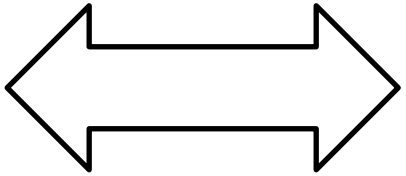
Symbolic Music Genre Transfer Insights

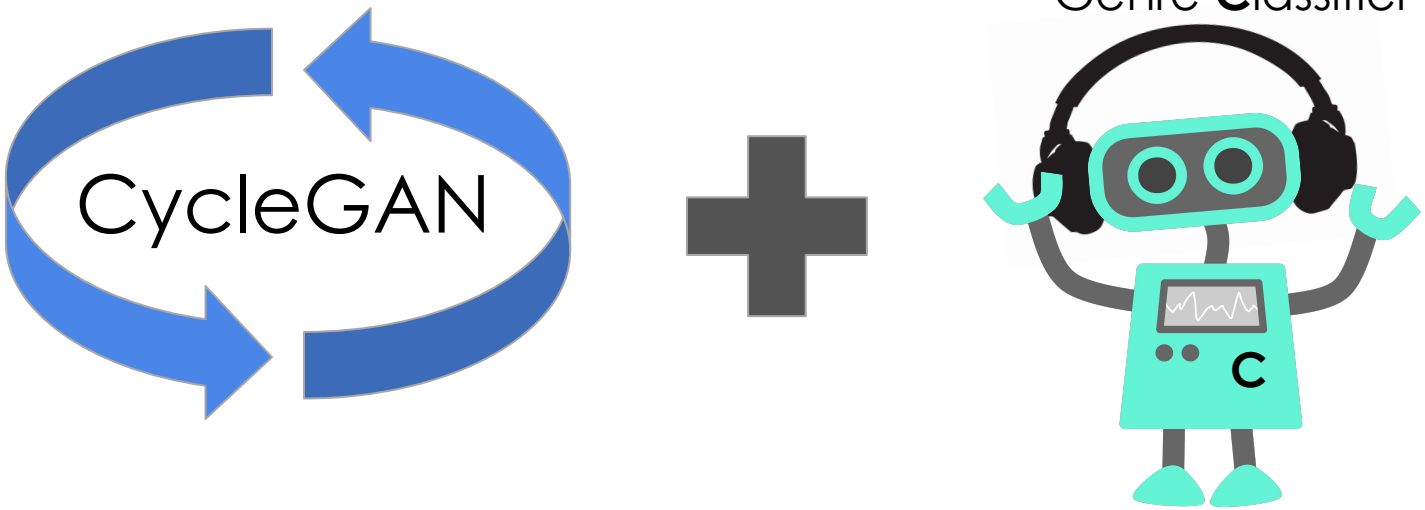
Gino Brunner, *Mazda Moayeri*, **Oliver Richter**, *Roger Wattenhofer, Chi Zhang**

**Alphabetical order*

ETH Zurich – Distributed Computing – www.disco.ethz.ch



Jazz, Classic, Pop…?

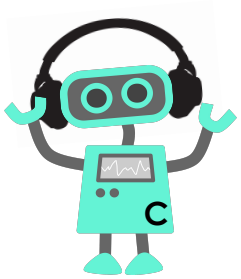

CycleGAN
Genre Classifier
C

Self Attention

Spectral Normalization

$$A = \begin{bmatrix} 2 & 1 & -2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad Det\,|A - \boldsymbol{\lambda I}\,|$$

CycleGAN

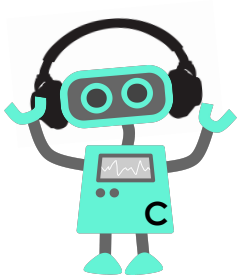


Self Attention

Spectral Normalization

$$A = \begin{bmatrix} 2 & 1 & -2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad Det\,|A - \boldsymbol{\lambda I}\,|$$

