

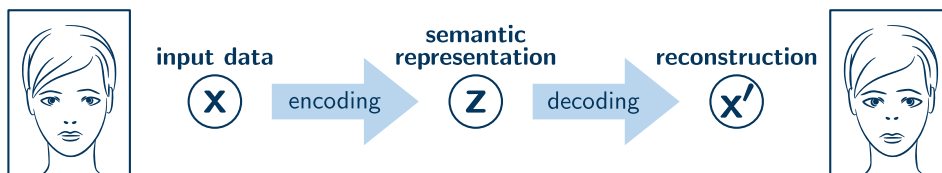
Lecture 10:

Lossy Compression: From VAEs to Rate/Distortion Theory

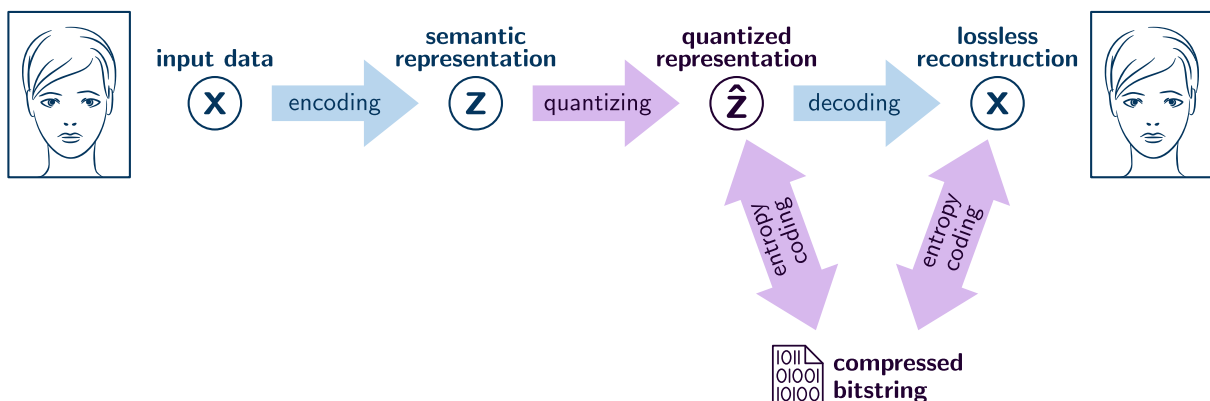
Robert Bamler · Summer Term of 2023

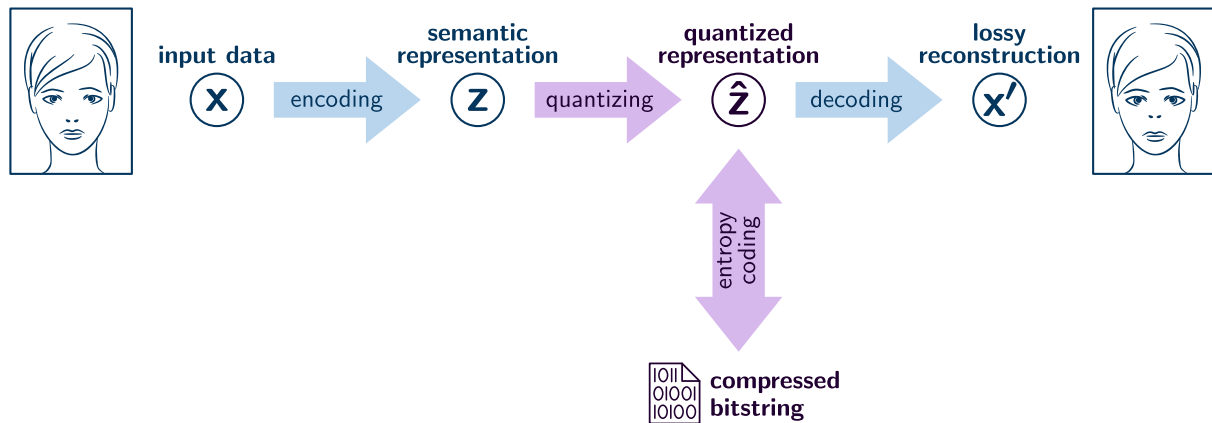
These slides are part of the course “Data Compression With and Without Deep Probabilistic Models” taught at University of Tübingen. More course materials—including video recordings, lecture notes, and problem sets with solutions—are publicly available at <https://robamler.github.io/teaching/compress23/>.

