文章编号：1003-0077(2021)04-0016-07

一种改进的 GloVe 词向量表示学习方法

石隽锋，李济洪，王瑞波

（山西大学 现代教育技术学院，山西 太原 030006）

摘 要：GloVe 模型是一种广泛使用的词向量表示学习的模型。许多研究发现，学习得到的词向量维数越大，性能越好；但维数越大，模型学习耗时越长。事实上，GloVe 模型中，耗时主要表现在两方面，一是统计词对共现矩阵，二是训练学习词向量表示。该文在利用 GloVe 模型统计语料中词对共现时，基于对称或非对称窗口得到两个共现矩阵，然后分别学习得到较低维度的词向量表示，再拼接得到较高维度的词向量表示。从计算的复杂度来看，该文方法并不会产生多的计算量，但显然统计共现矩阵和训练学习可通过并行方式实现，能够显著提高计算效率。在使用大规模语料的实验中，以对称和非对称窗口分别统计得到共现矩阵，分别学习得到 300 维词向量表示，再使用拼接方式得到 600 维词向量表示。与 GloVe 模型对称和非对称的 600 维的词向量相比，在中文和英文的词语推断任务上，显著地提高了预测的准确率，在词语聚类任务上，有较好的聚类效果，验证了该文方法的有效性。

关键词：GloVe 模型；拼接的词向量；词语推断任务

中图分类号：TP391 **文献标识码**：A

An Improved Learning Method for GloVe Word Vector Representation

SHI Junfeng, LI Jihong, WANG Ruibo

(School of Modern Educational Technology, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: GloVe model is a popular word vector model, which is revealed a better performance with the increase of word vector dimensions. To avoid intolerable time cost in training high-dimension word vector, we propose an improved method of GloVe which is easily implemented in a parallel manner. We first construct a co-occurrence matrix with symmetrical windows and a co-occurrence matrix with asymmetrical windows on a corpus respectively, and apply the original GloVe model for two low-dimension word vectors. Then we concatenate the word vectors as the final high-dimension vectors. Tested in large-scale corpus, the 600-dimension word vector trained by the proposed method achieves better performance in the Chinese and English word analogy task and word clustering task, compared with the 600-dimension word vector trained directly by GloVe model.

Keywords: GloVe model; concatenated word vector; word analogy task

0 引言

近些年来，在自然语言处理领域，预训练词语分布式表示在很多任务中有很好的性能，这些任务包括文本分类[1]、问答系统[2]、命名实体识别[3]、语义角色标注[4]等。为此，研究人员提出了很多预训练的语言模型[5-9]，较为典型的方法有 SENNA 模型[5]、Word2Vec 模型[6]、GloVe 模型[7]、ELMo 模型[8]、BERT 模型[9]。其中，GloVe 模型基于任意两个词之间的全局共现信息，再采用 Log-Bilinear 模型学习得到词向量表示，在多项评测任务中表现良好。通常情况下，得到的词向量的维数越高模型性能越好，但维数越大则训练耗时越长。一个自然的想法是，利用并行计算分别学习较低维数的词向量再拼接，能否得到性能更高的词向量？事实上，在 GloVe 模型中，共现矩阵的统计方法采用了两种，一种是对称窗口的方法，即取目标词两侧固定窗口内的词语作为其上下文；另一种是非对称窗口的方法，即取目标词左侧的固定窗口内的词语作为其上下

收稿日期：2019-12-15 **定稿日期**：2020-01-12

基金项目：国家自然科学基金(61806115)

文，不同的共现矩阵会反映不同的句法和语义信息。为此，在 GloVe 模型中，本文以对称和非对称窗口统计得到两个共现矩阵，分别学习得到词向量表示，然后再采用拼接的方式，得到较高维度的词向量表示。在验证实验中，我们分别学习得到的 300 维向量，再拼接得到 600 维向量表示，在中文和英文的词语推断任务的评测集上，预测的准确率得到显著提升。

