Soft then Hard: Rethinking the Quantization in Neural Image Compression

Zongyu Guo, Zhizheng Zhang, Runsen Feng, Zhibo Chen*
University of Science and Technology of China

38th International Conference on Machine Learning

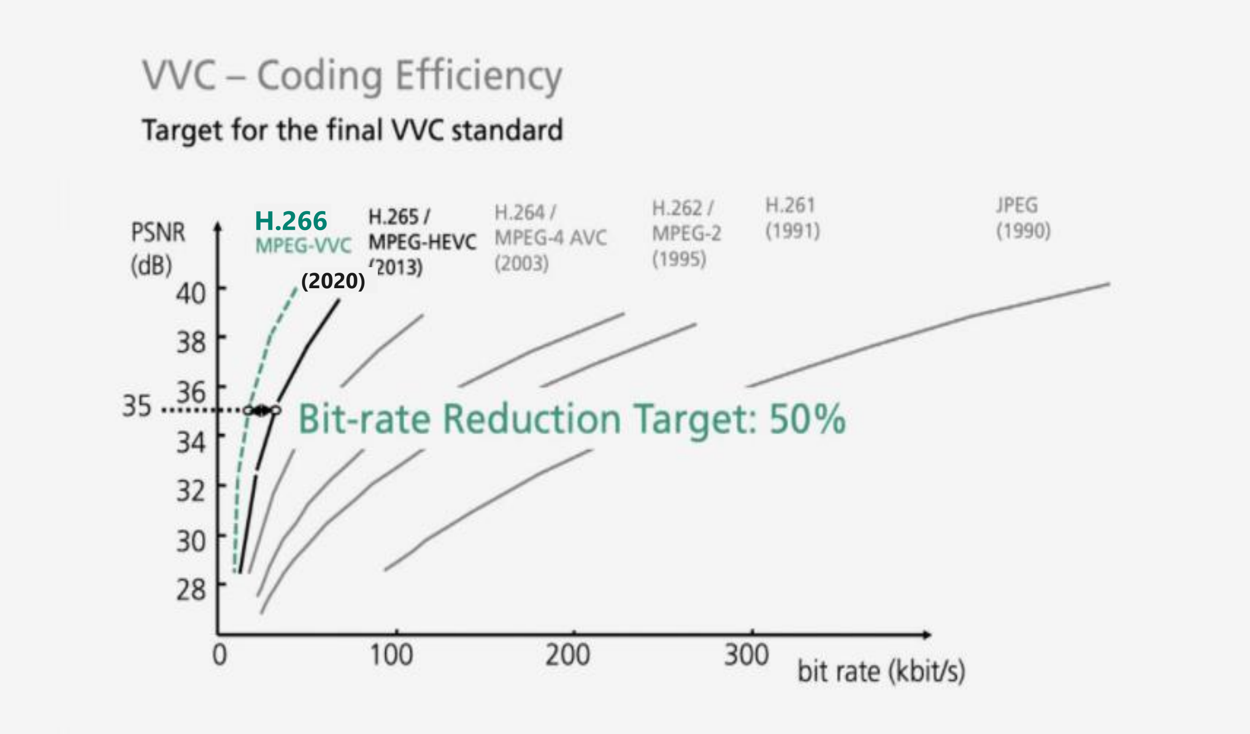
Task: Image Compression

- Image/video compression techniques are important for image/video transmission and storage.
  - Video contributes to more than 75% Internet traffic.

- In particular, image compression is the foundation of video compression.
  - Image compression aims to use limited number of bits to represent an image with desirable reconstruction quality.

- Traditional compression standards.
  - From JPEG (image), to H.26x (image/video).
  - The latest standard is H.266/VVC.

- The development of compression standards
  - follows Moore's Law as well.



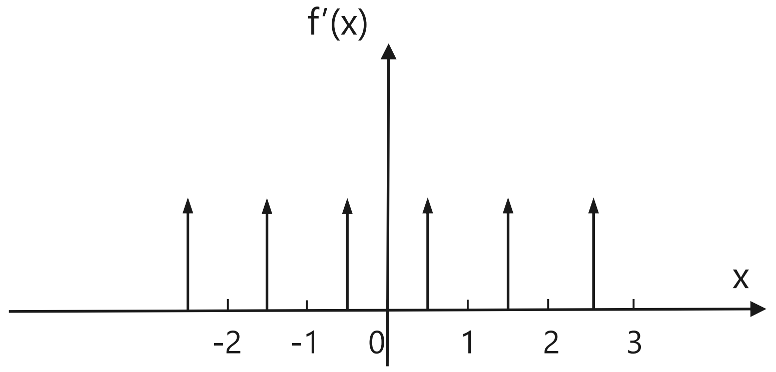
Task: Neural Image Compression

- Neural image compression has surpassed traditional compression standards.
  - The current learned image compression models outperform H.266 intra in terms of PSNR.
  - They are promising to achieve much better perceptual quality. (e.g., saves 40% bits against VVC in terms of MS-SSIM)

- Quantization is indispensable for lossy image compression.
  - The distortion introduced by quantization has been studied well with **Shannon's rate-distortion theory.**

- Quantization influences the performance of an neural image compression model obviously.

