

Low Bitrate Image Compression with Discretized Gaussian Mixture Likelihoods

Zhengxue Cheng¹, Heming Sun^{2,3*}, Jiro Katto¹

¹ Department of Computer Science and Communication Engineering, Waseda University, Tokyo, Japan

² Waseda Research Institute for Science and Engineering, Tokyo, Japan ³ JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama, Japan

{zxcheng@asagi., hemingsun@aoni., katto@}waseda.jp

Abstract

In this paper, we provide a detailed description on our submitted method Kattolab to Workshop and Challenge on Learned Image Compression (CLIC) 2020. Our method mainly incorporates discretized Gaussian Mixture Likelihoods to previous state-of-the-art learned compression algorithms. Besides, we also describes the acceleration strategies and bit optimization with the low-rate constraint. Experimental results have demonstrated that our approach Kattolab achieves 0.9761 in terms of MS-SSIM at the rate constraint of 0.15 bpp during the validation phase.

1. Introduction

Image compression is a fundamental research topic in the field of image signal processing for many decades to achieve efficient image transmission and storage. Traditional image compression standards have been developed for a long time, such as JPEG [1], JPEG2000 [2], HEVC/H.265 [3] and ongoing Versatile Video Coding (VVC) [4]. Typically they rely on hand-crafted creativity to present a fixed encoder/decoder (codec) block diagrams. They use predefined transform matrix, intra prediction, quantization, arithmetic coders and various post filters to reduce spatial redundancy and improve the coding efficiency. The standardization of a traditional codec has historically spanned many years. Along with the fast development of new image formats and the proliferation of high-resolution mobile devices, existing image compression standards are not expected to be an optimal and general solution for all kinds of image contents.

Recently, various approaches has been investigated for end-to-end learned image compression such as early-stage differentiable quantization for end-to-end training [5, 6, 7], recurrent neural networks-based methods [8, 9, 10], some generative models [11, 12, 13], content-weighted strategy [14], conditional probability models [15], de-correlating

different channels using principle component analysis [16, 17], or energy compaction based approach [18, 19]. The most representative approaches are adaptive entropy models for rate estimation, including a hyperprior [20] and its variants [21, 22, 23] to achieve state-of-the-art performance. Specifically, the work [20] proposed a scale hyperprior, by encoding additional bits to build the entropy model for latent codes. The work [21] jointly combined an autoregressive mask convolution and a mean-scale hyperprior to make entropy model more accurate. The work [22] proposed a quite similar idea by considering two types of contexts, bit-consuming contexts (i.e., hyperprior) and bit-free contexts (i.e., mask convolution model) to realize a context-adaptive entropy model. The work [23] further extended the single Gaussian model to Gaussian mixture likelihoods to further improve the accuracy of entropy models. Our method is based on these recent techniques and apply them to low bitrate image compression.

In this paper, we present a detailed description on our submitted method to Workshop and Challenge on Learned Image Compression (CLIC) 2020. The network architecture combines recent techniques, including deep residual blocks, subpixel convolution and attention modules. The entropy model utilizes discretized Gaussian mixture likelihoods to achieve more accurate entropy model than single Gaussian model. Besides, we also apply some acceleration strategies and bit optimization to meet the limit of 10 hours decoding time and 0.15 *bpp* rate constraint in the CLIC low-rate track. Experimental results have demonstrated that our approach *Kattolab* achieves 0.9761 in terms of MS-SSIM at the rate constraint of 0.15 *bpp* during the validation phase.

2. Learned Low Bitrate Image Compression

2.1. Network Architecture

The network architecture we used is shown in Fig. 1, referring to [23]. Compared to the work [20], the backbone network architecture has been improved by using residual blocks, subpixel convolution and attention modules. Based on the observations of [24], deep residual blocks can

*Corresponding author.

