑暗时期 (90年代)

- 梯度消失或爆炸问题
- 神经网络的层数增多
效果却不尽如人意
- 计算机能力有限

字长：32位;
晶体管数目：9.5M
峰值主频： 450 MHz

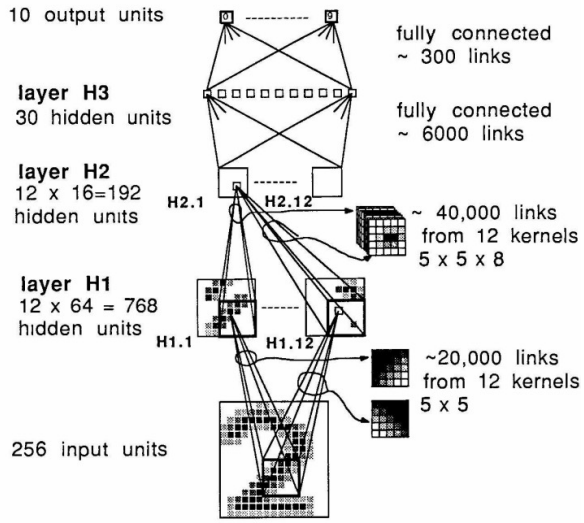
Y. Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition.1989.

J. Schmidhuber. Deep Learning in Neural Networks: An Overview. arxiv, 2014.

G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. Science, 2006.

发展历程 --- 总览

卷积神经网络 (80年代)



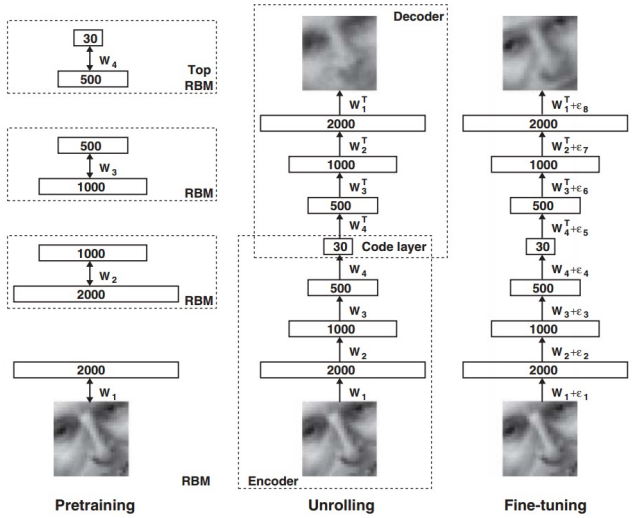
字长：32位;
晶体管数目：275K
峰值主频： 40 MHz

黑暗时期 (90年代)

- 梯度消失或爆炸问题
- 神经网络的层数增多效果却不尽如人意
- 计算机能力有限

字长：32位;
晶体管数目：9.5M
峰值主频： 450 MHz

复兴时期 (2006至今)



字长：32位;
晶体管数目：120G
峰值主频：1531 MHz

Y. Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition.1989.

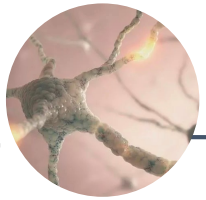
J. Schmidhuber. Deep Learning in Neural Networks: An Overview. arxiv, 2014.

G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. Science, 2006.

发展历程 --- 里程碑

时间线

1943
第一个神经元模型



1941
第一台电子计算机

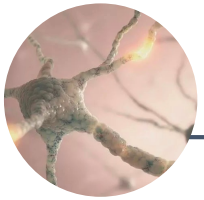
1950 图灵测试

“人工智能之父”

发展历程 --- 里程碑

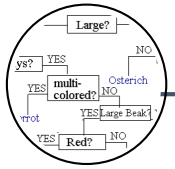
时间线

1943
第一个神经元模型



1941
第一台电子计算机

1970 专家系统



1950 图灵测试

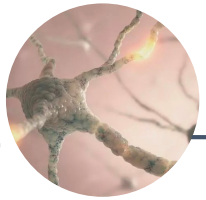
“人工智能之父”

1997 深蓝击
败卡斯巴罗夫

发展历程 --- 里程碑

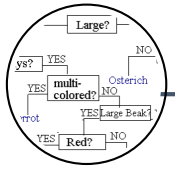
时间线

1943
第一个神经元模型



1941
第一台电子计算机

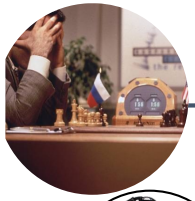
1970 专家系统



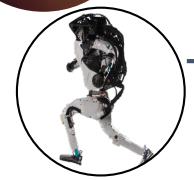
1950 图灵测试

“人工智能之父”

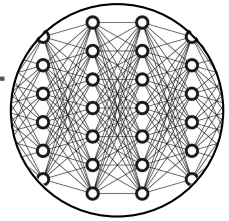
1997 深蓝击
败卡斯巴罗夫

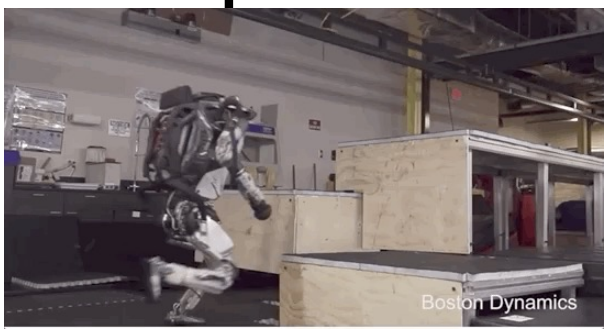


2013 波士顿动力



2006 深度学习概念

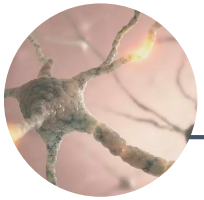




发展历程 --- 里程碑

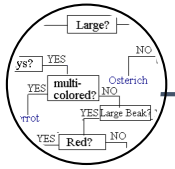
时间线

1943
第一个神经元模型



1941
第一台电子计算机

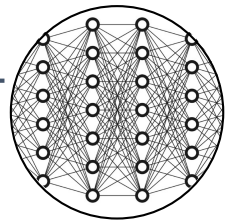
1970 专家系统



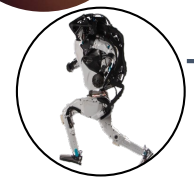
1950 图灵测试

1997 深蓝击
败卡斯巴罗夫

2006 深度学习概念



2013 波士顿动力

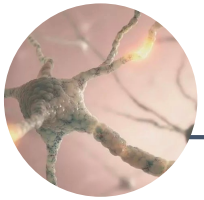


2016
Alpha Go 击败李世
石

发展历程 --- 里程碑

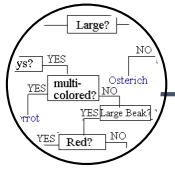
时间线

1943
第一个神经元模型



1941
第一台电子计算机

1970 专家系统

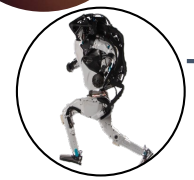


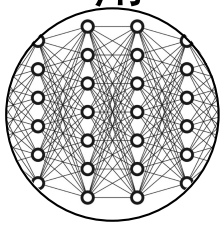
1950 图灵测试

1997 深蓝击败卡斯巴罗夫

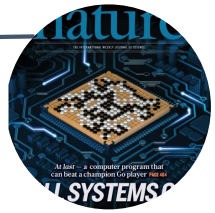
1990 人工智能在军队应用

2013 波士顿动力





2006 深度学习概念

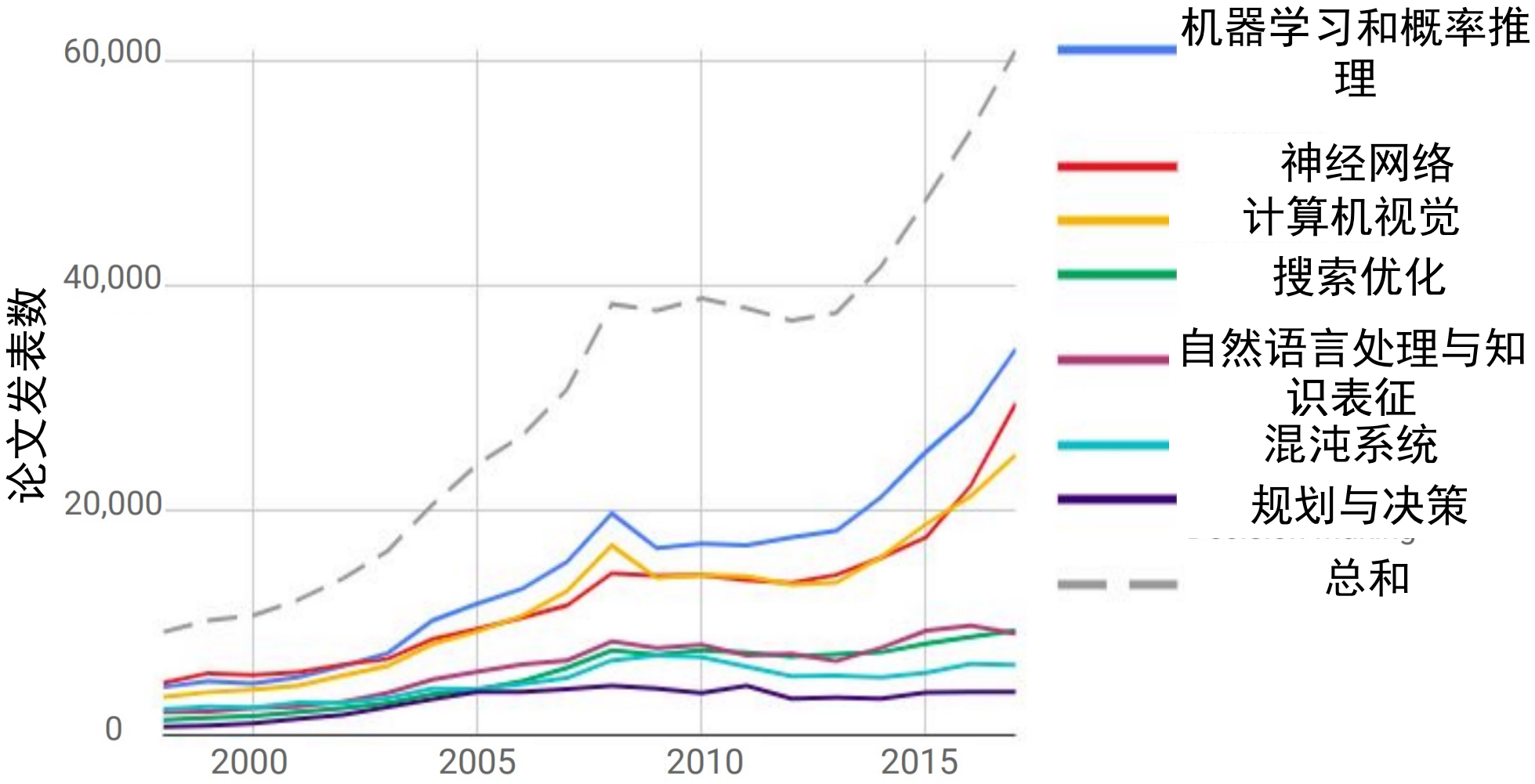


2023 Chatgpt横空出世

2016
Alpha Go 击败李世石

发展历程 --- 学术文章

人工智能论文发表数 （Scopus）



Source of image: http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf

发展历程 ---科技公司

Baidu 百度

Duer | AR Lab | IDL | Apollo | BIL .....


