# 《SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS》

# 《基于图卷积网络的半监督分类》

**摘要：**

1. **本文的背景：**由于卷积神经网络（convolutional neural network, CNN）在深度学习领域取得了很大成功，因此研究者们迫切希望也能够在图上定义卷积运算。而在本文此之前，已经有大量图卷积处理的方法被提出，如 Spectral Network、ChebNet，它们为本文提供了坚实的理论基础。
2. **本文的贡献：**进一步简化了基于谱分解相关方法的图卷积公式，使其能够更好地与神经网络结合，并且本篇论文正式开启了基于图卷积神经网络的应用潮。
3. **主要创新点：**1）GCN 通过谱图卷积（spectral graph convolutions）的局部一阶近似来确定卷积网络的结构；2）GCN 模型能够学习隐藏层表示，而这些表示既编码了局部图结构，也编码了节点特征；3）通过图结构数据中部分有标签的节点数据对 GCN 模型的训练，使模型能够对其余无标签的数据进行进一步分类。
4. **实验结果：**在对大量关于引用网络和知识图谱网络数据集进行对比实验后，研究者发现 GCN 模型的有效性显著优于其他图学习方法。

**背景知识——谱图卷积：**

（1）Spectral graph convolutions(第一代卷积公式)：

$$g_\theta \star x = U g_\theta U^T x$$

其中，$x$为图节点的特征向量，$x \in \mathbb{R}^N$；$g_\theta = diag(\theta)$，为图卷积核；$\theta \in \mathbb{R}^N$为待训练参数；$U$为图的归一化拉普拉斯矩阵$L$的特征向量矩阵，而归一化拉普拉斯矩阵$L = I_N - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = U \Lambda U^T$。【其中，$I_N$为单位矩阵，$D$、$A$和$\Lambda$分别为图的度矩阵、邻接矩阵和特征值的对角矩阵。】

<u>但由于这个公式的计算太过复杂，且卷积核的选取并不合适，需要改进。</u>

（2）Chebyshev polynomials(第二代卷积公式)：

$$g_\theta \star x \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{L})x$$

2011 年，David Hammond 等人提出可以用切比雪夫多项式的前 K 阶$T_k(x)$来逼近$g_\theta$，因此有了如上公式。【https://zhuanlan.zhihu.com/p/106687580 相关公式推导参考 https://www.zhihu.com/question/54504471/answer/332657604】其中，$\tilde{L} = \frac{2}{\lambda_{max}} L - I_N$，$\lambda_{max}$是$L$的最大特征值，$\theta \in \mathbb{R}^K$是切比雪夫多项式系数组成的向量。而切比雪夫多项式的定义为：

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$$

其中，$T_0(x) = 1$，$T_1(x) = x$。而由于上述公式是拉普拉斯矩阵的 K 阶多项式，所以上述的卷积公式运算也是 K 阶局部化的。

**该方法的好处在于省去了计算拉普拉斯矩阵特征向量的过程。**

**基于图的卷积公式：**

本文在上述理论的基础上做出了进一步优化，主要改进有：

1）将层级卷积运算的$K$限制为 1，以此来缓解模型在节点的度分布范围较大的图上存在的局部结构过拟合问题；

2）进一步假定$\lambda_{max} = 2$来简化公式。因此，可将上述公式$g_\theta \star x \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{L})x$化简为：

$$g_\theta \star x \approx \theta_0 x + \theta_1 (L - I_N)x = \theta_0 x - \theta_1 D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

其中，$\theta_0$和$\theta_1$为可调节参数。

3）对参数$\theta_0$和$\theta_1$做出进一步限制，使得$\theta = \theta_0 = -\theta_1$，因此上述公式又可简化为：

$$g_\theta \star x \approx \theta \left( I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \right) x$$

4）注意，在节点上迭代执行上述运算时可能会导致数值不稳定和梯度爆炸或梯度消失问题，因此为了解决这些问题，进一步引入**重归一化操作**，使得：

$$I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \rightarrow \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$$

其中，$\tilde{A} = A + I_N$，$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$。

5）最后，将该定义推广到具有$C$个输入通道（即每个节点的$C$维特征向量）的信号$X \in \mathbb{R}^{N\times C}$和$F$个滤波器上，则特征映射（feature maps）如下：

$$Z = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta$$

其中，$\Theta \in \mathbb{R}^{C\times F}$是卷积核参数矩阵，$Z \in \mathbb{R}^{N\times F}$是卷积后的信号矩阵。

