(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118247603 A
(43)申请公布日 2024.06.25

(21)申请号 202410425526.8

(22)申请日 2024.04.10

(71)申请人 复旦大学
地址 200433 上海市杨浦区邯郸路220号
申请人 中电金信数字科技集团股份有限公司

(72)发明人 吕智慧 郭旭 郭恒其 况文川 单海军

(74)专利代理机构 上海德昭知识产权代理有限公司 31204
专利代理师 程宗德

(51)Int.Cl.
G06V 10/774(2022.01)
G06V 10/82(2022.01)
G06N 3/045(2023.01)
G06N 3/0475(2023.01)
G06N 3/094(2023.01)
G06N 3/084(2023.01)
G06N 3/0985(2023.01)
G06V 10/764(2022.01)

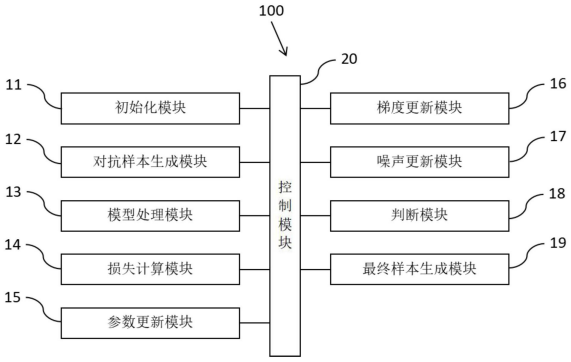
权利要求书3页 说明书8页 附图5页

(54)发明名称

基于梯度修改的黑盒对抗样本生成方法及装置

(57)摘要

本发明提供了一种基于梯度修改的黑盒对抗样本生成方法及装置,具有这样的特征,包括以下步骤:步骤S1初始化参数;步骤S2根据干净样本和代理模型得到干净输出;步骤S3根据噪声和干净样本得到对抗样本;步骤S4根据对抗样本和代理模型得到对抗输出;步骤S5根据干净输出和对抗输出得到损失;步骤S6根据损失得到更新参数;步骤S7根据更新参数和梯度得到新梯度;步骤S8根据新梯度和噪声得到新噪声;步骤S9判断迭代次数是否已满,若是则进入步骤S10,若否则将新梯度作为梯度并将新噪声作为噪声进入步骤S3;步骤S10根据新噪声和干净样本得到最终对抗样本。总之,本方法能够快速生成高攻击成功率的对抗样本。



CN 118247603 A

1.一种基于梯度修改的黑盒对抗样本生成方法,用于根据代理模型$\tilde{F}$和干净样本x生成对应的最终对抗样本x',其特征在于,包括以下步骤:

步骤S1,设置当前迭代数i为0、迭代总次数为N、噪声Δx_i为$\vec{0}$以及梯度g_i为$\vec{0}$;

步骤S2,将所述干净样本x输入所述代理模型$\tilde{F}$,得到干净输出$\tilde{F}(x)$;

步骤S3,将所述噪声Δx_i和所述干净样本x相加,得到对抗样本x_i;

步骤S4,将所述对抗样本x_i输入所述代理模型$\tilde{F}$,得到对抗输出$\tilde{F}(x_i)$;

步骤S5,根据所述干净输出$\tilde{F}(x)$和所述对抗输出$\tilde{F}(x_i)$计算得到损失$L(\Delta x_i)$;

步骤S6,根据所述损失$L(\Delta x_i)$,通过反向传播计算得到更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$;

步骤S7,根据所述更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$更新所述梯度$g_i$,得到梯度$g_{i+1}$;

步骤S8,根据所述梯度g_{i+1}更新所述噪声Δx_i,得到噪声Δx_{i+1};

步骤S9,所述当前迭代数i加1,判断所述当前迭代数i是否等于所述迭代总次数,若是,则进入步骤S10,若否,则将所述梯度g_{i+1}作为所述梯度g_i并将所述噪声Δx_{i+1}作为所述噪声Δx_i,进入步骤S3;

步骤S10,将所述噪声Δx_{i+1}与所述干净样本x相加,得到所述最终对抗样本x',

其中,所述代理模型$\tilde{F}$通过对具有Transformer结构的神经网络模型插入梯度修改层构成,

所述梯度修改层仅在所述反向传播中工作,根据输入的梯度G输出对应的梯度G_out,

所述梯度G'的计算表达式为:

$G' = \max(\min(G, \mu+\lambda*\sigma), \mu-\lambda*\sigma)$,

$G\_out = G' * \gamma$,

式中μ为所述梯度G的均值,σ为所述梯度G的标准差,λ和γ均为预设的超参数。

2.根据权利要求1所述的基于梯度修改的黑盒对抗样本生成方法,其特征在于:

其中,所述具有Transformer结构的神经网络模型为ViT模型,

所述ViT模型包含至少一个Transformer块,

每个所述Transformer块包括注意力层、投影层和多层感知机,

所述代理模型$\tilde{F}$的所述注意力层与所述投影层之间插入有所述梯度修改层,

所述代理模型$\tilde{F}$的所述投影层与所述多层感知机之间插入有所述梯度修改层。

3.根据权利要求1所述的基于梯度修改的黑盒对抗样本生成方法,其特征在于:

其中,所述对抗样本x_i的计算表达式为:

$x_i = \Delta x_i + x$。

4.根据权利要求1所述的基于梯度修改的黑盒对抗样本生成方法,其特征在于:

其中,所述损失$L(\Delta x_i)$的计算表达式为:

$$L(\Delta x_i) = CE(\tilde{F}(x), \tilde{F}(x_i)) - \beta \cdot \|\Delta x_i\|_2,$$

$$CE(\tilde{F}(x),\tilde{F}(x_i)) = -\sum_{k=1}^{D}\tilde{F}(x)_k \cdot \log \tilde{F}(x_i)_k,$$

$$\|\Delta x_i\|_2 = \sqrt{\sum_{m=1}^{H}(\Delta x_i)_m^2},$$

式中D为代理模型输出的维度，$\tilde{F}(x)_k$为经过代理模型$\tilde{F}$判断干净样本x属于第k类的预测值，$\tilde{F}(x_i)_k$为经过代理模型$\tilde{F}$判断第i轮迭代得到的对抗样本x_i属于第k类的预测值，H为输入样本的维度，$(\Delta x_i)_m$为输入样本维度m上的数值。

5.根据权利要求1所述的基于梯度修改的黑盒对抗样本生成方法，其特征在于：

其中，所述梯度g_{i+1}的计算表达式为：

$$g_{i+1} = g_i + \alpha \cdot \frac{\partial L(\Delta x_i)}{\partial \Delta x_i},$$

