

Assignment 1: Deep N-grams

Welcome to the first graded assignment of course 3. In this assignment you will explore Recurrent Neural Networks RNN .

In this notebook you will apply the following steps:

- Convert a line of text into a tensor
- Create a tensorflow dataset
- Define a GRU model using TensorFlow
- Train the model using TensorFlow
- Compute the accuracy of your model using the perplexity
- Generate text using your own model

Before getting started take some time to read the following tips:

TIPS FOR SUCCESSFUL GRADING OF YOUR ASSIGNMENT:

- All cells are frozen except for the ones where you need to submit your solutions.
- You can add new cells to experiment but these will be omitted by the grader, so don't rely on newly created cells to host your solution code, use the provided places for this.
- You can add the comment # grade-up-to-here in any graded cell to signal the grader that it must only evaluate up to that point. This is helpful if you want to check if you are on the right track even if you are not done with the whole assignment. Be sure to remember to delete the comment afterwards!
- To submit your notebook, save it and then click on the blue submit button at the beginning of the page.

Table of Contents

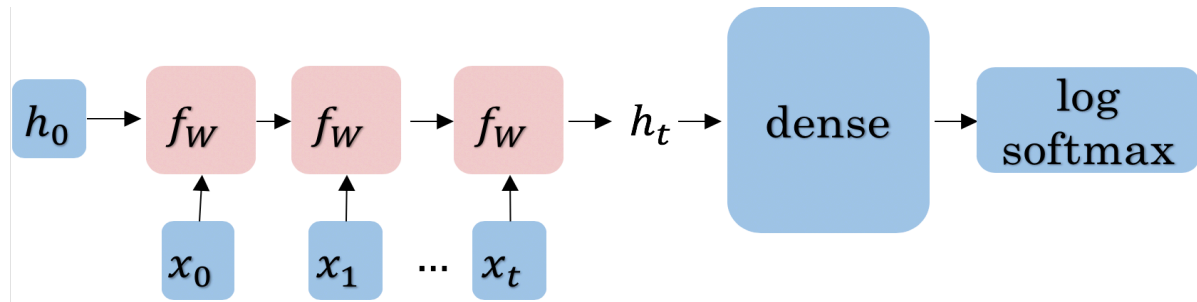
- [Overview](#)
- [1 - Data Preprocessing Overview](#)
 - [1.1 - Loading in the Data](#)
 - [1.2 - Create the vocabulary](#)
 - [1.3 - Convert a Line to Tensor](#)
 - [Exercise 1 - line_to_tensor](#)
 - [1.4 - Prepare your data for training and testing](#)
 - [1.5 - Tensorflow dataset](#)
 - [1.6 - Create the input and the output for your model](#)
 - [Exercise 2 - data_generator](#)
 - [1.7 - Create the training dataset](#)
- [2 - Defining the GRU Language Model \(GRULM\)](#)
 - [Exercise 3 - GRULM](#)
- [3 - Training](#)
 - [Exercise 4 - train_model](#)
- [4 - Evaluation](#)
 - [4.1 - Evaluating using the Deep Nets](#)
 - [Exercise 5 - log_perplexity](#)
- [5 - Generating Language with your Own Model](#)
 - [Optional Exercise 6 - GenerativeModel \(Not graded\)](#)
- [On statistical methods](#)

Overview

In this lab, you'll delve into the world of text generation using Recurrent Neural Networks (RNNs). Your primary objective is to predict the next set of characters based on the preceding ones. This seemingly straightforward task holds immense practicality in applications like predictive text and creative writing.

The journey unfolds as follows:

- **Data Preprocessing:** You'll start by converting lines of text into numerical tensors, making them machine-readable.
- **Dataset Creation:** Next, you'll create a TensorFlow dataset, which will serve as the backbone for supplying data to your model.
- **Neural Network Training:** Your model will be trained to predict the next set of characters, specifying the desired output length.
- **Character Embeddings:** Character embeddings will be employed to represent each character as a vector, a fundamental technique in natural language processing.
- **GRU Model:** Your model utilizes a Gated Recurrent Unit (GRU) to process character embeddings and make sequential predictions. The following figure gives you a summary of what you are about to implement.



- **Prediction Process:** The model's predictions are achieved through a linear layer and log-softmax computation.

This overview sets the stage for your exploration of text generation. Get ready to unravel the secrets of language and embark on a journey into the realm of creative writing and predictive text generation.

And as usual let's start by importing all the required libraries

```
import os
import traceback
os.environ['TF_CPP_MIN_LOG_LEVEL'] = '3'

import numpy as np
import random as rnd

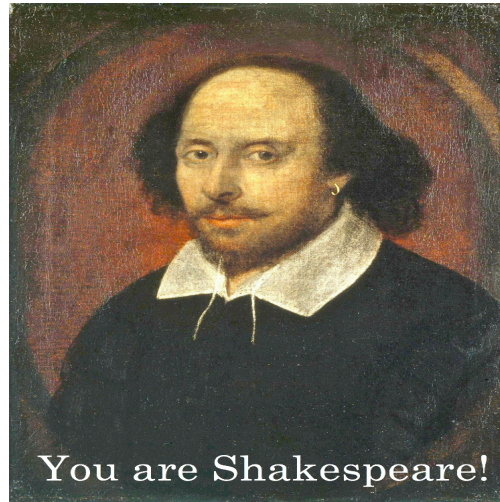
import tensorflow as tf
from tensorflow.keras import layers
from tensorflow.keras.layers import Input

from termcolor import colored

# set random seed
rnd.seed(22)
```

```
import sys, unittest
```

1 - Data Preprocessing Overview



In this section, you will prepare the data for training your model. The data preparation involves the following steps:

- **Dataset Import:** Begin by importing the dataset. Each sentence is structured as one line in the dataset. To ensure consistency, remove any extra spaces from these lines using the `strip` function.
- **Data Storage:** Store each cleaned line in a list. This list will serve as the foundational dataset for your text generation task.
- **Character-Level Processing:** Since the goal is character generation, it's essential to process the text at the character level, not the word level. This involves converting each individual character into a numerical representation. To achieve this:
 - Use the `tf.strings.unidecode_split` function to split each sentence into its constituent

```
dirname = 'data/'
filename = 'shakespeare_data.txt'
lines = [] # storing all the lines in a variable.

counter = 0

with open(os.path.join(dirname, filename)) as files:
    for line in files:
        # remove leading and trailing whitespace
        pure_line = line.strip().lower()

        # if pure_line is not the empty string,
        if pure_line:
            # append it to the list
            lines.append(pure_line)

n_lines = len(lines)
print(f"Number of lines: {n_lines}")
```

Number of lines: 125097

Let's examine a few lines from the corpus. Pay close attention to the structure and style employed by Shakespeare in this excerpt. Observe that character names are written in uppercase, and each line commences with a capital letter. Your task in this exercise is to construct a generative model capable of emulating this particular structural style.

```
print("\n" .join(lines[500:514]))
```

BENVOLIO Here were the servants of your adversary,
And yours, close fighting ere I did approach:
I drew to part them: in the instant came
The fiery Tybalt, with his sword prepared,
Which, as he breathed defiance to my ears,
He swung about his head and cut the winds,
Who nothing hurt withal hiss'd him in scorn:
While we were interchanging thrusts and blows,

1.2 - Create the vocabulary

In the following code cell, you will create the vocabulary for text processing. The vocabulary is a crucial component for understanding and processing text data. Here's what the code does:

- Concatenate all the lines in our dataset into a single continuous text, separated by line breaks.
- Identify and collect the unique characters that make up the text. This forms the basis of our vocabulary.
- To enhance the vocabulary, introduce two special characters:
 - [UNK]: This character represents any unknown or unrecognized characters in the text.
 - "" (empty character): This character is used for padding sequences when necessary.
- The code concludes with the display of statistics, showing the total count of unique characters in the vocabulary and providing a visual representation of the complete character set.

```
text = "\n".join(lines)
# The unique characters in the file
vocab = sorted(set(text))
vocab.insert(0, "[UNK]") # Add a special character for any unknown
vocab.insert(1, "") # Add the empty character for padding.
```

```
print(f'{len(vocab)} unique characters')
```

```
print("\n".join(vocab))
```

82 unique characters

```
[UNK]
! $ & ' ( ) , - . 0 1 2 3 4 5 6 7 8 9 : ; ? A B C D E F G H I J K L M N
O P Q R S T U V W X Y Z [ ] a b c d e f g h i j k l m n o p q r s t u v w
x y z |
```

1.3 - Convert a Line to Tensor

Now that you have your list of lines, you will convert each character in that list to a number using the order given by your vocabulary. You can use `tf.strings.unicode_split` to split the text into characters.

```
line = "Hello world!"
chars = tf.strings.unicode_split(line, input_encoding='UTF-8')
print(chars)

tf.Tensor([b'H' b'e' b'l' b'l' b'o' b' ' b'w' b'o' b'r' b'l' b'd' b '!'], shape=(12,), dtype=string)
```

Using your vocabulary, you can convert the characters given by `unicode_split` into numbers. The number will be the index of the character in the given vocabulary.

```

print(vocab.index('a'))
print(vocab.index('e'))
print(vocab.index('i'))
print(vocab.index('o'))
print(vocab.index('u'))
print(vocab.index(' '))
print(vocab.index('2'))
print(vocab.index('!'))

```

```

55
59
63
69
75
4
16
17

```

Tensorflow has a function `tf.keras.layers.StringLookup` that does this efficiently for list of characters. Note that the output object is of type `tf.Tensor`. Here is the result of applying the `StringLookup` function to the characters of "Hello world"

```

ids = tf.keras.layers.StringLookup(vocabulary=list(vocab), mask_token=None)(chars)
print(ids)

```

```
tf.Tensor([34 59 66 66 69  4 77 69 72 66 58  5], shape=(12,), dtype=int64)
```

Exercise 1 - line_to_tensor

Instructions: Write a function that takes in a single line and transforms each character into its unicode integer. This returns a list of integers, which we'll refer to as a tensor.

```

# GRADED FUNCTION: line_to_tensor
def line_to_tensor(line, vocab):
    """
    Converts a line of text into a tensor of integer values representing characters.

    Args:
        line (str): A single line of text.
        vocab (list): A list containing the vocabulary of unique characters.

    Returns:
        tf.Tensor(dtype=int64): A tensor containing integers (unicode values) corresponding to the characters in the line.
    """
    ### START CODE HERE ###

    # Split the input line into individual characters
    chars = tf.strings.unicode_split(line, input_encoding='UTF-8')
    # Map characters to their respective integer values using StringLookup
    ids = tf.keras.layers.StringLookup(vocabulary=vocab, mask_token=None)(chars)

    ### END CODE HERE ###

    return ids

```

```
# Test your function
tmp_ids = line_to_tensor('abc xyz', vocab)
print(f"Result: {tmp_ids}")
print(f"Output type: {type(tmp_ids)}")
```

Result: [55 56 57 4 78 79 80]

Output type: <class 'tensorflow.python.framework.ops.EagerTensor'>

Expected output

Result: [55 56 57 4 78 79 80]

Output type: <class 'tensorflow.python.framework.ops.EagerTensor'>

UNIT TEST

```
if __name__ == '__main__':
    unittest.main()
```

All test passed!

You will also need a function that produces text given a numeric tensor. This function will be useful for inspection when you use your model to generate new text, because you will be able to see words rather than lists of numbers. The function will use the inverse Lookup function `tf.keras.layers.StringLookup` with `invert=True` in its parameters.

```
def text_from_ids(ids, vocab):
    """
    Converts a tensor of integer values into human-readable text.

    Args:
        ids (tf.Tensor): A tensor containing integer values (unicode IDs).
        vocab (list): A list containing the vocabulary of unique characters.

    Returns:
        str: A string containing the characters in human-readable format.
    """
    # Initialize the StringLookup Layer to map integer IDs back to characters
    chars_from_ids = tf.keras.layers.StringLookup(vocabulary=vocab, invert=True, mask_token='')

    # Use the layer to decode the tensor of IDs into human-readable text
    return tf.strings.reduce_join(chars_from_ids(ids), axis=-1)
```

Use the function for decoding the tensor produced by "Hello world!"

```
text_from_ids(tmp_ids, vocab)
```

b'Hello world!'

