Newsletter Data & AI

10/02/2025

THE SHORT OF IT ⏱️

- **R3GAN Revives GANs :** A new modernized GAN, R3GAN, simplifies training, outperforms StyleGAN2, and is fully open-source.
- **Self-Adaptive Transformers :** Transformer² enables LLMs to adapt to new tasks in real-time with minimal adjustments—code available.

Glossary 📖

- **GAN:** An AI model that generates images, videos or text.
- **Transformer:** A model that powers AI chatbots (like ChatGPT) and translation.

Trends

- [Paper] The GAN Is Dead; Long Live The GAN! A Modern Baseline GAN

  Researchers from Brown and Cornell University challenge the belief that GANs are inherently difficult to train by introducing a principled, regularized relativistic loss that ensures local convergence without ad-hoc tricks. Their R3GAN framework modernizes GAN architectures, simplifying training and outperforming StyleGAN2 while achieving competitive results against state-of-the-art GANs and diffusion models across multiple datasets.

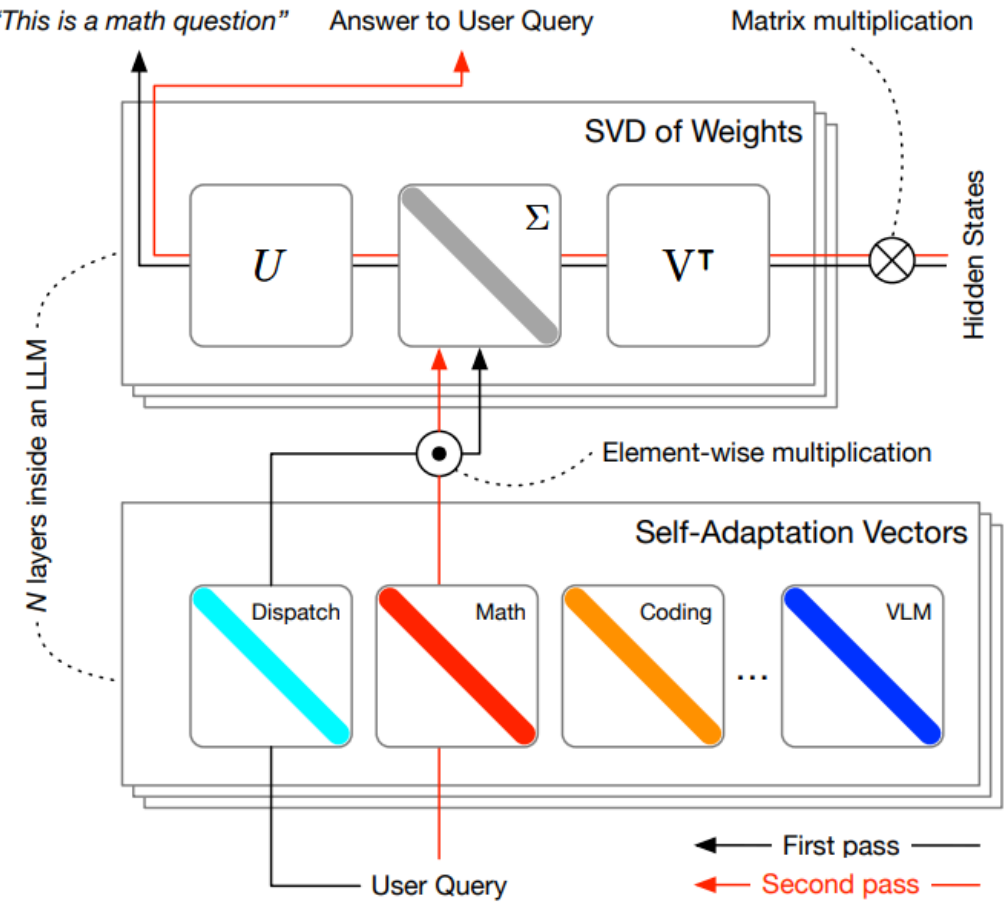
- [Paper] Entropy-Guided Attention for Private LLMs

  The paper introduces an information-theoretic framework to optimize decoder-only language models for private inference (PI), addressing the computational overhead imposed by nonlinear operations. Using Shannon entropy as a guiding measure, they uncover the dual role of nonlinearities in stabilizing training and preserving attention head diversity, proposing an entropy-guided attention mechanism and regularization technique to enhance PI-friendly transformer architectures.

