LFDP：融合低频信息的差分隐私鲁棒性增强方法

王 豪[1,2], 许 强[3], 张清华[4], 李开菊[1,5]

[1] 重庆邮电大学 计算机科学与技术学院 重庆 中国 400065
[2] 旅游多源数据感知与决策技术文化和旅游部重点实验室 重庆 中国 400065
[3] 香港城市大学 电机工程系 香港 中国 999077
[4] 重庆邮电大学 计算智能重庆市重点实验室 重庆 中国 400065
[5] 重庆大学 计算机学院 重庆 中国 400044

摘要 机器学习模型由于其预测和分类的高精度和各种应用场景的普适性, 在图像处理、自动驾驶、自然语言处理等领域得到广泛应用。但机器学习模型容易遭受对抗样本攻击, 在遭受对抗样本攻击时, 预测和分类的精度会大幅下降。目前, 数据增强方法通过改变或者扰动原始图像的方式, 使得机器学习模型具有更强的泛化能力, 在保护隐私的同时, 能够增强其抵御对抗样本攻击的鲁棒性, 是当前机器学习模型鲁棒性增强的主流方法之一。但基于差分隐私的鲁棒性增强方法面临加入的高频噪声容易被滤除, 导致鲁棒性增强效果下降的问题。针对这一问题, 结合信号处理的知识, 本文从频域角度阐述差分隐私能够增强机器学习模型鲁棒性的原理, 从理论上证明其有效性。设计了一种高频噪声滤波器 HFNF, 能够将差分隐私加入的高频高斯噪声滤除, 使得差分隐私的鲁棒性增强效果下降, 从理论上分析差分隐私鲁棒性增强方法存在缺陷的原因。提出了一种普适的融合低频信息的差分隐私鲁棒性增强算法 LFDP, 通过对图像不同频域部分加入生成的高低频噪声, 即使存在高频噪声滤波攻击, 仍然能够保证模型的鲁棒性, 弥补了差分隐私原有高频高斯噪声的不足。从理论上分析并给出所提出方案的鲁棒性和误差边界, 并在实际的数据集中进行测试。实验结果表明, 与直接加入高频噪声的差分隐私鲁棒性增强方法相比, LFDP 在不增大噪声尺度的同时能够起到更好的鲁棒性增强效果。

关键词 机器学习; 鲁棒性; 差分隐私; 低频噪声
中图法分类号 TP309.2 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2025.01.04

LFDP: A Differentially Private Robustness Augmentation Method Combining Low-Frequency Information

WANG Hao[1,2], XU Qiang[3], ZHANG Qinghua[4], LI Kaiju[1,5]

[1] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] Key Laboratory of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[3] Department of Electrical Engineering, City University of Hong Kong, Hongkong 999077, China
[4] Key Laboratory of Big Data Intelligent Computing, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[5] College of Computer Science, Chongqing University, Chongqing 400044, China

Abstract Machine learning model has been widely used in image processing, automatic driving, natural language processing and other fields because of its high accuracy of prediction and classification and the universality of various application scenarios. However, the machine learning model is vulnerable to counter sample attacks. When it is attacked by counter sample attacks, the accuracy of prediction and classification will be greatly reduced. At present, the data enhancement method makes the machine learning model have stronger generalization ability by changing or disturbing the original image, and can enhance its robustness against sample attacks while protecting privacy, which is one of the mainstream methods for enhancing the robustness of machine learning models. However, the robustness enhancement method based on differential privacy is faced with the problem that the added high-frequency noise is easy to be filtered out, resulting in a decline in the robustness enhancement effect. Aiming at this problem, combined with the knowledge of signal processing, this paper expounds the principle that differential privacy can enhance the robustness of machine learning models from the perspective of frequency domain, and proves its effectiveness in theory. A high frequency

通讯作者： 李开菊, 博士, 讲师, Email: likaiju@mail.gufe.edu.cn。

本课题得到国家自然科学基金(No. 42001398，No.62402150，No. 62276038)、国家重点研发计划课题(No. 2020YFC2003502)、重庆市教委科学技术研究重点项目(No. KJZD-K202300601)、贵州财经大学引进人才科研启动基金(No. 2023YJ10)、旅游多源数据感知与决策技术文化和旅游部重点实验室开放基金资助项目(No. TMDPD-2023N-002)、贵州省教育厅青年科技成长项目(黔教技[2024]86)、重庆邮电大学计算机学院人才梯队提升计划项目(No. JKY-202423)、贵州省科技计划项目(黔科合成果[2024]重大 018)资助
收稿日期: 2022-12-11; 修改日期: 2023-03-11; 定稿日期: 2024-11-20

noise filter HFNF is designed, which can filter out the high frequency Gaussian noise added by differential privacy and reduce the robustness enhancement effect of differential privacy. The reason for the defects of the robustness enhancement method of differential privacy is analyzed theoretically. This paper proposes a universal differential privacy robustness enhancement algorithm LFDP, which fuses low frequency information. By adding high and low frequency noise generated in different frequency domain parts of the image, even if there is high frequency noise filtering attack, the robustness of the model can still be guaranteed, making up for the deficiency of the original high frequency Gaussian noise in differential privacy. The robustness and error boundary of the proposed scheme are theoretically analyzed and given, and tested in actual data sets. The experimental results show that compared with the difference privacy robustness enhancement method directly adding high-frequency noise, LFDP can play a better robustness enhancement effect without increasing the noise scale.

Key words machine learning; robustness; differential privacy; low-frequency noise

1 引言

机器学习算法由于能够提供较好的训练和预测精确度, 以及不同领域的高度适用性, 目前已经广泛地应用于图像处理、自动驾驶、自然语言处理等领域。但机器学习算法在进行广泛应用的同时, 最新的研究表明, 机器学习算法可能会遭受到对抗性样本的影响[1-3], 例如, 可以在原始图像中加入肉眼几乎不可见的扰动, 从而改变机器学习算法的分类或预测结果。目前, 利用 FGSM[4-5]、PGD[6-7]等经典对抗样本方法进行攻击, MNIST 数据集的错误分类率可以达到 99%, 机器学习算法的对抗样本安全威胁问题受到越来越多研究者和工业界的关注。

为此, 研究人员已经探索了多种方案来避免对抗样本的安全威胁。其中, 数据增强方法通过改变或者扰动原始图像的方式, 使得机器学习模型具有更强的泛化能力, 从而增强其抵御对抗样本攻击的鲁棒性, 也是当前机器学习模型鲁棒性增强的主流方法之一。在基于数据增强的方法中, 差分隐私(DP)[8-10]通过对原始图像加入噪声, 能够在保护隐私的同时, 增强机器学习模型的鲁棒性, 展示了其在隐私保护和鲁棒性增强两个方面的优越性。

虽然现有研究已经证实[11-12], 差分隐私在保护图像隐私的同时, 能够增强机器学习算法的鲁棒性, 但仍然面临以下两个方面的开放性问题有待解决:

1) 差分隐私增强机器学习模型鲁棒性的原理尚不清晰。由于机器学习模型的不可解释性, 现有研究仅仅从实验结果说明差分隐私鲁棒性增强的有效性, 但目前尚未有研究者能够从理论上证明或阐述其有效性的本质原因;

2) 差分隐私增强机器学习模型鲁棒性的普适性有限。由于差分隐私是一种基于高频噪声扰动的数据增强方法, 而图像数据由高频和低频两部分组成, 因此, 差分隐私加入的高频噪声面临被滤除的问题。

在本文中, 我们试图回答上述开放性问题, 并对差分隐私的鲁棒性增强原理和限制进行深入研究。本文首先分析了差分隐私能够增强机器学习鲁棒性的原理, 设计了一种高频噪声滤波方法(High frequency noise filtering, HFNF), 指出基于差分隐私的数据增强方法所面临的问题, 接着设计了融合低频噪声的差分隐私噪声生成方法(Low frequency differential privacy, LFDP), 增强了差分隐私的实用性。本文的贡献主要包括以下 4 个方面:

1) 结合信号处理的知识, 从频域角度阐述差分隐私能够增强机器学习模型鲁棒性的原理, 从理论上证明其有效性;

2) 设计了一种高频噪声滤波器 HFNF, 能够将差分隐私加入的高频高斯噪声滤除, 使得差分隐私的鲁棒性增强效果下降, 从理论上分析差分隐私鲁棒性增强方法存在缺陷的原因;

3) 提出了一种普适的融合低频信息的差分隐私鲁棒性增强算法 LFDP, 通过对图像不同频域部分加入生成的高低频噪声, 即使存在高频噪声滤波攻击, 仍然能够保证模型的鲁棒性, 弥补了差分隐私原有高频高斯噪声的不足;

4) 从理论上分析并给出所提出方案的鲁棒性和误差边界, 并在实际的数据集中进行测试。实验结果表明, 相比于直接加入高频噪声的方式, 融合低频信息的方法能够显著增强机器学习模型的鲁棒性。

本文第 2 节介绍了与机器学习鲁棒性增强的相关工作; 第 3 节阐述了差分隐私鲁棒性增强的原理和不足, 提出了融合低频信息的差分隐私鲁棒性增强方法, 并给出了所提出方法性能的理论分析; 第 4 节给出实验结果与对比分析; 第 5 节总结全文并对未来的研究工作进行展望。

