基于纹理特征约束的神经网络模型鲁棒性提升方法

杨中国[1,2] 张 镌[1], 王丽君[1,2]

[1]北方工业大学信息学院 北京 中国 100144
[2]大规模流数据集成与分析北京市重点实验室 北京 中国 100144

摘要 深度学习模型可以从原始数据中自动学习到数据的纹理特征和形态特征, 使得其在安全验证、识别分类、语音人脸识别等不同领域取得远远超过人工特征方法的性能。虽然深度学习在图像分类和目标检测等方向上取得了较好成效, 但是通过在输入上添加难以察觉的微小扰动形成的对抗样本导致深度学习模型在实际使用中存在巨大的风险。因此, 提高单个模型的鲁棒性是重要的研究方向。前人在时序数据分类模型的鲁棒性研究中, 对抗样本的解释性研究较为欠缺。目前较为常见的防御对抗样本的方法是对抗训练, 但是对抗训练有着非常高的训练代价。本文以时序数据分类模型为研究对象, 定义了时序数据的纹理特征和形态特征, 并基于理论证明和可视化特征层方式, 说明了纹理特征是被攻击的关键因素。同时, 提出了一种基于特征约束的模型鲁棒性提升方法。该方法结合多任务学习, 通过在误差函数中增加特征的平滑约束项, 引导模型在分类的同时尽可能学习到原始数据的形态特征。在保证分类精度的同时, 降低对抗样本存在的空间, 从而训练出更加鲁棒的模型。算法在经典分类模型和多个时序数据集进行了大量的实验, 实验结果表明了本文方法的有效性, 在多种对抗攻击下, 能较好的提高单个模型的鲁棒性。

关键词 时序数据分类; 对抗样本; 纹理特征; 鲁棒性
中图法分类号 TP183; TP391.4 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2023.01.03

Robust Enhancement Method of Depth Model Based on Texture Feature Constraint

YANG Zhongguo[1,2], ZHANG Juan[1], WANG Lijun[1,2]

[1] School of Information Science and Technology, North China University of Technology, Beijing 100144, China
[2] Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, Beijing 100144, China

Abstract Deep learning model can automatically learn the texture and morphological features from original data, which makes it achieve far better performance than the manual features based method in many fields such as security verification, recognition and classification, voice and face recognition, etc. Although deep learning has achieved good performance in image classification and object detection, the existence of adversarial examples formed by adding imperceptibly small perturbations to the input leads to huge risks in the practical use of deep learning models. Among them, the improvement of the robustness of a single model is an important research field. In the previous research on the robustness of time-series data classification model, the explanatory research against samples is relatively lacking. At present, the most common method to defend against adversarial samples is adversarial training, but adversarial training has a very high training cost. Take the time-series data classification model for example, we define the texture features and morphological features of the time series data. Additionally, based on the theoretical proof and feature visualize method, we explain that the texture feature is the key factor to be attacked. At the same time, a method for improving model robustness based on feature constraints is proposed. This method combines multi-task learning to guide the model to learn the morphological features of the original data as much as possible. While ensuring the classification accuracy, the space of adversarial samples is reduced, so as to train a more robust model. A large number of experiments on classical classification models and multiple time-series datasets were conducted, and the experimental results show the effectiveness of the method. Moreover, it can better improve the robustness of a single model under a variety of adversarial attacks.

Key words time-series classification; adversarial attack; textural features; robustness

通讯作者: 杨中国, 博士, 助理研究员, Email: yangzhongguo@ncut.edu.cn。
本课题得到“融合业务过程和物联大数据的服务抽象与编程机制研究”国家自然科学基金委重点国际(地区)合作研究项目(No. 62061136006)和北京市自然科学基金项目(No. 4202021)资助。
收稿日期: 2021-09-24; 修改日期: 2022-02-24; 定稿日期: 2022-11-03

1 引言

随着智能传感器的出现以及大规模数据收集和存储技术的进步, 可以从各种传感器收集时间序列数据进行分析学习, 进而得到有用的模式。这使得应用系统更加智能化[1]。其中, 时序数据的分类是一个研究热点[2-3]。

典型的时间序列分类应用包括, 基于心电图数据检测有认知缺陷的患者[4]、基于音频对单词进行分类、基于运动轨迹数据识别手势动作[5]、基于用电时序数据判定是否有窃电行为[6]等, 此外还有基于制造厂的传感器数据, 对设备进行实时分析, 并自动预警, 以避免出现设备的重大问题[7]。

对抗攻击是通过一定手段将输入数据施加微小的变化, 使得已经训练好的模型做出错误的决策。对抗样本使得深度模型会在看似正确的输入数据上做出错误的预测, 从而产生有目标的攻击行为。在神经网络应用中, 是一个严重的安全问题, 目前广泛存在于视觉任务中。在输入图像上添加轻微的扰动或精心设计的噪声可能会误导图像分类算法, 使其做出置信度很高但完全错误的判断[8]。

虽然在多个领域的分类任务中, 深度模型取得了较好的效果, 时间序列分类问题也大多使用深度模型来解决[3,9-10], 但对抗样本的漏洞使得安全性要求高的领域面临极大的风险。Papernot 等人[11]最近的研究表明, 针对特定计算机视觉分类器的对抗性攻击可以很容易地转移到其他类似的分类器中, 使得这个问题的求解更加复杂。考虑到物联网数据的广泛使用, 时序数据分类模型的对抗防御也成为了研究重点[12]。

对抗样本的生成方法大多针对深度神经网络模型(Deep Neural Networks, DNNs)的梯度信息, 使得 DNNs 容易受到攻击[13−15]。Carlini 和 Wagner[16]展示了如何攻击文本到语音分类器。此外, 它们还提供各种音频片段, 将这些片段输入到文本到语音分类器 DeepSpeech 中, 均无法正确检测语音。在使用时间序列分类算法的医疗设备中, 可能还存在其他安全问题, 在这些设备中, 它可能会被欺骗, 导致误诊, 从而影响对患者疾病的诊断。用于检测和监测地震活动的时间序列分类算法可能被操纵, 导致社会性恐惧。使用时间序列数据对佩戴者的活动进行分类的可穿戴设备可能会被欺骗, 以至让用户相信他们正在做其他动作。

针对时序分类模型, 时序数据的对抗样本研究工作主要分为两类。从攻击角度, 包括对抗样本构造、迁移性; 从防御角度, 包括如何提高模型的鲁棒性。但是, 针对时序数据分类模型的对抗样本解释性和原理研究较少。

本文从特征提取的角度解释对抗样本存在的合理性和危害性, 进一步提出了基于特征约束的训练方法, 用于提高模型的鲁棒性。本文贡献可以概括为以下三点:

1) 从特征变化角度, 给出了纹理特征放大噪声, 引起了巨大偏差的理论说明。

2) 验证了特征约束的作用, 通过设置约束误差, 大大降低了纹理特征的比例, 提高了单个模型的鲁棒性。

3) 本文方法可以作为其他方法的叠加使用, 只需要修改误差函数的构成。

本文结构如下: 第二部分回顾了针对时序数据分类模型的对抗攻击防御的相关工作。第三部分介绍了几种时间序列分类模型的背景知识、对抗攻击方法。从理论上介绍了时序分类模型的对抗样本存在的原因, 并提出基于特征约束方法的防御方法。第四部分给出了两种攻击方法下的多个时序分类数据集的分类准确率结果, 并分析了提出的方法的有效性。第五部分对全文进行了总结, 并展望了今后的工作。

2 相关工作

本文针对时序分类任务的对抗样本可解释性和防御算法两方面进行回顾总结, 分析相关问题。深度学习的对抗样本一直是一个研究热点, 段广晗等人[17]总结了几种对抗样本的存在性解释, 包括盲区假说、线性假说、边界倾斜假说、决策面假说、流形假说。该文献指出当前的各种假说对于对抗样本的生成机理缺乏共识, 由于深度学习模型的不可解释性以及数据流形几何结构的高度复杂性, 不同假说对于对抗样本的生成机理研究具有不同的侧重点, 缺乏数理完备的统一理论解释。

其中, 我们从数据特征学习的角度进行了相关研究总结, 关注到不同的研究者从模型学习的角度指出深度模型会偏向学习到数据中的高频特征, 或者说是纹理特征。

Galloway 等人[18]指出批归一化 BN(Batch Normalization, BN)层是造成模型鲁棒性差的一个原因。在论文中, 作者通过增加 BN 层和去除 BN 层的实验, 给出了多个模型的分类准确率和对抗准确率, 验证了文章的结论。但是, 文章中没有分析数据特征的变化规律。

