

MEMORY-EFFICIENT LEARNED IMAGE COMPRESSION WITH PRUNED HYPERPRIOR MODULE

Ao Luo¹, Heming Sun^{2,3}, Jinming Liu¹, Jiro Katto¹

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

²Waseda Research Institute for Science and Engineering, Tokyo, Japan

³JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama, Japan

ABSTRACT

Learned Image Compression (LIC) gradually became more and more famous in these years. The hyperprior-module-based LIC models have achieved remarkable rate-distortion performance. However, the memory cost of these LIC models is too large to actually apply them to various devices, especially to portable or edge devices. The parameter scale is directly linked with memory cost. In our research, we found the hyperprior module is not only highly over-parameterized, but also its latent representation contains redundant information. Therefore, we propose a novel pruning method named ERHP in this paper to efficiently reduce the memory cost of hyperprior module, while improving the network performance. The experiments show our method is effective, reducing at least 22.6% parameters in the whole model while achieving better rate-distortion performance.

Index Terms— Learned Image Compression, Hyperprior Module, Model Pruning

1. INTRODUCTION

Image compression is one of the important part for various applications. The conventional compression standards, such as JPEG [1], JPEG 2000 [2], BPG (Better Portable Graphics) [3] and VVC (Versatile Video Coding) [4] mainly use linear transform with hand-designed codes to compress the images. In recent several years, deep-learning based image compression (Learned Image Compression, LIC) methods take use of neural network, which has non-linear activation, to compress the images. These methods gradually outperform the classic ones.

One famous LIC method is Hyperprior [5], which utilizes a hyperprior module to capture the spatial redundancy among neighboring elements. The whole model consists of two parts: main path g_a, g_s and hyper path h_a, h_s , both of which are pairs of encoder $*_a$ and decoder $*_s$, as shown in Fig. 2. The main path receives input image, generating its latent representation y , which is assumed to obey arbitrary zero-mean Gaussian distribution. Then the hyper encoder uses y to calculate the side information z , which helps the

hyper decoder to generate scale of y . With scale information, y is rescaled to standard normal distribution and transmitted with z together to decoder part. Regarding to scale generated by z , y is generated from transmitted information and is given to main decoder to reconstruct the image. With the help of hyper path, Hyperprior model achieved dramatic progress, exceeding conventional methods. Regarding to its good performance, hyperprior module became an important part in latter methods, such as the methods improving performance [6, 7, 8], the ones making LIC models more applicable [9, 10].

However, one problem for LIC is that the requirements for memory are much larger than the conventional methods. With the development of mobile internet, more and more people tend to display and store images on portable or edge devices, whose memory is not sufficient for an image compression algorithm, such as mobile phones, advertising screens and so on. Therefore, it is difficult for LIC methods to widespread in practice.

The memory cost is directly linked with parameter scale. There are some former works focusing on lightweight image compression models with much lower FLOPs and parameter scale. [11] proposed several lightweight components, which decreased FLOPs and memory cost. [12] implements group Lasso loss to prune convolution layers' channels in decoding part, by which the researchers obtained lightweight models. However, the rate-distortion performance of these works dropped distinctly. In the meanwhile, [12] also found that their pruning method has little impact on hyper path.

In this paper, based on ResRep [13], we propose a pruning method called ERHP (Enhanced Resrep on Hyper Path in learned image compression) to prune LIC models. We finetune the pruned network to recover its performance, since the image compression task requires higher quality than image classification task on which ResRep is proposed. In this way, our method reduces parameter scale distinctively while even improving the performance. Our experiments on Hyperprior model[5] and Cheng[8] show the efficiency of that our method.

To summarize, our contributions are listed as below:

- We confirmed that hyper path is severely over-parameterized, which can be pruned to reduce memory consumption and redundancy.
- We propose a ERHP, which adapts ResRep pruning method to the LIC task by implementing PixelShuffle [14] layer and deconv layer, and prunes the LIC models efficiently.
- The experiments on Hyperprior model[5] and Cheng[8] show our method achieves much lower parameter scale and even improves the performance of the pruned model. In this way, the LIC models are more applicable for edge devices and perform better than the former models.

The following parts of this paper are arranged as below. In Sec. 2, we introduce the former methods directly linked with our work. After that, our proposed ERHP is introduced in Sec. 3. Then, the experiment results are shown in Sec. 4. Finally, we summarize our work in Sec. 5.

