解析 | 人工智能发展及技术架构

数邦客 昨天

一、 人工智能技术发展概述

（一） 人工智能技术流派发展简析

让机器实现人的智能，一直是人工智能学者不断追求的目标，不同学科背景或应用领域的学者，从不同角度，用不同的方法，沿着不同的途径对智能进行了探索。其中，符号主义、连接主义和行为主义是人工智能发展历史上的三大技术流派。

符号主义又称为逻辑主义，在人工智能早期一直占据主导地位。该学派认为人工智能源于数学逻辑，其实质是模拟人的抽象逻辑思维，用符号描述人类的认知过程。早期的研究思路是通过基本的推断步骤寻求完全解，出现了逻辑理论家和几何定理证明器等。上世纪70年代出现了大量的专家系统，结合了领域知识和逻辑推断，使得人工智能进入了工程应用。PC机的出现以及专家系统高昂的成本，使符号学派在人工智能领域的主导地位逐渐被连接主义取代。

连接主义又称为仿生学派，当前占据主导地位。该学派认为人工智能源于仿生学，应以工程技术手段模拟人脑神经系统的结构和功能。连接主义最早可追溯到1943年麦卡洛克和皮茨创立的脑模型，由于

受理论模型、生物原型和技术条件的限制，在20世纪70年代陷入低潮。直到1982年霍普菲尔特提出的Hopfield神经网络模型和 986年鲁梅尔哈特等人提出的反向传播算法，使得神经网络的理论研究取得了突破。2006年，连接主义的领军者 Hinton 提出了深度学习算法，使神经网络的能力大大提高。2012年，使用深度学习技术的AlexNet模型在 ImageNet 竞赛中获得冠军。

行为主义又称为进化主义，近年来随着AlphaGo取得的突破而受到广泛关注。该学派认为人工智能源于控制论，智能行为的基础是“感知—行动”的反应机制，所以智能无需知识表示，无需推断。智能只是在与环境交互作用中表现出来，需要具有不同的行为模块与环境交互，以此来产生复杂的行为。

在人工智能的发展过程中，符号主义、连接主义和行为主义等流派不仅先后在各自领域取得了成果，各学派也逐渐走向了相互借鉴和融合发展的道路。特别是在行为主义思想中引入连接主义的技术，从而诞生了深度强化学习技术，成为AlphaGo战胜李世石背后最重要的技术手段。

（二） 深度学习带动本轮人工智能发展

深度学习已经在语音识别、图像识别等领域取得突破。深度学习全称深度神经网络，本质上是多层次的人工神经网络算法，即从结构上模拟人脑的运行机制，从最基本的单元上模拟了人类大脑的运行机制。深度学习已经开始在计算机视觉、语音识别、自然语言理解等领域取得了突破。在语音识别领域，2010年，使用深度神经网络模型的语音识别相对传统混合高斯模型识别错误率降低超过 20%，目前所有的商用语音识别算法都基于深度学习。在图像分类领域，目前针对ImageNet数据集的算法分类精度已经达到了 95%以上，可以与人的分辨能力相当。深度学习在人脸识别、通用物体检测、图像语义分割、自然语言理解等领域也取得了突破性的进展。

海量的数据和高效的算力支撑是深度学习算法实现的基础。深度学习分为训练(training)和推断(inference)两个环节。训练需要海量数据输入，训练出一个复杂的深度神经网络模型。推断指利用训练好的模型，使用待判断的数据去“推断”得出各种结论。大数据时代的到来，图形处理器（Graphics Processing Unit，GPU）等各种更加强大的计算设备的发展，使得深度学习可以充分利用海量数据（标注数据、弱标注数据或无标注数据），自动地学习到抽象的知识表达，即把原始数据浓缩成某种知识。当前基于深度学习的人工智能技术架构如图1所示。

二、基于深度学习的人工智能技术现状

(一)基于深度学习的人工智能技术体系综述

当前，基于深度学习的人工智能算法主要依托计算机技术体系架构实现，深度学习算法通过封装至软件框架1的方式供开发者使用。软件框架是整个技术体系的核心，实现对人工智能算法的封装，数据的调用以及计算资源的调度使用。为提升算法实现的效率，其编译器及底层硬件技术也进行了功能优化。具体架构请见图1中的基础硬件层、深度神经网络模型编译器及软件框架三层。

本章所探讨的人工智能技术体系主要包含三个维度，一是针对人工智能算法原理本身的探讨，二是对算法实现所依托的技术体系进行概述，三是针对深度学习所需的数据进行分析。

1. 基础硬件层

基础硬件层为算法提供了基础计算能力。硬件层涵盖范围除了中央处理器（Central Processing Unit，CPU）及GPU外，还包括为特定场景应用而定制的计算芯片，以及基于计算芯片所定制的服务器，包括GPU 服务器集群，各类移动终端设备以及类脑计算机等。

2. 深度神经网络模型编译器

深度神经网络模型编译器是底层硬件和软件框架、以及不同软件框架之间的桥梁。该层旨在为上层应用提供硬件调用接口，解决不同上层应用在使用不同底层硬件计算芯片时可能存在的不兼容等问题。其涵盖范围包括针对人工智能计算芯片定向优化的深度神经网络模型编译器，以及针对不同神经网络模型表示的规定及格式。

3. 软件框架层

软件框架层实现算法的模块化封装，为应用开发提供集成软件工具包。该层涵盖范围包括针对算法实现开发的各类应用及算法工具包，为上层应用开发提供了算法调用接口，提升应用实现的效率。

