Single-Shot Lossy Compression for Joint Inference and Reconstruction

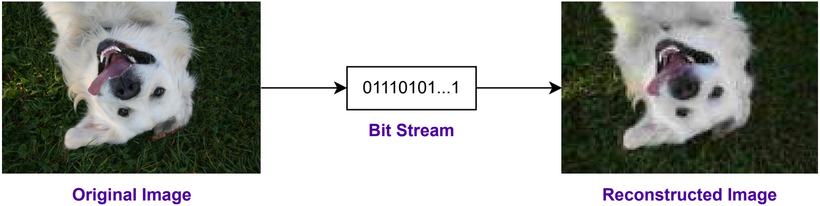
Oğuzhan Kubilay Ülger, Elza Erkip

New York University

09/28/2023

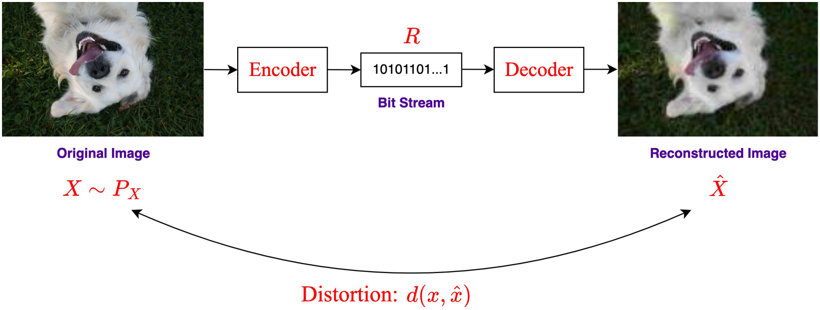
59th Annual Allerton Conference On Communication, Control, and Computing

Compression and Reconstruction



- We compress data to store and transmit it efficiently.
- Encoder: Observes data, turns it into a bit stream.
- Decoder: Takes the bit stream and reconstructs the source.
- Reconstruction may be **lossy**.

Compression and Reconstruction



- Loss is determined by a suitable distortion metric.
- Trade-off between Loss and Rate (# of bits).

Compression and Reconstruction

Rate-Distortion Function:

$$R(D) = \min_{\substack{P_{\hat{X}|X}: \\ \mathbb{E}[d(X,\hat{X})] \leq D}} I(X;\hat{X})$$

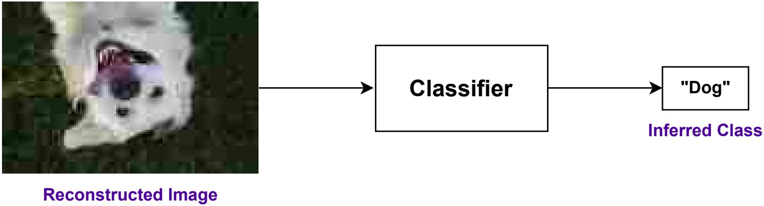
- Optimal asymptotic rate-distortion trade-off
- n i.i.d. samples X^n, are compressed together
- Per Sample Rate vs. Expected Distortion

Compression and Reconstruction

What next?

- In many cases we will use the reconstructed source in **further tasks**.
- We may want to **infer** more information.
- This **inference** is usually done at the decoder.

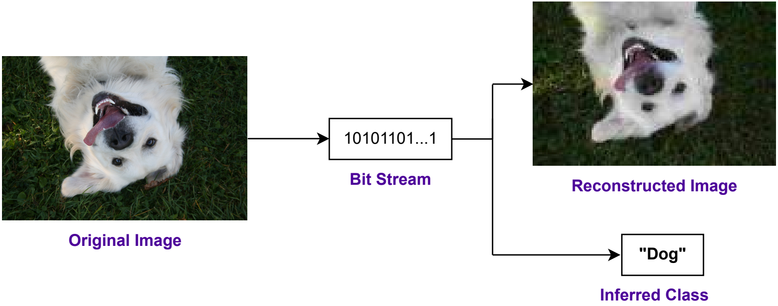
Inference After Reconstruction



- This is suboptimal!
- We did not consider the inference task when compressing.
- Possibly results in high classification error.

Joint Inference and Reconstruction

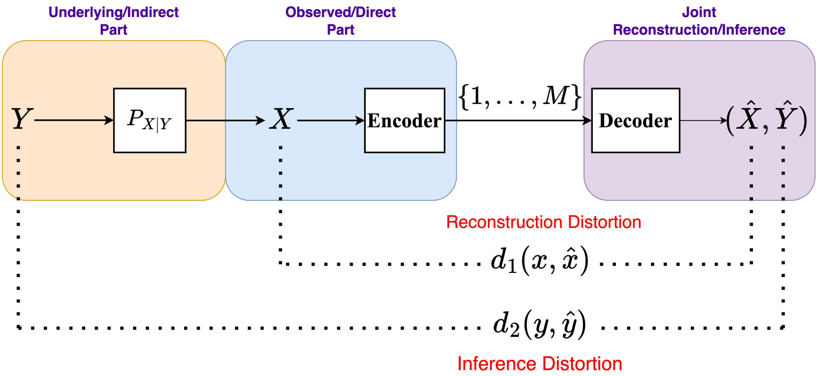
We can do better if we know the task beforehand!



