University of Bergen
Department of Informatics

Automatic Drum Transcription using Deep Learning

Author: Runar Fosse
Supervisor: Pekka Parviainen

UNIVERSITETET I BERGEN
Det matematisk-naturvitenskapelige fakultet

June, 2025

Abstract

Automatic Drum Transcription (ADT) remains a challenging task within the field of Music Information Retrieval (MIR), especially when drum sounds are mixed with melodic instruments, otherwise called Drum Transcription in the Presence of Melodic Instruments (DTM). The current standard methodology for solving such tasks is through deep learning. This thesis investigates how different architectures and dataset composition affect ADT performance. Two studies were conducted: one comparing the performances of five different neural network architectures over four standard ADT datasets, and another evaluating how training models on different dataset combinations impacts their generalization ability both on- and Out-Of-Distribution (OOD). A novel dataset, SADTP, is introduced for this latter study is utilized for OOD evaluation. The first study concluded convolutional recurrent architectures perform the best across datasets, however strictly recurrent architectures and the Vision Transformer show promising performance on larger datasets. The second study shows that training on large and diverse datasets from multiple sources could improve both on- and Out-Of-Distribution (OOD) generalization. These findings offer insight into how architecural and dataset choices influence generalization for ADT.

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.

Write proper acknowledgements

`Runar Fosse`
Monday 23rd June, 2025

Contents

Chapter 1

Introduction

Within the field of Music Information Retrieval (MIR), the task of Automatic Music Transcription (AMT) is considered to be a challenging research problem. It describes the process of generating a symbolic notation from audio. The majority of instruments are melodic, where key information for transcription would be to discern pitch, onset time, and duration. This stands in contrast to percussive instruments, where instead of pitch and duration one would focus on instrument classification and onset detection. This sets the stage for Automatic Drum Transcription (ADT), which is a subfield of AMT, specifically focusing on transcribing drums and percussive instruments [44]. Specifically, this thesis will focous on Drum Transcription in the Presence of Melodic Instruments (DTM), the hardest subtask of ADT.

Previously, a popular approach to ADT was using signal processing, which later developed into using classical machine learning methods [44]. In later years, deep learning has shown to be quite effective, evolving into becoming the standard. Therefore, the recent focus of most authors has been to find the best performing deep learning approaches by either; constructing and analysing the best performing model architectures, or by finding datasets which allow models to generalize the best [47].

1.1 Thesis statement

This leads us to two primary questions. Which deep learning architecture is the best suited for solving a task like this? And, what makes a dataset optimal by making models generalize? These are two of the questions we will try to answer in this thesis.

For the former, we will train different model architectures on different, well-known ADT datasets. Specifically, recurrent neural networks, convolutional neural networks, convolutional-recurrent neural networks, convolutional transformers and, novel to the field of ADT, vision transformers. By comparing their performances we could be able to gauge the one best suited for an ADT task.

For the latter, we will select the best performing model architecture from the first question, and train it over several different combinations of the ADT datasets. By performing cross-dataset evaluations, we could analyse and figure out what makes a good ADT, specifically DTM, dataset and how it would enhance a suitable model architecture. For this I also introduce SADTP, a novel dataset solely used for Out-Of-Distribution (OOD) evaluation purposes.

Remember the concrete What do we want to figure out.

1.2 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2: Background — Covers information needed to fully understand the ADT inference pipeline, as well as understanding how performance is represented.

Chapter 3: Architectures — Presents and goes in-depth into each of the different deep learning architectures trained for the first study.

Chapter 4: Datasets — Presents each of the different datasets used within this thesis, as well as comparing their characteristics. In addition, I introduce the novel DTM dataset SADTP.

Chapter 5: Methodology — Covers the specific methods used in this thesis, from dataset preparation to model selection and training procedure.

Chapter 6: Architecture Study — Compares the performances different architectures trained over each dataset, discusses the different results and concludes by selecting the best performing one.

Chapter 7: Dataset Study — Compares the best performing model from the previous study trained over different combination of datasets, and discusses the results when performing on- and OOD evaluation.

Chapter 8: Conclusion — Concludes this thesis by reflecting on the results of the different studies, as well as giving an outlook into what should be covered in future work.

Chapter 2

Background

2.1 Automatic Drum Transcription

As mentioned, ADT describes the task of transcribing symbolic notation for drums from audio. To be even more descriptive, ADT can be split into further tasks. From least to most complex we have: Drum Sound Classification (DSC), where we classify drum instruments from isolated single event recordings. Drum Transcription of Drum-only Recordings (DTD), where we transcribe audio containing exclusively drum instruments. Drum Transcription in the Presence of Additional Percussion (DTP), where we transcribe audio containing drum instruments, and additional percussive instruments which the transcription should exclude. Finally, we have DTM, which describes the task of drum transcription with audio containing both drum, and melodic instruments. [44]

As mentioned, this thesis will focus on the most complex of these, namely DTM. Intuitively, we want to develop a deep learning model which, given input audio, has the ability to detect and classify different drum instrument onsets (events), while selectively ignoring unrelated, melodic instruments.

This task comes with difficulties not seen in the less complex tasks. Zehren et al. [47] describes one example, in where *"melodic and percussive instruments can overlap and mask eachother..., or have similar sounds, thus creating confusion between instruments"*.

Deep learning has shown to be a promising method to solve such a task, and several different approaches have been tried, many with great success. Vogl et al. [42, 41] displayed good results with both a convolutional, and a convolutional-recurrent neural

network. Zehren et al. [47, 48] focused on datasets, showing that the amount of data and quality of data are equally important to get good performance. Most recently, Chang et al. [10] explored an autoregressive, language model approach. This approach explored multi-instrument transcriptions, but their results regarding DTM were notable.

This reinforces the fact that there still exist many approaches to attempt, which could lead to a general improvement for both general ADT and DTM models.

2.2 The Drum Set

The drum set is a collection of percussive instruments like different drums, cymbals, and possibly different auxillary percussions. A drum set can vary in what it is composed of, however a standard kit usually consists of a snare drum, a bass drum, one or more tom-toms (toms), one or more cymbals (crash and ride), and a hi-hat cymbal [30].

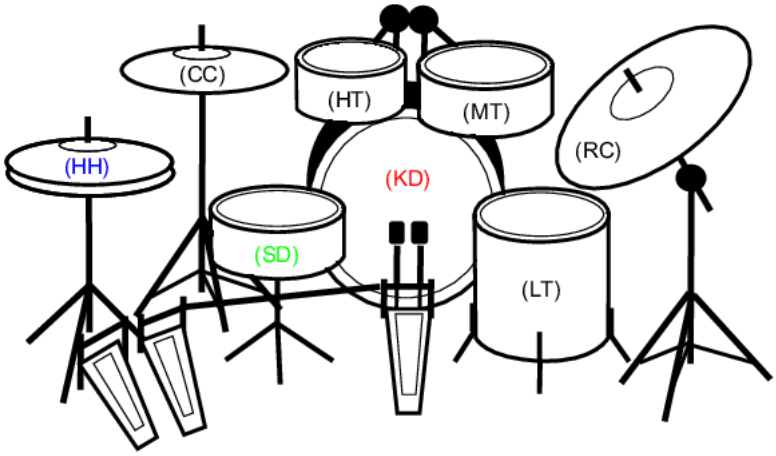

Figure 2.1: Example of the different instruments on the drumset. They are the Kick Drum (KD), Snare Drum (SD), Hi-Hat (HH), Crash Cymbal (CC), Ride Cymbal (RC), High Tom (HT), Mid Tom (MT), Low Tom (LT).

As mentioned, percussion like the drum set, stands in contrast to other musical instruments in that the different ways of playing the same instrument often differ a lot in their *"audible footprint"*. The snare drum, bass drum and hi-hat all have quite different timbres, frequency span, volume, and all in all fundamentally are different instruments.

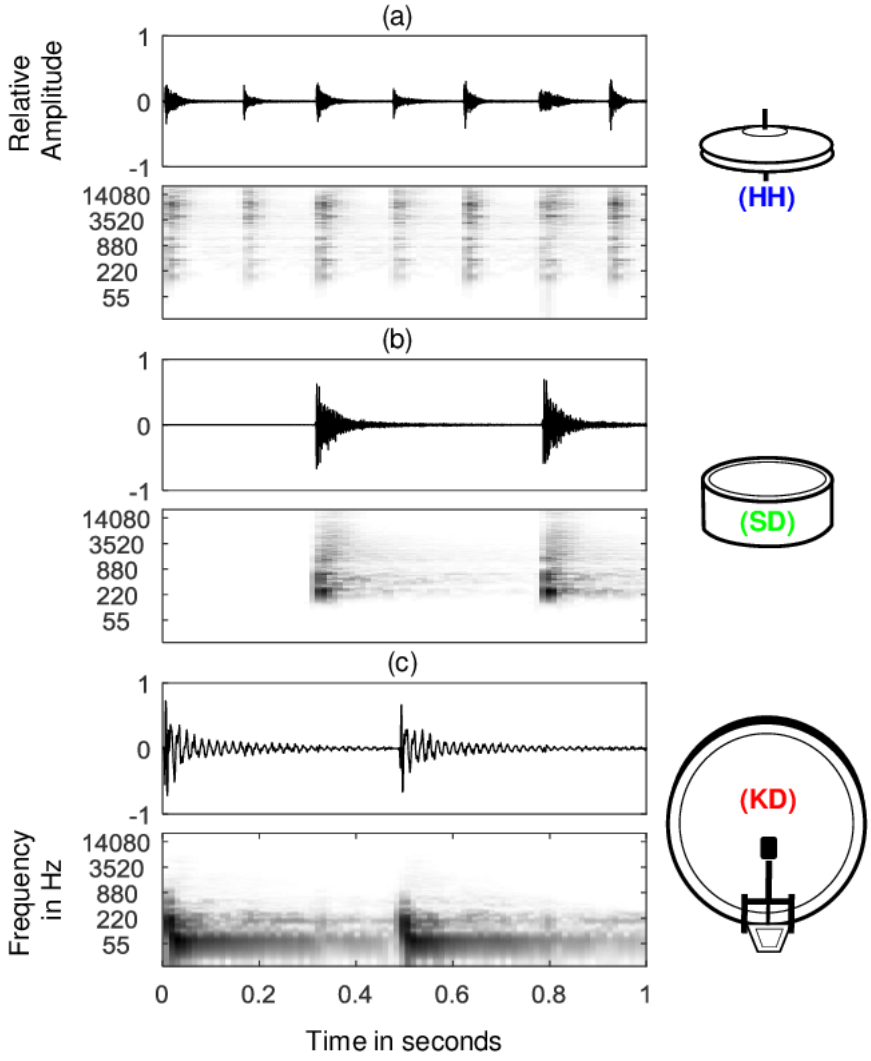

Figure 2.2: Example of the different audible footprint for drum set percussion. Plotted are the waveforms of three different drum instruments played at different speeds, together with its corresponding spectrogram. As we can see, each instrument event different significantly in how they look in and affect the spectrogram.

Mention the different drum set instruments. Mention how they all have different musical properties, like frequency, timbre, etc (show waveforms maybe?). Also mention the most fundemental ones, and how Bass, snare and hi-hat are more important than e.g. the mid-tom or something.

2.3 Transcription Task

Understand the pipeline for a transcription task, specifically ADT, is crucial for this thesis. It all starts with an initial audio waveform representing the audio track we want

to transcribe, usually split into smaller non-overlapping partitions [42, 15]. This is parsed into a spectrogram, which unravels the frequencies of the audio wave while condensing information across time, making the input features easier for a model to handle and interpret. The spectrogram is then input into an ADT model, such as a Deep Neural Network (DNN), which reads this input spectrogram and for each timeframe predicts the probability that a certain instrument is played. These continuous likelihood predictions are then postprocessed into a more readable and intended format, such as drum notation.

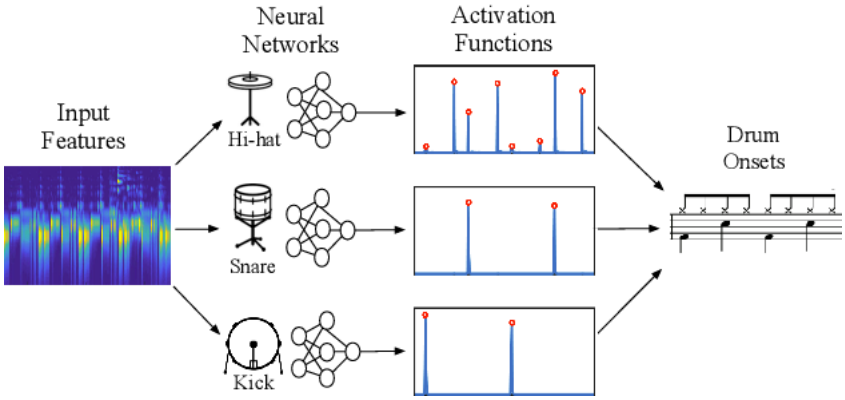

Figure 2.3: Example of what the prediction pipeline of an ADT model could look like. Given an input spectrogram, a model predicts a likelihood distribution of each instrument appearing (called the activation functions), which then is quantized into drum onsets and sheet music notation.

Following this pipeline there are certain parameters which need to be kept in mind when constructing the DNN, namely: what does the input and ouput of model look like? The input is given as a spectrogram, but these can vary in size and format based different parameters. These will be discussed further in a later section. The output of the model has a sequence size based on the input spectrogram, but also based on the number of instruments we want to transcribe.

Early ADT literature used a 3-instrument approach, predicting the basic Kick Drum (KD), Snare Drum (SD) and Hi-Hat (HH) [40]. These give a useful basis in investigating wether ADT problems are solvable using deep learning methods, but are too sparse as this leaves out instruments crucial to the basic drum set. To address this, we expand to a 5-instrument approach, including cymbals (capturing both Crash Cymbal (CC) and Ride Cymbal (RC)) and Tom-Tom (TT) (capturing all toms) in addition to the three prior instruments. This makes the problem slightly more complex, but allows for denser coverage by representing a full drum set.

