计算机科学
COMPUTER SCIENCE

基于近端线性组合的信号识别神经网络黑盒对抗攻击方法

郭宇琦, 李东阳, 闫镔, 王林元

引用本文

郭宇琦, 李东阳, 闫镔, 王林元. 基于近端线性组合的信号识别神经网络黑盒对抗攻击方法[J]. 计算机科学, 2024, 51(10): 425-431.

GUO Yuqi, LI Dongyang, YAN Bin, WANG Linyuan. Black-box Adversarial Attack Methods on Modulation Recognition Neural Networks Based onSignal Proximal Linear Combination [J]. Computer Science, 2024, 51(10): 425-431.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900054

基于近端线性组合的信号识别神经网络黑盒对抗攻击方法

郭宇琦　李东阳　闫　镔　王林元
战略支援部队信息工程大学成像与智能处理实验室　郑州 450001
(guoyuqi728@foxmail.com)

摘　要　随着深度学习在无线通信领域特别是信号调制识别方向的广泛应用,神经网络易受对抗样本攻击的问题同样影响着无线通信的安全。针对无线信号在通信中难以实时获得神经网络反馈且只能访问识别结果的黑盒攻击场景,提出了一种基于近端线性组合的黑盒查询对抗攻击方法。该方法首先在数据集的一个子集上对每个原始信号样本进行近端线性组合,即在非常靠近原始信号的范围内与目标信号进行线性组合(加权系数不大于0.05),并将其输入待攻击网络以查询识别结果。通过统计网络对全部近端线性组合识别出错的数量,确定每类原始信号最容易受到线性组合影响的特定目标信号,将其称为最佳扰动信号。在攻击测试时,根据信号的类别选择对应最佳扰动信号执行近端线性组合,生成对抗样本。实验结果显示,该方法在选定子集上将每种调制类别的最佳扰动信号添加在全部数据集上能将神经网络识别准确率从94%降至50%,且相较于添加随机噪声攻击的扰动功率更小。此外,生成的对抗样本对于结构近似的神经网络具有一定迁移性。这种方法在统计查询后生成新的对抗样本时,易于实现且无需再进行黑盒查询。

关键词:深度学习;对抗样本;信号识别;黑盒攻击;对抗信号
中图分类号　TP391

Black-box Adversarial Attack Methods on Modulation Recognition Neural Networks Based on Signal Proximal Linear Combination

GUO Yuqi, LI Dongyang, YAN Bin and WANG Linyuan
Laboratory of Imaging and Intelligent Processing, PLA Strategy Support Force Information Engineering University, Zhengzhou 450001, China

Abstract　With the extensive application of deep learning in the field of wireless communication, especially in signal modulation recognition, the vulnerability of neural networks to adversarial example attacks poses challenges to the security of wireless communication. Addressing the black-box attack scenario in wireless signals, where real-time feedback from the neural network is hard to obtain and only recognition results can be accessed, a black-box query adversarial attack method based on proximal linear combination is proposed. Initially, on a subset of the dataset, each original signal undergoes a proximal linear combination with target signals, where they are linearly combined within a range very close to the original signal (with weighting coefficients no greater than 0.05) and then input into the neural network to query. By counting the number of misrecognitions by the network for all proximal linear combinations, specific target signals most susceptible to linear combination effects for each original signal category are determined, which is termed the optimal perturbation signals. During attack testing, adversarial examples are generated by executing proximal linear combinations using the optimal perturbation signal corresponding to the signal category. Experimental results demonstrate that using the optimal perturbation signal for each modulation category on the chosen subset, the recognition accuracy of the neural network dropped from 94% to 50% when applied to the entire dataset, with a lower perturbation power compared to adding random noise attacks. Furthermore, the generated adversarial examples exhibit some transferability to structurally similar neural networks. This method, which generates new adversarial examples after statistical queries, is easy to implement and eliminates the need for further black-box queries.

Keywords　Deep learning, Adversarial examples, Signal recognition, Black-box attack, Adversarial signal

到稿日期:2023-09-11　返修日期:2024-03-05
基金项目:国家自然科学基金(62271504)
This work was supported by the National Natural Science Foundation of China(62271504).
通信作者:王林元(wanglinyuanwly@163.com)

1 引言

深度学习在无线通信领域得到广泛应用，包括信号识别[1]、频谱感知、无线资源管理等方面。然而深度学习模型普遍存在易受对抗样本攻击的问题[2]，即通过对输入数据进行微小的修改，攻击者可以使模型产生错误的输出结果。对抗样本的存在同样给无线通信安全带来隐患。例如，对抗样本可以用于攻击基于神经网络的通信系统，攻击者对无线信号添加扰动可以使接收端产生错误的识别结果，进而导致后续通信失败甚至信息泄露[3]。此外，对抗样本也可以用于防御，通过扰动信号，欺骗非合作接收端，提高通信安全性[4]。

Sadeghi 等[5]首次将对抗样本应用于同相/正交(In-phase/Quadrature，I/Q)信号调制识别的卷积神经网络，研究发现，即使给信号添加很小的对抗扰动也能显著降低分类器的性能。对抗攻击相较于传统的噪声干扰攻击效果更为显著，且所需的扰动功率更小。这些微小的修改往往难以察觉，例如功率足够小或者波形上无明显失真，但这些修改却足以欺骗神经网络使其误判。