→ **End-to-end optimization requires differentiable approximations of quantization.**



The derivative of rounding

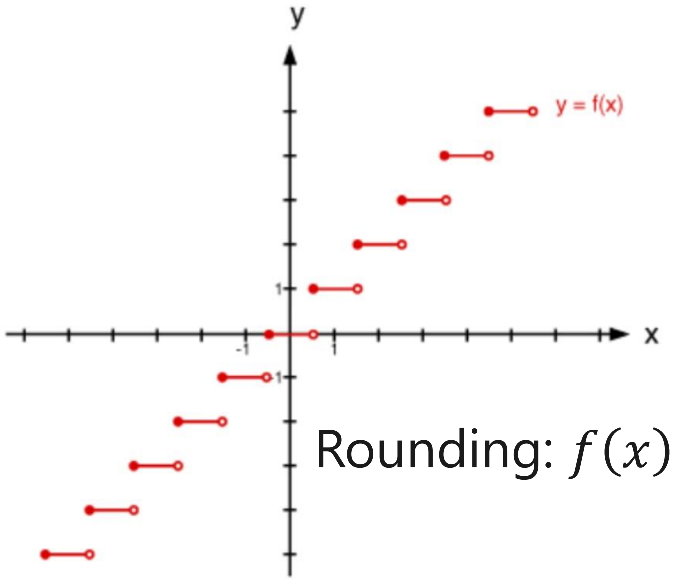
The gradient of quantization is zero almost everywhere, how to make back-propagation applicable?

Background

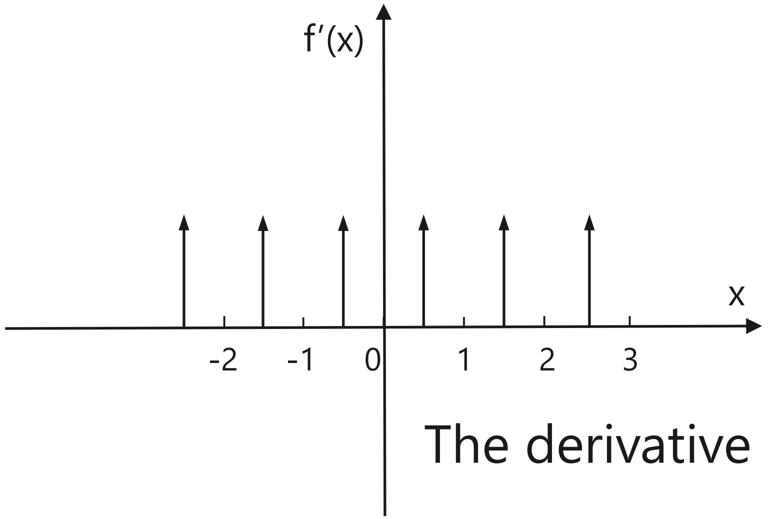
- Problem: How to end-to-end optimize a neural network that consists of quantization layer?

- 1. Scalar Quantization
- 2. Vector Quantization (more complicated)

a typical scalar quantization method



Rounding: $f(x) = \lfloor x \rceil$

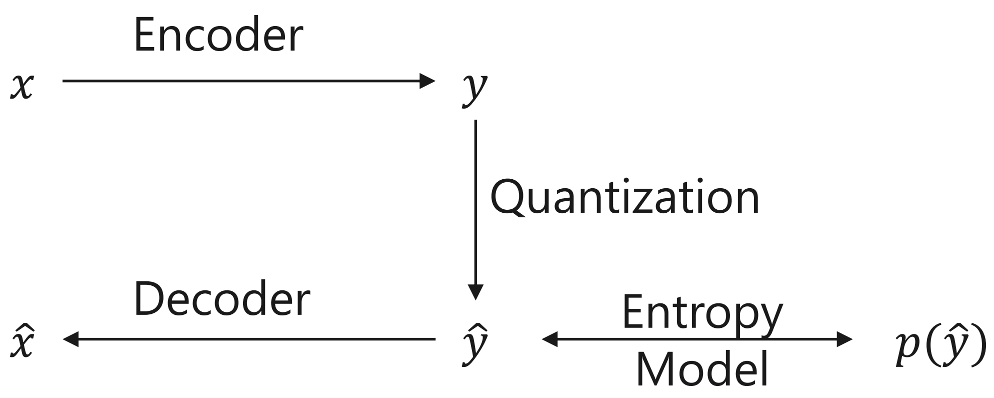


The derivative of rounding

This is a basic problem for neural image compression.

Notations

x: a natural image
y: latent variable after encoding transform. $y = Encoder(x)$.
$\hat{y}$: discrete latent variable after quantization. $\hat{y} = Round(y)$.
$\hat{x}$: the reconstruction image after decoding transform. $\hat{x} = Decoder(\hat{y})$
$p(\hat{y})$: the probability of discrete latent variable.
$R(\hat{y})$: the rate that describes the transmission cost. $R(\hat{y}) = -\log_2 p(\hat{y})$