Wang 等人[19]从数据的角度入手, 探讨了深度模型的泛化能力。当训练一个模型的时候, 没有指明模型应该学习数据想表达的内容还是这些高频信号, 模型就会无差别地学习数据本身的信号和这些高频信号。该文作者在训练好的模型中, 调整其权重, 使卷积核变得更加平滑; 直接在训练好的卷积核上将高频信息过滤掉; 在训练卷积神经网络的过程中增加正则化, 使得相邻位置的权重更加接近, 从而使得模型在对抗鲁棒性有一定程度的提高。他们指出高频信息是对抗攻击成功的一部分原因, 并不是全部的因素。

研究者[20-21]尝试让卷积神经网络(Convolutional Neural Networks, CNN)学到形状, 而不仅是纹理。通过风格迁移, 作者创造了新的图像, 在猫的形状上披上大象的纹理。对于机器学习模型而言, 该图像被认为是大象, 但是对于人而言还是猫。该文献说明, 现有的机器学习模型会大量学习到图像的纹理特征, 并作为分类的重要依据。同时该文作者也通过生成不同纹理的数据集来增强训练集的方式, 来扭转模型对纹理的偏好, 实现更好的形状学习。

针对时序数据集, Fawaz 等人[22]给出了时序数据对抗攻击的形式化定义, 验证了潜移学习方式下也可以产生对抗样本, 并在多个时序数据集上给出了不同攻击方法的分类准确率。

Sarkar 等人[23]尝试从数据增强的角度来提高分类模型的鲁棒性, 他们提出了基于局部梯度的方法来增强原始数据。

Wang 等人[24] 针对时序分类任务, 提出时间序列的子序列是具有辨别力的组成部分, 但是他们与原始时间序列却不完全相似。这就使得很难解释对分类起作用特征的工作机理, 即原始序列中到底什么样的特征能完成分类任务。作者在文中采用一些约束方法让模型学习到可解释性的独特子序列。

Yang 等人[25]通过在模型最后的特征层中增加识别网络来辨别学习对抗样本, 从而完成防御的方法。该方法适用于白盒模型, 对黑盒攻击和迁移攻击的效果有限。之后, 他们[26]又提出使用编码的方式对时序数据进行预处理, 一定程度上提高了模型的鲁棒性。该方法中需要构造新的训练数据, 成本高且在某些场景下不能实现。

Jiang 等人[27]提出一种基于时序数据表示方式, 并基于该表示来识别对抗样本和普通样本, 该方法没有直接提高模型的鲁棒性, 且计算复杂度较高。

3 纹理特征与对抗样本

3.1 基于深度学习时序数据分类模型

时序数据分类任务中大量使用了深度学习方法[4,6,28], 并且取得了很好的性能, 包括健康诊断、恶意行为识别、窃电行为检测等。通常, 这些模型是通过在输入时间序列上滑动一维滤波器(1-CNN), 使得网络能够学习对分类有用且有区别的、非线性的特征。例如: 它[29]包含 3 个卷积层, 卷积核大小分别为 8、5 和 3, 并分别发射 128、256 和 128 个滤波器。每个卷积层后面是一个批量规范化层[30], 该层应用了 ReLU 激活层。在最后一个 ReLU 激活层之后采用全局平均池化层, 最后应用 softmax 确定分类概率向量, 典型的分类网络结构见图 1。

3.2 对抗样本及鲁棒性

论文[31]介绍了时序分类模型及其对抗样本的形式化描述。时序数据的一个样本可以表示为 $X=[x_1,x_2,...,x_T]$, 其中 T 是样本中采样点的个数。一个训练好的模型表示为 $f(\cdot)\in F:R\wedge T\to Y$, 它把输入数据从 T 维实数空间转化为离散的类别空间中, 并表示为 Y。令对抗样本为 X', 需要满足 $\|X-X'\|_p<\varepsilon$, 且 $Y\neq Y'$, 其中 ε 是一个较小的正数, P 可以取 1, 2 或者 ∞。通常 ε 取值为 0.01 或者 0.1, 即

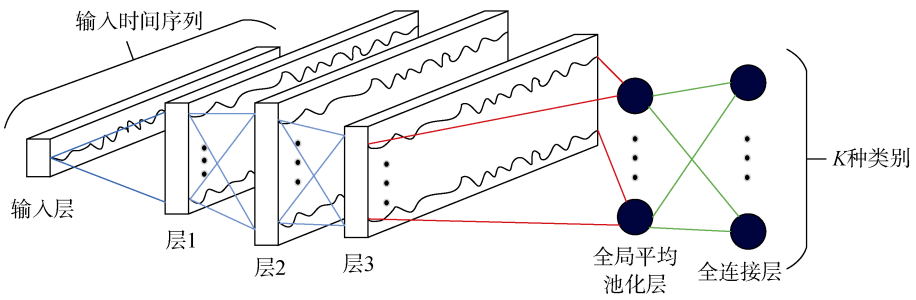

图 1 基于残差网络的时序分类模型示意图[3,31]

Figure 1 Schematic diagram of time series classification model based on residual network[3,31]

对抗样本与原始样本的差异会约束在一个较小的范围, 甚至连肉眼都难以观察到。在训练很好的模型上, 原始样本 X 和对抗样本 X' 的识别结果却不一样, 示意图见图 2。

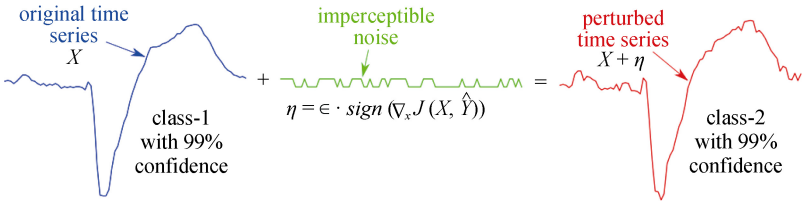

图 2 基于快速符号梯度法的对抗样本[22]

Figure 2 Adversarial samples based on fast sign gradient method [22]

当样本中增加了一定噪声, 会导致分类正确率急速下降。在白盒攻击下, 这些噪声由梯度信息计算得到。在鲁棒模型的定义下, 鲁棒模型的决策并不会因为对抗噪声的加入而改变, 因此鲁棒模型在对抗样本上依然会保留可观的准确率。或者说, 对抗噪声的强度需要达到一定程度, 才可以极大的改变模型的决策。

3.3 常见的三种基于梯度的攻击方法

快速梯度符号攻击(Fast Gradient Sign Method, FGSM)[32], 是一种通过反向传播, 达到将损失函数反传到输入层, 并且通过梯度方向更新原样本, 完成生成对抗样本、欺骗分类器的目的。从攻击可以应用于任何可以计算梯度的深度学习模型。具体攻击公式如下:

$$x^{adv} = x^{clean} + \varepsilon sign(\nabla_x L(x^{clean}, y)) \tag{1}$$

其中, $\nabla_x L$ 是反传回输入层的梯度, $sign(\bullet)$是符号函数, ε一般用来调整 FGSM 攻击的扰动程度, x^{clean}是原样本, y 是真实类别标签。FGSM 算法较为简单, 即针对一个样本仅加一次梯度, 生成的对抗样本大多攻击能力比较一般。

映射梯度下降法(Projected Gradient Descend, PGD)[33], 是 FGSM 算法的改进, 可以理解为 FGSM 的迭代。当模型复杂时, 极小范围的变化也能攻击, 同时为了保证 扰动满足 L_2 或者 L_∞ 的限制, 在迭代的过程中引入了映射的过程。具体公式如下:

$$x_0^{adv} = x^{clean} \tag{2}$$

$$x_{t+1}^{adv} = proj_x(x_t^{clean} + \varepsilon sign(\nabla_x L(x^{clean}, y)) \tag{3}$$

