CONTENT-WEIGHTED AUTONCODER FOR IMAGE COMPRESSION

Xiyang Wu

Georgia Institute of Technology

ABSTRACT

This paper presents a novel neural network structure for image compression tasks. The network structure utilizes the ResNet autoencoder for the encoder and decoder part, and incorporates the importance map to enhance the details in the image. The result shows that the compression performance of the convolutional autoencoder with the importance map is superior to the traditional 3-layer convolutional autoencoder and the ResNet autoencoder without the importance map under multiple image compression quality metrics.

Index Terms— Image Compression, Autoencoder, Importance Map

1. INTRODUCTION

As a heated topic in image processing, image compression is a fundamental and well-investigated topic. The general idea of image compression is to use fewer symbols to represent the given image. The traditional methods generally rely on the image transformation and manually designed block diagrams. For example, JPEG [1] is based on the discrete cosine transform, and the JPEG 2000 [2] is based on the multi-scale orthogonal wavelet decomposition. By reducing the internal relativity and quantizing the result, the transformation methods could represent the image with fewer features. Though almost effective, these manually designed compression methods may cause severe information loss under some certain circumstances.

Deep learning-based image compression methods, on the other hand, provides the adaptive image compression methods through the neural network. Mainly relying on the end-to-end model, modern deep learning based image compression method could be divided into two major genre: convolutional autoencoder with importance map and the super-resolution method. For the first method, L.Thesis et al [3] incorporates the convolutional autoencoder and the binary quantization layer in the lossy image compression. M. Li et al [4] first introduces the concept 'Importance Map' to the image compression task that achieves a better reconstruction performance by enhancing the importance parts in the image. F. Mentzer [5] further investigates the importance map method, and incorporates it with the ResNet-based autoencoder. C. Dong [6] presents the super-resolution method based on the deep convolutional neural network. As for the super-resolution methods, C. Ledig [7] and X. Wang [8] introduces the GAN method into the super-resolution field, and its performance is better than the BPG method. Methods based on the deep learning methods could not only achieve better reconstruction performance with more effective feature encoding, but also adapt to new media formats and contents [4], which indicates that the deep learning-based compression methods are more effective and general than the traditional methods.

In this paper, we investigate the start-of-art of the image-compression and propose a novel neural network structure that bases on the ResNet autoencoder and the importance map. Section 2 presents the motivation for this network. Section 3 presents the network structure, while Section performs the experiment on this network. The result shows that the compression performance of the convolutional autoencoder with the importance map is superior to the traditional 3-layer convolutional autoencoder and the ResNet autoencoder without the importance map.

2. MOTIVATION

The encoder of the super-resolution compression network, like the SRCNN, resembles a uniform down-sampling system while all parts in the image will be compressed by an equivalent and fixed rate. However, different parts in the given image may not contain the equivalent amount of information. The uniform downsampling method may deteriorate the reconstruction quality, since the SR-based decoder may not be able to recover the important details of the image.

To overcome the blindness of the simple down-sampling method, a novel idea called varying length quantization is proposed. First introduced by [3], the idea of lossy compression in the convolutional autoencoder is realized by encoding the output of the original encoder with 0 and 1. Obviously, this binary quantization method could deeply compress the image, but the quality of the reconstructed image could be serious deteriorated due to the information loss. To overcome this problem, [2] presents a novel method that uses quantizer varying in length to represent different parts of the image. This content adaptive method could enhance the important content in the image by encoding with more bits and achieve a good compression rate by maintaining a high compression rate, but its performance is greatly determined by the selection of the parameter, like the quantization length and the size of the encoding result.

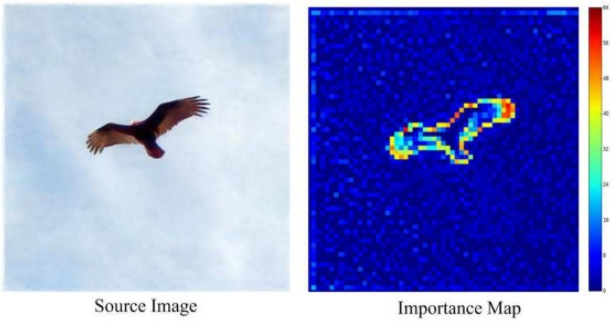


Fig. 1. The sample importance map and its corresponding original image [4]

As for the image compression tasks, since the distortion in the high-frequency part of the image may seriously affect the reconstruction quality, the enhanced parts in the proposed network mainly consist the high-frequency component of the image. To differentiate the components and enhance the important one, the important features will be encoded with more bits during the coding process, while the less important one will be encoded with fewer bits.

