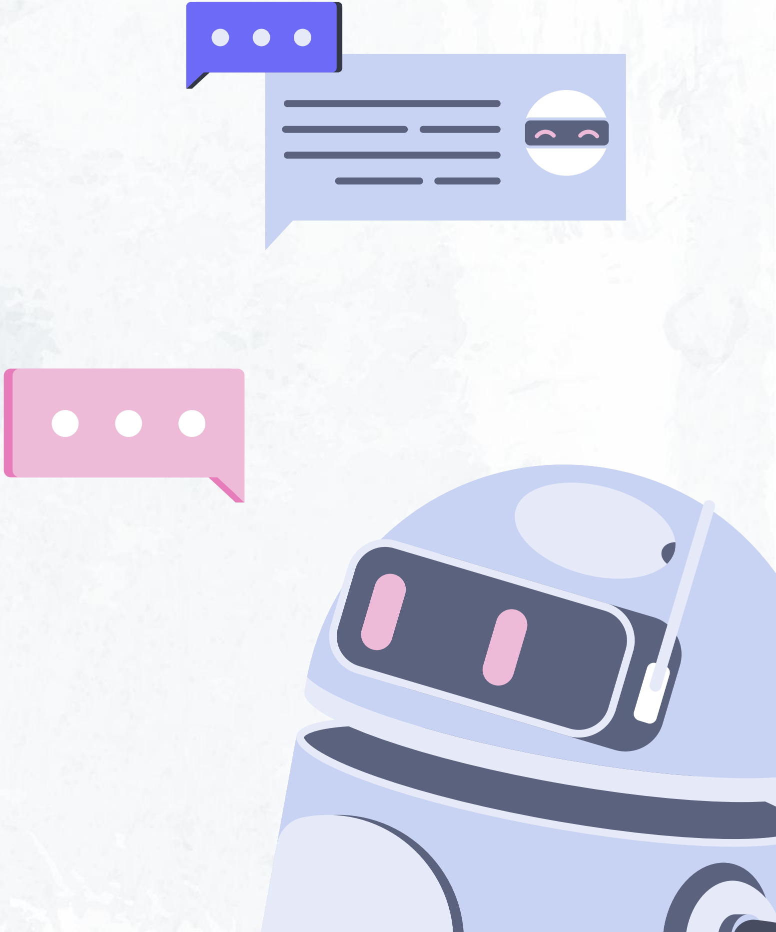
CS 681 - Deep Learning for NLP Final Report

Muhamamd Adnan R 2403851
Supervised By Dr. Fawaz Al-Salmi



FACULTY OF COMPUTING & INFORMATION TECHNOLOGY
KING ABDULAZIZ UNIVERSITY
FCIT
كلية الحاسبات وتقنية المعلومات
جامعة الملك عبدالعزيز

Research Question

How effectively can Microsoft's Phi-2 LLM be fine-tuned for abstractive text summarization?

How does DistilBERT's performance on sentiment analysis change after task-specific fine-tuning?

Introduction

Text Summarization: Extracting key information from text to produce a concise version while retaining core meaning.

Sentiment Analysis: Determining the emotional tone or opinion expressed in text (e.g., positive, negative, neutral).

Large Language Model: A neural network trained on massive text data to understand and generate human-like text.

Introduction

Distillation on Transformer-Based Model: Reducing the size of a transformer model by transferring knowledge to a smaller model while retaining performance.

Fine-Tuning LLM: Adapting a pre-trained large language model to specific tasks or domains by training it on a smaller, task-specific dataset.

Related Works

Textbooks Are All You Need:

Researchers developed phi-1, a relatively small 1.3B-parameter language model that achieves state-of-the-art performance on Python coding tasks by using high-quality "textbook-style" data, demonstrating that superior data quality can dramatically outperform traditional scaling approaches that rely on larger models and more compute.

Related Works

DistilBERT:

DistilBERT is a smaller general-purpose language model created through knowledge distillation during pre-training that reduces BERT's size by 40% while retaining 97% of its language understanding capabilities, running 60% faster, and enabling efficient on-device computations for edge environments with constrained resources.

Methodology (Text Summarization)

Model Overview: Microsoft's Phi-2 (2.7B parameters) optimized for efficiency while maintaining strong performance capabilities

Architecture Design: Decoder-only transformer architecture similar to GPT models but with optimizations for smaller computational footprints

Data Pipeline: Utilized CNN/DailyMail v3.0.0 dataset with 70/15/10/5 split for training, validation, development, and testing respectively. Used subset of dataset (210 train, 45 val, 30 dev, 15 test)

Technical Implementation: Implemented LoRA (Low-Rank Adaptation) with 4-bit quantization to enable processing on less powerful computing resources

Methodology (Text Summarization)

