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

Authors of the paper


Shahin Amiriparian


Michael Freitag


Nicholas Cummins


Björn Schuller

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

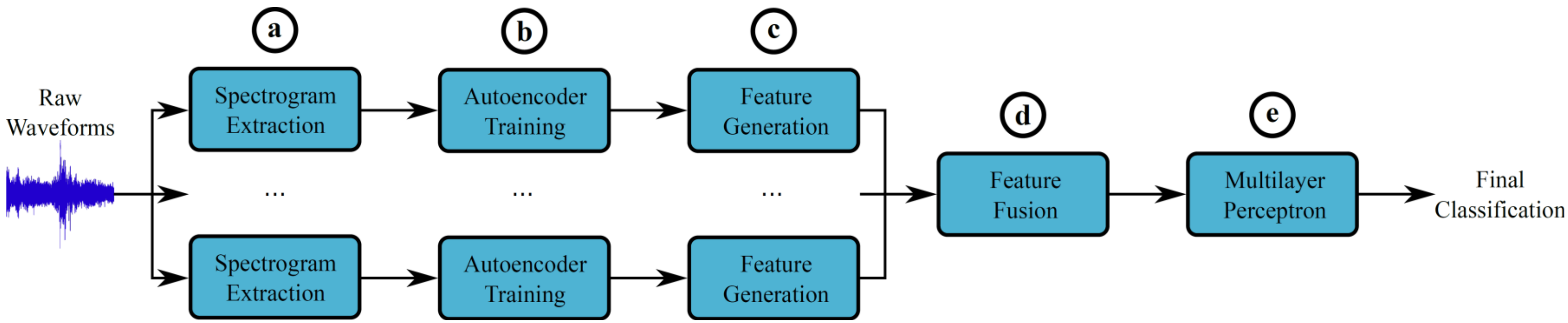
Most current representation learning approaches

- 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

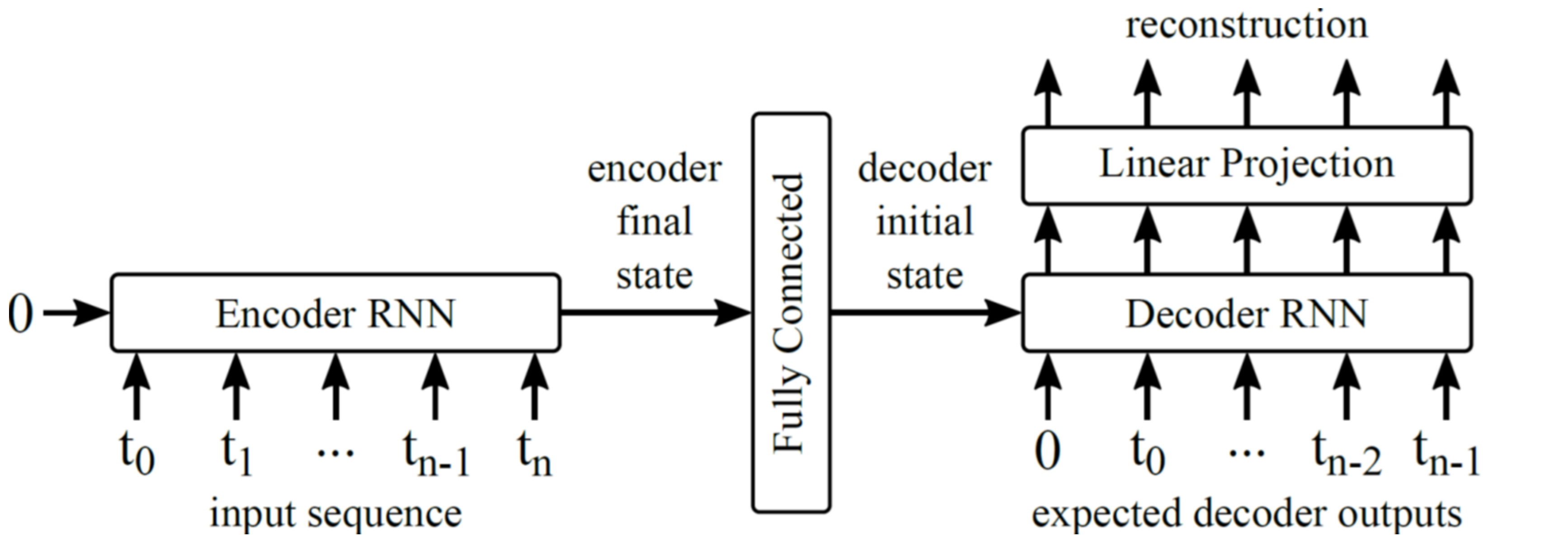


Spectrogram Extraction

- Hann windows with width $\boldsymbol{w}$ and overlap $\mathbf{0.5}\boldsymbol{w}$
- Computing a given number $\boldsymbol{N_m}$ of log-scaled Mel frequency bands
- Normalising the Mel-spectra $[-\mathbf{1}, \mathbf{1}]$
- Stereo data
  - Right, left, mean, and difference of the channels

- H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G.Widmer, "CPJKU submissions for DCASE-2016: A hybrid approach usingbinaural i-vectors and deep convolutional neural networks," Detection and Classification of Acoustic Scenes and Events 2016 IEEE AASP Challenge (DCASE 2016), Sep 2016

Recurrent Sequence to Sequence Autoencoders



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

- M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B.Schuller. auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks, Journal of Machine Learning Research, 2017, submitted, 5 pages
- D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014, 15 pages.

Experimental Settings

Settings for the **autoencoders**

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

- I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q.Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112.

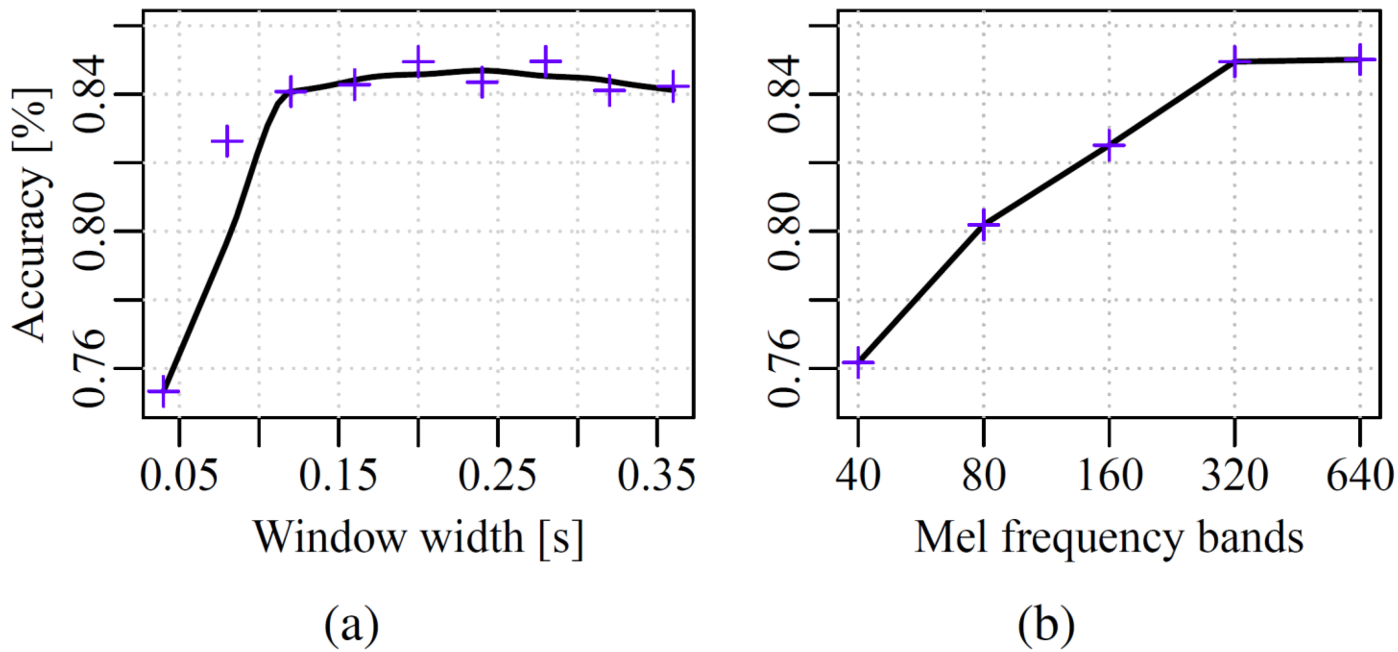
Experimental Settings

Settings for the **MLPs**

- Number of epochs: 400
- Without batching
- Without gradient clipping
- Droupout: 40%
  - Applied to the hidden layers

Experimental Settings

Hyperparameter Optimisation



Results

Fusion Experiments

System	Features	Accuracy [%] Devel.	Eval.
Baseline	200 (per frame)	74.8	61.0
Proposed: Individual Feature Sets			
Mean (M)	1 024	85.0	–
Left (L)	1 024	84.6	–
Right (R)	1 024	83.8	–
Difference (D)	1 024	82.0	–
Proposed: Fused Feature Sets			
Mean, Left	2 048	86.2	–
Mean, Left, Right	3 072	86.9	–
All (M + L + R + D)	4 096	**88.0**	67.5

Conlusions and Future Work

Conclusions

- Promising results with sequence to sequence autoencoders
- Effective alternative to expert-designed feature sets
- Fully unsupervised training
- Variable-length input

Conlusions and Future Work

Further research

- Comparison/fusion with Deep Convolutional Generative Adversarial Networks
- Feature selection and dimensionality reduction
- Using CAS^2T 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), AAAC, IEEE, October 2017. 6 pages

References

Dropbox download link for:

- Presentation slides
- Paper
- References