式中α为给定的超参数。

6.根据权利要求1所述的基于梯度修改的黑盒对抗样本生成方法，其特征在于：

其中，所述噪声Δx_{i+1}的计算表达式为：

$\Delta x_{i+1} = \Delta x_i + \in /N \cdot sign(g_{i+1})$，

式中∈为给定的超参数。

7.一种基于梯度修改的黑盒对抗样本生成装置，用于根据代理模型$\tilde{F}$和干净样本x生成对应的最终对抗样本x'，其特征在于，包括：

初始化模块、对抗样本生成模块、模型处理模块、损失计算模块、参数更新模块、梯度更新模块、噪声更新模块、判断模块和最终样本生成模块，

其中，所述初始化模块用于设置当前迭代数i为0、迭代总次数为N、噪声Δx_i为$\vec{0}$以及梯度g_i为$\vec{0}$，

所述对抗样本生成模块用于将所述噪声Δx_i和所述干净样本x相加，得到对抗样本x_i，

所述模型处理模块存储有所述代理模型$\tilde{F}$，用于根据所述干净样本x生成干净输出$\tilde{F}(x)$，根据所述对抗样本x_i生成对抗输出$\tilde{F}(x_i)$，

所述损失计算模块用于根据所述干净输出$\tilde{F}(x)$和所述对抗输出$\tilde{F}(x_i)$计算得到损失$L(\Delta x_i)$，

所述参数更新模块用于根据所述损失$L(\Delta x_i)$，通过反向传播计算得到更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$，

所述梯度更新模块用于根据所述更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$更新所述梯度$g_i$，得到梯度$g_{i+1}$，

所述噪声更新模块用于根据所述梯度g_{i+1}更新所述噪声Δx_i，得到噪声Δx_{i+1}，

所述判断模块用于将所述当前迭代数i加1，判断所述当前迭代数i是否等于所述迭代总次数，若是，则控制所述最终样本生成模块运行，若否，则将所述梯度g_{i+1}作为所述梯度g_i并将所述噪声Δx_{i+1}作为所述噪声Δx_i，控制所述对抗样本生成模块运行，

所述最终样本生成模块用于将所述噪声Δx_{i+1}与所述干净样本x相加，得到所述最终对

抗样本x',

其中,所述代理模型$\tilde{F}$通过对具有Transformer结构的神经网络模型插入梯度修改层构成,

所述梯度修改层仅在所述反向传播中工作,根据输入的梯度G输出对应的梯度G_out,

所述梯度G'的计算表达式为:

$G' = \max(\min(G, \mu+\lambda*\sigma), \mu-\lambda*\sigma)$,

$G\_out = G' * \gamma$,

式中μ为所述梯度G的均值,σ为所述梯度G的标准差,λ和γ均为预设的超参数。

基于梯度修改的黑盒对抗样本生成方法及装置

技术领域

[0001] 本发明涉及一种对抗样本生成方法,具体涉及一种基于梯度修改的黑盒对抗样本生成方法及装置。

背景技术

[0002] 现有研究表明,深度学习模型容易受到对抗样本的欺骗。在图像分类领域,对抗攻击就是对干净的图片即干净样本进行扰动,将其加上对抗噪声得到对抗样本。图1是干净样本与其对应的对抗样本的示意图。如图1所示,添加对抗噪声后的对抗样本与其对应的干净样本之间的差异是微小且不易被人眼察觉的。这样的对抗样本可以误导深度学习模型将其错误预测,导致其错误地通过恶意访问或拒绝正常访问。实际上,不论什么类型的深度学习模型,都会受到对抗样本的攻击。因此,带有深度学习模型的产品应用落地之前,有必要对其评估,确认其是否能够有效防御对抗攻击。所以,基于对抗攻击方法评估模型的鲁棒性变得至关重要。

[0003] 根据攻击方法的不同,现有的评估模型鲁棒性的方法可以分为白盒、灰盒、黑盒三种类型:

[0004] (1)白盒方法:待检测方提交了待检测模型,评估方拥有待检测模型的结构和参数,可以有针对性地生成对抗样本,观察待检测模型在白盒场景下鲁棒性表现;现有的白盒攻击方法包括FGSM、I-FGSM、PGD、L-BFGS、C&W、UAP、JSMA等。

[0005] (2)灰盒方法:待检测方提供了待检测模型的访问接口,评估方可以多次通过该接口得到待检测模型的输出,依靠这些信息生成相应的对抗样本,观察待检测模型在灰盒场景下的鲁棒性表现。现有的灰盒方法包括:使用有限的查询训练替代模型以改善目标模型的逼近的方法、基于进化策略的算法、使用差分进化的单像素黑盒攻击以实现隐蔽的方法等。

[0006] (3)黑盒方法:待检测方未提供目标模型的相关信息,评估方只能借助本地已知的代理模型生成对抗样本,观察待检测模型在黑盒场景下的鲁棒性表现。图2是黑盒攻击的原理示意图。如图2所示,代理模型是攻击者已知的本地模型,攻击者通过干净样本结合代理模型生成对抗噪声,并通过干净样本与对抗噪声的结合生成对抗样本,从而通过该对抗样本成功欺骗完全未知的目标模型。现有的黑盒方法包括:动量迭代方法、跳过梯度法、Nesterov迭代快速梯度符号方法、方差调整方法、梯度正则化方法等。

[0007] 然而,上述现有技术还存在如下缺点:

[0008] (1)白盒方法要求攻击者具有目标模型的全部知识,灰盒方法要求攻击者能够大量地访问目标模型并且得到模型输出。因此,实际中,白盒方法与灰盒方法均难以有效应用。

