GNNExplainer

本文是对GNNE的精读，本人能力有限，请多多指点。

> 原论文名：《GNNExplainer:Generating Explanations for Graph Neural Networks》
>
> 作者：Rex Ying Dylan Bourgeois Jiaxuan You Marinka Zitnik Jure Leskovec
>
> 时间:2019
>
> 期刊会议：NeurIPS

Abstract

GNN虽然是个处理图上的机器学习问题非常强大的工具，但因为同时结合结点特征和图的结构会形成非常复杂的模型。这种复杂性导致GNN做的预测非常难以解释。本文提出的GNNE可以识别出一个紧凑的子图结构和一部分在GNN预测中起到至关重要作用的结点特征，还可以对一整个实例类作出简洁的解释。

我们把GNNE表示为一个优化任务，该任务最大化GNN预测和可能子图结构分布之间的相互信息。GNNE有一系列优点，比如可视化语义相关结构以实现可解释性，给予我们深入洞察GNN的缺陷的机会等。

Introduction

其他的神经网络模型的可解释性工作是建立在模型的相关特征上的检查，和找到高层特征的良好定性解释，或者识别具有影响力的输入实例。这些方法在整合关系信息方面存在缺陷，但关系信息是图的本质，想要对图做解释，必须充分利用丰富的关系信息和结点特征。

GNNE将解释定义为GNN所训练的整个图的一个丰富子图，该子图最大化与GNN的预测之间的互信息。为了实现这一目标，通过构建一个均场变分近似（mean field variational approximation）并学习一个实值图掩码(graph mask)，选择GNN计算图的重要子图。同时，GNNEXPLAINER还学习一个特征掩码(feature mask)，用于屏蔽不重要的节点特征。

Related work

我们把非图神经网络的可解释性分为两大类。

1.**构建完整神经网络的简化代理模型**

通过与模型不相关的方式，围绕预测去学习一个局部可信的近似，比如通过线性模型或一系列规则，来对预测进行充分的表征。

问题是：图的关系信息不能仅仅被线性模型表示

2.从计算方面入手

比如特征梯度，神经元对输入特征的反向传播，和反事实推理等等。

问题是：

所产生的显著图在某些实例上具有误导性，并且容易产生梯度饱和。这些问题在图神经网络上更为严重，因为图的邻接矩阵作为离散输入，梯度值非常大，且只在一个小区间变化。

GAT也许可以用来增强解释性，但是因为注意力系数对于所有结点的预测都是相同的，所以与很多场景矛盾，比如该边对于预测一个结点的标签重要，但也许对预测另一个就不重要。

Formulating explanations for GNN

首先，设定G为有边集E和定点集V的图，其中结点特征有d维 $\chi = \{x_1, ..., x_n\}, x_i \in \mathbb{R}^d$ 。 f 是一个在结点上的标签函数，满足 $f : V \mapsto \{1, ..., C\}$,将v中的每一个结点映射道C个类别上去。GNN模型Φ在训练集中的所有节点上进行优化，然后用于预测，即在新节点上近似 f 。

Background on GNN

GNN模型在 l 层的更新遵循3个关键计算，

1.计算对每对结点的神经元消息 m_{ij}^l

其中的MSG代表结点对的各自上一层的表示以及关系的综合函数，

$$m_{ij}^{l}=MSG\left(h_{i}^{l-1},h_{j}^{l-1},r_{ij}\right)$$

2.对每一个结点 v_i ,GNN通过聚合函数AGG来聚合全部邻居的消息

$$M_{i}^{l}=AGG\left(\left\{m_{ij}^{l}|v_{j}\in\mathcal{N}_{v_{i}}\right\}\right)$$

3.对得到的聚合信息 M_i^l 和 v_i 自己上一层的表示 $\boldsymbol{h}_i^{l-1}$ 结合得到这一层的表示

$$\boldsymbol{h}_{i}^{l}=\mathrm{U}_{PDATE}\left(M_{i}^{l},\boldsymbol{h}_{i}^{l-1}\right)$$