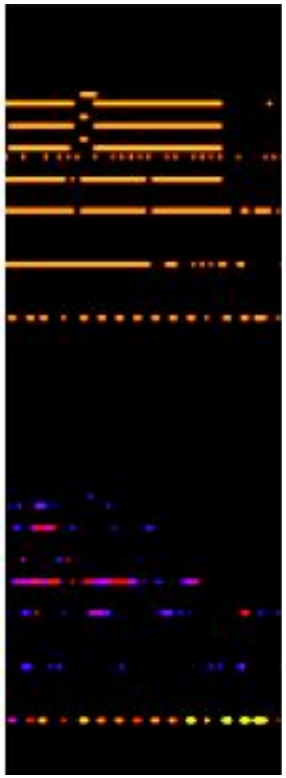
CycleGAN

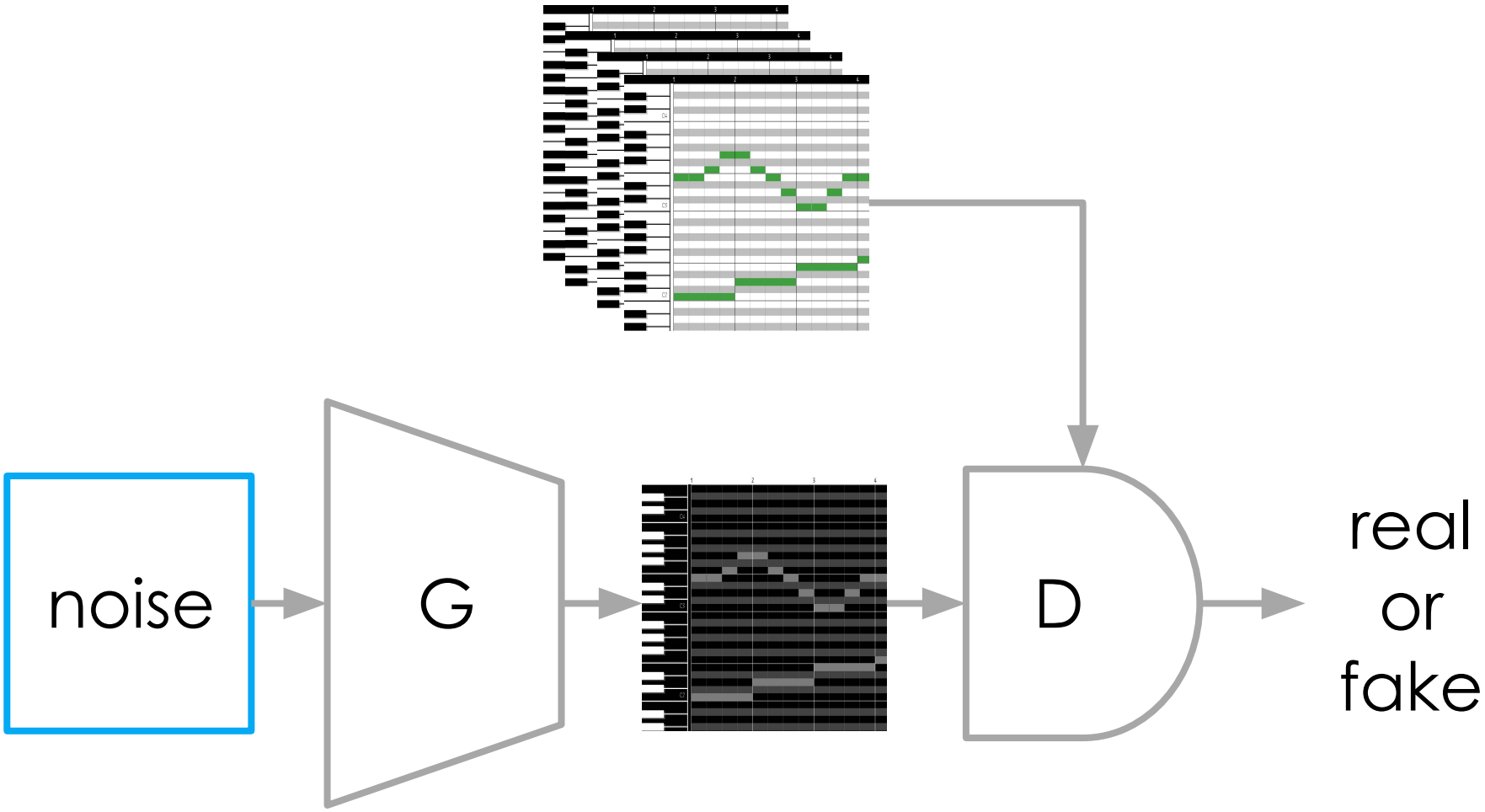


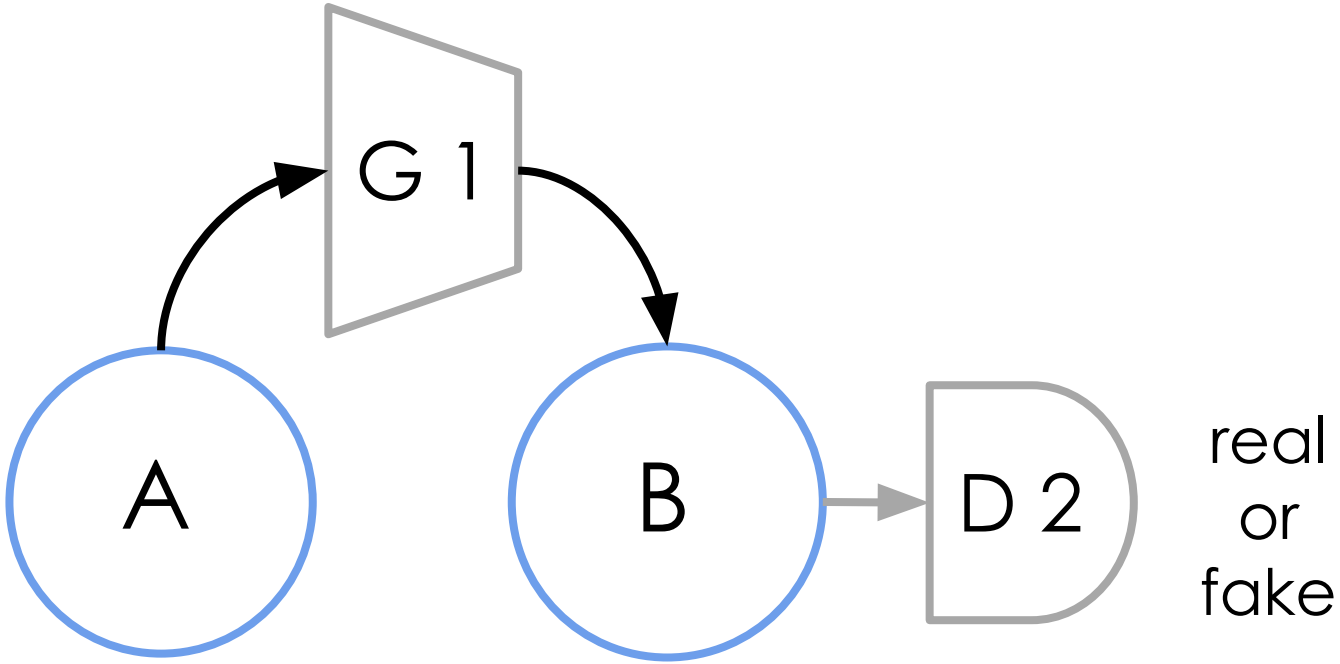
Self Attention

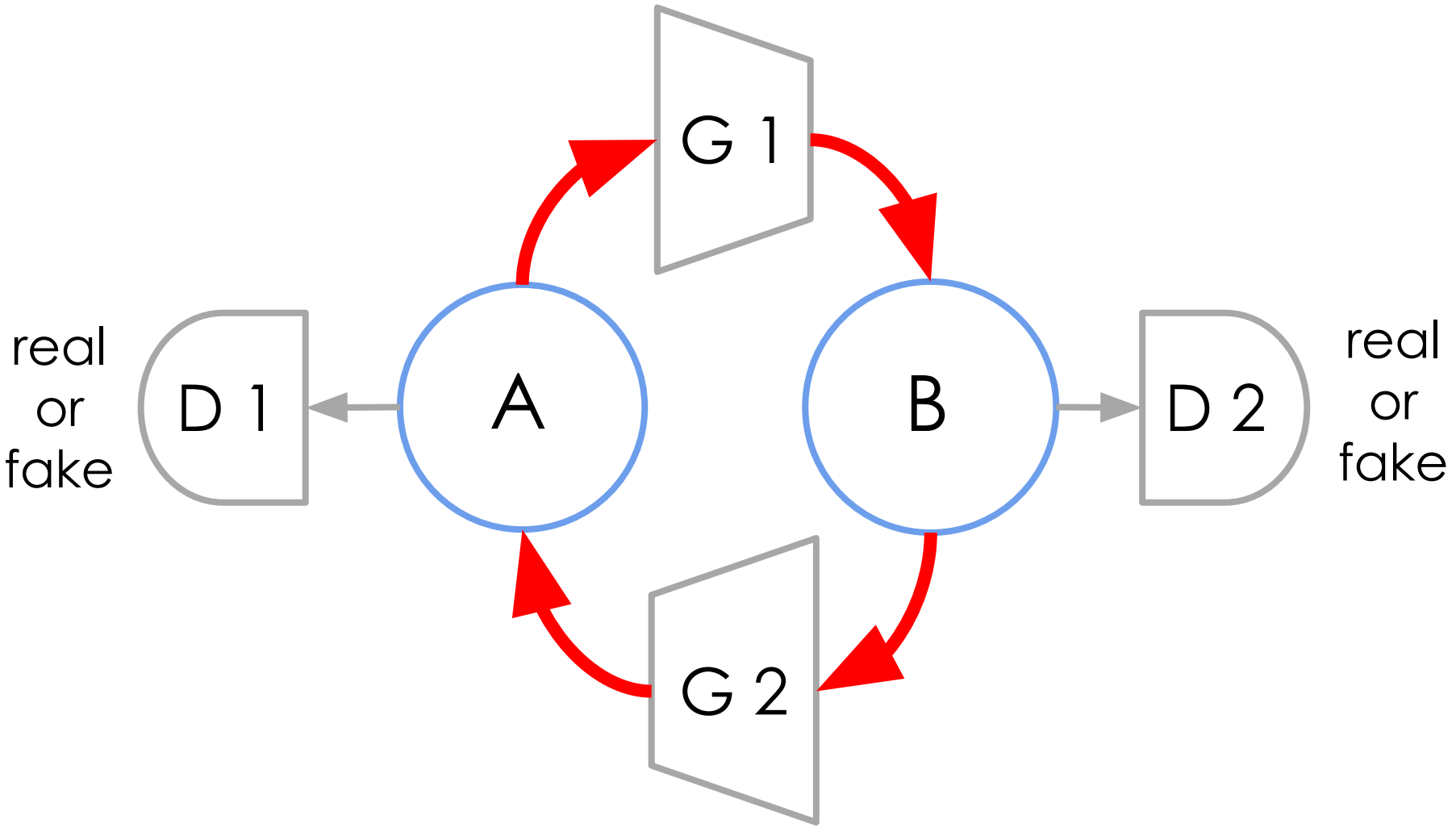
Spectral Normalization

$$A = \begin{bmatrix} 2 & 1 & -2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad Det\,|A - \boldsymbol{\lambda I}\,|$$

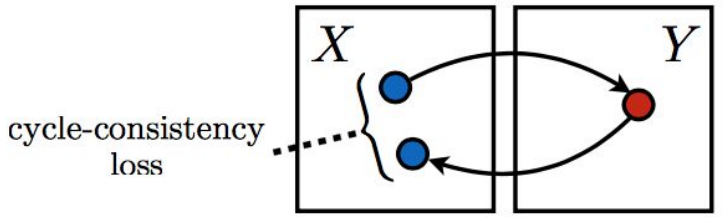


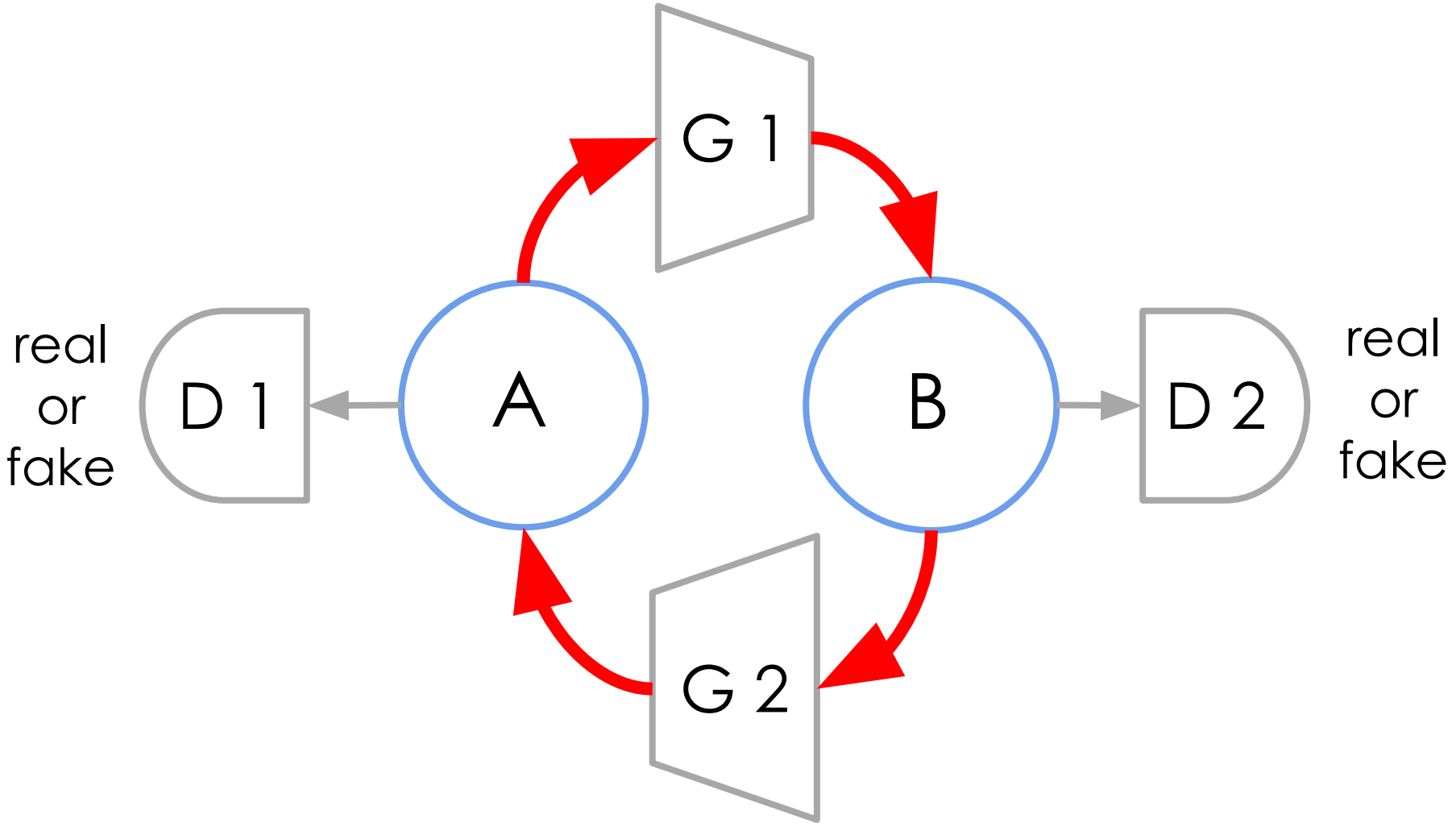

noise
G
D
real
or
fake


G 1
A
B
D 2
real
or
fake


G 1
real
or
fake
D 1
A
B
D 2
real
or
fake
G 2

Cycle Consistency!

