A hybrid deep neural network compression approach enabling edge intelligence for data anomaly detection in smart structural health monitoring systems

Tarutal Ghosh Mondal[a], Jau-Yu Chou[b], Yuguang Fu[c*], Jianxiao Mao[d]

[a]*Civil, Architectural and Environmental Engineering, Missouri University of Science and Technology, Rolla, 65401, MO, United States*
[b]*Department of Civil Engineering, National Taiwan University, Taipei, 10617, Taiwan*
[c]*School of Civil and Environmental Engineering, Nanyang Technological University, 639798, Singapore*
[d]*Key Laboratory of C&PC Structures of Ministry of Education, Southeast University, 211189, China*

(Received keep as blank , Revised keep as blank , Accepted keep as blank)

Abstract. This study explores an alternative to the existing centralized process for data anomaly detection in modern Internet of Things (IoT)-based structural health monitoring (SHM) systems. An edge intelligence framework is proposed for the early detection and classification of various data anomalies facilitating quality enhancement of acquired data before transmitting to a central system. State-of-the-art deep neural network pruning techniques are investigated and compared aiming to significantly reduce the network size so that it can run efficiently on resource-constrained edge devices such as wireless smart sensors. Further, depthwise separable convolution (DSC) is invoked, the integration of which with advanced structural pruning methods exhibited superior compression capability. Last but not least, quantization-aware training (QAT) is adopted for faster processing and lower memory and power consumption. The proposed edge intelligence framework will eventually lead to reduced network overload and latency. This will enable intelligent self-adaptation strategies to be employed to timely deal with a faulty sensor, minimizing the wasteful use of power, memory, and other resources in wireless smart sensors, increasing efficiency, and reducing maintenance costs for modern smart SHM systems. This study presents a theoretical foundation for the proposed framework, the validation of which through actual field trials is a scope for future work.

Keywords: deep neural network compression; edge intelligence; data anomaly detection; structural health monitoring; network pruning.

1. Introduction

1.1 Background

Future smart cities will see the deployment of a broad range of IoT-based wireless sensors for intelligent monitoring of civil infrastructure (Alavi et al. 2018, Fu et al. 2018, Bisio et al. 2022, Haque et al. 2020, Hou and Wu 2019, Fu et al. 2022, Mishra et al. 2022). However, sensors are the weak links in an IoT network. Any anomaly in the sensor data may degrade the system performance significantly. Therefore, early detection and isolation of various data anomalies are

*Corresponding author, Assistant Professor, E-mail: yuguang.fu@ntu.edu.sg

crucial to protect an IoT network from the adverse impact of the same (Fu et al. 2019, Peng et al. 2017). Detecting data anomalies is also important for improving the overall reliability of sensor data. Traditionally, data acquired by peripheral sensor nodes are sent to a central station, where all the incoming data are accumulated and processed en masse to remove any noises and anomalies (Chatterjee and Ahmed 2022, Cook et al. 2019). This process makes it difficult to isolate the anomalous data. Unnecessary storage and transmission of erroneous sensor data also lead to the misuse of energy and bandwidth. An alternate framework for data anomaly detection is presented by the latest advancements in the area of edge intelligence techniques. The sensor data can be processed on the edge leveraging state-of-the-art AI techniques (Al-amri et al. 2021, Peralta Abadia et al. 2022), and any anomaly in the data can be identified and cleansed before transmitting to a central station. This will lead to the quality enhancement of the acquired data. This will also reduce the amount of data sent to the central system mitigating the network overload, latency, and maintenance costs, and minimizing the wasteful use of energy and other resources.

The First International Project Competition for Structural Health Monitoring (IPC-SHM 2020) (Bao et al. 2021) gave rise to a number of studies focusing on advanced deep learning techniques for autonomous detection and classification of a variety of anomalies in structural condition assessment data (Bao et al. 2019, Chou et al. 2022, Du et al. 2022, Gao et al. 2022, G. Liu et al. 2022, Martakis et al. 2021, Shajihan et al. 2022, Xu et al. 2022, Yang et al. 2022). The success of this competition is testimony to the relevance, popularity, and appropriateness of this topic which is being investigated in this study. Chou et al. (2022) relied on GoogLeNet (Szegedy et al. 2015) for identifying anomalies in the time-history representations of accelerometer data. G. Liu et al. (2022), on the other hand, resorted to the ResNet-18 (K. He et al. 2016) architecture for distinguishing between different types of data anomalies. However, the state-of-the-art IoT end devices have limited capabilities to host such high-demand computing as entailed by the advanced techniques adopted in these studies. This highlights the need to develop an efficient edge intelligence framework that is computationally light and compliant with a resource-constrained IoT environment, which is the focus of this study.

1.2 Contribution

This study aims at bridging the prevailing gap between the resource requisitions of modern AI-based solutions for data anomaly detection and the computing capabilities of decentralized IoT edge devices. A relatively new and faster convolution approach called DSC (Chollet 2017) is invoked and integrated with advanced network pruning methods and QAT to formulate a hybrid three-pronged compression strategy that yields superior compression and acceleration ratios. It is appropriate to note here that pruning refers to a class of compression techniques that involves removing the redundant parts of a neural network. A set of four state-of-the-art pruning algorithms based on filter sketch (M. Lin, Cao, et al. 2021), adaptive exemplar filters (M. Lin, Ji,

et al. 2021), high-rank feature map (M. Lin et al. 2020), and generative adversarial learning (S. Lin et al. 2019), are probed in this regard and the best approach is identified by means of a comparative assessment. It's worth mentioning that although network pruning and quantization approaches have been around for quite some time now, their potential to expedite various structural health monitoring applications has not yet been fully exploited (Wu et al. 2019). The present study serves to fill this knowledge gap by being the first ever study to propose an efficient edge intelligence solution to the crucial data anomaly detection problem. Notwithstanding that the task of anomalous data detection is used as a case study, the proposed technique can however be extended to many other structural health monitoring applications with appropriate modifications.

1.3 Scope

The remainder of the paper is organized as follows. Section 2 summarizes various elements of the network compression strategy proposed in this study. Section 3.1 describes the source and preprocessing of the data used in this study. Section 3.2 introduces the CNN architecture which is used as a baseline to demonstrate the feasibility of the proposed framework. The results are presented and discussed in Section 3.4. Finally, the conclusions of this study are summarized in Section 4.

2. Deep Neural Network Compression

This section provides a synopsis of various network compression approaches investigated in this study. Pruning constitutes the core of various model compression techniques that seek to reduce the computational complexity of traditional deep learning-based methods. Previous studies have indicated that the traditional convolutional neural networks (CNN) often tend to be over-parameterized (Denil et al. 2013). This makes it important to get rid of the redundant parameters for the sake of attaining computational efficiency. Network pruning serves as a means to this end. Various pruning techniques available in the existing literature can be broadly classified into unstructured pruning and structured pruning. Unstructured pruning aims to remove the individual elements of the weight matrices (Ding et al. 2019, Frankle and Carbin 2018, Han et al. 2015, Hassibi and Stork 1992, LeCun et al. 1989). This sparsifies the remaining weights and leads to irregular memory access, necessitating specialized hardware (Han et al. 2016) and software (Park et al. 2016) accelerators to speed up the inference. Structured pruning, on the contrary, is more favorable to general-purpose inference platforms, as it removes a filter and the corresponding channel in its entirety, obliterating the need for specialized hardware or software support. This major advantage makes structured pruning more popular in the scientific community than the unstructured pruning approaches (Y. He et al. 2017, Hu et al. 2016, Huang and Wang 2018, Li et al. 2016, S. Lin et al. 2019, Z. Liu et al. 2019, Z. Liu et al. 2017, Luo et al. 2017, Singh et al. 2019, Wen et al. 2016, Yu et al. 2018, Zhao et al. 2019). Therefore, the latest

structured pruning techniques are taken up in this study for feasibility analyses, as described in the following sections. Apart from that, DSC and QAT are invoked to formulate a hybrid three-pronged compression strategy which affords better compressibility than any individual pruning method.

2.1 Network Pruning

Previous studies (Sun et al. 2017) indicated that the correlation between original pre-trained model weights and corresponding pruned model weights is maximized when the latter preserves the second-order covariance of the former. In mathematical terms:

$$\Sigma_{W^i} = \Sigma_{\Omega^i} \tag{1}$$

where, Σ_{W^i} and Σ_{Ω^i} represent the covariance matrices of the i-th layer filters in the original and pruned models, respectively. Based on this observation, M. Lin, Cao, et al. (2021) proposed a novel network pruning approach called FilterSketch (FS), which is free from expensive data-driven optimization. In this technique, a set of new parameters is learnt for each layer in the pruned model by minimizing the following objective function:

$$\arg - \min_{\Omega^i} \left\| \Sigma_{W^i} - \Sigma_{\Omega^i} \right\|_F \tag{2}$$

where, $\|\cdot\|_F$ denotes the Frobenius norm. These new parameters serve as initial estimates, which are subsequently fine-tuned to stack up against the original model performance. This technique formulates the task of information preservation as a matrix sketch problem, which is eventually solved using the Frequent Direction (Liberty 2013) method, leading to several order-of-magnitude compression in the network size.

Different filters in a CNN layer impact the network inference to various extents. The filters which do not contribute adequately induce informational redundancy, the removal of which is a key to attaining filter optimality. This is tantamount to identifying the most impacting filters, also known as the exemplars, from a large filter pool. However, the number of filters in different layers of a CNN are different. Therefore, it is of utmost importance to formulate an efficient but adaptive approach for identifying the most informative exemplars in a unified way that enables end-to-end fine-tuning without human intervention. The algorithm proposed by M. Lin, Ji, et al. (2021) achieves this objective by employing affinity propagation (Frey and Dueck 2007), which is a clustering technique based on the notion of message passing between data points. In contrast to many other popular clustering techniques such as k-means clustering, affinity propagation does not require any prior knowledge about the number of clusters in the data. This pruning approach is referred to in this study as Adaptive Exemplar Filters (AEF).

