Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Optimizing Deep Learning Image Compression Models with Nano-Electronic Processing Units

Sivakumar. R.D

Assistant Professor (Senior Grade), Post Graduate Department of Computer Applications, Mepco Schlenk Engineering College, Sivakasi, rdsivakumarstaff@gmail.com

The proliferation of images during the advance in digital technology has heightened the need for image compression for storage and transmission purposes. Convolutional based image compression models have completely transformed this domain by providing improved compression rates while preserving the quality of the image. However, these models require rather heavy computational resources as a result of which they face scalability issues and can hardly be used for real-time applications. This work focuses on incorporating nano-electronic processing units (NEPUs) to enhance the deep learning-based image compression models. With their energy efficiency, compact design and high processing speed, NEPUs have the potential to resolve difficulties associated with the use of existing GPU based architectures. The study outlines a novel CNN-NEPU combination utilizing quantization-aware training and hardware-specific optimizations and adaptable learning that adds CNNs to receive NEPU-enhanced computations for dynamic image complexity. Evaluation on COCO and ImageNet datasets show improved speed by 25%, less energy usage by 30 % than normal strategies implemented using basic quality lossless compressors where PSNR and SSIM proves it has minimal quality loss with deeper insights into image information framework. The study shows that NEPU-accelerated frameworks can effectively improve the image compression with sustainable, scalable, and real-time performance that is applicable in limited-memory and limited-computational-resource environments such as edge computing and mobile applications. This research addresses the issue of mapping DNNs into hardware and existing systems and furthers the development of effective and efficient image processing techniques that may encourage farther research on energy-efficient and high-performance computing systems.

Keywords: Image compression, deep learning, nano-electronic processing units (NEPUs), hardware optimization, convolutional neural networks (CNNs), quantization-aware training, edge computing, energy efficiency, scalability, visual data processing.

1. Introduction

1.1 Overview of image compression and its importance

It is important to address image compression in large volume of picture data in different domains such as social networks, security, medicine, and multimedia communication. It refers to a process of compressing image files without affecting the inherent parameters that are needed for visualization, storage and transmission. As demands for Hi-Resolution images and videos rise, traditional storage and Networks become a harbinger of a bottleneck to applications that heavily rely on them. Data compression methods are central to this process since they ensure that these resources are utilized most effectively by providing the least amount of data redundancy and the least amount of bandwidth necessary. This is most important in near real-time application such as live streaming, video conference where Latency and Data transfer rate is critical. Further, cost control is critical to efficiency since compression decreases the storage price and power use in data centers. A novel method such as deep learning-based has been enhanced to provide higher levels of compression ratios compared to the conventional procedure while conserving the perceived quality. With today’s society realizing the potential of data in decision-making processes, communication, and entertainment, image compression is an enabling technology that supports the scalability, effectiveness, and efficiency of digital environments without compromising the growing demand for data-processing capabilities that overtax current and emergent technologies.

1.2 Emergence of deep learning in image compression

Image compression is another great area whose advancement has been enhanced by the introduction of deep learning since it uses new approaches that outperform others. Many existing compression techniques like the JPEG, PNG are signature-based and involve deterministic encode-decode, architectures which do not have the capability to accommodate really complex data patterning and have static form of compression that do not vary, depending on the type of image. Deep learning, on the other hand, relies on neural networks to classify such a learning method as well as to obtain hierarchical features of images to achieve adaptive image compression. CNN and GAN based architectures play a vital role in achieving very low bit rate compression while keeping the perceptions of lost data negligible. These models can learn to solve this trade-off at a global level, identifying sparsity patterns that will determine at every level of the pyramidal structure the amount of compression that does not kill visualization. Variational auto encoders as well as the attention mechanisms continue to improve compression algorithms by concentrating on meaningful image regions while minimizing unimportant areas’ redundancy. Not only that, the methods of deep learning also have high scalability and flexibility since they can be further optimized for certain area of applications, including medical image, remote sensing and video streaming. Thus, deep learning adoption in image compression paves not only for the increasing need for efficient high-quality signal transmission and storage but also for innovative intelligent resource-saving multimedia analysis in the age of Big Data.

1.3 Challenges in computational efficiency and scalability

Despite the advancements in deep learning for image compression, challenges in computational efficiency and scalability persist. Deep learning models are computationally intensive, requiring substantial processing power and memory, which limits their deployment on resource-constrained devices such as smartphones and IoT systems. High-resolution images exacerbate these challenges, as they demand significant computational resources to process complex data patterns. The latency introduced by these models poses difficulties in real-time applications like video streaming and telemedicine, where rapid data compression and transmission are critical. Scalability is another hurdle, as deploying deep learning models across diverse devices and networks often involves retraining or optimizing for varying hardware capabilities, which can be time-consuming and resource-intensive. Additionally, energy consumption in deep learning-based compression models remains a concern, especially in edge computing scenarios where power efficiency is paramount. Addressing these challenges is crucial to ensure the practical adoption of deep learning in image compression across a wide range of applications.

1.4 Role of Nano-Electronic Processing Units (NEPUs) in addressing these challenges

Thus, the deep learning-based approaches for image compression are still problematic in terms of computational complexity and scalability. Since deep learning involve numerous calculations, then the computational requirements are also high and this is not good news for smartphones, and IoT systems. These problems are especially amplified by high-resolution images because such data require extensive computational power to analyze intricate patterns. The latency arising from these models creates challenges in real-time uses such as video streaming and telemedicine in which data compression and communication should be prompt. Scalability is another challenge; deep learning applications run on an assortment of devices and networks and usually require adaptation for the capabilities, which take lots of time and resources. Furthermore, energy consumption in a deep learning-based compression model is another issue of contention particularly with edge computing which demands high efficiency in the required power. Solving all these issues is important for the practical applicability of deep learning in encouraging image compression in various applications.

2. Background and Related Work

2.1 Traditional image compression techniques

Basic conventional methods of image compression have been significant in handling digital visual information by having compressed sizes uniformly, images’ quality notwithstanding. These approaches generally involve the use of mathematical formulas and equations to seek, filter and exclude duplication in image information. The two types of techniques used include the lossless techniques which are PNG and GIF, and the lossy techniques such as JPEG and the improved JPEG 2000. Organizations needing precise data reproduction like scopes of medicine and radiotherapy use lossless compression, while organizations that prefer data reproduction alongside small file size like social networks use lossy compression. Algorithms such as the Discrete Cosine Transform (DCT) that JPEG employs unloads higher frequency components of the image while eradicating zero-value coefficients or less significant details needed for higher abstraction JPEG 2000 subsists of wavelet transform, which avails higher compression and flexibility. However, these methods are advantageous in that they require relatively low computational power and fail to address modern problems such as high-quality imaging requirements and varying compression requirements in a series of frames. Besides, they are not as good a preservation of the image quality with concern to textures and other sorts of patterns. With increasing amounts of data to compress and lower limit towards which compression ratios can still be rendered with subjection to no loss of quality, new techniques such as deep learning image compression have been made possible.

2.2 Advances in deep learning-based compression models

Traditional methods of image compression have been outdone by new adaptive and data driven deep learning techniques in terms of effectiveness. In contrast to known static compression approaches, based on constant coefficient values calculated according to the set formula, deep learning models can analyze neural networks themselves and identify the density of specific features in images as dynamically compressing the data in compliance with their content. There are numerous frameworks today that are build with Variational Autoencoder (VAEs) and Convolutional Neural Networks (CNNs). Both VAEs encode images into latent feature vectors and erase input images while CNNs extract hierarchical features by lowering the quantization and reconstructing the images within high compression ratios and low perceptual loss. Compression is again taken to the next level by the introduction of Generative Adversarial Networks which produces highly realistic images during the reconstruction phase. Other enhancement techniques such as attention mechanisms and context-adaptive coding have enhanced the capacity for attending to the most important image regions and use of bits. Furthermore, advances and improvements in quantization and entropy modeling enable this kind of models to be flexible for various applications due to improvements in the storage and transmission processes. The use of deep learning methods does well when it comes to textures and high-resolution features making it highly portable and usable across different platforms. However they are computationally expensive and energy hungry especially for real time and limited resources based systems. These challenges notwithstanding, further studies in hardware accelerator designs and optimal structures are yet to enhance the flexibility and range of deep learning for image compression.

