计算机科学
COMPUTER SCIENCE

一种面向脑疾病诊断的图卷积网络对抗攻击方法

王晓明, 温旭云, 徐梦婷, 张道强

引用本文

王晓明, 温旭云, 徐梦婷, 张道强.一种面向脑疾病诊断的图卷积网络对抗攻击方法[J]. 计算机科学, 2022, 49(12): 340-345.

WANG Xiao-ming, WEN Xu-yun, XU Meng-ting, ZHANG Dao-qiang. Graph Convolutional Network Adversarial Attack Method for Brain Disease Diagnosis [J]. Computer Science, 2022, 49(12): 340-345.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500185

一种面向脑疾病诊断的图卷积网络对抗攻击方法

王晓明　温旭云　徐梦婷　张道强
南京航空航天大学计算机科学与技术学院　南京 211111
(leowxm@nuaa.edu.cn)

摘　要　近年来,利用静息态功能磁共振成像的脑功能网络分析已被广泛应用于各类脑疾病的计算机辅助诊断任务中。结合临床表型测量与脑功能网络构建的图卷积神经网络框架,提高了智能医学疾病诊断模型对现实世界的适用性。但是,基于脑功能网络的疾病诊断模型的可信度研究是一个重要但仍被广泛忽视的部分。对抗攻击技术在医疗机器学习中对模型的"欺骗"进一步引发了模型应用于临床实际中的安全与信任问题。基于此,在这项工作中,首次提出了一种面向脑疾病诊断的图卷积网络对抗攻击方法BFGCNattack,结合临床表型测量构建了疾病诊断模型,探索评估了智能诊断模型在面临对抗攻击时的鲁棒性。在自闭症脑成像数据集上的实验结果表明,使用图卷积网络构建的诊断模型在面临提出的对抗攻击时是脆弱的,即使只执行少量(10%)的扰动,模型的准确率和分类裕度均显著下降,同时愚弄率也显著提高。

关键词:对抗攻击方法;脑疾病诊断;图卷积网络;脑功能网络分析;模型鲁棒性
中图法分类号　TP181

Graph Convolutional Network Adversarial Attack Method for Brain Disease Diagnosis

WANG Xiao-ming,WEN Xu-yun,XU Meng-ting and ZHANG Dao-qiang
College of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211111,China

Abstract　In recent years,brain functional networks analysis using the resting state functional magnetic resonance imaging data has been widely used in computer-aided diagnosis tasks of various brain diseases. The graph convolutional network framework integrating clinical phenotypic measurements and brain functional networks improves the applicability of intelligent medical disease diagnosis models to the real world. However,the trustworthiness study is an important but still widely neglected component of disease diagnosis models based on brain functional networks. Adversarial attack techniques in medical machine learning can deceive models,which further leads to the security and trust issues of the model applied in clinical practice. Based on this,this paper proposes an adversarial attack method BFGCNattack on graph convolutional network for brain disease diagnosis,constructs a disease diagnosis model integrating clinical phenotypic measurements,and evaluates the robustness of brain functional networks-based disease diagnosis model in the face of adversarial attacks. Experimental results on the autism brain imaging data exchange dataset suggest that the models constructed using graph convolutional networks are vulnerable to the proposed adversarial attack. Even if only a small number(10%) of perturbations are performed,the model's accuracy and classification margin significantly decrease,while the fooling rate significantly increases.

Keywords　Adversarial attack method,Brain disease diagnosis,Graph convolutional network,Brain functional networks analysis,Model robustness

近年来,机器学习模型因其性能高超正被广泛应用于各种挑战性任务中。神经影像学研究也越来越多地转向机器学习方法,以检测微弱的、广泛分布的大脑回路和表型测量之间的关联。基于脑功能网络的脑疾病诊断模型处于这一趋势的前沿[1-2],在理解认知[3]、心理健康[4]以及疾病诊断[5]等方面显示了有前景的结果。基于脑功能网络的模型在准确性[6]和公平性[7]方面的提高,代表了实际应用这些模型到计算机辅助诊断任务上的重要一步。

然而,随着机器学习的发展,相应的风险也随之而来。机器学习模型中一个重要却被广泛忽视的部分是模型可信度。在这项工作中,我们将其定义为面对对抗攻击时的鲁棒性。对抗攻击指对输入数据添加微小扰动,旨在使训练良好的模型可以被对抗性输入轻易"欺骗"[8]。如今,对抗攻击已广泛存在于网络安全[9]、ImageNet分类挑战[10],甚至是医疗机器

到稿日期:2022-05-20　返修日期:2022-08-27
基金项目:国家自然科学基金(62136004,61876082,61732006);中央高校基本科研业务费专项资金(3082020NZ2020018)
This work was supported by the National Natural Science Foundation of China(62136004,61876082,61732006) and Fundamental Research Funds for the Central Universities(3082020NZ2020018).
通信作者:张道强(dqzhang@nuaa.edu.cn)

