Music Transcription Using Recurrent Neural Networks

Sara Hahner

March 29, 2019

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

Inhaltsverzeichnis

Introduction

The Dataset: MusicNet

The Task: Music Transcription

Translation-Invariant Neural Network
- Training Results
- Variation

Recurrent Neural Network
- Motivation
- Arquitecture
- Training Results

Conclusion

References

Introduction

Given Dataset (MusicNet) of aligned recording and MIDI scores for 330 classical music pieces ($\approx$ 34h)

Introduction

Given Dataset (MusicNet) of aligned recording and MIDI scores for 330 classical music pieces ($\approx$ 34h)

Task Given a sequence of the recording identify the notes played in the middle of the sequence.

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Introduction

Given Dataset (MusicNet) of aligned recording and MIDI scores for 330 classical music pieces ($\approx$ 34h)

Task Given a sequence of the recording identify the notes played in the middle of the sequence.

Idea Use a Feed Forward Translation Invariant Neural Network ([THFK18])

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Introduction

Given Dataset (MusicNet) of aligned recording and MIDI scores for 330 classical music pieces ($\approx$ 34h)

Task Given a sequence of the recording identify the notes played in the middle of the sequence.

Idea Use a Feed Forward Translation Invariant Neural Network ([THFK18])

$\Rightarrow$ Try a Recurrent Neural Network

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

MusicNet

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

MusicNet

- Dataset introduced in ([THK17])
- 330 music recordings of solo and chamber music pieces
- Sample rate of 44,100Hz

MusicNet

- Dataset introduced in ([THK17])
- 330 music recordings of solo and chamber music pieces
- Sample rate of 44,100Hz
- Labels: digital MIDI scores which contain notes and the instrument by which it is played
- Alignment using Dynamic Time Warping minimizing a cost function in the performance space

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

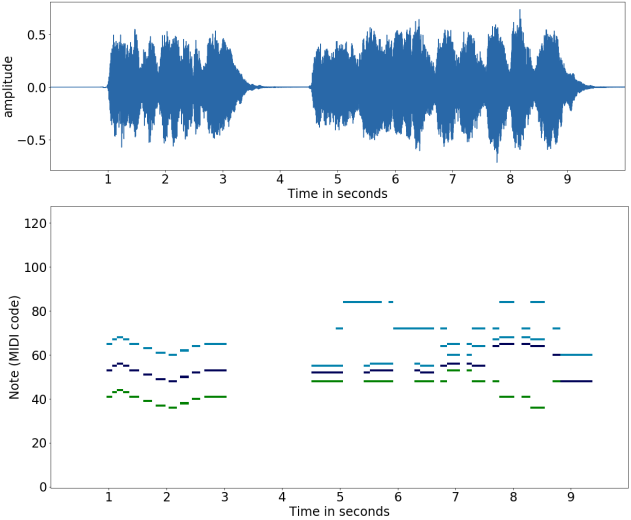


Figure: First 10 seconds of String Quartet No 11 in F minor from Beethoven, first movement Allegro con brio (recording-ID 2494)

MusicNet: Data Augmentation

Application of data augmentation to every minibatch [THFK18]

1. **Pitch-shift in the frequency domain**
   generally between -5 and $+5$ semitones
   $s_{pitchshift} \in \mathbb{Z} \cap [-5, +5]$

MusicNet: Data Augmentation

Application of data augmentation to every minibatch [THFK18]

1. **Pitch-shift in the frequency domain**
   generally between -5 and $+5$ semitones
   $s_{pitchshift} \in \mathbb{Z} \cap [-5, +5]$
2. **Jittering**
   continuous shift to each data point (generally between -0.1 and $+0.1$ semitone) acting as Tuning Variations
   $s_{jittering} \in [-.1, .1]$

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

MusicNet: Data Augmentation

Application of data augmentation to every minibatch [THFK18]

1. **Pitch-shift in the frequency domain**
   generally between -5 and $+5$ semitones
   $s_{pitchshift} \in \mathbb{Z} \cap [-5, +5]$
2. **Jittering**
   continuous shift to each data point (generally between -0.1 and $+0.1$ semitone) acting as Tuning Variations
   $s_{jittering} \in [-.1, .1]$

