Joint learned and traditional image compression for transparent coding

Pikpik Inc.
pikpiktech@gmail.com

Abstract

This paper proposes a novel image compression framework, which consists of a CNN-based method and a versatile video coding (VVC) based method. The CNN-based method uses the auto-encoder to learn the quantized latent representation of the image and joints the autoregressive and hierarchical priors to exploit the probabilistic structure. We also design a post-processing network for VVC to further improve the quality of compressed images. We find that CNN-based method and VVC-based method are complementary to each other in terms of MS-SSIM and PSNR. Thus, we combine the two methods together to obtained better coding performance. Furthermore, to select the best compression parameter, an optimal coding mode selection algorithm is introduced. Experimental results indicate that the proposed image compression scheme can achieve significantly better rate-distortion (RD) performance than other methods.

1. Introduction

The fast development of image capture and display devices has brought a dramatic demand for high definition (HD) and Ultra high definition (UHD) images. At present, there are many image coding standards, such as JPEG 2000 [13] and BPG [6]. But these traditional standards rely heavily on manual operations and are still not efficient enough to compress images, which become the burden of image storage and transmission. Moreover, the streaming of digital media is expected to reach 80% of Internet traffic by 2020 [7]. Therefore, lossy image compression is becoming more and more important in saving transmission bandwidth and hardware storage.

Recently, lossy image compression based on depth neural network (DNN) has attracted significant attention. Compared with JPEG 2000, some of them have competitive or even higher coding performance, which shows that DNN-based image compression has great potential. These work can be broadly categorized into two types according to the network architecture: recursive neural network (RNN) [17, 19, 9] and convolutional neural network (CNN) [15, 2, 4, 5, 10, 8, 11, 18, 12, 2, 16]. The RNN-based method provides a variable rate, but the iterative mode is very complex, and both the encoding and decoding processes are iterative. As a result, both in training and application, the requirements of hardware storage and performance are extremely high.

On the contrary, the CNN-based method compresses the image effectively. The transform in traditional coding algorithm is based on a linear orthogonal transform. However, the research shows that there are still many high-dimensional correlation redundancies in the nature image after the linear transform, which can be significantly reduced by using a nonlinear transform. Fortunately, CNN-based method can map pixels to a more compressible potential space than the linear transform used in traditional image codecs by learning nonlinear functions. This nonlinear transform coding method is similar to auto-encoder composed of an encoder and a decoder. The encoder converts the pixels to a reduced-dimensional latent space, while the decoder maps the latents back to the pixels. Some standard CNN modules such as res-block are applied to create an auto-encoder in [16]. Agustsson *et al.* [3] proposes to use the vector quantization to replace the scalar quantization in [16]. Ballé *et al.* [4] firstly designs the generalized divisive normalization (GDN) and its inverse (IGDN) to obtain the local joint statistics of images. All of these CNN-based approaches optimize the model by minimizing the trade-off between rate and distortion between the original and reconstructed images.

The proposed image compression framework is based on the work of Minnen *et al.* [11], which studies autoregressive, hierarchical and combinatorial priors as alternatives, and weighs their costs and benefits in the context of image compression. According to the quality requirement of challenge (PSNR $\geq$ 40dB, MS-SSIM $\geq$ 0.993), an optimal mode selection algorithm combining the traditional coding method (versatile video coding, VVC)[1] and the CNN-based coding method is designed. Usually, VVC has better coding performance in terms of PSNR, while CNN-based method has better coding performance in terms of

MS-SSIM. The proposed algorithm makes full use of the advantages of the two coding methods. Specifically, for the transparent track of the challenge, a high bit rate model is trained for CNN-based coding method. Meanwhile, a post-processing network is trained for VVC, and the bit rate is further reduced when the quality requirement is satisfied. Finally, an optimal mode selection algorithm is introduced to select the best compression parameter for each image. As a result, when the quality requirements of both PSNR and MS-SSIM are satisfied at the same time, the bit rate of the proposed algorithm is lower than that of using the two methods alone.