学习[11]中，这将在临床实践中造成严重的安全与信任问题。据我们所知，基于脑功能网络的脑疾病诊断模型可信度研究工作仍然很缺乏。

由此，本文的主要贡献包括 3 个方面：

(1)首次将对抗攻击应用到脑功能网络上，提出了一种面向脑疾病诊断的图卷积网络对抗攻击方法 BFGCNattack，评估了诊断模型面对对抗攻击时的鲁棒性。图 1 给出了本文方法的框架。

(2)在自闭症脑成像数据集[12]上结合表型测量与脑功能网络，构建了基于图卷积网络的疾病诊断模型，通过对脑功能网络中的边执行少量扰动(增加或删除)生成对抗样本，使用代理模型简化计算，最终实现了对目标脑网络的对抗性攻击。

(3)本文的实验研究表明，即使只执行少量的扰动，模型对目标脑网络分类的准确性和分类裕度均显著下降，对模型的愚弄率也显著提升，并且本文方法的攻击效果优于其他对比方法。

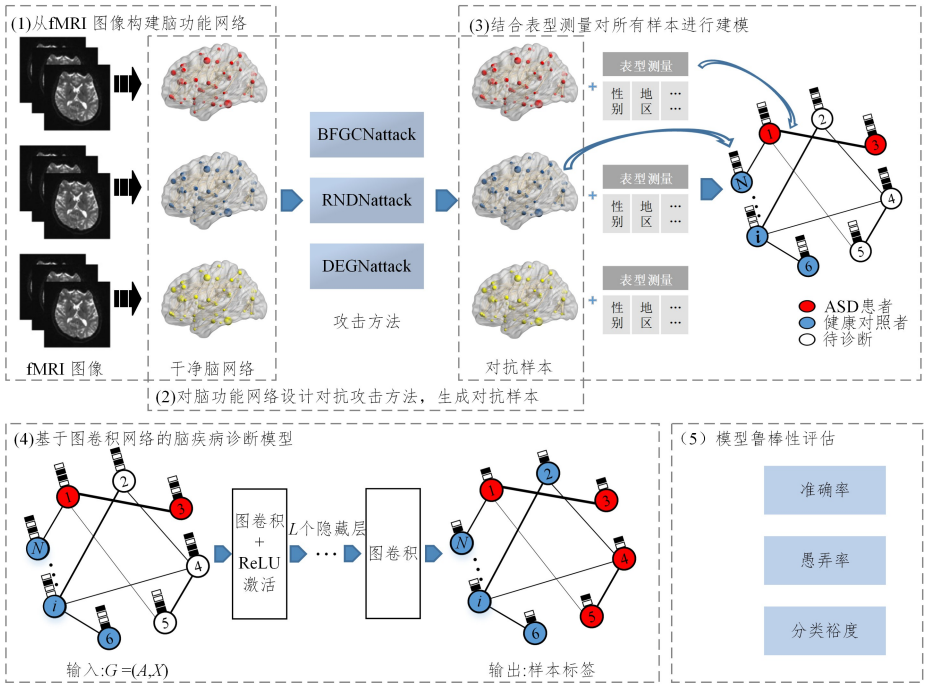


图 1 本文方法的总体框架

Fig. 1 Overall framework of this work's methods

1 相关工作

1.1 面向图的对抗攻击

与以往基于图像的对抗攻击相比，基于图的对抗攻击面临两个独特的挑战[13]：1)与由连续特征组成的图像不同，图的结构和节点特征是离散的，在离散空间中设计对抗攻击的有效算法是困难的；2)图像的对抗攻击被设计为人类无法察觉，可以强制一个特定的距离函数，例如对抗性与良性实例之间的 L_p 范数距离很小。然而，在图中，如何定义“不可察觉的”，需要进一步的分析与研究。

当前基于图的对抗攻击研究仍处于起步阶段。已有的攻击方法主要有边级别的扰动[14-15]、节点级别的扰动[16]以及节点属性级别的扰动[17]等。当前研究最多的是边级别的扰动，主要集中在对图中的边进行增加、删除、重连等。节点级别的扰动和节点属性级别的扰动的相关研究较少，主要集中在对节点进行增加、删除和更改节点属性等。关于攻击的任务，主要有节点相关的任务，如节点分类[14,16-17]、社区检测[18]、链路相关的任务(如链路预测[19])以及图相关的任务(如图聚类[20])。但是，目前的研究主要集中在节点相关的任务，对图相关的任务的研究较少。随着图神经网络的兴起与应用，当前基于图的对抗攻击的研究也较多集中在图神经网络的模型上，文献[21]证明了图神经网络模型在面对对抗扰动时的脆弱性。关于攻击的阶段，主要有投毒攻击(Poisoning Attacks)与规避攻击(Evasion Attacks)，两者的区别在于对图执行扰动后，投毒攻击是在被扰动的图上训练并测试模型，规避攻击则是对训练好的模型进行攻击，此时攻击者不能修改模型的参数或结构，而是用被扰动的图去测试模型。关于微小扰动的原则，在图结构中，如何定义“不可察觉的”，目前主要是通过设置攻击预算的方法来对扰动进行限制。

