Attentional Multi-scale Back Projection Network for Low Bit-rate Image Compression

Ge Gao, Rong Pan, Pei You, Shunyuan Han, Yuanyuan Zhang, Hojae Lee
Samsung R&D Institute China Xi’an
{ge1.gao,rong.pan,pei.you,shuny.han,yuan2.zhang,hojae72.lee}@samsung.com

Abstract

In this paper, we provide the description of our approach for participating Workshop and Challenge on Learned Image Compression (CLIC) 2021: SRCX_DLIC. Our method is an end-to-end network based on variational autoencoder, for which we develop a novel back projection method with attentional and multi-scale feature fusion. Our back projection method recalibrates the current estimation by establishing feedback connections between high-level and low-level attributes in a discriminative and efficient manner. Further, our network recovers the fine spatial details by decomposing the input image and separately processing the distinct frequency components, whose derived latents are recombined using a novel dual attention module, so that details inside regions of interest could be explicitly manipulated. We also describe the training strategies for variable bit-rates and perceptual quality enhancement.

1. Introduction

Lately, the demand for image compression has increased dramatically to cope with the enormous amount of high-resolution images. Based on deep neural networks (DNNs), neural image compression has reinvigorated this domain with its superb capacity to learn in a data- and metric-driven manner, as opposed to their conventional counterparts [13].

Neural image compression typically employs autoencoders to model image down-sampling and up-sampling as a unified task and optimize the rate-distortion trade-off jointly. Such methods map the input image into a latent intermediate via an encoder and inversely transform the quantized latent back to generate the reconstructed image on the decoder side. Many researches concentrate on optimizing the network architecture, e.g., GDN [4], residual blocks [22, 18], RNNs [23, 16, 24], to facilitate both decorrelation and recovery of the image signal. Meanwhile, some other works focus on further reducing the entropy of the latent representations to attain fewer encoding bits. Earlier works [6, 22] in this respect incorporated elementwise entropy models to encode each element independently. Later advancements introduced hierarchical hyperprior networks [5] and autoregressive components [14, 20] into the VAE framework to explicitly estimate the entropy of the latent representation by utilizing prior information. Currently, the rate-distortion performance of the state-of-the-art methods have surpassed that of reigning compression codecs, such as BPG [7] and VVC [21], on both PSNR and MS-SSIM.

Nonetheless, existing schemes are limited in capturing the mappings between the source image and its compact form, leading to over-smoothed reconstructions at low compression rates. One major issue is that, while the autoencoder excels at extracting contextualized, non-linear information for effective decorrelation, it stumbles in preserving spatial image details that are crucial to faithful reconstruction. In addition, the fact that the input image is usually processed in its RGB format, in which these easily-lost high-frequency details are mingled with large-scale variations, makes it even harder for the network to preserve or infer fine-grained details for optimal reconstruction.

In this paper, to enable mutual facilitation between low- and high-level image properties, we replace the standard feedforward up-sampling layers with a novel Attentional Multi-scale Back Projection (AMBP) module. Our AMBP module aggregates intermediate features from higher to lower layers of the network, allowing it to attain semantically-rich features, on the one hand, and extrapolate fine spatial details, on the other. Retaining the desired properties of both gives the network a greater flexibility to decide which information should be preserved for better rate-distortion trade-off. Our simplified design discards the iterative up- and down-sample procedure [11, 17] by including both in-scale and cross-scale feature fusion within one back projection step. To extract richer visual representations, we further leverage a soft attention mechanism that consolidates the input feature maps in a weighted average fashion.

Moreover, we propose to decompose the original image by extracting and processing its frequency components separately so that the network could yield further efficiencies in representation by exploiting different pieces of information
that contains distinctive frequency characteristics. Besides, to enhance the perceptual quality, we finetune the model by adding lpips [25] to the loss function, which enforces the network to focus on enhancing perceptual quality.

2. Proposed Method
2.1. Network Architecture

The network architecture is shown in Fig.1. The encoder side of our design consists of a decomposition module, a dual-branch encoder and a dual attention module. Instead of processing the input image in its RGB form, we extract its low- and high-frequency components and compress them separately using the dual-branch encoder, in which each branch contains four down-sampling modules/blocks and two spatial attention modules/blocks in between [9]. The down-sampled latents of the frequency components are then rescaled and combined into the complete latent representation $\hat{y}$ via the dual attention module. The hyperprior model and the context model follows the same design as [5]. The single-branch decoder consisting of four AMBP module that up-sample the quantized latent $\hat{y}$ into the reconstruction image. We further adopt dense connections where the current AMBP module process the concatenation of outputs of all previous modules [11]. Gain and Inverse-Gain refers to the gain units used to accomplish variable bit-rates in a single model [10].

2.2. Attentional Multi-scale Back Projection

Back projection was first put forward in DBPN [11] for image super-resolution. The back projection technique iteratively utilizes the feedback residual to refine HR images, based on the assumption that the projected, down-sampled version of a SR image should be as close as possible to the original LR image. We extend the similar idea to image compression task and construct our building blocks entitled as the AMBP modules.

Figure 2. Illustration of the back projection procedure using feature maps sampled from the decoder when reconstructing *kodim21.png*. The updated HR map $\hat{H}_{t+1}$ contains better defined details than the initial HR map $H_{t+1}$.

Convolution layers of the autoencoder trade fine spatial details for copious semantic information through the repeated use of down-sampling operations, making it less reliable for faithful image reconstruction. To address this issue, AMBP aggregates multi-scale features across stages in a trainable way. That is, the current stage features are consolidated by the complementary information (spatially accurate) from later computations. The refined feature
maps in turn produce features of higher quality in the next stage, thereby achieving progressive improvement to the intermediate features that propagate throughout the computation.

As shown in Fig. 2, our AMBP module refines a HR map \( H_{t+1} \), up-sampled from \( H_t \), by applying reverse mapping to recover its original resolution. Despite having the same resolution, the re-sampled feature map \( Y_t \) encloses details from \( H_{t+1} \) that are then integrated into \( H_t \) using a fusion module \( R_b \), and the updated LR map is up-sampled again to yield a re-sampled HR map \( Y_{t+1} \). To facilitate in-scale feature fusion, we leverage another fusion module \( R_a \) that aggregates \( H_{t+1} \) and \( Y_{t+1} \) to update the later as \( H_{t+1} \) that contains more detailed features. The described process can be written as:

\[
\begin{align*}
Y_t &= \downarrow_s (H_{t+1}) = \downarrow_s (\uparrow_s (H_t)) \\
H_t &= R_b(H_t, Y_t) \\
Y_{t+1} &= \uparrow_s (H_t) \\
H_{t+1} &= R_a(H_{t+1}, Y_{t+1}).
\end{align*}
\]

The feature fusion is based on residual calculation. As shown in Fig. 1, the residual fusion module \( R_b \) aggregates \( H_t \) and \( Y_t \) according to their residual \( e_t \). Intuitively, the residual \( e_t = H_t - Y_t \) represents distinctive information available in one source while missing in the other. The key modification we make is to incorporate another attentional fusion module \( R_a \) for the \((H_{t+1}, Y_{t+1})\) pair. We leverage channel attention [12] to capture the channel-wise dependencies of the residual \( e_{t+1} = H_{t+1} - Y_{t+1} \) at a global scale, and adopt a soft fusion scheme that reweights the respective inputs by \( W \) and \((1 - W)\), where \( W \) is the normalized attention map. The residual fusion modules is formulated as:

\[
\begin{align*}
R_b(H_t, Y_t) &= \mathcal{R}(Y_t + \mathcal{R}(e_t)) \\
W &= S(\mathcal{R}(e_{t+1})) \\
R_a(H_{t+1}, Y_{t+1}) &= \mathcal{R}(W \odot H_{t+1} + (1 - W) \odot Y_{t+1}),
\end{align*}
\]

where \( \mathcal{R} \) denotes residual blocks, \( W \) denotes the attention map, \( S \) denotes channel attention, and \( \odot \) denotes element-wise multiplication.

