人工智能领域研究

上海人大人科技创新服务有限公司 主办

2017 年 3 月 6 日　　第 3 期　　（总第 3 期）

本期主题

2017 MIT 人工智能 5 大趋势预测（三）

百家评说

人工智能的发展未来与创业　　胡郁　科大讯飞执行总裁

通讯地址：上海市南京东路 800 号
新一百大厦 17 楼

联系人：陈海燕

联系方式：chenhy@chinardr.net

本期主题

2017 MIT 人工智能 5 大趋势预测（三）

技术奇点（technological singularity）是一个根据技术发展史总结出的观点，认为未来将要发生一件不可避免的事件——技术发展将会在很短的时间内发生极大而接近于无限的进步。当此转捩点来临的时候，旧的社会模式将一去不复返，新的规则开始主宰这个世界。

50 多年来，（希望模仿人类大脑的思考操作的）人工智能（Artificial Intelligence）经历了“爆发到寒冬再到野蛮生长”的历程，伴随着人机交互、机器学习、模式识别等人工智能技术的提升，机器人与人工智能成了这一技术时代的新趋势。关于人工智能的各级规划、各种预测，成为一股新的策划趋势。

本期，我们结合 MIT Technology Review 最近发布的 2017 年人工智能的五大趋势预测，探索 2017 年人工智能的发展路径和方向。

4 趋势四：语言学习 (Language learning)

人们希望，有朝一日，计算机可以通过语言与我们交流和互动。AI 研究人员正致力于提升语音和图像识别等领域的技术，帮助计算机更好的理解语言的上下文含义（从而更有效地分析和生成语言），进而对自己的决策行为作出说明（可以反过来给予科学家更多的灵感），提升 AI 系统的实用性。

例 1：AlphaGo 与李世石世纪大战的第二盘第 37 步的困惑

AlphaGo 与李世石世纪大战的第二盘第 37 步，机器选择了一个不同寻常的落子点。DeepMind 当时只能看到 AlphaGo 的实时胜率预判，并不理解 AlphaGo 当时这一不同寻常的选择的含义（事后，经过几天的系统分析，才理解其选择的含义）。如果 DeepMind 落实“将 AlphaGo 的决策系统开源出来，找到（智能助理的改良、当作医疗诊断的工具等）可商业化项目”的考虑，则该系统能否“使用人类的语言，向（医护人员等）相关人员解释他们作出的决策的依据”，在（医疗诊断等）实际业务过程中显得更为重要。

但鉴于语言的复杂性、微妙性、多语种歧义性，短时间内，用户和智能手机还不能进行深入和有意义的对话，但语音识别和语音接口在技术和应用场景方面均较为成熟，（谷歌助理、亚马逊 Alexa 等）一些令人印象深刻的进步正在进行。

4.1Google Home 与亚马逊 Echo 的正面交锋

谷歌在为智能手机、平板、智能手表配备谷歌助理之后，（2016.10.04）进一步推出内置了谷歌助理的无线音箱 Google Home（售价 129 美元），完善并吸引用户进入其智能生态圈。这款由 Chromecast 团队主导开发的谷歌智能生活入口设备，像一个随时待命的具象化的虚拟助理，能够调用谷歌搜索以及其他应用程序，用户通过语音指令，控制它执行播放音乐、关闭房间照明、回答知识性问题、查询交通状况、更改预约等任务。（承载着谷歌在物联网和智能家居领域新希望的）该产品和（亚马逊广受欢迎的智能音箱）Echo 均主打语音控制、人工智能助理、将各类用户常用的第三方生活场景应用接入自身智能生态圈，因而成为直接对手，但两者在语音接收处理、功能略有不同。

