NLP Homework 2 Report

王宇轩 2022E8001082025

1 问题描述

利用前馈神经网络、循环神经网络和自注意力机制网络对比语言模型的困惑度。

2 方法简述

2.1 前馈神经网络

前馈神经网络的结构与实现参考了[1]，通过一个embedding层生成词向量，将n步内的词向量拼接起来，通过一个线性层得到隐层输出，在得到最后输出时，对词向量再次进行映射，加到输出中，整体的计算过程如下：

$$
\begin{aligned}
Embed &= \mathbf{Concat}(Embedding(input)) \\
hidden &= tanh(\mathbf{W_0} + b_0) \\
output &= \mathbf{W_1} hidden + \mathbf{W_2} Embed + b_1
\end{aligned}
$$

在具体的参数设置方面，令词向量维度为30，隐层维度为50.

2.2 循环神经网络

一般来说，循环神经网络可以直接接受关于词表的one-hot编码作为输入，但是在实现过程中，由于词表较大，同时one-hot编码作为一个稀疏矩阵，比较浪费空间，所以先对输入求词向量进行降维，再使用RNN进行计算，取最后一个节点的输出作为输出，计算过程如下：

$$
\begin{aligned}
Embed &= Embedding(input) \\
outputs &= RNN(Embed, h_0) \\
output &= \mathbf{W} outputs[-1] + b
\end{aligned}
$$

其中，h_0为全0初始化的隐状态，用于在RNN结构中记录历史信息。词向量的维度为30，隐层维度为50.

2.3 自注意力机制

自注意力机制的模型实现，参考了Transformer decoder[2]的结构，但进行了简化。使用了两层attention，保留了残差连接，取最后一个输入token对应的输出经过映射得到输出。计算过程如下：

$$
\begin{aligned}
E &= Embedding(input) \\
hidden_1 &= Attn(\mathbf{Q_0}E, \mathbf{K_0}E, \mathbf{V_0}E) \\
E_1 &= layer\_norm(hidden_1 + E) \\
hidden_2 &= Attn(\mathbf{Q_1}E_1, \mathbf{K_1}E_1, \mathbf{V_1}E_1) \\
E_2 &= layer\_norm(hidden_2 + E_1) \\
output &= \mathbf{W}E_2[-1] + b
\end{aligned}
$$

其中词向量的维度设置为30.

3 实验

3.1 数据预处理

本次实验数据为统一要求，是2018年的中文新闻语料。不同于英文语料，要对中文语料构建语言模型，需要先进行分词。本次实验使用开源中文分词器jieba进行分词，并根据分词结构构建词表。为了控制词表的大小，去除了词频小于5的词，并将所有未出现于词表中的词默认映射到特殊词$< unk >$上。取语料的前80%的词作为训练集，剩下的20%作为测试集。设置模型考虑的上文范围为5，即，每个样本取5个词，前4个词作为输入数据，最后1个词作为标签。

3.2 实验结果

所有的模型都固定学习率为0.01，使用Adam优化器。我们计算了实验结果的困惑度作为评价指标。首先，每种模型都训练5轮的结果在表1中呈现。

Model	PPL
NNLM	1017.07
RNN	613.72
Attention	**530.49**

表 1: 训练5轮的结果

由于Attention具有更强的特征提取能力，自注意力机制模型取得了最好的效果，这也与预期符合。在这个结果的基础上，保持参数不变，再训练5轮，结果在表2中展示。

Model	PPL
NNLM	1423.94
RNN	718.05
Attention	**599.65**

表 2: 训练10轮的结果

可以看到，继续训练下去，每个模型都出现了比较明显的过拟合现象，这可能源于数据量有限。但是不同模型的语言模型同样表现出了与5轮的结果一致的性能，即自注意力机制的性能最强，RNN其次，NNLM性能最差。

同时，我们统计了不同模型每轮训练的平均时间，在表3中展示。

Model	Avg Time(s)
NNLM	2181
RNN	1137
Attention	1059

表 3: 训练1轮的平均时间

训练Attention网络的效率最高，其次是RNN，最后是NNLM，这可能是由于NNLM中，将词向量直接拼接再通过线性层，这一操作大大增加了矩阵计算的维度，降低了计算速度。

参考文献

[1] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," vol. 3, 01 2000, pp. 932–938.

[2] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.