



北京大学
PEKING UNIVERSITY

基于Looped Transformer架构的优化理论与 关键技术研究

项目负责人：贺笛

项目参与人：陆一平

参与单位：北京大学智能学院

北京大学国际数学研究中心



北京大学
PEKING UNIVERSITY

目录

一

国内外现状及发展趋势

二

科学问题与研究目标

三

研究内容与计划进度

四

研究团队与工作基础

五

预期成果与考核指标



北京大学
PEKING UNIVERSITY

目录

一

国内外现状及发展趋势

二

科学问题与研究目标

三

研究内容与计划进度

四

研究团队与工作基础

五

预期成果与考核指标

大语言模型成为智能应用的重要基础，能力扩展伴随资源需求增长

主流生成范式：标准自回归



代表性模型的规模扩展

模型 / 年份	参数量
GPT-2 2019	1.5B
Llama-2 2023	70B
Kimi k3 2026	2.8T

关键转向：在有限资源下，同时提升计算利用率（计算瓶颈）与参数利用率（存储瓶颈）



计算瓶颈：是否每个词元生成都需要相同计算深度？

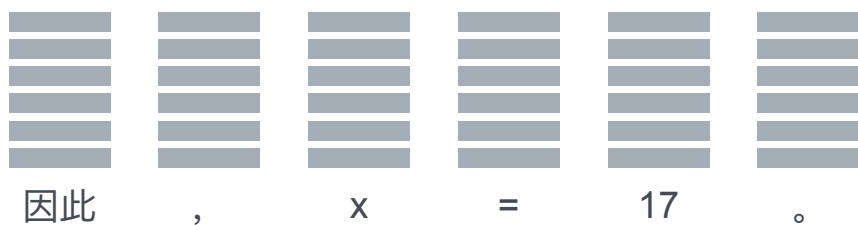
生成位置的预测难度并不一致，固定深度难以匹配差异化计算需求

从“一律算满”到“按需分配”

同一段输出，在不同生成位置采用不同的内部计算深度

固定深度

所有位置
完整计算



自适应深度

不同位置
按需计算



色块数量表示内部深度；划分与难度仅作机制示意

代表性研究与启示

• 谷歌 CALM · 2022

依据置信度提前退出，减少部分生成位置的层计算。

• 谷歌 MoR · 2025

按词元路由递归深度，联合参数共享与自适应计算。

研究启示：把计算用于关键位置



存储瓶颈：独立参数是否都有存在意义？

跨层参数共享，使参数量与深度解耦

标准Transformer：独立层堆叠



存储：4份模块权重

计算：4次模块调用

改进：共享模块循环的编码器



存储：1份模块权重

计算：R次模块调用

代表性进展

2018 · Universal Transformer

循环深度模型及其计算理论

2019 · StackingBERT(申请人工作)

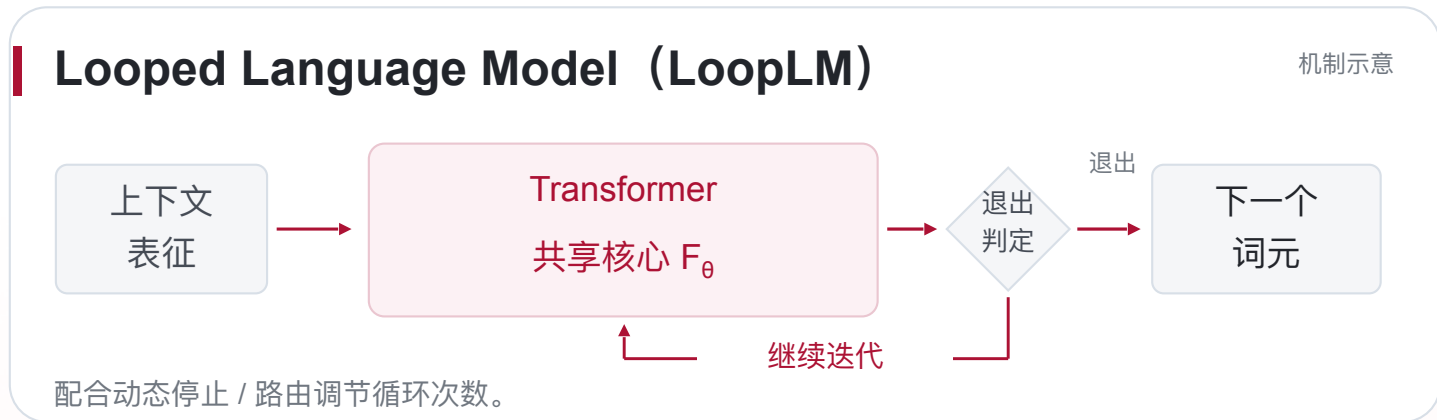
循环深度编码器训练策略

2019 · ALBERT

循环深度编码器预训练模型



前沿热点：深度循环语言模型解决计算瓶颈与存储瓶颈



媒体动态 | 2026.09

OpenAI Astra

被报道采用“深度循环”