根据攻击者所掌握的被攻击神经网络模型的信息，对抗样本生成方法可以划分为白盒攻击和黑盒攻击[6]。在白盒攻击中，攻击者了解神经网络的结构和参数，可通过梯度信息生成对抗样本。许多图像上的对抗样本生成算法被广泛应用于I/Q信号的对抗，例如基于快速梯度符号方法(Fast Gradient Sign Method，FGSM)的攻击[5]，基于基本迭代方法(Basic Iterative Method，BIM)、动量迭代方法(Momentum Iterative Method，MIM)和投影梯度下降(Projected Gradient Descent，PGD)的攻击[7]等。这些方法根据调制识别网络的梯度计算扰动方向。在黑盒攻击中，攻击者仅能访问神经网络的输出，而无法获得网络内部结构、参数等详细信息。黑盒攻击的常见方法包括基于迁移和基于查询这两类。近年来，advGAN(Adversarial Generative Adversarial Networks)作为一种新兴的对抗样本生成方法，通过生成对抗网络(GAN)来生成对抗样本，也被应用于黑盒迁移对抗攻击[8]。信号识别的黑盒迁移攻击利用目标模型的基本信息训练替代模型，再进行白盒攻击[9-10]。信号识别的黑盒查询攻击则主要依赖查询被攻击模型的损失、置信度或识别结果来估计梯度[11-12]或者进行整体优化[13]。这些黑盒方法在生成对抗样本时，需要频繁查询目标模型的输出，这与典型的计算机视觉和自然语言处理应用通过互联网应用程序接口(Application Programming Interface，API)进行查询不同。在无线通信场景中，攻击者很难频繁查询分类器获得其分类结果[14]。这种查询受限的情况，也降低了其他领域基于查询的方法直接在生成信号对抗样本上应用的可行性。

考虑到信号对抗攻击应用的实际情况，攻击者会面临更多的限制。首先，他们可能无法访问模型的置信度和损失函数等信息，只能获取模型返回的识别结果。其次，无线通信中的信号数据由天线接收并转换为数字格式。这意味着在实际环境中攻击者很难对每个信号数据点进行精确的修改，因为这种操作需要对物理层的信号进行精确控制，在许多情况下较难实现。因此，在信号对抗样本的生成中，除了关注攻击效果外，还需要考虑其生成的难度、效率，以及在实际无线通信环境下的可实现性。

对抗样本的存在引发了安全性和鲁棒性问题，也给无线通信带来潜在威胁。关于对抗样本存在的根本原因，研究者们提出了多种解释[15-16]。其中 Shamir 等[17]证明，对于任意 m 分类问题，只需修改样本 $m+1$ 个点即可找到对抗样本，并提出基于 0 范数的攻击方法。该方法通过将不同类别样本的连线路径输入神经网络，选择样本 $m+1$ 个点的子集作为待扰动的点，沿输出路径找到对抗样本使网络输出错误类别。然而在验证 0 范数攻击的实验中，生成的对抗样本的其他范数可能非常大，与信号实际情况不符。尽管这种方法不能直接使用，但启发我们研究不同类别样本连线路径上的线性组合。我们在实验中观察到，当不同样本的连线段输入神经网络时，即使在与原始样本非常接近的地方(如线性加权系数小于 0.05)，其线性组合依然可能引发神经网络的分类错误。我们将这种非常靠近原始样本的线性组合称为近端线性组合。有研究表明，神经网络的决策边界可能会非常接近给定的样本数据[18]，这意味着近端线性组合中就可能含有对抗样本。

受此启发，本文提出了一种基于信号近端线性组合的黑盒查询对抗样本生成方法。该方法利用神经网络决策边界可能就在样本数据附近的特性，对原始样本进行近端线性组合，查询统计其识别结果，从而确定每类原始信号最容易受到线性组合影响的特定目标信号。在测试攻击时，根据待攻击信号的类别选择对应目标信号进行近端线性组合生成对抗样本。这种近端线性组合在信号处理中是一个基础且易实现的操作，可以直接生成对抗样本。下文将介绍相关工作、方法和实验设计，并讨论实验结果和分析结论。

2 相关工作

信号识别深度神经网络分类器为：

$$f(\cdot;\boldsymbol{\theta}):\chi\to\boldsymbol{R}^m \tag{1}$$

其中，$\boldsymbol{\theta}$ 为模型的参数，$\chi\subset R^n$ 是模型输入空间，n 是输入的维度，m 是类别总数。对于任意输入 $\boldsymbol{x}$，神经网络预测结果为：

$$\hat{l}(\boldsymbol{x},\boldsymbol{\theta})=\arg\max_k f_k(\boldsymbol{x},\boldsymbol{\theta}) \tag{2}$$

与图像对抗样本的定义方式相同，一个非定向的信号对抗扰动定义如下：

$$\begin{aligned}&\arg\min_{\boldsymbol{r}_x}\|\boldsymbol{r}_x\|_p\\&\text{s.t. }\hat{l}(\boldsymbol{x},\boldsymbol{\theta})\neq\hat{l}(\boldsymbol{x}+\boldsymbol{r}_x,\boldsymbol{\theta})\end{aligned} \tag{3}$$

其中，$\boldsymbol{x}_r=\boldsymbol{x}+\boldsymbol{r}_x$ 即为要寻找的对抗样本，且满足条件 $\boldsymbol{x}+\boldsymbol{r}_x\in\chi$。为了确保难以察觉，扰动还需要满足一定的约束条件。$\|\cdot\|_p$ 表示 p 范数约束，常见的对抗样本范数可以是 0，1，2 或无穷。对于信号数据，2 范数的平方值代表离散信号的短时能量，因此可以使用 2 范数来直观刻画扰动的大小。

