大数据和深度学习介绍

张潼

2013年11月3日

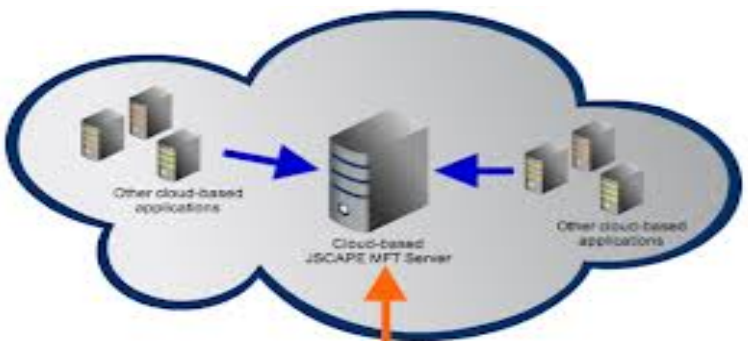
大数据在互联网

数据是互联网公司的最大战略资源
- 创造用户体验
- 创造商业价值

核心技术
- 大数据管理： infrastructure
- 大数据分析: machine learning
- 应用： system integration

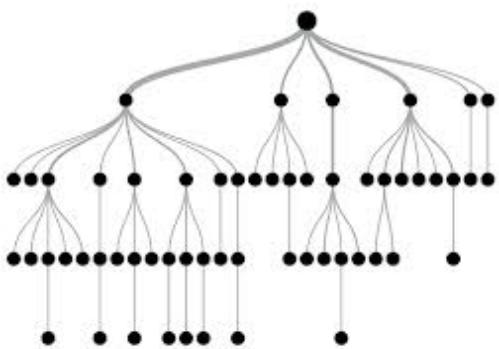
机器学习

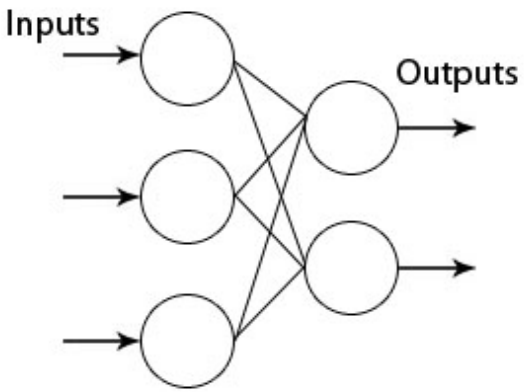
- 目标：让计算机系统更智能
- 方法： 大数据+计算能力+复杂模型+高效算法➔智能

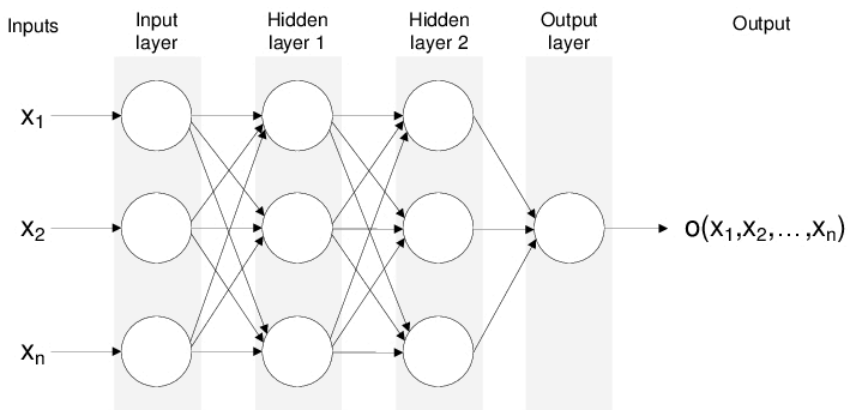


常用机器学习模型

观察量 ➔ 决策的数学模型







主要讨论监督模型

搜索广告

Baidu 百度

新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

北京美食 | 百度一下

推荐：用手机随时随地上百度

北京美食 首选国内领先的吃喝玩乐信息平台-易吃易乐 bj.echiele.com 推广链接
北京美食首选国内领先的吃喝玩乐信息平台-易吃易乐,每天有上百万网
易吃易乐 餐饮美食 休闲娱乐 美容美发

找北京美食?来DaoDao.com www.daodao.com
找北京美食?DaoDao.com为您提供210000条北京市旅游点评/攻略.

北京美食-大众点评网
根据合理的商区、地标和美食商户分类系统,为你提供北京83892家美食商户,并通过海量亲身消费者的点评聚合,以各种评分、星级的标准让你选择。
www.dianping.com/beijing/f... 2013-7-5 - 百度快照

北京美食攻略_北京美食推荐_美食街,小吃,指南-驴妈妈旅游网
驴妈妈旅游网关于北京美食攻略,包含更多北京特色美食小吃(美食,餐饮,娱乐),【旅游预订】打折门票,周边酒店,自由行及跟团游信息,就在(www.lvmama.com)
www.lvmama.com/travel/zhongguo_beiji... 2013-6-29 - 百度快照

北京有什么特色美食?_百度知道
13个回答 - 提问时间: 2011年12月25日
最佳答案: 1.烤鸭:在北京您要是想吃到便宜实惠的烤鸭,您可以去便宜坊、大鸭梨、安贞烤鸭店。当然您要是想吃最地道的烤鸭那就去和平门的全聚德。 2.涮羊肉:地...
zhidao.baidu.com/question/3585625... 2013-1-27 - 百度快照