2.4 Audio

Sound has be described as *"the sensation caused in the nervous system by vibration of the delicate membranes of the ear."* [1]. In short, sound is the human perception of acoustic waves in a transition medium, like air. These waves, consisting of vibrating molecules, get sensed by our auditory organs and perceived by the brain.

Thus sound can be described as the propogation and perception of waves. Mathematically, waves can be studied as signals [9]. To represent these sounds digitally, as *audio*, one can express these waves as a signal, giving rise to the *waveform*. The waveform is a representation of a signal as a graph, and charts the amplitude, or strength of the signal, over time.

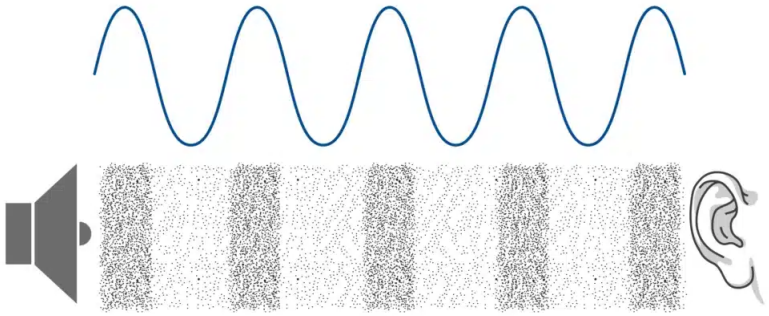

Figure 2.4: Soundwave to waveform relationship

For monophonic sound, this waveform is a one-dimensional representation. Even though this is an excellent way of storing audio digitally, it is very compact. There have been deep learning models working directly with these waveforms, e.g. Oord et al.'s WaveNet [38], however the task of parsing and perceiving such a signal is a complex one.

2.4.1 Fourier Transform

The Fourier Transform is a mathematical transformation which, given a frequency, computes its significance, or intensity, in a given signal. As we've established, audio is

represented as a signal, and we can therefore use this transform to turn this audio signal into frequency space.

The fourier transform is a complex transformation. Given a signal f, we can compute the integral

$$\widehat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi\xi x} dx$$

for a frequency ξ, resulting in a *complex* number. This number consists of a *real* part and an *imaginary* part. The real part consists of the amplitude of a certain frequency, where as the imaginary part consists of the phase. This information is what allows us to, for a given signal, figure out which frequencies it is made out of and how much each frequency contributes.

By doing such a transform, we turn our temporal data into spectral data. This initively *untangles* our signal into its respective base frequencies. Such an transformation could lessen the complexity of the task, making *understanding* of audio easier.

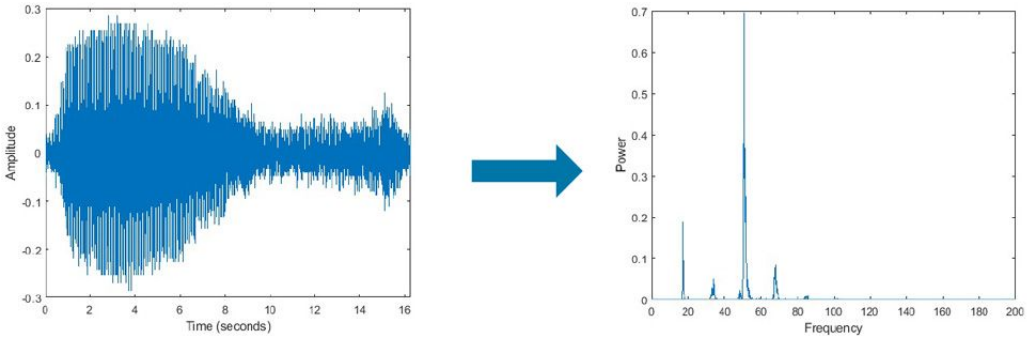

Figure 2.5: Application of a Fourier Transform

Note that the Fourier Transform is invertible, meaning that, given information about each frequency, we can perform a similar integral and reconstruct the original signal. In signal processing, this property is exploited heavily.

2.4.2 Discrete Fourier Transform

The Fourier Transform is defined as an integral over continuous time. On computers, instead of storing signals continuously we store signals using a discrete number of samples. Each signal's *sampling rate* describes how many samples a signal contains per second of audio, and is denoted in *Hz*.

To extract frequency values from these signals, we instead have to use the Discrete Fourier Tranform (DFT). Intuitively this works as the normal Fourier Transform, but ported to work on discrete-valued signals. It is given by the formula

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi \frac{k}{N} n},$$

where k denotes the frequency and N the number of discrete samples.

Add a example figure of FT vs DFT

2.4.3 Nyquist frequency

When we discretize a signal, e.g. when going from continuous audio waves in the air to discrete audio signals on a computer, we could lose some information. The discrete representation of the signal is an *approximation* which quality is directly dependent on the sampling rate. The higher the sampling rate, the *closer* we are to the original, continuous signal. However a higher sampling rate comes at the cost of needing to store these signals at a higher precision. A lower sampling rate would need less information stored, but this could also mean a less precise signal approxmation.

Aliasing is the phenomena where new frequencies seem to emerge in undersampled signals. For a given discrete signal, the *Nyquist frequency*, equal to half the sampling rate, is the maximum frequency a signal accurately can represent. Thus to prevent aliasing, one would need to store a signal with a sampling rate of at least double the maximum frequency.

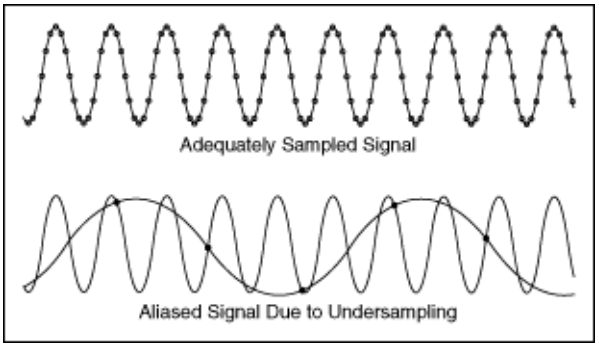

Figure 2.6: Example of aliasing in an undersampled signal.

Regarding the DFT, it here directly follows that the maximum frequency we accurately could extract information about is proportional to the sampling rate of the signal, that being it is equal to the nyquist frequency of the signal.

2.4.4 Fast Fourier Transform

Keen-eyed computer scientists may have spotted that the DFT runs in $\mathcal{O}(n^2)$ time as we, for every frequency in the range $[0, N]$ have to sum over N different values. In other words, the DFT algorithm scales quite poorly. Take into account that the standard sampling rate for audio is 44.1kHz, i.e. 44100Hz, then we can see that the DFT could be inefficient. [3]

The Fast Fourier Transform (FFT) is an algorithm which solves this problem, and instead computes the DFT of a signal within $\mathcal{O}(n \log n)$ time. Described by Gilbert Strang as *"the most important numerical algorithm of our lifetime"* [37], this practically solves our scaling problem, and allows us to efficiently extract spectral information from a signal regardless of sampling rate.

There exist many different implementations of the FFT. However the Cooley-Tukey algorithm is by far the most used FFT and optimizes calculations through a *divide and conquer* approach, utilizing previous calculations to compute others. [12]

2.4.5 Short-time Fourier Transform

The Fourier Transform comes with some drawbacks, notably how by moving from time space into frequency space, we lose temporal information. For certain tasks this might be sufficient, but the temporal dimension is vital when working with transcriptions and ADT tasks. We've seen how the Fourier Transform computes the frequencies of a signal, but what happens if we had applied the same transform to smaller, *partitions* of a signal.

This leads us to the Short-time Fourier Transform (STFT). By instead of transforming the whole signal, we transform smaller *windows*, we could gain insight into the frequency space while keeping temporal information relatively intact. This turns our data from being one-dimensional into two-dimensional, giving us insight into the intensities of different frequencies, along different timesteps.

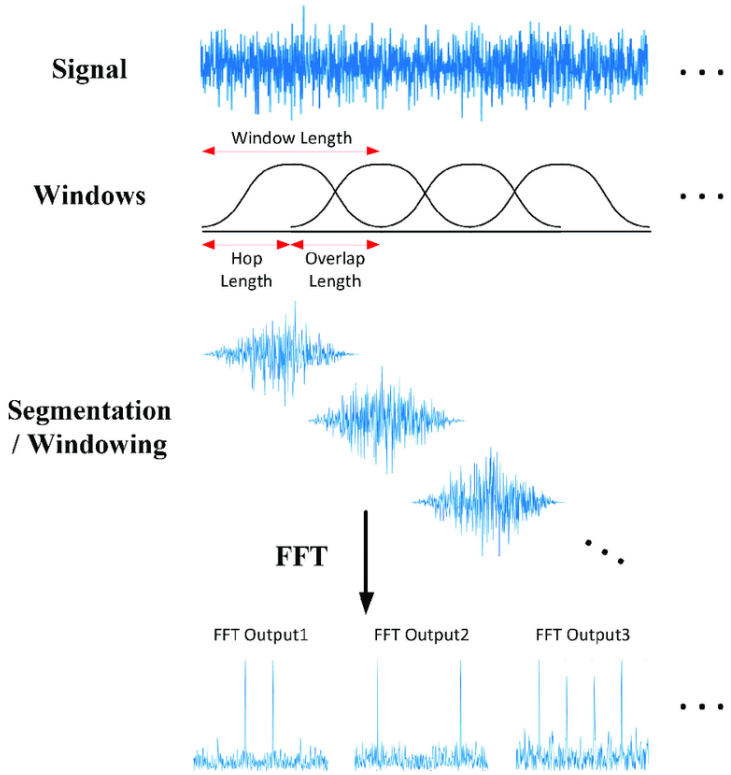

Figure 2.7: Example of the STFT. Given a signal we partition into separate windows, with an applied windowing function. Each window is then applied an FFT.

The STFT comes with several parameters which affect the output format, most importantly the *window function*, *window length* and *hop length*. Due to a phenomena called *spectral leakage*, where spectral information spread larger than a window bleeds into other frequencies, a windowing function is applied per window. A usual function for this is the "Hann window", as seen in figure 2.8. The window length is crucial regarding the time-frequency resolution, where a larger window length provides better frequency resolution, but a worse time resolution. This because it decides how much of the original signal each window FFT should cover. The last, hop length, also affects the temporal resolution, by deciding how many windows the signal should be partitioned into. This can also be used to give each window a temporal index. By using a lower hop length, each window would represent a smaller time window, but would require a higher number of FFTs due to increasing the number of windows.

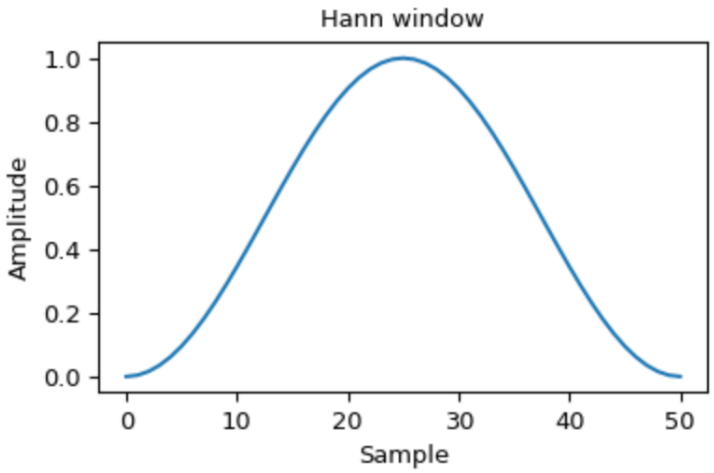

Figure 2.8: The Hann window function, a function applied to each window during the STFT to lessen the effects of spectral leakage.

For ADT, a common approach is to use the previously mentioned Hann-window, along with a window size of 2048 input samples, and a hop length representing 10ms, equal to the sampling rate divided by 100 [44, 40, 42, 47].

2.4.6 Spectrogram

The STFT, similar to the standard Fourier Transform, returns the data as complex values. To turn these into strictly real values without discarding data, we could compute the spectrogram. This is done by squaring the absolute value of each complex number.

This results in a 2-dimensional, real representation of our signal. A representation like this is equivalent to an *image*, but can also still be modelled as a time series. In this way, we've converted our audible information into visual information. Naturally, these spectrograms can be visualized using e.g. a heatmap.

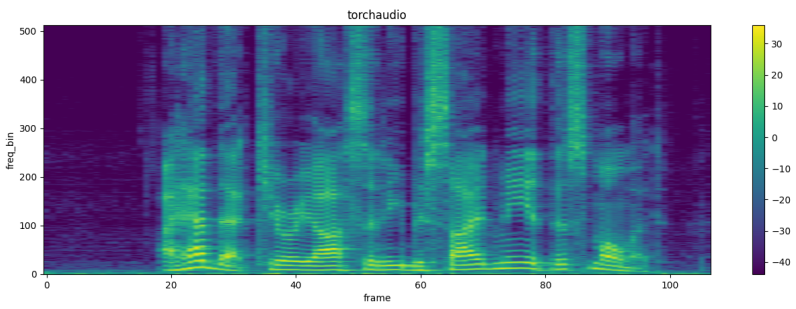

Figure 2.9: Heatmap of an audio spectrogram.

One drawback about the spectrogram is that it contains no information about the phase of the signal it represents. That means it will not be possible to reverse the process and recreate the exact original signal. However, one could try to create an approximation like is done with the Griffin-Lim algorithm [19].

2.4.7 Filters

Signal frequencies and human perception have a special relationship. We humans percieve logarithmic differences in frequencies as a linear difference in pitch, and we tend to be better at distinguishing differences in lower frequencies than higher. E.g., the notes A_2 and B_2 have the same perceptual pitch difference as D_7 and E_7, even though their difference in frequency, $B_2 - A_2 \approx 13.471$Hz and $E_7 - D_7 \approx 287.703$Hz, are vastly different. As the frequency bins in a spectrogram are linearly spaced, this leads to the spectrogram not representing each frequency equally compared to our perception.