2. RELATED WORK

There have been plenty of works on lightweight models and cropping parameters. There are lots of former works taking use of Lasso loss penalty, such as [15, 16]. The Lasso loss penalty calculates l_1 regularization of weights ($\|W\|_1$ in Eq. 1) and add this penalty term to loss function with a coefficient β . In training step, the penalty term suppresses weights to zero, and the other terms in loss function amplify the weights. With the Lasso loss, part of weights are reduced to zero, while keeping others valid enough.

$$\mathcal{L}_{prune1} = \mathcal{L}_{quality} + \beta \cdot \|W\|_1 \quad (1)$$

Johnston et. al [12] applied to LIC the Group Lasso loss [17], which groups the weights in a convolution kernel (3x3 or 5x5 and so on) together by l_2 norm, as defined in Eq. 2,

$$\mathcal{L}_{prune2} = \mathcal{L}_{quality} + \beta \cdot \sum \|w_i\|_2 \quad (2)$$

where w_i is the i -th kernel in the network. This work cropped a lot of parameters, but the performance of output models dropped too.

ResRep [13] splits the weights for remembering and pruning. It add an 1x1 convolution layer behind each of the convolution layers to be pruned. This 1x1 convolution layer, called compactor, has the same input and output channel. The weights of compactor are initialized as identity matrices, whose output is the same as its input. The Group Lasso loss is only applied to the compactor, while the normal loss is applied to all the components. When finishing the pruning, ResRep combines each pair of original convolution layer and compactor together. In this way, the network retains good performance and is pruned efficiently. However, ResRep just

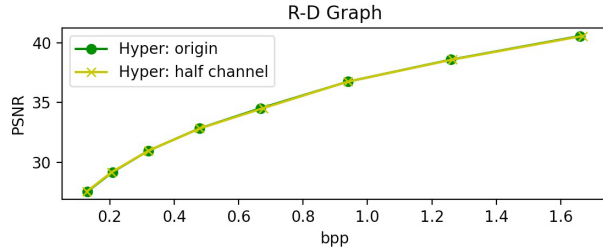


Fig. 1. Comparison between original and half-channel-pruned hyperprior model.

provided components for normal CNN, which cannot satisfy LIC models.

3. PROPOSED METHOD

3.1. Preliminary Analysis

In this section, we illustrate our analysis on hyper path, introducing the differences between main path and hyper path.

The parameter scale determines the upper bound of information volume can be obtained by the network. In our experiments, we compared the bit rate and parameter scale between main path (y) and hyper path (z). In hyperprior model [5], the parameter ratio of hyper path is approximately 40%, while the bit-rate ratio of z is only 1.7%. When inferring, the feature map of z is much smaller than that of y , which means the information z carries is distinctively lower than y . This phenomenon reveals the hyper path is heavily over-parameterized. We confirmed this conclusion in our experiments. As shown in Fig. 1, even cropped half of channels in hyper path, the models still achieves almost the same performance as original models.

On the other hand, [5] mentioned the relationship between channel numbers and rate-distortion performance. With more channels in the network, more details are extracted and obtained, finally lead to better performance. However, they set the same channels for main path and hyper path in their experiments, which ignored the difference between them. It is true that, with more channels in main path, the network extracts more details from the original image, which significantly improve the reconstructed image. However, hyper path generates the scale information to rescale y into an approximate range, which requires not so much details. In our experiments, more channels in hyper path even cause performance decline because of information redundancy.

3.2. Problem Formulation

We formulate the LIC task and introduce ResRep to better illustrate our proposed method.

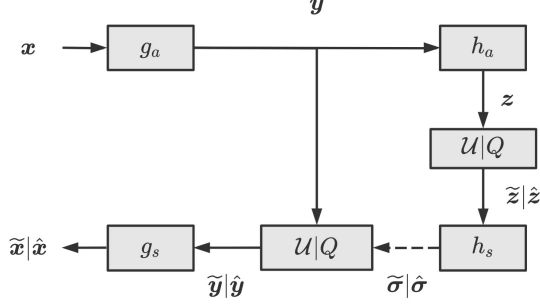


Fig. 2. Sketch of hyperprior model.

3.2.1. Learned Image Compression with Hyperprior

As mentioned before, the hyperprior model [5] is shown in Fig. 2. g_a , g_s , h_a , h_s are nonlinear neural networks. x is the input image, $y = g_a(x)$ and $z = h_a(y)$ are latent representation and hyper latent, respectively. $\hat{y} = Q(y)$ and $\hat{z} = Q(z)$ are quantized y and z . \hat{z} is taken as side information for generating the scale parameter $\hat{\sigma}$ for the entropy model of latent \hat{y} . $\hat{x} = g_s(\hat{y})$ is the reconstructed image. In training step, the quantization operation is applied by adding uniform noise, which produces differentiable variables \tilde{y} , \tilde{z} , \tilde{x} and $\tilde{\sigma}$. For simplicity, we represent $\tilde{x}|\hat{x}$, $\tilde{y}|\hat{y}$, $\tilde{z}|\hat{z}$ and $\tilde{\sigma}|\hat{\sigma}$ as \hat{x} , \hat{y} , \hat{z} and $\hat{\sigma}$. The loss function can be written as the trade-off of rate R and distortion D :

