深入理解word2vec

小小挖掘机 今天

以下文章来源于机器学习实验室 ， 作者louwill

机器学习实验室

统计学出身的深度学习算法工程师。进击的Coder。

深度学习

Author：louwill

From：深度学习笔记

语言模型是自然语言处理的核心概念之一。word2vec是一种基于神经网络的语言模型，也是一种词汇表征方法。word2vec包括两种结构：**skip-gram（跳字模型）和CBOW（连续词袋模型）**，但本质上都是一种词汇降维的操作。

word2vec

我们将NLP的语言模型看作是一个监督学习问题：即给定上下文词，输出中间词，或者给定中间词，输出上下文词。基于输入和输出之间的映射便是语言模型。这样的一个语言模型的目的便是检查和放在一起是否符合自然语言法则，更通俗一点说就是和搁一起是不是人话。

所以，基于监督学习的思想，本文的主角——word2vec便是一种基于神经网络训练的自然语言模型。word2vec是谷歌于2013年提出的一种NLP分析工具，其特点就是将词汇进行向量化，这样我们就可以定量的分析和挖掘词汇之间的联系。因而word2vec也是我们上一讲讲到的词嵌入表征的一种，只不过这种向量化表征需要经过神经网络训练得到。

word2vec训练神经网络得到一个关于输入和输出之间的语言模型，我们的关注重点并不是说要把这个模型训练的有多好，而是要获取训练好的神经网络权重，这个权重就是我们要拿来对输入词汇的向量化表示。一旦我们拿到了训练语料所有词汇的词向量，接下来开展 NLP 研究工作就相对容易一些了。

word2vec包括两种模型。一种是给定上下文词，需要我们来预测中间目标词，这种模型叫做连续词袋模型（Continuous Bag-of-Words Model，以下简称CBOW），另一种是给定一

个词语，我们根据这个词来预测它的上下文，这种模型叫做skip-gram模型，也有种翻译称之为“跳字”模型。

CBOW模型的应用场景是要根据上下文预测中间词，所以我们的输入便是上下文词，当然原始的单词是无法作为输入的，这里的输入仍然是每个词汇的one-hot向量，输出为给定词汇表中每个词作为目标词的概率。CBOW模型的结构如图1所示。