In another compression technique based on high-rank feature maps (HR), the identification of important filters is guided by the rank of feature maps (M. Lin et al. 2020). This approach is based on the deduction that the high-rank feature maps contain more useful information and are therefore more important to preserve accuracy. Low rank feature maps, on the other hand, contain less information, and can therefore be pruned to achieve network compression. It is observed that the average rank of feature maps generated by a filter does not change considerably as the network sees more and more data during the training process. In other words, the rank of feature maps produced by different filters in a CNN can be estimated accurately and efficiently by training the network with only a small portion of the available training data. Following this, the low rank filters, which do not contribute significantly to model accuracy, can be removed to achieve the desired compression ratio. This can be posed as an optimization problem as:

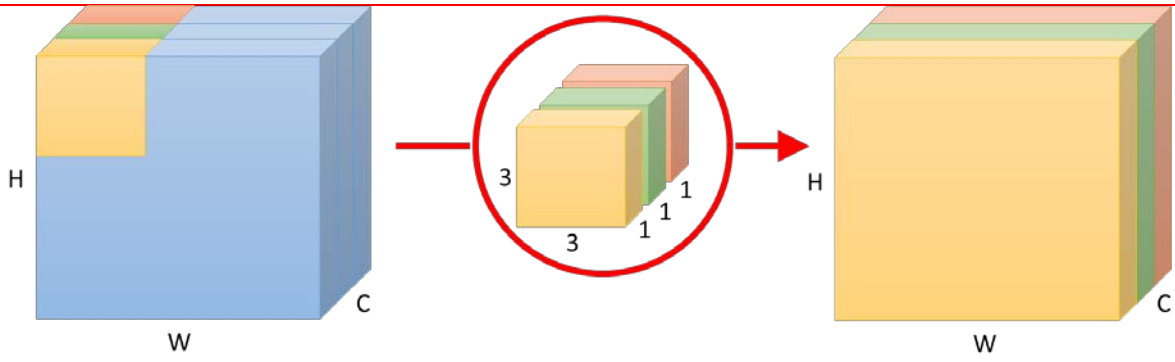
$$\begin{aligned} &\min_{\delta_{ij}} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \delta_{ij} \mathcal{L}\left(w_j^i\right) \\ &s.t. \quad \sum_{j=1}^{n_i} \delta_{ij} = n_{i2} \end{aligned} \tag{3}$$

where, K is the number of convolutional layers in a given CNN model. n_i is the number of filters at layer i. w_j^i denotes the parameters of filter j at layer i. n_{i2} is the number of unimportant filters at layer i. δ_{ij} is a binary indicator variable which assumes a value of 1 when j belongs to the set of unimportant filters, and 0 otherwise. $\mathcal{L}(\cdot)$ measures the importance of an input filter, which is manifested in the information that feature maps generated by it contains. To this end, the rank of feature maps is considered to be an effective indicator of the information content and is used in this study to define $\mathcal{L}(\cdot)$. Further details on this algorithm are included in M. Lin et al. (2020).

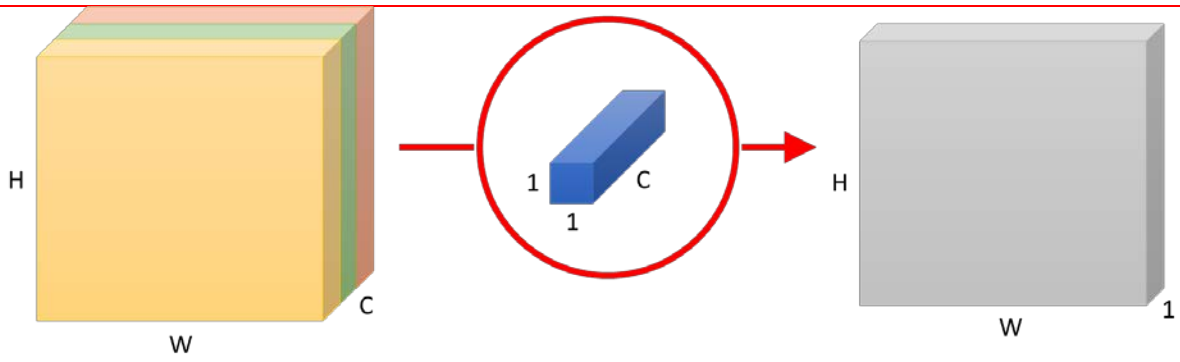
Last but not least, a generative adversarial learning (GAL)-based compression technique proposed by S. Lin et al. (2019) enables the pruning of heterogeneous structures (e.g. channels, branches, and blocks) in an end-to-end manner. This approach is based on the idea of knowledge distillation, where a large pretrained model (teacher) is used to teach a smaller student model. In other words, knowledge is distilled/transferred from a baseline teacher model to a significantly smaller counterpart model in a way that ensures that the smaller student model can accurately mimic the output of the original teacher model. The student model is obtained in this technique by introducing a trainable soft mask to scale the output of each structure. The scaling factors for

the redundant or less important structures are forced to zero by the optimization process involved in the network training, and the corresponding structures are removed to obtain a pruned model. Additionally, adversarial learning (Goodfellow et al. 2014) is invoked where the pruned model is treated as a generator, and a discriminator is used to distinguish between features generated by the original baseline model and its pruned counterpart. This ensures that the pruned model can effectively substitute the original baseline model without compromising on the accuracy. The readers may refer to the original paper by S. Lin et al. (2019) to know more about this algorithm.

The unique characteristics and underlying principles distinguish the four pruning algorithms considered in this study. For example, minimizing the Frobenius norm of the difference between the filter covariance matrices of the original and pruned models is the most critical factor in the FS-based pruning approach. On the other hand, identifying the most informative exemplars by employing an affinity propagation-based clustering technique is key to the AEF-based pruning method. Similarly, distinguishing the high-rank feature maps from the low-rank ones is crucial for achieving optimum network compression in the HR-based pruning technique. Last but not least, an essential element in the GAL-based pruning algorithm is distilling knowledge from a baseline teacher model to a significantly smaller counterpart model to ensure that the smaller student model can accurately reproduce the output of the original teacher model.



(a) Depthwise convolution



(b) Pointwise convolution

Fig. 1 Depthwise separable convolution

2.2 Depthwise Separable Convolution (DSC)

Separable convolution is a computationally cheaper alternative to normal convolution, which is central to any CNN-based algorithm. A separable convolution is a process in which a single convolution is split into two or more convolutions to produce the same effect. Along this line, DSC separates a regular convolution into depthwise convolution and pointwise convolution applied in sequence (Chollet 2017). Depthwise convolution uses the same number of filters as the number of input channels (Fig. 1a). These filters have a single depth and work separately on respective input channels, yielding one output feature corresponding to each input channel. The output features are stacked along the depth dimension producing an intermediate output feature map having the same number of channels as the input features. So this process does not change the depth dimension of convolutional features. However, it should be noted that regular convolution often involves changing the number of channels in the convolutional features. Pointwise convolution, which utilizes 1×1 kernels, is invoked to this end. In this process, each kernel has a depth equal to the number of input channels. The number of such kernels to be used is determined by the desired depth of the output features. Compared to regular convolution, DSC significantly downsizes the number of trainable parameters and thereby reduces the processing time. It also helps reduce overfitting leading to improved accuracy. The parameters for DSC were initialized in this study using the Kaiming uniform initialization technique (K. He et al. 2015) and were subsequently finetuned during the network training process.

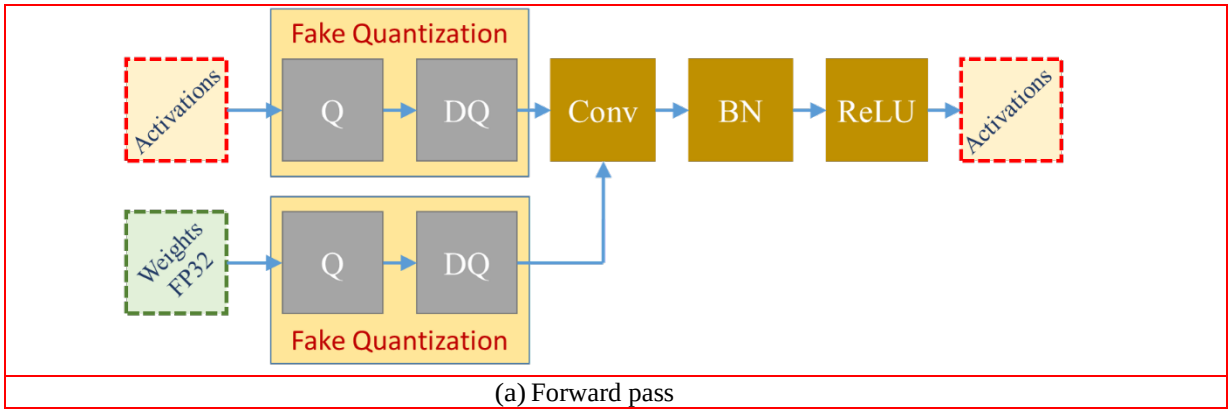
2.3 Quantization-aware Training (QAT)

Another important element of the proposed network compression strategy is network quantization. Quantization is the process of reducing the numerical precision of weights and activations without significantly compromising the network performance (Jacob et al. 2018). This

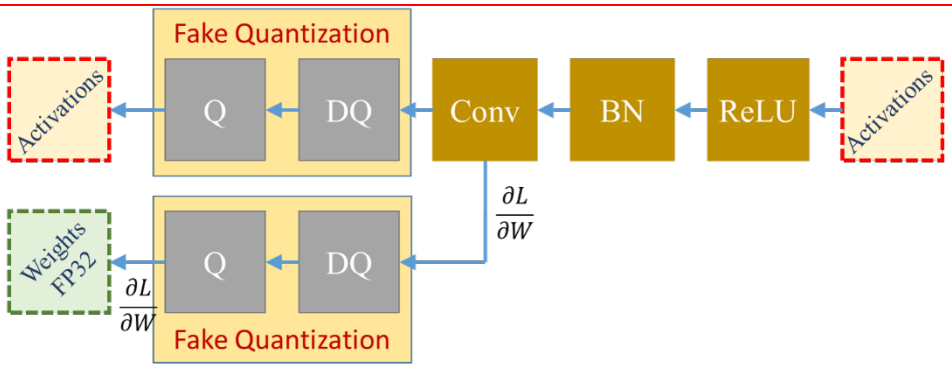
leads to faster processing and higher throughput. Besides, quantized models have a lower memory footprint and power consumption, which are crucial for edge deployment. Among various quantization techniques available in the literature, QAT is known to produce the highest accuracy and is therefore considered in this study. In this quantization scheme, the model learns the quantization effect by including the quantization error in the training loss. Quantization is simulated by inserting fake quantization modules (Fig. 2) in the training graph which quantize and de-quantize the data and weights in immediate succession, as shown in Eq. (4).

$$\hat{x} = f_d\big(f_q(x, s_x, z_x), s_x, z_x\big) + \Delta_x \tag{4}$$

where, x is a floating-point value; f_q and f_d are the quantization and de-quantization functions, respectively; Δ_x is the quantization error. On one hand, a fake quantization module enables the layer's operations to take place in floating-point precision (32-bits); on the other hand, it induces some quantization noise akin to what is generally encountered in quantized inference. This ensures that the prediction loss accounts for any expected quantization error. In the backward pass, a straight-through estimator (Bengio et al. 2013) is used to compute the quantization gradients and update the floating-point weights. At the end of the training process, the quantization scales are collected from the trained fake-quantization modules and are used to quantize the weights and activations to 8-bit integer precision at the time of inference.



(a) Forward pass



(b) Backward pass

Fig. 2 Fake quantization in forward and backward passes; Q: Quantization; DQ: De-quantization; Conv: Convolution; BN: Batch Normalization; ReLU: Rectified Linear Unit; L: Loss; W: Weights.

2.4 Proposed Hybrid Approach

It may be noted that the network compression induced by DSC, pruning, and QAT are orthogonal to each other. Therefore, an integration of these three approaches can potentially lead to greater network compression and speed-up, benefiting from the best of three worlds. This hypothesis was put to test in this study by formulating a hybrid three-pronged compression technique by combining DSC with the state-of-the-art pruning and quantization algorithms. In this new approach, all standard convolutions in the baseline CNN, where the kernel size is greater than one, are replaced by DSC (Fig. 3). It can be shown that replacing the 1 × 1 convolutions by DSC does not reduce the number of parameters. Therefore, the 1 × 1 convolutions, if any, are not altered in the proposed hybrid approach. The resulting network, which is already considerably downsized due to the induction of DSC, is then subjected to structured pruning. At this stage, the network is also trained simultaneously, producing a pruned version of the baseline model. This pre-trained pruned model is subsequently fine-tuned by QAT, resulting in a quantized network that is computationally light yet accurate, as described in the following sections. The proposed framework can be potentially extended to other types of neural networks with appropriate case-specific modifications.

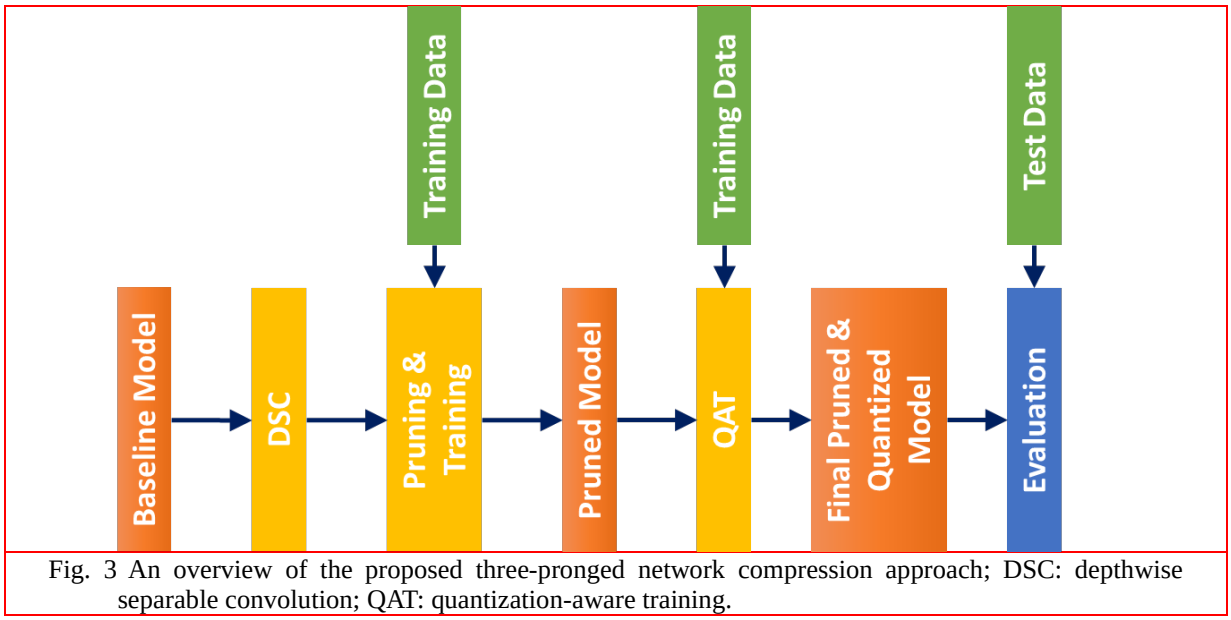


Fig. 3 An overview of the proposed three-pronged network compression approach; DSC: depthwise separable convolution; QAT: quantization-aware training.

3. Case Study

3.1 Data Description

The data used in this study were acquired on a long-span cable-stayed bridge in China from January to February 2012 (Tang et al. 2019). 38 channels of accelerometers were distributed uniformly across the bridge slabs, pylon, and pier, to collect data at a sampling rate of 20 Hz for long-term health monitoring of the bridge, as denoted by the red dots in Fig. 4. The measured accelerations were manually categorized into seven types of data anomalies, namely 'Normal', 'Missing', 'Minor', 'Outlier', 'Square', 'Trend', and 'Drift' (Bao et al. 2021). Each labeled signal was 1 hour long comprising 12000 data points. The 7 types of data anomalies are depicted as acceleration time histories in Fig. 5. As shown, each type of data anomaly is characterized by distinct features, indicating the potential of successful classification by deep learning techniques. A usual oscillation response is designated as 'Normal'. 'Missing' denotes the case where most/all of the vibration signal is missing. 'Minor' corresponds to a situation where the amplitude is very small relative to the normal sensor data. A time-history response containing one or more outliers is categorized as an 'Outlier'. On the other hand, a time-series signal resembling a square wave is labeled as 'Square'. A 'Trend' is a data anomaly that has a clear trend in the time domain. Last but not least, a 'Drift' indicates a non-stationary vibration response with random drift. The dataset was released during the 1st International Project Competition for Structural Health Monitoring (IPC-SHM 2020) (Bao et al. 2021) aiming to promote deep learning-based approaches for data anomaly classification. In this competition, the dataset collected in January

2012 was used for the training and validation of deep learning models. Additionally, the data collected in February 2012 were used as a blind dataset to test the model performance. This study followed the same scheme for splitting the data into training and test sets. A small subset of data (15%) was sliced off from the training set randomly to serve as a validation set during the training process. A classwise distribution of the size of the datasets used in this study for training, validation, and testing of deep learning models is shown in Fig. 6.

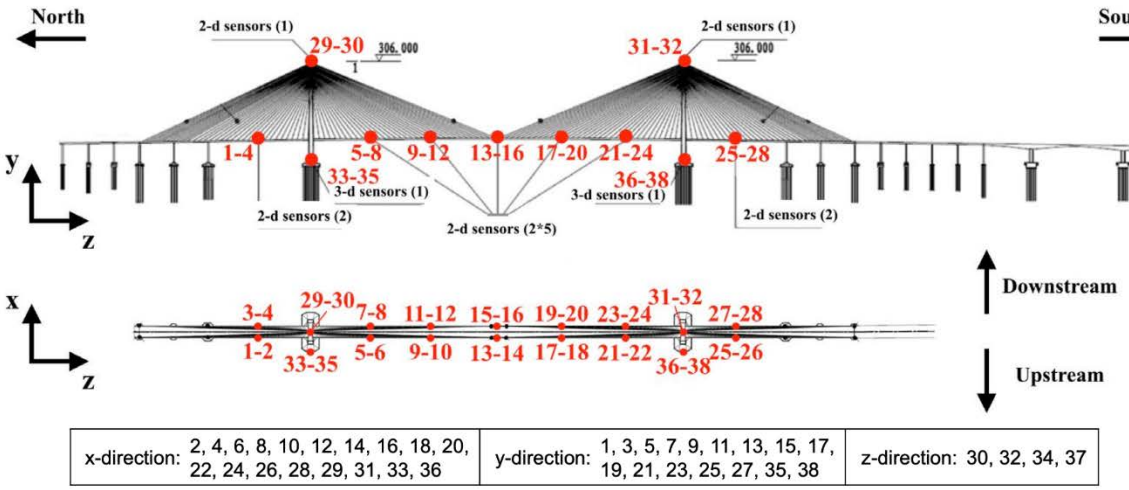
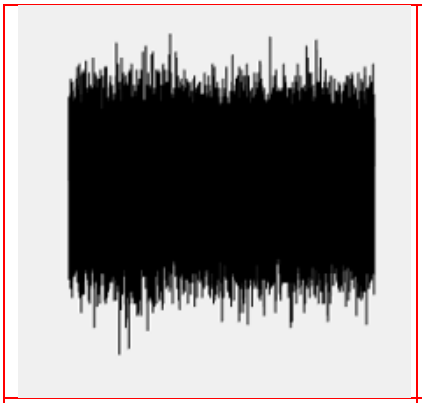
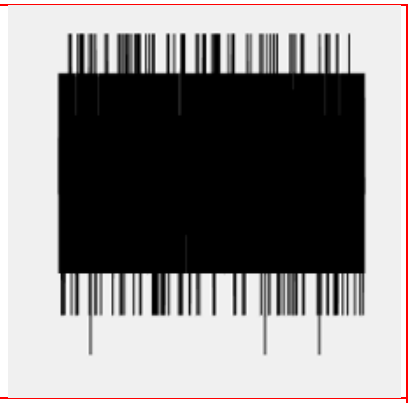


Fig. 4 Depiction of sensor locations used for data acquisition (Bao et al. 2021)

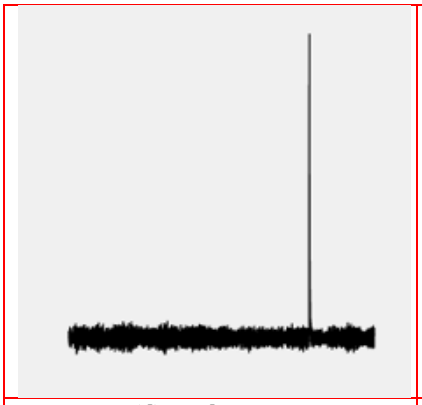


(a) Normal

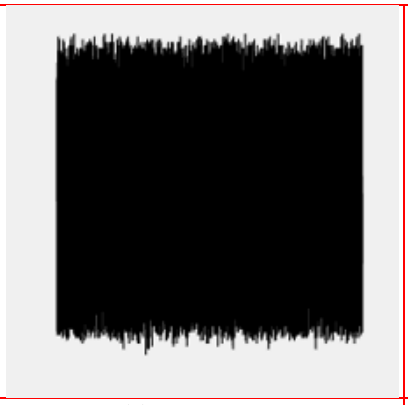
(b) Missing



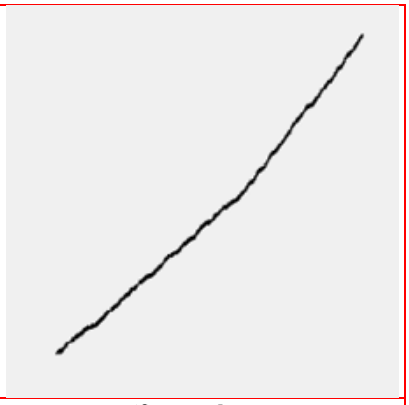
(c) Minor



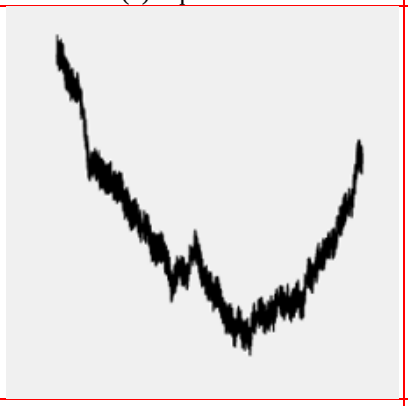
(d) Outlier



(e) Square



(f) Trend



(g) Drift

Fig. 5 Illustration of different types of data anomalies

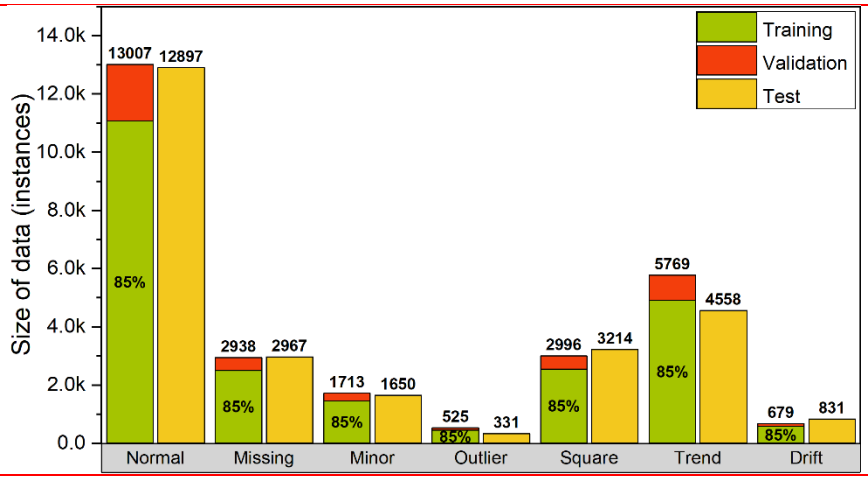


Fig. 6 Size of data for different types of data anomalies

3.2 Baseline Deep CNN and Data Representation

This study exploited GoogLeNet as the baseline model to evaluate the efficacy of various network compression strategies taken up in this study. This network is chosen in this study as it was proven to be one of the best performing models that IPC-SHM 2020 had produced (Chou et al. 2022). This CNN is based on Inception architecture (Szegedy et al. 2015) which stacks filters of various sizes on the same level, making the network wider rather than deeper (Fig. 7). The outputs from different filters are concatenated and sent to the next layer. Additionally, 1 × 1 convolutions are employed aiming at reducing the dimension of input channels and thereby reducing the computational cost. More details about the network can be found in Szegedy et al. (2015). Besides, the time series data should be represented in a proper way to ensure a high level of accuracy for the CNN model. Chou et al. (2022) demonstrated that the time-frequency representation of a time series can serve as a more informative CNN input as compared to the direct transformation of the time series to an image, and the same strategy is adopted in this study. Therefore, the time-frequency response is first obtained through short-time Fourier transform with a Kaiser window using time series normalized between 0 and 1. The frequency range in this study is selected to vary from 0 to 20 Hz, resulting in a Nyquist frequency of 10 Hz with a frequency resolution of 0.0017 Hz/sample. The three parameters, namely time, frequency, and magnitude are then converted to a 2D heat map where the time and frequency points are represented in the horizontal and vertical axes, respectively; the magnitudes are shown in gradient colors (Fig. 8). Next, the time history is added to the time-frequency heat map along the time axis, creating two different scales of vertical axes (i.e., the frequency points and the acceleration magnitude). Since low frequencies are more observable in bridge vibrations, the time history is shifted away from the low-frequency range. Finally, the image is resized to a resolution of 224 pixels × 224 pixels using bi-cubic interpolations before being used as input to the GoogLeNet model, which reduces the frequency resolution to some extent.

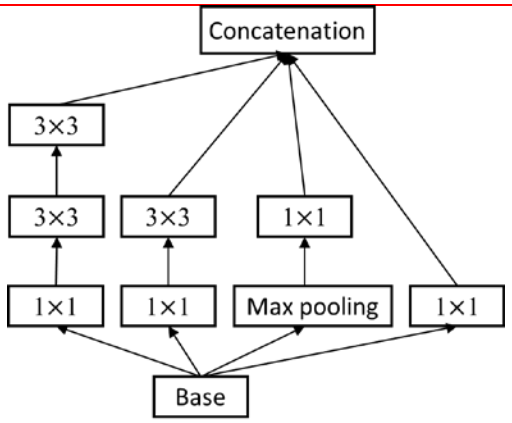
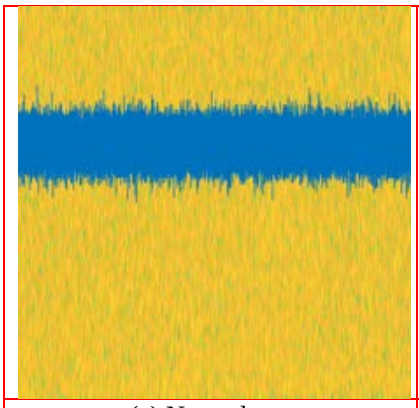
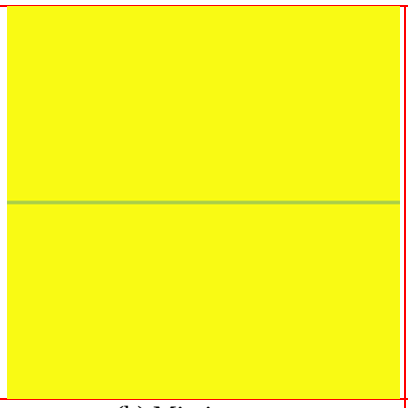


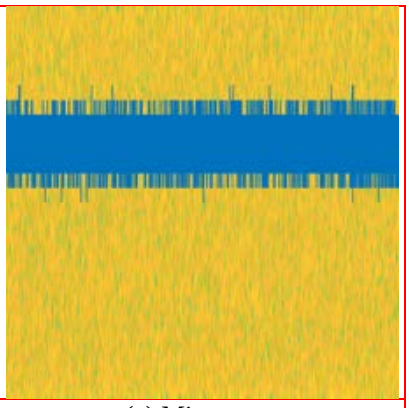
Fig. 7 Inception module used in GoogLeNet



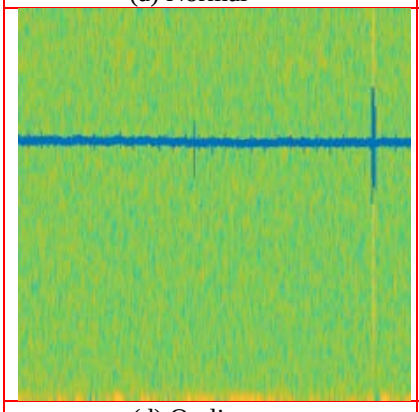
(a) Normal



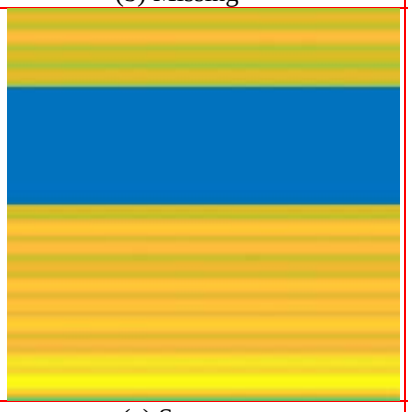
(b) Missing



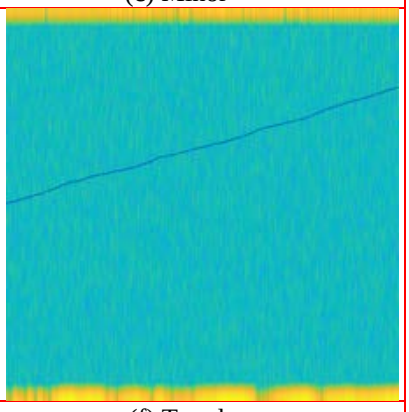
(c) Minor



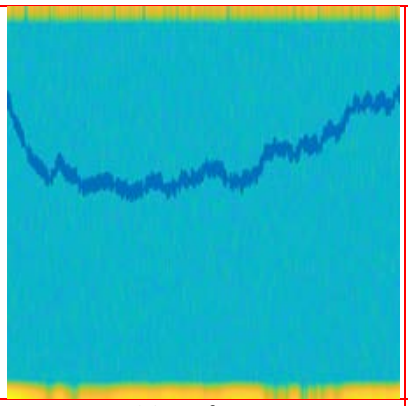
(d) Outlier



(e) Square



(f) Trend



(g) Drift

Fig. 8 Time-frequency representation of different types of data anomalies

3.3 Implementation Details

To implement the proposed hybrid compression strategy, all the 3 × 3 convolutions in GoogLeNet were first replaced by DSC. The resulting network was then pruned using the structured pruning techniques described in Section 2, and trained simultaneously to produce a pruned version of the baseline GoogLeNet model. This pre-trained model was then finetuned through QAT, resulting in a model which is pruned and quantized and, as a result, significantly reduced in size and computation cost. The model was trained and finetuned by minimizing a cross-entropy loss between the target and predicted class labels using a stochastic gradient descent optimizer. Moreover, this study employed a class weighting strategy to deal with the menace of class imbalance. In this technique, the cost function is adjusted in such a way that misclassifying an anomaly from a minority class is more heavily penalized than that from a majority class. This is achieved by assigning weights to the classification loss which are inversely proportional to the class frequencies in the training data. This approach rebalances the class distribution and assigns equal importance to all classes during gradient updates. This prevents the network from over-classifying the more frequent class based on its higher prior probability.

3.4 Results and Discussions

In this study, the performance of various compression methods is evaluated in terms of retention ratio, which measures the percentage of parameters or floating point operations (FLOPs) that are retained after the application of a compression technique. The lower the retention ratio higher is the degree of compression, and therefore, the more efficient is the compression technique. A sensitivity analysis is carried out to identify the best set of parameters

for each pruning technique. In the case of FS, the ratio of the number of channels in the pruned model to that in the original unpruned model is designated as the sketch rate. Therefore, a lower sketch rate leads to higher compression (Table 1). On the other hand, the AEF-based pruning is sensitive to a hyper-parameter denoted by β, which is closely linked to the complexity reduction. A large β leads to higher compression and vice-versa (Table 2). Similarly, the sensitivity of various compression ratios for the HR-based pruning method is shown in Table 3. Finally, the influence of the sparsity regularization factor (λ) on the pruning rate for the GAL-based approach is summarized in Table 4. The best hyper-parameter for each pruning technique is highlighted in bold font in Tables 1-4.

Table 1 Sensitivity analysis of sketch rate for FS-based pruning; The most ideal sketch rate is shown in bold font.

Sketch Rate	Parameters	Retention	Accuracy
0.10	**1905745**	**30.9%**	**98.9%**
0.15	2131991	34.6%	98.8%
0.20	2359227	38.3%	99.0%
0.30	2817737	45.7%	99.2%
0.40	3279607	53.2%	99.4%
0.50	3763615	61.1%	99.2%
0.60	4217609	68.4%	99.1%
0.70	4692187	76.1%	99.2%
0.80	5172225	83.9%	99.1%
0.90	5655191	91.8%	99.1%

Table 2 Sensitivity analysis of β, which is a hyper-parameter linked to the complexity reduction for AEF-based pruning; The most ideal value for β is shown in bold font.

β	Parameters	Retention	Accuracy
0.75	6156284	99.9%	98.5
0.8	6061279	98.3%	98.5
0.85	5456282	88.5%	99.0
0.9	3442057	55.8%	99.0
0.95	2222929	36.1%	98.7
1	1993410	32.3%	98.6
1.2	1728537	28.0%	98.4
1.4	1658883	26.9%	98.7
2	1576811	25.6%	98.6
2.2	1562962	25.4%	98.7
3	1540584	25.0%	98.5

4	**1521657**	**24.7%**	**98.0**

Table 3 Sensitivity analysis of various compression ratios for HR-based pruning; *a*, *b*, *c*, and *d* denote the compression rates for the initial convolutional layer and the inception blocks 1-2, 3-7, 8-9, respectively. The most ideal parameter values are shown in bold font.

