Multi-level adaptive networks in tandem and hybrid ASR systems

Peter Bell, Pawel Swietojanski, and Steve Renals
{peter.bell,p.swietojanski,s.renals}@ed.ac.uk
Centre for Speech Technology Research, University of Edinburgh, UK

This work
- We investigate the use of multi-level adaptive networks (MLAN) to incorporate out-of-domain (OOD) data when training speech recognition systems.
- We compare the use of MLAN in both tandem and hybrid deep neural network (DNN) based systems.
- The MLAN approach yields consistent word error rate reductions of 5-10% relative, in both tandem and hybrid configurations.
- Tandem and hybrid systems are close in performance – it is not clear whether one is better than the other.

Motivation: neural network features for domain adaptation
- Features derived from neural networks are known to provide a degree of domain-independence.
- Features trained on one domain may be used add discriminative ability to another domain – perhaps one where data is limited. Adaptation could be performed by:
  - carrying out additional training iterations on in-domain data (but pre-training already provides a degree of regularisation when training data is limited)
  - using a merger MLP to combine nets trained on different domains
  - in tandem framework, retraining the GMMs.
- We propose the MLAN scheme, where a second DNN is used to discriminatively select which OOD features are most effective in the new domain.

Tandem and hybrid DNN systems

Tandem:
- Neural networks are used to derive features for training data, which are augmented with the standard acoustic features and used to train GMMs.
- Networks typical model posterior probabilities over monophones.
- The output features may be either decorrelated posterior features, or bottleneck features.

Hybrid:
- Neural networks are used to generate posterior probabilities over states, which are used as pseudo-likelihoods in an HMM-based decoder, scaled by the state priors.
- Recent systems model the probabilities of tied triphone states (“senones”), giving a much wider output layer than in the tandem case.

MLAN: Multi-level adaptive networks

The MLAN scheme is used to incorporate information from multiple OOD features:
- Generate OOD tandem features for the in-domain data by combining posterior features generated using OOD nets, and standard acoustic features.
- Train tandem and hybrid DNNs on the OOD tandem features:
  - tandem DNNs are used to generate MLAN posterior features, which are then used to generate final MLAN GMMs.
  - hybrid MLAN DNNs directly model tied triphone-states using a decision-tree clustering derived from the in-domain data.

Experiments

Data
- We investigated the MLAN technique on two ASR tasks:
  - TV and radio broadcasts in British English from the British Broadcasting Corporation (BBC), including studio speech, programmes made on location, and a TV drama series; with 20.7 hours of training data.
  - The TED lectures transcription task used in the IWSLT 2012 evaluation campaign, with 143 hours of training data.

Results

<table>
<thead>
<tr>
<th>System setup</th>
<th>Tandem Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBC systems</td>
<td>33.0 33.1</td>
</tr>
<tr>
<td>AMI</td>
<td>32.5 -</td>
</tr>
<tr>
<td>CTS tandem</td>
<td>34.1 -</td>
</tr>
<tr>
<td>AMI MLAN</td>
<td>31.0 31.0</td>
</tr>
<tr>
<td>CTS MLAN</td>
<td>30.7 30.1</td>
</tr>
<tr>
<td>AMI + CTS MLAN</td>
<td>29.9 29.6</td>
</tr>
</tbody>
</table>

- For out of domain data we used 276 hours of US-English conversational telephone speech (CTS) from Switchboard, and 127 hours of multi-party meetings from the AMI corpus. Both were highly mismatched to the BBC and TED tasks.

- Tandem DNNs used four hidden layers with 1024 hidden units per layer; hybrid DNNs used 6-7 hidden layers with 2048 units per layer. All DNNs used nine frames of acoustic context.
- DNNs were initialised with RBM pre-training and fine-tuned with stochastic gradient descent.
- Tandem features were projected to 30 dimensions – tandem GMMs were trained on the resulting 69-dimensional features.
- The triphone state clustering used for the each hybrid system was obtained from the equivalent tandem system.
- There was no speaker adaptation on the BBC task. On the TED task:
  - tandem systems were trained with SAT, using 32 block-diagonal CMLLR transforms per speaker.
  - for hybrid systems we investigated SAT with a single feature-space transform per speaker, computed using a GMM trained on the same features.
- All tandem systems use MPE training.

- We compared the use of MLAN in both tandem and hybrid (MLAN) to incorporate information from multiple OOD features for the in-domain data by combining posterior features generated using OOD nets, and standard acoustic features.
- Networks typical model posterior probabilities over monophones.
- The output features may be either decorrelated posterior features, or bottleneck features.

- Recent systems model the probabilities of tied triphone states (“senones”), giving a much wider output layer than in the tandem case.

- We investigated the MLAN technique on two ASR tasks:
  - TV and radio broadcasts in British English from the British Broadcasting Corporation (BBC), including studio speech, programmes made on location, and a TV drama series; with 20.7 hours of training data.
  - The TED lectures transcription task used in the IWSLT 2012 evaluation campaign, with 143 hours of training data.

- For out of domain data we used 276 hours of US-English conversational telephone speech (CTS) from Switchboard, and 127 hours of multi-party meetings from the AMI corpus. Both were highly mismatched to the BBC and TED tasks.

- Tandem DNNs used four hidden layers with 1024 hidden units per layer; hybrid DNNs used 6-7 hidden layers with 2048 units per layer. All DNNs used nine frames of acoustic context.
- DNNs were initialised with RBM pre-training and fine-tuned with stochastic gradient descent.
- Tandem features were projected to 30 dimensions – tandem GMMs were trained on the resulting 69-dimensional features.
- The triphone state clustering used for the each hybrid system was obtained from the equivalent tandem system.
- There was no speaker adaptation on the BBC task. On the TED task:
  - tandem systems were trained with SAT, using 32 block-diagonal CMLLR transforms per speaker.
  - for hybrid systems we investigated SAT with a single feature-space transform per speaker, computed using a GMM trained on the same features.
- All tandem systems use MPE training.

- We investigated the MLAN technique on two ASR tasks:
  - TV and radio broadcasts in British English from the British Broadcasting Corporation (BBC), including studio speech, programmes made on location, and a TV drama series; with 20.7 hours of training data.
  - The TED lectures transcription task used in the IWSLT 2012 evaluation campaign, with 143 hours of training data.