

智简语义通信的链路设计及关键技术研究



Link Design and Key Technologies of Intellicise Semantic Communication System

孙梦颖/SUN Mengying, 熊华超/XIONG Huachao,
王怡宁/WANG Yining, 韩书君/HAN Shujun,
许晓东/XU Xiaodong

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

DOI: 10.12142/ZTETJ.202302009

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20230411.1438.018.html>

网络出版日期: 2023-04-11

收稿日期: 2023-01-28

摘要: 智简语义通信是一种模型驱动的语义通信新范式, 融合了人工智能与通信技术, 实现通信对象间高效的语义交互。提出了智简语义通信系统的链路结构和关键技术, 从链路结构、语义模型增强、模型传输3个角度, 实现语义通信系统的整体性能增强。提出了4种智简语义通信的关键技术, 为未来6G赋能多种垂直行业和新场景应用提供了参考: 语义知识图谱增强的智简通信技术通过增加语义知识这一信息维度, 提升了语义知识恢复准确度和传输效率; 语义知识图谱的云-边-端协同预缓存技术可以实现语义知识图谱的高效获取, 辅助语义恢复性能; 模型传输与部署技术可以实现模型、网络资源与终端能力的有效适配; 语义模型传输与恢复级联过程的部署成功率大规模语义模型传输及资源部署提供理论依据。

关键词: 智简语义通信; 模型处理; 模型传输; 语义知识库

Abstract: Intellicise semantic communications is a new paradigm of model-driven semantic communication, which integrates artificial intelligence and communication technologies to realize efficient semantic interaction between communication objects. In general, the link structure and key technologies of the intellicise semantic communication system are proposed from three perspectives: link structure, semantic model enhancement, and model transmission, and the overall performance of the semantic communication system is enhanced. Four key technologies of intellicise semantic communications are proposed, which provide a reference for the future 6G empowerment of various vertical industries and new scene applications: a semantic knowledge graph-enhanced intellicise communication technology improves the accuracy of semantic knowledge recovery and the transmission efficiency by increasing the information dimension of semantic knowledge; the cloud-edge-device collaborative pre-cache technology of semantic knowledge graph can realize the efficient acquisition of semantic knowledge graph and assist semantic recovery; model transmission and deployment technology can achieve the effective adaptation of the models, network resources and terminal capabilities; the deployment success probability of the semantic model related to the transmission-recovery cascade process is analyzed, which provides a theoretical basis for large-scale semantic model transmission and resource deployment.

Keywords: intellicise semantic communication; model processing; model transmission; semantic knowledge base

随着传输能力的不断增强, 移动通信系统以系统复杂度换取性能增益的堆叠式发展将无法满足日益增长的网络智能化和服务多样化的业务需求^[1]。因此, 6G亟需发展面向未来多类型应用的移动通信新范式。

语义通信是一种全新的通信范式, 通过将任务和需求与信息传输有机融合, 大幅度提升通信效率, 改善用户的业务体验。依托智能与通信技术、网络技术等的深度融合, 通过

语义特征提取、语义信息传输以及语义信息恢复, 语义通信可实现信息的高效准确传达和精准控制^[2]。语义通信以节点智能为基础, 赋予节点类人感知能力, 实现达意传输, 满足人-机-物之间的敏捷交互。同时, 语义通信也将赋能智慧车联网、元宇宙、智能制造、多维感官互联、新兴多媒体、智慧城市等新场景、新业态^[3-4], 如图1所示。

1 智简语义通信系统

1953年, W. WEAVER^[5]提出了通信传输的3个层次, 即语法层、语义层和语用层。其中, 语法层解决的是符号级通

基金项目: 国家重点研发计划(2020YFB1806900); 国家自然科学基金(62201079、61871045)



▲图1 语义通信赋能新场景、新业态

信中信息如何精准传输的问题,语义层关注的是所传递的符号信息如何准确地传达所需的含义,语用层针对的是所接收的语义信息是否能够以最佳的方式被利用。这3个层次之间相互重叠又相互关联。从1G到5G,学术界和工业界聚焦于语法信息层次,以期增强通信系统的传输性能。近年来,面向语义层的语义通信成为学术界关注的焦点,也成为6G极具竞争力的关键技术^[6]。