1. 基础应用技术

当前人工智能的商业化实现主要是基于计算机视觉、智能语音、自然语言处理等基础应用技术实现，并形成了相应的产品或服务。本部分将在第三章进行详细讨论。

（二） 算法发展趋势

当前，人工智能算法已经能够完成智能语音语义、计算机视觉等智能化任务，在棋类、电子游戏对弈，多媒体数据生成等前沿领域也取得了一定进展，为人工智能应用落地提供了可靠的理论保障。

1. 算法的设计逻辑

人工智能算法的设计逻辑可以从“学什么”、“怎么学”和“做什么”三个维度进行概括。

首先是学什么。人工智能算法需要学习的内容，是能够表征所需 完成任务的函数模型。该函数模型旨在实现人们需要的输入和输出的映射关系，其学习的目标是确定两个状态空间（输入空间和输出空间） 内所有可能取值之间的关系；其次是怎么学。算法通过不断缩小函数 模型结果与真实结果误差来达到学习目的，一般该误差称为损失函数。损失函数能够合理量化真实结果和训练结果的误差，并将之反馈给机 器继续作迭代训练，最终实现学习模型输出和真实结果的误差处在合 理范围；最后是做什么。机器学习主要完成三件任务，即分类、回归

和聚类。目前多数人工智能落地应用，都是通过对现实问题抽象成相应的数学模型，分解为这三类基本任务进行有机组合，并对其进行建模求解的过程。

2. 算法的主要任务

人工智能实际应用问题经过抽象和分解，主要可以分为回归、分类和聚类三类基本任务，针对每一类基本任务，人工智能算法都提供了各具特点的解决方案：

一是回归任务的算法。回归是一种用于连续型数值变量预测和建模的监督学习算法。目前回归算法最为常用的主要有四种，即线性回归（正则化）、回归树（集成方法）、最邻近算法和深度学习。二是分类任务的算法。分类算法用于分类变量建模及预测的监督学习算法，分类算法往往适用于类别（或其可能性）的预测。其中最为常用的算法主要有五种，分别为逻辑回归（正则化）、分类树（集成方法）、支持向量机、朴素贝叶斯和深度学习方法。三是聚类任务的算法。聚类算法基于数据内部结构来寻找样本集群的无监督学习任务，使用案例包括用户画像、电商物品聚类、社交网络分析等。其中最为常用的算法主要有四种即 K 均值、仿射传播、分层/ 层次和聚类算法

(Density-Based Spatial Clustering of Applications with Noise，DBSCAN)。

1. 新算法不断提出

近年来，以深度学习算法为代表的人工智能技术快速发展，在计算机视觉、语音识别、语义理解等领域都实现了突破。但其相关算法目前并不完美，有待继续加强理论性研究，也不断有很多新的算法理论成果被提出，如胶囊网络、生成对抗网络、迁移学习等。

胶囊网络是为了克服卷积神经网络的局限性而提出的一种新的网络架构。卷积神经网络存在着难以识别图像中的位置关系、缺少空间分层和空间推理能力等局限性。受到神经科学的启发，人工智能领军人物 Hinton 提出了胶囊网络的概念。胶囊网络由胶囊而不是由神经元构成，胶囊由一小群神经元组成，输出为向量，向量的长度表示物体存在的估计概率，向量的方向表示物体的姿态参数。胶囊网络能同时处理多个不同目标的多种空间变换，所需训练数据量小，从而可以有效地克服卷积神经网络的局限性，理论上更接近人脑的行为。但胶囊网络也存在着计算量大、大图像处理上效果欠佳等问题，有待进一步研究。

生成对抗网络(GAN: Generative Adversarial Networks)是于 2014 年提出的一种生成模型。该算法核心思想来源于博弈论的纳什均衡，通过生成器和判别器的对抗训练进行迭代优化，目标是学习真实数据的分布，从而可以产生全新的、与观测数据类似的数据。与其他生成模型相比，GAN 有生成效率高、设计框架灵活、可生成具有更高质量的样本等优势，2016 年以来研究工作呈爆发式增长，已成为人工智能一个热门的研究方向。但GAN 仍存在难以训练、梯度消失、模式崩溃等问题，仍处于不断研究探索的阶段。

迁移学习是利用数据、任务或模型之间的相似性，将学习过的模型应用于新领域的一类算法。迁移学习可大大降低深度网络训练所需的数据量，缩短训练时间。其中，Fine-Tune 是深度迁移学习最简单的一种实现

方式，通过将一个问题上训练好的模型进行简单的调整使其适用于一个新的问题，具有节省时间成本、模型泛化能力好、实现简单、少量的训练数据就可以达到较好效果的优势，已获得广泛应用。

（三） 软件框架成为技术体系核心

当前，人工智能基础性算法已经较为成熟，各大厂商纷纷发力建设算法模型工具库，并将其封装为软件框架，供开发者使用，可以说软件框架是算法的工程实现。企业的软件框架实现有闭源和开源两种形式：苹果公司等少数企业选择闭源方式开发软件框架，目的是打造技术壁垒；目前业内主流软件框架基本都是开源化运营。本篇主要关注开源软件框架的技术特点，对闭源软件框架不做过多讨论。

1. 开源软件框架百花齐放各具特点

人工智能国际巨头企业将开源深度学习软件框架作为打造开发及使用生态核心的核心。总体来说开源软件框架在模型库建设及调用功能方面具有相当共性，但同时又各具特点。业界目前主要有深度学习训练软件框架和推断软件框架两大类别。

