西安电子科技大学学报
Journal of Xidian University
ISSN 1001-2400,CN 61-1076/TN

《西安电子科技大学学报》网络首发论文

题目：针对 ASR 系统的快速有目标自适应对抗攻击
作者：张树栋，高海昌，曹曦文，康帅
收稿日期：2020-08-14
网络首发日期：2020-10-22
引用格式：张树栋，高海昌，曹曦文，康帅．针对 ASR 系统的快速有目标自适应对抗攻击．西安电子科技大学学报.
https://kns.cnki.net/kcms/detail/61.1076.tn.20201022.1106.002.html

针对 ASR 系统的快速有目标自适应对抗攻击

张树栋，高海昌，曹曦文，康帅

（西安电子科技大学 计算机科学与技术学院，陕西 西安 710071）

摘要：对抗样本是一种恶意输入，通过在输入中添加人眼无法察觉的微小扰动来误导深度学习模型产生错误的输出。近年来，随着对抗样本研究的发展，除了大量图像领域的对抗样本工作，在自动语音识别领域也开始有一些新进展。目前，针对自动语音识别系统的最先进的对抗攻击来自 Carlini&Wagner，其方法是通过获得使模型被错误分类的最小扰动来生成成功的对抗样本。因为这种方法需要同时优化两个损失函数项，通常需要进行数千次迭代，效率低下，因此提出了 A-FTA 方法。该方法通过最大化自动语音识别模型关于对抗样本的预测和目标短语之间的相似度来快速生成对抗样本，并且在攻击过程中根据是否攻击成功自适应地调整扰动大小，从而生成较小扰动的对抗样本。实验结果表明，这种方法相比于目前最先进的方法，用更少的迭代次数取得了更好的攻击结果，在高效的同时依然能保证很高的攻击成功率。

关键词：深度神经网络；对抗样本；语音识别；机器学习；对抗攻击

中图分类号：TN912.34;TP183 **文献标识码**：A

Adaptive fast and targeted adversarial attack for speech recognition

ZHANG Shudong, GAO Haichang, CAO Xiwen , KANG Shuai

(School of Computer Science and Technology, Xidian University, Xi'an 710071, China)

Abstract: Adversarial examples are malicious inputs designed to induce deep learning models to produce erroneous outputs, which make humans imperceptible by adding small perturbations to the input. Most research on adversarial examples is in the domain of image. Recently, studies on adversarial examples have expanded into the automatic speech recognition domain. The state-of-art attack on the ASR system comes from C&W, which aims to obtain the minimum perturbation that causes the model to be misclassified. However, this method is inefficient since it requires the optimization of two terms of loss functions at the same time, and usually requires thousands of iterations. In this paper, we propose an efficient approach based on strategies that maximize the likelihood of adversarial examples and target categories. A large number of experiments show that our attack achieves better results with fewer iterations.

Key Words: Deep Neural Networks(DNN); Adversarial examples; Speech Recognition; Machine Learning; Adversarial attacks

收稿日期： 2020-08-14
基金项目：国家自然科学基金（61972306）
作者简介：张树栋（1993-），男，西安电子科技大学博士研究生，E-mail：sdong_zhang@163.com.
通信作者：高海昌（1978-），男，教授，博士，E-mail：hchgao@xidian.edu.cn.

神经网络的快速发展，使其应用于多种领域，如自动驾驶，面部识别，目标检测，语音识别和图像分类等。但是，最近的研究[1-3]已经表明神经网络容易受到对抗样本的影响。攻击者可以通过在输入中添加一些人类不容易感知的微小扰动，使得神经网络输出任何想要的结果。对抗样本的发现对深度神经网络在现实生活中的应用构成了极大的安全威胁。攻击者可以通过生成对抗样本[4]来欺骗面部识别系统，入侵移动设备获取相关的隐私信息；或是对道路标识符进行篡改，促使自动驾驶汽车将右转弯的交通标志识别为笔直，由此引发交通事故[5]。

