MUSIC SKETCHNET

CONTROLLABLE MUSIC GENERATION VIA FACTORIZED REPRESENTATION OF PITCH AND RHYTHM

Autores

- Ke Chen,
- Cheng-i Wang,
- Taylor Berg-Kirkpatrick,
- Shlomo Dubnov

- CREL, Music Department (Center of Research in Entertaining and Learning)
- University of California, San Diego
- Smule, Inc

Motivación

Generación automática de música para estudiar y expandir la expresividad/creatividad humana

Permitir a usuarios controlar de manera flexible e intuitiva el resultado de la generación automática de música.

Permitir especificar ideas musicales parciales en términos de representaciones de ritmo y altura incompletas.

Music Sketching Task

Completar una pieza musical generando una secuencia de compases faltantes dado un contexto circundante.

Para esto es necesario solucionar:

1. ¿Cómo representar ideas musicales?
2. ¿Como generar material nuevo dado el contexto pasado y futuro?
3. ¿Cómo procesar input de usuario e integrarlo al sistema?

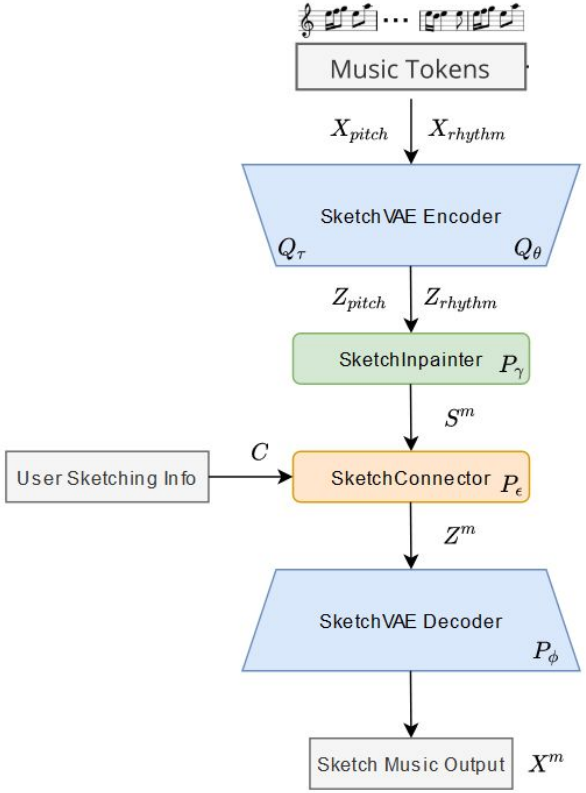
Solución Propuesta

Solución Propuesta

Music SketchNet resuelve estos problemas a través de tres estructuras:

1. SketchVAE
2. SketchInpainter
3. SketchConnector

Solución Propuesta



Formalmente

El framework propuesto es un modelo de la probabilidad conjunta del contenido musical faltante Xm, condicionado al pasado, futuro y input de bosquejo del usuario.

Formalmente

$$P_{\phi,\varepsilon,\gamma,\theta,\tau}(X^m, Z, S | X^p, X^f, C) =$$

$$P_\phi(X^m|Z^m) \qquad \text{(SketchVAE Decoder)}$$

$$* \; P_\varepsilon(Z^m|S^m, C) \qquad \text{(SketchConnector)}$$

$$* \; P_\gamma(S^m_{pitch}|Z^p_{pitch}, Z^f_{pitch}) \qquad \text{(SketchInpainter)}$$

$$* \; P_\gamma(S^m_{rhythm}|Z^p_{rhythm}, Z^f_{rhythm}) \qquad \text{(SketchInpainter)}$$

