

This is a note for the paper: Graph Attention Networks ( ICLR 2018 )

## 问题背景

问题的抽象描述：有一张无向图 $G(V, E)$ ，节点数 $N$ ，边数 $E$ ，每个节点各有一个特征向量 $x \in R^d$ ， $x$ 表示了某种属性，比如一个词的嵌入向量。现在已有图上一些节点的标签，希望对图上剩余的节点进行分类（作者提出的未来工作：将注意力引入对图的分类）。

问题的一个实例：文献分类。每个节点代表一篇文献，如果两篇文献有引用关系（不管谁引用谁），则两个节点连一条边。现在已经对一部分文献进行了类别标注（每个类别都标注了一些），希望知道剩余文献的类别。每个节点的特征向量为文献的词向量，维度为词典大小，某一维为1则这个词在此文献中出现过。

## 注意力机制

聚集周围节点的信息时，权重不仅取决于图本身的连接，还取决于节点的特征。取决于节点特征的部分采用注意力机制。注意力系数为

$$e_{ij} = a(Wh_i, Wh_j)$$

$W$ 为线性特征变换，是为了网络有足够的表达能力。 $a$ 为注意力函数，输出为一个标量，即注意力系数。此处，注意力函数采用投影系数，即

$$e_{ij} = \vec{a}^T [Wh_i || Wh_j]$$

$||$ 表示拼接。 $e_{ij}$ 经过leaky relu，再通过一个softmax进行归一化得到注意力系数 $\alpha_{ij}$

最终特征向量的更新为

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

其中 $\sigma$ 为激活函数，此处采用了ELU（relu将负值部分改为衰减的指数函数，正值部分还是单位映射）

另外，作者宣称可以用此机制不依赖与邻接矩阵，用注意力当做连接系数。但这样就丧失了结构信息，不能称之为图网络，也不合理。实际中，作者还是采用了邻接矩阵，在邻接矩阵上乘以注意力系数。