Universal Rate-Distortion-Classification Representations for Lossy Compression

Nam Nguyen[1], Thuan Nguyen[2], Thinh Nguyen[1], and Bella Bose[1]

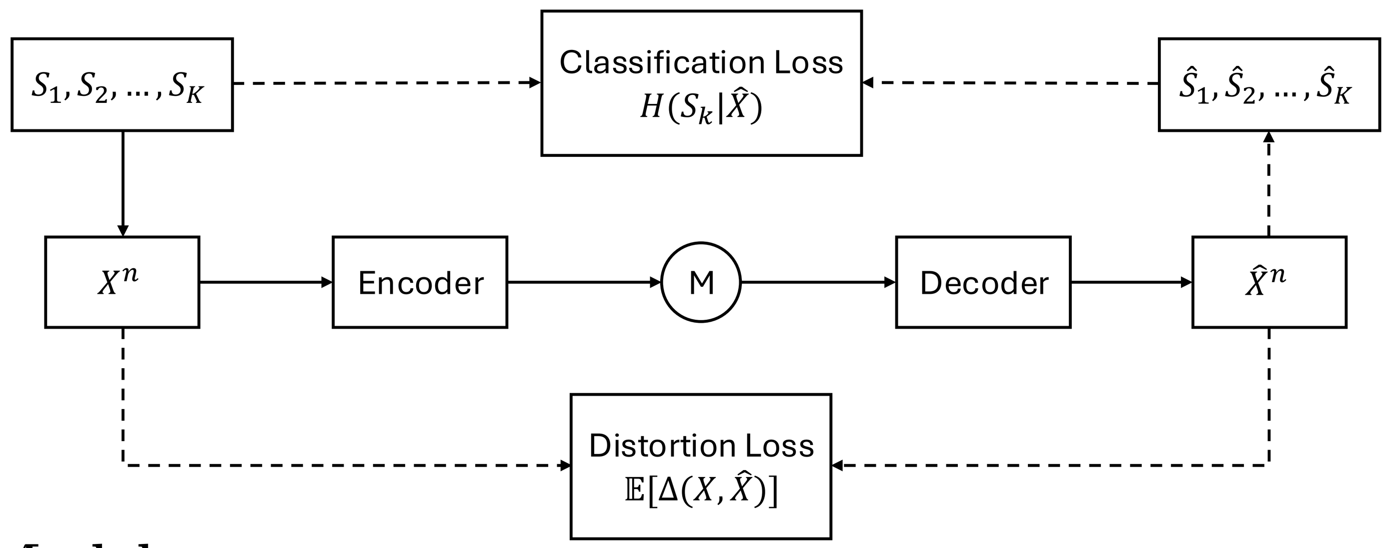
[1]School of Electrical and Computer Engineering, Oregon State University, Corvallis, OR, 97331

[2]Department of Engineering, Engineering Technology, and Surveying, East Tennessee State University, Johnson, TN, 37604

Oregon State University

IEEE Information Theory Society

Multi-Task Learning Lossy Compression



Model:

- Source: $X \sim p_X(x)$.
- Target labels: $S_1, \cdots, S_K \sim p_S(s_1, \cdots, s_K)$, where $p_{X,S}(x, s_1, \cdots, s_K)$.

Lossy Compression: $X_1, X_2, \cdots, X_n \overset{\text{i.i.d}}{\sim} p(x)$.

- Encoder: $f : \mathcal{X}^n \mapsto \{1, 2, \cdots, 2^{nR}\}$ maps the source X^n to a message M.
- Decoder: $g : \{1, 2, \cdots, 2^{nR}\} \mapsto \hat{\mathcal{X}}^n$ reproduces data $\hat{X}^n$.

Universal Rate-Distortion-Classification

Definition 1. DCR Function

$$D(C,R) = \min_{p_{\hat{X}|X}} \mathbb{E}[(X-\hat{X})^2] \quad (1a)$$
$$\text{s.t.} \quad I(X;\hat{X}) \le R, \quad (1b)$$
$$H(S|\hat{X}) \le C. \quad (1c)$$

Let $\Omega(R) = \{(D,C) : R(D,C) \le R\}$.