2. Framework of the Proposed Image Compression Method

This section will give a detailed description of the proposed image compression framework. The following subsections briefly describe the key elements of the design.

2.1. Problem modeling

Image compression is primarily characterized in terms of bit rate and perceived distortion of the reconstructed image. The main task of this work is to convey the sequence of images with minimum possible bit rate while maintaining a specific perceived distortion level. For this problem, the fundamental issue is to obtain the best trade-off between the rate and perceived distortion. The process used to achieve this objective is commonly known as rate-distortion optimization (RDO), which can be expressed by minimizing the bit rate R with the perceived distortion D subjected to a constraint D_c. Thus, the problem of the transparent track of the challenge can be expressed as follows:

$$min\{R\} \qquad subject\ to\ D \leq D_c \tag{1}$$

This is a typical constrained optimization problem which is usually solved by Lagrangian optimization and dynamic programming. In practical applications, the computational complexity of dynamic programming is often too high, which is only used when direct Lagrangian optimization is difficult. Thus, this paper adopts the Lagrangian optimization technique to convert the constrained optimization problem to an unconstrained optimization problem, which can be expressed as:

$$min\{J\}\ where\ J = R + \lambda_{mse} D_{mse} + \lambda_{msssim} D_{msssim} \tag{2}$$

where J is called the rate-distortion (RD) cost and the rate R is measured in the number of bits per pixel. Since this task needs to meet the quality requirements of PSNR and MS-SSIM at the same time, we consider two perceived distortion in this work. D_{mse} represents the mean squared error. D_{msssim} is used to define the measure of perceived distortion of MS-SSIM and can be calculated by $D_{msssim} = 1 - MSSSIM$. λ_{mse} and λ_{msssim} are the Lagrange multiplier which controls the trade-off between bit rate and distortion. λ_{mse} and λ_{msssim} are adapted for each image by taking the properties of the input images into consideration.

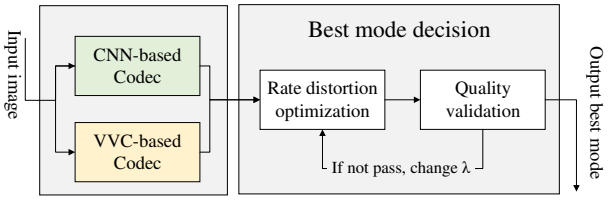

Figure 1. The framework of the proposed image compression method.

In order to achieve optimal RD performance, it is very important to carefully choose λ_{mse}, λ_{msssim} and the best coding mode. For each image, we define two types of coding modes: CNN-based coding mode and VVC-based coding mode. Within the framework, we firstly learn multiple models for CNN-based method. These models have different bit rates and will be used for optimal mode selection. Similarly, for VVC-based method, we also train multiple models with different bit rates. Finally, all of these coding results are combined as different coding modes. The switching between the traditional and CNN is done per image. The training process will keep adjusting Lagrange multiplier until the quality just meets the requirements via optimizing Eq.(2), and then the best bit allocation for each image is obtained. Based on the above analysis, the framework of the proposed image compression method is shown in Fig. 1.

2.2. CNN-based coding mode

In order to get the CNN-based coding mode, we design an end-to-end trainable image compression algorithm for transparent coding. Following the work of Minnen *et al.* [11], a high-level overview of the proposed image compression model is shown in Fig. 2, which consists of two sub-networks. Different from previous work, we increase the number of filter channels for the high bit rate coding and adopt the range code for the entropy coding.

The first sub-network is the core auto-encoder, which learns the quantized potential representation of the image. The encoder and decoder blocks in the first sub-network are composed of convolutions and GDN/IGDN. Within this module, the input is recursively analyzed by linear filters to extract the primary features which include the most basic information of the input. The primary features are normalized to form the first scale representation, which can be used for reconstruction with basic quality. Inspired by Ballé's work

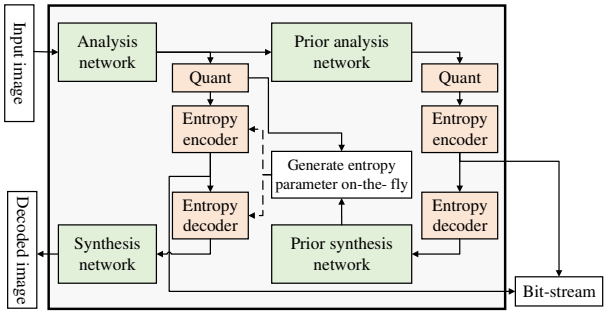

Figure 2. Overview of the proposed CNN-based image compression model.