在经过 L 层之后 v_i 的最后embedding是 $\mathbf{z}_i=\mathbf{h}_i^L$,本文的GNNE模型可以对任何在以上基础上的GNN模型进行解释。

GNNExplainer:Problem formulation

GNN的预测是由图结构信息 $\mathrm{G_c}\left(v\right)$ 和结点特征信息 $X_{\mathrm{c}}\left(v\right)$ 共同决定的。那我们就只需要考虑这两个因素来进行解释即可，可知我们的GNN预测结果为 $\hat{y}=\boldsymbol{\Phi}\left(G_c\left(v\right),X_c\left(v\right)\right)$ ，GNNE生成的解释为 $\left(G_S,X_S^F\right)$ 。其中 G_S 是计算图的一部分子图， X_S 是 G_S 的相关特征， X_S^F 是对于结点特征中对于解释 $\hat{y}$ 最重要的一部分子集，由maskF来选取， $X_S^F=\left\{x_j^F|v_j\in G_S\right\}$

GNNExplainer

GNNEXPLAINER将通过识别计算图的子图和模型Φ预测中最具影响力的节点特征的子集来生成解释。

Single-instance explanations

对于一个结点v，我们的任务就是识别

$$G_S\subseteq G_C\text{ 和 }X_S^F=\left\{x_j^F|v_j\in G_S\right\}$$

如何找到最重要的结点特征呢？使用互信息指标MI来打造一个优化任务，

$$\max_{G_S}MI\left(Y,\left(G_S,X_S\right)\right)=H\left(Y\right)-H\left(Y|G=G_S,X=X_S\right)$$

MI量化了当结点v的计算图限制在 G_S 上，结点特征限制在 X_S 上的预测结果 $\hat{y}$ 的变化。例如，考虑这样一种情况： $v_j\subseteq G_C\left(v_i\right),v_j\neq v_i$ 。那么，如果从 $G_C\left(v_i\right)$ 中移除 v_j 会显著降低预测 $\hat{y}$ 的概率，那么节点 v_j 就是对于节点 v_i 的预测的一个良好的反事实解释。类似地，考虑这样一种情况： $\left(v_j,v_k\right)\in G_c\left(v_i\right),v_i,v_k\neq v_i$ 。那么，如果移除 v_j 和 v_k 之间的边会显著降低预测 $\hat{y}$ 的概率，那么这条边的缺失就是对于节点 v_i 的预测的一个良好的反事实解释。

因为 $H(Y)$ 在训练好的GNN的控制下是固定的，所以最大化任务同等于最小化条件熵 $H(Y|G=G_S, X=X_S)$,由信息熵公式表示，

$$H(Y|G=G_S, X=X_S) = -\mathbb{E}_{Y|G_S,X_S}\left[\log P_{\Phi}(Y|G=G_S, X=X_S)\right]$$

G_S 和 X_S 的选择会最大化 $\hat{y}$ 的概率，左式就会越小。我们设定 G_S 最多的结点为 K_M 个，为了限定邻居的影响。

GNNExplainer's optimization framework

因为 G_S 的个数太多，我们用分数邻接矩阵来表示它。

$$A_S \in [0,1]^{n\times n}$$

如果我们把 G_S 当做一个在G上的随机变量，则目标函数变成，

$$\min_{\mathcal{G}} \mathbb{E}_{G_S\sim\mathcal{G}} H(Y|G=G_S, X=X_S)$$

我们使用Jensen不等式和假设H是凸函数得到目标函数的上界为，

$$\min_{\mathcal{G}} H(Y|G=\mathbb{E}_{\mathcal{G}}[G_S], X=X_S)$$

> 凸函数

凸函数任意两点的割线位于函数图形上方

> Jensen不等式

任意点集 $\{x_i\}$ ，有 $\lambda_i \geqslant 0$且$\sum_i \lambda_i = 1$,有凸函数 $f(x)$ 满足，

$$f\left(\sum_{i=1}^{M}\lambda_i x_i\right) \leqslant \sum_{i=1}^{M}\lambda_i f(x_i)$$

意思所有采样点的加权和的函数值小于函数值的加权和

confussion：

In practice, due to the complexity of neural networks, the convexity assumption does not hold. However, experimentally, we found that minimizing this objective with regularization often leads to a local minimum corresponding to high-quality explanations.

为了便于估计 E_G ，把 $\mathcal{G}$ 这个分布分解为多元伯努利分布，使用平均场变分近似得到：

$$P_{\mathcal{G}}(G_S) = \prod\nolimits_{(j,k)\in G_C} A_S[j,k]$$

> 平均场变分近似

复杂的概率模型包含很多随机变量，目标计算后验分布，需要将这种分布分解成多个较小的因子，每个因子涉及一个或多个随机变量，引入变分参数调整因子分布的形状逼近真实后验分布，接下来就是迭代优化变分参数的问题。

对于 $E_{\mathcal{G}}\left(G_S\right)$ ，采用掩码 $M \in \mathbb{R}^{n \times n}$ 来实现，具体操作为

$$A_c \odot \sigma\left(M\right)$$

因为用户对模型为何被分类为某一个类别更加感兴趣，所以使用标签和模型预测的交叉熵目标函数更加合适，对目标的优化采用SGD，

$$\min_{M} -\sum_{c=1}^{C} \mathbb{I}\left[y=c\right] \log P_{\Phi}\left(Y=y|G=A_c \odot \sigma\left(M\right), X=X_c\right)$$

在计算的时候涉及阈值抹去M矩阵中非常小的数字，最终得到 G_S 作为对 v 结点的预测的解释

Joint learning of graph structural and node feature information

为了得到影响结点预测的最重要的结点特征，GNNE为其结点 v 学习一个选择器 F ，称为特征掩码，来选择重要的特征，

$$X_S^F = \left\{x_j^F | v_j \in G_S\right\}$$

$$x_j^F = \left[x_{j,t_1}, ..., x_{j,t_k}\right] for\ F_{ti} = 1$$

其中 $F \in \left\{0,1\right\}^d$ 是一个需要被学习的d维的特征掩码，于是总的目标函数考虑到结构解释和结点特征解释后如下，

$$\max_{G_S,F} MI\left(Y,\left(G_S,F\right)\right) = H\left(Y\right) - H\left(Y|G=G_S, X=X_S^F\right)$$

Learning binary feature selector F.

作者计算 $X_S^F = X_S \odot F$

一般情况下， F 中很小的值代表该特征被去掉并不影响预测结果的准确度。但也有一些情况导致预测忽略掉了值很小但却很重要的特征。为了在训练中全方位观察每一个特征的影响程度，在采样策略里使用蒙特卡洛采样法对所有的特征子集进行采样。

