This CVPR 2020 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

A Hybrid Image Codec with Learned Residual Coding

Wei-Cheng Lee
weicheng.ee07g@nctu.edu.tw

Hsueh-Ming Hang
hmhang@nctu.edu.tw

Department of Electronics Engineering
National Chiao Tung University, Taiwan

Abstract

We propose a three-layer image compression system consisting of a base-layer VVC (intra) codec, a learning-based residual layer codec, and a learnable hyperprior. This proposal (Team: NCTU_Commlab) is submitted to the Challenge on Learned Image Compression (CLIC) in March 2020*. Our contribution is developing a data fusion attention module and integrating several known components together to form an efficient image codec, which has a higher compression performance than the standard VVC coding scheme. Unlike the conventional residual image coding, both our encoder and decoder take inputs also from the base-layer output. Also, we construct a refinement neural network to merge the residual-layer decoded residual image and the base-layer decoded image together to form the final reconstructed image. We tested two autoencoder structures for the encoder and decoder, namely, CNN with GDN [5, 6], and the generalized octave CNN [4]. Our results show that the transmitted latent representations are very efficient in coding the residuals because the object boundary information can be provided by the proposed spatial attention module. The experiments indicate that the proposed system achieves better performance than the single-layer VVC at both PSNR and subjective quality at around* 0.15 *bit-per-pixel.*

1. Introduction

Recent researches have revealed that the deep learning-based image compression methods potentially outperform the traditional coding scheme, JPEG2000, the H.265/HEVC-based BPG [7] image codec, and the latest versatile video coding (VVC)[8].

In participating in the low-rate track of CLIC 2020, we propose a hybrid coding scheme, which consists of a VVC (intra) codec as the base-layer, and a residual-layer which uses an autoencoder architecture and a conditional entropy model based on Gaussian mixture model. The function of the residual-layer is to reduce the artifacts generated by the base-layer at lower bitrates. This method is inspired by the conventional layered coding concept, and the recent deep learning-based codec in [6], [12], [3], and [13]. At the receiver, the decoder network takes the inputs from the outputs of the base-layer and the residual-layer to produce an intermediate output, which is then processed by a refinement network to resynthesize the final output image. This refinement network can be viewed as a post-processing procedure to further improve the image quality. At a total bit rate around 0.15 bit-per-pixel (bpp), our system allocates about 0.125 bpp to the VVC (intra) base-layer coding and about 0.025 bpp to the residual-layer coding.

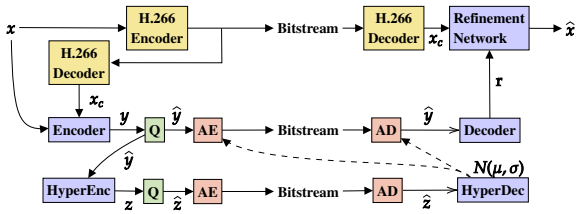

Figure 1. The architecture of the proposed image compression system with residual coding.

Our contribution in this study is developing a data fusion attention module and integrating several known components together to form an efficient image codec, which has a higher compression performance than the standard VVC coding scheme. Unlike the conventional multi-layer coding, both our encoder and decoder take inputs also from the base-layer output. In addition, we construct a refinement neural network to merge the residual-layer decoded residual image and the base-layer decoded image together to form the final reconstructed image. We tested two autoencoder structures for the encoder and decoder, namely, CNN with GDN (Generalize Divisible Normalization) [5, 6], and the generalized octave CNN [4]. Our results show that in both cases the transmitted latent representations are very

efficient in coding the residuals because the object boundary information can be provided by the proposed spatial attention module. The experiments indicate that the proposed system achieves better performance than the single-layer VVC at both PSNR and subjective quality at around 0.15 bit-per-pixel. Because the above two autoencoder structures have similar compression performance, only the CNN with GDN is submitted to the CLIC contest (Team: NCTU_Commlab).

2. Related work

The recent promising learning-based image compression studies [5, 6, 10, 11, 12, 14, 15, 16] adopt the autoencoder structure to extract the latent representation of input data. Tt was shown that a CNN neural net with GDN nonlinearity is able to achieve a high image compression efficiency comparable to that of BPG [5, 6]. Furthermore, a recent proposal using the generalized octave convolution can achieve a comparable performance with the best conventional codec, VVC [8].

In addition to the trainable autoencoder, many learned codecs contain a highly efficient entropy coder (arithmetic coder) and an accurate rate estimator. In [6], the Gaussian scalar mixture (GSM) model is used to model the entropy probability distribution, and it is implemented by using the so-called hyperprior network to estimate entropy in an end-to-end training manner. Furthermore, a complicated conditional entropy model based on the Gaussian mixture model (GMM) was mentioned in [14],and proposed in [10, 4]. For the multi-layer coding, a deep semantic segmentation-based layered image compression (DSSLIC) was proposed in [3], which integrates the semantic segmentation task with image compression; it intends to take the advantages of the learning-based system and the BPG-based codec.

