


```
In [1]: !pip install transformers datasets rouge-score nltk torch
```

Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: transformers in ./local/lib/python3.9/site-packages (4.46.2)
Requirement already satisfied: datasets in ./local/lib/python3.9/site-packages (3.1.0)
Requirement already satisfied: rouge-score in ./local/lib/python3.9/site-packages (0.1.2)
Requirement already satisfied: nltk in /apps/cent7/jupyterhub/lib/python3.9/site-packages (3.6.5)
Requirement already satisfied: torch in ./local/lib/python3.9/site-packages (2.5.1)
Requirement already satisfied: filelock in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from transformers) (3.3.1)
Requirement already satisfied: safetensors>=0.4.1 in ./local/lib/python3.9/site-packages (from transformers) (0.4.5)
Requirement already satisfied: pyyaml>=5.1 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from transformers) (6.0)
Requirement already satisfied: numpy>=1.17 in ./local/lib/python3.9/site-packages (from transformers) (1.21.2)
Requirement already satisfied: regex!=2019.12.17 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from transformers) (2021.8.3)
Requirement already satisfied: tqdm>=4.27 in ./local/lib/python3.9/site-packages (from transformers) (4.67.0)
Requirement already satisfied: packaging>=20.0 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from transformers) (21.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.23.2 in ./local/lib/python3.9/site-packages (from transformers) (0.26.2)
Requirement already satisfied: tokenizers<0.21,>=0.20 in ./local/lib/python3.9/site-packages (from transformers) (0.20.3)
Requirement already satisfied: requests in ./local/lib/python3.9/site-packages (from transformers) (2.32.3)
Requirement already satisfied: pandas in ./local/lib/python3.9/site-packages (from datasets) (2.0.3)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in ./local/lib/python3.9/site-packages (from datasets) (0.3.8)
Requirement already satisfied: multiprocessing<0.70.17 in ./local/lib/python3.9/site-packages (from datasets) (0.70.16)
Requirement already satisfied: pyarrow>=15.0.0 in ./local/lib/python3.9/site-packages (from datasets) (18.0.0)
Requirement already satisfied: xxhash in ./local/lib/python3.9/site-packages (from datasets) (3.5.0)
Requirement already satisfied: fsspec[http]<=2024.9.0,>=2023.1.0 in ./local/lib/python3.9/site-packages (from datasets) (2024.9.0)
Requirement already satisfied: aiohttp in ./local/lib/python3.9/site-packages (from datasets) (3.10.10)
Requirement already satisfied: absl-py in ./local/lib/python3.9/site-packages (from rouge-score) (2.1.0)
Requirement already satisfied: six>=1.14.0 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from rouge-score) (1.16.0)
Requirement already satisfied: click in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from nltk) (8.0.3)
Requirement already satisfied: joblib in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from nltk) (1.1.0)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in ./local/lib/python3.9/site-packages (from torch) (12.3.1.170)
Requirement already satisfied: sympy==1.13.1 in ./local/lib/python3.9/site-packages (from torch) (1.13.1)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in ./local/lib/python3.9/site-packages (from torch) (2.21.5)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in ./local/lib/python3.9/site-packages (from torch) (11.2.1.3)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in ./local/lib/python3.9/site-packages (from torch) (12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in ./local/lib/python3.9/site-packages (from torch) (9.1.0.70)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in ./local/lib/python3.9/site-p

