Project Title: Benchmarking LLMs: A Study on Document Understanding Capabilities
Student: Vivek Vardhan Reddy Kumtam
Supervisor: Prof. Zeyun Yu

Table of Contents

Abstract

This project provides an in-depth exploration of the capabilities of large language models (LLMs) in addressing three core natural language processing (NLP) tasks: summarization, question answering, and information extraction. With the increasing complexity of modern data, these tasks have become vital across industries such as healthcare, finance, and legal systems. This study evaluates state-of-the-art models like PEGASUS, BERT, and BioBERT, fine-tuning them on domain-specific datasets to maximize task performance. Employing advanced optimization strategies like gradient checkpointing and mixed-precision training, the study emphasizes computational efficiency alongside accuracy. The results showcase the potential of LLMs to revolutionize NLP applications, highlighting the importance of domain adaptation and optimization for achieving state-of-the-art outcomes.

Introduction

The advent of large language models (LLMs) has revolutionized the field of natural language processing, enabling sophisticated solutions to tasks like summarization, question answering, and information extraction. These tasks play pivotal roles in industries ranging from healthcare and finance to legal and scientific research. However, while general-purpose models like GPT and BERT achieve remarkable results on diverse datasets, their performance often falters when applied to domain-specific tasks without fine-tuning.

This project aims to bridge this gap by benchmarking state-of-the-art LLMs on specialized datasets across three tasks. Summarization condenses large texts into concise summaries, making it invaluable for information-dense domains. Question answering enables retrieval of precise answers from context, facilitating decision-making. Information extraction identifies and categorizes entities in text, structuring unorganized data for further analysis. Through systematic evaluation, this study seeks to optimize LLMs for these tasks and explore their potential for domain-specific applications.

Objectives

The overarching goals of this project are:

1. To evaluate the performance of cutting-edge LLMs on summarization, question answering, and information extraction tasks.
2. To examine the impact of domain-specific fine-tuning on model performance and reliability.
3. To identify optimization techniques that enhance computational efficiency and accuracy.

Background and Related Work

The rapid advancements in NLP have been driven by the emergence of LLMs like BERT, GPT, and T5, which leverage large-scale pre-training on diverse corpora to capture intricate linguistic patterns. These models excel in general-purpose applications but often require fine-tuning to address the nuances of domain-specific tasks. Domain-adapted models such as BioBERT and LegalBERT have demonstrated the value of pre-training on specialized corpora, showcasing superior performance in biomedical and legal text analysis compared to their general-purpose counterparts.

Previous research has extensively highlighted the capabilities of abstractive summarization models, with PEGASUS emerging as a leader in generating coherent and concise summaries. Studies indicate that models like BERT perform exceptionally well in extractive question answering, thanks to their robust contextual embeddings. For information extraction, models such as BioBERT and SciBERT have proven effective in recognizing entities in structured formats, particularly in scientific and medical domains.

Despite these advancements, challenges persist. Generalization across diverse domains remains a significant hurdle, as models fine-tuned on specific datasets often struggle with unseen contexts. Additionally, computational efficiency is a growing concern, necessitating the development of techniques like gradient checkpointing and mixed-precision training to manage memory usage and accelerate training. This project builds upon these findings to benchmark LLMs across tasks, offering a comprehensive analysis of their strengths, limitations, and optimization potential.

Approach

The approach to benchmarking large language models (LLMs) in this project was meticulously structured, integrating well-defined stages of task selection, dataset preparation, model fine-tuning, and evaluation. Each stage was tailored to align with the specific needs of summarization, question answering, and information extraction tasks while addressing challenges of domain-specific applications and computational constraints.

Task Selection
The project focused on three pivotal NLP tasks that represent key challenges and practical applications in natural language understanding:

1. **Summarization**: This task was designed to condense long, verbose texts into concise, coherent summaries. The primary goal was to evaluate how well different models perform on generating abstractive summaries that preserve the semantic integrity of the original text. Models such as PEGASUS, BART, and DistilBART were selected for their strong capabilities in both abstractive and extractive summarization techniques.