Having scaling factor $s_{scaling} = s_{pitchshift} + s_{jittering}$ apply the multiplication in the frequency domain by $2^{s_{scaling}/12}$

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

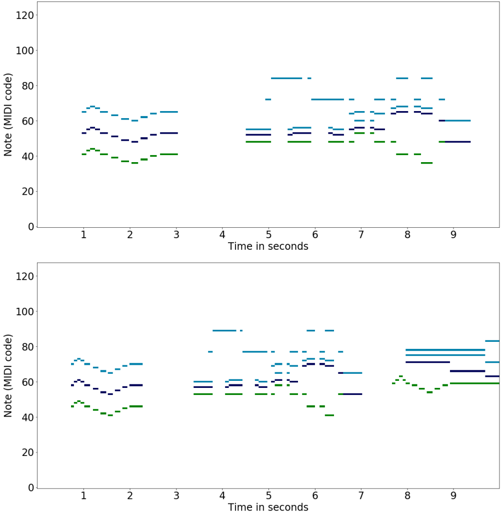


Figure: First 10 seconds of recording No. 2494: before and after applying a 5 semitone pitch shift

MusicNet: Normalization

- Amplitude: volume of the music
- Preprocessing of each window: $x \mapsto x/||x||_2$
  "Normalizing the audible volume of each frame" ([THFK18])

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

MusicNet: Normalization

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

- Amplitude: volume of the music
- Preprocessing of each window: $x \mapsto x/||x||_2$
  "Normalizing the audible volume of each frame" ([THFK18])

The mean of the training data set has been calculated and is approximately zero (-3.5×10^{-7}).

Music Transcription: Multi-Label Classification Problem

Music Transcription: Multi-Label Classification Problem

To every segment $x \in \mathcal{X}$ of an audio is assigned a binary label vector $y \in \mathcal{H} = \{0, 1\}^{128}$. Every dimension corresponds to a frequency value of a note and $y_n = 1$ if and only if the note n is present at the midpoint of x. [THK17]

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

Music Transcription: Multi-Label Classification Problem

To every segment $x \in \mathcal{X}$ of an audio is assigned a binary label vector $y \in \mathcal{H} = \{0,1\}^{128}$. Every dimension corresponds to a frequency value of a note and $y_n = 1$ if and only if the note n is present at the midpoint of x. [THK17]

$\Rightarrow$ Learn the feature map $f : \mathcal{X} \to \mathcal{H}$ by minimizing square loss

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

Translation-Invariant Neural Network

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

Translation-Invariant Neural Network [THFK18]

1. Receive a window of size 16,384 (≈ 0.37 seconds)

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

Translation-Invariant Neural Network [THFK18]

1. Receive a window of size 16,384 (≈ 0.37 seconds)
   1.1 Strided convolution over the time dimension with a 4,096-sample receptive field and a 512 sample stride
   1.2 For each region compute a filterbank representation by applying 512 sine and cosine filters with logarithmically spaced frequencies

Translation-Invariant Neural Network [THFK18]

1. Receive a window of size 16,384 (≈ 0.37 seconds)
    1.1 Strided convolution over the time dimension with a 4,096-sample receptive field and a 512 sample stride
    1.2 For each region compute a filterbank representation by applying 512 sine and cosine filters with logarithmically spaced frequencies
2. Convolution along the log-frequency axis

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Translation-Invariant Neural Network [THFK18]

1. Receive a window of size 16,384 (≈ 0.37 seconds)
   1.1 Strided convolution over the time dimension with a 4,096-sample receptive field and a 512 sample stride
   1.2 For each region compute a filterbank representation by applying 512 sine and cosine filters with logarithmically spaced frequencies
2. Convolution along the log-frequency axis
   $\Rightarrow$ Learned translation-invariant filters

Translation-Invariant Neural Network [THFK18]

1. Receive a window of size 16,384 (≈ 0.37 seconds)
   1.1 Strided convolution over the time dimension with a 4,096-sample receptive field and a 512 sample stride
   1.2 For each region compute a filterbank representation by applying 512 sine and cosine filters with logarithmically spaced frequencies
2. Convolution along the log-frequency axis
   $\Rightarrow$ Learned translation-invariant filters
3. Convolution using filters of height 1 along the log-frequency axis

