WikiHow Heading Generation: Explanatory Documentation

Project Overview

The WikiHow Heading Generation project is a machine learning application designed to automatically generate headings for WikiHow articles. This tool can be incredibly useful for content creators, editors, and anyone working with enormous amounts of text.

Key Features

1. Utilizes a fine-tuned LED (Language Model for Dialogue Applications) model from Hugging Face
2. Trains on WikiHow article data
3. Generates headings for new paragraphs

Technical Details

Model Architecture

The project uses the LED model, which is particularly well-suited for processing long documents. It's fine-tuned on WikiHow data to specialize in generating concise, informative headings.

Data Structure

The training data is stored in a CSV file with three main columns:

1. Article Title: The overall title of the WikiHow article

2. Subheading: The heading of a specific section within the article

3. Paragraph: The actual text content of the section

This structure allows the model to learn the relationship between paragraph content and appropriate headings.

Setup and Installation

Prerequisites

Before using this project, ensure you have Python installed on your system. The code is designed to run in a Jupyter notebook environment, which is excellent for interactive development and testing.

Required Packages

The project depends on several Python packages:

- transformers: For accessing and using the LED model
- datasets: For handling and processing the WikiHow dataset
- pandas: For data manipulation and analysis
- rouge_score: For evaluating the quality of generated headings
- matplotlib: For any visualization needs
- torch: The underlying deep learning framework

Install these packages using pip:

```
pip install transformers datasets pandas rouge_score matplotlib torch
```

Usage Guide

Training the Model

1. Prepare your WikiHow dataset in CSV format.

2. Load the data using pandas:

df = pd.read_csv('./wikiHow.csv')

3. Set up the LED model and tokenizer:

tokenizer = LEDTokenizer.from_pretrained("allenai/led-base-16384")

model = LEDForConditionalGeneration.from_pretrained("allenai/led-base-16384")

4. Configure and initialize the Seq2SeqTrainer:

training_args = Seq2SeqTrainingArguments(...)

trainer = Seq2SeqTrainer(model=model, args=training_args, ...)

5. Start the training process:

trainer.train()

Generating Headings

1. **Load the trained model and tokenizer:**

tokenizer = LEDTokenizer.from_pretrained("/content/checkpoint-60")

model = LEDForConditionalGeneration.from_pretrained("/content/checkpoint-60").to("cuda").half()

2. **Prepare your input paragraph:**

```
  sample_paragraph = "Your paragraph text here"

df = pd.DataFrame([sample_paragraph], columns=['Paragraph'])

df_test = Dataset.from_pandas(df)
```

3. **Use the `generate_answer` function to create headings:**

```
result = df_test.map(generate_answer, batched=True, batch_size=2)

print(result["generated_heading"])
```

Evaluation

The project uses the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric to assess the quality of generated headings. ROUGE compares the generated headings with reference headings, measuring their similarity in terms of overlapping words or n-grams.

To evaluate your model:

1. Generate headings for a test set
2. Compare the generated headings with the actual headings using the ROUGE metric
3. Analyze the ROUGE scores to understand the model's performance

By following this documentation, you should be able to understand, set up, and use the Wiki How Heading Generation project effectively. <The repo doesn't contain model files due to its storage size.>