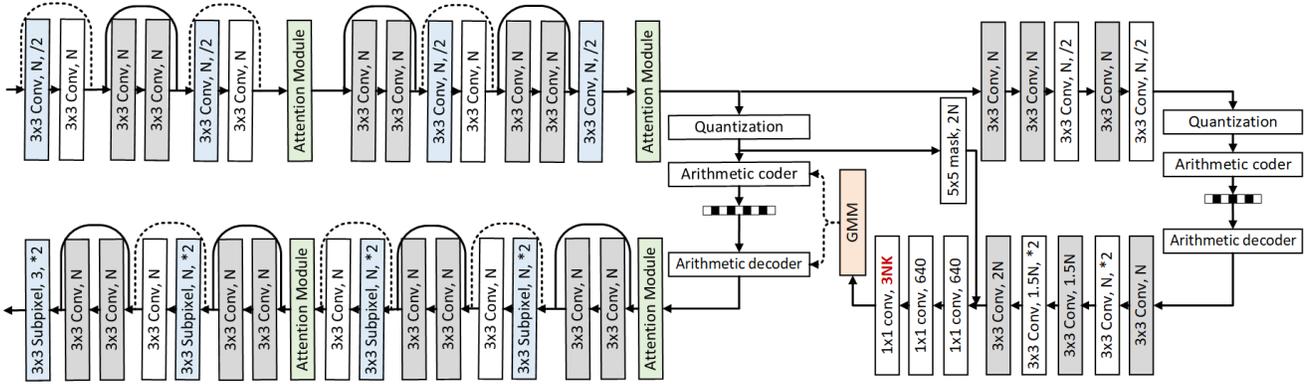


Figure 1: The overall network architecture we used in *Kattolab*, where the green blocks denote attention modules, the orange block denotes Gaussian mixture model (K denotes the number of mixtures, N denotes the number of filters) and the blue blocks represent the downsampling and upsampling units, implemented by stride-2 convolutions and sub-pixel convolutions.

achieve more larger and effective receptive field than 5×5 filters, therefore, we used the residual block implemented by stacked 3×3 filters to replace 5×5 filters as downsampling units at the encoder side and mirrored them at the decoder. GDN and IGDN [25] are only followed by the convolution with the stride of 2, and ReLU is used after other convolution filters. Besides, [24] also found subpixel convolution could maintain more details compared to transposed convolution to improve the quality of reconstructed images, so we used subpixel convolution to upsample the feature sizes at the decoder side.

Attention module can increase the values of responses which are originally large and decrease the values of response which are originally small, thus it forces models to pay more attention to complex regions instead of simple regions to improve the coding performance, indicated by [26, 27], although the structures of attention modules are slightly different as shown in Fig. 2. By experiments, we find non-local block (NLB), proposed by [28] and used in [26] is time-consuming for training and also memory-consuming when the resolution of input image is very large during inference. The work [27] used a variant of attention module as Fig. 2(b) by removing NLBs, but introduced a pair of downscale and upscale convolution in the attention branch motivated from [29]. The key point is to grasp information with larger receptive field size and large-stride convolution can increase receptive field to obtain more sophisticated attention map and capture long-range dependencies for image restoration task. Because our network for image compression already used deep residual blocks to capture large enough receptive field, so we used a more simplified version as Fig. 2(c). Different from [26, 27], we also modified the residual block in attention modules with $1 \times 1 \times \frac{N}{2} - 3 \times 3 \times \frac{N}{2} - 1 \times 1 \times N$ to replace $3 \times 3 \times N - 3 \times 3 \times N$ which they used, to avoid too much overhead of increasing number of param-

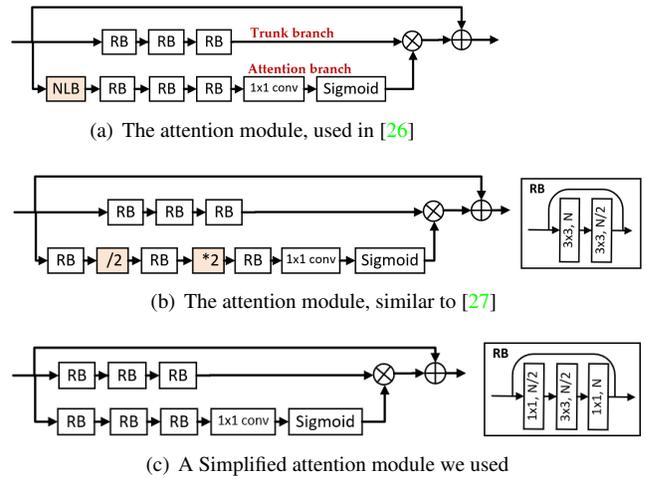


Figure 2: Different attention modules.

eters. Then we insert our simplified attention module into encoder-decoder network as Fig. 1.