ackages (from torch) (12.4.127)
Requirement already satisfied: jinja2 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from torch) (2.11.3)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in ./local/lib/python3.9/site-packages (from torch) (12.4.127)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in ./local/lib/python3.9/site-packages (from torch) (10.3.5.147)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in ./local/lib/python3.9/site-packages (from torch) (12.4.5.8)
Requirement already satisfied: typing-extensions>=4.8.0 in ./local/lib/python3.9/site-packages (from torch) (4.12.2)
Requirement already satisfied: networkx in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from torch) (2.6.3)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in ./local/lib/python3.9/site-packages (from torch) (12.4.127)
Requirement already satisfied: triton==3.1.0 in ./local/lib/python3.9/site-packages (from torch) (3.1.0)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in ./local/lib/python3.9/site-packages (from torch) (11.6.1.9)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in ./local/lib/python3.9/site-packages (from torch) (12.4.127)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from sympy==1.13.1->torch) (1.2.1)
Requirement already satisfied: aiohttp>=1.1.2 in ./local/lib/python3.9/site-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: yarl<2.0,>=1.12.0 in ./local/lib/python3.9/site-packages (from aiohttp->datasets) (1.17.1)
Requirement already satisfied: multidict<7.0,>=4.5 in ./local/lib/python3.9/site-packages (from aiohttp->datasets) (6.1.0)
Requirement already satisfied: attrs>=17.3.0 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from aiohttp->datasets) (21.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in ./local/lib/python3.9/site-packages (from aiohttp->datasets) (1.5.0)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in ./local/lib/python3.9/site-packages (from aiohttp->datasets) (2.4.3)
Requirement already satisfied: async-timeout<5.0,>=4.0 in ./local/lib/python3.9/site-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: pyparsing>=2.0.2 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from packaging>=20.0->transformers) (3.0.4)
Requirement already satisfied: idna<4,>=2.5 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from requests->transformers) (3.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from requests->transformers) (2.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from requests->transformers) (2021.10.8)
Requirement already satisfied: urllib3<3,>=1.21.1 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from requests->transformers) (1.26.7)
Requirement already satisfied: propcache>=0.2.0 in ./local/lib/python3.9/site-packages (from yarl<2.0,>=1.12.0->aiohttp->datasets) (0.2.0)
Requirement already satisfied: MarkupSafe>=0.23 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from jinja2->torch) (1.1.1)
Requirement already satisfied: python-dateutil>=2.8.2 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /apps/cent7/jupyterhub/lib/python3.9/site-packages (from pandas->datasets) (2021.3)
Requirement already satisfied: tzdata>=2022.1 in ./local/lib/python3.9/site-packages (from pandas->datasets) (2024.2)

```
In [2]: # Imports
from datasets import load_dataset
from transformers import T5Tokenizer, T5ForConditionalGeneration, AdamW
from rouge_score import rouge_scorer
from torch.utils.data import DataLoader
from torch import nn
import torch
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.translate.bleu_score import sentence_bleu, SmoothingFunction
import os

# Ensure necessary NLTK data is downloaded
nltk.download('stopwords')
nltk.download('wordnet')

# Device setting
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
print(f"Using device: {device}")
```

```
/home/desai226/.local/lib/python3.9/site-packages/transformers/loss/loss_for_object_detection.py:28: UserWarning: A NumPy version >=1.22.4 and <2.3.0 is required for this version of SciPy (detected version 1.21.2)
  from scipy.optimize import linear_sum_assignment
```

Using device: cuda

```
[nltk_data] Downloading package stopwords to
[nltk_data]   /home/desai226/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /home/desai226/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

```
In [3]: # Load and Prepare Model
t5_model_name = 't5-base'
print(f"Loading T5 model: {t5_model_name}")
tokenizer = T5Tokenizer.from_pretrained(t5_model_name)
model = T5ForConditionalGeneration.from_pretrained(t5_model_name).to(device)
```

Loading T5 model: t5-base

spiece.model: 100% 792k/792k [00:00<00:00, 13.1MB/s]

tokenizer.json: 100% 1.39M/1.39M [00:00<00:00, 28.5MB/s]

config.json: 100% 1.21k/1.21k [00:00<00:00, 74.7kB/s]

You are using the default legacy behaviour of the <class 'transformers.models.t5.tokenization_t5.T5Tokenizer'>. This is expected, and simply means that the `legacy` (previous) behavior will be used so nothing changes for you. If you want to use the new behaviour, set `legacy=False`. This should only be set if you understand what it means, and thoroughly read the reason why this was added as explained in <https://github.com/huggingface/transformers/pull/24565> (<https://github.com/huggingface/transformers/pull/24565>)

model.safetensors: 100% 892M/892M [00:04<00:00, 193MB/s]

generation_config.json: 100% 147/147 [00:00<00:00, 8.04kB/s]

