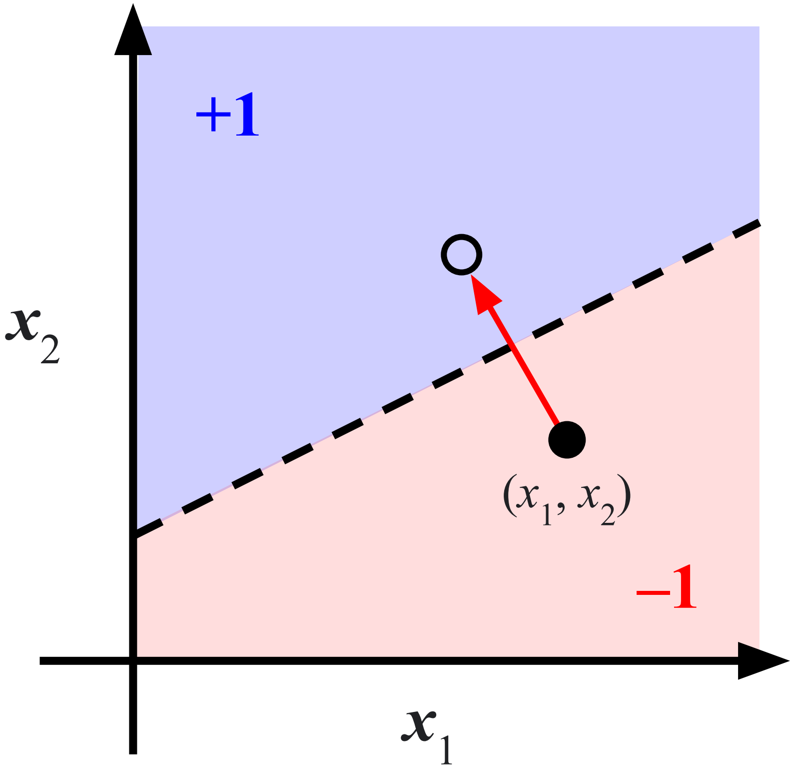
Adversarial Examples, Part II

4.27.2022

Patrick O'Reilly

Adversarial Examples: View 1



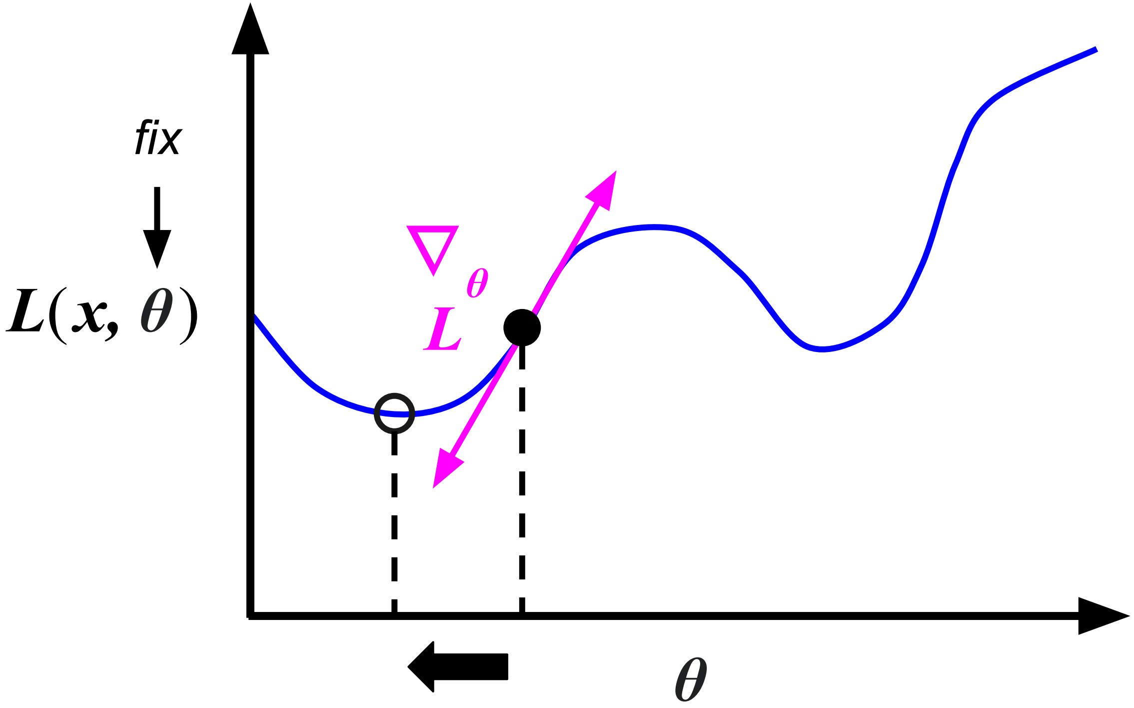
$$f(x) = \text{sign}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$$

$$\boldsymbol{L}(\boldsymbol{x}, \boldsymbol{w}) = -(w_1 x_1 + w_2 x_2 + b)$$

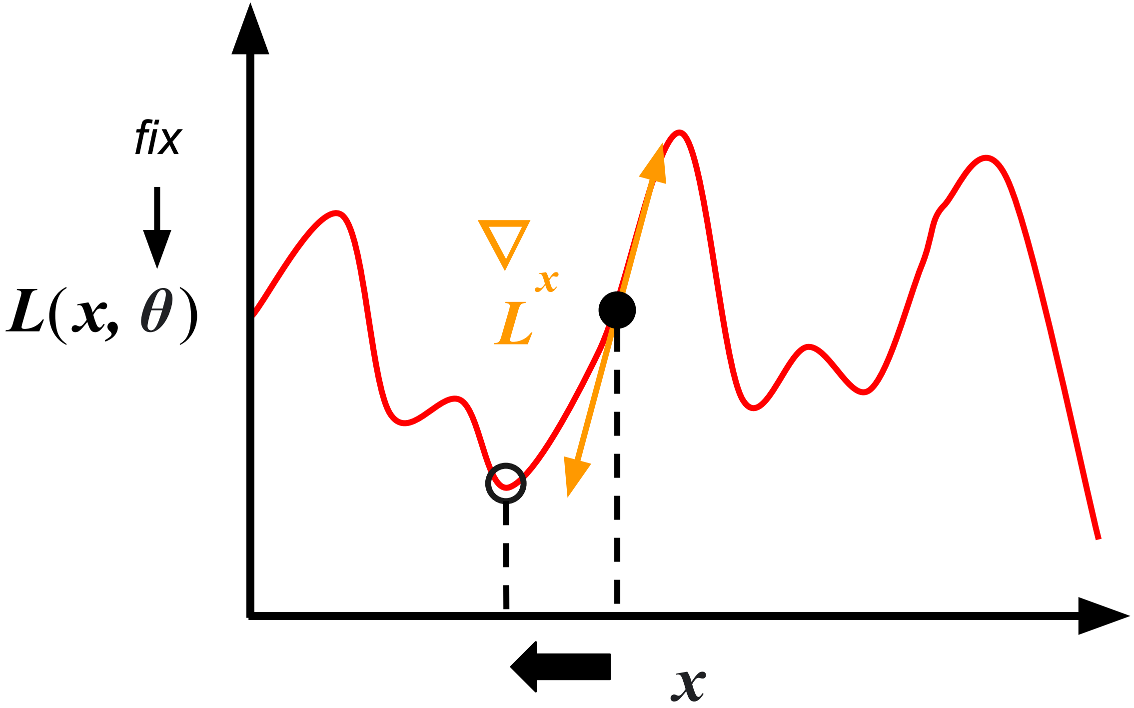
$$\nabla \boldsymbol{L}_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{w}) = -(w_1, w_2)$$

$$(w_1, w_2)$$

Adversarial Examples: View 2



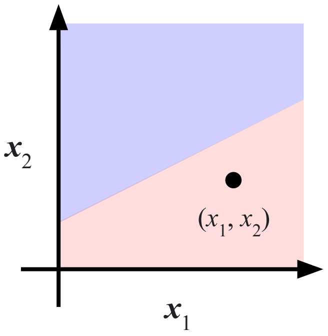
Adversarial Examples: View 2

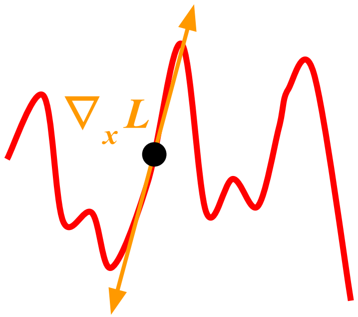


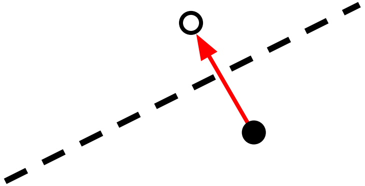
Building Intuition

If any of this gets confusing, remember:

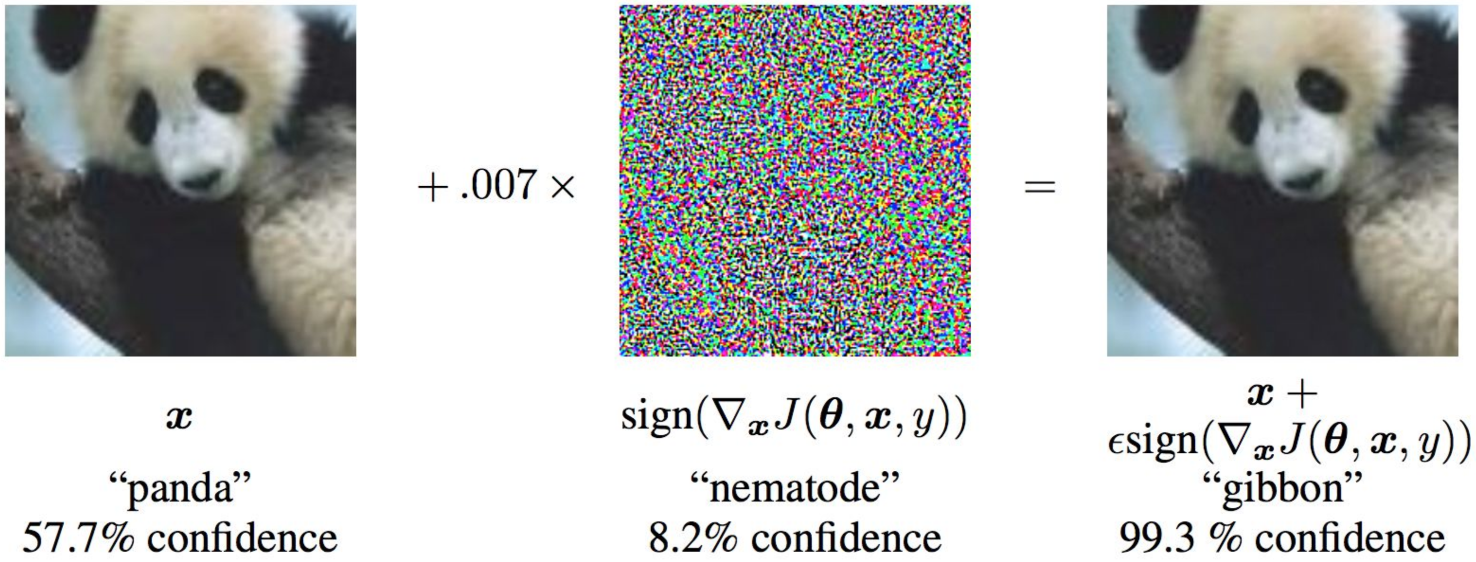
- we're modifying points in a model's *input space…*
- …using the gradients of some *loss function…*
- …such that the modified points cross a *decision boundary…*
- …while keeping the modifications as *inconspicuous* as possible







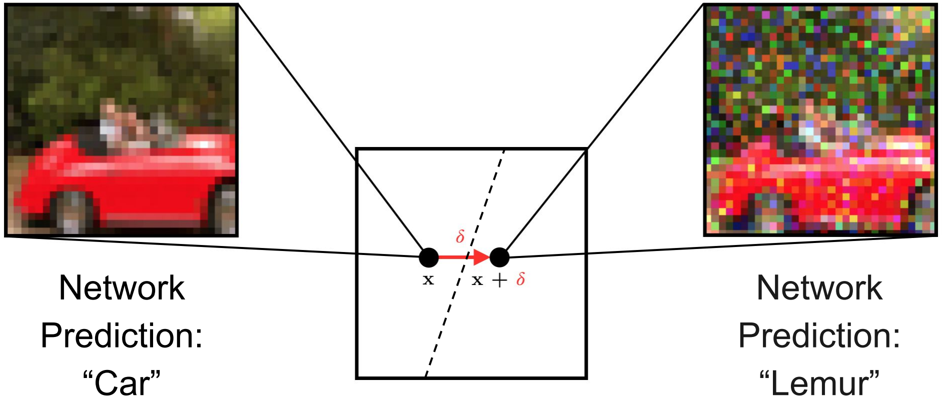
Building Intuition



(Goodfellow et al. 2014)

