

# Report of Deep Learning for Natural Language Processing

冯佳铂  
775193638@qq.com

## Abstract

利用给定语料库（金庸语小说语料链接），用 Seq2Seq 与 Transformer 两种不同的模型来实现文本生成的任务（给定开头后生成武侠小说的片段或者章节），并对比与讨论两种方法的优缺点

## Introduction

文本生成旨在根据输入的信息生成连贯、准确且自然的文本。

Seq2Seq 模型输入是一个序列，输出也是一个序列，是一种序列到序列的编码器-解码器结构，由一个编码器和一个解码器组成。通过深度神经网络模型（RNN 或 LSTM），编码器将输入序列(如源语言文本)编码为固定长度的向量，解码器则将这个向量解码为目标序列。大致可理解为 encoder 编码器负责对输入句子的理解，decoder 解码器负责对理解后的句子进行处理，组装回答。

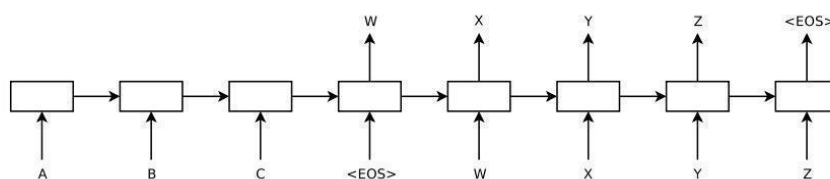
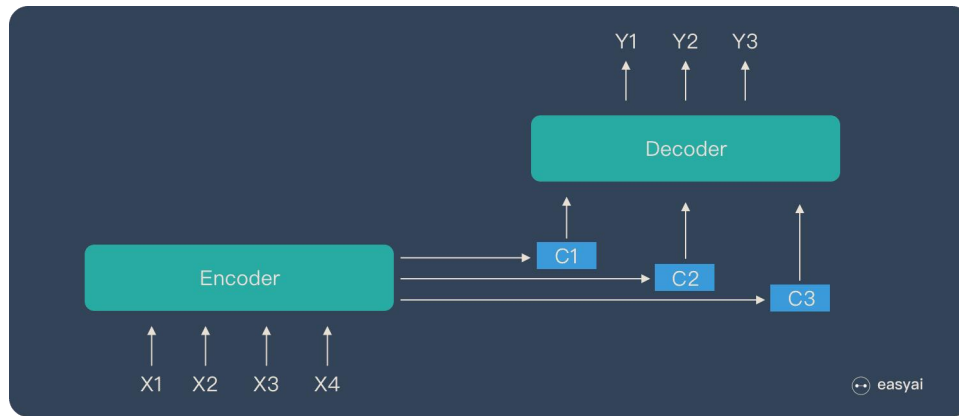


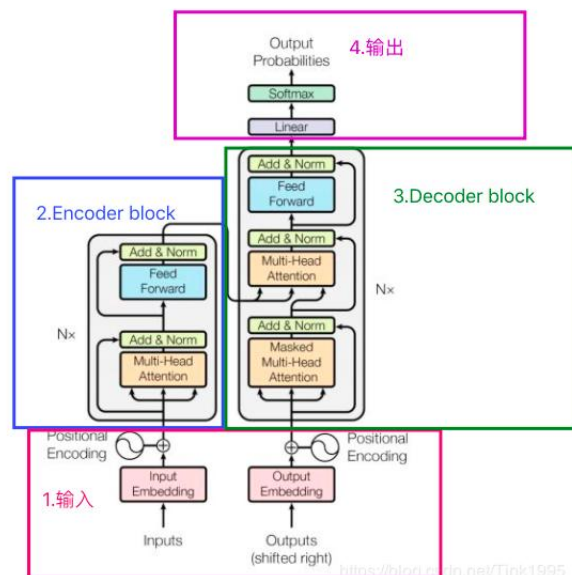
Figure 1: Our model reads an input sentence “ABC” and produces “WXYZ” as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

[http://blog.csdn.net/Sorhhxg\\_30](http://blog.csdn.net/Sorhhxg_30)

在此基础上，引入 attention 注意力机制，可理解为对于输入信息，根据其重要程度进行不同权重的加权处理。解决当信息过长时信息丢失的问题，此时 encoder 不再将整个序列编码为固定长度的中间向量，而是编码成一个向量序列，根据信息的重要程度有的放矢的进行特征提取。



