Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

域自适应网络嵌入-DANE

硕士研究生 吴杭颐

2021年02月28日

内容提要

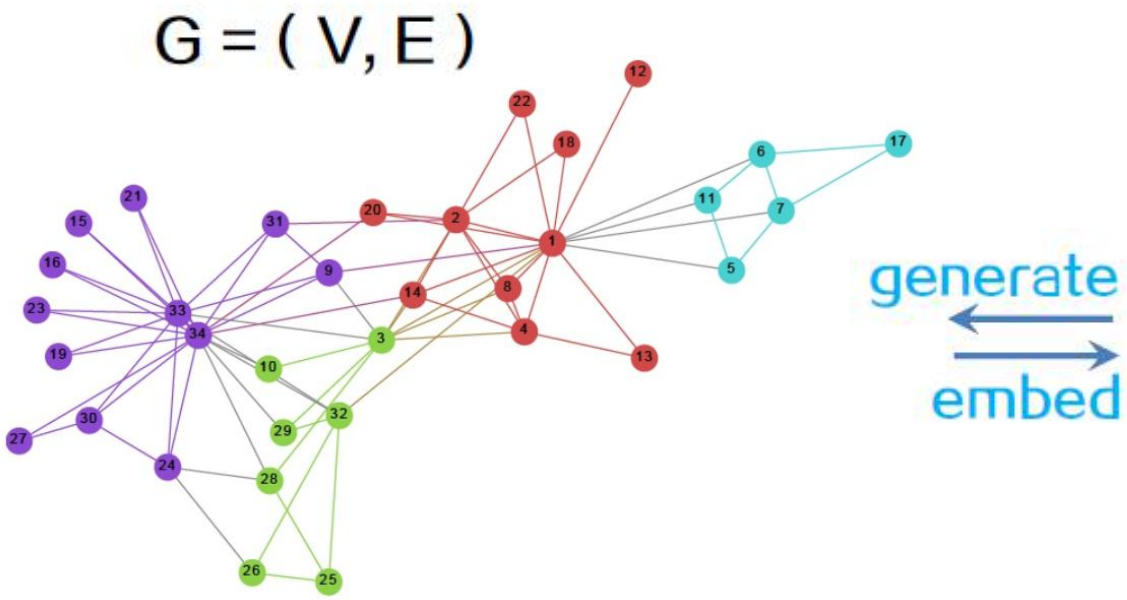
- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1.理解图嵌入的意义与挑战
  - 2.理解DANE的算法原理
  - 3.了解DANE的应用

背景简介

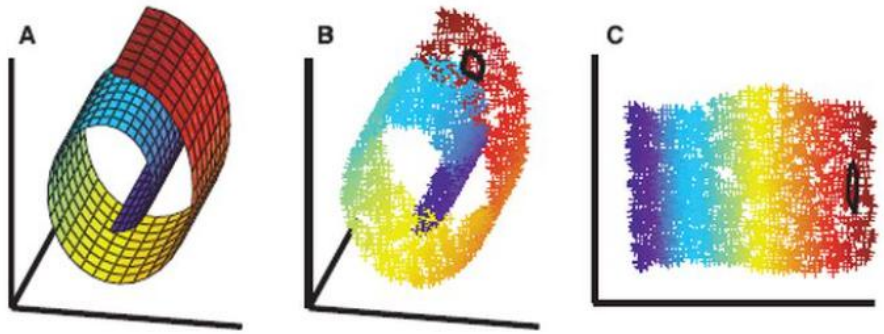
- 图嵌入 or 网络嵌入 or 网络表征学习：
  - 是一种将图数据（通常为高维稀疏的矩阵）映射为低维稠密向量的过程；
  - 需要捕捉图的拓扑结构，顶点与顶点的关系，以及其他信息（如子图，连边等）。



