Hierarchical Image Compression Framework Conference Submissions

Anonymous authors
Paper under double-blind review

Abstract

In learning-based image compression approaches, compression models are based on variational autoencoder(VAE) framework and optimized by a rate-distortion objective function, which achieve better performance than hybrid codecs. However, VAE maps the input to a lower dimensional latent space which becomes a bottleneck of reconstruction. In this paper, we propose a deep Hierarchical Compression(HC) model, which can achieve good compression performance from low-bit to very high-bit. HC model consists of two closely-related modules, including hierarchical latent compression module and Hierarchical Conditional Entropy(HCE) module. Such a design transmits the details in the shallower layers and coarse information in the deeper layers and conditions the shallower entropy estimation on the deeper information. Extensive experiments show that HC model could breakthrough the AE limit and achieve significant improvements over state-of-the-art approaches in the high quality regime.

1 Introduction

In recent years, end-to-end optimized image compression methods based on variational autoencoder(VAE) have achieved better performance than hybrid codecs. The methods utilize an end-to-end trainable model to jointly optimize the rate and distortion. The state-of-the-art networks for image compression are (Agustsson et al., 2017; Theis et al., 2017; Ballé et al., 2016a;b; 2018; Mentzer et al., 2018; Lee et al., 2019; Minnen et al., 2018a; Cheng et al., 2020; Qian et al., 2020; Hu et al., 2020). A common approach in VAE is to map the images into a lower dimensional latent space in which a probability distribution is learned to allow for entropy coding, such as hyper-prior and context model.

However, the operation of the existing VAE models that mapping an image to a lower dimensional latent space impose an implicit limit on the reconstruction quality. As such, they cannot address high quality levels well. VAE models are limited on the reconstruction quality by the AE Limit line (Helminger et al., 2020).

According to the Nyquist–Shannon sampling theorem (Shannon, 1949), high-frequency signals and low-frequency signals need smaller and larger receptive fields respectively. Shallower layer has a smaller receptive field, which is good for capturing high-frequency information, and the deeper layer has a larger receptive field, which is good for capturing low-frequency signals. Besides, there is a correlation between deep and shallow features. For these reasons, we explore the HC model. The contributions of our work are as follows:

(1)We propose a hierarchical compression framework, the model allows us to transmit details in the shallower layers and coarse information in the deeper layers. Entropy estimation of shallower features are conditioned on deeper features. (2)We propose a residual compression module, which is different from the widely used VAE framework, and could achieve better performance. (3)The experiment results show that HC model breaks the AE Limit, and outperforms the widely used VAE models with lower FLOPs.

The rest of this work is organized as follows. In Section 2, we introduce the key approaches of hierarchical compression, and the experimental results are shown in section 3. Finally, in Section 4, we discuss the current work and future improvements.

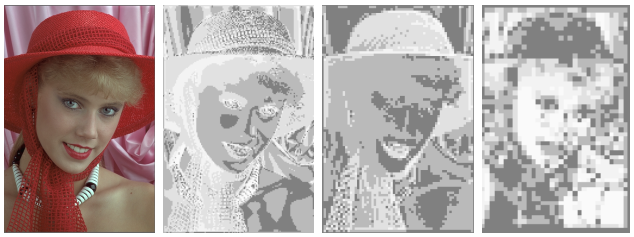

Figure 1: From left to right: original image and latent from 2rd layer to 4th layer.

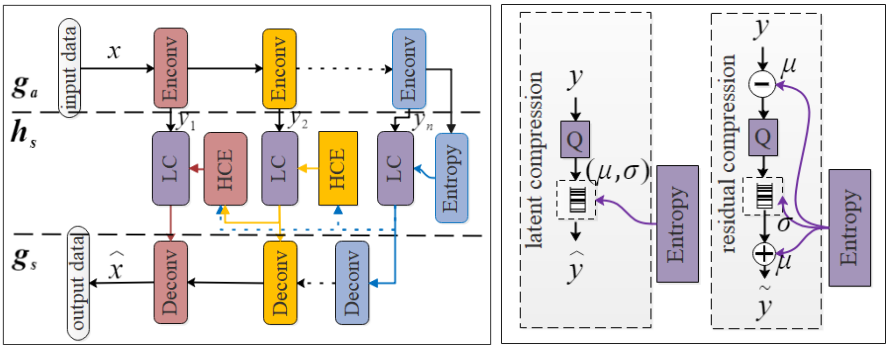

Figure 2: The overall architecture of our method. Left: the HC model. Right: Lossless compression(LC) module, the comparison of latent compression and residual compression.

2 Hierarchical Compression

2.1 Previous Works

Neural multi-scale image compression was first proposed by (Nakanishi et al., 2018), which maps the image to multi-scale latents and estimates the probability by a parallel multi-scale PixelCNN. (Helminger et al., 2020) proposed normalizing flows for lossy image compression and is superior to the existing VAE solutions in the high quality regime but at low bit-rates the performance is much lower.

