

Image Compression using Neural Networks

Usman Kaiser Moghal

Computer Sciences

FAST University

Islamabad, Pakistan

usmank.moghal@gmail.com

Abstract—Image compression plays a major role in helping to manage data in this age of digital media. As media platforms grow and more powerful hardware gets cheaper, image data on the internet will continue to grow exponentially. As a means of curbing this growth, image compression plays a heavy hand, by decreasing the file sizes and increasing transfer speeds.

The use of neural networks as approximate functions to help with image compression is a growing field, one where the generic nature of neural networks can be exploited to achieve better gains in compression as compared to older methods.

In this research study, we look into the more recent innovations in the field of image compression via neural networks, and exactly how these new methods can be used practically in the real world. After critical analysis we individually compare the various methods in terms of model type, real world implementation, and expected results.

Index Terms—Image Compression, Neural Networks, Comparison

I. INTRODUCTION

A. Image Compression

Image compression is a function in image processing, that is used for data compression, with the intent to reduce the cost for storage or transmission of the digital image. Several methods exist to achieve this effect, however superior results can be achieved by certain algorithms which take advantage of visual perception and the statistical properties of the image.

There are two types of image compression, lossy or lossless. In lossless compression, data is compressed in a fashion that prevents data loss, and is preferred for archival, medical imaging, clip art, technical drawings, or even comics. Lossy compression on the other hand aims to compress the image to greater degrees, even at the expense of some data loss, and even introduces compression artifacts especially at lower bitrates. Due to the nature of the method, Lossy compression is preferred for natural images such as photographs, or in applications where minor (sometimes imperceptible) loss of features is acceptable, so long as it can achieve a substantial reduction in file size. Such Lossy compression techniques that cause negligible degradation in quality may be called visually lossless.

Methods for lossy compression:

- Transform coding
 - Discrete Cosine Transform (DCT)
 - Wavelet Cosine transform (WCT)
- Color space reduction
- Chroma sub-sampling

- Fractal compression.

Methods for lossless compression:

- Run-length encoding
- Area image compression
- Predictive coding
- Entropy encoding
- Adaptive dictionary algorithms
- DEFLATE
- Chain codes

B. Neural Networks

A neural network is a network of neurons composed of artificial nodes. Hence a neural network, from the perspective of computer science, is an artificial neural network, for solving artificial intelligence (AI) problems. The connections of the biological neuron are modeled as weights, either positive or negative. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs to a neural network are altered by a weight and summed, an activity referred to as a linear combination. Finally, the amplitude of the output is controlled by an activation function, eg. a range of output between 0 and 1, or it could be negative 1 and positive 1.

There are many uses for this artificial network, such as predictive modeling, adaptive control and applications where they can be trained via a dataset. Networks can enable self-learning as a result of experience, which can derive conclusions from a complex and seemingly unrelated set of information.

Neural networks are essentially universal approximators, which work best assuming the intended environment the system will be run on has a high tolerance to error. Neural networks work well under the following cases:

- to capture associations or uncover regularities within a set of patterns
- where either the volume, number of variables or diversity of the data is too great
- the relationships between different variables are unclear
- conventional approaches cannot adequately describe the relationships

Some types of neural networks are listed below:

- Feedforward Neural Network
- Convolutional Neural Network
- Multilayer Perceptron
- Recurrent Neural Network(RNN)

- Radial Basis Function Neural Network
- Sequence-To-Sequence Models
- Modular Neural Network

II. CRITICAL REVIEW OF RELATED LITERATURE

Over a 100 publications were combed through, before 10 papers were selected, such that these were not just strongly relevant to the topic, but were all published recently by reputable institutes, and have already been cited by others. Below is a critical review of the respective 10 papers.

A. *Learned Neural Iterative Decoding for Lossy Image Compression Systems [1]*

The authors of [1] propose a lossy compression algorithm, for improving the reconstruction process of the decoder, and for it to work with any encoder.

The model was tested on several datasets, and evaluated using Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), Multi-Scale Structural Similarity (MS-SSIM), and MSE. The results show that the proposed method yields marginally better PSNR and SSIM, while being simpler and faster than other complex end-to-end models.

