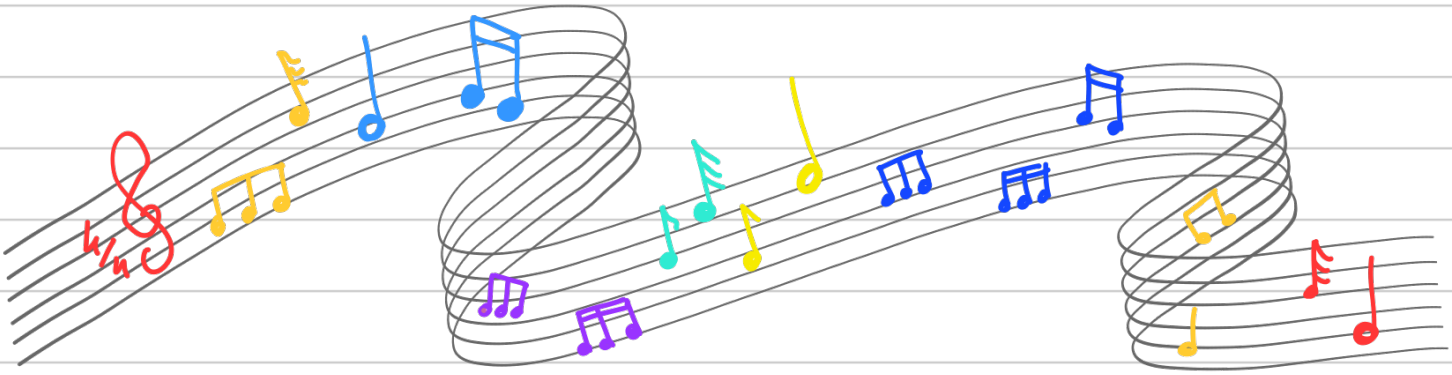
MUSIC GENERATION



WHAT

WHO

me

HOW

AUTOENCODERS

RECURSIVE NETWORKS

TRANSFORMERS

THE MAESTRO DATASET

1250+ POLYPHONIC






THE MIDI FORMAT 1983 !!

⇒ PITCH

.mid ⇒ START

⇒ DURATION

ALONG WITH OTHER DATA ...

INSTRUMENTS

VOLUME

EFFECTS

META-DATA

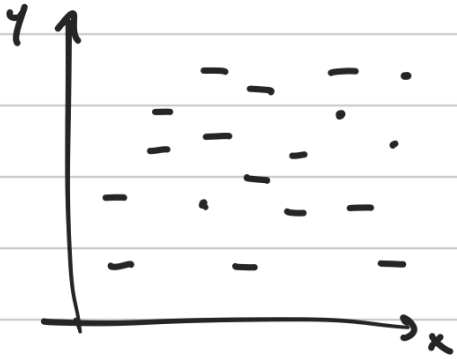
CHANNELS

BUT WE KEEP IT SIMPLE !

AUTOENCODER

TRAINING DATASET

THE PIANO ROLL REPRESENTATION

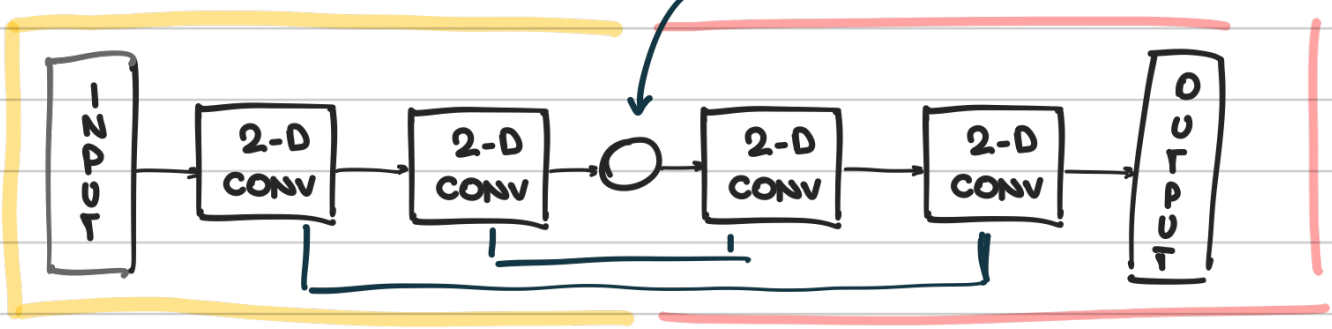


where:
- x = time
- y = pitch

SIMPLE!

