SemQNet: Semantic-Aware Quantised Network for mmWave Beam Prediction

Ahsan Raza Khan, Poonam Yadav
Department of Computer Science, University of York, YO10 5GH, United Kingdom
Email: {ahsan.khan, poonam.yadav}@york.ac.uk

***Abstract*—Millimetre-wave (mmWave) communication systems use large antenna arrays and narrow beams to achieve strong signal power. However, this approach requires extensive beam training, which leads to high overhead. Recently proposed vision-aided beam prediction methods show promising results, reducing this overhead. However, these techniques have considerable computational complexity, hindering practical deployment. To address this issue, we propose a Semantic-Aware Quantised Network (SemQNet) framework that leverages image compression and a lightweight computer vision model to extract semantic information used for training a fully connected neural network (FCNN). Additionally, the proposed SemQNet also uses quantisation-aware training (QAT), which enables low-precision arithmetic operation, reducing the model size in the training process. Our tests on the DeepSense 6G dataset show that SemQNet achieves almost the same top-1 accuracy as existing vision-based methods while reducing the model size by 74.21%. This smaller model size reduces the communication overhead, making SemQNet a practical and efficient solution for energy-constrained mmWave communication systems.**

***Index Terms*—Millimetre wave, semantic communication, beam prediction, computer vision, deep learning**

I. Introduction

The future of wireless communication is rapidly advancing toward higher frequency bands, such as millimetre-wave (mmWave) and sub-terahertz (THz). This transition is driven to enable services like ultra-reliable low-latency communication (URLLC), massive machine-type communication (MTC), and enhanced mobile broadband (eMBB) [1]. Fifth-generation (5G) networks have established a foundation for applications such as autonomous driving, augmented or virtual reality, and Industry 4.0. However, emerging technologies and applications like mixed reality, 8K video streaming, and telepresence are imposing even stricter requirements, pushing the limits of 5G capabilities [2]. Beyond 5G (B5G) and sixth-generation (6G) networks are envisioned to meet these demands by leveraging the larger bandwidths available in mmWave and THz frequencies. However, these high-frequency bands are inherently more sensitive to physical obstructions, requiring precise beamforming with large antenna arrays to maintain stable connections [4]. This precision comes at the cost of significant beam training overhead, creating challenges for highly mobile, low-latency applications in dynamic environments.

Various techniques have been proposed to reduce the beam training and channel estimation overhead in mmWave communication systems. These approaches mainly focus on three key strategies [3], [4]. The first approach involves developing adaptive or hierarchical beam codebooks to efficiently narrow down the set of candidate beams with fewer measurements [5]. The second leverages compressive sensing techniques, which take advantage of the sparse nature of mmWave channels, to estimate the full channel with fewer observations [6]. The third strategy focuses on beam-tracking methods that use user mobility information to predict future beams, reducing the need for exhaustive searches. While these classical methods achieve some level of improvement, their effectiveness is limited in real-world systems, especially with larges number of antenna arrays and applications with stringent low-latency requirements [3].

The limitations of classical approaches have driven the adoption of machine learning (ML) techniques to address the beam training and channel estimation challenges in mmWave. These methods leverage multi-modal data, such as user position and orientation, LiDAR point clouds, radar measurements, and RGB images, to significantly reduce training overhead [3], [4]. Among these, vision-aided beam prediction has shown great potential to predict optimal beams directly from raw RGB images. While effective, the reliance on raw image data introduces high storage and computational demands, making this solution less practical [4]. As a result, further advancements are needed to balance efficiency and performance in vision-aided wireless systems.

To address the above mentioned challenge, we propose a Semantic-Aware Quantised Network (SemQNet) framework that combines image compression, lightweight semantic extraction, and quantisation-aware training (QAT) to enhance efficiency and scalability. Unlike prior work, such as [4], which directly uses extracted semantic features, our framework introduces an additional image preprocessing step, where RGB images are converted to grayscale and compressed using lossy JPEG encoding. This preprocessing reduces the data size and computational requirements while preserving essential spatial features. MobileNetV2 a lightweight computer vision (CV) model then processes the compressed images to extract bounding box (bbox) features and inputs to a fully connected neural network (FCNN). To further optimise performance, we integrate QAT, enabling low-precision arithmetic during training and significantly reducing the model size without compromising accuracy. The key contributions of this work are highlighted as:

- Proposed SemQNet framework that uses compressed grayscale images with lossy JPEG encoding for semantic information extraction in a practical mmWave communication, significantly reducing data size and computational requirements.
- Integrated QAT into the training process of FCNN to enable low-precision arithmetic operations, reducing the model size and training overhead while maintaining competitive accuracy.
- Validated the proposed framework on a publicly available dataset DeepSense 6G [7], demonstrating competitive accuracy, improved accuracy-to-model-size ratio, and enhanced energy efficiency compared to existing methods.

II. System Model and Problem Formulation

This section first presents the system model adopted for this work, followed by problem formulation for semantics-aware beam prediction.

A. System Model

This research considers mmWave communication system with a base station (BS) equipped with an N-element uniform linear array (ULA) and a vision sensor to capture the surrounding environment. The user is a mobile node equipped with a single antenna and a GPS module to collect real-time positional data. The communication system employs Orthogonal frequency-division multiplexing (OFDM) for transmission, with K sub-carriers and a cyclic prefix of duration τ. To ensure robust communication, the BS utilises a predefined beamforming codebook $\mathcal{B} = \{\mathbf{w}_q\}_{q=1}^{Q}$, where $\mathbf{w}_q \in \mathbb{C}^{N\times 1}$ represents the beamforming vectors and Q is the total number of beams available. Let's assume that the $\mathbf{h}_k[t] \in \mathbb{C}^{N\times 1}$ denotes the channel response at the k-th sub-carrier and time instance t. The downlink signal received by the mobile user is expressed as [4]:

$$r_k[t] = \mathbf{h}_k^{\mathsf{H}}[t]\mathbf{w}_t s + n_k[t], \tag{1}$$

where $\mathbf{w}_t \in \mathcal{B}$ is the beamforming vector selected at time t. The $s \in \mathbb{C}$ is the complex transmitted symbol with a power constraint $\mathbb{E}[|s|^2] = P$, and $n_k[t] \sim \mathcal{CN}(0, \sigma^2)$ represent the noise with Gaussian distribution. The optimal beamforming vector $\mathbf{w}_t^\star$ is chosen to maximise the average received signal-to-noise ratio (SNR) across all sub-carriers mathematically given as:

$$\mathbf{w}_t^\star = \underset{\mathbf{w}_q\in\mathcal{B}}{\operatorname{argmax}} \frac{1}{K}\sum_{k=1}^{K} \mathrm{SNR}|\mathbf{h}_k^{\mathsf{H}}[t]\mathbf{w}_q[t]|^2, \tag{2}$$

where P/σ^2 denotes the transmit SNR. This mathematical formulation provides a theoretical approach to identifying optimal beams.

B. Beam Prediction Problem Formulation

Given the system model, the goal of beam prediction is to identify the optimal beamforming vector $\mathbf{w}_t^\star \in \mathcal{B}$, from a pre-defined codebook at any time t, maximising the average received power. Solving optimal beam prediction problems typically requires precise channel state information, often challenging to obtain in practical scenarios. An alternative approach is an exhaustive search over the pre-defined beam codebook. This exhaustive search incurs significant overhead due to the large antenna arrays and narrow beams in mmWave systems [8]. ML-based solutions have been proposed to mitigate this overhead, leveraging multi-sensor data and prior observations to facilitate rapid beam prediction [4]. Inspired by this, we propose a semantics-aided approach where environmental information derived from RGB images assists in beam index prediction. Instead of directly using raw RGB images, we compressed the images, extracted high-level semantic features, such as object masks and bbox and trained a DL model to predict optimal beams efficiently.

