A Learning Framework for Bandwidth-Efficient Distributed Inference in Wireless IoT

Mostafa Hussien, Graduate Student Member, IEEE, Kim Khoa Nguyen, Member, IEEE, and Mohamed Cheriet, Senior Member, IEEE

Abstract—The limited bandwidth and power resources of wireless sensors in distributed environments have resulted in new challenges in handling the ever-growing volume of transmissions generated by the Internet-of-Things (IoT) applications. To overcome these challenges, each sensor should compress and quantize its observations before sending them to a fusion center (FC) for global decision inference. Unfortunately, most of the conventional compression techniques and entropy quantizers only focus on reconstruction fidelity as a performance measure, neglecting the sensing goal. In this article, we propose a joint design of compression mechanisms and entropy quantizers with the sensing goal of machine-to-machine (M2M) communications. We define a deep learning-based framework for compressing and quantizing observations from correlated sensors. Unlike traditional methods, our proposed method not only maximizes the reconstruction fidelity but also optimally compresses sensor observations in terms of the accuracy of the inferred decision (i.e., the sensing goal) at the FC. The proposed framework is widely applicable as it does not impose any assumptions on observation distribution. Furthermore, a novel loss function has been proposed to focus on learning complementary features at each sensor. Our experimental results demonstrate that the framework outperforms other benchmark models.

Index Terms—Data compression, deep learning, distributed inference, sensor networks, wireless communications.

NOMENCLATURE

| Symbol | Description |
|--------|-------------|
| $x^i$  | Current observation of the $i$th sensor. |
| $z^i$  | Compressed and quantized representation for the current observation at the $i$th sensor. |
| $\phi_i$ | Encoder parameters of the $i$th sensor. |
| $f_{\phi_i}$ | Encoder function at the $i$th sensor given by a neural network parameterized by parameters $\phi_i$. |
| $y_j$ | Label of the $j$th data point. |
| $\hat{y}_j$ | Predicted label of the $j$th data point. |
| $\theta$ | Parameters of the decision function (i.e., decision rule) at the FC. |
| $\omega$ | Parameters of the decision function given the raw-observations. |
| $S$ | Total number of sensors. |
| $C$ | Number of possible classes (i.e., decisions) to be predicted at the FC. |
| $d$ | Dimension of the raw observations. |
| $n$ | Dimension of the compressed and quantized observations. |
| $R$ | Bandwidth (in bps) assigned to each sensor. |
| $\chi$ | Observation space, $\mathbb{R}^d$. |
| $Z$ | Latent space, $\{0, 1\}^n$. |
| $S_n$ | $n$-dimensional vector where each element belongs to the set $S$. |
| CE($\cdot$) | Crossentropy loss function, given in (7). |
| KL($\cdot$) | KL-divergence loss given in (3). |

I. INTRODUCTION

NUMEROUS applications of wireless Internet of Things (IoT) utilize a mechanism of distributed inference such as radar systems, multiview surveillance systems, and multisensory human activity recognition. For example, in...
multisensory human activity recognition, a human wears multiple, spatially distributed, sensors (e.g., gyroscope and accelerometer). A decision about human activity (e.g., walking, running, etc.) is inferred from the received sensor signals. In such a scenario, if each sensor considered only its local observations for inferring a decision, the error probability would be much higher compared to the scenario in which a global decision is inferred from the aggregated sensor data [1].

To tackle this problem, a spatially distributed setup may be employed in which the sensed data (a.k.a, environment observations) are sent to a central node called fusion center (FC). The FC infers a global decision based on the aggregated data received from all sensors. However, the sensors usually have limited power and bandwidth resources. For example, each sensor may have a fixed data rate of $R$ bps. Therefore, it should compress and quantize its sensed observation to fit the assigned bit rate. The FC, then, performs a specific inference task (i.e., the sensing goal). However, the FC infers the decision from only partial information due to the compression and quantization steps. This may reduce decision accuracy at the FC [1]. Optimally processing the observations at each sensor can minimize the degradation of the decision accuracy [2]. For conditionally independent sensor observations, an optimal decision can be easily reached using Bayesian inference theory [3]. However, the conditional-independence assumption does not hold for many real-world problems. In prior work, Chamberland and Veeravalli [4] and Tay et al. [5], assume that the statistical distribution of sensor observations is priorly known. In this case, the goal is to design an optimal decision rule that maximizes the likelihood of the correct decisions. However, in many practical applications, this distribution is not priorly known which increases the problem’s complexity. In this case, data-driven solutions provide practical and efficient alternatives.

Although different works in the literature propose compression and quantization techniques for sensor data, their goal was mainly obtaining a high-fidelity reconstruction at the FC. This seems relevant for human-consumed data such as images and videos. However, for machine-consumed data, adopting reconstruction fidelity as a metric is doubtful. Indeed, the accuracy of the inferred decisions is more crucial than having a good reconstruction.

This article aims to explore the issue of compressing and quantizing observations from correlated sensors in the context of distributed inference tasks. The main objective of our study is to maximize the accuracy of the inferred decision instead of minimizing the reconstruction loss. Prior research has generally assumed sensor independence for tractability of mathematical analyses; however, we tackle the more complex scenario of correlated sensors. We argue that this correlation can be utilized to achieve higher compression rates while preserving decision accuracy. Specifically, we propose and study transmitting each sensor’s unique features while avoiding the transmission of redundant features that other nodes in the network are likely to transmit. Our primary research question concerns the feasibility of using distributed redundancy screening of sensor observations to transmit only informative data without imposing any assumptions regarding the distribution of observations.

To address the question of whether distributive redundancy screening of sensor observations can be used to transmit only informative data, we leverage the recent developments in statistical learning techniques, particularly deep learning. Leveraging a novel deep-learning encoder, we compress and quantize observations at each sensor. This encoder is jointly trained with the decision rule at the FC to maximize inferred decision accuracy. End-to-end learning involves training a complex system by applying gradient-based learning to the system as a whole [6]. To enhance the sensors’ ability to learn decision-aware representations of observations, we propose a new loss function. We also present a training algorithm that efficiently trains the proposed framework. Our extensive experiments demonstrate the robustness and superiority of our framework compared with various benchmark models.