北京美食_百度百科
北京美食guide是一款让你随时随地掌握北京美食信息的手机软件。北京美食拥有详尽的地图，十多种美食分类。
基本信息 - 软件介绍 - 安装指南 - 分辨率 - 软件截图
baike.baidu.com/ 2013-07-03

在北京市搜索北京美食_百度地图


A. 辣尚瘾(人大店) - (010)82650566
★★★★★ 1229条评论

北京美食
外文名: Guide
版本: V1.8.0
软件大小: 3941KB
来自百度百科>>

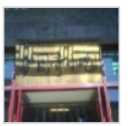
相关食物


北京小吃


北京烤鸭


门钉肉饼


褡裢火烧


护国寺小吃

推广链接
北京美食餐厅预定有折扣
上咕嘟妈咪,方便轻松享优惠北京美食;
咕嘟妈咪,不让亲朋排队等.
www.gudumami.cn

北京美食,金鼎鱼香渔村欢迎..
金鼎鱼香生态渔村,特色全鱼宴,灶台柴锅水库鱼,柴锅柴鸡,特色烧鸽
www.myjdyx.com

北京美食 刷雅酷卡 乐享无限..
找北京美食,精选北京美食折扣优惠!吃喝玩乐尽在雅酷卡网!
www.yacol.com

机器学习问题

- 点击率（CTR）预估
- 问题规模：

  数据存储和管理：上万台机器

  数据量：百亿到千亿级

  特征数：百亿到千亿级 （稀疏离散值特征）
- 大型线性Logistic Regression模型
- 计算技术：分布式同步CPU并行计算

语音识别

机器学习问题

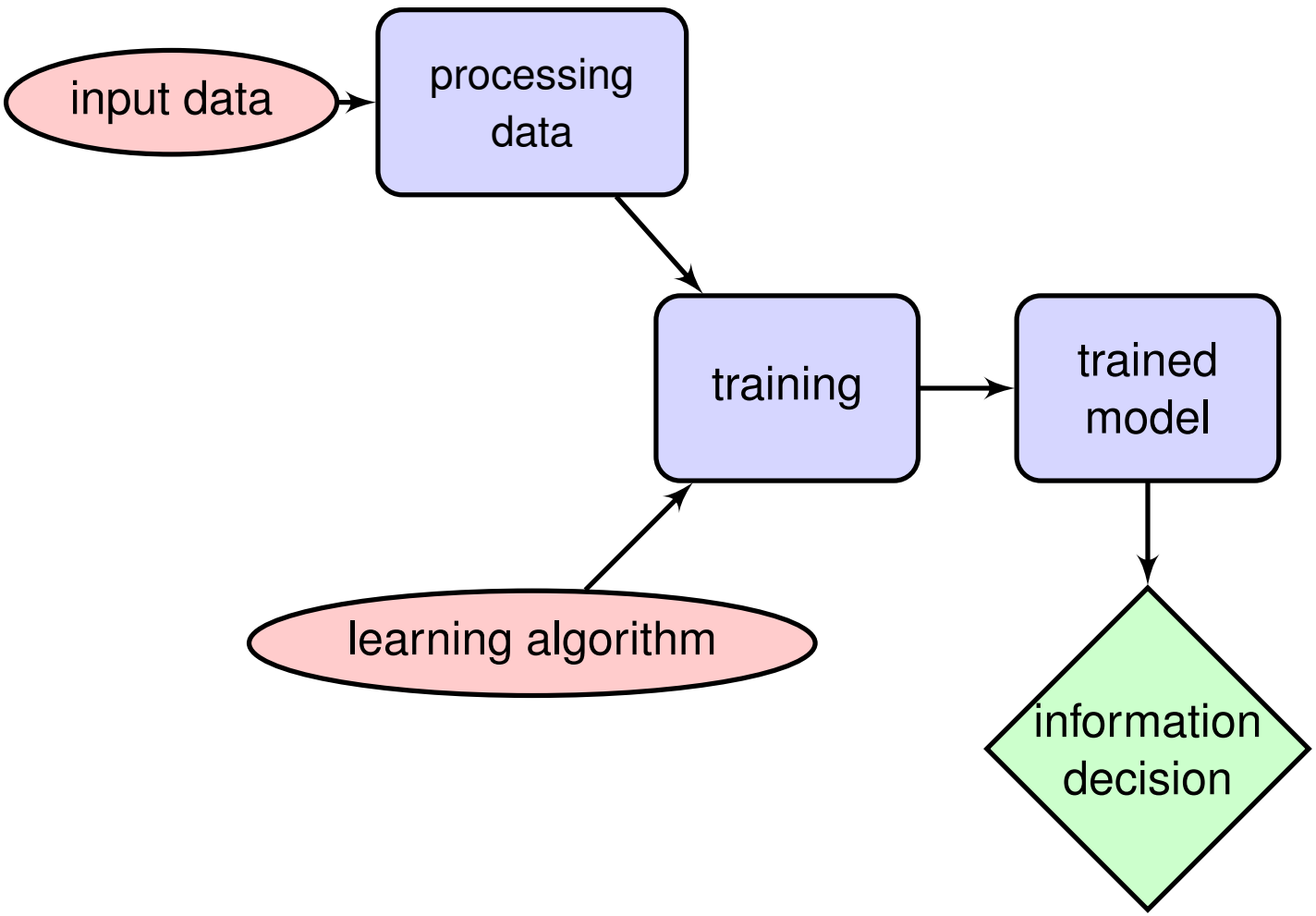
- 把声学信号变成文字：多分类问题
- 问题规模

  万小时级语料

  百亿级训练数据

  上万类别；几百维特征 （稠密连续值特征）
- 深度神经网络模型
- 计算技术：分布式异步GPU计算

机器学习流程


input data
processing data
training
trained model
learning algorithm
information decision