Definition 2. Universal RDC Function

Let Z be a **representation** of X by $p_{Z|X}$. For each $(D,C) \in \Theta$, $\exists p_{\hat{X}_{D,C}|Z}$: $\mathbb{E}[\Delta(X,\hat{X}_{D,C})] \le D$ and $H(S|\hat{X}_{D,C}) \le C$.

$$R(\Theta) = \inf_{p_{Z|X} \in \mathcal{P}_{Z|X}(\Theta)} I(X;Z). \quad (2)$$

Definition 3. Rate Penalty

$$A(\Theta) = R(\Theta) - \sup_{(D,C)\in\Theta} R(D,C), \quad (3)$$

where $\Omega(p_{Z|X}) = \left\{(D,C) : \exists p_{\hat{X}_{D,C}|Z} \text{ s.t. } \begin{array}{l} \mathbb{E}[\Delta(X,\hat{X}_{D,C})] \le D, \\ H(S|\hat{X}_{D,C}) \le C \end{array}\right\}.$

- Ideally, $A(\Theta) = 0$ for each R, meaning a **single encoder** suffices for the entire tradeoff.

Theoretical Results

Theorem 1. DCR for a Gaussian Source

A source $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and a classification variable $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$ with $\mathrm{Cov}(X,S) = \theta_1$.

$$D(C,R) = \begin{cases} \sigma_X^2 e^{-2R}, \;\; C > \frac{1}{2}\log\left(1 - \frac{\theta_1^2(\sigma_X^2 - \sigma_X^2 e^{-2R})}{\sigma_S^2\sigma_X^4}\right) + h(S) \\ \sigma_X^2 - \frac{\sigma_S^2\sigma_X^4}{\theta_1^2}\left(1 - e^{-2h(S)+2C}\right), \\ \quad \frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right) + h(S) \le C \le \frac{1}{2}\log\left(1 - \frac{\theta_1^2(\sigma_X^2 - \sigma_X^2 e^{-2R})}{\sigma_S^2\sigma_X^4}\right) + h(S) \\ 0, \quad C > h(S) \text{ and } R > h(X). \end{cases}$$

Theorem 2. No Rate-Penalty for a Gaussian Source

A source $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and a classification variable $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$ with $\mathrm{Cov}(X,S) = \theta_1$.

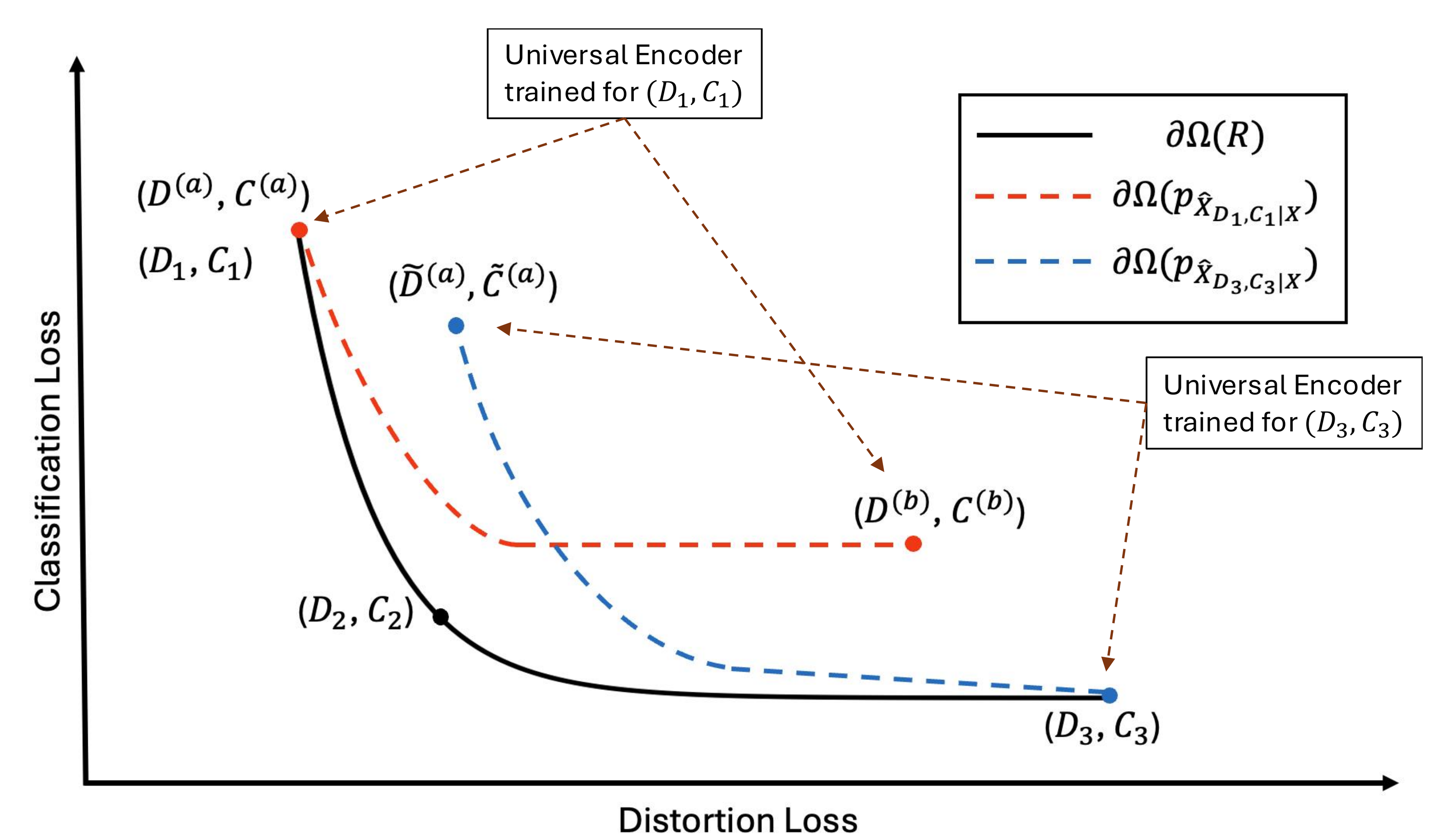
$$A(\Theta) = 0; \quad I(X;Z) = \sup_{(D,C)\in\Theta} R(D,C); \quad \Theta \subseteq \Omega(p_{Z|X}) = \Omega(I(X;Z)). \quad (4)$$

