Extended End-to-End optimized Image Compression Method based on a Context-Adaptive Entropy Model

Jooyoung Lee*, Seunghyun Cho, Se-Yoon Jeong, Hyoungjin Kwon,
Hyunsuk Ko, Hui Yong Kim & Jin Soo Choi
Broadcasting and Media Research Laboratory
Electronics and Telecommunications Research Institute
Daejeon, Korea
leejy1003@etri.re.kr

Abstract

In this paper, we propose an extended compression method using a context-adaptive entropy model. Based on the Lee et al. *[11]'s approach, we extend the network structure so that compression and quality enhancement methods are jointly optimized. In terms of contexts for estimating distributions, we additionally use offset information. By exploiting the extended structure and the additional contexts, we obtain substantially improved compression performance, in terms of multi-scale structural similarity (MS-SSIM) index, compared to the model without the extensions.*

1. Introduction

Recently, artificial neural networks (ANNs) based image compression methods have been studied in various ways. Some approaches have been proposed for enhancing each tool of conventional compression codecs, and other approaches have been studied for post-processing of the reconstructed such as artifact reduction or super-resolution methods. Meanwhile, end-to-end image compression approaches [15, 8, 4, 14, 5, 11, 13] have been proposed based on strong optimization capabilities of neural networks. These approaches can be divided into two classes, distinguished based upon whether entropy models are used or not. Toderici *et al*. [15] introduced a novel ANN-based image compression method using a small number of latent binary representations, and Johnston *et al*. [8] improved the compression performance by enhancing the model operation methods. On the other hand, Other approaches [4, 14, 5, 11, 13] view the image compression problem as entropy minimization. They use entropy models to approximate expected data rates, and the rates are used in the objective functions so that the trained model generates latent variables having as low entropy as possible. Ballé *et al*. (2017) [4] and Theis *et al*. [14] introduced the first entropy minimization approaches and leaded many researchers to study more efficient entropy models. Ballé *et al*. (2018) [5] enhanced entropy models by adopting hierarchical networks for estimating scales of hidden representations, whereas the former two approaches train fixed entropy models. Minnen *et al*. [13] and Lee *et al*. [11] assume entropy-coding and decoding process of latent variables are conducted in a sequential manner (e.g. a raster scanning order), so they utilize known neighborhood components of hidden variables as additional contexts for estimating distributions of latent variables. Both approaches enhanced the compression performance of ANN-based image compression, and obtained the results better than BPG [6], a HEVC (ISO/IEC 23008-2, ITU-T H.265) [7] based image compression method. In this paper, we propose a extended compression method based on Lee *et al*. [11]'s approach, and demonstrate substantially improved compression results in terms of multi-scale structural similarity (MS-SSIM) index.

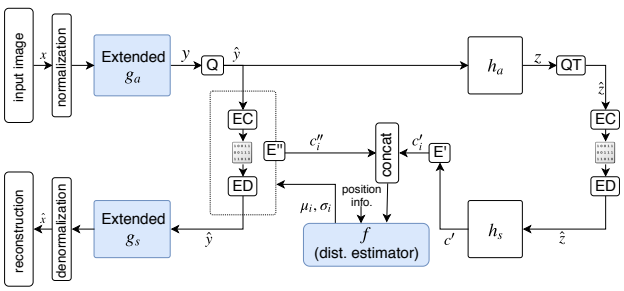

Figure 1. Overall structure of the proposed method.