2 相关工作

本文主要研究基于差分隐私的机器学习模型鲁棒性增强方法, 与本文研究相关的工作主要包括 3 个方面, 分别是数据增强、差分隐私以及频域鲁棒性

增强三个方面, 下面分别对这 3 个方面的相关工作进行阐述。

2.1 数据增强

数据增强技术是利用反转、加噪等方式来扩充训练数据的特征空间, 从而增强训练模型的鲁棒性。如Pinot等人[13]将对抗性攻击和防御问题描述为一个无穷和博弈问题, 他们证明了当分类器和对手都是确定性的时候, 博弈中不存在纳什均衡, 从而在确定性的情况下对上述问题给出了否定的回答。Araujo等人[14]证明了 l_∞ 的保护机制无法针对 l_2 范数的攻击提供很好的防御, 反之亦然。李等人[15]介绍了一个可扩展的框架, 并为构造对抗性示例提供输入操作规范的证明边界。Cohen 等人[16]展示了如何将在高斯噪声下分类良好的分类器转换成一个新的分类器, 该分类器在 l_2 范数条件下对不利干扰具有鲁棒性。此外, Pinot 等人[13]研究了对抗性攻击的鲁棒性理论。

2.2 差分隐私

作为一种噪声加扰方法, 差分隐私从直觉上来说可以增强机器学习模型的鲁棒性, 现有学者研究了对抗学习与差分隐私之间的关系。例如, Lecuyer等人[17-18]提出了 PixelDP 机制, 分析了对抗性实例和差分隐私之间的关系, 提出了一种通用、灵活的差分隐私鲁棒性增强机制。Phan 等人[12,18]开发出一种可扩展的算法, 以保证深层神经网络(DNN)对抗性学习中的差分隐私(DP), 算法从数学上验证了对对抗性示例的鲁棒性 Xu 等人[19]提出了 GanobFouscator, 一种有区别的隐私保护器, 它可以通过在学习过程中向梯度中添加精心设计的噪声来实现不同等级的隐私保护。Pinot 等人[20-21]证明, 差分隐私和对抗性实例的鲁棒性建立在相同的理论基础上, 因此, 在一个领域获得的结果可以转移到另一个领域。

2.3 频域鲁棒性增强

由于机器学习模型的不可解释性, 现有研究试图从频域角度探索图像频率与数据增强方法之间的关系。Yin 等人[22]发现高斯数据增强和对抗训练都提高了图片高频部分的模型鲁棒性, 同时降低了图片低频部分的模型鲁棒性。Wang 等人[23]研究了图像数据的频谱与卷积神经网络(CNN)泛化行为之间的关系。郭等人[24]建议将对抗性图像的搜索限制在低频部分, 从而以较高的效率检测出对抗样本。

差分隐私[25-27]作为一种隐私保护方法, 在机器学习领域中具有很好的保护效果, 同时, 差分隐私作为一种噪声扰动机制, 本身也是一种数据增强方法。本节分别阐述了几个子领域的工作进展, 但差分隐私生成的高斯噪声是高频噪声, 直接将差分隐私生成的高频噪声加入到原始数据中, 数据增强的效果有限, 这需要研究几个子领域的内在关系, 从而提出相应的增强方法。

3 方案原理

本节首先从理论上分析证明差分隐私能够增强机器学习模型鲁棒性的原因, 接着指出差分隐私鲁棒性增强的缺陷, 提出了高频噪声滤波的攻击算法(HFNF), 最后提出了融合低频信息的差分隐私高斯噪声生成方法(LFDP), 以提高差分隐私的鲁棒性。

3.1 差分隐私鲁棒性增强原理和缺陷

在利用差分隐私保护图像隐私时, 现有方法首先将原始图片进行傅里叶变换, 然后在变换后的频域系数上加入高斯噪声, 最后进行反变换得到含有噪声的图像[28-29]。首先, 我们分析在这个过程中, 对系数加入噪声进行反变换之后得到的噪声尺度的形式和大小, 如定理 1 所阐述。

定理 1. 记原始图像的尺寸为 $M\times N$, 如果把标准差为 σ 的高斯噪声添加到傅里叶变换系数中, 则等同于把标准差为 $\frac{\sigma}{\sqrt{M\times N}}$ 的高斯噪声添加到原始图像中。

证明. 如果用 $z[n]$ 表示噪声的离散域形式, 那么 $z[n]$ 可以表示为:

$$z[n]=a[n]+j\cdot b[n]$$

其中, $a[n]$ 和 $b[n]$ 分别为

$$f_{a[n]}(z)=f_{b[n]}(z)=\frac{1}{\sqrt{2\pi/M\times N}\sigma}e^{-\frac{a^2[n]}{2\sigma^2/(M\times N)}}$$

于是, 高斯分布的傅里叶变换为:

$$Z[k]=A[k]+j\cdot B[k]=\sum_{n=0}^{M\times N-1}(a[n]+j\cdot b[n])\cdot e^{-2\pi j\frac{nk}{M\times N}}$$

如果我们将 $Z[n]$ 的实部和虚部分离, 可以得到

$$A[k]=\sum_{n=0}^{M\times N}(a[n]\cos\frac{2\pi nk}{M\times N}+b[n]\sin\frac{2\pi nk}{M\times N})$$

$$B[k]=\sum_{n=0}^{M\times N}(b[n]\cos\frac{2\pi nk}{M\times N}-a[n]\sin\frac{2\pi nk}{M\times N})$$

接下来我们关注 $a[n]$, 因为 $a[n]\cos\frac{2\pi nk}{M\times N}$ 中的 $\cos\frac{2\pi nk}{M\times N}$ 是一个与随机变量 $a[n]$ 无关的常量, 根据这些等式, 我们有:

$$f_{a[n]\cos\frac{2\pi nk}{M\times N}}(y)=\frac{1}{|\cos\frac{2\pi nk}{M\times N}|\sigma\sqrt{2\pi/M\times N}}e^{\frac{-a^2[n]}{\frac{2\sigma^2}{M\times N}\cos^2\frac{2\pi}{nk}}}$$

如果我们将上述等式进行拉普拉斯变换, 可以得到:

$$M_{a[n]\cos\frac{2\pi nk}{M\times N}}(s)=\int_{-\infty}^{+\infty}f_{a[n]\cos\frac{2\pi nk}{M\times N}}(y)\cdot e^{sy}dy=e^{\frac{s^2\sigma^2}{2M\times N}\cos^2\frac{2\pi nk}{M\times N}}$$

类似的, 有:

$$M_{b[n]\cos\frac{2\pi nk}{M\times N}}(s)=e^{\frac{s^2\sigma^2}{2M\times N}\sin^2\frac{2\pi nk}{M\times N}}$$

于是, 通过拉普拉斯变换, $r[n]=a[n]\cos\frac{2\pi nk}{M\times N}+b[n]\sin\frac{2\pi nk}{M\times N}$ 可以转化为:

$$\begin{aligned}M_{r[n]}(s)&=M_{a[n]\cos\frac{2\pi nk}{M\times N}}(s)\cdot M_{b[n]\sin\frac{2\pi nk}{M\times N}}(s)\\&=e^{\frac{s^2\sigma^2}{2M\times N}(\sin^2\frac{2\pi nk}{M\times N}+\cos^2\frac{2\pi nk}{M\times N})}\\&=e^{\frac{s^2\sigma^2}{2M\times N}}\end{aligned}$$

由于在 $r[n]$ 中的变量是各自独立的, $A[k]$ 的变换可表示为:

$$M_{A[k]}(s)=\prod_{n=0}^{M\times N-1}M_{r[n]}(s)=e^{\frac{s^2\sigma^2}{2}}$$

如果我们对最后一个等式进行逆变换, 可以得到关于 $A[k]$ 的概率密度函数公式:

$$f_{A[k]}(z)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}$$

同理, 因为概率密度函数中的 $a[n]$ 和 $b[n]$ 相同, 则有

$$f_{B[k]}(z)=f_{A[k]}(z)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}$$

定理 1 证明, 在图像的变换系数上加入高斯噪声等同于在原始图像中加入噪声, 并给出了等价的噪声尺度。从定理 1 可以看出, 从信号处理的角度来看, 差分隐私实际上也是一种数据增强的鲁棒性方法, 其原理等价于将差分隐私的高斯噪声加入到原始图像中, 从而在保护隐私的同时增强训练模型的鲁棒性, 下面我们分析这种噪声形式所存在的缺陷。

由于差分隐私加入的噪声是高斯噪声, 而图像数据往往由高频噪声和低频数据组成, 其中低频数据代表其轮廓而高频数据代表其细节, 下面我们探索高斯噪声加入图像之后的机器学习模型的鲁棒性变化情况, 如定理 2 所述。

定理 2. 如果 X_l 和 X_h 分别代表图片的低频和高频分量, R_l 和 R_h 分别代表分量的鲁棒性半径, 那么把标准方差为 $\frac{\sigma}{\sqrt{M\times N}}$ 的高斯噪声加入到原始图像 X 之后, 机器学习模型的鲁棒性半径为

$$R_h<\frac{\sigma}{2\sqrt{M\times N}}(\varphi^{-1}(p_A)-\varphi^{-1}(p_B)) \quad (1)$$

其中 p_A 是分类结果中的最大概率, p_B 是分类结果中第二大的概率。

证明. 令 $g(x)=\arg\max_c P(h(x+\eta)=c)$, 对于任意的 $c_A\in\gamma$, 其中 γ 是模型输出的数值域, 则定理 2 等价于证明:

$$P(h(x+\eta)=c)\geqslant p_A\geqslant p_B\geqslant\max_{c\neq c_A}P(h(x+\eta)=c) \quad (2)$$