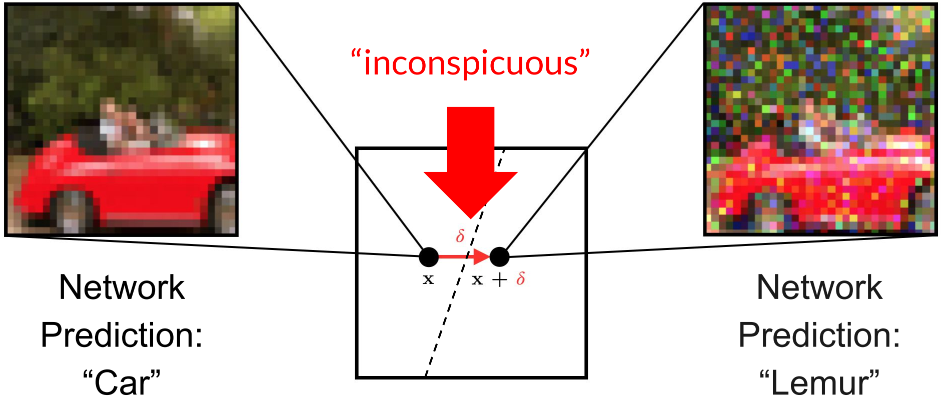
Getting Formal

In discriminative tasks such as image classification, deep neural networks have been shown to be vulnerable to **adversarial examples** - artificially-generated perturbations of natural instances that cause a network to make incorrect predictions (thereby “evading” the network).



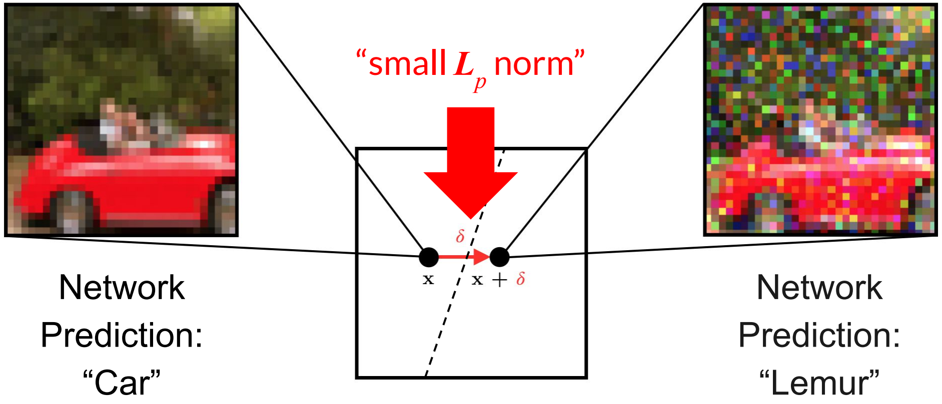
Getting Formal

In discriminative tasks such as image classification, deep neural networks have been shown to be vulnerable to **adversarial examples** - artificially-generated perturbations of natural instances that cause a network to make incorrect predictions (thereby “evading” the network).



Getting Formal

In discriminative tasks such as image classification, deep neural networks have been shown to be vulnerable to **adversarial examples** - artificially-generated perturbations of natural instances that cause a network to make incorrect predictions (thereby “evading” the network).



Getting Formal

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

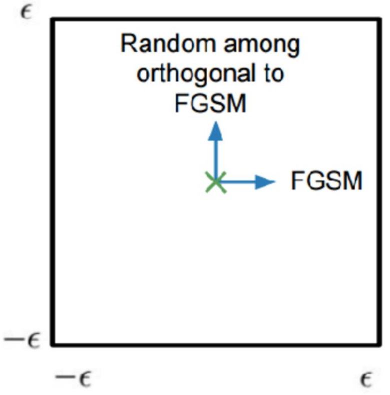
Jargon Watch

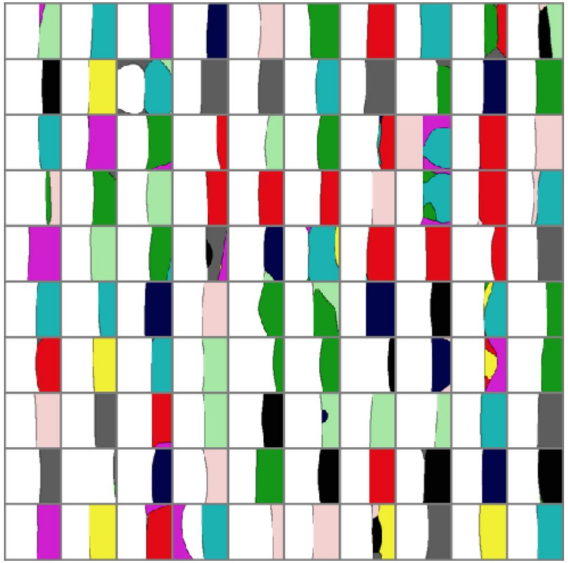
- **Attack** – an algorithm for crafting adversarial examples
- **Targeted adversarial examples** – designed to fool a model in a specific way chosen by the adversary
- **Untargeted adversarial examples** – designed to cause general misclassifications but no particular outcome
- **White-box attacks** – the adversary has complete knowledge of the model they are trying to fool
- **Black-box attacks** – the adversary has no knowledge of the model they are trying to fool

Question Time

1. **Are adversarial examples rare? That is, can most data be minimally perturbed to evade classification?**

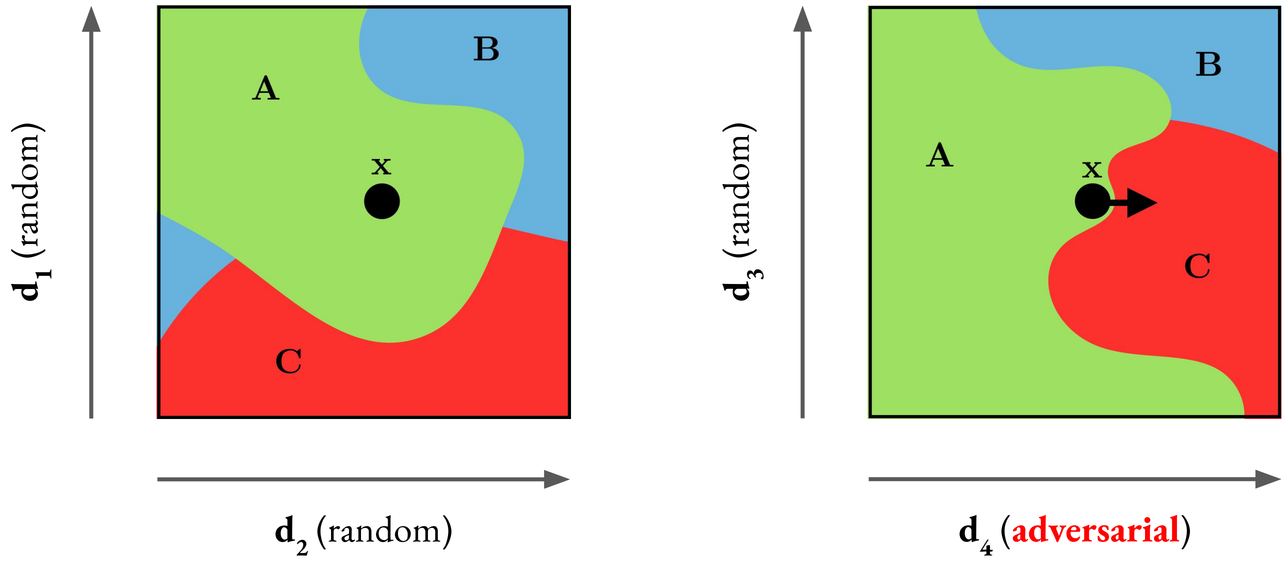
Q1: Are Adversarial Examples Rare?





(Goodfellow, Warde-Farley, Papernot 2016)

Q1: Are Adversarial Examples Rare?



Question Time

1. Are adversarial examples rare? That is, can most data be minimally perturbed to evade classification?
2. **Are adversarial examples limited to neural networks?**

Q2: Neural Networks Only?

No. Attacks have been demonstrated against:

- SVM (linear & RBF kernel)
- KNN
- Decision trees (gradient-boosted, random forests)

Gradient-free methods (e.g. decision-based attacks) rely on predictions rather than gradients and can therefore generalize beyond neural networks.

Question Time

1. Are adversarial examples rare? That is, can most data be minimally perturbed to evade classification?
2. Are adversarial examples limited to neural networks?
3. **Are adversarial examples effective if the adversary does not have white-box access to the model?**

Q3: Black-Box Attacks?

Yes. Aside from gradient-free methods, an adversary can perform a **transfer attack**:

- Obtain a neural network with “similar” architecture to victim model
- Craft white-box attacks against this **surrogate** model
- These attacks will often be effective against the real (unseen) model!

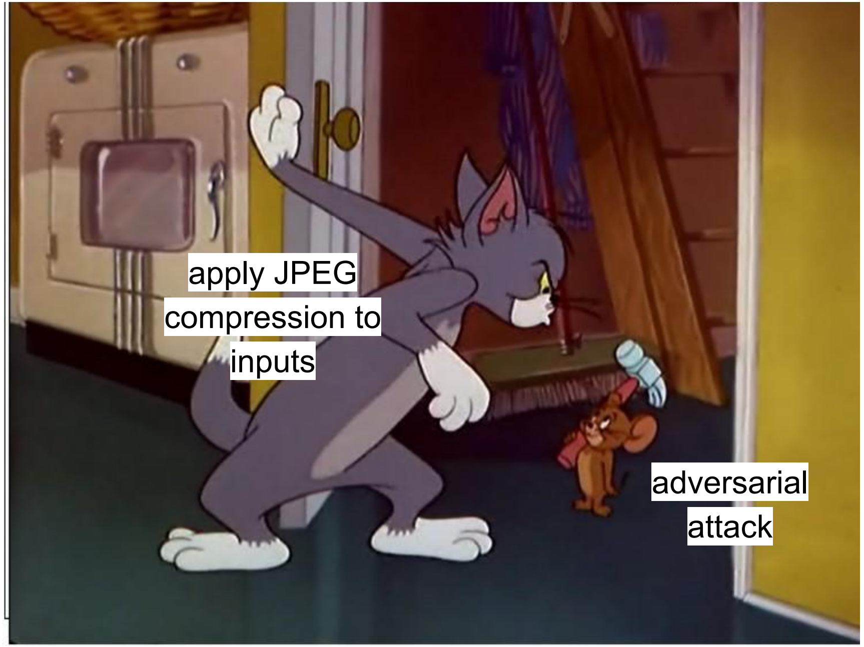
Question Time

1. Are adversarial examples rare? That is, can most data be minimally perturbed to evade classification?
2. Are adversarial examples limited to neural networks?
3. Are adversarial examples effective if the adversary does not have white-box access to the model?
4. **How can we defend against adversarial examples?**

Q4: Adversarial Defenses?

Lots of **heuristic** defenses are proposed every year, but an adversary with knowledge of a defense can often break it.

Q4: Adversarial Defenses?