A. Related Work

Similar work in the literature has been proposed for specific problems. For example, a line of work has been proposed for the problem of human activity recognition [7], [8], [9], [10]. In this problem, the hypotheses are the different human actions, while the data come from multiple sensors fixed on the actor’s body (e.g., gyroscope, accelerometer, etc). Yang et al. [7] aimed to achieve a high action-classification accuracy with minimum bandwidth consumption. At each sensor, a decision is inferred from the local information. The FC then takes a global decision using a majority-voting mechanism. Although they obtained good results, this approach ignores any complementary information captured by other sensors. Another work has been proposed for the problem of earthquake detection from wireless IoT sensors’ network [11]. They presented a distributed approach for the rapid detection of earthquakes using cell phone accelerometers, consumer USB devices, and cloud computing-based sensor fusion. They proposed to learn a threshold for each sensor involved in the network in a way that maximizes the performance of the anomaly detection algorithm employed at the FC. Experimental results showed that this approach successfully distinguished between seismic motion and acceleration due to normal daily activities.

The work in [12] studied the problem of binary hypothesis testing with two observers, where the collected observations are assumed to be statistically correlated. To infer a decision, one of three solutions can be adopted. The first is a centralized solution in which the observations collected by both observers are sent to the FC. A global decision is inferred at the FC from the aggregated sensor observations. The main concern of this solution is the huge bandwidth incurred in fusing the raw observations to the FC. The second solution makes each sensor rely on its own locally collected observation. Then, each sensor exchanges its locally inferred decision with the other sensor to reach a global decision. The main limitation of this solution is that each sensor depends only on its local observation and ignores any complementary information captured by the other sensor. In the last solution, each sensor formulates the problem as a sequential hypothesis-testing problem.
In their work, Bouchoucha et al. [13] presented a framework aimed at utilizing the correlation between observations to minimize the mean square error of a distributed estimation problem. They proposed a framework in which nodes predict their future observations and send quantized prediction errors rather than quantized observations to the FC to accomplish this goal.

In the context of task-aware compression, a similar problem has been addressed in [14], [15], and [16]. For example, Chinchali et al. [14] used a reinforcement learning (RL) agent at each sensor to compress the observations before fusing them to the FC. The reward function at each agent considers its commitment to the assigned bandwidth. Although they achieved a good performance, there is a probability that the agent does not meet the bandwidth constraints after deployment. While Hu et al. [15] proposed Starfish, an image compression framework that outperforms JPEG by up to 3x in terms of bandwidth consumption and up to 2.5x in power consumption. Hu et al. [15] used an AutoML technique to search for tiny ML models that can work on power AoI accelerators.

We can summarize the limitations of the literature work, which we addressed in our work, as: 1) the conditional-independence assumption of the sensor observations is not always held; 2) the conditional-independence assumption ignores the potential opportunity to benefit from complementary features captured by different sensors; 3) the compression algorithms are designed independently of the sensing goal; and 4) the limited power of analytical-based techniques in dealing with a large number of possible decisions and correlated sensors.

1) Contribution: This article presents a novel deep learning-based compression framework for correlated-sensors data compression and quantization. Discrete-representation autoencoders (AEs) are adopted at each sensor to generate a compressed quantized representation of the observations. At the FC, a multi-layer perceptron (MLP) architecture is adopted to jointly learn the decision rule with the sensor encoders. The main contribution of this work comes in three folds.

1) Proposing an extension of AEs to efficiently learn compressed and quantized representations of correlated-sensor observations. The learned representation entails complementary features extracted from each sensor observation, thereby improving the likelihood of correct decisions at the FC, subject to a communication constraint. Furthermore, the representation is jointly learned with the decision rule at the FC in an end-to-end fashion to maximize decision accuracy.

2) Introducing a novel loss function to encourage the model to learn each sensor’s unique features. The function learns the soft probabilities of a knowledgeable model trained using the raw observations. Moreover, we present a training algorithm that efficiently works in a wide range of applications.

3) Eliminating the conditional-independence assumption between sensor observations which has been widely adopted for mathematical tractability. Moreover, we consider a multi-hypothesis problem, which is more complex and realistic than the simple binary hypothesis problem assumed in most of the literature work.

The article is organized as follows: Section II formulates the problem. In Section III, we describe the various elements of the proposed framework. The discussion and Section IV discusses the obtained results. In Section V, we conclude our work.

II. Problem Statement

Notation: Through this text, we refer to random variables by italic capital letters (e.g., X). Small letters refer to one realization of a random variable (e.g., x). Superscripts denote the sensor number. For example, x\textsuperscript{t} denotes the observation at sensor i. The observations are referred to by X while Y refers to the random variable of the labels (i.e., the target decisions at the FC). The parameters of the encoder at the ith sensor is referred to as \( \theta_i \). The parameter set of the decision rule at the FC is referred to by \( \omega \). The \log(\cdot) function uses a base of 2. Nomenclature summarizes the used symbols and notations. Through the text, we use the terms observations and data-points interchangeably. Moreover, the terms decision and sensing goal have the same meaning.

Suppose that there exists a discrete random variable, Y, that describes a hypothesis about a given environment. The values of the variable correspond to a finite set of classes, with \( y \in \{1, 2, \ldots, C\} \), where \( C \) represents the total number of possible hypotheses. Our objective is to create an estimate, denoted by \( \hat{Y} \), of the true hypothesis, based on a set of observations gathered from a set of \( S \) sensors. Let \( x^t \) represent the observation at sensor \( t \) for \( t = 1, \ldots, S \), where \( x^t \in \mathbb{R}^d \) within some observation space, \( \chi \). The collection of all observations is an \( S \)-dimensional random vector, \( X = (x^1, x^2, \ldots, x^S) \in \chi^S \), drawn from the joint distribution \( P(X, Y) \).

The primary objective of this study is to achieve an optimal estimate \( \hat{Y} \) for the true labels \( Y \) at an FC. If the FC has knowledge of the observation distribution, \( P(X, Y) \), it can easily formulate a decision rule, such as the likelihood ratio test for binary hypotheses, defined as \( P(X|Y = 1)/P(X|Y = -1) \). However, in practical scenarios, the FC does not have prior knowledge of the joint distribution \( P(X, Y) \). Rather, it has access to only compressed forms of observations, represented by \( z^t \), where \( t \in \{1, 2, \ldots, S\} \). Moreover, each sensor has a restricted bandwidth of \( \mathcal{R} \) bps and can transmit an \( n \)-dimensional message, represented by \( z^t \in \mathbb{Z}^n \), which takes values in a space \( Z \) such that \( n \leq \mathcal{R} \). An encoder, represented by \( q: x \to Z \), converts observations from the observation space, \( \chi \), to the \( Z \)-space. The encoder maps an input observation, denoted by \( x \), in \( \chi \)-space to a codeword, represented by \( z \), in \( Z \)-space, which is sent to the FC. To compute the estimate \( \hat{Y} \), the FC applies a decision rule, represented by \( g_\theta \), to the aggregated received messages, i.e., \( \hat{Y} = g_\theta(z^1, z^2, \ldots, z^S) \). According to the rate-distortion theory, the rate, denoted by \( \mathcal{R} \), and the distortion at the receiver, measured in terms of reconstruction loss, are inversely proportional. Hence, a higher rate, \( \mathcal{R} \), results in better reconstruction fidelity at the receiver. However, the main goal of our study is not to maximize reconstruction fidelity but rather to optimize the accuracy of the inferred decisions.