其中, 对于 $g(x+\eta)=c_A$, 任意 $\|\delta\|_2<R_h$, 有 $R_h=\frac{\sigma}{2\sqrt{M\times N}}(\varphi^{-1}(p_A)-\varphi^{-1}(p_B))$。根据 $g(\cdot)$ 函数的定义, 有

$$R_h=\frac{\sigma}{2\sqrt{M\times N}}(\varphi^{-1}(p_A)-\varphi^{-1}(p_B))$$

因为只解决某一类 C_B 有失普遍性, 所以对于每一类 $C_B\neq C_A$, 我们都会证明 $P(f(x+\delta+\varepsilon)=c_B)$。为简便起见, 定义随机变量

$$X':=x+\eta=N(x',\frac{\sigma^2I}{M\times N}),$$

$$\tilde{X}:=x+\delta+\eta=N(x'+\delta,\frac{\delta^2I}{M\times N})$$

根据公式(2)可得

$$P(f(x+\varepsilon)=c_A)\geqslant\underline{p_A}$$
$$P(f(x+\varepsilon)=c_B)\geqslant\overline{p_B}$$

那么, 我们的目标是证明

$$P(f(\tilde{X})=c_A)>P(f(\tilde{X})=c_B)$$

接着定义半空间:

$$A:=\{o:\delta^T(o-x)\leqslant\sigma\|\delta\|\varphi^{-1}(p_A)\}$$
$$B:=\{o:\delta^T(o-x)\leqslant\sigma\|\delta\|\varphi^{-1}(1-p_B)\}$$

由上述两个公式可以得到 $P(X'\in A)=\underline{p_A}$。因此, 通过公式我们得知 $P(f(X')=c_A)\geqslant P(X'\in A)$。因此, 我们运用引理 2[16]和 $h(o):=\mathbb{1}[f(o)=c_A]$ 来综合得出:

$$P(f(\tilde{X})=c_A)\geqslant P(\tilde{X}\in A)$$

类似的, 代数式表明 $P(X'\in B)=\overline{p_B}$。因此, 通过公式(2)可以得到

$$P(f(X')=c_B)\leqslant P(X'\in B)$$

另外, 根据文献[16]中的引理 2 和 $h(o):=1[f(o)=c_B]$ 得到

$$P(f(\tilde{X})=c_B)\leqslant P(\tilde{X}\in B)$$

根据上述公式可以得到 $P(\tilde{X}\in A)>P(\tilde{X}\in B)$, 由此可以推导出下述不等式:

$$P(f(\tilde{X})=c_A)\geqslant P(\tilde{X}\in A)>P(\tilde{X}\in B)\geqslant P(f(\tilde{X})=c_B)$$

另外, 由于

$$P(\tilde{X}\in A)=\varphi(\varphi^{-1}(p_A)-\frac{\|\delta\|}{\sigma\sqrt{M\times N}})$$

$$P(\tilde{X}\in B)=\varphi(\varphi^{-1}(p_B)+\frac{\|\delta\|}{\sigma\sqrt{M\times N}})$$

最后, 当且仅当 $\|\delta\|<\frac{\sigma}{2\sqrt{M\times N}}(\varphi^{-1}(\underline{p_A})-\varphi^{-1}(\overline{p_B}))$ 成立时, 有

$$P(\tilde{X}\in A)>P(\tilde{X}\in B)$$

由定理 2 可知, R_h 正比于 σ 的值, 且值恒大于 0, 说明在变换系数中加入高斯噪声后, 图像高频部分的鲁棒性半径增大, 这也是差分隐私的噪声扰动能够增强鲁棒性的原因。

3.2 高频噪声滤波攻击算法 HFNF

定理 2 给出了差分隐私的高斯噪声加入变换系数后, 机器学习模型鲁棒半径的理论值。但由于图像由低频和高频分量组成, 而噪声只是高频分量, 根据现有的图像处理方法, 我们可以设计特定的滤波器[30]将高斯噪声滤除, 如定理 3 所述。

定理 3. 在将标准方差为 $\frac{\sigma}{\sqrt{M\times N}}$ 的高斯噪声加入原始图像, 再通过特定的低通滤波器后, 机器学习模型的鲁棒性半径为:

$$\hat{R}_h=\frac{\hat{\sigma}}{2\sqrt{M\times N}}(\varphi^{-1}(\underline{p_A})-\varphi^{-1}(\overline{p_B})) \tag{3}$$

其中, $\hat{\sigma}=\sqrt{\frac{2\varepsilon}{\|X''\|}}$。

证明. 用 $\hat{X}$ 表示过滤后的图像, 根据 Hodson 等人[28]的结论, $\hat{X}$ 可被扩展为:

$$\hat{X}=X+\frac{X''}{2}\hat{\sigma}^2+\cdots+\frac{X^{2m}}{\prod_{p=1}^{m}2p}+\cdots$$

其中, $m=1,2,\cdots,N,X''$ 是 X 的二阶导数。于是上面的方程可以近似表达为 $\hat{X}\cong X+\frac{X''}{2}\hat{\sigma}^2$。

过滤后噪声的标准方差是 $\hat{\sigma}=\sqrt{\frac{2\alpha}{\|X''\|}}$, 其中 $\|\hat{X}-X\|\leqslant\varepsilon$。根据定理 2, 有

$$\hat{R}_h=\frac{\hat{\sigma}}{2\sqrt{M\times N}}(\varphi^{-1}(\underline{p_A})-\varphi^{-1}(\overline{p_B}))$$

定理 3 表明, 如果加入的噪声是差分隐私噪声, 加入噪声后, 攻击者可以设计相应的高频滤波器, 滤除部分噪声, 导致鲁棒性半径减少, 即鲁棒性降低。

在实际应用中设计滤波器时, 我们可以直接采用高斯滤波器, 高斯滤波器可看成是一种带权的均值滤波器, 而均值滤波器的特点是用以某点为中心的窗口内所有像素点的平均值来代替该点的像素, 从而达到平滑噪声的目的。

在图像处理领域, 一般将高斯滤波器默认为二维高斯滤波器, 通常情况下, 在使用高斯滤波器对二维图像进行平滑处理时, 高斯函数的方差 σ^2 是一个固定的值。在图像的细节区域使用较大的 σ^2 会造成过度平滑, 使之失去细节信息; 在平坦的区域, 使用较大的 σ^2 却是一个比较好的选择, 因此, σ^2 的选择对于滤波结果至关重要。

为此, 本文设计了一种自适应高斯滤波器, 可以根据被平滑的图像的局部特征, 选择不同的 σ^2, 使得在平滑后的结果图像中保持细节信息。其主要思想是要在最平滑的结果和最佳的保持细节信息的结果中取一个折中方案, 它使用了一个能量函数, 该函数的形式为:

$$\sigma_f=\left[\arg\min_{\sigma}\frac{c}{\sigma^2}+\varsigma^2\right] \tag{4}$$

其中, σ_f 是最佳的高斯滤波器方差, c 是常数, ς 表示原始图像的像素值与高斯平滑灰度值的差值。

下面给出能够滤除差分隐私加入的高斯噪声的高频噪声高斯滤波算法 HFNF, 如算法 1 所述。

算法 1: 高频噪声滤波算法 HFNF

输入: 加入差分隐私扰动的 $M\times N$ 图像 $\hat{X}$, 常数 c

输出: 过滤后的加扰图像 $\tilde{X}$

1: 初始化常数 c

2: 计算参数 ς

3: 计算高斯滤波器核 $\sigma_f\leftarrow\left[\arg\min_{\sigma}\frac{c}{\sigma^2}+\varsigma^2\right]$

4: FOR $i\leftarrow 0$ to M DO

5: FOR $j\leftarrow 0$ to N DO

6: 根据高斯滤波器核的尺寸对加扰像素滤

波, 得到过滤后的加扰图像 $\tilde{X}$

7: END FOR

8: END FOR

9: Return $\tilde{X}$

3.3 融合低频信息的差分隐私鲁棒性增强算法 LFDP

通过 3.2 节的分析可知, 在图像变换系数中加入高斯噪声, 会被高频滤波器滤除, 导致鲁棒性降低, 下面阐述本文提出的融合低频信息的差分隐私鲁棒性增强算法 LFDP。LFDP 包含两个部分, 分别是低频和高频信息部分, 高频噪声易于生成, 而低频高斯噪声较难生成。本文首先给出低频高斯噪声的生成方法, 如定理 4 所述。

定理 4. 对服从 $x_{i,j} \sim N(\mu,\sigma^2), X \in R^{d\times d}$, $\hat{\eta}$ 的高斯噪声 $\hat{\eta}$ 进行逆离散余弦变换:

$$\hat{\eta} = \begin{cases} x_{i,j} \sim N(\mu,\sigma^2), if 1 \leqslant i,j \leqslant rd \\ 0, otherwize \end{cases} \tag{5}$$

$$\eta_{i1,i2} = \sum_{j_1=0}^{d-1}\sum_{j_2=0}^{d-1} N_{j1}N_{j2}\hat{\eta}_{j_1,j_2}\varphi_d(i_1,j_1)\varphi_d(i_2,j_2) \tag{6}$$

得到的噪声 η 即为低频高斯噪声。

证明. 用 X 表示原始二维图像, 即 $X \in R^{d\times d}$。通过离散余弦变换可以将 X 变换到频域, 我们定义基础函数如下:

$$\varphi_d(i,j) = \cos[\frac{\pi}{d}(i+\frac{1}{2})j], 1 \leqslant i,j \leqslant d$$

在 $1 \leqslant i,j \leqslant d$ 范围内, 离散余弦变换 $V = DCT(X)$ 如下:

$$V_{j1,j2} = N_{j1}N_{j2}\sum_{i_1=0}^{d-1}\sum_{i_2=0}^{d-1} X_{i1,i2}\varphi_d(i_1,j_1)\varphi_d(i_2,j_2)$$

