International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307 (Online), Volume-13 Issue-3, July 2023

Implications of Deep Compression with Complex Neural Networks

Lily Young, James Richardson York, Byeong Kil Lee

Abstract*: Deep learning and neural networks have become increasingly popular in the area of artificial intelligence. These models can solve complex problems, such as image recognition or language processing. However, the memory utilisation and power consumption of these networks can be substantial for many applications. This has led to research into techniques to compress the size of these models while retaining accuracy and performance. One of the compression techniques is the deep compression three-stage pipeline, which includes pruning, trained quantisation, and Huffman coding. In this paper, we apply the principles of deep compression to multiple complex networks to compare the effectiveness of deep compression in terms of compression ratio and the quality of the compressed network. While the deep compression pipeline is effectively working for CNN and RNN models to reduce network size with minimal performance degradation, it is not working correctly for more complex networks, such as GAN. In our GAN experiments, performance degradation is excessive due to compression. For complex neural networks, careful analysis is necessary to identify which parameters enable a GAN to be compressed without compromising output quality.

***Keywords*: Neural Network, Network Compression, Pruning, Quantization, CNN, RNN, GAN.**

I. INTRODUCTION

In recent years, deep learning and neural networks have gained significant popularity in the field of artificial intelligence. These models can solve complex problems, such as image recognition or language processing. However, the memory utilization and power usage of these networks can be prohibitively extensive for many applications. This has led to research into techniques to compress the size of these models while retaining accuracy and performance [1][2][3]. While many pruning and compression techniques have been developed, they often require specialised tools to achieve their full effect. One of these techniques is the deep compression three-stage pipeline, including pruning, trained quantization, and Huffman coding, which can be implemented through the use of widely available tools [1].

Manuscript received on 12 May 2023 | Revised Manuscript received on 22 May 2023 | Manuscript Accepted on 15 July 2023 | Manuscript published on 30 July 2023.

*Correspondence Author(s)
Lily Young, Department of Electrical and Computer Engineering, University of Colorado Colorado Springs, 1420 Austin Bluffs Parkway, Colorado Springs, CO 80918, USA. Email: ayoung4@uccs.edu
James Richrdson York, Department of Electrical and Computer Engineering, University of Colorado Colorado Springs, 1420 Austin Bluffs Parkway, Colorado Springs, CO 80918, USA. Email: jyork2@uccs.edu
Byeong Kil Lee*, Department of Electrical and Computer Engineering, University of Colorado Colorado Springs, 1420 Austin Bluffs Parkway, Colorado Springs, CO 80918, USA. Email: blee@uccs.edu, ORCID ID: https://orcid.org/0000-0002-0260-2238.

In this paper, we apply the principles of deep compression to multiple complex networks using Keras with Tensorflow 2 to make the models more suitable for deployment on embedded devices and other devices with limited resources. We utilise deep compression for three complex neural networks: CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), and GAN (Generative Adversarial Network). Based on the experimental results, we compare the effectiveness of deep compression in terms of compression ratio and the quality of the compressed network. While the deep compression pipeline is effectively working for CNN and RNN models to reduce network size with minimal performance degradation, it is not working correctly for more complex networks, such as GAN. In our GAN experiments, performance degradation is excessive due to compression. For complex neural networks, we need to develop various compression methodologies.

The rest of the paper is organized as follows. In Section II, we describe related work. The deep compression pipeline and implementation for this research are presented in Section III. In Section IV, we describe the modelling of complex neural networks. Experimental results are presented in Section V, and we conclude with Section VI. Section VII includes future work.

II. DEEP COMPRESSION PIPELINE

A deep compression pipeline is a powerful tool for reducing the size of large deep learning models while retaining their accuracy. It consists of three steps: pruning, trained quantization, and Huffman coding as shown in Figure 1 below.