1、外观和语音接收方面的差别

亚马逊 Echo 是一个黑色的柱状音箱，同时配有一个内置麦克风的无线遥控器。音箱（内置的 7 个麦克风接收器组成的矩阵）使用音波聚束技术探测远场声音、配合增强的噪音消除技术，使得 Echo 即使在播放音乐时也能听清用户的语音提问指令；当用户所在位置的语音指令不能被 Echo 接收到时，无线遥控器就显得非常便利。

Google Home 使用了“花瓶型”的更圆润而精致线条设计，该机身顶部斜面（隐藏了四种颜色 LED 灯的）可触控表面显示当前的音量级别（当有用户语音指令处理时，4 色的 LED 灯就会亮起），背面设置的“关闭麦克风”按钮（同时加入了手指在按钮上旋转来控制音量的操作）可以用来暂停或播放音乐，（音箱）底部（多种颜色和材质）的扬声器格栅使用磁铁吸附。此外，Google Home 使用定制的 AC 电源（取代 USB 电源）保证内置 3 英寸扬声器的声音足够持续稳定填满整个房间。

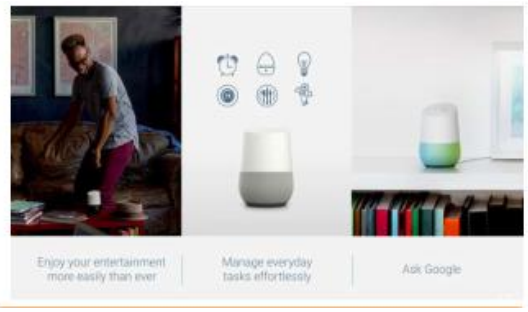

图 1：Google Home 产品宣传图（左图）、Google Home 和亚马逊 Echo 外观对比（右图）（资料来源：谷歌 2016 I/O 大会现场照片、Wired 及 Engadget 等）

Echo 与 Google Home 最关键的区别在于麦克风的数量与阵列：Echo 使用了 7 个麦克风的结构，而 Google Home 只有 2 个（原理上，麦克风越多，越能收集到来自不同方位的远场声音，并从环境噪音中识别出用户指令，例如 Echo 的远程声音识别）。谷歌表示，他们通过云端机器学习算法（例如自然语意处理）对 2 个麦克风进行的调试显示：他们使用的 2 个麦克风能达到（Echo）7 麦克风相同的效果。此外，外媒测评显示，同一间屋子里的几个不同的 Google Home 可以同时响应用户语音指令（例如，同时播放歌曲），这是谷歌从一开始就设计的多房间支持(Multi-room capability)。

图 2：Google Home 2 个麦克风设计（左图）、Echo 的 7 个麦克风矩阵：远程语音识别（右图）（资料来源：Wired、Engadget 等 ）

2、功能方面的差别

亚马逊 Echo 最精明的地方是可以出现在第三方设备和服务中：（1）不仅可以在亚马逊平台上购物和播放 Prime 音乐，还可以让用户选择 Pandora、Spotify 等娱乐，购买达美乐披萨外卖、获得 Yelp 点评的功能、Uber 打车服务等；（2）在智能家居应用方面，与三星、飞利浦、Belkin、Ecobee 等合作，将他们的智能家居设备整合到 Echo 的控制系统中。也正是因为亚马逊基于其电商基础，将 Echo 做成 Prime 电商服务的语音入口，使用户在 Echo 上可以要求 Alexa 重新下单已经购买过的商品、为用户推荐亚马逊 Prime 类别下的各类商品（并由亚马逊管理配送，在 2 天内送达），所以吸引了更多用户在亚马逊上购物及参与成为 Prime 会员。Slice Intelligence 的报告显示：Echo 用户都是“亚马逊重度消费者”，他们比非 Echo 用户在亚马逊上的花费多 7%（这也给了亚马逊更多的用户消费数据，从而提高消费者体验）。