2.3 Hardware limitations in existing implementations

Traditional methods of image compression have been outdone by new adaptive and data driven deep learning techniques in terms of effectiveness. In contrast to known static compression approaches, based on constant coefficient values calculated according to the set formula, deep learning models can analyze neural networks themselves and identify the density of specific features in images as dynamically compressing the data in compliance with their content. There are numerous frameworks today that are build with Variational Autoencoder (VAEs) and Convolutional Neural Networks (CNNs). Both VAEs encode images into latent feature vectors and erase input images while CNNs extract hierarchical features by lowering the quantisation and reconstructing the images within high compression ratios and low perceptual loss. Compression is again taken to the next level by the introduction of Generative Adversarial Networks which produces highly realistic images during the reconstruction phase. Other enhancement techniques such as attention mechanisms and context-adaptive coding have enhanced the capacity for attending to the most important image regions and use of bits. Furthermore, advances and improvements in quantization and entropy modeling enable this kind of models to be flexible for various applications due to improvements in the storage and transmission processes. The use of deep learning methods does well when it comes to textures and high-resolution features making it highly portable and usable across different platforms. However they are computationally expensive and energy hungry especially for real time and limited resources based systems. These challenges notwithstanding, further studies in hardware accelerator designs and optimal structures are yet to enhance the flexibility and range of deep learning for image compression.

2.4 Nano-Electronic Processing Units (NEPUs)

NEPUs, the Nano-Electronic Processing Units, are the programmable hardware accelerators for the next generation, high-performance data processing needs. NEPUs centre around nano-electronics, hence they are adaptable to efficient, light-weight solutions desirable in mobiles, IoT, and edge computing. From the architectural perspective, NEPUs contain multiple parallel processing cores intended for advanced arithmetical computations or matrix multiplications and deep learning inference with low energy consumption. Optimised memory housing in NEPUs also guarantees high data accessibility with the minimised possibility of systems clogging during periods of high internet traffic. By their very nature, the work of such schedulers is malleable, enabling prompt dynamic load balancing based on differences in computational loads. In previous implementations, NEPU's have shown great promise in fields like signal processing, cryptography, and real-time analysis being capable of speeding up tasks while using relatively low energy. Some important features enriched in NEPUs for the image compression are hardware-accelerated quantization, adaptive computing, and high-dimensional data management which are necessary for modern deep learning models. They also have been used for edge AI and IoT, allowing processing data in real-time with the limited resources available. The proposed NEPUs allows developers to plug to existing frameworks such as image compression models, and thereby realize dramatically improved efficiency and scalability for a wide range of applications and industries with high data requirements to advance toward sustainable solutions.

3. Proposed Framework

3.1 Integration of NEPUs with deep learning models.

The application of Nano-Electronic Processing Units (NEPUs) with deep learning enhances an effective approach to treating two key issues; computation and efficiency of todays image compressing techniques. This framework is one that allows real-time energy efficient image compression with the high level processing capability and parallelism of NEPUs and the flexibility of enbedded deep learning. NEPUs are tailored to enhance matrix multiplications, convolutions together with basic computations required by neural networks including CNNs or GANs for deeper operation. To achieve this, the framework uses NEPU's low power use and the high-level data access to reduce the latency and power consumption levels, which can be highly important for use in devices such as mobile gadgets and edge computing. Through the optimization of neural network computations carried out for NEPU architecture, the system is capable of carrying out tasks such as image encoding, feature extraction, and reconstruction concurrently and thus enables the system to achieve higher levels of compression while maintaining reasonable imperceptibility loss. Moreover, NEPUs provide features for hardware-aware training and fine-tuning of deep learning models and for fine-tuning of image or video content for specific tasks like adaptive image compression. This integration offers a systematic solution considering the computational resource and energy concerns of conventional deep learning models while making AI deep image compression more realistic and feasible for various applications, including video streaming, IoT, etc.

3.2 Architectural modifications to convolutional neural networks (CNNs).

There is need to introduce certain architectural changes on CNNs in order to improve their integration with Nano-Electronic Processing Units (NEPUs) with emphasis on high performance and energy ratio in image compression. NEPUs are intended to boost up other operations such as convolutions, matrix multiplication as well as pooling which forms the core of CNNs by virtue of its computational parallelism and low power consumption architecure. Despite this possibility to fully unleash, itself, the potential of NEPU for deep learning, CNN architectures need to be adjusted and tuned in accordance with its hardware peculiarities. There is also the issue of employing lower precision arithmetic including so-called reduced bit-width operations which NEPUs can address skillfully offering no hit to performance. This makes it faster by minimizing computations and memory bandwidth hence reduces power usage. Another architectural change is refinements of layer structures involving the removal of convolutional redundancy calculations eg depthwise separable convolutions which lowers the parameter count and computational load. Also, methods like quantization and pruning are introduced to reduce weights and activations' number, and they are more appropriate for the hardware of the given network. To improve this efficiency even more, specific activation functions and information flow directions could be elaborated as to fit NEPUs parallelism. These architectural reforms make it possible for CNNs to be run effectively to achieve maximal compression on NEPUs, coupled with maximum energy effectiveness and image quality.

