CS464
Fall 2019
Term Project
Presentation

Listen to What Poster Says: Music Generation Based on Movie Posters

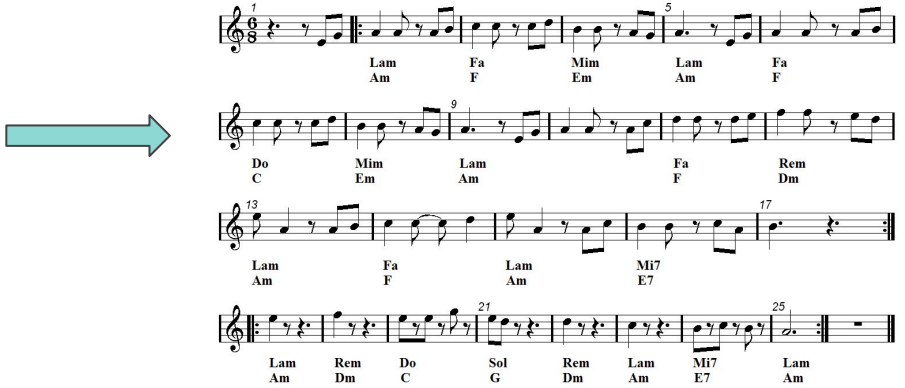
Outline

- Introduction
- Project Description
- Datasets
- Methods & Results for Poster Recognition
- Methods & Results for Soundtrack Generation
- Conclusion

Introduction

The aim of the project:

Predicting the genre of the movie poster, and producing a convenient soundtrack according to the genre.



Project Description

The main problem:

Can an original soundtrack be generated for a movie just by looking at the movie's poster?

By using a Machine Learning pipeline,

First, inference of a poster (with CNN)

Then, generating an original soundtrack for that particular genre (with RNN).

Poster Recognition

Dataset

http://www.imdb.com/title/tt114709	Toy Story (1995)	8.3	Animation|Adventure|Comedy	https://images-na.ssl-images-amazon.com/images/M/MV5BMDU2ZWJlMjktMTRhMy00ZTA5LWEzNDgtYmNmZTEwZTViZWJkXkEyXkFqcGdeQXVyNDQ2OTk4MzI@._V1_UX182_CR0,0,182,268_AL_.jpg
http://www.imdb.com/title/tt113497	Jumanji (1995)	6.9	Action|Adventure|Family	https://images-na.ssl-images-amazon.com/images/M/MV5BZTk2ZmUwYmEtNTcwZS00YmMyLWFkYjMtNTRmZDA3YWExMjc2XkEyXkFqcGdeQXVyMTQxNzMzNDI@._V1_UY268_CR10,0,182,268_AL_.jpg
http://www.imdb.com/title/tt113228	Grumpier Old Men (1995)	6.6	Comedy|Romance	https://images-na.ssl-images-amazon.com/images/M/MV5BMjQxM2YyNjMtZjUxYy00OGYyLTg0MmQtNGE2YzNjYmUyZTY1XkEyXkFqcGdeQXVyMTQxNzMzNDI@._V1_UX182_CR0,0,182,268_AL_.jpg

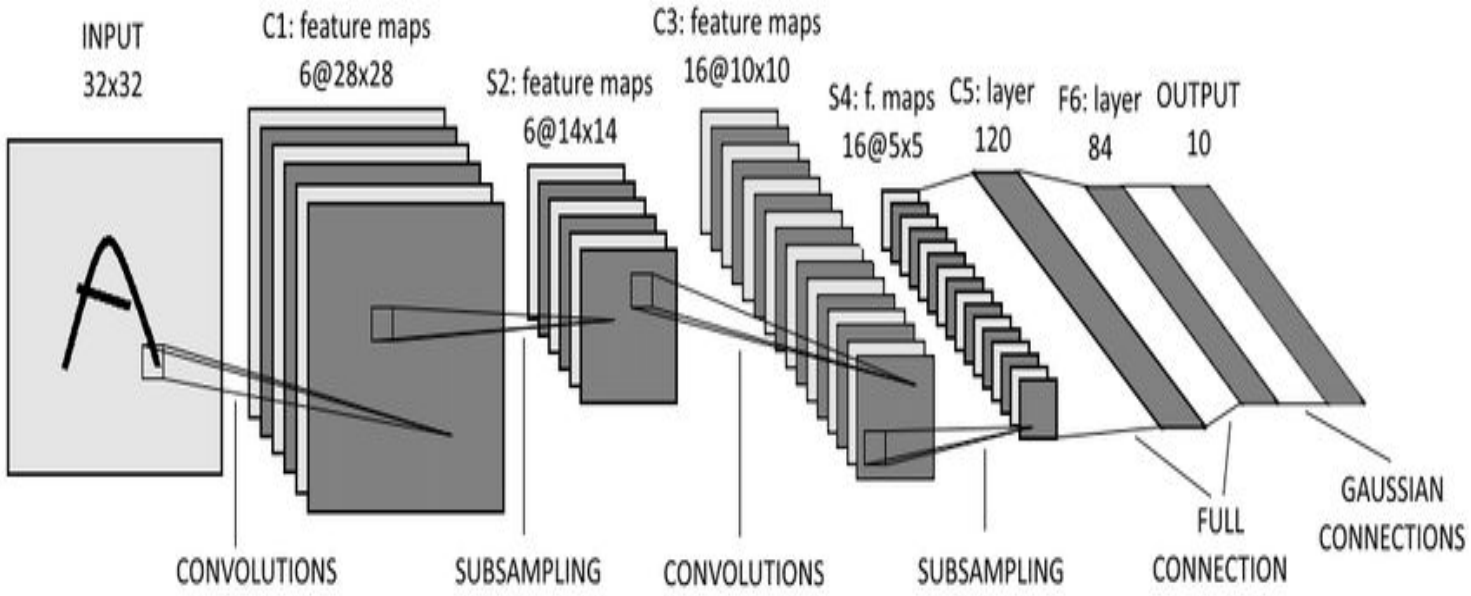
Initial Genre Distribution:

4412 action, 1545 animation, 9136 comedy, 9297 drama, 1859 horror images

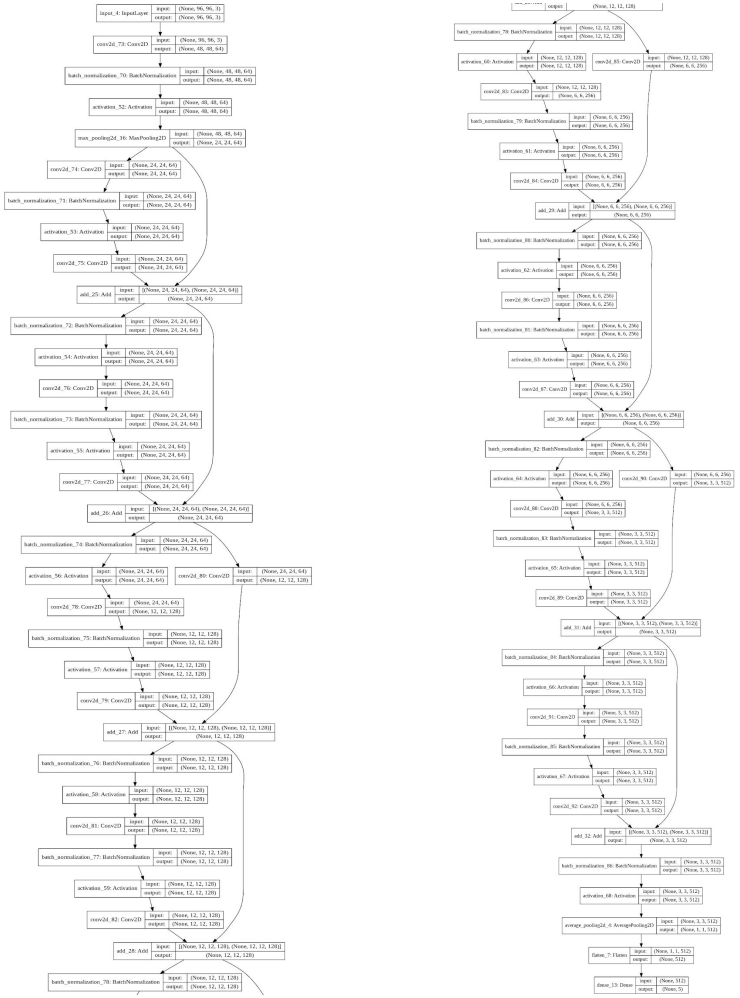
After Data Manipulation:

1445 action, 1688 animation, 1700 comedy, 1335 drama, 1243 horror with a total of 7254 images

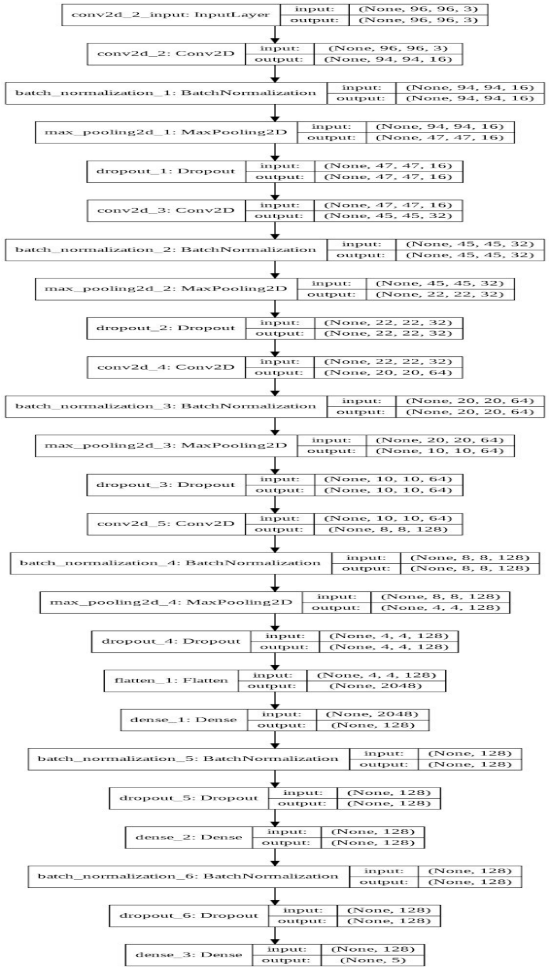
LeNet Architecture



RasNet20 Architecture



Custom Model Architecture



Adam Optimizer

$$For\ each\ Parameter\ w^j$$

$(j\ subscript\ dropped\ for\ clarity)$

$$\nu_t = \beta_1 * \nu_{t-1} - (1 - \beta_1) * g_t$$

$$s_t = \beta_2 * s_{t-1} - (1 - \beta_2) * g_t^2$$

$$\Delta\omega_t = -\eta \frac{\nu_t}{\sqrt{s_t + \epsilon}} * g_t$$

$$\omega_{t+1} = \omega_t + \Delta\omega_t$$

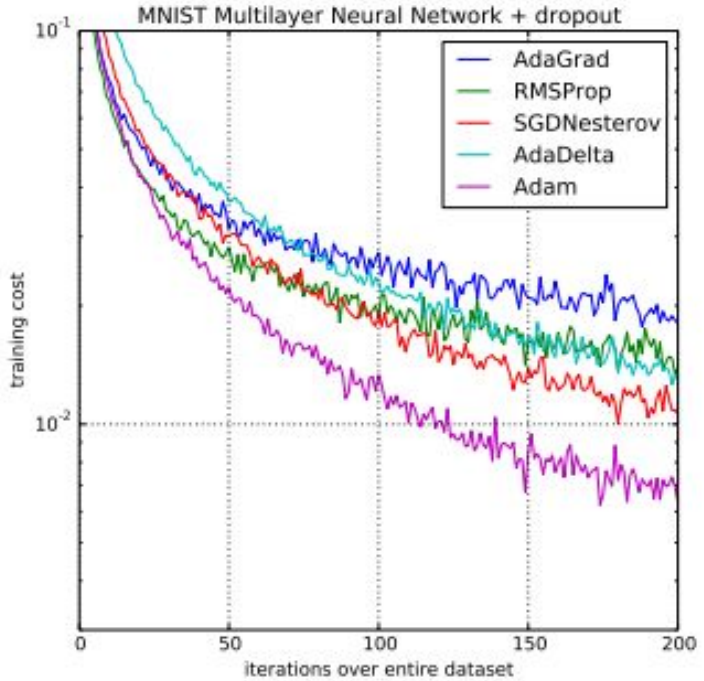
$\eta : Initial\ Learning\ rate$

$g_t : Gradient\ at\ time\ t\ along\ \omega^j$

$\nu_t : Exponential\ Average\ of\ gradients\ along\ \omega_j$

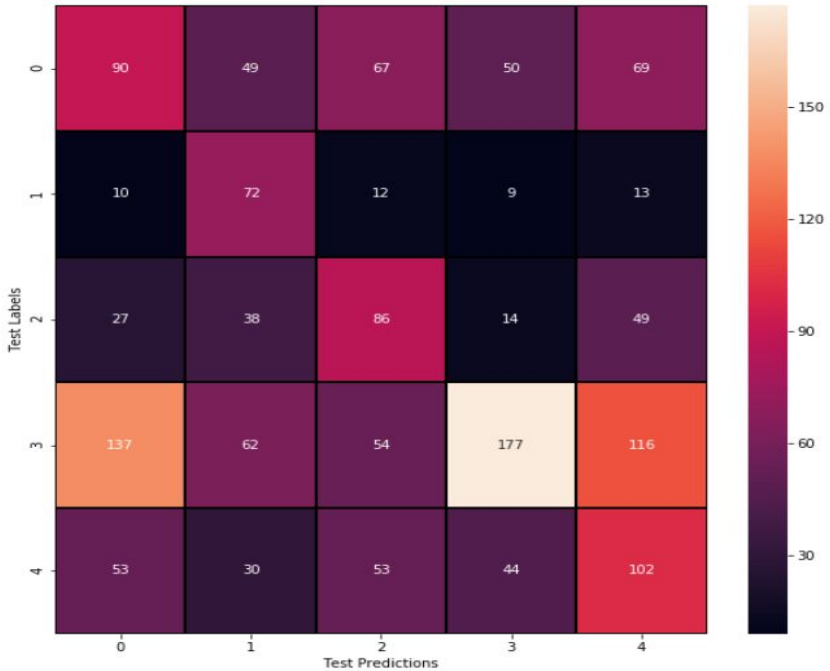
$s_t : Exponential\ Average\ of\ squares\ of\ gradients\ along\ \omega_j$

$\beta_1, \beta_2 : Hyperparameters$



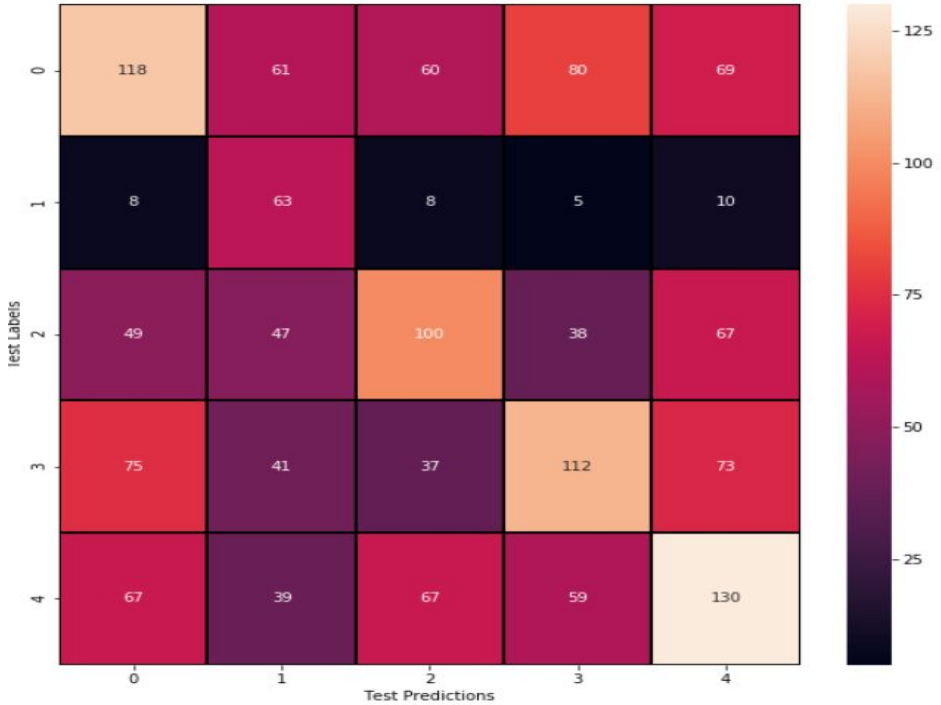
RESULTS

LeNet



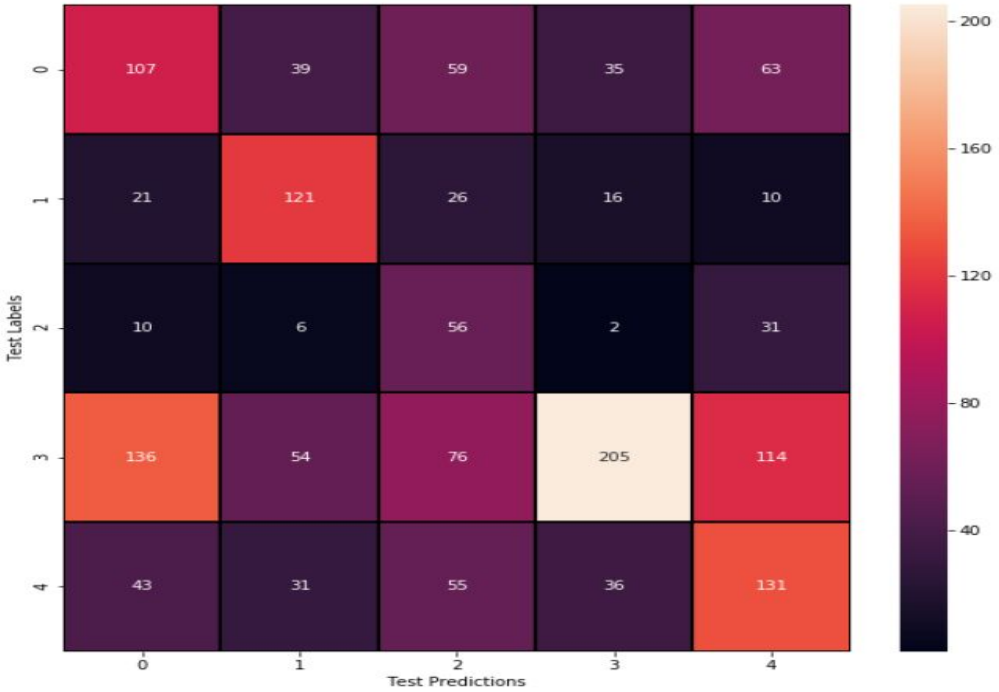
	precision	recall	f1-score	support
0	0.28	0.28	0.28	325
1	0.29	0.62	0.39	116
2	0.32	0.40	0.35	214
3	0.60	0.32	0.42	546
4	0.29	0.36	0.32	282
accuracy			0.36	1483
macro avg	0.36	0.40	0.35	1483
weighted avg	0.41	0.36	0.36	1483

RasNet20



	precision	recall	f1-score	support
0	0.37	0.30	0.33	388
1	0.25	0.67	0.37	94
2	0.37	0.33	0.35	301
3	0.38	0.33	0.35	338
4	0.37	0.36	0.37	362
accuracy			0.35	1483
macro avg	0.35	0.40	0.35	1483
weighted avg	0.37	0.35	0.35	1483