王海峰


Yuanqing Lin


Xu Wei


Jun Huan


Hui Xiong

Tencent 腾讯

AI Lab | Youtu Lab | WeChat AI .....


张潼


Dong Yu


Han Liu


贾佳亚

Alibaba.com

AI Lab | iDST | Alibaba Cloud | Ali Pay .....


王刚


漆远


金榕


任小枫


华先胜


贾扬清

JD.COM 京东

JD Finance, AI, and Big Data


周伯文

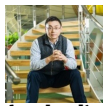

薄列峰


裴建


周志华

360 www.360.com

AI Institute .....


颜水成

Google

Google Brain | AI Lab | Tensorflow | TPU

Google Image | Google Translate .....


Feifei Li


Geoffrey Hinton

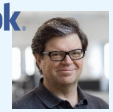

Ian Goodfellow


Samy Bengio


Dave Patterson

facebook

Facebook AI Research (FAIR) | Caffe | Torch .....


Yann LeCun


Jerome Pesenti


Kaiming He

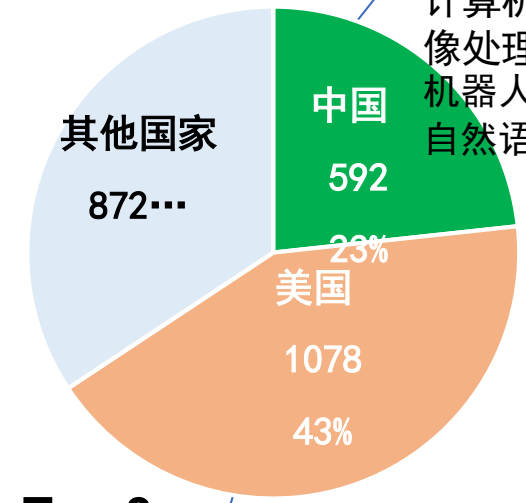
amazon

Amazon Go | Amazon Alexa | Amazon AI .....


Ganapathy (Krish) Krishnan


Ranju Das

发展历程---创业公司

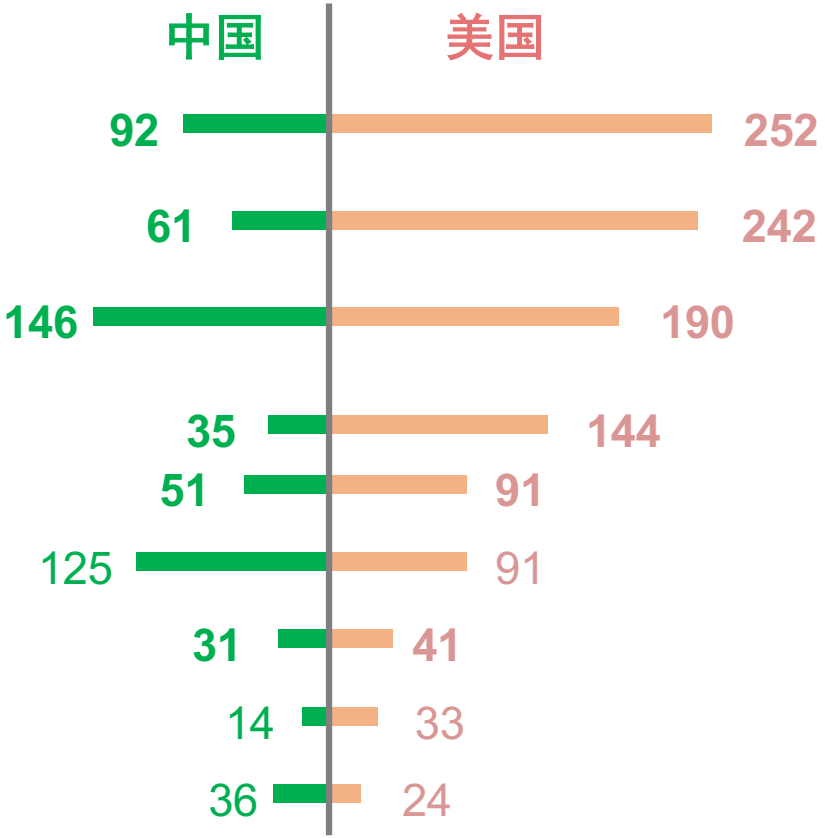


前三

计算机视觉及图像处理： 146
机器人：125
自然语言处理: 92

Top 3

自然语言处理: 252
机器学习应用: 242
计算机视觉及图像处理： 190



2017年6月前的数据

Source: http://www.cbdio.com/BigData/2017-08/14/content_5576175.htm

人社部拟发布“人工智能训练师”新职业

2020-01-02 16:32　来源：OFweek人工智能网

近日，中国就业培训技术指导中心发布《关于拟发布新职业信息公示的通告》，人工智能训练师以及智能制造和工业互联网领域的16个新职业已评审论证，经人社部同意进入公示阶段。

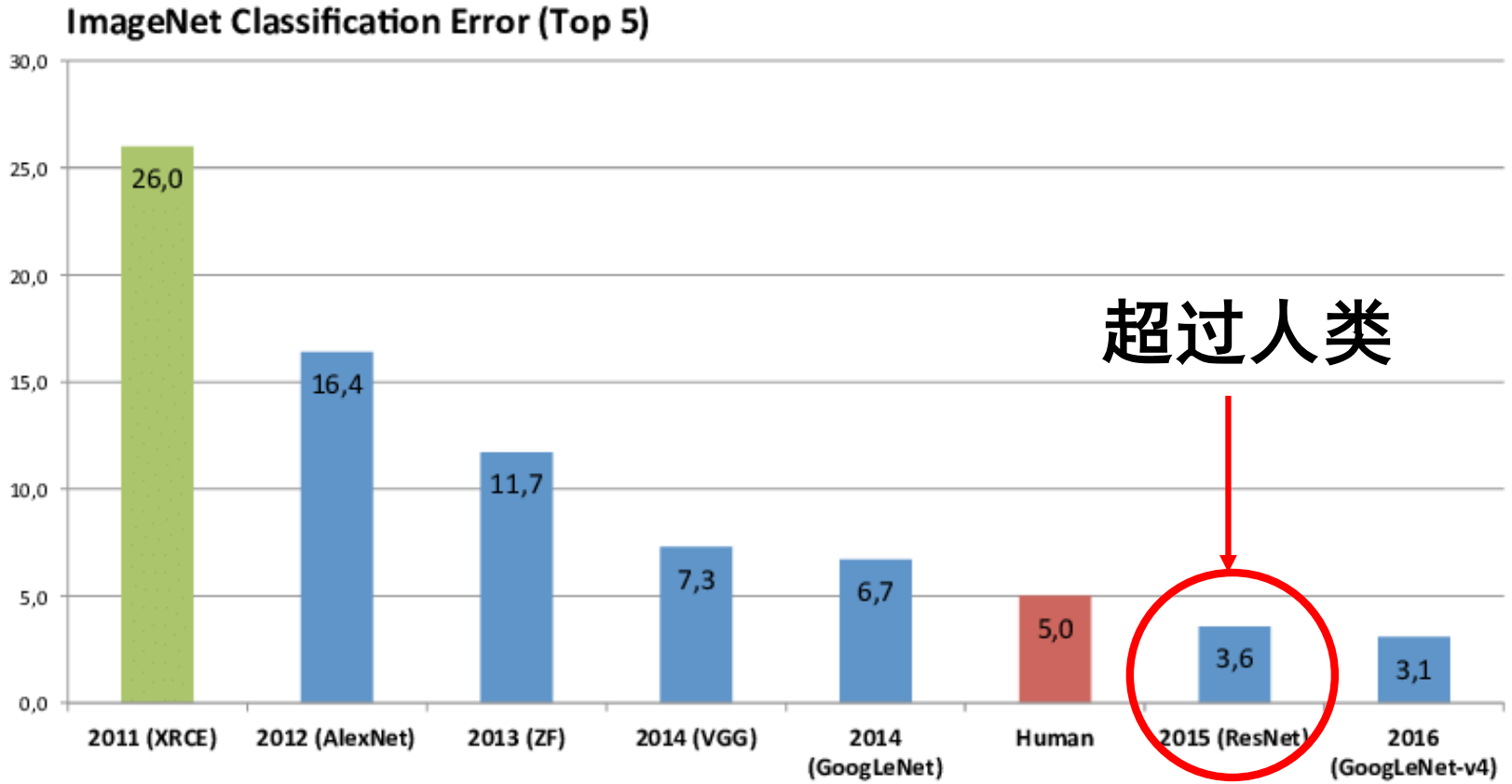
公示内容显示，人工智能训练师新职业属于软件和信息技术服务业，主要工作任务包括：标注、加工原始数据、分析提炼专业领域特征，训练和评测人工智能产品相关的算法、功能和性能，设计交互流程和应用解决方案，监控分析管理产品应用数据、调整优化参数配置等。

一般而言，AI公司从客户（用户）那里获取到的原始数据无法直接用于模型训练，在“人工智能训练师”出现以前，是由AI产品经理先用相关工具简单处理，再交给数据标注人员进行标注加工，但因为标注人员对数据的理解和标注质量差异很大，导致整体标注工作的效率和效果都不够理想。

