网络表达学习系列（一）：深度游走（Deepwalk）

原创 Xiaohan 论文收割机 2019-01-12

这次我们准备的主题是关于最近很火的主题：网络表达学习（Network Representation Learning），或者称之为网络嵌入（Network Embedding）。我们会把近年来经典的网络表达学习算法，包括Deepwalk，LINE，Node2vec等逐一解释，做成一个系列分享给大家。

什么是网络表达学习？

网络表达学习吗，顾名思义，是指学习出一个网络的低维度隐含表达。网络表达学习的目标是将一个复杂网络中的结点和边的信息都通过向量来表达，我们通过算法可以学习出这些向量，因此这些向量中包含着网络的结构信息。学习得到的向量可以作为图的特征用在基于图的各种任务上，比如连接预测，节点分类，社区发现以及可视化等问题上。

网络表达学习本质上是将网络进行降维从而更好得利用其它模型来处理网络数据。对于一个网络G=(V,E)来说，网络表达的结果是将这个网络变成维度为|V|xd的一个二维矩阵，|V|是指网络节点个数，d是每个节点表达的维度。网络表达学习相当于一个映射函数，将每个节点映射到一个向量中去。原本网络中单个节点包含的特征是复杂且难以把控的，节点的特征包括周边邻接点的信息（结构信息），节点的属性以及边的属性等等，如果我们要同时直接使用节点的所有特征是非常困难的，每个特征都有自己的含义，所以我们希望把这些特征都封装到一起成为一个向量，这样既方便读取也方便模型运算。

深度游走（DeepWalk）算法

深度游走算法是近年来第一个有影响力的大规模网络表达学习算法，它的本质是将随机游走（Random Walk）和自然语言处理中的skip-gram算法作组合所产生的算法。我们首先来介绍一下随机游走和skip-gram，然后再来看deepwalk是如何工作的。

随机游走（Random Walk）

随机游走是一个非常基础的基于网络的算法。它的本质就是从一个节点出发，随机选择它的一个邻接点，再从这个邻接点出发到下一个节点，重复这个步骤然后记录下所经过的所有节点。这个算法的变种在Google搜索和金融领域应用广泛。通过随

机游走我们可以得到从每个节点出发的一条路径，这条路径就代表了这个节点的结构信息。

Skip-gram算法

Skip-gram算法是自然语言处理中常用的一种方法，它是用一个词来预测这个词前后可能会出现哪些词。举个例子，对于I can do all things这句话，我们希望学出do这个词的表达，因此我们要定一个窗口，如图1所示，1就是我们要学的词，而uk就是窗口，代表我们要用哪些词来训练1的表达。若窗口大小定为3，则对于do这个词而言，训练数据为(do, can)，(do, all)，若窗口大小定为5，则对于do这个词而言，训练数据为(do, can)，(do, I)，(do all)，(do things)。

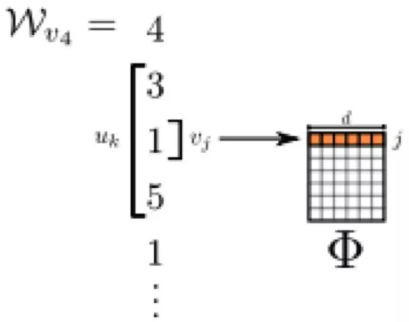


图1. skip-gram窗口

我们是通过神经网络来学出词的表达。网络的结构如图2。输入是每个词的one-hot向量，即，如果字典里一共有10000个词，那输入就是一个10000维的向量，只在这个词所处的位置上置1其余全部置0。输出也是训练集中对应词的one-hot向量，此时输入层和输出层之间隐含层的值就是我们学习出的词的表达。对于skip-gram模型，我们采用随机梯度下降进行参数学习。我们希望将这种方法用在学习网络节点表达上，接下来将描述如何将随机游走和skip-gram结合起来。

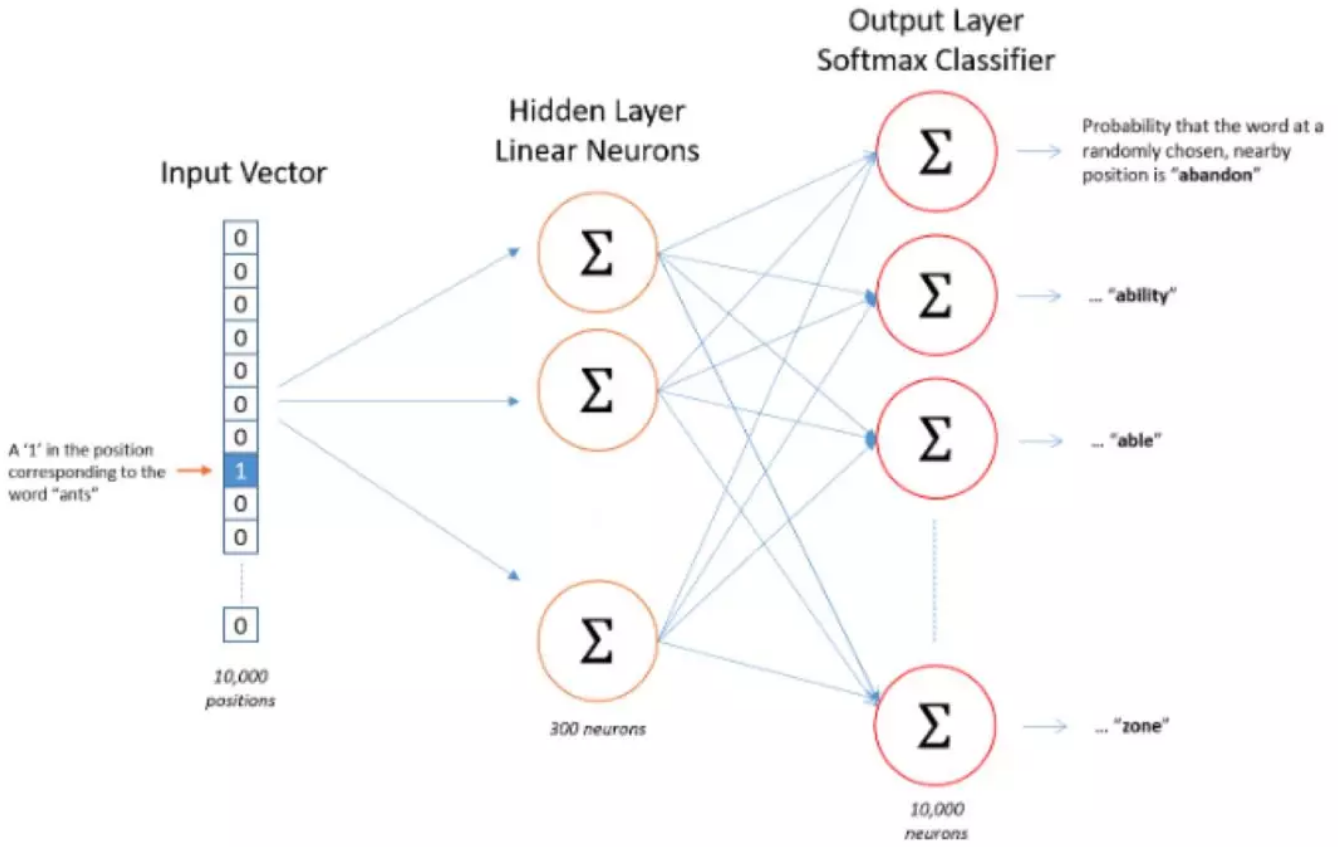
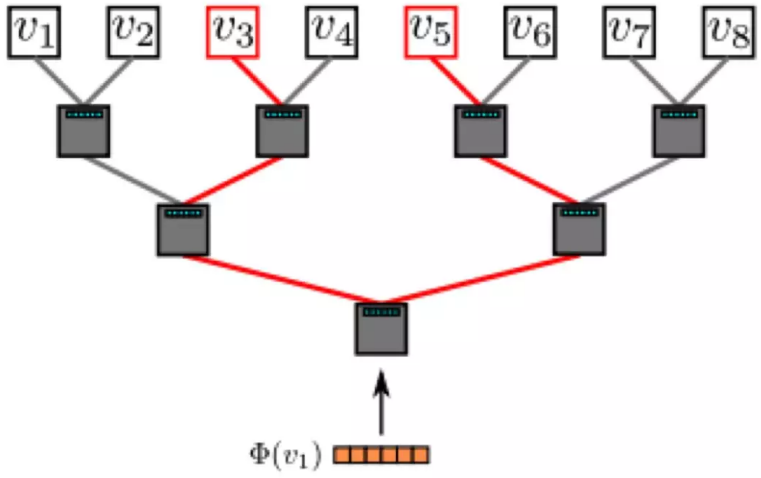


图2. skip-gram模型

深度游走算法

深度游走的核心思想总结成一句话就是，短的随机游走路径=句子（short random walk = sentence，quoted from Bryan Perozzi），因此我们只需要设定一个随机游走的步数r，通过随机游走我们就可以得到一个长度为r的路径（节点集），此时我们将其中每一个节点都看成单词，每个节点也都有对应的one-hot编码，这样就可以直接用skip-gram模型学出节点的表达。

由于网络的节点数目可能达到百万甚至千万级，节点的one-hot编码会过于稀疏，不同于传统的skip-gram模型直接用softmax函数得到输出，deepwalk采用的是层级softmax方法，如图3所示，每个节点对应一个完全二叉树的叶子节点，根节点输入的是我们目标节点的表达，此时就变成一个二分类问题，我们只需要判断二叉树的左右子树就可以学出节点的表达。



(c) Hierarchical Softmax.

图3. 层级softmax模型

总结

Deepwalk是一个非常简单但很有创意的方法，它将基于图的经典方法随机游走和自然语言处理中的skip-gram模型结合，得到了一个简单好用的网络表达学习方法。这也是第一篇将深度学习应用在大规模网络上，因此这个方法具有很强的可拓展性。Deepwalk作为网络表达学习中一个开创性工作，从一个简单的角度切入，用现有的成熟的方法，在一个全新的且尚未成为主流的问题中得到一个行之有效的解，不仅对网络表达这个问题带来极大的发展，同时也为我们做研究提供了一个很好的思路。一个优秀的方法是要建立在前人工作的基础上而不是凭空产生，并且当前最热门的问题并不一定是最值得研究的，冷门的问题也有其潜在的研究价值。

参考文献

[1] DeepWalk: Online Learning of Social Representations，KDD' 14

[2] Skip-gram. http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

阅读原文