Faculty of Science · Department of Computer Science · Group of Prof. Robert Bamler

Lecture 10, Part 1:

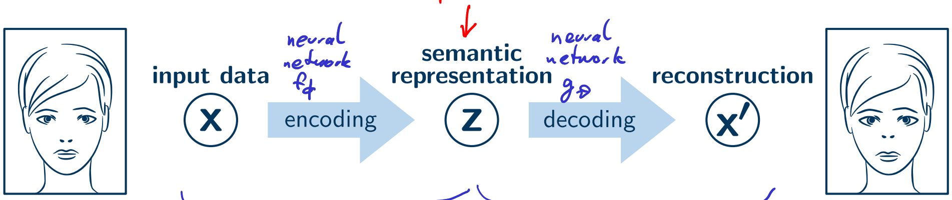
Lossy Compression: From VAEs to Rate/Distortion Theory

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)

prior $P(Z)$



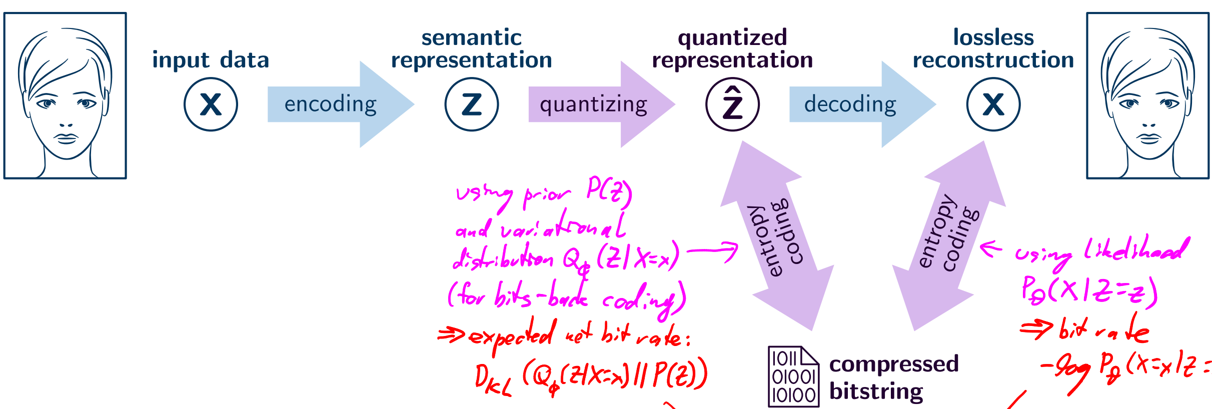
neural network f_ϕ

neural network g_θ

variational distribution $Q_\phi(Z|X)$

likelihood $P_\theta(X|Z)$

Problem Set 9: Lossless Compression With a VAE



using prior $P(Z)$ and variational distribution $Q_\phi(Z|X{=}x)$ (for bits-back coding)

