On the Potential of Entropy-Constrained Vector Quantization in Semantic Communication

Junyong Shin and Yo-Seb Jeon
Department of Electrical Engineering, POSTECH, Pohang, Gyeongbuk 37673, Republic of Korea
Email: {sjyong, yoseb.jeon}@postech.ac.kr

***Abstract*—This paper investigates the benefit of integrating entropy coding into the vector quantization (VQ) framework for semantic communication. When using a learnable VQ codebook jointly trained with the semantic encoder and decoder, the resulting codeword distribution is typically non-uniform. Motivated by this observation, we incorporate entropy coding into the learned VQ structure to enhance compression efficiency. Specifically, we adopt an entropy-constrained vector quantization (ECVQ) framework that leverages probabilistic modeling to improve rate-distortion performance. Simulation results show that ECVQ outperforms conventional VQ-based methods, validating its effectiveness in improving rate-distortion efficiency.**

I. Introduction

Semantic communication (SC) has recently emerged as a paradigm shift in wireless communication, aiming to transmit the semantic meaning of data rather than its raw form. This shift allows communication systems to significantly reduce transmission overhead while preserving task-relevant information, making it particularly effective in many applications [1]. Within this context, digital semantic communication has garnered increasing attention as it enables semantic transmission over existing digital infrastructures. This is typically achieved by extracting semantic features using neural networks in deep learning, and then mapping these features into discrete representations through quantization [2]–[4].

Among quantization approaches, vector quantization (VQ) has proven to be an effective tool for compressing high-dimensional semantic features into finite-bit representations [5]. One promising methodology is to jointly train a neural encoder, decoder, and a learnable VQ module, which allows the quantization process to adapt to the feature distribution in a data-driven manner. By integrating the VQ module directly within the semantic feature space, this approach enables the design of quantizers that are well-matched to the underlying data distribution, thereby enhancing compression efficiency and overall system performance [6]–[9].

One critical challenge in applying VQ to SC systems is achieving rate-distortion optimality. Entropy coding, when applied to the output of a VQ module, can effectively reduce the average number of bits required for transmission by assigning shorter bit sequences to frequently used codewords. However, simply applying entropy coding as a post-processing step does not guarantee optimal compression efficiency, as conventional VQ schemes are not designed to account for the underlying codeword distribution during quantization. To address this limitation, entropy-constrained vector quantization (ECVQ) integrates entropy modeling directly into the quantization process [10]. By incorporating codeword probabilities into the quantization criterion, ECVQ guides the quantizer to favor high-probability codewords, thereby achieving more efficient bit allocation and improved rate-distortion performance. This approach has proven effective in various deep learning-based VQ frameworks, where quantization and entropy coding are jointly optimized for end-to-end compression efficiency [11].

In this paper, we investigate the potential of ECVQ to enhance communication efficiency in semantic communication. Following the approach in [11], we adopt entropy coding as a post-processing step of the VQ module, which is jointly trained with the semantic encoder and decoder. To support this process, we employ a modified VQ criterion that incorporates a rate bias term. This rate bias is modeled using trainable parameters to reflect the distribution of codewords. Simulation results demonstrate that the ECVQ framework achieves superior rate-distortion performance compared to existing VQ-based methods, validating its effectiveness in improving rate-distortion efficiency.

II. System Model

In this section, we present a digital semantic communication system considered in our work. To perform an image reconstruction task with a given data sample $\mathbf{x} \in \mathbb{R}^{M_0}$, the transmitter first employs a semantic encoder network $f_{\rm enc}(\cdot, \boldsymbol{\theta})$ to extract a semantic feature $\mathbf{z} \in \mathbb{R}^{M}$, given by

$$\mathbf{z} = f_{\rm enc}(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^{M}. \tag{1}$$

To represent $\mathbf{z}$ with finite bit precision, we assume that a quantization method is applied to the semantic feature $\mathbf{z}$, producing a quantized semantic feature $\mathbf{z}_q$. The bit sequence associated with $\mathbf{z}_q$ is then transmitted to the receiver over an error-free communication channel. The specific quantization approach adopted in our proposed framework will be described in detail in the following section. At the receiver, the quantized feature $\mathbf{z}_q$, is processed by a semantic decoder network $f_{\rm dec}(\cdot, \boldsymbol{\phi})$ to reconstruct the source sample, denoted by

$$\hat{\mathbf{x}} = f_{\rm dec}(\mathbf{z}_q, \boldsymbol{\phi}), \tag{2}$$

This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00453301) and in part by a grant of Korean ARPA-H Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare (No. RS-2025-25454431).

where $\hat{\mathbf{x}}$ is the final reconstructed output.