$$\mathcal{L}_{LIC} = R + \lambda D = \mathbb{E}_{x \sim p_x} [-\log_2 p_{\tilde{y}|\tilde{z}}(\tilde{y}|\tilde{z}) - \log_2 p_{\tilde{z}}(\tilde{z})] + \lambda \cdot \mathbb{E}_{x \sim p_x} [d(x, \hat{x})] \quad (3)$$

where the bit rate of \hat{y} and \hat{z} is evaluated by entropy, λ controls the trade-off of rate and distortion, p_* is probability of $*$, and $d(x, \hat{x})$ is the distortion, which is MSE in our work.

3.2.2. ResRep Pruning Method

The weight and bias of the i -th convolution layer can be shown as $\mathbf{W}_i \in \mathbb{R}^{C_{i+1} \times C_i \times ks \times ks}$ and $\mathbf{b}_i \in \mathbb{R}^{C_{i+1}}$, where C_{i+1} and C_i is the input channel number of layer i and $i-1$, and ks is the kernel size of this layer. The weight of compactor of the i -th convolution layer is $\mathbf{R}_i \in \mathbb{R}^{C_{i+1} \times C_{i+1} \times 1 \times 1}$, which is initialized as an identity matrix. After pruning, the \mathbf{R}_i is cropped as $\mathbf{R}'_i \in \mathbb{R}^{C'_{i+1} \times C_{i+1} \times 1 \times 1}$, where C'_{i+1} is the pruned channel number, satisfying $C'_{i+1} \leq C_{i+1}$. Then ResRep combines the normal convolution layer and compactor layer, as shown in Eqs. 4 and 5:

$$\begin{aligned} \mathbf{O} &= \mathbf{I} \otimes \mathbf{W}'_i + B(\mathbf{b}'_i) \\ &= (\mathbf{I} \otimes \mathbf{W}_i + B(\mathbf{b}_i)) \otimes \mathbf{R}'_i \end{aligned} \quad (4)$$

$$\begin{aligned} &= \mathbf{I} \otimes \mathbf{W}'_i \otimes \mathbf{R}'_i + B(\mathbf{b}_i) \otimes \mathbf{R}'_i \\ \mathbf{W}'_i &= T(T(\mathbf{W}_i) \otimes \mathbf{R}'_i) \end{aligned} \quad (5)$$

$$\mathbf{b}'_{i;j} = \mathbf{b}_i \cdot \mathbf{R}'_{i;j,;,;,}, \forall 1 \leq j \leq C'_{i+1} \quad (6)$$

where O and I are the output and input of the network, \cdot is element-wise multiply, \otimes is the convolution operation, $B(*)$ is the duplication for bias, and $T(*)$ is transposition.

3.3. Enhanced ResRep on Hyper Path in Learned Image Compression

The ResRep pruning method only implemented compactor for standard convolution layer and convolution layers with batch normalization, which cannot satisfy LIC models. In this paper, we propose an ERHP (Enhanced Resrep on Hyper Path in learned image compression), which implements compactors for PixelShuffle [14] layer and deconvolution layer for adaptation to LIC models.

The PixelShuffle component expands feature map by α times, and arranges neighbour α^2 channels into one plain. For example, the shape of input for PixelShuffle operation is $\alpha^2 C_{i+1} \times H \times W$, where the output should be $C_{i+1} \times \alpha H \times \alpha W$. Therefore, every α^2 channels in convolution layer of PixelShuffle generate a small patch, and should be reserved or removed together. We put the compactor after the shuffling operation, instead of after the convolution layer directly, as shown in Eq. 7:

$$\mathbf{O} = PS(\mathbf{I} \otimes \mathbf{W}_i^{ps} + B(\mathbf{b}_i^{ps})) \otimes \mathbf{R}'_i \quad (7)$$

where $PS(*)$ is the PixelShuffle operation, and the weight of convolution layer is $\mathbf{W}_i^{ps} \in \mathbb{R}^{\alpha^2 C_{i+1} \times C_i \times ks \times ks}$.

