[Samsung PRISM] Preliminary Discussion

Deep Cache Implementation

Team

1. College Professor(s):
   1. Dr. G.K.Sandhia / sandhiag@srmist.edu.in
   2. Dr.D.Viji / vijid@srmist.edu.in
2. Students:
   1. Abhishek Soni / as8819@srmist.edu.in
   2. Arnav Agarwal / aa5579@srmist.edu.in
   3. Devanand K / dk7180@srmist.edu.in
   4. Kushagra Saxena / ks3780@srmist.edu.in
3. Department: CTECH

Date: 6th Feb 2024

DeepCache Implementation

SAMSUNG

Problem Statement

Context

DeepCache helps improve inference speed of Stable Diffusion model. It does this by skipping few steps in U-Net during inference. The results are faster with very little degradation in quality. We can have this worklet to compare this method as well as combining this with other methods to improve inference speed of SD.

Statement

Implement DeepCache on latest Stable Diffusion model.

Worklet Details

6 — Duration (Months)

4 — Members Count

Mentors

☎ ✉
Pranal Prasad Dongare
+91-7022250561
pranal.p@samsung.com

☎ ✉
Tushar Madaan
+91-9205301569
Tushar.m2@samsung.com

Pre-Requisite

- 2312.00858.pdf (arxiv.org)

Expectations

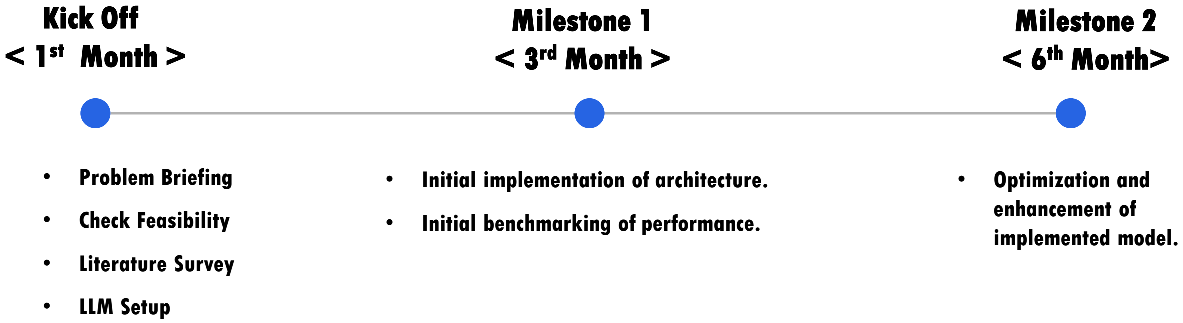
Undertaken Tasks

- Conduct Literature survey.
- Implement DeepCache.
- Identify whether LoRA technique post DeepCache implementation is feasible and helps in getting more inference time.

KPI

- DeepCache implementation
- Inference speed should have at least 25% improvement for similar prompt on vanilla Stable Diffusion model.
- LoRA implementation.

Timeline



Kick Off
< 1st Month >

- Problem Briefing
- Check Feasibility
- Literature Survey
- LLM Setup

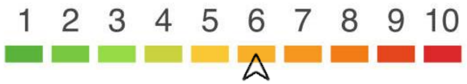
Milestone 1
< 3rd Month >

- Initial implementation of architecture.
- Initial benchmarking of performance.

Milestone 2
< 6th Month>

- Optimization and enhancement of implemented model.

Complexity



Literature survey and study

Literature Surveys

I. Stable Diffusion Model

Explored foundational papers on the Stable Diffusion (SD) model, including:

- "**Diffusion Models Beat GANs on Image Synthesis** " by Prafulla Dhariwal and Alex Nichol.
- "**Denoising Diffusion Probabilistic Models**" by Pieter Abbeel, Ajay Jain and Jonathan Ho.
- Studied the architecture, training process, and use cases of SD in image generation.
- Identified the importance of optimizing inference speed without compromising image quality.

II. DeepCache

Researched papers and resources related to the DeepCache technique, including:

- "**DeepCache: Accelerating Diffusion Models for Free** " Xinyin Ma Gongfan Fang Xinchao Wang.
- "**DeepCache: A Deep Learning Based Framework For Content Caching**" by Arvind Narayanan, Saurabh Verma, Eman Ramadan, Pariya Babaie, Zhi-Li Zhang.
- Understood DeepCache principles and its potential role in accelerating inference.
- Acknowledged considerations such as limitations and trade-offs associated with DeepCache.