对抗样本的研究最初主要集中在图像空间上，除了少部分目标检测[6]，语义分割[7]，人脸识别[8]和强化学习[9]的研究内容外，大部分都是针对图像分类任务[1-4,10]。在其他领域，对抗样本也有相应的研究，如：文本分类[11]，恶意软件检测[12]和语音识别[13]等。本文着重于语音识别领域的对抗样本攻击研究。

通常，根据攻击者攻击目标的不同，对抗攻击可以分为两种类型。一种常见的攻击方式是找到使目标模型分类错误的最小扰动。第二，在最大允许扰动范围内，最大化目标模型将扰动样本分类为目标类别的概率。最近，在获得具有最小扰动的对抗样本的攻击下，Carlini 等人[14]证明，对于任何音频样本，仅添加少量扰动就可以使自动语音识别模型将音频转录为攻击者指定的任意短语。尽管这种攻击产生的样本具有较低的噪声，但它需要进行大量的迭代，这对于实际场景中的自动语音识别攻击是不切实际的。

为了产生使对抗扰动范数值最小的攻击，需要优化两个目标，即在模型对输入进行了错误分类的同时还要保证尽量小的添加的扰动范数值。当前最先进的方法 C&W [7]通过设计使用两个损失函数项来解决此问题，该方法攻击效果的好坏取决于平衡两个损失函数的超参数 c 的选择，在此基础上，还需要通过大量迭代来实现攻击。笔者提出了一种 A-FTA 方法，该方法基于最大化对抗样本和目标类别相似度的策略。具体来说，使用投影梯度下降法来生成目标音频对抗样本。在每次迭代中，针对每个音频值在损失函数中进行梯度下降操作，以使损失函数最小化，同时根据样本是否具有对抗性来改变最大扰动范数值。攻击的步长则随着迭代次数的增加而逐渐减小。最后，将生成的扰动音频映射到固定的可行解空间中。所提出的方法可以大大减少攻击所需的迭代次数，并且还可以避免由于扰动范数较小而导致最优解在两点之间振荡问题。大量实验表明，在 300 次迭代中，A-FTA 方法的攻击效果要优于现有的方法。即使仅进行 100 次迭代，依然能保证非常高的攻击成功率。

1.背景及相关工作

1.1 对抗样本

给定模型 $f(\cdot)$ 和输入样本 $x \in \mathbb{R}^n$，其对应的标签 $y \in \mathbb{R}^m$。存在微小扰动 δ，使得 $x' = x + \delta$ 在某个距离度量 $d(x, x')$ 中与 x 相似，但分类结果 $f(x') \neq y$。这样的样本 x' 称为无目标对抗样本。除此之外，还有一种更强大的攻击，称为有目标对抗样本，它不仅可以使目标模型针对 x 和 x' 输出不同的结果，而且还可以使模型对输入样本 x' 误分类为特定的标签 t（由攻击者选择），即 $y \neq f(x') = t$。在图像领域，研究人员选择 l_p 距离作为 $d(x, x')$，最常使用 l_∞ 距离来评估添加的扰动的大小。根据攻击者是否对分类器 $f(\cdot)$ 的参数和结构有足够的了解，对抗样本的攻击方法可以进一步分为白盒攻击和黑盒攻击。在白盒攻击中，攻击者知道分类器 $f(\cdot)$ 的所有知识。在黑盒攻击中，除了输入和输出外，攻击者对分类器 $f(\cdot)$ 一无所知。

