Report

ANLP
Assignment 3
Mihika Sanghi 2021113014

Theory Questions:

1. Concept of Soft Prompts: The introduction of soft prompts addresses several key limitations of discrete text prompts:
    - Continuous Optimization: Unlike discrete text prompts that are limited to actual words, soft prompts exist in a continuous embedding space, allowing for fine-grained optimization through gradient descent.
    - Task-Specific Learning: Soft prompts can learn optimal task-specific representations that might not be expressible using natural language tokens.
    - Memory Efficiency: They require storing only a small number of learned vectors rather than long text sequences.
    - Flexibility: Soft prompts can adapt to capture complex task requirements that might be difficult to express in natural language.
2. Scaling and Efficiency in Prompt Tuning: The efficiency of prompt tuning improves with model scale because:
    - Larger models have richer internal representations, making them better at utilizing learned prompt embeddings
    - The parameter efficiency becomes more significant as models grow (updating a few hundred parameters vs billions)
    - The relative cost of storing separate fine-tuned versions increases with model size, making prompt tuning more attractive
    - Larger models can better leverage the compressed task-specific information encoded in soft prompts
3. Understanding LoRA: Key principles of LoRA:
    - Weight updates during fine-tuning often have low intrinsic rank
    - Instead of updating full weight matrices, LoRA decomposes updates into products of smaller matrices ($\Delta W = BA$)
    - Original weights remain frozen, only low-rank update matrices are trained
    - Updates are scaled by a factor α to control their magnitude

    Improvements over traditional fine-tuning:

    - Drastically reduced memory requirements
    - Faster training times
    - Ability to switch between different adaptations efficiently
    - Comparable performance with significantly fewer parameters
4. Theoretical Implications of LoRA: The success of LoRA has several theoretical implications:

- Task adaptation primarily occurs in low-dimensional subspaces of the parameter space
- Most weight updates during fine-tuning are redundant or unnecessary
- There exists a trade-off between expressiveness (rank size) and generalization
- The effectiveness of low-rank updates suggests that the important dimensions for task adaptation are much fewer than the full parameter space
- This aligns with theoretical work on the "intrinsic dimension" of neural network optimization

The model's expressiveness under LoRA depends on:

- The chosen rank (r) of the update matrices
- The architecture layers where LoRA is applied
- The scaling factor α that controls update magnitude
- The initialization of the low-rank matrices