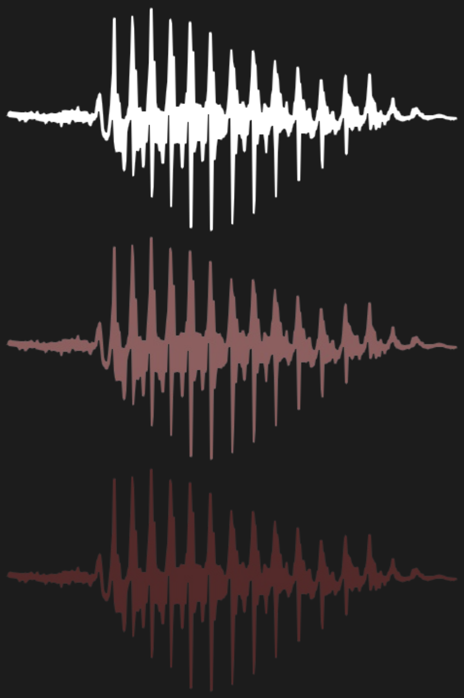
Adversarial Attacks

on Audio Speech Recognition systems

LINKMEDIA
TEAM

Student : Yoann Lemesle
Supervisor : Laurent Amsaleg



Deep Learning & applications

- **Machine Learning**
  ML

  Learning **automatically** from a set of data to perform a task without explicit programmation.

- **Deep Learning**
  DL

  Methods of ML that use **neural networks**.

NEURAL NETWORK

0.01 | **ASSAULT RIFLE**
0.05 | **HORSE**
0.04 | **SHEEP**
0.11 | **DUCK**
0.08 | **FROG**
0.03 | **DOG**
0.84 | **CAT**

Input
x

Model
f

Predictions
f(x)

Deep Learning & security

- **Adversarial Example** | A **seemingly benign** input that fools a neural network.

→ **NEURAL NETWORK** →

0.99 | **ASSAULT RIFLE**
0.05 | **HORSE**
0.04 | **SHEEP**
0.11 | **DUCK**
0.08 | **FROG**
0.03 | **DOG**
0.02 | **CAT**

Input
x

Model
f

Predictions
f(x)

Neural Networks

- **Powerful graphs**

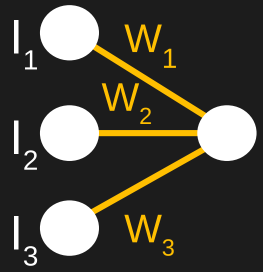
  Weighted / Directed

Layers of **neurons** with **weighted connections**.

Acts as an **ajustable** function f(**θ**, **x**)

- **Forward** propagation

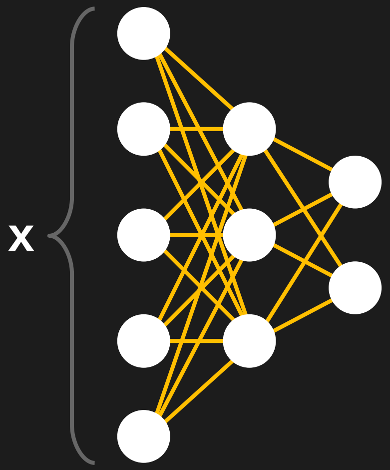
  How information is processed.



$$out = \Phi\left(I_1 * W_1 + I_1 * W_2 + I_1 * W_3 + Bias \right)$$

- **Back** propagation

  How a neural network is trained.



predictions f(**θ**, **x**)

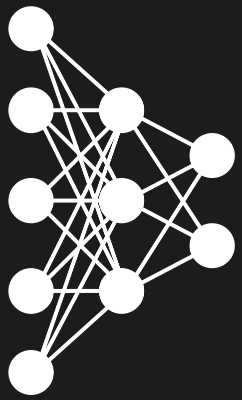
truth **y**

Loss(θ, x, y)

