Perceptual Image Compression using Relativistic Average Least Squares GANs

Zhengxue Cheng, Ting Fu, Jiapeng Hu, Li Guo, Shihao Wang, Xiongxin Zhao, Dajiang Zhou, Yang Song
Ant Group, Xihu District, Hangzhou, China.
{zhengxue.czx, ft224739, libby.hjp, li.gl, shihao.wsh, xiongxin.zxx, dajiang.zdj, zhaoshan.sy}@antgroup.com

Abstract

In this work, we provide a detailed description on our submitted methods ANTxNN and ANTxNN\_SSIM to Workshop and Challenge on Learned Image Compression (CLIC) 2021. We propose to incorporate Relativistic average Least Squares GANs (RaLSGANs) into Rate-Distortion Optimization for end-to-end training, to achieve perceptual image compression. We also compare two types of discriminator networks and visualize their reconstructed images. Experimental results have validated our method optimized by RaLSGANs can achieve higher subjective quality compared to PSNR, MS-SSIM or LPIPS-optimized models.

1. Introduction

Image compression is a fundamental topic in the field of image processing for many decades owing to its significant effect to transmission and storage. Many representative compression standards have been widely used in industries, such as JPEG [1], JPEG2000 [2], High Efficiency Video Coding (HEVC/265) [3] and latest Versatile Video Coding (VVC/266) [4]. These standards have achieved impressive coding efficiency with more and more complex hand-engineered coding tools. Although the computational complexity is significantly increasing, further improvement on coding efficiency becomes more and more difficult relative to its predecessor.

Recent learned image compression has made tremendous progress and has achieved impressive coding efficiency regarding well-known PSNR and MS-SSIM quality metrics in [5, 6, 7, 8, 9, 10] etc. One main advantage of learned image compression is that it is flexible to be adapted to any quality metrics as long as this quality metric is differentiable. Different quality metrics leads to different reconstruction artifacts. Some of artifacts are easily captured by human visual system such as blurring, while some of them are not. For instance, LPIPS [11] is a neural-network based quality metric, which is found to correlate much better with human evaluations in CLIC 2020 low-rate image compression track [12] compared with PSNR and MS-SSIM. Generative adversarial models also exhibit a superior performance in terms on image subjective quality. However, GANs sometimes lead to the unstable training process in rate-distortion optimization problem. In [13] and [14], authors either used a frozen learned encoder or a fixing VVC encoder and only learned the decoder using VGG-19 and ESRGAN losses. In [15], selective detail decoding was realized by an additional selective decoder and stable training could be achieved by using a softplus-based discriminator loss function. In [16], authors present very competitive perceptual qualities, and found MSE is likely to be the driving force for stability, and jointly learning encoder without MSE led to a collapse of GAN training. However, there is a large room to explore the discriminator architectures and GAN training strategies.

This work presents a detailed description on our submitted methods ANTxNN and ANTxNN\_SSIM to Workshop and Challenge on Learned Image Compression (CLIC) 2021. The network architecture combines recent techniques, including split attention and Gaussian mixture models. To achieve a higher perceptual quality, we apply Relativistic Average Least Squares GANs to achieve a stable rate-distortion. We also incorporate LPIPS, a learned quality metric as one of distortion metrics. Experimental results demonstrate our models can achieve higher subjective quality than optimizing LPIPS, or MSE or MS-SSIM only.

2. Related Work

Learned Image Compression Related studies have validated image compression can be formulated as a rate-distortion optimization problem and then solved benefiting recent deep learning techniques in [5, 6, 7, 8, 9, 10] etc. Some other works explored the approximations of non-differentiable quantization, such as uniform noise [17], straight-through proxy [18], soft quantization [19], universal quantization [20], for the better gradient back-propagation. Some works proposed different network struc-
3. Proposed Method