$$L = R(\hat{y}) + \lambda \cdot D(x, \hat{x})$$

Optimization goal: rate-distortion tradeoff

Differentiable approximations of quantization

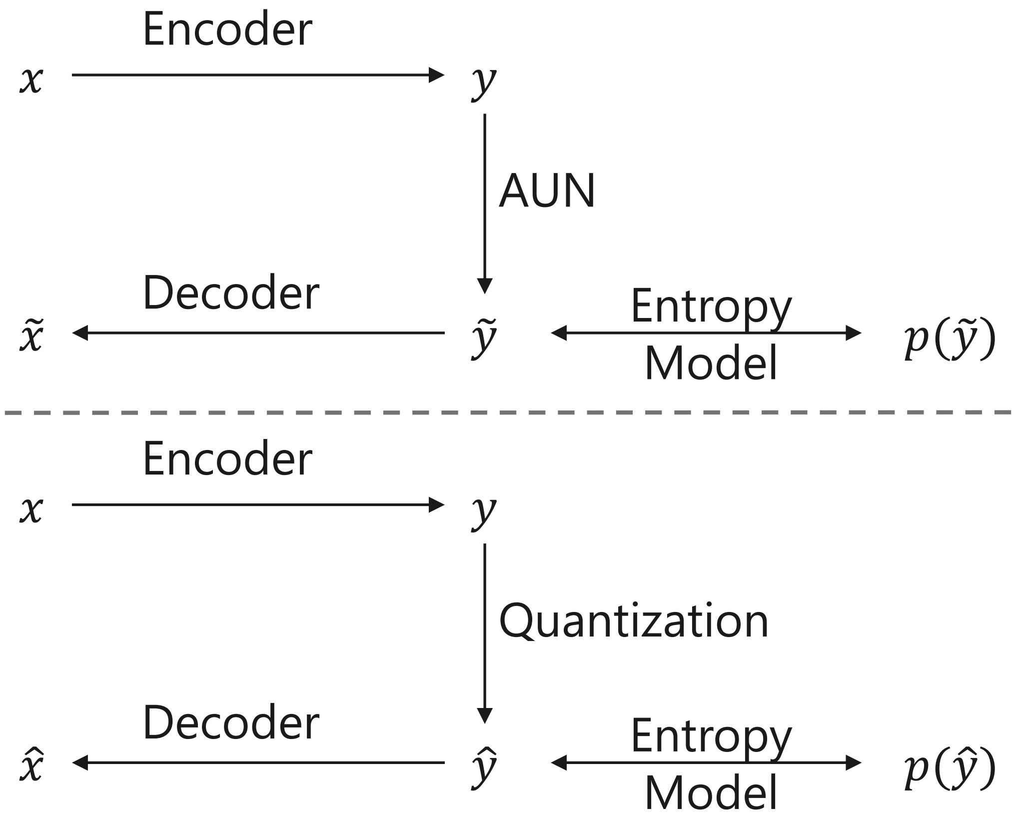
Method 1: additive uniform noise (AUN), proposed in [3], a popular approximation method

Training:

$$\tilde{y} = y + u, u \in (-0.5, 0.5]$$

Testing:

$$\hat{y} = \lfloor y \rceil$$



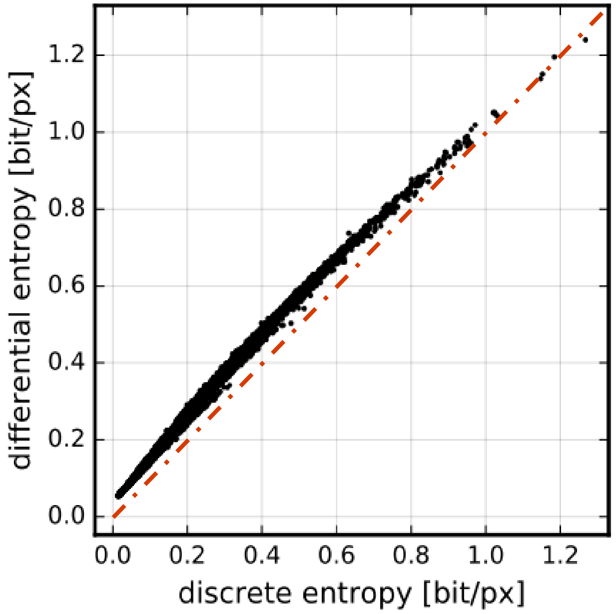
[3] End-to-end optimized image compression, Balle et al., in ICLR 2017.

Why AUN works?

Quantization with additive uniform noise (AUN) can be interpreted as variational compression.

(i) The optimization goal of image compression, i.e., the *rate-distortion* tradeoff, is associated with variational inference.

(ii) The rate-distortion loss is actually the upper bound of actual rate-distortion value.