Theorem 3. Universality for a General Source

Let $X \sim p_X$ and $S \sim p_S$ with $\mathrm{Cov}(X,S) = \theta_1$. Define $\tilde{X} = \mathbb{E}[X|Z]$ as MMSE estimator.

$$\Omega(p_{Z|X}) \subseteq \left\{(D,C) : D \ge \mathbb{E}\|X - \tilde{X}\|^2 + \begin{array}{c} \inf_{p_{\hat{X}}} W_2^2(p_{\tilde{X}}, p_{\hat{X}}) \\ \text{s.t. } H(S|\hat{X}) \le C \end{array}\right\} \subseteq \mathrm{cl}(\Omega(p_{Z|X})). \quad (5)$$

.



Theorem 4. Quantitative Results

Let $\hat{X}_{D_1,C_1}$ be optimal reconstruction at (D_1, C_1) on the conventional RDC trade-off curve, satisfying $I(X;\hat{X}_{D_1,C_1}) = R(D_1,C_1)$. Then, $\Omega(p_{\hat{X}_{D_1,C_1}|X})$ satisfies $(D^{(a)}, C^{(a)}) = (D_1, C_1)$. Now, consider $(D^{(b)}, C^{(b)}) \in \Omega(p_{\hat{X}_{D_1,C_1}|X})$ and $(D_3, C_3) \in \Omega(R)$:

$$D_3 - D^{(b)} \ge \sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - 2\sigma_{\hat{X}_{D_3,C_3}}\sqrt{\sigma_X^2 - D_1} - 2D_1 \text{ and } \frac{D_3}{D^{(b)}} \ge \frac{\sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - 2\sigma_{\hat{X}_{D_3,C_3}}\sqrt{\sigma_X^2 - D_1}}{2D_1}. \quad (6)$$

If $W_2^2(p_X, p_{\hat{X}_{D_3,C_3}}) = 0$, i.e., $\sigma_X^2 = \sigma_{\hat{X}_{D_3,C_3}}^2$, the **distortion gap becomes small** under:

$$D_3 - D^{(b)} \approx 0 \text{ if } D_1 \approx 0 \text{ or } D_1 \approx \sigma_X^2 \quad \text{and} \quad \frac{D_3}{D^{(b)}} \approx 1 \text{ if } D_1 \approx \sigma_X^2. \quad (7)$$