2. **Question Answering (QA)**: QA involves retrieving precise answers from a given context, a critical task in domains requiring fact verification and automated assistance. Pre-trained models like BERT, ALBERT, and DistilBERT were chosen for their

demonstrated ability to excel in extractive question answering tasks, especially on datasets like SQuAD v2 and MedMCQA.

3. **Information Extraction**: This task concentrated on named entity recognition (NER), which involves identifying and categorizing entities such as names, dates, or biomedical terms in text. BioBERT, DistilBERT, and BERT-base were utilized for their robustness in handling structured and domain-specific data.

Dataset Preparation

Comprehensive dataset preparation was a cornerstone of the project, ensuring compatibility and meaningful evaluation for each task:

- **Summarization**: Text summarization tasks used datasets such as XSum, CNN/DailyMail, SAMSum, and PubMed. Preprocessing involved tokenizing the text, truncating long articles to fit model input limits, and preparing corresponding summaries for model training.
- **Question Answering**: QA datasets like SQuAD v2 were aligned with their respective contexts and answers. Special attention was given to handling unanswerable questions, ensuring that start and end positions for tokens were accurately calculated.
- **Information Extraction**: For NER tasks, annotated datasets like CoNLL-2003 and biomedical-specific datasets were utilized. The preprocessing included mapping entities to their labels and ensuring token alignment for accurate entity recognition.

Model Fine-Tuning

Fine-tuning was the core experimental phase, leveraging advanced frameworks and techniques to adapt pre-trained models to specific tasks and datasets. The Hugging Face Transformers library was used extensively for this purpose. Several strategies were employed to enhance performance:

1. **Gradient Checkpointing**: Enabled efficient memory management, allowing larger models and batch sizes to be utilized on GPUs with limited memory.
2. **Mixed Precision Training**: Used FP16 computations to accelerate training and reduce resource consumption.
3. **Dynamic Learning Rates**: Cosine learning rate schedulers were implemented to achieve gradual convergence, stabilizing model training across epochs.
4. **Task-Specific Adjustments**: For summarization, beam search parameters were optimized to balance length and coherence of outputs. For QA, metrics like F1-score and exact match guided the evaluation. In NER, token alignment and class-weighted loss functions were explored to address class imbalances.

Evaluation Framework

Evaluation metrics were selected to align with task-specific objectives:

- **Summarization**: Metrics such as ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum were used to quantify lexical overlap between generated summaries and reference texts.
- **Question Answering**: Performance was assessed using F1-score and exact match (EM), which measured the overlap and correctness of retrieved answers.
- **Information Extraction**: NER models were evaluated based on precision, recall, and F1-score, providing a balanced measure of both the accuracy and completeness of

extracted entities.

Optimization and Efficiency Techniques
To overcome computational constraints, several optimization strategies were implemented:

- Freezing model layers, such as the encoder layers in PEGASUS, to reduce training time without significant performance loss.
- Utilizing subsets of datasets for initial experiments to identify optimal configurations before scaling up.
- Employing distributed training methods and gradient accumulation to achieve effective larger batch sizes.

This structured approach ensured a rigorous yet efficient evaluation of LLMs across the chosen NLP tasks, providing actionable insights into their strengths, limitations, and potential for domain-specific applications.

Results

Summarization

The summarization task demonstrated varying model performances across domains. PEGASUS excelled in news summarization, achieving the highest ROUGE scores, while DistilBART performed well in scientific summarization.

Domain	**Model**	**Dataset**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**ROUGE-Lsum**
News	PEGASUS	XSum	46.56	24.83	39.33	39.36
Scientific	DistilBART	PubMed	31.42	11.63	19.27	27.77
Conversational	BART-large	SAMSum	40.10	19.60	30.70	30.70

Question Answering

Question answering models exhibited diverse performance levels. BERT achieved moderate success on general-purpose datasets, while domain-specific datasets posed challenges.