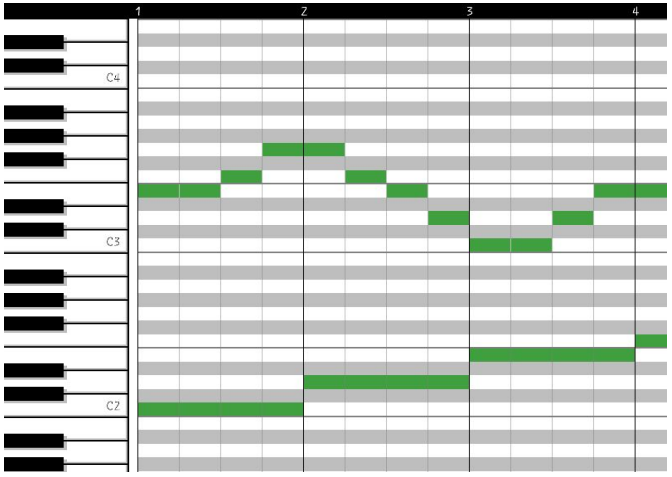


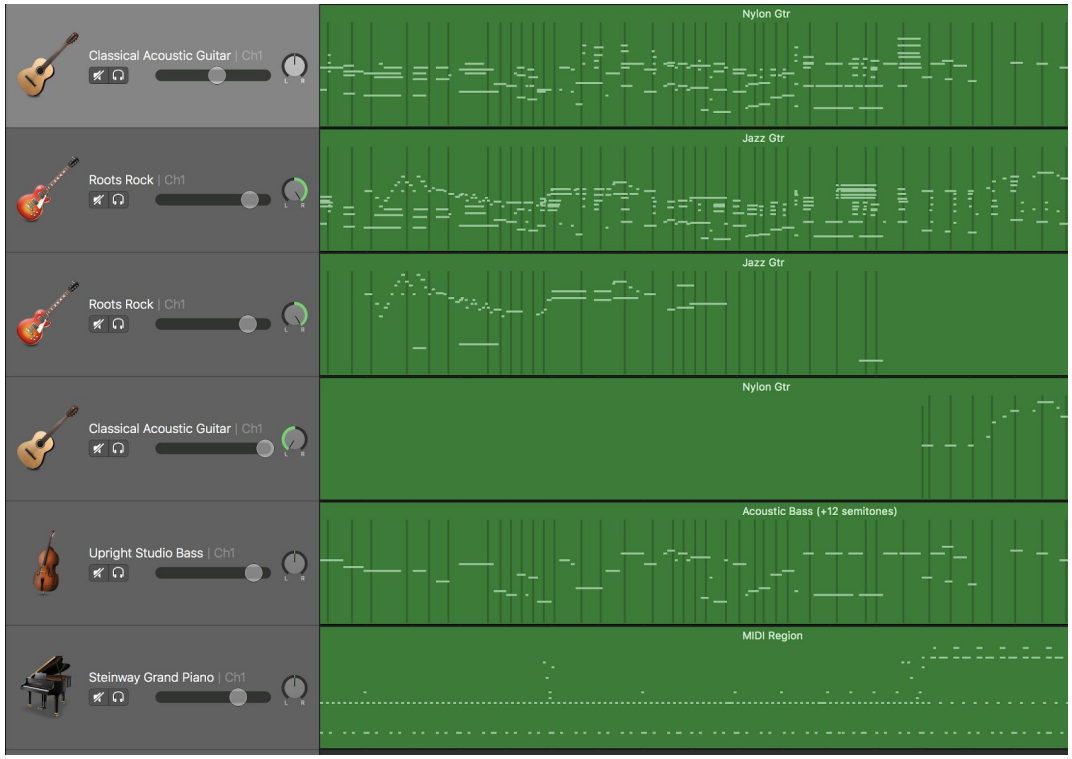


≈

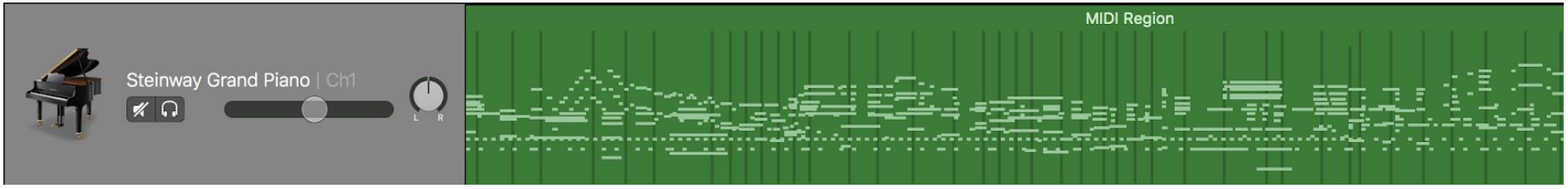
MIDI

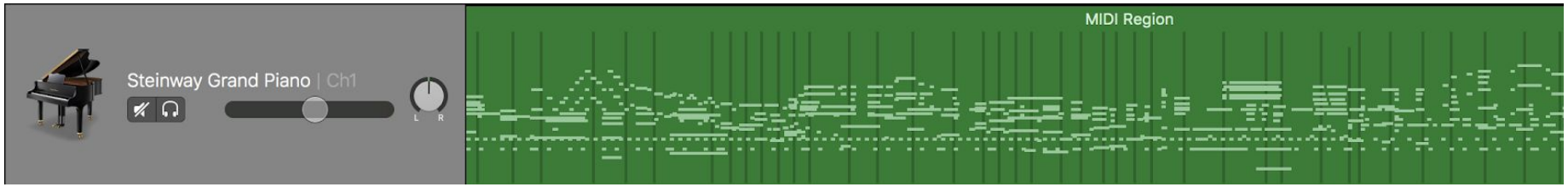
Piano Roll





Discard Drums and Merge

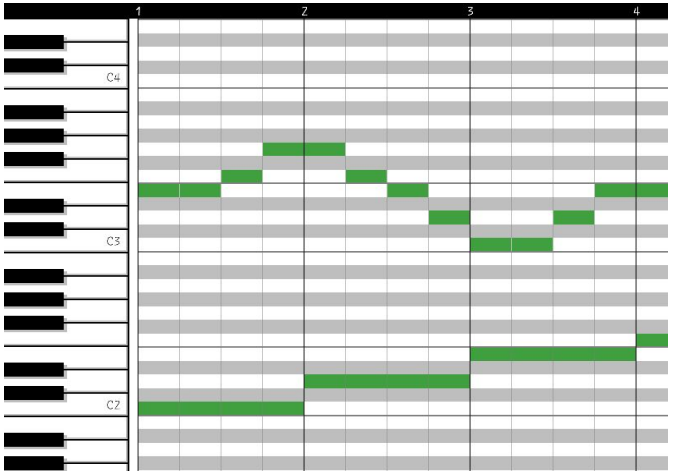




Discard velocity

Piano Roll

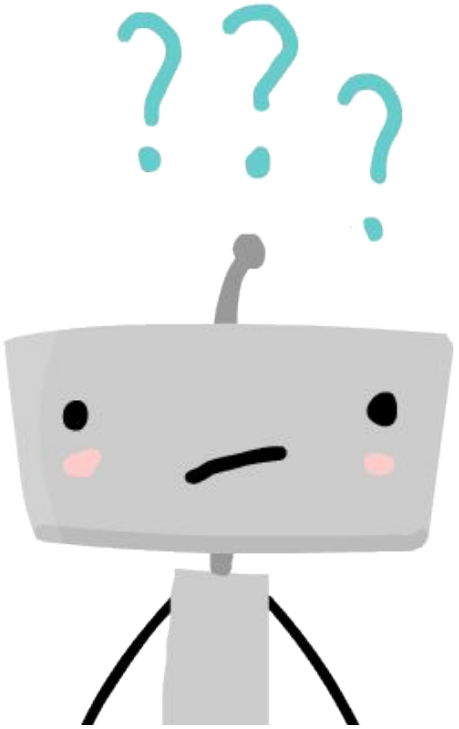
84 pitches

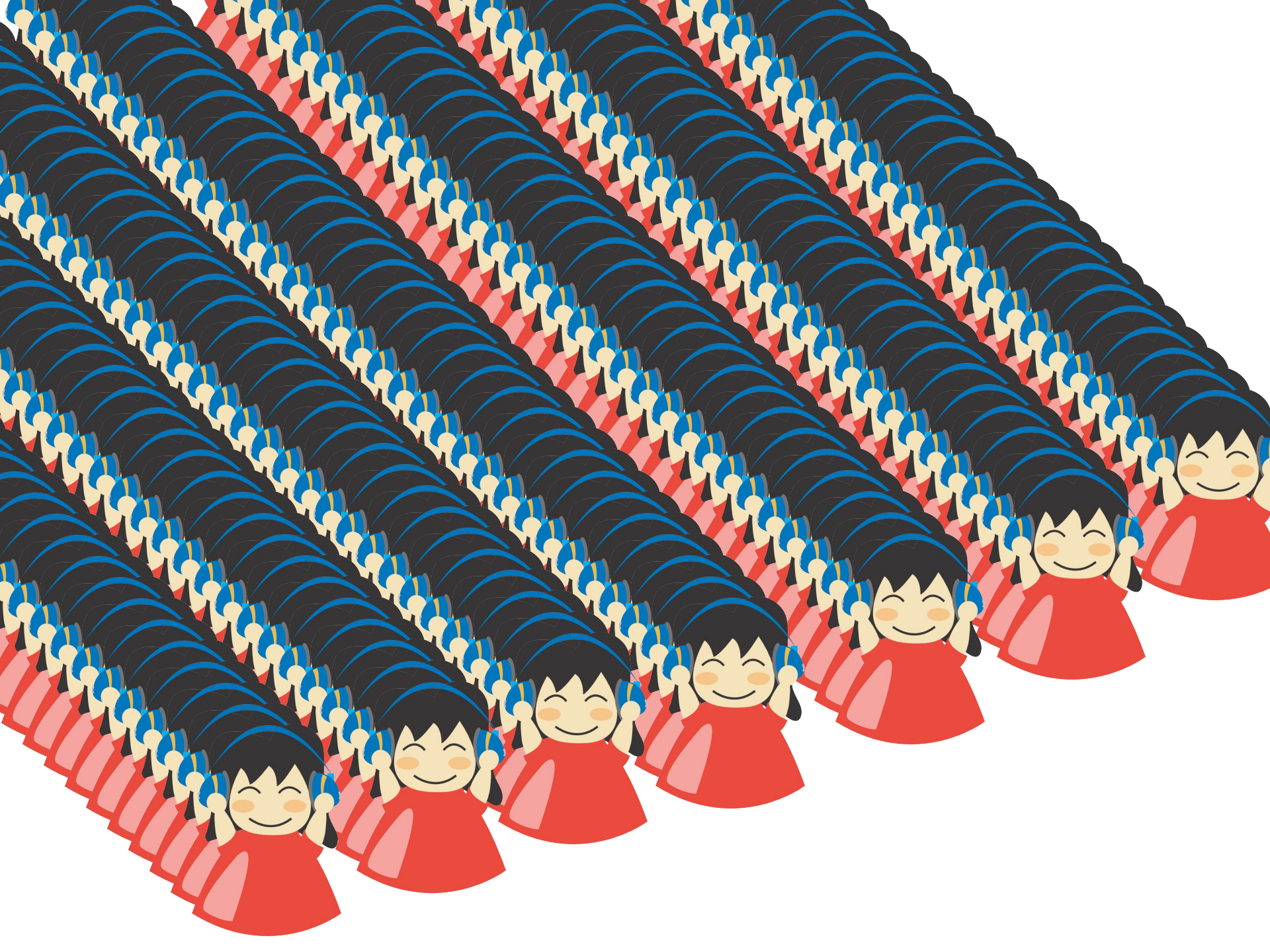