Inherent to the nature of communication systems, increasing the rate, \( \mathcal{R} \), of transmitted messages, \( z^t \), also increases the
amount of information conveyed, and therefore, enhances the accuracy of the FC. This is reflected in the mutual information, $I$, between the joint distributions $P(\hat{Y}|Z)$ and $P(\hat{Y}|X)$, which increases with higher rates. However, when operating within the constraints of limited bandwidth, the bandwidth cannot be increased beyond its assigned value, and each sensor must adhere to its allotted bandwidth. In such scenarios, exploiting the redundancy among the correlated sensor observations can enable more efficient compression while maintaining the decision accuracy at the FC. This optimization is achieved by maximizing the function described in (1). Additionally, the objective of (1) can be achieved by minimizing the KL-divergence between the two distributions, as explained below

$$\min_{\phi, \theta} \frac{1}{N} \sum_{j=1}^{N} - \log (g_\theta (z_j) = y_j)$$

s.t. $\phi = [\phi_1, \phi_2, \ldots, \phi_S]$

$$z_j = (f_{\phi_1}(x_1^1, f_{\phi_2}(x_2^2), \ldots, f_{\phi_S}(x^S)))$$

$$f_\phi \in \{0, 1\}^n \quad \forall i \in \{1, 2, \ldots, S\}$$

$$n \leq R$$

(1)

where $N$ refers to the total number of points in a test set. The decision function at the FC is denoted as $g_\theta$, where $\theta$ is a set of parameters. Furthermore, $f_\phi$, denotes the encoder function at the $i$th sensor, which is parameterized by $\phi_i$, and $R$ represents the assigned bandwidth for each sensor. The objective of the function given in (1) is to optimize the encoder parameters $\phi = [\phi^1, \phi^2, \ldots, \phi^S]$ and the decision rule parameters $\theta$ to minimize the negative log-likelihood loss. The function $f_\phi(x^i)$ represents the compressed and quantized version of the observation at the $i$th sensor, with $n$ being the dimensionality of the compressed representation. The output of the quantizer is binary quantized and belongs to $\{0, 1\}^n$, where $n$ must be less than or equal to the assigned bandwidth $R$.

The objective of the system can also be expressed using the Kullback–Leibler divergence between the two conditional distributions of the decision, given the raw observations and the compressed messages. These conditional distributions are specified by (2) and (4)

$$\min_{\omega, \theta, \phi} \text{KL} \left( P(\hat{Y}|X) \| P(\hat{Y}|Z) \right)$$

s.t. $\phi = [\phi_1, \phi_2, \ldots, \phi_S]$

$$P(\hat{Y}|X) = f_\omega(x^1, x^2, \ldots, x^S)$$

$$P(\hat{Y}|Z) = g_\theta(f_{\phi_1}(x_1^1), f_{\phi_2}(x_2^2), \ldots, f_{\phi_S}(x^S))$$

$$f_\phi \in \{0, 1\}^n \quad \forall i \in \{1, 2, \ldots, S\}$$

$$n \leq R$$

(2)

where $\omega$ is the parameters of a benchmark model (i.e., a larger neural network model trained to classify the raw observations without compression). But the KL-divergence is given by the following equation:

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

(3)

Substituting the KL term in (2) by (3), we get the following equation:

$$\min_{\omega, \theta, \phi} \sum_{i} P(\hat{Y}_i|X_i) \log \left( \frac{P(\hat{Y}_i|X_i)}{P(\hat{Y}_i|Z_i)} \right)$$

s.t. $\phi = [\phi_1, \phi_2, \ldots, \phi_S]$

$$P(\hat{Y}|X) = f_\omega(x^1, x^2, \ldots, x^S)$$

$$P(\hat{Y}|Z) = g_\theta(f_{\phi_1}(x_1^1), f_{\phi_2}(x_2^2), \ldots, f_{\phi_S}(x^S))$$

$$f_\phi \in \{0, 1\}^n \quad \forall i \in \{1, 2, \ldots, S\}$$

$$n \leq R.$$

(4)

Two important considerations should be taken into account when dealing with the problem at hand. First, the message space $[0, 1]$ is significantly smaller than the observation space $\mathbb{R}$. Second, the dimensionality required for the message, denoted by $n$, is substantially smaller than that of the raw observation, denoted by $d$, that is $n \ll d$. Consequently, the problem can be conceptualized as the search for an optimal encoder/quantizer tuple that maximizes the mutual information between the two conditional distributions $P(Y|X)$ and $P(\hat{Y}|Z)$ under a given communication rate $R$ for each sensor. It is worth noting that although (4) is proposed mainly for correlated observations, it can also be applied to independent observations. However, this approach would not provide significant benefits in terms of either the compression ratios (CRs) or the accuracy of the inferred decision.

### III. PROPOSED FRAMEWORK

#### A. Background on AE’s

One of the powerful deep-learning architectures that achieved the state-of-the-art results in different contexts is the AE. AE is a neural-network architecture consisting of two models namely, encoder and decoder models. The encoder maps an $I$-dimensional input to an $O$-dimensional codeword, where $O \ll I$. The decoder then reconstructs the input from this, compressed, codeword. This codeword is usually referred to as the latent representation and it belongs to a space called the latent space. This process is performed in an end-to-end fashion which implies that the encoder learns to compress the data in a way that helps the decoder in the reconstruction process. If the codeword is quantized (binary or multi-level), then the architecture is referred to as discrete representation’s AE. For further details on AE architecture, we refer to [17].