2.1 Mixup 数据线性组合

Mixup 是一种用于增强深度学习模型的鲁棒性和泛化能力的数据增强方法[19]。该方法通过将两个不同的样本 x_1 与 x_2 进行全局的线性组合，生成一个新样本 x_α，对原始两个标签 y_1 和 y_2 进行对应线性插值得到其标签 y_α：

$$\begin{aligned}\boldsymbol{x}_\alpha&=\alpha\boldsymbol{x}_1+(1-\alpha)\boldsymbol{x}_2\\y_\alpha&=\alpha y_1+(1-\alpha)y_2\end{aligned} \tag{4}$$

Mixup 方法通过全局线性组合和标签插值提高模型的泛化能力并防止过拟合，增加训练集的多样性。然而 Mixup 的线性组合并不适用于所有任务和数据集。一是对于不平衡数据集，例如常见的 I/Q 信号调制识别数据集 RML2016[20]，RML2018[21]中，新生成的样本可能会偏向于数量较多的数字调制类别，从而影响模型的泛化能力；二是在连续型数据如信号上使用 Mixup 可能会扭曲数值，进而改变其携带信息；三是对于有噪声的数据，如无线信号上的噪声较大时，使用 Mixup 可能导致性能下降或结果不稳定。

与此不同，本文方法仅在与原始样本非常接近的局部范围内进行了微小的线性组合，而不是涵盖整个数据范围。由于这种近端线性组合与原始样本非常接近，因此认为信号携带数据本身可以保持不变。仅进行样本的混合，而不涉及标签的。这种专注于原始样本附近的策略与 Mixup 有明显的不同。

3 基于近端线性组合的对抗样本生成方法

本文研究的是在无法直接访问信号调制识别模型参数和梯度的情况下的黑盒对抗攻击，在此类攻击条件设定下，提出了一种利用原始信号样本近端线性组合查询黑盒模型对抗样本生成方法。对于给定的两个不同类别的信号数据 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$，定义其对于 $\boldsymbol{x}_i$ 的近端线性组合表示为：

$$prox(\boldsymbol{x}_{ij})=(1-\alpha)\cdot\boldsymbol{x}_i+\alpha\cdot\boldsymbol{x}_j \tag{5}$$

其中，α 是一个较小的系数，本文确定为 $0<\alpha\leqslant 0.05$，以确保 $prox(\boldsymbol{x}_{ij})$ 与原始样本 $\boldsymbol{x}_i$ 之间的距离非常接近，从而使对信号数据本身的修改尽可能微小。因为它仅在原始样本附近的局部范围内进行线性组合操作，所以称为近端。

信号调制识别的神经网络为 $f(\boldsymbol{x};\boldsymbol{\theta})$，其中 $\boldsymbol{\theta}$ 是神经网络训练参数，在黑盒攻击条件下无法直接获取，只能得到网络对于给定信号 $\boldsymbol{x}$ 的预测标签 $\hat{l}(\boldsymbol{x},\boldsymbol{\theta})$。基于近端线性组合的对抗样本生成方法整体流程如图 1 所示，主要包括以下步骤：

步骤 1　子集选择与近端线性组合。选择合适子集，将每个样本作为原始信号，遍历其他类别样本作为扰动信号进行近端线性组合。

步骤 2　黑盒查询与确定最佳扰动信号。查询所有近端线性组合识别结果，统计每个类别所有信号的近端线性组合导致识别结果出错的次数。确定每一类样本最容易受到影响的目标样本，作为最佳扰动信号。

步骤 3　生成对抗样本。根据待攻击的信号类别，使用对应的最佳扰动信号进行近端线性组合，从而生成对抗样本。

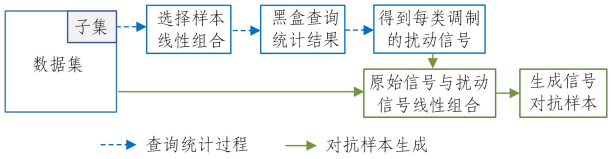


图 1　信号对抗样本生成流程图

Fig. 1　Process of signal adversarial examples generation

3.1 生成样本的近端线性组合

首先，根据调制类别和信噪比对数据集 C 进行分层采样，得到一个随机子集 C^{sub}。在这个子集中将每一个样本视为原始样本 $\boldsymbol{x}_i$，并从其他调制类别的样本中选择目标扰动样本 $\boldsymbol{x}_j$。得到的原始信号的全部近端线性组合的集合为：

$$\boldsymbol{X}_{ij}=\left\{\boldsymbol{x}_{ijk}\mid\boldsymbol{x}_{ijk}=\left(1-\frac{k}{K}A\right)\cdot\boldsymbol{x}_i+\frac{k}{K}A\cdot\boldsymbol{x}_j,k=1,\cdots,K\right\} \tag{6}$$

其中，权重参数 A 设置为 0.05，整个线性组合集合满足近端线性组合条件。为了探索神经网络在原始样本附近的决策变化，以及寻找潜在的对抗样本，对子集中所有样本进行近端线性组合，并在接下来的步骤中进行黑盒查询统计。