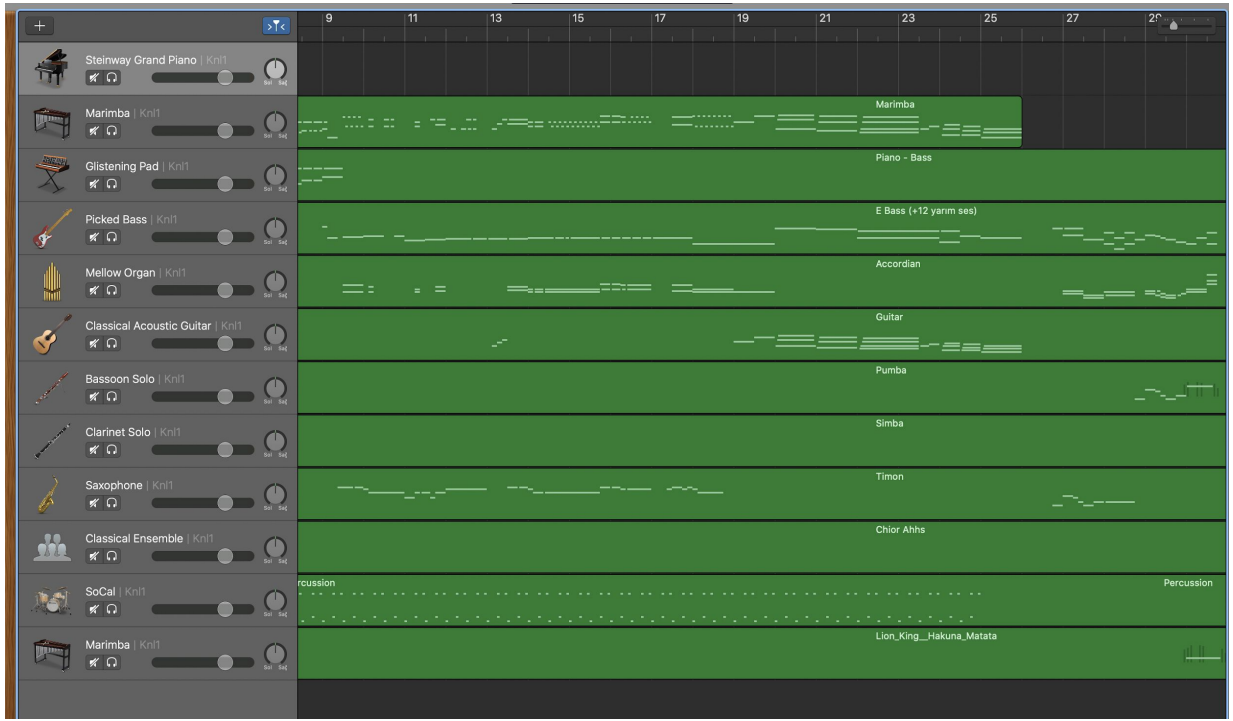
Custom Model



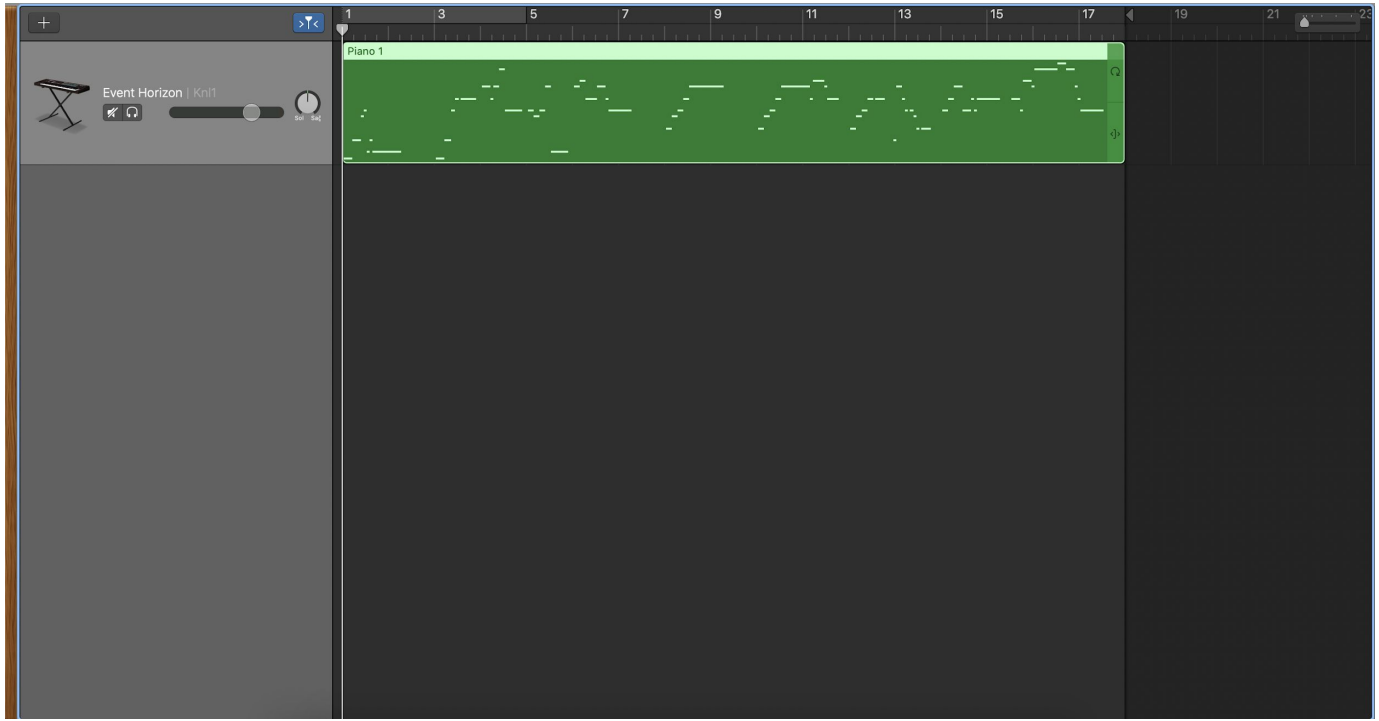
	precision	recall	f1-score	support
0	0.34	0.35	0.35	303
1	0.48	0.62	0.54	194
2	0.21	0.53	0.30	105
3	0.70	0.35	0.47	585
4	0.38	0.44	0.41	296
accuracy			0.42	1483
macro avg	0.42	0.46	0.41	1483
weighted avg	0.50	0.42	0.43	1483

