TU WIEN Informatics

[1]Faculty of Informatics, TU Wien, Austria
[2]School of Music, Georgia Institute of Technology, USA

On the Impact and Interplay of Input Representations and Network Architectures for Automatic Music Tagging

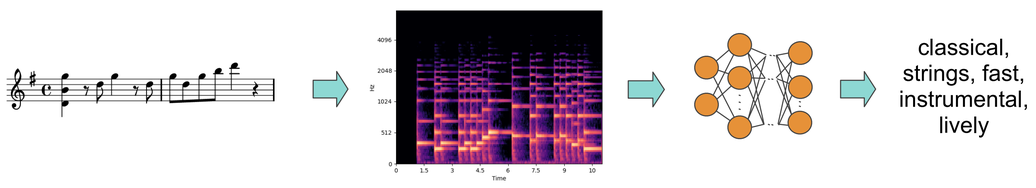
Maximilian Damböck[1]
maxi.damboeck@gmail.com

Richard Vogl[1]
richard.vogl@tuwien.ac.at

Peter Knees[1,2]
peter.knees@tuwien.ac.at

Automatic Music Tagging

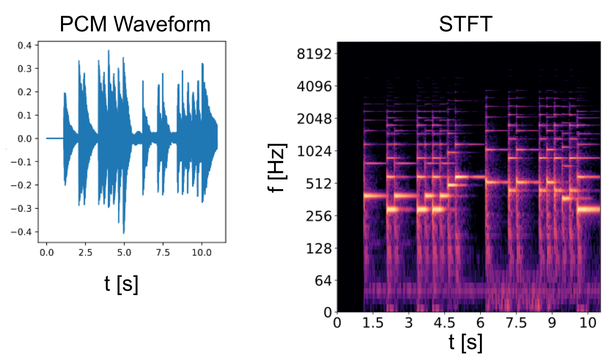
E.g.: Eine kleine Nachtmusik - W. A. Mozart



Preprocessing Methods

Comparison of **five** input representations:

- **Raw** PCM Waveform
- **STFT** magnitude spec.
- **Mel** spectrogram
- **CQT** spectrogram
- **MFCC**s



Network Architectures

Comparison of **five** network architectures:

- **VGG-16** [1]
- **ResNet** [2]
- **SENet** [3]
- **Musicnn** [4]
- Musicnn w/ dilated CNN frontend (**Dilated CNN**)

Experiment Setup

Datasets:

- MagnaTagATune (~25 000 examples)
- MTG-Jamendo (~55 000 examples)

Evaluation:

- Multiple runs (5-7) per configuration (5x5 configurations)
- ROC-AUC and PR-AUC metrics

Statistical analysis:

- Impact model — preprocessing method (Two-Way ANOVA)
- Rank preprocessing methods (Tukey-HSD range test)

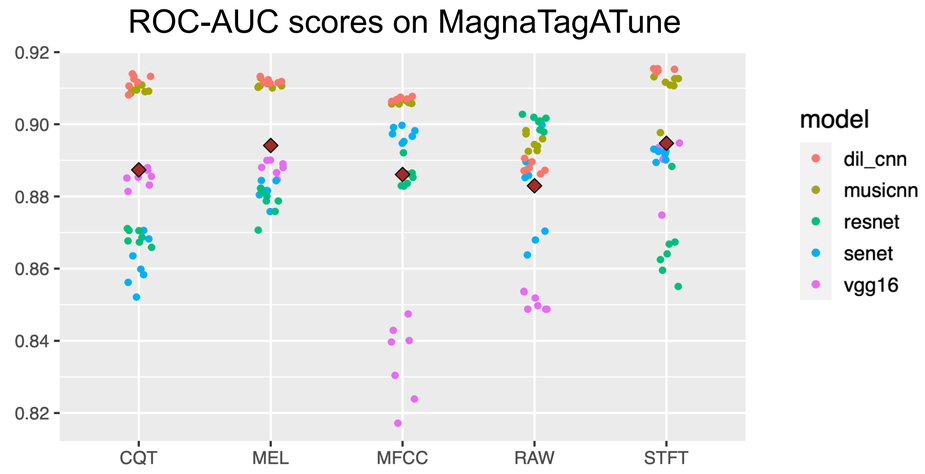
[1] K. Simonyan and A. Zisserman, "**Very deep convolutional networks for large-scale image recognition**," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

