GANs for Speech Recognition Data Augmentation

Tianyu Wu
Supervised by Phil Woodland & Chao Zhang

Objectives
Investigate the use of Generative Adversarial Networks (GANs) for acoustic data generation given an initial small training set.
- Build a trainable generator for phone units and context windows.
- Develop the training pipeline to improve the pre-trained acoustic model’s performance with augmented data.

Background
Performance of a speech recogniser improves with increased training data size, **but**
- collecting a large matched training dataset is difficult,
- manually transcribing speech data is expensive.
Generative Adversarial Networks are a method which can generate simulated data [1].

Methodology
- Perform frame level speech data generation.
- Train separate GANs for each phonetic unit.
- Continue training the pre-trained acoustic model with augmented data (supervised training).

Generative Adversarial Networks
GANs are a framework to estimate generative models via an adversarial-process:
- Discriminator is trained to perform classification between true data and fake data.
- Generator is trained to generate fake samples to fool the discriminator.

The configuration of our GAN:
- Use deep convolutional structure.
- Use spectral normalization technique [2].
- Condition the model on phone states [3].
- Generate data on the speech feature level (FBANK).

Methodology
- Perform frame level speech data generation.
- Train separate GANs for each phonetic unit.
- Continue training the pre-trained acoustic model with augmented data (supervised training).

Generated Fake Data
Evaluate the generated data based on the pre-trained acoustic model.

![Generated Fake Data](image-url)

Table 1: Top1, Top3 and Top5 Classification Accuracy

<table>
<thead>
<tr>
<th>Phone</th>
<th>Generated fake data</th>
<th>Test set data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%top1</td>
<td>%top3</td>
</tr>
<tr>
<td>aa[2]</td>
<td>30.2</td>
<td>60.7</td>
</tr>
<tr>
<td>aa[3]</td>
<td>28.8</td>
<td>60.8</td>
</tr>
<tr>
<td>t[2]</td>
<td>35.2</td>
<td>57.8</td>
</tr>
<tr>
<td>th[2]</td>
<td>45.1</td>
<td>69.6</td>
</tr>
</tbody>
</table>

Upcoming Work
- Improve acoustic model performance with augmented data.
- Extend to triphone system.

References