Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio

Shahin Amiriparian $^{1,2,3}$, Michael Freitag $^3$, Nicholas Cummins $^{1,2}$, Björn Schuller $^{1,4}$

$^1$ Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
$^2$ Machine Intelligence & Signal Processing Group, TU München, Germany
$^3$ Chair of Complex and Intelligent Systems, University of Passau, Germany
$^4$ GLAM – Group on Language, Audio & Music, Imperial College London, London, UK
Introduction

Why unsupervised representation learning?
• Tedious to manually design feature sets
• Abundant unlabelled data
• More robust to overfitting

Current state-of-the-art: Deep Neural Networks
• Stacked Autoencoders
• Restricted Boltzmann Machines
• (Deep Convolutional) Generative Adversarial Networks
Representation Learning from Acoustic Data

Most current representation learning approaches

– Inputs of fixed dimensionality
– No explicit consideration of the sequential nature of audio

Alternative: Sequence to sequence learning models

– Proposed in machine translation
– Based on Recurrent Neural Networks (RNNs)
– Learn fixed-length representations of variable-length input
System Architecture

Sequence to Sequence Autoencoders for Unsupervised Representation Learning

Amiriparian, et al.
Spectrogram Extraction

• Hann windows with width $w$ and overlap $0.5w$

• Computing a given number $N_m$ of log-scaled Mel frequency bands

• Normalising the Mel-spectra $[-1, 1]$

• Stereo data
  – Right, left, mean, and difference of the channels

---

Recurrent Sequence to Sequence Autoencoders

Sequence to Sequence Autoencoders for Unsupervised Representation Learning

Amiriparian, et al.
Experimental Settings

Common Experimental Settings

• Implementation: as part of auDeep toolkit
  – for deep representation learning from audio
    https://github.com/auDeep/auDeep

• The autoencoders and MLPs are trained using the Adam optimizer
  – fixed learning rate of 0.001

---

Experimental Settings

Settings for the autoencoders

• Number of epochs: 50
• Batch size: 64
• Dropout: 20%
  – Applied to the output of each recurrent layer
• Gradients with absolute value above 2 were clipped

---

Experimental Settings

Settings for the MLPs

• Number of epochs: 400
• Without batching
• Without gradient clipping
• Dropout: 40%
  – Applied to the hidden layers
Experimental Settings

Hyperparameter Optimisation

(a)

Accuracy [%]

0.76 0.80 0.84

Window width [s]

0.05 0.15 0.25 0.35

(b)

0.76 0.80 0.84

Mel frequency bands

40 80 160 320 640
# Results

## Fusion Experiments

<table>
<thead>
<tr>
<th>System</th>
<th>Features</th>
<th>Accuracy [%]</th>
<th>Devel.</th>
<th>Eval.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>200 (per frame)</td>
<td>74.8</td>
<td>61.0</td>
<td></td>
</tr>
<tr>
<td>Proposed: Individual Feature Sets</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (M)</td>
<td>1024</td>
<td>85.0</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Left (L)</td>
<td>1024</td>
<td>84.6</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Right (R)</td>
<td>1024</td>
<td>83.8</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Difference (D)</td>
<td>1024</td>
<td>82.0</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Proposed: Fused Feature Sets</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean, Left</td>
<td>2048</td>
<td>86.2</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Mean, Left, Right</td>
<td>3072</td>
<td>86.9</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>All (M + L + R + D)</td>
<td>4096</td>
<td><strong>88.0</strong></td>
<td>67.5</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions and Future Work

Conclusions

• Promising results with sequence to sequence autoencoders

• Effective alternative to expert-designed feature sets

• Fully unsupervised training

• Variable-length input
Conclusions and Future Work

Further research

• Comparison/fusion with Deep Convolutional Generative Adversarial Networks

• Feature selection and dimensionality reduction

• Using CAS²T to gather more “in-the-wild” data

---

* S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, “CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms,” in Proc. 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017), (San Antonio, TX), AAAI, IEEE, October 2017. 6 pages
References

Dropbox download link for:

- Presentation slides
- Paper
- References