1　相关工作

在自然语言处理领域，词语的分布式表示(distributional representation)是将词的上下文信息表示为词向量的形式，这种词向量构建的基础是 1957 年 Firth 提出的分布式假说(distributional hypothesis)[10]，即一个词语的语义信息是由其周围的词语来刻画的(a word is characterized by the company it keeps)。科研人员提出了多种词向量的构造方法。Burgess 等[11]构造的词向量的每一维上表示目标词和其上下文词语共现的频次，而有些研究人员[12-13]用目标词和它的上下文的逐点互信息(pointwise mutual information，PMI)或正逐点互信息(positive pointwise mutual information，PPMI)代替了频次。词语的共现范围通常用滑动窗口的方法来实现[11]，给定窗口的大小为 w，通过在语料上逐词地滑动窗口。在每个窗口里，共现的词对的频次的和形成共现矩阵，词对是有序的，即只统计目标词左侧上下文的频次，而把目标词和上下文交换角色后，就可以得到目标词右侧上下文的频次。文献[14]系统地比较了不同的距离测度对不同的共现矩阵(PMI 共现矩阵、PPMI 共现矩阵)得到的词向量在各种任务上的性能。在 PPMI 共现矩阵中，分出了四种共现矩阵，即基于左侧共现、右侧共现、及左右侧共现相加、左右侧共现拼接的共现矩阵，依次表示为：L，R，L+R，L&R，在语义聚类任务和句法聚类任务上比较了基于四种共现矩阵的性能，发现在语义聚类任务上，基于 L&R 的词向量性能略高于基于 L+R 的词向量；在句法聚类任务上，基于 L&R 的词向量在维数较高的情况下，性能显著高于基于 L+R 的词向量。词语的分布式表示是高维的、稀疏的向量，不利于进行语义计算。为此，科研人员提出了一些降低维度的方法，文献[15]对词对的频次排序，设定阈值，删掉词对频次低于阈值的维数，使得词向量的维数大大降低。文献[16]提出了奇异值分解方法，将文档矩阵进行分解，降低了词向量的维数，文献[17]是对共现的 PPMI 矩阵进行因式分解。近些年来，科研人员通过神经网络训练词语的低维表示。Word2Vec 模型[6]包括 CBOW 模型和 Skip-gram 模型，目标函数为目标词和上下文的关系，CBOW 模型的目标函数为通过上下文预测目标词，而 Skip-gram 模型的目标函数为通过目标词预测上下文。文献[18-21]都是在 CBOW 和 Skip-gram 模型基础上进一步考虑了词语在句子中的位置以及和目标词的关联程度提出的改进模型，这些模型在句法任务上性能均有所提升。文献[18]采用了基于句法关系的上下文训练的词向量作为依存句法解析的特征，来提高模型性能。文献[19]在 CBOW 模型和 Skip-gram 模型的基础上添加更多的参数，保留上下文和目标词之间的位置信息。但模型的复杂度会随着窗口的增大线性增加。文献[20]在 CBOW 模型的基础上，根据上下文的不同类型以及和目标词的相对位置的不同，为上下文分配不同的权重。文献[21]引入一个方向向量来表示上下文是在目标词的左边还是右边，从而提高 Skip-gram 模型的性能。文献[22]提出了采用基于句法关系的上下文训练词向量的方法，在 Skip-gram 模型上，比较了基于句法关系的上下文和基于滑动窗口的上下文训练得到的词向量，发现通过基于句法关系的词向量找到的相似词语中功能型相似(functional similarity)的词语比较多，基于滑动窗口的词向量找到的相似词语中主题相似(topical similarity)的词语比较多，例如，“佛罗里达州”在第一种上下文的词向量下得到的相似的词语为其所属的国家或者它包含的城市，在第二种上下文的词向量下得到的相似词语是美国的一些其他的州。因此基于滑动窗口上下文的词向量表示和基于句法上下文的词向量表示各有优劣。应当把这两种词向量表示结合起来使用。基于 GloVe 模型有两种统计共现矩阵的方式，一种是对称窗口方式，没有考虑词语顺序；另一种是非对称窗口方式，考虑了上下文在目标词的前后顺序。因此，我们有必要将两种共现矩阵得到的词向量结合起来，得到精度更高的词向量表示，来更好地完成语义和句法任务。

2　GloVe 模型

GloVe 模型可以分别训练出基于对称共现矩阵的低维词向量和基于非对称共现矩阵的低维词

向量。

GloVe 模型训练基于对称共现矩阵的低维词向量的步骤如下：

(1) 从语料库统计出词表。从给定语料库统计每个不同的词语出现的次数，按照频次从高到低排序，c_i 表示第 i 个词，f_i 表示第 i 个词的频次，$1\leqslant i\leqslant n$，其中 n 为语料库中不同的词语个数。

