Simplifying Graph Convolutional Networks
https://arxiv.org/abs/1902.07153
https://github.com/Tiiiger/SGC
Abstract
图卷积神经网络及其变种已经受到许多关注，并且已经成为学习图表示的一种事实上的方法。图卷积神经网络主要从最近的深度学习方法中获得灵感，因此也可能继承了许多不必要的复杂性和冗余计算。在本文中我们通过连续移除非线性但愿并破坏连续层间的权重矩阵减少其额外的复杂性。我们从理论上分析了得到的线形模型，并证明其相当于一个固定的低通滤波器加一个线性分类器。并且进一步的实验表明这些简化没有对下游应用的准确性产生负面影响。更进一步来说，我们的模型可以扩展到大规模数据集，具有自然的可解释性，比 FastGCN 还要快两个数量级。
1. Introduction
图卷积神经网络是卷积神经网络在图上的一个高效变种。图卷积神经网络把学到的谱卷积核和非线性激活函数作为神经网络层堆叠在一起来学习图表示。最近，GCN 及其变种都在许多领域的应用中都取得了最前沿的成果，包括但不限于文献引用网络、社交网络、应用化学、自然语言处理、计算机视觉等。
历史上，机器学习算法的发展从最开始的简洁到后来的因需求驱动造成的复杂遵循着清晰的脉络。例如，线性感知机的局限性就促进了更为复杂但表达能力更强的多层感知机的发展。相似的，简单的预定义线性图像滤波器最终促进了具有可学习卷积核的非线性卷积神经网络的发展。因为更复杂的算法设计会导致理论分析的困难，阻碍人们理解算法，所以通常在更简单的算法不足以应对某些实际应用时我们才引入复杂的模型。有争议的是，许多现实世界中的分类器仍然是线性分类器（通常是逻辑回归），我们可以很简单的优化并理解。
然而，也许是因为 GCN 是在最近神经网络的文艺复兴之后才提出的，GCN 可以看作神经网络发展潮流的一个例外，GCN 的设计基于多层神经网络，而不是对图上对应的简单线性网络的拓展。
本文中，我们观察到图神经网络从他们的深度学习家系中继承了相当多的复杂性，这可能给没有这么多需求的实际任务带来不必要的负担。因为缺少一种更简单的对应前身，我们从这种历史遗漏中可以受到启发，我们的目标是推导最简单的线性模型，我们希望推导得到一个更简单的线形模型，如果沿传统发展路线，其“本应”先于图神经网络出现。通过不断的减少图卷积神经网络层之间的非线性激活单元来减少网络复杂性，并把最终的结果函数简化为线性函数。我们从经验上展示了最终的线形模型在许多任务中表现出与图神经网络相当甚至更优秀的性能，同时在计算上更加高效、使用更少的参数。我们把这种简化线形模型称为简化图卷积。
与其非线性模型对应相比，简化图卷积具有直觉上的可解释性，从图卷积的视角提供理论上的分析。值得注意的是，简化图卷积中的特征提取相当于一个施加在每个特征纬度上的单独的固定滤波器。Kipf & Welling (2017) 从经验上观察到“再正则化技巧”，即在图中增加自环，可以提升准确率。我们证明了这种方法有效缩减了图的谱域，在施加在图上时导致了一个低通滤波器。至关重要的是，这种滤波操作提升了整张图的局部光滑特征。
通过在包括文献引用、社交网络等节点分类基准测评数据集上的经验评价，我们展示了简化图卷积展现出和图卷积神经网络或其他前沿图神经网络相当的性能。然而，简化图卷积明显更快，甚至在大规模测评数据集（Reddit）超过了 FastGCN 模型两个数量级。最终，我们证实了简化图卷积可以将其有效性推广到大量的下游任务。尤其是，简化图卷积可以在文本分类、用户地理位置、关系提取、zero-shot 图像分类等任务上与基于图卷积神经网络方法一较高下，甚至可能取得更好的效果。可以在 Github 上找到相应的代码。
2. Simplified Graph Convolution
