RATE-DISTORTION-OPTIMIZATION FOR DEEP IMAGE COMPRESSION

Fraunhofer
HHI

SVCP 2022
Ilmenau
04.07.2022

Sophie Pientka

CONTENTS

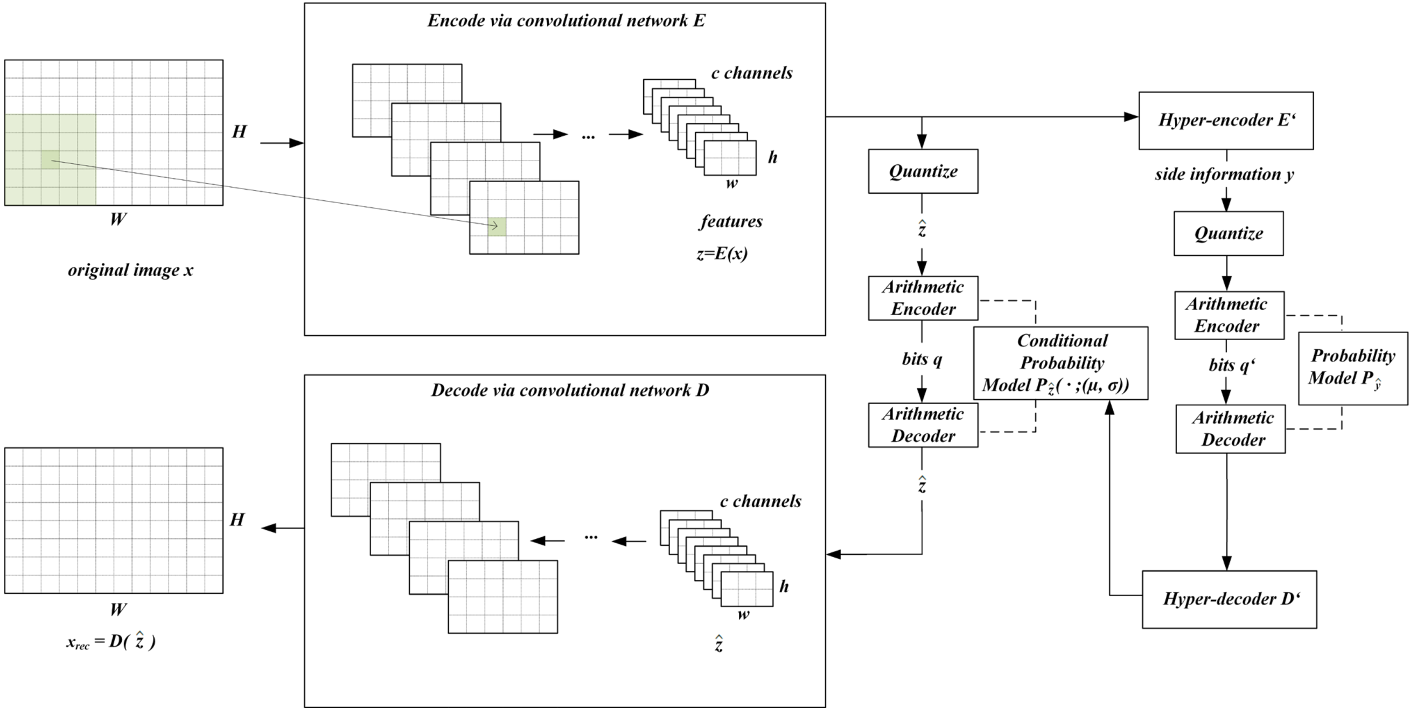
MOTIVATION OF RESEARCH

- Variational auto-encoders have shown promising progress in still image compression
  - Typically use convolutional neural networks (CNN) to find efficient feature representation
  - Can be characterized as non-linear transform coding
  - Trained with stochastic gradient descent on large data sets
  - Ballé et al. [14] employed hyper latents as side information to more efficiently compress extracted image features

- In RGB settings, VAEs can keep up with HEVC in terms of compression efficiency
- Conventional codecs employ orthogonal linear transforms and signal-dependant encoder optimizations

- Can we improve the coding gain of a deep-learned image compression network by rate-distortion optimized quantization?
  - Exhaustively checking a 768x512 luma-only image would take ~10 million decoder network executions
  - Estimate impact of quantization on bitrate and sample distortion

DESCRIPTION OF THE NETWORK ARCHITECTURE



- Based on the network architecture from Ballé et al. [14], 2018
- Multi-scale convolutions with different resolutions
- 256 channels, 3 enc. layers, 3 dec. layers

Image compression process

- Encoder E is trained to find a representation of the input x as features, which are quantized with step size $\Delta>0$:

$$z = E(x); \quad \hat{z}(\Delta) = \Delta \left\lfloor \frac{z}{\Delta} + \frac{1}{2} \right\rfloor$$