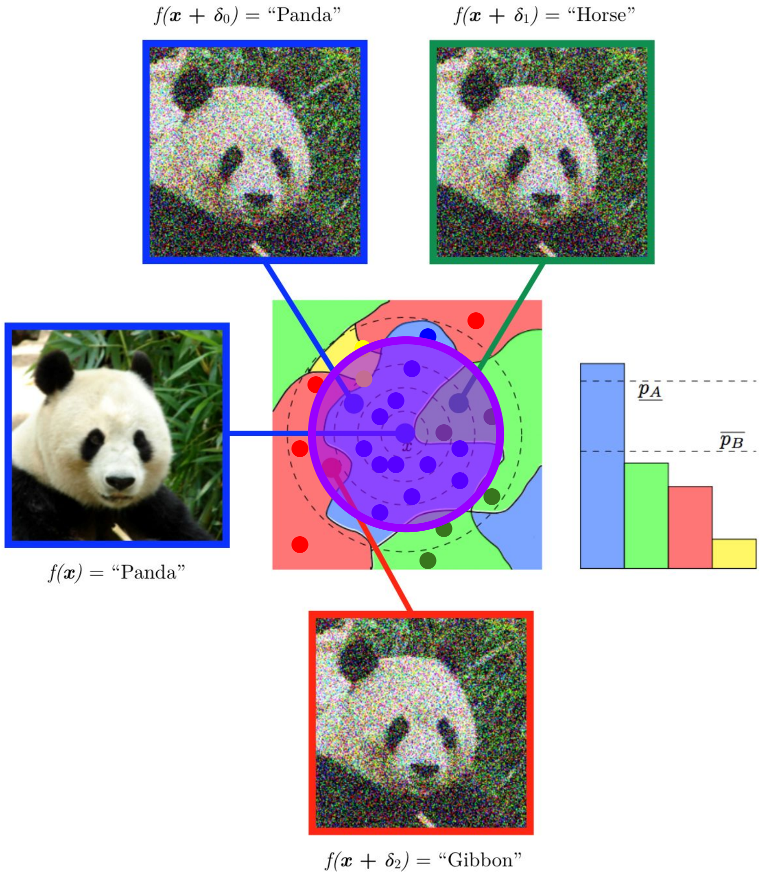


Q4: Adversarial Defenses?

Q4: Adversarial Defenses?

Certified robustness methods provide mathematically provable guarantees on the behavior of classification models.

For example, **randomized smoothing** guarantees that adversarial examples cannot exist within a certain distance of “clean” inputs



Q4: Adversarial Defenses?

However, the guarantees provided by these methods are often of little practical value

For example, randomized smoothing can only certify very small L_p radii in the input space

Question Time

1. Are adversarial examples rare? That is, can most data be minimally perturbed to evade classification?
2. Are adversarial examples limited to neural networks?
3. Are adversarial examples effective if the adversary does not have white-box access to the model?
4. ~~**How can we defend against adversarial examples?**~~

   Should we defend against adversarial examples?

Public Perception of Adversarial Examples

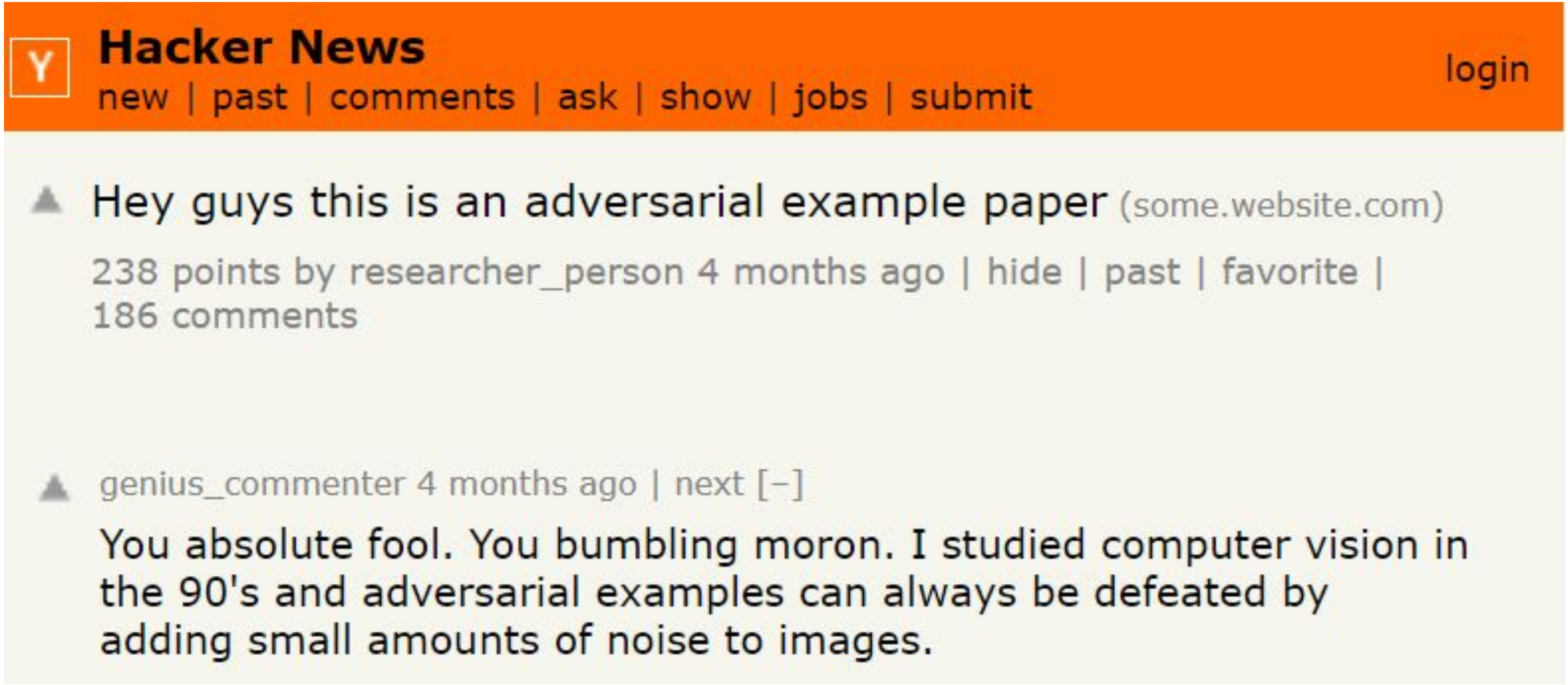
theory

I can make self-driving cars crash, defeat biometric security, and bypass content-detection systems

practice

haha that's not a gibbon it's a panda

Most ML Practitioners Don’t Care About Adversarial Examples



The Danger of Adversarial Examples Is Often Exaggerated

Y **Hacker News**
new | past | comments | ask | show | jobs | submit
login

▲ Hey guys this is an adversarial example paper (some.website.com)
238 points by researcher_person 4 months ago | hide | past | favorite | 186 comments

▲ genius_commenter 4 months ago | next [–]
Adversarial examples are literally the apocalypse and we should abandon neural nets.

Opportunity Cost

Hacker News

new | past | comments | ask | show | jobs | submit

login

Hey guys this is an adversarial example paper (some.website.com)

238 points by researcher_person 4 months ago | hide | past | favorite | 186 comments

genius_commenter 4 months ago | next [–]

I just don't think that neural networks are 'intelligent' per se. Can a neural network ever truly create a work of art? In my opinion symbolic-reasoning-based approaches to AI are the best bet for creating artificial general intelligence. In this essay I will

However…

- There are many applications in which adversarial examples pose a credible threat
- New adversarial attack algorithms can help expose vulnerabilities in neural network systems, ultimately making them more robust
- To illustrate these points, I’ll be talking about some of my recent work on adversarial attacks **in the audio domain**

Stay Tuned!

~~I can make self-driving cars crash~~, defeat biometric security, and bypass content-detection systems

Audio Adversarial Examples

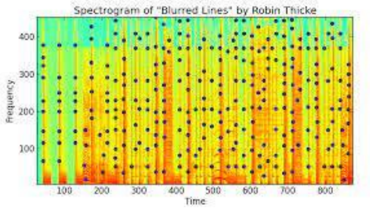
Neural Networks Power Audio Interfaces

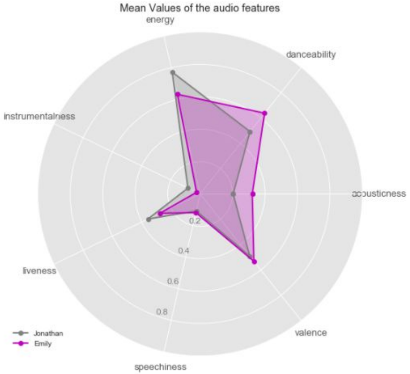
Voice-based machine-learning systems for authentication and control are common in products such as mobile devices, vehicles, and household appliances.

Neural Networks Power Audio Interfaces

Machine learning is also prevalent in audio analysis tasks such as **copyright detection.**

YouTube Content ID



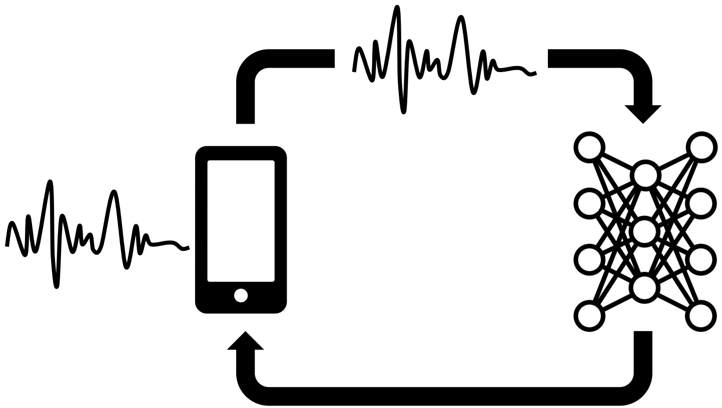


LA cops tried using Instagram's copyright filter to stop someone from filming them

Beverly Hills Police can be seen playing songs from Sublime and The Beatles.

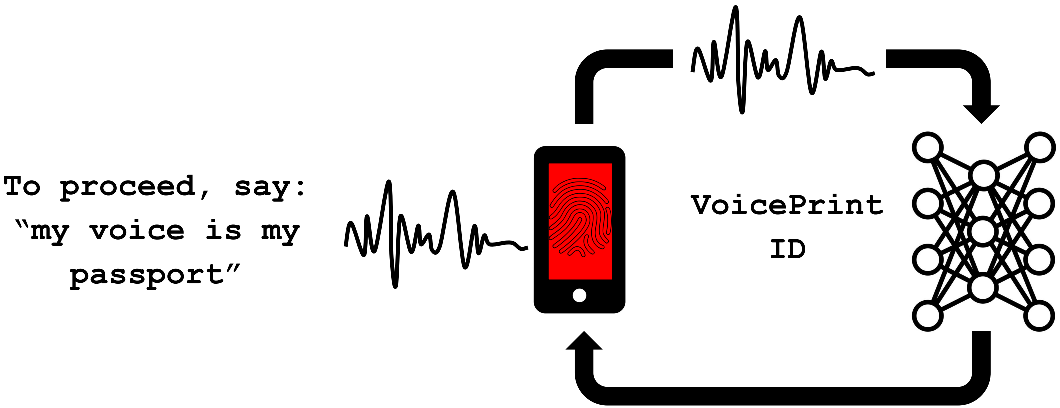
Neural Networks Power Audio Interfaces

In many applications, user-supplied audio is passed to a remote neural network system for prediction.



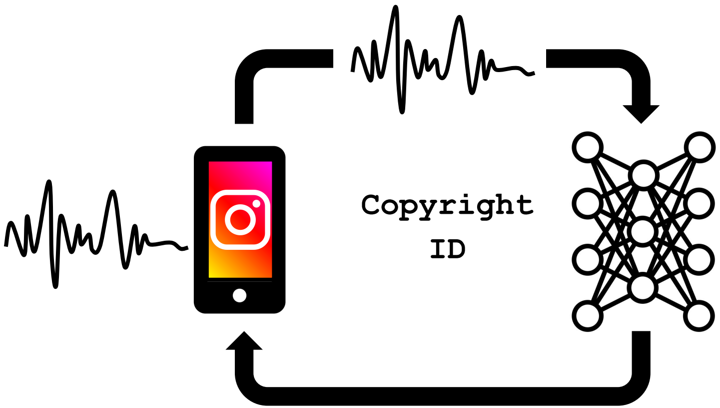
Neural Networks Power Audio Interfaces

“Voiceprint” authentication can be used to screen VoIP transactions in mobile banking applications



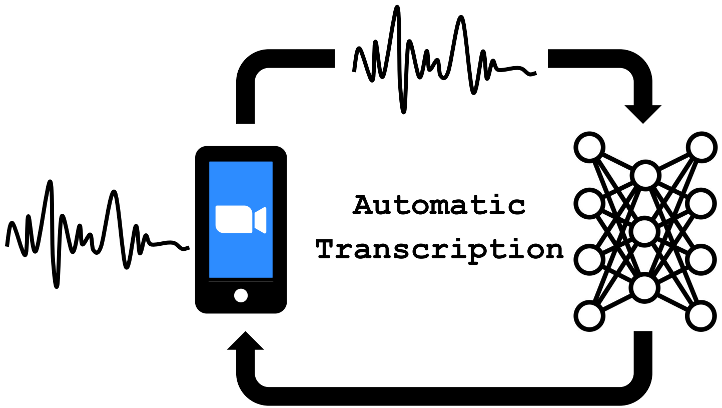
Neural Networks Power Audio Interfaces

Live-streaming applications can flag content for suspected copyright infringement by running algorithms to match audio against a database



Neural Networks Power Audio Interfaces

Video-conferencing software can automatically transcribe user speech



Evading Audio Interfaces

Many different parties may be interested in **evading** (or fooling) such systems.

