个人简历 | PERSONAL RESUME

徐名宇

出生年月： 1999.05　　邮箱： xumingyu2021@ia.ac.cn

联系方式： 18801200516　　职位： 大模型算法工程师

教育背景

中国科学院自动化研究所　多模态人工智能系统全国重点实验室　　2021.09-2024.06

模式识别与智能系统专业　工学硕士　保研　学术型硕士

北京大学　数学科学学院　　2017.09-2021.06

信息与计算科学专业　理学学士　统招　全日制本科

实习和工作经历

- 2023.11-至今　百川智能　实习转正 大语言模型基座预训练

1.模型结构优化

a) 从多头注意力热启动到推理友好的分组注意力(GQA)结构。探索基于最小二乘法的多头合并方法,相比google原始的直接平均的方法收敛速度快3倍。

b) 在业内率先探索高效的模型，如滑动窗口注意力、分组查询注意力和分层注意力的混合，在保证性能不怎么影响的情况下在长窗口下降低kv cache32-128倍，降低注意力机制的计算量8倍左右。(在5月份就已基本完成这种高效结构的验证，而业内最早是c.ai在6月份发布的blog, 目前yi最新发布的Lightning和混元的moe模型均采用了这种Attention结构)。除此之外还进行了moe的共享专家池实验，在降低了总参数量的需求。

c) 除此之外进行很多认知和探索的实验。譬如关注transformer的理论限制，展开对loop transformer的实验，利用数据内生的统计特性设计循环次数的监督信号。研究LLM的特征，关注非transformer类模型结构的进展与限制。

2.长窗口训练和优化

a)从理论上研究了关于rope中关键参数base和可支持的上下文关系之间的关系，为模型中位置编码base的设置提供了理论和经验上的见解。这种关系不仅存在于长文本微调阶段，同样也存在于预训练阶段。

b)对长窗口训练进行了varlen和ulysses的适配, 以及探索varlen对长窗口训练的影响。

c)探索了长文本模型所需的数据配比进行了研究，以及高效地利用短数据训练长窗口模型。

部分论文

读研期间的研究方向为机器学习 (弱监督学习、分布外检测) 和多模态 (情感识别、幽默检测) ,本人学习能力强, 实习工作期间转向大模型方向 (模型结构优化、长窗口训练和优化), 已正式发表的一作CCF-A文章一篇。并担任TKDE等期刊、会议审稿人。本人总共以第一作者或者共同第一作者发表CCF-A论文3篇, CCF-B论文1篇。以下是我的以第一作者或者共同第一作者发表的文章:

大语言模型

1.Base of RoPE Bounds Context Length. （NeurIPS2024）**Mingyu Xu***, Xin Men*, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, weipeng chen

2.Shortgpt: Layers in large language models are more redundant than you expect. （Submit to ICLR2025）Xin Men*, **Mingyu Xu***, Qingyu Zhang*, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, Weipeng Chen

机器学习与多模态

1.ALIM: Adjusting Label Importance Mechanism for Noisy Partial Label Learning. (NeurIPS2023)**Mingyu Xu***,Zheng Lian*, Lei Feng, Bin Liu, Jianhua Tao

2.VRA: Variational Rectified Activation for Out-of-distribution Detection.（NeurIPS2023）**Mingyu Xu**,Zheng Lian, Bin Liu, Jianhua Tao

3.Pseudo Labels Regularization for Imbalanced Partial-Label Learning.(ICASSP2024,oral)**Mingyu Xu;** Zheng Lian; Bin Liu; Zerui Chen; Jianhua Tao

4.Humor Detection System for MuSe 2023: Contextual Modeling, Pseudo Labelling, and Post-smoothing.（ACMMM2023workshop）**Mingyu Xu**, Shun Chen, Zheng Lian, Bin Liu

5.Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent model.（ACMM2021workshop) Licai Sun*, **Mingyu Xu***, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, Yuan Cheng

技能

- 熟悉python/matlab/C/Linux系统的操作;
  熟练使用Pytorch, deepspeed等深度学习和机器学习框架;
  在千卡集群上进行过 2B (50次左右)、7B（10次左右）、20B（10次左右）参数模型的预训练;

相关竞赛

- ACMMM2024 workshop MUSE HUMOR（多模态幽默检测竞赛） 第一名
- ACMMM2021 workshop MUSE CAR（多模态情感识别竞赛） 第一名
- 2016中国数学奥林匹克（CMO）一等奖