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Translation-Invariant Neural Network [THFK18]

1. Receive a window of size 16,384 (≈ 0.37 seconds)
   1.1 Strided convolution over the time dimension with a 4,096-sample receptive field and a 512 sample stride
   1.2 For each region compute a filterbank representation by applying 512 sine and cosine filters with logarithmically spaced frequencies
2. Convolution along the log-frequency axis
   $\Rightarrow$ Learned translation-invariant filters
3. Convolution using filters of height 1 along the log-frequency axis

Prediction at output layer by linear classification

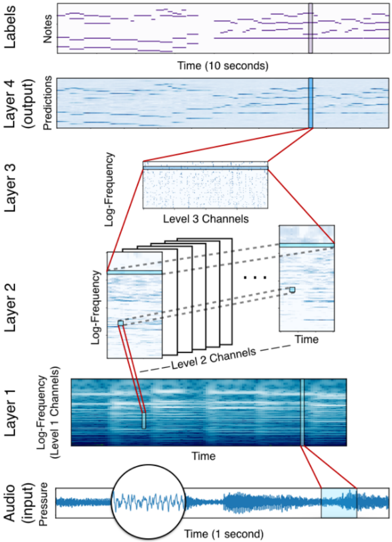


Figure: The translation-invariant network for note classification. Figure from [THFK18]

Music Transcription Using RNN
Sara Hahner
Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

The Training

- Momentum Optimizer (Momentum = 0.95)
- 300,000 iterations using batches of size 150
- Initial learning rate .00001 (apply learning rate decay)
- Moving Average of the weights for evaluation

The Training

- Momentum Optimizer (Momentum $= 0.95$)
- 300,000 iterations using batches of size 150
- Initial learning rate .00001 (apply learning rate decay)
- Moving Average of the weights for evaluation

Training Results

Analysis:

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

Training Results

Analysis:

- Average Precision: scikit-learn version 0.19.1
- Accuracy and Error: mireval [RMH+14]
  global prediction threshold of 0.4

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Training Results

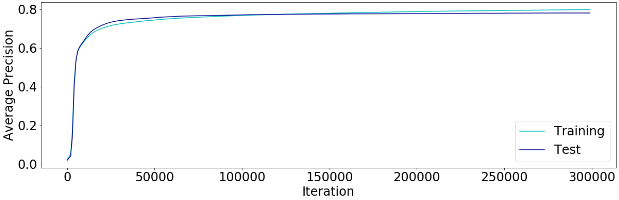


Figure: Development of the average precision during training of replica of the translation-invariant neural network.

Training Results

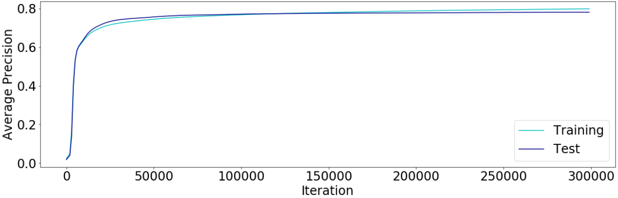


Figure: Development of the average precision during training of replica of the translation-invariant neural network.

Algorithm	Avg. Prec.	Acc.	Err.	Reference
Melodyne	57.9%	.395	.744	[Cel], [THFK18]

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Training Results

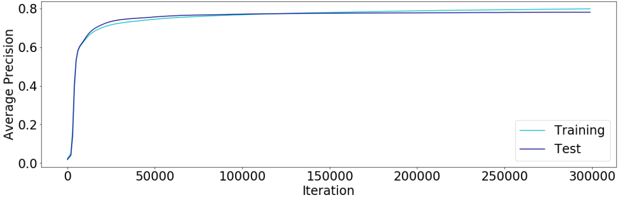


Figure: Development of the average precision during training of replica of the translation-invariant neural network.