其中, $j=0$ 时, $N_j = \sqrt{\frac{1}{d}}$; 否则, $N_j = \sqrt{\frac{2}{d}}$, 这保证了 $\| X \|_2 = \| DCT(X) \|_2$。同时, 离散余弦变换也是可逆的, 即 $X = IDCT(V)$,

$$X_{i1,i2} = \sum_{j_1=0}^{d-1}\sum_{j_2=0}^{d-1} N_{j_1}N_{j_2}V_{j_1,j_2}\varphi_d(i_1,j_1)\varphi_d(i_2,j_2)$$

定理 4 即为低频高斯噪声的生成方法, 对于包含多个颜色通道的图像, DCT 和 IDCT 可以独立作用在每个通道上, 从而在每个通道上生成所需的低频高斯噪声。

LFDP 在处理图像数据 X 时, 先将其进行二维傅立叶变换, 将图像变换到频域 $Z(X)$, 然后寻找频谱的零点, 作为高低频的分界参数 k, 然后根据定理 4 生成低频的高斯噪声, 加入到 $Z(X)$ 的低频部分(分界参数 k 的左上角), 根据原始差分隐私机制生成高频高斯噪声, 加入到 $Z(X)$ 的高频部分(分界参数 k 的其余部分), 最后将加入噪声后的频域 $Z'(X)$ 进行二维傅立叶逆变换, 得到扰动的图像数据。

算法 2 是在图像变换系数中加入混合频率高斯噪声的算法流程。算法 2 描述了本文提出的完整的融合低频高斯噪声的差分隐私拉普拉斯噪声机制。通过在图像变换系数中分别加入低频和高频噪声, 即使是拥有图像频率等相关背景知识的攻击者依然无法通过滤波等手段缩小噪声的强度。

算法 2: 混合噪声生成算法 LFDP

输入: 尺寸为 $M \times N$ 的原始图像 X, 高低频分界参数 k, 高斯噪声的方差 σ

输出: 加扰图像 X'

1: $Z(X) \leftarrow$ 二维傅里叶变换 (X)

2: $n \leftarrow$ 生成尺度为 $M \times N$, 方差为 σ 的高斯噪声

3: FOR $i \leftarrow 0$ to M DO

4: FOR $j \leftarrow 0$ to N DO

5: IF i in range(0, k) and j in range(0, k) THEN

6: $n[i][j] \leftarrow 0$

7: END IF

8: END FOR

9: END FOR

10: 低频高斯噪声 $n_{low} \leftarrow$ 二维离散余弦逆变换 (n)

11: $Z'[X] \leftarrow Z[X] + n_{low}$

12: $X' \leftarrow$ 二维傅里叶逆变换 $Z'(X)$

3.4 性能分析

本节从理论上分析融合低频噪声的差分隐私机制的隐私保护性能、算法复杂度和对应的分类误差。

(1) 隐私性能分析

本文提出的混合噪声添加机制(LFDP)是为了克服原始差分隐私机制加入的高频高斯噪声在图像数据中容易被滤除, 从而造成隐私保护强度下降的缺陷, 根据图像的高低频组成成分的特点, 设计的一种由高频和低频两种高斯噪声组成的混合噪声添加机制, 从噪声尺度上来看, LFDP 并没有改变加入的噪声大小, 而只是改变了加入噪声的频率, 因此, LFDP 满足 (ε,δ)-DP。

(2) 算法复杂度分析

根据算法 2, 本文提出的方案的算法复杂度主要集中在傅里叶变换和余弦变换等步骤, 由于傅里叶变换和余弦变换都是一种线性变换, 因此并不会大幅增加算法的复杂度, 定理 5 给出了本文提出的方案的复杂度。

定理 5. 若原始图像尺寸是 $M \times N$, 则本文提出的算法复杂度是

$$O(M^2 * N^2) + 2O(M * N * \log(M * N)) + 3O(M * N) \approx O(M^2 * N^2)$$

证明. 从算法 1 可知, 本文提出的方案中步骤 1 和 12 分别是二维傅里叶变换和逆变换, 复杂度为 $2O(M * N * \log(M * N))$, 步骤 10 是二维离散余弦逆变换, 复杂度为 $O(M^2 * N^2)$, 步骤 3 和 4 是 *for* 循环, 复杂度是 $2O(M * N)$, 步骤 11 是在变换系数中加入噪声, 复杂度是 $O(M * N)$, 复杂度一共是 $O(M^2 * N^2) + 2O(M * N * \log(M * N)) + 3O(M * N)$.

从定理 5 可知, 相比于原始经过傅里叶变换后加入高频噪声的算法相比, 本文提出的方案增加的算法复杂度是线性复杂度, 并不会对整体的训练和预测时间噪声很大影响。

(3) 分类误差分析

定理 6. 若分类错误用 C_e 表示, 则 $C_e \leqslant 1 - e^{-}E_y[e^{-\varepsilon H(P(y))}]$, 其中表示图像的正确分类。

证明.

$$\begin{aligned}
C_e &= E_{(y,y')}[\sup_{\|\delta\| \leqslant R} E_{P(y')}[1(P(y_i^{'}) \neq y)] - E_{P(y)}[1(P(y_i) \neq y)]] \\
&= E_{(y,y')}[\sup_{\|\delta\| \leqslant R} E_{P(y'),P(y)}[1(P(y_i^{'}) \neq y) - 1(P(y_i) \neq y)]] \\
&\leqslant E_{(y,y')}[\sup_{\|\delta\| \leqslant R} E_{P(y'),P(y)}[1(P(y_i^{'}) \neq P(y_i))]] \\
&= E_{(y,y')}[\sup_{\|\delta\| \leqslant R} E_{P(y'),P(y)}[1(P(y_i^{'}) \neq P(y_i))]]
\end{aligned}$$

对于两个 IID 的变量和, 根据 Jensen 不等式, 我们有

$$P(L = Q) = \sum_{i=1}^{K} l_i q_i \geqslant e^{\sum_{i=1}^{K} l_i \log q_i} = e^{-d_{KL}(L,Q) - H(L)}$$

于是有

$$\begin{aligned}
&E_{(y,y')}[\sup_{\|\delta\| \leqslant R} P_{P(y'),P(y)}[1(P(y_i^{'}) \neq P(y_i))]] \\
&\leqslant E_{(y,y')}[\sup_{\|\delta\| \leqslant R} 1 - e^{-d_{KL}(P(y),P(y')) - H(P(y))}] \\
&\leqslant E_{(y,y')}[1 - e^{-H(P(y))}] \\
&= 1 - e^{-}E_y[e^{-H(P(y))}]
\end{aligned}$$

4 实验评估

本节评估了所提出的融合低频高斯噪声差分隐私鲁棒性增强方法 LFDP 的性能, 主要包括不同数据集上的精确度、鲁棒性半径和计算复杂度评估, 并与当前的代表性算法进行了对比分析。

4.1 数据集和配置

本文实验环境是 Intel(R) Core(TM) Xeon W2223 CPU @3.60GHz Windows 10 工作站, 32 GB 内存, RTX TITAN 24 GB 显卡, 本文在真实的数据集中对提出的算法进行测试, 包括 MNIST 和 CIFAR-10 图像数据集。

为了测试本文提出的方法的有效性, 我们从以下 3 个方面进行性能评估:

(1) 精确度评估。由于差分隐私是一种噪声增强机制, 本文提出的 LFDP 方案将低频和高频噪声融合, 在满足差分隐私的同时增强鲁棒性。因此, 需要测试所提出的方法对模型精确度的影响, 以确保提出的模型不会对模型造成很大的影响。

(2) 鲁棒性能评估。模型鲁棒性测试主要评估在经典攻击模型下, 分别对比原始图像、原始差分隐私噪声机制以及融合低频信息的噪声机制的鲁棒性能。

(3) 计算复杂度评估。本部分主要测试提出的方案的算法复杂度, 主要包括训练和预测所用的时间, 评估提出的方案对训练和预测时间造成的影响。

实验的测试算法包括原始差分隐私噪声机制、GanobFouscator 和 PixelDP, 对比其在有滤波和无滤波状态下的性能。其中 PixelDP 的结构中包括一个 DP 噪声层, 它使网络的计算随机化, 以强制执行 DP 界限, 即使得预测结果的分布在小的规范约束下变化。在推理时, PixelDP 利用 DP 界限来实现对单个预测结果的鲁棒性检查。对于一个给定的输入, 通过鲁棒性检查可以保证其在某一特定大小的扰动下, 不存在噪声导致网络改变其预测结果的情况。按照文献[18]的建议, PixelDP 的噪声尺度计算采用 1-范数, 噪声采用高斯噪声, 在神经网络的第 1 层加入扰动的噪声, 敏感度函数设置为 1。

4.2 精确度评估

在精确度评估方面, 我们利用经典的 FGSM 和 PGD 对抗样本攻击方法进行攻击, 分别评估加入高频噪声后的原始图片、加入高频噪声滤波后的图片以及融合低频信息的噪声扰动图片三种情形下的预测精确度。实验过程主要包括如下三个过程:

(1) 将参数不同的高频噪声加入到数据集中进行差分隐私保护, 以保护原始数据集。在加入高斯噪

声时, 首先对原始图像进行快速傅里叶变换, 并且将零频点放到频谱的中间以便观察傅立叶变换并加入噪声。

以MNIST数据集[31]为例, 将零频点放到中间后, 左上角的区域为图像的高频部分, 我们对左上角10×10 频域的实部加入高斯噪声。这样做的原因是, 将零频点放到频谱的中间后, 低频部分集中在左上角, 而在频域中, 低频部分的实部较高频部分更大, 对低频部分加入高斯噪声, 对图像的改变是微小的, 同时又能进行差分隐私保护。之后, 我们通过快速傅里叶逆变换即可得到带高频噪声的数据集, 图 1 展示了原始图像和加入差分隐私高斯噪声后的图像。

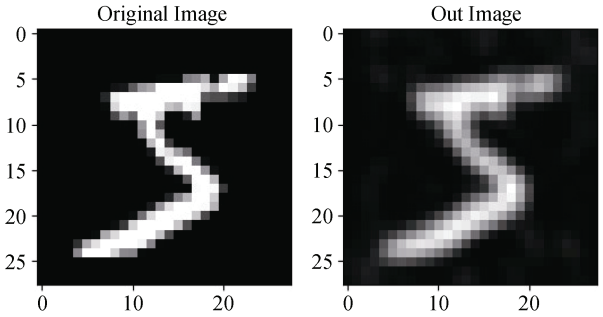


图 1　原始图像和加入高频高斯噪声后的对比图

Figure 1　Comparison between the original image and the high frequency Gaussian noise

(2) 利用高斯滤波对数据集发起攻击。在进行高频噪声攻击时, 恶意攻击者可以用高斯滤波器来滤除加入的差分隐私噪声。

(3) 在融合低频噪声时, 将由高斯噪声经过逆离散余弦变换而来的低频噪声与高频噪声混合, 加入到数据集中, 同样, 我们需要对噪声进行剪裁(=10), 再将生成的低频高斯噪声加入到离散余弦变换后的系数中, 对其进行离散傅立叶逆变换, 得到加入低频和高频混合噪声的数据集。

在实验中, 噪声的裁剪参数是根据数据集的高低频成分特征进行设置的。在本文中, 我们以MNIST 数据集为例来说明裁剪参数的设置方法, 首先对选取的数字为 5 的图片进行傅立叶变换, 然后寻找频域中零频谱的位置, 发现以 10×10 的频域为界, 左上角是低频部分, 我们在低频部分加入低频高斯噪声, 而其余频域部分为高频部分, 则加入差分隐私生成的高频高斯噪声。

对于 MNIST 数据集, 按照上述方法加入方差不同的高频噪声和高低频混合噪声后, 利用卷积神经网络进行训练。我们选取的噪声方差范围在 0～20 之间, batch-size 选取 100, 利用交叉熵损失函数进行网络的反向传播, 通过网络的 10 轮迭代, 得到测试准确率如图 2 所示。对于 CIFAR-10 数据集, 按照上述方法加入方差不同的高频噪声和高低频混合噪声后, 利用残差神经网络 ResNet18 进行训练。由于 CIFAR-10 数据集为三通道的图片, 因此对每张图片的 3 个通道都要进行加噪保护处理。加入噪声后, 用周期性学习率技术设置学习率, batch-size 选取 128, 利用交叉熵损失函数进行网络的反向传播, 通过 15 轮迭代, 得到测试准确率如图 3 所示。

从图 2 和图 3 可以看出, 在训练开始噪声较小时, 滤波与未滤波的训练结果相差较小, 但随着噪声逐渐加大, 滤波后的准确率渐渐高于未滤波的准确率。例如, 在 MNIST 原始数据集中, 噪声方差等于 20.0, 采用 FGSM 对抗样本攻击时, 滤波后的预测精确度是 0.1014, 未滤波的预测精确度是 0.2739, 降低 63.0%, 在其他对比算法 GanobFouscator 和 PixelDP 上也表现出同样的趋势。同样的规律在 CIFAR10 数据集的实验结果上也能够发现。这说明 HFNF 在一定程度上可以滤除差分隐私在图像中加入的高频噪声, 使得保护效果降低, 鲁棒性能也随之下降。

另外, 我们注意到, 在噪声方差较小时, 数据集经过自适应滤波后, 训练准确率仍然低于原始图像。这是因为, 在加入高频噪声过后, 对图像进行滤波时, 高斯滤波无法区分出原始图像中的高频噪声与后来加入的高频噪声, 于是只要遇到高频噪声, 就会将其进行过滤, 丢失了原始图像中带有的一部分高频信息, 因此准确率比原始图像低。由图 2~3 可见, 滤波后的准确率在很大范围内仍然略低于未滤波准确率, 说明高斯滤波并不能将加入的低频噪声完全滤去。

不同于无攻击方式, 随着噪声方差的增大, 模型的训练准确率不断提升, 例如在 σ^2 =0.1 时, 在无滤波情况下, MNIST 数据集上 FGSM 攻击时的预测准确率是 14.71%, 而当 σ^2 =15.0 时, FGSM 攻击的准确率是 29.18%, 在 PGD 以及 CIFAR10 数据集上也能发现同样的规律。这是因为, 随着加入的噪声尺度增加, 训练的模型得到了更好的泛化, 从而对于 FGSM 以及 PGD 等对抗样本攻击方式不敏感, 即模型的鲁棒性得到了增强。

对比图 2～3 可以发现, 无论在有滤波还是无滤波情况下, 混合噪声都要比高频噪声的预测精确度高。例如, 在 MNIST 数据集上采用 FGSM 攻击时, 当 σ^2 =20 时, 混合噪声的 LFDP 的预测精确度是 68.34%, 而现有最优的 PixelDP 的精确度是 56.47%, 提高了 17.4%; 而高频噪声情况下 LFDP 和 PixelDP 的预测精确度分别是 57.73%和 46.96%, 提高了

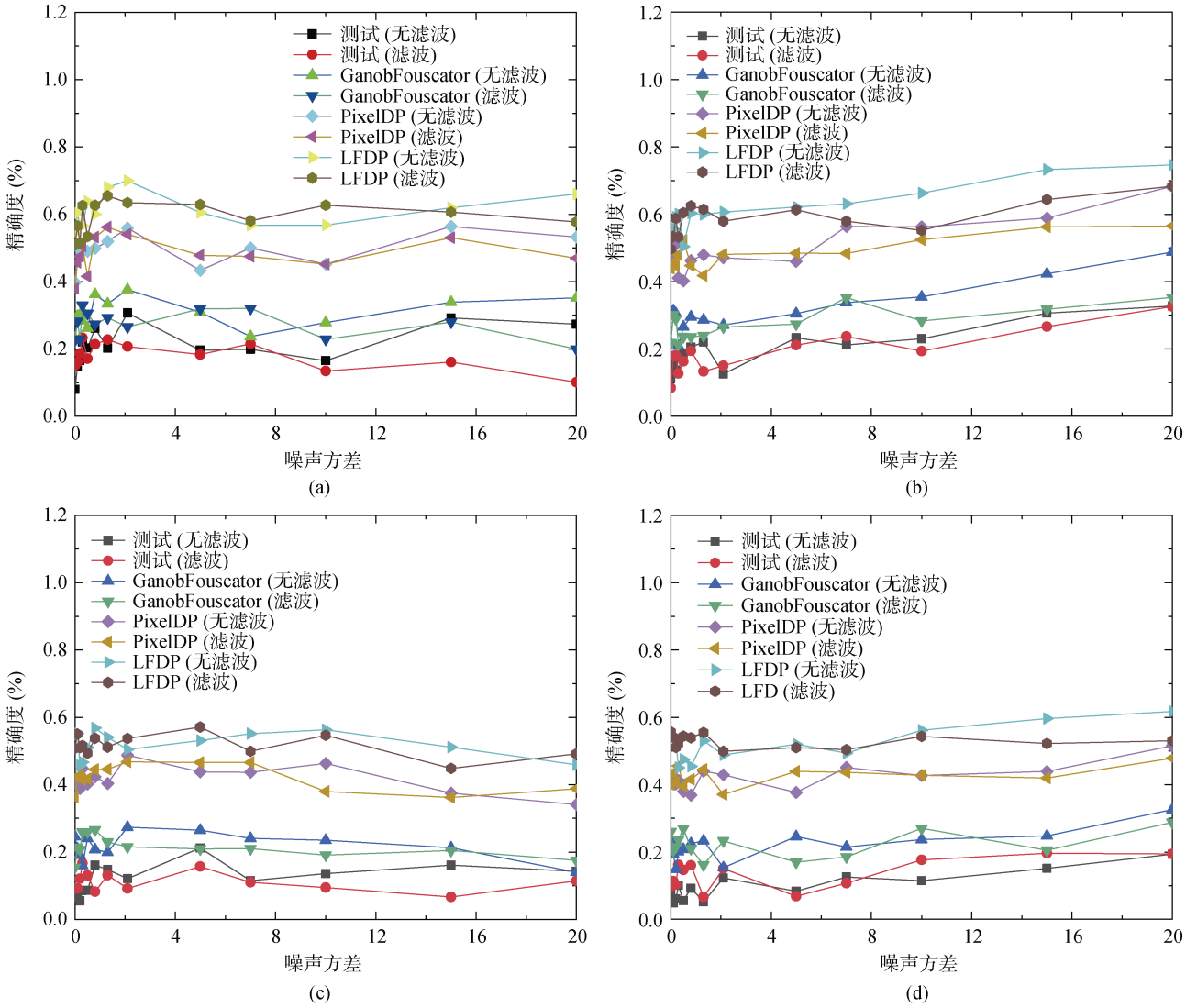


图 2　MNIST 数据集加入不同噪声后的精确度对比图: (a)高频噪声在 FGSM 攻击下的精确度 (b)混合噪声在 FGSM 攻击下的精确度 (c)高频噪声在 PGD 攻击下的精确度 (d)混合噪声在 PGD 攻击下的精确度

