词向量

One－hot 向量：

记词典里有 |V| 个词，每个词都被表示成一个 |V| 维的向量，设这个词在字典中相应的顺序为 i，则向量中 i 的位置上为 1，其余位置为 0.

词－文档矩阵：

构建一个矩阵 X，每个元素 X_{ij} 代表 单词 i 在文档 j 中出现的次数。

词－词共现矩阵：

构建矩阵 X，每个元素 X_{ij} 代表 单词 i 和单词 j 在同一个窗口中出现的次数。

word2vec

word2vec 是一套能将词向量化的工具，它将文本内容处理成为指定维度大小的实数型向量表示，并且其空间上的相似度可以用来表示文本语义的相似度。

Word2vec 的原理主要涉及到统计语言模型（包括 N-gram 模型和神经网络语言模型），continuousbag-of-words 模型以及 continuous skip-gram 模型。

N-gram 的意思就是每个词出现只看其前面的 n 个词，可以对每个词出现的概率进行近似。

比如当 n=2 的时候：

$$P(I,saw,the,red,house) \approx P(I|<s>,<s>)P(saw|<s>,I)$$
$$\times P(the|I,saw)P(red|saw,the)P(house|the,red)$$
$$\times P(</s>|red,house)$$

神经网络语言模型（NNLM）用特征向量来表征每个词各个方面的特征。NNLM 的基础是一个联合概率：

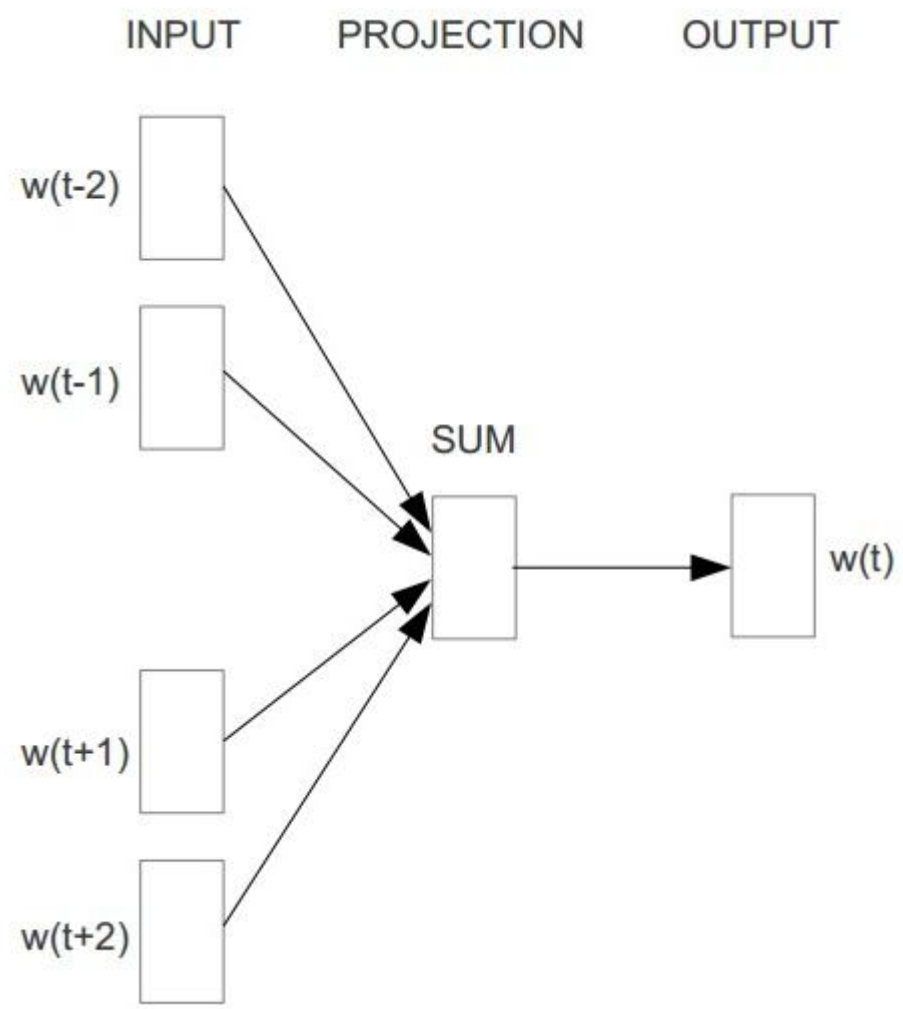
$$P(Z_1=z_1,\ldots,Z_n=z_n)=\prod_i P(Z_i=z_i \mid g_i(Z_{i-1}=z_{i-1},Z_{i-2}=z_{i-2},\ldots,Z_1=z_1))$$

