Grenze International Journal of Engineering and Technology, January Issue

The Role of Compression Models in Human Activity Recognition: A Comprehensive Review

Jaspreet Kaur[1] and Veenu Mangat[2]
[1-2]University Institute of Engineering and Technology (UIET) Panjab University Chandigarh, India
Email: jassimathoda@gmail.com, vmangat@pu.ac.in

***Abstract*— Human activity recognition has gained significant interest in a number of fields, such as wearable technologies, smart environments, and healthcare. However, the growing complexity of HAR models and the large amount of sensor and video data pose some issues related to storage, computation, and real-time processing. To address these challenges, compression techniques have been explored to reduce model size while persevering accuracy of recognition. This paper compares different compression techniques based on key performance metrics, including compression ratio, accuracy retention, inference time, and energy efficiency. In addition, we discuss how they affect real-time HAR applications, especially in contexts with limited resources and edge computing. Our research demonstrates the trade-offs between performance and compression, the applicability of different approaches for distinct HAR situations, and the challenges associated with computational overhead and information loss. In order to increase the effectiveness of HAR models, we highlight the need for adaptive and hybrid compression techniques in our final research directions. The goal of this review is to provide insight in creating HAR systems that are more efficient and scalable.**

***Index Terms*— Human activity recognition, machine learning, deep learning, Compression Techniques, edge computing.**

I. Introduction

An important field of study in artificial intelligence and machine learning is Human Activity Recognition (HAR), which focuses on recognizing and categorizing human actions using sensor or video data. To track and examine human motions, HAR systems use information from wearable sensors, smartphones, ambient sensors, and camera-based vision systems.it plays a critical role in many fields, including smart environments, fitness, healthcare, and security [1], [2]. The majority of early Human Activity Recognition (HAR) systems were vision-based, tracking human movement with cameras. Despite their effectiveness, these methods required a lot of processing power and raised privacy issues due to sensitive nature of video data. The emphasis changed to sensor-based HAR systems with the advent of inertial sensors like gyroscopes and accelerometers. These sensors provide a more energy-efficient and privacy-friendly option and these sensors commonly found in wearable and mobile devices. Sensor-based HAR uses machine learning methods as shown in figure [?] to analyse features extracted from unprocessed sensor data in order to identify activities. Although they were first widely used, traditional techniques like decision trees, k-nearest neighbours (KNN), and support vector machines (SVM) mainly relied on human feature engineering, which is time consuming and restricts scalability [3].

Grenze ID: 01.GIJET.12.1.47

In recent years, deep learning models that automatically learn characteristics from raw data, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have improved performance. But the Deployment of these models on low-power devices, such as wearables and IoT nodes, is complicated because they require high computational processing [4].

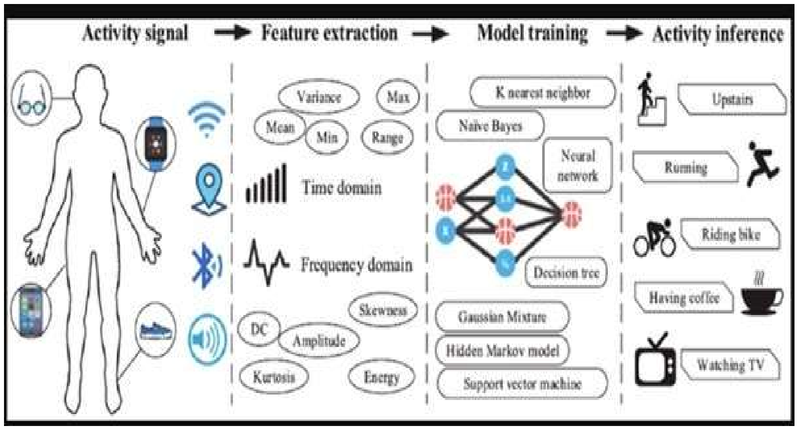

Fig. 1. Human Activity recognition Process

The ability to observe and interpret human actions can enhance many applications, ranging from real-time health monitoring to intelligent surveillance systems. As wearable technology and the Internet of Things (IoT) gain popularity, HAR has been developed to provide more precise and reliable insights on human behaviour. The growing use of HAR across a range of industries has increased the need for scalable, effective, and real-time solutions [5], [6]. Despite its advancements, HAR still has a number of issues that prevent widespread use, especially in settings with limited resource such as such as edge devices and Internet of Things(IoT) applications. Important difficulties include:

- Large volumes of data are produced by HAR from sensors and video streams, that requires a substantial amount of processing and storage capacity.
- Another challenge associated with HAR is energy consumption in Wearable and portable devices. Due to their high computational demands and frequent reliance on cloud-based processing, traditional machine learning models for HAR raise the requirement for wireless data transfer and, as a result, energy consumption. The batteries of wearable and mobile devices get depleted by constant data collection and model inference [7].
- Deep learning-based HAR models offer very accurate results but they are computationally costly and inappropriate for use on embedded or low-power systems [8].
- Maintaining accuracy across diverse environments and users is another major issue. As HAR models are trained on specific dataset and they behave poorly when exposed to variation in environmental factors, user behaviour, or sensor placement [9].

In order to overcome these difficulties, compression models can be used as they have ability to maintain recognition accuracy while lowering model size, memory footprint, and computational complexity. HAR's primary compression methods include:

Model Compression: Model compression is the process of lowering the memory and computational footprint of deep learning models by using methods like quantisation, pruning, and knowledge distillation. These techniques greatly reduce the model size and inference latency while preserving a respectable level of accuracy, which makes them ideal for deployment on edge devices with limited resources [9]. In applications like wearables, mobile health monitoring, and smart home automation, model compression makes real-time Human Activity Recognition (HAR) possible by eliminating redundancies and streamlining network architecture.

Data Compression: In data Compression methods raw input data (such as sensor signal or video frames) is reduced before that is fed into a machine learning model using approaches like dimensionality reduction (e.g. Principal Component Analysis [PCA] and autoencoders). These approaches extract the most useful features from high dimensional data, therefore reducing computational cost while preserving essential activity related patterns. Furthermore, to maximise storage and transmission, entropy coding techniques (like run-length encoding or Huffman coding) compress data at the bit level. These techniques are especially helpful in HAR systems, which need the fast processing of large volume of multimodal sensor data on edge or mobile devices [10].

Hybrid Approach: In order to increase the overall efficiency of HAR systems, hybrid approaches combining data and model compression techniques can be used. These strategies balance performance and resource utilization by concurrently optimising model architecture through quantisation, pruning, or knowledge distillation, and reducing the size of input data through dimensionality reduction or entropy coding. With the help of this dual level optimization HAR models can maintain excellent recognition accuracy while significantly lowering computing load, memory utilisation, and energy consumption. These hybrid techniques can be very beneficial for deployment in real-time, resource-constrained situations, such as wearable technology, smartphones, and IoT enabled smart homes [11].

HAR systems can achieve real-time processing, reduced energy consumption, and enhanced deployment practicality by incorporating compression techniques in mobile and embedded applications. The paper presents an overview of HAR and its challenges and explore different compression techniques on HAR system.

II. RELATED WORK

A lot of research has been done to increase the effectiveness of HAR systems, especially by compressing the data and models. In this section we review the literature that examine the effects of methods like dimensionality reduction, quantisation, pruning, and knowledge distillation on HAR performance. In this work, author used TensorFlow Lite compression techniques to investigate the effectiveness of deploying deep learning (DL) models for Human Activity Recognition (HAR) on devices with limited resources. On an ESP32 chip, the techniques were tested in cascade and standalone compression situations for CNNs, LSTMs, and CNN-LSTM hybrids. Although dynamic quantisation preserves accuracy with little loss in stand-alone mode, the results demonstrate that cascade compression is not practical for performance evaluation. Comparing dynamic and complete integer quantisation to uncompressed models, both show comparable reductions in inference delay and a significant reduction in energy consumption (up to 45%) [12]. HAR has attracted a lot of interest because of its many uses and the accessibility of time-series data from wearable sensors such as gyroscopes and accelerometers. Traditional HAR approaches is time-consuming because they require manual feature engineering. The author proposed a hybrid model that can independently analyse raw sensor data and attained 99. 74% accuracy on the WISDM data set and also addressed the issue of fitting these models into small devices by employing compression pruning and quantization techniques. This model had achieved very good results but had used basic compression techniques [13]. In this study, author provide C-HAR, a sub-Nyquist compressive sensing based lightweight human activity recognition (HAR) system. It has used information from a smartwatch's integrated gyroscope and accelerometer to categorise a variety of everyday tasks. C-HAR operates at one-fourth the sampling rate while achieving comparable accuracy to conventional techniques. Additionally, it performs better than current HAR systems based on compressive sensing since it allows multiclass classification at a lower sample rate. The feature extraction and classification parts of C-HAR operate effectively on microcontroller units (MCUs), providing over twice the speed of similar systems and doing away with the requirement to offload to external devices like servers or smartphones. One disadvantage of this system is that it causes addresses privacy concerns [14]. This study employs compressed deep learning models to recognise human movement and gestures. Three architectures—MLP, CNN-LSTM, and CNN-GRU—were used to classify five gestures: leap, run, walk, squat, and other. The initial accuracies were 88%, 88.1%, and 89%, respectively. Model size was reduced by a factor of 12 while classification accuracy increased to over 95% with the use of model compression techniques including pruning and quantisation. The author used basic techniques of compression some advanced techniques can also be used in future to further reduce the model size while maintain the accuracy [15]. This work introduces a human activity recognition (HAR) framework that has used stretch and accelerometer sensors for implementation on a wearable IoT device. The approach begins with a novel method for non-uniformly segmenting sensor data based on user motion. From these segments, features are extracted using Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT). These features are then utilized for both online inference and training via a neural network—making this the first solution to enable online training directly on the device. Experiments conducted on a TI-CC2650 microcontroller with data from nine users demonstrated a high accuracy of 97.7% in recognizing six activities and their transitions, while maintaining power consumption under 12.5 mW. In this work only basic activities used. For future more complex or multi step activities can also be used [16]. The focus of author [17] is on optimising models for the classification of human activities. This study evaluated the effects of pruning and quantisation techniques on accuracy and model size in existing HAR models. Based on the findings, the paper compares three models (CNN, LSTM, and Multi-layer LSTM) using the UCIHAR dataset. Models can be compressed up to ten times using either method without compromising accuracy. The multilayer LSTM did have one peculiarity, though, in that the model size was compressed up to 4.5 times after applying quantisation and optimised almost three times after pruning, unlike the other two models.

III. Experimental Work

This section describes the experimental approach used to assess the performance of several compression techniques on a hybrid CNN-LSTM model for Human Activity Recognition (HAR). The UCI HAR dataset was used for the studies, and the performance was assessed based on the accuracy of the model and the use of memory after compression.

A. Baseline Model

A hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM)architecture was used in the baseline HAR model. The CNN layers were responsible for extracting local spatial features from sensor signal windows. The extracted features were sent to a series of LSTM layers after CNN blocks. These recurrent layers were configured to learn long-term temporal dependencies and sequential relationships in the activity patterns. In order to capture transitions and recurrent behaviour in human motion, LSTM units have to maintain a memory of past states. The model was trained using categorical cross-entropy loss with the Adam optimizer over multiple epochs, and its performance was used as a benchmark for further compression techniques as shown in figure 2.

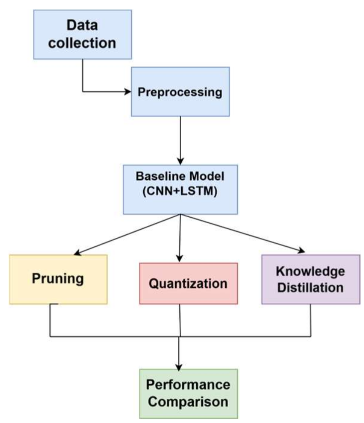

Fig. 2. Flowchart of Work

B. Compression Techniques

To evaluate the trade-off between accuracy and model efficiency different compression methods have been used. These techniques were applied individually to a baseline CNN-LSTM HAR model, the effects on accuracy and model size were thoroughly examined. The experiments aim is to identify which combination of methods will be feasible for deployment in real world applications, particularly those requiring wearable sensors and real-time decision-making. Three widely used compression techniques were implemented that are as follow:

Quantization: TensorFlow Lite was used after training to apply dynamic range quantization. This technique greatly reduced the model size without the need for retraining by converting the 32-bit floating-point weights to 8-bit integers.

Pruning: TensorFlow Model Optimisation Toolkit was used to implement magnitude-based weight pruning. The pruning schedule successfully eliminated less important weights while maintaining model accuracy by progressively raising sparsity from 20% to 80% over a predetermined number of training steps.

Knowledge distillation: The baseline CNN-LSTM model’s output (teacher) was replicated by training a smaller student model. The training phase employed the teacher model’s soft goal labels to effectively transmit knowledge while preserving the generalisability of the model.

IV. Evaluation Metrics

To evaluate the performance of baseline model and compressed HAR model two key metrics were used: Accuracy and model size. These metrics were used to demonstrate the models’ predictive power as well as their viability for deployment on edge devices with limited recourses.

