

Text Generation with GPT-2

This repository contains scripts for fine-tuning the GPT-2 model for text generation tasks, particularly focusing on natural language generation using the Hugging Face Transformers library. The process includes data preparation, model training, and evaluation using ROUGE metrics for generated text.

Requirements

Before you run the scripts, ensure you have the following installed:

- Python 3.6+
- PyTorch
- Transformers
- Pandas
- NumPy
- Scikit-Learn

You can install the necessary libraries using pip:

```
>>pip install torch transformers pandas numpy scikit-learn
```

Files Description

train.py: Contains the main script for training the GPT-2 model using custom datasets.

dataset.py: Contains a custom PyTorch dataset class CustomTextDataset, which prepares text data for training.

utils.py: Utility functions for data collation and manipulation.

evaluate.py: Functions to calculate and print average ROUGE scores for model evaluation.

Usage

Training the Model

Prepare your dataset in a CSV format and name the text column as "Text".

Modify the script train.py to point to your dataset file path.

Set the desired training parameters in the script.

Evaluating the Model

After training, use the generated text to evaluate the model using ROUGE metrics included in the script.

Ensure that your evaluation dataset includes both the generated summaries and reference summaries.

```
# Train the model
```

```
train(train_data, "gpt2", "./gpt2_finetuned", True, 8, 10, 10000)
```

```
# Calculate average ROUGE scores
```

```
average_rouge_scores = calculate_average_rouge_scores(rouge_scores)
```

```
-----  
-----
```

```
# Print the average ROUGE scores
```

```
print("Average ROUGE-1:")
```

```
print(f"Precision: {average_rouge_scores['rouge-1']['precision']:.2f}, "
```

```
      f"Recall: {average_rouge_scores['rouge-1']['recall']:.2f}, "
```

```
      f"F1-Score: {average_rouge_scores['rouge-1']['f1']:.2f}")
```