Fault-Tolerant 1-bit Representation for Distributed Inference Tasks in Wireless IoT

Mostafa Hussien[1,2], Kim Khoa Nguyen[2], and Mohamed Cheriet[2]
[1] Resilient Machine-learning Institute (ReMI)
[2] École de technologie supérieure (ÉTS), Univeristy of Québec
Montréal, QC, Canada

***Abstract*—In IoT applications, the sensors usually have limited bandwidth and power resources. Therefore, the sensed data should be mapped to a low-bit representation by means of compression and quantization before being transmitted to a central node, called the fusion center (FC). At the FC, a global decision is inferred from this data. In many cases, this data is intended for machine consumption, not for human perception. However, the compression techniques are mainly designed for reconstruction fidelity. The accuracy of the inferred decision at the FC is less considered. In this work, we present an end-to-end framework for learning a 1-bit representation of correlated-sensors data. We also propose a novel loss function and a three-stage training algorithm for learning discriminative binary features at each sensor. Extensive experiments show the proposed framework achieves high compression ratios with a marginal loss in the inferred decision accuracy. Comparatively, the obtained results outperform other benchmark models in the literature.**

***Index Terms*—Deep learning, wireless sensor networks, distributed inference.**

I. Introduction

Many wireless Internet of things (IoT) applications employ a distributed inference mechanism such as radar systems, surveillance systems, or multi-sensory human activity recognition [1]. In these cases, the sensed data are not entirely processed locally at each node. However, the data is offloaded to a central, and more powerful, node called fusion center (FC), where a global decision is inferred from this data [2]. The main challenge of this centralized scenario is the limited bandwidth and power resources of the nodes [3], [4]. Therefore, the observations should be compressed and quantized before transmission, see Fig. 1. These compression and quantization processes introduce a loss in the transmitted information, and the FC infers the decision based on partial information [5].

We tackle the problem of compressing and quantizing correlated-sensor observations for distributed inference tasks. While most of the literature assumes sensors independence for mathematical tractability, we address the more complicated scenario of correlated sensors. We argue that this correlation can be exploited to obtain higher compression ratios without considerable loss in the accuracy of the inferred decision. More specifically, we distributively screen redundancies in sensor observations to transmit only informative data without imposing any assumptions on the distribution of the observations. In this case, analytical solutions are intractable and hard to generalize. On the other hand, data driven solutions are achieving outstanding results in different domains and applications [6].

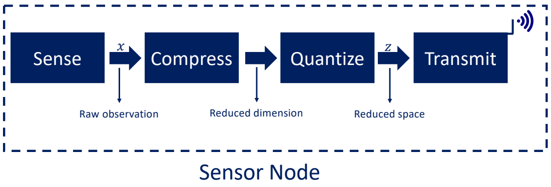

Fig. 1: The different operations inside each sensor.

Motivated by the breakthrough in statistical learning techniques, in this work, we propose a novel deep learning framework that learns to compress and quantize the observations of each sensor. In addition, the framework is jointly trained with the decision rule at the FC in an end-to-end fashion to maximize the accuracy of the inferred decision. We also propose a new loss function that helps the sensors learn a task-aware 1-bit representation for the observations. A three-stage training algorithm to train the framework is proposed. Extensive results show the superiority of our proposed framework compared with different bench-mark models. The results also confirm the robustness of the framework against nodes failure or significant delay caused by high-latency channels.

The main contributions of this work can be summarized as follows:

1) Extending autoencoders to learn a compressed and quantized representation for correlated-sensor observations. This learned representation conveys the complementary features at each sensor observation, which help maximize the likelihood of the correct decisions while communicating efficiently.
2) Proposing a powerful loss function that encourages the model to learn the unique features at each sensor. Furthermore, we present a training algorithm that efficiently works in a wide range of applications.

The rest of this paper is organized as follows: section II, provides the problem formulation. Section III presents the implementation details. The experimental results are given in section IV while section V concludes the work.

II. Problem Formulation

Suppose Y is a discrete random variable, representing a hypothesis about an environment. The variable takes values in: $y \in \{c_1, c_2, \dots, c_C\}$ where C is the number of possible

978-3-903176-36-2 ©2021 IFIP

hypothesis or classes. Our goal is to form an estimate, $\hat{Y}$, for the true hypothesis, based on observations collected from a set of S sensors. Accordingly, for each $t = 1, \dots, S$, let x^t represents the observation at node t, where $x^t \in \chi$ belongs to $\mathcal{R}^d$ and χ represents the observation space.