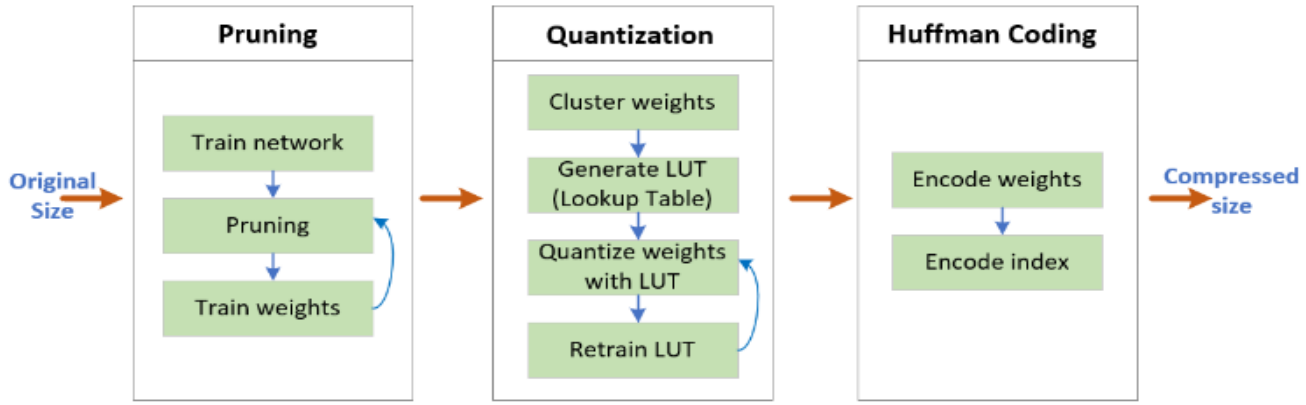


Fig. 1. Deep Compression Pipeline Steps [1]

A. Implementation of Pruning

Pruning is the first step in the deep compression pipeline, involving the removal of redundant or unnecessary parameters from the model to reduce its size without compromising any essential nodes or weights. In general, pruning can be performed through weight pruning, which removes connection weights between nodes, or through neuron pruning, which eliminates entire nodes from layers of the network. When pruning a network, weights near zero are selected for removal because they have a minimal effect on the network's output. Once weights or neurons are pruned, the network can be retrained to compensate for the changes caused by removing the selected weights. Keras contains a compression interface that allows trained pruning on most layer types. The code for pruning used in our experiment is shown in Figure 2 [5].

```
pruning_params = {'pruning_schedule':
    tfmot.sparsity.keras.ConstantSparsity (.5, 0, end_step=-1,
    frequency=1)}

def apply_pruning(layer):
    if isinstance (layer, tf.keras.layers.Dense):
        return prune_low_magnitude (layer, **pruning_params)
    if isinstance (layer, tf.keras.layers.Conv2DTranspose):
        return prune_low_magnitude (layer, **pruning_params)
return layer

model_pruned = tf.keras.models.clone_model (model,
    clone_function=apply_pruning)
train_model (model_pruned)
model_pruned = tfmot.sparsity.keras.strip_pruning(model_pruned)
```

Fig. 2. Keras code for pruning

B. Implementation of Trained Quantization

The second step in the compression pipeline is called trained quantization [6]. In this step, n centroids are uniformly selected between the minimum and maximum network weights, and a lookup table is created for these values. For each weight, if the nearest value entry in the lookup table is found, the weight is replaced by a reference to the table. Because the quantization lookup table space is much smaller than the real numbers, the bit width of each weight not including lookup table overhead can be reduced to ceil (logbase(2, num_centroids)) bits. Once the initial quantization step is complete, the centroids can then be trained in the same way as normal weight values. An example of quantized weights is shown in Table 1.

C. Implementation of Huffman Coding

The final step in the deep compression pipeline is Huffman coding. In this step, a model is losslessly compressed using variable-length codes and a prefix-free binary tree. When the frequency of each byte in the saved model is determined, the high-frequency values are then converted into shorter values that can be referenced back to a code tree to reconstruct the original model. Huffman coding is used in many standard file compression algorithms, such as deflate.

Keras also includes an interface that allows a model to be quantised and trained. The code for quantization in Keras is included in Figure 3.