[0009] (2)执行效率略低,攻击成功率不高。过往的方法在进行梯度操作时,需要对梯度信息进行复杂的分析,需要运行的时间较长。

[0010] 综上,黑盒方法如何快速生成高攻击成功率的对抗样本还存在较大的提升空间。

发明内容

[0011] 本发明是为了解决上述问题而进行的,目的在于提供一种基于梯度修改的黑盒对抗样本生成方法及装置。

[0012] 本发明提供了一种基于梯度修改的黑盒对抗样本生成方法,用于根据代理模型$\tilde{F}$和干净样本x生成对应的最终对抗样本x',具有这样的特征,包括以下步骤:步骤S1,设置当前迭代数i为0、迭代总次数为N、噪声Δx_i为$\vec{0}$以及梯度g_i为$\vec{0}$;步骤S2,将干净样本x输入代理模型$\tilde{F}$,得到干净输出$\tilde{F}(x)$;步骤S3,将噪声Δx_i和干净样本x相加,得到对抗样本x_i;步骤S4,将对抗样本x_i输入代理模型$\tilde{F}$,得到对抗输出$\tilde{F}(x_i)$;步骤S5,根据干净输出$\tilde{F}(x)$和对抗输出$\tilde{F}(x_i)$计算得到损失$L(\Delta x_i)$;步骤S6,根据损失$L(\Delta x_i)$,通过反向传播计算得到更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$;步骤S7,根据更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$更新梯度$g_i$,得到梯度$g_{i+1}$;步骤S8,根据梯度$g_{i+1}$更新噪声$\Delta x_i$,得到噪声$\Delta x_{i+1}$;步骤S9,当前迭代数i加1,判断当前迭代数i是否等于迭代总次数,若是,则进入步骤S10,若否,则将梯度g_{i+1}作为梯度g_i并将噪声Δx_{i+1}作为噪声Δx_i,进入步骤S3;步骤S10,将噪声Δx_{i+1}与干净样本x相加,得到最终对抗样本x',其中,代理模型$\tilde{F}$通过对具有Transformer结构的神经网络模型插入梯度修改层构成,梯度修改层仅在反向传播中工作,根据输入的梯度G输出对应的梯度G_out,梯度G'的计算表达式为:G'=max(min(G,μ+λ*σ),μ-λ*σ),G_out=G'*γ,式中μ为梯度G的均值,σ为梯度G的标准差,λ和γ均为预设的超参数。

[0013] 在本发明提供的基于梯度修改的黑盒对抗样本生成方法中,还可以具有这样的特征:其中,具有Transformer结构的神经网络模型为ViT模型,ViT模型包含至少一个Transformer块,每个Transformer块包括注意力层、投影层和多层感知机,代理模型$\tilde{F}$的注意力层与投影层之间插入有梯度修改层,代理模型$\tilde{F}$的投影层与多层感知机之间插入有梯度修改层。

[0014] 在本发明提供的基于梯度修改的黑盒对抗样本生成方法中,还可以具有这样的特征:其中,对抗样本x_i的计算表达式为:$x_i=\Delta x_i+x$。

[0015] 在本发明提供的基于梯度修改的黑盒对抗样本生成方法中,还可以具有这样的特征:其中,损失$L(\Delta x_i)$的计算表达式为:$L(\Delta x_i)=$

[0016] $$CE(\tilde{F}(x),\tilde{F}(x_i))-\beta\cdot\|\Delta x_i\|_2,\ CE(\tilde{F}(x),\tilde{F}(x_i))=-\sum_{k=1}^{D}\tilde{F}(x)_k\cdot\log\tilde{F}(x_i)_k,$$

[0017] $$\|\Delta x_i\|_2=\sqrt{\sum_{m=1}^{H}(\Delta x_i)_m^2}$$,式中D为代理模型输出的维度,$\tilde{F}(x)_k$为经过代理模型$\tilde{F}$判断干净样本x属于第k类的预测值,$\tilde{F}(x_i)_k$为经过代理模型$\tilde{F}$判断第i轮迭代得到的对抗样本x_i属于第k类的预测值,H为输入样本的维度,$(\Delta x_i)_m$为输入样本维度m上的数值。

[0018] 在本发明提供的基于梯度修改的黑盒对抗样本生成方法中,还可以具有这样的特征:其中,梯度g_{i+1}的计算表达式为:$g_{i+1}=g_i+\alpha\cdot\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$,式中α为给定的超参数。

[0019] 在本发明提供的基于梯度修改的黑盒对抗样本生成方法中,还可以具有这样的特

征:其中,噪声Δx_{i+1}的计算表达式为:$\Delta x_{i+1}=\Delta x_i+\in/N\cdot sign(g_{i+1})$,式中∈为给定的超参数。

[0020] 本发明还提供了一种基于梯度修改的黑盒对抗样本生成装置,用于根据代理模型$\tilde{F}$和干净样本x生成对应的最终对抗样本x',具有这样的特征,包括:初始化模块、对抗样本生成模块、模型处理模块、损失计算模块、参数更新模块、梯度更新模块、噪声更新模块、判断模块和最终样本生成模块,其中,初始化模块用于设置当前迭代数i为0、迭代总次数为N、噪声Δx_i为$\vec{0}$以及梯度g_i为$\vec{0}$,对抗样本生成模块用于将噪声Δx_i和干净样本x相加,得到对抗样本x_i,模型处理模块存储有代理模型$\tilde{F}$,用于根据干净样本x生成干净输出$\tilde{F}(x)$,根据对抗样本x_i生成对抗输出$\tilde{F}(x_i)$,损失计算模块用于根据干净输出$\tilde{F}(x)$和对抗输出$\tilde{F}(x_i)$计算得到损失$L(\Delta x_i)$,参数更新模块用于根据损失$L(\Delta x_i)$,通过反向传播计算得到更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$,梯度更新模块用于根据更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$更新梯度$g_i$,得到梯度$g_{i+1}$,噪声更新模块用于根据梯度$g_{i+1}$更新噪声$\Delta x_i$,得到噪声$\Delta x_{i+1}$,判断模块用于将当前迭代数i加1,判断当前迭代数i是否等于迭代总次数,若是,则控制最终样本生成模块运行,若否,则将梯度g_{i+1}作为梯度g_i并将噪声Δx_{i+1}作为噪声Δx_i,控制对抗样本生成模块运行,最终样本生成模块用于将噪声Δx_{i+1}与干净样本x相加,得到最终对抗样本x',其中,代理模型$\tilde{F}$通过对具有Transformer结构的神经网络模型插入梯度修改层构成,梯度修改层仅在反向传播中工作,根据输入的梯度G输出对应的梯度G_out,梯度G'的计算表达式为:$G'=\max(\min(G,\mu+\lambda*\sigma),\mu-\lambda*\sigma)$,$G\_out=G'*\gamma$,式中μ为梯度G的均值,σ为梯度G的标准差,λ和γ均为预设的超参数。

[0021] 发明的作用与效果

[0022] 根据本发明所涉及的基于梯度修改的黑盒对抗样本生成方法及装置,因为,一方面,对具有Transformer结构的神经网络模型的Transformer块插入梯度修改层,使其在反向传播中进行梯度修改;另一方面,通过梯度修改层对输入的梯度依次进行梯度裁剪和梯度缩放,从而能够更快得到对抗样本,并使其具有更高的攻击成功率。所以,本发明的基于梯度修改的黑盒对抗样本生成方法及装置能够快速生成高攻击成功率的对抗样本。

附图说明

[0023] 图1是本发明中干净样本与其对应的对抗样本的示意图;

[0024] 图2是本发明中黑盒攻击的原理示意图;

[0025] 图3是本发明的实施例中黑盒对抗样本生成装置的框图;

[0026] 图4是本发明的实施例中ViT模型的Transformer块的结构示意图;

[0027] 图5是本发明的实施例中代理模型$\tilde{F}$的Transformer块的结构示意图;