背景简介

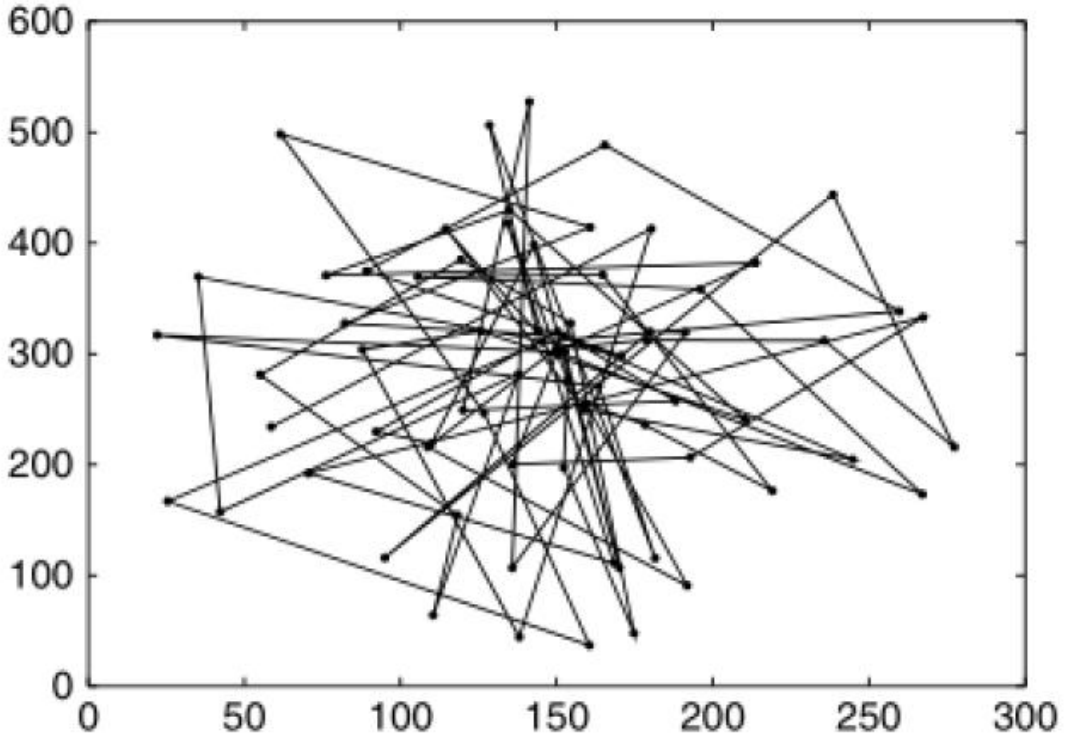
- 图嵌入的意义：
  - 机器学习在图上的应用是有限的；
  - 嵌入是压缩的表示，能够将节点属性打包到一个维度更小的向量中；
  - 向量运算比图上的运算更简单和快捷。

- 图嵌入的要求/挑战：
  - 属性选择：确保嵌入能很好地描述图的属性，使得后期预测或可视化获得较好表现；
  - 可扩展性：嵌入方法应具有可扩展性，能够处理大型的图；
  - 嵌入的维度：根据需求权衡决定嵌入维数。



