UNIVERSAL EFFICIENT VARIABLE-RATE NEURAL IMAGE COMPRESSION

Shanzhi Yin, Chao Li, Youneng Bao, Yongsheng Liang[*] *Fanyang Meng , Wei Liu*

Harbin Institute of Technology, Shenzhen Peng Cheng Laboratory

ABSTRACT

Recently, Learning-based image compression has reached comparable performance with traditional image codecs(such as JPEG, BPG, WebP). However, computational complexity and rate flexibility are still two major challenges for its practical deployment. To tackle these problems, this paper proposes two universal modules named Energy-based Channel Gating(ECG) and Bit-rate Modulator(BM), which can be directly embedded into existing end-to-end image compression models. ECG uses dynamic pruning to reduce FLOPs for more than 50% in convolution layers, and a BM pair can modulate the latent representation to control the bit-rate in a channel-wise manner. By implementing these two modules, existing learning-based image codecs can obtain ability to output arbitrary bit-rate with a single model and reduced computation.

Index Terms— image compression, dynamic pruning, variable-rate

1. INTRODUCTION

Image compression is a fundamental technology in signal processing and computer vision. It reduces the required bits for image transmission and storage while maintains its reconstruction quality as much as possible. In recent years, many learning-based image compression methods have achieved the state-of-the-art performance comparing to traditional image codecs [1–3]. However, there are still some challenges for its practical deployment.

With a predefined trade-off factor, bit-rate and reconstruction quality are fixed for a single trained model. Various requirements needs a potentially large amount of models and corresponding storage budget.To tackle this disadvantage, quantization steps [4], decomposition methods [5, 6]or trade-off factor [7–9] are leveraged to obtrain variable-rate ability. However, these rate-control methods involve many modifications in original architectures and is hard to be adopted by existing models.

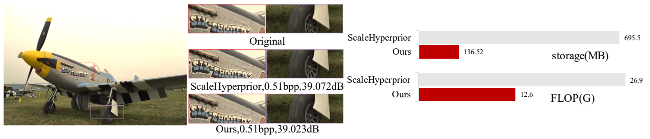

Fig. 1. Visualization of reconstructed images from Kodak dataset and corresponding storage&computation budget

Due to their complex network architectures, computational cost in learning-based compression models is relatively high. In addition, advanced neural network modules are introduced to further boost their performance [2–4, 10, 11]. Heavy computational burden is unfriendly to circumstances like portable electrical devices or edge computing. To tackle this disadvantage, Johnston et al. [12]proposed a comprehensive evaluation on compression models, combing bit-rate, distortion and computation efficiency. Guo et al. [13]design a complexity adaptive module to replace the decoder. However, the computational complexity levels are predefined in these models, arbitrary complexity is not achieved.

Under real world circumstances, the computation capacity of a given device is usually fixed, but the available transmission bandwidth and reconstruction requirements are varies. To deal with such situation and corresponding challenges of learning-based image compression, this paper proposes a universal variable-rate efficient method for neural image compression(NIC).

Fig.2 illustrates the architecture of the model. We achieve bit-rate flexibility and reduce computational complexity by embedding two portable modules i.e. Energy-based channel gating module and Bit-rate modulator into the existing NIC networks. Energy-based channel gating modules are inserted before convolution layers to introduce sparsity, it uses simple 1-D convolution to generate channel-wise threshold and implements dynamic pruning on input feature map. By tuning the learnable adjustment vector, the compression model can achieve arbitrary computational complexity in convolution operations. Bit-rate modulator regards trade-off factor as the input to generate channel-wise multiplier with only two full-connected layers in a plug-in manner. By channel-wisely product the latent representation of pretrained fixed-rate model with the output of the bit-rate modulator, it can modulate the compression bit-rate outside the entropy coding

[*] Corresponding author, this work was supported by National Natural Science Foundation of China (Grant No.62031013, Grant No.61871154), Youth Program of National Natural Science Foundation of China (Grant No.61906103, Grant No.61906124), Basic and Applied Basic Research Fund of Guangdong Province (2019A1515011307).

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

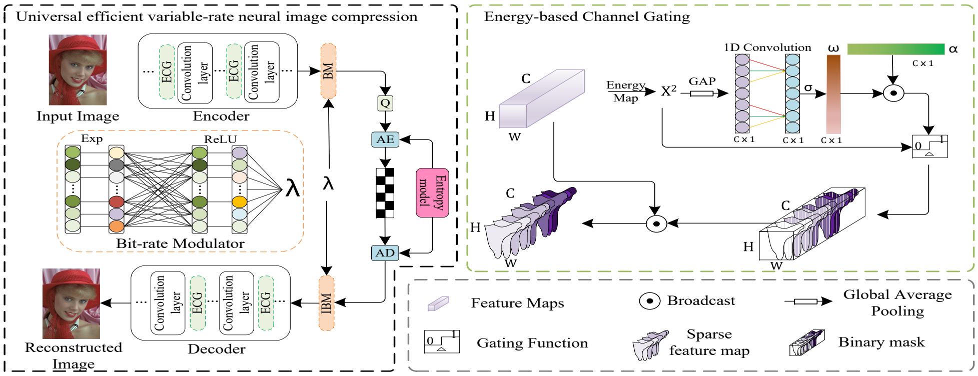

Fig. 2. Architecture of universal variable-rate efficient compression model.

process and generate arbitrary bit-rate. To end-to-end optimize the proposed compression model, we formulate a comprehensive optimization problem including bit-rate, distortion and computational complexity. We implement our method on three classical NIC models and conduct comprehensive experiments to prove the effectiveness of our method.Visual results and efficiency comparison are shown in Fig.1. There is no obvious qualitative degredation on the reconstructed image, while the storage and computation saving is quite impressive.

In this paper, our main contribution is to present a method to built variable-rate and low-computation image compression based on existing models. We design a dynamic feature map pruning module that can reduce the FLOP of NIC models up to 3× and we also design a the simple but effective plug-in bit-rate modulator that can obtain arbitrary bit-rate of NIC models. We use modules mentioned-above to form universal efficient variable-rate NIC and generalize it to three existing models. By solving a comprehensive optimization problem, experiments show that our method can achieve continuous bit-rate flexibility in a single model with less than half of the original computation.

2. PROPOSED METHOD

2.1. Energy-based channel gating

Sparsity is an effective way to reduce computational complexity. In convolution operations, inputs of different intensities have different influence on the results, sparsity can be introduced by pruning the inputs with less significance, which will be judged by energy-based channel gating module(ECG). Inspired by [14], ECG uses learnable dynamic feature map pruning with channel-wise threshold. The energy of the input is evaluated within each channel and between its neighbour channels [15] to obtain threshold for every channel.

As shown in Fig.2, given the input x, the energy map x^2 is first adaptive average pooled(GAP) to obtain the intensity information of a single channel, then 1-D convolution with adaptive kernel size is conducted following [15] to integrate the intensity information of neighbour channels, the convolution result is activated by Sigmoid function to form a importance vector ω:

$$\omega = \sigma\{C1[P(x^2)]\} \tag{1}$$

in which P is adaptive average pooling, $C1$ is 1-D convolution, σ is Sigmoid activation. The importance vector then dot-products a learnable adjustment vector α to form a learnable threshold $th = \omega \odot \alpha$. A binary mask is generated by gating the energy map x^2 through this channel-wise threshold th, the gating function can be described as:

$$gt(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } otherwise, \end{cases} \tag{2}$$

For this gating function is not differentiable, Sigmoid function $\sigma(x) = \frac{1}{1+e^{-\epsilon x}}$ is used to replace it during the training process [14]. The final output of ECG x_{mask} is the original input x dot-product the binary mask:

$$x_{mask} = x \odot gt(x^2 - \omega \odot \alpha) \tag{3}$$

x_{mask} is the sparse version of the original input x and the computation can be reduced with its sparsity.

2.2. Bit-rate modulator

Following the work in [9, 16], we try to affect the R-D performance of the compression model through the quantization process. Instead of learning a gain matrix [16], we use bit-rate modulator shown in Fig.2 to implement the bottleneck scaling in a channel-wise and plug-in manner.

