数字强国①

大模型引领AI新十年，如何潮涨不落

林咏华：打造基座大模型等同于AI中的CPU，以开源、评测推动产业落地

6月20日，163期文汇讲堂“数字强国”系列启动，首期《AIGC驱动生产力跃升与良好未来塑造》在华东师大举办。北京智源人工智能研究院副院长兼总工程师林咏华应邀作主讲，华东师大哲学系杨国荣致辞《科技发展与人类生活》，计算机学院贺樑、哲学系鄔全民、付长珍、潘斌、刘梁剑参与数字&人文圓桌对谈，13位听友现场互动。

校内外7600余人观看华东师大微信号直播。现场发放了50枚NFT数字徽章。

本次讲座由文汇报社、上海树图区块链研究院、华东师大中国现代思想与文化研究所、华东师大哲学系伦理与智慧研究中心联合主办。

本版整理 李念 摄影 周文强 版式 李洁

①林咏华结合智源研发过程的讲解，让听友知其然并知其所以然。②杨国荣致辞提出，从人是目的和具有原创性来规范技术发展。③现场50位听友有幸获取上海树图区块链研究院铸造的NFT数字徽章。

嘉宾主讲

非常高兴借此机会和大家分享我过去多年在AI（人工智能）领域经历的机遇和挑战：从小模型走到大模型，从科研成果落地到产业。在过去几十年，人工智能起起落落。去年6月之前，整个人工智能处在前一波浪潮往下落的一个区间。去年下半年，出现了两个现象级的应用：一是文生图，二是以ChatGPT为代表的大模型技术的涌现和爆发。这两个事件把整个AI从一个拐点引向下一个起点，而这个新起点是由大模型引领未来人工智能发展的十年。

思考一：带来研发范式改变

为什么研发范式很重要？因为当科研界将一个技术做到突破和创新后，它们如何广泛地落地到各行各业，与其研发范式、研发产品的代价息息相关。

至今，AI研发范式经历了三个阶段的变化。

过去十年：预训练模型+微调训练的迁移学习

第一个阶段是从头开始训练领域模型。最初深度学习出现时，大家考虑的都是如何利用手上海量的数据，通过诸多计算资源，把模型从头到尾训练出来，然后再将它部署到各行各业。因为需要大量数据、算力，尤其是所需的整个AI全栈的技术人才特别昂贵，因此，这种范式无法持久。

2014年，在几个AI顶级峰会上分别出现了描述预训练模型+微调的迁移学习技术的文章。利用拥有1000多万张图片、涵盖常见的2万种物品的图片库，训练出通用的视觉分类基础模型，其规模是中小量级的模型。此后，大家利用医疗影像分析、工业的缺陷检测等自己领域的数据对它进行训练。这一过程是从一个通用领域到另一个专用领域的迁移学习。从今天视角来看，相当于一个初中毕业生通过三年的专科培训，成为了一个具有专业技能的专员。

由此，研发范式进入第二个阶段——由预训练的基础模型加上小批量的数据和少量的算力的微调训练，就可以形成企业要落地到不同场景的不同模型。这种范式中，行业企业只需要做少量数据收集和处理、模型微调训练、部署模型服务等部分工作，从人力、物力、财力上来看，投入量减少了几倍、甚至十倍。

计算机视觉领域的迁移学习，带动了过去十年的AI潮起潮落。这整个过程可称为小模型的阶段。

从2013到2015年，人工智能因为迁移学习的出现，让基于深度学习的计算机视觉分析，在多个领域落地变得似乎更加容易，人工智能被认为有望大范围成功。商汤、云从、依图、格灵深瞳等在内的众多AI公司纷纷创立，受到投资界的普遍追捧。

但从2017年之后，人工智能从高潮慢慢缓落。到2020年，拿到融资而成立AI公司的数字从2007年顶峰时的4000家落至600-700家，以至于在过去一两年甚至出现了AI泡沫破灭的众多说法。

为什么跟大家分享这些？眼看AI又一个新的十年潮起涌现，作为从业者需要深入思考：为何前一个十年出现万众期待，最后并未如想象在各行各业广泛落地？而在未来十年，该做对什么，使得新一轮技术潮起后能得到更好的发展，而非快速潮落。

当下阶段：基础大模型+应用提示

在当下的第三阶段研发范式中，基础大模型很重要的是基座，一是需要用海量的预训练数据去训练它，通常是千亿级以上的数据。二是参数量很大，几十亿参数是入门，很多时候会达到百亿级参数，甚至千亿级参数。三是所需要的算力更大。这种基础大模型帮助我们学习各种通用的知识，包括实现各种模型的能力，如理解能力、生成能力，甚至涌现能力。目前看到的GPT-4、GPT-3.5、LLaMA、北京智源人工智能研究院（以下简称“智源”）新研发出来的天鹰Aquila等，都是基础大模型。它最突出的能力是提示学习能力，跟人很像，可以做到有样学样。

