Multi-task ensembles with teacher-student training

Jeremy H. M. Wong and Mark J. F. Gales
Department of Engineering, University of Cambridge
jhmw2@cam.ac.uk, mffg@eng.cam.ac.uk

MOTIVATION
- Ensemble combination gives gains over single model.
- Can use diversity of output targets in ensemble.
- Computationally expensive to perform recognition.
- Paper investigates reducing computational cost by:
  - Training single student model to emulate ensemble.
  - Merging hidden layers into multi-task architecture.

OUTPUT TARGET DIVERSITY
- Phonetic Decision Tree (PDT) defines output targets
  \( s_e = T(c) \)
- Use random forest method to generate multiple PDTs.
- Use different PDT for each model in the ensemble.
- Models discriminate between different sets of states.

TEACHER-STUDENT TRAINING
- Train single student to emulate combined ensemble.
- Propagate information from teachers to student.
- Map posteriors across PDts.

\[
\mathcal{F}_T = -\sum_{i} \sum_{c} Q \left( s_i | \alpha, \phi \right) \log P \left( s_i | \alpha, \phi \right)
\]

\[
Q \left( s_i | \alpha, \phi \right) = \sum_{m} \lambda_m \sum_{c \in C_T} P \left( s_i | c \right) P \left( s_i | \alpha, \phi \right)
\]

- Lose output target diversity and phonetic resolution.

ENSEMBLE CONSTRUCTION
- Compress ensemble by merging hidden layers into multi-task ensemble.
  - Improve recognition efficiency.
  - Retain output target diversity.
- Previous work has investigated:
  - Multi-task CE training using hard targets.
  - Yields limited diversity and combination gains.
- Propose to:
  - Do teacher-student training of multi-task ensemble.
  -explicitly capture diversity of separate models.

Multi-task with hard targets

\[
\mathcal{F}_{MT} = -\sum_{i} \sum_{c \in C_T} \sum_{s \in S} \delta \left( s_i, s_{i}^{*} \right) \log P \left( s_i | \alpha, \phi, \Xi \right)
\]

Multi-task teacher-student

\[
\mathcal{F}_{MT-ST} = -\sum_{i} \sum_{c \in C_T} \sum_{s \in S} \log P \left( s_i | \alpha, \Phi^{m} \right) \log P \left( s_i | \alpha, \phi, \Xi \right)
\]

- Further refine ensemble using sequence training.
- Back-propagate gradient through frame combination.

COMBINATION

- Hypothesis level
  - Separate feed-forward, separate decoding runs.
  - \( H^* = \arg \min_{H} \sum_{m} \lambda_m \sum_{c} \mathcal{L} \left( H, H^* \right) P \left( H^{m} | \omega, \Phi^{m*} \right) \)

Frame level
  - Separate feed-forward, single decoding run.
  \[
  \hat{p} \left( s_i | \alpha, \hat{\phi} \right) = \sum_{m} \lambda_m \sum_{c \in C_T} P \left( s_i | \alpha, \Phi^{m} \right) P \left( s_i | \alpha, \phi \right)
  \]

Teacher-student
  - Single feed-forward, single decoding run.
  - Student may require large output layer to effectively capture phonetic resolution of the ensemble.

Number of NN weights for Tok Pisin VLLP ensembles.

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Combined WER (%)</th>
<th>Hypothesis</th>
<th>Frame</th>
<th>Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>207V separate</td>
<td>45.8</td>
<td>46.8</td>
<td>46.6</td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td>47.7</td>
<td>47.8</td>
<td>47.3</td>
<td></td>
</tr>
<tr>
<td>MT-TS</td>
<td>45.7</td>
<td>45.7</td>
<td>46.3</td>
<td></td>
</tr>
<tr>
<td>AMI separate</td>
<td>24.5</td>
<td>24.6</td>
<td>24.6</td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td>25.4</td>
<td>25.3</td>
<td>25.1</td>
<td></td>
</tr>
<tr>
<td>MT-TS</td>
<td>24.3</td>
<td>24.4</td>
<td>24.6</td>
<td></td>
</tr>
<tr>
<td>HUB4 separate</td>
<td>8.7</td>
<td>8.7</td>
<td>9.1</td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td>9.1</td>
<td>8.8</td>
<td>8.8</td>
<td></td>
</tr>
<tr>
<td>MT-TS</td>
<td>8.8</td>
<td>8.7</td>
<td>8.9</td>
<td></td>
</tr>
</tbody>
</table>

- Ensembles and students have been sequence trained.
- Can compress separate models into MT-TS ensemble without performance loss.
- 207V MT-TS student with intersect PDT: 45.8 %

MEASURING DIVERSITY

- cross-WER = \( \frac{1}{M(M-1)} \sum_{m=1}^{M} \sum_{r=1}^{M} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L} \left( H^{m_i}, H^{r_i} \right) \)
- Minimum edit distance between the transcriptions given by two models.
- Average over all pair-wise combinations of models within the ensemble.

CONCLUSION

This work has proposed to:
- Compress ensemble with output target diversity, by merging hidden layers.
- Use teacher-student training to allow multi-task ensemble to learn from diversity of separate models.
- Jointly train multi-task ensemble toward sequence discriminative criterion.
- Further compress ensemble into student with single output layer.

EXPERIMENTS

Datasets:
- 207V: IARPA Babel Tok Pisin VLLP (3 hours)
- AMI: Augmented multi-party interaction (81 hours)
- HUB4: English broadcast news (144 hours)

Diversity of Tok Pisin VLLP ensembles

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Mean WER (%)</th>
<th>cross-WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>separate</td>
<td>45.8</td>
<td>46.6</td>
</tr>
<tr>
<td>MT</td>
<td>47.7</td>
<td>47.3</td>
</tr>
<tr>
<td>MT-TS</td>
<td>45.7</td>
<td>46.3</td>
</tr>
<tr>
<td>207V separate</td>
<td>45.8</td>
<td>46.6</td>
</tr>
<tr>
<td>MT</td>
<td>47.7</td>
<td>47.8</td>
</tr>
<tr>
<td>MT-TS</td>
<td>45.7</td>
<td>46.3</td>
</tr>
<tr>
<td>AMI separate</td>
<td>24.5</td>
<td>24.6</td>
</tr>
<tr>
<td>MT</td>
<td>25.4</td>
<td>25.3</td>
</tr>
<tr>
<td>MT-TS</td>
<td>24.3</td>
<td>24.6</td>
</tr>
<tr>
<td>HUB4 separate</td>
<td>8.7</td>
<td>9.1</td>
</tr>
<tr>
<td>MT</td>
<td>9.1</td>
<td>8.8</td>
</tr>
<tr>
<td>MT-TS</td>
<td>8.8</td>
<td>8.9</td>
</tr>
</tbody>
</table>