Algorithm	Avg. Prec.	Acc.	Err.	Reference
Melodyne	57.9%	.395	.744	[Cel], [THFK18]
Transl.-inv.	79.9%	.599	.423	[THFK18]

Training Results

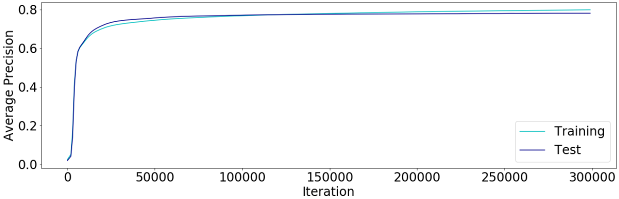


Figure: Development of the average precision during training of replica of the translation-invariant neural network.

Algorithm	Avg. Prec.	Acc.	Err.	Reference
Melodyne	57.9%	.395	.744	[Cel], [THFK18]
Transl.-inv.	79.9%	.599	.423	[THFK18]
Transl.-inv.	78.1%	.583	.427	Replica

Table: Test results from the authors of [THFK18] and my replica.

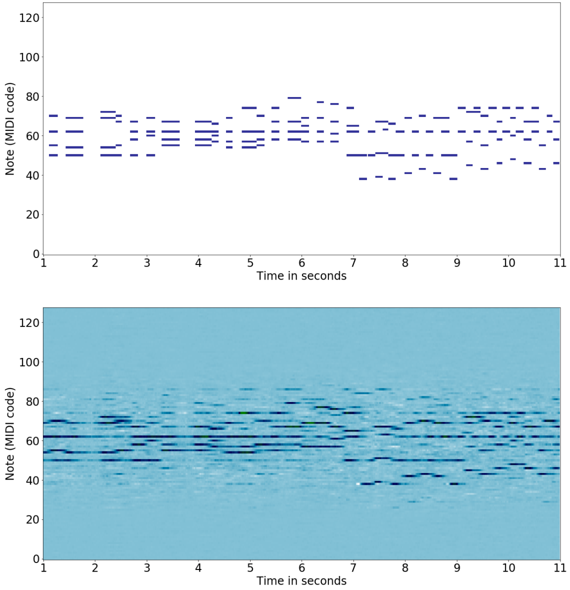


Figure: Direct Output and original score for recording No. 2628 (Violin Sonata No. 10 in G major from Beethoven, 3rd movement (Scherzo))

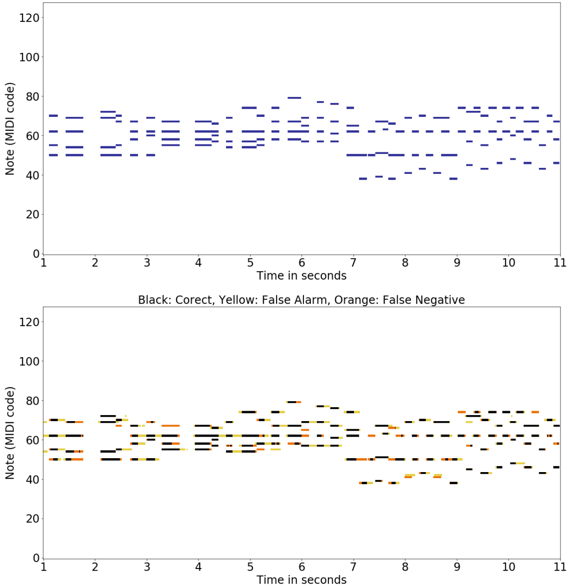


Figure: Comparison of original/predicted score for recording No. 2628

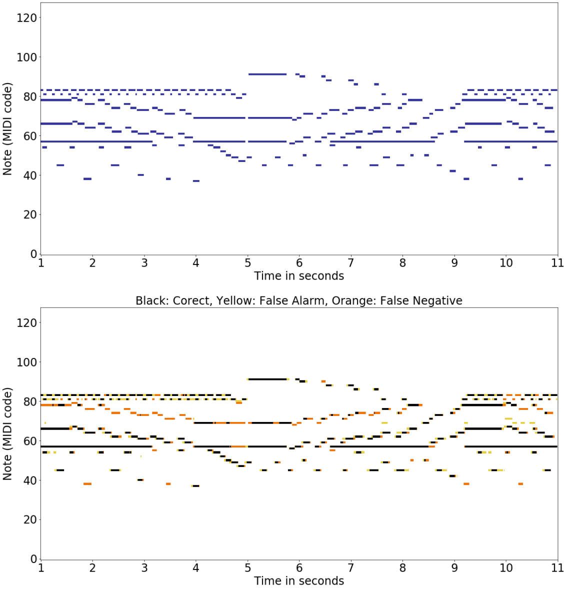


Figure: Direct Output and comparison of original/predicted score for recording No. 2718