Experimental Results

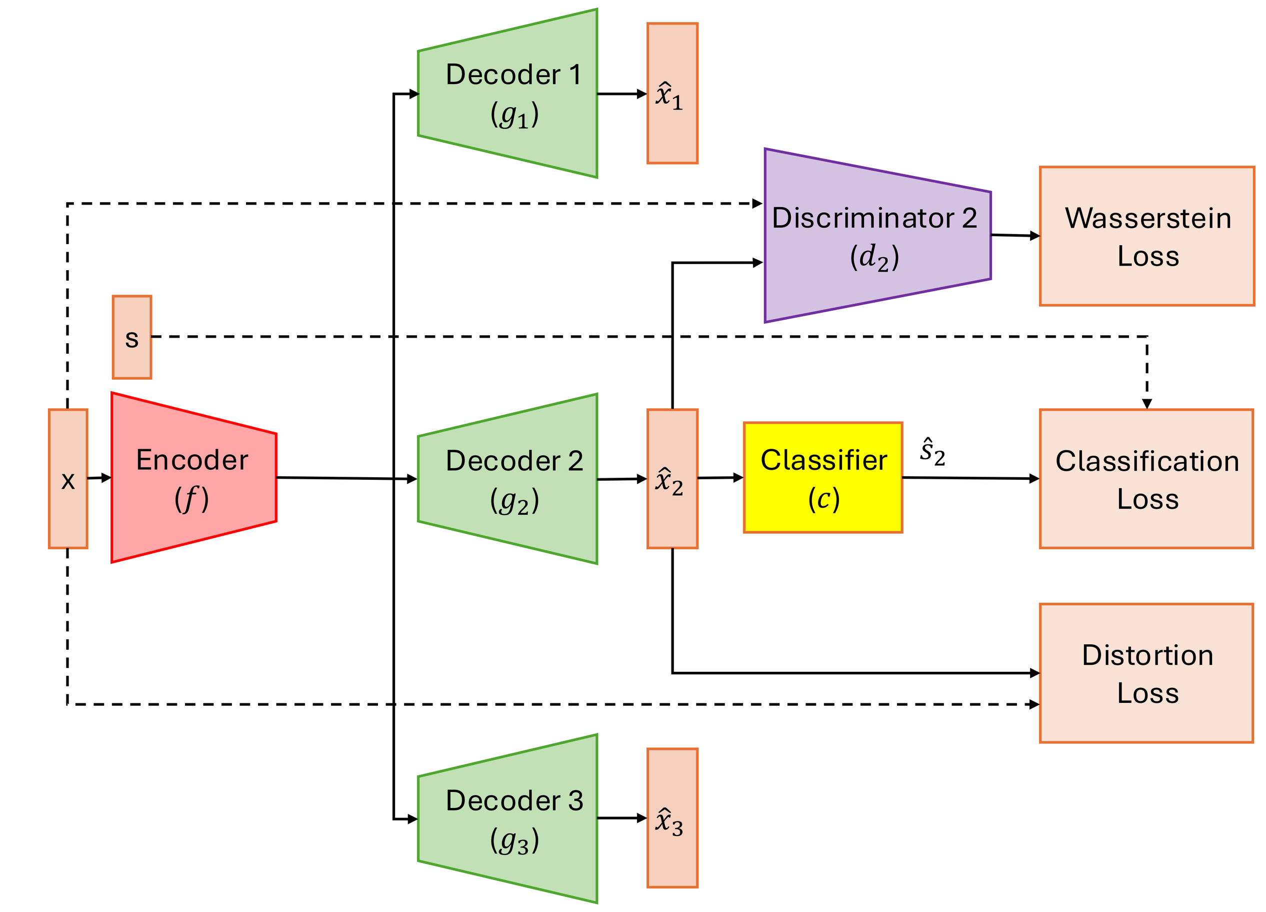
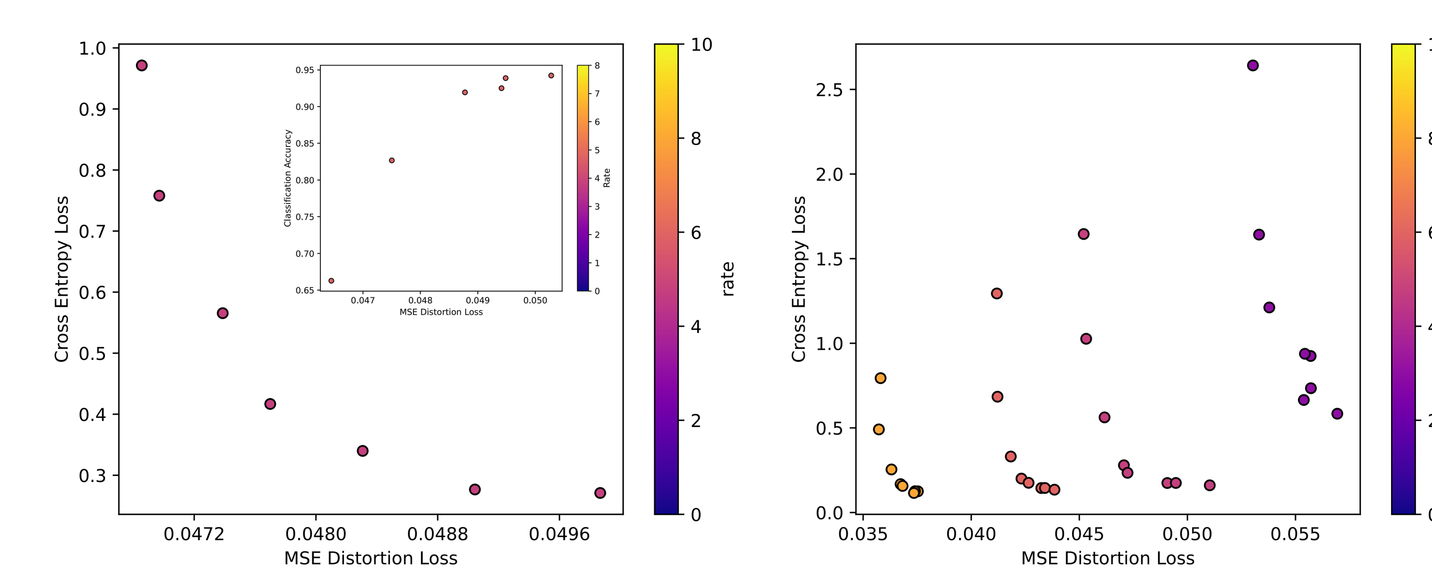