Let $\mathbf{I}[t] \in \mathbb{R}^{W\times H\times C}$ represent the RGB image captured by the camera, placed at the BS on given time instant t, where W, H, and C denote the width, height, and colour channels of the image, respectively. The captured images are preprocessed and compressed to optimise storage and processing efficiency. This involves converting the RGB images to single-channel grayscale and encoding them in a lossy JPEG format to reduce their size significantly. This preprocessing step not only reduces storage requirements but also streamlines the subsequent extraction of high-level semantics, ensuring computational efficiency. Once compressed, the images are used to extract semantic features, denoted by $\mathbf{S}[t]$ from $\mathbf{I}[t]$. The task of beam prediction can then be expressed as finding a mapping function $\mathcal{F}_\Theta$ that maps the extracted semantics $\mathbf{S}[t]$ to an estimated beam index $\hat{\mathbf{w}}_t \in \mathcal{B}$:

$$\mathcal{F}_\Theta : \mathbf{S}[t] \rightarrow \hat{\mathbf{w}}_t. \tag{3}$$

The ML model is parameterised by Θ, which is optimised using a dataset $\mathcal{D} = \{(\mathbf{S}_u, \mathbf{w}_u^\star)\}_{u=1}^{U}$, where U is the number of samples. The dataset consists of labelled image beam pairs, and the objective of the ML model is to maximise the prediction accuracy across the dataset:

$$\Theta^\star = \underset{\Theta}{\operatorname{argmax}} \prod_{u=1}^{U} \Pr(\hat{\mathbf{w}}_u = \mathbf{w}_u^\star \mid \mathbf{S}_u). \tag{4}$$

This work develops a deep learning (DL) framework to learn the mapping function $\mathcal{F}_\Theta$ for mmWave beam prediction, leveraging extracted semantic features from visual data for efficient and accurate inference.

III. Proposed Sementic Extraction and QAT

This section presents the proposed SemQNet framework, which comprises three main components: image preprocessing, lightweight semantic information extraction using MobileNetV2, and DL model training with QAT. The block diagram of the proposed model is shown in Fig. 1.

A. Data Compression

The input data consists of RGB images captured by a vision sensor placed at the BS. These images undergo preprocessing

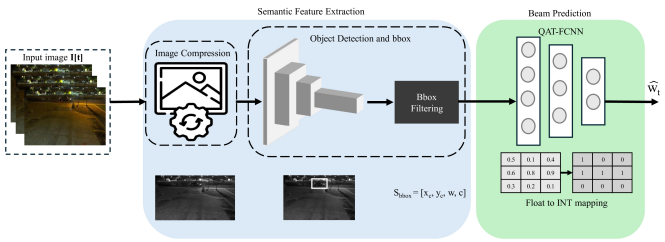

Fig. 1. Proposed system model with QAT.

to optimise the computational efficiency and storage requirements, including conversion to grayscale and compression into a lossy JPEG format. This preprocessing pipeline ensures that the meaningful spatial features of the visual data are preserved while reducing the data size significantly. The system prepares the input for the subsequent semantic extraction process by transforming the RGB images into compact representations.

B. Semantic Information Extraction

For semantic information extraction, we employ MobileNetV2 [9], a lightweight and efficient CV model optimised for edge processing. Instead of training MobileNetV2 from scratch, we incorporate a pre-trained version into our architecture with minor adjustments. This approach provides two primary benefits: (i) enhanced detection performance through transfer learning and (ii) accelerated training convergence. By utilising transfer learning, MobileNetV2 leverages prior knowledge gained from large, diverse datasets such as the COCO dataset [10], which contains object classes commonly encountered in outdoor environments.

