解决一般图卷积网络的过度平滑问题

黄文兵*，于荣*，徐廷阳，孙富春，黄俊洲

摘要--增加GCN的深度，预计会允许更多的表现力，但却显示出对性能的损害，特别是在节点分类上。其主要原因在于过度平滑。过度平滑的问题促使GCN的输出走向一个包含有限的节点区分信息的空间，导致表达能力差。一些关于改进深度GCN结构的工作已经被提出，但在理论上，这些改进是否能够缓解过度平滑的问题仍然是未知的。在本文中，我们首先从理论上分析了一般的GCN是如何随着深度的增加而行动的，包括通用GCN、带偏置的GCN、ResGCN和APPNP。我们发现，所有这些模型都有一个普遍的过程：所有的节点都收敛到一个立方体。根据这个定理，我们提出了DropEdge，通过在每个训练周期随机删除一定数量的边来缓解过度平滑的问题。理论上，DropEdge要么降低了过度平滑的收敛速度，要么缓解了维度坍缩造成的信息损失。对模拟数据集的实验评估已经直观地显示了不同GCN之间过度平滑的差异。此外，在几个真实基准上的广泛实验支持DropEdge在各种浅层和深层GCN上持续改善性能。

Index Terms-Graph Convolutional Networks, Over-Smoothing, DropEdge, Node Classification.

1 简介

很多数据都是图结构的形式，其中一定数量的节点通过边不规则地联系在一起。例如社交网络[1]、知识库[2]、分子[3]、场景图[4]等。在图上学习是至关重要的，不仅对于图数据本身的分析，而且对于一般的数据形式，因为图提供了强大的归纳偏见，使关系推理和组合概括成为可能[5]。最近，图神经网络（GNN）[6]已经成为图学习的最理想工具。发明GNN的最初动机是为了推广神经网络（NN）的成功经验从表格/网格数据到图域。

GNN的关键精神在于，它利用递归邻域聚合函数来结合来自一个节点以及其邻域的特征向量，直到一个固定的迭代次数d（*又称*网络深度）。给定一个适当定义的聚合函数，这种消息传递被证明可以捕获每个节点在其*d-hop*邻域内的结构，就像Weisfeiler- Lehman（WL）图的同构性测试[7]一样强大，已知它可以区分一大类图[8]。在本文中，我们主要关注的是图卷积网络(GCNs）[1], [9], [10], [11], [12], [13], [14], 是GNN的一个核心系列，将卷积操作从图像扩展到图形。GCN已经被成功地用于节点分类的任务，这也是本文的主要焦点。

在视觉领域已经众所周知，卷积神经网络（CNN）的深度在性能上起着关键作用。受CNN成功的启发，人们可能期望通过堆叠更多的层来使GCN具有更多的表达能力，以描述更丰富的邻居拓扑结构。开发深层GCN的另一个原因是，描述图的拓扑结构需要足够深的架构。[15]和[16]的工作表明，如果深度受到限制，GCN就不能学习图的时刻或估计某些图的属性。

然而，制定深层和ex- pressive GCN的期望并不容易实现。这是因为深度GCN实际上受到了主要由*过度平滑*造成的表达能力的损害[17]。过度平滑的一个直观概念是，通过图卷积对邻域特征的混合，促使无限深的GCN的输出走向一个包含节点间有限区分信息的空间。从训练的角度来看，过度平滑会从输入中抹去重要的区分信息，从而导致集合训练性。我们在图1中展示了一个实验实例，其中观察到深度GCN的训练收敛效果很差。

已经提出了一些尝试来探索如何构建深度GCN[1]、[12]、[18]、[19]。然而，他们都没有提供足够有表现力的架构，而且这些架构是否在理论上保证了防止（或至少缓解）过度平滑，仍然不清楚。Li等人最初将GCN线性化为拉普拉斯平滑，并发现图的每个连接组件内的顶点特征将收敛到相同的值。在[17]的基础上向前迈进了一步，Oono

- *黄文兵(hwenbing@126.com)和孙富春(fcsun@mail.tsinghua.edu.cn)在北京国家研究中心工作。*
  清华大学计算机科学与技术系智能技术与系统国家重点实验室信息科技中心（BNRist）。
- *于荣（yu.rong@tencent.com）、徐廷阳（tingyangxu@tencent.com）和黄俊洲（joehhuang@tencent.com）是腾讯人工智能实验室的成员。中国，深圳。*
- *黄文兵和于荣对这项工作贡献相同。*
- *孙富春是通讯作者。*

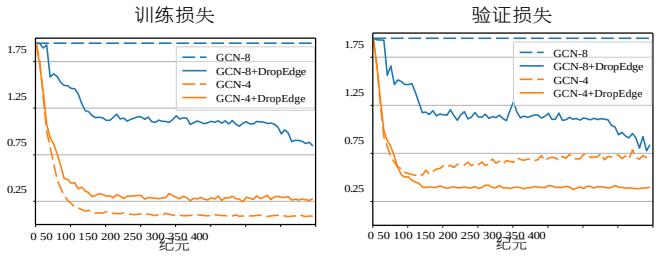


图1：Cora上GCN的性能。我们实现了4层和8层的GCNs，包括DropEdge和不包括DropEdge。GCN-4陷入了过度拟合的问题，获得了较低的训练误差，但验证误差较高；GCN-8的训练由于过度平滑而无法令人满意地收敛。通过应用DropEdge，GCN-4和GCN-8在训练和验证中都表现良好。请注意，这里的GCN是没有偏差的。

& Suzuki[20]同时考虑了非线性（ReLU函数）和卷积滤波器，并证明GCN收敛于以节点度数为基础制定的子空间，但这一结果仅限于通用的GCN[1]，没有讨论其他架构。

因此，*在理论上，对于一般的GCN家族来说，为什么以及何时会发生过度平滑？*

为此，我们首先以一般方式重新审视过度平滑的概念。除了通用的GCN[1]，我们还探索了通常在实践中实现的带偏置的GCN[15]，以及通过涉及跳过连接来细化GCN的ResGCN[1]和APPNP[12]。我们从数学上证明，如果我们使用无限的层数，所有这些模型将收敛到一个*立方体*，该*立方体*扩展了[20]提出的子空间，直到一定半径r。首先，收敛到立方体意味着收敛到子空间，但反之则不然。其次，与现有方法[17]，[20]不同的是，我们的结论显示，加入偏见会导致半径不为零，有趣的是，这将在某种程度上阻碍过度平滑。最后，我们的定理表明，ResGCN减缓了过度平滑，而APPNP总是保持某些输入信息，这两点与我们的本能理解一致，但以前没有严格的探索。

对立方体而不是子空间的过度平滑，尽管不是那么糟糕，但仍然限制了表达能力，需要加以缓解。为此，我们提出了DropEdge。术语 "DropEdge "指的是在每个训练时间内随机放弃输入图的某些边的比率。在它的特殊形式中，每条边都以固定的概率p被独立地丢弃，p是一个超参数，由验证决定。在GCN训练中应用DropEdge有几个好处（见图1中DropEdge的实验改进）。首先，DropEdge可以被看作是一个消息传递减速器。在GCN中，相邻节点之间的信息传递是沿着边缘路径进行的。去除某些边缘是使节点连接更加稀疏，因此在GCN非常深入时，在一定程度上避免了过度平滑。事实上，正如我们将在本文中从理论上得出的，DropEdge要么减慢过度平滑的退化速度，要么减少信息损失。

DropEdge的另一个优点是，它也可以被看作是一种数据增强技术。通过DropEdge，我们实际上是在生成原始图形的不同随机变形副本；因此，我们增强了输入数据的随机性和多样性，从而能够更好地防止过度拟合。这类似于执行随机
旋转、裁剪或拍打，用于强大的CNN训练。
图像的背景。请注意，DropEdge与随机图生成方法（如Erdos-Renyi（ER）模型[21]或稀疏图学习方法（如GLASSO[22]）在边缘修改方面有关系。然而，DropEdge根据均匀分布为不同的训练迭代去除不同的边缘子集，而ER模型或GLASSO一旦创建了随机/稀疏图，就会在所有训练迭代中采用相同的图。这就是为什么DropEdge能够减轻每个训练迭代的过度平滑，但是，在整个训练过程中，它仍然保留了概率意义上的基础图核的全部信息。

我们提供了一套完整的实验来验证我们的
与我们对过度平滑的重新思考有关的结论，以及DropEdge在四个节点分类的基准上的功效。特别是，我们的DropEdge--作为一种灵活和通用的技术，能够提高各种流行的骨干网络的性能，包括GCN[1]、ResGCN[19]、JKNet[18]和APPNP[12]。它证明了DropEdge在各种浅层和深层的GCN上都能持续提高性能。完整的细节将在第5节提供。

总而言之，我们的贡献如下。

- 我们研究了一般深层GCN（通用GCN、带偏置的GCN、ResGCN和APPNP）输出的渐进行为，其中涉及非线性问题。我们从理论上表明，这些GCN将收敛到一个具有无限层堆叠的立方体。
- 我们提出了DropEdge，一种新的技术，在每次训练迭代中统一丢弃一定数量的边，这被证明可以缓解一般深度GCN的过度平滑问题，即减慢收敛速度或减少信息损失。
- 在四个节点分类基准上的实验支持了我们提出的定理的合理性，并表明DropEdge能够增强各种GCN的浅层和深层变体。

