Fine-Tuning T5 for Mental Health Explanations: Workflow and Findings

1. Introduction

The goal of this project was to fine-tune the T5 (Text-to-Text Transfer Transformer) model to generate natural language explanations for predicted mental health conditions and suggest coping mechanisms. This approach leverages large language models (LLMs) to enhance conversational AI capabilities in mental health applications.

2. Dataset and Preprocessing

We used a mental health dialogue dataset consisting of user queries (contexts) and professional responses. The dataset underwent the following preprocessing steps:

- **Loading Data:** CSV files containing conversation contexts and responses were loaded.
- **Cleaning Data:**
  - Removed missing values and duplicates.
  - Standardized text formatting.
- **Splitting Data:**
  - Divided into training (80%) and evaluation (20%) sets using train_test_split.

3. Model Selection and Fine-Tuning

- **Model Used:** google/t5-base
- **Tokenization:** Applied AutoTokenizer from the Hugging Face Transformers library to tokenize input texts and responses.
- **Training Process:**
  - Used Seq2SeqTrainer with the following parameters:
    - Epochs: 1+1
    - Batch size: 24
    - Learning rate: 1e-3
    - Weight decay: 0.01
  - Training was performed on GPU for efficiency.
- **Checkpoint Handling:**
  - Saved model and tokenizer after each epoch.
  - Resumed training from the checkpoint of epoch 1 to continue training for epoch 2.
  - Used stored optimizer and scheduler states to maintain training consistency.

4. Evaluation Metrics and Results

- **Metrics Used:**
  - **ROUGE Scores** (measuring text similarity to reference responses)
  - **Loss Reduction Across Epochs**
- **Findings:**
  - **Epoch 1 Results:**
    - Training Loss: 0.88
    - Validation Loss: 0.84
    - ROUGE-1: 0.417
  - **Epoch 2 Results:**
    - Training Loss: 0.8142
    - Validation Loss: 0.7883
    - ROUGE-1: 0.402

5. Predictions and Analysis

- The model generated insightful, human-like responses aligned with therapist-style communication.
- Improvements observed after the second epoch suggest the model benefits from additional training.
- The generated responses were contextually relevant, with minor hallucinations observed in a few cases.
- Predictions improved significantly with checkpoint-based training, reducing inconsistencies and improving response fluency.

6. Conclusion

This fine-tuned T5 model demonstrates promising capabilities in mental health applications by providing empathetic and informative responses. By utilizing checkpoint-based training and careful evaluation, the model has shown consistent improvements across epochs. Further refinements and testing in real-world scenarios can enhance its effectiveness in therapeutic settings.