与 FGSM 相比, PGD 算法攻击强度更强, 但是生成对抗样本的耗时也是 FGSM 的数倍, 通常 PGD 迭代次数越多, 生成的对抗样本攻击能力越强。

动量迭代梯度标志攻击(Momentum Iterative Fast Gradient Sign Method, MI-FGSM)是由 Dong 提出的一种动量方法, 通过在迭代过程中沿损失函数的梯度方向累积速度矢量来加速梯度下降算法的技术。比起 PGD 方法能够更好的缓解过拟合现象, 提升算法黑盒迁移效果[34]。其中梯度方法的累积公式如下:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x L(x_t^{adv}, y)}{\left|\nabla_x L(x_t^{adv}, y)\right|_1} \tag{4}$$

$$x_{t+1}^{adv} = proj_x(x_t^{adv} + \varepsilon sign(g_{t+1})) \tag{5}$$

3.4 纹理特征定义与对抗样本存在性

经过前人研究, 可知时序数据分类模型中, 会有大量的纹理特征和全局特征存在。如何精确的刻画和辨别它们也是研究热点, 本节形式化地定义了时序数据的纹理特征, 并给出辨别方法。

简要分析时序分类模型的特征计算过程, 如下图 3 所示, 展示了一维卷积的计算过程, 也是时序数据特征提取的过程。

图 3 中蓝色表示的特征中每个点是由上一层特征的一个滑动窗口的数据进行卷积操作得到, 其计算公式可以表示为公式 6。

$$W^T X + b = Y \tag{6}$$

在时序数据集 Coffee[7]的分类模型中展示了几种典型的特征, 具体表现为部分 W 会提取出全局特征, 见图 4 左部分, 有部分的 W 会提取出很细微的纹理, 见图 4 右部分。

当模型学习到的特征提取器, 即 W 不同时, 提取的特征会有很大的不同, 大致可以分为形态特征和纹理特征。纹理特征可以定义为 $f\_texture$, 它需要满足公式 7 和公式 8。

$$W^T X_i + b = h_i, i \in [1, N] \tag{7}$$

$$s.t. \|w\|_p < \theta \tag{8}$$

纹理特征的每个点相差范围很小, 相对集中, 统计方差较小。而对于形态特征, 它的分布范围和原

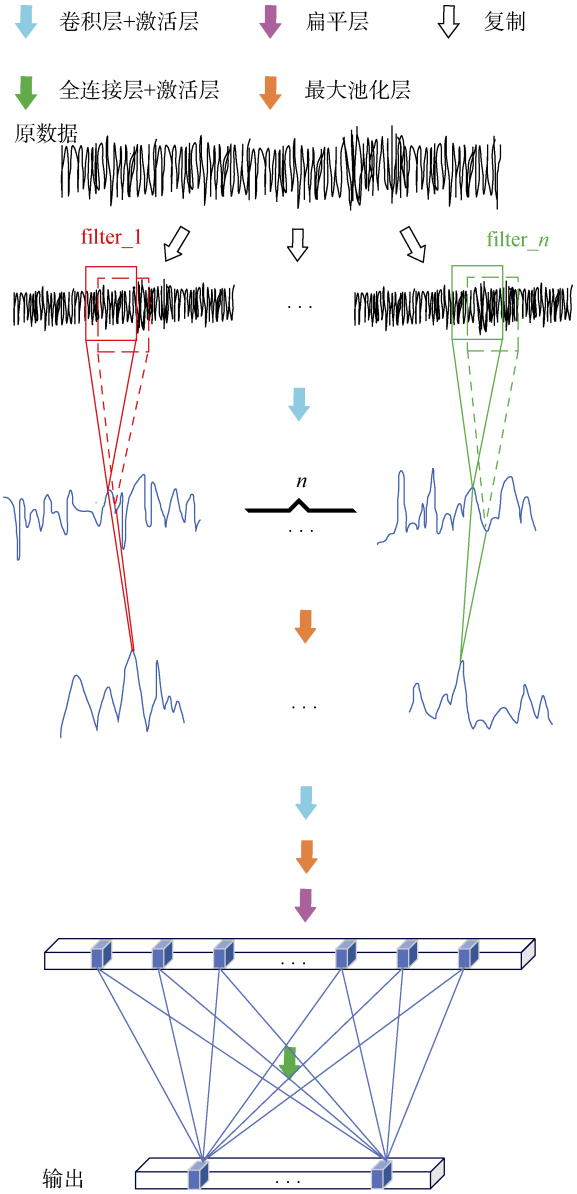

图 3 一维卷积的计算过程

Figure 3 Calculation process of one-dimensional convolution

始数据较为相似, 体现了原始数据形态的变化趋势。

纹理特征的提取在经过 BN 层后会放大误差, 是造成对抗样本存在的一个原因。首先, BN 层通过 $y_i = \gamma \cdot \frac{h_i - \mu}{\sigma} + \beta$ 计算, 把所有提取的特征都归一到同一尺度下。其中 μ 是特征的均值, σ 是特征的标准差, γ 和 β 是训练得到的参数。

Johan[35]发现通过简单的从批归一化的算法中删除参数 γ 和 β, 再用训练网络研究影响, 对最终的准确度没有太大的影响。本文通过统计归一化层的参数也发现 BN 层的 λ 数值相似, 实验是在 VGGnet 的基础上, 以 Coffee 数据训练的模型精度达到 92% 的模型进行可视化。 图 5、图 6 统计并展示了 BN 层的参数。

所以, 特征是否缩放主要取决于特征的方差, 而纹理特征和形态特征的方差相差巨大。具体实验结果统计如图 6 所示:

特征尺度放大、缩小的系数主要由 σ/γ 决定。纹理特征会通过 BN 层的缩放, 极大的将细小的误差放大。在 *eps* 不等的 FGSM 攻击下, 鲁棒性最好的模型进行可视化的展示, 并选取其中一个样本举例。经过卷积之后的特征层主要呈现两种趋势, 把剧烈波动的特征称为高频特征, 而平缓的特征称为低频特征。纹理特征放大了误差, 而形态特征压制了误差。

证明: 纹理型特征在深度神经网络中会被逐步放大该证明说明在一次卷积->批归一化的运算中, 纹理型特征会由于数据本身的分布导致被放大。

(1) 计算过程展示

$$W^T X_i + b = h_i, i \in [1, N] \tag{9}$$

$$X = (x_{i-L/2}, x_{i-L/2+1}, ..., x_i, ..., x_{i+L/2})^T \tag{10}$$

X 的每个元素是上一层特征数据的一段, 该段

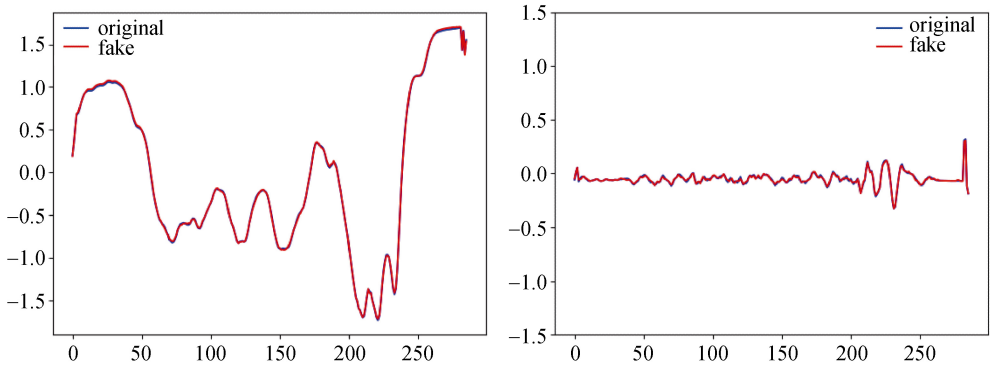

图 4 时序分类模型中的形态特征与纹理特征, 以 Coffee 数据集为例说明

Figure 4 The morphological features and texture features in the temporal classification model are illustrated by taking the coffee data set as an example