😈 **Malicious actors** may wish to bypass an authentication system by impersonating a verified user

🔒 **Privacy-minded individuals** may wish to avoid eavesdropping and automatic transcription from an application, or to confuse a content-detection system

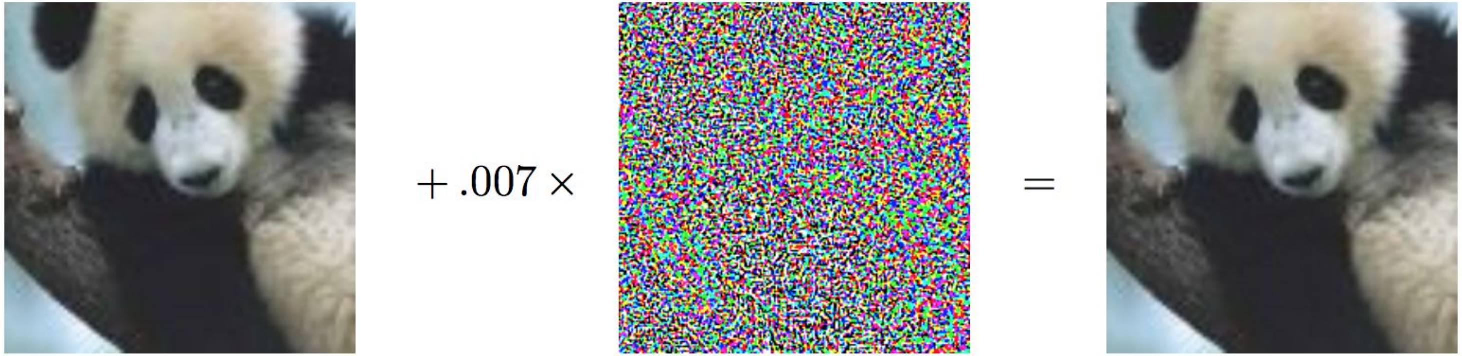
👩‍💻 **System designers** may wish to understand the vulnerabilities of these systems by finding ways to fool them

But How?

Adversarial Examples Fool Neural Networks

Neural networks are known to be vulnerable to **adversarial examples –** inputs that have been *slightly* altered to force incorrect predictions

Adversarial Examples Fool Neural Networks

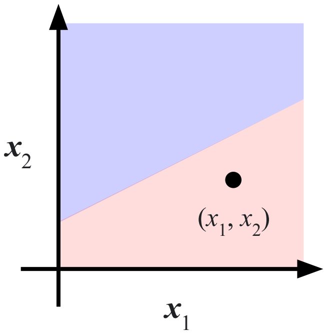


$+ .007 \times$ … $=$

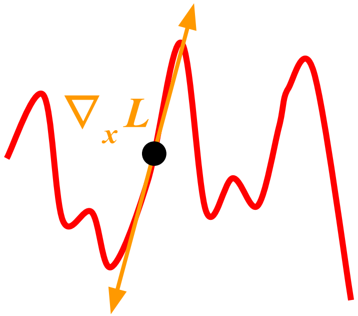
$\boldsymbol{x}$	$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$	$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"panda"	"nematode"	"gibbon"
57.7% confidence	8.2% confidence	99.3 % confidence

(Goodfellow et al. 2014)

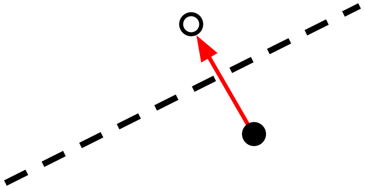
But How?



x_2
(x_1, x_2)
x_1



$\nabla_x L$

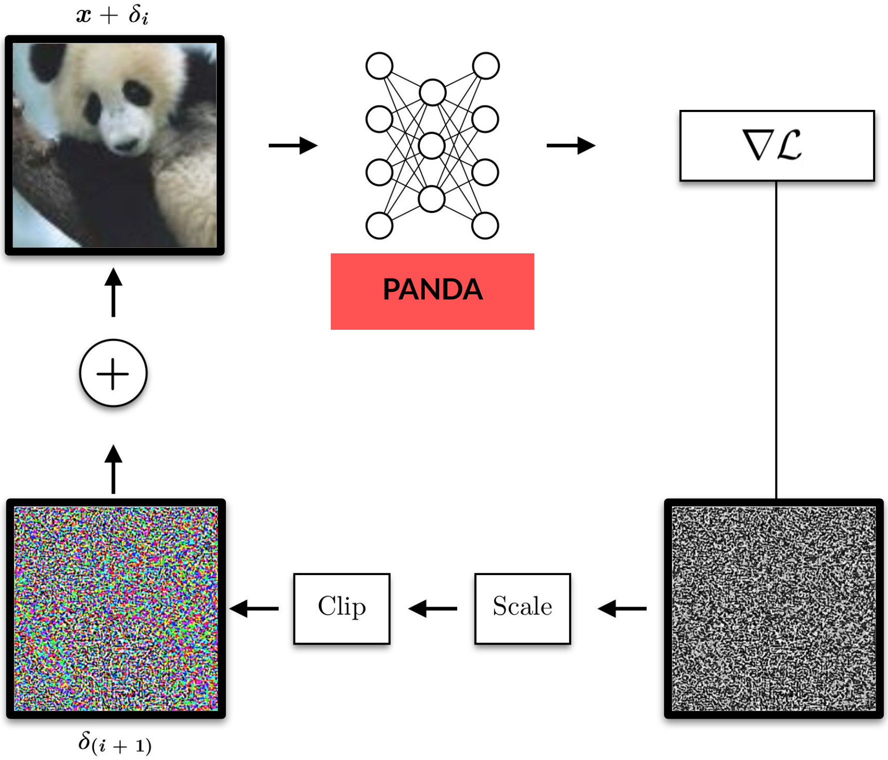


Adversarial Examples Fool Neural Networks

$f(x + \delta_0)$ = PANDA

$f(x + \delta_1)$ = PANDA

$f(x + \delta_{...})$ = PANDA



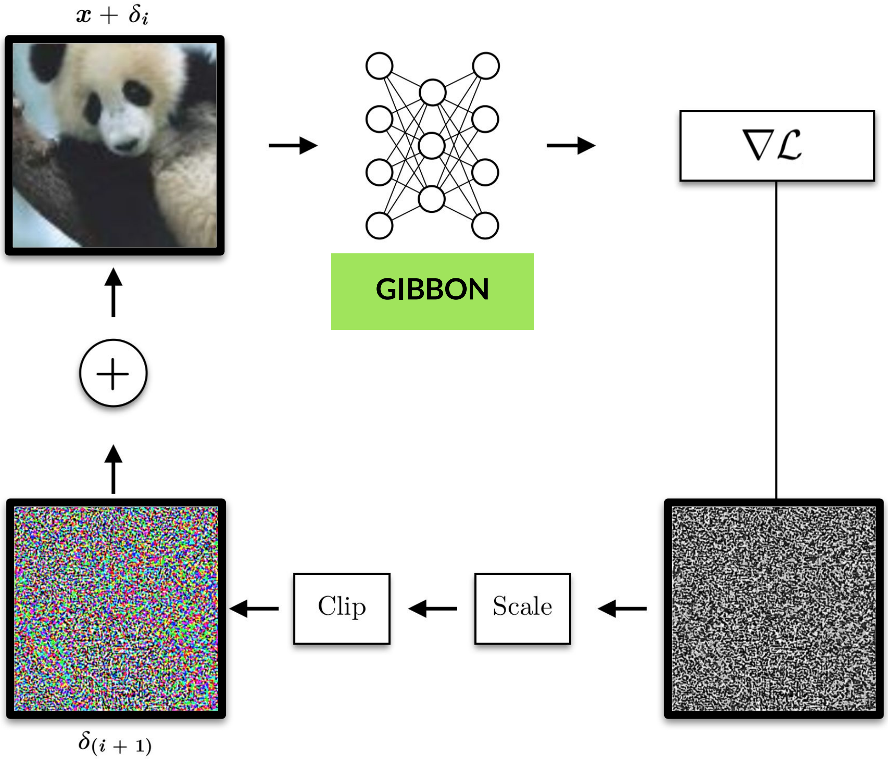
Adversarial Examples Fool Neural Networks

$f(x + \delta_0)$ = PANDA

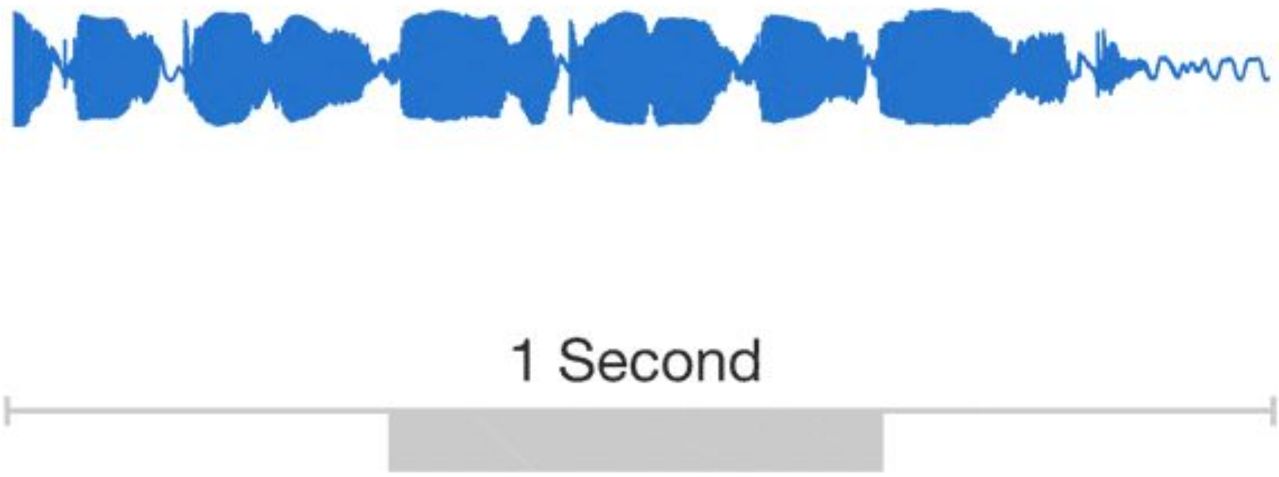
$f(x + \delta_1)$ = PANDA

$f(x + \delta_{\dots})$ = PANDA

$f(x + \delta_{final})$ = GIBBON



Moving to Audio



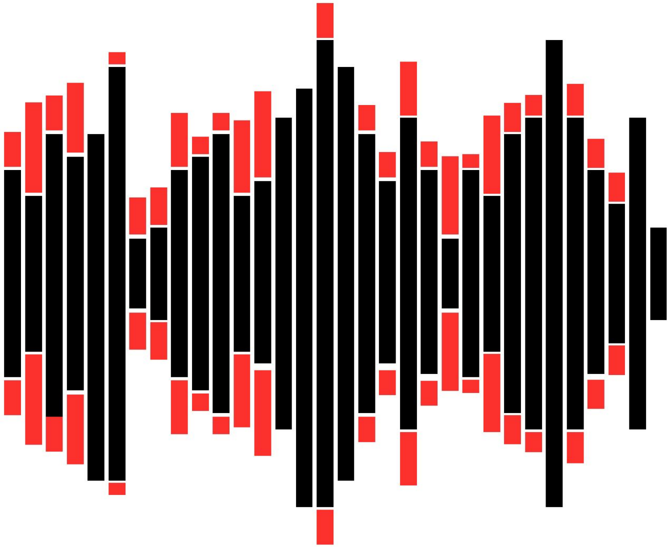
(Oord et al. 2016)