2.2 Compression Based on Hierarchical Condition

The latent representations in deeper and shallower layers have a strong correlation. We show three latent representations from the 2rd layer to the 4th layer in figure 1. Visually, it is obvious that the features of different layers are similar. Shallower and deeper features store details and structural information respectively.

For this reason, we explore a HC model, as shown in the figure 2(left). Our framework has three fundamental parametric transform functions: an analysis transform $g_a(x;\varphi_g)$, a synthesis transform $g_s(\hat{y};\theta_g)$, and a group of conditional transform $h_s(\hat{y};\theta_h)$.The encoding process is described as follows: on the encoder side, g_a encodes the input image into hierarchical features $y_i, i = 1, 2, ..., n$ which then quantized to form $\hat{y}_i$ and lossless compressed using entropy coding techniques. On the decoder side, an entropy decoder recovers $\hat{y}_i$ from the compressed signal, and g_s recovers the reconstructed image $\hat{x}$ from $\hat{y}$. The entropy coding needs the probability distribution of $\hat{y}$, where $\hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$. We assumed a Gaussian distribution for the Hierarchical Conditional Entropy(HCE), that is,

$$p(\hat{y}_i|\hat{y}_{i+1}, ..., \hat{y}_n) \sim N(\mu_i, \sigma_i) \tag{1}$$

Since there is no prior for $\hat{y}_n$, so a factorized density model (Ballé et al., 2018) is used to estimate the probability distribution. Subsequently, the joint probability $p(\hat{y})$ is represented as the product of the hierarchical conditional distributions:

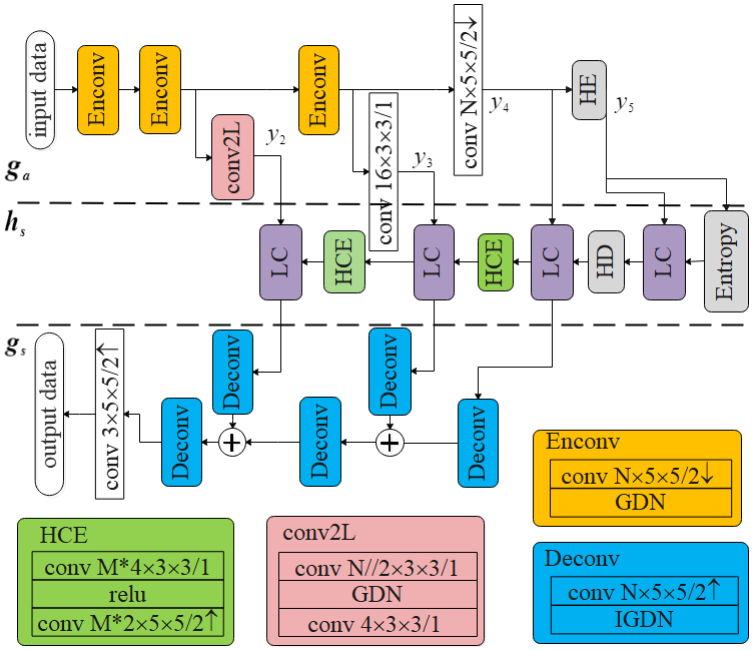

Figure 3: The detailed architecture of HC. Convolution parameters are denoted as number of fileters×kernel height×kernel width / stride, where ↑ and ↓ represent upsampling and downsampling respectively. GDN and IGDN represent generalized divisive normalization and the inverse counterpart respectively (Ballé et al., 2017). LC represents the lossless compression module in figure 2(right). HE and HD module are the same as h_a and h_s in (Ballé et al., 2018).

$$p(\hat{y}) = p(\hat{y}_n) \prod_{i=1}^{n-1} p(\hat{y}_i | \hat{y}_{i+1}, ..., \hat{y}_n) \tag{2}$$

Rate is the expected code length (bit rate) of the compressed representation: assuming the entropy coding technique is operating efficiently, it can be written as a cross entropy:

$$R = \mathbb{E}_{\mathrm{x} \sim px}[-log_2 p_{\hat{y}}] \tag{3}$$

The loss function is defined as below:

$$L = \lambda D + R \tag{4}$$

where D is distortion, λ is a hyperparameter that controls the bpp.

2.3 Residual compression

As shown in figure 2(right), in latent compression, the quantization of the latents y can be expressed as follows: $\hat{y} = round(y)$. This process will inevitably lead to a residual error $r = y - \hat{y}$ in the latent space that manifests as extra distortion when $\hat{y}$ is transformed back into $\hat{x}$. In response to this problem, we use residual compression in our approach to reduce the residual error. The quantization of residual compression can be expressed as follows: $\hat{y} = round(y - \mu) + \mu$, where μ is is the mean value of Gaussian distribution output from the entropy estimation network. We will show in our experiments that residual compression can reduce the residual error, and achieve better performance.