1.1 智简语义通信系统的概念

基于上文所提的通信三大层次,智简语义通信系统着眼于语义传输,以人工智能和信息论等多学科的科学理论为指导,实现面向需求和任务的通信。智简语义通信系统深挖语法、语义与语用之间的转换关系,将需求信息进行特征提取后转化为语义信息,通过编码,再转化成语法信息。借助深度神经网络的刻画能力,智简语义通信系统对信源信息的需求及任务进行准确地识别和特征提取,并在接收端进行原始信息重建,实现精准高效的传输。同时,智简语义通信也进行了基础理论的创新:以香农信息论为基础,拓展了语义熵和系统熵^[7]的概念,从理论上度量语义信息量。在网络传输层面上,“节点极智、链路极柔、网络极简”将是智简通信的主要特征,这有助于网络性能的全面提升。通过柔性简约的通信链路和按需编排的资源配置,网络中的智能节点将自主执行数据感知、处理、应用和传输等功能,实现网络由繁变简、多维资源的高效利用。智简语义通信系统将颠覆原有的通信模式,赋予信源到信宿包括感知能力、通信能力和理解能力在内的更多新能力。

1.2 智简语义通信系统的需求和挑战

智简语义通信系统旨在建立一套全新的通信体系,通过海量的语义模型和语义知识图谱对智能性进行支撑,实现人-机-物及智能体之间的智能交互。语义模型训练及语义知识图谱的构建将消耗巨大的时间与计算资源,因此,提升模型的训练效率,降低模型的训练成本,实现模型在网络中

高效的传输和部署,将是智简语义通信的重要基础,也是所面临的关键挑战之一。

未来6G网络的信息多样性将得到进一步拓展,存在文本、语音、图像、触觉等多模态的信息形式,这给语义特征提取与建模带来巨大挑战。在面向需求和任务的通信中,发送端需要对接收端的需求或任务进行精准预测,以保障对重要信息的准确表达,而这进一步增加了语义信息的提取难度。多类型的信息具有不同的数据格式,如何对语义信息进行一致化建模,实现不同模态间的语义信息的高效转换,也将是一大挑战。

语义信息的传输与准确恢复是智简语义通信系统的重要过程。原始信息的准确传达,不仅需要语义信息的可靠传输,也需要语义解码模型的准确适配。在这种双重的语义信息恢复条件制约下,如何实现语义信息及语义模型的高效准确的传输也是智简语义通信系统的一大挑战。

智简语义通信是建立在节点智能性的基础上的。在当前阶段,模型的体量无法适配于节点受限的处理能力和网络资源。在保障模型性能的前提下,如何将大型模型转换为轻量级的小型模型是语义通信的迫切需求。不同节点的能力也存在不均衡性,因此如何根据不同节点的需求和能力部署相应的模型,以提升语义恢复准确度,并降低服务响应时延,也是智简语义通信的一个关键挑战。

1.3 智简语义通信系统的链路结构

利用深度神经网络技术,人们对文字、图像、视频以及多模态业务的语义通信技术进行了探索^[8-11]。在此基础上,业界提出了基于人工智能模型的以网络智能决策和模型传输为特征的智简语义通信系统,如图2所示。在智简语义通信系统中,收发端分别部署语义编码模块和译码模块,模块对应的语义模型分别用于提取和恢复语义信息。边缘或云端生成语义模型,并根据收发端的请求,部署适配的语义模型。同时,收发端将语义模型存储在各自的语义模型库中。发送端将原始信息输入到语义提取与表征模块,得到语义信息,并通过联合的语义编码和信道编码将语义信息转化为比特数据,再进行传输。接收端对接收到的比特数据进行联合信道译码、语义译码,以及语义信息恢复重建,恢复出原始信息。由于上下文、通信环境等背景因素会影响语义信息的恢复,语义译码模块对背景因素带来的误差可进行补偿。