64 beats = 4 bars = 1 phrase

	Jazz	Classic	Pop
Songs	559	2714	1069
Phrases	**12341**	16545	20780

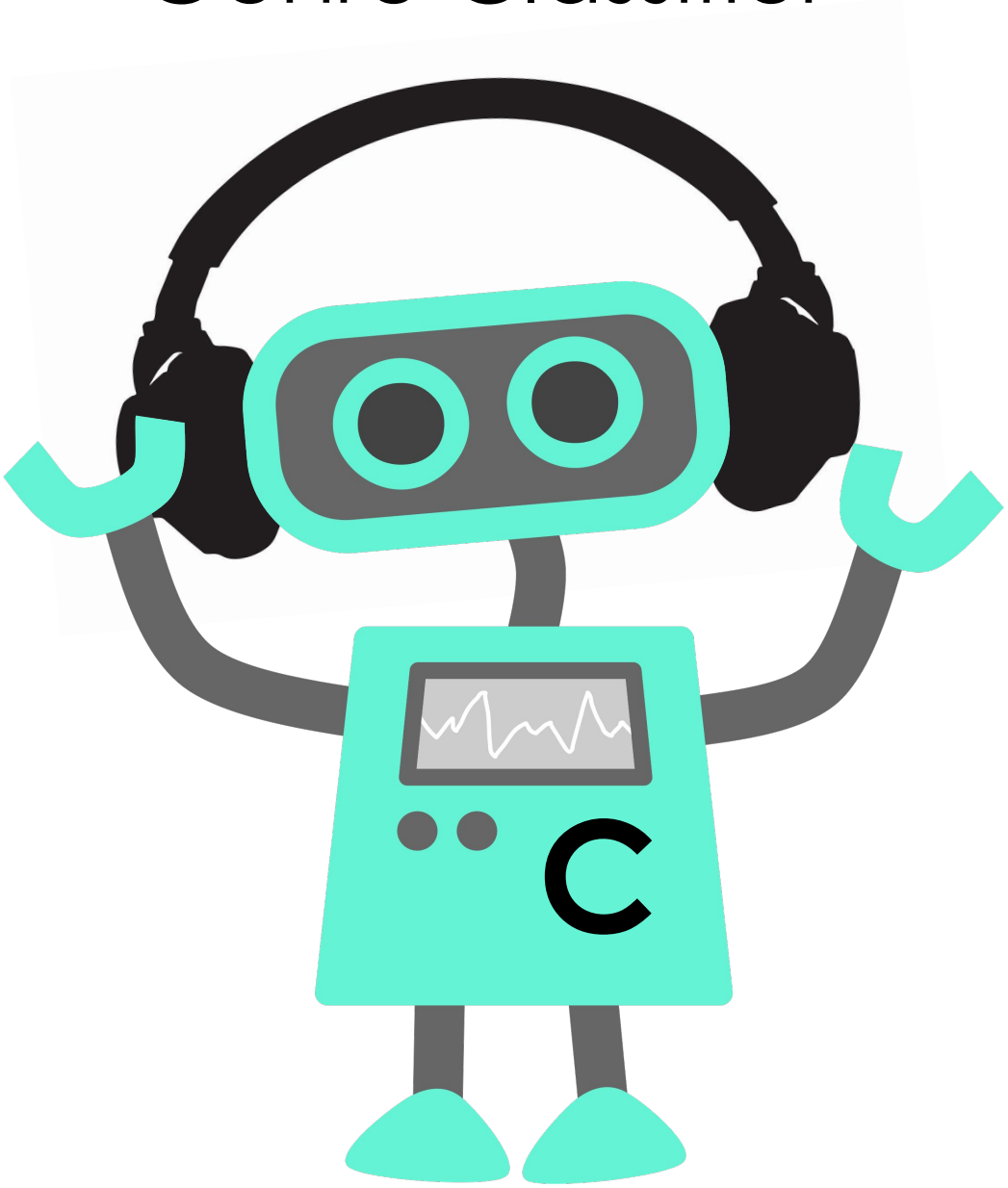

evaluation

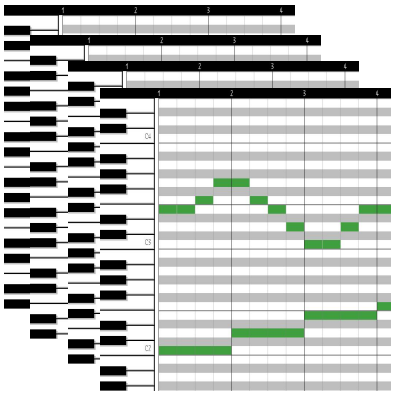


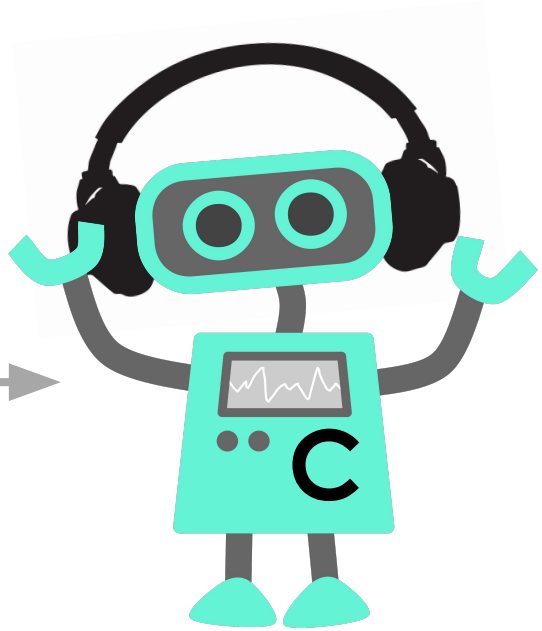


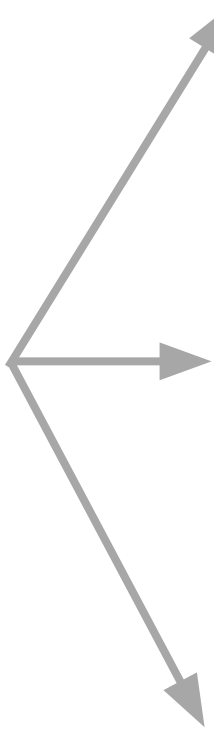
Let me do it!

Genre **C**lassifier

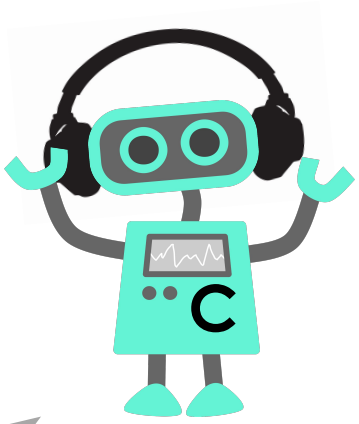




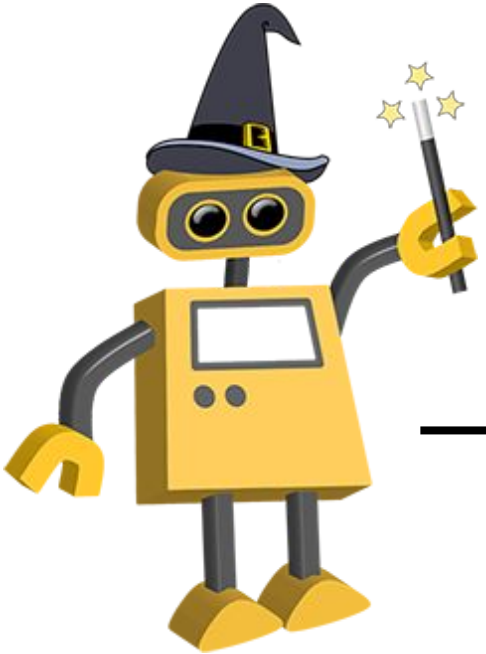




90% Jazz!

