# Transcribing Audio Recordings to the Guitar Fretboard Using the Transformer Encoder Architecture

Tyler Hattori

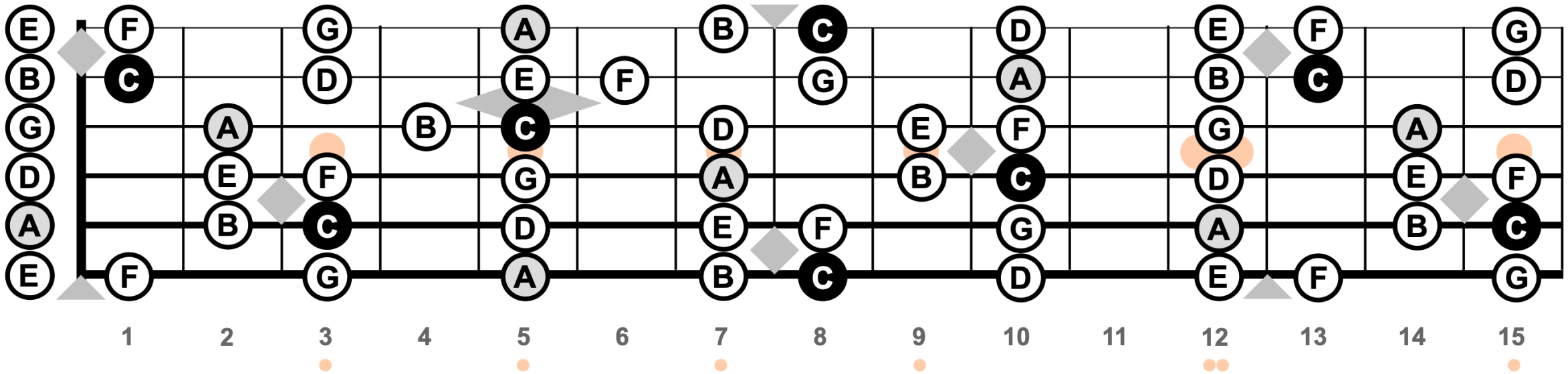
2024

# Outline

- Project overview
- Related work
- Model architecture
- Results and key takeaways
- Code demo

# Overview

The Guitar Fretboard [1]



# Overview

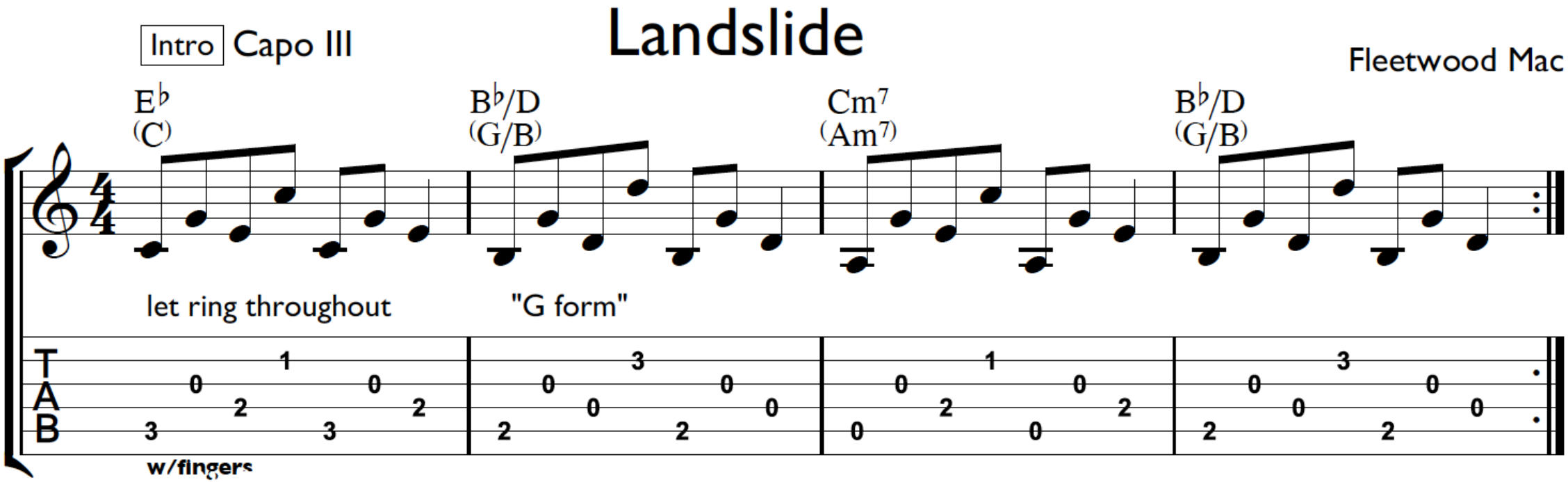
Fretboard Frequencies [2]

- Moving forward, I assume standard tuning
- Human hearing is about 20Hz to 20kHz
- Guitar frequency range is only 80Hz to 1kHz

