周报\_2018\_9\_30

1. node2vec:Scalable Feature Learning for Networks (2016)

监督学习算法往往需要格式化的、有识别力的、独立的特征。一种典型的解决方法是让在问题所在的领域的专家进行人工制作特征，但是这种特征既不通用效果也不是很好。因此该文章提出的是一种将图中的节点转化成embedding向量，这个embedding 向量在一定程度上表达了该图的结构以及节点之间的相互关系。这个embedding向量因为是独立于具体的监督学习任务的，它仅仅是用于对节点进行表达的，因此它具有通用性。

该文章对图的处理也是使用random walk将它采样成一系列walks的集合，然后计算得到node的embedding（文章中并没有说明具体对walks如何处理得到embedding，只有等之后看代码再理解这块了）。它的随机采样与一般的random walk有一定的区别，它引入了两个变量p，q可以使得采样可以兼顾到网络的同质性以及结构等价性（类似于同时考虑了DFS与BFS），这就使得它可以一定程度上控制搜索空间。它还提供了计算edge的embedding向量的方法，即是利用两个node的embedding做二元运算即可。（这就是“Semantic Proximity Search on Heterogeneous Graph by Proximity Embedding”中所说的间接的方法，即是并不是直接计算出两个节点之间的embedding，而是先计算出node embedding，再在此基础上计算出edge embedding）。该文章的局限性在于它仅仅是用于同构图，异构图不适用。

2、metapath2vec\_ Scalable Representation Learning for Heterogeneous Networks (2017)

该文章也是介绍的在网络或者说图上的节点的representation learning，学习节点的embedding向量，然后再利用节点分类等任务来评估所得到的embedding vector的效果。该文章与node2vec以及DeepWalk的区别在于，它是在异构图上进行的节点嵌入。为了处理异构图的问题，该文章提出了meta-path-based random walk，它引入了meta-path scheme，能够指导random walker进行路径的采样，使得采样的时候能够考虑到节点的类型，从而解决异构图节点类型不同的问题。（例如：定义一个模式”APVPA”—A=Author，P=Paper，V=Venue，那么进行路径采样的时候，所采集的路径的节点的类型就依次是A,P,V,P,A，其他类型的路径不会被采集，对于APVPA的解释：它表示了author发表的paper在同一venue）。

