

语义通信性能评估体系及指标



Performance Evaluation Systems and Metrics for Semantic Communications

郑远/ZHENG Yuan, 王凤玉/WANG Fengyu,
许文俊/XU Wenjun

(北京邮电大学, 中国 北京 100876)
(Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTETJ.202302010

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20230411.1340.008.html>

网络出版日期: 2023-04-11

收稿日期: 2023-02-20

摘要: 鉴于语义通信系统缺乏统一的、具泛化价值的性能评估体系, 对现有研究中评估方法的使用情况与不足进行分析, 并提出语义通信效率指标 E_{sc} 和语义通信效用指标 U_{sc} 。其中, 指标 E_{sc} 用于衡量通信系统在给定通信资源下单位时间开销内的任务完成准确性, 指标 U_{sc} 用于衡量通信系统在给定通信资源下与任务性能上限的接近程度。与现有评估指标相比, 这两种指标具有更好的泛用性, 可为不同任务场景、模态信息下的语义通信模型横向性能对比提供指导。此外, 以语义图像重建与语音重建任务为例, 分别搭建端到端语义通信仿真模型, 并基于所提 E_{sc} 与 U_{sc} 指标对仿真模型性能进行评估。

关键词: 语义通信; 人工智能; 性能评估指标

Abstract: The absence of a unified and generalized performance evaluation system brings challenges to the realization of semantic communication systems. The usages and shortcomings of evaluation methods in existing research are analyzed. Two evaluation metrics, namely the semantic communication efficiency metric E_{sc} and the semantic communication utility metric U_{sc} , are proposed. E_{sc} measures the accuracy of task completion within unit time cost of a communication system under certain resources. U_{sc} measures how close a communication system is to the upper limit of task performance under certain resources. Compared with the existing evaluation methods, the two proposed metrics are of better generality, which can provide guidance for the performance comparison of semantic communication models under different scenarios and information modalities. In addition, the semantic image reconstruction and speech reconstruction are taken as examples to build end-to-end semantic communication simulation models respectively, and the performance of models based on the proposed E_{sc} and U_{sc} metrics is evaluated.

Keywords: semantic communication; artificial intelligence; performance evaluation metric

从20世纪80年代后期1G的诞生到如今5G商业化进程的稳步展开, 移动通信经历了近半个世纪的高速发展。得益于高峰值速率、低传输时延与高连接密度, 5G系统支持包含虚拟/增强现实、智能驾驶、智慧城市等在内的多样化垂直应用^[1]。随着全球新一轮科技与产业革命的加速发展, 信息通信技术将进一步重构人与人、人与物、物与物之间的复杂联系, 为工业、医疗、交通、教育、金融、娱乐、零售等千行百业带来革新。当前通信系统主要基于香农信息论研究设计。从1G到5G, 通信系统虽然在业务形式、服务对象、网络架构和承载资源等方面进行了技术变革, 但都依赖于增加信息传输的物理维度、通过技术堆叠处理来逼近香农信息论极限。

随着通信系统与各垂直行业领域的紧密结合, 面向个人、行业等用户的智能需求将被进一步挖掘, 这对通信系统准确传递信息含义提出更高要求。而以香农信息论为基础的传统通信系统并不关注信息中承载的含义, 只关注每个传输比特的正确接收。这导致不必要的通信资源耗费, 难以满足未来通信持续发展的需求。语义通信通过交叉融合人工智能与通信技术, 深度挖掘信息本身的语义维度, 引入语义层次的信息, 关注传输信息内容而非编码符号, 更加满足未来通信需求。通过充分利用语义层面信息的高度抽象、智能简约等特性, 语义通信将变革经典香农信息论框架, 有望突破经典通信系统的传输瓶颈, 形成智能化的新型通信体系^[2]。

当前, 语义通信尚处于研究初期, 构建统一的、具有泛化价值的性能评估体系对语义通信的发展至关重要。本文围绕评估体系在语义通信系统设计实现中的作用展开探讨, 并在分析现有通信系统评估指标的基础上提出具有泛化性的语

基金项目: 国家自然科学基金(62293485); 中央高校基本科研业务费专项资金(2022RC18)

义通信评估体系,为语义通信的发展奠定基础。

1 语义通信性能评估体系需求

当前,语义通信尚处于研究初期,语义通信理论在代表性的人-机、机-机等智能交互场景中快速发展,关键性成果不断涌现,系统性能增益更加显著。然而,由于缺乏统一且具有泛化价值的性能评估体系,语义通信研究成果的可理解性有所欠缺,成果间的横向对比难以实现。

一般而言,准确性与时效性是评价通信系统性能的两个重要指标。现有语义通信相关研究所采用的评估体系主要由传统通信系统评估体系演化而来,或由下游语义任务评估体系迁移而来。对于由传统通信系统评估体系演化而来的情况,相应的准确性评估指标常采用误比特率,时效性评估指标常采用比特传输速率。这类评估指标的问题在于无法有效反映通信收发端语义信息传递的能力。对于由下游语义任务评估体系迁移而来的情况,由于系统模型多采用通信模块与下游语义任务模块组合确定结构、端到端训练固定参数的方式,这类评估指标具有与通信任务场景、信息模态高度关联的特征。如面向语义重建的语义通信系统普遍采用原始信息与重建信息的误差函数作为准确性度量,采用压缩比的函数作为时效性度量;而面向文本翻译、图像分割等语义任务的语义通信系统则普遍将模型输出与监督标签的差异作为准确性度量。

