中国科学院团队发布类脑脉冲大模型“瞬悉”

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/35511.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

中国科学院团队发布类脑脉冲大模型“瞬悉”。中新网北京9月8日电 (记者 孙自法)中国科学院自动化研究所9月8日宣布，该所研究团队最新成功研发并发布类脑脉冲大模型“瞬悉”1.0(SpikingBrain-1.0)，实现两个首次和多项核心性能突破，为新一代人工智能发展提供了新技术路线，并将启迪更低功耗的下一代神经形态计算理论和芯片设计。

两个首次

这是中国首次提出大规模类脑线性基础模型架构、并首次在国产图形处理器(GPU)算力集群上构建类脑脉冲大模型的训练和推理框架。



最新研发发布的类脑脉冲大模型“ 瞬悉 ”1.0界面。中国科学院自动化研究所供图

“ 瞬悉 ”大模型由中国科学院自动化所李国齐和徐波研究员领导团队基于“ 内生复杂性 ”理论构建，已在国产GPU平台上完成全流程训练与推理，实现大模型在超长序列推理上数量级的效率和速度提升，展示出构建国产自主可控的新型大模型架构生态的可行性。

“ 瞬悉 ”大模型解决了脉冲驱动限制下的大规模类脑模型性能退化问题，其超长序列处理能力在法律/医学文档分析、复杂多智能体模拟、高能粒子物理实验、DNA序列分析、分子动力学轨迹等超长序列任务建模场景中，具有显著的潜在效率优势。

目前，研究团队已开源“ 瞬悉 ”大模型并开放测试网址，还同步公开经工业界大规模验证的该类脑脉冲大模型中英文技术报告。

四项突破

研究团队指出，“ 瞬悉 ”大模型主要在高效训练、推理效率、类脑大模型生态构建、多尺度稀疏机制等四项核心性能上实现突破：

极低数据量上的高效训练方面，训练阶段具有线性或近线性复杂度，显著提升长序列训练效率，并能依托高效转换训练范式，以约为主流大模型2%的预训练数据量，实现与众多开源Transformer(一种基于自注意力机制的深度学习架构)模型在多任务语言理解、中文多任务语言理解、常识推理能力任务上相媲美的性能。

推理效率的数量级提升方面，推理阶段结合脉冲神经元事件驱动特性，“ 瞬悉 ”大模型具有常数或部分层常数级别的复杂度和存储开销，在超长序列处理能力上展现出数量级的效率和速度提升。

国产自主可控类脑大模型生态的构建方面，“ 瞬悉 ”大模型适配了面向国产GPU集群的高效训练和推理框架、有关算子库、模型并行策略以及集群通信原语，表明构建国产自主可控的新型非Transformer大模型架构生态的可行性。

基于动态阈值脉冲化的多尺度稀疏机制方面，“ 瞬悉 ”大模型设计细粒度的两阶段动态阈值脉冲化策略，结合粗粒度的混合专家模型方案，实现超过69.15%的稀疏度，长序脉冲占比约1.85%，为低功耗的类脑大模型运行提供有力支撑。

新型路径

研究团队介绍，当前基于Transformer架构的主流大模型，通过增加网络规模、算力资源和数据量提升模型智能水平，其基本计算单元为简单的点神经元模型，此路径被称为“ 基于外生复杂性 ”的通用智能实现方法。

不过，Transformer架构的固有缺点，是训练时开销随序列长度呈平方级增长以及推理时显存占用也随序列长度线性增加，构成资源消耗的主要瓶颈，导致其处理超长序列的能力受限。

借鉴大脑神经元内部复杂工作机制，研究团队提出“ 基于内生复杂性 ”的大模型构架方式，成功研发出类脑脉冲大模型“ 瞬悉 ”，在理论上建立脉冲神经元内生动力学与线性注意力模型之间的

联系，揭示现有线性注意力机制是树突计算的特殊简化形式，从而清晰展示出一条不断提升模型复杂度和性能的新型可行路径。

根据新型路经，研究团队构建并开源基于脉冲神经元、具有线性及混合线性复杂度的新型类脑基础模型“瞬悉”，它仅需约主流模型2%的数据量，就能在多项语言理解和推理任务中媲美众多主流模型。(完)

（原标题：中国团队发布类脑脉冲大模型“瞬悉”实现两个首次和多项核心突破）

作者：孙自法 来源：中新网

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发