Recall: Variational Autoencoder (VAE)



Problem Set 9: Lossless Compression With a VAE





VAEs for Lossy (Image) Compression

- ▶ **Example:** data $\mathbf{X} = (X_{i,j,k})$ are color images.
 - ▶ $(i, j, k) = (\text{x-position}, \text{y-position}, \text{red/green/blue})$; $X_{i,j,k} \in [0, 1]$ is a (continuous) RGB value
 - ▶ Likelihood is again parameterized by a neural network g_θ : $P_\theta(\mathbf{X} | \mathbf{Z}) = \prod_i P_\theta(X_{i,j,k} | \mathbf{Z})$
 - ▶ This time: Gaussian likelihood, i.e., density function $p_\theta(x_{i,j,k} | \mathbf{Z} = \mathbf{z}) = \mathcal{N}(x_{i,j,k}; g_\theta(\mathbf{z})_{i,j,k}, \frac{\beta}{2}I)$
- ▶ **Idea:** just use $g_\theta(\mathbf{z})$ as the reconstruction of an image.
(Don't bother using the likelihood $P_\theta(\mathbf{X} | \mathbf{Z} = \mathbf{z})$ to encode the true image.)
- ▶ $\text{ELBO}_{\beta}(\theta, \phi, \mathbf{x}) =$

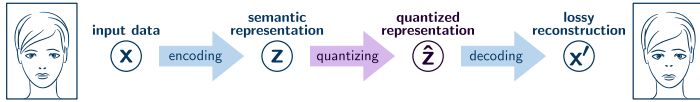
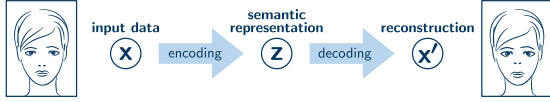
Quantizing Latent Space

- ▶ Latents $\mathbf{z} \in \mathbb{R}^d$ are *continuous* \implies can't be entropy coded
- ▶ Problem 9.2: for *lossless* compression with bits-back coding, we can simply quantize \mathbf{z} to an arbitrarily fine grid.

$$D_{\text{KL}}(Q_\phi(\mathbf{Z} | \mathbf{X} = \mathbf{x}) \parallel P_\theta(\mathbf{Z})) = \mathbb{E}_{Q_\phi(\mathbf{Z} | \mathbf{x} = \mathbf{x})} [\log Q_\phi(\mathbf{Z} | \mathbf{X} = \mathbf{x}) - \log P_\theta(\mathbf{Z})]$$
 - ▶ bit rate for encoding $\hat{\mathbf{z}} = \lceil \mathbf{z} \rceil_\delta$ with quantized prior $P_\theta(\hat{\mathbf{Z}})$:
 - ▶ bit rate for decoding $\hat{\mathbf{z}}$ with quantized var. dist. $Q_\phi(\hat{\mathbf{Z}} | \mathbf{X} = \mathbf{x})$:
$$\implies \text{for } \delta \rightarrow 0, \text{ expected net bit rate is } D_{\text{KL}}(Q_\phi(\mathbf{Z} | \mathbf{X} = \mathbf{x}) \parallel P_\theta(\mathbf{Z})) \text{ (independent of } \delta)$$
- ▶ **Problem:** bits-back coding does not work (out of the box) for lossy compression.
 - ▶ Receiver never recovers the exact message $\mathbf{x} \implies$ can't encode $\hat{\mathbf{z}}$ with $Q_\phi(\hat{\mathbf{Z}} | \mathbf{X} = \mathbf{x})$.
 - ▶ Thus, bit rate would depend on how fine we make the grid.

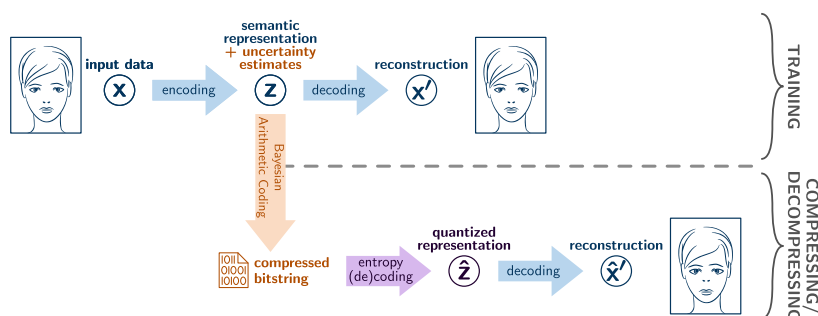
- ▶ **Idea:** take quantization into account already during training.
 - ▶ **Goal:** model should learn to encode important information on length scales $\geq \delta$.
 - ▶ **Problem:** quantization $\mathbf{z} \mapsto \hat{\mathbf{z}} := \lceil \mathbf{z} \rceil_\delta$ is not differentiable.
 - ▶ **Observation:** $(\hat{\mathbf{z}} - \mathbf{z}) \in [-\frac{\delta}{2}, \frac{\delta}{2}]^d$ and (empirically) approximately uniformly distributed.
- ▶ **Proposal:** at training, replace quantization by adding uniform noise $\epsilon \sim \mathcal{U}([-\frac{\delta}{2}, \frac{\delta}{2}]^d)$
 - ▶ Equivalent to using a box-shaped variational distribution with fixed width:
 - ▶ δ is fixed (i.e., data-independent)
 - \Rightarrow might as well set $\delta = 1$ as long as the prior $P_\theta(\mathbf{Z})$ does not impose any fixed length scale.
 - \Rightarrow at deployment, encode each component $\hat{z}_i := \lceil z_i \rceil_\mathbb{Z}$ with model $P_\theta(\hat{\mathbf{Z}}_i = \hat{z}_i) =$
 - \Rightarrow bit rate $P_\theta(\hat{\mathbf{Z}} = \hat{\mathbf{z}}) =$