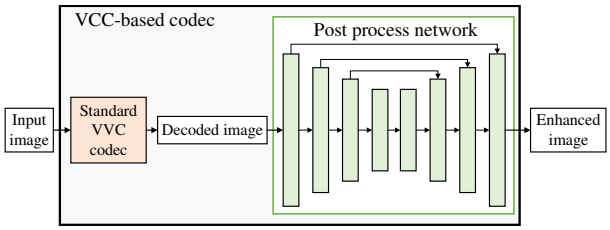

Figure 3. The architecture of the proposed post-processing network.

et al. [4, 5], GDN/IGDN is a special form of joint local gain control and is used as nonlinear normalization. Compared with the point-wise nonlinearity following by "batch normalization", GDN/IGDN has higher efficiency in representing the local probability structure of photographic images and provides higher nonlinearity and spatial adaptability.

In order to allow optimization by stochastic gradient descent, the quantization approximations studied include the gradient of the replacement quantizer [16] and the substitution of additive uniform noise for the quantizer itself in the course of training [4]. According to the experiment, we find that the latter approximation method with uniform additive noise achieves higher RD performance than the alternative quantization gradient method. Therefore, the following approximate methods are used in this paper:

$$\hat{F} \approx F + \mu, \ \mu \sim \boldsymbol{U}(0,1) \tag{3}$$

where μ is a uniform noise with zero distribution centre and the same width as the quantization bin, $\hat{F}$ is the representation out of the forward transform.

The second sub-network is responsible for learning the quantized probability model for entropy coding. It combines the context model and an autoregressive model with hyper-networks, which learns to represent information used to correct context-based predictions. The data from these two sources is combined with entropy parameter network to generate the mean and scale parameters of the conditional Gaussian entropy model.

2.3. VVC-based mode

As stated above, the CNN-based and VVC-based compression methods are complementary to each other in terms of PSNR and MS-SSIM, and better coding performance can be achieved by mixing the two methods. In this paper, we adopt the VVC encoder [1] to obtain the VVC-based mode. However, the reconstructed images of VVC encoder still contain compression artifacts, such as blocking artifacts, ringing effects, blurring, etc.. It is desired to study on improving the visual quality of the decoded image. Thus, we designed a residual-learning CNN as post-processing techniques to further improve the quality of the compressed images in VVC. The architecture of the proposed post-processing network is illustrated in Fig. 3. The proposed method is based on residual blocks, which is widely used in de-noising[20], super-resolution[14] etc.. We totally adopt 10 residual blocks, which makes a good trade-off between performance and computational complexity.

3. Experimental Results

In this section, we firstly introduce the experiment configurations and the training details. Then we present the coding performance of the proposed method (CNN+VVC+POST), which is compared with four methods, i.e., the VVC-based method without the post-processing (VVC) [1], the individual performance of the CNN-based method (CNN) and VVC-based method with post-processing (VVC+POST), and the joint performance of the CNN-based method and the VVC-based method without the post-processing (CNN+VVC).

All experiments use a training dataset of 1500 high quality natural images which are downloaded from flickr.com. Each mini-batch used four 256×256 patches randomly cropped from these images. During the training, the mini-batch size is 4 and an initial learning rate of 0.01 is adopted. The network parameters are learned by minimizing the loss function with stochastic gradient descent.