According to the aforementioned problem formulation, our objective is to jointly learn an optimal encoder and quantizer at each sensor, $q^i: q^i(x_i) = z_i$, and an optimal decision rule at the FC $g_\theta(x^1, z^2, \ldots, z^S)$. To this end, we adopt a discrete-representation AE at each sensor node to compress and quantize the sensor observations, see Fig. 1. In this context, we distinguish between reducing the input size and reducing the input space from a continuous space with uncountable values to a discrete countable space. We refer to the former one as compression: $f: \mathbb{R}^d \rightarrow \mathbb{R}^n$, with $n < d$. While the latter one is referred to as quantization and represents the mapping, $f: \mathbb{R}^d \rightarrow S^n$ where $S$ is a countable set.
Each sensor transmits the output of its encoder model to the FC. The output of the encoder model at sensor $i$ is given by: $f_{\phi_i}(\cdot)$, where $\phi_i$ is the parameters of the $i$th sensor. At the FC, an MLP neural network parameterized by parameters, $\theta$, is used to approximate the optimal decision rule as depicted in Fig. 2. The decision rule at the FC is given by the following equation:

$$g_{\theta} \left( \left[ f_{\phi_1}(x^1), f_{\phi_2}(x^2), \ldots, f_{\phi_S}(x^S) \right] \right)$$  \hspace{1cm} (5)

where $x^i$ is the current observation at the $i$th sensor.

### B. Description of the Implementation

At each sensor, the encoder architecture consists of a multilayer perceptron (MLP) with three fully connected layers employing rectified linear unit (ReLU) activation functions. The output layer of the encoder utilizes a quantized sigmoid (QSmog) activation as proposed by Moons et al. [18]. In the FC, six fully connected layers are employed with ReLU activation functions in the hidden layers and a Softmax activation in the output layer.

The $He$ initializer [19] is used to initialize the model weights. The proposed loss function given in (8) is minimized using the Adam optimizer [20] with a learning rate of 0.01. It is worth noting that the end-to-end training enables the encoder to capture the unique characteristics of each sensor, aiding the FC in making accurate decisions. Additionally, the FC weights are optimized to enhance the probability of accurate decisions given the compressed observations. As a result, the optimization of the FC’s classifier weights can be interpreted as learning a threshold function for the decision rule.

### C. Learning Algorithm

We propose a three-phase training algorithm for the proposed framework. The first stage involves the training of an AE at every sensor, where the objective is to reconstruct the input from a compressed codeword by minimizing l2-loss using the following equation:

$$\min_{\phi_i, \theta} \frac{1}{N} \sum_{i=1}^{N} \| x_i - \hat{x}_i \|^2$$  \hspace{1cm} (6)

where $x_i$ is the $i$th observation, $\hat{x}_i$ is the reconstruction, and $N$ is the total number of observations in a dataset. In the second stage, we train a distinct inference model, denoted as $I_1$, that accepts raw observations, $X = [x^1, x^2, \ldots, x^S]$, as input and generates the corresponding decision. It is important to note that the input to this model is the raw observations, without any compression or quantization. The optimization of this model is performed using the classical Crossentropy function, as shown in the following equation:

$$\min_{\theta} - \sum_{i=1}^{C} y_i \log (\hat{y}_i) + (1 - y_i) \log (1 - \hat{y}_i)$$  \hspace{1cm} (7)

where $y_i$ and $\hat{y}_i$ are the true and predicted one-hot encoded label vectors of the $i$th data point, and $C$ is the number of classes. The output of the model $I_1$ approximates the conditional distribution $P(Y|X)$. The set of parameters in the model $I_1$, denoted by vector $\omega$, is then frozen and it will be used only for computing the value of the inference model loss function at the FC. This model represents the benchmark model that we aim to mimic after the compression and the quantization take place. We elaborate more on this part in Section III-D.

In the third and final phase, the encoder model at each sensor is utilized to compress the captured observations. The encoder output at sensor $i$ is represented as $z^i$ and the corresponding encoder model parameters are denoted by $\phi_i$. Therefore, $z^i = f_{\phi_i}(x^i)$. The outputs of all encoders are aggregated and input into an inference model, $I_2$, with parameters $\theta$, to make predictions. Specifically, the input to the model $I_2$ is the concatenation of all encoders output, $Z = [z^1, z^2, \ldots, z^S]$. The output of $I_2$ approximates the conditional distribution $P(\hat{Y}|Z)$. It is important to note that $I_2$ weights, $\theta$, and the encoder weights, $\phi_i$, at each sensor are trained jointly.

The training in the final phase is performed in an end-to-end manner between the encoder weights and the decision rule parameters at the FC. The training procedure is summarized in Algorithm 1.

![Algorithm 1](image-url)

It is worth mentioning that although we used the concatenation operator to aggregate the sensor outputs, there are several data aggregation techniques that can be used to combine several input vectors into one vector. The choice and optimization of the aggregation technique are out of this work’s scope and have been left to future work. However,

---

**Fig. 1.** Figure that explains the system model. On the left, we see the sensor observations going through the discrete encoders to obtain the compressed quantized form of the observations. Then these messages are sent to the FC which passes the aggregated message to the neural network architecture to get a hypothesis estimation.

**Fig. 2.** The decision rule at the FC is given by the following equation.
we mention some of the widely studied aggregation techniques such as follows.

1) **Mean Pooling**: In this technique, the mean value of each element in the input vectors is computed, and a new vector is created with these mean values.

2) **Max Pooling**: In this technique, the maximum value of each element in the input vectors is computed, and a new vector is created with these maximum values.

3) **Min Pooling**: In this technique, the minimum value of each element in the input vectors is computed, and a new vector is created with these minimum values.

4) **Sum Pooling**: In this technique, the sum of each element in the input vectors is computed, and a new vector is created with these sum values.

5) **Concatenation**: In this technique, the input vectors are concatenated to form a single longer vector.

6) **Weighted Pooling**: In this technique, each input vector is assigned a weight, and the final vector is computed as a weighted average of the input vectors, where the weights are used as the coefficients.

7) **Attention Pooling**: In this technique, an attention mechanism is used to compute a weight for each element in the input vectors, and the final vector is computed as a weighted sum of the input vectors, where the attention weights are used as the coefficients.