The set of all observations correspond to an S-dimensional random vector $X = (x^1, x^2, \dots, x^S) \in \chi^S$ drawn from the conditional distribution $P(X|Y)$. We assume that an optimal estimate, $\hat{Y}$, is computed at the FC. If the FC has access to the distribution of the observations, $P(X|Y)$, then an optimal decision rule can be easily formulated. For example, with a binary hypothesis, an optimal decision rule can be reached using the likelihood ratio test: $P(X|Y = 1)/P(X|Y = -1)$. However, in real-world problems, the FC does not have access to this distribution, and it only has access to summarized forms of the original observations, z^t, for all values of t. More specifically, we assume that each sensor, t, is restricted to a given bandwidth of R bps. Therefore, each sensor is allowed to transmit an *n-dimensional* message, z^t, taking values in some space $Z = \{0, 1\}$, such that $n \leq R$. The conversion from the observation space, χ, to Z-space is carried out by an encoder $q : \chi \rightarrow Z$. The encoder, q, maps an input observation, x, in χ-space, to a codeword, z, in Z-space. This encoded observation, z, will be sent to the FC. To compute the estimate $\hat{Y}$, the FC applies a certain decision rule, ψ, on the aggregated received messages such that $\hat{Y} = \psi(z^1, z^2, \dots, z^S)$. It is known from the rate-distortion theory that the rate, R, and the distortion are inversely proportional. Therefore, a larger rate implies better reconstruction fidelity at the receiver end. However, in our problem, we are not concerned about reconstruction fidelity as our main objective. Instead, we are more interested in maximizing the accuracy of the inferred decision. Inherently, increasing the rate, R, will increase the information included in a message, z^t, which increases the FC accuracy. In other words, increasing the rate increases the mutual information between the joint distributions $P(\hat{Y}|Z)$ and $P(\hat{Y}|X)$. However, for limited bandwidth systems, we can not violate the bandwidth constraint. Therefore, for correlated sensor observations, the redundancy between the different sensor observations could be exploited to obtain more efficient compression with a marginal reduction in the inferred decision accuracy at the FC. This can be formalized by minimizing the function in (1).

$$\begin{aligned} \min_{\theta, \phi_i} \quad & KL\left(P(\hat{Y}|X)\,||\,P(\hat{Y}|Z)\right) \\ \text{s.t.} \quad & P(\hat{Y}|X) = f_\omega\left(x_1, x_2, \dots, x_S\right), \\ & P(\hat{Y}|Z) = f_\theta\left(f_{\phi_1}(x_1), f_{\phi_2}(x_2), \dots, f_{\phi_S}(x_S)\right), \\ & f_{\phi_i} \in \{0,1\}^n \quad \forall i \in \{1, 2, \dots, S\}, \\ & n \leq R \end{aligned} \tag{1}$$

θ is the parameters of the decision rule at the FC, ϕ_i is the encoder parameters at the i^{th} sensor, and R is the bandwidth assigned to each sensor. Substituting the KL term in (1) by (2), we get the objective function in (3).

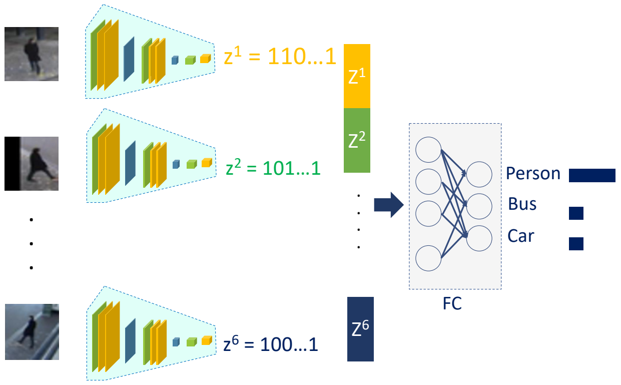

Fig. 2: Diagram demonstrating the system model. In this example, the cameras represent the sensors, observations are the images, and the decision is the predicted class.

$$KL(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right) \tag{2}$$

$$\begin{aligned} \min_{\theta, \phi_i} \quad & \sum_i P(\hat{Y}_i|X_i) \log\left(\frac{P(\hat{Y}_i|X_i)}{P(\hat{Y}_i|Z_i)}\right) \\ \text{s.t.} \quad & P(\hat{Y}|X) = f_\omega\left(x_1, x_2, \dots, x_S\right), \\ & P(\hat{Y}|Z) = f_\theta\left(f_{\phi_1}(x_1), f_{\phi_2}(x_2), \dots, f_{\phi_S}(x_S)\right), \\ & f_{\phi_i} \in \{0,1\}^n \quad \forall i \in \{1, 2, \dots, S\}, \\ & n \leq R \end{aligned} \tag{3}$$

