智能技术趋势的若干思考

张亚勤
清华大学

关键词：智能技术　数字化

近几十年来，随着计算机技术的快速发展，我们经历了两波数字化浪潮，这些浪潮中的产品、技术等至今影响着世界和我们的生活。现在随着智能技术的不断迭代，人类社会即将迎来第三波数字化浪潮，我们将拥抱怎样的全新智能化技术？智能化未来的发展又会遇到什么样的瓶颈？解决方法是什么？智能技术的发展会给产业带来哪些新的机遇？中国在这一次浪潮中又将担任什么样的角色？本文将对以上几个问题展开探讨。

数字化 3.0——信息、物理和生物世界的融合

过去 30 余年，IT 行业最大的发展就是数字化。第一波数字化 1.0 从 20 世纪 80 年代中期开始，主要是内容数字化。数字化的主要对象为语音、音乐、视频、图像和数字文档等，主要成果有 MP3/4、JPEG、MPEG、H.26X、AVS、HDTV、Word、PowerPoint 和 Excel 等。相关技术的发展也非常迅速，例如视频编码标准从最初的 H.261 发展到现在的 H.266。目前内容数字化已经完全实现。

第二波数字化 2.0 从 90 年代中期开始，主要是互联网和企业数字化。首先，前期的内容数字化产生了消费互联网，包括从门户网站、搜索引擎、社交平台、电商，到 O2O、分享经济，再到如今以 ZOOM 为代表的视频通讯、数字货币和移动支付等。其次，企业也在数字化，包括最初的 ERP、CRM、HR，再到供应链（supply chain）、Workflow、商业智能，以及大型数据库和云计算。中国在互联网移动支付的多个方面处于世界领先水平，包括技术、规模、用户群体和产品体验，但是中国在消费类软件和企业级的软件产品方面一直没有进入主流市场。在企业软件领域，中国与美国相比落后 5 年左右。在企业层面，相比于已开始形成规模的 IaaS（Infrastructure as a Service，基础设施即服务）和 PaaS（Platform as a Service，平台即服务），未来 10 年在软件方面最大的增长点是在 SaaS（Software as a Service，软件即服务）层。

第三波数字化 3.0 就是现在，主要是信息、物理和生物世界的融合，包括物理世界和生物世界的数字化。物理世界包括家庭、交通、城市和工业，比如车、船、飞行器、道路、工厂、机器等都在实现数字化。物理世界数字化可以看成物理世界和数字世界形成一对一的影射和孪生，比如过去提到的数字高速公路，现在高速公路已经可以数字化。物理世界数字化会产生天文级别的数据和信息，比如一辆无人驾驶汽车每天产生的数据已经达到了 TB 数量级。这些数据大部分不是给人“看”，而主要是给机器“看”的，比如在无人驾驶中，90% 的数据是提供给机器做决策的。

生物世界也在数字化。我们的身体乃至每个器官，甚至 DNA、蛋白质，都可以通过生物电子芯片（bio-electron chips）、神经元传感器（neuron sensors）等接口与世界相连。这样产生的数据量将会更大，比如一个蛋白质空间结构的信息量是 10^{300}，一个人体的高通量测序技术（High-Throughput Sequencing，HTS）基因测序产生的数据量是 5PB~6PB，这个量

