A Practical Approach for Rate-Distortion-Perception Analysis in Learned Image Compression

Ogün Kirmemis, A. Murat Tekalp*
Electrical and Electronics Engineering
Koc University, Istanbul, Turkey
Email: okirmemis16@ku.edu.tr, mtekalp@ku.edu.tr

Abstract—Rate-distortion optimization (RDO) of codecs, where distortion is quantified by the mean-square error, has been a standard practice in image/video compression over the years. RDO serves well for optimization of codec performance for evaluation of the results in terms of PSNR. However, it is well known that the PSNR does not correlate well with perceptual evaluation of images; hence, RDO is not well suited for perceptual optimization of codecs. Recently, rate-distortion-perception trade-off has been formalized by taking the Kullback–Leibler (KL) divergence between the distributions of the original and reconstructed images as a perception measure. Learned image compression methods that simultaneously optimize rate, mean-square loss, VGG loss, and an adversarial loss were proposed. Yet, there exists no easy approach to fix the rate, distortion or perception at a desired level in a practical learned image compression solution to perform an analysis of the trade-off between rate, distortion and perception measures. In this paper, we propose a practical approach to fix the rate to carry out perception-distortion analysis at a fixed rate in order to perform perceptual evaluation of image compression results in a principled manner. Experimental results provide several insights for practical rate-distortion-perception analysis in learned image compression.

Index Terms—learned image compression, PSNR, learned entropy models, perceptual quality evaluation, rate-distortion-perception optimization

I. INTRODUCTION

The third deep learning revolution, which started in the early 2010s, not only changed the course of artificial intelligence and machine learning but also had a serious impact on the signal, image, and video processing. One of the first successful applications of deep learning to image processing was in single-image super-resolution (SISR) [1], which started the new era of learned SISR techniques. Deep learning also had a significant impact on image/video compression enabling end-to-end optimization of image/video codecs. Early works on end-to-end optimized learned codecs considered traditional rate-distortion optimization, while more recent works have started to investigate rate-distortion-perception optimization.

Initial works on deep learned image compression relied on pixel-recurrent RNNs, where Toderici et al. [2] managed to exceed the performance of JPEG. Later, it became clear that better performance can be achieved by end-to-end rate-distortion optimized learned image compression based on jointly learning the parameters of a variational autoencoder [3] and an entropy model. The state-of-the-art in learned image compression has advanced fast after 2018 through annually organized Challenge on Learned Image Compression (CLIC), where the participants were asked to compress images below 0.15 bpp and have better performance than BPG in terms of either peak signal-to-noise ratio (PSNR) or structural similarity index measure (SSIM). There are several advantages of end-to-end optimized learned image compression compared to traditional engineered codecs. First, we are not restricted to using linear and block-based transforms and engineered context models for arithmetic coding. Second, the whole system can be optimized end-to-end according to a desired non-convex (but differentiable) loss function, which can be any combination of $l_1/l_2$ loss, a feature-driven visual distortion loss, such as SSIM, or a no-reference perceptual loss.

Compression algorithms are optimized and evaluated based on Shannon’s rate-distortion framework, where the goal is to minimize the distortion, typically represented by the mean-square error (MSE), at a given bitrate or vice versa. Although this framework helps design codecs that yield the best PSNR performance at a given bitrate, it is well known that PSNR does not correlate well with human perception of image quality since optimizing MSE or PSNR can result in blurry images. The standard approach for perceptual quality evaluation in the image compression community is to conduct subjective tests based on the Mean Opinion Score (MOS). However, this is a time-consuming process, it is not scalable to large datasets, and the results have not been optimized for perceptual quality. With the advent of learned image compression, we can now end-to-end optimize codecs with respect to perceptual distortion losses, such as SSIM [4], or no-reference perceptual losses in addition to MSE. Recent works of Matsumoto [5], [6] and Blau et al. [7] have shown that there exist a triple trade-off between rate, distortion (fidelity), and perception (RDP). However, there exist two important challenges with employing rate-distortion-perception optimization in practice: i) there is no universally agreed objective measure of perceptual quality, ii) there is no practical approach to fix or control rate, distortion, and perception separately in order to employ this framework in learned codecs at the moment.