Strengths

- Lower distortion
- Improved decoding accuracy
- Functions with any encoder
- Can handle wide range of bitrates

Weaknesses

- Improved accuracy only for specific bitrates

B. *Learning-Based Image Compression using Convolutional Autoencoder and Wavelet Decomposition[2]*

The authors of [2] propose a convolutional auto-encoder architecture which employs wavelet decomposition of all input channels (WCAE), in an attempt to showcase the emphasis of utilizing a wavelet decomposition as a pre-processing step to improve learning based compression.

The experimentation shows that the proposed method outperforms both the standard JPEG and the no wavelet architecture. WCAE resulted in fewer low-frequency errors as compared to JPEG, however as compared to JPEG2000, showed greater high frequency errors.

Strengths

- Able to tweak the image characteristics based on frequency
- Custom tweaking can allow better subjective quality

Weaknesses

- Unable to outperform JPEG2000
- Requires custom tweaking

C. *Context-adaptive neural network based prediction for image compression[3]*

The authors of [3] proposes an intra image prediction architecture called Prediction Neural Networks Set (PNNS) and showcases the benefits of using neural networks for intra image prediction.

The use of neural networks for intra prediction is justified through this paper, by demonstrating performance gains over state of the art methods, even though the proposed method takes longer and has a higher memory requirement. **Strengths**

- Better performance than state-of-the-art methods
- Robust against noise

Weaknesses

- Works on gray-scale only
- Limited context sizes possible
- Greater memory consumption
- Slower performance

D. *UltraCompression: Framework for High Density Compression of Ultrasound Volumes using Physics Modeling Deep Neural Networks[4]*

The authors of [4] propose a framework for ultrasound image compression, with the ability to retain realistic speckle appearance despite high compression factors. Their goal is to produce a framework capable of maintaining key information rendered through speckles and for the framework to become a guide for future speckle based compression.

The experimentation showed that the proposed model achieved a very high compression ratio with the decompressed images having a low inter-tissue divergence, indicating a very close proximity to the original image. The proposed method outperforms standard methods by far. **Strengths**

- Computationally efficient
- High compression ratio
- Accurate results
- Can work with other segmentation and generation methods

Weaknesses

- Requires powerful/expensive hardware
- Requires GPU

E. *Learning a virtual codec based on deep convolutional neural network to compress image[5]*

The authors of [5] propose a new auto-encoder framework for mixed-resolution image compression involving image representation network (IRN) and post-processing neural network (PNN), which are helped by a virtual codec network (VCN). The goal is to develop a new framework that can optimally learn non-differentiable components via a virtual differentiable networks.

The proposed method is compared with JPEG and JPEG2000, as well as other methods involving JPEG compression and artifact removal, and evaluated over SSIM, sharpness and PSNR. Results show that with enough iterations, the proposed method has greater gains over other methods in both SSIMs and PSNR. The proposed solution, at least on paper, outperforms the major state-of-the-art methods. **Strengths**

- Better results

Weaknesses

- Difficult to learn full framework
- Long learning time
- Computationally expensive

F. Learning a Deep Vector Quantization Network for Image Compression[6]

The authors of [6] propose a Deep Convolutional Neural Network (DCNN) framework of image compression, which consists of an encoder, quantizer, and a decoder, which are jointly optimized via a fully convolutional vector quantization network (VQNet). The authors intend to provide a compression methods that would outperform the standard techniques.

The experimentation show that the proposed method outperforms all standard methods, and based on the type of image input, the custom optimization further increases performance. However this performance gain comes at a cost of expensive hardware, and slow processing speeds. **Strengths**

- Better performance than state-of-the-art methods
- Simple implementation

Weaknesses

- Requires powerful hardware
- Requires GPU
- Expensive
- Slow

G. Semantic Perceptual Image Compression with a Laplacian Pyramid of Convolutional Networks[7]

The authors of [7] propose a novel image compression framework, involving the construction of a multi-scale image representation through use of a Laplacian pyramid, with the aim of providing an alternative tool that would allow high compression with good visual quality.

According to the metrics defined, the proposed method performed poorly, despite being visually better, especially at lower bitrates. As a result, a mean opinion score (MOS) test was performed, that resulted in a higher score for the proposed method as compared to JPEG and JPEG2000. **Strengths**

- Better visual quality
- Simple implementation
- Strong foundation

Weaknesses

- Bad metric score
- Bad visual quality on high bitrates

H. Spatially adaptive image compression using a tiled deep network[8]

The authors of [8] propose a tiled network approach to image compression which combines DNN'S and bit-rate adaptors, with the intention of fruitfully extending the work of another method proposed last year.