[0028] 图6是本发明的实施例中基于梯度修改的黑盒对抗样本生成方法的流程示意图。

具体实施方式

[0029] 为了使本发明实现的技术手段、创作特征、达成目的与功效易于明白了解,以下实

施例结合附图对本发明基于梯度修改的黑盒对抗样本生成方法及装置作具体阐述。

[0030] 本实施例中提供一种基于梯度修改的黑盒对抗样本生成装置,用于根据代理模型$\tilde{F}$和干净样本x生成对应的最终对抗样本x',该最终对抗样本x'用于金融领域中基于图像分类的深度学习模型的鲁棒性评估。

[0031] 图3是本发明的实施例中黑盒对抗样本生成装置的框图。

[0032] 如图3所示,黑盒对抗样本生成装置100包括初始化模块11、对抗样本生成模块12、模型处理模块13、损失计算模块14、参数更新模块15、梯度更新模块16、噪声更新模块17、判断模块18、最终样本生成模块19以及控制上述各模块运行的控制模块20。

[0033] 初始化模块11用于设置当前迭代数i为0、迭代总次数为N、噪声Δx_i为$\vec{0}$以及梯度g_i为$\vec{0}$。

[0034] 对抗样本生成模块12用于将噪声Δx_i和干净样本x相加,得到对抗样本x_i。其中,对抗样本x_i的计算表达式为:$x_i = \Delta x_i + x$。

[0035] 模型处理模块13存储有代理模型$\tilde{F}$,用于根据干净样本x生成干净输出$\tilde{F}(x)$,根据对抗样本x_i生成对抗输出$\tilde{F}(x_i)$。

[0036] 其中,代理模型$\tilde{F}$通过对具有Transformer结构的神经网络模型插入梯度修改层构成,梯度修改层仅在反向传播中工作,根据输入的梯度G输出对应的梯度G_out,梯度G'的计算表达式为:

[0037] $G' = \max(\min(G, \mu+\lambda*\sigma), \mu-\lambda*\sigma)$,

[0038] $G\_out = G*\gamma$,

[0039] 式中μ为梯度G的均值,σ为梯度G的标准差,λ和γ均为预设的超参数。本实施例中梯度修改层先对输入的梯度G进行梯度裁剪得到梯度G',再对梯度G'进行梯度缩放得到梯度G_out作为输出,从而实现对代理模型的优化,从而提高执行效率和攻击成功率。

[0040] 本实施例中具有Transformer结构的神经网络模型为ViT模型,ViT模型包含至少一个Transformer块。

[0041] 图4是本发明的实施例中ViT模型的Transformer块的结构示意图。

[0042] 如图4所示,ViT模型的Transformer块200包括注意力层201、投影层202和多层感知机203,Transformer块200的输入依次经由注意力层201、投影层202和多层感知机203处理,得到对应的输出。

[0043] 图5是本发明的实施例中代理模型$\tilde{F}$的Transformer块的结构示意图。

[0044] 如图5所示,代理模型$\tilde{F}$的Transformer块300包括注意力层301、投影层302、多层感知机303、第一梯度修改层304和第二梯度修改层305。模型前向输出过程中,Transformer块300的输入依次经由注意力层301、投影层302和多层感知机303处理得到对应的输出,即该过程中第一梯度修改层304和第二梯度修改层305不起任何作用,Transformer块300的输出与Transformer块200一致。反向传播中,依次经由多层感知机303、第二梯度修改层305、投影层302、第一梯度修改层304和注意力层301进行反向传播计算,该过程中第一梯度修改层304和第二梯度修改层305用于梯度修改。

[0045] 现有的梯度裁剪方法的具体为:在执行反向传播算法结束后,裁剪限制对应参数

的梯度数值,即在反向传播结束后仅对损失函数中的自变量的梯度进行一次裁剪。本实施例的梯度修改层则设置于Transformer块内,在反向传播过程中对损失函数中隐含的中间变量进行多次梯度修改,在一次反向传播中进行了多次梯度修改,从而实现对变量的多层次微调。

[0046] 损失计算模块14用于根据干净输出$\tilde{F}(x)$和对抗输出$\tilde{F}(x_i)$计算得到损失$L(\Delta x_i)$。

[0047] 其中,损失$L(\Delta x_i)$的计算表达式为:

[0048] $$L(\Delta x_i) = CE(\tilde{F}(x), \tilde{F}(x_i)) - \beta \cdot \|\Delta x_i\|_2,$$

[0049] $$CE(\tilde{F}(x), \tilde{F}(x_i)) = -\sum_{k=1}^{D} \tilde{F}(x)_k \cdot \log \tilde{F}(x_i)_k,$$

[0050] $$\|\Delta x_i\|_2 = \sqrt{\sum_{m=1}^{H} (\Delta x_i)_m^2},$$

[0051] 式中D为代理模型输出的维度,$\tilde{F}(x)_k$为经过代理模型$\tilde{F}$判断干净样本x属于第k类的预测值,$\tilde{F}(x_i)_k$为经过代理模型$\tilde{F}$判断第i轮迭代得到的对抗样本x_i属于第k类的预测值,H为输入样本的维度,$(\Delta x_i)_m$为输入样本维度m上的数值。本实施例中代理模型$\tilde{F}$输出的结果为1000种分类的预测值时,则维度D=1000。输入样本为通道数、宽度和高度依次为3、224、224的RGB图片时,则维度H=3*224*224。

[0052] 参数更新模块15用于根据损失$L(\Delta x_i)$,通过反向传播计算得到更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$。

[0053] 梯度更新模块16用于根据更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$更新梯度$g_i$,得到梯度$g_{i+1}$。

[0054] 其中,梯度g_{i+1}的计算表达式为:

[0055] $$g_{i+1} = g_i + \alpha \cdot \frac{\partial L(\Delta x_i)}{\partial \Delta x_i},$$

[0056] 式中α为给定的超参数。

[0057] 噪声更新模块17用于根据梯度g_{i+1}更新噪声Δx_i,得到噪声Δx_{i+1}。

[0058] 其中,噪声Δx_{i+1}的计算表达式为:

[0059] $$\Delta x_{i+1} = \Delta x_i + \in / N \cdot sign(g_{i+1}),$$

[0060] 式中∈为给定的超参数,其代表噪声预算,sign为符号函数,只取向量或矩阵的方向。本实施例中$sign(g_{i+1})$为带符号的向量,模长为1。

[0061] 判断模块18用于将当前迭代数i加1,判断当前迭代数i是否等于迭代总次数,若是,则控制最终样本生成模块19运行,若否,则将梯度g_{i+1}作为梯度g_i并将噪声Δx_{i+1}作为噪声Δx_i,控制对抗样本生成模块12运行。