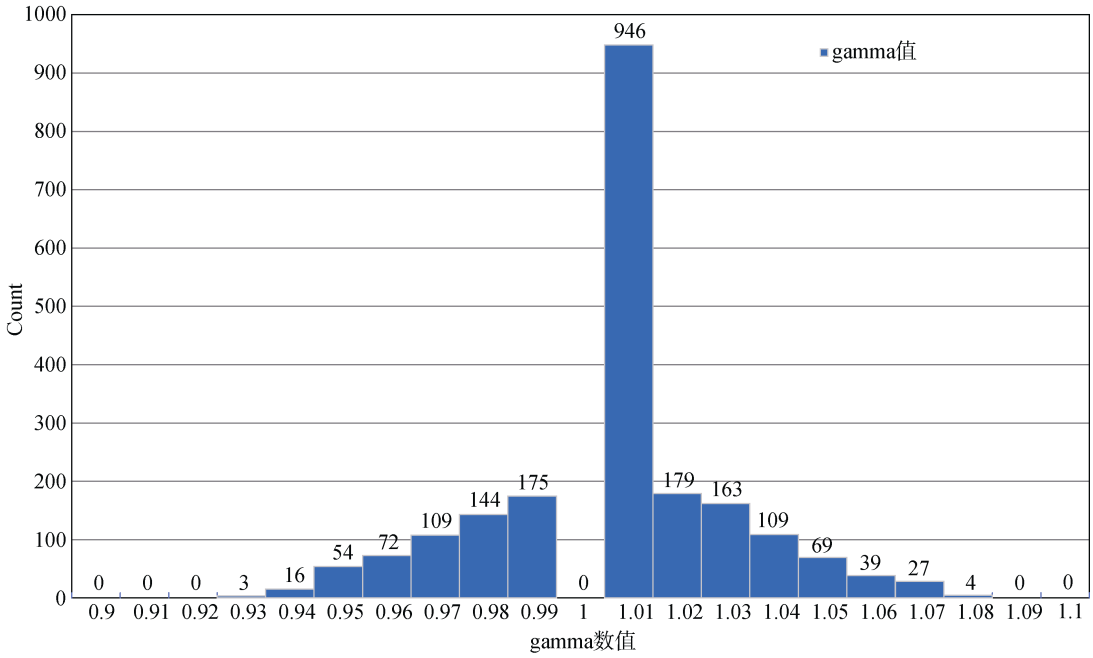

图 5　时序分类模型中 BN 层的γ, 以 Coffee 数据集为例说明

Figure 5　The γ of the BN layer in the temporal classification model are illustrated by taking the coffee data set as an example

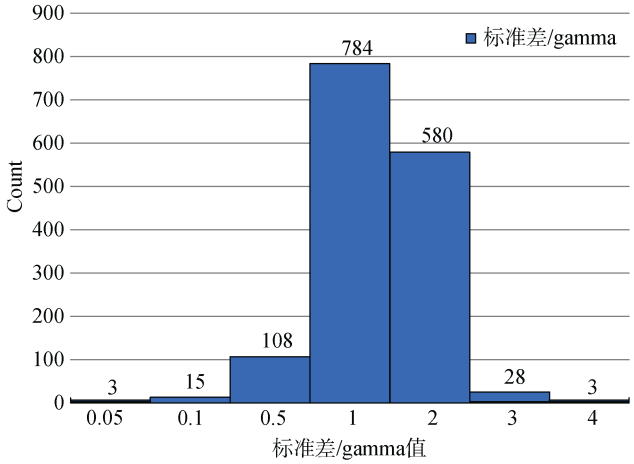

图 6　时序分类模型中 BN 层的σ/γ, 以 Coffee 数据集为例说明

Figure 6　The σ/γ of the BN layer in the temporal classification model are illustrated by taking the coffee data set as an example

数据以 x_i 为中心的列量, 其长度为 L, L 是卷积核 W 的长度, N 为输入特征的长度。

$$y_i = \gamma \cdot \frac{h_i - \mu}{\sigma} + \beta \tag{11}$$

其中 H 是经过卷积后得到的特征向量, σ 是 H 的标准差。

(2) 标准差 σ 的取值范围, 不妨令偏置项 b 为零。

$$\|H\| \leqslant \|W\| \|X\| \leqslant \theta \|X\| \tag{12}$$

H 是卷积后的特征, 基于 H 统计出均值和标准差。当 W 是纹理型特征, 它的范数会小于给定的正数 θ, 使得 $\|H\|$ 的取值范围变得很小, 从而导致标准差 σ 很小。

(3) y_i 取值范围。

γ 的取值是训练得到的, 经过统计得知, $\gamma \notin (0,2)$, γ 的取值与 $\|W\|$ 没有关系, 且相差不大。

所以, y_i 的取值与 W 的范数有直接关系。当 $\|W\| \leqslant \theta$, 导致很小 σ, 从而使得 y_i 被放大很多倍。

(4)噪声会随 y_i 的变化而变化

$$W^T(X_i + \Delta x) + b = h_i, i \in [1, N] \tag{13}$$

$$y_i = \gamma \cdot \frac{h_i - \mu}{\sigma} + \beta \tag{14}$$

当噪声 Δx 加入到特征 X_i 中,

$$\|H\| \leqslant \|W\| \|X + \Delta x\| \leqslant \theta(\|X\| + \|\Delta x\|) \tag{15}$$

$$y_i^{'} = \gamma \cdot \frac{W^T(X_i + \Delta x) - \mu}{\sigma} + \beta \tag{16}$$

$$= \gamma \cdot \frac{W^T X_i - \mu}{\sigma} + \gamma \cdot \frac{W^T \Delta x - \mu}{\sigma} + \beta$$

$$= y_i + \gamma \cdot \frac{W^T \Delta x - \mu}{\sigma}$$

$$= y_i + \gamma \cdot \frac{W^T \Delta x - \mu}{\sigma}$$

后面这一项中的 $\gamma \cdot \frac{W^T \Delta x - \mu}{\sigma}$ 表明噪声在线性变换后, 会经过 σ 的放缩而产生巨大的偏差。

上述公式证明了噪音在 BN 层会发生变化, 是高频特征放大误差, 低频特征抑制误差的最主要原因。波动大的高频特征会被放大, 产生的误差也会被放大。低频特征本身范围大, 在 BN 层后, 会被缩小尺度, 同时误差也被缩小。实验以很小的随机噪声[0.001,– 0.001,0.001,0.001,0.001, – 0.001,⋯,0.001]输入到网络中, 观察数据的变化, 辅助说明数学公式。图 7 中展现了微小噪声逐层经过 BN 之后, 微小的误差经过多层放大、收缩, 变成了十分明显的误差。在最后一层之前, 前层的每一个特征都用最大池化方法变成了一个点, 把多个特征的点画在一起作对比。在不同攻击率下采用 FGSM 方法, 具体方法细节见下节。

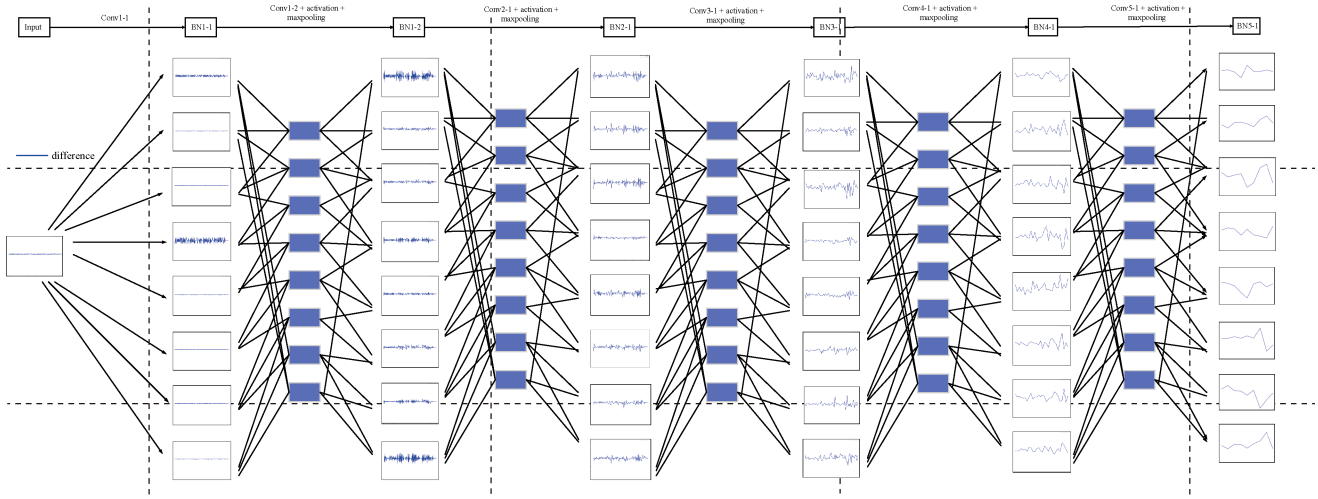

图 7　不同种类的特征在 BN 层作用下的误差变化

Figure 7　Error variation of different types of features under the action of BN layer