谷歌根据其收集的（用户每日的日程安排、地图搜索、邮件收发等行为等）用户日常生活行为习惯数据，为用户提供更多网络服务内容，并致力于智能家居

方面的拓展*（由于谷歌在用户消费数据上无法与亚马逊相比，Google Home 目前暂不支持软件内支付，因此目前无法通过 Home 进行网购消费）*：（1）谷歌遍布全世界的（音频、视频）网络服务内容给用户带来更多的可选性：①谷歌 2015 年 1 月推出的 Google Cast 软件，整合了（自己的）Play、Spotify、Pandora、iHeart 广播等音乐服务，不仅为用户提供了丰富的音乐内容，还便于使用 iOS 和 Android 设备的用户将手机中的音乐推送到 Home 中播放；②用户可以指挥（由 Chromecast 团队主导开发的）Home 搜索播放 YouTube、Netflix 上的视频，并通过安装了 Chromecast 的电视屏幕自动播放出来，从而使得 Chromecast 与电视屏幕成为 Google Home 的一个可视化的界面。（2）在智能家居应用方面，谷歌计划率先将（其拥有的智能家居市场最受关注品牌）Nest 旗下的（包括智能学习恒温器、烟雾探测器、智能监控摄像头等）器件整合进 Home 智能家居系统平台，并与飞利浦 Hue、IFTTT、三星旗下的 SmartThings 平台等建立了合作（希望在年内指导更多的第三方厂商将智能家居设备和应用整合到谷歌助理中）。

虽然两家公司都在建立自己的智能生态圈，但是，布局先人一步的亚马逊很可能依靠 Echo 的明星效应和 Alexa 的开源布局抢滩智能家居市场。市场调研机构 CIRP 的统计显示：Echo 自 2014 年底推出以来，已经卖出了约 510 万台（其中，2016 年前九个月卖出约 200 万台），其新近推出的 Echo Dot 和 Echo Tap（这两者比传统 Echo 更小更便宜）在过去六个月贡献了至少 33％的销售额。

图 3：内置 Alexa 的 LG 智能冰箱（左图）及其配置的 29 寸触摸屏（右图）（资料来源：Engadget、CNET）

4.2 谷歌的 Allo 智能回复

谷歌在将智能回复应用到邮件服务 Gmail / Inbox 中后，进一步将该功能应用到聊天软件而推出 Allo：该聊天软件可以先通过对用户的对话记录来生成“标准化”智能回复选项，再在随后的不断学习“用户的个人说话方式”过程中（更优

化地理解用户的对话语义）逐渐生成“私人定制”的智能回复。

1、生成“标准化”智能回复

Allo 团队使用了一个类似“编码－解码”两步模型的方法：（1）首先使用一个递归神经网络将对话语句一个词一个词进行编码生成对应口令(token)，然后口令进入长短期记忆神经网络(Long-short term memory, LSTM)生成一个连续向量，随后这个连续向量进一步通过 softmax 模型生成一个（包含一组可以用来回复的可选择单词组的）离散语义结构(discretized semantic class)。（2）接着使用第二个递归神经网络从可选择单词组中挑出最合适的回复、并让离散语义结构进入长短期记忆神经网络(LSTM)（一次一个口令的）生成完整的回复消息，然后解码成为自然语义单词。如下图所示，当提问句为“Where are you?”时，神经网络会将问句三个单词生成 3 个口令，然后进行下一步处理；经过第二个递归神经网络中的长短期记忆神经网络(LSTM)处理，系统生成了对刚才“Where are you?”提问的回答“I’m at work”。