To obtain rate-flexibility in a convenience but effective way, we design the bit-rate modulator(BM) to be a simple and portable module with two full-connected layer only [7]. It maps a trade-off factor λ into a vector which has the same channel number as the entropy coding process:

$$f_1(\lambda) = ReLU(\mathbf{H^{(1)}}\lambda + \mathbf{b^{(1)}}) \tag{4}$$

$$bm(\lambda) = exp(\mathbf{H^{(2)}}f_1 + \mathbf{b^{(2)}}) \tag{5}$$

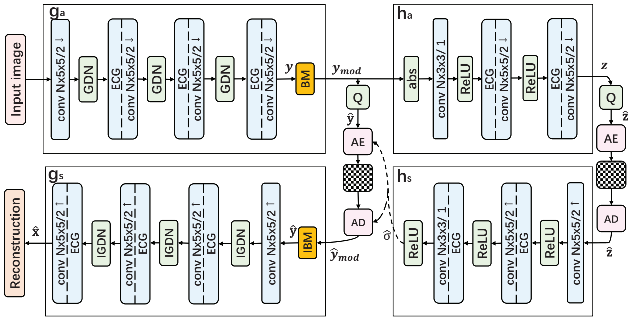

Fig. 3. Implementation on ScaleHyperprior model

in which the exponential mapping before the final output is to maintain the positive value of the vector and expand the dynamic range of the modulator.

During the compression process, the latent representation y channel-wisely products the $bm(\lambda)$ after the encoder to obtain the modulated latent y_{mod}. While before the decoder, y_{mod} products the inverse module $ibm(\lambda)$ to restore the original feature map.

2.3. Implementation on existing compression models

To illustrate the effectiveness of our method, we implement it on several classical neural image compression models [1, 17] denoted as ScaleHyperprior model, MeanscaleHyperprior model and JointAutoregressive model respectively.

In this section, we use ScaleHyperprior model as an example to show the implementation details and optimization strategies of our method.

As shown in Fig.3, we embedded totally 10 ECG modules with convolution layers to reduce the computational complexity, while 4 convolution layers remain unchanged. These 4 layers receive the input image or feature map of main codecs and hyper codecs respectively. Keeping these layers can maintain the receptive ability of the codecs to avoid dramatic performance degradation. Bit-modulator is inserted as the last layer of the main encoder and the first layer of the main decoder to modulate the feature map for ScaleHyperprior entropy model. The feature map of hyper codecs remains unchanged, for the bit-rate produced by it is relatively low [17]and the predicted scales are significant for keeping the ScaleHyperprior model as accurate as possible.

After adopting our method into the original ScaleHyperprior model, it needs to be trained in an end-to-end manner. Similar to most neural image compression model, two loss terms rate R and distortion D need to be optimized. As the actual arithmetic coding is bypassed [18], the rate term is given by the cross entropy of the estimated distribution of y and the its actual distribution:

$$R(\hat{y};\theta,\phi,\xi,\lambda) = \mathbb{E}_{\hat{y}\sim p_y}\{log_2 q_y[Q(y \odot bm(\lambda))]\} \tag{6}$$

in which θ , ϕ ,ξ denote the parameter of codecs, ECG Modules, BM modules respectively and λ is the trade-off factors as the input of the BM module, Q represents the quantizaiton process, p and q are actual and predicted distribution of image data respectively. In our work, the distortion metric is the mean square error measured on the test set:

$$D(x,\hat{x};\theta,\phi,\xi,\lambda) = \mathbb{E}_{x\sim p_x}[||x-\hat{x}||^2] \tag{7}$$

The R and D loss term are weighted-summed by a trade-off factor λ. When we try to obtain the rate flexibility through multiple R-D trade-offs, the variable-rate compression optimization problem can be form as:

$$\underset{\theta,\phi,\xi,\lambda}{\operatorname{argmin}} \sum_{\lambda\in\Lambda}[R(\hat{y};\theta,\phi,\xi,\lambda) + \lambda D(x,\hat{x};\theta,\phi,\xi,\lambda)] \tag{8}$$

in which Λ is the set of all possible values of λ, i.e $\Lambda = \{\lambda_0, \lambda_1 \cdots \lambda_n\}$

Unlike traditional neural image compression, our method introduce sparsity through learnable ECGs which need to be optimized during the training process as well. In ECG, the learnable adjustment vector α affect the final gating threshold th, larger the α is, higher the final threshold on each channels will be, and the output feature map of ECG will be sparser. During the training process, we set a relatively high target value α_t for the learnable adjustment vector and take the mean square error of present α and α_t as an additional loss term. With this loss term being optimized, we can ensure that, the sparsity of the ECG output can gradually reach an expected level and overall optimization under such multiple trade-offs can be achieved. The final optimization problem can be formulated as:

$$\underset{\theta,\phi,\xi,\lambda}{\operatorname{argmin}} \sum_{\lambda\in\Lambda}[R + \lambda D + \gamma\sum_{i=1}^{n}(\alpha_n - \alpha_t)^2] \tag{9}$$

in which α_i represents the ith ECG module, γ is the trade-off factor for computation effeciency and R, D are defined as equation(6)and(7). When tuning the γ, arbitrary computational comlexity can be achieved, so that the model can be adopted on various computational environments.

