
Compressing RNNs for IoT devices by 15-38x using Kronecker Products

Urmish Thakker

Arm ML Research
Austin

urmish.thakker@arm.com

Jesse Beu

Arm ML Research
Austin

jesse.beu@arm.com

Dibakar Gope

Arm ML Research
Austin

dibakar.gope@arm.com

Chu Zhou

Arm ML Research
Boston

chu.zhou@arm.com

Igor Fedorov

Arm ML Research
Boston

igor.fedorov@arm.com

Ganesh Dasika

Arm ML Research
Austin

ganesh.dasika@arm.com

Matthew Mattina

Arm ML Research
Boston

matthew.mattina@arm.com

Abstract

Recurrent Neural Networks (RNN) can be large and compute-intensive, making them hard to deploy on resource constrained devices. As a result, there is a need for compression technique that can significantly compress recurrent neural networks, without negatively impacting task accuracy. This paper introduces a method to compress RNNs for resource constrained environments using Kronecker products. We call the RNNs compressed using Kronecker products as Kronecker product Recurrent Neural Networks (KPRNNs). KPRNNs can compress the LSTM[22], GRU [9] and parameter optimized FastRNN [30] layers by $15 - 38\times$ with minor loss in accuracy and can act as in-place replacement of most RNN cells in existing applications. By quantizing the Kronecker compressed networks to 8-bits, we further push the compression factor to $50\times$. We compare the accuracy and runtime of KPRNNs with other state-of-the-art compression techniques across 5 benchmarks spanning 3 different applications, showing its generality. Additionally, we show how to control the compression factors achieved by Kronecker products using a novel hybrid decomposition technique. We call the RNN cells compressed using Kronecker products with this control mechanism as hybrid Kronecker product RNNs (HKPRNN). Using HKPRNN, we compress RNN Cells in 2 benchmarks by $10\times$ and $20\times$ achieving better accuracy than other state-of-the-art compression techniques.

1 Introduction

Recurrent Neural Networks (RNNs) have shown state-of-the-art accuracy for many applications that use time-series data. As a result, RNNs can greatly benefit important Internet-of-Things (IoT) applications like wake-word detection [54], human activity recognition [18, 38, 39] and predictive maintenance [3, 42]. IoT applications typically run on highly constrained devices. Due to their energy, power, and cost constraints, IoT devices frequently use low-bandwidth memory technologies and smaller caches compared to desktop and server processors. For example, some IoT devices

have 2KB of RAM and 32 KB of Flash Memory [17, 29]. The size of typical RNN layers can prohibit deployment of these networks on IoT devices or reduce the efficiency of the execution of these networks on devices with small capacity caches [43]. Thus, there is a need for a compression technique that can drastically compress RNN layers without sacrificing the task accuracy. Our results (section 4) show that popular compression techniques like pruning [6, 19, 56] and low-rank matrix factorization (LMF) [7, 16, 28] can lead to significant loss in accuracy ($>3\%$) when IoT applications are compressed by factors of $15\times$ or more. Additionally, a compression technique should not sacrifice run-time during inference as some of these applications can have hard real time deadlines. We provide an alternative to pruning and LMF that can achieve these objectives.

This paper makes the following key contributions:

- To the best of our knowledge, this is the first paper that shows how to use Kronecker products to compress RNN layers. We call the RNN cells compressed using Kronecker products as KPRNNs. We compress the LSTM [22], GRU [9] and parameter optimized FastRNN [30] layers across 5 benchmarks by $15\text{--}38\times$ using Kronecker products. We were able to push the compression factor to $50\times$ by quantizing the networks to 8-bits.
- We show a novel way to control the compression factors of layers compressed using Kronecker products. We call the layers compressed using controlled Kronecker product compression as Hybrid Kronecker product RNNs (HKPRNN) and compress LSTM layers in 2 benchmarks by $10\times$ and $20\times$.
- We compare the accuracy and speed-up over baseline during inference for the compressed networks with magnitude pruning and low-rank matrix factorization (LMF) showing that KPRNNs outperform these state-of-the-art compression techniques.

2 Related work

The research in neural network (NN) compression can be roughly categorized under 4 topics: Pruning, structured matrix based techniques, quantization, and tensor decomposition. **Pruning** [6, 19, 36, 56] has been a prominent compression technique. **Structured matrices** have shown significant potential for compression of neural network [12, 41]. Block circular compression [11, 48] is an extension of structured matrix based compression, converting every block in a matrix into a structured matrix. Compression using structured matrices translates to inference speed-up over baseline on CPU for larger matrices only [44] or when using specialized hardware [8, 41]. **Tensor decomposition** (CP decomposition [31], Tucker decomposition [5], etc.) based methods have also shown significant reduction in parameters [45]. Matrix Factorization [7, 16, 28] can also be categorized under the tensor decomposition topic. Lastly, **quantization** is another popular technique for compression [20, 23, 33, 47, 55]. The benefits of quantization can be orthogonal to the compression techniques discussed above.

GRUs and LSTMs have $3\times\text{--}4\times$ more parameter than RNN. Another way to compress RNNs is to replace the LSTM and GRU cells with lightweight RNN Cells. However, RNN Cells are hard to train and can lead to vanishing and exploding gradients. Thus, any work that leads to stable training of RNN Cells can potentially compress neural networks by factor of $3\times\text{--}4\times$. Various techniques [4, 25, 26, 30, 35, 51, 53] have been proposed to stabilize RNN training. FastRNN cells [30] shows the most promise amongst these set of work and has shown promising results in the IoT domain. In this paper, we further compress FastRNN cells by $16\times$ using the proposed technique.

Kronecker products have been used in NN before in [26, 52]. Zhou et al. [52] use Kronecker products to compress fully connected layers in AlexNet. They start with a pre-trained model and use a low rank decomposition technique to find the sum of Kronecker products that best approximate the FC layer. We deviate from their work as we use Kronecker products to compress RNNs and learn the matrices using back-propagation. Jose et al. [26] use Kronecker products to stabilize RNN training by adding the unitary constraint. A detailed discussion of how this work differs from [26] can be found in section 3.

3 Kronecker Product Recurrent Neural Networks

3.1 Background

Let A , B and C be three matrices, then the Kronecker product is expressed as

$$A = B \otimes C \quad (1)$$

$$A = \begin{bmatrix} b_{1,1} \circ C & b_{1,2} \circ C & \dots & b_{1,n1} \circ C \\ b_{2,1} \circ C & b_{2,2} \circ C & \dots & b_{2,n1} \circ C \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1,1} \circ C & b_{m1,2} \circ C & \dots & b_{m1,n1} \circ C \end{bmatrix}$$

where, $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{m1 \times n1}$, $C \in \mathbb{R}^{m2 \times n2}$, $m = m1 \times m2$, $n = n1 \times n2$ and \circ is the hadamard product. The variables B and C will be referred to as the Kronecker factors of A in this paper. The algorithm to calculate the Kronecker product of two matrices in Tensorflow is given in Algorithm 3 in Appendix C.

An RNN layer has two sets of weight matrices - input-hidden and hidden-hidden (also known as recurrent). The input-hidden matrix gets multiplied with the input, while the hidden-hidden (or recurrent) matrix gets multiplied by the hidden vector. Jose et al. [26] use Kronecker factors of size 2×2 to replace the hidden-hidden matrices of every RNN layer. Thus a traditional RNN Cell, represented by:

$$h_t = f([W_x \ W_h] * [x_t; h_{t-1}]) \quad (2)$$

$$(3)$$

Is replaced by,

$$h_t = f([W_x \ W_0 \otimes W_2 \dots \otimes W_{F-1}] * [x_t; h_{t-1}]) \quad (4)$$

where W_x (input-hidden matrix) $\in \mathbb{R}^{m \times n}$, W_h (hidden-hidden or recurrent matrix) $\in \mathbb{R}^{m \times m}$, $W_0 \dots W_{F-1} \in \mathbb{R}^{2 \times 2}$, $x_t \in \mathbb{R}^{n \times 1}$, $h_t \in \mathbb{R}^{m \times 1}$, and $F = \log_2(m) = \log_2(n)$. Thus a 256×256 sized matrix will be expressed as a product of 8 matrices of size 2×2 . This can potentially lead to approximately $2 \times$ compression. The aim of Jose et al. [26] was to stabilize RNN training to avoid vanishing and exploding gradients. They add a unitary constraint to these 2×2 matrices, stabilizing RNN training. However, in order to regain the baseline accuracy, they needed to increase the size of the RNN layers. Thus, they do not achieve significant compression.

We tried using 2×2 Kronecker factor matrices for hidden-hidden/recurrent matrices of GRU layers [9] of the key-word spotting network [54]. This resulted in an approximately $2 \times$ reduction in the number of parameters. However, the accuracy dropped by 3% relative to the baseline. When we examined the 2×2 matrices, we observed that, during training, the values of some of the matrices hardly changed after initialization (see Appendix A). Additionally, using 2×2 matrices leads to significant slow-down during inference [26]. We leverage this observation in developing the method discussed in this paper to compress the RNN layers using Kronecker products.