The benefits of the additional feature fusion operations are threefold. First, it further optimizes the re-sampled feature map \( Y_{t+1} \) and facilitates both in-scale and cross-scale feature fusion without iterations. Second, the proposed soft content selection scheme efficiently enables more adaptive feature fusion by implicitly re-guarlizing the values of the attention map. Third, fusion based on residual allows the network to focus only on distinctive information, making the gradient update better guided, and incorporating another residual-based fusion operation could further accelerate the training procedure.

### 2.3. Decomposition

“High-frequency content will get lost during sample rate conversion”, as pointed out by the Nyquist-Shannon Sampling Theorem. Decomposition could help alleviate this issue by enabling the network to explicitly manipulate the high-frequency contents.

As illustrated in Fig. 3, the low-frequency components across scales are obtained using average pooling with various kernel sizes. The high-frequency components are attained by subtracting the corresponding low-frequency component from the input image \( x \). To produce the base layer \( x_b \), we pass the concatenated low-frequency components to a residual block. The detail layer \( x_d \) containing high-frequency information is attained in a similar way. As the original image \( x \) also contains rich information, it is concatenated with \( x_b \) and \( x_d \) and then processed separately by the dual-branch encoder, which progressively down-samples the respective components into their latent representations \( y_{lf} \) and \( y_{hf} \).

Afterwards, the latents \( y_{lf} \) and \( y_{hf} \) are aggregated using a dual attention module (Fig. 4). The latents of the respective components are concatenated along the channel dimension to produce feature map \( F \), which is then transformed by a residual block and passed to the channel and the spatial attention.
attention module. To reduce computation, the spatial attention module independently applies global average pooling and global max pooling to \( F \) along the channel dimension and concatenates the results to form feature map \( F_x \in \mathbb{R}^{H \times W \times 2} \), from which is the spatial attention map \( W_c \in \mathbb{R}^{H \times W \times 1} \) extracted. The channel attention feature map \( W_c \in \mathbb{R}^{1 \times 1 \times C} \) is generated using SE blocks [12]. We adopt the soft selection trick to improve representations as well. The low-frequency latent \( y_{l,f} \) is rescaled by \( W_c \) and then \( W_c \) while the high-frequency latent \( y_{h,f} \) is rescaled by \((1-W_c)\) and then \((1 - W_c)\). The re-weighted latents are then summed to yield the final latent representation \( y \).

2.4 Variable Bit Compression

Considering the bit-rates constraint, we adopt the gain units [10] to achieve continuously variable rate with a single model. As shown in Fig.1, the gain unit and the reverse gain unit are added after the encoder and before the decoder, respectively. The pairs are inserted into the hyperprior model in a similar manner.

2.5. Finetune Strategies

Enhancing the perceptual quality of our method is imperative, as the challenge performs evaluation based on human perception. Thus, we finetune the decoder \( D \) in our baseline model by incorporating the up-to-date perceptual loss - LPIPS [25] to improve the perceptual quality. Besides, we notice that the reconstruction quality could be considerably enhanced by decoding the original rather than quantized latent without tuning, so we added a rounding loss \( \hat{d} = MSE(y, \hat{y}) \) to the loss function, where \( \hat{y} \) is the updated latent map by the residual fusion module. Intuitively, variations of a pixel’s neighboring pixels in the high-resolution map could help the network infer the pixel’s value before quantization. The loss function for finetuning becomes:

\[
L_f = d(x, \hat{x}) + \beta \cdot d_x + \gamma \cdot d_{l_{pips}},
\]

where \( \beta \) and \( \gamma \) controls the weight of the rounding loss term and perceptual loss term.