III. Entropy-constrained Vector Quantization (ECVQ) in SC

The ECVQ method was developed to account for the rate-distortion optimality in VQ under entropy-coded outputs. This is motivated by the fact that simply applying entropy coding to conventional VQ—based solely on minimum Euclidean distance—does not inherently guarantee rate efficiency [10]. To address this point in SC framework, we first divide the semantic feature $\mathbf{z}$ into N sub-vectors, each of dimension D, such that $M = N \times D$. Let $\mathbf{z}_i$ denote the i-th sub-vector of $\mathbf{z}$, defined as $\mathbf{z}_i = [z_{(i-1)D+1}, \cdots, z_{iD}]$, where z_j is the j-th entry of $\mathbf{z}$. Each sub-vector is quantized independently using a D-dimensional VQ codebook $\mathcal{B} = \{\mathbf{b}_k\}_{k=1}^{K}$. Let $\mathbf{z}_{q,i}$ be the quantization output of the VQ module for $\mathbf{z}_i$. The quantized sub-vectors are then entropy coded to produce corresponding bit sequences $\hat{\mathbf{z}}_{q,i}$, which are concatenated to form the final transmission sequence as $\hat{\mathbf{z}}_q = [\hat{\mathbf{z}}_{q,1}, \cdots, \hat{\mathbf{z}}_{q,N}]$.

To quantize each sub-vector, the ECVQ method adopts the following VQ criterion:

$$\mathbf{z}_{q,i} = \underset{\mathbf{b}_k \in \mathcal{B}}{\operatorname{argmin}} \left\{ -\nu \log_2 P(k) + \|\mathbf{z}_i - \mathbf{b}_k\|^2 \right\}, \tag{3}$$

where $P(k)$ denotes the probability distribution of codeword $\mathbf{b}_k$ on the VQ codebook. This distribution can be modeled as

$$P(k) = \frac{e^{-w_k}}{\sum_{p=1}^{K} e^{-w_p}}, \tag{4}$$

where $\{w_k\}_{k=1}^{K}$ are trainable parameters. In this formulation, the first term in (3), referred to as the *rate bias* term, contributes to shift the quantization boundaries from high-probability to low-probability regions in the VQ criterion. By introducing this term, the quantization results can be more concentrated on high-probability codewords, ultimately reducing the bit length after entropy coding. This term is regulated by the hyperparameter ν, which controls the extent to which the rate is reduced within the overall VQ criterion. When $\nu = 0$, the rate bias is ignored, and the criterion reduces to the conventional VQ objective without rate considerations.

To jointly train the encoder and decoder networks with the ECVQ module, we design the loss function $\mathcal{L}_{\text{ecvq}}$ as follows:

$$\mathcal{L}_{\text{ecvq}} = \mathcal{L}_{\text{vq}} - \nu(1+\beta)\mathbb{E}_{k|\mathbf{z}}\left[\log_2 P_{k|\mathbf{z}}\right], \tag{5}$$

$$\mathcal{L}_{\text{vq}} = \left\|\hat{\mathbf{x}} - \mathbf{x}\right\|^2 + \left\|\text{sg}(\mathbf{z}) - \mathbf{z}_q\right\|^2 + \beta \left\|\mathbf{z} - \text{sg}(\mathbf{z}_q)\right\|^2, \tag{6}$$

where β is a hyperparameter and $\text{sg}(\cdot)$ denotes the stop-gradient operator, which treats its input as a constant during backpropagation, thereby preventing gradient updates. Moreover, $P_{k|\mathbf{z}}$ represents the conditional probability distribution of $\mathbf{b}_k$ given $\mathbf{z}$, corresponding only to the codewords that are actually selected as quantization outputs.