3.2 KPRNN cells

KPRNN cells are RNN cells with all of the matrices compressed by replacing them with Kronecker products of smaller matrices. We restrict the number of Kronecker factors to two. We use Algorithm 4 in Appendix C to find the dimensions of the Kronecker factors. The algorithm takes in the prime factors of the dimensions of the input matrix and returns the dimensions of the two Kronecker factor matrices by converting the list of prime factors for each input dimension into the smallest two numbers, whose product will return a value equal to that dimension. For example, for an input matrix, A , of dimension 154×164 , this algorithm would suggest creation of Kronecker factor matrices of dimension 11×41 and 14×4 , where $11 \times 14 = 154$ and $41 \times 4 = 164$. This leads to a $50 \times$ reduction in the number of parameters required to store A .

Instead of starting with a trained network and decomposing its matrices into the Kronecker factors, we replace the RNN/LSTM/GRU cells [9, 22] in a neural network with its Kronecker equivalent and

train the entire model from the beginning. We call these cells Kronecker product RNN cells. Below are the equations for the RNN cells and the KPRNN cells:

$$RNN \text{ Cell} : h_t = f([W_x \ W_h] * [x_t; h_{t-1}]) \quad (5)$$

$$KPRNN \text{ Cell} : h_t = f((W_1 \otimes W_2) * [x_t; h_{t-1}]) \quad (6)$$

where $W_x \in \mathbb{R}^{m \times n}$, $W_h \in \mathbb{R}^{m \times m}$, $x_t \in \mathbb{R}^{n \times 1}$, $h_t \in \mathbb{R}^{m \times 1}$, $W_{x1} \in \mathbb{R}^{m1 \times n1}$, $W_{x2} \in \mathbb{R}^{m2 \times n2}$, $m1 \times m2 = m$ and $n1 \times n2 = (m + n)$. Thus, KPRNN replaces the W_x and W_h matrices in the RNN cells with a Kronecker product of two smaller matrix. LSTM, GRU and FastRNN cells are compressed in a similar fashion, by replacing the matrices in these layers by Kronecker products of two smaller matrices.

3.2.1 Matrix Vector Product calculation in KPRNN cells

Algorithm 1 Implementation of Matrix Vector Product, when matrix is expressed as a Kronecker product of two matrices

Input: Matrices A of dimension $m1 \times n1$, B of dimension $m2 \times n2$ and x of dimension $n \times 1$.
 $m = m1 \times m2$, $n = n1 \times n2$

Output: Matrix y of dimension $m \times 1$

- 1: $X = \text{reshape}(x, n2, n1)$ {reshapes the x vector to a matrix of dimension $n2 \times n1$ }
- 2: $At = A.\text{transpose}()$
- 3: $Y = B \times X \times At$
- 4: $y = \text{reshape}(Y, m, 1)$ {reshapes the y vector to a matrix of dimension $m \times 1$ }

For inference on IoT devices, it is safe to assume that the batch size will be one [43]. When the batch size is one, the RNN cells compute matrix vector product during inference. In case of KPRNN cells, this will turn out to be:

$$y = (A \otimes B) \times x \quad (7)$$

where, $y \in \mathbb{R}^{m \times 1}$, $x \in \mathbb{R}^{n \times 1}$, $A \in \mathbb{R}^{m1 \times n1}$, $B \in \mathbb{R}^{m2 \times n2}$ and $m = m1 \times m2$, $n = n1 \times n2$. One possible way to calculate this matrix vector product is to expand the Kronecker product between the A and B matrices and to calculate the matrix-vector product between the resultant matrix and x . However, this will lead to an increase in the number of computations required. A better method that exploits the block structure of Kronecker product and avoids expanding the matrix is shown in Algorithm 1. This leads to significant speed-up during inference (section 4.1). The derivation of the algorithm can be found in [1] and is included in Appendix (C.1).

3.3 Hybrid Kronecker Product Recurrent Neural Network

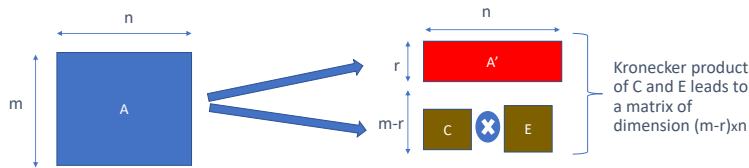


Figure 1: Matrix representation for matrices in a HKPRNN cells

KPRNN can be an extremely effective compression technique, as we will illustrate in Section 4.1. However, sometimes the accuracy loss induced by KPRNN compression may be too large for the technique to be useful. Other compression techniques like pruning [56] and LMF [10] have fine-grained control to set the amount of compression via pre-determining the sparsity (pruning) or setting the rank of the matrix. These control mechanisms can help regain some of the lost accuracy by increasing the number of parameters in the layer via decreasing the sparsity or increasing the rank of the matrix.

Currently, the factor by which a network is compressed using KPRNN can only be controlled by two ways – by increasing the size of the Kronecker factor matrices or by increasing the number of layers of KPRNN cells. Increasing the size of the Kronecker factor leads to an RNN layer with a

Algorithm 2 Implementation of Matrix Vector Product, when matrix is expressed as a Hybrid of unconstrained upper part and a lower part created using Kronecker product of two matrices

Input: Matrices A' of dimension $r \times n$, C of dimension $m1 \times n1$, E of dimension $m2 \times n2$ and x of dimension $n \times 1$. $m1 \times m2 = (m - r)$, $n = n1 \times n2$

Output: Matrix y of dimension $m \times 1$

```

 $y_{upper} = A' \times x$ 
 $X = \text{reshape}(x, n2, n1)$  {reshapes the x vector to a matrix of dimension  $n2 \times n1$ }
 $At = A.\text{transpose}()$ 
 $Y_{lower} = B \times X \times At$ 
 $y_{lower} = \text{reshape}(Y, m, 1)$  {reshapes the x vector to a matrix of dimension  $(m - r) \times 1$ }
 $y = \text{concat}(y_{upper}, y_{lower})$ 

```

larger hidden vector. While this might well lead to a valid solution, it removes the possibility of using KPRNN as an in-place replacement in an existing RNN. This also increases the size of the softmax layers or the subsequent RNN layers that usually follow an RNN layer. Alternatively, increasing the number of layers leads to a deeper network which can be hard to train. An additional constraint on the use of KPRNNs is that they cannot compress RNNs in an existing application, if one of the dimensions of the matrix of an RNN layer is a prime number.

In order to solve these issues, we propose the Hybrid Kronecker Product mechanism to compress RNNs. We refer to RNNs compressed using this mechanism as, “HKPRNN”, in this paper. HKPRNN divides a matrix in a neural network into two parts – an unconstrained upper part and a lower part created using the Kronecker product of two matrices. This is illustrated in Figure 1. Below are the equations for the RNN cells and the HKPRNN cells:

$$RNNCell : h_t = f([W_x \ W_h] * [x_t; h_{t-1}]) \quad (8)$$

$$HKPRNNCell : h_t = f([A'; (W_1 \otimes W_2)] * [x_t; h_{t-1}]) \quad (9)$$

where $W_x \in \mathbb{R}^{m \times n}$, $W_h \in \mathbb{R}^{m \times m}$, $A \in \mathbb{R}^{r \times (m+n)}$, $W_{x1} \in \mathbb{R}^{m1 \times n1}$, $W_{x2} \in \mathbb{R}^{m2 \times n2}$, $m1 \times m2 = (m-r)$ and $n1 \times n2 = (m+n)$. Thus, by cleverly selecting r , we can tune the amount of compression. LSTMs and GRUs are compressed in a similar fashion, by replacing the matrices in these layers with their hybrid Kronecker product representation. Algorithm 2 shows how to calculate the matrix vector product without expanding the matrix into its full representation.

4 Results

	MNIST-LSTM	USPS-FastRNN	KWS-LSTM	KWS-GRU	HAR1-BiLSTM
Application Domain	Image Classification	Image Classification	Key-word spotting	Key-word spotting	Human Activity Recognition
Reference Paper		[30]	[54]	[54]	[18]
Cell Type	LSTM [22]	FastRNN [30]	LSTM [22]	GRU [9]	Bidirectional LSTM [40]
Dataset	[32]	[24]	[49]	[49]	[39]
Accuracy	99.40%	93.77%	92.50%	93.50%	91.90%
#Parameters	11,450	1,856	62,316	78,090	374,468
Size, assuming 32b weights	44.73 KB	7.25 KB	243.42 KB	305.04 KB	1,462.77 KB
Runtime (ms)	6.4	1.175	26.8	67	470

Table 1: Benchmarks evaluated in this paper. These benchmarks represent some of the key applications in the IoT domain. We cover a wide variety of applications and RNN cell types.

Other compression techniques evaluated: We compare networks compressed using KPRNN and HKPRNN with magnitude pruning [56] and low-rank matrix factorization (LMF). While there are multiple possible ways to prune [34, 37], magnitude pruning has shown comparable or better accuracy

compared to other pruning techniques [14]. For an additional comparison point, we also train a smaller baseline with the same number of parameters as the compressed baseline.

Training platform, infrastructure and measuring inference run-time: We use Tensorflow 1.12[2] as the training platform and 4 Nvidia RTX 2080 GPUs to train our benchmarks. To measure the inference run-time, we implement the baseline and the compressed cells in C++ using the Eigen library [13] and run them on the Arm Cortex-A73 cores on a Hikey 960 development board.

Dataset and data pre-processing: We evaluate the impact of compression using the techniques discussed in section 3 on a wide variety of benchmarks spanning applications like key-word spotting, human activity recognition, and image classification. The details regarding the datasets used; and the size of the train, test, and validation sets can be found in Appendix B.1. The details regarding input pre-processing for various benchmarks can be found in Appendix B.2.

Benchmarks: Table 1 shows the benchmarks used in this work. The hyperparameters used for baseline networks are discussed in Appendix B.3. Appendix D.1 and Appendix E.1 discuss the hyperparameters of the KPRNN and HKPRNN networks, respectively, and their corresponding comparison techniques. The appendix also discusses the mean and variance of the accuracy of these networks after compression.

Evaluation Criteria: We evaluate and compare the compressed networks based on the final accuracy of the network on the held out test set. We also measure the run-time (wall clock time taken to execute a single inference) on the Hikey platform and report the speed-up over the baseline. Together, these two metrics help us evaluate whether the proposed training technique can help recover accuracy after significant compression without sacrificing any real-time deadlines these applications may have.

4.1 KPRNN networks

Figure 2 shows the results of applying the Kronecker Product technique across a wide variety of applications and RNN cells. As mentioned in Section 3, using only two matrix factors, only one level of compression is feasible. The compression achieved for each network is mentioned in the captions and is quite substantial – ranging from $16\times$ to $38\times$ for our benchmarks. The KPRNN networks are compared to the uncompressed baseline and the networks generated when alternative compression techniques are used to achieve the same compression ratio as KPRNNs. This allows for a fair comparison of accuracy and run-time across the different techniques. We find that KPRNNs are consistently the most accurate of the compressed networks and are faster than the baseline.

A few results are of particular note. The USPS network uses FastRNN cells which are highly optimized RNN cells that avoid exploding and vanishing gradient problems associated with other RNN cells, and they do so without adding additional computation. Given that the FastRNN cells are not over-parameterized, they represent a great benchmark to identify whether a compression technique is effective. As shown in Figure 2b, using the Kronecker products these highly optimized cells are compressed by a factor of $16\times$ with minimal loss in accuracy, unlike the alternative compression techniques. Figure 2d is the only result that does not have a data point for magnitude pruning. This is because the magnitude pruning infrastructure we used [56] is not available for GRU-based networks. The Kronecker Product-based network is still more accurate than the remaining alternatives.

Additional details about how these experiments were run, the mean and variance of the accuracy, etc. can be found in Appendix D.1.3 and D.1.4.

Relationship between accuracy, rank, condition number, singular values and the compression techniques : In general, the poor performance of the LMF technique can be attributed to “rank-collapse”. For all of the benchmarks, LMF will only achieve the required compression by reducing the rank of the matrix significantly (generally < 10). Kronecker Products, on the other hand, will create a full rank matrix [50], if the Kronecker factors are fully ranked

$$\text{rank}(A \otimes B) = \text{rank } A \cdot \text{rank } B. \quad (10)$$

We observe that, Kronecker factors of all the compressed benchmarks are fully-ranked. A full-rank matrix can also lead to poor accuracy if it is ill-conditioned [15]. However, KPRNN learns matrices that do not exhibit this behavior. The condition numbers of the matrices of the best-performing KPRNN compressed networks discussed in this paper are in the range of 1.2 to 7.3.

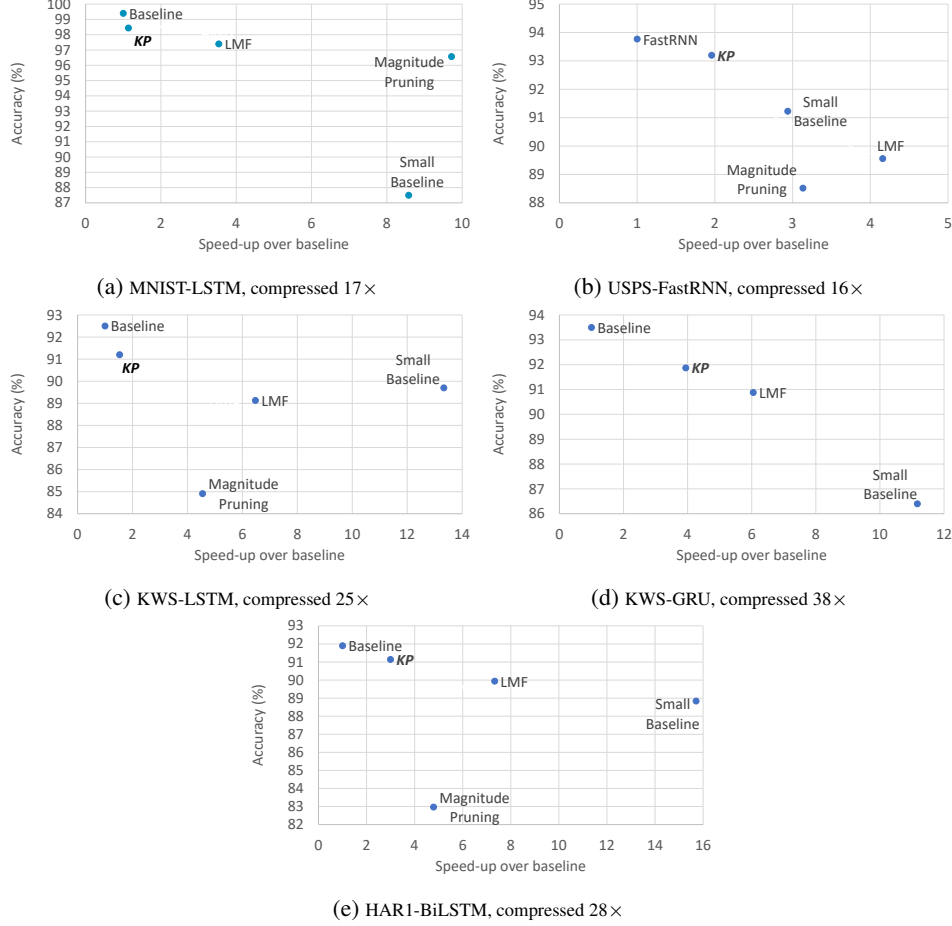


Figure 2: Model accuracy vs inference speed-up for our benchmarks. The baseline networks are compared to networks compressed by either the Kronecker Product (KP), magnitude pruning, low-rank matrix factorization (LMF), or by scaling the network size. Each compressed network has fewer RNN parameters than the baseline by the amount indicated. The Kronecker Product-based networks are consistently the most accurate alternative while still having speed-up over the baseline.

To prune a network to the same compression factor as KPRNN, networks need to be pruned to 94% sparsity or above. It has been observed that pruning leads to significant accuracy drop beyond 90% sparsity for parameter efficient models [14]. Pruning FastRNN cells to the required compression factor leads to an ill-conditioned matrix. This might explain the poor accuracy of sparse FastRNN network. However, for other pruned networks, the resultant sparse matrices have a condition number less than 20 and are fully-ranked. Thus, condition number does not explain the loss in accuracy for these benchmarks.

To further understand the loss in accuracy of pruned LSTM networks, we looked at the singular values of the resultant sparse matrices in the KWS-LSTM network. Let $y = Ax$. The largest singular value of A upper-bounds $\|y\|_2$, i.e. the amplification applied by A . Thus, a matrix with larger singular value can lead to an output with larger norm [46]. Since RNNs execute a matrix-vector product followed by a non-linear sigmoid or tanh layer, the output will saturate if the value is large. The matrix in the LSTM layer of the best-performing pruned KWS-LSTM network has its largest singular value in the range of 48 to 52 while the baseline KWS-LSTM network learns a LSTM layer matrix with largest singular value of 19 and the Kronecker product compressed KWS-LSTM network learns LSTM layers with singular values less than 15. This might explain the especially poor results achieved after pruning this benchmark. Similar observations can be made for the pruned HAR1 network.