More Examples

- Compression for Humans + Machines
- Speech compression + underlying text
- Speaker Identification

Joint Inference and Reconstruction



Asymptotic Rate Distortion

Asymptotic rate-distortion result with expected distortion constraints [Liu et al. 2022]

Asymptotic RD

$$R(D_1, D_2) = \min_{P_{\hat{X}\hat{Y}|X}:} I(X; \hat{X}, \hat{Y})$$
$$\mathbb{E}[d_1(X, \hat{X})] \leq D_1$$
$$\mathbb{E}[\tilde{d}_2(X, \hat{Y})] \leq D_2.$$

where $\tilde{d}_2(x, \hat{y}) = \mathbb{E}[d_2(Y, \hat{y})|X = x]$.

- We are interested in **single-shot compression**: Low latency applications, modern neural compressors.

Goal

- Encoder and decoder pair:

$$f : \mathcal{X} \longrightarrow \{1, \ldots, M\}$$
$$g : \{1, \ldots, M\} \longrightarrow \hat{\mathcal{X}} \times \hat{\mathcal{Y}}.$$

- The output of encoder and decoder $g(f(X)) = (\hat{X}, \hat{Y})$.
- **Excess distortion probability**:

$$\mathbb{P}[\{d_1(X, \hat{X}) > D_1\} \cup \{d_2(Y, \hat{Y}) > D_2\}] \leq \epsilon.$$

- **Main goal** is to characterize:

$$\epsilon^*(M, D_1, D_2) : \text{minimum achievable } \epsilon \text{ given } (M, D_1, D_2).$$

Achievability: Proof Sketch

- Fix a codebook distribution $P_{\hat{X}\hat{Y}}$.
- Generate M codeword pairs $\{(c_{1,1}, c_{1,2}), \ldots, (c_{M,1}, c_{M,2})\}$.
- Encoder simply sends the index.

$$i^* \in \arg\min_{i \in 1,\ldots,M} \pi(x, c_{i,1}, c_{i,2})$$

$$\pi(x, \hat{x}, \hat{y}) = \mathbb{P}\left[\{d_1(X, \hat{x}) > D_1\} \cup \{d_2(Y, \hat{y}) > D_2\} | X = x\right].$$

- Decoder outputs $(c_{i*,1}, c_{i*,2})$
- We then take average over all random codebooks and optimize over distributions.

Achievability

Theorem (Achievability)

$$\epsilon^*(M, D_1, D_2) \leq \inf_{P_{\hat{X}\hat{Y}}} \int_0^1 \mathbb{E}\left[\mathbb{P}\left[\pi(X, \hat{X}, \hat{Y}) > t \middle| X\right]^M\right] dt$$

where the infimum is taken over all distributions $P_{\hat{X}\hat{Y}}$ independent of X and

$$\pi(x, \hat{x}, \hat{y}) = \mathbb{P}\left[\{d_1(X, \hat{x}) > D_1\} \cup \{d_2(Y, \hat{y}) > D_2\} | X = x\right].$$

Converse

Definition

The joint (D_1, D_2)-tilted information is defined as:

$$\jmath_{X;\hat{X}\hat{Y}}(x, y, \hat{x}, \hat{y}, D_1, D_2) = \imath_{X;\hat{X}\hat{Y}}(x; \hat{x}, \hat{y}) + \lambda_1(d_1(x, \hat{x}) - D_1) + \lambda_2(d_2(y, \hat{y}) - D_2)$$

$$\imath_{X;Y}(x; \hat{x}, \hat{y}) = \log \frac{P_{X|Y}(x|\hat{x}, \hat{y})}{P_X(x)}.$$

Converse

Theorem (Converse)

$$\epsilon^*(M, D_1, D_2) \geq \inf_{P_{\hat{X}\hat{Y}|X}} \sup_{\gamma \geq 0} \left\{ \mathbb{P}\left[\jmath_{X;\hat{X}^*,\hat{Y}^*}(X, Y, \hat{X}, \hat{Y}, D_1, D_2) \geq \log M + \gamma \right] - 2^{-\gamma} \right\}$$

where $\jmath_{X;\hat{X}^*,\hat{Y}^*}(x, y, \hat{x}, \hat{y}, D_1, D_2)$ *is defined according to some distribution* $P_{\hat{X}^*\hat{Y}^*|X}$ *that achieves the asymptotic rate-distortion function* $R(D_1, D_2)$.

