Short Introduction ...

I am a PhD Candidate in the Department of Computational Perception at Johannes Kepler University Linz (JKU).

My supervisor
Prof. Gerhard Widmer
Short Introduction ...

I am a PhD Candidate in the Department of Computational Perception at Johannes Kepler University Linz (JKU).

My supervisor
Prof. Gerhard Widmer

"Basic and applied research in machine learning, pattern recognition, knowledge extraction, and generally Artificial and Computational Intelligence. ... focus is on intelligent audio (specifically: music) processing."
This Talk Is About ...

Multi-Modal Neural Networks
This Talk Is About ...

Multi-Modal Neural Networks

Audio-Visual Representation Learning
This Talk Is About ...

Multi-Modal Neural Networks

Audio-Visual Representation Learning

Learning Correspondences between Audio and Sheet-Music
OUR TASKS
Our Tasks

Score Following (Localization)

Cross-Modality Retrieval

Task: predict sheet position

Multi-modal audio-to-sheet matching model

**Spectrogram**

**Sheet image**

**View 1**

**View 2**

**Embedding Layer**

**Ranking Loss**

\[ \hat{x} = f(x, \Theta_y) \quad \hat{y} = g(a, \Theta_x) \]
Task - Score Following

Score Following is the process of following a musical performance (audio) with respect to a known symbolical representation (e.g. a score).

You are here!
The Task: Audio to Sheet Matching
The Task: Audio to Sheet Matching
The Task: Audio to Sheet Matching

Spectrogram

Sheet image

Spectrogram Excerpt

Multi-modal audio-to-sheet matching model
The Task: Audio to Sheet Matching

**Task:** predict sheet position

Multi-modal audio-to-sheet matching model
The Task: Audio to Sheet Matching

Simultaneously **learn** (in end-to-end neural network fashion) to

- read notes from images (pixels)
- listen to music
- match played music to its corresponding notes

**Task:** predict sheet position

Multi-modal audio-to-sheet matching model
METHODS
Spectrogram to Sheet Correspondences

- Rightmost onset is target note onset
- Temporal context of 1.2 sec into the past
Multi-modal Convolution Network

The output layer is a $B$-way soft-max!
Multi-modal Convolution Network

The output layer is a $B$-way soft-max!
Multi-modal Convolution Network

The output layer is a $B$-way soft-max!
Multi-modal Convolution Network

The output layer is a $B$-way soft-max!
Multi-modal Convolution Network

The output layer is a $B$-way soft-max!
Soft Target Vectors

- Staff image is quantized into buckets
- Each bucket is represented by one output neuron
- Buckets hold probability of containing the note
- Neighbouring buckets share probability → soft targets
Soft Target Vectors

- Staff image is quantized into buckets
- Each bucket is represented by one output neuron
- Buckets hold probability of containing the note
- Neighbouring buckets share probability → soft targets

- Used as target values for training our networks
Optimization Objective

Output activation: $B$-way soft-max

$$\phi(y_{j,b}) = \frac{e^{y_{j,b}}}{\sum_{k=1}^{B} e^{y_{j,k}}}$$
Optimization Objective

Output activation: $B$-way soft-max

$$
\phi(y_{j,b}) = \frac{e^{y_{j,b}}}{\sum_{k=1}^{B} e^{y_{j,k}}}
$$

Soft targets $t_j$
Optimization Objective

Output activation: $B$-way soft-max

\[ \phi(y_{j,b}) = \frac{e^{y_{j,b}}}{\sum_{k=1}^{B} e^{y_{j,k}}} \]

Soft targets $t_j$

Loss: Categorical Cross Entropy

\[ l_j(\Theta) = -\sum_{k=1}^{B} t_{j,k} \log(p_{j,k}) \]
Discussion: Choice of Objective

- Allows to model uncertainties (e.g. repetitive structures in music)
- Our experience: Much nicer to optimize than MSE regression or Mixture Density Networks
Sheet Location Prediction

At test time: Predict expected location $\hat{x}_j$ of audio snippet with target note $j$ in sheet image.
Sheet Location Prediction

At test time: Predict expected location $\hat{x}_j$ of audio snippet with target note $j$ in sheet image.

Probability weighted localization

$$\hat{x}_j = \sum_{k \in \{b^*-1,b^*,b^*+1\}} w_k c_k$$

- bucket $b^*$ with highest probability $p_j$
- weights $w = \{p_j,b^*-1, p_j,b^*, p_j,b^*+1\}$,
- bucket coordinates $c_k$
EXPERIMENTS / DEMO
Train / Evaluation Data


- Trained on monophonic piano music
- Localization of staff lines
- Synthesize midi-tracks to audio
- Signal processing
  - Spectrogram (22.05 kHz, 2048 window, 31.25 fps)
  - Filterbank: 24 band logarithmic (80 Hz to 8 kHz)
# Model Architecture and Optimization