1.2 基于脑功能网络的疾病诊断研究进展

静息态功能磁共振成像(Resting State Functional Magnetic Resonance Imaging，rs-fMRI)是测量人类大脑在静息状态下自发神经活动的有力工具[22]。利用 rs-fMRI 数据构建脑功能网络可以揭示脑疾病的病理基础以及发现生物标志物[23]。基于 rs-fMRI 的脑功能网络分析已被广泛应用于各类脑疾病的计算机辅助诊断任务中。近年来，随着深度学习的发展，图卷积神经网络(Graph Convolutional Network，GCN)被逐渐用于分析大脑这种高度非线性结构的数据。例如，Ktena 等[24]用暹罗 GCN 学习脑功能网络之间的相似性度量，并将其用于自闭症谱系障碍疾病的诊断。

以往的方法大多只关注受试者之间的成对相似性。在真正的临床应用中，为了更加准确地对患者进行诊断，专家通常

会参考患者的临床表型测量，如年龄、性别、智商等。近年来，许多研究表明，临床表型测量有助于脑疾病诊断。Parisot等[25]提出了一种较为全面的通用框架，利用图卷积神经网络将种群表示为一个稀疏图，其中节点与基于图像的特征向量相关联，而表型测量集成为边权，改善了疾病诊断的结果。

2 面向脑疾病诊断的图卷积网络对抗攻击方法

2.1 疾病诊断模型的构建

2.1.1 从fMRI构建脑功能网络

我们选择用于连接组分析的可配置管道(Configurable Pipeline for the Analysis of Connectomes，C-PAC)处理fMRI，随后将其配准到标准解剖空间(MNI152)以允许跨受试者比较。使用哈佛-牛津分割模板(Harvard-Oxford，HO)划分了110个大脑区域，提取每组脑区的平均时间序列，并将其归一化。通过计算每个脑区时间序列之间的Fisher变换皮尔逊相关来估计受试者的脑功能网络连接矩阵。对连接矩阵进行Fisher变换以提高正态性。由于连接矩阵的高维性，我们使用与Parisot等[25]相同的降维策略，并加入了对脑网络比例阈值化的处理，比例取值为0.1。

2.1.2 结合表型测量对受试者进行建模

我们将所有的受试者建模在图 $\boldsymbol{G}=(\boldsymbol{A},\boldsymbol{X})$ 上，图中每个节点代表一个受试者。图1中(3)提供了模型构建的简单示例。这里要注意的两个关键问题是：1)定义图中节点的特征，我们提取相关矩阵的上三角元素作为特征向量，这种方法已经在基于fMRI的分类中取得了许多成功；2)定义图的边以及权值，我们将受试者脑网络之间的成对相似性与表型测量结合后获得边。考虑一组 H 个非成像表型测量 $\boldsymbol{M}=\{\boldsymbol{M}_h\}$，如性别、年龄。构建的邻接矩阵 $\boldsymbol{A}$ 定义为：

$$\boldsymbol{A}(v,w)=Sim(\boldsymbol{S}_v,\boldsymbol{S}_w)\sum_{h=1}^{H}\gamma(\boldsymbol{M}_h(v),\boldsymbol{M}_h(w)) \tag{1}$$

$$Sim(\boldsymbol{S}_v,\boldsymbol{S}_w)=\exp\left(-\frac{[\rho(x(v),x(w))]^2}{2\sigma^2}\right) \tag{2}$$

其中，$Sim(\cdot,\cdot)$ 是受试者脑网络之间的相似性度量，$\boldsymbol{S}_v$ 和 $\boldsymbol{S}_w$ 是样本 v 和样本 w 构建的脑网络，ρ 是相关距离，σ 决定了核的宽度。而 γ 是表型测量之间的度量，我们选择了性别和中心两种表型，因此将 γ 定义为克罗内克函数，这意味着如果受试者之间的性别或中心一致，则他们的边缘权重将增加。Parisot等[25]的实验也充分证明了图结构在结合性别、中心与受试者相似性3种指标时分类表现是最好的。

2.1.3 使用图卷积网络构建疾病诊断模型

我们使用带有一个隐藏层的GCN作为分类模型。考虑到脑功能网络复杂网络的特性，在滤波器的选择上，我们参考Defferrard等[26]与Parisot等[25]的方法，将滤波器类限制为多项式滤波器。

$$g_\theta(\Lambda)=\sum_{k=0}^{K}\theta_k\Lambda^k \tag{3}$$

由于 K 阶多项式滤波器是严格 K 局部化的，这种方法产生了在空间中严格局部化的滤波器，显著降低了卷积算子的复杂度。同样，我们使用切比雪夫多项式的截断展开逼近的方式递归计算该多项式。