$$\boldsymbol{\theta' = \theta - \nabla_\theta \text{Loss}}$$

Fast Gradient Sign Method

BENIGN INPUT
x

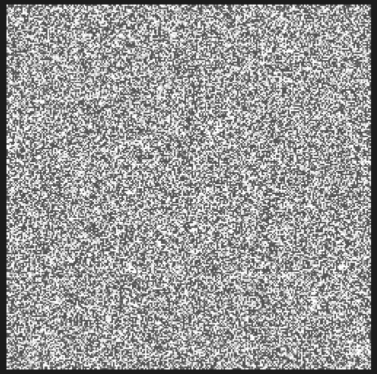


predictions $f(\mathbf{x})$

target y

$$\mathrm{Loss}(f(x), y)$$

$- 0.01 *$



$=$

x

$$\delta = \nabla_x \mathrm{Loss}$$

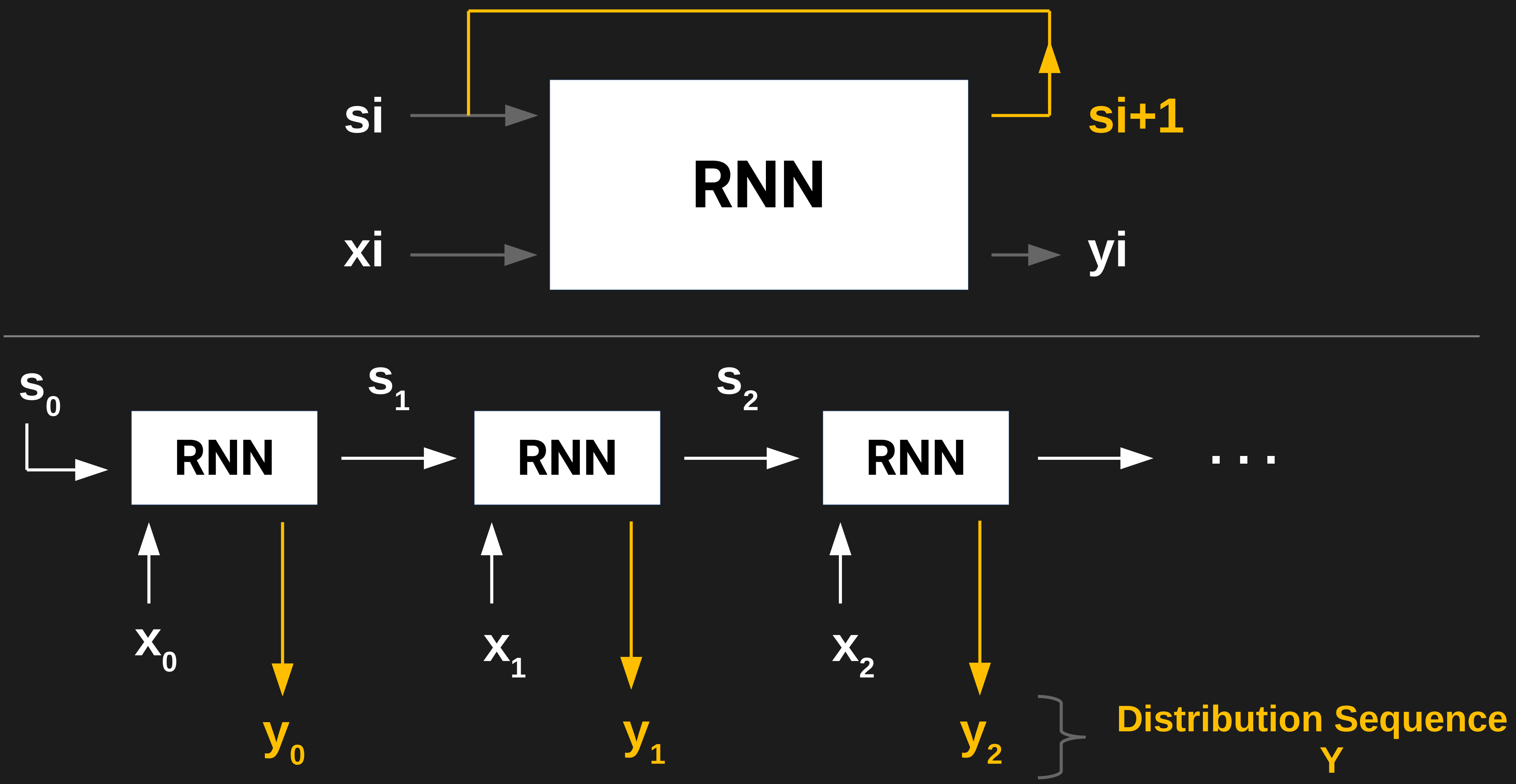
x'