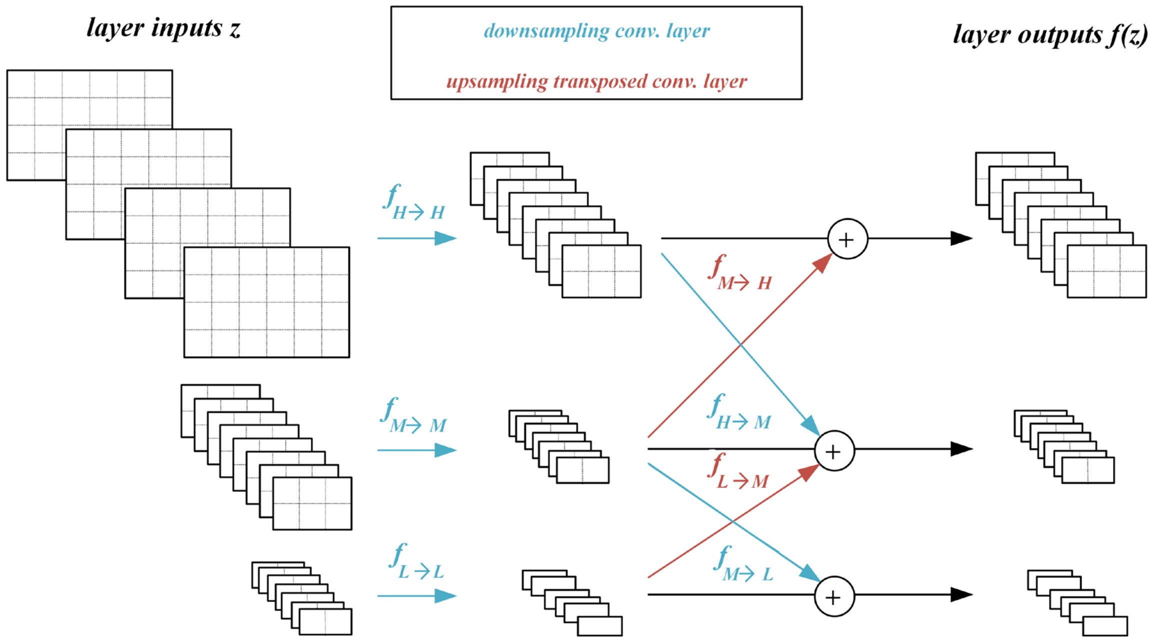
- Decoder D reconstructs the image from the quantized features

$$x_{rec} = D(\hat{z}).$$

- We assume $z \sim \mathcal{N}(\mu, \sigma^2)$
- Hyper encoder E' extracts side information y from the features, which are also quantized and transmitted
- Hyper decoder D' reconstructs the parameters (μ, σ) from $\hat{y}$

Multi -scale convolutional layers

- Separate channels into high (192 channels), middle (48) and low resolution (16)



- $f_{H\to H}$, $f_{M\to M}$, $f_{L\to L}$ are convolutional layers (5×5 kernels, stride 2) with non-linear activation
- $f_{H\to M}$, $f_{M\to L}$ are convolutional layers (5×5 kernels, stride 2)
- $f_{M\to H}$, $f_{L\to M}$ are transposed convolutional layers (5×5 kernels, stride 2)

RD-OPTIMIZED QUANTIZATION

- Fixed side information $\hat{y}$ and probability parameters (μ, σ^2)
- Consider the set of quantization indices

$$\omega \in \mathbb{Z}^{h \times w \times c_0} \oplus \mathbb{Z}^{\frac{h}{2} \times \frac{w}{2} \times c_1} \oplus \mathbb{Z}^{\frac{h}{4} \times \frac{w}{4} \times c_2}.$$

- Minimize the RD cost via

$$\min_{\omega} [\mathrm{MSE}(x, D(\omega)) + \lambda R(\omega)].$$

- The encoder E typically does not find a global solution.
- Goal: understand the impact of selecting different quantization indices.

Distortion estimation I

- Given the error due to quantization $h = \hat{z} - z$, we define an auxiliary function as

$$\epsilon(h) := \text{MSE}\big(D(z), D(z+h)\big)$$

while assuming that the reconstruction quality of unquantized features is as least as good as the quantized features

$$0 \leq \text{MSE}\big(x, D(z)\big) \leq \text{MSE}\big(x, D(\hat{z})\big)$$

- A minimum is at $\epsilon(0) = 0 \implies \nabla\epsilon(0) = 0$
- We can approximate ϵ by a polynomial of degree 2 or higher

Distortion estimation II

- We estimate ϵ by using randomly chosen images and evaluating $\epsilon(h)$ at different quantization step sizes per channel
- Then, we fitted the following biquadratic polynomial to the data by a least-squares approximation

$$\epsilon(h) \approx \sum_{j=1}^{256} \left(\gamma_1^{(j)} \left\|h^{(j)}\right\|^2 + \gamma_2^{(j)} \left\|h^{(j)}\right\|^4\right). \qquad (1)$$

- j denotes the channel index and $h^{(j)}$ is the quantization error in the j-th feature channel.

- Further, $\gamma_i = (\gamma_i^{(1)}, \dots, \gamma_i^{(256)}), i = 1{,}2$ are the coefficients to be determined from the data.

