Improvise a Jazz Solo with an LSTM Network

Welcome to your final programming assignment of this week! In this notebook, you will implement a model that uses an LSTM to generate music. You will even be able to listen to your own music at the end of the assignment.

You will learn to:

- Apply an LSTM to music generation.
- Generate your own jazz music with deep learning.

Please run the following cell to load all the packages required in this assignment. This may take a few minutes.

In [51]:
```
from __future__ import print_function
import IPython
import sys
from music21 import *
import numpy as np
from grammar import *
from qa import *
from preprocess import *
from music_utils import *
from data_utils import *
from keras.models import load_model, Model
from keras.layers import Dense, Activation, Dropout, Input, LSTM, Reshape, Lambda, RepeatVector
from keras.initializers import glorot_uniform
from keras.utils import to_categorical
from keras.optimizers import Adam
from keras import backend as K
```

1 - Problem statement

You would like to create a jazz music piece specially for a friend's birthday. However, you don't know any instruments or music composition. Fortunately, you know deep learning and will solve this problem using an LSTM netwok.

You will train a network to generate novel jazz solos in a style representative of a body of performed work.

1.1 - Dataset

You will train your algorithm on a corpus of Jazz music. Run the cell below to listen to a snippet of the audio from the training set:

In [52]:
```
IPython.display.Audio('./data/30s_seq.mp3')
```

Out[52]:

0:00 / 0:29

We have taken care of the preprocessing of the musical data to render it in terms of musical "values." You can informally think of each "value" as a note, which comprises a pitch and a duration. For example, if you press down a specific piano key for 0.5 seconds, then you have just played a note. In music theory, a "value" is actually more complicated than this--specifically, it also captures the information needed to play multiple notes at the same time. For example, when playing a music piece, you might press down two piano keys at the same time (playng multiple notes at the same time generates what's called a "chord"). But we don't need to worry about the details of music theory for this assignment. For the purpose of this assignment, all you need to know is that we will obtain a dataset of values, and will learn an RNN model to generate sequences of values.

Our music generation system will use 78 unique values. Run the following code to load the raw music data and preprocess it into values. This might take a few minutes.

In [53]:
```
X, Y, n_values, indices_values = load_music_utils()
print('shape of X:', X.shape)
print('number of training examples:', X.shape[0])
print('Tx (length of sequence):', X.shape[1])
print('total # of unique values:', n_values)
print('Shape of Y:', Y.shape)
```

```
shape of X: (60, 30, 78)
number of training examples: 60
Tx (length of sequence): 30
total # of unique values: 78
Shape of Y: (30, 60, 78)
```

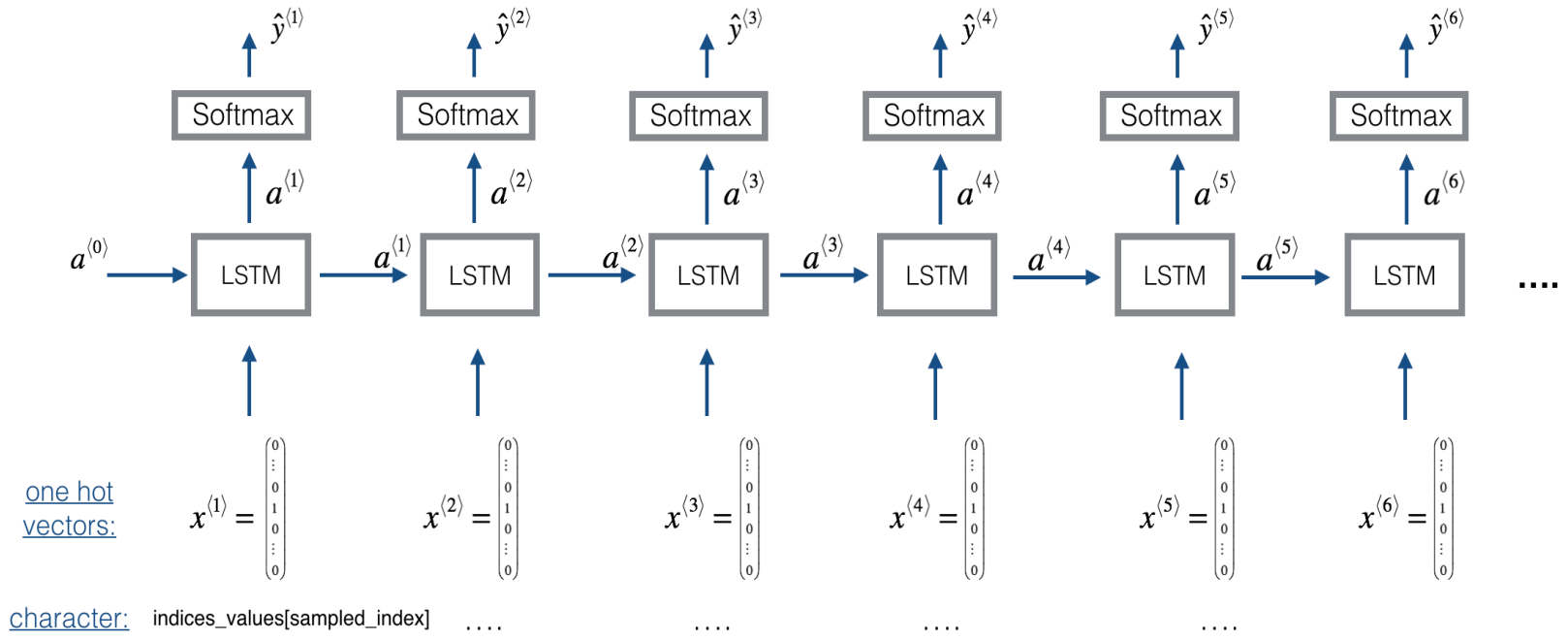
You have just loaded the following:

- X: This is an (m, T_x, 78) dimensional array. We have m training examples, each of which is a snippet of $T_x = 30$ musical values. At each time step, the input is one of 78 different possible values, represented as a one-hot vector. Thus for example, X[i,t,:] is a one-hot vector representating the value of the i-th example at time t.
- Y: This is essentially the same as X, but shifted one step to the left (to the past). Similar to the dinosaurus assignment, we're interested in the network using the previous values to predict the next value, so our sequence model will try to predict $y^{\langle t \rangle}$ given $x^{\langle 1 \rangle}, \ldots, x^{\langle t \rangle}$. However, the data in Y is reordered to be dimension $(T_y, m, 78)$, where $T_y = T_x$. This format makes it more convenient to feed to the LSTM later.
- n_values: The number of unique values in this dataset. This should be 78.
- indices_values: python dictionary mapping from 0-77 to musical values.

1.2 - Overview of our model

Here is the architecture of the model we will use. This is similar to the Dinosaurus model you had used in the previous notebook, except that in you will be implementing it in Keras. The architecture is as follows:

$$x^{\langle i+1 \rangle} = y^{\langle i \rangle}$$



We will be training the model on random snippets of 30 values taken from a much longer piece of music. Thus, we won't bother to set the first input $x^{\langle 1 \rangle} = \vec{0}$, which we had done previously to denote the start of a dinosaur name, since now most of these snippets of audio start somewhere in the middle of a piece of music. We are setting each of the snippts to have the same length $T_x = 30$ to make vectorization easier.

2 - Building the model

In this part you will build and train a model that will learn musical patterns. To do so, you will need to build a model that takes in X of shape $(m, T_x, 78)$ and Y of shape $(T_y, m, 78)$. We will use an LSTM with 64 dimensional hidden states. Lets set `n_a = 64`.

In [54]:
```
n_a = 64
```

Here's how you can create a Keras model with multiple inputs and outputs. If you're building an RNN where even at test time entire input sequence $x^{\langle 1 \rangle}, x^{\langle 2 \rangle}, \ldots, x^{\langle T_x \rangle}$ were *given in advance*, for example if the inputs were words and the output was a label, then Keras has

simple built-in functions to build the model. However, for sequence generation, at test time we don't know all the values of $x^{\langle t \rangle}$ in advance; instead we generate them one at a time using $x^{\langle t \rangle} = y^{\langle t-1 \rangle}$. So the code will be a bit more complicated, and you'll need to implement your own for-loop to iterate over the different time steps.

The function `djmodel()` will call the LSTM layer T_x times using a for-loop, and it is important that all T_x copies have the same weights. I.e., it should not re-initiaiize the weights every time---the T_x steps should have shared weights. The key steps for implementing layers with shareable weights in Keras are:

1. Define the layer objects (we will use global variables for this).
2. Call these objects when propagating the input.

We have defined the layers objects you need as global variables. Please run the next cell to create them. Please check the Keras documentation to make sure you understand what these layers are: Reshape() (https://keras.io/layers/core/#reshape), LSTM() (https://keras.io/layers/recurrent/#lstm), Dense() (https://keras.io/layers/core/#dense).