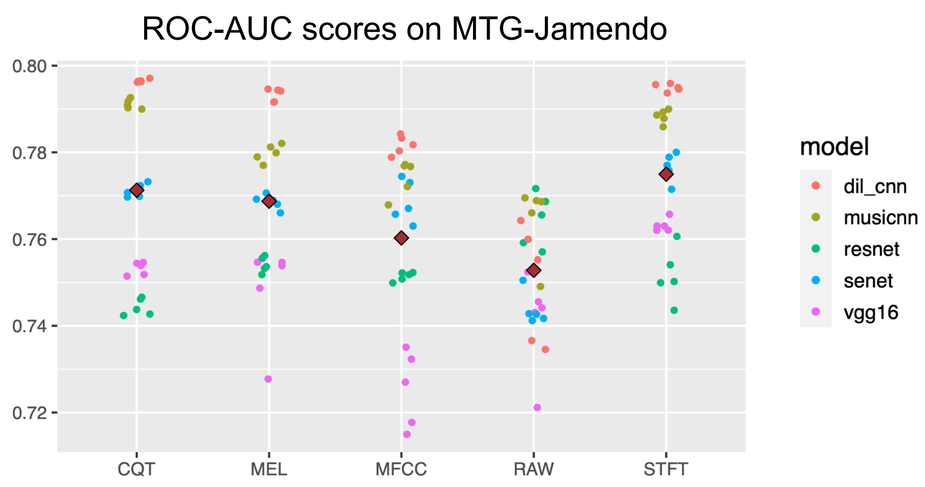
[2] K. He, X. Zhang, S. Ren, and J. Sun, "**Deep residual learning for image recognition**," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, 2016.

[3] T. Kim, J. Lee, and J. Nam, "**Sample-level CNN architectures for music auto-tagging using raw waveforms**," in *Proceedings of the 43th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018.

[4] J. Pons and X. Serra, "**Musicnn: Pre-trained convolutional neural networks for music audio tagging**," in *Late Breaking and Demos of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

Results for Input Representations





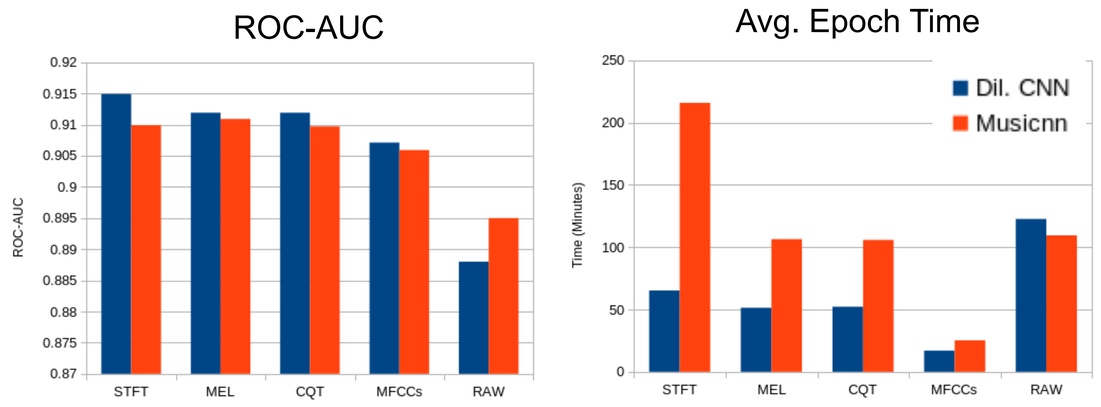
Two-way ANOVA

- Architectures and input representations **significantly influence performance**

Tukey-HSD:

- **STFT best** on both datasets
- Raw waveform worst
- Good average results for "lightweight" MFCC

Results for Dilated CNN



- **Significantly better** on 2D input representations
- Worse on raw waveform
- **Faster training**

Conclusions

- Input representation **does influence performance**
- **No** consistent **trends for individual tag** categories
- **Dilated CNN can improve** performance and training time

Source code: https://gitlab.com/MaxDamb/dl-autotagging
Additional materials: https://tinyurl.com/ismir22-317