$\Rightarrow$ expected net bit rate: $D_{KL}\big(Q_\phi(Z|X{=}x)\,\|\,P(Z)\big)$

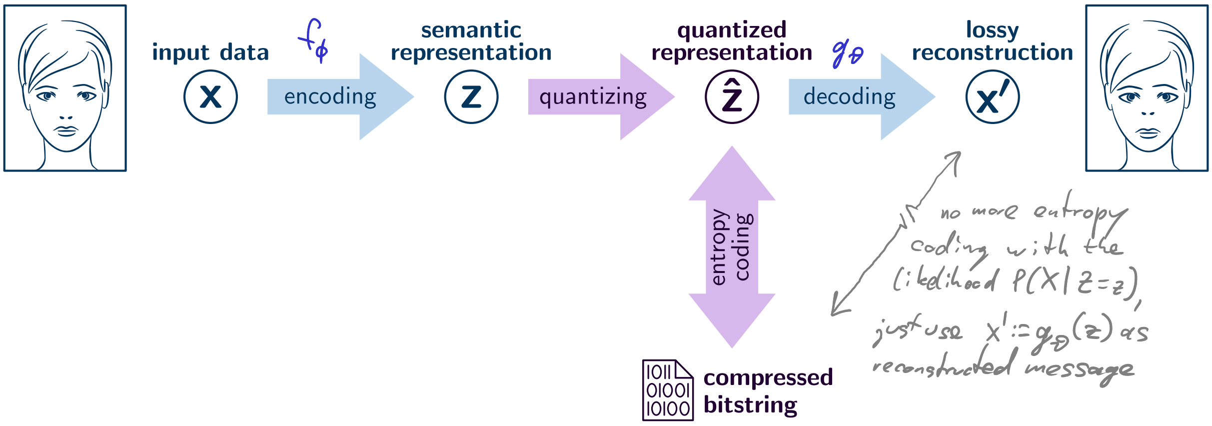
using likelihood $P_\theta(X|Z{=}z)$

$\Rightarrow$ bit rate $-\log P_\theta(X{=}x|Z{=}z)$

training objective ("loss function"):

$$\mathcal{L}(\theta,\phi) = \mathbb{E}_{x\sim\text{training set}}\Big[\mathbb{E}_{Q_\phi(Z|X=x)}\big[-\log P_\theta(X{=}x|Z{=}z)\big] + D_{KL}\big(Q_\phi(Z|X{=}x)\,\|\,P(Z)\big)\Big] = -\mathbb{E}_x\big[\text{ELBO}(\theta,\phi,x)\big]$$

Today: *Lossy* Compression With a VAE



VAEs for Lossy (Image) Compression

- **Example:** data $\mathbf{X} = (X_{i,j,k})$ are color images.
  - $(i,j,k) = $ (x-position, y-position, 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: $P_\theta(\mathbf{X} \mid \mathbf{Z}) = \prod_i P_\theta(X_{i,j,k} \mid \mathbf{Z})$
  - This time: Gaussian likelihood, i.e., density function $p_\theta(x_{i,j,k} \mid \mathbf{Z}=\mathbf{z}) = \underbrace{\mathcal{N}(x_{i,j,k}; g_\theta(\mathbf{z})_{i,j,k}, \tfrac{\beta}{2} I)}$

$$= \prod_{i,j,k} \frac{1}{\sqrt{2\pi \beta/2}} \exp\Big[-\frac{1}{\beta}\big(x_{i,j,k} - g_\theta(z)_{i,j,k}\big)^2\Big]$$

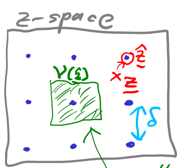
- **Idea:** just use $g_\theta(\mathbf{z})$ as the reconstruction of an image.
  (Don't bother using the likelihood $P_\theta(\mathbf{X} \mid \mathbf{Z}=\mathbf{z})$ to encode the true image.)
- $\text{ELBO}_\beta(\theta, \phi, \mathbf{x}) = \mathbb{E}_{Q_\phi(\underline{z} \mid \underline{X}=\underline{x})}\Big[\sum_{i,j,k} \log p_\theta(x_{i,j,k} \mid \underline{z})\Big] - D_{KL}\big(Q_\phi(\underline{Z} \mid \underline{X}=\underline{x}) \,\|\, P_\theta(\underline{Z})\big)$

$$\propto - \underbrace{\mathbb{E}_{\underline{z} \sim Q_\phi(\underline{z} \mid \underline{X}=\underline{x})}\big[\|\underline{x} - g_\theta(\underline{z})\|_2^2\big]}_{\text{"distortion" (reconstruction error)}} - \beta \underbrace{D_{KL}\big(Q_\phi(\underline{Z} \mid \underline{X}=\underline{x}) \,\|\, P_\theta(\underline{Z})\big)}_{\text{(bit) rate}}$$

Problem 10.1 (a)

