ANLP

Ex1

Shir Babian, id: 322314303

public Github repository path -

https://github.com/shir-babian/ANLP_ex1.git

Part 1: Open Questions

1. QA can be a powerful framework for evaluating intrinsic properties of **language understanding, such as comprehension, inference, and semantic interpretation**. Below are three QA datasets that focus on intrinsic language capabilities:

a. SQuAD (Stanford Question Answering Dataset)

SQuAD evaluates the model's ability to understand and extract relevant information from a given passage. Since the answers are spans from the text, it directly measures intrinsic comprehension and reasoning based on context.

b. BoolQ (Boolean Questions)

BoolQ consists of yes/no questions paired with short passages. The task requires models to determine the truth value of a question given the text, which assesses their ability to perform natural language inference—an intrinsic aspect of language understanding.

c. ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset)

ReCoRD challenges models to answer cloze-style questions by reasoning over a passage and applying commonsense knowledge. This tests deep reading comprehension and the ability to resolve references and understand implicit information, making it an intrinsic evaluation task.

2. a.

1. Self-Consistency

- ***Description:***
  This method involves sampling multiple reasoning paths from the model (using stochastic decoding) and selecting the most frequent final answer among them.
- ***Advantages:***
  It increases reliability by reducing the likelihood of spurious errors in a single decoding path, especially in complex reasoning tasks.
- ***Computational Bottlenecks:***
  It requires generating many complete outputs per question, leading to significantly higher computational costs.
- ***Parallelizable:***
  Yes – all sampled generations can be executed in parallel.

2. Verifier-Based Methods

- ***Description:***
  A separate verifier model is used to assess and rank candidate outputs generated by the main model, selecting the most likely correct answer.
- ***Advantages:***
  Allows for more accurate selection without modifying the base model, and can help filter out incorrect or nonsensical answers.
- ***Computational Bottlenecks:***
  The verifier adds an additional inference pass for each candidate, which can be expensive if many candidates are considered.

- ***Parallelizable:***
  Yes – verifier evaluations can be parallelized across candidates.

3. Chain-of-Thought (CoT) Prompting

- ***Description:***
  CoT prompting adds step-by-step reasoning to the prompt, guiding the model to follow a logical path to the answer.
- ***Advantages:***
  Enhances performance on reasoning-heavy tasks, such as arithmetic or multi-hop questions, by encouraging intermediate reasoning steps.
- ***Computational Bottlenecks:***
  Increases prompt and output lengths, leading to higher memory usage and latency per example.
- ***Parallelizable:***
  Yes – despite longer sequences, multiple queries can be processed simultaneously.

4. Planning & Self-Correction (e.g., O1)

- ***Description:***
  This method involves first generating a high-level plan, then executing it step by step. The model monitors its own reasoning and may revise incorrect steps.
- ***Advantages:***
  Leads to more structured and coherent outputs, and allows for error correction mid-process.
- ***Computational Bottlenecks:***
  Involves multiple stages (planning, generation, evaluation, correction), each requiring additional inference steps.

- ***Parallelizable:***

  Partially – the planning and correction steps are sequential, but execution and evaluation of steps can be parallelized.

*b. Given a complex scientific task that requires strong reasoning capabilities and access to a single GPU with large memory, I would choose **Self-Consistency** as the preferred inference-time scaling method.*

Self-Consistency significantly improves the reliability of the model's outputs by generating multiple reasoning paths and selecting the most frequent final answer. This is particularly valuable in scientific tasks where correctness and robustness are crucial. Since I have access to a high-memory GPU, I can take advantage of its capacity to generate multiple samples in parallel, which helps mitigate the added computational cost. Unlike methods that require sequential planning and correction, Self-Consistency can fully utilize the GPU through parallelized sampling, making it both effective and feasible under the given resource constraints.

Part 2: Programming Exercise

Hyperparameter Comparison and Test/Validation Results Analysis

I conducted three experiments to fine-tune a BERT model on the Microsoft Research Paraphrase Corpus (MRPC) dataset. Each experiment used different hyperparameters to evaluate their impact on model performance. The experiments varied in learning rates and number of epochs while maintaining a consistent batch size. Below is a summary of the configurations tested and the results obtained.

Results Summary

Experiment	Epochs	Learning Rate	Batch Size	Validation Accuracy	Test Accuracy	Training Loss
1	1	0.01	16	0.6838	0.6649	0.75716
2	3	4e-5	16	0.8603	**0.8441**	0.00502
3	5	5e-5	16	**0.8652**	0.8307	0.00305

Analysis of Results

Did the configuration that achieved the best validation accuracy also achieve the best test accuracy?

No, the configuration that achieved the best validation accuracy did not achieve the best test accuracy. Experiment 3 (5 epochs, learning rate 5e-5) achieved the highest validation accuracy at 86.52%, but Experiment 2 (3 epochs, learning rate 4e-5) achieved the highest test accuracy at 84.41%. This suggests a potential case of overfitting in Experiment 3, where the model performed better on the validation set but did not generalize as well to the unseen test data compared to Experiment 2.

The difference is particularly notable when comparing Experiment 1 (1 epoch, learning rate 0.01), which performed significantly worse on both validation (68.38%) and test sets (66.49%). This demonstrates that while additional training epochs can improve validation performance (as seen in the progression from Experiment 2 to 3), it doesn't necessarily translate to better generalization on unseen data. The results indicate that the moderate approach of Experiment 2, with 3 epochs and a carefully tuned learning rate of 4e-5, provides the optimal balance between fitting to the training data and generalizing to new examples.

