Attack on Speech Recognition

Todo List

1. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.
2. Biggio, B., Corona, I., Maiorca, D., Nelson, B., ˇSrndi´c, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In Joint European conference on machine learning and knowledge discovery in databases, pp. 387–402. Springer, 2013.
3. N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In Security and Privacy (SP), 2017 IEEE Symposium on, pages 39–57. IEEE, 2017.
4. N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou. Hidden voice commands. In 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, 2016.
5. A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in Conference on Computer Vision and Pattern Recognition. IEEE, Jun. 2015, pp. 427–436.
6. N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in Symposium on Security and Privacy. IEEE, May 2017, pp. 39–57.
7. I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, “Robust physical-world attacks on machine learning models,” CoRR, vol. abs/1707.08945,

pp. 1–11, Jul. 2017.
8. G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, “DolphinAttack: Inaudible voice commands,” in Conference on Computer and Communications Security. ACM, Oct. 2017, pp. 103–117.
9. Moustapha Ciss´e, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems, pages 6980–6990, 2017.

* Did you hear that? Adversarial Examples Against Automatic Speech Recognition

Contribution

1. 针对关键词识别模型进行黑盒攻击；

Notes

1. **黑盒**、**有目标**的**针对语音关键词识别**的对抗攻击算法。攻击的模型是 **Speech Commands Classification Model**，其中涉及的关键词有 `yes`、`no`、`up`、`down`、`left`、`right`、`on`、`off`、`stop` 和 `go`；
2. 算法流程：

Algorithm 1: Generation of Targeted Adversarial Audio Files using Genetic Algorithm

Inputs : Original benign example $\mathbf{x}$
target classification label t
Output : Targeted attack example $\mathbf{x}_{adv}$

```
/* Initialize the population of candidate solutions                    */
population ⟵ InitializePopulation(x)
iter_num = 0
while iter_num < max_iter do
    scores ⟵ ComputeFitness(population)
    x_adv ⟵ population [argmax(scores)]
    if argmax f(x_adv) = t then
        break // Attack succeeded, Stop early.
    end
    /* Compute selection probabilities.                                 */
    select_probs ⟵ Softmax(scores/Temp)
    Next population ⟵ { }
    for i ← 1 to size do
        Select parent_1 from population according to probabilities select_probs
        Select parent_2 from population according to probabilities select_probs
        child = Crossover(parent_1, parent_2)
        Next population = Next population ⋃ {child}
    end
    foreach child of Next population do Mutate(child)
    population ⟵ Next population
    iter_num = iter_num + 1
end
return x_adv
```

使用**遗传算法**生成对抗样本；

Links

- 论文链接：Alzantot, Moustafa, Bharathan Balaji, and Mani Srivastava. "Did you hear that? adversarial examples against automatic speech recognition." *NIPS Machine Deception Workshop* (2017).
- 论文主页：Adversarial Speech Commands | adversarial_audio (nesl.github.io)
- 论文代码：nesl/adversarial_audio (github.com)

Audio Adversarial Examples: Targeted Attacks on Speech-to-Text

Contribution

1. 白盒、有目标的、攻击端到端 DeepSpeech 模型（CTC）的对抗攻击算法；

Notes

1. **白盒**、**有目标**的对抗攻击算法。攻击的模型为 **DeepSpeech** 模型，攻击的指令为**任意长度**；
2. **基础**的攻击方法，loss 函数（ 后半部分为 CTC-Loss ）如下：

$$\begin{aligned} &\text{minimize } |\delta|_2^2 + c \cdot \ell(x+\delta, t) \\ &\text{such that } dB_x(\delta) \leq \tau \end{aligned}$$

   作者提到使用 2 范数而不用无穷范数的原因是，无穷范数可能会导致不收敛的问题，难以训练。在参数的选择上，作者使用 Adam 算法，学习率为 5，迭代论述为 5000。在实验过程中，作者发现**目标指令越长**，需要添加越多的扰动来生成对抗样本；而如果**原始指令越长**，似乎更加容易生成对抗样本（这一点我的想法是，**如果原始指令越长，原始存在更多的音素和能量可以被梯度下降过程利用**）。

3. **改进**的攻击方法（作者称：这种改进的攻击方法只能在 DeepSpeech 使用 Greedy-Search 的情况下有效），loss 函数如下：

$$\begin{aligned} &\text{minimize } |\delta|_2^2 + \sum_i c_i \cdot L_i(x+\delta, \pi_i) \\ &\text{such that } dB_x(\delta) < \tau \end{aligned}$$

   其中$L_i(\boldsymbol{x}, \pi_i) = \ell(f(\boldsymbol{x})^i, \pi_i)$表示对于当前 alignment，第 i 帧的 loss 值。作者这样修改 loss 函数的原因大致有两个：

