AICIT 2024

2024 Hong Kong, Guangzhou and Taipei Joint Workshop on Artificial Intelligence, Communications and Information Theory

Source Coding Theorems with Semantic Computing-Oriented Criterion

Tingting Zhu and Xiao Ma

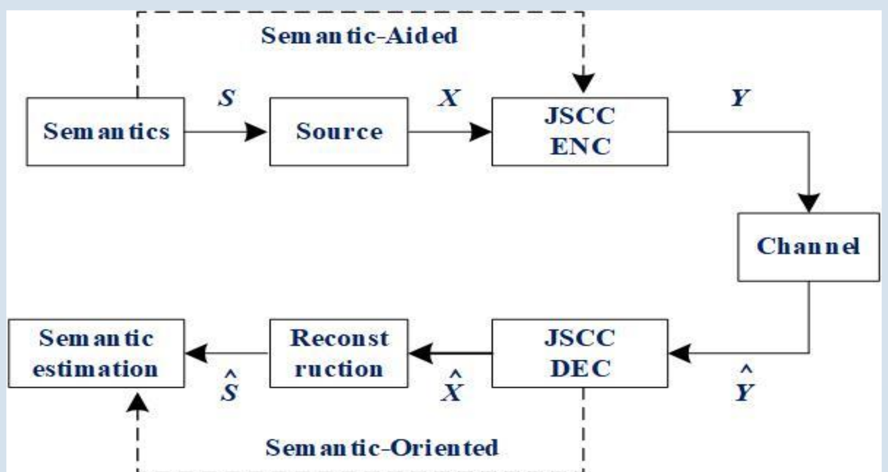
School of Computer Science and Engineering, Sun Yat-sen University

Motivation

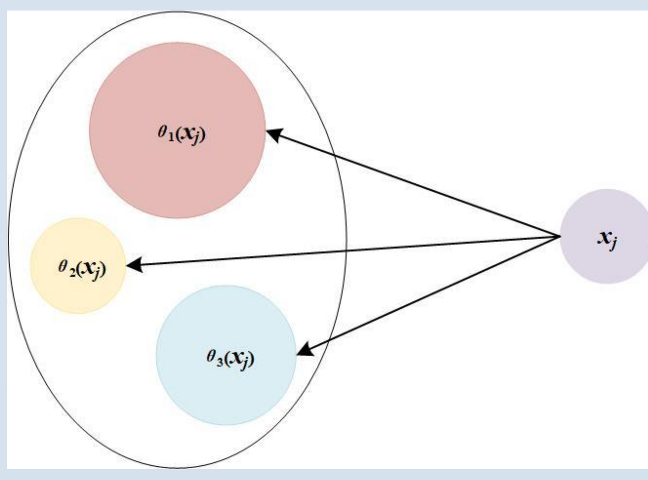
Based on the semantic communications system model, a new data compression problem arises: what is the theoretical limit of the source coding when the goal of communication is to compute semantic functions rather than reconstructing source symbols?

System Model

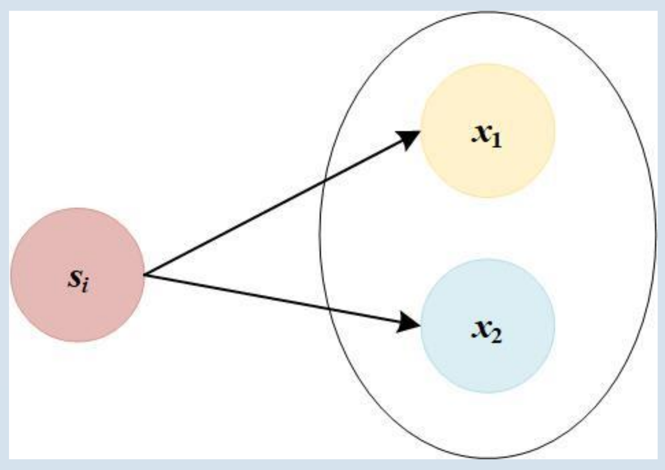
- **Observable source:** X
- **Unobservable semantics:** $S = \theta(X)$
- **JSCC Decoder:** $\hat{X}$
- **Semantic Decoder:** $\hat{S} = \theta(\hat{X})$
- **Distortion:** $d(S, \hat{S}) = d(\theta(X), \theta(\hat{X}))$
- **Goal:** $\min d(\theta(X), \theta(\hat{X}))$