主要内容

- 课程介绍
- 深度神经网络
  - 发展历程
  - 应用
  - 类别
  - 重要指标
  - 软件框架
  - 硬件平台

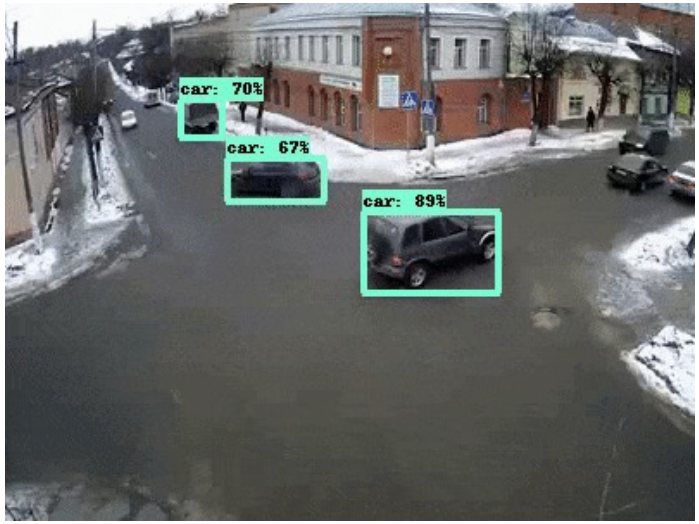
应用一：图像任务



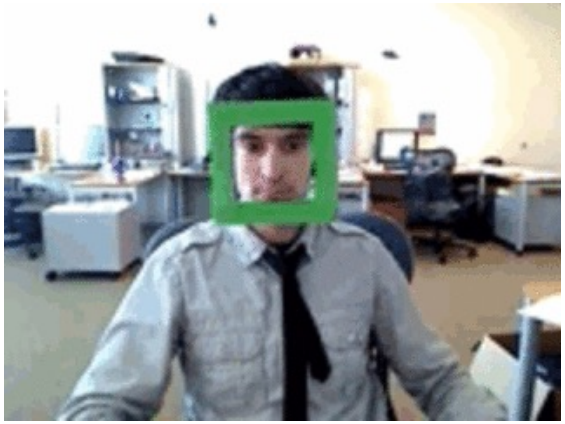
这是哪种乌龟？

1. 棱皮龟
2. 蠵（Xi）龟
3. 肯氏龟
4. 太平洋丽龟

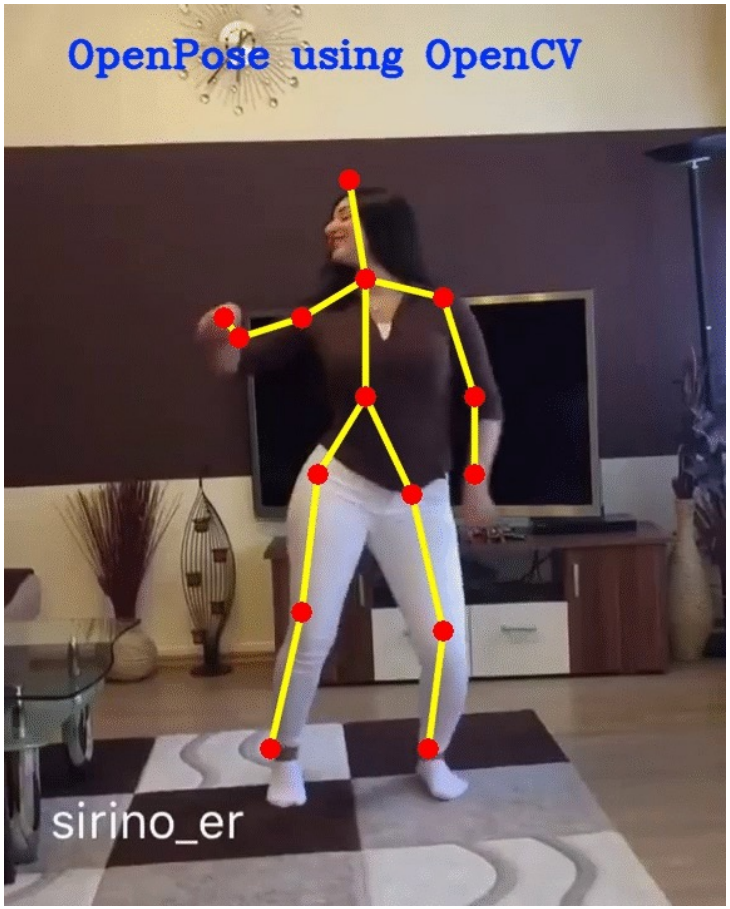
应用一：视频



物体检测

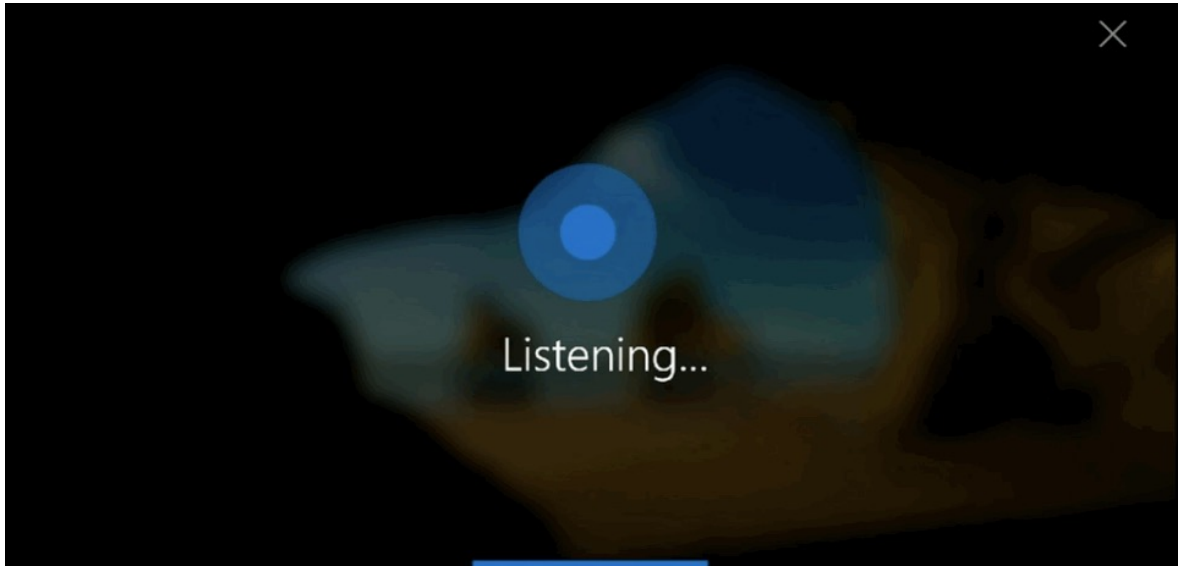


~~实时面部追踪~~

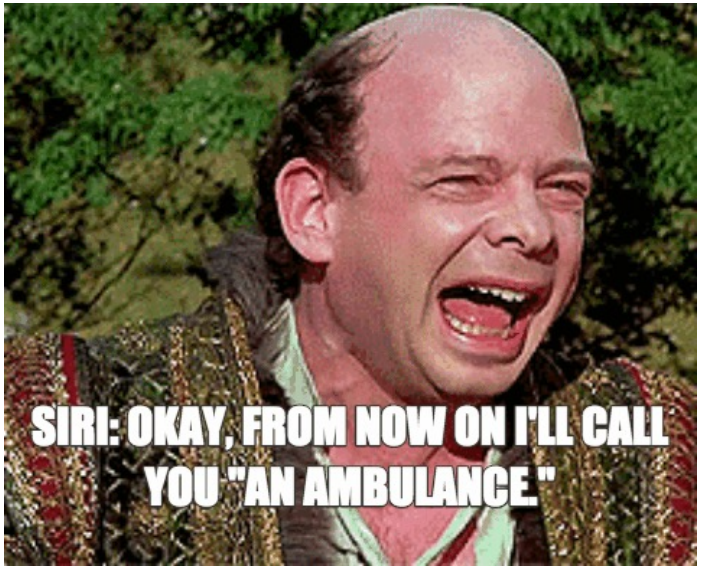


人体姿势估计

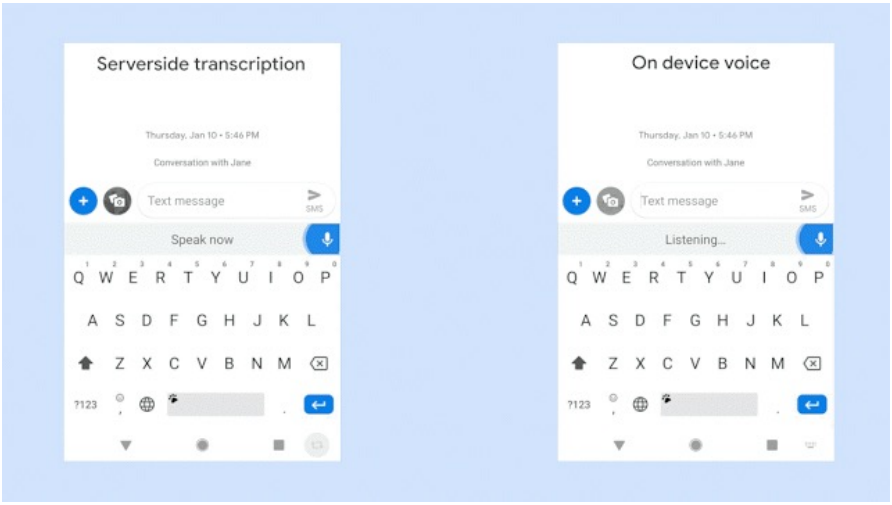
应用三：语音



Cortana



“Siri, 给我叫救护车”

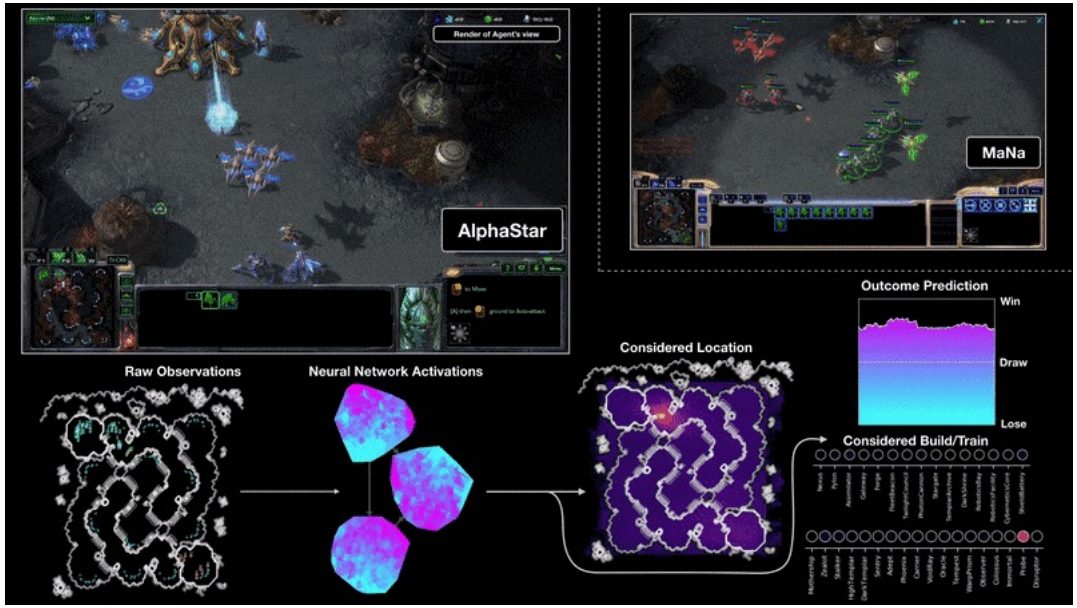


语音转换文字

“声音比打字快.”

