·科技纵横·

语音交互为何迭代如此之快

最近,亚马逊公司旗下的云计算服务平台(AWS)发布了Amazon Polly和Amazon Lex。Amazon Polly的发音与人声已经非常相像,很多时候很难分辨机器与人声的界限。Amazon Lex是亚马逊的人工智能助手Alexa的内核,而Alexa已经被应用于亚马逊的Echo系列智能音箱。

2016年9月初,谷歌的DeepMind实验室公布了其在语音合成领域的最新成果WaveNet——一种原始音频波形深度生成模型,能够模仿人类的声音,生成的原始音频质量优于目前常用的语音合成方法,包括参数化合成(Parameric TTS,合成的语音听起来不自然)与拼接式合成(Concatenative TTS,对语料库的要求非常大,处理不好会产生语音毛刺和语调的诡异变化,且无法调整语音的抑扬顿挫)。

WaveNet引入了一种全新的思路,区别于上面两种方法。WaveNet利用真实的人类声音剪辑和相应的语言、语音特征来训练其卷积神经网络,让其能够辨别语音和语言的模式,其输出的音频明显更接近自然人声。WaveNet技术无疑是计算机语音合成领域的一大突破,但是其最大缺点是计算量太大,而且存在很多工程化问题。

语音交互是一个较长的生态链条,包括麦克风阵列、语义识别、语义理解和语音合成等,其中麦克风阵列技术主要解决远距离语音识别的问题,以保证真实场景下的语音识别率。这涉及了语音交互用户场景的变化,当用户从手机切换到类似智能音箱或者机器人的时候,实际上麦克风面临的环境就完全变了,这就如同两个人窃窃私语和大声嘶喊的区别。麦克风阵列技术的成熟,实际上补全了语音交互的最后一个技术链条,奠定了语音交互快速迭代的基础。

事实上,从2016年下半年语音交互市场的突然爆发起,几乎每隔一个多月,语音交互的效果都会出现较大的提升。为何语音交互技术的迭代会如此迅速?可以从下面几点来窥得一斑。

1 语音交互技术链条的成熟

深度学习带给了语音识别巨大的进步,但是以Siri(苹果公司的一项智能语音控制功能)为代表的手机语音交互一直不温不火,直到Echo和车载这类智能设备的出现,语音识别才突破手机的限制,真正落地到真实的垂直场景。这个转变不仅仅是场景的转变这么简单,实际上这从认知和技术上都是一个巨大的变化。当前的用户对于人工智能的要求其实并不高,他们希望确实能够解决一些具体问题,但是通用的语音交互总是伴随着智慧的概念,根本无法做到令用户满意。因此语音交互的落地首先就要考虑是否能够先服务好用户,这是一个关键的认知变化,而且基于这种认知,语音交互的免费策略似乎就不重要了,用户更为关注的是性能而非低价。另外一点是技术链条的成熟,语音识别从手机转向垂直场景,需要解决远场语音识别和场景语言理解的问题,亚马逊率先解决了这些问题,国内科大讯飞和声智科技也随后补齐了这个链条。目前来看,智能语音交互的技术链条趋于成熟,已经不存在较大的障碍。

2 真实场景数据规模的扩大

随着Echo的热卖,对于场景交互尤为重要的真实数据急剧增加,原先训练可能只有几千或者几万个小时,但是亚马逊已经从已售设备中获取了几千万的数据,而当前的训练已经是十万级数据的规模,将来百万级的数据训练也会出现。事实上,这些庞大的数据中囊括了用户时间长度和空间维度的信息,这是手机时代做不到的,从这些丰富信息之中,即便简单搜索提升的效果都是惊人的。

3 云端计算能力的不断提高

拥有了庞大的数据量,自然就急需计算能力的不断提升,2016年底,Intel召开发布会,中央处理器(CPU)和图形处理器(GPU)的综合计算能力再次提升了20多倍,这相当于原先需要训练20天的数据,现在可能不到1天就能完成,这是语音交互产业链条的根本性保证。

4 深度学习人才聚集的效应

技术、数据、计算链条的相对完善,核心还需要人才的驱动,而随着人工智能的热潮,不断有更多相关人才从科院机构和院校走出来加入这个行业。创业公司的竞争是可怕的,这些才华横溢的人才日夜拼搏,其效率提升到其他任何时代可能都难以匹及的程度。

总之,智能语音交互链条已经具备了大规模普及的基础,等待的只是用户习惯的改变,而这种改变正在逐步发生。可预见的几年,语音交互相对于其他人工智能技术,应该是最先落地的一种技术,而且其迭代的速度可能会超过我们的预期。但是语音交互仍然还有很多问题需要解决,包括终端技术的低功耗和集成化、语音识别的场景化和一体化以及语言理解的准确性和引导性。

未来几年,智能语音交互的迭代至少还要解决如下几个问题:一是如何基于用户提出的多种多样的、基于情感的、语意模糊的需求进行深刻分析,精确理解用户的实际需求;二是如何将各种结构化、非结构化、半结构化的知识进行组织与梳理,最终以结构化、清晰化的知识形式完整地呈现给用户;三是如何猜测用户可能会有什么未想到、未提出的需求,从而先人一步为用户提供相关的扩展信息;四是如何将信息进行有效地组织与整理,以条理化、简洁化、直接化的形式呈现给用户。

当前来看,语音交互需要一个屏幕,这虽然会使语音交互产品的品类属性减弱,但是在AR(增强现实技术)还没有发展起来之前,确实也没有更好的办法。毕竟单纯的语音交互缺少一个使得人机交互更完整的重要的组件——视觉交互,没有用户界面或上下文元素的基于语音交互的系统是不完整的。因为用户倘若想播放音乐、定时、控制灯光、获得新闻头条等时,可以通过聊天的方式来控制,然而若要在线比较一下两种产品的价格、各种性能参数,或者想看一下未来一周天气预报的温度趋势,就需要一块显示屏了。

文/陈孝良
作者简介 中国科学院声学研究所副研究员,声智科技创始人。

(责任编辑 王丽娜)