Compression Rate: [*a*]+[*b*]*2+[*c*]*5+[*d*]*2				Parameters	Retention	Accuracy
a	b	c	d			
0.3	0.6	0.7	0.8	2854685	46.3%	99.2
0.3	0.66	0.75	0.8	2695862	43.7%	99.3
0.35	0.72	0.8	0.85	2473665	40.1%	99.3
0.35	0.78	0.85	0.85	2322381	37.7%	99.4
0.4	0.85	0.9	0.9	2096191	34.0%	99.3
0.4	0.9	0.95	0.95	1901966	30.9%	99.3
0.45	**0.95**	**0.97**	**0.97**	**1798717**	**29.2%**	**99.3**

Table 4 Sensitivity analysis of sparsity factor (λ) for GAL-based pruning; The most ideal λ value is shown in bold font.

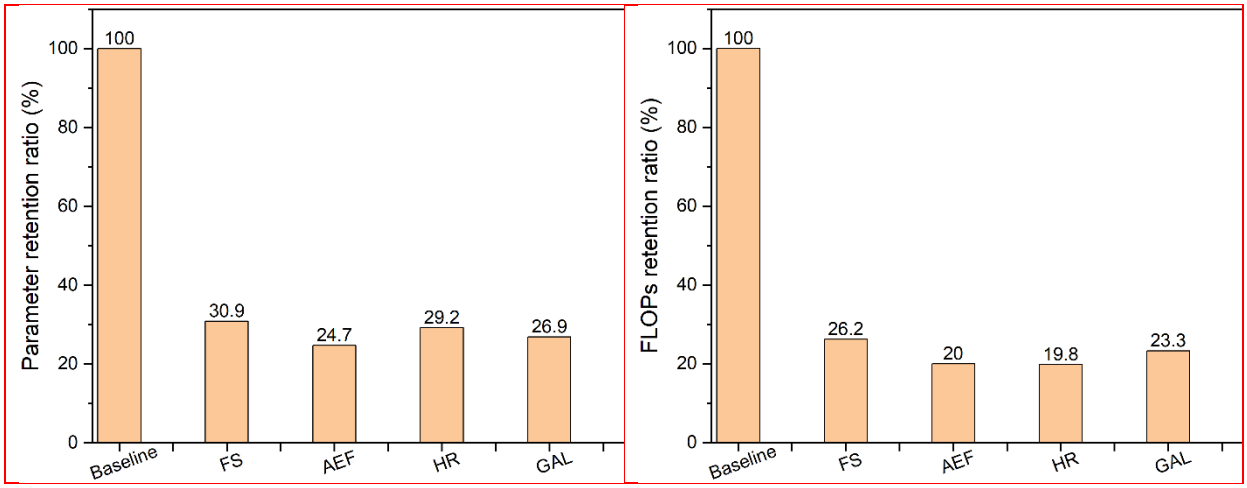
λ	Parameters	Retention	Accuracy
0	4053679	65.8%	98.7%
0.05	2283687	37.1%	98.9%
0.1	**1657487**	**26.9%**	**98.9%**
0.15	2537943	41.2%	98.7%

The model accuracies on validation and test data for the best set of pruning parameters are reported in Table 5. It should be noted in this context that accuracy is the ratio of correct predictions to the total number of cases examined. Various pruning techniques were first applied to the baseline GoogLeNet model to evaluate the efficacy of each method. It was observed that AEF was the most efficient in terms of parameter retention ratio retaining about 24.7% of the network parameters (Fig. 9a). The AEF technique also secured a test accuracy of 96%, which is at par with the original baseline model. On the other hand, FS was the least efficient as it retained about 30.9% of the model weights. However, it achieved a test accuracy of 97%, which is 1% higher than the baseline accuracy. The other two pruning techniques, namely HR and GAL exhibited a parameter retention ratio of 29.2% and 26.9%, respectively. The corresponding test accuracies for these techniques were 97.5% and 96.5%, respectively. This implies that network compression does not always lead to a reduction in accuracy. On the contrary, in some cases, it may increase the accuracy by reducing the overfitting. However, it should be noted in this context

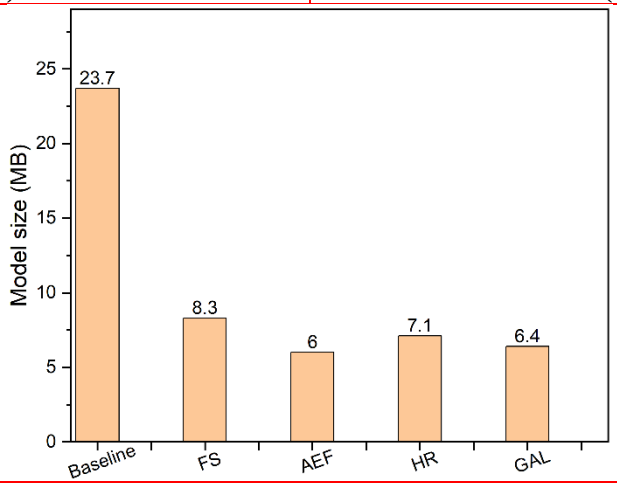
that pruning is designed not to improve accuracy but to reduce the computational overhead and thus make it suitable for deployment in resource-constrained edge environments. Network pruning aims to eliminate the redundant and less informative parameters that contribute little to the final output. This makes the network lightweight and fast while closely retaining the original model performance. This is why the pruning techniques did not offer any significant advantages in terms of accuracy compared to the unpruned baseline. A slightly different trend was observed in the FLOP retention ratio, where the HR came out as the most efficient pruning technique having a FLOP retention ratio of 19.8% (Fig. 9b). The FS still remained the least efficient producing a FLOP retention ratio of 26.2%. The remaining two pruning algorithms, namely AEF and GAL could retain about 20% and 23.8% of the original FLOPs. One can argue that using a simple neural network may turn out more advantageous than pruning a deep CNN. However, Chou et al. (2022) have shown that simple neural networks (overall accuracy on blind dataset = 75.6%) cannot generally match the accuracy of deep CNNs in the context of data anomaly detection. Therefore, pruning a deep CNN becomes inevitable for balancing high accuracy with computation cost.

Table 5 Performance of various pruning techniques in terms of parameter retention ratio, FLOP retention ratio, and accuracy.

Method	Parameters	Retention	FLOPs	Retention	Accuracy	
					Validation	Blind Test
Baseline	6163175	100 %	74.9 B	100%	99.2%	96.0%
FS	1905745	30.9%	19.7 B	26.3%	98.9%	97.0%
AEF	1521657	24.7%	15.0 B	20.0%	98.0%	96.0%
HR	1798717	29.2%	14.8 B	19.8%	99.3%	97.5%
GAL	1657487	26.9%	17.5 B	23.3%	98.9%	96.5%



(a) (b)



(c)

Fig. 9 (a) Parameter retention ratio, (b) FLOP retention ratio, and (c) model size for different pruning techniques.

Apart from parameter retention ratio and FLOP retention ratio, another parameter which is of utmost importance, particularly from the perspective of edge analysis, is model size. Edge devices are generally starved of storage space. Therefore, any reduction in the memory requirement immensely benefit the resource-constrained edge applications. Therefore, model size is also used as an indicator in this study to evaluate the performance of the compression techniques (Fig. 9c). It was found that the pattern observed in the model size is no different than that in the parameter retention ratio. The most efficient approach, namely AEF calls for 6 megabytes of space to store all the model weights. Whereas, FS, which is the least efficient technique, consumes about 8.3 megabytes of space for the same purpose. The other two pruning techniques, namely HR and GAL require about 7.1 and 6.4 megabytes of spaces, respectively, for the storage of the network parameters.

Table 6 Performance of various hybrid compression strategies in terms of parameter retention ratio, FLOP retention ratio, and accuracy.

Method	Parameters	Retention	FLOPs	Retention	Accuracy	
					Validation	Blind Test
Baseline	6163175	100 %	74.9 B	100%	99.2%	96.0%
DSC	2733954	44.4%	29.2 B	39.0%	99.2%	96.2%
DSC-FS	2411838	39.1%	25.3 B	33.8%	99.1%	95.2%
DSC-AEF	1479578	24%	12.1 B	16.1%	96.9%	94.9%
DSC-HR	1143835	18.6%	11.6 B	15.5%	99.2%	96.4%

DSC-GAL	509234	8.3%	6.9 B	9.3%	97.9%	94.6%

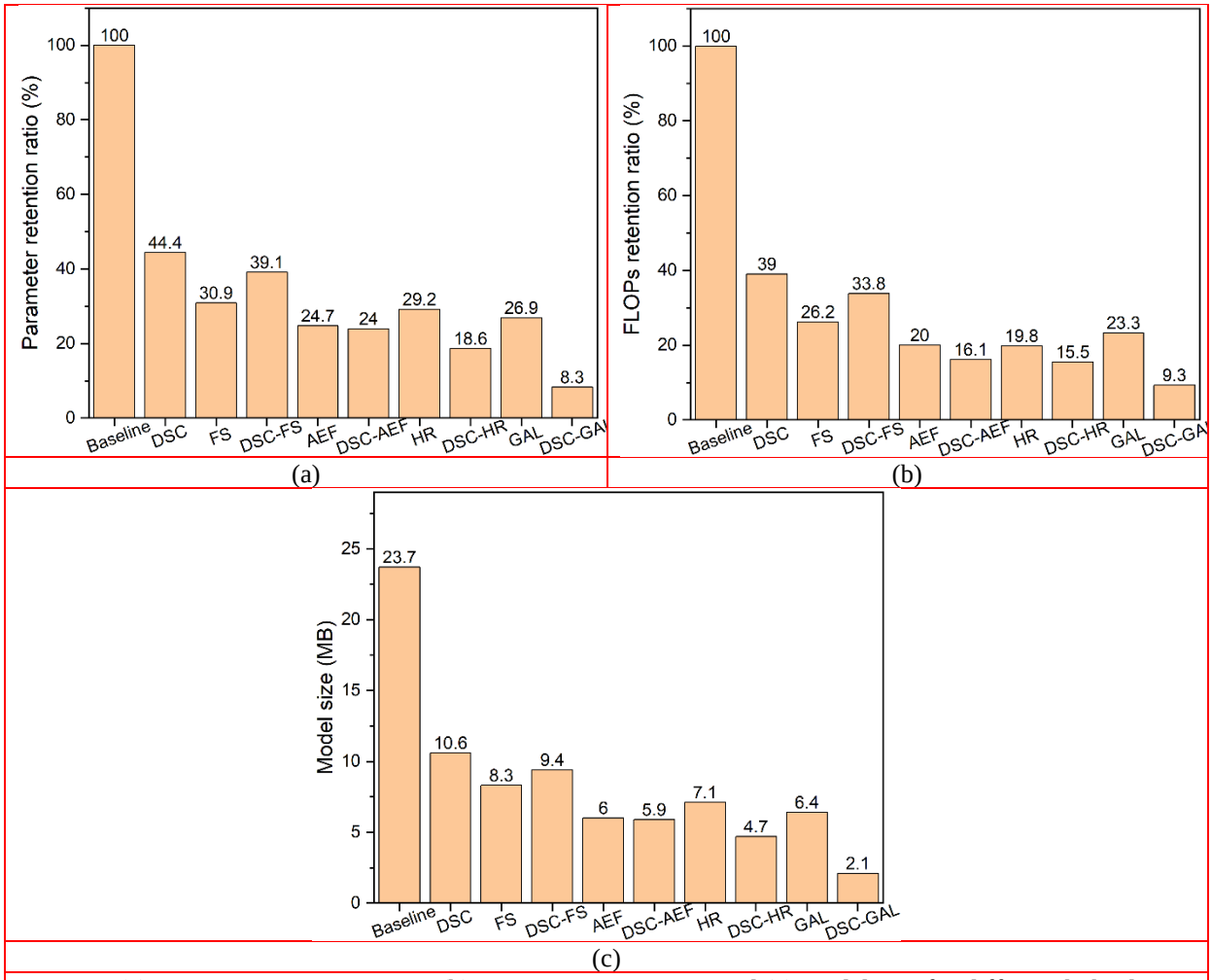


Fig. 10 (a) Parameter retention ratio, (b) FLOP retention ratio, and (c) model size for different hybrid compression strategies.