*Corresponding author

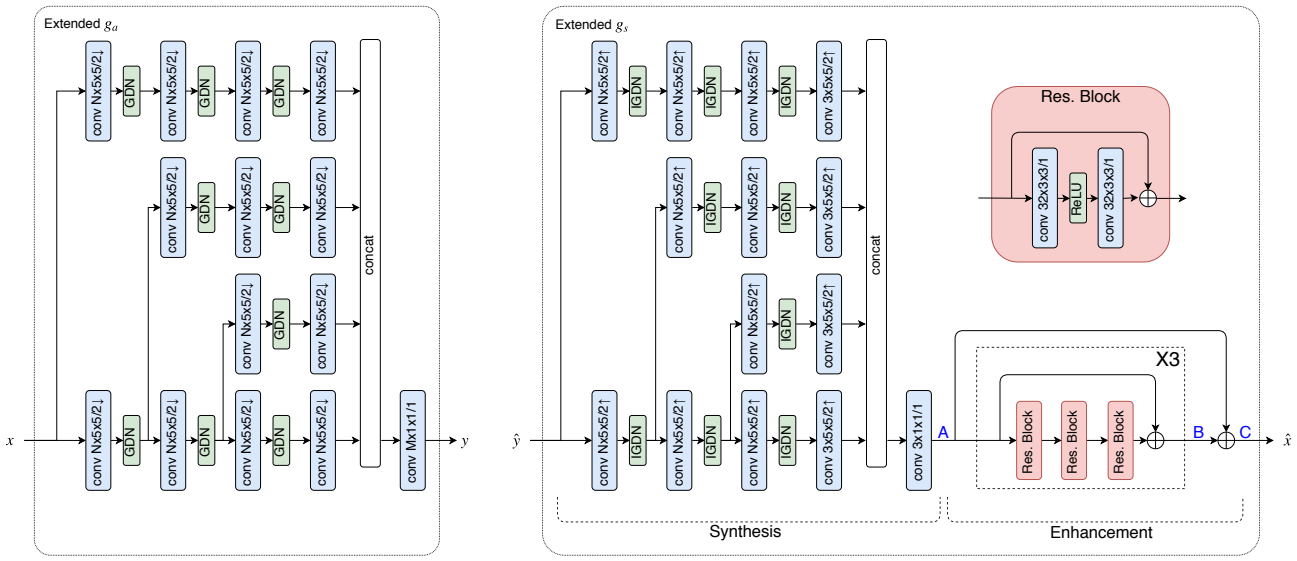

Figure 2. Structure of the extended analysis and synthesis transform networks.

2. Proposed model

We basically follow Lee *et al*. [11]'s model, so our method conceptually has the same entropy model. In our model, as in Lee *et al*. [11]'s model, input $\mathbf{x}$ is transformed into $\mathbf{y}$, and some spatial correlations of $\mathbf{y}$ are further transformed into $\hat{\mathbf{z}}$. Four fundamental parametric transform functions, an analysis transform $g_a(\mathbf{x};\phi_\mathbf{g})$, a synthesis transform $g_s(\hat{\mathbf{y}};\theta_\mathbf{g})$, an analysis transform $h_a(\hat{\mathbf{y}};\phi_\mathbf{h})$, and a synthesis transform $h_s(\hat{\mathbf{z}};\theta_\mathbf{h})$ are used, likewise. However, we extend the compression model in terms of both network structure and contexts as represented in Figure 1. The extended parts in the compression structures are highlighted with blue color.

2.1. Network structure extension

Lee *et al*. [11]'s approach has the same structures of g_a and g_s as in Ballé *et al*. (2018) [4]'s approach, which arranges convolutional layers and non-linear layers in an alternate order. Based on this structure, we change two structural parts of the transform networks. First of all, we connect each non-linear layer (GDN [3]) in g_a and g_s to a designated network, and all the results from the designated networks are aggregated by 1x1 convolution. Note that normalization and denormalization steps are omitted in Figure 2 for simple illustration. Although total number of layers are higher than that of Lee *et al*. [11]'s model, we minimize increase in complexity by using smaller number of filters at each layer: We set N to 64.

In addition to the sub-networks designated for each scale of transforms, we also append a quality enhancement part at the end of the g_s synthesis network, inspired by ANN-based quality enhancement methods [9, 12]. We built this quality enhancement part using residual building blocks, and we incorporate hierarchical skip connections at three different levels of the structure. Consequently, our synthesis model can be viewed as a synthesis transform function jointly optimized with a quality enhancement method. A, B, and C in Figure 2, highlighted with blue font, represent intermediate images, generated residuals, and the final reconstructions, respectively. Some samples of the three types of images are shown in Figure 3.

As described above, our g_s network consists of two parts, synthesis part and enhancement part. One advantage of this structure is that it allows us to adopt various new quality enhancement structures, regardless of synthesis method. However, although the enhancement part is replaceable from a structural viewpoint, joint optimization is inevitable at the moment.

