图卷积网络与自注意机制的对比分析

蒋浩泉[1,2], 张儒清[1,2], 郭嘉丰[1,2], 范意兴[1,2], 程学旗[1,2]
（1. 中国科学院 计算技术研究所 网络数据科学与技术重点实验室，北京 100190；
2. 中国科学院大学，北京 100049）

论文摘要

图卷积网络近来受到大量关注；同时，自注意机制作为**Transformer**结构及众多预训练模型的核心之一也得到了广泛运用。该文从原理上分析发现，自注意机制可视为图卷积网络的一种泛化形式，其以所有输入样本为节点，构建有向全连接图进行卷积，且节点间连边权重可学。多个文本分类数据集上的对比实验一致显示，使用自注意机制的模型较使用图卷积网络的对照模型分类效果更佳，甚至超过了目前图卷积网络用于文本分类任务的最先进水平。并且随着数据规模的增大，两者性能差距也随之扩大。这些证据表明，自注意机制更具表达能力，或可替代图卷积网络，带来潜在的性能提升。

论文简介

深度学习可被纳入机器学习中表示学习（**representation learning**）的研究范畴，相对传统特征工程需要人工设计特征，它能够利用神经网络技术、根据设定的任务目标自动学习得到输入及中间环节良好的分布式向量表示。多层感知机（**Multi-Layer Perceptron**）、卷积神经网络（**Convolutional Neural Networks，CNN**）和循环神经网络（**Recurrent Neural Networks，RNN**）是此前最常用的神经网络结构，被广泛用于自然语言处理、计算机视觉、语音识别等领域文本、图像及语音等信息的表示。

随着深度学习研究的发展，近期图卷积网络（**Graph Convolutional Networks，GCN**）和自注意（**self-attention**）机制获得了机器学习从业者们的大量关注。两者作为深度学习领域的最新进展，近期众多相关研究成果纷纷涌现。

目前最常采用的图卷积网络形式作为谱图卷积的局部一阶近似被提出，是一种简单而有效的图神经网络（**Graph Neural Networks，GNN**）。相比传统神经网络只能用于一般如文本序列、图像栅格等的网格状数据，图神经网络能够对非欧氏度量空间的数据进行建模。图卷积网络是传统神经网络中卷积神经网络结构在图数据上的推广，它本质是一种可直接作用于图上的多层神经网络。图卷积网络基于每个节点的邻居节点生成该节点的嵌入向量表示，该嵌入表示能将局部的图结构以及临近节点的特征信息编码入其中。图卷积网络通过一层卷积操作只能获取到其直接邻居节点的信息，而通过多层图卷积网络堆叠，就能整合更大范围的临近信息。

自注意机制是一种特殊的注意力（**attention**）机制。注意力机制现已成为神经网络领最重要的概念之一，它使得神经网络模型能够根据自身需求灵活自动地关注输入数据或特征中重要的部分，极大提高了模型的表达能力。注意力机制最初是与编码器-解码器（**encoder-decoder**）架构相结合被用于机器翻译领域，其要求输入与输出都是一个序列。但是对于诸如文本分类等任务，其输入是一个序列，而输出并不是序列的形式。故自注意机制的思想被提出，直接在一个序列内部实现注意力机制的运用。此后，**Transformer**架构横空出世，为自然语言处理领域预训练模型研究的热潮奠定了基础，而这些模型的共同核心思想之一就是自注意机制。

在将图卷积网络与自注意机制在工作原理上进行对比分析后，我们发现两者在形式上极其相似。一次图卷积操作和一次自注意步骤之间的区别仅在于对表示节点间连接关系的邻接矩阵的计算方式。从某种角度上我们可以认为，自注意机制也以其所有输入样本为节点，构建了一个全连接的图。并且，图卷积网络中用于表示节点间关系的邻接矩阵往往是训练之前人为预先给定的，而自注意机制中与之相对应表示节点之间连接关系的矩阵则是由可学习的参数根据不同任务特性学习得到，亦即各节点之间连边的权重甚至图的结构（连边权重为**0**则代表无连接，否则代表有连接）都是由可学习的参数根据不同任务的学习目标自适应地决定的。因此，自注意机制在某种程度上可以说是图卷积网络的一种泛化，具有较图卷积网络更强的表达能力。我们进一步推测，相比图卷积网络，在实际任务中使用更具表达能力的自注意机制将可能带来性能上的提升。

为了在实际任务中验证上述想法，我们选择了自然语言处理最基本问题之一的文本分类任务，将其作为代表进行对比实验。最终的实验结果显示，在相对对等的条件下，采用自注意机制的模型在多个文本分类数据集上的表现都显著优于使用图卷积网络的对照模型，甚至超过了目前图卷积网络用于文本分类任务的最先进水平。除此之外我们还观察到，随着数据规模的增加，使用自注意机制的模型和使用图卷积网络模型之间的性能差距也逐渐扩大。这些结果从实际任务表现方面，验证了自注意机制确实具有较图卷积网络更强表达能力的观点。