2 相关的工作

GCNs。第一个关于GCN的突出研究是在[9]中提出的，它发展了基于频谱和空间视图的图卷积。后来，[1]、[23]、[24]、[25]、[26]对基于谱的GCN进行了改进、扩展和近似。为了解决基于频谱的GCN在大图上的可扩展性问题，基于空间的GCN得到了快速发展[11], [27], [28], [29]。最近，一些基于抽样的方法被提出来用于快速的图表示学习，包括节点抽样方法[11]，层抽样方法[10]，[13]。
和图解法[30], [31]。具体来说，GAT[32]

已经讨论了在边缘注意力上应用Dropout。虽然它实际上是注意力计算前的DropEdge的后导版本，但在[32]中从未探讨过与过度平滑的关系。然而，在我们的论文中，我们已经正式提出了DropEdge的表述，并为其在缓解过度平滑方面的好处提供了严格的理论依据。

深层GCNs。尽管取得了丰硕的成果，但以前的大多数工作只关注浅层GCN，而对深层的扩展很少讨论。构建深层GCN的尝试可以追溯到GCN论文[1]，其中应用了残差机制；出乎意料的是，正如他们的实验所示，当深度为3及以上时，残差GCN的性能仍然较差。作者在[17]中首先指出了构建深度网络的主要困难在于过度平滑，但遗憾的是，他们从未提出任何方法来解决这个问题。后续研究[12]通过使用个性化的PageRank来解决过度平滑的问题，该方法还将根节点纳入了信息传递循环。JKNet[18]采用密集连接进行多跳消息传递，这与DropEdge兼容，用于制定深度GCN。最近，DAGNN[33]完善了GCN的架构，首先将表示转换与传播解耦，然后利用自适应调整机制来平衡每个节点的本地和全球邻域的信息。

作者在[20]中从理论上证明了深层GCN的节点特征将收敛到一个子空间并产生信息损失。它通过考虑ReLU函数和卷积滤波器，概括了[17]中的结论。在本文中，我们研究了更广泛的一类GCN的过度平滑行为，并表明一般的GCN会收敛到子空间以外的立方体。Chen等人[34]根据[17]的结论开发了一个过度平滑的测量方法，并提出通过使用基于监督的优化方法来缓解过度平滑，而我们的DropEdge被证明只需通过随机边缘采样就能缓解一般GCN的问题，简单而有效。其他最近的研究为了防止过度平滑，采用了激活归一化[35]和双残差连接[36]，这与我们的DropEdge是互补的。最近的一种方法[19]将剩余层、密集连接和扩张卷积纳入GCNs，以促进深度架构的发展，其中没有讨论过度平滑。

其他相关的工作。在DropEdge中，从输入图中去除边缘的想法与稀疏图学习方法[22]，[37]相似但不同。通过DropEdge，我们并不意味着底层图核的边缘是充分无信息的；相反，DropEdge仍然保留了这种信息，因为它**以一种随机但无偏见的方式**行事。我们将在下文中提供更多解释§ 第4.2节和第5.3.3节的评价。著名的Perron- Fronbenius定理（PFT）[38]和谱图定理[39]描述了图上随机漫步的收敛行为。然而，这些结果并不直接适用于我们的情况，因为这里的相邻关系被增加了自循环，而且模型比随机漫步更复杂，*例如涉*及非线性。

3 初步了解

3.1 图形表示法和光谱分析。

让$G = (V, E)$代表大小为N的输入图，节点$v_i \in V$ 边$(v_i, v_j) \in E$。我们用
$\boldsymbol{X} = \{x_1, \ldots, \boldsymbol{x}_N\} \in \mathbb{R}^{N \times C}$的节点特征，并通过
$\boldsymbol{A} \in \mathbb{R}^{N \times N}$毗连矩阵，其中元素$\boldsymbol{A}(i, j)$
返回每条边（v_i, v_j）的权重。节点度由$\boldsymbol{d} = \{d_1, \ldots, d_N\}$给出，其中$d_i$ 计算与节点i相连的边权重之和。我们定义$\boldsymbol{D}$
按照以前的研究[1], [12], [18], 边缘权重应该是非负的, 只捕捉节点之间的相似性, 而不是它们的负相关关系.

我们将在后面介绍，GCN[1]通过添加自循环和增强的度数归一化来应用规范化的增强邻接，这将导致
$\hat{\boldsymbol{A}} = \hat{\boldsymbol{D}}^{-1/2}(\boldsymbol{A} + \boldsymbol{I})\hat{\boldsymbol{D}}^{-1/2}$, 其中$\hat{\boldsymbol{D}} = \boldsymbol{D} + \boldsymbol{I}$. 我们定义增强的归一化拉普拉斯[20]为$\hat{\boldsymbol{L}} = \boldsymbol{I} - \hat{\boldsymbol{A}}$通过建立与通用光谱理论的关系Laplacian[39]，Oono & Suzuki[20]推导出增强的归一化Laplacian及其邻接的谱系，从而。我们将该结果总结如下。

定理1（增强的光谱特性[20]）。*由于$\hat{\boldsymbol{A}}$是对称的，让$\lambda_1 \leq \cdots \leq \lambda_N$ 是$\hat{\boldsymbol{A}}$的实数特征值，按升序排序。假设最大的特征值λ_N 的倍数是M，*即，$\lambda_{N-M} < \lambda_{N-M+1} = \cdots = \lambda_N$。*那么我们有。*

- $-1 < \lambda_1, \lambda_{N-M} < 1$。
- $\lambda_{N-M+1} = \cdots = \lambda_N = 1$。
- *M是由G中的连接组件的数量给出的，$\hat{\boldsymbol{e}}_m := \hat{\boldsymbol{D}}^{1/2}\boldsymbol{u}_m$ 是与特征值λ_{N-M+m} 相关的特征向量，其中$\boldsymbol{u}_m \in \mathbb{R}^N$ 是第m个连接组件的指标，*即，$\boldsymbol{u}_m(i) = 1$ *如果节点i属于第m个成分，否则$\boldsymbol{u}_m(i) = 0$*。

定理1主要关注的是正化增强邻接$\hat{\boldsymbol{A}}$的特征值。著名的Perron-Fronbenius定理（PFT）[38]指出，对于一个行随机的、不可还原的概率矩阵$\boldsymbol{P}$，光谱半径正好是1，所有其他特征值的大小都严格小于1。然而，这个结果不能应用于
因为$\hat{\boldsymbol{A}}$不是一个不可还原的随机矩阵。的
[39]的结论适用于邻接矩阵$\boldsymbol{A}$的谱分析，但它没有考虑到增强的自循环（*即*$\boldsymbol{A}+\boldsymbol{I}$）。定理1严格地将[39]的结论从$\boldsymbol{A}$*推广*到其增强的
版本$\hat{\boldsymbol{A}}$。

3.2 GCN的变体

在此，我们介绍GCN的几个典型变体。

通用的GCN。正如最初由[1]开发的，GCN中的前馈传播是递归进行的，即

$$\boldsymbol{H}_{l+1} = \sigma(\hat{\boldsymbol{A}}\boldsymbol{H}_l\boldsymbol{W}_l)。\tag{1}$$

其中，$\boldsymbol{H}_l = [\boldsymbol{h}_{1,l}, \ldots, \boldsymbol{h}_{N,l}]$ 是隐性向量的
*第l*层，$\boldsymbol{h}_{i,l}$ 是节点i的隐藏特征；$\sigma(\cdot)$是一个非线性函数（它被实现为ReLU

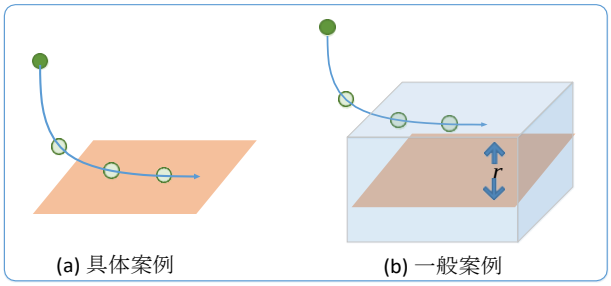


图2：（a）通用GCN的过度平滑[20]：收敛到一个多维子空间M；（b）过度平滑的
一般模型：收敛到一个立方体$O(M, r)$。

贯穿本文）；而$\boldsymbol{W}_l \in \mathbb{R}^{C_l \times C_{l+1}}$是第$l$层的滤波矩阵。在第4节的分析中，我们把
为了简单起见，所有层的尺寸都是一样的 $C_l = C$。除非另有说明，我们此后将通用GCN简称为GCN。

带偏置的GCN（GCN-b）。在大多数文献中，GCN是以公式1的形式引入的，没有明确涉及偏置项，但在实际实施中是必要的。如果加上偏置，公式1被更新为

$$\boldsymbol{H}_{l+1} = \sigma\left(\boldsymbol{A}\boldsymbol{H}_l\boldsymbol{W}_l + \boldsymbol{b}_l\right)。 \tag{2}$$

其中偏差的定义是：$\boldsymbol{b}_l \in \mathbb{R}^{1\times C}$。

ResGCN。通过借用ResNet[40]的概念，Kipf & Welling[1]利用隐藏层之间的残余连接，通过携带前一层输入的信息来促进更深层次模型的训练。

$$\boldsymbol{H}_{l+1} = \sigma\left(\boldsymbol{A}\boldsymbol{H}_l\boldsymbol{W}_l + \alpha H_l\right)。 \tag{3}$$

