【动手做】用DGL库和GraphSAGE完成链接预测

爱科技勤折腾的苹果二 2023-11-16

讨论内容

1. 工具准备
- Python和Anaconda
- DGL-Deep Graph Library
2. 知识准备
- 链接预测
- GraphSAGE的原理
3. 案例演示
- 代码运行流程
- 代码演示
4. 参考资料

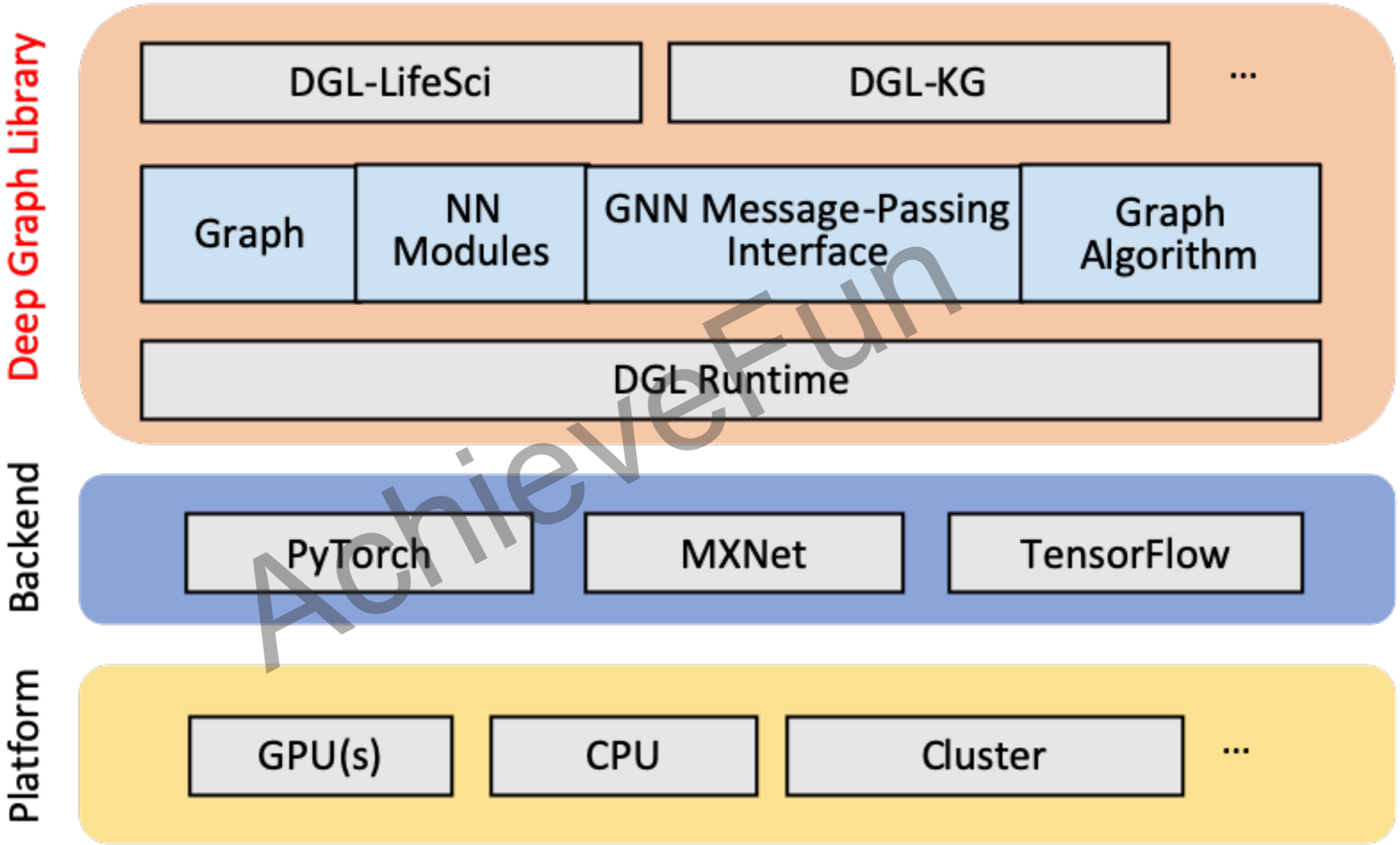
1 工具准备：Python和Anaconda

- https://github.com/bettermorn/IAICourse/wiki/Python%E5%AD%A6%E4%B9%A0%E5%8F%82%E8%80%83%E8%B5%84%E6%96%99 Python学习参考资料
- https://github.com/bettermorn/IAICourse/wiki/%E5%AD%A6%E4%B9%A0%E4%B9%8B%E5%89%8D%E7%9A%84%E5%87%86%E5%A4%87%E5%B7%A5%E4%BD%9C%E5%92%8CAnaconda%E5%B8%B8%E8%A7%81%E9%97%AE%E9%A2%98%E8%A7%A3%E5%86%B3%E6%96%B9%E6%B3%95 学习之前的准备工作和Anaconda常见问题解决方法，例如：使用pip 下载依赖包安装较慢，如何解决?
