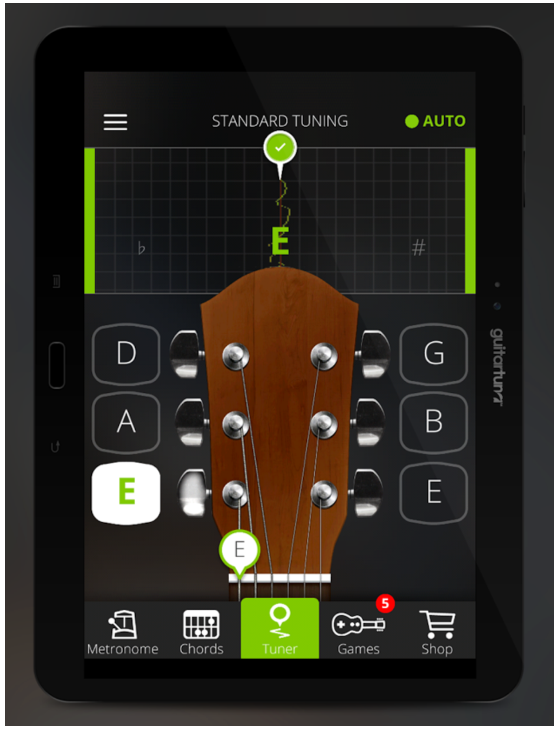
A Universal Music Translation Network

NOAM MOR, LIOR WOLF, ADAM POLYAK, YANIV TAIGMAN
FACEBOOK AI RESEARCH

Liron London

Some images were taken from [Jaley Dholakiya's blog post]

Computers Love Music



Can Computers Mimic Music?

Music Translation

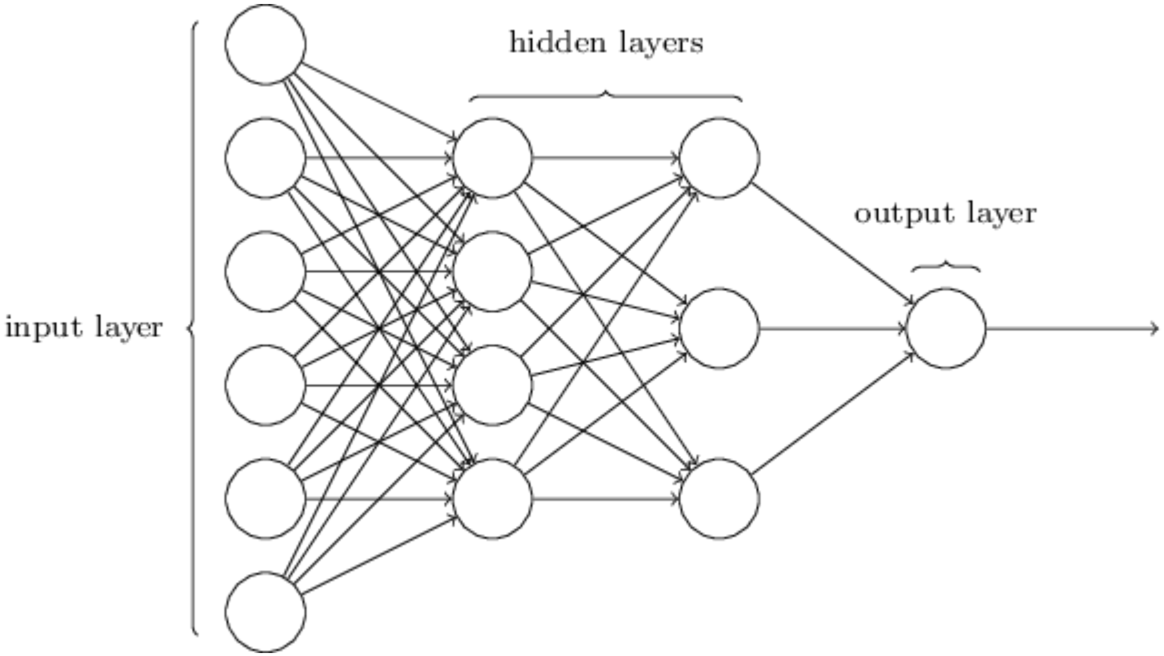
- The goal: translating music across instruments, genres and styles
- The method: neural networks – multi-domain wavenet autoencoder
- The challenge: no data!