In this paper, we introduce a practical approach to perform rate-distortion-perception analysis in learned image compression. In particular, we perform perception-distortion analysis at various fixed bitrates sequentially. To this effect, we fix the
bitrate by freezing the training of the encoder and hyperprior networks at successive desired bitrate points. Once the best perception-distortion point is found at a fixed rate, we continue training the encoder and hyperprior networks to reach another rate point. We continue this procedure of finding the best perception-distortion points at certain fixed bitrates until the set of desired rate points are covered. The remainder of the paper is organized as follows: Section III discusses related works. The details of the proposed procedure are presented in Section IV. Experimental results are shown in Section V. Finally, Section VI concludes the paper.

II. Related Work

Ballé et al. [3], [8], [9] laid the foundation of end-to-end optimized image compression using an auto-encoder by introducing joint optimization of rate, distortion, and a probability model. They estimate the rate by the entropy of a non-parametric prior distribution on the latent representation. During the training, they model quantization by adding a uniform distributed noise to deal with the non-differentiability of quantization. Ballé et al. [8] introduced forward adaptation of the entropy model by a learnable scale hyperprior, where they assume the latents are independent Laplace random variables and estimate the scale of the variables through a hyperprior network. Minnen et al. [10] included a backward adaptation step by adding a context model, where they used PixelCNN [11] architecture for pixel-wise context model. In addition to the scale, their network also estimated the mean of latent representation. However, PixelCNN dramatically slows the decoding process, because it gives a single pixel output for a channel at a time. This problem is addressed by the channel-wise autoregressive entropy model [12], which slices the latent representation in the channel dimension and every slice is conditioned on the prior slice. This process is easily parallelizable due to the nature of CNNs; hence, it is much faster than the PixelCNN. These works have considered optimization with respect to MSE or SSIM losses; hence, they did not explicitly address the perception dimension.

Although the works of Matsumoto [5], [6] and Blau et al. [7] have introduced the rate-distortion-perception trade-off framework, they did not propose practical approaches to achieve combined optimization of rate, distortion, and perceptual measures. Mentzer et al. [13] proposed optimization of a combination of rate, MSE, and perceptual distortion measure, together with an adversarial loss in order to encode images at very low bitrates with an acceptable rate-distortion-perception trade-off. However, they employed ad-hoc tricks to keep the rate approximately constant. Prior works on generative image compression include [14] and [15]. One advantage of generative compression is that it can encode images with extremely low bitrates (less than 0.1 bpp) and decode perceptually meaningful images since GANs are able to learn the image manifold and create real looking images [16]. Agustsson et al. [17] propose using conditional adversarial training which helps decoder to identify which regions to hallucinate and which regions to decode faithfully. However, these methods did not consider a formal approach for rate-distortion-perception trade-off and decoded images might not be faithful to the original images by some desired amount. Our main contribution in this paper is to propose a practical approach to fix the rate exactly at the desired value for a proper rate-distortion-perception analysis.

III. Rate-Constrained Optimization of Perception and Distortion

A traditional codec can be optimized for a hard constrained rate-distortion optimization with an exhaustive search of compression parameters. However, their neural network counterparts are trained with soft constrained loss functions. This leads to the same model achieving different rate-distortion points on different images. Ballé et al. optimizes the rate-distortion trade-off by minimizing the following equation:

$$L = d(x, \hat{x}) + \lambda H(y)$$

where $x$ is the original image, $\hat{x}$ is the decoded image, $d(x, \hat{x})$ denotes a distortion function, such as the MSE, and $H(y)$ is the entropy (bitrate) of the compressed representation $y$.

Inclusion of a perceptual loss term to the overall loss function further complicates the optimization process since we need to weigh the three terms with proper weights. There are different approaches to optimizing these quantities.