Two main points should be considered in this formulation. First, the message space $\{0, 1\}$ is significantly smaller than the observation space $\mathcal{R}$. Second, the dimension for the compressed message, n, is substantially smaller than that of the raw observation, d (i.e., $n \ll d$). Therefore, the problem is finding an optimal quantizer q: $q(x^t) = z^t$ for each sensor, t, that maximizes the mutual information between the conditional distributions of the true and estimated hypothesis $P(Y|X)$ and $P(\hat{Y}|Z)$ under a certain communication constraint $n \leq R$.

III. Proposed Framework

To learn an optimal quantizer at each sensor, q^t: $q^t(x_i) = z_i$, and an optimal decision rule at the FC $\psi(z^1, z^2, \dots, z^S)$, we adopt a low-bit encoder (namely 1-bit encoder) at each sensor node to compress and quantize sensor observations, see Fig. 2. Its worth to differentiate between compression and quantization in this context. The compression means the mapping from a higher-dimensional to a lower-dimensional space, $f : \mathbb{S}^d \rightarrow \mathbb{S}^n$, where $n \ll d$. On the other hand, quantization is mapping the values of individual dimensions from a set $\mathbb{S}_1$ to a set $\mathbb{S}_2$, where the cardinality of $\mathbb{S}_1$ is smaller than that of $\mathbb{S}_2$, i.e., $|\mathbb{S}_1| < |\mathbb{S}_2|$. The output of the encoder model at sensor i is given by: $f_{\phi_i}(\cdot)$, where ϕ_i is the encoder parameters at sensor i. At the FC, a multilayer perceptron (MLP) architecture parameterized by parameters, θ, is employed to approximate the optimal decision rule. The decision rule at the FC is given by:

$$f_\theta([f_{\phi_1}(x^1), f_{\phi_2}(x^2), \ldots, f_{\phi_S}(x^S)]). \tag{4}$$

where x^i is the current observation at sensor, i.

A. Implementation Details

The encoder architecture at each sensor is an MLP of three fully-connected layers with *ReLU* activations. In the output layer of the encoder, a *QSigmoid* activation is used to generate binary activations [7]. In the FC, we used six fully connected layers with *ReLU* activations in the hidden layers and *Sigmoid* activation in the output layer. The model weights are initialized using *He* initializer [8]. The models are trained using *Adam* optimizer [9], with (0.01) learning-rate and optimized to minimize the *crossentropy* loss function (5).

$$f(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i), \tag{5}$$

where y_i is the target label of the i^{th} data point. We can interpret the optimization of the classifier weights at the FC as a threshold optimization for the decision rule.

B. Training Procedure

The training of the proposed framework consists of three phases. In the first phase, we train an autoencoder at each sensor. The autoencoders are classically trained to reconstruct their input from compressed codewords by minimizing the $L2$-norm between the input and the reconstruction given in (6).

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2 \tag{6}$$

In the second phase, we train an inference model, I_1, that takes as input the raw observations, X = $[x_1, x_2, \ldots, x_S]$, and outputs the corresponding decision. Note that the inputs to this model are the raw observations without either compression or quantization. The model is trained to optimize the classical Crossentropy function (5). The output of the model I_1 approximates the conditional distribution $P(\hat{Y}|X)$. This model represents our baseline where the FC has access to the full information. The weights of this model I_1 are then frozen, and its output will be used for computing the loss value of the inference model at the FC.

In the third and last phase, we use the encoder part of the autoencoder at each sensor to compress the observation at this sensor. The output of the encoder at the i^{th} sensor is denoted by z_i. The parameters of an encoder model of the autoencoder at sensor, i, denoted by ϕ_i, such that $z^i = f_{\phi_i}(x)$. The outputs of all encoders are concatenated and fed to an inference model, I_2, with parameters θ to infer the global decision. The output of I_2 in this case approximates the conditional distribution $P(\hat{Y}|Z)$. Note that the weights of I_2, θ, are trained jointly with the encoder's weights, ϕ_i, at each sensor. Algorithm 1 summarizes the training procedure.

