COMBINING HIGH-LEVEL FEATURES OF RAW AUDIO WAVES AND MEL-SPECTROGRAMS FOR AUDIO TAGGING

Universität Konstanz

Marcel Lederle and Benjamin Wilhelm

University of Konstanz

Overview

Objective:

- Classification of a diverse set of audio files: Task 2 of the DCASE 2018 Challenge
- Achieving high accuracy without excessive ensembling of different models

Approach:

- One-dimensional CNN trained on raw-audio data
- Two-dimensional CNN trained on mel-spectrograms
- Combining both CNNs by densely connected layers within a single network
- Data augmentation

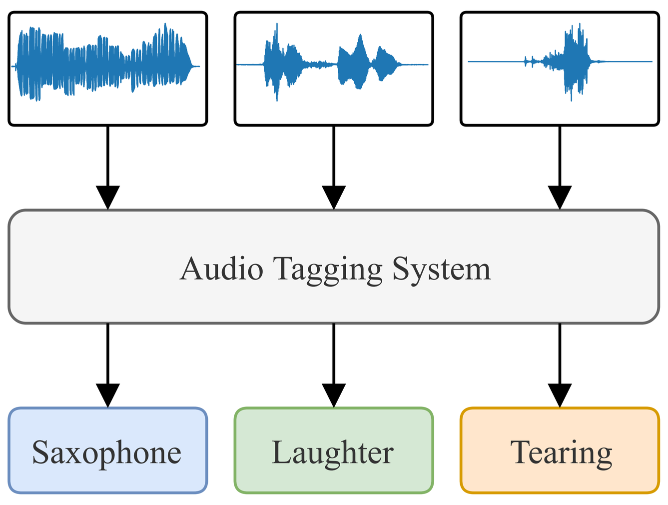

Figure 1: Illustration of the task by [1].

Data Augmentation

- **Time shifting (1)**
  Random shift in time dimension
- **Random cropping (2)**
  Crop input to match input size of CNNs
- **Random padding (3)**
  Pad input with zeros to match input size of CNNs
- **Replication (4)**
  Replicate input several times
- **Mixup (5)**
  Blend multiple audio clips of same or different classes [4]

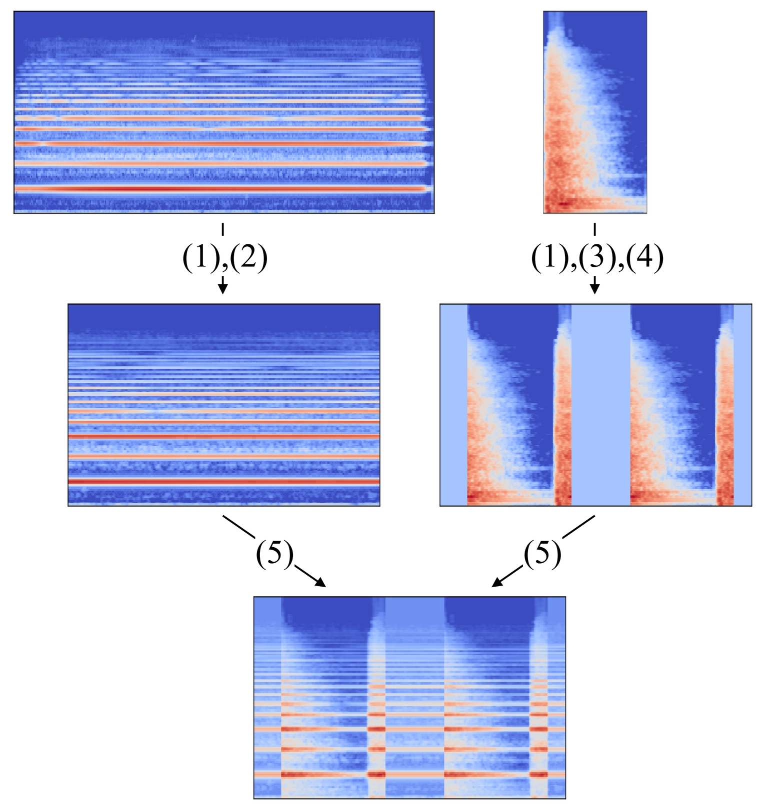

Figure 3: Visualization of augmentation techniques for mel-spectrograms.