To solve this, we can filter the spectrogram into different bins, more suited to represent our perception of sound. This filtering is done by matrix multiplying our spectrogram with a *filterbank*; a matrix representation of different filters.

Mel Spectrograms

The mel scale, presented by Stevens, Volkmann, and Newmann in 1937, is a transformation from the frequency scale to the mel scale. These mels have the property such that a linear difference in mels are percieved as linear differences in pitch. Application of mel-filters result in the *mel spectrogram*, and are widely used when dealing with audio in machine learning, and successful applications have been seen in AMT. [43, 15, 10, 44, 18, 48]

Logarithmic Filters

The mel scale was created to mimic human perception of sound, however within ADT there is a different trend. By instead using logarithmically spaced filters, centered on the note A_4, we get a *logarithmically filtered spectrogram*. Intuitively one could assume this, instead of mimicing human perception, ports the spectrogram into a format preserving musical relationship and information. This seems to be a standard for ADT and has been used extensively by the likes of Vogl et al. Specifically, one constructs 12 different

logarithmically spaced filters per octave, for a frrequency range from 20 to 20,000 Hz. This results in a spectrogram with $D_{STFT} = 84$ frequency bins [44, 42, 41, 47].

Add a example figure of Spectrogram vs Mel vs Logarithmic

2.5 Transcription

Transcription refers to a process in which we convert information from an audible format, like music, to another medium. This medium then contains a *description* of said audio. As we focus on a musical context, there are a few notable such mediums.

2.5.1 Sheet Music

Sheet music is a written transcription using musical notation that, for a given instrument, contains the *recipe* for a musician to play parts of the original recording. This is the standard when it comes to printing arrangements, and is extensively used by musicians.

Sheet music is typically descriptively exhaustive, and could contain information about musical properties like instrument onsets, tempo, velocity, etc.

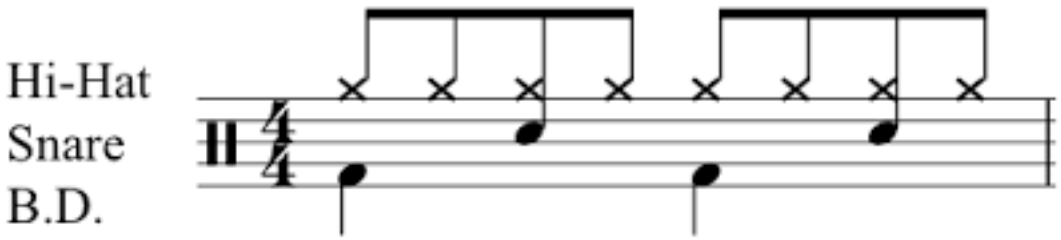

Figure 2.10: Example sheet music for a drumset

2.5.2 MIDI Annotations

Musical Instrument Digital Interface (MIDI) is the industry standard for handling music digitally. It is a binary format, containing sequences of commands that allow digital interfaces to *synthesize* music. As it is binary, it is unreadable to us humans without

translating it into another format. When computers play MIDI arrangements, the MIDI sequences are parsed at a constant speed, playing different sounds through *note on*/*note off* events, delayed by time *deltas*. Similar to sheet music, MIDI is also very descriptive. And one could say that, intuitively, MIDI is to a computer what sheet music is to a musician.

Recently, outputting transcriptions in a MIDI-like format has been attempted in DTM, and has shown to be promising. Utilizing a sequence-to-sequence Natural Language Processing (NLP) approach, Gardner et al. presented MT3 [15], a model inputting spectrograms and outputting MIDI events autoregressively. This format was expanded on by Chang et al.'s YourMT3+ [10], using a Large Language Model (LLM) instead.

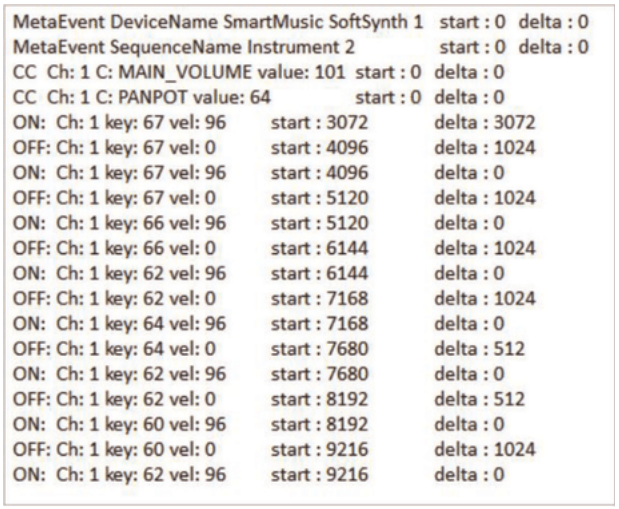

Figure 2.11: Example MIDI arrangement in a readable format

2.5.3 Activation Functions

In machine learning, the task of detecting instrument onsets could be described as a multi-label sequence labeling task. This involves, for each timeframe in a sequence, predicting a probability, or rather confidence value, that a certain instrument onset happens. In the domain of MIR and AMT, it has become common place to describe these confidence distributions as *activation functions*; not to be confused with the general deep learning term, activation functions like ReLU or sigmoid. [44, 35, 42]

This way of frame-level prediction is extensively used within onset detection in ADT and is the approach we will be taking in this thesis.

Need an isolated example of activation functions and respective labels.

Figure 2.12: Example of ADT activation function output

Peak-picking

When predicting activation functions, we need a separate post-processing step to turn these confidence distributions into onset events. By utilizing a standard *peak-picking* algorithm, we can isolate and enhance peaks in these activation functions, and go from a continuous distribution to a collection of discrete events.

The peak-picking algorithm, introduced in its current form by Böck et al. [5], defines that a prediction $\hat{y}_n$ at timeframe n is a *peak* if it fulfills the three conditions:

$$\begin{aligned}\hat{y}_n &= \max(\hat{y}_{n-m}, ..., \hat{y}_n, ...\hat{y}_{n+m}),\\ \hat{y}_n &\geq \text{mean}(\hat{y}_{n-a}, ..., \hat{y}_n, ...\hat{y}_{n+a}) + \delta,\\ n &\geq n_{\text{last onset}} + w.\end{aligned}$$

For appropriately trained deep learning models, Vogl et al. [42] showed that the peak-picking parameters which gave the best results were $m = a = w = 2$ and $\delta = 0.1$.

2.6 Performance Measure

2.6.1 Correct Predictions

Our machine learning models predict instrument onset events on a frame-level basis. In other words, are predictions are very granular, and we need some way to decide when a prediction is correct versus incorrect. In ADT, a standard has become to allow a *tolerance window* where event predictions are correct if they lie within a certain time window, often between 25ms and 50ms. A side effect of this is that, by shifting our focus to predicted events, we lose information about *not* predicting any events [40].

2.6.2 Accuracy

For classification tasks, a standard performance measure would be *accuracy*:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

Summing up correct predictions, True Positives (TP) and True Negatives (TN), and dividing by total number of predictions, sum of TP, TN, False Positives (FP) and False Negatives (FN), we find a model's probability of having a correct prediction.

This performance measure falls short in that it is very susceptible to imbalanced datasets. In ADT, most timeframes contain no onset, meaning a naïve predictor would get a high accuracy by never predicting any onsets. Another problem with accuracy is that, due to our tolerance window approach we do not have quantities for TN, such that the standard accuracy computation is incomputable.

2.6.3 F1-score

Mentioned above are some of the reasons why *F1-score* has become the typical performance measure within ADT. F1-score combines and tries to maximize two different performance measures, namely *precision*;

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

and *recall*;

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The precision of a model can tell us how good it is at *hitting* predictions. *Perfect precision* happens when a model has no FP, i.e. never predicting an event where one doesn't happen. Recall is similar, but represents the other end of the stick. It tells us how good a model is at *not missing* predictions. *Perfect recall* happens when a model has no FN, i.e. never *not* predicting an event where one does happen.

As mentioned, F1-score combines these two measures in an aggregate performance measure by computing their harmonic mean:

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

By maximizing F1, we simultaneously maximize both precision and recall as well, reaping all their benefits.

2.6.4 Micro vs. Macro

There are different ways of computing and combining F1-score on multi-label data. Even though they might seem similar, they fundamentally represent different information, and thus the choice in which one to select is crucial.

Macro F1-score is computed through the arithmetic mean of the classwise computed F1-scores. Finding a model which maximizes this measure would be similar to finding the model which performes best on each of the separate classes, preventing a class from taking priority due to imbalanced datasets. Relating this to ADT, it would mean focusing on transcribing each instrument equally well.

Micro F1-score is computed through finding the F1-score with global TP, FP, FN values. Maximizing this would mean prioritizing classes that occur more frequently in the datasets. Such as in ADT, this would mean focusing on transcribing instruments which appear often, like the snare or base drum, over rarer instruments like the toms.

For ADT, the trend has been to select Micro F1-score as the main performance measure, due to its ability to show a model's *general* performance on musical pieces. We want our model to maximize their ability to transcribe music, not maximize their ability to transcribe each instrument in said music. ADT, prioritizing frequent instruments is relevant. As mentioned previously, the more frequent instruments lay the ground work for the fundamentals, and could be said to be more important than scarcely occuring ones.

Chapter 3

Architectures

Finding a suitable architecture is a vital step in creating a well-performant deep learning model. By leveraging different techniques, we balance introduction of inductive biases and possibility for complexity, which hopefully can help us end up with a generalizing model.

3.1 Recurrent Neural Network

The Recurrent Neural Network (RNN) is a standard architecture when it comes prediction on sequence data. It has been tried and tested, showing promising results for audio tasks.

The fundamental building block for RNNs is the *recurrent unit*. It iterates the whole input sequence, storing information from previous timesteps in a form of memory, through maintenance of a *hidden state*. This can be extended to gaining information about future timesteps by using *bidirectional* versions. In this way, prediction on current timesteps are affected by the information from surrounding timesteps. This is relevant in tasks such as ADT as auditory information usually spreads over several timesteps, e.g. the timbre of an instrument event lingering after onset.

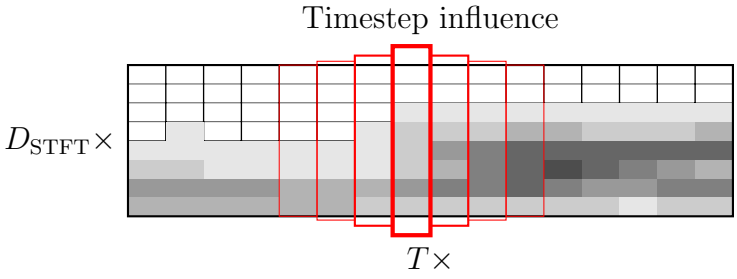

Figure 3.1: An example of how bidirectional RNNs allow for adjacent timesteps to influence the prediction of the current timestep. The background is a spectrogram, and the red boxes represent influence in the middle timeframe. The height of the box help visualize influence strength.

However, traditional RNNs suffer from the *vanishing gradient problem* due to a timestep's influence diminishing with distance, making *long range dependencies* harder to learn. Different architectures have been developed to try to overcome these issues, such as the Gated Recurrent Unit (GRU) by Cho et al. [11], and Long Short-Term Memory (LSTM) by Hochreiter and Schmidhuber [22].

It has been shown that GRUs and LSTMs are capable of learning ADT related tasks, and is therefore in interest to comparatively measure how their efficiency stands in regards to other architectures [35, 40, 41, 47].

3.1.1 Implementation

Our RNN architecture consists of several bidirectional recurrent units, ending in a frame-wise linear layer. For the Bidirectional Recurrent Unit (BiRU), we train either a GRU or an LSTM model as hyperparameters, in addition to search over number of layers L and hidden size H, selecting the one with best performance. At last, we have a linear layer, outputting onset probabilities for each of the drums per timeframe.

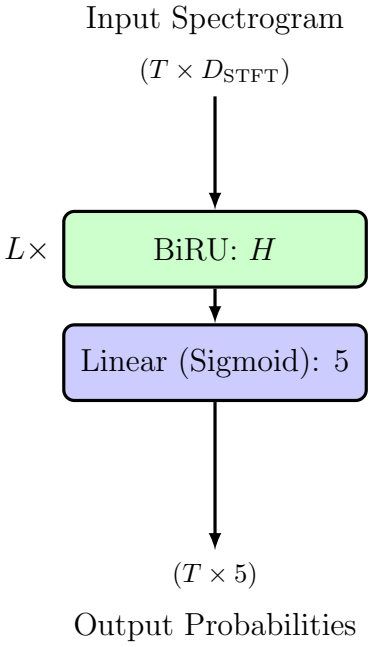

Figure 3.2: RNN architecture structure.

Hyperparameter		Values
L	Number of layers	{2, 3, 4, 5, 6}
H	Hidden size	{72, 144, 288, 576}
BiRU	Bidirectional Recurrent Unit	{GRU, LSTM}

Table 3.1: The different hyperparameters and their respective values tuned to train the Recurrent Neural Network.

3.2 Convolutional Neural Network

We've mentioned that spectrograms can be treated as images. Therefore it would make sense to try an image focused approach, by utilizing *convolutions*. By applying convolutional layers, each timestep gets access to information around itself, a *context*. These convolutional layers make up the primary building blocks for the Convolutional Neural Network (CNN).

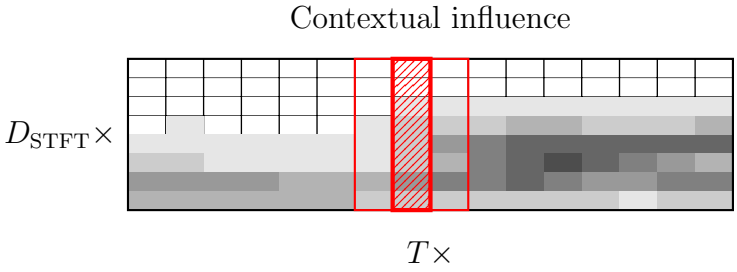

Figure 3.3: An example of how CNNs allow for a fixed context of influence from neighbouring timesteps when predicting each timestep. The background is a spectrogram, and the red boxes represent influence in the middle timeframe, being described by the red shaded region.