In [55]:
```
reshapor = Reshape((1, 78))                       # Used in Step 2.B of djmodel(), below
LSTM_cell = LSTM(n_a, return_state = True)        # Used in Step 2.C
densor = Dense(n_values, activation='softmax')    # Used in Step 2.D
```

Each of `reshapor`, `LSTM_cell` and `densor` are now layer objects, and you can use them to implement `djmodel()`. In order to propagate a Keras tensor object X through one of these layers, use `layer_object(X)` (or `layer_object([X,Y])` if it requires multiple inputs.). For example, `reshapor(X)` will propagate X through the `Reshape((1,78))` layer defined above.

Exercise: Implement `djmodel()`. You will need to carry out 2 steps:

1. Create an empty list "outputs" to save the outputs of the LSTM Cell at every time step.
2. Loop for $t \in 1, \ldots, T_x$:

   A. Select the "t"th time-step vector from X. The shape of this selection should be (78,). To do so, create a custom Lambda (https://keras.io/layers/core/#lambda) layer in Keras by using this line of code:

   ```
   x = Lambda(lambda x: X[:,t,:])(X)
   ```

   Look over the Keras documentation to figure out what this does. It is creating a "temporary" or "unnamed" function (that's what Lambda functions are) that extracts out the appropriate one-hot vector, and making this function a Keras `Layer` object to apply to `X`.

   B. Reshape x to be (1,78). You may find the `reshapor()` layer (defined below) helpful.

C. Run x through one step of LSTM_cell. Remember to initialize the LSTM_cell with the previous step's hidden state a and cell state c. Use the following formatting:

```
a, _, c = LSTM_cell(input_x, initial_state=[previous hidden state, previous cell state])
```

D. Propagate the LSTM's output activation value through a dense+softmax layer using `densor`.

E. Append the predicted value to the list of "outputs"

In [56]:
```
# GRADED FUNCTION: djmodel

def djmodel(Tx, n_a, n_values):
    """
    Implement the model

    Arguments:
    Tx -- length of the sequence in a corpus
    n_a -- the number of activations used in our model
    n_values -- number of unique values in the music data

    Returns:
    model -- a keras model with the
    """

    # Define the input of your model with a shape
    X = Input(shape=(Tx, n_values))

    # Define s0, initial hidden state for the decoder LSTM
    a0 = Input(shape=(n_a,), name='a0')
    c0 = Input(shape=(n_a,), name='c0')
    a = a0
    c = c0

    ### START CODE HERE ###
    # Step 1: Create empty list to append the outputs while you iterate (≈1 line)
    outputs = []

    # Step 2: Loop
    for t in range(Tx):

        # Step 2.A: select the "t"th time step vector from X.
        x = Lambda(lambda x: X[:,t,:])(X)
        # Step 2.B: Use reshapor to reshape x to be (1, n_values) (≈1 line)
        x = reshapor(x)
        # Step 2.C: Perform one step of the LSTM_cell
        a, _, c =  LSTM_cell(x, initial_state=[a, c])
        # Step 2.D: Apply densor to the hidden state output of LSTM_Cell
        out = densor(a)
        # Step 2.E: add the output to "outputs"
        outputs.append(out)
```

```
    # Step 3: Create model instance
    model = Model(inputs=[X, a0, c0], outputs=outputs)

    ### END CODE HERE ###

    return model
```

Run the following cell to define your model. We will use `Tx=30`, `n_a=64` (the dimension of the LSTM activations), and `n_values=78`. This cell may take a few seconds to run.

In [57]:
```
model = djmodel(Tx = 30 , n_a = 64, n_values = 78)
```

You now need to compile your model to be trained. We will Adam and a categorical cross-entropy loss.

In [58]:
```
opt = Adam(lr=0.01, beta_1=0.9, beta_2=0.999, decay=0.01)

model.compile(optimizer=opt, loss='categorical_crossentropy', metrics=['accuracy'])
```

Finally, lets initialize `a0` and `c0` for the LSTM's initial state to be zero.