Method and Network Architecture

***cnn-spec* and *cnn-audio*:**

- Architecture is similar to common image classification CNNs (VGG19 [2], AlexNet [3])
- Batch Normalization after each block and dense layer
- ReLU after each convolutional and dense layer
- *cnn-audio* and *cnn-spec* are trained separately on raw-audio waves and log-scaled mel-spectrograms, respectively

***cnn-comb*:**

- *cnn-audio* and *cnn-spec* are joined by removing the softmax and dense layers, concatenating the output features, and adding a densely connected neural network
- The transferred weights are kept fixed while training the new layers

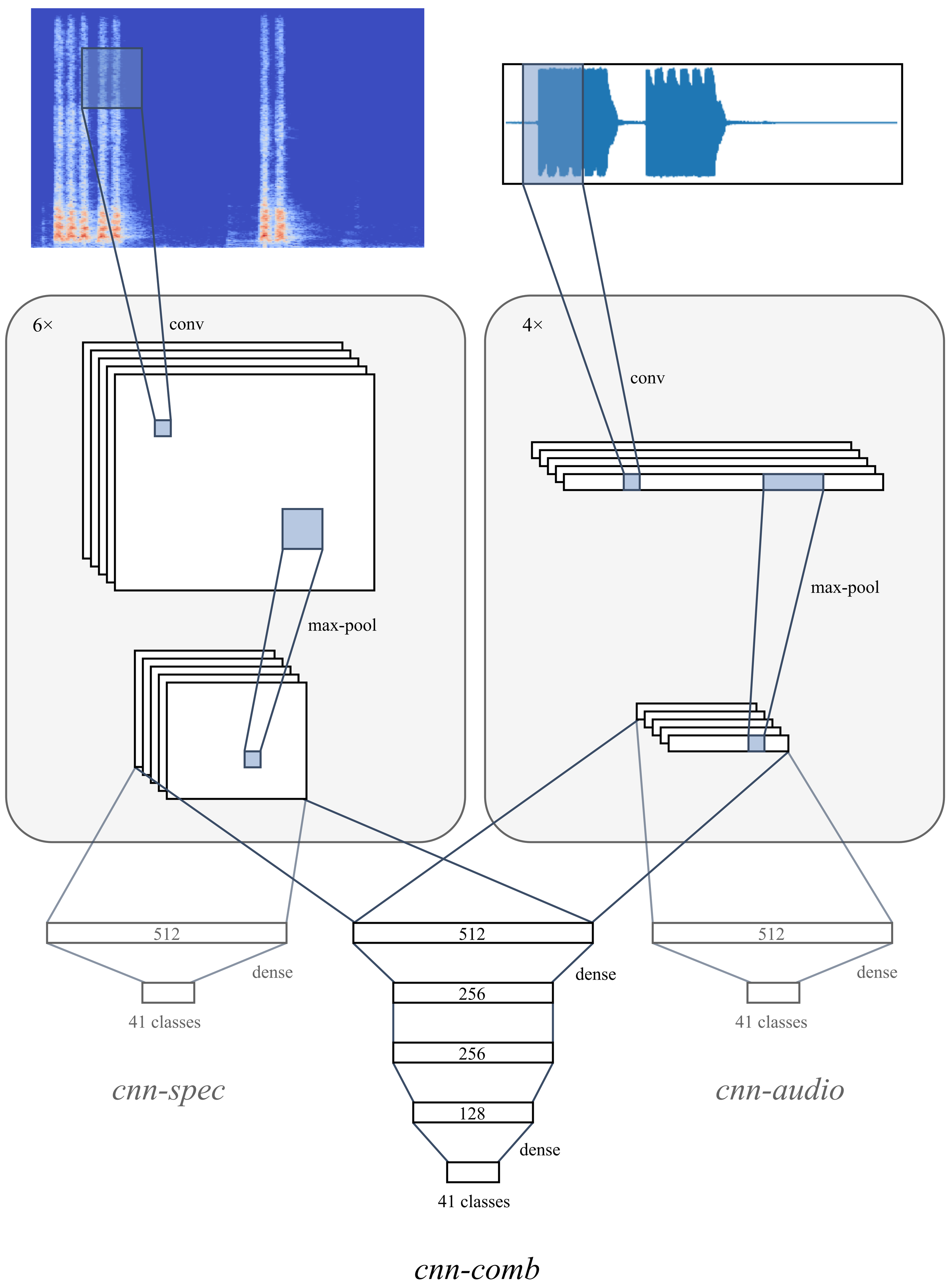

Figure 2: Illustrated architecture of the models.

Evaluation

Model	Input length	Public score	Private score	Total score
cnn-audio	1 sec	0.920	0.888	0.894
	2 sec	0.921	0.884	0.891
	3 sec	0.935	0.889	0.898
cnn-spec	1 sec	0.930	0.923	0.924
	2 sec	0.950	0.928	0.932
	3 sec	0.935	0.930	0.931
cnn-comb	1 sec	0.955	0.939	0.942
	2 sec	**0.966**	**0.944**	**0.948**
	3 sec	0.956	**0.944**	0.946

Table 1: Evaluation results (MAP@3) of the individual models on the public (301 samples), private (1299 samples), and full test set of the DCASE 2018 Challenge on Kaggle.

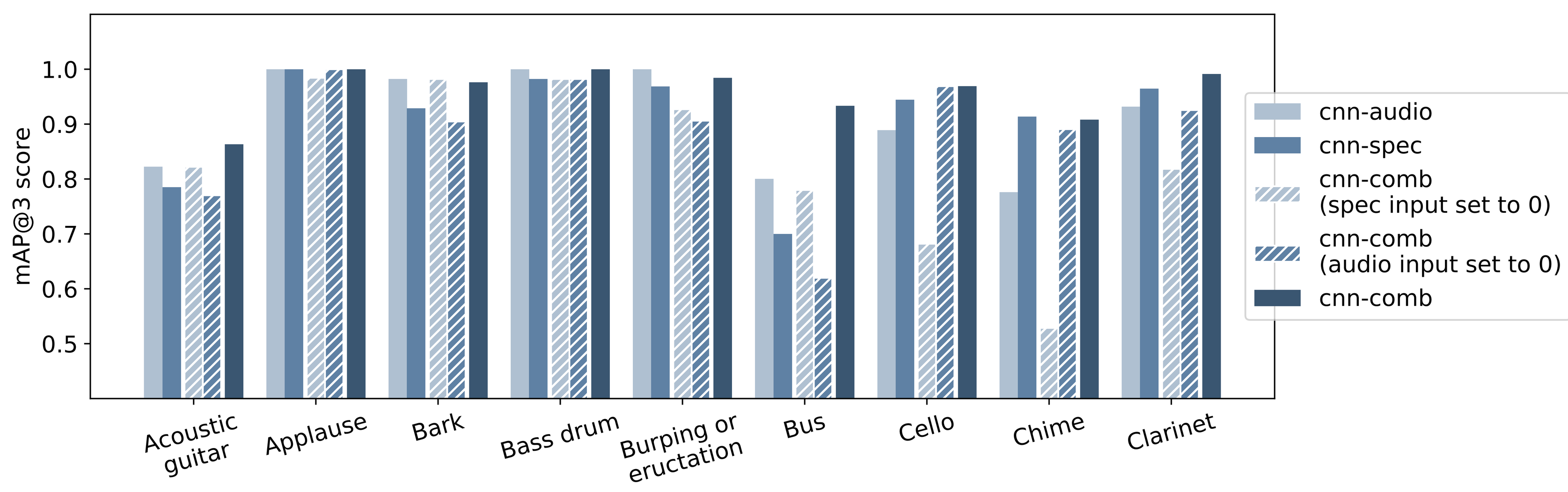

Figure 4: Comparison of per-category scores of single-input models, combined models with one input alternately set to zero, and the combined model with both inputs. The mAP@3 score is reported on a single fold for each model.

- *cnn-audio* $\overset{\text{MAP@3}}{>}$ *cnn-spec* : $\textit{cnn-comb}_{\text{spec\_input=0}} \overset{\text{MAP@3}}{>} \textit{cnn-comb}_{\text{audio\_input=0}}$
- *cnn-spec* $\overset{\text{MAP@3}}{>}$ *cnn-audio* : $\textit{cnn-comb}_{\text{audio\_input=0}} \overset{\text{MAP@3}}{>} \textit{cnn-comb}_{\text{spec\_input=0}}$
- *cnn-comb* with both inputs performs best

$\Longrightarrow$ *cnn-comb* uses high-level features of both models

$\Longrightarrow$ *cnn-comb* focuses on the features of the superior model

Conclusion

Extending current Convolutional Neural Network approaches that only make use of a frequency representation by adding a second input that incorporates the raw audio wave, has improved the mAP@3 score significantly. We have demonstrated the capabilities of our model by competing in the Freesound General-Purpose Audio Tagging Challenge on Kaggle and ranking in the top two percent of all participants.

References

[1] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," *arXiv preprint arXiv:1807.09902*, 2018.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.