Sequence-discriminative training of deep neural networks

Joint work with Karel Vesely, Lukas Burget, and Daniel Povey

23 May 2013
Sequence-discriminative training of HMMs

- Conditional maximum likelihood (Nádas, 1983) and Maximum mutual information (Bahl et. al 1986) for training GMM parameters
  - Both are identical when LM is fixed during training
  - MI between word and acoustic sequences
- Efficient extended Baum-Welch (EBW) training proposed by Normandin in 1991
- First practical large-scale system in 2000 from Cambridge (Povey & Woodland)
  - Posteriors computed over lattices
  - Sequence-discrimination better than frame-discrimination
Training an MLP

- Trained using *error backpropagation*
- Requires gradient of output w.r.t. activations $a_{ut}(s)$ at the output layer

$$y_{ut}(s) \triangleq P(s | o_{ut}) = \frac{\exp\{a_{ut}(s)\}}{\sum_{s'} \exp\{a_{ut}(s')\}}$$
The cross-entropy criteria

\[ F_{CE} = - \sum_u \sum_t \log y_{ut}(s_{ut}), \]

\[ \frac{\partial F_{CE}}{\partial a_{ut}(s)} = y_{ut}(s) - \delta_{s; s_{ut}} \]

- Expected cross-entropy between reference labels and predicted distribution \( y(t) \)
- Minimizing CE same as maximizing mutual information between \( y(t) \) and reference
Training an MLP sequence-discriminatively

- Bridle & Dodd’s Alphanet (1991) was trained in a sequence-discriminative fashion
- Revived by Kingsbury (2009) using lattice-based computations

\[
\frac{\partial F_{CE}}{\partial a_{ut}(s)} = y_{ut}(s) - \delta_{s;s_{ut}} \\
\frac{\partial F_{MMI}}{\partial a_{ut}(s)} = \kappa(\delta_{s;s_{ut}} - \gamma_{ut}^{DEN}(s))
\]
Maximum-mutual information

\[
\mathcal{F}_{MMI} = \sum_u \log \frac{p(O_u | S_u)^\kappa P(W_u)}{\sum_W p(O_u | S)^\kappa P(W)},
\]

\[
\frac{\partial \mathcal{F}_{MMI}}{\partial a_{ut}(s)} = \kappa (\delta_{s; s_{ut}} - \gamma_{ut}^{DEN}(s))
\]

- Here we assume that numerator stats collected through forced alignment
- Possible to use forward-backward and use \(\gamma_{ut}^{NUM}(s)\)
Minimum Bayes Risk

\[ F_{MBR} = \sum_u \frac{\sum_W p(O_u|S)^\kappa P(W) A(W, W_u)}{\sum_{W'} p(O_u|S)^\kappa P(W')} \],

\[ \frac{\partial F_{MBR}}{\partial a_{ut}(s)} = \kappa \gamma_{ut}^{DEN} (s) \{ \bar{A}_u(s_t = s) - \bar{A}_u \} \]

- \( A(W, W_u) \) is the accuracy of phone labels (for MPE) or state labels (for sMBR)
- \( A_u(s_t = s) \) is the average accuracy of all paths passing through state \( s \) at time \( t \)
Boosted MMI

\[ \mathcal{F}_{BMMI} = \sum_u \log \frac{p(O_u|S_u)^\kappa P(W_u)}{\sum_W p(O_u|S)^\kappa P(W) e^{-b A(W,W_u)}} \]

- “Boosts” the likelihood of paths that contain more errors --- state-of-the-art training criteria for GMMs
- May also be interpreted as incorporating a margin term in the MMI objective (Heigold, et al. 2008)
- Gradient computation identical to that for MMIE
Switchboard setup

- 300-hours Switchboard conversational telephone speech recognition
- Trained on Switchboard-1 Release 2 (LDC97S62)
- MSU transcripts, 30K-word lexicon
- Tested on Hub5 ’00 and Hub5 ’01 test sets
- 3-gram LM trained on 3M words of training transcripts interpolated with 11M words of Fisher-1
  - Interpolated KN smoothing; 950K 3-grams, 1064K 2-grams
- AMs trained on 40-dim LDA+STC features from 7 frames (±3) of 13-dim MFCC (C0-C12)
GMM-HMM Baselines (on Hub5 ’00)