1) 深度学习训练软件框架

基于深度学习的训练框架主要实现对海量数据的读取、处理及训练，主要部署在 CPU 及 GPU 服务集群，主要侧重于海量训练模型实

现、系统稳定性及多硬件并行计算优化等方面的任务。目前主流的深度学习训练软件框架主要有 TensorFlow，MXNet，Caffe/2+PyTorch 等。

TensorFlow 以其功能全面，兼容性广泛和生态完备而著称。该软件框架由谷歌大脑（Google Brain）团队主要支撑，实现了多 GPU 上运行深度学习模型的功能，可以提供数据流水线的使用程序，并具有模型检查，可视化和序列化的配套模块。其生态系统已经成为深度学习开源软件框架最大的活跃社区。

MXNet 以其优异性能及全面的平台支持而著称。该软件框架是由亚马逊公司（Amazon）主导的深度学习平台，目前已经捐献到阿帕奇软件基金会（Apache）进行孵化。其主要特点包括：一是可以在全硬件平台（包括手机端）运行，提供包括 Python、R 语言、Julia、C++、Scala、Matlab 以及 Javascript 的编程接口；二是具有灵活的编程模型，支持命令式和符号式编程模型；三是从云端到客户端可移植，可运行于多 CPU、多 GPU、集群、服务器、工作站及移动智能手机；四是支持本地分布式训练，在多 CPU/GPU 设备上的分布式训练，使其可充分利用计算集群的规模优势。

Caffe/2+PyTorch 以其在图像处理领域的深耕和易用性而著称。该软件框架是由脸书公司（Facebook）主导的平台，目前 Caffe 1/2 两个项目已经合并到 PyTorch 统一维护。在图像处理领域Caffe 有着深厚的生态积累，结合 PyTorch 作为一个易用性很强的软件框架，越来越受到数据科学家的喜爱。我国很多人工智能图像处理团队选择PyTorch 作为主要工作平台。

Microsoft Cognitive Toolkit (CNTK)以其在智能语音语义领域的优势及良好性能而著称。该软件框架由微软公司于2016 年基于 MIT 协议开源，它具有速度快、可扩展性强、商业级质量高以及 C++和Python 兼容性好等优点，支持各种神经网络模型、异构及分布式计算，依托于微软的产品生态，在语音识别、机器翻译、类别分析、图像识别、图像字幕、文本处理、语言理解和语言建模等领域都拥有良好应用。

PaddlePaddle 以其易用性和支持工业级应用而著称。该软件框架是百度旗下的深度学习开源平台，是我国自主开发软件框架代表。其最大特点就是易用性，得益于其对算法的封装，对于现成算法(卷积神经网络 VGG、深度残差网络 ResNet、长短期记忆网络 LSTM 等) 的使用可以直接执行命令替换数据进行训练。非常适合需要成熟稳定的模型来处理新数据的情况。

除上之外，业界及学术界还存在着多个机器学习及深度学习软件框架，如 Scikit-learn，Theano 等。这些软件框架在其专长领域仍然发挥重要作用。但由于各软件框架的维护力量及发展思路不同，同时缺少贡献人员，导致软件框架发展水平略显滞后，存在着包括算法库扩展不及时，API 水平较低以及不支持分布式任务等问题。

2) 深度学习推断软件框架基于深度学习的推断的计算量相对训练过程小很多，但仍涉及到大量的矩阵卷积、非线性变换等运算，为了满足在终端侧限定设备性能及功耗等因素的场景下，业界也开发了众多开源的终端侧软件框架。

Caffe2go 是最早出现的终端侧推断软件框架，能够让深层神经网络在手机上高效的运行。由于终端侧的 GPU 设备性能有限，Caffe2go 是基于 CPU 的优化进行设计。TensorFlow Lite 可以运行在 Android 和 iOS 平台，结合 Android 生态的神经网络运行时能够实现较为高效的 AI 移动端应用速度。NCNN 是腾讯开源的终端侧 AI 软件框架， 支持多种训练软件框架的模型转换，是主要面向CPU 的AI 模型应用， 无第三方依赖具有较高的通用性，运行速度突出，是国内目前较为广泛使用的终端侧 AI 软件框架。Core ML 是苹果公司开发的 iOS AI 软件框架，能够对接 Caffe、PyTorch、MXNet、TensorFlow 等绝大部分 AI 模型，并且自身提供了常用的各种手机端 AI 模型组件，目前也汇集了众多开发者及贡献力量。Paddle-mobile 是百度自研的移动端深度学习软件框架，主要目的是将 Paddle 模型部署在手机端，其支持 iOS GPU 计算。但目前功能相对单一，支持较为有限。TensorRT 是英伟达（NVIDIA）开发的深度学习推断工具，已经支持 Caffe、Caffe2、TensorFlow、MXNet、PyTorch 等主流深度学习库，其底层针对NVIDIA 显卡做了多方面的优化，可以和 CUDA 编译器结合使用。

目前主要产业巨头均推出了基于自身技术体系的训练及推断软件框架，但由于目前产业生态尚未形成，深度学习模型表示及存储尚未统一，训练软件框架及推断软件框架尚未形成一一对应关系，技术生态争夺将继续持续。

2. 巨头以开源软件框架为核心打造生态

人工智能开源软件框架生态的核心，是通过使用者和贡献者之间的良好互动和规模化效应，形成现实意义的标准体系和产业生态，进而占据人工智能核心的主导地位。开源软件框架的用户包括最终服务及产品的使用者和开发者。当前开源软件框架的技术发展呈现出以下几方面的特点：

