Candidate Name: Salim Abboudi
Candidate email: salim1abboudi@gmail.com

Trimble / Bilberry : AI Engineer
Task2: Paper Review

1 Introduction

- Title : A Survey of Methods for Low-Power Deep Learning and Computer Vision [1]
- Authors : Goel, Abhinav and Tung, Caleb and Lu, Yung-Hsiang and Thiruvathukal, George K
- Motivation : The motivation behind the paper is to address the challenge of deploying deep learning models, particularly Deep Neural Networks (DNNs), on low-power devices with limited compute resources. While DNNs have achieved remarkable success in various computer vision tasks, their accuracy often requires large numbers of parameters and operations, leading to significant energy consumption, computation, and memory usage. Such resource-intensive models hinder their practical deployment in real-world scenarios, especially in the context of resource-constrained devices like Internet of Things (IoT) devices and edge computing platforms. The authors aim to explore and evaluate methods for compacting and accelerating DNN models, thereby reducing the computational burden while preserving accuracy, with the overarching goal of fostering the development of energy-efficient solutions for low-power deep learning and computer vision applications.

2 Low-Power Computer Vision Techniques

This paper conducts a comprehensive survey of the existing literature, presenting cutting-edge solutions in low-power computer vision. The focus is specifically on low-power Deep Neural Network (DNN) inference, aiming to achieve high throughput. The paper categorizes the low-power inference methods into four distinct categories

2.1 Pruning Parameters and Connections

Pruning Parameters and Connections in Deep Neural Networks (DNNs) reduces memory accesses by removing unimportant parameters and connections. Techniques like Hessian-weighted distortion measure identify importance and enable pruning in fully-connected layers. Extension to convolutional layers is achieved through particle filtering and sample input data. Deep Compression combines pruning, quantization, and encoding to significantly reduce model size by 95%. Advantages include performance gains and reduced overfitting, while disadvantages involve considerable training effort and challenges with sparse matrices. Channel-level pruning is suggested as an improvement to existing techniques for easier implementation without special data structures.

2.2 Convolutional Filter Convolutional

Filter Compression replaces large filters with smaller ones, reducing parameters and computation costs in DNNs. SqueezeNet and MobileNets achieve significant parameter reduction while maintaining high accuracy through 1×1 convolutions and depthwise separable convolutions. Matrix factorization accelerates DNNs up to 4× by creating smaller matrices, but challenges include exponential increase in hyper-parameters with DNN depth. Inclusion of hyper-parameters in the training process can improve implementation and training for large DNNs. Both techniques offer advantages in terms of improved memory, latency, and computation requirements, paving the way for energy-efficient deep learning on resource-constrained devices.

2.3 Network Architecture Search

This method automates DNN architecture design, exploring various possibilities to find efficient architectures for low-power computer vision applications. Network Architecture Search (NAS) uses a Recurrent Neural Network (RNN) controller with reinforced learning to compose candidate architectures, achieving state-of-the-art accuracy. Techniques like MNasNet and Proxyless-NAS further optimize NAS for mobile devices, reducing parameters and operations with improved accuracy. However, NAS algorithms are computationally intensive, requiring significant GPU hours. Researchers propose proxy-based approaches (e.g., FBNet) to reduce computation, but accuracy might suffer. Improvements include parallel training and adaptive learning rates to enhance accuracy while reducing training time.

2.4 Knowledge Transfer

Knowledge Transfer (KT) and Knowledge Distillation (KD) techniques enable small DNNs to learn complex functions by leveraging knowledge from larger pre-trained models. KT involves training the small DNN on data labeled by a larger DNN to benefit from its insights. KD employs a teacher-student paradigm, simplifying training and achieving comparable accuracy to ensembles. While effective, KD may have strict assumptions and rely heavily on softmax outputs. Future research could explore approaches that allow the student to learn neuron activation sequences, enhancing flexibility and generalizability.
In table 1 we can see a general comparison between all suggested methods.

Table 1: Comparison of different techniques for performing low-power computer vision.

Technique	Advantages	Disadvantages
Quantization and Pruning	Negligible accuracy loss with small model size. Highly efficient arithmetic operations.	Difficult to implement on CPUs and GPUs because of matrix sparsity. High training costs.
Filter Compression and Matrix Factorization	High accuracy. Compatible with other optimization techniques.	Compact convolutions can be memory-inefficient. Matrix factorization is computationally expensive.
Network Architecture Search	State-of-the-art accuracy with low energy consumption.	Prohibitively high training costs.
Knowledge Distillation	Low computation cost with few DNN parameters.	Strict assumptions on DNN structure. Only compatible with softmax outputs.

3 Interest in the Paper

The paper offers valuable insights into the task of achieving energy efficiency in Deep Neural Networks (DNNs) for computer vision applications. It provides guidelines for using various techniques and emphasizes the evaluation of performance beyond just accuracy. The paper covers a diverse set of techniques, such as quantization, pruning, compression, factorization, and knowledge distillation, showcasing its thorough exploration of the research landscape. These guidelines are particularly useful for researchers and practitioners aiming to design efficient DNNs for deployment on embedded and mobile devices. The paper's focus on energy consumption aligns well with the increasing demand for sustainable AI solutions, making it highly relevant and appealing to the current research community.

References

[1] Abhinav Goel, Caleb Tung, Yung-Hsiang Lu, and George K Thiruvathukal. A survey of methods for low-power deep learning and computer vision. In *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, pages 1–6. IEEE, 2020.