为了保障模型的适配性,需要不断对模型进行优化和更新。迁移学习是一种高效的方式,只需要对部分语义编码器和译码器的神经网络层进行重新设计和训练,重用预训练模型中某些网络层的权值,就可以减少模型再训练的时间、资

源等消耗。智简通信系统结合“残差传播”理念，对智简模型进行切分传输，即信源端根据模型切分技术将智简模型分为多个切片。这些切片在多路进行传输，并于信宿端重新组建。模型切分传输的方式一方面平衡无线网络负载，减少网络拥塞，提升模型并行传播速率；另一方面促进不同节点进行模型共享，利于后续模型的传输。模型压缩技术在保障信息恢复准确性的前提下，可有效地提升网络传输效率。

2 知识库增强的智简语义通信系统设计

知识图谱是一种语义表示，可以对现实世界中的实体、概念、属性以及它们之间的关系进行建模，具有极强的灵活性和表达能力。语义知识库是由诸多的语义知识图谱组成的。在上一节所提的智简语义通信链路架构的基础上，利用知识图谱的灵活精简表征的功能来部署语义知识库可有效地增强语义通信性能^[12]。

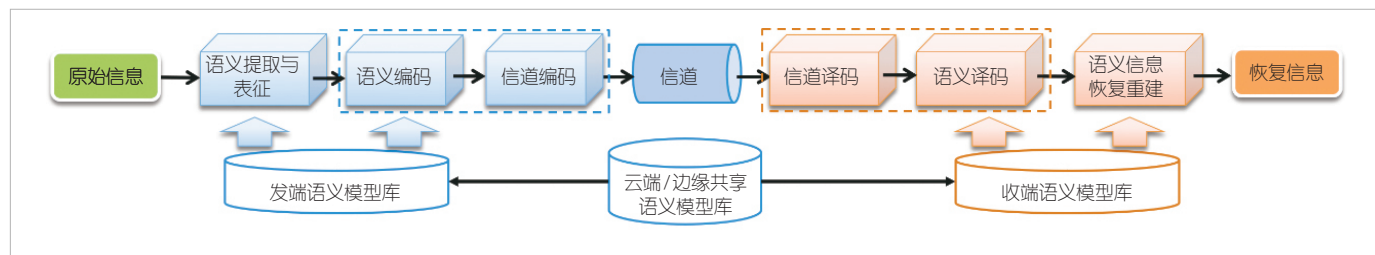
2.1 知识库增强的智简语义通信关键技术

当前，针对文本、语音、图像等多种信息类型的语义通信的研究工作已取得诸多进展，如基于深度学习的语义通信（Deep-SC）^[8-9]和基于深度学习的信源信道联合编码（Deep-JSCC）^[10-11]等。大部分的语义编译码是基于Transformer模型来实现的。以Transformer模型为基础，智简语义通信系统引入了语义知识图谱以增强语义通信的性能，构建出如图3所示的基于知识增强的智简语义通信系统链路结构。

