

## 解决的问题

对于学习图上顶点的embedding，现有的方法多为直推式学习，学习目标是直接生成当前顶点的embedding，不能泛化到未知顶点上

## 做法及创新

论文提出一种归纳式学习方法GrdaphSAGE，不为每个顶点学习单独的embedding，而是学习一种聚合函数AGGREGATE，从一个顶点的局部邻域聚合特征信息，为未知的顶点直接生成embedding，因此旧的顶点只要邻域发生变化也能得到一个新的embedding

GCN不是归纳式，因为每次迭代会用到整个图的邻接矩阵A；而GraphSAGE可以对GCN做了精简，每次迭代只抽样取直接相连的邻居

## 算法流程

1. 给定顶点 $v$ 及其特征 $x_v$ ，作为它的初始表示 $h_v^0 = x_v$ 。
2. 计算邻域向量 $h_{N(v)}^k = \text{AGGREGATE}(h_u^{(k-1)}, \forall u \in N(v))$ ，当前层顶点的邻居从上一层采样，且邻居个数固定，非所有邻居，这样每个顶点和采样后邻居的个数都相同，可以直接拼成一个batch送到GPU中进行批训练
3. 将邻域向量与自身上一层的表示拼接，通过非线性激活函数 $\sigma$ 后作为这一层的表示 $h_v^k = \sigma(W^k \text{CONCAT}(h_v^{(k-1)}, h_{N(v)}^k))$
4. 标准化  $h_v^k = h_v^k / \|h_v^k\|_2$

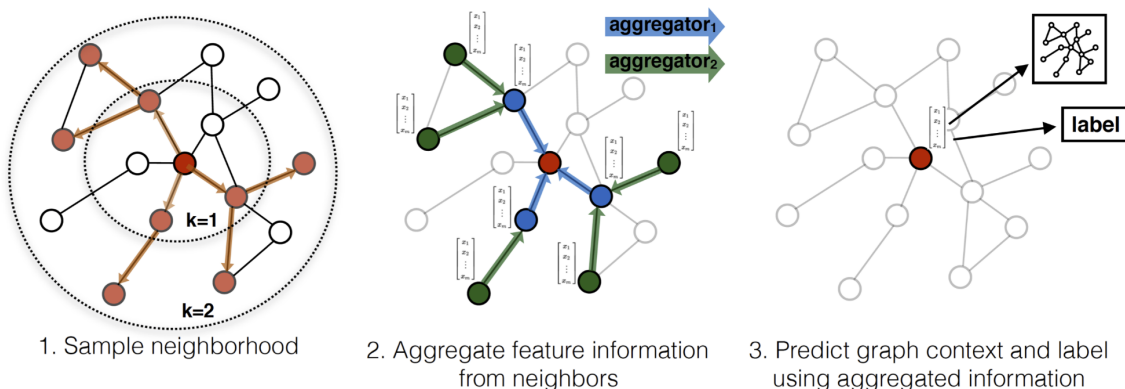
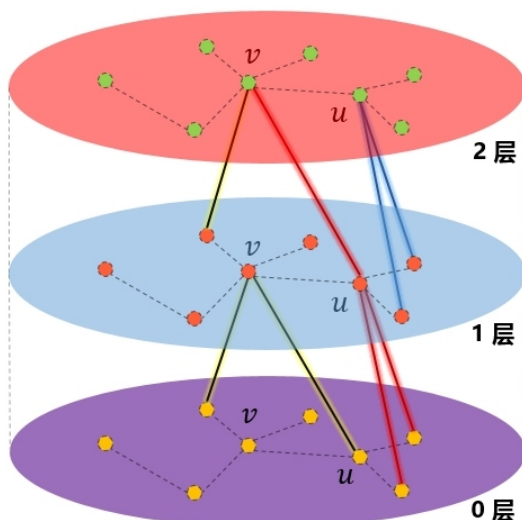


Figure 1: Visual illustration of the GraphSAGE sample and aggregate approach.



## 聚合函数

### 1. MEAN

$$h_v^k = \sigma(W \cdot \text{MEAN}(\{h_v^{k-1}\} \cup \{h_u^{k-1}, \forall u \in N(v)\}))$$

### 2. LSTM

### 3. Pooling

GraphSAGE采用的max-pooling策略能够隐式地选取领域中重要的顶点：

$$\text{AGGREGATE}_k^{pool} = \max(\{\sigma(W_{pool} h_u^k + b), \forall u \in N(v)\})$$

## 数据集

BioGRID、Reddit