2.2. Discretized Gaussian Mixture Model

Following the work [23], we utilize discretized Gaussian mixture likelihoods to replace single Gaussian model. The motivation is to consider more flexible parameterized distributions to achieve arbitrary likelihoods, to fully utilize the contexts and information from neighboring elements and additional bits \hat{z} . The Gaussian mixture model is formulated by

$$p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}) \sim \sum_{k=1}^K w^{(k)} \mathcal{N}(\mu^{(k)}, \sigma^{2(k)}) \quad (1)$$

where \hat{y} is discrete-valued after quantization. The reason why we did not use Logistic mixture likelihoods is that

Gaussian achieves slightly better performance than logistic [21]. Then the entropy model in end-to-end learned image compression is calculated as

$$\begin{aligned}
 p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}) &= \prod_i p_{\hat{y}_i|\hat{z}}(\hat{y}_i|\hat{z}) \\
 p_{\hat{y}_i|\hat{z}}(\hat{y}_i|\hat{z}) &= \left(\sum_{k=1}^K w_i^{(k)} \mathcal{N}(\mu_i^{(k)}, \sigma_i^{2(k)}) * \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right) \right)(\hat{y}_i) \\
 &= \sum_{k=1}^K w_i^{(k)} \left(c^{(k)}\left(\hat{y}_i + \frac{1}{2}\right) - c^{(k)}\left(\hat{y}_i - \frac{1}{2}\right) \right)
 \end{aligned} \tag{2}$$

where i specifies the location in feature maps, and k denotes the index of mixtures. Each mixture is characterized by a Gaussian distribution with 3 parameters, i.e. weights $w_i^{(k)}$, means $\mu_i^{(k)}$ and variances $\sigma_i^{2(k)}$ for each element \hat{y}_i and weights are normalized by passing through a softmax layer. $c^{(k)}$ is the cumulative distribution function for each mixture. The range of \hat{y} is automatically learned and unknown ahead of time. To achieve stable training, we clip the range of \hat{y} to $[-255, 256]$ because empirically \hat{y} would not exceed this range. For the edge case of -255 , replace $c(\hat{y}_i - \frac{1}{2})$ by zero, i.e. $c(-\infty) = 0$. For the edge case of 256 , replace $c(\hat{y}_i + \frac{1}{2})$ by one, i.e. $c(+\infty) = 1$. It provides a numerically stable implementation for training.

3. Implementation Details and Results

For training, we used a subset of OpenImage database [31] and CLIC train dataset [32]. To train our image compression models, the model was optimized using Adam [33] with a batch size of 8. N is set as 128 for low bitrate models. The learning rate was maintained at a fixed value of 1×10^{-4} during the training process, and was reduced to 1×10^{-5} for the last $80k$ iterations. Each model was trained to a total of 1.8×10^6 iterations for each λ to achieve stable performance.

We optimized our models using MS-SSIM quality metrics [34] to achieve better visual quality and distortion term is defined by $\mathcal{D}(\mathbf{x}, \hat{\mathbf{x}}) = 1 - \text{MS-SSIM}(\mathbf{x}, \hat{\mathbf{x}})$, where the weights in multi-scale SSIM is defined as the default values [0.0448, 0.2856, 0.3001, 0.2363, 0.1333]. Finally, the loss function is defined as

$$\begin{aligned}
 \mathcal{L} &= \mathcal{R}(\hat{y}) + \mathcal{R}(\hat{z}) + \lambda \cdot \mathcal{D}(\mathbf{x}, \hat{\mathbf{x}}) \\
 &= \mathbb{E}[-\log_2(p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}))] + \mathbb{E}[-\log_2(p_{\hat{z}|\psi}(\hat{z}|\psi))] \\
 &\quad + \lambda \cdot \mathcal{D}(\mathbf{x}, \hat{\mathbf{x}})
 \end{aligned} \tag{3}$$

3.1. Acceleration strategy

To make the autoregressive model faster during the decoding, we apply two acceleration strategies referring to [30]. The first strategy is to use 5×5 window to feed