[0062] 最终样本生成模块19用于将噪声Δx_{i+1}与干净样本x相加,得到最终对抗样本x'。

[0063] 控制模块20存储有控制各个模块运行的控制程序。

[0064] 以下结合附图,对采用本实施例的黑盒对抗样本生成装置100进行基于梯度修改的黑盒对抗样本生成方法的过程进行说明。

[0065] 图6是本发明的实施例中基于梯度修改的黑盒对抗样本生成方法的流程示意图。

[0066] 如图6所示,本实施例的基于梯度修改的黑盒对抗样本生成方法包括以下步骤:

[0067] 步骤S1,采用初始化模块11设置当前迭代数i为0、迭代总次数为N、噪声Δx_i为$\vec{0}$以及梯度g_i为$\vec{0}$。

[0068] 步骤S2,采用模型处理模块13根据干净样本x得到干净输出$\tilde{F}(x)$。

[0069] 步骤S3,采用对抗样本生成模块12将噪声Δx_i和干净样本x相加,得到对抗样本x_i。

[0070] 步骤S4,采用模型处理模块13根据对抗样本x_i得到对抗输出

[0071] $\tilde{F}(x_i)$。

[0072] 步骤S5,采用损失计算模块14根据干净输出$\tilde{F}(x)$和对抗输出$\tilde{F}(x_i)$计算得到损失L(Δx_i)。

[0073] 步骤S6,采用参数更新模块15根据损失L(Δx_i)通过反向传播计算得到更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$。

[0074] 步骤S7,采用梯度更新模块16根据更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$更新梯度$g_i$,得到梯度$g_{i+1}$。

[0075] 步骤S8,采用噪声更新模块17根据梯度g_{i+1}更新噪声Δx_i,得到噪声Δx_{i+1}。

[0076] 步骤S9,采用判断模块18将当前迭代数i加1,判断当前迭代数i是否等于迭代总次数,若是,则进入步骤S10,若否,则将梯度g_{i+1}作为梯度g_i并将噪声Δx_{i+1}作为噪声Δx_i,进入步骤S3。

[0077] 步骤S10,采用最终样本生成模块19将噪声Δx_{i+1}与干净样本x相加,得到最终对抗样本x'。

[0078] 本实施例中将基于梯度修改的黑盒对抗样本生成方法即本方法与现有的基础方法即MIM方法以及现有性能最好的TGR方法进行对比测试,得到在不同代理模型上生成的对抗样本对不同目标模型的攻击成功率如下所示:

[0079]

代理模型	方法	ViT-B/16	PiT-B	CaiT-S/24	Visformer-S	DeiT-B	TNT-S	LeViT-256	ConViT-B
ViT-B/16	MIM	99.3	41.5	71.0	42.1	69.3	56.4	39.4	71.5
	TGR	99.2	51.4	83.0	56.5	83.5	73.9	60.6	83.7
	本方法	99.6	54.9	86.6	61.0	87.5	78.4	64.4	87.5
PiT-B	MIM	27.0	99.0	39.5	47.6	38.2	47.1	39.6	40.3
	TGR	65.8	100.0	82.9	88.5	83.2	91.2	87.8	82.4
	本方法	67.4	100.0	83.7	89.8	85.4	92.1	87.3	83.9
CaiT-S/24	MIM	66.8	52.5	96.5	55.7	80.2	72.7	54.9	80.6
	TGR	87.1	70.4	100.0	81.3	98.8	94.6	82.7	97.9
	本方法	90.2	72.6	100.0	82.3	99.6	95.9	84.3	98.7

[0080] 上表中第一列为各个代理模型,第二列为采用的各个方法,第三列至第十列依次为以ViT-B/16模型、PiT-B模型、CaiT-S/24模型、Visformer-S模型、DeiT-B模型、TNT-S模型、LeViT-256模型和ConViT-B模型作为目标模型时各个不同方法在不同代理模型上训练的对抗样本的攻击成功率。上述代理模型和目标模型均为具有Transformer结构的神经网

络模型。由此可见,本方法相较于现有的两种方法,具有更优的攻击成功率。

[0081] 本实施例中将本方法与上述现有方法进行跨模型攻击性能对比测试,得到在不同代理模型上生成的对抗样本对不同目标模型的攻击成功率如下所示:

[0082]

代理模型	方法	Inc-v3	Inc-v4	IncRes-v2	Res-v2	Inc-v3_ens3	Inc-v3_ens4	IncRes-v2_adv
ViT-B/16	MIM	33.6	31.3	27.6	32.0	23.8	22.5	17.8
	TGR	50.5	44.1	39.2	45.4	34.4	31.6	28.9
	本专利	53.8	48.5	43.7	47.8	37.8	34.9	31.0
PiT-B	MIM	35.5	32.3	28.5	31.4	19.7	17.0	14.8
	TGR	80.0	73.5	69.3	71.9	51.1	51.5	40.5
	本专利	78.0	74.3	70.4	71.4	51.7	50.2	41.7
CaiT-	MIM	46.7	43.9	40.1	43.8	32.7	29.2	26.5

[0083]

S/24	TGR	68.6	61.2	59.4	62.8	49.1	47.1	38.3
	本专利	69.5	63.3	62.3	64.7	51.6	49.9	40.2

[0084] 上表中第一列为各个代理模型,第二列为采用的各个方法,第三列至第九列依次为以Inc-v3模型、Inc-v4模型、IncRes-v2模型、Res-v2模型、Inc-v3_ens3模型、Inc-v3_ens4模型和IncRes-v2_adv模型作为目标模型时各个不同方法在不同代理模型上训练的对抗样本的攻击成功率。上述目标模型均为不具有Transformer结构的神经网络模型。由此可见,本方法相较于现有的两种方法,在跨模型攻击上同样具有更优的攻击成功率。

[0085] 此外,本方法与TGR方法在相同条件下进行1000个批次的处理,本方法平均每个批次耗时为0.63s,而TGR方法平均每个批次耗时为1.59s,即本方法在运行效率上也优于TGR方法。

[0086] 本实施例中MIM方法在不同代理模型上生成的对抗样本对不同目标模型的攻击成功率可以视为采用现有的梯度裁剪方法处理后的攻击成功率,则可见,通过设置梯度修改层后即本方法的攻击成功率,相较于MIM方法的攻击成功率有较大的提升,可见本方法相较于现有的梯度裁剪方法具有更好的对抗样本生成效果。

[0087] 综上所述,本方法相较于现有的对抗样本生成方法,能够更加快速的生成对抗样本,且该对抗样本具有更高的攻击成功率。因此,因此本方法能够进一步用于对模型的鲁棒性进行测试,提高测试结果的准确度。

