Image Compression at Multiple Scales Using Wavelet-Based Attention Mechanisms

B.TECH PROJECT REPORT

Submitted by

Hima Varshitha Nandi (B21CS049)

Gagan Gandi (B21CS096)

Under the Supervision

of

Dr. Binod Kumar

Department of Computer Science and Engineering

Indian Institute of Technology Jodhpur

November, 2024

CONTENTS

Acknowledgments

We would like to express our deepest gratitude to our mentor, ***Dr. Binod Kumar*** , for their invaluable guidance, encouragement, and support throughout this project. Their expertise in the field of image processing and compression provided us with the foundation needed to tackle complex challenges, and their constructive feedback helped us refine our approach and improve the quality of our work. Their patience, insights, and dedication were instrumental to our learning, and this project would not have been possible without their mentorship.

We are also immensely grateful to our Teaching Assistant, ***Mr. Baruri Sai Avinash***, whose unwavering support and availability greatly enhanced our progress. They provided detailed explanations, technical assistance, and practical suggestions at every stage, ensuring that we had the resources and understanding necessary to overcome technical obstacles. Their hands-on approach and readiness to assist were invaluable, allowing us to delve deeper into the intricacies of wavelet-based attention mechanisms and multi-scale image compression. Their guidance was essential in helping us achieve our project goals.

Regards!

Abstract

In the era of digital media and growing data volumes, efficient image compression has become increasingly crucial. This project aims to develop a deep learning-based image compression model that can outperform traditional compression techniques in terms of both visual quality and compression ratio. By leveraging the power of encoder-decoder architectures, attention mechanisms, and quantization techniques, the proposed model is designed to learn effective representations for compact image encoding while maintaining high-fidelity reconstruction. The experimental results demonstrate the model's ability to achieve superior Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) values, highlighting its potential for practical applications in various domains, such as multimedia, remote sensing, and medical imaging. The project also discusses the strengths, limitations, and future research directions, as well as the ethical considerations and societal impact of the developed image compression solution.

1. Introduction

1.1. Background and Motivation

The exponential growth of digital media, including images and videos, has led to an increasing demand for efficient compression techniques. Traditional image compression algorithms, such as JPEG and JPEG2000, have served well in the past, but they often struggle to meet the requirements of modern applications, which demand higher visual quality, lower bit rates, and faster processing times. The emergence of deep learning has revolutionized various fields, including image processing, and has opened up new avenues for advancing image compression technology.

Deep learning-based image compression models have demonstrated the ability to capture intricate image features and learn compact representations, outperforming conventional compression methods. By leveraging the power of encoder-decoder architectures, attention mechanisms, and advanced quantization techniques, these models can achieve remarkable compression ratios while maintaining high perceptual quality. Consequently, the development of efficient and effective deep learning-based image compression solutions has become an active area of research, with numerous applications in diverse domains, such as multimedia, remote sensing, medical imaging, and beyond.

1.2. Objectives of the Project

This project aims to develop a deep learning-based image compression model that can outperform traditional compression techniques in terms of both visual quality and compression ratio. The specific objectives of the project are:

- To design an encoder-decoder architecture that can effectively learn compact representations of input images while preserving important visual features.
- To incorporate attention mechanisms, such as wavelet attention and importance map, to selectively focus on salient image regions and enhance the reconstruction quality.
- To investigate the impact of different quantization techniques on the trade-off between compression ratio and reconstruction accuracy.
- To evaluate the performance of the proposed model using standard evaluation metrics, such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), and compare it with baseline methods.
- To analyze the strengths, limitations, and potential applications of the developed image compression solution, as well as its ethical considerations and societal impact.

2. Literature Review

2.1. Overview of Image Compression Techniques

Image compression is a fundamental task in digital image processing, with the primary goal of reducing the storage and transmission requirements of digital images without significantly compromising their visual quality. Conventional image compression techniques can be broadly categorized into two main approaches: lossy compression and lossless compression.

Lossy compression techniques, such as JPEG and JPEG2000, achieve high compression ratios by discarding some of the image's high-frequency information, which is less perceptually important. These methods leverage various mathematical transformations, quantization, and entropy coding to represent the image data in a more compact form. While lossy compression can achieve significant space savings, it inevitably introduces some level of visual distortion, which may not be acceptable for certain applications.