(2) 设定固定窗口大小为 w，依次遍历语料库中的词语，统计目标词两侧固定窗口内的词语的频次，生成对称共现矩阵，表示为 $\boldsymbol{X}^S$。矩阵的大小为 $n\times n$。矩阵的行和列为词表中的每个词的序号。用 $\boldsymbol{X}_{ij}^S$ 表示对称共现矩阵第 i 行第 j 列的元素。

(3) 用 $\boldsymbol{v}^S$ 表示基于对称共现矩阵训练得到的低维词向量。训练 $\boldsymbol{v}^S$ 的目标函数如式(1)所示。

$$J^S=\sum_{i,j=1}^{n}f(\boldsymbol{X}_{ij}^S)((\boldsymbol{v}_i^S)^T\tilde{\boldsymbol{v}}_j^S+\boldsymbol{b}_i^S+\boldsymbol{b}_j^S-\log\boldsymbol{X}_{ij}^S)^2 \tag{1}$$

其中，$\boldsymbol{v}_i^S$ 和 $\tilde{\boldsymbol{v}}_j^S$ 分别表示词 c_i 和 c_j 的对称低维词向量表示，$\boldsymbol{b}_i^S$ 和 $\boldsymbol{b}_j^S$ 为 $\boldsymbol{v}_i^S$ 和 $\tilde{\boldsymbol{v}}_j^S$ 对应的偏置项，$f(X_{ij}^S)$为权重函数。

GloVe 模型训练基于非对称共现矩阵的低维词向量的步骤如下：

(1) 从语料库统计出词表。从给定语料库统计每个不同的词语出现的次数，按照频次从高到低排序，c_i 表示第 i 个词，f_i 表示第 i 个词的频次，$1\leqslant i\leqslant n$，其中 n 为语料库中不同的词语个数。

(2) 设定固定窗口大小为 w，依次遍历语料库中的词语，统计目标词左侧固定窗口内的词语的频次，生成左侧共现矩阵，表示为 $\boldsymbol{X}^L$。用 $\boldsymbol{X}_{ij}^L$ 表示左侧共现矩阵第 i 行第 j 列的元素。

(3) 用 $\boldsymbol{v}^A$ 表示基于左侧共现矩阵训练得到的低维词向量。训练 $\boldsymbol{v}^A$ 的目标函数如式(2)所示。

$$J^A=\sum_{i,j=1}^{n}f(\boldsymbol{X}_{ij}^L)((\boldsymbol{v}_i^A)^{\mathrm{T}}\tilde{\boldsymbol{v}}_j^A+\boldsymbol{b}_i^A+\tilde{\boldsymbol{b}}_j^A-\log\boldsymbol{X}_{ij}^L)^2 \tag{2}$$

其中，$\boldsymbol{v}_i^A$ 和 $\tilde{\boldsymbol{v}}_j^A$ 分别表示词 c_i 和 c_j 的非对称低维词向量表示，$\boldsymbol{b}_i^A$ 和 $\tilde{\boldsymbol{b}}_j^A$ 为 $\boldsymbol{v}_i^A$ 和 $\tilde{\boldsymbol{v}}_j^A$ 对应的偏置项，$f(X_{ij}^L)$为权重函数。

3 GloVe 词向量拼接模型

本文提出了 GloVe 词向量拼接模型，该模型并行训练出只有一半维数的 $\boldsymbol{v}^A$ 和 $\boldsymbol{v}^S$，再将它们拼接起来，完成词语推断任务。具体步骤如下：

(1) 从语料库统计出词表。从给定语料库统计每个不同的词语出现的次数，按照频次从高到低排序，c_i 表示第 i 个词，f_i 表示第 i 个词的频次，$1\leqslant i\leqslant n$，其中 n 为语料库中不同的词语个数。

(2) 设定固定窗口大小为 w，依次遍历语料库中的词语，并行统计出左侧共现矩阵和对称共现矩阵 $\boldsymbol{X}^L$ 和 $\boldsymbol{X}^S$。两个矩阵的大小都为 $n\times n$。$\boldsymbol{X}^L$ 和 $\boldsymbol{X}^S$ 都是按词频排序的。

(3) 并行打乱 $\boldsymbol{X}^L$ 和 $\boldsymbol{X}^S$ 的顺序。

(4) 在两个处理器上，设置维数为 GloVe 模型的一半，分别用式(1)训练出 $\boldsymbol{v}^S$，用式(2)训练出 $\boldsymbol{v}^A$。

