Background

- Deep learning has made a significant difference

... and many more!

... and many more!
Background

- The current business model

Big model on the server
Background

• However, the problems ...

Big model on the server
Background

• However, the problems ...
Background

• However, the problems ...

No server!
Background

- Small-footprint models
- Running speech recognition locally
Background: smaller models

- LSTMs and CNNs vs. DNNs
- Low-ranks matrices for DNNs
  - J. Xue, J. Li, and Y. Gong, “Restructuring of deep neural network acoustic models with singular value decomposition.” in Proc. INTERSPEECH, 2013
Background: smaller models

- LSTMs and CNNs vs. DNNs
- Low-ranks matrices for DNNs
- FitNet by teacher-student training
Background: smaller models

- LSTMs and CNNs vs. DNNs
- Low-ranks matrices for DNNs
- FitNet by teacher-student training
- Structured linear layers
This paper

- FitNet – Thinner and deeper networks
- No teacher-student training
- Highway connections
This paper

• Based on papers addressing *How deep is deep?*

• Image recognition now employs 100+ convolutional layers.

• Our experience
  ○ Reducing the error rate is difficult
  ○ Reducing the mode size is much easier
Model

\[ h_l = \sigma(h_{l-1}, \theta_l) \circ T(h_{l-1}, W_T) + h_{l-1} \circ C(h_{l-1}, W_C) \] (1)

- Shortcut connections with gates
- Similar to Residual networks
- \( W_T \) and \( W_C \) are layer independent
Experiments

- AMI meeting speech transcription with 80h training data
- Using the standard Kaldi recipe
  - fMLLR acoustic features
  - 3-gram language models
- CNTK was used to build HDNN models
- The same decision tree was used
Experiments – Depth and Width

- DNNs were trained using Kaldi with RBM pretraining
Experiments – Depth and Width

- DNNs were trained using Kaldi with RBM pretraining
Experiments – Convergence Rate

![Graph showing convergence rate with frame error rate (%) on the y-axis and number of epochs on the x-axis. The graph compares transform gate only, carry gate only, and transform + carry gate scenarios.]

- **Transform gate only**
- **Carry gate only**
- **Transform + carry gate**
Experiments – Gates

Table: With and without the transform and/or carry gate.

<table>
<thead>
<tr>
<th>System</th>
<th>#Layer</th>
<th>Dim</th>
<th>Transform</th>
<th>Carry</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN*</td>
<td>10</td>
<td>512</td>
<td>×</td>
<td>×</td>
<td>28.8</td>
</tr>
<tr>
<td>HDNN</td>
<td>10</td>
<td>512</td>
<td>√</td>
<td>√</td>
<td>27.2</td>
</tr>
<tr>
<td>HDNN</td>
<td>10</td>
<td>512</td>
<td>√</td>
<td>×</td>
<td>27.6</td>
</tr>
<tr>
<td>HDNN</td>
<td>10</td>
<td>512</td>
<td>×</td>
<td>√</td>
<td>27.5</td>
</tr>
</tbody>
</table>

• Works best with both gates
Experiments – Constraint Gates

Table: Results of using constrained carry gate, where $C(\cdot) = 1 - T(\cdot)$.

<table>
<thead>
<tr>
<th>System</th>
<th>#Layer</th>
<th>Dim</th>
<th>Constrained</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDNN</td>
<td>10</td>
<td>1024</td>
<td>×</td>
<td>26.8</td>
</tr>
<tr>
<td>HDNN</td>
<td>10</td>
<td>1024</td>
<td>√</td>
<td>28.0</td>
</tr>
<tr>
<td>HDNN</td>
<td>10</td>
<td>512</td>
<td>×</td>
<td>27.2</td>
</tr>
<tr>
<td>HDNN</td>
<td>10</td>
<td>512</td>
<td>√</td>
<td>27.4</td>
</tr>
</tbody>
</table>

- Constraint carry gate results in accuracy loss.
Experiments – Sequence Training

Table: $\theta_h$ denotes all the model parameters of the hidden layers, $\theta_g = (W_T, W_c)$, and $\theta_c$ is the parameters in the softmax layer.

<table>
<thead>
<tr>
<th>Model</th>
<th>sMBR Update</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta_h$</td>
<td>$\theta_g$</td>
</tr>
<tr>
<td>HDNN-$H_{512}L_{10}$</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>√</td>
</tr>
</tbody>
</table>

Experiments – Adaptation

Table: Results of unsupervised speaker adaptation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Seed</th>
<th>Update</th>
<th>WER (eval) SI</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDNN-H&lt;sub&gt;512&lt;/sub&gt;L&lt;sub&gt;10&lt;/sub&gt;</td>
<td>sMBR</td>
<td>θ&lt;sub&gt;g&lt;/sub&gt;</td>
<td>24.9</td>
<td>24.1</td>
</tr>
<tr>
<td>HDNN-H&lt;sub&gt;256&lt;/sub&gt;L&lt;sub&gt;10&lt;/sub&gt;</td>
<td></td>
<td></td>
<td>26.0</td>
<td>25.0</td>
</tr>
<tr>
<td>HDNN-H&lt;sub&gt;512&lt;/sub&gt;L&lt;sub&gt;10&lt;/sub&gt;</td>
<td></td>
<td>{θ&lt;sub&gt;h&lt;/sub&gt;, θ&lt;sub&gt;g&lt;/sub&gt;, θ&lt;sub&gt;c&lt;/sub&gt;}</td>
<td>24.9</td>
<td>24.5</td>
</tr>
<tr>
<td>HDNN-H&lt;sub&gt;256&lt;/sub&gt;L&lt;sub&gt;10&lt;/sub&gt;</td>
<td></td>
<td></td>
<td>26.0</td>
<td>25.4</td>
</tr>
</tbody>
</table>

Conclusion

- Small-footprint models using highway networks
  - 2M HDNN model $\approx$ 30M DNN model after sequence training

- More adaptable and controllable
  - The tied gates largely controls the whole network

- Teacher-student training can further improve the accuracy
  - L. Lu, et al, “Knowledge Distillation for Small-footprint Highway Networks”, arXiv 2016. (improves the model with $< 0.8M$ parameters)
Thank you! Questions?