Quantizing Latent Space

- Latents $\mathbf{z} \in \mathbb{R}^d$ are *continuous* $\Longrightarrow$ can't be entropy coded ($\mathbb{R}^d$ is not countable but $\{0,1\}^*$ is $\Rightarrow \nexists$ injective mapping $\mathbb{R}^d \to \{0,1\}^*$)
- Problem 9.2: for *lossless* compression with bits-back coding,
  we can simply quantize $\mathbf{z}$ to an arbitrarily fine grid.
  $D_{KL}\big(Q_\phi(\mathbf{Z} \mid \mathbf{X}=\mathbf{x}) \,\big|\, P_\theta(\mathbf{Z})\big) = \mathbb{E}_{Q_\phi(\mathbf{Z} \mid \mathbf{X}=\mathbf{x})}\big[\log Q_\phi(\mathbf{Z} \mid \mathbf{X}=\mathbf{x}) - \log P_\theta(\mathbf{Z})\big]$



$$P_\theta(\hat{\underline{Z}} = \hat{\underline{z}}) := \int_{V(\hat{\underline{z}})} p_\theta(\underline{z})\, d\underline{z} \underset{\text{(for small } \delta)}{\approx} \delta^d\, p_\theta(\hat{\underline{z}}); \quad \text{analogously for } Q_\phi(\hat{\underline{z}} \mid \underline{X}=\underline{x})$$

pdf of $P(\underline{z})$

all points $\underline{z}$ in this area would be rounded to the grid point $\hat{\underline{z}}$ in its center; area has volume δ^d ← dim($\underline{z}$) ↖ grid spacing

  - bit rate for encoding $\hat{\mathbf{z}} = \lceil \mathbf{z} \rfloor_\delta$ with quantized prior $P_\theta(\hat{\mathbf{Z}})$: $-\log P_\theta(\hat{\underline{Z}}=\hat{\underline{z}}) \approx -d\log\delta - \log p_\theta(\hat{\underline{z}})$
  - bit rate for decoding $\hat{\mathbf{z}}$ with quantized var. dist. $Q_\phi(\hat{\mathbf{Z}} \mid \mathbf{X}=\mathbf{x})$: $-d\log\delta - \log q_\phi(\hat{\underline{z}} \mid \underline{X}=\underline{x})$

  $\Longrightarrow$ for $\delta \to 0$, expected net bit rate is $D_{KL}\big(Q_\phi(\mathbf{Z} \mid \mathbf{X}=\mathbf{x}) \,\big|\, P_\theta(\mathbf{Z})\big)$ (independent of δ)
  ($-d\log\delta$ cancels)
- **Problem:** bits-back coding does not work (out of the box) for lossy compression.
  - Receiver never recovers the exact message $\mathbf{x} \Longrightarrow$ can't encode $\hat{\mathbf{z}}$ with $Q_\phi(\hat{\mathbf{Z}} \mid \mathbf{X}=\mathbf{x})$.
  - Thus, bit rate would depend on how fine we make the grid. (finer grid → higher bit rate)

Simplest Solution: Uniform Quantization [Ballé et al., 2017]

- **Idea:** take quantization into account already during training.
  - **Goal:** model should learn to encode important information on length scales $\geq \delta$.
  - **Problem:** quantization $\mathbf{z} \mapsto \hat{\mathbf{z}} := \lceil \mathbf{z} \rfloor_\delta$ is not differentiable.
  - **Observation:** $(\hat{\mathbf{z}} - \mathbf{z}) \in \big[-\frac{\delta}{2}, \frac{\delta}{2}\big]^d$ and (empirically) approximately uniformly distributed.

