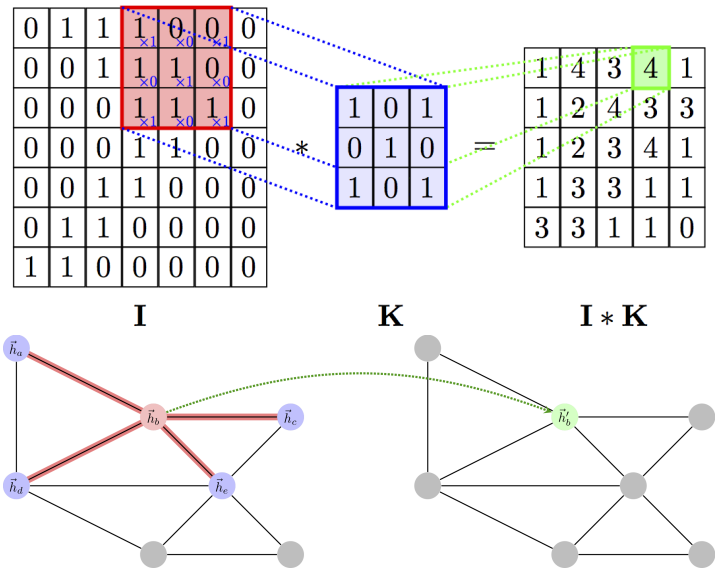
ICLR18一篇解决GCN聚合信息时无法区分信息重要性的论文

解决的问题

如何将attention机制应用于图类型的数据上。

做法及创新

图卷积



给定一个含n个顶点的图，其中顶点的特征构成的集合为$(\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_n})$，$\overrightarrow{h_i} \in \mathbb{R}^F$且邻接矩阵为$A$。一个图卷积层根据已有的顶点特征和图的结构来计算一个新的特征集合$(\overrightarrow{h'_1}, \overrightarrow{h'_2}, \ldots, \overrightarrow{h'_n})$，$\overrightarrow{h'_i} \in \mathbb{R}^{F'}$

每个图卷积层首先会进行特征转换，以特征矩阵W表示，$W \in \mathbb{R}^{F' \times F}$它将特征向量线性转换为$\overrightarrow{g_i} = W\overrightarrow{h_i}$，再将新得到的特征向量以某种方式进行结合。为了利用邻域的信息，一种典型的做法如下：

$$\overrightarrow{h_i}' = \sigma\left(\sum_{j \in N_i} \alpha_{ij} \overrightarrow{g_j}\right)$$

其中N_i表示顶点i的邻域（典型的构造方式是选取直接相连的顶点，包括自身），α_{ij}表示顶点j的特征对于顶点i的重要程度，也可以看成一种权重。

现有的做法都是显式地定义α_{ij}，本文的创新之处在于使用attention机制隐式地定义α_{ij}。所使用的attention机制定义为$a : R^{F'} \times \mathbb{R}^{F'} \to \mathbb{R}$，以一个权重向量$\overrightarrow{a} \in \mathbb{R}^{2F'}$表示，对应于论文中的self-attention。

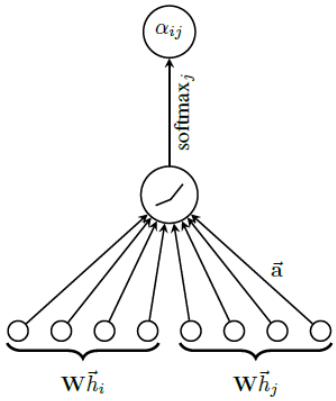
Self-attention

1. 基于顶点的特征计算系数e_{ij}

$$e_{ij} = a(W\overrightarrow{h_i}, W\overrightarrow{h_j})$$

2. 以顶点的邻域将上一步计算得到的系数正则化，这么做能引入图的结构信息：

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in N_i}\exp(e_{ik})}$$
$$= \frac{\exp(\text{LeakyReLU}(\overrightarrow{a}^T[W\overrightarrow{h}_i||W\overrightarrow{h}_j]))}{\sum_{k\in N_i}\exp(\text{LeakyReLU}(\overrightarrow{a}^T[W\overrightarrow{h}_i||W\overrightarrow{h}_k]))}$$



> 次序不变性：给定$(i,j),(i,k),(i',j),(i',k)$表示两个顶点间的关系，可以为边或自环。$a$为对应的attention系数，如果$a_{ij} > a_{ik}$，则有$a_{i'j} > a_{i'k}$

DeepInf中给出了证明：

将权重向量$\overrightarrow{a} \in \mathbb{R}^{2F'}$重写为$\overrightarrow{a} = [p^T, q^T]$，则有

$$e_{ij} = \text{LeakyReLU}(p^T W h_i + q^T W h_j)$$

由softmax与LeakyReLU的单调性可知，因为$a_{ij} > a_{ik}$，有$q^T W h_j > q^T W h_k$，类似地就可以得到$a_{i'j} > a_{i'k}$。

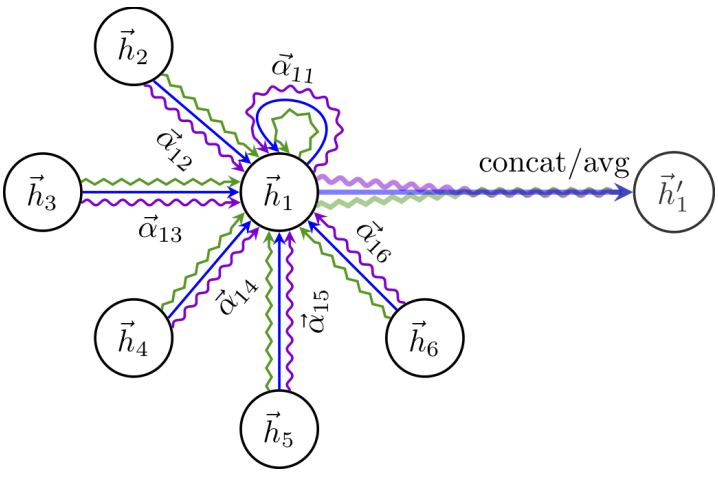
这意味着，即使每个顶点都只关注于自己的邻域，但得到的attention系数却具有全局性。

3. 以上一步得到的系数α_{ij}作为顶点j的特征对顶点i的重要程度，将领域中各顶点的特征做一个线性组合以作为顶点i最终输出的特征表示：

$$\overrightarrow{h'_i} = \sigma\left(\sum_{j\in N_i}\alpha_{ij}W\overrightarrow{h_j}\right)$$

Multi-head attention

为了稳定self-attention的学习过程，论文引入了multi-head attention，即由K个相互独立的self-attention得到各自的特征，再进行拼接：



$$\overrightarrow{h'_i} = \Big\|_{k=1}^{K}\sigma\left(\sum_{j\in N_i}\alpha_{ij}^k W^k\overrightarrow{h_j}\right)$$

其中α_{ij}^k是第k个attention机制(a^k)计算出来的正则化系数，W^k是对应的将输入进行线性转化的权重矩阵。论文选取的拼接操作为求平均：

$$\overrightarrow{h_i'} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^k W^k \overrightarrow{h_j}\right)$$

数据集

Cora、Citeseer、Pubmed、PPI