其中我们进一步增加了权重$0\le\alpha\le1$，以便更灵活地在GCN传播之间取得平衡
和残留的信息。

APPNP。由于深层GCN会因为过度平滑而使输出与输入隔离，Klicpera等人[12]建议明确地从输入层到每个隐藏层进行跳转连接，以保留输入信息。

$$\boldsymbol{H}_{l+1} = (1-\beta)\boldsymbol{A}\boldsymbol{H}_l + \beta H_0。 \tag{4}$$

其中$0\le\beta\le1$是权衡权重。请注意，[12]的原始版本不涉及非线性问题
和每个隐层的权重矩阵。GCNII的工作[36]通过在传播中加入ReLU函数和可训练的权重来寻求更大的容量。在这里，我们采用了原始版本，并发现它工作得很好。

4 分析和方法学

在这一节中，我们首先推导出普遍定理（Theo- rem 2），以解释为什么以及何时会发生过度平滑的情况，适用于§3.2中介绍的所有四个模型。然后，我们介绍了DropEdge，它被证明可以缓解所有模型的过度平滑现象。我们还认为我

4.1 一般模型的过度平滑

过度平滑"的概念最初由[17]提出，后来由[20]和许多其他最近的工作[34]、[35]、[36]、[41]解释。一般来说，过度平滑现象意味着节点表征在经过许多层信息传递后变得混杂，彼此无法区分。因此，它削弱了深度GCN的可训练性和表现力。请注意，在我们的背景下，如果频率被理解为邻接/拉普拉斯矩阵的特征值，加入[42]的定义，那么过度平滑与古典信号处理中的"频率平滑"有关。过度平滑可以解释为小特征值的图形频率过滤。然而，与频率平滑的传统研究相比[42]，考虑到复杂的结构和非线性的参与，GCN模型中过度平滑的分析更具挑战性，这也是本文理论研究的动机。

在我们下面的分析中，我们利用了[20]的结论，因为它同时考虑到了非线性（即ReLU函数）和卷积滤波器的通用性。作者在[20]中解释了深度GCN的过度平滑为收敛到一个多维的子空间。落入这个子空间会遇到信息损失：同一连接部件内的节点是可以区分的

只由它们的度数决定。换句话说，如果两个节点在同一组件中拥有相同的度数，它们的表示方法
在无限层传播之后将是相同的，甚至它们的
们的DropEdge能够防止过度拟合，并涉及DropEdge与其他相关概念的讨论和扩展。

初始特征和局部拓扑结构明显不同。如果连接成分的数量，即子空间的维度（定义见下文）较小，这种信息损失将变得更加严重。因此，我们可以利用每个GCN层和子空间之间的距离来衡量过度平滑有多严重。

子空间的定义如下。

定义1（子空间）。*我们定义* $\mathcal{M} := \{H \in \mathbb{R}^{N\times C} \mid H = \hat{E}C, C \in \mathbb{R}^{M\times C}\}$ *为* $\mathbb{R}^{N\times C}$ *中的* M *维（* $M \le N$ *）子空间，其中* $\hat{E} = \{\hat{e}_1, \ldots, \hat{e}_M\} \in \mathbb{R}$ *收集了定理1中* $\hat{A}$ *最大特征值的基数，即* $\hat{e} = \hat{D}^{1/2}u$ *。* $\in \mathbb{R}^{N\times M}$ *收集了定理1中* $\hat{A}$ *的最大特征值的基数，即* $\hat{e}_m = \hat{D}^{1/2}u_m$ *。*

子空间$\mathcal{M}$是由$\hat{E}$的列展开的。定义1遵循[43]中关于子空间的常规定义，即$\mathcal{M}$在加法和标量下是封闭的乘法。具体来说，如果$H_1, H_2 \in \mathcal{M}$，对于任何$a, b \in \mathbb{R}$，$aH_1 + bH_2 = a\hat{E}C_1 + b\hat{E}C_2 = \hat{E}(aC_1 + bC_2) = \hat{E}C \in \mathcal{M}$，因为$C := aC_1 + bC_2 \in \mathbb{R}^{M\times C}$。我们将矩阵$H \in \mathbb{R}^{N\times M}$和$\mathcal{M}$之间的距离定义为$d_{\mathcal{M}}(H) := \inf_{Y \in \mathcal{M}} \|H - Y\|_F$，其中$\|\cdot\|_F$计算Frobenius规范。然而，[20]的结论只是适用于一般的GCN。在这一节中，我们推导出一个通用定理来描述一般GCN的行为，表明它们会随着深度的增加而收敛到一个立方体而不是子空间。我们首先定义立方体如下。

定义2（长方体）。*我们将* $\mathcal{O}(\mathcal{M}, r)$ *定义为将* $\mathcal{M}$ *扩展到半径* $r \ge 0$ *的立方体，即* $\mathcal{O}(\mathcal{M}, r) := \{H \in \mathbb{R}^{N\times C} \mid d_{\mathcal{M}}(H) \le r\}$ *。*

我们现在设计出关于过度平滑的一般定理。

定理2（一般过度平滑定理）。*对于公式1至公式4中定义的GCN模型，我们普遍有*

$$d_M(\boldsymbol{H}_{l+1}) - r \le v\,(d_M(\boldsymbol{H}_l) - r)\text{。} \tag{5}$$

其中$v \ge 0$和r分别描述了收敛因子和半径，这取决于具体的模型是什么。具体而言。

- *对于一般的GCN（公式1），$v = s\lambda$，$r = 0$。*
- *对于GCN-b(公式2)[1], $v = s\lambda$, $r = \frac{d_M(b_l)}{1-v}$。*
- *对于ResGCN（公式3），$v = s\lambda + \alpha$，$r = 0$。*
- *对于APPNP（公式4），$v = (1-\beta)\lambda$，$r = \frac{\beta d_M(\boldsymbol{H}_0)}{1-v}$。*

其中，$s > 0$是所有$\boldsymbol{W}_l$的所有奇异值的最高值，$\lambda := \max_{n \le 1}^{N-M} |\lambda_n|$是$\hat{\boldsymbol{A}}$的第二大特征值。公式5中的平等在某些规范下是可以实现的。

与权力法[44]和基于随机行走的方法[17]相比，定理2的主要特点是它考虑了非线性ReLU函数σ。由于涉及到非线性，分析收敛行为确实具有挑战性，它需要借助某些棘手的变换和不等式，如附录中的公式17-20所示。此外，定理2适用于一般模型，使其比针对某些特定情况提出的方法[20]更强大。

证明在附录A中提供。根据公式5，我们递归推导出$d_M(\boldsymbol{H}_l) - r \le v(d_M(\boldsymbol{H}_{l-1}) - r) \le \cdots \le v^l(d_M(\boldsymbol{H}_0) - r)$。我们假设任何$v \in$ 都<1 $\{s\lambda, s\lambda + \alpha, (1-\beta)\lambda\}$在定理2中，通过观察$\lambda < 1$，$s \le 1$（由于训练期间的$\pounds_2$惩罚，通常是这种情况）和$\alpha$可以被设置为足够小的情况[2].在此情况下

假设，定理2指出，一般的GCN模型实际上是向立方体$O(M, r)$收敛的，如图2所描述。我们进一步分析了以下的收敛行为

每个特定的模型都有以下说明。

备注1。*对于一般的无偏差的GCN，半径变成$r = 0$，我们有$\lim_{l\to\infty} d_M(\boldsymbol{H}_{l+1}) \le \lim_{l\to\infty} v^l d_M(\boldsymbol{H}_0) = 0$，表明$\boldsymbol{H}_{l+1}$指数地收敛于$M$，从而导致过度平滑，正如[20]已经研究的那样。*

备注2。*对于GCN-b，半径不为零：$r>0$，我们有$\lim_{l\to\infty} d_M(\boldsymbol{H}_{l+1}) \le \lim_{l\to\infty} r + v^l(d_M(\boldsymbol{H}_0) - r) = r$，即，$\boldsymbol{H}_{l+1}$指数地收敛到立方体$O(M, r)$。与$M$不同的是，$O(M, r)$与$\mathbb{R}^{N\times C}$和$O(M, r)$具有相同的维度。*
可能包含有用的信息（除节点度之外），用于
节点代表。

备注3.*对于ResGCN来说，它最终收敛到M的速度与通用GCN相似。然而，由于$v = s\lambda + \alpha \ge s\lambda$，与通用GCN相比，它对$M$的收敛速度较慢，这与我们对增加的好处的直观理解一致。*
残留的连接。

备注4.*对于APPNP，它收敛于M以外的$O(M, r)$。*
这就解释了为什么将输入层加入到每个
APPNP中的隐藏层有助于阻碍过度平滑。请注意
增加β将扩大r，但同时减少v，从而导致更快地收敛到一个更大的立方体。

1. 为了简单起见，我们假设所有层的距离$d_M(b_l)$保持一致；否则，我们可以将其定义为上位数。

2. 否则，如果$v \ge 1$，将有可能导致梯度爆炸和深度GCN的不稳定训练，这不是本文的重点。

上面备注1-4的讨论表明，λ的值在影响不同模型的过度平滑方面起着重要作用，较大的λ意味着较少的过度平滑。在下一小节中，我们将介绍我们提出的DropEdge方法能够增加λ，从而防止过度平滑。

