New York University

Automatic Piano Music Transcription

GitHub: https://github.com/shriyakalakata/automatic-piano-music-transcription

Argy Sakti, Sean Wiryadi, Shriya Kalakata

May 07, 2024

Executive Summary

Problem Statement: Automatic music transcription, which involves converting audio signals into musical notation, is a challenging task due to the complexities involved in representing and analyzing audio data.

Goal: develop a machine learning model capable of predicting the activations of MIDI notes from an audio file containing a musical piece.

Technical Challenges

- Representing audio data in a suitable format
- Capturing long-range dependencies and musical intricacies
- Large number of classes (88 MIDI notes)
- Imbalanced data

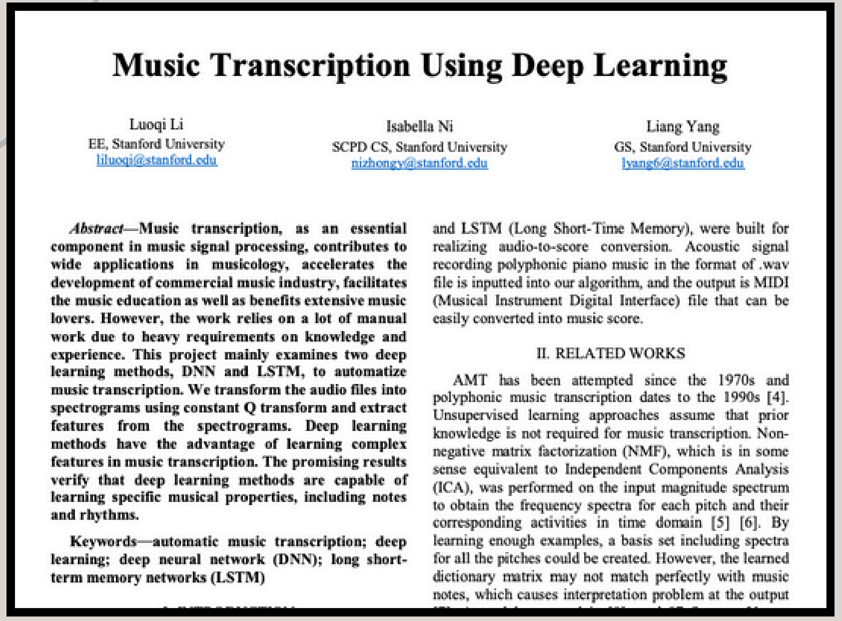
Solution Approach

- Deep neural network (DNN) with dropout and early stopping
- Long Short-Term Memory (LSTM) networks
- Experimented with transfer learning

Value/Benefit

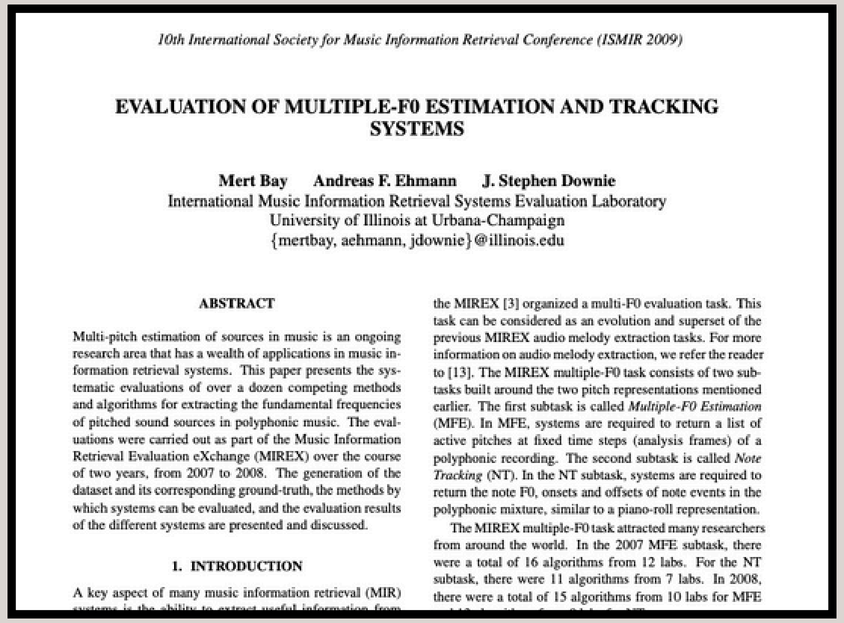
- Facilitate computational musicology and music information retrieval
- Enable applications like score generation, music education, and audio-to-score conversion