- https://github.com/bettermorn/IAICourse/wiki/%E3%80%90%E5%8F%82%E8%80%83%E4%BE%8B%E5%AD%90%E3%80%91%E5%AE%89%E8%A3%85%E8%AE%AD%E7%BB%83%E7%8E%AF%E5%A2%83%E5%92%8C%E8%AE%B0%E5%BD%95%E6%B5%8B%E8%AF%95%E7%BB%93%E6%9E%9C 【参考例子】安装训练环境和记录测试结果

1 工具准备：DGL-Deep Graph Library

- 正式网站 https://www.dgl.ai/
- GitHub https://github.com/dmlc/dgl/
- Overview of Deep Graph Library (DGL)
- 开始 https://www.dgl.ai/pages/start.html
- Deep Graph Library（DGL）是一个Python软件包，用于在现有DL框架（当前支持PyTorch，MXNet和TensorFlow）之上轻松实现图神经网络模型系列。 它提供消息传递的通用控制，通过自动分批处理和高度可调的稀疏矩阵内核进行速度优化以及多GPU / CPU训练，以缩放到数亿个节点和边缘的图形。

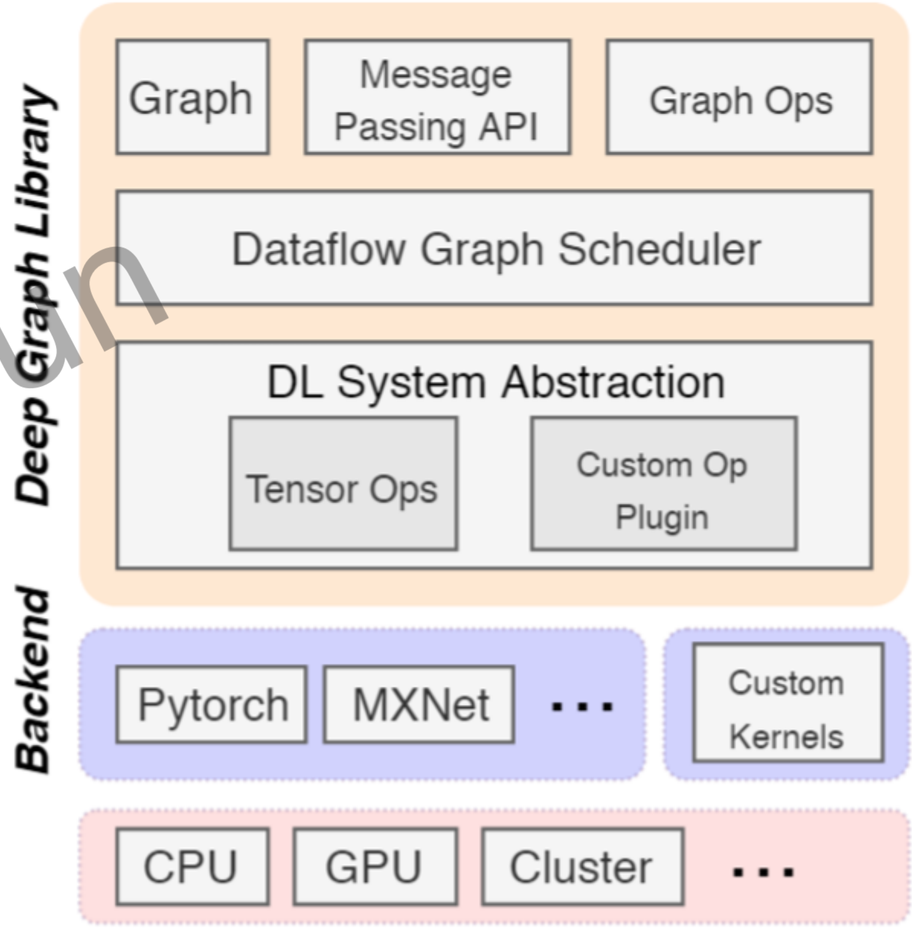
DGL架构



图片来源：Overview of Deep Graph Library (DGL)