At the inference stage of MobileNetV2, a sequence of input compressed images $\mathbf{I}$ is processed; it extracts key semantic features such as object classes and bbox coordinates. This capability allows for the efficient conversion of raw images into compact semantic representations denoted as $\mathbf{S}[t]$, which serve as input to the beam prediction model. The bbox features are represented as $\mathbf{S}_{\text{bbox}} = [x_c, y_c, w, h]$, where $\mathbf{S}_{\text{bbox}} \in \mathbb{R}^{4\times 1}$. The x_c, y_c are the coordinates of the object's center, and w and h denote its width and height, respectively. These features are normalised to the range $[0, 1]$ to ensure compatibility with the training pipeline and improve model convergence.

C. Deep Learning with QAT

The semantic features $\mathbf{S}_{\text{bbox}}$ are fed into a 2-layer fully connected neural network (FCNN) designed for efficient and accurate beam prediction. Each layer in the FCNN contains 175 neurons and utilises the rectified linear unit (ReLU) activation function to model complex relationships between the input features and the beam indices.

1) Model Architecture:

- **Input Layer:** Takes the 4-dimensional semantic vector (bounding box coordinates) as input.
- **Hidden Layers:** Consist of two fully connected layers with 175 neurons each, designed to learn complex relationships between the semantic features and beam indices. This architecture is intentionally chosen for fair comparison given in one of the pioneering works in this domain [4].
- **Output Layer:** Produces a probability distribution over the available beam indices, enabling the beam selection that maximises the expected received SNR.

2) Quantisation-Aware Training: To achieve both accuracy and efficiency in computation and storage, we integrate QAT, which performs low-precision arithmetic operations, such as using 8-bit integers instead of 32-bit floats, during training. Unlike post-training quantisation, which applies quantisation after a model has been fully trained in floating-point precision, QAT includes quantisation steps directly within the training loop [11]. By doing so, the model parameters and intermediate features learn to be robust under low-precision arithmetic, thereby reducing model size and computational overhead when deployed on energy-constrained or latency-sensitive platforms.

a) Setup for QAT: Recall from our system model discussed in Section. II-A, a dataset $\mathcal{D} = \{(\mathbf{S}_u, w_u^\star)\}_{u=1}^{U}$, where each $\mathbf{S}_u \in \mathbb{R}^4$ represents the extracted semantic features at instance u, and $w_u^\star \in \{1, \ldots, Q\}$ denotes the optimal beam index selected from a predefined beamforming codebook $\mathcal{B}$. The neural network $f_\Theta : \mathbb{R}^4 \to \mathbb{R}^Q$ is parameterised by Θ (comprising weights and biases) and maps the semantic input $\mathbf{S}_u$ to a set of logits $\hat{\mathbf{z}}_u$, where Q is the number of possible beam indices.

Under standard (full-precision) training, the forward pass through a two-layer FCNN can be represented as:

$$\mathbf{z}_u^{(1)} = \mathbf{S}_u \mathbf{W}_1 + \mathbf{b}_1, \quad \mathbf{a}_u^{(1)} = \text{ReLU}(\mathbf{z}_u^{(1)}), \tag{5}$$

$$\mathbf{z}_u^{(2)} = \mathbf{a}_u^{(1)} \mathbf{W}_2 + \mathbf{b}_2, \quad \hat{\mathbf{z}}_u = \mathbf{z}_u^{(2)}.$$

Here, $\mathbf{W}_l$ and $\mathbf{b}_l$ denote the weights and biases of the l-th layer, and $\mathbf{a}_u^{(l)}$ the activation output from that layer.

b) Incorporating Quantisation into the Forward Pass: The quantisation operations is introduced in the training loop, where $Q(\cdot)$ represent the quantisation operator and $DQ(\cdot)$ the corresponding dequantisation operator. These operators approximate the process of mapping floating-point values to lower-precision integers and then back again. During training, the parameters and activations are passed through these quantisation steps so that the model learns robust weight distributions and feature representations to reduced precision. For a given variable x (which could be a weight or an activation), the quantisation process can be conceptually described as [11]:

$$\widehat{x} = DQ(Q(x)), \tag{6}$$

where, $Q(x)$ converts the floating-point value x into an integer approximation based on observed ranges and a scale factor

and $DQ(\cdot)$ maps the integer back to a floating-point representation, simulating the effect of low-precision inference. When quantisation applied to weights and activations, these become:

$$\widehat{\mathbf{W}}_l = DQ(Q(\mathbf{W}_l)), \quad \widehat{\mathbf{a}}_u^{(l)} = DQ(Q(\mathbf{a}_u^{(l)})). \tag{7}$$

