DeepWalk：图网络与NLP的巧妙融合

原创 kaiyuan NewBeeNLP 2020-09-21

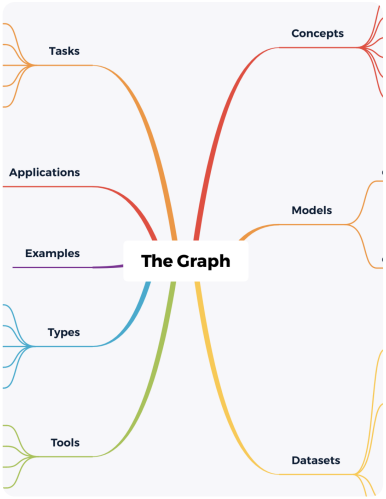
收录于话题
#图网络学习

5个

听说星标这个公众号⏫
模型效果越来越好嘞 🥳

作者 | kaiyuan
整理 | NewBeeNLP公众号

最近这段时间一直在做图网络相关，也差不多收尾了，有空整体复盘了下，大致以下几个主题，不过没整理完全哈哈（下次一定😜）



顺手再安利几份资料吧 👏

- 斯坦福的CS224W课程
- 清华大学唐杰老师的很多分享
- 清华大学 thunlp/GNNPapers
- 一些大佬们的新书：《Graph Representation Learning》、《Deep Learning on Graphs》
- 等等

ok，回到正题，今天要介绍的这篇是『Graph Embedding』系列第一篇，十分经典

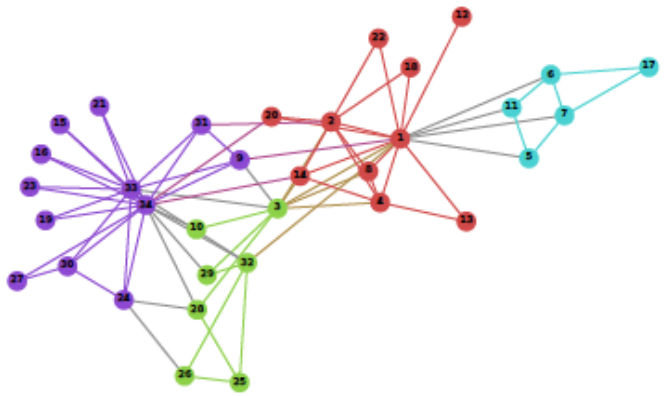
- 论文：DeepWalk: Online Learning of Social Representations[1]

- 代码：https://github.com/phanein/deepwalk

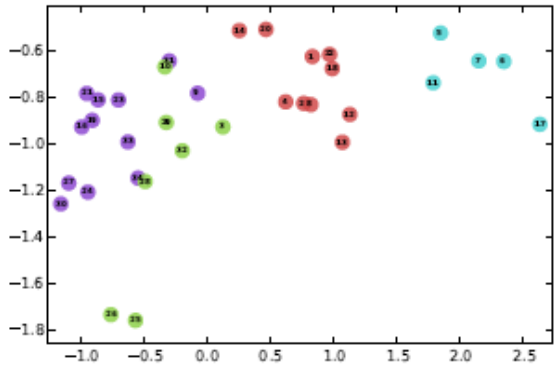
enjoy~

TL;DR

DeepWalk是首次将深度学习技术（无监督学习）引入到网络分析（network analysis）中的工作，它的输入是一个图，最终目标就是获得网络图中每个结点的向量表示 $\mathbf{X}_e \in \mathbb{R}^{|V| \times d}$。毕竟万物皆可向量，得到向量之后能做的事情就非常多了。如下所示是论文中给出的 Karate network 例子。



(a) Input: Karate Graph

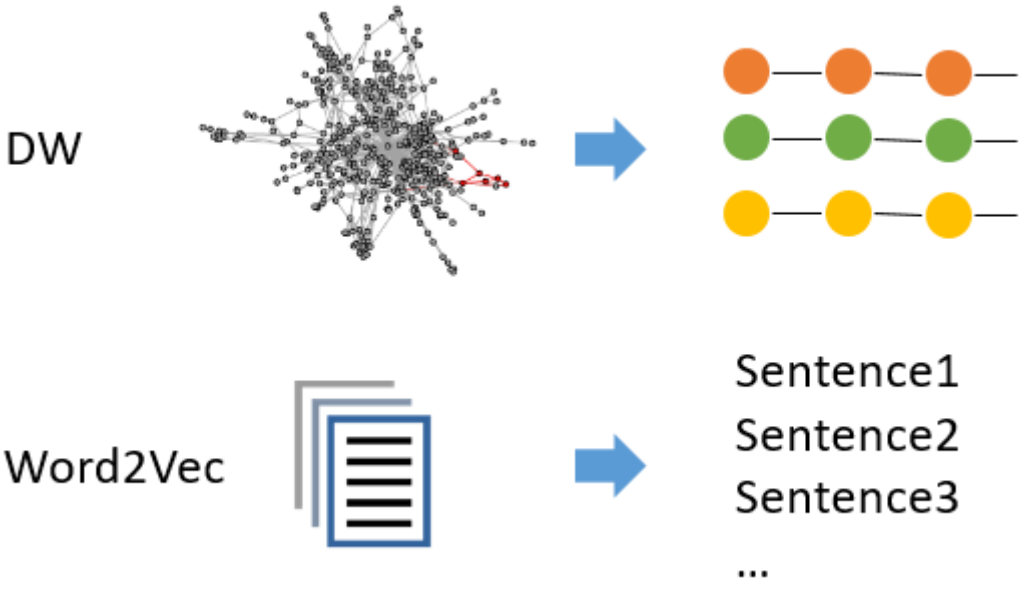


(b) Output: Representation

先验知识

说到生成向量表示，最有名的莫过于 Word2Vec 了，那么是不是可以将 network embedding 的问题转化为熟悉的 word embedding 形式呢？这样我们就可以借用word2vec的思想来解决了。

转化的方式就是 Random Walk ，通过这种方式将图结构表示为一个个序列，然后我们就可以把这些序列当成一个个句子，每个序列中的结点就是句子中的单词。



简单的说， DeepWalk = RandomWalk + SikpGram ， 下面我们来具体介绍下两种技术。

Random Walk

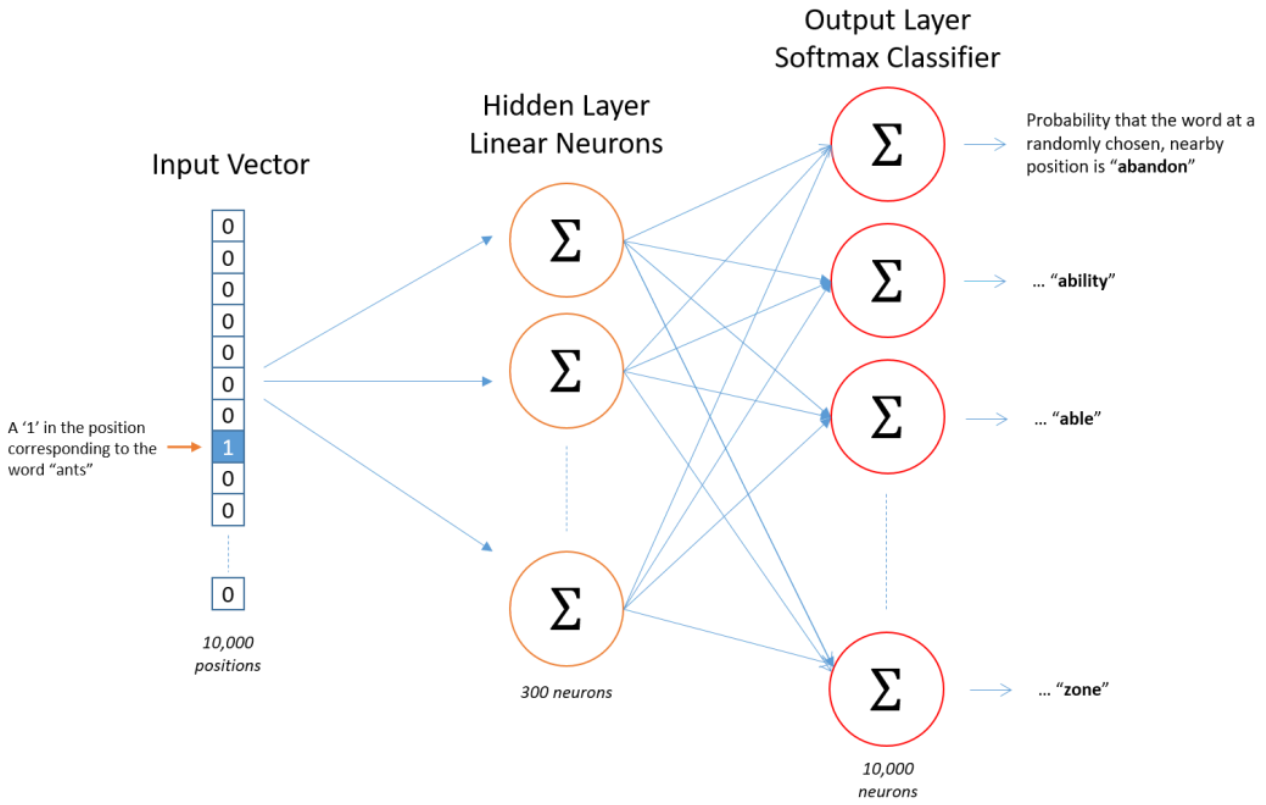
随机游走，顾名思义，就是从输入图中的任意一个结点 v_i 开始，随机选取与其邻接的下一个结点，直至达到给定长度 t， 生成的序列 $\tilde{\mathcal{W}}_{v_i} = \left(\mathcal{W}_{v_i}^1, \cdots, \mathcal{W}_{v_i}^k, \cdots, \mathcal{W}_{v_i}^t\right)$ 。在论文中，对于每一个顶点 v_i， 都会随机游走出 γ 条序列。