一是谷歌与其他公司间持续竞争。巨头公司在技术上将积极探寻包括模型互换，模型迁移等技术联合，以对抗谷歌公司。例如脸书

（Facebook）和微软已经合作开发了一个可互换的人工智能软件框架解决方案。二是开源软件框架在向统一和标准化方向发展。随着人工智能应用的爆发，开发人员在不同平台上创建模型及部署模型的需求愈发强烈，在各类软件框架间的模型迁移互换技术研发已经成为重点。三是更高级的 API2逐渐占据主导地位。以 Keras 为例，它是建立在TensorFlow、Theano、CNTK、MXNet 和 Gluon 上运行的高级开源神经网络库，以其高级 API 易用性而得到了广泛的使用。四是模型的集群并发计算成为业界研究热点。当前人工智能网络对于单计算节点的 算力要求过高，但当前主流开源软件框架对于模型分割进行计算并没有实现，而这个问题也将随着应用场景的不断丰富而不断引起重视，成为开源软件框架下一个核心竞争点。

（四） 编译器解决不同软硬件的适配问题

在实际工程应用中，人工智能算法可选择多种软件框架实现，训练和开发人工智能模型也可有多种硬件选项，这就开发者带来了不小

的挑战。原因一是可移植性问题，各个软件框架的底层实现技术不同，导致在不同软件框架下开发的模型之间相互转换存在困难；二是适应性问题，软件框架开发者和计算芯片厂商需要确保软件框架和底层计算芯片之间良好的适配性。解决以上两个挑战的关键技术之一就是深度神经网络模型编译器，它在传统编译器功能基础上，通过扩充面向深度学习网络模型计算的专属功能，以解决深度学习模型部署到多种设备时可能存在的适应性和可移植性问题。

1. 深度学习网络模型编译器解决适应性问题

传统编译器缺少对深度学习算法基础算子（卷积、残差网络及全连接计算等）的优化，且对人工智能多种形态的计算芯片适配缺失，针对人工智能底层计算芯片及上层软件框架进行适配优化的编译器需求强烈。目前

业界主要采用依托传统编译器架构进行演进升级的方式来解决这个问题。当前业界主流编译器主要包括英伟达公司的CUDA 编译器，英特尔公司开发的 nGraph 以及华盛顿大学团队开发的 NNVM 编译器。

目前产业界绝大多数编译器都是按照 LLVM 体系架构设计的。LLVM 全称 Low Level Virtual Machine，是一种应用广泛的开源编译器架构。该项目由伊利诺伊大学发起，由于其开源特性，目前已有基于这个软件框架的大量工具可以使用，形成了具有实际标准意义的生态。

英伟达通过提供针对 LLVM 内核的 CUDA 源代码及并行线程执行后端打造了 CUDA 编译器。该编译器可支持C、C++以及 Fortran

语言，能够为运用大规模并行英伟达 GPU 的应用程序加速。英特尔基于 LLVM 架构打造 nGraph 计算库，为深度学习提供优化方法，可以处理所有的计算芯片抽象细节，目前已经开发了 TensorFlow/XLA、MXNet 和 ONNX 的软件框架桥梁；华盛顿大学基于LLVM 架构打造了 NNVM/TVM 编译器，能够直接从多个深度学习前端将工作负载编译成为优化的机器代码。实现端到端的全面优化。

2. 中间表示层解决可移植性问题

在工程实践中，人工智能软件框架训练完成的模型将按照中间表示层的规定进行表达和存储。中间表示层（Intermediate Representation， IR）是编译器用来表示源码的数据结构或代码，可以看作是不同中间件的集合，在性能优化及通信方面有着非常重要的作用。上文介绍的LLVM 架构最大优点之一就是其有一个表达形式很好的中间表示层语言，这种模块化设计理念能够支撑各种功能扩充，三大主流深度学习网络模型编译器均是通过在中间表示层中增加专属优化的中间件 来实现功能演进创新的。

扩充性能的中间表示层是打通深度学习计算中多种不同前端训练软件框架和多种不同后端的表达桥梁，使深度学习网络模型编译器更有效实现二者之间的优化和影射。在深度学习网络模型编译器中， 中间表示层的核心思想借鉴了 LLVM 架构设计，新增加的专属中间件是解决推断侧模型运行在不同硬件平台的重要描述方法。当前深度学习网络模型编译器的中间表示层主要分为 NNVM/TVM 和TensorFlow XLA 两大阵营，但实际上类似 ONNX、NNEF 等模型交换格式也是各种对中间层表示的定义。业界共识“IR”的竞争，将是未来软件框架之争的重要一环。

3. 未来亟需模型转换及通用的模型表示

在工程实践上，除了上文提到使用统一的中间表示层对模型进行表达及存储外，输入数据格式以及模型表示规范也同样是重要的影响因素。

主流软件框架输入数据集格式各有不同。由于在训练中已经过清洗和标注的数据依然面临着多线程读取、对接后端分布式文件系统等实际操作问题，各主流人工智能软件框架均采用了不同的技术和数据集格式来实现

此类数据操作。如TensorFlow 定义了 TFRecord、MXNet 及 PaddlePaddle 使用的是 RecordIO 等。

深度学习网络模型的表示规范分为两大阵营。第一阵营是 Open Neural Network Exchange（ONNX，开放神经网络交换），是一个用于表示深度学习模型的标准，可使模型在不同软件框架之间进行转移。ONNX 由微软和 Facebook 联合发布，该系统支持的软件框架目前主要包括 Caffe2，PyTorch，Cognitive Toolkit 和 MXNet，而谷歌的TensorFlow 并没有被包含在内。第二阵营是 Neural Network Exchange Format（NNEF，神经网络交换格式），是由 Khronos Group 主导的跨厂商神经网络文件格式，计划支持包括 Torch, Caffe, TensorFlow, 等 几乎所有人工智能软件框架的模型格式转换，目前已经有 30 多家计算芯片企业参与其中。