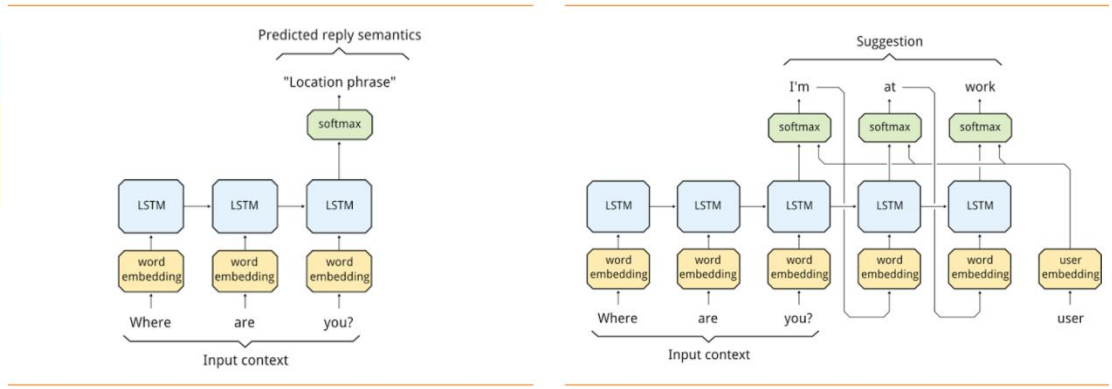

图 4：神经网络将问句三个单词生成 3 个口令（左图）、谷歌语音识别神经网络的输出示意图（右图）（资料来源：谷歌研究所官方博客）

Allo 团队提及，为了保证软件及时（不使用户失去使用的耐心）生成长度适中（如果过长就不能适应手机屏幕，如果过短会造成可用性不强）的回复选项，他们（1）将模型第一部分中的 softmax 算法改成分层式 softmax 算法（即，将对可选择单词组的遍历从单词列表遍历改为了单词树遍历），成功将递归神经网络的延迟时间从 0.5 秒缩短到 200 毫秒以下；（2）在第二部分使用定向搜索(beam search)技术（该技术一般用来对搜索域中最优解进行向下拓展的启发式搜索）挑选离散语义结构所包含的可选单词组、并将定向搜索算法的倾向调整为去搜索使

用效率更高的单词组路径，提高了回复单词组的选择效率、保证了生成的回复选项长度适中。

2、“私人定制”回复及多语言关联

Allo 团队表示，他们在神经网络训练中添加了“用户嵌入”(user embedding)项学习用户的“说话风格”，使用了 L-BFGS(Limited-memory Broyden Fletcher Goldfarb Shanno 或在受限内存时的拟牛顿算法)迅速生成海量“用户嵌入”数据，使得 Allo 的智能回复会随着用户的使用时间增加而更加反映用户的说话习惯。例如，当用户在回答“How are you?”时习惯使用“Fine”而不是“I'm good”，Allo 会把这些习惯添加到神经网络中，把“说话风格”作为神经网络的一个参数项来进行回复推荐（如下图所示）。

此外，开发团队使用基于（半监督学习(Semi-supervised)语义理解的）图表关联(graph-based)的机器学习技术进行多语言之间的相互关联，并连接了谷歌机器翻译团队的模型来进行单词翻译，使得 Allo 的智能回复适用于所有语言。

（注：半监督学习(Semi-supervised Learning)技术是监督式学习(Supervised 与非监督学习(Unsupervised Learning)相结合的一种学习方法，它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。）

