CS 59000 Application of Deep Learning

Homework 4: Hugging Face and Google Cloud

Due on **October 2**, 2024, 11:59pm

The goal of this homework assignment is to help you get familiar with Hugging Face Transformers pipeline, Serverless Inference API, Spaces, and Google Cloud VM. Please create an account at https://huggingface.co/. For parts 1 – 3, please refer to OpenAI Whisper Small model at https://huggingface.co/openai/whisper-small.

1. **Transformers Pipeline**. Create a Jupyter Notebook "parts_1_2.ipynb" file at Google Colab and run the code in the section of "Long-Form Transcription" for the OpenAI Whisper Small model. Please play out the audio inside the Jupyter Notebook.
2. **Serverless Inference API**. In the same Jupyter Notebook, i.e., "parts_1_2.ipynb", apply whisper-small model to the same audio as Part 1 through serverless inference API. You need to create an access token through your Hugging Face account.
3. **Hugging Face Spaces**. Create a public Hugging Face Space for the application of automatic speech recognition through the OpenAI Whisper Small model. Specifically, the model should apply Transformers pipeline in the app.py code (i.e., refer to Part 1), instead of relying on the "Deploy as a Gradio app in one click" functionality provided by the page. Please put the link to your application at Hugging Face Spaces in the file "parts_3_4.docx". Moreover, upload the entire project, such as "whisper_gradio", containing all files, to your GitHub homework repo under "hw4" folder.
4. **Google Cloud VM**. Refer to the provided file "run_website_at_Google_Cloud.pdf" to run the Gradio app in a Google Cloud VM. Please put the "gradio.live" link (like "https://1c166334a8ebfa2c5c.gradio.live") in the "parts_3_4.docx" file. Please take a screenshot of running the Gradio app (i.e., when run "python3 app.py") in VM and put it in the "parts_3_4.docx". Please keep running the application in your VM at Google Cloud until you receive the score of hw4.

Please upload both "parts_1_2.ipynb" and "parts_3_4.docx" files, as well as the entire project for Part 3, to your GitHub homework repo under "hw4". For the "parts_1_2.ipynb", please make sure to keep all outputs in the file.

Grading rubric:

- Part 1 – 25pts
- Part 2 – 25pts
- Part 3 – 25pts
- Part 4 – 25pts

Total: 100pts