3. EXPERIMENTS

The NIC_dataset [1] is used for training, which contains 607,714 256x256 patches cropped from the 1,600 original images and the 2x, 4x down-sampled versions using bicubic interpolation. We adapted CompressAI [19] re-implementation of neural compression models and followed their training parameters settings and training strategies. All trained models are evaluated on Kodak dataset.

Table 1. The PSNR and FLOP reduction of models with ECG module

Model	Performance	Quality							
		1	2	3	4	5	6	7	8
ScaleHyperprior	PSNR drop(%)	0	0.346	0.228	0.269	0.336	0	0	0.216
	FLOP reduction	**2.54**×	**2.86**×	**2.60**×	**2.54**×	**2.54**×	**2.07**×	**2.14**×	**2.03**×
MeanscaleHyperprior	PSNR drop(%)	0.39	0.22	0.77	0.69	0.37	0.61	0.71	0.78
	FLOP reduction	**2.34**×	**2.50**×	**2.56** ×	**2.68**×	**2.33**×	**2.12**×	**2.12**×	**2.24**×
JointAutoregressive	PSNR drop(%)	0.207	0.335	0.437	0.807	0.465	0.150	0.354	0.553
	FLOP reduction	**2.43**×	**2.67**×	**2.48**×	**2.23**×	**2.29**×	**2.02**×	**2.06**×	**2.02**×

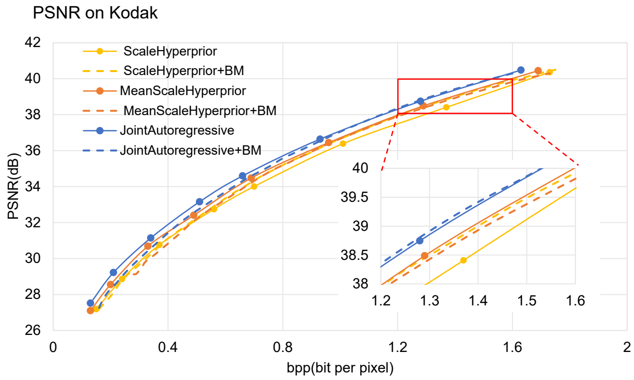

Fig. 4. Bit-rate flexibility on three classical neural image compression model

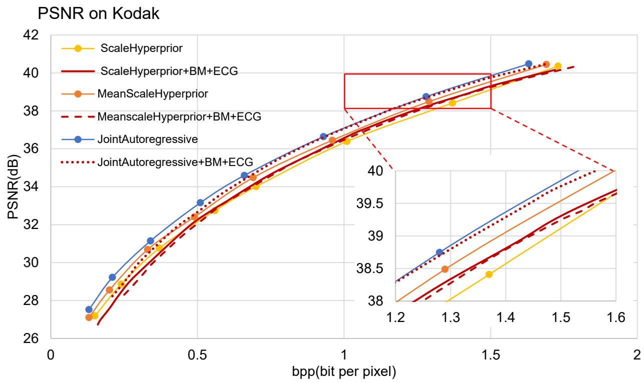

Fig. 5. Performance of universal variable-rate efficient models

3.1. Ablation study

We first evaluate the effectiveness of our ECG module by embedding it into ScaleHyperprior, MeanscaleHyperprior and JointAutoregressive models.γ and α_t was empirically set to be 0.0001. We follow the 8 different trade-off factors in [19] for training the original models. When embedded with ECGs, we observe that the bit-rate of efficient models fluctuate around their original bit-rate. For fair comparison, we fine-tunned the trade-off factor for efficient models so that their bit-rates are the same (accurate to two decimal places) as their original counterparts. PSNR and FLOP reduction are shown in Table.1. We can see that the FLOP reduction of more than 2× can be achieved in three neural image compression models with very slight PSNR degradation around 0.5% and no more than 1%.