FRAGMENTS OF 50 ms (time sampling)

THE MODEL



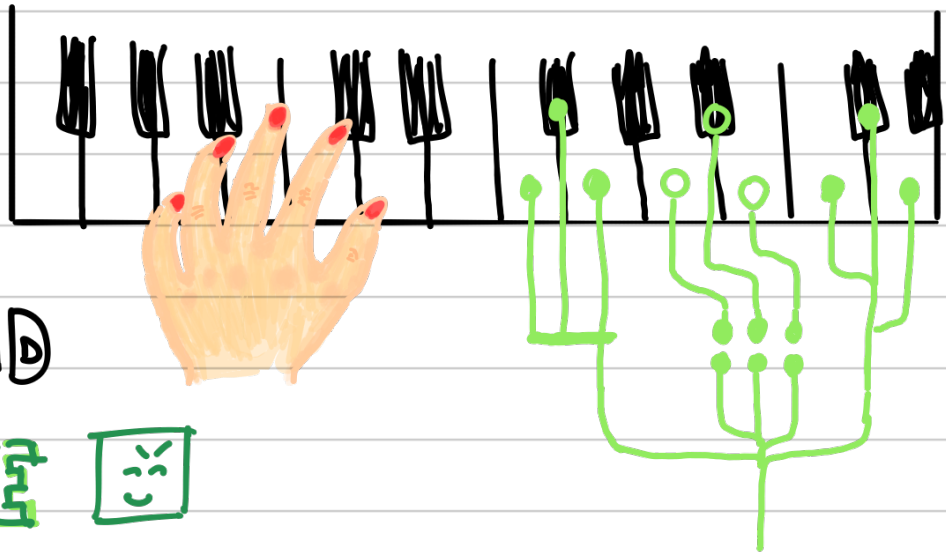
OUTPUT IS ACTIVATION OF PITCHES

⇒ THRESHOLD IS SET TO GEN.

- RELU ACTIVATIONS
- SKIP CONNECTIONS
- KERNELS 3x3
- CONV-2D 32 AND 64 FILTERS

MUSICA MAESTRO!!!

LISTEN...



... AND FIND

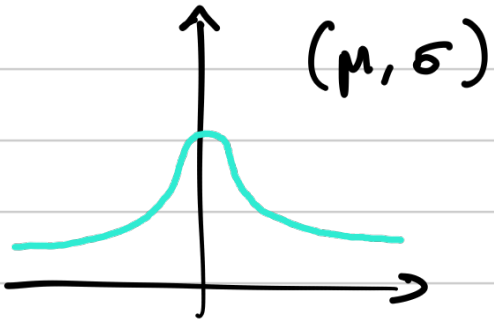
THE FAKE

WELL... NOT WHAT WE WOULD CALL "NICE"

MAYBE A BETTER LATENT SPACE...

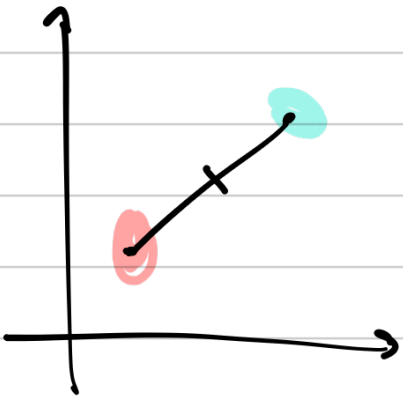
- Random value of latent space => meaningfull output
- Give the decoder "generative" capabilities