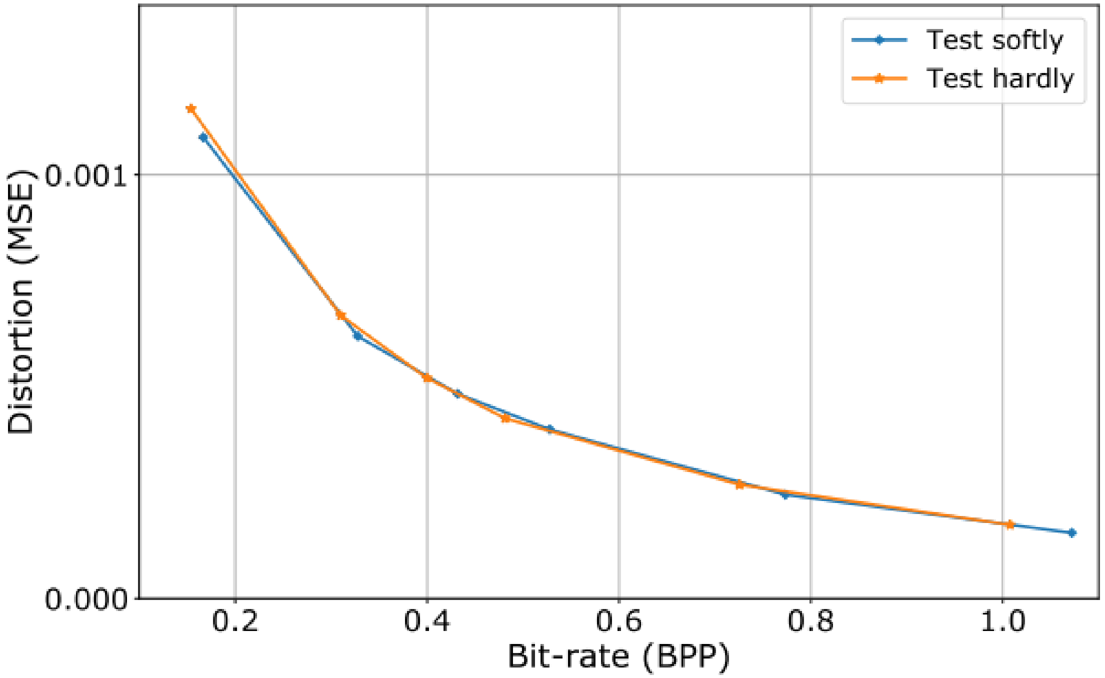


$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}\sim q}[-\log P(\hat{\boldsymbol{y}})] &\approx \mathbb{E}_{\boldsymbol{y}\sim q}[-\log \int_{[-0.5,0.5]} p(\boldsymbol{y}+\boldsymbol{u})d\boldsymbol{u}] \\ &\leq \mathbb{E}_{\boldsymbol{y}\sim q}[-\int_{[-0.5,0.5]} \log p(\boldsymbol{y}+\boldsymbol{u})d\boldsymbol{u}] \\ &= \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}[-\log p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}})]. \end{aligned} \quad (3)$$

A statistical explanation of this variational relaxation in
[3] End-to-end optimized image compression, Balle et al., in ICLR 2017.

The train-test mismatch issue of AUN

Stochastic training with soft approximation
Deterministic testing with hard quantization



The rate point shift issue caused by AUN

The train-test mismatch issue would hurt the rate-distortion performance of a compression model.

Some competitive alternatives for quantization include:

- Straight-through estimator
- Soft-to-hard annealing.

Both of them achieve train-test consistency.

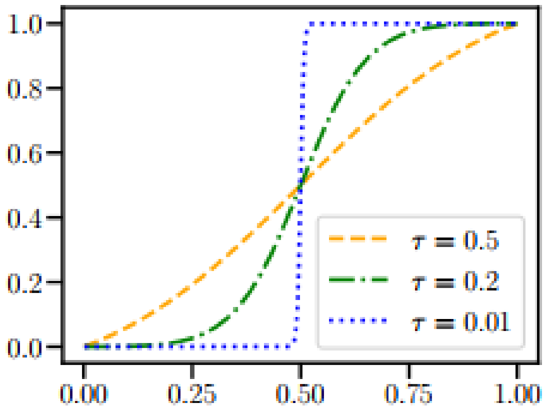
Differentiable approximations of quantization

Method 2: straight through estimator (STE) and its variants

- STE applies the identity gradients to pass through the hard quantization layer.
- The backward and forward passes do not match, the coarse gradient before the quantization layer is certainly not the gradient of loss function.

Method 3: soft-to-hard annealing.

- The differentiable function goes towards the shape of hard quantization gradually.
- Previous works using soft assignment [4] or soft simulation [5] initially.



→ An illustration of the annealing-based approximation function in [5]

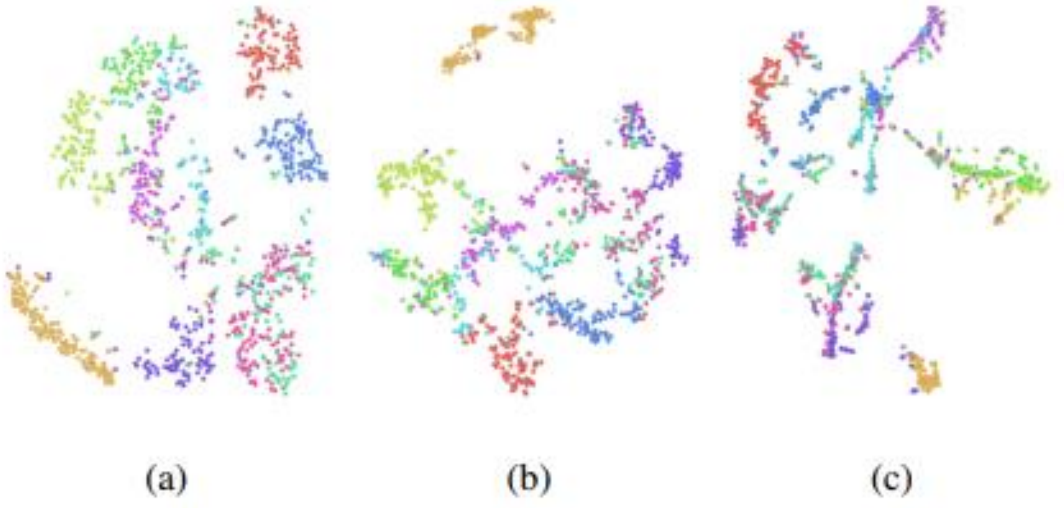
[4] Soft-to-hard vector quantization for end-to-end learning compressible representations, Agustssons et al., in NeurIPS 2017.
[5] Improving inference for neural image compression, Yang et al., in NeurIPS 2020.

