面向智能任务的语义通信：理论、技术和挑战

刘传宏[1]，郭彩丽[1,2]，杨洋[2]，陈九九[1]，朱美逸[1]，孙鲁楠[1]

（1. 北京邮电大学先进信息网络北京实验室，北京 100876；
2. 北京邮电大学网络体系构建与融合北京市重点实验室，北京 100876）

摘 要：未来机−机、人−机万物智能互联对传统通信方式提出了挑战，提取信源语义信息进行传输的语义通信方法为 6G 提供了新的解决方法。首先，综述了语义通信的发展历程和研究现状，分析了语义通信目前面临的两大瓶颈问题，提出了面向智能任务的语义通信架构，给出了面向智能任务的语义信息熵和语义信道容量的度量方法；其次，针对不同的智能任务，分别提出了语义编码和语义联合信源信道编码方案；再次，搭建了语义通信平台，对所提方法进行实验验证；最后，对语义通信未来的挑战和开放性问题进行了总结。语义通信方法相较于传统通信方法可以大大降低传输数据量和传输时延，将在未来万物智联的通信中发挥重要作用。

关键词：6G；语义熵；语义通信；语义编码；智能任务

中图分类号：TN929.5

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2022117

Intelligent task-oriented semantic communications: theory, technology and challenges

LIU Chuanhong[1], GUO Caili[1,2], YANG Yang[2], CHEN Jiujiu[1], ZHU Meiyi[1], SUN Lu'nan[1]

1. Beijing Laboratory of Advanced Information Networks, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. Beijing Key Laboratory of Network System Construction and Integration, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: In the future, intelligent interconnection of all things, such as machine-to-machine and human-to-machine, poses challenges to traditional communication methods. The semantic communication method that extracts semantic information from source information and transmits them provides a novel solution for the 6G communication system. First, the development process and research status of semantic communications were reviewed, the two bottleneck problems faced by semantic communications were analyzed, and an intelligent task-oriented semantic communication architecture was proposed. The measurement methods of task-oriented semantic entropy and semantic channel capacity were given. For different intelligent tasks, semantic coding and semantic source-channel joint coding schemes were proposed respectively. Besides, a semantic communication platform was built to verify the proposed method. Finally, the future challenges and open issues of semantic communications were summarized. Compared with traditional communication methods, semantic communication can significantly reduce the amount of transmitted data and the transmission delay, and it will play an important role in the future communication of the Internet of everything.

Keywords: 6G, semantic entropy, semantic communication, semantic coding, intelligent task

收稿日期：2022−02−18；**修回日期**：2022−05−09

通信作者：郭彩丽，guocaili@bupt.edu.cn

基金项目：中央高校基本科研业务费专项资金资助项目（No.2021XD-A01-1）；北京市自然科学基金资助项目（No.4202049）；北京邮电大学博士生创新基金资助项目（No.CX2022101）

Foundation Items: The Fundamental Research Funds for the Central Universities (No.2021XD-A01-1), The Beijing Natural Science Foundation (No.4202049), BUPT Excellent Ph.D. Students Foundation (No.CX2022101)

0 引言

目前，以 5G/6G 通信和人工智能（AI, artificial intelligence）为代表的新一轮信息技术革命和产业变革席卷全球，通信与 AI 已成为国家战略的两大重要组成部分。随着通信与 AI 紧密融合的智能社会到来，传统的人−人通信将延伸到机−机、人−机、人−人多种方式智能互联，通信的信源和信宿将变成具有智能分析和处理能力的人、机等智能体[1]。不同类型的智能体之间如何实现更高效的通信？经典香农信息论的基本假设和理论结果是否可以支撑海量智能体互联的需求？这些都是在以传统通信为主导的香农时代没有被考虑的场景和问题，亟须挖掘通信基础理论的新突破和新指引。

当面临智能体之间的通信时，需要重新审视通信的目的、方式和过程。通信真正的目的是通过通信双方交互使接收方理解发送方的信息内容，即“达意”通信[2]。香农早已定义信息所表示的内容为“语义”。Weaver[3]进一步对通信的认识做出重要补充，将通信问题归为 3 个层面。

1) 语法（技术）层面。这一层面是经典香农信息论涉及的范畴，解决通信符号如何准确地加以传输。

2) 语义层面。这一层面解决传输的符号如何精确地传达内容含义（也就是本文所说的语义信息）。

3) 语用层面。这一层面解决如何对接收的语义信息以最佳的方式加以利用，即通信的目的。

尽管消息蕴含的语义信息和语用价值早已被察觉并定义，但受当时技术发展水平和通信场景需求的限制，人们在根据香农信息论进行通信系统的相关研究时，主要专注于语法（技术）层面的问题，仅以可靠有效传输比特数据为目标[4]。时至今日，有关通信可靠性及有效性的问题已基本得到解决。随着人工智能技术与通信技术的融合日益紧密，未来的通信趋于万物智联，过去暂时搁置的语义层面的问题重新凸显，以“达意”为目标的语义通信成为下一个研究热点[2]。本文旨在综述已有语义通信相关工作并为语义信息的度量和编码提出一些可行的思路。本文的主要贡献如下。

1) 综述现有语义通信的研究现状，分析现有语义通信在语义度量和语义编码方面存在的瓶颈问题，提出面向智能任务的语义通信网络架构，将语义信息和语用价值进行深度融合。

2) 针对非统计型语义信息难以度量的问题，借鉴模糊数学理论和方法对面向智能任务的语义信息量进行定性和定量分析，在智能任务的约束下，将模糊集、隶属函数和模糊度等赋予实际物理意义，使语义熵的计算可以通过经典信息论扩展得到。

3) 针对如何高效压缩语义信息的问题，基于信息瓶颈理论和扩展的信息瓶颈理论，分别提出面向智能任务的语义编码和信源信道联合编码方案，显著提升语义编码压缩的有效性和语义信息传输的可靠性。

4) 搭建语义通信平台，对所提方法进行验证和测试；实验结果证明所提方法的可行性和优越性，相较于传统通信方法，所提方法可以有效提升智能任务性能和带宽利用率。

5) 分析并总结语义通信未来的主要挑战和开放性问题，旨在为语义通信后续的探索提供思路和方向。

1 语义通信的研究现状及瓶颈问题

1.1 语义通信的研究现状

语义通信的研究致力于解决 Weaver[3]定义的通信系统语义和语用层面的问题，现有研究主要分为两类，一是为解决语义层面的问题——如何精确地传达内容含义，考虑从通信系统传输需求出发，研究语义信息的度量；二是为解决语用层面的问题——如何最佳利用语义信息，考虑从智能系统应用需求出发，研究语义信息理论的延伸扩展和实际应用。

在语义信息度量方面，主要面临以下问题。①与语法信息相比，语义信息涉及内容含义，存在对错偏差等问题，所以其概率是非统计型的，导致语义信息难以表示；②语义信息的不确定性不仅来自事件发生概率，还来自语义概念及其外延的模糊性，导致具有随机和模糊双不确定性的语义信息难以度量。后续语义信息度量相关研究重点在于解决这两大问题，主要分为以下两类。

1) 针对语义信息的非统计特性和模糊性等问题，从信息哲学的角度对语义信息进行表示。Carnap 等[5]提出用逻辑概率代替统计概率来描述语义信息的对错偏差和非统计特性等问题。Popper[6]提出用逻辑概率和信息准则检验语义信息，并对逻辑概率进行数学描述。另一方面，Zadeh[7-8]提出模糊集合论和模糊事件来描述语义信息的模糊性等问题。

Luca 等[9]提出用来测度模糊事件信息量的模糊信息熵公式。

2) 针对具有双不确定性的语义信息难度量问题，引入广义信息论对语义信息进行度量。Floridi[10]将语义信息进行分类并提出基于逻辑真值的语义信息度量公式，但该信息公式与传统的香农信息公式相距太远，无法在同一个通信系统中统一表示和计算。Lu[11-12]引入广义信息论，考虑在香农信息论的基础上对语义信息统一度量，并基于贝叶斯公式、逻辑概率和模糊集合等理论对语义信息进行数学度量。上述研究都是在香农经典信息论的基础上进一步扩展，侧重研究语义信息的表示和度量，未考虑信息语义理解的智能化需求，很难在实际场景中进行应用。

在语义信息理论延伸扩展和实际应用方面，主要面临以下问题。①语义信息面向不同的智能化需求，蕴含不同的效用价值，忽略语用价值的语义信息的度量缺乏物理含义，难以对实际应用场景进行指导；②传统的通信系统和网络架构难以满足智能化需求，在实际应用中智能任务的实现难以达到预期效果；③如何高效提取不同模态信源中的主观语义信息存在挑战。因此，近年来，大量研究关注智能化需求，对语义通信理论进行扩展并考虑其实际应用，研究工作主要分为以下三类。

1) 针对语义信息的效用价值等问题，引入全信息论等对传统信息论进行延伸和扩展。钟义信[13]提出在人工智能领域更应该关注信息的内容含义和效用价值，并提出了全信息的概念；从语法信息、语义信息和语用信息 3 个方面对信息进行统一描述，并在后续研究工作中阐述了 3 种信息形式的相互关系并推导了全信息的初步测度公式[14]；钟义信等[15]利用信息生态理论解释了语义信息的生成机理，然后对其进行定义，并给出了人工智能与信息科学的交叉研究模型。石光明等[16]从智能感知的角度给出了新的语义通信方式，并讨论了语义编译码机制。上述研究对语义信息的定义和表示形式进行了有益的探讨，指出了信息论与人工智能结合的信息科学理论的研究方向。

2) 针对传统通信网络难以满足智能化需求的问题，利用人工智能技术赋能，设计面向未来智慧场景的语义通信系统和网络架构[17-26]。在语义网络架构设计方面，Bao 等[17]首先提出了在网络中加入语义处理和知识共享机制，并给出了基本的三层通信模型，强调了本地知识和共享知识对收发双方语义通信的辅助作用。Strinati 等[2]提出在 6G 的语义网络中加入语义学习机制。Popovski 等[18]提出在 6G 协议栈中引入语义平面，以实现语义滤波和面向特定目标的语义控制。在语义通信系统设计方面，Kountouris 等[19]探讨了面向智能系统互联的语义使能通信场景，并给出了语义使能通信的基本模型。Yang 等[20]将语义通信和智能任务结合，提出了面向智能任务的语义通信架构。Kalfa 等[21]提出了语义信号处理框架，可以适用于接收方不同的通信任务。牛凯等[22]和 Zhang 等[23]探讨了语义信息的度量，并提出了智能高效的语义通信系统架构。石光明等[24]将语义通信作为未来万物智联网络的新型基础范式，提出了与万物智联网络融合的语义通信的基本模型和组成。