[0088] 实施例的作用与效果

[0089] 根据本实施例所涉及的基于梯度修改的黑盒对抗样本生成方法及装置,一方面,对具有Transformer结构的神经网络模型的Transformer块插入梯度修改层,使其在反向传播中进行梯度修改;另一方面,通过梯度修改层对输入的梯度依次进行梯度裁剪和梯度缩放,从而能够更快得到对抗样本,并使其具有更高的攻击成功率。总之,本方法能够快速生成高攻击成功率的对抗样本。

[0090] 本行业的技术人员应该了解,本发明不受上述实施例的限制,上述实施例和说明书中描述的只是说明本发明的原理,在不脱离本发明精神和范围的前提下,本发明还会有各种变化和改进,这些变化和改进都落入要求保护的本发明范围内。本发明要求保护范围

由所附的权利要求书及其等效物界定。

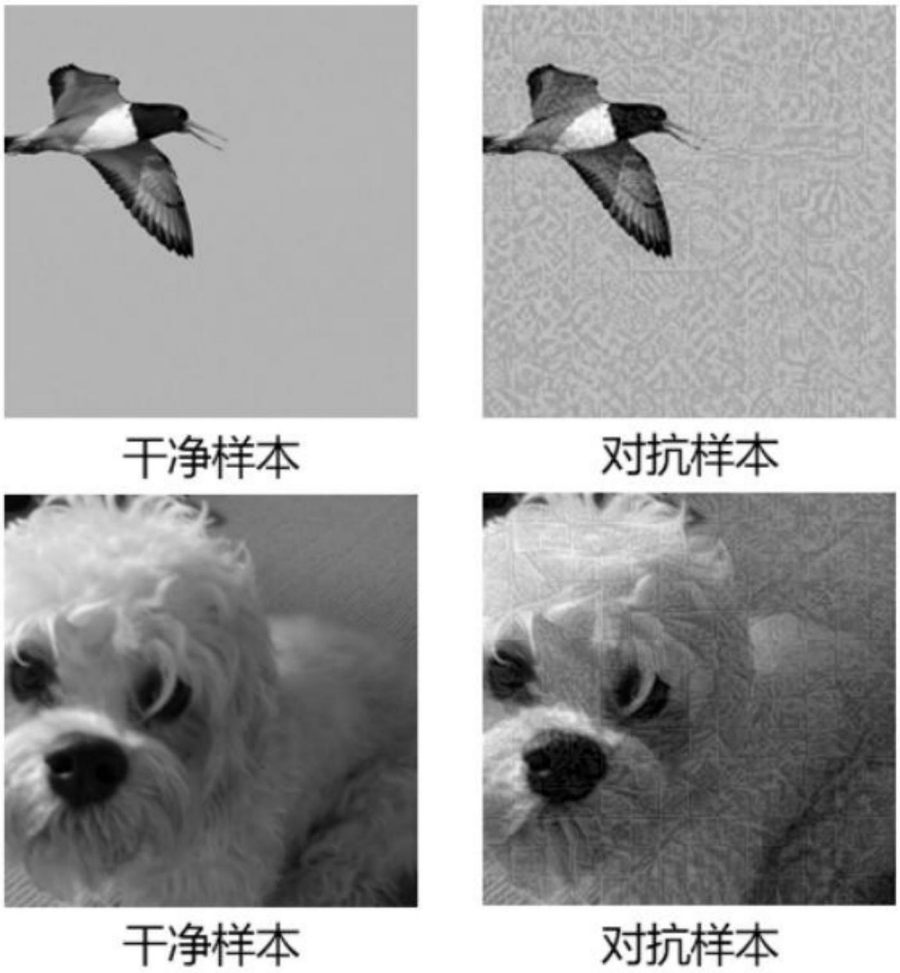


图1

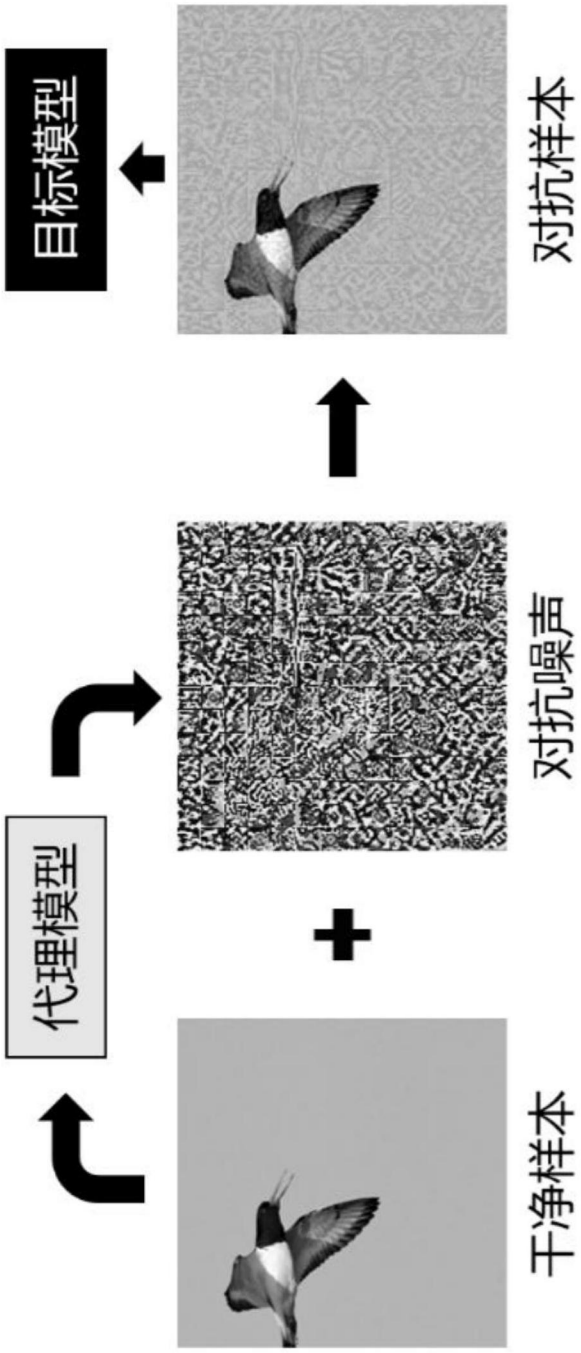


图2

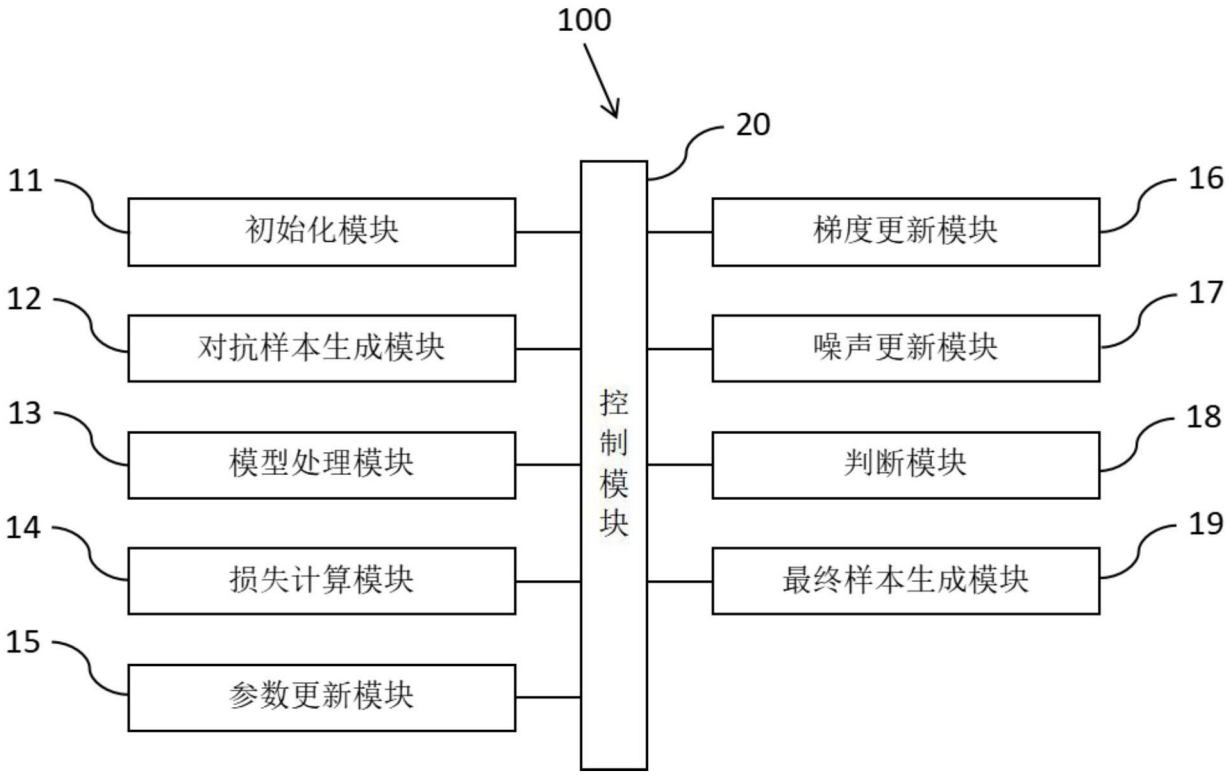

100
20
11
初始化模块
12
对抗样本生成模块
13
模型处理模块
14
损失计算模块
15
参数更新模块
控制模块
16
梯度更新模块
17
噪声更新模块
18
判断模块
19
最终样本生成模块

