宋浩浩

男 · 25 岁 · NLP、ML、DL

(+86) 189-378-88773 · songhaohao2018@cqu.edu.cn · https://songs18.github.io

***2021* 年博士申请**

教育经历

重庆大学 · 计算机学院（保送） *2018.9 - 至今*

计算机科学与技术 · 学硕

河南大学 · 计算机与信息工程学院 *2014.9 - 2018.6*

网络工程

科研成果

Embedding Compression with Right Triangle Similarity Transformations [一作]

accepted by ICANN2020 (CCF-C)

- **问题背景**：词嵌入作为 NLP 的底层技术被普遍使用，然而，传统的词嵌入是一个高维的浮点数矩阵，占据大量的内存和计算资源，限制了 NLP 的应用场景和计算速度。
- **任务目标**：学习低维的、二值化词嵌入表示，降低内存使用和提高计算速度。
- **提出方法**：基于直角三角形相似性变换的词嵌入压缩方法 (RTST)。学习一个从原始空间的一组直角三角形到目标子空间对应三角形的相似性映射，借助一个带有离散化项的损失函数的孪生神经网络来实现。
- **有效性**：对主流的 GloVe 和 Word2Vec（变体）离散化学习，从内部任务（语义相似度、词类比）和外部性质（文档/问题分类、情感分析）评估学习到的编码。实验结果显示，在 RTST 方法的帮助下，当压缩率未 2.7% 时，性能下降仅 1.2%，超出了该领域的最好结果。进一步的消融实验验证了方法的合理性。
- **创新性**：基于直角三角形相似性变换的嵌入压缩方法可以看作是机器学习中的流形学习和神经网络的结合。该方法独立提出，并非在已有方法上的改进。并且方法有效。

Learning Discrete Sentence Representations via Construction & Decomposition [一作]

accepted by ICONIP2020 (CCF-C)

- **问题背景**：虽然资源消耗型的深度学习模型推动了 AI 各个领域的发展，然而，昂贵的硬件设备门槛阻碍了这些模型的实际应用。本文提出一种新的离散编码学习方法，旨在解决模型压缩中的一个基本问题：向量表示压缩。
- **提出方法**: 生活中，我们可以使用直尺度量盒子的大小，然后根据得到的尺寸和直尺建造一个相同尺寸的盒子出来。受到这种方法的启发，这篇论文中，我们构造性生成一批一致的、无偏见的、完备的锚向量作为我们的尺子，然后在构造（Construction）阶段建立锚向量和输入向量的相似度矩阵，然后在分解（Decomposition）阶段中，将 PCA 处理过的锚向量作为相似矩阵的一个固定因子分解。我们使用数学推导计算出每一个比特的最优解，然后使用离散坐标下降法解决这个混合整数的优化问题。
- **有效性**: 我们选择句子向量压缩作为我们的评估任务，在和该领域的最新进展 ACL2019 的对比中全面获胜。相应的消融实验证明了方法的合理性。
- **创新性**: 先构建、再分解的方法在离散化学习中还没有出现过，完整的数学推导保证了该算法的有效性，论文中提出的锚向量假设的合理性在消融实验中得到了证实。在 ICONIP 的论文评审中，该方法获得了 2 位审稿人一致的高度评价（一个强烈接收、一个接收）。

此外，作为联作参与论文有：

Four-way Bidirectional Attention for Multiple-choice Reading Comprehension *CoNLL2020 在审*