Distortion estimation III

- With $\hat{z} = z + h$, the triangle inequality yields

$$\mathrm{MSE}\big(x, D(\hat{z})\big) \lesssim \mathrm{MSE}\big(x, D(z)\big) + \epsilon(h). \qquad (2)$$

- the upper bound is used for estimating $\mathrm{MSE}\big(x, D(\hat{z})\big)$

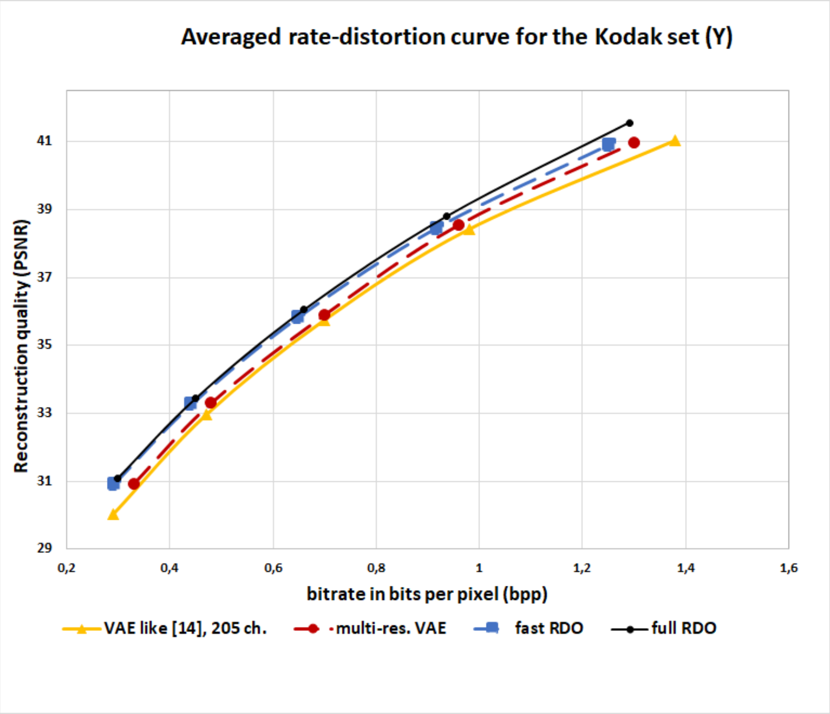
Fast RDO algorithm I

- Compute $z = E(x)$ and proceed for each multi-index position l in the feature representation of $\omega^* := \hat{z}$ as follows:
  - Compute the bitrate of the entry ω_l^* and the auxiliary value $\epsilon^* = \epsilon(h^*)$
  - For candidate test the upper and lower neighboring quant.-level and the mean value from the entropy model, denoted as ω^k:
    - Compute the updated bitrate and auxiliary value $\epsilon^k = \epsilon(h^k)$.
    - Pre-estimate the distortion by using (1) and (2) with the values ϵ^k and ϵ^* .
    - When the pre-estimated RD cost is less than the current minimum, execute the decoder network.
    - Set $\omega^* := \omega^k$, when the true RD cost is a minimum.

Fast RDO algorithm II

- The RD trade-off is measured as Lagrangian cost $d + \lambda R$, where the MSE is used as distortion measure
  - For every λ, a different quantization parameter was used

EXPERIMENTAL RESULTS



- Averaged RD curves over a luma-only version of the Kodak set.
- The average PSNR is computed by from the average MSE over the Kodak set
- As benchmark, a VAE like in [14] with 205 channels has been trained on luma-only data from Imagenet.

CONCLUSIONS

- The quantization error of the features can be used to estimate the sample distortion via a suitably -determined polynomial.

- By using (1) to search suitable candidates, a fast RDO algorithm with significantly less decoder executions than fully testing each candidate can be designed.

- The application of the fast RDO improves the coding gain while keeping the same decoder.

REFERENCES

Rate -Distortion Optimization and Quantization

[7] G. J. Sullivan and T. Wiegand, "Rate -distortion optimization for video compression," IEEE Signal Processing Magazine, vol. 15, no. 6, pp. 74–90, 1998.

[8] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," IEEE Transactions on Image Processing, vol. 3, no. 5, pp. 533–545, 1994.

[10] M. Karczewicz, Y. Ye, and I. Chong, "Rate distortion optimized quantization," ITU -T SG16/Q6 (VCEG), January 2008.

[22] J. Stankowski, C. Korzeniewski, et al., "Rate -distortion optimized quantization in HEVC: Performance limitations," 2015 Picture Coding Symposium (PCS), pp. 85–89, 2015.

Deep Image Compression

[11] J. Ballé, P. Chou, et al., "Nonlinear Transform Coding," IEEE Journal of Selected Topics in Signal Processing, vol. PP, pp. 1–1, 10 2020.

[13] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in International Conference on Learning Representations (ICLR), Toulon, France, April 2017.

[14] J. Ballé, D. Minnen, et al., "Variational image compression with a scale hyperprior," in International Conference on Learning Representations, 2018.

[15] M. Akbari, J. Liang, et al., "Generalized octave convolutions for learned multifrequency image compression," 2020, Available at https://arxiv.org/abs/2002.10032.