引用格式:张思成,张海超,史明佳,等. 基于信息瓶颈准则约束的对抗鲁棒语义通信方法[J]. 信息对抗技术,2024, 3(6):10-18. [ZHANG Sicheng, ZHANG Haichao, SHI Mingjia, et al. Adversarial robust semantic communication method based on information bottleneck criterion constraint[J]. Information Countermeasure Technology, 2024, 3(6):10-18. (in Chinese)]

基于信息瓶颈准则约束的对抗鲁棒语义通信方法

张思成,张海超,史明佳,林 云*

(哈尔滨工程大学信息与通信工程学院,黑龙江哈尔滨 150001)

摘 要 基于深度学习的语义通信旨在传递用户意图和语义信息,有望成为 6G 网络“内生智能”架构的重要技术支撑,但语义通信系统的对抗鲁棒性及其安全性尚未得到充分研究。为此,提出了一种基于信息瓶颈准则约束的对抗鲁棒语义通信方法,给出了语义通信系统模型,并从互信息理论的角度分析了系统模型发射端、信道以及接收端中语义信息的任务相关和任务无关特征。在保留原始语义相似性的前提下,加入信息瓶颈准则约束,抑制解码器中间表征的任务无关特征,从而增强语义通信模型的抗干扰能力。通过实验以及综合分析,证明了该方法在提高基于深度学习的多级语义通信系统的对抗鲁棒性方面的优越性能。

关键词 6G 内生智能技术;语义通信;对抗鲁棒性;信息瓶颈准则;语义相似性

中图分类号 TN 929.5 **文章编号** 2097-163X(2024)06-0010-09

文献标志码 A **DOI** 10.12399/j.issn.2097-163x.2024.06.002

Adversarial robust semantic communication method based on information bottleneck criterion constraint

ZHANG Sicheng, ZHANG Haichao, SHI Mingjia, LIN Yun*

(College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract Semantic communication based on deep learning aims to convey user intentions and semantic information, and is expected to become an important technical support for the “endogenous intelligence” architecture of 6G network. However, the adversarial robustness and security of semantic communication systems have not been fully studied. To this end, we proposed an adversarial robust semantic communication method based on information bottleneck criterion constraint. We firstly presented a semantic communication system model and then analyzed the task-relevant and task-irrelevant features of the semantic information in the transmitter, channel and receiver of the semantic communication system model from the perspective of mutual information theory. While preserving the original semantic similarity, we incorporated the information bottleneck criterion constraint to suppress the task-irrelevant features of the intermediate representation in the decoder, thereby enhancing the adversarial robustness of the semantic communication model. Through experiments and comprehensive analysis, we have demonstrated the superior performance of this method in improving the adversarial robustness of multi-level semantic communication systems based on deep

收稿日期:2024-07-16 **修回日期**:2024-09-29

通信作者:林云,E-mail:linyun@hrbeu.edu.cn

基金项目:国家自然科学基金资助项目(U23A20271,62201172)

learning.

Keywords　6G endogenous intelligent technology; semantic communication; adversarial robustness; information bottleneck criterion; semantic similarity

0　引言

随着深度学习技术的发展和物联网设备的大规模部署，网络提供了更多智能化的服务。与此同时，这些服务产生了大量数据，使得传统通信系统可能无法满足这种大规模数据传输的需求[1]。

现有大多数通信系统都是在香农方法和理念的指导下，以速率为中心指标进行设计的，如吞吐量、频谱效率等。传统通信系统更加关注比特级符号传输的准确性，而非准确传输真实的有效信息，这种“逐比特”的传输方式在处理复杂、高维度信息时效率较低，延迟较大，暴露出其应用在第六代移动通信（6th-generation mobile communications，6G）方面的局限性[2]。语义通信作为 6G 移动通信的潜在关键技术之一，其利用传输内容的语义信息进行编码，可以去除冗余数据，减少传输数据量，满足 6G 时代的智能通信需求，受到了学术界和产业界的广泛关注。中国国际移动通信 IMT-2030（6G）推进组和欧盟的 SONATA 计划等都将语义通信视为一种能够打破跨域通信壁垒、提高传输效率的 6G 关键技术[3-4]。与传统无线通信专注于减少传输符号错误不同，语义通信的目标是准确地提取和解释符号背后的含义[5]，因此，语义通信的优化目标是缩小发射信号和接收信号之间的语义差距。

在语义通信领域，早期的研究主要集中在语音处理和自然语言处理（natural language processing，NLP）中的语义提取，提出了通过语义层对信息进行压缩和选择的思路，例如在语音识别系统中，通过只传递语义相关信息来减少带宽占用[6]，在减少数据量的同时保证了通信的有效性。在深度学习技术的推动下，近年来的研究更加倾向于通过神经网络提取语义信息，极大提高了系统处理复杂数据的能力。WANG 等[7]提出了一种基于注意力机制的语义提取模型，通过关注语义信息的关键部分，有效提升了通信效率。XIE 等[8]针对文本信息传输提出了基于深度学习的语义通信系统，初步考虑了信源-信道联合编码，并结合迁移学习来确保该语义通信系统适用于不同的通信环境且加速模型的训练，进一步提升了模型的泛化能力。