上述各评估指标的定义不同,取值范围存在显著差异,导致诸多研究的系统模型之间无法横向贯通。即使针对相同评估指标,由于信道、干扰、信源语义特征等差异性存在,不同场景下的语义传输方案也无法进行有效对比。统一且具泛化价值的语义通信评估体系,从整体视角对通信系统性能进行直观的评估,不仅能验证系统的合理性,还可以为系统的优化改进提供参考,进而为系统演进提供方向。语义通信评估体系相关研究亟待开展,以便为语义通信的稳步发展夯实基础。

2 现有语义通信性能评估指标

语义通信性能评估指标与具体通信任务场景、信息模态高度关联。其中,信息重建任务广泛出现在以人类作为最终接收用户的应用场景中,如音视频通话、多媒体文件云备份等,该类任务要求在收端将信息恢复至发端信息模态,并尽量保证信息的准确,性能评估指标基于发送信息与重建信息的差异进行定义;非信息重建任务多出现于人-机、机-机交互场景中,如智能驾驶、工业物联网等,该类任务要求所发信息的关键内容能够被理解和应用,性能评估指标基于所

输出信息与相应监督标签之间的差异定义。对于信息模态而言,人类针对文本、图像/视频、语音等模态信息采用不同的重点捕捉方式,各模态信息的失真无法简单通过欧氏距离等方式统一度量。各模态信息所衍生的后续任务存在很大差异,文本情感分析、智能问答、图像分割、语音识别等任务均需要定义各自的性能评估指标。

基于重建任务与其他代表性人工智能任务(例如文本分类、图像目标检测、语音识别)的划分,本节对文本、图像/视频、语音3种常见模态信息传输模型的性能评估指标进行分析总结,为后续构建统一、具有泛化性的语义通信性能评估体系奠定基础。

2.1 面向重建任务的评估指标

2.1.1 文本信息重建

文本信息重建的现有常用衡量指标主要为双语替换评测分数^[3](BLEU),该指标最初被用于文本翻译的质量评估。BLEU基于加权 n 元模型(n -gram)精确度进行定义,其具体形式为:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right), \quad (1)$$

其中, $\text{BP} = \begin{cases} 1 & \text{if } l_{\hat{x}} > l_x \\ e^{1 - \frac{l_x}{l_{\hat{x}}}} & \text{if } l_{\hat{x}} \leq l_x \end{cases}$ 为针对重建文本语句长度过

短的惩罚项, $l_{\hat{x}}$ 、 l_x 分别为重建文本与原始文本语句的长度, w_n 为归一化权重, $N=4$ 为经验取值。

$$p_n = \frac{\sum_{n-\text{gram} \in \hat{x}} \text{Count}_{\text{clip}}(n-\text{gram})}{\sum_{n-\text{gram}' \in \hat{x}} \text{Count}(n-\text{gram}')}, \quad (2)$$

其中, p_n 表示 n -gram精确度, $\text{Count}_{\text{clip}}(n-\text{gram})$ 、 $\text{Count}(n-\text{gram}')$ 分别表示 n 长词组在原始文本 X 与重建文本 $\hat{X}$ 中的出现次数。BLEU的物理意义是衡量重建文本的所有 n 长词组中同时出现在原始文本中的数量占比。尽管作为最为常用的文本信息重建衡量指标,BLEU仍无法实现对于词汇语义的理解^[4],存在未考虑到词汇语料的召回率^[5]等不足。

2.1.2 图像/视频信息重建

现有图像/视频信息重建的常用衡量指标包括峰值信噪比(PSNR)与多尺度结构相似性^[6](MS-SSIM)。PSNR定义为均方误差(MSE)的对数函数。对于动态范围 N 的图像/视频帧,PSNR计算如下:

$$\text{PSNR} = 10 \lg \frac{N^2}{\text{MSE}(X, \hat{X})}, \quad (3)$$

其中, $\text{MSE}(X, \hat{X}) = \sum_{i=1}^H \sum_{j=1}^W (X_{ij} - \hat{X}_{ij})^2 / (HW)$ 表示尺寸为 $H \times W$ 的原始与重建图像/视频帧的均方误差。PSNR 是对图像/视频帧像素级重建误差的精确反映, 因定义简单且易于计算, 在早期工作^[7]中被广泛应用。但由于未考虑人眼生理特性等因素, PSNR 的评估准确性不如 MS - SSIM。

MS - SSIM 将人眼生理特性纳入考虑: 人眼在衡量两幅图的相似性时, 对于两幅图的局部结构差异更为敏感。故一方面, MS - SSIM 的运算基于一定尺寸的图像块; 另一方面, MS - SSIM 通过分别定义亮度、对比度、结构对比函数, 将 3 类信息拆分, 赋予它们不同的权重并独立进行对比。相较于 PSNR, MS - SSIM 的评估结果更符合人眼感知。MS - SSIM 的计算方式为:

$$\text{MS - SSIM}(X, \hat{X}) = \left[L_M(X, \hat{X}) \right]^{\alpha_M} \prod_{l=1}^M \left[C_J(X_M, \hat{X}_M) \right]^{\beta_J} \left[S_J(X, \hat{X}_M) \right]^{\gamma_J}, \quad (4)$$

