董汉德

手机：(+86) 17737085538 · **邮箱：**donghd66@gmail.com

个人主页：https://donghande.github.io/

教育背景

中国科学技术大学，电子工程与信息科学系，导师：何向南教授，*硕士* 2019.09 - 2022.06

- **研究方向**：图神经网络、数据挖掘、推荐系统中的偏差与去偏。

中国科学技术大学，近代物理系，**赵忠尧英才班**，*本科* 2015.09 - 2019.06

工作经历

idea 研究院，跟进代码理解与生成的研究进展，打通代码预训练大模型的流程 2022.07 - 至今

实习经历

idea 研究院，学习并参与代码搜索的前沿技术调研与探索 2022.04 - 2022.06

美团-NLP 中心，使用用户行为数据及评论数据挖掘商品之间相关性研究 2021.09 - 2022.01

华为-无线部门，使用机器学习技术在无线信号之间进行预测 2018.07 - 2018.09

主要发表论文

On the Equivalence of Decoupled Graph Convolution Network and Label Propagation; WWW 2021; **Hande Dong**, Jiawei Chen, Fuli Feng, Xiangnan He, e.t.c.; CCF-A 类会议, 第一作者

AutoDebias: Learning to Debias for Recommendation; SIGIR 2021; Jiawei Chen*, **Hande Dong***, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, Keping Yang; CCF-A 类会议, 共同一作

Bias and Debias in Recommender System: A Survey and Future Directions; TOIS; Jiawei Chen, **Hande Dong**, Xiang Wang, Fuli Feng, Meng Wang, Xiangnan He; CCF-A 类期刊, 第二作者

谷歌学术显示，**本人论文被引用数达 480** 次，github 仓库被 star 达 136 次。全部论文见谷歌学术主页。

主要项目经历

预训练大语言模型的代码生成探索及落地，*idea 研究院工作项目，主要负责人* 2023.03-至今

随着预训练大语言模型的兴起，代码生成领域取得了巨大的突破。我跟进了当前商用、开源的各种预训练代码生成模型，**打通了预训练大模型的代码生成的全套流程**：下载并处理 The Stack 数据集；熟悉了 deepspeed, Megatron-LM, peft 等训练框架；从头预训练代码生成模型，并在 100M, 300M, 600M, 1B 四个模型规模下达到了和 Codex 相当的性能；部署模型层面，并基于 triton, fastAPI, uvicorn 框架部署模型，并和前端同事一起完成了 vscode 代码生成插件的内测版本。

代码搜索模型的微调方式优化，*idea 研究院研究项目，主要参与人* 2022.10-2023.01

目前代码搜索多采用对偶的模型来微调预训练模型，分别编码自然语言和代码，这无法建模自然语言和代码的细粒度关联。我提出了使用交叉模型来建模自然语言和代码之间的相关性，并提升了模型的性能约 4 个百分点。相关研究申请了专利一项（专利号 202310015054.4）。

利用用户数据挖掘商品关系，*美团实习项目，主要参与人* 2021.09-2022.01

美团运行多年积累了丰富的用户数据，挖掘用户数据有助于建模商品之间的关系。我参与指定落实了两项策略：通过挖掘用户行为数据（例如点击、购买）之间时间的时间差挖掘相关的商品及品类，并通过建立图结构进一步优化方案；此外，我基于用户评论数据微调 T5 模型，并用此模型挖掘潜在图谱关系。以上策略挖掘了可能存在关系的商品，并送给标注人员，提高了整体的标注效率。

推荐系统中偏差与去偏，*研究生科研项目，主要参与人* 2020.05-2021.02

数据驱动的推荐系统范式导致了偏差的产生，需要合理的策略来去除偏差。我广泛阅读并整理相关文献，从风险与经验风险不一致的角度定义推荐系统中的偏差，并参与提出了自动去偏算法（AutoDebias），实验验证其优越性。相关论文发表于 SIGIR 2021 和 TOIS，并申请专利一项（专利号 202110424290.2）。

图卷积网络的机理研究，*研究生科研项目，主要负责人* 2019.02-2020.10

图卷积网络的工作机理不透明，缺乏可解释性。我通过梯度的角度证明了解耦图神经网络的工作机理在于利用标签传播进行的数据增强，并指出了当前模型存在的问题而且提出了解决方案，提高了稳定性和鲁棒性差，并做实验进行验证。相关论文发表于 WWW 2021，并申请专利一项（专利号 202011591264.0）。

主要技能

深度学习、机器学习：掌握主流的机器学习及深度学习算法，尤其是在自然语言处理、推荐系统、图卷积网络等领域。熟悉 pytorch，transformers 等训练框架。
大模型：紧跟大模型前沿技术，具有一定训练大模型的相关经验。掌握数据并行、模型并行、张量并行等并行化技术的原理，并能够使用 deepspeed，Megatron-LM 分布式训练大模型。
代码智能：深入了解预训练技术在代码智能中的应用，并掌握了数据、模型训练、模型部署的全套流程。
英语水平：六级 516 分。

个人荣誉

研究生学业奖学金一等，中国科学技术大学 2019、2020 年 12 月
优秀学生奖学金金奖，中国科学技术大学 2017 年 12 月
校级优秀毕业生、省级品学兼优毕业生，科大，安徽省 2019 年 6 月
优秀学生干部及优秀团干，中国科学技术大学 2015 年-2020 年多次

其他经历

网页维护与服务器管理，数据科学实验室。维护实验室官网和实验室服务器 2021.03-2022.03
助教，课程“数据科学基础”。批改学生作业，上习题课，解答同学问题 2020 年秋季学期
助教，课程“计算机程序设计”。指导学生上机，批改作业，回答学生问题 2017 年秋季学期

学生工作及所获荣誉

班长，6 系 2019 级硕士 2 班。任职期间，班级获得校级先进班集体 2019.09-2022.06
团委学生助理，物理学院 2018.09-2019.06
团支书，物院 15 级本科 1 班。任职期间，团支部获得安徽省五四红旗团支部 2015.09-2019.06