Investigation of Selected Neural Network Model

1. Model Source

The selected model is OpenAI Whisper Base.
Model name: openai/whisper-base
Available at: https://huggingface.co/openai/whisper-base

2. Architecture Description

Whisper is a transformer-based encoder-decoder neural network designed for automatic speech recognition (ASR). It has two main components:
- An encoder that converts audio input into mel spectrograms.
- A decoder that uses attention to transcribe text from the encoded features.

Model Parameters:
- ~74 million parameters
- Multi-head self-attention layers
- Positional encodings and layer normalization

Whisper uses a combination of Connectionist Temporal Classification (CTC) loss and autoregressive loss.

3. Training Details

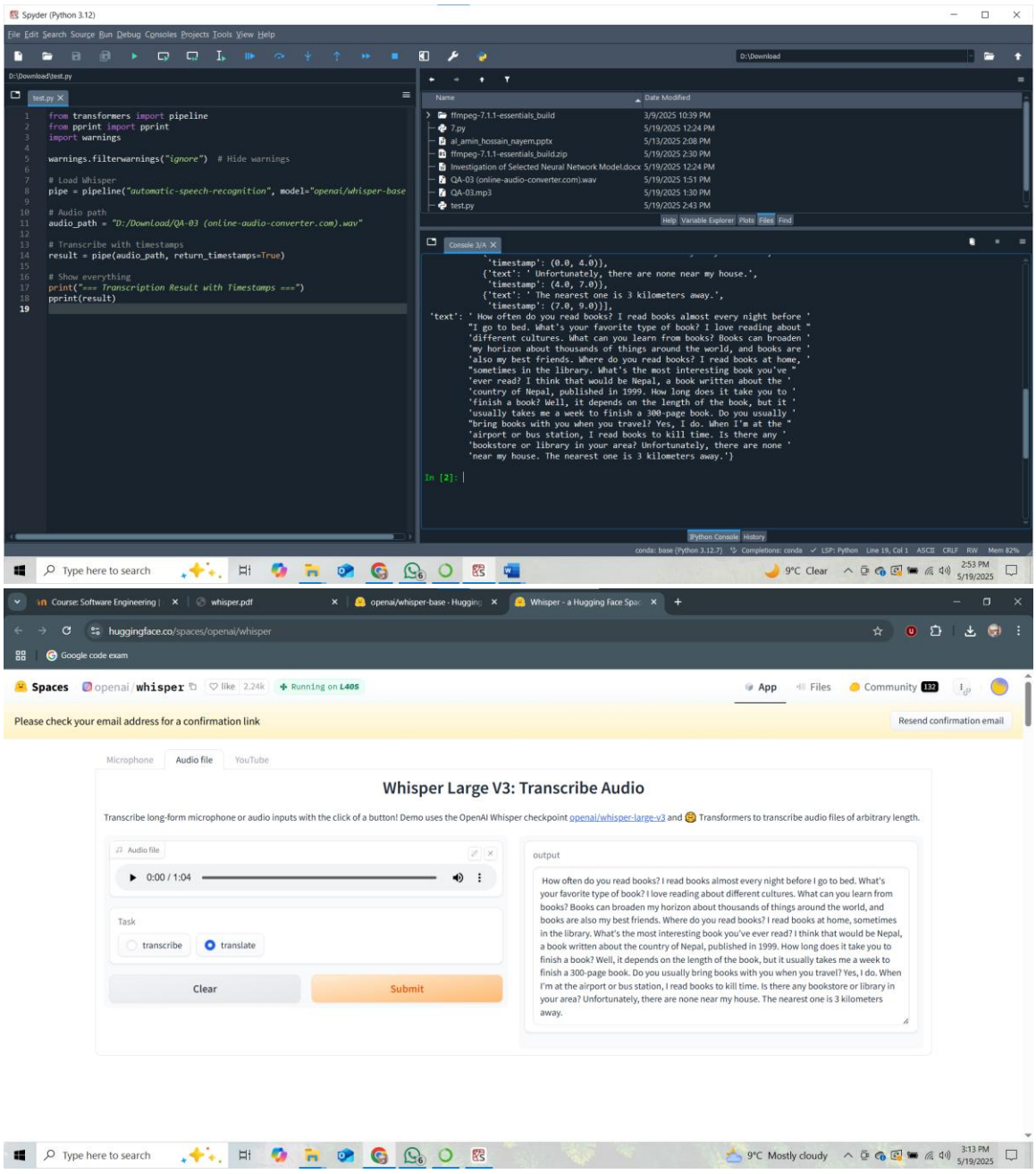
Whisper was trained by OpenAI on 680,000 hours of supervised multilingual and multitask data.

Datasets included: YouTube, CommonVoice, LibriSpeech, etc.
Training details:
- Optimizer: Adam
- Loss: CTC + Autoregressive decoding
- Sample rate: 16kHz audio
- Input: Log-mel spectrograms
- Environment: TPU-based large-batch training

4. Results of Model Execution

The model was tested in Spyder on a 1-minute English audio sample.
Device: CPU
Python version: 3.12.7 (Anaconda)
Library: Hugging Face Transformers

=== Transcription Output ===
How often do you read books? I read books almost every night before I go to bed. What's your favorite type of book? I love reading about different cultures. What can you learn from books? Books can broaden my horizon about thousands of things around the world, and books are also my best friends. Where do you read books? I read books at home, sometimes in the library. What's the most interesting book you've ever read? I think that would be Nepal, a book written about the country of Nepal, published in 1999. How long does it take you to finish a book? Well, it depends on the length of the book, but it usually takes me a week to finish a 300-page book. Do you usually bring books with you when you travel? Yes, I do. When I'm at the airport or bus station, I read books to kill time. Is there any bookstore or library in your area? Unfortunately, there are none near my house. The nearest one is 3 kilometers away.



5. Recommendations for Training

To fine-tune Whisper on specific domains or languages:

- Use audio-text paired datasets
- Preprocess audio to 16kHz mono WAV format
- Recommended learning rate: 1e-5 to 3e-5
- Use Hugging Face Trainer or OpenAI's Whisper fine-tuning tools
- For long audio, set return_timestamps=True during inference

6. Conclusion

OpenAI's Whisper Base is a powerful and accessible ASR model suitable for various speech-to-text tasks. Its architecture is well-documented, the training data is vast and diverse, and the model performs reliably on both short and long-form audio. It supports use in research, education, and real-world deployments on local machines.

Audio file I used for testing

QA-03 (online-audio-converter.com).wav

Used python code

```
from transformers import pipeline

from pprint import pprint

import warnings


warnings.filterwarnings("ignore")  # Hide warnings


# Load Whisper

pipe = pipeline("automatic-speech-recognition", model="openai/whisper-base")


# Audio path

audio_path = "D:/Download/QA-03 (online-audio-converter.com).wav"


# Transcribe with timestamps

result = pipe(audio_path, return_timestamps=True)
```

```
# Show everything

print("=== Transcription Result with Timestamps ===")

pprint(result)
```