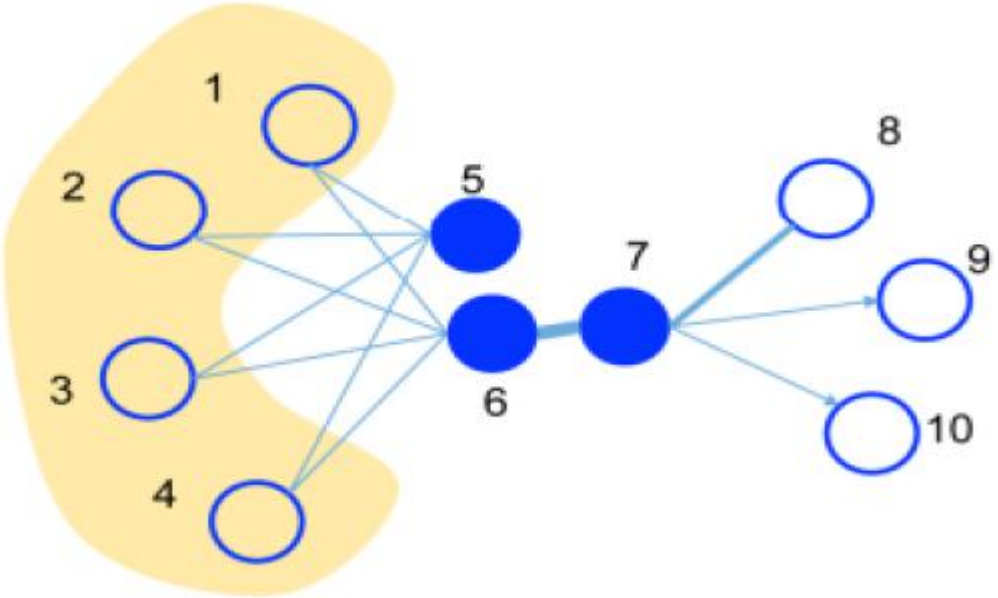
背景简介

- 随机游走
  - 基于过去的表现，无法预测将来的发展步骤和方向。



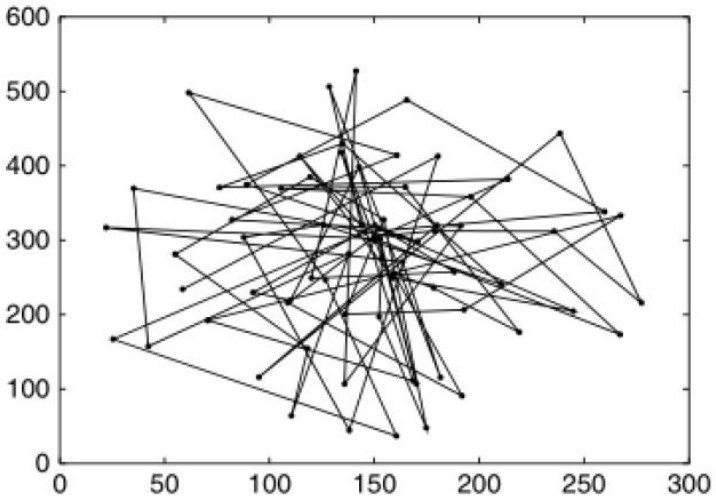
背景简介

- 一阶相似度
  - 用于描述图中成对顶点之间的局部相似度，形象化描述为若v_i，v_j之间存在直连边，则边权$v_{i,j}$即为两个顶点的相似度，若不存在直连边，则1阶相似度为0。
- 二阶相似度
  - 描述一对节点的邻域结构的接近程度。即两个节点的共同上下文节点越多，这两个节点越相似。

