CS 613: NLP Assignment 3 Report

Fine Tuning & Evaluation of Pre-trained Models

Course Code:	CS 613
Assignment:	Assignment 3
Submission Date:	November 20, 2024

Indian Institute of Technology Gandhinagar

Contents

1 Introduction

This report presents the fine-tuning and evaluation of the pre-trained model **meta-llama/Llama-3.2-1B** for two NLP tasks: Classification on the SST-2 dataset and Question-Answering on the SQuAD dataset. The tasks were performed on Kaggle Notebooks, and the results include both zero-shot and fine-tuned performance scores. For ease of computation, we have frozen the parameters of all layers and added an appropriate trainable layer at the end of the model.

2 Model Details

2.1 Pre-trained Model Selection

Selected model: **meta-llama/Llama-3.2-1B**.
The Meta Llama 3.2-1B is a compact, transformer-based language model with 1 billion parameters, designed for efficient and versatile natural language processing.

2.2 Parameter Calculation

The calculated parameters of the model are as follows:

	SST-2	**SQuAD**
Total Parameters (base Model)	1,235,814,400	1,235,814,400
Total Parameters (after adding one layer)	1,235,818,496	1,235,818,498
Trainable Parameters	4,096	4,098
Freezed Parameters (all except last layer)	1,235,814,400	1,235,814,400
Reported Parameters (from paper)	1.23B	

Table 1: Parameter Counts for SST-2 and SQuAD Using Base Llama Model

Please refer Appendix 1 for more details

3 Fine-Tuning Details

3.1 Task 1: Classification (SST-2)

- Dataset: SST-2
- Train-test split: 80:20 (seed=1)
- Metrics: Accuracy, Precision, Recall, F1

Fine-tuning process and hyperparameters:

- Learning rate: 2 x 10^{-5}
- Batch size: 16
- Epochs: 3

3.2 Task 2: Question-Answering (SQuAD)

- Dataset: SQuAD
- Train-test split: 80:20 (seed=1)
- Metrics: squad_v2, F1, BLEU, ROUGE, exact-match

Fine-tuning process and hyperparameters:

- Learning rate: 5 x 10^{-5}
- Batch size: 32
- Epochs: 2

4 Results and Analysis

4.1 Results

Refer Appendix 2 and Appendix 3 for model results

4.2 Performance Scores Before and After Fine-tuning

Table 2: Performance Metrics for Classification (SST-2)

Metric	Zero-Shot	Fine-Tuned
Accuracy	0.473	0.8451
Precision	0.4369	0.8503
Recall	0.1171	0.8445
F1	0.1847	0.8474

Table 3: Performance Metrics for Question-Answering (SQuAD)

Metric	Zero-Shot	Fine-Tuned
F1	75.0	2.748
BLEU	0.00416	0.00419
ROUGE-1	0.02885	0.02889
ROUGE-2	0.00930	0.00938
ROUGE-L	0.02859	0.02873
Exact-Match	0.0	7.2

Ideally the F1 score should be less then 1, but the values here suggest otherwise. This are the values the code gave. Refer to the github links provided for details, we have used code provided in the official PyTorch documentation.

4.3 Analysis of Results

- We freezed the model's parameters and only trained the final layer which is same as a training a NN for classification. We leverage the pre-trained model's general language understanding and fine tuned it to the specific requirements of binary sentiment classification.

- For SQuAD, there is not much imporvement, this is because question answering is a difficult task, and just training the last layer is not enough. The model captures biases from the original train dataset (of the Llama-3.2-1B Model). However, due to time and computational constraints, we could only train the last layer.

Data and Fine-Tuning Issues

The minimal improvements in BLEU, ROUGE, and Exact Match suggest that the fine-tuning process was suboptimal. Possible reasons include:

- Insufficient or low-quality training data.
- A mismatch between the pre-trained model's capabilities and the target task.

Model's Poor Task Adaptation

The consistently low scores across all metrics indicate the model might not be well-suited for the task without significant adjustments. This might require:

- Architecture modifications tailored to the task.
- More comprehensive or task-specific training data.

5 Model Parameters After Fine-Tuning

We add an extra layer for specific tasks (Classification and Question Answering). We freeze all the other layers and just train the last (fine-tuning) layers.