| | String Number (Note) | | | | | |
|---|---|---|---|---|---|---|
| Fret | 1 (E) | 2 (A) | 3 (D) | 4 (G) | 5 (B) | 6 (Ė) |
| 1 | 82.41 | 110.00 | 146.83 | 196.00 | 246.94 | 329.63 |
| 2 | 87.31 | 116.54 | 155.56 | 207.65 | 261.63 | 349.23 |
| 3 | 92.50 | 123.47 | 164.81 | 220.00 | 277.18 | 369.99 |
| 4 | 98.00 | 130.81 | 174.61 | 233.08 | 293.66 | 392.00 |
| 5 | 103.83 | 138.59 | 185.00 | 246.94 | 311.13 | 415.30 |
| 6 | 110.00 | 146.83 | 196.00 | 261.63 | 329.63 | 440.00 |
| 7 | 116.54 | 155.56 | 207.65 | 277.18 | 349.23 | 466.16 |
| 8 | 123.47 | 164.81 | 220.00 | **293.66** | 369.99 | 493.88 |
| 9 | 130.81 | 174.61 | 233.08 | 311.13 | 392.00 | 523.25 |
| 10 | 138.59 | 185.00 | 246.94 | 329.63 | 415.30 | 554.37 |
| 11 | 146.83 | 196.00 | 261.63 | 349.23 | 440.00 | 587.33 |
| 12 | 155.56 | 207.65 | 277.18 | 369.99 | 466.16 | **622.25** |
| 13 | 164.81 | 220.00 | 293.66 | 392.00 | 493.88 | 659.26 |
| 14 | 174.61 | 233.08 | 311.13 | 415.30 | 523.25 | 698.46 |
| 15 | 185.00 | 246.94 | 329.63 | 440.00 | 554.37 | 739.99 |
| 16 | 196.00 | 261.63 | 349.23 | 466.16 | 587.33 | 783.99 |
| 17 | 207.65 | 277.18 | 369.99 | 493.88 | 622.25 | 830.61 |
| 18 | 220.00 | 293.66 | 392.00 | 523.25 | 659.26 | 880.00 |
| 19 | 233.08 | 311.13 | 415.30 | 554.37 | 698.46 | 932.33 |
| 20 | 246.94 | 329.63 | 440.00 | 587.33 | 739.99 | 987.77 |

# Overview

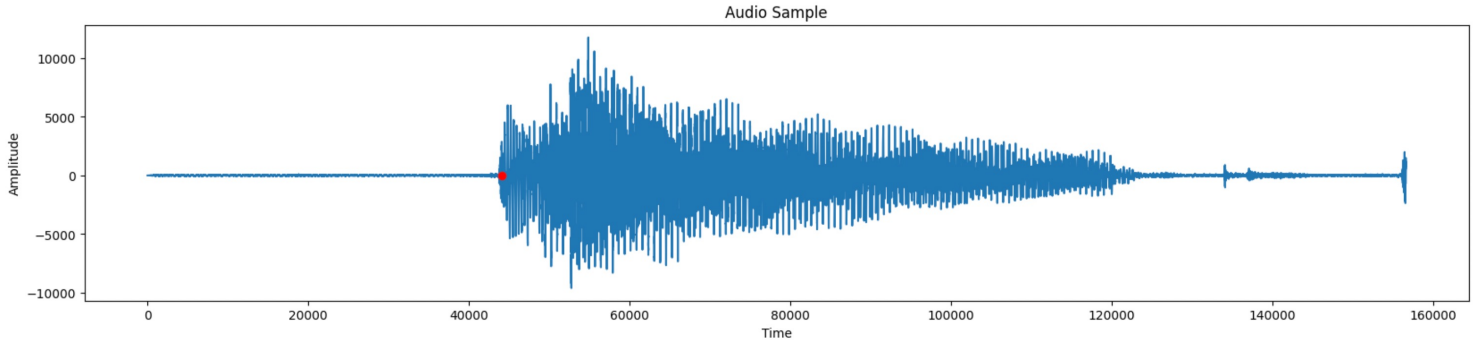
Guitar Tablature Example [3]



# Overview

## Music Signal Representations

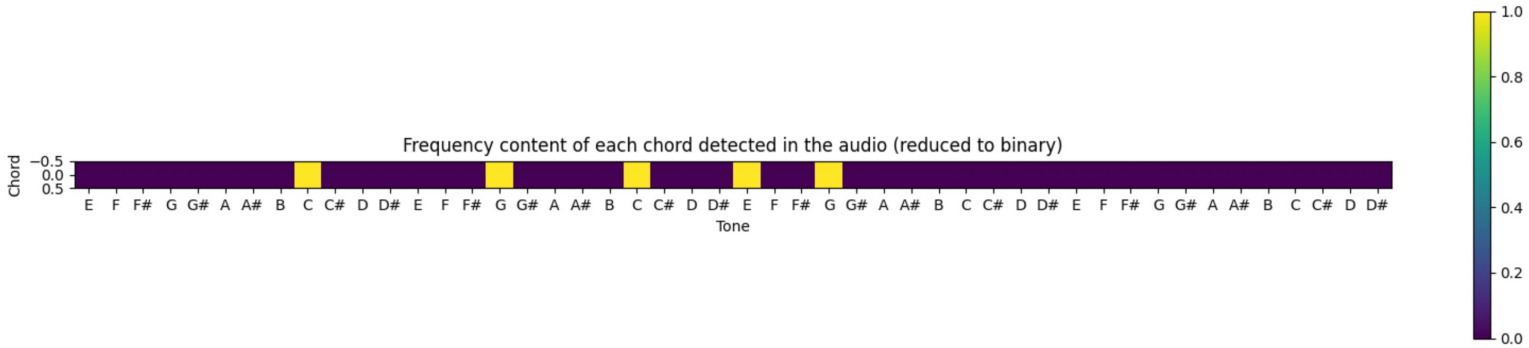
- Time signal
- Frequency Content
- Guitar tablature
- Pattern characteristics
- Key label