Problem 10.1 (b)

- **Proposal:** at training, replace quantization by adding uniform noise $\boldsymbol{\epsilon} \sim \mathcal{U}\Big(\big[-\frac{\delta}{2}, \frac{\delta}{2}\big]^d\Big)$
  - Equivalent to using a box-shaped variational distribution with fixed width:

$$Q_\phi(\underline{Z} \mid \underline{X} = \underline{x}) = \prod_i Q_\phi(Z_i \mid \underline{X} = \underline{x}) \quad \text{with pdfs} \quad q_{\phi,i}(z_i \mid \underline{X} = \underline{x}) = \mathcal{U}\Big(z_i; \big[f_\phi(\underline{x})_i - \tfrac{\delta}{2}, f_\phi(\underline{x})_i + \tfrac{\delta}{2}\big]\Big)$$

  - δ is fixed (i.e., data-independent)

    $\Longrightarrow$ might as well set $\delta = 1$ as long as the prior $P_\theta(\mathbf{Z})$ does not impose any fixed length scale.

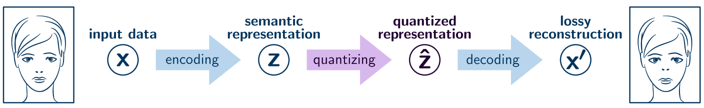
    i.e., length scales in $P_\theta(\underline{z})$ need to be learnable (Problem 10.1(c))

    $\Longrightarrow$ at deployment, encode each component $\hat{z}_i := \lceil z_i \rfloor_{\mathbb{Z}}$ with model $P_\theta(\hat{Z}_i = \hat{z}_i) = \underbrace{\int_{\hat{z}_i - \frac{1}{2}}^{\hat{z}_i + \frac{1}{2}} p_{\theta,i}(z_i)\, dz_i}_{=: \tilde{p}_{\theta,i}(\hat{z}_i)}$ → Problem 10.1(c)

    $\Longrightarrow$ bit rate $-\log P_\theta(\hat{\mathbf{Z}} = \hat{\mathbf{z}}) = -\sum_i \log \tilde{p}_{\theta,i}(\hat{z}_i) \approx \mathbb{E}_{\underline{z} \sim Q_\phi(\underline{Z} \mid \underline{X} = \underline{x})}\Big[-\sum_i \log \tilde{p}_{\theta,i}(z_i)\Big]$

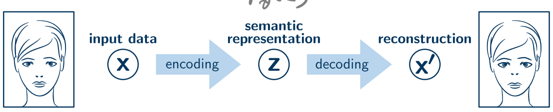
Model Training: Rate/Distortion Trade-Off

- **In deployment:**



  - quantized latent representation: $\hat{\mathbf{z}} := \lceil \mathbf{z} \rfloor = \lceil f_\phi(\mathbf{x}) \rfloor$
  - bit rate: $-\log P_\theta(\hat{\mathbf{Z}} = \hat{\mathbf{z}}) = -\log \Big(\underbrace{\int_{\mathbf{z} \in [\hat{\mathbf{z}} - \frac{1}{2}, \hat{\mathbf{z}} + \frac{1}{2}]} p_\theta(\mathbf{z})\, d\mathbf{z}}_{= \tilde{p}_\theta(\underline{z})}\Big)$
  - reconstructed message: $\mathbf{x}' = g_\theta(\hat{\mathbf{z}})$