- KL-DIVERGENCE



MABYE INTERPOLATING CLASSES:

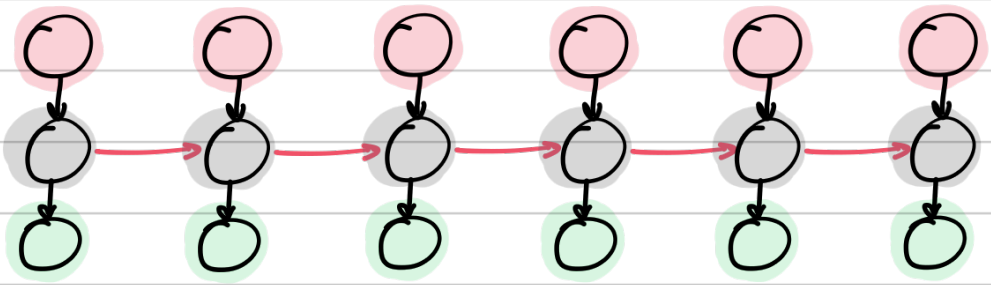
- Higher chance of a meaningfull output!
- Both for AE and VAE



SOME CONSIDERATINS

- PIANO-ROLL DOESN'T WORK
- THE AUTO-ENCODER IS NOT 'SPECIALIZED'

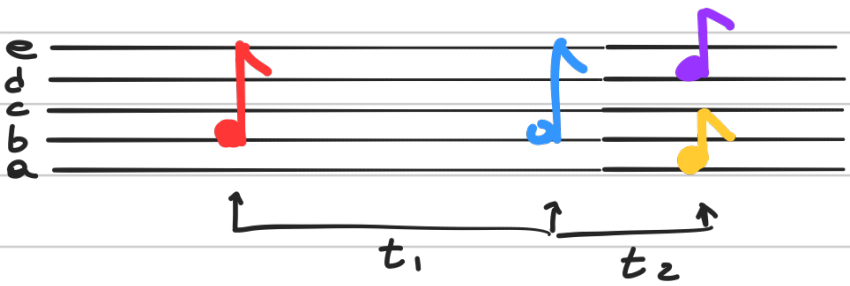
RECURSIVE NETWORKS



MUSIC IS RELATED TO TIME !!

TRAINING DATASET

- TUPLES: (PITCH, STEP, DURATION)



(b, 0, 1)
(b, t_1, 2)
(d, t_2, 1)
(b, 0, 1)

- SEQUENCE LENGHT: 25
- INPUT SHAPE: (25, 3)

⇓

- OUTPUT SHAPE: (1 , 3)

THE MODEL

STEP ↓ 1.0 *

DURATION ↓ 1.0 *

PITCH ↓ 0.05 *

POSITIVE PRESSURE

$mse + 10 \cdot \max(-y, 0)$

↑ "ONLY IF y IS NEGATIVE"