# Overview

## Music Signal Representations

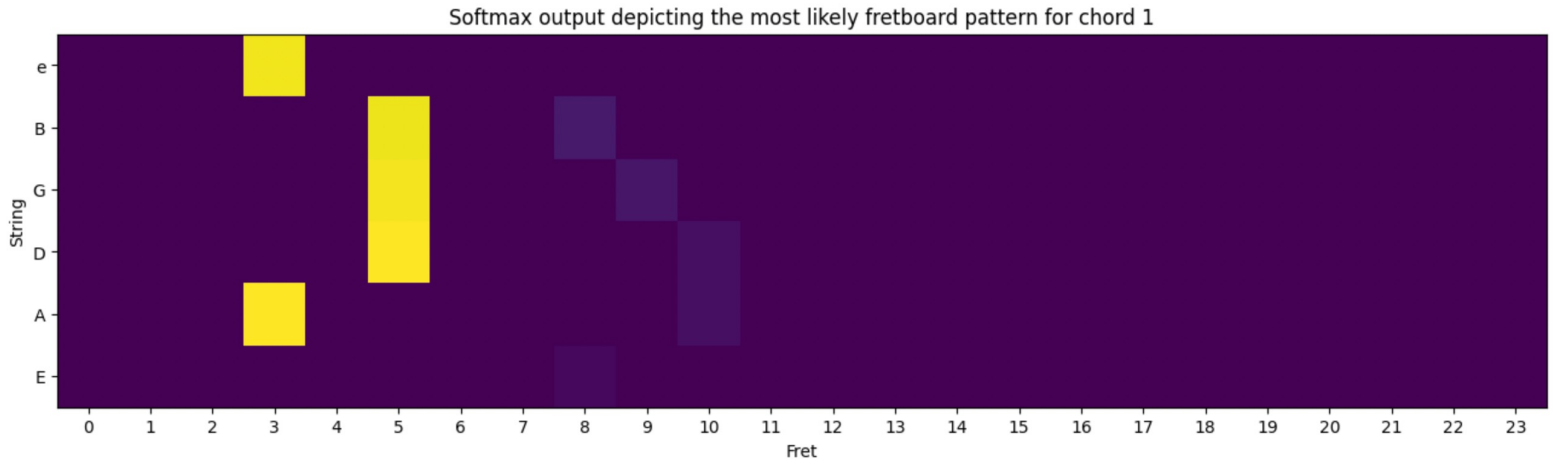
- Time signal
- Frequency Content
- Guitar tablature
- Pattern characteristics
- Key label



# Overview

## Music Signal Representations

- Time signal
- Frequency Content
- Guitar tablature
- Pattern characteristics
- Key label

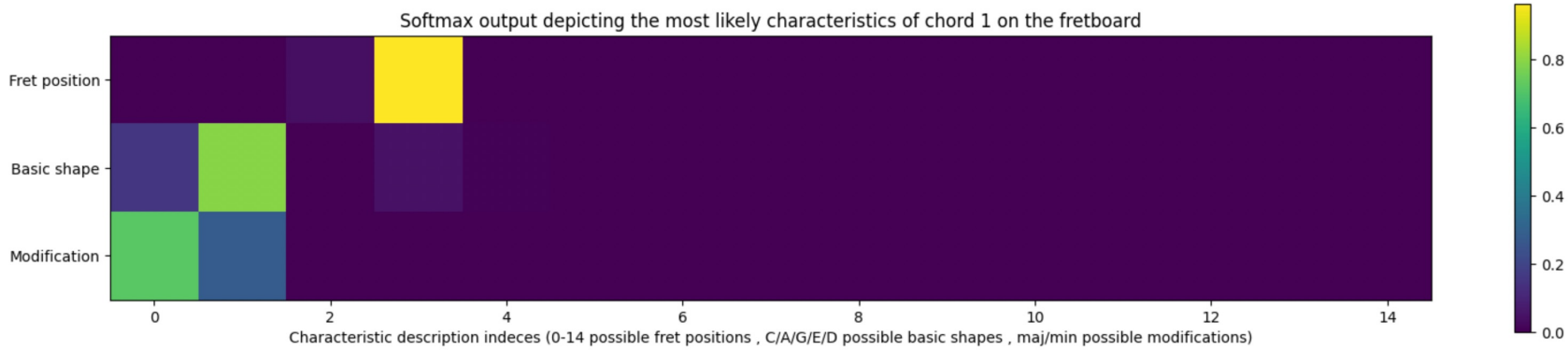



```
Guitar tablature:
 3
 5
 5
 5
 3
 –
```

# Overview

## Music Signal Representations

- Time signal
- Frequency Content
- Guitar tablature
- Pattern characteristics
- Key label

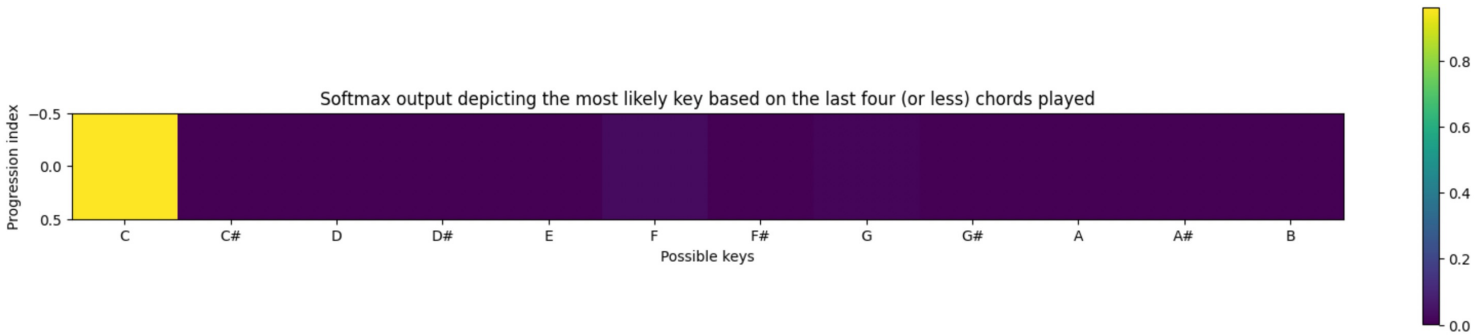



```
The model guesses that chord 1 has a bar at fret 3, using the A chord shape in its major form.

Chord 1 guess: C
```