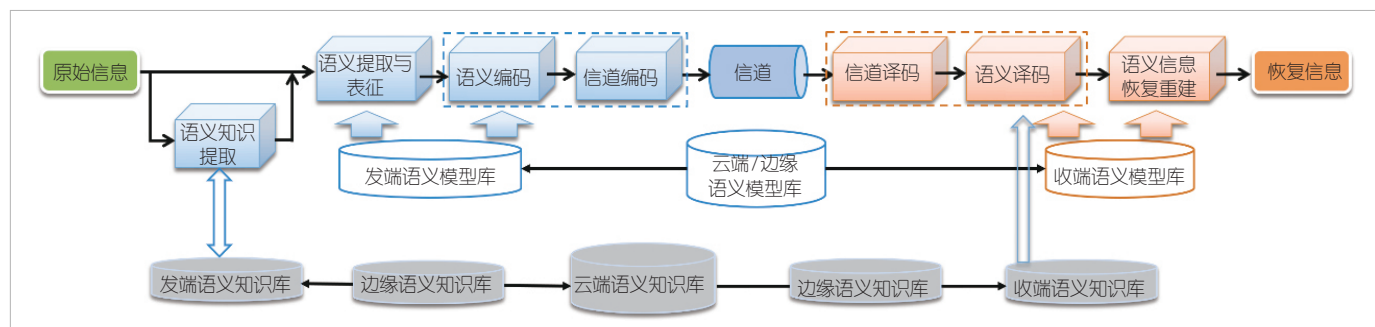
语义知识是从原始信息的语义知识中提取得到的。语义

提取和表征编码器将语义知识和原始信息进行联合语义提取。所提取出来的语义信息将会根据重要性进行排序，重要性程度较高的语义信息将会以更可靠的方式进行联合语义信道编码。接收端利用相匹配的语义知识图谱来增强语义译码，辅助恢复信息。语义知识图谱需进行云-边-端同步，以保障收发端能够基于相匹配的语义知识图谱来增强语义提取与恢复。为了使接收端能够从语义知识库中快速查找到相应的语义知识图谱，我们在发送端的信道导频之前增加了语义导频，以使接收端基于语义导频快速查找到相应的语义知识图谱，增强语义恢复效率和准确性。从语义提取模块输出的数据格式为浮点型，我们需对浮点型数据进行量化和数学变换以降低编码冗余。根据信道状态及语义信息重要程度，我们对提取的语义数据进行可变码长的方式进行编码，在保障语义传输准确度的前提下，提升传输效率。

性能评估也是语义通信系统设计中的重要一环。在面向文本的语义通信中，误字率（WER）^[13]和双语替换评价（BLEU）^[14]是两种经典的文本语义信息恢复准确度的评估指标。基于语义知识增强的智简语义通信系统已被验证能够有效地提升语义信息恢复准确性能和效率。此外，针对图像传输，峰值信噪比（PSNR）和结构相似指数（SSIM）是常用的信息相似度评估的方法^[15]。针对语音传输，音频信号转为文本过程采用WER来评估，转成的文本之间采用Transformer模型的双向编码器（BERT）方法^[16]来评估。如何高效地构建多种模态信源信息的知识图谱，提升语义知识图谱的更新效率，降低传输冗余，仍然是智简语义通信设计中亟待



▲图2 智简语义通信系统链路结构



▲图3 知识库增强的智简语义通信链路

解决的问题。

2.2 智简语义通信中基于缓存增强的语义知识库构建方法

如2.1节所述,智简语义通信系统需在云-边-端协同部署语义知识库,以提升接收端的语义信息恢复准确度。语义知识图谱的生成往往需要大量的计算和存储资源,在云端生成知识图谱是一种行之有效的方式。然而,受到多维资源的制约,云与用户之间的多跳传输,会给语义知识库的传输时延、传输成功率和语义恢复准确度带来较大的影响。每个用户所需的语义知识图谱繁杂且具有时变性,因此边缘服务器或用户进行语义知识库预缓存是一种必然趋势,可有效降低语义知识库获取时延,具体的云-边-端协同语义知识图谱的预缓存机制如图4所示。基于大数据分析和人工智能技术,根据用户侧的历史请求构建知识图谱的预测模型,动态调整和定期更新语义知识库,是语义知识图谱高效谱缓存和快速获取的关键^[17]。采用数学拟合的方式,得出针对特定应用场景的语义知识图谱的恢复准确度与知识图谱大小之间的数学表达,这对定量地缓存语义知识图谱来说具有重要意义。如何利用用户偏好和能力,实现语义知识图谱的分布式协同存储,也是进一步改善智简语义通信系统性能的有效方式之一。