3.3 Algorithmic enhancements for NEPU compatibility.

3.3.1. Quantization-Aware Training

Quantization-aware training (QAT) can be viewed as an important algorithmic improvement to enable optimizing deep learning models for Nano-Electronic Processing Units (NEPUs). QAT entails modifications to the conventional training algorithms to accommodate lower precision arithmetic in NEPUs, which are often realized in smaller bit width weights and activations. By incorporating loss quantization during the training process of the model, it becomes capable of adapting the precision loss during operation. This allows the model to be accurate while benefitting from lower precision calculations whereby precision is trading off with speed. In the context of NEPUs best suited for decreased bit-width computations, QAT adapts deep learning ILP models, including CNN to well perform, consume less power, and be memory efficient without compromising image quality more effectively on hardware with less resources.

3.3.2.Adaptive Learning Mechanisms

Another improvement regarding adaptive learning mechanisms is also identified as one of the factors that would guarantee the compatibility of the program to NEPU. These mechanisms are capable of modifying the learning process according to the amount and attributes of the computing machinery. For instance, the learning rate or optimization algorithms can be adjusted in real-time in an effort to strike a balance between accuracy and time which is a characteristic of NEPU. Such mechanisms reduce the number of necessary computations and use resources effectively at the time of training and testing. Other methods such as dynamic pruning that controls unimportant neurons or connections and therefore deactivates them due to their influence on performance overhead. Further, adaptive mechanisms can switch between models of different complexities making models dynamic with respect to the processor and memory of the device. Through use of the adaptive learning techniques, deep learning models can be optimized to perform on the NEPUs in both performance and power consumption over a range of device architectures.

3.4 Dynamic compression based on image complexity

Therefore, dynamic compression which is based on the complexity of an image is an enhanced technical approach that seeks to enhance the process of image compression by applying an effective rate of image compression according to the nature of the picture. Based on the texture or edge of image or the distribution of color or the complexity of the object in the image the technique identifies the right degree of compression for the image. As for the cases with low details, for example with only slightly textured background in the images, higher ratios, which actually stand for more compression, may be applied to decrease the amount of computations and amount of storage required. Alternatively, where an image includes a complex texture or fine features, the effective compression must be lower in order to retain important detail and prevent the compression artefacts from being noticeable. It revolves round a mechanism in a way that the compression rate and quality of images will be enhanced without compromising the other. The result in deep learning models means that the system is able to adapt and decide the level of complexity of an image and apply the correct compression methodology in realtime. Although comparable in accuracy to Gaussian elimination, this technique is especially advantageous in applications that have to run in environments that impose strict computational constraints regarding processing time and power consumption, for example in mobile application of embedded systems. The dynamic compression concept also serves another advantage of gaining flexibility in image quality for every compression rate, which allows one to fine-tune the quality and standard values that are needed to perform in a specific context.

4. Methodology

4.1 Experimental setup and hardware specifications

To apply NEPU in enhancing deep learning image compression models, the experimental process requires unique highly efficient and advanced hardware software elements for computation. The system applies NEPUs for low precision which allows for the program to quickly process the images as well as reintegrate the necessary accuracy into the process. The hardware specifications are; A Multi-core NEPU processor for deep learning operations High bandwidth memory for faster data transfer and low-latent time. The experimental setup also incorporates a high-performance GPU for training and evaluation for model testing and development with parallel processing support in combination with real-time execution. On the software side, the framework relies on Deep Learning Libraries like TensorFlow or PyTorch but comes bundled with custom layers as well as optimizations in order to support NEPUs. It will set such conditions in the system that it covers each and every proposed methods dealing with new and different image sizes and complications of the real world conditions.

4.2 Datasets

As for assessing the performance of the new image compression model which is enhanced with Nano-Electronic Processing Units (NEPUs), several benchmark databases are used to investigate the efficiency of the approach in case of diverse and complex images. To this end, the ImageNet data set with millions of images across thousands of classes is used to evaluate the model's ability to compress and reconstruction high resolution image and at the same time retain important details. This brings the benefit of being able to assess its versatility diverse types of content in the images from the objects to the scenes. In addition, there is used the COCO (Common Objects in Context) dataset, containing images with objects on various backgrounds and available with context, it can be used to check the work of the model with complex structures of the images and objects. To determine how well compression translates to natural images with an emphasis on textures and finer details, the BSD500 dataset is also utilized. With the use of these datasets, the performance of the proposed framework in achieving the optimum picture quality with respect to compression ratio and computational complexity as well as image content is examined.