Moving to Audio

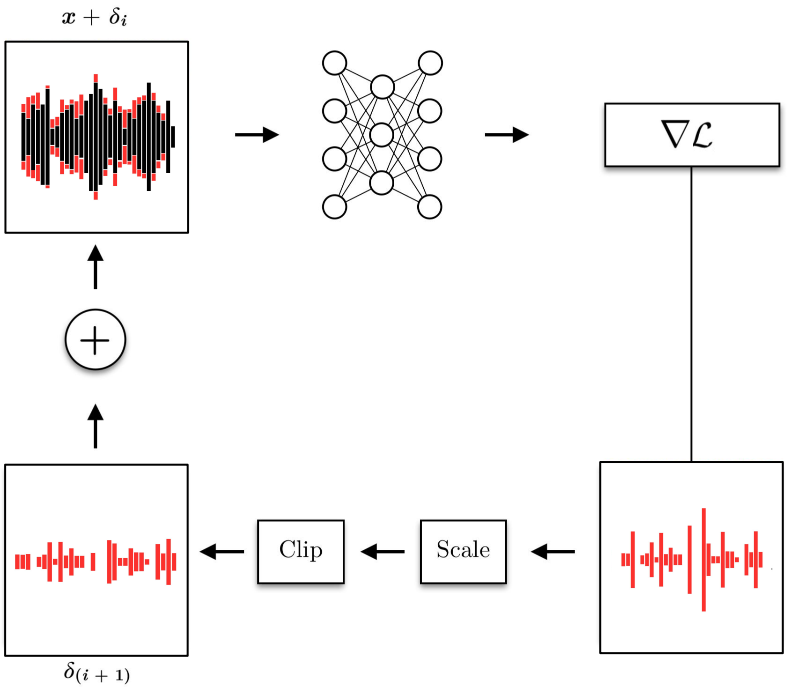
x

Moving to Audio

$x + \delta$



Attacking with Waveform Perturbations



Attacking with Waveform Perturbations

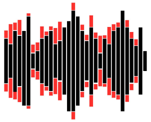
original

+

10% scale!

waveform perturbation

=



adversarial

Attacking with Waveform Perturbations

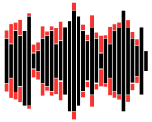
1% scale, with spectral loss!

original

\+

waveform perturbation

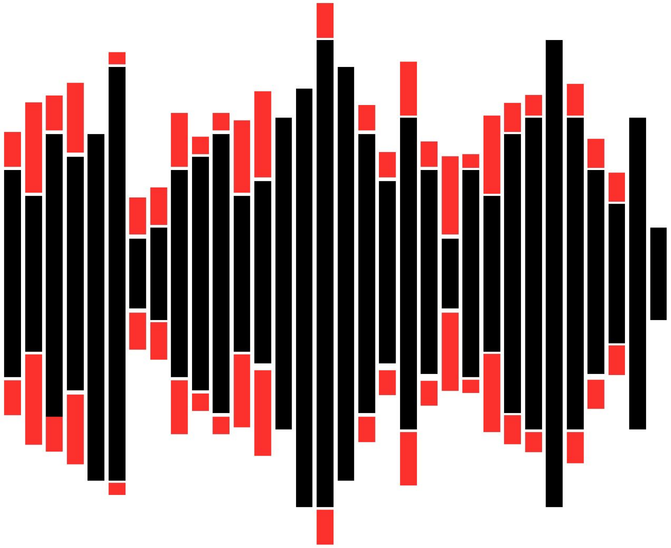
=



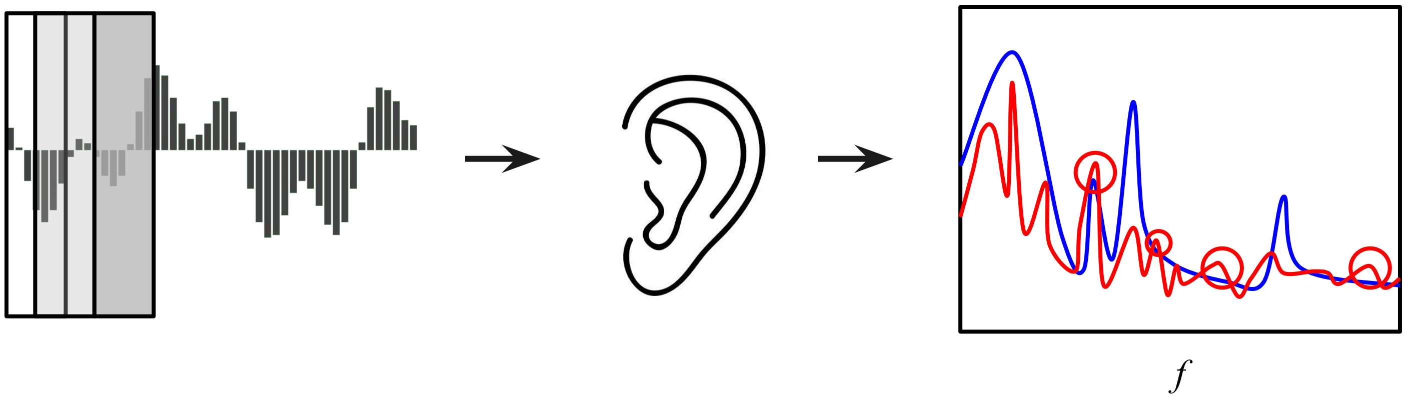
adversarial

Additive Attacks Introduce Noise

$x + \delta$

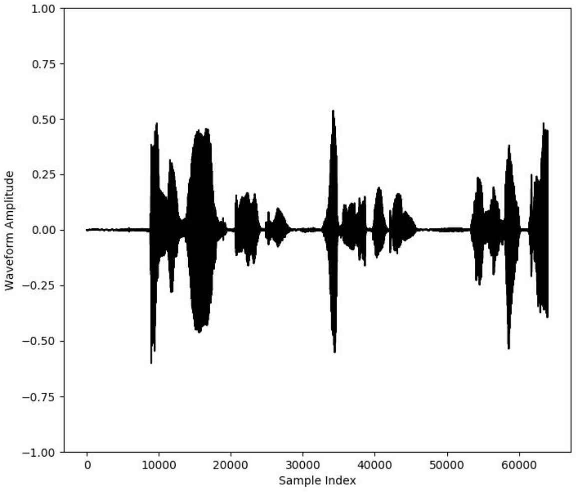


A “Perceptual” Frequency-Masking Loss

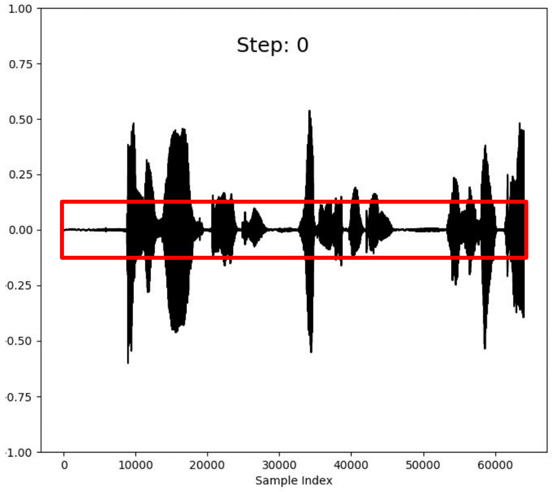


Attacking with Waveform Perturbations

Original Audio

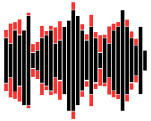


Qin et al.*



Attacking with Waveform Perturbations

1% scale, with
“perceptual” loss!



original

+

waveform
perturbation

=

adversarial

Effective and Inconspicuous Over-the-Air Adversarial Examples with Adaptive Filtering

Patrick O'Reilly[1], Pranjal Awasthi[2], Aravindan Vijayaraghavan[1], Bryan Pardo[1]

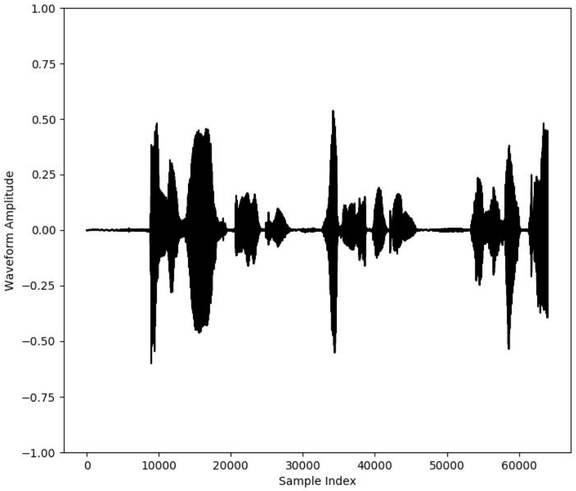
ICASSP '22

1. Northwestern University
2. Google Research

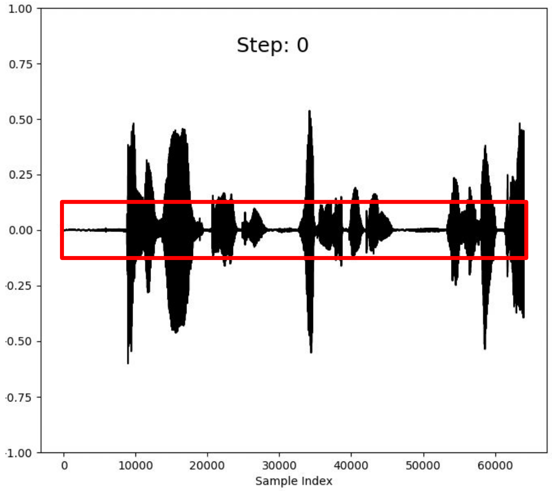
interactiveaudiolab.github.io/project/audio-adversarial-examples.html

Beyond Waveform-Additive Perturbations

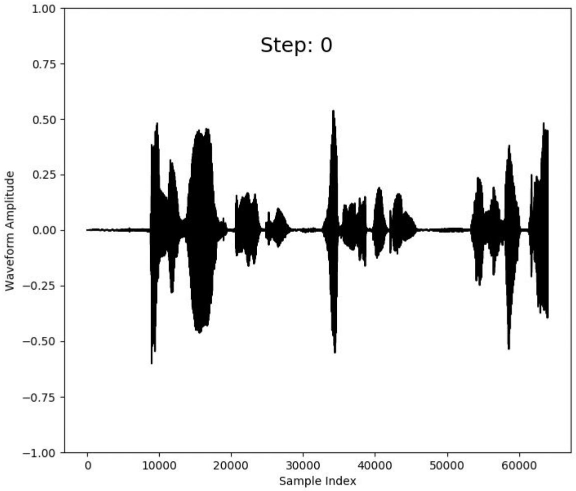
Original Audio



Qin et al.*



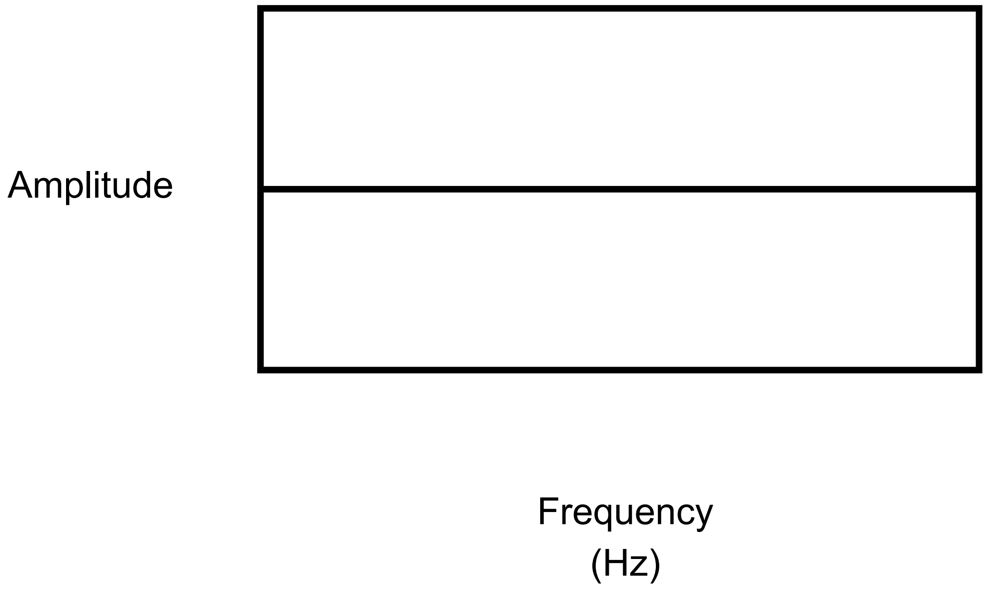
Proposed



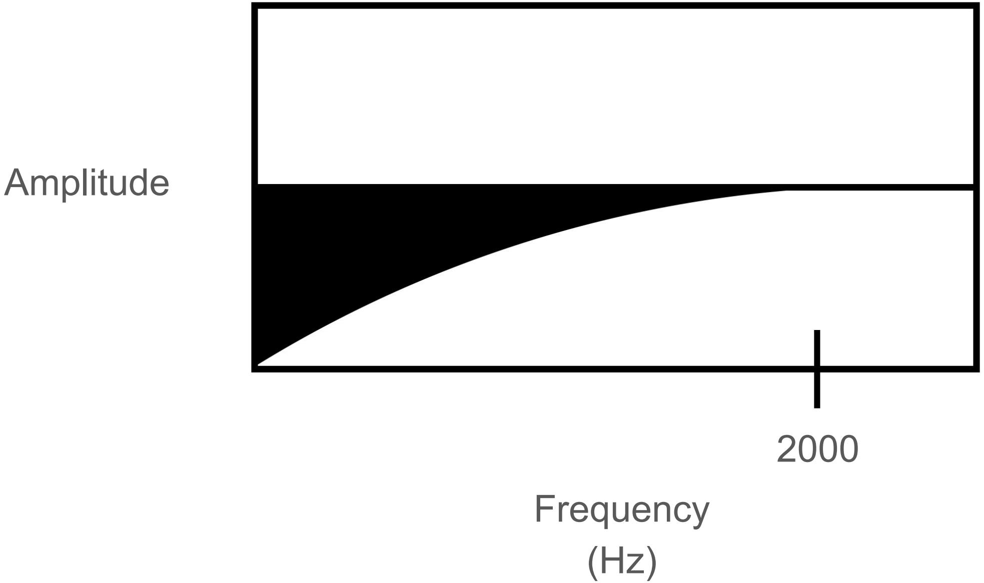
Our proposal: a **different way of modifying audio**, so we don’t need a complicated scheme to conceal attacks