Evaluation Metrics: ROUGE, BLEU, Perplexity, and Human Evaluation

Baseline: Non-fine-tuned Phi-2 model (zero-shot summarization)

Notable Hyperparams:

- Epochs: 1
- Max steps: 10
- Optimizer: adam

Methodology (Sentiment Analysis)

Model Overview: DistilBERT - a compressed BERT variant retaining 97% language understanding while being 40% smaller and 60% faster

Architecture Design: Decoder-only transformer architecture similar to GPT models but with optimizations for smaller computational footprints

Data Pipeline: Stanford Sentiment Treebank (SST-2) with strategic 70/15/10/5 split for training, validation, development, and testing. Used subset of dataset (3500 train, 750 val, 500 dev, 250 test)

Technical Implementation: Leveraged Hugging Face Transformers for model fine-tuning while establishing SVM with TF-IDF vectorization as baseline

Methodology (Sentiment Analysis)

Evaluation Metrics: Accuracy, precision, recall, F1 score and Human Evaluation

Baseline: Non-fine-tuned DistilBERT (zero-shot classification) + SVM with TF-IDF features

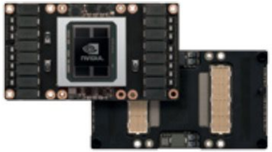
Notable Hyperparams:

- Epochs: 5
- Optimizer: adam

Methodology

Used Kaggle with GPU P100 Accelerator

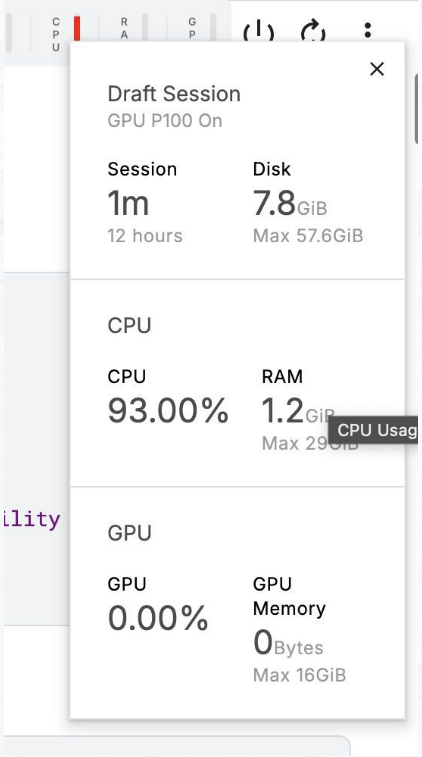
```
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                4
On-line CPU(s) list:   0-3
Thread(s) per core:    2
Core(s) per socket:    2
Socket(s):             1
NUMA node(s):          1
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 79
Model name:            Intel(R) Xeon(R) CPU @ 2.20GHz
Stepping:              0
CPU MHz:               2199.998
BogoMIPS:              4399.99
Hypervisor vendor:     KVM
Virtualization type:   full
L1d cache:             32K
L1i cache:             32K
L2 cache:              256K
L3 cache:              56320K
NUMA node0 CPU(s):     0-3
```