In [*]:
```
m = 60
a0 = np.zeros((m, n_a))
c0 = np.zeros((m, n_a))
```

Lets now fit the model! We will turn `Y` to a list before doing so, since the cost function expects `Y` to be provided in this format (one list item per time-step). So `list(Y)` is a list with 30 items, where each of the list items is of shape (60,78). Lets train for 100 epochs. This will take a few minutes.

In [*]:
```
model.fit([X, a0, c0], list(Y), epochs=100)
```

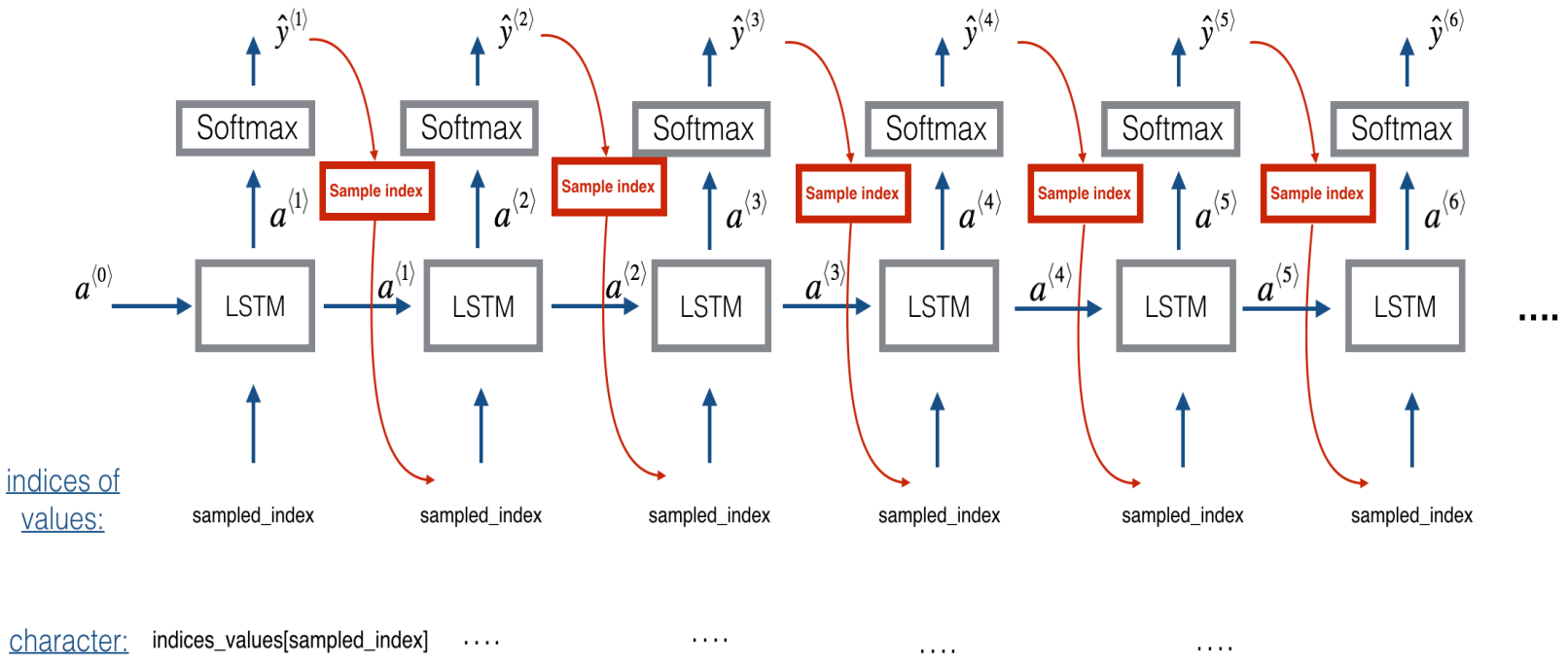
```
Epoch 1/100
```

You should see the model loss going down. Now that you have trained a model, lets go on the the final section to implement an inference algorithm, and generate some music!

3 - Generating music

You now have a trained model which has learned the patterns of the jazz soloist. Lets now use this model to synthesize new music.

3.1 - Predicting & Sampling



At each step of sampling, you will take as input the activation `a` and cell state `c` from the previous state of the LSTM, forward propagate by one step, and get a new output activation as well as cell state. The new activation `a` can then be used to generate the output, using `densor` as before.

To start off the model, we will initialize `x0` as well as the LSTM activation and and cell value `a0` and `c0` to be zeros.

Exercise: Implement the function below to sample a sequence of musical values. Here are some of the key steps you'll need to implement inside the for-loop that generates the T_y output characters:

Step 2.A: Use `LSTM_Cell`, which inputs the previous step's `c` and `a` to generate the current step's `c` and `a`.

Step 2.B: Use `densor` (defined previously) to compute a softmax on `a` to get the output for the current step.

Step 2.C: Save the output you have just generated by appending it to `outputs`.

Step 2.D: Sample `x` to the be "out"'s one-hot version (the prediction) so that you can pass it to the next LSTM's step. We have already provided this line of code, which uses a Lambda (https://keras.io/layers/core/#lambda) function.

```
x = Lambda(one_hot)(out)
```

[Minor technical note: Rather than sampling a value at random according to the probabilities in `out`, this line of code actually chooses the single most likely note at each step using an argmax.]

In [*]:
```
# GRADED FUNCTION: music_inference_model

def music_inference_model(LSTM_cell, densor, n_values = 78, n_a = 64, Ty = 100):
    """
    Uses the trained "LSTM_cell" and "densor" from model() to generate a sequence of values.

    Arguments:
    LSTM_cell -- the trained "LSTM_cell" from model(), Keras layer object
    densor -- the trained "densor" from model(), Keras layer object
    n_values -- integer, umber of unique values
    n_a -- number of units in the LSTM_cell
    Ty -- integer, number of time steps to generate

    Returns:
    inference_model -- Keras model instance
    """

    # Define the input of your model with a shape
    x0 = Input(shape=(1, n_values))

    # Define s0, initial hidden state for the decoder LSTM
    a0 = Input(shape=(n_a,), name='a0')
    c0 = Input(shape=(n_a,), name='c0')
    a = a0
    c = c0
    x = x0

    ### START CODE HERE ###
    # Step 1: Create an empty list of "outputs" to later store your predicted values (≈1 line)
    outputs = []

    # Step 2: Loop over Ty and generate a value at every time step
    for t in range(Ty):

        # Step 2.A: Perform one step of LSTM_cell (≈1 line)
        a, _, c = LSTM_cell(x, initial_state=[a, c])

        # Step 2.B: Apply Dense layer to the hidden state output of the LSTM_cell (≈1 line)
        out = densor(a)

        # Step 2.C: Append the prediction "out" to "outputs". out.shape = (None, 78) (≈1 line)
        outputs.append(out)
```

```
        # Step 2.D: Select the next value according to "out", and set "x" to be the one-hot representation of the
        #           selected value, which will be passed as the input to LSTM_cell on the next step. We have prov
        #           the line of code you need to do this.
        x =  Lambda(one_hot)(out)

    # Step 3: Create model instance with the correct "inputs" and "outputs" (≈1 line)
    inference_model = Model(inputs=[x0, a0, c0], outputs=outputs)

    ### END CODE HERE ###

    return inference_model
```

Run the cell below to define your inference model. This model is hard coded to generate 50 values.

In [*]:
```
inference_model = music_inference_model(LSTM_cell, densor, n_values = 78, n_a = 64, Ty = 50)
```

Finally, this creates the zero-valued vectors you will use to initialize `x` and the LSTM state variables `a` and `c`.

In [*]:
```
x_initializer = np.zeros((1, 1, 78))
a_initializer = np.zeros((1, n_a))
c_initializer = np.zeros((1, n_a))
```

Exercise: Implement `predict_and_sample()`. This function takes many arguments including the inputs [x_initializer, a_initializer, c_initializer]. In order to predict the output corresponding to this input, you will need to carry-out 3 steps:

1. Use your inference model to predict an output given your set of inputs. The output `pred` should be a list of length T_y where each element is a numpy-array of shape (1, n_values).
2. Convert `pred` into a numpy array of T_y indices. Each index corresponds is computed by taking the `argmax` of an element of the `pred` list. Hint (https://docs.scipy.org/doc/numpy/reference/generated/numpy.argmax.html).
3. Convert the indices into their one-hot vector representations. Hint (https://keras.io/utils/#to_categorical).

