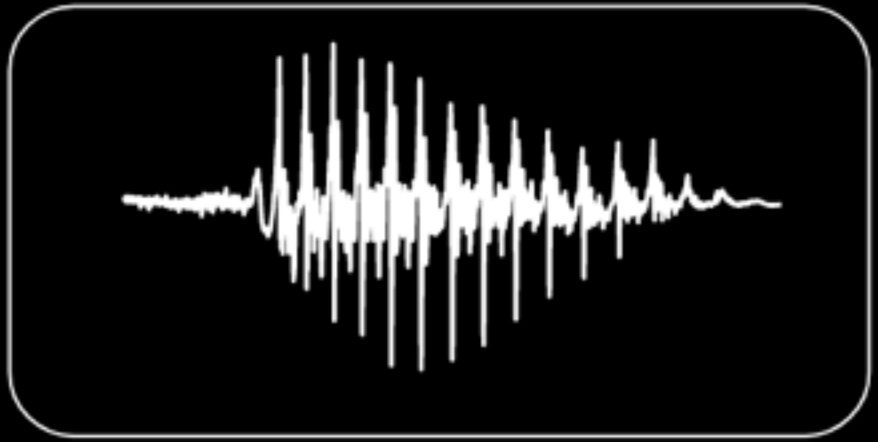
Audio Adversarial Examples: Targeted Attacks on Speech-To-Text

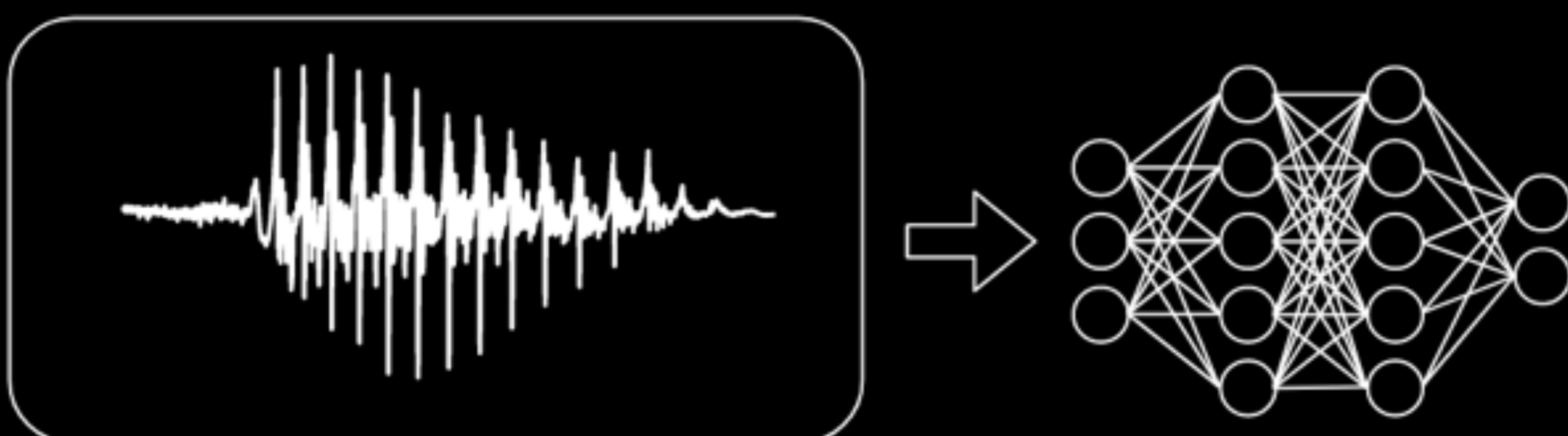
Nicholas Carlini and David Wagner
University of California, Berkeley

Background

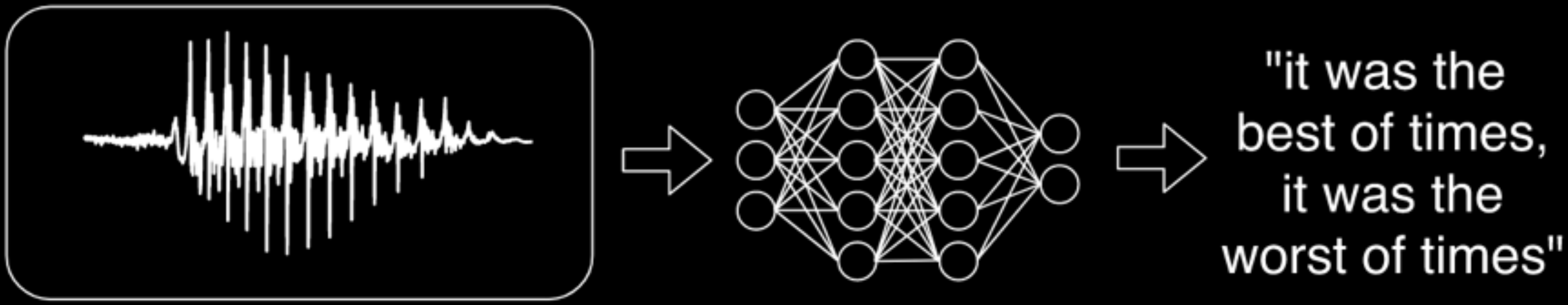
Neural Networks for Automatic Speech Recognition