Logarithmic Loss

- $\hat{\mathcal{X}}$ is the set of probability distributions on $\mathcal{X}$
- $\hat{X}$ is a distribution, soft decision.
- Logarithmic loss (log-loss) is defined as:

$$\mathrm{d}(x, \hat{x}) = \log \frac{1}{\hat{x}(x)}$$

Example

$\mathcal{X} = \{0, 1\}$
If $x = 1$ and our soft decisions are $\hat{x}(1) = 0.8$ and $\hat{x}(0) = 0.2$
So our log-loss will be $\mathrm{d}(x, \hat{x}) = \log \frac{1}{0.8} = 0.32$

Logarithmic Loss

- We set the direct distortion metric $d_1(x, \hat{x})$ as log-loss.
- This gives us some freedom on our encoder design.
- For a distortion threshold D_1,

$$\hat{x}(x) \geq \exp(-D_1)$$

- A single $\hat{x}$ can cover $\lfloor \exp(D_1) \rfloor$, x values

Example

$\mathcal{X} = \{1, 2, 3, 4, 5\}$ and $D_1 = \log 4$.
If $\hat{x}(1) = \hat{x}(2) = \hat{x}(3) = \hat{x}(4) = 0.25$.
$d_1(x, \hat{x}) \leq D_1$ for $x = 1, 2, 3, 4$.

Achievability: Log-Loss

Theorem (Achievability for Log-Loss)

$$
\begin{aligned}
\epsilon^*(M, D_1, D_2) \leq \inf_{P_{\hat{Y}}} \inf_{\gamma \geq 0} \inf_{0 \leq \epsilon' \leq 1} \Big\{ & \epsilon' \left(1 - \mathbb{E}\left[\eta(\epsilon')^M\right]\right) \\
& + \mathbb{E}\left[\eta(\epsilon')^M\right]\left(1 + 2^{1-\gamma}\right) \\
& + 2^{1-\gamma} \sum_{k=1}^{M} \binom{M}{k} \frac{M}{k} \mathbb{E}[\eta(\epsilon')^{M-k}(1-\eta(\epsilon'))^k] \\
& + \mathbb{P}[\imath_X(X) > D_1 + \log M - \gamma] \Big\}
\end{aligned}
$$

where

$$
\begin{aligned}
\eta(\epsilon') &= \mathbb{P}[\pi'(X, \hat{Y}) > \epsilon' | X] \\
\pi'(x, \hat{y}) &= \mathbb{P}[d_2(Y, \hat{y}) > D_2 | X = x].
\end{aligned}
$$

Converse: Log-Loss

Converse also simplifies using another property of log-loss.

- Reconstruction only: $R_1(D_1) = R(D_1, D_2 = \infty)$ (Direct RD).
- Inference only :$R_2(D_2) = R(D_1 = \infty, D_2)$ (Indirect RD).
- **Log-loss property**: $R(D_1, D_2) = \max(R_1(D_1), R_2(D_2))$.
- Not true for all distortion metrics!

Converse: Log-Loss

Theorem (Converse for Log-Loss)

For $R_2(D_2) < H(X) - D_1$,

$$\epsilon^*(M, D_1, D_2) \geq \sup_{\gamma \geq 0} \left\{ \mathbb{P}\left[\imath_X(X) \geq D_1 + \log M + \gamma\right] - 2^{-\gamma} \right\}$$

and for $R_2(D_2) \geq H(X) - D_1$,

$$\epsilon^*(M, D_1, D_2) \geq \inf_{P_{\hat{Y}|X}} \sup_{\gamma \geq 0} \left\{ \mathbb{P}\left[\jmath_{X;\hat{Y}^*}(X, Y, \hat{Y}, D_2) \geq \log M + \gamma\right] - 2^{-\gamma} \right\}.$$

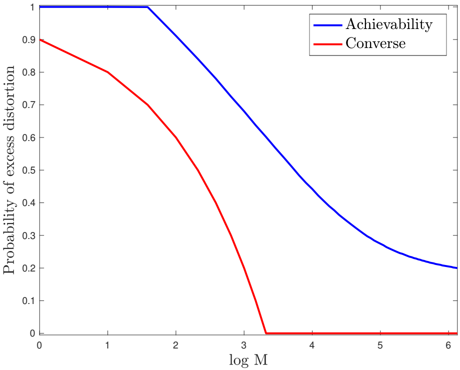
Numerical Example

- $|\mathcal{Y}| = 10$ and $|\mathcal{X}| = 7|\mathcal{Y}| = 70$.
- $Y \sim \text{Uniform}\{0, \ldots, 9\}$.

$$P_{X|Y}(x|y) = \begin{cases} \phi(x - 6y), & x \in [7y, 7y + 6] \\ 0, & \text{otherwise} \end{cases}$$

$$\phi(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x \in [0, n] \\ 0, & \text{otherwise.} \end{cases}$$

- X is a Binomial RV, alphabet is determined by Y which represents its class.
- $d_1(x, \hat{x})$ is Log-loss while $d_2(y, \hat{y})$ is Hamming distortion.



Conclusion & Future Work

This work:

- We explored a single shot compression setting that jointly considers direct and indirect source coding.
- We provided some achievability and converse bounds for excess distortion probability.

In Future:

- Improve on the achievability result, especially for log-loss
- Consider a case where the inference task is unknown at the encoder (among many tasks).

Thank you for listening!
Q&A

For further questions: kubi@nyu.edu