为了生成对抗样本，攻击者通常需要预先设置损失函数，然后通过执行基于梯度的优化过程来最小化

损失函数。根据不同的目标，生成对抗样本的方法可以表示为在成功攻击的前提下保证扰动$\|\delta\|$最小，或者在最大扰动预算范围$\|\delta\| \le \epsilon$中最大化对抗样本被分类为目标类别的概率。例如，获得具有最小失真的目标对抗样本的优化过程可以表示为：

$$min\ d(x, x+\delta),\ s.t.\ f(x+\delta) = t,\ x+\delta \in [0, M] \tag{1}$$

，其中[0,M]是有效输入的阈值范围。但是，由于约束$f(x+\delta) = t$是高度非线性的，因此现有的基于梯度的算法很难直接求解上述公式1。因此，研究人员将其表达为更适合优化的另一种形式[17]：

$$min\ d(x, x+\delta) + c \cdot l(f(x+\delta), t),\ s.t.\ x+\delta \in [0, M] \tag{2}$$

其中$c > 0$是平衡攻击成功率和扰动δ大小的超参数。

还有另一个更简单的目标函数，该函数最小化模型关于对抗样本的预测与目标标签之间的差异。与最小扰动优化不同，基于 p 范数的扰动固定为小于指定的阈值。对应的优化公式为

$$min\ l(f(x+\delta), t),\ s.t.\ \|\delta\| \le \epsilon \tag{3}$$

，其中ϵ是预设的添加的最大扰动值。与前面的公式 2 相比，此优化更为简单有效，因为它不需要搜索其他超参数c。

1.2 音频对抗样本

Cisse M 等人[4]成功生成了语音对抗样本导致 Google 的语音应用模型对输入的音频进行了错误的转录。2018 年，针对 Mozilla 对于 DeepSpeech 端到端的实现[5]，Carlini 等人[14]使用基于优化的最小扰动白盒攻击方法来构建目标音频对抗样本，生成的样本作为输入可以被模型转录为他们想要的任何目标短语。Schonherr 等人[6]通过使用“心理声学掩膜”对 Kaldi 上的 ASR 系统进行了人耳难以察觉的攻击。上述方法都直接将 wav 音频文件输入到模型中，而这在实际应用中是不现实的。Yuan 等人[7]提出“ CommanderSong”的方法来评估 Kaldi 模型，并使用歌曲作为载体来构建对抗攻击，所生成的对抗样本在空气传播中也同样有效。Yakura 等人[8]生成了可以无线播放的对抗样本。该攻击对短的两个单词或三个单词的短语非常有效，但对较长的句子短语则没有什么效果。Qin 等人[9]将基于优化的最小扰动攻击与听觉掩盖的心理声学原理[10]结合，并针对 Lingvo ASR 系统[11]生成了难以察觉的音频对抗样本。Liu X 等人[12]提出了加权采样音频对抗攻击方法，该方法可以在几分钟内生成低扰动和高鲁棒性的音频样本。Li J 等人[13]提出了对抗音乐的方法，并成功地欺骗了 Amazon Alexa 唤醒词检测系统。

1.3 威胁模型

针对音频领域中自动语音识别任务进行有目标音频对抗攻击。给定一个音频输入 x，目标是生成一个听起来类似于x的新音频$x' = x + \delta$，但是$f(x') = t$。仅当目标模型预测的短语与攻击者选择的目标短语完全匹配时，攻击才会成功。选择攻击 DeepSpeech 模型，这是由 Mozilla 使用 TensorFlow 实现的开源语音文本引擎。该模型使用 Mel 频率倒谱（MFC）转换作为输入的预处理，然后是递归神经网络（RNN）使用 LSTM[14]将音频波形映射到单个字符上的概率分布序列。与大多数以前的工作中使用的威胁模型一样，笔者假设白盒攻击设置，其中攻击者可以完全访问模型，并且知道模型的所有参数和体系结构。此威胁模型用于评估最坏情况下的系统安全性。