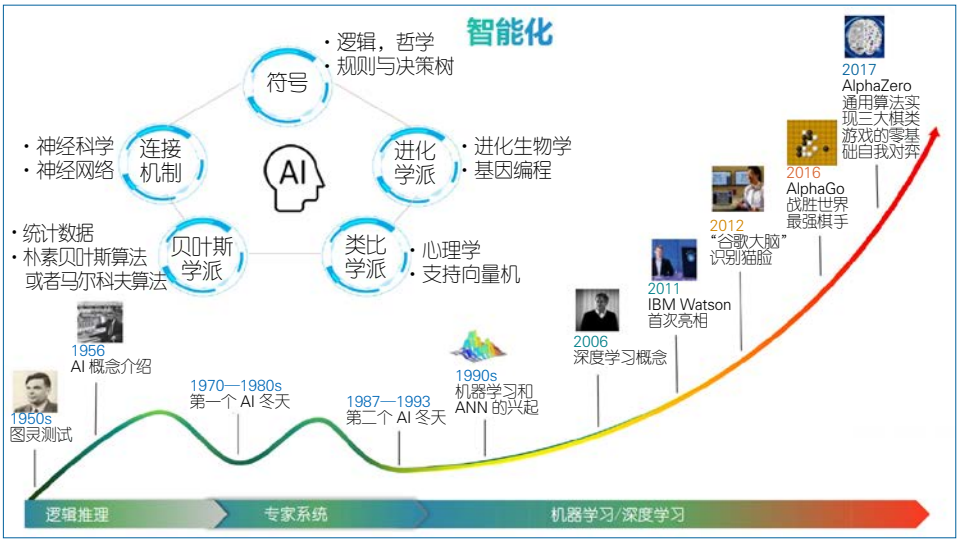

图 1　人工智能起伏的发展历程和五大流派

了 AI 的五大流派，包括符号、进化学派、类比学派、贝叶斯学派和连接机制。未来 AI 的发展趋势最大的可能是借各种流派之长创造新的 AI 算法，既包含逻辑符号也有数据和知识，还要借鉴人类的进化和大脑的特点。所以，当前无论是科学研究还是产业发展，都在思考下一轮 AI 的突破点在什么地方。

级的数据很难再用普通的方式去处理计算。总之，现在的世界是信息、物理和生物世界的融合，将经历数字化、互联化，最终达到智能化。

图 1 展示了人工智能（AI）从 20 世纪 50 年代开始，经历了 60 年起伏的发展历程。其中，AI 共经历了 2 次发展的春天和冬天，现在正在经历第 3 个春天，即以数据为驱动力的深度学习。谷歌的 AlphaGo、AlphaZero、AlphaFold 是很好的标志性产品。

佩德罗·多明戈斯（Pedro Domingos）曾总结

计算体系与通讯范式的瓶颈与展望

算力需求飞速增长

对深度学习来说，半导体与芯片架构领域的进步是不可或缺的发展动力。谷歌公司的杰夫·狄恩（Jeff Dean）曾说过：“数据 + 算法 + 算力 = 数据 +100 × 算力”。也就是说，他认为在数据、算法和算力三大因素中，算力占据着绝对的主导地位，算法则相对来说没有那么重要。这样的说法或许有些过于绝对，笔者并不完全同意，但算力在当前的技术领域的确有十分重要的地位。

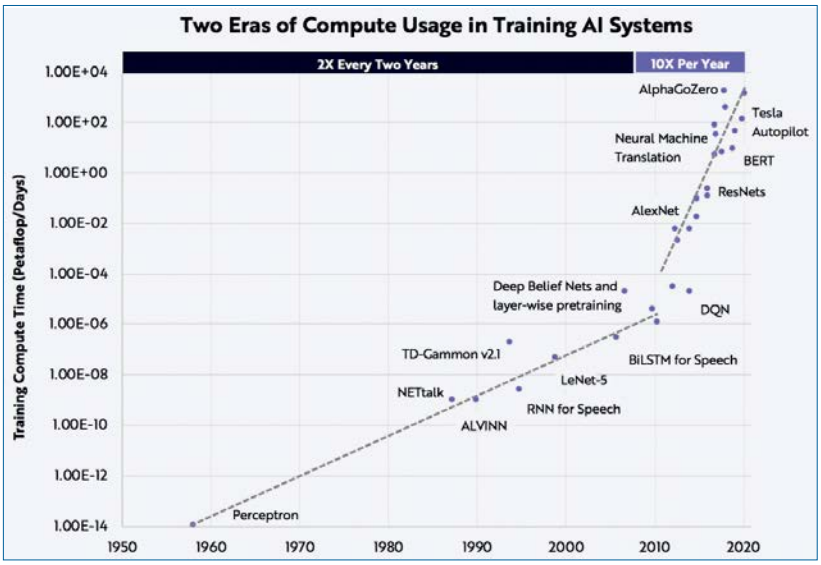

图 2　深度学习训练过程中计算量的发展

从图 2 来看，随着时代的发展，深度学习在训练过程中产生的计算量可以分成两个阶段：在深度学习发展的初期阶段，训练产生的计算量的增长速率相对较慢；近 10 年间，计算量以每年 10 倍的速率增长，远远超过摩尔定