Related Works



Music Transcription Using Deep Learning

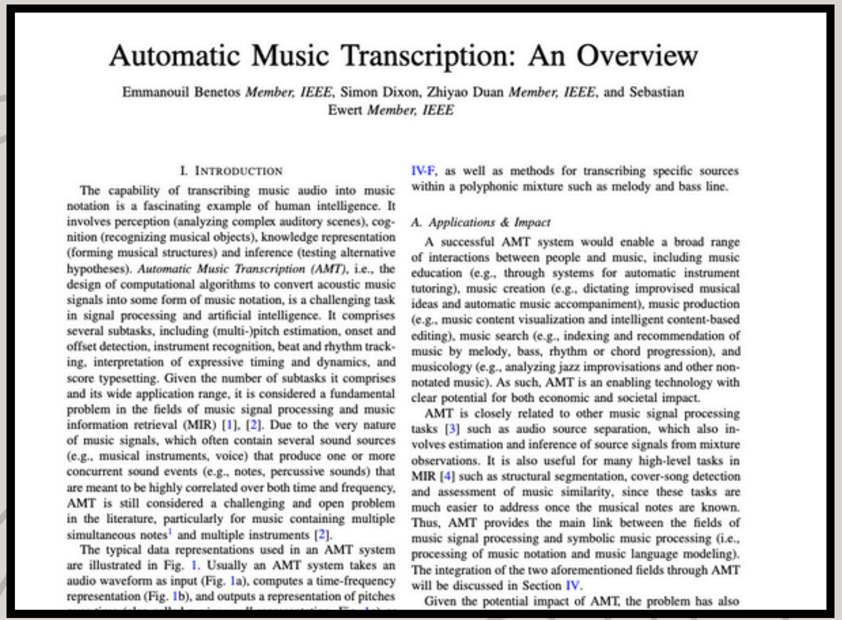
Luoqi Li, EE, Stanford University, liluoqi@stanford.edu
Isabella Ni, SCPD CS, Stanford University, nizhongy@stanford.edu
Liang Yang, GS, Stanford University, lyang6@stanford.edu



10th International Society for Music Information Retrieval Conference (ISMIR 2009)

EVALUATION OF MULTIPLE-F0 ESTIMATION AND TRACKING SYSTEMS

Mert Bay, Andreas F. Ehmann, J. Stephen Downie
International Music Information Retrieval Systems Evaluation Laboratory
University of Illinois at Urbana-Champaign
{mertbay, aehmann, jdownie}@illinois.edu



Automatic Music Transcription: An Overview

Emmanouil Benetos *Member, IEEE*, Simon Dixon, Zhiyao Duan *Member, IEEE*, and Sebastian Ewert *Member, IEEE*

MT3: Multi-Task Multitrack Music Transcription

Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, Jesse Engel
Google Research, Brain Team

What have others done?

Representing audio data with numerical representation

- **Non-negative matrix factorization (NMF)**, decomposing into basis components representing frequency spectra for each pitch and their activations in time
- **Constant Q-Transform (CQT)**, to represent audio signal by decomposing the signal into frequency components over time, with a time domain that is segmented through a frame-based approach

Model Architectures

- Recurrent Neural Networks
- Deep Neural Networks
- Long Short-Term Memory

Model Evaluation

- Frame-based Evaluation
- Note-based Evaluation

Limitations

- With NMF, the learned dictionary matrix may not match perfectly with music notes, which causes interpretation problems at the output
- Scarce data on annotations of ground-truth music transcriptions
- Overlapping sound events incites challenges with harmonics overlap in frequency