	SST-2	SQuAD
Parameters before Fine Tuning	1235814400	1235814400
Additional Parameters	4096	4096
Additional Biases	0	2
Parameters after Fine Tuning	1235818496	1235818498
Freezed-Layes Parameters	1235814400	1235814400
Trainable Parameters	4096	4098

After adding the relavant last layer, fine tuning the model does not change the number of parameters.

6 Model Upload to Hugging Face

The fine-tuned model is available on Hugging Face - SST2 Link and Hugging Face - SQuAD Link.

The code notebooks are available on Github - SST2 Link and Github - SQuAD Link.

7 Conclusion

Part A

Classification Task (Accuracy, Precision, Recall, F1)

Higher scores indicate better performance Expectations: Higher scores for fine-tuned models compared to pre-trained (zero-shot) models Rationale: Fine-tuned models are specifically trained on the task-specific dataset, learning relevant features and patterns, leading to improved predictions. Pre-trained models lack task-specific adaptation, so their performance is generally lower in zero-shot settings.

Question Answering Task (SQuAD v2, F1, BLEU, ROUGE, Exact Match)

Higher scores indicate better performance Expectations: Higher scores for fine-tuned models compared to pre-trained (zero-shot) models Rationale: Metrics like BLEU, ROUGE, and METEOR directly measure textual overlap and semantic similarity. These values improve after fine-tuning since the model has learned to predict answers closer to the reference. Exact Match (EM) and F1 improve as fine-tuning helps the model align predictions more closely with ground truth.

Part B

The number of parameters remains the same for both pre-trained and fine-tuned models because fine-tuning does not alter the model's architecture (e.g., number of layers, hidden units, or attention heads). Instead, it just updates the weights of final layers to adapt the model to the specific task.

Part C

Fine-tuned models trained on task-specific datasets outperform zero-shot models because they are additionally trained for the specific task. The zero-shot model, limited by its lack of task-specific optimization, demonstrates poor performance on task specific datasets as compared to finetuned models but displays strength of the pre-trained representations for general tasks.

Work Distribution

- **Aditya Mehta (Roll Number: 22110017)**
  - Developed the code for SQuAD, with the notebook and trained model available in the links above.
  - Computed various matrices for SQuAD.
  - Compiled and wrote the entire report.
- **Daksh Jain (Roll Number: 22110066)**
  - Developed the code for SQuAD, with the notebook and trained model available in the links above.
  - Conducted research on the Llama-3.2-1B model and calculated its parameters.
- **Hrriday Ruparel (Roll Number: 22110099)**
  - Fine-tuned the SQuAD model.
  - Analyzed the results for Question 7.
  - Provided debugging assistance and support as needed.
- **Kishan Ved (Roll Number: 22110122)**
  - Developed the code for SST-2, with the notebook and trained model available on Hugging Face (link above).
  - Computed various matrices for both SST-2 and SQuAD.
  - Wrote scripts to upload the model to Hugging Face.
- **Summet Sawale (Roll Number: 22110234)**
  - Developed the code for SST-2, with the notebook and trained model available on Hugging Face (link above).
  - Computed various matrices for SST-2.

Appendix 1

Add all references cited in the report.