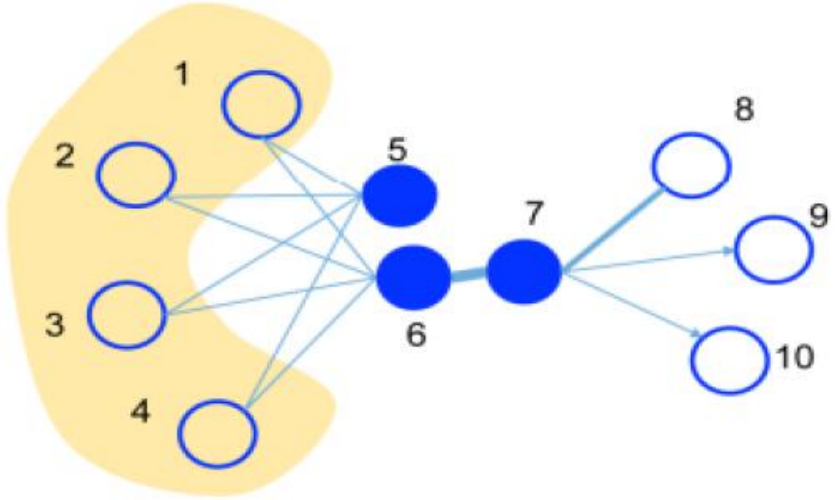


背景简介

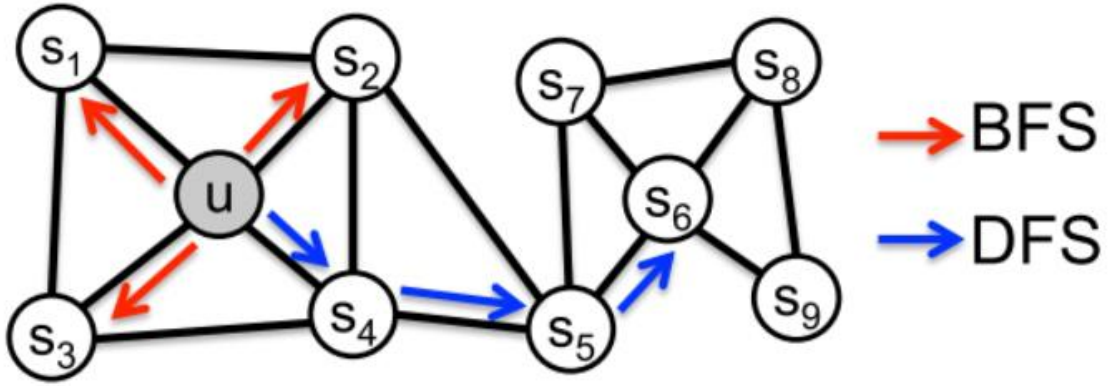
- 网络嵌入算法
  - Deepwalk（2014）
  - Line（2015）
  - Node2vec（2016）



Deepwalk



Line



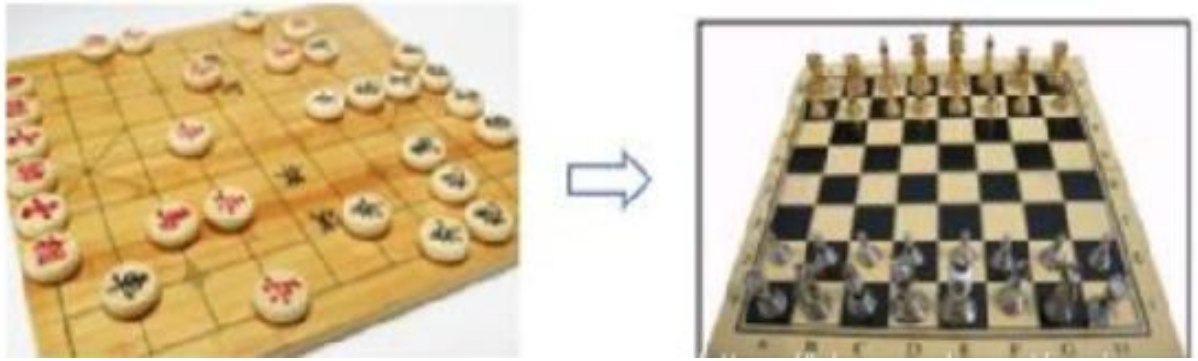
Node2vec

基本概念

基本概念

- 迁移学习
  - 利用数据、任务或模型之间的相似性，将在旧领域学习过的模型，应用于新领域的一种学习过程。

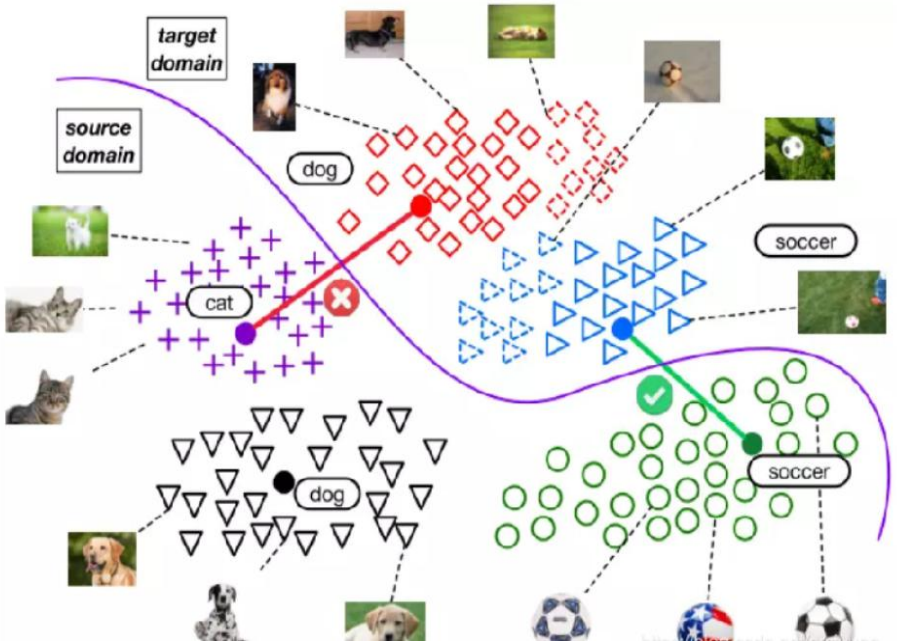
- 领域是学习的主体，由数据特征和特征分布组成。
  - 源领域：有知识、有大量数据标注的领域，要迁移的对象；
  - 目标领域：要赋予知识、赋予标注的对象；
  - 知识从源领域传递到目标领域，就完成了迁移。