3) 针对主观的语义信息难以提取的问题，基于深度学习技术，面向不同类型的信源提出了一系列工程可实现的语义通信方法[27-35]。针对文本和语音信息传输，Farsad 等[27]基于长短期记忆（LSTM, long short-term memory）网络提出了联合信源信道编码（JSCC, joint source channel coding）方法，并证明了基于深度学习的 JSCC 的优越性。Xie 等[4]基于 Transformer 提出了一种用于文本信息传输的语义通信系统 DeepSC（deep learning based semantic communication），首次在句子层级上对语义信息进行了区分。在文献[4]的基础上，Xie 等[28]进一步提出了一个轻量化的分布式语义通信系统，使其更容易部署在物联网设备上。Weng 等[29]将 DeepSC 扩展到语音信号传输中，设计了一个基于注意力机制的语义编解码器 DeepSC-S。针对图像信源传输，Bourtsoulatze 等[30]基于卷积神经网络（CNN, convolutional neural network）提出了 JSCC 来实现无线信道中的图像传输，同时优化语义编解码器来提升图像传输的性能。后续工作将图像信源和视觉任务进行结合，Lee 等[31]设计了一个联合图像传输和分类的语义通信系统，接收端直接输出分类任务的结果。

针对图像检索任务，Jankowski 等[32]提出了边端协同的语义通信方法，大大提升了图像检索任务的性能。刘传宏等[33]考虑在智能物联网场景中，提出了智能任务导向的语义通信系统，并基于语义概念和语义特征之间的重要性关系实现对语义信息的进一步压缩。针对多模态信源数据传输，Xie 等[34]

考虑视觉问答任务，一位用户传输文本问题，另一位用户传输待询问的图像，提出了针对多模态数据传输的多用户语义通信系统。基于文献[34]的研究，Xie 等[35]进一步考虑多用户及多任务的语义通信方法，提出了基于 Transformer 的语义通信框架，并在机器翻译、图像检索和视觉问答智能任务上验证了所提方法的优越性。上述研究主要包括基于深度学习的语义编码方法和语义通信的框架及思路，为工程应用提供了可供参考的系统和网络模型，但是暂未将智能任务和语义信息熵的度量相结合，同时暂无基于深度学习的语义编码和语义通信平台的落地实现，这些都需要进一步研究和探讨，对推动语义通信未来的进一步发展和完善具有重要意义。

1.2 语义通信面临的瓶颈问题

综上，自 1950 年以来，人们期待对语义信息建立相应的理论，包括语义信息的度量、语义信息的压缩及语义信息在语义信道中的有效传输。现有国内外相关研究还处于语义信息度量公式的探讨或是基于深度学习的语义通信工程技术层面的尝试，语义通信的进一步发展面临以下瓶颈问题亟须突破。

1) 如何量化语义信息。语义信息的度量问题是语义信息理论的基础。经典的香农信息论以概率论为基础来表征和度量信息，其信息概率是固定可统计的，而语义信息经过传输和智能计算后不仅存在统计判决对错问题，还存在语义理解偏差带来的模糊表征及度量问题，属于非统计型信息，其概率是不确定的[5]。

2) 如何压缩语义信息。对信源进行语义层面的信息提取与编码表示，有助于进一步压缩语义信息的冗余，提高语义传输的有效性。语义编码充分利用信源的语义冗余，提取最重要的语义特征，属于有损压缩。一般用深度神经网络（DNN, deep neural network）提取的语义信息具有不可解释性，如何设计最优语义编码方法，探寻有损压缩的极限问题难度很大。有损压缩极限问题的另一个关键是对语义失真的度量，但语义的失真度量存在很大主观性。

2 面向智能任务的语义通信

2.1 语义与语用信息的结合

值得一提的是，语义信息往往隐含了语用的目的，而语用也往往包含了语义信息，正如 Weaver[3] 在论文中所提，语义信息与语用信息是难以分割的。北京邮电大学的钟义信教授[14,36]也指出，语义信息的内涵是语法信息和语用信息。在未来人工智能时代，智能体之间通信的最终目的往往为完成智能任务，即通信终端不仅接收语义信息，还需要理解发送方的语义信息并加以利用。因此智能体间通信的本质是通信与智能任务的融合。

以完成目标检测智能任务为目标的语义通信示例如图 1 所示。从图 1 可以看出，当智能任务的目标（即语用信息）分别为检测狗、检测猫和同时检测猫和狗时，对应的通信内容（即语义信息）随之发生变化。面向特定的智能任务，语义的模糊性以及不可解释性限制语义度量和压缩难题有望得到突破。

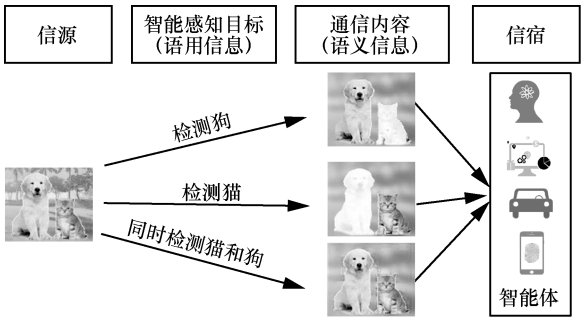


图 1 面向智能任务的语义通信示例

2.2 面向智能任务的语义通信架构

图 2 对比了面向智能任务的传统通信系统和语

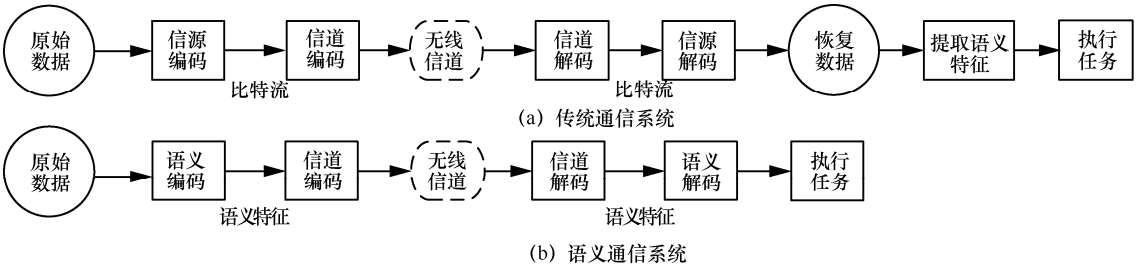


图 2 传统通信系统和语义通信系统对比

义通信系统。传统通信系统中，信源在编码和传输的过程被转换成比特流，接收机追求准确恢复代表信源的比特流以恢复原始数据，接着进行语义提取并执行任务。语义通信系统传输信源的语义信息和传统通信系统最主要的区别之一是引入了语义编码，语义编码依据要执行的任务对信源进行语义特征提取，仅传输语义特征，大大降低了对通信资源的需求。接收机的任务包括信源重建和其他智能任务，如图像分类和目标检测等。语义编码的引入实现了信源传输和信源理解过程的融合，即“先理解后传输”；而传统通信方法中，信源传输和理解相对独立，为“先传输后理解”的通信模式。

本文基于 DNN 设计语义特征提取网络，主要有以下 2 个原因：1) 传统机器学习通过人工设定规则提取特征，而 DNN（如 CNN 等）通过卷积池化等多层网络提取语义特征，虽然 DNN 缺乏可解释性，但是大量的实验结果证明 DNN 提取的特征比人工设定规则提取的特征在完成智能任务方面表现更佳；2) 语义信息本身具有主观和不确定特性，特别是针对图像等信源，难以设计合适的手工提取规则，无法直接用传统机器学习的方法提取信源数据中的语义特征，而通过可视化证明了 CNN 等方法与人类神经系统极为相似，能够提取局部语义特征，因此基于 DNN 的语义特征提取方法更加适合语义通信的场景。此外，实际应用中提取语义信息的编解码网络结构设计需要综合考量信源模态和具体的智能任务等，如针对文本信源，循环神经网络（RNN, recurrent neural network）能提取上下文相关语义信息；针对图像信源，CNN 表现更加优越。然而，要设计最优的面向智能任务的语义通信系统，语义信息的度量是基础。

3　面向智能任务的语义信息度量

3.1　语义信息熵的度量

信源产生的消息序列不仅服从一定的概率分布，还蕴含丰富的语义信息。现有语义信息熵度量相关研究一方面基于模糊数学理论，用隶属函数刻画信源的语义模糊测度[37-38]；另一方面基于命题逻辑理论，用逻辑概率描述消息语义为真的概率[17,39]。上述研究均借鉴传统香农熵的形式，虽分别构建了模糊熵和语义熵来度量信源的语义信息量，但未考虑面向不同智能任务时，因关注的语用价值不同，同一消息蕴含的语义信息量不同。因此，度量所得的语义信息量数值缺乏物理意义，难以对面向智能任务的语义通信技术进行有效指导。

因此，本文针对特定智能任务，度量基于本地知识库理解的消息所蕴含的语义信息量。定义构成语义消息的最小基本单元为语义元（例如用知识图谱表示语义消息时，语义元为三元组），从语义元层面考虑本地知识库的辅助作用以及智能任务对语义理解的影响。具体表现为本地知识库提供构成消息的所有可能的语义元，协助将消息提炼为多个语义元，每个语义元对不同类别的智能任务进行决策的贡献程度不同，同时在特定智能任务条件下，每个语义元对每个正确决策结果的隶属程度具有模糊性。因此，基于模糊数学理论刻画语义理解的模糊程度，实现语义信息熵的度量。