Technical Background

In order to understand the research, we'll discuss some concepts and terms first:

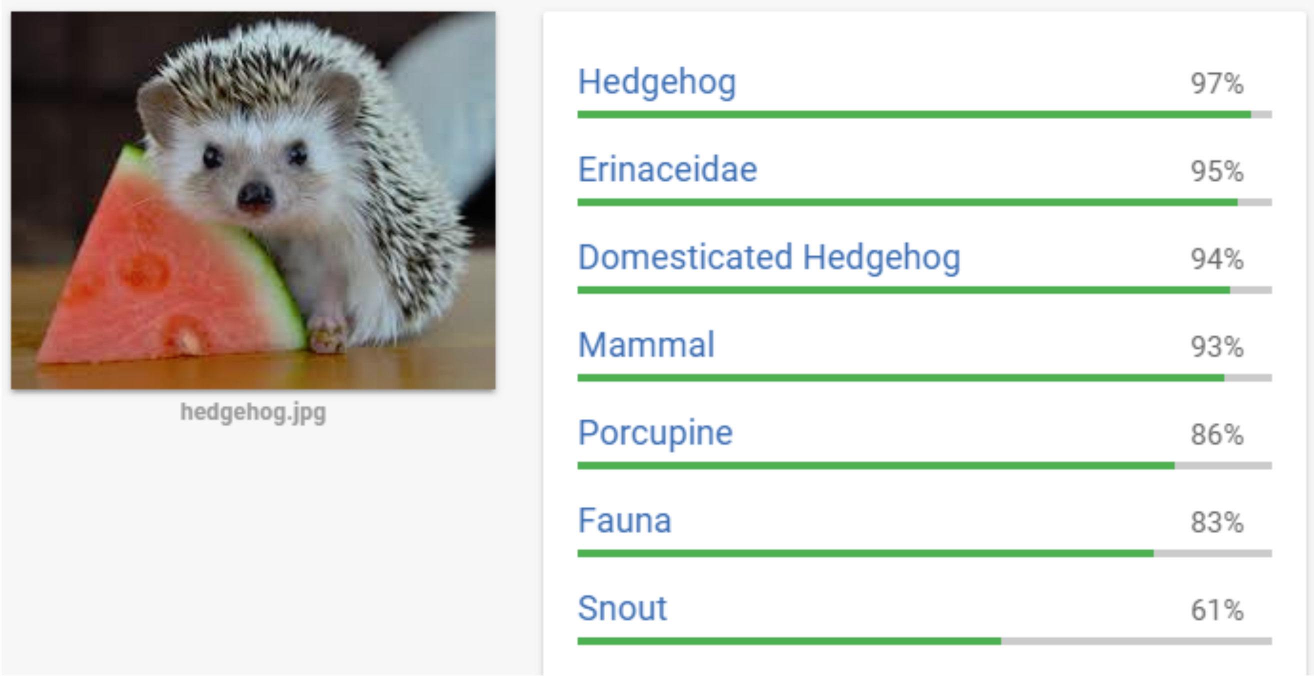
- Neural networks
- Domain transfer

Neural Networks



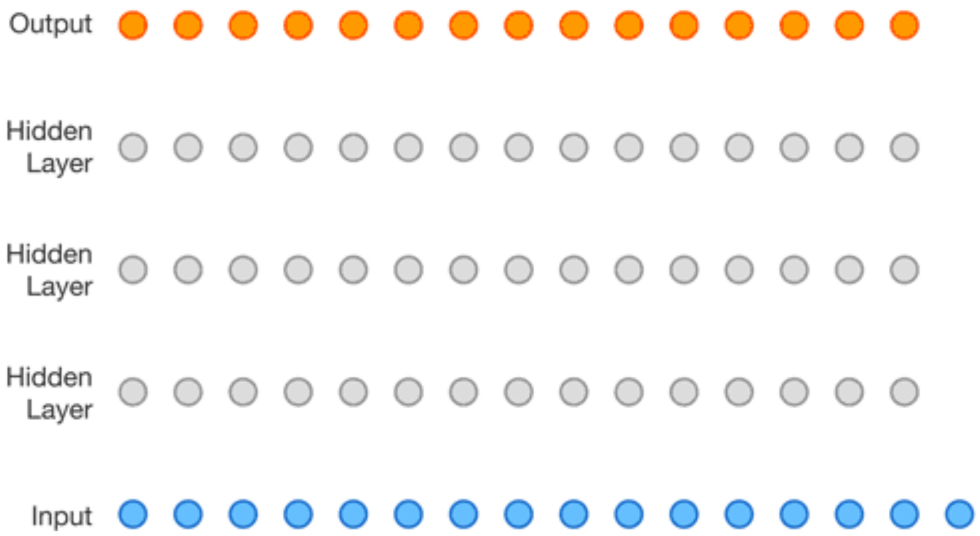
Neural Networks - Types

- Convolutional (CNN): in our case – a classifier that