# Overview

## Music Signal Representations

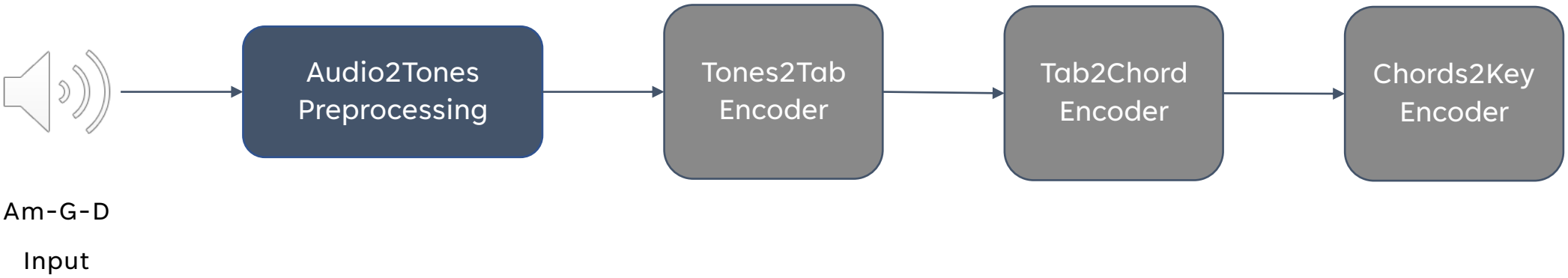
- Time signal
- Frequency Content
- Guitar tablature
- Pattern characteristics
- Key label



# Related Work

- Dong et al. MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. (2017).
- Vaswani et al. Attention Is All You Need. (2017).
- Dosovitski et al. An Image is Worth 16x16 Words: Transformers Image Recognition at Scale. (2021).
- Liu et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. (2021).
- Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs. (2021).

# My Model


Audio2Tones Preprocessing
Tones2Tab Encoder
Tab2Chord Encoder
Chords2Key Encoder
Am-G-D
Input

# My Model

## Audio2Tones Pre-processing

- $N$-sample windowing
- Tone extraction
- Chord onset detection
- Resize windows around chords
- Tone extraction
- Tokenize tone vectors

# My Model

Audio2Tones Pre-processing

- $N$-sample windowing
- Tone extraction
- Chord onset detection
- Resize windows around chords
- Tone extraction
- Tokenize tone vectors

$$H = \begin{bmatrix} W_0^0 & \cdots & W_0^{N-1} \\ \vdots & \ddots & \vdots \\ W_{47}^0 & \cdots & W_{47}^{N-1} \end{bmatrix}$$

$$W_k^n = e^{-j2\pi nF(k)/f_s}$$

$$F(k) = 440 * 2^{(k-29)/12}$$

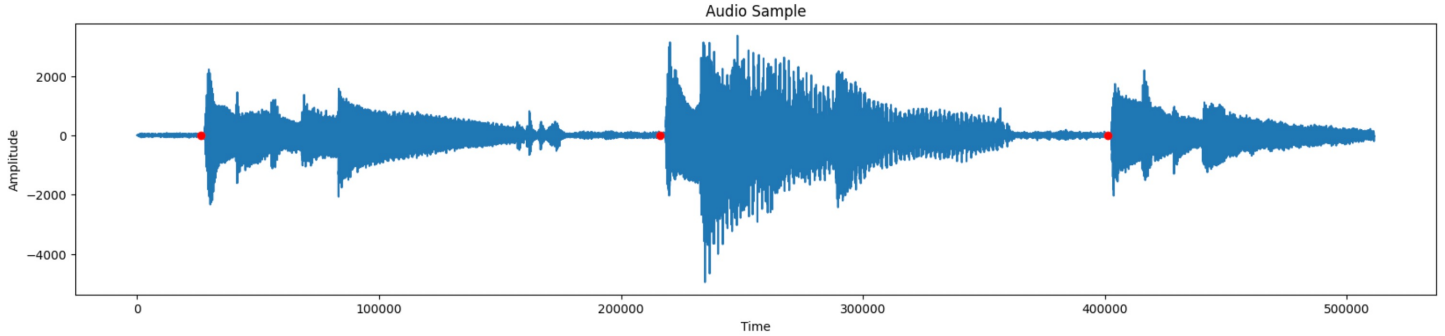
$$y_{tones} = Hx_{audio}$$

$$H \in \mathbb{R}^{48\times N}$$
$$y_{tones} \in \mathbb{R}^{48\times 1}$$
$$x_{audio} \in \mathbb{R}^{N\times 1}$$

