Distant Speech Recognition (DSR)

- DSR is to recognize human speeches in the presence of various noise sources caused by the large distance between speakers and microphones.
- Traditional speech recognizers trained with clean data often fail to recognize due to signal quality mismatch between training and test conditions.
- Proposed a new recursive architecture that can iteratively improve signal denoising and recognition in BridgeNet.

Main Contribution

- Proposed a new student-teacher paradigm for DSR: BridgeNet
- BridgeNet provides teacher’s intermediate features as additional hints, which can properly regularize a student network.
- Proposed a new recursive architecture that can iteratively improve signal denoising and recognition in BridgeNet.

BridgeNet

- Loss function is a weighted sum of all error measures:

\[ L(\phi) = \sum_{i=1}^{n} \alpha_i c_i(\phi) + \sum_{i=1}^{n} (p_i(x_{\text{train}}))^{-1} \log p_i(x_{\text{noisy}}; \phi) \]

where the second term is the cross-entropy between label and student softmax output.

Recurrent Architecture

- Composed of four sub-blocks: I, F, M, L
- I and F take two inputs: acoustic input \((x_t)\) and output \((x_{t-1})\) from prior recursion.
- M merges two independent paths.

\[ m_{i}^{n} = g(W_{i}x_{i} + W_{i}f_{i}^{(n-1)} + b) \]

- For each new recursion, the same \(x_{i}\) is fed into \(L\), which acts as a new global shortcut path.
- The global shortcut paths act as highway paths that facilitate gradient flows and helps to have deep recursive architecture.

Existing Approaches for DSR

- Multi-task denoising (MTD):
  - Jointly optimize denoising (DE) and recognition (RE) subnetworks integrated within the unified neural network.
  - Minimizing MSE between raw acoustic data and high-level abstracted features in MTD is unsuccessful.

- BridgeNet provides the similar high-level features to guide a student network.

Knowledge distillation (KD):

- Knowledge bridges (hints) provide an error measure to guide intermediate feature representation of a student network.

- Knowledge bridges (hints) provide additional supervision to a student net.

Knowledge distillation (KD):

- BridgeNet provides multiple hints from teacher’s intermediate layers.

Main Result

- BridgeNet presented 5.29% accuracy improvements over the baseline CNN-LSTM model on AMI corpus.
- Compared with KD, it showed 2.72% relative WER reduction.
- Recursive architecture further improved BridgeNet: 13.24% improvement of relative WER over CNN-LSTM, 10.88% over KD.

Experiments

- Multi-Task Denoising on AMI SDM corpus: CNN-LSTM* is trained with clean alignment. Rest of them used noisy alignment

<table>
<thead>
<tr>
<th>Acoustic Model</th>
<th>WER(all)</th>
<th>WER (main)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-LSTM(baseline), R0</td>
<td>46.5%</td>
<td>37.7%</td>
</tr>
<tr>
<td>KD, R0</td>
<td>44.8%</td>
<td>35.7%</td>
</tr>
<tr>
<td>KD+DR, R0</td>
<td>44.3%</td>
<td>35.0%</td>
</tr>
<tr>
<td>CNN-LSTM, denoised</td>
<td>41.8%</td>
<td>32.2%</td>
</tr>
<tr>
<td>CNN-LSTM*, denoised</td>
<td>46.9%</td>
<td>38.2%</td>
</tr>
</tbody>
</table>

- CNN-LSTM* is our baseline model: two layers of CNN layers are stacked with 3 layers of LSTM. CNN model has 8 layers.
- Multi-task denoising showed marginal improvement for CNN and CNN-LSTM.
- CNN-LSTM using clean alignment showed degradation with MTD.

BridgeNet: single channel SDM corpus is used for training a student network

<table>
<thead>
<tr>
<th>Acoustic Model</th>
<th>WER(all)</th>
<th>WER (main)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-LSTM(baseline), R0</td>
<td>44.8%</td>
<td>34.0%</td>
</tr>
<tr>
<td>KD, R0</td>
<td>42.8%</td>
<td>33.1%</td>
</tr>
<tr>
<td>KD+DR, R0</td>
<td>42.3%</td>
<td>32.5%</td>
</tr>
<tr>
<td>CNN-LSTM(baseline), R2</td>
<td>43.0%</td>
<td>33.3%</td>
</tr>
<tr>
<td>KD, R1</td>
<td>40.4%</td>
<td>30.8%</td>
</tr>
<tr>
<td>KD+DR, R1</td>
<td>39.5%</td>
<td>29.9%</td>
</tr>
<tr>
<td>KD+DR+LSTM3, R1</td>
<td>39.3%</td>
<td>29.5%</td>
</tr>
</tbody>
</table>

- KD, DR and LSTM3 are knowledge bridges between student and teacher networks.
- Each added bridge incrementally improves BridgeNet: KD+DR+LSTM3 provided 6.9% gain over CNN-LSTM and 1.6% gain over KD.
- BridgeNet with recursion presented huge gain: 13.24% and 10.88% WER reduction over CNN-LSTM and KD.