**D. Proposed Objective Function**

In the course of our training, we first optimize the MSE loss function and subsequently the cross-entropy (CE) loss function. However, in the third phase of our training, which entails the joint training of the sensor encoders and FC inference model, we find that the traditional CE loss function only is inadequate for the objective of this work. As discussed in the preceding sections, our framework aims to enable the encoders to leverage redundancies between sensor observations to achieve high compression rates without sacrificing decision accuracy. This means that the encoders should learn to encode the complementary features of their respective observations. Accordingly, we propose a novel loss function, which is formulated as follows:

$$L(Y, \hat{Y}) = CE(Y, \hat{Y}) + KL(P(\hat{Y} | X) \| P(\hat{Y} | Z)).$$  \hspace{1cm} (8)

The purpose of the introduced function is to help in learning a joint conditional distribution for the decision, based on compressed observations, $P(\hat{Y} | Z)$. The goal of the function is to make this distribution as close as possible to the joint conditional distribution for the decision based on uncompressed observations, $P(\hat{Y} | X)$. By doing so, the function helps to minimize the degradation in decision accuracy resulting from compressing sensor observations. Due to the limited bit rate available to encode observations, the proposed function incentivizes the encoders to concentrate only on relevant features that are essential for maximizing the likelihood of an accurate decision at the FC.

During the end-to-end optimization of the proposed loss function, the encoders will tend to eliminate the mutual information between the correlated sensors’ observations and encode only the relevant features to satisfy the bandwidth constraint. Although this setting is derived from correlated data, it works also for independent sensors. However, the inter-observation redundancy is much less, and the compression can hurt the inferred decision accuracy.

Note that we handcrafted a model for each dataset to achieve the highest possible accuracy. The models have been selected according to the proposed loss function (8), such that it emphasizes learning complementary features at each sensor. To this

---

**Fig. 2.** Proposed framework for deep distributed inference in wireless sensor networks.
end, the second term in (8) adds a regularization term based on the KL-divergence between the conditional probability distribution of the decision given the full observation $P(\hat{Y}|X)$, and the distribution of the decision given the compressed and quantized version of the observations $P(\hat{Y}|Z)$. Moreover, our model jointly learns a quantizer function (entropy encoder) along with the source encoder. Jointly learning the encoders with the decision rule encourages the model to learn only the complementary features at each sensor. The proposed models work well with each problem without overwhelming the framework with complex architectures such as AlexNet, ResNet, GoogleNet, etc. [21]. The power of these complex models is required mainly for high-dimensional observation space, such as surveillance cameras. In this case, the hidden (deep) convolutional layers can extract spatial features in the observations in an efficient way. However, in lower-dimensional observation space, as in our case, handcrafted models are good enough. This conclusion is compatible with the findings reported in [22].

### E. Dataset Preparation

The proposed framework is general and widely applicable to different problems. For the framework to be employed in a certain distributed inference task, a dataset should be prepared for training purposes. A typical dataset consists of $N$ data points along with the associated labels $\{x_i, y_i\}_{i=1}^N$. Each data point, $x_i$, represents the concatenation of simultaneous readings from $S$ sensors such that $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,S}]$. The label $y_i \in \{1, 2, \ldots, C\}$ is the target hypothesis (i.e., class) associated with these sensor readings. It is worth noting that these readings are assumed to be perfectly synchronized and each data point represents the readings at the same time step.

### IV. RESULTS AND DISCUSSION

Various datasets have been used to evaluate the performance of the proposed framework. Each dataset represents a different environment setting and generating distribution.

#### A. Distributed Inference Accuracy

1) **Comparative Evaluation**: A public dataset for human action recognition called wearable action recognition database (WARD) has been utilized to evaluate our proposed framework. The dataset was released [25]. The obtained results are compared with three other models using the same dataset. This dataset is collected from five sensor boards attached to different points in the human body. Each sensor board has a tri-axial accelerometer and a bio-axial gyroscope with three and 2-D outputs, respectively. Each human operator performs 13 different actions which represent the labels (classes) to be predicted by the classifier at the FC.

Table I and Fig. 3 present a comparative analysis of the performance of the proposed framework and the baseline models at varying CRs. Examination of Table I reveals that the proposed framework surpasses other models in terms of accuracy across all CRs. Notably, the proposed framework maintains a high level of accuracy even when subjected to high CRs. For example, increasing the CR from 2 to 8 decreased the accuracy by 4.1% only (i.e., from 99.7% to 95.6%). This is a small margin compared with 11% loss in Cheng et al. [23] (ASRCM), and 7% in Cheng et al. [23] (NN) and Zhang and Sawchuk [24].

Table II shows the classification accuracy of the proposed framework compared with the accuracy of other works in the literature. The table reports results for Zhu et al. [3], Guo et al. [8], Huynh [9], He et al. [10], Yang et al. [25], Oniga and Jozsef [26], and Sheng et al. [27]. It is clear from the table that the proposed framework achieves state-of-the-art accuracy compared with the aforementioned works. In addition, the proposed framework involves the minimal required bit rate, $R$, from the sensors to the FC, which highly contributes to power saving and prolonging the sensors’ lifetime. These results can be attributed to the fact that we learn complementary features between correlated sensors that highly contribute to improving the decision accuracy rather than learning local features for each sensor. This learning behavior is motivated by the proposed loss function in (8).

#### Table I

**Classification Accuracy of the Proposed Framework Under Different CRs Compared With Benchmark Models on WARD Dataset**

| Method                  | CR=2 | CR=4 | CR=8 |
|-------------------------|------|------|------|
| Cheng et al. (ASRCM)    | 94%  | 88%  | 83%  |
| Cheng et al. (NN)       | 82%  | 78%  | 75%  |
| Zhang et al. [24]       | 87%  | 83%  | 80%  |
| Our Framework           | 99.7%| 97.4%| 95.6%|

#### Table II

**Comparing the Proposed Framework With Prior Work Using WARD Dataset at a CR (CR = 2)**

| Method                  | Detection Accuracy |
|-------------------------|--------------------|
| Zhu et al. [3]          | 99.00%             |
| Yang et al. [26]        | 93.60%             |
| Huynh [9]               | 96.97%             |
| He et al. + PCA [10]    | 76.31%             |
| He et al. + LDA [10]    | 40.30%             |
| He et al. + GDA [10]    | 99.20%             |
| Guo (Majority voting) [8]| 94.96%             |
| Guo (Maximum) [8]       | 96.20%             |
| Guo (WLOP) [8]          | 98.02%             |
| Guo (WLOGP) [8]         | 98.78%             |
| Sheng et al. [27]       | 95.90%             |
| Oniga and Jozsef [28]   | 98.10%             |
| Our Framework           | 99.77%             |

---

Fig. 3. Comparison of model accuracy under different CRs.
Moreover, our framework jointly learns a quantizer function \( q: \mathcal{X} \rightarrow \mathcal{Z} \) with the encoder function which minimizes the end-to-end error and improves the accuracy of the sensing task. Note that the work in [3] explores the correlation between the sensor observations to disable the transmission on the sensors that did not capture new relevant features and thus save the consumed bandwidth. Comparatively, in our work, we exploit this correlation to transmit only the relevant complementary features. Consequently, we contribute in two directions, namely, saving the consumed bandwidth and, at the same time, improving the decision accuracy.