III. LoRA Technique

Investigated the LoRA (Low Rank Adaptation) technique and its relevance to deep learning, referencing:

" **Lora: Low-rank Adaptation Of Large Language Models**" by Edward Hu, Yelong Shen ,Phillip Wallis ,Zeyuan Allen-Zhu ,Yuanzhi Li, Shean Wang, Lu Wang and Weizhu Chen

- Explored potential applications and benefits of LoRA, particularly in improving inference speed.

Literature survey and study

- **Major Observations / Conclusions:**

(provide details about your findings, experimental opinion – Use separate slide if necessary)

Observations :

I. Stable Diffusion (SD):

- Identified the potential of SD in capturing complex data distributions for high-quality image generation.
- Acknowledged the evolving landscape of real-world applications demanding enhanced efficiency in SD models.

II. DeepCache:

- Identified DeepCache as a promising technique for accelerating inference in complex models.
- Emphasized the need for novel strategies to address the unique challenges posed by complex generative models like SD.

III. LoRA Technique:

- Explored LoRA as an intriguing approach to refining feature representations, with promising implications for further optimizing SD models.
- Considered the integration of LoRA as an innovative step towards achieving superior inference speed without compromising on quality.

Conclusions :

I. Need for Optimization:

- Highlighted the pressing demand for breakthroughs in optimizing SD models, especially as they find applications in dynamic, real-time environments.

II. DeepCache Integration:

- Concluded that integrating DeepCache into SD represents a leap forward, offering a paradigm shift in accelerating inference while maintaining model robustness.

III. Feasibility of LoRA:

- Acknowledged the potential of LoRA as a novel tool for refining feature representations, presenting an exciting opportunity to synergize with DeepCache and amplify speed gains.

IV. Balancing Trade-offs:

- Emphasized the critical importance of finding a delicate balance between inference speed and image quality, envisioning a future where both are elevated simultaneously through innovative techniques.

Queries

- **Challenges** :
  (Discuss in the form of bullets, what are the next action steps, any road blocks / bottlenecks)

I. Integration Complexity:

- **Challenge:**
  - Integrating DeepCache into the existing U-Net architecture of the Stable Diffusion model may be complex, potentially leading to compatibility issues.
- **Next Action Steps:**
  - Conduct a thorough code review to understand the existing U-Net implementation.
  - Collaborate with experts or research groups specializing in both Stable Diffusion and DeepCache for insights.

II. Performance Trade-Offs:

- **Challenge**:
  - Achieving a balance between improved inference speed and maintaining high-quality image generation might pose challenges.
- **Next Action Steps**:
  - Experiment with different configurations and hyper parameters to optimize DeepCache's impact on speed without sacrificing too much in terms of image quality.
  - Utilize quantitative metrics for comprehensive evaluation.

III. LoRA Integration Challenges:

- **Challenge:**
  - Integrating LoRA post-DeepCache implementation might introduce complexities in feature alignment and model stability.
- **Next Action Steps:**
  - Explore and implement LoRA in a controlled environment before integrating it with the main project.
  - Collaborate with researchers familiar with LoRA for guidance.

Queries

- **Challenges :**
  (Discuss in the form of bullets, what are the next action steps, any road blocks / bottlenecks)

IV. Experimental Validations:

- **Challenge**:
  - Ensuring the reliability and reproducibility of experimental results may be challenging due to the complexity of deep learning models.
- **Next Action Steps**:
  - Implement rigorous experimental protocols and documentation to enhance reproducibility.
  - Consider peer reviews or collaborations to validate findings.

V. Implementation Complexities:

- **Challenge:**
  - The complexity of implementing DeepCache and LoRA may lead to coding challenges and errors.
- **Next Action Steps:**
  - Break down the implementation process into manageable tasks with clear documentation.
  - Foster collaboration within the team to address coding challenges effectively.

VI. Unanticipated Model Behaviour:

- **Challenge:**
  - The Stable Diffusion model may exhibit unexpected behavior during integration with DeepCache or LoRA.
- **Next Action Steps:**
  - Implement extensive model testing with various datasets to identify and address any unforeseen issues.
  - Collaborate with the research community for insights into model behavior.

Thank you