# My Model

## Audio2Tones Pre-processing

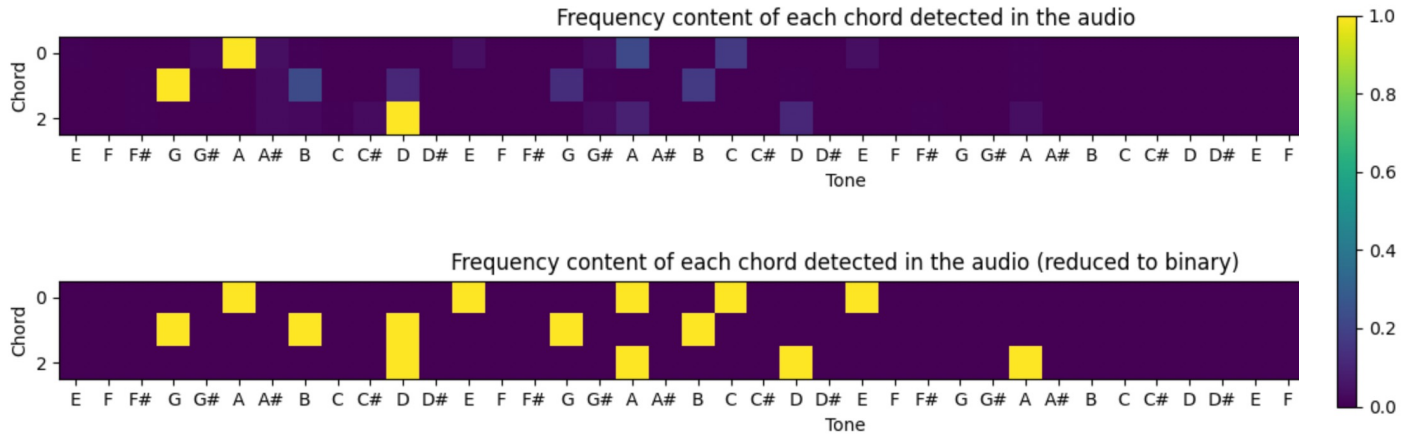
- Windowing (0.1 sec)
- Tone extraction
- Chord onset detection
- Resize windows around chords
- Tone extraction
- Tokenize tone vectors



# My Model

## Audio2Tones Pre-processing

- Windowing (0.1 sec)
- Tone extraction
- Chord onset detection
- Resize windows around chords
- Tone extraction
- Tokenize tone vectors



# My Model

## Audio2Tones Pre-processing

- Windowing (0.1 sec)
- Tone extraction
- Chord onset detection
- Resize windows around chords
- Tone extraction
- Tokenize tone vectors

$$y_{tones} \in \mathbb{R}^{48\times 1}$$

$$y_{tones}^{r,c} \in \{0,1\}$$

$$y_{tones} \in \mathbb{R}^{6\times 8} \text{ (reshape)}$$

$$y_{tokens} = \begin{bmatrix} b(y_{tones}^{0}) \\ b(y_{tones}^{1}) \\ \vdots \\ b(y_{tones}^{5}) \end{bmatrix} \in \mathbb{R}^{6\times 1}$$

| Tone vector | [00000000 00100000 01000010 01000100 00000000 00000000] |
|---|---|
| Output | [0 32 66 68 0 0] |

# My Model

Tones2Tab Encoder

- Dataset
- Parameters
- Am-G-D example

| Size | 179712 |
|---|---|
| Input length | 6 |
| Input vocab | 256 |
| Output length | 6 |
| Output vocab | 25 |

| Input example | [0 32 66 68 0 0] |
|---|---|
| Output | [0 6 8 8 7 6]* |

* Fret held on each string; 0 need to be a mute token so fret tokens shifted up by 1

# My Model

## Tones2Tab Encoder

- Dataset
- Parameters
- Am-G-D example

| | |
|---|---|
| $d_{model}$ | 64 |
| $d_{ff}$ | 256 |
| $d_{softmax}$ | 6x25 |
| $H$ | 8 |
| $L$ | 6 |
| Batch size | 32 |
| Dropout rate | 0.1 |