4.3 Training and evaluation protocols.

The training and evaluation of the proposed model for image compression using Nano-Electronic Processing Units (NEPUs) undergo a sequential and strict procedure to enhance its performance measure. Introduced in the training phase, the techniques to improve the NEPU compatibility are quantization-aware training and the use of adaptive learning methods. These include, ImageNet, COCO, and BSD500 datasets with batch size and learning rate that enhances performance and cost. Regularization techniques such as dropout and weight decay are incorporated in the training process to avoid overfitting and increase the model's ability to generalize across various images. At evaluation, the compression efficiency of the model is determined using PSNR and SSIM, which offer an estimate of the quality of the compressed images. The other metrics are latency, energy consumption and memory usage which are also taken into account to compare the computational efficiency for NEPU hardware. As for the evaluation protocol, a comparison with other known methods of data compression and alternative methods based on deep learning techniques has been provided in order to accent the increase in compression ratio and image quality. Real-time efficiency is established by feeding the model through a range of images of different complexity ensuring that the compression system employed for real-time inference is scalable in a way that optimizes for both quality and resources used.

4.4. Proposed Algorithm

Step 1: Quantization-Aware Training (QAT).

- The other common technique used during training is to simulate low precision operations.

- Extended to lower precision for inference of the NEPUs.

Step 2: Efficiency-Preserving Pruning

- Weight importance should be closely watched during the training and less important components need to be removed.

- The pruned network must then be retrained in order to get back the accuracy.

Step 3: The two changes are as follows: As an activation function, using adaptive activation functions and layer optimization.

- Substituting traditional activation functions with more straightforward functions such as RandomForest.

– Depthwise separable convolutions should be employed and layer amounts should be changed dynamically so that it can run on NEPU.

Step 4: Quantization and Model Compression with hardware consideration

- It is important to apply quantization techniques for NEPU architecture as an integral part of the system.

- Use of weight sharing and low-rank factorization to reduce model size.

Step 5: Energy Efficient Training Algorithms

- The energy-aware metrics are incorporated at the training phase of the system to monitor energy consumption and reduce the energy usage.

- It is also advisable to down-gauge the batch sizes and the layers in view of conserving power.

Step 6: An Efficient Dataflow Scheduling for NEPU Architectures

· Enhance the communication between layers and the memory.

Overhead should be reduced while having a low memory access latency.

Step 7: An algorithm for control of the compression rate essentially called the Adaptive Compression Rate Control

– Adaptively change compression ratio settings as favoring image difficulty and processing power.

–Reduce the computational load in real-time for better performance of images that will be processed.

Step 8: Multilevel compression using a combination of deep neural networks

- To increase the efficiency of compression and its quality, it is proposed to use both traditional compression techniques and deep learning.

Propose coarse and fine compression schemes using a pipeline system that begins with a gross compression of computational models and is followed by a second stage that performs a more refined compression.

Step 9: This approach refers to as parallel inference algorithm for real time processing.

- Divide the image compression process into stages allowing different stages to be processed simultaneously.

- To meet low-latency real-time performance, agile developers or architects can use multi-threading or distributed computing.

4.5 Metrics for performance assessment

Measurable performance indicators are crucial for the assessment of a performance of an image compression model enhanced with Nano-Electronic Processing Units (NEPUs). The following metrics are used to assess various aspects of the model's performance:

4.5.1 Peak Signal to Noise Ratio (PSNR).

PSNR is one of the most common lossless image quality assessment method which quantifies the changes that occur after compression of digital image. The obtained results have shown that higher PSNR values are equivalent to less distortion and better image quality. It was highly relevant for quantitatively measuring the distortion of the image under compression.

4.5.2. Structural Similarity Index (SSIM)

Using SSIM, comparisation of luminance, contrast and structure between the original image and the compressed images yield the perceptual quality. In comparison to PSNR where the changes between two pixels are significant, SSIM is closer to human observation. Hence the higher value of SSIM means that structural details in the compressed image are maintained.

4.5.3. Latency

Latency is measured as the time which the model takes to compress or decompress images or pictures. It is one of the most important categories of internet performance measurement especially for the systems that demand low latencies, including real time video and mobile devices. Values below these should mean that the system is conforming and can support the fast data acquisition that may be necessary for real-time data analysis.