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Observations

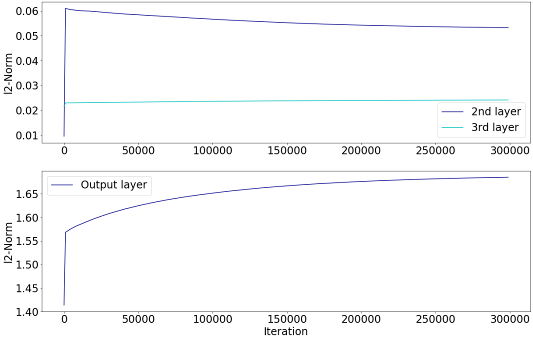


Figure: Norm of the weights during the training of second, third and output layer.

Observations

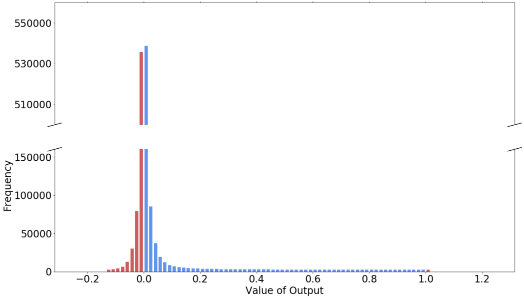


Figure: Histogram showing the distribution of the output values when applying my replica of the translation-invariant neural network to test set.

Variations

- Additional Regularization: L^2 parameter norm penalty
  Add the term
  $$\frac{1}{2}\sum_i \|w^{(i)}\|_2^2 ,$$
  with $w^{(i)}$ being the trainable weights from layer i to the training loss

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Variations

- Additional Regularization: L^2 parameter norm penalty
  Add the term
  $$\frac{1}{2}\sum_i \|w^{(i)}\|_2^2 \ ,$$
  with $w^{(i)}$ being the trainable weights from layer i to the training loss
- Sigmoid function in output layer
  $\Rightarrow$ Predicted scores lie in interval $[0, 1]$

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Variations

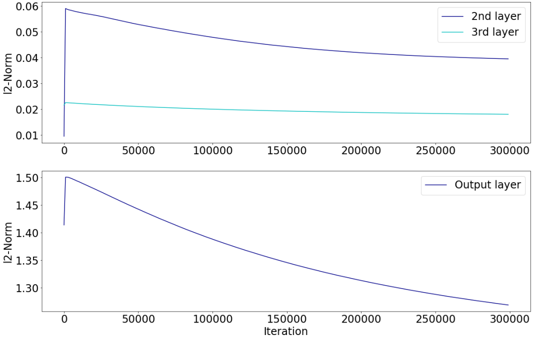


Figure: Norm of the weights of second, third and output layer when applying L^2 parameter norm penalty.

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

Results

Algorithm	Avg. Prec.	Acc.	Err.	Reference
Melodyne	57.9%	.395	.744	[Cel], [THFK18]
Transl.-inv.	79.9%	.599	.423	[THFK18]
Transl.-inv.	78.1%	.583	.427	Replica

Results

Algorithm	Avg. Prec.	Acc.	Err.	Reference
Melodyne	57.9%	.395	.744	[Cel], [THFK18]
Transl.-inv.	79.9%	.599	.423	[THFK18]
Transl.-inv.	78.1%	.583	.427	Replica
Transl.-inv.	78.5%	.589	.424	Regularization

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Results

Algorithm	Avg. Prec.	Acc.	Err.	Reference
Melodyne	57.9%	.395	.744	[Cel], [THFK18]
Transl.-inv.	79.9%	.599	.423	[THFK18]
Transl.-inv.	78.1%	.583	.427	Replica
Transl.-inv.	78.5%	.589	.424	Regularization
Transl.-inv.	76.9%	.566	.452	Sigmoid
Transl.-inv.	71.1%	.512	.512	Sigmoid-Reg.

Table: Test results from the authors of [THFK18] and my replicas. For the average Precision scikit-learn version 0.19.1 was used. Accuracy and Error are computed by mireval [RMH$^+$14] assuming a global prediction threshold of 0.4.

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Recurrent Neural Network

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

Motivation

- Harmony of notes sounding at time t:
  perceived by translation-invariant convolution along the log-frequency axis in layer two
- Harmony of melodies over time:
  - pitch differences between $t-1$ and t
  - typical chord sequences in melodic passages

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Pitch differences

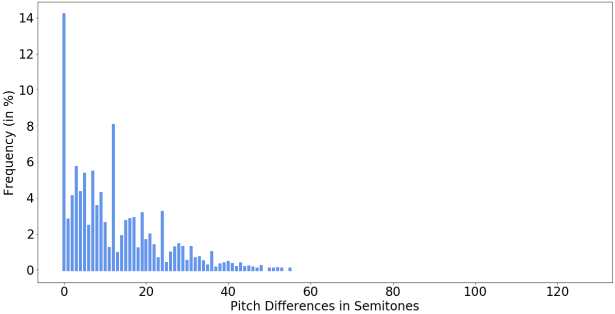


Figure: Relative Frequency of pitch differences after 0.37 seconds in MusicNet.

Pitch differences

Semitones	Chord	Relative Frequency of Interval
0	Perfect unison	14.16%
12	Perfect octave	8.0%
3	Minor third	5.68%
7	Perfect fifth	5.41%
5	Perfect fourth	5.32%
4	Major third	4.28%
9	Major sixth	4.23%
2	Major second	4.06%
8	Minor sixth	3.49%
24	Double octave	3.18%
19		3.11%
17		2.84%
16		2.79%
1	Minor second	2.77%
15		2.67%
10	Minor seventh	2.56%
6	Tritone	2.41%
21		1.93%
14		1.83%
20		1.62%
28		1.39%
22		1.31%
29		1.23%
31		1.23%
27		1.22%
11	Major seventh	1.19%
18		1.16%

Table: Harmonic pitch differences are more probable [Hin40]. Pitch differences after 0.37 seconds in MusicNet.

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Bidirectional Recurrent Neural Networks

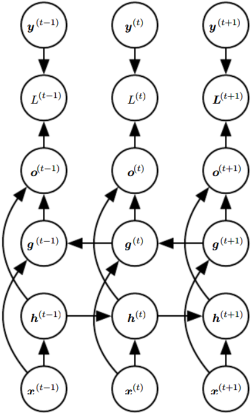


Figure: Arquitecture of the utilized bidirectional recurrent neural network: Given input $x^{(t)}$, the value $o^{(t)}$ of the output unit is calculated considering the information from timestep $t-1$, namely $h^{(t-1)}$, and the information from the next timestep $g^{(t-1)}$. The loss $L^{(t)}$ is calculated comparing $o^{(t)}$ to the target $y^{(t)}$. Figure from [GBC16]

The Realization

- Use tensorflow LSTM-cell (tf.nn.rnn_cell.LSTMCell) enabling peephole connections variating the number of units m

The Realization

- Use tensorflow LSTM-cell (tf.nn.rnn_cell.LSTMCell) enabling peephole connections variating the number of units m
- Dynamic version of bidirectional recurrent neural network (tf.nn.bidirectional_dynamic_rnn) considering s_{time} timesteps

The Realization

- Use tensorflow LSTM-cell (tf.nn.rnn_cell.LSTMCell) enabling peephole connections variating the number of units m
- Dynamic version of bidirectional recurrent neural network (tf.nn.bidirectional_dynamic_rnn) considering s_{time} timesteps
- Maintain a window size of approximately 0.37 seconds $\Rightarrow$ downsample from 44,100 Hz to 11,025 Hz

The Realization

- Use tensorflow LSTM-cell (tf.nn.rnn_cell.LSTMCell) enabling peephole connections variating the number of units m
- Dynamic version of bidirectional recurrent neural network (tf.nn.bidirectional_dynamic_rnn) considering s_{time} timesteps
- Maintain a window size of approximately 0.37 seconds $\Rightarrow$ downsample from 44,100 Hz to 11,025 Hz
- Normalize the norm of every window

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Reduction of the Feed-Forward Network

		Parameter	Sample Rate 44.100 Hz	Sample Rate 11.025 Hz
Input	Window size		16384 ~ 0.37 seconds	4096 ~ 0.37 seconds
	Sample rate		44100 Hz	11025 Hz
First Layer	Receptive field		4096	1024
	Stride		512 → 25 Regions	256 → 13 Regions
	Filter		512 sine & cosine filterbank on frequencies	256 sine & cosine filterbank on frequencies
	Layer-Output		1 x 25 x 512	1 x 13 x 256
Second Layer	Time	Receptive field	512	256
		Stride	1 → 25 Regions	1 → 13 Regions
	Freq	Receptive field	128	128
		Stride	2 → 193 Regions	2 → 65 Regions
		Filter	trainable filter of size (1, 128, 1, 128)	trainable filter of size (1, 128, 1, 128)
	Layer-Output		128 x 25 x 193	128 x 13 x 65
Third Layer	Time	Receptive field	25	13
		Stride	1 → 1 region	1 → 1 Region
	Freq	Receptive field	1	1
		Stride	1 → 193 Regions	1 → 65 Regions
		Filter	trainable filter of size (25, 1, 128, 4096)	trainable filter of size (13, 1, 128, 1024)
	Layer-Output		4096 x 1 x 193	1024 x 1 x 65
Output Layer	Feed Forward		Reshape into shape (790528) weights of size: (790528, 128)	Reshape into shape (66560) weights of size: (66560, 128)
Output			128	128

Figure: Illustration of the chosen size reductions in all layers of the neural network when applying downsampling from 44,100 Hz to 11,025 Hz.

The Training

- Use Pre-Trained weights
- Use Momentum and Adam Optimizer

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

The Training

- Use Pre-Trained weights
- Use Momentum and Adam Optimizer
- Start learning rate at .001, apply learning rate decay
- 100,000 iterations

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

The Training

- Use Pre-Trained weights
- Use Momentum and Adam Optimizer
- Start learning rate at .001, apply learning rate decay
- 100,000 iterations
- Variate number of units m of the LSTM-cell and number of timesteps s_{time} to include

Introduction

MusicNet

Music Transcription

Translation-Invariant Neural Network

Training Results

Variation

Recurrent Neural Network

Motivation

Arquitecture

Training Results

Conclusion

References

Training Results

s_{time}	Units m	Opt.	Avg. Prec. Train	Test	Acc.	Err.	Runtime for 1000 iter.
1	-	MomOpt	79.8%	76.1%	.559	.463	75 sec

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Training Results

s_{time}	Units m	Opt.	Avg. Prec. Train	Avg. Prec. Test	Acc.	Err.	Runtime for 1000 iter.
1	-	MomOpt	79.8%	76.1%	.559	.463	75 sec
3	128	MomOpt	58.9%	57.6%	.384	.655	135 sec
3	128	AdamOpt	58.3%	58.0%	.390	.640	138 sec
3	256	MomOpt	59.2%	57.7%	.393	.636	140 sec

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

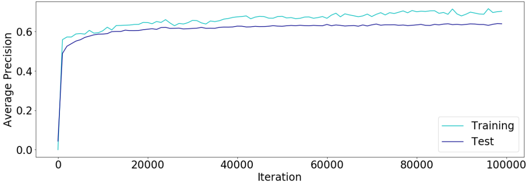
Training Results

s_{time}	Units m	Opt.	Avg. Prec. Train	Test	Acc.	Err.	Runtime for 1000 iter.
1	-	MomOpt	79.8%	76.1%	.559	.463	75 sec
3	128	MomOpt	58.9%	57.6%	.384	.655	135 sec
3	128	AdamOpt	58.3%	58.0%	.390	.640	138 sec
3	256	MomOpt	59.2%	57.7%	.393	.636	140 sec
9	1024	MomOpt	67.8%	60.6%	.401	.626	235 sec
9	1024	AdamOpt	70.3%	64.0%	.433	.589	360 sec
9	1024	AdamOpt*	70.2%	63.7%	.436	.592	350 sec
15	2048	AdamOpt*	78.7%	63.9%	.433	.589	700 sec