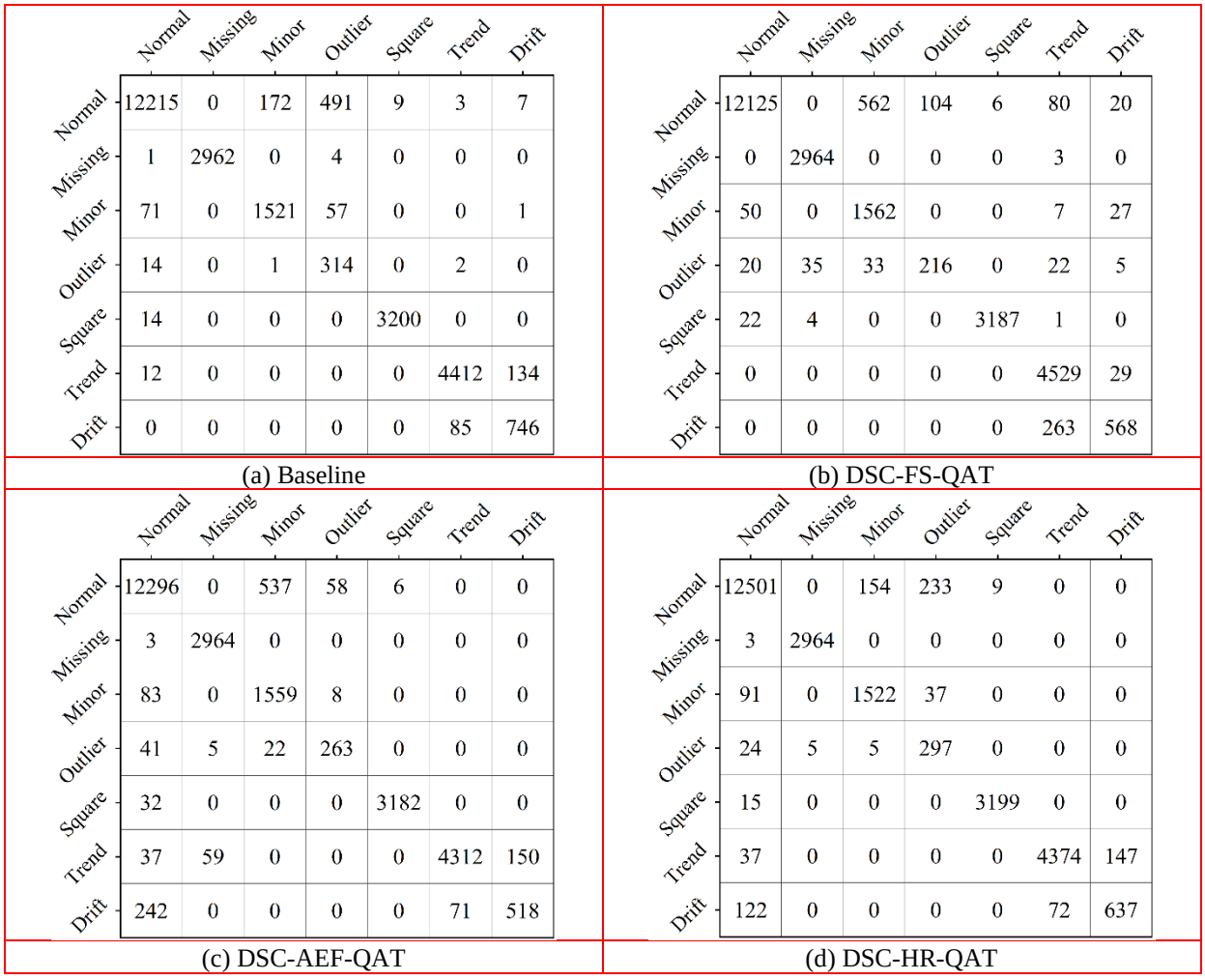
In the best of cases, the pruning algorithms still retained 24.7% of parameters and 20% of FLOPs, consuming 6 megabytes of space in the memory. In a quest to achieve a higher compression ratio, a hybrid approach was developed by integrating DSC with the aforementioned pruning algorithms. This reduced the number of parameters and FLOPs by 55.6% and 61%, respectively, even without the application of any pruning technique, and without compromising

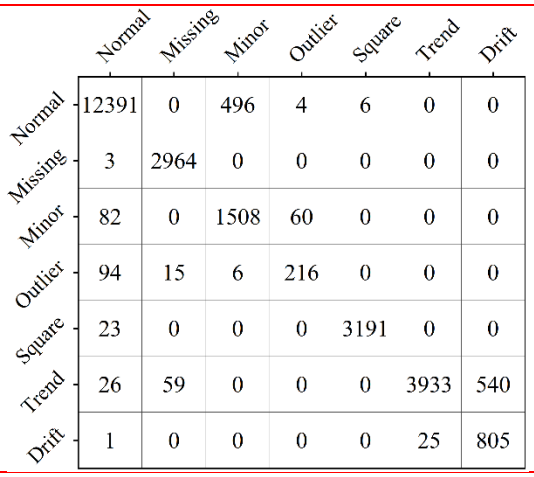
on the test accuracy (Table 6). Following this, the pruning techniques were employed on top of this downsized network. It was observed that the hybrid approach can compress the network more than any pruning technique can do independently, with a possible exception of the FS. The FS involves learning a new set of parameters that preserve the second-order information of the original unpruned model weights. However, as noted earlier, the application of DSC ends up removing a lot of parameter redundancy, leaving little scope for FS to induce further compression. As a result, the integration with DSC proved to be disadvantageous for the FS approach. However, for all other cases, a greater degree of compression was achieved due to the incorporation of DSC. In terms of parameter retention ratio, an additional 0.7%, 10.6%, and 16.6% compression was achieved by the hybrid approach as compared to the individual pruning techniques, that is to say AEF, HR, and GAL, respectively (Fig. 10a). On the other hand, the additional reductions in FLOPs due to the hybrid technique were 3.9%, 4.3%, and 14% for AEF, HR, and GAL, respectively (Fig. 10b). The same advantage was also reflected in the model size, where a 0.1 megabyte, 2.4 megabyte, and 4.3 megabyte reduction in the memory requirement was observed in the case of AEF, HR, and GAL, respectively, owing to the hybrid method (Fig. 10c). On the downside, incorporation of DSC was seen to reduce the test accuracy to a certain degree. In more specific terms, a 1.1%, 1.1%, and 1.9% diminution in the test accuracy was noted in the case of AEF, HR, and GAL, respectively, which are attributable to the hybrid approach (Table 6).

Subsequent to this, the hybrid approach was extended to include QAT. The aforementioned trained DSC-based pruned models were further finetuned using QAT, leading to additional compression in the model size, as shown in Table 7. It was observed that QAT reduced the model size by 70.21%, 71.19%, 65.96%, and 66.67% for models subjected to FS, AEF, HR, and GAL-based pruning, respectively. The sizes of the resulting pruned and quantized models were only 11.81%, 7.17%, 6.75%, and 2.95% of the baseline model. At the same time, QAT did not lead to any appreciable loss of accuracy. This is a win-win on all counts, as the model size is considerably reduced at almost no cost of accuracy.

Table 7 Results of quantization-aware training (QAT)

Method	Model Size	Accuracy
Baseline	23.7 MB	96.0%
DSC-FS	9.4 MB	95.2%
DSC-FS-QAT	2.8 MB	95.1%
DSC-AEF	5.9 MB	94.9%
DSC-AEF-QAT	1.7 MB	94.9%
DSC-HR	4.7 MB	96.4%
DSC-HR-QAT	1.6 MB	96.4%
DSC-GAL	2.1 MB	94.6%
DSC-GAL-QAT	0.7 MB	94.5%

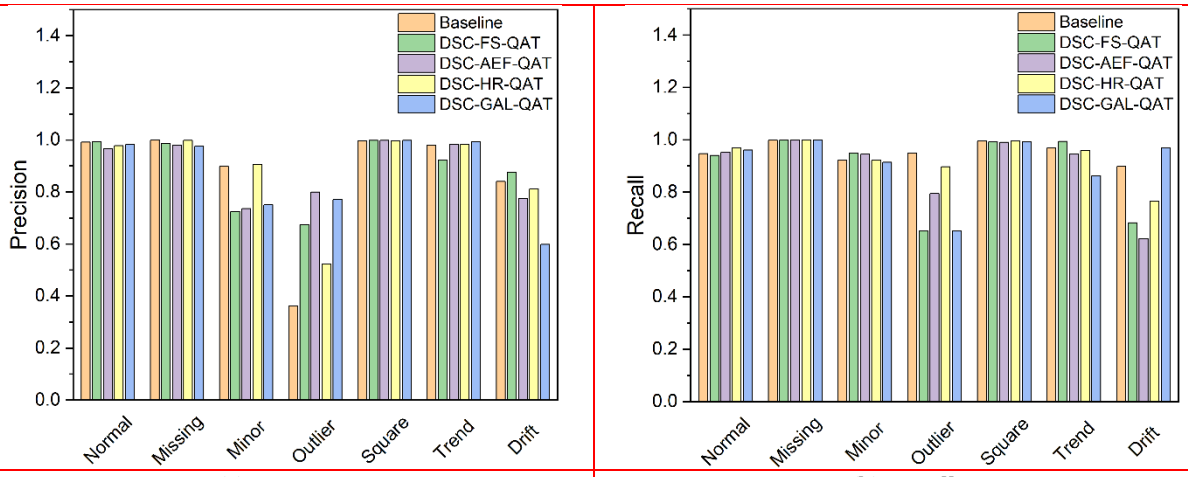

Normal
Missing
Minor
Outlier
Square
Trend
Drift
12215 0 172 491 9 3 7
1 2962 0 4 0 0 0
71 0 1521 57 0 0 1
14 0 1 314 0 2 0
14 0 0 0 3200 0 0
12 0 0 0 0 4412 134
0 0 0 0 0 85 746
(a) Baseline
12125 0 562 104 6 80 20
0 2964 0 0 0 3 0
50 0 1562 0 0 7 27
20 35 33 216 0 22 5
22 4 0 0 3187 1 0
0 0 0 0 0 4529 29
0 0 0 0 0 263 568
(b) DSC-FS-QAT
12296 0 537 58 6 0 0
3 2964 0 0 0 0 0
83 0 1559 8 0 0 0
41 5 22 263 0 0 0
32 0 0 0 3182 0 0
37 59 0 0 0 4312 150
242 0 0 0 0 71 518
(c) DSC-AEF-QAT
12501 0 154 233 9 0 0
3 2964 0 0 0 0 0
91 0 1522 37 0 0 0
24 5 5 297 0 0 0
15 0 0 0 3199 0 0
37 0 0 0 0 4374 147
122 0 0 0 0 72 637
(d) DSC-HR-QAT



(e) DSC-GAL-QAT

Fig. 11 Confusion matrices corresponding to the baseline and compressed models

The efficiency of the pruned and quantized models is also presented in the form of confusion matrices to enable a more comprehensive performance assessment (Fig. 11). The horizontal and vertical dimensions of the tables correspond to the predicted and ground truth labels, respectively. A common weakness in all the models concerned a normal signal being misclassified as a 'Minor' anomaly or an 'Outlier'. This can be attributed to the great degree of visual similarities between these anomaly classes. For the same reason, a number of signals with 'Trends' were mislabeled as 'Drift', and vice-versa. Another prominent soft spot in most networks was that a number of anomalous signals were predicted as normal. These instances mostly pertained to the situations where the anomaly signature was not very vivid and clear. All the same, it was seen that the diagonal entries in the confusion matrices are significantly larger than the nondiagonal elements signifying a high degree of accuracy for the baseline and compressed models. Including additional training samples representing the under-represented anomaly classes is likely to mitigate the problem of class imbalance, leading to improved detection performance. Moreover, class-specific precision and recall values were computed, as shown in Fig. 12. For a given class, precision indicates what percentage of the predicted anomaly are true positives. On the other hand, recall implies the percentage of the actual anomalies successfully classified. In a general sense, the performance of the pruned and quantized models was not significantly different from the baseline network, barring a few exceptions concerning 'Outlier' and 'Drift'.



