Tutorial
Cross-Modal Music Retrieval and Applications

Part III: Machine Learning Approaches

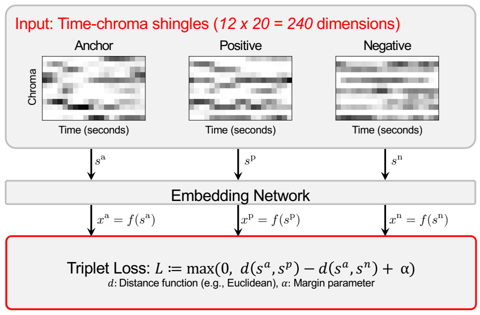
Meinard Müller
International Audio Laboratories Erlangen
meinard.mueller@audiolabs-erlangen.de

Andreas Arzt, Stefan Balke
Johannes Kepler University
andreas.arzt@jku.at, stefan.balke@jku.at

Shingle-Based Retrieval

[Schroff et al. CVPR 2015, FaceNet]
[Zalkow, Müller; submitted]

Strategy: Learn embedding using Siamese NN with triplet loss



Triplet Loss

$$L := \max(0, d(s^a, s^p) - d(s^a, s^n) + \alpha)$$

d: Distance function (e.g., Euclidean), α: Margin parameter

- **Goal:** Support that positive samples lie together ($L = 0$), penalize if not ($L > 0$).
- $L = 0$ is fulfilled if:

$$d(s^a, s^p) - d(s^a, s^n) + \alpha < 0$$

$$d(s^a, s^n) > d(s^a, s^p) + \alpha$$

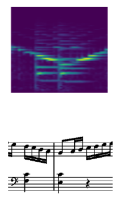
- **Intuition:** Push away negative samples w.r.t. distance in embedding space

---

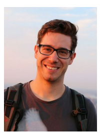
What if input modalities differ?

Image caption:
Mummified cat wrapped in undyed and red linen. Eyes and mouth painted in black.

Left ear missing.



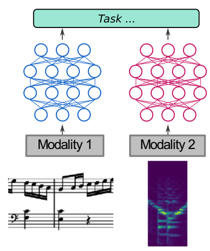
Credits
Many thanks to Matthias Dorfer for sharing his slides!



CROSS-MODAL AUDIO–SHEET MUSIC RETRIEVAL

Multi-Modal Deep Representation Learning

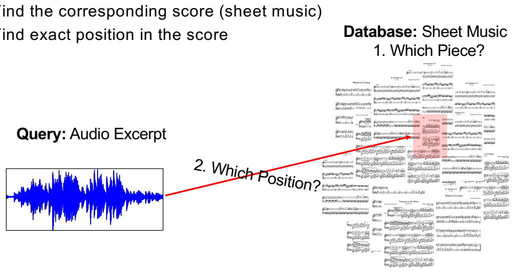
- Learning task specific representations from two or potentially more inputs.
- Images & Text
- **Images/Videos & Audio**
- …basically any modality pairs
- Very similar to Siamese Networks but weights are no longer shared!



Audio–Sheet Music Retrieval

Given: Short audio snippet as query
Goal: Retrieve relevant counterpart in sheet music collection

1. Find the corresponding score (sheet music)
2. Find exact position in the score



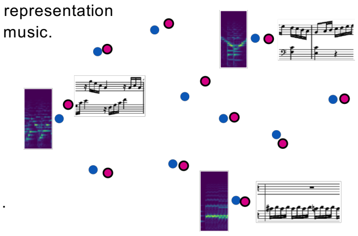
Task Description: ML Perspective

Type of Problem: Cross-Modality Retrieval

State-of-the-Art Approaches:
Learn a common vector representation of both audio and sheet music.

Retrieval:
Nearest neighbor search in the embedding space

Problem:
ML is usually data hungry…

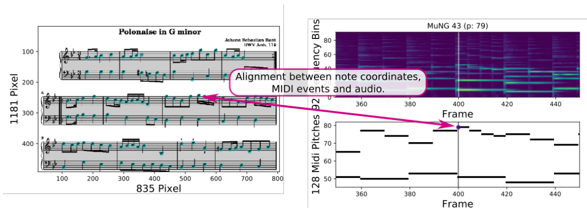


Multimodal Audio–Sheet Music Data

- **M**ultimodal Audio **S**heet **M**usic **D**ataset (MSMD)
- Obtained from the Mutopia project https://www.mutopiaproject.org
- Synthesized MIDI data of ca. 500 solo piano pieces
- Approximately 15 hours of music!

[Dorfer et al., TISMIR, 2018]
M. Dorfer, J. j. Hajič, A. Arzt, H. Frostel, and G. Widmer.
Learning audio–sheet music correspondences for cross-modal retrieval and piece identification.
Transactions of the International Society for Music Information Retrieval, 2018.