（五） AI 计算芯片提供算力保障

现有深度神经网络需要用更短的时间、更低功耗完成计算，这就给深度学习计算芯片提出了更高的要求。

1. 深度学习对 AI 计算芯片的需求

总体来看，深度神经网络对计算芯片的需求主要有以下两个方面：一是计算芯片和存储间海量数据通信需求，这里有两个层面，一个是缓存（Cache）和片上存储（Memory）的要大，另一个是计算单元和存储之间的数据交互带宽要大。二是专用计算能力的提升，解决对卷积、残差网络、全连接等计算类型的大量计算需求，在提升运算速度的同时实现降低功耗。总的来说，AI 计算芯片的发展过程可以总结为一直在围绕如何有效解决存储与计算单元的提升这两个问题而展 开，成本问题则作为一个舵手控制着最终的航向。

2. 典型 AI 计算芯片的使用现状

在深度学习训练环节，除了使用 CPU 或 GPU 进行运算外，现场可编程门阵列（ Field——Programmable Gate Array，FPGA）以及专用集成电路（Application Specific Integrated Circuit，ASIC）也发挥了重大作用；而用于终端推断的计算芯片主要以ASIC 为主。

CPU 在深度学习训练场景下表现不佳。最初的深度学习场景是使用CPU 为架构搭建的，如最初GoogleBrain 就是基于CPU 组成的。但由于 CPU 其本身是通用计算器，大量芯片核心面积要服务于通用场景的元器件，导致可用于浮点计算的计算单元偏少，无法满足深度学习特别是训练环节的大量浮点运算需求，且并行计算效率太低，很快被具有数量众多的计算单元、具备强大并行计算能力的 GPU 代替。

GPU 成为目前深度学习训练的首要选择。GPU 的关键性能是并行计算，适合深度学习计算的主要原因一是高带宽的缓存有效提升大量数据通信的效率。GPU 的缓存结构为共享缓存，相比于 CPU，GPU 线程（Thread）之间的数据通讯不需要访问全局内存，而在共享内存中就可以直接访问。二是多计算核心提升

并行计算能力。GPU 具有数以千计的计算核心，可实现 10-100 倍于CPU 的应用吞吐量。同时，基于由 NVIDIA 推出的通用并行计算架构 CUDA，使 GPU 能够解决复杂的计算问题。其包含的 CUDA 指令集架构（ISA）以及 GPU 内部的并行计算引擎可针对当前深度学习计算进行加速，但是由于深度学习算法还未完全稳定，若深度学习算法发生大的变化，则 GPU 存在无法灵活适配问题。

FPGA 在深度学习加速方面具有可重构、可定制的特点。因 FPGA 没有预先定义的指令集概念，也没有确定的数据位宽，所以可以实现应用场景的高度定制。但FPGA 的灵活性（通用性）也意味着效率的损失。由于FPGA 应用往往都需要支持很大的数据吞吐量，这对于内存带宽和I/O 互连带宽要求很高。同时由于逻辑利用率低，引发无效功耗大。

FPGA 市场化的主要阻碍是成本高昂，价格在几十到几万美元一片不等，且应用者必须具备电路设计知识和经验。由于FPGA 省去了流片过程，在深度学习发展初期成为计算芯片主要解决方案之一， 在GPU 和ASIC 中取得了权衡，很好的兼顾了处理速度和控制能力。

ASIC（专用集成电路，Application Specific Integrated Circuit）是不可配置的高度定制专用计算芯片。ASIC 不同于 GPU 和 FPGA 的灵活性，定制化的 ASIC 一旦制造完成将不能更改，所以初期成本高、开发周期长，使得进入门槛高。但ASIC 作为专用计算芯片性能高于FPGA，相同工艺的ASIC 计算芯片比 FPGA 计算芯片快5-10 倍，同时规模效应会使得 ASIC 的成本降低。但高昂的研发成本和研发周期是未来广泛应用的阻碍。ASIC 主要生产企业包括如 Google 的TPU 系列计算芯片，以及国内的寒武纪、地平线等公司。

TPU 的核心为脉动阵列机，其设计思想是将多个运算逻辑单元（ALU）串联在一起，复用从一个寄存器中读取的结果，从而有效平衡了运算和 I/O 的需求。但其只适合做信号处理的卷积、信号和图像处理（signal and image processing），矩阵算术（matrix arithmetic） 和一些非数值型应用（non-numeric application）。

另一类 ASIC 代表企业为国内寒武纪，其 DianNao 系列核心思想为结合神经网络模型的数据局部性特点以及计算特性，进行存储体系以及专用硬件设计，从而获取更好的性能加速比以及计算功耗比。

（六） 数据为算法模型提供基础资源

基于深度学习的人工智能技术，核心在于通过计算找寻数据中的规律，运用该规律对具体任务进行预测和决断。源数据需要进行采集、标注等处理后才能够使用，标注的数据形成相应数据集。业务类型主要包括数据采集、数据处理、数据存储以及数据交易等环节。

当前，人工智能数据集的参与主体主要有以下几类:一是学术机构，为开展相关研究工作，自行采集、标注，并建设学术数据集。这类数据集以 ImageNet 为代表，主要用于算法的创新性验证、学术竞赛等，但其迭代速度较慢，难用于实际应用场景。二是政府，等中立机构，他们以公益形式开放的公共数据，主要包括政府、银行机构等行业数据及经济运行数据等，数据标注一般由使用数据的机构完成。三是人工智能企业，他们为开展业务而自行建设数据集，企业一般自行采集，标注形成自用数据集，或采购专业数据公司提供的数据外包服务。四是数据处理外包服务公司，这类公司业务包括出售现成数据，训练集的使用授权，或根据用户的具体需求提供数据处理服务（用户提供原始数据、企业对数据进行转写、标注），具体业务服务形式包括且不限于提供数据库资源、提供数据采集服务，提供数据转写标注服务等。