The results of the experimentation show that the proposed model outperformed the standard JPEG and other method, at a specific bitrate, while solving issues during the training of the other method. **Strengths**

- Better performance as compared to JPEG

Weaknesses

- Specific to certain bitrates

I. Efficient deep neural network for digital image compression employing rectified linear neurons[9]

The authors of [9] propose a DNN framework for image compression with the aim of reduced training time despite better generalization property, such that it could support real time applications.

The results of the experiments showed that the proposed network is trained faster, and achieves comparable results to its more complex sister network, and by further tweaking the hidden layer, achieved higher compression ratio than the sister network. **Strengths**

- Simple implementation
- Fast training
- Good compression results

Weaknesses

- Limited to specific resolution
- Limited to specific bitrate
- Limited to gray-scale images

J. Image compression using neural networks and haar wavelet[10]

The authors of [10] propose a neural network that can find the optimum ratio between image intensity and compression ratios in a Haar Wavelet transform based compression model.

The experiment results showed that calculating the optimum compression ratio helped achieved better results in a smaller time period, and that the whole process could be adopted in real-time application after significant training for the particular environment. **Strengths**

- Simple implementation
- Fast results
- Applicable in other areas

Weaknesses

- Limited to certain image resolutions
- Large variance in input images will degrade performance
- Single outlier can disrupt performance
- Limited to gray-scale images

III. COMPARATIVE STUDY OF RELATED LITERATURE

After critically analysing the research papers, certain key notes come to light when comparing the different works to each other. For simplicity the key points are listed below tabularly.

The table I categorises the studied papers into their "Focus Area" and details its "Network Type" as well as the novel "Idea" presented in the individual papers.

The table II and III shows the crucial details of the studied papers required when attempting to replicate the authors proposed solutions. This means the implementors need to know exactly what type of real-world data the implementing system will be able to handle as well as the proposed speed, complexity and hardware requirements of the system. ("✓" in "Image Support" indicates whether the proposed system will support an input of that class. Values in "Implementation" are numbered 1-5, with 1 being low and 5 being high, eg.

TABLE I
BREAKDOWN OF MODELS STUDIED

Paper	Breakdown		
	Focus Area	Network Type	Idea
[1]	Decoder	RNN	Global context per patch
[2]	Encoder	CAE	Apply wavelet transform on 3 classes of frequency
[3]	Encoder	PNNs	Spacial context above and on left
[4]	Encoder	CNN	Segmentation based encoding
[5]	Training	IRN+PNN+VCN	Use of virtual network for adjusting parameters
[6]	Training	DCNN+VQNet	Use of virtual network for adjusting parameters
[7]	Encoder+Decoder	DNN	Using an altered mathematical model
[8]	Encoder	CNN	Tiled Prediction
[9]	Encoder+Decoder	DNN	Alter layer and neuron size based on input size
[10]	Training	ANN	Find optimum compression ratio

TABLE II
MODEL REQUIREMENTS

Paper	Image Support			
	RGB	Super Resolution	Variable Resolutions	Variable Bitrates
[1]			✓	✓
[2]	✓			✓
[3]				✓
[4]			✓	
[5]		✓	✓	✓
[6]	✓	✓	✓	✓
[7]	✓	✓	✓	✓
[8]	✓			✓
[9]			✓	✓
[10]				

^aAll proposed methods support gray-scale images

TABLE III
SOLUTION IMPLEMENTATION FEASIBILITY

Paper	Implementation		
	Training Speed	Complexity	Hardware Requirement
[1]	2	2	2
[2]	3	1	2
[3]	5	2	3
[4]	4	1	5
[5]	2	3	2
[6]	5	1	5
[7]	4	1	2
[8]	4	2	2
[9]	5	1	2
[10]	5	1	2

^aScaled 1-5, 1 being lowest, 5 being highest

"Training Speed" of 2 vs 5 implies that the first one will train slower than the second one, "Complexity" of 2 vs 5 means the first one is easier to implement in code, "hardware Requirement" of 2 vs 5 means the first one is low powered and hence cheaper than the second one)

The table IV and V shows the results to expect if the proposed system is implemented, and on what basis were the results verified. ("✓" in "Evaluation Metric" means the metrics were used when testing the system, while in "Result" it means that the proposed system outperformed the respective model, eg. "✓" in "WebP" means that the proposed system outperformed "WebP" model.)

