# RNN Compression using Hybrid Matrix Decomposition

Urmish Thakker, Jesse Beu, Dibakar Gope, Ganesh Dasika, Matthew Mattina
Arm ML Research Lab

## Overview

- Compression techniques should not impact the inference run-time and task accuracy.
- Hybrid Matrix Decomposition (HMD) can compress RNNs by 2x while being **2x faster than pruning** and **more accurate than a traditional matrix factorization technique**, better enabling the deployment of TinyML applications.

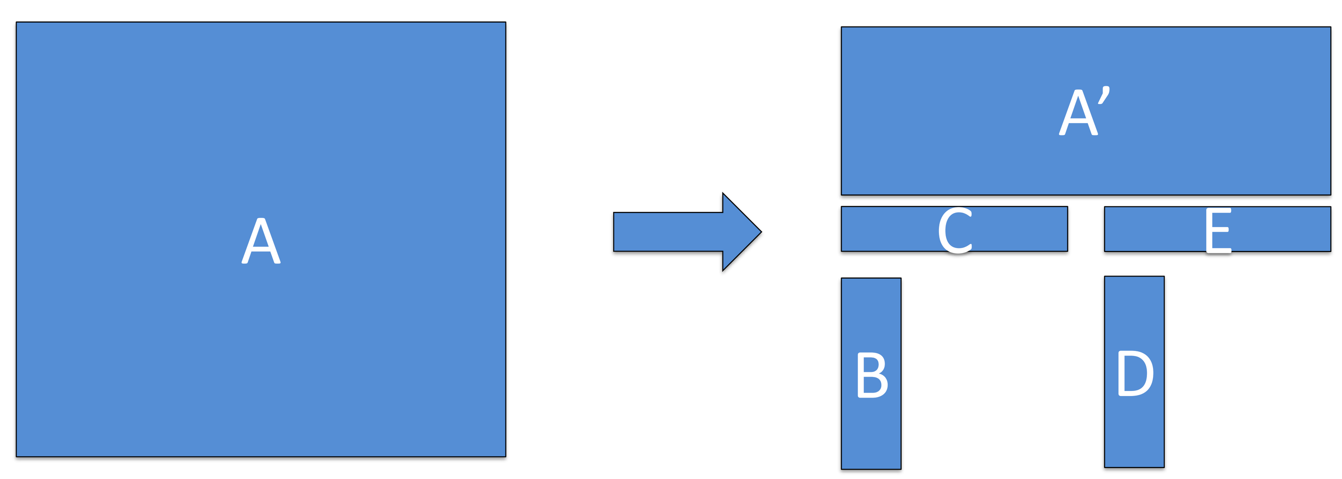
## Motivation

- In an RNN, every element of an output vector is connected to every element of the input and hidden vectors of that RNN layer. However, there are **many output vector elements with sparse dependence on the input and hidden vectors**.
- Most RNN networks are followed by a fully connected softmax layer or another RNN layer. Thus, the **order of the elements in an output vector of RNN hidden layer is not strictly important**.

## Hybrid Matrix Decomposition

- HMD **breaks a matrix into two parts** - a fully parameterized upper half and a constrained lower half.
- This creates a dense matrix representation making it **more hardware-friendly than pruning**. Additionally, it creates a **higher rank matrix than low rank matrix factorization (LMF)**, giving it more expressibility

Dim(A) = mxn




Dim(A') = rxn
Dim(B) = (m-r)x1
Dim(C) = (n/2)x1
Dim(D) = (m-r)x1
Dim(E) = (n/2)x2

Total Parameters = m*n

Total Parameters = r*n + 2*(m-r + n/2)

- Additionally, HMD requires fewer operations to compute the matrix-vector product, as shown in Algorithm 1

**Algorithm 1** Matrix vector product when a matrix uses the HMD technique

**Input 1:** Matrices $A'$, $B$, $C$, $D$, $E$
**Input 2:** Vector $I$ of dimension $n \times 1$
**Output:** Matrix $O$ of dimension $m \times 1$

$O_{1:r} \leftarrow A' \times I$
$Temp1Scalar \leftarrow C \times I_{1:n/2}$
$Temp1 \leftarrow B \circ Temp1Scalar$
$Temp2Scalar \leftarrow E \times I_{1+n/2:n}$
$Temp2 \leftarrow D \circ Temp2Scalar$
$O_{r+1:m} \leftarrow Temp1 + Temp2$
$O = concatenate\{O_{1:r}, O_{r+1:m}\}$