We looked into the condition number and largest singular value of small baseline networks also. However, we did not see a consistent story. The small baseline for KWS-LSTM network learned a

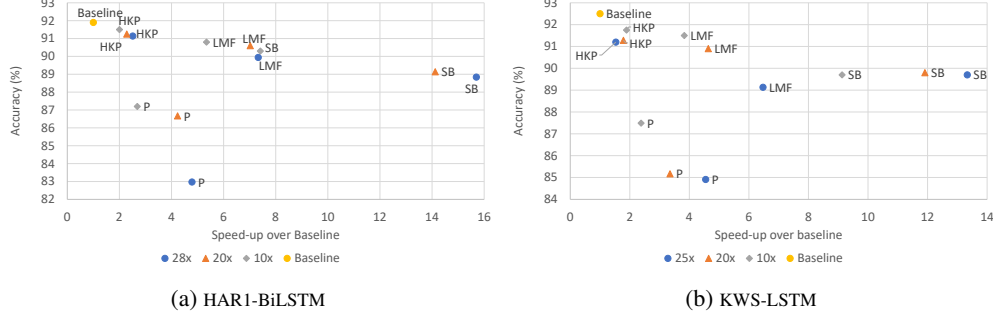


Figure 3: Model accuracy vs inference speed-up for HAR1-BiLSTM and KWS-LSTM using varying compression ratios. The baseline networks are compared to networks compressed by either the Hybrid Kronecker Product (HKP), magnitude pruning (P), low-rank matrix factorization (LMF), or by scaling the network size (SB). Each compressed network has fewer RNN parameters than the baseline by the amount indicated. The Kronecker Product-based networks are consistently the most accurate alternative, at all compression levels, while still having speed-up over the baseline.

RNN layer with a matrix whose condition number is > 90 and largest singular value is > 50 . But for other networks, small baseline learned well conditioned matrices with small singular values in the RNN layers.

Quantization: One of the most commonly used techniques to reduce the size and computation of a neural network is quantization. To check whether using Kronecker Products conflicts with quantization, we quantized the HAR1 and KWS-LSTM networks to 8 bits. This led to an overall compression factor of $50\times$ and $30\times$, and a corresponding accuracy loss of 0.24% and 0.16%, respectively. Based on the minimal loss in accuracy, we feel that in addition to compressing a network using Kronecker products, additional savings can still be had through the use of quantization.

4.2 HKPRNN Networks

As mentioned in Section 3.3, using the two-matrix Kronecker Product technique results in only one possible compression ratio, and using the hybrid HKPRNN technique is a useful way to control the level of compression and the corresponding reduction in accuracy and run-time. Figure 3 shows the results from using HKPRNN. These are similar graphs to those shown in Figure 2, but rather than using the only compression factor allowed by KPRNN, three possible compression ratios were explored – $10\times$, $20\times$, and the maximum compression ratio – resulting in the three data points for each compression scheme. The maximum compression ratio for the Kronecker Product technique is when a hybrid scheme is not used at all (i.e., $r = 0$), so the HKPRNN data points at the maximum compression ratio are equivalent to the corresponding KPRNN data points in Figure 2.

Even at non-maximal compression ratios, the Hybrid Kronecker Product technique consistently results in superior accuracy to the alternative techniques. This illustrates that HKPRNN can be effectively used to modulate the accuracy loss from compression.

Additional details about the training hyperparameters used, the mean and variance of the accuracy and the specific model sizes and run-times can be found in Appendix E.1.1 and E.1.2.

5 Conclusion

We show how to compress RNN Cells by $15\times$ to $38\times$ using Kronecker products. We call the cells compressed using Kronecker products as KPRNNs. KPRNNs can act as a drop in replacement for most RNN layers and provide the benefit of significant compression with marginal impact on accuracy. Additionally, we show how to control the compression achieved by KPRNN by suggesting a novel hybrid compression technique. We call this family of controlled Kronecker compressed network as HKPRNN and show how we can compress the network by a factor of $10 - 20\times$. None of the other compression techniques (pruning, LMF) match the accuracy of the Kronecker compressed

networks. We show that this compression technique works across 5 benchmarks that represent key applications in the IoT domain.

References

- [1] Introduction to kronecker products. <http://www.mathcs.emory.edu/~nagy/courses/fall10/515/KroneckerIntro.pdf>. Accessed: 2019-05-20.
- [2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [3] Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134 – 147, 2017. Online Real-Time Learning Strategies for Data Streams.
- [4] Martín Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. *CoRR*, abs/1511.06464, 2015.
- [5] Giuseppe Giovanni Calvi, Ahmad Moniri, Mahmoud Mahfouz, Zeyang Yu, Qibin Zhao, and Danilo P. Mandic. Tucker tensor layer in fully connected neural networks. *CoRR*, abs/1903.06133, 2019.
- [6] Soravit Changpinyo, Mark Sandler, and Andrey Zhmoginov. The power of sparsity in convolutional neural networks. *CoRR*, abs/1702.06257, 2017.
- [7] Ting Chen, Ji Lin, Tian Lin, Song Han, Chong Wang, and Denny Zhou. Adaptive mixture of low-rank factorizations for compact neural modeling. *Advances in neural information processing systems (CDNNRIA workshop)*, 2018.
- [8] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. Choudhary, and S. Chang. An exploration of parameter redundancy in deep networks with circulant projections. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2857–2865, Dec 2015.
- [9] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.
- [10] Misha Denil, Babak Shakibi, Laurent Dinh, Marc’Aurelio Ranzato, and Nando de Freitas. Predicting parameters in deep learning. *CoRR*, abs/1306.0543, 2013.
- [11] Caiwen Ding, Siyu Liao, Yanzhi Wang, Zhe Li, Ning Liu, Youwei Zhuo, Chao Wang, Xuehai Qian, Yu Bai, Geng Yuan, Xiaolong Ma, Yipeng Zhang, Jian Tang, Qinru Qiu, Xue Lin, and Bo Yuan. Circnn: Accelerating and compressing deep neural networks using block-circulant weight matrices. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO-50 ’17*, pages 395–408, New York, NY, USA, 2017. ACM.
- [12] Caiwen Ding, Ao Ren, Geng Yuan, Xiaolong Ma, Jiayu Li, Ning Liu, Bo Yuan, and Yanzhi Wang. Structured weight matrices-based hardware accelerators in deep neural networks: Fpgas and asics. In *Proceedings of the 2018 on Great Lakes Symposium on VLSI, GLSVLSI ’18*, pages 353–358, New York, NY, USA, 2018. ACM.
- [13] Benoit Jacob Gael Guennebaud. Eigen library. <http://eigen.tuxfamily.org/>. Accessed: 2018-12-21.
- [14] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574, 2019.

- [15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [16] Artem M. Grachev, Dmitry I. Ignatov, and Andrey V. Savchenko. Neural networks compression for language modeling. In B. Uma Shankar, Kuntal Ghosh, Deba Prasad Mandal, Shubhra Sankar Ray, David Zhang, and Sankar K. Pal, editors, *Pattern Recognition and Machine Intelligence*, pages 351–357, Cham, 2017. Springer International Publishing.
- [17] Chirag Gupta, Arun Sai Suggala, Ankit Goyal, Harsha Vardhan Simhadri, Bhargavi Paranjape, Ashish Kumar, Saurabh Goyal, Raghavendra Udupa, Manik Varma, and Prateek Jain. ProtoNN: Compressed and accurate kNN for resource-scarce devices. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1331–1340, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.
- [18] Nils Y Hammerla, Shane Halloran, and Thomas Ploetz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *IJCAI 2016*, 2016.
- [19] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *International Conference on Learning Representations (ICLR)*, 2016.
- [20] Qinyao He, He Wen, Shuchang Zhou, Yuxin Wu, Cong Yao, Xinyu Zhou, and Yuheng Zou. Effective quantization methods for recurrent neural networks. *CoRR*, abs/1611.10176, 2016.
- [21] Qinyao He, He Wen, Shuchang Zhou, Yuxin Wu, Cong Yao, Xinyu Zhou, and Yuheng Zou. Effective quantization methods for recurrent neural networks. *CoRR*, abs/1611.10176, 2016.
- [22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.
- [23] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *J. Mach. Learn. Res.*, 18(1):6869–6898, January 2017.
- [24] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, May 1994.
- [25] Li Jing, Yichen Shen, Tena Dubcek, John Peurifoy, Scott A. Skirlo, Max Tegmark, and Marin Soljacic. Tunable efficient unitary neural networks (EUNN) and their application to RNN. *CoRR*, abs/1612.05231, 2016.
- [26] Cijo Jose, Moustapha Cissé, and François Fleuret. Kronecker recurrent units. *CoRR*, abs/1705.10142, 2017.
- [27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [28] Oleksii Kuchaiev and Boris Ginsburg. Factorization tricks for LSTM networks. *CoRR*, abs/1703.10722, 2017.
- [29] Ashish Kumar, Saurabh Goyal, and Manik Varma. Resource-efficient machine learning in 2 kb ram for the internet of things. May 2017.
- [30] Aditya Kusupati, Manish Singh, Kush Bhatia, Ashish Kumar, Prateek Jain, and Manik Varma. Fastgrnn: A fast, accurate, stable and tiny kilobyte sized gated recurrent neural network. *CoRR*, abs/1901.02358, 2019.
- [31] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, and V. Lempitsky. Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition. *arXiv e-prints*, December 2014.
- [32] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