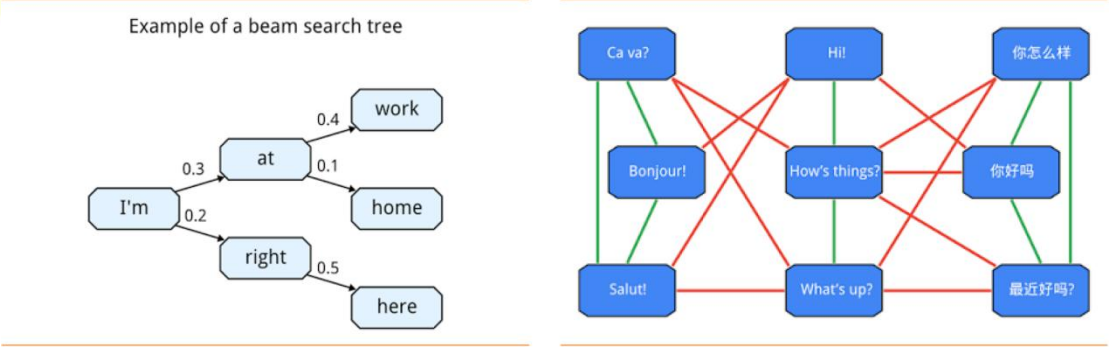

图 5：神经网络将问句三个单词生成 3 个口令（左图）、谷歌语音识别神经网络的输出示意图（右图）（资料来源：谷歌研究所官方博客）

4.3 神经机器翻译系统

1、翻译基础

10 年前，谷歌发布了谷歌翻译（该系统的核心算法是基于短语的机器翻译技术(Phrase-Based Machine Translation, PBMT)）并随即将递归神经网络 RNN 加入机器翻译中，直接学习输入端（一种语言的一个句子）到输出端（另一种语言的

同一句子）的映射，实现了机器翻译，但（这种将句子中的词和短语拆分进行独立翻译的做法）却很容易出现罕见词不识别、上下文意不通的情况。

最近，谷歌发布了谷歌神经机器翻译(Google Neural Machine Translation system, GNMT)，该系统使用的神经机器翻译系统(NMT)将整个句子视作翻译的基本输入单元（*NMT 相对于 PBMT 的优势在于能够减少工程设计。并且，随着 NMT 的不断改进，研究人员又加入了外部对准模型(External Alignment Model)来标记罕见词*），进行直接的端到端训练，实现了机器翻译技术迄今为止的最大进步。谷歌翻译、有道翻译、百度翻译分别对“小偷偷偷偷东西”的翻译结果显示：基于句子的（谷歌）机器翻译优于基于短语的（有道、百度）机器翻译。

图 6：谷歌翻译、有道翻译、百度翻译实例对比（资料来源：各翻译软件）

2、翻译机制

谷歌神经机器翻译系统使用了深度长短期记忆神经网络 LSTM，该神经网络由 8 个编码器和 8 个解码器组成、使用注意链接(attention connections)和残差连接(residual connections)连接编码器与解码器。例如，汉英翻译时，系统先将输入的汉语句子的词编码成一个向量列表（其中每个向量都表征了到目前为止所有被读取到的词的含义(即编码器)），读取完整个句子后，解码器重点“注意”与生成英语词最相关的编码的汉语向量的权重分布，一次生成英语句子的一个词（即解码器）。

此外，在该翻译系统中，（1）注意连接机制将解码器的底层连接到了编码器的顶层，提升了并行性并降低训练时间；（2）谷歌在推理运算时使用低精度算法，增加了最终翻译速度；（3）谷歌将词组分为由常见词组成的子词单元(sub-word units)的有限集合，同时作为输入和输出内容，有效平衡了“字符(character)”限定模型(delimited models)的灵活性与“词(word)”限定模型的有效性，自然地处