In [*]:
```
# GRADED FUNCTION: predict_and_sample

def predict_and_sample(inference_model, x_initializer = x_initializer, a_initializer = a_initializer,
                       c_initializer = c_initializer):
    """
    Predicts the next value of values using the inference model.

    Arguments:
    inference_model -- Keras model instance for inference time
    x_initializer -- numpy array of shape (1, 1, 78), one-hot vector initializing the values generation
    a_initializer -- numpy array of shape (1, n_a), initializing the hidden state of the LSTM_cell
    c_initializer -- numpy array of shape (1, n_a), initializing the cell state of the LSTM_cel

    Returns:
    results -- numpy-array of shape (Ty, 78), matrix of one-hot vectors representing the values generated
    indices -- numpy-array of shape (Ty, 1), matrix of indices representing the values generated
    """

    ### START CODE HERE ###
    # Step 1: Use your inference model to predict an output sequence given x_initializer, a_initializer and c_ini
    pred = inference_model.predict([x_initializer, a_initializer, c_initializer])
    # Step 2: Convert "pred" into an np.array() of indices with the maximum probabilities
    indices = np.argmax(pred, axis=-1)
    # Step 3: Convert indices to one-hot vectors, the shape of the results should be (1, )
    results = to_categorical(indices, num_classes=78)
    ### END CODE HERE ###

    return results, indices
```

In [*]:
```
results, indices = predict_and_sample(inference_model, x_initializer, a_initializer, c_initializer)
print("np.argmax(results[12]) =", np.argmax(results[12]))
print("np.argmax(results[17]) =", np.argmax(results[17]))
print("list(indices[12:18]) =", list(indices[12:18]))
```

Expected Output: Your results may differ because Keras' results are not completely predictable. However, if you have trained your LSTM_cell with model.fit() for exactly 100 epochs as described above, you should very likely observe a sequence of indices that are not all identical. Moreover, you should observe that: np.argmax(results[12]) is the first element of list(indices[12:18]) and np.argmax(results[17]) is the last element of list(indices[12:18]).

np.argmax(results[12]) =	1

np.argmax(results[12]) =	42
list(indices[12:18]) =	[array([1]), array([42]), array([54]), array([17]), array([1]), array([42])]

3.3 - Generate music

Finally, you are ready to generate music. Your RNN generates a sequence of values. The following code generates music by first calling your `predict_and_sample()` function. These values are then post-processed into musical chords (meaning that multiple values or notes can be played at the same time).

Most computational music algorithms use some post-processing because it is difficult to generate music that sounds good without such post-processing. The post-processing does things such as clean up the generated audio by making sure the same sound is not repeated too many times, that two successive notes are not too far from each other in pitch, and so on. One could argue that a lot of these post-processing steps are hacks; also, a lot the music generation literature has also focused on hand-crafting post-processors, and a lot of the output quality depends on the quality of the post-processing and not just the quality of the RNN. But this post-processing does make a huge difference, so lets use it in our implementation as well.

Lets make some music!

Run the following cell to generate music and record it into your `out_stream`. This can take a couple of minutes.

In [*]:
```
out_stream = generate_music(inference_model)
```

To listen to your music, click File->Open... Then go to "output/" and download "my_music.midi". Either play it on your computer with an application that can read midi files if you have one, or use one of the free online "MIDI to mp3" conversion tools to convert this to mp3.

As reference, here also is a 30sec audio clip we generated using this algorithm.

In [*]:
```
IPython.display.Audio('./data/30s_trained_model.mp3')
```

Congratulations!

You have come to the end of the notebook.

Here's what you should remember:

- A sequence model can be used to generate musical values, which are then post-processed into midi music.
- Fairly similar models can be used to generate dinosaur names or to generate music, with the major difference being the input fed to the model.
- In Keras, sequence generation involves defining layers with shared weights, which are then repeated for the different time steps $1, \ldots, T_x$.

Congratulations on completing this assignment and generating a jazz solo!

References

The ideas presented in this notebook came primarily from three computational music papers cited below. The implementation here also took significant inspiration and used many components from Ji-Sung Kim's github repository.

- Ji-Sung Kim, 2016, deepjazz (https://github.com/jisungk/deepjazz)
- Jon Gillick, Kevin Tang and Robert Keller, 2009. Learning Jazz Grammars (http://ai.stanford.edu/~kdtang/papers/smc09-jazzgrammar.pdf)
- Robert Keller and David Morrison, 2007, A Grammatical Approach to Automatic Improvisation (http://smc07.uoa.gr/SMC07%20Proceedings/SMC07%20Paper%2055.pdf)
- François Pachet, 1999, Surprising Harmonies (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.7473&rep=rep1&type=pdf)

We're also grateful to François Germain for valuable feedback.