首先，将本地知识库 K 表示为语义元的集合，记作 $K=\{c_1,c_2,\cdots,c_n,\cdots,c_N\}$，其中，$c_n$ 表示本地知识库 K 中的语义元，语义元 c_n 表示一条知识。其次，用重要度 $\omega_{\Gamma}(c_n)$ 衡量语义元 c_n 协助智能任务 Γ 进行决策的贡献程度，且满足 $\sum_{n=1}^{N}\omega_{\Gamma}(c_n)=1$。同时，智能任务 Γ 具有多个推理结果，记作 $\Gamma=\{T_1,T_2,\cdots,T_m,\cdots,T_M\}$，例如在监督学习中，$T_m$ 表示标签。每个语义元是否属于某种正确决策结果具有模糊性，因此，T_m 定义了一个模糊集合，采用 Zadeh 记法，写作 $T_m=\frac{\mu_{T_m}(c_1)}{c_1}+\frac{\mu_{T_m}(c_2)}{c_2}+\cdots+\frac{\mu_{T_m}(c_n)}{c_n}+\cdots+\frac{\mu_{T_m}(c_N)}{c_N}$，其中，$\mu_{T_m}(c_n)$ 表示 c_n 对 T_m 的隶属度，且满足 $0\leqslant\mu_{T_m}(c_n)\leqslant 1$。语义元 c_n 对正确决策结果 T_m 的模糊熵 $T_m(c_n)\in[0,a](a\geqslant 0)$ 反映了 c_n 属于 T_m 的模糊程度，模糊熵 $T_m(c_n)$ 与隶属度 $\mu_{T_m}(c_n)$ 存在以下函数关系

$$T_m(c_n)=-\Big[\mu_{T_m}(c_n)\log\mu_{T_m}(c_n)+\big(1-\mu_{T_m}(c_n)\big)\log\big(1-\mu_{T_m}(c_n)\big)\Big] \tag{1}$$

则语义元 c_n 对智能任务 Γ 的模糊熵为 c_n 对每个正确结果 T_m 的模糊熵之和，即

$$\Gamma(c_n)=\sum_{m=1}^{M}T_m(c_n) \tag{2}$$

语义信源产生的每条语义消息可被理解为若干语义元的集合 X，且 $X\subseteq K$。针对每一个 $c_n\in K$，

若 $c_n \in X$，则 $\rho_X(c_n)=1$，否则 $\rho_X(c_n)=0$。则在给定本地知识库 K 时，面向智能任务 Γ 的语义消息 X 的语义熵为

$$H_s(X|\ \Gamma)=\sum_{n=1}^{N}\rho_X(c_n)\omega_\Gamma(c_n)I(c_n) \tag{3}$$

该语义熵度量了在给定本地知识库 K 时，面向智能任务 Γ 的语义消息 X 蕴含的语义信息量。

3.2 语义信道容量的度量

传统通信系统信道容量仅与信道转移概率有关，可通过优化编码器寻求最佳码字分布，使收发端码字间的互信息达到上确界，则信息传输速率逼近信道容量。与传统信道容量追求最大化传输速率不同，语义信道容量追求最大化语义信息传输能力。针对语义通信系统，文献[17]提出降低语义编码的模糊度，提高物理信道的传输速率和语义解码器的理解能力，来优化语义信息传输速率达到语义信道容量。但在实际语义通信场景中，通信双方的知识共享程度越高，语义信道传输相同码字承载的语义信息就越多；同时，只有面向特定智能任务时语义编解码器理解的语义信息才具有语用价值。考虑图 3 所示的语义通信过程，设通信双方共享知识库为 K，面向智能任务 Γ 的语义通信过程描述为发送端将语义信源产生的消息 X（标签为 Y，同一信源 X 在不同智能任务 Γ 下有不同的标签 Y）经语义编码生成语义表示 Z，经语义信道传输后，接收端将接收到的 $\hat{Z}$ 经语义解码得到智能任务处理结果 $\hat{Y}$。

在上述通信过程中，语义信道容量与智能任务的类别、收发双方共享知识库的协同程度以及语义信道条件有关。

面向智能任务 Γ 时，当共享知识库 K 的协同程度和语义信道转移概率 $p(\hat{z}\,|\,z)$ 一定时，语义通信系统的语义信道容量为定值，反映了该系统的语义信息传输能力上限。语义信道容量 C_s 的计算式为

$$C_s=\sup_{p(z),p(\hat{y})}\left\{H_s(Z|\ \Gamma,K)+H_s(\hat{Y}|\ \Gamma,K)+I(Z;\hat{Z})\right\} \tag{4}$$

其中，$H_s(Z|\Gamma,K)$ 为在智能任务 Γ 和共享知识库 K 条件下语义表示 Z 的语义熵，即 Z 蕴含的语义信息量，反映了语义编码器的语义信息提取能力；同理，$H_s(\hat{Y}|\ \Gamma,K)$ 反映了语义解码器的语义信息理解能力；$I(Z;\hat{Z})$ 为 Z 与 $\hat{Z}$ 之间的互信息，表示经过语义信道后，接收的语义表示 $\hat{Z}$ 所保留的关于 Z 的语义信息，反映了语义编码器的抗语义信道干扰能力。此时，可通过优化语义编解码器，寻求最佳语义表示 Z 的分布 $p(z)$，以及最佳智能任务处理结果 $\hat{Y}$ 的分布 $p(\hat{y})$，使上述三项之和达到上确界，则语义信息传输速率达到语义信道容量。

4 面向智能任务的语义编码

由语义信道容量度量计算式可以看出，如何设计最优的语义编码方案对语义通信系统至关重要。语义编码的本质是在一定的语义失真前提下，尽可能多地压缩信源信息，语义编码与传统信源编码的区别如表 1 所示。

面向智能任务的语义编码示意如图 4 所示，以广义的率失真理论为语义编码的指导理论，可以表示为

$$\min I(X;O)+\lambda D(X;O) \tag{5}$$

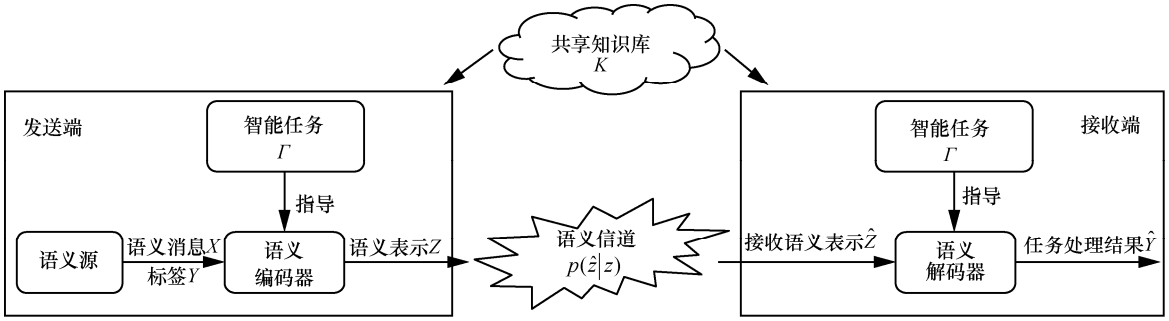


图 3 语义通信过程示意

表 1 语义编码与传统信源编码的区别

	指导理论	编码方式	算力需求	评价指标	编码目标
语义编码	广义率失真理论[40]	基于神经网络	GPU/CPU	智能任务性能	提升码元的语义信息量
传统信源编码	率失真理论	非神经网络	CPU	编码效率	提高码元的平均信息量

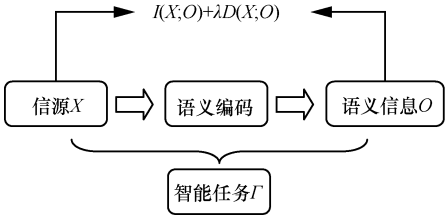


图 4　面向智能任务的语义编码示意

其中，$I(X;O)$ 表示 O 保留的关于 X 的信息量，衡量编码前后语义信息量的压缩程度；$D(X;O)$ 表示 O 与 X 之间的信息差别，衡量编码前后语义信息的失真程度；λ 表示权重参数。针对是否需要重建信源，O 可以有不同表示，当需要重建信源时，O 为由码字重建的信源 $\hat{X}$；当不需要重建信源时，O 为语义编码后的码字 Z。本节以智能任务为导向，分别从语义失真度量和语义编码方案两方面综述现有研究及可行的方向。

4.1　语义失真度量

目前，语义失真评价的目标是研究人在观察图像或视频时的视觉感受，而非机器的理解[41-43]。近年来，基于深度学习的失真评估方法开始大范围应用[44-46]。这些方法一般分两步评估样本的失真程度，首先设计适当的特征提取网络提取语义特征，然后使用这些特征进行回归或分类来评估失真程度。然而在智能互联时代，大量的失真图像和视频需要被输入机器中执行各种智能任务，语义失真度量应该趋近于机器理解并以智能任务为导向，使机器可以更好地理解图片和视频中的语义信息，从而更好地完成智能任务。

基于此，为了更好地度量机器对于语义理解的失真，可以基于孪生语义编码器实现语义失真度量。孪生语义编码器 Φ 对一组样本同时提取语义特征，这里的孪生语义编码器网络模型的选择和信源模态等有关（如文本信源可以选择 RNN，图像信源可以选择 CNN 等），接着语义失真度量模型度量语义重建的数据 X' 和原数据 X 的语义域距离 $d\left(\Phi(X),\Phi(X')\right)$；同时，将失真前后的信源数据输入智能任务处理网络，得到智能任务的性能损失 Δd，并以 Δd 作为语义失真度量模型的参考和标签，指导语义失真度量模型拟合智能任务的性能损失，从而实现以智能任务为导向、趋近于机器理解的语义失真度量。语义失真度量模型如图 5 所示。

此外，考虑到互信息能够衡量一个随机变量包含的关于另一个随机变量的信息量，因此还可以借鉴互信息度量语义失真程度。$I(X;Y)-I(X;Z)$ 度量将信源 X 压缩为 Z 所丢失的关于重要语义信息 Y 的信息量，即压缩过程中的语义失真程度。其中，Y 与智能任务相关，例如在分类任务中 Y 为类别标签。针对具体的智能任务，还可以在收发双方分别对信源产生的消息和信宿接收的消息提取语义信息，将收发双方语义信息熵的差值 $H(X)-H(X')$ 作为语义失真的度量。3 种语义失真度量方法的总结与对比如表 2 所示。

表 2　3 种语义失真度量方法的总结与对比

语义失真度量方法	数学表达形式	优缺点
基于孪生神经网络	$d\left(\Phi(X),\Phi(X')\right)$	便于计算，但算力需求较大
基于语义互信息	$I(X;Y)-I(X;Z)$	难以计算，可基于神经网络估计
基于语义信息熵	$H(X)-H(X')$	度量结果准确，但难以计算