基本概念

- 域自适应：
  - 迁移学习中的一种代表性方法；
  - 利用信息丰富的源域样本来提升目标域模型的性能。

- 域自适应学习：
  - 能有效地解决训练样本和测试样本概率分布不一致的问题。



基本概念

- 域自适应网络嵌入：
  - 使模型能够在其他网络上重用，降低了训练下游机器学习模型的成本；
  - 将在标记网络上训练好的下游模型转移到未标记网络；
  - 以无监督方式学习表示，从而实现双向域适配。

- 两个挑战：
  - 嵌入空间对齐，来自不同网络的结构相似的节点应在嵌入空间中具有相似的表示；
  - 分布对齐，嵌入向量的分布偏移会影响模型在目标网络上的性能。

算法原理

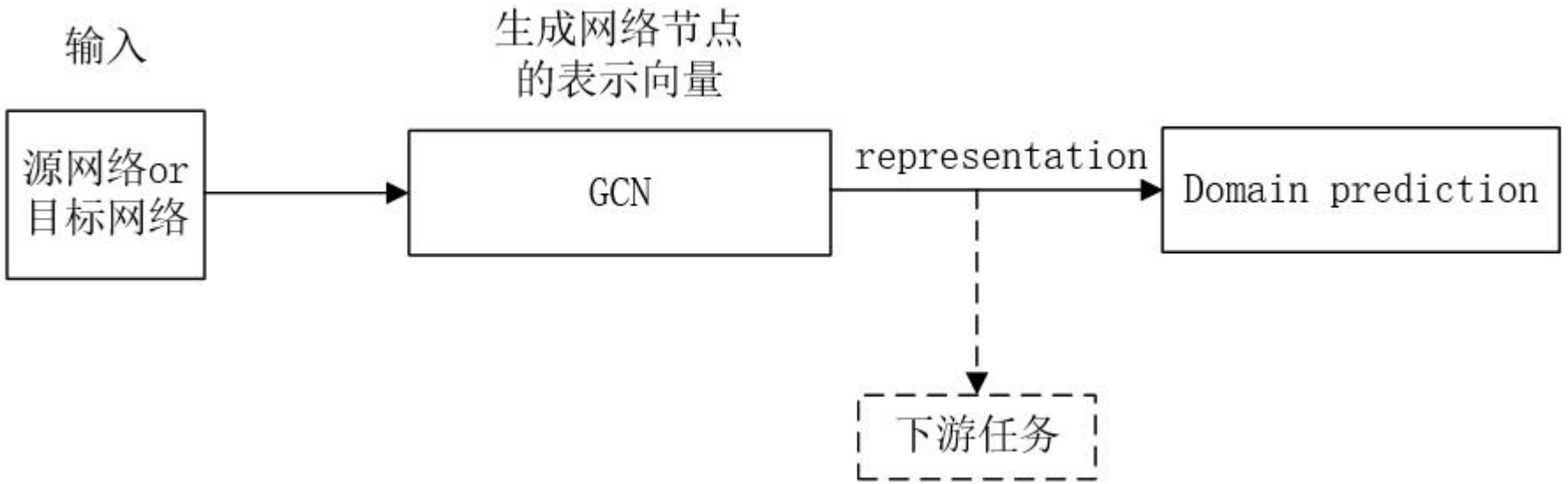
算法原理

T	支持不同网络上下游模型传输的网络嵌入算法
I	来自多个网络的节点
P	1.节点通过共享的可学习参数集被编码成向量； 2.通过对抗学习正则化进一步调整嵌入在不同网络上的分布。
O	共享对齐嵌入空间的向量