2) Artificial Problem: The proposed framework underwent several evaluations using four datasets: street view houses (SVH) [28], CIFAR-10 [29], Modified National Institute of Standards and Technology (MNIST) [30], and Fashion-MNIST [31]. For each dataset, we tested various CRs, which denote the ratio of uncompressed dimension to compressed dimension [32]. It should be noted that prior research only considered dimensionality reduction for compression, where input or output dimensions could be any value in \( \mathbb{R} \). As a result, the input and output spaces remained unaltered. In contrast, our framework goes beyond this by taking into account the quantization of the codewords, since input space is \( \mathbb{R}^d \) while output dimensions are quantized to belong to the space \( 0, 1^d \).

In this experiment, we simulate (an artificial problem) where two sensors (\( s_1, s_2 \)) transmit their data to an FC. Suppose the observations at sensor \( s_1 \) correspond to class \( C_i \), and those at sensor \( s_2 \) correspond to class \( C_j \). The decision rule at the FC can be formulated as follows:

\[
\psi(z_1, z_2) = \begin{cases} 
  i, & \text{if } i = j \\
  -1, & \text{if } i \neq j.
\end{cases}
\]

In essence, the decision is the class label when the two observations belong to the same label and \(-1\) otherwise. As each dataset contains images belonging to one of the ten total classes, the classifier is expected to have 11 classes (i.e., ten original classes and the “\(-1\)” class). To ensure a fair comparison, we employed the same classifier capacity (e.g., the number of layers, number of nodes per layer, activation functions, etc.) for all datasets. We compared the results obtained from our proposed framework with the accuracy of a baseline model, defined as a neural network that takes raw observations as input. In this case, the FC has complete knowledge of the sensed data, which represents the ideal scenario in terms of data availability.

The results of the proposed framework are illustrated in Fig. 4. The performance of the framework closely approaches that of the baseline when the CR is the lowest, i.e., CR = 2. A marginal decrease in accuracy is observed at higher CRs, such as CR = 4 and CR = 8. Nevertheless, the obtained accuracy is still remarkable, even at the highest CR. For instance, the framework achieved 95.3% accuracy in the MNIST dataset with CR = 8, which implies a reduction of only 4.7% in accuracy after compressing the observations to only 12.5% of their original dimension with quantization.

In the reconstruction of the training dataset, we randomly shuffle the datasets at each sensor. Consequently, most of the observation combinations fall in the class of \(-1\) (i.e., the two observations are not in the same class). This produces an imbalanced class distribution. Due to this imbalance, we report the confusion matrix of the framework classifier for the MNIST dataset and 98-dimension latent code in Fig. 5. We can see from Fig. 5 that the proposed framework is capable of inferring the right decision with high accuracy even with imbalanced data.

The key idea of compressing correlated sensor data is extracting complementary information from correlated observations and ignoring any redundancies. Our proposed loss function (8) achieves this goal by incorporating a KL divergence term to the loss function between the soft labels generated by a baseline model (e.g., a large model trained on raw observations to predict \( P(Y|X) \)) and the decision function at the FC, \( P(Y|Z) \). To minimize this term, we encode only the complementary features that help the FC to mimic the behavior of the baseline model. As described in Algorithm 1, we jointly train the encoder models at each sensor with the decision function at the FC in an end-to-end fashion. This end-to-end training makes the encoders jointly learn features with the decision function as they receive penalization based on the distance between the predicted distribution and that of the baseline model.

We also consider a convex combination between these two terms in (8) as follows:

\[
\mathcal{L}(\hat{Y}, \hat{Y}) = \phi \text{ CE}(\hat{Y}, \hat{Y}) + (1 - \phi) \times \left( \text{KL}(P(\hat{Y}|X) \mid \mid P(\hat{Y}|Z)) \right).
\]

We experimented different values for \( \phi \) in the range (0, 1) with 0.1 step size. We note that the best results are obtained at different values for \( \phi \) for the different datasets. For example,
Fig. 6. Accuracy of indoor localization problem using the proposed framework.

Fig. 7. Interpolation between two points in the latent space. We randomly choose a start point and an endpoint. The start point is shown in the top-left corner and the endpoint is shown in the bottom-right. Each time, we gradually flip a bit along the different bits between the two vectors.

using $\phi$ equals 0.4 gave the best results with MNIST while 0.5 and 0.7 gave the best results with Fashion-MNIST and CIFAR10. This can be attributed to the different distribution of the observation and the class weights in each dataset. Therefore, we recommend trying different values of $\phi$ and using the value with the best results.

3) Indoor Localization: We test our proposed framework on a dataset for indoor localization using WiFi fingerprint. The dataset consists of 7175 fingerprints collected from 489 different locations (almost 15 fingerprints per location). The training dataset was compiled by taking samples at every 3 m on average with 15 samples per location. The time at each location was approximately 40 s performing consecutive scans with a bq Aquaris E5 4G device using Android stock 6.0.1 without making any movements during the process. For a complete description of the dataset and the dataset collection protocol, we refer the reader to [33]. Fig. 6 shows the degradation in the accuracy due to the increase in the CR. Note that in this experiment, we are not attempting to achieve the state-of-the-art classification results. Rather, we aim to highlight how much accuracy we can lose due to the observation compression. It is clear from Fig. 6 that the decrease in the accuracy due to observation compressing using our proposed framework is much less compared with compression using traditional AEs.

B. Semantics of the Latent Representation

AE-based architectures for dimensionality reduction often prioritize the resilience of the learned codewords in the latent space [34]. To evaluate the robustness of these codewords, interpolation between various points in the latent space is conducted, and the gradual changes in the reconstructed data are qualitatively observed. This experimentation serves to demonstrate that the model has: 1) integrated sufficient redundancies into the codewords, allowing for input reconstruction despite the presence of codeword errors and 2) acquired relevant features that align with the underlying structure of the data.