Soundtrack Generator

Dataset


9
11
13
15
17
19
21
23
25
27
Steinway Grand Piano | Knl1
Marimba | Knl1
Glistening Pad | Knl1
Picked Bass | Knl1
Mellow Organ | Knl1
Classical Acoustic Guitar | Knl1
Bassoon Solo | Knl1
Clarinet Solo | Knl1
Saxophone | Knl1
Classical Ensemble | Knl1
SoCal | Knl1
Marimba | Knl1
Marimba
Piano - Bass
E Bass (+12 yarım ses)
Accordian
Guitar
Pumba
Simba
Timon
Chior Ahhs
rcussion
Percussion
Lion_King__Hakuna_Matata

Dataset



Dataset

Action: 12,000 => 10,000 => 120,000 notes

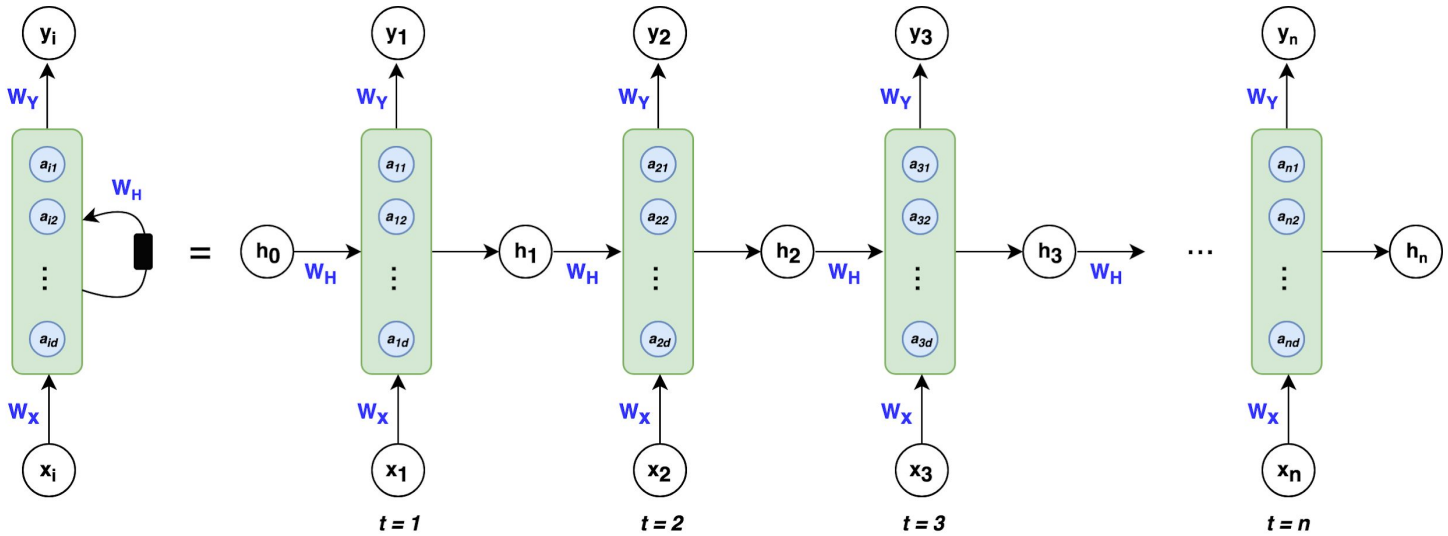
Animation: 22,203 notes

Comedy: 6,600 notes

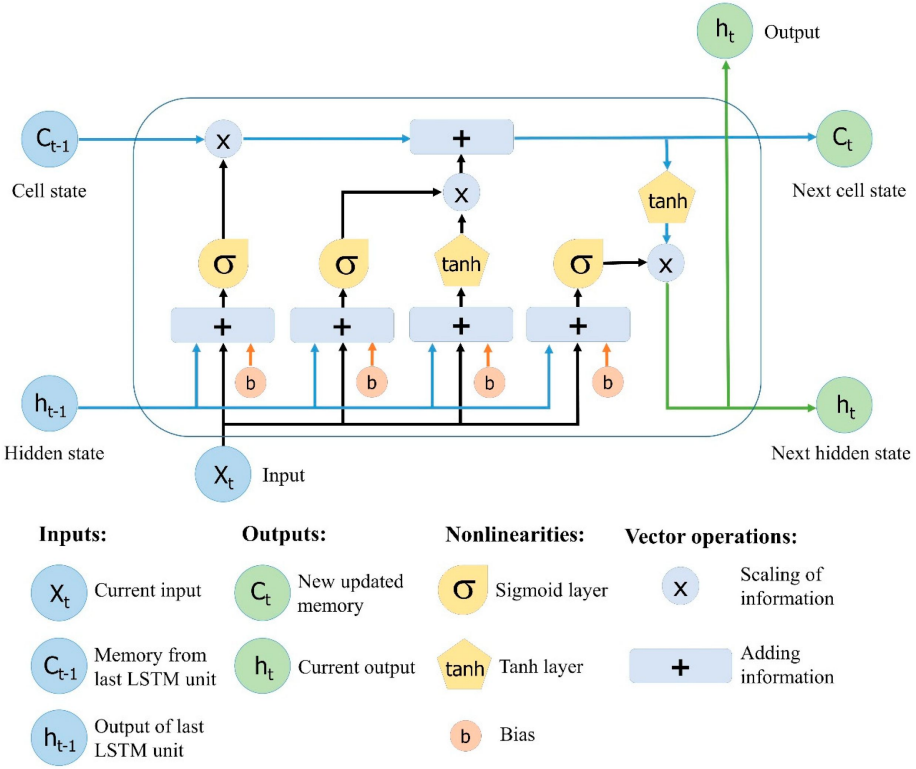
Drama: 6,000 notes

Horror: 10,000 notes

RNN


y_i
W_Y
a_{i1}
a_{i2}
a_{id}
W_H
W_X
x_i
=
h_0
W_H
y_1
W_Y
a_{11}
a_{12}
a_{1d}
W_X
x_1
t = 1
h_1
W_H
y_2
W_Y
a_{21}
a_{22}
a_{2d}
W_X
