PERSONAL REPORT

Wang Yuqi, [redacted]

1. Primary Role and Responsibilities

- I led the group project from system design to report writing.
- I was solely responsible for the Deep Learning component, which encompassed both the development of code and the generation of innovative ideas.
- I assisted Liu Chen in enhancing his statistical learning methods in terms of rigorous benchmarking and accurate visualization

2. Exploration of Alternative Methodologies

Bi-directional Attention:

Our team evaluated the use of bi-directional large language models, such as the Nomic-embed-text. However, results were subpar compared to autoregressive LLMs like Llama. We hypothesize that Llama's larger model capacity enabled better utilization of extensive world knowledge, facilitating superior generalization beyond the limited dataset of one million headlines.

Joint-Embeddings Architecture:

We also explored a joint-embedding architecture by eliminating the decoder component of the autoencoder. The initial approach involved training two encoders in a manner akin to Barlow Twins, focusing on embedding similarity. This method required predefined similarity metrics between data points, prompting a shift to a non-contrastive training strategy. Despite implementing variance-invariance-covariance regularization (VICReg), the model tended to collapse to an identity function without contrastive samples, leading to the eventual abandonment of this approach.

3. Challenges Encountered

The primary challenge was managing and processing large datasets exceeding 90GB, which imposed significant memory constraints. To address this, I extensively optimized the code to facilitate dynamic data loading from disk to GPU, thereby eliminating performance bottlenecks and enabling efficient experimentation.

3.1. Learnings from the Project

Theoretical Insights: I expanded my knowledge of representation learning through comprehensive literature reviews, gaining exposure to various novel methodologies.

Practical Skills: I acquired valuable engineering skills. For example, I was at first confused why validation loss was lower than training loss, and later learned that this was common when dropout is used.