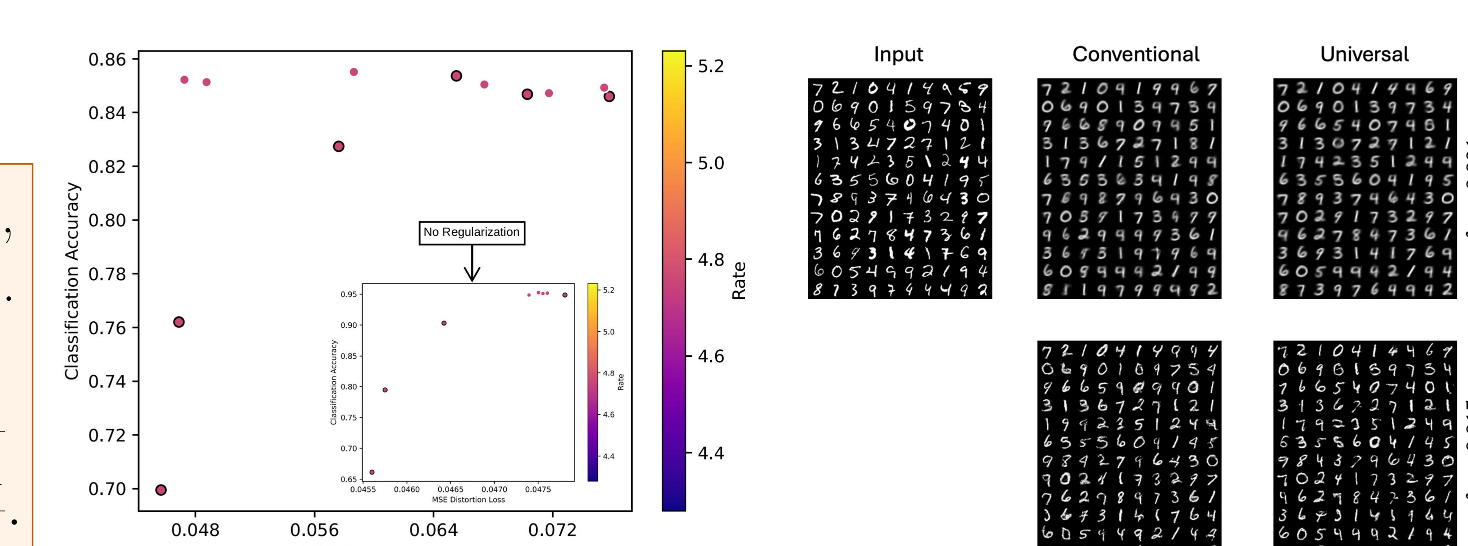
Figure: An illustration of the universal RDC scheme.

- Dataset: MNIST; compression rate: $R = d \times \log_2(L)$.
- Classifier (c) is pre-trained. Training Encoder (f), Decoder (g), and Discriminator (d) with this loss function:

$$\mathcal{L} = \lambda_d \mathbb{E}[\|X - \hat{X}\|^2] + \lambda_c \mathrm{CE}(S,\hat{S}) + \lambda_p W_1(p_X, p_{\hat{X}}).$$



(a) The RDC curve under rate $R = 4.75$. (b) The RDC curves at multiple rates.



(a) The RDC curve under rate $R = 4.75$. (b) Decompressed outputs.

Acknowledgments: This work is supported in part by the National Science Foundation Grant CCF-2417898.