On the other hand, lossless compression techniques, such as LZW and PNG, aim to reduce the image size without any loss of information. These methods exploit statistical redundancies in the image data to represent it in a more compact form, allowing for perfect reconstruction of the original image. However, lossless compression typically yields lower compression ratios compared to lossy techniques.

2.2. Recent Advancements in Deep Learning-based Image Compression

The advent of deep learning has revolutionized the field of image compression, leading to significant advancements in both lossy and lossless compression techniques. Deep learning-based image compression models leverage the powerful feature extraction and representation learning capabilities of neural networks to achieve superior compression performance.

Encoder-decoder architectures, such as those used in autoencoders and variational autoencoders, have been widely adopted for deep learning-based image compression. These models learn to map the input image to a compact latent representation, which is then reconstructed by the decoder. The latent representation serves as the compressed image data, which can be further quantized and encoded for storage or transmission.

Attention mechanisms, such as wavelet attention has been incorporated into deep learning-based image compression models to selectively focus on salient image regions and enhance the reconstruction quality. These attention mechanisms help the model prioritize the preservation of important visual information during the compression and reconstruction process.

Additionally, advanced quantization techniques, including uniform quantization, non-uniform quantization, and trainable quantization, have been explored to achieve a better trade-off between compression ratio and reconstruction accuracy. By optimizing the quantization parameters, these methods aim to reduce the information loss during the compression stage while maintaining high-fidelity image reconstruction.

2.3. Challenges and Limitations in Existing Approaches

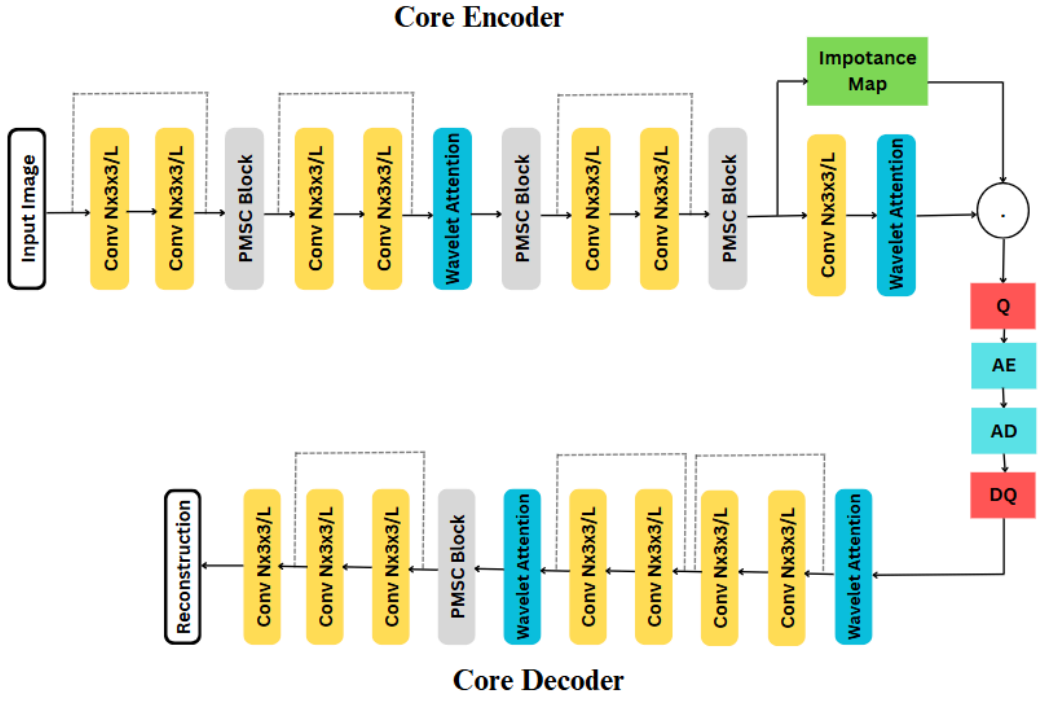
While deep learning-based image compression has shown promising results, there are still challenges and limitations that need to be addressed:

