

A Design Space Exploration Framework for DNN Compression using Low Rank Factorization

Milad Kokhazadeh¹, Georgios Keramidas^{1,2}, Vasilios Kelefouras³

1 Aristotle University of Thessaloniki, Thessaloniki, Greece

2 Think Silicon S.A., An Applied Materials Company, Patras, Greece

3 University of Plymouth, Plymouth, United Kingdom



Motivation

- ❖ DNNs are **computationally expensive**, limiting deployment on edge devices
- ❖ **Low-rank factorization (LRF)** decomposes weight tensors to **reduce storage and computation**
- ❖ **Optimal rank selection, inter-layer dependencies, and balance compression and accuracy**

Problem Statement

- ❖ **DNNs Overparameterization**
- ❖ **Large Design Space**
- ❖ **Suboptimal Compression**
- ❖ **The Lack of a Unified Framework**
- ❖ **Training Overheads**
- ❖ **Computational Complexity of Decomposition**

DSE Methodology (for Dense Layers)

- ❖ **Fully parameterized** and multi-step methodology
- ❖ A stand-alone tool on top of **TensorFlow2.x** and **T3F library**
- ❖ **Grid-Based Design Space Exploration (DSE)** for Optimal Low-Rank Factorization (LRF)
- ❖ Generates a **set of optimal solutions** that satisfy user-defined metrics

- ❖ Keep only solutions with **less memory and FLOPs** compared to the original layer
- ❖ The **memory vs. FLOPs** design space is broken down into tiles, e.g., 8x8 grid
- ❖ Select **multiple solutions** from each tile

Similarity-Based Methodology (for Convolution Layers)

- ❖ **Unified framework**, compatible with all layer types and LRF methods
- ❖ A **dynamic compression ratio** selection strategy based on feature map similarity.
- ❖ Layers are compressed **adaptively**, considering **cosine similarity** of activations.
- ❖ A **step-wise method** refines the compression ratio, ensuring minimal accuracy loss.
- ❖ Evaluates similarity based on **feature maps** instead of weights

- ❖ The main problem for multiple layers: different layers might include tiles of different scales
- ❖ First, the design space and tiling is created for **each layer separately**
- ❖ Then, **corresponding** grid cells must be selected for all layers
- ❖ Fine-tune each solution for a couple of epochs, e.g., **5 epochs**
- ❖ **High-level criteria(s)** defined by user
- ❖ Extract a suitable set of LRF solutions in a **reasonable time**

High Compression with Minimal Accuracy Drop Up to **94.6%** parameter reduction (**82.3%** on average) and **90.7%** FLOPs reduction (**59.6%** on average) accuracy loss below **1.5%**.

Outperformed **VBMF** and **FBP** in most of the cases, providing higher compression while maintaining accuracy.

One-Shot fine-tuning reduces retraining costs

More than 90% compression only in conv. part

Conclusion

Robust methodologies are proposed to facilitate the deployment of DNN models on edge devices by framing LRF-based compression as a DSE problem, utilizing a novel per-layer rank selection strategy that leverages feature map similarities. The approach achieves up to **94.6%** parameter reduction and **90.7%** FLOPs reduction, significantly minimizing retraining needs through an efficient one-shot fine-tuning process.

Future works

- ❖ Extend the methodology to combine multiple LRF techniques.
- ❖ Explore adaptive threshold selection policies for enhanced compression.
- ❖ Optimize the methodology for RISC-V processors.
- ❖ Extend the methodology to combine with other compression techniques.

Seven different DNN models
Up to **99.8%** compression with **minimal** impact in validation accuracy

Publications

1. M. Kokhazadeh, G. Keramidas, V. Kelefouras, I. Stamouli, "A CNN Compression Methodology for Layer-Wise Rank Selection Considering Inter-Layer Interactions," Accepted in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2025
2. M. Kokhazadeh, G. Keramidas, V. Kelefouras, I. Stamouli, "Denseflex: A Low Rank Factorization Methodology for Adaptable Dense Layers in DNNs," Presented in *ACM International Conference on Computing Frontiers (CF'24)*, 2024
3. M. Kokhazadeh, G. Keramidas, V. Kelefouras, I. Stamouli, "A Practical Approach for Employing Tensor Train Decomposition in Edge Devices," Published in *International Journal of Parallel Programming (IJPP)*, Springer, 2024
4. D. Gkountelos, M. Kokhazadeh, C. Bournas, G. Keramidas, V. Kelefouras, "Towards Highly Compressed CNN Models for Human Activity Recognition in Wearable Devices," Presented in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2023
5. G. Keramidas, D. Georgakakis, I. Stamouli, C. Bournas, I. Vasileiou, V. Andreoutsopoulos, D. Tsaligkos, U. Mueller, M. Kokhazadeh, Z. Kokhazadeh, "NEOX TM AI-SDK: Enabling the Deploying of TinyML Models in Wearable Devices," Presented in *Embedded World Exhibition & Conference*, 2023
6. M. Kokhazadeh, G. Keramidas, V. Kelefouras, I. Stamouli, "A Design Space Exploration Methodology for Enabling Tensor Train Decomposition in Edge Devices," Presented in *22th International Conference on Embedded Computer Systems: Architecture, Modeling, and Simulations (SAMOS XXII)*, 2022