2.方法

2.1 动机

目前，针对 ASR 系统的最先进的音频对抗攻击来自 Carlini 等人[14]，他们通过求解公式 2 来获得最小扰动，成功地攻击了 DeepSpeech 模型。该方法以可微的方式实现了 MFC 的预处理，并将l设置为 Connectionist Temporal 分类（CTC）损失[15]，然后在整个音频输入上对其进行优化来获得对抗样本。找到最接近的对抗样本是困难的，因此有必要找到一个合适的参数c来共同优化分类项$l_{ctc}\left(f\left(x+\delta\right),t\right)$和扰动$\delta$的范数。在约束优化的一般情况下，这种基于惩罚的方法是众所周知的一般准则。解决公式 2 中最优化问题的主要困难是如何找到合适的参数c来平衡扰动δ和分类损失l_{ctc}。如果c太小，则增加的对抗扰动将非常小，但是生成的样本可能不是具有对抗性的。如果c太大，攻击总是会成功，但是添加的对抗扰动不是最佳的。此外，惩罚方法通常会导致收敛缓慢。在图像领域，Carlini 等人使用改进的二分搜索来选择c。这种方法可以找到最佳的c，但以降低攻击效率为代价。在音频领域中，c设置为固定常数。尽管它可以节省二分搜索带来的额外计算量，但由于针对 ASR 对抗攻击的内在复杂性，该方法仍然效率不高。此外，常数c的选择对于这种攻击的成功至关重要。

2.2 损失函数

根据在图像领域中生成对抗样本的经验和公式 3 的定义，很容易找到在固定扰动范围内最差的对抗样本。在公式 3 中，两个约束都可以用δ表示，并且可以使用投影梯度下降法（PGD）来优化所得的公式。笔者依照 Carlini 的设置，并将损失函数 l 设置为 CTC 损失。最终，优化公式为

$$min\ l_{ctc}\left(f\left(x+\delta\right),t\right),\ s.t.\left\|\delta\right\|\le\epsilon \tag{4}$$

可以使用标准的 PGD 方法来解决公式 4 中的优化问题，来构造目标音频对抗样本。 具体地，在每次迭代中，该方法基于攻击者选择的输出短语，对每个音频值在损失函数中执行梯度下降步骤，以使损失函数最小化。接下来，再将产生的扰动音频投影到可行解空间内（在每个原始音频预设的最大扰动范围ϵ内）。迭代过程可以表示为:

$$x^{i+1}=C_{lip}\left\{x^{i}-\alpha\cdot sign\left(\nabla_{x}l_{ctc}\left(f\left(x^{i}\right),t\right)\right)\right\} \tag{5}$$

其中，α是步长，x_i是第i次迭代生成的对抗样本。

2.3 自适应的快速有目标攻击

尽管标准的 PGD 方法可以解决公式 4 的优化问题，但发现生成的扰动大小和攻击成功率受到最大扰动预算和步长α的限制，即使对于凸优化问题，也不能保证去得到一个最优解。如果扰动预算太大，则产生的扰动也将相应较大，使得人耳很容易察觉。如果扰动预算过小，则因为固定步长的设置可能导致梯度下降法在两点之间振荡，优化将不会收敛。

鉴于上述的缺点，提出了一种自适应快速有目标攻击（A-FTA）方法来解决这两个问题。首先，根据目标模型的预测结果动态更改 A-FTA 方法的最大扰动范数值，而不是像标准 PGD 这样使用一个预设的固定扰动约束。具体地，通过在原始音频x周围的最大扰动范围内投影对抗性扰动δ来约束规范。然后，根据二分决策的结果对范数约束值进行修改。如果样本x_i在步骤i中不是对抗样本，则增加步骤$i+1$下的约束值，否则会相应减少。其次，在优化过程中，动态调整步长，而不是固定不变。具体来说，在迭代开始时，将初始化较大的步长α，并且通过扰动范数约束ϵ的动态转换，可以快速找到成功的对抗样本。接着，