**Graph 中的快速近似卷积：**

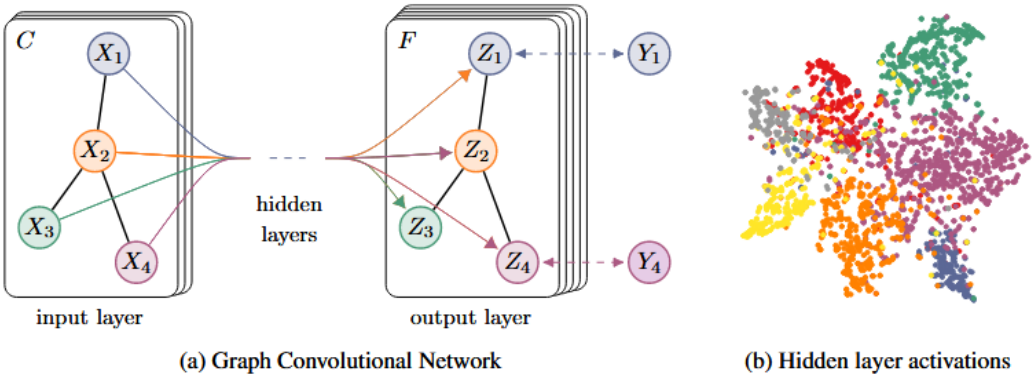
在定义了上述卷积运算后，我们就能构造出具有以下分层传播规则的**多层图形卷积网络（GCN）**：

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

其中，$\tilde{A} == A + I_N$为无向图G的带自环邻接矩阵，$I_N$为单位矩阵，$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$，$W^{(l)}$为 layer-specific 的可训练权重向量，$\sigma(\cdot)$为激活函数，$H^{(l)} \in \mathbb{R}^{N\times D}$为第$l$层的激活矩阵，而$H^{(0)} = X$。

**半监督节点分类模型：**

在有了 GCN 模型后，我们已经可以在图上有效的传播信息，因此对于半监督节点分类问题，一个整体的多层半监督 GCN 模型如下图所示：




(a) Graph Convolutional Network　　(b) Hidden layer activations

Figure 1: *Left*: Schematic depiction of multi-layer Graph Convolutional Network (GCN) for semi-supervised learning with $C$ input channels and $F$ feature maps in the output layer. The graph structure (edges shown as black lines) is shared over layers, labels are denoted by $Y_i$. *Right*: t-SNE (Maaten & Hinton, 2008) visualization of hidden layer activations of a two-layer GCN trained on the Cora dataset (Sen et al., 2008) using 5% of labels. Colors denote document class.

在上图中，左(a)是一个 GCN 网络示意图，它在输入层拥有$C$个输入，中间有若干隐藏层，在输出层有$F$个特征映射；图的结构（边用黑线表示）在层之间共

享；标签用$Y_i$表示。右(b)是一个两层 GCN 在 Cora 数据集上（使用了 5%的标签）训练得到的隐藏层激活值的形象化表示，颜色表示文档类别。

**GCN 模型训练实例：**

考虑一个用于半监督图节点分类问题的两层 GCN，邻接矩阵为$A$（二进制/加权）。

1）预处理操作

在预处理操作中，计算$\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$，因此前述模型可进一步简化为：

$$Z = f(X, A) = softmax(\hat{A}ReLU(\hat{A}XW^{(0)})W^{(1)})$$

其中，$W^{(0)} \in \mathbb{R}^{C\times H}$为输入层到隐藏层的权重矩阵，该隐藏层具有$H$个特征映射，$W^{(1)} \in \mathbb{R}^{H\times F}$为隐藏层到输出层的权重矩阵。

2）损失函数计算

对于半监督多类别分类，我们评估所有标记标签的交叉熵误差，即：

$$\mathcal{L} = -\sum_{l\in y_L}\sum_{f=1}^{F} Y_{lf} lnZ_{lf}$$

其中，$y_L$为带标签的节点集。

3）模型训练

网络中的权重$W^{(0)}$和$W^{(1)}$通过梯度下降训练；并且对每个【训练迭代】使用完整的数据集执行【批量梯度下降】；对于$A$使用稀疏表示，即边数是线性的；通过添加 Dropout 操作引入训练过程中的随机性。

**实验及结果：**

1）数据集

Table 1: Dataset statistics, as reported in Yang et al. (2016).

| Dataset | Type | Nodes | Edges | Classes | Features | Label rate |
|---|---|---|---|---|---|---|
| Citeseer | Citation network | 3,327 | 4,732 | 6 | 3,703 | 0.036 |
| Cora | Citation network | 2,708 | 5,429 | 7 | 1,433 | 0.052 |
| Pubmed | Citation network | 19,717 | 44,338 | 3 | 500 | 0.003 |
| NELL | Knowledge graph | 65,755 | 266,144 | 210 | 5,414 | 0.001 |

如上图所示，在方法对比实验中，研究主要采用了引文网络数据集 Citeseer、Cora 和 Pubmed，以及从知识图中提取的二分图 NELL。它们分别对应了**半监督文档分类问题**和**半监督实体分类问题**。

2）实验设置

采用前述的两层 GCN 网络模型，并在 1000 个标记示例的测试集上评估预测准确性。此外，还使用了 500 张带标签的验证集来为超参数进行优化，其中包含两层 GCN 的 dropout 率，第一层 GCN 的$L2$正则化因子以及隐藏层单元的个数。