9V
TREBLE
MID
BASS

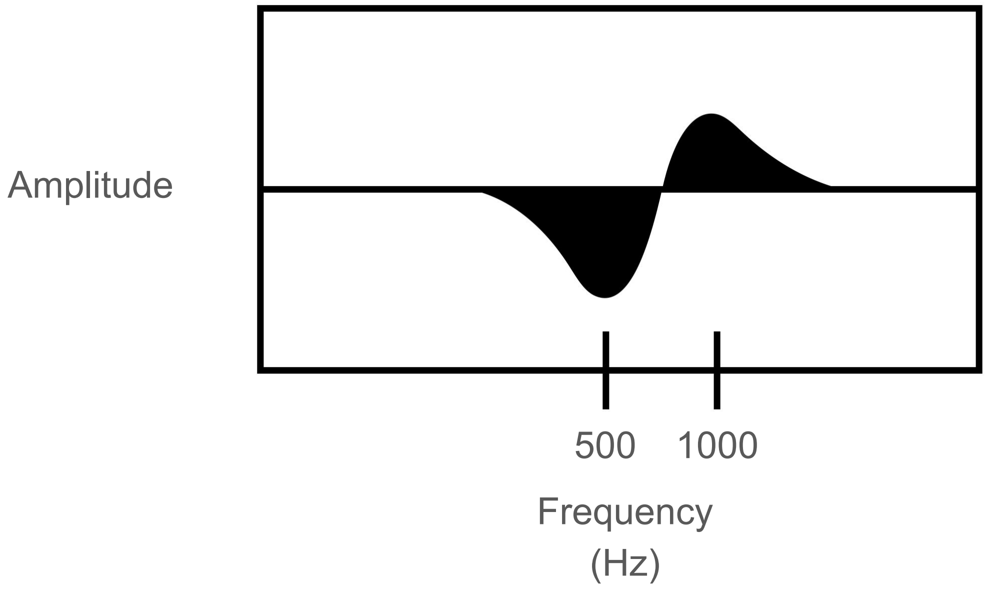
Filters Let Us Shape Frequency Content



Filters Let Us Shape Frequency Content

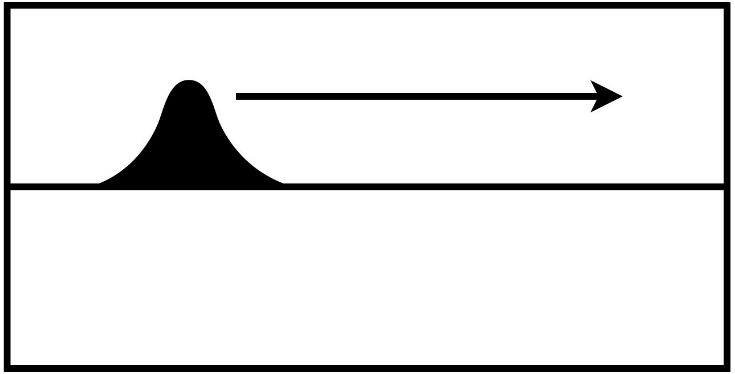


Filters Let Us Shape Frequency Content



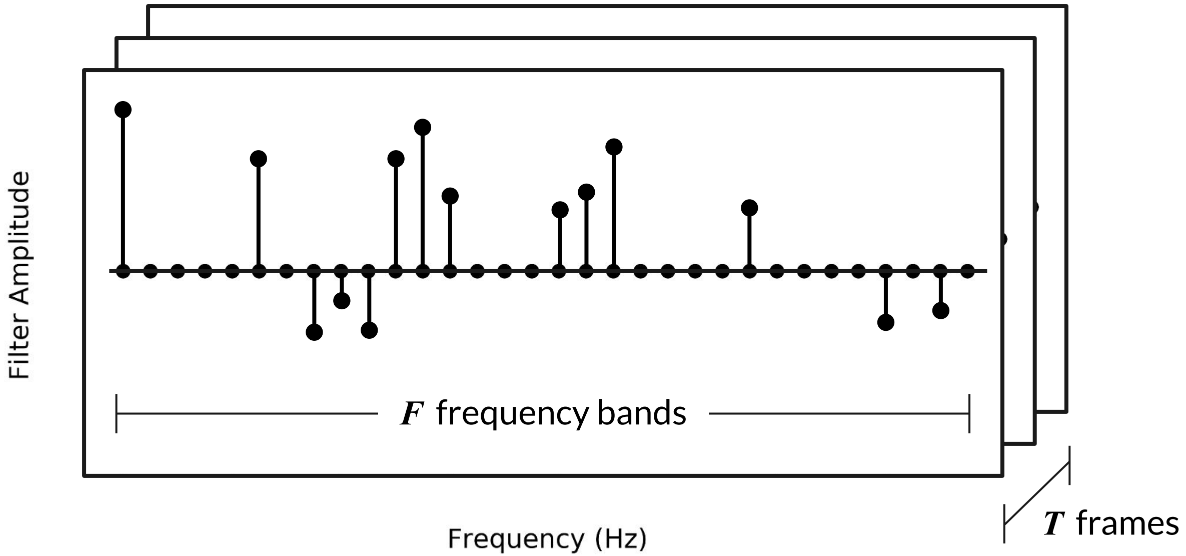
Filters Let Us Shape Frequency Content

Amplitude

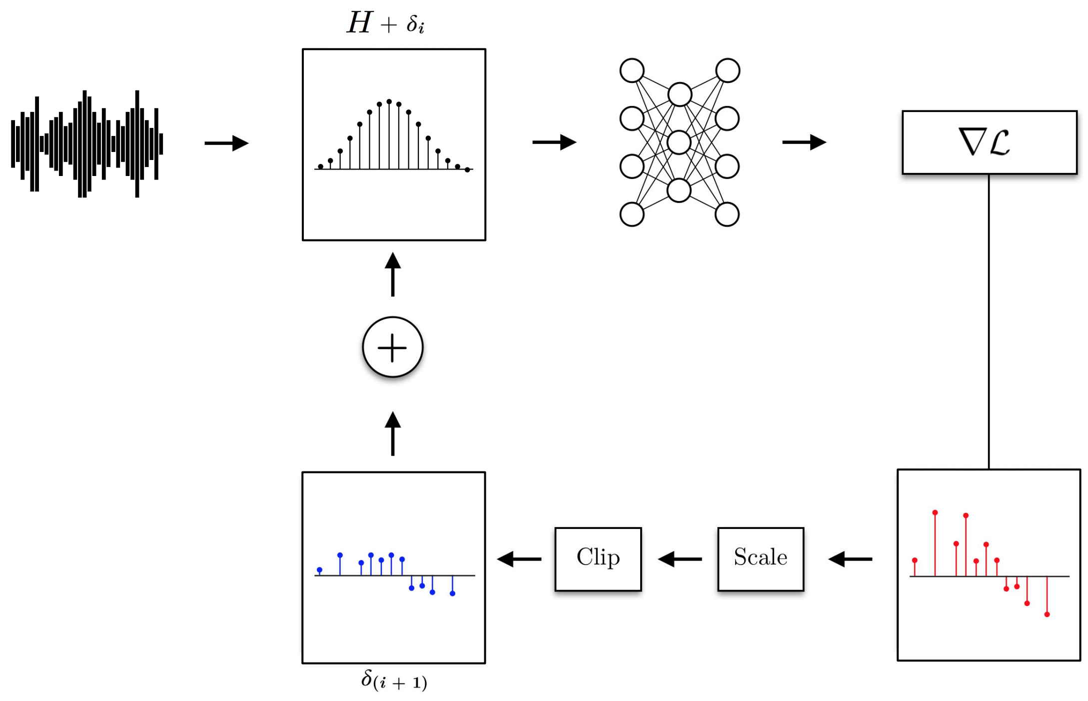


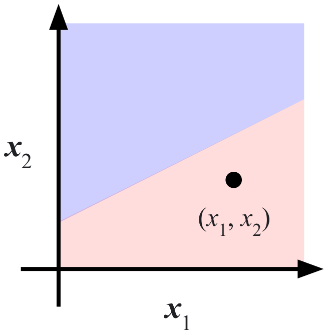
Frequency
(Hz)