Our analyses

We analyze these three quantization methods, AUN, STE and soft-to-hard annealing.

STE or annealing-based quantization seems good by achieving train-test consistency?
Our argument: no, these two quantization methods hurt the latent representation ability.

- The t-SNE visualization of the continuous latent space before quantization.
- Here we use stochastic Gumbel annealing (SGA) [5] for this illustrative task.



AUN STE SGA

[5] Improving inference for neural image compression, Yang et al., in NeurIPS 2020.

Our analyses

Expressive latent space is significant for image compression, where we always expect the transmitted symbol to convey more effective information.

Table 1. Comparisons of three quantization methods. Improper training strategy may cause unstable convergence of STE-based model (Yin et al., 2019), thus represented as -. Our proposed soft-then-hard (STH) strategy and scaled uniform noise (SUN) are meaningful.

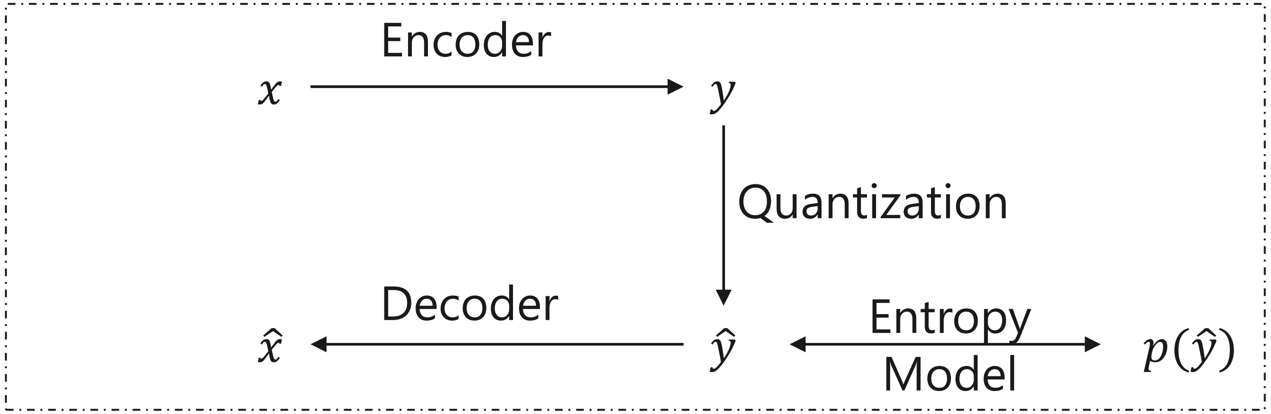
	AUN	STE	Annealing-Based	STH (Ours)	STH + SUN (Ours)
Train-Test Consistency	✗	✓	✓	✓	✓
Latent Expressiveness	✓	✗	✗	✓	✓
Variational Compression	✓	✗	✗	✓	✓(more flexible)
Exact Gradient	✓	✗	✓	✓	✓
Stable Training	✓	-	✗	✓	✓

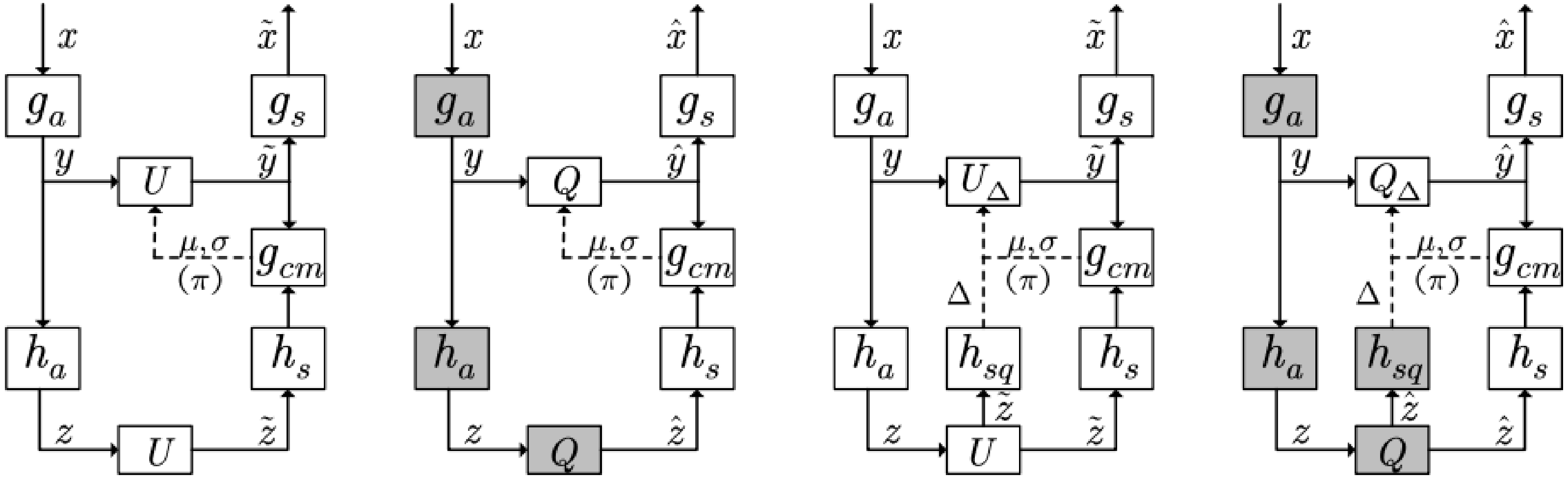
- Previous methods cannot achieve train-test consistency and latent expressiveness at the same time.
- But our proposed soft-then-hard (STH) can.