The combination of convolution layer and compactor is

$$\mathbf{W}_{i;k::\alpha^2,;,;}^{ps'} = T(T(\mathbf{W}_{i;k::\alpha^2,;,;}^{ps}) \otimes \mathbf{R}'_i), \forall 1 \leq k \leq \alpha^2 \quad (8)$$

$$\begin{aligned} \mathbf{b}'_{i;k \times \alpha^2 + j} &= \mathbf{b}_{i;k::\alpha^2} \cdot \mathbf{R}'_{i;j,;,;,}, \forall 1 \leq k \leq \alpha^2, \\ &1 \leq j \leq C'_{i+1} \end{aligned} \quad (9)$$

where the $k :: \alpha^2$ means that, starting from the k -th element, select an element every α^2 . For example, the k -th, $k + \alpha^2$ -th, $k + 2\alpha^2$ -th elements are selected.

The deconvolution layer is utilized in decoder part of hyper path. It upsamples the feature map of \hat{z} from bit stream. The structure of its weight is $\mathbf{W}_i^{de} \in \mathbb{R}^{C_i \times C_{i+1} \times ks \times ks}$, where the input and output channels are transposed, while the bias keeps the same. Therefore, the combination of its weight and corresponding compactor is written as

$$\mathbf{W}_i^{de'} = \mathbf{W}_i^{de} \otimes \mathbf{R}'_i \quad (10)$$

where the converting of bias is the same with Eq. 6.

With the newly implemented two components, we apply ResRep pruning method on hyper path. The whole loss func-

Table 1. Pruning results of ERHP. Quality (λ) is the trade-off of rate and distortion, as shown in Eq. 3. Corresponding PSNRs are the same because of frozen main path and context model.

Quality(λ)	Results of Hyperprior Model[5]			Results of Cheng[8]		
	Origin Performance PSNR@BPP	ERHP Performance PSNR@BPP	Parameter Scale pruned/origin	Origin Performance PSNR@BPP	ERHP Performance PSNR@BPP	Parameter Scale pruned/origin
0.0483	36.706@0.937	36.706@ 0.936	7.748M/11.582M(33.1%↓)	36.898@0.823	36.898@ 0.818	21.867M/28.244M(22.6%↓)
0.0250	34.501@0.667	34.501@ 0.667	3.705M/4.969M(25.4%↓)	35.282@0.603	35.282@ 0.601	21.576M/28.244M(23.6%↓)
0.0130	32.823@0.478	32.823@ 0.476	3.651M/4.969M(26.5%↓)	33.521@0.433	33.521@ 0.429	21.327M/28.244M(24.5%↓)
0.0067	30.962@0.319	30.962@ 0.319	3.589M/4.969M(27.8%↓)	31.318@0.292	31.318@ 0.290	9.700M/12.563M(22.8%↓)
0.0035	29.192@0.209	29.192@ 0.208	3.264M/4.969M(34.3%↓)	29.763@0.199	29.763@ 0.197	9.663M/12.563M(23.1%↓)
0.0018	27.578@0.131	27.578@ 0.131	3.318M/4.969M(33.2%↓)	28.233@0.130	28.233@ 0.127	9.526M/12.563M(24.2%↓)

Table 2. Comparisons of different pruning methods on Cheng[8].

Quality(λ)	PSNR	BPP		
		Origin	Manual	ERHP
0.0483	36.898	0.823	0.983	0.818
0.0250	35.282	0.603	0.601	0.601
0.0130	33.521	0.433	0.429	0.429
0.0067	31.318	0.292	0.292	0.290
0.0035	29.763	0.199	0.197	0.197
0.0018	28.233	0.130	0.129	0.127

tion for pruning is written as

$$\begin{aligned}
 \mathcal{L} &= R + \lambda D + \beta \mathcal{L}_{lasso} \\
 &= \mathbb{E}_{x \sim p_x} [-\log_2 p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}) - \log_2 p_{\hat{z}}(\hat{z})] \\
 &\quad + \lambda \cdot \mathbb{E}_{x \sim p_x} [d(x, \hat{x})] \\
 &\quad + \beta \cdot \sum_{i=1}^L \sum_{j=1}^{C_i} \|\mathbf{R}_{i;j,\dots}\|_2
 \end{aligned} \tag{11}$$

where λ is the rate-distortion trade-off, β is the lasso penalty strength, $\mathbf{R}_{i;j,\dots}$ is the j -th channel of the i -th layer to be pruned, and L means the total number of layers to be pruned.

In the original ResRep, the authors directly use the pruned model to classify images. However, the image compression task requires higher latent representation qualities than image classification task, so we finetune the pruned model to recover the original performance.

4. EXPERIMENTS

In this section, we show our implement details and experiment results of ERHP-pruned models on Hyperprior model[5] and Cheng[8], which achieve distinctively lower memory cost and improve the rate-distortion performance.

