University of Bergen
Department of Informatics

Automatic Drum Transcription using Deep Learning

Author: Runar Fosse
Supervisor: Pekka Parviainen

Add correct NT faculty, instead of MAT-NAT

UNIVERSITETET I BERGEN
Det matematisk-naturvitenskapelige fakultet

March, 2025

Abstract

Lorem ipsum dolor sit amet, his veri singulis necessitatibus ad. Nec insolens periculis ex. Te pro purto eros error, nec alia graeci placerat cu. Hinc volutpat similique no qui, ad labitur mentitum democritum sea. Sale inimicus te eum.

No eros nemore impedit his, per at salutandi eloquentiam, ea semper euismod meliore sea. Mutat scaevola cotidieque cu mel. Eum an convenire tractatos, ei duo nulla molestie, quis hendrerit et vix. In aliquam intellegam philosophia sea. At quo bonorum adipisci. Eros labitur deleniti ius in, sonet congue ius at, pro suas meis habeo no.

Write proper abstract

Acknowledgements

Est suavitate gubergren referrentur an, ex mea dolor eloquentiam, novum ludus suscipit in nec. Ea mea essent prompta constituam, has ut novum prodesset vulputate. Ad noster electram pri, nec sint accusamus dissentias at. Est ad laoreet fierent invidunt, ut per assueverit conclusionemque. An electram efficiendi mea.

`Runar Fosse`
Monday 10th March, 2025

Write proper acknowledgements

Contents

Chapter 1

Introduction

Within the field of Music Information Retrieval (MIR), the task of Automatic Music Transcription (AMT) is considered to be, both an important, and challenging research problem. It describes the process of generating a symbolic notation from audio. Most instruments are melodic, where key information for transcription would be to discern pitch, onset time, and duration. This stands in contrast to percussive instruments, which are inherently event based. This sets the stage for Automatic Drum Transcription (ADT), which is a subfield of AMT, specifically focusing on drums and percussive instruments. [2]

Previously, a popular approach to ADT was using signal processing, which later developed to using classical machine learning methods [2]. However in later years, deep learning has shown to be quite effective [3].

Provide a good introduction into the master thesis, menitoning AMT, ADT and why deep learning is suited for such a task.

1.1 Thesis statement

Present the aim of the thesis here. And the questions! How do we train a model capable of solving such a task at a high performing level. More specifically:

What architectures are suited for learning such a task? What datasets / combination of datasets makes the model generalize best? Of the many techniques made to help models learn this task, which ones actually help? (Ablation)

Remember the concrete What do we want to figure out.

Chapter 2

Background

2.1 Transcription Task

Mention some more detailed background into ADT, Drum Transcription in the Presence of Melodic Instruments (DTM), etc. Mention some more details into how it started, how it is going, etc. Also why this is of interest. (Maybe also what is missing)

Mention the sequence to sequence prediction.

2.2 Related work?

Maybe mention some ADT related work.

Chapter 3

Datasets

Mention the different already existing datasets used.

3.1 ENST

3.2 MDB

3.3 EMG-D

3.4 ADTOF-YT

Chapter 4

Methods

4.1 Task

Precisely explain the task we are solving. Explain what the input data is, what the labels are. Give intuition into what exactly we want our model to predict.

Here we also explain the input and output, i.e. the data and the labels. What do they look like in their un-preprocessed form?

Here we can also give a figure into the pipeline itself, for better intuition.

4.2 Pipeline

Talk about the general ADT pipeline.

4.2.1 Preprocessing

Now explain what we do to the data before prediction. Explain why we use log-spectrograms / log-mel-spectrograms and how they are computed. Shortely explain what they mean, intuition. Explain the preprocessing step we do afterwards.

And explain how we preprocess the labels (target widening, etc.).

Figures?

4.2.2 Training

Mention the loss function used, and why we use this (BCEWithLogitsLoss).

Mention the computation of infrequency weights, i.e. how they are computed, why they are computed, the intuition into how they will help us...

4.2.3 Postprocessing

Mention how model outputs a ”confidence in event happening” distribution, which we want to discretize into events. I.e. explain Vogl’s peak picking algorithm [1].

Need some figures here.

4.2.4 Performance Measures

Mention how F-measure (F1-score) is the most used and why. Compare this to accuracy, balanced accuracy. Mention precision, recall and their meaning.