尽管现有研究在语义通信的效率方面取得了重要进展，但在面对复杂通信环境（如噪声干扰和对抗性攻击）时，语义通信系统的鲁棒性仍然面临挑战。由于无线信道的开放性以及深度神经网络的脆弱性，语义通信系统极易遭受对抗样本的恶意攻击。对于深度神经网络，在输入中添加微小的扰动就可能会误导模型，使其输出错误结果，从而对自动驾驶、无人机操作和智能手机应用等关键任务造成重大安全问题。

在计算机视觉和机器学习等领域，对抗性训练有很多方法[9-13]，这些方法的成功应用为基于深度神经网络的语义通信系统的鲁棒性和安全性提供了重要参考。HU 等[14]提出对输入中语义噪声频繁出现的部分进行掩蔽，并结合噪声相关掩蔽策略设计了掩蔽适量量化-变分自编码器，结合带有权重扰动的对抗训练，显著提高了语义通信系统对语义噪声的鲁棒性。TANG 等[15]设计了一种能够在语义干扰存在的情况下准确地恢复语义信息的鲁棒接收器，通过交替优化干扰器和接收器，提高语义通信系统的安全性。NAN 等[16]建立了一个物理层攻击者模型，其利用物理层对抗扰动生成器产生面向语义的、可控的扰动以更好地模拟真实物理环境的效果，并通过对抗训练方法增强语义通信系统对攻击的抵御能力。PENG 等[17]利用校准的自注意力机制和对抗训练来抵抗语义噪声，对于对抗性语义噪声，通过对抗性训练方法找到主要干扰语义通信系统的扰动，并训练该系统抵抗这些扰动，提高了模型的泛化能力，使得该系统可以对抗不同形式的语义噪声并提高各种无线环境下的鲁棒性。在提高深度学习模型针对对抗样本的鲁棒性方面，还有输入去噪[18]、防御蒸馏[19]、梯度正则化[20]、权重扰动[21]等一些方法。虽然它们有效地提高了语义通信系统的鲁棒性，进一步从总体上提高了语义通信的安全性，但是并没有深入考虑语义信息的特性。语义信息可以分为 2 个部分，一部分

是与任务有关的语义信息；另一部分则是与任务无关的信息，可以在接收端被抑制掉，从而增强语义通信系统的抗噪声能力，提高语义通信的鲁棒性。

针对以上问题，本文提出一种基于信息瓶颈准则约束的对抗鲁棒语义通信方法 ARSC（adversarial robust semantic communication）。

1 系统模型

本文研究在对抗攻击场景下，基于深度学习的语义通信系统的对抗鲁棒性。含有对抗攻击的语义通信系统主要包括发送端和接收端 2 个部分。其中，发送端和接收端均采用了联合信源信道编解码的方式，充分挖掘图像中的语义信息，从而提升通信的效率和可靠性。攻击者使用对抗攻击方法，通过增加任务无关特征，意图导致接收信号的错误分类，从而实现恶意攻击的目的。根据攻击发生的位置，对抗攻击可以分为直接攻击、间接攻击和叠加攻击。其中，直接攻击为攻击者直接侵入接收器设备物理层发起的攻击。本文采用了直接攻击的方式，攻击目标是经过物理信道后的语义编码信号。

对抗鲁棒语义通信系统模型如图 1 所示。令 Y 表示语义编码信号的数据空间，T 表示语义信息对应的标签空间，它们的联合分布用 P_{YT} 表示。通过从数据和标签空间中进行成对采样得到一个域 $D=\{(y_i,t_i)\}_{i=1}^{n}\sim P_{YT}$，其中，$y_i\in Y$，$t_i\in T$ 是数据与标签的采样，n 是样本数量。在该域内完成语义信息分类任务的模型为 $f:Y\to T$。定义原始信号域 $S=\{(y_i^s,t_i^s)\}_{i=1}^{n_s}\sim P_{Y^sT^s}$，对抗样本域 $A=\{(y_i^a,t_i^a)\}_{i=1}^{n_a}\sim P_{Y^aT^a}$。对抗防御的目标是针对大多数的原始编码语义信号 X，使含有对抗干扰的语义信息尽可能接近原始信息，该目标形式化表示为：

$$\max_f \mathbf{E}_{(y,t)\in S,A} f(y+\delta)=t \tag{1}$$

式中，δ 是对抗攻击设备发送的对抗扰动。另外，本文考虑以图像数据作为传输任务，传输的数据和信道的输入均采用 RGB 格式。在传输过程中定义了压缩比（compression ratio，CR）以对图像进行压缩，压缩比为图像分辨率和通道输入序列长度之比：