<table>
<thead>
<tr>
<th>Sheet-Image $40 \times 390$</th>
<th>Spectrogram $136 \times 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VGG style image model</strong></td>
<td><strong>VGG style audio model</strong></td>
</tr>
<tr>
<td>$3 \times 3$ Conv, BN, ReLU</td>
<td>$3 \times 3$ Conv, BN, ReLU</td>
</tr>
<tr>
<td>Max pooling</td>
<td>Max pooling</td>
</tr>
<tr>
<td>Dense, BN, ReLu, Drop-Out</td>
<td>Dense, BN, ReLu, Drop-Out</td>
</tr>
<tr>
<td><strong>Multi-modality merging</strong></td>
<td></td>
</tr>
<tr>
<td>Concatenation-Layer</td>
<td></td>
</tr>
<tr>
<td>Dense, BN, ReLu, Drop-Out</td>
<td></td>
</tr>
<tr>
<td>Dense, BN, ReLu, Drop-Out</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$B$-way Soft-Max Layer</td>
</tr>
</tbody>
</table>
Model Architecture and Optimization

<table>
<thead>
<tr>
<th>Sheet-Image 40 × 390</th>
<th>Spectrogram 136 × 40</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VGG style image model</strong></td>
<td><strong>VGG style audio model</strong></td>
</tr>
<tr>
<td>3 × 3 Conv, BN, ReLU</td>
<td>3 × 3 Conv, BN, ReLU</td>
</tr>
<tr>
<td>Max pooling</td>
<td>Max pooling</td>
</tr>
<tr>
<td>Dense, BN, ReLu, Drop-Out</td>
<td>Dense, BN, ReLu, Drop-Out</td>
</tr>
</tbody>
</table>

**Multi-modality merging**
- Concatenation-Layer
- Dense, BN, ReLu, Drop-Out
- Dense, BN, ReLu, Drop-Out

**B-way Soft-Max Layer**

- Mini-batch stochastic gradient descent with momentum
  - Mini-batch size: 100
  - Learning rate: 0.1 (divided by 10 every 10 epochs)
  - Momentum: 0.9
  - Weight decay: 0.0001
Demo with Real Music

Minuet in G Major (BWV Anhang 114, Johann Sebastian Bach)

- Played on Yamaha AvantGrand N2 hybrid piano
- Recorded using a single microphone
Demo with Real Music
So far so good ...

Model works well on monophonic music and seems to learn reasonable representations.

Important observation: No temporal model required!

What to do next?
Switch to "Real Music"

Minuet in G-major

Johann Sebastian Bach
Switch to "Real Music"
Composers, Sheet Music and Audio

- Pieces from **MuseScore** (annotating becomes feasible)

- **Classical Piano Music** by Mozart (14 pieces), Bach (16), Beethoven (5), Haydn (4) and Chopin (1)

- Experimental Setup:
  
  **train / validate**: Mozart | **test**: all composers

- Audio is synthesized
Fully Convolutional Segmentation Networks

Optical Music Recognition (OMR) Pipeline

1. Input Image
Fully Convolutional Segmentation Networks

Optical Music Recognition (OMR) Pipeline

1. Input Image
2. System Probability Maps
Fully Convolutional Segmentation Networks

Optical Music Recognition (OMR) Pipeline

1. Input Image
2. System Probability Maps
3. Systems Recognition
Fully Convolutional Segmentation Networks

Optical Music Recognition (OMR) Pipeline

1. Input Image
2. System Probability Maps
3. Systems Recognition
4. Regions of Interest
Fully Convolutional Segmentation Networks

Optical Music Recognition (OMR) Pipeline

1. Input Image
2. System Probability Maps
3. Systems Recognition
4. Regions of Interest
5. Note Probability Maps
Fully Convolutional Segmentation Networks

Optical Music Recognition (OMR) Pipeline

1. Input Image
2. System Probability Maps
3. Systems Recognition
4. Regions of Interest
5. Note Probability Maps
6. Note Head Recognition
Now we know

- the locations of staff systems and note heads and for each note head its onset time in the audio.
- overall 63836 annotated correspondences of 51 pieces.
Annotation Pipeline

Now we know

- the locations of staff systems and note heads and for each note head its onset time in the audio.
- overall **63836** annotated correspondences of **51 pieces**.
Train Data Preparation

We unroll the score and have the relations to the audio

This is all we need to train our models!
W.A. Mozart
Piano Sonata K545, 1st Movement

Plain, Frame-wise
Multi-Modal Convolution Network
Observations

- Sometimes a bit shaky
- **Score following fails** at the beginning of second page!

But why?
Failure
Failure
Failure
Failure

Sheet Image (539)

Spectrogram

Network Prediction

prediction

truth

25/39
Failure
Failure
Failure
Guided Back-Propagation


**Saliency Maps** for **understanding trained models**
Guided Back-Propagation


**Saliency Maps for understanding trained models**

Given a trained network $f$ and a fixed input $X$ we compute the gradient of network prediction $f(X) \in \mathbb{R}^k$ with respect to its input

$$\frac{\partial \max(f(X))}{\partial X}$$

Determines those parts of the input having the highest effect on the prediction when changed.
Guided Back-Propagation


**Saliency Maps for understanding trained models**

Given a **trained network** $f$ and a **fixed input** $X$ we compute the gradient of network prediction $f(X) \in \mathbb{R}^k$ with respect to its input

$$\frac{\partial \max(f(X))}{\partial X}$$

Determines those parts of the input having the highest effect on the prediction when changed.