图 8 展示了对抗样本在全局平均池化层(GAP, Global average pooling)的表现, 其中, 大部分对抗样本特征与原样本的特征都发生了一定的偏移, 最为明显的是方框里的特征, 发生的偏移量最大。

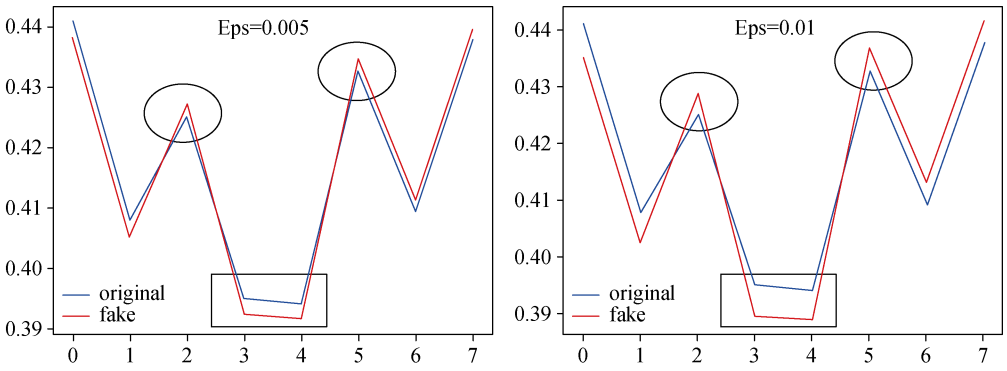

图 8　GAP 层特征在对抗样本下的表现

Figure 8　Features of GAP layer under adversarial samples

4　基于特征约束的对抗防御方法

为了提高模型的鲁棒性, 本文尝试约束模型学习到的特征类型。采用的方法如图 9 所示, 在模型的误差中, 加入了特征正则化约束误差。该误差项的目的就是控制特征层中学习到的特征更趋近于形态。关于纹理特征和形态特征, 肉眼很容易观察到, 但是在模型中需要量化的方法来区别。朴素的想法就是让纹理型的特征, 拥有较大的误差值, 而形态型特征拥有较小的误差值。通过在误差函数中引入这个形态相关的误差值, 从而在误差减少的过程中, 约束纹理型特征的比例。再结合训练过程中的多任务训练框架, 自动调整特征约束的权重, 让模型既

能正确分类, 又能保留较多的形态特征。最后, 通过大量对抗攻击实验的方法验证本文方法的有效性, 以及在多个数据集上测试本文方法是否能有效的改善模型的鲁棒性。也就是采用同样的训练环境, 增加和不增加特征约束误差两种情况下得到模型。再利用同样的攻击方法, 设置相同参数的情况下, 来验证模型的分类正确率。由于模型训练具有一定的随机性, 本文实验均是在多次训练基础上, 采用统计方法得到的精度。

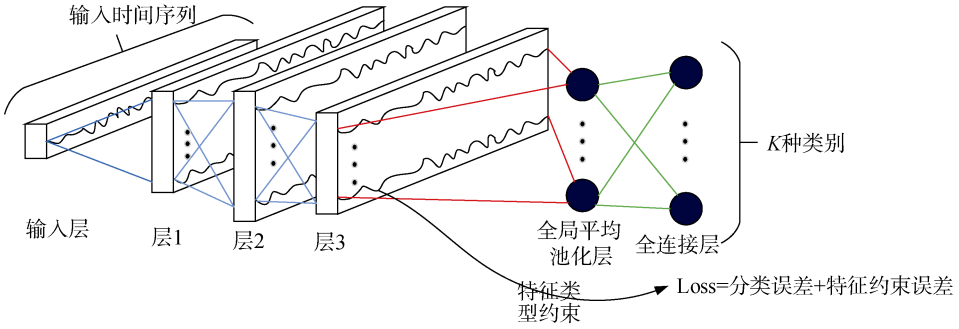

图 9　基于特征层正则化的对抗防御算法流程图

Figure 9　Flow chart of countermeasure defense algorithm based on feature layer regularization

4.1　特征约束方法

4.1.1　特征差异定义

为了验证不同类型的特征在量化指标上的差异。将训练好的分类模型中的任意一层特征可视化出来, 并计算了原始样本与对抗样本在同一位置上的特征差异。这里采用了两个序列差异的二范数, 计算方式如下。

$$\| f - f' \|^2 = \sum_{i=1}^{K} (f_i^2 - f_i'^2) \tag{17}$$

上述式子计算的值越大, 表明原始样本和对抗样本在该处提取的特征差异越大。也表明, 该处的特征发生了更大的偏差。

在 Coffee 数据集上得到的结果如图 10 所示。其中蓝色的是原始样本, 红色的是基于 FGSM 攻击方法得到的对抗样本, 它们的差值的二范数标于图中。

图 10 中的第一排的两个特征明显倾向于形态特征, 而第二排的两个特征更倾向于纹理特征。而第二排的两个特征的二范数也明显大于第一排的两个特征。实验表明, 局部波动剧烈的纹理特征下, 其偏差越大; 在平缓变化, 但整体取值范围更大的形态特征下, 其偏差越小。这与纹理特征更容易放大对抗样本中的误差是吻合的, 也验证了本文对时序分类模型中的误差积累的说明。

4.1.2　特征约束方法

上述实验用第二范数通过差值定义了高低频特征。此节通过加入平滑层, 用来约束最后一层的特征层达到约束纹理特征的效果。平滑层采用一维卷积核, 利用经验参数进行实验尝试, 将平滑因子初步设置为[0.1201,0.2339,0.2921,0.2339,0.1201]。平滑层不参与后续训练, 仅将第二范数设置为另一任务 loss。Loss 越趋近 0, 高频越趋近低频。具体流程如下图 11 所示。

4.2　基于特征约束的模型鲁棒性提升

4.2.1　多任务训练

多任务训练是把多个相关的任务放在一起学习的一种机器学习方法。主任务利用相关任务的训练信号所拥有的领域相关信息, 作为主导偏差来提升主任务泛化效果。多任务训练涉及多个相关的任务同时并行训练, 梯度同时反向传播, 多个任务通过底层的共享表示来互相帮助学习, 提升泛化效果。

鲁棒性和模型的准确率是天然相关的任务, 它们都是利用神经网络自动提取特征的能力完成分类, 有共同的特征提取基础和鲁棒性要求。要求提取的特征具有较好的抗干扰能力, 同时分类准确性需要提取的特征有足够的分辨能力。本文借助多任务训练的方式, 在误差反向传播中, 学习到能分辨多个类别样本的特征, 同时确保这些特征有一定的抗干扰能力。由此可以通过协同训练的方式提高模型的鲁棒性。

本文考虑基于多任务训练思想, 通过特征约束, 降低易受到干扰的特征比例, 甚至降为 0。特征的约束可以增加到任意的特征抽取层中, 此实验是在特征抽取的高层特征输出上进行的, 即在最后一层激活层的特征层进行重组、平滑计算, 产生平滑后的特征层。对平滑后的特征与平滑前的特征的差值进行定义, 作为误差加入到原来的误差函数中, 进行协同训练。这里使用的是第二范数作为两个特征层曲线的差值定义, 添加为新的误差项进行模型训练。