2.2 对抗攻击方法设计

攻击者的目标是对图 $\boldsymbol{G}^{(0)}=(\boldsymbol{A},\boldsymbol{X}^{(0)})$ 中的一个目标节点 t 的脑网络执行简单的扰动，得到图 $\boldsymbol{G}'=(\boldsymbol{A},\boldsymbol{X}')$，从而实现误分类。为了确保攻击者不能完全修改脑网络，我们首先限制允许修改的次数为 Δ，即对于 t，有：

$$\sum_t\sum_i|\boldsymbol{X}_{ti}^{(0)}-\boldsymbol{X}'_{ti}|\leqslant\Delta \tag{4}$$

其中，i 表示特征矩阵 $\boldsymbol{X}$ 中对应于目标节点 t 所拥有的 i 个特征，即式(4)限制对目标节点 t 所有特征的修改小于 Δ。但是，我们认为在面向脑功能网络的对抗攻击设计时，只考虑限制允许修改的次数在“不可察觉的”这个关键问题上是不够的。图2总结了本文提出的BFGCNattack方法的示意图，主要包括对脑网络的边进行联合分布建模和对抗攻击方法的设计。

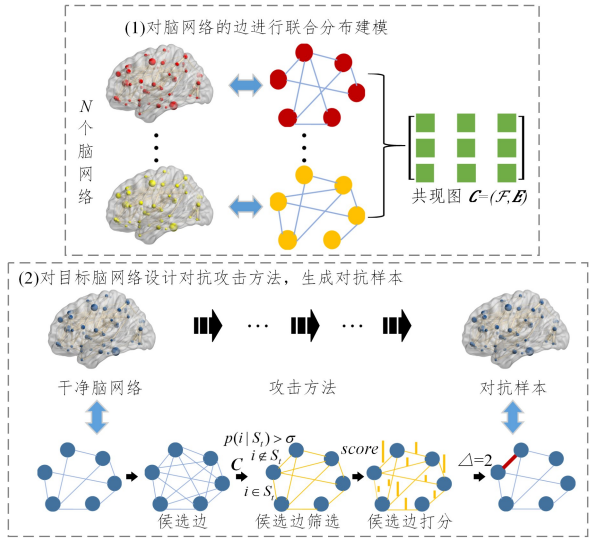


图2 提出的BFGCNattack方法的示意图

Fig. 2 Illustration of the proposed BFGCNattack

2.2.1 对脑网络的边进行联合分布建模

考虑到疾病诊断背景下，统计学的先验知识往往是重要的，因此我们对所有脑网络的边的联合分布进行建模。我们在图 $\boldsymbol{G}^{(0)}$ 上定义一个特征共现图 $\boldsymbol{C}=(\mathcal{F}\times\boldsymbol{E})$，其中 $\mathcal{F}$ 是特征集合，$\boldsymbol{E}\subseteq\mathcal{F}\times\mathcal{F}$ 被定义为目前为止所有脑网络中一起出现的边。定义一个随机漫步器，如果随机漫步器从 t 存在的边开始并执行一步随机游走到达边 i 的概率非常大，则添加 i 是不引人注意的。形式上，令 $S_t=\{j|\boldsymbol{X}_{tj}\neq0\}$ 为 t 最初存在的所有特征的集合。我们认为将特征 $i\notin S_t$ 添加到节点 t 是“不可察觉的”，当且仅当：

$$p(i|S_t)=\frac{1}{|S_t|}\sum_{j\in S_t}\frac{1}{d_j}\boldsymbol{E}_{ij}>\sigma \tag{5}$$

其中，d_j 为共现图 $\boldsymbol{C}$ 的度。在我们的实验中，将 σ 设置为最大可能达到的一半，即：

$$\sigma=0.5\cdot\frac{1}{|S_t|}\sum_{j\in S_t}\frac{1}{d_j} \tag{6}$$

因此，我们限制扰动图 $\boldsymbol{G}'=(\boldsymbol{A},\boldsymbol{X}')$ 还需要满足：

$$\forall t\in\mathcal{V},\forall i\in\mathcal{F}:\boldsymbol{X}'_{ti}=1\Rightarrow(i\in S_t\vee p(i|S_t)>\sigma) \tag{7}$$

其中，$\mathcal{V}$ 是图 G 中的节点集合，i 对应于特征集合中的特征，

$\boldsymbol{X}'_u$ 表示对目标 t 执行扰动后的特征矩阵。

因此，我们定义候选边集合 $\mathcal{P}_{\Delta}^{G^{(0)}}$ 为满足式(4)、式(7)的扰动图集合。这里，候选边集合指执行对抗攻击时可供选择的边。

