Modeling Long Temporal Contexts in Convolutional Neural Network-Based Phone Recognition

László Tóth
Research Group on Artificial Intelligence, Hungarian Academy of Sciences and University of Szeged, Hungary

ABSTRACT

The DNN component of current hybrid ASR systems is trained on a context of consecutive feature vectors. Here, we propose the training of the combined model using a longer context of the input. The improvement is achieved by a combination of a network that splits the frequency axis into several bands. We report a phone error rate of 17.1% on the TIMIT core test set, which is among the best scores published.

1. MOTIVATION

Our baseline system is a Convolutional Neural Network (CNN) that splits the frequency axis into several bands [1].

• We observed that splitting the input along the frequency axis was beneficial even without convolution [2].
• Schwarz et al. found that splitting the input along time was advantageous when using shallow networks [3]. Similar results were recently reported for fully connected DNNs [4].

QUESTIONS

• How does the split temporal context approach work for CNNs that split the frequency axis as well?
• Can it be combined with hierarchical modeling (which seeks to extend the time-span of the acoustic model in a different way)?

Main References


2. EXPERIMENTAL CONTEXT (STC)

• The temporal context of the input is split into ‘left’ and ‘right’ parts.
• The lowermost hidden layers are also split. That is, the left and right parts are processed by separate, dedicated sub-networks.
• The left and right sub-networks are trained to recognize the same label, which belongs to the center frame of the full context.
• The left and right estimates are merged via additional layers.

3. EXPERIMENTAL RESULTS ON TIMIT

First, we compared how the phone error rate of the baseline (JTC) and the STC networks varies with the observation context size (given in frames). The JTC and the STC nets had 4 hidden layers, and the same total number of weights.

<table>
<thead>
<tr>
<th>No. of split layers</th>
<th>Dev. set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.7%</td>
<td>17.2%</td>
</tr>
<tr>
<td>2</td>
<td>15.4%</td>
<td>17.0%</td>
</tr>
<tr>
<td>3</td>
<td>15.1%</td>
<td>17.0%</td>
</tr>
<tr>
<td>4</td>
<td>15.1%</td>
<td>17.0%</td>
</tr>
</tbody>
</table>

4. HIERARCHICAL MODELING [2,5]

• The temporal context is split into several (in this case, 5) parts.
• The five parts are processed by 5 replicas of the same sub-network, using the same, shared weights.
• The sub-network is trained to estimate different labels that belong to the center frames of their local context.
• The five local estimates are merged via additional layers.

5. EXPERIMENTAL RESULTS ON TIMIT

To be comparable with STC, we evaluated the hierarchical model with a lower sub-network part consisting of 4 layers, while the merging was performed by the softmax output layer.

<table>
<thead>
<tr>
<th>Network type</th>
<th>Dev. set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Core baseline (49 frames)</td>
<td>15.7%</td>
<td>17.7%</td>
</tr>
<tr>
<td>STC (49 frames)</td>
<td>14.8%</td>
<td>17.1%</td>
</tr>
<tr>
<td>Hierarchical (69 frames)</td>
<td>14.8%</td>
<td>17.0%</td>
</tr>
<tr>
<td>STC + hierarchical (69 frames)</td>
<td>14.0%</td>
<td>16.6%</td>
</tr>
<tr>
<td>STC + hierarchical + dropout</td>
<td>13.4%</td>
<td>16.2%</td>
</tr>
</tbody>
</table>

We repeated the training of the combined model using dropout, and we evaluated the model on the core test set as well. The phone error rate of the combined model is significantly lower than that of the baseline, and it compares favorably with state-of-the-art models.

CONCLUSION

The DNN component of current hybrid ASR systems is trained on a context of consecutive feature vectors. We propose the training of the combined model using a longer context of the input. The improvement is achieved by a combination of a network that splits the frequency axis into several bands.