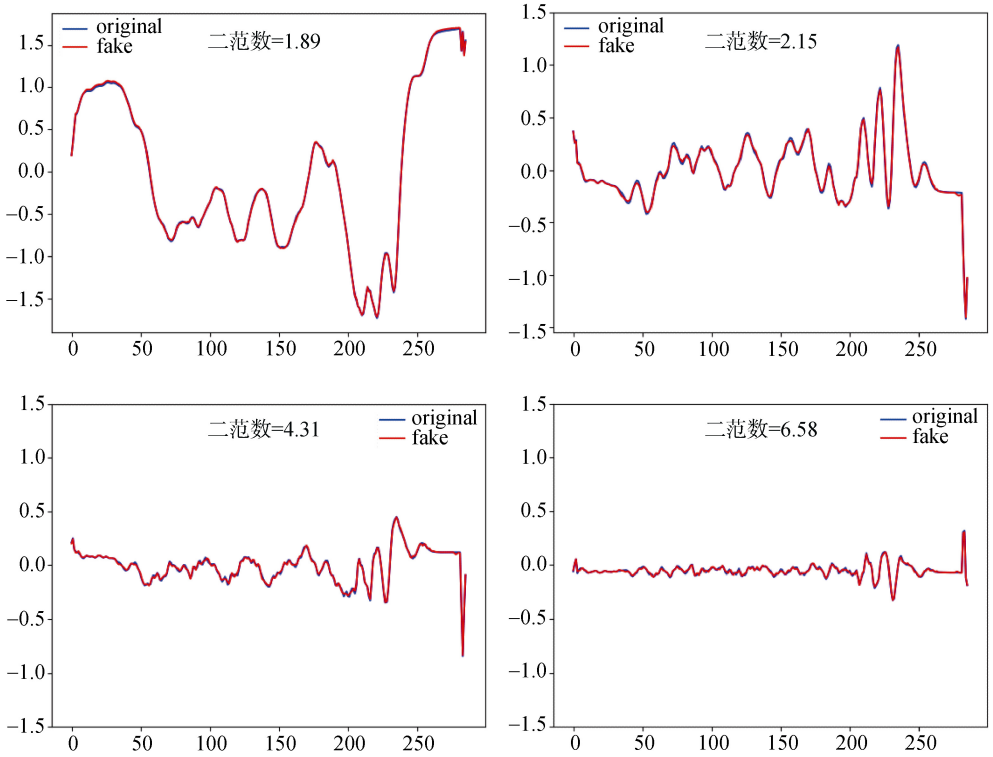

图 10　Coffee 数据集的特征及对应的第二范数值

Figure 10　Characteristics of coffee dataset and corresponding second norm values

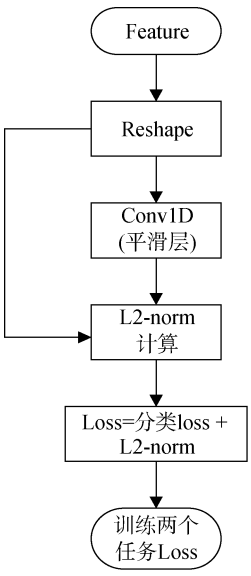

图 11　特征约束的算法流程图

Figure 11　Algorithm flowchart of feature constraints

当最后一层的特征在平滑后与原始特征进行对比, 如果其差值的二范数越小, 则表明该特征越是倾向于形态特征。相关的解释在图 10 中也有说明。

本文实验的网络架构图见表 1。其中借鉴了经典残差网络的跨层连接方式。其中每个 Block 是一个残差块, 里面由 7*7, 5*5 和 3*3 的三个卷积层构成。每个卷积层后(Conv)都接入了批归一化层(Batch Normalization)和激活层(Relu)。在残差块后接入了广泛使用的全局归一化层(Global Average Pooling)和全连接层(Dense Layer)。模型的最后是分类网络的误差函

表 1　网络结构

Table 1　Network structure table

Input		
Conv7*7 Batch Normalization Relu	Conv5*5 Batch Normalization Relu	Conv3*3 Batch Normalization Add Input
	Relu	
Conv7*7 Batch Normalization Relu	Conv5*5 Batch Normalization Relu	Conv3*3 Batch Normalization Add Block1
	Relu	
Conv7*7 Batch Normalization Relu	Conv5*5 Batch Normalization Relu	Conv3*3 Batch Normalization Add Block2
	Relu	
	Global Average Pooling	
	Dense	
	Softmax	

数计算层(Softmax Layer)。该模型的设计考虑的是最常用的模块组合，能代表一大部分基于卷积操作的分类网络。

4.2.2 特征可视化

图 12 将原模型的特征以及优化后模型的特征进行可视化。图左边的源特征层高频特征较为明显的为 4 个，而图右边高频特征减少为 2 个。

通过特征类型约束下的训练，可以有效的降低纹理特征的比例，从而降低对抗样本攻击的成功率。但是，由图 12 也发现，某些特征不能完全变为低频特征的原因在于实验选取数据集 Coffee 分类 *class-0* 和 *class-1* 的差别过小，不能直接区分出来，需要细节对比。本文阐述的设计思路，是从高频特征趋近低频特征入手。高频对应着纹理细节，低频特征对应着整体形态。而差别过小的 Coffee 数据集分类需要依靠这些纹理特征进行分类，达到分类准确，但是这就造成了本文论点的矛盾，即分类准确率和鲁棒性的矛盾。

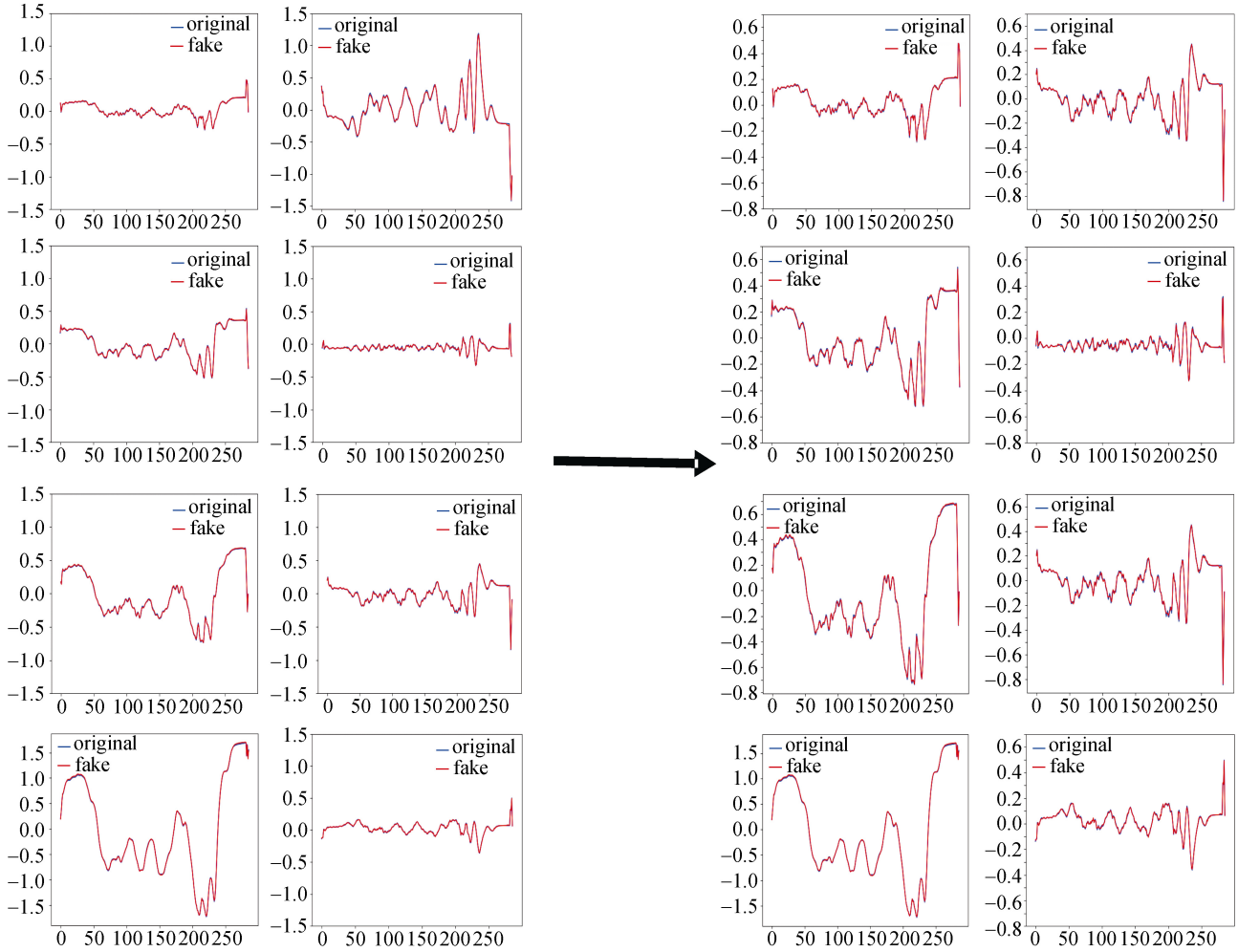

图 12 原模型和优化模型的特征前后对比, (左)原模型特征层特征(右)优化模型特征层特征

Figure 12 Comparison of the features of the original model and the optimized model, (left) features of the feature layer of the original model (right) features of the feature layer of the optimized model