Two test points are randomly chosen as the starting and ending points for the interpolation experiment. The process involves flipping a bit in the latent codeword at each step and then passing the resulting altered codeword to the decoder model and observing the resulting changes in the reconstruction. The detailed steps of this experiment are outlined in Algorithm 2. Gradual changes in the shape of the digit are depicted in Fig. 7, showing that decreasing the Hamming distance between the start and end points by flipping bits results in a gradual alteration of the digit characteristic until it reaches the endpoint.

Algorithm 2 The Procedure for Evaluating the Semantics of the Latent Codewords

Randomly select two random points $x_1, x_2$;
Encode each data point using the encoder function, $f_\phi$: $z_1, z_2 = f_\phi(x_1), f_\phi(x_2)$;
$h = z_1 \oplus z_2$;
i=0;
while i < len(h) do
  if $h[i]$ equals 1 then
    Flip the bit at $z_1[i]$;
    i = i + 1;
  end if
  $x_{i1} = f_\theta(z_1)$; where $f_\theta$ is the decoder function.
end while
Plot $x_{i1}$;
C. Rate/Computation Tradeoff

While compressing the observations reduces bandwidth consumption for transmission, it comes at a cost in accuracy and computation. The required computation resources (measured by the floating-point operations (FLOPS)) increase according to the model complexity (measured by the number of weights). Moreover, increasing the model complexity leads to improved compression, and consequently improved decision accuracy at the FC. Therefore, a design decision should compromise between the model complexity, on the one hand, and the consumed computations and FC accuracy, on the other hand. However, the training phase can be done offline (before the deployment of the sensors), and only the inference will take place during the operation which requires only one forward pass (a very small number of FLOPS) to predict the encoded messages. Fig. 8 shows this trade-off trend between the computation requirement (measured by FLOPS) and the model accuracy. In this figure, we can see that increasing the model accuracy requires adopting smaller CRs which imply higher data transmission. On the other hand, a smaller CR requires transmitting more data and requires more computational resources at each sensor. The optimization of CRs is out of the scope of this article and will be explored in our future work.

D. On Quantizer Design

Different techniques for quantizer design have been proposed in the literature. Some work proposed analytical techniques for the quantizer design. These techniques give a precise description of the optimality of the quantizer and, in some cases, they specify the quantizer performance (in the form of the error bound or other criteria). However, to derive this mathematical analysis, these techniques assume certain statistical properties in the sensor data. There are many cases in which we do not have prior knowledge of this information. For example, the quantizer proposed on [35] studied Gaussian observations. On the other hand, our proposed quantizer design does not impose any restrictions or assumptions on the distribution of the raw observations. Other learning-based quantizer designs either do not address the case of correlated observations or do not consider the accuracy of the inferred decision on the loop. Choi et al. [34] proposed a learning-based quantizer in the channel coding regime. However, they did not consider the accuracy of the decisions inferred from this compressed data. In this work, we address this gap.

E. Applicability

Our framework along with the proposed loss function in (8) and the training procedure given in Algorithm 1 can work with any type of parallel distributed detection network. This type of setting has various applications in wireless IoT. Although minor customizations are required to fit each specific problem, the framework is still widely applicable to different problems from various domains. In this article, we reported the experimental results on various types of sensors and applications (e.g., image classification, human activity recognition, etc.). Specifically, we experimented with five different datasets (MNIST, Fashion MNIST, Street View House Numbers (SVHN), CIFAR-10, WARD) representing three different types of sensors (cameras, gyroscope, and accelerometer). To further evaluate the generality of our framework, we evaluated a completely different domain (i.e., wireless link adaptation) using three datasets combined in a global one [36]. In this scenario, the sensors are the antennas at each mobile node, the observations are the channel state information (CSI) captured at each mobile, and the environment is the wireless channel [37]. The sensors send their observations to an FC to infer a global decision. The base station (BS) acts as an FC in this case, and the decision is the selected modulation and coding scheme (MCS). The results shown in Fig. 9 show a minor loss in the adaptation decision at the FC with the increase in the adopted CR. For example, when compressing the original raw observations (i.e., CSI in this case), the accuracy only drops from 94.25% to 93.7%. This means only a 0.55% loss in accuracy is achieved while saving 75% of the original bandwidth. The obtained results confirm the general applicability of our proposed method in different domains and problems.

F. Results of Input Reconstruction

A task of input reconstruction was performed to evaluate the robustness of the learned features. In this experiment, MNIST and Fashion-MNIST datasets are used in the evaluation. We used a CR = 8, which corresponds to a latent code of 98-bit. Fig. 10 shows the result of the input reconstruction.

V. Conclusion

In this article, we proposed a deep-learning framework for compressing correlated sensor observations in distributed inference problems. The proposed framework employs discrete representation AEs to encode the observations at each sensor. A novel loss function has been proposed to improve the
accuracy of the framework. A MLP architecture has been used at the FC to jointly learn the decision rule. The proposed framework addresses the hard-to-tackle problem of correlated sensor observations and does not assume any prior knowledge about the distribution of the observations. The framework has been extensively tested using different datasets and has demonstrated significant performance improvements.

In the future, we plan to optimize the various hyperparameters in the proposed loss function and training algorithm (e.g., the contribution of each term in 10). We believe that optimizing these values can lead to substantial improvements. Furthermore, to obtain the highest radio resource utilization, we can tackle the problem of assigning different rates to different sensors based on the relevancy of the information captured at each sensor. In addition, incorporating the prior knowledge of other FCs in the training as a type of transfer learning can be of substantial contribution. Another research direction can be dedicated to optimizing the data aggregation direction can be dedicated to optimizing the data aggregation technique based on the inference task, the number of sensors, the dimensionality of each sensor codeword, and other factors relevant to the problem design. This can significantly contribute to improving the inferred decision accuracy as well as resource utilization.

REFERENCES

[1] S. Salehkalaibar, M. Wigger, and R. Timo, “On hypothesis testing against conditional independence with multiple decision centers,” IEEE Trans. Commun., vol. 66, no. 6, pp. 2409–2420, Jun. 2018.

[2] Y. Abdi and T. Ristaniemi, “The max-product algorithm viewed as linear data-fusion: A distributed detection scenario,” IEEE Trans. Wireless Commun., vol. 19, no. 11, pp. 7585–7597, Nov. 2020.

[3] P. Zhu et al., “Cost aware inference for IoT devices,” in Proc. Int. Conf. Artif. Intell. Statist., 2019, pp. 2770–2779.