- **At training time:**

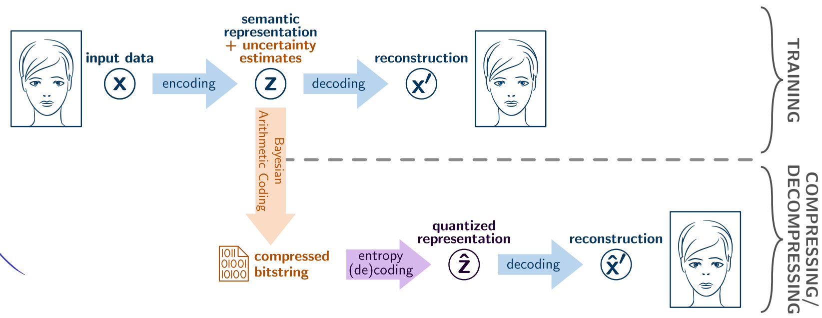


  - added uniform noise: $\mathbf{z} = f_\phi(\mathbf{x}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{U}\big(\big[-\frac{1}{2}, \frac{1}{2}\big]^d\big)$
  - approximated bit rate: $\mathcal{R}(\theta, \phi, \mathbf{x}) := -\log \tilde{p}_\theta(\mathbf{z})$ where $\tilde{p}_\theta(\mathbf{z}) := \int_{\mathbf{z}' \in [\mathbf{z} - \frac{1}{2}, \mathbf{z} + \frac{1}{2}]} p_\theta(\mathbf{z}')\, d\mathbf{z}'$
  - reconstruction error (distortion): e.g., MSE: $\mathcal{D}(\theta, \phi, \mathbf{x}) := \|g(\mathbf{z}) - \mathbf{x}\|_2^2$
  - loss function: rate/distortion trade off: $\mathcal{L}_\beta(\theta, \phi, \mathbf{x}) = \mathbb{E}_{\boldsymbol{\epsilon}}\big[\beta\, \mathcal{R}(\theta, \phi, \mathbf{x}) + \mathcal{D}(\theta, \phi, \mathbf{x})\big]$

    $\longrightarrow$ Problem 10.1: $\exists$ probabilistic model such that $\mathcal{L}_\beta(\theta, \phi, \mathbf{x}) \propto -\mathrm{ELBO}(\theta, \phi, \mathbf{x}) + \text{const}$

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*. ⇒ decoder must have several trained models with various β-values saved, and then load the appropriate one for a given message
  - less studied in the literature.
  - variational inference can help: [Yang et al., 2020, Tan & Bamler, 2022]



Lecture 10, Part 2:

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: $\mathbb{E}[\text{bitrate}(\mathbf{X})] \geq H[\mathbf{X}] = \mathbb{E}[-\log P(\mathbf{X})] \quad \forall$ lossless codes
  - achievable with negligible overhead: $\exists$ lossless code : $\text{bitrate}(\mathbf{x}) < -\log P(\mathbf{x}) + 1 \; \forall \mathbf{x}$
- *Lossy* compression can have bit rates $< H[\mathbf{X}]$.
  - today and problem set: lower bound
  - next week: achievability of lower bound (+ implications on channel coding)

Lower Bound on the Bit Rate of Lossy Compression

(encoder & decoder are usually deterministic, but treating them as conditional prob. dist is more general & turns out to simplify the discussion)

- Encoder/decoder form a *Markov chain:*

  original message $X \xrightarrow[P(S\,|\,X)]{\text{encoder}}$ bit string $S \xrightarrow[P(X'\,|\,S)]{\text{decoder}}$ reconstruction X'

- Problem 10.3: data processing inequality:

  $\forall$ Markov chains $X_1 \to X_2 \to X_3$: $I_p(X_1; X_3) \leq \begin{cases} I_p(X_1; X_2) \\ I_p(X_2; X_3) \end{cases}$ (both hold)

- Thus, lower bound on expected bit rate:
  - consider data source $P(\mathbf{X})$ and fixed mapping $P(\mathbf{X}' \,|\, \mathbf{X})$ from messages to reconstructions;
  - encoder $P(\mathbf{S} \,|\, \mathbf{X})$ and decoder $P(\mathbf{X}' \,|\, \mathbf{S})$ satisfy: $\sum_s P(X', S \,|\, X) = \sum_s P(S{=}s \,|\, X)\, P(X' \,|\, S{=}s) = P(X' \,|\, X)$

• source coding theorem: $\mathbb{E}_p[\underbrace{\text{length}(S)}_{\text{"bit rate"}}] \geq H_p(S)$ (assuming unique decodability)

• data processing ineq.: $I_p(X; X') \leq I(X; S) \underset{\text{Problem 4.4(c)}}{=} H_p(S) - H_p(S \,|\, X) \leq H_p(S)$

$\Rightarrow \boxed{\mathbb{E}_p[\text{bit rate}] \geq I_p(X; X')}$ ← lower bound on expected bit rate of lossy compression