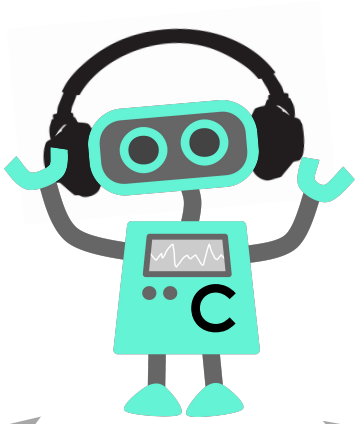


Before



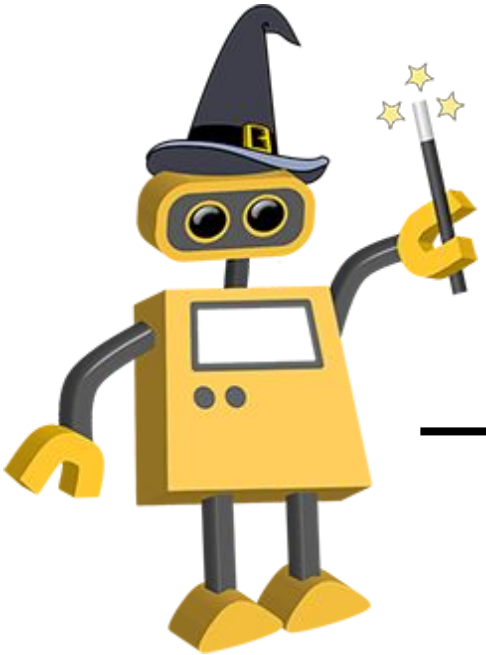
After

90% Jazz!



80% Classic!

Before



After

1. classic: 86%

2. classic -> pop: 16%

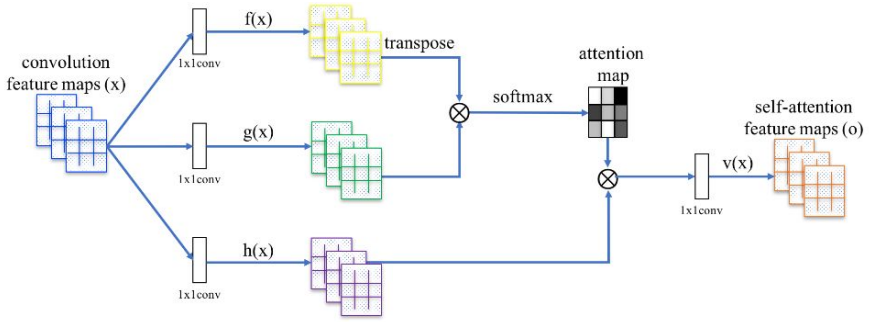
3. classic -> pop -> classic: 80%

$$s_1 = \frac{1}{2}\big((86\% - 16\%) + (80\% - 16\%)\big) = 67\%$$

$$s_2 = 60\%$$

$$s_{tot}^{D} = \frac{1}{2}(s_1 + s_2) = 63.5\%$$

Self-Attention



Spectral Normalization

Algorithm 1 SGD with spectral normalization

- Initialize $\tilde{\boldsymbol{u}}_l \in \mathcal{R}^{d_l}$ for $l = 1, \ldots, L$ with a random vector (sampled from isotropic distribution).
- For each update and each layer l:
  1. Apply power iteration method to a unnormalized weight W^l:

$$\tilde{\boldsymbol{v}}_l \leftarrow (W^l)^{\mathrm{T}} \tilde{\boldsymbol{u}}_l / \|(W^l)^{\mathrm{T}} \tilde{\boldsymbol{u}}_l\|_2 \tag{20}$$

$$\tilde{\boldsymbol{u}}_l \leftarrow W^l \tilde{\boldsymbol{v}}_l / \|W^l \tilde{\boldsymbol{v}}_l\|_2 \tag{21}$$

  2. Calculate $\bar{W}_{\mathrm{SN}}$ with the spectral norm:

$$\bar{W}^l_{\mathrm{SN}}(W^l) = W^l / \sigma(W^l), \text{ where } \sigma(W^l) = \tilde{\boldsymbol{u}}_l^{\mathrm{T}} W^l \tilde{\boldsymbol{v}}_l \tag{22}$$

  3. Update W^l with SGD on mini-batch dataset $\mathcal{D}_M$ with a learning rate α:

$$W^l \leftarrow W^l - \alpha \nabla_{W^l} \ell(\bar{W}^l_{\mathrm{SN}}(W^l), \mathcal{D}_M) \tag{23}$$

	J vs. P	C vs. P	J vs. C
Baseline	28.49%	**64.62%**	57.64%
SN	32.16%	61.88%	63.98%
SA	**44.85%**	59.35%	63.56%
SN+SA	33.23%	53.07%	**66.76%**

	J vs. P	C vs. P	J vs. C
Baseline	28.49%	**64.62%**	57.64%
SN	32.16%	61.88%	63.98%
SA	**44.85%**	59.35%	63.56%
SN+SA	33.23%	53.07%	**66.76%**

Need to train more models!


