Weekly Report – Week 1

Ahmedabad University

Project Topic:

Investigation of Attention Mechanism for Multimodal Transformers

Course Name: CSE623 Machine Learning Theory and Practice
Professor Name: Prof Mehul Raval
University: Ahmedabad University

Team Name: FrameTrackers

Team Members:

1. **Shlok Shelat** - AU2240025
2. **Shrey Salvi** - AU2240033
3. **Rushi Moliya** - AU2240020

Summary of Work Done This Week: This week, we mainly focused on understanding the fundamentals of attention mechanisms in multimodal transformers, particularly in the CLIP model for text-based person search.

Understanding Attention Mechanisms:

- We examined three attention variants including self-attention as well as cross-attention while sparse attention received special attention.
- The research focused on analyzing attention system implementation in the CLIP model.

Review of Reference Materials:

- Read the paper "Attention is All You Need" to grasp the Transformer architecture concept.
- Students studied tutorials about vision transformers plus multi-head attention.
- I reviewed the CLIP research paper together with additional multimodal models that exist in this field.

Analysis of Attention Efficiency:

- The study focused on analyzing the computational expenses affiliated with distinct attention approaches.
- An analysis was conducted to determine both memory usage and processing duration among the available mechanisms.

Challenges Identified:

- High computational cost of self-attention.
- Trade-offs between accuracy and efficiency in different attention mechanisms.
- The optimization of attention mechanisms requires attention because they consume excessive resources.

Plan for Next Week: For the upcoming week, we will mainly be focused on mid-semester preparation and will be busy with the same, but in the mean while continuing to expand our understanding of CLIP and attention mechanisms.

Deep Dive into CLIP Implementation:

- The CLIP Code Notebook demands work from users while they learn its structural elements.
- Begin by working with the Flickr dataset until you proceed to the PRS dataset.

Feature Extraction and Optimization:

- Examines how different variations of attention mechanisms influence the extraction process of features.
- Study several methods which enhance operational efficiency.

Preliminary Implementation:

- Start your CLIP implementation by conducting first experiments with various attention methods.
- Assess early results that measure the effectiveness of text-based person search.

Conclusion: We have worked together to increase the theoretical understanding of attention mechanisms and their role in multimodal transformers, all in order to increase the overall understanding of the project. Moving forward, we will apply this knowledge gain to optimize attention mechanisms for efficient text-based person search using CLIP and the look through the code provided.