Transformer 是一种特殊的 seq2seq 模型，完全基于自注意力机制，没有递归结构。输入序列中每个位置单词都以各自单独的路径流入编码器，使得编码器在对每个单词编码时会关注输入句子中的其他单词，位置编码保证可正确联系上下文，将单词放置到整个语境之中理解。



## Methodology






流程如下：

1. 语料导入：使用金庸先生的《鹿鼎记》为语料。
2. 文本预处理：将语料中多余符号进行删除，如“本书来自 [www.cr173.com](http://www.cr173.com) 免费 txt 小说下载站”、“更多更新免费电子书请关注 [www.cr173.com](http://www.cr173.com)”等，使用 jieba 库进行分词。使用 word2vec 计算词向量。
3. 训练模型：使用 encoder 与 decoder 均使用 LSTM 建立相应 seq2seq 模型。

4. 任意给出一段文字开头，使用模型自动生成后续文字段落。

## Experimental Studies

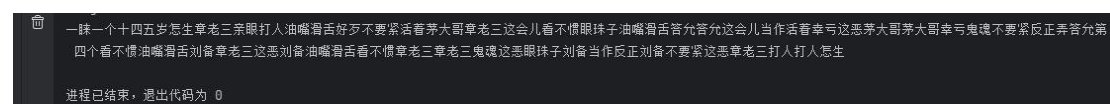
将训练所得模型。

	model.model	2024/6/16 21:38	MODEL 文件	1,409 KB
	model.model.syn1neg.npy	2024/6/16 21:38	NPY 文件	192,697 KB
	model.model.wv.vectors.npy	2024/6/16 21:38	NPY 文件	192,697 KB
	model.pth	2024/6/16 21:57	PTH 文件	6,169 KB
	seq2seq	2024/6/16 22:01	JetBrains PyChar...	7 KB

所给开头:

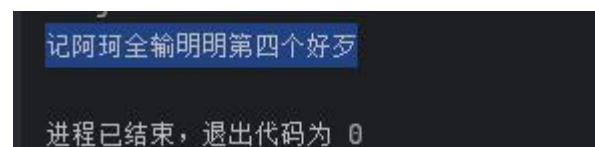
①“韦小宝本盼他一刀砍在自己胸口，自己有宝衣护身，不会丧命，便可将四名喇嘛吓得逃走，哪知他不砍胸而砍背，将自己推入他怀中，正好乘机用匕首戳他几剑，只是在对方胯下爬出，未名太过狼狈，临危逃命，也顾不得英雄还是狗熊了。”

运行结果：一瞅一个十四五岁怎生章老三亲眼打人油嘴滑舌好歹不要紧活着茅大哥章老三这会儿看不惯眼珠子油嘴滑舌答允答允这会儿当作活着幸亏这恶茅大哥茅大哥幸亏鬼魂不要紧反正弄答允第四个看不惯油嘴滑舌刘备章老三这恶刘备油嘴滑舌看不惯章老三章老三鬼魂这恶眼珠子刘备当作反正刘备不要紧这恶章老三打人打人怎生



②“行出十余里，身后马蹄声响，数十乘马追了上来，阿珂脸上登现喜色。”

运行结果：记阿珂全输明明第四个好歹



## Conclusions

上面 2 个输出都很差，不能构成通顺的语义。不过倒也稍微存在些许联系，①中生成的文字中反复出现油嘴滑舌，与①中所描述场景稍微吻合。

seq2seq 模型中 encoder 所有信息压缩至一个固定长度的向量中，以此向量为隐藏输入，预测 decoder 端输出，会损失较多信息，同时也会由于信息过于庞杂，导致不能得到理想的

输出。

transformer 模型通过引入多头注意力机制、前馈神经网络和残差连接，对 seq2seq 模型的缺点有实质性改进，使用 transformer 可以生成更加通顺的段落文字。

## References

- [1] Transformer 通俗笔记：从 Word2Vec、Seq2Seq 逐步理解到 GPT、BERT
- [2] 【NLP】一文理解 Seq2Seq-CSDN 博客
- [3] 史上最小白之 Transformer 详解-CSDN 博客
- [4] PyTorch 从零开始实现 Transformer\_transformer 代码 pytorch-CSDN 博客
- [5] 深度学习与自然语言处理——段落分析模型-CSDN 博客