State Of The Art

- [Paper] Transformers-Squared: Self-Adaptive LLMs

  Sakana AI and Tokyo Institute of Science researchers introduce *Transformer²*, a self-adaptive framework that enables LLMs to handle unseen tasks in real time by adjusting only key components of their weight matrices. With a two-pass inference mechanism and task-specific expert vectors trained via reinforcement learning, *Transformer²* outperforms LoRA in efficiency and scalability while demonstrating versatility across architectures and vision-language tasks, advancing toward self-organizing AI systems.



- [Paper] X-Sample Contrastive Loss: Improving Contrastive Learning with Sample Similarity Graphs

  Meta FAIR, New York University, and Brown University researchers introduce *X-Sample Contrastive*, a novel objective that refines contrastive learning by encoding nuanced relationships between samples rather than relying on a binary similarity graph. Tested across datasets from ImageNet-1k to CC12M, it surpasses both self-supervised and vision-language models, outperforming CLIP by up to 18% in low-data settings and improving object-background separation in learned representations.

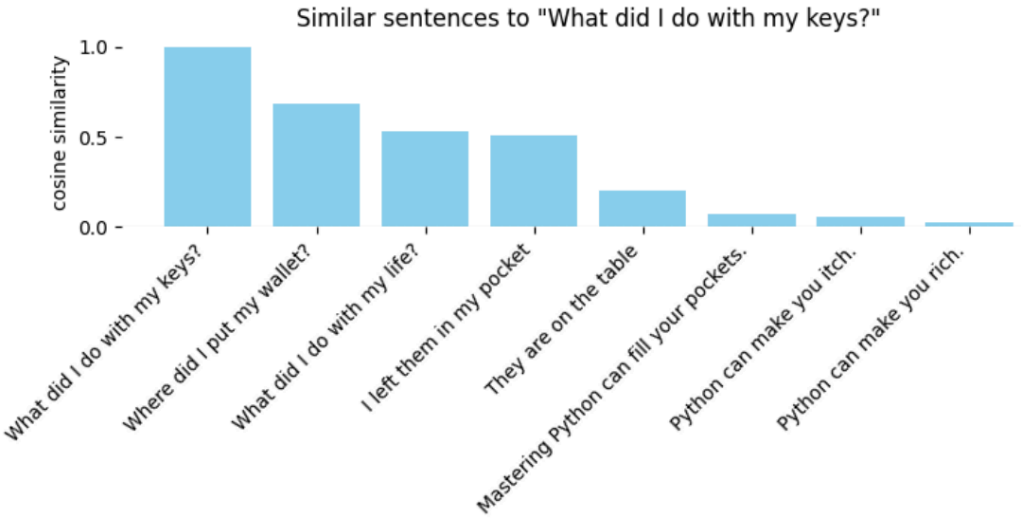
Miscellaneous

- [Blog] Batch Inference vs Online Inference

  The article explores the trade-offs between batch and online inference for deploying machine learning models, emphasizing that the choice depends on product needs and latency requirements. While batch inference simplifies infrastructure and suits non-urgent predictions like recommendation systems, online inference is essential for real-time applications such as ETA predictions in UberEats, requiring robust monitoring and system optimization to meet strict latency constraints.

- [Blog] Don't Use Cosine Similarity Carelessly

  Piotr Migdał warns against the careless use of cosine similarity, highlighting its tendency to capture surface-level patterns rather than true semantic meaning. While embeddings are powerful, blindly applying cosine similarity can lead to misleading matches, and he advocates for alternatives like task-specific embeddings, prompt engineering, and context-aware transformations to improve vector-based comparisons.



- [Blog] Data Have a Limited Shelf Life

  The blog challenges the idea that data remain valuable forever, showing how repeated use can obscure biases. Through cases like Quetelet's soldiers and von Bortkiewicz's cavalry deaths, it highlights how selection and context can render datasets "dead," raising concerns about bias and outdated data in science.

- [Package] Bespoke Curator

  Bespoke Curator is a Python library for efficient synthetic data curation, optimizing post-training and structured data extraction. It supports batch processing, structured outputs, and seamless integration with LiteLLM and vLLM for high-quality dataset generation.

Events

- [Conference] AAAI 2025

  The Thirty-Ninth Annual AAAI Conference on Artificial Intelligence (AAAI-25) is scheduled from February 25 to March 4, 2025, at the Pennsylvania Convention Center in Philadelphia, Pennsylvania.or more details and highlights, visit the AAAI-25 website.

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