CNN-Optimized Image Compression with Uncertainty based Resource Allocation

Zhenzhong Chen*, Yiming Li, Feiyang Liu, Zizheng Liu,
Xiang Pan, Wanjie Sun, Yingbin Wang, Yan Zhou, Han Zhu†
Wuhan University, Wuhan, China
zzchen@whu.edu.cn

Shan Liu
Tencent Media Lab, Palo Alto, CA, USA
shanl@tencent.com

Abstract

In this paper, we provide the description of our approach designed for participating the CVPR 2018 Challenge on Learned Image Compression (CLIC). Our approach is a hybrid image coder based on CNN-optimized in-loop filter and mode coding, with uncertainty based resource allocation for compressing the task images. Two solutions were submitted, i.e., "iipTiramisu" and its speedup version "iipTiramisuS", resulting in 32.14 dB and 32.06 dB in PSNR, respectively. These two results have been ranked No. 1 and 2 on the leaderboard.

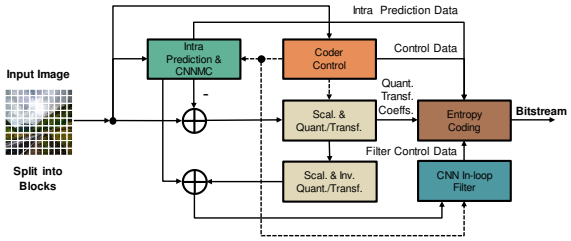


Figure 1: The architecture of our image codec.

1. Introduction

Recently, the popularity of image sharing on the social network or instant messenger such as QQ and WeChat in daily lives has brought explosive growth of image data, which brings great challenges in applications and services due to the limited communication bandwidth and storage resources. Many image compression methods have been developed to efficiently compress the image, e.g. JPEG, JPEG 2000 , WebP, H.265/HEVC Main Still Picture (HEVC-MSP) profile etc. For example, BPG is an image compression scheme developed based on H.265/HEVC. It includes some optimization based on H.265/HEVC intra coding, which achieves higher compression performance than JPEG or JPEG 2000 at similar quality.

In this paper, we design a hybrid block-based image compression approach based on conventional neural network (CNN). First, we introduce CNN into two important parts of the traditional hybrid block-based encoder, mode coding and in-loop filter, to guarantee high coding efficiency. We design a CNN based method to predict probability distribution of syntax element and boost the performance of in-loop filtering with a novel convolutional network that incorporates dense connection and identity skip connection. Moreover, the uncertainty based resource allocation is used to solve the constrained minimum distortion optimization problem, which can determine appropriate quantization parameter (QP) for each image efficiently. The experimental results demonstrate the superior performance of the proposed approach.

2. Our Image Compression Approach

2.1. Hybrid block-based image codec

In the proposed approach, we develop our codec based on the JEM platform, i.e., Joint Exploration Model (JEM) 7.1 [1], which is the codec developed based on H.265/HEVC structure. We design CNN based in-loop filter (CNNIF) and CNN based mode coding (CNNMC) to enhance the coding performance. Our image compression

*Corresponding Author.
†In alphabet order.

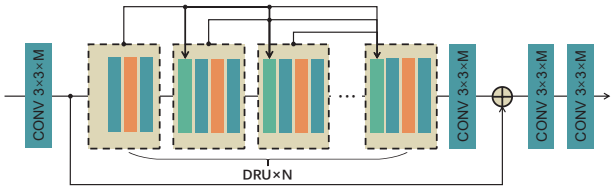


Figure 2: The architecture of CNNIF. N and M denote the number of DRU and the feature maps generated by the convolutional layer, respectively.

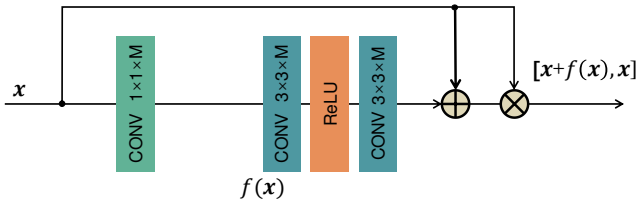


Figure 3: Dense residual unit in our network.

framework is shown in Fig. 1. Each image is split into block-shaped regions, and coded using intra prediction and other coding modules. The residual signal of intra prediction is transformed by a linear spatial transformation. The transform coefficients are then scaled, quantized and coded with entropy coding. Moreover, CNNMC and CNNIF are performed based on rate-distortion optimized coder control. In addition, relevant and irrelevant syntax elements are revisited to further improve the coding performance.

2.2. CNN based In-Loop Filter (CNNIF)

In-loop filtering is an important technique in current mainstream codec for improving the quality of compressed image. Motivated by the latest advances of CNNs in image classification [2, 3] and image restoration [4, 5], we design a novel CNN architecture and further boost the performance of in-loop filtering.

The whole architecture of our network is shown in Fig. 2. The input of the network is the decoded image in RGB color space. Apart from the last convolution layer that outputs a 3-channel image, other layers generate the same number of feature maps. The network is mainly constructed by stacking many dense residual units (DRU) as shown in Fig. 3. We employ the residual block structure with identity shortcuts and further improve it for better performance on in-loop filtering. The goal of our network is to restore the information lost in encoding process, where the input and the target are highly similar. Since most of the residuals between the input and the target are small or zero, the residual learning contributes to faster convergence and easier training without introducing extra parameters. The rectified linear unit after the shortcut connection of the residual block is removed for fast convergence as indicated in [4]. There is a highway connection in the unit that allows the signal to bypass the convolutional layers and directly propagate to the next unit. The outputs of the first and the last convolutional layers are firstly added and then concatenated with the original input to generate the final output of the unit. Then each DRU further learns the residuals of above feature maps and propagates them to the next unit. Since each DRU receives the outputs from all preceding units, we append a 1×1 convolutional layer as the bottleneck layer for saving computational resources. No activation layer is appended after the bottleneck layer so that the bottleneck layer simply generates the linear combination of the inputs. The first 1×1 convolutional layer inside each unit plays a role of combining the information from all those inputs by weights and reducing the number of parameters of the network model at the same time.

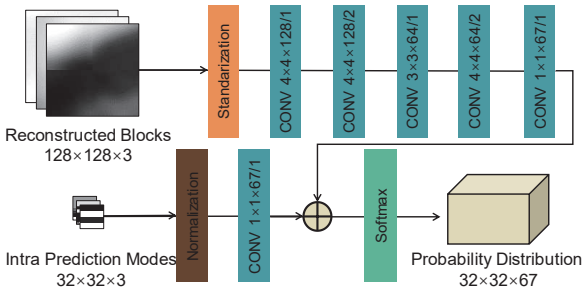


Figure 4: Architecture of the proposed network for probability distribution prediction. Convolutions are followed by tanh activations.

Note that the batch normalization (BN) layers are removed in our network, since it normalizes the input signals and may lead to difference between the input and the target. No activation layer is appended in the network except in the DRU and the first DRU has no 1×1 convolutional layer. Stacked feature maps and dimension reduction achieved by 1×1 convolutional layer are the key of keeping balance between promoting filtering performance and saving computational resources.

2.3. CNN based Mode Coding (CNNMC)

For modern image compression, intra prediction is a key role to provide high efficiency in compression. The corresponding syntax elements, e.g., intra prediction mode, will be encoded by Context-Based Adaptive Binary Arithmetic Coding (CABAC) in JEM. To compress it efficiently, a heuristic method to derive the Most Probable Modes (MPMs) has been introduced since H.264/AVC. The opti-

mal intra prediction mode, which is decided by RDO, is encoded based on the MPMs to help reduce the number of encoded symbols. Hence the accuracy of the MPMs is of great importance to guarantee the compression performance and the CNN based method could improve the performance eventually [6]. In this work, we design a CNN-based probability distribution prediction for intra prediction mode coding.

The architecture of the proposed network is shown in Fig. 4. It utilizes the above, the left and the above-left reconstructed blocks. Reconstructed blocks and intra prediction modes of them are used as the input of the proposed network. Note that intra prediction information is stored every 4×4 block, so there are totally 32×32 units in a 128×128 block. Hence the size of input reconstructed blocks is $128 \times 128 \times 3$ and the size of intra prediction information is $32 \times 32 \times 3$. On the other hand, the directional intra modes in JEM are extended from 33, as defined in HEVC, to 65 [1]. Therefore, the output of the network is a probability distribution P of each unit of all the modes with size of $32 \times 32 \times 67$, which means that $P(i, j)$ is a 67-dimension (with DC mode and planar mode) non-negative vector whose sum is 1 and $P(i, j, k)$ denotes probability of the kth prediction mode of the unit located at ith row and jth column. The MPMs are derived from P according to the location and size. For example, if the block's upper left point is located at (r, c) with size $m \times n$ in the block, then the corresponding probability p is obtain as:

$$p = \frac{\sum_{i=0}^{m} \sum_{j=0}^{n} P(i + r, j + c)}{m \times n} \tag{1}$$

which is used to indicate the probabilities of prediction modes. Then the MPMs are derived and sorted based on p.

2.4. Uncertainty based Resource Allocation (UN-RA)

As the task of the challenge is to minimize the overall distortion of the image set under the given rate constraint, we can find the allocated rate for each image by solving the corresponding optimization problem. For this, we need several tuples of R-D data to obtain R-D relationship for each image. However, there usually exists difference between the estimated values and the real values of parameters due to the inevitable estimation error, which is regarded as uncertainty in this paper. After statistical analysis of the coding results of the training set, we can apply the normal distribution to express the uncertainty. Thus, the objective of uncertainty based resource allocation is to minimize the total expectation distortion subjective to the rate constraint:

Method	bpp	PSNR
BPG	0.149	30.85
UN-RA+CNNIF_S (iipTiramisuS)	**0.149**	**32.06**
UN-RA+CNNIF+CNNMC (iipTiramisu)	**0.148**	**32.14**

Table 1: The performance of the proposed approaches. The details could be found on the leaderboard (http://www.compression.cc/leaderboard/).

$$\arg\min_{Q_1 \cdots Q_N} \sum_{i=1}^{N} \mathbb{E}(D_i(Q_i) \times P_i) \text{ s.t.} \sum_{i=1}^{N} \mathbb{E}(R_i(Q_i) \times P_i) \leqslant T \tag{2}$$

where Q_i means the QP of ith image. $D_i(Q_i)$ and $R_i(Q_i)$ stand for the distortion (MSE) and rate (bpp) of the ith image encoded by Q_i, respectively. P_i represents the number of pixels in the ith image and T means the total target bits. Specifically, we apply a hyperbolic function based R-D model [7], and the relationship between the optimal QP for a given λ to obtain R-QP and D-QP models. This optimization problem can be solved by applying dynamic programming. We first construct a graph where the node represents the rate and distortion of an image encoded by a specific QP, then we can use Depth-First-Search (DFS) method to search the optimal QP combinations in the constructed graph. For practice, we divide the test image set into group of pictures (GOP) with M frames ($M > 1$) in each GOP, which can reduce the complexity to find the optimal QP combination. After obtaining the optimal QP combinations for each GOP, they can be applied in practical image compression.

3. Experiments

To evaluate the performance and the effectiveness of our method, we participate in the Workshop and Challenge on Learned Image Compression (CLIC) of CVPR 2018. This challenge provides two image datasets: Dataset P ("professional") and Dataset M ("mobile") that are collected to be representative for images commonly used in the wild, containing around two thousand images (in which about 1633 images are used for training, 102 for validation and 286 for test). The participants are required to submit an encoded file for each test image and the total file size should be less than 0.15 bpp. Due to the limit of space, we only provide the results of our approaches and BPG on the validation dataset under the challenge requirements. For the proposed method, we first apply the UN-RA approach and set 0.15 bpp as the rate constraint. Then, we enable the designed CNN based in-loop filter (CNNIF) and CNN based mode coding (CNNMC) tools for evaluation. We also designed a simplified version of the CNNIF, named CNNIF_S, where

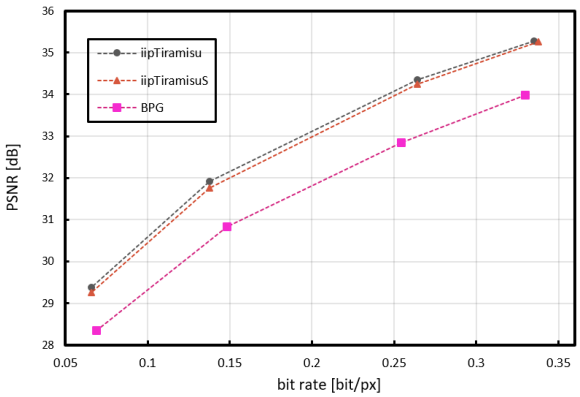


Figure 5: The Rate-PSNR curve of different encoders. (bit rate range 0.05-0.35 bpp)

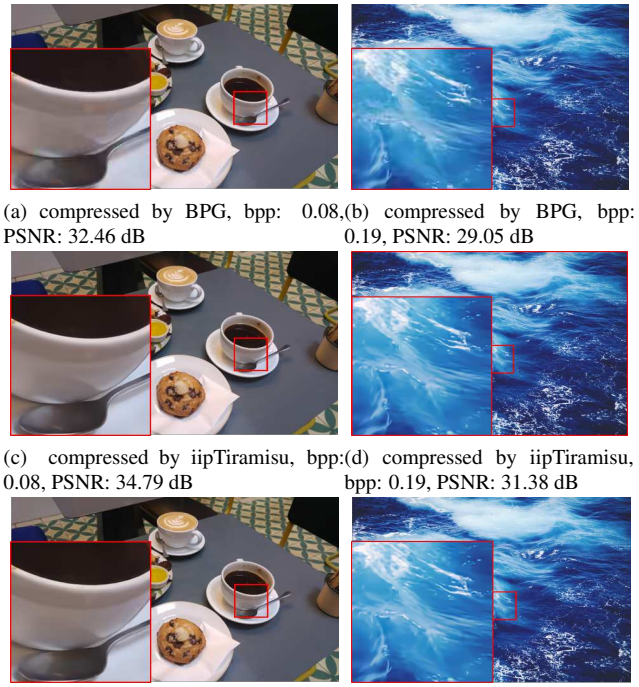
Compared Anchor	Proposed Scheme	
	iipTiramisu	iipTiramisuS
BPG	-30.80%	-27.90%

Table 2: The BD-rate comparison of the proposed schemes and BPG. (bit rate range 0.05-0.35 bpp)

the number of DRU is reduce from 8 to 2, and the number of feature maps is reduced from 64 to 32, respectively. In the experiment of speedup version, we only enabled CNNIF_S. From Table. 1, we can see the PSNR results of proposed approaches are both higher than BPG for more than 1 dB, whose PSNR are 32.06 dB and 32.14 dB respectively while that of BPG is 30.85 dB. To further verify the performance of our proposed schemes, the Rate-PSNR curve of different encoders is shown in Fig. 5, and the corresponding Bjϕntegaard-Delta rate (BD-rate) gain is listed in Table 2, where we can see our two schemes both outperform BPG significantly. When compared with BPG, the proposed scheme can gain up to 30.80%. Fig. 6 also shows some compressed images for visualization.

4. Conclusion

In this paper, we have designed a hybrid block-based image compression approach. Based on traditional hybrid block-based encoder, a novel convolutional network that incorporates dense connection and identity skip connection is proposed to enhance the performance of in-loop filtering, then a CNN based method is employed to predict probability distribution of syntax elements and reduce the volume of symbols to be encoded. Finally we solve the constrained minimum distortion optimization problem and find the appropriate QP for each image based on an uncertainty based resource allocation method. As shown in the results of the challenges, our approach achieves the best performance among all submissions.



(a) compressed by BPG, bpp: 0.08, PSNR: 32.46 dB
(b) compressed by BPG, bpp: 0.19, PSNR: 29.05 dB
(c) compressed by iipTiramisu, bpp: 0.08, PSNR: 34.79 dB
(d) compressed by iipTiramisu, bpp: 0.19, PSNR: 31.38 dB
(e) compressed by iipTiramisuS, bpp: 0.08, PSNR: 34.31 dB
(f) compressed by iipTiramisuS, bpp: 0.19, PSNR: 31.18 dB

Figure 6: Visualization of some compressed images and their enlarged details.

References

[1] "JVET Reference software Ver. 7.1 (JEM-7.1)," 2018. https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/tags/HM-16.6-JEM-7.1/.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[3] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[4] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, 2017.

[5] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPR Workshop*, 2017.

[6] R. Song, D. Liu, H. Li, and F. Wu, "Neural network-based arithmetic coding of intra prediction modes in HEVC," in *VCIP*, 2017.

[7] S. Mallat and F. Falzon, "Analysis of low bit rate image transform coding," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 1027–1042, 1998.