CompSci Journal
Double Blind Peer Reviewed, Open Access International Research Journal.
ISSN: 3067-3089 | *Website: https://csjour.com* | *Volume-1, Issue-5* | *October - 2025*

Original Article

Recent Advances in Machine Learning Technologies

Shailaja S. Shelke
Computer Science & Engineering, NBNSCOE, Solapur

Manuscript ID:
CSJ-2025-010502

Volume 1

Issue 5

Pp. 6-8

October 2025

ISSN: 3067-3089

Submitted: 05 Sept. 2025

Revised: 15 Sept. 2025

Accepted: 02 Oct. 2025

Published: 31 Oct. 2025

Correspondence Address:
Shailaja S. Shelke
Computer Science & Engineering, NBNSCOE, Solapur
Email: shailajashelke5@gmail.com

Quick Response Code:

Web. https://csjour.com/

DOI: 10.5281/zenodo.17710709

DOI Link:
https://doi.org/10.5281/zenodo.17710709

Abstract

The last three years have experienced a massive consolidation and diversification of machine-learning research and deployment. Breakthroughs in language, vision and multimodal tasks continue to rely on foundation models and transformers, whereas efficiency, adaptively and retrieval have emerged as themes. The paper is a survey of the main technological advances (20232025): foundation and multimodal models, retrieval-augmented generation (RAG), parameter-efficient fine-tuning (PEFT) models such as LoRA variants, run-time/compute optimizations such as FlashAttention-2, and generative vision model advances (e.g., SDXL) or promptable segmentation (SAM). We talk about practical use, evaluation issues (robustness, alignment, hallucination), and research directions such as continual learning, causal representation learning and greener training. Simultaneously, image understanding and generation have been extended by vision technologies like DINOv2, the Segment Anything Model (SAM), and image generation models like high-fidelity diffusion models such as SDXL.

These innovations have facilitated their use in knowledge assistants, content creation, perception of robotics, and deployment of edges. Although things have been developing at a very fast pace, there are still difficulties related to the reduction in hallucination, assessment reliability, ecological sustainability, and congruence. The paper is a systematic review of some of the important technologies, system applications, constraints, and future research trends that define the present machine learning landscape.

Keywords: *Promt, PEFT, RAG, Machine Learning, Foundation Models, Multimodal Models, Retrieval-Augmented Generation (RAG), Parameter-Efficient Fine-Tuning (PEFT), Low-Rank Adaptation (LoRA)*

Introduction

Machine learning refers to an artificial intelligence (AI) technology based on algorithms that can be used to analyze and learn on the basis of data and make predictions or decisions without the need to write any explicit code. It allows systems to be able to perform more effectively as time progresses as they learn more information and has been used in various areas, such as image recognition, speech recognition, and recommendation systems. Machine learning (ML) is moving away from task-oriented networks to large, generalist models (foundation models) that are fine-tuned, downstream adapted, or augmented. At the same time, studies have worked on the cost and safety constraints of large models through the enhancement of: (1) inference/training efficiency, (2) modular retrieval and grounding to external knowledge, and (3) parameter-efficient adaptation. The integration of these threads allows real-life systems that are more precise, less expensive to operate and less likely to be covert by disastrous crashes like hallucinations.

General Overview of Key Technologies

The foundation and multimodal models provide a framework for researching the problem of motivation and its impact on the work environment.<|human|>The foundation and multimodal models offer an opportunity to study the issue of motivation and its influence on the work environment.

Large pretrained models (also known as foundation models) queriable on a variety of downstream problems are still at the center. Self-supervised methods such as **DINOv2** were used in vision and trained visual features that could transfer to other tasks; in segmentation, Meta's **Segment Anything Model (SAM) introduced promptable segmentation and was found to transfer well to a wide range of image tasks without training on any examples of those tasks. These are the examples of transition to large pretraining models and generalized interfaces to use downstream. [1].

How to cite this article:
Shelke, S. S. (2025). Recent Advances in Machine Learning Technologies. CompSci Journal, 1(5), 6–8. https://doi.org/10.5281/zenodo.17710709

Retrieval-Augmented Generation (RAG)

RAG systems are parametric (LLMs) combined with non-parametric retrieval over external corpora to achieve factuality, reduce hallucination, and also provide up-to-date answers. Recent surveys and system architecture formalize the taxonomy and strength provisions of the retriever-generator pipes and emphasize their increasing significance in production systems (to improve accuracy and traceability). [2])

Parameter-Efficient Fine-Tuning (PEFT)

Rather than performing complete fine-tuning, PEFT models such as LoRA and its recent extensions (e.g., VB-LoRA, B-LoRA, LA-LoRA) adjust or introduce a small number of additional trainable parameters to large models, which reduces memory and compute expenses by a significant margin but does not impact performance. Such techniques are currently the norm to reduce large models in resource-constrained environments. [3].

Optimization of Attention and Compute

The scale has been necessary in optimizing core transformer operations. FlashAttention-2 enhanced attention calculations, lessening non-matmul FLOPs and enhanced parallelism, with quite great acceleration and approaching hardware peak performance - faster training and reduced inference of big transformer designs.

Generative Vision Models and High-Fidelity Synthesis

Image generators based on diffusion developed quickly — **Stable Diffusion XL (SDXL) (and subsequent versions) demonstrate better composition, text clarity, and photorealism, and add model and pipeline features that allow high-quality image generation to be more accessible to applications of design, media, and content creation. [5].