Carlini & Wagner's adversarial attack (1)

- **Targeted** — The resulting adversarial example has a **desired** classification.
- **White Box** — Requires **full** knowledge of the model.
- **Minimally Perceptible** — Trying to minimize the **perceptibility** of the **adversarial noise δ**.

« Hello » « Goodbye »

Recurrent Neural Networks



Connectionist Temporal Classification

Alignment π — **Reduces to** — **Sentence p**

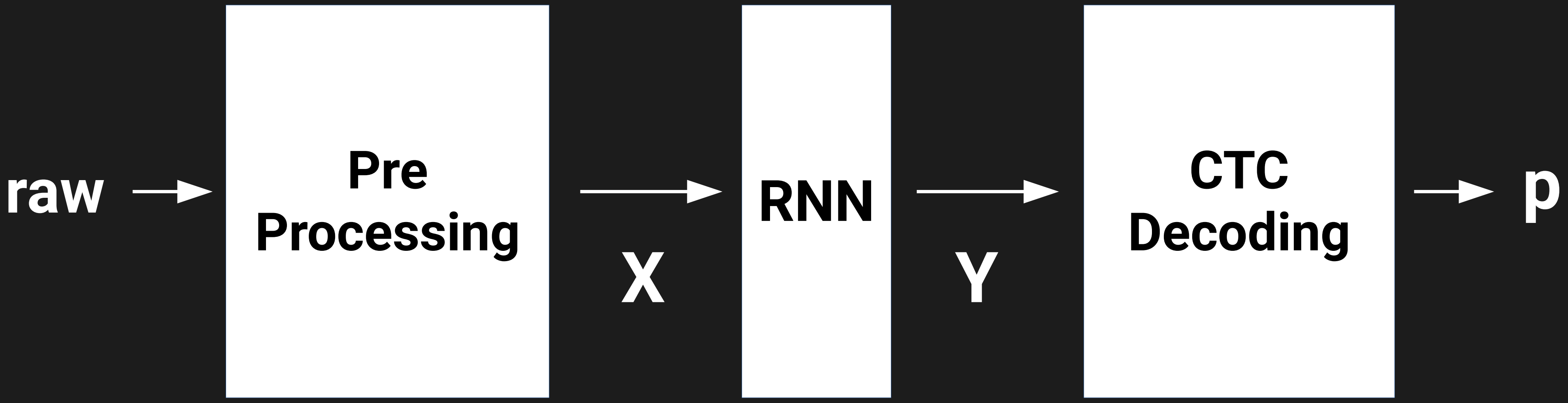
BBBεAεεBBεAAA $\longrightarrow$ BABA

$$
\begin{array}{c}
\varepsilon \\ A \\ B \\ C
\end{array}
\underbrace{
\begin{pmatrix} 0.1 \\ 0.2 \\ 0.6 \\ 0.1 \end{pmatrix}
\begin{pmatrix} 0.5 \\ 0.1 \\ 0.2 \\ 0.3 \end{pmatrix}
\begin{pmatrix} 0.1 \\ 0.7 \\ 0.1 \\ 0.1 \end{pmatrix}
\begin{pmatrix} 0.1 \\ 0.8 \\ 0.0 \\ 0.1 \end{pmatrix}
\begin{pmatrix} 0.0 \\ 0.1 \\ 0.0 \\ 0.9 \end{pmatrix}
}_{\text{Distribution Sequence } \mathbf{Y}}
\longrightarrow
$$

$$Pr(\pi|\mathbf{Y}) = \prod_i \mathbf{Y}^i_{\pi^i}$$

$$Pr(\mathbf{p}|\mathbf{Y}) = \sum_{\pi \in \Pi(\mathbf{p},\mathbf{Y})} Pr(\pi|\mathbf{Y})$$