3.2 确定最佳扰动信号

对子集 C^{sub} 中不同类别的原始信号 $\boldsymbol{x}_i$ 和目标信号 $\boldsymbol{x}_j$ 执行近端线性组合操作并生成 $\boldsymbol{X}_{ij}$。随后将 $\boldsymbol{X}_{ij}$ 全部近端线性组合输入神经网络进行黑盒查询，统计这 K 个近端线性组合结果中模型识别出现错误的数量，记为 $d(\boldsymbol{x}_{ij})$：

$$d(\boldsymbol{x}_{ij})=|\{\boldsymbol{x}_{ijk}\mid\hat{l}(\boldsymbol{x}_{ijk},\boldsymbol{\theta})\neq y_i\}| \tag{7}$$

其中，$|\{\cdot\}|$ 为集合的势，表示集合中所包含元素个数。遍历子集中每个原始样本 $\boldsymbol{x}_i$ 以及作为扰动样本的 $\boldsymbol{x}_j$，统计子集中所选类别的所有近端线性组合在神经网络中的识别结果，可以得到不同类别信号使用近端线性组合的方法攻击模型总的出错次数。针对某个扰动目标信号 $\boldsymbol{x}_j$，统计第 p 类所有信号与其进行近端线性组合发生识别错误的总次数：

$$d_p(\boldsymbol{x}_j)=\sum_{i=1}^{N_p}d(\boldsymbol{x}_{ij}) \tag{8}$$

对于每类信号，统计出不同扰动目标信号造成的识别错误总次数，选取导致一类信号出错最多的扰动目标信号，将其称为“最佳扰动信号”。

$$\boldsymbol{x}^{p\prime}=\arg\max_{\boldsymbol{x}_j}(d_p(\boldsymbol{x}_j)) \tag{9}$$

对于子集中每一种调制类别，找到一个最佳的扰动信号，以最大程度地干扰神经网络的决策。

3.3 生成对抗样本

完成上述统计后，在黑盒攻击测试中，当需要为某一信号生成对抗样本时，选择对应类别的最佳扰动信号，生成原始样本的近端线性组合，作为对抗样本 $\boldsymbol{x}_i^*$：

$$\boldsymbol{x}_i^*=(1-\alpha)\cdot\boldsymbol{x}_i+\alpha\cdot\boldsymbol{x}_p{}' \tag{10}$$

生成对抗样本时，在满足近端线性组合的前提下，可以参考信号所能容忍的最大无穷范数扰动来设定近端线性组合的系数 α，该系数的设定值决定了与最佳扰动信号的混合程度，即它会直接影响对抗样本生成的效果。具体而言，通过调整 α 的大小，可以控制扰动的强度及相应的攻击效果，当 α 较大时，能够获得更明显的攻击效果；而在 α 较小时，扰动更小，攻击更为隐蔽。图 2 展示了对抗样本的生成过程。将第 p 类样本的最佳扰动信号 $\boldsymbol{x}_p{}'$ 添加到这类样本上，得到的近端线性组合中，有些可以成功改变网络识别结果。

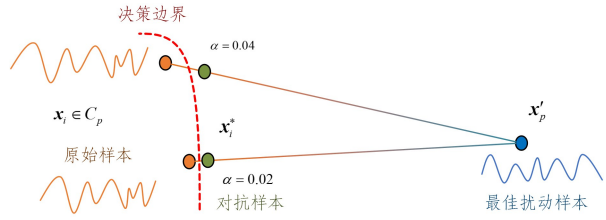


图 2　近端线性组合生成对抗样本的示意图

Fig. 2　Generation of signal adversarial examples via proximal linear combination

虽然遍历子集进行近端线性组合并查询模型识别结果可能会有一定的开销，但一旦完成统计，后续生成对抗样本时无需再查询。对每类信号中的任意新样本，都可以通过与最佳扰动信号进行近端线性组合生成对抗样本，这种方式在信号处理领域非常常见且易于实施。

4 实验和结果

本章使用公开信号数据集和典型神经网络结构进行实验，生成对抗样本并与随机噪声攻击对比。在信号处理领域，添加随机噪声是一种常见的攻击方法，其通过添加随机噪声使模型难以识别信号特征，从而降低预测准确性。这种不针对特定网络的攻击方法简单迅速，并且能够在一定程度上干扰神经网络的性能。实验表明所提方法与随机噪声扰动相比有显著优势。

4.1 实验设置

实验使用了 I/Q 信号调制识别数据集 RML2018[21]，其中每个信号尺寸为 1 024 * 2，该数据集包含数字调制和模拟调制共 24 个调制类别。每种调制类别都包含 26 种信噪比，范围为 −20～30 dB。每种信噪比下有 4 096 个 I/Q 信号样本。考虑到在信噪比过低的情况下，网络识别的出错概率会增加，实验将重点放在信噪比非负的信号上，一共选取 16 种信噪比，以保证实验结果的稳定性和可靠性。因此数据集的规模被调整为(24 * 16 * 4 096, 1 024, 2)。

在训练识别网络之前，对原始数据集进行了预处理。原始数据集中的信号并未进行归一化或功率归一化处理，导致不同类别样本间的取值范围存在明显差异。例如部分模拟信号的幅值可达 20，而某些数字信号的幅值仅为 1～2。虽然直接训练可以获得较高的准确率，但实际情况中信号幅值的差异通常没有这么大，且这样生成的对抗样本更容易被检测出来。为了避免过大的信号幅值差异影响对抗样本，我们对数据集进行了最小-最大归一化处理。