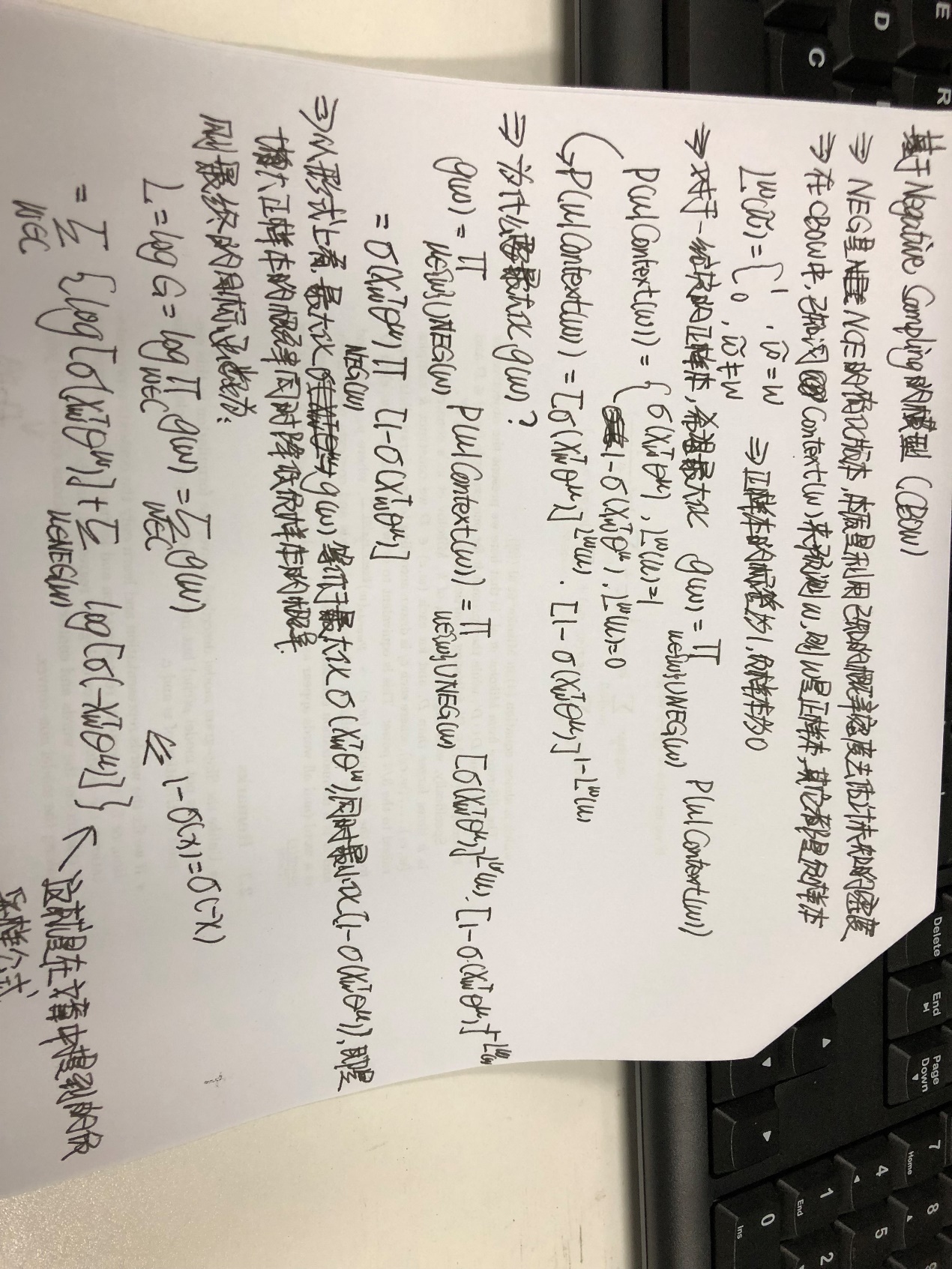
该算法的缺点是：这样的解决方案使得对于每一个不同的网络我们还需要去定义它特定的scheme，很显然对于scheme不同，那么最终得到的embedding的性能也不同。

（再回去看semantic proximity search on heterogeneous graph by proximity embedding，发现它似乎对于random walk并没有做特殊的处理，那么它为什么能够处理异构图的问题呢？：它的LSTM的输入x，构造了与节点类型相关的特征）

1. Distributed Representations of Words and Phrases and their Compositionality

在该文章之前，使用skip-gram来训练词向量需要使用分层softmax来加快训练速度，该文章提出了一种用来替代分层softmax的方法叫做负采样（negative sampling）。该文章还发现，对频率较高的词进行二次抽样，可以提高算法的速度以及生成的词向量的质量。该文章也如“DAG-Structure Long Short-Term Memory for Semantic Compositionality”一样，也认为日常用的短语不能仅仅由组成短语的单词的词向量相加得到，因此它提出了设置一个threshold，判断[count(w1w2)-a]/(count(w1)\*count(w2))与threshold的关系，如果大于则认为它是一个短语，否则不是一个短语，然后对于短语的训练，则将其当做一个整体进行训练即可。该文章还发现了另一种representation的线性结构，通常简单的向量的加法能够生成有意义的结果，例如：Vec(Czech)+Vec(currency)=Vec(koruna)。

以下是文章中所提到的负采样公式的推导过程：



4、PTE:Predictive Text Embedding through Large-scale （2015）

相比于复杂的DL框架而言，无监督学习的embedding向量往往在应用到具体的任务的时候是性能较差的。这是因为这些embedding方法只是完全的无监督学习，忽略了具体labeled data。

本文提出的方法就是对于text分类的具体任务，进行embedding，通过了三个网络（word-word Network，word-document Network，word-label Network）的结合，利用了labeled data以及unlabeled data来生成embedding向量，这样得到的embedding向量通过最后的测试发现：在Long Document上，性能是要比CNN要好的；在Short Document上，性能和CNN类似。