5 实验结果与分析

5.1 ECG200 数据集

选取 UCR 数据集中的另一数据集 ECG200，此数据集和 Coffee 相比，类别间形态差异更为明显，如图 13 所示。

同样采用了三种攻击方法来验证优化模型和原模型的抗干扰能力。FGSM、20-PGD、MI-FGSM 攻击结果见图 14、图 15 和图 16 所示。

红色线(优化模型)精度明显优于蓝色线(原模型)的精度。即随着攻击强度的增加，两者在抗干扰能力上有明显的差异。说明在该数据集上，利用本文提出的纹理特征约束方法能较好的提高模型的鲁棒性。

相比较于 Coffee 数据集，该方法在 ECG200 数据集上更有效的限制住纹理特征的比例，降低干扰噪

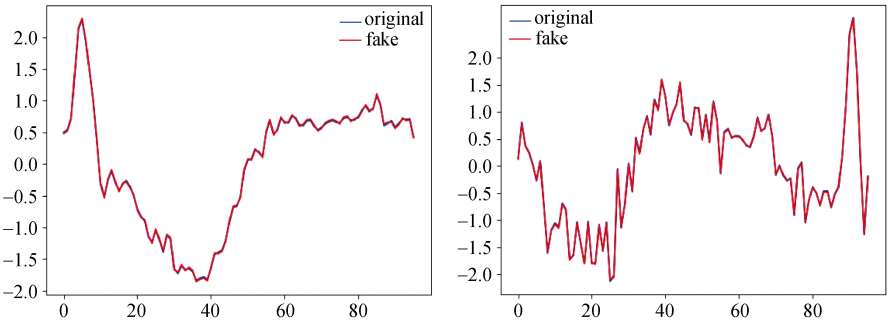

图 13　ECG200 数据集两个类别的样本

Figure 13　Example samples from ECG200 data set

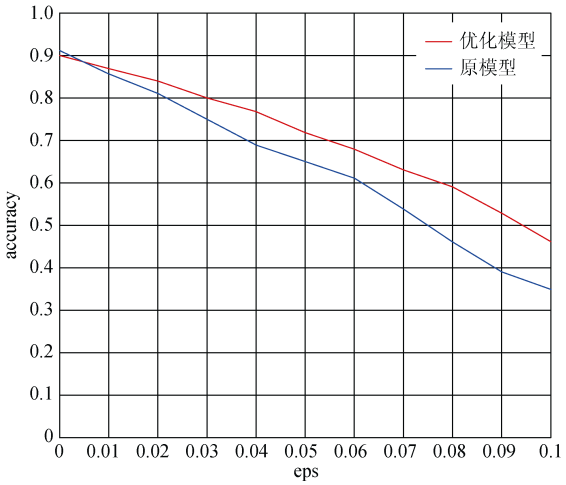

图 14　ECG200 数据集在不同攻击率 FGSM 设置下的准确率, 红色为优化模型准确率, 蓝色为原模型的准确率

Figure 14　The accuracy of ECG200 dataset under different attack rate FGSM settings. Red represents the accuracy of the optimization model and blue represents the accuracy of the original model

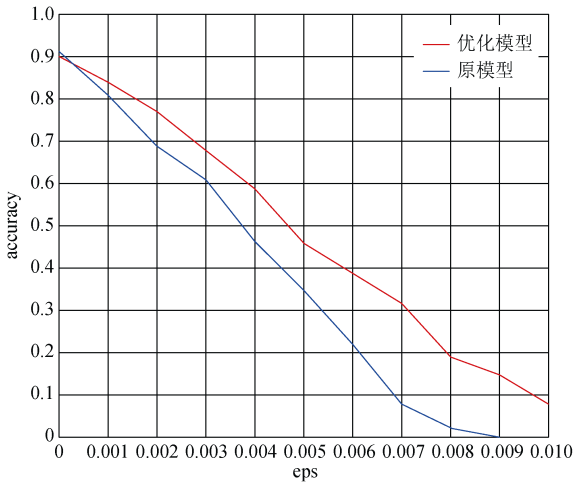

图 15　ECG200 数据集在不同攻击率 20-PGD 设置下的准确率, 红色为优化模型准确率, 蓝色为原模型的准确率

Figure 15　The accuracy of ECG200 dataset under different attack rate 20-PGD settings. Red represents the accuracy of the optimization model and blue represents the accuracy of the original model

声的影响。同时由于该数据集的类别间样本的差异大, 在训练模型的时候, 能达到准确率和鲁棒性同时提高的效果。

5.2　更多数据集

为了验证本文方法的有效性, 在更多数据集上测试本文方法, 其中部分数据集的攻击防御效果见表 2。

表 2 结果显示, 针对不同类型的数据集, 本文方法可以在一定程度上提高模型的抗干扰能力, 即模型鲁棒性。但在应对不同噪声等级的攻击时, 抗干扰能力有明显的不同。有的数据集可以保障较高准确率, 有的数据集无明显效果。这与数据集的各类别间的差异有很大关系, 如果数据集类别之间本身的差异较大, 则很容易通过约束模型学习到形态特征,而不是纹理特征。因此, 本文方法主要针对那些类别间差异较大的数据集, 在该类数据集上可以取得较好效果。针对类别间差异较小的数据集, 效果有限。

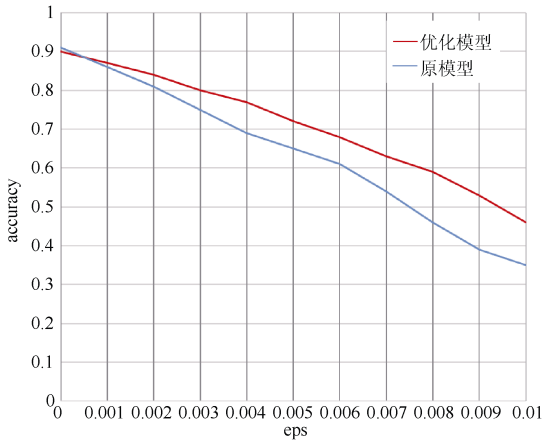

图 16　ECG200 数据集在不同攻击率 MI-FGSM 设置下的准确率，红色为优化模型准确率，蓝色为原模型的准确率

Figure 16　The accuracy of ECG200 dataset under different attack rate MI-FGSM settings. Red represents the accuracy of the optimization model and blue represents the accuracy of the original model

表 2　本文方法在更多数据上测试结果

Table 2　The experimental results for more datasets

数据集名称	数据集描述	攻击方法(ε)	准确率%	
			原始模型	本文方法
UMD	类型: 模拟数据; 类别: 4; 序列长度 150; 训练集: 36; 测试集: 144	FGSM(0)	92	92
		FGSM(0.01)	92	92
		FGSM(0.05)	64	83
		FGSM(0.1)	36	50
Trace	类型: 传感器; 类别: 4; 序列长度 275; 训练集: 100; 测试集: 100	FGSM(0)	100	100
		FGSM(0.01)	89	100
		FGSM(0.05)	62	69
		FGSM(0.1)	48	48
PowerCons	类型: 能源; 类别: 2; 序列长度 144; 训练集: 180; 测试集: 180	FGSM(0)	92	92
		FGSM(0.1)	81	83
		FGSM(0.5)	46	49
ToeSegmentation1	类型: 运动轨迹; 类别: 2; 序列长度 277; 训练集: 40; 测试集: 228	FGSM(0)	93	93
		FGSM(0.05)	80	82
		FGSM(0.1)	38	40
		FGSM(0.5)	30	35
ToeSegmentation2	类型: 运动轨迹; 类别: 2; 序列长度 343; 训练集: 36; 测试集: 130	FGSM(0)	97	97
		FGSM(0.05)	75	78
		FGSM(0.1)	69	72
		FGSM(0.5)	50	53
ShapeletSim	类型: 模拟数据; 类别: 2; 序列长度: 500; 训练集: 20; 测试集: 180	FGSM(0)	90	100
		FGSM(0.01)	85	95
		FGSM(0.03)	70	75

6　结论

本文分析了时序分类模型中对抗样本存在的理论依据。从纹理特征和形态特征的处理, 分析了噪声逐步放大的依据, 从数据的角度说明了对抗样本存在的可能性, 并借助形式化表达从理论上说明了该过程。基于纹理特征的脆弱性, 设计了相应的约束方法, 在训练过程中, 通过约束纹理特征的比例, 有方向性的引导模型学习到更多的形态特征, 从而在保证分类精度的同时, 降低对抗样本存在的空间, 提高模型的鲁棒性。

本文方法在多个攻击方法和多个时序数据验证下, 得到了较好的精度, 证明了该方法的效果。但是, 该方法应用在一些类别间样本形态差距很小的数据集上, 效果有限。其原因使模型为了得到足够高分类的准确率, 需要学习到足够多的纹理特征, 而纹理特征的存在又使得细小噪声被逐层放大。未来的研究将考虑针对此类数据集, 修改训练过程, 学习到足够好的形态特征, 从而提高模型鲁棒性。

参考文献

[1] Karim F, Majumdar S, Darabi H. Adversarial Attacks on Time Series[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(10): 3309-3320.

