Journal of Machine Learning Research 18 (2018) 1-5 Submitted 7/17; Revised 12/17; Published 4/18

auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks

Michael Freitag FREITAGM@FIM.UNI-PASSAU.DE
Shahin Amiriparian SHAHIN.AMIRIPARIAN@TUM.DE
Sergey Pugachevskiy PUGACHEV@FIM.UNI-PASSAU.DE
Nicholas Cummins NICHOLAS.CUMMINS@IEEE.ORG
Björn Schuller SCHULLER@IEEE.ORG
Chair of Embedded Intelligence for Health Care & Wellbeing
Augsburg University, Augsburg, Germany &
Chair of Complex & Intelligent Systems
Universität Passau, 94032 Passau, Germany &
GLAM – Group on Language, Audio & Music
Imperial College London, London, UK

Editor: Geoff Holmes

Abstract

auDeep is a Python toolkit for deep unsupervised representation learning from acoustic data. It is based on a recurrent sequence to sequence autoencoder approach which can learn representations of time series data by taking into account their temporal dynamics. We provide an extensive command line interface in addition to a Python API for users and developers, both of which are comprehensively documented and publicly available at `https://github.com/auDeep/auDeep`. Experimental results indicate that auDeep features are competitive with state-of-the art audio classification.

Keywords: deep feature learning, sequence to sequence learning, recurrent neural networks, autoencoders, audio processing

1. Introduction

Machine learning approaches for audio processing commonly operate on a variety of hand-crafted features computed from raw audio signals. Considerable effort has been put into developing high-performing feature sets for specific tasks. Recently, representation learning, in particular deep representation learning, has received significant attention as a highly effective alternative to using such conventional feature sets (Bengio et al., 2013; Schmitt and Schuller, 2017). These techniques have been shown to be superior to feature engineering for a plethora of tasks, including speech recognition and music transcription (Bengio et al., 2013; Amiriparian et al., 2016). Sequential data such as audio, however, poses challenges for deep neural networks, as they typically require inputs of fixed dimensionality. In this regard, sequence to sequence learning with *recurrent neural networks* (RNNs) has been proposed in machine translation, for learning fixed-length representations of variable-length sequences (Sutskever et al., 2014).

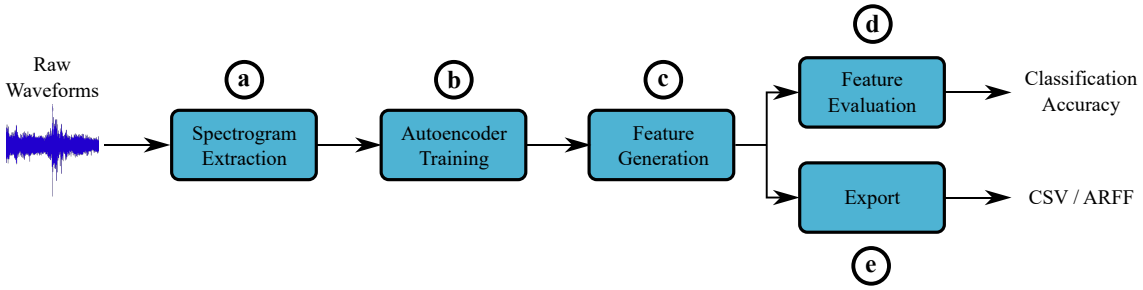

Figure 1: Illustration of the feature learning procedure with auDeep. A detailed description of the procedure is given in Section 3.1.

In this paper, we present auDeep, a first of its kind TensorFlow based Python toolkit for deep unsupervised representation learning from acoustic data. This is achieved using a recurrent sequence to sequence autoencoder approach. auDeep can be used both through its Python API as well as through an extensive command line interface.

