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.

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

<table>
<thead>
<tr>
<th>Natural example</th>
<th>Adversarial example</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>( x + \delta )</td>
</tr>
<tr>
<td>( F )</td>
<td>( F )</td>
</tr>
<tr>
<td>( y )</td>
<td>( y )</td>
</tr>
<tr>
<td>( t )</td>
<td>( t )</td>
</tr>
<tr>
<td>( t \neq y )</td>
<td></td>
</tr>
</tbody>
</table>

Correct classification

Incorrect classification

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

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

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

2. Defences

A. One-hot thermometer encoding of input:

<table>
<thead>
<tr>
<th>Real value</th>
<th>Quantised</th>
<th>Discretised (one-hot)</th>
<th>Discretised (thermometer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.13</td>
<td>0.15</td>
<td>[0100000000]</td>
<td>[0111111111]</td>
</tr>
<tr>
<td>0.66</td>
<td>0.65</td>
<td>[0000000100]</td>
<td>[0000001111]</td>
</tr>
</tbody>
</table>

B. Stochastic Activation Pruning (SAP):

1. Prune activations
2. Re-scale activations

C. Adversarial Training:

Training set = \{Natural examples\} \cup \{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.

E. Non-differentiable Transform of Input:

Weierstrauss function is non-differentiable everywhere.

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?