边圣陶

- 出生年月：1995-06
- 新南威尔士大学 计算机科学硕士
- Github

联系方式

- 手机：15557537168
- Email：bian.shengtao@gmail.com
- 微信号：15557537168

教育经历

- **硕士　人工智能与数据科学　工程学院　　新南威尔士大学**
  2020.09 ~ 2023.02
- **本科　电子商务　法学（双学位）　信息管理学院　山西财经大学**
  2014.09 ~ 2018.07

实习经历

京东集团股份有限公司 京东工业
NLP算法实习生
起止时间：2022-5～至今

一、用tensorflow2复现simCSE模型。
二、对simCSE模型用京东内部数据进行二次训练。
三、参与对京东工业大脑的产品迭代，将神经网络检索准确率从76%提高到97%。
四、用CLIP模型，做图像-文本跨模态检索。将大量电商产品图片与它的描述生成图片文本对，投入CLIP做训练。补全这一文字-图像搜索的空缺。

杭州海康威视数字技术股份有限公司 研究院
NLP算法实习生
起止时间：2021-2～5

一、研究金融领域事件抽取方向论文，如Doc2EDAG，GIT等，并撰写算法综述。
二、研究超长文本表征的方法，如roformer，层次分解位置编码等让bert突破512个字符限制的方法，复现代码并撰写算法综述。
三、研究并整理文本匹配方向的算法方向，如sbert，simCSE，esimCSE等，复现代码并撰写算法综述。

树根互联信息科技有限公司 AI lab
NLP算法实习生
起止时间：2021-10-01～2022-01-15

一、负责整理清洗问答对，并录入数据库。
二、用elasticsearch库建立问答对的倒排索引。用elasticsearch做短语层面的相似性计算。
三、用simbert将题库中的标准问题和用户提问转化成句向量，用faiss存储标准问题向量并对用户提问通过余弦相似性计算获得相近语义的标准问题。

四、统合两者返回的topk以及score，对超过期待阈值的则返回问答库中的标准问题和回答，并对相似度低的问题录入陌生问题数据库中。
五、用自产数据对原有simbert预训练模型继续fintune。
六、基于自研数据进行实体和关系抽取，并整理。
七、将抽取的实体和关系存入Neo4j中，建立知识图谱。

项目经历

- 基于Jina和streamlit的 Video Clip Extraction by description
  起止时间：2022.03 ~ 05

  项目描述
  基于streamlit的框架搭建前端，基于Jina的神经搜索框架搭建后端，部署CLIP模型。对上传的视频进行抽帧，并暂存为一系列图片，通过CLIP的强大图文或文图搜索能力，实现通过文字描述定位视频的关键帧，映射到视频的start时间节点和end时间节点，剪辑视频并返回。

- 自然语言理解进阶探索
  起止时间：2022.01 ~ 04

  项目描述
  基于bert-style预训练模型，Dataset: Multi-genre natural language inference (MNLI) dataset ，对前沿sentence embedding做进阶探索。目前主要工作 对Separate encoding -> Interactive embedding concatenation + 3-way MLP classifier范式，simCSE和promt-Bert进行复现，并探索进一步的改进

- 新冠疫情相似句对文本匹配
  起止时间：2021.12 ~ 2022.01

  项目描述
  基于新冠疫情相似句对pointwise数据集，bert以及ernie预训练模型，用sentence-bert双塔模型以及simCSE模型，对模型进行fine-tune，然后将title和query转化为句向量，存入数据库中，工问答系统使用，这个体系比相对于单纯用基于bert的句向量有4%的提升。

技能清单

- 编程语言：Python/C++/SQL/CYPHER
- 机器学习框架：PyTorch/Tensorflow/Keras/Scikit-learn
- Web框架：Flask/Scrapy/fastapi
- 数据库相关：MySQL/PgSQL
- No-SQL：Redis/mongodb/neo4j
- 版本管理、文档和自动化部署工具：Git/docker

Shengtao Bian

Email: bian.shengtao@gmail.com

Home phone: +8615557537168

Social URL: www.bianst.cn

Education

University of New South Wales

Masters (or equivalent), Artificial Intelligence

2020 – 2023 February | New South Wales , Australia

Shanxi University of Finance and Economics

Bachelors (or equivalent), E-Commerce/Electronic Commerce

2014 – 2018 | Shanxi , China

Experience

MLE Intern (NLP)

JD.com, Inc.

May 2022 – Present

1. Replicating the simCSE model with Tensorflow2
2. Fine-tuning the simCSE model with JD's internal data
3. Participated in the product iteration of JD Industrial Brain, improving the accuracy of neural network retrieval from 76% to 97%.
4. Construct video-text cross-modal retrieval with CLIP model. Generate image text pairs from a large number of e-commerce product images with its descriptions and put them into CLIP for fine-tuning. Fill in the gaps of text-image search.

MLE Intern (NLP)

Hikvision Digital Technology Co.,Ltd.

February 2022 – May 2022

1. Research on event extraction direction papers in finance, such as Doc2EDAG GIT, etc., and write algorithm reviews.
2. Research on methods of ultra-long text characterization, such as roformer, hierarchical decomposition position coding, etc. to let bert break the 512 character limit, reproduce the code and write a review of the algorithm.
3. Research and organize the algorithm direction of text matching, such as sbert simCSE esimCSE, etc., reproduce the code and write a review of the algorithm.

MLE Intern (NLP)

Sany Heavy Industry Co.,Ltd

October 2021 – January 2022

1. Responsible for collating and cleaning Q&A pairs and entering them into the database.
2. Build inverted index of Q&A pairs with elasticsearch. Use elasticsearch to do similarity calculation at phrase level.
3. Use simbert to transform the standard questions and user questions in the question database into sentence vectors. Use faiss to store the standard question vectors and obtain the standard questions with similar

semantics by cosine similarity calculation for the user's questions.

4. Combine the topk and score returned by the above two methods. Return the standard questions and answers from the Q&A database if they exceed the expectation threshold, and enter the questions with low similarity into the database of unfamiliar questions.
5. Fine-tune the original simbert pre-training model with self-produced data.
6. Entity relationship extraction based on self-generated data.
7. Store the extracted entities and relations in Neo4j to build the knowledge graph.