机器学习的大数据思辨

刘铁岩
微软亚洲研究院 副院长

AlphaGo Zero当真不需要大数据吗？

- 如何界定机器学习中的大数据？
  - 大数据 vs. 大量有标签数据
  - 有标签数据 vs. 带反馈数据
  - 历史数据 vs. 仿真数据，……
- AlphaGo具有特殊性
  - 确定性游戏规则：天然的终止条件和胜负判决准则（本质上是搜索问题）
  - 更一般的学习任务可能无法利用确定性游戏规则生成大量带反馈信号的仿真数据以驱动强化学习。

今天的深度学习仍离不开大数据

深度学习技术依赖大规模数据

- 深度学习利用复杂的模型实现超强的拟合能力
- 大模型的训练离不开大量训练样本

反思

- 人类智能是否同样依赖大数据？
- 是否存在不那么依赖大数据的机器学习方法？
- 如何改造深度学习以减少其对大数据的依赖？

人类是否同样离不开大数据？

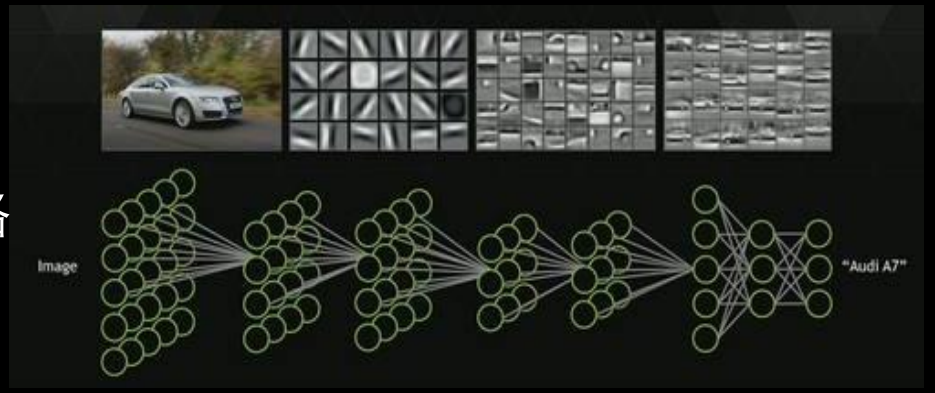
- 众多研究表明：
  - 人类在很多时候表现出极强的小样本学习能力
  - 但人类也并非对所有任务都能实现小样本学习，当面临不熟悉、非自然存在的学习任务时（如对二进制序列进行分类），也会束手无策。
- “迁移学习”假说：
  - 认为人类之所以可以仅利用少量样本就实现对某个任务的学习，是因为很多其他相关或相似的任务为其表示和结构的学习提供了帮助。
- “基因先验”假说：
  - 认为人类通过世代遗传，获得了高效的表示和结构基础，为其处理小样本学习任务提供了坚实的物质基础。

机器学习 >> 深度神经网络

Unsupervised learning （无监督学习利用数据的相似性挖掘有用的信息）	Support Vector Machines (支持向量机为小样本而生，通过正则化实现较强的泛化能力）	Semi-supervised Learning （半监督学习利用数据相似性，生成伪标签或正则项）	Transfer learning, zero/few shot learning （迁移学习利用任务之间的相关性减少单个任务对有标签数据的依赖）
Bayesian learning （贝叶斯学习利用先验知识降低机器学习对经验数据的依赖）	Generative adversarial networks （对抗生成网络通过训练生成器，创造数据用于后续训练）	Dual learning （对偶学习利用AI任务天然结构对偶性，创造强化学习回路，减少对有标签数据的依赖）	……

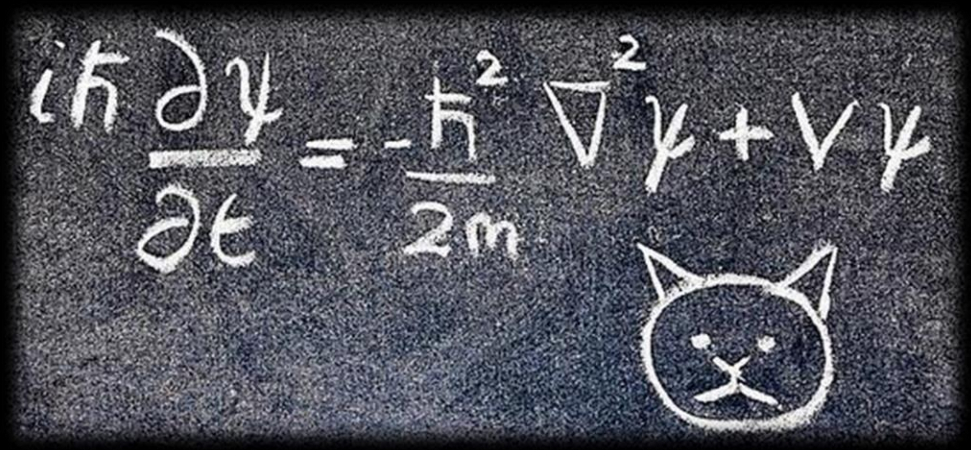
如何改造深度学习，减弱其对大数据的依赖？

- 深度学习引以为傲的是其极强的函数拟合能力
  - Universal Approximation Theorem
- 然而，该定理是存在性的，无法保证函数拟合的效率
  - 因为深度神经网络中常用的激活函数表达能力较差，对于复杂函数可能需要指数多的神经元才能实现有效拟合。
  - 通过提高激活函数自身的表达能力，对同样的函数拟合任务，可大幅减少神经元数目；而一旦网络规模下降，则不再需要那么多的训练样本。
  - 例：ArithNet通过使用复杂激活函数，帮助神经网络以线性的模型复杂度解决Parity问题（Parity问题被认为是DNN的failure case之一）



如何改造深度学习，减弱其对大数据的依赖？

- 在训练数据上直接进行函数拟合本身可能就是一个误区！
- 科学家向来从数据表象中发现背后的规律，而非简单拟合数据表象本身
  - 牛顿定律
  - 薛定谔方程
- 变换学习思路，可以使现有深度学习技术换发青春
  - 拟合复杂动态系统背后的简单规律（偏微分方程），逼近数据生成方式，而非数据本身！



劝君莫要一味追求大数据

- 当学习任务逐渐精细化（个性化），所有数据都将变成小数据
  - 需要追求共性和个性的平衡
- 大数据并不见得一定能带来大信息
  - 并非所有知识都能通过数据形式进行表达（如物理定律、数学理论等）
  - 大数据看起来丰富很可能是因为噪声，其背后的真实驱动系统可能非常简单
- 数据其实只是问题的一个侧面
  - 机器学习在建模方面的过度简化（例如假设样本是独立同分布的、数据分布式静态、不随模型而变化的），可能导致再多数据也无法取得更好的学习效果（模型假设失配）。

以全局的眼光看待机器学习，不要为数据所困，寻找新的突破点。