理了罕见词翻译，进而提升了整体翻译质量。

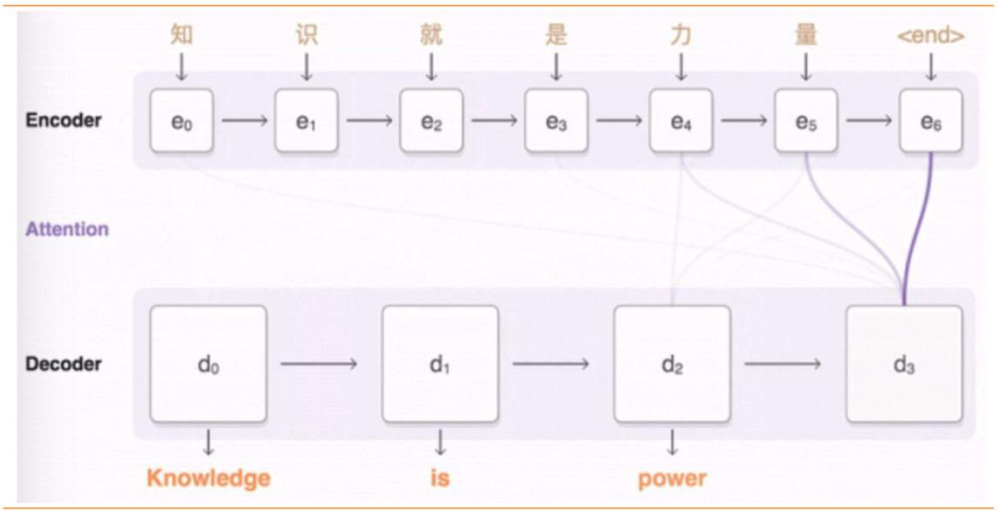

图 7：谷歌神经机器翻译系统(GNMT)翻译机制（资料来源：谷歌研究所官方博客）

3、翻译系统的适用范围

此次由 Google Brain 和谷歌翻译团队共同开发完成的系统，使用了谷歌的开源机器学习平台 TensorFlow 和张量处理单元 TPU，保证了系统的计算能力和严格的延迟要求。测试结果表明：新系统在多个主要语言的翻译中将翻译误差降低了 55%-85%以上。特别是在英语到西班牙语的翻译中（满分 6 分），新系统的平均得分（5.43 分）与人类翻译的平均得分（5.55 分）相差无几。目前，谷歌翻译的汉英翻译已经在使用这套系统完成所有的翻译请求（约 1800 万条/天），未来几个月，谷歌将会把 GNMT 扩展到更多的语言翻译上。

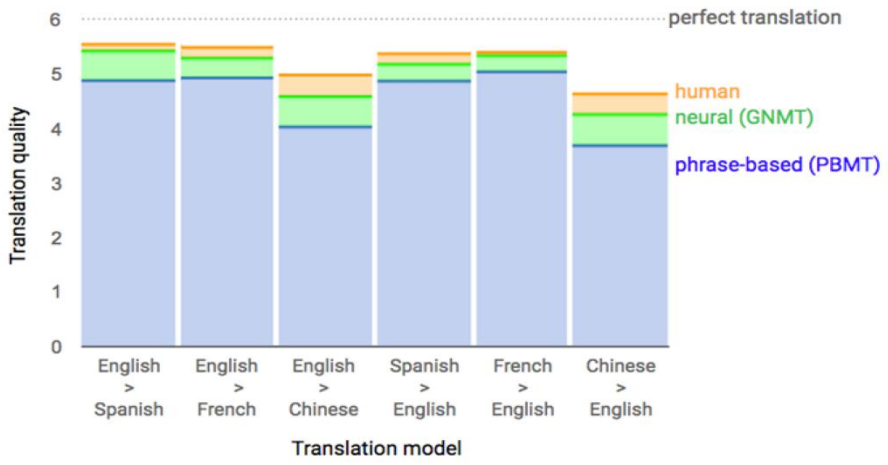

图 8：满分 6 分记，人类翻译、谷歌神经翻译与 PBMT 的得分对比（资料来源：谷歌研究所官方博客）

不过，Google Brain 的成员同时表示，由于人类语言在不断进步、不断出现新生词汇，无法在任何情况下都使用机器翻译替换人类翻译，但在（论文、科技文献等）结构化比较高、（新闻时事短讯等）写作思路比较固定且读者不太关注文笔而更注重信息传达的文章中，（基于已经出现过的语言现象的）机器翻译能够非常快的提高翻译精确度。

