周 晨 星

年龄：24
185 8305 3106
哈尔滨理工大学
cxzhou7@163.com

教 育 背 景

2017.09 – 2021.06	哈尔滨理工大学	自动化(自动化学院)	一本
2021.09 – 至今	哈尔滨理工大学	电子信息（自动化学院）	硕士在读

技 能 情 况

- 熟练掌握 python 基础知识，熟悉 linux 指令，具有良好的面向对象编程思想。
- 熟练掌握 pytorch 和 Tensorflow 深度学习框架，并对网络训练流程有深入理解。
- 精通 CNN、RNN、GRU、LSTM、Transformer、BERT、HMM、CRF、InstructGPT 等深度学习和机器学习算法的应用和原理，并在项目中熟练运用。
- 熟练掌握 NLP 关系抽取、文本分类、实体识别、答案评分等任务，并有相关的实战经验。
- 有复现开源社区代码能力，熟练使用 Vscode、Pycharm 等开发工具，熟练使用 git 命令。
- 熟练使用 fairseq 工具包，熟悉在集群中训练网络模型。
- 熟练使用 Excel、PPT、Word 等办公软件。

实 习 经 历

科大讯飞股份有限公司 **助理研究算法工程师** **2023.1.3-2023.4.3**

英语口语考试评分项目

项目描述：通过使用深度学习算法，搭建端到端系统。实现输入为学生作答的口语考试答案，输出为学生的考试分数。

主要技术：pytorch + Huggingface + fairseq + pandas + 集群训练

责任描述：1、对历史模型进行改进，历史模型采用 pipeline 形式，先训练由学生作答到给出人工评语（主语缺失、谓语形式错误等），再通过人工评语得到最终得分。首先将人工评语从中文改成更专业的英文人工评语；使用“[cls]问题[sep]学生作答[sep]评语[sep]”和“[cls]答案[sep]学生作答[sep]评语[sep]”来替换只使用评语作为输入；再通过历史数据训练基底模型，并在各省份的定标集上进行 finetune。2、使用端到端的系统完成口语考试的评分，在 huggingface 上调研适合的模型，使用“[cls]答案[sep]学生作答”作为模型的输入，直接输出考试得分。使用数据增强的方法扩充数据，采用 bert 的 mask 方法对正确答案进行 mask 操作以降低过拟合，并在集群上进行训练。

工作结果：1、历史模型改进后有提升但不明显，在 finetune 前改进前相关度 0.793，改进后为 0.818。2、端到端系统改进后的模型在 finetune 前，bert-large 的相关性从 0.866 到 0.877，albert-large 从 0.887 到 0.891，roberta-large 从 0.868 到 0.876。结果表明端到端系统相关性更高，且经过数据增强后相关性有提升。

项 目 经 历

“贴吧文本关系抽取”项目

项目描述：通过深度学习算法对百度贴吧文本中的主体、客体以及他们之间的关系进行学习，从而实现对未见过的句子进行三元组关系抽取。详细步骤已上传博客：
https://blog.csdn.net/weixin_49327481/article/details/128092238?spm=1001.2014.3001.5502

主要技术：pytorch + Hugging face + Self-Attention + Bert +服务器训练

责任描述：1、自定义 Dataset 函数实现对数据的**批量提取**，方便后续训练。2、使用 transformers 库中 BertTokenizerFast 模块对文本序列进行中文分词，返回分词后文字 id 值和偏移量，从而解决在输入文本中的中英文混用情况时**分词不可逆问题**。3、对预测值的损失设置不同的平衡权重系数，避免**数据不均衡**带来的训练问题。4、在预测主体时利用自注意力机制来增强上下文语义信息，提高主体识别准确度。5、在 kaggle 云服务器上使用 GPU 加速网络模型的训练。

工作结果：1、对预测出来的损失值使用不同的权重系数后，f1 分值从 0.27 提升至 0.62。2、通过分词时的偏移量信息可以正常返回原文中的位置。

“商品评价实体情感识别”项目

项目描述：通过深度学习算法，对商品评价进行分析，得到评价中的实体位置和实体对应的情感分析(好评/差评)。详细步骤已上传博客：
https://blog.csdn.net/weixin_49327481/article/details/127578363?spm=1001.2014.3001.5502

主要技术：pytorch + Hugging face + Bert + Self-Attention + Bi-LSTM + CRF + 服务器训练

责任描述：1、对训练样本进行预处理，形成统一格式。2、自定义 dataset 类来进行数据的批量读取。3、使用 **BERT** 和 **Bi-LSTM** 网络对实体位置进行预测，后接上 **CRF 层**来进行校正。4、模型采用分块思想进行训练，先要得出实体的准确位置，再通过准确位置对情感进行预测，故调高实体预测部分的 loss 权重。5、在进行商品情感预测时，将句子向量和实体附近的特征进行拼接后输入到 **self-attention** 层来提取整个句子的情感特征。

工作结果：1、实体位置预测部分的 loss 权重不调高时，在训练集上的 f1 分值为 0.8 左右，验证集上 f1 值为 0.7 左右，调高时效果虽有提升但很小。2、将 bert 参数都设置为可学习的后，在训练集上 f1 分值为 0.95 左右，验证集上为 0.8 左右。

获 奖 情 况

硕士一年级一等学业奖学金　　硕士二年级一等学业奖学金　　Cet6

校 园 经 历

2017.09 – 2021.06　　学习委员　　2023.08 – 至今　　深信服 24 秋招校园大使