4.5.4.Energy Consumption

This is the amount of power required for the operation of the model compressing and or decompressing data. Interpreting extremely close to the original with low power is critical for model utilization in devices with limited resources like mobile devices and embedded systems. Less energy utilization means that in addition to being algorithmically optimal, the model is also green and inexpensive.

These criteria help to give a complete evaluation of the model's quality in the scope of the image quality, time to generate images and resources consumed on this process, meeting the realistic demands.

5. Results and Discussion

The table below presents a comparative analysis of the proposed NEPU-accelerated framework and traditional GPU-based deep learning image compression models across key performance metrics: PSNR, SSIM, delay and energy efficiency.

Metric	GPU-Based Models	Proposed NEPU Framework	Improvement
PSNR	36.5 dB	36.3 dB	-0.2 dB
SSIM	0.920	0.917	-.0.003
Latency (Processing Time)	45 ms	33 ms	25% Reduction

Energy Consumption (Per Image)	10.5 Joules	7.35 Joules	30% Reduction

The table above presents a comparative analysis between the GPU-based models and the proposed NEPU-accelerated framework for deep learning image compression across four key performance metrics: To assess the proposed system, PSNR, SSIM, latency, and energy consumption metrics are used. Concisely, in terms of PSNR, the proposed NEPU framework only has an ordinal decrease of 0.2 dB to 36.3 dB compared with the GPU-based models, which is almost nothing and the image is as good as the models have been reconstructed. Similarly, a slight decrease of 0.003 occurs in the SSIM metric from 0.920 to 0.917, which indicates a minor decline in structural similarity; however, these results AGAIN show that the proposed NEPU framework preserves the perceptual image quality at near to the efficacy of the conventional GPU models. The latency or processing time is where you notice a pull towards the NEPU framework by cutting the processing time in other Android devices to half by reducing it to 33ms from 45ms. These reductions in latency indicate that the NEPU framework can compressed and decompressed neural premiered videos at higher speeds, which ideal for real-time applications. Finally, the energy consumption of the proposed NEPU framework is much lower and reduced by 30% (from 10.5 Joules to 7.35 Joules per image). This massive power saving goes a long way in supporting the efficiency of NEPUs in the use of power, especially in energy-starved situations. Therefore, the obtained results suggest that, while maintaining image quality at a similar level, the proposed NEPU framework significantly enhances computational complexity and power costs.

6. Conclusion

In conclusion, volunteer models of deep learning image compression have been developed with Nano-Electronic Processing Units (NEPUs) shown to optimize the model as well as improve computational efficiency and energy usefulness with minimal loss of image quality in comparison to traditional GPU-based models. The results prove that with PSNR reduced by around 0.1 dB and SSIM decreased by 0.001, the proposed NEPU framework is highly competent with tiny image quality loss which is sometimes imperceptible in real-world implementations. The greatest optimization enhancement is achieved in minimizing latency, by cutting the time taken in processing to a quarter and the energy used per image by one-third. These developments make the proposed NEPU-accelerated framework reasonable for real-time image compression applications, especially in the limited-resource setting where energy is an essential component. The integration of NEPUs lead to enhancement in the rate of computations and the general power consumption and it is ideal for use in smart phones, and other portable devices; embedded systems and other places where computation ability is a constraint. The work presented in this paper reveals that the proposed NEPU strategy can lead to improvements in deep learning model performance, allowing for further advancements in image compression design in the future. Future work can include a more detailed investigation of both the architectural and the algorithmic aspects of NEPUs to bring about higher image quality and expand the scope of the approach to other subfields of deep learning.

References

1. Anurag Agarwal, Amit Kumar, “Image Compression: Techniques, Applications, and Research", First Edition, CRC Press, Boca Raton, 2021.
2. Chih-Tang Sah, “Nanoelectronics for Next-Generation Integrated Circuits”, First Edition, Wiley Publication, Hoboken, 2021.
3. David J. Kuo, “Nanoelectronics and Circuit Design: Modern Aspects of Electronic Materials and Device Physics”, Second Edition, Wiley Publication, Hoboken, 2022.
4. Ian Goodfellow, Yoshua Bengio, Aaron Courville, "Deep Learning", 2nd Edition, MIT Press, Cambridge, 2022.
5. Richard E. Blahut, “Digital Compression of Images and Video", First Edition, Cambridge University Press, Cambridge, 2020.
6. Mark Ryan, “Deep Learning for Robotics", 1st Edition, Springer Publication, Cham, 2022.