3 智简语义通信中的模型传输关键技术

面向人-机-物智能化应用,节点极智是必然的发展趋势。节点利用其智能性进行网络感知与推理。智简语义通信需要通信对象配置相同的语义知识库和语义模型。语义模型

的训练和传输部署也是未来网络的一项基本业务。模型的训练精度和更新频率将影响智能节点语义感知性能,轻量化且适配的语义模型部署是智简语义通信系统的基石。

3.1 语义模型处理方法

当前比较成熟的语义模型往往体量巨大,以谷歌推出的Transformer模型为例,其参数高达1.6万亿个,训练时间和成本以及模型运行所耗费的计算资源都是终端设备所无法承受的。因此,需要针对特定的应用和业务,并考虑终端设备的处理能力,对模型进行压缩、剪枝^[18]、量化、蒸馏^[19]、切片等处理。处理后的轻量级模型将被分发至边缘服务器及终端设备进行存储。终端设备利用部署的模型,处理感知数据,执行推理应用。

接下来,我们会对这几种模型处理技术进行简要介绍。

1) 模型剪枝是通过去除两个神经元之间多余的或不重要的连接,或直接去除不需要的神经元及其相关分支,减小模型规模。

2) 模型量化主要是通过减小权重参数,在服务质量(QoS)容忍范围内降低模型的精度。这种方式也是通过减小模型规模以节省缓存空间,加速模型训练,并减少传输的比特数据量。

3) 模型蒸馏是基于大规模的教师模型,在满足QoS的情况下,训练一个轻量级的学生模型。该方法可以将预先训练的复杂模型的知识与泛化能力转移到一个结构更简单的网络上。具体来说,该方法利用教师模型的输出作为监督信号来训练学生模型,实现一个具有相似知识的简单网络。

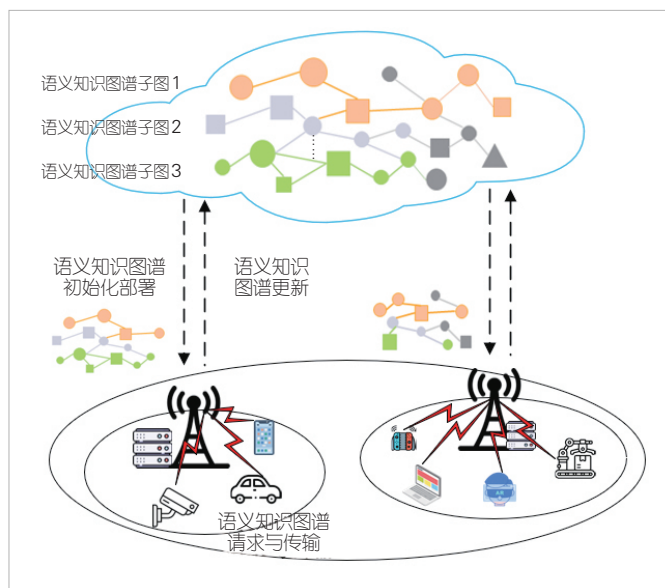
4) 模型选择是根据终端的QoS要求选择合适的模型,进一步优化整个系统的延迟、能效、能耗和决策精度。该方法是将相同功能的不同尺寸和准确度的模型全部存储下来,以按需分配,但难免造成缓存冗余。

5) 语义切片技术是将模型进行分层化设计(包括基础模型和增强切片),基础模型以较低的通信和计算开销独立完成语义通信,也可以根据QoS要求、信道条件和传输目标选择合适的模型切片来增强基础模型的性能。

6) 模型处理是智简语义通信系统实现模型有效传输及部署的关键,值得我们进行深入探索。

3.2 语义模型传输方法

面向智简语义通信的模型传输方案旨在将面向数据认知的智能能力引入数据生成节点和服务生成节点,对模型和网络资源进行端到端编排,并充分利用网络中的分布式计算资源和数据,在架构层面实现通信连接和智能能力的融合。语