CNNs have been shown to give reasonable performance within ADT. This could be due to contextual information being important for identifying instrument onsets, and making learning easier for our models [41].

3.2.1 Implementation

Our CNN architecure consists of I initial convolutional blocks. Inside this block, convolutional layers have an increasing number of kernels $C = \{32, 64, 96\}$, intuitively leading to an increase in complexity along with depth. Note that this is not a hyperparameter, but just denotes that we increase the amount of kernels the deeper into the convolutional block we go. Then we have a varying amount of fully connected layers L, projecting into a latent space sized H. Both the convolutional and fully connected layers are followed by a Rectified Linear Unit (ReLU) activation function. Lastly, an output layer computed each instrument's onset probabilities.

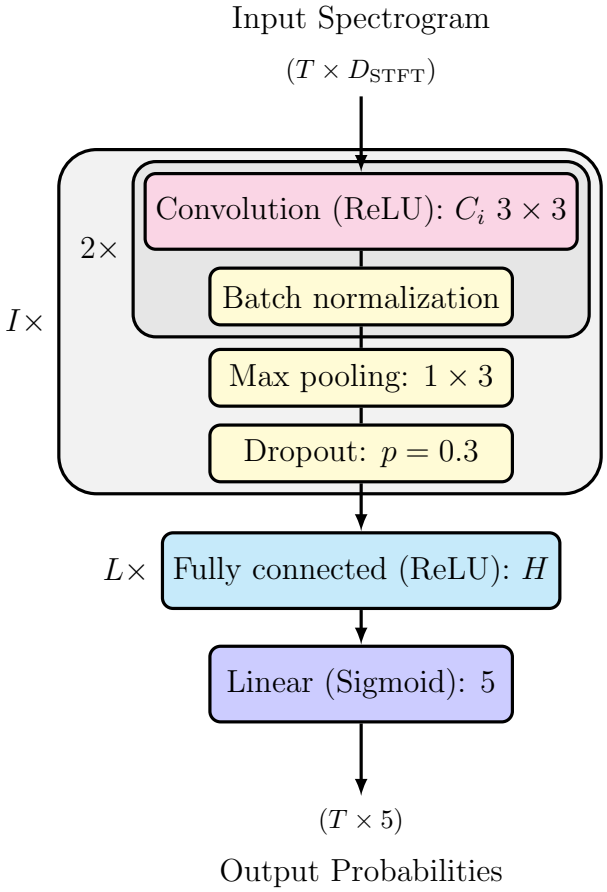

Figure 3.4: CNN architecture structure.

	Hyperparameter	Values
I	Number of convolutions	{1, 2, 3}
L	Number of layers	{2, 3, 4}
H	Hidden size	{72, 144, 288, 576}

Table 3.2: The different hyperparameters and their respective values tuned to train the Convolutional Neural Network.

3.3 Convolutional Recurrent Neural Network

The previous features, recurrent layers and convolutions, are not mutually exclusive. Theoretically they can harmonize together, complementing eachother for easier learning. This results in the Convolutional Recurrent Neural Network (CRNN) architecture.

Intuitively, the CNNs ability to process images like the spectrogram, together with the RNNs ability to understand temporal sequence data should prove beneficial for ADT

tasks. And indeed, this combination of cross-timestep memory and contextual data representation has shown to be insightful [41, 42, 47].

3.3.1 Implementation

We begin with a fixed-size convolutional block with $I = 2$, as used by several ADT authors [41, 47]. Following the CNN, it has an increasing number of kernels $C = \{32, 64\}$ the deeper into the convolutions we go. We then, similar to the RNN, have a BiRU layer of either a GRU or LSTM, with number of layers $L \in \{2, 3, 4, 5\}$ and hidden size $H \in \{72, 144, 288\}$. Output probabilities are then computed through the final linear layer.

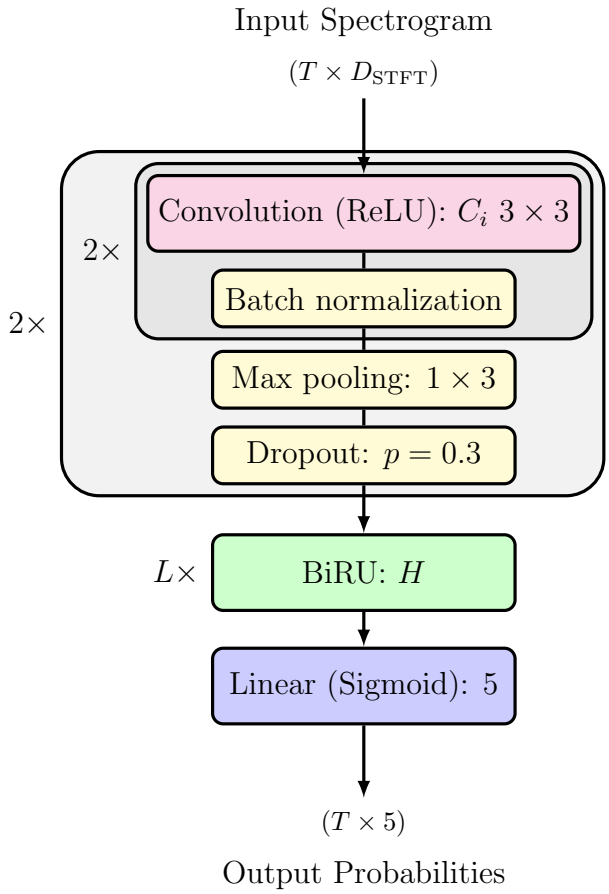

Figure 3.5: Convolutional RNN architecture structure.

Hyperparameter		Values
L	Number of layers	{2, 3, 4, 5}
H	Hidden size	{72, 144, 288, 576}
BiRU	Bidirectional Recurrent Unit	{GRU, LSTM}

Table 3.3: The different hyperparameters and their respective values tuned to train the Convolutional Recurrent Neural Network.

3.4 Convolutional Transformer

Google's "Attention Is All You Need" [39] made headway in regards to sequence prediction. It introduced the *Attention* layer, making a model capable of learning to *attend* to different elements in a sequence, and learning the relationship between them. Models dropping the recurrent units in favour of attention blocks are called *transformers*.

As mentioned, the RNN displays difficulty in sustaining long range dependencies through its hidden state, and information further away tends to become attenuated. The attention layer solves this by allowing each element to individually attend to each other element in the sequence separately. Intuitively, it allows each element to *"intelligently"* pick and choose where it wants to look, and what elements it wants to be influenced by. This stands in contrast to the recurrent units, where each element has to learn and predict what information about itself other elements could find useful, and *"remembering"* that, adding it to the hidden state.

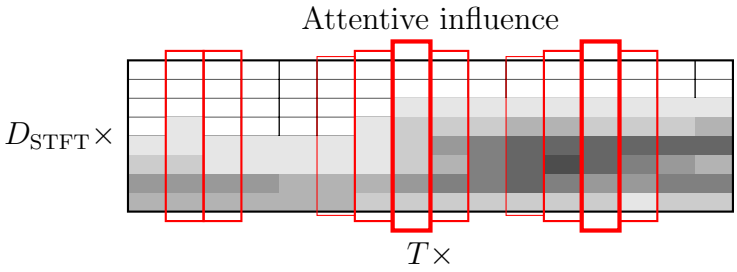

Figure 3.6: An example of how attention layers allow for attentive influence from from other timeframes spanning large distances. The background is a spectrogram, and the red boxes represent influence in the middle timeframe. The height of the box help visualize influence strength.

Recently, the attention layers have shown great success in AMT and ADT tasks, in some cases proving superior to the RNN [15, 10, 48].

Simply replacing the recurrent layers with attention blocks could allow our model to reap the reward by increasing its ability to understand sequences, while keeping the previously gotten gains from the convolutional layer. Both inside and outside of ADT, combining convolutional layers with transformers has seemed beneficial [48, 20].

3.4.1 Implementation

Similar to the CRNN, an initial fixed-size convolutional block with $I = 2$ is used, with an increasing number of kernels $C = \{32, 64\}$. Then, we project the resulting latent space into a separate, lower dimensional embedding space with dimension D_e, before combining it with a sinusoidal positional encoding.

Following this, the model contains L standard pre-layer normalization attention block, as these recently have been shown to be more stable during learning than the post-layer versions [45]. These attention blocks contain multi-head self-attention layers with H number of heads. Note that the first layer of the feedforward layer inside these attention blocks uses the Gaussian Error Linear Unit (GELU) activation function due to it being the standard within transformers and for its possible performance improvements over ReLU [13, 21]. Lastly, the linear layer outputs onset probabilities.

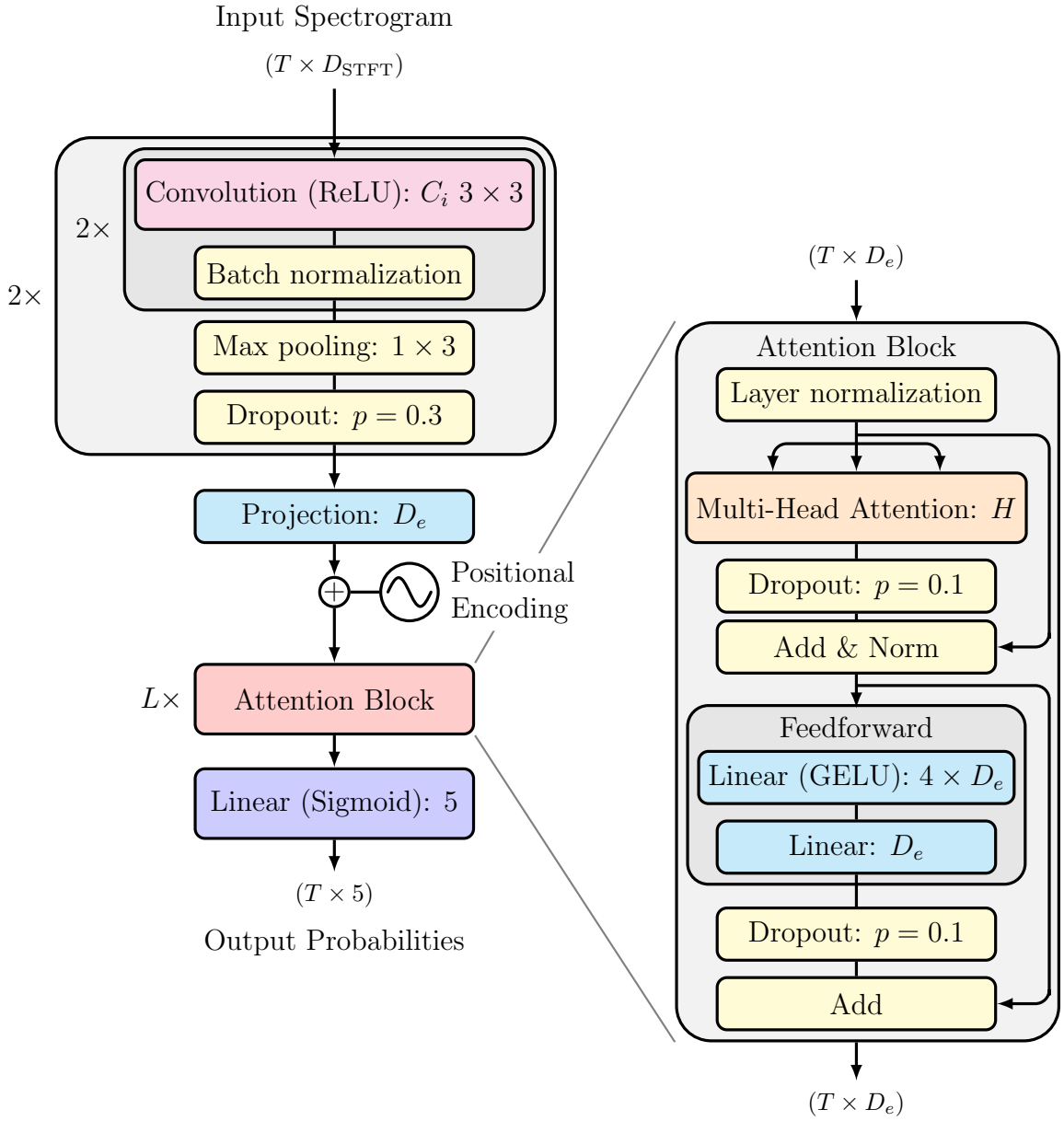

Figure 3.7: Convolutional Transformer architecture structure.

Hyperparameter		Values
H	Number of heads	{2, 4, 6, 8}
L	Number of layers	{2, 4, 6, 8, 10}
D_e	Embedding dimension	{72, 144, 288, 576}

Table 3.4: The different hyperparameters and their respective values tuned to train the Convolutional Transformer.

3.5 Vision Transformer

Introducing the Transformer raises a question. Would it be possible for an architecture to be comprised of solely attention layers, removing convolutions and breaking this need for a mixed convolutional-transformer architecture. This question was answered by Google's "An Image Is Worth 16x16 Words" [14], introducing the Vision Transformer.

The Vision Transformer was created to tackle image recognition tasks, though it has been applied to audio classification, displaying great performance on both [14, 18]. However, application of the Vision Transformer on an ADT task is a novel approach.

It is worth to note that Vision Transformers have been shown to display excellent performance, however they usually need more significantly more data than other architectures to function optimally [14].

3.5.1 Patch Embedding

A key component of the Vision Transformer is the creation of a patch embedding. First, we split the input image into different non-overlapping patches, and flatten them. These patches are linearly projected into a latent space, and a positional encoding is added to retain positional information. This resulting sequence of patches is what is referred to as a patch embedding.

We usually say that patch embeddings eliminate the use of convolutions in the Vision Transformer. However, the actual implementation of splitting the image into patches and linearly projecting each patch are usually implemented with a single 2D convolutional layer. Note that it is a linear projection, meaning the convolutional layer is absent of an activation function.