TABLE IV
EVALUATION METRICS USED FOR VALIDATION

Paper	Metrics				
	PSNR	SSIM	MS-SSIM	Compression Ratio	Other
[1]	✓	✓	✓		
[2]	✓		✓		
[3]	✓				
[4]				✓	✓
[5]	✓	✓			✓
[6]	✓		✓		
[7]	✓	✓	✓	✓	✓
[8]	✓				
[9]	✓				
[10]				✓	✓

TABLE V
COMPARISON AGAINST STANDARD MODELS

Paper	Models					
	JPEG	JPEG2000	BPG	WebP	HEVC	Other
[1]	✓	✓				✓
[2]	✓					
[3]					✓	
[4]	✓	✓	✓	✓		
[5]	✓	✓				✓
[6]	✓	✓	✓			✓
[7]	✓					✓
[8]	✓					
[9]	✓					✓
[10]						✓

IV. CONCLUSIONS AND FUTURE WORK

From this brief study it is clear that the importance of image compression will only grow in years to come as data continues to accumulate. New methods need to be found to curb this huge growth. Neural networks is a valid approach. The field itself is growing, and new ideas continue to fester on how to exploit the unique characteristics of neural networks for the sake of improving compression techniques.

In this paper 10 different papers were looked at, all of which had a unique idea or a unique implementation that improved the compression of images as compared to standards like JPEG. While some were simple changes, others were revolutionary ideas, all of which targeted different areas of the image compression model such that even state-of-the-art techniques like WebP and HEVC were changeable.

However, despite such optimistic results, it is clear that the proposed solutions are not yet ready for general use in the real world, or even to become a standard themselves such as WebP and HEVC. They need further development.

Regarding work extension, the obvious comes to mind. To find a way to refine the proposed solutions, such that they require less time to train, and can function on relatively cheaper hardware. Though it may seem a simple solution, it is one of the more difficult options available, as such refinements in the proposed models would involve mathematical and coding intricacies.

Another idea for future work would be to combine the respective proposed solutions. The 10 papers discussed all focus on varying areas of the compression model. If we can combine the methods, then a new method would emerge such that it would have greater accuracy, lower distortion, and decreased hardware constraints.

REFERENCES

- [1] Alexander G Ororbia, Ankur Mali, Jian Wu, et al. "Learned Neural Iterative Decoding for Lossy Image Compression Systems". In: *2019 Data Compression Conference (DCC)*. IEEE. 2019, pp. 3–12.
- [2] Pinar Akyazi and Touradj Ebrahimi. "Learning-Based Image Compression using Convolutional Autoencoder and Wavelet Decomposition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0.
- [3] Thierry Dumas, Aline Roumy, and Christine Guillemot. "Context-adaptive neural network based prediction for image compression". In: *IEEE Transactions on Image Processing* (2019).
- [4] Debarghya China, Francis Tom, Sumanth Nandamuri, et al. "UltraCompression: Framework for High Density Compression of Ultrasound Volumes using Physics Modeling Deep Neural Networks". In: *arXiv preprint arXiv:1901.05880* (2019).
- [5] Lijun Zhao, Huihui Bai, Anhong Wang, et al. "Learning a virtual codec based on deep convolutional neural network to compress image". In: *Journal of Visual Communication and Image Representation* 63 (2019), p. 102589.
- [6] Xiaotong Lu, Heng Wang, Weisheng Dong, et al. "Learning a Deep Vector Quantization Network for Image Compression". In: *IEEE Access* 7 (2019), pp. 118815–118825.
- [7] Juan Wang, Xiaoming Tao, Mai Xu, et al. "Semantic Perceptual Image Compression with a Laplacian Pyramid of Convolutional Networks". In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 699–703.
- [8] David Minnen, George Toderici, Michele Covell, et al. "Spatially adaptive image compression using a tiled deep network". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 2796–2800.
- [9] Farhan Hussain and Jechang Jeong. "Efficient deep neural network for digital image compression employing rectified linear neurons". In: *Journal of Sensors* 2016 (2016).
- [10] Adnan Khashman and Kamil Dimililer. "Image compression using neural networks and haar wavelet". In: *WSEAS Transactions on Signal Processing* 4.5 (2008), pp. 330–339.