百家评说

人工智能的发展未来与创业

胡郁　科大讯飞执行总裁

1 人工智能的前世今生

人工智能这一话题，最早可以回溯到 1946 年世界上第一台电子计算机 ENIAC 的诞生。ENIAC 产生以后，很多计算机科学家对于计算机将来能够代替人类做什么事情有很多联想，其中最著名的一个人是图灵，他在 1950 年左右在人工智能领域进行了很多的探讨，并且提出了著名的“图灵测试”。而“人工智能”一词真正被提出来，是在 1956 年 Dartmouth 的会议上，由四位图灵奖得主、信息论创始人和一位诺贝尔奖得主一起将“人工智能”定义出来，包括明斯基、西蒙、麦卡塞等人，这次会议也被公认为人工智能研究的出生典礼。

人工智能 (Artificial Intelligence) 是指，能够和人一样进行感知、认知、决策、执行的人工程序或系统。然而，人工智能发展的 60 年不是一帆风顺的，起起伏伏共经历了三次浪潮。

（1）1970 年第一次黄金期。自从 Dartmouth 会议以后，人们陆续发明了第一款感知神经网络软件和聊天软件，那时大家都惊呼“人工智能来了，再过十年机器要超越人类了”。不过，很快到了 70 年代后期，人们发现过去的理论和模型，只能解决一些非常简单的问题，很快人工智能进入了第一次的冬天。

（2）1990 年第二次黄金期。随着 1982 年 Hopfield 神经网络和 BP 训练算法的提出，大家发现人工智能的春天又来了。80 年代又兴起一波人工智能的热潮，包括语音识别、语音翻译以及日本提出的第五代计算机。不过，到了 90 年代后期，人们发现这种东西离我们的实际生活还很遥远。比如 IBM 在 90 年代时提出了一款语音听写的软件叫 IBM Viavoice，在演示当中效果不错，但是真正用时却很难使用。因此，在 2000 年左右第二次人工智能的浪潮又淹没了。

（3）现在到了人工智能真正爆发的前夜。随着 2006 年 Hinton 提出的深度学习技术，以及在图像、语音识别和其他领域内取得的一些成功，大家认为经过了两次起伏，人工智能开始进入了真正爆发的前夜。

就国内外人工智能公司这么多年的发展来看，使命是内在的，阶段性目标是变化的。

2 人工智能何时到来

在我看来，人工智能时代的到来离不开人机交互模式的变革。可以看到，自60年代至今，IT产业已经历硬件、软件、互联网、移动互联网与人工智能这五大浪潮，当前已进入物联网产业万物互联的时代。在无屏、移动、远场状态下，以语音为主，键盘、触摸等为辅的人机交互时代正在到来。目前主要面临两种交互：一种是只需要语音即可，比如蓝牙音箱、手环等，语音之外，不需要看到任何信息；另一种是语音+图像，比如电视上的语音交互、手机等。在这种情况下，触摸交互的学术名词应该叫做强视觉呈现的触摸交互；而语音作为人机交互最自然的方式，将有效促进人工智能与各行业的结合，让人工智能更容易进入大家的生活。除了语音交互，科大讯飞也在研究人脸识别技术，其特色是可以将人脸识别与声纹识别结合在一起，将声音与图片混合，来做活性检测。

由此，人工智能也将进入“智能+”的时代，人工智能与各个行业的深入结合蕴含着巨大的机会。除了交互，人工智能还可以用在教育、医疗、智慧城市、出行、司法、安全、金融等众多领域；同时，它在各个行业里可以做一个最简单的事情：就是替代人工。在未来的10年，人工智能会像技术的服务一样，进入到我们的生活当中，每个人都将离不开。

那么人工智能如何得以实现？在这里，我将人工智能的演进发展分成三个阶段：计算智能（能存会算）、感知智能（能听会说，能看会认）和认知智能（能理解会思考）。计算智能就是计算机与人类比存储、比记忆，在此方面已经远远超过人类了。不过，在感知层面，计算机在语音、图像识别等方面与人类还有较大差距，让计算机真正能理解、会思考、进行自我学习，还是很欠缺的。只有实现认知智能的突破，AI才能部分取代脑力劳动。