Method/Approach

Data Preprocessing & Normalization

- Represented audio using **Constant-Q Transform (CQT)**
- **Normalized** CQT vectors within the limits of the training set
- **Divided the audio into frames** of 32 ms (frame-based approach)
- **Aligned** CQT vectors with MIDI annotations using **one-hot encoding** representing note activations

Regularization/Optimization Techniques

- **Dropout** - reduce overfitting
- **Early stopping**
- **Cyclical learning rates** (Max LR 0.1, Min LR 0.0001)
- **Mini Batch Size** (Tradeoff between Accuracy and Efficiency)

Model Architectures

- Baseline Logistic Regression
- Deep Neural Network (DNN)
- Long Short-Term Memory (LSTM)

- Used **binary cross-entropy loss** for multi-label classification
- Employed **Adam optimizer** for training

Implementation/Experimentation

Model Architectures

- Deep Neural Network (DNN)
    - Explored 1, 2, 3, 4 hidden layers with ReLU activations
- Long Short-Term Memory (LSTM)
    - Attempted transfer learning by initializing weights from pre-trained DNNs

Hyperparameter Tuning

- Grid search for optimal hyperparameters
- Tuned dropout rates: 0.05, 0.15, 0.25
- Tuned minibatch sizes: 250, 500, 1000, 1500
- Best configuration: dropout = 0.05, minibatch = 1500
- (for LSTM): window size

Results – Baseline Model

Baseline Logistic Regression

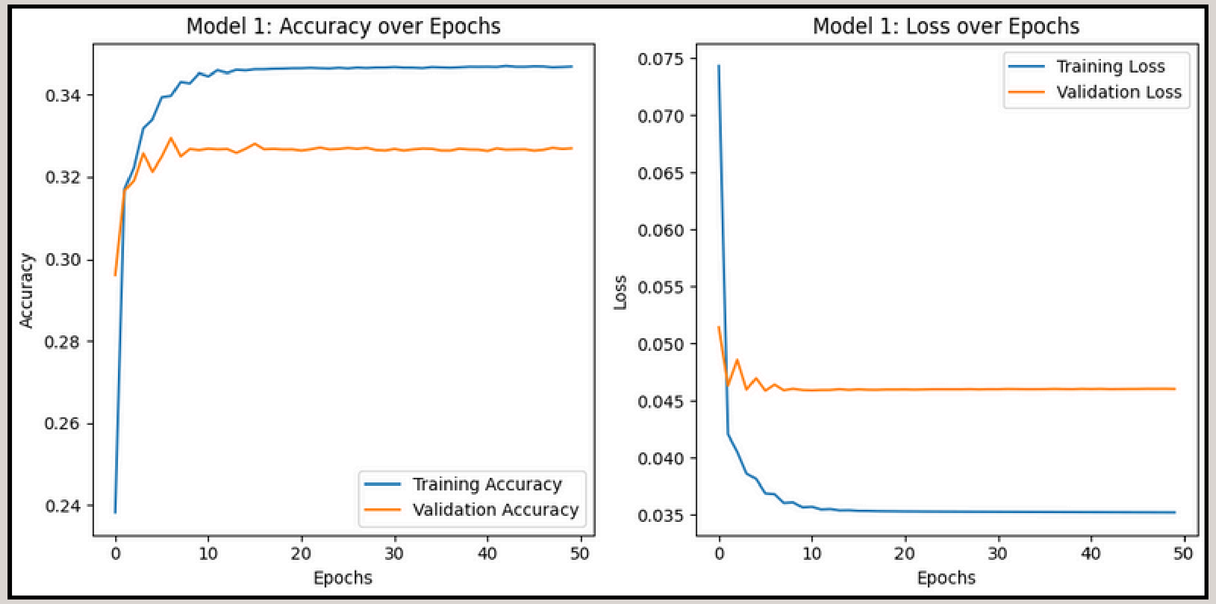
- Implemented a baseline model to serve as a comparison to our neural networks model
- **Accuracy** = **10.75%**
- Extremely **high recall of 85.17%**, at the cost of a **low precision of 10.96%**, indicates that the model overpredicts on note activations