Neural Networks for Automatic Speech Recognition



Neural Networks for Automatic Speech Recognition

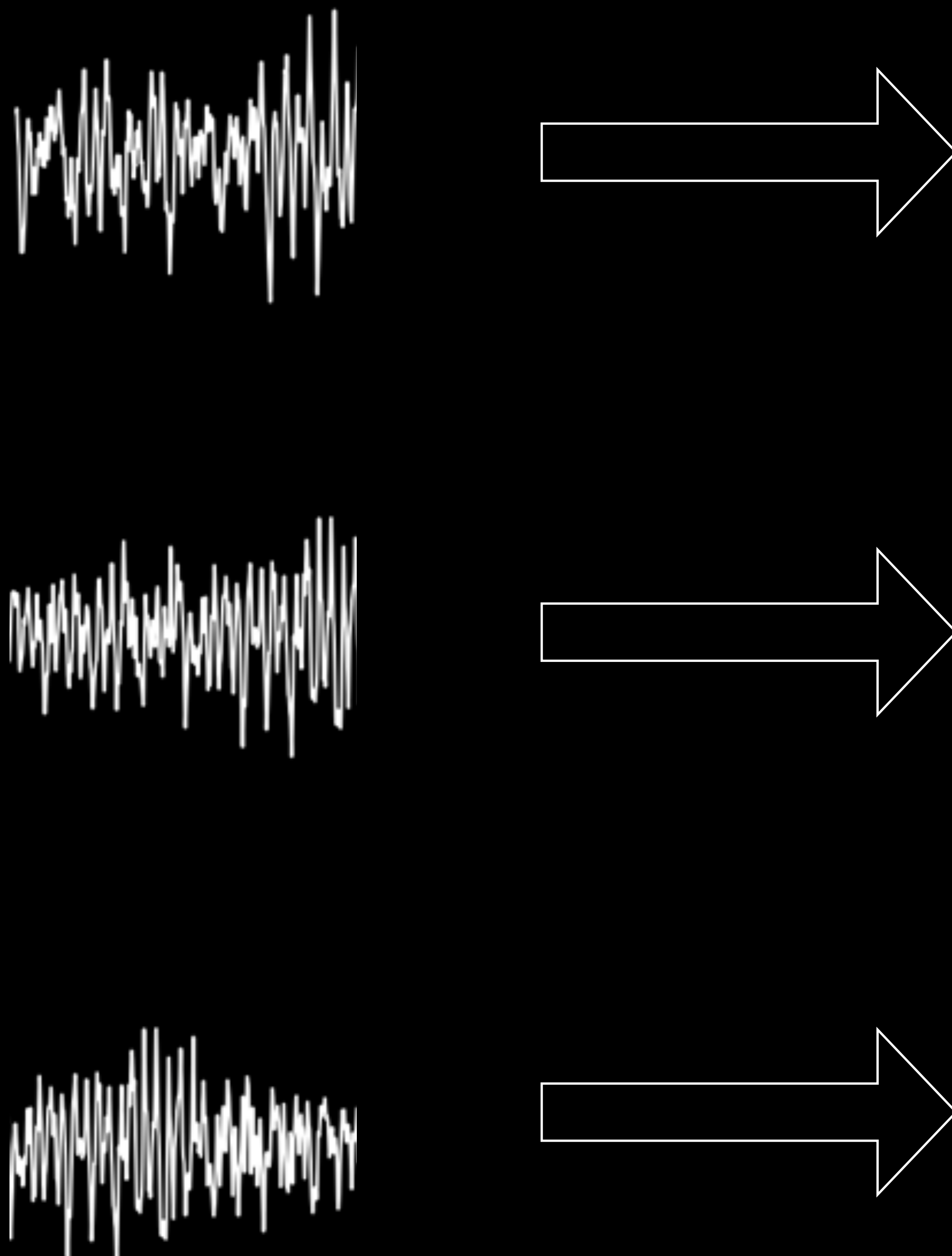


(slightly) More Formally

- Let an audio waveform X be a sequence of values [-1,1]
- Let F(X) be a neural network that outputs a sequence of probability distributions over characters *a-z* (and *space*)
    - (F is often a recurrent neural network)
- A decoder converts this sequence of probability distributions to the final output string

Training for Automatic Speech Recognition

Training Data:



"pairs of audio and text"

"of variable length"

"with no alignment"

New function:

CTC Loss

A differentiable measure of distance from $F(x)$ to the true target phrase

Training objective

Minimize CTC Loss between training audio and corresponding transcriptions

Background: Targeted Adversarial Examples

- Given an input X, classified as F(X) = L ...
- ... it is easy to find an X′ close to X
- ... so that F(X′) = T [for **any** T != L]

</background>

This Talk:

Can we construct targeted adversarial examples for automatic speech recognition?

Concretely,

Can we make a neural network
recognize this audio as any
target transcription?
(e.g., "okay google, browse to evil.com")

Why?