(5) 将 $\boldsymbol{v}^A$ 和 $\boldsymbol{v}^S$ 拼接起来作为词语的低维词表示。

4 实验

实验环境为山西大学高性能计算平台。

4.1 在英文词语推断任务上比较

从 English Wikipedia 语料分割出三个不同大小的语料，分别包含 2 亿、5 亿、10 亿个单词，文件大小分别为 1.09 GB、2.71 GB、5.42 GB。滑动窗口大小(window-size)设置为 10，词典中的最大词数(max-vocab)设为 100 000，用 GloVe 模型训练出 600 维的 $\boldsymbol{v}^S$ 和 $\boldsymbol{v}^A$，用 GloVe 词向量拼接模型训练出 600 维的 $\boldsymbol{v}^S$ 和 $\boldsymbol{v}^A$ 的拼接向量($\boldsymbol{v}^S$ 和 $\boldsymbol{v}^A$ 的维数都是 300 维)，在词语推断任务[3]上比较它们的准确率，实验结果如下，词语推断任务的测试集包括语义任务(capital：country，city：state，family)和句法任务(adjective：adverb，opposite，comparative 等)，结果如表 1～表 3 所示。

从表 1 可以看出，GloVe 词向量拼接模型得到的词向量在语义任务、句法任务和总任务上的准确率均有不同程度的提升，句法任务和总任务上提升

表 1 1.09 GB English Wikipedia 语料下的比较结果

(单位：%)

词向量	准确率		
	语义任务	句法任务	总任务
$\boldsymbol{v}^S$	79.95	47.85	62.15
$\boldsymbol{v}^A$	79.90	48.55	62.52
$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$	82.01	50.85	64.74

表 2　2.71 GB English Wikipedia 语料下的比较结果

（单位：%）

词向量	准确率		
	语义任务	句法任务	总任务
$\boldsymbol{v}^S$	84.17	56.94	69.11
$\boldsymbol{v}^A$	85.01	56.58	69.29
$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$	84.62	58.93	70.42

表 3　5.42 GB English Wikipedia 语料下的比较结果

（单位：%）

词向量	准确率		
	语义任务	句法任务	总任务
$\boldsymbol{v}^S$	84.41	60.94	71.41
$\boldsymbol{v}^A$	84.78	60.45	71.31
$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$	84.40	62.01	72.00

较大。从表 2 和表 3 可以看出，GloVe 词向量拼接模型得到的词向量在句法任务上有较大提升，在总任务上准确率也有所提升。综合表 1 到表 3，GloVe 词向量拼接模型在句法任务上性能较好，在较小的语料库上性能提升得较大。随着语料规模的扩大，在“$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$”词向量下，语义任务上的准确率先升后降（82.01%→84.62%→84.40%），这是因为 max-vocab 参数的设置，该参数限制了词典的最大词数，在不同大小的语料上，词典里的词按照频次从高到低排序，词数相同，使得保留下来的词并不相同，较大的语料保留了词频较高的词，但可能删去了一些有意义的上下文词语。因此，语料大也可能使准确率下降。由于实验目的是比较在相同语料规模下，GloVe 模型训练出词向量和 GloVe 词向量拼接模型训练出的词向量的性能，因此，没有考虑三个语料下要统一词表。

4.2　在中文词语推断任务上比较

本文在中文的词语推断任务上也做了相同的实验，中文语料采用 1998 年和 2000 年人民日报语料合并后的语料，大小为 186 MB，中文的词语推断任务的测试集是文献[23]提供的，只包含语义任务（首都：国家，省会：省，家庭关系），用 GloVe 模型训练出 600 维的 $\boldsymbol{v}^S$ 和 $\boldsymbol{v}^A$，用 GloVe 词向量拼接模型训练出 600 维的 $\boldsymbol{v}^S$ 和 $\boldsymbol{v}^A$ 的拼接向量（$\boldsymbol{v}^S$ 和 $\boldsymbol{v}^A$ 的维数都是 300 维），在中文的词语推断任务上进行比较，实验结果如表 4 所示。

表 4　人民日报语料下的比较结果（单位：%）

词向量	语义任务准确率
$\boldsymbol{v}^S$	82.27
$\boldsymbol{v}^A$	85.07
$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$	87.32