我们遵循 Kipf & Welling (2017) 的工作在文本分类任务情景下介绍图卷积神经网络，之后介绍简单图卷积。这里，图卷积神经网络将部分节点带有标签的图作为输入，对图上所有节点标签生成预测。让我们对这种图下一个正式的定义：$\mathcal{G} = (\mathcal{V}, \mathbf{A})$，其中 $\mathcal{V}$ 代表节点集合 $\{v_1, \ldots, v_n\}$，并且 $\mathbf{A} \in \mathbb{R}^{n \times n}$ 表示对称邻接矩阵（通常是稀疏矩阵），其中 $a_{i,j}$ 表示节点 v_i 和节点 v_j 之间的边的权重。不存在的边通过 $a_{i,j} = 0$ 表示。我们定义度矩阵 $\mathbf{D} = \mathrm{diag}(\mathrm{d}_1, \ldots, \mathrm{d}_\mathrm{n})$ 为一个对角矩阵，对角线上的每个元素都等于邻接矩阵对应行的行和 $d_i = \sum_{j=1}^{n} a_{i,j}$. 图上的每个节点 v_i 都有一个相应的 d 维特征向量 $\mathbf{x}_i \in \mathbb{R}^d$，完整的特征矩阵 $\mathbf{X} \in \mathbb{R}^{n \times d}$ 将 n 个特征向量堆叠起来，即 $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^\mathrm{T}$. 每个节点都属于 C 类中的一类，所以可以用 C 维独热向量 $\mathbf{y}_i \in \{0,1\}^C$ 表示其标签。我们只知道部分节点的标签，希望能预测剩余的部分。
2.1 Graph Convolutional Networks
与卷积神经网络或者多层感知机相似，图卷积网络通过多层结构为每个节点的特征 $\mathbf{x}_i$ 学习一个新的特征表示。 其之后将用于作为线性分类器的输入。对于第 k 个图卷积层，我们将所有节点的输入节点表示为矩阵 $\mathbf{H}^{(k-1)}$ 并将输出节点表示记为 $\mathbf{H}^{(k)}$。自然的，最初的节点表示只是原始的输入特征：$\mathbf{H}^{(0)} = \mathbf{X}$，其作为图卷积网络的第一层。
一个 K 层图卷积等同于在图中每个节点的特征向量 $\mathbf{x}_i$ 上施加 K 层多层感知机，除了每个节点的隐藏表示在每层开始都由其邻居的表示求平均得到。在每个图卷积层，节点表示的更新分三个阶段：特征传播，线性变换，逐点非线性激活（如图一所示）。为了保持清晰性，我们将描述每一步的细节。
Feature Propagation 是图卷积网路区别于多层感知机的关键。在每层的开始每个节点 v_i 的特征 $\mathbf{h}_i$ 与其局部邻居的特征向量求平均，$\bar{\mathbf{h}}_i^{(k)} \leftarrow \frac{1}{d_i + 1}\mathbf{h}_i^{(k-1)} + \sum_{j=1}^{n} \frac{a_{ij}}{\sqrt{(d_i+1)(d_j+1)}}\mathbf{h}_j^{(k-1)}$. 我们可以通过简单的矩阵操作更紧凑的在整个图上表示这个更新过程，令 $\mathbf{S}$ 表示带有附加自环的正则化邻接矩阵，$\mathbf{S} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}$，其中 $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ 并且 $\tilde{\mathbf{D}}$ 是 $\tilde{\mathbf{A}}$ 的度矩阵。对所有