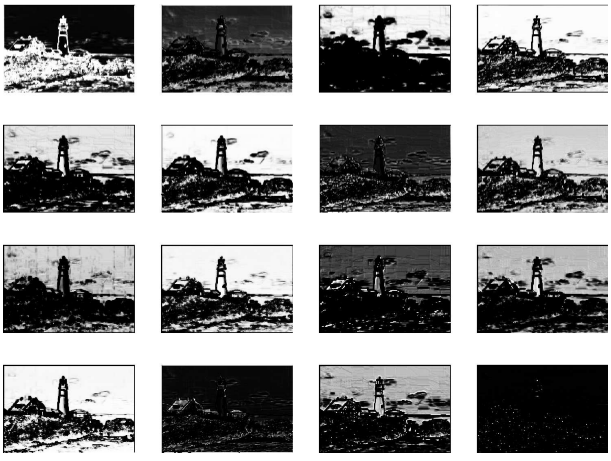

Figure 2. Visualization of attention masks.

3. Proposed Methods

3.1. Overall architecture and design philosophy

Fig.1 show the proposed our multi-layer image compression system. As shown in Fig.3, the input to the encoder is a six-channel tensor, which is a concatenation of a VVC (intra) decoded image and the original image in RGB format. Given an input image x, the VVC (intra) encoder produces a compressed image x_c. With both x and x_c, a learning-based encoder generates a set of residual feature maps $z = E(x; x_c)$. To simulate the quantization effect in the training process, we adopt the additive noise technique proposed in [5] to prevent the zero gradient problem caused by quantization. Let $\hat{y} = Q(y)$; we adopt the technique in [14] , which uses the Gaussian mixture model $N(\mu, \sigma)$ to approximate the probability distribution of $\hat{y}$. The parameters of the Gaussian model are estimated by a hyperprior model (neural net). The probability model is used to estimate the entropy value in the training phase and it estimates the conditional probability and used by the arithmetic coder in the testing phase. At the beginning of the decoding stage, the residual feature sample $\hat{y}$ is first recovered by the decoder network $D(\hat{z}; x_c)$. And then the refinement network generates the final reconstructed image $\hat{x} = R(D(\hat{z}; x_c); x_c)$ using both decoded VVC (intra) image and decoded residuals.

Noted that instead of using the difference between the original image x and the decoded base-layer image $\boldsymbol{x_c}$, the encoder takes inputs of both these two images. Because our residual-layer encoder input is not identical to the residual image and it also makes use of the base-layer decoded output, it may be able to provide some additional "side-information" for reconstructing the original image at the decoder. Furthermore, we include a spatial attention mechanism at both encoder and decoder, which takes input only from $\boldsymbol{x_c}$ and produces an attention mask. With the help of the attention mask, the autoencoder can predict the location of residuals quite precisely, and thus only the most meaningful information is coded. An example of attention mask is shown in Fig.2.

The decoder is roughly the reverse of the encoder but the decoder has an additional refinement network, which has two inputs: the base-layer reconstructed image and the output of the residual-layer decoder. Therefore, the refinement network plays the role of postprocessing. Nevertheless, because it is included in the training loop, its function is more than simply combining the base-layer image and the residuals.

We tried two CNN structures for the residual encoder and decoder. The first structure is modified from the autoencoder architecture, convolutional layers with GDN nonlinearity (Fig.3), proposed by [5]. The second structure is the generalized octave convolution (GoConv) adopted from [4].

