Improved Neural Image Compression: A New Two-Stage Quantization Strategy

Zongyu Guo, Runsen Feng, Yaojun Wu, Zhibo Chen
University of Science and Technology of China
guozy@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

Abstract

This one-page paper describes our scheme in the track of image compression. We target at low-rate compression and pursue the optimal MS-SSIM value when the bit-rate is constrained. Our submission IMCL_IMG_MSSSIM achieves 0.96092, 0.97834, 0.98886 MS-SSIM at 0.075, 0.15 and 0.30 bpp during the validation phase. The compression network is almost the same as our solution in the last year but we use some new techniques that improve the MS-SSIM value a lot. In particular, we use a two-stage quantization strategy: soft-then-hard, which is able to eliminate the quantization mismatch between training and test phases. We briefly introduce the ablation results of this quantization strategy and provide some discussions here.

1. Introduction

MS-SSIM-optimized or even MSE-optimized model is able to evaluate the compression ability of networks. We can train a powerful image compression model optimized for MS-SSIM in advance, and then combine it with GAN-based optimization or other perceptual metrics.

In this short paper, we introduce our scheme optimized for MS-SSIM in the track of image compression. The main network structures are almost the same as our previous submission in CLIC2020 [3]. It contains a causal context model that leverages the serial decoding process to conduct separate entropy coding across channels, which is termed as 3-D context model in [3]. We replace the spatial-channel attention by group-separate attention module to enhance the transform network, which is the only difference in terms of network structure. The group-separate attention adopts group convolution to strengthen the attention module that is similar to [6]. In addition, we apply a two-stage quantization strategy termed *soft-then-hard* to eliminate the mismatch between training and test phases.

This new quantization strategy is plug-and-play in all previous compression models that are optimized with additive uniform noise. Although additive uniform noise approximates the quantization error variationally, it introduces stochasticity during training and leads to the train-test mismatch. Some recent works try to close the mismatch by adopting annealing-based quantization method including [5] and [1]. However, according to our analysis in [4], quantization with straight-through estimator or annealing may degrade to optimizing a deterministic autoencoder such as [5] and suffers from some training troubles such as [1]. In contrast, quantization with additive uniform noise is superior in learning an expressive latent space since the noise works as a regularization term to stabilize training.

Soft-then-hard is a two-stage quantization strategy. At the first stage, the compression model is optimized softly with additive uniform noise as usual. At the second stage, we fix the encoder and directly quantize the latent variables y, z with hard rounding. The decoder and the hyper decoder are then finetuned to be optimized for the actual rate-distortion value. The detailed description of this quantization strategy can be found in [4].

2. Some Ablation Results

As shown in table 1, the soft-then-hard quantization strategy improves the MS-SSIM value a lot. We should note that it is different and more effective than the quantization strategy in [2], which sends the discrete latent variable as the input of the decoder but adopts universal quantization for entropy estimation. If optimized for MSE, this strategy can improve performance for more than 0.3 dB even though the main compression model is already powerful enough. The second tuning stage takes 200,000 iterations.

Table 1. The MS-SSIM metric here is different from competition.

Additive Uniform Noise		Soft-then-Hard	
bpp	MS-SSIM	bpp	MS-SSIM
0.0957	0.96024	**0.0920**	**0.96080**
0.1778	0.97538	**0.1726**	**0.97580**
0.3182	0.98478	**0.3117**	**0.98511**

References

[1] Eirikur Agustsson and Lucas Theis. Universally quantized neural compression. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[2] Tiansheng Guo, Jing Wang, Ze Cui, Yihui Feng, Yunying Ge, and Bo Bai. Variable rate image compression with content adaptive optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 122–123, 2020. 1

[3] Zongyu Guo, Yaojun Wu, Runsen Feng, Zhizheng Zhang, and Zhibo Chen. 3-d context entropy model for improved practical image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 116–117, 2020. 1

[4] Zongyu Guo, Zhizheng Zhang, Runsen Feng, and Zhibo Chen. Soft then hard: Rethinking the quantization in neural image compression. *arXiv preprint arXiv:2104.05168*, 2021. 1

[5] Yibo Yang, Robert Bamler, and Stephan Mandt. Improving inference for neural image compression. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[6] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 1