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Log Regression	<u>10.75</u>	10.96	85.17	19.41

$$Accuracy\,(Modified) = \frac{TP}{(TP + FP + FN)}$$

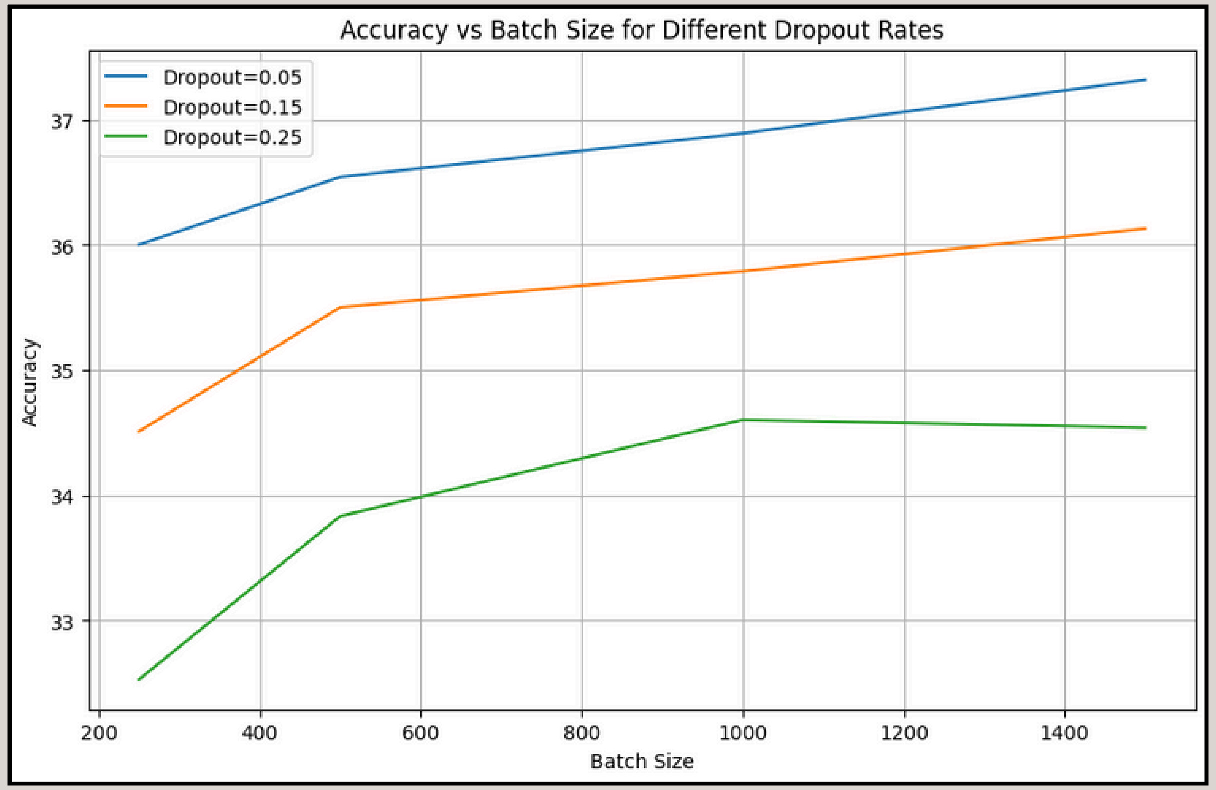
Results - DNN Models

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
DNN 1 Layer Base	37.01	68.64	44.54	54.02
DNN 2 Layer Base	35.67	65.13	44.10	52.59
DNN 3 Layer Base	34.53	64.65	42.56	51.33
DNN 4 Layer Base	30.89	64.37	37.26	47.20
DNN 1 Layer Dropout 0.05 Batch 1500	37.32	69.91	44.46	54.36
DNN 1 Layer Tuned	37.06	69.92	44.09	54.08
DNN 2 Layer Tuned	37.59	67.20	46.03	54.64
DNN 3 Layer Tuned	36.58	67.13	44.56	53.57
DNN 4 Layer Tuned	35.40	66.93	42.90	52.28