其中, 降采样次数 M 、亮度权重 α_M 、对比度权重 β_J 、结构对比权重 γ_J 为超参数, 其常用取值为: $M = 5, \beta_1 = \gamma_1 = 0.0448, \beta_2 = \gamma_2 = 0.2856, \beta_3 = \gamma_3 = 0.3001, \beta_4 = \gamma_4 = 0.2363, \alpha_5 = \beta_5 = \gamma_5 = 0.1333$ 。 X_M 、 $\hat{X}_M$ 分别为原图及重构图像/视频帧的 M 级降采样结果。

$$L_M(X_M, \hat{X}_M) = \frac{2\mu_{X_M}\mu_{\hat{X}_M} + C_1}{\mu_{X_M}^2 + \mu_{\hat{X}_M}^2 + C_1}, \quad (5)$$

$$C_J(X_M, \hat{X}_M) = \frac{2\sigma_{X_M}\sigma_{\hat{X}_M} + C_2}{\sigma_{X_M}^2 + \sigma_{\hat{X}_M}^2 + C_2}, \quad (6)$$

$$S_J(X_M, \hat{X}_M) = \frac{\sigma_{X_M\hat{X}_M} + C_3}{\sigma_{X_M}\sigma_{\hat{X}_M} + C_3}, \quad (7)$$

其中, $L_M(X_M, \hat{X}_M)$ 、 $C_J(X_M, \hat{X}_M)$ 、 $S_J(X_M, \hat{X}_M)$ 分别为指标的亮度对比因子、对比度因子、结构对比因子, μ_{X_M} 、 $\mu_{\hat{X}_M}$ 、 σ_{X_M} 、 $\sigma_{\hat{X}_M}$ 分别为 X_M 、 $\hat{X}_M$ 的均值与标准差, $\sigma_{X_M\hat{X}_M}$ 为两者的协方差。

$C_1 = (k_1 N)^2$ 、 $C_2 = (k_2 N)^2$ 、 $C_3 = C_2/2$ 为固定参数, 用于防止除法运算问题的产生。 N 为图像/视频帧的动态范围, $k_1 = 0.01$ 、 $k_2 = 0.03$ 为经验取值。

近年的相关工作^[8-9]多采用 PSNR 与 MS - SSIM 指标并行的方式。此外, 文献^[8]还采用了对数形式的 MS - SSIM 指

标, 以应对 MS - SSIM 取值范围过小的问题。

2.1.3 语音信息重建

语音信息重建的现有衡量指标包括语音质量感知评估^[10] (PESQ) 与短时客观可懂度^[11] (STOI) 等。早期语音质量衡量主要通过主观打分方式获得平均意见值 (MOS)。PESQ 是用于模拟 MOS 打分的音频质量评价算法, 其定义为:

$$\text{PESQ} = 4.5 - 0.1d_{\text{SYM}} - 0.0309d_{\text{ASYM}}, \quad (8)$$

其中, d_{SYM} 与 d_{ASYM} 分别为算法求得的语音信息对称与非对称干扰, 反映重建模型的预测精度和概括能力。PESQ 是对重建语音信号可理解程度的衡量, 侧重人的主观感受, 在早期的无线通信系统中广泛使用。

同样被广泛使用的指标还包括 STOI。STOI 是对重建语音信息失真程度的客观度量, 其取值与基于人工智能的语音识别等下游任务性能相关。具体而言, 首先定义 $\hat{x}(k, m)$ 为第 m 帧语音信号的第 k 个离散傅里叶变换 (DFT) 波瓣, 计算第 m 帧语音信号的第 j 个 1/3 倍频程范数 $X_j(m)$ 。

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k, m)|^2}, \quad (9)$$

其中, $k_1(j)$ 、 $k_2(j)$ 为定义的取值边界。同理, 可定义重建语音信号的相应范数 $Y_j(m)$, 计算 $Y_j(m)$ 的归一化形式 $Y_j(m)'$:

$$Y_j(m)' = \max \left(\min \left(\alpha Y_j(m), X_j(m) + 10^{-\frac{\beta}{20}} X_j(m) \right), X_j(m) - 10^{-\frac{\beta}{20}} X_j(m) \right), \quad (10)$$

其中, $\beta \in \{-\infty, -30, -20, -15, -10\}$ 与 $N \in \{20, 30, 40, 50, 60\}$

为可选参数, $\alpha = \left(\frac{\sum_n X_j(n)^2}{\sum_n Y_j(n)^2} \right)^{1/2}$ 为功率归一化参数,

$n \in \mathcal{M} = \{(m - N + 1), \dots, m\}$ 。

之后, 定义第 m 帧语音信号的第 j 个 1/3 倍频程损失 $d_j(m)$:

$$d_j(m) = \frac{\sum_n (X_j(n) - \frac{1}{N} \sum_l X_j(l)) ((Y_j'(n) - \frac{1}{N} \sum_l Y_j'(l)))}{\sqrt{\sum_n \left(X_j(n) - \frac{1}{N} \sum_l X_j(l) \right)^2 \sum_n \left(Y_j'(n) - \frac{1}{N} \sum_l Y_j'(l) \right)^2}}, l \in \mathcal{M}. \quad (11)$$

