**Overview**

Accent mismatching is a critical problem for ASR. Commercial speech applications typically only model varieties associated with major countries. In real-world smart speaker devices, users set up their language preferences regardless of whether they are native speakers or not.

Our work: We aim to advance accent-invariant modeling with RNN-T based on the domain adversarial training (DAT). We propose reDAT, a novel technique based on DAT, which relabels data using either unsupervised clustering or soft labels.

**Contributions**

(1) We lay out the theory behind DAT and provide, for the first time, a theoretical guarantee that DAT learns accent-invariant representations.

(2) We also prove that performing the gradient reversal in DAT is equivalent to minimizing the Jensen-Shannon divergence between different output distributions.

(3) We introduce reDAT, a novel technique based on DAT, which refines accent classes with either unsupervised clustering or soft labels. It yields significant improvements over strong baselines.

**Theoretical Guarantee of Accent-Invariance for DAT**

**DAT Framework**

Our DAT framework (illustrated in right part of Figure 1), consists of an accent invariant feature generator \( G \), an English accent classifier \( C \), and RNN-T model \( R \).

During training, weight is updated by the following gradient descent rules:

\[
\begin{align*}
\theta_G &\leftarrow \theta_G - \alpha \left( \frac{\partial L_R}{\partial \theta_G} - \lambda \frac{\partial L_C}{\partial \theta_G} \right), \\
\theta_C &\leftarrow \theta_C - \alpha \frac{\partial L_C}{\partial \theta_C}, \\
\theta_R &\leftarrow \theta_R - \alpha \frac{\partial L_R}{\partial \theta_R}.
\end{align*}
\]

Claims: Performing gradient reversal is equivalent to minimizing Jensen-Shannon divergence (JSD) among multiple domain distributions. The optimization rule for the accent classifier \( C \):

\[
C^* = \arg \max_{\theta_C} \sum_{i=1}^{N} E(z_i) \log C_i(z_i).
\]

Hence, for the optimization of \( G \), by taking the optimal \( C^* \) into the previous expression, it can be deduced to

\[
G^* = \arg \min_{\theta_G} \left( -N \log P_G + \sum_{i=1}^{N} \text{KLD} \left( P_G, \frac{1}{N} \sum_{i=1}^{N} P_{G_i} \right) \right),
\]

where it's equivalent to the JSD between the distributions of all accents:

\[
G^* = \arg \min_{G} \left( -N \log P_G + \text{JSD} \left( P_{G_1}, P_{G_2}, \ldots, P_{G_N} \right) \right).
\]

The global minimum is achieved if and only if

\[
P_{G_1} = P_{G_2} = \ldots = P_{G_N},
\]

which indicates that the embeddings \( z \) are accent-invariant. Please refer to our paper for more mathematical details.

**RedAT: DAT with Relabeling**

**Motivation**: From the theoretical guarantees of DAT, we could get more invariant training results by pre-defining more detailed acoustic information, such as a refined accent label for utterances.

**reDAT**: The DAT approach with relabeling of domain classes either by unsupervised clustering or with soft labels.

**Relabeling with Unsupervised Clustering**

<Step 1>: Train an accent classifier.

<Step 2>: Generate accent embeddings.

<Step 3>: Perform unsupervised clustering on accent embeddings.

<Step 4>: Perform DAT on new labels.

**Relabeling with Soft Labels**

<Step 1>: Train an accent classifier.

<Step 2>: Generate soft labels.

<Step 3>: Perform DAT on new labels.

**Experiments**

**Data set**: 23K-hour en-X data: 13K hours of en-US data, 6K hours of en-GA data, and 4K hours of en-IN. en-AU is set as unseen test set.

**Acoustic features**: 64-dimensional log-Mel features, computed over 25ms windows with 10ms hop length. Feature vector stacked 2 frames to the left.

**Model**: RNN-Transducer. Encoder has 5 LSTM layers with 1024 units, prediction network has 2 LSTM layers with 1024 units.

**Normalized WER**: WER percentage over the reference. For example, Data Pooling is chosen as the reference so that its WER is 1.000, and DAT, as a control, is 0.985.

<table>
<thead>
<tr>
<th>Approach</th>
<th>en-US %</th>
<th>en-GA %</th>
<th>en-AU %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data pooling</td>
<td>1.000</td>
<td>1.472</td>
<td>1.027</td>
</tr>
<tr>
<td>AIPNet-s</td>
<td>0.997</td>
<td>1.425</td>
<td>1.023</td>
</tr>
<tr>
<td>One-hot emb</td>
<td>0.981</td>
<td>1.528</td>
<td>1.010</td>
</tr>
<tr>
<td>Linear emb</td>
<td>0.991</td>
<td>1.442</td>
<td>1.017</td>
</tr>
<tr>
<td>DAT</td>
<td>0.985</td>
<td>1.448</td>
<td>1.012</td>
</tr>
<tr>
<td>reDAT-unsup8</td>
<td>0.969</td>
<td>1.472</td>
<td>0.996</td>
</tr>
<tr>
<td>reDAT-unsup20</td>
<td>0.980</td>
<td>1.470</td>
<td>1.006</td>
</tr>
<tr>
<td>reDAT-soft</td>
<td>0.973</td>
<td>1.409</td>
<td>0.997</td>
</tr>
</tbody>
</table>