本文第**1**节将简要介绍图卷积网络和自注意机制的一些相关工作。第**2**节对比图卷积网络与自注意机制的工作原理，证明后者在某种程度上可被视为前者的一种泛化。第**3**节通过文本分类任务对比实验的实际表现，验证自注意机制较图卷积网络具有更强的表达能力，可获得更佳性能。最后，我们对本文进行总结，并提供一些未来可能进行的工作和研究方向。

模型分析

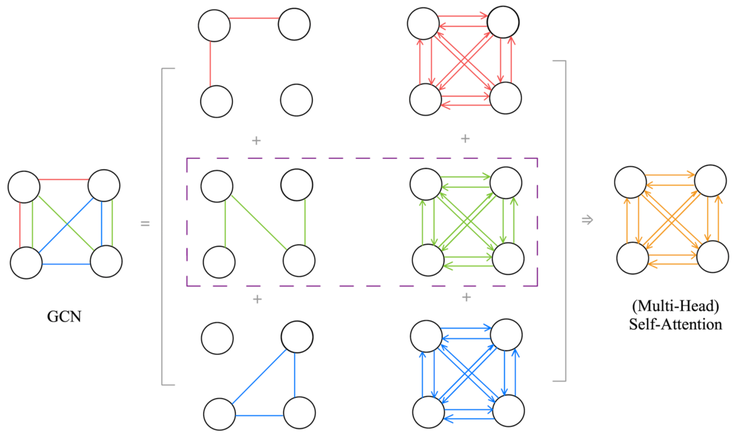

图 1　图卷积网络（左）与（多头）自注意机制（右）的对比示意图

引入处理连边方向和标签种类机制的图卷积网络的工作原理可大致由图**1**的左半部分示意。由于不同指向不同类型的连边其节点变换参数矩阵也不相同，因此总的图卷积网络可按照不同方向不同类型边拆分成多个子图分别进行图卷积，最后再将所有子图相加得到最终新的节点表示。为了简化表示，这里忽略了边的方向，只按照边的类型即三种不同颜色对图卷积网络进行了分解。

由于篇幅限制，相应公式请见原论文。对比代表图卷积网络原理的公式**(2)(4)**和代表自注意机制原理的公式**(6)(9)(8)**，我们已经可以非常清晰地看出图卷积网络和自注意机制在形式上的相似性。通过对比在形式上完全一致的公式**(2)**和公式**(6)**，可以看出图卷积网络和自注意机制都是首先对所有输入样本的特征表示向量组成的矩阵**H**进行一次线性变换，之后再利用各个样本变换后的特征表示向量加权求和得到各样本新的特征表示向量，而加权求和的权重则由矩阵中对应位置的数值决定。既然如此，那么参考图卷积网络的解释，从某种角度上我们可以认为，自注意机制也以其所有输入样本为节点，构建了一个图进行卷积操作。但由于代表图卷积网络的公式**(2)**和自注意机制**(6)**中对邻接矩阵的定义不尽相同，因此两者所构建图各节点之间的连接方式是不同的。

首先，从对原始邻接矩阵的归一化方式上看，图卷积网络与自注意机制存在细微差异，这也导致了两者所构建图连边的类型不同。通过表示两者对原始邻接矩阵的归一化的公式**(4)**和公式**(9)**可以看出，图卷积网络采用的是对称归一化，而自注意机制采用的则是行归一化。因此，图卷积网络所构建的是一个无向图，而自注意机制构建的则是一个有向图。

其次，相比图卷积网络中原始邻接矩阵**A**往往是训练之前人为预先给定的，自注意机制中与之相对应表示节点之间连接关系的矩阵则是由可学习的参数根据不同任务特性学习得到，亦即各节点之间连边的权重甚至图的结构（连边权重为**0**则代表无连接，否则代表有连接）都是由可学习的参数根据不同任务的学习目标自适应地决定的。

我们可以通过图**1**中央紫色虚线方框部分清晰地看到，单层图卷积网络与单头注意力机制之间的关系。左侧的图卷积网络构建了一个无向图，且节点之间的连接关系及权重是人为预先给定的。而自注意机制则是构建了一个有向全连接图，其各节点之间连边的权重甚至图的结构都是由可学习的参数根据不同任务的学习目标自适应地决定的。图中只表示了结构，没有表示边的相对权重大小。

即使是对于引入处理连边方向和标签种类机制的图卷积网络，通过图**1**我们也可以直观地看到，多头自注意机制也能通过多个并行的自注意模块以及可学习的参数对图卷积网络进行表达。实际上，每个自注意头并不一定如图中所示是一一对应。因为相比于左侧图卷积网络中边的类型需要提前定义，自注意机制则可以更具不同任务目标、通过可学习的参数自动学习捕获到模型各部分的良好表示，而这并不一定与前者预先定义的一致。