receives an input and determines which class it belongs to
  - Can provide a clear-cut or a probable answer



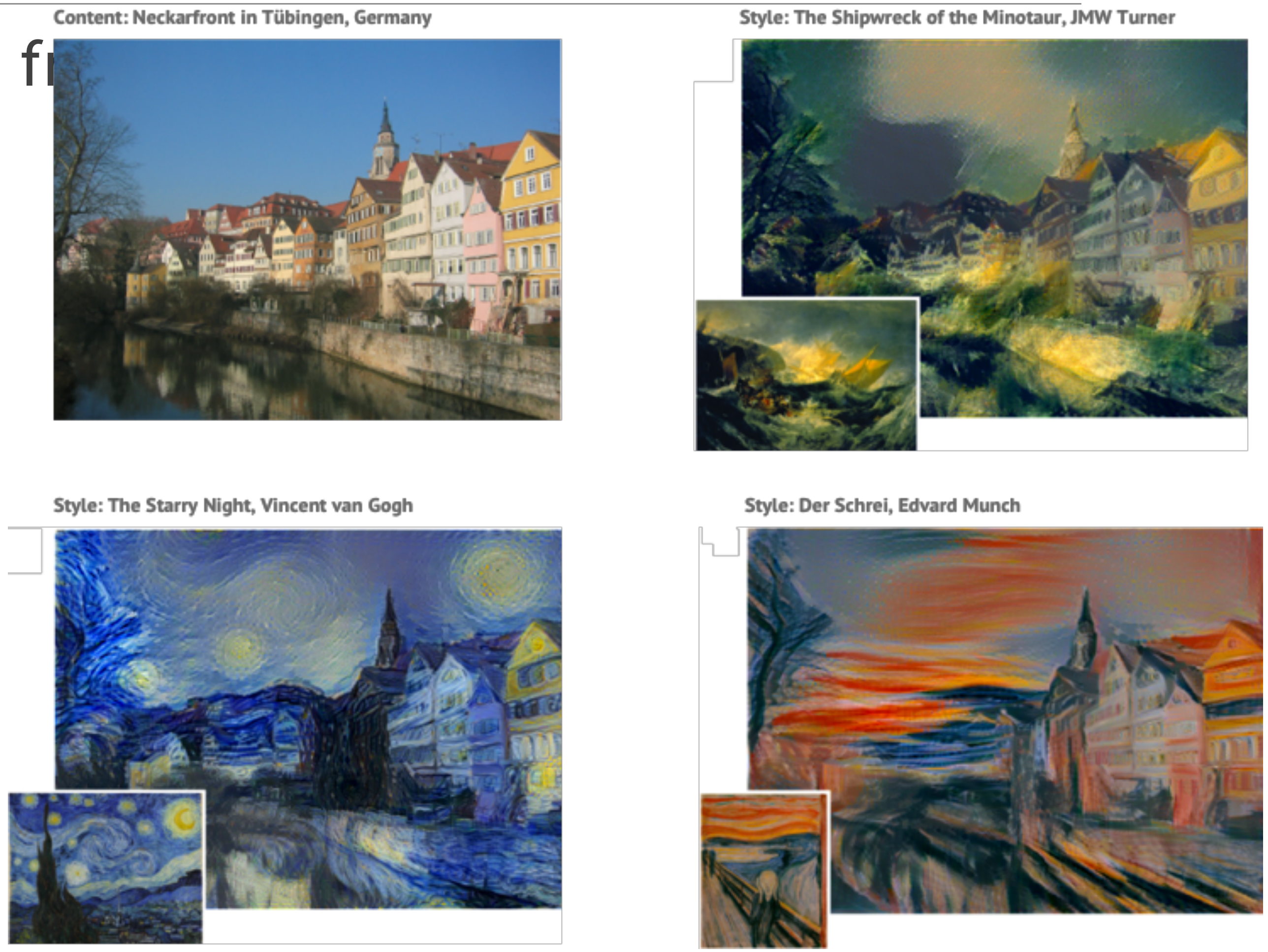
Neural Networks - Types

- Auto-regressive (AR): creates the next frame in time, adds it to history, thus lengthening the history and building the “future” upon it.