(a) Precision (b) Recall

Fig. 12 Class-specific performance evaluation in terms of precision and recall

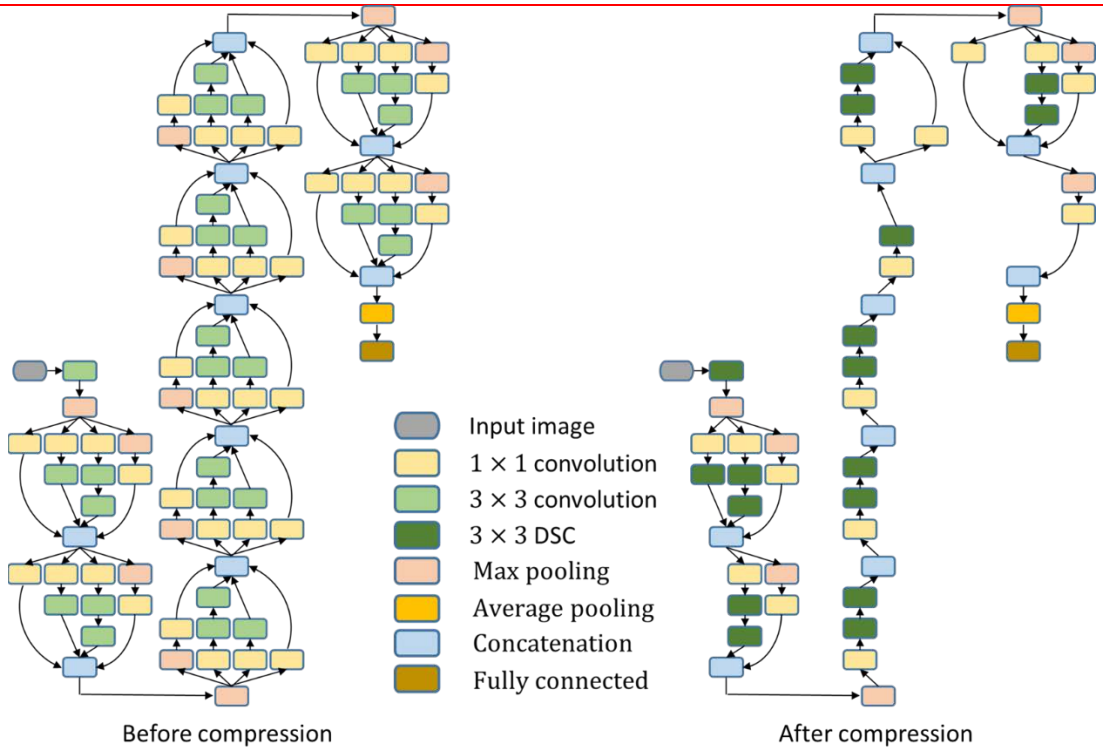


Fig. 13 A depiction of the GoogLeNet architecture before and after the application of the hybrid compression technique.

On the whole, the hybrid approach entailing DSC and GAL produced the highest compression ratio among all techniques, retaining only 8.3% of parameters, 9.3% of FLOPs, and requiring only 0.7 megabytes of space in the memory. The GAL is unique among all pruning strategies examined in this study in the sense that it can effectively deal with branch redundancy, which is a common menace in any multi-branch network, by completely eliminating a redundant branch without cutting off the information flow. In the aforementioned best case scenario, it was observed that 21 out of the 36 branches present in the baseline GoogLeNet architecture were removed (Fig. 13), leading to such a high degree of compression. However, the generalizability of the observed results is contingent upon a range of different factors. For example, the conclusions may not hold in the event of a significant domain shift. In this regard, a more balanced training dataset with an equitable representation of different anomaly classes will add to the reliability of the inferences drawn in this study.

A variant of the proposed strategy has also been applied to a solution of the third project in the IPC-SHM 2020, which focused on the condition assessment of stay cables of a large-span cable-stayed bridge in Mainland China. The first prize winner for this project, Zhang et al. 2021, developed a hybrid deep learning model comprising a long short-term memory network and a fully convolutional network to accurately identify the damaged cables based on cable force and cable force ratio. The strategy proposed in this study was used to compress the deep network by replacing the regular convolutions with DSC, applying a state-of-the-art pruning method, and conducting a post-training quantization. As a result, the model size was reduced to 258 KB, which implies a more than five times reduction compared to the original baseline model while encountering only a 1% drop in accuracy. As the scope of the third project does not entail data anomaly detection, the detailed results are therefore not included in this paper but can be found in Aishah (2022). Likewise, the proposed framework can be extended to many other relevant application areas, including but not limited to real-time visual data analytics in edge environments provided by mobile robotic inspection platforms and head-mounted augmented reality devices (Mondal 2021, Huang et al. 2023).

4. Conclusions

Conventional studies deal with the problem of data anomaly detection as a part of big data challenges, where data cleansing and mining are performed after the data are collected in a central station. However, the transmission of misleading anomalous data in such systems leads to misuse of power and memory. This study aimed to provide an alternative framework based on edge intelligence to minimize the resource consumption and maximize the effective information in a smart infrastructure monitoring system. Latest deep learning-based approaches for autonomous data anomaly detection are computationally expensive, and therefore not suitable to be deployed in resource-constrained edge environments. This study addressed this challenge by

proposing an efficient and hybrid deep neural network compression approach leveraging DSC along with state-of-the-art pruning and quantization methods. The combination of DSC, GAL, and QAT demonstrated the greatest efficacy by compressing a deep CNN like GoogLeNet by more than 90%. The proposed framework will help mitigate the impact of sensor malfunction, optimize data transmission, and minimize the energy drainage. The problem of data anomaly detection was considered as a case study in this work to showcase the feasibility of this approach. However, the proposed network compression and acceleration strategies can be potentially extended to other IoT-based SHM applications with necessary modifications. This study presents a proof of concept for the proposed edge analysis framework for anomalous data detection. Future studies should focus on validating this approach through actual field experiments. Evaluating the inference time and power consumption on various edge computing platforms is another scope for future work.

5. Acknowledgement

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-001), the Ministry of Education Tier 1 Grants, Singapore (No. RG121/21), and the start-up grant at Nanyang Technological University, Singapore (03INS001210C120).

References

Al-amri, R., Murugesan, R. K., Man, M., Abdulateef, A. F., Al-Sharafi, M. A. and Alkahtani, A. A. (2021), "A review of machine learning and deep learning techniques for anomaly detection in IoT data", *Applied Sciences*, **11**(12), 5320. https://doi.org/10.3390/app11125320.

Alavi, A. H., Jiao, P., Buttlar, W. G. and Lajnef, N. (2018), "Internet of things-enabled smart cities: State-of-the-art and future trends", *Measurement*, **129**, 589–606. https://doi.org/10.1016/j.measurement.2018.07.067.

Bao, Y., Chen, Z., Wei, S., Xu, Y., Tang, Z. and Li, H. (2019). "The state of the art of data science and engineering in structural health monitoring", *Engineering*, **5** (2), 234–242. https://doi.org/10.1016/j.eng.2018.11.027.

Bao, Y., Li, J., Nagayama, T., Xu, Y., Spencer Jr, B. F. and Li, H. (2021), "The 1st international project competition for structural health monitoring (IPC-SHM, 2020): A summary and benchmark problem", *Structural Health Monitoring*, **20**(4), 2229– 2239. https://doi.org/10.1177/14759217211006485.

Bengio, Y., L´eonard, N. and Courville, A. (2013), "Estimating or propagating gradients through stochastic neurons for conditional computation", arXiv preprint arXiv:1308.3432. https://doi.org/10.48550/arXiv.1308.3432.

Bisio, I., Garibotto, C., Lavagetto, F. and Sciarrone, A. (2022), "A novel IoT-based edge sensing platform for structure health monitoring", *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 1–6, NY, USA, May.

Chatterjee, A. and Ahmed, B. S. (2022), "IoT anomaly detection methods and applications: A survey", *Internet of Things*, **19**, 100568. https://doi.org/10.1016/j.iot.2022.100568.

Chollet, F. (2017), "Xception: Deep learning with depthwise separable convolutions", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258, Honolulu, Hawaii, July.

Chou, J.Y., Fu, Y., Huang, S.K. and Chang, C.M. (2022), "SHM data anomaly classification using machine learning strategies: A comparative study", *Smart Structures and Systems*, **29**(1), pp.77-91. https://doi.org/10.12989/sss.2022.29.1.077.

Cook, A. A., Mısırlı, G. and Fan, Z. (2019), "Anomaly detection for IoT time-series data: A survey", *IEEE Internet of Things Journal*, **7**(7), 6481–6494. 10.1109/JIOT.2019.2958185.

Denil, M., Shakibi, B., Dinh, L., Ranzato, M. and De Freitas, N. (2013), "Predicting parameters in deep learning", *Advances in Neural Information Processing Systems*, **26**.

Ding, X., Zhou, X., Guo, Y., Han, J., Liu, J., et al. (2019), "Global sparse momentum SGD for pruning very deep neural networks", *Advances in Neural Information Processing Systems*, **32**.

Du, Y., Li, L.-f., Hou, R.-r., Wang, X.-y., Tian, W. and Xia, Y. (2022), "Convolutional neural network-based data anomaly detection considering class imbalance with limited data", *Smart Structures and Systems*, **29**(1), 63–75. https://doi.org/10.12989/sss.2022.29.1.063.

Frankle, J. and Carbin, M. (2018), "The lottery ticket hypothesis: Finding sparse, trainable neural networks", arXiv preprint arXiv:1803.03635. https://doi.org/10.48550/arXiv.1803.03635.

Frey, B. J., and Dueck, D. (2007), "Clustering by passing messages between data points", *Science*, **315**(5814), 972–976. https://doi.org/10.1126/science.1136800.