2.2.2 对抗攻击方法设计

通过对脑网络的边进行联合分布建模，我们可以得到一个候选边集合，即满足 $\mathcal{P}_{\Delta}^{G^{(0)}}$ 的边。接下来，我们将基于模型损失对候选边做进一步选择，以执行攻击。

首先参考 Kipf 等[27]的方案，设计只有一个隐藏层的 GCN 为：

$$\boldsymbol{Z}=f_\theta(\boldsymbol{A},\boldsymbol{X})=\mathrm{softmax}(\hat{\boldsymbol{A}}\sigma(\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}^{(1)})\boldsymbol{W}^{(2)}) \tag{8}$$

其中，$\boldsymbol{A}$ 为邻接矩阵，$\boldsymbol{X}$ 为特征矩阵，$\hat{\boldsymbol{A}}=\tilde{\boldsymbol{D}}^{-1/2}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-1/2}$，$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N$，$\boldsymbol{I}_N$ 为单位矩阵，$\tilde{\boldsymbol{D}}_{ii}=\sum_j\tilde{\boldsymbol{A}}_{ij}$，$\sigma(\cdot)$是激活函数，$\boldsymbol{Z}_{vc}$ 是将节点 v 分配给类 c 的概率。另外，$\boldsymbol{W}^{(1)}$ 与 $\boldsymbol{W}^{(2)}$ 是可训练的权重矩阵，使用 θ 表示所有参数集合 $\theta=\{\boldsymbol{W}^{(1)},\boldsymbol{W}^{(2)}\}$。通过最小化标记样本集 $\mathcal{Y}$ 输出的交叉熵来得到最优参数。

$$L(\theta;\boldsymbol{A},\boldsymbol{X})=-\sum_{v\in\mathcal{Y}}\ln Z_{v,c_v},\boldsymbol{Z}=f_\theta(\boldsymbol{A},\boldsymbol{X}) \tag{9}$$

基于此，我们对脑功能网络的对抗攻击问题被定义为问题 1。

问题 1 给定图 $\boldsymbol{G}^{(0)}=(\boldsymbol{A},\boldsymbol{X}^{(0)})$，其中图 $\boldsymbol{G}^{(0)}$ 中每个顶点对应脑网络，图中的每条边对应脑网络间的相似度与表型测量计算，给定目标脑网络 t，给定攻击预算 Δ。令 c_{old} 为 t 的真实标签，使得：

$$\arg\max_{(\boldsymbol{A},\boldsymbol{X}')\in\mathcal{P}_{\Delta}^{\boldsymbol{G}^{(0)}}}\max_{c\neq c_{\mathrm{old}}}\ln\boldsymbol{Z}^*_{t,c}-\ln\boldsymbol{Z}^*_{t,c_{\mathrm{old}}}$$

$$\text{s.t. } \boldsymbol{Z}^*=f_{\theta^*}(\boldsymbol{A},\boldsymbol{X}') \text{ with } \theta^*=\arg\min_\theta L(\theta;\boldsymbol{A},\boldsymbol{X}') \tag{10}$$

其中，c 是样本的类别，$\boldsymbol{Z}=f_\theta(\boldsymbol{A},\boldsymbol{X}')$代表了模型的输出，而 $\boldsymbol{Z}^*$ 则对应于模型的最优参数 θ^* 下的输出。攻击者的目标是找到一个扰动图 $\boldsymbol{G}'=(\boldsymbol{A},\boldsymbol{X}')$，它将 t 分类为 c_{new}，与原类别 c_{old}之间对应的损失差值最大。需要注意的是，式(10) 引入了双层优化问题，与多数研究一致，我们选择了一种简单的变体，考虑规避攻击，假设参数是静态的并且基于旧图学习 $\theta^*=\arg\min_\theta L(\theta;\boldsymbol{A},\boldsymbol{X})$。另外，我们参考 Zügner 等[17]的方案，先攻击代理模型，从而导致被攻击的图求解。这个图随后被用来训练最终的模型。

于是，我们对式(8)进行线性化，得到如下的代理模型。

$$\boldsymbol{Z}'=\mathrm{softmax}(\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}^{(1)}\boldsymbol{W}^{(2)})=\mathrm{softmax}(\hat{\boldsymbol{A}}^2\boldsymbol{X}\boldsymbol{W}) \tag{11}$$

我们将对数概率简化为 $\hat{\boldsymbol{A}}^2\boldsymbol{X}\boldsymbol{W}$ 后，定义代理损失为：

$$\mathscr{L}_s(\boldsymbol{A},\boldsymbol{X};\boldsymbol{W},t)=\max_{c\neq c_{\mathrm{old}}}[\hat{\boldsymbol{A}}^2\boldsymbol{X}\boldsymbol{W}]_{tc}-[\hat{\boldsymbol{A}}^2\boldsymbol{X}\boldsymbol{W}]_{tc_{\mathrm{old}}} \tag{12}$$