Parameter details for Llama 3.2 1B Base Model

Modules	Parameters
model.embed$_{t}$*okens.weight*	262668288
model.layers.0.self$_{a}$*ttn.q$_{p}$roj.weight*	4194304
model.layers.0.self$_{a}$*ttn.k$_{p}$roj.weight*	1048576
model.layers.0.self$_{a}$*ttn.v$_{p}$roj.weight*	1048576
model.layers.0.self$_{a}$*ttn.o$_{p}$roj.weight*	4194304
model.layers.0.mlp.gate$_{p}$*roj.weight*	16777216
model.layers.0.mlp.up$_{p}$*roj.weight*	16777216
model.layers.0.mlp.down$_{p}$*roj.weight*	16777216
model.layers.0.input$_{l}$*ayernorm.weight*	2048
model.layers.0.post$_{a}$*ttention$_{l}$ayernorm.weight*	2048
model.layers.1.self$_{a}$*ttn.q$_{p}$roj.weight*	4194304
model.layers.1.self$_{a}$*ttn.k$_{p}$roj.weight*	1048576
model.layers.1.self$_{a}$*ttn.v$_{p}$roj.weight*	1048576
model.layers.1.self$_{a}$*ttn.o$_{p}$roj.weight*	4194304
model.layers.1.mlp.gate$_{p}$*roj.weight*	16777216
model.layers.1.mlp.up$_{p}$*roj.weight*	16777216
model.layers.1.mlp.down$_{p}$*roj.weight*	16777216
model.layers.1.input$_{l}$*ayernorm.weight*	2048
model.layers.1.post$_{a}$*ttention$_{l}$ayernorm.weight*	2048
model.layers.2.self$_{a}$*ttn.q$_{p}$roj.weight*	4194304
model.layers.2.self$_{a}$*ttn.k$_{p}$roj.weight*	1048576
model.layers.2.self$_{a}$*ttn.v$_{p}$roj.weight*	1048576
model.layers.2.self$_{a}$*ttn.o$_{p}$roj.weight*	4194304
model.layers.2.mlp.gate$_{p}$*roj.weight*	16777216
model.layers.2.mlp.up$_{p}$*roj.weight*	16777216
model.layers.2.mlp.down$_{p}$*roj.weight*	16777216
model.layers.2.input$_{l}$*ayernorm.weight*	2048
model.layers.2.post$_{a}$*ttention$_{l}$ayernorm.weight*	2048
model.layers.3.self$_{a}$*ttn.q$_{p}$roj.weight*	4194304
model.layers.3.self$_{a}$*ttn.k$_{p}$roj.weight*	1048576
model.layers.3.self$_{a}$*ttn.v$_{p}$roj.weight*	1048576
model.layers.3.self$_{a}$*ttn.o$_{p}$roj.weight*	4194304
model.layers.3.mlp.gate$_{p}$*roj.weight*	16777216
model.layers.3.mlp.up$_{p}$*roj.weight*	16777216
model.layers.3.mlp.down$_{p}$*roj.weight*	16777216
model.layers.3.input$_{l}$*ayernorm.weight*	2048
model.layers.3.post$_{a}$*ttention$_{l}$ayernorm.weight*	2048
model.layers.4.self$_{a}$*ttn.q$_{p}$roj.weight*	4194304
model.layers.4.self$_{a}$*ttn.k$_{p}$roj.weight*	1048576
model.layers.4.self$_{a}$*ttn.v$_{p}$roj.weight*	1048576
model.layers.4.self$_{a}$*ttn.o$_{p}$roj.weight*	4194304

Modules	**Parameters**
model.layers.4.mlp.gate$_proj.weight$	16777216
model.layers.4.mlp.up$_proj.weight$	16777216
model.layers.4.mlp.down$_proj.weight$	16777216
model.layers.4.input$_layernorm.weight$	2048
model.layers.4.post$_attention_layernorm.weight$	2048
model.layers.5.self$_attn.q_proj.weight$	4194304
model.layers.5.self$_attn.k_proj.weight$	1048576
model.layers.5.self$_attn.v_proj.weight$	1048576
model.layers.5.self$_attn.o_proj.weight$	4194304
model.layers.5.mlp.gate$_proj.weight$	16777216
model.layers.5.mlp.up$_proj.weight$	16777216
model.layers.5.mlp.down$_proj.weight$	16777216
model.layers.5.input$_layernorm.weight$	2048
model.layers.5.post$_attention_layernorm.weight$	2048
model.layers.6.self$_attn.q_proj.weight$	4194304
model.layers.6.self$_attn.k_proj.weight$	1048576
model.layers.6.self$_attn.v_proj.weight$	1048576
model.layers.6.self$_attn.o_proj.weight$	4194304
model.layers.6.mlp.gate$_proj.weight$	16777216
model.layers.6.mlp.up$_proj.weight$	16777216
model.layers.6.mlp.down$_proj.weight$	16777216
model.layers.6.input$_layernorm.weight$	2048
model.layers.6.post$_attention_layernorm.weight$	2048
model.layers.7.self$_attn.q_proj.weight$	4194304
model.layers.7.self$_attn.k_proj.weight$	1048576
model.layers.7.self$_attn.v_proj.weight$	1048576
model.layers.7.self$_attn.o_proj.weight$	4194304
model.layers.7.mlp.gate$_proj.weight$	16777216
model.layers.7.mlp.up$_proj.weight$	16777216
model.layers.7.mlp.down$_proj.weight$	16777216
model.layers.7.input$_layernorm.weight$	2048
model.layers.7.post$_attention_layernorm.weight$	2048
model.layers.8.self$_attn.q_proj.weight$	4194304
model.layers.8.self$_attn.k_proj.weight$	1048576
model.layers.8.self$_attn.v_proj.weight$	1048576
model.layers.8.self$_attn.o_proj.weight$	4194304
model.layers.8.mlp.gate$_proj.weight$	16777216
model.layers.8.mlp.up$_proj.weight$	16777216
model.layers.8.mlp.down$_proj.weight$	16777216
model.layers.8.input$_layernorm.weight$	2048
model.layers.8.post$_attention_layernorm.weight$	2048
model.layers.9.self$_attn.q_proj.weight$	4194304
model.layers.9.self$_attn.k_proj.weight$	1048576
model.layers.9.self$_attn.v_proj.weight$	1048576
model.layers.9.self$_attn.o_proj.weight$	4194304