3 人工智能与创业

2016年，人工智能产业得到了长足的发展，收获了不少成功的案例。这里，我认为至少有三个因素促进了人工智能在产业界的成功：深度神经网络、大数据

以及涟漪效应：（1）深度神经网络。其模型和算法相对于传统的方法，有着本质的不同；虽然它与我们人类的神经网络相比，还有很多不足，但是确实在架构和描述方面有其强大之处。（2）大数据。随着移动互联网的迅猛发展，数据每天都是以指数级增加：通过手机、微信等工具和软件，人们可以随时随地把视觉、听觉上的这些数据轻松地传到网上，汇聚起来形成大数据。（3）涟漪效应。随着移动互联网的发展，各种软件、各种设备接触用户的门槛极大地降低了。例如，当一款新的 APP 找到第一批用户时，他们的使用行为和个人数据就被后台记录下来，开发者再对这种行为和记录进行迭代改进；当再把 APP 投向第二批用户时，软件的性能已经比第一代产品有了较大提升，这就是涟漪效应。

可以说，涟漪效应推动了语音辨识与图片识别的发展，特别是语音识别的实用化，更是得益于“涟漪效应”。科大讯飞在 2010 年推出语音识别产品时，识别率只有 60%左右，刚开始大家都觉得很难用，但是有一批尝鲜的用户。随着技术的迭代、更新，以及数据持续的迭代，如今讯飞语音识别率已经提高到 95%以上，达到了完全实用的状态。图像识别技术也同样如此，ImageNet 图像识别任务在 2012 年时错误率高达 26.2%，但是到 2015 年底已经降到了 3.57%。基本上可以说，图像识别技术的发展使得我们只要通过一个摄像头，就能将家中的各种物体很轻易地分辨出来。

因此，可以得出两点结论：深度神经网络与大数据的结合已成为当前主流路径；而基于互联网和移动互联网的“研究—工程—产品—用户”的闭环优化加速了产品迭代优化的进程。

当然，对于人工智能领域的创业者来说，产品创新、系统创新以及商业模式的创新也都是非常重要的。从技术层面看，产品创新与系统创新是相对立存在的，产品创新可以是一些微创新，而系统创新所需的资金和时间耗费都很大，从没有到开始立项，到最后商用需要 15~20 年，基本上创业者一辈子只能做出一个。从公司竞争角度看，现在的人工智能公司竞争不是单独两个公司，而是生态系统的竞争。比如创业公司很难独立把人工智能做好，于是各大公司都要做人工智能平台，包括科大讯飞的语音开放平台，现在已有 23 万开发者，每天服务 30~35 亿次，连接的数目达 90 多亿。

同时，在这个过程中，商业模式的创新非常重要，即好的技术创新一定要配

合好的商业模式创新。高科技企业的早期市场和主流市场之间存在着一条巨大的“鸿沟”，能否顺利跨越鸿沟并进入主流市场，成功赢得实用主义者的支持，就决定了一项高科技产品的成败。破坏性创新之父——克里斯坦森提出：“大公司卓越有效的管理对于延续性创新的成功具有决定性的作用而破坏式创新能够让创业公司和小公司拥有颠覆现有产业链的能力！”

最后，我想给创业者提点个人建议：去玩儿的事业一定是你真心喜欢的事情，如果你去玩儿还不选你喜欢的事情，我想你一定是神经病；去玩儿的事业，一定要跟你喜欢的人一起去做，玩耍的过程比结果更重要。谋事在人成事在天，能成为马云和马化腾是历史的必然，但成为这两个具体的人一定有很多未然的因素；改变你能改变的，接受你不能改变的。所以，我觉得人工智能创业不管是做系统创新，还是做产品创新、微创新，我们要以这样的心态，真正去享受到我们生活中的每一个小细节，同时要有使命感与宏伟蓝图！

（根据胡郁在 2017 年小饭桌人工智能创业班上的主题演讲整理）