Hence, the forward pass with QAT integrated is:

$$\mathbf{z}_u^{(1)} = \widehat{\mathbf{S}}_u \widehat{\mathbf{W}}_1 + \widehat{\mathbf{b}}_1, \quad \widehat{\mathbf{a}}_u^{(1)} = \text{ReLU}(\mathbf{z}_u^{(1)}), \tag{8}$$
$$\mathbf{z}_u^{(2)} = \widehat{\mathbf{a}}_u^{(1)} \widehat{\mathbf{W}}_2 + \widehat{\mathbf{b}}_2, \quad \hat{\mathbf{z}}_u = \mathbf{z}_u^{(2)}.$$

Here, $\widehat{\mathbf{S}}_u$, $\widehat{\mathbf{W}}_l$, and $\widehat{\mathbf{b}}_l$ represent the quantised-and-dequantised inputs, weights, and biases. During training, these quantisation steps ensure the network learns how to operate effectively in the quantised domain.

c) Training Objective with QAT: As beam prediction is a multi-class classification problem, a standard cross-entropy loss function is used, mathematically denoted as:

$$\mathcal{L}(\Theta) = -\frac{1}{U} \sum_{u=1}^{U} \log \left(\frac{\exp(\hat{z}_{u,w_u^\star})}{\sum_{q=1}^{Q} \exp(\hat{z}_{u,q})} \right), \tag{9}$$

where $\hat{z}_{u,q}$ denotes the logit corresponding to the q-th beam index for the u-th sample. Gradient-based optimisers (e.g., Adam) are employed to update the full-precision parameters. The quantisation operations are treated as part of the computational graph, and straight-through estimators are used to handle the non-differentiability of the rounding steps in $Q(\cdot)$. This process ensures that the model parameters Θ are adjusted to minimise $\mathcal{L}(\Theta)$ while simultaneously adapting to the constraints imposed by quantisation.

IV. Dataset and Simulation Setup

This section presents the dataset, simulation environment and performance metrics used to evaluate the proposed SemQNet for beam prediction in mmWave communication systems.

A. Dataset Description

The performance of the proposed SemQNet framework is evaluated using a publically available DeepSense 6G dataset [7], a widely recognised benchmark for sensing-aided wireless communication research. The dataset contains multi-modal data collected in a real-world wireless environment, including mmWave wireless communication signals, GPS coordinates, and RGB images. Data acquisition was performed using a sophisticated hardware testbed consisting of a stationary unit (BS) and a mobile unit (vehicle). The stationary unit is equipped with a 16-element 60 GHz mmWave phased array and an RGB camera. In comparison, the mobile unit acts as a transmitter with a 60 GHz quasi-omni antenna and a GPS receiver for logging real-time location information.

Each sample in the dataset includes the transmitter's GPS position, an RGB image capturing the surrounding environment, and the mmWave receive power vector corresponding to a predefined beamforming codebook. Our analysis focused exclusively on scenario 5 of the DeepSense 6G dataset.

TABLE I
Simulation Parameters

Parameter	Value
Batch Size	128
Learning Rate	1×10^{-2}
Learning Rate Decay	Epochs 15 and 30
Total Epochs	50
Learning Rate Reduction Factor	0.1

This scenario features measurements collected at night in an urban setting, creating diverse visual and communication conditions for testing the framework's robustness. Scenario 5 contains a total of 2,300 samples, which are divided into training, validation, and testing subsets in a 70/20/10 ratio to ensure a balanced evaluation. The dataset provides a rich and challenging testbed for assessing the proposed semantic-aware quantised network, leveraging multi-modal insights under varying conditions.

