Musik in bayern
ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*352*

Real-Time Deep Learning Systems: Challenges in Deploying DL Models for High-Speed Applications

[1] Dr.M.Charles Arockiaraj
Associate Professor
Dept.of MCA

[2] Dr. R. Amutha
Associate Professor
Dept.of ISE

Dr.M.Umadevi
Assistant Professor
Dept.of Comp.Sci

[1,2] AMC Engineering College, Bangalore
[3]Sri Adi Chunchanagiri Women's College,Cumbum.

To Cite this Article

Dr. M. Charles Arockiaraj, Dr. R. Amutha, Dr. M. Umadevi,**" Real-Time Deep Learning Systems: Challenges in Deploying DL Models for High-Speed Applications"** ***Musik In Bayern, Vol. 89, Issue 11, Nov 2024, pp38-47***

Article Info

Received: 27-10-2024 Revised: 08-11-2024 Accepted: 16-11-2024 Published: 26-11-2024

Abstract

Real-time applications of deep learning (DL) have become pivotal in industries such as autonomous vehicles, financial services, and industrial automation. However, the integration of DL models into high-speed environments poses significant challenges due to computational complexity, latency constraints, and scalability requirements. This paper explores the key obstacles in deploying DL models for real-time applications and discusses potential solutions, emphasizing the interplay of hardware, software, and system design. Deep learning (DL) has revolutionized industries ranging from healthcare to finance by enabling unprecedented levels of automation, accuracy, and insight extraction. However, deploying these systems in real-time applications, such as autonomous vehicles, financial trading, and industrial automation, presents unique challenges. This article explores the technical and practical hurdles of deploying deep learning systems in high-speed environments.

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*352*
Keywords: Data quality, model complexity, and interpretability, deep learning, software deployment, Stack Over-ow.

1. Introduction

Deep learning has transformed various domains by enabling powerful predictive and analytical capabilities. However, its adoption in real-time systems introduces stringent requirements that conflict with the resource-intensive nature of DL models. Real-time systems, by definition, must process and respond to inputs within tight deadlines, often measured in milliseconds. Deploying DL models in such settings requires overcoming challenges in computational latency, energy efficiency, and reliability. The twenty-first century has witnessed a transformative shift in the industrial landscape, largely driven by the rapid advancements in artificial intelligence (AI). This change is often called the "Fourth Industrial Revolution" or "Industry 4.0", which focuses on digital interconnectivity, automation, and intelligent decision-making. However, this revolution has further evolved into what is now referred to as "Industry X.0." Industry X.0 encompasses not only the advancements of Industry 4.0 but also integrates new innovation, sustainability, and resilience challenges, making it a more comprehensive concept. This shift is well-articulated in the work by Gallab and Di Nardo, who discuss the new challenges and opportunities in the X.0 era. Unlike its predecessors, this revolution doesn't just change how things are made but alters the very nature of the products and services themselves.

This paper identifies the critical challenges in deploying real-time DL systems and explores potential solutions and emerging trends to address these issues.

2 Distinctive Features of Deep Learning

Generally, Deep Learning is used to extract meaningful data from data (collected from smart things/ internet of things internet connected things) a machine learning method. In general, deep learning is a subset of machine –learning techniques. Initially, concept of deep learning was taken from machine learning. As biggest advantage of deep learning and its popularity is: it works with large amounts of data, while machine learning techniques does not. The "Big Data Era" of technology will provide huge amounts of opportunities for new innovations in deep learning. According to a member of Google-Brain Project, "The analogy to deep learning is that

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*352*
the rocket engine is the deep learning models and the fuel is the huge amounts of data we need to feed to these (existing) algorithms" (Sambit Mahapatra, towardsdatascience.com, (2018)). Deep Learning requires high-end machines contrary to traditional Machine Learning algorithms.