1.4 - Prepare your data for training and testing

As usual, you will need some data for training your model, and some data for testing its performance. So, we will use 124097 lines for training and 1000 lines for testing.

```
train_lines = lines[:-1000] # Leave the rest for training
eval_lines = lines[-1000:] # Create a holdout validation set

print(f"Number of training lines: {len(train_lines)}")
print(f"Number of validation lines: {len(eval_lines)}")
```

Number of training lines: 124097

Number of validation lines: 1000

1.5 - TensorFlow dataset

Most of the time in Natural Language Processing, and AI in general you use batches when training your models. Here, you will build a dataset that takes in some text and returns a batch of text fragments (Not necessarily full sentences) that you will use for training.

- The generator will produce text fragments encoded as numeric tensors of a desired length

Once you create the dataset, you can iterate on it like this:

```
data_generator.take(1)
```

This generator returns the data in a format that you could directly use in your model when computing the feed-forward of your algorithm. This batch dataset generator returns batches of data in an endless way.

So, let's check how the different parts work with a corpus composed of 2 lines. Then, you will use these parts to create the first graded function of this notebook.

In order to get a dataset generator that produces batches of fragments from the corpus, you first need to convert the whole text into a single line, and then transform it into a single big tensor. This is only possible if your data fits completely into memory, but that is the case here.

```
all_ids = line_to_tensor("\n".join(["Hello world!", "Generative AI"]), vocab)
all_ids
<tf.Tensor: shape=(26,), dtype=int64, numpy=
array([34, 59, 66, 66, 69,  4, 77, 69, 72, 66, 58,  5,  3, 33, 59, 68, 59,
       72, 55, 74, 63, 76, 59,  4, 27, 35])>
```

Create a dataset out of a tensor like input. This initial dataset will dispatch numbers in packages of a specified length. For example, you can use it for getting the 10 first encoded characters of your dataset. To make it easier to read, we can use the `text_from_ids` function.

```
ids_dataset = tf.data.Dataset.from_tensor_slices(all_ids)
print([text_from_ids([id1, vocab), numpy() for id1 in ids_dataset.take(10)])
[b'H', b'e', b'l', b'l', b'o', b' ', b'w', b'o', b'r', b'l']
```

But we can configure this dataset to produce batches of the same size each time. We could use this functionality to produce text fragments of a desired size (`seq_length + 1`). We will explain later why you need an extra character into the sequence.

```
seq_length = 10
data_generator = ids_dataset.batch(seq_length + 1, drop_remainder=True)
```

You can verify that the data generator produces encoded fragments of text of the desired length. For example, let's ask the generator to produce 2 batches of data using the function `data_generator.take(2)`

```
for seq in data_generator.take(2):
    print(seq)
```

```
tf.Tensor([34 59 66 66 69  4 77 69 72 66 58], shape=(11,), dtype=int64)
tf.Tensor([ 5  3 33 59 68 59 72 55 74 63 76], shape=(11,), dtype=int64)
```

But as usual, it is easier to understand if you print it in human readable characters using the 'text_from_ids' function.

```
i = 1
for seq in data_generator.take(2):
    print(f"{i}. {text_from_ids(seq, vocab).numpy()}")
    i = i + 1
```

1. b'Hello world'
2. b'!\nGenerativ'

1.6 - Create the input and the output for your model

In this task you have to predict the next character in a sequence. The following function creates 2 tensors, each with a length of `seq_length` out of the input sequence of length `seq_length + 1`. The first one contains the first `seq_length` elements and the second one contains the last `seq_length` elements. For example, if you split the sequence `['H', 'e', 'l', 'l', 'o']`, you will obtain the sequences `['H', 'e', 'l', 'l']` and `['e', 'l', 'l', 'o']`.

```
def split_input_target(sequence):
    """
    Splits the input sequence into two sequences, where one is shifted by one position.

    Args:
        sequence (tf.Tensor or list): A list of characters or a tensor.

    Returns:
        tf.Tensor, tf.Tensor: Two tensors representing the input and output sequences for
    """
    # Create the input sequence by excluding the last character
    input_text = sequence[:-1]
    # Create the target sequence by excluding the first character
    target_text = sequence[1:]

    return input_text, target_text
```

Look the result using the following sequence of characters

```
split_input_target(list("Tensorflow"))
([ 'T', 'e', 'n', 's', 'o', 'r', 'f', 'l', 'o'],
 [ 'e', 'n', 's', 'o', 'r', 'f', 'l', 'o', 'w'])
```

The first sequence will be the input and the second sequence will be the expected output

Now, put all this together into a function to create your batch dataset generator

Exercise 2 - data_generator

Instructions: Create a batch dataset from the input text. Here are some things you will need.