MSMD Annotations

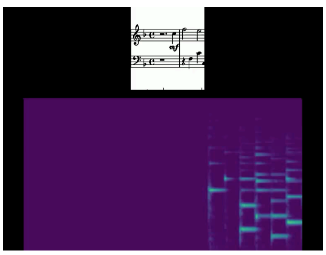


344,742 note head correspondences

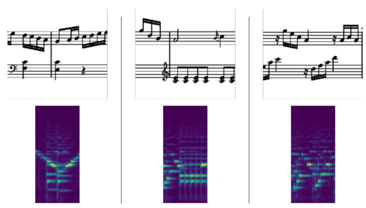
This is exactly the kind of data we need to explore the potential of powerful machine learning methods.

MSMD Example

J. S. Bach – Air (BWV Anh. 131)

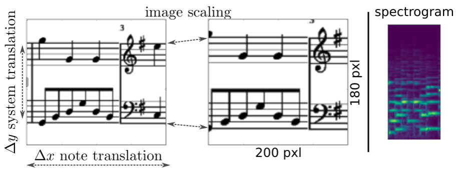


Training Data



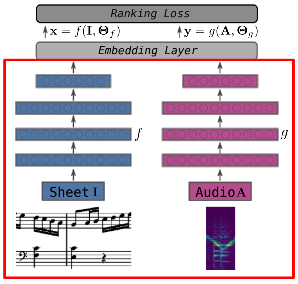
- Corresponding snippets of audio and sheet music
- Sheet Music (200 x 160 px), Audio Snippet (92 x 42, 2.1 s)

Data Augmentation



- **Image Augmentation**
  Rescaling and translation
- **Audio Augmentation**
  Synthesized with different tempi and MIDI-Soundfonts

Network Architecture



- Sheet image and and audio spec as input
- 32 dim. embeddings for both modalities
- Embedding Layer (CCA)
- Ranking Loss

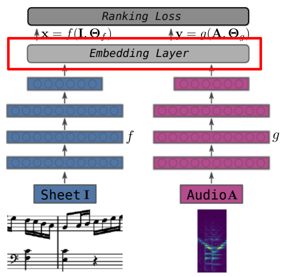
Network Architecture

2x Conv(3, pad-1)-24 – BN
MaxPooling(2)
2x Conv(3, pad-1)-48 – BN
MaxPooling(2)
2x Conv(3, pad-1)-96 – BN
MaxPooling(2)
2x Conv(3, pad-1)-96 – BN
MaxPooling(2)
Conv(1, pad-0)-32 - Linear – BN
Dense(32) + Linear

2x Conv(3, pad-1)-24 – BN
MaxPooling(2)
2x Conv(3, pad-1)-48 – BN
MaxPooling(2)
2x Conv(3, pad-1)-96 – BN
MaxPooling(2)
2x Conv(3, pad-1)-96 – BN
MaxPooling(2)
Conv(1, pad-0)-32 - Linear – BN
Dense(32) + Linear

- VGGish architecture:
  - Convolutional blocks followed by pooling
  - Number of filters increases with depth (e.g., 24, 28, 96)

Embedding Layer



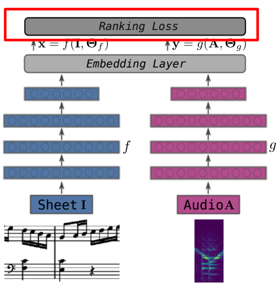
- **C**anonical **C**orrelation **A**nalysis

Recap CCA

- Find projection to maximize correlation of two vectors.
- Translated to special NN layer that allows to back-propagate a pairwise Ranking Loss through the analytical projections of CCA.

[Dorfer et al., JMIR, 2018]
M. Dorfer, J. Schlüter, A. Vall, F. Korzeniowski, G. Widmer.
End-to-End Cross-Modality Retrieval with CCA Projections and Pairwise Ranking Loss.
International Journal of Multimedia Information Retrieval, 2018.

Ranking Loss



- **Triplet Loss:**
  $\max(0, d(x^a, y^p) - d(x^a, y^n) + \alpha)$
- $d(\cdot)$: Cosine-Distance
- Margin $\alpha = 1$

Experiment: Snippet Retrieval

R@k: Recall at k
MRR: Mean Rec. Rank
MR: Median Rank

- **Setting:** 10,000 audio–sheet music pairs

Model	R@1	R@5	R@25	MRR	MR
CNN-2s	48.91	67.22	78.27	0.57	2

Problems:

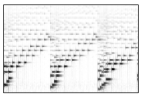
- Temporal context of audio input (2.1s) is not enough for all queries

Experiment: Snippet Retrieval

- **Setting:** 10,000 audio–sheet music pairs

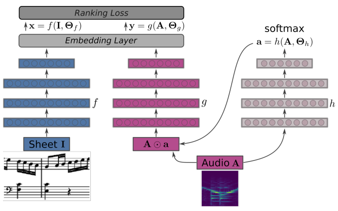
Model	R@1	R@5	R@25	MRR	MR
CNN-2s	48.91	67.22	78.27	0.57	2
CNN-4s	47.08	68.19	80.82	0.57	2
CNN-8s	43.46	68.38	82.84	0.55	2

