学　　期 2021-2022（2）

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理

第五次大作业

Seq2Seq 小说文本生成

院（系）名称	自动化科学与电气工程学院
专业名称	电子信息
学生姓名	孙茗逸
学号	ZY2103113
指导老师	秦曾昌

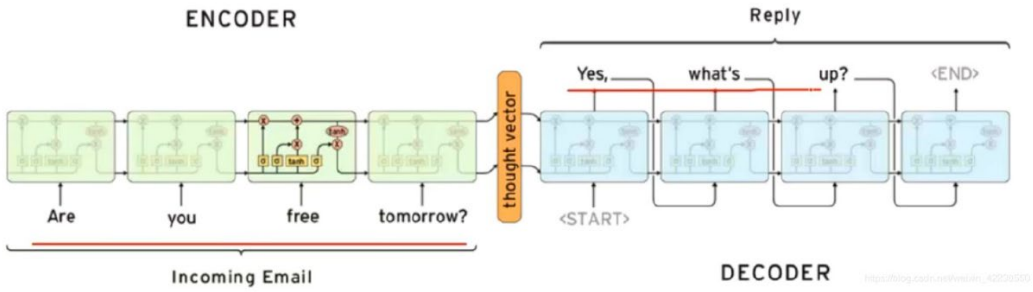
2022 年 6 月

一、任务要求

基于 Seq2seq 模型来实现文本生成的模型，输入可以为一段已知的金庸小说段落，来生成新的段落并做分析。

二、实验原理

Seq2seq 即序列到序列模型，可用于机器翻译、文本摘要、会话建模、图像字幕等任务。其主要结构如下图所示。



模型大致可以分为四个部分，输入、编码（encode）部分、解码（decode）、输出。输入部分主要为文本嵌入，为了将文本中词汇的数字表示转变为向量表示，希望在这样的高维空间捕捉词汇间的关系。

中间部分属于 encoder-decoder 结构，基本思想就是利用两个循环神经网络（RNN、LSTM 等等），一个网络作为 encoder，另一个网络作为 decoder。encoder 负责将输入序列压缩成指定长度的向量，这个向量就可以看成是这个序列的语义，这个过程称为编码，用于对输入进行指定的特征提取过程。

Encoder 通过学习输入，将其编码成一个固定大小的状态向量 S，继而将 S 传给 Decoder，Decoder 再通过对状态向量 S 的学习来进行输出。

输出部分通过对上一步的线性变化得到指定维度的输出，也就是转换维度的作用。Softmax 函数使最后一维的向量中的数字缩放到 0-1 的概率值域内，并满足他们的和为 1，得到对应的概率。

三、实验过程

3.1 文本预处理

过程与前几次实验大体相同，包括文本的读取，去除特殊标点符号，去除停词，分词等操作。为了让分词更准确，在网站上下载了人名、门派、武功的专有词汇，用于分词过程中。

3.2 模型定义

模型的定义与训练包括Word2Vec模型以及Seq2Seq模型。
在对seq2seq模型进行训练前，采用基于CBOW方法的Word2Vec模型，通过对金庸小说文本进行训练，生成文本信息的编码，用词向量来表示文本信息。
Seq2Seq模型编码器和解码器均采用LSTM，在模型的输入和输出前增加线性映射层。

3.3 模型的训练和预测

简单起见，模型的训练loss采用计算余弦相似度的方法，即通过衡量预测词向量与目标词向量之间的余弦相似度，若相似度较大，则损失较小，反之亦然。
在模型的预测过程中，通过设定预测结束的条件，即对输出的总词数以及输出句子的数量进行限制，得到最后的输出。
采用《天龙八部》的全部内容作为训练数据，对模型进行训练，共训练100epoch，采用SGD优化器，学习率为0.01。测试过程中挑选书中的某半句话作为测试输入。

3.4 模型效果

采用《天龙八部》对模型进行训练，并摘取其中某一句话作为引导词，观察模型的输出。

提示字符：虚竹恍然
生成文本：虚竹心下恍然，铁丑怕羞。朱四哥缝套粗心，腐骨丸无法无天，无意之中痛快小贼，毒得朱四哥饮水。
原文对照：生成文本：虚竹心下恍然，知道童姥为了恼他宁死不肯食荤，却去掳了一个少女来，诱得他破了淫戒，不由得又是悔恨，又是羞耻，突然间纵起身来，脑袋疾往坚冰上撞去，砰的一声大响，掉在地下。

分析：总体来看，模型的输出语句与金庸风格比较相近，学会了基本的形容词-名词，动词-副词等语法，并且学会了书中的一些特有词汇，比如腐骨丸、铁丑等词的词性和用法。但是，内容上缺乏实际含义，前后语言不搭，说明模型还没有理解语言背后的深层含义。

四、总结体会

基于 Seq2seq 模型实现了文本生成的模型，可以通过输入一段已知的小说段落来生成新的段落。由于时间和算力的限制，没能对模型进行很好的改进和调

节，但是对 LSTM 以及 Seq2Seq 模型的训练和测试过程有了一定的体会，也对文本生成的任务有了一定的了解。