- Join all the input lines into a single string. When you have a big dataset, you would better use a flow from directory or any other kind of generator.
- Transform your input text into numeric tensors
- Create a TensorFlow DataSet from your numeric tensors: Just feed the numeric tensors into the function `tf.data.Dataset.from_tensor_slices`
- Make the dataset produce batches of data that will form a single sample each time. This is, make the dataset produce a sequence of `seq_length + 1`, rather than single numbers at each time. You can do it using the `batch` function of the already created dataset. You must specify the length of the produced sequences (`seq_length + 1`). So, the sequence length produced by the dataset will `seq_length + 1`. It must have that extra element since you will get the input and the output sequences out of the same element. `drop_remainder=True` will drop the sequences that do not have the required length. This could happen each time that the dataset reaches the end of the input sequence.
- Use the `split_input_target` to split each element produced by the dataset into the mentioned input and output sequences. The input will have the first `seq_length` elements, and the output will have the last `seq_length`. So, after this step, the dataset generator will produce batches of pairs (input, output) sequences.
- Create the final dataset, using `dataset_xy` as the starting point. You will configure this dataset to shuffle the data during the generation of the data with the specified `BUFFER_SIZE`. For performance reasons, you would like that tensorflow pre-process the data in parallel with training. That is called `prefetching`, and it will be configured for you.

```

# GRADED FUNCTION: create_batch_dataset
def create_batch_dataset(lines, vocab, seq_length=100, batch_size=64):
    """
    Creates a batch dataset from a list of text lines.

    Args:
        lines (list): A list of strings with the input data, one line per row.
        vocab (list): A list containing the vocabulary.
        seq_length (int): The desired length of each sample.
        batch_size (int): The batch size.

    Returns:
        tf.data.Dataset: A batch dataset generator.
    """
    # Buffer size to shuffle the dataset
    # (TF data is designed to work with possibly infinite sequences,
    # so it doesn't attempt to shuffle the entire sequence in memory. Instead,
    # it maintains a buffer in which it shuffles elements).
    BUFFER_SIZE = 10000

    # For simplicity, just join all lines into a single line
    single_line_data = "\n".join(lines)

    ### START CODE HERE ###

    # Convert your data into a tensor using the given vocab
    all_ids = line_to_tensor(single_line_data, vocab)
    # Create a TensorFlow dataset from the data tensor
    ids_dataset = tf.data.Dataset.from_tensor_slices(all_ids)
    # Create a batch dataset
    data_generator = ids_dataset.batch(seq_length + 1, drop_remainder=True)
    # Map each input sample using the split_input_target function
    dataset_xy = data_generator.map(split_input_target)

    # Assemble the final dataset with shuffling, batching, and prefetching
    dataset = (
        dataset_xy
        .shuffle(BUFFER_SIZE)
        .batch(batch_size, drop_remainder=True)
        .prefetch(tf.data.experimental.AUTOTUNE)
    )

    ### END CODE HERE ###

    return dataset

```

```
# test your function
tf.random.set_seed(1)
dataset = create_batch_dataset(train_lines[1:100], vocab, seq_length=16, batch_size=2)

print("Prints the elements into a single batch. The batch contains 2 elements: ")

for input_example, target_example in dataset.take(1):
    print("\n\033[94mInput0\t:", text_from_ids(input_example[0], vocab).numpy())
    print("\n\033[93mTarget0\t:", text_from_ids(target_example[0], vocab).numpy())

    print("\n\n\033[94mInput1\t:", text_from_ids(input_example[1], vocab).numpy())
    print("\n\n\033[93mTarget1\t:", text_from_ids(target_example[1], vocab).numpy())
```

Prints the elements into a single batch. The batch contains 2 elements:

Input0 : b'and sight distra'

Target0 : b'nd sight distrac'

Input1 : b'when in his fair'

Target1 : b'hen in his fair '

Expected output

Prints the elements into a single batch. The batch contains 2 elements:

Input0 : b'and sight distra'

Target0 : b'nd sight distrac'

Input1 : b'when in his fair'

Target1 : b'hen in his fair '

UNIT TEST

```
if __name__ == '__main__':
    unittest.main()
```

All test passed!

1.7 - Create the training dataset

Now, you can generate your training dataset using the functions defined above. This will produce pairs of input/output tensors each time the batch generator creates an entry.

Batch size

```
BATCH_SIZE = 64
```

```
dataset = create_batch_dataset(train_lines, vocab, seq_length=16, batch_size=BATCH_SIZE)
```

2 - Defining the GRU Language Model (GRULM)

Now that you have the input and output tensors, you will go ahead and initialize your model. You will be implementing the GRULM, gated recurrent unit model. To implement this model, you will be using TensorFlow. Instead of implementing the GRU from scratch (you saw this already in a lab), you will use the necessary methods from a built-in package. You can use the following packages when constructing the model:

- `tf.keras.layers.Embedding` : Initializes the embedding. In this case it is the size of the vocabulary by the dimension of the model. [docs](#)
 - `Embedding(vocab_size, embedding_dim)` .
 - `vocab_size` is the number of unique words in the given vocabulary.
 - `embedding_dim` is the number of elements in the word embedding (some choices for a word embedding size range from 150 to 300, for example).
- `tf.keras.layers.GRU` : TensorFlow GRU layer. [docs](#)) Builds a traditional GRU of `rnn_units` with dense internal transformations. You can read the paper here: <https://arxiv.org/abs/1412.3555>
 - `units` : Number of recurrent units in the layer. It must be set to `rnn_units`
 - `return_sequences` : It specifies if the model returns a sequence of predictions. Set it to `True`
 - `return_state` : It specifies if the model must return the last internal state along with the prediction. Set it to `True`
- `tf.keras.layers.Dense` : A dense layer. [docs](#). You must set the following parameters:
 - `units` : Number of units in the layer. It must be set to `vocab_size`
 - `activation` : It must be set to `log_softmax` function as described in the next line.
- `tf.nn.log_softmax` : Log of the output probabilities. [docs](#)
 - You don't need to set any parameters, just set the activation parameter as `activation=tf.nn.log_softmax` .

```

# GRADED CLASS: GRULM
class GRULM(tf.keras.Model):
    """
    A GRU-based language model that maps from a tensor of tokens to activations over a vocabulary.

    Args:
        vocab_size (int, optional): Size of the vocabulary. Defaults to 256.
        embedding_dim (int, optional): Depth of embedding. Defaults to 256.
        rnn_units (int, optional): Number of units in the GRU cell. Defaults to 128.

    Returns:
        tf.keras.Model: A GRULM language model.
    """
    def __init__(self, vocab_size=256, embedding_dim=256, rnn_units=128):
        super().__init__(self)

        ### START CODE HERE ###

        # Create an embedding layer to map token indices to embedding vectors
        self.embedding = tf.keras.layers.Embedding(vocab_size, embedding_dim)
        # Define a GRU (Gated Recurrent Unit) layer for sequence modeling
        self.gru = tf.keras.layers.GRU(rnn_units,
                                         return_sequences=True,
                                         return_state=True)

        # Apply a dense layer with log-softmax activation to predict next tokens
        self.dense = tf.keras.layers.Dense(vocab_size, activation=tf.nn.log_softmax)

        ### END CODE HERE ###

    def call(self, inputs, states=None, return_state=False, training=False):
        x = inputs
        # Map input tokens to embedding vectors
        x = self.embedding(x, training=training)
        if states is None:
            # Get initial state from the GRU layer
            states = self.gru.get_initial_state(x)
        x, states = self.gru(x, initial_state=states, training=training)
        # Predict the next tokens and apply log-softmax activation
        x = self.dense(x, training=training)
        if return_state:
            return x, states
        else:
            return x

