作业3 机器翻译项目说明文档

本次作业选用RNN + Attention模型，平台使用pytorch

一、项目结构说明

├─dataset 数据集

│ │ input.cn.txt

│ │ input.en.txt

│ │ train.tsv

│ │ prepare.py 构造训练样例，生成train.tsv文件

│ │ sgns.renmin.bigram-char 从网上下载的预训练好的中文词向量

│ │ glove.6B.300d.txt 从网上下载的预训练好的英文词向量

│

├─model

│ │ NMT\_Atten.py

│ EncoderRNN encoder模块

│ AttentionDecoderRNN 带attention的decoder模块

│

├─results

│ │ rnn\_atten\_nmt\_Model.pt

│

├─utils

│ │ data\_preprocess.py 数据预处理

│ 导入中文词向量sgns.renmin.bigram-char和英文词向量

│ glove.6B.300d.txt，根据train.tsv构建词典，将原始

│ train.tsv分成训练集（90%）和验证集（10%）

│

│ │ config.py 设置具体的参数

│ │ evaluate.py 用BLEU算法评价机器翻译效果

│

├─main.py 模型训练，模型验证，模型推理（预测）

│

├─README.md

│

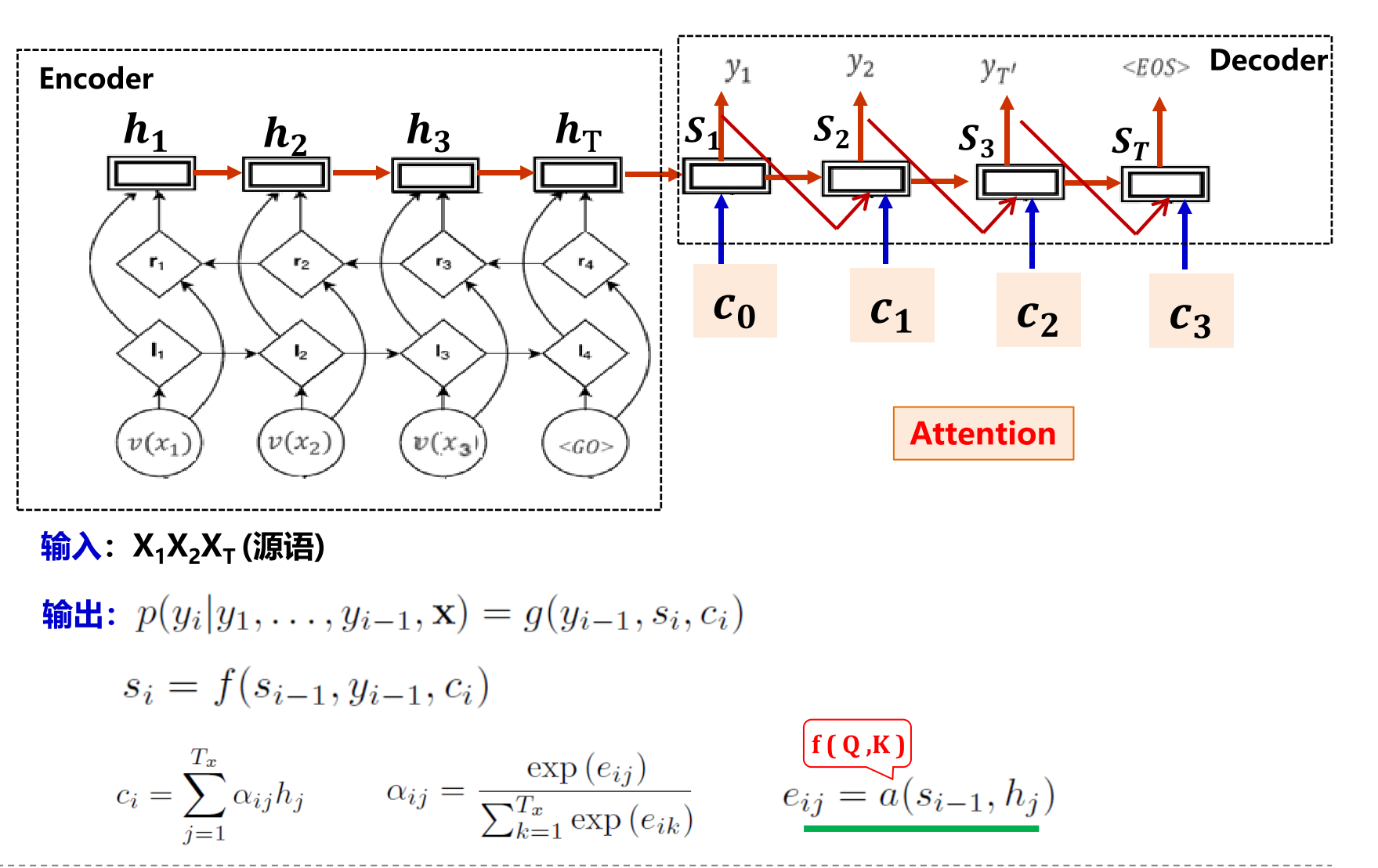
├─requirments.txt

│

├─doc

│ │ NMT说明文档

二、模型结构



三、模型参数设置

**1、Train&Dev**

parameter settings：

lr-- 2e-3

train\_data\_path-- dataset/train.tsv

src\_embedding\_path-- dataset/sgns.renmin.bigram-char

tgt\_embedding\_path-- dataset/glove.6B.300d.txt

train\_batch\_size-- 128

src\_embedding\_dim-- 300

tgt\_embedding\_dim-- 300

注：

sgns.renmin.bigram-char是源语言预训练词向量文件， glove.6B.300d.txt目标语言预训练词向量文件，其维度应与各自的 embedding\_dim 保持一致。在本项目没有训练词向量，使用了已经预训练好的词向量。

**2、Test**

parameter settings：

train\_data\_path-- dataset/train.tsv

load\_model-- results/rnn\_atten\_nmt\_Model.pt

注： rnn\_atten\_nmt\_Model.pt 为在验证集（训练集的 10%用于验证调参）上的最优模型

二、模型训练/测试过程

在Spyder运行即可

1、模型训练过程文件

train.html

2、模型测试过程文件

test.html

附：代码参考来源：

* <https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html?highlight=machine%20translation>
* [GitHub - AIRobotZhang/RNN\_Attention\_NMT: RNN+Attention for Neural Machine Translation](https://github.com/AIRobotZhang/RNN_Attention_NMT)
* <https://www.cnblogs.com/jiangxinyang/p/10523585.html>
* [GitHub - Embedding/Chinese-Word-Vectors: 100+ Chinese Word Vectors 上百种预训练中文词向量](https://github.com/Embedding/Chinese-Word-Vectors)
* <https://nlp.stanford.edu/projects/glove/>
* [GitHub - Wasim37/nmt: 神经机器翻译-seq2seq](https://github.com/Wasim37/nmt)
* [GitHub - AIRobotZhang/RNN\_Attention\_NMT: RNN+Attention for Neural Machine Translation](https://github.com/AIRobotZhang/RNN_Attention_NMT)