[4] J.-F. Chamberland and V. V. Veeravalli, “Wireless sensors in distributed detection applications,” IEEE Signal Process. Mag., vol. 24, no. 3, pp. 16–25, May 2007.

[5] W. Peng Tay, J. N. Tsitsiklis, and M. Z. Win, “Bayesian detection in bounded height tree networks,” IEEE Trans. Signal Process., vol. 57, no. 10, pp. 4042–4051, Oct. 2009.

[6] T. Glasmachers, “Limits of end-to-end learning,” in Proc. Asian Conf. Mach. Learn. (ACML), 2017, pp. 17–32.

[7] A. Y. Yang, S. Eyvindar, S. Sastry, R. Bajcsy, P. Kuryloski, and R. Jafari, “Distributed segmentation and classification of human actions using a wearable motion sensor network,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2008, pp. 1–8.

[8] Y. Guo, W. He, and C. Gao, “Human activity recognition by fusing multiple sensor nodes in the wearable sensor systems,” J. Mech. Med. Biol., vol. 12, no. 5, Dec. 2012, Art. no. 12500084.

[9] D. T. G. Huynh, “Human activity recognition with wearable sensors,” Ph.D. dissertation, Technische Universität Darmstadt, Germany, 2008.

[10] W. He, Y. Guo, C. Gao, and X. Li, “Recognition of human activities with wearable sensors,” EURASIP J. Adv. Signal Process., vol. 2012, no. 1, p. 108, Dec. 2012.

[11] M. Faulkner, M. Olson, R. Chandy, J. Krause, K. Mani Chandy, and A. Krause, “The next big one: Detecting earthquakes and other rare events from community-based sensors,” in Proc. 10th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw., Apr. 2011, pp. 13–24.

[12] A. Raghavan and J. S. Baras, “Binary hypothesis testing by two collaborating observers: A fresh look,” in Proc. 27th Medit. Conf. Control Autom. (MED), Jul. 2019, pp. 244–249.

[13] T. Bouchoucha, M. F. A. Ahmed, T. Y. Al-Naffouri, and M. Alouini, “Distributed estimation based on observations prediction in wireless sensor networks,” IEEE Signal Process. Lett., vol. 22, no. 10, pp. 1530–1533, Oct. 2015.

[14] S. P. Chinchali, E. Cidon, E. Pergament, T. Chu, and S. Katti, “Neural networks meet physical networks: Distributed inference between edge devices and the cloud,” in Proc. 17th ACM Workshop Hot Topics Netw., Nov. 2018, pp. 50–56.

[15] P. Hu, J. Im, Z. Asgar, and S. Katti, “Starfish: Resilient image compression for AoIoT cameras,” in Proc. 18th Conf. Embedded Networking Sensor Syst., Nov. 2020, pp. 395–408.

[16] H. Amer, “Image/video compression: Human and computer vision perspectives,” Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Waterloo, Waterloo, ON, Canada, 2020.

[17] A. Majumdar, “Blind denoising autoencoder,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 1, pp. 312–317, Jan. 2019.

[18] B. Moons, K. Goetschalckx, N. Van Berckelaer, and M. Verhelst, “Minimum energy quantized neural networks,” in Proc. 51st Asilomar Conf. Signals, Syst., Comput., Oct. 2017, pp. 1921–1925.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1026–1034.

[20] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[21] Y. LeCun et al., “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[22] J. Suto and S. Oniga, “Efficiency investigation from shallow to deep neural network techniques in human activity recognition,” Cognit. Syst. Res., vol. 54, pp. 37–49, May 2019.

[23] L. Cheng, Y. Guan, K. Zhu, Y. Li, and R. Xu, “Accelerated sparse representation for human activity recognition,” in Proc. IEEE Int. Conf. Inf. Reuse Integ. (IRI), Aug. 2017, pp. 245–252.

[24] W. Zhang and A. S. A. Seah, “Human daily activity recognition with sparse representation using wearable sensors,” IEEE J. Biomed. Health Informat., vol. 17, no. 3, pp. 553–560, May 2013.

[25] A. Y. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy, “Distributed recognition of human actions using wearable motion sensor networks,” J. Ambient Intell. Smart Environments, vol. 1, no. 2, pp. 103–115, 2009.

[26] S. Oniga and S. Jozsef, “Optimal recognition method of human activities using artificial neural networks,” Meas. Sci. Rev., vol. 15, no. 6, pp. 323–327, Dec. 2015.

[27] M. Sheng, J. Jiang, B. Su, Q. Tang, A. A. Yahya, and G. Wang, “Short-time activity recognition with wearable sensors using convolutional neural network,” in Proc. 15th ACM SIGGRAPH Conf. Virtual-Reality Continuum Its Appl. Ind., Dec. 2016, pp. 413–416.

[28] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” in Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn., 2011. [Online]. Available: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf

[29] A. Krizhevsky et al., “Learning multiple layers of features from tiny images,” Toronto, ON, Canada, 2009.

[30] Y. LeCun. (1998). The MNIST Database of Handwritten Digits. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[31] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms,” 2017, arXiv:1708.07747.

[32] K. Sayood., Introduction to Data Compression. San Mateo, CA, USA: Morgan Kaufmann, 2017.

[33] J. L. S. González, L. M. S. Morillo, J. A. Álvarez-García, F. E. De Salamanca Ros, and A. R. J. Ruiz, “Energy-efficient indoor localization WiFi-fingerprint system: An experimental study,” IEEE Access, vol. 7, pp. 162664–162682, 2019.

[34] K. Choi et al., “Neural joint source-channel coding,” in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 1–11.

[35] A. Vempaty, H. He, B. Chen, and P. K. Varshney, “On quantizer design for distributed Bayesian estimation in sensor networks,” IEEE Trans. Signal Process., vol. 62, no. 20, pp. 5359–5369, Oct. 2014.

[36] M. Hussien, M. F. A. Ahmed, G. Dhunna, K. K. Nguyen, M. Cheriet, and G. Poitau, “Towards more reliable deep learning-based link adaptation for WiFi 6,” in Proc. IEEE Int. Conf. Commun. (ICC), Jun. 2021, pp. 1–6.

[37] M. Hussien, K. K. Nguyen, and M. Cheriet, “PRVNet: A novel partially regularized variational autoencoders for massive MIMO CSI feedback,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Apr. 2022, pp. 2286–2291.