Then we evaluate the effectiveness of our BM module. We use pre-trained highest quality original models and fine-tune them with BM module. The R-D performance is shown in Fig.4. We can see that continuous rate flexibility can be achieved in three architecture with only one trained model. Compared with previous 8 models files of 695.5MB,1125.6MB, 1916MB, one single model only takes 135MB,201MB and 311MB storage, saving up to 83.8%.

3.2. R-D and efficiency performance

We implement both ECG and BM modules to form variable-rate and low- computational complexity efficient compression model on above-mentioned three architectures. We use ECG embedded models with highest quality as pre-trained models and fine-tuned them with BM module. The R-D performance is shown in Fig.5. We can see that three variable-rate efficient models are able to approach the performances of the original models. The models can achieve sparsity around **0.5** in convolution operations and achieve storage saving of **80.42%**,**82.04%** and **83.07%** respectively.

4. CONCLUSION

We proposed two simple modules, energy-based channel gating module and bit-rate modulator, which can be adopted into most of the existing models. Continuous bit-rate flexibility can be achieved with reduced FLOP. In our experiment, existing compression models can generate arbitrary bit-rate and reconstruction quality in a single trained model with 80% storage saving and 2-3×FLOP saving, which makes it easier for their pratical deployment. Potential future work includes better joint training strategies for overall performance and adaptive computational cost in single model.

[1] The NIC_Dataset can be accessed at https://www.bitahub.com/dataset, which is opened as a public dataset.

5. REFERENCES

[1] David Minnen, Johannes Ballé, and George Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems*, sep 2018, vol. 2018-Decem, pp. 10771–10780.

[2] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, jan 2020, pp. 7936–7945.

[3] Tong Chen, Haojie Liu, Zhan Ma, Qiu Shen, Xun Cao, and Yao Wang, "End-to-End Learnt Image Compression via Non-Local Attention Optimization and Improved Context Modeling," *IEEE Transactions on Image Processing*, vol. 30, pp. 3179–3191, 2021.

[4] Haichuan Ma, Dong Liu, Ning Yan, Houqiang Li, and Feng Wu, "End-to-End Optimized Versatile Image Compression With Wavelet-Like Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8828, no. c, pp. 1–1, 2020.

[5] Chunlei Cai, Li Chen, Xiaoyun Zhang, and Zhiyong Gao, "Efficient variable rate image compression with multi-scale decomposition network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3687–3700, 2019.

[6] Zhizheng Zhang, Zhibo Chen, Jianxin Lin, and Weiping Li, "Learned scalable image compression with bidirectional context disentanglement network," *Proceedings - IEEE International Conference on Multimedia and Expo*, vol. 2019-July, pp. 1438–1443, 2019.

[7] Fei Yang, Luis Herranz, Joost van de Weijer, José A. Iglesias Guitián, Antonio M. López, and Mikhail G. Mozerov, "Variable rate deep image compression with modulated autoencoder," *IEEE Signal Processing Letters*, vol. 27, pp. 331–335, 2020.

[8] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee, "Variable rate deep image compression with a conditional autoencoder," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 3146–3154, 2019.

[9] Chuanmin Jia, Ziqing Ge, Shanshe Wang, Siwei Ma, and Wen Gao, "Rate distortion characteristic modeling for neural image compression," 2021.

[10] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli, "Density modeling of images using a generalized normalization transformation," in *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, nov 2016.

[11] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár, "Lossy image compression with compressive autoencoders," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–19, 2017.

[12] Nick Johnston, Elad Eban, Ariel Gordon, and Johannes Ballé, "Computationally efficient neural image compression," 2019.

[13] Jinyang Guo, Dong Xu, and Guo Lu, "Cbanet: Towards complexity and bitrate adaptive deep image compression using a single network," 2021.

[14] Weizhe Hua, Yuan Zhou, Christopher M De Sa, Zhiru Zhang, and G. Edward Suh, "Channel gating neural networks," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[15] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11531–11539.

[16] Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, and Bo Bai, "Asymmetric gained deep image compression with continuous rate adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10532–10541.

[17] 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.

[18] J Ballé, V Laparra, and E P Simoncelli, "End-to-end optimized image compression," in *Int'l Conf on Learning Representations (ICLR)*, Toulon, France, April 2017, Available at http://arxiv.org/abs/1611.01704.

[19] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja, "Compressai: a pytorch library and evaluation platform for end-to-end compression research," *arXiv preprint arXiv:2011.03029*, 2020.