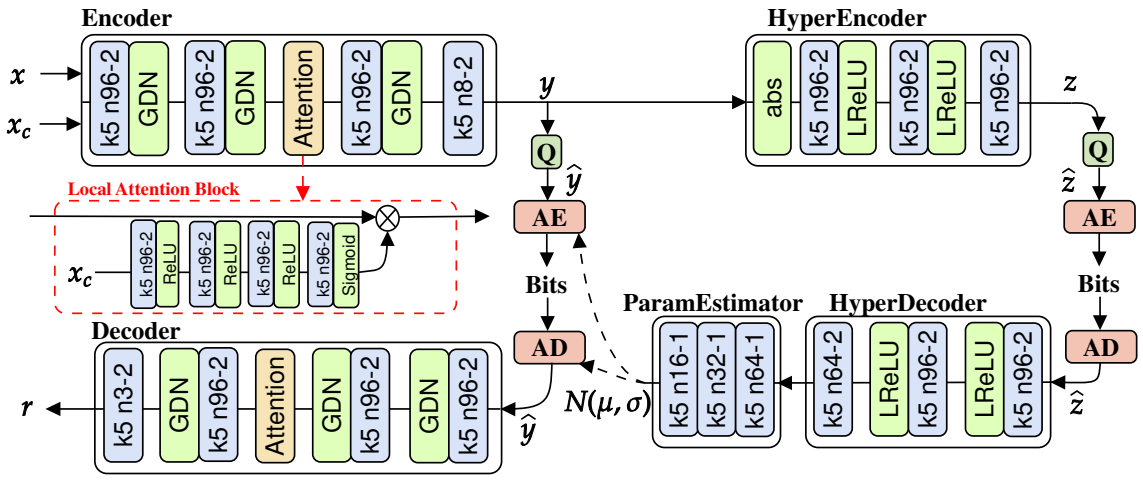

Figure 3. The architecture of proposed autoencoder and hyperprior

In the implementation of the generalized octave network, we replace all the convolution and transpose-convolution layers by GoConv and GoTConv layer (Fig.4) in our architecture.

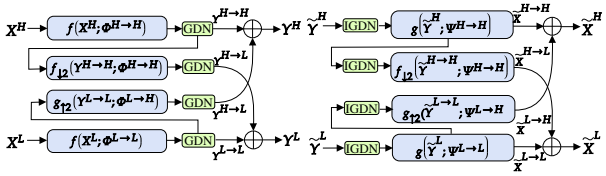

Figure 4. Architecture of generalize octave convolution (GoConv) shown in the left figure, and transposed-convolution (GoTConv) shown in the right figure.

3.2. The loss function

In this work, the loss function is defined as:

$$\mathcal{L} = \lambda \times L_R(\boldsymbol{q}) + L_D(\boldsymbol{x}, \hat{\boldsymbol{x}}), \tag{1}$$

where L_R is the entropy estimation by the Gaussian mixture model using the parameters produced by the hyperprior and $L_D(x; \hat{x})$ is defined by some measure of reconstruction error, such as mean square error (MSE) or multiscale structure similarity (MS-SSIM).

In the conventional compression schemes, the reconstructed errors usually have large magnitude response in the high frequency region. Since the residual signal has generally the properties of small variation and spatial sparsity, the general-purpose network may not be suitable for its representation. With the help of base-layer decoded image as the extra information, the autoencoder can put more attention on the high frequency region for bit allocation. In our structure, the entire encoder produces two separate bit streams: the VVC (intra) codes and residual codes.

3.3. Local Attention Module

Fig.3 also shows the proposed local attention module. It consists of 4 layers of CNN. Fig.2 shows the local attention masks produced by our attention module. These attention masks signal the edges of objects, where often the significant coding errors locate. The idea of using attention masks is to reduce the transmission of the duplicated information among the feature maps. Some portions of the high-level feature maps are suppressed in processing and thus they reduce the overlapped information. And the ultimate goal is to reduce the bits in transmission.

(a) (b) (c)

Figure 5. Comparison between images using VVC (intra) and using our system at around 0.12bpp. (a)Original image (b)VVC (intra): 0.184 bpp, MS-SSIM 0.9468, PSNR 31.53dB;; (c)Our system: 0.181 bpp, MS-SSIM 0.9642, PSNR 32.01dB

4. Experiments

4.1. Autoencoder and refinement network

Our autoencoder and the hyperprior structure is similar to that in [6] shown in Fig 3.

In our system, the bitrate ratio of the base-layer and the residual-layer is about 5 : 1. The intermediate feature map channel number is 96 inside the autoencoder and 8 chan-

nels for the transmitted feature maps due to the 0.15 bpp constraint, which is an extremely low bitrate constrain. The notation "k5 n96-2" represents a convolution layer with kernel size 5, 96 output channels, and a stride of 2. The hyper autoencoder takes absolute value of latent representation y as input, and produces 16 output channels representing the corresponding mean and sigma pair of y for the entropy coding.

4.2. Training

Our training dataset contains 1672 images provided by CLIC 2020 [1]. In the training phase, we randomly crop these images into 256×256 patches as inputs to our autoencoder. For validation, we use the Kodak dataset [2]. The training procedure of our proposed method consists of two steps. The first step is to code the base-layer images using the VVC (intra) codec with appropriate quantization parameters. The generated decoded images are used in the second phase learnable training. In the second step, we train the parameters in the autoencoder and the refinement nets using the cost function defined by (1) with MS-SSIM distortion.

We use Adam optimizer [9] with a mini-batch size 8 to train the model optimized for MS-SSIM. Setting a initial learning rate at $1 \cdot 10^{-3}$ and for $1 \cdot 10^{-4}$ for the autoencoder and the hyperprior respectively. The training procedure is similar for both CNN with GDN and the GoConv network.

Model	Bits/Pixel	PSNR	MS-SSIM
BPG420	0.1478	28.372	0.9182
VTM7.1	0.1482	29.141	0.9305
Ours(MS-SSIM)	0.1439	27.472	0.9579
Ours(PSNR)	0.1361	30.498	0.9578

Table 1. Compression result comparison on CLIC validation set and Kodak dataset.

4.3. Experiment results

For the low rate track, we focus on the perception quality. The distortion loss is defined as $D = 1 - L_(MS-SSIM)$, and then four models with λ=0.2/0.3/0.4/0.5 in the loss fuction 1 are trained to match the target bitrate. Then, the individual image feature map optimization technique is used to fine-tune the feature map for each image. Table 1 shows the comparison of our method (CNN with GDN) optimized for MS-SSIM around 0.15 bpp. Fig.5 compares the compressed images at around the same rate using VVC (intra) and using our scheme.

The rate-distortion curves of several schemes are shown in Fig 6. Our hybrid coding scheme can achieve higher MS-SSIM value without sacrificing too much PSNR value.

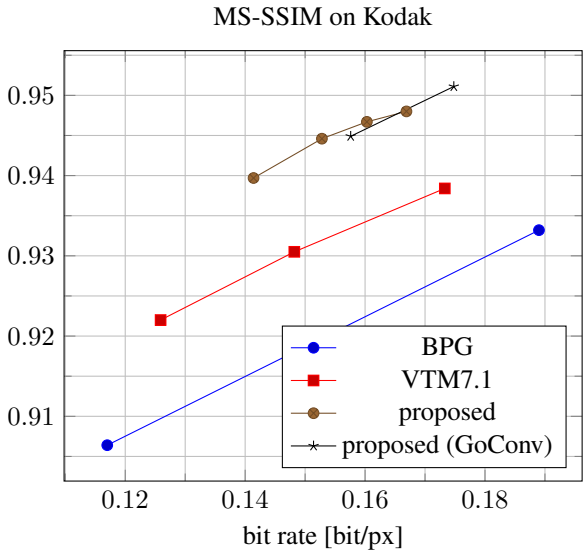

Figure 6. RD curve of on Kodak dataset.

5. Conclusions

In this paper, we propose an end-to-end trainable hybrid image coding scheme. The base-layer is the standard VVC (intra) codec, and the residual-layer is a deep-learning based codec. Our learning-based encoder and decoder is modified from [5]. We add a local attention module to enhance the RD performance, and insert a refinement net to synthesize the final reconstructed image. We also integrate the latest GoConv component in the architecture. At the end, our method outperforms the original VVC (intra) particularly on the MS-SSIM index.

6. Acknowledgment

We are grateful to the National Center for High-performance Computing for computer time and facilities. This work was supported in part by Ministry of Science and Technology, Taiwan under Grant MOST 108-2634-F-009-007 through Pervasive AL Research (PAIR) Labs, National Chiao Tung University, Taiwan.

References

[1] Challenge on Learned Image Compression. http://compression.cc/. 4

[2] Kodak PhotoCD dataset. http://r0k.us/graphics/kodak/. 4

[3] M. Akbari, J. Liang, and J. Han. Dsslic: deep semantic segmentation-based layered image compression. In *ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2042–2046. IEEE, 2019. 1, 2

[4] M. Akbari, J. Liang, J. Han, and C. Tu. Generalized octave convolutions for learned multi-frequency image compression. *arXiv preprint arXiv:2002.10032*, 2020. 1, 2

[5] J. Ballé, V. Laparra, and E. P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016. 1, 2, 4

[6] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018. 1, 2, 3

[7] F. Bellard. BPG image format. http://bellard.org/bpg. 1

[8] HHI. Fraunhofer. VVC official test model VTM. https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM. 1, 2

[9] D. P Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[10] J. Lee, S. Cho, and M. Kim. A hybrid architecture of jointly learning image compression and quality enhancement with improved entropy minimization. *arXiv preprint arXiv:1912.12817*, 2019. 2

[11] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3214–3223, 2018. 2

[12] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. V. Gool. Conditional probability models for deep image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4394–4402, 2018. 1, 2

[13] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. V. Gool. Practical full resolution learned lossless image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10629–10638, 2019. 1

[14] D. Minnen, J. Ballé, and G. D. Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, pages 10771–10780, 2018. 2

[15] O. Rippel and L. Bourdev. Real-time adaptive image compression. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2922–2930. JMLR. org, 2017. 2

[16] L. Theis, W. Shi, A. Cunningham, and F. Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017. 2