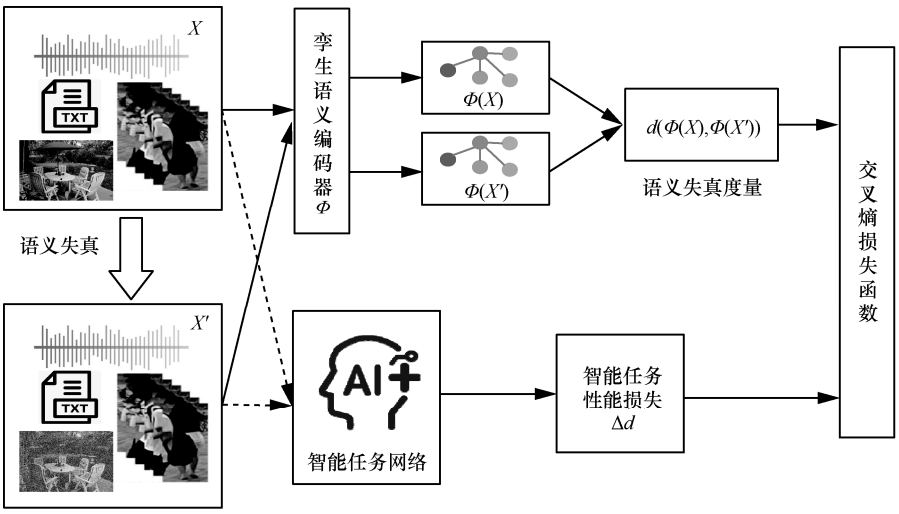


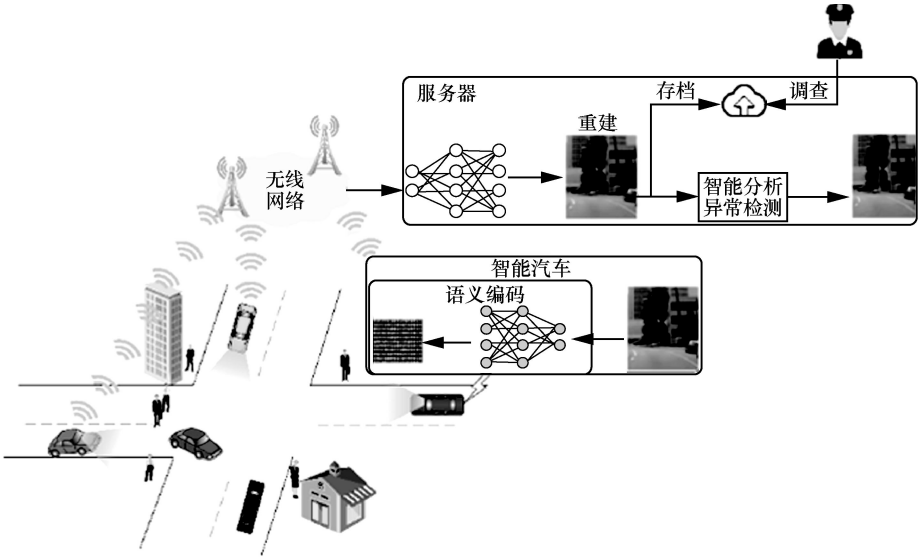
图 5　语义失真度量模型

4.2　面向智能任务的语义编码方案

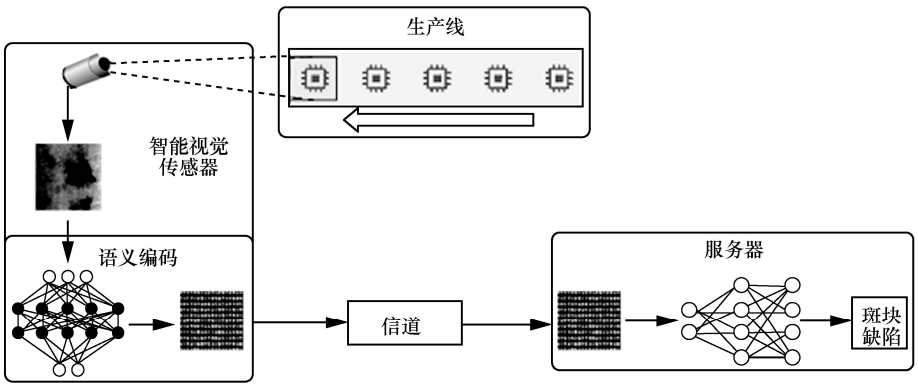
基于广义率失真理论，不仅需要考虑语义失真的度量，还需要进一步研究面向智能任务的语义编码方案，其可以根据完成智能任务的接收端是否需要重建信源分成需要重建信源和不需要重建信源的语义编码方案，如图 6 所示。

4.2.1　需重建信源的语义编码方案

在需要同时满足人和机器理解的场景（如虚拟现实）中，语义编码传输方案应具备在接收端恢复信源的能力，从而不仅可以直接面向智能任务，还可以用于人类的视觉理解与分析，适用于人–物交互的通信场景中。以图 6(a)中移动监控场景为例，不仅需要实现行人和车辆检测，还需要保存移动监控记录视频，以备后续人工查阅。目前，基于深度学习的信源编码方案发展迅速，在图像压缩编码领域，现有大多数工作都仅以图像像素级的差异作为编码失真度量，仅以恢复图像为目标，未引入对下游智能任务的考量[47]。文献[48]提出了一种可辨别的图像压缩方法，旨在保持下游 AI 任务的特征级一致性。然而，下游 AI 任务的性能最终更多地取决于语义级信息。像素级、特征级和语义级信息的关系示意如图 7 所示。与现有工作思路类似，需要重建信源的语义编码主要侧重于给定压缩比时，最小化语义失真 $D(X;\hat{X})$。一方面，可以将智能任务作为先验信息，在信源重建中考虑对后续智能任务的影响，在收发双方分别提取信源、信宿针对特定任务的语义信息，将损失函数设计为信源在收发双方的语义失真，基于此完成语义编解码网络的训练。由于考虑到了语用层，该编码思路不仅可以提升接收端的智能任务性能，还可以保证信源恢复的



(a) 需要重建信源的语义编码方案（如移动监控场景）



(b) 不需要重建信源的语义编码方案（如工业互联网场景中残次品检测任务）

图 6　面向智能任务的语义编码方案

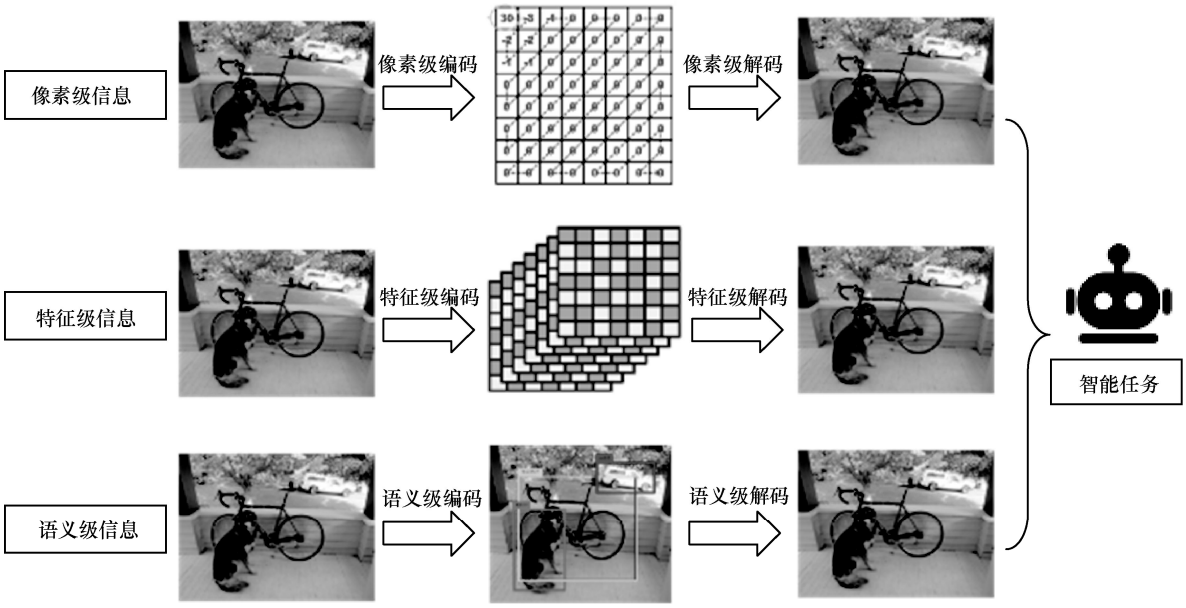


图 7　像素级、特征级和语义级信息的关系示意

质量与传统编码方法相当。另一方面，还可以考虑将像素级的失真均方误差（MSE, mean squared error）和智能任务的性能联合考虑作为语义的失真度量。具体来说，在考虑图像重构质量的基础上，同时考虑了重构图片在后续智能任务上的性能，系统的优化目标加权考虑了重构质量和智能任务性能，即 $\mathrm{MSE}(X,\hat{X})+\alpha\mathrm{CrossEntropy}$，其中 α 为权重系数，直接将智能任务的影响引入编码压缩过程中，从而可以在编码过程中有效保留智能任务需要的语义信息；CrossEntropy 为分类任务的损失函数。

4.2.2　*不需要重建信源的语义编码方案*

与需要重建信源的语义编码方案固定压缩比不同，不需要重建信源的语义编码方案不追求对信源的无失真重建，主要适用于物–物互联的场景，因此有更高的压缩空间，其旨在找到语义失真 $D(X;Z)$ 和语义压缩 $I(X;Z)$ 的最优折中。以图 6(b) 中工业互联网场景中残次品检测任务为例，工业相机捕捉产品图像仅需要完成残次品检测智能任务，而不需要人工进行图像识别，因此不需要重建信源。此时，语义编码应尽可能保留信源中关于智能任务全部的语义信息，语义编码的目的与数理统计中充分统计量的思路一致，充分统计量是对信源数据的总结，包含了完成智能任务所需的语义信息，极小充分统计量是最小化的充分统计量[49]。最优的语义编码应包含所有语用信息，即信源中与智能任务相关的信息，同时应是对信源中与智能任务无关信息的最大程度压缩，所以可以从极小充分统计量的角度出发研究针对智能任务最优的语义编码方案。

1) 基于近似极小充分统计量的语义编码

由于在信源分布未知的情况下直接求解信源的极小充分统计量非常困难，而且只有非常特殊的分布才有精确的极小充分统计量[50]，因此可以通过信息瓶颈[52]的方法来求解近似的极小充分统计量，如图 8 所示。具体来说，就是放宽优化条件，要求尽可能压缩信源互信息 $I(X;Z)$ 同时尽可能多地保留与智能任务相关的语义信息 $I(Z;Y)$，而非全部信息。通过信息瓶颈方法得到的有效表示就是近似的极小充分统计量，是实现高效语义编码的一种可行方案。

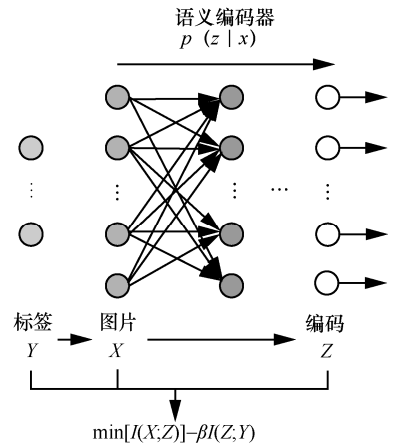


图 8　基于信息瓶颈求解极小充分统计量示意

2) 基于可解释性的语义编码

上述语义编码均基于神经网络实现，考虑到语义信息的抽象属性和神经网络的黑盒属性，为了更好地利用神经网络提取语义信息，应进一步研究神经网络提取语义信息的可解释性，从而实现高效的语义编码方案。

针对具体的智能任务，可以从智能任务对应的语义概念（指智能任务中客观表示的某一具体事物，如猫狗分类任务中的猫和狗）出发，如图 9 所示，首先利用神经网络对信源进行编码特征提取，然后利用语义概念 c 的得分对第 k 个特征图 $\boldsymbol{A}^k$ 求梯度，最后经过全局平均池化即可得到针对语义概念 c 的第 k 个特征图的重要性权重 ω_k^c，可表示为

$$\omega_k^c = \frac{1}{wh}\sum_i\sum_j \frac{\partial y^c}{\partial \boldsymbol{A}_{ij}^k} \tag{6}$$

其中，得分 y^c 定义为最后一层全连接经过 Softmax 之前语义概念对应的神经元激活值；$\boldsymbol{A}^k$ 的宽度和高度分别为 w 和 h；$\boldsymbol{A}_{ij}^k$ 为特征图第 i 行 j 列的激活值。基于此，将语义概念和特征进行关联，提取可解释的语义关系，建立特征对概念的重要性排序，并基于重要性排序对特征进行裁剪压缩，从可解释角度实现对语义信息的进一步压缩，从而在保证后续智能任务性能的前提下，大大提高语义编码的效率。

5　面向智能任务的语义联合信源信道编码

为了实现最优的语义通信系统，不仅需要设计语义编码，还需要考虑语义信道传输对语义通信的影响，以提升语义通信系统的稳健性。近年来，基于深度学习的 JSCC 在通信性能上展现了强大的优势，其可能成为未来通信系统的主要编解码方式。然而现有的联合信源信道编码主要考虑信源传输及重建的场景，缺乏与智能任务的结合，因此仍然存在大量的冗余，可以进一步进行压缩并提升系统的稳健性。本文分别从是否需要重建信源的 JSCC 两方面综述现有研究及可行方向。

5.1　需要重建信源的 JSCC

在需要重建信源的 JSCC 中，原始图片经过编码器提取到的特征不仅需要在接收端尽可能无失真地恢复出图片，还需要能够抵抗信道噪声的干扰。然而，目前针对图片重建的 JSCC 主要考虑在固定的网络结构下，尽可能多地保留特征中关于原始图片的信息。这导致提取的语义信息中仍存在一定的语义冗余，因此应在现有基础上研究如何在传输更少的信息的同时更好地抵抗信道噪声干扰。需要重建信源的语义 JSCC 方案如图 10 所示，基于扩展的信息瓶颈理论，在最大化接收特征与图像间的互信息 $I(X;\hat{Z})$ 的同时最小化发送特征和图像的互信息 $I(X;Z)$，以 $I(X;Z)-\lambda I(X;\hat{Z})$ 为网络训练的损失函数，其中 λ 为超参数。在尽可能压缩图像的语义信息量的同时，保留足够的语义信息恢复图片，同时考虑信道噪声的影响，提升语义信息传输的稳健性和有效性。在实际应用过程中，超参数 λ 控制压缩和重建质量之间的平衡，如果 λ 过大，则 Z 有可能丢失过多关于 X 的信息；如果 λ 过小，则 Z 中会包含更多的冗余信息，增加传输所需比特数。针对损失函数中超参数 λ 调参烦琐且费时费力的问题，基于比例积分微分控制（PID, proportional-integral- derivative control）理论引入了超参数自适应算法[51]，更好地平衡了语义压缩和重建质量。

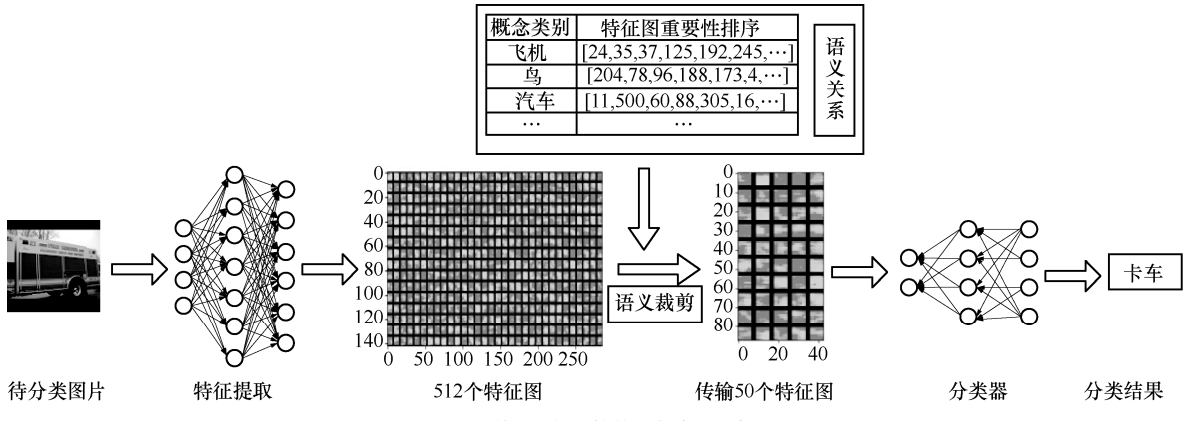


图 9　基于可解释性的语义编码方案

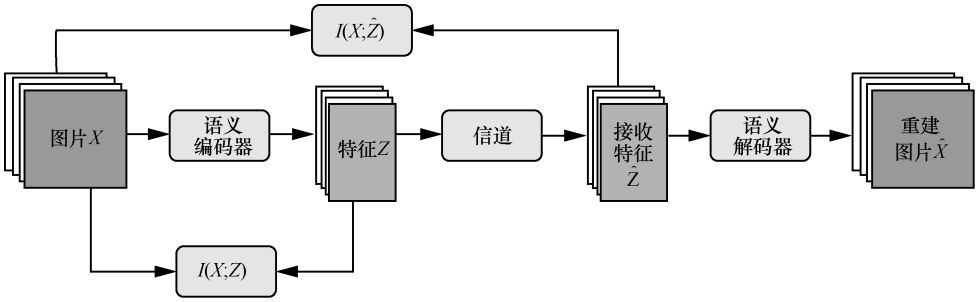


图 10　需要重建信源的语义 JSCC 方案

5.2　不需要重建信源的 JSCC

在不需要重建信源的 JSCC 中，语义编码器提取信源的语义信息后经信道传输，接收端不需要重建信源而是直接完成下游智能任务。目前相关的工作旨在完成智能任务的同时保证香农信息熵最小[52]，难以保证在考虑信道传输的影响后接收端仍为最优的语义表示。面向智能任务的 JSCC 方案如图 11 所示，基于扩展的信息瓶颈理论，通过最小化 $I(X;Z)$ 和最大化 $I(\hat{Z};Y)$ 寻求语义压缩与经信道传输后语义信息失真的最优权衡，指导语义信源信道编码器既能实现面向智能任务的最优压缩，又能抵抗物理信道干扰。同时，考虑最大化收发端传输语义信息间的互信息 $I(Z;\hat{Z})$，从而提高语义通信系统的信息传输速率。当面向图像分类任务时，最大化 $I(\hat{Z};Y)$ 可转化为最小化交叉熵，指导分类器进行智能任务处理。

6　面向智能任务的语义通信性能测试

目前，语义通信相关的实验结果均处于仿真阶段，缺乏硬件平台的支撑，为了测试语义通信方法在实际应用中的可行性和优势，本文设计并搭建了语义通信平台，对语义通信的性能进行了测试，以图像分类这一智能任务为例，面向智能任务的语义通信平台架构如图 12 所示。语用层明确通信所面向的智能任务；语义层主要完成语义信息的编译码；技术层主要实现信道编译码、调制等传统通信中的其他功能。考虑到语义通信平台和传统通信平台最大的区别在于语义编解码部分，为了更好地提取语义，语义通信平台中语义编解码器均需基于 DNN 实现。首先基于 PyTorch 深度学习框架实现了语义编解码器，为了将提取的语义信息转化为可传输的波形信号，需要设计语义层和技术层的接口，本文搭建的平台利用通用软件无线电外设（USRP, universal software radio peripheral）从计算机的用户数据报协议（UDP, user datagram protocol）端口读取语义信息，然后基于 LabVIEW 软件控制 USRP 对语义信息进行信道编码和 64QAM 调制，最后经天线发出；接收端为一系列逆过程，即利用 USRP 恢复接收到的语义信息，并通过技术层和语义层的接口将语义信息输入语义解码器，输出智能任务结果。

实验数据集采用 STL-10 图像分类数据集，包含 10 类图像，图像分辨率为 96 像素×96 像素，每类 500 张用于训练，800 张用于测试。语义层 DNN 初始化参数设置如表 3 所示。通信系统设计套件为 LabVIEW Communications System Design Suite 2.0。技术层基于 LTE 协议实现，信道编码采用 Turbo 码。技术层参数设置如表 4 所示。