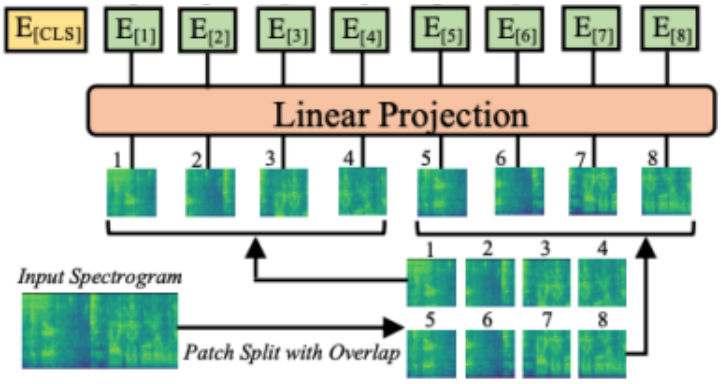

Figure 3.8: The creation of a patch embedding from an input spectrogram.

3.5.2 Architecture Modifications

Originally, the Vision Transformer has been used on classification tasks where the output is not a sequence. ADT is a sequence labeling task, and requires that the output is a sequence, where the time dimension matches the size of the input.

This can be solved by treating a group of patches from the patch embedding together as a timeframe, as long as we ensure that there the number of timeframes are a factor of the number of patches. Then the output of the Vision Transformer could be construed to match our intended output sequence.

3.5.3 Implementation

Initally, we transform the input spectrogram into a $(T \times D_e)$ patch embedding. The Convolutional layer splits the spectrogram into $(T \times D_e/P \times P)$ different patches. Added to these are a $(D_e/P \times P)$ learnable positional embedding, providing positional information to each patch. These are then permuted and flattened, transforming them into our final patch embedding.

Afterwards, we combine it with a sinusoidal positional encoding, and successively apply L attention blocks with number of heads H, identical in structure with those used in the Convolutional Transformer. Lastly, the linear layer outputs onset probabilities.

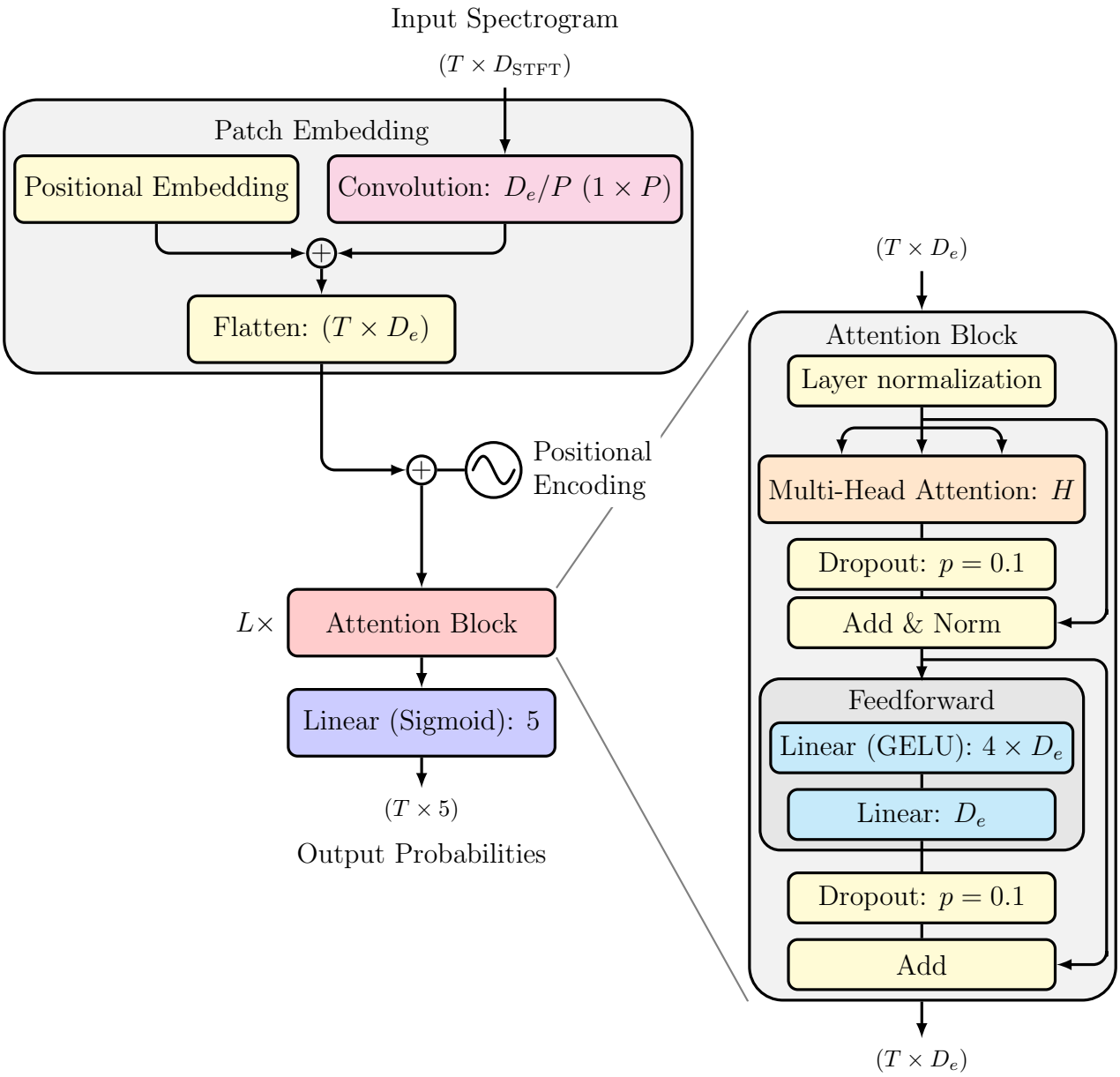

Figure 3.9: Vision Transformer architecture structure.

Hyperparameter		Values
P	Patch height	{7, 14, 21}
H	Number of heads	{2, 4, 6, 8}
L	Number of layers	{2, 4, 6, 8, 10}
D_e	Embedding dimension	{72, 144, 288, 576}

Table 3.5: The different hyperparameters and their respective values tuned to train the Vision Transformer.

Chapter 4

Datasets

4.1 ENST+MDB

The ENST-Drums dataset by Gillet and Richard [16] has been one of the most commonly used ADT datasets [44]. It features thorougly annotated drum samples by three drummers over different musical genres. Most of the tracks contain drum-only recordings, except the *minus-one* subset, which is played together with a music accompaniement. As this thesis attends to DTM tasks, we isolate our focus to this subset of tracks.

As this dataset contains separate audio files for performance and accompaniement, we additively combine them to create a singular respective mixture track. There exist many recordings from differently placed microphones of a single performance, however in this thesis, we solely selected the *"wet mix"* due to it being a combined recording of all other microphone recordings, and its *mix* showing resembelance of a polished performance track.

ENST-Drums contains 1.02 hours of music over 64 tracks.

Another well-known MedleyDB Drums dataset, from Southall et al. [36]. This dataset is built on top of Bittner et al.'s MedleyDB (MDB) dataset [4], but re-annotated and specialized for ADT related tasks. This dataset is, similar to ENST-Drums, also split into different stem tracks, such as isolated drum recordings and accompaniement. However, they also contain already-mixed *full mix* tracks, which are the ones we use in this thesis.

MDB-Drums contains 0.35 hours of music over 23 tracks.

Both of these datasets distributes their audio in waveform files, and their annotations in text files. Annotations are formatted by onset time and instrument label. They are also relatively small, and contain thorough, real, annotated data. Due to these similarities, in this thesis they are combined together into a slightly larger ENST+MDB dataset.

In total, this dataset contains 1.37 hours of music over 87 tracks.

4.1.1 Splits

These two datasets do not have predefined train/validation/test splits, such that we decided to construct our own splits. From ENST-Drums, *drummer1* and *drummer2* make our training split. The remaining drummer, *drummer3* is split in half, each for validation and test respectively. From MDB-Drums we do not have different explicit drummers, but instead split on specific genres. The different splits are specified in Appendix A.

4.1.2 Mapping

Maybe (or maybe not) explain the mapping from this to 5-instrument mapping. Also put this under a Appendix.

4.2 E-GMD

The Expanded Groove MIDI Dataset (E-GMD) from Callender et al. [7] is a large ADT dataset consisting of audio recordings from human drum performances annotated in MIDI. It is an expansion of Gillick et al.'s Groove MIDI Dataset (GMD) [17].

GMD was created through recording human performances in MIDI format through a Roland TD-11 electronic drumkit. This dataset does not contain audio recordings, only MIDI files, such that it is not applicable for ADT. To expand this dataset to be used with ADT, Callender et al. re-recorded the MIDI sequences on a Roland TD-17 electronic drumkit in real-time on a Digital Audio Workstation (DAW). These re-recordings were done over a large amount of differing soundfonts (confusingly also called *drumkits* on electronic drumkits), synthesizing several differently sounding audio recordings from a single MIDI performance [17, 7].

In contrast to the other dataset's utilized in this thesis, this dataset is not created for a DTM task, but rather a DTD task. This is due to all recordings containing solo, drum-only performances. In addition, as this data is recorded from human performances in a semi-manual nature, there exist some errors from the recording process *"that resulted in unusable tracks."* [7]. The magnitude of these errors are not stated, however other authors propose that it might be as high as 20.5% of the tracks with varying amounts of discrepancies [23].

This dataset contains 444.5 hours of audio recordings, from 1,059 unique drum performances resampled to 45,537 MIDI sequences.

4.2.1 Mapping

Maybe (or maybe not) explain the mapping from this to 5-instrument mapping. Also put this under a Appendix.

4.3 Slakh

The Synthesized Lakh 2100 (Slakh) Dataset from Manilow et al. [29] is a synthesized version of a subset of the Lakh Dataset from Raffel [31], that subset being a 2100 randomly selected tracks from Lakh where the MIDI files contain at least a piano, bass, guitar, and drums. These MIDI files are rendered and mixed into combined audio files, stored together with their respective original MIDI performances.

As mentioned, this dataset contains more instruments than just the drumset, such that it is originally meant for a AMT task. However, due to each track being guaranteed to contain a drum performance, converting it to an ADT task is trivial, and is done by selectively only utilizing the MIDI files corresponding to the drumset as our labels.

The original Slakh dataset was found to have data leakage between the different splits, and it is therefore recommended for transcription tasks to use a smaller subset, Slakh2100-redux, where this issue has been solved. Therefore, this is the dataset used for this thesis. Needs citation, other than the github and zenodo?
`https://github.com/ethman/slakh-utils`, `https://zenodo.org/records/4599666`

This dataset contains 115 hours of audio recordings over 1709 different songs.

4.3.1 Mapping

Maybe (or maybe not) explain the mapping from this to 5-instrument mapping. Also put this under a Appendix.

4.4 ADTOF-YT

The Automatic Drums Transcription On Fire YouTube (ADTOF-YT) dataset from Zehren et al. [47] is a large ADT dataset containing crowdsourced data, hence the YouTube suffix. Due to the crowdsourced nature of this dataset, it is possible to utilize large amounts of non-synthetic, human data, but with the tradeoff in which we can not guarantee its quality.

Contrary to the other datasets, this one is distributed in prepackaged TensorFlow datasets for each split, with datapoints as pairs of logarithmically filtered log-spectograms and sequence of instrument onset probabilities. This dataset also comes with a vocabulary of 5, and is thus the least diverse dataset in this thesis.

In their original paper, "High-Quality and Reproducible Automatic Drum Transcription From Crowdsourced Data", they state that their spectrograms are stored as log-magnitude, log-frequency spectrograms [47]. This contradicts information from where they distribute their dataset, where they state that *"The data as mel-scale spectrograms."* [46]. Through manual verification and approximation of the original waveforms, we propose that this latter information is an error, and confirms that the spectrograms are indeed stored as logarithmically filtered spectrograms, following the same format used by Vogl et. al [41] and what is otherwise used in this thesis. Is this enough justification? Should I provide more information? If so, what exactly?

The dataset contains 245 hours of music over 2924 tracks.

4.4.1 Mapping

Due to this dataset being distributed in a preprocessed nature, no re-mapping has been done, and the dataset is used "as is".

4.5 SADTP

The SADTP (Small Automatic Drum Transcription Performance) dataset is a novel dataset introduced in this thesis. It is a small dataset comprised of 16 songs with corresponding MIDI transcriptions.

The *performance* name alludes to the transcription being recorded live while listening to the songs on playback, with only minor post-processing. The transcriptions were recorded on a Roland TD-11 electric drumset, recording the MIDI perfomance to Apple's Garageband DAW, and extracting them to separate MIDI files. This comes with a similarly to E-GMD, as this dataset also was recorded in a semi-manual nature, which opens the possibility for slight, human induced errors. The magnitude of such errors are however not known, but we speculate that it is small but not insignificant.

This dataset stands out, as it is the only one in this thesis not used for training. Its sole purpose is for cross-dataset evaluation, and to provide information on the generalization ability for models trained on data from other sources.

The dataset contains 1.08 hours of music, which can be split into 977 non-overlapping 4 second datapoints (includes zero padding certain pieces for even partitioning).

4.5.1 Mapping

Maybe (or maybe not) explain the mapping from this to 5-instrument mapping. Also put this under a Appendix.

4.6 Summary

Dataset	Duration (h)	Number of tracks	Vocabulary	Melodic	Synthetic
ENST+MDB	1.37	87	20/21	✓	×
E-GMD	444.5	45,537	7	×	✓
Slakh	115	1709	?	✓	✓
ADTOF-YT	245	2915	5	✓	×
SADTP	1.08	977	?	✓	×

Table 4.1: Comparison of each dataset's characteristics like total duration, number of tracks, and vocabulary size. Melodic datasets denote being a part of DTM, otherwise they are DTD. Synthetic datasets contain music which is synthesized digitally, with an often automatic generation [48].

Compute specific vocabulary for the missing datasets, labeled (?)