在这个阶段，只要做API（应用程序编程接口）调用，就可适用到下游行业企业的应用领域，成本大幅度降低。

思考二：大模型如何产业落地？

大模型如何产业落地？这一步走好才能让上亿甚至数十亿、数百亿元在大模型上的研发投入，能够真正带领所有行业的智能化提升。

从通用的问答聊天功能通向专业行业的应用

大模型的应用方式有两种：一种是提示学习，另一种是指令微调训练。

大模型是“记不住”提示学习的过程的，如果仅靠提示学习中的“提示”，势必每一次的API调用都得带上冗长、而且越来越长的提示，这在实际产品中很难满足。因此在产品真正落地时，必须要引入指令微调。指令微调就是利用基础模型的知识完成指定的任务。就像本科生学了大量知识后，需要一个上岗培训。

今天看到的ChatGPT不是一个基础模型，它是一个经过很多指令对它进行微调的对话模型，所以它似乎做什么都很在行。智源的天鹰AquilaChat对话模型，也是在Aquila基础模型之上经过指令微调才可回答人类的各种问题。比如6月8日正好是全国高考，测试时它在10秒内就完成了当天的高考作文。

通过指令微调的大模型还只具备通用的能力，即主要是面对互联网的应用，如闲聊、问答。如果希望大模型能够真正服务于更多的实体经济，就需要考虑如何把大模型落地到专业行业里。很重要的一点是要在通用能力的基础模型之上，通过加入大量专业领域知识进行持续训练，形成专业领域的基础模型。

所以，综合来看，基础模型训练相当于通用领域的本科生学习，基础模型在专业知识数据的持续训练相当于专业领域的研究生深造学习，之后再进行指令微调训练，相当于专业领域的上岗培训。

如何克服落地时遇到的遗忘率和幻觉率

在现实落地中会遇到大模型的遗忘率和幻觉率。

遗忘率就是“记不住”。无论模型大小，如果只让模型看过2-3遍的数据，它能记住的只有百分之几的数据量。这就产生了一对矛盾。首先从版权保护的角度看，如果它因为读了大量文章，而产生大篇幅与原文相同的内容，是否会导致版权问题？这是有待解决的问题。其次，如果模型的记忆力只有百分之几，版权问题就不会那么严重。但当真正产业落地时，这又会成为较大的问题，即模型训练了半天却记不住。

“幻觉率”就是我们常说的一本正经的胡说八道。成因是什么？第一，预训练的数据集可能会包含某一些错误的信息，很多来自二十年前、三十年前，会昨是今非。第二，更多可能是模型的数据预训练的上亿、几亿的数据里没有直接包含相关信息。这会导致我们面对严肃的行业时，如医疗、金融、法律等，必须考虑用什么额外的技术来降低幻觉率。

未来十年，大模型和小模型如何并存

我个人认为，未来十年大模型和小模型必定会共存，这两个技术更多时候可以相互融合。

对过去十年发展起来的小模型的AI公司、科研团队，在大模型时代是否都需要迁往大模型？应该如何利用已有积累做得更好？

第一，可以把原有在小模型时代的算法进行更新换代，把大模型新的技术融入到小模型。举例，Transformer模型结构被认为是大模型时代重要的技术标志。用Transformer为基础的VIT计算机视觉模型，来替代小模型时代的CNN网络，发现在达到差不多准确率的情况下，大模型在预训练阶段可节省1/4的显存，推理速度只需要ResNet50的58%时延，上线所需要的资源更少。这的确打破了大模型技术必须是资源消耗高的定律。

第二，应用新的方法解决以前的难题。比如Meta公司在今年3月发布的视觉分割大模型SAM，能做到视觉范围内各种物体被精准地分割出来。这种技术可以用于清点超市、仓库等的货物数量。此前一直很难做到，有一些小模型公司已将SAM大模型落地。

第三，大模型中的小模型。例如我们新发布的AquilaChat天鹰对话模型，仅70亿参数，通过int4量化技术，就可在4G的显存上运行起来。而当前国产边缘侧的芯片都已经有8G显存。所以，大模型浪潮下，很多AI小模型赛道的公司，完全可以焕发一种更新的活力。

思考三：打造基础大模型重如CPU

大模型中最重要的是下面的基座模型。打造基座大模型就等同于AI中的CPU一样的重要。

投入昂贵，百亿参数动辄千万元以上

第一，除了做芯片、CPU的流片以外，基础模型已经成为AI大模型时代单一产品投入最大的部分。一般而言，300亿参数的模型，包括数据、训练、评测的成本、所有的人力、物力、算力加起来，要耗资2000万元；而上千亿参数的模型，则约在4000多万元甚至更高。所以动辄几千万元训出一个模型，投入十分高昂。

第二，基础大模型决定了下游各种模型的重要能力。从能力来看，大模型的理解能力、涌现能力、上下文学习能力都是由这个基础模型的结构、尺寸等等决定。从知识来看，无论是通用知识还是专业知识都是在基础模型训练过程中学习到的。

价值观的保证首先需要干净的语料库

第三，从合规性和安全性来看，对于内容生成的模型，其生成的内容是否积极阳光，无偏见和伦理问题等，很大程度是由基础模型决定。

基础模型如何获得人类的价值观？通过训练语料。国内外一些科研机构、公司训练基础模型，通常应用到Common crawl语料库，这是互联网训练语料全球最大的集合。但其中只有很少的是中文数据，而所有中文数据中，又只有17%的网源、网站、网址是来自于国内。