The Information 报道 · The Verge 跟进

媒体消息；官方材料未确认具体架构。

模型规模化

2025.10—2026.10

字节Ouro, SMELT模型

1B - 54B深度循环语言模型。

高效部署

2026.09.15 / 10.04

LoopSpec / Loopy

自推测解码加速；面向
深度循环设计低比特量化方法。

本项目切入：从循环架构设计，走向稳定训练与高效优化



北京大学
PEKING UNIVERSITY

目录

一

国内外现状及发展趋势

二

科学问题与研究目标

三

研究内容与计划进度

四

研究团队与工作基础

五

预期成果与考核指标

前向传播：循环深度增加，并不自动保证前向传播的表示稳定

同一变换反复作用，局部效应持续累积



函数压缩过强

表示趋同，信息逐轮损失

函数波动放大

扰动累积，残差与损失不稳定

核心矛盾：

既要数值可控

又要保留有效特征更新

过强约束可能抑制信息加工；

多步放大可能破坏数值稳定。

需明确稳定性边界

关联循环深度、参数尺度与更新幅度

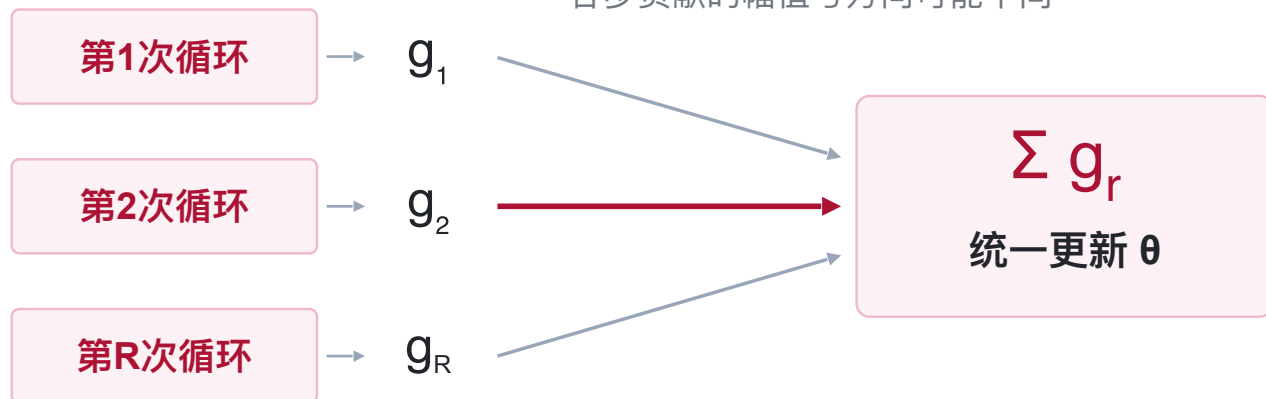
研究观察

Parcae (2026) 报告有循环导致的残差爆炸与损失尖峰出现，需要优化稳定方案。[1]