DGL meta-objectives & architecture

- Forward and backward compatible
  - ***Forward***: easy to develop new models
  - ***Backward***: seamless integration with existing frameworks (MXNet/Pytorch/Tensorflow)
- **Fast** and **Scalable**

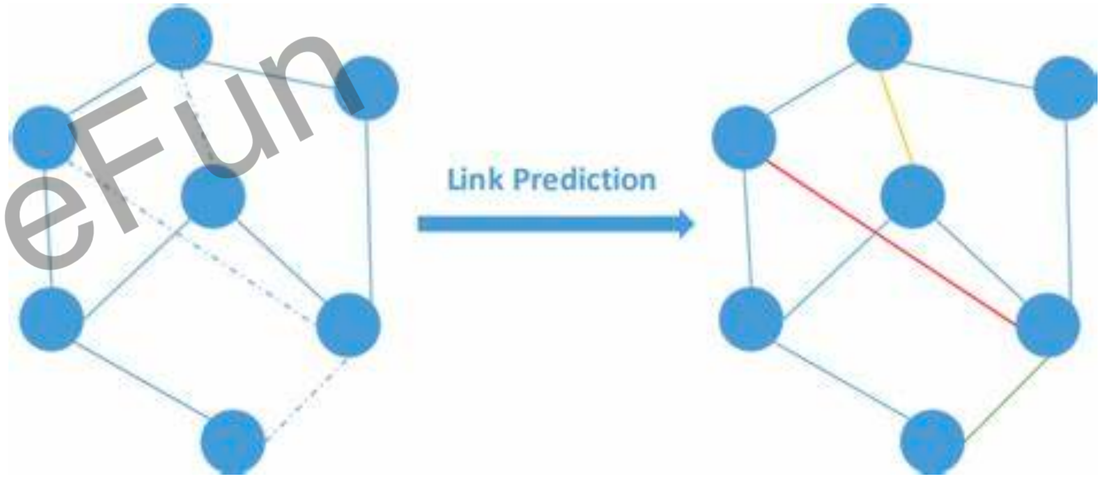


aws

图片来源：Overview of Deep Graph Library (DGL)

2 知识准备：链接预测

- 基于已有的知识图谱，预测和推断出实体之间可能存在的关系，或者预测可能存在的实体和关系。预测两个特定节点之间是否存在边。
- 有助于丰富和完善知识图谱的内容，使其更加全面和精确。
- 应用于：推荐系统、社交网络分析、生物信息学、知识图谱补全



图片来源：https://www.researchgate.net/publication/350110330/figure/fig1/AS:1002329977393155@1615985497381/An-example-of-link-prediction.ppm

2 知识准备：GraphSAGE的原理

- 对节点的周围邻居节点进行随机采样，使得图网络模型可以应用到大规模图谱上。
- 基于采样（**SA**mple）和聚集（Aggre**G**at**E**）操作
- 对红色节点采样到其周围的蓝色节点，并进行特征聚合。使用更通用的聚合函数AGG（Aggregation）：任意地可将一组向量聚合成一个向量的可微分方程。
- paperswithcode：https://paperswithcode.com/method/graphsage
- https://snap.stanford.edu/graphsage/