律每18~24个月提高2倍的增长速率。

我们以OpenAI发布的预训练模型GPT为例，来说明近几年来机器学习领域对算力需求的飞速增长。2018年6月发布的GPT-1是在约5GB的文本上进行无监督训练，针对具体任务，在小的有监督数据集上做微调，得到包含1.1亿参数的预训练模型；而2019年2月发布的GPT-2则是在约40GB文本上进行无监督训练，得到具有15亿参数的预训练模型；而2020年5月公布的GPT-3则是在499B tokens（令牌）的数据基础上训练，得到包含1750亿参数的模型。在不到2年的时间内，模型参数从1.1亿的规模增长至1750亿，而单次训练GPT-3就需要花费1200万美元，模型在飞速发展的同时，带来的是巨大的算力要求和高成本的代价。

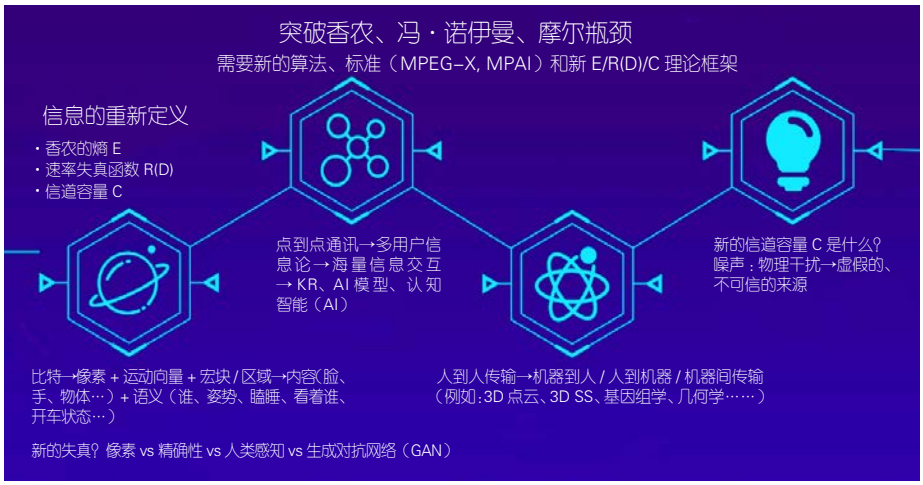

图3 信息论的研究方向

传统计算与通讯范式的瓶颈

人工智能/深度学习领域对算力的需求驱动了新算力的发展。要想谋求更高效率的计算，我们需要回到计算和通讯领域最基本的理论和范式。在过去的几十年间，涌现出了许许多多的定律和体系，而其中有三个定律和体系被视为计算与通讯范式的根本。

第一个是香农定律（Shannon Theroy）。香农是信息论的奠基者，他引入了信息熵的概念，为数字通信奠定了基础。其实香农定律定义了三个极限，分别为无损压缩极限E、信道传输极限C、有损压缩极限R(D)。目前，我们已经接近这些极限。

第二个是冯·诺伊曼架构（Von neumann Structure）。在冯·诺伊曼架构中，计算机由运算器、控制器、存储器、输入设备和输出设备5个基本部分组成，具有程序存储、共享数据、顺序执行的特点。冯·诺伊曼架构简单且漂亮，是图灵机的优秀范例，至今仍被广泛地应用。然而，冯·诺伊曼架构的设计构成了运算器和存储器间的瓶颈，这对深度学习的发展造成了一定的限制。

第三个是摩尔定律（Moore's Law）。戈登·摩尔（Gordon Moore）总结认为，集成电路上可以容纳的晶体管数目大约在18个月左右便会增加一倍。而现在晶体管数目的增长越来越慢，摩尔定律逐步趋向于饱和阶段，而我们对计算能力的需求却飞速提升，不断提升的算力需求与芯片技术发展趋缓的矛盾日趋显现。