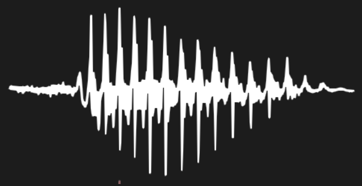
DeepSpeech



$$\mathrm{CTCLoss}(\,\mathrm{raw},\ \mathbf{p}\,) = -\log\ Pr(\mathbf{p}|\mathrm{Y})$$

Carlini & Wagner's adversarial attack (2)

- **Targeted** — The resulting adversarial example has a **desired** classification.
- **White Box** — Requires **full** knowledge of the model.
- **Minimally Perceptible** — Trying to minimize the **perceptibility** of the **adversarial noise δ**.



+ δ =

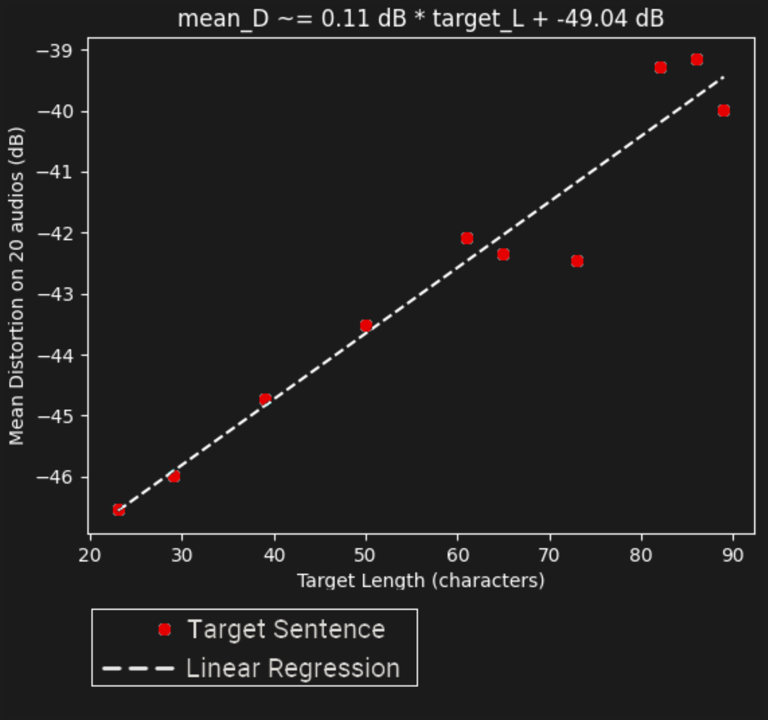
« Hello »

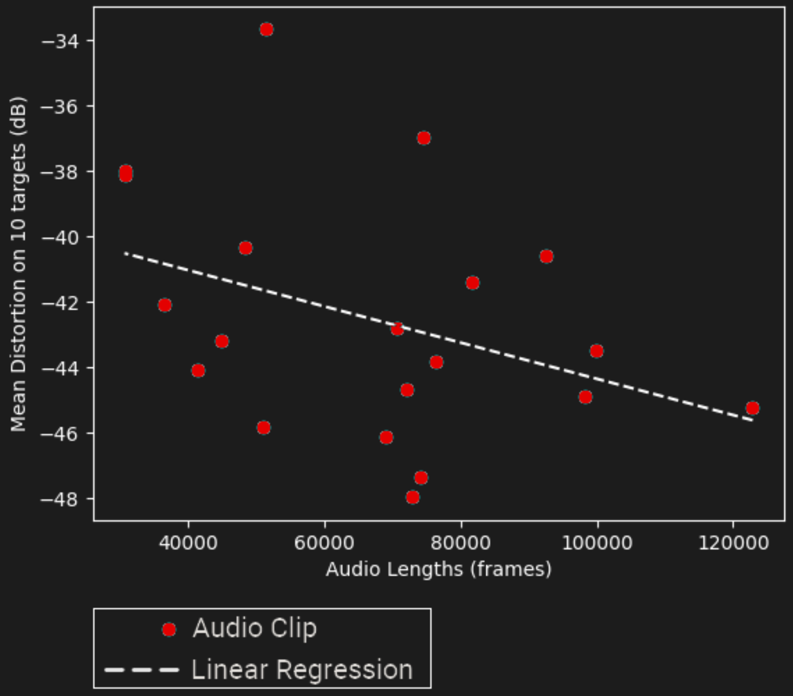
« Goodbye »

$$\text{minimize}\ \ \mathrm{CTCLoss}(\mathbf{raw} + \boldsymbol{\delta}, p)$$
$$\text{such that}\ \ \mathrm{dB}_{raw}(\delta) \leq \tau$$

