INSTITUT Louis Bachelier LABS

Newsletter Data & AI

29/11/2024

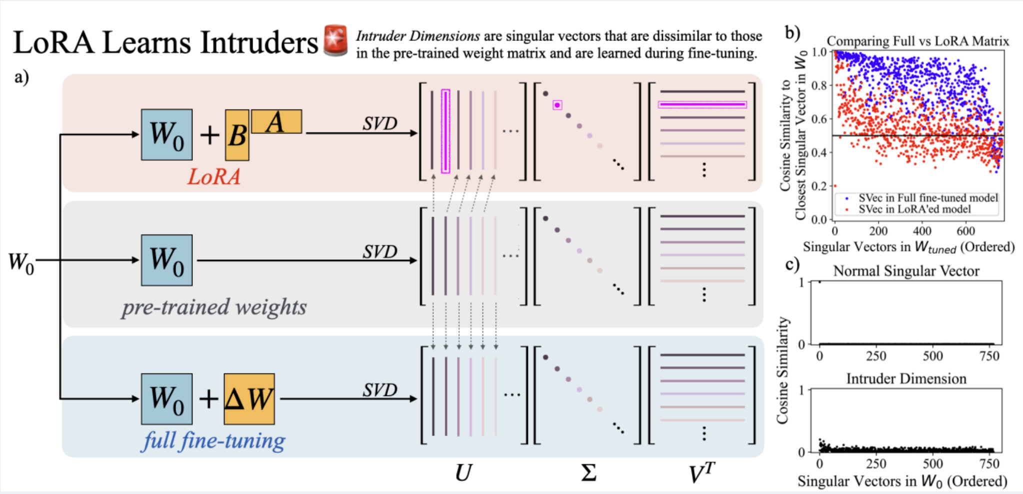
THE SHORT OF IT ⏱️

- **Streamlining Deep Learning Optimization :** Researchers found that stabilizing LoRA (Low-Rank Adaptation) fine-tuning enhances robustness and efficiency, while a schedule-free optimization method simplifies training by eliminating learning rate schedules and achieving state-of-the-art performance.
- **Improving Model Efficiency:** Mixture of Transformers reduces multi-modal pretraining costs by 55.8%, while SeerAttention achieves 90% sparsity and speeds up long-context tasks by 5.67×.

Trends

- [Paper] LoRA vs Full Fine-Tuning: An Illusion of Equivalence

  MIT researchers revealed that Low-Rank Adaptation (LoRA) fine-tuning, despite matching full fine-tuning in task-specific performance, introduces "intruder dimensions" in weight matrices that weaken robustness and generalization across tasks. Their analysis shows that higher-rank, stabilized LoRA models mitigate these issues, aligning more closely with the robustness of full fine-tuning while retaining efficiency.



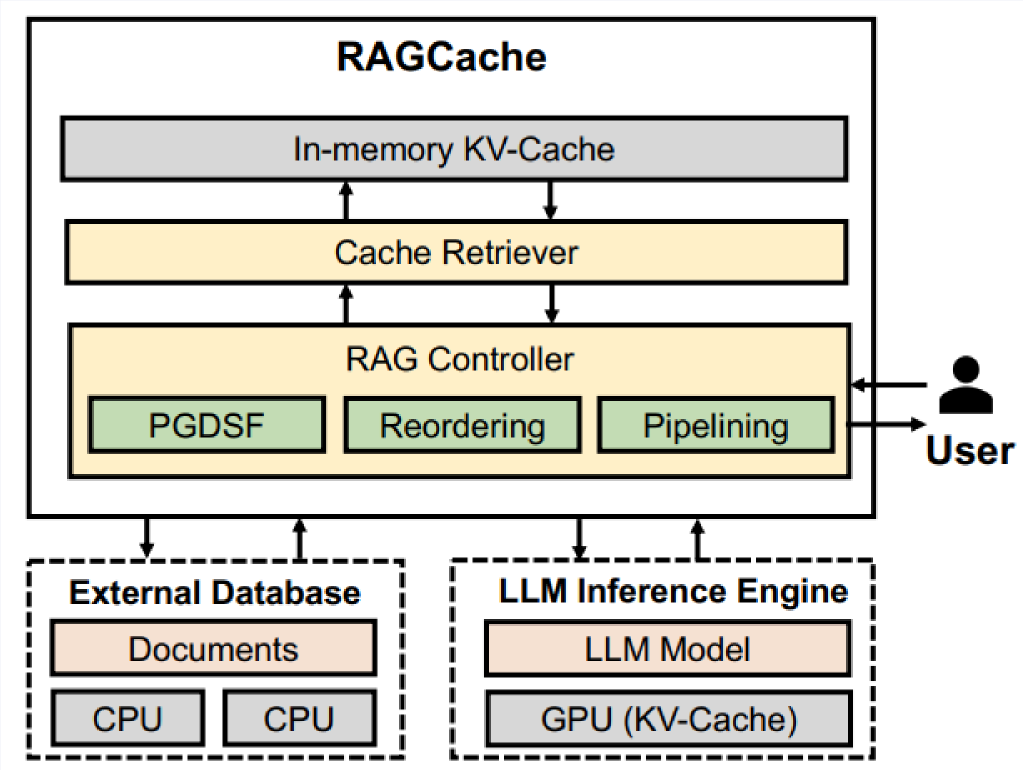
LoRA Learns Intruders 🚨 *Intruder Dimensions* are singular vectors that are dissimilar to those in the pre-trained weight matrix and are learned during fine-tuning.