Approach taken by Agustsson et al. [17] is assuming an upper bound on the entropy of the latent code. When the number of quantization levels is finite and the dimension of the latent code is fixed, the upper bound on the entropy can be estimated by assuming the latent code has independent uniform distributions. It should be noted that this would be a very loose bound since the actual latent code would be neither independently nor uniformly distributed. Perceptual loss is added in the means of adversarial loss where a discriminator $D$ is added to differentiate between the decoded image and the original image. Due to adversarial training, two different loss functions has to be optimized concurrently. These loss functions are:

$$L = d(x, \hat{x}) + \lambda H(y) - \beta \log(D(\hat{x}, y))$$

$$L_D = -\log(1 - D(\hat{x}, y)) - \log(D(x, y))$$

The new terms in (2) come from the discriminator $D$, $\lambda$ and $\beta$ controls the trade-off between rate, distortion and perception. Equation 3 is the standard binary cross-entropy loss used for training the discriminator.

Agustsson et al. [13] also uses the loss functions $L$ and $L_D$, however, with a significant difference: In [17] $\lambda$ is set to 0 and thus the rate is not optimized directly. Instead, as mentioned above the authors rely on the loose upper bound. Since the entropy is not directly optimized, a distribution on the latent code cannot be forced as in the RDO framework provided by [3]. This approach simplifies optimization greatly. On the other hand, [13] uses both loss functions in (2) and (3) with non-zero hyperparameters. However, with this formulation, optimizing for different bitrates is more difficult.
since changing one of the hyperparameters in $\mathbf{2}$ actually changes the relative weight of the other two terms. In order to constrain the rate, authors use two hyperparameters $\lambda^a$ and $\lambda^b$ with $\lambda^a \gg \lambda^b$ to change the weight of the entropy term adaptively. If the target rate is exceeded then $\lambda^a$ is used to bring the model back to the target bitrate, otherwise $\lambda^b$ is used.

Our approach is different from the aforementioned methods since we fix the bitrate by freezing the encoder. First, we train the network with $\mathbf{1}$ to achieve a target bitrate. Since the entropy only depends on the encoder and the entropy model defined by the hyperprior network, the bitrate can be fixed by freezing these two networks. At this point, we can optimize for distortion-perception trade-off directly since we have a single hyperparameter in the loss function:

$$L = \gamma d(x, \hat{x}) + L_p(\hat{x}, x)$$  \hspace{1cm} (4)

where $L_p$ is a generic perceptual loss, $\gamma$ is a hyperparameter controlling the perception-distortion trade-off.

The proposed procedure for Distortion-Perception optimization at a fixed bitrate with our practical approach for fixing the bitrate can be summarized as:
1: Randomly initialize the network
2: Train the network to reach the desired rate-distortion point using $\mathbf{1}$
3: Freeze the weights of the encoder and entropy model estimator sub-networks to fix the encoding rate
4: Start the search for optimal $\gamma$ with the loss function in $\mathbf{4}$
5: Find the $\gamma$ that corresponds to the knee point (see Figures $\mathbf{1}$ and $\mathbf{2}$)

This procedure is repeated for as many different rate points as desired.

IV. EXPERIMENTAL RESULTS

In this section, we provide experimental results to demonstrate that there exists a best perception-distortion trade-off point when encoding at a fixed bitrate. To this effect, we generate multiple decoded images from the same encoded bitstream at a fixed rate by optimizing the decoder network parameters according to different weighting of distortion and perceptual loss functions according to the procedure described above.

Bégaint et al. [18] provides pretrained models of $\mathbf{10}$ for eight different bitrates. We picked the models with the four lowest bitrates since the perception-distortion trade-off is more apparent in the low bitrate regime as suggested by $\mathbf{2}$. The encoder and the entropy models are frozen before training. For the perceptual loss, we employed LPIPS v0.1 [19] with VGG features [20]. VGG features are used widely for perceptual improvement since Ledig et al. [21] showed that the use of VGG loss improves MOS significantly. However, perceptual evaluation is done with LPIPS with AlexNet features [22] since authors state that this network correlates more with MOS. Using a different perceptual metric in test time also provides some impartialness in evaluation. The network is trained with Adam optimizer [23] with a learning rate of $10^{-5}$ and default parameters. We train the network for 400000 iterations on Vimeo-90k dataset [24] with a batch size of 16 on $256 \times 256$ patches.