In Table 4.1 we have a comprehensive summary of the different aspect of each of the dataset, as well as presenting them in a comparable format. These dataset span a good range of different characteristics, which could prove beneficial for training generalizing ADT models.

Chapter 5

Methodology

Though we perform 2 different studies, with 2 different intentions, the dataset preparation and model selection pipeline remains the same.

5.1 Data Preparation

As mentioned, the different datasets are distributed in differening formats. A few transformations are done to unify them all as PyTorch datasets.

Due to the preprocessed nature of the ADTOF-YT dataset, the unification is trivial and simply denotes a transformation from a stored TensorFlow dataset to a PyTorch dataset. Otherwise it is kept "as is".

5.1.1 Audio Files

The others are however distributed in the Waveform Audio File Format (with the suffix *.wav*) or using the Free Lossless Audio Codec (with the suffix *.flac*). Both of these formats are loaded using the PyTorch library `Torchaudio` and converted to monophonic format through meaning over each side's waveform. If any datasets contain distinct drum and accompaniement audio (like e.g. ENST-Drums), these are additively mixed together.

After each track is loaded into a waveform, a zero-padding is added to the end of each sequence allowing for even partitioning into 4 second partitions. Then we turn the track

into a spectrogram with 2048 fft's, and a window length of 2048. By keeping the hop length equal to the sampling rate divided by 100, the resulting spectrogram's timesteps represent a 10ms window of the original waveform.

After this, a filterbank is computed by generating 12 normalized logarithmically spaced filters, centered at 440Hz, and bounded over the interval [20Hz, 20,000Hz]. Applying this filterbank as a simple matrix multiplication over the spectrogram results in a logarithmically filtered spectrogram with $D_{\text{STFT}} = 84$ number of frequency bins. Lastly, we turn it into a log-spectrogram by applying a $\log_{10}$ operation to each cell, following an addition of 1 (preventing $\lim_{x \to 0} \log_{10}(x) = -\infty$ situations).

5.1.2 Annotations

The annotations are either distributed in specific formats as text files (with the suffix *.txt*), or in MIDI files (with the suffix *.mid* or *.midi*), each one having a different transformation into a sequence of instrument onset probabilities.

The datasets declaring onsets in text files (ENST-Drums and MDB-Drums) follow a similar format, storing onsets on separate lines, each one containing the time in seconds for the onset, and its respective instrument ID, separated by a space or tab respectively. To convert this into the onset probability sequence, we convert the time into a timeframe index by turning the time into milliseconds, dividing by 10 to group into 10ms intervals, and rounding to the nearest integer. After this, we map the instrument ID into its respective class, and set that specific (timeframe, class) cell value to 1.

The data given in MIDI format, the annotations are parsed using the library `Partitura`, and loads the information into an array of MIDI events called *notes*. These *notes* contain information for each event, importantly time, pitch, and velocity. These events are very thorough, but strict instrument onsets can be isolated by restricting our view to notes with a non-zero velocity. Instruments are denoted by the event's pitch, and a mapping is done from each pitch to a respective class. The time is converted identically to the loading of the text annotations, turning them into timeframe indices. At last, we also here set each specific registered onset (timeframe, class) cell to 1.

In addition to this, we apply a *target widening* step, setting values in timeframes adjacent to an instrument onset with a lower weight, equal to 0.5. These additional neighbouring *soft labels* have shown to be benifical in countering sparsity in our labels following multiple works on beat transcription [25, 47].

5.1.3 Splitting and Storing

These spectrogram/onset sequence pairs are stored together in PyTorch's TensorDatasets, separated into each track's respective train/validation/test split, and stored into PyTorch pickle files (with the suffix *.pt*). By doing all this preprocessing in advance, minimal preprocessing has to be done during runtime, increasing the efficiency of training.

5.2 Preprocessing

Most of the preprocessing is done during the data preparation step, however there are some remaining. Most importantly, a given model computes the mean and standard deviation of its training dataset, and uses these parameters to standardize its input data before prediction during runtime.

Data normalization like this has been shown to increase the speed and stability of convergence during training and, in summary, producing models which better generalize to unseen data. Although the specific benefits depend on the normalization technique used, the general consensus is that normalization in itself is beneficial in machine learning, hence their ubiquitous use in state-of-the-art models [24].

Another preprocessing step motivated by ADT specific methods is the use of infrequency weights, which framewise weighs the loss based on the instrument onsets that are present at each frame. These weights are precomputed from the training dataset, and are, for each instrument, computed by what Cartwright and Bello call *"the inverse estimated entropy of their event activity distribution"* [8]. Although they apply this to account for sparsity in data along different tasks, Zehren et al. [47] applied it to give more weight to infrequent instruments.

These weights are computed by calculating the probability of an instrument i appearing $p_i = \frac{n_i}{T}$ as the total number of onsets n_i divided by the total number of timesteps T. With this probability we compute its inverse entropy, giving us our final weights $w_i = (-p_i \log p_i - (1 - p_i) \log 1 - p_i)^{-1}$. Note that our probability computation differs from the work of Cartwright and Bello, as we do not divide by the number of instruments [8].

Should I mention anything on how the entropy is symmetric over 0.5? Such that a probability over 0.5 would lower our weights again, but how that is not a problem in ADT due to instrument onset inherently being sparse?

5.3 Training

As mentioned, the ADT task could be thought of as a sequence labeling task, where we each timeframe could have several instrument onsets present, taking the form of a 0 if an instrument is not present, and a 1 if it is. A natural loss function for this, where each value is handled as a separate independent probability distribution is the binary cross-entropy loss. Due to the numerical instability which can appear by applying a sigmoid activation function to our logits before computing the loss, we instead output our logits directly and utilize PyTorch's `BCEWithLogitsLoss` loss function, as recommended in their documentation Do I need to cite this?. It increases the numerical stability by taking advantage of the log-sum-exp trick, increasing numerical precision by avoiding underflow or overflow problems followed by significantly small or large input values.

With the choice by what to use, Adam was considered. This is an optimizer so ubiquitously used within the field of deep learning that when questioned with what optimizer to generally use authors like Sebastian Ruder state that *"., Adam might be the best overall choice* [33]. However contrary to this, the Adam optimizer has been shown to contain some issues, like its coupling of the weight decay term inside its gradient-based updates. Due to this, the choice instead fell on AdamW, a modified Adam implementation decoupling weight decay in whole from the gradient-based updates, and displaying a better ability to generalize. [26, 6, 28]

It was observed during training that the magnitude of the loss values could vary greatly and often displayed a tendancy to explode. To counteract this observation, we clip the gradients with a maximum norm set to 2. This addition significantly lowered the observed chance of exploding gradients occuring.

Another addition which is frequenct in other ADT works is the use of a learning rate scheduler. A learning rate scheduler keeps track of recent validation loss values, and if the minimum loss achieved plateaus (meaning it stop decreasing) for a certain number of epochs, the learning rate gets reducing by a given factor. In this thesis we reduce the learning rate by a factor of 5 if we observe 5 epochs of plateauing, with no improvement to our minimal validation loss [10, 47]. We also keep track of the general count of epochs since validation loss last improvement and perform an early stop if we ever observe 15 epochs without improvement.

5.4 Postprocessing

As mentioned, the model outputs a sequence of activation values, a 2 dimensional matrix with values on the interval $(0, 1)$ interpreted as the model's confidence in an instrument onset being present per frame. This can be utilized directly when computing our loss during training, however it is a difficult format to work with when talking about general performance, due to it rather representing a continuous confidence rather than discrete predictions. To suit this purpose, additional postprocessing is performed on the output.

First, we apply the aforementioned peak picking algorithm to isolate peaks in the model's onset confidence, intuitively being frames where the model is most confident in an instrument onset happening [5, 42]. Afterwards, we count a predicted onset if the given peak has a value larger or equal to 0.5. From this, it is trivial to compare predicted onsets with actual onsets, by greedily iterating our output sequence from the beginning, counting a prediction as a true positive if it happens within a 5 frame (50 ms) interval of a true onset, false positive if it happens outside such an interval, or false negative if a true onset happens with no prediction within said interval.

Could be useful showing transformation from output, to peak picked, to correct vs incorrect prediction.

These TP, FP, FN predictions are then added together and used to compute the previously mentioned micro F1-score. For additional analysis, we also compute and store instrumentwise F1-scores.

5.5 Model Selection

For model training and selection we utilize the RayTune library [27]. It simplifies the training of models by allowing us to input a training function, which metric to optimize for, hyperparameter spaces and search strategy, and additional configs. This simplifies our training, and through per-epoch reporting, it handles both checkpointing of model weights and best performing model selection for us.

Through RayTune, we train 15 different models (25 solely for the smallest dataset ENST+MDB), with hyperparameters tuned through bayesian optimization (see the next section 5.6). Each model is also ran for at most 100 epochs. As mentioned, we perform an

early stop if performance stops increasing. We utilize PyTorch's dataloaders for iterating the datasets, with a batch size of 128. And using the AdamW optimizer, as previously mentioned.

During each epoch of training, we evaluate the model on the training dataset's corresponding validation data. After training, the model with the highest Micro F1-score is selected. Due to the pre-split nature of most of our datasets, we utilize hold-out validation, with separate train, validation and test datasets. This is not only very common within ADT [40, 44, 10], but throughout the whole field of deep learning. Raschka mentions that *"The holdout method is inarguably the simplest model evaluation technique"* [32], and might be one of the reasons for its popularity. Then, we estimate its performance on unseen data through evaluating it on the test dataset. At last, said model is stored together with its corresponding weights, training config, and metrics.

5.6 Hyperparameter Tuning

5.6.1 Search Strategies

There are several hyperparameters that are tuned for each model. One of the most thorough ways to tune these would be through a grid-search regime. Then, one would train and evaluate each possible hyperparameter combination, to find the best performing. However, this grows out of proportion fast, with an exponentially growing number of combinations based on the number of hyperparameters and their values, making it computationally expensive. Another way would be to use a random-search regime, where one randomly picks the combination of hyperparamters. This regime sacrifices hyperparameter combination coverage by increasing computational efficiency. The drawback with this approach however is its reliance on probability. During a given training trial, we might be unlucky with all our randomly selected hyperparameter combinations, leading to a suboptimal performing model.

A combination of these two regimes, almost utilizing "the best of both worlds" would be to use a bayesian optimization regime. It chooses its hyperparameter combinations in a random fashion, like random-search, but also uses previously trained models' performance values to "intelligently" perform later choices, focusing in on promising combinations. In this way, we reap the rewards of the random-search regime's computational efficiency, while keeping some of the thoroughness from the grid-search regime.

We utilize RayTune's implementation of `OptunaSearch`, a hyperparameter searching regime which uses Optuna, an automatic hyperparameter optimization software framework based on bayesian optimization [2]. It has shown to be efficient in finding a good tradeoff between computation and performance, with authors like Shekhar et al. performing benchmarks on neural networks and showiung that *"The performance score of Optuna is the highest for all datasets"* [34].

How much do I need to cite here?

5.6.2 Hyperparameters

RayTune allows one to easily declare the search space for each of the hyperparameters. Every architecture-specific hyperparameter is chosen based on a random choice. In addition to this, we also tune the optimizer-specific learning rate and weight decay using a logarithmically uniform random sample. The specific spaces can be found below, in Table 5.1.

Hyperparameter	Search Space
Learning Rate	Log-uniform over $[1 \cdot 10^{-4}, 5 \cdot 10^{-3}]$
Weight Decay	Log-uniform over $[1 \cdot 10^{-6}, 1 \cdot 10^{-2}]$
Architecture-specific Hyperparameters	Random choice over each

Table 5.1: The search space for each of the different hyperparameters. The architecture-specific hyperparameters can be found under each architecture in the Architecture chapter 3

Chapter 6

Architecture Study

This study's main purpose is to figure out how suited each architecture is for Automatic Drum Transcription (ADT), more specifically Drum Transcription in the Presence of Melodic Instruments (DTM) tasks. This could help us figure out which architecture is superior for ADT, if there are any similarities between architectures who perform similarly well, or if there are any architectures who perform poorly.

6.1 Methodology

We perform hyperparameter tuning and model selection to train a separate model for each architecture over each dataset. At last we test the model on each dataset's respective test split. As a result, we are left with performance measures on unseen data from the same distribution as those each model was train on. This will give us a good intuition into each architecture's ability to learn the task of ADT and could help us estimate their generalization ability.

6.2 Results

Architecture	ENST+MDB	E-GMD	Slakh	ADTOF-YT
Recurrent Neural Network	0.6682	0.889	0.864	**0.9635**
Convolutional Neural Network	0.7797	0.8744	0.8318	0.844
Convolutional Recurrent Neural Network	**0.8132**	**0.8935**	**0.8959**	0.9333
Convolutional Transformer	0.776	0.8831	0.8826	0.9535
Vision Transformer	0.5426	0.8779	0.879	**0.9635**

Table 6.1: The Micro F1-score for each architecture, trained and tested on each dataset. The performances which are **bolded** represent the highest F1-score, and thus best performance, for that respective dataset.