B. Simulation Setup

The simulations for the proposed SemQNet framework were conducted to evaluate its performance in predicting optimal beam indices using semantic features extracted from visual data. As outlined in the section III, image preprocessing involves converting RGB images to grayscale and compressing them into a lossy JPEG format. Semantic extraction is performed using a pre-trained MobileNetV2 model, while the extracted bounding box features are fed into a 2-layer fully connected DL model trained with QAT, which employs 8-bit integer precision to simulate low-precision arithmetic, ensuring model robustness and efficiency for resource-constrained platforms. The reason for choosing this DL model architecture to perform a fair comparison with the pioneer work proposed in [4]. Additionally, we also utilised the other semantic representation image mask, extracted using the MobileNetV2. The model configuration and simulation setup is same as given in [4]. Moreover, the DL model uses the ReLU activation function and is trained using the Adam optimiser with a cross-entropy loss function. The simulations were implemented in PyTorch with GPU acceleration enabled, utilising an Intel RTX 4060 GPU. The key simulation parameters are summarised in Table I. This simulation setup ensures a rigorous evaluation of the proposed framework's performance under realistic conditions, leveraging state-of-the-art tools and techniques for semantic-aware beam prediction.

C. Performance Metric

The evaluation of the proposed SemQNet framework for semantic-aware beam prediction is based on three key performance metrics: top-k accuracy and energy efficiency. These metrics provide a comprehensive understanding of the framework's predictive accuracy, resource efficiency, and computational feasibility in real-world scenarios.

1) Top-k Accuracy: Top-k accuracy is the primary metric used to evaluate the predictive performance of the proposed framework. It is defined as the percentage of test samples with the ground-truth beam index within the top-k predicted beam indices. This metric assesses the ability of the framework to rank the optimal beam index highly among its predictions. For comprehensive evaluation, we report top-1, top-2, and top-3 accuracies.

2) Energy Efficiency: Energy efficiency is critical for deploying the proposed framework in energy-constrained environments. It is measured as the average electrical energy consumption required to transfer raw data or model parameters over a wireless link, expressed in kilowatt-hours per gigabyte (kWh/GB). The energy efficiency is estimated using the following equation:

$$E_{\text{est}} = N\big[(\alpha \times t_c) + (\beta \times P_{\text{trn}})\big], \tag{10}$$

where, t_c is the computation time for training, N is the number of sharing iteration, P_{trn} denotes size of model parameters, α is computation constant (0.003), and β is communication constant (0.0001) [12].

V. Results and Discussions

This section focuses on the performance evaluation of proposed SemQNet framework using the metrics discussed in Section IV-C.

A. Top-k Beam Prediction Performance

The results in Fig. 2 present the top-k beam prediction accuracies achieved by four modalities: Vision, bbox, Mask, and the proposed SemQNet framework. The comparison evaluates the effectiveness of each approach in identifying the optimal beam index under a consistent simulation setup for fair benchmarking, as used in the base paper [4]. For Vision, a ResNet-50 model processes raw RGB images; for bbox, a two-layer FCNN is employed; Mask utilises a LeNet-based CNN architecture; and SemQNet leverages QAT-FCNN with semantic information from compressed greyscale images.

The top-1 accuracy results highlight that SemQNet achieves an accuracy of 57.02%, closely matching the performance of Vision 58.1% and bbox 57.75%, with only a negligible drop. This performance underscores the ability of SemQNet to maintain accuracy despite the use of compressed input data and an energy-efficient architecture. Conversely, the Mask-based approach achieves a significantly lower top-1 accuracy of 47.5%, highlighting the limitations of binary mask representations for beam prediction. Additionally, the increasing trend in accuracy from top-1 to top-3 across all modalities demonstrates the ability of these models to capture a broader range of plausible beam indices. This is particularly important in practical scenarios where the wireless environment is dynamic, and providing multiple high-probability beam options can enhance decision-making reliability.

