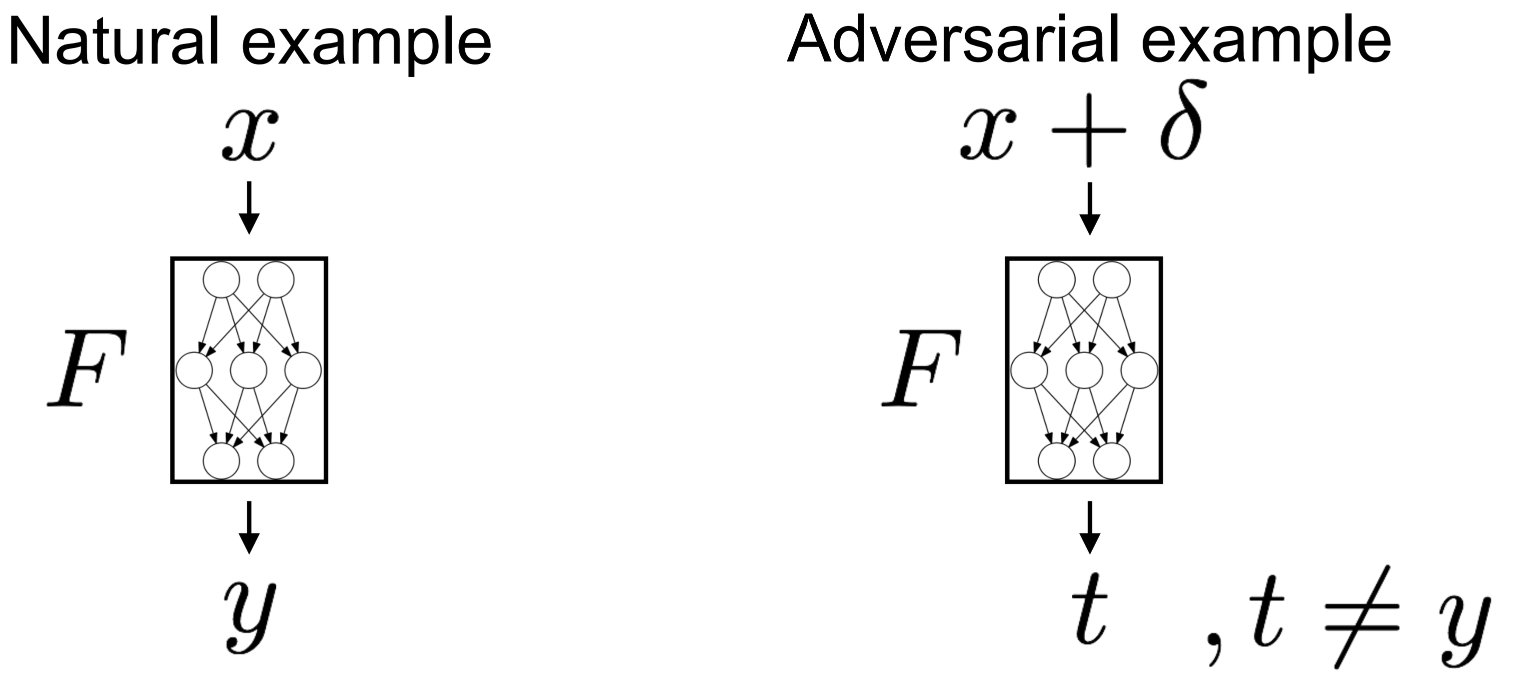
Defending a Speech Recogniser against Adversarial Examples

Áine Cahill, Matt Seigel, Rogier van Dalen
Cambridge University Engineering Department, Apple
aec83@cam.ac.uk

Motivation

- Adversarial examples pose a security threat to neural networks.
- An adversarial example is a malicious input to a neural network which causes the network to misclassify.
- Adversarial examples are easily computed on all neural networks, with or without knowledge of model parameters

Adversarial Examples



Correct classification | Incorrect classification

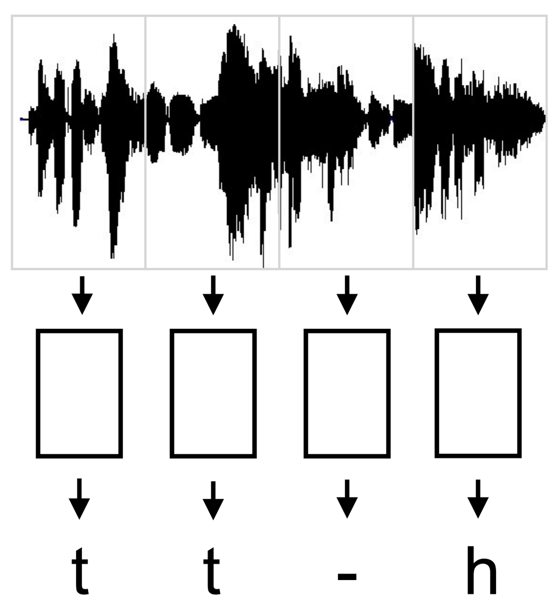
- During an attack, noise, δ, is computed using gradient-descent,

$$\boxed{\text{minimise } |\delta|^2 + L(x+\delta, t)}$$

Distortion of input, x

Loss function: performance of targeted misclassification, $L(x+\delta,t) \leq 0 \Leftrightarrow F(x+\delta) = t$

