Weekly Report (April 14,2025 - April 20,2025)

1st Xiaosong Yuan
AISIG
Shanghai Jiao Tong University
Shanghai, China
yuanxiaosong1999@gmail.com

***Abstract*—This weekly report summarizes my research progress from April 14 to April 20, 2025. The focus includes analyzing the DistServe paper about optimizing LLM serving through prefill-decoding disaggregation, studying fundamental neural network architectures (ResNet, Batch Normalization), and identifying current challenges in research methodology. Key findings reveal the trade-offs in distributed LLM serving systems and the importance of normalization techniques in deep learning. The report concludes with plans for strengthening the foundational knowledge and improving paper reading efficiency.**

***Index Terms*—LLM serving, distributed systems, neural networks, Batch Normalization, ResNet, research methodology**

I. Paper Reading

- **DistServe**: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving.
  - **Core Problem**: Improves TTFT and TPOT by separating prefill and decoding phases while maintaining SLO constraints to increase effective throughput.
  - **Key Innovations**:
    - Dedicated GPU groups: Prefill group (compute-intensive) uses tensor parallelism.
    - The Decoding group (I/O-intensive) uses pipeline parallelism.
    - NVLINK minimizes communication overhead between groups.
  - **Open Questions**: How to handle inter-group communication without expensive NVLINK? Alternative solutions may be needed.

  Detailed analysis is available on the iCloud forum.

II. Knowledge Acquisition

- **Key Concepts**:
  - **Batch Normalization**:
    - Addresses gradient explosion/vanishing similarly to weight decay.
    - Weight decay regularizes via L2-norm constraints.
    - BN standardizes layer distributions through learnable parameters (α, β).
    - Typically placed between linear and activation layers.
  - **ResNet**:
    - Enables ultra-deep networks (1000+ layers) through skip connections.
    - Additive operations prevent gradient vanishing in backward propagation.
    - Lower layers remain trainable despite small gradients in the upper layers.
- **Studied Architectures**: LeNet, AlexNet, VGG, NiN, GoogLeNet, ResNet.

III. Current Challenges and Next Steps

1) **Foundational Knowledge Enhancement**
   - Complete "Dive into Deep Learning" (by Mu Li) within the coming week.
   - Continue systematic study of "LLM BOOK" for large language model fundamentals.
   - Initiate parallel computing studies through:
     - Stanford CS149: Parallel Computing.
     - CSAPP (Computer Systems: A Programmer's Perspective).
2) **Paper Reading Methodology Improvement**
   - Strengthen literature accumulation through daily paper analysis.
   - Develop critical reading skills via:
     - Regular identification of paper contributions/limitations.
     - Comparative analysis with state-of-the-art works.
   - Maintain consistent practice (Practice makes perfect).
3) **Research Practice Transition**
   - Experimental replication:
     - Select 1-2 key papers for implementation.
     - Focus on reproducible components (e.g., prefill-decoding separation).
   - Academic writing preparation:
     - Draft methodology sections for potential publications.
     - Participate in vertical research projects for hands-on experience.