3.1. Problem Formulation and Network

Learned image compression is formulated as a Lagrangian multiplier-based rate-distortion optimization problem. This problem can be solved by minimizing its loss function:

\[
\mathcal{L} = \lambda \times \mathcal{D}(x, \hat{x}) + \mathcal{R}(\hat{y}) + \mathcal{R}(\hat{z})
\]

where \(\lambda\) controls the rate-distortion trade-off. Different \(\lambda\) values are corresponding to different bit rates. \(\mathcal{R}(\hat{y})\) and \(\mathcal{R}(\hat{z})\) are rate terms, calculated by \(\mathbb{E}[-\log \mathcal{N}(\hat{y} | \hat{z})]\) and \(\mathbb{E}[-\log \mathcal{N}(\hat{z} | \psi)]\). The probabilities are further calculated by

\[
p_{\hat{y} | \hat{z}}(\hat{y} | \hat{z}) = \prod_{i} \mathcal{N}(\hat{y}_i | \mu_i^{(k)}, \sigma_i^{2(k)}) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2})(\hat{y}_i)
\]

\[
p_{\hat{z} | \mu_i, \sigma_z}(\hat{z} | \mu_i, \sigma_z) = \prod_{j \in \mathbb{N}} \mathcal{N}(\mu_{z_i}^{(k)}, \sigma_z^{2}) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2})(\hat{z}_j)
\]

where \(y_i, z_i\) denote the \(i\)-th element of \(y\) and \(z\). The likelihoods of \(y_i\) are estimated by predicted \(w_i^{(k)}, \mu_i^{(k)}\) and \(\sigma_i^{2(k)}\), which are parameter of Gaussian Mixture Models (GMM). Typical \(K\) is 3. The entropy models for \(\hat{z}\) is a channel-wise single Gaussian model given by a pair of mean \(\mu_z\) and scale \(\sigma_z\) with the size of \([N, 1]\), where \(j\) denotes the index of channels, \(N\) denotes the number of total channels. \(\mathcal{D}(x, \hat{x})\) denotes the distortion term and will be discussed in next section.

The network architecture we used is shown in Fig. 1. We follows the latest learned image compression approach with a hyperprior and autoregressive model, and uses a Gaussian mixture models (GMMs) with \(K\) mixtures [8, 9], which are reported to have state-of-the-art performance. But, we replace the attention module by grouped-separated attention block in [32] to further enhance the performance. This block can enable separate groups attention and improve the PSNR by about 0.1dB at the same bit rate. All the parameters in Fig. 1 are end-to-end learned to achieve the best rate-distortion optimization.
Table 1: Two different types of discriminators, where all the layers use leakyReLU with alpha of 0.2 as activation functions except for the last layer with no activation.

<table>
<thead>
<tr>
<th>disc_{4,4}</th>
<th>disc_{2B}</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y \in \mathbb{R}^{122 \times 122 \times 192}$</td>
<td></td>
</tr>
<tr>
<td>Conv 3 x 3, stride 1, 12</td>
<td></td>
</tr>
<tr>
<td>UpSampling2D (size=(16, 16))</td>
<td></td>
</tr>
<tr>
<td>Concat(x, features), $x \in \mathbb{R}^{192 \times 192 \times 4}$</td>
<td></td>
</tr>
<tr>
<td>Conv 4 x 4, stride 2, 64</td>
<td>RB 3 x 3, stride 2, 64</td>
</tr>
<tr>
<td>Conv 4 x 4, stride 2, 128</td>
<td>RB 3 x 3, stride 2, 128</td>
</tr>
<tr>
<td>Conv 4 x 4, stride 2, 256</td>
<td>RB 3 x 3, stride 2, 256</td>
</tr>
<tr>
<td>Conv 4 x 4, stride 2, 512</td>
<td>RB 3 x 3, stride 2, 512</td>
</tr>
<tr>
<td>Conv 4 x 4, stride 2, 512</td>
<td>RB 3 x 3, stride 2, 512</td>
</tr>
<tr>
<td>Flatten</td>
<td>Flatten</td>
</tr>
<tr>
<td>Dense layer, 1024</td>
<td>Dense layer, 1024</td>
</tr>
<tr>
<td>Dense layer, 1</td>
<td>Dense layer, 1</td>
</tr>
</tbody>
</table>

3.2. Least Squares Relativistic Generative Adversarial Learning

As Fig. 1, we add a discriminator network to tell the difference between raw images $x$ and compressed images $y$ are also fed into the discriminator network as the conditional information, similar to the design of conditional GANs (CGANs). We tried two designs of discriminators, as listed in Table 1. Different discriminator structures lead to different artifacts, which will be visualized in Section 5.