   (1) 如果使用 CTC-Loss，会添加不必要的修改。**如果已经解码出 "ABCX"，目标指令为 "ABCD"，在使用 CTC-Loss时，梯度下降算法仍然会在 "A" 上添加扰动使得其变得更像 "A"**；

   (2) **不同的字符生成的难易程度是不同的**，所以把权重系数 c 移到了累加的里面。(**这一点作者称是在 Hidden Voice Command 中发现的规律，但其实只是在附录中给出了不同的单词可能需要的最短音素帧的数量是不同的，并没有给出字符难易程度的结论；并且这篇文章开源的代码中也没有给出这个改进的 loss 函数，所以可以直接把这个 c 移出去作为单个参数进行调参**)；

   训练的 **trick**：首先用 CTC-Loss 生成一个对抗样本，以这个对抗样本为参照固定 alignment（在 CTC 中，可能有许多种alignment，作者通过这种方法来确定选择其中一种），然后用改进的 loss 函数来生成；（这边，我的想法是，**改进的攻击方法会使得对抗样本丧失其迁移性，因为它只是恰好将特征拟合到模型的边界而已，而没有去进一步地逼近泛化的特征上**）

4. Evaluation：

   (1) 作者在原始指令的基础上通过非常小（约为 -30dB）的扰动生成对抗样本，并且在白盒的情况下最多可以在 **1s** 的语音中插入 **50** 个字符；

   (2) **对于 Non-Speech，作者发现更难生成对抗样本**；

   (3) 作者还对比了 FGSM 和 Iterative Optimization 两种生成对抗样本的算法，发现**在语音识别领域 FGSM 只适合生成 un-targeted 样本，而不适合生成 targeted 样本（或者说生成的效率很差，几乎没有办法生成）**；

   (4) 作者发现**这种方法生成的对抗样本是对噪声不鲁棒的**；

Links

- 论文链接：Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018.
- 论文主页：Audio Adversarial Examples (carlini.com)
- 论文代码：carlini/audio_adversarial_examples: Targeted Adversarial Examples on Speech-to-Text systems (github.com)

Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding

Contribution

1. 白盒、有目标的、针对API的、针对Kaldi DNN-HMM模型的对抗攻击算法；
2. 首次提出使用声学掩蔽效应；
3. 实验部分做的很全面，值得借鉴；

Notes

1. **白盒**、**有目标**的、**只针对API**的对抗攻击算法。攻击的模型为 Kaldi 的 **WSJ** 模型（或称为 recipe）；
2. 攻击方法整体架构图如下：

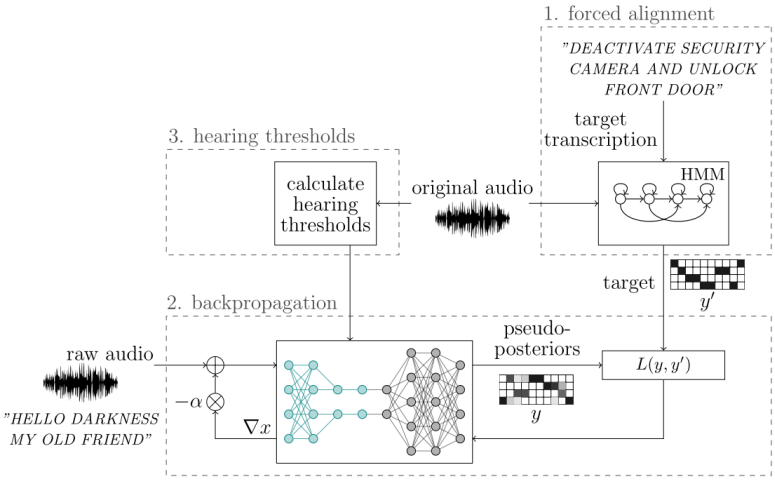


Fig. 3: The creation of adversarial examples can be divided into three components: (1) *forced alignment* to find an optimal target for the (2) backpropagation and the integration of (3) the hearing thresholds.