Mention the difference in class-wise, micro- and macro-F1, and why we choose to focus on class-wise and micro in this thesis. Also mention how these are all computed.

Lastly, mention what denotes a True Positive. Mention why we allow a window for a prediction to be correct.

Figures!

4.3 Architectures

Mention the different architectures trained in the first task. Mention something about why we chose them, (e.g. have they shown to be promising in the past? In other fields?).

Figures!

4.4 Datasets

Talk about the different datasets used, how they differ and why we use them.

4.5 Experiments

Here we mention the setups for each of the experiments.

Mention that we use RayTune to train, with PyTorch models. Mention that we only used RayTune's FIFOScheduler, and how for random search / grid search we used their built in parameter space functionality.

Mention that every single experiment was trained for at most 100 epochs, with a early stopping if validation loss didn't decrease within 10 epochs. Mention the learning rate scheduler, where we reduce the learning rate by a factor of 5 if the model hasn't improved in the last 3 epochs.

Mention that we perform early stopping on the validation loss (and why, like the smooth nature, overfitting prevention, etc.), where as we store the best performing model based on the validation F1-score (due to this representing overall prediction performance).

Mention that every experiment is model selected using hold-out validation, and best model is chosen based on micro F1-score.

4.5.1 Architecture experiment

Shortly mention what we do, what the goal is, what we want to figure out.

Architectures

Mention the different architectures trained, and at what hyperparameters they were trained over.

Datasets

Mention the different datasets used, and tested over.

4.5.2 Dataset generalization experiment

Shortly mention why, what, like in the previous experiment.

Architectures

Mention which architectures we now use, and why we chose them. (And hyperparameters)

Datasets

Now mention which datasets / combination of datasets we use. Mention how we now use zero-shot testing (and maybe why).

4.5.3 Ablation experiments?

Technique 1

Technique 2

Technique 3

Chapter 5

Results

5.1 Architecture experiment

Display a table of results, class-wise and micro-F1: Best archicecture per dataset is bolded.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
CNN	0.5			
RNN	0.4			
Conv-RNN	0.8			
Conv-Attention	0.9			
Transformer	**0.95**			

Display the results in a barplot, to easily capture well-performing models.

Also plot enough information to be able to conclude about overall performance of models, performance on rarer instruments, etc.

5.2 Dataset generalization experiment

Display a table of results, possibly both class-wise and micro-F1 (or maybe just micro-F1): Zero-shot tests have a grayed background, best zero-shot test are bolded.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Dataset 1	0.5	0.3		
Dataset 2	0.4	0.8		
Dataset 1+2	0.8	0.7		
Dataset 3	**0.9**	0.6		
Dataset 1+2+3	0.95	0.8		
Dataset 1+4	0.95	**0.75**		
Dataset 1+2+3+4	0.95	0.82		

Display the results in a barplot, to easily capture well-performing models.

Also plot enough information to be able to conclude about overall performance of models, performance on rarer instruments, etc.

5.3 Ablation experiments

Display results and data to be able to conclude if techniques help training / give better end results.

I.e., do we converge faster? Do we converge to a better minimum?

Could plot some loss over epochs? Need to be thorough (or average) to ensure that gains/losses are due to technique (and not hyperparameter choice, etc.).

List of Acronyms and Abbreviations

ADT Automatic Drum Transcription.

AMT Automatic Music Transcription.

DTM Drum Transcription in the Presence of Melodic Instruments.

MIR Music Information Retrieval.

Bibliography

[1] Richard Vogl, Gerhard Widmer, and Peter Knees. Towards multi-instrument drum transcription, 2018.
URL: `https://arxiv.org/abs/1806.06676`.

[2] Chih-Wei Wu, Christian Dittmar, Carl Southall, Richard Vogl, Gerhard Widmer, Jason Hockman, Meinard Müller, and Alexander Lerch. A review of automatic drum transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1457–1483, 2018. doi: 10.1109/TASLP.2018.2830113.

[3] Mickaël Zehren, Marco Alunno, and Paolo Bientinesi. High-quality and reproducible automatic drum transcription from crowdsourced data. *Signals*, 4(4):768–787, 2023. ISSN 2624-6120. doi: 10.3390/signals4040042.
URL: `https://www.mdpi.com/2624-6120/4/4/42`.