x_2
t = 2
h_2
W_H
y_3
W_Y
a_{31}
a_{32}
a_{3d}
W_X
x_3
t = 3
h_3
W_H
...
y_n
W_Y
a_{n1}
a_{n2}
a_{nd}
W_X
x_n
t = n
h_n

LSTM


h_t Output
C_{t-1} Cell state
C_t Next cell state
tanh
σ
tanh
σ
σ
h_{t-1} Hidden state
h_t Next hidden state
b
X_t Input
Inputs:
X_t Current input
C_{t-1} Memory from last LSTM unit
h_{t-1} Output of last LSTM unit
Outputs:
C_t New updated memory
h_t Current output
Nonlinearities:
σ Sigmoid layer
tanh Tanh layer
b Bias
Vector operations:
Scaling of information
Adding information

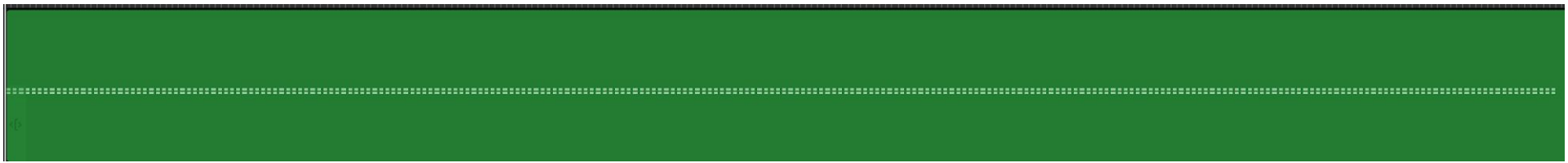
First Model

- LSTM(512)
- Dropout(0.3)
- LSTM(256)
- Dense(256)
- Dropout(0.3)
- Dense(number of notes)
- Activation(softmax)
- Categorical cross-entropy, rmsprop as optimizer

Result

- Loss value: 11.0
- Meaningless output
- Same notes through the whole soundtrack

