WID 3003: Neural Computing

Explore/Play around with Large Language Models (LLMs)

Name: Aiman Syazwan bin Adam
Student ID: 17201819/2

GPT2 Classical Malay Text Generation with KerasNLP

For this tutorial, I am using KerasNLP to load a pre-trained Large Language Model (LLM) - GPT-2 Model, initially invented by OpenAI and then fine-tune the model using a specific text style which is Classical Malay text. The model can generate text by continuing the input.

Setup Google Colab

Starts with changing the runtime type to GPU accelerator runtime since I'm going to fine-tune the GPT-2 model and it's going to take a lot of time and memory.

Install KerasNLP and Import Dependencies

```
!pip install -q keras-nlp
```

```
import keras_nlp
from keras.models import load_model
import tensorflow as tf
from tensorflow import keras
import time
import gspread
import pandas as pd
import numpy as np
```

KerasNLP

Large Language Models are complex to build and expensive to train from scratch. Luckily there are pre-trained LLMs available for use right away. KerasNLP provides a large number of pre-trained models that allow one to experiment with SOTA models without needing to train them.

Load a pre-trained GPT-2 model

Load the pre-trained models available from the KerasNLP repository.

```
preprocessor = keras_nlp.models.GPT2CausalLMPreprocessor.from_preset(
    "gpt2_base_en",
    sequence_length=128,
)
gpt2_lm = keras_nlp.models.GPT2CausalLM.from_preset(
    "gpt2_base_en", preprocessor=preprocessor
)
```

Generate Some Text

The model can use to generate some text right away using generate().

```
start = time.time()

output = gpt2_lm.generate("My trip to Kuala Lumpur was so stressfull", max_length=200)
print("\nGPT-2 output:")
print(output)

end = time.time()
print(f"TOTAL TIME ELAPSED: {end - start:.2f}s")
```

Finetune on Classical Malay dataset

The model can be trained to generate text in a specific style, short or long depending on the dataset. The pre-trained GPT-2 models can also be finetuned by non-English datasets.

Load the Classical Malay Story Dataset

The dataset I read it using a data frame from Google Sheets.

```
# The google sheet link https://docs.google.com/spreadsheets/d/191YRBsdUEGtgXvWl428L9xt48Ah5zcyoRK1pvNhUDPM/edit#gid=1080950255

gsheetid = "191YRBsdUEGtgXvWl428L9xt48Ah5zcyoRK1pvNhUDPM"
sheet_name = "Sheet2"

gsheet_url = "https://docs.google.com/spreadsheets/d/{}/gviz/tq?tqx=out:csv&sheet={}".format(gsheetid, sheet_name)
```

Store the Data in the Data Frame

```
df = pd.read_csv(gsheet_url)
df = df.iloc[:, 0:3]
print(df['content'])
```

```
0     Dia ternganga. Tidak menyangka akan ditinggalk...
1     Angin menderu-deru, biasan pertembungan antara...
2      “Di manakah budak itu?” \r\nBagai halilintar,...
3     Api marak meliang-liuk diserbu angin, menerang...
4     “Allahu akbar.”\r\n\tJeda sedetik.\r\n\t“Attah...
```

Pre-process the Data

Pre-processing the data by removing new lines, tabs, lowercasing and more.

```
df['content'] = df['content'].str.lower().str.replace('\r\n', '').str.replace('\t', '').str.replace('\n', '').str.replace('      ','')
```

Sample Classical Malay Text

```
df['content'][1]
```

‘’’

angin menderu-deru, biasan pertembungan antara dia dan satu lembaga di hadapannya. semakin lama mereka bertarung, semakin jelas susuk lawannya itu. seorang lelaki, berwajah garang dengan misai dan janggut yang sedikit tebal. rambutnya pendek dan kemas, bertanjak megah, berpakaian pendekar yang diperbuat daripada kain yang terbaik lagi mahal. kerisnya bukan senjata biasa-biasa. hanya mereka yang bertaraf laksamana mampu mempunyai keris sedemikian rupa. mereka sudah bertarung garang semenjak kelam dinihari sehingga terang mula merajai alam. sekali-sekala, mereka bagaikan melayang saat menyerang atau mengelak serangan lawan. ibarat punya kesaktian tersendiri. masing-masing tidak terkalahkan. “engkau boleh bertahan sebegini lama ya! tidak hairanlah engkau mampu menumbangkan nama-nama besar di selatan ini!” saukan cuba mencapai kepalanya. dia tunduk tetapi dijamu pula dengan tujahan lutut lawan. tangkas dua lengan dinaikkan untuk mempertahankan diri. tubuhnya terangkat sekitar sehasta dari bumi. “tetapi engkau masih belum mampu menghadapi aku!” tangan yang tadi cuba menyauk kepalanya, kini melawan arah dengan melepaskan pukulan silang. pantas dia mengangkat sebelah tangan dan kaki, mempertahankan kepala dan tubuh daripada menjadi mangsa. namun, serangan pantas itu berisi tenaga yang begitu tinggi, membuatkan dia terpelanting ke tepi sebelum bergolek beberapa kali di lantai bumi. bingkas dia bangkit dengan sebelah lutut dan tangan menongkat tubuh. lawannya membuka tari silat, bagaikan memancing untuk dia memulakan serangan. langkah si lawan itu kemas dan teratur. daripada buah-buah persilatannya, tidak mempamerkan sebarang kelemahan yang terbuka. tangannya perlahan-lahan beralun. seiring dia menarik nafas, tangan kiri merapati hadapan

dada, sementara tangan kanan terjulur keluar. ibu jari dilipat ke dalam telapak sedang jari-jemari lain terbuka lurus. pandangan redupnya memerhati lawan. dia dan tetamu tanpa undangan itu telah bertukar puluhan jurus untuk beberapa ketika. “sudah bersedia untuk tamatkan pertembungan ini?”
‘”

Training the Classical Malay Text

```
train_ds = (
    tf.data.Dataset.from_tensor_slices(df['content'])
    .batch(8)
    .cache()
    .prefetch(tf.data.AUTOTUNE)
)

# Running through the whole dataset takes long, only take `500` and run 1
# epochs for demo purposes.
train_ds = train_ds.take(500)
num_epochs = 10

learning_rate = keras.optimizers.schedules.PolynomialDecay(
    5e-4,
    decay_steps=train_ds.cardinality() * num_epochs,
    end_learning_rate=0.0,
)
loss = keras.losses.SparseCategoricalCrossentropy(from_logits=True)
gpt2_lm.compile(
    optimizer=keras.optimizers.Adam(learning_rate),
    loss=loss,
    weighted_metrics=["accuracy"],
)

gpt2_lm.fit(train_ds, epochs=num_epochs)
```

Generate Some Classical Malay Text

```
output = gpt2_lm.generate("raja dan permaisuri", max_length=200)
print(output)
```

‘”
raja dan permaisuri perlahan, sahaja tidak dapat digambarkan oleh kata-kata. langit dihiasi awan gemawan gebu, meredupkan pakaian dan sedikit beberapa orang tentera istana santubong.“tuanku, jebat. apakah engkau sudah kersama-sama guru.”“sang kelembai?”“telah berhadapan den

‘’’

Develop Simple UI

Using gradio to develop a simple input and output UI.

```
import gradio as gr

def predict(starts_words):
    predict_next = gpt2_lm.generate(starts_words, max_length=200)



    return predict_next

demo = gr.Interface(
    fn=predict,
    inputs=gr.Textbox(lines=2, placeholder="Start your story here:"),
    outputs=["text"]
)
demo.launch(share=True)
```