基于已出现的这种风险，我们训练的中文语料均来自智源从2019年积累至今的数据，99%以上是来自咱们国内的站源。国内站源具有的优点是都有ICP备案许可，所以也规范了网络内容的可靠性和可信度。

开发中英文、可商用基础模型造福下游企业

第四，从版权和商用许可来看，非商用许可的大模型，企业后续无法商用。据统计，今年1月至5月，新发布的国外开源语言大模型共有39个，其中可直接商用的是16个；而国内开源语言大模型只有11个，且仅有1个对话模型是有可商用许可。基础模型对整个产业的发展价值更大。截至5月底，国外发布的开源语言大模型里只有5个是基础模型，而国内发布的开源语言大模型里只有2个是基础模型，是复旦的MOSS和清华的CPM—Bee。

因此，作为非盈利科研机构的智源，我们倡导更多的力量做这样的投入：第一，支持中英双语的基座模型。中文里有很多知识是需要直接被训练到模型，依靠翻译无法将许多中文知识纳入其中。第二，希望可以支持商用许可协议，在预训练阶段就可以开源，这才能够避免众多企业重复资源建造基座模型。第三，符合国内数据合规需要，尤其是纳入优秀、高质量的中文内容。

另外，代码模型是大模型产业落地的一种很重要的模型，具有广阔应用前景。基于Aquila-7B强大的基础模型能力，我们用更少的代码训练数据，小参数量，高效实现了目前性能最好的中英双语代码模型。我们分别在英伟达和国产芯片上完成了代码模型的训练，通过支持不同芯片架构的代码+模型的开源，推动芯片创新和百花齐放。

思考四：评测变得无比重要

大模型训练要紧抓两头：一头是数据，一头是评测。

为什么评测很重要？一个300亿参数的模型，每天对它投入的算力是10万元，十分昂贵。而正因为它大，在整个过程中更需要关注所有的细节，一旦出现问题，要及时做出调整。

需要各种评测方法评测大模型复杂的能力

此外，大模型的能力很复杂，很难用单一指标表明这个模型在未来使用的各种能力，所以要使用各种评测方法和评测集对它进行评测。在大模型训练稳定之后，就要开始指令微调训练，再进行循环迭代，不断调整。如果在过程中只用计算机客观评测，很难准确及系统性地看到主观的生成能力，因此还必须加入主观评测。而主观评测至今还只能由人类进行，我们也尝试用ChatGPT配合人类进行评测。最后优选的模型还要进入red-team评测，也就是找一组未参与模型研发的人员扮演用户群，对这个模型进行各种提问，包括各种恶意、刁钻的提问，来评估这个模型的效果。

智源打造了FlagEval天秤大模型评测系统，包括了中、英双语的客观、主观22个评测集合，8万多个评测项。基于目前最新的评测，AquilaChat以大约相当于其他模型50%的训练数据量达到了最优性能。

评测已经演进到认知能力和人类思维能力

大模型从去年进入所有人的视野，其能力发展迅速。同时评测的难度也一路攀高，相当于不断地拉长尺子，才能更好地量度大模型的能力。

随着大模型能力的提升，对评测产生了四个台阶的演进：

第一，理解能力。过去十年、二十年，AI一直是以理解能力评测为主，无论是计算机视觉还是自然语言处理。

第二，生成能力。现在已经出现了AI生成内容，这不得不依靠人类的主观进行评测。主观评测的质量很难完全保障一致和客观性，虽然也逐步引入一些AI的辅助手段，但在很多的测试案例上仍然偏差很大。

第三，认知能力。当前人们考量各种大模型，已经不认为它们只是一个能说会写的语言模型，而是希望看到各种各样的知识能力、认知能力。因此，对评测来说，更大的挑战是如何刻画一个全人类的认知能力。另外，很多人用各类考题考验这些模型，但这些考题很多都已经被泄露到模型的训练语料，所以这种认知能力的评测也有失偏颇。

第四，人类思维能力。更困难的是，很多人希望这个模型更加像一个人类的思维一样去理解、去思辨。所以对于模型的心智能力应该怎样评判、评价，就需要多学科的交叉。

思考五：保持工匠精神与好奇心

在大模型时代，拥有近200个全职研究人员的智源看到各种各样的现实问题、技术的问题，亟需AI领域内外学科的科研人员团结协作，共同去突破。无论文生图还是ChatGPT的应用，都离不开冰山下整个大模型全技术栈的积累，而这正是智源一直致力于打造的部分——所有的基础模型，包括数据集、数据工具、评测工具，甚至包括AI系统、多种的跨芯片技术的支撑。这是我们的使命，既要打造冰山以下的大模型技术栈，同时以可商用的形式全部开源出来。

大模型时代一方面需要以工匠的精神锻造每一个大模型，每一步都要精雕细琢；另一方面，大模型里有太多的未知，需要以追星逐月的好奇心去探究。只有我们探究得更好，才能让它在产业落地得更稳，未来的十年才能是潮起后不断地稳步向前发展。