$STOI = \frac{1}{JM} \sum_{j,m} d_j(m)$ 为 $d_j(m)$ 的均值, 其中 M 、 J 分别表示语音信号的总帧数与 1/3 倍频程数。现有工作^[12]多同时采用 PESQ 与 STOI, 以获得对于所提模型更为全面的性能认知。

2.2 面向其他人工智能任务的评估指标

2.2.1 文本分类

文本分类是自然语言处理领域的典型任务, 其目的在于通过人工智能为文本信息按照一定的标准进行标记, 可细分为情感分析、主题分类、问答任务、意图识别等多种应用场景。以单标签二分类的情感分析任务为例, 现有评价指标主要为分类准确率与 F 分数^[13] (F - score), 具体计算逻辑如下:

按照各样本的分类结果与真实标签, 可将各样本的分类结果分为真阳性 (TP)、真阴性 (TN)、假阳性 (FP)、假阴性 (FN), 如表 1 混淆矩阵所示。在此基础上, 准确率 $ACCURACY = \frac{TP + TN}{TP + FP + TN + FN}$ 定义为经模型正确分类的样本占总样本的比例, 是对任务性能最简单直观的度量。在实际应用场景中, FN 与 FP 的代价可能是不同的, 准确率指标无法对其进行细化评估, 而 F 分数弥补了这方面的不足。

$$F - score = \frac{(1 + \beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$
 是基于所有样本的

分类精确率 $precision = \frac{TP}{TP + FP}$ 与召回率 $recall = \frac{TP}{TP + FN}$ 的调和平均数定义的度量, 超参数 β 赋予了指标为精确率与召回率针对性加权的能力。在现今的相关研究中, $F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall}$ ($\beta = 1$) 是常用的一种形式。对于多分类任务, marco - F1、weighted - F1 等变体指标被广泛应用^[14]。当数据集的分布不均衡时, F 分数仍存在主要类别的评估值偏高^[15]等弊端。相应的改进方法有基于接收者操作特征曲线 (ROC) 的度量等。

2.2.2 图像目标检测

图像目标检测的目的在于通过人工智能判断出图像中物

▼表1 二分类任务混淆矩阵

	分类结果为真	分类结果为假
真实标签为真	TP	FN
真实标签为假	FP	TN
FN: 假阴性	FP: 假阳性	TN: 真阴性
	TP: 真阳性	

体的类别与相应位置。图像目标检测任务的现有评价指标主要为平均精确度均值 (mAP), 其定义为各类目标查准率 (P) - 查全率 (R) 曲线下方面积的均值。

具体而言, 首先, 根据目标检测结果 B_p 的置信度, 以及 B_p 与真实结果 B_{gt} 的重合程度 (可用两者的交并比 IoU 度量, $IoU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}$), 可将检测结果分为如表 2 混淆矩阵所示的 TP、TN、FP、FN; 其次, 由检测顺序中 TP、FP 数量的累加情况计算查准率与查全率的累加值:

$$P = \frac{TP}{TP + FP}, \quad (12)$$

$$R = \frac{TP}{TP + FN}; \quad (13)$$

然后, 对于各类目标样本, 以查全率为横坐标, 以查准率为纵坐标, 绘制 $P-R$ 曲线并计算精确度均值 (AP); 最后, 由

$$AP \text{ 加权平均可得 mAP, 即 } mAP = \frac{\sum_1^M N_M AP_M}{\sum_1^M N_M}$$

为目标类别数量与该类别样本数量。

目前, 基于上述定义的 mAP 指标在现有工作^[16]中已被广泛使用。值得注意的是, 在如图像检索等实例级图像任务中, 存在另一种基于整图匹配结果定义的 mAP 指标 (已应用在文献[17]的端到端车辆重识别系统中)。两者的运算逻辑一致, 仅在混淆矩阵定义上存在差异。

2.2.3 语音识别

语音识别的目的在于通过人工智能实现语音信息到文本信息的模态转换过程。语音识别任务的现有评价指标主要为准确率 $ACCURACY$ 。准确率可基于字错误率 (WER) 定义, 即:

$$ACCURACY = 1 - WER, \quad (14)$$

其中, $WER = \frac{N_{sub} + N_{ins} + N_{del}}{N}$, N_{sub} 、 N_{ins} 、 N_{del} 分别为识别文本到达参考文本所需的单词替换、插入与删除操作的次数, N 为参考文本的总单词数。ACCURACY 已在文献[18]等工作中获得应用。

▼表2 目标检测任务混淆矩阵

	B_p 置信度 大于阈值	B_p 置信度 小于阈值
B_{gt} 与 B_p 的 IoU 大于阈值	TP	FN
任意 B_{gt} 与 B_p 的 IoU 均小于阈值	FP	TN
FN: 假阴性	IoU: 交并比	TP: 真阳性
FP: 假阳性	TN: 真阴性	

整体而言, 语义通信评价指标存在以下问题:

1) 现有评价指标普遍从具体的重建任务/人工智能任务直接迁移而来, 侧重于反映任务完成的准确性维度, 忽视了时效性维度。语义通信系统追求的是在有限带宽资源消耗与时延下的任务完成准确性, 而非不计成本地逼近准确性上限。