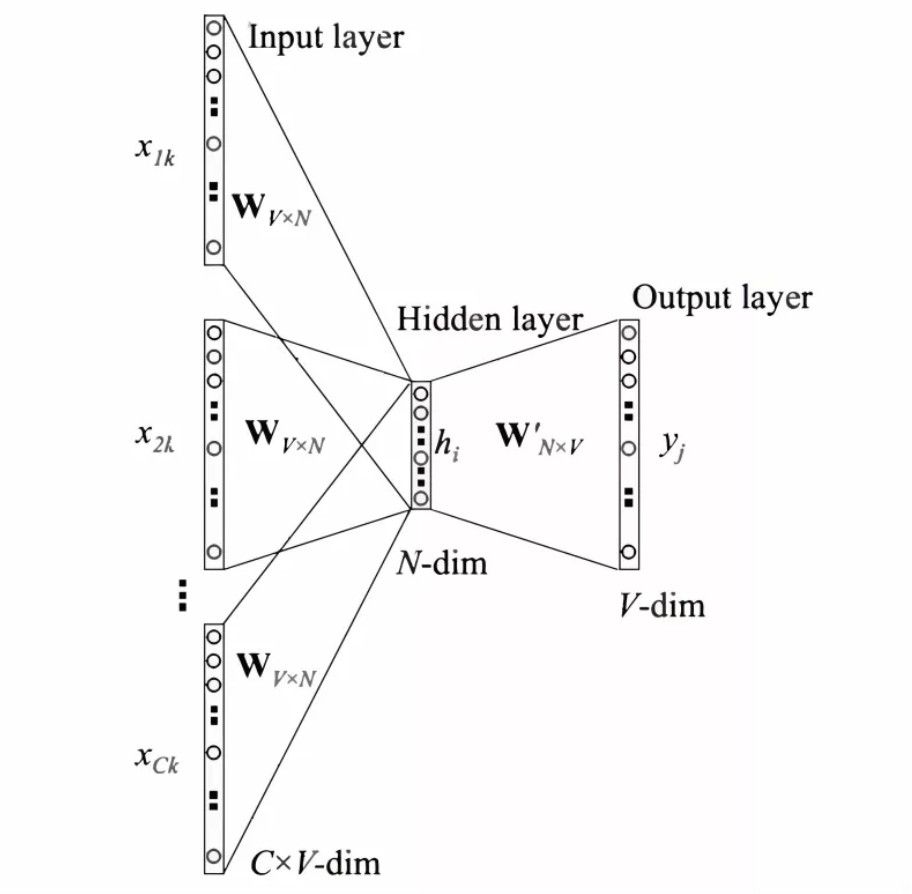

图1 CBOW模型

Skip-gram模型的应用场景是要根据中间词预测上下文词，所以我们的输入是任意单词，输出为给定词汇表中每个词作为上下文词的概率。Skip-gram模型的结构如图2所示。

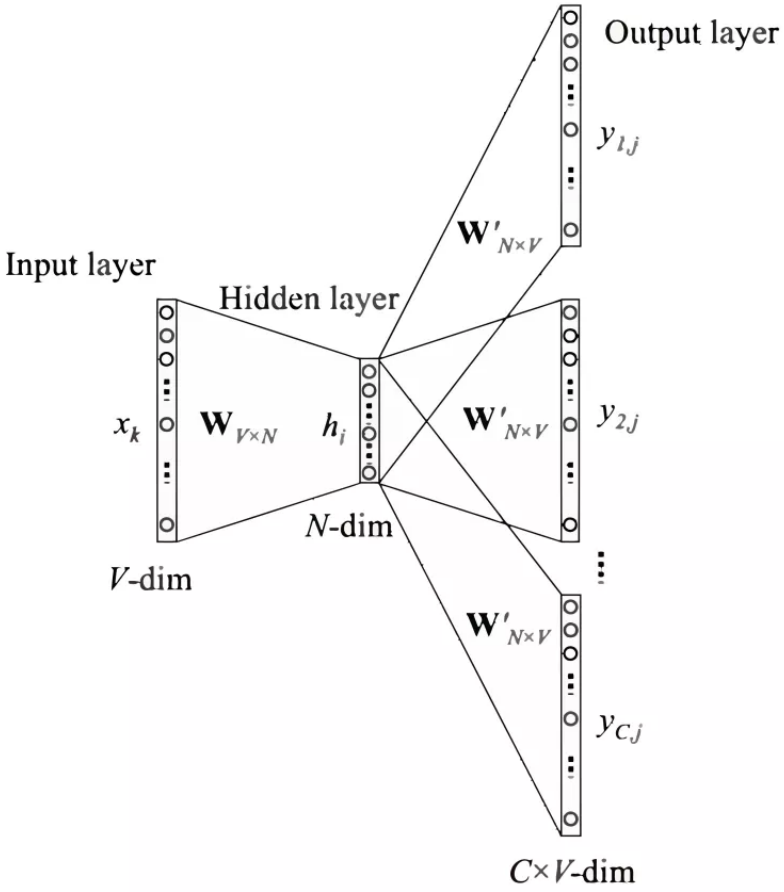

图2 skip-gram模型

从CBOW和skip-gram模型的结构图可以看到，二者除了在输入输出上有所不同外，基本上没有太大区别。将CBOW的输入层换成输出层基本上就变成了 skip-gram 模型，二者可以理解为一种互为翻转的关系。

从监督学习的角度来说，word2vec本质上是一个基于神经网络的多分类问题，当输出词语非常多时，我们则需要一些像 Hierarchical Softmax （分层 Softmax） 和 Negative Sampling（负采样）之类的技巧来加速训练。但从自然语言处理的角度来说，word2vec关注的并不是神经网络模型本身，而是训练之后得到的词汇的向量化表征。这种表征使得最后的词向量维度要远远小于词汇表大小，所以word2vec从本质上来讲是一种降维操作。我们把数以万计的词汇从高维空间中降维到低维空间中，对NLP的下游任务大有裨益。

word2vec的训练过程：以CBOW为例

因为skip-gram和CBOW的相似性，本小节仅以CBOW模型为例说明word2vec是如何训练得到词向量的。图3标出了CBOW模型要训练的参数，很明显我们要训练得到输入层到隐藏层的权重以及隐藏层到输出层的权重。