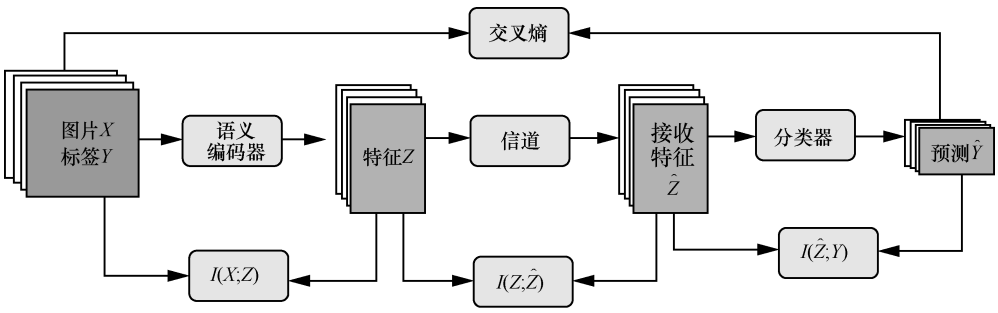


图 11　面向智能任务的 JSCC 方案

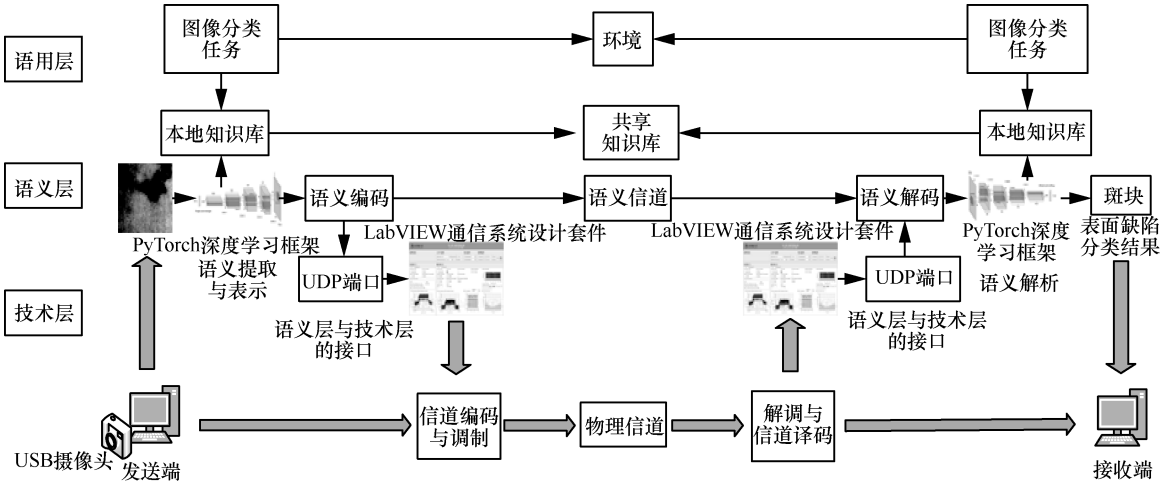


图 12 面向智能任务的语义通信平台架构

表 3 语义层 DNN 初始化参数设置

训练参数	参数取值
Epochs	10
Batchsize	20
优化器	随机梯度下降
学习率	0.001
动量	0.9

表 4 技术层参数设置

实验参数	参数取值
天线距离/m	1.8
发收频率/GHz	5
调制方式	64QAM
码率	0.43
带宽/MHz	40

结合未来的通信需求，语义通信方法及本文所提的语义通信平台可以应用到大量的通信场景中，尤其是数据量大且通信资源相对紧张的环境中，如物联网、车联网、扩展现实、工业互联网、远程医疗和智慧城市等[26,53]。本文以工业互联网场景中的残次品检测任务为例，说明所搭建的语义通信平台的应用方法，其主要分为 2 个阶段。1) 离线训练阶段。首先建模工业互联网的通信传输场景，在离线训练语义通信模型时考虑信道噪声等影响，得到该场景下最优的网络模型参数。2) 在线测试阶段。利用工业摄像头捕捉流水线上的待检测产品图像，图像在本地输入语义编码器提取语义信息，后经量化调制等传输到服务器端，服务器端解调恢复语义信息以完成分类任务，得到是否为残次品的结果，并做出相应决策。

为了验证语义通信平台的性能优势，采取 JPEG 信源编码和 Turbo 信道编码的传统通信方法作为对比方案。传统通信方法中，发送端采集的图像经过 JPEG 压缩编码、信道编码后进行传输，接收端进行信道译码和 JPEG 译码恢复图像后在本地完成分类任务。需要重建信源的语义通信方法选择面向智能任务的图像语义重构（ITOISR, intelligent tasks-oriented image semantic reconstruction）方法[54]（4.2.1 节）；不需要重建信源的语义通信方法选择基于可解释性的语义通信（SCBI, semantic communication based on interpretability）方法[33]（4.2.2 节）与基于信息瓶颈的语义通信（SCBIB, semantic communication based on information bottleneck）方法（5.2 节）。接下来分别对重建信源的 ITOISR、直接面向分类任务的 SCBI 和 SCBIB 在平台上的实验细节进行介绍。

ITOISR。语义编解码器的深度学习模型采用基于 CNN 的自动编码器架构，并在接收端解码器后级联一个基于 ResNet18 的分类网络。发送端首先利用语义编码器提取图像的语义信息，然后将语义信息经过二值量化、64QAM 调制后经天线发出；接收端解调恢复语义信息，并将接收到的语义特征输入语义解码器中，以恢复原始图像，最后将恢复的图像输入 ResNet18 分类网络以完成图像分类任务。

SCBI 和 SCBIB。语义编解码器的深度学习模型基于 ResNet18 网络，根据文献[33]可知，VGG16 等其他深度学习模型同样适用。SCBI 和 SCBIB 的主要区别体现在以下两点：1) 2 种方法训练过程中的损失函数设计不同；2) 2 种方法调节压缩率的方式不同。SCBI 通过在测试过程中依据语义信息重要性不同进行语义裁剪来调整编码码率，不同码率时不需要重复训练网络参数；SCBIB 通过修改编码器输出的神经元

个数来调整编码码率，不同码率时需要重新训练网络。2 种方法的实现流程类似，发送端利用 ResNet18 的卷积层提取语义信息，并对语义信息进行压缩，然后将压缩后的语义信息进行二值量化和 64QAM 调制，接着经天线发出；接收端解调恢复接收到的语义信息，并直接输入全连接层分类器，输出分类结果。

6.1　需要重建信源的语义通信方法性能分析

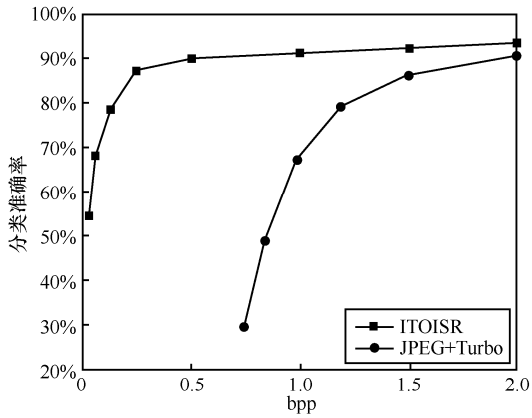
面向智能任务需重建信源的语义通信方法性能不仅关注重建图像的性能，如峰值信噪比（PSNR, peak signal-to-noise ratio）和结构相似性（SSIM, structural similarity）等，同时关注重建图像后续完成智能任务的性能，如分类任务的分类准确率。

图 13(a)对比了不同压缩比时传统通信方法（JPEG+Turbo）与需要重建信源的语义通信方法（ITOISR）的分类性能。像素深度（bpp, bits per pixel）表示存储每个像素所用的位数。对于相同尺寸的图像，bpp 越小就意味着图像可以使用越少的比特来表示，在传输时占用越少的带宽资源。如图 13(a)所示，在所有 bpp 设置下，所提 ITOISR 方法的分类性能都优于 JPEG+Turbo 方法，尤其是在 bpp 较低的情况下。

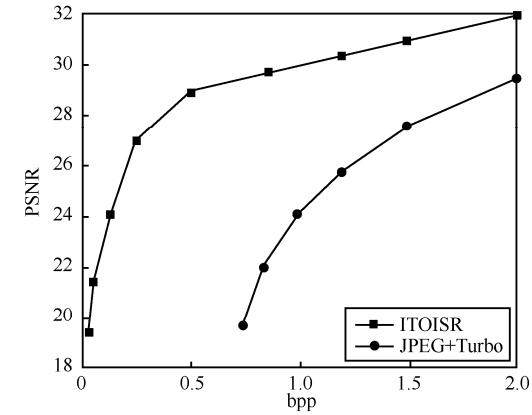
图 13(b)和图 13(c)对比了不同压缩比时传统通信方法（JPEG+Turbo）与需要重建信源的语义通信方法（ITOISR）的图像重建性能，与传统方法相比，语义通信方法图像重建性能也更优，在相同压缩比下有更高的 PSNR 和 SSIM，这是因为语义通信方法可以在保留智能任务所需的语义信息的同时更好地恢复信源数据。

图 14 从直观上将传统通信方法（JPEG+Turbo）和语义通信方法（ITOISR）的图像重建性能进行了对比。从图 14 中可知，语义通信方法重建的图像在主观上也有更高的清晰度，目标物体轮廓更清晰。这与客观指标的评估结果一致，即语义通信方法不仅可以保证更高的智能任务性能，还可以获得更优的重建性能。

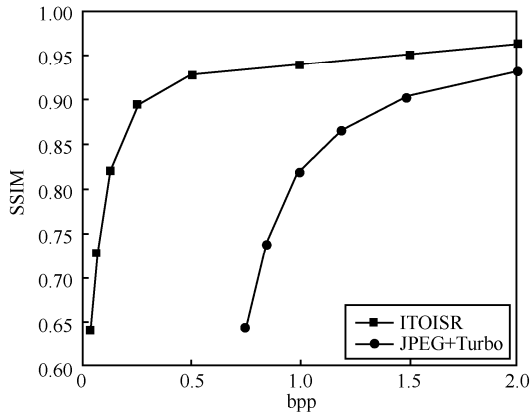
超参数 α 的选择会直接影响语义通信的性能，为了探究重建和分类性能随超参数的变化情况，实验中将 α 的值设置为从小到大依次增加 10 倍，不同 α 值时重建和分类性能变化如图 15 所示。从图 15 中可知，随着 α 的增大，分类任务的准确率逐渐上升，但是图像重构的质量逐渐下降，这是因为 α 越大，$\mathrm{MSE}(X,\hat{X})+\alpha\mathrm{CrossEntropy}$ 中分类部分占比就越大，网络将更加关注分类任务的性能，由于分类和重构所需要的特征分布差异较大，因此更加关注分类的特征将会导致重构性能下降。为了尽量平衡重构性能和分类任务的性能，本节实验中 α 的取值设定为 0.01。