试想在极端的情况下，如果完全符合任务的目标，那么自注意机制中的矩阵**A**也可通过可学习参数的改变表示为与图卷积网络中邻接矩阵**A**数值相等。由此可以看出，可学习参数的引入，使得自注意机制较图卷积网络具有更强的模型表达能力。换而言之，在某种程度上，我们可以说自注意机制实际上是图卷积网络的一种泛化。

实验分析

我们和多个图卷积网络相关的文本分类基准模型进行比较，为了对图卷积网络和自注意机制进行较为公平的对比，我们还设计了实验组。我们在三个被广泛使用的文本分类数据集上进行了对比实验，即**R8**、**R52**，以及**Ohsumed**。详细的模型设计及试验结果数据请参见原论文。

对比代表图卷积网络的**GCN**模型和代表自注意机制的**SA**模型的分类准确率我们可以看到，使用自注意机制的模型在多个文本分类数据集上表现都明显超越了使用图卷积网络的模型，且存在较大差距。**GCN**模型相对**SA**模型的区别仅在于：**GCN**模型的邻接矩阵计算方式即节点间的连接关系及权重是提前根据经验人为设计并固定的，而**SA**模型由于采用的是自注意机制，相对应表示节点之间连接关系的矩阵则是由可学习的参数根据不同任务特性学习得到，亦即各节点之间连边的权重甚至图的结构都是由可学习的参数根据不同任务的学习目标自适应地决定的。该现象表明，可学习参数的引入，使得自注意机制较图卷积网络具有更强的模型表达能力。

继续观察**GCN**和**SA**这两组实验，我们发现随着数据集的规模的增大，两模型的性能差距也相应扩大。如图**2**所示，横轴依次代表的是**R8**、**R52**以及**Ohsumed**数据集构图所含的节点总数依次增大，纵轴表示分类准确率，绿色部分即表示了使用自注意机制的**SA**模型与使用图卷积网络的**GCN**模型之间的分类准确率差异。可以直观地看到，随着代表数据集规模的构图节点数目的增加，绿色部分的纵向距离不断增大。这说明，随着数据规模的增加，更具表达泛化能力的自注意机制能够更好地捕获并表示数据中与任务目标一致的特征，从而取得较图卷积网络更优异的性能表现。

我们还发现，使用自注意机制的**SA**模型在三个文本分类数据集上的分类准确率也都显著优于目前使用图卷积网络最先进的文本分类模型**Text-GCN**（使用学生氏**T**检验，**p<0.05**的条件下）。在**R52**和**Ohsumed**两个数据集上更是超过了**Conv-GCN**，达到了目前最先进的文本分类水平。两者在**R8**数据集上的表现也相当，可归因为参数优化上的随机性。综合之前的分析结果，这些证据表明自注意机制更具表达能力，或可替代图卷积网络，带来潜在的性能提升。

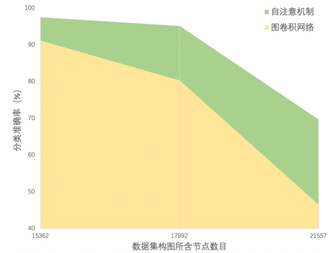

图 2　分类准确率差距与数据集规模的关系

论文结论

本文对深度学习的两项前沿技术即图卷积网络和自注意机制进行了比较研究，从原理上分析发现，自注意机制可视为图卷积网络的一种泛化形式，具有较图卷积网络更强的表达能力。自注意机制可认为也以所有输入样本为节点，构建有向全连接图进行卷积。并且，图卷积网络中用于表示节点间关系的邻接矩阵往往是训练之前人为预先给定的，而自注意机制中与之相对应表示节点之间连接关系的矩阵则是由可学习的参数根据不同任务特性学习得到，亦即各节点之间连边的权重甚至图的结构都是由可学习的参数根据不同任务的学习目标自适应地决定的。最后，在多个文本分类数据集上进行了图卷积网络与自注意机制的对比实验。结果显示，使用自注意机制的模型较使用图卷积网络的对照模型分类效果更佳，甚至超过了目前图卷积网络用于文本分类任务的最先进水平。并且随着数据规模的增大，两者性能差距也随之扩大。这些证据表明自注意机制更具表达能力，或可替代图卷积网络，带来潜在的性能提升。

需要指出，本文对于图卷积网络和自注意机制的讨论实际上是限制在自然语言处理领域内的。除本文的内容之外，还有许多角度可以对两者之间的关系进行探讨，例如在非欧式度量空间的数据上两者表现的比较、两者的计算代价等，还可以利用可视化的方法在原理上探索两者表现差异的缘由。接下来的工作可以尝试在其他更多任务上进行图卷积网络与自注意机制的对比研究和实验，例如机器翻译、文本摘要等，甚至是其他领域的任务。考虑到所采用技术的相似性，也可以进一步对融入了注意力机制的许多图神经网络和自注意机制进行对比研究。

西安电子科技大学 XIDIAN UNIVERSITY