Advanced Techniques in NLP: Summarisation with In-Context Learning and LoRA (2024)

*University of Edinburgh*
*Pasquale Minervini*

**Tutorial 6: Implementing Text Summarisation Systems (Week 12)**

In our course, we explored the basics and some advanced topics in NLP, including abstractive and extractive text summarisation. In this tutorial, we delve into two cutting-edge methods: in-context learning for summarisation and parameter-efficient fine-tuning with LoRA (Low-Rank Adaptation). Both techniques represent the forefront of making NLP models more flexible and efficient.

**Question 1: In-Context Learning for Summarisation**

In-Context Learning allows models to perform tasks without explicit retraining or fine-tuning, using a prompt that includes examples of the task to guide the model’s predictions. This question focuses on applying in-context learning to summarisation using the CNN/Daily Mail (CNN/DM) or XSum datasets.

**Question 1:** In this task, you will leverage LLaMA, a decoder-only transformer-based language model, for in-context learning applied to summarisation. Implement a summarisation system using in-context learning with either the CNN/DM or XSum dataset. Outline the steps you would take to format the data, design your prompts, and evaluate the system’s performance. Consider the impact of the number and quality of in-context examples on the system’s output.

**Question 2: Parameter-Efficient Fine-Tuning with LoRA**

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning technique that adapts pre-trained models to specific tasks without modifying the original model weights. Instead, LoRA introduces trainable parameters that interact with the pre-trained weights through low-rank matrices. This approach significantly reduces the number of parameters needed to fine-tune large models on tasks like summarisation.

**Question 2:** Using the CNN/DM or XSum dataset, implement a summarization system by applying LoRA for fine-tuning a pre-trained decoder-only transformer-based language model, such as LLaMA. Outline the steps to integrate LoRA into the LLM architecture, focusing on the specific adaptations needed for a decoder-only transformer-based model. Describe the modifications necessary to integrate LoRA into the model architecture, the training process, and how you would assess the effectiveness of this approach compared to traditional fine-tuning methods.