计算体系与通讯架构的革新

在过去的60年里，这三个基本理论在计算和通讯领域建立了决定性的基础，然而日趋逼近的极限也使得当前AI技术的发展逐步接近瓶颈。为了避免技术的停滞不前，我们或许可以从以下三个方面做出一些突破和革新。

首先，对信息重新定义。香农于上世纪40年代对信息熵、速率失真函数R(D)和信道容量C做出了定义，而这些定义是基于比特的基础实现的。以视频图像举例，过去我们一直采用比特来描述信息；后来我们从数字的层面使用像素、运动向量、宏块（macroblock）和区域（regions）结合的方式来描述图像；之后我们上升到从内容层面来描述图像，比如一个身体部位是脸部还是手部等；现在我们对图像的描

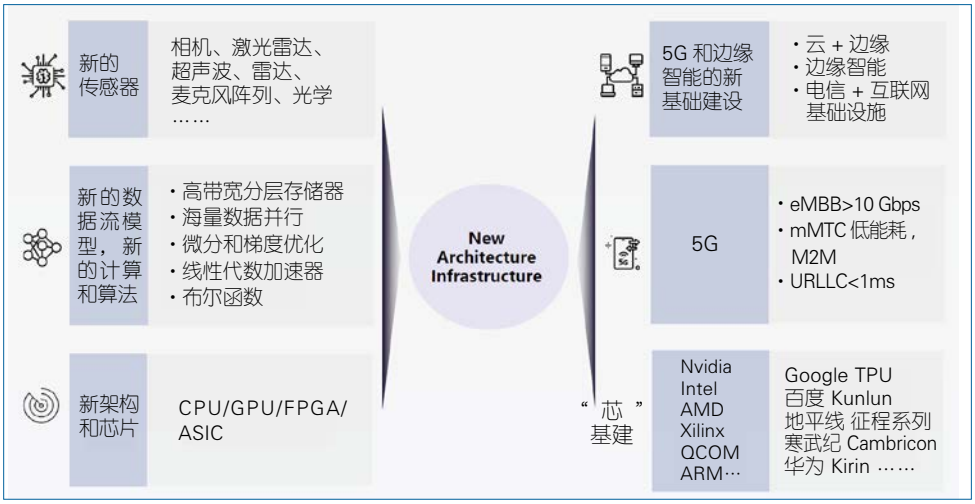

图 4　计算体系与通讯架构的研究方向

述上升到语义层面，比如“是谁”“在做什么动作”“是否在睡觉”“眼睛在看什么”等，这些问题从语义的层面描述了图像传达的信息。当信息的描述方式发生变化时，熵的概念也发生了变化。比方说，过去我们用比特的形式来描述图像失真现象，而现在我们用生成对抗网络（Generative Adversarial Networks，GAN）生成图像，用肉眼来看 GAN 输入的图片和生成的图像几乎是一致的，但是从比特层面来比较，会发现二者十分不同。因此，如何从语义、特征和内容的角度来定义熵与速率失真函数是我们未来需要研究的问题。另外，香农理论从最开始的点到点通讯，扩展到后来的多用户信息论。但是在当下的互联网时代，面对海量的交互信息，部分香农理论已不再适用，学术界却没有提出一个新的完善的理论。过去的信息更多的是人与人之间的传输，而现在的信息则更多地面向机器，比如 3-D Point Cloud、3D SS、Genomics、Geometry 等，所以我们需要新的算法和新的标准。近期，一个新的组织 MPAI 开始研究面向 AI 的针对机器和生物信息的算法和表征。我们也同样需要在表征理论方面做更多的工作，建立新的“信息论”。信息论的研究方向如图 3 所示。