反向传播：一组参数承担多轮计算，参数更新须兼顾不同循环位置

不同循环的梯度，汇聚为同一次更新

各步贡献的幅值与方向可能不同



同一参数改动，在不同隐状态下产生不同影响

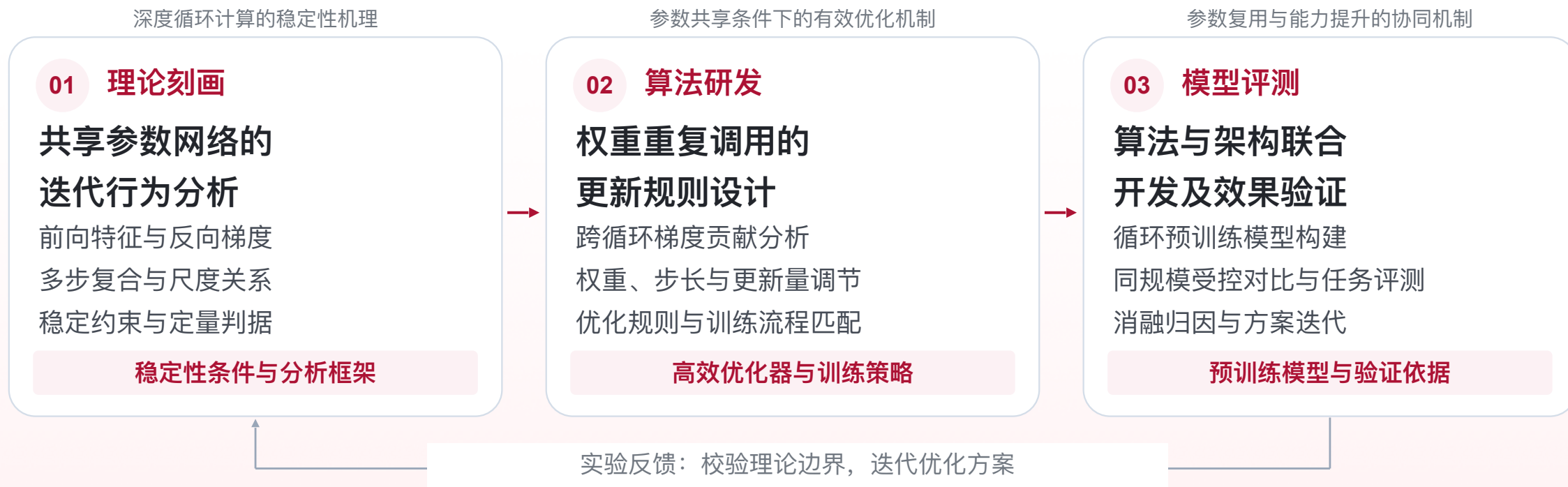
三项相互耦合的优化难点

- 1 贡献不均**
各步贡献可能失衡或抵消，整体更新难兼顾所有位置。
- 2 尺度耦合**
循环深度改变，步长与更新强度难以直接沿用。
- 3 开销仍在**
共享权重不等于反向传播时间与激活开销同比下降。

研究观察

Loop Dropout (2026.09) 在LoRA适配中观察到“循环偏置”；不同循环次数的梯度贡献不同。[2]

