ChatInterface

October 21, 2024

[42]:
```
import gradio as gr
import torch
from transformers import AutoModelForCausalLM, AutoTokenizer

# The tokenizer from Hugging Face's pre-trained DialoGPT model
tokenizer = AutoTokenizer.from_pretrained("microsoft/DialoGPT-small")

# Load the model architecture (DialoGPT)
model = AutoModelForCausalLM.from_pretrained("microsoft/DialoGPT-small")

# Fine Tuned models path
model_path = r'./fine_tuned_dialoGPT/fine_tuned_dialoGPT_epoch3_step3600.pt'
#model_path = r'C:\Temp\USD\AAI-520\Final␣
 ↪Project\Chatbot\fine_tuned_dialoGPT\fine_tuned_dialoGPT_epoch3_step3600.pt'

# Move model to GPU if available
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
# model.to(device)

# Load the fine-tuned model weights using torch.load (since it's a .pt file)
model_state_dict  = torch.load(model_path,map_location=torch.device(device))

# Load the state dictionary into the model
model.load_state_dict(model_state_dict['model_state_dict'])

# Set the model to evaluation mode
model.eval()

# Function to generate a response using the trained DialoGPT model
def generate_response(message, history):

    print("Me: {}".format(message))
    # Tokenize the input and convert it to input_ids
    new_input_ids = tokenizer.encode(message + tokenizer.eos_token,␣
 ↪return_tensors='pt').to(device)

    # Generate a response from the model
```

```
    # Increase the penalty to avoid repetition
    #The higher the value, the more the model is penalized for repeating tokens.
    #Top-k sampling limits the next token to be chosen from the top k most␣
 ↪probable tokens.
    #Top-p (nucleus) sampling chooses the next token from a dynamically sized␣
 ↪group of tokens with cumulative probabilities adding up to p.
    #The temperature parameter controls the randomness of predictions by␣
 ↪scaling the logits before applying softmax.
    response_ids = model.generate(new_input_ids,
                                  max_length=128,
                                  pad_token_id=tokenizer.eos_token_id,
                                  repetition_penalty=1.2,    # Repetition penalty
                                  top_k=50,                  # Top-k sampling
                                  top_p=0.95,                # Nucleus sampling
                                  do_sample=True,            # Enable sampling
                                  no_repeat_ngram_size=3,    # Prevent repeating␣
 ↪3-word sequences
                                  temperature=0.7            # Lower temperature␣
 ↪for more focused response, Control randomness
                                 )

    # Decode the generated response and clean up special tokens like <EOS>
    response = tokenizer.decode(response_ids[:, new_input_ids.shape[-1]:][0],␣
 ↪skip_special_tokens=True)

    # Clean the response to remove any unexpected tokens or unwanted EOS␣
 ↪markers. This limited the display within the chatbot
    #to the EOS marker otherwise.
    response = response.replace('<EOS>', '').strip()

    print("ChatFlix: {}".format(response))

    return response

# Chat Interface in Gradio
interface = gr.ChatInterface(
    fn=generate_response,
    title="ChatFlix using DialoGPT",
    description="A chatbot based on the DialoGPT model, fine-tuned for␣
 ↪multi-turn conversations.",
    examples=[{"text": "Hello", "files": []}],
    multimodal=False
)

interface.launch()
```

```
C:\Users\mthiruma\AppData\Local\Temp\ipykernel_23900\3223938444.py:20:
```

```
FutureWarning: You are using `torch.load` with `weights_only=False` (the current
default value), which uses the default pickle module implicitly. It is possible
to construct malicious pickle data which will execute arbitrary code during
unpickling (See
https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models for
more details). In a future release, the default value for `weights_only` will be
flipped to `True`. This limits the functions that could be executed during
unpickling. Arbitrary objects will no longer be allowed to be loaded via this
mode unless they are explicitly allowlisted by the user via
`torch.serialization.add_safe_globals`. We recommend you start setting
`weights_only=True` for any use case where you don't have full control of the
loaded file. Please open an issue on GitHub for any issues related to this
experimental feature.
  model_state_dict  = torch.load(model_path,map_location=torch.device(device))
C:\Users\mthiruma\AppData\Local\Programs\Python\Python312\Lib\site-
packages\gradio\components\chatbot.py:228: UserWarning: The 'tuples' format for
chatbot messages is deprecated and will be removed in a future version of
Gradio. Please set type='messages' instead, which uses openai-style 'role' and
'content' keys.
  warnings.warn(

* Running on local URL:  http://127.0.0.1:7886

To create a public link, set `share=True` in `launch()`.

<IPython.core.display.HTML object>
```

[42]:

```
Me: hi
ChatFlix: how did you know?  i am the one with a job. but do not tell anybody!
it is all about me, man… what are we doing here at home tonight anyway and how
will they make them feel like this for us to come back on time again tomorrow
night captain morgan?! he got nothing better than that right now.. wait till
next year there was no other choice yet after midnight in your office room…now
why does everything have any purpose of my own life these days ago last week
evening or something crazy too soon before lunchtime today afternoon morning
daynight never mind birthday nights later
Me: what are birthday nights?
ChatFlix: i am at a party. you know, my…  i do not like that one very much? it
is the only way in! what about other days of your life for me to go home and
watch television with him her back there on this day afternoon night time too
late tomorrow morning dinner tonight evening lunch midnight mary peter
killnight..! any advice out here yet or something around then again right now…
hehe will be fine up by 1230 we shall have an alligator hunting accident soon
enough usway down town later today baby jake nicklin wakeup 002037
Me: how about evening lunch , lengthy inputs, peter killnight?  what is
happening in downtown?
ChatFlix: yes.  i am on the way to work… this will not be easy for me if you
```

```
do it right now! how are we going?! can somebody help us out here please and see
where all these people come from around town tonight with their new job tomorrow
morning afternoon at noon start time of day night early dinner timeside working
today midnight wake up already… so ready yet again have a date there too
late no thanks sir jackie barney just like that one did last week ago never mind
him anyway
```

[ ]: