World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:08, 2024

Exploring Deep Neural Network Compression: An Overview

Authors : Ghorab Sara, Meziani Lila, Rubin Harvey Stuart

Abstract : The rapid growth of deep learning has led to intricate and resource-intensive deep neural networks widely used in computer vision tasks. However, their complexity results in high computational demands and memory usage, hindering real-time application. To address this, research focuses on model compression techniques. The paper provides an overview of recent advancements in compressing neural networks and categorizes the various methods into four main approaches: network pruning, quantization, network decomposition, and knowledge distillation. This paper aims to provide a comprehensive outline of both the advantages and limitations of each method.

Keywords : model compression, deep neural network, pruning, knowledge distillation, quantization, low-rank decomposition
Conference Title : ICCIT 2024 : International Conference on Computing and Information Technology
Conference Location : Venice, Italy
Conference Dates : August 15-16, 2024