面向参数共享带来的训练稳定性与优化效率瓶颈，建立循环计算优化理论，
研制高效优化器与训练策略，构建兼具稳定性、可训练性与表达能力强的语言模型。



目标落点：同等参数规模下，推动循环语言模型性能超越标准非共享 Transformer



北京大学
PEKING UNIVERSITY

目录

一

国内外现状及发展趋势

二

科学问题与研究目标

三

研究内容与预算分配

四

研究团队与工作基础

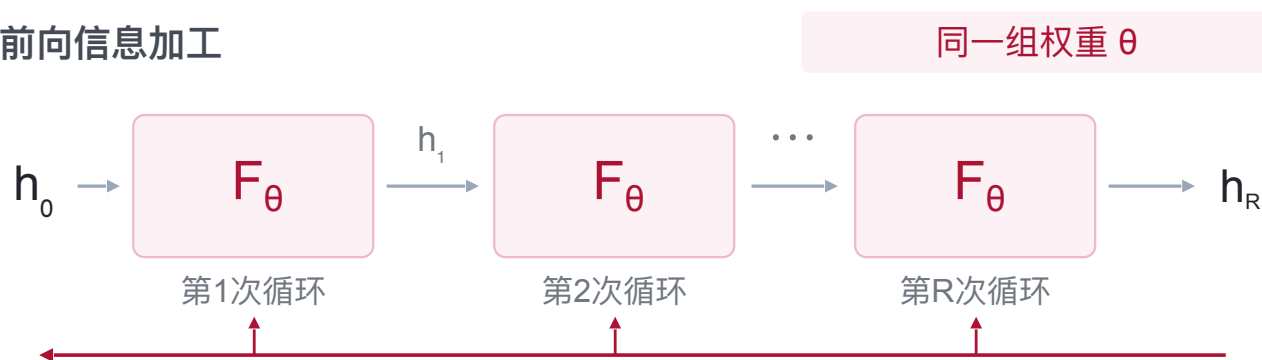
五

预期成果与考核指标

从前向表示与反向梯度出发，提炼深度循环计算的稳定性条件

共享模块多步复合：前向 / 反向联合刻画

前向信息加工



反向误差传递：雅可比连乘与梯度贡献累积

局部线性化

多步复合分析

稳定约束推导

三个分析重点

1 前向特征演化

隐状态范数与特征相关性
识别特征趋同和信息损失

2 反向信号传播

分析梯度衰减、放大与抵消
区分误差校正与贡献失配

3 稳定与更新兼顾

关联权重尺度、更新量与步长
避免过强收缩抑制特征更新

理论分析框架

循环次数

梯度传播

参数 / 更新范数

有效学习率

预期产出：稳定性条件与定量判据，指导优化器构造并明确适用范围



研究内容二：面向权重重复调用的更新规则设计

以循环传播特征指导更新调节，实现优化器与训练流程的协同设计



预期产出：适配参数共享机制的高效优化器与配套训练策略

在受控对比与分层评测中，检验参数复用转化为能力提升的实际收益

模型构建与公平对照

约5亿参数 · 相同训练语料量 · 统一评测设置

本项目
方案

新优化器
训练策略

共享模块
循环配置

Looped
预训练模型

标准
基线

标准非共享 Transformer

独立层堆叠；参数规模与训练语料量一致

分层评测与贡献归因

01 预训练效果

比较验证集损失

02 零样本任务

HellaSwag / PIQA / ARC-Easy

03 消融与重复试验

替换优化器、改变循环次数、
移除训练策略；

预期：验证集损失
降低不低于

0.02

预期：三项平均准确率
提升不低于

3%

预期产出
模型、设计准则与可复用组件



科目名称		申请经费（万）
设备费	租赁设备费	36
业务费	差旅/会议/国际合作交流费	3
	出版/文献/信息传播/知识产权事务费	1
劳务费	人员劳务费	30
	专家咨询费	2
间接经费	间接经费	28
总计		100



北京大学
PEKING UNIVERSITY

目录

一

国内外现状及发展趋势

二

科学问题与研究目标

三

研究内容与计划进度

四

研究团队与工作基础

五

预期成果与考核指标



北京大学智能学院

北京大学智能学院提供了丰富的研究配套资源，包括优秀的生源和强大的算力平台。研究院是北京大学通用人工智能班（“通班”）的主要教学与科研执行单位。针对本项目，预计多名博士学生都会分配具体的研究子课题中，此外通过“通班”科研轮转项目另有学术能力突出的本科生参与到课题中来。

北京大学国际数学研究中心

北京大学国际数学研究中心包括优秀的本硕博生源、雄厚的数学研究力量以及北京大学跨学科科研平台的支持。中心汇聚了众多数学及交叉学科领域的优秀学者，为本项目的理论与算法研究提供有力支撑。此外，依托北京大学数学学科拔尖人才培养计划，项目还将吸纳一批数学基础扎实、科研能力突出的本科生，通过科研训练和专题讨论班参与课题研究，为项目实施提供持续的人才支持。



贺笛
项目负责人

北京大学 研究员

- ▶ 2020年于北京大学信息科学技术学院获得计算机科学与技术专业博士学位
- ▶ ICLR'23杰出论文奖
- ▶ ICLR'24杰出论文奖提名
- ▶ 曾主持或参与超10项国家自然科学基金项目和企业合作项目
- ▶ 深耕深度学习理论与应用，发表国际高水平论文50余篇，引用量超过17000次
- ▶ 开创“层归一化前置”技术，被OpenAI采纳并在其技术报告中明确标明被使用

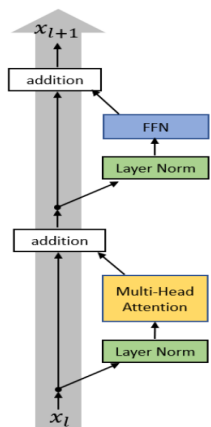


陆一平
课题骨干

北京大学 研究员

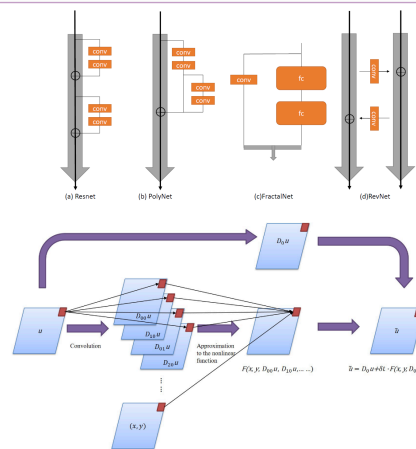
- ▶ 2023年于美国斯坦福大学获得计算数学工程博士学位
- ▶ 2024-2026年在美国西北大学工业工程与管理科学系任助理教授
- ▶ 提出PDE-Net、ODE-Net等模型，将偏微分方程与动力系统结构融入神经网络，相关成果成为合作者董彬教授国际数学家大会报告的重要内容。
- ▶ 获多项数据科学领域新星奖，运筹学顶级期刊Mathematics of Operations Research副编辑