[2] Sirisambhand K, Ratanamahatana C A. A Dimensionality Reduction Technique for Time Series Classification Using Additive Representation[M]. Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 2018: 717-724.

[3] Ismail Fawaz H, Forestier G, Weber J, et al. Deep Learning for Time Series Classification: A Review[J].*Data Mining and Knowledge Discovery*, 2019, 33(4): 917-963.

[4] Ma T F, Xiao C, Wang F. Health-ATM: A Deep Architecture for Multifaceted Patient Health Record Representation and Risk Prediction[M]. Proceedings of the 2018 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2018: 261-269.

[5] Yan L M, Li Y, Du B, et al. Dynamic Gesture Recognition Based on Key Feature Points Trajectory[J]. *Optoelectronic Technology*, 2015, 35(3): 187-190.
(严利民, 李跃, 杜斌, 等. 基于关键特征点运动轨迹的动态手势识别[J]. *光电子技术*, 2015, 35(3): 187-190.)

[6] Zheng Z B, Yang Y T, Niu X D, et al. Wide and Deep Convolu-

tional Neural Networks for Electricity-Theft Detection to Secure Smart Grids[J]. *IEEE Transactions on Industrial Informatics*, 2018, 14(4): 1606-1615.

[7] Dau H A, Bagnall A, Kamgar K, et al. The UCR Time Series Archive[J]. *IEEE/CAA Journal of Automatica Sinica*, 2019, 6(6): 1293-1305.

[8] Dong Y P, Su H, Wu B Y, et al. Efficient decision-based black-box adversarial attacks on face recognition[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 7706-7714.

[9] Karim F, Majumdar S, Darabi H, et al. LSTM Fully Convolutional Networks for Time Series Classification[J]. *IEEE Access*, 6: 1662-1669.

[10] Karim F, Majumdar S, Darabi H, et al. Multivariate LSTM-FCNS for Time Series Classification[EB/OL]. 2018: arXiv: 1801.04503. https://arxiv.org/abs/1801.04503.

[11] Papernot N, McDaniel P, Goodfellow I. Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples[EB/OL]. 2016: arXiv: 1605.07277. https://arxiv.org/abs/1605.07277

[12] Oregi I, del Ser J, Perez A, et al. Adversarial Sample Crafting for Time Series Classification with Elastic Similarity Measures[M]. Intelligent Distributed Computing XII. Cham: Springer International Publishing, 2018: 26-39.

[13] Akhtar N, Mian A. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey[J]. *IEEE Access*, 6: 14410-14430.

[14] Tramèr F, Kurakin A, Papernot N, et al. Ensemble Adversarial Training: Attacks and Defenses[EB/OL]. 2017: arXiv: 1705.07204. https://arxiv.org/abs/1705.07204

[15] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[EB/OL]. 2017: arXiv: 1706.06083. https://arxiv.org/abs/1706.06083

[16] Carlini N, Wagner D, Communication N A B T, et al. Audio adversarial examples: Targeted attacks on speech-to-text[C]. *2018 IEEE Security and Privacy Workshops*, 2018: 1-7.

[17] Duan G H, Ma C G, Song L, et al. Research on Structure and Defense of Adversarial Example in Deep Learning[J]. *Chinese Journal of Network and Information Security*, 2020, 6(2): 1-11.
(段广晗, 马春光, 宋蕾, 等. 深度学习中对抗样本的构造及防御研究[J]. *网络与信息安全学报*, 2020, 6(2): 1-11.)

[18] Galloway A, Golubeva A, Tanay T, et al. Batch Normalization is a Cause of Adversarial Vulnerability[EB/OL]. 2019: arXiv: 1905.02161. https://arxiv.org/abs/1905.02161

[19] Wang H H, Wu X D, Huang Z Y, et al. High-frequency component helps explain the generalization of convolutional neural networks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 8681-8691.

[20] Yosinski J, Clune J, Bengio Y, et al. How Transferable are Features in Deep Neural Networks? [C]. *The 27th International Conference on Neural Information Processing Systems - Volume 2*, 2014: 3320-3328.

[21] Shi B F, Zhang D H, Dai Q, et al. Informative Dropout for Robust Representation Learning: A Shape-Bias Perspective[C]. *The 37th International Conference on Machine Learning*, 2020: 8828-8839.

[22] Ismail Fawaz H, Forestier G, Weber J, et al. Adversarial attacks on deep neural networks for time series classification[C]. *2019 International Joint Conference on Neural Networks*, 2019: 1-8.

[23] Sarkar A, Raj AS, Iyengar RS. Improving Robustness of time series classifier with Neural ODE guided gradient based data augmentation[EB/OL]. 2019: ArXiv Preprint ArXiv:1910.06813.

[24] Wang Y C, Emonet R, Fromont E, et al. Learning Interpretable Shapelets for Time Series Classification through Adversarial Regularization[EB/OL]. 2019: arXiv: 1906.00917. https://arxiv.org/abs/1906.00917

[25] Yang Z G, Li H, Zhang M Z, et al. A Method for Resisting Adversarial Attack on Time Series Classification Model in IoT System[M]. Web Information Systems and Applications. Cham: Springer International Publishing, 2020: 559-566.

[26] Yang Z G, Abbasi I A, Algarni F, et al. An IoT Time Series Data Security Model for Adversarial Attack Based on Thermometer Encoding[J]. *Security and Communication Networks*, 2021, 2021: 1-11.

[27] Jiang H, Nai H, Jiang Y, et al. An Adversarial Examples Identification Method for Time Series in Internet-of-Things System[J]. *IEEE Internet of Things Journal*, 2021, 8(12): 9495-9510.

[28] Tobiyama S, Yamaguchi Y, Shimada H, et al. Malware detection with deep neural network using process behavior[C]. *2016 IEEE 40th Annual Computer Software and Applications Conference*, 2016: 577-582.

[29] Wang Z G, Yan W Z, Oates T, et al. Time series classification from scratch with deep neural networks: A strong baseline[C]. *2017 International Joint Conference on Neural Networks*, 2017: 1578-1585.

[30] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[EB/OL]. 2015: arXiv: 1502.03167. https://arxiv.org/abs/1502.03167

[31] Ismail Fawaz H, Forestier G, Weber J, et al. Deep neural network ensembles for time series classification[C]. *2019 International Joint Conference on Neural Networks*, 2019: 1-6.

[32] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. 2014: arXiv: 1412.6572. https://arxiv.org/abs/1412.6572

[33] Shafahi A, Saadatpanah P, Zhu C, et al. Adversarially Robust Transfer Learning[EB/OL]. 2019: arXiv: 1905.08232. https://arxiv.org/abs/1905.08232

[34] Dong Y P, Liao F Z, Pang T Y, et al. Boosting Adversarial Attacks with Momentum[EB/OL]. 2017: arXiv: 1710.06081. https://arxiv.org/abs/1710.06081

[35] Bjorck J, Gomes C, Selman B. Understanding Batch Normalization[J/OL]. 32nd Conference on Neural Information Processing Systems.2018. arXiv:1806.02375v1.

杨中国 于2018年在中国石油大学(北京)计算机技术与资源信息工程专业获得博士学位。现任北方工业大学信息学院人工智能系助理研究员。研究领域为对抗攻击、流数据处理。研究兴趣包括: 模型安全、服务计算。Email: yangzhongguo@ncut.edu.cn

王丽君 于 2020 年在济宁医学院计算机科学与技术专业获得学士学位。现在北方工业大学学校电子信息专业攻读硕士学位。研究领域为大规模流数据集成与分析。研究兴趣包括: 对抗攻击、服务计算。Email: wlijun808@163.com

张镌 于 2021 年在北方工业大学信息安全获得学士学位。研究领域为信息安全, 对抗攻击。研究兴趣包括: 深度学习模型安全。Email: 1979174801@qq.com