Source of Gifs: http://www.impactlab.net/2016/10/26/microsofts-speech-recognition-is-now-as-accurate-as-a-humans/ ; https://www.nextbigfuture.com/2019/04/google-has-80-mb-speech-recognizer-that-can-work-offline-on-your-smartphone.html ; https://media.giphy.com/media/LY1oUG96n20h2/giphy.gif

应用四：游戏



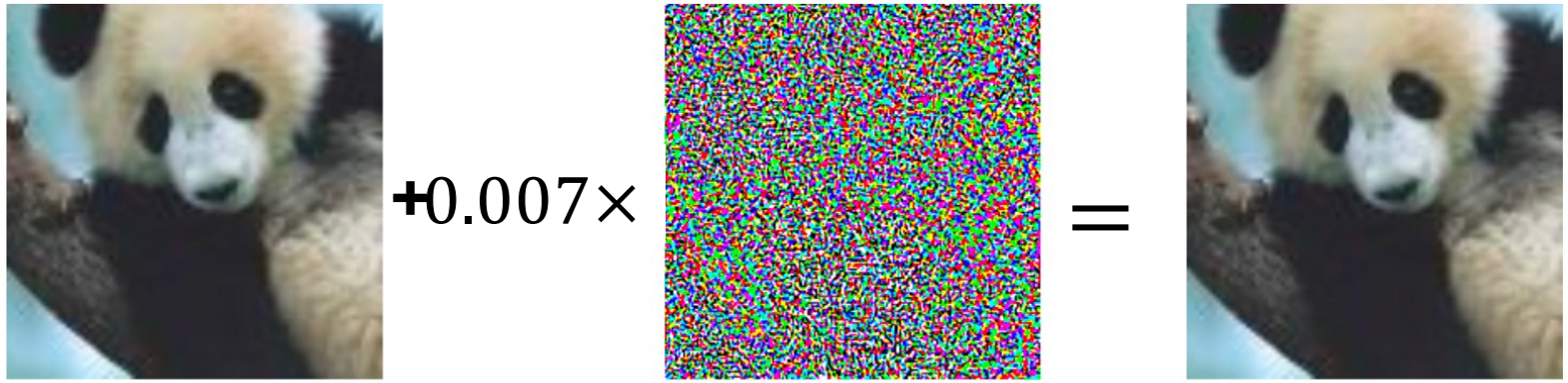
AlphaStar：星际争霸

Alpha Go

Source of Gifs: https://www.techspot.com/news/78431-human-player-finally-beat-deepmind-alphastar-ai-starcraft.html; https://imgur.com/gallery/aPw2D;

人工智~~能~~ or 人工智障?

- 非常脆弱…



57. 7% 置信度

99. 3% 置信度

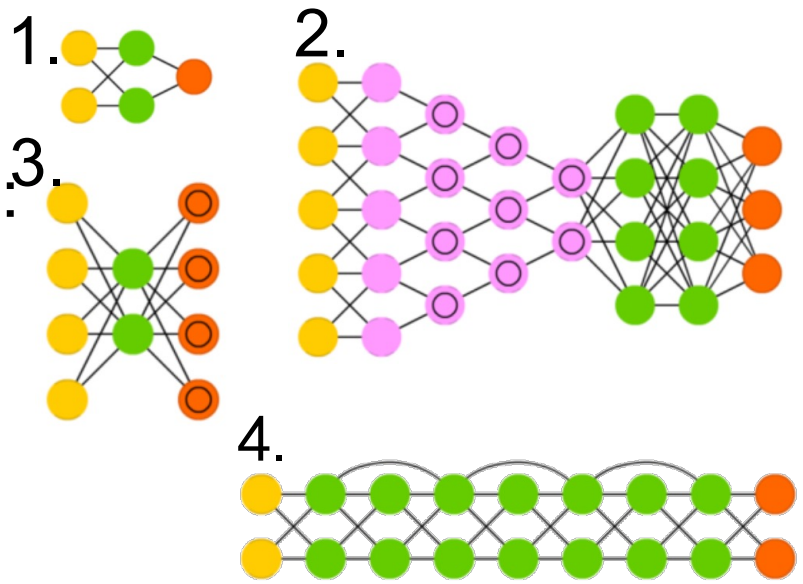
长臂猿

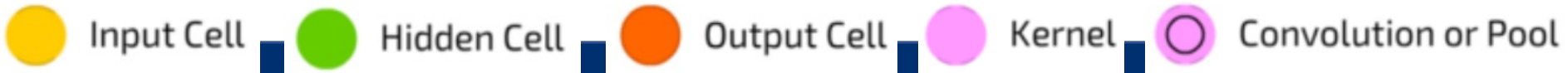
Outline

- 课程介绍
- 机器学习&深度神经网络
  - 应用
  - 类别
  - 一些重要指标
  - 软件框架 & 硬件平台

类别-网络结构

- 前馈神经网络(FeedForward):
  1. 多层感知机
  2. 卷积神经网络
  3. 自动编码器
  4. 深度残差网络

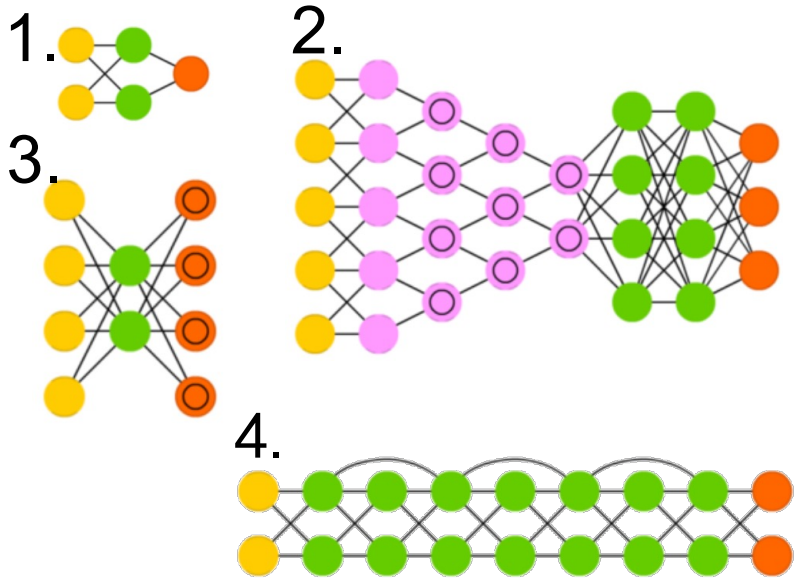


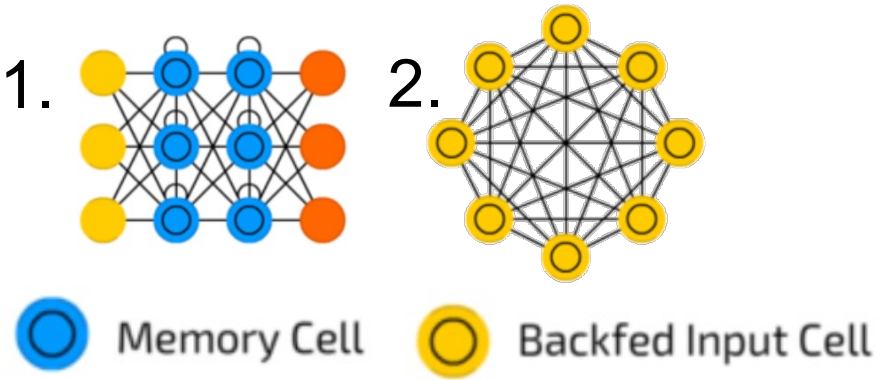


Source of image: https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

类别-网络结构

- 前馈神经网络(FeedForward)
  1. 多层感知机
  2. 卷积神经网络
  3. 自动编码器
  4. 深度残差网络
- 循环神经网络:
  1. 长短期记忆网络LSTM
  2. Hopfield网络
  3. …



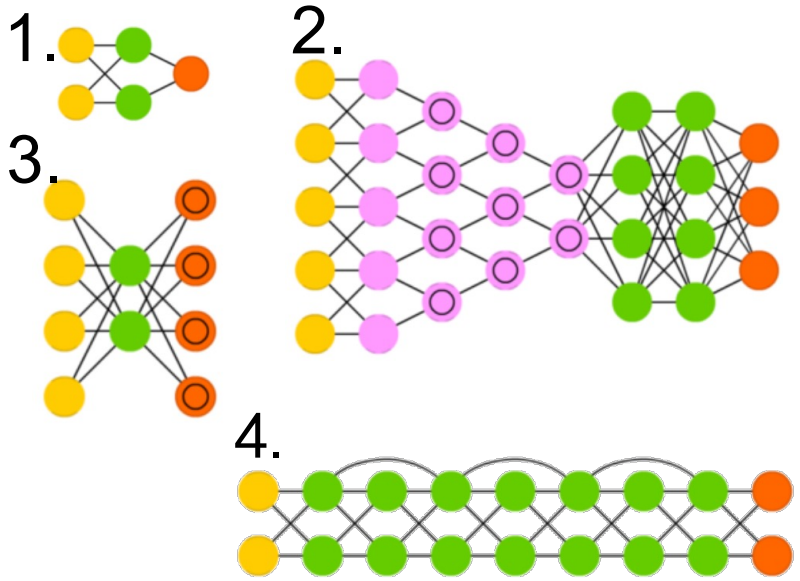


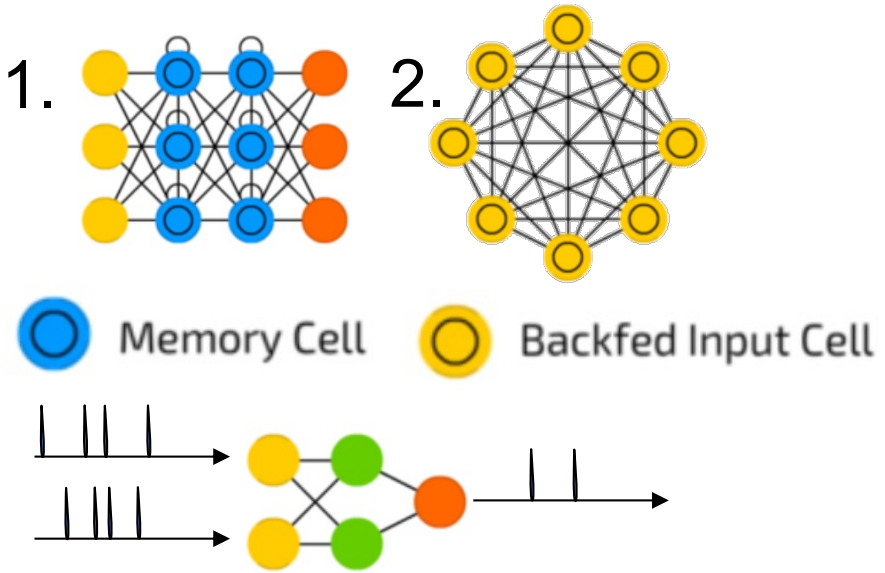
Input Cell Hidden Cell Output Cell Kernel Convolution or Pool