采用随机游走有两个好处：

- **「利于并行化」**：随机游走可以同时从不同的顶点开始采样，加快整个大图的处理速度；
- **「较强适应性」**：可以适应网络局部的变化；

Skip Gram

word2vec的skip-gram相信大家都非常熟悉了，这里就不再赘述，放一张图。



DeepWalk

结合上面两点， deepwalk其实就是首先利用random walk来表示图结构，然后利用skip-gram模型来更新学习节点表示。算法流程如下所示：

Algorithm 1 DeepWalk(G, w, d, γ, t)

Input: graph $G(V, E)$
window size w
embedding size d
walks per vertex γ
walk length t
Output: matrix of vertex representations $\Phi \in \mathbb{R}^{|V| \times d}$

Initialization: Sample Φ from $\mathcal{U}^{|V| \times d}$
Build a binary Tree T from V
for $i = 0$ to γ **do**
  $\mathcal{O} = \text{Shuffle}(V)$
  for each $v_i \in \mathcal{O}$ **do**
    $\mathcal{W}_{v_i} = RandomWalk(G, v_i, t)$
    SkipGram(Φ, $\mathcal{W}_{v_i}$, w)
  end for
end for

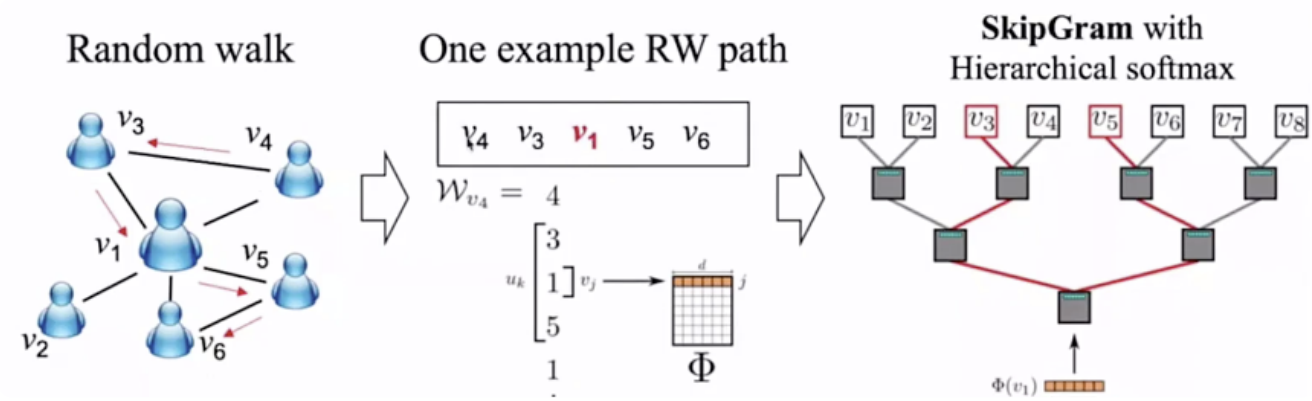
算法有两层循环，第一层循环采样 γ 条路径，第二层循环遍历图中的所有结点随机采样一条路径并利用skip-gram模型进行参数更新。

其中第2步构建二叉树的目的是为了方便后续 SkipGram模型的层次softmax算法。

参数更新的流程如下：

Algorithm 2 SkipGram(Φ, $\mathcal{W}_{v_i}$, w)

```
for each v_j ∈ W_{v_i} do
    for each u_k ∈ W_{v_i}[j − w : j + w] do
        J(Φ) = − log Pr(u_k | Φ(v_j))
        Φ = Φ − α ∗ ∂J/∂Φ
    end for
end for
```



一个小结

deepwalk可以说给网络学习方向打开了一个新思路，有很多优点：

- 支持数据稀疏场景
- 支持大规模场景（并行化）

但是仍然存在许多不足：

- 游走是完全随机的，但其实是不符合真实的社交网络的；
- 未考虑有向图、带权图

本文参考资料

[1] **DeepWalk: Online Learning of Social Representations:** https://arxiv.org/abs/1403.6652

- END -