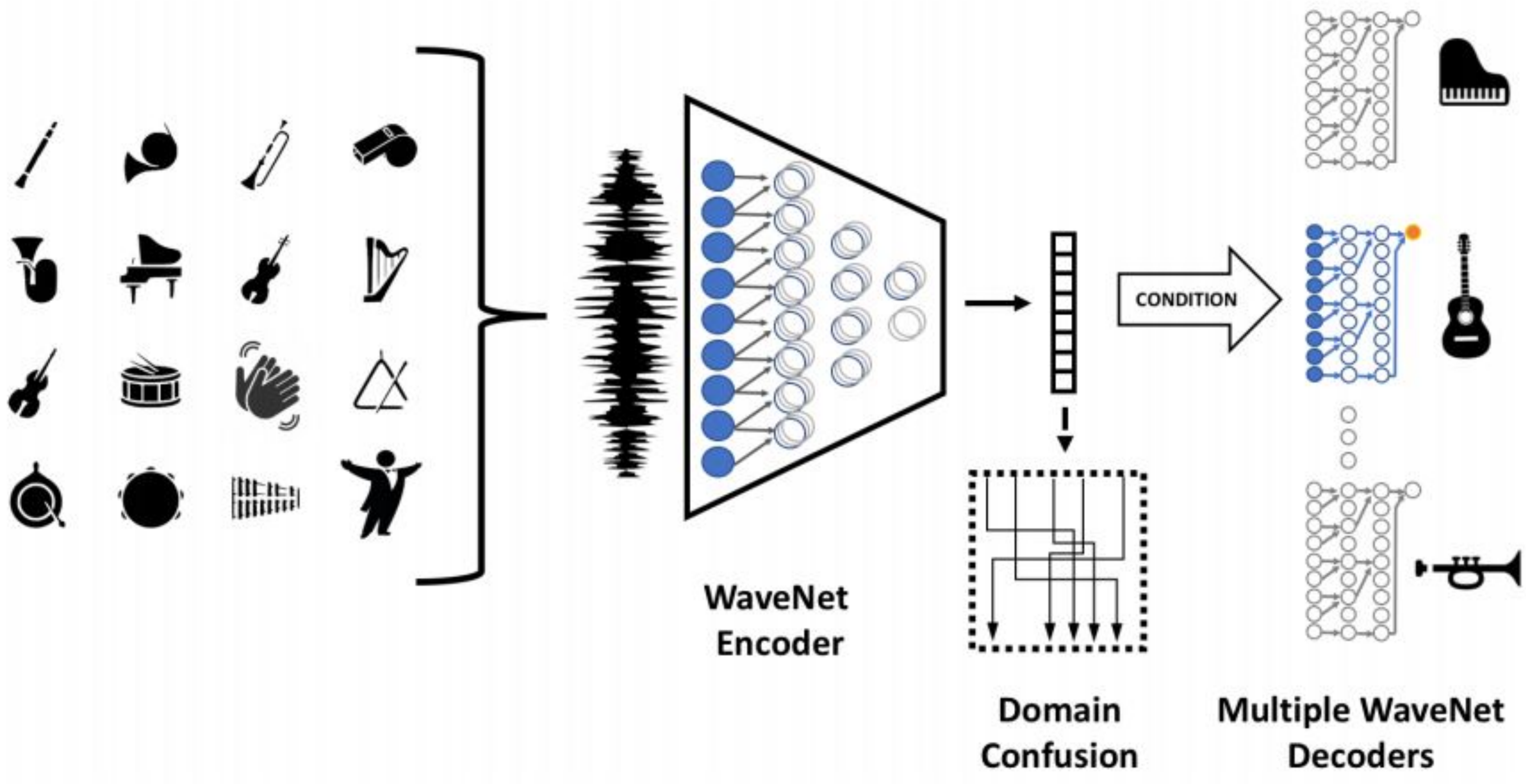
Domain Transfer

- The challenge of translating input fr
  one domain to another
- Can be unsupervised