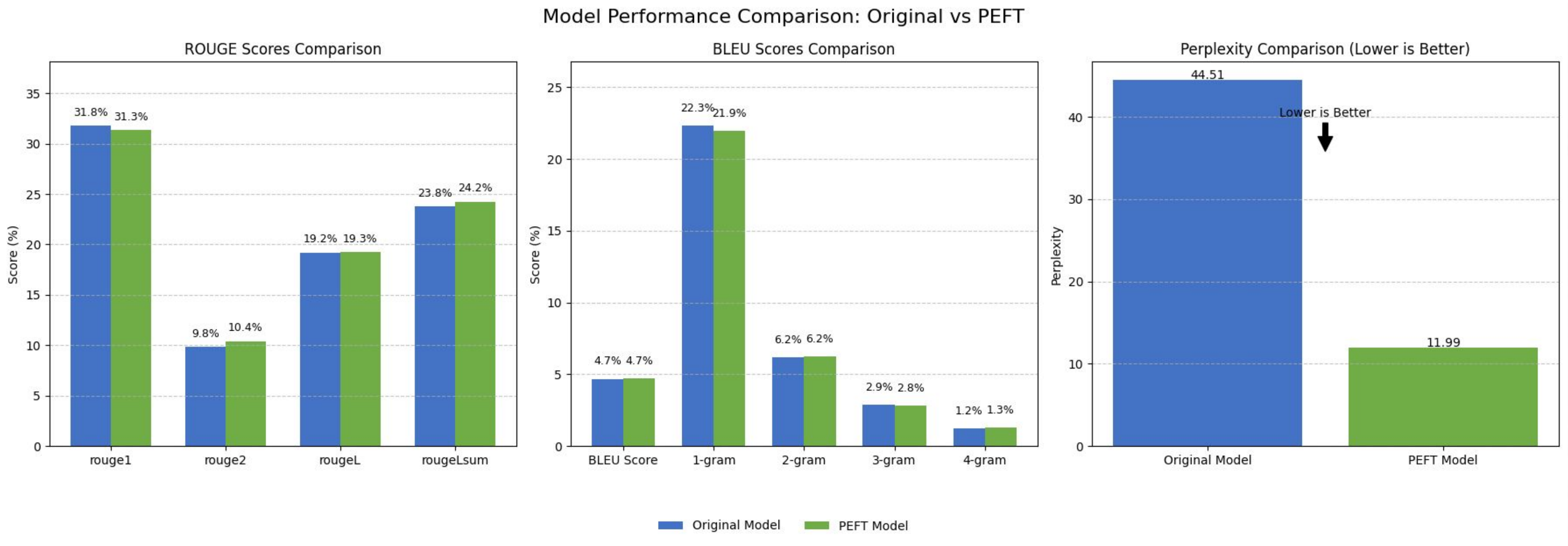
SPECIFICATIONS

GPU Architecture	**NVIDIA Pascal**
NVIDIA CUDA® Cores	**3584**
Double-Precision Performance	**5.3 TeraFLOPS**
Single-Precision Performance	**10.6 TeraFLOPS**
Half-Precision Performance	**21.2 TeraFLOPS**
GPU Memory	**16 GB CoWoS HBM2**
Memory Bandwidth	**732 GB/s**
Interconnect	**NVIDIA NVLink**
Max Power Consumption	**300 W**
ECC	**Native support with no capacity or performance overhead**
Thermal Solution	**Passive**
Form Factor	**SXM2**
Compute APIs	**NVIDIA CUDA, DirectCompute, OpenCL™, OpenACC**

TeraFLOPS measurements with NVIDIA GPU Boost™ technology



Results (Text Summarization)



Results (Text Summarization)

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Perplexity
Baseline Phi-2	0.32	0.10	0.19	0.05	44.51
Fine-tuned Phi-2	0.31	0.10	0.19	0.05	11.99

ROUGE Scores: Baseline model shows slightly better performance than the fine-tuned model (31.80% vs 31.32% for ROUGE-1), indicating minimal differences in recall-oriented metrics.

BLEU Score: Both models demonstrate nearly identical lexical precision with scores of 4.67% (baseline) and 4.68% (fine-tuned).

Perplexity: Fine-tuned model significantly outperforms the baseline (11.99 vs 44.51),

Results (Text Summarization)

Example	Human	Baseline	Fine tuned
Archaeological Site	Archaeologists believe site was centre of a network of long-distance trade in metals.	The article describes how archaeologists have discovered a prehistoric site that was once at the center of a network of long-distance trade.	The article discusses the latest discovery of a prehistoric site that was once at the center of a network of long-distance trade.
Jay-Z/Rihanna	Jay-Z reported to be 'deeply disappointed' by Rihanna's reunion with Chris Brown.	Rihanna was pictured out in London today looking happy and carefree, amid reports that her mentor Jay-Z is 'deeply disappointed' by her reunion with Chris Brown.	Rihanna was pictured out in London today looking happy and carefree, amid reports that her mentor Jay-Z is 'deeply disappointed' by her reunion with Chris Brown.