$$C_{\mathrm{CR}}=\lg|\Gamma X|/\lg|\Gamma S| \tag{2}$$

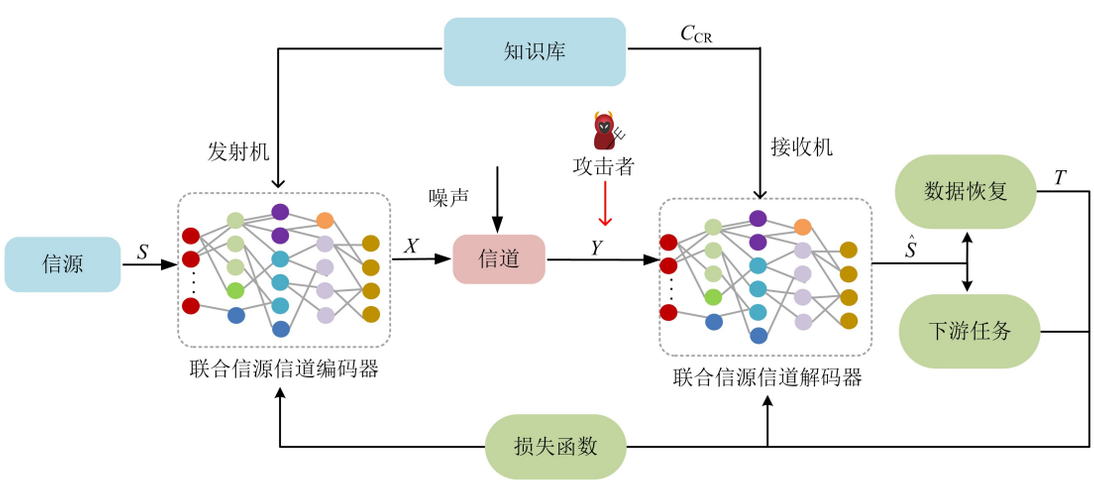

图 1 对抗鲁棒语义通信系统模型

Fig. 1 Adversarial robust semantic communication system model

2 方法描述

2.1 研究目的

在数据分析领域，从自然空间获取的样本具有丰富的内在特征，这些特征可能包含与任务相关或任务无关的成分。举例来说，在图像分类领域，与标签相关的特征被认为是任务相关特征，能保证模型的预测准确性；相反，那些与标签不相关或呈现负相关的特征则被认为是任务无关特征，可能会降低模型的泛化能力。需要注意的是，与任务无关的特征不仅有噪声，还可能包含着不同图像类别之间共享、相似或模糊的特征，这些特征可能会影响模型的有效性。本文提出的基于深度学习的语义通信系统架构可以分为联合编码器和联合解码器。联合解码器的作用之一是将输入信号 y 转换为潜在空间中的特征

向量 $\boldsymbol{z}$。建模的首要目标是微调整个模型的参数，以最大化 Z（代表 $\boldsymbol{z}$ 的随机变量）和 $\hat{S}$（信号标签 $\hat{s}$ 的随机变量）之间的相关性，从而提高分类精度。在这种情况下，互信息（mutual information，MI）作为基本的相关性度量，通过 Z 和 $\hat{S}$ 边际分布的联合和乘积之间的相对熵，即 KL 散度（Kullback-Leibler divergence）进行量化，表示为：

$$I(Z;\hat{S})=\int p(\boldsymbol{z},\hat{s})\lg\left(\frac{p(\boldsymbol{z},\hat{s})}{p(\boldsymbol{z})p(\hat{s})}\right)\mathrm{d}\boldsymbol{z}\,\mathrm{d}\hat{s} \quad (3)$$

过分依赖互信息可能会导致过拟合，因为它会利用任务相关和不相关的特征来增强 Z 和 $\hat{S}$ 之间的相关性。因此，经过精心设计的对抗性扰动可能会影响任务无关的特征，使得这些原本不会影响模型识别结果的任务无关特征变得具有“恶意攻击性”。为了应对这项挑战，本文提出了一种基于信息瓶颈准则约束的对抗鲁棒语义通信方法，旨在保持分类性能的同时，降低联合解码器对任务无关特征在 Z 中的映射。方法着重于 2 个方面：一是减少信号数据 Y（信号数据 y 的随机变量）与 Z 之间的互信息，二是增强 Z 与 $\hat{S}$ 之间的互信息，这样的双重目标与信息瓶颈准则的基本原理相契合。图 2 为强对抗和弱对抗鲁棒模型的 Venn 图比较，可以清晰地呈现这些变量之间的概念相互作用。通过这一方法，即使对数据施加不会从根本上改变数据本质的对抗性扰动，也不太可能影响 Z，从而避免导致错误分类。