Figure 2　Comparison of accuracy of MNIST data set with different noises: (a) Accuracy of high frequency noise under FGSM attack (b) Accuracy of mixed noise under FGSM attack (c) Accuracy of high frequency noise under PGD attack (d) accuracy of mixed noise under PGD attack

18.7%; 在 PGD 攻击条件下, 混合噪声的精确度是 53.01%, 现有最优的 PixelDP 的精确度 47.90%, 提高了 9.6%, 同样的趋势在 CIFAR10 数据集中也能发现。因此, LFDP 的准确率总体上比只加入高频滤波的准确率要高, 即使是在有滤波的情况下也是如此, 说明混合噪声的鲁棒性效果更好, 精确度更高, 这与理论上的预期结果相符。

4.3　鲁棒性评估

此节进行攻击测试, 分别用两种经典的对抗样本攻击方法 FGSM、PGD 对提出的方案进行测试, 以比较不同情形下的模型鲁棒性。

模型鲁棒性半径是指模型能够容忍的对抗样本尺度, 在容忍半径范围内对模型的精确度没有影响。从图 4 可以看出, 模型的鲁棒性半径随着噪声尺度的增加不断增大。例如, 在 MNIST 数据集上采用 FGSM 攻击时, 在无滤波的情况下, 当噪声尺度为 0.1 时, 鲁棒性半径为 0.2051, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.321; 在有滤波的情况下, 当噪声尺度为 0.1 时, 鲁棒性半径为 0.2041, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.2969; 在 MNIST 数据集上采用 PGD 攻击时, 在无滤波的情况下, 当噪声尺度为 0.1 时, 鲁棒性半径为 0.0003, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.116; 在有滤波的情况下, 当噪声尺度为 0.1 时, 鲁棒性半径为 0, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.09。这说明滤波确实能够减少模型的鲁棒性半径, 即降低模型的鲁棒性。

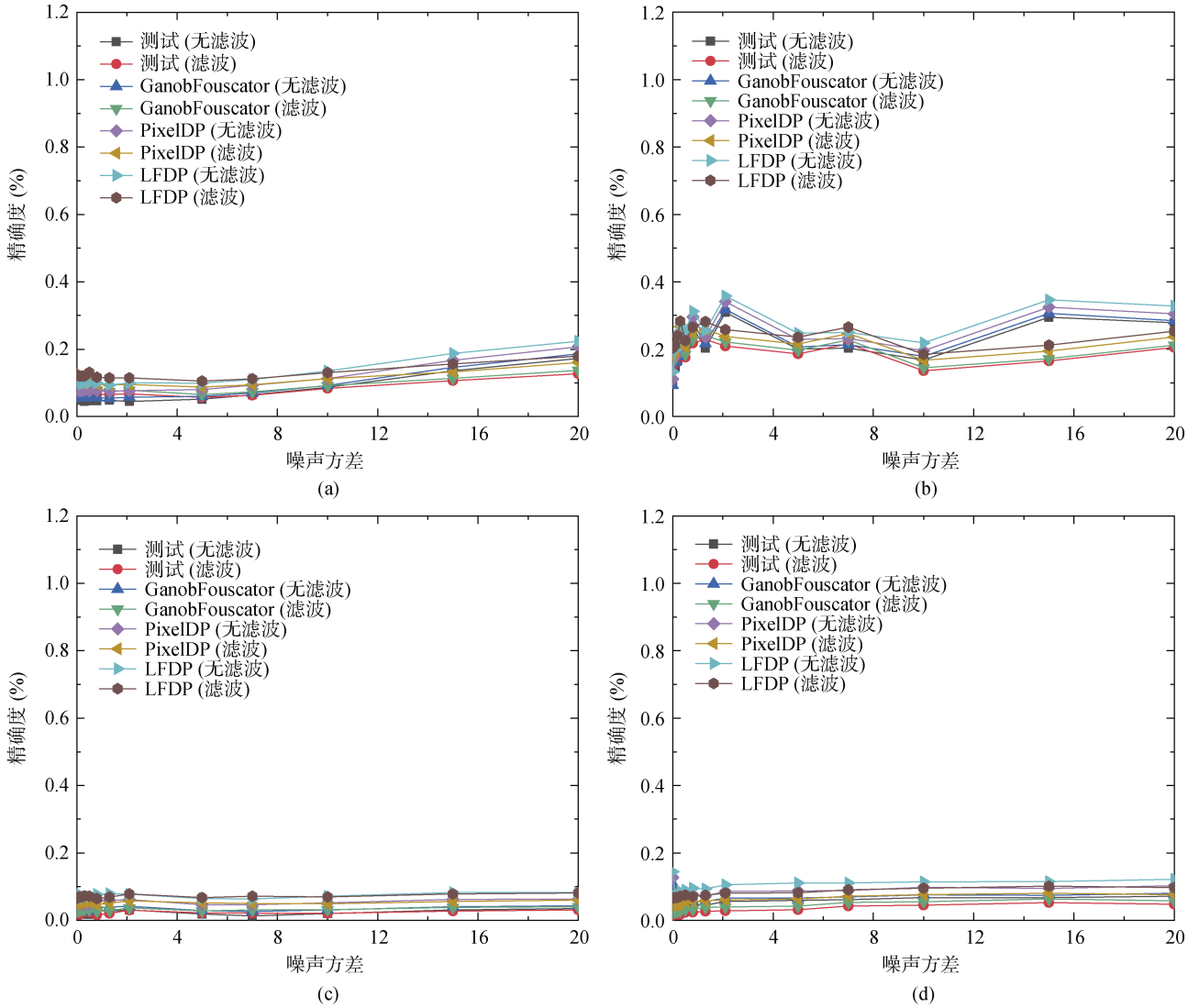
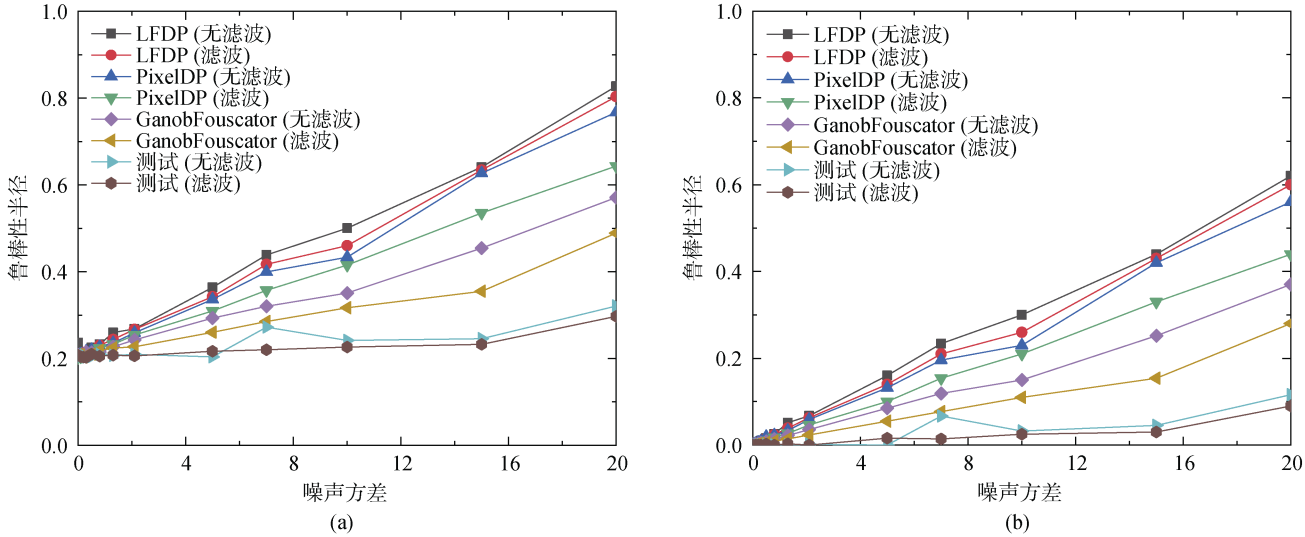


图 3　CIFAR-10 数据集加入不同噪声后的精确度对比图: (a)高频噪声在 FGSM 攻击下的精确度 (b)混合噪声在 FGSM 攻击下的精确度 (c)高频噪声在 PGD 攻击下的精确度 (d)混合噪声在 PGD 攻击下的精确度

Figure 3　Accuracy comparison of CIFAR-10 data set with different noises: (a) Accuracy of high frequency noise under FGSM attack (b) Accuracy of mixed noise under FGSM attack (c) Accuracy of high frequency noise under PGD attack (d) Accuracy of mixed noise under PGD attack