CTC Speech Recogniser



- Connectionist Temporal Classification (CTC) uses RNNs to label unsegmented data sequences.
- CTC speech recogniser predicts a sequence of labels (letters and space symbols) from unsegmented audio.
- Input = MFCC feature vectors.
- Softmax output layer predicts label at each time instance.
- Decoder finds most likely label sequence at output.
- Mozilla DeepSpeech is a CTC speech recogniser.

Questions to Answer

1. What is the most suitable measure of robustness against adversarial examples on a speech recogniser?
2. How do state-of-the-art defences against adversarial examples perform on a speech recogniser?
3. Are audio adversarial examples transferable between speech recognisers?

1. Measures of Robustness

- Number of training iterations until successful attack found
- Mean distortion of adversarial examples.
- Success rate of adversarial examples.
- Model accuracy vs. % adversarial examples in test set.
- Formal verification methods, e.g. Reluplex, CLEVER.

Planned Experiments

		MNIST	DeepSpeech
	Undefended model		
A	One-hot Thermometer Encoding of Input		
B	Stochastic Activation Pruning (SAP)		
C	Adversarial Training		
D	Linear Region Compression		
E	Non-differentiable Transform of Input		
F	Randomised Sequence of Networks from Ensemble		

2. Defences

A. *One-hot thermometer encoding of input*:

Real value	Quantised	Discretised (one-hot)	Discretised (thermometer)
0.13	0.15	[0100000000]	[0111111111]
0.66	0.65	[0000001000]	[0000001111]

B. *Stochastic Activation Pruning (SAP)*:

Dropout activations from each layer post-training.

1. Prune activations
2. Re-scale activations

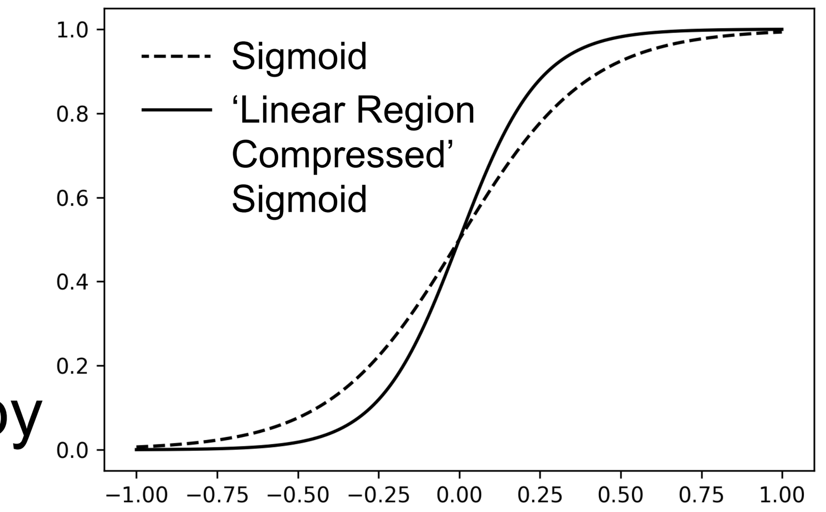
$$p_j^{(i)} = \frac{|h_j^{(i)}|}{\sum_k |h_k^{(i)}|}, \forall j \text{ in layer } i$$

C. *Adversarial Training:*

$$\text{Training set} = \{\text{Natural examples}\} \cup \{\text{Adversarial examples}\}$$

D. *Linear Region Compression:*

Push operation of network into more non-linear regions of activation functions by multiplying weights by a factor > 1.0.



Sigmoid activation function

E. *Non-differentiable Transform of Input:*

Weierstrauss function is non-differentiable everywhere. $\frac{df(\mathbf{x})}{d\mathbf{x}}$ undefined

Adversarial examples cannot be computed if network is non-differentiable.

F. *Randomised Sequence of Networks from Ensemble*

1. Train an ensemble of networks
2. Deploy networks in a random sequence during inference; harder to find adversarial examples.

3. Transferability

- Can audio adversarial examples trained on one speech recogniser successfully fool another speech recogniser trained separately?