其神经网络的目的是要学习：

$$f(i,w_{t-1},\ldots,w_{t-n+1})=g(i,C(w_{t-1}),\ldots,C(w_{t-n+1}))$$

$$P(w_t=i \mid w_1^{t-1})$$

Continuous Bag-of-Words(CBOW) 模型与 NNLM 类似，结构如下：



CBOW 是通过上下文来预测中间的词，如果窗口大小为 k，则模型预测：

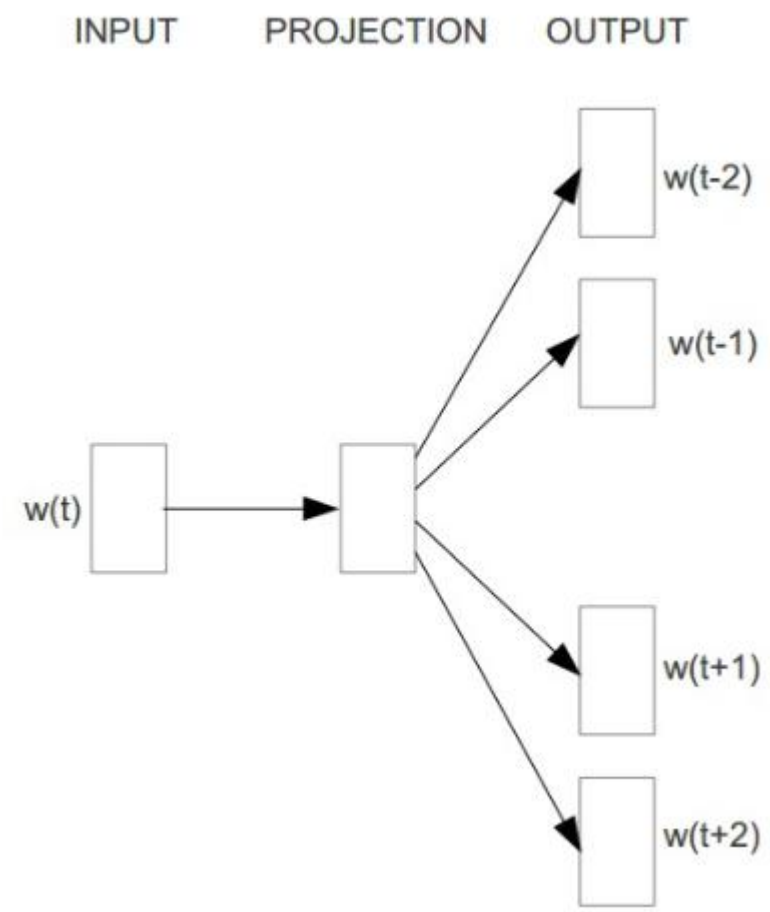
$$P(w_t \mid w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots w_{t+k})$$

其神经网络就是用正负样本不断训练，求解输出值与真实值误差，然后用梯度下降的方法求解各边权重参数值的。

Continuous skip-gram 模型与 CBOW 正好相反，是通过中间词来预测前后词，一般可以认为位置距离接近的词之间的联系要比位置距离较远的词的联系紧密。目标为最大化：

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log P(w_{t+j}\mid w_t)$$

结构为：



Doc2Vec

Doc2Vec 或者叫做 paragraph2vec, sentence embeddings, 是一种非监督式算法, 可以获得 sentences/paragraphs/documents 的向量表达, 是 word2vec 的拓展。

学出来的向量可以通过计算距离来找 sentences/paragraphs/documents 之间的相似性, 或者进一步可以给文档打标签。

Doc2Vec 实现的两种方法:

dbow (distributed bag of words)

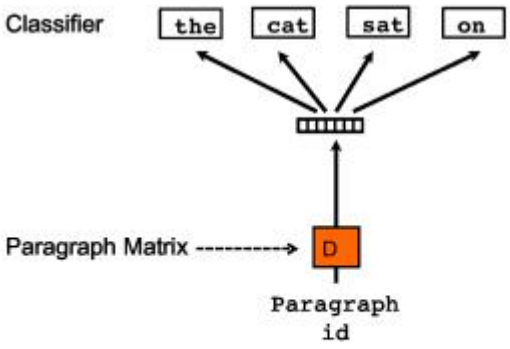


Figure 3. Distributed Bag of Words version of paragraph vectors. In this version, the paragraph vector is trained to predict the words in a small window.

dm (distributed memory)

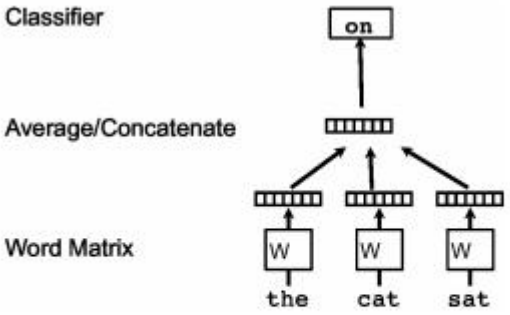


Figure 1. A framework for learning word vectors. Context of three words ("the," "cat," and "sat") is used to predict the fourth word ("on"). The input words are mapped to columns of the matrix W to predict the output word.

二者在 gensim 实现时的区别是 dm = 0 还是 1。

Doc2Vec 的目的是获得文档的一个固定长度的向量表达。

数据：多个文档，以及它们的标签，可以用标题作为标签。

影响模型准确率的因素：语料的大小，文档的数量，越多越高；文档的相似性，越相似越好。