▲图4 云-边-端协同语义知识图谱缓存机制

义模型的训练过程是网络中的单个设备无法承担的,往往需要特定的训练中心来执行。网络针对特定的任务进行模型的定制化处理,并基于用户的请求,将语义模型传输给各个终端设备。基于云-边-端3层网络架构,语义模型可实现多级分布式的缓存和管理,边缘及终端也可基于本地信息进行模型的微调,并将微调信息反馈到网络,实现全局的更新。边缘和设备根据模型流行度及用户偏好提前缓存部分语义模型,终端设备之间也可以共享语义模型,这将实现模型及时获取和缓存效率提升。联邦学习是进行模型分布式训练及更新的有效方式之一,设备之间通过参数交互进行共同的模型优化,保障了数据的隐私和安全。

模型选择是适配网络通信能力和节点算力的重要方式,即对相同功能且不同大小和精度的模型进行选择并传输。为了验证模型选择及模型传输在语义通信中的优势,我们构建了一个模型选择与通信资源配置的联合优化方案。该方案提出了一个新的评估指标,即模型有效性,其定义为模型传输成功率与语义恢复准确率的乘积。其中,传输成功率被定义为模型在容忍时延内传输至接收端的概率。在仿真验证中,主要参数设置为:对于相同功能的模型集合,对应的恢复错误率为{0.125, 0.150, 0.175, 0.100, 0.250, 0.300},相应的模型大小分别为{82.274 6, 46.671 9, 28.899 3, 19.079 4, 9.532 9, 5.407 8} Mbit,上述参数由针对图像传输的语义模型训练所得。如图5所示,在给定通信资源的情况下,与固定模型大小和准确度的方案对比,模型选择和通信资源联合配置可有效地提升语义模型的有效性。随着用户数增多,单用户获取的资源减少,所提方案可有效平衡模型大小与准确度。上述研究的论文将于后续发表。为了实现模型驱动的智简语义通

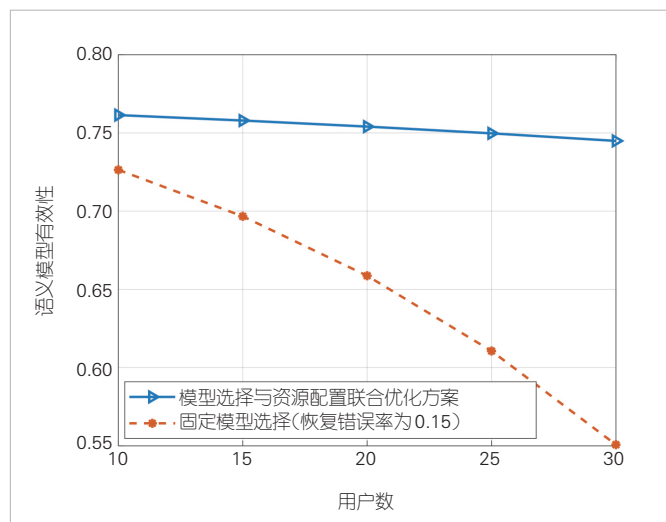
信系统设计,一方面需要模型预处理,使其适配于终端设备;另一方面也需要针对模型传输进行模型选择和资源配置的联合优化,以增强总体网络性能。

4 结束语

随着元宇宙、智慧车联网、新媒体等新场景和新业务的出现,智简语义通信对6G关键服务起到重要的支撑作用。业界也正从不同的角度探索语义通信基础理论及关键技术,并设计原型系统,以实现多类型信息的高效语义传输。依托人工智能与模型传输技术,智简语义通信将实现网元之间高效敏捷交互,有望成为6G的颠覆性技术之一。