SPARSE CATEGORICAL CROSS ENTROPY

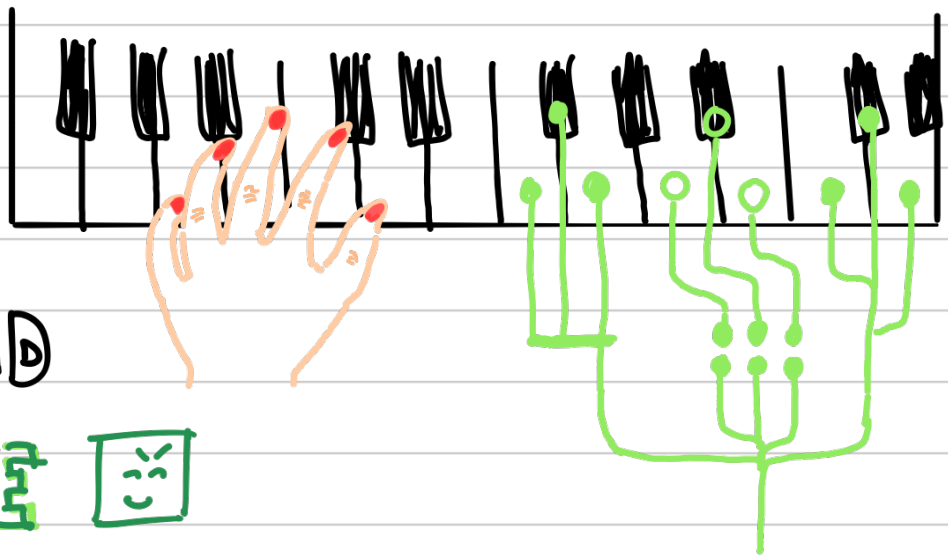
- LOSS WEIGHTS *

- HOT OR COLD: TEMPERATURE

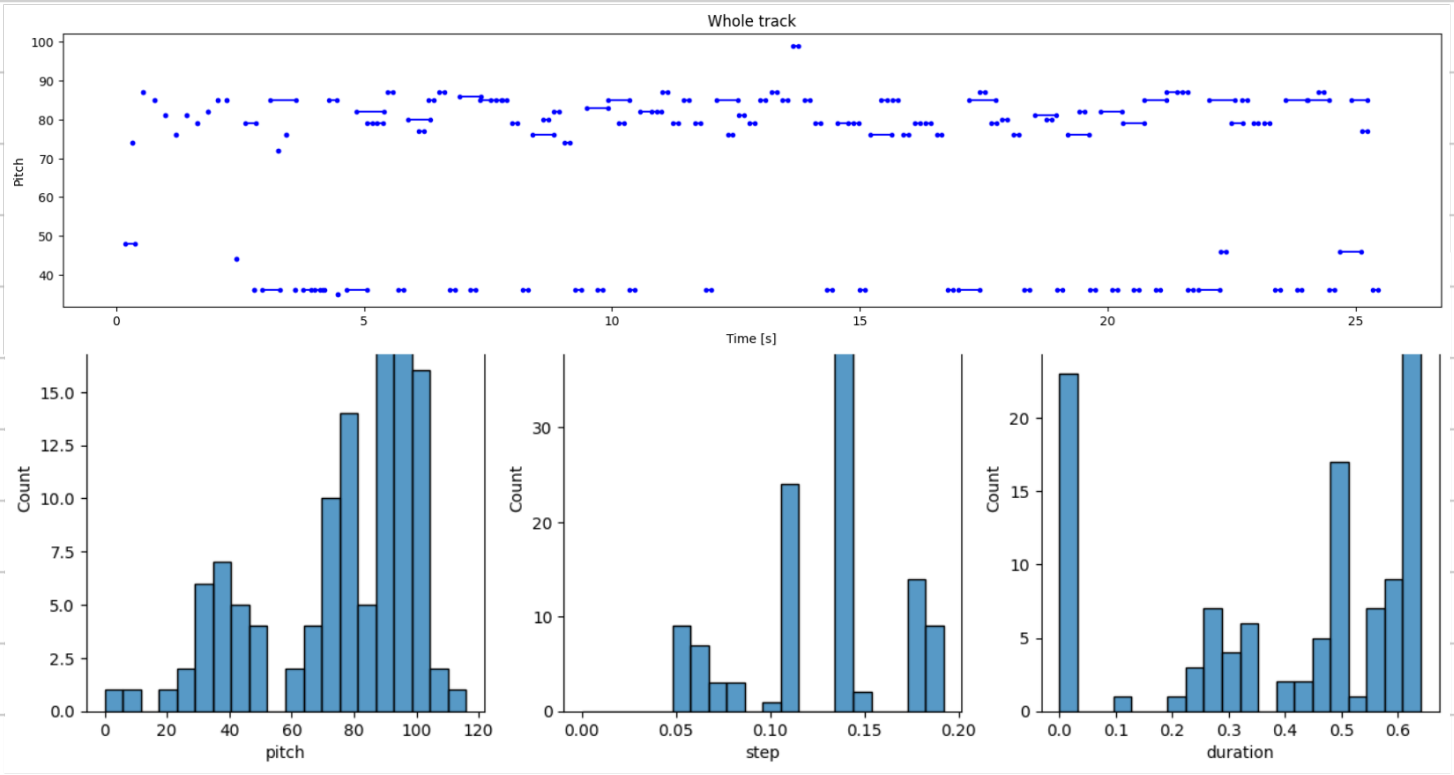
↑ ACTS ON PITCH

MUSICA MAESTRO!!!

