Artificial Neural Networks
人工神经网络

权小军教授
中山大学计算机学院

quanxj3@mail.sysu.edu.cn

2023 年 6 月 1 日

Final-term Project: Chinese-to-English Machine Translation

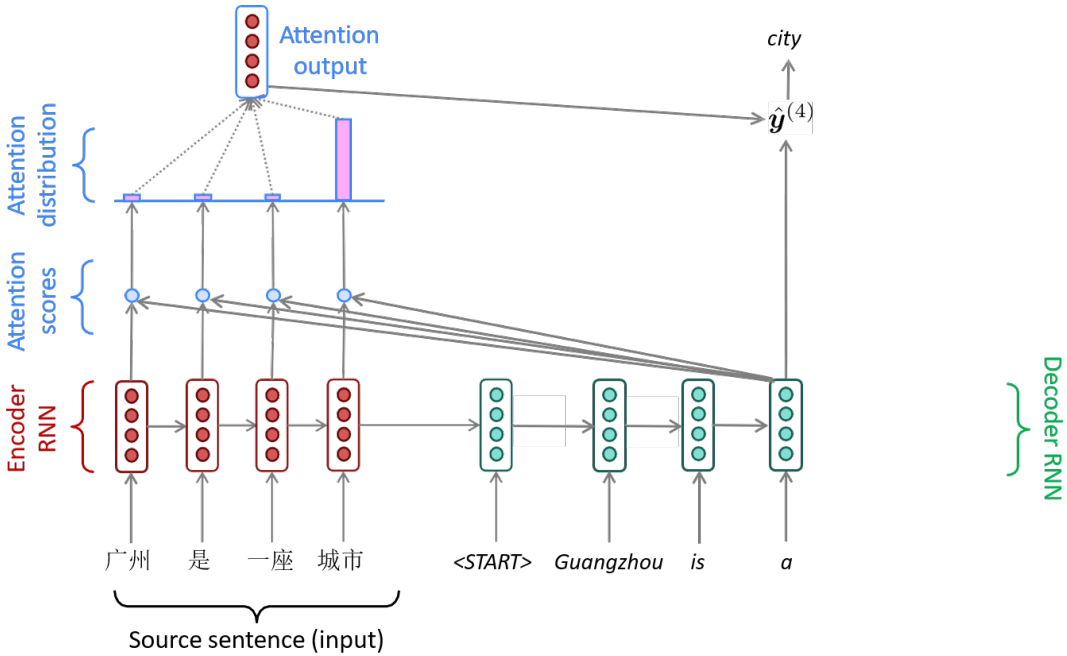
数据集简介

- 数据集说明：压缩包中共有 3 个文件夹，分别对应着训练集、评估集和测试集，它们的大小分别是 10000、1000、1000。每个文件夹中都含有 src data (中文) 和 target data (英文) 构成平行语料对。模型的性能以测试集的结果为最终标准
- 数据下载地址：课程百度网盘

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：将输入句子切分为一个个子串，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 (编码器和解码器各 3-4 层；单向或者双向)
- 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

编程语言与环境

- 编程语言：python
- 深度学习框架：pytorch

评估指标

- BLEU

$$\text{precision}_n = \frac{\sum_{C \in corpus} \sum_{\text{n-gram} \in C} \text{count-in-reference}_{\text{clip}}(\text{n-gram})}{\sum_{C \in corpus} \sum_{\text{n-gram} \in C} \text{count}\ (\text{n-gram})}$$

$$\text{BLEU-4} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right) \prod_{i=1}^{4} \text{precision}_i$$

重点考查

- 搭建 Seq2Seq 模型
- Attention 机制的实现
- 训练 Seq2Seq 模型的技巧
- 分类结果性能评估

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：“2023ANN-final-term-project-学号-姓名.zip/rar”
- 邮件主题：2023ANN-final-term-project-学号-姓名
- 提交邮箱：sysucsers@163.com
- Deadline: 2023-06-18 24pm

参考资料

- pytorch 框架：
  https://pytorch.org/tutorials/
- 模型搭建及训练：
  https://github.com/pcyin/pytorch_basic_nmt
- 分词工具使用:
  https://zhuanlan.zhihu.com/p/146792308

Thank you!