从表中的数据可以看出，GloVe 词向量拼接模型得到的词向量准确率有大幅提高。

4.3　显著性检验

本文对表 1 中的数据用 χ^2 检验方法进行了显著性检验，如式(3)所示。

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \tag{3}$$

在本实验中，n_{01} 表示使用 GloVe 词向量拼接模型预测错误而 GloVe 模型预测正确的词语个数，n_{10} 表示使用 GloVe 词向量拼接模型预测正确而 GloVe 模型预测错误的词语个数，通过计算得到的 χ^2 值如表 5 所示。

表 5　“$\boldsymbol{v}^S$”和“$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$”在各个任务上的 χ^2 值

词向量	语义任务	句法任务	总任务
“$\boldsymbol{v}^S$”和“$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$”	58.9	99.1	158.4
“$\boldsymbol{v}^A$”和“$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$”	56.3	52.1	105.1

表中的 χ^2 值都远大于 $\chi^2_{0.05} = 3.84$，因此，认为在显著性水平 0.05 下，GloVe 词向量拼接模型的拼接向量和 GloVe 模型的对称和非对称词向量的实验结果是有显著性差异的，即 GloVe 词向量拼接模型比 GloVe 模型的性能有显著提升。

4.4　词语聚类的示例

词语聚类的效果可以检验词向量性能。通常可以通过计算词语向量的相邻词，观察这些学习到的词向量表示的好坏。本文采用词向量的余弦相似度来度量词语的相邻程度。采用 4.2 节训练的词向量。表 6 和表 7 分别列出了在“$\boldsymbol{v}^S$”“$\boldsymbol{v}^A$”和“$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$”的词向量下，英国、德国最相邻的 10 个词。

可以看出，与“$\boldsymbol{v}^S$”与“$\boldsymbol{v}^A$”词向量相比，在“$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$”词向量下，词语的余弦相似度较大，说明聚在一起的相似的词语比较多。

通过列出的 10 个近邻词语可以看出，在“$\boldsymbol{v}^S$”词向量下，列出了更多语义上比较接近的词，在“$\boldsymbol{v}^A$”词向量下，列出了更多句法上接近的词语，在“$\boldsymbol{v}^A$ 拼

表6 “英国”在“v^S”、“v^A”和“v^A 拼接 v^S”词向量下的10近邻词及余弦相似度

英国					
v^S		v^A		v^A 拼接 v^S	
法国	0.547 7	法国	0.550 5	法国	0.701 8
伦敦	0.536 0	布莱尔	0.527 6	德国	0.646 1
布莱尔	0.525 7	德国	0.483 7	**美国**	**0.619 2**
德国	0.504 1	西班牙	0.479 5	布莱尔	0.602 8
西班牙	0.493 1	伦敦	0.479 2	西班牙	0.599 0
意大利	0.460 1	意大利	0.463 5	意大利	0.575 1
牛津	**0.459 5**	加拿大	0.459 9	**日本**	**0.573 9**
荷兰	0.451 3	**美国**	**0.446 7**	加拿大	0.565 1
爱丁堡	0.450 8	**日本**	**0.427 8**	伦敦	0.565 0
加拿大	0.449 2	**澳大利亚**	**0.424 7**	荷兰	0.534 2

表7 “德国”在“v^S”、“v^A”和“v^A 拼接 v^S”下的10个近邻词及余弦相似度

德国					
v^S		v^A		v^A 拼接 v^S	
法国	0.607 5	法国	0.600 1	法国	0.738 5
意大利	0.555 0	意大利	0.560 3	意大利	0.699 6
欧洲	0.514 6	施罗德	0.495 7	英国	0.646 1
英国	0.504 1	英国	0.483 7	欧洲	0.600 0
施罗德	**0.482 2**	欧洲	0.481 3	加拿大	0.589 1
奥地利	0.477 1	奥地利	0.467 7	奥地利	0.580 3
加拿大	0.474 4	比利时	0.456 3	**日本**	**0.578 7**
澳大利亚	0.461 3	**荷兰**	**0.452 6**	**荷兰**	**0.562 2**
纳粹	**0.439 5**	加拿大	0.444 3	美国	0.552 9
比利时	0.439 3	**日本**	**0.439 7**	**施罗德**	**0.549 0**