大规模机器学习

- 基础架构

  分布式数据存储，管理，和分析

  分布式CPU/GPU计算平台

- 算法

  模型和特征提取

  数据抽样

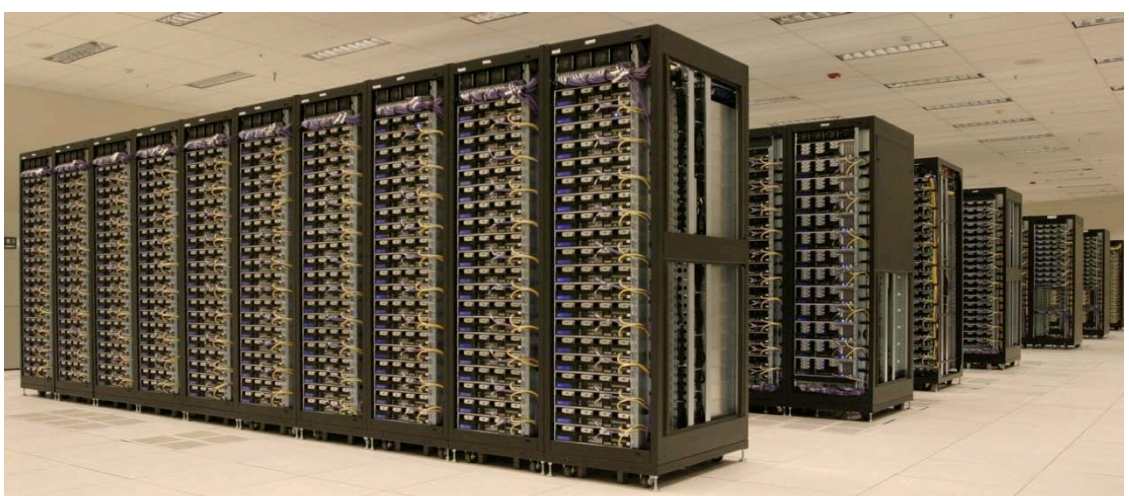
  大型机器学习模型优化

数据管理理

- Hadoop: 大数据存储 de facto standard

  分布式文件系统（HDFS）

  Map-Reduce



可用于机器学习特征提取

计算性能

CPU: complex tasks
Large memory (128G)
Few cores (8)
Peak 100+Gflops

适合：
稀疏离散特征
树模型

GPU: simple tasks extreme parallel
Small memory (5G)
Many cores (2K cores)
Peak 3Tflops

适合：
稠密连续值
深层神经网络计算

数学问题

- 大型机器学习训练优化问题:

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

  - 分布式多机并行训练

问题的分配方式

分配数据到多机

- 每个机器有所有模型参数
- 每个机器也不同数据

分配特征到多机

- 每个机器有所有数据的一些特征
- 每个机器有不同参数

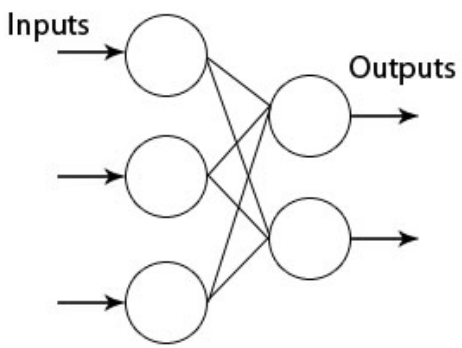
分配数据和特征到多机

- 每个机器有一些特征和一些参数

大型线性模型

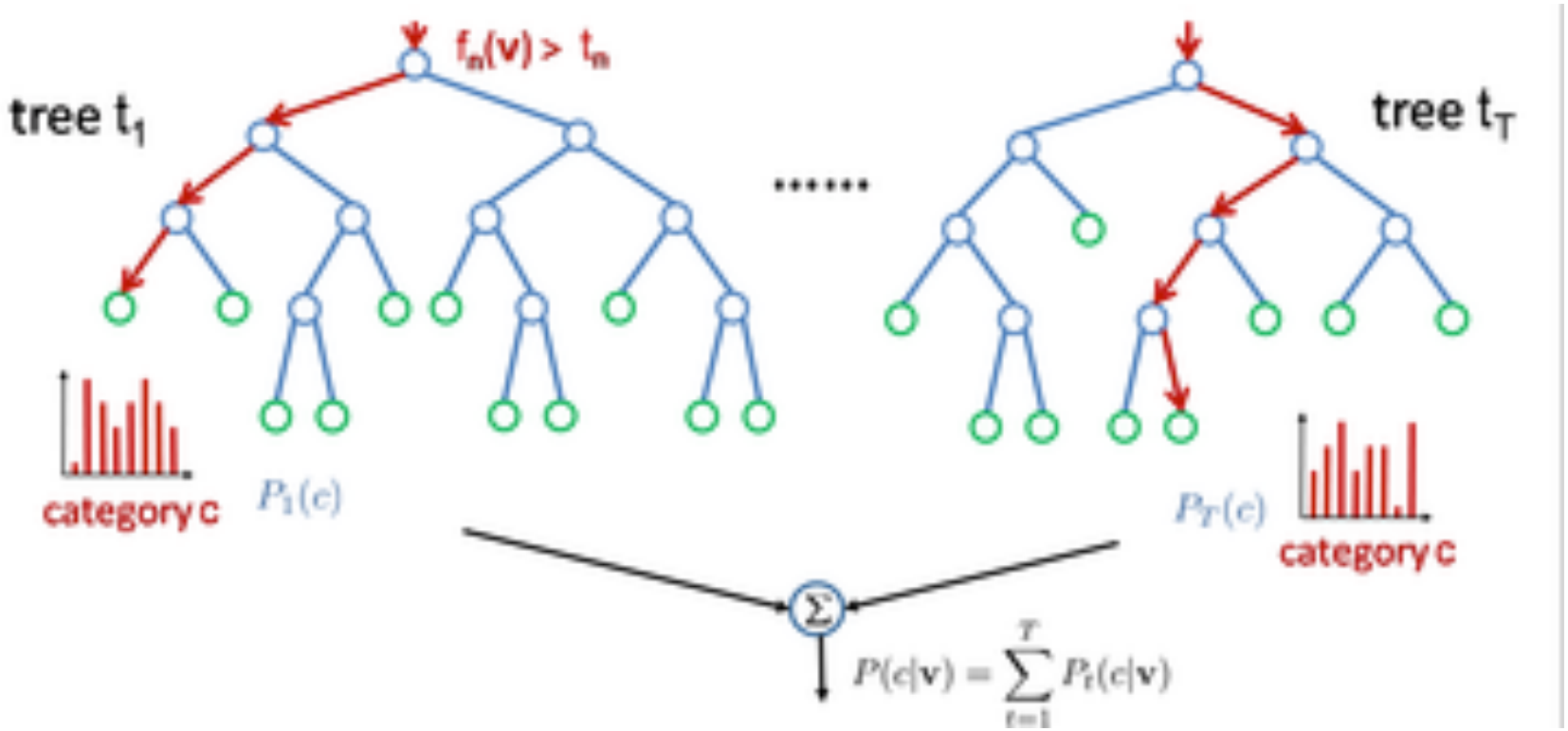
$$\min_w \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

$$f_i(w) = \ln(1 + e^{-w^\top x_i y_i})$$



多机CPU分布式计算

树模型

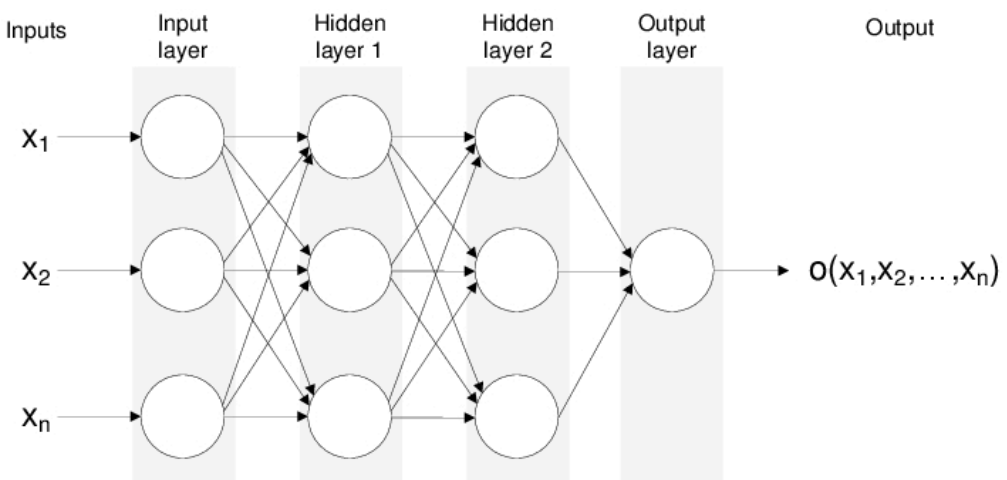


可用CPU分布式计算

把特征发到不同机器

深度神经网络

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$



多机GPU/CPU分布式计算

大数据算法研究

问题

用什么数据解决什么问题

数据

数据融合，结构；噪声过滤和纠偏; 数据抽样方法; 数据降维

模型

图模型，树模型，深度神经网络

分布式计算

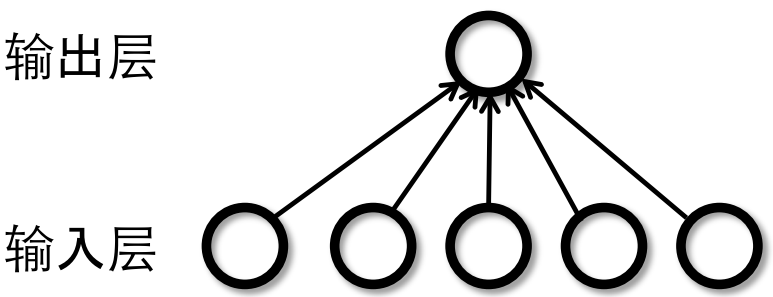
同步和异步；鲁棒性; 大模型; 理论分析

从浅层到深度学习

浅层网络:

人工特征抽取

学习线性组合

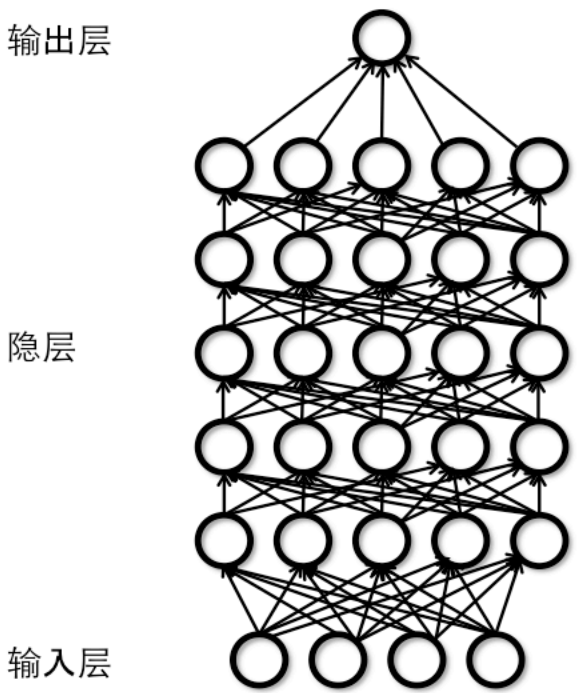


不含隐层的浅层学习模型

深层网络:

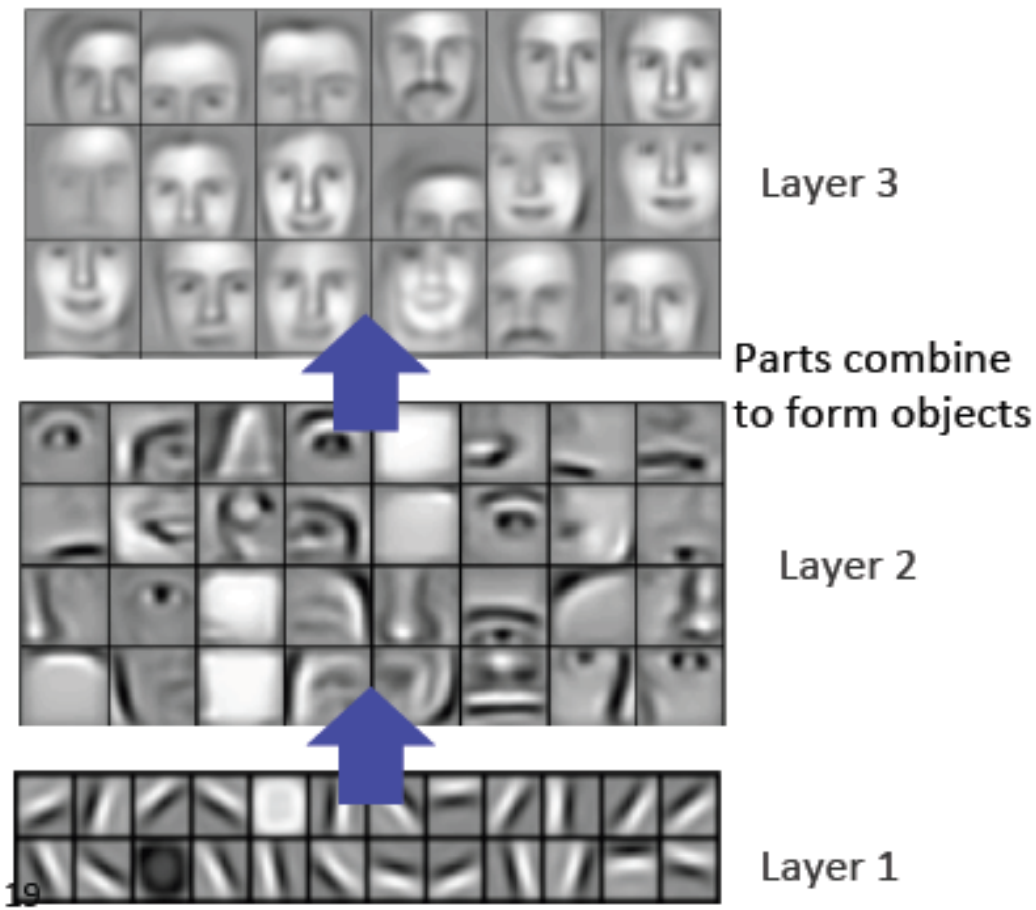
从原始特征出发

自动学习高级特征组合

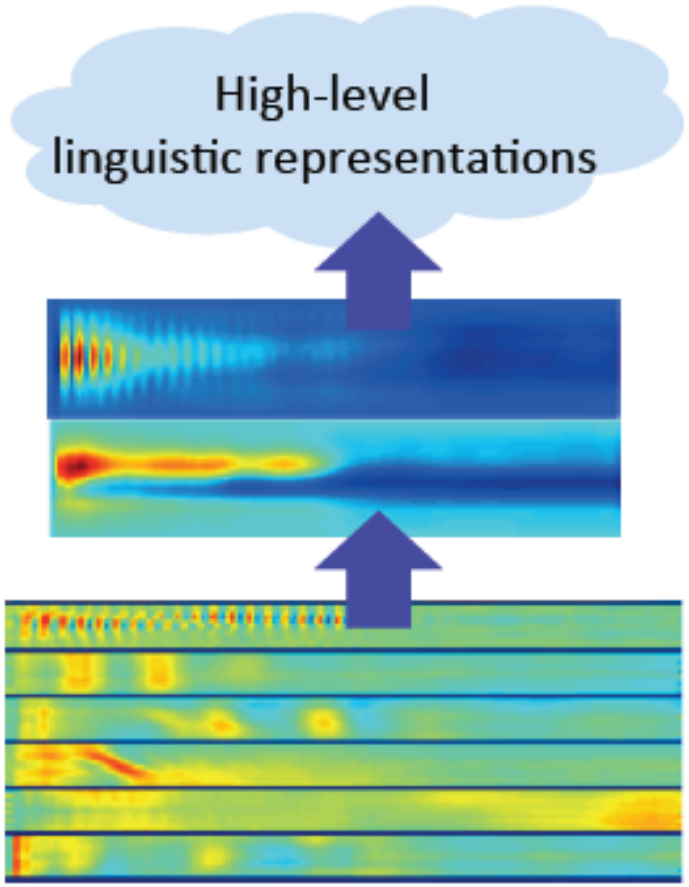


含多个隐层的深度学习模型

高级特征



Image



Speech

(Lee, Largman, Pham & Ng, NIPS 2009)
(Lee, Grosse, Ranganath & Ng, ICML 2009)

神经网络的发展

- Perceptron (1958-1969)
- Neural Networks (mid 1980 – early 1990)
- Deep Learning (2006 – now)
- 1995—2006
  - SVM，Kernel Machines
  - Convex; Linear
    - 好的理论分析
    - 容易调参

深度学习成功条件

- 2010-今：在工业界取得巨大成功
  - 复杂模型
- 大数据：100x
- 大规模计算能力：1000x

- 大数据+计算能力+复杂模型+高效算法➔智能

深度学习在百度

- 2012年夏天投入研发
- 用GPU提升计算效率，处理海量训练数据
- 语音识别，OCR识别，人脸识别，图像搜索等巨大提升

- 到目前，超过8项技术在产品上线

百度深度学习成果

- 语音：错误率相对降低20-30%
- OCR：错误率相对降低30%
- 人脸识别：世界最好结果
- 全网相似图像搜索：效果显著超谷歌同类产品
- 全流量上线广告CTR预估，显著提升广告点击率

语音产品


语音搜索
输入法
地图搜索
语音助手

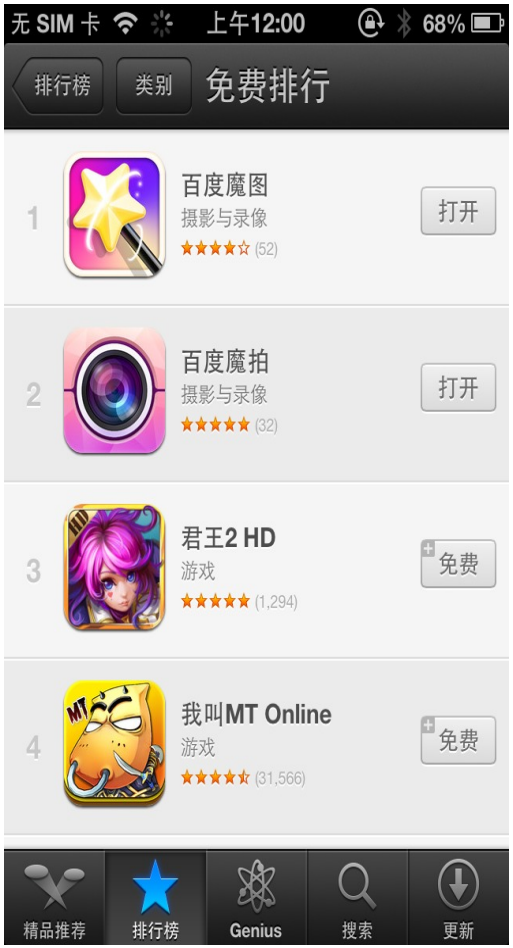

请说话
取消
完毕
搜索动口不动手


评论微博
我觉得中国好声音还不错，特别喜欢刘欢老师。


百度语音助手
说到，做到


从五道口到三里屯
查看方案

百度魔图

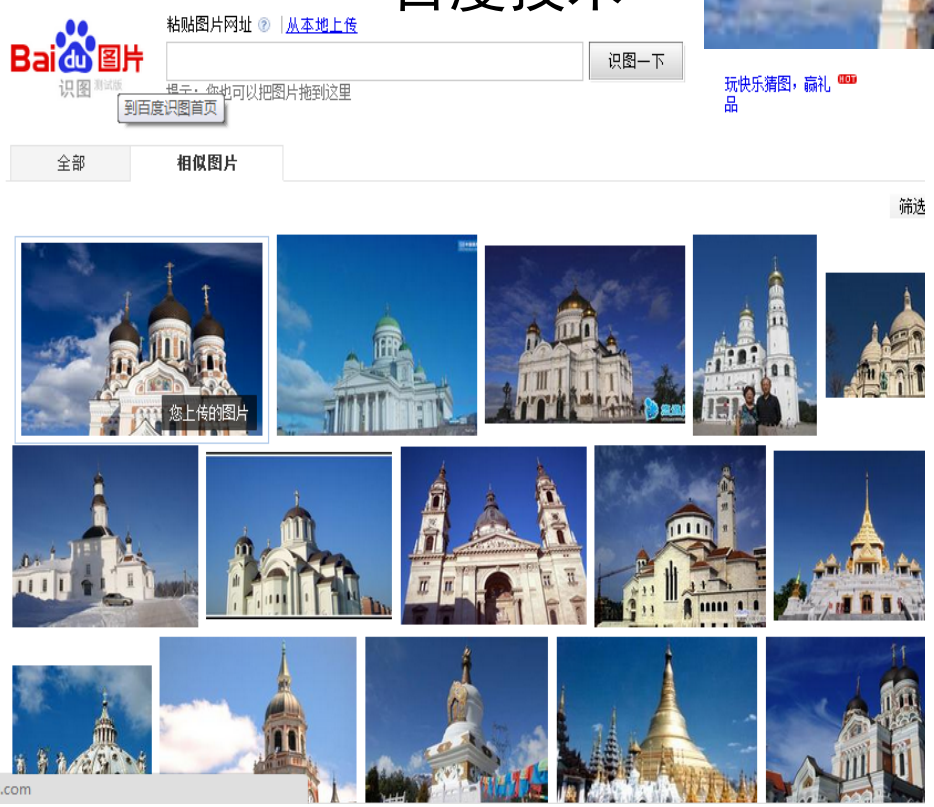


单日最高上载9000张图片，在IOS APP排行榜总榜排名第一达3周之久

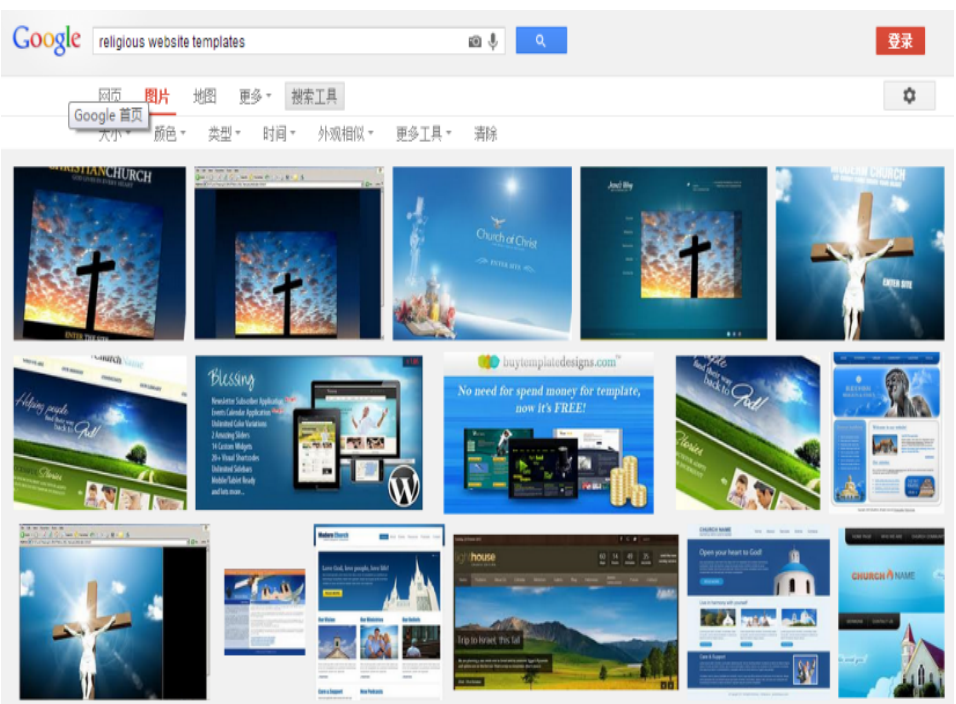
图片搜索

检索图片

百度技术



Google搜索结果

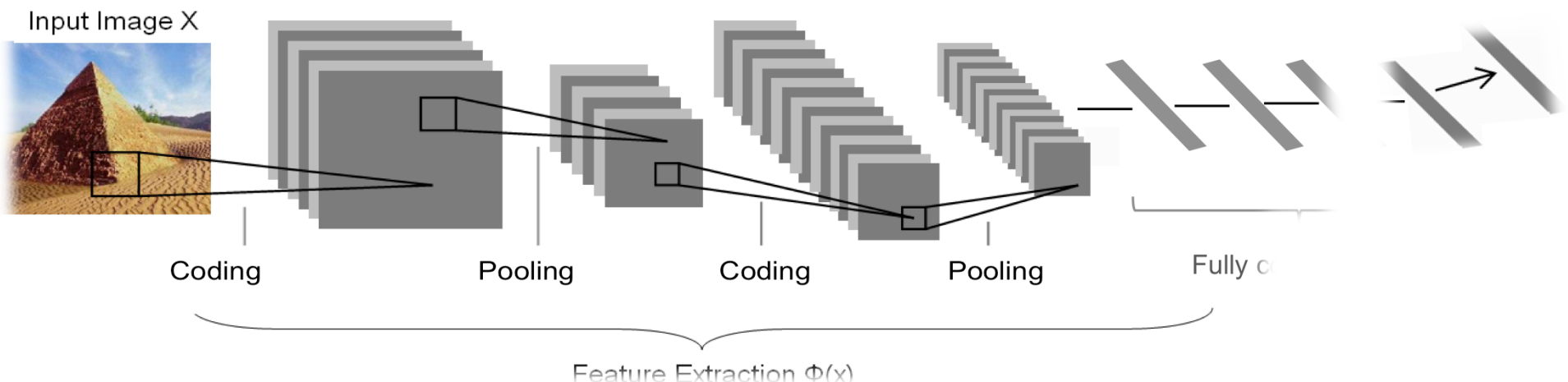


数据规模

- 图像识别： 数千万训练样本
- OCR： 数千万训练样本
- 语音识别： 数百亿训练样本
- 广告：千亿训练样本
- …