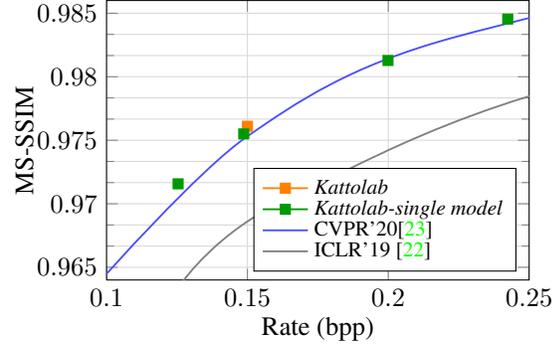


Figure 3: Performance Comparison on CLIC Validation.

in the context model when decoding. The mask convolution needs sequence decoding, while each time only 5×5 centered at a specific point is needed to update the value of \hat{y} at this point, instead of feeding the whole size of \hat{y} to the network. The second strategy is to add some flags to denote all-zero channels. For 0.15bpp, we have found many channels are quantized to all zeros. Therefore, we can skip the arithmetic coding for these all-zero channels to save time. The overhead bit is only N bits, and in our case N is equal to 128, so only a total of 16 bytes per image is required.

3.2. Bit optimization with the rate constraint

To reach the rate constraint of 0.15bpp, we have trained four models with λ in the set I of $\{4.5, 6, 10, 14\}$ to increase the flexibility. The results with single model are shown in Table 1.

Table 1: Results on CLIC validation dataset [32].

λ	MS-SSIM	Rate (bpp)
4.5	0.9716	0.1254
6	0.9755	0.1487
10	0.9813	0.1999
14	0.9845	0.2424

We formulate this problem as a multiple-choice knapsack problem, and solved it by dynamic programming.

$$\max_{\lambda \in I} \sum_i^N \text{MS-SSIM} \quad s.t. \quad \sum_i^N R_{i,\lambda} \leq R_{\text{Thre}}. \tag{4}$$

After bit allocation, MS-SSIM reaches 0.9761 at the rate of 0.15bpp. Because our submitted method is mainly based on [23], so we also list the RD curve comparisons as Fig. 3. Result of *Kattolab* is equal to the original results of [23].

4. Conclusion

In this paper, we have described our method *Kattolab* for challenge on learned image compression (CLIC) 2020,

which includes the network architecture, Gaussian mixture model, acceleration strategy and implementation details. Results have shown our approaches achieve 0.9761 of MS-SSIM at the rate of 0.15 *bpp* during the validation phase.

5. Acknowledgement

This work was supported in part by JST, PRESTO Grant Number JPMJPR19M5, Japan and Japan Society for the Promotion of Science (JSPS) Research Fellowship Grant Number 201914620.