Source of image: https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

类别-网络结构

- 前馈神经网络(FeedForward)
  1. 多层感知机
  2. 卷积神经网络
  3. 自动编码器
  4. 深度残差网络
- 循环神经网络:
  1. 长短期记忆网络LSTM
  2. Hopfield网络
  3. …
- 脉冲神经网络（SNN）





Input Cell Hidden Cell Output Cell Kernel Convolution or Pool

Source of image: https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

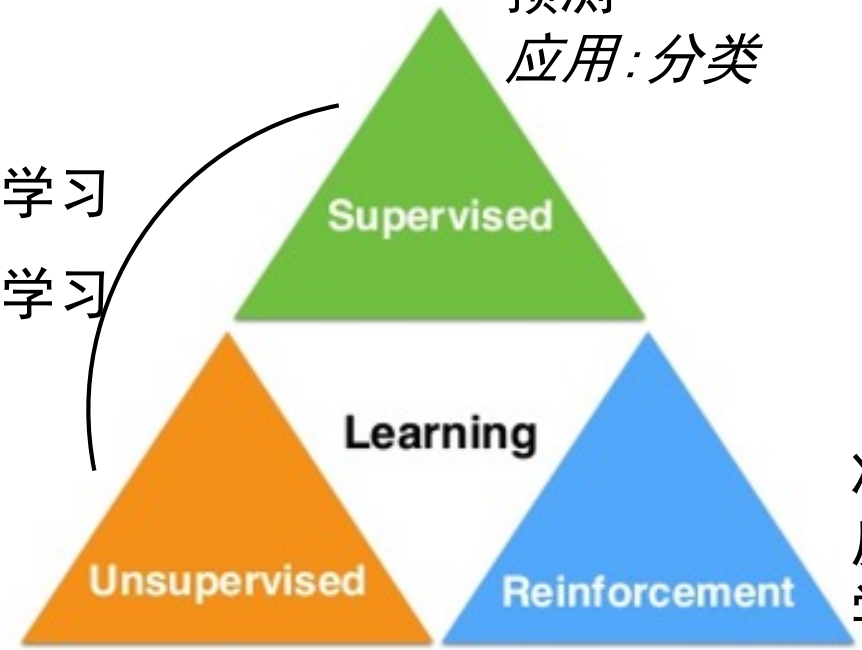
类别-学习模式

- 监督学习
- 半监督学习
- 无监督学习
- 强化学习

带标记的数据
直接反馈
预测
应用：分类

半监督学习

弱监督学习



数据无标记
结果无反馈
找到隐含特征
应用：重建

决策过程
反馈系统
学习一系列行为
应用：决策

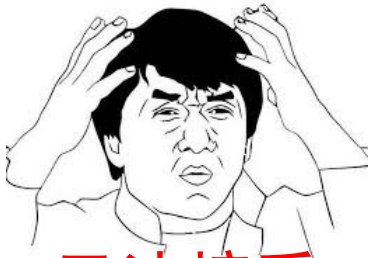
以后课程
详细介绍

Outline

- 课程介绍
- 机器学习&深度神经网络
  - 应用
  - 类别
  - 重要的评价标准(**LASER，激光**)
    - 延迟（**L**atency）
    - 准确度（**A**ccuracy）
    - 模型大小（**S**ize of model）
    - 能效（**E**nergy efficiency）
    - 鲁棒性 （**R**obustness）
  - 软件框架 & 硬件平台

延迟 Latency

- 从输入数据到在输出端得到结果花费的时间
- 准确度越高?延迟越大!
- VGG-16 在手机上处理一张图片需要 ~3秒.



无法接受

网络层数越来越多
网络越来越深

识别错误率

16.4% AlexNet (2012)

7.3% VGG (2014)

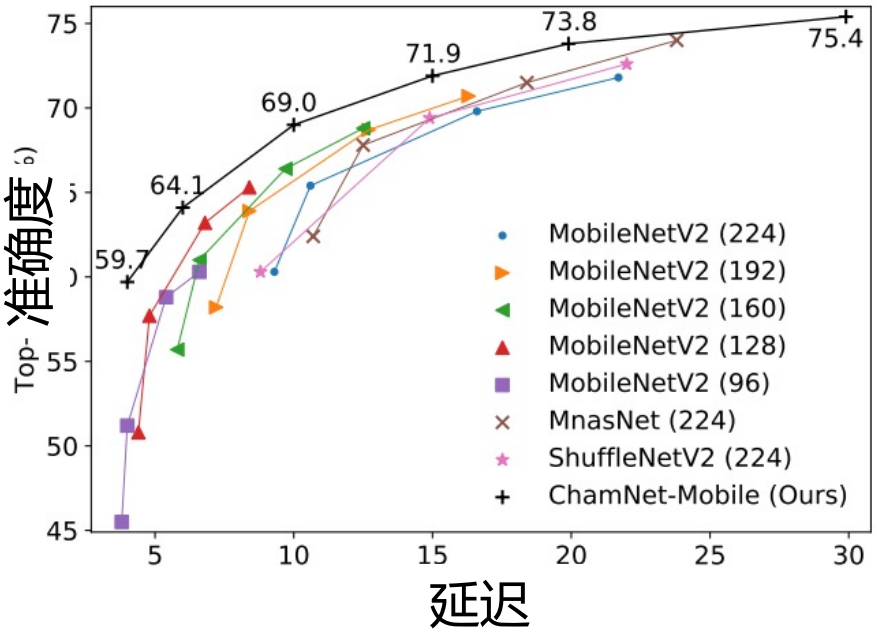
6.7% GoogleNet (2014)

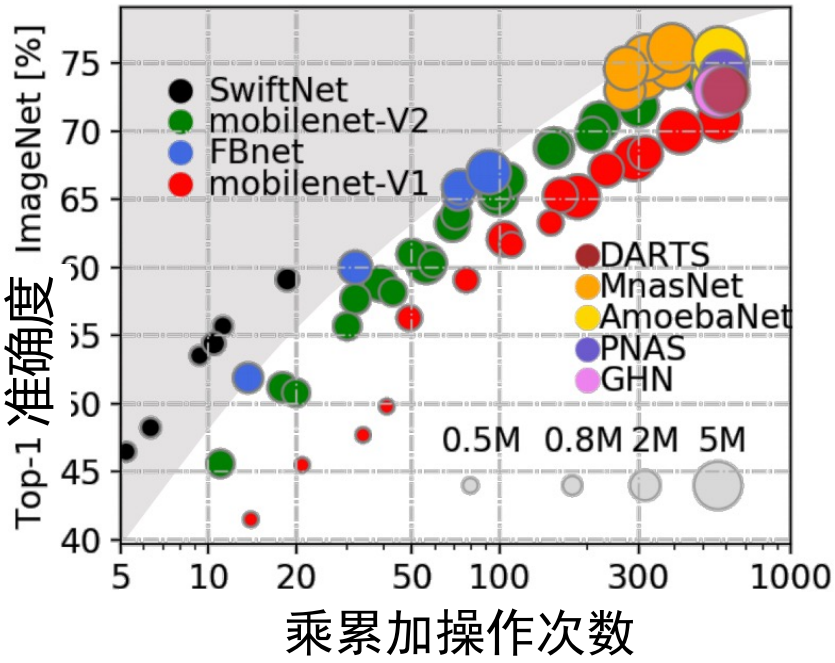
3.57% Residual Net (2015)

152 层

准确率 Accuracy

- 对分类问题的评价标准
- 我们通常会说“Top-*K* Accuracy”
- 准确率越高越好, 但是需要相应的付出。
  - 任何事情都是折衷.





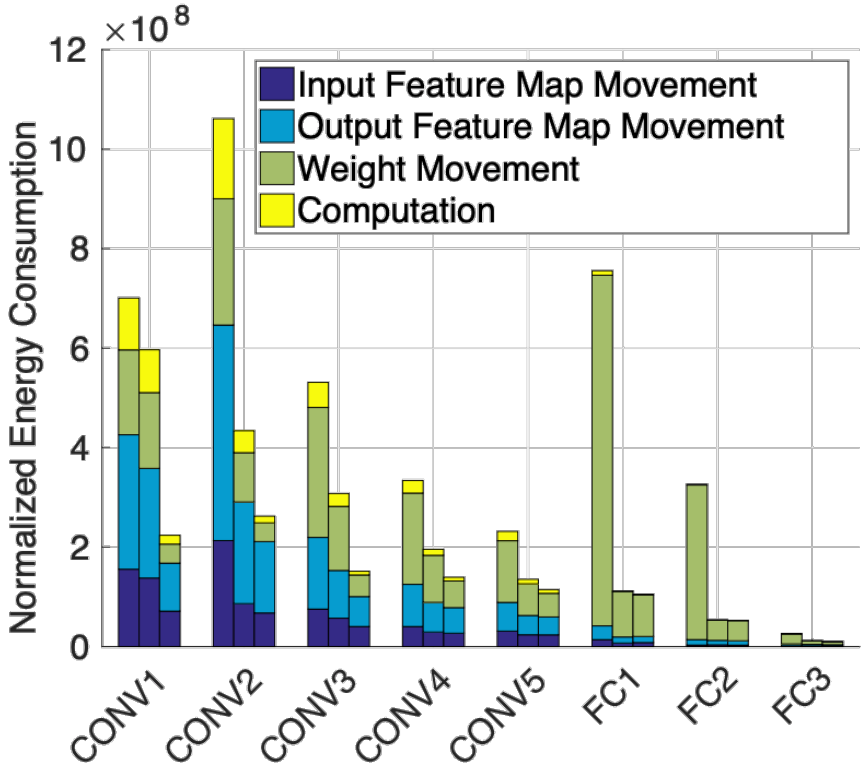
Source:
1. Dai, Xiaoliang, et al. "Chamnet: Towards efficient network design through platform-aware model adaptation.” (2019)
2. Cheng, Hsin-Pai et al. “SwiftNet: Using Graph Propagation as Meta-knowledge to Search Highly Representative Neural Architectures” (2019)

