Emotion and Theme Recognition in Music with Frequency-Aware Receptive Field Regularized CNNs

Khaled Koutini, Shreyan Chowdhury, Verena Haunschmid,
Hamid Eghbal-Zadeh, Gerhard Widmer

MediaEval 2019
10th Anniversary Workshop
27-29 October 2019
EURECOM, Sophia Antipolis, France

JOHANNES KEPLER UNIVERSITY LINZ

Institute of Computational Perception

Agenda

- Overview
- Approaches
  - What worked
  - What didn’t work
- Results
- Conclusion

Takeaways

- **Insights**

    Make the network see less of the input to avoid overfitting.

    Frequency context helps.

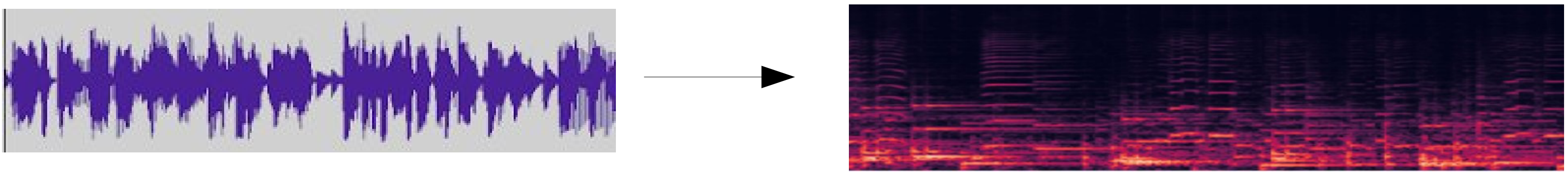
    Ensembling helps.

- **Results**

Validation PR-AUC	Testing PR-AUC
0.1189	0.1546
-	0.1077 (baseline)

Overview

- Emotion and theme recognition task – instance of auto-tagging.
- Current state-of-the-art in many audio tasks, including audio-tagging uses CNNs in a VGG-like architecture.
- Input: Time-Frequency representation of audio (*Spectrograms*)



Overview

- The success of CNNs started in computer vision tasks.
- Later improvements on CNNs architectures made the networks deeper.
- Deeper architectures such as ResNet and DenseNet don’t perform as well in audio processing tasks.

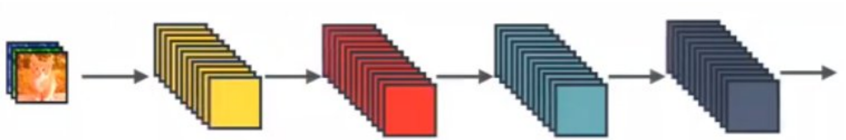
Approach: Baseline Models

- VGG-like
- ResNet18, ResNet34, ResNet50
- CRNN – Convolutional Recurrent Neural Network

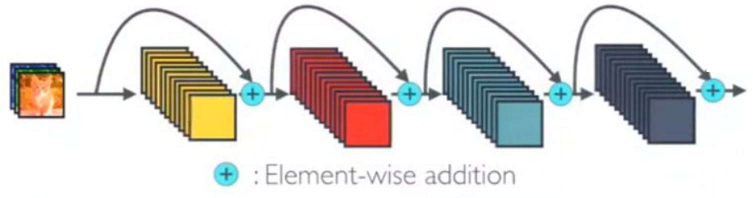
Model	Validation PR-AUC	Testing PR-AUC
VGG-like	-	0.1077
ResNet34	0.0924	0.1021
CRNN	0.0924	0.1172

Deeper = Better?

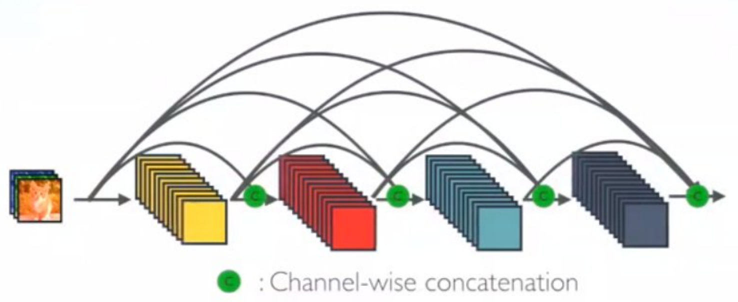
- *ResNet* [1] and *DenseNet* [2] variants outperform earlier (and shallower) *VGG*-based [3] variants by a significant margin (Vision tasks).
- They address shortcomings of VGG such as the vanishing gradient.