## Results

- Compressed human activity recognition networks (HAR1 and HAR2) in [1] and [2] by a factor of 2
- HMD leads to better runtime than pruning with equivalent accuracy and better accuracy than low rank matrix factorization

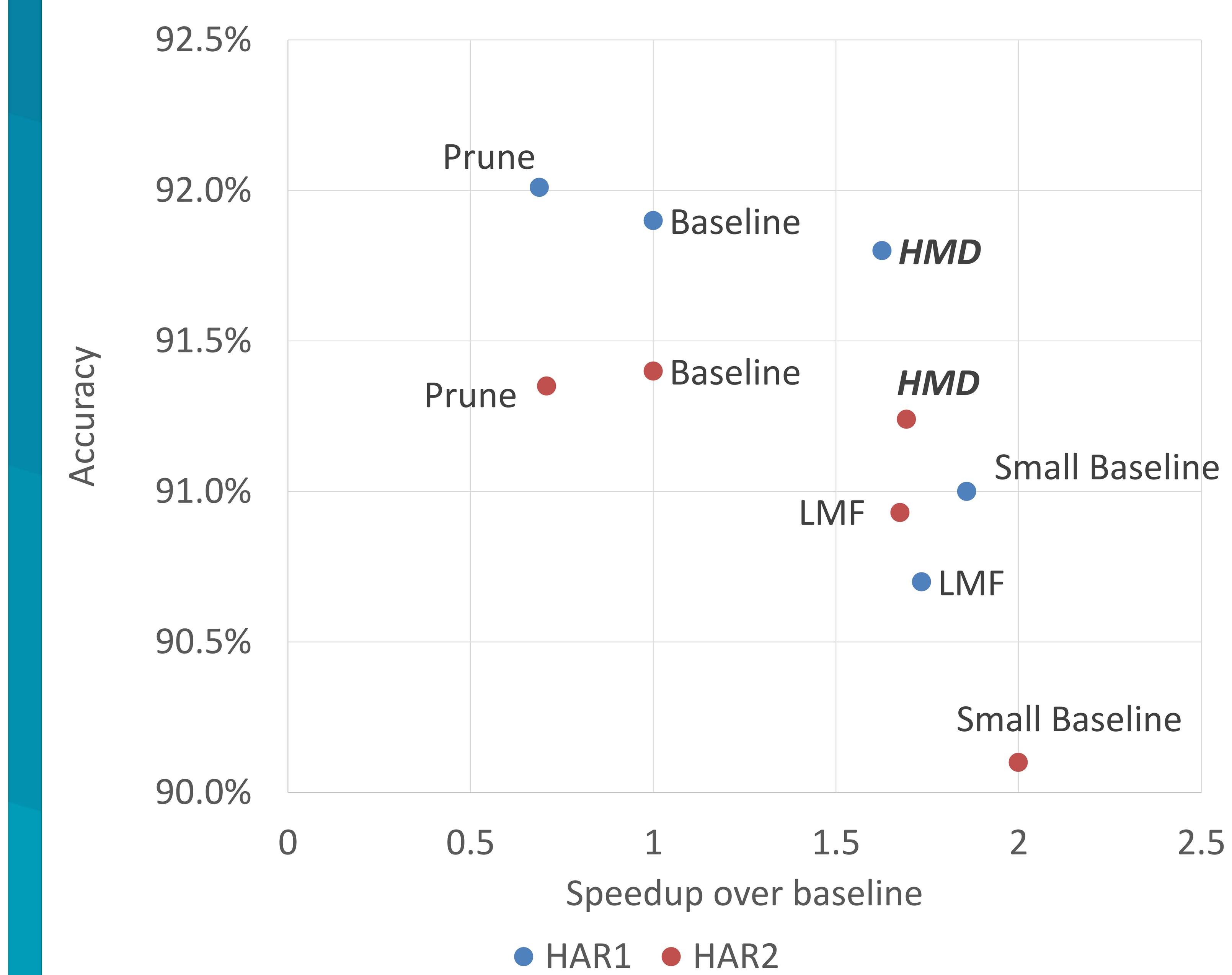



Fig: Accuracy vs speed-up over baseline when HAR1 and HAR2 networks are compressed using 3 different compression techniques

## References

[1] N. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables", IJCAI 2016

[2] F. Ordez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition", Sensors 2016