Table I. Quantization Example

Weights	32-bit floating-point value	Quantized (2 bits)
weight 1	1.127	0
weight 2	5.133	1
weight 3	2.769	0
weight 4	6.953	2
weight 5	13.444	3
…	…	…
weight N	12.763	3

Lookup Table	Centroid value (32-bit floating-point number)
0	1.872
1	5.102
2	7.003
3	13.120

```
def apply_quant (layer):
    if isinstance (layer, tf.keras.layers.Dense):
        return quantize_l (layer)
    if isinstance (layer, tf.keras.layers.Conv2D):
        return quantize_l (layer)
return layer

model_quantized = tf.keras.models.clone_model(model_pruned,
        clone_function=apply_quant,)
tfmot.quantization.keras.quantize_apply(generator_quantized)
train_model(model_quantized)
```

Fig. 3. Keras code for trained quantization

When evaluating the effectiveness of pruning and quantization in Tensorflow 2, compressing the model using Huffman coding is necessary. This is because TensorFlow

Retrieval Number: 100.1/ijsce.C36130713323
DOI: 10.35940/ijsce.C3613.0713323
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

saves all parameters of a model, including those which are zero due to pruning. To evaluate each model, the gzip compression algorithm was used in our experiment, and the size of the compressed model was used as a measure of evaluation. The code for compressing a model is shown in Figure 4.

```
def get_gzipped_model_size (file):
    # Returns size of gzipped model, in bytes.
    import os
    import zipfile

    _, zipped_file = tempfile.mkstemp ('.zip')
        with zipfile.ZipFile (zipped_file, 'w',
        compression=zipfile.ZIP_DEFLATED) as f:
    f.write (file)

return os.path.getsize (zipped_file)

print (get_gzipped_model_size ("./saved/unpruned.h5"))
print (get_gzipped_model_size ("./saved/pruned.h5"))
```

Fig. 4. Model compression using gzip

III. MODELING OF COMPLEX NETWORKS

A. Convolutional Neural Network

Convolutional Neural Networks are designed to utilise a stack of convolutional filters to extract features from images. For the experiment of deep compression, we choose a non-sequential CNN which uses residual blocks [4]. These blocks combine the input data with the convolved output of a convolutional layer, essentially passing the input forward to the next layer. The structure of a residual layer is shown in Figure 5. This enables the network to be trained more efficiently, resulting in lower training errors for dense networks. The reason we chose this type of network is that it achieves better performance than a standard convolutional neural network. It is also a good test for the compression pipeline because of its increased complexity.

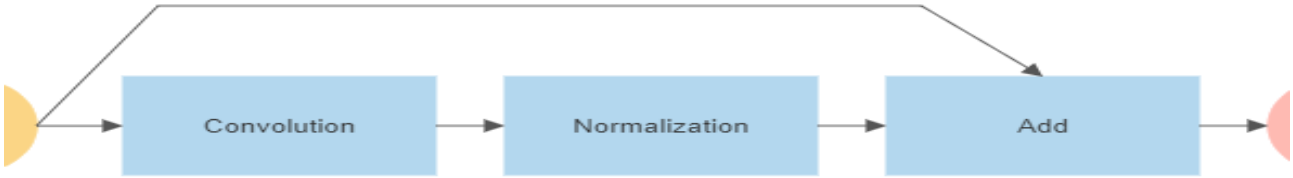


Fig. 5. Residual layer structure

The final network used is a 12-layer network with 15,634,994 parameters. It accepts RGB colour, 227x227 pixel images. These images are first passed through 2 standard convolutional layers. The outputs are then passed through 5 residual convolutional layers. The output from these layers is finally passed through 4 dense layers that classify the images into 10 different categories of the CIFAR-10 image data.

B. Recurrent Neural Network

We applied deep compression to investigate its effectiveness in Recurrent Neural Networks (RNNs). RNNs are a type of neural network that is particularly well-suited for processing sequential data, making them useful in a wide range of applications such as natural language processing, speech recognition, machine translation, and image captioning. By compressing the size of RNNs, we hoped to improve their portability and performance. To evaluate the deep compression pipeline for the RNN, we utilise a Long Short-Term Memory (LSTM) model to perform sentiment analysis on a dataset of movie reviews. LSTM networks are a type of RNN that is well-suited for this task [7]. Due to the memory cells in LSTM models, they can better analyse the overall sentiment of a longer piece of text, as the sentiment of a text may not be immediately apparent from individual words or short phrases. Additionally, LSTM networks can effectively handle input data of varying lengths, unlike RNN networks, which are limited to fixed-sized inputs. LSTMs have more logic to remember or forget some information, which means LSTMs have more complexity. The movie review dataset we used is a built-in Keras dataset of 50,000 movie reviews, classified as either positive or negative sentiment, sourced from the IMDb website. The words from the review were filtered to include the first 20,000 most frequent words, while eliminating the 10 most frequent words. The reviews were then padded to a maximum length of 500 words, meaning that any review shorter than 500 words was padded with empty values, while any review longer than 500 words was truncated. The RNN platform has three layers: Embedding, LSTM, and a Dense layer. In total, there were 2,690,433 parameters. The embedding layer hosts 2,560,000 parameters, the LSTM layer holds 131,584 parameters, and the dense layer holds 129 parameters. The embedding layer is massive due to the 20,000-long library, each with 128 nodes.