最后我们使用一种贪婪求解的方案，在向任意脑网络执行攻击时，利用式(12)对满足 $\mathcal{P}_{\Delta}^{G^{(0)}}$ 的候选边进行打分 $score=\mathscr{L}_s(\boldsymbol{A},\boldsymbol{X};\boldsymbol{W},t)$，选择 $score$ 较高的边执行攻击。

3 实验

3.1 数据集

自闭症谱系障碍(Autism Spectrum Disorders，ASDs)由于其具有高患病率、终身性质、复杂性和异质性，是精神病学和神经科学面临的一个巨大挑战。我们在自闭症脑成像数据集(Autism Brain Imaging Data Exchange，ABIDE)上进行了实验。该数据集汇聚了国际不同采集站点的数据，公开共享了 1 112 名受试者的 fMRI 和表型数据。我们选择了 871 名符合成像质量和表型信息标准的受试者，包括 403 名 ASD 患者和 468 名健康对照者。

3.2 评价指标

本文采用模型对脑功能网络的分类准确率(Accuracy，Acc)、愚弄率(Fooling Rate，FR)和分类裕度(Classification Margin，CM)作为实验的主要评价指标。其中，愚弄率指当脑功能网络变成对抗样本时，模型对改变其预测结果的样本数与总数的比率，愚弄率越高，相应地，模型的鲁棒性表现就越差。分类裕度是模型对目标节点的真实标签的输出概率减去非真实标签中最大概率类别的概率，表示为：

$$CM=\boldsymbol{Z}^*_{t,c_{\mathrm{old}}}-\max_{c\neq c_{\mathrm{old}}}\boldsymbol{Z}^*_{t,c} \tag{13}$$

其中，c_{old}为 t 的真实标签，分类裕度越低则攻击能力越强。

3.3 实验设置

我们使用十折交叉验证对数据集进行划分训练与测试。对每次不同的划分都执行以下步骤：1)训练疾病诊断模型后，从测试集中被正确分类的节点中选择 5 个 CM 最大的节点、5 个 CM 最小的节点和随机 10 个其他节点；2)使用本文的攻击方法对目标脑网络进行扰动，将扰动后的脑网络重新传入疾病诊断模型。

在疾病诊断模型构建时，设置的主要参数为：$L=1$，$dropout\ rate=0.3$，$l2\ regularisation=5\times10^{-4}$，$learning\ rate=5\times10^{-3}$，$epochs=150$，$K=4$。

关于对抗攻击方法，我们设置了总边数的 0%，2.5%，5%，7.5%，10%等不同的预算次数分别进行实验。

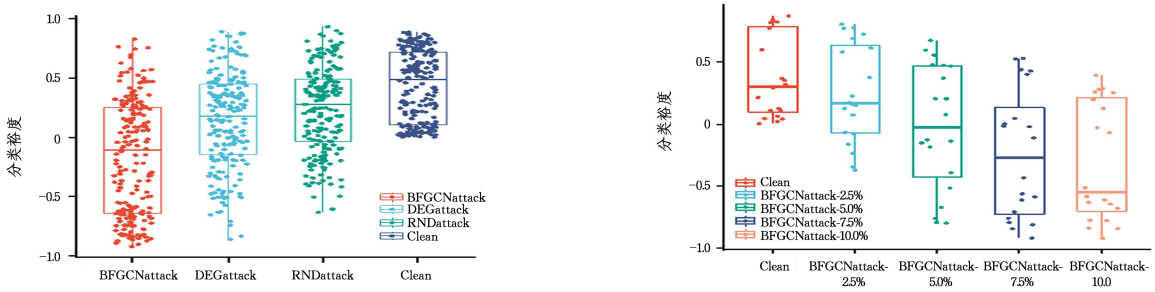
3.4 对比方法

由于当前基于图的对抗攻击仍处于起步阶段，并且这是首次将对抗攻击技术应用到基于脑功能网络的疾病诊断模型上，因此我们比较了以下两种对比方法：1)RNDattack[17]算法随机选择目标脑网络的边执行攻击；2)DEGattack[19]算法选择目标脑网络中度中心性较大的边执行攻击。

3.5 实验结果

本文的实验结果表明，即使只执行少量的扰动，模型对目标脑网络分类的准确性和分类裕度均显著下降，并且愚弄率显著提高，这在一定程度上说明了基于脑功能网络的疾病诊断模型在面临对抗攻击时的脆弱性。对比其他攻击方法，我们提出的 BFGCNattack 方法拥有更强的攻击能力。

图 3 总结了不同的攻击方法与攻击程度对分类裕度的影响。其中，图 3(a)中的 3 种攻击方法允许修改的次数均设置为 $\Delta=10\%$。相比 RNDattack，针对脑网络中度中心性较高的边执行攻击也表现出了较好的攻击能力，这在一定程度上说明了度中心性较高的边对模型的影响较大。图 3(b)中，以十折交叉验证中的一折 fold6 为例，显示了本文提出的 BFGCNattack 方法随着允许修改的次数的提升，攻击能力也不断提升。