Then we further explore different types of GAN losses, including standard GAN, HingeGAN, WGAN and LS-GAN. Experimental results show LSGAN [33] shows the most stable performance for end-to-end training. So we use LSGAN. Beside, we use the relative average GAN by assuming the real samples should be better than average fake ones, and vice versa. Denote the $C(x)$ and $C(\hat{x})$ are the output of discriminator with the input of raw images $x$ and compressed images $\hat{x}$. The relative output is computed as:

$$C(x) = C(x) - E(C(\hat{x}))$$

$$\frac{C(x)}{C(\hat{x})} = \frac{C(x) - E(C(\hat{x}))}{E(C(\hat{x}))}$$

(4)

Then the Rate-distortion loss in Eq. (1) is incorporated into the generator loss. The final generator losses and discriminator losses are listed as

$$L_g = \mathbb{E}[(C(x) - 1)^2] + \mathbb{E}[(C(\hat{x}) + 1)^2] \ast \beta + \lambda \ast D(x, \hat{x}) + R(y) + R(\hat{z})$$

$$L_d = \mathbb{E}[(C(x) - 1)^2] + \mathbb{E}[(C(\hat{x}) + 1)^2]$$

(5)

where $\beta$ defines the weight of generator, and is set as 0.015 in our experiments. The distortion term is a combination of MSE and LPIPS, as

$$D(x, \hat{x}) = \text{MSE} + k_p \ast \text{LPIPS}$$

(6)

where $k_p$ controls the weight of LPIPS and $k_p$ is set as $10^3$ in our experiments. $\lambda$ controls the rate and will be given in Section 4.

4. Implementation Details

Training Details For training, we used a 2K resolution high-quality DIV2K dataset [35] with 800 images to avoid over-fitting to JPEG compressed artifacts and distributions. We cropped them into samples with the size of 192 x 192 x 3, instead of 256 x 256 to reduce the training time. The number of channels $N$ in Figure 1 is set as 192.

Our submitted method ANTxNN_SSIM is optimized by only MS-SSIM quality metric [36], i.e. $D(x, \hat{x}) = 1 - \text{MS-SSIM}(x, \hat{x})$. When optimized by MS-SSIM, $\lambda$ belongs to the set {1, 2, 5, 10, 6, 12, 16, 20, 24} to get five models. They were trained using the Adam [37] algorithm with a batch size of 8. The learning rate was kept at a fixed value of $1 \times 10^{-4}$ for 500k iterations, and was reduced to $1 \times 10^{-5}$ for the last 80k iterations. Then the bit allocation is solved as a knapsack problem to push the total bits to the targeted bpp.

Our submitted method ANTxNN is optimized by a combined distortion metrics, including LPIPS, mean square error (MSE) and GAN loss. $\lambda$ belongs to the set {0, 10, 50, 0.45, 0.5, 0.575, 1, 2, 4, 6, 8} to get eight models. Both discriminator and generator were trained using the Adam [37] algorithm with a batch size of 8. All the parameters are firstly initialized by a model trained for 500k iterations by MSE only. Then, the learning rate was kept at a fixed value of $1 \times 10^{-4}$ for 300k iterations, and was reduced to $1 \times 10^{-5}$ for the last 80k iterations. Learning rate decay can train the models to the convergence.

Textual Region Enhancement Due to the special attention of human visual system on textual regions, we used a simple OCR detection algorithm to get the enhanced mask and merge sub-masks into a whole region of interest (ROI). Then we crop the ROI from reconstructed images and encode the residual image using VVC [39] encoder, as a ROI-based enhancement.