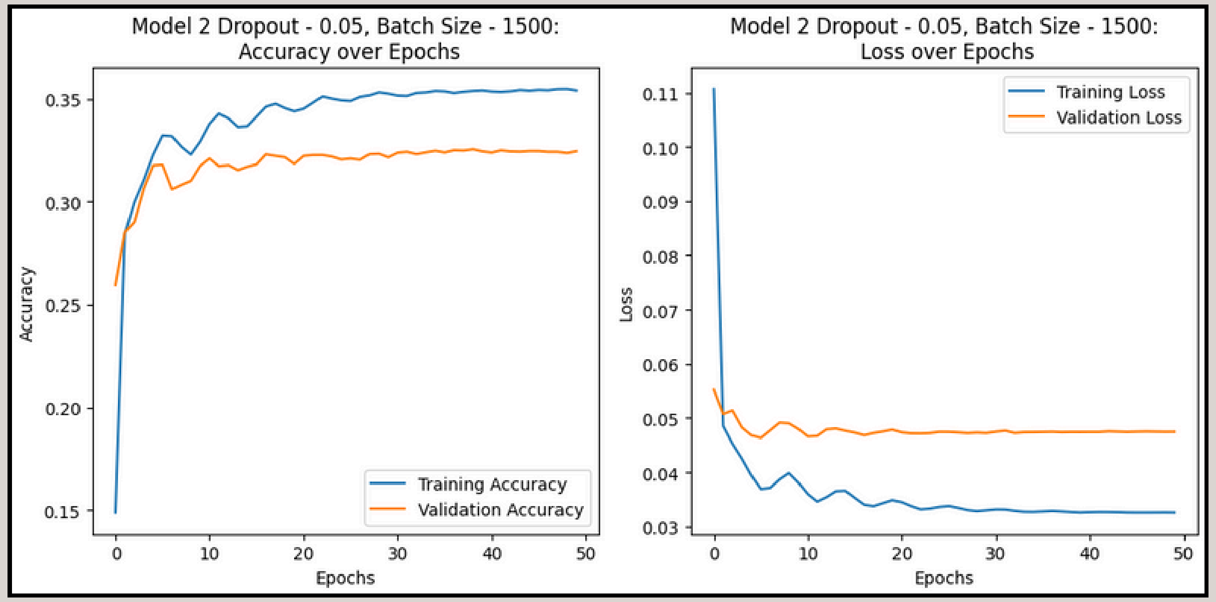


Performance of best "baseline" DNN Model (1 layer)

Accuracy = 37.01%



Finetuning Dropout Rate and Batch Size based on best "baseline" DNN Model



Performance of best tuned model (2 layers) with best hyperparameter configuration