Qualitative Analysis of BERT Model Performance on Paraphrase Detection

To conduct a thorough qualitative analysis comparing the best and worst performing model configurations, I implemented additional code specifically designed to identify and analyze cases where the best model succeeded but the worst model failed. Based on the final experimental results (epoch_num: 3, lr: 4e-05, batch_size: 16 **vs.** epoch_num: 1, lr: 0.01, batch_size: 16), the enhanced analysis identified 128 examples where the best model correctly classified pairs while the worst model misclassified them (You can found it in the file "complete_disagreement_examples.txt"). By examining these examples in detail, I was able to categorize them into distinct patterns that reveal the specific challenges faced by the lower-performing model.

Categorization of Challenging Examples

1. Sentence Length and Structural Complexity

Example #47 (Long, Complex Sentences):

- Sentence 1: "President Bush raised a record-breaking $49.5 million for his re-election campaign over the last three months, with contributions from 262,000 Americans, the president's campaign chairman said Tuesday."
- Sentence 2: "President Bush has raised $83.9 million since beginning his re-election campaign in May, and has $70 million of that left to spend, his campaign said Tuesday."
- True Label: 0 (Not a paraphrase)
- Best Model Prediction: 0 (Correct) with high confidence
- Worst Model Prediction: 1 (Incorrect) with moderate confidence

In this example, both sentences are long (30+ words) with multiple clauses about similar topics but with critically different information (different amounts of money over different time periods). The worst model with its high learning rate (0.01) and limited training (1 epoch) failed to distinguish these complex structures while the best model correctly identified them as non-paraphrases.

2. Subtle Semantic Differences

Example #16 (Critical Detail Differences):

- Sentence 1: "SARS has killed about 800 people and affected more than 8400 since being detected in China in November."
- Sentence 2: "SARS has killed about 800 people and sickened more than 8,400 worldwide, mostly in Asia."
- True Label: 0 (Not a paraphrase)
- Best Model Prediction: 0 (Correct) with high confidence
- Worst Model Prediction: 1 (Incorrect) with moderate confidence

These sentences contain subtle but important differences in details - the first specifies the disease was detected in China in November, while the second mentions "worldwide, mostly in Asia" instead. The worst model, with its very high learning rate and single training epoch, failed to capture this significant difference in geographical and temporal information, likely focusing instead on the high lexical overlap in the first part of both sentences.

3. Similar Named Entities with Different Information

Example #85 (Same Entities, Different Facts):

- Sentence 1: "GM, the world's largest automaker, has 115,000 active UAW workers and another 340,000 retirees and spouses."
- Sentence 2: "They cover more than 300,000 UAW workers and 500,000 retirees and spouses."
- True Label: 0 (Not a paraphrase)
- Best Model Prediction: 0 (Correct) with high confidence
- Worst Model Prediction: 1 (Incorrect) with moderate confidence

Both sentences mention UAW workers and retirees but present entirely different numbers and contexts. The worst model appears to have focused on the entity overlap rather than the meaningful differences in the facts presented. The much higher learning rate (0.01) likely caused the model to overfit to simple patterns like entity matching without developing nuanced understanding of context.

4. Temporal and Numerical Differences

Example #42 (Different Numerical Values):

- Sentence 1: "The Dow Jones industrial average <.DJI> added 28 points, or 0.27 percent, at 10,557, hitting its highest level in 21 months."
- Sentence 2: "The Dow Jones industrial average <.DJI> rose 49 points, or 0.47 percent, to 10,578."
- True Label: 0 (Not a paraphrase)
- Best Model Prediction: 0 (Correct) with high confidence
- Worst Model Prediction: 1 (Incorrect) with moderate confidence

This example shows different numerical values (28 vs. 49 points, 0.27% vs. 0.47%) that significantly change the factual content despite similar wording. The worst model, with its high learning rate (0.01) and single training epoch, consistently struggled with such numerical distinctions, suggesting it failed to develop adequate representations for comparing precise numerical information.

5. High Lexical Overlap with Semantic Divergence

Example #121 (Similar Structure, Different Meaning):

- Sentence 1: "On the stand Wednesday, she said she was referring only to the kissing."
- Sentence 2: "On the stand Wednesday, she testified that she was referring to the kissing before the alleged rape."
- True Label: 0 (Not a paraphrase)
- Best Model Prediction: 0 (Correct) with high confidence
- Worst Model Prediction: 1 (Incorrect) with moderate confidence

These sentences begin identically but end with critically different information - the second sentence adds "before the alleged rape," completely changing the meaning. The worst model appears to have been misled by the high lexical overlap. The high learning rate (0.01) and limited training (1 epoch) likely caused the model to develop an overreliance on surface-level features rather than deeper semantic understanding.

6. Context-Dependent Information

Example #10 (Context Changes Meaning):

- Sentence 1: "The driver, Eugene Rogers, helped to remove children from the bus, Wood said."
- Sentence 2: "At the accident scene, the driver was 'covered in blood' but helped to remove children, Wood said."

- True Label: 0 (Not a paraphrase)
- Best Model Prediction: 0 (Correct) with high confidence
- Worst Model Prediction: 1 (Incorrect) with moderate confidence

The second sentence adds crucial contextual information ("covered in blood" and "at the accident scene") that significantly alters the scenario being described. The worst model, with its exceptionally high learning rate (0.01) and limited training time, failed to develop the capacity to recognize that this additional context changes the meaning.