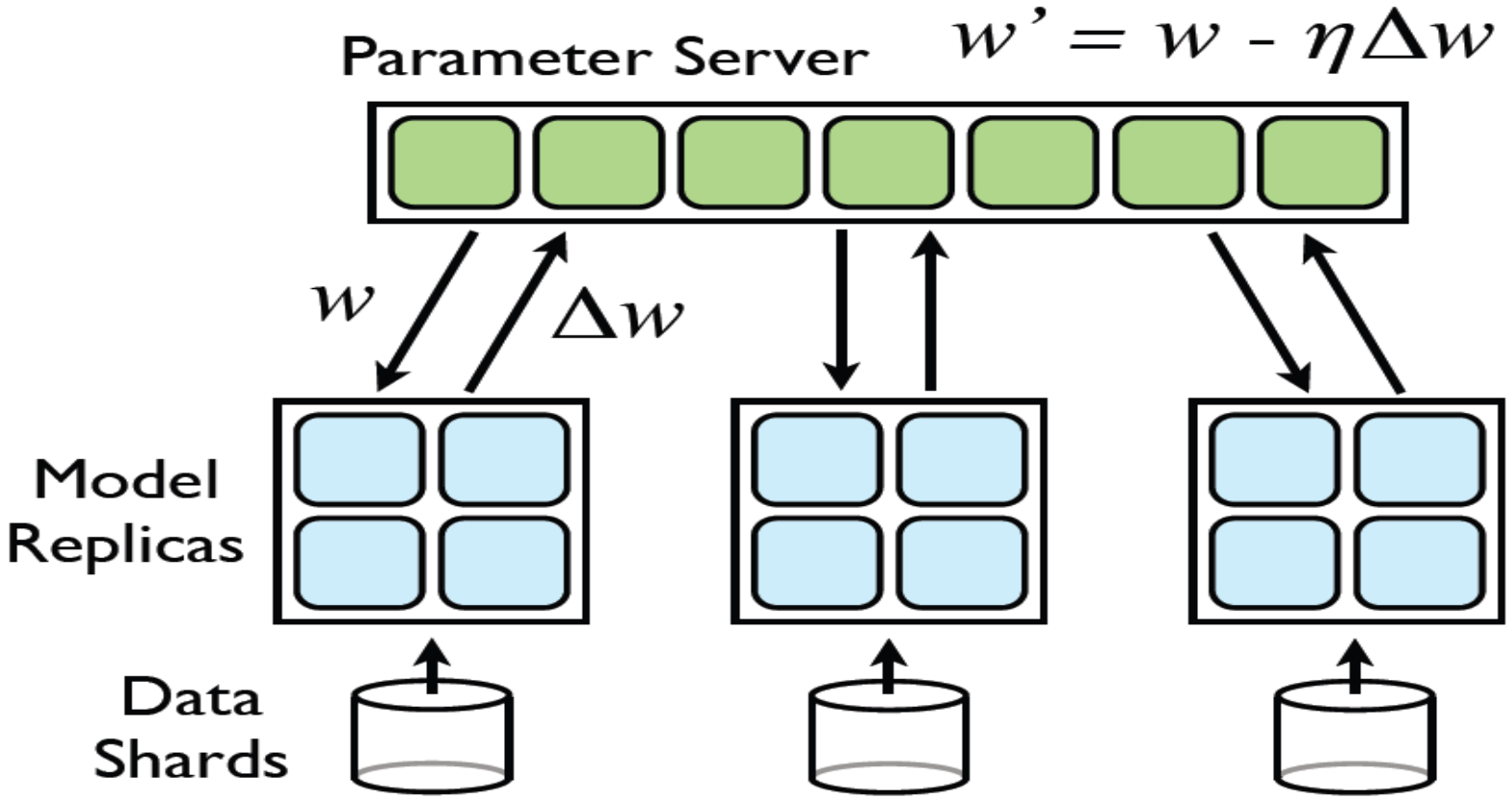
训练数据每年成倍增长 …

深度学习模型：图像


Input Image X
Coding
Pooling
Coding
Pooling
Fully c
Feature Extraction Φ(x)

计算能力和算法

- 几十台GPU并行计算
- 并行算法



深度学习研究

- 大数据的问题
- 基于问题的模型结构
  - Nonlinear feature discovery
  - Knowledge representation
  - Forming high level semantics
- 大规模分布式算法
  - Platform + engineering + algorithm

大数据和深度学习的意义

- 目标：计算机智能和人工智能
- 手段：
  - 大数据
  - 复杂模型
  - 计算能力和算法
  - 系统合成
- 深度学习：最接近人脑的复杂模型
  - 目前向人工智能走得最近的方法