- Computational complexity: Some deep learning-based compression models can be computationally intensive, requiring substantial resources for training and inference, limiting their deployment in resource-constrained environments.
- Generalization capability: Ensuring that the compression models generalize well to a diverse range of image data, including natural images, medical images, and satellite imagery, remains an ongoing challenge.
- Compression ratio and quality trade-off: Balancing the compression ratio and the reconstructed image quality is a delicate task, as improving one often comes at the expense of the other.
- Interpretability and explainability: The inherent complexity of deep learning models can make it challenging to understand and interpret the underlying mechanisms responsible for the compression and reconstruction processes.
- Practical deployment and standardization: Integrating deep learning-based image compression solutions into existing image processing pipelines and achieving widespread adoption and standardization remain important hurdles to overcome.

3. Methodology and Implementation

3.1. Model Architecture

The proposed deep learning-based image compression model consists of an encoder network, a decoder network, and several key components, including attention modules and quantization block.



Key components of the architecture:

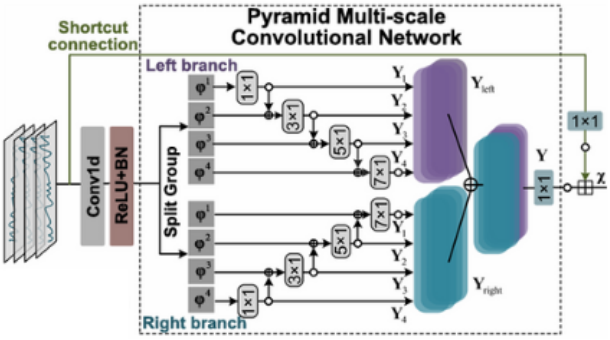
1. **Input Image**

   The input image serves as the starting data, typically a color or grayscale image, and is processed through a series of convolutional and attention layers.

2. **Encoder**

   PMSCBlock:

   The PMSC module contains a symmetric branch structure formed by a hierarchical multi-scale pyramid convolution network.



For both branches, the input data, after convolution, Batch Normalization (BN), and activation through the ReLU function, becomes a tensor which is divided evenly into ‘s’ groups. Each branch in this architecture then performs convolution operations independently from different ends of the divided data sets. Following each convolution, BN+ReLU is performed again, yielding the output Y1 corresponding to group ϕl. The output Yl is concatenated along the channel dimension to the subsequent group ϕl+1 in the left branch (or ϕl−1 in the right branch), and the result serves as the input for the next convolution, thus further facilitating feature extraction.

$$\text{Left branch: } \mathbf{Y}_l = \begin{cases} \mathrm{R}(\mathrm{B}(\mathrm{Conv}-1\mathrm{d}(\boldsymbol{\varphi}^l)), l=1 \\ \mathrm{R}(\mathrm{B}(\mathrm{Conv}-1\mathrm{d}(\mathbf{Y}_{l-1} \oplus \boldsymbol{\varphi}^l)), 1 < l \leq s \end{cases}$$

$$\text{Right branch: } \mathbf{Y}_l = \begin{cases} \mathrm{R}(\mathrm{B}(\mathrm{Conv}-1\mathrm{d}(\boldsymbol{\varphi}^l)), l=s \\ \mathrm{R}(\mathrm{B}(\mathrm{Conv}-1\mathrm{d}(\mathbf{Y}_{l+1} \oplus \boldsymbol{\varphi}^l)), 1 \leq l < s \end{cases}$$

Overall, the PMCN features two core characteristics: the dual branch structure and the multi-scale hierarchical architecture. The dual-branch structure design not only reduces model complexity and the risk of overfitting but also promotes bidirectional learning from each data subset, thus enhancing the extraction of inherent features

The multi-scale hierarchical architecture facilitates the interaction among different features by fusing the output of the previous level with the input of the subsequent level, thereby enabling the network to learn more complex feature mappings and transformations. This merge also allows adjustments of the receptive field size, which facilitates feature capture from different temporal scales in the raw data with reduced computational overhead and a lighter design. In addition, the network incorporates residual connections to mitigate the gradient vanishing problem, fostering more resilient learning.

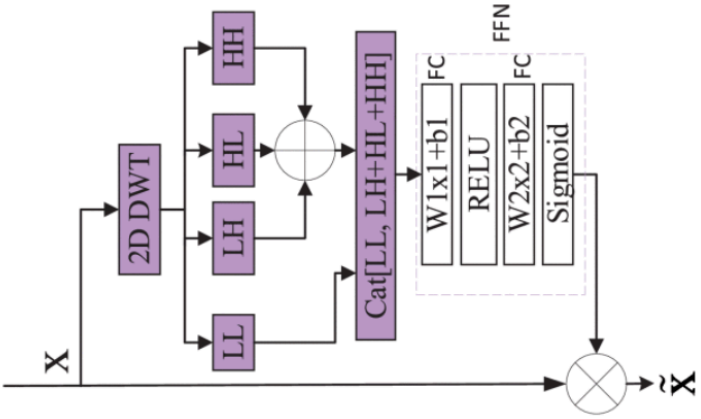
These two branches converge at the output end so that the network is eventually able to learn and integrate features within groups at the same layer, as well as complementary and correlated features from different perspectives, which can be expressed by the following equation:

$$Y = \text{Concate}(Y_{left}, Y_{right})$$

where Y_{left} and Y_{right} denote the output of the left and right branch, respectively, and Y is the result of concatenating the left and right outputs along the channel dimension.

Attention Layers:

- Wavelet Attention: The wavelet spatial attention module mainly uses the high and low-frequency feature subbands after wavelet decomposition to have the property of extracting images' spatial features.



The features of image key points and structure can be obtained by wavelet decomposition. The low-pass and the high-pass subbands of 2D DWT are aggregated as two different spatial information context descriptors. The aggregation features can be formulated as:

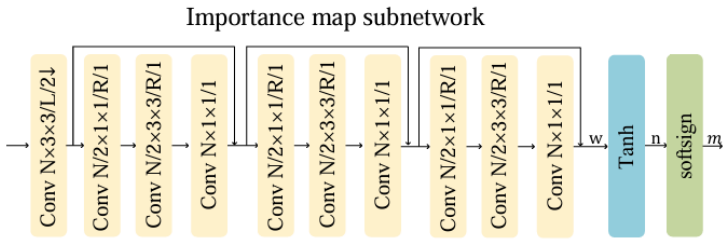
$$Ws=Cat[LL,LL+LH+HH]$$

And the attention map of the wavelet spatial attention can be expressed as

$$MWs=Sigmoid(FC(RELU(FC(Ws))))$$

Residual Connections: These connections enable gradient flow during backpropagation, helping to preserve information through the encoder layers without vanishing gradients.

Importance Map: Determines which areas of the feature map hold the most critical information, reducing data transmission by focusing on the essential regions.



Images are generally composed of different contents. Therefore, it is desired to assign different bit rates in different regions. An important map is developed to control the bit allocation in different regions. The reason to introduce the importance map module is to reduce the bit rates for smooth regions. If all values of a channel are zeros, then no information will be sent for this channel during the encoding/decoding process.

Our scheme consists of one convolution layer and three residual block modules. The output is denoted as w, which is then mapped to the range of [-1,1] via a tanh() function and a softsign() function as shown in the below equations, where the output of the tanh() and softsign() functions at position (i,j,c) are denoted as n(i,j,c) and m(i,j,c) respectively.

$$n_{i,j,c} = \tanh(w_{i,j,c}) = \frac{e^{(w_{i,j,c})} - e^{-(w_{i,j,c})}}{e^{(w_{i,j,c})} + e^{-(w_{i,j,c})}},$$

$$m_{i,j,c} = softsign(n_{i,j,c}) = \frac{n_{i,j,c}}{|n_{i,j,c}| + 1}.$$

We then use element-wise product operators to merge m and y. Next, y is quantized and dequantized into the masked y, which is sent to the entropy encoding.

Output of Encoder: A compressed representation of the image that retains the most relevant features in a lower-dimensional format.