Accuracy = 37.59%

Results – LSTM

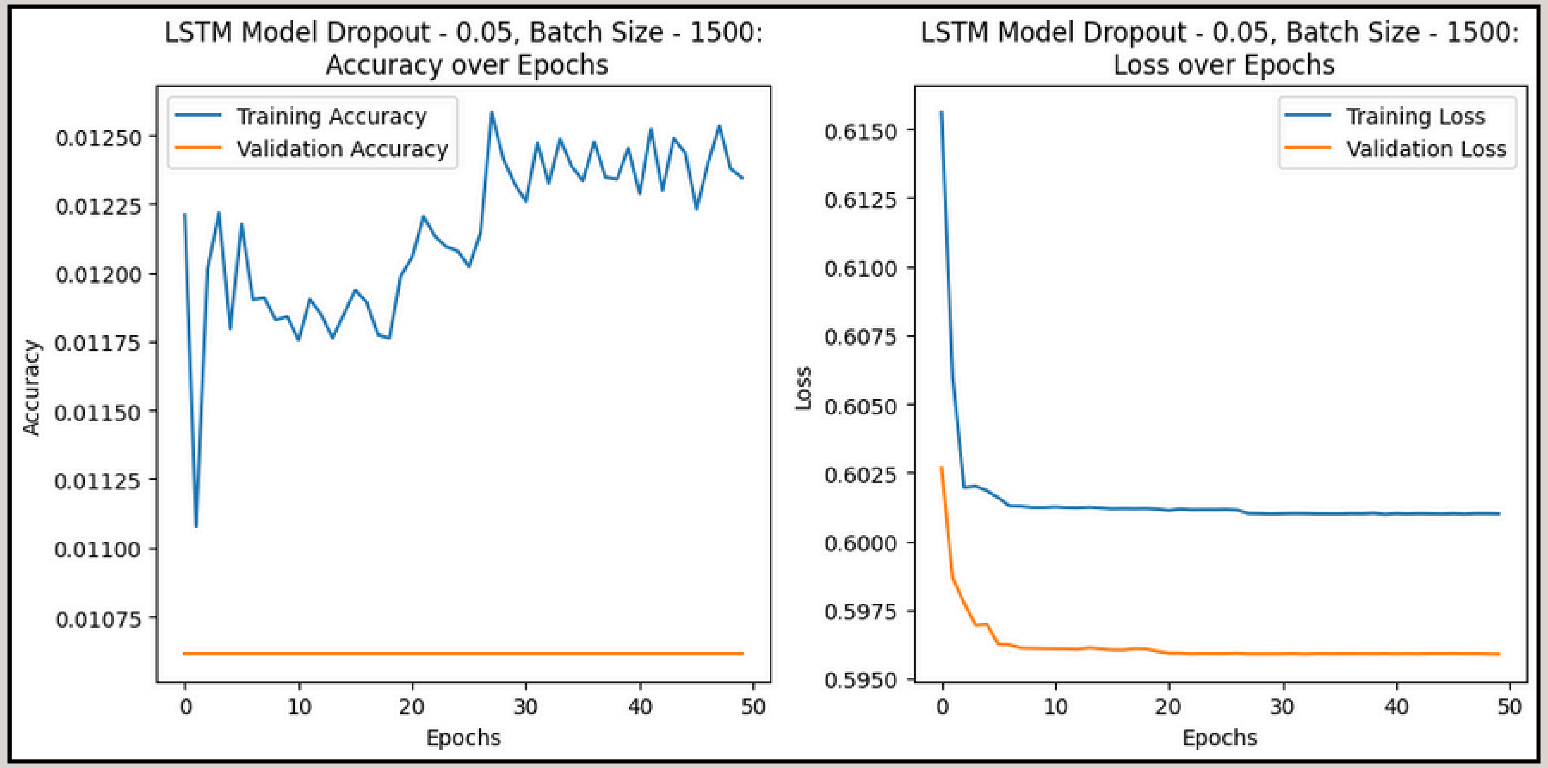
- Poor Performance: 2.67%
- Weight Sharing/Transfer Learning (Weights from best performing DNN)

- Issues:
  - Data Reshaping

- Optimization Methods:
  - Learning Rate
  - Hidden Layers
  - Window Size

- Improvements:
  - Investigate reshaping
  - Bidirectional LSTM

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
LSTM	2.67	-	-	-



LSTM Model Performance

Conclusion

- Automatic piano transcription is challenging due to complexities in converting audio to notation
- Explored deep neural networks (DNNs) and LSTMs to capture patterns and long-range dependencies

Future Works:

- Attention-based models and transformers
- Investigate data augmentation and representation learning techniques
- Leverage multi-modal approaches combining audio with symbolic/visual data

Conclusion

- Limitations in achieving high accuracy
- Exploration of architectures and techniques provided valuable insights
- Need for innovative approaches to effectively capture audio information and translate to notation

New York University

Thank You

Argy Sakti, Sean Wiryadi, Shriya Kalakata