P	解决嵌入空间偏移和嵌入分布偏移问题
C	网络嵌入方法在支持域自适应任务上的性能和跨网络传送下游机器学习模型处理相同任务时的准确性
D	提出无监督的网络嵌入框架
L	IGCAI 2019

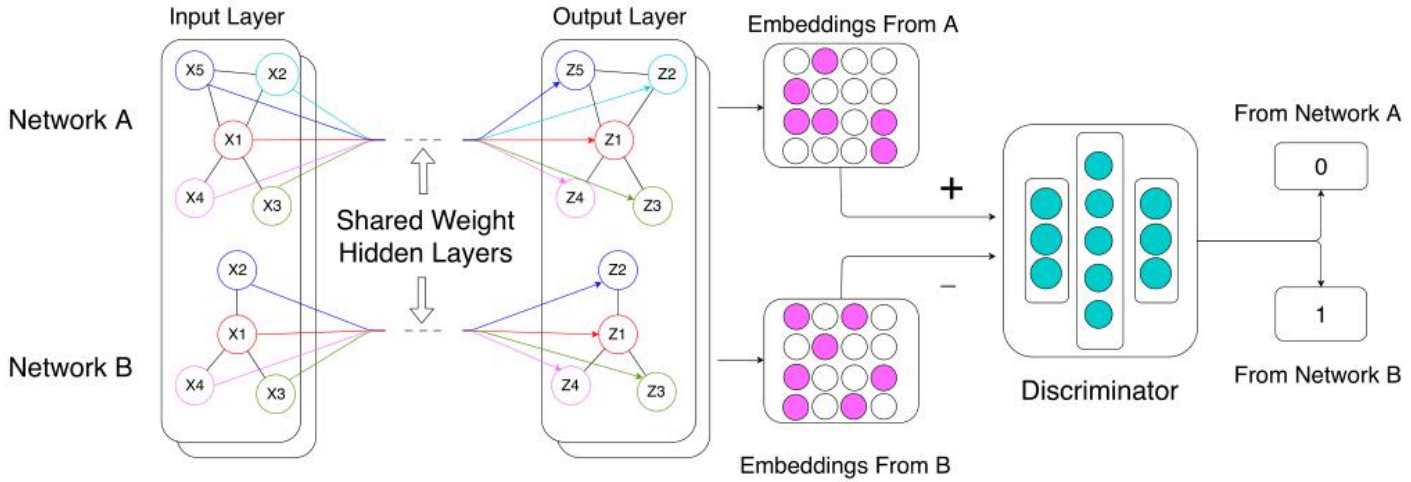
算法原理

- 核心思想：
  - 提出一个无监督的网络嵌入框架DANE，可以通过GCN和adversarial learning解决嵌入空间偏移和嵌入分布偏移的问题；
  - 不是所有的网络对都可以进行双向的域自适应，本文主要处理边为同质的，点的特征具有相似意义的网络。



算法原理

- 总体框架：
  - 共享权重的GCN网络：网络中节点通过 Shared Weight Graph Convolution Network 被编码为向量，来取得图的嵌入空间平行，保存交叉网络节点对之间的结构相似性；
  - 对抗学习模块：使用了Adversarial Learning Regularization来保证不同网络中嵌入向量分布的平行。



算法原理

- 共享权重的GCN网络：
  - 两个图网络对应源域和目标域，完全使用GCN网络，得到两个域的节点嵌入；
  - 在GCN中，每一层表示为：

$$H^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W_l)$$

  - multi-task loss function：

$$L_{gcn} = L_{G_{src}} + L_{G_{tgt}}$$

  - first-order loss function：

$$L_G = -\sum_{(i,j)\in E} \log\sigma(v_j \cdot v_i) - Q \cdot \mathbb{E}_{k\sim P_{neg}(N)} \log\sigma(-v_i \cdot v_k)$$