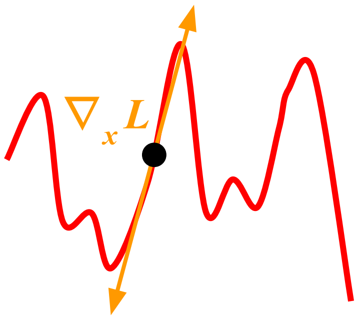
Filters Let Us Shape Frequency Content

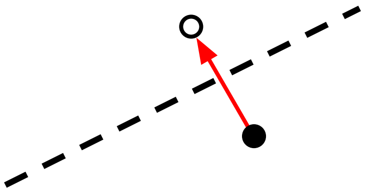


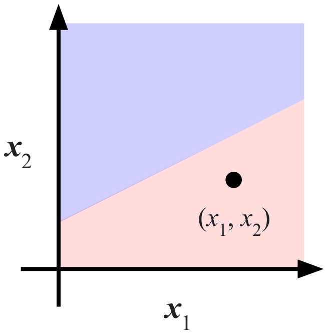
Attacking with Filter Perturbations



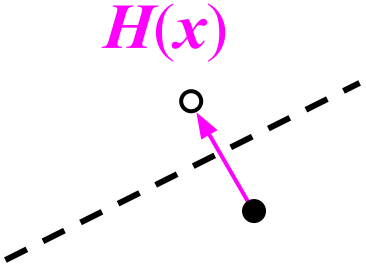

x_2
(x_1, x_2)
x_1


$\nabla_x L$



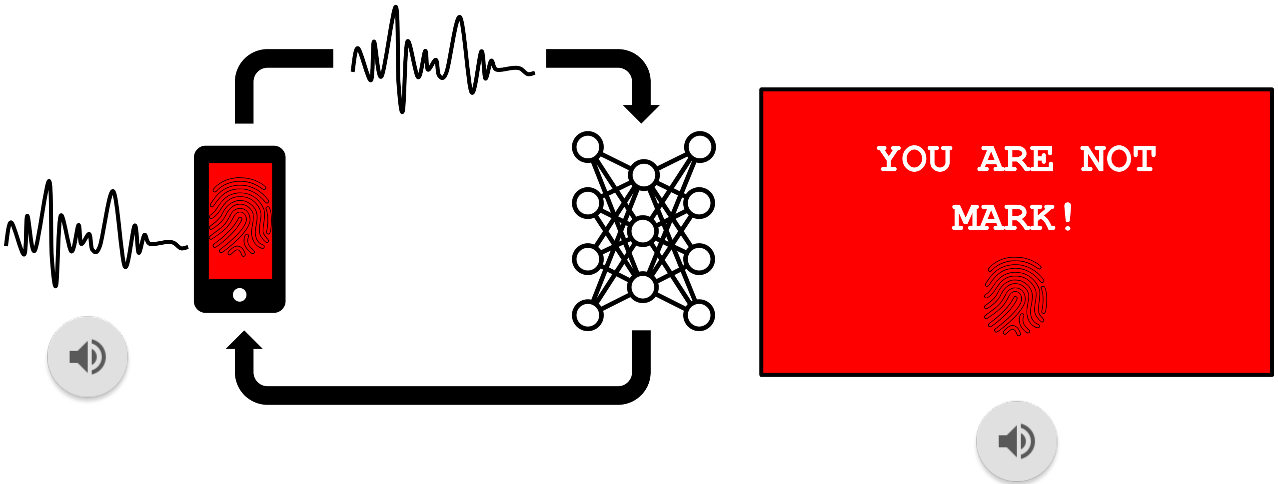


$$\nabla_H L$$



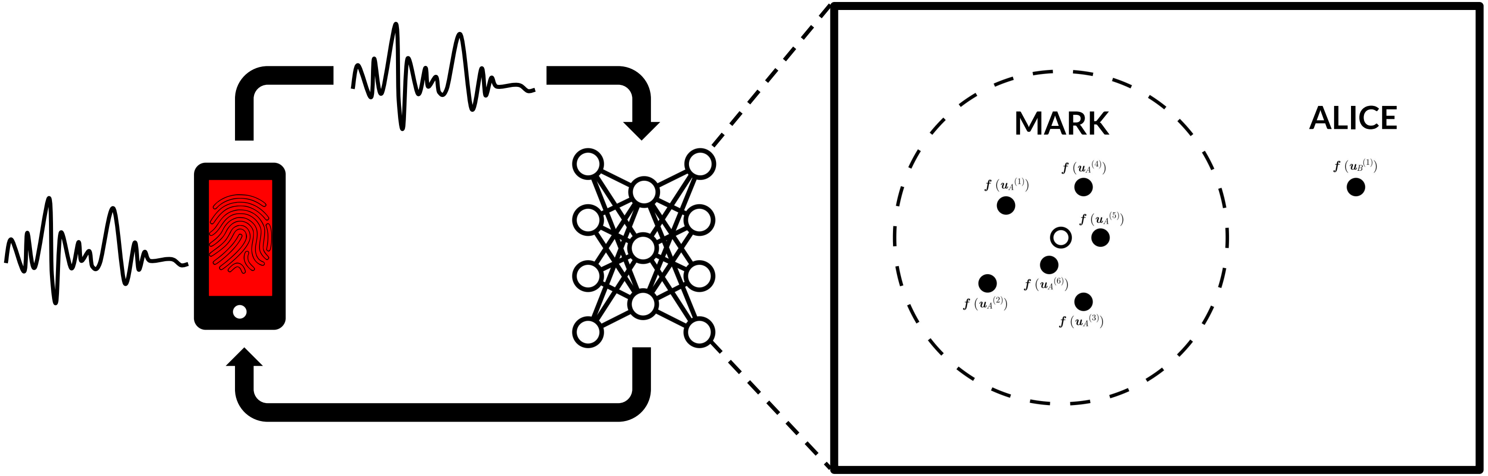
Attacking Speaker Verification

We explore impersonation attacks on a **speaker-verification system**.



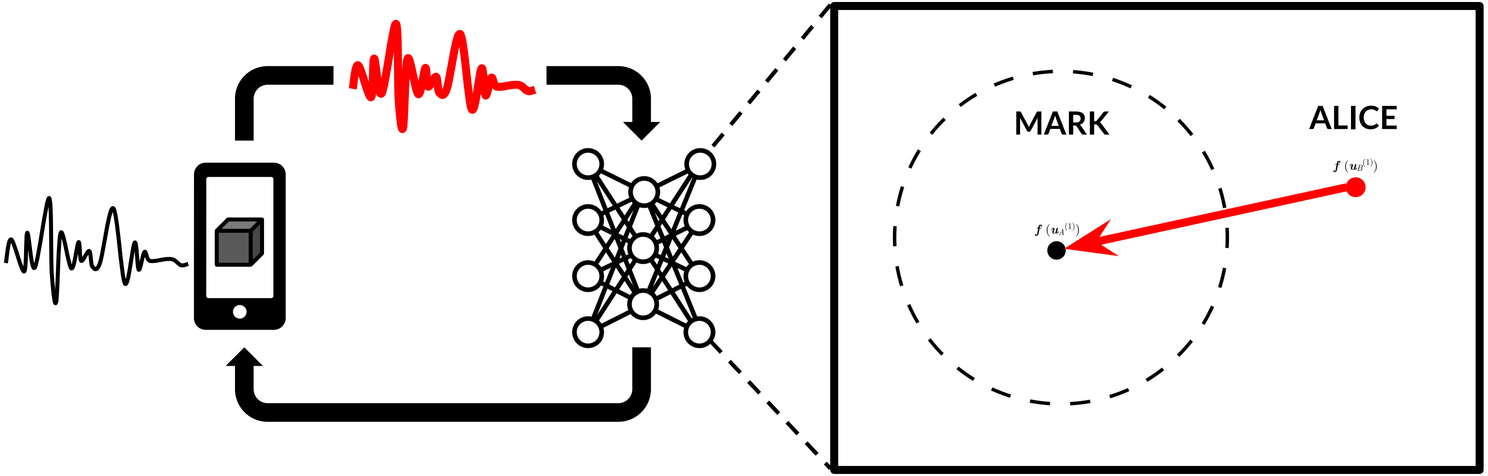
Attacking Speaker Verification

We explore impersonation attacks on a **speaker-verification system**.



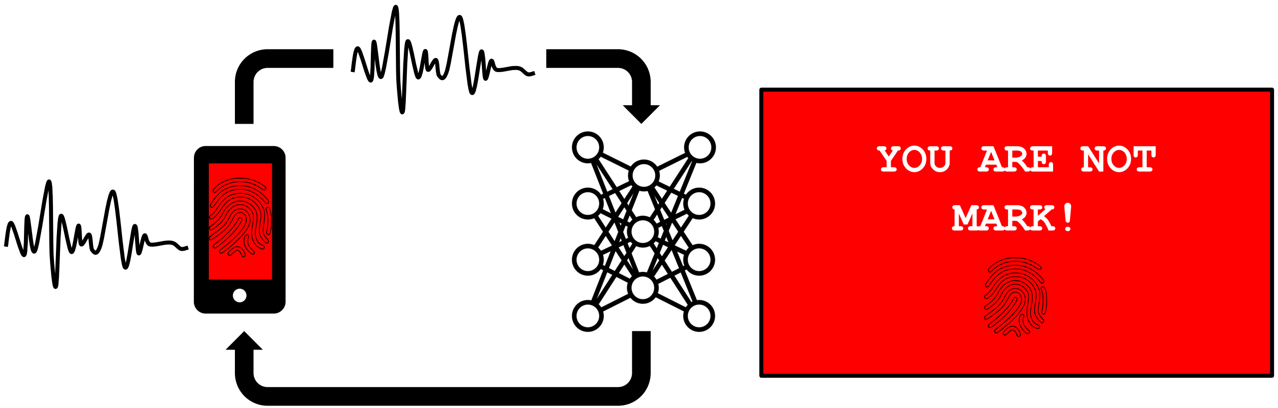
Attacking Speaker Verification

We explore impersonation attacks on a **speaker-verification system**.



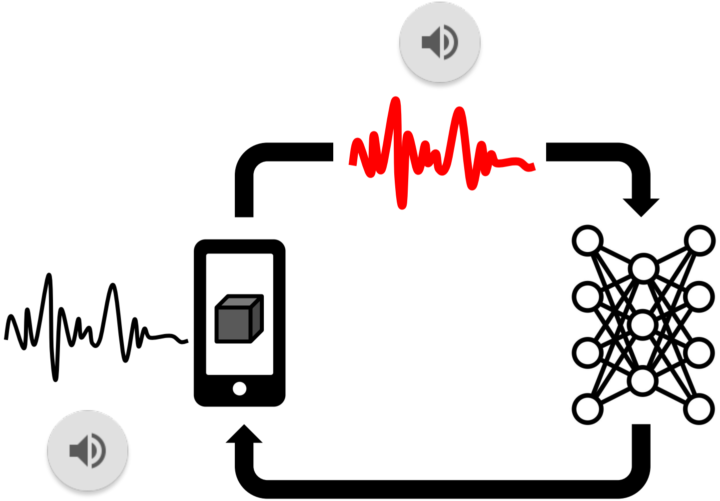
Attacking Speaker Verification

We explore impersonation attacks on a **speaker-verification system**.



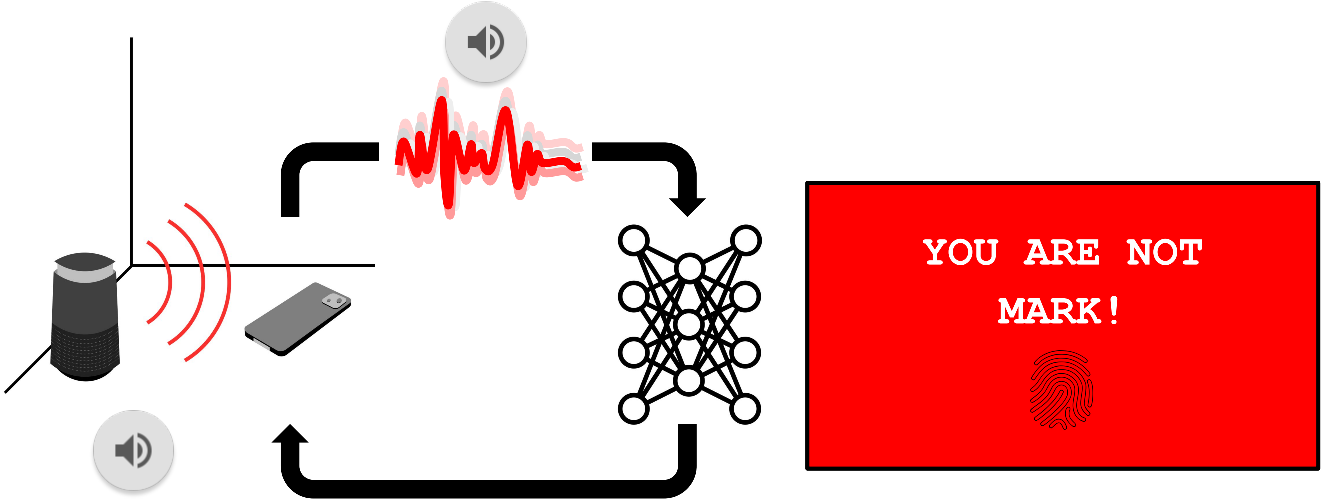
Attacking Speaker Verification

We explore impersonation attacks on a **speaker-verification system**.