- [33] Zhouhan Lin, Matthieu Courbariaux, Roland Memisevic, and Yoshua Bengio. Neural networks with few multiplications. *ArXiv e-prints*, abs/1510.03009, October 2015.
- [34] Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization, 2017.
- [35] Zakaria Mhammedi, Andrew D. Hellicar, Ashfaqur Rahman, and James Bailey. Efficient orthogonal parametrisation of recurrent neural networks using householder reflections. *CoRR*, abs/1612.00188, 2016.
- [36] Sharan Narang, Eric Undersander, and Gregory F. Diamos. Block-sparse recurrent neural networks. *CoRR*, abs/1711.02782, 2017.
- [37] Kirill Neklyudov, Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Structured bayesian pruning via log-normal multiplicative noise, 2017.
- [38] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 2016.
- [39] D. Roggen, A. Calatroni, M. Rossi, T. Holleczeck, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkel, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. d. R. Millán. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh International Conference on Networked Sensing Systems (INSS)*, pages 233–240, June 2010.
- [40] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November 1997.
- [41] Vikas Sindhwani, Tara Sainath, and Sanjiv Kumar. Structured transforms for small-footprint deep learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3088–3096. Curran Associates, Inc., 2015.
- [42] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi. Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, 11(3):812–820, June 2015.
- [43] Urmish Thakker, Ganesh Dasika, Jesse G. Beu, and Matthew Mattina. Measuring scheduling efficiency of rnns for NLP applications. *CoRR*, abs/1904.03302, 2019.
- [44] Anna Thomas, Albert Gu, Tri Dao, Atri Rudra, and Christopher Ré. Learning compressed transforms with low displacement rank. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9066–9078. Curran Associates, Inc., 2018.
- [45] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Compressing recurrent neural network with tensor train. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 4451–4458. IEEE, 2017.
- [46] Lloyd Trefethen and David Bau. *Numerical Linear Algebra*. SIAM: Society for Industrial and Applied Mathematics, 1997.
- [47] Vincent Vanhoucke, Andrew Senior, and Mark Z. Mao. Improving the speed of neural networks on cpus. In *Deep Learning and Unsupervised Feature Learning Workshop, NIPS 2011*, 2011.
- [48] Shuo Wang, Zhe Li, Caiwen Ding, Bo Yuan, Qinru Qiu, Yanzhi Wang, and Yun Liang. C-lstm: Enabling efficient lstm using structured compression techniques on fpgas. In *Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, FPGA ’18*, pages 11–20, New York, NY, USA, 2018. ACM.
- [49] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209, 2018.
- [50] Wikipedia contributors. Kronecker product — Wikipedia, the free encyclopedia, 2019. [Online; accessed 19-May-2019].

- [51] Scott Wisdom, Thomas Powers, John R. Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks, 2016.
- [52] J. Wu. Compression of fully-connected layer in neural network by kronecker product. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*, pages 173–179, Feb 2016.
- [53] Jiong Zhang, Qi Lei, and Inderjit S. Dhillon. Stabilizing gradients for deep neural networks via efficient SVD parameterization. *CoRR*, abs/1803.09327, 2018.
- [54] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. Hello edge: Keyword spotting on microcontrollers. *CoRR*, abs/1711.07128, 2017.
- [55] Chenzhuo Zhu, Song Han, Huizi Mao, and William J. Dally. Trained ternary quantization. *CoRR*, abs/1612.01064, 2016.
- [56] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv e-prints*, page arXiv:1710.01878, October 2017.

Appendix A Background

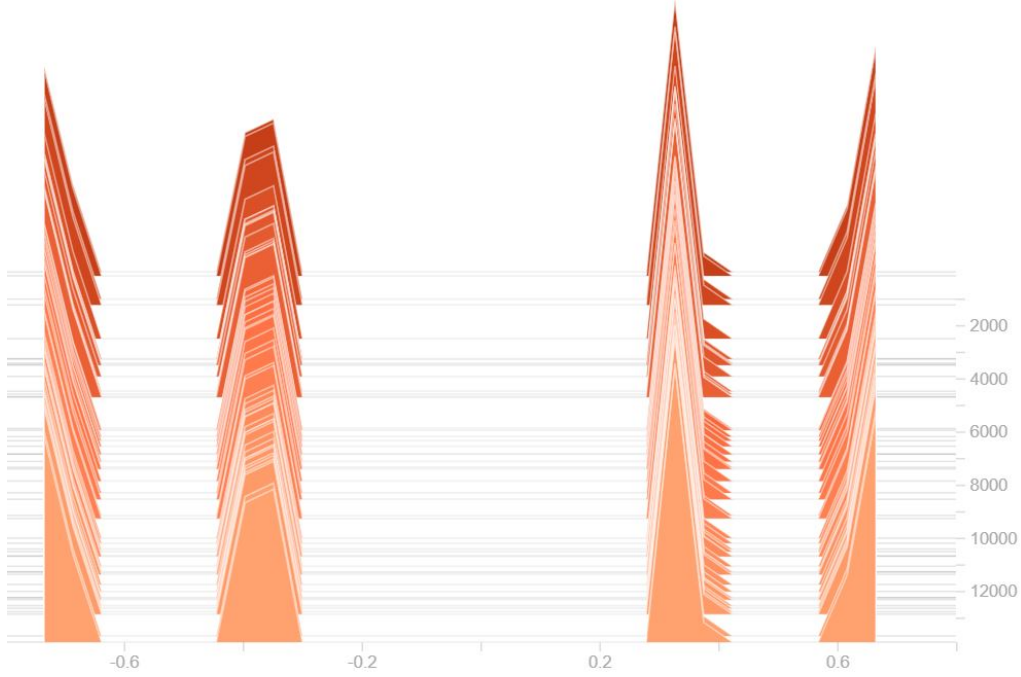


Figure 4: The values of a 2x2 matrix across multiple epochs

We tried using the framework provided by [26] to compress the GRU matrix in the small GRU baseline in [54] by a factor of 2. We used a GRU with hidden vector size of 256 and replaced the hidden-hidden matrix with Kronecker product of 8, 2x2 matrices as described in [26]. The resultant network lost 3% accuracy. On inspecting the 2x2 matrices, we found that the matrices hardly changed after initialization. Figure 4 shows the values of a 2x2 matrix across multiple epochs during training. We see the values in the matrix do not change after initialization.

Appendix B Dataset details and baseline implementation

B.1 Datasets

We evaluate the impact of compression using the techniques discussed in section 3 and 3.3 on a wide variety of benchmarks spanning applications like key-word spotting, human activity recognition, image classification and language modeling.

- **Human Activity Recognition:** We use the [39] dataset for human activity recognition. We split the benchmark into training, validation and test dataset using the procedure described in [18]. They use a subset of 77 sensors from the dataset. They use run 2 from subject 1 as their validation set, and replicate the most popular recognition challenge by using runs 4 and 5 from subject 2 and 3 in the test set. The remaining data is used for training. For frame-by-frame analysis, they created sliding windows of duration 1 second and 50% overlap leading to input vector of size 81×77 i.e. 81 dimensional input is fed to the network over 77 time steps. The resulting training-set contains approx. 650k samples (43k frames).
- **Image Classification:** We use the MNIST [32] and USPS [24] dataset for image classification. The USPS dataset consists of 7291 train and 2007 test images while the MNIST dataset consists of 60k training and 10k test images. We split the publicly available training set into 80% training set and 20% validation set and use the selected set of hyperparameters on the test set.

- **Key-word Spotting:** We use the [49] dataset for key-word spotting. The entire dataset consists of 65K different samples of 1-second long audio clips of 30 keywords, collected from thousands of people. We split the benchmark into training, validation and test dataset using the procedure described in [54].

B.2 Data Pre-processing

For the key-word spotting benchmarks, we reuse the framework provided by [54]. Thus we pre-process the data as suggested by them. For the human activity recognition dataset, we follow the pre-processing procedure described in [18]. We reuse the framework provided by [30] for the USPS dataset, thus using the pre-processing procedure provided by them.