The importance of different features in the image is learnt by a separate network call importance map network. The relative importance of each encoded bit is presented by 8-bit integer, for example. During the quantization process, the first several bits of the original data will be extracted as the code of the input image, while its length is determined by the learnt importance. In this case, the original 32-bit float number is substituted by the 8-bit integer in the worst case. The size of the encoding result turns to 1/4 of the original one or even less.

A sample importance map and its corresponding original image are shown in Fig. 1. Details, like feather of the eagle, are considered as the important feature and marked with red in the importance map. Blocky features, like sky background, are considered as less important relatively marked with blue in the map. Represented by more bits, the important parts could be less likely to suffer from the quantization error and reserved better than the background during the compression process.

Though the quantization process can achieve better compression performance, the quality of the reconstruction image may be seriously deteriorated. Because of this, the encoder and decoder network must be carefully designed to minimize the quantization distortion and achieve a trade-off between the compression ratio and the reconstruction performance.

3. METHODS

As Section 2 mentions, our work mainly focuses on the content-weighted ResNet autoencoder. The whole structure of the autoencoder is presented in Fig. 2. The whole network could be divided into three parts: the ResNet encoder and decoder that follow a symmetric structure, and the importance network and quantization layer for major compression. To compare the compression performance, two baseline autoencoder are presented in our project: a 3-layer traditional convolutional autoencoder and a ResNet autoencoder using the same structure except the importance map and the quantization layer. The compression comparison bases on the size of the encoding output, which are for all three autoencoders.

3.1. ResNet Autoencoder

The ResNet autoencoder is used to coarsely compress the image. Inspired by the structure of SRGAN encoder, the ResNet autoencoder mainly consists two parts: the down-sampling network and the residual block group. The down-sampling network contains two convolutional layers. The

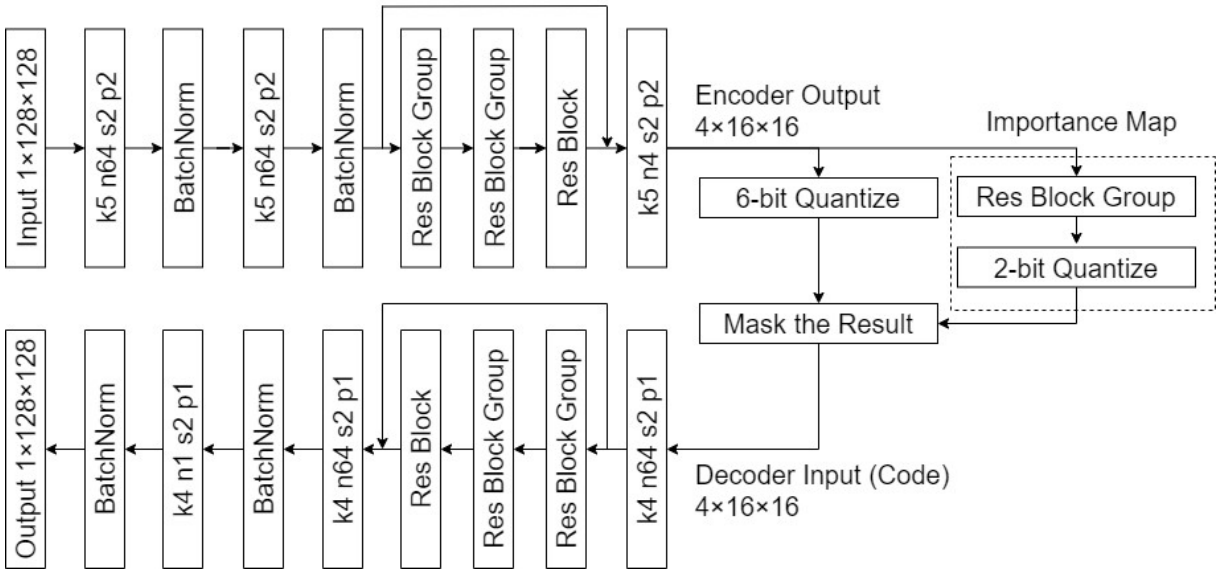


Fig. 2. The structure of the content-weighted ResNet autoencoder

goal of this part is to make the initial image fit the input size of the residual block groups. The residual block group, on the other hand, incorporates two residual blocks made up by three convolutional layers. The purpose of this is to preserve the image quality from the down-sampling loss by encoding the image with massive parameters. After the residual block group, the encoding result is processed by a convolutional layer to down-sample by 2.