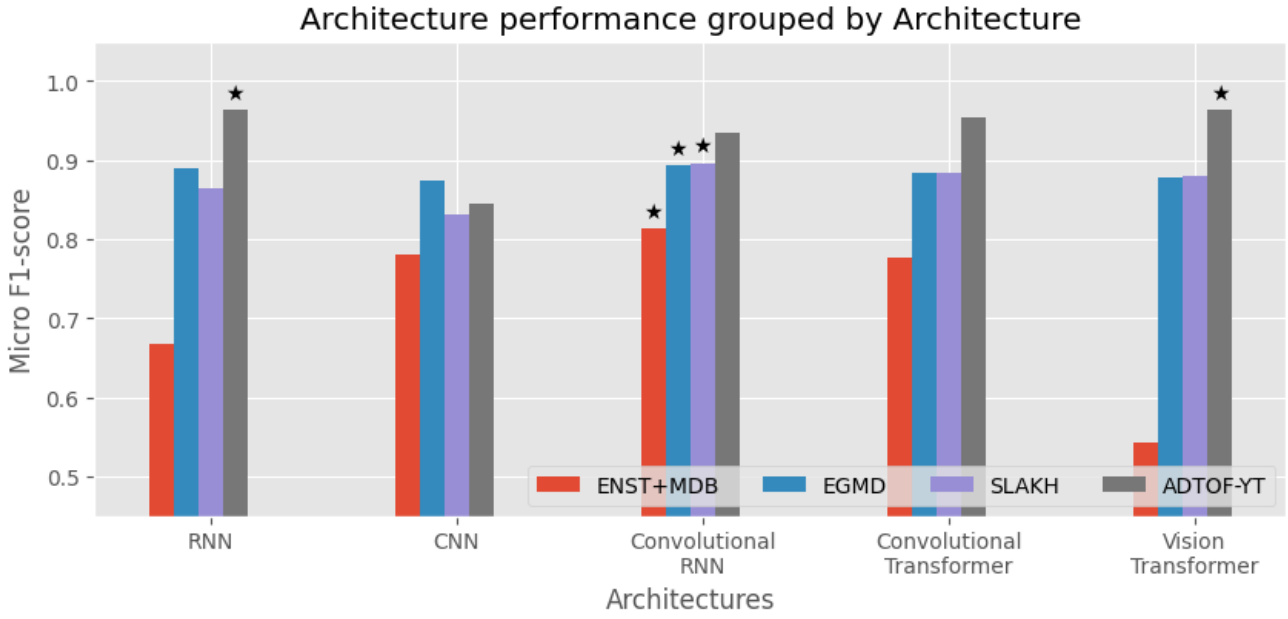

Figure 6.1: Comparison of Micro F1-scores for each dataset across the different architectures. Bars marked with a ($\star$) indicate the best performing architecture for each respective dataset.

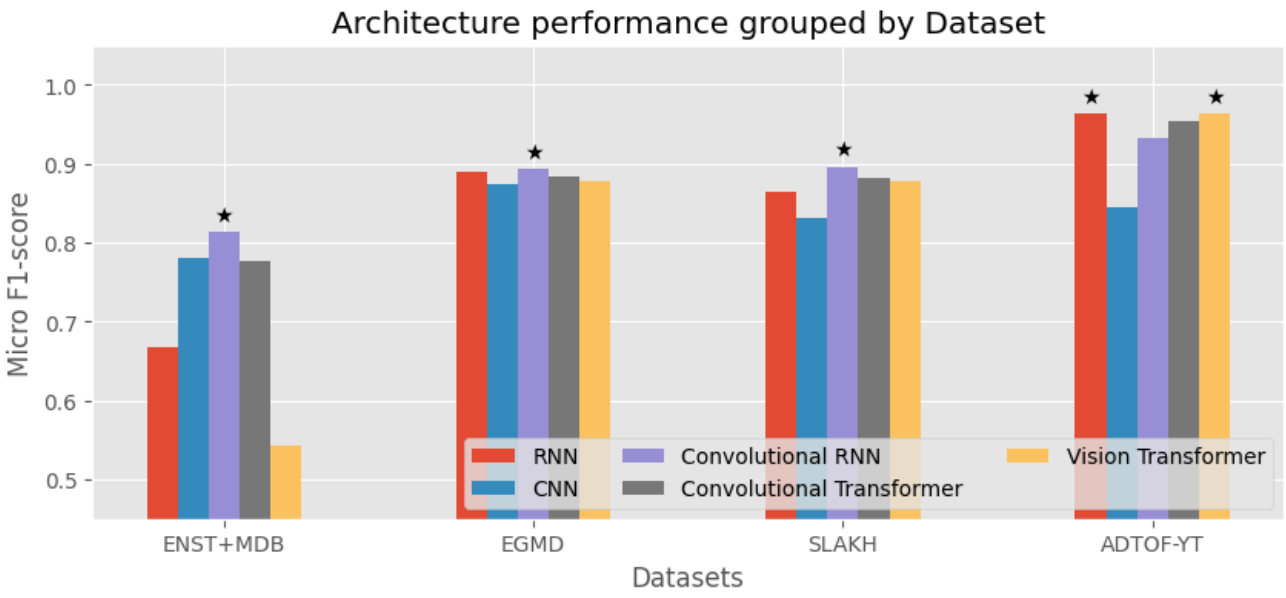

Figure 6.2: Comparison of Micro F1-scores for each architectures across the different dataset. Bars marked with a ($\star$) indicate the best performing architecture for each respective dataset.

6.3 Discussion

The results from the architecture study, summarized in Table 6.1 and Figures 6.1 and 6.2, indicate that there is no single superior architecture for ADT. Instead, performance is variant on the properties of each dataset, as well as the inherent inductive biases of each architecture.

Firstly, it is evident that there does not seem to be one superior architecture when it comes to ADT. There does not exist a single architecture outperforming the others across all the datasets, and the different architectures often share similarly high performances on most of the datasets.

However, the convolutional recurrent neural network demonstrates the highest Micro F1-score on three of the four datasets (namely ENST+MDB, E-GMD and Slakh). It also provides a high, but not exceptional F1-score on the fourth (ADTOF-YT). The consistency of its high performance across the different characteristics of each dataset suggests that it is able to handle a wide variety of ADT tasks, and that its performance may be high independent on the training dataset's size and complexity. In other words, it displays properties of being a architecture highly suitable for ADT tasks. One could speculate

that this suitability is provided by the strong inductive bias from the combination of convolutions and recurrent units, allowing for both initial spatial feature extraction as well as short-range temporal modelling.

The convolutional neural network displays a moderate performance across all datasets. It shows adequate performance on the smallest dataset (ENST+MDB), having the second highest F1-score. However, across all others (E-GMD, Slakh and ADTOF-YT) it displays the lowest. This relatively poor overall performance seems to indicate that the convolutional neural network currently is an inferior architecture for ADT tasks, unless dataset size is limited, in which it could provide a viable architecture. It also shows the importance of explicitly leveraging the temporal dependencies within ADT, as solely relying on the inductive bias of the convolutions didn't prove sufficient here.

This importance seems to be strengthened by the performance of the purely recurrent neural network, which has a surprisingly high performance across the datasets. In a way, it displays the opposite behaviour compared to the CNN, displaying inadequate performance for the smallest dataset (ENST+MDB), but a relatively high performance for two of the others (E-GMD and Slakh), sharing the crown for the highest F1-score on the last (ADTOF-YT). The first dataset's low performance suggests that the inductive bias of the recurrent layers are "weaker" than the ones from convolutions, relying on a larger amount of data to accurately learn the task. Notably, it significantly outperforms the CRNN on the last dataset, hinting that the inductive bias of the convolutions might "overpower" the model, proving so strong that it ends up hindering its performance.

The convolutional transformers exhibits performance comparable to something in between the RNN and CRNN, and shows what can be described as the most consistent relatively high performance across all the four datasets. It puts itself just under the CRNN for the first three datasets (ENST+MDB, E-GMD and Slakh), and a good step over for the last (ADTOF-YT). This does seem to indicate that the transformer models provide enough of a temporal inductive bias to accurately perform well on ADT tasks, however it might also hint that the long range dependencies are more easily modelled with attention layers are not as important as the accurate short range aggregation from the recurrent layers on most ADT datasets. Although by the performance displayed, the convolutional transformer could be chosen as a viable architecture for ADT.

At last we have the vision transformer, which performs similar to the RNN. For the smallest dataset (ENST+MDB) it displays a subpar performance with by far the lowest F1-score. For the two middle ones (E-GMD and Slakh) it exhibits a relatively average

performance. For the last dataset (ADTOF-YT) it shows excellent performance, having an identical F1-score to the RNN. This seems to agree with the previous conclusion that the inductive bias of the convolutional layers are "strong", putting a higher responsibility on a large amount of quality data when omitting it. The significantly poor performance on the small dataset might also indicate that the attention layers and transformer blocks come with an even "weaker" inductive bias than the recurrent layers, hightening the reliance on data amount. This aligns well with existing literature, where vision transformers generally require extensive training data or pre-training on larger datasets to obtain optimal performance [14].

In summary, our results strongly indicate that the Convolutional Recurrent Neural Network is the current most suited architecture for ADT tasks across different datasets. However, further research should be done into both the Vision Transformer and Recurrent Neural Network, gauging how these models scale with larger datasets.

Is it relevant to talk about future research here or should it be in the final conclusion?

Chapter 7

Dataset Study

This study's main purpose is to figure out how each dataset, and their characteristics, combined can influence a model's performance, both on- and OOD. This could help us figure out how to utilize and combine current datasets, or how to intelligently construct future datasets, to optimize models' performance and generalization ability for Automatic Drum Transcription (ADT) and Drum Transcription in the Presence of Melodic Instruments (DTM).

7.1 Methodology

We train convolutional recurrent neural networks, the best performing architecture from the first study, over several different combinations of the first four datasets ENST+MDB, E-GMD, Slakh, and ADTOF-YT. Datasets are combined as a union, where one epoch over the combined dataset would be equivalent to one epoch of both datasets separately. Each model is also evaluated on the remaining datasets, in addition to the SADTP dataset.

To get a comprehensive overview over how the different datasets could complement or contrast eachother's performances, without doing redundant experiments with diminishing returns, we select and train on a subset of all possible dataset combination. More specifically, we select 10 different combinations of datasets in such a way that we could extract valuable information about how their characteristics affect the resulting model's generalization ability.

7.2 Results

Training Dataset	ENST+MDB	E-GMD	Slakh	ADTOF-YT	SADTP
ENST+MDB	0.8132	0.5328	0.5294	0.5983	0.4165
E-GMD	0.4208	**0.8935**	0.3532	0.3011	0.1833
SLAKH	0.7987	0.6676	0.8959	0.5922	0.4769
ADTOF-YT	0.8403	0.6253	0.6546	0.9333	0.607
ENST+MDB + SLAKH	0.8394	0.6669	0.8983	0.6324	0.475
ENST+MDB + ADTOF-YT	0.863	0.6368	0.6342	0.9404	**0.6189**
SLAKH + ADTOF-YT	0.8552	0.6493	**0.9008**	**0.965**	0.6178
ENST+MDB + SLAKH + ADTOF-YT	**0.8772**	0.6768	0.8942	0.9502	0.6122
E-GMD + SLAKH + ADTOF-YT	0.8576	0.8911	0.897	0.9401	0.6165
ENST+MDB + E-GMD + SLAKH + ADTOF-YT	0.8646	0.8919	0.8932	0.9335	0.6169

Table 7.1: The Micro F1-score for a convolutional recurrent neural network trained over different combination of datasets, and tested on all datasets. The performances which are **bolded** represent the highest overall F1-score for the given dataset. Cells which are coloured light blue represent OOD evaluations. Values which are underlined represent the highest possible F1-score of OOD evaluations for the respective dataset.

Model	ENST+MDB	Slakh	ADTOF-YT
SLAKH + ADTOF-YT	0.8552	**0.9008**	**0.965**
ENST+MDB + SLAKH + ADTOF-YT	**0.8772**	0.8942	0.9502
ADTOF-RGW + ADTOF-YT [47]	0.78/0.81*	-	0.85
MT3 (mixture) [15]	-	0.76	-
YPTF.MoE+M [10]	0.8727**/-	0.8456	-

Table 7.2: The Micro F1-score of selectively chosen convolutional recurrent neural networks from this dataset study, compared with best performing models from other literature over the three of the test sets. The performances which are **bolded** represent the highest overall F1-score for the given dataset for all compared models. (*) Zehren et. al tests on ENST-Drums and MDB-Drums separately and tests on different test splits than ours [47]. (**) Chang et. al only tests ENST-Drums and tests on a different test split than ours [10].

7.3 Discussion

The results from the dataset study, summarized in Table 7.1 highlight important insights regarding how dataset composition could affect and impact model generalization. These results clearly demonstrate that strategically combining ADT datasets improve both on- and Out-Of-Distribution (OOD) generalization.

Firstly, it is evident that each dataset exhibits great performance on test splits which distribution overlaps with their own training dataset, but a lesser performance when testing OOD. There are several hypotheses one could propose for why this is the case, however one plausible one is that the datasets are characteristically varied enough in such a way that they each contrast eachother. This could give rise to this generalization penalty we see and explains the drop in performance. This generalization penalty also seem to be very uniform for different datasets, except one (E-GMD) which is talked about in the following paragraph. The best performing OOD evaluations all lie in the Micro F1-score range of 0.6 to 0.7, except for the smallest dataset (ENST+MDB) which displays an F1-score of around 0.87. Interestingly, the best performing OOD evaluation for ENST+MDB is significantly close to its best performance overall. As we know, in contrast to the other datasets ENST+MDB consist of high quality, really precise DTM data. A hypothesis is that this transcription quality of ENST+MDB more easily show the strength of the other model's, which might be hidden in the lesser quality of other test splits.

Another important observation, as mentioned, is that of E-GMD. As known, this dataset differs from the others by being a Drum Transcription of Drum-only Recordings (DTD) dataset, instead of a DTM one. Despite its large size, it is the worst at generalizing to the others, and it is the worst performing model on all other datasets. However, it is worth to note that it also displays the best performance on itself. This helps reinforcing the differences between the tasks of DTD and DTM, showing how model's trained on differently tasked datasets are not directly applicable interchangeably.

Building on this information we could inspect the opposite relationship. Notably, the other model's OOD evaluation on E-GMD is comparable to that of their OOD evaluations on themselves. This also helps strengthen the hypothesis that these tasks, DTD and DTM, display a complexity hierarchy where one seems to build upon the other. In other words, DTM might be a more difficult transcription task than DTD, but in return, a model trained for DTM might be sufficient for DTD tasks, and exhibits a zero-shot generalization ability which does not hold for the opposite.

One results that might merit discussion is the correlation between high ADTOF-YT performance and high SADTP performance. A high F1-score in the ADTOF-YT test split is accompanied by a relatively high OOD F1-score for SADTP. Due to the crowdsourced nature of ADTOF-YT, and the SADTP consisting of contemporary and public music tracks, there might exist some overlap in their distribution which could explain this behaviour.Maybe show correlation graph?.