2.2. Context extension

Lee *et al*. [11]'s approach exploits two different types of contexts, bit-consuming contexts and bit-free contexts. For bit-consuming contexts, they use additional side information based on hierarchical priors, similar to Ballé *et al*. (2018) [5]'s approach, and they utilize the known neighborhood components of hidden variables as bit-free contexts. We basically follow the same framework, but we utilize one more additional bit-free context, positional information of the current spatial point of the latent variable $\hat{y}_i$. The MS-SSIM metric has a characteristic that as positions of image components are close to borders of images, fidelity of those components contribute less to the final MS-SSIM value, compared to those in the middle of the images. Therefore, reconstructed images from the image com-

pression approaches optimized for MS-SSIM tend to have higher-fidelity in the middle area than the area close to borders. Considering hidden representations preserve spatially corresponding relationship with input space more or less, we can intuitively expect that the distributions of the hidden representations vary according to their coordinates. However, because the contexts utilized in Lee *et al*. [11]'s approach are subset of results coming from convolutional layers that have a translation-invariant characteristic, the estimator f cannot reflect the characteristic of $\hat{y}_i$'s distribution varying according to its position. Therefore, we provide the distribution estimator f with a position of a current representation. More specifically, we utilize offsets from four spatial borders of $\hat{\mathbf{y}}$ as position information. The offsets of current representation $\hat{y}_i$ is transformed into the form of one-hot vector, and then concatenated to the results of the convolutional layers in the estimator f, as shown in Figure 4.

3. Experiments

3.1. Experimental environments

We trained all the models using MS-SSIM based distortion term. We trained the model using 51,141 256×256 patches extracted from CLIC [2] 2019 trainset images. We

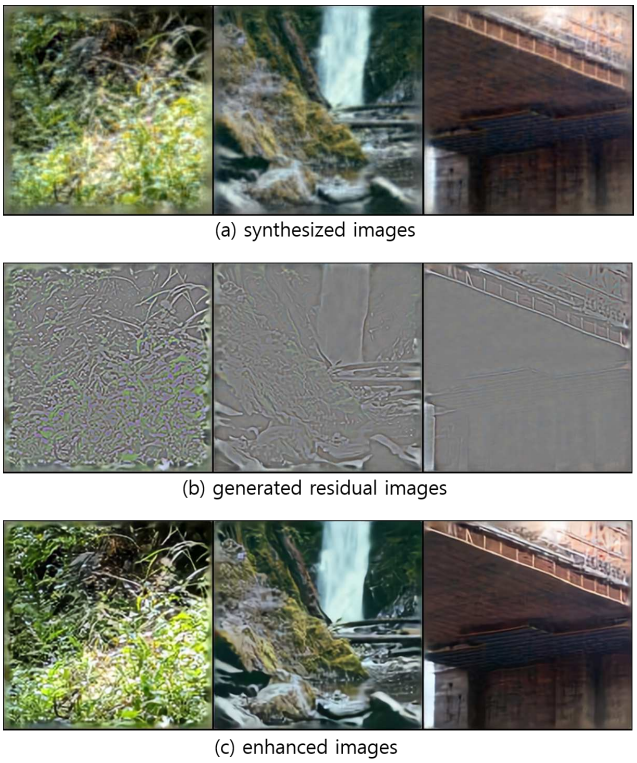

Figure 3. Examples of synthesized images, generated residual images, and enhanced images.

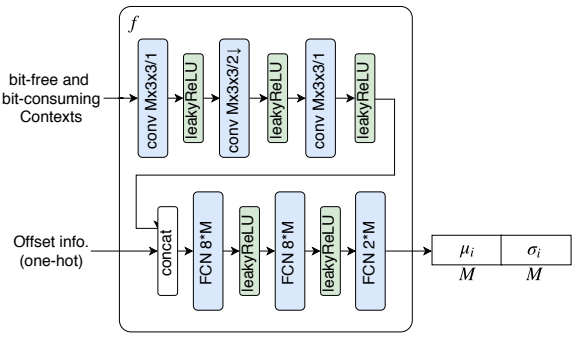

Figure 4. Structure of the extended distribution estimator f

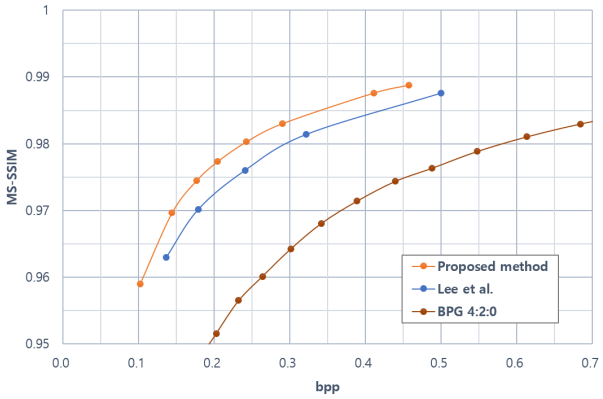

Figure 5. Experimental results over CLIC test set.