ConvNet



ResNet



DenseNet

Sources: Figures from https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803

Deeper = Better? Can lead to overfitting

Hershey et al. [8] compared various vision CNNs on a large-scale dataset of 70M audio clips from YouTube.

- ResNet-50 can perform very well.
- However, training such deep architectures on smaller datasets results in heavy overfitting on the training samples.

The Receptive Field in CNNs

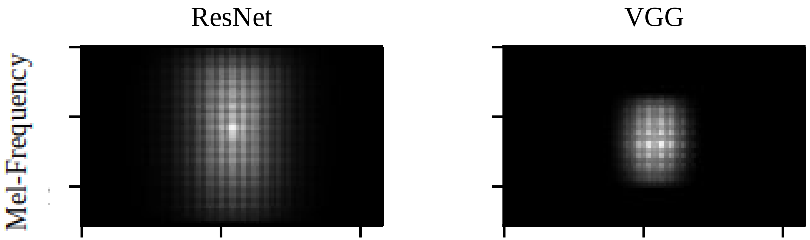
- In fully-connected layers, each neuron is affected by the whole input. In contrast, in convolutional layers each neuron has a strictly limited 'field of view' (*RF*).
- Input values outside of this RF cannot influence the neuron's activation.
- The maximum RF can be calculated:

$$S_n = S_{n-1} * s_n$$
$$RF_n = RF_{n-1} + (k_n - 1) * S_n$$

s_n, k_n are stride and kernel size of layer n, respectively, and S_n, RF_n are cumulative stride and RF of a unit from layer n to the network input.

The Effective Receptive Field in CNNs

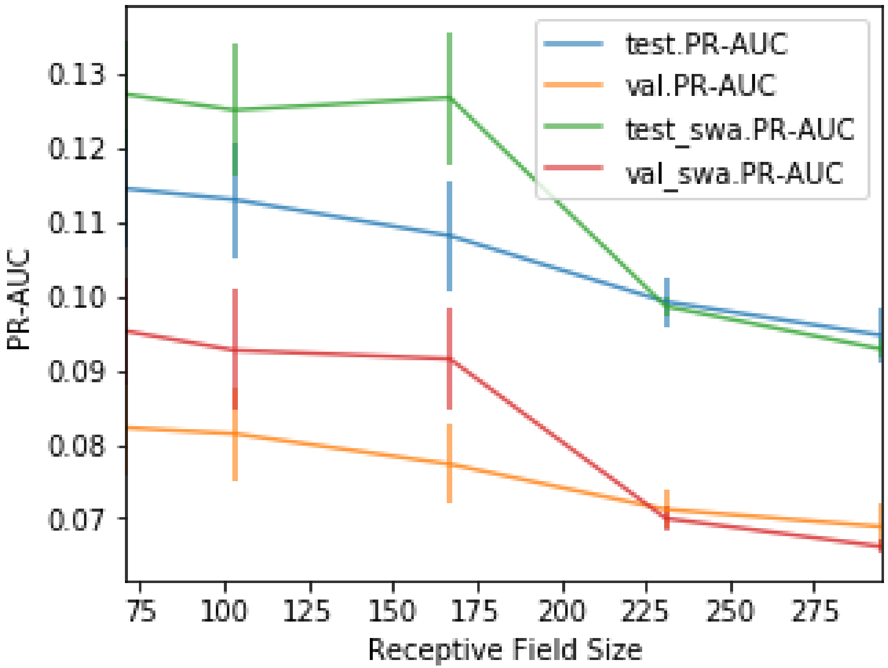
- A neuron may not actually use all the available receptive field.
- The set of input pixels or units that effectively influence a neuron is called its *Effective Receptive Field (ERF) by* Luo et al. [14]