C. Proposed Loss function

In the first and second phases of the training, we optimize the MSE and Crossentropy loss functions, respectively. However, for the third phase in which we jointly train the encoders and the inference model at the FC, minimizing the traditional Crossentropy is found to be limited in solving the complex problem being addressed. Recall from the previous section that the objective of the proposed framework is to make the encoders benefit from the redundancies (between the sensor observations) to obtain high compression ratios without considerable reduction in the decision accuracy (i.e., the sensing goal). This implies that encoders should learn to encode the complementary features of their observation.

Algorithm 1: The training procedure for the proposed framework, S, sensors.

```
Input: Dataset $D$, consisting of $N$ tuples of
       observations acquisted from $S$ sensors and the
       corresponding label $y$.
Output: Model parameters, $\theta$, and $\phi_i$ for
        $i \in \{1, 2, \ldots, S\}$
  - At each sensor, $s_i$, train an autoencoder to reconstruct
    its input using observations in D;
  - Train an inference model, $I_1$, to approximate the
    conditional distribution $p(\hat{y}|X)$;
  - Freeze the weights of $I_1$;
  - Train an inference model, $I_2$, (jointly with the
    encoders weights) to approximate the conditional
    distribution $p(\hat{y}|Z)$;
```

$$\mathcal{L}(Y, \hat{Y}) = CE(Y, \hat{Y}) + KL(P(\hat{Y}|X)||P(\hat{Y}|Z)) \tag{7}$$

To this end, we propose a novel loss function given in (7). The proposed function helps the model to learn a conditional distribution for the decision given the compressed observations, $P(\hat{Y}|Z)$, which is as similar as possible to the conditional distribution for the label given the uncompressed observations, $P(\hat{Y}|X)$. This term reduces the loss in the decision accuracy due to the compression of the sensor observations. Given the limited budget of data bits to encode the observations, we argue the proposed function encourages the encoders to encode only the relevant features that help in maximizing the likelihood of the correct decision at the FC.

IV. Results and Discussion

A. Distributed Inference Accuracy

In this subsection, we show the results of the inferred decision accuracy using our proposed compression and quantization technique.

1) Comparative Evaluation

To evaluate the effectiveness of the proposed framework, we used a publicly available dataset called *Wearable Action Recognition Database (WARD)* presented in [10]. The obtained performance is compared against three other literature works applied to the same dataset. A data point represents the readings of five accelerometer/gyroscope pairs. Each point belongs to one of 13 different classes. For more information we refer to [10].

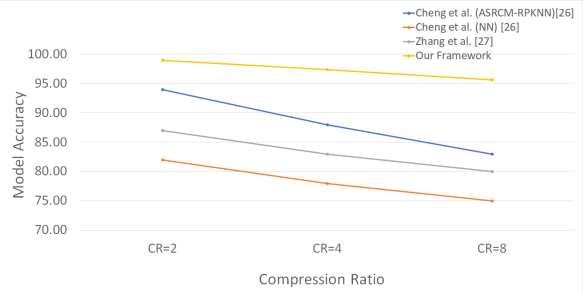

Fig. 3: Comparison of model accuracy under different compression ratios.

Fig. 3 shows a comparison between the performance of the proposed framework and other works in the literature under different compression ratios. We can see from the figure that the performance of our framework outperforms other works under all compression ratios.

Table I shows the classification accuracy of the proposed framework compared with the accuracy of other literature works. The table reports results for Yang et al. [10], Huynh [11], He et al. [12] with different dimensionality reduction techniques, and Guo et al. [13] with different fusion methods. It is clear from the table that the proposed framework achieves state-of-the-art accuracy compared with other works. In addition, the proposed framework respects the bit rate constraint assigned for each sensor, R, which highly contributes to power-saving and sensor lifetime.

2) *A Toy Problem*

We tested the proposed framework with four datasets: 1) MNIST, 2) Fashion-MNIST, 3) Street View Houses (SVH), 4) CIFAR-10. For each dataset, we used different Compression Ratios, CR. CR is defined as the ratio between the uncompressed dimension and compressed dimension [14]. It is worth noting that the compression ratios of other methods consider compression by dimensionality reduction only (i.e., any input or output dimension $\in \mathcal{R}$). Based on that, the input and output space remains the same. Unlike these methods, our work goes beyond to count for the quantization (since the input dimension is $\in \mathcal{R}$ while the output dimension is quantized $\in \{0, 1\}$).

TABLE I: The classification accuracy of the proposed framework compared with different work from the literature on WARD dataset.