模型规模 Size of model

- 参数个数
- 浮点操作次数（ floating point operations, FLOPS）
- 乘累加操作次数（ multiply-and-accumulate , MAC）
  - 通常 1次浮点乘累加操作约等于2次FLOPS

能效 Energy efficiency

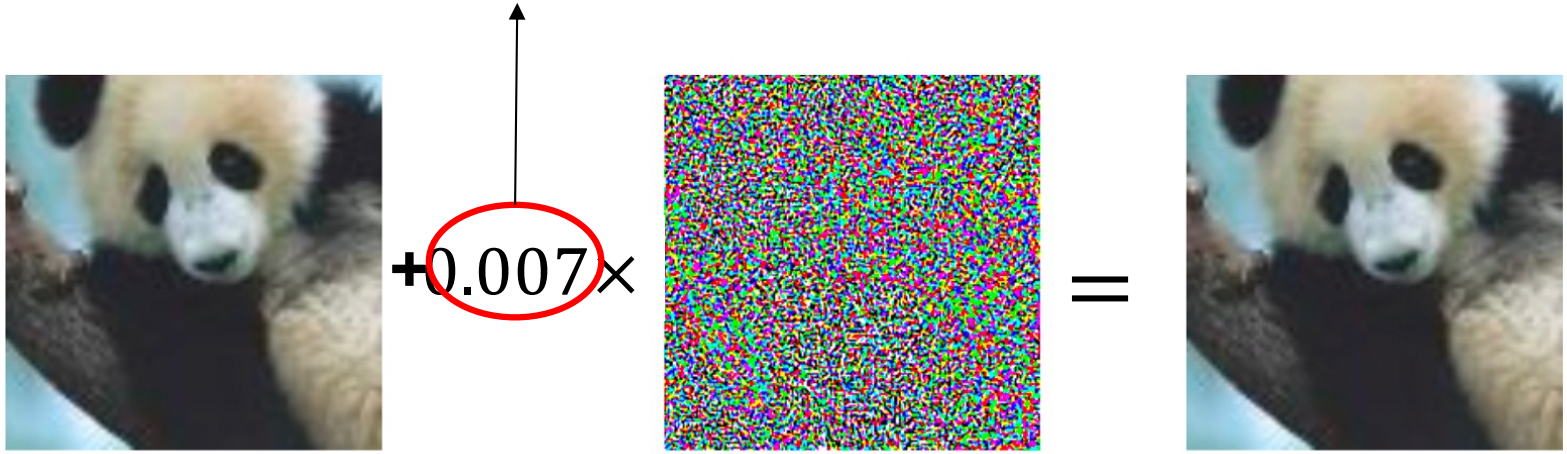
- 功耗 [mW]
- 能耗用于评估以下开销
  - 计算
  - 数据传输
- 能耗不同:
  - 网络的MAC数少 不一定 会有低能耗。
  - 卷积网络层的计算比全连接层的计算层消耗更多的能量。
  - 对卷积网络来说，层数深，参数少并不一定比层数少，参数多的网络能耗低。



Source of image: http://eyeriss.mit.edu/

鲁棒性 Robustness

- 这个参数, 用来评估网络的鲁棒性.



57.7% 置信度　　99.3% 置信度

- 通常, 准确率高的模型并不鲁棒.
- 模型规模和模型结构, 模型结果与模型的鲁棒性关系更紧密.

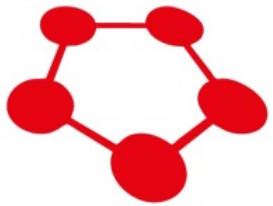
任何事情都是折衷

Source of image: Explaining and harnessing adversarial examples

Outline

- 课程介绍
- 机器学习&深度神经网络
  - 应用
  - 类别
  - 一些重要评价指标 (**LASER，激光**)
  - 软件开发框架 & 硬件平台
    - 软件开发框架
    - 硬件计算平台

软件框架



Chainer

PyTorch

TensorFlow

mxnet

Microsoft CNTK

Caffe2

开发框架对比

https://en.wikipedia.org/wiki/Comparison_of_deep-learning_software

	PyTorch	mxnet	Chainer	TensorFlow
支持的编程接口	C++, Python, Java, Rust, Go	C++, Python, Scala, Julia, Perl, MATLAB	C++, Python	C++, Python, Go, Java, Swift, JavaScript
是否支持多GPU 并行	是	是	是	是
开发商	facebook AI research group	THE APACHE SOFTWARE FOUNDATION	Preferred Networks, NVIDIA, intel, IBM	Google Brain team

S. Bahrampour, et al. Comparative Study of Deep Learning Software Frameworks. *arxiv*, 2016.

Tensorflow 优缺点

优点

- TensorBoard 可视化
- 数据和模型并行化；
- 大公司喜欢

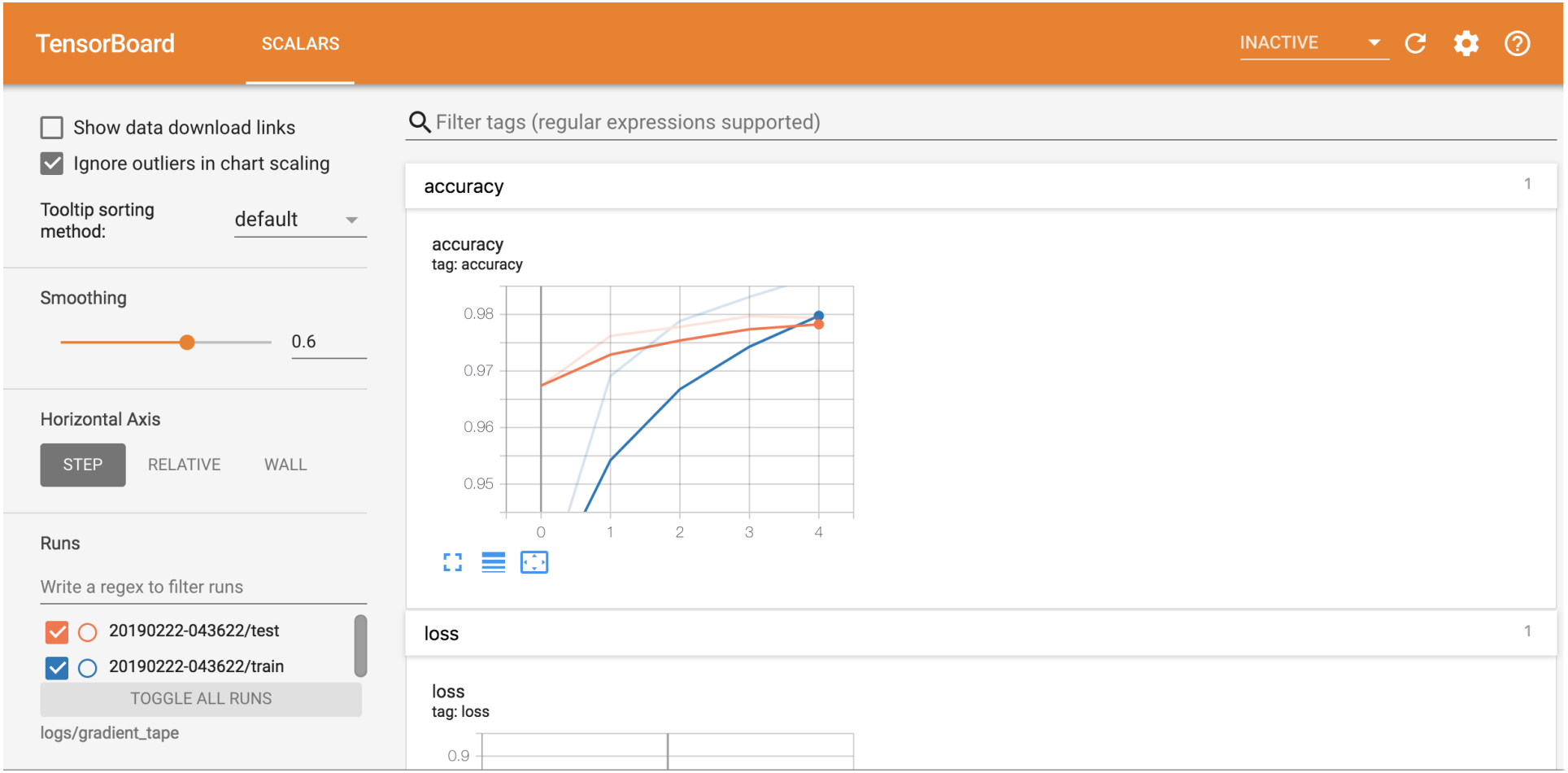
缺点：

- 运行速度比不上其他的框架
- 学习曲线陡峭

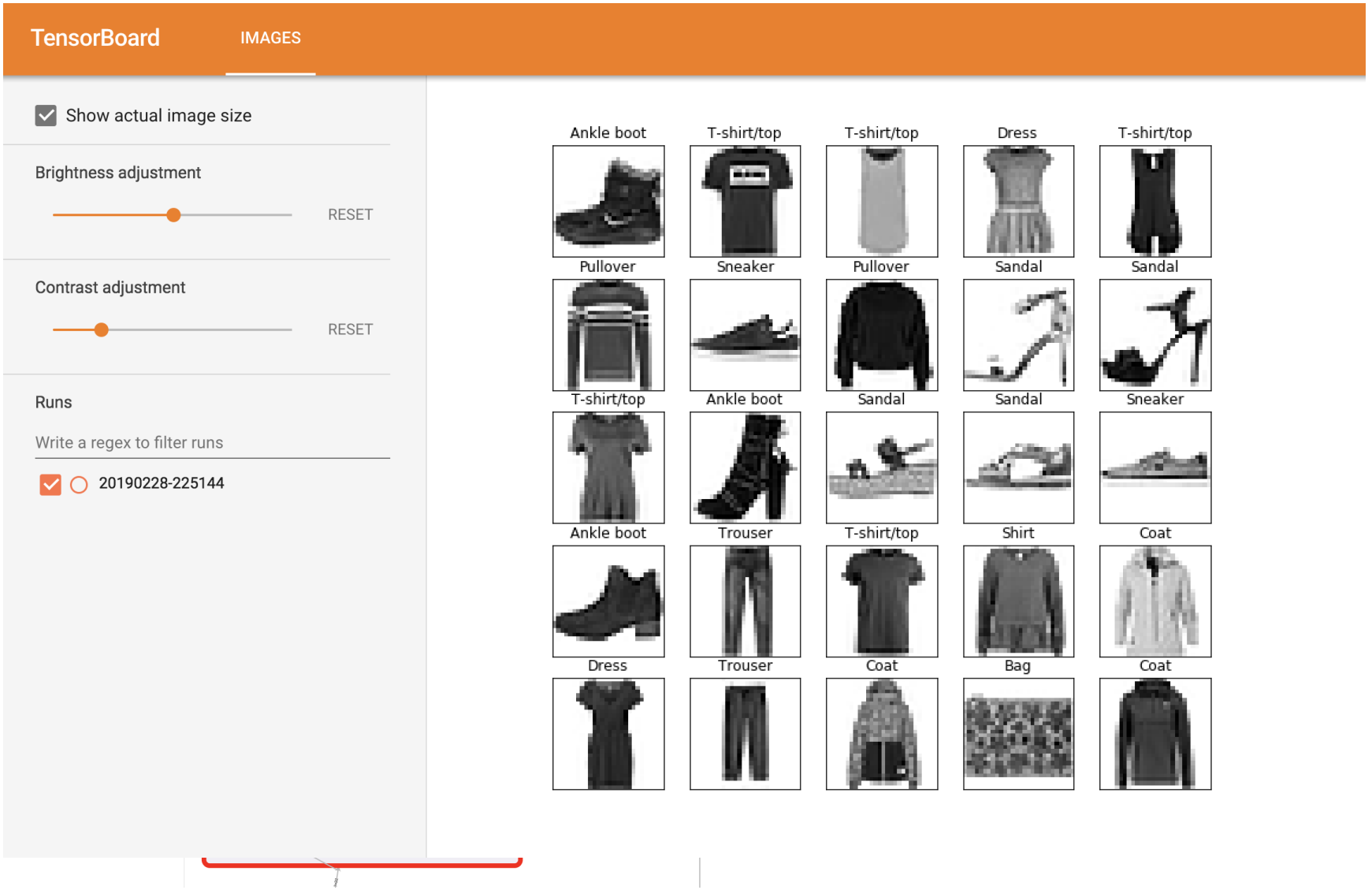
TensorFlow 操作举例.