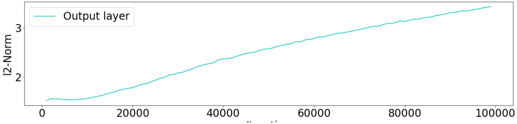
Table: Results for different variable choices for the recurrent neural network. (* indicates the implementation of L^2 parameter norm penalty with $\beta_{reg} = 0.01$)

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Training Results



(a) Average Precision



(b) Norm of the weights of the output layer

Figure: Test statistics from the training of the three-layer translation-invariant network in the bidirectional recurrent version. $s_{time} = 9$ and $m = 1024$, trained without L^2 parameter norm penalty.

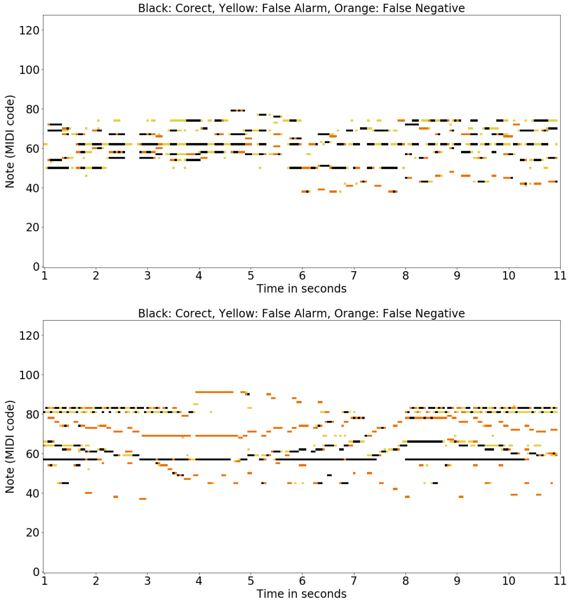


Figure: Comparison of original/predicted score for recording No. 2628 and 2718

Conclusions

- Successful implemented Translation Invariant Neural Network allows a good recognition of the music notes' pitches
- Struggle with the recognition of notes' beginning and ending

Conclusions

- Successful implemented Translation Invariant Neural Network allows a good recognition of the music notes' pitches
- Struggle with the recognition of notes' beginning and ending
- Using downsampled dataset the results are very similar and the training faster

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

Conclusions

- Successful implemented Translation Invariant Neural Network allows a good recognition of the music notes' pitches
- Struggle with the recognition of notes' beginning and ending
- Using downsampled dataset the results are very similar and the training faster
- RNN was not able to recognize the rhythm of melody despite its high model capacity

Celemony.
Melodyne.
http://www.celemony.com/en/melodyne/what-is-melodyne.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
MIT Press, 2016.
http://www.deeplearningbook.org.

P. Hindemith.
Unterweisung im Tonsatz.
Number Bd. 1 in Edition Schott. B. Schott's Söhne, 1940.

Colin Raffel, Brian Mcfee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel P. W. Ellis, C Colin Raffel, Brian Mcfee, and Eric J. Humphrey.
mir_eval: a transparent implementation of common mir metrics.

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References

In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, 2014.

John Thickstun, Zaid Harchaoui, Dean P. Foster, and Sham M. Kakade.
Invariances and data augmentation for supervised music transcription.
In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

John Thickstun, Zaid Harchaoui, and Sham M. Kakade.
Learning features of music from scratch.
In *International Conference on Learning Representations (ICLR)*, 2017.

Introduction
MusicNet
Music Transcription
Translation-Invariant Neural Network
Training Results
Variation
Recurrent Neural Network
Motivation
Arquitecture
Training Results
Conclusion
References