content


content

3 added + **3** removed

= **6** total

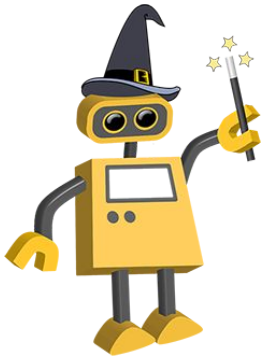
1 → 2 → 3

Baseline

SN

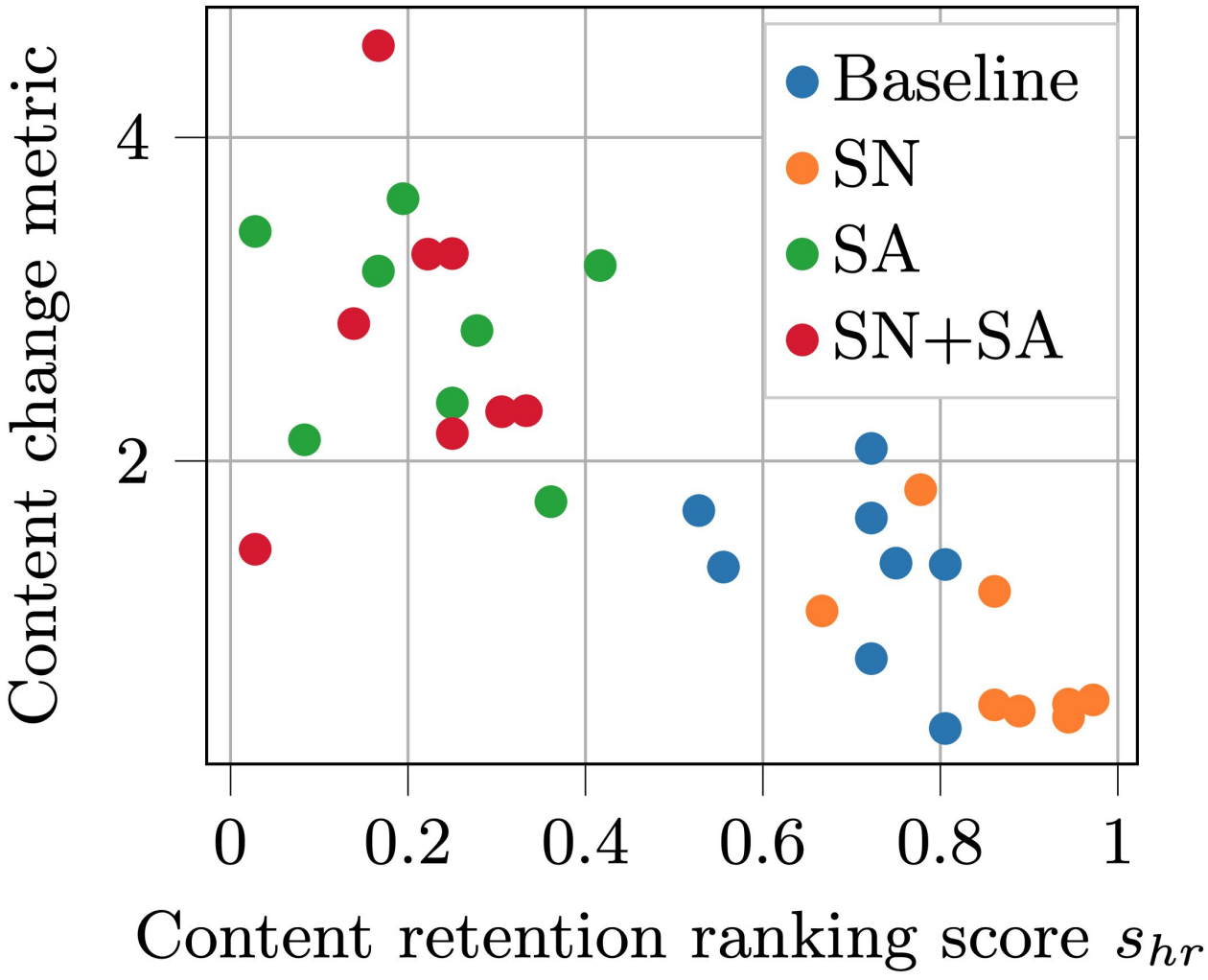
SA

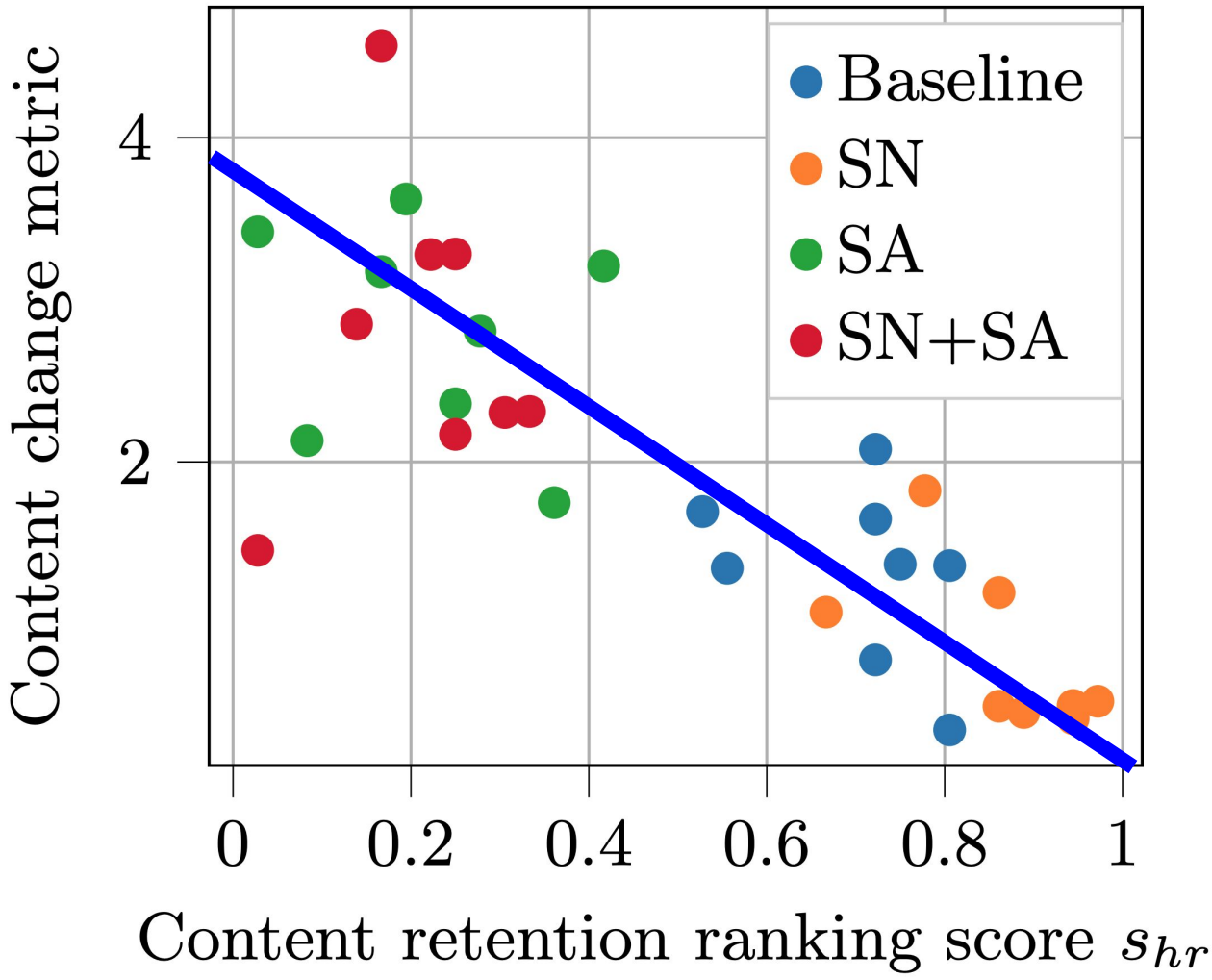
SN+SA



Recognizable?

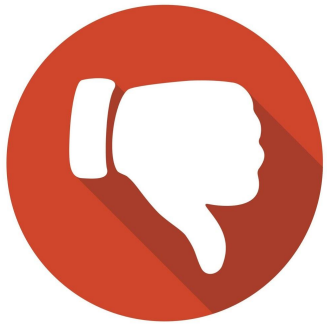
Fidelity?


Content change metric
4
2
0
0.2
0.4
0.6
0.8
1
Content retention ranking score s_{hr}
Baseline
SN
SA
SN+SA



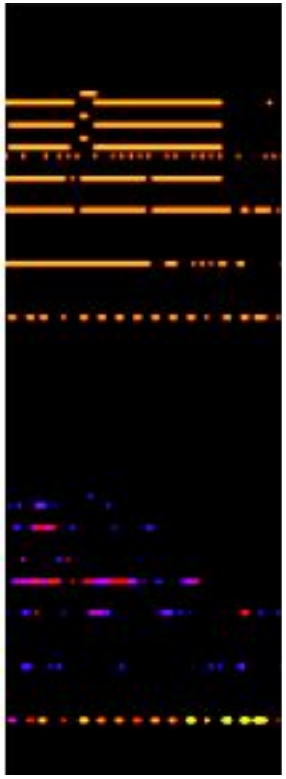
Pearson correlation: **-0.805**

	Baseline		SN	
	$C \to P$	$P \to C$	$C \to P$	$P \to C$
added	0.82 ± 0.45	0.28 ± 0.17	0.57 ± 0.38	0.08 ± 0.09
removed	0.27 ± 0.17	0.46 ± 0.12	0.91 ± 0.07	0.3 ± 0.11
total	1.10 ± 0.5	0.75 ± 0.25	0.66 ± 0.38	0.38 ± 0.16
	SA		SN + SA	
added	1.47 ± 0.41	0.85 ± 0.79	1.78 ± 1.33	0.72 ± 0.66
removed	0.95 ± 0.04	0.95 ± 0.04	0.95 ± 0.05	0.93 ± 0.05
total	2.42 ± 0.42	1.79 ± 0.78	2.73 ± 1.33	1.65 ± 0.66

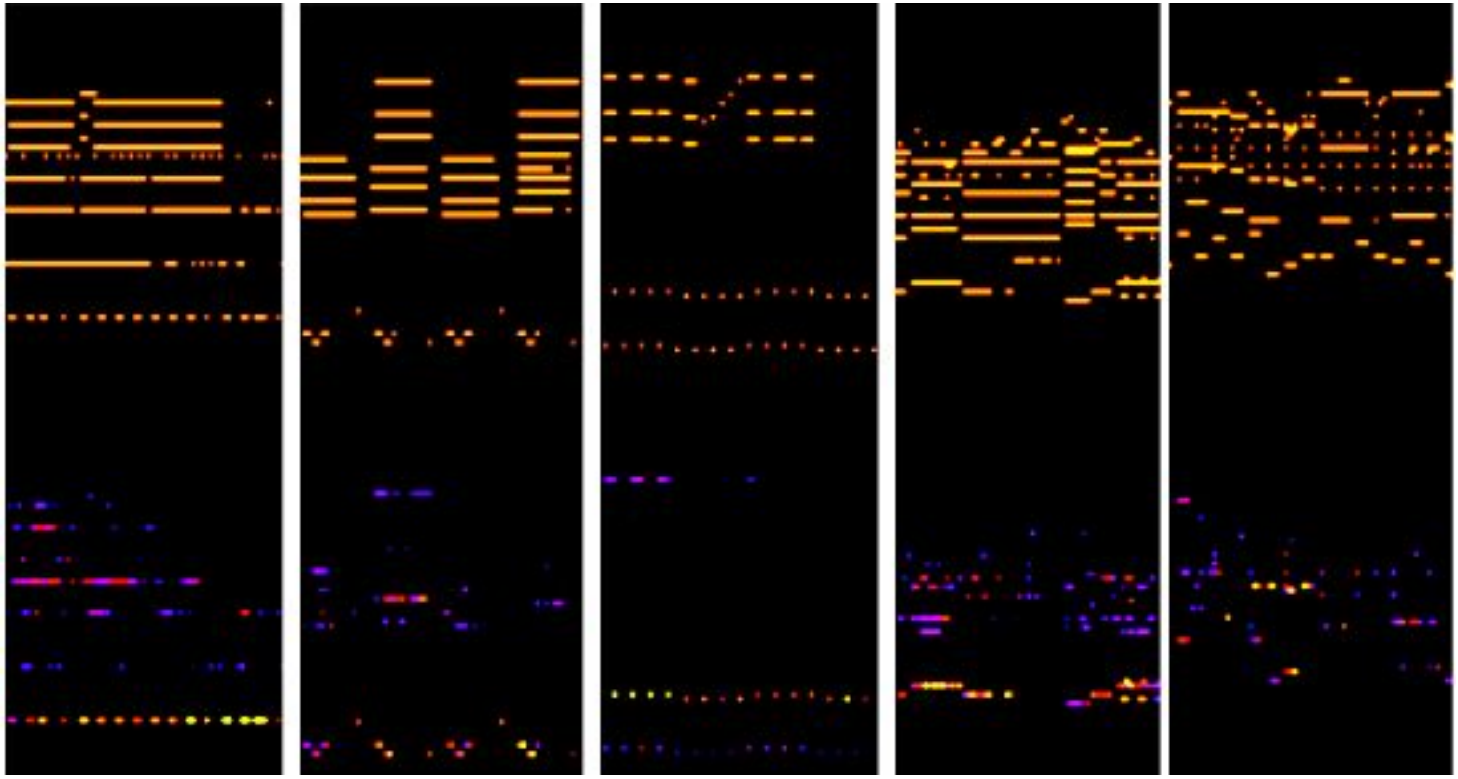


1. SN	0.6
2. SN+SA	0.51
3. Baseline	0.47
4. SA	0.42

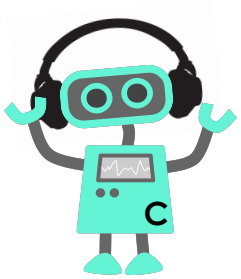
1.0: Always ranked best
0.0: Always ranked worst