而对于引文网络数据集的模型训练，它的最大迭代次数为 200，梯度下降算法采用 Adam 算法，学习率为 0.01，训练的停止条件为验证集$loss$连续十个迭代期没有明显下降。此外，（按行）对输入特征向量进行归一化。

3）实验结果

Table 2: Summary of results in terms of classification accuracy (in percent).

| Method | Citeseer | Cora | Pubmed | NELL |
|---|---|---|---|---|
| ManiReg [3] | 60.1 | 59.5 | 70.7 | 21.8 |
| SemiEmb [28] | 59.6 | 59.0 | 71.1 | 26.7 |
| LP [32] | 45.3 | 68.0 | 63.0 | 26.5 |
| DeepWalk [22] | 43.2 | 67.2 | 65.3 | 58.1 |
| ICA [18] | 69.1 | 75.1 | 73.9 | 23.1 |
| Planetoid* [29] | 64.7 (26s) | 75.7 (13s) | 77.2 (25s) | 61.9 (185s) |
| **GCN (this paper)** | **70.3** (7s) | **81.5** (4s) | **79.0** (38s) | **66.0** (48s) |
| GCN (rand. splits) | $67.9 \pm 0.5$ | $80.1 \pm 0.5$ | $78.9 \pm 0.7$ | $58.4 \pm 1.7$ |

如图所示，在各个数据集的对比实验中，GCN 都取得了最好成绩。

4）对卷积传播公式的评价

Table 3: Comparison of propagation models.

| Description | | Propagation model | Citeseer | Cora | Pubmed |
|---|---|---|---|---|---|
| Chebyshev filter (Eq. 5) | $K = 3$ | $\sum_{k=0}^{K} T_k(\tilde{L})X\Theta_k$ | 69.8 | 79.5 | 74.4 |
| | $K = 2$ | | 69.6 | 81.2 | 73.8 |
| 1st-order model (Eq. 6) | | $X\Theta_0 + D^{-\frac{1}{2}}AD^{-\frac{1}{2}}X\Theta_1$ | 68.3 | 80.0 | 77.5 |
| Single parameter (Eq. 7) | | $(I_N + D^{-\frac{1}{2}}AD^{-\frac{1}{2}})X\Theta$ | 69.3 | 79.2 | 77.4 |
| **Renormalization trick** (Eq. 8) | | $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta$ | **70.3** | **81.5** | **79.0** |
| 1st-order term only | | $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}X\Theta$ | 68.7 | 80.5 | 77.8 |
| Multi-layer perceptron | | $X\Theta$ | 46.5 | 55.1 | 71.4 |

如图所示，本文所采用的卷积传播公式$\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta$在引文网络数据集的实验上取得了最好成绩。

5）对模型训练时间的评价

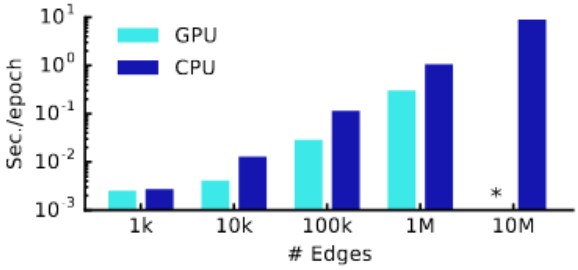



Figure 2: Wall-clock time per epoch for random graphs. (*) indicates out-of-memory error.

如图所示，GCN 模型的每次迭代时间与训练图的大小呈线性相关，因此它具有较好的扩展能力。

**总结：**

1）优点：

图卷积神经网络 GCN 能够以非顺序方式学习，即不以节点的输入顺序为转移；GCN 能够以更加灵活的方式进行节点之间特征的学习；且和 GNN 不同的是，GCN 能够关心以某节点作为中心的 K 阶邻居之内的信息。

2）缺点：

对内存的需求过大，论文中提及的方法均涉及整个 batch 的训练，需要大量的显存，即需要知道整个图的归一化邻接矩阵等信息才能进行计算。此外，原始的 GCN 对带有权重的有向图的支持并不太好。还有，论文中的一些假设可能具有局限性，例如原文假设了局部性（K 阶领域）以及自身节点和临近节点同等重要，实际上，这种假设有可能不成立，我们可以引入权重参数$\lambda$来进行调整：

$$\tilde{A} = A + \lambda I_N$$

**注意：**参数$\lambda$和半监督以及非监督损失的权衡参数类似，同样可以通过梯度下降来进行学习。

**对于本文的感悟：**

GCN 的提出引发了在图神经网络领域上现象级的研究热潮。然而，GCN 模型的提出也并不是空穴来风的。总的来说，它是在大量前人研究的基础上不断改进而得的，而或许这就是量变引起质变的最好例证，一切付出终究是有意义的。