第二，我们需要新的计算范式。包括量子计算、类脑计算和生物计算等在内的新的计算范式能够为计算瓶颈提供解决途径。

第三，我们需要新的计算体系和通讯架构（见图 4），突破冯·诺伊曼体系架构的限制。首先，我们需要新的传感器、新的数据流架构和计算模式，以及高速的存储，这些都与传统的冯·诺伊曼架构不同。我们还需要新的通讯架构，即 5G 技术和边缘计算。5G 技术首次在应用层上实现了“三网合一”，比提升传输的速度更加有效。此外，5G 技术更好地解决了延时问题并带来了新的应用，如百度的阿波罗项目中有一个服务叫做“云代驾”，通过 5G 技术让远程的安全操作员实时了解车辆所处的环境与状态，在自动驾驶无法完成的场景下接管车辆，完成远程协助。但 3G 和 4G 网络的延迟使得“云代驾”模式无法成为现实，必须通过 5G 网络来解决延迟问题。很多人认为，当前的 5G 技术在能耗和覆盖率等方面还没有达到预期，但笔者认为任何新技术的发展都需要时间，相信在未来的三五年后，5G 技术能够为用户、工业和产业界带来巨大的变革。

芯片的升级对产业界的作用是显著的，近年来国内有许多公司在芯片领域有所成就。以百度的昆仑 AI 芯片为例（见图 5），第一代昆仑芯片采用 14nm 先进工艺，2.5D 封装，使用 HBM 内存，可以达到 512GB/s 的带宽。而预计于 2021 年量产的第二代昆仑芯片，采用 7nm 先进工艺，性能是第一代昆仑芯片的 3 倍，同时耗能减少，具备了大规模片间互联的能力，进步显著。

另一个芯片领域的成果是地平线自动驾驶芯片（见图 6）。其中，AI 芯片 Journey 系列从第一代研发到第三代，严格按照车规级可靠性要求设计，符合 ISO 26262 标准，同时在质量、性能、耗能等方面的表现都十分优秀，和许多国际顶尖芯片相比也

图 5　百度昆仑 AI 芯片路线图

图 6　地平线自动驾驶芯片发展图

毫不逊色。这些芯片领域的成果将推动算力的进步，为 AI 技术的发展进步提供更强的驱动力。

智能技术变革带来的新机遇

智能技术的蓬勃发展为包括 IT 在内的诸多产业带来新的机遇，主要包括三个方面。

1. AI 技术将使 IT 产业本身成为最大的受益者，无论是芯片技术、操作系统，还是云平台和应用，都在不断地快速迭代。

2. AI 技术的发展将改变甚至颠覆目前的产业，届时，教育、医疗、金融、制造、交通等各个产业都会与 AI 相结合，AI 将像过去 20 年来互联网的经历一样，逐渐渗透并融入到每个行业中。

3. AI 技术将会创造更多新的产业，典型的例子有：自动驾驶、工业物联网、AI+ 医疗，这三个领域在未来都很有潜力。智能交通未来的研究方向将集中在环境建模、协同感知与融合、驾驶行为分析和 V2X 车路协同领域等；工业物联网则集中于万物的安全互联、自我认知的环境计算、超低功耗的边缘计算、基于模型的感知与推理等；AI+ 医疗将聚焦于新药物研发、影像学与医疗诊断、基因组编码与分析、AI 辅助的神经科学等。不同的研究领域虽然有不同的研究方向，但是 AI 技术为它们提供了共同的技术支撑，可以简单归纳为“ABCD”，即人工智能算法（Artificial Intelligence）、大数据（Big Data）、云（Cloud）和设备（Device）。

总结

现在，我们正在面临一次新的工业革命。在前三次工业革命中，中国都是旁观者，仅在第三次工业革命的后期，中国才在很多方面开始崛起。如今的第四次工业革命，我们有机会把握这些新机遇，成为领先者。 ■

（本文根据 CNCC 2020 特邀报告整理而成）

致谢：感谢百度公司首席芯片架构师欧阳剑和地平线公司 CEO 余凯博士为本文提供部分图片。

张亚勤

清华大学讲席教授，清华大学智能产业研究院院长。IEEE Fellow，美国艺术与科学院院士。主要研究方向为数字视频和人工智能。

整理者：王　哲　陈小雪