AAI520_Final_Group_Inference

October 16, 2023

0.1 AAI-520

0.2 Final Project - Group 6

0.3 Chatbot for Movie Info utilizing the Cornell Movie Dialogs Corpus

This Jupyter Notebook is used for inference of the chatbot. It is used to load the trained model and generate responses to user input. The model is trained on a custom training set derived from the Cornell Movie Dialogs Corpus.

Install the needed libraries to load and run the fine tuned Transformer.

[ ]:

```
!pip install transformers
!pip install accelerate
!pip install peft
```

You must enter the HuggingFace token to be able to retrieve the model stored in the repository.

[ ]:

```
!git config --global credential.helper store
```

[ ]:

```
!huggingface-cli login
```

The below code will load the adapter PEFT model we have created. The HuggingFace transformers library handles loading the base model and applying the adapter fine tuned settings.

[ ]:

```
from transformers import AutoTokenizer
import transformers
import torch

model = "guitarnoob/msaai520_with_HF_format_v1"

tokenizer = AutoTokenizer.from_pretrained(model)
pipeline = transformers.pipeline(
    "text-generation",
    model=model,
    torch_dtype=torch.float16,
    device_map="auto",
)
```

Below is the code to run a Chat prompt. The format used is the HuggingFace prompt format detailed in the HF blog: https://huggingface.co/blog/llama2

For multi-turn chat converstations, we will append the answer and follow up question to the instruction.

[ ]:

```
instruction = """
<s>[INST] <<SYS>>
You are a knowledgeable movie bot who can only answer questions related to movies.

Below is an instruction that describes a movie related question. Write a response that appropriately answers the question using the Cornell Movie-Dialog Corpus.

If you cannot answer the question or the question is not related to movies repond with "Sorry I am a movie bot and I am not sure the answer to that question."

Keep your answer concise and less that 20 words.
<</SYS>>

What movies came out in 1980?  [/INST]
"""

sequences = pipeline(
    instruction,
    do_sample=True,
    top_k=10,
    num_return_sequences=1,
    eos_token_id=tokenizer.eos_token_id,
    max_length=200,
)
for seq in sequences:
    print(f"Result: {seq['generated_text']}")
```