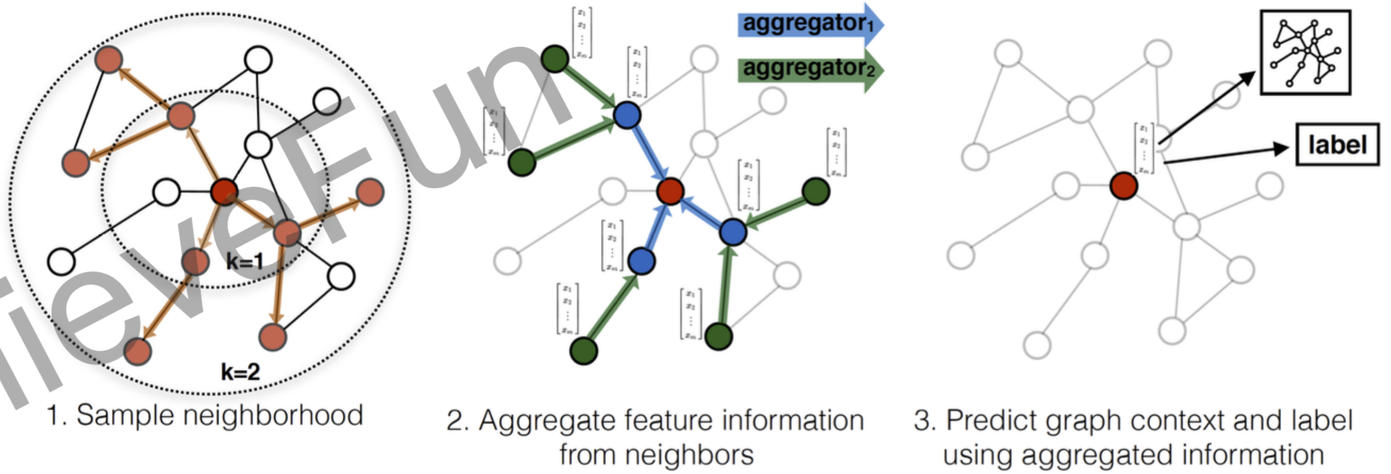


Figure 1: Visual illustration of the GraphSAGE sample and aggregate approach.

图片来源：https://arxiv.org/pdf/1706.02216.pdf

2 知识准备：GraphSAGE的聚合函数

基本邻域聚合

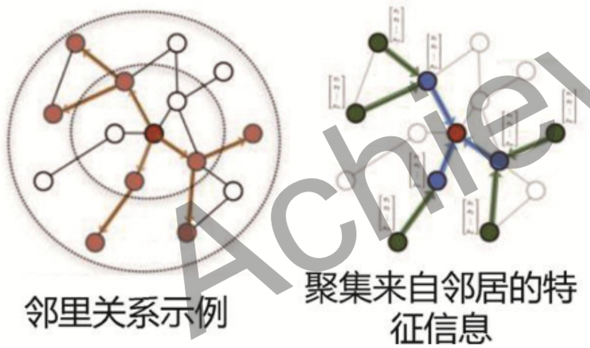
$$\boldsymbol{h}_v^k = \sigma\left(\boldsymbol{W}_k \sum_{u \in N(v)} \frac{\boldsymbol{h}_u^{k-1}}{|N(v)|} + \boldsymbol{B}_k \boldsymbol{h}_v^{k-1}\right)$$

连接自嵌入和邻域嵌入

$$\boldsymbol{h}_v^k = \sigma\left(\left[\boldsymbol{W}_k \cdot \mathrm{AGG}\left(\left\{\boldsymbol{h}_u^{k-1}, \forall u \in N(v)\right\}\right), B_k \boldsymbol{h}_v^{k-1}\right]\right)$$

广义聚合

任意可将一组向量合成一个向量的可微分方程都可应用



邻里关系示例

聚集来自邻居的特征信息

求均值 $\mathrm{AGG} = \sum_{u \in N(v)} \frac{\boldsymbol{h}_u^{k-1}}{|N(v)|}$

元素平均值/最大值

池化操作 $\mathrm{AGG} = \gamma\left(\left\{\boldsymbol{Q}\boldsymbol{h}_u^{k-1}, \forall u \in N(v)\right\}\right)$

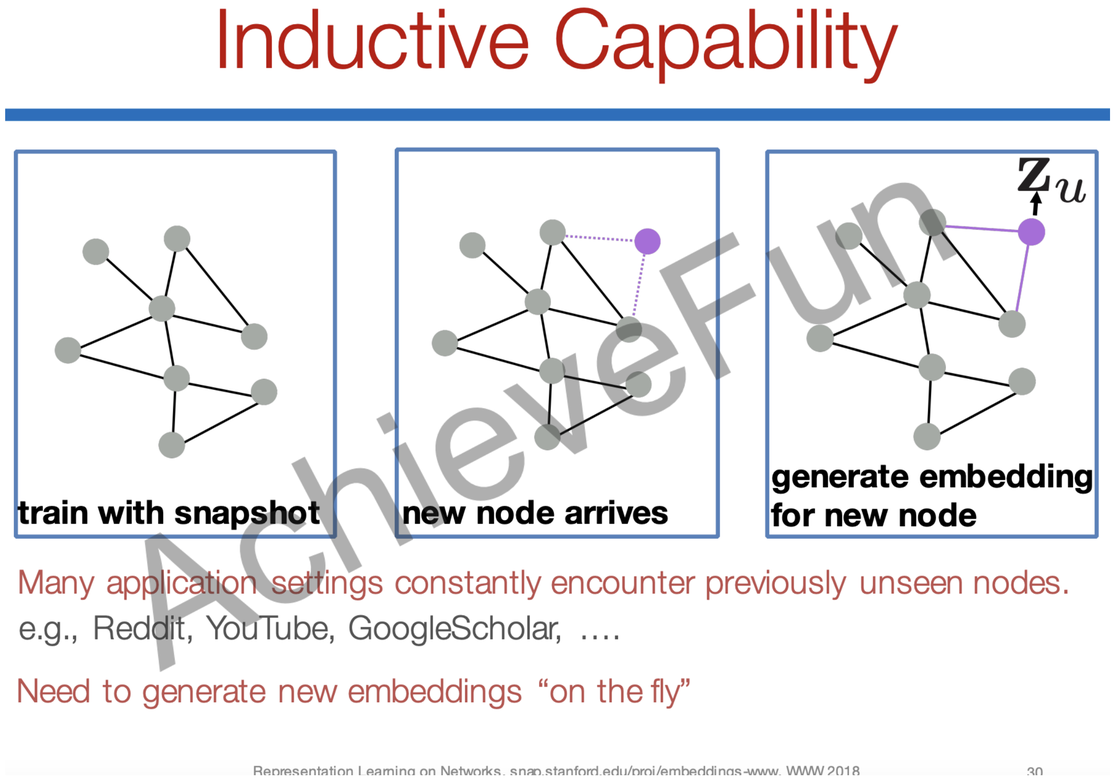
LSTM $\mathrm{AGG} = LSTM\left(\left[\boldsymbol{h}_u^{k-1}, \forall u \in \pi(N(v))\right]\right)$