| | |
|---|---|
| Accuracy* | 0.986 |

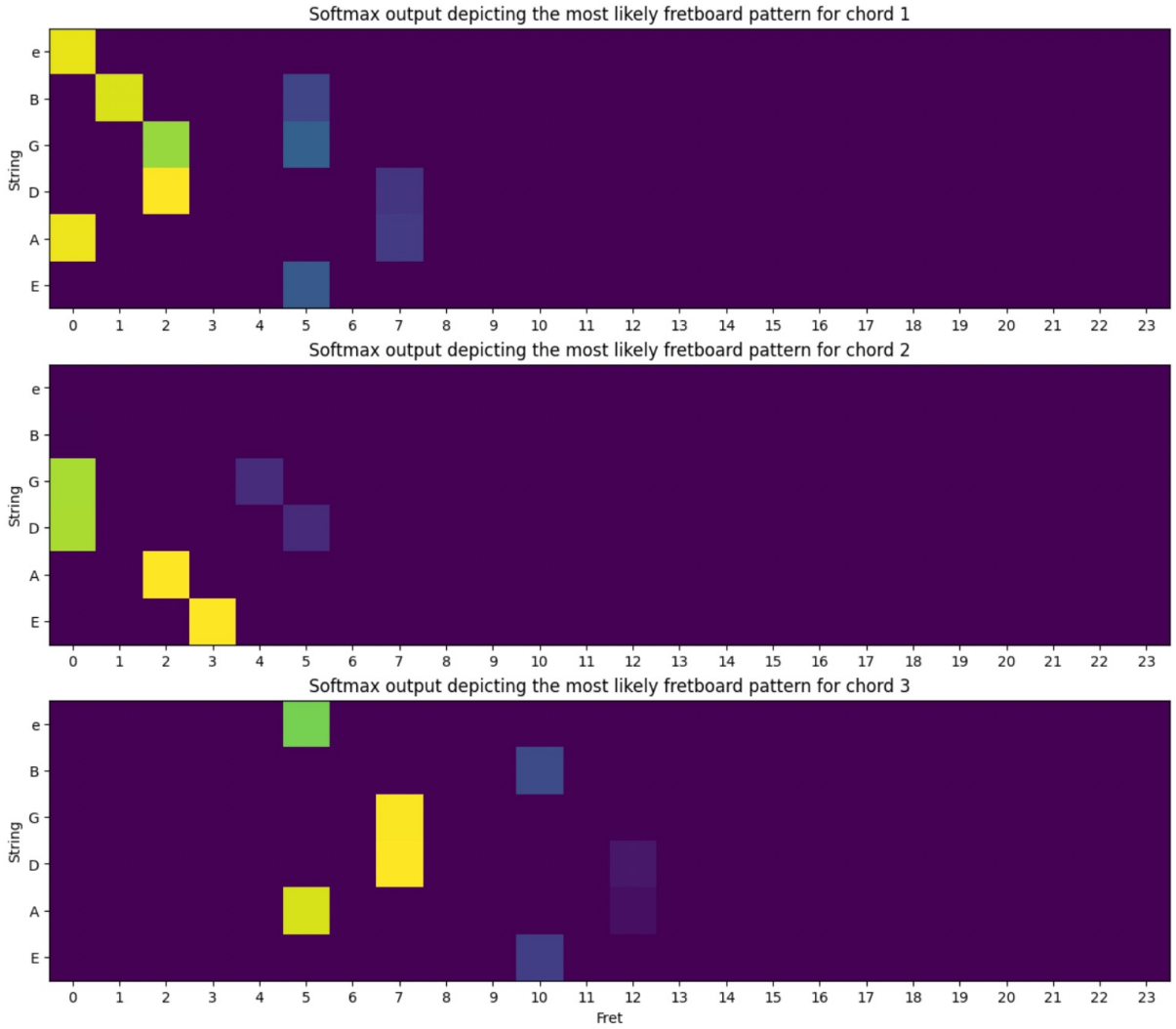
* for common chords

```
Layer (type)                 Output Shape              Param #
=================================================================
t2t_positional_embedding_3   multiple                  16384
 (T2TPositionalEmbedding)

t2t_encoder_8 (T2TEncoder)   multiple                  166016

t2t_encoder_9 (T2TEncoder)   multiple                  166016

t2t_encoder_10 (T2TEncoder   multiple                  166016
)

t2t_encoder_11 (T2TEncoder   multiple                  166016
)

t2t_encoder_12 (T2TEncoder   multiple                  166016
)

t2t_encoder_13 (T2TEncoder   multiple                  166016
)

dropout_25 (Dropout)         multiple                  0

sequential_20 (Sequential)   (32, 25)                  104985

sequential_21 (Sequential)   (32, 25)                  104985

sequential_22 (Sequential)   (32, 25)                  104985

sequential_23 (Sequential)   (32, 25)                  104985

sequential_24 (Sequential)   (32, 25)                  104985

sequential_25 (Sequential)   (32, 25)                  104985

=================================================================
Total params: 1642390 (6.27 MB)
Trainable params: 1642390 (6.27 MB)
```

# My Model

## Tones2Tab Encoder

- Dataset
- Parameters
- Am-G-D example




```
Guitar tablature:

 0  –  5

 1  –  –

 2  0  7

 2  0  7

 0  2  5

 –  3  –
```

# My Model

Tab2Chord Encoder

- Dataset
- Parameters
- Am-G-D example

| Size | 179712 |
|---|---|
| Input length | 6 |
| Input vocab | 25 |
| Output length | 3 |
| Output vocab | 15 |

| Input example | [0 6 8 8 7 6] |
|---|---|
| Output | [5 1 1]* |

* The output describes the chord pattern's characteristics [fret shape modification]

# My Model

## Tab2Chord Encoder

- Dataset
- Parameters
- Am-G-D example

| | |
|---|---|
| $d_{model}$ | 512 |
| $d_{ff}$ | 2048 |
| $d_{softmax}$ | 3x15 |
| $H$ | 8 |
| $L$ | 6 |
| Batch size | 32 |
| Dropout rate | 0.1 |