The method

LET’S HEAR SOME MUSIC ♫

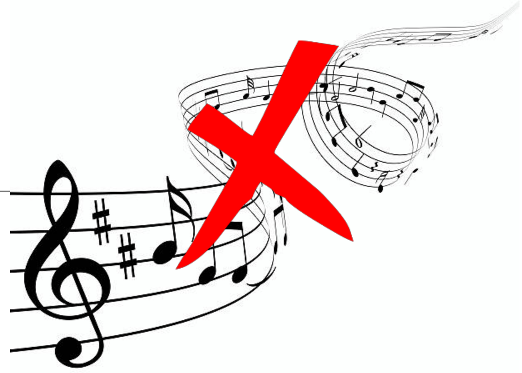

CONDITION
WaveNet
Encoder
Domain
Confusion
Multiple WaveNet
Decoders

Data

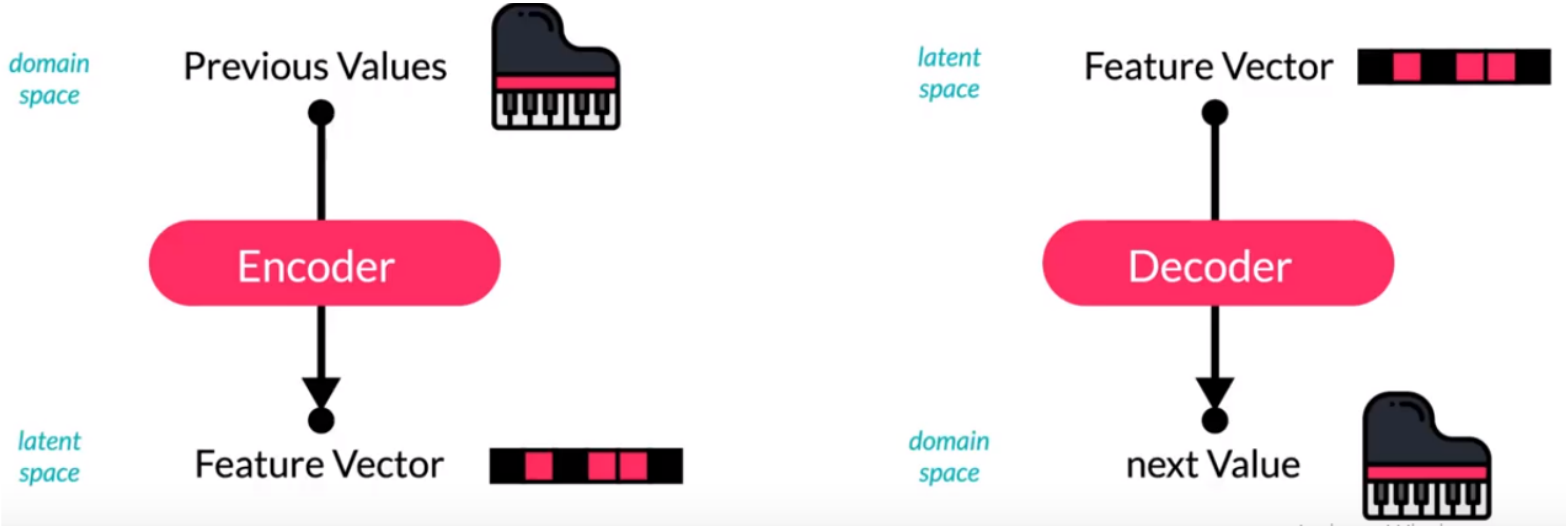
- 6 input musical domains: Mozart - symphonies, Bach – orchestra and choir, Bach – organ, Bach – harpsichord, Beethoven – piano
- Data separated to train and test sets
- Each musical piece split to 1-second segments

(harpsichord)

Encoding



- NNs work on numbers, not music
- Need to encode the music to numbers
- Can’t do notes – too specific, too complicated, existing results for simpler tasks are not good enough
- One encoder to rule them all