Accuracy: Accuracy is a key parameter used to evaluate the performance of a classification model. It calculates the percentage of true predictions the model made out of total number of predictions [18]. The formula for accuracy is:

$$Accuracy = \frac{TP+TN}{P+N} \quad (1)$$

where TP is true positive computed as the sum of all true positive instances across the dataset and is given by

$$TP = \sum_{k=1}^{n} TP_k \quad (2)$$

similarly, TN represents the addition of all true negative samples and given by

$$TN = \sum_{k=1}^{n} TN_k \quad (3)$$

for misclassification False positive (FP) denotes the addition of false positive and given by

$$FP = \sum_{k=1}^{n} FP_k \quad (4)$$

And,

$$FN = \sum_{k=1}^{n} FN_k \quad (5)$$

Model Size: Model size refers to the total amount of memory required to store a trained machine learning model. It is typically measured in kilobytes (KB) or megabytes (MB), depending on the scale of the model.
It is given by

$$Modelsize(KB) = \frac{Total\ Number\ of\ Parameters*Precision(inbits)}{8*1024} \quad (6)$$

V. RESULT

In this work we applied different compression model on baseline model. These findings are shown in table 1 that quantisation is appropriate for deployment in extremely resource-constrained contexts since it produces the largest model size reduction with the least amount of accuracy loss. Pruning produced a balanced size reduction while maintaining performance close to baseline. Real-time applications were favoured by the knowledge distillation strategy, which produced the smallest model with intermediate accuracy and rapid inference. Figure3 shows the performance of different compression techniques when that are applied to HAR model. Baseline model achieves highest accuracy, the distilled and quantized models significantly reduced complexity while maintaining competitive accuracy.

TABLE I: COMPARISON OF MODEL PERFORMANCE METRICS

Model Type	Accuracy (%)	Model Size (KB)
Baseline	93	250
Quantized	91	120
Pruned	92	90
Distilled	90	70

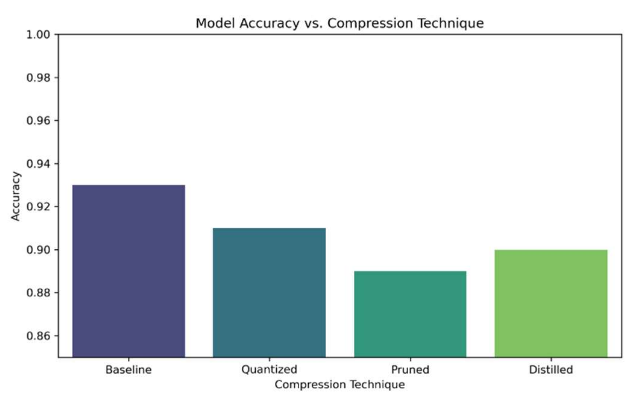

Figure 3. Accuracy comparison of different compression techniques

Figure 4 highlights model size with respect to compression techniques applied and it shows that quantization has achieved smallest model size.

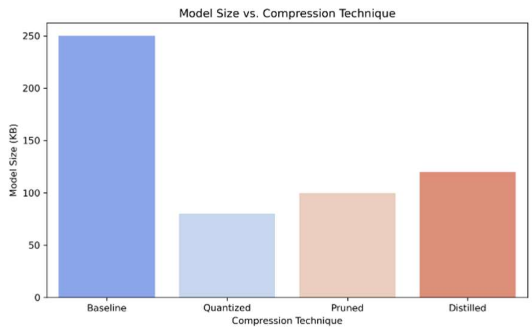

Figure 4. Model size comparison of different compression techniques

Figure 5 highlights trade-off between accuracy and model size and shows that quantization has achieved the smallest model size with the least amount of accuracy loss, while distillation provides a good balance between both metrics.