➢ System Model



➢ Semantic representation of source symbol



➢ Semantic representation redundancy: one-to-many relationship between semantics and the observed sources

Discrete Source

➢ Definition 1 (Semantic entropy):

$$H(\theta_i(X)) = -\sum_j p(\theta_i(x_j)) \log_2 p\left(\theta_i(x_j)\right), 1 \le i \le |\mathcal{S}|, 1 \le j \le |\mathcal{X}|$$

➢ Theorem 1 (Lossless Source Coding Theorem with Semantic Computing-Oriented Criterion):

Given the observable source X and its semantic feature function $S_i = \theta_i(X)$, $1 \le i \le |\mathcal{S}|$, for each semantics S_i, all rates above the semantic entropy are achievable, and all rates below the semantic entropy are not; that is, for $R_i \ge H(\theta_i(X)) + \varepsilon, \varepsilon > 0$, there exists a sequence of codes $\mathcal{C}$ such that $R_{\mathcal{C}} \ge R_i$ and $\lim_{k\to\infty} \Pr\{\theta_i(\hat{X}^k) \ne \theta_i(X^k)\} = 0$. Conversely, for $R_{\mathcal{C}} < R_i$, the error probability is bounded away from 0.

➢ Corollary 1: $H(\theta_i(X)) \le H(X), 1 \le i \le |\mathcal{S}|$

➢ Corollary 2: $H(S_1, \ldots, S_M) \le H(S_1) + \cdots + H(S_M) \le MH(X)$

➢ Corollary 3: $H(S) < H(S) + H(X|S) = H(X)$

Table 1 PMF and Entropy of discrete source X

X	1	2	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/6	1/6	1/6
$H(X)$	2.59					

Table 2 Semantic feature function, semantic space and its probability distribution

$\theta_i(X)$	Semantic Space $\mathcal{S}_i$	Semantics s	$p(s)$	$H(S) = H(\theta_i(X))$
$\theta_1(X) = X \bmod 2$	$\mathcal{S}_1 = \{0, 1\}$	0	1/2	1
		1	1/2	
$\theta_2(X) = X \bmod 4$	$\mathcal{S}_2 = \{0, 1, 2, 3\}$	0	1/6	1.92
		1	1/3	
		2	1/3	
		3	1/6	

Table 3 Joint distribution of X and S

S \ X	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$
$S = 0$	0	1/6	0	1/6	0	1/6
$S = 1$	1/6	0	1/6	0	1/6	0

Continuous Source

➢ Definition 2 (Semantic rate distortion function):

$$R_{\theta_i}(D_{\theta_i}) = \min_{p(\hat{s}_i|s_i): \sum_{s_i, \hat{s}_i} p(s_i) p(\hat{s}_i|s_i) d(s_i, \hat{s}_i) \le D_{\theta_i}} I(S_i; \hat{S}_i), 1 \le i \le |\mathcal{S}|$$

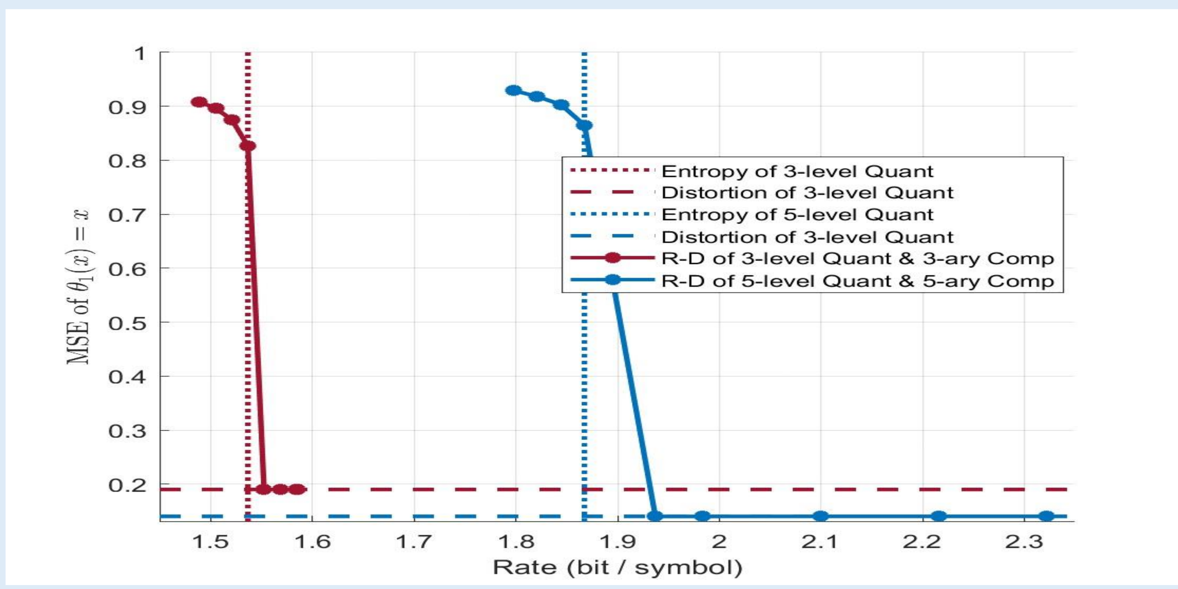
➢ Theorem 2 (Lossy Source Coding Theorem with Semantic Computing-Oriented Criterion):