```

Now, you can define a new GRULM model. You must set the `vocab_size` to 82; the size of the embedding `embedding_dim` to 256, and the number of units that will have you recurrent neural network `rnn_units` to 512

```
# Length of the vocabulary in StringLookup Layer
```

```
vocab_size = 82
```

```
# The embedding dimension
```

```
embedding_dim = 256
```

```
# RNN Layers
```

```
rnn_units = 512
```

```
model = GRUMLP(
```

```
    vocab_size=vocab_size,
```

```
    embedding_dim=embedding_dim,
```

```
    rnn_units=rnn_units)
```

```
# testing your model
```

```
try:
```

```
    # Simulate inputs of length 100. This allows to compute the shape of all inputs and c
```

```
    model.build(input_shape=(BATCH_SIZE, 100))
```

```
    model.call(Input(shape=(100)))
```

```
    model.summary()
```

```
except:
```

```
    print("\033[91mError! \033[0mA problem occurred while building your model. This error
```

```
    traceback.print_exc()\n
```

```
Model: "grulm"
```

| Layer (type) | Output Shape | Param # |
|-------------------------------------|------------------------------------|---------|
| ===== | | |
| embedding (Embedding) | (None, 100, 256) | 20992 |
| gru (GRU) | [(None, 100, 512), (None, 512)] | 1182720 |
| dense (Dense) | (None, 100, 82) | 42066 |
| ===== | | |
| Total params: 1245778 (4.75 MB) | | |
| Trainable params: 1245778 (4.75 MB) | | |
| Non-trainable params: 0 (0.00 Byte) | | |

Expected output

```
Model: "grulm"
```

| Layer (type) | Output Shape | Param # |
|-------------------------------------|------------------------------------|---------|
| ===== | | |
| embedding (Embedding) | (None, 100, 256) | 20992 |
| gru (GRU) | [(None, 100, 512), (None, 512)] | 1182720 |
| dense (Dense) | (None, 100, 82) | 42066 |
| ===== | | |
| Total params: 1245778 (4.75 MB) | | |
| Trainable params: 1245778 (4.75 MB) | | |
| Non-trainable params: 0 (0.00 Byte) | | |

UNIT TEST

```

import unittest
class GPTLM(GPTLM):

```

```

Test case 1:

```

```

All tests passed!

```

```

Test case 2:

```

```

All tests passed!

```

Now, let's use the model for predicting the next character using the untrained model. At the beginning the model will generate only gibberish.

```

for input_example_batch, target_example_batch in dataset.take(1):
    print("Input: ", input_example_batch[0].numpy()) # Lets use only the first sequence of the dataset
    example_batch_predictions = model(tf.constant([input_example_batch[0].numpy()]))
    print("\n" + "example_batch_predictions shape: ", example_batch_predictions.shape)

```

```

Input:  [73 62 55 66 74  4 62 69 66 58  4 74 62 59  4 69 70 63 68 63 69 68
 4 69
 60  4 42 79 74 62 55 61 69 72 55 73  4 59 72 59  4 35  4 77 63 66 66  3
 55 66 66 69 77  4 69 60  4 74 62 79  4 77 63 74 73 11  4 55 68 58  4 60
 59 55 72  4 74 69  4 65 63 66 66  4 55  4 77 69 69 58 57 69 57 65 11  4
 66 59 73 74]

```

```

(1, 100, 82) # (batch_size, sequence_length, vocab_size)

```

The output size is (1, 100, 82). We predicted only on the first sequence generated by the batch generator. 100 is the number of predicted characters. It has exactly the same length as the input. And there are 82 values for each predicted character. Each of these 82 real values are related to the logarithm likelihood of each character to be the next one in the sequence. The bigger the value, the higher the likelihood. As the network is not trained yet, all those values must be very similar and random. Just check the values for the last prediction on the sequence.

```

example_batch_predictions[0][99].numpy()
array([-4.394292, -4.40476, -4.412511, -4.3860574, -4.4002004,
       -4.406017, -4.391721, -4.412148, -4.398938, -4.409028,
       -4.4036393, -4.4032044, -4.412924, -4.4294558, -4.4058275,
       -4.42735, -4.4058037, -4.3803573, -4.400791, -4.400802,
       -4.406753, -4.3941684, -4.3932967, -4.4120283, -4.387955,
       -4.401387, -4.409998, -4.3959923, -4.3911653, -4.3927126,
       -4.418993, -4.3965154, -4.4084926, -4.411369, -4.3952107,
       -4.4010906, -4.3987713, -4.438392, -4.395659, -4.382313,
       -4.41321, -4.3956714, -4.407373, -4.4170127, -4.4092526,
       -4.3940024, -4.412472, -4.4142866, -4.4282956, -4.4088607,
       -4.4279985, -4.412727, -4.4195285, -4.4102244, -4.3989367,
       -4.4095216, -4.4144945, -4.3992176, -4.413693, -4.400564,
       -4.4158244, -4.428811, -4.41504, -4.419681, -4.416288,
       -4.4038076, -4.423907, -4.413642, -4.394232, -4.419655,
       -4.4194393, -4.425377, -4.40321, -4.3787003, -4.411811,
       -4.386497, -4.405401, -4.3831735, -4.40585, -4.428001,
       -4.410224, -4.4233975], dtype=float32)

```

And the simplest way to choose the next character is by getting the index of the element with the highest likelihood. So, for instance, the prediction for the last character would be:

```

last_character = tf.math.argmax(example_batch_predictions[0][99])
print(last_character.numpy())