To to differentiate
properties of adversarial examples *on images*
from
properties of adversarial examples *in general*

Key Finding:

Most results on images hold true on audio, without (much) modification.

(Background) Constructing Adversarial Examples

- Formulation: given input x, find x′ where
  minimize   d(x,x′)
  such that   F(x′) = T
                     x′ is "valid"

Aside: what is our distance metric?

Magnitude of perturbation (in dB) relative to the source audio signal

(Background) Constructing Adversarial Examples

- Formulation: given input x, find x′ where
  minimize $\quad d(x,x')$
  such that $\quad F(x') = T$
  $\quad\quad\quad\quad\quad x'$ is "valid"

- Gradient Descent to the rescue?

- No. Non-linear constraints are hard

(Background) Reformulation

- Formulation:
  minimize $d(x,x') + g(x')$
  such that x' is "valid"

- Where $g(x')$ is some kind of loss function for how close $F(x')$ is to target T
  - $g(x')$ is small if $F(x') = T$
  - $g(x')$ is large if $F(x') \neq T$

What loss function $g(x')$ should we use?

CTC Loss!

Reformulation

- Formulation:
  minimize $\quad d(x,x') + \text{CTC-Loss}(x')$
  such that $\quad x'$ is "valid"

The only necessary change to get adversarial examples on speech-to-text

Despite the simplicity, if you do this, then things basically works as I said.

Despite the simplicity, if you do this, then things basically works as I said.

Okay, there are some details that are necessary but basically what I've said here is true, and if you apply gradient descent to the CTC loss and add some hyperparameter tuning then you can generate adversarial examples with low distortion. In order to make these samples remain adversarial when quantizing to 16-bit integers you have to add some Gaussian noise during the attack generation process to help prevent overfitting. And when you do this, the full process still often requires many thousand iterations to achieve which can take almost an hour when operating over very large audio samples, but can be sped up significantly by generating multiple adversarial examples simultaneously and then performing one final fine-tuning step that deals with some implementation difficulties of attacking variable length audio samples. But if you do all of this then things actually will work out and everything is fine with the adversarial examples. and now because I can I will just start to dump random text that seems like it might be relevant. We construct targeted audio adversarial examples on automatic speech recognition. Given any audio waveform, we can produce another that is over 99.9% similar, but transcribes as any phrase we choose (recognizing up to 50 characters per second of audio). We apply our white-box iterative optimization-based attack to Mozilla's implementation DeepSpeech end-to-end, and show it has a 100% success rate. The feasibility of this attack introduce a new domain to study adversarial examples As the use of neural networks continues to grow, it is critical to examine their behavior in adversarial settings. Prior work [8] has shown that neural networks are vulnerable to adversarial examples [40], instances x′ similar to a natural instance x, but classified by a neural network as any (incorrect) target t chosen by the adversary. Existing work on adversarial examples has focused largely on the space of images, be it image classification [40], gener- ative models on images [26], image segmentation [1], face detection [37], or reinforcement learning by manipulating the images the RL agent sees [6, 21]. In the discrete domain, there has been some study of adversarial examples over text classification [23] and malware classification [16, 20]. There has been comparatively little study on the space of audio, where the most common use is performing automatic speech recognition. In automatic speech recognition, a neural network is given an audio waveform x and perform the speech-to-text transform that gives the transcription y of the phrase being spoken (as used in, e.g., Apple Siri, Google Now, and Amazon Echo). Constructing targeted adversarial examples on speech recognition has proven difficult. Hidden and inaudible voice commands [11, 39, 41] are targeted attacks, but require synthesizing new audio and can not modify existing audio (analogous to the observation that neural networks can make high confidence predictions for unrecognizable images [33]). Other work has constructed standard untargeted adversarial examples on different audio systems [13, 24]. The current state-of-the-art targeted attack on automatic speech recog- nition is Houdini [12], which can only construct audio adversarial examples targeting phonetically similar phrases, leading the authors to state

Now for the fun part.

Mozilla's DeepSpeech

Mozilla's DeepSpeech
transcribes this as

"most of them were staring
quietly at the big table"

[adversarial]

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity"

It works on music, too

DeepSpeech transcribes
"speech can be embedded in music"

And can "hide" speech

DeepSpeech does not hear any speech in this audio sample

Key Limitation:

Only works when used directly as an audio waveform, *not* if played over-the-air

However,

Prior work (**Hidden Voice Commands** and **DolphinAttack**) are effective over-the-air;

Physical world adversarial examples exist on deep learning for image recognition

Also,

These audio adversarial examples are robust to *synthetic* forms of noise (sample-wise noise, MP3 compression)

Future Work:

New research questions
for audio adversarial examples

Can these attacks be played over-the-air?

Does the transferability property still hold?

Which defenses work on the audio domain?

Conclusion

- Most things we know about adversarial examples apply to audio without significant modification
  - Optimization-based attacks are effective
- Exciting opportunities for future work

https://nicholas.carlini.com/code/audio_adversarial_examples

New domain to compare

neural networks

to

traditional methods

State-of-the-art attack on "traditional" methods

Audio adversarial examples (so far) do not exist on audio using traditional machine learning methods