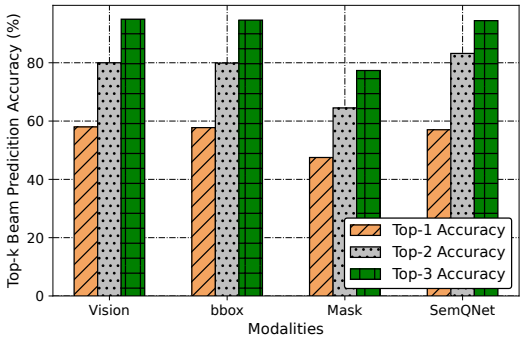

Fig. 2. This figure plots top-k accuracies for four modalities. For comparison, the modalities vision, bbox, and Mask from base paper [4] are also plotted with our proposed SemQNet.

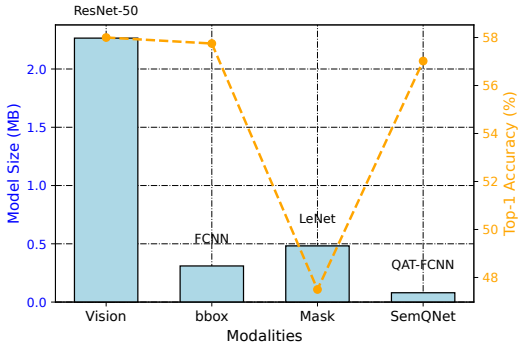

Fig. 3. This figure plots top-1 accuracy vs model size in (MB) for four modalities.

B. Model Size vs. top-1 Accuracy

The results in Fig. 3 illustrate the relationship between model size (in MB) and top-1 accuracy for four modalities used in the simulation setup. This comparison highlights the trade-off between accuracy and model complexity, showcasing the efficiency of SemQNet in achieving competitive accuracy with significantly reduced model size. For instance, the Vision modality processes raw RGB images using a ResNet-50 architecture and achieves the top-1 accuracy of 58.1%, but at the cost of the largest model size, 2.265 MB. In contrast, the bbox modality, utilising a two-layer FCNN for bounding box features, achieves a comparable top-1 accuracy of 57.75% with a much smaller model size of 0.31 MB, making it significantly more efficient. However, our proposed SemQNet framework achieves a top-1 accuracy of 57.02%, closely matching Vision and bbox, with the smallest model size of only 0.08 MB. This result underscores the impact of incorporating QAT, which reduces the memory and energy footprint of model training without compromising accuracy.

C. Energy Efficiency

As wireless systems are deployed over large geographical areas, data collection and model training are often performed in centralised locations, requiring edge nodes to transmit raw

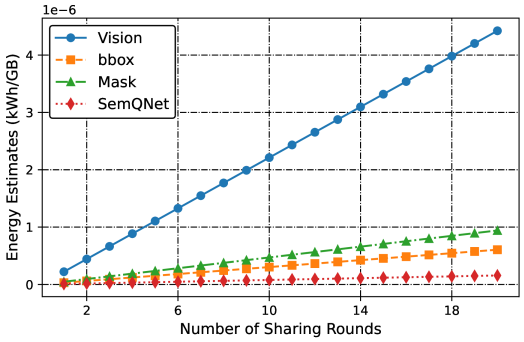

Fig. 4. This figure plots the energy estimates in (kWh/GB) vs number of share rounds during training.

data to a central server. This centralised approach, however, incurs significant communication overhead, particularly as the volume of data and number of iterations grow. An alternative, energy-efficient paradigm is distributed learning, where edge nodes perform semantic extraction and local model training before sharing only the model parameters with a central server for aggregation. Hence, to evaluate the effectiveness of our proposed SemQNet framework, we performed a simulation in a distributed manner where the edge node shared the model parameter to a centralised server over 20 iterations. The cumulative energy consumption was estimated by simplifying the energy estimate energy given in Section IV-C. Here, we considered the communication overhead only resulting the $E_{est} = N(\beta \times P_{\text{trn}})$.