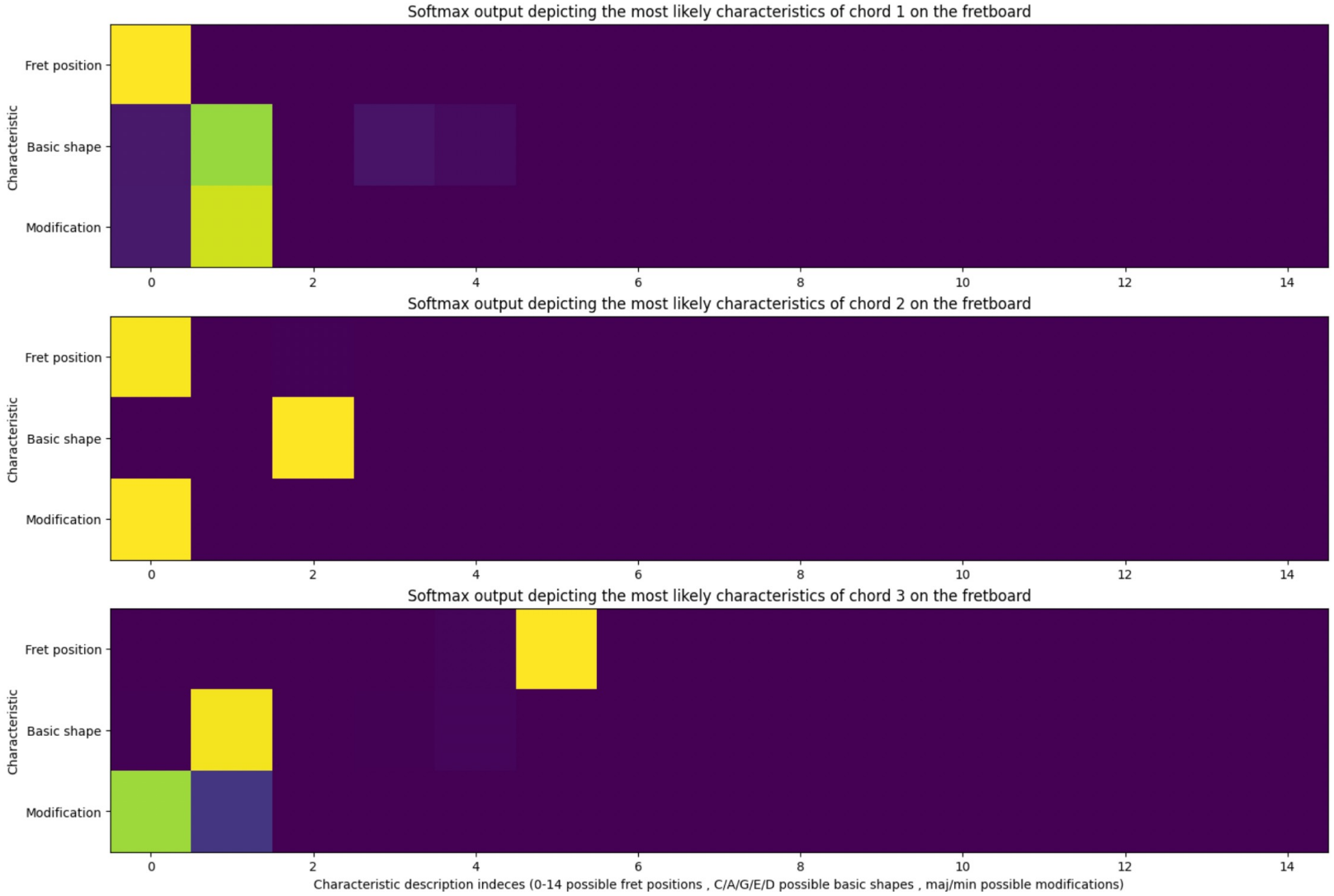
| | |
|---|---|
| Accuracy | 0.954 |
| Chord accuracy | 0.896 |

```
Layer (type)                 Output Shape              Param #
=================================================================
t2c_positional_embedding_4   multiple                  12800
7 (T2CPositionalEmbedding)

t2c_encoder_262 (T2CEncode   multiple                  10503168
r)

t2c_encoder_263 (T2CEncode   multiple                  10503168
r)

t2c_encoder_264 (T2CEncode   multiple                  10503168
r)

t2c_encoder_265 (T2CEncode   multiple                  10503168
r)

t2c_encoder_266 (T2CEncode   multiple                  10503168
r)

t2c_encoder_267 (T2CEncode   multiple                  10503168
r)

dropout_864 (Dropout)        multiple                  0

sequential_582 (Sequential   (32, 15)                  6324239
)

sequential_583 (Sequential   (32, 15)                  6324239
)

sequential_584 (Sequential   (32, 15)                  6324239
)

=================================================================
Total params: 82004525 (312.82 MB)
Trainable params: 82004525 (312.82 MB)
```

# My Model

## Tab2Chord Encoder

- Dataset
- Parameters
- Am-G-D example



# My Model

Tab2Chord Encoder

- Dataset
- Parameters
- Am-G-D example

```
The model guesses that chord 1 is played open, using the A chord shape in its minor form.

The model guesses that chord 2 is played open, using the G chord shape in its major form.

The model guesses that chord 3 has a bar at fret 5, using the A chord shape in its major form.

Chord 1 guess: Am
Chord 2 guess: G
Chord 3 guess: D
```

# My Model

Chords2Key Encoder

- Dataset
- Parameters
- Am-G-D example

| Size | 180000 |
|---|---|
| Input length | 4 |
| Input vocab | 150 |
| Output length | 1 |
| Output vocab | 12 |

| Input example | [96 96 96 96] |
|---|---|
| Output | [0]* |

* [key_label] from 0 to 11

# My Model

## Chords2Key Encoder

- Dataset
- Parameters
- Am-G-D example

| | |
|---|---|
| $d_{model}$ | 512 |
| $d_{ff}$ | 2048 |
| $d_{softmax}$ | 12 |
| $H$ | 8 |
| $L$ | 6 |
| Batch size | 32 |
| Dropout rate | 0.1 |

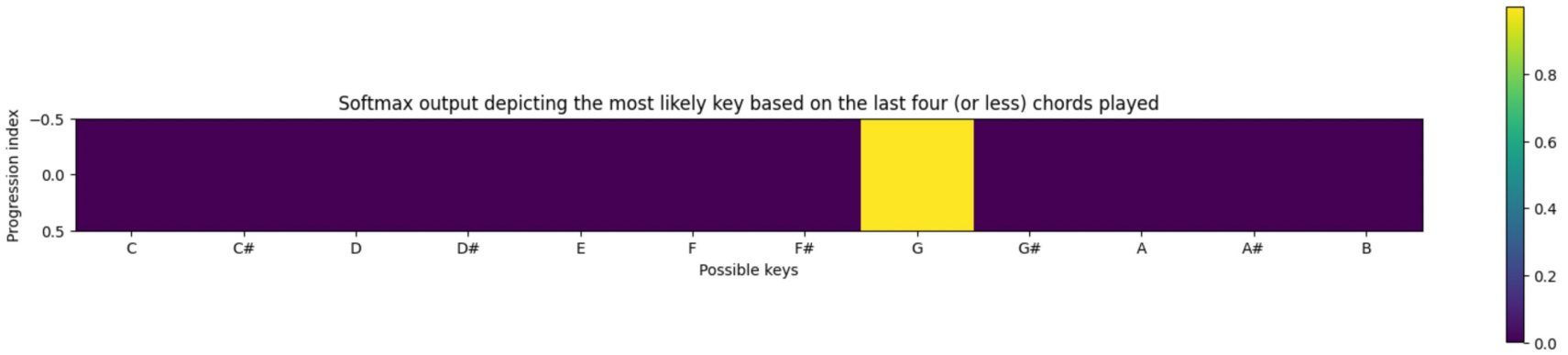
| | |
|---|---|
| Accuracy | 0.985 |

```
Layer (type)                 Output Shape              Param #
=================================================================
s2k_positional_embedding_2   multiple                  76800
7 (S2KPositionalEmbedding)

s2k_encoder_147 (S2KEncode   multiple                  10503168
r)

s2k_encoder_148 (S2KEncode   multiple                  10503168
r)

s2k_encoder_149 (S2KEncode   multiple                  10503168
r)

s2k_encoder_150 (S2KEncode   multiple                  10503168
r)

s2k_encoder_151 (S2KEncode   multiple                  10503168
r)

s2k_encoder_152 (S2KEncode   multiple                  10503168
r)

dropout_835 (Dropout)        multiple                  0

sequential_567 (Sequential   (1, 12)                   4220940
)

=================================================================
Total params: 67316748 (256.79 MB)
Trainable params: 67316748 (256.79 MB)
```