Category	Examples
Element-wise mathematical operations	Add, Sub, Mul, Div, Exp, Log, Greater, Less, Equal, ...
Array operations	Concat, Slice, Split, Constant, Rank, Shape, Shuffle, ...
Matrix operations	MatMul, MatrixInverse, MatrixDeterminant, ...
Stateful operations	Variable, Assign, AssignAdd, ...
Neural-net building blocks	SoftMax, Sigmoid, ReLU, Convolution2D, MaxPool, ...
Checkpointing operations	Save, Restore
Queue and synchronization operations	Enqueue, Dequeue, MutexAcquire, MutexRelease, ...
Control flow operations	Merge, Switch, Enter, Leave, NextIteration

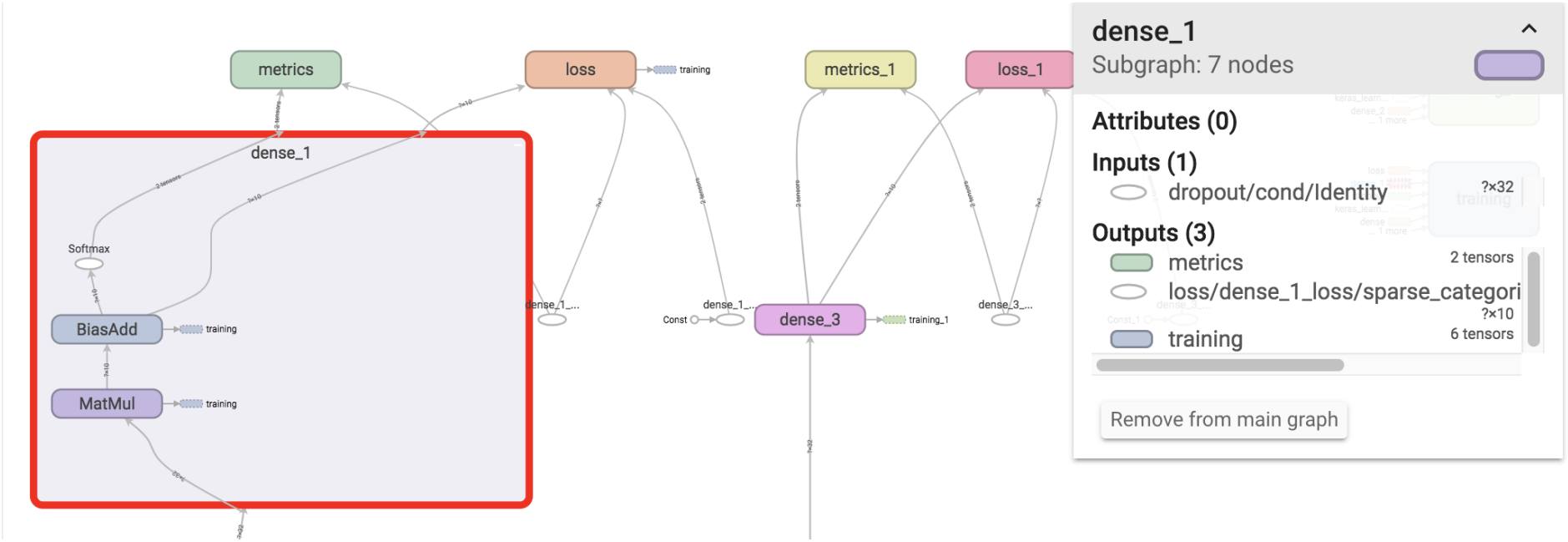
TensorBoard 展示



TensorBoard 展示



TensorBoard 展示



Pytorch 优缺点

优点

- 动态计算图的编程思想；
- 相比其他DNN工具库，速度更快；
- 用法方便

缺点：

- 相比Tensorflow,开发者人数少
- 研究者用的多；产品人员用得少

Pytorch vs. Tensorflow: 代码量对比

```python
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.conv1 = nn.Conv2d(1, 20, 5, 1)
        self.conv2 = nn.Conv2d(20, 50, 5, 1)
        self.fc1 = nn.Linear(4*4*50, 500)
        self.fc2 = nn.Linear(500, 10)

    def forward(self, x):
        x = F.relu(self.conv1(x))
        x = F.max_pool2d(x, 2, 2)
        x = F.relu(self.conv2(x))
        x = F.max_pool2d(x, 2, 2)
        x = x.view(-1, 4*4*50)
        x = F.relu(self.fc1(x))
        x = self.fc2(x)
        return F.log_softmax(x, dim=1)
```

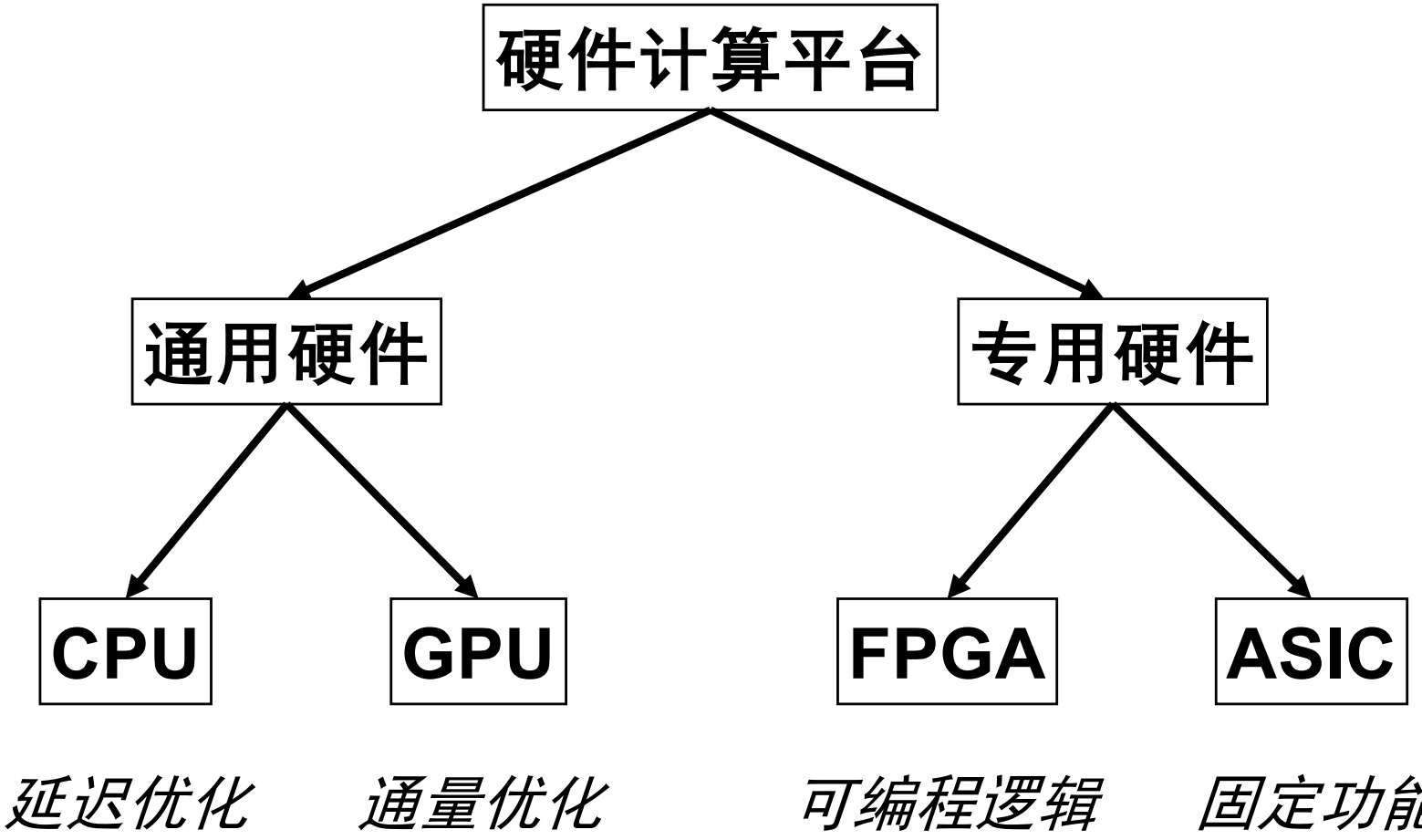
https://github.com/pytorch/examples/blob/master/mnist/main.py

```python
def cnn_model_fn(features, labels, mode):
  """Model function for CNN."""
  input_layer = tf.reshape(features["x"], [-1, 28, 28, 1])
  conv1 = tf.layers.conv2d(
      inputs=input_layer,
      filters=32,
      kernel_size=[5, 5],
      padding="same",
      activation=tf.nn.relu)
  pool1 = tf.layers.max_pooling2d(inputs=conv1, pool_size=[2, 2], strides=2)
  conv2 = tf.layers.conv2d(
      inputs=pool1,
      filters=64,
      kernel_size=[5, 5],
      padding="same",
      activation=tf.nn.relu)
  pool2 = tf.layers.max_pooling2d(inputs=conv2, pool_size=[2, 2], strides=2)
  pool2_flat = tf.reshape(pool2, [-1, 7 * 7 * 64])
  dense = tf.layers.dense(inputs=pool2_flat, units=1024, activation=tf.nn.relu)
  dropout = tf.layers.dropout(
      inputs=dense, rate=0.4, training=mode == tf.estimator.ModeKeys.TRAIN)
  logits = tf.layers.dense(inputs=dropout, units=10)

  predictions = {
      # Generate predictions (for PREDICT and EVAL mode)
      "classes": tf.argmax(input=logits, axis=1),
      # Add `softmax_tensor` to the graph. It is used for PREDICT and by the
      # `logging_hook`.
      "probabilities": tf.nn.softmax(logits, name="softmax_tensor")
  }
```

https://github.com/tensorflow/tensorflow/blob/master/tensorflow/examples/tutorials/layers/cnn_mnist.py