```

```

73

```

And the prediction for the whole sequence would be:

```
sampled_indices = tf.math.argmax(example_batch_predictions[0], axis=-1)
print(sampled_indices.numpy())
```

```
[21 27 71 12 73 41  6 75 75 48 72 77  6  6 41 75  4  3 28 28  3 14 18 75
 65 41 29 29 73  6 71  6 14 14 23 36 41 41 73 44 41 73 41 41 22 75  8 28
 25 12 12 75  5 41 75 74 41 29  6  6 41 41 22 77  3 17 54 41 50 48 41 29
 65 41 73 41 73 75 54  5  1 75  8 12 41 41 41 75 75 48 29 14 45  5 54 54
 12 70 21 73]
```

Those 100 numbers represent 100 predicted characters. However, humans cannot read this. So, let's print the input and output sequences using our `text_from_ids` function, to check what is going on.

```
print("Input:\n", text_from_ids(input_example_batch[0], vocab))
print()
print("Next Char Predictions:\n", text_from_ids(sampled_indices, vocab))
```

Input:

```
tf.Tensor(b'shalt hold the opinion of Pythagoras ere I will\nallow of thy
wits, and fear to kill a woodcock, lest', shape=(), dtype=string)
```

Next Char Predictions:

```
tf.Tensor(b"7Aq-s0$uuVrw$$0u \nBB\n04uk0CCs$q$009J00sR0s008u'B;--u!0ut0C
$$008w\n3]0XV0Ck0s0su]!u'-000uuVC0S!]]-p7s", shape=(), dtype=string)
```

As expected, the untrained model just produces random text as response of the given input. It is also important to note that getting the index of the maximum score is not always the best choice. In the last part of the notebook you will see another way to do it.

3 - Training

Now you are going to train your model. As usual, you have to define the cost function and the optimizer. You will use the following built-in functions provided by TensorFlow:

- [`tf.losses.SparseCategoricalCrossentropy\(\)`](#) : The Sparse Categorical Cross Entropy loss. It is the loss function used for multiclass classification.
 - `from_logits=True` : This parameter informs the loss function that the output values generated by the model are not normalized like a probability distribution. This is our case, since our GRULM model uses a `log_softmax` activation rather than the `softmax`.
- [`tf.keras.optimizers.Adam`](#) : Use Adaptive Moment Estimation, a stochastic gradient descent method optimizer that works well in most of the cases. Set the `learning_rate` to 0.00125.

Exercise 4 - train_model

Instructions: Compile the GRULM model using a `SparseCategoricalCrossentropy` loss and the `Adam` optimizer


```
# GRADED FUNCTION: Compile model

def compile_model(model):
    """
    Sets the loss and optimizer for the given model

    Args:
        model (tf.keras.Model): The model to compile.

    Returns:
        tf.keras.Model: The compiled model.
    """
    ### START CODE HERE ###

    # Define the loss function. Use SparseCategoricalCrossentropy
    loss = tf.losses.SparseCategoricalCrossentropy(from_logits=True)
    # Define and Adam optimizer
    opt = tf.keras.optimizers.Adam(learning_rate=0.00125)
    # Compile the model using the parametrized Adam optimizer and the SparseCategoricalCr
    model.compile(optimizer=opt, loss=loss)

    ### END CODE HERE ###

    return model
```

UNIT TEST

```
unittest.test_compile_model(compile_model)
```

All test passed!

Now, train your model for 10 epochs. With GPU this should take about one minute. With CPU this could take several minutes.

EPOCHS = 10

```
# Compile the model
model = compile_model(model)
# Fit the model
history = model.fit(dataset, epochs=EPOCHS)
```

```
Epoch 1/10
790/790 [=====] - 12s 11ms/step - loss: 2.0145
Epoch 2/10
790/790 [=====] - 9s 10ms/step - loss: 1.4798
Epoch 3/10
790/790 [=====] - 9s 10ms/step - loss: 1.3790
Epoch 4/10
790/790 [=====] - 9s 10ms/step - loss: 1.3326
Epoch 5/10
790/790 [=====] - 9s 10ms/step - loss: 1.3032
Epoch 6/10
790/790 [=====] - 9s 10ms/step - loss: 1.2819
Epoch 7/10
790/790 [=====] - 9s 9ms/step - loss: 1.2656
Epoch 8/10
790/790 [=====] - 9s 10ms/step - loss: 1.2518
Epoch 9/10
790/790 [=====] - 9s 10ms/step - loss: 1.2406
Epoch 10/10
790/790 [=====] - 9s 10ms/step - loss: 1.2310
```

You can uncomment the following cell to save the weights of your model. This allows you to use the model later.

```
# If you want, you can save the final model. Here is deactivated.
#output_dir = './model/'
#
#try:
#    shutil.rmtree(output_dir)
#except OSError as e:
#    pass
#model.save_weights(output_dir)
```

The model was only trained for 10 epochs. We pretrained a model for 30 epochs, which can take about 5 minutes in a GPU.

4 - Evaluation

4.1 - Evaluating using the Deep Nets

Now that you have learned how to train a model, you will learn how to evaluate it. To evaluate language models, we usually use perplexity which is a measure of how well a probability model predicts a sample. Note that perplexity is defined as:

$$P(W) = \frac{1}{N} \prod_{i=1}^N P(w_i | w_1, \dots, w_{n-1})$$

As an implementation hack, you would usually take the log of that formula (to enable us to use the log probabilities we get as output of our RNN, convert exponents to products, and products into sums which makes computations less complicated and computationally more efficient).

$$\begin{aligned} \log P(W) &= \log \left(\frac{1}{N} \prod_{i=1}^N P(w_i | w_1, \dots, w_{n-1}) \right) \\ &= \log \left(\prod_{i=1}^N \frac{1}{P(w_i | w_1, \dots, w_{n-1})} \right)^{\frac{1}{N}} \\ &= \log \left(\prod_{i=1}^N P(w_i | w_1, \dots, w_{n-1}) \right)^{-\frac{1}{N}} \\ &= -\frac{1}{N} \log \left(\prod_{i=1}^N P(w_i | w_1, \dots, w_{n-1}) \right) \\ &= -\frac{1}{N} \sum_{i=1}^N \log P(w_i | w_1, \dots, w_{n-1}) \end{aligned}$$

Exercise 5 - log_perplexity

Instructions: Write a program that will help evaluate your model. Implementation hack: your program takes in `preds` and `target`. `preds` is a tensor of log probabilities. You can use `tf.one_hot` to transform the `target` into the same dimension. You then multiply them and sum them. For the sake of simplicity, we suggest you use the NumPy functions `sum`, `mean` and `equal`, Good luck!