3. Implementation Details and Results

We trained the proposed networks using cropped images of size 256x256 from DIV2K [2], Flickr2K [15], and CLIC training dataset [1] without augmentation. Formulating \( E \), \( D \) and \( P(y) \) in our network allows them to be trained jointly by minimizing the rate-distortion trade-off:

\[
\mathcal{L} = \sum_{i=0}^{n} \lambda_i \cdot d(x, \hat{x}) + R,
\]

where \( d \) represents distortion, \( R \) represents the required number of bits, and \( \lambda_i \) represents the index of the gain vectors in the gain matrix [10]. We used the Adam algorithm to jointly optimize the networks for 1.2M iterations with a mini-batch size of 4. The initial learning rate was set to \( 1 \times 10^{-4} \) and decreased to \( 5 \times 10^{-5} \) at 800k iterations. The networks were optimized with respect to two quality metrics, i.e., mean square error (MSE) and multi-scale structural similarity index (MS-SSIM). The distortion \( d \) is defined as \( d = MSE(x, \hat{x}) \) and \( d = 1 - MS-SSIM(x, \hat{x}) \), respectively. When optimized by MSE, the value of \( \lambda \) belongs to the set \( \{0.0004, 0.0025, 0.009\}, \{0.0009, 0.0085, 0.02\}, \{0.005, 0.009, 0.045\} \) for the three different bit-rates, respectively. As to networks optimized for MS-SSIM, the value of \( \lambda \) belongs to the set \( \{1, 4, 9\}, \{2, 10, 32\}, \{10, 35, 120\} \) for the three different bit-rates, respectively. After that, we finetuned the decoder for 50k iterations at the learning rate \( 5 \times 10^{-5} \), the coefficient \( \beta \) and \( \gamma \) was set to 1 and 100 for MSE optimized model.

The results on CLIC validation dataset are summarized in Table 1, we could attain the bit-rates close to the constraint value through the gain unit. It can be seen that, while MSE optimized models attain higher PSNR than MS-SSIM optimized models, they decode images of lower MS-SSIM in comparison. Further, finetuned models achieve higher perceptual scores (in this case, the lower the lpips the better), which suggests the effectiveness of our strategies in enhancing the perceptual quality of reconstructed images.

<table>
<thead>
<tr>
<th>Method</th>
<th>BPP</th>
<th>PSNR</th>
<th>MSSIM</th>
<th>LPIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.0732</td>
<td>27.836</td>
<td>0.9279</td>
<td>0.1676</td>
</tr>
<tr>
<td>Optimized</td>
<td>0.1481</td>
<td>30.825</td>
<td>0.9605</td>
<td>0.0946</td>
</tr>
<tr>
<td></td>
<td>0.2964</td>
<td>34.014</td>
<td>0.9783</td>
<td>0.1059</td>
</tr>
<tr>
<td>Finetuned</td>
<td>0.0732</td>
<td>26.866</td>
<td>0.9513</td>
<td>0.1602</td>
</tr>
<tr>
<td>MSE</td>
<td>0.1481</td>
<td>29.124</td>
<td>0.9694</td>
<td>0.0927</td>
</tr>
<tr>
<td>Optimized</td>
<td>0.2964</td>
<td>30.355</td>
<td>0.9799</td>
<td>0.1013</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.0743</td>
<td>26.375</td>
<td>0.9555</td>
<td>-</td>
</tr>
<tr>
<td>Optimized</td>
<td>0.1475</td>
<td>28.621</td>
<td>0.9718</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.2974</td>
<td>29.819</td>
<td>0.9813</td>
<td>-</td>
</tr>
</tbody>
</table>

4. Conclusion

In this paper, we propose a neural image compression scheme using back projection techniques and frequency decomposition. We reformulate the iterative projection operations into a multi-scale feature fusion module and incorporate channel attention with soft content selection. We also enable the network to focus on respective frequency components of the input image via decomposition, where their derived latents are adaptively rescaled and integrated using an efficient dual attention module. Further, we adopt a finetune strategy that helps enhance the perceptual quality and reduce the latent residual.
References