实验中使用的调制识别网络是基于残差网络(ResNet)的 I/Q 信号识别模型[21]。在归一化后的测试集所有信噪比中，干净样本的准确率达到了 86.5%。随后根据调制类别和信噪比进行分层均匀抽样，从每种调制类别和每种信噪比中均匀抽取 64 个样本作为子集，即总共 24 * 16 * 64 = 24 576 个样本，占总数据集的 12.5%。在后续实验中，近端线性组合的参数权重不超过 0.05，设定子集查询时扰动最大值范围 α 为 0.02, 0.03, 0.04 和 0.05。作为对比，随机噪声扰动的最大值范围也同样包括这些值。通过这样的设置，可以评估所提方法在生成对抗样本方面的性能，并与随机扰动进行比较。最后将该方法得到的对抗样本在另一个神经网络上进行结果的验证。该神经网络结构基于 ResNet[21]，去掉了网络第二个全连接层，并使用了不同的优化器和训练参数。接下来，进行子集近端线性组合的黑盒查询结果统计。

4.2 子集上查询结果统计

为了确定每个类别中各个样本的近端线性组合对于模型分类的影响，接下来对子集进行了查询统计。在原始样本附近根据式(6)选取了 K 个近端线性组合进行实验。在这里 K 被设定为 128。然后统计了这 128 个近端线性组合中导致模型分类错误的次数 $d(\boldsymbol{x}_{ij})$。

针对所有样本的近端线性组合识别结果，从中选择干净样本网络分类正确而近端线性组合分类有出错的情况来计算攻击成功率 SR。

$$n_p = N_p - |\{\hat{l}(\boldsymbol{x}_i,\theta) | \hat{l}(\boldsymbol{x}_i,\theta) \neq y_i\}| \tag{11}$$

$$SR = \frac{n_p - |\{\boldsymbol{x}_i | \boldsymbol{x}_i \in C_p^{\text{sub}}, \boldsymbol{x}_j \in C^{\text{sub}}, (\Sigma d(\boldsymbol{x}_{ij})) = 0\}|}{n_p} \tag{12}$$

其中，n_p 为第 p 类干净样本识别准确的个数，N_p 是子集上第 p 类样本的数量。结果如图 3 所示，对于某些高阶调制的样本，近端线性组合能轻易导致黑盒模型分类错误。对于大多数调制，在信噪比较低时，近端线性组合具有较好的攻击效果。

接下来对于某个类别，根据式(9)和式(10)统计不同目标扰动信号下该类近端线性组合黑盒识别结果发生错误的总次数。对于不同的原始类别，选择对网络识别结果影响更大的样本，并将其被视为黑盒模型对应类别的最佳扰动信号 $\boldsymbol{x}_p'$。

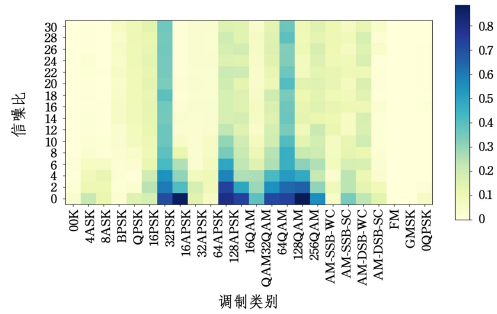


图 3 子集上每类样本近端线性组合的攻击成功率

Fig. 3 Attack success rate of proximal linear combinations on subsets

4.3 测试阶段对抗样本生成

完成黑盒查询统计后，得到全部类别对应的最佳扰动信号 $\boldsymbol{x}_p'$。对所有待生成对抗样本的信号，根据统计结果选择对应类别的最佳扰动信号，并选择具体参数进行近端线性组合，从而生成对抗样本。接着将这些对抗信号输入待攻击模型中进行测试。在攻击扰动参数 $\alpha = 0.03$ 的情况下，模型准确率随信噪比变化趋势如图 4 所示，当信噪比较高时，模型的识别准确率可以从 94%下降至 50%。

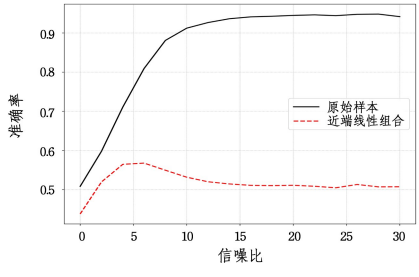


图 4 黑盒攻击下模型准确率随信噪比变化图

Fig. 4 Model accuracy varies with SNR under black-box attacks

当 $\alpha = 0.03$ 时，分析使用最佳扰动信号进行攻击的成功率。对每个调制类别(共 24 种)和每个信噪比(共 24 种)总计

384 种数据场景进行了测试，每种数据包括 512 个样本。如图 5 所示，在全部 384 个数据场景中，有将近 1/3（大约 29.7%）的场景下，攻击成功率超过了 50%，说明近端线性组合是一种有效的攻击方法。近端线性组合对部分调制类（如 QPSK 和 8PSK）攻击效果显著。

此外，实验还观察了单个信号样本在生成对抗样本后波形变化是否显著，以 $\alpha=0.03$ 为例，图 6 展示了两个信号样本生成近端线性组合前后的波形对比。其中图 6(a)和图 6(d)是原始信号的波形图，图 6(b)和图 6(e)是最佳扰动信号的波形图，图 6(c) 和图 6(f)是近端线性组合生成的对抗样本的波形图。在第一个示例中，图 6(a)是一个调制类别为 4ASK 的原始信号，其与对应的最佳扰动信号（见图 6(b)）进行近端线性组合，生成了新的信号（见图 6(c)），这个新信号被黑盒模型识别为了 8ASK 调制类别。在第二个示例中，图 6(d)是一个调制类别 AM-DSB-WC 的原始信号，其与对应的最佳扰动信号（见图 6(e)）进行近端线性组合，生成的对抗信号（见图 6(f)）被模型识别为 AM-DSB-SC 调制类别。