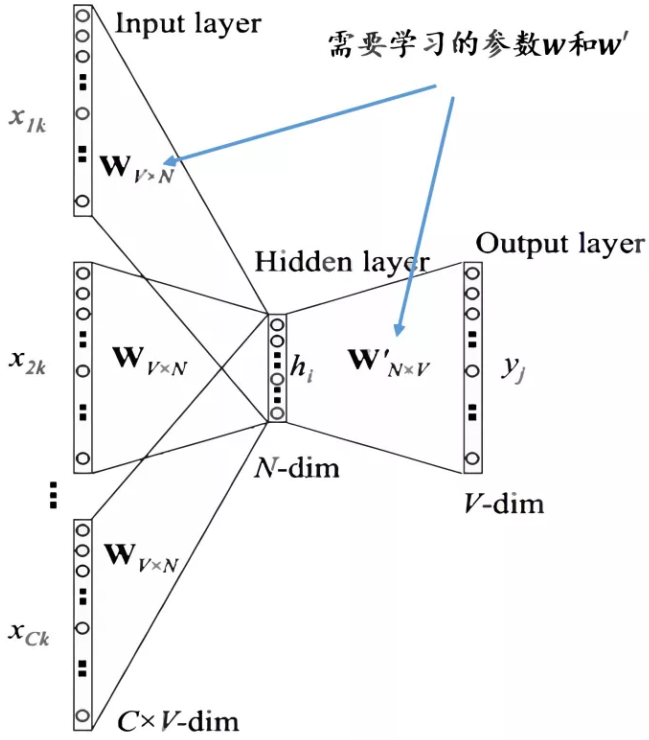

图3 CBOW的训练权重

CBOW模型训练的基本步骤包括：

1. 将上下文词进行one-hot表征作为模型的输入，其中词汇表的维度为，上下文单词数量为；
2. 然后将所有上下文词汇的one-hot向量分别乘以共享的输入权重矩阵；
3. 将上一步得到的各个向量相加取平均作为隐藏层向量；
4. 将隐藏层向量乘以共享的输出权重矩阵；
5. 将计算得到的向量做softmax激活处理得到维的概率分布，取概率最大的索引作为预测的目标词。

下面以具体例子来说明。假设语料为I learn NLP everyday, 以I learn everyday作为上下文词，以NLP作为目标词。将上下文词和目标词都进行one-hot表征作为输入，如图4所示。