算法原理

- 对抗学习正则化：
  - 同GAN类似，以第一部分的共享权重的图卷积网络为生成器，设计一个判别器分辨嵌入向量是来自哪个域。

$$D(x) = \begin{cases} 0 & x \in V_{src} \\ 1 & x \in V_{tgt} \end{cases}$$

  - 判别器的损失函数

$$L_D = \mathbb{E}_{x \in V_{src}}[(D(x) - 0)^2] + \mathbb{E}_{x \in V_{tgt}}[(D(x) - 1)^2]$$

  - 对抗训练的损失函数

$$L_{adv} = \mathbb{E}_{x \in V_{src}}[(D(x) - 1)^2] + \mathbb{E}_{x \in V_{tgt}}[(D(x) - 0)^2]$$

  - 整体的损失函数

$$L = L_{gcn} + \lambda L_{adv}$$

算法原理

- 数据集：
  - Paper Citation Networks： 包含两个网络A和B，每个节点是一篇文章，标签为所属领域，节点特征为由摘要构成的词频向量；
  - Co-author Networks： 包含两个网络A和B，每个节点是一个作者，每个作者都有一个或多个所属研究领域的标签，节点特征为由作者论文关键词构成的词频向量。

Network Name	Nodes	Edges
Paper Citation A	2277	8245
Paper Citation B	3121	7519
Co-author A	1500	10184
Co-author B	1500	10606

Table : Network size of two datasets.

算法原理

- 实验目的：
  - 比较DANE相比Baseline算法的提升效果；
  - 对比分析对抗学习正规化的重要性。
- Baselines：
  - DeepWalk
  - LINE
  - Node2vec
  - unsupervised GraphSAGE(GCN version)

算法原理

- **节点分类：**

Methods	Paper Citation Network (Single-label)				Co-author Network (Multi-label)			
	A→B		B→A		A→B		B→A	
	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
DeepWalk	0.282	0.381	0.22	0.32	0.517	0.646	0.502	0.620
LINE	0.156	0.214	0.175	0.272	0.525	0.634	0.506	0.601
Node2vec	0.147	0.196	0.248	0.32	0.513	0.632	0.520	0.627
GraphSAGE Unsup	0.671	0.703	**0.861**	0.853	0.724	0.809	0.741	0.832
DANE	**0.797**	**0.803**	0.852	**0.872**	**0.785**	**0.847**	**0.776**	**0.849**

Table 2: Micro and macro F1 score of different network embedding methods in unsupervised domain adaptation

