Artificial Neural Networks

期末项目: 中-英机器翻译

数据集简介

- 数据集说明：压缩包中共有 4 个 jsonl 文件，分别对应着训练集（大）、训练集（小）、验证集和测试集，它们的大小分别是 100k、10k、500、200。jsonl 文件中的每一行包含一个平行语料样本。模型的性能以测试集的结果为最终标准。
- 若计算设备受限，可以仅使用训练集（小）中的 10k 平行语料进行训练；鼓励探索使用训练集（大）；
- 数据下载地址（尽快下载以免过期）：
  - 链接
  - 也可以从QQ群文件下载

数据预处理

- 数据清洗：非法字符，稀少字词的过滤；过长句子的过滤或截断。
- 分词：将输入句子切分为 tokens，每个子串相对有着完整的语义，便于学习 embedding 表达
  - 英文: 词语之间存在天然的分隔 (空格、标点符号)，可以直接利用 NLTK 或 BPE、WordPiece 等统计方法分词
  - 中文：可以借助分词工具，诸如 Jieba(轻量型), HanLP(大体量但效果好)
- 构建词典：利用分词后的结果构建统计词典，可以过滤掉出现频次较低的词语，防止词典规模过大
- 建议用预训练词向量初始化，在训练的过程中允许更新

NMT 模型

- 自行构建基于 GRU 或者 LSTM 的 Seq2Seq 模型
- 自行实现 attention 机制
- 自行探索 attention 机制中不同对齐函数 (dot product, multiplicative, additive) 的影响

训练和推理

- 定义损失函数（例如交叉熵损失）和优化器（例如 Adam）。
- 将双语平行语料库处理成中译英数据，训练模型的中译英能力。
- 在训练过程中，对比 Teacher Forcing 和 Free Running 策略的效果。
- 对比 greedy 和 beam-search 解码策略；

编程语言与环境

- 编程语言：python
- 深度学习框架：pytorch

评估指标

- BLEU

报告要点

- 如何在本实验中搭建seq-seq框架并实现 Attention 机制
- 分享你在训练模型时学到的的技巧
- 比较使用不同基础架构或者训练推理策略时的分类效果
- Attention 可视化（少量案例进行分析） 独立完成，不得抄袭！

提交

- 源代码和训练好的 checkpoint
- 文档（PDF）（至少包含方法、实验结果分析以及心得体会）
- 压缩文件并命名："2025ANN-project2-学号-姓名.zip/rar"
- 邮件主题：2025ANN-project2-学号-姓名
- 提交邮箱：
  - 请学号为单数的同学发送到liuym87@mail2.sysu.edu.cn
  - 请学号为双数的同学发送到fengwc5@mail2.sysu.edu.cn
- Deadline: 6 月 22 日 23:59:59

参考资料

- seq2seq 机器翻译教程 (来自 pytorch 官方教程，需对数据预处理方式进行更改)：链接
- 分词工具使用: jieba 中文分词工具：链接 sentencepiece 英文分词工具： 链接
- 参考论文 Bahdanau 原版 seq2seq+attention 论文 (ICLR2015)：Neural Machine Translation by Jointly Learning to Align and Translate