接 v^S”词向量下，列出了更多句法和语义上接近的词语。

比如，在“英国”的10个近邻词中，在“v^S”词向量下，“英国”的相邻词中包括“牛津”，而在“v^A”词向量下没有这个词；在“v^A”词向量下，“英国”的相邻词中包括“美国”“日本”“澳大利亚”，而在“v^S”词向量下没有这些词。在“v^A 拼接 v^S”词向量下，“英国”的10个近邻词中包括“美国”“日本”，不包括“牛津”“澳大利亚”。但“英国”的第14近邻词为“澳大利亚”，和“英国”词向量的余弦相似度为0.502 7，“英国”的第15近邻词为“牛津”，和“英国”词向量的余弦相似度为0.495 3。虽然这两个词不在“英国”的前10个近邻祠内，但是，在“v^A 拼接 v^S”下，这两个词和“英国”的余弦相似度分别比在“v^A”和“v^S”词向量下的大。例如，在“v^A 拼接 v^S”下，“澳大利亚”和“英国”的词向量的余弦相似度为0.502 7，0.502 7>0.424 7(“v^A”下“英国”和“澳大利亚”的余弦相似度)，同样，在“v^A 拼接 v^S”下，“牛津”和“英国”的词向量的余弦相似度为0.495 3，0.495 3>0.459 5(v^S 下“英国”和“牛津”的余弦相似度)。同样，在“v^A 拼接 v^S”下，“美国”“日本”和“英国”的余弦相似度比“v^A”词向量下的余弦相似度大。

比如，在“德国”的10个近邻词中，在“v^S”词向量下，“德国”的相邻词中包括“施罗德”(德国前总理)、“纳粹”，而在“v^A”词向量下没有这两个词。在“v^A”词向量下，“德国”的相邻词中包括“荷兰”“日本”，而在“v^S”词向量下没有这些词。在“v^A 拼接 v^S”词向量下，“德国”的10个近邻词中包括“施罗德”“荷兰”“日本”，不包括“纳粹”。但“德国”的第22近邻词为“纳粹”，余弦相似度比“v^S”词向量下的大，为0.482 9，0.482 9>0.439 5(“v^S”下“德国”和“纳粹”的余弦相似度)。在“v^A 拼接 v^S”词向量下，“德国”的第7近邻词为“日本”，余弦相似度为0.578 7，0.578 7>0.439 7(“v^A”下“德国”和“日本”的余弦相似度)。在“v^A 拼接 v^S”词向量下，“施罗德”“荷兰”的余弦相似度分别比在“v^S”和“v^A”词向量下的大，由于篇幅所限，在此不一一列举。

总的来说，在“$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$”下，词语的近邻词中包括了更多语义和句法上相近的词语。“$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$”得到的词向量在词语聚类上的表现优于“$\boldsymbol{v}^S$”和“$\boldsymbol{v}^A$”词向量。

4.5 运行时间

本文统计了4.1节在1.09 GB的English Wikipedia语料下完成词语推断任务时，GloVe模型和GloVe词向量拼接模型运行的时间，如表8所示。

表8 1.09 GB English Wikipedia语料下的运行时间

词向量	运行时间	是否并行
$\boldsymbol{v}^S$	3h56min	否
$\boldsymbol{v}^A$	2h45min	否
$\boldsymbol{v}^A$ 拼接 $\boldsymbol{v}^S$	1h58min	是

因此，对GloVe模型，采用并行的训练学习方法，既可以提高词向量的性能，又能节省训练时间。

4.6 实验结果分析

从大部分的词语推断任务和聚类任务的实验结果可以看出，“v^A 拼接 v^S”词向量在语义任务和句法任务上都超过了“v^S”词向量和“v^A”词向量。原因是“v^S”词向量和“v^A”词向量共现矩阵构造过程不同，反映的句法和语义信息也不同。“v^A 拼接 v^S”词向量能够体现更完整的句法和语义信息。

“v^S”词向量的共现矩阵构造方法为：在语料库上，从开始位置滑动固定大小的窗口，统计目标词两侧固定窗口内的词语的频次，生成对称共现矩阵。

“v^A”词向量的共现矩阵构造方法为：在语料库上，从开始位置滑动固定大小的窗口，统计目标词左侧固定窗口内的词语的频次，生成左侧共现矩阵。左侧共现矩阵的转置即为右侧共现矩阵，因此右侧共现矩阵不需要单独统计。

“v^A”词向量的共现矩阵保存了词语在目标词左右的位置信息，而“v^S” 词向量的共现矩阵将目标词左侧和右侧的相同词语的频次求和，使得共现矩阵中混合了目标词之前和之后的上下文词语。“v^A”词向量聚类能将句法相近的词语更好地聚在一起，而“v^S”词向量聚类能将语义相近的词更好地聚在一起。因此，“v^A”词向量更多地体现句法信息，而“v^S”词向量更多地体现语义信息。