Graphical Processing Unit (GPU) has become an essential part in execution of a Deep Learning algorithm. GPU is used to handle large amount of data, in which processor works in parallel. In traditional Machine learning techniques, most of the applied features need to be identified by a domain expert in order to reduce the complexity of the data and make patterns more visible to learning algorithms to work. On another side, the biggest advantage Deep Learning algorithm also is that they try to learn high-level features from data in an incremental manner. This eliminates the need of domain expertise (skilled people) and hard-core feature extraction, i.e., in this extraction of meaningful information is done automatically by reward based learning (without human-intervention).

2.1 Why do we Require Deep Learning?

We used deep learning to process a large amount of data (big data), especially images, prediction of prices, stock prices, etc. This algorithm is used to solve complex problems using neurons or a concept of hidden layers in its working. Apart from above importance, it is far better than machine learning techniques in producing results, i.e., produce results in minimum times. But, a problem with deep learning is that, it does not tell the way of output, i.e., "How it produced", whereas in machine learning, we know "How an output is produced" with having every step of inputs or processing. Another major difference between Deep Learning and Machine Learning technique is the problem-solving approach. Deep Learning techniques tend to solve the problem end to end, whereas Machine learning techniques need the problem statements to break down to different parts to be solved first and then their results to be combine at final stage. For example, for a multiple object detection problem, Deep Learning techniques like Yolo net take the image as input and provide the

3. Characteristics of Real-Time Systems

Real-time systems are broadly categorized into:

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*352*

- Hard real-time systems, where missing deadlines leads to catastrophic outcomes (e.g., autonomous vehicle control).
- Soft real-time systems, where delays degrade performance but do not result in failure (e.g., video streaming analytics).

Key requirements for real-time systems include:

- Low latency: Processing times must meet application-specific deadlines.
- High throughput: Systems must handle a high volume of requests per second.
- Reliability: Accuracy and consistency of DL predictions are crucial.
- Energy efficiency: Especially important for edge devices in mobile or embedded environments.

4. Challenges in Real-Time DL Deployment

4.1 Computational Latency

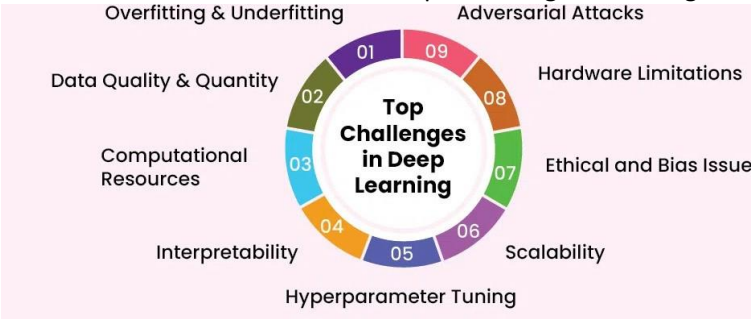
DL models, especially deep architectures like transformers and CNNs, require significant computational resources. Real-time applications cannot tolerate the latency induced by extensive computations.

- Approaches to Mitigate Latency:

Model compression: Techniques like pruning, quantization, and distillation reduce model size and complexity without compromising accuracy significantly.

Hardware acceleration: GPUs, TPUs, and application-specific integrated circuits (ASICs) provide parallelized computations to speed up inference.

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*352*



4.2 Scalability

Real-time systems often experience variable workloads, requiring scalable solutions to maintain performance during peak demand.

- Approaches to Enhance Scalability:

Cloud-edge integration: Offload computationally intensive tasks to the cloud while processing critical tasks on edge devices.

Dynamic load balancing: Allocate resources dynamically based on demand, using containerized deployments and orchestration tools like Kubernetes.

4.3 Energy Efficiency

Power consumption is a critical challenge, particularly for edge devices and IoT systems running DL models.

Approaches to Reduce Energy Consumption:

- Efficient architectures: Use low-power neural networks such as MobileNet and SqueezeNet.
- Neuromorphic computing: Leverage hardware inspired by biological neural networks for energy-efficient processing.

4.4 Robustness and Reliability

Real-time systems operate in unpredictable environments, requiring models to perform consistently despite noise, hardware failures, or adversarial attacks.