> 服从经验分布的Mont Carlo采样法

蒙特卡洛方法基于随机抽样原理，通过在设定的分布下生成大量随机样本来近似真实的概率分布，经验分布是由实际观测数据中得到的分布。

> Reparametrization trick

重参数化，因原模型的参数不好优化，采用另外的随机变量（一般是服从正态分布）来模拟该参数，可以表示为新参数=新参数+随机变量*一个函数。

重参数化之后的X变成了，

$$X = Z + (X_S - Z) \odot F s.t. \sum_j F_j \leq K_F$$

其中 Z 是d维的表示服从经验分布的随机变量， K_F 表示最大的特征选择数。

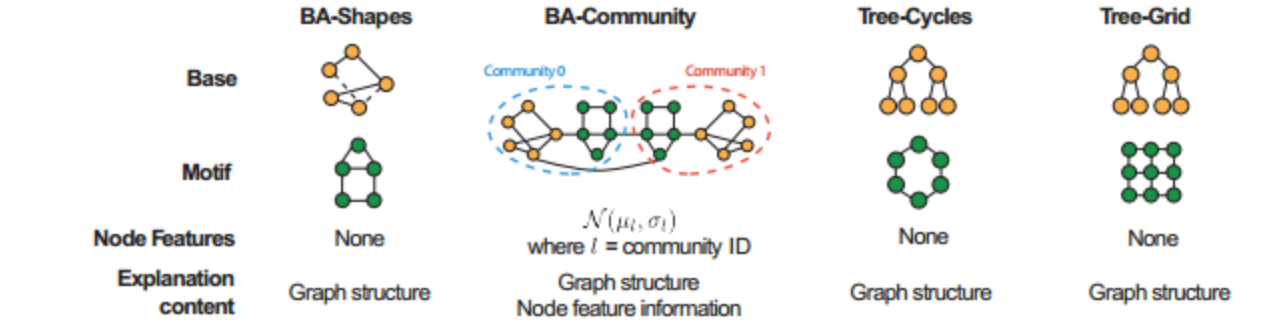
Integrating additional constraints into explanations

使用各种正则化技，比如交叉熵来使得学习到的掩码离散化（靠近0或者1），添加掩码全部参数的和来控制解释图不要太大等等。

Experiments

使用GNNE对GNN在节点分类和图分类任务进行解释。

Synthetic datasets



Explanation accuracy				
Att	0.815	0.739	0.824	0.612
Grad	0.882	0.750	0.905	0.667
GNNExplainer	**0.925**	**0.836**	**0.948**	**0.875**

Table 1: Illustration of synthetic datasets (refer to "Synthetic datasets" for details) together with performance evaluation of GNNEXPLAINER and alternative baseline explainability approaches.

数据集解读

BA-Shapes（节点分类）:将以5个节点形成的房子图案的Motif随机连接到300个结点形成的base图上形成大图，base上的节点标号为0，房子顶部为1，中间为2，底部为3

BA-Community(节点分类)：两个BA图合起来，节点特征服从高斯分布，根据不同社区，可以分为8个类

Tree-Cycles(节点分类)：以二叉树为base图（层数可控制），将6节点形成的循环图加上去形成大图

Tree-Grid(节点分类)：以二叉树为base图（层数可控制），将3x3的网格图加上去形成大图

Mutag（图分类)：4337个分子图组成的188个图，分子图可视碳环为base图，NO_2 和 NH_2 为附加图，含有附加图的为一类，不含的为一类

Reddit-Binary(图分类)：2000个帖子图，在帖子图中，节点为用户，用户对另一个用户进行评论则形成一条边。

解释结果

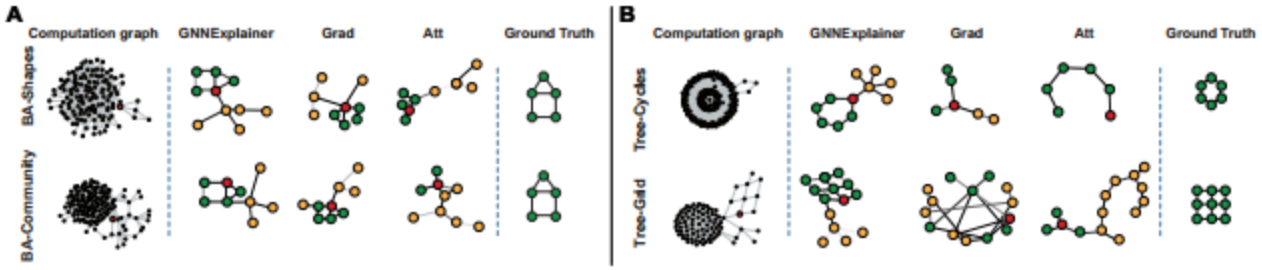


Figure 3: Evaluation of single-instance explanations. **A-B.** Shown are exemplar explanation subgraphs for node classification task on four synthetic datasets. Each method provides explanation for the red node's prediction.