IV. Simulation Results

In this section, we evaluate the performance of the ECVQ method in SC. We utilize the CIFAR-10 image dataset in simulation. The sizes of training and inference datasets are configurated as 50,000 and 10,000, respectively. The Adam optimizer is employed with 10^{-3} learning rate, 128 training epochs, and 64 batch size. Other key parameters for the ECVQ method are set as follows: $D = 4$, $K = 256$, $\beta = 0.25$, and $\nu \in \{\frac{1}{8}, \frac{1}{16}, \frac{1}{24}\}$. In our simulation, peak signal-to-noise ratio (PSNR) is adopted as a performance measure, which is defined as

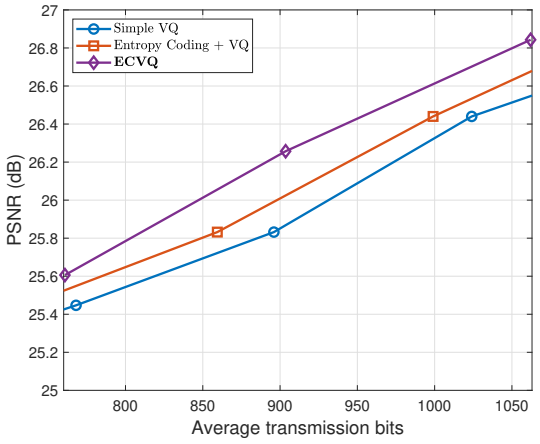

Fig. 1. PNSR performances of various VQ frameworks for semantic communications across different average transmission bits.

$$\text{PSNR} = 10 \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}^2}\right) \text{ (dB)}, \tag{7}$$

where MAX denotes the maximum possible pixel value of the image (e.g., 255 for 8-bit image), and MSE represents the mean-squared-error between the input image and reconstructed image. To perform the entropy coding in the inference stage, a mapping table for the Huffman coding is constructed for each VQ codebook based on the empirical distribution of the outputs of each trained VQ module from all training data.

Fig. 1 presents the PSNR performance of the ECVQ method in SC across varying average transmission bit rates. As illustrated, the ECVQ method consistently outperforms other baselines. The performance gap between ECVQ-based structures and the simple entropy-coded VQ baseline underscores the effectiveness of the ECVQ criterion and its corresponding training strategy. Merely applying entropy coding to the VQ output yields only marginal performance gains, as this approach does not account for rate-distortion optimality. In contrast, ECVQ explicitly incorporates rate-distortion trade-offs into both its quantization criterion and loss function, leading to substantial improvements over standard entropy-coded VQ.

V. Conclusion

This paper examined the use of entropy-constrained vector quantization (ECVQ) for semantic communication. By incorporating entropy modeling into the quantization process, ECVQ achieves superior rate-distortion efficiency compared to conventional VQ. Simulation results confirmed that explicit entropy-aware design substantially improves compression per-

formance, highlighting ECVQ as a promising approach for efficient semantic communication.

REFERENCES

[1] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 210–219, Feb. 2022.

[2] M. Rao, N. Farsad, and A. Goldsmith, "Variable length joint source-channel coding of text using deep neural networks," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Aug. 2018, pp. 1–5.

[3] K. Choi, K. Tatwawadi, A. Grover, T. Weissman, and S. Ermon, "Neural joint source-channel coding," in *Proc. Int. Conf. Machine Learning (ICML)*, Jun. 2019, pp. 1182–1192.

[4] J. Park, H. Kim, J. Shin, Y. Oh, and Y.-S. Jeon, "End-to-end training and adaptive transmission for OFDM-based semantic communication," *ICT Exp.*, early access, doi: 10.1016/j.icte.2025.05.005.

[5] J. Shin, Y. Oh, J. Park, J. Park, and Y.-S. Jeon, "ESC-MVQ: End-to-end semantic communication with multi-codebook vector quantization," *IEEE Trans. Wireless Commun.,* early access, doi: 10.1109/TWC.2025.3605838.

[6] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6306–6315.

[7] J. Shin, Y. Kang, and Y.-S. Jeon, "Vector quantization for deep-learning-based CSI feedback in massive MIMO systems," *IEEE Wireless Commun. Lett.,* vol. 13, no. 9, pp. 2382–2386, June 2024.

[8] J. Shin and Y.-S. Jeon, "Error-robust deep learning-based CSI feedback in massive MIMO systems: A multi-rate vector quantization approach," in *Proc. 15th Int. Conf. on ICT Convergence (ICTC),* Oct. 2024, pp. 1–3.

[9] J. Shin, E. Jeon, I. Kim, and Y.-S. Jeon "Deep learning-based CSI feedback for Wi-Fi systems with temporal correlation," *IEEE Trans. Commun.,* early access, doi: 10.1109/TCOMM.2025.3600202.

[10] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 37, no. 1, pp. 31–42, Jan. 1989.

[11] J. Shin, J. Park, and Y.-S. Jeon, "Entropy-constrained VQ-VAE for deep-learning-based CSI feedback," *IEEE Trans. Veh. Technol.*, vol. 74, no. 6, pp. 9870–9875, June 2025.