used eight patches per batch, and 1M iterations of training steps were conducted using ADAM optimizer[10]. Initial learning rate was set to 1×10^{-4}, and then reduced in half every 25,000 steps for the last 200,000 steps. For each trained model, we measured the average bit per pixel (BPP) and average MS-SSIM over the CLIC 2019 challenge testset that consists of 330 images, and compared the results with Lee *et al*. [11]'s approach and BPG 4:2:0 [6].

3.2. Experimental results

Figure 5 demonstrates experimental results of the proposed model, Lee *et al*. [11]'s approach and BPG [6]. the proposed model outperforms both of the other approaches. In terms of MS-SSIM BD-rate, we obtained 19.38% and 60.40% of compression gains over Lee *et al*. [11]'s approach and BPG 4:2:0, respectively. However, note that we measured the results of Lee *et al*. [11]'s approach using their publicly distributed general models [1], whereas our method can be viewed as a specialized model for the CLIC challenge. In other words, some level of compression gains of our model, more or less, may come from the dedicated environments or fine tuning processes, such as use of CLIC trainset or padding optimization. Therefore, to

Image set	Average MS-SSIM
CLIC validation set	0.975208
CLIC test set	0.972710

Table 1. Average MS-SSIM results within 0.15 bpp.

reach a generalized conclusion, further experiments, including oblation studies, are required on the same experimental environments.

In addition, we also measured maximized average MS-SSIM values over CLIC validation set and test set, respectively. To maximize the average MS-SSIM values, we used 30 trained models, and chose the best model in terms of quality to rate. Table 1 shows the optimized average MS-SSIM values over CLIC validation set and test set images, given the constraint that compression is to less than 0.15 bpp across the full image set.

4. Conclusion

In this paper, we extended image compression models based on Lee *et al*. [11]'s work in two different aspects, network structure and utilized contexts. We used designated networks for each hidden layers of transform functions, g_a and g_s, and in addition, we extended the synthesis transform g_s by incorporation skip-connection utilizing layers, which are inspired by the quality enhancement studies. To the best of our knowledge, our model is the first joint optimization model including compression and quality enhancement functions together, and thereby we obtained the superior results as shown in the experimental results. However, as mentioned, the challenge-dedicated environments contribute to the obtained compression gains more or less. Therefore, as a future work, we will provide an ablation study on various components proposed in this paper, and MSE-optimized version of our model will also be studied further.

Acknowledgments

This work was supported by Institute for Information and communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) 2017-0-00072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media.

References

[1] Repository of the paper "context-adaptive entropy model for end-to-end optimized image compression", 2018. https://github.com/JooyoungLeeETRI/CA_Entropy_Model. 3

[2] Workshop and challenge on learned image compression, 2019. https://www.compression.cc/. 3

[3] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. Density modeling of images using a generalized normalization transformation. In *the 4th Int. Conf. on Learning Representations*, 2016. 2

[4] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *the 5th Int. Conf. on Learning Representations*, 2017. 1, 2

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *the 6th Int. Conf. on Learning Representations*, 2018. 1, 2

[6] Fabrice Bellard. Bpg image format, 2014. http://bellard.org/bpg/. 1, 3

[7] Information technology – high efficiency coding and media delivery in heterogeneous environments – part 2: High efficiency video coding. Standard, ISO/IEC, 2013. 1

[8] Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, and George Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[9] Jiwon Kim, Jung Kwon Lee, , and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *the 3rd Int. Conf. on Learning Representations*, 2015. 3

[11] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In *the 7th Int. Conf. on Learning Representations*, May 2019. 1, 2, 3, 4

[12] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 2

[13] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, May 2018. 1

[14] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszr. Lossy image compression with compressive autoencoders. In *the 5th Int. Conf. on Learning Representations*, 2017. 1

[15] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1