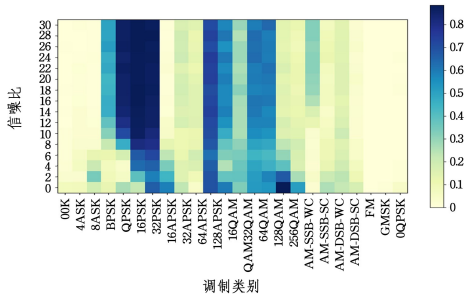
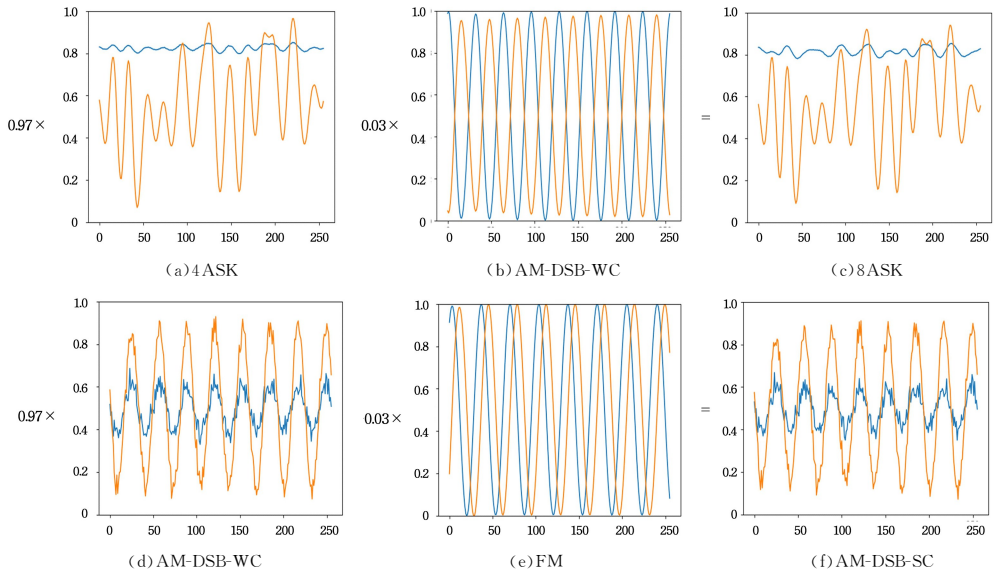


图 5 全部数据集上最佳扰动信号的攻击成功率

Fig. 5 Attack success rate using optimal perturbation signals on the entire dataset



(a)4ASK (b)AM-DSB-WC (c)8ASK

(d)AM-DSB-WC (e)FM (f)AM-DSB-SC

图 6 添加对抗样本前后攻击方法的波形变化图

Fig. 6 Waveform changes of attack methods before and after adding adversarial samples

4.4 与随机噪声的效果比较

为了进一步评估本文提出的方法在直接生成对抗样本上的实际效益，实验将其与随机噪声攻击在同等无穷范数扰动下的攻击效果进行了比较。如图 7 所示，在相同扰动强度条件下，近端线性组合的攻击相比随机噪声攻击能更大幅度地降低目标模型的识别准确率，而且本文对抗样本攻击与噪声攻击同样具有快速、易实现的特点，这进一步表明了本文方法的有效性和优越性。特别是在信噪比较高的情况下，当扰动强度较低时（如 $\alpha=0.02$），随机噪声攻击几乎没有改变目标网络准确率。当扰动较大时（如 $\alpha=0.04$），近端线性组合攻击可以大幅降低目标模型准确率。

接下来对扰动大小进行统计分析。通过从对抗信号中减去原始的干净信号得到对抗扰动。计算 2 范数平均值可以发现，与随机噪声相比，近端线性组合方法产生的扰动具有更小的 2 范数，即扰动幅度更小，扰动功率占干净信号功率的比例也更小。不同扰动强度对比结果如表 1 所列。从表中可以看出，使用近端线性组合的方式添加扰动，扰动噪声功率更小。

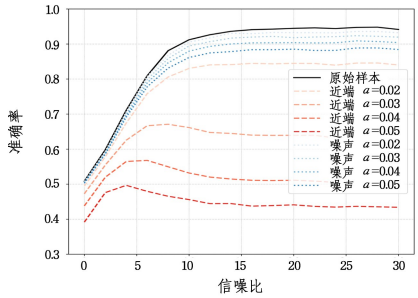


图 7 不同扰动强度下的准确率对比

Fig. 7 Accuracy comparison under different perturbation

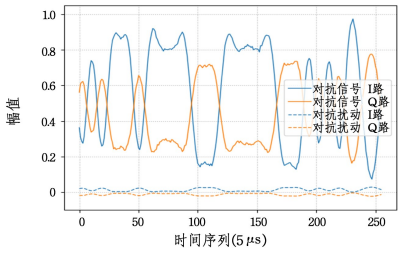
表 1 不同扰动强度下的攻击效果对比

Table 1 Comparison of attack effects under different perturbation

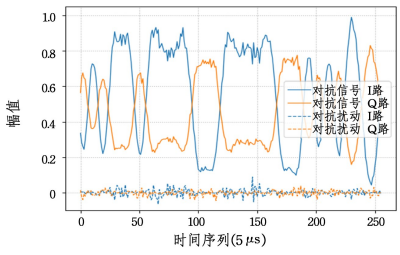
扰动强度 α	噪声扰动 L2	噪声扰动功率占比	对抗扰动 L2	对抗扰动功率占比
0.02	0.523	0.045	0.459	0.035
0.03	0.784	0.101	0.689	0.078
0.04	1.045	0.180	0.918	0.140
0.05	1.306	0.280	1.147	0.219