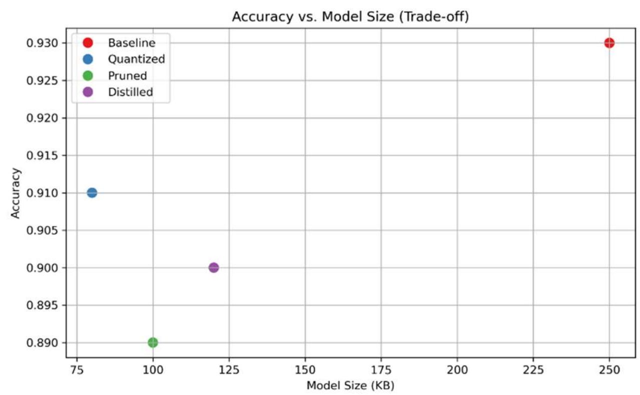

Figure 5. Trade-off between accuracy and model size

V. Conclusion

In recent decades, identifying human activity from sensor data has proven to be a very challenging research area. In this paper we have applied different compression techniques on baseline model of HAR. The analysis reveals a crucial trade-off between accuracy and model size. When it comes to deploying on devices with limited resources, quantisation works effectively because it can drastically reduce model size with no influence on performance. Pruning provides a moderate compression ratio with strong accuracy, making it perfect for situations where efficiency and accuracy must be balanced. Because knowledge distillation strikes a fair balance between precision and compactness, it is particularly well-suited for real-time applications. Finally, the particular requirements of the target environment should guide the compression strategy selection. Future studies should concentrate on adaptive or hybrid compression techniques to further maximise the use of HAR models in various real-world scenarios.

References

[1] Gupta, N., Gupta, S. K., Pathak, R. K., Jain, V., Rashidi, P., & Suri, J. S. (2022). Human activity recognition in artificial intelligence framework: a narrative review. Artificial intelligence review, 55(6), 4755-4808.

[2] Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., Alshurafa, N. (2022). Deep learning in human activity recognition with wearable sensors: A review on advances. Sensors, 22(4), 1476.

[3] Ramanujam, E., Perumal, T., Padmavathi, S. J. I. S. J. (2021). Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. IEEE Sensors Journal, 21(12), 1302913040.

[4] Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., & Alshurafa, N. (2022). Deep learning in human activity recognition with wearable sensors: A review on advances. Sensors, 22(4), 1476.
[5] Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artificial Intelligence Review, 54(3), 2259-2322.
[6] Mohsen, S., Elkaseer, A.,& Scholz, S. G. (2021, September). Human activity recognition using k-nearest neighbor machine learning algorithm. In Proceedings of the International Conference on Sustainable Design and Manufacturing (pp. 304-313). Singapore: Springer Singapore.
[7] Islam, M. M., Nooruddin, S., Karray, F., & Muhammad, G. (2022). Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects. Computers in biology and medicine, 149, 106060.
[8] H. Ren, D. Anicic, and T. A. Runkler, ”Tinyol: Tinyml with onlinelearning on microcontrollers,” in 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, Jul. 2021, pp. 1–8.
[9] C. Tanama, K. Peng, Z. Marinov, R. Stiefelhagen, and A. Roitberg, ”Quantized distillation: Optimizing driver activity recognition models for resource-constrained environments,” in 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, Oct. 2023, pp. 5479–5486.
[10] Contoli, C., & Lattanzi, E. (2023). A study on the application of tensorflow compression techniques to human activity recognition. IEEE Access, 11, 48046-48058.
[11] Rashid, N., Demirel, B. U., & Al Faruque, M. A. (2022). AHAR: Adaptive CNN for energy-efficient human activity recognition in lowpower edge devices. IEEE Internet of Things Journal, 9(15), 1304113051.
[12] Contoli, C., & Lattanzi, E. (2023). A study on the application of tensorflow compression techniques to human activity recognition. IEEE Access, 11, 48046-48058.
[13] Semwal, V. B., & Soni, V. (2024, February). Compressed deep learning model for human activity recognition. In 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-5). IEEE.
[14] Dawton, B., Ishida, S.,& Arakawa, Y. (2023, March). C-HAR: Compressive measurement-based human activity recognition. In 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) (pp. 601607). IEEE.
[15] Kumari, N., Yadagani, A., Behera, B., Semwal, V. B., & Mohanty, S. (2024). Human motion activity recognition and pattern analysis using compressed deep neural networks. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 12(1), 2331052.
[16] Bhat, G., Deb, R., Chaurasia, V. V., Shill, H., & Ogras, U. Y. (2018, November). Online human activity recognition using low-power wearable devices. In 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD) (pp. 1-8). IEEE.
[17] S. Gupta, S. Jain, B. Roy, and A. Deb, ”A TinyML approach to human activity recognition,” in Journal of Physics: Conference Series, vol. 2273, no. 1, p. 012025, May 2022.
[18] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, ”Introduction to machine learning, neural networks, and deep learning,” Translational Vision Science Technology, vol. 9, no. 2, pp. 14–14, 2020.