- Simply extending the input context does not improve results!
- Introduces too much “confusion” through irrelevant information



TEMPO-INVARIANCE THROUGH INPUT ATTENTION

Embedding Model with Attention



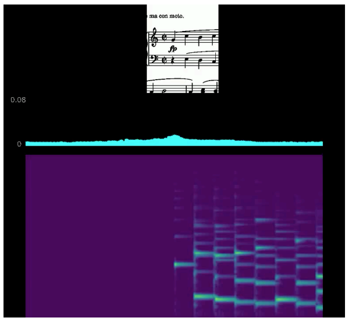
Soft Attention Mechanism
Additional audio network path with softmax output (attention mask).

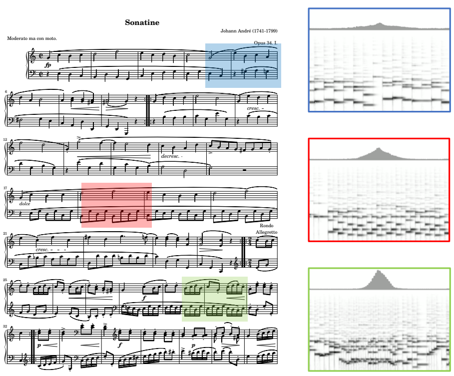
Weight input spectrogram with attention mask.

Intuition
Model encodes only relevant parts of the spectrogram.

[Dorfer et al., ICML Workshop 2018]
[Balke et al., submitted]

Johann André – Sonatine





Experiment: Snippet Retrieval

[Balke et al., submitted]

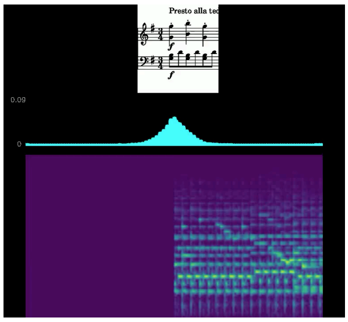
- **Setting:** 10,000 audio–sheet music pairs

Model	R@1	R@5	R@25	MRR	MR
CNN-2s	48.91	67.22	78.27	0.57	2
CNN-4s	47.08	68.19	80.82	0.57	2
CNN-8s	43.46	68.38	82.84	0.55	2
CNN-2s-AT	55.43	72.64	81.05	0.63	1
CNN-4s-AT	58.14	76.50	84.60	0.67	1
CNN-8s-AT	66.71	84.43	91.19	0.75	1

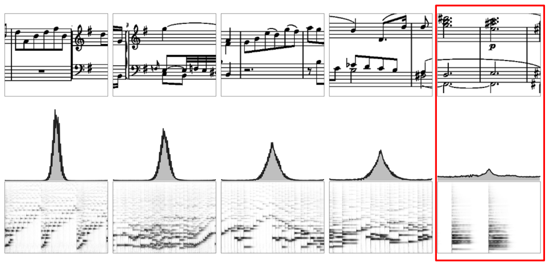
- **Attention enables us to use a longer temporal context!**

Attention Examples

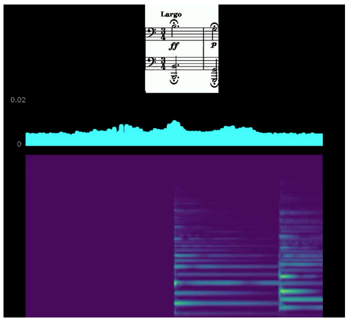
L. van Beethoven – Op. 79, 1st Mvt.



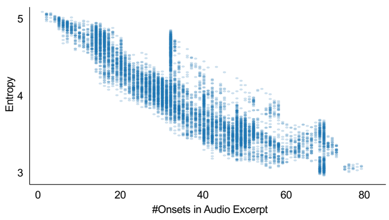
Attention Examples



M. Mussorgsky – Catacombae



Attention Summary



- **Very intuitive:**
  The more onsets in the audio snippet, the narrower (lower entropy) the attention output!
- Give it a try on your time series!

Summary Machine Learning Approaches

- End-to-end Embedding Space Learning for retrieval:
  - Siamese networks for single modality tasks
  - Separate embedding networks for cross-modality tasks
  - In each case, retrieval is simple nearest neighbor search
- Separate attention network enhances tempo robustness
- However:
  - Methods are very data hungry…
  - We still work with synthetic data…*)

*) Dorfer showed experiments with good performance on real performances (augmentation is imortant)

Overview of Challenges

- **Cross-modality**
  Symbolic vs. audio data
- **Tuning**
  Deviations from standard tuning
- **Transposition**
  Played key vs. written key
- **Tempo**
  Local & global tempo deviations
- **Polyphony**
  Monophonic query vs. polyphonic audio

Ultimate Retrieval Challenge