层归一化前置结构设计 (ICML 2020)



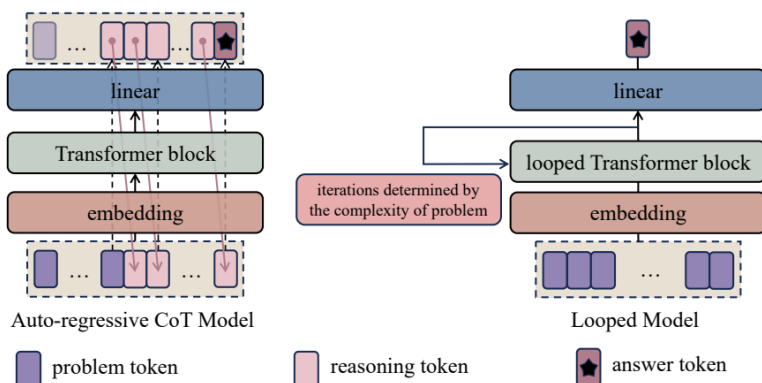
针对Transformer训练稳定性不足问题，提出层归一化前置技术，理论证明其训练稳定性。目前被包括OpenAI在内的开源大模型使用。

基于数值微分方程的网络架构设计(ICML'18 a, ICML' 18 b)

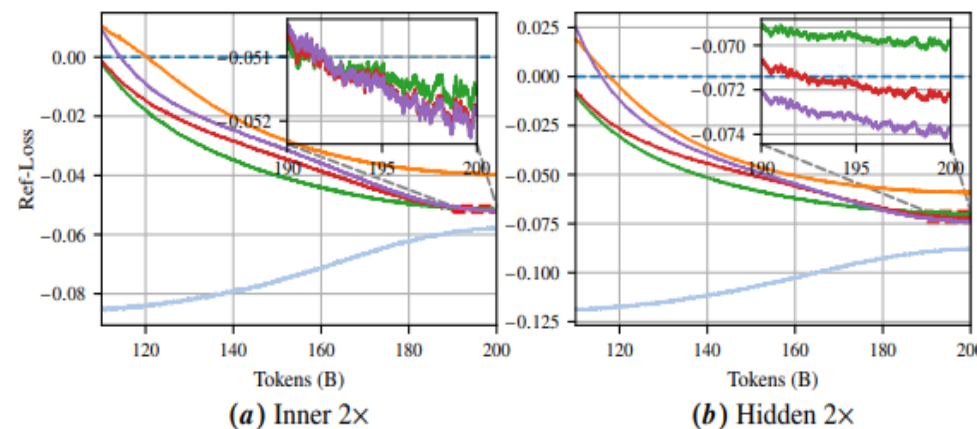


开创性地将数值微分方程理论引入深度神经网络架构设计，建立残差网络与数值离散方法的系统联系，提出具有更高稳定性的网络结构，推动了神经网络与动力系统交叉研究的发展，成为深度学习连续动力系统理论的重要奠基性工作，被后续 Neural ODE 等研究广泛借鉴与拓展。

Looped model高效推理(EACL'26 oral)



跨宽度跨深度参数共享的高效循环语言模型预训练 (ICML'26)



观测到参数共享的深层循环模型在训练中会遇到信号保存损失及RMS层扩展不一致问题，提出针对性解决方案，使深度模型训练稳定。

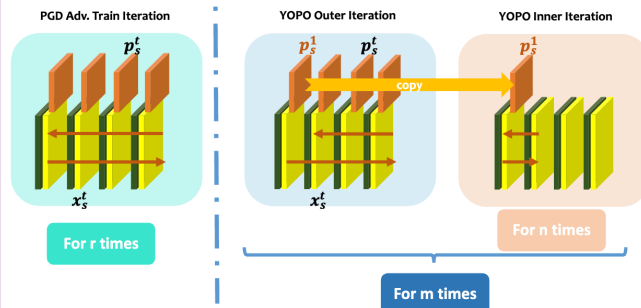


前期工作积累（神经网络优化理论）

神经网络宽度扩展理论（2026）

建立统一的矩阵算子范数优化框架，证明网络宽度无关的光滑性界，揭示 AdamW、Muon 等优化器的宽度依赖机制，实现跨模型宽度的稳定超参数迁移与高效预训练。相关研究成果被 DeepSeek-V4.1-Flash 采用，应用于模型优化算法设计。

