Deep Image Compression with Latent Optimization and Piece-wise Quantization Approximation

Yuyang Wu[1,2], Zhiyang Qi[1,2], Huiming Zheng[1,2], Lvfang Tao[1,2], Wei Gao[1,2,*]
[1]School of Electronic and Computer Engineering, Shenzhen Graduate School,
Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
wuyy1234@stu.pku.edu.cn,{zhiyangqizero,hmzheng1999}@gmail.com,{ltao,gaowei262}@pku.edu.cn

Abstract

Benefit from its capability of learning high-dimensional compact representation from raw data, the auto-encoders are widely used in various tasks of data compression. In particular, for deep image compression, auto-encoders generally take the responsibility of mapping original images to the latent representation to be coded. In this paper, we propose a new framework for deep image compression by devising a loss function for latent optimization, and adopting the differentiable approximation of quantization. In our experiments, both subjective and objective results can confirm the effectiveness of our contributions.

1. Introduction

In recent years, with the development of deep learning technology, researchers have begun to apply the related technologies to image and video compression. Ballé et al. first proposed an end-to-end joint Rate-Distortion optimization for image compression in [1] and integrated the quantization function into the CNN based network. Theis et al. used sub-pixel architecture and designed a new image compression network in [15] to process high-resolution images more efficiently. However, there are problems for CNN-based deep image compression, such as the need to train different models for different compression ratios or the inability to dynamically adjust the size of the input image. To avoid deficiencies of CNN-based auto-encoders, Toderici et al. [16] proposed a deep image compression model based on RNN. Rippel et al. proposed a lightweight and quickly deployable image compression model [14] using GAN, which achieves real-time encoding and decoding.

Auto-encoder has been widely used in the field of deep learning including image reconstruction, data compression [7] and data generation [17, 13, 12]. In the field of deep image compression, auto-encoder is applied to remove image redundant information. At present, most of the optimization goals in the field of deep compression are to reduce the error between the input original image and the output reconstructed image. These errors include image reconstruction error (MSE, 1-MS-SSIM) or perceptual loss (LPIPS) [18]. Among them, the perceptual loss is considered to be closer to the perceptual evaluation of the human eye. However, the latent high-dimensional feature representation learned by auto-encoder also contains a lot of useful information which has been rarely mentioned in the field of deep compression before. In this article, we will use the latent information for optimization and training.

Due to the discrete and non-differentiable nature of quantization, the continuous optimization method like stochastic gradient descent (SGD) cannot be directly used to optimize quantized representation of network outputs. Previous works have proposed many methods to solve this problem. Uniform noise was added to replace quantization in the training process in [2] . Uniform noise can be used to simulate the difference between the quantized value and the real-value, and the gradient can be calculated during the back propagation. Linear approximation of the quantization process was proposed in [8] where normal quantization was kept for the forward propagation, and a linear model was designed to approximate the quantization for the backward propagation. More methods of quantization approximation are proposed and validated in the field of deep compression, such as differentiable soft quantization [6], piecewise polynomial function [10] and so on. In this work, we adopt a

[*] Corresponding author: Wei Gao (gaowei262@pku.edu.cn). This work was supported by Ministry of Science and Technology of China - Science and Technology Innovations 2030 (2019AAA0103501), Natural Science Foundation of China (61801303 and 62031013), Guangdong Basic and Applied Basic Research Foundation (2019A1515012031), Shenzhen Science and Technology Plan Basic Research Project (JCYJ20190808161805519), Open Projects Program of National Laboratory of Pattern Recognition (NLPR) (202000045), and CCF-Tencent Open Fund (RAGR20200114).

piecewise function to replace the discrete quantization during training.

2. Proposed Method

2.1. Overview

Our autoencoder architecture consists of two parts, one is the main network and the other is the hyperprior which is similar to [4]. The main network is shown in Figure 1, which is composed by residual blocks, attention modules and convolutional layers. For the encoder, two types of residuals block are devised. Both of them are composed of two 3x3 convolution kernels, and one of them includes the sub-sampling operation. Meanwhile, we use skip-connections in both sides of the encoder and the decoder to facilitate feature aggregation.

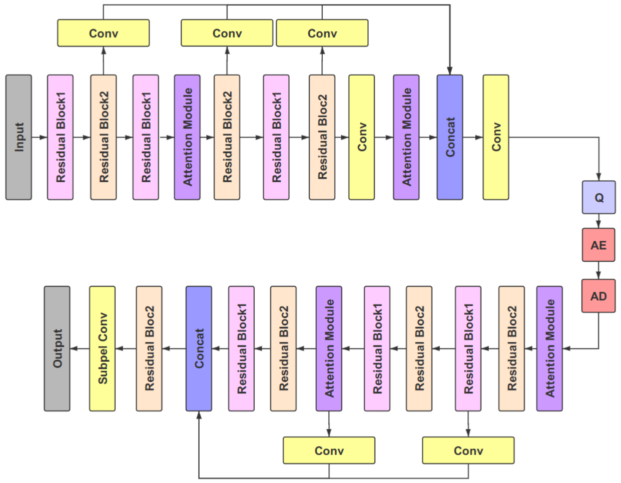


Figure 1. The encoder and decoder of the main network.