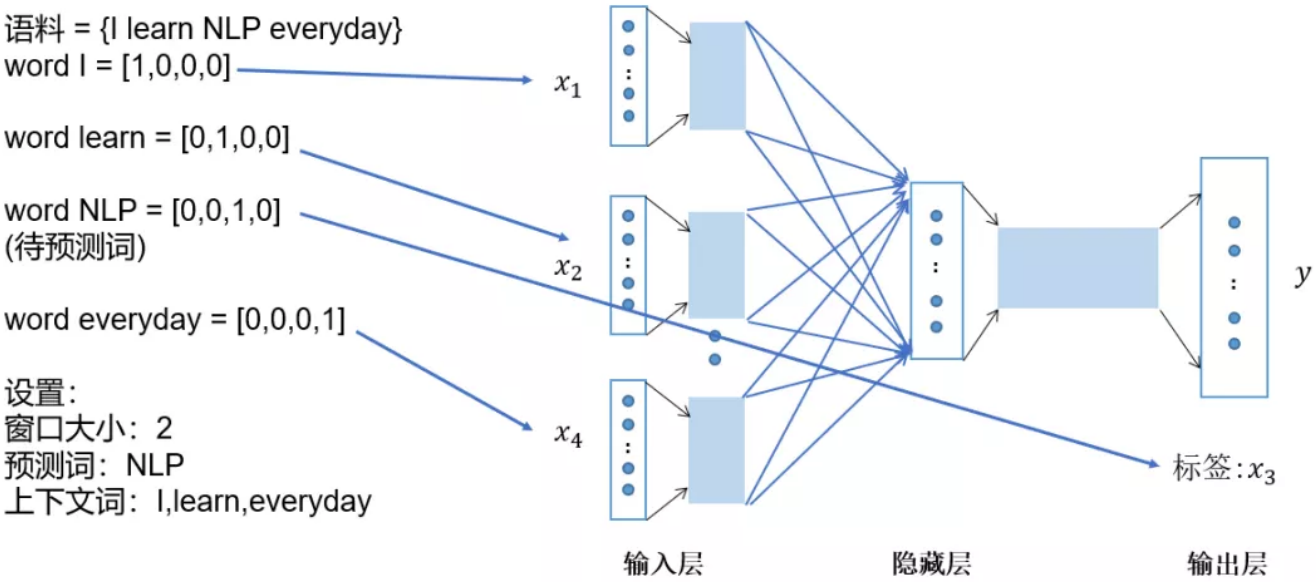

图4 CBOW训练过程1：输入one-hot表征

然后将one-hot表征分别乘以输入层权重矩阵，这个矩阵也叫嵌入矩阵，可以随机初始化生成。如图5所示。

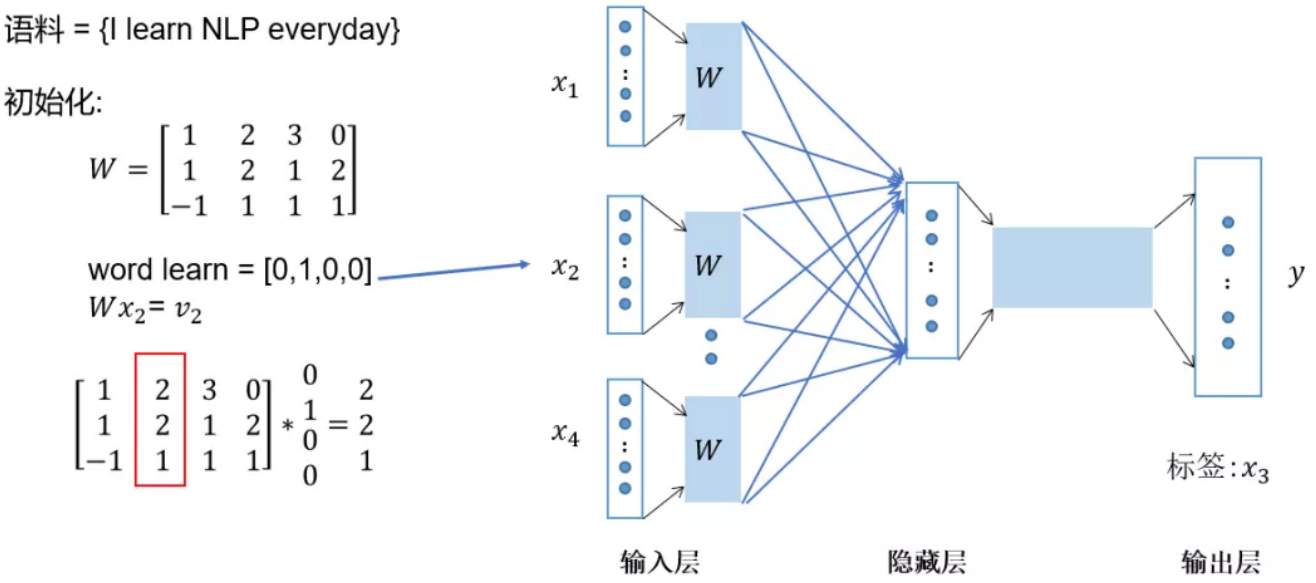

图5 CBOW训练过程2：one-hot输入乘以嵌入矩阵

然后将得到的向量结果相加求平均作为隐藏层向量，如图6所示。