References

- [1] G. K Wallace, "The JPEG still picture compression standard", IEEE Trans. on Consumer Electronics, vol. 38, no. 1, pp. 43-59, Feb. 1991. [1](#)
- [2] Majid Rabbani, Rajan Joshi, "An overview of the JPEG2000 still image compression standard", ELSEVIER Signal Processing: Image Communication, vol. 17, no. 1, pp. 3-48, Jan. 2002. [1](#)
- [3] G. J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard", IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649-1668, Dec. 2012. [1](#)
- [4] G. J. Sullivan and J. R. Ohm, "Versatile video coding Towards the next generation of video compression", Picture Coding Symposium, Jun. 2018. [1](#)
- [5] Lucas Theis, Wenzhe Shi, Andrew Cunningham and Ferenc Huszar, "Lossy Image Compression with Compressive Autoencoders", Intl. Conf. on Learning Representations (ICLR), pp. 1-19, April 24-26, 2017. [1](#)
- [6] J. Ballé, Valero Laparra, Eero P. Simoncelli, "End-to-End Optimized Image Compression", Intl. Conf. on Learning Representations (ICLR), pp. 1-27, April 24-26, 2017. [1](#)
- [7] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, L. V. Gool, "Soft-to-Hard Vector Quantization for End-to-End Learning Compressible Representations", Neural Information Processing Systems (NIPS) 2017, arXiv:1704.00648v2. [1](#)
- [8] G. Toderici, S. M.O'Malley, S. J. Hwang, et al., "Variable rate image compression with recurrent neural networks", arXiv: 1511.06085, 2015. [1](#)
- [9] G. Toderici, D. Vincent, N. Johnson, et al., "Full Resolution Image Compression with Recurrent Neural Networks", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, July 21-26, 2017. [1](#)
- [10] Nick Johnson, Damien Vincent, David Minnen, et al., "Improved Lossy Image Compression with Priming and Spatially Adaptive Bit Rates for Recurrent Networks", arXiv:1703.10114, pp. 1-9, March 2017. [1](#)
- [11] Ripple Oren, L. Bourdev, "Real Time Adaptive Image Compression", Proc. of Machine Learning Research, Vol. 70, pp. 2922-2930, 2017. [1](#)
- [12] S. Santurkar, D. Budden, N. Shavit, "Generative Compression", Picture Coding Symposium, June 24-27, 2018. [1](#)
- [13] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative Adversarial Networks for Extreme Learned Image Compression", arXiv:1804.02958. [1](#)
- [14] M. Li, W. Zuo, S. Gu, D. Zhao, D. Zhang, "Learning Convolutional Networks for Content-weighted Image Compression", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), June 17-22, 2018. [1](#)
- [15] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, L. V. Gool, "Conditional Probability Models for Deep Image Compression", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), June 17-22, 2018. [1](#)
- [16] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, "Deep Convolutional AutoEncoder-based Lossy Image Compression", Picture Coding Symposium, pp. 1-5, June 24-27, 2018. [1](#)
- [17] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, "Performance Comparison of Convolutional AutoEncoders, Generative Adversarial Networks and Super-Resolution for Image Compression", CVPR Workshop and Challenge on Learned Image Compression (CLIC), pp. 1-4, June 17-22, 2018. [1](#)
- [18] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, "Energy Compaction-Based Image Compression Using Convolutional AutoEncoder", IEEE Transactions on Multimedia, vol.22, no. 4, April 2020. [1](#)
- [19] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, "Learning Image and Video Compression through Spatial-Temporal Energy Compaction", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019, arXiv.1906.09683 [1](#)
- [20] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, N. Johnston, "Variational Image Compression with a Hyperprior", Intl. Conf. on Learning Representations (ICLR), 2018. [1](#)
- [21] D. Minnen, J. Ballé, G. Toderici, "Joint Autoregressive and Hierarchical Priors for Learned Image Compression", NeurIPS 2018, arXiv.1809.02736. [1, 3](#)
- [22] J. Lee, S. Cho, S-K Beack, "Context-Adaptive Entropy Model for End-to-end optimized Image Compression", Intl. Conf. on Learning Representations (ICLR) 2019. [1, 3](#)
- [23] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, "Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2020, arXiv.2001.01568. [1, 2, 3](#)
- [24] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, "Deep Residual Learning for Image Compression", CVPR Workshop, pp. 1-4, June 16-20, 2019. [1, 2](#)
- [25] J. Ballé, "Efficient Nonlinear Transforms for Lossy Image Compression", Picture Coding Symposium, 2018. [2](#)
- [26] H. Liu, T. Chen, P. Guo, Q. Shen, X. Cao, Y. Wang, Z. Ma, "Non-local Attention Optimized Deep Image Compression", arXiv.1904.09757. [2](#)
- [27] L. Zhou, Z. Sun, X. Wu, J. Wu, "End-to-end Optimized Image Compression with Attention Mechanism", CVPRW 2019. [2](#)
- [28] X. Wang, R. Girshick, A. Gupta, K. He, "Non-local Neural Networks", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 7794-7803, 2018. [2](#)
- [29] Y. Zhang, K. Li, K. B. Zhong, Y. Fu, "Residual non-local attention networks for image restoration", Intl. Conf. on Learning Representations (ICLR) 2019. [2](#)
- [30] J. Zhou, S. Wen, A. Nakagawa, K. Kazui, Z. Tan, "Multi-scale and Context-adaptive Entropy Model for Image Compression", CVPR workshop 2019. [3](#)
- [31] OpenImages, dataset available from <https://storage.googleapis.com/openimages/web/download.html>. [3](#)
- [32] Workshop and Challenge on Learned Image Compression (CLIC), <http://www.compression.cc/> [3](#)
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", arXiv:1412.6980, pp.1-15, Dec. 2014. [3](#)
- [34] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multiscale structural similarity for image quality assessment", The 36-th Asilomar Conference on Signals, Systems and Computers, Vol.2, pp. 1398-1402, Nov. 2013. [3](#)