KaggleX- Showcase

Text Summarization Using Large Language Models (LLMs)

Lochan Basyal

Background

- Machine Learning Engineer.

- Exploring deep learning, computer vision and natural language processing.

- MS in Computer Engineering (Artificial Intelligence) from Stevens Institute of Technology.

Socials

https://www.linkedin.com/in/lochanbasyal/

https://github.com/lbasyal

kaggle

Project Definition

This project primarily focuses on a comparative study of text summarization inferences based on two open-source datasets and three fine-tuned Large Language Models (LLMs).

Datasets:

- CNN Dailymail and Extreme Summarization(Xsum)

Models:

- MPT-7b-instruct, Falcon-7b-instruct, and Text-davinci-003

Furthermore, this project explores the application side of Text Summarization.

Project Details

LLM Model	Dataset	Avg. Word Count	ROUGE-1	ROUGE-2	ROUGE-L	BERT Score (P/R/F1)
falcon-7b-instruct	CNN (n=25)	784.24	0.226	0.053	0.197	0.818 / 0.860 / 0.838
falcon-7b-instruct	XSum (n=25)	410.44	0.139	0.014	0.113	0.787 / 0.863 / 0.823
mpt-7b-instruct	CNN (n=25)	784.24	0.236	0.060	0.213	0.839 / 0.864 / 0.851
mpt-7b-instruct	XSum (n=25)	410.44	0.159	0.024	0.133	0.828 / 0.871 / 0.848
text-davinci-003	CNN (n=25)	784.24	0.272	0.096	0.255	0.854 / 0.883 / 0.868
text-davinci-003	XSum (n=25)	410.44	0.206	0.053	0.173	0.844 / 0.893 / 0.868

TABLE I: Performance Metrics of LLMs on "CNN/Daily Mail 3.0.0" and "XSum" Datasets

LLM Model	Dataset	Avg. Word Count	BLEU Score
falcon-7b-instruct	CNN (n=25)	784.24	9.4726138403298e-232
falcon-7b-instruct	XSum (n=25)	410.44	9.225829346520394e-232
mpt-7b-instruct	CNN (n=25)	784.24	9.35328936831654e-232
mpt-7b-instruct	XSum (n=25)	410.44	9.542118736121376e-232
text-davinci-003	CNN (n=25)	784.24	0.4896200481408649
text-davinci-003	XSum (n=25)	410.44	0.48979461356547943

TABLE II: Performance Metrics, BLEU Score of LLMs on "CNN/Daily Mail 3.0.0" and "XSum" Datasets

kaggle

Project Examples

Reference summary and Text-davinci-003 model generated summary

In [19]:

```
for i in range(25):
    model_summary = test_df['model_generated'][i]
    reference_summary = test_df['summary'][i]
    print(f"{i + 1} — Reference Summary: {reference_summary}\nModel Summary: {model_summary}\n")
```

1 – Reference Summary: There is a "chronic" need for more housing for prison leavers in Wales, according to a chari
ty.
Model Summary: Prison Link Cymru, a charity that helps people find accommodation after their release, had 1,099 ref
errals in 2015–16 and said some ex-offenders were living rough for up to a year before finding suitable accommodati
on. Changes to the Housing Act in Wales, introduced in 2015, removed the right for prison leavers to be given prior
ity for accommodation. Andrew Stevens, who works in Welsh prisons trying to secure housing for prison leavers, said
the need for accommodation was "chronic"

2 – Reference Summary: A man has appeared in court after firearms, ammunition and cash were seized by police in Edi
nburgh.
Model Summary: On Wednesday, police officers searched properties in two areas of Edinburgh, Scotland and recovered
three firearms, ammunition, and a large sum of money. A 26-year-old man was arrested and charged in connection with
the incident and appeared in court the following day.

kaggle

Project Examples

Summary of YouTube Video Comments.

In [54]:
```python
import transformers
text = fetch_comments(url)
tokenizer = transformers.GPT2TokenizerFast.from_pretrained("gpt2")
chunks = text_to_chunks(text, tokenizer)
print("Chunks list size: ", len(chunks))    # TODO: Remove this
# summary = summarize_comment(chunks,524,0.1)
# print(summary)
```

Chunks list size: 1

In [55]:
```python
summary = summarize_comment(chunks,524,0.1)
print(summary)
```

The comments are generally positive and praise Steve Jobs for his inspiring speech. Many people mention the impact of ISKCON (Hare Krishna) on his thoughts and life, and comment on his blend of artistic and scientific abilities. The comments also mention the impact of the speech on their lives, and how it has helped them stay focused and motivated.

In [56]:
```python
summary = summarize_comment(chunks,524,0.7)
print(summary)
```

The comments are in response to a commencement speech given by Steve Jobs. The comments praise Jobs for his contributions to the world, his blend of artistic and scientific thinking, his inspirational message, and his intellect. They also mention his speech's impact on the lives of those who have watched it.

Project Links

- Research Paper: https://arxiv.org/abs/2310.10449

- GitHub Repo: https://github.com/lbasyal/LLMs-Text-Summarization

Thank You