参考文献

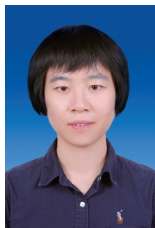
- [1] 张平, 陶运铮, 张治. 5G若干关键技术评述[J]. 通信学报, 2016, 37(7): 15–29. DOI: 10.11959/j.issn.1000-436x.2016130
- [2] 石光明, 肖泳, 李莹玉, 等. 面向万物智联的语义通信网络[J]. 物联网学报, 2021, 5(2): 26–36. DOI: 10.11959/j.issn.2096-3750.2021.00209
- [3] 牛凯, 戴金晟, 张平, 等. 面向6G的语义通信[J]. 移动通信, 2021, 45(4): 85–90. DOI: 10.3969/j.issn.1006-1010.2021.04.014
- [4] 徐英姿, 刘原, 时梦然, 等. 语义在通信中的应用综述[J]. 电信科学, 2022, 38(S1): 43–59. DOI: 10.11959/j.issn.1000-0801.2022107
- [5] WEAVER W. Recent contributions to the mathematical theory of communication[J]. ETC: a review of general semantics, 1953, 10(4): 261–281
- [6] 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. DOI: 10.1016/j.eng.2021.11.003
- [7] 张平, 许文俊, 王凤玉, 等. 智简空地一体化网络[J]. 无线电通信技术, 2022, 48(3): 381–384. DOI: 10.3969/j.issn.1003-3114.2022.03.001
- [8] 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
- [9] 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. DOI: 10.1109/JSAC.2021.3087240
- [10] DAI J C, WANG S X, TAN K L, et al. Nonlinear transform source-channel coding for semantic communications[J]. IEEE journal on selected areas in communications, 2022, 40(8): 2300–2316. DOI: 10.1109/JSAC.2022.3180802
- [11] WANG S, DAI J, LIANG Z, et al. Wireless deep video semantic transmission [EB/OL]. [2023-01-22]. <https://arxiv.org/abs/2205.13129>
- [12] XU X, XIONG H, WANG Y, et al. Knowledge-enhanced semantic communication system with OFDM transmissions [EB/OL]. [2022-07-01] [2023-01-23]. <https://www.sciengine.com/SCIS/doi/10.1007/s11432-022-3624-4>
- [13] FARSAFAD 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 (ICASSP). IEEE, 2018: 2326–2330. DOI: 10.1109/ICASSP.2018.8461983
- [14] 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. ACM, 2002: 311–318. DOI: 10.3115/1073083.1073135
- [15] SARA U, AKTER M, UDDIN M S. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study[J]. Journal of computer and communications, 2019, 7(3): 8–18. DOI: 10.4236/jcc.2019.73002
- [16] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2018-10-11] [2023-01-22]. <https://arxiv.org/abs/1810.04805>
- [17] CHE Y, XIONG H C, HAN S J, et al. Cache-enabled knowledge base construction strategy in semantic communications [C]//Proceedings of 2022 IEEE Globecom Workshops (GC Wkshps). IEEE, 2023: 1507–1512. DOI: 10.1109/GCWkshps56602.2022.10008494



▲图5 语义模型有效性随用户数的变化曲线

- [18] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural networks [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems – Volume 1. ACM, 2015: 1135–1143. DOI: 10.5555/2969239.2969366
- [19] WANG C, CHEN D F, MEI J P, et al. SemCKD: semantic calibration for cross-layer knowledge distillation [J]. IEEE transactions on knowledge and data engineering, 2022, 99: 1. DOI: 10.1109/TKDE.2022.3171571

作者简介



孙梦颖，北京邮电大学博士后；主要研究方向为面向6G的语义通信与边缘智能技术；发表论文10余篇，申请专利10余项。



熊华超，北京邮电大学在读硕士研究生；主要研究方向为智简语义通信及组网技术、6G原型系统搭建；发表论文3篇，申请专利2项。



王怡宁，北京邮电大学在读博士研究生；主要研究方向为智简语义通信及组网技术；发表论文1篇，申请专利3项。



韩书君，北京邮电大学博士后；主要研究方向为面向B5G/6G的新一代移动通信理论与技术；主持和参加项目10余项，获得1项科研成果奖；发表论文20余篇，申请专利60余项。



许晓东，北京邮电大学教授；主要研究方向为无线组网及覆盖扩展技术；作为负责人完成多项国家级项目，获北京市科学技术奖二等奖、中国通信学会科学技术奖一等奖，出版的专著获得“中国出版政府奖”；近5年发表重要期刊论文50余篇，申请专利80余项。