(1) **Forced Alignment**：时序信号经过分帧后，每一帧都有对应音素的概率分布（Kaldi 中使用 tri-phone，直接说音素比较好理解）。作者根据目标指令和原始语音找到一个最好的对齐方式，目的是为了**使得修改量最小**；

(2) **Back-propagation in Feature Extraction**：语音识别过程给网络的一般是 MFCC、Mel-log Filter Bank 等语音特征，把它简单地理解成是一张**二维热力图**，算法需要把梯度从这个特征回传到时域信号。（Kaldi 不像 tensorflow 那样直接就帮你把梯度计算好了，所以作者去推导了相关的梯度计算公式。不过，这里作者只推导了对数能量谱的梯度，但是 WSJ 里面用的应该是 MFCC 才对。另外不清楚作者用的是优化器，还需要看一下 Kaldi 代码。）。

(3) **Hearing Thresholds**：**心理声学掩蔽效应**，可以计算出音频各时间点、各个频率的**能量掩蔽值**，只要修改量不超过这个值，那么人就不会察觉。作者计算样本的能量变化 D，期望 D 在任何时间点、频率点均小于掩蔽阈值 H，公式如下（论文中的公式有个小错误，f 应该是 k）：

$$D(t,f) \leq H(t,k), \quad \forall t,k,$$
$$\text{with} \quad D(t,k) = 20 \cdot \log_{10} \frac{|S(t,k) - M(t,k)|}{\max_{t,k}(|\mathbf{S}|)}$$

变量 $\mathbf{\Phi}$ 度量能量变化 D 和 掩蔽阈值 H 之间的差值。如果 D 在任何点都不能超过 H，这样的限制条件过于苛刻，可能会导致无法生成对抗样本。故作者添加一个系数来放宽这个限制条件，公式如下：

$$\mathbf{\Phi} = \mathbf{H} - \mathbf{D}$$
$$\mathbf{\Phi}^* = \mathbf{\Phi} + \lambda$$

将 $\mathbf{\Phi}$ 小于 0 的值置为 0 并归一化到 0~1，公式如下：

$$\Phi^*(t,k) = 0, \ if \ \Phi^*(t,k) < 0$$
$$\hat{\Phi}(t,k) = \frac{\Phi^*(t,k) - \min_{t,k}(\mathbf{\Phi}^*)}{\max_{t,k}(\mathbf{\Phi}^*) - \min_{t,k}(\mathbf{\Phi}^*)}, \quad \forall t,k$$

只添加 $\mathbf{\Phi}$ 到梯度回传中时，作者发现差点意思。将 H 归一化到 0~1，公式如下：

$$\hat{H}(t,k) = \frac{H(t,k) - \min_{t,k}(H)}{\max_{t,k}(H) - \min_{t,k}(H)}$$

最后作者将这两个系数结合到 DFT 的梯度回传上（声学特征的计算这里不做解释了，推荐 Mel Frequency Cepstral Coefficient (MFCC) tutorial），公式如下：

$$\nabla X^*(t,k) = \nabla X(t,k) \cdot \hat{\Phi}(t,k) \cdot \hat{H}(t,k), \quad \forall t,k$$

我对这一块的理解：**整体来看，作者想要使用 "心理声学掩蔽效应" 来生成更具隐藏性（或者说修改量小）的对抗样本，他认为 "当前掩蔽值大" 、并且"修改量远小于掩蔽值" 的点可以添加更多的扰动，回传的梯度可以更大；而 "当前掩蔽值小"、或者是 "修改量已经接近掩蔽值" 的点不应该再做更多的修改，回传的梯度趋近于 0 。相对而言，我更喜欢 "Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition" 的工作，他们直接将 "心理掩蔽效应" 添加到了 loss 函数中去，让模型自己来选择梯度的变化。**

最后思考一个问题：**这样用掩蔽阈值和 perturbation 的差值来度量真的是一种好的方法吗？可能不是，我们其实更希望的是去度量 频率掩蔽曲线 的变化有多大。举例来说，计算掩蔽阈值的时候首先得到的是 masker（可以理解为一个频率点，其能量是个极值点），我们在 masker 处增加能量来抬高 masker（完全可以做到增加的能量低于 masker 处的掩蔽值，因为这个点的掩蔽值等于 masker 的能量，这个值是很大的），这样人耳的听觉感受已经发生了改变。但是如果要这么来做，就要用可求解梯度的方法来实现 "计算掩蔽值" 的过程，过程实在是很复杂，这也可能是大家不这么做的原因（代价太大，做出来还不知道能不能收敛，效果好不好）。**

3. Evaluation：

(1) 目标指令：

Target Transcriptions.

```
01: DO NOT BLAME YOU
02: THE COMMAND IS PLANTED
03: THE CAKE IS A LIE
04: THE COMMAND IS IN MY BRAIN
05: I'M AN INVADER COMING FOR YOU
06: WINTER IS COMING ZOMBIE COMING
07: IN MY RIGHT HAND
08: PRINCESS IN THE CASTLE
09: THEY DON'T BLAME YOU FIND A BOY
10: WELCOME TO THE JUNGLE ZOMBIE COMING WINTER IS COMING
11: THE CAKE IS A LIE DON'T BLAME YOU
12: I BELIEVE MOST PEOPLE ARE GOOD
13: THE HEAD THEY ARE STILL FIGHTING
14: I BELIEVE ALL PEOPLE ARE GOOD
15: THE SOUND OF SILENCE
16: IN THE MONEY CASTLE
17: WINTER IS COMING
18: DEACTIVATE SECURITY CAMERA AND UNLOCK FRONT DOOR
19: HE IS A MAN HE'S A GHOST
20: INTO YOUR FACE
21: TODAY I AM GOING NOWHERE
```