随着攻击的不断迭代，更多的对抗样本成功生成，扰动范数约束值ϵ和步长α也在动态调节下变得越来越小。最终，可以成功生成具有微小扰动的对抗样本。

算法1中描述了完整的攻击过程。从原始音频输入x开始，并根据目标模型的预测结果迭代更新最大干扰参数ϵ。在第i次迭代中，如果x_i仍然不是对抗样本，则通过公式$\epsilon_{i+1}=(1+\lambda)\times\epsilon_i$来增加扰动约束$\epsilon$；如果$x_i$已经是对抗样本了，则通$x_0\leftarrow x,\epsilon_0\leftarrow\epsilon,\alpha_i\leftarrow\alpha$过公式：$\epsilon_{i+1}=(1-\lambda)\times\epsilon_i$来减小$\epsilon$的值。在两种情况下，都从点$x_i$开始来进行公式5的迭代操作，将扰动音频投影到可行解空间中，并得到x_{i+1}。然后，通过余弦退火的方法减小步长α。

算法1 A-FTA 攻击

输入: 音频样本x，扰动预算ϵ，迭代次数K，步长α，目标短语t，余弦下降函数$cosine(\cdot)$，调整扰动范数的超参数λ，ASR模型$f(\cdot)$，损失函数$l_{ctc}(x,y)$.

输出: 音频对抗样本x'.

Initialize

For $i=0$ **to** $K-1$ **do**

$grad\leftarrow\nabla_{x_i}(l_{ctc}(x_i,t))$

$x_{i+1}\leftarrow x_i-\alpha_i\cdot sign(grad)$

$x_{i+1}\leftarrow clip(x_{i+1},x_i-\epsilon_i,x_i+\epsilon_i)$

$x_{i+1}\leftarrow clip(x_{i+1},-2^{15},2^{15}-1)$

If $f(x_{i+1})==t$ **then**

$\epsilon_{i+1}\leftarrow(1-\lambda)\times\epsilon_i$

else

$\epsilon_{i+1}\leftarrow(1+\lambda)\times\epsilon_i$

end if

$\alpha_{i+1}\leftarrow cosine(\alpha_i)$

End for

返回 具有最小扰动的对抗样本x_i

3 实验

3.1 数据集和评估指标

数据集: 在实验中使用两个开源的语音数据集，Mozilla Common Voice (MCV)和LibriSpeech (LS)。对于

MCV 数据集，选择干净测试集的前 100 个音频样本以生成目标音频对抗样本，并从测试集标签中随机选择 100 个不正确的转录文本作为目标转录文本。对于 LS 数据集，从测试集中选择 50 个样本，并从测试集标签中随机选择 50 个不正确的转录文本作为目标转录文本。确保所选择的目标转录文本长度不会超过目标音频转录长度的最大值。为了保证攻击的多样性，根据音频帧的长度和目标文本的长度将攻击分为四类，即短到短，短到长，长到短，长到长。 所有音频样本均以 16Khz 采样。表 1 中展示了原始短语和目标短语的示例。

表 1 原始/目标短语示例

原始短语	目标短语
the boy knew a lot of people in the city	strange images passed through my mind
all they think about is food and water	they were men of the desert and they were fearful of sorcerers
we are refugees from the tribal wars and we need money and other figure said	the boy reminded the old man that he had said something about hidden treasure
he was going to miss the place and all the good things he had learned	i told you to have the ice box fixed

评估指标：使用单词错误率（W_{er}）来评估转录结果与目标短语之间的差异。W_{er} 是语音识别或机器翻译系统性能的通用指标，用于解决识别的单词序列的长度与参考单词序列的长度不同的问题。$W_{er} = \frac{S + D + I}{N} \times 100\%$，其中 S 是替换数，$D$ 是删除数，I 是通过动态字符串对齐计算的插入数，N 是在文本中单词的总数。对于对抗性攻击，W_{er} 值越小，模型预测的短语越接近目标短语，并且表明攻击方法越好。使用 $S_{uc} = \frac{N_o}{N_a} \times 100\%$ 来计算攻击成功率，其中 N_o 是被识别为目标短语的音频样本的数量，N_a 是所有用来测试的音频样本数量。此外，遵循 Carlini 等人在论文中的设置[14]，同时使用分贝（dB）来测量音频扰动的大小，$d_B(x) = \max_i 20 \cdot \log_{10}(x_i)$。将扰动 δ 的 dB 水平与原始音频输入 x 进行比较。具体地，$d_{B\delta} = d_B(\delta) - d_B(x)$。希望添加尽可能小的扰动，以使人类无法感知，因此 $d_{B\delta}$ 通常为负数。$d_{B\delta}$ 越小，表明增加的扰动越小。此外，计算完成攻击所需的平均迭代次数 I_{ter}，以评估算法的效率。

3.2 实现细节

对于 C&W 实验，将 c 设置为：0.1,1.0,10,100 和 1000；相应地，将学习率设置为 10，迭代次数设置为 5000，分别对 MCV 和 LS 数据集进行攻击。在成功攻击的前提下，记录最小扰动的迭代次数及其对应的扰动值 δ，用于计算平均迭代次数和 $d_{B\delta}$。

对于 PGD 实验，针对 MCV 数据集进行了评估。首先将最大扰动预算 ϵ 设置为 300、500、800、1000、1500 和 2000，固定步长 $\alpha = 10$，以评估不同预设扰动对产生的扰动的影响。然后，确定最大预设扰动，并将步长分别设置为 10、20，...，100，以评估不同步长对所产生的扰动的影响。由于 PGD 的目的不是寻找最小的扰动，而是使模型预测短语和目标短语在最大允许扰动范围内尽可能相似。因此，将记录第一次成功攻击的迭代次数和扰动大小。

对于所提出的 A-FTA 方法，对两个数据集分别进行 100、300、500 和 1000 次迭代的攻击，以证明攻击的有效性。在所有情况下，设置初始化参数：$\epsilon_0 = 1000$，$\lambda = 0.2$ 和 $\alpha = 20$。在成功攻击的前提下，记录下最小扰动的迭代次数及其对应的扰动值大小。

对于所有实验，如果在迭代结束时未生成成功的对抗示例，则会记录最后一次迭代的次数和扰动。所有实验均在配备 Intel Xeon CPU E5-2620 v4、64G 内存和一个 GTX TITAN XP GPU 的 Ubuntu 16.04 工作站

上进行。

3.3 实验结果

1）**对 C&W 方法的评估**：为了说明所提出的方法的有效性，笔者将 A-FTA 与 C&W 方法进行了比较。对于 C&W 方法，使用了作者发布的源码，并用与笔者提出方法相同的测试集进行了评估。表 2 显示了在不同c值的情况下 5000 次迭代的评估结果。从表 2 中可以看出，随着分类损失值的比例增加，$d_{B\delta}$的值也逐渐增加，攻击成功率显着提高。W_{er}分别从最初的 102.92%、135.23%降低到 0.29%和 0.00%，这表明随着c值的增加，预测短语与目标短语越来越相似。成功攻击所需的平均迭代次数也随着c的增加而逐渐减少。根据表 2 中的结果，选择$c=1000$的结果作为基线与笔者提出的方法进行比较。

表 2 不同 c 值的 5000 次迭代评估结果

c	$d_{B\delta}$		W_{er} / %		I_{ter}		S_{uc} / %	
	MCV	LS	MCV	LS	MCV	LS	MCV	LS
0.1	-48.73	-47.10	102.92	135.23	4950	5000	1	0
1	-46.51	-46.58	80.86	122.30	4869	5000	6	0
10	-44.22	-42.81	36.60	76.70	4178	4798	42	6
100	-43.12	-42.68	5.77	3.94	3495	3960	85	84
1000	-41.17	-40.58	0.29	0.00	2802	2688	99	100