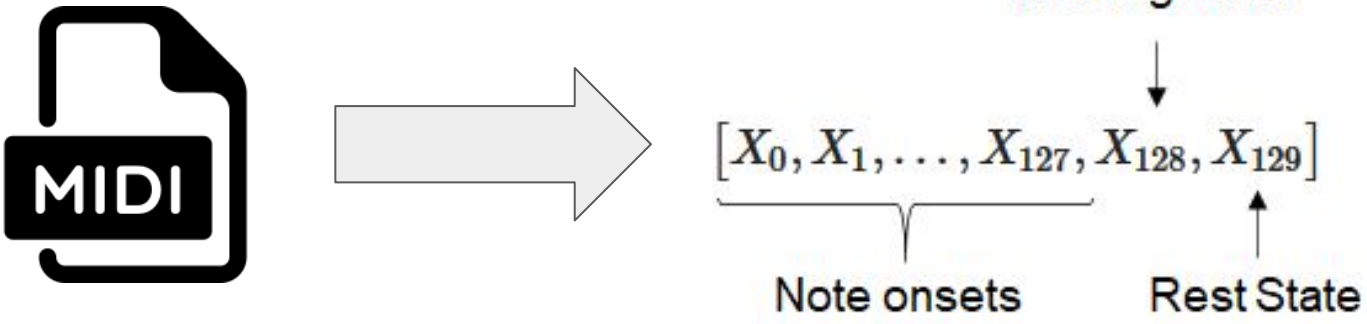
$$* \; Q_\theta(Z^p_{pitch}, Z^f_{pitch}|X^p_{pitch}, X^f_{pitch})$$

$$* \; Q_\tau(Z^p_{rhythm}, Z^f_{rhythm}|X^p_{rhythm}, X^f_{rhythm})$$

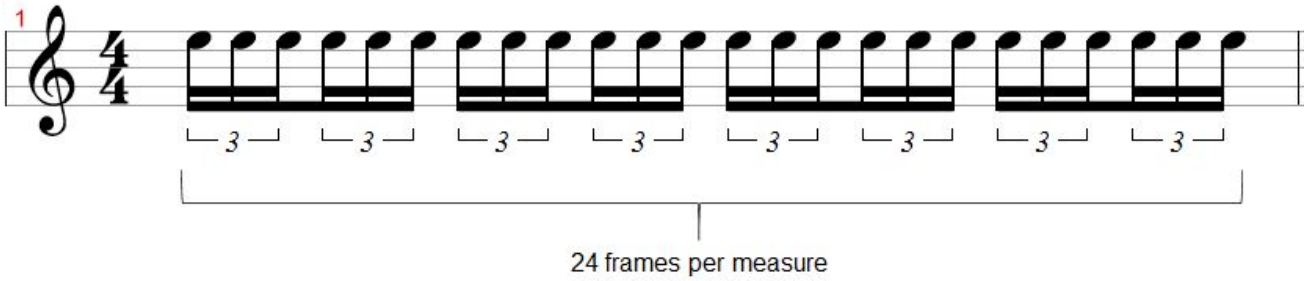
$$\text{(SketchVAE Encoders)}$$

Encoding

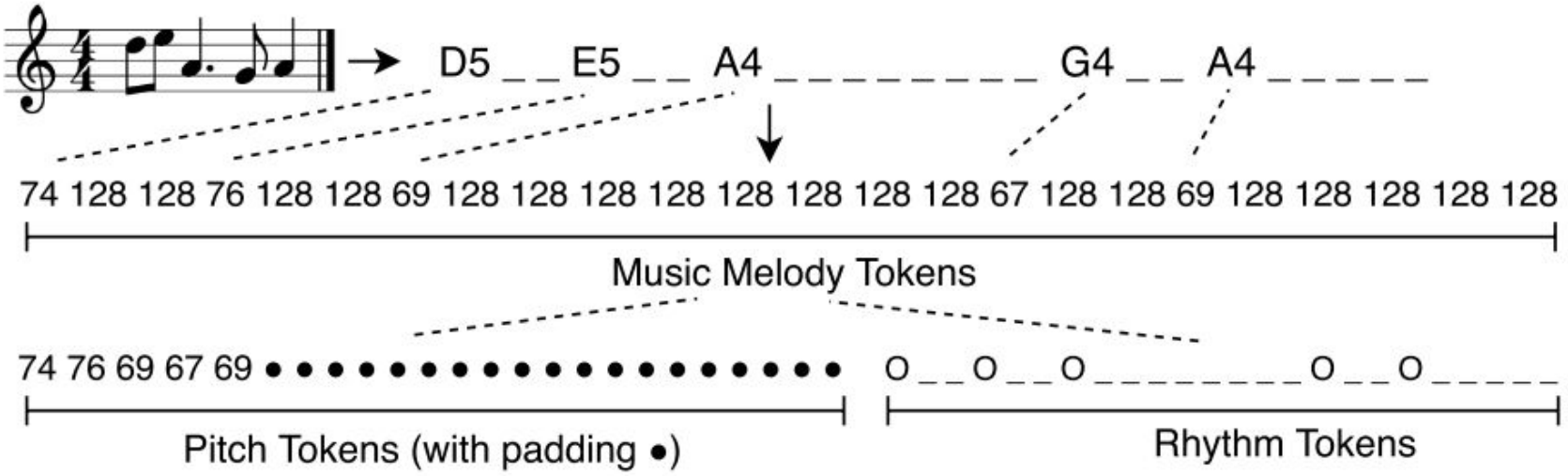
Encoding


MIDI
Holding State
$[X_0, X_1, \ldots, X_{127}, X_{128}, X_{129}]$
Note onsets
Rest State