注：分类裕度越低对应攻击能力越强，模型鲁棒性越差

图 3 不同的攻击方法与攻击程度对分类裕度的影响

Fig. 3 Effect of different attack methods and attack levels on classification margin

表 1 列出了 3 种攻击方法下，相比干净脑网络模型准确率与愚弄率的结果统计。

表 1 不同的攻击方法下模型的准确率与愚弄率统计

Table 1 Statistical results of accuracy and fooling rate of model with different attack methods

策略	愚弄率	准确率
Clean	0	1
RNDattack-10%	0.285	0.715
DEGattack-10%	0.355	0.645
BFGCNattack-10%(Ours)	0.560	0.440

注：愚弄率越高，准确率越低，对应的攻击能力越强，模型的鲁棒性越差

实验结果再次一致表明了我们提出的 BFGCNattack 方法，仅仅执行了 10% 的扰动，就使得模型的愚弄率提高了 56%。

图 4 给出了十折交叉验证下本文攻击方法 BFGCNattack-10% 与干净脑网络的对比实验结果，进一步充分显示了对比模型在面临我们设计的攻击方法时的脆弱性。另外，在实验中还发现，在对目标节点进行攻击时，其他节点也会受到影响而造成整体分类准确率的下降。

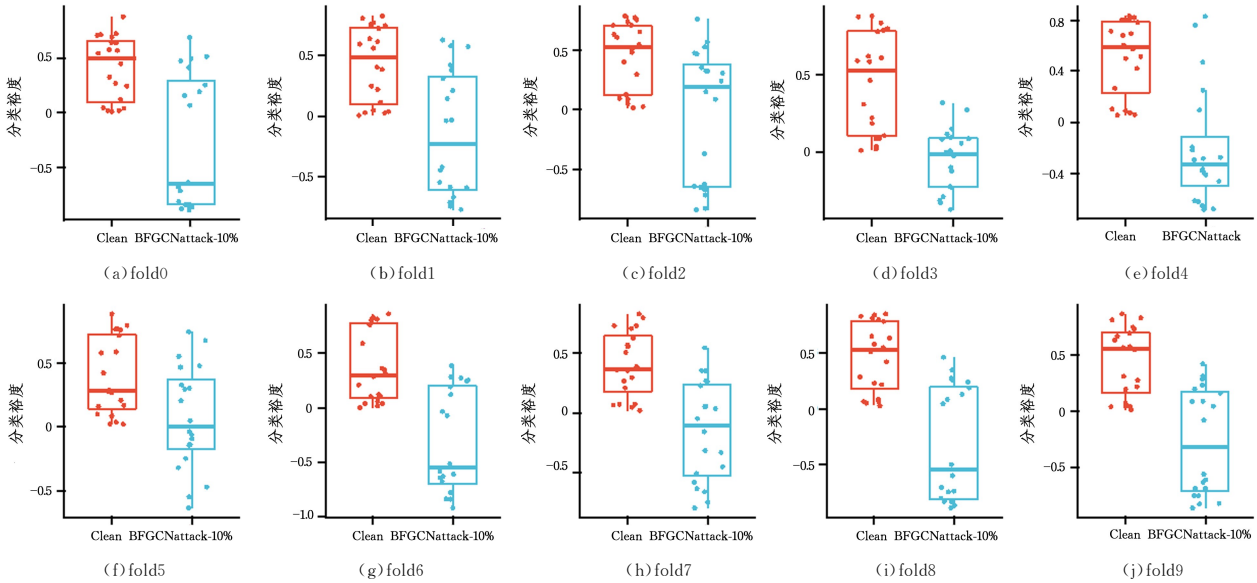


图 4 十折交叉验证下 BFGCNattack-10% 与干净脑网络的对比实验

Fig. 4 Comparative experiment between BFGCNattack-10% method and clean brain networks under 10-fold cross validation

结束语 随着基于功能磁共振成像的脑功能网络分析被广泛应用于各类脑疾病的计算机辅助诊断任务中，模型的可靠性、安全性、可信任性都需要我们时刻关注。本文提出了一种面向脑疾病诊断的图卷积网络对抗攻击方法，探索评估了智能诊断模型在面临对抗性扰动时的鲁棒性。模型准确率、分类裕度的降低与愚弄率的提高均说明了基于脑功能网络的模型在面对对抗性扰动时的脆弱性。

在未来的工作中，我们将进一步完善优化对抗攻击方法的设计，如考虑添加真实场景噪声，也将评估更多的医学诊断模型的鲁棒性。另外，增强模型鲁棒性也是未来研究工作的一个重点，未来的工作可以在构建模型时考虑进行扩展，加入防御或检测手段，使其能够更健壮地抵御攻击，以助力提升医学诊断模型的安全性和可靠性。

参 考 文 献

[1] FINN E S, ROSENBERG M D. Beyond fingerprinting: Choosing predictive connectomes over reliable connectomes[J]. NeuroImage, 2021, 239: 118254.