Jazz Samples



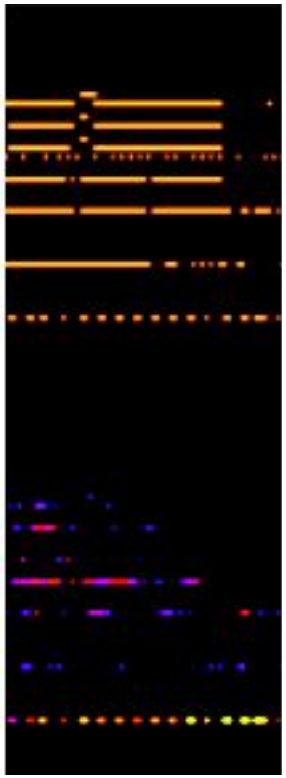
CycleGAN



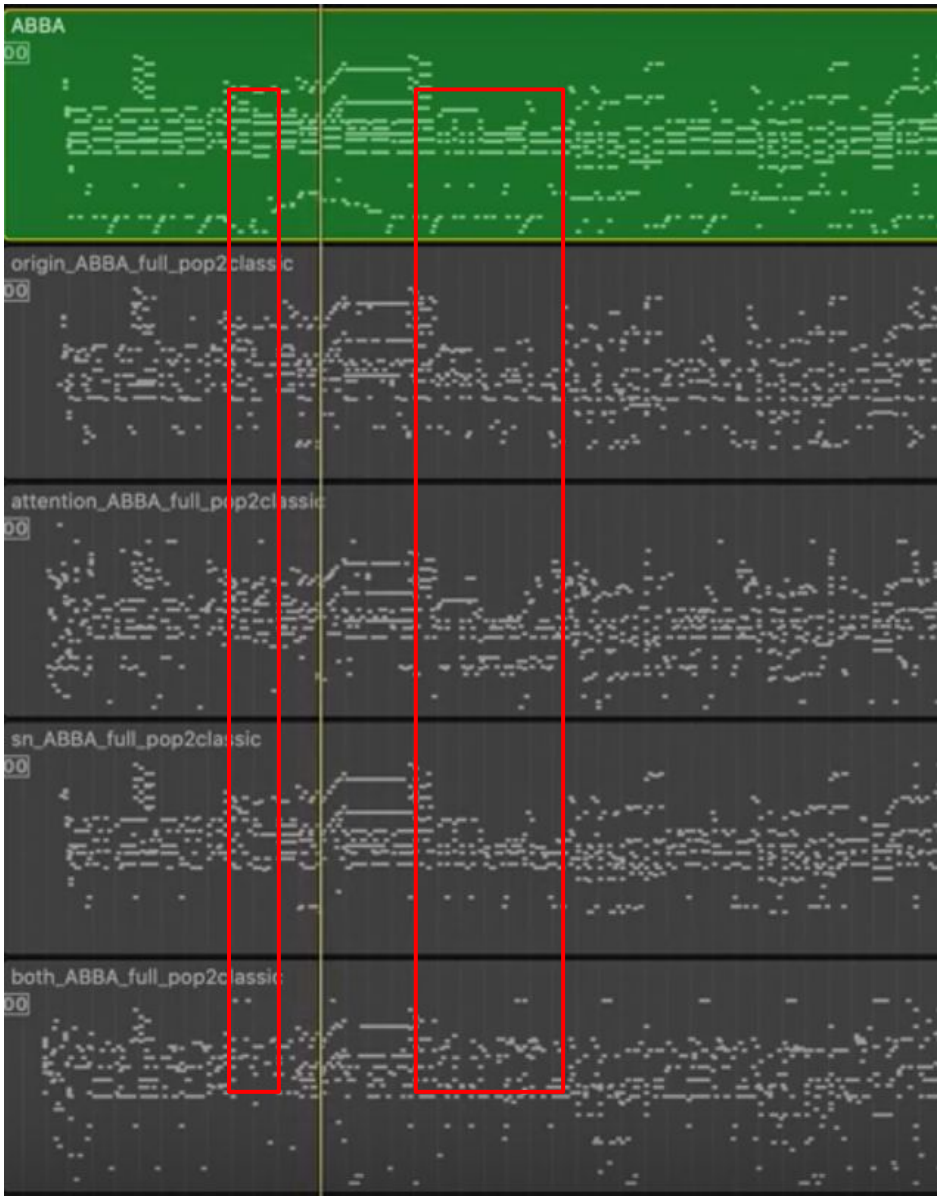
Self Attention

Spectral Normalization

$$A = \begin{bmatrix} 2 & 1 & -2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad Det\,|A - \boldsymbol{\lambda I}\,|$$