4.2 缓解过度平滑的DropEdge

在每个训练周期，DropEdge技术以随机方式丢弃一定比例的输入图的边。形式上，它随机地将邻接矩阵A的$V\ p$非零元素强制为零，其中V是边缘的总数，p是丢弃率。如果我们把得到的邻接矩阵表示为$\boldsymbol{A}_{\text{drop}}$，那么它与$\boldsymbol{A}$的关系就变成了

$$\boldsymbol{A}_{\text{drop}} = \text{Unif}(\boldsymbol{A}, 1 - p)。 \tag{6}$$

其中$\text{Unif}(\boldsymbol{A}, 1 - p)$对$\boldsymbol{A}$中的每条边进行统一采样，属性为$1-p$，即$\boldsymbol{A}_{\text{drop}}(i, j) = \boldsymbol{A}(i, j) * \text{Bernoulli}(1 - p)$。在我们的实现中，为了避免多余的采样
边缘，我们通过从$\boldsymbol{A}$中以非替换的方式抽取大小为V（$1-p$）的边缘子集来创建$\boldsymbol{A}_{\text{drop}}$。按照[1]的想法，我们也对$\boldsymbol{A}_{\text{drop}}$进行重新规范化的技巧，以获得$\hat{\boldsymbol{A}}_{\text{drop}}$。我们在公式1中用$\hat{\boldsymbol{A}}_{\text{drop}}$代替$\hat{\boldsymbol{A}}$
用于传播和训练。在验证和测试时，不利用DropEdge。

定理2告诉我们，深层GCN的退化表现力与v密切相关，因此λ--$\hat{\boldsymbol{A}}$的绝对第二大特征值。在此，我们将证明
采用DropEdge降低了λ，并缓解了过度平滑。在我们之前的会议版本[41]中，我们只讨论了DropEdge如何影响$\boldsymbol{A}$的频谱$_{\text{drop}}$，而没有考虑重新归一化的技巧。在本文中，我们将直接得出归一化的结论
扩充后的邻接矩阵$\hat{\boldsymbol{A}}$下降

定理3. *我们把$\lambda(p)$表示为重新规范化后的下降率p下的预期$\boldsymbol{A}$的任何绝对特征值$_{\text{drop}}$。我们可以通过以下方式来约束$\lambda(p)$的值*

$$\mu(p) \leq \lambda(p) \leq \gamma(p)。 \tag{7}$$

此外，$\gamma(p)-\mu(p)$之间的差距随着p单调地减少，当$p=1$时，差距减少到零，导致$\mu(1)= \lambda(1)= \gamma(1)= 1$。

我们在附录B中提供了全部细节。定理3告诉我们，执行DropEdge能够增加λ的上界和下界（同时减少它们的差距），这将迫使λ走向一个更大的值，特别是当p接近1时（有足够的边缘被放弃）。这在一定程度上可以减慢公式5中的过度平滑速度。图3说明了λ和p之间的关系，其中λ最初可能会波动，但当p增大时，其值最终会增加。

DropEdge也能够增加子空间的维度，从而减轻信息损失，这一点已经被我们的会议论文[41]所证明。我们将这一特性总结为以下定理。

定理4. *关于公式1到公式4中的GCN模型，我们用M假设定义1中定义的收敛子空间，在*

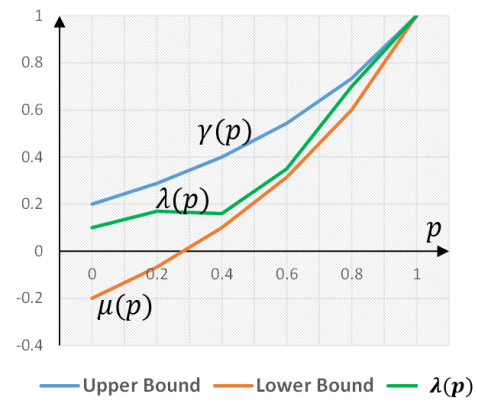


图3：公式7的说明，其中λ被$\mu(p)$和$\gamma(p)$所约束。附录B中给出了$\mu(p)$和$\gamma(p)$的推导。

在原图上，在DropEdge之后的图上，由M来决定。那么，在删除某些边之后，信息损失只是减少了。$N - dim(M) \geq N - dim(M\ \ )$[3].

定理3-4确实表明DropEdge能够缓解过度平滑，但它们并不意味着通过DropEdge防止过度平滑总是能够提供更高的分类性能。例如，放弃所有的边会完全解决过度平滑的问题，但这也会削弱模型的表达能力，因为GCN模型已经退化为MLP，而没有考虑拓扑结构建模。一般来说，如何在防止过度平滑和表达图的拓扑结构之间取得平衡是很关键的，我们应该注意选择一个适当的边缘丢弃率p来反映这一点。在我们的实验中，我们通过使用验证数据来选择p的值，并发现它在一般情况下效果很好。

我们想强调的是，DropEdge是**以随机但无偏见的方式**行事的。它确实在每个特定的训练迭代中删除了一定数量的边，以缓解过度平滑，但不同的边在不同的迭代中按照统一的概率被删除。在期望中，从整个训练过程的角度来看，图的边缘（和底层内核）的信息仍然被保留。简而言之，第$(l+1)$层中节点i的隐藏特征是由它在第l层中的所有邻域聚合而成的，这意味着全聚合被定义为 $\boldsymbol{h}_{i,l+1} = \sum_{j\in N(i)} \boldsymbol{A}(i,j)\boldsymbol{h}_{j,l}$ 。在DropEdge中，每条边都是从伯努利分布中抽取的，表示为伯努利（1-p），其中p是放弃率。那么，关于所有的聚合的期望值为边，由$\mathbb{E}[\sum_{j\in N(i)} h_{i,l+1}|H_l] = \mathbb{E}[\sum_{j\in N(i)} A_{i,j}\boldsymbol{h}_{j,l}|H_l] = (1-p)\sum_{j\in N(i)} A_{i,j}\boldsymbol{h}_{j,l}$，这与原来的完全聚合是一样的，最多是一个乘数1-$p$。如果和到1的归一化，这个乘数就会被抹去。这种无偏的抽样行为使我们的DropEdge有别于随机抽样。

3. 在一般意义上，数据空间的维度不一定反映信息量，但在本文中，考虑到定义1给出的子空间的特殊结构，收敛到较小维度的子空间确实表明了更严重的信息损失。

图的生成方法[21]或稀疏图学习方法[22]，其中图一旦生成/修改就会在所有训练迭代中保持固定，导致边缘连接的确切信息损失。顺便说一下，随机放弃边缘能够创造出输入图的不同随机变形。通过这种方式，DropEdge能够防止过度拟合，类似于典型的图像增强技能（*例如*旋转、裁剪和拍打），以阻碍训练CNN的过度拟合。我们将在第5.3.3节提供实验验证。

层-明智的DropEdge。上述DropE- dge的表述是一次性的，所有层共享相同的扰动邻接矩阵。事实上，我们可以对以下情况进行DropEdge
每个单独的层。具体来说，我们得到$\boldsymbol{A}^{(l)}$ 下降 作者：in-
依靠计算每个l层的公式6。不同的层可以有不同的邻接矩阵$\boldsymbol{A}^{(l)}$。这样的层- 下降
明智的版本带来了更多的随机性和原始数据的变形，我们将在第5.3节中实验性地比较它与原始DropEdge的性能。

4.3 讨论

本节对比了DropEdge和其他相关概念的区别，包括Dropout、DropNode和Graph Sparsification。我们还讨论了节点分类和图分类之间过度平滑的区别。

DropEdge vs. Dropout。Dropout技巧[45]试图通过随机设置特征尺寸为零来扰乱特征矩阵，这可能会减少过度拟合的影响，但对防止过度平滑没有帮助，因为它没有对邻接矩阵做任何改变。作为参考，DropEdge可以被看作是Dropout的一代，从放弃特征维度到放弃边缘，它可以减轻过度拟合和过度平滑。事实上，Dropout和DropEdge的影响是相互补充的，它们的兼容性将在实验中显示。

DropEdge vs. DropNode。另一个相关的脉络属于那种基于节点采样的方法，包括GraphSAGE [11], FastGCN [10], 和AS-GCN [13].我们将这一类方法命名为DropNode。就其原始动机而言，DropNode对子图进行采样以进行小批量训练，它也可以被视为一种特殊形式的丢弃边缘，因为与丢弃节点相连的边缘也被移除。然而，DropNode对丢弃边缘的影响是面向节点的和间接的。相比之下，DropEdge是面向边缘的，它有可能为训练保留所有的节点特征（如果它们能一次装入内存的话），表现出更大的灵活性。此外，为了保持理想的性能，目前的DropNode方法中的采样策略通常是低效的，例如GraphSAGE受到指数级增长的层大小（采样节点的数量）的影响，而AS-GCN需要逐层递归地进行采样。然而，我们的DropEdge既没有随着深度的增加而增加层的大小，也没有要求递归的进展，因为所有边的采样是平行的。

DropEdge与图谱解析。 图-

稀疏化[46]是图域中一个古老的研究课题。它的目标是去除图中不必要的边。

表1:数据集统计

数据集	节点	边缘	课堂	特点	培训/验证/测试	类型
Cora	2,708	5,429	7	1,433	1,208/500/1,000	横向的
馨香坊	3,327	4,732	6	3,703	1,812/500/1,000	横向的
公共医学杂志	19,717	44,338	3	500	18,217/500/1,000	横向的
睿迪特（Reddit）公司	232,965	11,606,919	41	602	152,410/23,699/55,334	感应式

压缩的同时几乎保留了输入图的所有信息。这显然与DropEdge的目的不同，DropEdge不需要优化目标。具体来说，DropEdge会在每次训练时随机删除输入图的边，而Graph- Sparsification则采用繁琐的优化方法来确定哪些边要被删除，一旦这些边被丢弃，输出图就会保持不变。