The attention module help the model pay more attention to significant regions of the image and improve the coding performance of areas with richer textures. In [19], a residual non-local attention structure is proposed for high-quality image restoration, and we notice that this structure is also used as an attention module in [4] and [5]. In the proposed network, we refer to the design of the attention module mentioned above and make a minor adjustment as follows: we discard the non-local module and keep the residual block. Our attention module is shown in Figure 2.

The hyperprior network is designed to extract side information from the encoded representation y. We use the side information to improve the entropy estimation of the latent representation. In the proposed network, we use sub-pixel convolution and down-sampling to preserve more details. At the end of the hyperprior decoder, there are the Gaussian entropy model and the 3D context model. We adopt the Gaussian modeling for accurate estimation of the entropy parameters. We use 3D context model as introduced in [9]. The input tensors are converted to a 3D tensor and then processed by 3D CNN. 3D context model can make use of feature representations from different channels and facilitate the estimation of Gaussian entropy parameters.

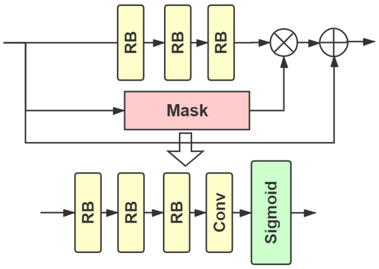


Figure 2. The attention module.

The whole hyperprior network is illustrated in Figure 3 and the entropy estimation module is depicted in Figure 4.

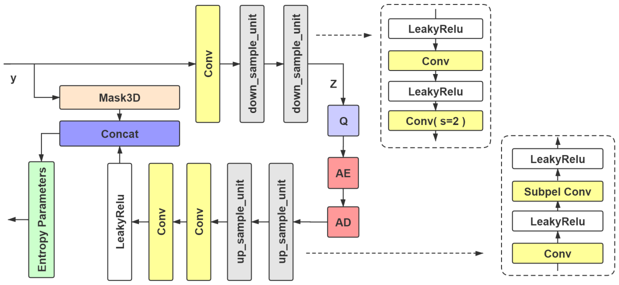


Figure 3. The encoder and decoder of the hyperprior network.

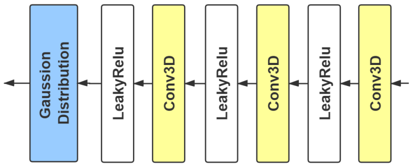


Figure 4. The entropy estimation module.

2.2. Loss Function and Latent Optimization

In the proposed framework, our training and optimization use the structure of auto-encoder and latent representations encoded by the encoder. Since our network adopts the skip-connections structure, the latent representations encoded by the encoder of our network contain rich multi-dimensional features of the image. We hope to make full use of these feature representations that are more efficient than the original image data for our network training and optimization.

For the calculation of the loss of latent variables, we adopt the following calculation process: the original latent variables are passed through the decoder to obtain the reconstructed image and the reconstructed image is input into the encoder again to obtain the reconstructed latent variable.

We define the error between the original latent variable and the reconstructed latent variable as latent loss. The difference between the calculation method of latent loss and image reconstruction loss can be seen in Figure 5.

We hope that our network can preserve latent feature representations as much as possible. With this goal, we use the L2 norm for the calculation of latent loss $\mathcal{L}_l$. The formula is as follows:

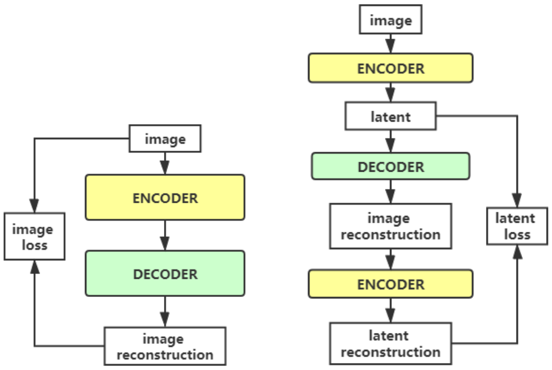
$$\mathcal{L}_l = \|E(x) - E(D(E(x)))\|_2, \tag{1}$$

in which x stands for the input image, E stands for the encoder, D stands for the decoder.

The loss function of the experiment includes four parts, namely perceptual loss, reconstruction loss, entropy coding loss and latent loss. We denote the loss function as follows:

$$\mathcal{L} = \lambda_p \mathcal{L}_p + \mathcal{L}_e + \lambda_l \mathcal{L}_l + \lambda_r \mathcal{L}_r. \tag{2}$$

For the perceptual loss function $\mathcal{L}_p$, we use LPIPS. For the reconstruction loss function $\mathcal{L}_r$ we use MSE and a differentiable implementation of (1-MS-SSIM), and the entropy coding loss $\mathcal{L}_e$ includes two parts of the entropy coding length of the main network and the hyperprior. The latent loss $\mathcal{L}_l$ includes that of the main network and that of the hyperprior part.



(a) (b)

Figure 5. Difference between (a) image reconstruction loss and (b) latent loss.

2.3. Piece-wise Function for Quantization Approximation

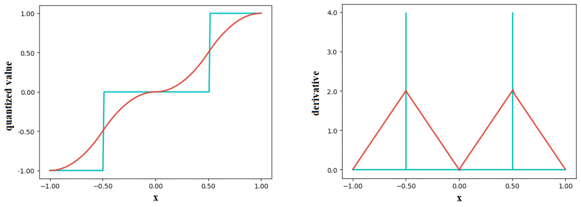
In the back propagation, unlike the previous method of adding random noise, we use a nonlinear piece-wise function to approximate the quantization process so that parameters of the entire network can be updated by gradient descent. We also notice that most values of the latent representation are ranged between -1 and 1. Therefore, we design the piece-wise function for quantization approximation:

$$G(x) = \begin{cases} x + noise, & x < -1, \\ 2 * x^2 + 4 * x + 1, & -1 \le x < -\frac{1}{2}, \\ -2 * x^2, & \frac{1}{2} \le x < 0, \\ 2 * x^2, & 0 \le x < \frac{1}{2}, \\ -2 * x^2 + 4 * x + 1, & \frac{1}{2} \le x < 1, \\ x + noise, & otherwise. \end{cases} \tag{3}$$

Meanwhile, the derivative of this equation is formulated as the following piece-wise linear function:

$$\frac{\partial G(x)}{\partial x} = \begin{cases} 1, & x < -1, \\ 4 * x + 4, & -1 \le x < -\frac{1}{2}, \\ -4 * x, & \frac{1}{2} \le x < 0, \\ 4 * x, & 0 \le x < \frac{1}{2}, \\ -4 * x + 4, & \frac{1}{2} \le x < 1, \\ 1, & otherwise. \end{cases} \tag{4}$$

Random noise can only roughly simulate the quantization process. The approximation with random noise may fail in modeling the exact rounding direction for different sub-intervals. It also cannot accurately simulate the derivative of the quantization function. On the contrary, the piece-wise function can better reflect the trend of data changes in the quantization process. Its triangular derivative is closer to the derivative of the discrete quantization process than other methods [10], as illustrated in Figure 6.



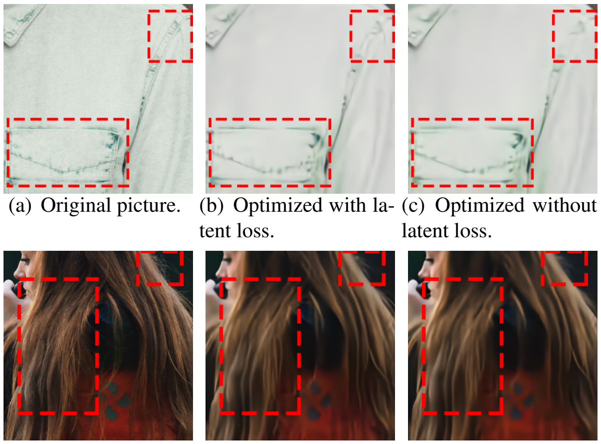
(a) (b)

Figure 6. (a) The discrete quantization function and the piece-wise function (Equation 3) for quantization approximation, (b) The derivative of each function.

3. Experiment

We set the number of channel as 128 for all three different bit-rates, namely 0.075 bpp, 0.15 bpp, and 0.3 bpp. In the training phase, we use 192 × 192 patches randomly cropped from the CLIC 2021 training dataset. We adapt a two-step training process: in the first step, our loss function only includes $\mathcal{L}_e$, $\mathcal{L}_r$ and $\mathcal{L}_p$. In the second step, we add latent loss $\mathcal{L}_l$ into the loss function. The learning rate is set 0.0001 initially and gradient decay is applied by decreasing the learning rate by half every 6 epochs. The batch size is set to 4.

In order to verify the effectiveness of the proposed latent optimization, we conduct an ablation study. Note that during the calculation process, we use piece-wise quantization approximation instead of direct quantization. All parameters are set under training mode. Under the extremely low bit-rate of 0.075 bpp, we select two patches from CLIC 2021 validation dataset for comparison, which are visualized in Figure 7. Although the model optimized with latent loss has lower PSNR (28.279 versus 28.898) scores and lower MS-SSIM (0.942 versus 0.944) scores than the one optimized without latent loss, it provides better perceptual quality.



(a) Original picture. (b) Optimized with latent loss. (c) Optimized without latent loss.

(d) Original picture. (e) Optimized with latent loss. (f) Optimized without latent loss.

Figure 7. Comparison between reconstructed images with the model optimized with (0.073 bpp) and without latent loss (0.078 bpp).

Besides, we also conduct an ablation study for piece-wise function for quantization approximation. We train two models with the same hyper-parameters. One is trained with piece-wise quantization approximation and the other is trained with uniform noise. Both of the models are optimized with latent loss. Result shows that piece-wise quantization approximation can improve PSNR by 0.13 and MSSSIM by 0.0002 with lower bit rate (0.2568 bpp versus 0.2579 bpp).

As shown in Figure 8, we adopt MS-SSIM as the quality metric. Our proposed method is compared with other deep learning based competitive methods for image compression such as Balle's work [2] (bmshj2018_factorized means model with factorized prior and bmshj2018_hyperprior means model with scale hyperprior), Minnen's work [11] (mbt2018_mean means model with mean scale hyperprior and mbt2018 means model with joint autoregressive hierarchical priors) and Cheng's work [4] (cheng2020). At low bit rate (about 0.075 bpp), our proposed method has surpassed Balle's work by a large margin. When compared with Minnen's work , we also achieve comparable results. At middle bit rate (about 0.15 bpp), our proposed method has surpassed Minnen's work and the results are comparable with Cheng's work. At high bit rate (about 0.3 bpp), we achieve excellent performance comparing with other state-of-art models. Our test code is derived from [3]. Finally, Table 1 shows the results using the CLIC 2021 validation dataset. Our submission team's name is “IVPG”.

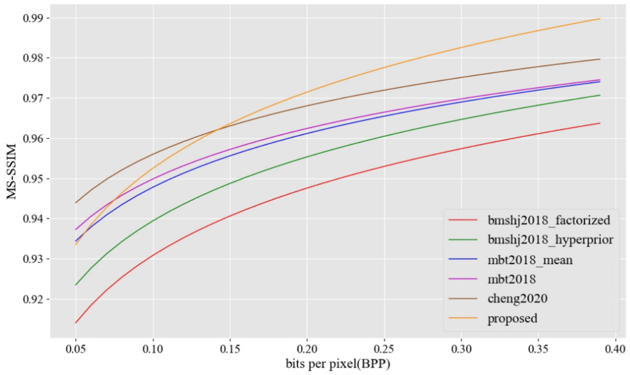


Figure 8. Rate-distortion curves of the proposed method and other competitive methods for image compression (Quality measured with MS-SSIM).

bpp	FID	PSNR	MSSSIM
0.282	192.101	32.987	0.979
0.143	218.328	30.490	0.966
0.073	258.160	28.279	0.942

Table 1. Evaluation results using the CLIC 2021 validation dataset. Note that FID score is not deterministic because of the random crop selection.

4. Conclusion

In this paper, we propose a novel method for deep image compression. Our framework includes latent optimization and quantized approximation function. Making use of the new method, we further improve the quality of reconstructed images with respect to perception. In the experiment section, our ablation study and comparison results prove the effectiveness of our framework.

References

[1] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016. 1

[2] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. 1, 4

[3] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. pages arXiv–2011, 2020. 4

[4] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *2020*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7936–7945. IEEE, 2020. 2, 4

[5] Wei Gao, Lvfang Tao, Linjie Zhou, Dinghao Yang, Xiaoyu Zhang, and Zixuan Guo. Low-rate Image Compression with Super-resolution Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 607–610, Seattle, WA, USA, June 2020. IEEE. 2

[6] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4851–4860. IEEE, 2019. 1

[7] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 1

[8] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3214–3223. IEEE Computer Society, 2018. 1

[9] Haojie Liu, Tong Chen, Qiu Shen, and Zhan Ma. Practical stacked non-local attention modules for image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2

[10] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-Real Net: Enhancing the Performance of 1-bit CNNs With Improved Representational Capability and Advanced Training Algorithm. *arXiv:1808.00278 [cs]*, Sept. 2018. arXiv: 1808.00278. 1, 3

[11] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *NeurIPS*, 2018. 4

[12] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 1

[13] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020. 1

[14] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *International Conference on Machine Learning*, pages 2922–2930. PMLR, 2017. 1

[15] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017. 1

[16] George Toderici, Sean M O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Variable rate image compression with recurrent neural networks. *arXiv preprint arXiv:1511.06085*, 2015. 1

[17] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. It takes (only) two: Adversarial generator-encoder networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1

[19] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 2