LISTEN...



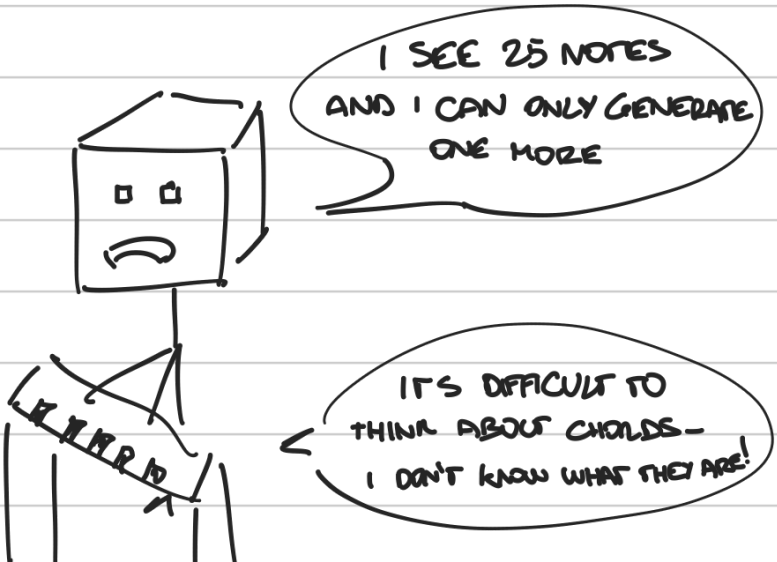
... AND FIND THE FAKE



CONSIDERATIONS

- MODEL CREATES PAUSES
- SIMPLE ARMONIES
- BUT IT FEELS A BIT MONOPHONIC...

CHORDS ARE NOT ENFORCED



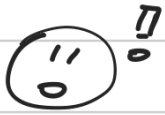
SOLUTIONS:

- EMBED CHORDS AS INPUTS
- GENERATE MORE THEN 1 NOTE ON EACH ITERATION

TRANSFORMERS

TRAINING DATASET

TOKENIZER: 340+ EVENTS



TRAIN 90% 10% TEST

NOTES BATCHING → SEQUENCING

THE MODEL

"MUSIC TRANSFORMER" 12 SEP 2018

↑ GENERATING MUSIC USING LONG-TERM STRUCTURE

- RELATIVE ATTENTION
  - STD. RELATIVE POSITION ENCODING
  - ENABLE LONGER SEQUENCES - MINUTES
- TEMPORAL SENSITIVITY
  - SINE POSITION ENCODING
- TOP-K SAMPLING
  - CONSTRAINED RANDOMNESS

PREPROCESS:

MIDI INPUT → EMBEDDING (VECTORIZE INPUTS) → POSITIONAL ENCODING (PRESERVE TIME) → DROPOUT (GENERALISE)

DECODER:

RELATIVE GLOBAL ATTENTION (KEY PART) → NORM → FFN (NON-LINEARITY) → NORM (VANISHING GRADIENTS)

GENERATION:

OUTPUT LAYER (NOTES TOKENS) — TOP-k SAMPLING (RANDOMNESS)

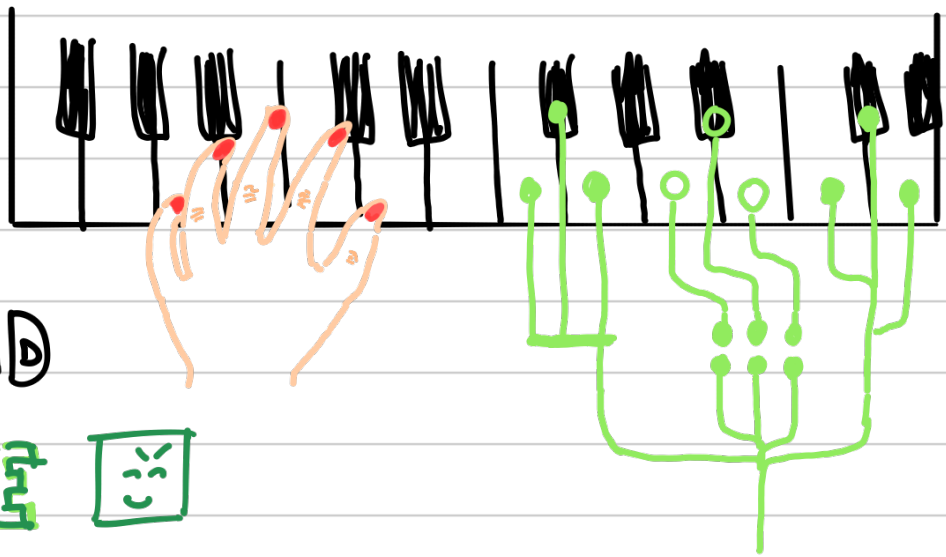
1 DURING TRAINING 2

RELATIVE ATTENTION

MASKED CATEGORICAL CROSS ENTROPY

MUSICA MAESTRO!!!

LISTEN...



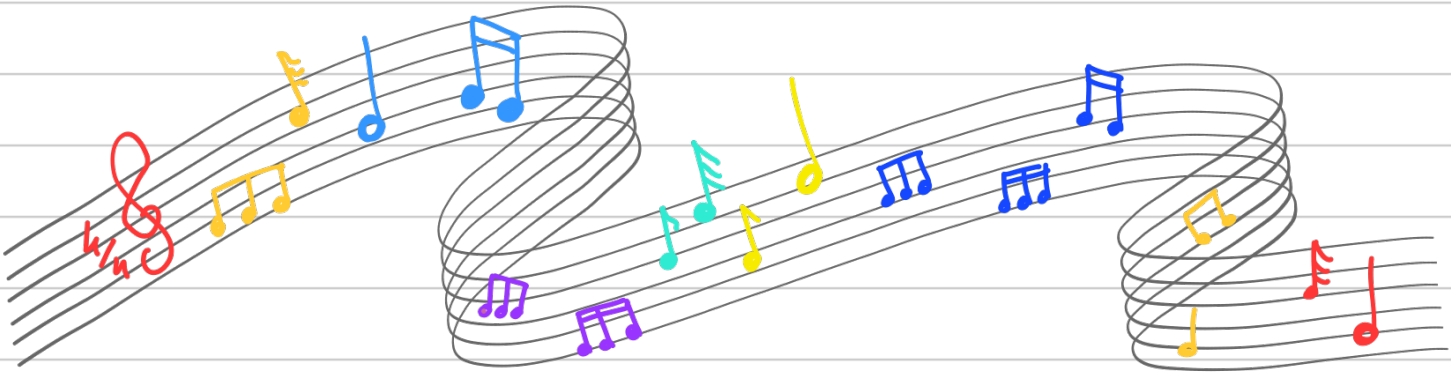
... AND FIND

THE FAKE

CONSIDERATIONS

- CHORDS ✓
- ARMONIES ✓
- PAUSES ✓
- TOKENIZATION WORKS

CONCLUSION



A JOURNEY THROUGH:

AEs — NOT THE RIGHT APPROACH

RNNs — A JUMP FORWARD ON TEMPORAL DATA

VAEs — BETTER RESULTS ON DIFFERENT SCENARIOS

TRANSFORMERs — HIGH POTENTIAL AND BEST RESULTS

AND I HAD A LOT OF FUN!