![Fig. 1. LPIPS vs. PSNR plots of models trained at different fixed bitrates on Kodak dataset. The best perception (LPIPS)-distortion (PSNR) trade-off point at each fixed bitrate are determined as the knee-point of the respective curves.](image1)

![Fig. 2. LPIPS vs. SSIM plots of models trained at different fixed bitrates on Kodak dataset. The best perception (LPIPS)-distortion (SSIM) trade-off point at each fixed bitrate are determined as the knee-point of the respective curves.](image2)
this network MSE only. After freezing the encoder side and hyperprior sub-networks, we start the perception-distortion optimization. The other points on the curves are the results of the experiments with \( \gamma \) values of \( 10^{-2}, 10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}, \) and 0 from right to left, respectively. \( \gamma = 10^{-3} \) improves LPIPS at the cost of a slight decrease in PSNR. The model trained with \( \gamma = 10^{-4} \) still shows some improvement in LPIPS but marginal gain is much lower, also evident from the slope between the points \( \gamma = 10^{-3} \) and \( \gamma = 10^{-4} \). The extreme case of \( \gamma = 0 \) is trained to display the nature of the optimization problem. Since MSE minimizes the pixel-wise differences, the network’s outputs start to differ too much from the input without MSE loss, this becomes harmful for both PSNR and LPIPS. Figure 1 shows the optimal point to stop the hyperparameter search is at the knee point of the curves.

Figure 2 shows the perception-distortion trade-off in terms of LPIPS and SSIM. Although we have not explicitly optimized for SSIM, our observations on Figure 1 also applies here. After reaching the knee of the curve, decreasing \( \gamma \) more harms the network since it only increases distortion without an improvement in perception. Both figures clearly show there is an optimal perception-distortion point for every fixed bitrate.

Figure 3 shows an example image from Kodak dataset compressed with the trained networks. The compressed images are at the same bitrate and have similar distortion levels, however, Figure 3c looks sharper on the grass and the stones in the water. This is evident from the LPIPS scores: (b) scored 0.41 whereas (c) scored 0.35. (Lower LPIPS is better.)

Fig. 3. Example crop from Kodak dataset: (a) original image, (b) output of decoder model optimized with respect to MSE only, (c) output of decoder model with \( \gamma = 5 \times 10^{-4} \). Both images (b) and (c) are reconstructed from the same encoded bitstream at 0.40 bpp. PSNR of (b) and (c) are 24.58 dB and 24.35 dB, respectively. SSIM scores are also similar: 0.7967 and 0.7920, respectively. Although (b) has slightly less distortion in terms of MSE, (c) looks sharper especially on the grass and the stones in the water. This is evident from the LPIPS scores: (b) scored 0.41 whereas (c) scored 0.35. (Lower LPIPS is better.)

Experimental results demonstrate that there exists a best perception-distortion trade-off point at a fixed rate. This paper has two main contributions:

- We propose a practical approach to fix the bitrate of the encoded image at a desired value to perform perception-distortion analysis at a fixed bitrate.
- We propose a principled approach to determine the best perception-distortion trade-off point at a fixed rate.

V. Conclusion

This paper has two main contributions:

- We propose a practical approach to fix the bitrate of the encoded image at a desired value to perform perception-distortion analysis at a fixed bitrate.
- We propose a principled approach to determine the best perception-distortion trade-off point at a fixed rate.

Experimental results demonstrate that there exists a best perception-distortion trade-off operating point for each fixed bitrate. We also propose a framework and a procedure to conduct perception-distortion analysis at a succession of fixed bitrates to achieve the best rate-distortion-perception optimization for learned image compression. In the future, we would like to extend our work to include GAN based networks.