基于极大值原理的高效对抗训练（NeurIPS 2019）

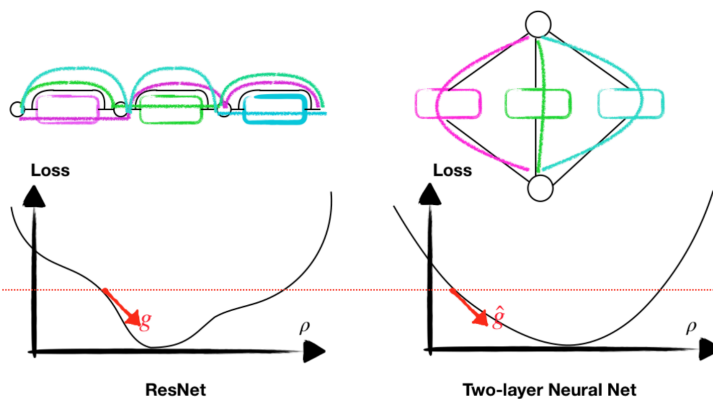


提出 YOPO 对抗训练算法，利用庞特里亚金极大值原理解耦对抗扰动更新与网络反向传播，在保持相近鲁棒精度的同时，将训练速度提升约 **4–5 倍**。

生成式人工智能推理复杂度 (ICML'26)

提出面向大语言模型与扩散模型的 Bellman 误差校正推理框架，利用序贯蒙特卡洛方法融合概率校正机制。理论证明可将大模型推理复杂度从指数级降低至多项式级；进一步提出基于 Girsanov 测度变换的无梯度扩散采样算法，实现生成分布的无偏校正与高效推理。

深层 ResNet 的平均场优化理论 (ICML 2020, ICLR 2024 Spotligh)



建立深层 ResNet 的平均场理论，证明其连续极限下局部极小值均为全局极小值，首次建立多层神经网络在平均场框架下的全局优化收敛保证。



北京大学
PEKING UNIVERSITY

目录

一

国内外现状及发展趋势

二

科学问题与研究目标

三

研究内容与计划进度

四

研究团队与工作基础

五

预期成果与考核指标



优化 方法

从循环动力学与梯度传播机制出发，建立“循环次数—梯度传播—参数/更新范数—有效学习率”理论分析框架，设计适配参数共享机制的优化器与训练策略



预训练 模型

相较于参数量相同规模的标准非共享参数Transformer架构，新方法训练出的Looped架构模型在典型任务中有准确率提升





考核指标(12个月)

优化效率指标	在约2亿参数规模的模型上开展对比实验，以Adam和Muon为基线优化器，分别使用20B tokens进行预训练；新优化方法在达到同等精度的前提下，所需训练数据量较基线减少5%。
准确率指标	在约5亿参数规模的模型上开展对比实验，以参数规模相同的标准非共享Transformer为基线，在相同训练数据量（20B tokens）下，新方法训练的Looped Transformer模型验证集损失（Loss）降低不低于0.02，且在HellaSwag、PIQA、ARC-Easy等涵盖常识推理、物理常识与基础科学知识的典型零样本（Zero-shot）评测任务上，平均准确率提升不低于3%。
专利	申请一项相关技术专利。



中期指标(6个月)

优化效率指标

在约2亿参数规模的模型上开展对比实验，以Adam和Muon为基线优化器，分别使用20B tokens进行预训练；新优化方法在达到同等精度的前提下，所需训练数据量较基线减少2%。

准确率指标

在约5亿参数规模的模型上开展对比实验，以参数规模相同的标准非共享Transformer为基线，在相同训练数据量（20B tokens）下，新方法训练的Looped Transformer模型验证集损失（Loss）降低不低于0.01，且在HellaSwag、PIQA、ARC-Easy等涵盖常识推理、物理常识与基础科学知识的典型零样本（Zero-shot）评测任务上，平均准确率提升不低于1%。



北京大学
PEKING UNIVERSITY

感谢

—— 衷心感谢各位专家，敬请批评指正 ——