2. Recurrent Sequence to Sequence Autoencoders

Our implementation of sequence to sequence autoencoders extends the RNN encoder-decoder model proposed by Sutskever et al. (2014). The input sequence is fed to a multi-layered *encoder* RNN which collects key information about the input sequence in its hidden state. The final hidden state of the encoder RNN is then passed through a fully connected layer, the output of which is used to initialise the hidden state of the multilayered *decoder* RNN. The function of the decoder RNN is to reconstruct the input sequence based on the information contained in the initial hidden state. The network is trained to minimise the root mean squared error between the input sequence and the reconstruction. In order to accelerate model convergence, the expected decoder output from the previous step is fed back as the input into the decoder RNN (Sutskever et al., 2014). Once training is complete, the activations of the fully connected layer are used as the representation of an input sequence.

For the purposes of representation learning from acoustic data, we train sequence to sequence autoencoders built of *long short-term memory* cells or *gated recurrent units* on spectrograms, which are viewed as time dependent sequences of frequency vectors. Two of the key strengths of this approach are (i) fully unsupervised training, and (ii) the ability to account for the temporal dynamics of sequences.

3. System Overview

auDeep contains at its core a high-performing implementation of sequence to sequence autoencoders which is not specifically constrained to acoustic data. Based on this domain-independent implementation, we provide extensive additional functionality for representation learning from audio.

3.1 Practical Usage

An illustration of the feature learning procedure with auDeep is shown in Figure 1. First, spectrograms are extracted from raw audio files (cf. Figure 1a). Then, a sequence to sequence autoencoder, as previously described, is trained on the extracted spectrograms (cf. Figure 1b), and the learned representation of each instance is extracted as its feature vector (cf. Figure 1c). If instance labels are available, a classifier can then be trained and evaluated on the extracted features (cf. Figure 1d). Finally, the extracted features, and any associated metadata, can be exported to CSV or ARFF for further processing, such as classification with alternate algorithms (cf. Figure 1e).

While our system is capable of learning representations of the extracted spectrograms entirely without additional metadata, we provide the possibility to parse instance labels, data set partitions, or a cross-validation setup from a variety of common formats. If available, metadata is stored alongside the extracted spectrograms, and can be used, e. g. for evaluation of a classifier on the learned representations. To demonstrate the strength of the learned representations two conventional classifiers are built into the toolkit: a *Multilayer Perceptron* (MLP) with softmax output, and an interface to LibLINEAR (Fan et al., 2008).

3.2 Design

auDeep provides a highly modularised Python library for deep unsupervised representation learning from audio. The core sequence to sequence autoencoder models are implemented using TensorFlow. This implementation substantially extends the built-in sequence to sequence learning capabilities of TensorFlow; for example, the RNNs allow probabilistic feedback and are also reusable, self-contained modules. Furthermore, diversely structured data sets are handled by the system in a unified way without requiring time-consuming manual adjustments. The topology and parameters of autoencoders are stored as TensorFlow checkpoints which can be reused in other applications. This enables users, e. g. to pretrain the encoder RNN with auDeep and subsequently apply custom retraining. Data sets are represented in binary format using the platform-independent NetCDF standard, but we provide tools for converting data between NetCDF and CSV/ARFF.

Users are given fine-grained control over the representation learning process through the Python API and the command line interface. The system is platform-independent, and has been tested on Windows and various Linux distributions on desktop PCs and in a cluster environment. auDeep is capable of running on CPU only, and GPU-acceleration is leveraged automatically when available.

4. Experiments

We demonstrate the capabilities of auDeep on three audio classification tasks. First, we perform acoustic scene classification on the development partition of the TUT Acoustic Scenes 2017 (TUT AS 2017) data set (Mesaros et al., 2017). Furthermore, we conduct environmental sound classification (ESC) on the ESC-10 and ESC-50 data sets (Piczak, 2015b), and, finally, we perform music genre classification on the GTZAN data set (Tzanetakis and Cook, 2002).

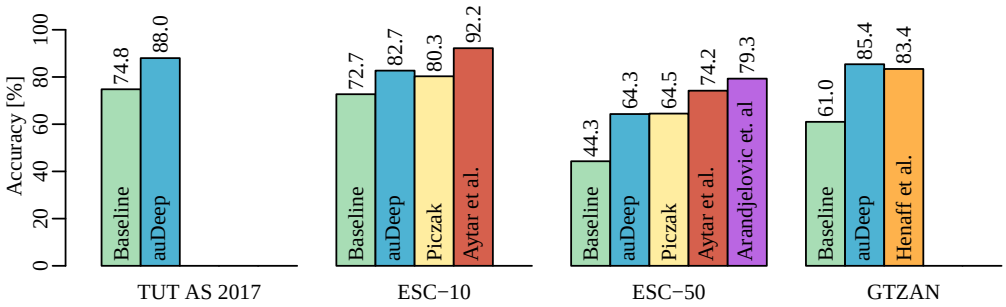

Figure 2: Comparison of auDeep to baselines and state-of-the-art on four data sets.

We train multiple autoencoder configurations using auDeep, and perform feature-level fusion of the learned representations. The fused representations are evaluated using the built-in MLP with the same cross-validation setup as used for the baseline systems on the TUT AS 2017, ESC-10, and ESC-50 data sets. For GTZAN, no predefined cross-validation setup is available, therefore we randomly generate five stratified cross-validation folds (cf. Figure 2, auDeep). Due to space limitations, we refrain from detailing our full experimental setup. However, to ensure reproducibility, we distribute the codes and parameter choices used for these experiments with auDeep.

Finally, we compare the performance of auDeep with baseline and state-of-the-art approaches for the different datasets (cf. Figure 2, identified by authors' names). We observe that auDeep either matches or outperforms a *convolutional neural network* approach (Piczak, 2015a) and a representation learning approach for ESC-10 and ESC-50, and a sparse coding approach for GTZAN (Henaff et al., 2011). The SoundNet (Aytar et al., 2016) and L3 (Arandjelovic and Zisserman, 2017) systems did achieve stronger performances on ESC-10 and ESC-50. However, this is not a straightforward comparison, as auDeep was trained using ESC-10 and ESC-50 data only whilst L3 and SoundNet were pre-trained on external corpora of 500 000 and 2+ million videos, respectively. For further details and comparisons with state-of-the-art for the TUT Acoustic Scenes 2017 corpus, the reader is referred to Amiriparian et al. (2017).

5. Conclusions

auDeep is an easy-to-use, open-source toolkit for deep unsupervised representation learning from audio with competitive performance on various audio classification tasks. Our long-term goal is to grow auDeep into a general-purpose deep audio toolkit, by integrating other deep representation learning algorithms such as conditional variational sequence to sequence autoencoders or *deep convolutional generative adversarial networks*, and by extending the feature learning functionality to regression tasks on continuously labelled data.

Acknowledgments

This research has received funding from the EU's 7^{th} Framework Programme through the ERC Starting Grant No. 338164 (iHEARu) and from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902.

References

S. Amiriparian, J. Pohjalainen, E. Marchi, S. Pugachevskiy, and B. Schuller. Is deception emotional? An emotion-driven predictive approach. In *INTERSPEECH*, pages 2011–2015. ISCA, 2016.

S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller. Sequence to sequence autoencoders for unsupervised representation learning from audio. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, pages 17–21. IEEE, 2017.

R. Arandjelovic and A. Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision*, pages 609–617. IEEE, 2017.

Y. Aytar, C. Vondrick, and A. Torralba. SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems 29*, pages 892–900. Curran Associates, Inc., 2016.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828, 2013.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. LeCun. Unsupervised learning of sparse features for scalable audio classification. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 681–686. ISMIR, 2011.

A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*. IEEE, 2017.

K. J. Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2015a.

K. J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018. ACM, 2015b.

M. Schmitt and B. W. Schuller. openXBOW—Introducing the Passau open-source crossmodal bag-of-words toolkit. *Journal of Machine Learning Research*, 18:1–5, 2017.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.