# My Model

## Chords2Key Encoder

- Dataset
- Parameters
- Am-G-D example



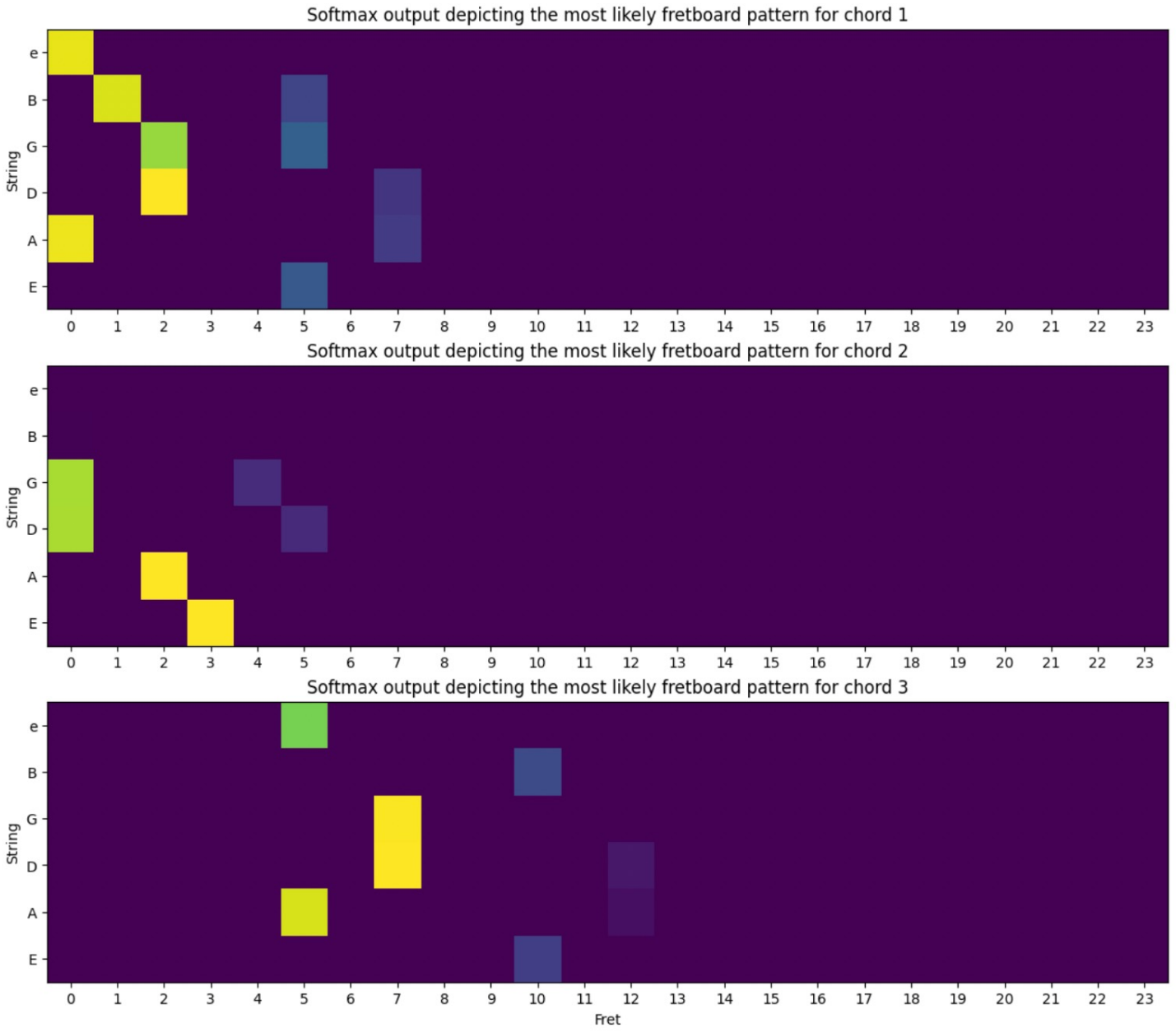
# References

[1] https://upload.wikimedia.org/wikipedia/commons/1/1b/C_Major_Scale_on_fretboard.svg

[2] https://www.researchgate.net/figure/Guitar-Fretboard-frequencies_tbl1_311707611

[3] https://www.guitarmusictheory.com/landslide-music-tab-chords-guitar-fingerpicking-songs/

# THANK YOU


Softmax output depicting the most likely fretboard pattern for chord 1
Softmax output depicting the most likely fretboard pattern for chord 2
Softmax output depicting the most likely fretboard pattern for chord 3
String
e
B
G
D
A
E
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Fret