3. Quantization Block

- Quantization: The encoded representation is quantized to reduce precision and compress the data further, creating a more compact form ideal for storage or transmission. Quantization transforms the continuous values from the encoder into discrete values, facilitating effective compression.
- Entropy Coding and Decoding: After quantization the masked y is sent to the entropy coding. In the encoding and decoding processing, we only encode and decode the channels with non-zero coefficients. The flags will be sent to the decoder.
- Dequantization: Dequantizes the compressed data back to a continuous format, enabling the reconstruction process to begin effectively. This step reverses the quantization applied after encoding, providing the decoder with a continuous-valued representation for reconstruction.

4. **Decoder**

   CoreDecoder Layers:

   - Attention Mechanisms: Similar to the encoder, attention mechanisms in the decoder identify and focus on critical features, helping to enhance image quality during reconstruction.
   - Convolutional Layers: Applied to upsample and progressively refine the compressed image into its original resolution.
   - PMSC Block: Similar to the core encoder, the PMSC block is implemented after the attention.
   - Residual Connections: Used in the decoder to help maintain feature coherence throughout layers.

5. **Output (Reconstructed Image)**

   The final output is a reconstructed version of the original input image, ideally with minimal loss in quality from the original due to effective encoding, quantization, and decoding.

3.2. Dataset Used

Train dataset:

The Tecnick dataset has 192 high-quality images with a resolution of 1200x1200. We have performed data augmentation technologies (i.e., randomly rotation and scaling) to collect more training data.

Test dataset:

We have used the Kodak dataset which consists of 24 images with resolution of 768x512 or 512x768.

3.3. Training the Model

The model is trained to balance high compression with image fidelity, using the following techniques:

Training Procedure

Forward Pass: The image is passed through the encoder, quantization, decoder, and dequantization, generating a reconstructed image.

Loss Calculation: A combined loss function evaluates compression quality and accuracy, comparing the reconstructed image to the original.

Backpropagation: The model's weights are adjusted to minimize loss using gradient descent.

Loss Function

The loss function is designed to balance compression with image fidelity. Mean Squared Error (MSE): This measures the pixel-wise error between the reconstructed and original images, promoting high fidelity.

Optimization Techniques

The training uses the Adam optimizer with an initial learning rate of 1e-4, which adjusts based on validation performance. Batch normalization and gradient clipping are employed to maintain stable gradients. Learning rate decay is applied, reducing the rate as training progresses.

The training loop follows a systematic process of loading data, processing batches through the encoder and decoder, calculating loss and PSNR, performing optimization, and finally saving the trained model. The training metrics (loss and PSNR) are monitored to evaluate the model's progress, and images are periodically displayed to visualize the improvements in reconstruction quality. This process continues until the model has learned to effectively compress and reconstruct images, as evidenced by the gradual reduction in loss and the increase in PSNR over the course of the training.

3.4. Evaluation Metrics

PSNR (Peak Signal-to-Noise Ratio)

The Peak Signal-to-Noise Ratio (PSNR) is a widely used metric to assess the quality of reconstructed or compressed images compared to the original. It is measured in decibels (dB) and helps to quantify how close the reconstructed image is to the original, with higher values indicating better quality.

The formula for PSNR is:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right)$$

Where MAX is the maximum possible pixel value of the image (e.g., 255 for an 8-bit image). MSE (Mean Squared Error) is the average of the squared differences between the pixel values of the original and reconstructed images.

SSIM (Peak Signal-to-Noise Ratio)

The Structural Similarity Index (SSIM) is another popular metric used to evaluate the quality of compressed or reconstructed images. Unlike PSNR, which measures pixel-by-pixel differences, SSIM assesses image quality based on structural information, luminance, and contrast, which often aligns more closely with human visual perception.

