



Chapter 12

Lightweight Deep Learning: Introduction, Advancements, and Applications

Hari Kishan Kondaveeti

 <https://orcid.org/0000-0002-3379-720X>
Vellore Institute of Technology, India

Valli Kumari Vatsavayi

 <https://orcid.org/0000-0002-7252-8301>
Andhra University, India

Srileakhana Mangapathi

Vellore Institute of Technology, India

Reddy M. Yasaswini

Vellore Institute of Technology, India

ABSTRACT

Lightweight deep learning is a subfield of artificial intelligence and machine learning that prioritises efficiency and compactness while developing deep learning models. It is ideal for low-powered mobile phones, embedded systems, and internet-of-things devices due to their speed and low latency. To make lightweight deep learning models, pruning and quantization are used to remove unnecessary parameters and reduce model weight accuracy. Transfer learning is used to fine-tune a pre-trained deep learning model on a smaller dataset. This chapter introduces the fundamentals of lightweight deep learning, including various lightweight models and their applications across different industries.

DOI: 10.4018/978-1-6684-8386-2.ch012

INTRODUCTION

Lightweight Deep Learning is a branch of Deep Learning that strives to construct neural network models that can perform well even on low-powered platforms such as smartphones, Internet of Things (IoT) devices, and embedded systems. Due to the restricted memory, processing power, and battery life of such devices, it is impossible to install complex deep learning models that need vast computing resources. Lightweight Deep Learning is essential as a result of the growing interest in deploying deep learning models on edge devices. Edge devices often have limited capabilities in terms of their capacity to do sophisticated calculations, hence this limitation necessitates the need for lightweight deep learning. Smart homes, fitness monitors, and smartphone applications that can distinguish faces and voices are just a few examples of emerging technologies.

The Model compression, parameter pruning, quantization, and knowledge distillation are some of the approaches that are employed in Lightweight Deep Learning. These techniques are utilised in order to optimise both the size of the model and its performance. These methods make it possible to minimise the overall size of the model, get rid of any superfluous parameters, and make the necessary computations during inference much easier to do. Reduced model size is one of the many advantages of Lightweight Deep Learning, along with decreased costs, expanded accessibility, and enhanced levels of privacy. However, there are also constraints, such as lower precision, restricted flexibility, complexity, and trade-offs between model size, accuracy, and speed. Additionally, there is a reduction in the overall speed of the model.

ADVANTAGES OF LIGHTWEIGHT DEEP LEARNING

Lightweight deep learning has several advantages over traditional deep learning techniques. In this section, we will discuss some of the significant benefits of lightweight deep learning.

Faster Inference Times

Lightweight deep learning models are significantly smaller than their full-scale counterparts, resulting in faster inference times. This makes them ideal for use cases where real-time processing is critical, such as image or video analysis, natural language processing, and other applications that require quick decision-making.

Lower Memory Usage

Smaller models also require less memory to operate, making them more efficient in resource-constrained environments. This is particularly beneficial for mobile and embedded devices, which have limited resources.

16 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/lightweight-deep-learning/328531

Related Content

Towards Clinical and Operational Efficiency through Healthcare Process Analytics

Vassiliki Koufi, Flora Malamateniou and George Vassilacopoulos (2016). *International Journal of Big Data and Analytics in Healthcare* (pp. 1-17).

www.irma-international.org/article/towards-clinical-and-operational-efficiency-through-healthcare-process-analytics/171401

Artificial Intelligence-Based Sustainable Tourism Planning: A Conceptual Model Proposal

Yunus Topsakal (2025). *Advancing Smart Tourism Through Analytics* (pp. 65-94).

www.irma-international.org/chapter/artificial-intelligence-based-sustainable-tourism-planning/362478

Big Data Analytics in Supply Chain Management

Nenad Stefanovic (2022). *Research Anthology on Big Data Analytics, Architectures, and Applications* (pp. 1801-1816).

www.irma-international.org/chapter/big-data-analytics-in-supply-chain-management/291066

Big Data Analytics in Social Media: An Overview

Janani Balakumar and Vijayarani Mohan (2019). *Machine Learning Techniques for Improved Business Analytics* (pp. 107-124).

www.irma-international.org/chapter/big-data-analytics-in-social-media/207382

Genetic Diagnosis of Cancer by Evolutionary Fuzzy-Rough based Neural-Network Ensemble

Sujata Dash and Bichitrnanda Patra (2020). *Data Analytics in Medicine: Concepts, Methodologies, Tools, and Applications* (pp. 645-662).

www.irma-international.org/chapter/genetic-diagnosis-of-cancer-by-evolutionary-fuzzy-rough-based-neural-network-ensemble/243138