Guided back-propagation with rectified linear units only back-propagates positive error signals $\delta_{l-1} = \delta_l \mathbf{1}_{x>0} \mathbf{1}_{\delta_l>0}$
Net Debugging

Sheet Image (Error: -9.01)

Spectrogram

Network Prediction
Net Debugging
Net Debugging
Net Debugging
Net Debugging

Sheet Image (Error: -59.13)

Network Prediction
Net Debugging
Failure Analysis Continued

- Network pays attention to note heads but does not seem to be pitch sensitive.

- However, exploiting **temporal relations** inherent in music could fix the problem!
RECURRENT NEURAL NETWORKS!
RNN Training Examples
RNN Training Examples

sheet image

classification target

spectrogram
RNN Training Examples

sheet image

classification target

spectrogram
RNN Training Examples

Sheet image

Classification target

Spectrogram

29/39
RNN Training Examples
RNN Learning Curves

![RNN Learning Curves Graph]

- more_conv_musescore_results_tr
- more_conv_musescore_results_va
- rnn_more_conv_musescore_results_tr
- rnn_more_conv_musescore_results_va

Epoch
Loss
0 20 40 60 80 100
1.0
1.5
2.0
2.5
3.0
3.5

0 20 40 60 80 100
1.0
1.5
2.0
2.5
3.0
3.5
HIDDEN MARKOV MODELS (HMMS)
Hidden Markov Models

Enforce spatial and temporal structure into single-time-step prediction score-following-model.
HMM - Design
HMM - Design

States
HMM - Design

States

Observations

0.75
0.25
HMM - Design

Map Local Predictions to Global Sheet Image and use them as Observations
HMM - Design

Apply HMM Filtering / Tracking Algorithm
W.A. Mozart
Piano Sonata K545, 1st Movement

HMM-Tracker
Multi-Modal Convolution Network
Conclusions

Learning multi-modal representations in the context of music-audio and sheet-music is a challenging application.

■ Learning Temporal Relations from training data
■ Real audio and real performances, (asynchronous onsets, pedal, and varying dynamics)
■ More training data!
■ ...
Conclusions

Learning multi-modal representations in the context of music-audio and sheet-music is a challenging application.

Multi-Modal Convolution Networks are the right direction.
Conclusions

Learning multi-modal representations in the context of music-audio and sheet-music is a challenging application.

Multi-Modal Convolution Networks are the right direction.

However there are many open problems left:

- **Learning Temporal Relations** from training data
- Real audio and real performances, (asynchronous onsets, pedal, and varying dynamics)
- **More training data!**
- ...
Data Augmentation

Image augmentation:

- Image scaling
- Note translation ($\Delta x$)
- System translation ($\Delta y$)
- Spectrogram
Data Augmentation

Image augmentation:

- \( \Delta y \) system translation
- \( \Delta x \) note translation
- Image scaling
- Spectrogram
- 200 pxl
- 180 pxl
Data Augmentation

Image augmentation:

- Image scaling
- 
- 200 pxl
- 180 pxl
- spectrogram
- 180 pxl
- 200 pxl
- 
- Audio augmentation
- Different tempi and sound fonts
Data Augmentation

Image augmentation:

- Image scaling
- $\Delta y$ system translation
- $\Delta x$ note translation
- Spectrogram
- Audio augmentation: different tempi and sound fonts

200 pxl
180 pxl
Data Augmentation

Image augmentation:

Audio augmentation
Different tempi and sound founts
AUDIO - SHEET MUSIC
CROSS-MODALITY
RETRIEVAL
The Task

*Our Goal:* Find a common vector representation of both audio and sheet music (low dimensional embedding)
The Task

*Our Goal:* Find a common vector representation of both audio and sheet music (low dimensional embedding)
The Task

*Our Goal:* Find a common vector representation of both audio and sheet music (low dimensional embedding)

*Why would we like this:* to make them comparable.
Cross-Modality Retrieval Neural Network

Optimizes the similarity (in embedding space) between corresponding audio and sheet image snippets.
Model Details and Optimization

- Uses CCA Embedding Layer
- Trained with Pairwise Ranking Loss
- 32-dimensional embedding
Model Details and Optimization

- Uses CCA Embedding Layer
- Trained with Pairwise Ranking Loss
- 32-dimensional embedding

Encourage an embedding space where the distance between matching samples is lower than the distance between mismatching samples.
Cross-Modality Retrieval

Audio query point of view:

- blue dots: embedded candidate sheet music snippets
- red dot: embedding of an audio query.
Cross-Modality Retrieval

Audio query point of view:
- blue dots: embedded candidate sheet music snippets
- red dot: embedding of an audio query.

→ Retrieval by nearest neighbor search