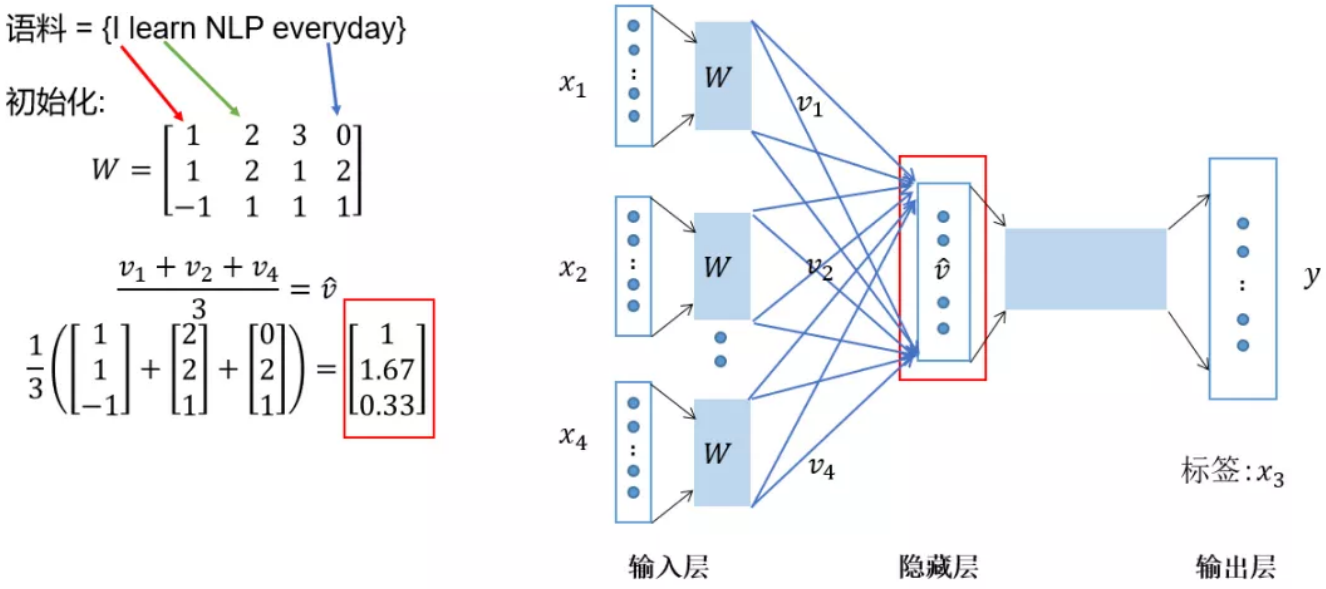

图6 CBOW训练过程3：求平均

然后将隐藏层向量乘以输出层权重矩阵，这个矩阵也是嵌入矩阵，可以初始化得到。得到输出向量，如图7所示。

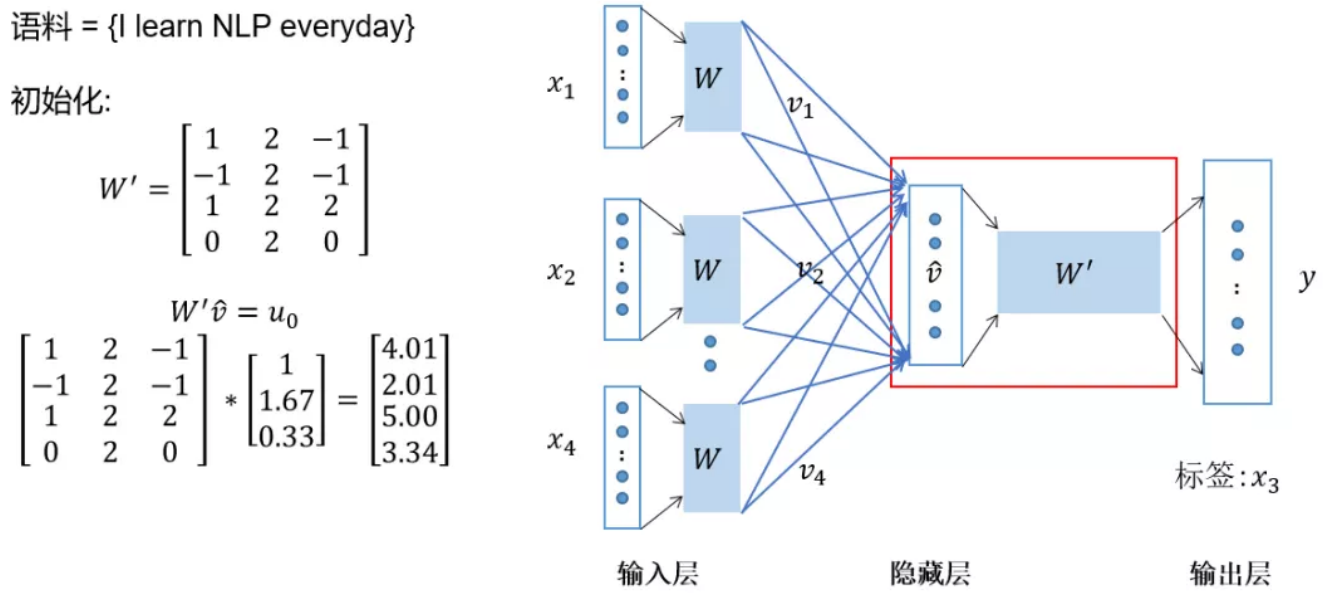

图7 CBOW训练过程4：隐层向量乘以嵌入矩阵

最后对输出向量做Softmax激活处理得到实际输出，并将其与真实标签做比较，然后基于损失函数做梯度优化训练。