节点进行同步更新在这种记号下可以简单的写成矩阵乘法 $\bar{\mathbf{H}}^{(k)} \leftarrow \mathbf{S}\mathbf{H}^{(k-1)}$。直觉上讲，这一步局部的沿着边平滑了隐藏表示并且在局部相连的节点中鼓励相似预测结果。

Feature transformation and nonlinear transition. 在进行局部平滑之后，一个图卷积层等同于一个标准多层感知机。每个层与一个可学习的参数矩阵 $\boldsymbol{\Theta}^{(k)}$ 关联，并且平滑后的隐藏特征表示被线性的转换。最终，在输出特征表示 $\mathbf{H}^{(k)}$ 前一个非线性激活函数例如 ReLU 施加在每个点。总而言之，第 k 层的表示更新规则为 $\mathbf{H}^{(k)} \leftarrow \mathrm{ReLU}\left(\bar{\mathbf{H}}^{(k)}\boldsymbol{\Theta}^{(k)}\right)$. 第 k 层的逐点非线性变换之后跟着第 $(k+1)$ 层的特征传播。

Classifier. 对于节点分类，类似于标准多层感知机，图卷积网络的最后一层使用 softmax 分类器预测标签。记对 n 个节点的分类预测为 $\hat{\mathbf{Y}}_{GCN} \in \mathrm{softmax}\left(\mathbf{S}\mathbf{H}^{(K-1)}\boldsymbol{\Theta}^{(K)}\right)$，其中 $\mathrm{softmax}(\mathbf{x}) = \frac{\exp(\mathbf{x})}{\sum_{c=1}^{C}\exp(x_c)}$ 对所有类进行正则化。

2.2 Simple Graph Convolution

作为传统的多层感知机，更深的层增加了表达能力，因为其允许创造特征的层级，即第二层的特征建立在第一层特征的基础上 。在图卷积神经网络中，网络层有第二个重要的功能，在每层中，隐含表示在一跳邻居中被平均。这里的隐含意义是 k 层之后一个节点可以从所有图上 k 跳距离之内的节点获得特征信息。这个效果相似于卷积神经网络，其中随着深度的增加内部特征感受野也随之增加。尽管卷积网络可以从深度增加中显著受益，通常的多层感知机在超过 3 或 4 层之后受益就很少增加了。

Linearization. 我们假设图卷积神经网络层之间的非线性单元不关键，其大部分受益都来自局部平均操作。我们因此移除每层之间的非线性转移函数只留下最终的 softmax 来得到概率化的输出。最后得到的模型是线性，但是仍旧有和 K 层图卷积神经网络相同的扩大的「感受野」。$\hat{\mathbf{Y}} = \mathrm{softmax}\left(\mathbf{S}\ldots\mathbf{S}\mathbf{S}\boldsymbol{\Theta}^{(\mathbf{1})}\boldsymbol{\Theta}^{(\mathbf{2})}\ldots\boldsymbol{\Theta}^{(\mathbf{K})}\right)$. 为了简化记号，我们可以将正则化邻接矩阵的重复乘法缩减为单独的一个矩阵，即 $\mathbf{S}^K$. 进一步的，我们可以把权重重新参数化为一个单独的矩阵 $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(1)}\boldsymbol{\Theta}^{(2)}\ldots\boldsymbol{\Theta}^{(K)}$. 得到的分类器就变为 $\hat{\mathbf{Y}}_{SGC} = \mathrm{softmax}\left(\mathbf{S}^K\mathbf{X}\boldsymbol{\Theta}\right)$，我们将其称为 Simple Graph Convolution (SGC).

Logistic Regression. 上面的等式为 SGC 给出了一个自然的直觉上的解释：通过区分特征提取和分类器，SGC 包括了两个模块：一个固定的（例如非参的）特征提取/平滑模块 $\bar{\mathbf{X}} = \mathbf{S}^K\mathbf{X}$ ，之后跟着一个线性的逻辑回归分类器 $\hat{\mathbf{Y}} = \mathrm{softmax}(\bar{\mathbf{X}}\boldsymbol{\Theta})$ 模块。因为 $\bar{\mathbf{X}}$ 的计算不需要权重，实际上就等同于一个特征预处理步骤，并且模型的整个训练过程化简为在预处理后的特征 $\bar{\mathbf{X}}$ 上进行直接的多类逻辑回归。

Optimization Details. 训练逻辑回归是一个经过充分研究的凸优化问题，并且可以由任何高效的二阶方法或者随机梯度下降方法来实现。假定图联通模式足够稀疏，SGD 可以自然的扩展到非常大规模的图上，并且训练 SGC 比 GCN 要快很多。

3. Spectral Analysis

我们现在从图卷积的视角来研究一下 SGC. 我们证明 SGC 相当于图谱域的一个固定卷积核。进一步的，我们证明了在原图上增加自环，即重正则化技巧，有效收缩了隐含的图谱。在调整后的图谱域，SGC 相当于一个低通滤波器，生成整张图上的光滑特征。作为结果，临近节点倾向于共享相似的表示以及之后的预测结果。