Fu, Y., Peng, C., Gomez, F., Narazaki, Y. and Spencer Jr, B. F. (2019), "Sensor fault management techniques for wireless smart sensor networks in structural health monitoring", *Structural Control and Health Monitoring*, **26**(7), e2362. https://doi.org/10.1002/stc.2362.

Fu, Y., Zhu, L., Hoang, T., Mechitov, K., & Spencer Jr, B. F. (2018, March). "Demand-based wireless smart sensors for earthquake monitoring of civil infrastructure", *In Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2018*, 10598, pp. 245-251, SPIE. https://doi.org/10.1117/12.2296634.

Fu, Y., Zhu, Y., Hoang, T., Mechitov, K. and Spencer Jr, B. F. (2022). "xImpact: Intelligent Wireless System for Cost-Effective Rapid Condition Assessment of Bridges under Impacts", *Sensors*, **22**(15), 5701. https://doi.org/10.3390/s22155701.

Gao, K., Chen, Z.-D., Weng, S., Zhu, H.-P. and Wu, L.-Y. (2022), "Detection of multi-type data anomaly for structural health monitoring using pattern recognition neural network", *Smart Structures and Systems*, **29**(1), 129–140. https://doi.org/10.12989/sss.2022.29.1.129.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014), "Generative adversarial nets", *Advances in Neural Information Processing Systems*, **27**. https://doi.org/10.1145/3422622.

Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M. A. and Dally, W. J. (2016), "EIE: efficient inference engine on compressed deep neural network", *ACM SIGARCH Computer Architecture News*, **44**(3), 243–254. https://doi.org/10.1145/3007787.3001163.

Han, S., Mao, H. and Dally, W. J. (2015), "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding", arXiv preprint arXiv:1510.00149. https://doi.org/10.48550/arXiv.1510.00149.

Haque, M. E., Asikuzzaman, M., Khan, I. U., Ra, I.-H., Hossain, M. S. and Shah, S. B. H. (2020), "Comparative study of IoT-based topology maintenance protocol in a wireless sensor network for structural health monitoring", *Remote Sensing*, **12**(15), 2358. https://doi.org/10.3390/rs12152358.

Hassibi, B. and Stork, D. (1992), "Second order derivatives for network pruning: Optimal brain surgeon",

Advances in Neural Information Processing Systems, **5**.
He, K., Zhang, X., Ren, S. and Sun, J. (2015), "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification", *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034, Santiago, Chile, December.
He, K., Zhang, X., Ren, S. and Sun, J. (2016), "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778, Las Vegas, USA, June.
He, Y., Zhang, X. and Sun, J. (2017), "Channel pruning for accelerating very deep neural networks", *Proceedings of the IEEE International Conference on Computer Vision*, 1389–1397, Venice, Italy, October.
Hou, S. and Wu, G. (2019), "A low-cost IoT-based wireless sensor system for bridge displacement monitoring", *Smart Materials and Structures*, **28**(8), 085047. 10.1088/1361-665X/ab2a31.
Hu, H., Peng, R., Tai, Y.-W. and Tang, C.-K. (2016), "Network trimming: A data-driven neuron pruning approach towards efficient deep architectures", arXiv preprint arXiv:1607.03250. https://doi.org/10.48550/arXiv.1607.03250.
Huang, Z. and Wang, N. (2018), "Data-driven sparse structure selection for deep neural networks", *Proceedings of the European Conference on Computer Vision (ECCV)*, 304–320, Munich, Germany, September.
Huang, Y. T., Jahanshahi, M. R., Shen, F. and Mondal, T. G. (2023). "Deep Learning–Based Autonomous Road Condition Assessment Leveraging Inexpensive RGB and Depth Sensors and Heterogeneous Data Fusion: Pothole Detection and Quantification", *Journal of Transportation Engineering, Part B: Pavements*, **149**(2), 04023010. https://doi.org/10.1061/JPEODX.PVENG-1194.
Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H. and Kalenichenko, D. (2018), "Quantization and training of neural networks for efficient integer-arithmetic-only inference", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2704–2713, Salt Lake City, USA, June.
LeCun, Y., Denker, J. and Solla, S. (1989), "Optimal brain damage", *Advances in Neural Information Processing Systems*, **2**.
Li, H., Kadav, A., Durdanovic, I., Samet, H. and Graf, H. P. (2016), "Pruning filters for efficient convnets", arXiv preprint arXiv:1608.08710. https://doi.org/10.48550/arXiv.1608.08710.
Liberty, E. (2013), "Simple and deterministic matrix sketching", *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 581–588, Chicago, USA, August.
Lin, M., Cao, L., Li, S., Ye, Q., Tian, Y., Liu, J., Tian, Q. and Ji, R. (2021), "Filter sketch for network pruning", *IEEE Transactions on Neural Networks and Learning Systems*. 10.1109/TNNLS.2021.3084206.
Lin, M., Ji, R., Li, S., Wang, Y., Wu, Y., Huang, F. and Ye, Q. (2021), "Network pruning using adaptive exemplar filters", *IEEE Transactions on Neural Networks and Learning Systems*. 10.1109/TNNLS.2021.3084856.
Lin, M., Ji, R., Wang, Y., Zhang, Y., Zhang, B., Tian, Y. and Shao, L. (2020), "HRank: filter pruning using high-rank feature map", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1529–1538, June.
Lin, S., Ji, R., Yan, C., Zhang, B., Cao, L., Ye, Q., Huang, F. and Doermann, D. (2019), "Towards optimal structured CNN pruning via generative adversarial learning", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2790–2799, Long Beach, USA, June.
Liu, G., Niu, Y., Zhao, W., Duan, Y., and Shu, J. (2022), "Data anomaly detection for structural health monitoring using a combination network of GANomaly and CNN", *Smart Structures and Systems*, **29**(1), 53-62. https://doi.org/10.12989/sss.2022.29.1.053.
Liu, Z., Mu, H., Zhang, X., Guo, Z., Yang, X., Cheng, K.-T. and Sun, J. (2019), "Metapruning: Meta

learning for automatic neural network channel pruning", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3296– 3305, Seoul, Korea, October.

Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S. and Zhang, C. (2017), "Learning efficient convolutional networks through network slimming", *Proceedings of the IEEE International Conference on Computer Vision*, 2736–2744, Venice, Italy, October.

Luo, J.-H., Wu, J. and Lin, W. (2017), "ThiNet: a filter level pruning method for deep neural network compression", *Proceedings of the IEEE International Conference on Computer Vision*, 5058–5066, Venice, Italy, October.

Martakis, P., Movsessian, A., Reuland, Y., Pai, S. G., Quqa, S., Cava, D. G., Tcherniak, D. and Chatzi, E. (2021), "A semi-supervised interpretable machine learning framework for sensor fault detection", *Smart Structures and Systems An International Journal*, **29**, pp. 251-266. https://doi.org/10.12989/sss.2022.29.1.251.

Mishra, M., Louren¸co, P. B. and Ramana, G. V. (2022), "Structural health monitoring of civil engineering structures by using the internet of things: A review", *Journal of Building Engineering*, **48**, 103954. https://doi.org/10.1016/j.jobe.2021.103954.

Mondal, T. G. (2021). Development of Multimodal Fusion-Based Visual Data Analytics for Robotic Inspection and Condition Assessment, Doctoral dissertation, Purdue University.

Nur, A. B. S. (2022). Deep neural network compression for artificial intelligent of things, Final Year Project Report, Nanyang Technological University.

Park, J., Li, S., Wen, W., Tang, P. T. P., Li, H., Chen, Y., and Dubey, P. (2016), "Faster CNNs with direct sparse convolutions and guided pruning", arXiv preprint arXiv:1608.01409. https://doi.org/10.48550/arXiv.1608.01409.

Peng, C., Fu, Y. and Spencer, B. F. (2017), "Sensor fault detection, identification, and recovery techniques for wireless sensor networks: A full-scale study", *In Proceedings of the 13th International Workshop on Advanced Smart Materials and Smart Structures Technology*, pp. 22-23, Tokyo, Japan, July.

Peralta Abadia, J., Fritz, H., Dragos, K. and Smarsly, K. (2022), "Sensor fault diagnosis coupling deep learning and wavelet transforms", *13th International Workshop on Structural Health Monitoring, IWSHM 2021*, 779–785, Standford, USA, September.

Shajihan, S. A., Wang, S., Zhai, G. and Spencer Jr, B. F. (2022), "CNN based data anomaly detection using multi-channel imagery for structural health monitoring", *Smart Structures and Systems*, **29**(1), 181–193. https://doi.org/10.12989/sss.2022.29.1.181.

Singh, P., Verma, V. K., Rai, P. and Namboodiri, V. P. (2019), "Play and prune: Adaptive filter pruning for deep model compression", arXiv preprint arXiv:1905.04446. https://doi.org/10.48550/arXiv.1905.04446.

Sun, Y., Zheng, L., Deng, W. and Wang, S. (2017), "SVDNet for pedestrian retrieval", *Proceedings of the IEEE International Conference on Computer Vision*, 3800–3808, Venice, Italy, October.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015), "Going deeper with convolutions", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9, Boston, USA, June.

Tang, Z., Chen, Z., Bao, Y., and Li, H. (2019), "Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring", *Structural Control and Health Monitoring*, **26**(1), e2296. https://doi.org/10.1002/stc.2296.

Wen, W., Wu, C., Wang, Y., Chen, Y. and Li, H. (2016), "Learning structured sparsity in deep neural networks", *Advances in Neural Information Processing Systems*, **29**.

Wu, R.-T., Singla, A., Jahanshahi, M. R., Bertino, E., Ko, B. J., and Verma, D. (2019), "Pruning deep convolutional neural networks for efficient edge computing in condition assessment of infrastructures", *Computer-Aided Civil and Infrastructure Engineering*, **34**(9), 774–789.

https://doi.org/10.1111/mice.12449.
Xu, J., Dang, D., Ma, Q., Liu, X. and Han, Q. (2022), "A novel and robust data anomaly detection framework using LAL-AdaBoost for structural health monitoring", *Journal of Civil Structural Health Monitoring*, **12**(2), 305–321. https://doi.org/10.1007/s13349-021-00544-2.
Yang, K., Jiang, H., Diang, Y., Wang, M. and Wan, C. (2022), "Data abnormal detection using bidirectional long-short neural network combined with artificial experience", *Smart Structures and Systems*, **29**(1), 117–127. https://doi.org/10.12989/sss.2022.29.1.117.
Yu, R., Li, A., Chen, C.-F., Lai, J.-H., Morariu, V. I., Han, X., Gao, M., Lin, C.-Y. and Davis, L. S. (2018), "NISP: pruning networks using neuron importance score propagation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9194–9203, Salt Lake City, USA, June.
Zhang, Z., Yan, J., Li, L., Pan, H. and Dong, C. (2021), "Condition assessment of stay cables through enhanced time series classification using a deep learning approach", *Smart Structures and Systems*, **29** (1), 105-116. https://doi.org/10.48550/arXiv.2101.03701.
Zhao, C., Ni, B., Zhang, J., Zhao, Q., Zhang, W. and Tian, Q. (2019), "Variational convolutional neural network pruning", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2780–2789, Long Beach, USA, June.