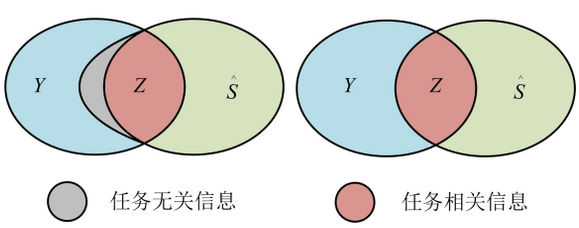

图 2　强对抗和弱对抗鲁棒模型的 Venn 图比较

Fig. 2　Comparison of Venn diagrams between strong and weak adversarial robust models

2.2　方法框架与训练过程

图 3 为所提出的信息瓶颈准则约束方法构建的框架图，该框架是在图 1 对抗鲁棒语义通信模型的基础上，把信息瓶颈准则约束方法添加到模型中对应位置。信源 S 经过联合信源信道编码器 E，得到信道输入 X，X 经过信道之后得到输出 Y，Y 作为联合信源信道解码器 D 的输入，经过信道解码和语义解码，得到解码后的信宿 $\hat{S}$，其中，信道包含高斯白噪声 $n(n\sim N(0,\sigma^2))$。令 X_{s} 表示浅层联合信源信道解码器，X_{d} 表示深层联合信源信道解码器，Z 表示中间特征。

本文的研究目标是增强语义通信系统解码器的对抗鲁棒性，这里具有 3 个隐含条件：1）输入信源 S 和解码后的信宿 $\hat{S}$ 应该尽可能在语义层面上相似；2）加入对抗扰动的 Y 恢复出来的数据与未加入对抗扰动的恢复出来的图像数据在语义层面上也应该保持一致；3）增强解码器的对抗鲁棒性。

对于第 1 个目标，等价于优化参数 θ_1 和 θ_2，将损失函数表示为 $L(S,\hat{S})$，利用 Adam 算法把参数更新为：

$$\theta_{i,t+1}=\theta_{i,t}-\eta\frac{\rho_t}{\sqrt{\nu_t+\varepsilon}} \quad (4)$$

式中，ρ_t 和 ν_t 分别是梯度的一阶和二阶动量，ε 是防止第二项的分母为 0 的平滑项，η 是学习率。所以，损失函数可以表示如下：

$$L_1=L_{\theta_1,\theta_2}(S,\hat{S})=L_{\mathrm{CE}}(S,\hat{S}) \quad (5)$$

式中，L_{CE} 为采用交叉熵损失函数。

对于第 2 个目标，定义没有添加对抗扰动的信号经过解码器 D 之后的输出为 $D(Y)$，添加了对抗扰动的信号经过解码器 D 之后输出为 $D(Y+\delta)$。本文的目标是最小化语义相似度变化幅度，即：

$$L_2=L_{\mathrm{CE}}(D(Y),D(Y+\delta)) \quad (6)$$

如 2.1 中的解释，对于第 3 个目标，本文通过利用信息瓶颈准则对解码器的中间表征 Z 进行约束。在计算 2 个随机变量的互信息时需要首先对它们的概率密度进行估计，而这是相当困难的。目前有许多致力于互信息估计的研究[22-23]。受文献[24]的启发，与互信息类似，希尔伯特-施密特独立性准则（Hilbert-Schmidt independence criterion，HSIC）是一种衡量 2 个随机变量独立性的统计方法，其理论基础是希尔伯特-施密特空间。当 HSIC 接近于 0 时，可能暗示着 2 个变量之间的独立性较高；反之，当 HSIC 值远离 0 时，则可能表明这 2 个变量之间存在一定程度的相关性。与互信息不同，HSIC 不需要对 2 个变量的

概率密度进行估计，而是通过直接采样计算得出。当样本数量足够多时，$H_{\mathrm{HSIC}}(M;N)$ 表示如下：

$$\begin{aligned}H_{\mathrm{HSIC}}(M;N)&=\mathbf{E}_{mnm'n'}\left[k_m(m,m')k_{n'}(n,n')\right]\\&+\mathbf{E}_{mm'}\left[k_m(m,m')\right]\mathbf{E}_{nn'}\left[k_n(n,n')\right]\\&-2\mathbf{E}_{mn}\left\{\mathbf{E}_{m'}\left[k_m(m,m')\right]\mathbf{E}_{n'}\left[k_n(n,n')\right]\right\}\end{aligned}\tag{7}$$