2) 不同任务场景/模态信息下的评估指标不互通, 性能评估结果不易理解。在语义通信过程中, 受通信意图变化的影响, 系统的评估指标会不断发生变化, 造成通信质量评估混乱。

3) 各评价指标的取值范围、量纲不一致, 给性能的横向对比带来困难。如上文中所提到的 PESQ、PSNR 指标与文本分类和语音识别准确率指标存在明显差异。这种差异主要表现在取值范围与物理意义方面。

综上所述, 对于语义通信系统的评价指标, 一方面, 应在选取应用任务导向的准确性指标的同时, 引入时效性指标以进行制约; 另一方面, 应筛选物理意义相似的指标以组成指标集合, 并设计指标的标准化逻辑统一取值范围, 使在不同任务场景下构建的语义通信系统之间能够进行横向性能对比。对于包含不同通信意图的语义通信系统, 它们的具体实现需要与场景、应用紧密结合。从整体视角进行直接的性能评估, 有助于系统之间横向贯通, 促进系统的迭代演进。

3 语义通信性能评估新角度

为解决现有语义通信性能评估指标存在的问题, 我们提出语义通信效率指标 E_{sc} 与语义通信效用指标 U_{sc} , 并以语义图像重建与语音重建任务为例, 搭建端到端语义通信仿真系统, 基于所提指标对仿真系统性能进行评估。

3.1 新评估指标定义

首先, 我们定义标准化准确性指标 $\overline{ACC} \in [0, 1]$ 与标准化时效性指标 $\overline{TIM} \in [1, +\infty)$:

$$\overline{ACC}(\gamma) = \min \left\{ \frac{ACC(\gamma) - ACC_{\min}(\gamma)}{ACC_{th}(\gamma)}, 1 \right\}, \quad (15)$$

$$\overline{TIM}(\gamma) = \max \left\{ \frac{TIM(\gamma)}{TIM_{th}(\gamma)}, 1 \right\}, \quad (16)$$

其中, γ 表示给定的通信资源条件, 包含信噪比 (SNR) 与带宽等参数; $ACC \in \{g(\text{BLEU}, \text{MS-SSIM}, \text{TOP-n ACCURACY}, \text{mAP}, \dots)\}$ 、 $TIM \in \{f(\text{计算时延}, \text{传输时延},$

端到端时延, $\dots)\}$ 分别为语义任务相关的准确性与时效性评价指标, $g(\cdot)$ 与 $f(\cdot)$ 分别为相应的变换函数; $ACC_{\min}$ 与 ACC_{th} 分别为特定语义任务的准确性下界与优化边界, TIM_{th} 为特定语义任务时效性优化边界。这里, 超越边界的准确性或时效性指标的取值不存在或不具备进一步优化的需求。

标准化操作旨在规范不同评估指标的取值尺度, 剔除语义任务难易度差异对语义通信系统性能评估产生的影响。进一步地, 基于上述标准化指标, 我们定义语义通信效率指标 E_{sc} :

$$E_{sc}(\gamma) = \frac{\overline{ACC}(\gamma)}{\overline{TIM}(\gamma)}. \quad (17)$$

语义通信效率指标 $E_{sc} \in [0, 1]$ 旨在衡量通信系统在给定通信资源下、在单位时间开销内的任务完成准确性。当通信系统的准确性指标增大或时效性指标减小时, E_{sc} 会增大, 即 E_{sc} 越接近 1, 系统的语义通信效率越高。

语义通信效用指标 U_{sc} 为:

$$U_{sc}(\gamma) = \frac{\lambda}{1 + \lambda} \cdot \overline{ACC}(\gamma) + \frac{1}{1 + \lambda} \cdot \frac{1}{\overline{TIM}(\gamma)}, \quad (18)$$

其中, λ 表示加权因子, 用于权衡准确性需求和时效性需求的比重。语义通信效用指标 $U_{sc} \in [0, 1]$ 旨在衡量通信系统在给定通信资源下, 对于任务性能上限的接近程度。当系统侧重于任务完成准确性或时效性时, U_{sc} 的前项或后项会增大, 但另一项会相应减小。 U_{sc} 越接近 1, 系统在准确性与时效性两方面的综合效用表现越高。

3.2 新性能指标分析与评估

本文首先以图像重建任务为例, 选取车联网领域的传感器扩展应用作为具体验证场景, 基于上述评估指标进行仿真分析。传感器扩展应用要求车辆、行人、交通设施等通信节点之间能够实现传感器所采集的图像/视频信息的实时交互, 对数据传输的准确性、时效性均有较高要求。

