张配天

namespace.pt@gmail.com namespace-Pt Google Scholar 北京, 深圳

教育经历

中国人民大学 - 高瓴人工智能学院 - 人工智能 - 学术型硕士 (导师：窦志成) 2022.9 – 2025.6

中国人民大学 - 信息学院 - 计算机科学与技术 - 本科 2018.9 – 2022.6

项目经历

检索模型 & 检索索引

- **智源通用向量表征模型: FlagEmbedding**
  - *(介绍)* 一系列精准通用的向量表征模型，适用于用于通用检索和检索增强任务，包括：
    - BGE: 轻量、强大、通用的表征模型；
    - BGE-M3: 支持多种语言、多种检索方式、多种检索粒度的表征模型；
    - LLM-Embedder：支持 LLM 多样化的检索增强场景的表征模型。
  - *(角色)* 参与 BGE、BGE-M3 的数据构建和模型训练，主导 LLM-Embedder 的全部流程。
  - *(成效)* BGE 在权威向量表征模型榜单 MTEB 上达到 SOTA 效果；整个模型系列在 Huggingface 上**下载量全中国第一**，被 LlamaIndex、Azure 等数个大模型框架和云服务集成；开源仓库获得了 6K+ 标星；**对应 3 篇论文均被 CCF A 类会议录用** (一篇一作)。
- **华为-人大合作项目基于与训练生成模型的端到端信息检索系统: TSGen**
  - *(介绍)* 使用一组关键词集合作为文档标识符，设计序列等变解码使得自回归模型可以用任何顺序生成相关文档对应的词集合，并优化训练方式令模型能够自行探索词集合的最优生成顺序。
  - *(介绍)* 主导算法设计、数据构建、模型训练、测评等全部流程。
  - *(成效)* 克服了传统的序列标识符一步错步步错的问题，从而获得显著的检索精度提升，并对训练阶段未见过的新文档有更强的建模能力；**对应一作论文被 CCF A 类会议录用**。
- **高效向量检索索引: Hybrid Inverted Index**
  - *(介绍)* 使用向量聚类中心和关键词倒排列表协同工作的近似最近邻搜索索引。
  - *(介绍)* 主导算法设计、数据构建、模型训练、测评等全部流程。
  - *(成效)* 该索引无需监督训练即达到和 HNSW 达到接近的精度和时延，而空间占用仅有其十分之一；经过监督训练后，精度能够显著超过 HNSW；**对应一作论文被 CCF B 类会议录用**。
- **工程实践：中国人民大学智能类案检索系统**
  - *(介绍)* 该系统能够基于关键词匹配和语义向量相似度完成对超过 10M 裁判文书的检索，同时支持分片搜索、解释搜索结果等高级功能。
  - *(角色)* 独立完成数据收集，模型训练，前后端开发，系统部署。
  - *(成效)* 为中国人民大学第一届法律大数据分析挑战赛提供支持，被教师和学生广泛使用。

长文本语言模型

- **长上下文窗口扩展: LongLLM-QLoRA**
  - *(介绍)* 通过优化 RoPE 位置外推技术和构造长依赖训练数据，在扩展语言模型上下文窗口的同时增强其对窗口内部信息的利用能力。
  - *(角色)* 主导算法设计、数据构建、模型训练、测评等全部流程。
  - *(成效)* 基于 4.5K 长依赖数据，在 Llama-3 发布一周内将其上下文窗口扩展 10 倍，在下游长文本榜单上效果远超社区内同期工作；优化后的位置外推技术简单易用，效果超越 YaRN。
- **长上下文高效计算: Activation Beacon**
  - *(介绍)* 从序列维度压缩 KV cache，灵活处理各种压缩率、输入长度，使得 LLM 能够在有限的上下文窗口中处理更多的内容，同时节省显存占用、减少自注意力计算量。
  - *(角色)* 主导算法设计、数据构建、模型训练、测评等全部流程。
  - *(成效)* 在较短文本上的轻量化微调即可建立压缩能力，在 Mistral、Llama-2/3、Qwen-2 等 LLM 上能够以 **x4 乃至更高压缩率**完成高质量的长上下文 KV 压缩，从而等比率扩展上下文长度；兼容于其他维度的 KV 压缩算法和上下文窗口扩展技术；对应一作论文 CCF A 类会议在投。

实习经历

百度 (文心 Lite 团队) 2024.6 – 至今

主要方向: 长文本语言模型的技术研究与应用。

北京智源人工智能研究院 (知识检索与计算组) 2023.6 – 2024.6

主要方向: 向量表征模型的研究与应用，长文本语言模型技术研究。

微软亚洲研究院 (社会计算组) 2021.6 – 2022.4

主要方向：高效向量索引技术研究，高效新闻推荐技术研究。

学术经历

发表一作论文 3 篇 (2 篇 CCF-A，1 篇 CCF-B)，一篇一作论文在投 (CCF-A)，Google Scholar引用量 305。

[1] *(Under Review)* Soaring from 4K to 400K: Extending LLM's Context with Activation Beacon
Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, Zhicheng Dou

[2] *(ACL'24)* Retrieve Anything To Augment Large Language Models
Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, Jian-Yun Nie

[3] *(EMNLP'23)* Hybrid Inverted Index is A Rubust Accelerator for Dense Retrieval
Peitian Zhang, Zheng Liu, Shitao Xiao, Zhicheng Dou, Jing Yao

[4] *(SIGIR'24)* Term-Sets Can Be Strong Document Identifiers For Auto-Regressive Search Engines
Peitian Zhang, Zheng Liu, Yujia Zhou, Zhicheng Dou, Zhao Cao

[5] *(SIGIR'24)* C-pack: Packaged Resources to Advanced General Chinese Embedding
Shitao Xiao, Zheng Liu, **Peitian Zhang**, Niklas Muennighof

[6] *(ACL'24)* BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation
Jianlv Chen, Shitao Xiao, **Peitian Zhang**, Kun Luo, Defu Lian, Zheng Liu

[7] *(ACL'24)* LM-Cocktail: Resilient Tuning of Language Models via Model Merging
Shitao Xiao, Zheng Liu, **Peitian Zhang**, Xingrun Xing

[8] *(ACL'24)* INTERS: Unlocking the Power of Large Language Models in Search with Instruction Tuning
Yutao Zhu, **Peitian Zhang**, Chenghao Zhang, Yifei Chen, Binyu Xie, Zhicheng Dou, Zheng Liu, Ji-Rong Wen

[9] *(Under Review)* Are Long-LLMs A Necessity For Long-Context Tasks?
Hongjin Qian, Zheng Liu, **Peitian Zhang**, Kelong Mao, Yujia Zhou, Xu Chen, Zhicheng Dou

技能

编程技能	Python, C++, HTML, CSS
专业技能	PyTorch, Transformers, Faiss, Elasticsearch, Django

获奖

中国人民大学三好学生，中国人民大学优秀毕业生，大学生创新实验计划国家级立项（优秀结项）负责人，文体优秀奖学金