节点分类与图分类。我们论文的主要重点是节点分类，即所有节点都在一个相同的图中。在图分类中，节点分布在不同的图中；在这种情况下，定理2仍然适用于每个图，同一图实例中无限深的GCN的节点激活只能通过节点度来区分。然而，对于不同图中的那些节点来说，这并不是真的，因为它们会收敛到$\boldsymbol{M}$中的不同位置（*即*定义1中的$\boldsymbol{C}$）。为了说明这一点，我们假设所有的图形实例都是完全连接的图，并且共享相同形式的$\boldsymbol{A} = \{\mathbf{1}_{N\times N}\}$，图$i$内的节点特征$\boldsymbol{X}_i$ (≥ 0)是相同的，但与不同图中的节点特征不同，且权重矩阵在公式1中被固定为$\boldsymbol{W} = \boldsymbol{I}$。然后，通用GCN的任何一层对图$i$不断输出$\boldsymbol{X}_i$，这表明在不同的图之间不会发生信息混淆。请注意，对于图形分类来说，每个图形的过度平滑仍然会阻碍GCN的表达能力，因为它将导致输入数据的维度崩溃。

5 实验

我们进行实验评估的目的是为了回答以下问题。

- 我们提出的普遍过度平滑定理是否与实验观察相一致？
- 我们的DropEdge是如何帮助缓解不同GCN的过度平滑的？

为了解决第一个问题，我们在一个模拟数据集上显示了当深度增长时，节点激活将如何表现。我们还计算了节点激活和子空间之间的距离以显示收敛的动态。至于第二个问题，我们在几个真实的节点分类基准上，对比不同深度的模型的分类性能，有无DropEdge。同时也涉及到与最先进的方法的比较。

节点分类数据集。加入以前的作品的在实践中，我们重点研究了四个在图的大小和特征类型上各不相同的基准数据集：(1) 对三个引用数据集中的论文研究主题进行分类。Cora、Citeseer和Pubmed[47]；(2) 预测Reddit社交网络中不同的帖子属于哪个社区[11]。请注意，在Cora、Citeseer和Pubmed中的任务都是过渡性的，即所有的节点特征在训练期间都可以获得，而

Reddit中的任务是归纳性的，意味着测试节点在训练中是不可见的。我们在实验中对所有数据集采用了[13]和[10]中使用的完全监督训练方式。所有数据集的统计资料列于表1。1.

小科拉。我们从Cora中构建了一个小数据集。具体来说，我们从Cora的训练图中抽取了两个连接组件，节点数分别为654和26。节点的原始特征维度为1433，不适合在二维平面上进行可视化。因此，我们应用截断的SVD进行降维，输出维度为2。 图4 (a)显示了节点特征的分布。我们称这个模拟数据集为Small-Cora。

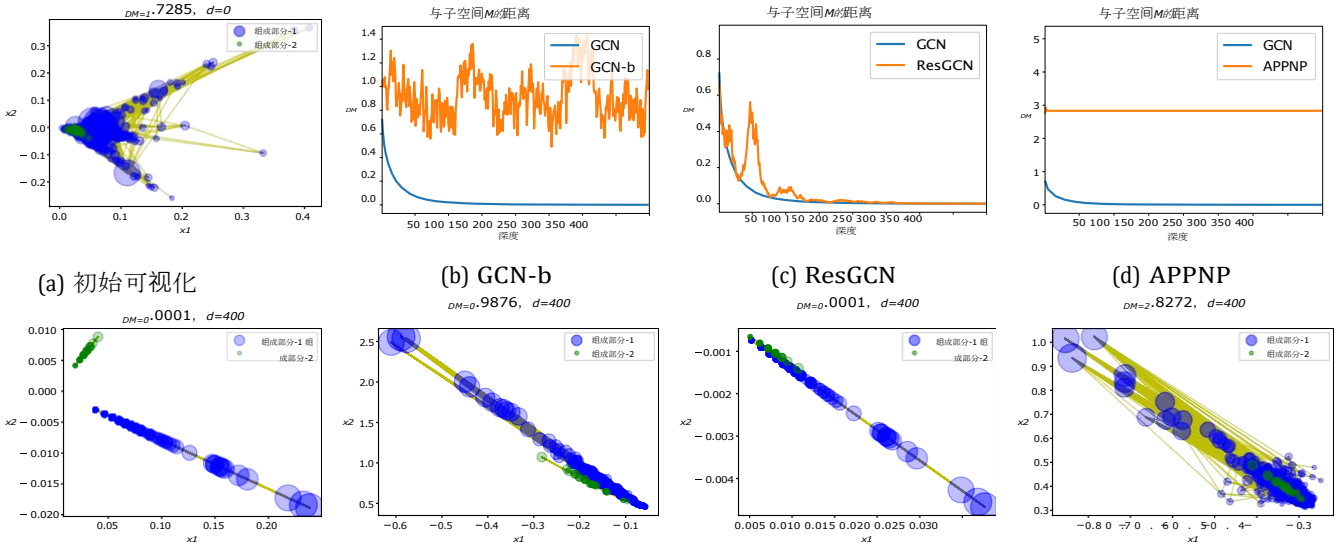
5.1 视觉化的过度平滑在小可乐上的应用

定理2得出了四个模型的过度平滑的普遍性。GCN、GCN-b、ResGCN和APPNP。在这里，为了检查它是否与经验观察一致，我们在Small-Cora上可视化了节点激活的动态。

实施。为了更好地关注不同GCN的不同结构如何影响过度平滑，本节的实验将所有GCN的隐藏维度固定为2，为每一层随机初始化一个正交的权重矩阵$\boldsymbol{W}$，并保持其未被训练，这导致公式5中的$s = 1$。我们还删除了ReLU函数，因为我们发现，有了ReLU，节点的激活将退化为当层数增加时，就会出现零，这将阻碍可视化的实现。关于GCN-b，每层的偏差被设定为0.05。我们对ResGCN和APPNP分别固定α=0.2和β=0.5。由于节点总数较少（*即*680个），我们能够准确设计出子空间的基数。*即*根据定理1的$\boldsymbol{E}$，用附录A的公式11计算节点激活$\boldsymbol{H}$和子空间$\boldsymbol{M}$之间的距离：$d_{\boldsymbol{M}}$。图4展示了输出动态的所有模型。我们有以下发现。

对于GCN来说。当d=0时，节点一般是可以区分的（见（a））。增加d后，距离$d_{\boldsymbol{M}}$ 急剧下降，最后达到非常小的值，当d=400（见（b）为例）。我们可以清楚地看到，当d=400时，不同成分中的节点会碰撞到不同的线上，而且节点越大（度数越大），它离零点越远（见（e））。这种观察与备注1一致，因为不同的线确实代表了子空间的不同基点。

对于GCN-b。GCN-b的输出动态与GCN不同。事实证明，当d=400时，同一分量内的节点保持非共线性，如（f）所示。在（b）中，与GCN相反，GCN-b的曲线在一定范围内波动。这一结果与备注2相吻合，并支持加入偏置项使节点激活收敛到围绕子空间的一定镭的立方体。



(a) 初始可视化 (b) GCN-b (c) ResGCN (d) APPNP

(e) GCN (f) GCN-b (g) ResGCN (h) APPNP

图4：GCN的输出动态。(a) Small-Cora上的节点分布的初始可视化，其中显示的每个节点的大小反映了其程度。$_M$(b-d)当深度d在0到400之间时，GCN和GCN-b、ResGCN和APPNP分别与子空间d的距离的比较；(e-h)GCN的输出的可视化。GCN-b、ResGCN和APPNP，当d=400时。

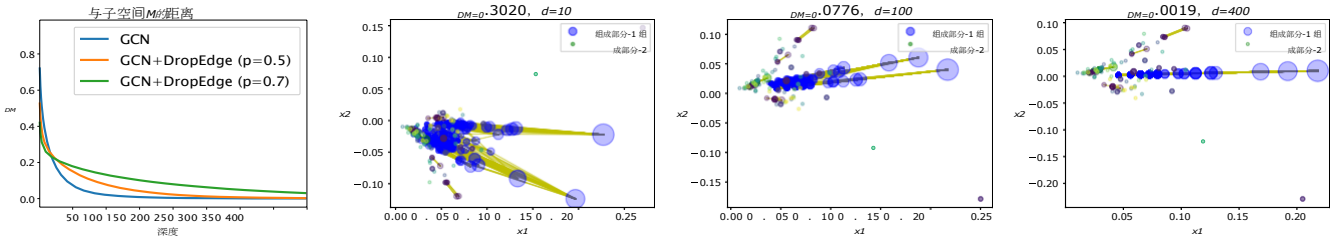


图5：带DropEdge的GCN的输出动态。左边的子图是GCN和带DropEdge的GCN（下降率$p = 0.5, 0.7$）在不同深度下的d_M。其他子图描述了深度为10、100和400时的输出（p=0.5）。

对于ResGCN。与GCN相似，ResGCN的输出最终接近子空间，但它的收敛动态如（c）所示有点不同。在最终退化之前，曲线上下晃动了几轮。这可能是因为ResGCN中的跳过连接有助于防止过度平滑，甚至在早期阶段扭转收敛方向。当d=400时（在(g)中），每个节点将以$\lambda+\alpha$的速度指数化地落入子空间，如备注3所证明。请注意，ResGCN的平均速度小于GCN（回顾$\lambda+\alpha>\lambda$）。

对于APPNP。APPNP的行为与（d）中的GCN完全不同。它很快就变得静止了，而且这个静止点超出了子空间，直到固定的距离r>0，这证实了备注4的结论。在APPNP中，由于速率$v = \lambda\ \beta$比GCN小，它的收敛速度更快。