当前，人工智能基础数据类型主要包括语音语言类（包括声音、文字、语言学规则）、图像识别类（包括自然物体、自然环境、人造物体、生物特征等）以及视频识别类三个大类，从世界范围来看，数据服务商总部主要分布在美国、欧洲等发达国家。但其数据处理人员则大多数分布在第三世界国家；我国语音、图像类资源企业机构正处于快速发展阶段，为产业发展增添了动力。

（七） 高性能计算服务器和服务平台快速发展

深度学习使用GPU 计算具有优异表现，催生了各类GPU 服务器， 带动了 GPU 服务器的快速发展；同时，也带动了以服务的形式提供人工智能所需要的能力，如深度学习计算类的计算平台，以及语音识别，人脸识别等服务，这也成为人工智能企业打造生态的重要抓手。

1. GPU 服务器服务器厂商相继推出了专为 AI 而设计的、搭载 GPU 的服务器。GPU 服务器是基于GPU 应用于视频编解码、深度学习、科学计算等多种场景的计算服务设备。GPU 服务器为 AI 云场景对弹性配置能力予以优化，以增强 PCI-E 拓扑和数量配比的弹性，增加适配多种软件框架的运算需求，可以支持 AI 模型的线下训练和线上推理两类场景， 能够让 AI 模型训练性能最大化或 AI 在线推断效能最大化，一般分为 4 卡，8 卡，10 卡等多种类型。

另外，英伟达等公司推出了专用的 GPU 一体机。例如DGX-1 系列深度学习一体机，采用定制的硬件架构，并使用 NVlink 提升了CPU、GPU 以及内存之间的通信速度和带宽；同时搭载了集成了NVIDIA 开发的操作系统，NVIDIA docker 环境和很多常用的 Framework 的Docker 镜像，实现了从底层硬件到上层软件的紧密耦合。类似的产品还有浪潮的 AGX-1 系列服务器。

2. 以服务的形式提供人工智能能力成为趋势

为了解决企业自行搭建 AI 能力时遇到的资金、技术和运维管理等方面困难，人工智能企业纷纷以服务的形式提供 AI 所需要的计算资源、平台资源以及基础应用能力。这类服务的意义在于一是有效推动社会智能化

水平的提升，降低企业使用人工智能的成本，推动人工智能向传统行业融合。二是人工智能服务化转型的重要基础。服务平台使人工智能服务和应用不再封装于具体产品中，而可以在以线、随用随取的服务形式呈现。三是服务平台成为垂直行业落地的重要基础。近两年，教育、医疗、金融等传统行业对人工智能相关技术和应用需求的不断提升，而服务平台是解决技术和应用的基础。

以服务形式提供人工智能服务主要有两类，即平台类的服务和软件 API 形式的服务。平台类服务主要包含 GPU 云服务，深度学习平台等，类似云服务的基础设施即服务（Infrastructure as a Service，IaaS）和平台即服务（Platform as a Service，PaaS）层。GPU 云服务是以虚拟机的形式，为用户提供 GPU 计算资源，可适用于深度学习、科学计算、图形图像渲染、视频解码等应用场景。

深度学习平台是以TensorFlow、Caffe、MXNet、Torch 等主流深度学习软件框架为基础， 提供相应的常用深度学习算法和模型，组合各种数据源、组件模块，让用户可以基于该平台对语音、文本、图片、视频等海量数据进行离线模型训练、在线模型预测及可视化模型评估。软件 API 服务主要分为智能语音语类服务和计算机视觉服务。其中智能语音语类服务主要提供语音语义相关的在线服务，可包括语音识别、语音合成、声纹识别、语音听转写等。计算机视觉类服务主要提供物体检测、人脸识别、人脸检测、图像识别、光学字符识别（Optical Character Recognition， OCR）识别、智能鉴黄等服务。

三、 基于深度学习的基础应用技术现状

目前随着深度学习算法工程化实现效率的提升和成本的逐渐降低，一些基础应用技术逐渐成熟，如智能语音，自然语言处理和计算机视觉等，并形成相应的产业化能力和各种成熟的商业化落地。同时，业界也开始探索深度学习在艺术创作、路径优化、生物信息学相关技术中的实现与应用，并已经取得了瞩目的成果。

本章主要分析目前商业较为成熟的智能语音、自然语言处理和计算机视觉技术的情况，如图 2 所示，每个基础应用技术各分为若干应用类别。

（一） 智能语音技术改变人机交互模式

智能语音语义技术主要研究人机之间语音信息的处理问题。简单来说，就是让计算机、智能设备、家用电器等通过对语音进行分析、理解和合成，实现人“能听会说”、具备自然语言交流的能力。

1. 智能语音技术概述

按机器在其中所发挥作用的不同，分为语音合成技术、语音识别技术、语音评测技术等。语音合成技术即让机器开口说话，通过机器自动将文字信息转化为语音，相当于机器的嘴巴；语音识别技术即让机器听懂人说

话，通过机器自动将语音信号转化为文本及相关信息， 相当于机器的耳朵；语音评测技术通过机器自动对发音进行评分、检错并给出矫正指导。此外，还有根据人的声音特征进行身份识别的声纹识别技术，可实现变声和声音模仿的语音转换技术，以及语音消噪和增强技术等。

2. 智能语音产品和服务形态多样

智能语音技术会成为未来人机交互的新方式，将从多个应用形态成为未来人机交互的主要方式。

智能音箱类产品提升家庭交互的便利性。智能音箱是从被动播放音乐，过渡到主动获取信息、音乐和控制流量的入口。当前智能音箱以语音交互技术为核心，成为作为智能家庭设备的入口，不但能够连接和控制各类智能家居终端产品，而且加入了个性化服务，如订票、查询天气、播放音频等能力。

