CS 681 - Deep Learning for NLP
Project Proposal

Muhammad Adnan Rizqullah (2403851)

Supervised by Dr. Fawaz Al-Salmi

Literature Review

DistilBERT - A Compressed and Efficient Language Model

Introduction

The advancement of large language models has led to significant concerns regarding computational efficiency and environmental impact. This literature review examines DistilBERT, a notable contribution to addressing these challenges through model compression techniques.

Challenges with Large Language Models

Recent trends toward increasingly large language models have raised several concerns. First is the environmental cost of exponentially scaling these models' computational requirements, as noted by Schwartz et al. (2019) and Strubell et al. (2019). Second, while running these models on-device in real-time could enable novel language processing applications, their growing computational and memory demands may limit widespread adoption.

DistilBERT: A Distilled BERT Model

Sanh et al. (2020) introduced DistilBERT, a distilled version of BERT that addresses these efficiency challenges. DistilBERT is characterized as "smaller, faster, cheaper and lighter" than its predecessor, offering significant advantages:

1. 40% reduction in model size compared to the original BERT
2. 60% faster inference speed
3. Retention of 97% of BERT's language understanding capabilities

Methodological Approach

Unlike previous work that focused on task-specific model distillation, DistilBERT leverages knowledge distillation during the pre-training phase to create a general-purpose language representation model. This approach allows the resulting model to be fine-tuned for various downstream tasks while maintaining strong performance.

The authors validated their approach through proof-of-concept experiments and a comparative on-device study, demonstrating DistilBERT's capabilities for resource-constrained environments and real-time applications.

Limitations

1. Pre-training distillation focus: The research primarily applies knowledge distillation during the pre-training phase, while noting that other compression techniques (pruning, quantization) remain complementary and could be applied alongside their method.

2. Performance trade-off: Despite its reduced size, DistilBERT preserves 97% of BERT's language understanding capabilities, representing a modest performance sacrifice for significant computational efficiency gains.

3. Limited generalization evidence: The study demonstrates successful distillation specifically for BERT, but does not empirically verify whether the same approach would yield comparable results when applied to other transformer architectures.

Conclusion

DistilBERT represents a significant advancement in creating efficient language models through knowledge distillation. By substantially reducing computational requirements while preserving most of the original model's capabilities, it addresses critical concerns about the environmental impact and accessibility of large language models. The work demonstrates that effective general-purpose language models can be successfully created through distillation techniques, opening avenues for on-device deployment and more sustainable NLP applications.

Textbooks Are All You Need - The Impact of High-Quality Data on Language Models

Introduction

Research in large language models (LLMs) has traditionally focused on scaling laws, where performance improvements correlate with increases in model size and computational resources. However, a new direction explores how data quality, rather than quantity, can yield significant efficiency and performance gains.

Data Quality as a Scaling Dimension

As highlighted in "Textbooks Are All You Need" (2023), there has been a consistent pattern in the development of language models where "performance improves somewhat predictably as one scales up either the amount of compute or the size of the network" (Hestness et al., 2017), a phenomenon known as scaling laws (Kaplan et al., 2020). Building on the work of Eldan and Li (2023), this research investigates improvements obtainable through a different dimension: data quality.

High-Quality Data for Model Efficiency

The authors demonstrated that using "textbook quality" data enabled the training of a model that surpasses almost all open-source models on coding benchmarks such as HumanEval and MBPP. This achievement is particularly notable given that their model is 10x smaller in size and trained on a dataset 100x smaller than typical approaches.

This efficiency gain addresses a significant concern in the field: the environmental impact of training large language models. As noted by Bender et al. (2021), smaller models requiring less training can substantially reduce the environmental cost associated with developing and deploying LLMs.

The Phi-1 Model

The paper introduces phi-1, a coding-focused model that achieves remarkable results through the use of high-quality training data. The authors hypothesize that "high quality data dramatically improves the learning efficiency of language models for code as they provide clear, self-contained, instructive, and balanced examples of coding concepts and skills."

Limitations

Despite its impressive performance, phi-1 has several limitations compared to larger models:

1. It is specialized in Python coding, limiting its versatility compared to multi-language models
2. It lacks domain-specific knowledge for programming with specific APIs or using less common packages
3. Due to the structured nature of the training datasets and limited diversity in language and style, phi-1 is less robust to stylistic variations or errors in prompts
4. Performance degrades substantially when prompts contain grammatical mistakes

Conclusion

"Textbooks Are All You Need" demonstrates that high-quality, carefully curated training data can lead to significant performance improvements in language models while dramatically reducing computational requirements. This approach not only produces more efficient models but also addresses environmental concerns associated with training large language models.

Tasks

1. Sentiment analysis
    a. Tasks: Will fine tune a transformer pre-trained model for sentiment analysis
    b. Data source: Stanford Sentiment Treebank https://huggingface.co/datasets/stanfordnlp/sst2
    c. Pretrained model: DistilBERT https://huggingface.co/distilbert/distilbert-base-uncased
    d. Baselines:
        i. DistilBERT without fine tuning
        ii. A machine learning model baseline
    e. Evaluation metrics:
        i. Accuracy
        ii. Recall

iii. Precision
iv. F1

2. Creative text generation
   a. Tasks: Will fine tune a transformer pre-trained model for text summarization
   b. Data source: DialogSum https://huggingface.co/datasets/neil-code/dialogsum-test
   c. Pretrained model: Microsoft’s phi-2 1b
   d. Baselines:
      i. phi-2 without fine tuning zero-shot
      ii. phi-2 without fine tuning with 1-shot
   e. Evaluation metrics:
      i. Rouge
      ii. BLeU
      iii. Human evaluation on 10 random generation
3. Efficiency challenges
   a. Tasks: Will use quantization to reduce the size of model on fine tuning for text summarization
   b. Data source: DialogSum https://huggingface.co/datasets/neil-code/dialogsum-test
   c. Pretrained model: Microsoft’s phi-2 1b
   d. Baseline:
      i. Phi-2 1b without quantization
      ii. Phi-2 1b with lower quality quantization
   e. Evaluation:
      i. Model size
      ii. Fine-tuning training time
      iii. Evaluation results