- Previous methods cannot adaptively determine the quantization granularity.
- But our proposed scaled uniform noise (SUN) can.

Our methods

Framework





(a) Compression with AUN

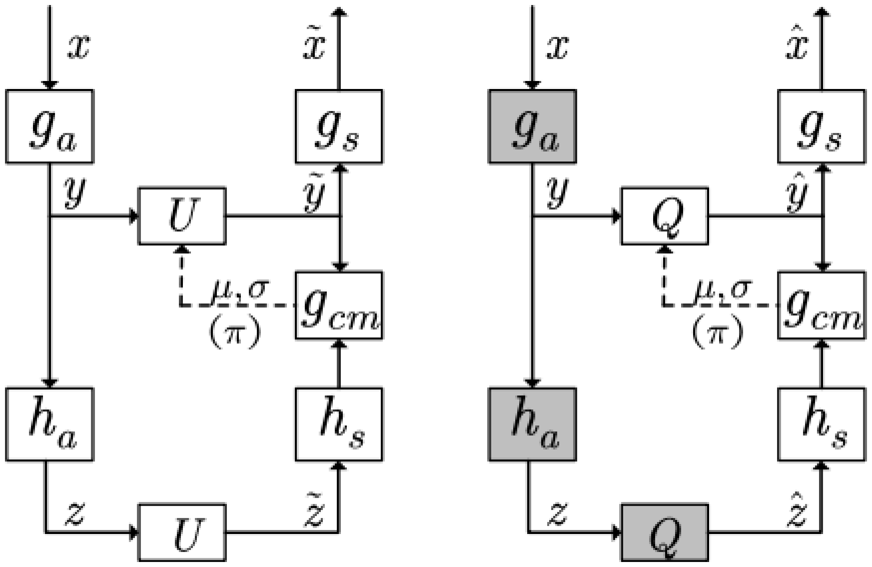
(b) Our proposed soft-the-hard (STH)

(c) Our proposed scaled uniform noise (SUN)

(d) STH + SUN

Soft then hard (STH)

Motivated by the two-stage training in some VQ-VAE works



(a)
Compression with AUN

(b)
Our proposed
soft-the-hard (STH)

- Fix the encoder.
- Tune the decoder.
- Apply the unbiased rate-distortion loss for finetuning.

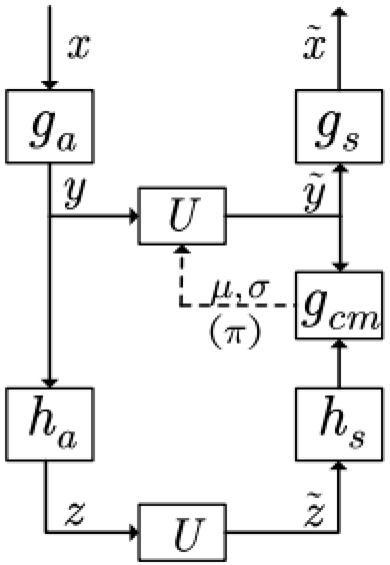
$$\mathcal{L} = \mathbb{E}_{\boldsymbol{y}\sim q}[-\log P(\hat{\boldsymbol{y}}) - \log p_{\boldsymbol{x}|\hat{\boldsymbol{y}}}(\boldsymbol{x}|\hat{\boldsymbol{y}})]. \quad (4)$$

By detaching the decoder from the encoder, the ex-post tuning stage can be regarded as a joint optimization of two independent tasks:

1) Optimize a reconstruction model, input is $\hat{y}$, output is $\hat{x}$.
2) Learn a prior likelihood model to estimate density $P(\hat{y})$.

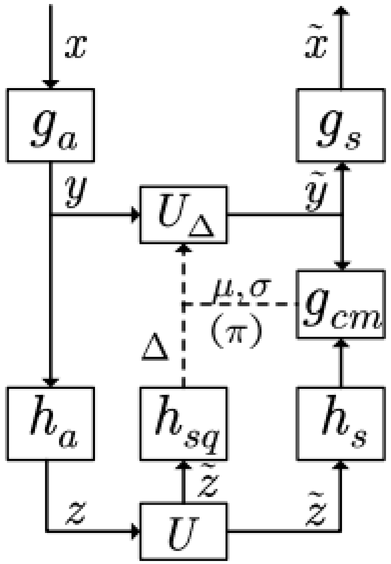
Scaled uniform noise (SUN)

Motivated by the variational dequantization in flow models



(a)

Compression with AUN



(c)

Our proposed scaled uniform noise (SUN)

Beyond fixed integer quantization, we propose to learn the noise scale to adaptively control the quantization step.