Adapting the RF of Vision Architectures

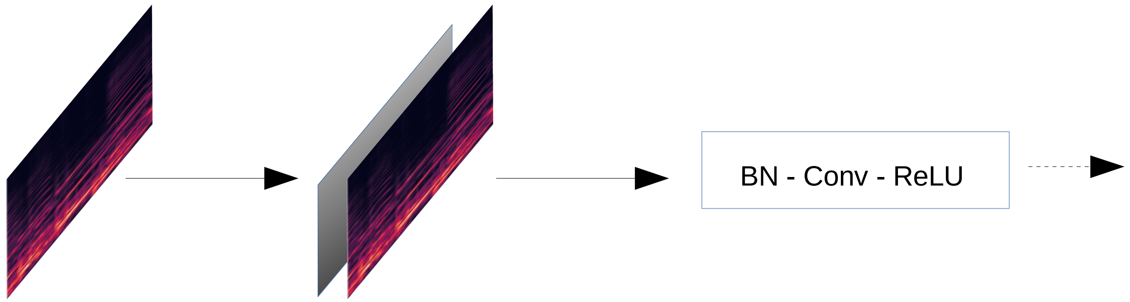
- **Changing filter sizes** to change the maximum receptive field.
- We changed some filter sizes from 3×3 to 1×1.
- Filter size is a hyperparameter.

Results with Receptive Field Adaptation

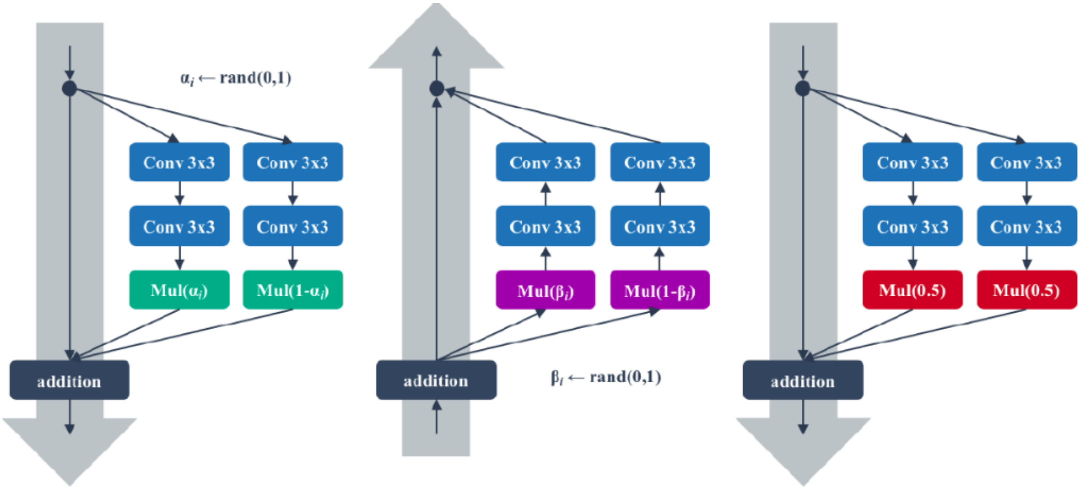


Frequency Aware Networks

- Drawback of CNN for audio domain: lack of spatial ordering in convolutional layers.
- Solution: Add a channel that encodes frequency as a value in [-1, 1].



Shake-Shake Regularization



Left: Forward training pass. **Center:** Backward training pass. **Right:** At test time.

Image source: Gastaldi, X., 2017. Shake-shake regularization. arXiv preprint arXiv:1705.07485.

Ensembling

- Stochastic Weight Averaging
  - Maintain a paired network with running average of weights
- Snapshot Averaging
  - Average the predictions of the 5 snapshots of the model during training
- Multi-model averaging
  - Average predictions from models with different initializations, RFs, etc.

Results

Submission	Validation PR-AUC	Testing PR-AUC
ShakeFAResNet*	.1132	.1480
FAResNet*	.1149	.1463
Avg_ensemble*	**.1189**	**.1546**
ResNet34	.0924	.1021
CRNN	.0924	.1172
CP_ResNet	**.1097**	**.1325**
VGG-ish-baseline	-	.1077
popular baseline	-	.0319

*: indicates an ensemble.

Conclusions

- Receptive field regularization is useful to avoid overfitting.
- Frequency-aware networks improve performance (spectrogram as input).
- Ensembling improved results.

- Future work:
  - Temporal context?
  - Perceptual features?

Thank you!
Questions?

MediaEval 2019
10th Anniversary Workshop
27-29 October 2019
EURECOM, Sophia Antipolis, France

shreyan.chowdhury@jku.at

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Str. 69
4040 Linz, Austria
www.jku.at