个人智能语音助手重塑了人机交互模式。个人语音助手，特别是嵌入到手机、智能手表、个人电脑等终端中的语音助手，将显著提升这类产品的易用性。如苹果虚拟语音助手Siri 与苹果智能家居平台Homekit 深度融合，用户可通过语音控制智能家居。Google Now 为用户提供关心的内容，如新闻、体育比赛、交通、天气等等。微软的Cortana 主要优势在于提升个人计算机的易用性。

以 API 形式提供的智能语音服务成为行业用户的重要入口。智能语音 API 主要提供语音语义相关的在线服务，可包括语音识别、语音合成、声纹识别、语音听转写等服务类型，并且可以嵌入到各类产品，服务或 APP 中。在商业端，智能客服、教育（口语评测）、医疗（电子病历）、金融（业务办理）、安防、法律等领域需求强烈；在个人用户领域，智能手机、自动驾驶及辅助驾驶、传统家电、智能家居等领域需求强烈。

（二） 计算机视觉技术已在多个领域实现商业化落地

计算机视觉识别这一人工智能基础应用技术部分已达商业化应用水平，被用于身份识别、医学辅助诊断、自动驾驶等场景。

1. 计算机视觉概述

一般来讲，计算机视觉主要分为图像分类、目标检测、目标跟踪和图像分割四大基本任务。

图像分类是指为输入图像分配类别标签。自2012 年采用深度卷积网络方法设计的 AlexNet 夺得 ImageNet 竞赛冠军后，图像分类开始全面采用深度卷积网络。2015 年，微软提出的 ResNet 采用残差思想，将输入中的一部分数据不经过神经网络而直接进入到输出中，解决了反向传播时的梯度弥散问题，从而使得网络深度达到 152 层，将

错误率降低到 3.57%，远低于 5.1% 的人眼识别错误率，夺得了ImageNet 大赛的冠军。2017 年提出的 DenseNet 采用密集连接的卷积神经网络，降低了模型的大小，提高了计算效率，且具有非常好的抗过拟合性能。

目标检测指用框标出物体的位置并给出物体的类别。2013 年加州大学伯克利分校的 Ross B. Girshick 提出 RCNN 算法之后，基于卷积神经网络的目标检测成为主流。之后的检测算法主要分为两类，一是基于区域建议的目标检测算法，通过提取候选区域，对相应区域进行以深度学习方法为主的分类，如 RCNN、Fast-RCNN、Faster-RCNN、SPP-net 和 Mask R-CNN 等系列方法。二是基于回归的目标检测算法，如 YOLO、SSD 和DenseBox 等。

目标跟踪指在视频中对某一物体进行连续标识。基于深度学习的跟踪方法，初期是通过把神经网络学习到的特征直接应用到相关滤波或 Struck 的跟踪框架中，从而得到更好的跟踪结果，但同时也带来了计算量的增加。最近提出了端到端的跟踪框架，虽然与相关滤波等传统方法相比在性能上还较慢，但是这种端到端输出可以与其他的任务一起训练，特别是和检测分类网络相结合，在实际应用中有着广泛的前景。

图像分割指将图像细分为多个图像子区域。2015 年开始，以全卷积神经网络（FCN）为代表的一系列基于卷积神经网络的语义分割方法相继提出，不断提高图像语义分割精度，成为目前主流的图像语义分割方法

2. 计算机视觉技术应用领域广阔

在政策引导、技术创新、资本追逐以及消费需求的驱动下，基于深度学习的计算机视觉应用不断落地成熟，并出现了三大热点应用方向。

一是人脸识别抢先落地，开启“刷脸”新时代。目前，人脸识别已大规模应用到教育、交通、医疗、安防等行业领域及楼宇门禁、交通过检、公共区域监控、服务身份认证、个人终端设备解锁等特定场景。从2017 年春运，火车站开启了“刷脸”进站，通过摄像头采集旅客的人脸信息，与身份证人脸信息进行验证；2017 年 9 月苹果公司发布的iPhone X 第一次将 3D 人脸识别引入公众视线，迅速引发了“移动终端+人脸解锁”的布局风潮。

二是视频结构化崭露头角，拥有广阔应用前景。视频结构化就是将视频这种非结构化的数据中的目标贴上相对应的标签，变为可通过某种条件进行搜索的结构化数据。视频结构化技术的目标是实现以机器自动处理为主的视频信息处理和分析。从应用前景看，视频监控技术所面临的巨大市场潜力为视频结构化描述提供了广阔的应用前景，很多行业需要实现机器自动处理和分析视频信息，提取实时监控视频或监控录像中的视频信息，并存储于中心数据库中。用户通过结构化视频合成回放，可以快捷的预览视频覆盖时间内的可疑事件和事件发生时间。

三是姿态识别让机器“察言观色”，带来全新人机交互体验。在视觉人机交互方面，姿态识别实际上是人类形体语言交流的一种延伸。它的主要方式是通过对成像设备中获取的人体图像进行检测、识别和跟踪，并对人体行为进行理解和描述。从用户体验的角度来说，融合姿态识别的人机交互能产品够大幅度提升人机交流的自然性，削弱人们对鼠标和键盘的依赖，降低操控的复杂程度。从市场需求的角度来说，姿态识别在计算机游戏、机器人控制和家用电器控制等方面具有广阔的应用前景，市场空间十分可观。