C. Generative Adversarial Network

The third type of network evaluated is a Generative Adversarial Network (GAN) [8]. This model is designed for generating images that are similar to an input dataset. To do this, it utilizes two separate models. The first of these models is called the discriminator. This model is designed to determine if an input image is real or fake. The second model is called the generator. The generator is creating images that are indistinguishable from real photos. An example of the GAN architecture is shown in Figure 6.

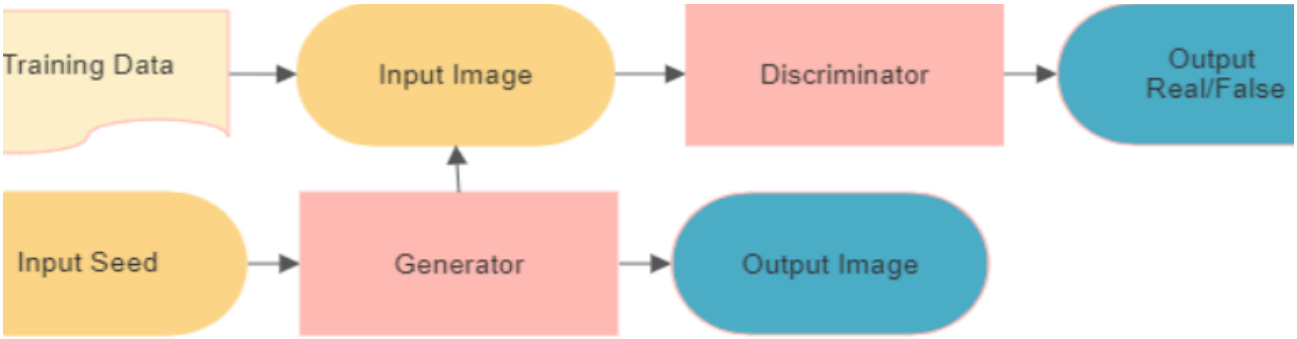


Fig. 6. Generative Adversarial Network Architecture

This type of network is particularly complex to train because it requires two passes through the data. In the first pass, the discriminator network is trained to distinguish between real and fake images. In the second pass, the generator is trained to create images that the discriminator believes are real. Because two networks are trained against each other, the loss for each can change dramatically during each training step. This can make it challenging to determine network convergence. Once the network is trained, the discriminator can be discarded, and just the generator is used. This network is chosen for evaluation of the compression pipeline due to its training behaviour. Since it trains two networks simultaneously, we expect the result of trained pruning and quantisation to be worse than expected. Due to the complexity of the desired output, we observe that small changes caused by pruning might result in an unrecognisable output. Unfortunately, while this network type offers good performance due to its adversarial structure, it is very challenging to evaluate quantitatively because of its complex structure. The GAN we applied is a convolutional generative adversarial neural network. For the generator, a 6-layer network was used. The input layer accepted 100 randomly generated values. It then passes the values through a dense layer of size 4x4x256. This initial image is then upscaled through 4 sets of *conv2d* transpose layers to create the final 32x32x3 image. For the discriminator, another 6-layer network is used. The input to this network is a 32x32x3 colour image. This is then passed through 4 layers of convolutional filters. Finally, a single dense layer is used to output whether the input was generated or real. The initial parameter count for this network is 1,988,612.

IV. EXPERIMENTAL RESULTS

We investigate the effectiveness of deep compression, including the impact on network size and its subsequent performance.

