Rate-Distortion-Optimization for Learning-Based Image Compression using Adaptive Hierarchical Autoencoders

Fabian Brand
fabian.brand@fau.de

Chair of Multimedia Communications and Signal Processing
Rate-Distortion Optimization

• Multiple decisions in traditional video coding
  – Block partitioning
  – Prediction mode selection
  – Quantization stepsize
  – Many more ...

• Selection according to rate distortion cost function
  \[ J = R + \lambda D \]

• Typically done by testing multiple settings
Adaptive Block Partitioning

- Block-based image and video compression (e.g. HEVC, VVC)
- Block-size determines
  - Context for prediction
  - Transform lengths
  - Quantization stepsize
- Rule of thumb:
  - Small blocks for detailed content
  - Large blocks for stationary content
- Adaptive partitioning greatly increases coding efficiency
Learning-Based Image Compression

- Main tool: Autoencoder
- Consists of Encoder network $e(x)$ and decoder network $d(z)$
- Trained on equal input and output, e.g. using $MSE$
  \[ L_D = MSE[x; \tilde{x}] \]
- Entropy bottleneck between encoder and decoder
  \[ L_R = H(\hat{z}) \]
- End-to-end training possible

\[ x \]
\[ z = e(x) \]
\[ \tilde{z} = Q(z) \]
\[ \hat{x} = d(\tilde{z}) \]
Rate-Distortion Optimization in E2E Compression

• Training on joint loss function
  \[ L = L_D + \lambda L_R \]

• „Static“ RDO

• No free parameters after training
  – No possibility for „dynamic“ RDO
RDONet

- Compressive Autoencoder capable of coding at adaptive depth
- Compression after 4, 5 or 6 downsampling steps
- Decision on block-level
- Compression as whole image
- No block division
- No block artifacts

Side Information per block

Brand, Fischer, Kaup: „Rate-Distortion-Optimized Image Compression using an adaptive hierarchical autoencoder with conditional hyperprior“, CVPR 2021
Latent Space Units
Latent Space Units

- Downsampling on encoder side
- Upsampling on decoder side
- Transmitting masked latent space
- Redundancy from lower LS-Unit
  - Transmit conditional to previous layer

![Diagram of Latent Space Units]
Conditional Hyperprior

- Compression of each latent space with hyperprior
  - Separate autoencoder transmitting pdf for latent space
- Replace hyperprior autoencoder with conditional autoencoder
- Reducing redundancy from previous latent space
Summary of Network

Coding on different levels of autoencoder possible

Level externally adjustable on block-level

Conditional coding to reduce redundancy between levels
Rate-Distortion Optimization in E2E Compression

- Training on joint loss function
  \[ L = L_D + \lambda L_R \]
- „Static“ RDO
- No free parameters after training
  - No possibility for „dynamic“ RDO
- Externally controlled depth
- Test different depth configurations and pick best
RDO Search

• Initialize image coding with lowest latent space
• Test if higher latent space yields better RD-behavior
• Optimizing each 64x64 area individually
  – Global search not feasible
  – Global optimum not found
• Solution: 2-pass RDO
  – Initialize with result of first pass
Experiments

Training

• Train on CLIC Intra + DIV2K + TECNICK
• Random choice of LS depth
• Train for 2000 epochs
• Train on MS-SSIM and MSE
  – MSE needed for stability

Test

• Evaluate on CLIC Intra test set
• Compare 1-pass and 2-pass RDO
• Compare against standard 4 layer autoencoder with hyperprior and context model
• MS-SSIM as distortion metric

\[
L_{\text{train}} = D_{\text{ms-ssim}} + 0.1 \cdot D_{\text{mse}} + \lambda_t R
\]

\[
L_{\text{RDO}} = D_{\text{ms-ssim}} + \lambda_e R
\]
Visual Example

Example Block Partitioning (Dark: Small blocks, high level)
Visual Example

\[ \lambda_e = 1 \quad r = 0.096 \text{bpp} \]
Visual Example

\[ \lambda_e = 0.5 \]

\[ r = 0.104 \text{bpp} \]
Visual Example

\[ \lambda_e = 0.25 \quad r = 0.119\text{bpp} \]
Visual Example

\[ \lambda_e = 0.125 \quad r = 0.136 \text{bpp} \]
Visual Example

\[ \lambda_e = 0.0625 \quad r = 0.141 \text{bpp} \]
Visual Example

Standard Autoencoder

\[ r = 0.141 \text{bpp} \]
Results

• Five rate-points per model
  – RDO allows multiple rate points per model
  – Optimal behavior if $\lambda$ match in training and RDO
• Pick one rate point per model for final evaluation
Results

- Rate savings over compression without RDO
- Additional gains by 2-pass RDO
- 7.7% rate saving on average
- Up to 22.5% for single images

<table>
<thead>
<tr>
<th></th>
<th>1-pass</th>
<th>2-pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worst Case</td>
<td>+7.5%</td>
<td>+3.5%</td>
</tr>
<tr>
<td>Best Case</td>
<td>-18.8%</td>
<td>-22.5%</td>
</tr>
<tr>
<td>Average</td>
<td>-4.1%</td>
<td>-7.7%</td>
</tr>
</tbody>
</table>

BD-Rates for entire CLIC validation set
Conclusion

- RDONet enables RDO similar to adaptive block partitioning
- Saving 7.7% rate compared to standard autoencoder
- Increase visual quality by adaptive bit allocation
- Transferring concept from traditional video compression to end-to-end image coding