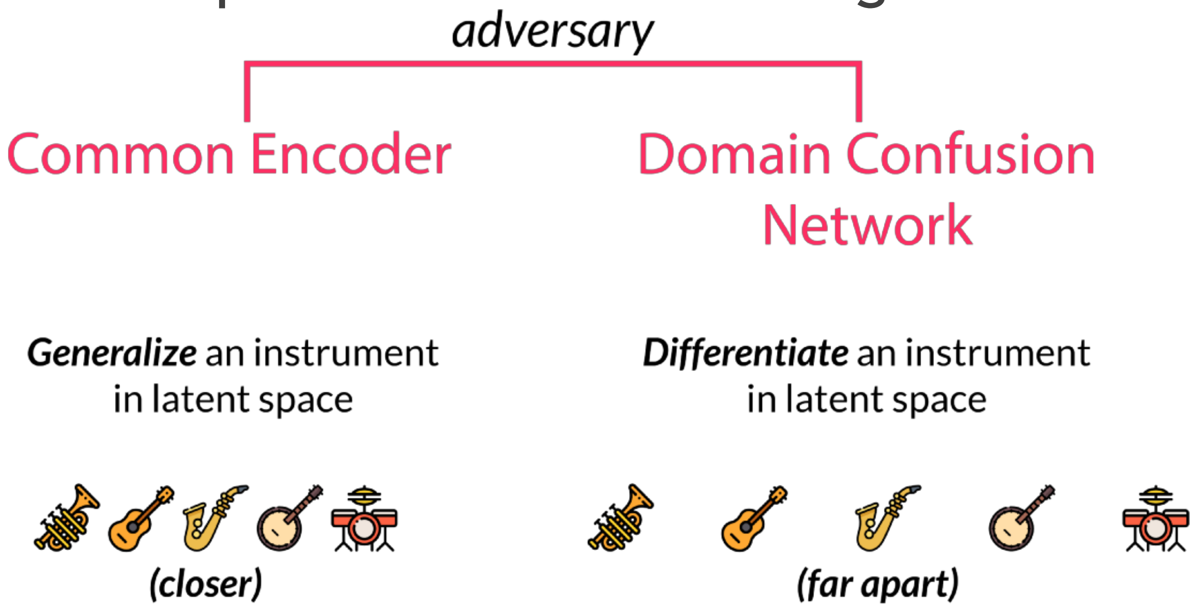

domain space
Previous Values
Encoder
latent space
Feature Vector
latent space
Feature Vector
Decoder
domain space
next Value

Encoding

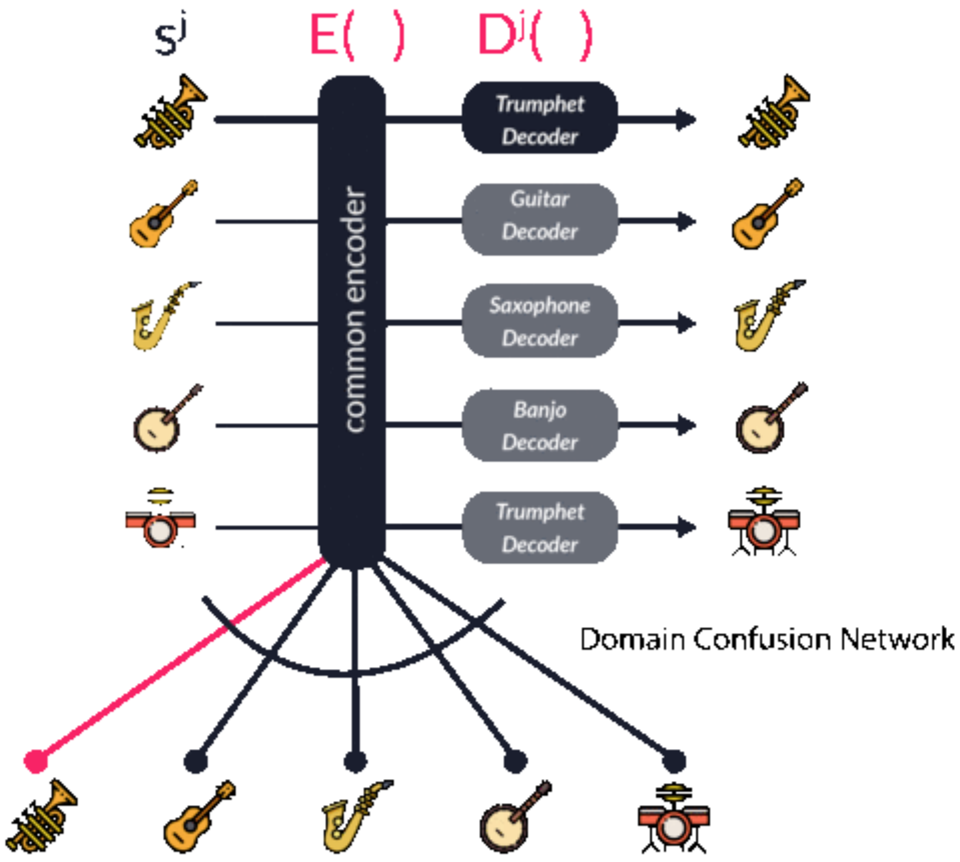
- Based on WaveNet
- Input music is encoded to latent space
- In order to prevent the encoder from memorizing music – noise was added to the data
  - In each 1-sec file, the pitch of a randomly chosen segment length of between 0.25-0.5 seconds gets modulated by a -0.5 to 0.5 half-tone

Data Augmentation

- The goal: prevent the system from encoding data that is domain-specific
- The means: confusion network – another network, used only during training, which is responsible for minimizing the classification loss



Training



(main loss)

$$\sum_{j} \sum_{s^j} E_{r}[\, L(\, D^j(\, E(\, O(s^j, r))), s^j)]$$

(domain confusion network loss)

$$- \lambda \sum_{j} \sum_{s^j} E_{r}[\, L(C(\, E(\, O(s^j, r))), j)]$$

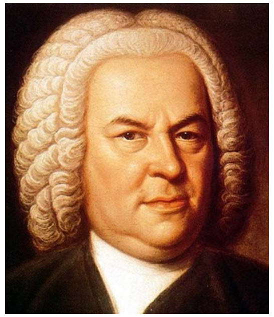
Loss Function, Explained

In red – the decoder is given an encoded sample, outputs a “cover” in the same style

In blue – the domain confusion network is given an encoded sample, and outputs which domain it belonged to

Evaluating the New Music

- How do you give a score to a cover version?



- Compare the network's results to the same task performed by human musicians
  - The task – convert 60 segments of 1 second each, to piano
- Comparison done by both human listeners and automatic score

Results

- The human scoring was done using CrowdMOS (mean opinion score), an open source tool for Mechanical Turk that helps detect and discard inaccurate scores
- The users were asked 2 questions: on a scale of 1 to 5 -
  - what's the quality of the audio?
  - How well does the converted version match the original?

Table 1: MOS scores (mean± SD) for the conversion tasks.

Converter	Harpsichord→ Piano		Orchestra→ Piano		New domains→ Piano	
	Audio quality	Translation success	Audio quality	Translation success	Audio quality	Translation success
E	3.89 ± 1.06	4.10± 0.94	4.02± 0.81	4.12± 0.97	4.44±0.82	4.13± 0.83
M	3.82 ± 1.18	3.75± 1.17	4.13± 0.89	4.12± 0.98	4.48±0.72	3.97± 0.88
A	3.69 ± 1.08	3.91± 1.16	4.06± 0.86	3.99± 1.08	4.53±0.79	3.93± 0.95
Our	2.95 ± 1.18	3.07± 1.30	2.56± 1.04	2.86± 1.16	2.36±1.17	3.18± 1.14

Results

- The automatic scoring was done by pitch matching
- The system was more true-to-source than the pianists

Table 2: Automatic quality scores for the conversion task.

Converter	Harpsichord→ Piano		Orchestra→ Piano		New domains→ Piano	
	NCC	DTW	NCC	DTW	NCC	DTW
E	0.82	0.98	0.78	0.97	0.76	0.97
M	0.69	0.96	0.65	0.95	0.72	0.95
A	0.76	0.97	0.73	0.95	0.75	0.94
Our	0.84	0.98	0.82	0.97	0.88	0.98

Significance of This Research

- Superior results compared to existing methods
- Breaking ground in the field of musical AI
- Democratization of music
- Changing what was considered possible