References

[1] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 38, no. 2, pp. 295–307, 2016 (first published in ECCV 2014).

[2] 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 *IEEE/CVF Conf. on Comp. Vis. and Patt. Recog. (CVPR)*, July 2017, pp. 5435–5443.

[3] J. Ballé, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” in *Int. Conf. Learning Representations (ICLR)*, 2017.

[4] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[5] R. Matsumoto, “Introducing the perception-distortion tradeoff into the rate-distortion theory of general information sources,” *IEICE Communications Express*, vol. 7, no. 11, p. 427–431, 2018. [Online]. Available: [http://dx.doi.org/10.1587/comex.2018XBL0109](http://dx.doi.org/10.1587/comex.2018XBL0109)

[6] ——, “Rate-distortion-perception tradeoff of variable-length source coding for general information sources,” *IEICE Communications Express*, vol. 8, no. 2, p. 38–42, 2019. [Online]. Available: [http://dx.doi.org/10.1587/comex.2018XBL0139](http://dx.doi.org/10.1587/comex.2018XBL0139)
Fig. 4. Example crop from Set14 dataset [25]: (a) original image, (b) output of decoder model optimized with respect to MSE only, (c) output of decoder model with $\gamma = 5 \times 10^{-4}$. Both images (b) and (c) are reconstructed from the same encoded bitstream at 0.16 bpp. PSNR of (b) and (c) are 25.14 dB and 25.11 dB, respectively. SSIM scores are 0.8253 and 0.8243, respectively. Although (b) has slightly less distortion in terms of MSE, (c) looks sharper in the area on the left side of the woman’s hand. This is also evident from the LPIPS scores: (b) scored 0.359 whereas (c) scored 0.347. (Lower LPIPS is better.)

[7] Y. Blau and T. Michaeli, “Rethinking lossy compression: The rate-distortion-perception tradeoff,” in eprint arXiv:abs/1901.07821, May 2019.

[8] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” in eprint arXiv:abs/1802.01436, May 2018.

[9] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, “Nonlinear transform coding,” IEEE Jour. Special Topics in Signal Processing, vol. 15, no. 2, pp. 339–353, 2021.

[10] D. Minnen, J. Ballé, and G. D. Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” in Advances in Neural Information Processing Systems 31. Curran Associates, Inc., 2018, pp. 10 771–10 780.

[11] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, “Conditional image generation with pixelcnn decoders,” in eprint arXiv:abs/1606.05328, 2016.

[12] D. Minnen and S. Singh, “Channel-wise autoregressive entropy models for learned image compression,” in eprint arXiv:abs/2007.08739, 2020.

[13] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, “High-fidelity generative image compression,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[14] O. Rippel and L. Bourdev, “Real-time adaptive image compression,” in eprint arXiv:abs/1705.05821, 2017.

[15] S. Santurkar, D. Budden, and N. Shavit, “Generative compression,” in eprint arXiv:abs/1703.01467, 2017.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[17] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, “Generative adversarial networks for extreme learned image compression,” in eprint arXiv:abs/1804.02958, 2019.

[18] J. Bégain, F. Rapacé, S. Feltman, and A. Pushparaja, “Compressai: a pytorch library and evaluation platform for end-to-end compression research,” arXiv preprint arXiv:2011.03029, 2020.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in IEEE/CVF Conf. Comp. Vision and Patt. Recog. (CVPR), June 2018.

[20] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in International Conference on Learning Representations, 2015.

[21] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in eprint arXiv:abs/1609.04802, May 2017. [Online]. Available: https://arxiv.org/abs/1609.04802

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in 3rd Int. Conf. on Learning Representations, ICLR, May 2015.

[24] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” International Journal of Computer Vision (IJCV), vol. 127, no. 8, pp. 1106–1125, 2019.

[25] R. Zeyde, M. Elad, and M. Protter, “Single image scale-up using sparse-representations,” in Curves and Surfaces. Springer, 2012, pp. 711–730.