Method	**Detection Accuracy**
Yang et al. [10]	93.6%
Huynh [11]	96.97%
He et al. + PCA [12]	76.31%
He et al. + LDA [12]	40.3%
He et al. + GDA [12]	99.2%
Guo (Majority voting) [13]	94.96%
Guo (Maximum) [13]	96.20%
Guo (WLOP) [13]	98.02%
Guo (WLOGP) [13]	98.78%
Our Framework (CR=2)	**99.7%**

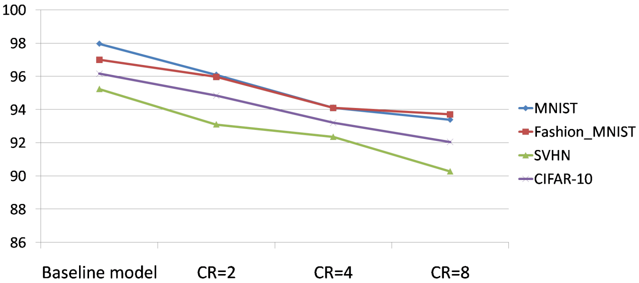

Fig. 4: The decision accuracy of the inferred decision under different compression ratios. The baseline represents sending the raw observations to the FC without compression.

In these experiments, we simulate two sensors (s_1,s_2) sending their data to a FC. Assume the observations at sensor s_1 belongs to a class C_i and at sensor s_2 belongs to a class C_j. The decision rule at the FC can be defined as:

$$\psi(z^1, z^2) = \begin{cases} i & if i = j \\ -1 & if i \neq j \end{cases}$$

In other words, the decision will be the class label if the two observations belong to the same label, and -1 otherwise. Since each dataset consists of images belonging to one out of 10 total classes, we expect the classifier to have 11 classes.

To make a fair comparison between the framework accuracies with different compression ratios, we used the same classifier capacity (in terms of the number of layers, the nodes in each layer, the activation functions used, etc.) for each dataset.

We compared the obtained results with the baseline model accuracy. The baseline model is defined as the accuracy of a neural network classifier taking as input the raw observations without compression or quantization, x^t. In this case, the FC has the complete vector of sensed data, which represents the optimal case in terms of the data availability at the FC.

Fig. 4 shows the obtained results for each case. We can see that the performance of our framework approaches the baseline model with the lowest compression ratio, $CR = 2$, in the table. A small reduction in the accuracy has been reported with higher CR (i.e., $CR = 4$ and 8). However, we still obtain high accuracy even with the highest compression ratio, $CR = 8$. For example, we obtained 95.3% of the baseline with $CR = 8$ in MNIST dataset. This means compressing the observations to only 12.5% of its original dimension with quantization results in only 4.7% reduction in accuracy.

B. *Fault-Tolerance Evaluation*

One powerful characteristic behind any distributed system is the high fault tolerance. Fault tolerance means the system keeps working (maybe with a little performance degradation) even when some of the nodes are down. Due to the learning nature of our proposed framework and the proposed training algorithm, our framework has a high fault tolerance which makes it resilient against node failures. To evaluate this property, we used a multi-view dataset presented in [15]. The

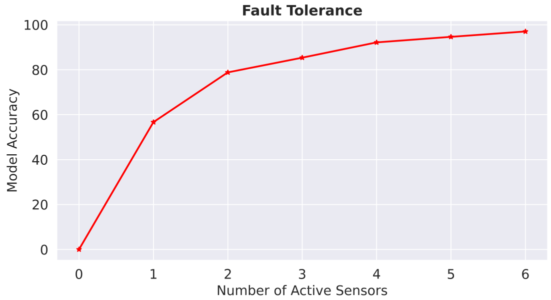

Fig. 5: Evaluating the fault-tolerance by monitoring the model accuracy as the number of failed sensors increase.

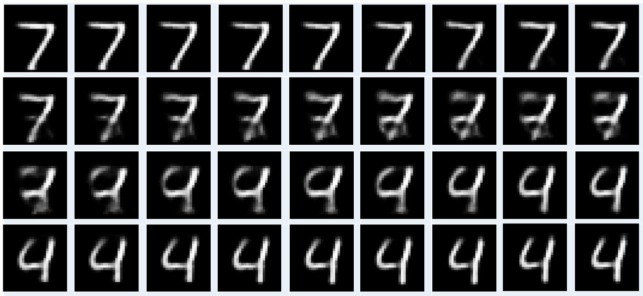

Fig. 6: Interpolation between two points in the latent space.

dataset contains images captured from six different synchronized cameras and three classes (person, bus, and car).