Encoding



Encoding


D5 _ _ E5 _ _ A4 _ _ _ _ _ _ _ _ _ G4 _ _ A4 _ _ _ _ _
74 128 128 76 128 128 69 128 128 128 128 128 128 128 128 67 128 128 69 128 128 128 128 128
Music Melody Tokens
74 76 69 67 69 •
O _ _ O _ _ O _ _ _ _ _ _ _ _ O _ _ O _ _ _ _ _
Pitch Tokens (with padding •)
Rhythm Tokens

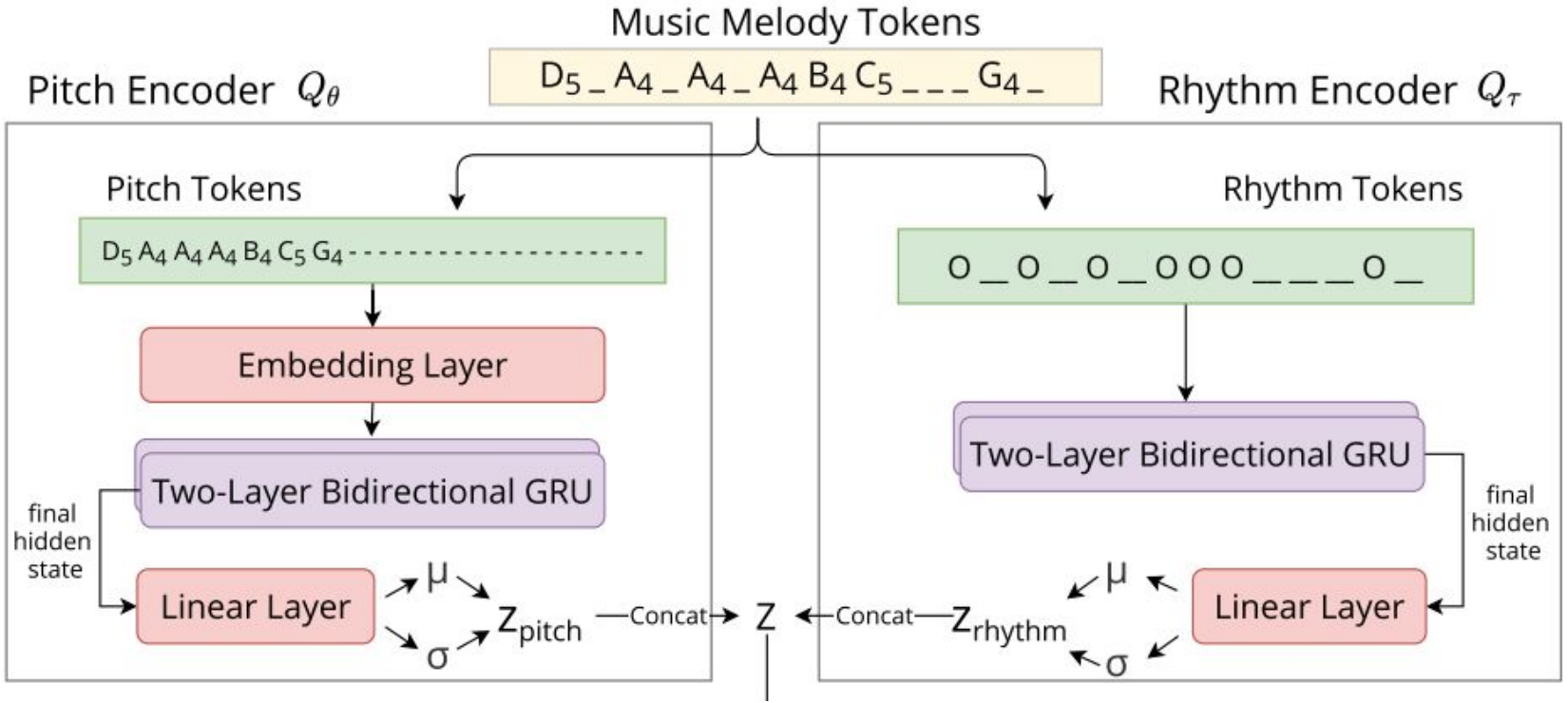
SketchVAE

SketchVAE

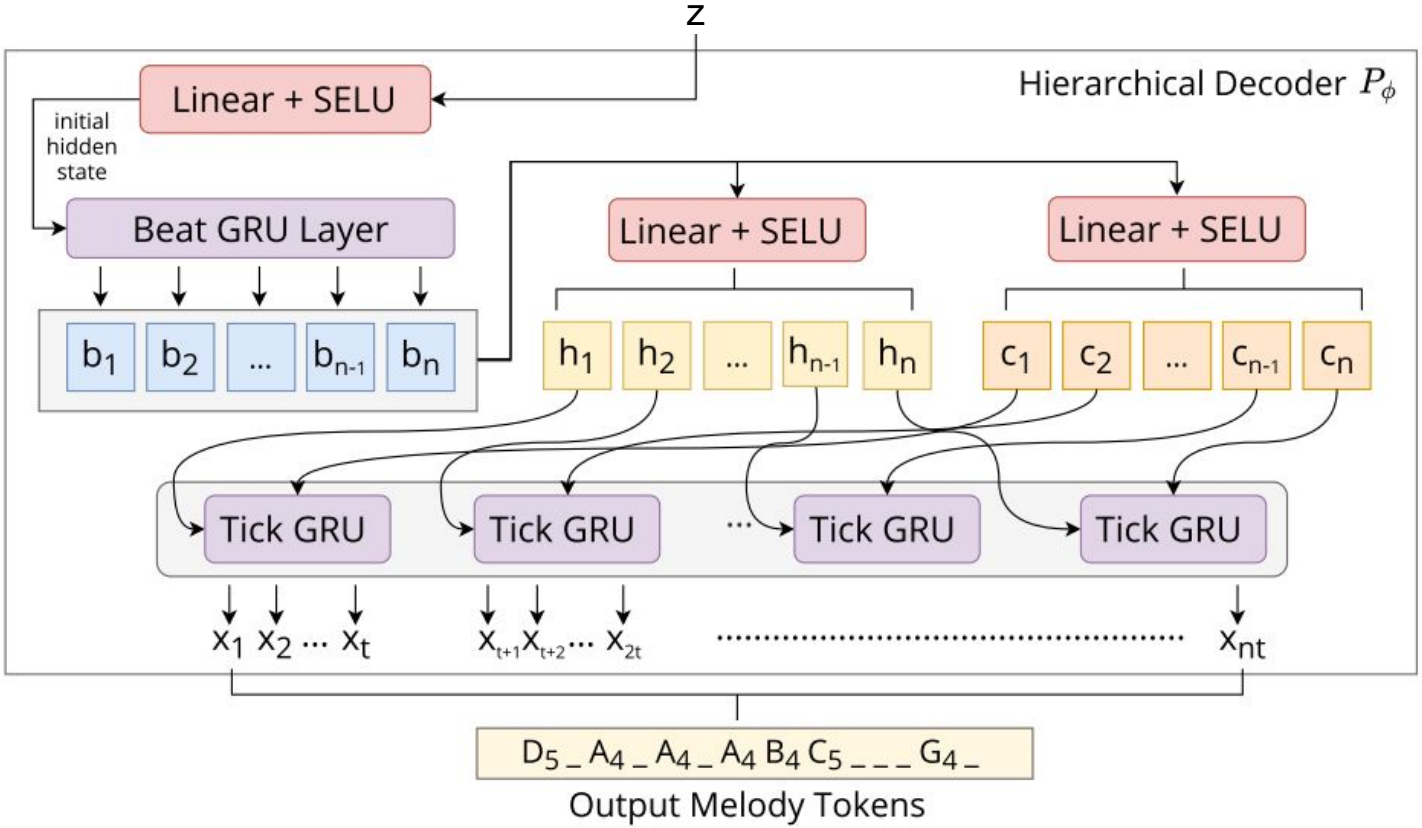
Busca representar un único compas musical como una variable latente z que codifique el ritmo y la altura en dimensiones separadas (zpitch, zrhythm)