“v^A 拼接 v^S”词向量是将“v^A”词向量和“v^S”词向量拼接起来，融入了“v^A”词向量和“v^S”词向量的信息，因此该词向量能体现更多的句法和语义信息。

5 结论与展望

事实上，表示学习的理论依据是词的意义是由与其共现的词来体现的，意义的不同体现了其共现词语的差异。GloVe 模型中共现是以滑动窗口的方式来统计的，显然，对许多词，使用词的左侧、右侧窗口或对称窗口来计算共现能够体现词组合的不同分布特性。因此，采用多种方式而不是仅仅用对称窗口方式得到共现矩阵，应该可以学习到更为准确的词表示向量。

本文提出了 GloVe 词向量拼接模型，使用不同的共现矩阵，并采用并行处理分别学习较低维度的词向量，再采用拼接方式得到较高维度的词向量表示，减少了词向量的训练时间。实验结果表明，由 GloVe 拼接模型得到的词向量在词语推断任务上和词语聚类任务上性能有显著提升。下一步我们将研究如何得到反映多种层面信息的共现矩阵，有效集成多种词表示向量，提高表示学习的性能。

参考文献

[1] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[2] Tellex S, Katz B, Lin J, et al. Quantitative evaluation of passage retrieval algorithms for question answering [C]//Proceedings of the 26th Annual International ACM SIGIR Conference, Toronto, Canada, 2003.

[3] Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 384-394.

[4] He L, Lee K, Lewis M, et al. Deep semantic role labeling: What works and what's next[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 473-483.

[5] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2461-2505.

[6] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]// Proceedings of the International Conference on Learning Representations, 2013.

[7] Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014: 1532-1543.

[8] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2018: 2227-2237.

[9] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805v1, 2018.

[10] Firth J R. A synopsis of linguistic theory 1930-1955 [G]. Selected Papers of J R Firth, Longman, London. 1957: 168-205.

[11] Burgess C, Lund K. Modeling cerebral asymmetries in high-dimensional semantic space[G]. Right Hemisphere Language Comprehension: Perspectives from Cognitive Neuroscience, Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 1998: 215-244.

[12] Church K W, Hanks P. Word association norms, mutual information, and lexicography[J]. Computational Linguistics, 1990, 16(1):22-29.

[13] Baroni M, Dinu G, Kruszewski G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, USA, 2014: 238-247.

[14] Bullinaria J A, Levy J P. Extracting semantic representations from word co-occurrence statistics: A computational study[J]. Behavior Research Methods, 2007, 39(3): 510-526.

[15] Padró M, Idiart M, Ramisch C, et al. Nothing like good old frequency: Studying context filters for distributional thesauri[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (Short Papers), Doha, Qatar, 2014: 419-424.

[16] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[17] Salle A, Idiart M, Villavicencio A. Matrix factorization using window sampling and negative sampling for improved word representations[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, 2016: 419-424.

[18] Bansal M, Gimpel K, Livescu K. Tailoring continuous word representations for dependency parsing [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics Baltimore, Maryland, 2014: 809-815.

[19] Wang L, Dyer C, Black A W, et al. Two/Too simple adaptations of Word2Vec for syntax problems[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, 2015: 1299-1304.

[20] Wang L, Tsvetkov Y, Silvio Amir, et al. Not all contexts are created equal: Better word representations with variable attention[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015: 1367-1372.

[21] Song Y, Shi S, Li J, et al. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings[C]//Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 175-180.

[22] Levy O, Goldberg Y. Dependency-based word embeddings[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 302-308.

[23] Chen X, Xu L, Liu Z, et al. Joint learning of character and word embeddings[C]//Proceedings of the 24th International Joint Conference on Artificial Intelligence, San Francisco, CA: Morgan Kaufmann, 2015: 1236-1242.

石隽锋(1979—),硕士,讲师,主要研究领域为统计自然语言处理。
E-mail: sjf@sxu.edu.cn

王瑞波(1985—),博士研究生,工程师,主要研究领域为统计自然语言处理。
E-mail: wangruibo@sxu.edu.cn

李济洪(1964—),通信作者,博士,教授,主要研究领域为统计机器学习、统计自然语言处理。
E-mail: lijh@sxu.edu.cn