(2) 原始音频：Speech (从 WSJ 数据集中获取) + Music

(3) 评估指标：**WER** 和 **平均修改能量**，前者越小越好，后者越大越好；

(4) 分析 Hearing Threshold 和 Forced Alignment 的效果，学习率 **0.05 (这个和其他的工作相差挺大，猜测可能是因为有像 librosa 那样的 normalization)**，迭代 **500** 轮：

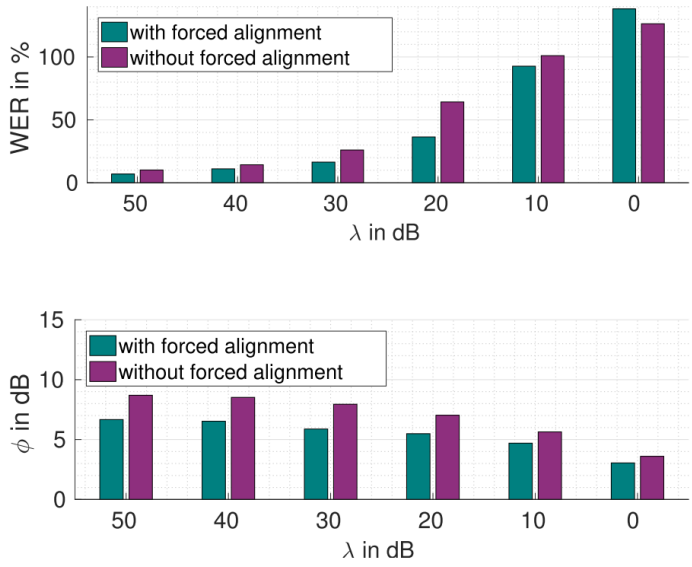


Fig. 6: Comparison of the algorithm with and without forced alignment, evaluated for different values of λ.

(5) 分析单位时间嵌入词数量的影响：

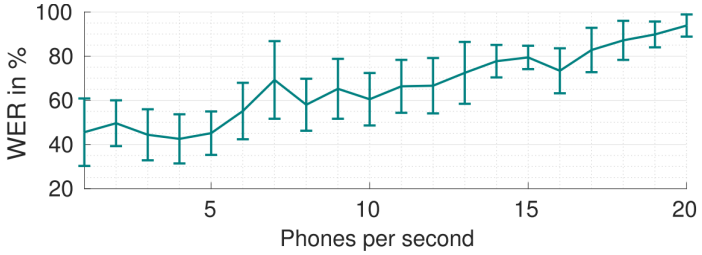
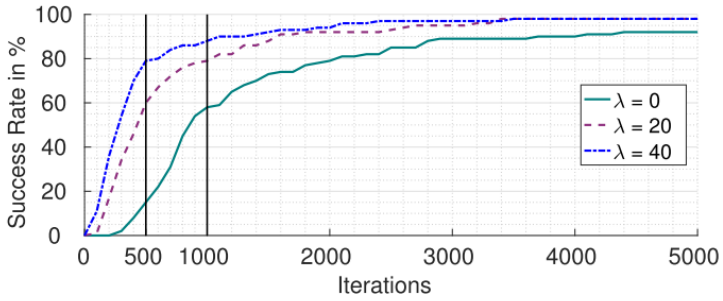
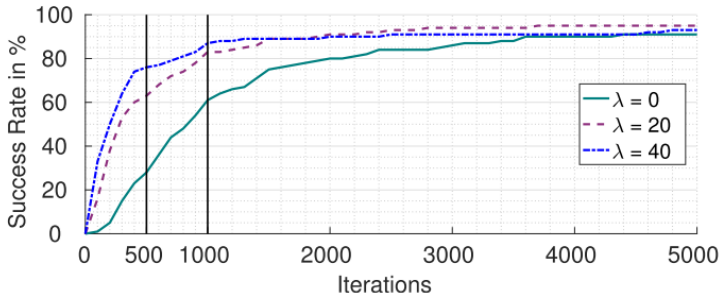


Fig. 7: Accuracy for different phone rates. To create the examples, 500 iterations of backpropagation and $\lambda = 20$ are used. The vertical lines represent the variances.