科研经历

国家重点支撑计划（编号：JG2019071） *2019.7 - 至今* *科研项目*

从白话文到诗词的风格迁移 *2019.7* *自我实践*

- **背景**：在大组讨论会上听取了作诗机报告，回去后做了背景调研，发现从白话文到诗词的风格迁移工作还是空白。考虑到当前白话文使用普遍，诗词高雅，在不改变语义的情况下实现从白话文到诗词的风格迁移在朋友相聚即兴赋诗等场景有实用价值。属于应用型创新
- **提出方法**：对任务进行抽象。该任务可以看作一个 seq2seq 任务，输入白话文，输出诗词。然而，该任务的难点在于训练语料的稀缺。通过网络爬虫、从《唐诗三百首》抽取翻译两种手段，获取到 2262 条翻译双语预料。使用 seq2seq 模型学习：双向 GRU 对白话文进行编码，然后对中间隐状态进行解码。其中涉及到注意力机制、集束搜索、teacher forcing 等经典技术。
- **效果**：由于训练语料质量不高和语料数量稀少，在后续的评估中模型表现差强人意。

新闻核心实体词识别和情感分析 *2019.4* *小组合作*

- **简介**：必修课程要求自由选择数据集，然后在该数据集上进行数据挖掘分析作业。我们选择了搜狐公开的新闻数据集，任务是找出一篇新闻中的不超过 3 个核心实体词，并识别出该实体词对应的情感类别。我们小组在官方 baseline 上提出了手工设计的 5 个特征和设计的 RNN 网络，在 F1 评估指标上实现了一定提升。因为我们设计的 5 个特征和详尽的实验（不同模型间的对比实验、验证特征有效的消融实验），我们小组获得了课程班级最高分（94.4）。

项目及实习

- **京东爬虫** *项目* *2019.4 - 至今*

京东 6 品类（洗衣机、牙膏、单反相机、数据线、大米、红酒）自营商品每日排名、价格、当日商品评论等信息爬虫及 Linux 服务器线上维护，涉及代理、MySQL、多线程、异常处理、日志记录、crontab 定时、日邮件统计等技术，月爬取 300w+ 条记录。

- **重庆市肿瘤医院细胞计数系统** *项目* *2019.7 - 2019.8*

根据染色后的切片图片，统计上皮型、间质型、混合型细胞计数，自动生成 word 报告文档。使用 cv2 读取图片并转换为灰度图，实现不相交集算法计数，编辑 word 模板文档并生成检测结果报表。

- **北京智能一点算法实习生 [智能客服方向]** *实习* *2018.6 - 2018.9*
- **开封市通许县大数据监管 WPF 客户端二期** *项目* *2017.9 - 2018.4*

能力

- **创新能力**

在算法创新上，在新闻核心实体词提取上提出 5 个有效特征，基于直角三角形相似性变换的嵌入离散化方法结合了机器学习中的流形学习和神经网络方法，先构建再分解的 C&D 方法新颖并具有数学和实验上的两重保证；在应用创新上，根据一次作诗机讲座提出白话文到诗词的风格迁移应用性创新。

- **技术及工程能力**

超过 3 年的 Python 使用经验，熟悉 Numpy、Pandas、Tensorflow、Pytorch、gensim、NLTK 等数据建模分析包；本科接受专业的计算机教育，对 Java、C++、C#、matlab、MySQL 等语言有较长时间的学习过程；熟练掌握数据结构、面向对象等概念，对代码重构、并发、异步、git 熟悉，熟练使用 Linux 服务器并有一定运维能力，有良好的代码风格和总结撰写 wiki 能力。

- **复现论文能力**

对公布完整代码的论文可以迅速 clone 仓库并开展实验；对公布部分代码的论文可以快速定位缺少的模块补全；对三方实现的代码严格对比论文方法描述和实现是否一致；对未提供代码的论文可以根据方法描述使用 Python 实现。已实现的包括 fasttext、基于矩阵分解的系统过滤、离散协同过滤等。

- **问题抽象能力**

具有理清问题主体和主体之间关系的能力，和解决问题的能力。

- **办事及团队合作能力**

思路清晰，有毅力，具备抗压能力；在国家重点支撑计划、开封市通许县项目等中表现出良好的合作能力。

奖励及其他

- 英语水平：6 级
- 研一、研二全额学业奖学金
- 重庆大学优秀研究生