Targeted Adversarial Examples for Black Box Audio Systems

Amog Kamsetty, Rohan Taori, Nikita Vemuri, Brenton Chu
Who Are We?

- Students at UC Berkeley
- Work done at Machine Learning @ Berkeley (ML@B)
  - ml.berkeley.edu
  - Aim to provide AI/ML opportunities at the undergraduate level
What is an Adversarial Example?
Adversarial Example

\[ x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \]

\[ x \quad \text{“panda”} \quad 57.7\% \text{ confidence} \]

\[ \text{sign}(\nabla_x J(\theta, x, y)) \quad \text{“nematode”} \quad 8.2\% \text{ confidence} \]

\[ x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \quad \text{“gibbon”} \quad 99.3\% \text{ confidence} \]
Untargeted vs Targeted

- Untargeted: Provide input to the model such that it misclassifies the adversarial input
- Targeted: Provide input to the model so it classifies it as a predetermined target class
White Box vs Black Box

- **White box:** complete knowledge of model architecture and parameters, allows for gradient computation
- **Black box:** no knowledge of model or parameters except for output logits of model
Why does this matter?

- Black box attacks can be of particular interest in ASR systems

- If we can create an adversarial audio file, we can trick the model into translating what we want

- If we do this with a black box approach, we can apply this to proprietary systems (ex. Google or IBM APIs)
Classical Adversarial Attacks

- Taking gradient iteratively
- FGSM - Fast Gradient Sign Method
- Houdini
Prior Work in Audio

- UCLA - Black box genetic algorithm on single word classes $\rightarrow$ softmax loss
- Carlini & Wagner: white box attack
  - CTC loss allows for comparison with arbitrary length translations
- Our project: Black box genetic algorithm on sentences using CTC Loss
Problem Statement

without the dataset the article is useless

ok google browse to evil.com
Problem Statement

- Black-box Targeted Attack
  - Given a target $t$, a benign input $x$, and model $M$, perturb $x$ to form $x'=x+\delta$
  - S.t. $M(x')=t$ while maximizing $\text{cross\_correlation}(x, x')$
  - Only have access to logits of $M$
    - Not given gradients!
Datasource: DeepSpeech

- The model we are targeting is DeepSpeech
  - Architecture created by Baidu
  - Tensorflow implementation by Mozilla; available on Github

- Utilize Common Voice dataset by Mozilla
  - Consists of voice samples
  - Sampling rate of 16 KHz
Final Algorithm: Guided Selection

- Genetic Algorithm approach
- Given the benign input, generate population of size 100
- On each iteration select the best 10 samples using scoring function
- Perform crossover and momentum mutation
- Apply high pass filter to added noise
Genetic Algorithm with Momentum

Population

Evaluate fitness with CTC loss

Elite

Crossover and Momentum Mutation (ours)

Perform this population size times
Momentum Mutation

\[ p_{new} = (\alpha \times p_{old}) + \frac{\beta}{|curr\,Score - prev\,Score|} \]

- Probability of mutation is function of difference in scores across iterations
- If little increase in score between iterations, increase “momentum” by increasing probability of mutation
- Encourages decoding to build up to target after making input similar to silence
Gradient Estimation

- Genetic algorithms work best when search space is large
- However, when adversarial sample is near target, only few key perturbations are necessary
- Apply gradient estimation at 100 random indices

\[ FD_x(g(x), \delta) = \begin{bmatrix} \frac{(g(x_1 + \delta) - g(x_1))}{\delta} \\ \vdots \\ \frac{(g(x_n + \delta) - g(x_n))}{\delta} \end{bmatrix} \]
Results

- Tested on first 100 samples of CommonVoice dataset
- Randomly generated 2 target words
- Targeted attack similarity: 89.25%
  - Algorithm could almost always reach the target
- Average similarity score: 94.6%
  - Computed via wav-file cross-correlation
## Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>White Box Attacks</th>
<th>Our Method</th>
<th>Single Word Black Box</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted attack success rate</td>
<td>100%</td>
<td>35%</td>
<td>87%</td>
</tr>
<tr>
<td>Average similarity score</td>
<td>99.9%</td>
<td>94.6%</td>
<td>89%</td>
</tr>
<tr>
<td>Similarity score method</td>
<td>cross-correlation</td>
<td>cross-correlation</td>
<td>human study</td>
</tr>
<tr>
<td>Loss used for attack</td>
<td>CTC</td>
<td>CTC</td>
<td>Softmax</td>
</tr>
<tr>
<td>Dataset tested on</td>
<td>Common Voice</td>
<td>Common Voice</td>
<td>Speech Commands</td>
</tr>
<tr>
<td>Target phrase generation</td>
<td>Single sentence</td>
<td>Two word phrases</td>
<td>Single word</td>
</tr>
</tbody>
</table>

![Histogram of Levenshtein distances of attacks](chart.png)
Example

Original file: “and you know it”

Adversarial target: “hello world”

Audio Similarity: 94%

(cross-correlation)
Future Work

● Attack a broader range of models
  ○ Transferability across models
● Increasing sample efficiency to target
  ○ API call costs can be prohibitive
● Computational Efficiency
Thank You!

Code and samples: https://github.com/rtaori/Black-Box-Audio