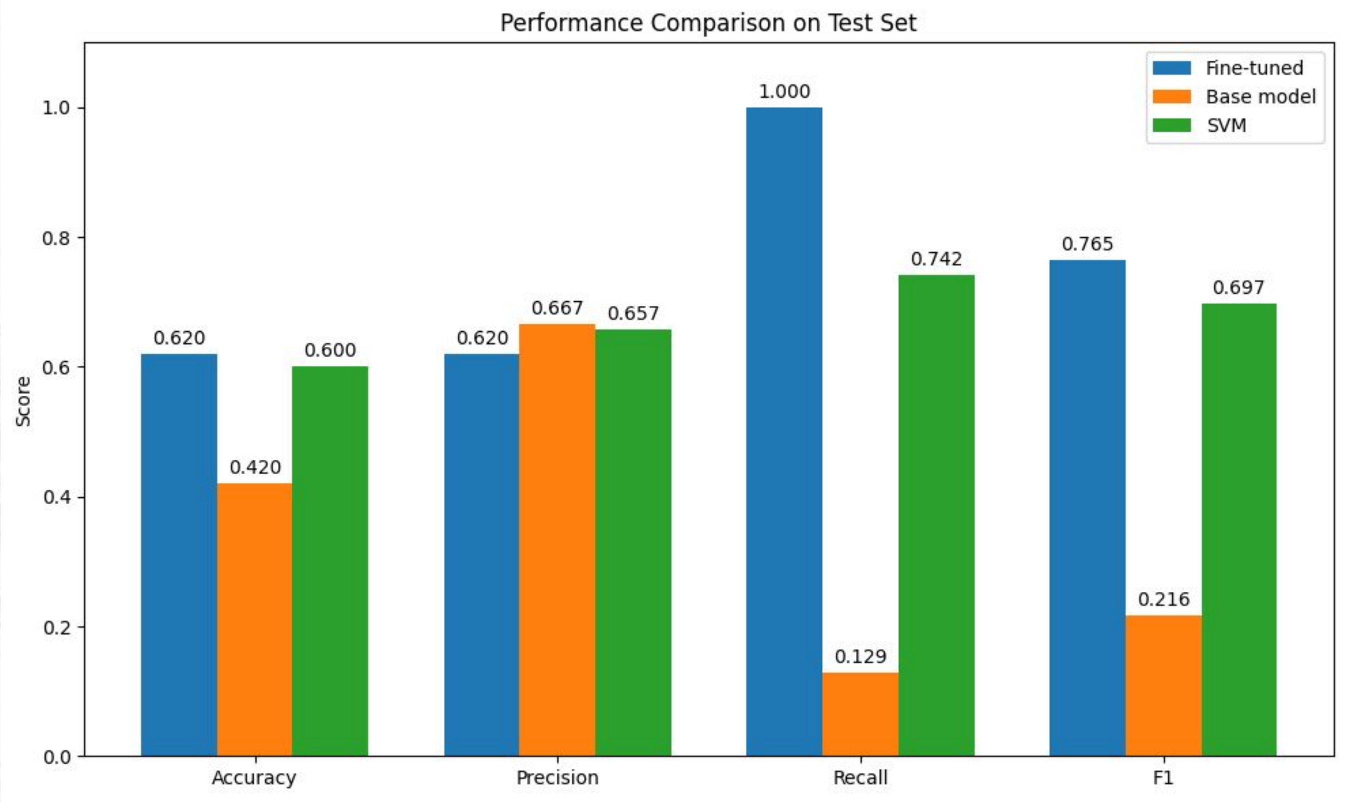
Results (Text Summarization)

Similarity Between Model Outputs: The original and fine-tuned model summaries show remarkable similarity in many cases

Stylistic Differences: Human summaries tend to be more concise and often use telegram-style writing (omitting articles), while both model versions generate more verbose, complete sentences.

Fluency: The fine-tuned model maintains the strong natural language fluency of the original model

Results (Sentiment Analysis)



Results (Sentiment Analysis)

Model	Accuracy	Precision	Recall	F1
SVM with TF-IDF	0.60	0.66	0.74	0.70
Baseline DistilBERT	0.42	0.67	0.13	0.22
Fine-tuned DistilBERT	0.62	0.62	1.00	0.77

Fine-tuned DistilBERT demonstrates the best overall performance with the highest F1 score (0.77) and perfect recall (1.00), meaning it identified all positive instances.

Baseline DistilBERT performs significantly worse, with particularly poor recall (0.13), suggesting it fails to identify most positive instances despite reasonable precision.

SVM with TF-IDF shows competitive performance, with more balanced precision and recall than either DistilBERT model.

Results (Sentiment Analysis)

Text	Ground Truth	Fine tuned	Baseline	SVM
A masterpiece of modern cinema with stunning visuals	Positive	Positive	Negative	Positive
The director failed to engage the audience	Negative	Positive	Negative	Negative
Not the best film I've seen, but still enjoyable	Positive	Positive	Negative	Negative
A complete waste of time and money	Negative	Positive	Negative	Positive

Results (Sentiment Analysis)

Model agreement: Fine-tuned/Base agreement: 4/24 (16.7%), Fine-tuned/SVM agreement: 14/24 (58.3%), Base/SVM agreement: 12/24 (50.0%)

Model accuracy on test examples: Fine-tuned model: 15/24 (62.5%), Base model: 7/24 (29.2%), SVM model: 13/24 (54.2%)

Model Prediction Patterns: Fine-tuned DistilBERT shows perfect recall but lower precision indicating to its positive prediction bias, the Baseline DistilBERT demonstrates very low recall indicating negative bias, while the SVM with TF-IDF exhibits the most balanced performance with more evenly distributed predictions across classes.

Conclusion & Future Works

Key Findings

Fine-tuning provides substantial performance improvements over zero-shot approaches

Even smaller models (Phi-2, DistilBERT) can achieve strong results when properly adapted

Conclusion & Future Works

Limitations

Our experiments were limited by computational resources, using smaller training sets than might be optimal

We focused on English-language datasets, limiting linguistic diversity

The summarization evaluation could be enhanced with human assessments of quality

Conclusion & Future Works

Future Works

More extensive training

Cross-lingual adaptation

Use more models

Conclusion & Future Works

Future Works

More extensive training

Cross-lingual adaptation

Use more models

Thanks!

Any questions?