The results, plotted in Fig. 4, shows the linear relationship between energy consumption and communication rounds. The model size P_{trn} emerges as the primary factor influencing energy expenditure, with larger models incurring greater communication overhead. For instance, Vision demonstrates the steepest growth with a cumulative energy estimate of 4.42×10^{-6} due to its large model size. In contrast, the bbox and Mask exhibit moderate energy consumption, with cumulative of 6.05×10^{-7}, and 9.41×10^{-7}, respectively. Finally, the proposed SemQNet, with its highly compact model size, achieves the lowest cumulative energy consumption of 1.56×10^{-7}, minimising communication overhead while maintaining competitive predictive performance.

VI. Conclusion

In this work, we proposed SemQNet, a novel framework for semantic-aware beam prediction in mmWave communication systems. The framework incorporates an efficient image preprocessing pipeline that converts RGB images to grayscale and applies lossy JPEG encoding to enable semantic information extraction using a lightweight MobileNetV2 model. The extracted bounding box features are then used to train a FCNN optimised with QAT, which reduces the model size and enables low-precision arithmetic, transitioning from float32 to int8 operations. The proposed framework was evaluated on the DeepSense 6G dataset, achieving top-1 accuracy comparable to Vision-based (ResNet-50) and bbox-based (FCNN) approaches. Despite using compressed images and lightweight architecture, SemQNet demonstrated competitive predictive performance, validating its effectiveness. Furthermore, SemQNet achieved a 74.21% reduction in model size, significantly lowering communication overhead in distributed learning scenarios. This work highlights the potential of SemQNet for resource-constrained, low-latency applications, paving the way for energy-efficient and scalable solutions in next-generation wireless communication systems. Future efforts will focus on extending the latency analysis and exploring further optimisation to enhance real-time performance in highly dynamic environments.

Funding and Support

The research is funded by EPSRC & DSIT funded projects EP/X040518/1, EP/Y037421/1, and EP/Y019229/1.

References

[1] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless Communications and Applications Above 100 GHz: Opportunities and Challenges for 6G and Beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.

[2] D. Nguyen, M. Ding, P. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated Learning for Internet of Things: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, pp. 1622–1658, 2021.

[3] S. Imran, G. Charan, and A. Alkhateeb, "Environment Semantic Communication: Enabling Distributed Sensing Aided Networks," *arXiv preprint arXiv:2402.14766*, 2024.

[4] S. Imran, G. Charan, and A. Alkhateeb, "Environment Semantic Aided Communication: A Real World Demonstration for Beam Prediction," in *2023 IEEE ICC Workshops*, pp. 48–53.

[5] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Multilevel Millimeter Wave Beamforming for Wireless Backhaul," in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, 2011, pp. 253–257.

[6] A. Alkhateeb, O. El Ayach, G. Leus, and R. Heath, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems," *IEEE Journal of Selected Topics in Signal Processing*, pp. 831–846, Oct. 2014.

[7] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "DeepSense 6G: A Large-Scale Real-World Multi-Modal Sensing and Communication Dataset," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 122–128, 2023. doi: 10.1109/MCOM.006.2200730.

[8] G. Charan, M. Alrabeiah, T. Osman, and A. Alkhateeb, "Camera Based mmWave Beam Prediction: Towards Multi-Candidate Real-World Scenarios," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2024.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[10] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, Springer, 2014, pp. 740–755.

[11] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "HAWQ: Hessian AWare Quantization of Neural Networks With Mixed-Precision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[12] A. N. Mian, S. W. H. Shah, S. Manzoor, A. Said, K. Heimerl, and J. Crowcroft, "A Value-Added IoT Service for Cellular Networks Using Federated Learning," *Computer Networks*, vol. 213, p. 109094, 2022.