此外近端线性组合生成的对抗样本在波形上看起来更加

平滑自然。如图 8 所示，一个真实标签为 BPSK 的信号添加扰动后被识别为其他调制。相比随机噪声攻击，使用近端线性组合的方式直接叠加最佳扰动信号，波形更加自然。



(a)BPSK 识别为 0QPSK



(b)BPSK 识别为 8PSK

图 8 近端线性组合和随机噪声攻击波形对比

Fig. 8 Waveform comparison between proximal linear combination and random noise attacks

4.5 其他神经网络上的攻击效果验证

为了验证该方法生成的对抗样本在其他神经网络上的迁移能力，实验选择了与原始网络结构近似的另一个神经网络进行验证。原始目标网络对干净样本的识别准确率为 86.49%；新目标网络为 86.18%，两者性能相当。实验将原目标网络生成的信号对抗样本输入新的网络中，并计算其准确率，以观察对抗样本的迁移性。实验结果如图 9 所示，生成的对抗样本在结构类似且任务相同的神经网络上具有一定的有效性，这证明了该方法能够生成具有迁移性的对抗样本。但相较于图 7 所示的原始目标网络，迁移攻击性能略显不足，在信噪比较高且扰动较大的情况下，原始网络准确率下降了 50%，新的网络下降了 45%。这表明，即使无法对目标网络进行黑盒查询，也可以基于类似结构和任务的替代模型进行查询统计，然后使用本方法进行攻击。但需注意，如果替代网络结构与目标网络差异较大，对抗样本的迁移性可能会明显降低。

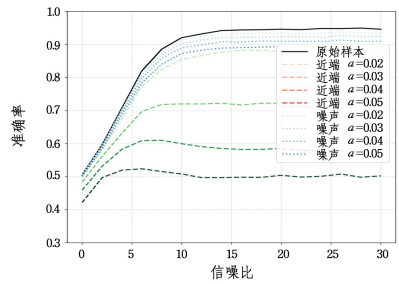


图 9 对抗样本在其他神经网络上的攻击效果

Fig. 9 Attack effects on other neural networks

4.6 与 HopSkipJumpAttack 方法的比较

本文提出的黑盒攻击方法适用于仅能访问网络识别结果而无法获取置信度或损失函数等信息的场景。因此，许多基于有限差分的梯度估计方法难以直接应用。现有的黑盒攻击方法中，利用网络识别结果的方法主要有决策边界法和其改进版 HopSkipJumpAttack(HSJA)。实验使用 HSJA 的黑盒攻击方法与本文方法进行比较。

本文方法在生成对抗样本前，需要对目标网络进行预先查询以获取最佳扰动信号。一旦完成查询，后续生成对抗样本时则不再需要额外的查询。而 HSJA 在攻击过程中需要对每一批信号进行多次查询以更精确地估计模型梯度，从而迭代生成效果更好的对抗样本。HSJA 查询时间会随着待生成的对抗样本数量的增加而增加。实验结果显示，当信号数据规模为 24×16×512 个样本时，本文方法可以将目标网络的准确率降至 44%，而 HSJA 则可以将其降至 15.8%。但本文方法在攻击过程中不再需要额外的查询，因此当需要生成大量对抗样本时，本文方法的攻击耗时几乎不会增加，而 HSJA 的攻击时间则与待生成的信号数量直接相关。

表 2 列出了在不同数据规模下两种方法在 NVIDIA TITAN RTX 的 GPU 上的运行耗时对比。可以看出，随着数据规模的增大，本文方法在攻击耗时上的增长几乎可以忽略不计，而 HSJA 的攻击耗时则线性增加。

表 2 两种黑盒攻击方法的耗时对比

Table 2 Time consumption of two black-box attacks

数据规模	方法	统计耗时/h	攻击耗时/h
当前规模	本文方法	130	可忽略
	HSJA	无	76.5
2 倍规模	本文方法	130	可忽略
	HSJA	无	153
5 倍规模	本文方法	130	可忽略
	HSJA	无	382.5

5 结论

针对无法实时访问信号识别神经网络模型参数和梯度的情况，本文提出了一种新的对抗样本生成方法。该方法利用信号样本之间的近端线性组合生成对抗样本。通过在一部分子集上进行近端线性组合与黑盒查询，能够实现针对不同调制类别的最佳扰动信号的选择。对于要生成对抗样本的数据，在原始信号附近叠加这些最佳扰动信号进行近端线性组合。实验中，该方法显著降低了黑盒模型识别的准确率，从 94%降至 50%。相比随机噪声扰动，对抗样本攻击的成功率更高，同时扰动功率也更低。该方法得到的对抗样本在结构近似的神经网络中也具有一定迁移性。