A. Convolutional Neural Network

The CNN model can be compressed successfully over multiple stages of deep compression. Table 2 displays the network size at each step, along with the total compression achieved up to that point and the compression resulting from that specific step. Quantisation exhibits the most excellent compression ratio, followed by pruning, and then gzip Huffman coding. The network started at 125.25 MB and is reduced to 62.66 MB after pruning. Quantisation was then applied, further reducing the network size to 25.68 MB. Finally, Huffman coding is used to compress the network to 13.05 MB, achieving a total compression of 9.6 times, which saves 112.20 MB of memory, as shown in Figure 7. This significant size reduction allows the network model to be more easily deployed on resource-constrained devices, improving its usability and accessibility. The compression of the network had a minimal impact on accuracy. The accuracy drop of the network after all compression steps is 2.6%, which is not negligible but a relatively small amount.

Table II. CNN Compression Results

Compression Stage	Size (MB)	Stage Compression	Total Compression (accumulated)
Original	125.252	1.000	1.000
Pruning	62.665	1.998	1.998
Quantization	15.678	3.996	7.988
Huffman coding	13.047	1.201	9.599

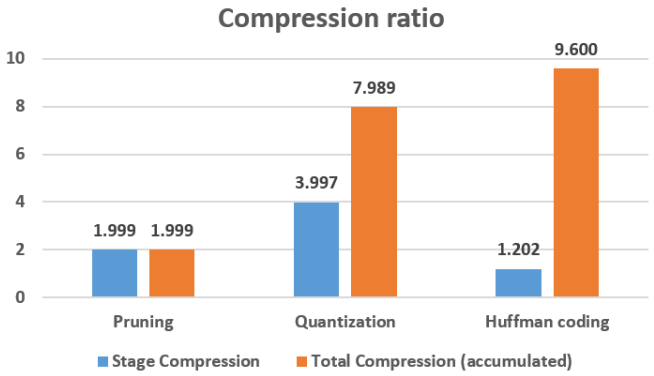


Fig. 7. Comparison Ratio for CNN

B. Recurrent Neural Network

The LSTM model was compressed very well by the deep compression pipeline. In total, the deep compression framework compressed the LSTM network by a factor of 21.69. This is mainly due to Quantisation, which decreased the size by the most significant amount of all stages. As shown in Table 3, Pruning compressed the file from 32.32 MB to 10.78 MB (2.99x compression). Quantisation compressed the file further to 2.7 MB, achieving a 3.98x compression ratio from the Pruning step. Finally, Huffman coding compressed it to the final 1.49 MB, achieving a 1.81x compression from the Quantisation step, and resulting in a final 21.69x compression rate, as shown in Figure 8. Additionally, the compression affected the accuracy, decreasing from 87.49% to 85.38%, resulting in a 2.11% loss in accuracy. Similarly to the CNN network, this is not

Retrieval Number: 100.1/ijsce.C36130713323
DOI: 10.35940/ijsce.C3613.0713323
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

negligible but is a small amount that can be ignored for most applications.

Table III. RNN Compression Results

Compression Stage	Size (MB)	Stage Compression	Total Compression (accumulated)
Original	32.319	1.000	1.000
Pruning	10.781	2.997	2.997
Quantization	2.704	3.985	11.948
Huffman coding	1.490	1.815	21.690

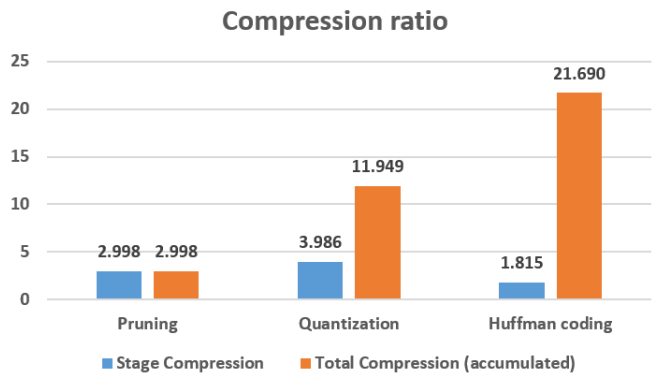


Fig. 8. Comparison Ratio for RNN

C. Generative Adversarial Network

Since the GAN network generates new images from random values, it is challenging to perform quantitative analysis on these images. To analyse the network's performance, five images were selected from the best outputs. The GAN was initially trained on the CIFAR-10. This dataset consists of colour images of animals, cars, and planes. Some example images are shown in Figure 9.