To verify the performance of the proposed image compression method, we conduct an experiment for the validation dataset. Table 1 demonstrates the coding performance of different method for validation dataset. Among these methods, the proposed whole algorithm has the lowest bits when the quality requirement is satisfied. In other words, the proposed method has the best coding performance. Specifically, when the quality requirement is satisfied, the consumed bits of the proposed method, VVC, VVC+POST, CNN, CNN+VVC are 28378366, 31968164, 29945191, 32285753 and 30506672, respectively. In the

	PSNR	MS-SSIM	bits	bpp
VVC	40.6396	0.9930	31968164	1.0154
VVC+POST	40.5802	0.9930	29945191	0.9512
CNN	40.0238	0.9934	32285753	1.02553
CNN+VVC	40.0287	0.9934	30506672	0.9690
CNN+VVC+POST(Pikpik)	40.0000	0.9930	28378366	0.9055

Table 1. Evaluation results on CLIC 2019 validation dataset.

	PSNR	MS-SSIM	bits	bpp
CNN+VVC+POST(Pikpik)	40.0010	0.9930	107402743	1.0229

Table 2. Evaluation results on CLIC 2019 test dataset.

test stage, we have submitted one result named Pikpik for compressing the images in the test dataset. Table 2 shows the final result for test dataset. Generally, the performance is improved by two main reasons. Firstly, our algorithm combines the advantages of VVC-based and CNN-based coding methods, which can achieve better coding performance. Secondly, in order to further improve the compression efficiency of the VVC coding method, we propose a post-processing module in VVC.

4. Conclusion

In this paper, a novel image compression framework which combines the VVC-based method and CNN-based method is proposed. In details, the CNN-based method adopts the auto-encoder to learn the approximate invertible mapping from pixels to quantized latent representations. Meanwhile, the CNN-based method joints the autoregressive and hierarchical priors to further improve compression performance. The VVC-based method uses the VVC encoder with a post-processing module to encoding the images. Usually, CNN-based method has better coding performance in terms of MS-SSIM, while VVC-based method has better compression performance in terms of PSNR. We find that in terms of compression performance, CNN-based method and VVC-based method are complementary to each other and can be combined to make better use of bit allocations. Thus, an optimal mode selection algorithm is presented to select the best compression for each image. Experimental results have demonstrated that the proposed overall algorithm can significantly reduce the bits while keeping the same video quality.

References

[1] H266 (https://de.wikipedia.org/wiki/h.266/), 2018. 1, 3

[2] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Advances in Neural Information Processing Systems*, pages 1141–1151, 2017. 1

[3] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. Van Gool. Soft-to-hard vector quantization for end-to-end learned compression of images and neural networks. 04 2017. 1

[4] J. Ballé, V. Laparra, and E. P. Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016. 1, 3

[5] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018. 1, 3

[6] F. Bellard. Bpg image format (http://bellard. org/bpg/), 2017. 1

[7] V. N. Index. White paper: Cisco vni forecast and methodology, 2015-2020. 2016. 1

[8] F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, and D. Zhao. An end-to-end compression framework based on convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3007–3018, Oct 2018. 1

[9] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S. Jin Hwang, J. Shor, and G. Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[10] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang. Learning convolutional networks for content-weighted image compression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[11] D. Minnen, J. Ballé, and G. Toderici. Joint autoregressive and hierarchical priors for learned image compression. *CoRR*, abs/1809.02736, 2018. 1, 2

[12] O. Rippel and L. Bourdev. Real-time adaptive image compression. *arXiv preprint arXiv:1705.05823*, 2017. 1

[13] D. S. Taubman and M. W. Marcellin. *JPEG2000 Image Compression Fundamentals, Standards and Practice*. 01 2002. 1

[14] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2017. 3

[15] L. Theis, W. Shi, A. Cunningham, and F. Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017. 1

[16] L. Theis, W. Shi, A. Cunningham, and F. Huszár. Lossy image compression with compressive autoencoders. 03 2017. 1, 3

[17] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar. Variable rate image compression with recurrent neural networks. *CoRR*, abs/1511.06085, 2016. 1

[18] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell. Full resolution image compression with recurrent neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5435–5443. IEEE, 2017. 1

[19] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, and M. Covell. Full resolution image compression with recurrent neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[20] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 3