Modules	**Parameters**
model.layers.9.mlp.gate$_{p}roj.weight$	16777216
model.layers.9.mlp.up$_{p}roj.weight$	16777216
model.layers.9.mlp.down$_{p}roj.weight$	16777216
model.layers.9.input$_{l}ayernorm.weight$	2048
model.layers.9.post$_{a}ttention_{l}ayernorm.weight$	2048
model.layers.10.self$_{a}ttn.q_{p}roj.weight$	4194304
model.layers.10.self$_{a}ttn.k_{p}roj.weight$	1048576
model.layers.10.self$_{a}ttn.v_{p}roj.weight$	1048576
model.layers.10.self$_{a}ttn.o_{p}roj.weight$	4194304
model.layers.10.mlp.gate$_{p}roj.weight$	16777216
model.layers.10.mlp.up$_{p}roj.weight$	16777216
model.layers.10.mlp.down$_{p}roj.weight$	16777216
model.layers.10.input$_{l}ayernorm.weight$	2048
model.layers.10.post$_{a}ttention_{l}ayernorm.weight$	2048
model.layers.11.self$_{a}ttn.q_{p}roj.weight$	4194304
model.layers.11.self$_{a}ttn.k_{p}roj.weight$	1048576
model.layers.11.self$_{a}ttn.v_{p}roj.weight$	1048576
model.layers.11.self$_{a}ttn.o_{p}roj.weight$	4194304
model.layers.11.mlp.gate$_{p}roj.weight$	16777216
model.layers.11.mlp.up$_{p}roj.weight$	16777216
model.layers.11.mlp.down$_{p}roj.weight$	16777216
model.layers.11.input$_{l}ayernorm.weight$	2048
model.layers.11.post$_{a}ttention_{l}ayernorm.weight$	2048
model.layers.12.self$_{a}ttn.q_{p}roj.weight$	4194304
model.layers.12.self$_{a}ttn.k_{p}roj.weight$	1048576
model.layers.12.self$_{a}ttn.v_{p}roj.weight$	1048576
model.layers.12.self$_{a}ttn.o_{p}roj.weight$	4194304
model.layers.12.mlp.gate$_{p}roj.weight$	16777216
model.layers.12.mlp.up$_{p}roj.weight$	16777216
model.layers.12.mlp.down$_{p}roj.weight$	16777216
model.layers.12.input$_{l}ayernorm.weight$	2048
model.layers.12.post$_{a}ttention_{l}ayernorm.weight$	2048
model.layers.13.self$_{a}ttn.q_{p}roj.weight$	4194304
model.layers.13.self$_{a}ttn.k_{p}roj.weight$	1048576
model.layers.13.self$_{a}ttn.v_{p}roj.weight$	1048576
model.layers.13.self$_{a}ttn.o_{p}roj.weight$	4194304
model.layers.13.mlp.gate$_{p}roj.weight$	16777216
model.layers.13.mlp.up$_{p}roj.weight$	16777216
model.layers.13.mlp.down$_{p}roj.weight$	16777216
model.layers.13.input$_{l}ayernorm.weight$	2048
model.layers.13.post$_{a}ttention_{l}ayernorm.weight$	2048
model.layers.14.self$_{a}ttn.q_{p}roj.weight$	4194304
model.layers.14.self$_{a}ttn.k_{p}roj.weight$	1048576
model.layers.14.self$_{a}ttn.v_{p}roj.weight$	1048576
model.layers.14.self$_{a}ttn.o_{p}roj.weight$	4194304