4.1. Implement Details

We set the lasso strength β to 1e-9, which is small enough to keep convolution layers available, while effectively pruning the model. We set the preliminary pruning target to 0.7, which

means the initial aim is to prune 70% parameters in hyper path and is appropriate for most of the models. After pruning, we use [18] to finetune our model, with learning rate of 1e-4.

4.2. Experiments of Pruned Models

We trained our models on OpenImage [19] and tested them on Kodak [20]. The PSNR ($10 \log_{10} \frac{255^2}{mse}$) and bit-per-pixel (bpp) are taken as our evaluating metrics.

We treat models as two parts: hyper path to prune; main path and context model (Hyperprior model[5] does not have context model) to freeze. To be concrete, when pruning, we initialize the model with pretrained models from [18], then train the hyper path only, while main path and context model frozen. We apply compactors to the whole hyper path except the last layer in hyper decoder, since the final output channel cannot be changed according to the fixed dimension of y . Table 1 shows the general results of ERHP on Hyperprior model and Cheng model. The results prove the efficiency of our ERHP, achieving at least 22.6% parameter reduction in the whole model.

We did ablations on ERHP and manually pruned models at the same parameter scale with ERHP results but uniform channel number for each layer in hyper path, as shown in Table 2. Because the main path and context model are frozen, the PSNR of same-quality models are the same in our experiments. Our method not only solves the over-parameterize problem, but also reduces the redundancy in z , improving the performance of the models slightly. All the results of ERHP are better than or equal to the corresponding manually pruned models, which shows the effectiveness of our method.

5. CONCLUSIONS

In this paper, we propose a novel ERHP (Enhanced Resrep on Hyper Path in learned image compression) for reducing the memory cost of LIC models by pruning channels of hyper path. We perform compactors for PixelShuffle and deconvolution layer, which are used in LIC models. The experiments on Hyperprior model and Cheng model show that our method is effective, pruning a large amount of parameters while improving the rate-distortion performance.

6. REFERENCES

- [1] Gregory K Wallace, “The jpeg still picture compression standard,” *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.
- [2] Majid Rabbani, “Jpeg2000: Image compression fundamentals, standards and practice,” *Journal of Electronic Imaging*, vol. 11, no. 2, pp. 286, 2002.
- [3] Fabrice Bellard, “Bpg image format,” <https://bellard.org/bpg>, 2015.
- [4] Benjamin Bross, Jianle Chen, Shan Liu, and Ye-Kui Wang, “Jvet-s2001 versatile video coding (draft 10),” in *Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11*, 2020.
- [5] 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.
- [6] David Minnen, Johannes Ballé, and George D Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” in *Advances in Neural Information Processing Systems*, 2018, vol. 31.
- [7] Zhao Zan, Chao Liu, Heming Sun, Xiaoyang Zeng, and Yibo Fan, “Learned image compression with separate hyperprior decoders,” *IEEE Open Journal of Circuits and Systems*, vol. 2, pp. 627–632, 2021.
- [8] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7939–7948.
- [9] Fei Yang, Luis Herranz, Yongmei Cheng, and Mikhail G Mozerov, “Slimmable compressive autoencoders for practical neural image compression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4998–5007.
- [10] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin, “Checkerboard context model for efficient learned image compression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14771–14780.
- [11] Zekun Zheng, Xiaodong Wang, Xinye Lin, and Shaohe Lv, *Get The Best of the Three Worlds: Real-Time Neural Image Compression in a Non-GPU Environment*, p. 5400–5409, Association for Computing Machinery, New York, NY, USA, 2021.
- [12] Nick Johnston, Elad Eban, Ariel Gordon, and Johannes Ballé, “Computationally efficient neural image compression,” Tech. Rep., Google Research, 2019.
- [13] Xiaohan Ding, Tianxiang Hao, Jianchao Tan, Ji Liu, Jungong Han, Yuchen Guo, and Guiguang Ding, “Resrep: Lossless cnn pruning via decoupling remembering and forgetting,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4510–4520.
- [14] Wenzhe et. al. Shi, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.
- [15] Yihui He, Xiangyu Zhang, and Jian Sun, “Channel pruning for accelerating very deep neural networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1389–1397.
- [16] Zheng et. al. Zhan, “Achieving on-mobile real-time super-resolution with neural architecture and pruning search,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4821–4831.
- [17] Ming Yuan and Yi Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pp. 49–67, 2006.
- [18] 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.
- [19] Candice Schumann et. al., “A step toward more inclusive people annotations for fairness,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.
- [20] “The kodak photocd dataset,” <http://r0k.us/graphics/kodak/>.