PyTorch Implementation

- **Methodology** | Targeting 10 different sentences on 20 audios.





Specificities of audio adversarial attacks

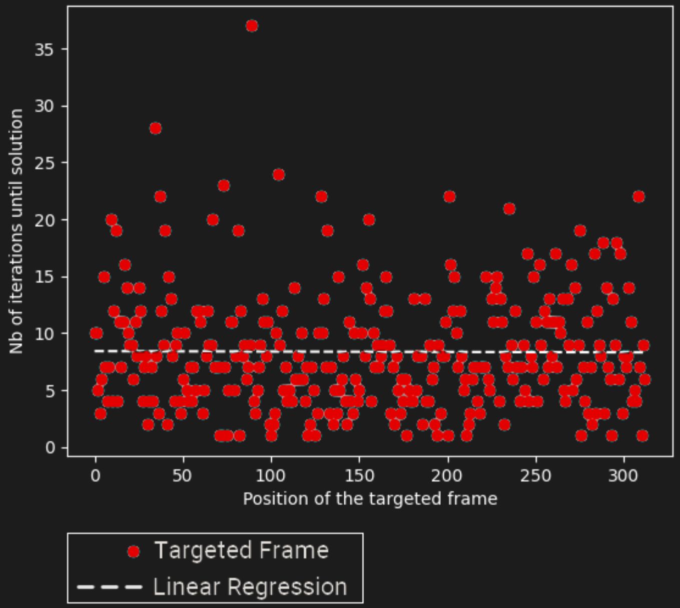
- **Distortion Metrics**

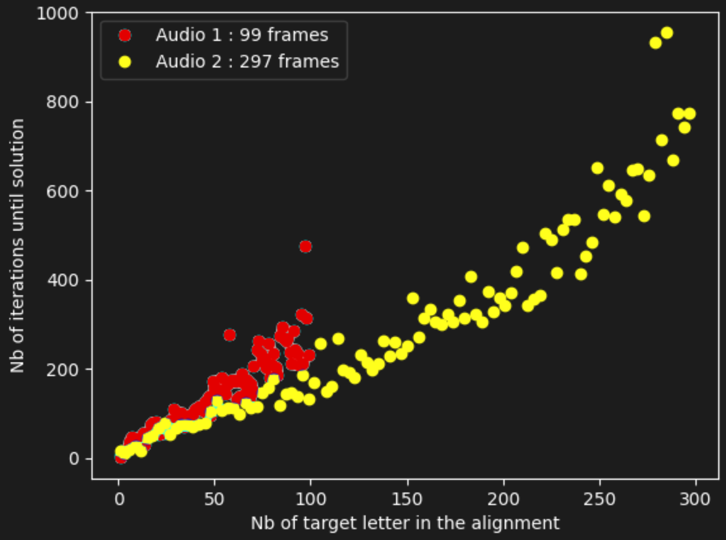
L_∞ & L_2 norms work well for images, not for audio !

- **Degrees of non linearity**

Differenciating through the **pre-processing** and **CTC decoding** step is not easy.

- **RNNs ?**





Conclusion

- Adversarial Attacks are a security and scientific challenge
- Neural Networks are not well understood
- Adversarial Attacks are harder on sequence input

Main References

Intriguing properties of neural networks (2013)
C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus.

Explaining and harnessing adversarial examples (2014)
I. J. Goodfellow, J. Shlens, and C. Szegedy

Audio adversarial examples: Targeted attacks on speech-to-text (2018)
Carlini, N. and Wagner