(a) 分类性能对比



(b) PSNR性能对比



(c) SSIM性能对比

图 13　不同压缩比时传统通信方法与需要重建信源的语义通信方法的性能对比

(a) 原图　　(b) JPEG + Turbo　　(c) ITOISR

图 14　2 种通信方法图像重建性能对比

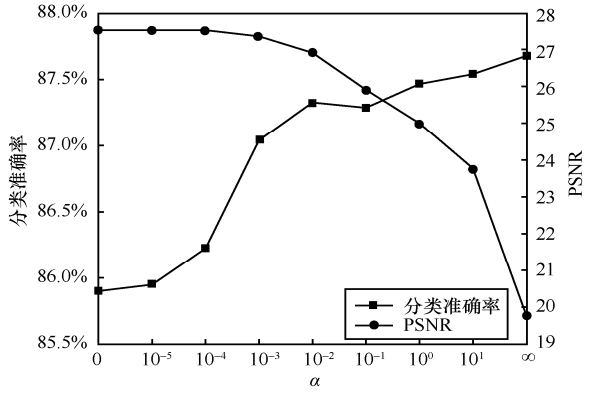


图 15　不同 α 值时重建和分类性能变化

6.2　不需要重建信源的语义通信方法性能分析

不需要重建信源的语义通信系统的性能会受到带宽资源和信噪比的影响，并且时延也是评估通信系统性能的重要指标。因此对于不同通信方式的性能比较主要考虑 3 个方面：1) 完成智能任务的带宽需求；2) 通信系统的抗噪声性能；3) 完成智能任务的时延。

图 16 对比了不同压缩比条件下传统通信方法（JPEG+Turbo）与不需要重建信源的语义通信方法（SCBI 和 SCBIB）的分类性能。实验中将信噪比固定为 20 dB，2 种通信方法分别对相同测试集的所有图像进行传输，计算分类准确率。传统通信方法通过改变 JPEG 压缩率来获得不同 bpp 的图像，然而无论怎么提高压缩率，图像经过 JPEG 压缩后的数据量始终远大于语义通信中特征提取的数据量，并且当压缩率较高时，图像失真严重，分类准确率急剧下降。语义通信方法可以在极大压缩比的情况下，较好地完成分类任务，这是因为语义通信传输图像的语义信息而非图像的所有数据，大大减小了其带宽需求。语义通信的带宽利用率超出传统通信方式的 100 倍，并且随着 bpp 下降，少量数据仍然包含重要的语义信息。

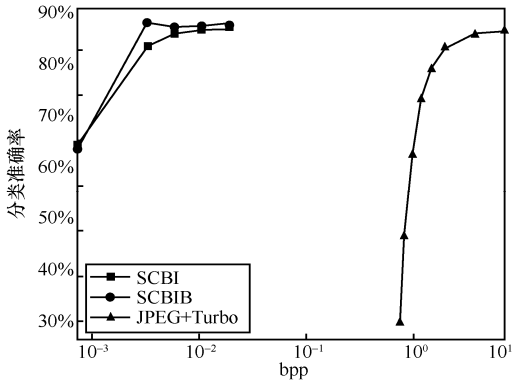


图 16　不同压缩比条件下传统通信方法与不需要重建信源的语义通信方法分类性能比较

图 17 对比了不同信噪比条件下传统通信方法与不需要重建信源的语义通信方法的分类性能。传统通信方法固定 JPEG 压缩率，传输图像经 JPEG 编码后的数据；语义通信方法固定压缩比，传输图像经特征提取的重要语义信息。训练完成后，分别对测试集所有图像进行传输，计算分类准确率。由于 JPEG 图像压缩算法未考虑语义信息，少量的误码会使恢复的图像产生较大的失真，在信道条件恶劣时甚至会出现图像格式错误并且无法恢复的情况，分类性能同样受到较大影响。而语义通信方法抗噪声性能远好于传统通信方法，这是因为语义通信方法传输的数据保留了图像的语义特征，且模型训练时考虑了信道噪声的影响，使其分类性能更优，具有更好的稳健性。

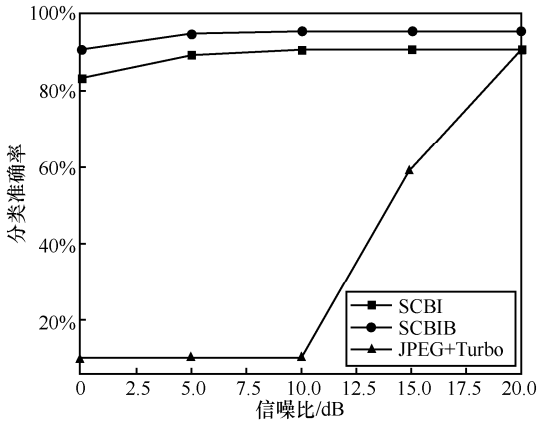


图 17　不同信噪比条件下传统通信方法与不需要重建信源的语义通信方法分类性能比较

图 18 对比了传统通信方法与不需要重建信源的语义通信方法完成智能任务的时延，分别从传输时延与处理时延 2 个方面衡量。传输时延即在单次任务中发收天线间信号传输所用的时间。传统通信方法的处理时延主要包括单次任务 JPEG 图像压缩、USRP 基带处理、图像重构、PyTorch 本地网络计算 4 个部分；语义通信方法的处理时延主要包括单次任务 PyTorch 语义提取、USRP 基带处理、PyTorch 分类网络计算 3 个部分。将信噪比固定为 20 dB，分别传输相同数量的图像，计算单次任务平均时延。相较于传统通信方法，语义通信方法由于传输数据量大大减少，因此在带宽资源相同的情况下，传输时延显著下降；此外，由于不需要进行图像的重构，软硬件的处理负荷减小，处理时延也有所下降。可见本文平台在保证高精度分类性能的同时，大幅减少了端到端智能任务的时延。

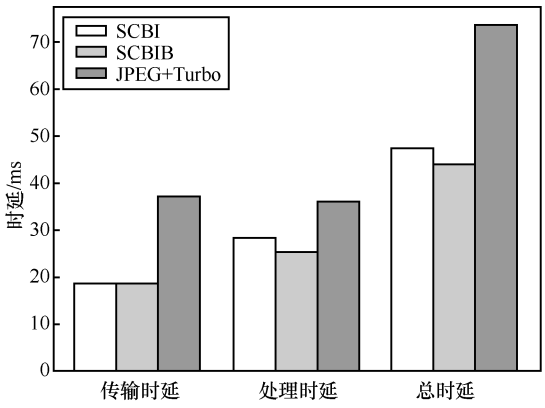


图 18　传统通信方法与不需要重建信源的语义通信方法时延比较

7　语义通信的未来挑战及开放性问题

7.1　语义信息理论研究

本文从智能任务的角度出发，基于模糊数学理论对语义信息量和语义信道容量进行了初步度量，但是如何利用得到的度量方法对语义编解码进行指导存在挑战；如何基于语义熵和语义信道容量公式设计最优的语义编解码架构还不明晰；此外，智能任务性能和语义通信的传输速率之间的理论关系仍有待进一步探究。

7.2　语义通信系统架构设计

语义信息是信源数据简洁有效的表示，可以使语义通信系统在保证智能任务性能的同时大大节省通信资源。但是，目前还没有针对语义通信方法的通用架构，缺乏语义通信系统统一的模型设计。此外，如何对语义通信中的语义噪声建模仍然存在挑战，设计一种对语义噪声具有强稳健性的语义通信架构具有重要意义。

7.3　语义通信知识库的设计与更新

语义通信性能很大程度上取决于通信双方本地及共享知识库的完备性，合理设计语义通信知识库至关重要。此外，语义知识会随着人类和社会的发展而变化，同时需要考虑语义实体和其他实体间的关系变化，因此如何基于终身学习的概念设计语义通信知识库的更新机制也需要进一步研究。

7.4　多模态信源语义编解码方法

目前，大多数语义通信相关的工作缺乏对输入信源模态的考量，都是针对某一种信源（如图像）设计的通信架构和训练策略，而在实际通信场景中，单纯针对一种模态（如图像）的信源难以完成很多智能任务（如视觉智能问答），因此如何设计合理的语义编解码方法以融合多模态信源语义信息至关重要，如何设计适用于不同模态信源的统一的编解码架构也需要进一步探索。

7.5　语义通信系统中的资源分配策略研究

目前，语义通信方法相关工作侧重于端到端的通信架构设计，忽略了资源分配对语义通信性能的影响。与传统通信系统中以最大化系统速率或最小化时延为优化目标的资源分配不同，语义通信系统中应综合考虑通信和智能任务[55]，以提升语义通信系统的效率为目标。首先需要探究如何定义和量化语义通信的性能和效率，其次设计面向智能任务的语义通信系统的资源分配优化问题存在挑战。

8　结束语

本文首先从语用的角度出发提出了面向智能任务的语义通信方法的基本架构，并简要给出了面向智能任务的语义信息度量的基本方法；然后，综述了现有语义编码相关的工作以及未来与智能任务结合后可行的语义编码思路；接着，综述了现有联合信源信道编码的相关工作，提出了从信息瓶颈理论出发的联合信源信道编码的语义通信方法；然后，搭建了端到端的语义通信平台，实验验证了所提思路的有效性；最后，对语义通信未来的挑战和开放性问题进行了进一步的思考和总结。毫无疑问，语义通信将继续保持快速发展，其中大量的基础概念和基础问题亟须进一步讨论和完善，极有可能成为 6G 时代技术创新和突破的主要领域，需要学术同仁共同推动实现。

参考文献：

[1] CHEN M Z, CHALLITA U, SAAD W, et al. Artificial neural networks-based machine learning for wireless networks: a tutorial[J]. IEEE Communications Surveys & Tutorials, 2019, 21(4): 3039-3071.

[2] STRINATI E C, BARBAROSSA S. 6G networks: beyond Shannon towards semantic and goal-oriented communications[J]. Computer Networks, 2021, 190: 107930.

[3] WEAVER W. Recent contributions to the mathematical theory of communication[J]. ETC: A Review of General Semantics, 1953, 10(4): 261-281.

[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.

[5] CARNAP R, BAR-HILLEL Y. An outline of a theory of semantic information[R]. 1952.

[6] POPPER K . Conjectures and refutations[M]. New York: Routledge, 2002.

[7] ZADEH L A. Fuzzy sets[J]. Information and Control, 1965, 8(3):

338-353.

[8] ZADEH L A. Probability measures of fuzzy events[J]. Journal of Mathematical Analysis and Applications, 1968, 23(2): 421-427.