Hints

```
# GRADED FUNCTION: log_perplexity
def log_perplexity(preds, target):
    """
    Function to calculate the log perplexity of a model.

    Args:
        preds (tf.Tensor): Predictions of a list of batches of tensors corresponding to lines of text
        target (tf.Tensor): Actual list of batches of tensors corresponding to lines of text

    Returns:
        float: The log perplexity of the model.
    """
    PADDING_ID = 1

    ### START CODE HERE ###

    # Calculate log probabilities for predictions using one-hot encoding
    log_p = np.sum(preds * tf.one_hot(target, preds.shape[-1]), axis=-1) # HINT: tf.one_hot
    # Identify non-padding elements in the target
    non_pad = 1.0 - np.equal(target, PADDING_ID) # You should check if the target is not padding
    # Apply non-padding mask to log probabilities to exclude padding
    log_p = log_p * non_pad # Get rid of the padding
    # Calculate the log perplexity by taking the sum of log probabilities and dividing by the number of non-padding elements
    log_ppx = np.sum(log_p, axis=-1) / np.sum(non_pad, axis=-1) # Remember to set the axis correctly
    # Compute the mean of log perplexity
    log_ppx = np.mean(log_ppx) # Compute the mean of the previous expression

    ### END CODE HERE ###
    return log_ppx
```

#UNIT TESTS

All test passed!

Now load the provided pretrained model just to ensure that results are consistent for the upcoming parts of the notebook. You need to instantiate the GRULM model and then load the saved weights.

Load the pretrained model. This step is optional.

```
vocab_size = len(vocab)
embedding_dim = 256
rnn_units = 512

model = GRULM(
    vocab_size=vocab_size,
    embedding_dim=embedding_dim,
    rnn_units = rnn_units)
model.build(input_shape=(100, vocab_size))
model.load_weights('model.weights.h5')

<tensorflow.python.checkpoint.checkpoint.CheckpointLoadStatus at 0x7148440c4370>
```

Now, you will use the 1000 lines of the corpus that were reserved at the beginning of this notebook as test data. You will apply the same preprocessing as you did for the train dataset: get the numeric tensor from the input lines, and use the `split_input_target` to generate the inputs and the expected outputs.

```
#for line in eval_lines[1:3]:
eval_text = "\n".join(eval_lines)
eval_ids = line_to_tensor([eval_text], vocab)
input_ids, target_ids = split_input_target(tf.squeeze(eval_ids, axis=0))

preds, status = model(tf.expand_dims(input_ids, 0), training=False, states=None, return_s

#Get the log perplexity
log_ppx = log_perplexity(preds, tf.expand_dims(target_ids, 0))
print('The log perplexity and perplexity of your model are {log_ppx} and {perplexity}')
```

The log perplexity and perplexity of your model are 1.2239635591264044 and 3.400639693684683 respectively

Expected Output: The log perplexity and perplexity of your model are around 1.22 and 3.40 respectively.

So, the log perplexity of the model is 1.22. It is not an easy to interpret metric, but it can be used to compare among models. The smaller the value the better the model.

5 - Generating Language with your Own Model

Your GRULM model demonstrates an impressive ability to predict the most likely characters in a sequence, based on log scores. However, it's important to acknowledge that this model, in its default form, is deterministic and can result in repetitive and monotonous outputs. For instance, it tends to provide the same answer to a question consistently.

To make your language model more dynamic and versatile, you can introduce an element of randomness into its predictions. This ensures that even if you feed the model in the same way each time, it will generate different sequences of text.

To achieve this desired behavior, you can employ a technique known as random sampling. When presented with an array of log scores for the N characters in your dictionary, you add an array of random numbers to this data. The extent of randomness introduced into the predictions is regulated by a parameter called "temperature". By comparing the random numbers to the original input scores, the model adapts its choices, offering diversity in its outputs.

This doesn't imply that the model produces entirely random results on each iteration. Rather, with each prediction, there is a probability associated with choosing a character other than the one with the highest score. This concept becomes more tangible when you explore the accompanying Python code.

```
def temperature_random_sampling(log_probs, temperature=1.0):
    """Temperature Random sampling from a categorical distribution. The higher the temper
    random the output. If temperature is close to 0, it means that the model will just
    of the character with the highest input log_score

    Args:
        log_probs (tf.Tensor): The log scores for each character in the dictionary
        temperature (number): A value to weight the random noise.

    Returns:
        int: The index of the selected character
    """
    # Generate uniform random numbers with a slight offset to avoid log(0)
    u = tf.random.uniform(minval=1e-6, maxval=1.0 - 1e-6, shape=log_probs.shape)

    # Apply the Gumbel distribution transformation for randomness
    g = -tf.math.log(-tf.math.log(u))

    # Adjust the logits with the temperature and choose the character with the highest score
    return tf.math.argmax(log_probs * g, axis=-1)
```

Now, it's time to bring all the elements together for the exciting task of generating new text. The `GenerativeModel` class plays a pivotal role in this process, offering two essential functions:

1. `generate_one_step` : This function is your go-to method for generating a single character at a time. It accepts two key inputs: an initial input sequence and a state that can be thought of as the ongoing context or memory of the model. The function delivers a single character prediction and an updated state, which can be used as the context for future predictions.
2. `generate_n_chars` : This function takes text generation to the next level. It orchestrates the iterative generation of a sequence of characters. At each iteration, `generate_one_step` is called with the last generated character and the most recent state. This dynamic approach ensures that the generated text evolves organically, building upon the context and characters produced in previous steps. Each character generated in this process is collected and stored in the result list, forming the final output text.