针对上述应用场景, 准确性指标 ACC 转换为对数形式 MS-SSIM, 时效性指标 TIM 为传输时延, 变换函数 $g(\cdot)$ 与 $f(\cdot)$ 均选取恒等函数, 即:

$$ACC(\gamma) = \text{MS-SSIM}(\gamma), \quad (19)$$

$$TIM(\gamma) = \text{传输时延}(\gamma). \quad (20)$$

根据第3代合作伙伴计划 (3GPP) 制定的《5G NR Rel-16 V2X 车联网标准》, 准确性边界、时效性边界与通信参数

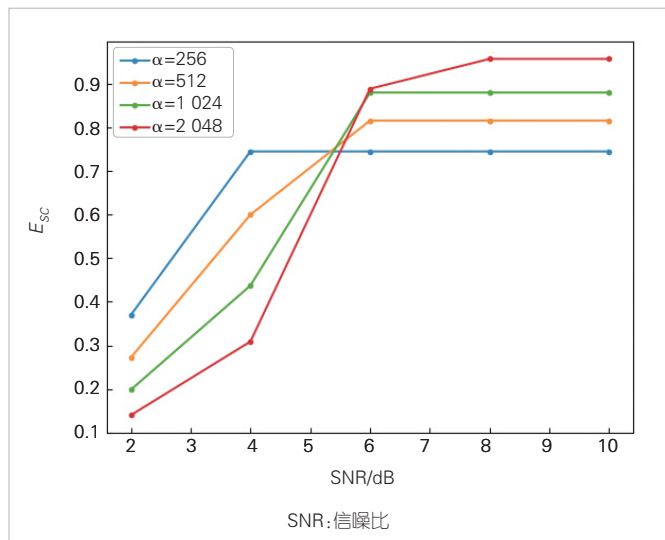
如表3所示。

进一步地，我们搭建包含图像压缩重建、信道编码与调制模块在内的端到端通信仿真模型，基于选定指标进行验证。图像压缩重建模块基于文献[19]的网络结构实现，并分别采用 ImageNet 数据集、Cityscapes 数据集进行训练和验证。图像压缩重建模块在训练过程中采用了4组不同的率失真参数 $\alpha = \{256, 512, 1\,024, 2\,048\}$ 。采用更高的 α 参数训练的模型会更倾向于降低图像压缩比以换取更高的重建质量。模型采用低密度奇偶校验码 (LDPC) 进行信道编码，并根据信噪比条件将 LDPC 码率控制在 1/5 ~ 8/9 之间。调制方式采用二进制相移键控 (BPSK)。

图1给出了语义通信效率 E_{sc} 与通信系统信噪比的关系曲线。随着信噪比的上升，包含不同参数的语义任务模型均表现出更高的语义通信效率。此外，采用高 α 参数训练的模型对信噪比的变化更为敏感，在低信噪比条件下表现出明显的性能劣化，同时在高信噪比条件下具有更高的性能上限。这是由于此类模型在低信噪比条件下产生了过多的编码开销以保护信息不被噪声破坏，从而无法满足任务的时效性需

▼表3 通信模型仿真参数

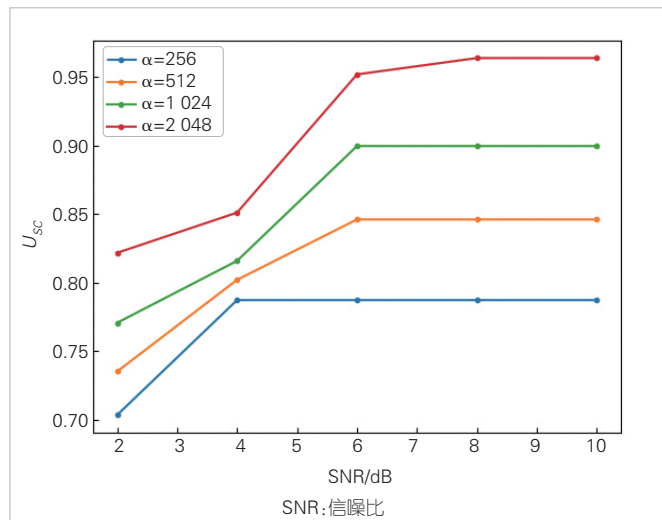
	图像重建任务	语音重建任务
准确性下界 ACC_{min}	0	1
准确性优化边界 ACC_{th}	0.99	4.5
时效性优化边界 TIM_{th}	100 ms	与所传输语音信息时长相等
通信参数 γ	信噪比 2 ~ 10 dB, 带宽 3 MHz	信噪比 2 ~ 10 dB, 带宽 10 kHz
信道编码与调制方式	LDPC+BPSK	LDPC+BPSK
LDPC: 低密度奇偶校验码 BPSK: 二进制相移键控		