此外，图5展示了DropEdge如何改变GCN的动态。很明显，结果验证了定理3-4，当我们在GCN上执行DropEdge的下降率p=0.5时，子空间的收敛速度会变慢，连接部件的数量会变大。如果我们进一步增加p到0.7，收敛速度将是进一步减少。

5.2 在不同的GCN上对DropEdge进行评估

在这一节中，我们感兴趣的是，如果应用DropEdge可以促进上述四个GCN模型在真实节点分类基准上的性能。Cora, Cite-seer, Pubmed, 和Reddit。我们进一步实现JKNet[14]并对其进行DropEdge。在下文中，我们将深度为d的每个模型X简称为X-d，*例如*GCN-4表示4层的GCN。

实施。与第5.1节不同的是，所有模型的参数都是可训练的，并采用[1]提出的方法进行初始化，同时增加了ReLU函数。我们在所有的数据集上实现所有的模型，其深度为$d\in\{2, 4, 8, 16, 32, 64\}$，隐藏维度为128。对于Reddit来说，考虑到mem-的最大深度是32。

我们对每个模型进行了随机超参数搜索，并报告了在每个基准数据集上获得最佳精度的情况。由于不同的结构在不同的数据集上表现出不同的训练动态，为了能够进行更稳健的比较，我们对每个模型进行随机超参数搜索，并报告在每个基准的验证集上给出最佳精度的情况。超参数的搜

索空间和更多的细节在附录D的表7中提供。附录D中的表

7。表8描述了不同类型的超参数。8描述了不同类型的

表2：不同骨架上的测试精度（%）。

层数			2	4	8	16	32	64	平均改进
Cora	GCN	原创	85.8	85.6	83.2	81.2	69.8	42.1	-
		滴水不漏	**86.4**	**86.6**	**85.5**	**84.2**	**71.3**	**50.6**	**+2.8**
	GCN-b	原创	86.1	85.5	78.7	82.1	71.6	52.0	-
		滴水不漏	**86.5**	**87.6**	**85.8**	**84.3**	**74.6**	**53.2**	**+2.7**
	ResGCN	原创	-	86.0	85.4	85.3	85.1	79.8	-
		滴水不漏	**-**	**87.0**	**86.9**	**86.9**	**86.8**	**84.8**	**+2.2**
	JKNet	原创	-	86.9	86.7	86.2	87.1	86.3	-
		滴水不漏	**-**	**87.7**	**87.8**	**88.0**	**87.6**	**87.9**	**+1.2**
	APPNP	原创	-	87.9	87.7	87.5	87.8	87.5	-
		滴水不漏	**-**	**88.6**	**89.0**	**88.8**	**88.9**	**89.1**	**+1.2**
馨香坊	GCN	原创	76.8	72.7	72.6	72.2	56.5	43.8	-
		滴水不漏	**78.1**	**79.0**	**75.4**	**67.5**	**60.5**	**46.4**	**+2.1**
	GCN-b	原创	75.9	76.7	74.6	65.2	59.2	44.6	-
		滴水不漏	**78.7**	**79.2**	**77.2**	**76.8**	**61.4**	**45.6**	**+3.8**
	ResGCN	原创	-	78.9	77.8	78.2	74.4	21.2	-
		滴水不漏	**-**	**78.8**	**78.8**	**79.4**	**77.9**	**75.3**	**+11.9**
	JKNet	原创	-	79.1	79.2	78.8	71.7	76.7	-
		滴水不漏	**-**	**80.2**	**80.2**	**80.1**	**80.0**	**80.0**	**+3.0**
	APPNP	原创	-	80.3	80.5	80.2	79.9	80.4	-
		滴水不漏	**-**	**80.8**	**80.9**	**81.1**	**81.2**	**81.3**	**+0.8**
公共医学	GCN	原创	89.5	89.2	88.3	87.7	78.6	72.7	-
		滴水不漏	**89.7**	**90.9**	**91.0**	**90.5**	**80.1**	**77.5**	**+2.3**
	GCN-b	原创	90.2	88.7	90.1	88.1	84.6	**79.7**	-
		滴水不漏	**91.2**	**91.3**	**90.9**	**90.3**	**86.2**	79.0	**+1.2**
	ResGCN	原创	-	90.7	89.6	89.6	90.2	87.9	-
		滴水不漏	**-**	**90.7**	**90.5**	**91.0**	**91.1**	**90.2**	**+1.1**
	JKNet	原创	-	90.5	90.6	89.9	89.2	90.6	-
		滴水不漏	**-**	**91.3**	**91.2**	**91.5**	**91.3**	**91.6**	**+1.2**
	APPNP	原创	-	90.4	90.3	89.8	90.0	90.3	-
		滴水不漏	**-**	**90.7**	**90.4**	**90.3**	**90.5**	**90.5**	**+0.3**
睿迪特（	GCN	原创	95.75	96.08	96.43	79.87	44.36	-	-
		滴水不漏	**95.93**	**96.23**	**96.57**	**89.02**	**66.18**	-	**+6.3**
	GCN-b	原创	96.11	96.62	96.17	67.11	45.55	-	-
		滴水不漏	**96.13**	**96.71**	**96.48**	**90.54**	**50.51**	-	**+5.8**
	ResGCN	原创	-	96.13	96.37	96.34	93.93	-	-
		滴水不漏	**-**	**96.33**	**96.46**	**96.48**	**94.27**	**-**	**+0.2**
	JKNet	原创	-	96.54	96.82	95.91	95.42	-	-
		滴水不漏	**-**	**96.75**	**97.02**	**96.39**	**95.68**	**-**	**+0.3**
	APPNP	原创	-	95.84	95.77	95.64	95.59	-	-
		滴水不漏	**-**	**95.89**	**95.91**	**95.76**	**95.73**	-	**+0.1**

对邻接矩阵进行归一化，而归一化的选择被视为一个超参数。关于相同的架构，不管是有还是没有DropEdge，我们应用相同的超参数集，除了用于公平评估的下降率p。我们采用亚当优化器进行模型训练。为了确保结果的再生产性，所有实验的随机数的种子被设置为相同。我们将所有数据集的训练周期数固定为400。所有实验都在具有24GB内存的NVIDIA Tesla P40 GPU上进行。附录中的Tab.附录9总结了每个骨干网在不同数据集上具有最佳精度的超参数。

总体结果。Tab.2总结了所有数据集的结果。我们有以下的观察。

- 在所有的情况下，DropEdge始终提高了没有DropEdge的模型的测试准确性。例如，在Citeseer上，ResGCN-64未能产生有意义的分类性能，而带有DropEdge的ResGCN-64仍然提供有希望的结果。
- 对于深度模型，GCN-b、ResGCN和APPNP的表现普遍优于通用GCN，无论是否有DropEdge，这与我们之前的retical分析是一致的。特别是，APPNP+DropEdge在Cora和Citeseer上表现最好，而JKNet+DropEdge在Pubmed和Reddit上产生了最好的结果，表明DropEdge对现代架构的兼容性。
- 在使用DropEdge之后，所有的模型通常在深度足够的情况下达到最高的精度。大。例如，当d增加时，GCN和GCN-b的表现都更差，但使用DropEdge时，它们都在d=4时达到了峰值，这可能是由于DropEdge减轻了过度舒缓的作用。

此外，所有4层和6层模型在Cora和Citeseer上的验证损失显示在图6。应用DropEdge后，训练和验证的曲线都被大幅拉低，这也解释了DropEdge的好处。

与SOTA的比较 我们为每个骨干网选择最佳性能的DropEdge，并与现有的艺术状态（SOTA）进行对比，包括KLED [48]，DCNN [49]，FastGCN [10]，AS-GCN [13]，，

GraphSAGE [50] 。

表3：与SOTA的测试准确率（%）比较。括号中的数字表示网络深度。

	Cora	馨香坊	公共医学杂志	睿迪特（Reddit）公司
KLED[48]	82.3	-	82.3	-
DCNN[49]	86.8	-	89.8	-
GAT [32]	87.4	78.6	89.7	-
FastGCN [10]	85.0	77.6	88.0	93.7
AS-GCN [13]	87.4	79.7	90.6	96.3
GraphSAGE [11]	82.2	71.4	87.1	94.3
DAGNN [33]	88.4	78.6	86.4	OOM
GCN+DropEdge	86.6(4)	79.0(4)	91.0(8)	96.57(8)
GCN-b+DropEdge	87.6(4)	79.2(4)	91.3(4)	96.71(4)
ResGCN+DropEdge	87.0(4)	79.4(16)	91.1(32)	96.48(16)
APPNP+DropEdge	**89.1(64)**	**81.3(64)**	90.7(4)	95.91(8)
JKNet+DropEdge	88.0(16)	80.2(8)	**91.6(64)**	**97.02(8)**
DAGNN+DropEdge	88.7(8)	79.5(8)	87.1(16)	OOM

和DAGNN[33]在Tab.3；对于SOTA方法，我们重新使用了[13]中报告的结果。我们在表3中列出了这些结果。3:

- 显然，我们的DropEdge在与SOTA的竞争中获得了明显的提升；特别是在Cora和Cite-seer上，APPNP+DropEdge的最佳准确率为89.10%和81.30%，明显优于之前的最佳成绩（87.44%和79.7%），并且与没有Drop的APPNP相比获得了约1%的提升。考虑到对这些基准的挑战，这样的改进被认为是一个显著的提升。
- 对于大多数带有DropEdge的模型，在深度超过4的情况下获得了最好的准确性，这再次验证了DropEdge对制定深度网络的影响。
- 如4.3节所述，FastGCN、AS-GCN和Graph-SAGE被认为是GCN的DropNode扩展。如表3所示，基于DropEdge的方法优于基于DropNode的变体。3，这证实了DropEdge的有效性。
- DAGNN是最近提出的一个模型，能够缓解过度平滑的问题。表3也报告了DAGNN的性能，在这里我们进行了不同深度的DAGNN，并在每个数据集上收集了最佳案例。据观察，在DAGNN上添加DropE- dge可以进一步提高性能，这意味着我们提出的方法的通用性。

5.3 消融研究

本节继续进行几项消融研究，以评估DropEdge中每个拟议组件的重要性。我们采用GCN作为骨干。隐蔽维度、学习率和权重衰减被固定为256、0.005和0.0005，接受性的。随机种子是固定的。除非另有提及，我们不使用 "withloop "和 "withbn "操作（见附录D中的表7的定义）。

5.3.1 关于与辍学的相容性

§ 4.3节已经讨论了DropEdge和Dropout之间的区别。因此，我们对GCN-4进行了消融研究，验证损失在图7（a-b）中展示。图中显示，虽然Dropout和DropEdge都能促进GCN的训练，但DropEdge的改进更为显著，如果我们同时采用它们，损失会进一步减少，这说明DropEdge与Dropout的兼容性。

5.3.2 关于分层的DropEdge

§ 第4.2节描述了DropE- dge的Layer-Wise（LW）扩展。在这里，我们提供了关于评估其效果的实验评估。从图7（c-d）可以看出，LW DropEdge实现了比原始版本更低的训练损失，而两个模型的验证值是相当的。这意味着LW DropEdge可以比原始DropEdge进一步促进训练。然而，我们更倾向于使用DropEdge而不是LW变体，这样不仅可以避免过度拟合的风险，还可以降低计算的复杂性，因为LW DropEdge需要对每一层进行采样，并花费更多时间。

表4:ER-Graph、GLASSO（GGL）和DropEdge的性能比较。

数据集骨干原始ER-图GLASSO DropEdge					
Cora	GCN-b	0.831	0.319	0.432	0.849
馨香坊	GCN-b	0.715	0.233	0.220	0.763
公共医学杂志	GCN-b	0.850	0.407	-	0.861

表5:GLASSO（GGL）和DropE- dge的运行时间。

数据集	骨干网	原创	格拉索	滴水不漏
Cora	GCN-b	7.77s	95.27s	8.73s
馨香坊	GCN-b	13.49s	328.63s	15.71s
公共医学杂志	GCN-b	33.26s	>40h	35.98s

5.3.3 关于与ER-Graph和GLASSO的比较

我们对使用由Erdos-Renyi模型[21]创建的随机图进行了消融研究，其边数与DropEdge之后的图相等。我们按照与DropEdge相同的设置，在这些随机图上训练6层的模型GCN-b，以进行公平的比较，并将20次运行的平均结果总结在表4中。这表明所有的结果都比DropEdge差得多，可能是因为ER方法生成的图与原始图中的边缘的通用统计不一致。相反，DropEdge的表现总是很好。

此外，我们在表4中提供了DropEdge和GLASSO[22]的实验对比。对于GLASSO，我们计算了经验协方差矩阵

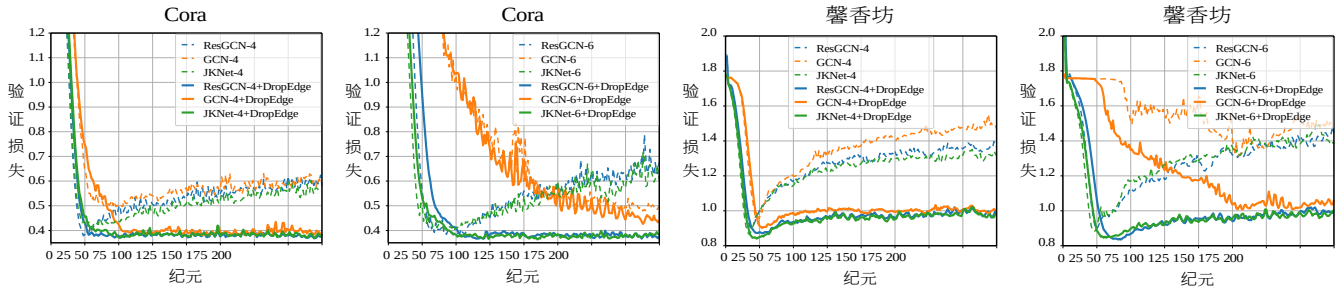
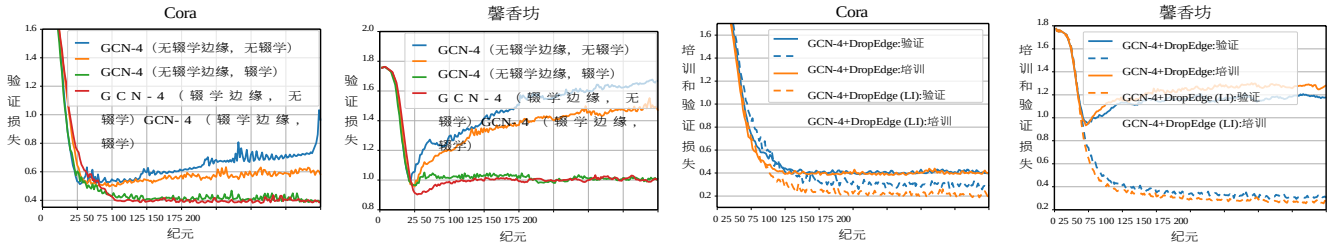


图6：不同骨干网的验证损失，有DropEdge和无DropEdge。GCN-n表示深度为n的GCN；其他骨干网也有类似的表示方法。



(a) (b) (c) (d)

图7：（a-b）DropEdge与Dropout的兼容性；（c-d）Layer-wise DropEdge的性能。

基于节点特征向量，并利用源码[4][2]构建的源代码来执行受限的GLASSO。特别是，我们选择了GGL设置（[2]中的问题1）来强制学习图拉普拉斯的对角线外元素为非正，以便按照预期输出一个正图。为了保持公平的比较，我们选择了适当的超参数（my_eps_outer=1.00E-05；Cora上的alpha=4.00E-05，Citeseer上的alpha=2.00E-06）以确保GGL的边数与DropEdge产生的边数相等。然后我们对学到的稀疏图进行GCN。从表4来看，在Cora和Citeseer数据集上，GGL的性能比DropE- dge和全图版本差很多。由于缺乏原始邻接矩阵的信息，GGL无法保持边缘连接的全部信息，而DropEdge能够通过边缘采样缓解过度平滑，并在整个训练阶段仍然保持全部信息。GLASSO的另一个缺点是稀疏化的优化过程很耗时。表5显示了400个历时的训练时间。可以看出，GLASSO涉及大量的额外运行时间，这使得它对大规模数据集不实用；我们无法获得
在Pubmed数据集上进行GGL，即使经过40小时的计算，也有合理的结果。

表6:统一丢弃和特征加权丢弃的性能比较。

数据集	骨干网	特征加权	统一科
~~拉~~	~~GCN-b~~	~~0.856~~	~~0.876~~
城市猎人	GCN-b	0.797	0.792
公共医学	GCN-b	0.888	0.913

4. https://github.com/STAC-USC/Graph_Learning

5.3.4 关于与基于特征的抽样调查的比较

在DropEdge中考虑到成对的相关性/相似性是可能的。然而，这将使DropE- dge有偏见，更多关注节点特征的相关性，而较少关注邻接矩阵$\boldsymbol{A}$中提供的真实边缘连接。正如之前所解释的，我们假设DropEdge的采样是无偏的，以便保持边缘连接的原始分布。如果使用有偏见的版本，它将在某种程度上污染$\boldsymbol{A}$中的信息，并可能导致性能受损。为了证明这一点，我们在具有4层的GCN-b上进行了实验，首先通过RBF核计算节点特征之间的相似度分数，然后用与相似度分数负相关的概率丢弃边缘。所有其他设置与我们在第5节中的实验保持一致。从表6可以看出，有偏见的版本只在Citeseer上比我们的无偏见方法表现得略好，但在其他情况下总是更差。这一观察结果支持了无偏取样在总体上的优越性。我们非常感谢审稿人的建议，并在第5.3节中加入了相应的讨论，使我们的论文更加完善。

6 结论

我们分析了4种流行的GCN模型的过度平滑的普遍过程，包括通用GCN、带偏置的GCN、ResGCN和APPNP。根据我们的分析，我们提出了DropEdge，一种新的和有效的技术来促进通用GCN的发展。在Cora、Citeseer、Pubmed和Reddit上进行的大量实验证明，DropEdge可以普遍和持续地提高当前流行的GCN的性能。请注意，从浅层2层模型的实验改进到

深度模型（带DropEdge）并不大，但仍然有意义。例如，在Pubmed上，带有DropEdge的8层GCN（91.0%）比没有DropEdge的2层GCN（89.5%）高出1.5%，这被认为是对这个基准的显著改善。更重要的是，除了实验上的改进，我们还从理论上解释了为什么深度GCN会失败，以及一般GCN模型是如何发生过度平滑的。即使我们没有像在其他领域（如图像上的CNN）那样从深度GCN中获得足够的好处，本文的理论分析和实验评估仍然是有价值的，可以促进未来在图学习方面更广泛的工作。

参考文献

[1] T.N. Kipf和M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the International Conference on Learning Representations*, 2017.

