

来自启智会的报告：

# 超大规模预训练模型的优势、局限与未来趋势

撰稿：黄民烈

组织：唐杰 文继荣

近几年来，预训练技术在自然语言处理领域蓬勃发展，在多个子方向取得了颠覆性的成果，得到了广泛应用。然而，预训练技术仍然有许多未被发掘的有效方法，以及很多亟待解决的重要问题。例如，预训练模型的理论基础是什么？如何理解预训练模型中存储的“知识”？下一代的预训练模型是什么样的，是多模态融合的模型，结合外部知识的模型，抑或是面向认知的模型？如何设计深度学习框架，实现更高效的大规模预训练？如何降低预训练模型的使用门槛，推动“预训练社区”的发展？

这些问题都是现在自然语言处理领域中预训练技术的核心问题，对技术本身甚至整个自然语言处理领域都将带来深远影响。深入探讨这些问题不仅有助于我们更好地理解预训练语言模型，将现有的技术潜力发挥到极致，也有助于我们把握未来的发展方向，做出超前的、具有突破意义的工作。

2021年1月1~3日，CCF第8期启智会在北京召开，来自国内5家单位的11位专家出席会议，围绕上述问题展开了深入讨论。会议采取主题报告、辩论、自由讨论相结合的形式，从不同角度对超大规模预训练语言模型的优势、局限以及未来趋势进行了探讨。

预训练技术在机器学习中由来已久，相关理论也早有建立，深度学习最早的突破便采用了逐层预训练。预训练可以看成一种正则化，使微调过程更

加稳定，近期也有理论成果分析了无监督表示学习对下游任务的泛化性能的影响。然而，现有基于Transformer的预训练模型的理论仍没有被研究透彻，这种架构也没有被证明是最好的模型架构，因此我们需要从理论上探讨Transformer模型中几个模块的作用，以帮助人们找到更好的预训练模型基础结构。

尽管预训练模型具有很强的能力，但我们仍不清楚其中蕴含的内容，甚至没有一个明确的定义。现有的国内外文献中大多使用“知识”（knowledge）指代预训练模型蕴含的能力，但是这种指代通常会和人们普遍所说的知识图谱、百科网页等知识库中的“知识”产生混淆，而后者并不能涵盖预训练模型的全部能力。因此，专家们提议定义一个新的名词“模识”（modeledge）以描述预训练语言模型中的“知识”与能力，以便和传统意义上的“知识”进行区分。在拥有“模识”定义的基础上，我们将更加关注如何帮助模型更好地学习“模识”，帮助人们理解与利用“模识”，以构建下一代预训练模型。下面总结了几个相关的重要研究方向。

第一个方向是将多模态信息引入预训练过程。其出发点在于，人类的认知就是基于多模态的。在常识推理的过程中，很多常识并不直接表现于文本内容中，现有只基于自然语言处理的工作很难取得一个好的常识理解。而图像中的视觉信息提供了不同角度，利用多模态预训练模型可以将常识信息有



效发掘出来,促进常识推理等相关任务的发展。然而,现有的多模态预训练模型面临着数据和训练目标两方面问题:在数据方面,通用领域多媒体数据比文本数据体量更大、更难收集全,质量更难控制;在训练目标方面,多模态预训练需要更泛化、更明确的预训练任务。将来,预训练模型的应用应该更多地进行多模态融合,例如新一代中文多模态对话系统 Meena++, 等等。

第二个方向是在模型训练过程中引入外部的人类知识,辅助“模识”的学习与利用。人类知识包括实体知识图谱、常识知识、概念知识等,形成了多类型、有层次结构的知识体系。这类方法最终的目标是形成知识与数据的“双塔”架构,建立深度学习与知识图谱双轮驱动的自然语言深度理解框架。然而,目前融合知识图谱等结构化知识的预训练模型还存在一些问题,一方面,结构化的知识需要向量化,会造成信息损失;另一方面,大部分结构化知识也是从非结构化文档中抽取出来的,抽取的过程会造成较多信息损失,需要考虑直接融合“文档知识”来避免上述两个问题。此外,我们还需要设计能从数据中诱导出常识知识的预训练学习任务,实现更好的效果。

将预训练模型的研究与认知科学相结合,也是一个十分重要的研究方向。从认知科学的角度,“模识”不仅包含模型对于语言的使用能力,还包含将语言与现实世界以及人们已有的知识进行对应(grounding)的能力。但 GPT-3 这类超大规模预训

练模型只具有强大的语言使用能力,仍然缺乏与客观世界对应的能力。相比于理想化的通用人工智能,当前的机器学习算法缺乏更高级的认知能力,特别是持续学习和完成复杂任务的能力。我们可以参考认知科学中的工作记忆理论、全局工作理论(GWT),构建面向认知的AI架构,一方面探索训练更大规模的模型,另一方面创新算法架构,加强模型在多跳推理、符号推理、复杂决策方面的能力。

在研究预训练模型本身的能力之外,如何构建更高效、更平民化的预训练模型,降低预训练模型的使用门槛与使用隐患,也是一个重要课题。一方面,我们可以着眼于优化深度学习框架。虽然现在的学习框架通过数据并行、混合精度训练等方式,对小规模模型的性能优化已经基本收敛,但由于大规模预训练模型无法存储在单张显卡上,因此需要使用模型并行、流水并行等技术,这便对显存利用率优化、网络传输带宽提出了新的挑战。未来,深度学习框架的设计会尝试使用编译器自动优化底层代码,静态执行与动态执行协同融合的技术路线,并最终实现深度学习框架之间的平滑迁移以及AI流程自动化。

另一方面,我们可以建立“预训练模型开放社区”。在这个过程中,需要实现以下几方面:(1)构建更高效的模型,更合理地分配计算资源,让普通科研人员得以使用预训练模型;(2)定义一套预训练模型的通用接口或者使用规范,让人们更方便发布和使用不同来源的预训练模型;(3)合理处理预训练模型可能涉及的版权和隐私问题,从而更为安全地使用预训练模型。■

出席本次启智会的专家(按姓氏拼音排序):

黄民烈 兰艳艳 刘洋 刘知远 邱锡鹏  
宋睿华 唐杰 文继荣 袁进辉 赵鑫  
朱军

(上述观点均来自各位专家,由撰稿者汇总整理)