Optional Exercise 6 - GenerativeModel (Not graded)

Instructions: Implementing the One-Step Generator

In this task, you will create a function to generate a single character based on the input text, using the provided vocabulary and the trained model. Follow these steps to complete the `generate_one_step` function:

1. Start by transforming your input text into a tensor using the given vocab. This will convert the text into a format that the model can understand.
2. Utilize the trained model with the input_ids and the provided states to predict the next characters. Make sure to retrieve the updated states from this prediction because they are essential for the final output.
3. Since we are only interested in the next character prediction, keep only the result for the last character in the sequence.
4. Employ the temperature random sampling technique to convert the vector of scores into a single character prediction. For this step, you will use the predicted_logits obtained in the previous step and the temperature parameter of the model.
5. To transform the numeric prediction into a human-readable character, use the `text_from_ids` function. Be mindful that `text_from_ids` expects a list as its input, so you need to wrap the output of the `temperature_random_sampling` function in square brackets [...]. Don't forget to use `self.vocab` as the second parameter for character mapping.
6. Finally, return the predicted_chars, which will be a single character, and the states tensor obtained from step 2. These components are essential for maintaining the sequence and generating subsequent characters.


```

# UNGRADED CLASS: GenerativeModel
class GenerativeModel(tf.keras.Model):
    def __init__(self, model, vocab, temperature=1.0):
        """
        A generative model for text generation.

        Args:
            model (tf.keras.Model): The underlying model for text generation.
            vocab (list): A list containing the vocabulary of unique characters.
            temperature (float, optional): A value to control the randomness of text generation.
        """
        super().__init__()
        self.temperature = temperature
        self.model = model
        self.vocab = vocab

    @tf.function
    def generate_one_step(self, inputs, states=None):
        """
        Generate a single character and update the model state.

        Args:
            inputs (string): The input string to start with.
            states (tf.Tensor): The state tensor.

        Returns:
            tf.Tensor, states: The predicted character and the current GRU state.
        """
        # Convert strings to token IDs.

        ### START CODE HERE ###

        # Transform the inputs into tensors
        input_ids = line_to_tensor(inputs, self.vocab)
        # Predict the sequence for the given input_ids. Use the states and return_state=True
        predicted_logits, states = self.model(input_ids, states=states, return_state=True)
        # Get only last element of the sequence
        predicted_logits = predicted_logits[0, -1, :]
        # Use the temperature_random_sampling to generate the next character.
        predicted_ids = temperature_random_sampling(predicted_logits, self.temperature)
        # Use the chars_from_ids to transform the code into the corresponding character
        predicted_chars = text_from_ids([predicted_ids], self.vocab)

        ### END CODE HERE ###

        # Return the characters and model state.
        return tf.expand_dims(predicted_chars, 0), states

    def generate_n_chars(self, num_chars, prefix):
        """
        Generate a text sequence of a specified length, starting with a given prefix.

        Args:
            num_chars (int): The length of the output sequence.
            prefix (string): The prefix of the sequence (also referred to as the seed).

        Returns:
            str: The generated text sequence.
        """
        states = None
        next_char = tf.constant([prefix])

```

```

result = [next_char]
for n in range(num_chars):
    next_char, states = self.generate_one_step(next_char, states=states)
    result.append(next_char)

```

```

return tf.strings.join(result+[0], numpy().decode('utf-8'))

```

UNIT TEST

Fix the seed to get replicable results for testing

```
tf.random.set_seed(272)
```

```
gen = GenerativeModel(model, vocab, temperature=0.5)
```

```
print(gen.generate_n_chars(32, " "), '\n\n' + '_'*80)
```

```
print(gen.generate_n_chars(32, "Dear"), '\n\n' + '_'*80)
```

```
print(gen.generate_n_chars(32, "KING"), '\n\n' + '_'*80)
```

hear he has a soldier.

Here is a

Dear gold, if thou wilt endure the e

KING OF THE SHREW

IV I beseech you,

Expected output

hear he has a soldier.

Here is a

Dear gold, if thou wilt endure the e

KING OF THE SHREW

IV I beseech you,

```

1. unit test test_GenerativeModel(GenerativeModel_model_vocab)

```

All test passed!

Now, generate a longer text. Let's check if it looks like Shakespeare fragment


```
tf.random.set_seed(np.random.randint(1, 1000))
gen = GenerativeModel(model, vocab, temperature=0.8)
import time
start = time.time()
print(gen.generate_n_chars(1000, "ROMEO "), '\n\n' + '_'*80)
print('\nRun time: %s' % (time.time() - start))
```

ROMEO and GUILDENSTERN]
 KING HENRY V Then my lord keep of them I smoke thy breath,
 Nor more gifts and Valentine i' the
 stemberent, sir, to stay to live in single and the
 world and counterfeit than she is.
 BUCKINGHAM Well, I have a goodly best food in his blood he set
 desires, that he did look upon penial
 Silvia, but the short are at my pity;
 Even in the former friend, young Pretentio,
 Whom I would encounter'd by the Reward
 Does and his throad of life to him, being presumes.
 [Enter SILVIA, and BEATRICE]
 What do you have made me effect?
 BRUTUS Your own wife and dunghill!
 [Exit]
 PERICLES, PRINCE OF TYRE
 ACT II
 SCENE I Milia.
 SHALLOW |
 | sir.
 CLARENCE I find another man.
 PAROLLES I know the best,
 The fearful thoughts shall buy in the time.
 GLOUCESTER It is King Henry to his brother Julia:
 But thou becomes thy father, and these deserts,
 Eiver'd from the town of a kingdom,
 And certain, devise to prove against the hour,
 I hope to hang the door of mine again,
 Is letting our side gibed the sea and h

Run time: 2.5923538208007812

In the generated text above, you can see that the model generates text that makes sense capturing dependencies between words and without any input. A simple n-gram model would have not been able to capture all of that in one sentence.

On statistical methods

Using a statistical method like the one you implemented in course 2 will not give you results that are as good as you saw here. Your model will not be able to encode information seen previously in the data set and as a result, the perplexity will increase. Remember from course 2 that the higher the perplexity, the worse your model is. Furthermore, statistical ngram models take up too much space and memory. As a result, they will be inefficient and too slow. Conversely, with deepnets, you can get a better perplexity. Note, learning about n-gram language models is still important and allows you to better understand deepnets.