式中，m 和 n 分别是 M 和 N 中的随机样本，k_m 和 k_n 是核。根据上述 HSIC 准则，$H(Y;Z)$ 表示接收信号 Y 和中间特征 Z 的互信息，$H(Z;T)$ 表示中间特征 Z 和真实标签 T 的互信息。本文的目标是增强解码器的对抗鲁棒性，即通过减少接收信号 Y 和中间特征 Z 的互信息，增大中间特征 Z 和真实标签 T 的互信息，可以表示为：

$$L_3=H_{\mathrm{HSIC}}(Z;T)-\lambda H_{\mathrm{HSIC}}(Y;Z)\tag{8}$$

式中，λ 是拉格朗日乘子，用于调节 2 个互信息约束的相对权重。

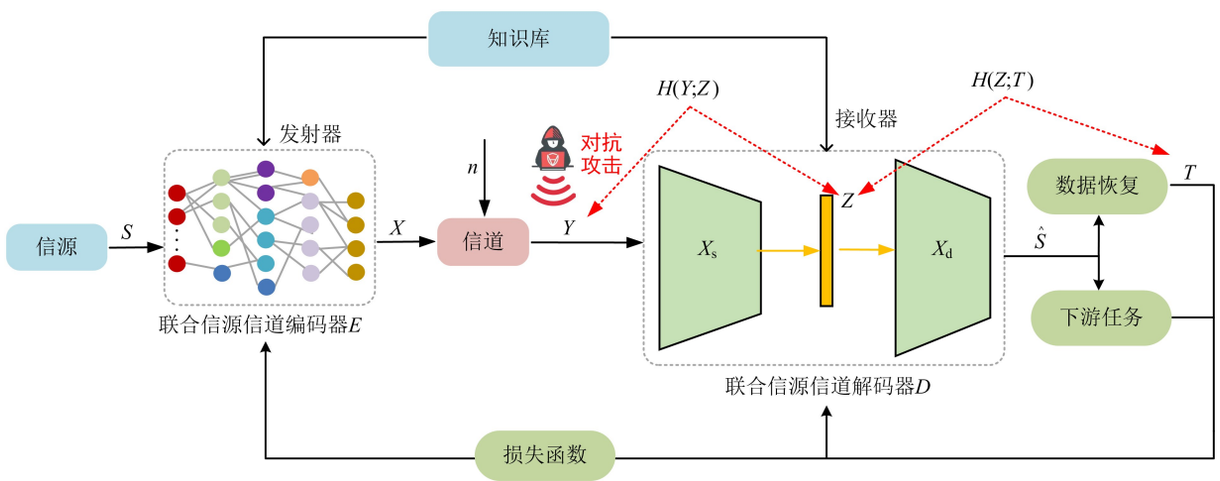

图 3 具有信息瓶颈准则约束的解码器框图

Fig. 3 Decoder block diagram constrained by information bottleneck criterion

最后，联合以上约束，构成本文中提出的基于信息瓶颈准则约束的对抗鲁棒语义通信方法。联合约束的表达式如下：

$$L_{\mathrm{total}}=L_1+\alpha L_2+\beta L_3\tag{9}$$

式中，α 和 β 为调节因子，在优化模型训练效果的同时，也用来权衡 3 个优化目标之间的比例。

2.3 方法流程

对抗鲁棒语义通信方法的核心思想是通过语义信息传输提升通信效率，同时增强系统在对抗性攻击下的鲁棒性。该方法的流程如图 4 所示，可以从以下几个关键步骤来详细描述：信源首先经过联合信源信道编码器进行语义提取，接着在接收端对通过信道后的信号添加对抗样本，然后对添加对抗样本的信号进行对抗训练，最后经过信息瓶颈准则约束保留任务相关特征，抑制无关特征，从而增强系统的对抗鲁棒性。