算法原理

- 实验结果：

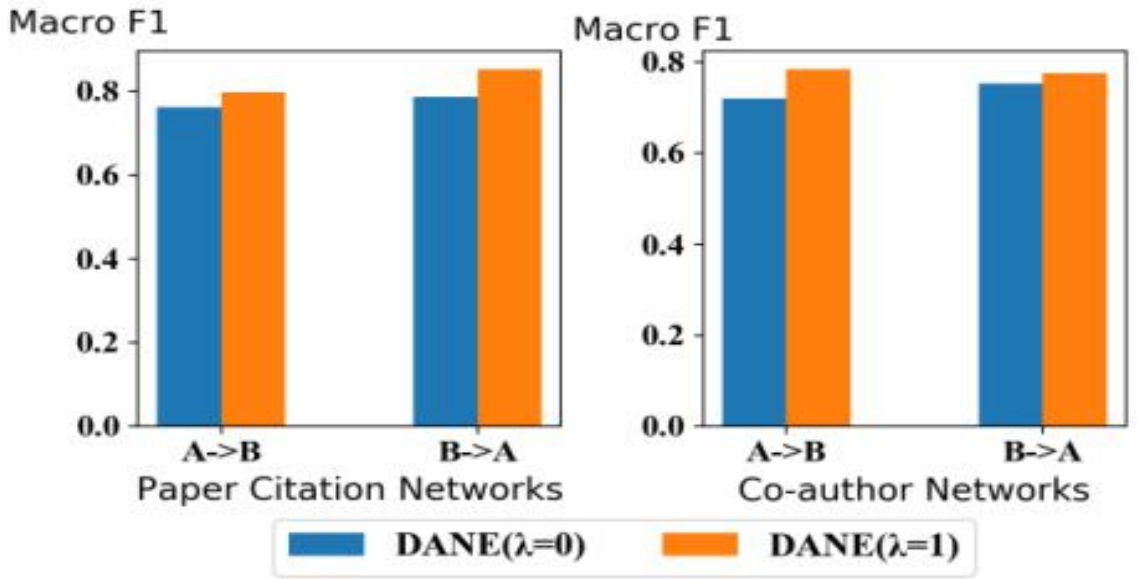
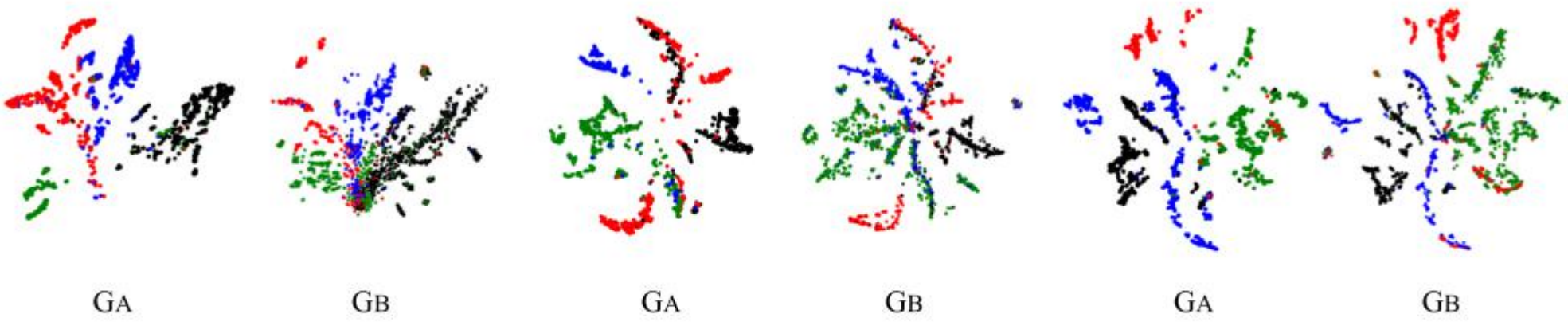


Figure 2: Node classification performance of DANE ($\lambda = 0$) and DANE ($\lambda = 1$)

算法原理

- 可视化：



(a) Randomly Initialized GCN (b) DANE (λ=0) (c) DANE (λ=1)

Figure 4: Visualization of Paper Citation Networks generated by a randomly initialized GCN, DANE ($\lambda = 0$) and DANE ($\lambda = 1$).

优劣分析

优劣分析

- 横向对比
  - DANE是一种支持不同网络上下游模型传输的网络嵌入框架。

- 纵向对比
  - DANE的表现优于其他网络嵌入算法；
  - 是最早提出新的基于多图的图卷积网络去学习可迁移的embedding的团队。

应用总结

- 应用领域
  - 节点分类
  - 可视化任务
  - …
- 未来的发展
  - 推广DANE来解决异构网络上的域自适应；
  - 拓展到更多场景，改进DANE框架以解决半监督和监督域自适应问题。

参考资料

[1] Yizhou Zhang, Guojie Song, Lun Du, Shuwen Yang, Yilun Jin. DANE: Domain Adaptive Network Embedding [C]. IJCAI 2019: 4362-4368.
[2] Zhang Z, Cui P, Zhu W. Deep learning on graphs: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2020.
[3] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020.
[4] Ganin, Yaroslav, et al. Domain-adversarial training of neural networks[J]. *The Journal of Machine Learning Research* 17.1 (2016): 2096-2030.
[5] https://zhuanlan.zhihu.com/p/62629465.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。