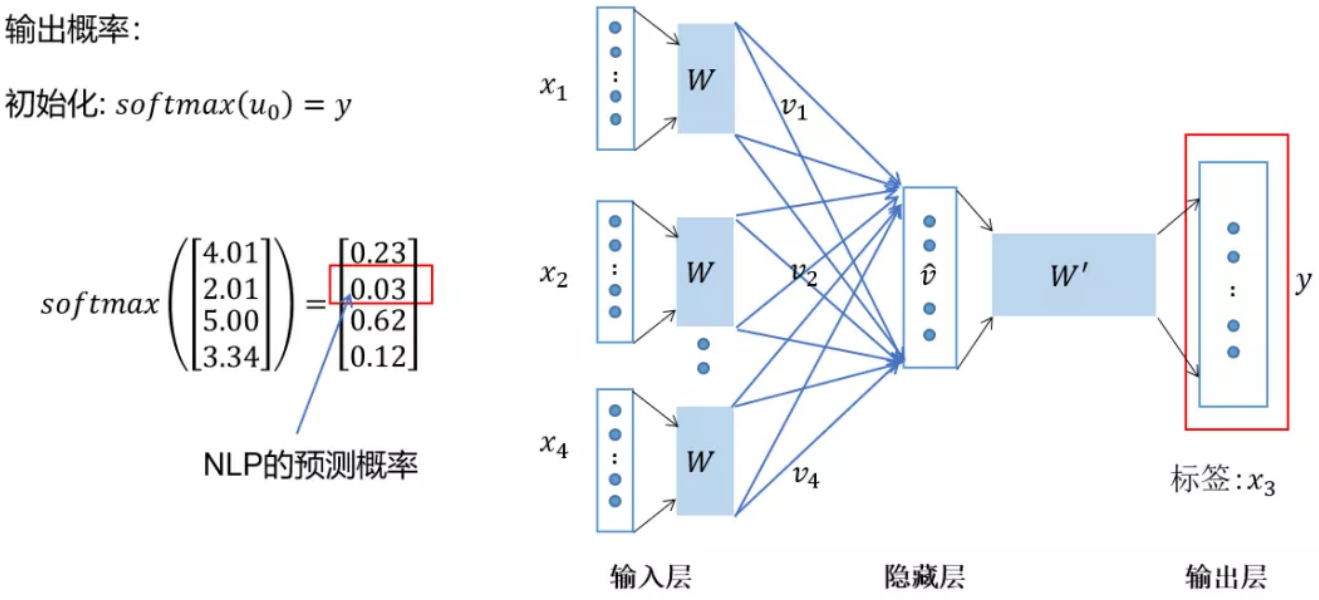

图8 CBOW训练过程5：Softmax激活输出

以上便是完整的CBOW模型计算过程，也是word2vec将词汇训练为词向量的基本方法之一。无论是Skip-gram模型还是CBOW模型，word2vec一般而言都能提供较高质量的词向量表达，图9是以50000个单词训练得到的128维的skip-gram词向量压缩到2维空间中的可视化展示。