- [Paper] The Road Less Scheduled

  Researchers introduced a *Schedule-Free optimization approach* that achieves state-of-the-art performance without relying on predefined learning rate schedules or stopping times. By unifying scheduling and iterate averaging, their method eliminates extra hyper-parameters, making it versatile across tasks from convex optimization to deep learning. This approach powered their winning entry in the MLCommons 2024 Algorithmic Efficiency Challenge.

- [Paper] RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation

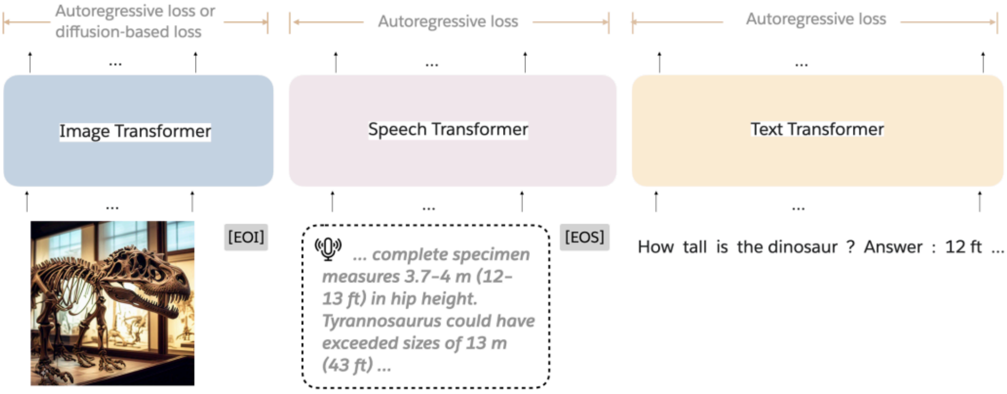
Researchers from ByteDance and Peking University introduced *RAGCache*, a dynamic caching system for Retrieval-Augmented Generation (RAG), to tackle the computational challenges of long sequence generation. By organizing retrieved knowledge into hierarchical caches and optimizing retrieval-inference overlap, RAGCache significantly reduces latency, achieving up to 4× faster token generation and 2.1× higher throughput compared to state-of-the-art systems.



State Of The Art

- [Paper] Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models

  Meta and Stanford researchers introduced *Mixture-of-Transformers (MoT)*, a sparse multi-modal transformer architecture that reduces pretraining costs for models processing text, images, and speech. By decoupling modality-specific parameters while maintaining global self-attention, MoT achieves dense model performance with significantly fewer FLOPs: 55.8% for text-image tasks and 37.2% for speech. In tasks like image generation, MoT delivers comparable or superior results in less wall-clock time, highlighting its efficiency and scalability for multi-modal LLMs.


Autoregressive loss or diffusion-based loss
Autoregressive loss
Autoregressive loss
Image Transformer
Speech Transformer
Text Transformer
[EOI]
... complete specimen measures 3.7–4 m (12–13 ft) in hip height. Tyrannosaurus could have exceeded sizes of 13 m (43 ft) ...
[EOS]
How tall is the dinosaur ? Answer : 12 ft ...

- [Paper] SEERATTENTION: Learning Intrinsic Sparse Attention in Your LLMs

  SeerAttention is a new attention mechanism for Large Language Models (LLMs) that tackles the inefficiency of quadratic complexity by dynamically learning attention sparsity. Instead of relying on predefined patterns, it uses a learnable gate to focus on significant blocks in the attention map, enhancing accuracy and speed. With a customized FlashAttention implementation, SeerAttention achieves a 90% sparsity ratio, a 5.67× speedup in long-context scenarios, and outperforms state-of-the-art sparse attention methods in adaptability and efficiency.

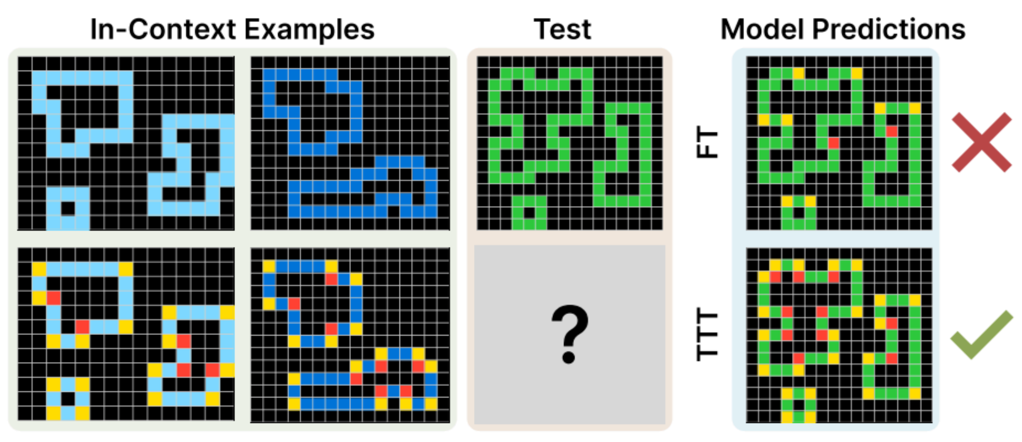
Miscellaneous

- [Blog] Evaluating feature steering: A case study in mitigating social biases

  Anthropic researchers examined *feature steering* in Claude 3 Sonnet to mitigate social biases while maintaining capabilities. They found a "sweet spot" where steering reduces biases (e.g., age, gender) without harming performance, though off-target effects and unpredictability remain challenges. Promisingly, features like "Neutrality and Impartiality" reduced biases across nine dimensions with minimal capability loss, showing potential for safer model outputs with further refinement.

- [Paper] The Surprising Effectiveness of Test-Time Training for Abstract Reasoning

  Test-time training (TTT) enhances language models' reasoning on novel tasks, achieving up to 6× accuracy improvement on the Abstraction and Reasoning Corpus (ARC). By finetuning on similar tasks, using auxiliary formats, and applying per-instance training, an 8B-parameter model reaches 53% ARC validation accuracy, a 25% state-of-the-art improvement for neural methods, and 61.9% when combined with program generation, matching human-level performance.



- [Paper] “Give me BF16 or Give Me Death”? Accuracy-Performance Trade-Offs in LLM Quantization

  This study evaluates FP8, INT8, and INT4 quantization formats on the Llama-3.1 model family across benchmarks and real-world tasks. FP8 is found to be lossless, INT8 incurs only 1-3% accuracy degradation with tuning, and INT4 performs competitively with 8-bit formats. Performance analysis highlights INT4's cost-efficiency for synchronous and mid-tier deployments, while FP8 and INT8 excel in large-scale asynchronous setups, offering clear guidelines for efficient LLM deployment.

- [Blog] Why are ML Compilers so Hard?

Pete Warden’s **“Why are ML Compilers so Hard?”** highlights the challenges of building efficient and scalable ML compilers, including the vast and evolving set of deep learning layers, complex computation graphs, and reliance on Python. He contrasts a "Matlab-like" future requiring manual optimization with an "LLVM-like" ecosystem featuring a universal intermediate representation for portability. Warden calls for training environments that balance researcher flexibility with compatibility to advance ML compilation.

- [Package] Outlines

  Outlines is a Python library that enables structured text generation with large language models (LLMs), ensuring outputs adhere to formats like JSON schemas, regex, or grammars. It supports multiple LLM integrations and efficient features like type constraints, caching, and batch inference. Designed for reliability and performance, Outlines is ideal for tasks requiring precise, predictable outputs.

EVENTS

- [Conference] NeurIPS 2024

  The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024) is scheduled to take place from Tuesday, December 10, to Sunday, December 15, 2024, at the Vancouver Convention Center. For more details and highlights, visit the NeurIPS 2024 site.

Thank you for your engagement. We eagerly anticipate sharing further advancements in AI with you.