```
In [11]: # Initialize Lemmatizer and stopwords
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))

# Preprocessing function
def preprocess_text(text):
    text = re.sub(r'^[a-zA-Z\s]', '', text) # Remove special characters
    text = re.sub(r'\s+', ' ', text).strip() # Remove extra spaces
    words = text.split()
    words = [lemmatizer.lemmatize(word.lower()) for word in words if word.lower() not in stop_words]
    return ' '.join(words)

# Function to remove rows with missing data
def remove_missing_data(dataset, columns):
    return dataset.filter(lambda x: all(x[col] is not None for col in columns))

# Function to load and clean dataset
def load_and_clean_data():
    train_data = load_dataset("ragha92/FNS_Summarization", split="train")
    validation_data = load_dataset("ragha92/FNS_Summarization", split="validation")
    test_data = load_dataset("ragha92/FNS_Summarization", split="test")

    columns = ['Annual Reports', 'Gold Summaries']
    train_data = remove_missing_data(train_data, columns)
    validation_data = remove_missing_data(validation_data, columns)
    test_data = remove_missing_data(test_data, columns)

    return train_data, validation_data, test_data
```

```

In [5]: # Summarization function for T5 with updated hyperparameters
def summarize_text_t5(article):
    input_text = f"summarize: {article}"
    input_ids = tokenizer.encode(input_text, return_tensors="pt", max_length=512, truncatio
    summary_ids = model.generate(
        input_ids,
        max_length=512,
        min_length=200,
        length_penalty=1.5,
        num_beams=3,
        early_stopping=True
    )
    summary = tokenizer.decode(summary_ids[0], skip_special_tokens=True)
    return summary

# Checkpoint save/load functions
def save_checkpoint(model, optimizer, epoch, path="t5_checkpoint.pt"):
    torch.save({
        'epoch': epoch,
        'model_state_dict': model.state_dict(),
        'optimizer_state_dict': optimizer.state_dict(),
    }, path)
    print(f"Checkpoint saved at epoch {epoch}")

def load_checkpoint(path="t5_checkpoint.pt"):
    checkpoint = torch.load(path)
    model.load_state_dict(checkpoint['model_state_dict'])
    optimizer.load_state_dict(checkpoint['optimizer_state_dict'])
    epoch = checkpoint['epoch']
    print(f"Checkpoint loaded. Starting from epoch {epoch + 1}")
    return epoch

```

```

In [6]: # Function to fine-tune the model
def fine_tune_model(train_data, epochs=3, learning_rate=5e-5, save_interval=1):
    model.train()
    optimizer = AdamW(model.parameters(), lr=learning_rate)
    dataloader = DataLoader(train_data, batch_size=2, shuffle=True)

    starting_epoch = 0
    if "t5_checkpoint.pt" in os.listdir():
        starting_epoch = load_checkpoint("t5_checkpoint.pt")

    for epoch in range(starting_epoch, epochs):
        total_loss = 0
        for batch_idx, batch in enumerate(dataloader):
            optimizer.zero_grad()
            input_texts = batch['Annual Reports']
            target_summaries = batch['Gold Summaries']

            inputs = tokenizer(input_texts, return_tensors="pt", max_length=512, truncation=True)
            targets = tokenizer(target_summaries, return_tensors="pt", max_length=512, truncation=True)

            outputs = model(input_ids=inputs, labels=targets)
            loss = outputs.loss
            loss.backward()
            optimizer.step()
            total_loss += loss.item()

        if (epoch + 1) % save_interval == 0:
            save_checkpoint(model, optimizer, epoch, path="t5_checkpoint.pt")

        print(f"Epoch {epoch+1}/{epochs} Average Loss: {total_loss / len(dataloader)}")

```