3.1 Preliminaries on Graph Convolutions

与欧几里得域可以类比，图傅立叶分析依靠的是对图拉普拉斯矩阵的谱分解。图拉普拉斯算子为 $\Delta = \mathbf{D} - \mathbf{A}$ （一种正则化形式为 $\Delta_{sym} = \mathbf{D}^{-1/2}\Delta\mathbf{D}^{-1/2}$） ，是对称半正定矩阵，特征分解为 $\Delta = \mathbf{U}\Lambda\mathbf{U}^T$，其中 $\mathbf{U} \in \mathbb{R}^{n\times n}$ 由正交特征向量组成，$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ 是一个由特征值构成的对角矩阵。拉普拉斯矩阵的特征分解让我们能定义定义在图谱域对应的傅立叶变换，其中特征向量表示傅立叶模式，特征值表示图的频率。在这种意义下，令 $\mathbf{x} \in \mathbb{R}^n$ 表示定义在图中顶点上的信号。我们定义对向量 $\mathbf{x}$ 的图傅立叶变换为 $\hat{\mathbf{x}} = \mathbf{U}^{\mathrm{T}}\mathbf{x}$，逆变换为 $\mathbf{x} = \mathbf{U}\hat{\mathbf{x}}$. 因此，在信号 $\mathbf{x}$ 和卷积核 $\mathbf{g}$ 之间的图卷积操作为 $\mathbf{g} * \mathbf{x} = \mathbf{U}\left((\mathbf{U}^{\mathbf{T}}\mathbf{g}) \odot (\mathbf{U}^{\mathbf{T}}\mathbf{x})\right) = \mathbf{U}\hat{\mathbf{G}}\mathbf{U}^{\mathbf{T}}\mathbf{x}$，其中 $\hat{\mathbf{G}} = \mathrm{diag}(\hat{g}_1, \cdots, \hat{g}_n)$ 表示对角矩阵，对角线元素是相对应的谱卷积系数。图卷积操作可以由 k 阶拉普拉斯多项式近似 $\mathbf{U}\hat{\mathbf{G}}\mathbf{U}^{\mathbf{T}}\mathbf{x} \approx \sum_{i=0}^{k}\theta_i\Delta^i\mathbf{x} = \mathbf{U}\left(\sum_{i=0}^{k}\theta_i\boldsymbol{\Lambda}^i\right)\mathbf{U}^{\mathrm{T}}\mathbf{x}$. 其中 θ_i 表示系数。在这种情形下，卷积核系数相当于拉普拉斯多项式的特征值，即 $\hat{\mathbf{G}} = \sum_i \theta_i\boldsymbol{\Lambda}^i$ 或者等价的 $\hat{g}(\lambda_j) = \sum_i \theta_i\lambda_j^i$.

图卷积网络使用仿射近似（k=1）等式 $\mathbf{U}\hat{\mathbf{G}}\mathbf{U}^{\mathbf{T}}\mathbf{x} \approx \sum_{i=0}^{k}\theta_i\Delta^i\mathbf{x} = \mathbf{U}\left(\sum_{i=0}^{k}\theta_i\boldsymbol{\Lambda}^i\right)\mathbf{U}^{\mathrm{T}}\mathbf{x}$，其中$\theta_0 = 2\theta$ 并且 $\theta_1 = -\theta$，我们可以得到基础图卷积神经网络的卷积操作 $\mathbf{g} * \mathbf{x} = \theta(\mathbf{I} + \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2})\mathbf{x}$. 在他们的最终设计中，Kipf 用 正则化版本的 $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ 取代了矩阵 $\mathbf{I} + \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$，其中 $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ 并且 $\tilde{\mathbf{D}} = \mathbf{D} + \mathbf{I}$，被称为再正则化技巧。最终，通过推广卷积操作，来在 d 通道输入上处理多个卷积核，同时在模型网络层之间添加非线性激活函数，我们就可以得到传播方程 $\mathbf{H}^{(k)} \leftarrow \mathrm{ReLU}\left(\bar{\mathbf{H}}^{(k)}\boldsymbol{\Theta}^{(k)}\right)$.

3.2 SGC 和低通滤波器

最初的由GCN得到的一阶切比雪夫滤波器相当于传播矩阵 $\mathbf{S}_{1-\mathrm{order}} = \mathbf{I} + \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. 因为正则化拉普拉斯算子 $\Delta_{sym} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\Delta\mathbf{D}^{-1/2}$，那么 $\mathbf{S}_{1-\mathrm{order}} = 2\mathbf{I} - \Delta_{sym}$. 因此，使用 $\mathbf{S}_{1-\mathrm{order}}^K$ 进行特征传播蕴含着卷积核系数 $\hat{g}_i = \hat{g}(\lambda_i) = (2-\lambda_i)^K$，其中 λ_i 表示拉普拉斯算子 Δ_{sym} 的特征值。Figure 2 展示了这种与 $\mathbf{S}_{1-\mathrm{order}}$ 相关的针对不同传

播步骤 $K \in \{1,\cdots,6\}$ 的滤波操作。正如我们可以观察到的，高阶的 $\mathbf{S}_{1-\mathrm{order}}$ 会导致爆炸的滤波器系数以及我们不希望看到的对于频率 $\lambda_i < 1$ 被过分放大的信号。

为了解决与一阶切比雪夫滤波器相关的数值问题，Kipf 提出了再正则化技巧。基本上来说，就是用增加了自环后再正则化的邻接矩阵来代替 $\mathbf{S}_{1-\mathrm{order}}$，我们将得到的邻接矩阵称为增强正则邻接矩阵 $\tilde{\mathbf{S}}_{\mathrm{adj}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$，其中 $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\tilde{\mathbf{D}} = \mathbf{D} + \mathbf{I}$，相应的，我们定义增强正则拉普拉斯算子为 $\tilde{\Delta}_{\mathrm{sym}} = \mathbf{I} - \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$，因此我们可将与 $\tilde{\mathbf{S}}_{\mathrm{adj}}$ 相关的谱卷积核表示为与隐含拉普拉斯算子特征值相关的特征多项式，即 $\hat{g}(\tilde{\lambda}_i) = (1 - \tilde{\lambda}_i)^K$，其中 $\tilde{\lambda}_i$ 表示 $\tilde{\Delta}_{\mathrm{sym}}$ 的特征值。

我们现在分析了 $\tilde{\Delta}_{\mathrm{sym}}$ 的谱并且展示了在图上增加自环可以收缩相应正则化拉普拉斯矩阵的谱（特征值）。

Theorem 1 令 $\mathbf{A}$ 表示没有孤立节点的无向简单带权图 $\mathcal{G}$ 的邻接矩阵，$\mathbf{D}$ 表示度矩阵，令 $\tilde{\mathbf{A}} = \mathbf{A} + \gamma\mathbf{I}$，其中 $\gamma > 0$，为增强邻接矩阵，其度矩阵为 $\tilde{\mathbf{D}}$. 同样，令 λ_1, λ_n 分别表示 $\Delta_{\mathrm{sym}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ 的最小和最大特征值，相似的，令 $\tilde{\lambda}_1, \tilde{\lambda}_n$ 分别表示 $\tilde{\Delta}_{\mathrm{sym}} = \mathbf{I} - \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$ 的最小和最大特征值。我们有 $0 = \lambda_1 = \tilde{\lambda}_1 < \tilde{\lambda}_n < \lambda_n$. Theorem 1 说明正则化图拉普拉斯矩阵的最大特征值在加上自环 $\gamma > 0$ 后变小了（证明见补充材料）。

Figure 2 绘制了在 Cora 数据集上进行的与正则化邻接矩阵 $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ 及其增强变种 $\tilde{\mathbf{S}}_{\mathrm{adj}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$ 相关的滤波操作。使用 $\mathbf{S}_{\mathrm{adj}}$ 进行的特征传播对应于在谱范围 $[0,2]$ 上的滤波器$g(\lambda_i) = (1 - \lambda_i)^K$. 因此，奇数次幂的 $\mathbf{S}_{\mathrm{adj}}$ 对频率 $\lambda_i > 1$ 得到负滤波器系数。通过增加自环（$\tilde{\mathbf{S}}_{\mathrm{adj}}$），最大的特征值从 2 收缩到大致 1.5，并且消除了负系数的影响。更进一步的，这个调整尺度后的谱允许通过对 $\tilde{\mathbf{S}}_{\mathrm{adj}}$ 取 $K > 1$ 次幂定义滤波器以表现的像个低通滤波器。在补充材料中，我们从经验上衡量对传播矩阵的不同选择。

4.1 Related Work