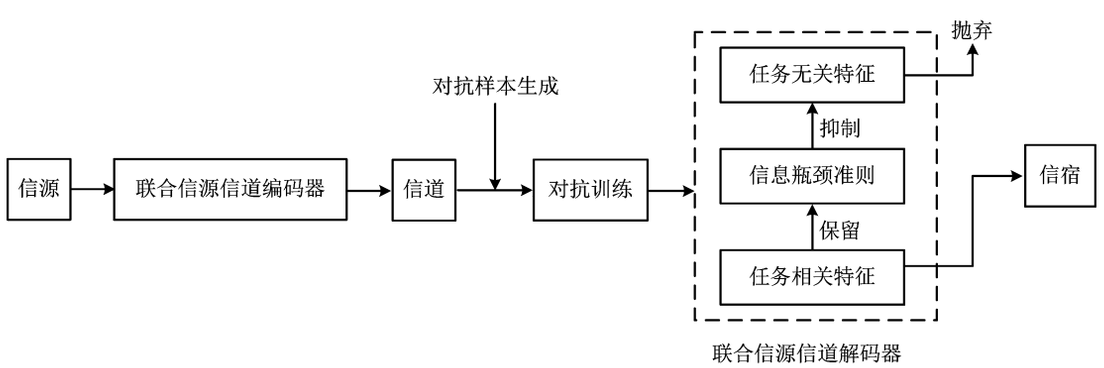

图 4 对抗鲁棒语义通信方法流程

Fig. 4 Process of adversarial robust semantic communication method

在语义通信中，传输的不是逐比特的原始数据，而是提取出来的语义信息。该阶段的主要任务是从原始数据中提取出最能代表数据内容和意义的语义特征。具体方法是先将原始输入数据通过联合信源信道编码器进行特征提取，通过网络的中间层提取语义向量，用于表示数据的高维语义信息，其中联合信源信道编码器为深度神经网络模型。在提取到语义信息后，接下来进行语义编码，这一步同样是通过该编码器模型对提取的语义特征进行压缩编码。编码后的语义特征向量具备一定的鲁棒性，但这对于稳健鲁棒的语义通信系统来说是远远不够的。在通过信道后，到达联合信源信道解码器，即接收端。在接收端含有对抗样本的语义向量经过信道解码后，恢复出含对抗样本的语义信息，通过信息瓶颈约束准则抑制任务无关特征，保留任务相关特征，进而通过信源解码，到达信宿。通过该方法，能够有效地提高语义通信系统的抗噪声能力，并且可以在面对不同攻击方式时能够保持良好的防御效果。同时，该方法通过降低传输数据量、提升传输效率，确保了系统在带宽受限或计算资源有限时依然能够高效运行。这些特点使得该方法在智能场景应用中具有广泛的应用前景。

2.4　测试过程

在测试阶段，严格遵循标准的测试流程，以确保模型的性能能够在真实环境中得到准确评估。测试过程包括将模型未知的测试数据输入训练完成的语义通信模型，通过前向传播计算输出，并根据标准评价指标（如准确率、峰值信噪比）对模型性能进行量化评估。

需要特别指出的是，本文提出的信息瓶颈准则方法仅在训练阶段引入了额外的约束项，用于优化模型的泛化能力，而它们在测试阶段并不参与计算。因此，该方法不会对测试阶段的模型推理造成额外的计算或存储负担。这一设计确保了该方法既能提升模型的训练效果，又能在测试过程中保持与常规模型相同的效率。这种分阶段的优化策略，兼顾了模型性能和推理效率，在实际应用中具有重要的实践价值。

3　实验及分析

3.1　实验方案

实验考虑将图像数据作为验证数据，使用的数据集为 CIFAR10（共有 60 000 张图片，每个类别有 6 000 张图像），在该数据集中有 50 000 张用于训练，10 000 张用于测试，每张图片的尺寸为 32×32。在图像传输过程中，利用深度神经网络提取图像中的高层语义特征。这些特征不仅包含视觉表象（如边缘、颜色等），还包括图像的核心语义信息（如物体类别、场景内容等）。通过这种方式，语义通信系统将冗余的像素级信息过滤掉，仅保留与当前传输任务密切相关的高维语义向量。这些向量包含了原始图像的核心语义，可以有效减少数据量并提升传输效率。在性能仿真中，使用为下游任务的数字识别的准确率和图像重建的峰值信噪比（peak signal-to-noise ratio，PSNR）作为性能指标。本文使用在相同条件下的 JPEG2000 的传统通信框架、基于变分自编码器（variational auto encoder，VAE）的语义编码和基于深度学习的深度联合信源信道编码（deep joint source-channel coding，Deep JSCC）语义通信算法作为对比方案。

3.2　实验结果分析

图 5～6 分别给出了 ARSC 和 3 种对比方案在不同压缩率下的识别准确率和峰值信噪比性能。从图 5 中可以看出，当 $C_{CR}<0.3$ 时，ARSC 在所有考虑的方案中具有最高的精度，并且 VAE 方法在所有不同的 C_{CR} 下表现最差。当 $C_{CR}\geqslant 0.3$ 时，所提系统与基于 JEPG2000 的方法之间存在微小的差异。特别是当 C_{CR} 在[0.3，0.5]范围内时，该方法的识别精度较好；而当 $C_{CR}\in[0.5,0.7]$ 时，基于 JPEG2000 方法的识别精度稍好。且 ARSC 在 $C_{CR}=0.7$ 时可以达到 88.52%的识别准确率，$C_{CR}=0.2$ 时可以达到 84.70%的识别准确率，仅损失不到 4%，在可接受范围内。从图 6 中可以看出，本文提出的 ARSC 在信噪比为 10 dB、$C_{CR}=0.2$ 时，PSNR 分别比 JPEG2000 平均高 7.60 dB，比 Deep JSCC 平均高 4.60 dB，比 VAE 平均高 11.60 dB；本文提出的 ARSC 在信噪比为 10 dB、$C_{CR}=0.3$ 时，PSNR 分别比 JPEG2000 平均高 4.45 dB，比

Deep JSCC 平均高 2.45 dB，比 VAE 平均高 13.45 dB。可以看出，相比于传统的压缩算法和基于深度设计的图像传输系统，所提出的 ARSC 传输性能有了显著的提升。综合分析可知，ARSC 可以有效地在低资源传输信息，符合语义通信的目的，因此本文在低压缩比（$C_{CR}=0.2$）的情况下进行对抗鲁棒性实验。

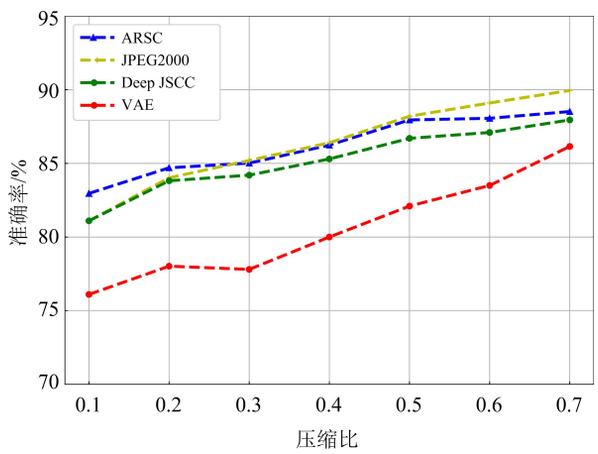

图 5 不同方法在识别准确率评估标准下的性能

Fig. 5 Performance of different methods under recognition accuracy evaluation criterion

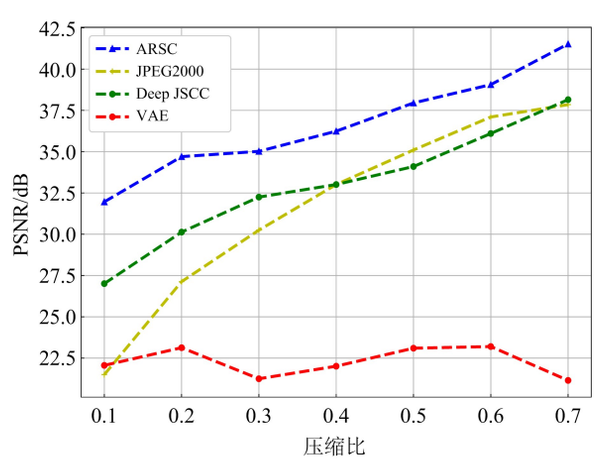

图 6 不同方法在 PSNR 评估标准下的性能

Fig. 6 Performance of different methods under PSNR evaluation criterion

为证明本文所提方法的有效性，图 7 展示出了 ARSC 和对比方案在 3 种对抗样本生成方法下识别准确率的变化情况。该实验使用了单步攻击方法——快速梯度符号法（fast gradient sign method，FGSM）和 2 种迭代攻击方法——基本迭代法（basic iterative method，BIM）和投影梯度下降法（projected gradient descent，PGD），并通过在不同扰动强度下的识别准确率来评估它们的效果。由图 7 可知，随着扰动强度的增加，识别准确率均有不同程度的下降，且在任意扰动强度下，迭代攻击方法的攻击效果均优于单步攻击法 FGSM 的攻击效果。这是因为迭代方法通过多次迭代可以充分利用模型的梯度信息，根据识别模型的反馈逐渐优化攻击，从而使生成的对抗样本更具有针对性。但在相同攻击强度变化下，所提出的 ARSC 的识别准确率下降较少，对抗鲁棒性更强。例如在 FGSM 攻击下，当攻击强度增加到 0.05 时，ARSC 识别准确率从 84.7%下降至 64.0%，下降了 20.7%，而对比模型的识别准确率从 82.8%下降至 56.2%，下降了 26.6%。证明了本文所提出的 ARSC 可根据压缩后的语义特征，优先还原有用语义特征，在面对恶意攻击时具有一定的对抗鲁棒性，在图像语义重建任务中可以取得更好的效果。

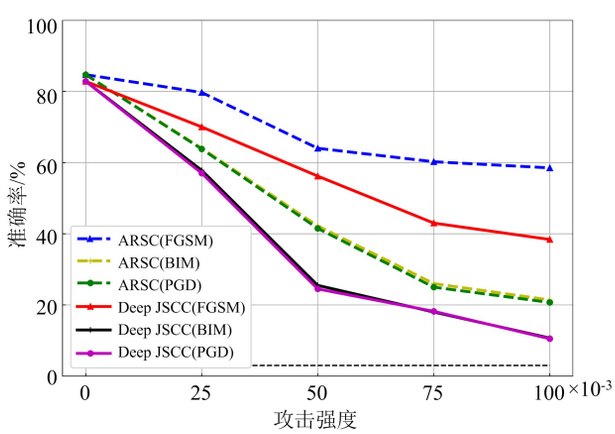

图 7 不同扰动强度下的识别准确率

Fig. 7 Recognition accuracy performance under different disturbance intensities

4 结束语

针对基于泛化能力有限的深度神经网络的语义通信系统在进行信息传输时易遭受恶意攻击进而导致模型错误分类的问题，本文提出了一种基于信息瓶颈准则约束的对抗鲁棒语义通信方法。该方法将互信息和语义通信系统深度融合，有效地增强了语义通信系统模型的对抗鲁棒性，不仅能够为 6G 内生智能网络提供技术支撑，还能为其通信安全保驾护航。

参 考 文 献

[1] MOHAMMADI M, AL-FUQAHA A, SOROUR S, et al. Deep learning for IoT big data and streaming analytics: a survey[J]. IEEE Communications Surveys &Tutorials, 2018, 20(4): 2923-2960.

[2] LAN Q, WEN D Z, ZHANG Z 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.

[3] LETAIEF K B, CHEN W, SHI Y M, et al. The roadmap to 6G: AI empowered wireless networks[J]. IEEE Communications Magazine, 2019(8): 84-90.

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

[6] ERDOGAN H, SARIKAYA R, CHEN S F, et al. Using semantic analysis to improve speech recognition performance[J]. Computer Speech&Language, 2005, 19(3): 321-343.

[7] WANG Y N, CHEN M Z, LUO T, et al. Performance optimization for semantic communications: an attention-based reinforcement learning approach[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(9): 2598-2613.

[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.

[9] MOOSAVI-DEZFOOLI S-M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2017: 1765-1773.

[10] WONG E, KOLTER J Z. Learning perturbation sets for robust machine learning[EB/OL]. (2020-10-08)[2024-09-28]. https://arxiv.org/abs/2007.08450.

[11] MAINI P, WONG E, KOLTER Z. Adversarial robustness against the union of multiple perturbation models[C]//Proceedings of the 37th International Conference on Machine Learning. [S. l.: s. n.], 2020: 6640-6650.

[12] MADAAN D, SHIN J, HWANG S J. Learning to generate noise for multi-attack robustness[C]//Proceedings of the 38th International Conference on Machine Learning. [S. l.: s. n.], 2021: 7279-7289.

[13] LEINO K, WANG Z F, FREDRIKSON M. Globally-robust neural networks[C]//Proceedings of the 38th International Conference on Machine Learning. [S. l.: s. n.], 2021: 6212-6222.

[14] HU Q Q, ZHANG G Y, QIN Z J, et al. Robust semantic communications with masked VQ-VAE enabled codebook[J]. IEEE Transactions on Wireless Communications, 2023, 22(12): 8707-8722.

[15] TANG R, GAO D H, YANG M X, et al. GAN-inspired intelligent jamming and anti-jamming strategy for semantic communication systems[C]//Proceedings of 2023 IEEE International Conference on Communications Workshops. Rome: IEEE, 2023: 1623-1628.

[16] NAN G S, LI Z C, ZHAI J L, et al. Physical-layer adversarial robustness for deep learning-based semantic communications[J]. IEEE Journal on Selected Areas in Communications, 2023, 41(8): 2592-2608.

[17] PENG X, QIN Z J, HUANG D L, et al. A robust deep learning enabled semantic communication system for tex[C]//Proceedings of 2022 IEEE Global Communications Conference. [S. l.]: IEEE, 2022: 2704-2709.

[18] LIAO F Z, LIANG M, DONG Y P, et al. Defense against adversarial attacks using high-level representation guided denoise[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2018: 1778-1787.

[19] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of 2016 IEEE Symposium on Security Privacy. [S. l.]: IEEE, 2016: 582-597.

[20] GU S X, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[EB/OL]. (2015-04-09)[2024-09-28]. https://arxiv.org/abs/1412.5068.

[21] WU D X, XIA S T, WANG Y S. Adversarial weight perturbation helps robust generalization[J]. Advances in Neural Information Processing Systems, 2020, 33: 2958-2969.

[22] BELGHAZI M I, BARATIN A, RAJESHWAR S, et al. Mutual information neural estimation[C]//Proceedings of the 35th International Conference on

Machine Learning. [S. l. : s. n.], 2018: 531-540.

[23] GAO W H, KANNAN S, OH S, et al. Estimating mutual information for discrete-continuous mixtures [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. [S. l. : s. n.], 2017: 5988-5999.

[24] MA W-D K, LEWIS J P, KLEIJN W B. The HSIC bottleneck: deep learning without back-propagation [C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. [S. l. : s. n.], 2020: 5085-5092.

作 者 简 介

张思成

男，1996 年生，博士研究生，研究方向为智能频谱感知技术、人工智能与模式识别、智能感知模型的对抗样本攻击与鲁棒防御

E-mail: 2015080325@hrbeu. edu. cn

张海超

男，2000 年生，硕士研究生，研究方向为通信技术、信号处理、语义通信的对抗攻防

E-mail: zhc418694586@hrbeu. edu. cn

史明佳

女，1999 年生，硕士研究生，研究方向为语义通信技术、信号处理和安全分析、电磁空间中人工智能模型的对抗安全问题

E-mail: shimingjia@hrbeu. edu. cn

林　云

男，1980 年生，教授，博士研究生导师，研究方向为智能无线电技术、人工智能与机器学习、大数据分析与挖掘、软件和认知无线电、信息安全与对抗、智能信息处理

E-mail: linyun@hrbeu. edu. cn

责任编辑　董　莉