```

In [7]: # Evaluation using ROUGE and BLEU with smoothing
def evaluate_model(dataset):
    scorer = rouge_scorer.RougeScorer(['rouge1', 'rouge2', 'rougeL'], use_stemmer=True)
    scores = {'rouge1': [], 'rouge2': [], 'rougeL': [], 'bleu': []}
    smooth = SmoothingFunction().method4

    for item in dataset:
        article = preprocess_text(item['Annual Reports'])
        gold_summary = preprocess_text(item['Gold Summaries'])
        if article:
            generated_summary = summarize_text_t5(article)
            score = scorer.score(gold_summary, generated_summary)
            scores['rouge1'].append(score['rouge1'].fmeasure)
            scores['rouge2'].append(score['rouge2'].fmeasure)
            scores['rougeL'].append(score['rougeL'].fmeasure)
            reference = gold_summary.split()
            candidate = generated_summary.split()
            bleu_score = sentence_bleu([reference], candidate, smoothing_function=smooth)
            scores['bleu'].append(bleu_score)

    avg_rouge1 = sum(scores['rouge1']) / len(scores['rouge1'])
    avg_rouge2 = sum(scores['rouge2']) / len(scores['rouge2'])
    avg_rougeL = sum(scores['rougeL']) / len(scores['rougeL'])
    avg_bleu = sum(scores['bleu']) / len(scores['bleu'])
    return avg_rouge1, avg_rouge2, avg_rougeL, avg_bleu

```



```
In [8]: train_data, validation_data, test_data = load_and_clean_data()
print("Datasets loaded and cleaned.")
```

Datasets loaded and cleaned.

```
In [9]: fine_tune_model(train_data, epochs=3)
```

```
/home/desai226/.local/lib/python3.9/site-packages/transformers/optimization.py:591: Future
Warning: This implementation of AdamW is deprecated and will be removed in a future versio
n. Use the PyTorch implementation torch.optim.AdamW instead, or set `no_deprecation_warnin
g=True` to disable this warning
```

```
warnings.warn(
Passing a tuple of `past_key_values` is deprecated and will be removed in Transformers v4.
48.0. You should pass an instance of `EncoderDecoderCache` instead, e.g. `past_key_values=
EncoderDecoderCache.from_legacy_cache(past_key_values)`.
```

```
Checkpoint saved at epoch 0
Epoch 1/3 Average Loss: 3.501926212461249
Checkpoint saved at epoch 1
Epoch 2/3 Average Loss: 3.083557219354852
Checkpoint saved at epoch 2
Epoch 3/3 Average Loss: 2.942903983101104
```

```
In [10]: avg_rouge1, avg_rouge2, avg_rougeL, avg_bleu = evaluate_model(validation_data)
print(f"Validation ROUGE-1: {avg_rouge1}, ROUGE-2: {avg_rouge2}, ROUGE-L: {avg_rougeL}, BLE
avg_rouge1, avg_rouge2, avg_rougeL, avg_bleu = evaluate_model(test_data)
print(f"Test Set ROUGE-1: {avg_rouge1}, ROUGE-2: {avg_rouge2}, ROUGE-L: {avg_rougeL}, BLEU:
```

```
The history saving thread hit an unexpected error (OperationalError('unable to open databa
se file')).History will not be written to the database.
```

```
Validation ROUGE-1: 0.2336938790736788, ROUGE-2: 0.1345571295661105, ROUGE-L: 0.1613417694
108391, BLEU: 0.08928334216352536
```

```
Test Set ROUGE-1: 0.21713754279472247, ROUGE-2: 0.11761398051899112, ROUGE-L: 0.1452819241
0298404, BLEU: 0.06995625174256394
```

```

In [ ]: # Function to display original reports, generated summaries, and true labels
def display_model_output(dataset, num_samples=3):
    for i in range(num_samples):
        # Get the ith sample
        report = dataset[i]['Annual Reports']
        true_summary = dataset[i]['Gold Summaries']

        # Preprocess and summarize
        generated_summary = summarize_text_t5(preprocess_text(report))

        # Display the output
        print(f"Sample {i + 1}")
        print("-" * 50)
        print("Original Report:")
        print(report[:1000] + "..." if len(report) > 1000 else report) # Display first 1000
        print("\nTrue Summary (Gold Label):")
        print(true_summary)
        print("\nGenerated Summary:")
        print(generated_summary)
        print("=" * 50)

# Display the output for three reports from the validation dataset
display_model_output(validation_data, num_samples=3)

```