The decoder is the inverse version of the encoder, while all the convolutional layers in the decoder are substituted by the corresponding transposed convolutional layers with the similar parameter combination. In this case, the input and output image size for the autoencoder could be the same.

3.2. Importance Map Network

Between the encoder and decoder of this network is the importance map network and the quantization layer, which plays a major role in the image compression. In this part, the encoding result is divided into two separate streams, and one of them will be sent into the importance network. The importance network mainly consists of a residual layer group, which extracts the details in images that seriously affects the quality of the reconstructed image. After the importance map, the quantization process for the result is also divided into two parts. The quantization layer with longer output length is used to encode the original output of the encoder, while the one with shorter output length is used to encode the output of the importance net. After quantization, the quantization result of the original encoding result is masked by the result of the importance net. Based on the content-weighted masking principle proposed in Section 2, the important features are presented with more bits, while the less important ones are presented with fewer bits. The masked result is the final encoding result.

3.3. Baseline Network

To illustrate the compression performance of the content-weighted autoencoder, two simplified baseline networks are designed. The first network is a 3-layer convolutional autoencoder. The encoder part consists three convolutional layer and two maxpooling layers. The maxpooling layers are used to down-sample the input image while the convolutional layers are used to compensate the number of the input and output channel and fit the proposed compression ratio. The decoder part of this network is symmetric version of the encoder. Batch normalization layers are used between each layer to prevent the gradient explosion. The encoding result of this autoencoder is the pure compression result of the encoder without extra quantization processing.

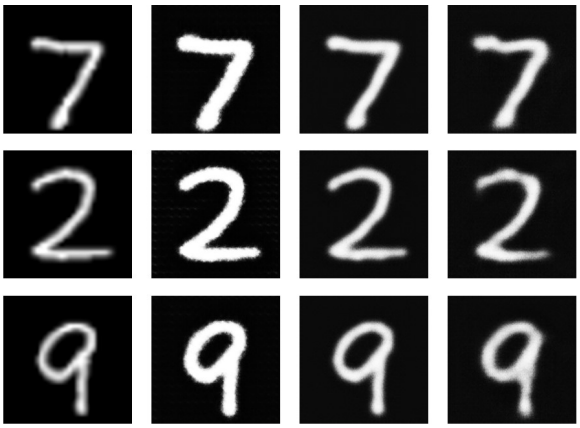
The other baseline network is the vanilla ResNet autoencoder without the importance map. Utilizing the same structure as the content-weighted ResNet autoencoder, the vanilla ResNet autoencoder uses the raw output of the encoder as the encoding result, without extra quantization processing. The evaluation for all the three network is based on the compression performance that bases on the size of encoding result, and the reconstruction quality reflected the image quality assessment metrics.

4. EXPERIMENTS

4.1. Experiment Design

Experiments for testing the performance of the content-weighted ResNet autoencoder incorporates its comparison with a pure 3-layer convolutional autoencoder and a ResNet autoencoder without the importance map. All networks are trained for 50 epochs. Training and testing datasets for all the models are MNIST. The size of the input and output image is 1×128×128, while the encoding result for all the three encoder networks is 4×16×16. The learning rate is 10-3. The optimizer is SGD. As for the quantization part, the output of the importance map is quantized as 2-bit integer, while the original encoder output is quantized as 6-bit integer. In this case, each pixel in the original encoder output is represented by an integer between 0 and 63, while the output of the importance map will be an integer between 0 and 3. During the masking process, first several bits of the quantization result of the original image will be extracted according to the result of the importance map, which will be sent into the decoder and calculate the backward gradient. The loss function of each network is MSE.

The performance evaluation mainly covers two aspects. The first is the compression performance that mainly relies on the output size of the encoder. In this project, the encoder result is simply encoded by the Huffman coding. To compare the code length of the encoding result, we will extract the first 6 digits of the encoding result of the ResNet autoencoder and the convolutional autoencoder, since the original data type of the output for these networks is float. Without utilizing sophisticated encoders, like the entropy encoder of JPEG2000, the actual compression performance



(a) (b) (c) (d)

Fig. 3. The reconstruction images (a) Original (b) CAE (c) ResNet AE (d) Content-weighted AE

of all three networks might be better than the result presented in this paper. The compression performance will be represented by BPP.

The other performance evaluation is the reconstruction quality, which mainly bases on the comparison between the original testing image and the reconstructed image. The testing images for this project include four handwritten digit images randomly extracted from the MNIST dataset. The evaluation metrics for the reconstruction performance of each network includes the MSE and PSNR between the original image and the reconstructed image.