图8-30 GraphSAGE：对邻居节点随机采样

图片来源：陈华钧. 知识图谱导论 (Chinese Edition) (p. 303). Kindle Edition

- 使用三种类型的聚合函数，即：求均值（Mean）、池化操作以及使用LSTM网络。

2 知识准备：GraphSAGE的归纳能力



图片来源：https://snap.stanford.ed/proj/embeddings-www/files/nrltutorial-part2-gnns.pdf

模型可以泛化到一些新的、未出现在训练集中的节点。

3 案例演示

- 案例来源：https://docs.dgl.ai/tutorials/blitz/4_link_predict.html
- 代码：
  - https://docs.dgl.ai/_downloads/ad3a2962e53be21909260c39376a994a/4_link_predict.py
  - https://docs.dgl.ai/_downloads/514d96075aeea53fc7a3d9873b7b963f/4_link_predict.ipynb

使用GNN进行链接预测的方法

- 例子：预测引用网络中两篇论文之间是否存在引用关系（引用或被引用）。
- 将链接预测问题表述为如下二元分类问题：

  a. 将图中的边视为正例。

  b. 抽取一些不存在的边（即节点对之间没有边）作为负例。

  c. 将正例和负例分为训练集和测试集。

  d. 用任何二元分类指标（如曲线下面积 (AUC)）评估模型。
- 这种做法来自 `SEAL `https://papers.nips.cc/paper/2018/file/53f0d7c537d99b3824f0f99d62ea2428-Paper.pdf，尽管这里的模型没有使用他们的节点标记思想。
- 在某些领域，如大规模推荐系统或信息检索，偏爱强调 top-K 预测的良好性能的指标。在这种情况下，可能需要考虑平均精度等其他指标，并使用其他负抽样方法。

环境准备

1. 安装Anaconda https://www.anaconda.com/download/
2. 在Anaconda建立环境dgl，根据dgl文档(https://www.dgl.ai/pages/start.html)确定**Python版本**
3. 根据代码，使用pip install -r requirements.txt 安装依赖库并验证库的存在pip list
    - matplotlib
    - torch
    - scikit-learn
4. 安装dgl https://www.dgl.ai/pages/start.html

例如：conda install -c dglteam dgl

代码运行流程

1. 加载图形和特征Loading graph and features
2. 准备训练集和测试集 Prepare training and testing sets https://graphsandnetworks.com/the-cora-dataset,https://github.com/topics/cora-dataset
3. 选择GraphSAGE 模型 Define a GraphSAGE model
4. 正图、负图和应用网格Positive graph, negative graph, and apply_edges
5. 训练模型

4 参考资料

- https://docs.dgl.ai/guide_cn/training-link.html
- https://docs.dgl.ai/tutorials/blitz/4_link_predict.html
- https://docs.dgl.ai/_downloads/ad3a2962e53be21909260c39376a994a/4_link_predict.py
- https://docs.dgl.ai/_downloads/514d96075aeea53fc7a3d9873b7b963f/4_link_predict.ipynb
- https://arxiv.org/pdf/1706.02216.pdf