1. Pitch Encoder Qo(zpitch|xpitch)
2. Rhythm Encoder Qt(zrhythm|xrhythm)
3. A hierarchical decoder Pphi(x|zpitch,zrhythm)

Pitch Encoder and Rhythm Encoder

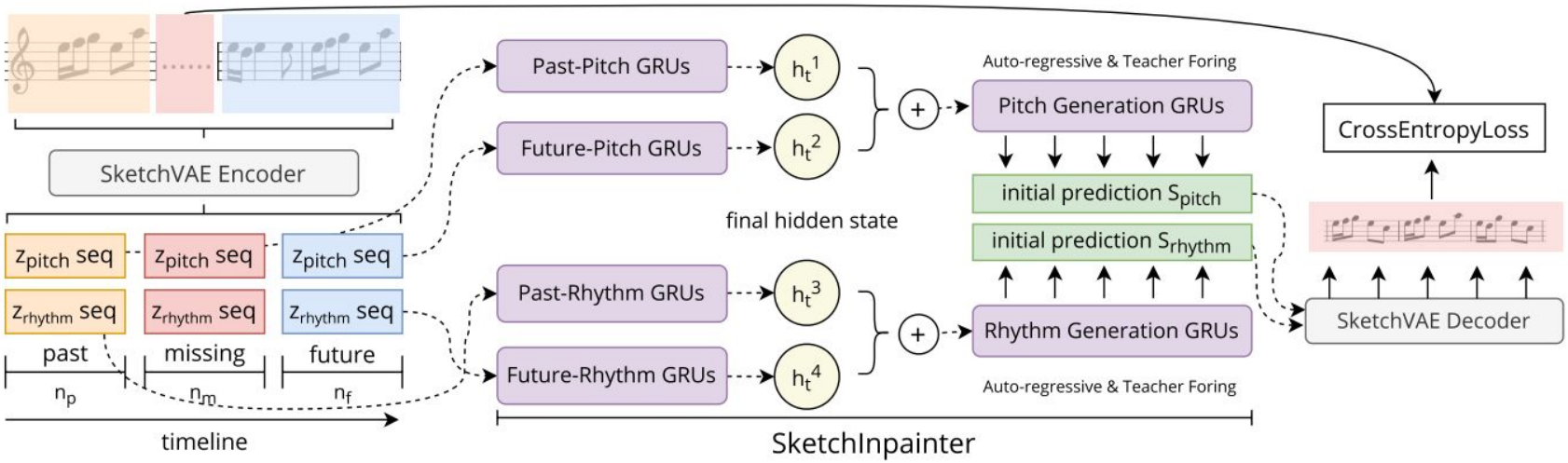


Hierarchical Decoder



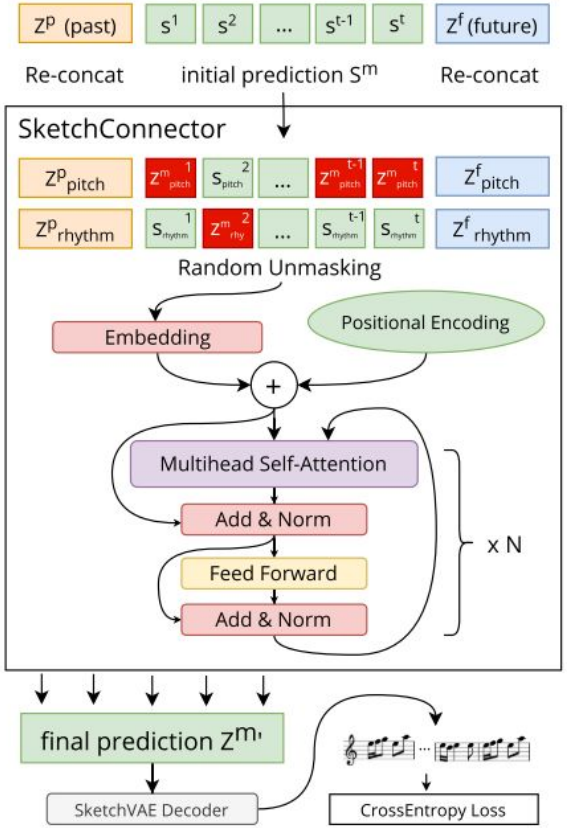
SketchInpainter

SketchInpainter



SketchConnector

SketchConnector



Experimentos

Dataset

- Irish and Scottish monophonic music Dataset
- Solo melodías en tiempos de 4/4
- ~16000 melodías para training
- ~2000 melodías para testing

Baseline (SketchVAE)

- Reconstrucción de input
- Modelos comparados:
  - SketchVAE
  - MeasureVAE
  - EC2-VAE
- Métricas
  - Reconstruction Rate (Accuracy)
- Hiperparámetros:
  - Latent variable |z| set to 256
  - Learning Rate 1e-4
  - Adam Optimization B1=0.9, B2=0.998

Resultados (SketchVAE)

	SketchVAE	MeasureVAE	EC2-VAE
Reconstruccion Rate (Accuracy)	98.8%	98.7%	99.0%

Resultados (SketchVAE)

- Todos los modelos de VAE son capaces de convertir las melodías a variables latentes (al rededor de 99% de precisión).
- SketchVAE es capaz de codificar/decodificar material musical en SketchNet.

Baseline (SketchNet)

- Predicción de 4 compases en el medio a partir de 6 compases de contexto pasado y futuro.
- Modelos comparados:
  - Music InpaintNet
  - SketchVAE + InpaintRNN
  - SketchVAE + SketchInpainter
  - SketchNet
- Métricas
  - Loss
  - Pitch Accuracy
  - Rhythm Accuracy
- Early stopping para todos los sistemas

Baseline (SketchNet)

- Se utilizan dos subsets de test adicionales
  - Irish-Test-R (repetition)
  - Irish-Test-NR (non-repetition)
- Calculados a partir de las similitudes entre el contexto pasado y futuro de cada canción.
  - El 10% de contextos más similares constituyen el primer subset
  - El 10% de contextos más disímiles constituyen el segundo subset
-

Resultados (SketchNet)

	Irish-Test			Irish-Test-R			Irish-Test-NR		
Model	loss ↓	pAcc ↑	rAcc ↑	loss ↓	pAcc ↑	rAcc ↑	loss ↓	pAcc ↑	rAcc ↑
Music InpaintNet	0.662	0.511	0.972	0.312	0.636	0.975	0.997	0.354	0.959
SketchVAE + InpaintRNN	0.714	0.510	0.975	0.473	0.619	0.981	1.075	0.374	0.964
SketchVAE + SketchInpainter	0.693	0.552	0.985	0.295	0.692	0.991	1.002	0.389	0.977
SketchNet	**0.516**	**0.651**	**0.985**	**0.206**	**0.799**	**0.991**	**0.783**	**0.461**	**0.977**

Resultados (SketchNet)

- SketchNet supera a todos los modelos en todos los sets de prueba.
- El desempeño mejoró más en la precisión de altura que en la de ritmo.
- La precisión es casi la misma entre el primer y segundo modelo.
- La precisión es ligeramente mejor si se usa SketchInpainter para tratar el ritmo y la altura de forma independiente durante la generación.
- Usando el transformador encoder y random unmasking se logra el mejor desempeño (SketchNet).
- Al aplicar Bootstrap Significance Test se concluye que SketchNet es diferente al resto de modelos con p-value menor a 0.05
- Se tienen las losses más grandes para todos los modelos en el subset de no-repetición.

Subjective Listening Test

- Cada sujeto escucha tres melodías de 32 segundos de piano renderizadas:
  - La versión original
  - La generación a partir de Music InpaintNet
  - La generación de SketchNet
- Tres criterios:
  - Complexity (cantidad de notas)
  - Structure (repetición de estructure)
  - Musicality (grado de armonía)
- Se asigna puntaje de 1.0 a 5.0 para cada criterio.
- 106 sujetos de prueba, 318 resultados consultados
- El inicio y el final son identicos para las 3 melodías

Subjective Listening Test

Model	Complexity↑	Structure↑	Musicality↑
Original	3.22	3.47	3.56
InpaintNet	2.98	3.01	3.09
SketchNet	3.04	3.29	3.26

Table 2. Results of the subjective listening test.

Resultados Subjective Listening

- SketchNet es mejor en los tres criterios.
- Complexity en los resultados generados por ambos modelos son similares en términos de riqueza de notas.
- SketchNet no incrementa de forma sustancial el número de notas generadas.