Applications & Systems

In search engines, assistants and knowledge workers, the efficiency of search engines has been a primary focus of attention.<|human|>3.1 Search, Assistants and Knowledge Work. The effectiveness of search engines has been the focus of attention in search engines.

The RAG + LLM pipelines have become the key actors in enterprise assistants and knowledge tools: the grounding (citations, documents) is offered by retrieval, and the synthesis and explanation are provided by LLMs. This mix is found in research assistants, medical and legal summary (with very strict guardrails), and customer support robotics[2].

Video Systems and Content generation

Image editing pipelines, medical imaging preprocessing pipelines, automated annotation pipelines and robotics perception pipelines have been integrated with SAM and DINOv2. SDXL and other diffusion models drive creative processes (illustration, concept art), synthetic data generation, and photorealism. ([

Edge and On-Device ML

Attention/compute optimizations combined with PEFT and quantization methods enabled adapted foundation models to be able to run on significantly smaller hardware or at far reduced server cost, which is critical in privacy-sensitive or latency-sensitive applications.

Assessment, Safety and Limitations

Hallucination makes lying true, so that the speaker can speak truthfully but misleadingly; and the opposite is also true.<|human|>The truthfulness and the hallucination are one, so that the speaker may speak truthfully, but deceptively; and vice versa.

Big models are able to make false statements which are plausible. RAG minimizes the hallucinations by giving a traceable evidence base but the quality of retrieval and passage ranking are position of failure; the industry and academic efforts concentrate on measuring confidence and exposing provenance. ([arXiv][2])

The 4.2 Robustness and Distribution Shift is another principle that is less obvious but equally important.<|human|>The next principle that is not as evident but equally crucial is the 4.2 Robustness and Distribution Shift.

Robustness is assisted by self-supervised pretraining (e.g., DINOv2), large segmentation datasets (the billion-mask dataset of SAM), yet the models continue to fail on rare edge cases, distribution shifts, and underrepresented populations. Stress tests and work on benchmarks is still a vital concern. ([arXiv][1].

Environmental and Economic Costs

The foundation models require training which is energy-consuming. System and innovations (FlashAttention-2, PEFT) are reduced in cost, but not to the extent of the environmental footprint - an operating field of community focus.

Representative Technical Details.

1. Flash Attention-2 (Efficiency)

FlashAttention-2 rearranges and combines attention computations to reduce memory traffic and non-matrix-multiply operations; it parallelizes attention between thread blocks and warps to get a high occupancy on the GPU, and achieves significant speedups over naive implementations. This technology in engineering directly saves on transformer training and cost.

2. VB-LoRA and PEFT Progress

Recent PEFT studies consider extreme parameter reduction, including, but not limited to, vector banks (VB-LoRA) and adaptive layer-wise rank allocation, demonstrating comparable performance to orders-of-magnitude fewer trainable parameters - allowing most downstream tasks to be fine-tuned on commodity GPUs.

3. RAG Architectures

Contemporary RAG systems can differ in the centrality of the retriever or generator; hybrid and robustness-aware systems use retrieval verification, multi-hop retrieval and selective grounding of generated claims. Best practices that are synthesized in surveys include indexing, chunking, and filtering of context. [2].

Open Research Directions

1. Constant and Computer-based Learning:

Techniques that enable models to update their knowledge without re-training them completely, and without forgetting the previous knowledge disastrously, are essential to long-lived systems.

2. Causal and Interpretable Representation:

It would be better to go beyond correlation to causal structure to enhance robustness and transferability.

3. The main issue lies in the fact that the current state does not align with the vision of a safer future.<|human|>Alignment and Safety at Scale:

Scalable oversight techniques (automated auditing, understandable justification, human-in-the-loop alignment) are required.

4. Green ML:

To minimize carbon footprint of large models, hardware-aware algorithms, sparse models and more efficient training primitives will be necessary.

5. Real World Reliability Benchmarking:

More realistic, multimodal, and adversarial tests to evaluate hallucination, robustness and fairness.

Conclusion

Since foundation models and self-supervised vision backbones to retrieval-grounded generation and efficiency engineering, recent ML studies (20232025) are providing increasingly competent models and also practical tools to deploy them responsibly. The ability to balance capability, cost and safety is actively balancing in the field with RAG, PEFT and system level optimizations finding the practical way forward between research milestones and the final product.

Acknowledgment

I express my sincere gratitude to all those who contributed to the successful completion of this paper. I am deeply thankful to the management and administration of NBNSCOE, Solapur, for providing the necessary facilities and a supportive academic environment.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Survey: *Retrieval-Augmented Generation: A Comprehensive Survey.* (2025).
2. Dao, T. et al., *FlashAttention-2: Faster Attention with Better Parallelism.* (2023).
3. Li, Y. et al., *Extreme Parameter Efficient Fine-Tuning with Vector Banks (VB-LoRA).* NeurIPS 2024. ([NeurIPS Proceedings])
4. Kirillov, A. et al., *Segment Anything (SAM).* arXiv: 2304.02643 (2023).
5. Oquab, M. et al., *DINOv2: Learning Robust Visual Features Without Labels.* (2023).
6. Stable Diffusion XL (SDXL) model pages and documentation.
7. Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2023). FlashAttention-2: Faster attention with better parallelism.arXiv preprint arXiv: 2307.08691.
8. Li, Y., Chen, W., Li, S., & Wang, Z. (2024). Extreme Parameter-Efficient Fine-Tuning with Vector Banks (VB-LoRA). Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS).
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., … Girshick, R. (2023). Segment Anything. arXiv preprint arXiv:2304.02643.