3 Experiments

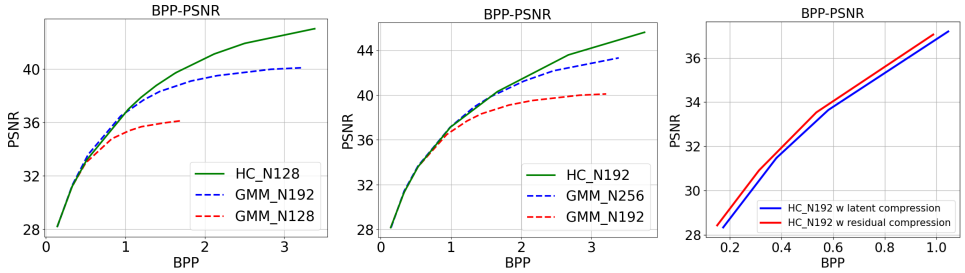

Figure 4: Rate–distortion curves aggregated over the Kodak dataset. The left two plots compare the performance of HC and GMM with different N, and the right plot compares the performance of residual compression and latent compression. In order to see the improvement clearly, we show a shorter BPP range in the right plot.

Table 1: Comparison of calculations (FLOPs)

Model	GMM_N128	GMM_N192	GMM_N256	HC_N128	HC_N192
Encoder	2.36E+09	5.19E+09	9.12E+09	2.70E+09	5.93E+09
Decoder	2.36E+09	5.19E+09	9.12E+09	2.51E+09	5.42E+09

3.1 Experimental Setup

The detailed framework of our approach is shown in figure 3. The backbone is similar to (Ballé et al., 2018) (denoted by GMM in the following). The channel of y_3, y_2, are set to 16, 4 respectively. Entropy estimation of shallow features are conditioned on deeper ones.

For training, we use Imagenet (Deng et al., 2009) as training dataset. The models are optimized using Adam (Kingma & Ba, 2015). For each input min-batch, we randomly crop 8 patches with size of 256×256 from the training dataset. It takes 6 epochs for training. We set the initial learning rate to 10^{-4} and reduced by 0.5 times at the 4th epoch. We optimized our models using mean square error (MSE), and λ belongs to the set $\{0.002, 0.007, 0.015, 0.005, 0.1, 0.15, 0.3, 0.5, 1.5, 3.0\}$

3.2 Experimental Results

In this section, we evaluate the effects of HC model. In the HC model with $N = 128(HC_N128)$, only y_4 is transmitted when $bpp \leq 0.5$, y_3, y_4 are transmitted when $0.5 < bpp \leq 1.0$, and y_2, y_3, y_4 are transmitted when $bpp > 1.0$. In the HC model with $N = 192(HC_N192)$, only y_4 is transmitted when $bpp \leq 1.0$, y_3, y_4 are transmitted when $1.0 < bpp \leq 2$, and y_2, y_3, y_4 are transmiteed when $bpp > 2.0$. Figure 4 compares the RD curves averaged over the Kodak image set. We can see from the left plot that, HC_N128 model is outperforms GMM_N128(GMM with N=128) and GMM_N192. From the middle plot, HC_N192 model outperforms GMM_N192 and GMM_N256. And from the right plot residual compression outperforms latent compression.

Take a 256×256 image as an example, The FLOPs of different models are shown in the table 1. It can be seen that our HC model is more lightweight.

4 Discussion

In this paper, we solved the problem of VAE limition by HC model. We also proposed a latent residual compression model that effectively improves the performance. Experimental results show that our approach can break the limitation of VAE and achieve much better performance at high bit rates with smaller FLOPs.

References

Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc VGool. Soft-to-hard vector quantization for end-to-end learning compressible representations. *Advances in Neural Information Processing Systems*, 2017.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. Density modeling of images using a generalized normalization transformation. *International Conference on Learning Representations*, 2016a.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimization of nonlinear transform codes for perceptual quality. *Picture Coding Symposium*, 2016b.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. *International Conference on Learning Representations*, 2017.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *International Conference on Learning Representations*, 2018.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. *Computer Vision and Pattern Recognition*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition*, 2009.

Leonhard Helminger, Abdelaziz Djelouah, Markus Gross, and Christopher Schroers. Lossy image compression with normalizing flows. *In arXiv e-prints*, 2020.

Yueyu Hu, Wenhan Yang, and Jiaying Liu. Coarse-to-fine hyper-prior modeling for learned image compression. *AAAI Conference on Artificial Intelligence*, 2020.

Diederik P. Kingma and Jimmy Lei Ba. A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. *International Conference on Learning Representations*, 2019.

Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. *Conference on Computer Vision and Pattern Recognition*, 2018.

David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *In Advances in Neural Information Processing Systems*, 2018a.

Ken Nakanishi, Shin ichi Maeda, Takeru Miyato, and Daisuke Okanohara. Neural multi-scale image compression. *In arXiv e-prints*, 2018.

Yichen Qian, Zhiyu Tan, Xiuyu Sun, Ming Lin, Dongyang Li, Zhenhong Sun, Hao Li, and Rong Jin. Learning accurate entropy model with global reference for image compression. *In arXiv e-prints*, 2020.

Claude Elwood Shannon. Communication in the presence of noise. *Institute of Radio Engineers*, 37, 1949.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *International Conference on Learning Representations*, 2017.