For each semantics S_i, $1 \le i \le |\mathcal{S}|$, given $\varepsilon > 0$, if $R_{\theta_i} \ge R_{\theta_i}(D^*_{\theta_i}) + \varepsilon$, there exists a sequence of codes $\mathcal{C}$, such that the code rate $R_{\mathcal{C}} \ge R_{\theta_i}$ and the average semantic distortion is bounded by $D_{\theta_i} \le D^*_{\theta_i} + \varepsilon$. Conversely, for $R_{\mathcal{C}} < R_{\theta_i}$, then for any $\mathcal{C}$, the average semantic distortion $D_{\theta_i} > D^*_{\theta_i}$.

➢ Corollary 4: $R_{\theta_i}(D_{\theta_i}) \le I(S_i; \hat{S}_i) \le I(X; \hat{X}) = R(D)$

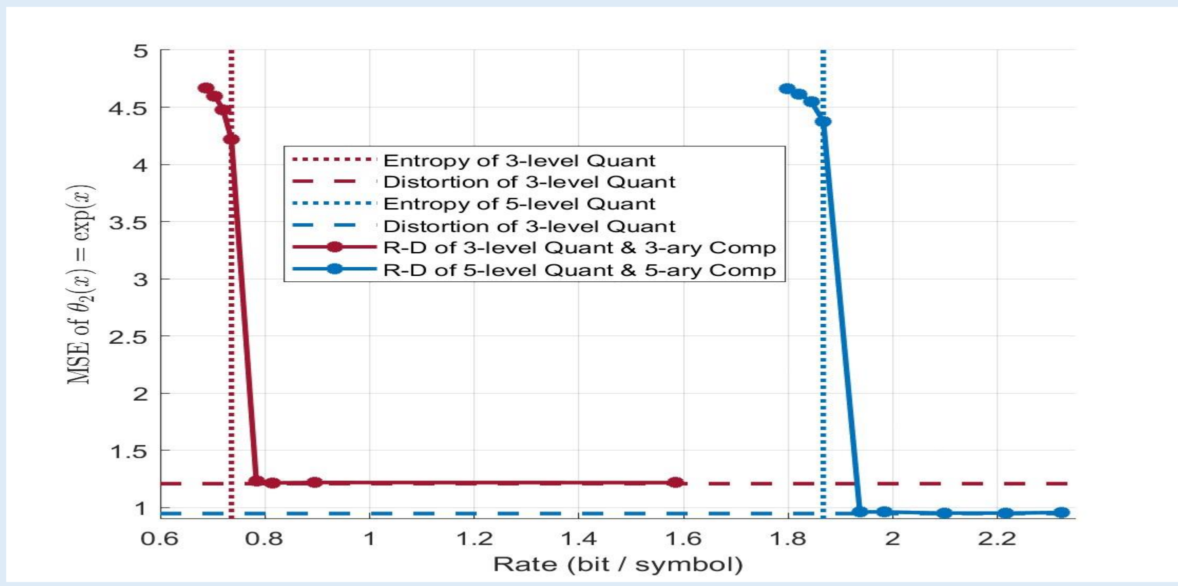
➢ **MSE of the semantic function**

$\theta_1(x) = x$



➢ **MSE of the semantic function**

$\theta_2(x) = \exp(x)$



➢ **Full paper is available: (in Chinese)**

Acknowledgement: This research was supported by the National Key R&D Program of China (No. 2021YFA1000500).