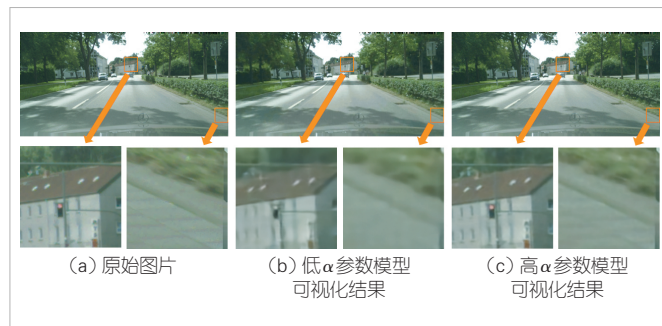
▲图1 图像重建任务语义通信效率 E_{sc} 与信噪比的关系曲线

求；在高信噪比条件下，通信所需的编码冗余显著减少。此类模型能够在时效性不超出阈值的前提下，实现更高的任务准确性。

图2给出了 $\lambda = 5$ 条件下语义通信效用 U_{sc} 与通信系统信噪比的关系曲线。考虑到实际语义任务场景中存在对准确性或时效性更为侧重的情况，参数 λ 使效用指标更为真实地反映语义任务需求。本部分以准确性需求导向的任务为例。随着信噪比的上升，4组模型的语义通信效用均表现出与直觉吻合的上升趋势。即使在低信噪比条件下，高 α 参数模型表现出较高的准确性与较低的时效性，其语义通信效率劣于低 α 参数模型。但由于任务对于准确性需求的侧重，高 α 参数模型具有更高的语义通信效用。图3与表4分别展示了典型图像经语义图像重构模型仿真传输后的可视化结果与具体性能参数。高 α 参数模型表现为更出色的图像重建质量（红绿灯的色彩更准确，建筑物、地砖的细节更清晰）与更高的图像传输时延开销，在较好的信道条件下，其传输时延未超出阈值，故高 α 参数模型具有更高的语义通信效率与语义通信效用。



▲图2 图像重建任务语义通信效用 U_{sc} 与信噪比的关系曲线 ($\lambda = 5$)



▲图3 通信模型可视化结果

▼表4 通信模型性能参数

性能参数	低 α 参数模型	高 α 参数模型
信噪比 SNR	10 dB	10 dB
率失真参数 α	256	2 048
MS-SSIM	0.961 2	0.986 3
传输时延	20.77 ms	73.20 ms
语义通信效率 E_{sc}	0.705 6	0.931 6
语义通信效用 U_{sc}	0.754 7	0.943 0

MS-SSIM:多尺度结构相似性 SNR:信噪比

接着我们以语音信息重建任务为例,构建应用自适应多速率宽带(AMR-WB)语音编码的高清语音通信模型,并基于上述评估指标进行仿真分析。准确性指标 ACC 转换为宽带语音质量感知评估(WB-PESQ),时效性指标 TIM 为传输时延,变换函数 $g(\cdot)$ 与 $f(\cdot)$ 均选取恒等函数,即:

$$ACC(\gamma) = WB - PESQ(\gamma), \quad (21)$$

$$TIM(\gamma) = \text{传输时延}(\gamma). \quad (22)$$

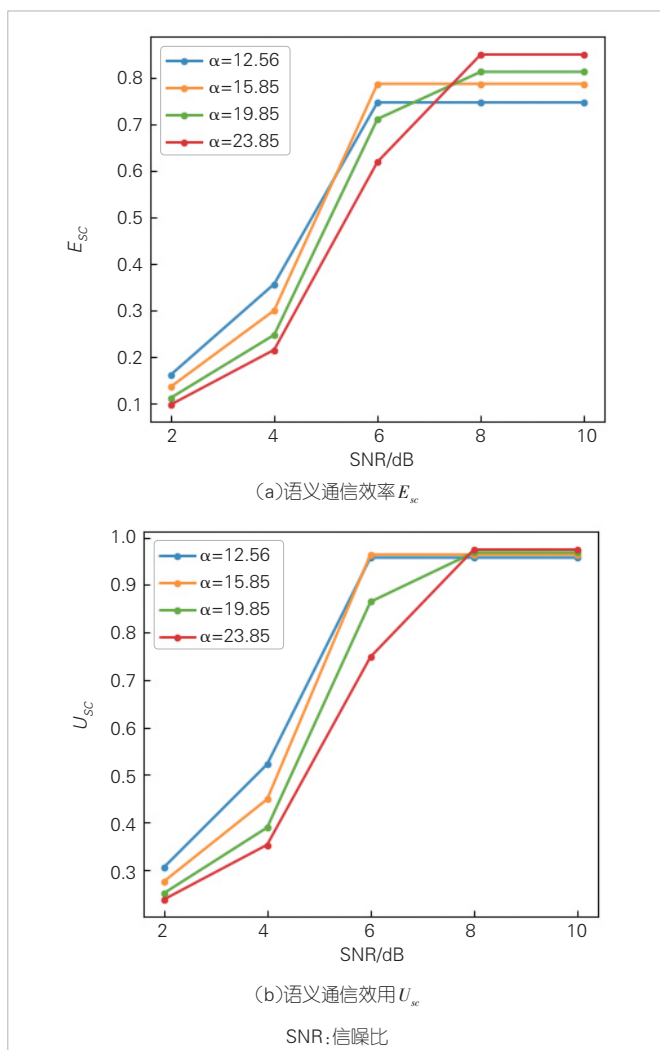
同时,准确性边界、时效性边界与通信参数如表3所示。

本文采用日本电报电话公司(NTT)宽带语音数据集^[20]的美式英语子集进行验证,同时为AMR-WB编码设定4组不同的速率参数 $\alpha = \{12.65, 15.85, 19.85, 23.85 \text{ (kbit/s)}\}$ 。采用更高码率的模型会更倾向于降低压缩比以获得更高的重建质量。模型同样采用LDPC信道编码与BPSK调制方式,并根据信噪比条件将LDPC的码率控制在1/5~8/9之间。