B.3 Baseline Algorithms and Implementation

- **MNIST:** For this benchmark, the 28×28 image is fed to a single layer LSTM network with hidden vector of size 40 over 28 time steps. The dataset is fed using a batch size of 128 and the model is trained for 3000 epochs using a learning rate of 0.001. We use the Adam Optimizer [27] during training. Additionally, we divide the learning rate by 10 after every 1000 epochs. The total size of the network is 44.72 KB.
- **HAR1:** We use the network described in [18]. Their network uses a bidirectional LSTM with hidden length of size 179 followed by a softmax layer to get an accuracy of 92.5%. Input is of dimension 77 and is fed over 81 time steps. The paper uses gradient clipping regularization with a max norm value of 2.3 and a dropout of value 0.92 for both directions of the LSTM network. The network is trained for 300 epochs using a learning rate of 0.025, Adam optimization [27] and a batch size of 64. We used their training infrastructure and recreated the network in tensorflow. The suggested hyperparameters in the paper got us an accuracy of 91.9%. Even after significant effort, we were not able to get the accuracy mentioned in the paper. Henceforth, we will use 91.9% as the baseline accuracy. The total size of the network is 1462.836 KB.
- **KWS-LSTM:** For our baseline Basic LSTM network, we use the smallest LSTM model in [54]. The input to the network is 10 MFCC features fed over 25 time steps. The LSTM architecture uses a hidden length of size 118 and achieves an accuracy of 92.50%. We use a learning rate of 0.0005, 0.0001, 0.00002 for 10000 steps each with ADAM optimizer [27] and a batch size of 100. The total size of the network is 243.42 KB.
- **KWS-GRU:** We use the smallest GRU model in [54] as our baseline. The input to the network is 10 MFCC features fed over 25 time steps. The GRU architecture uses a hidden length of size 154 and achieves an accuracy of 93.50%. We use a learning rate of 0.0005, 0.0001, 0.00002 for 10000 steps each with ADAM optimizer and a batch size of 100. The total size of the network is 305.03 KB.
- **USPS-FastRNN:** The input image of size 16×16 is divided into rows of size 16 that is fed into a single layer of FastRNN network [30] with hidden vector of size 32 over 16 time steps. The network is trained for 300 epochs using an initial learning rate of 0.01 and a batch size of 100. The learning starts rate declining by 0.1 after 200 epochs. The total size of the network is 7.54 KB.

Appendix C Kronecker Products - Implementation

Algorithm 3 Implementation of Kronecker Products in Tensorflow

Input: Matrices B of dimension $m1 \times n1$, C of dimension $m2 \times n2$

Output: Matrix A of dimension $m \times n$

```

1:  $b\_shape = [B.shape[0].value, B.shape[1].value]$ 
2:  $c\_shape = [C.shape[0].value, C.shape[1].value]$ 
3:  $temp1 = tf.reshape(B, [b\_shape[0], 1, b\_shape[1], 1])$ 
4:  $temp2 = tf.reshape(C, [1, c\_shape[0], 1, c\_shape[1]])$ 
5:  $A = tf.reshape(temp1 * temp2, [b\_shape[0] * c\_shape[0], b\_shape[1] * c\_shape[1]])$ 

```

Algorithm 4 Finding dimension of Kronecker Factors for a matrix of dimension $m \times n$

Input: *list1* is the sorted list of prime factors of m , *list2* is the sorted list of prime factors of n

Output: *listA* - Dimension of the first Kronecker factor. *listB* - Dimension of the second Kronecker factor

```

1: while (len(list1) > 2)
2:   temp1 = list1[0]
3:   list1.del(0) //Delete the element at position zero
4:   list1[0] = list1[0]*temp1
5:   list1.sort('ascending')
6: while (len(list2) > 2)
7:   temp1 = list2[0]
8:   list2.del(0) //Delete the element at position zero
9:   list2[0] = list2[0]*temp1
10:  list2.sort('ascending')
11: list1 = list1.sort('descending')
12: listA.add(list1[0])
13: listA.add(list2[0])
14: listB.add(list1[1])
15: listB.add(list2[1])

```

C.1 Proof of the Matrix-Vector Multiplication Algorithm when the Matrix is expressed as a Kronecker product of two matrices

Let,

$$y = (A \otimes B) \times x \quad (11)$$

where, $y \in \mathbb{R}^{m \times 1}$, $x \in \mathbb{R}^{n \times 1}$, $A \in \mathbb{R}^{m1 \times n1}$, $B \in \mathbb{R}^{m2 \times n2}$ and $m = m1 \times m2$, $n = n1 \times n2$.

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_{n1} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ \vdots \\ \vdots \\ y_{m1} \end{pmatrix}$$

where $x_i \in \mathbb{R}^{n2}$ and $y_i \in \mathbb{R}^{m2}$ Then,

$$y = (A \otimes B) \times x \quad (12)$$

$$y = \begin{bmatrix} a_{1,1}B & a_{1,2}B & \dots & a_{1,n1}B \\ a_{2,1}B & a_{2,2}B & \dots & a_{2,n1}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1,1}B & a_{m1,2}B & \dots & a_{m1,n1}B \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ x_{n1} \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ \vdots \\ y_{m1} \end{bmatrix} = \begin{bmatrix} a_{1,1}Bx_1 + a_{1,2}Bx_2 + \dots + a_{1,n1}Bx_{n1} \\ a_{2,1}Bx_1 + a_{2,2}Bx_2 + \dots + a_{2,n1}Bx_{n1} \\ \vdots \\ a_{m1,1}Bx_1 + a_{m1,2}Bx_2 + \dots + a_{m1,n1}Bx_{n1} \end{bmatrix}$$

Each y_i has the following form -

$$[a_{i,1}Bx_1 + a_{i,2}Bx_2 + \dots + a_{i,n1}Bx_{n1}]$$

$$= B \begin{bmatrix} x_1 & x_2 & \cdot & \cdot & \cdot & x_{n1} \end{bmatrix} \begin{bmatrix} a_{i,1} \\ a_{i,2} \\ \cdot \\ \cdot \\ a_{i,n1} \end{bmatrix}$$

Now, let

$$X = \begin{bmatrix} x_1 & x_2 & \cdot & \cdot & x_{n1} \end{bmatrix}$$

And,

$$\mathbf{a}_i = \begin{bmatrix} a_{i,1} & a_{i,2} & \cdot & \cdot & a_{i,n1} \end{bmatrix}^T$$

Then,

$$y_i = BX\mathbf{a}_i \text{ (for } i = 1, 2, \dots, m1 \text{)} \quad (13)$$

Let \mathbf{Y} be a concatenation of y_i Thus,

$$\begin{aligned} \mathbf{Y} &= \begin{bmatrix} y_1 & y_2 & \cdot & \cdot & y_{m1} \end{bmatrix} \\ \mathbf{Y} &= \begin{bmatrix} BX\mathbf{a}_1 & BX\mathbf{a}_2 & \cdot & \cdot & BX\mathbf{a}_{m1} \end{bmatrix} \\ \mathbf{Y} &= BX \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdot & \cdot & \mathbf{a}_{m1} \end{bmatrix} \\ \mathbf{Y} &= BXA^T \end{aligned}$$

Appendix D KPRNN - Additional Details

D.1 Hyperparameters

Algorithm 5 LRD1: Learning rate decay function

Input: curr_learning_rate, decay_rate, global_step, decay_steps

Output: new_learning_rate

- 1: $temp1 = global_step / decay_steps$
 - 2: $pow = decay_rate^{temp1}$
 - 3: $new_learning_rate = curr_learning_rate * pow$
-

D.1.1 MNIST-LSTM compressed using KPLSTM Cells

Hyperparameters: Table 2 shows the hyperparameters used for training the MNIST-LSTM baseline and the MNIST network compressed using pruning, LMF, KPLSTM and a smaller baseline with the number of parameters equivalent to the compressed network.

Mean and Std Deviation of the accuracy of the compressed network: Last three rows of Table 2 show the top test accuracy, mean test accuracy and standard deviation of test accuracy of the networks trained using top two sets of best performing hyper-parameters on a held out validation set.

Hyperparameter values explored: We explored a broad range of hyper-parameter that were the intersection of the following values -

- Initial Learning Rate - 0.01 to 0.001 in multiples of 3
- LR Decay Schedule - We experimented with a step function and exponential decay function as described in algorithm 5.

D.1.2 HAR1 compressed using KPLSTM Cells

Hyperparameters: Table 3 shows the hyperparameters used for training the HAR1 baseline and the HAR1 network compressed using pruning, LMF, KPLSTM and a smaller baseline with the number of parameters equivalent to the compressed network.

Mean and Std Deviation of the accuracy of the compressed network: Last three rows of Table 3 show the top test accuracy, mean test accuracy and standard deviation of test accuracy of the networks trained using top two sets of best performing hyper-parameters on a held out validation set.

Hyperparameter values explored: We explored a broad range of hyper-parameter that were the intersection of the following values -

- Initial Learning Rate - 0.0025 to 0.25 in multiples of 3
- Max Norm - 1, 1.5, 2.3 and 3.5
- Dropout - 0.3, 0.5, 0.7 and 0.9
- #Epochs - 200 to 400 in increments of 100 for all networks apart from pruning. For pruned networks, we increased the number of epochs to 600
- LR Decay Schedule - We experimented with a step function and exponential decay function as described in algorithm 5.
- Pruning parameters - We explored various pruning start_epoch and end_epoch. We looked at starting pruning after 25% to 33% of the total epochs in increments of 4% and ending pruning at 75% to 83% of the total epochs in increments of 4%

D.1.3 KWS-LSTM compressed using KPLSTM Cells

Hyperparameters: Table 4 shows the hyperparameters used for training the HAR1 baseline and the HAR1 network compressed using pruning, LMF, KPLSTM and a smaller baseline with the number of parameters equivalent to the compressed network.

Mean and Std Deviation of the accuracy of the compressed network: Last three rows of Table 4 show the top test accuracy, mean test accuracy and standard deviation of test accuracy of the networks trained using top two sets of best performing hyper-parameters on a held out validation set.