2）**对 PGD 方法的评估**：在这里，展示了在标准 PGD 攻击中常用的参数变动（步长，最大预设扰动）相对应的评估结果。对于最大预设扰动ϵ，将其设置为 300、500、800、1000、1500 和 2000，并固定步长$\alpha=10$。从表 3 中可以看出，$d_{B\delta}$的值与ϵ呈正相关，而迭代次数与ϵ为负相关。不难理解，ϵ的数值越大，所产生的扰动也就越大，损失收敛的速度也会更快。两个值都在$\epsilon=1000$左右时，逐渐稳定。对于所有不同的ϵ值，PGD 方法均达到了 100%的成功率且W_{er}值为 0%。C&W 方法在$d_{B\delta}$上则要优于 PGD 方法。这是因为 C&W 方法的目的是找到具有最小扰动的对抗样本，而 PGD 的目的是在最大预设扰动中最小化损失函数，从而高效地获取成功的对抗样本。

表 3 不同扰动预算ϵ的评估结果

ϵ	$d_{B\delta}$	W_{er} / %	I_{ter}	S_{uc} / %
300	-32.49	0	237	100
500	-28.65	0	227	100
800	-26.07	0	205	100
1000	-25.32	0	199	100
1500	-24.55	0	199	100
2000	-24.34	0	199	100

对于步长α，选择$\epsilon=300$和$\epsilon=2000$作为最大预设扰动，并设置α在范围 10-100 之间观察相应的结果。如表 4 所示，对于$\epsilon=300$，随着攻击步长α的增加，生成的音频对抗样本的W_{er}值也不断增加，攻击成功率则持续降低。攻击所需的平均迭代次数有所波动，但总体上是逐渐增加的。因为最大扰动ϵ限制为 300，所以样本的$d_{B\delta}$值几乎保持不变。对于$\epsilon=2000$，在所有不同步长α的情况下，攻击成功率均为 100%，并且成功进行一次攻击所需的迭代次数也随着α的增加而降低，并逐渐趋于稳定。$d_{B\delta}$值则是逐渐增加，最终同样也趋于稳定。

从 PGD 方法的实验结果可以看出，通过这种方式生成的对抗样本受预设扰动和不同固定步长的影响。对于较大的预设扰动值，使用较大的步长可以在短时间内生成对抗样本，但最终会给音频样本带来更大的扰动。而较小的预设值虽然可使所生成的对抗样本的$d_{B\delta}$变小，使得人耳难以察觉，但是生成一个成功的对抗样本需要更长的时间，并不高效。

表 4 不同步长α 的评估结果

α	$d_{B\delta}$		W_{er} / %		I_{ter}		S_{uc} / %	
	$\epsilon = 300$	$\epsilon = 2000$	$\epsilon = 300$	$\epsilon = 2000$	$\epsilon = 300$	$\epsilon = 2000$	$\epsilon = 300$	$\epsilon = 2000$
10	-32.49	-24.43	0.00	0.00	237	199	100	100
20	-32.44	-21.06	1.57	0.00	60	144	94	100
30	-32.37	-19.07	1.96	0.00	100	113	90	100
40	-32.37	-17.68	6.83	0.00	240	105	73	100
50	-32.37	-16.78	7.06	0.00	290	95	70	100
60	-32.37	-16.36	14.00	0.00	420	94	56	100
70	-32.37	-16.13	20.90	0.00	570	89	43	100
80	-32.37	-16.09	33.45	0.00	700	87	28	100
90	-32.37	-16.07	33.05	0.00	730	81	24	100
100	-32.37	-15.97	34.39	0.00	770	81	22	100

3）**A-FTA 方法的评估**：前两个实验的评估结果表明,标准的 PGD 方法可以在$\epsilon = 300$ 的设置下，通过 250 次迭代生成有效的目标音频对抗样本，但所产生的噪声大小和攻击成功率会受到最大扰动预算和步长选择的影响。

在 A-FTA 的实验中，为了证明 A-FTA 方法的有效性，分别选择迭代次数为 100、300、500 和 1000 次来评估攻击。实验结果在表 5 中给出。对于 MCV 数据集，所提出的方法的攻击成功率平均只需要迭代 97 次，就能达到 98%的攻击成功率,且$d_{B\delta}$ = -31.17，这与 PGD 方法在$\epsilon = 300$ 情况下进行了 237 次迭代的结果相近。随着迭代次数的增加，$d_{B\delta}$ 值逐渐减小。在仅迭代 300 次的限制下，$d_{B\delta}$ 值可以达到-44.00，要优于 C&W 方法中的-41.17，而这是该方法在平均迭代 2802 次的情况下才得到的。在迭代次数仅有 100 次的条件下，A-FTA 在 LS 数据集下依然达到了 68%的攻击成功率，虽然低于 MCV 数据集的 98%成功率，但依然远优于 C&W 方法。同样在迭代 300 次的限制下，达到了与 C&W 方法相近的效果。实验表明，A-FTA 方法在不同数据集上都有着良好的攻击效率与效果，证明了笔者提出方法的泛化性。

表 5 A-FTA 评估结果

迭代次数	$d_{B\delta}$		W_{er} / %		I_{ter}		S_{uc} / %	
	MCV	LS	MCV	LS	MCV	LS	MCV	LS
100	-31.17	-24.81	0.69	5.90	97	96	98	68
300	-44.00	-38.17	0.00	0.00	287	276	100	100
500	-46.83	-42.58	0.00	0.00	442	443	100	100
1000	-49.02	-45.60	0.00	0.00	660	772	100	100

4. 总　结

在本文中分析了比较了两种生成对抗样本的方法，并提出了 A-FTA 攻击，这是一种高效快速可以生成具有最小扰动的目标音频对抗样本方法。实验结果表明，与当前的最先进的方法相比，A-FTA 方法在成功攻击原始音频样本的前提下，所添加的扰动更加小，同时效率也更高，需要更少的迭代次数就可以实现有目标攻击。同时，笔者提出的方法具有良好的泛化性，不受数据集的影响。

由于设备的限制，笔者并没有进一步尝试将该方法改进应用在物理攻击（如所生成的对抗样本在空气

传播的过程中依然具有对抗性）中。但是，通过计算环境噪声转换的期望值，理论上 A-FTA 方法可以轻松地应用于物理攻击。

在后续的工作中，将专注于不同类型的音频对抗攻击及其相应的防御方法的研究。

参考文献:

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[2] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," arXiv preprint arXiv:1606.04435, 2016.

[3] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018, pp. 1–7.

[4] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured prediction models," arXiv preprint arXiv:1707.05373, 2017.

[5] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates et al., "Deep speech: Scaling up end-to-end speech recognition," arXiv preprint arXiv:1412.5567, 2014.

[6] L. Sch¨onherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," arXiv preprint arXiv:1808.05665, 2018.

[7] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," in 27th fUSENIXg Security Symposium (fUSENIXg Security 18), 2018, pp. 49–64.

[8] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," arXiv preprint arXiv:1810.11793, 2018.

[9] Y. Qin, N. Carlini, I. Goodfellow, G. Cottrell, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," arXiv preprint arXiv:1903.10346, 2019.

[10] M. Bosi and R. E. Goldberg, Introduction to digital audio coding and standards. Springer Science & Business Media, 2012, vol. 721.

[11] J. Shen, P. Nguyen, Y. Wu, Z. Chen, M. X. Chen, Y. Jia, A. Kannan, T. Sainath, Y. Cao, C.-C. Chiu et al., "Lingvo: a modular and scalable framework for sequence-to-sequence modeling," arXiv preprint arXiv:1902.08295, 2019.

[12] X. Liu, K. Wan, Y. Ding, X. Zhang, and Q. Zhu, "Weighted-sampling audio adversarial example attack." in AAAI, 2020, pp. 4908–4915.

[13] J. Li, S. Qu, X. Li, J. Szurley, J. Z. Kolter, and F. Metze, "Adversarial music: Real world audio adversary against wake-word detection system," in Advances in Neural Information Processing Systems, 2019, pp. 11 931–11 941.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] A. Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.