图4给出了 $\lambda = 1/5$ 条件下语义通信效率 E_{sc} 与效用 U_{sc} 随通信系统信噪比的变化趋势曲线。与面向图像重建任务的语义通信系统相似,随着信噪比的上升,效率与效用指标均表现出上升趋势。同时,在语义效用 U_{sc} 方面,低速率编码方案具有更显著的优势,能够反映出系统对时效性需求的侧重。在上述面向图像与语音重建任务的通信模型仿真中,除带宽之外的通信参数都是一致的。语义通信效率指标与效用指标的评估结果表现出统一性,为两者的横向性能对比提供指导。带宽参数的设定与信息模态相关,它所产生的影响可通过对变换函数进一步设计来消除。

4 结束语

通过融合人工智能与通信技术,充分利用语义层面信息的高度抽象、智能简约等特性,语义通信有望形成突破经典通信系统传输瓶颈的智能新型通信体系。由于语义通信尚处于研究初期,统一且具泛化价值的性能评估体系的缺失,阻



▲图4 语音重建任务语义通信效率 E_{sc} 与效用 U_{sc} 随信噪比的关系曲线($\lambda = 1/5$)

碍了关键性成果间的横向贯通与对比。本文通过分析语义通信评估痛点,提出语义通信效率指标 E_{sc} 与语义通信效用指标 U_{sc} 。该指标具有更好的泛用性,且为不同任务场景、模态信息下的语义通信模型的横向性能对比提供指导,进而促进语义通信的快速发展。

参考文献

- [1] 张平, 张建华, 戚琦, 等. Ubiquitous-X: 构建未来6G网络[J]. 中国科学: 信息科学, 2020, 50(6): 913-930
- [2] ZHANG P, XU W J, GAO H, et al. Toward wisdom-evolutionary and primitive-concise 6G: a new paradigm of semantic communication networks [J]. Engineering, 2022, 8(1): 60-73. DOI: 10.1016/j.eng.2021.11.003
- [3] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. New York: ACM, 2002: 311-318. DOI: 10.3115/1073083.1073135
- [4] XIE H Q, QIN Z J, LI G Y, et al. Deep learning enabled semantic communication systems [J]. IEEE transactions on signal processing, 2021, 69: 2663-2675. DOI: 10.1109/TSP.2021.3071210

- [5] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005: 65–72.
- [6] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment [C]//Proceedings of Thirty-Seventh Asilomar Conference on Signals, Systems & Computers. IEEE, 2004: 1398–1402. DOI: 10.1109/ACSSC.2003.1292216
- [7] JIANG J. Image compression with neural networks – A survey [J]. Signal processing: image communication, 1999, 14(9): 737–760. DOI: 10.1016/S0923-5965(98)00041-1
- [8] MINNEN D, BALLÉ J, TODERICI G. Joint autoregressive and hierarchical priors for learned image compression [C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. ACM, 2018: 10794 – 10803. DOI: 10.5555/3327546.3327736
- [9] LU G, OUYANG W L, XU D, et al. DVC: an end-to-end deep video compression framework [C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 10998–11007. DOI: 10.1109/CVPR.2019.01126
- [10] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs [C]//Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221). IEEE, 2002: 749–752. DOI: 10.1109/ICASSP.2001.941023
- [11] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech [C]//Proceedings of 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010: 4214–4217. DOI: 10.1109/ICASSP.2010.5495701
- [12] FU S W, WANG T W, TSAO Y, et al. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks [J]. IEEE/ACM transactions on audio, speech, and language processing, 2018, 26(9): 1570–1584. DOI: 10.1109/TASLP.2018.2821903
- [13] CHINCHOR N. MUC-4 evaluation metrics[C]//Conference on Message Understanding. Association for Computational Linguistics, 1992.
- [14] MA Y K, PENG H Y, CAMBRIA E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM [C]//Proceedings of the AAAI conference on artificial intelligence. ACM, 2018, 32(1): 5876 – 5883
- [15] BEKKAR M, DJEMAA H, ALITOCHE T A. Evaluation measures for models assessment over imbalanced data sets [J]. Journal of information engineering and applications, 2013, 3: 27–38
- [16] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137–1149. DOI: 10.1109/TPAMI.2016.2577031
- [17] XU W J, ZHANG Y M, WANG F Y, et al. Semantic communication for Internet of vehicles: a multi-user cooperative approach [EB/OL]. [2023–02–16]. <https://arxiv.org/abs/2212.03037>
- [18] CHIU C, RAFFEL C. Monotonic chunkwise attention [EB/OL]. [2023–02–16]. <https://arxiv.org/abs/1712.05382>
- [19] BALLE J, MINNEN D, SINGH S, et al. Variational image compression with a scale hyperprior [EB/OL]. [2023–02–16]. <https://arxiv.org/abs/2212.03037>
- [20] NTT Advanced Technology Corporation (NTT-AT). Multi-lingual speech database for telephony [DB]. 1994

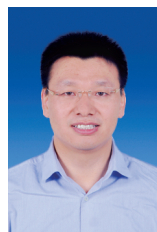
作者简介



郑远，北京邮电大学在读博士研究生；主要研究领域为语义通信。



王凤玉，北京邮电大学讲师；主要研究方向为语义通信、无线人工智能、通感一体化等。



许文俊，北京邮电大学教授、鹏城实验室兼职研究员；主要研究领域为语义通信、5G/6G 智能通信网络；曾获国家技术发明奖二等奖、教育部技术发明奖一等奖、IEEE ICC 2019 最佳论文奖等奖项；发表论文 130 余篇。