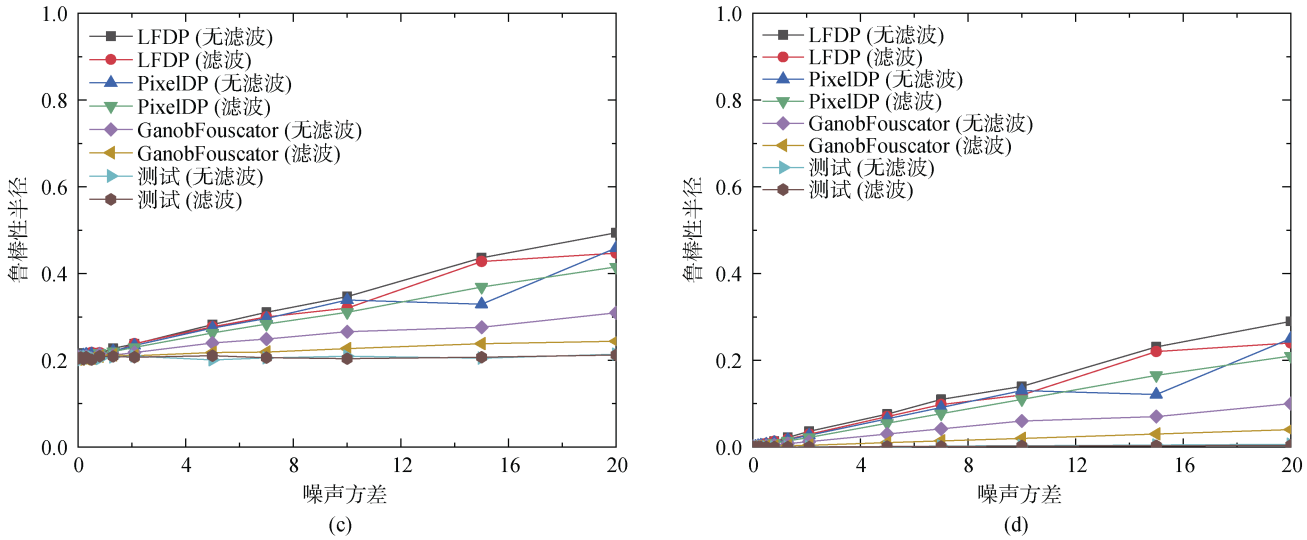


图 4　不同数据集加入不同噪声后的鲁棒性半径对比图: (a)FGSM 攻击时 MNIST 数据集上的鲁棒性半径 (b)PGD 攻击时 MNIST 数据集上的鲁棒性半径 (c)FGSM 攻击时 CIFAR-10 数据集上的鲁棒性半径 (d)PGD 攻击时 CIFAR-10 数据集上的鲁棒性半径

Figure 4　Comparison of robustness radii of different data sets with different noises: (a) Robustness radius on MNIST data set in case of FGSM attack (b) Robustness radius on MNIST data set in case of PGD attack (c) Robustness radius on CIFAR-10 data set in case of FGSM attack (d) Robustness radius on CIFAR-10 data set in case of PGD attack

类似地, 在 CIFAR10 数据集上采用 FGSM 攻击时, 在无滤波的情况下, 当噪声尺度为 0.1 时, 鲁棒性半径为 0.2008, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.2137; 在有滤波的情况下, 当噪声尺度为 0.1 时, 鲁棒性半径为 0.2068, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.2121; 在 CIFAR10 数据集上采用 PGD 攻击时, 在无滤波的情况下, 当噪声尺度为 0.1 时, 鲁棒性半径为 0.0003, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.006; 在有滤波的情况下, 当噪声尺度为 0.1 时, 鲁棒性半径为 0, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.004。在其他测试条件下以及其他算法上也能发现同样的规律, 这说明随着噪声尺度的增加, 模型的鲁棒性不断增强。

对比 LFDP 和目前最优的 PixelDP 算法, 在 PGD 攻击时 MNIST 数据集上无滤波的情况下, 当噪声尺度为 0.1 时, PixedlDP 的鲁棒性半径为 0.0028, 而 FLDP 的鲁棒性半径为 0.0032, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.56, 而 FLDP 的鲁棒性半径为 0.62, 鲁棒性半径增加了 10.7%; 在有滤波的情况下, 当噪声尺度为 0.1 时, PixedlDP 的鲁棒性半径为 0.0022, 而 FLDP 的鲁棒性半径为 0.003, 而当噪声尺度为 20.0 时, 鲁棒性半径为 0.44, 而 FLDP 的鲁棒性半径为 0.6, 鲁棒性半径增加了 36.4%。在 CIFAR10 数据集上以及 FGSM 攻击的情况下也能发现同样的趋势, 这表明 LFDP 的鲁棒性优于现有的最优方法 PixelDP。

需要注意的是, 结合图 2~5 可以看出, 虽然增加噪声尺度能够增强模型的鲁棒性, 但模型的训练精度也会下降, 这是由于加入噪声后会淹没一部分原始图像的像素, 导致训练精度下降, 在设计相应的噪声机制时, 需要同时关注噪声尺度对训练精度的影响。

另外, 我们注意到, 当噪声尺度为 0 时(即没有加入噪声时), MNIST 数据集和 CIFAR10 数据集的鲁棒性半径相同, 说明本文采用的鲁棒性半径指标与数据集无关, 是个独立指标, 因而能够客观的评估噪声对模型鲁棒性的影响。

4.4　计算复杂度评估

虽然 3.4 节从理论上分析了 LFDP 的计算复杂度, 证明本文提出的噪声生成方法是线性操作, 不会对计算复杂度造成很大影响。本节在实际数据集上评估 LFDP 的计算复杂度, 主要是训练时间的性能, 实验结果和对比图如图 5 所示。

从图 5 可以看出, 本文提出的算法对训练时间的影响较小。例如在 MNIST 数据集中, 当采用 FGSM 攻击时, 在无滤波的训练情形下, 训练时间为 6.7 s, 而 LFDP 的训练时间是 16.4 s, 最差的情形是在有滤波的测试情况下, 有滤波的测试时间是 11.3 s, 而 LFDP 的测试时间是 36.8 s, 但由于数值上只增加了 25.5 s 可以认为增加的测试时间在用户忍受范围

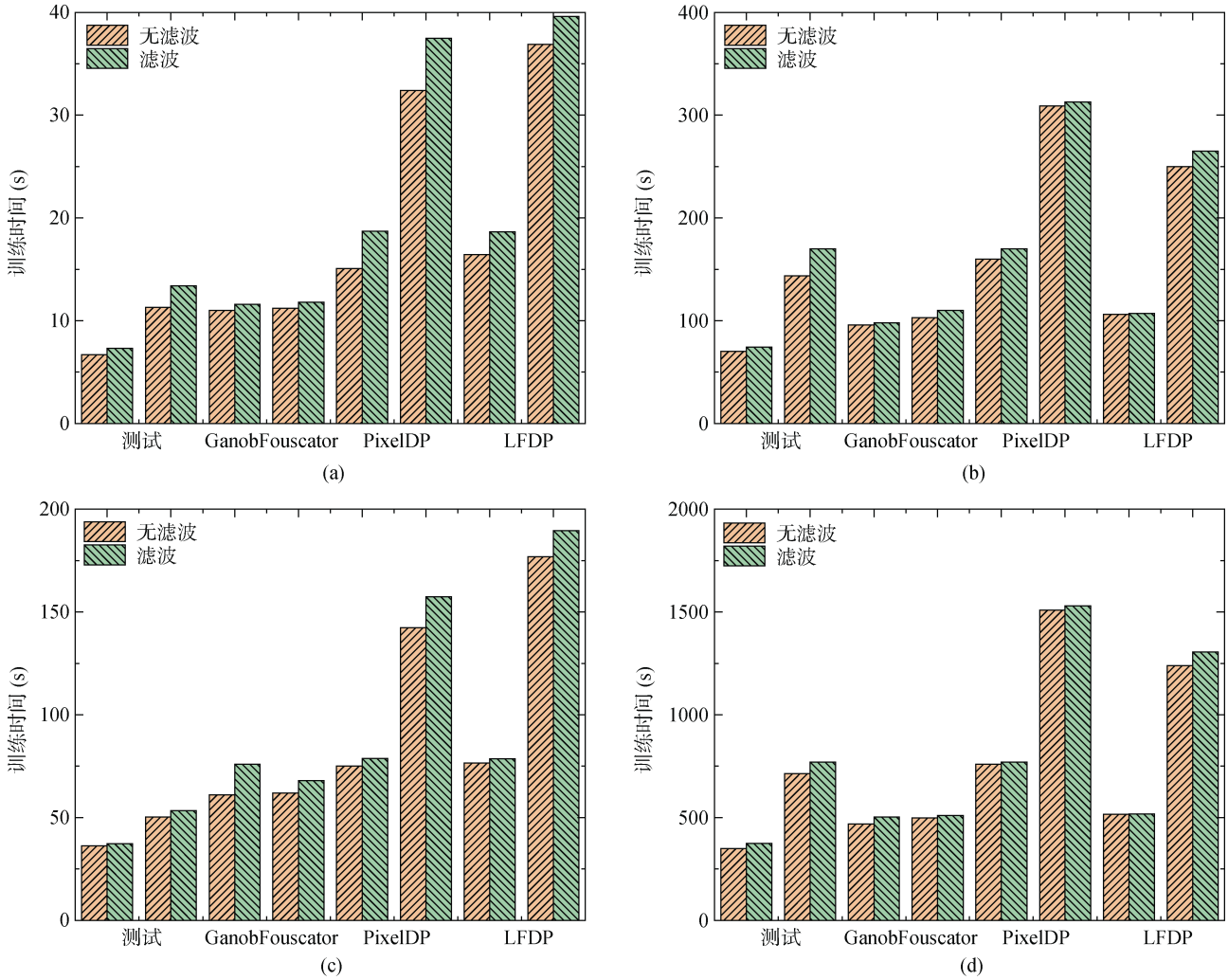


图 5　加入混合噪声后不同数据集在不同攻击方式下的训练时间对比图: (a)MNIST 数据集在 FGSM 攻击下的训练时间 (b) MNIST 数据集在 PGD 攻击下的训练时间 (c) CIFAR-10 数据集在 FGSM 攻击下的训练时间 (d) CIFAR-10 数据集在 PGD 攻击下的训练时间