[2] SHEN X L, FINN E S, SCHEINOST D, et al. Using connectome-based predictive modeling to predict individual behavior from brain connectivity[J]. Nature Protocols, 2017, 12(3): 506-518.

[3] SONG H, FINN E S, ROSENBERG M D. Neural signatures of attentional engagement during narratives and its consequences

for event memory[C]// Proceedings of the National Academy of Sciences. 2021.

[4] DU Y H,FU Z,CALHOUN V D. Classification and prediction of brain disorders using functional connectivity: promising but challenging[J/OL]. https://www. frontiersin. org/articles/10. 3389/fnins. 2018. 00525/full.

[5] ZHANG D Q,HUANG J S,JIE B,et al. Ordinal pattern: A new descriptor for brain connectivity networks[J]. IEEE Transactions on Medical Imaging,2018,37(7):1711-1722.

[6] GAN J Z,PENG Z W,ZHU X F,et al. Brain functional connectivity analysis based on multi-graph fusion[J/OL]. https://www. sciencedirect. com/science/article/abs/pii/S1361841521001031.

[7] BENKARIM O,PAQUOLA C,PARK B,et al. The cost of untracked diversity in brain-imaging prediction[J/OL]. https://www. biorxiv. org/content/10. 1101/2021. 06. 16. 448764v1.

[8] SZEGEDY C,ZAREMBA W,SUTSKEVER I,et al. Intriguing properties of neural networks[J]. arXiv:1312. 6199,2013.

[9] DEMONTIS A,MELIS M,BIGGIO B,et al. Yes,machine learning can be more secure! a case study on android malware detection[J]. IEEE Transactions on Dependable and Secure Computing,2017,16(4):711-724.

[10] GOODFELLOW I J,SHLENS J,SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv:1412. 6572,2014.

[11] FINLAYSON S G,BOWERS J D,ITO J,et al. Adversarial attacks on medical machine learning [J]. Science, 2019, 363(6433):1287-1289.

[12] DI MARTINO A,YAN C G,LI Q,et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism[J]. Molecular psychiatry, 2014,19(6):659-667.

[13] SUN L C,DOU Y T,YANG C,et al. Adversarial attack and defense on graph data: A survey[J]. arXiv:1812. 10528,2018.

[14] XU K D,CHEN H G,LIU S J,et al. Topology attack and defense for graph neural networks: An optimization perspective [J]. arXiv:1906. 04214,2019.

[15] DOU Y T,MA G X,YU P S,et al. Robust spammer detection by nash reinforcement learning[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020:924-933.

[16] WANG X Y,CHENG M H,EATON J,et al. Attack graph convolutional networks by adding fake nodes[J]. arXiv: 1810. 10751,2018.

[17] ZÜGNER D,AKBARNEJAD A,GÜNNEMANN S. Adversarial attacks on neural networks for graph data[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018:2847-2856.

[18] CHEN J Y,CHEN L H,CHEN Y X,et al. GA-based Q-attack on community detection[J]. IEEE Transactions on Computational Social Systems,2019,6(3):491-503.

[19] BOJCHEVSKI A,GÜNNEMANN S. Adversarial attacks on node embeddings via graph poisoning[C]// International Conference on Machine Learning. PMLR,2019:695-704.

[20] CHEN Y Z,NADJI Y,KOUNTOURAS A,et al. Practical attacks against graph-based clustering[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017:1125-1142.

[21] FENG F L,HE X N,TANG J,et al. Graph adversarial training: Dynamically regularizing based on graph structure[J]. IEEE Transactions on Knowledge and Data Engineering,2019,33(6): 2493-2504.

[22] BETZEL R F,BASSETT D S. Multi-scale brain networks[J]. Neuroimage,2017,160:73-83.

[23] GREICIUS M D,KRASNOW B,REISS A L,et al. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis[C]// Proceedings of the National Academy of Sciences. 2003:253-258.

[24] KTENA S I,PARISOT S,FERRANTE E,et al. Metric learning with spectral graph convolutions on brain connectivity networks [J]. NeuroImage,2018,169:431-442.

[25] PARISOT S,KTENA S I,FERRANTE E,et al. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease[J]. Medical Image Analysis,2018,48:117-130.

[26] DEFFERRARD M,BRESSON X,VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[J]. Advances in Neural Information Processing Systems,2016,29:3844-3852.

[27] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv:1609. 02907,2016.

WANG Xiao-ming, born in 1999, master, is a member of China Association of Artificial Intelligence. His main research interests include brain functional connectivity networks analysis and machine learning.

ZHANG Dao-qiang, born in 1978, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine learning, pattern recognition, data mining and medical image analysis.

(责任编辑:喻藜)