[2] H.Ren, W. Hu, and J. Leskovec, "Query2box:Reasoning over knowledge graphs in vector space using box embeddings," in *International Conference on Learning Representations*, 2019.

[3] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T.Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, 2015, pp. 2224-2232.

[4] D.Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410-5419.

[5] P.W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V.Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R.Faulkner *et al.* , "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[6] Z.Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *arXiv preprint arXiv:1901.00596*, 2019.

[7] B.Weisfeiler 和 A. A. Lehman, "A reduction of a graph to a canonical form and an algebra arising during this reduction," *Nauchno-Techicheskaya Informatsia*, vol. 2, no. 9, pp.

[8] K.Xu, W. Hu, J. Leskovec, and S. Jegelka, "图神经网络有多强大？" *arXiv preprint arXiv:1810.00826*, 2018.

[9] J.Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *Proceedings of International Conference on Learning Representations*, 2013.

[10] J.Chen, T. Ma, and C. Xiao, "Fastgcn:Fast learning with graph convolutional networks via importance sampling," in *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[11] W.Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, 2017, pp. 1025-1035.

[12] J.Klicpera, A. Bojchevski, and S. Günnemann, "Predict then propagate: 图形神经网络满足个性化的Pagerank，"在2019年*第七届学习代表国际会议论文集- tations*。

[13] W.Huang, T. Zhang, Y. Rong, and J. Huang, "Adaptive sampling towards fast graph representation learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 4558-4567.

[14] K.Xu, C. Li, Y. Tian, T. Sonobe, K.-i.Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," *arXiv preprint arXiv:1806.03536*, 2018.

[15] N.Dehmamy, A.-L. Barabási, and R. Yu, "Understanding the representation power of graph neural networks in learning graph topology," in *Advances in Neural Information Processing Systems*, 2019, pp.15 413-15 423.

[16] A.Loukas, "What graph neural networks cannot learn: depth vs width," in *International Conference on Learning Representations*, 2019.

[17] Q.Li, Z. Han, and X.-M.Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] K.Xu, C. Li, Y. Tian, T. Sonobe, K.-i.Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[19] G. Li, M. Müller, A. Thabet, and B. Ghanem, "Deepgcns:gcns能和cnns一样深入吗？"在*国际计算机视觉会议上*，2019。

[20] K.Oono和T. Suzuki, "图神经网络在节点分类中指数性地失去了表达能力"，在*国际学习代表会议上*，2020年。

[21] P.Erdos, A. Rényi *et al.* , "On the evolution of random graphs," *Publ.Math.Inst. Hung.Acad.科学*，第5卷，第1期，第17-60页，1960。

[22] J.Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso, " *Biostatistics*, vol. 9, no.3, pp. 432-441, 2008.

[23] M.Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems*, 2016, pp. 3844- 3852.

[24] M.Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[25] R.Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks, " in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[26] R.Levie, F. Monti, X. Bresson, and M. M. Bronstein, "Cayleynets:图卷积神经网络与复杂的有理频谱滤波器，"*IEEE信号处理的反应*，第67卷，第1期。pp.97-109, 2017.

[27] F.Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model cnns, " in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp.

[28] M.Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *International conference on machine learning*, 2016, pp.2014-2023.

[29] H.Gao, Z. Wang, and S. Ji, "Large-scale learnable graph convolutional networks," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.ACM，2018，第1416-1424页。

[30] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, "Cluster-gcn:An efficient algorithm for training deep and large graph convolutional networks," in *KDD*, 2019, pp.257-266.

[31] H.Zeng, H. Zhou, A. Srivastava, R. Kannan, and V. Prasanna, "Graphsaint:基于图形采样的归纳学习方法，" *ICLR*, 2020。

[32] P.Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y.Bengio, "Graph attention networks," in *ICLR*, 2018.

[33] M.Liu, H. Gao, and S. Ji, "Towards deeper graph neural networks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp.338-348.

[34] D.Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun, "Measure and relieving the over-smoothing problem for graph neural networks from the topological view." in *AAAI*, 2020, pp.3438-3445.

[35] L.Zhao and L. Akoglu, "Pairnorm:Tackling oversmoothing in gnns," in *International Conference on Learning Representations*, 2019.

[36] M.Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "简单和深度图卷积网络"。

[37] H.E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under structural and laplacian constraints," *arXiv preprint arXiv:1611.05181*, 2016.

[38] S.U. Pillai, T. Suel, and S. Cha, "The perron-frobenius theorem: some of its applications," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp.62-75, 2005.

[39] F.R. Chung and F. C. Graham, *Spectral graph theory*.American Mathematical Soc., 1997, no. 92.

[40] K.He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp.770-778.

[41] Y.Rong, W. Huang, T. Xu, and J. Huang, "Dropedge:Towards deep graph convolutional networks on node classification," in *International Conference on Learning Representations*, 2019.

[42] A.Ortega, P. Frossard, J. Kovacˇevic´, J. M. Moura, and P. Vandergheynst, "图信号处理。概述、挑战和应用"，*《IEEE论文集》*，第106卷，第。5, pp. 808-828, 2018.

[43] M.Vetterli, J. Kovacˇevic´, and V. K. Goyal, *Foundations of signal processing*.剑桥大学出版社，2014年。

[44] C.F. Van Loan和G. Golub, "Matrix computations (Johns Hopkins studies in mathematical sciences), 1996.

[45] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "通过防止共存问题来改进神经网络

特征检测器的适应性，" *arXiv预印本arXiv:1207.0580*， 2012。
[46] D.Eppstein, Z. Galil, G. F. Italiano, and A. Nissenzweig, "Sparsification-a technique for speeding up dynamic graph algorithms," *Journal of the ACM（JACM）*, vol. 44, no.5, pp. 669-696, 1997.
[47] P.Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data, " *AI magazine*, vol. 29, no.3, p. 93, 2008.
[48] F.Fouss, K. Fran?oisse, L. Yen, A. Pirotte, and M. Saerens, "An experimental investigation of graph kernels on a collaborative rec- ommendation task," in *Proceedings of the 6th International Conference on Data Mining (ICDM 2006*, 2006, pp. 863-868.
[49] J.Atwood 和 D. Towsley, "Diffusion-convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp.1993-2001.
[50] W.Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, 2017, pp.1024-1034.

黄文兵现在是清华大学计算机科学与技术系的助理研究员。他于2017年在清华大学获得计算机科学与技术专业博士学位。他目前的研究主要在机器学习、计算机视觉和机器人领域，特别是在不规则结构的学习方面，包括图形和视频。他已经发表了约30篇经同行评审的顶级会议和期刊论文，包括《NeurIPS论文集》。ICLR、ICML、CVPR等。他曾（将）担任AAAI 2021的高级程序委员会，ACMMMM研讨会HUMA 2020的区域主席，以及IJCAI 2019的会议主席。

于荣是腾讯AI实验室机器学习中心的高级研究员。他于2016年获得香港中文大学博士学位。他于2017年6月加入腾讯AI实验室。他的主要研究兴趣包括社会网络分析，图神经网络，以及大规模图系统。在腾讯AI实验室，他致力于构建大规模图学习框架，并将深度图学习模型应用于各种应用，如ADMET预测和恶意检测。他在数据挖掘、机器学习等顶级会议上发表了多篇论文，包括KDD、WWW、NeurIPS、ICLR、CVPR、ICCV等会议记录。

徐廷阳，腾讯AI实验室机器学习中心高级研究员。2017年在美国康涅狄格大学获得博士学位，2017年7月加入腾讯AI实验室。在腾讯AI实验室，他从事深度图学习、图代以及将深度图学习模型应用于各种应用，如分子生成和谣言检测。他的主要研究兴趣包括社会网络分析、图神经网络和图代，特别是专注于为分子代设计深度和复杂的图学习模型。他发表了多篇关于数据挖掘和机器学习的论文，包括KDD、WWW、NeurIPS、ICLR、CVPR、ICML等顶级会议。

孙福春，IEEE Fellow，1997年在中国北京清华大学获得计算机科学与技术博士学位。现任清华大学计算机科学与技术系教授，系学术委员会主任，中国北京智能技术与系统国家重点实验室副主任。他的研究兴趣包括告诉控制和机器人学，人工认知系统的信息传感和处理，以及其他方面。网络化控制系统。他在2006年被中国自然科学基金会认定为杰出青年学者。2006年，他成为IEEE控制系统协会智能控制技术委员会的成员。他担任《*认知计算与系统*国际期刊》的主编，以及一系列国际期刊的副主编，包括IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS，IEEE TRANSACTIONS ON FUZZY SYSTEMS，以及IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS。系统。

黄俊洲是德克萨斯大学阿灵顿分校计算机科学和工程系的副教授。他还担任过腾讯人工智能实验室机器学习中心的主任。他在中国武汉的华中科技大学获得工学学士学位，在中国北京的中国科学院自动化研究所获得硕士学位，并在罗格斯新泽西州立大学获得计算机科学博士学位。他的主要研究方向是包括机器学习、计算机视觉和成像信息学。他在2010年被IBM T.J. Watson研究中心选为多媒体和信号处理领域的10位新兴领导人之一。他的工作获得了2010年MICCAI青年科学家奖、2011年FIMH最佳论文奖、2011年MICCAI青年科学家奖入围奖、2012年STMI最佳论文奖、2013年NIPS最佳评论员奖、2014年MICCAI最佳学生论文奖入围奖和2015年MICCAI最佳学生论文奖。他在2016年获得了美国国家科学基金会的CAREER奖。