Report of Deep Learning for Natural Langauge Processing

Yiming Xu
1953999172@qq.com

Abstract

本文对金庸的 16 篇小说进行了语料分析，利用基于 Word2Vec 的神经语言模型来训练词向量,通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联、或者其他方法来验证词向量的有效性。

Introduction

Word2Vec 是一种通过训练神经网络学习词向量的模型，其核心是神经语言模型。Word2Vec 提供了两种训练模型：Skip-gram 和 Continuous Bag of Words (CBOW)。这两种模型都基于神经网络，旨在通过训练学习单词的向量表示，这些向量表示可以捕捉单词之间的语义和句法关系。

模型架构：

Word2Vec 的神经语言模型通常采用三层神经网络结构：输入层、隐藏层和输出层。

在 Skip-gram 模型中，给定一个目标词，模型预测其上下文词。而在 CBOW 模型中，则是通过上下文词来预测目标词。

训练过程：

训练数据通常是一系列的文本句子，模型通过滑动窗口在这些句子上移动，以生成训练样本。

对于每个训练样本，模型会调整其内部参数（主要是词向量），以最小化预测误差。

训练过程中，模型逐渐学习到单词之间的关联，并将这些信息编码到词向量中。

词向量：

在训练过程中，每个单词都会被分配一个向量表示。这个向量在神经网络的隐藏层中形成，并随着训练的进行而更新。

训练完成后，这些词向量可以用于各种自然语言处理任务，如文本分类、情感分析、机器翻译等。

Skip-gram 与 CBOW：

Skip-gram 模型通过给定目标词来预测其上下文词。这种模型在处理低频词时表现较好，因为它会尝试为每个目标词生成多个上下文词的预测。

CBOW 模型则是通过上下文词来预测目标词。这种模型在处理高频词时可能更有优势，因为它利用多个上下文词来预测一个目标词。

优化技巧：

Word2Vec 的训练过程涉及大量的计算，为了提高效率，通常会采用一些优化技巧，如层次 softmax、负采样等。

这些技巧有助于减少计算量，加速训练过程，同时保持良好的词向量质量。

总的来说，Word2Vec 的神经语言模型通过学习单词之间的关联，为自然语言处理任务提供了强大的词向量表示。这些词向量捕捉了单词之间的语义和句法关系，为各种 NLP 应用提供了有价值的信息。

Experimental Studies

实验步骤：

1、准备语料库:本次实验以金庸的 16 部武侠小说作为中文语料库。

2、预处理：对语料库进行预处理，删除标点符号，无意义的广告等。并使用 jieba 库对文本进行分词。

3、模型训练：通过 gensim 库中的 Word2Vec 模型对经过预处理的中文语料库进行训练,并通过 model. save 函数保存整个模型。

4、语意相似度计算。

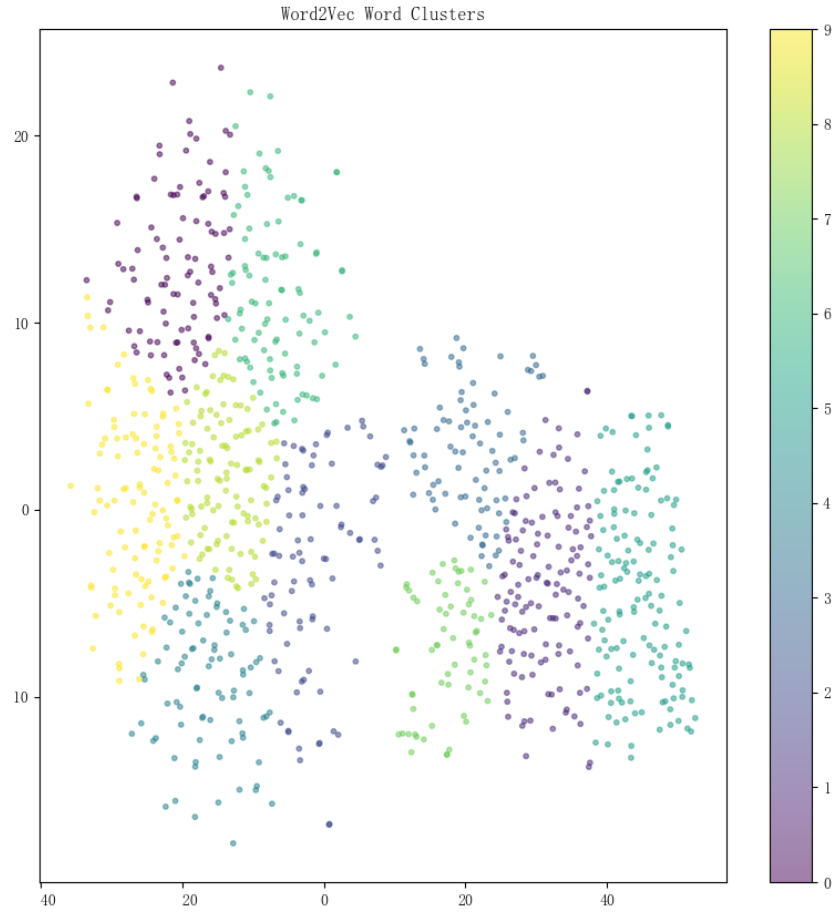
5、词类聚类并可视化。

6、词语类比。

Experimental Studies

词类聚类：

本实验采用 K-Means,n_clusters=10 进行聚类：



不同颜色代表不同的簇，轮廓系数为 0.3475.

词语类比：

通过两个词列表,其中一个正向加权,一个负向加权,指定两个正向词和一个负向词,从而找到最相似的词。本次实验所采用的词语对：(positive:女人,皇帝， negative:男人)，(positive:武林,江湖， negative:侠客)，(positive:马蹄,青石板， negative:黑衣)

词	相似词	相似度
(positive:女人,皇帝, negative:男人)	奸臣	0.7689
(positive:武林,江湖,	遭遇	0.8049

negative:侠客)		
(positive:马蹄,青石板, negative:黑衣)	隐隐	0.8246