Each time, we randomly select to deactivate one source and monitor the degradation in the framework accuracy. Then, we randomly deactivate two sources and monitor the accuracy, and so on. The results are shown in Fig. 5 where we can see that the framework maintains a high accuracy (around 80%) even when four cameras, out of six, cameras were in failure. We can see around 2.4% reduction in the achieved accuracy if only one sensor is down. These results suggest that our system is robust against multiple node failures.

C. Semantics of the Latent Representation

In autoencoder-based architectures for dimensionality reduction, a special interest is paid to the robustness of the learned codewords in the latent space [16]. To evaluate the robustness of such codewords, we interpolate between different points in the latent space and observe, qualitatively, the gradual changes in the reconstructed data. This widely used experiment verifies that the model: (a) has injected enough redundancies into the codewords and consequently the model is capable of reconstructing the input even in the presence of errors in the codeword, (b) has learned relevant features of the underlying structure of the data.

We randomly select two test points to represent the start and endpoints. In each step, we flip a bit in the latent codeword, fed the newly obtained codeword to the decoder model, and observe the gradual changes in the reconstruction. Fig. 6 shows the gradual transition in the digit shape with the gradual bit flipping. We can observe that decrementing the hamming distance between the start and endpoints, each bit-flip, slowly alters the characteristic features of the digit until the digit reaches the endpoint.

V. Conclusion

In this paper, we proposed a learning framework for task-aware compression of correlated-sensors data. The framework learns complementary binary features for sensor data to maximize the accuracy of the inferred decision at the FC. We proposed a powerful loss function and a three-stage training algorithm to maximize the framework's accuracy. Each training phase approaches better accuracy by incrementally learning more resilient unique features. By learning a low-bit representation for sensor data, we minimize the consumed bandwidth and the power consumption for each sensor. Moreover, our framework is a high fault-tolerant due to the proposed training algorithm. Extensive experiments confirm the superiority of our proposed framework in terms of achievable compression ratios, accuracy, and fault tolerance.

Acknowledgment

The authors thank Mitacs and Ciena for supporting this research in the IT13947 grant.

References

[1] Y. Abdi and T. Ristaniemi, "The max-product algorithm viewed as linear data-fusion: A distributed detection scenario," *IEEE Trans on Wireless Communications*, vol. 19, no. 11, pp. 7585–7597, 2020.

[2] S. Li and X. Wang, "Distributed sequential hypothesis testing with quantized message-exchange," *IEEE Trans on Information Theory*, vol. 66, no. 1, pp. 350–367, 2019.

[3] G. Katz and et al., "Distributed binary detection with lossy data compression," *IEEE Trans on Information Theory*, vol. 63, no. 8, pp. 5207–5227, 2017.

[4] M. Hussien and et al., "Towards more reliable deep learning-based link adaptation for WiFi 6," in *ICC*. IEEE, 2021.

[5] S. Salehkalaibar and et al., "On hypothesis testing against conditional independence with multiple decision centers," *IEEE Trans on Communications*, vol. 66, no. 6, pp. 2409–2420, 2018.

[6] M. Hussien and et al., "Prvnet: variational autoencoders for massive MIMO CSI feedback," *arXiv preprint arXiv:2011.04178*, 2020.

[7] B. Moons and et al., "Minimum energy quantized neural networks," in *Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2017.

[8] K. He and et al., "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *IEEE ICCV*, 2015.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[10] A. Y. Yang and et al., "Distributed recognition of human actions using wearable motion sensor networks," *Journal of Ambient Intelligence and Smart Environments*, vol. 1, no. 2, pp. 103–115, 2009.

[11] D. T. G. Huynh, "Human activity recognition with wearable sensors," Ph.D. dissertation, Technische Universität, 2008.

[12] W. He and et al., "Recognition of human activities with wearable sensors," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 108, 2012.

[13] Y. Guo and et al., "Human activity recognition by fusing multiple sensor nodes in the wearable sensor systems," *Journal of Mechanics in Medicine and Biology*, vol. 12, no. 05, p. 1250084, 2012.

[14] K. Sayood, *Introduction to data compression*. Morgan Kaufmann, 2017.

[15] Teerapittayanon and et al., "Distributed deep neural networks over the cloud, the edge and end devices," in *ICDCS*. IEEE, 2017, pp. 328–339.

[16] K. Choi and et al., "Neural joint source-channel coding," in *ICML*, 2019.