Backup Slides



original

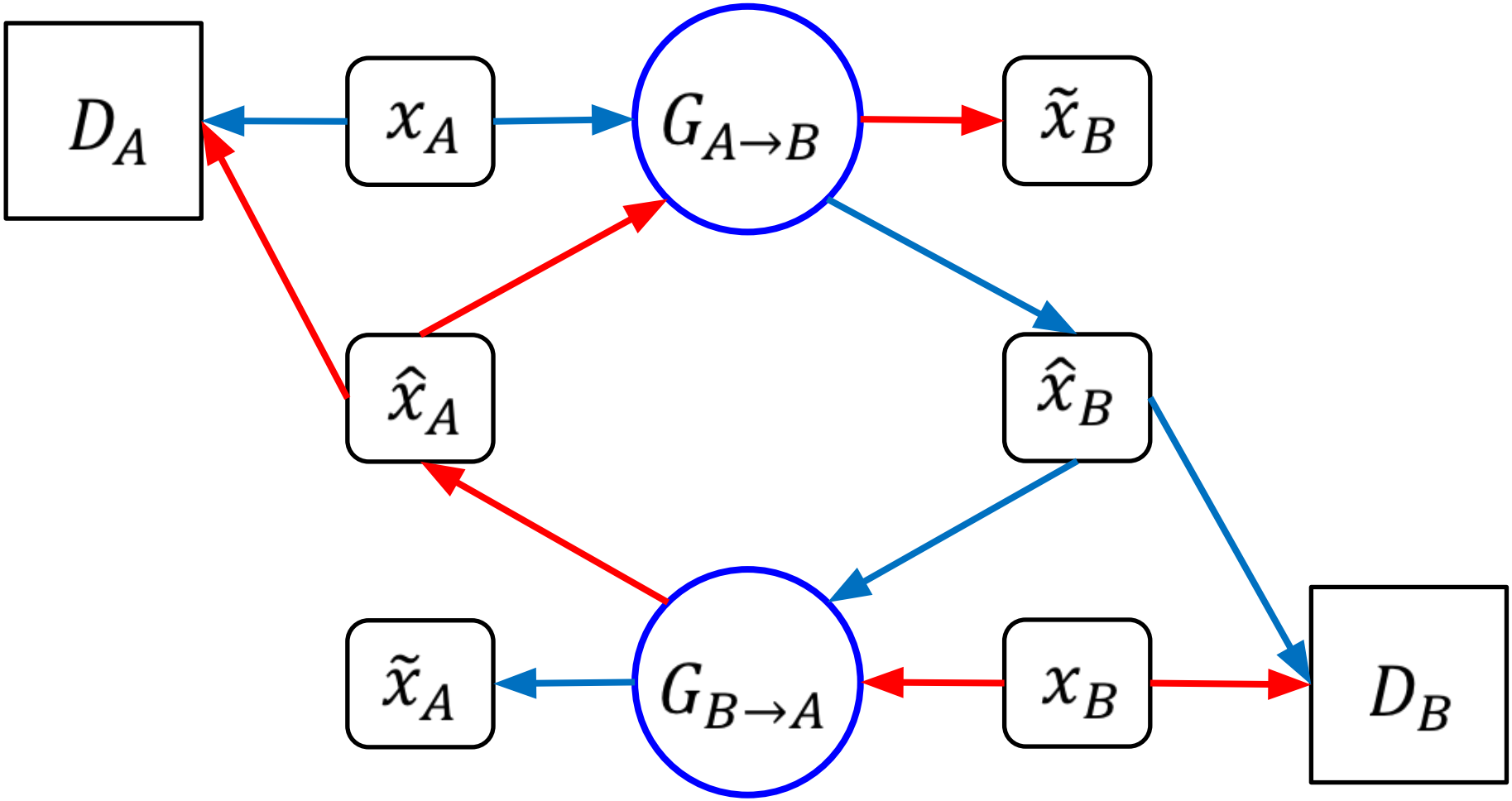
Baseline

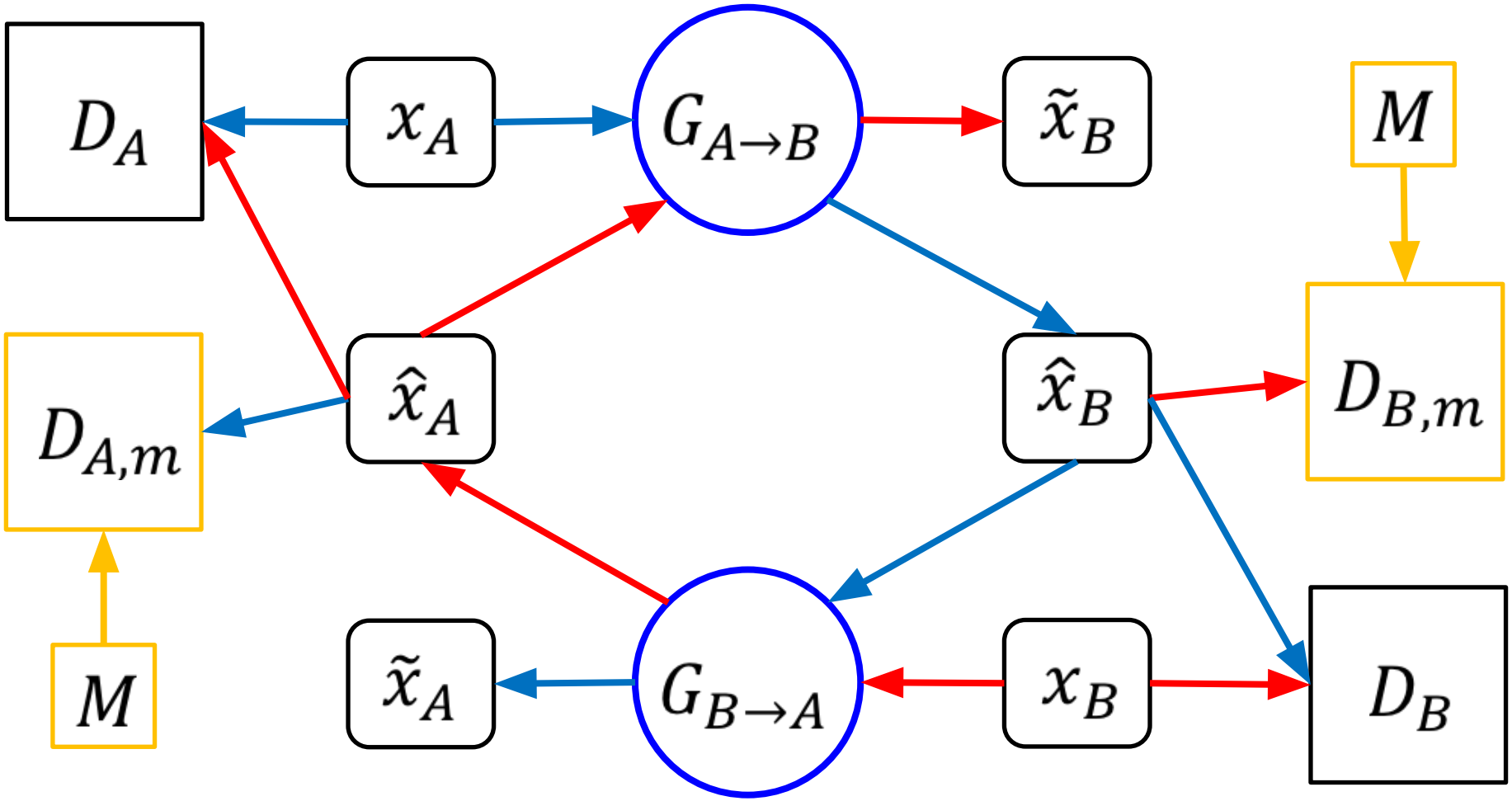
Self-Attention

Spectral Normalization

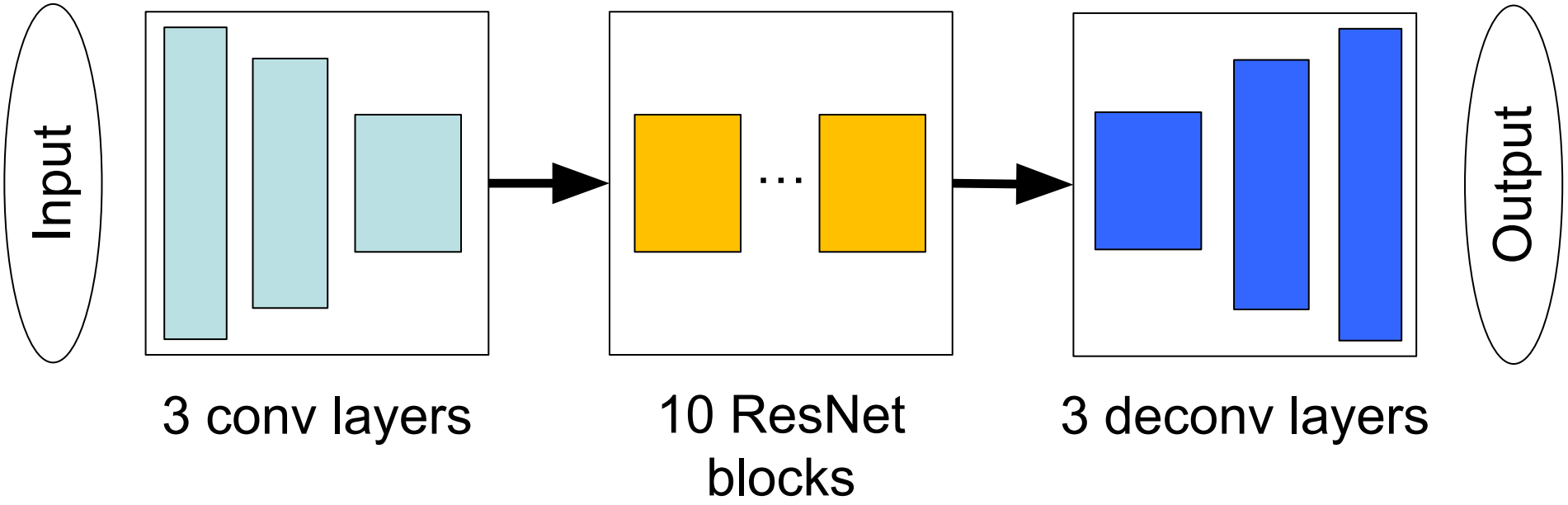
SN + SA

$$s_{hr}^{M} = \frac{1}{N} \sum_{r=1}^{K} \#\{\text{rank of } M = \text{r}\} \frac{K-r}{K-1}$$

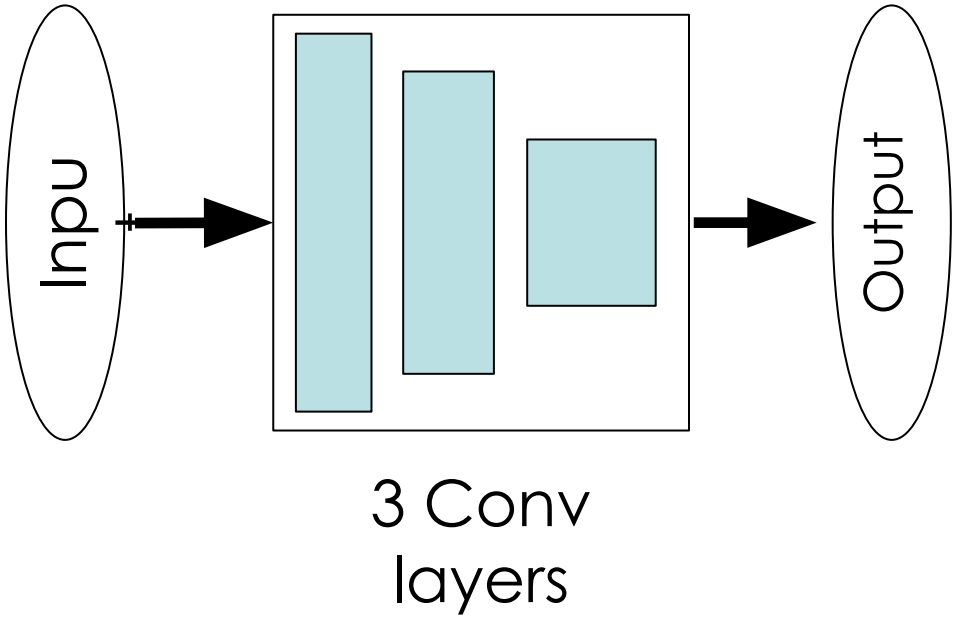

D_A
x_A
$G_{A\to B}$
$\tilde{x}_B$
$\hat{x}_A$
$\hat{x}_B$
$\tilde{x}_A$
$G_{B\to A}$
x_B
D_B


D_A
x_A
$G_{A\to B}$
$\tilde{x}_B$
M
$D_{A,m}$
$\hat{x}_A$
$\hat{x}_B$
$D_{B,m}$
M
$\tilde{x}_A$
$G_{B\to A}$
x_B
D_B

Generators

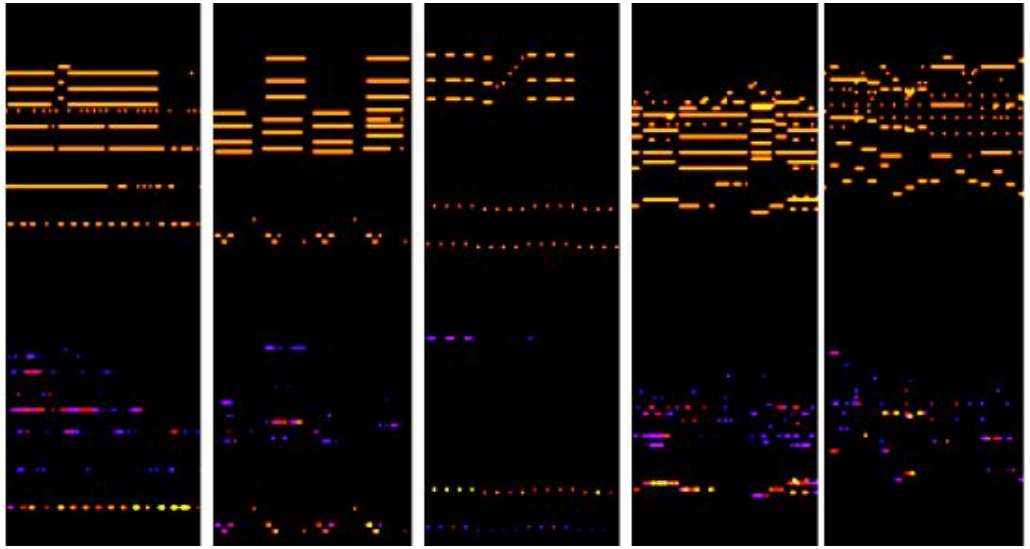


Discriminators

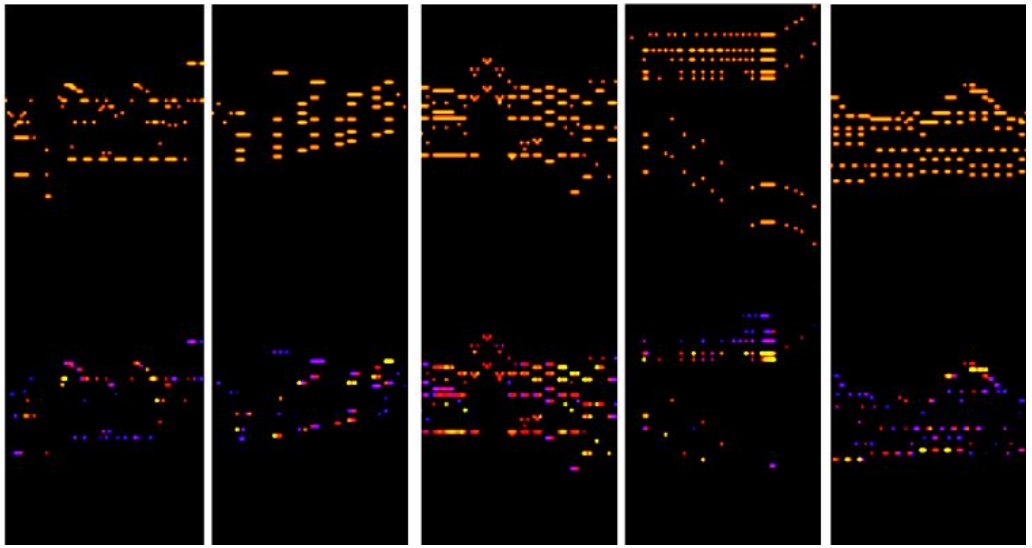


GRADIENT BASED ATTRIBUTION

Jazz samples



Classical samples



Original samples with saliency maps below.