Modules	Parameters
model.layers.14.mlp.gate$_p$*roj.weight*	16777216
model.layers.14.mlp.up$_p$*roj.weight*	16777216
model.layers.14.mlp.down$_p$*roj.weight*	16777216
model.layers.14.input$_l$*ayernorm.weight*	2048
model.layers.14.post$_a$*ttention$_l$ayernorm.weight*	2048
model.layers.15.self$_a$*ttn.q$_p$roj.weight*	4194304
model.layers.15.self$_a$*ttn.k$_p$roj.weight*	1048576
model.layers.15.self$_a$*ttn.v$_p$roj.weight*	1048576
model.layers.15.self$_a$*ttn.o$_p$roj.weight*	4194304
model.layers.15.mlp.gate$_p$*roj.weight*	16777216
model.layers.15.mlp.up$_p$*roj.weight*	16777216
model.layers.15.mlp.down$_p$*roj.weight*	16777216
model.layers.15.input$_l$*ayernorm.weight*	2048
model.layers.15.post$_a$*ttention$_l$ayernorm.weight*	2048
model.norm.weight	2048

All the above parameters were frozen while fine tuning, to reduce computation.

Trainable parameter details for fine tuned model for sentiment analysis

Modules	Parameters
score.weight	4096

Trainable parameter details for fine tuned model for question answering

Modules	Parameters
score.weight	4096
score.bias	2

Appendix 2

Sentiment analysis: Model Predictions and Actual Sentiments

The following are examples of model predictions compared to the actual sentiments for various sentences:

- **Sentence 1:** "it 's a charming and often affecting journey ."
  - **Predicted:** Positive (1)
  - **Actual:** Positive (1)
- **Sentence 2:** "unflinchingly bleak and desperate"
  - **Predicted:** Positive (1)
  - **Actual:** Negative (0)
- **Sentence 3:** "allows us to hope that nolan is poised to embark a major career as a commercial yet inventive filmmaker ."
  - **Predicted:** Positive (1)
  - **Actual:** Positive (1)
- **Sentence 4:** "the acting , costumes , music , cinematography and sound are all astounding given the production 's austere locales ."
  - **Predicted:** Positive (1)
  - **Actual:** Positive (1)
- **Sentence 5:** "it 's slow – very , very slow ."
  - **Predicted:** Negative (0)
  - **Actual:** Negative (0)
- **Sentence 6:** "although laced with humor and a few fanciful touches , the film is a refreshingly serious look at young women ."
  - **Predicted:** Positive (1)
  - **Actual:** Positive (1)
- **Sentence 7:** "a sometimes tedious film ."
  - **Predicted:** Negative (0)
  - **Actual:** Negative (0)
- **Sentence 8:** "or doing last year 's taxes with your ex-wife ."
  - **Predicted:** Negative (0)
  - **Actual:** Negative (0)

Appendix 3

Question Answering Task: Predicted vs Actual

The following shows the predicted and actual results for the first ten samples in the Question Answering task:

- **Sample 1:**
  - **Predicted:** ””
  - **Actual:** France
- **Sample 2:**
  - **Predicted:** 10
  - **Actual:** 10th and 11th centuries
- **Sample 3:**
  - **Predicted:** Normans
  - **Actual:** Denmark, Iceland and Norway
- **Sample 4:**
  - **Predicted:** The Normans
  - **Actual:** Rollo
- **Sample 5:**
  - **Predicted:** and 11th centuries gave
  - **Actual:** 10th century
- **Sample 6:**
  - **Predicted:** ””
  - **Actual:** ””
- **Sample 7:**
  - **Predicted:** Rollo
  - **Actual:** Rollo
- **Sample 8:**
  - **Predicted:** The
  - **Actual:** William the Conqueror