4.2. Result

The sample result for all the autoencoders are shown in Fig. 3. Images that locate from left to right are the original image, the result of the 3-layer convolutional autoencoder, the result of the ResNet autoencoder, and the result of the content-weighted ResNet autoencoder with the importance map method.

Metrics	CAE	ResNet	Content-weighted AE
BPP	0.5735	0.6245	0.3566

Table. 1. The BPP result for each model

Based on the evaluation metrics we proposed before, we can derive the evaluation for the output of each network. The BPP result for each network that relates to the compression performance is shown in Table. 1. According to the BPP result, we can find that content-weighted ResNet autoencoder can greatly decrease the length of the encoding result under the same training hyperparameters. Its BPP value is prior to the result of the convolutional autoencoder and the ResNet autoencoder, which reveals its great potential on deep image compression with high quality.

MSE	CAE	ResNet	Content-weighted AE
Image 1	44.42	129.50	127.05
Image 2	45.76	104.49	114.14
Image 3	46.51	118.86	115.64

Table. 2. The MSE result for each model on testing images

PSNR	CAE	ResNet	Content-weighted AE
Image 1	31.65	27.01	27.09
Image 2	31.53	27.94	27.56
Image 3	31.46	27.38	27.50

Table. 3. The PSNR result for each model on testing images

As for the quality metrics, the result of the MSE and PSNR result for each testing image has been shown in Table. 2. and Table. 3. According to the MSE and PSNR result of the reconstructed image, we can find that the MSE and PSNR value for the result of the convolutional autoencoder is better than the result of the ResNet autoencoder, which means that the result of the convolutional autoencoder is closer to the original image. The result of the ResNet autoencoder and the one with the importance map layer are approximately the same.

However, since the main-stream image assessment metrics, like the PSNR and the SSIM, all have their own weakness, which may not reflect the actual quality of the image, we will comment on the quality of the reconstruction image subjectively. Though the performance of the convolutional autoencoder's output is clearer and brighter than the other reconstruction result, its structure does not resemble the corresponding original image, since the gray-value distribution resembles the original image. Considering the superior compression performance, the compression result made by the content-weight ResNet autoencoder only slightly suffers from reconstruction error, which is tolerable based on the subjective assessment. Based on this, we can announce that the content-weight ResNet autoencoder reveals a better image compression performance than the other two methods.

5. CONCLUSION

This project presents a novel content-weighted autoencoder network structure for the deep image compression tasks. To deal with the serious information loss made by high compression ratio and missing details, this network presents an adaptive compression method that encode each pixel compression result with length-varying integers, while the important features in the image are encoded with more bits. By enhancing important features in the image, high frequency components that mainly determines the image reconstruction quality could be greatly preserved. The idea of importance map is implemented on the ResNet based autoencoder. During the experiment, the content-weighted autoencoder reveals its priority to the other two baseline networks on the given dataset.

However, the content-weighted ResNet autoencoder is not satisfying. Further investigation on this topic will focus on the following aspects, including introducing soft-quantization method [10] to avoid severe quantization error, or further optimizing the loss function and the model structure for a more robust image compression model.

7. REFERENCES

[1] Wallace G K. The JPEG still picture compression standard [J]. IEEE Transactions on Consumer Electronics, 1992, 38(1).

[2] Rabbani M, Joshi R. An overview of the JPEG 2000 still image compression standard [J]. Signal processing: Image communication, 2002, 17(1): 3-48.

[3] Theis L, Shi W, Cunningham A, et al. Lossy image compression with compressive autoencoders [J]. arXiv preprint arXiv:1703.00395, 2017.

[4] Li M, Zuo W, Gu S, et al. Learning convolutional networks for content-weighted image compression [C] Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3214-3223.

[5] Mentzer F, Agustsson E, Tschannen M, et al. Conditional probability models for deep image compression [C] Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4394-4402.

[6] Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 295-307.

[7] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network [C] Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[8] Wang X, Yu K, Wu S, et al. Esrgan: Enhanced super-resolution generative adversarial networks [C] Proceedings of the European Conference on Computer Vision (ECCV). 2018: 0-0.

[9] Agustsson E, Tschannen M, Mentzer F, et al. Generative adversarial networks for extreme learned image compression [C] Proceedings of the IEEE International Conference on Computer Vision. 2019: 221-231.

[10] Agustsson E, Mentzer F, Tschannen M, et al. Soft-to-hard vector quantization for end-to-end learning compressible representations [C] Advances in Neural Information Processing Systems. 2017: 1141-1151.