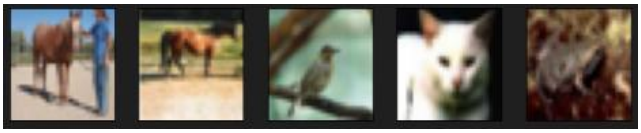


Fig. 9. CIFAR-10 dataset sample

When trained on this dataset, the GAN can create somewhat similar images. Note that, due to the way GAN functions, the generated images will not be of real animals or objects. The generated images feature shapes that resemble animals and vehicles. The photos also had colourations that were close to the input dataset and were detailed. A set of the five best-generated images is shown in Figure 10.

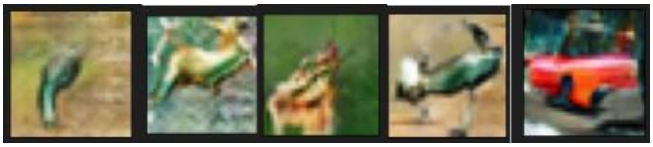


Fig. 10. GAN-generated images

The network's performance diminished significantly after pruning. Because a GAN is composed of two networks, and only the generator is used after training. Only the generator is pruned in our experiment. After pruning to a 0.6 sparsity, the output images show significantly lower quality. Even in the best output cases, the objects in the picture are less detailed and even look blocky. The output image colour is also degraded, and most images contain more blue and green than expected. Examples of these images are shown in Figure 11.

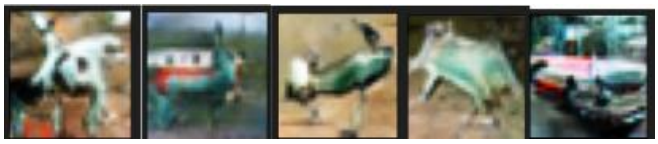


Fig. 11. Pruned GAN-generated images

The model was also negatively affected by quantization. In this step, only the generator was quantized for the same reason as in the pruning step. After quantization, the images lost even more detail and color depth. Many of the photos after quantization also became extremely noisy. Examples of the generated images are shown in Figure 12.

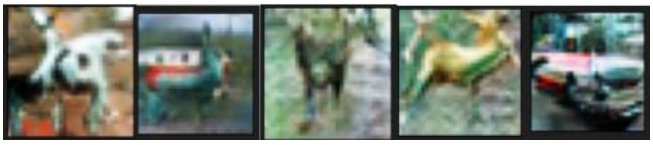


Fig. 12. Pruned + Quantized GAN generated images

The compression rate for the network after pruning and quantization is also lower than expected. During the pruning step, the network could not be reduced beyond .6 sparsity or its error would increase towards infinity during the fine-tuning step. While quantisation worked correctly, it also hurt the quality of the output. After pruning, quantization, and Huffman coding, a total compression of only 4.941 times smaller than the base model could be achieved.

V. CONCLUSION

Due to limitations within Keras and TensorFlow 2, the pruning and quantisation steps could not be finely controlled to create the smallest possible networks. One problem that occurred with using network compression in TensorFlow 2 is that saved networks include all parameters, including those removed by pruning and quantisation. Because of this, no benefits from compression could be seen until the saved output file is compressed using Huffman coding. The frameworks used also cause problems with compression because the software does not support quantisation beyond 32 8-bit. These problems cause our best results to have about 4x less compression than the theoretical max compression found in the Deep Compression paper [1].

Compressing the CNN network provides decent overall results. It was compressible down to 9.6x its original size. Additionally, accuracy decreased by only 1.6%. This makes the pipeline a vital tool for resource-constrained devices. While this is a great result, expanding the Keras tools for n-bit quantisation could have yielded even greater results. Compressing the RNN network produces the most impressive results of all the networks. Through the deep compression pipeline, the network, which was originally 32.32 MB in size, is reduced to just 1.49 MB, demonstrating the effectiveness of the pipeline. Furthermore, the experiment shows that the compressed LSTM can achieve comparable accuracy on sentiment analysis tasks, losing only 2.11% of the original accuracy.