Variational inference v.s. Neural image compression

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} D_{\mathrm{KL}}(q(\tilde{\boldsymbol{y}}|\boldsymbol{x})|p(\tilde{\boldsymbol{y}}|\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} \log p(\boldsymbol{x}) + \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}[\log q(\tilde{\boldsymbol{y}}|\boldsymbol{x}) - \log p_{\boldsymbol{x}|\tilde{\boldsymbol{y}}}(\boldsymbol{x}|\tilde{\boldsymbol{y}}) - \log p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}})]. \quad (1)$$

Not always 0 any more

$$\Delta = h_{sq}(\tilde{\boldsymbol{z}}),$$

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{u}, \boldsymbol{u} \sim \mathcal{U}(\frac{\Delta}{2}, \frac{\Delta}{2}),$$

$$q(\tilde{\boldsymbol{y}}|\boldsymbol{x}) = q(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = q(\boldsymbol{u}|\boldsymbol{y}) = \frac{1}{\Delta}, \quad (5)$$

We prove that we are still optimizing the upper bound of actual rate, and the rate-distortion optimization still conforms variational learning.

A short summary

We provide a new analysis of three quantization methods, AUN, STE and annealing-based, and argue that:

- Training with STE or annealing is equal to optimizing a deterministic autoencoder, in which it is hard to learn a smooth latent space due to the lack of regularization term.
- STE-based or annealing-based quantization suffers from some training troubles such as biased gradient or unstable gradient, rendering the encoder suboptimal.
- The above two methods cannot ensure the latent representation ability.
- In contrast, optimizing a compression model with additive uniform noise can be interpreted as variational optimization, but suffers from the train-test mismatch.

We thus propose:

- Soft-then-hard (STH) quantization strategy, which first learns an expressive latent space softly, then closes the train-test mismatch with hard quantization.
- Scaled uniform noise (SUN), by deriving a new variational upper bound on actual rate that incorporates the scale of additive uniform noise into optimization and thus enable flexible quantization.