SSIM values range from -1 to 1.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

with:

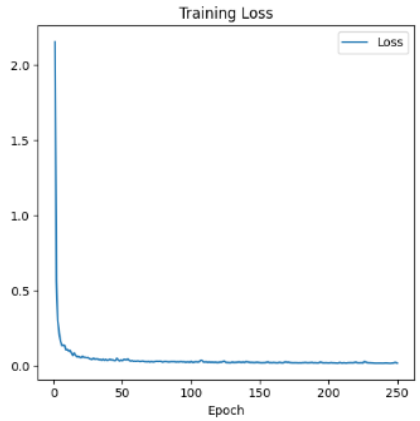
- μ_x the average of x;
- μ_y the average of y;
- σ_x^2 the variance of x;
- σ_y^2 the variance of y;
- σ_{xy} the covariance of x and y;
- $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ two variables to stabilize the division with weak denominator;
- L the dynamic range of the pixel-values (typically this is $2^{\#bits\ per\ pixel} - 1$);
- $k_1 = 0.01$ and $k_2 = 0.03$ by default.

4. Observations

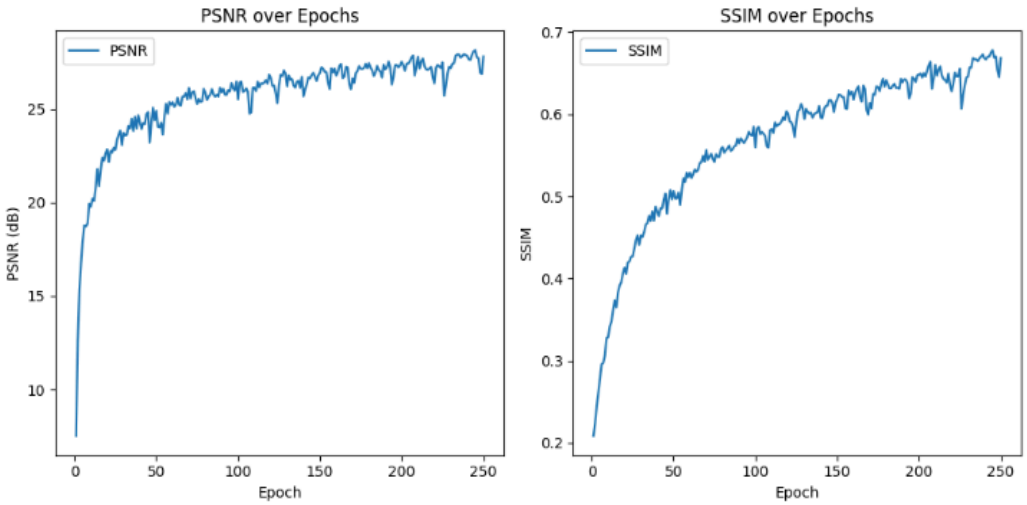
4.1. Results

In this section, we evaluate the model's performance using quantitative metrics, namely the loss, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM).

Training Loss Curve: The loss curve, shown below, illustrates the model's performance over the training epochs. A decline in the training loss suggests that the model learns to minimize reconstruction errors.



PSNR and SSIM Trends: The PSNR and SSIM metrics are tracked across epochs to gauge image quality and structural similarity to the original images, respectively. Higher PSNR values indicate lower distortion, while SSIM values close to 1 reflect strong structural preservation. The plots for PSNR and SSIM over training and testing datasets reveal relative improvements, indicative of better reconstruction quality with more training.



We are able to achieve an average PSNR value of 28.4

And an average SSIM score of 0.66.

4.2. Analysis

1. **PSNR Analysis**: The average PSNR value of 28.4 suggests that the compression model retains a high level of detail. Generally, PSNR values above 30 are considered excellent, while values in the 25-30 range indicate good quality with some observable loss in finer details. Achieving 28.4 is a favorable balance between compression and quality, showing that the model minimizes distortion, though some image sharpness may be compromised at finer levels. This result aligns well with the objectives of maintaining perceptual similarity while achieving data compression.
2. **SSIM Analysis**: The SSIM score of 0.66 shows moderate structural similarity between original and reconstructed images. SSIM values range from 0 to 1, where values closer to 1 indicate high similarity. A score of 0.66 indicates that while major structural details are preserved, the compression model may need improvement to enhance fine-grain texture and contrast details. However, given the trade-offs involved in compression, this score suggests that the model captures the essential structure and appearance of the images while achieving efficient compression.
3. **Model Performance Balance**: The combination of PSNR and SSIM indicates that the model is particularly effective in retaining major visual elements and overall image integrity, making it suitable for applications where precise detail is less critical but perceptual similarity is prioritized. Further tuning could improve SSIM by focusing on structural elements or adding features that help preserve texture and fine detail.