(6) 分析迭代轮数的影响：



(a) Speech



(b) Music

Fig. 8: Success rate as a function of the number of iterations. The upper plot shows the result for speech audio samples and the bottom plot the results for music audio samples. Both sets were tested for different settings of λ.

(7) 和 CommandSong 进行对比，对比的指标为 SNR：

TABLE III: Comparison of SNR with *CommanderSong* [61], best result shown in bold print.

	None	40 dB	20 dB	0 dB	*CommanderSong* [61]
SNR	15.88	17.93	**21.76**	19.38	15.32

$$\text{SNR(dB)} = 10 \cdot \log_{10} \frac{P_x}{P_\sigma},$$

Links

1. 论文链接：Schönherr, Lea, et al. "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding." *arXiv preprint arXiv:1808.05665* (2018).
2. 论文主页：Adversarial Attacks (adversarial-attacks.net)
3. 论文代码：rub-ksv/adversarialattacks: Adversarial Attacks (github.com)

* Targeted adversarial examples for black box audio systems

Contribution

1. 黑盒、有目标的攻击 DeepSpeech 的对抗攻击算法；
2. 结合遗传算法和梯度下降算法 (**在语音上没人这么来做过，但是其实在图像识别上面这种黑盒攻击不是新鲜事了，所以在算法上面没有创新型，而且直接应用到语义领域，出现了 query 次数巨大的问题**)；
3. 思路是值得借鉴的，黑盒攻击一定有比白盒攻击更加 interesting 的问题，但是不能照搬图像领域的方法；

Notes

1. **黑盒、有目标**的对抗攻击算法。攻击的模型为 **DeepSpeech** 模型，选择的方式是 **遗传算法和梯度下降算法的结合**；
2. 从 CommonVoice 测试机中挑选出前100个样本作为原始音频，目标指令都是 **2** 个词的指令，比较短；
3. 作者假设 DeepSpeech 模型是不可知探知的，但是知道模型最后的概率分布的输出，并且针对 Greedy Decoding 进行攻击 （我的想法：**这样的攻击场景其实是不常见的，所以这个工作可能指导意义不大，但是我们应该思考一下，如果 ASR 模型经过了 LM 模型的修饰，还能不能用黑盒探测的方法来生成对抗样本？如果能，代价又有多大？**）；
4. 算法流程：

Algorithm 1 Black box algorithm for generating adversarial audio sample

Require: Original benign input $\boldsymbol{x}$ Target phrase t
Ensure: Adversarial Audio Sample $\boldsymbol{x'}$
  population $\leftarrow [\boldsymbol{x}] * populationSize$
  while iter $< maxIters$ and $Decode$(best)! $= t$ **do**
    scores $\leftarrow -CTCLoss$(population, t)
    best $\leftarrow$ population$[Argmax$(scores)$]$

    if $EditDistance(t, Decode$(best)$) > 2$ **then**
      // phase 1 - do genetic algorithm
      while populationSize children have not been made **do**
        Select $parent1$ from $topk$(population) according to $softmax$(their score)
        Select $parent2$ from $topk$(population) according to $softmax$(their score)
        child $\leftarrow Mutate(Crossover$(parent1, parent2), p)
      end while
      newScores $\leftarrow -CTCLoss$(newPopulation, t)
      p $\leftarrow MomentumUpdate$(p, newScores, scores)

    else
      // phase 2 - do gradient estimation
      top-element $\leftarrow top$(population)
      grad-pop $\leftarrow$ n copies of top-element, each mutated slightly at one index
      grad $\leftarrow$ (-CTCLoss(grad-pop) - scores) / mutation-delta
      pop $\leftarrow$ top-element + grad
    end if
  end while
  return best

(1) 当样本解码的字符串距离目标指令较大时，使用遗传算法生成对抗样本，遗传算法的评分函数使用 CTC-Loss，其变异概率 p 由函数 `MomentumUpdate` 进行更新；

$$p_{new} = \alpha \times p_{old} + \frac{\beta}{|currScore - prevScore|}$$

(2) 当样本解码的字符串距离目标指令较小时，使用黑盒-梯度下降算法生成对抗样本，对每个序列样本点 (花费巨大，对于一个 16kHz 的 5s 语音，每轮都要调用目标模型进行解码 80k 次) 都分别添加小的扰动，根据 CTC-Loss 值的变化，确定扰动的影响是正面的还是负面的、是重要的还是不重要的；

$$FD_x(x, \delta) = \begin{bmatrix} (g(x+\delta_1) - g(x))/\delta \\ \vdots \\ (g(x+\delta_n) - g(x))/\delta \end{bmatrix}$$