值得强调的是，该方法在生成对抗样本时无需再进行查询，是一种高效攻击方法。与需要多轮迭代生成对抗样本的方法相比，特别是在实时高速处理的信号数据场景下，直接采用近端线性组合的方式更加简洁高效。相对于需要直接修改信号特定点值的方法，近端线性组合更易于实现。

结束语 在无线信号识别领域，对于很多黑盒对抗攻击方法来说，直接从天线获取的信号数据在攻击过程中通常较难实现反复查询和多次迭代修改。本文提出的攻击方法通过近端线性组合来生成对抗信号，避免了对抗攻击时繁琐的查询和迭代过程，对抗样本更简洁、高效且易于实现。未来，在防御上，可以在通信系统设计时考虑这类攻击的可能性。例如

增加近端线性组合数据进行对抗训练，或者集成结构更加复杂的神经网络模型。在攻击上，可以进一步拓展到其他无线通信应用场景，例如信号检测的对抗样本生成应用中。

参 考 文 献

[1] ZHANG F X，LUO C B，XU J L，et al. Deep learning based automatic modulation recognition：Models，datasets，and challenges [J]. Digital Signal Processing，2022，129：14.

[2] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. Intriguing properties of neural networks [C]// Proceedings of the 2nd International Conference on Learning Representations. 2014.

[3] ADESINA D，HSIEH C C，SAGDUYU Y E，et al. Adversarial machine learning in wireless communications using RF data：A review [J]. IEEE Communications Surveys & Tutorials，2023，25(1)：77-100.

[4] HAMEED M Z，GYORGY A，GUNDUZ D. The Best Defense Is a Good Offense：Adversarial Attacks to Avoid Modulation Detection [J]. IEEE Transactions on Information Forensics and Security，2021，16：1074-1087.

[5] SADEGHI M，LARSSON E G. Adversarial Attacks on Deep-Learning Based Radio Signal Classification [J]. IEEE Wirel Commun Lett，2019，8(1)：213-216.

[6] XU H，MA Y，LIU H C，et al. Adversarial Attacks and Defenses in Images，Graphs and Text：A Review [J]. Int J Autom Comput，2020，17(2)：151-178.

[7] LIN Y，ZHAO H，MA X，et al. Adversarial attacks in modulation recognition with convolutional neural networks [J]. IEEE Transactions on Reliability，2020，70(1)：389-401.

[8] HUANG S N，LI Y X，MAO Y Hg，et al. Black-box transferable adversarial attacks based on ensemble advGAN[J]. Journal of Jilin University (Engineering and Technology Edition)，2022，52(10)：2391-2398.

[9] LIN Y，ZHAO H J，TU Y，et al. Threats of Adversarial Attacks in DNN-Based Modulation Recognition [C]// Proceedings of IEEE INFOCOM 2020－IEEE Conference on Computer Communications. IEEE，2020：2469-2478.

[10] KIM B，SAGDUYU Y E，DAVASLIOGLU K，et al. Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels [C]// Proceedings of 2020 54th Annual Conference on Information Sciences and Systems. IEEE，2020：1-6.

[11] CHEN P Y，ZHANG H，SHARMA Y，et al. Zoo：Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]// Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017：15-26.

[12] CHEN J，JORDAN M I，WAINWRIGHT M J. Hopskipjumpattack：A query-efficient decision-based attack [C]// 2020 IEEE Symposium on Security and Privacy. IEEE，2020：1277-1294.

[13] BAHRAMALI A，NASR M，HOUMANSADR A，et al. Robust adversarial attacks against DNN-based wireless communication systems [C]// Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021：126-140.

[14] SAGDUYU Y E，SHI Y，ERPEK T. Adversarial deep learning for over-the-air spectrum poisoning attacks [J]. IEEE Transactions on Mobile Computing，2019，20(2)：306-319.

[15] GOODFELLOW I J，SHLENS J，SZEGEDY C. Explaining and Harnessing Adversarial Examples[C]// Proceedings of the International Conference on Learning Representations. 2014.

[16] ILYAS A，SANTURKAR S，TSIPRAS D，et al. Adversarial examples are not bugs，they are features [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc.，Red Hook，NY，USA，125-136.

[17] SHAMIR A，SAFRAN I，RONEN E，et al. A simple explanation for the existence of adversarial examples with small hamming distance [J]. arXiv：1901.10861，2019.

[18] SHAMIR A，MELAMED O，BENSHMUEL O. The dimpled manifold model of adversarial examples in machine learning [J]. arXiv：2106.10151，2021.

[19] ZHANG H，CISSE M，DAUPHIN Y N，et al. mixup：Beyond empirical risk minimization [J]. arXiv：1710.09412，2017.

[20] O'SHEA T J，CORGAN J，CLANCY T C. Convolutional Radio Modulation Recognition Networks [C]// Engineering Applications of Neural Networks：17th International Conference，EANN 2016，Aberdeen，UK，September 2－5，2016，Proceedings 17：Springer International Publishing，2016：213-226.

[21] O'SHEA T J，ROY T，CLANCY T C. Over-the-Air Deep Learning Based Radio Signal Classification [J]. IEEE J Sel Top Signal Process，2018，12(1)：168-179.

GUO Yuqi，born in 1991，postgraduate. Her main research interests include intelligent signal processing and artificial intelligence security.

WANG Linyuan，born in 1985，Ph.D，associate professor，master supervisor. His main research interests include sparse optimization theory，mathematical foundations of artificial intelligence，and hybrid intelligence in brain-computer interaction.

（责任编辑：何杨）