Model	Dataset	Exact Match (EM)	F1-Score
BERT	SQuAD v2	50.0	55.0
ALBERT	FinancialBank	0.0	2.9
DistilBERT	MedMCQA	0.0	24.0

Information Extraction

The NER task revealed that BioBERT outperformed other models, underscoring the value of domain-specific pre-training for biomedical text.

Model	Dataset	Precision	Recall	F1-Score
BioBERT	CoNLL-2003	48.4%	16.3%	24.4%
BERT-base	CoNLL-2003	50.5%	20.4%	29.1%
DistilBERT	CoNLL-2003	46.8%	15.5%	23.3%

Discussion

The findings from this project reveal both the power and the limitations of large language models in specialized applications. Summarization tasks, particularly abstractive ones, benefited greatly from models like PEGASUS, which achieved superior ROUGE scores on news datasets. However, these models exhibited variability when applied to more complex or niche datasets like PubMed for scientific summarization, reflecting the challenges in generalizing across domains.

In question answering, general-purpose models like BERT demonstrated reasonable performance on standard datasets such as SQuAD v2 but struggled with financial and medical datasets. This underscores the necessity for domain-specific fine-tuning, as models trained on general text corpora lack the nuanced understanding required for specialized fields. Models like BioBERT, designed for biomedical texts, illustrated this need by outperforming general-purpose models in extracting information from CoNLL-2003 datasets. However, the limited availability of high-quality labeled data and computational resources emerged as persistent bottlenecks in realizing the full potential of these models.

Optimization techniques played a crucial role in this project. Gradient checkpointing and mixed-precision training allowed the fine-tuning of large models on constrained computational resources, highlighting the importance of efficient training methodologies. Despite these advancements, the challenges of dataset alignment, computational overhead, and domain-specific model limitations call for further innovation. The role of transfer learning and ensemble approaches remains a promising avenue for future exploration.
The results underscore that no single model excels universally across all tasks and domains. Instead, leveraging the strengths of domain-specific models alongside task-specific optimizations provides a pathway for achieving high performance. This emphasizes the growing need for tailored solutions in NLP, reflecting the unique demands of various industries.

Conclusions

This project illustrates the transformative potential of large language models in advancing NLP applications. By benchmarking LLMs across summarization, question answering, and information extraction tasks, it becomes evident that domain-specific fine-tuning is pivotal in achieving high performance. Models like PEGASUS excelled in abstractive summarization for news datasets, while BioBERT demonstrated its utility in biomedical named entity recognition, showcasing the importance of specialized pre-training.

Optimization strategies, including gradient checkpointing and learning rate schedulers, significantly enhanced computational efficiency, enabling the training of large models on limited resources. These techniques ensure that cutting-edge models are accessible even with computational constraints, broadening their applicability in real-world scenarios.

While the findings highlight the impressive capabilities of LLMs, they also point to persistent challenges, such as generalization across domains and reliance on large labeled datasets. The importance of task-specific configurations and tailored training cannot be overstated, as evidenced by the variability in model performance across domains.

The study reaffirms that leveraging the right combination of models, datasets, and optimization strategies is essential for meeting the diverse demands of NLP applications. By bridging the gap between general-purpose and domain-specific models, this project lays the groundwork for further advancements in language understanding, benefiting industries and research fields alike. Future work should explore the integration of ensemble techniques, extended fine-tuning on diverse datasets, and advanced transfer learning methods to unlock the full potential of LLMs.

Future Work

1. **Extended Fine-Tuning:** Incorporate larger and more diverse datasets for enhanced model generalization.
2. **Ensemble Techniques:** Combine multiple models to leverage their individual strengths.
3. **Transfer Learning:** Explore pre-training on massive text datasets followed by fine-tuning for specialized domains.

References

1. Lewis, M., et al., "BART: Denoising Sequenceto-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," ACL, 2020.
2. Raffel, C., et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR, 2020.
3. Devlin, J., et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.