5. Evaluation：

(1) 使用 100 条原始语音，每个语音的目标指令是随机从 1000 个最常用的英语单词中抽取的 2 个单词，每个对抗样本，设置生成 3000 轮；

(2) 使用 Success Rate 来评估成功率，对抗样本的成功率为 35%；使用 Cross Correlation Coefficient 来评估相似性，对抗样本与原始语音的相似性为 94.6% (这里只看成功的对抗样本)；

Links

- 论文链接：Taori, Rohan, et al. "Targeted adversarial examples for black box audio systems." *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2019.
- 论文代码：rtaori/Black-Box-Audio: Targeted Adversarial Examples for Black Box Audio Systems (github.com)

Robust Audio Adversarial Example for a Physical Attack

Contribution

1. 引入脉冲响应；
2. 实现了较高的物理攻击成功率，并且用了两组播放和接收设备；
3. 指令过短，实验应该增加更多的物理环境；

Notes

1. **白盒、有目标**的**针对物理环境**的对抗攻击算法。攻击的模型为 **DeepSpeech** 模型，选取的指令都比较**短**；
2. 在图像领域的对抗攻击算法中，[Athalye et al., 2018] 等人提出了用一个抽象函数 t 来模拟物理环境下打印和拍照在样本上带来的扰动。将这个抽象函数 t结合到对抗样本的生成过程中去，可以大大增强生成的对抗样本的鲁棒性；
3. 作者提出的方法，关键点有三个：

   (1) **带通滤波器。**因为人的听觉频率范围是有限的，听筒-扬声器在工作的时候很多会直接丢弃其他频率的能量，所以作者设置了一个 **1000~4000** 范围的带通滤波器来减少修改量（我的看法：**我觉得 4000 这个上界是比较靠谱的，而 1000 这个下界可能并不合理，因为语音中低频的能量是比较多的，这部分的能量应该也是比较重要的；而这种带通滤波器的方法是否真的能够减少修改量也是存在问题的，因为依靠梯度下降算法，可能你限制了它修改的频带范围，需要的修改量可能是更多的**）。形式化公式如下：

$$\underset{\boldsymbol{v}}{\operatorname{argmin}} \underset{f}{Loss}\left(MFCC\left(\tilde{\boldsymbol{x}}\right), \boldsymbol{l}\right)+\epsilon\left\|\boldsymbol{v}\right\|$$
$$where\ \tilde{\boldsymbol{x}}=\boldsymbol{x}+\underset{1000\sim4000\mathrm{Hz}}{BPF}\left(\boldsymbol{v}\right)$$

   (2) **脉冲响应。**作者在生成对抗样本的过程中，添加脉冲响应的卷积来增强对抗样本对不同房间环境的鲁棒性。形式化公式如下：

$$\underset{\boldsymbol{v}}{\operatorname{argmin}}\, \mathbb{E}_{h\sim\mathcal{H}}\left[\underset{f}{Loss}\left(MFCC\left(\tilde{\boldsymbol{x}}\right), \boldsymbol{l}\right)+\epsilon\left\|\boldsymbol{v}\right\|\right]$$
$$where\ \bar{\boldsymbol{x}}=\underset{h}{Conv}\left(\boldsymbol{x}+\underset{1000\sim4000\mathrm{Hz}}{BPF}\left(\boldsymbol{v}\right)\right)$$

   (3) **高斯白噪声。**作者在生成对抗样本的过程中，添加高斯白噪声来增强对抗样本对背景白噪声的鲁棒性。形式化公式如下：

$$\underset{\boldsymbol{v}}{\operatorname{argmin}} \mathbb{E}_{h\sim\mathcal{H},\boldsymbol{w}\sim\mathcal{N}(0,\sigma^2)}\left[\underset{f}{Loss}\left(MFCC\left(\tilde{\boldsymbol{x}}\right),\boldsymbol{l}\right)+\epsilon\left\|\boldsymbol{v}\right\|\right]$$

$$where\, \bar{\boldsymbol{x}}=\underset{h}{Conv}\left(\boldsymbol{x}+\underset{1000\sim4000\mathrm{Hz}}{BPF}\left(\boldsymbol{v}\right)\right)+\boldsymbol{w} \qquad (7)$$

4. Evaluation:

(1) 作者其他的实现的细节与文章 "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text" 是一样的，Adam 迭代器和 CTC-Loss 函数；