在四个数据集上的节点分类解释中可以看到，对红色节点的预测解释，GNNE最为精准。

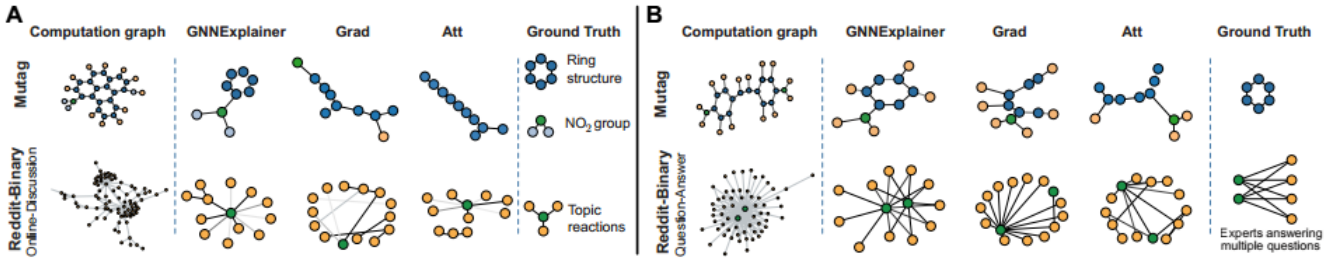
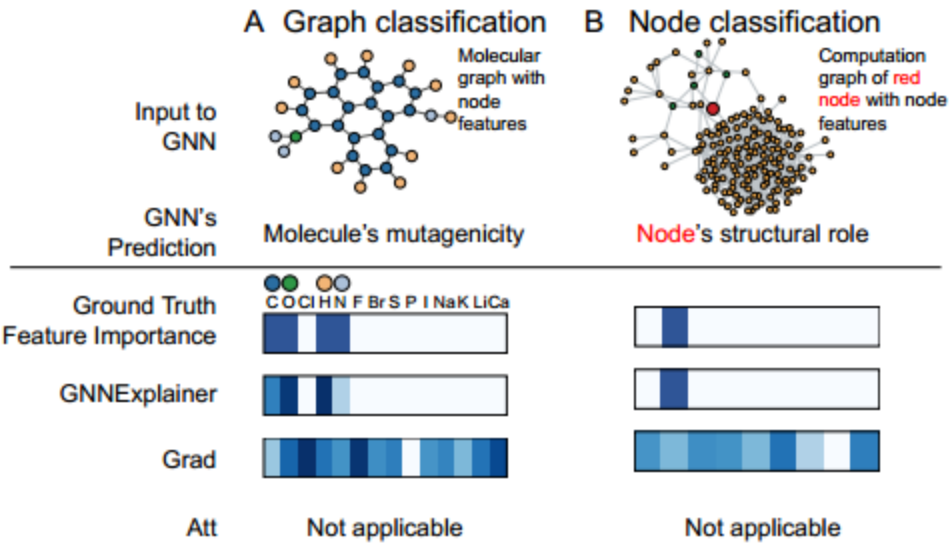


Figure 4: Evaluation of single-instance explanations. **A-B.** Shown are exemplar explanation subgraphs for graph classification task on two datasets, MUTAG and REDDIT-BINARY.

在两个数据集上的图分类的解释中，GNNE也最为精准



在对节点属性特征的重要性抓取上，GNNE也和真实解释最为接近

1.**Quantitative analysis**

在真实解释图中的边视为lable，模型解释图中的边如果在真实解释图中则拥有更高的分数。

2.**Qualitative analysis**

以上3张图可以作为定性分析的结果

Conclusion

本文引出了对任何GNN模型进行解释的GNNExplainer，最大化图的互信息提取图中最重要的节点特征和最重要的解释子图对预测结果进行解释。