Hyperparameter values explored: We explored a broad range of hyper-parameter that were the intersection of the following values -

- Initial Learning Rate - 0.001 to 0.1 in multiples of 10
- #Epochs - We trained the network for 30k-100k epochs with increments of 10k
- LR Decay Schedule - We experimented with a step function and exponential decay function as described in algorithm 5. For the step function we decremented the learning rate by 10 after every 10k, 20k or 30k steps depending on the improvement in held out validation accuracy. For the LRD1 algorithm, we tried decay_rate values of 0.03 to 0.09 in increments of 0.02.
- Pruning parameters - We explored various pruning start_epoch and end_epoch. We looked at starting pruning after 10k to 25k in increments of 5k and ending pruning at 60k to 90k in increments of 10k

D.1.4 KWS-GRU compressed using KPGRU Cells

Hyperparameters: Table 5 shows the hyperparameters used for training the HAR1 baseline and the HAR1 network compressed using pruning, LMF, KPGRU and a smaller baseline with the number of parameters equivalent to the compressed network.

Mean and Std Deviation of the accuracy of the compressed network: Last three rows of Table 5 show the top test accuracy, mean test accuracy and standard deviation of test accuracy of the networks trained using top two sets of best performing hyper-parameters on a held out validation set.

Hyperparameter values explored: We explored a broad range of hyper-parameter that were the intersection of the following values -

- Initial Learning Rate - 0.001 to 0.1 in multiples of 10
- #Epochs - We trained the network for 30k-100k epochs with increments of 10k
- LR Decay Schedule - We experimented with a step function and exponential decay function as described in algorithm 5. For the step function we decremented the learning rate by 10

after every 10k, 20k or 30k steps depending on the improvement in held out validation accuracy. For the LRD1 algorithm, we tried decay_rate values of 0.03 to 0.09 in increments of 0.02.

- Pruning parameters - We explored various pruning start_epoch and end_epoch. We looked at starting pruning after 10k to 25k in increments of 5k and ending pruning at 60k to 90k in increments of 10k

D.1.5 USPS-FastRNN compressed using KPFastRNN Cells

Hyperparameters: Table 6 shows the hyperparameters used for training the HAR1 baseline and the HAR1 network compressed using pruning, LMF, KPFastRNN and a smaller baseline with the number of parameters equivalent to the compressed network.

Mean and Std Deviation of the accuracy of the compressed network: Last three rows of Table 6 show the top test accuracy, mean test accuracy and standard deviation of test accuracy of the networks trained using top two sets of best performing hyper-parameters on a held out validation set.

Hyperparameter values explored: We explored a broad range of hyper-parameter that were the intersection of the following values -

- Initial Learning Rate - 0.01 to 0.001 in multiples of 3
- LR Decay Schedule - We experimented with a step function and exponential decay function as described in algorithm 5.

Table 2: Hyperparameters for MNIST baseline network, network compressed using KPLSTM and equivalent sized networks compressed using LMF, Pruning and Small Baseline. LRD1 refers to Algorithm 5.

Network	Baseline	Small Baseline	Pruning	LMF	KPLSTM
Batch Size	128	128	128	128	128
Optimizer	Adam	Adam	Adam	Adam	Adam
Weight Init	glorot_uniform				
#Epochs	3000	4000	10000	5000	5000
Initial LR	0.001	0.001	0.001	0.001	0.001
Decay Schedule	LR is divided by 10 after every #Epochs÷4 epochs				
Additional Details					
#Layers	1	1	1	1	1
Hidden Vector Size	40	40	40	40	40
Size of input	28	28	28	28	28
#Time Steps	28	28	28	28	28
Size (KB) for 32 bit weights	44.73	4.51	4.19	4.9	4.05
Mean Accuracy	-	87.20	96.49	97.24	98.28
Top Accuracy	99.40	87.50	96.81	97.40	98.44
Std Dev (Accuracy)	-	0.27	0.30	0.13	0.12
Runtime (ms)	6.4	0.8	0.66	1.8	5.6

Table 3: Hyperparameters for HAR1 baseline network, network compressed using KPLSTM and equivalent sized networks compressed using LMF, Pruning and Small Baseline. LRD1 refers to Algorithm 5.

Network	Baseline	Small Baseline	Pruning	LMF	KPLSTM
Optimizer	Adam	Adam	Adam	Adam	Adam
Batch Size	64	64	64	64	64
Weight Init	glorot_uniform				
#Epochs	300	300	600	300	300
Initial LR	0.025	0.025	0.025	0.025	0.025
Decay Schedule	LR reduced by a factor of 10 after every 100 epochs				
MaxNorm	2.3	2.3	2.3	3.5	2.3
Dropout	0.92	0.92	0.7	0.5	0.5
#Layers	1	1	1	1	1
Hidden Vector Size	179	179	179	179	178
Size of Input	77	77	77	77	77
#Time Steps	81	81	81	81	81
Additional Details			Pruning starts at Epoch #100 and ends at Epoch #500		
Size (KB) for 32-bit weights	1462.84	75.90	75.55	76.40	74.91
Mean Accuracy	-	88.39	89.63	89.63	90.95
Top Accuracy	91.90	88.84	89.94	89.94	91.14
Std Dev (Accuracy)	-	0.49	0.41	0.23	0.14
Runtime (ms)	470	29.92	98.2	64.12	187

Table 4: Hyperparameters for KWS-LSTM baseline network, network compressed using KPLSTM and equivalent sized networks compressed using LMF, Pruning and Small Baseline. LRD1 refers to Algorithm 5.

Network	Baseline	Small Baseline	Pruning	LMF	KPLSTM
Batch Size	100	100	100	100	100
Optimizer	Adam	Adam	Adam	Adam	Adam
Weight Init	glorot_uniform				
#Epochs	30k	80k	100k	80k	90k
Initial LR	5×10^{-4}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Decay Schedule	5×10^{-4} , 1×10^{-4} , 2×10^{-5} for 10k steps each	LRD1 with decay_rate 0.09	LRD1 with decay_rate 0.09	10^{-2} , 10^{-3} , 5×10^{-4} , 10^{-4} , 10^{-4} for 10k, 20k, 15k, 10k, 15k and 10k epochs each	LRD1 with decay_rate 0.09
#Layers	1	1	1	1	1
Hidden Vector Size	118	118	118	118	118
Size of Input	10	10	10	10	10
#Time Steps	25	25	25	25	25
Additional Details			Pruning starts at Epoch #15k and ends at Epoch #80k		
Size (KB) for 32 bit weights	243.42	15.66	15.57	16.80	15.30
Mean Accuracy	-	88.57	82.51	88.94	91.12
Top Accuracy	92.50	89.70	84.91	89.13	91.20
Std Dev (Accuracy)	-	0.67	2.72	0.16	0.07
Runtime (ms)	26.8	2.01	5.89	4.14	17.5

Table 5: Hyperparameters for KWS-GRU baseline network, network compressed using KPGRU and equivalent sized networks compressed using LMF, Pruning and Small Baseline. LRD1 refers to Algorithm 5.

Network	Baseline	Small Baseline (1L)	Small Baseline (2L)	LMF (1L)	LMF (2L)	KPGRU
Batch Size	100	100	100	100	100	100
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
Weight Init	glorot_uniform					
#Epochs	30k	70k	70k	90k	70k	90k
Initial LR	5×10^{-3}	10^{-2}	10^{-2}	10^{-2}	10^{-2}	0.01
Decay Schedule	5×10^{-3} , 10^{-3} , 2×10^{-4} for 10k steps each	10^{-2} , 5×10^{-3} , 10^{-3} , 2×10^{-4} for 15k, 15k, 15k and 10k epochs each	10^{-2} , 5×10^{-3} , 10^{-3} , 2×10^{-4} for 20k, 15k, 10k, 15k and 10k epochs each	10^{-2} , 5×10^{-3} , 10^{-3} , 10^{-4} for 20k, 30k, 20k and 20k epochs each	10^{-2} , 5×10^{-3} , 10^{-3} , 10^{-4} for 20k, 15k, 10k, 15k and 10k epochs each	LRD1 with decay_rate 0.01
Additional Details						
#Layers	1	1	2	1	2	2
Hidden Vector Size	154	154	154	154	154	154
Size of Input	10	10	10	10	10	10
#Time Steps	25	25	25	25	25	25
Size (KB) for 32 bit weights	305.04	22.63	22.27	24.50	25.47	22.23
Mean Accuracy	-	85.76	82.71	90.39	87.10	92.03
Top Accuracy	93.50	86.40	84.53	90.88	87.70	92.30
Std Deviation (Accuracy)	-	0.52	1.22	0.44	0.44	0.22
Runtime (ms)	67	6	10.13	7.16	11.1	34

Table 6: Hyperparameters for USPS-FastRNN baseline network, network compressed using KPGRU and equivalent sized networks compressed using LMF, Pruning and Small Baseline. LRD1 refers to Algorithm 5.