<table>
<thead>
<tr>
<th>System</th>
<th>Hours</th>
<th>SWB</th>
<th>CHE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML SAT GMM</td>
<td>300</td>
<td>21.2</td>
<td>36.4</td>
<td>28.8</td>
</tr>
<tr>
<td>BMMI SAT GMM</td>
<td>300</td>
<td>18.6</td>
<td>33.0</td>
<td>25.8</td>
</tr>
<tr>
<td>ML SAT GMM</td>
<td>110</td>
<td>23.8</td>
<td>38.6</td>
<td>31.2</td>
</tr>
<tr>
<td>BMMI SAT GMM</td>
<td>110</td>
<td>21.0</td>
<td>35.6</td>
<td>28.3</td>
</tr>
</tbody>
</table>

- 300-hour models have 8859 tied triphone states and 200K Gaussians
- 110-hour models have 4234 tied triphone states and 90K Gaussians
- Speaker adaptive training (SAT) with a single FMLLR transform
- BMMI uses a boosting factor of \( b = 0.1 \)
DNN-HMM results (on Hub5 ’00)

<table>
<thead>
<tr>
<th>System</th>
<th>Hours</th>
<th>SWB</th>
<th>CHE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMMI SAT GMM</td>
<td>300</td>
<td>18.6</td>
<td>33.0</td>
<td>25.8</td>
</tr>
<tr>
<td>BMMI SAT GMM</td>
<td>110</td>
<td>21.0</td>
<td>35.6</td>
<td>28.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>Init</th>
<th>Hours</th>
<th>SWB</th>
<th>CHE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN 7 layers</td>
<td>RBM</td>
<td>300</td>
<td>14.2</td>
<td>25.7</td>
<td>20.0</td>
</tr>
<tr>
<td>DNN 5 layers</td>
<td>Rand</td>
<td>110</td>
<td>17.1</td>
<td>29.6</td>
<td>23.4</td>
</tr>
</tbody>
</table>

- DNNs trained on 40dim LDA+STC+FMLLR features
- Input is 11 frames (± 5) of the 40-dim features for the 300hr model or 9 frames for the 110hr model
- Trained using SGD with 256-frame mini-batches and frame shuffling; alignments from GMM BMMI system
MMI-training of DNNs

- CE training using DNN alignments (better numerator) accounts for about half of MMI improvement
- MMI seen to overfit after 2 iterations
- Smaller learning rate (about 0.1 of CE lr) required
Segments with missing reference in lattice

- Reference for some segments missing from lattice
  - Search errors; poor match of acoustics; error in reference
- Mostly short segments (70% less than 0.5s) but they disproportionately affect the gradient
- Remove these from MMI training
The frame-rejection heuristic reduces the amount of training data by 2.5% but leads to more stable learning.
Comparing different criteria

- Not much difference between different sequence-discriminative criteria
Lattice regeneration

- Regenerating lattices after each iteration improves performance a little but it is computationally expensive
Results with 300-hour training set

| System       | Hub5 ’00 | | Hub5 ’01 | | |
|--------------|----------|----------|----------|----------|
|              | SWB      | CHE      | Total    | SWB      | SWB2P3   | SWB-Cell | Total    |
| GMM BMMI     | 18.6     | 33.0     | 25.8     | 18.9     | 24.5     | 30.1     | 24.6     |
| DNN CE       | 14.2     | 25.7     | 20.0     | 14.5     | 19.0     | 25.3     | 19.8     |
| DNN MMI      | 12.9     | 24.6     | 18.8     | 13.3     | 17.8     | 23.7     | 18.4     |
| DNN BMMI     | 12.9     | 24.5     | 18.7     | 13.2     | 17.8     | 23.5     | 18.3     |
| DNN MPE      | 12.9     | 24.1     | 18.5     | 13.2     | 17.7     | 23.4     | 18.2     |
| DNN sMBR     | 12.6     | 24.1     | 18.4     | 13.0     | 17.7     | 22.9     | 18.0     |

- sMBR found to be slightly better than other sequence-discriminative criteria (observation congruent to those of IBM)
- The current best results on Switchboard!
In poster session: cross-lingual pretraining

- Stacked RBMs trained on Spanish, Portuguese, Swedish, and German
- Finetuned using 1-hour, 5-hours, and 15-hours of German
In poster session: hat-swapping DNNs

The name hat-swapping was suggested by John Bridle
The Kaldi speech recognition toolkit

KALDI  http://kaldi.sf.net/

- Free (Apache v2.0 license), open-source, speech recognition toolkit
  - Written in C++ (supported on common *NIX platforms)
  - FST-based training & decoding using OpenFST
  - Supports standard GMMs, Subspace GMMs, and DNNs
  - Complete reproducible recipes with state-of-the-art results on several corpora

According to legend, Kaldi was the Ethiopian goatherd who discovered the coffee plant.