图3

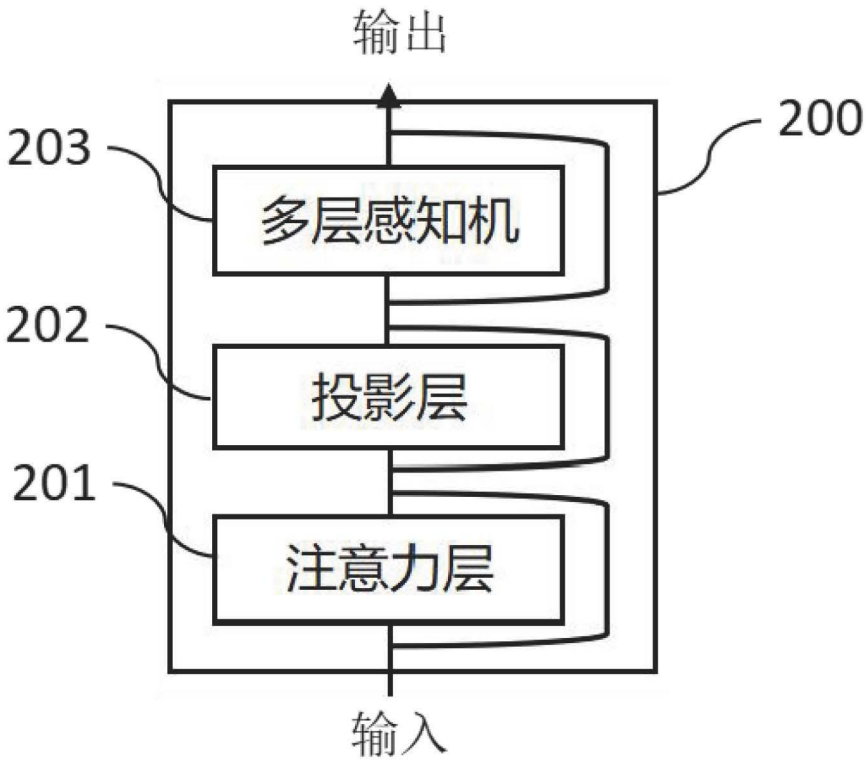

输出
200
203
多层感知机
202
投影层
201
注意力层
输入

图4

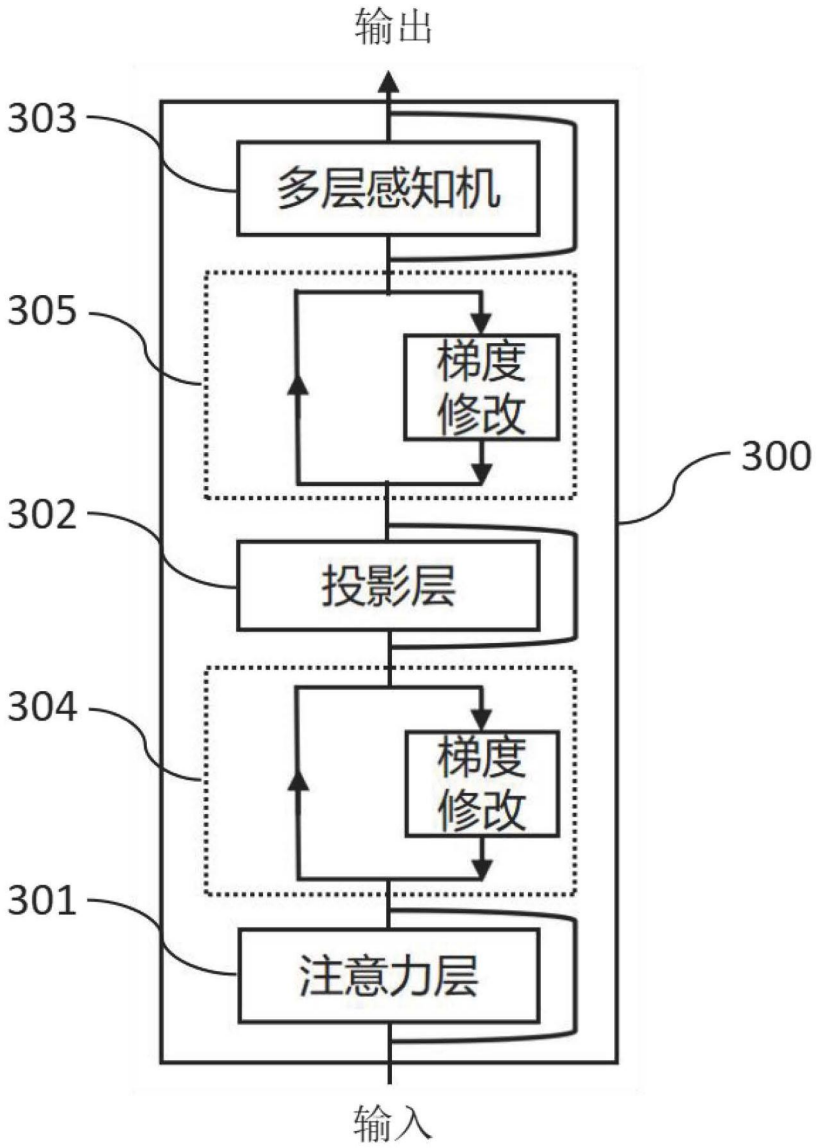


图5

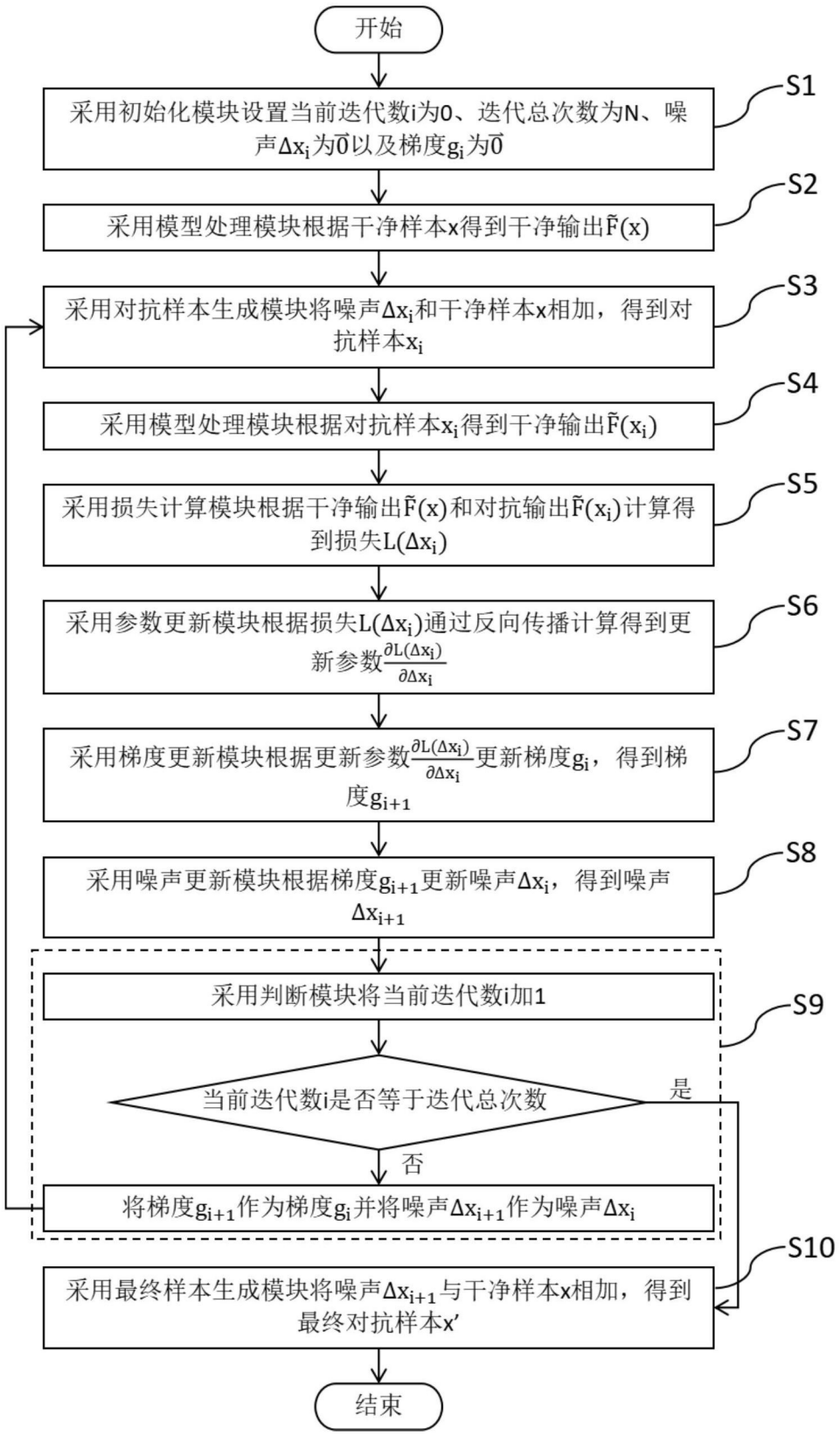

开始
采用初始化模块设置当前迭代数i为0、迭代总次数为N、噪声Δx_i为$\vec{0}$以及梯度g_i为$\vec{0}$
S1
采用模型处理模块根据干净样本x得到干净输出$\tilde{F}(x)$
S2
采用对抗样本生成模块将噪声Δx_i和干净样本x相加，得到对抗样本x_i
S3
采用模型处理模块根据对抗样本x_i得到干净输出$\tilde{F}(x_i)$
S4
采用损失计算模块根据干净输出$\tilde{F}(x)$和对抗输出$\tilde{F}(x_i)$计算得到损失$L(\Delta x_i)$
S5
采用参数更新模块根据损失$L(\Delta x_i)$通过反向传播计算得到更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$
S6
采用梯度更新模块根据更新参数$\frac{\partial L(\Delta x_i)}{\partial \Delta x_i}$更新梯度$g_i$，得到梯度$g_{i+1}$
S7
采用噪声更新模块根据梯度g_{i+1}更新噪声Δx_i，得到噪声Δx_{i+1}
S8
采用判断模块将当前迭代数i加1
当前迭代数i是否等于迭代总次数
是
否
S9
将梯度g_{i+1}作为梯度g_i并将噪声Δx_{i+1}作为噪声Δx_i
采用最终样本生成模块将噪声Δx_{i+1}与干净样本x相加，得到最终对抗样本x′
S10
结束

图6