Network	Baseline	Small Baseline	Pruning	LMF	KPFastRNN
Batch Size	100	100	100	100	100
Optimizer	Adam	Adam	Adam	Adam	Adam
Weight Init	random_normal				
#Epochs	300	400	500	500	500
Initial LR	0.01	0.01	0.01	0.01	0.01
Decay Schedule	Learning Rate declines by 10 after 200th epoch				
#Layers	1	1	1	1	1
Hidden Vector Size	32	32	32	32	32
Size of Input	16	16	16	16	16
#Time Steps	16	16	16	16	16
Additional Details					
Size (KB) for 32 bit weights	7.25	1.98	1.92	2.05	1.63
Mean Accuracy	-	91.13	86.57	89.39	93.16
Top Accuracy	92.50	91.23	88.52	89.56	93.20
Std Dev (Accuracy)	-	0.07	1.52	0.14	0.03
Runtime (ms)	1.175	0.4	0.375	0.283	0.6

D.2 Quantization

	HAR1			
	32-bit	Size (KB)	8-bit	Size (KB)
Baseline	91.90	1462.84	91.13	384.64
KPLSTM Compressed Network	91.14	74.91	90.90	28.22
	KWS-LSTM			
	32-bit	Size (KB)	8-bit	Size (KB)
Baseline	92.50	243.42	92.02	65.04
KPLSTM Compressed Network	91.20	15.30	91.04	8.01

Table 7: Accuracy of baseline HAR1, baseline KWS-LSTM, KPLSTM-HAR1 and KPLSTM-KWS network after quantization to 8-bits.

Quantization [21, 47] is another popular technique for compressing neural networks. It is orthogonal to the compression techniques discussed previously; prior work has shown that pruning [19] can benefit from quantization. We do a study to test whether KPRNNs are compatible with quantization. We use the quantization flow provided by the authors of [21]. We quantized the LSTM cells in the

baseline and the KPRNN compressed networks to 8-bits floating point representations to test the robustness of KPRNNs under reduced bit-precision. Table 7 show that quantization works well with KPRNN. The HAR1 and KWS-LSTM networks compressed using KPRNN can be further compressed using quantization.

Appendix E HKPRNN - Additional Details

E.1 Hyperparameters

E.1.1 HAR1 compressed using HKPLSTM

Hyperparameters: Table 8 shows the hyperparameters used for training the HAR1 baseline and the HAR1 network compressed using pruning, LMF, HKPLSTM and a smaller baseline with the number of parameters equivalent to the compressed network.

Mean and Std Deviation of the accuracy of the compressed network: Last three rows of Table 8 show the top test accuracy, mean test accuracy and standard deviation of test accuracy of the networks trained using top two sets of best performing hyper-parameters on a held out validation set.

Hyperparameter values explored: We explored a broad range of hyper-parameter that were the intersection of the following values -

- Initial Learning Rate - 0.0025 to 0.25 in multiples of 3
- Max Norm - 1, 1.5, 2.3 and 3.5
- Dropout - 0.3, 0.5, 0.7 and 0.9
- #Epochs - 200 to 400 in increments of 100 for all networks apart from pruning. For pruned networks, we increased the number of epochs to 600
- LR Decay Schedule - We experimented with a step function and exponential decay function as described in algorithm 5.
- Pruning parameters - We explored various pruning start_epoch and end_epoch. We looked at starting pruning after 25% to 33% of the total epochs in increments of 4% and ending pruning at 75% to 83% of the total epochs in increments of 4%

E.1.2 KWS-LSTM compressed using HKPLSTM

Hyperparameters: Table 10,9 shows the hyperparameters used for training the KWS-LSTM baseline and the KWS-LSTM network compressed using pruning, LMF, HKPLSTM and a smaller baseline with the number of parameters equivalent to the compressed network.

Mean and Std Deviation of the accuracy of the compressed network: Last three rows of Table 9,10 show the top test accuracy, mean test accuracy and standard deviation of test accuracy of the networks trained using top two sets of best performing hyper-parameters on a held out validation set.

Hyperparameter values explored: We explored a broad range of hyper-parameter that were the intersection of the following values -

- Initial Learning Rate - 0.001 to 0.1 in multiples of 10
- #Epochs - We trained the network for 30k-100k epochs with increments of 10k
- LR Decay Schedule - We experimented with a step function and exponential decay function as described in algorithm 5. For the step function we decremented the learning rate by 10 after every 10k, 20k or 30k steps depending on the improvement in held out validation accuracy. For the LRD1 algorithm, we tried decay_rate values of 0.03 to 0.09 in increments of 0.02.
- Pruning parameters - We explored various pruning start_epoch and end_epoch. We looked at starting pruning after 10k to 25k in increments of 5k and ending pruning at 60k to 90k in increments of 10k

Table 8: Hyperparameters for HAR1 baseline network, network with LSTM layers compressed using HKPLSTM by a factor of 10 and equivalent sized networks compressed using LMF, Pruning and Small Baseline. LRD1 refers to Algorithm 5.

Network	Baseline	Small Baseline	Pruning	LMF	HKPLSTM
Batch Size	64	64	64	64	64
Optimizer	Adam	Adam	Adam	Adam	Adam
Weight Init	glorot_uniform				
#Epochs	300	200	300	300	300
Initial LR	0.025	0.025	0.025	0.025	0.025
Decay Schedule	LR reduced by a factor of 10 after every 100 epochs				
MaxNorm	2.3	2.3	3.5	2.3	2.3
Dropout	0.92	0.8	0.92	0.5	0.5
#Bidirectional Layers	1	1	1	1	1
Hidden Vector Size	179	179	179	179	179
Size of Input	77	77	77	77	77
#Time Steps	81	81	81	81	81
Additional Details			Pruning starts at Epoch #100 and ends at Epoch #250		
Size (KB) assuming 32 bit weights	1462.84	173.94	169	167.53	159.83
Mean Accuracy	-	89.95	86.56	90.61	91.025
Top Accuracy	91.90	90.30	87.20	90.80	91.20
Std Dev (Accuracy)	-	0.22	0.34	0.17	0.14
Runtime (ms)	470	63.42	174.92	87.94	234.67

Table 9: Hyperparameters for KWS-LSTM baseline network, network with LSTM layers compressed using HKPLSTM by a factor of 10 and equivalent sized networks compressed using LMF, Pruning and Small Baseline. LRD1 refers to Algorithm 5.

Network	Baseline	Small Baseline	Pruning	LMF	HKPLSTM
Batch Size	100	100	100	100	100
Optimizer	Adam	Adam	Adam	Adam	Adam
Weight Init	glorot_uniform				
#Epochs	30k	80k	100k	80k	90k
Initial LR	5×10^{-4}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Decay Schedule	5×10^{-4} , 1×10^{-4} , 2×10^{-5} for 10k steps each	LRD1 with decay_rate 0.09	LRD1 with decay_rate 0.09	LRD1 with decay_rate 0.09	LRD1 with decay_rate 0.09
#Layers	1	1	1	1	1
Hidden Vector Size	118	118	118	118	118
Size of Input	10	10	10	10	10
#Time Steps	25	25	25	25	25
Additional Details			Pruning starts at Epoch #15k and ends at Epoch #80k		
Size (KB) assuming 32 bit weights	243.42	30.92	31.02	30.86	26.38
Mean Accuracy	-	88.69	87.25	91.26	91.66
Top Accuracy	92.50	89.80	87.49	91.40	91.75
Std Dev (Accuracy)	-	0.67	0.16	0.12	0.07
Runtime (ms)	26.8	3.2	11.26	6.99	14

Table 10: Hyperparameters for KWS-LSTM baseline network, network with LSTM layers compressed using HKPLSTM by a factor of 20 and equivalent sized networks compressed using LMF, Pruning and Small Baseline. LRD1 refers to Algorithm 5.

Network	Baseline	Small Baseline	Pruning	LMF	HKPLSTM
Batch Size	100	100	100	100	100
Optimizer	Adam	Adam	Adam	Adam	Adam
Weight Init	glorot_uniform				
#Epochs	30k	80k	100k	80k	90k
Initial LR	5×10^{-4}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Decay Schedule	5×10^{-4} , 1×10^{-4} , 2×10^{-5} for 10k steps each	LRD1 with decay_rate 0.09	LRD1 with decay_rate 0.09	LRD1 with decay_rate 0.09	LRD1 with decay_rate 0.09
#Layers	1	1	1	1	1
Hidden Vector Size	118	118	118	118	118
Size of Input	10	10	10	10	10
#Time Steps	25	25	25	25	25
Additional Details			Pruning starts at Epoch #15k and ends at Epoch #80k		
Size (KB) assuming 32 bit weights	243.42	17.34	17.9	16.8	16.76
Mean Accuracy	-		84.98	90.78	91.14
Top Accuracy	92.50	89.80	85.17	90.9	91.28
Std Dev (Accuracy)	-	0.58	0.16	0.11	0.11
Runtime (ms)	26.8	2.25	8	5.78	15