Model Training: Rate/Distortion Trade-Off

- ▶ **In deployment:**

 - ▶ quantized latent representation: $\hat{\mathbf{z}} := \lceil \mathbf{z} \rceil = \lceil f_\phi(\mathbf{x}) \rceil$
 - ▶ bit rate: $-\log P(\hat{\mathbf{Z}} = \hat{\mathbf{z}}) = -\log \left(\int_{\mathbf{z} \in [\hat{\mathbf{z}} - \frac{1}{2}, \hat{\mathbf{z}} + \frac{1}{2}]} p(\mathbf{z}) d\mathbf{z} \right)$
 - ▶ reconstructed message: $\mathbf{x}' = g_\theta(\hat{\mathbf{z}})$
- ▶ **At training time:**

 - ▶ added uniform noise: $\mathbf{z} = f_\phi(\mathbf{x}) + \epsilon$ where $\epsilon \sim \mathcal{U}([-\frac{1}{2}, \frac{1}{2}]^d)$
 - ▶ approximated bit rate: $\mathcal{R}(\theta, \phi, \mathbf{x}) := -\log \tilde{p}(\mathbf{z})$ where $\tilde{p}(\mathbf{z}) := \int_{\mathbf{z}' \in [\mathbf{z} - \frac{1}{2}, \mathbf{z} + \frac{1}{2}]} p(\mathbf{z}') d\mathbf{z}'$
 - ▶ reconstruction error (distortion): e.g., MSE: $\mathcal{D}(\theta, \phi, \mathbf{x}) := \|g(\mathbf{z}) - \mathbf{x}\|^2$
 - ▶ loss function: rate/distortion trade off: $\mathcal{L}_\beta(\theta, \phi, \mathbf{x}) = \mathbb{E}_\epsilon [\beta \mathcal{R}(\theta, \phi, \mathbf{x}) + \mathcal{D}(\theta, \phi, \mathbf{x})]$
 \rightarrow Problem 10.1: \exists probabilistic model such that $\mathcal{L}_\beta(\theta, \phi, \mathbf{x}) \propto -\text{ELBO}(\theta, \phi, \mathbf{x}) + \text{const}$

Limitations of Uniform Quantization

- ▶ quantization gap: rounding \neq adding uniform noise.
 - ▶ various proposals exist for better quantization at training time (Lecture 12)
- ▶ rate/distortion trade-off β must be set *at training time*.
 - ▶ less studied in the literature.
 - ▶ Variational Inference can help: [Yang et al., 2020, Tan & Bamler, 2022]





Lecture 10:

Lower Bound on the Bit Rate of Lossy Compression

Robert Bamler · Summer Term of 2023

These slides are part of the course “Data Compression With and Without Deep Probabilistic Models” taught at University of Tübingen. More course materials—including video recordings, lecture notes, and problem sets with solutions—are publicly available at <https://robamler.github.io/teaching/compress23/>.

Recall: Source Coding Theorem



- ▶ optimal expected bit rate of *lossless* compression: entropy $H[\mathbf{X}]$
- ▶ we proved that $H[\mathbf{X}]$ is both:
 - ▶ a lower bound: $\mathbb{E}[\text{bitrate}(\mathbf{X})] \geq H[\mathbf{X}] = \mathbb{E}[-\log P(\mathbf{X})] \quad \forall \text{lossless codes}$
 - ▶ achievable with negligible overhead: $\exists \text{lossless code : } \text{bitrate}(\mathbf{x}) \geq -\log P(\mathbf{x}) + 1 \quad \forall \mathbf{x}$
- ▶ *Lossy* compression can have bit rates $< H[\mathbf{X}]$.
 - ▶ today and problem set: lower bound
 - ▶ next week: achievability of lower bound

Lower bound on the Bit Rate of Lossy Compression



- ▶ Encoder/decoder form a *Markov chain*:
- ▶ Problem 10.3: data processing inequality: \forall Markov chains:
- ▶ Lower bound on bit rate: