# Générateur de musique à partir de paroles

Lucie GALLAND, Ahmed CHOUKARAH, Quentin VERMANDE

13/12/2019

# Modèle

Soit $P$ l'ensemble des hauteurs de note possibles (pitch) (silence inclus).
Soit $D$ l'ensemble des durées possible pour une note.
Soit $S$ l'ensembles des syllabes.
Le texte est découpé en phrases, chaque phrase est traitée séparément.

Entrée : séquence de syllabes : $X \in S^*$. Contexte : sortie $M$ de la production de mélodie de la phrase précédente.
Sortie : séquences de notes : $Y \in (P \times D \times \{0,1\})^*$. Pour $i \in [\![1, |Y|]\!]$, on note $Y_i = (Y_{i,p}, Y_{i,d}, Y_{i,l})$

$$\sum_{i=1}^{|Y|} Y_{i,l} = |X|$$

# GRU

Entrée : $h_{i-1}, x_i$.

$$z_i = \sigma(W_{hz} h_{i-1} + W_{xz} x_i + b_z)$$

$$r_i = \sigma(W_{hr} h_{i-1} + W_{xr} x_i + b_r)$$

$$\hat{h}_i = \tanh(W_h(r_i \times h_{i-1}) + W_x x_i + b)$$

$$h_i = (1 - z_i) \times h_{i-1} + z_i \times \hat{h}_i$$

# Encodeur de paroles

$$\vec{h}_{i,lrc} = GRU(\vec{h}_{i-1,lrc}, x_i)$$

$$\overleftarrow{h}_{i,lrc} = GRU(\overleftarrow{h}_{i+1,lrc}, x_i)$$

$$h_{i,lrc} = \begin{pmatrix} \vec{h}_{i,lrc} \\ \overleftarrow{h}_{i,lrc} \end{pmatrix}$$

# Encodeur de contexte

$$\vec{h}_{i,p} = GRU(\vec{h}_{i-1,p}, m_{i,p})$$

$$\overleftarrow{h}_{i,p} = GRU(\overleftarrow{h}_{i+1,p}, m_{i,p})$$

$$h_{i,p} = \begin{pmatrix} \vec{h}_{i,p} \\ \overleftarrow{h}_{i,p} \end{pmatrix}$$

$$\vec{h}_{i,d} = GRU(\vec{h}_{i-1,d}, m_{i,d}, h_{i,p})$$

$$\overleftarrow{h}_{i,d} = GRU(\overleftarrow{h}_{i+1,d}, m_{i,p}, h_{i,p})$$

$$h_{i,d} = \begin{pmatrix} \vec{h}_{i,d} \\ \overleftarrow{h}_{i,d} \end{pmatrix}$$

$$h_{i,con} = \begin{pmatrix} \vec{h}_{i,p} \\ \overleftarrow{h}_{i,d} \end{pmatrix}$$

# Décodeur

$$c_{i,con} = \sum_{t=1}^{|M|} a_{i,t} h_{t,con}$$

$$c_i = c_{i,con} + h_{j,lrc}$$

$$s_{i,p} = GRU(s_{i,p}, c_{i-1}, y_{i-1,p}, h_{j,lrc})$$

$$s_{i,d} = GRU(s_{i-1,d}, c_{i-1}, y_{i-1,d}, h_{j,rlc})$$

$$s_{i,l} = GRU(s_{i-1,l}, c_{i-1}, y_{i-1,l}, y_{i,p}, y_{i,d}, s_{i,d}$$

$$y_i = (argmax(s_{i,p}), argmax(s_{i,d}), argmax(s_{i,l}))$$

# Modèle

Soit $P$ l'ensemble des hauteurs de note possibles (pitch) (silence inclus).
Soit $D$ l'ensemble des durées possible pour une note.
Soit $M$ l'ensemble des mots.
Soit $S$ l'ensembles des syllabes.
Entrée : séquence de syllabes : $x \in S^*$. Contexte : sortie $m$ de la production de mélodie de la phrase précédente.
Sortie : séquences de notes : $y \in (P \times D \times D)^*$. Pour $i \in [\![1, |Y|]\!]$, on note $Y_i = (Y_{i,p}, Y_{i,d}, Y_{i,r})$

# Encodeur de paroles

$E_m, E_s$ : skip-gram model
L'encodage de la syllabe $s$ du mot $m$ est $E_m(m)E_s(s)$.

# LSTM

Portes internes (avec $x_t$ l'entrée et $h_{t-1}$ la sortie précédente) :

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

La sortie est :

$$\tilde{c}_t = \tanh(w_c[h_{t-1}, x_t] + b_c)$$

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t$$

$$h_t = o_t \times \tanh(c_t)$$

# GAN

Générateur :

- une couche ReLU, deux couches LSTM, une couche linéaire.
- Entrée : un vecteur de bruit et une syllabe encodée.
- Sortie : une note MIDI

Discriminateur :

- deux couches LSTM, une couche linéaire.
- Entrée : une note MIDI et une syllabe encodée.
- Sortie : 2