(2) 提到了一个比较有意思的攻击场景：FM radio；

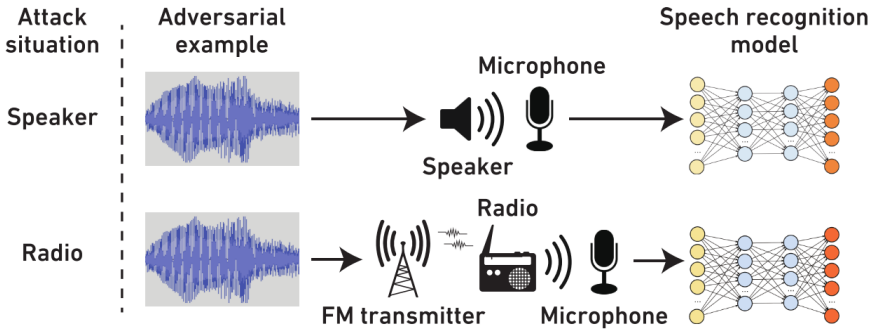


Figure 2: Two attack situations of the evaluation: speaker and radio. In the first situation, the adversarial examples were played and recorded by a speaker and a microphone. In the second situation, the adversarial examples were broadcasted using an FM radio.

(3) 分析针对 API 的攻击：

	Input sample	Target phrase	SNR
(A)	Bach	hello world	9.3dB
(B)	Bach	open the door	5.3dB
(C)	Bach	ok google	0.2dB
(D)	Owl City	hello world	11.8dB
(E)	Owl City	open the door	13.4dB
(F)	Owl City	ok google	2.6dB

Table 1: Details of the generated audio adversarial examples, which showed 100% success by both the speaker and the radio and having the maximum value of SNR[8].

(4) 分析针对物理的攻击：

	Input sample	Target phrase	SNR	Attack situation	Success rate	Edit dist.
(G)	Bach	hello world	11.9dB	Speaker	60%	1.1
				Radio	50%	1.3
(H)	Bach	open the door	6.6dB	Speaker	60%	1.8
				Radio	60%	1.8
(I)	Bach	ok google	4.2dB	Speaker	80%	0.6
				Radio	70%	0.9
(J)	Owl City	hello world	12.2dB	Speaker	70%	0.9
				Radio	50%	1.5
(K)	Owl City	open the door	14.6dB	Speaker	90%	0.2
				Radio	100%	0.0
(L)	Owl City	ok google	8.7dB	Speaker	90%	0.6
				Radio	70%	0.9

Table 2: Details of the generated audio adversarial examples, which showed at least 50% success by both the speaker and the radio and having the maximum value of SNR[8].

这里作者每个样本都尝试 10 次，统计成功率，并且发现只保证成功率高于 50% 的情况下，可以适当减少修改量。

Links

- 论文链接：Yakura, Hiromu, and Jun Sakuma. "Robust audio adversarial example for a physical attack." *arXiv preprint arXiv:1810.11793* (2018).
- 论文主页：Robust Audio Adversarial Example for a Physical Attack (yumetaro.info)
- 论文代码：hiromu/robust_audio_ae: Robust Audio Adversarial Example for a Physical Attack (github.com)
- 冲击响应：

  The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech (ieee.org)

  A binaural room impulse response database for the evaluation of dereverberation algorithms (ieee.org)

  Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition (naist.jp)

  Evaluation of speech dereverberation algorithms using the MARDY database (2006) (ist.psu.edu)

  Acoustic measurement data from the varechoic chamber (nist.gov)

Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition

Contribution

1. 白盒、有目标的、针对端到端 LAS 模型的对抗攻击算法；
2. 心理掩蔽效应；
3. 模拟房间声学响应；

Notes

1. **白盒**、**有目标**的对抗攻击算法。攻击的模型为 Lingvo 框架的 **LAS** 模型，攻击的指令选取了 1000 条**中等长度**的字符串;
2. 关键点在于两方面，一点是使用了**心理掩蔽效应**来提高对抗样本的隐蔽性，另一点是**模拟房间声学响应**来提高对抗样本的鲁棒性;
3. 心理掩蔽效应，简单来说就是能量大的声音可能会能量小的声音，主要分为时间掩蔽和频率掩蔽。其中频率掩蔽是作者在文章中用到的，可以将其理解为语音在各个时间各个频率的一个阈值，修改量超出这个阈值，那么人就可以听到这个改动。添加心理掩蔽效应后的 loss 函数：

$$\ell(x,\delta,y)=\ell_{net}(f(x+\delta),y)+\alpha\cdot\ell_\theta(x,\delta)$$

前面部分保证**样本的成功率**，后面部分保证**样本的隐藏性**，**alpha** 控制两者的权重。作者生成对抗样本的时候有一个 **trick**（因为作者把两个放在一起时发现很难生成对抗样本）：(**Stage-1**) 先根据前面的 loss 函数生成一轮对抗样本，(**Stage-2**) 然后根据后面的 loss 函数生成一轮对抗样本，如果 stage-2 迭代 **20** 轮后，成功生成了对抗样本，那就把 alpha 增大一些（**说明可以增加一些隐藏性**）；如果 stage-2 迭代 **50** 轮，都没能生成对抗样本，那就把 alpha 减小一些（**说明需要牺牲一些隐藏性**）。具体的迭代生成算法如下：

Algorithm 1 Optimization with Masking Threshold

Input: audio waveform x, target phrase y, ASR system $f(\cdot)$, perturbation δ, loss function $\ell(x, \delta, y)$, hyperparameters ϵ and α, learning rate in the first stage lr_1 and second stage lr_2, number of iterations in the first stage T_1 and second stage T_2.