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*352*
Approaches to Improve Robustness:

- Train models on diverse and representative datasets to handle real-world variability.
- Incorporate error detection and correction mechanisms.

4.5 Data Bandwidth and Communication Latency

Distributed systems often suffer from delays in transmitting large volumes of data between devices and central servers.

Approaches to Minimize Bandwidth Constraints:

- Edge processing: Perform inference directly on edge devices to reduce communication overhead.
- Efficient encoding: Use data compression techniques to reduce transmission time.

5. Emerging Trends and Solutions

5.1 Edge AI

Edge AI shifts computational workloads closer to data sources, enabling faster responses and reducing dependency on cloud resources. Advances in hardware, such as NVIDIA Jetson and Google Coral, support real-time inference on edge devices.

5.2 5G Networks

5G technology offers low latency and high bandwidth, enabling seamless communication between distributed devices and central processing units. This is particularly advantageous for real-time systems requiring coordination across multiple nodes.

5.3 Federated Learning

Federated learning enables decentralized training of DL models without transferring raw data, addressing privacy and bandwidth concerns. This approach is particularly useful in applications like healthcare and mobile systems.

5.4 Adaptive DL Models

Adaptive DL models can dynamically adjust their complexity based on real-time performance constraints. For example, early-exit architectures allow partial inferences when time constraints are tight.

6. Application Domains

6.1 Autonomous Vehicles

Autonomous systems require real-time processing for tasks such as object detection, path planning, and sensor fusion. Optimized DL models tailored for sensor-specific data (e.g., LiDAR, radar) are crucial in this domain.

6.2 Healthcare

Real-time diagnostic systems need to balance accuracy with interpretability while adhering to strict latency requirements. Lightweight models integrated with explainability frameworks are essential for deployment.

6.3 Financial Services

High-frequency trading platforms and fraud detection systems require ultra-low-latency predictions. Implementing lightweight models on specialized hardware ensures timely decision-making.

7. Problems/Issues with Deep Learning

We have discussed that deep learning technique is too useful (beneficial)in producing some predictions/ decisions based on some data-sets than its other family's fields like machine learning, and artificial intelligence. But it has also raised several issues in it time to time. Apart that, if some algorithms are unable to understand a data then even collected meaningful/ cleaned

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*352*
data also can give inaccurate results. Several problems have been investigated like this problem in deep learning, which will be discussed here. We need to look over such issues in near future.

1. Deep Learning is data hungry: In a world with infinite (i.e., huge) data, and infinite computational resources, we always require some efficient techniques to handle it (Tyagi, A. K., & Reddy, V. K. (2019)). But, were always fail to deliver appropriate tool or methods, because this data is growing everyday (by billions of internet of things or internet connected things). The tools proposed by several researchers do not provide every possible labelled sample of a problem space to a deep learning algorithm. Therefore, it will have to generalize or interpolate between its previous samples in order to classify data it has never seen before (i.e., a new image or sound that is not contained in its dataset). Currently, Deep Learning lacks a mechanism for learning abstractions through explicit, verbal definition, and works best when there are thousands, millions or even billions of training examples Marcus, G. (2018).

2. Deep Learning is Shallow: Deep Learning algorithms are very good at mapping inputs to outputs but not so much at understanding the context of the data they are handling. In fact, the word "deep" in deep learning is much more a reference to the architecture of the technology and the number of hidden layers it contains rather than an allusion to its deep understanding of what it does. For example, naturally apply to abstract concepts like 'justice,' 'democracy' or 'meddling'" (Ben Dickson, bdtechtalks.com, 2018). In an example like gaming, deep learning algorithms can become very good at playing games, and they can eventually beat the best human players in both video and board games. However, this does not mean that Artificial Intelligence (AI) algorithm has the same understanding as humans in the different elements of the game. It has learned through trial and error that making those specific moves will prevent it from losing. For example, Google DeepMind's mastering of the Atari game.

3. Deep Learning is Opaque: While decisions made by rule-based software can be traced back to the last, but in case of machine learning and deep learning algorithms, this facility is not available. This missing feature (of transparency) in deep learning is called as "black box" problem. Deep learning algorithms shift through millions of data points to find patterns and correlations that often go unnoticed to human experts. The decisions they make based on these findings often confound even the engineers who created them.

ISSN: 0937-583x Volume 89, Issue 11 (November -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*352*
4. Is Deep Learning doomed to fail? No, but it is bound for a reality check. In general, deep learning is "a perfectly fine way of optimizing a complex system for representing a mapping between inputs and outputs, given a sufficiently large data set" (Ben Dickson, bdtechtalks.com (2018), Marcus, G. (2018)). Deep learning must be acknowledged for what it is, a highly efficient technique for solving classification problems, which will perform well when it has enough training data and a test set that closely resembles the training data set. Note that if we do not have enough training data, or when our test data differs greatly from our training data, or when we are not solving a classification problem, then deep learning becomes a square peg slammed into a round hole, a crude approximation when there must be a solution elsewhere (Marcus, G. (2018)).

8. Future Directions

- Neuromorphic Hardware: Emerging technologies inspired by biological neural systems promise significant improvements in efficiency and latency for real-time DL.
- Self-Supervised Learning: Reduces reliance on labeled data, making it easier to adapt models to changing environments.
- Explainable AI (XAI): Enhancing transparency and reliability in critical real-time applications like healthcare and finance.

9. Conclusion

Deploying deep learning models for real-time systems is a multifaceted challenge requiring innovations in algorithms, hardware, and system design. While advancements in edge computing, 5G, and model optimization have addressed some barriers, continued research and development are essential to meet the evolving demands of high-speed applications. By leveraging emerging technologies and addressing current limitations, DL can be effectively harnessed for real-time systems, unlocking new possibilities across industries. Deep learning presents both incredible opportunities and significant challenges. Overcoming these challenges requires understanding the underlying issues and implementing effective strategies. By

enhancing data quality, leveraging advanced tools, and addressing ethical concerns, we can use deep learning's full potential. Continuous improvement and adaptation are key to success. Embracing these practices will lead to more robust and impactful deep learning models.

REFERENCES

[1] Arif Cam, Michael Chui, and Bryce Hall. 2019. Global AI survey: AI proves its worth, but few scale impact.

[2] Thomas H. Davenport and Rajeev Ronanki. 2018. Artificial intelligence for the real world. Harv. Bus. Rev. 96, 1 (2018),108–116.

[3] Royal Society (Great Britain). 2017. Machine Learning: The Power and Promise of Computers that Learn by Example:an Introduction. Royal Society.

[4] Michal Pěchouček and Vladimír Mařík. 2008. Industrial deployment of multi-agent technologies: Review and selected case studies. Auton. Agents and Multi-agent Syst. 17, 3 (2008), 397–431.

[5] Kyle Wiggers. 2019. Algorithmia: 50% of companies spend between 8 and 90 days deploying a single AI model. Retrieved from https://venturebeat.com/2019/12/11/algorithmia-50-of-companies-spend-upwards-of-three-monthsdeploying-a-single-ai-model/.

[6] Lawrence E. Hecht. 2019. Add It Up: How Long Does a Machine Learning Deployment Take? Retrieved from https: //thenewstack.io/add-it-up-how-long-does-a-machine-learning-deployment-take/.

[7] Kyle Wiggers. 2019. IDC: For 1 in 4 companies, half of all AI projects fail. Retrieved from https://venturebeat.com/2019/07/08/idc-for-1-in-4-companies-half-of-all-ai-projects-fail/.

[8] Ben Lorica and Nathan Paco. 2018. The State of Machine Learning Adoption in the Enterprise. O'Reilly Media.

[9] 2019. The state of development and operations of AI applications. Dotscience Retrieved from https://dotscience.com/assets/downloads/Dotscience_Survey-Report-2019.pdf.