计算平台



硬件计算平台

价格和速度是两个非常重要的考量。

GPU

- 功耗高
- 成本高
  - 数据中心，高冷却成本；电费和占地面积

CPU

- 成本适中
- 功耗适中
- 速度慢

ASIC

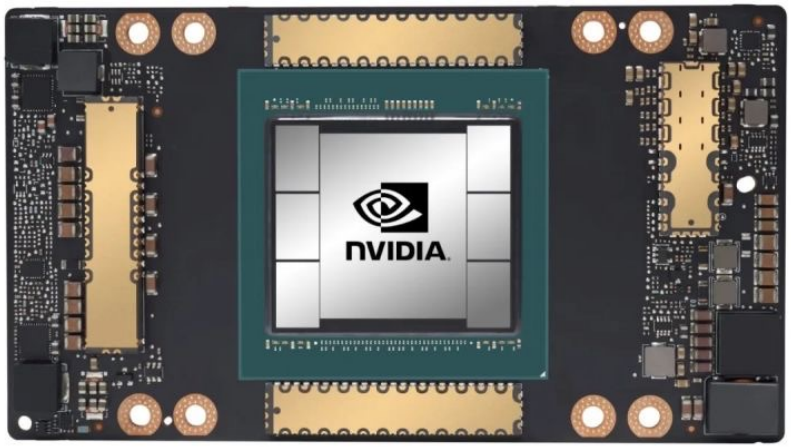
- 不可重复性工程
- 设计及开发周期长
- 不适合NN的快速迭代开发

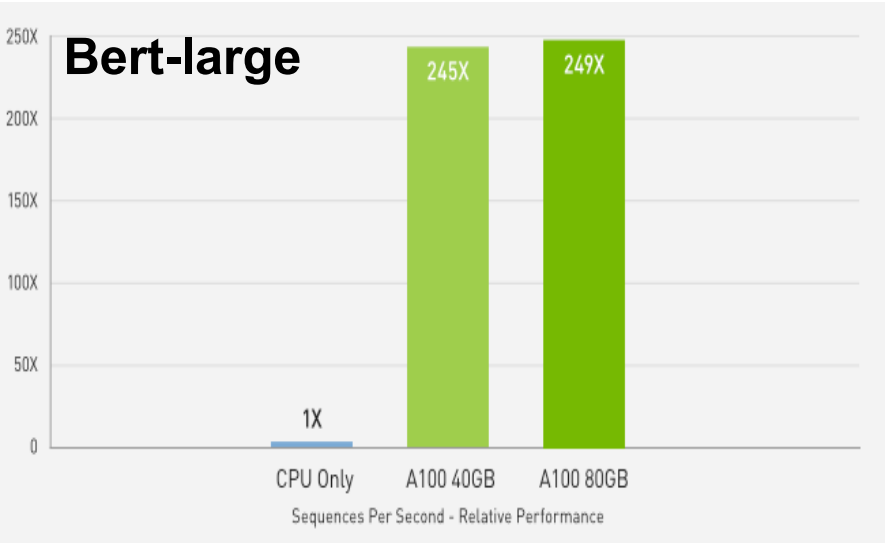
FPGA

- 功耗低
- 成本低, 几百到几千元
- 功能可重配置

GPU --- 英伟达A100

- 2020年5月推出，
- NVIDIA Ampere 架构
- 针对 AI、数据分析和 HPC 应用场景
- 相比CPU，>200x加速
- 相比上一代V100， 20x加速
- 2022年9月，对中国限购
- 核心面积826平方毫米，具有540亿个晶体管
- 助力ChatGPT





(Google, https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html)

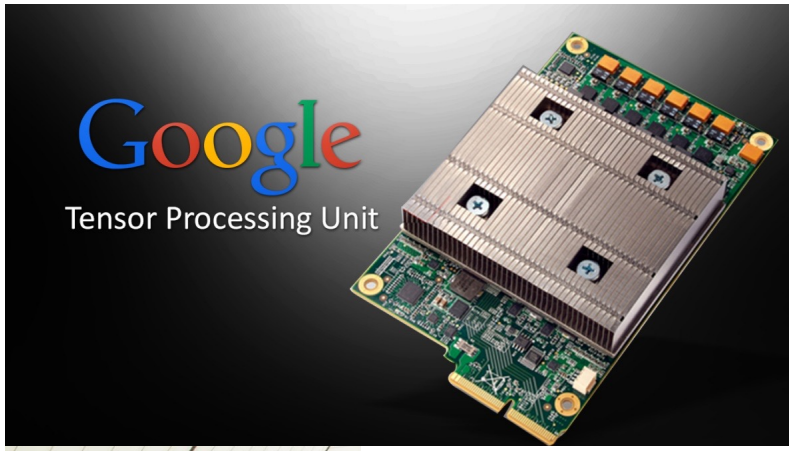
ImageNet-1K 分类任务测评

计算平台	推断任务吞吐量	TFLOPs (峰值)	TFLOPs (有效)	功耗	能效 (GOPs/J)
Intel Xeon E5-2450 **(CPU)**	53 张/秒	0.27T	0.074T (27%)	~225W	~0.3
Intel Altera Arria 10 GX1150 **(FPGA)**	369 张/秒	1.366T	0.51T (38%)	**~40W**	**~12.8**
NVIDIA Titan X **(GPU)**	4129 张/秒	6.1T	5.75T (94%)	~250W	~23.0

神经网络训练在云端GPU集群，执行在用户端设备；
FPGA适合执行低功耗推断任务

ASIC --- 谷歌 TPU

2016 谷歌开发者大会推出；
ASIC, 直接运行Tensorflow；
稠密矩阵计算加速。



应用:

1. RankBrain: 改善搜索结果.
2. 街景Street View: 改进谷歌导航.
3. AlphaGo: “想”更快，“考虑”更长远.

运行战胜李世石的AlphaGo 的TPU机器机架

(Google, https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html)

国内芯片公司

平头哥

•落地场景：
-阿里云：含光800，解决图像、视频识别、云计算等商业场景的AI推理运算问题。
-物联网：玄铁910，RISC-V 架构处理器

寒武纪

•落地场景：
-中科曙光：AI推理专用服务器Phaneron
-华为：麒麟970内人工处理器技术， Mate 手机

HISILICON

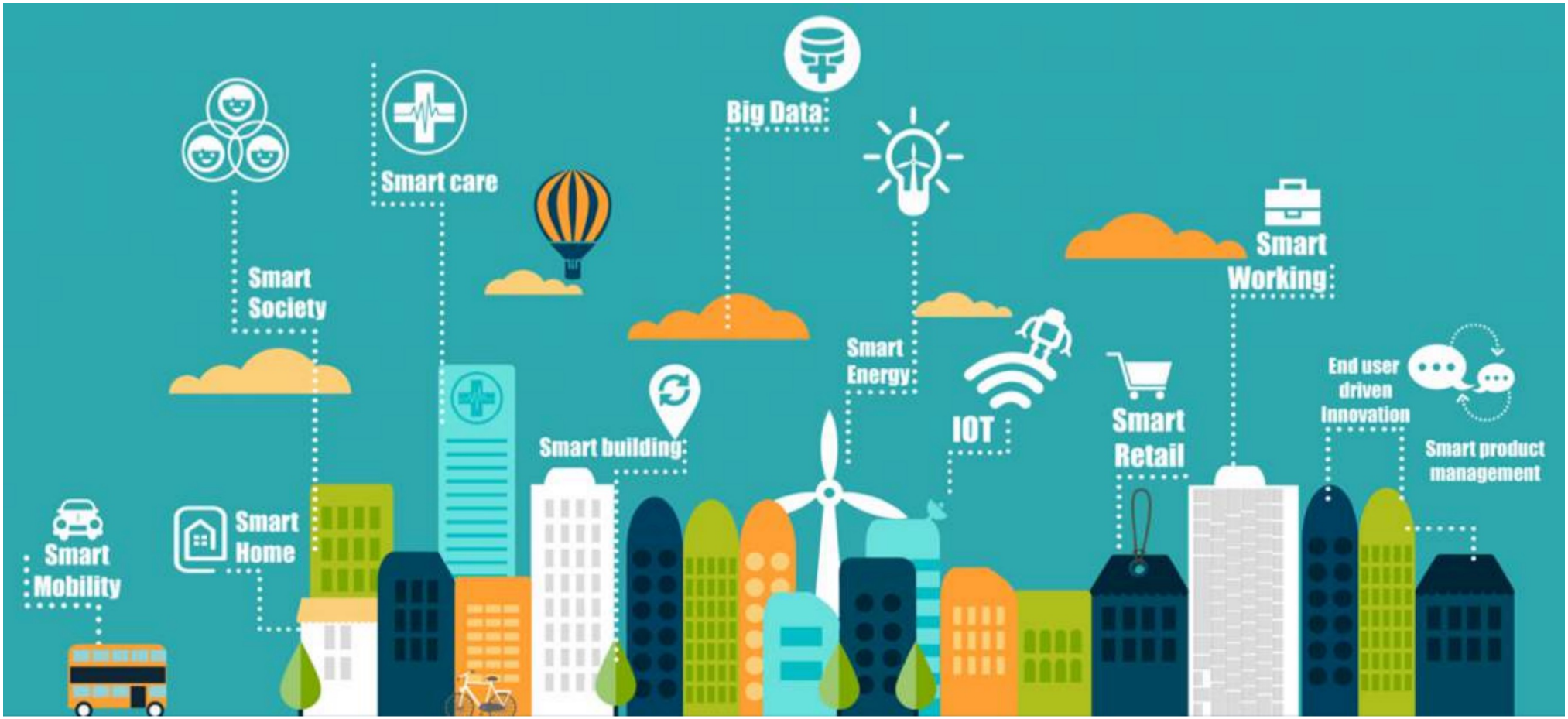
地平线 Horizon Robotics

天数智芯 Iluvatar CoreX

登临科技 DENGLIN

摩尔线程 MOORE THREADS

展望



智能　　更快　　更隐私　　更便携　　高能效

Backup Slides