This is significant enough for many networks to be implemented in memory-

Retrieval Number: 100.1/ijsce.C36130713323
DOI: 10.35940/ijsce.C3613.0713323
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

scarce infrastructure, which would not be able to handle the large size of the networks.

Compressing the GAN network generates poor overall results. It was only compressible to about 5x smaller than the original model. Throughout the compression process, a significant amount of quality is also lost. Some of these problems may have been due to the non-symmetrical compression done to the network. Because only the generator is compressed into a simpler network, the discriminator network may be able to adapt to the generator more quickly. The poor compression could also have been due to the overall complexity of producing convincing fake images. This complexity may have allowed small changes to the internal weights to cascade into significant problems with the output data.

While the deep compression pipeline is effectively working for CNN and RNN models to reduce network size with minimal performance degradation, it is not as effective for more complex networks, such as GANs. In our GAN experiments, performance degradation is excessive due to compression. For complex neural networks, we need to develop various compression methodologies.

VI. FUTURE WORK

To improve network compression performance, the Tensorflow 2 and Keras APIs can be modified. Within these libraries, the code for quantisation support can be updated to enable n-bit quantisation. To achieve this, the TensorFlow light-embedded kernel would also need to be updated to support the same levels of quantisation. Saved model support can also be improved to save weight matrices in a sparse row or sparse column format, rather than including every variable as a floating-point value. Due to the poor performance of the compression pipeline on the GAN network, a significant amount of work can be accomplished. More analysis and work should be conducted to identify the parameters that enable a GAN to be compressed without compromising output quality. Testing should also be done to determine if compressing both the generator and discriminator networks improves the output quality. For the CNN and RNN networks, future work would involve fine-tuning each stage of the compression pipeline further and developing more specialised tools to facilitate this.

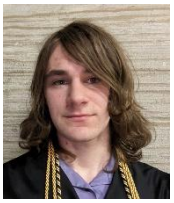
DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval or consent to participate, as it presents evidence that is already publicly available.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this article.

REFERENCES

1. S. Han, H Mao, and W J. Dally, “Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization, and Huffman Coding”, (ICLR) 2016
2. J.Luo, et al., “ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression”, IEEE International Conference on Computer Vision, 2017. [CrossRef]
3. H.Li, et al., “Pruning Filter for Efficient ConvNets”, International Conference on Learning Representations (ICLR), 2017.
4. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778
5. “Trim insignificant weights | TensorFlow Model Optimization,” https://www.tensorflow.org/model_optimization/guide/pruning (accessed Dec. 12, 2022).
6. “Quantization aware training in Keras example | TensorFlow Model Optimization,” https://www.tensorflow.org/model_optimization/guide/quantization/training_example (accessed Dec. 12, 2022).
7. “RNN, LSTM & GRU,” dProgrammer Lopez, Apr. 06, 2019. http://dprogrammer.org/rnn-lstm-gru
8. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. Commun. ACM 63, 11 (November 2020), 139–144. [CrossRef]

AUTHOR PROFILE



Lily Young received her Bachelor of Science in computer engineering from the University of Colorado, Colorado Springs, in May 2023. She will be joining Lockheed Martin. She is interested in applied machine learning, deep learning, FPGA design, microcontroller firmware design, operating system kernel and driver development, and cryptography. She intends to pursue an M.S. in computer engineering at the University of Denver.

James Richardson York earned his Bachelor of Science in Computer Engineering from the University of Colorado, Colorado Springs, in May 2023. He will be joining Northrop Grumman as an Associate Software Engineer before entering the US Air Force as a Pilot. During his studies, James completed a Senior Design Project with Semtech, optimizing their semiconductor silicon wafer processing using machine vision.

Byeong Kil Lee received a Ph.D. degree in Computer Engineering from the University of Texas at Austin in 2005. He is currently an assistant professor in the Department of Electrical and Computer Engineering at the University of Colorado, Colorado Springs. His current research interests include computer architecture, workload characterization of emerging applications, deep learning, low-power mobile processors, application-specific embedded systems, and cybersecurity.