Figure 5　Comparison of training time of different data sets under different attack modes after adding mixed noise: (a) Training time of MNIST data set under FGSM attack (b) Training time of MNIST data set under PGD attack (c) Training time of CIFAR-10 data set under FGSM attack (d) Training time of CIFAR-10 data set under PGD attack

内。如果从最耗时的 PGD 攻击来看, 与目前最优的 PixelDP 方案相比, 加入混合噪声的训练和测试总耗时为 313 s, 而本文方案的耗时为 265 s, LFDP 具有更少的训练和测试时间。在 CIFAR10 数据集上能够发现同样的趋势, 表明本文提出的 LFDP 虽然相比无噪声的情况增加了训练和预测时间, 但增加的复杂度在可忍受的范围内, 且比目前最优的方案具有很少的计算复杂度。

4.5　实验结论

通过在不同数据集上对精确度、鲁棒性以及计算复杂度进行评估, 从实验结果来看, 我们可以得到如下结论:

(1) 相比于原始差分隐私机制中加入的高频高斯噪声, 本文提出的 LFDP 方案能够作为一种通用方案, 仅仅通过改变噪声的形式, 就能从整体上提高现有方法的鲁棒性表现, 证明了 LFDP 的有效性;

(2) 更大尺度的噪声意味着更大的鲁棒半径, 但更大尺度的噪声可能会淹没原始图像数据, 导致训练和预测精度下降, 需要在这两者之间达到一个平衡;

(3) 基于差分隐私的扰动方式计算复杂度较低, 在增强模型鲁棒性能的同时, 对算法复杂度的影响很小, 基于噪声扰动的鲁棒性增强方法在实际应用中实用性较强。

5　结论

为了解决现有差分隐私机器学习模型鲁棒性增

强方法面临的训练和预测精度较低、噪声容易被滤除的问题, 本文提出了一种普适的融合低频信息的差分隐私鲁棒性增强方法 LFDP, 通过分离原始图片的频域, 在高低频部分分别加入生成的高低频差分隐私高斯噪声。实验结果表明, 与现有方法相比, 对于图像数据来说, 本文方法在保证差分隐私的前提下, 具有更高的训练和预测精度以及更大的鲁棒性半径, 且计算复杂度没有明显增加。同时也能够作为一种通用方案, 以改进现有的差分隐私噪声生成形式。

由于图像数据高低频的界限分割以及加入的噪声尺度会对模型鲁棒性以及训练和预测精度造成直接影响, 接下来的工作中将从理论上研究高低频的界限以及噪声尺度和模型鲁棒性之间的关系; 同时, 也会尝试寻求更加复杂场景下结合其他鲁棒性增强方法的差分隐私鲁棒性增强方法, 以期达到复杂场景下更加实用的鲁棒性增强效果。

参考文献

[1] Zhang S S, Zuo X, Liu J W. The Problem of the Adversarial Examples in Deep Learning[J]. *Chinese Journal of Computers*, 2019, 42(8): 1886-1904.
(张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. *计算机学报*, 2019, 42(8): 1886-1904.)

[2] Pan W W, Wang X Y, Song M L, et al. Survey on Generating Adversarial Examples[J]. *Journal of Software*, 2020, 31(1): 67-81.
(潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述[J]. *软件学报*, 2020, 31(1): 67-81.)

[3] 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.)

[4] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[J]. *ArXiv e-Prints*, 2014: arXiv: 1412.6572.

[5] ZHAO Z, LIU Z, LARSON M. On success and simplicity: A second look at transferable targeted attacks[C]. *35th Conference on Neural Information Processing Systems,* 2021, 34.

[6] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[EB/OL]. 2017: 1706.06083. https://arxiv.org/abs/1706.06083v4.

[7] Amiri M M, Gündüz D. Machine Learning at the Wireless Edge: Distributed Stochastic Gradient Descent Over-the-Air[C]. *2019 IEEE International Symposium on Information Theory*, 2019: 1432-1436.

[8] Dwork C. Differential Privacy[M]. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 1-12.

[9] Dwork C. A Firm Foundation for Private Data Analysis[J]. *Communications of the ACM*, 2011, 54(1): 86-95.

[10] Wang H, Wang H. Correlated Tuple Data Release *via* Differential Privacy[J]. *Information Sciences*, 2021, 560: 347-369.

[11] Lecuyer M, Atlidakis V, Geambasu R, et al. Certified Robustness to Adversarial Examples with Differential Privacy[C]. *2019 IEEE Symposium on Security and Privacy*, 2019: 656-672.

[12] Phan H, Thai M T, Hu H, et al. Scalable differential privacy with certified robustness in adversarial learning[C]. *International Conference on Machine Learning*, 2020: 7683-7694.

[13] Pinot R, Meunier L, Araujo A, et al. Theoretical evidence for adversarial robustness through randomization[C]. *33rd Conference on Neural Information Processing Systems*, 2019: 32.

[14] Araujo A, Meunier L, Pinot R, et al. Robust Neural Networks Using Randomized Adversarial Training[EB/OL]. 2019: 1903.10219. https://arxiv.org/abs/1903.10219v3.

[15] LI B, CHEN C, WANG W, et al. Certified adversarial robustness with additive noise[C]. *International Conference on Neural Information Processing Systems,* 2019: 1-11.

[16] Cohen J, Rosenfeld E, Kolter Z. Certified adversarial robustness via randomized smoothing[C]. *International Conference on Machine Learning*, 2019: 1310-1320.

[17] Lecuyer M, Atlidakis V, Geambasu R, et al. On the connection between differential privacy and adversarial robustness in machine learning[J]. *stat,* 2018, 1050: 9.

[18] Phan N, Vu M N, Liu Y, et al. Heterogeneous Gaussian Mechanism: Preserving Differential Privacy in Deep Learning with Provable Robustness[C]. *The Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019: 4753-4759.

[19] Xu C G, Ren J, Zhang D Y, et al. GANobfuscator: Mitigating Information Leakage under GAN *via* Differential Privacy[J]. *IEEE Transactions on Information Forensics and Security*, 2019, 14(9): 2358-2371.

[20] Pinot R, Yger F, Gouy-pailler C, et al. A unified view on differential privacy and robustness to adversarial examples[C]. *Workshop on Machine Learning for Cyber Security at ECMLPKDD,* 2019.

[21] Pinot R, Ettedgui R, Rizk G, et al. Randomization Matters. how to Defend Against Strong Adversarial Attacks[EB/OL]. 2020: 2002.11565. https://arxiv.org/abs/2002.11565v5.

[22] Yin D, Lopes R G, Shlens J, et al. A Fourier Perspective on Model Robustness in Computer Vision[EB/OL]. 2019: 1906.08988. https://arxiv.org/abs/1906.08988v3.

[23] 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.

[24] Guo C, Frank J S, Weinberger K Q. Low frequency adversarial perturbation[C]. *Uncertainty in Artificial Intelligence, PMLR*, 2020: 1127-1137.

[25] Xiong P, Zhu T Q, Wang X F. A Survey on Differential Privacy and Applications[J]. *Chinese Journal of Computers*, 2014, 37(1): 101-122.
(熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. *计算机学报*, 2014, 37(1): 101-122.)

[26] Zhang X J, Meng X F. Differential Privacy in Data Publication and Analysis[J]. *Chinese Journal of Computers*, 2014, 37(4): 927-949.
(张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. *计*

算机学报, 2014, 37(4): 927-949.)

[27] Zafarani F, Clifton C. Differentially Private Naïve Bayes Classifier Using Smooth Sensitivity[J]. *Proceedings on Privacy Enhancing Technologies*, 2021, 2021(4): 406-419.

[28] Zhang X J, Fu C C, Meng X F. Facial Image Publication with Differential Privacy[J]. *Journal of Image and Graphics*, 2018, 23(9): 1305-1315.
(张啸剑, 付聪聪, 孟小峰. 面向人脸图像发布的差分隐私保护[J]. *中国图象图形学报*, 2018, 23(9): 1305-1315.)

[29] Zhang X J, Fu C C, Meng X F. Private Facial Image Publication through Matrix Decomposition[J]. *Journal of Image and Graphics*, 2020, 25(4): 655-668.
(张啸剑, 付聪聪, 孟小峰. 结合矩阵分解与差分隐私的人脸图像发布[J]. *中国图象图形学报*, 2020, 25(4): 655-668.)

[30] Deng G, Cahill L W. An Adaptive Gaussian Filter for Noise Reduction and Edge Detection[C]. *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, 2002: 1615-1619.

[31] Deng L. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web[J]. *IEEE Signal Processing Magazine*, 2012, 29(6): 141-142.

王豪 于 2018 年在武汉大学通信与信息系统专业获得博士学位。现任重庆邮电大学副教授。研究领域为网络空间安全。研究兴趣包括: 隐私保护、对抗机器学习。Email: haowang@cqupt.edu.cn

许强 于 2021 年在上海交通大学计算机科学与技术专业获得博士学位。现任上海交通大学单位助理研究员。研究领域为网络空间安全。研究兴趣包括: 视频安全、对抗生成网络。Email: xuqiangwhu@sjtu.edu.cn

张清华 于 2009 年在西南交通大学计算机科学与技术专业获得博士学位。现任重庆邮电大学教授。研究领域为人工智能。研究兴趣包括: 智能计算、知识发现。Email: zhangqh@cqupt.edu.cn

李开菊 于 2023 年在重庆大学计算机科学与技术专业获得博士学位。现任贵州财经大学讲师。研究领域为高性能计算。研究兴趣包括: 联邦学习、通信优化。Email: likaiju@mail.gufe.edu.cn