An insightful observation is that the best performance of all DTM datasets from the architecture study, sees themselves beat by models trained on a combination of datasets. This is an important observation and shows that model performance often is enhanced by expanding the dataset size and variation. However, building upon what we discussed in the last paragraph, it is important to understand the differences between DTD and DTM datasets. As we observe, the best Micro F1-score for DTD dataset comes from training data solely of DTD data. The same holds the other way, in that the best performances on DTM datasets happen when solely trained on DTM data.

Secondly, take a look at the model trained on the largest amount of data, the combination of all datasets. This model exhibits superior performance across the board, however interestingly it does not achieve the highest Micro F1-score on any, though being very close. This strengthens the hypothesis that expanding the amount data will heighten generalization ability for a given model, and agrees with the current consensus that training dataset size is vital to achieving an optimally generalizing model [47, 25]. For future works it would be interesting to analyze how valuable data augmentation would be for ADT, and if it affects generalization positively.

Lastly, in Table 7.2 we compare our best performing model's with other literature and giving remarkable insight. Note that not much literature exist with models evaluated on thsee datasets. Gardner et. al's MT3 [15] and Chang et. al's YPTF.MOE+M [10] are general AMT models, predicting several different instruments in addition to drums. They also present using the Offset F1-score terminology, equivalent to our Micro F1-score (also used by Zehren et. al [47]). Zehren et. al's ADTOF-RGW + ADTOF-YT is a DTM model, similar to that of ours. Observably, we note that our model's outperform the others' by a singificant amount on both the Slakh and ADTOF-YT datasets. Due to the splits and datasets being a bit different for ENST+MDB, a direct comparison is a bit more difficult, but one could reason that our model's performance is comparable to that of YPTF.MOE+M [10]. By comparing with results from other literature, we can put our model's performance into perspective and strengthens the argument that our combinations could in fact provide valuable advancements for the state-of-the-art in ADT and DTM.

In summary we can conclude that our results strongly indicate and strengthens our hypothesis that combining different datasets with variable characteristics, will help a model perform better on ADT tasks, and could make them generalize better towards Out-Of-Distribution (OOD) datasets. We also note that applicability of different ADT tasks follow a relationship, where DTM datasets generalize better on DTD datasets than the inverse.

I'm unsure if I've discussed enough (or too much of certain aspects) around the results presented in 7.1?

Chapter 8

Conclusion

This thesis set out to investigate how to achieve optimal performance on Automatic Drum Transcription (ADT) tasks, specifically Drum Transcription in the Presence of Melodic Instruments (DTM). Through exploring and analysing how different deep learning architectures and different dataset compositions affect model performance.

In the first study we trained and compared the performance of five deep learning architectures, namely a recurrent neural network, convolutional neural network, convolutional recurrent neural network, convolutional transformer, and vision transformer, each trained and tested over different dataset. The best architecture showing the strongest overall performance was the convolutional recurrent neural network, which performed the best performance on three of the four datasets. However, we also note that more experimentation should be done with the recurrent neural network, and the vision transformer, specifically on larger dataset, which both identically achieved the best performance on the last dataset.

In the second study we explored how combining datasets of different characteristics influences a model's generalization ability. We found out that combining datasets substantially improves both on- and Out-Of-Distribution (OOD) performance. However, we also identified that carefully constructing these datasets is vital, but that an increase in data amount generally leads to better performance.

This thesis contributes not only by comparing different architectural choices for DTM tasks, but also through evaluating different ADT datasets' affection on model generalization, this on both public datasets, and on a novel dataset composed and transcribed

for this thesis, SADTP. These contributions could give valuable insight into ADT by emphasizing both choices on deep learning architectures as well as construction of datasets.

Future work should expand upon this research by performing rigorous architecture analyses over larger datasets, as well as optionally inspecting how the different number of drum instrument classes both affect and perform over these different datasets.

Lastly, this thesis demonstrates that deep learning model's can exhibit great performance on ADT and DTM tasks achievable through intelligent choices in both architecture and dataset.

List of Acronyms and Abbreviations

ADT Automatic Drum Transcription.

AMT Automatic Music Transcription.

BiRU Bidirectional Recurrent Unit.

CC Crash Cymbal.

CNN Convolutional Neural Network.

CRNN Convolutional Recurrent Neural Network.

DAW Digital Audio Workstation.

DFT Discrete Fourier Tranform.

DNN Deep Neural Network.

DSC Drum Sound Classification.

DTD Drum Transcription of Drum-only Recordings.

DTM Drum Transcription in the Presence of Melodic Instruments.

DTP Drum Transcription in the Presence of Additional Percussion.

FFT Fast Fourier Transform.

FN False Negatives.

FP False Positives.

GELU Gaussian Error Linear Unit.

GRU Gated Recurrent Unit.

HH Hi-Hat.

HT High Tom.

KD Kick Drum.

LLM Large Language Model.

LSTM Long Short-Term Memory.

LT Low Tom.

MIDI Musical Instrument Digital Interface.

MIR Music Information Retrieval.

MT Mid Tom.

NLP Natural Language Processing.

OOD Out-Of-Distribution.

RC Ride Cymbal.

ReLU Rectified Linear Unit.

RNN Recurrent Neural Network.

SD Snare Drum.

STFT Short-time Fourier Transform.

TN True Negatives.

TP True Positives.

TT Tom-Tom.

Bibliography

[1] *Fundamentals of Telephony*. United States, Department of the Army, 1953.
URL: https://books.google.no/books?id=8nvJ6qvtdPUC.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[3] Pras Amandine and Guastavino Catherine. Sampling rate discrimination: 44.1 khz vs. 88.2 khz. *Journal of the Audio Engineering Society*, (8101), may 2010.

[4] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *Ismir*, volume 14, pages 155–160, 2014.

[5] Sebastian Böck, Florian Krebs, and Markus Schedl. Evaluating the online capabilities of onset detection methods. In *International Society for Music Information Retrieval Conference*, 2012.
URL: https://api.semanticscholar.org/CorpusID:7379180.

[6] Sebastian Bock, Josef Goppold, and Martin Weiß. An improvement of the convergence proof of the adam-optimizer, 2018.
URL: https://arxiv.org/abs/1804.10587.

[7] Lee Callender, Curtis Hawthorne, and Jesse Engel. Improving perceptual quality of drum transcription with the expanded groove midi dataset, 2020.
URL: https://arxiv.org/abs/2004.00188.

[8] Mark Cartwright and Juan Pablo Bello. Increasing drum transcription vocabulary using data synthesis. In *Proc. International Conference on Digital Audio Effects (DAFx)*, pages 72–79, 2018.

[9] Pragnan Chakravorty. What is a signal? [lecture notes]. *IEEE Signal Processing Magazine*, 35(5):175–177, 2018. doi: 10.1109/MSP.2018.2832195.

[10] Sungkyun Chang, Emmanouil Benetos, Holger Kirchhoff, and Simon Dixon. Yourmt3+: Multi-instrument music transcription with enhanced transformer architectures and cross-dataset stem augmentation. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2024.

[11] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL, 2014. doi: 10.3115/V1/D14-1179.
URL: `https://doi.org/10.3115/v1/d14-1179`.

[12] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965. ISSN 00255718, 10886842.
URL: `http://www.jstor.org/stable/2003354`.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.
URL: `https://aclanthology.org/N19-1423/`.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
URL: `https://arxiv.org/abs/2010.11929`.

[15] Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse Engel. Mt3:

Multi-task multitrack music transcription, 2022.
URL: https://arxiv.org/abs/2111.03017.

[16] Olivier Gillet and Gaël Richard. Enst-drums: an extensive audio-visual database for drum signals processing. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2006.

[17] Jon Gillick, Adam Roberts, Jesse Engel, Douglas Eck, and David Bamman. Learning to groove with inverse sequence transformations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2269–2279. PMLR, 09–15 Jun 2019.
URL: https://proceedings.mlr.press/v97/gillick19a.html.

[18] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer, 2021.
URL: https://arxiv.org/abs/2104.01778.

[19] D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984. doi: 10.1109/TASSP.1984.1164317.

[20] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition, 2020.
URL: https://arxiv.org/abs/2005.08100.

[21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023.
URL: https://arxiv.org/abs/1606.08415.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.
URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[23] Thomas Holz. Automatic drum transcription with deep neural networks. Master's thesis, Technische Universitaet Berlin (Germany), 2021.

[24] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Normalization techniques in training dnns: Methodology, analysis and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10173–10196, 2023. doi: 10.1109/TPAMI.2023.3250241.

[25] Yun-Ning Hung, Ju-Chiang Wang, Xuchen Song, Wei-Tsung Lu, and Minz Won. Modeling beats and downbeats with a time-frequency transformer. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 401–405, 2022. doi: 10.1109/ICASSP43922.2022.9747048.

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
URL: `https://arxiv.org/abs/1412.6980`.

[27] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training, 2018.
URL: `https://arxiv.org/abs/1807.05118`.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.
URL: `https://arxiv.org/abs/1711.05101`.

[29] Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 45–49, 2019. doi: 10.1109/WASPAA.2019.8937170.

[30] Geoff Nicholls. *The Drum Handbook: Buying, maintaining, and getting the best from your drum kit*. San Francisco, CA: Backbeat Books, 2003.

[31] Colin Raffel. *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. Columbia University, 2016.

[32] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning, 2020.
URL: `https://arxiv.org/abs/1811.12808`.

[33] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2017.
URL: `https://arxiv.org/abs/1609.04747`.

[34] Shashank Shekhar, Adesh Bansode, and Asif Salim. A comparative study of hyper-parameter optimization tools. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE, 2021.

[35] Carl Southall, Ryan Stables, and Jason Hockman. Automatic drum transcription using bi-directional recurrent neural networks. In *International Society for Music*

Information Retrieval Conference, 2016.
URL: https://api.semanticscholar.org/CorpusID:2891003.

[36] Carl Southall, Chih-Wei Wu, Alexander Lerch, and Jason Hockman. Mdb drums: An annotated subset of medleydb for automatic drum transcription. 2017.

[37] Gilbert Strang. Wavelet transforms versus fourier transforms. *Bulletin of the American Mathematical Society*, 28(2):288–305, 1993.

[38] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.
URL: https://arxiv.org/abs/1609.03499.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[40] Richard Vogl, Matthias Dorfer, and Peter Knees. Recurrent neural networks for drum transcription. In *ISMIR*, pages 730–736, 2016.

[41] Richard Vogl, Matthias Dorfer, Gerhard Widmer, and Peter Knees. Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks. In *International Society for Music Information Retrieval Conference*, 2017.
URL: https://api.semanticscholar.org/CorpusID:21314796.

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

[43] Friedrich Wolf-Monheim. Spectral and rhythm features for audio classification with deep convolutional neural networks, 2024.
URL: https://arxiv.org/abs/2410.06927.

[44] 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.

[45] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR, 13–18 Jul 2020.

URL: https://proceedings.mlr.press/v119/xiong20b.html.

[46] Mickael Zehren, Marco Alunno, and Paolo Bientinesi. Adtof datasets, November 2023.

URL: https://doi.org/10.5281/zenodo.10084511.

[47] 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.

[48] Mickaël Zehren, Marco Alunno, and Paolo Bientinesi. Analyzing and reducing the synthetic-to-real transfer gap in music information retrieval: the task of automatic drum transcription, 2024.

URL: https://arxiv.org/abs/2407.19823.

Appendix A

ENST+MDB Splits

Split	Drummer	Track
Train	drummer_1	107_minus-one_salsa_sticks.wav
	drummer_1	108_minus-one_rock-60s_sticks.wav
	drummer_1	109_minus-one_metal_sticks.wav
	drummer_1	110_minus-one_musette_brushes.wav
	drummer_1	111_minus-one_funky_rods.wav
	drummer_1	112_minus-one_funk_rods.wav
	drummer_1	113_minus-one_charleston_sticks.wav
	drummer_1	114_minus-one_celtic-rock_brushes.wav
	drummer_1	115_minus-one_bossa_brushes.wav
	drummer_1	121_MIDI-minus-one_bigband_brushes.wav
	drummer_1	123_MIDI-minus-one_blues-102_sticks.wav
	drummer_1	125_MIDI-minus-one_country-120_brushes.wav
	drummer_1	127_MIDI-minus-one_disco-108_sticks.wav
	drummer_1	129_MIDI-minus-one_funk-101_sticks.wav
	drummer_1	131_MIDI-minus-one_grunge_sticks.wav
	drummer_1	133_MIDI-minus-one_nu-soul_sticks.wav
	drummer_1	135_MIDI-minus-one_rock-113_sticks.wav
	drummer_1	137_MIDI-minus-one_rock'n'roll-188_sticks.wav
	drummer_1	139_MIDI-minus-one_soul-120-marvin-gaye_sticks.wav
	drummer_1	141_MIDI-minus-one_soul-98_sticks.wav
	drummer_1	143_MIDI-minus-one_fusion-125_sticks.wav
	drummer_2	115_minus-one_salsa_sticks.wav
	drummer_2	116_minus-one_rock-60s_sticks.wav
	drummer_2	117_minus-one_metal_sticks.wav
	drummer_2	118_minus-one_musette_brushes.wav
	drummer_2	119_minus-one_funky_sticks.wav
	drummer_2	120_minus-one_funk_sticks.wav
	drummer_2	121_minus-one_charleston_sticks.wav
	drummer_2	122_minus-one_celtic-rock_sticks.wav
	drummer_2	123_minus-one_celtic-rock-better-take_sticks.wav
	drummer_2	124_minus-one_bossa_sticks.wav
	drummer_2	130_MIDI-minus-one_bigband_sticks.wav
	drummer_2	132_MIDI-minus-one_blues-102_sticks.wav
	drummer_2	134_MIDI-minus-one_country-120_sticks.wav
	drummer_2	136_MIDI-minus-one_disco-108_sticks.wav
	drummer_2	138_MIDI-minus-one_funk-101_sticks.wav