[9] LUCA A D, TERMINI S. Algebraic properties of fuzzy sets[J]. Journal of Mathematical Analysis and Applications, 1972, 40(2): 373-386.

[10] FLORIDI L. Outline of a theory of strongly semantic information[J]. Minds and Machines, 2004, 14(2): 197-221.

[11] LU C G. From Bayesian inference to logical Bayesian inference: a new mathematical frame for semantic communication and machine learning[J]. arXiv Preprint, arXiv: 1809.01577, 2018.

[12] LU C G. The third kind of Bayes' theorem links membership functions to likelihood functions and sampling distributions[C]//International Conference on Cognitive Systems and Signal Processing. Berlin: Springer, 2019: 268-280.

[13] 钟义信. 信息科学原理(第 3 版)[M]. 北京: 北京邮电大学出版社, 2002.
ZHONG Y X. Principles of information science(3rd edition)[M]. Beijing: Beijing University of Posts and Telecommunications Press, 2002.

[14] 钟义信. 面向智能研究的全信息理论: 纪念 Shannon 信息论 50 周年[J]. 北京邮电大学学报, 1998, 21(4): 1-6.
ZHONG Y X. Intelligence oriented comprehensive information theory: in memory of the 50th anniversary of Shannon information theory[J]. Journal of Beijing University of Posts and Telecommunications, 1998, 21(4): 1-6.

[15] 钟义信, 张瑞. 信息生态学与语义信息论[J]. 图书情报知识, 2017(6): 4-11.
ZHONG Y X, ZHANG R. Information ecology and semantic information theory[J]. Documentation, Information & Knowledge, 2017(6): 4-11.

[16] 石光明, 李莹玉, 谢雪梅. 语义通讯: 智能时代的产物[J]. 模式识别与人工智能, 2018, 31(1): 91-99.
SHI G M, LI Y Y, XIE X M. Semantic communications: outcome of the intelligence era[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(1): 91-99.

[17] BAO J, BASU P, DEAN M K, et al. Towards a theory of semantic communication[C]//Proceedings of 2011 IEEE Network Science Workshop. Piscataway: IEEE Press, 2011: 110-117.

[18] POPOVSKI P, SIMEONE O, BOCCARDI F, et al. Semantic-effectiveness filtering and control for post-5G wireless connectivity[J]. Journal of the Indian Institute of Science, 2020, 100(2): 435-443.

[19] KOUNTOURIS M, PAPPAS N. Semantics-empowered communication for networked intelligent systems[J]. IEEE Communications Magazine, 2021, 59(6): 96-102.

[20] YANG Y, GUO C L, LIU F F, et al. Semantic communications with AI tasks[J]. arXiv Preprint, arXiv: 2109.14170, 2021.

[21] KALFA M, GOK M, ATALIK A, et al. Towards goal-oriented semantic signal processing: applications and future challenges[J]. Digital Signal Processing, 2021, 119: 103134.

[22] 牛凯, 戴金晟, 张平, 等. 面向 6G 的语义通信[J]. 移动通信, 2021, 45(4): 85-90.
NIU K, DAI J C, ZHANG P, et al. 6G-oriented semantic communications[J]. Mobile Communications, 2021, 45(4): 85-90.

[23] 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: 60-73.

[24] 石光明, 肖泳, 李莹玉, 等. 面向万物智联的语义通信网络[J]. 物联网学报, 2021, 5(2): 26-36.
SHI G M, XIAO Y, LI Y Y, et al. Semantic communication networking for the intelligence of everything[J]. Chinese Journal on Internet of Things, 2021, 5(2): 26-36.

[25] UYSAL E, KAYA O, EPHREMIDES A, et al. Semantic communications in networked systems[J]. arXiv Preprint, arXiv: 2103.05391, 2021.

[26] LAN Q, WEN D, ZHANG Z, et al. What is semantic communication? a view on conveying meaning in the era of machine intelligence[J]. Journal of Communications and Information Networks, 2021, 6(4): 336-371.

[27] FARSAD N, RAO M, GOLDSMITH A. Deep learning for joint source-channel coding of text[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 2326-2330.

[28] XIE H Q, QIN Z J. A lite distributed semantic communication system for internet of things[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(1): 142-153.

[29] WENG Z Z, QIN Z J. Semantic communication systems for speech transmission[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(8): 2434-2444.

[30] BOURTSOULATZE E, KURKA D B, GÜNDÜZ D. Deep joint source-channel coding for wireless image transmission[J]. IEEE Transactions on Cognitive Communications and Networking, 2019, 5(3): 567-579.

[31] LEE C H, LIN J W, CHEN P H, et al. Deep learning-constructed joint transmission-recognition for Internet of things[J]. IEEE Access, 2019, 7: 76547-76561.

[32] JANKOWSKI M, GÜNDÜZ D, MIKOLAJCZYK K. Wireless image retrieval at the edge[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(1): 89-100.

[33] 刘传宏, 郭彩丽, 杨洋, 等. 人工智能物联网中面向智能任务的语义通信方法[J]. 通信学报, 2021, 42(11): 97-108.
LIU C H, GUO C L, YANG Y, et al. Intelligent task-oriented semantic communication method in artificial intelligence of things[J]. Journal on Communications, 2021, 42(11): 97-108.

[34] XIE H Q, QIN Z J, LI G Y. Task-oriented multi-user semantic communications for VQA[J]. IEEE Wireless Communications Letters, 2022, 11(3): 553-557.

[35] XIE H, QIN Z, TAO X, et al. Task-oriented multi-user semantic communications[J]. arXiv Preprint, arXiv: 2112.10255, 2021.

[36] 李蕾, 周延泉, 钟义信. 基于语用的自然语言处理研究与应用初探[J]. 智能系统学报, 2006, 1(2): 1-6.
LI L, ZHOU Y Q, ZHONG Y X. Pragmatic information based NLP research and application[J]. CAAI Transactions on Intelligent Systems, 2006, 1(2): 1-6.

[37] LUCA A D, TERMINI S. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory[J]. Information and Control, 1972, 20(4): 301-312.

[38] 吴伟陵. 广义信息源与广义熵[J]. 北京邮电大学学报, 1982(1): 29-41.
WU W L. Generalized information source and generalized entropy[J]. Journal of Beijing University of Posts and Telecommunications, 1982(1): 29-41.

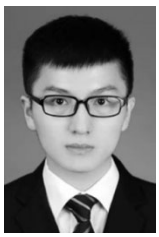
[39] BASU P, BAO J, DEAN M, et al. Preserving quality of information by using semantic relationships[J]. Pervasive and Mobile Computing, 2014, 11: 188-202.
[40] HOANG D T, LONG P M, VITTER J S. Rate-distortion optimizations for motion estimation in low-bit-rate video coding[C]//Digital Video Compression: Algorithms and Technologies. [S.l.:s.n.], 1996: 18-27.
[41] HUYNH-THU Q, GHANBARI M. Scope of validity of PSNR in image/video quality assessment[J]. Electronics Letters, 2008, 44(13): 800.
[42] WANG Z, BOVIK A C. A universal image quality index[J]. IEEE Signal Processing Letters, 2002, 9(3): 81-84.
[43] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
[44] KIM J, LEE S. Deep learning of human visual sensitivity in image quality assessment framework[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1969-1977.
[45] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 586-595.
[46] HU Y Q, ZHOU W, ZHAO S X, et al. SDM: semantic distortion measurement for video encryption[C]//Proceedings of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. Piscataway: IEEE Press, 2018: 764-768.
[47] HUSSAIN A J, AL-FAYADH A, RADI N. Image compression techniques: a survey in lossless and lossy algorithms[J]. Neurocomputing, 2018, 300: 44-69.
[48] YANG Z H, WANG Y H, XU C, et al. Discernible image compression[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 1561-1569.
[49] 韦来生. 数理统计(第 2 版)[M]. 北京: 科学出版社, 2015.
WEI L S. Mathematical statistics(2nd edition)[M]. Beijing: Science Press, 2015.
[50] SHWARTZ-ZIV R, TISHBY N. Opening the black box of deep neural networks via information[J]. arXiv Preprint, arXiv: 1703.00810, 2017.
[51] LI Y Y, ZHAO P P, WANG D Q, et al. Learning disentangled user representation based on controllable VAE for recommendation[C]// International Conference on Database Systems for Advanced Applications. Berlin: Springer, 2021: 179-194.
[52] TISHBY N, PEREIRA F C, BIALEK W. The information bottleneck method[J]. arXiv Preprint, arXiv: physics/0004057, 2000.
[53] QIN Z J, TAO X M, LU J H, et al. Semantic communications: principles and challenges[J]. arXiv Preprint, arXiv: 2201.01389, 2022.
[54] LIU F F, TONG W J, SUN Z F, et al. Task-oriented semantic communication with semantic reconstruction: an extended rate-distortion theory based scheme[J]. arXiv Preprint, arXiv: 2201.10929, 2022.
[55] 陈九九, 冯春燕, 郭彩丽, 等. 车联网中视频语义驱动的资源分配算法[J]. 通信学报, 2021, 42(7): 1-11.
CHEN J J, FENG C Y, GUO C L, et al. Video semantics-driven resource allocation algorithm in Internet of vehicles[J]. Journal on Communications, 2021, 42(7): 1-11.

[作者简介]

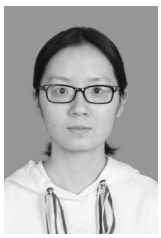
刘传宏（1998− ），男，安徽池州人，北京邮电大学博士生，主要研究方向为深度学习、语义通信、资源分配等。

郭彩丽（1977− ），女，山西太原人，博士，北京邮电大学教授、博士生导师，主要研究方向为语义通信、无线移动通信技术、认知无线电、信号检测与估值、车联网、可见光通信、视觉智能计算、社交跨媒体数据挖掘与分析等。

杨洋（1991− ），男，湖南娄底人，博士，北京邮电大学讲师，主要研究方向为可见光通信、室内定位技术、车联网技术、语义通信技术等。



陈九九（1994− ），男，湖南平江人，北京邮电大学博士生，主要研究方向为车联网资源分配、语义通信、强化学习算法等。



朱美逸（1999− ），女，湖北保康人，北京邮电大学博士生，主要研究方向为语义通信、车联网通信、强化学习算法等。

孙鲁楠（1996− ），女，辽宁葫芦岛人，北京邮电大学博士生，主要研究方向为语义通信、图像传输、信源信道编码等。