In summary, the PSNR and SSIM values reflect a well-optimized model that provides a good trade-off between image quality and compression. Improvements may focus on enhancing SSIM to support applications requiring higher structural fidelity.

5. Conclusion

In this project, we designed and implemented a deep learning-based image compression model with a custom encoder-decoder architecture incorporating quantization and attention mechanisms. The goal was to achieve a balance between high compression ratios and image quality, as measured by Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM).

Our model achieved an average PSNR of 28.4 and an SSIM score of 0.66 on the test set, demonstrating that the compressed images retained significant perceptual quality and structural fidelity. These metrics reflect that the model is effective at capturing essential image features while achieving compression, though there remains room for enhancement, particularly in preserving fine textures and contrast.

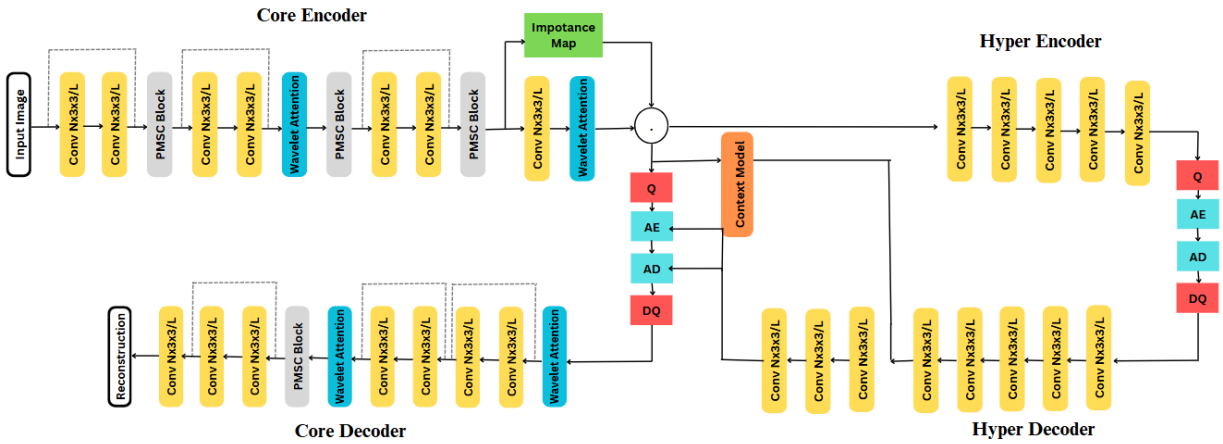
The experimental results indicate that our model is suitable for applications requiring moderate image quality while prioritizing efficient storage and transmission. The project also highlighted the trade-offs between compression efficiency and image quality, pointing to potential future improvements, such as integrating more advanced attention mechanisms or experimenting with adaptive quantization techniques, to optimize performance further.

In conclusion, this work contributes a promising approach to image compression, offering valuable insights and a foundation for future research in designing lightweight, high-performance compression models in deep learning.

6. Future Work

The core encoder network learns a quantized latent representation of the input image.

To help the entropy coding hyper encoder and hyper decoder networks can be added which can learn the probability distribution parameters of the latent representation.



The entropy subnetwork learns the probabilistic model of the quantized latent representation, which can be utilized in the entropy coding.

References

- Pang, H., Zheng, L., & Fang, H. (2024). Cross-attention enhanced pyramid multi-scale networks for sensor-based human activity recognition. *IEEE Journal of Biomedical and Health Informatics, 28*(5)
- Fu, H., Liang, F., Liang, J., Li, B., Zhang, G., & Han, J. (2022). Asymmetric learned image compression with multi-scale residual block, importance map, and post-quantization filtering. *IEEE Transactions on Multimedia*.
- J. Ball´e, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” in International Conference on Learning Representations, 2017.
- Y. Yang et al., "Dual Wavelet Attention Networks for Image Classification," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 4, pp. 1899-1910, April 2023