Stage 1: minimize $\|\delta\|$
Initialize $\delta = 0$, $\epsilon = 2000$ and $\alpha = 0$.
for $i = 0$ **to** $T_1 - 1$ **do**
  $\delta \leftarrow \delta - lr_1 \cdot \text{sign}(\nabla_\delta \ell(x, \delta, y))$
  Clip $\|\delta\| \leq \epsilon$
  if $i \,\%\, 10 = 0$ and $f(x + \delta) = y$ **then**
    if $\epsilon > \max(\|\delta\|)$ **then**
      $\epsilon \leftarrow \max(\|\delta\|)$
    end if
    $\epsilon \leftarrow 0.8 \cdot \epsilon$
  end if
end for

Stage 2: minimize the perceptibility
Reassign $\alpha = 0.05$
for $i = 0$ **to** $T_2 - 1$ **do**
  $\delta \leftarrow \delta - lr_2 \cdot \nabla_\delta \ell(x, \delta, y)$
  if $i \,\%\, 20 = 0$ and $f(x + \delta) = y$ **then**
    $\alpha \leftarrow 1.2 \cdot \alpha$
  end if
  if $i \,\%\, 50 = 0$ and $f(x + \delta) \neq y$ **then**
    $\alpha \leftarrow 0.8 \cdot \alpha$
  end if
end for
Output: adversarial example $x' = x + \delta$

4. 模拟房间声学响应，简单来说，当固定了房间的参数和你设备的参数，你可以将整个物理信道用一个函数 t(x) 来建模。添加房间声学响应后的 loss 函数：

$$\text{minimize} \;\; \ell(x, \delta, y) = \mathop{\mathbb{E}}_{t \sim \mathcal{T}} \left[\ell_{net}(f(t(x + \delta)), y) \right]$$
$$\text{such that} \;\; \|\delta\| < \epsilon.$$

   训练的 **trick**：(**Stage-1**) 使用 `lr_1=50` 迭代 *2000* 轮保证在其中 **1 个房间声学响应**下能够生成对抗样本，(**Stage-2**) 然后使用 `lr_2=5` 迭代 *5000* 轮来保证在另外随机采样的 **10 个房间声学响应**下都能够生成对抗样本（这个期间不再减小 **perturbation** 的上限）。

5. 结合心理掩蔽效应和模型房间声学响应。结合后的 loss 函数:

$$\ell(x, \delta, y) = \mathop{\mathbb{E}}_{t \sim \mathcal{T}} \left[\ell_{net}(f(t(x + \delta)), y) + \alpha \cdot \ell_\theta(x, \delta) \right]$$

   训练的 **trick**：(在 **4** 的对抗样本基础上) 结合整个 loss 函数来生成具有隐藏性的对抗样本，和 3 中的分两步生成不同。

6. Evaluation 部分，我觉得封面给的 100% 其实不是关键，因为该实验给的样本只是使用了心理掩蔽效应，考虑到这个攻击是个白盒攻击，所以白盒的非物理攻击其实在现实中意义不大，但是可以去论文主页听一下效果，确实添加的扰动几乎无法听到。所以主要还是关注它的第二个实验结果：

Input	Clean	Robust ($\Delta = 300$)	Robust ($\Delta = 400$)	Imperceptible & Robust
Accuracy (%)	31.37	62.96	64.64	49.65
WER (%)	15.42	14.45	13.83	22.98

   作者采用的对抗样本的评价指标分别是：Accuracy - 整句话的成功率，WER - 词错率，和隐藏性。其中隐藏性没有采用常用的 SNR 来度量，而是直接采用问卷调查的形式，作者的问卷调查的问题分别为：

   (1) 音频是否清晰；

   (2) 分辨两个音频哪一个是原始音频；

   (3) 判断两个音频是否相同；

Links

- 论文链接：Qin, Yao, et al. "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition." International Conference on Machine Learning. PMLR, 2019.

- 论文主页：Imperceptible, Robust and Targeted Adversarial Examples for Automatic Speech Recognition (ucsd.edu)
- 论文代码：cleverhans/examples/adversarial_asr at master · tensorflow/cleverhans (github.com)