Knowledge Distillation Across Neural Networks With Different Triphone Clusters in Speech Recognition

Jeremy Wong
Outline

- Motivation
  - Triphone clustering
  - Ensemble methods
  - Knowledge distillation

- Proposal
  - Modification to standard knowledge distillation
Triphone clustering

- Acoustic model gives probability of observation given current phone.

- Observation of current phone affected by neighbouring phones.

- Instead, model likelihood given triphone, \( p(o_t|\tau) \).

<table>
<thead>
<tr>
<th>Word:</th>
<th>good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phones:</td>
<td>g, uw, dcl</td>
</tr>
<tr>
<td>Triphones:</td>
<td>0–g–uw, g–uw–dcl, uw–dcl–0</td>
</tr>
</tbody>
</table>
Languages tend to have ~ 50 to 100 phones.

Leads to ~ 100000 possible triphones.

Cluster triphones together to get likelihood of observation given state cluster, \( p(o_t|s) \).

Many possible ways to get the state clusters.
Model

- Model produces $P(s|o_t)$.
- Convert to $p(o_t|s)$. 

\[ P(s|o_t) \]

\[ p(o_t|s) \]
Ensemble methods

- Do recognition using multiple models.
- Combine the outputs.
- Performs better than a single model.
- Models in ensemble should be diverse.
- Computation scales with ensemble size.
Knowledge distillation

- Train a single model to emulate the behaviour of the ensemble.
Knowledge distillation

- Propagate information from teachers to student.
- Minimise KL–divergence between posteriors.

\[ \theta^* = \arg \min_{\theta} \sum_t \sum_s \frac{1}{M} \sum_{m=1}^M P(s \mid o_t, \phi^{(m)}) \log P(s \mid o_t, \theta) \]
Ensemble with different clusterings

- What if teachers each have a different triphone clustering?
- Allow more diverse behaviours between models.
- Leads to better combined performance.
Knowledge distillation with different clusters

- Each teacher’s states represent different clusters.

\[ P(s|o_t, \phi^{(m)}) \rightarrow P(s^{(m)}|o_t, \phi^{(m)}) \]

- KL–divergence does not make sense.
Modification to knowledge distillation

- Minimise KL–divergence over all triphones, $\tau$, not clusters, $s$.

$$\theta^* = \arg\min_\theta \sum_t \sum_{\tau} \frac{1}{M} \sum_{m=1}^{M} P(\tau|o_t, \phi^{(m)}) \log P(\tau|o_t, \theta)$$

- Rewrite as

$$\theta^* = \arg\min_\theta \sum_t \sum_s \left[ \frac{1}{M} \sum_{m=1}^{M} \sum_{s^{(m)}} P(s|s^{(m)})P(s^{(m)}|o_t, \phi^{(m)}) \right] \log P(s|o_t, \theta)$$

Target
Modification to knowledge distillation

- Targets:

\[ \frac{1}{M} \sum_{m=1}^{M} \sum_{s^{(m)}} P(s|s^{(m)})P(s^{(m)}|o_t, \varphi^{(m)}) \]

Transformation Teachers’ outputs
Interpretation

\[ \varphi^{(1)} \quad \varphi^{(2)} \quad \varphi^{(3)} \quad \theta \]

\[ s^{(1)} \rightarrow s \quad s^{(2)} \rightarrow s \quad s^{(3)} \rightarrow s \]
Conclusion

- Standard knowledge distillation requires same state clusters.
- Propose minimising KL-divergence over all triphones.
- Leads to method of mapping the teachers’ outputs between clusters.
Thank you