Attacking Speaker Verification

We focus on a challenging **over-the-air setting**



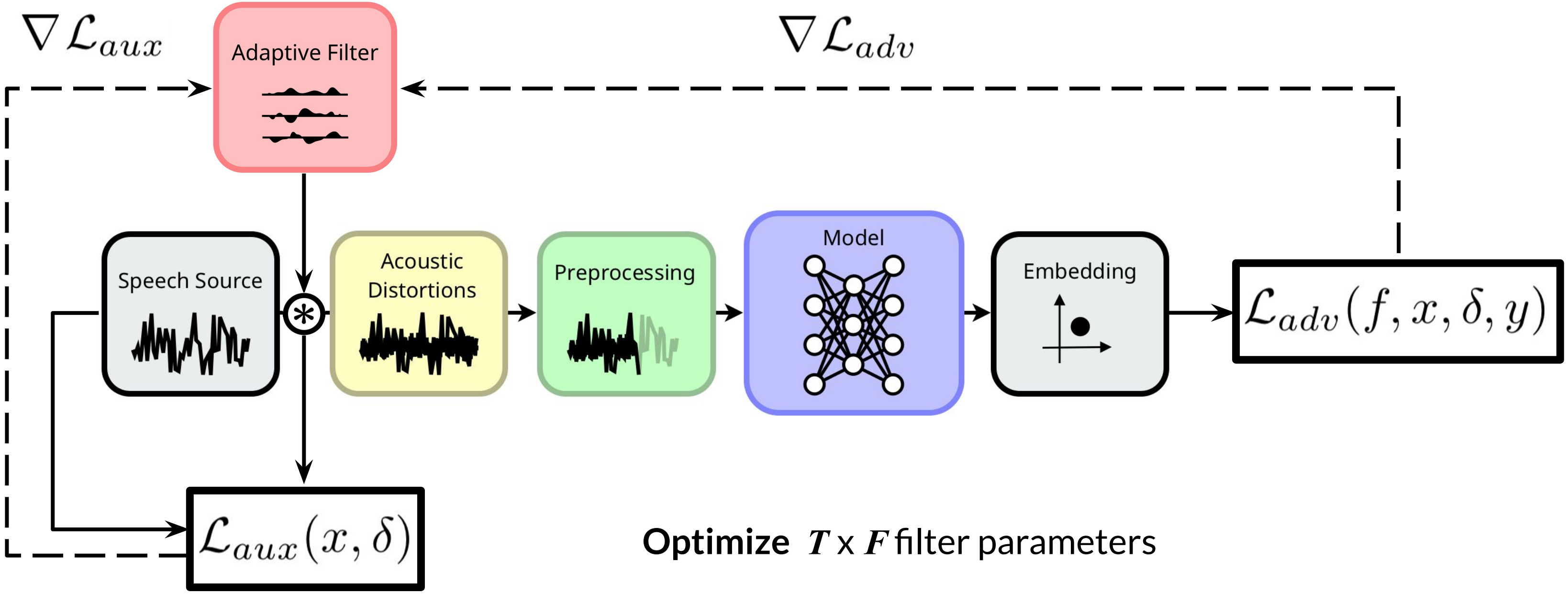
Over-the-Air Attacks Are Harder to Conceal

Qin et al. (2019): speech recognition

Li et al. (2020): speaker recognition

Chen et al. (2020): speech recognition

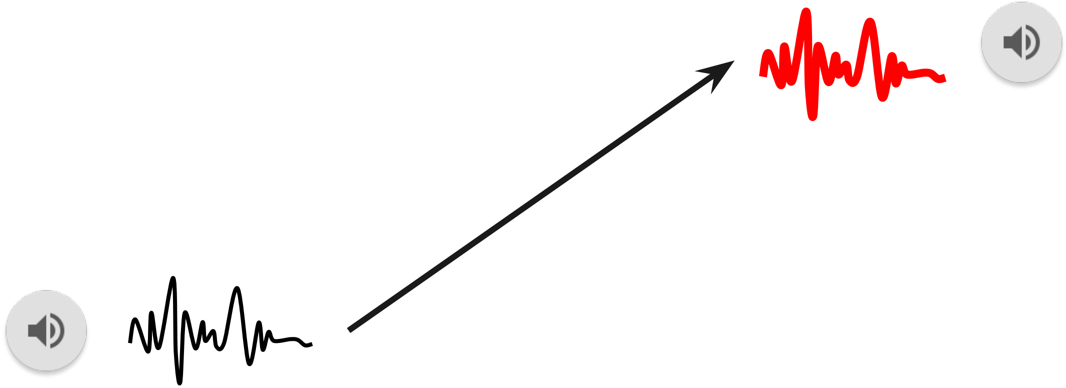
Over-the-Air Simulation



Optimize T x F filter parameters

(T frames and F frequency bands per frame)

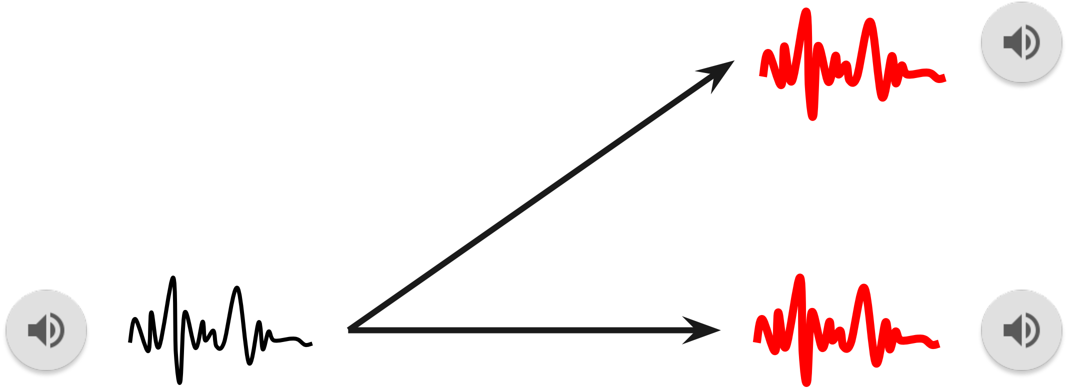
Stealth, Simplicity & Success



“Generic”

+ 89% effective
– easy to hear

Stealth, Simplicity & Success



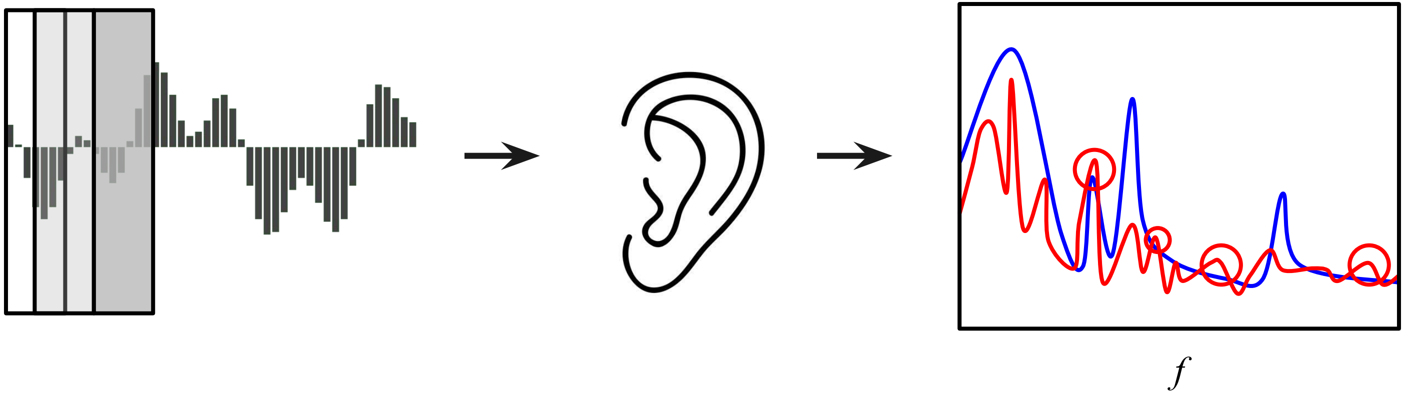
“Generic”

+ **89%** effective
– easy to hear

Qin et al.*

+ **93%** effective
+ hard to hear
– computationally expensive

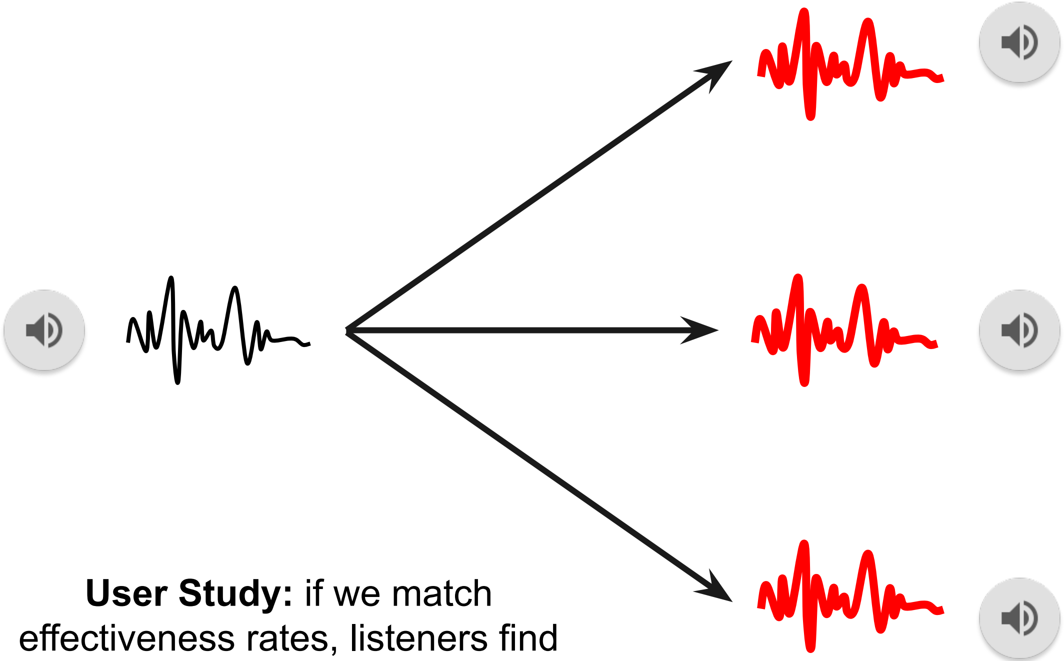
Over-the-Air Attacks Are Harder to Conceal



Two-stage frequency-masking attack: Qin et al. (2019), Szurley & Kolter (2019), Dörr et al. (2020), Wang et al. (2020)

😔 State-of-the-art approach for concealing attacks is expensive

Stealth, Simplicity & Success



“Generic”

+ **89%** effective
– easy to hear

Qin et al.*

+ **93%** effective
+ hard to hear
– computationally expensive

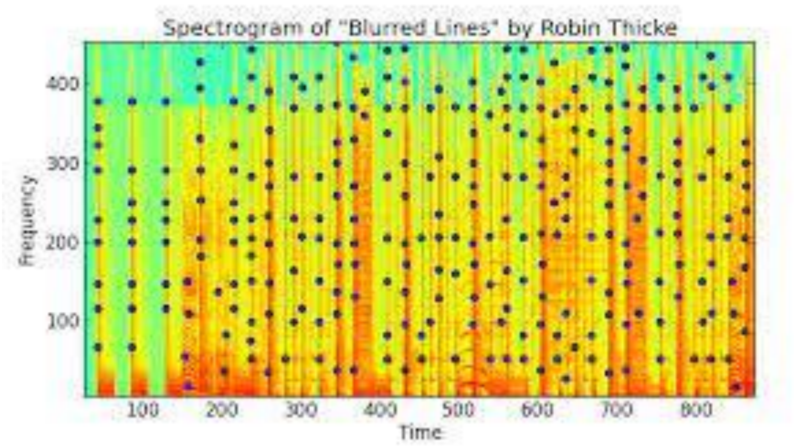
Adaptive Filtering (Ours)

+ **95%** effective
+ hard to hear
+ efficient

User Study: if we match effectiveness rates, listeners find our attack less conspicuous that Qin et al.* by a 2-to-1 margin

Beyond Authentication Attacks

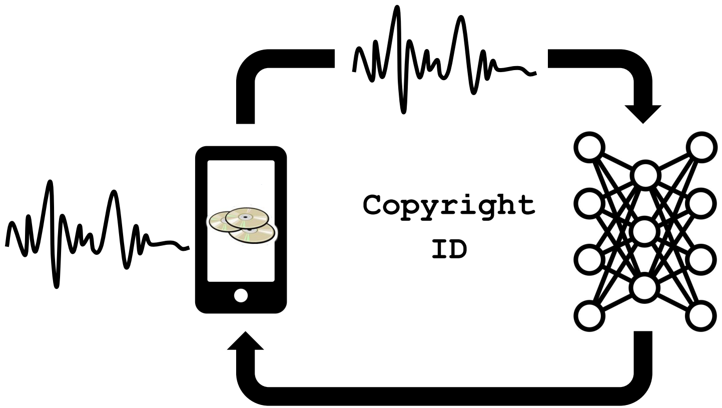
Our main contribution is a new method of perturbing audio adversarially. **We can apply it to any arbitrary task or victim model.**



For example, **copyright identification.**

Copyright Identification

We can perform a **transfer attack** on the *AudioTag* copyright-identification service by building an approximation of its underlying model.



We model our attack on the method of Saadatpanah et al., but because we use filters, our attacks are **noise-free**. See: https://www.cs.umd.edu/~tomg/projects/copyrightattack/

That’s All For Now!

- On Friday, we’ll learn how to code image and audio adversarial attacks!
- If you have any questions about the research, feel free to reach out!
  patrick.oreilly2024@u.northwestern.edu
- Currently working on more cool adversarial stuff with NU students Andreas Bugler (a PM for this course!) and Keshav Bhandari