Quantized rate computation We apply layer-wise fake quantization after each convolution in a hyper decoder and autoregressive model with a fixed encoder and decoder. And, we also apply fake quantization after the softmax operation to push the total bits to the targeted bpp.
In our experiments, $q_{\mu}$ is set as 0.2, $q_{\sigma}$ is set as 0.1 and $q_{w}$ is set as 0.0001. Quantized rate computation leads to less than 3% rate increment. But, quantized rate computation can avoid the floating-point calculation errors and improve the computation stability.

5. Results

In this section, experimental results and ablation studies are present. First, we have shown the results of our submitted methods on CLIC validation dataset in Table. 2 and Table 3. $ANTxNN$ ranks the 2-nd in terms of MS-SSIM. $ANTxNN$ ranks the 6-th, 2-nd and 3-rd in terms of FID at the rate of 0.075bpp, 0.15bpp and 0.30bpp, separately.

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR</th>
<th>MS-SSIM</th>
<th>FID</th>
<th>Rate (bpp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ANTxNN$</td>
<td>33.726</td>
<td>0.97748</td>
<td>144.509</td>
<td>0.30</td>
</tr>
<tr>
<td>$ANTxNN$</td>
<td>31.315</td>
<td>0.96098</td>
<td>158.102</td>
<td>0.15</td>
</tr>
<tr>
<td>$ANTxNN$</td>
<td>27.465</td>
<td>0.91785</td>
<td>192.697</td>
<td>0.075</td>
</tr>
</tbody>
</table>

Table 3: Results of our submitted methods $ANTxNN_{SSIM}$.

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR</th>
<th>MS-SSIM</th>
<th>FID</th>
<th>Rate (bpp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ANTxNN_{SSIM}$</td>
<td>33.726</td>
<td>0.97748</td>
<td>144.509</td>
<td>0.30</td>
</tr>
<tr>
<td>$ANTxNN_{SSIM}$</td>
<td>31.315</td>
<td>0.96098</td>
<td>158.102</td>
<td>0.15</td>
</tr>
<tr>
<td>$ANTxNN_{SSIM}$</td>
<td>27.465</td>
<td>0.91785</td>
<td>192.697</td>
<td>0.075</td>
</tr>
</tbody>
</table>

Figure 2: Comparison of R-D curves using LPIPS quality metric on the Kodak [38] dataset.

Besides, ablation study on rate-LPIPS rate curves is shown in Fig. 2. We can observe that the models optimized by MSE or MS-SSIM can not achieve good LPIPS values. By incorporating LPIPS into the loss function, the perceptual quality of reconstructed images can be significantly improved. We also compare our results with recent perceptual-optimized compression approach [16]. It can be seen our method achieves lower LPIPS performance compared to [16]. Models optimized by LPIPS+MSE achieved the lowest LPIPS, and adding GAN slightly increases the value of LPIPS.

Figure 3: Visualization of different artifacts.

Fig. 3 visualize the compressed artifacts by different methods using the same network, i.e optimizing by MSE, MS-SSIM, LPIPS+MSE, LPIPS+MSE+GAN. We can observe adding discriminator can improve the perceptual quality significantly than optimizing LPIPS, MSE or MS-SSIM only. Since $4 \times 4$ discriminator is slightly better than discriminator with residual block, so our submitted method $ANTxNN$ used $4 \times 4$ discriminator.

6. Conclusion

In this paper, we have described our methods $ANTxNN$ and $ANTxNN_{SSIM}$, which are submitted to challenge on learned image compression (CLIC) 2021. RaLSGANs are incorporated into rate-distortion optimization stably to improve perceptual quality of reconstructed images than other GAN losses. We also discuss two types of discriminator networks. Besides, visualization comparisons have demonstrated RaLSGANs-optimized compression model can generate higher subjective quality than optimizing LPIPS, PSNR, or SSIM only.
References


[38] Kodak Lossless True Color Image Suite, Download from http://r0k.us/graphics/kodak/ 4