（三） 自然语言处理成为语言交互技术的核心

自然语言处理（Natural Language Processing ，NLP）是研究计算机处理人类语言的一门技术，是机器理解并解释人类写作与说话方式的能力，也是人工智能最初发展的切入点和目前大家关注的焦点。

1. 自然语言处理技术现状

自然语言处理主要步骤包括分词、词法分析、语法分析、语义分析等。其中，分词是指将文章或句子按含义，以词组的形式分开，其中英文因其语言格式天然进行了词汇分隔，而中文等语言则需要对词组进行拆分。词法分析是指对各类语言的词头、词根、词尾进行拆分，各类语言中名词、动词、形容词、副词、介词进行分类，并对多种词义进行选择。语法分析是指通过语法树或其他算法，分析主语、谓语、宾语、定语、状语、补语等句子元素。语义分析是指通过选择词的正确含义，在正确句法的指导下，将句子的正确含义表达出来。

2. 自然语言处理技术的应用方向

自然语言处理的应用方向主要有文本分类和聚类、信息检索和过滤、信息抽取、问答系统、机器翻译等方向。其中，文本分类和聚类主要是将文本按照关键字词做出统计，建造一个索引库，这样当有关键字词查询时，可以根据索引库快速地找到需要的内容。此方向是搜索引擎的基础。信息检索和过滤是网络瞬时检查的应用范畴，在大流量的信息中寻找关键词，找到后对关键词做相应处理。信息抽取是为人们提供更有力的信息获取工具，直接从自然语言文本中抽取事实信息。机器翻译是当前最热门的应用方向，目前微软、谷歌的新技术是翻译和记忆相结合，通过机器学习，将大量以往正确的翻译存储下来。谷歌使用深度学习技术，显著提升了翻译的性能与质量。

四、 问题和趋势展望

（一） 主要问题

在算法层面，深度学习算法模型存在可靠性及不可解释性问题。首先是可靠性问题，深度学习模型离开训练使用的场景数据，其实际效果就会降低。由于训练数据和实际应用数据存在区别，训练出的模型被用于处理未学习过的数据时，表现就会降低。其次是不可解释性问题，深度学习计算过程为黑盒操作，模型计算及调试的执行规则及特征选取由机器自行操作，目前尚无完备理论能够对模型选取及模本身做出合理解释，随着相关算法在实际生产生活中的融合应用，存在产生不可控结果的隐患。

在数据层面，主要存在流通不畅、数据质量良莠不齐和关键数据集缺失等问题。

具体来看，一是数据流通不畅。目前人工智能数据集主要集中在政府和大公司手里，受制于监管、商业门槛等问题，数据无法有效流动；部分有价值数据，如监控、电话客服等数据目前没有合法渠道获得；二是数据质量良莠不齐。数据标注主要通过外包形式，劳动力水平决定了产出的标注数据质量。三是关键领域和学术数据集不足。计算机视觉、自然语言处理等领域的数据资源严重不足，同时目前我国产业数据主要供给给产业界，目前学术界数据集数量较少， 可能影响科研及前瞻性的技术研究。

在软件框架层面，实现深度学习应用落地的推断软件框架质量参差不齐，制约了业务开展。由于深度学习应用场景众多，相关应用呈现碎片化特点，用于实现最后应用落地的开源推断软件框架无论在功能还是性能层面距离实际需求还存在相当距离，与训练软件框架趋同趋势不同，产业界所使用的推断软件框架需要聚力研发，尚未形成具有实际标准意义的优秀实例。

在编译器层面，各硬件厂商的中间表示层之争成为技术和产业发展的阻碍。目前业界并没有统一的中间表示层标准，并且模型底层表示、存储及计算优化等方面尚未形成事实标准，导致各硬件厂商解决方案存在一定差异，导致应用模型迁移不畅，提高了应用部署难度。

在 AI 计算芯片层面，云侧和终端侧对计算芯片提出了不同的要求。对于云侧芯片，随着深度学习计算需求的逐渐增加，业界希望在提升云侧芯片运算效能的前提下，希望针对不同网络实现更优化的性能表现，而功耗比则不是首要关注的因素；对于终端侧芯片，在功耗为首要要求的情况下，更加注重的推断运算的性能，并且不同终端应用场景对芯片提出了更多个性化需求，如在人脸识别摄像头、自动驾驶汽车等场景。

（二） 趋势展望

迁移学习的研究及应用将成为重要方向。迁移学习由于侧重对深度学习中知识迁移、参数迁移等技术的研究，能够有效提升深度学习模型复用性，同时对于深度学习模型解释也提供了一种方法，能够针对深度学习算法模型可靠性及不可解释性问题提供理论工具。

深度学习训练软件框架将逐渐趋同，开源推断软件框架将迎来发展黄金期。随着人工智能应用在生产生活中的不断深入融合，对于推断软件框架功能及性能的需求将逐渐爆发，催生大量相关工具及开源推断软件框架，降低人工智能应用部署门槛。

中间表示层之争将愈演愈烈。以计算模型为核心的深度学习应用， 由于跨软件框架体系开发及部署需要投入大量资源，因此模型底层表 示的统一将是业界的亟需，未来中间表示层将成为相关企业的重点。

AI 计算芯片朝云侧和终端侧方向发展。从云侧计算芯片来看， 目前 GPU 占据主导市场，以 TPU 为代表的 ASIC 只用在巨头的闭环生态，未来 GPU、TPU 等计算芯片将成为支撑人工智能运算的主力器件，既存在竞争又长期共存，一定程度可相互配合；FPGA 有望在数据中心中以 CPU+FPGA 形式作为有效补充。从终端侧计算芯片来看，这类芯片将面向功耗、延时、算力、特定模型、使用场景等特定需求，朝着不同发展。

来源： 物联网报告中心