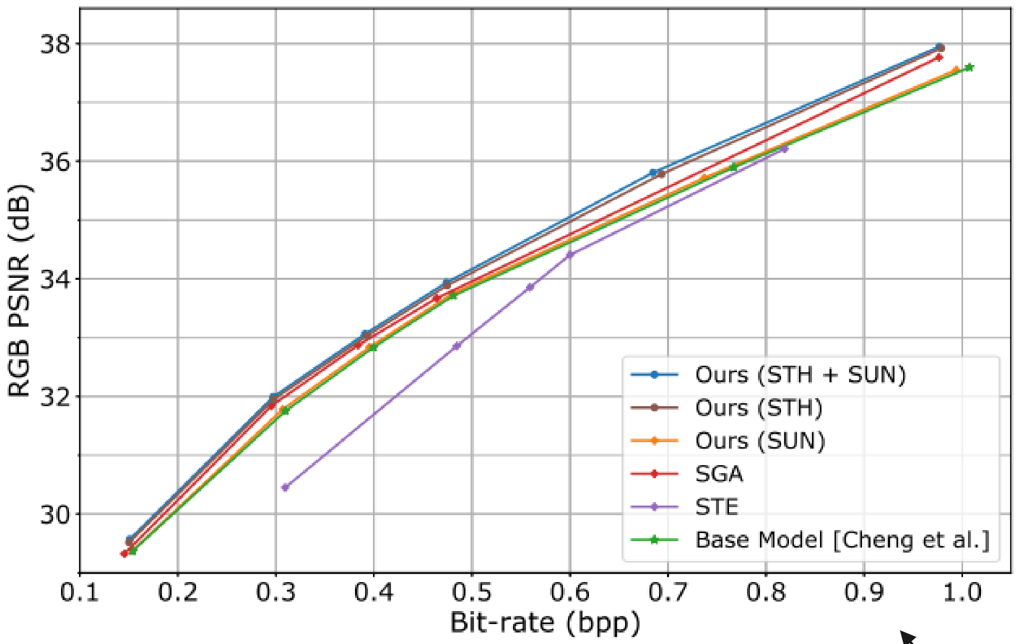
Experiments

Our methods, soft then hard (STH) and scaled uniform noise (SUN) are

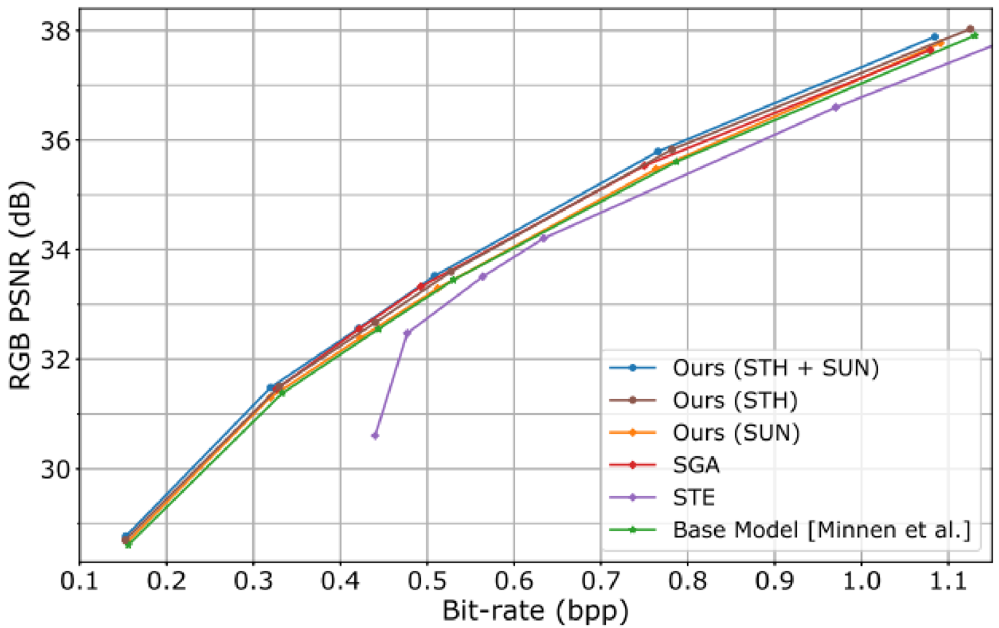
- Easy to adopt
  - STH does not requires additional parameters.
  - SUN requires minor additional parameters.
- Stable to train
  - The ex-post tuning stage of STH is more stable than the first-stage training with AUN.
  - If training with AUN starts to collapse on some models, we can even decrease the iteration number of the first stage but still conduct ex-post tuning.
- Highly effective especially on complex compression models
  - We empirically find that the train-test mismatch issue caused by additive uniform noise is more serious on complex compression models, presumably because of the posterior collapse issue.

Experiments

- Select two popular models as base models.



A powerful base model [6]



An early work as base model [7]

8.9% BD-rate savings

[6] Learned image compression with discretized Gaussian mixture likelihoods and attention modules, Cheng et al., in CVPR 2020.
[7] Joint autoregressive and hierarchical priors for learned image compression, Minnen et al., in NeurIPS 2018.

Experiments

Ground Truth — (a) 0.210 / 32.80 / 0.9689 — (b) 0.203 / **33.15** / 0.9710 — (c) 0.169 / 28.86 / **0.9767**

Figure 4. Qualitative comparisons. (a) Base model (Cheng et al., 2020) optimized for PSNR. (b) Employing our methods optimized for PSNR. (c) Employing our methods optimized for MS-SSIM. The statistics are the values of bit-rate (bpp) / PSNR (dB) / MS-SSIM.

[6] Learned image compression with discretized Gaussian mixture likelihoods and attention modules, Cheng et al., in CVPR 2020.

Experiments

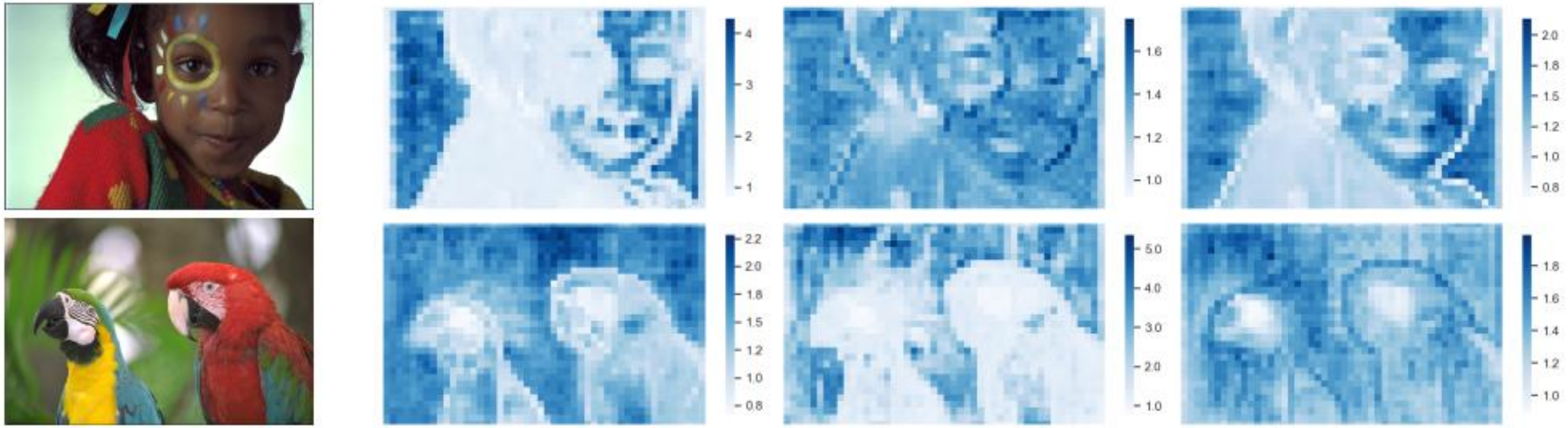


Figure 5. Visualizations of the noise scale. Left: ground truth. Right three columns: noise scale in different channels.

- Our proposed scaled uniform noise is used to determine the quantization step for effective spatial bit allocation.
- It is proved to generate image-adaptive quantization step.
- It is also promising to be extended into variable rate compression.

Thank you!

Author list:
- Zongyu Guo (guozy@mail.ustc.edu.cn)
- Zhizheng Zhang (zhizheng@mail.ustc.edu.cn)
- Runsen Feng (fengruns@mail.ustc.edu.cn)
- Prof. Zhibo Chen (chenzhibo@ustc.edu.cn)