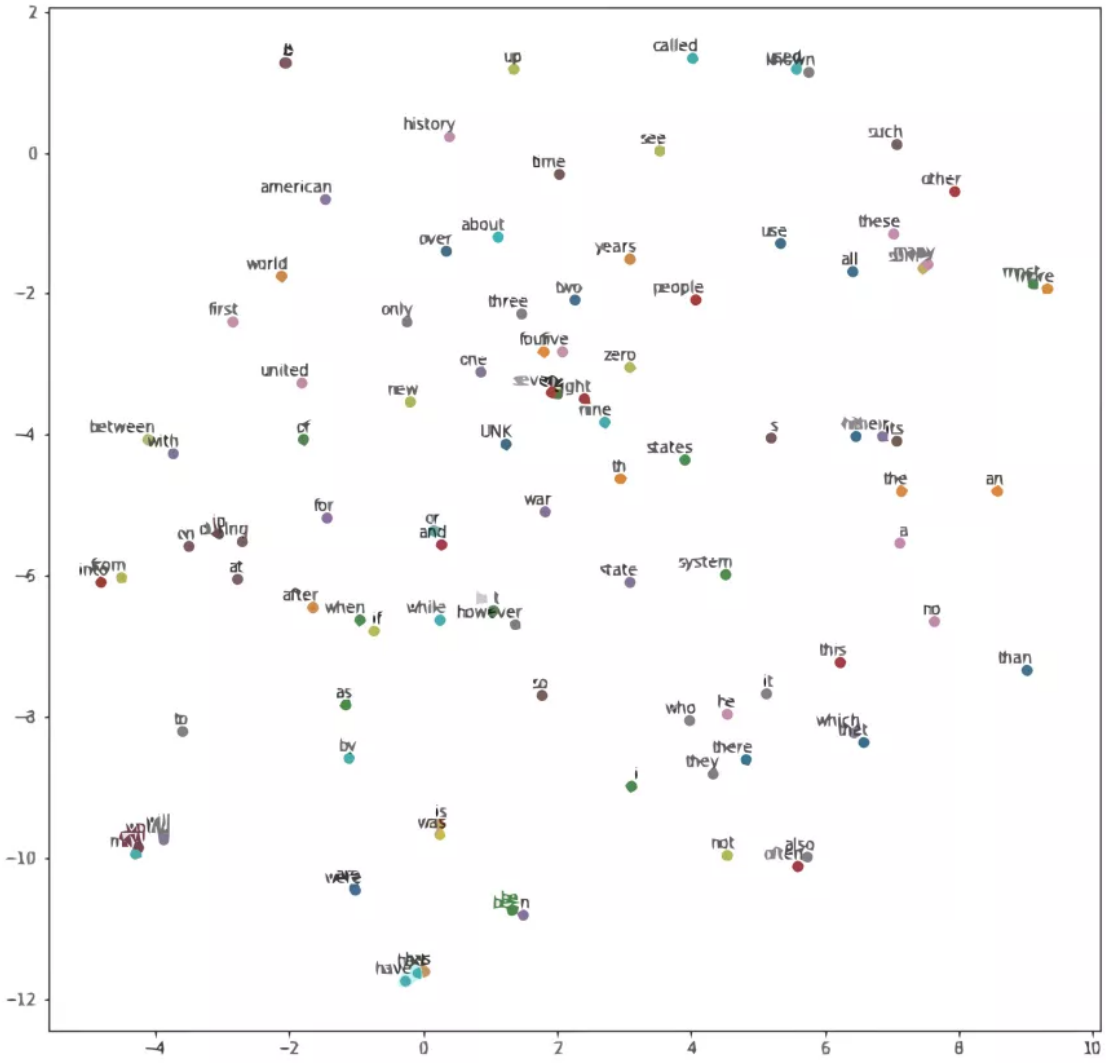

图9 word2vec的可视化效果

可以看到，意思相近的词基本上被聚到了一起，也证明了word2vec是一种可靠的词向量表征方式。