

姓名：衣景龙

email: 1026277317@qq.com tel: [17863136229](tel:17863136229)

个人简介

哈尔滨工业大学计算机专业本硕，5 年大模型、搜索与推荐算法研发经验。聚焦 Foundation Model 预训练、模型 Scaling、长上下文建模及高效推理，具备从数据构建、模型架构、规模化训练、性能优化到线上部署的完整经验。

主导 Foundation Model 从 0 到 1 建设，将 Dense 模型扩展至 4B 参数、稀疏模型扩展至百 T 参数规模、训练上下文扩展至 110K；系统实践 Transformer/HSTU、MoE、Flash Attention、Sparse Attention、KV Cache、PD 分离及模型蒸馏，具备 10 人+技术团队管理和万卡集群训练（NVIDIA、HWJ）推理协作经验。

核心能力

- 基础模型预训练：Foundation Model、数据构建、自监督预训练、Scaling Law、下游任务迁移
- 模型架构与 Scaling：Transformer/HSTU、MoE、Dense/Sparse 模型、参数与上下文长度 Scaling
- 长上下文建模：110K 序列训练、Flash Attention、Sparse Attention、SWA、随机长度训练
- 高效训练与推理：Dynamic Batch、KV Cache、PD 分离、量化、算子融合、显存及吞吐优化
- 后训练与知识迁移：SFT、QLoRA、知识蒸馏、对比学习、多任务学习
- LLM 业务应用：CoT Prompt、Few-shot、显式用户表征、冷启动兴趣推理

教育背景

※硕士：· 哈尔滨工业大学 · 计算机科学与技术学院 · 计算机科学与技术专业

● 成绩：90.75 排名：TOP 0.5%

● 研究方向：自然语言处理与人机对话

※本科：· 哈尔滨工业大学 · 计算机科学与技术学院 · 计算机科学与技术专业

● 成绩：88.98 排名：TOP 4%

工作经历

2023.12-今 字节跳动-抖音主站-大模型高级算法专家

- 荣誉：连续多次 top30% 绩效
- 项目：长上下文大模型与高效注意力 | 核心算法负责人
 - 面向推荐系统中用户行为序列中的长期依赖、兴趣探索和中长尾建模问题，设计基于 Transformer/HSTU 的长上下文基础模型，推动模型效果达到线上传统架构水平。
 - 将训练数据组织升级为 Varlen Flash Attention 与 Dynamic Batch，降低变长序列中的 Padding 计算浪费，训练吞吐提升 23%~37%。
 - 针对推荐序列中多类型 Token 及复杂可见性关系导致 Attention Mask 显存开销过大的问题，设计 **压缩 Mask 表示**，单任务显存占用降低 30GB 以上。
 - 设计随机上下文长度训练及推理长度外推策略，使模型适配不同长度分布，训练吞吐提升至基线的 2.2 倍，并取得在线 auc+0.4% 的收益。
 - 设计适配行为序列的 **Sparse Attention**，在保持模型效果的同时将推理成本降低 **50%**，支持更长上下文建模。
 - 通过序列压缩和多类型 Token 融合提升单位上下文的信息密度，扩展有效建模长度并提升 AUC 0.2%。
 - 最终完成生成式推荐系统上线，LT 提升 0.3%，完成全量推全部署。
- 项目：Foundation Model 预训练与 Scaling | 项目负责人，带队 7 人，万卡集群训练
 - 面向流式大模型训练和推理成本高、模型规模受在线算力约束的问题，设计全域、全场景批式预训练范式与 KV Cache 在线服务架构，从 0 搭建数据、模型、训练、评测和部署 Pipeline。
 - 主导 Dense 模型扩展至 4B 参数、稀疏模型扩展至百 T 参数规模、训练上下文扩展至 110K，解决大规模训练过程中的稳定性、显存、吞吐及效果退化问题。
 - 设计基于 KV Cache 的在线推理架构，系统评估 GQA、SWA、CLA、量化、BWA、Linear Attention 及

姓名：衣景龙

email: 1026277317@qq.com tel: [17863136229](tel:17863136229)

Cut-half Attention 等压缩方案，平衡模型效果、缓存规模、计算成本和数据时效性，完成 **4B+百T 模型+110k 上下文** 上线，且 ROI 打正。

- 设计 Foundation Model 向下游流式模型的知识迁移方案，通过预训练表示复用和模型蒸馏，以约 20% 的增量计算成本获得同等 Scale 方案约 80% 的效果收益，整体算力节省约 70%。
- 完成 PD 分离的算法方案设计，联合 NVIDIA 及工程架构团队推进 Attention、SWA、Fusion Bias 和 MoE 相关算子优化，相关能力被多个项目复用。
- Foundation Model 在线实验 AUC 和 LT 均正向显著，推全中。

➤ 项目：LLM 后训练与显式用户表征 | 项目负责人

- 面向新用户和低活用户观测信息不足的问题，提出利用 LLM 世界知识和推理能力生成显式用户表征的建模方案，打通“用户信息—LLM 兴趣推理—标签生成—下游召回”链路。
- 根据用户关键属性组合设计 CoT Prompt 和 Few-shot 示例，引导 LLM 推理潜在兴趣标签，并使用推荐模型完成候选内容召回。
- 在新用户、低活用户场景取得 LT 显著正向收益并完成全量部署，方案被多个内部业务复用。
- 第二阶段融合搜索与推荐行为构建训练数据，采用教师模型生成兴趣表征并进行知识蒸馏，基于 Qwen3 完成线上实验，LT 取得正向收益，推全中。

2021.07-2023.12 腾讯 算法研究员

- 荣誉：连续 **top10%** 绩效；**连续晋升三级**（最后一次晋升成功率 **10%**）；四篇专利；部门卓越之星
- 领域大模型（意图识别）与表征学习（相关性）| 项目负责人，带队 2 人
 - 带队 2 人完成视频搜索领域大模型训练，采用自监督预训练、SFT 和 QLoRA 提升 Query 实体与意图识别效果，线上准召率提升 11.2%。
 - 通过对比学习、精排蒸馏和交叉蒸馏优化搜索粗排模型，离线 AUC 提升 2.98%，在线用户时长提升 1.83%。
 - 将 Query 和 Document 由单向量升级为**多向量表征**，提升多义词及多意图建模能力，在线用户时长提升 2.59%、点击率提升 0.52%。

部分实习经历

2020.07-2020.09 Alibaba • 搜索推荐事业部 • 首页推荐 算法工程师

融合文本与图像信息识别 OCR 推荐文案质量，使可用率由 90% 提升至 95%，在线透出率提升 7%。

2020.04-2020.07 Netease • 伏羲实验室 • 算法工程师

负责对话平台 Slot Filling 模块，将准确率由 45% 提升至 86%，应用于舆情监控平台及游戏业务。

竞赛经历

全国大学生数学建模竞赛**国家二等奖**；CCF 语言与智能技术竞赛第一赛季 18/163，第二赛季 20/93；京东对话挑战大赛 16/256；天池文本对抗大赛 **14/1282**

获奖情况

- 本硕阶段累计获得包括国家奖学金、人民奖学金、企业奖学金等在内的奖学金 20w+。
- 累计获得**国家级**奖项 2 项，**省级**奖项 1 项，**校级**奖项 20 余项。

出版物

A 会一作 2 篇；一作期刊两篇（IEEE Spectrum、智能计算机与应用）；专利六篇（其中一作三篇）；软件著作权一篇；