

Skip-gram 词向量实验报告

王松宸 2024201594

2025 年 11 月 12 日

1 引言

词向量是将离散的词映射到连续向量空间的一种表示学习方法，相比于 one-hot 表示，词向量在维度上大幅压缩，并通过训练捕获词间的语义与句法关系，使得向量空间中距离与方向具有一定的语义意义，例如类比关系 $\text{king} - \text{man} + \text{woman} \approx \text{queen}$ 。

2 对词向量的理解

2.1 概述

把每个词放到同一个连续空间里，像在一张看不见的“语义地图”上给它一个坐标。意思越相近、在相似语境里常一起出现的词，坐标就越靠近；用法差很多的词，距离就更远。最终得到的是一张“词到点”的映射表，人们把这些点的坐标称为词向量。

2.2 词向量与 One-hot 的区别

One-hot 只会告诉你“词是不是同一个”，不同词之间全都等距，无法表达相似性；而词向量会让“相似的词更近、不相似的更远”，从而能做相似度、聚类、检索等语义相关的事情。另一方面，词向量把信息分布到少量维度上，参数可以共享，数据更容易泛化到没见过的句子里。

2.3 词向量能做什么

- 最近邻：找与某词最接近的词，常能得到同义或近义词；
- 类比：在向量空间里，类似于“国王：王后”与“男人：女人”的对应方向关系，经常能被保留下来；
- 作为下游任务特征：在情感分析、文本分类、序列标注等任务里，词向量可以作为良好的初始化或固定特征。

3 Skip-gram 模型

3.1 概述

Skip-gram 的目标是：给定一个“中心词”，让模型去预测它周围会一起出现的词。凡是经常出现在同一语境里的词，其向量会被推近；用法差异大的词会被推远。模型内部维护两套嵌入表：输入嵌入代表中心词，输出嵌入代表被预测的上下文词；训练结束后，通常把输入嵌入当作最终的词向量。

3.2 训练数据如何构造

遍历语料，按窗口大小在句子上滑动。对于每个中心词，都与窗口内的上下文词组成若干训练样本（中心词，外部词）。窗口越大，产生的样本越多，语义范围也更广。

3.3 对学习过程的理解

对每个样本，模型学习“让正确的上下文词得高分，让无关词得低分”。优化用随机梯度下降完成。

4 词向量可视化与图像解读

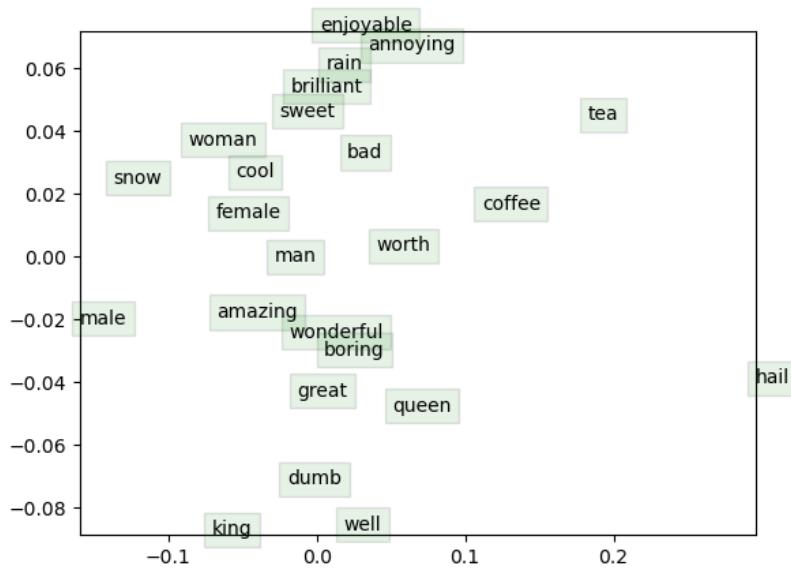


图 1: word_vectors

4.1 降维方法

使用 PCA 将 d -维嵌入映射到二维。PCA 保留最大方差方向，但轴本身没有固定语义，仅相对位置与方向可供解释。

4.2 结构观察

图中 king 与 queen 相对接近且与 man/woman 形成类比方向；积极情感词 (amazing, wonderful, great) 与消极词 (bad, annoying, dumb) 有一定分离。`male/female` 与 man/woman 邻近显示性别语义子空间。

4.3 潜在问题

观察到一些积极词与消极词混杂，可能因训练语料规模有限，导致某些词的上下文信息不足。PCA 投影可能掩盖部分高维结构，可是尝试结合其他降维方法（如 t-SNE）进行补充分析。

5 梯度结果分析

```
iter 29910: 9.829300
iter 29920: 9.785530
iter 29930: 9.722733
iter 29940: 9.677070
iter 29950: 9.709149
iter 29960: 9.724619
iter 29970: 9.759012
iter 29980: 9.740169
iter 29990: 9.731548
iter 30000: 9.798442
sanity check: cost at convergence should be around or below 10
training took 6078 seconds
```

图 2: 训练迭代后期的损失值

Loss 数值从最初的 20 左右下降到最终的 9.7 左右，符合预期的标准，显示出明显的收敛趋势。从图 2 可见损失在后期围绕一个缓慢下降的趋势波动，说明学习率退火后更新步长较小，进入稳定区域。

6 结论

本文以直观视角理解词向量与 Skip-gram，词向量可视化结果中特征较为明显，训练损失稳定收敛。