Meaningless Output



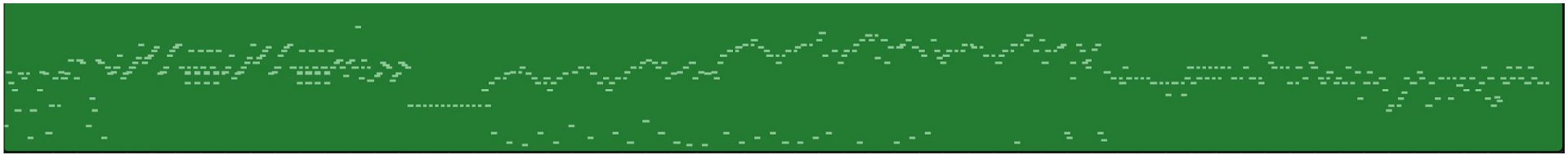
Second Model

- Bidirectional(512)
- Dropout(0.3)
- Bidirectional(512)
- Dense(number of notes)
- Activation(softmax)
- Categorical cross-entropy, rmsprop as optimizer

Results

- Loss value: 0.08
- Output sounds nice
- However, all of the outputs includes very similar melody to soundtrack of Pirates of Caribbean. Overfit!

Action Output



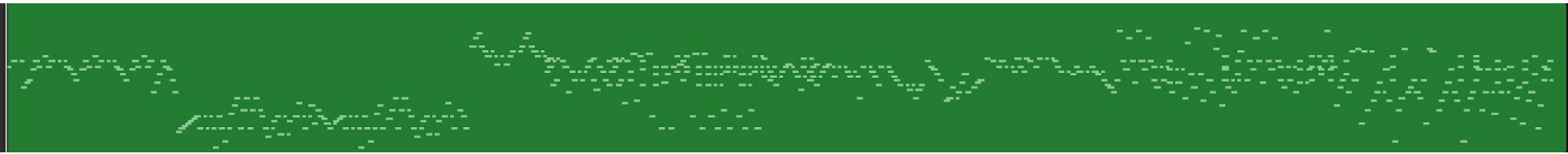
Dataset Review

- More data
- Balanced data

Results

This time, it sounds original and different than the songs from dataset

Comedy Output



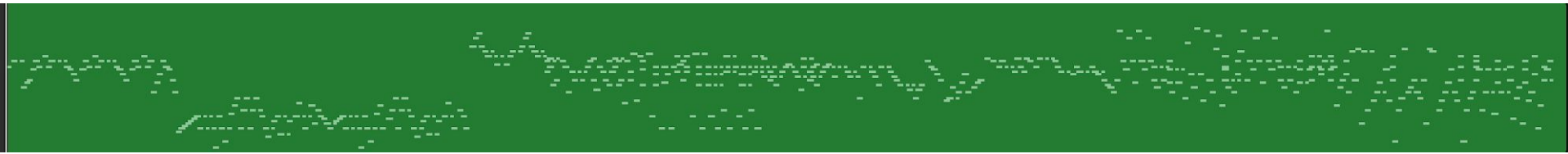
Final Model

- Bidirectional(512)
- Bidirectional(512)
- Bidirectional(512)
- BatchNorm()
- Dropout(0.3)
- Dense(256)
- Activation(relu)
- BatchNorm()
- Dropout()
- Dense(number of notes)
- Activation(softmax)
- Categorical cross-entropy, rmsprop as optimizer

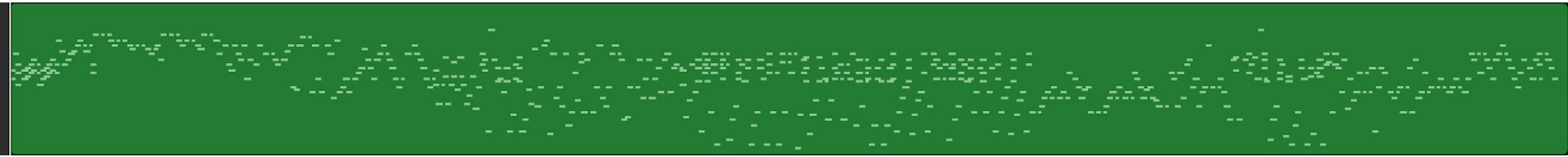
Result

- Unique outputs
- Increased performance vs low number of notes (overfit)

Previous Comedy Output



New Comedy Output



Example Soundtracks

Action (pirates of the caribbean - second model):
https://drive.google.com/file/d/1XjTmf1daCb1JLsw3pZv4hYoDUaKjyWis/view?usp=sharing

Action (second model):
https://drive.google.com/file/d/1sSGdeC_SYvSwx9C4lt845sEnbNvboYbb/view?usp=sharing

Horror (second model):
https://drive.google.com/file/d/1JFYrT3J1bTxGkt7njCp6sZEsCpMkhP_k/view?usp=sharing

Comedy (second model):
https://drive.google.com/file/d/1l48Y4LzXFneHvktSBcRKovbOdeiQYfsV/view?usp=sharing

Comedy (final model):
https://drive.google.com/file/d/1p60R99gRXLWf9For36lObNqv7sakdon8/view?usp=sharing

Conclusion

Poster Prediction:

44% accuracy, difficulties with posters that do not obey universal patterns

Music Generation:

Original soundtracks generated, belong to a certain genre