



基于自注意力机制的大规模 MIMO 信道 状态信息特征向量反馈方法

杨蓓¹, 梁鑫^{1,2}, 尹航¹, 蒋峥¹, 余小明¹

(1. 中国电信股份有限公司研究院, 北京 102209; 2. 北京邮电大学, 北京 100876)

摘要: 大规模多输入多输出 (MIMO) 系统可以为 5G 和未来无线通信系统提供令人满意的频谱效率的增益。在频分双工 (FDD) 模式下, 需要将下行信道状态信息 (CSI) 的特征向量精确地反馈到基站侧以获得这种增益。为了提升下行 CSI 特征向量的反馈精度, 提出了一种基于自注意力机制的 CSI 反馈方法 SA-CsiNet。SA-CsiNet 通过分别在编、解码器部署自注意力模块实现 CSI 的特征提取和重构。仿真实验结果表明, 相较于码本和传统的深度学习 CSI 反馈方案而言, SA-CsiNet 能够提供更高的 CSI 重建精度。

关键词: 信道反馈; 大规模 MIMO; 自注意力机制; 深度学习

中图分类号: TN929.5

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2023247

Self-attention mechanism-based CSI eigenvector feedback for massive MIMO

YANG Bei¹, LIANG Xin^{1,2}, YIN Hang¹, JIANG Zheng¹, SHE Xiaoming¹

1. Research Institute of China Telecom CO., Ltd., Beijing 102209, China

2. Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: Massive multiple-input multiple-output (MIMO) system can provide satisfying gain of spectrum efficiency for 5G and future wireless communication systems. In frequency-division duplex (FDD) mode, downlink channel state information (CSI) needs to be accurately fed back to the base station side to obtain this gain. To improve the feedback accuracy of downlink CSI eigenvector, a self-attention mechanism-based CSI feedback method named SA-CsiNet was proposed. SA-CsiNet respectively deployed self-attention modules at the encoder and the decoder to achieve feature extraction and reconstruction of CSI. Experimental results show that compared with codebook-based and conventional deep learning-based CSI feedback approaches, SA-CsiNet provides higher reconstruction accuracy of CSI.

Key words: channel feedback, massive MIMO, self-attention mechanism, deep learning

0 引言

大规模多输入多输出 (multiple-input multiple-output, MIMO) 已经成为 5G 及未来无线通信

的关键技术之一^[1]。在基站侧部署大规模天线阵列, 可以获得显著的下行吞吐量增益^[2]。而这种增益的实现往往需要基站获知准确的下行信道状态信息 (channel state information, CSI) 以进行下

行预编码^[3]。在频分双工（frequency-division duplex, FDD）模式下，由于上、下行工作在互不交叠的频段，上、下行 CSI 不具有显式的互易性，因此需要将下行 CSI 反馈到基站侧^[4]。

CSI 反馈方法大致分为 3 类，即基于码本的 CSI 反馈^[5]、基于压缩感知的 CSI 反馈^[6]和基于人工智能（artificial intelligence, AI）的 CSI 反馈^[7]。基于码本的 CSI 反馈方案在基站和用户侧部署预先设计好的预编码码本，然后根据当前 CSI 选取合适的码本，以预编码矩阵标识符（precoding matrix index, PMI）的形式进行反馈。在大规模 MIMO 场景下，基于码本的 CSI 反馈开销随着天线数目的增加而增加，消耗过多的物理资源。基于压缩感知的 CSI 反馈算法中的稀疏性假设和随机投影技术影响了 CSI 反馈的精度，并且压缩感知算法往往是迭代的，对于时间敏感的 CSI 反馈任务而言可能会影响其实时性。基于 AI 的 CSI 反馈具有反馈开销低、模型推理时间短等优势，已经成为一种具有前景的方案。目前，国际标准化组织 3GPP 也已经启动了智能 CSI 反馈增强技术的研究工作^[8]。文献[9]中首次提出大规模 MIMO 下基于 AI 的 CSI 反馈网络 CsiNet，通过自编码器 AI 架构实现 CSI 的压缩、反馈和重建。文献[10]提出了一种 CSI 反馈的 AI 模型架构 CRNet，相较于 CsiNet，CRNet 拥有更多的神经网络层数，且设计了多分支的网络结构，以提升 AI 模型在 CSI 反馈任务中的精度。文献[11-12]通过 AI 利用信道中的隐式相关性提升 CSI 的反馈精度。文献[13]将量化模块引入基于 AI 的 CSI 反馈任务中，降低了 CSI 反馈开销。文献[14]对基于 AI 的 CSI 反馈中可能出现的量化错误进行建模，并设计了一种纠错模块对错误进行纠正，提高了基于 AI 的 CSI 反馈的鲁棒性。

最近，自然语言处理领域的 Transformer 架构^[15]通过引入注意力机制提升 AI 的特征提取能力，从而提升 AI 在机器翻译、文本生成等任务中的性能。考虑 Transformer 中自注意力模型展现出的优

越性能，本文将自注意力模型引入 CSI 反馈问题中，提出了一种大规模 MIMO 下基于自注意力机制的 CSI 特征向量反馈方案。

1 系统模型

考虑一个单小区大规模 MIMO 正交频分复用（orthogonal frequency division multiplexing, OFDM）系统，基站侧部署 $N_t \gg 1$ 根发送天线，用户侧部署 $N_r > 1$ 根接收天线。OFDM 系统的传输流数为 1，子带数目为 N_s ，则第 i 个子带上的下行接收信号表示为：

$$\mathbf{y}_i = \mathbf{h}_i^H \mathbf{v}_i \mathbf{x}_i + \mathbf{n}_i \quad (1)$$

其中， $\mathbf{h}_i \in \mathbb{C}^{N_r \times N_t}$ 表示第 i 个子带的下行 CSI 向量， $\mathbf{v}_i \in \mathbb{C}^{N_t \times N_t}$ 表示预编码向量， $\mathbf{x}_i \in \mathbb{C}^{N_r \times 1}$ 和 $\mathbf{n}_i \in \mathbb{C}^{N_r \times 1}$ 分别表示下行发送符号和加性噪声， $(\cdot)^H$ 表示共轭运算。假设该 OFDM 系统具有理想的信道估计，则第 i 个子带的下行预编码向量为 $\mathbf{h}_i$ 的特征向量 $\mathbf{w}_i$ 可通过式（2）进行计算：

$$\mathbf{h}_i \mathbf{h}_i^H \mathbf{w}_i = \lambda_i \mathbf{w}_i \quad (2)$$

其中， λ_i 是 $\mathbf{h}_i \mathbf{h}_i^H$ 的最大特征值， $\|\mathbf{w}_i\|_2 = 1$ ， $\|\cdot\|_2$ 表示 2-范数。考虑 N_s 个子带的预编码矩阵为各个子带预编码向量的堆叠：

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{N_s}]^H \in \mathbb{C}^{N_s \times N_t} \quad (3)$$

需要将 $\mathbf{W}$ 从用户侧压缩反馈至基站侧。

本文采用 AI 中的自编码器架构实现 CSI 特征向量反馈。用于 CSI 特征向量反馈的 AI 自编码器结构如图 1 所示，自编码器通过编码器 $f_{\text{enc}}(\cdot)$ 进行 CSI 特征向量的特征提取，并将其压缩为一个 M 维码字向量：

$$\mathbf{c} = f_{\text{enc}}(\mathbf{W}) \quad (4)$$



图 1 用于 CSI 特征向量反馈的 AI 自编码器结构



然后，部署在用户侧的量化模块 $f_{\text{quan}}(\cdot)$ 将 $\mathbf{c}$ 量化后以比特 (bit) 流的形式反馈至基站侧，基站侧部署的去量化模块 $f_{\text{de-quant}}(\cdot)$ 将比特流恢复为量化后的码字：

$$\hat{\mathbf{c}} = f_{\text{de-quant}}(f_{\text{quan}}(\mathbf{c})) \quad (5)$$

最后，基站通过解码器 $f_{\text{dec}}(\cdot)$ 重建 CSI 特征向量：

$$\hat{\mathbf{W}} = f_{\text{dec}}(\hat{\mathbf{c}}) \quad (6)$$

其中， $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \dots, \hat{\mathbf{w}}_{N_s}]^H \in \mathbb{C}^{N_s \times N_t}$ 为重建的 CSI 特征向量。

在本文中，余弦相似度作为 AI 自编码器 CSI 特征向量反馈的目标函数，即：

$$\rho = \frac{1}{KN_s} \sum_{k=1}^K \sum_{i=1}^{N_s} \frac{\|\mathbf{w}_i^H \hat{\mathbf{w}}_i\|_2}{\|\mathbf{w}_i\|_2 \|\hat{\mathbf{w}}_i\|_2} \quad (7)$$

其中， K 为 AI 训练数据的样本数量。

2 SA-CsiNet

2.1 网络结构

不同于传统基于 AI 的 CSI 反馈网络架构进行特征提取时所使用的卷积神经网络 (convolutional neural network, CNN) [16]，本文使用自注意力机制进行 CSI 的特征提取。自注意力模块相较于 CNN 而言，能够更专注于整个输入中不同部分的相关性，从而提升 AI 的性能。基于自注意力机制的 CSI 特征向量反馈 AI 模型 SA-CsiNet 如图 2 所示，

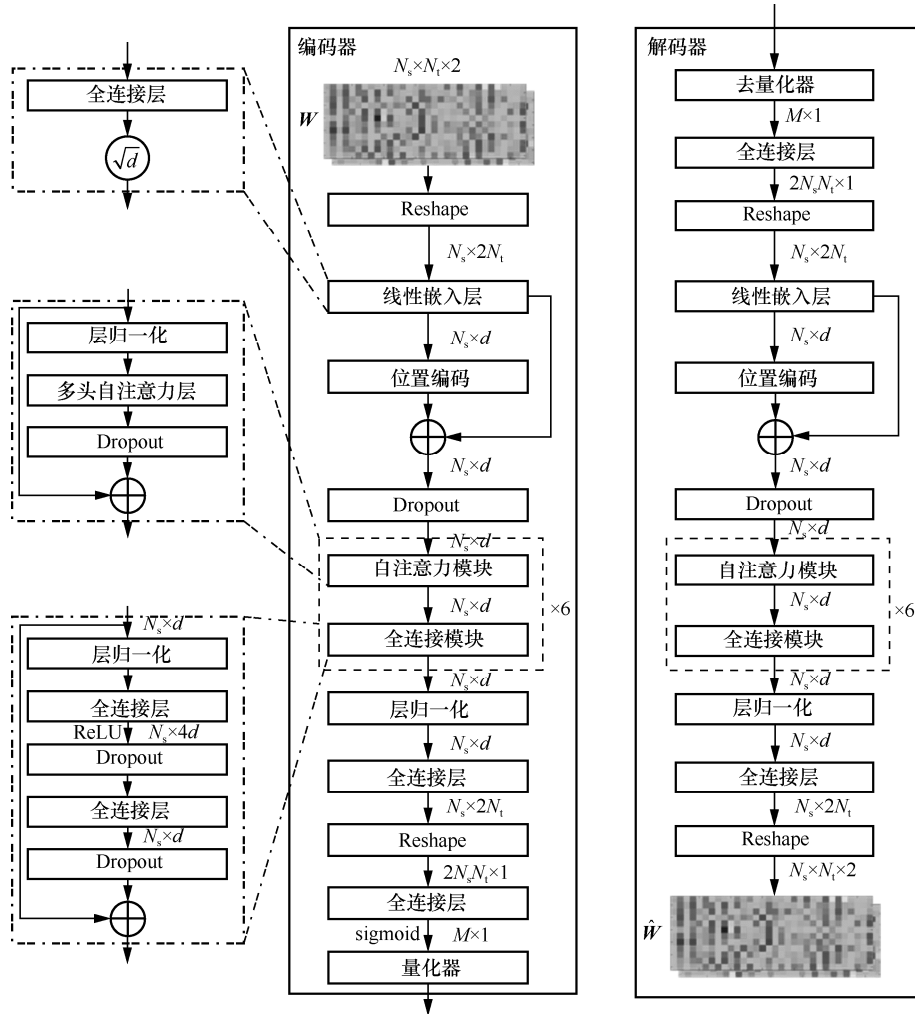


图 2 基于自注意力机制的 CSI 特征向量反馈 AI 模型 SA-CsiNet

在编码器一侧, 复数 CSI 矩阵首先被分为实部和虚部两部分, 其尺寸为 $N_s \times N_t \times 2$, 通过一个 Reshape 层调整尺寸为 $N_s \times 2N_t$, 输出张量通过线性嵌入层后进行位置编码以强调 CSI 的位置信息。其中, 线性嵌入层由一个全连接层 $f_{FC}(\cdot)$ 和一个缩放系数组成:

$$\mathbf{O}^i = \sqrt{d} f_{FC}(\mathbf{I}^i), i=1, 2, \dots, N_s \quad (8)$$

其中, $\mathbf{I}^i \in \mathbb{R}^{1 \times 2N_t}$ 和 $\mathbf{O}^i \in \mathbb{R}^{1 \times d}$ 分别是线性嵌入层的输入和输出张量的第 i 行。位置编码满足:

$$P_j^i = \begin{cases} \sin \frac{i}{10000^{\frac{2[j/2]}{d}}}, i \text{ 为偶数} \\ \cos \frac{i}{10000^{\frac{2[j/2]}{d}}}, i \text{ 为奇数} \end{cases} \quad (9)$$

$$i=1, 2, \dots, N_s; j=1, 2, \dots, d$$

其中, P 表示输出张量的位置编码结果。输出张量的值加上自身的位置编码结果后经过一个 Dropout 层随机失活, 以提高神经网络训练的鲁棒性, 然后将输出结果送入自注意力模块。本文所使用的自注意力模块通过多头自注意力模型实现, 具体而言, 若自注意力模块中有 n 个头(head), 则第 i 个头的输出为:

$$\mathbf{h}_i = \text{softmax} \left(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d/n}} \right) \mathbf{V}_i, i=1, 2, \dots, n \quad (10)$$

其中, d/n 为整数, $\text{softmax}(\mathbf{x}) = \frac{e^{x_j}}{\sum_j e^{x_j}}$,

$$\pm \mathbf{x} = [x_1, x_2, \dots, x_{d/n}],$$

$$\begin{aligned} \mathbf{Q}_i &= \mathbf{q} \mathcal{W}_i^{(q)} \\ \mathbf{K}_i &= \mathbf{k} \mathcal{W}_i^{(k)} \\ \mathbf{V}_i &= \mathbf{v} \mathcal{W}_i^{(v)} \end{aligned} \quad (11)$$

其中, $\mathcal{W}_i \in \mathbb{R}^{d \times d/n}$ 为可训练的权重, $\mathbf{q} \in \mathbb{R}^{N_s \times d}$ 、 $\mathbf{k} \in \mathbb{R}^{N_s \times d}$ 、 $\mathbf{v} \in \mathbb{R}^{N_s \times d}$ 均相等且等于自注意力模块的输入。 n 个头的输出 $\mathbf{h}_i \in \mathbb{R}^{N_s \times d/n}$ 被拼接后乘以一个新的可训练权重成为多头自注意力模型的输出:

$$\mathbf{O} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n] \mathcal{W}^{\text{out}} \quad (12)$$

其中, $\mathcal{W}^{\text{out}} \in \mathbb{R}^{d \times d}$ 。之后, 编码器使用一个全连接层将自注意力模块的输出尺寸变为 $N_s \times 2N_t$, 然后将其展平为尺寸为 $2N_s N_t \times 1$ 的一维向量, 再经过一个全连接层将其尺寸压缩为 $M \times 1$, 最后使用均匀量化器将 M 维的码字向量离散化, 以比特流的形式进行反馈:

$$f_{\text{quan}}(\mathbf{x}) = \text{round}(2^B \mathbf{x} - 0.5) \quad (13)$$

其中, B 为码字中每个元素的量化比特数, 则总反馈开销为 $O = M \times B$ 比特。

解码器端首先将接收到的比特流还原为 M 维码字向量:

$$f_{\text{de-quant}}(\mathbf{x}) = \frac{f_{\text{quan}}(\mathbf{x}) + 0.5}{2^B} \quad (14)$$

然后使用一个全连接层将其尺寸扩张为 $2N_s N_t \times 1$, 通过一个 Reshape 层调整其尺寸为 $N_s \times 2N_t$ 后送入线性嵌入层。输出张量经过位置编码后送入 Dropout 层进行随机失活, 之后进入一系列自注意力模块进行特征还原, 最后经过一个全连接层还原其尺寸为 $N_s \times N_t \times 2$ 并进行输出。

2.2 训练方法

本文采用端到端的训练方式对 SA-CsiNet 模型进行训练。SA-CsiNet 的输入是 CSI 的特征向量矩阵 $\mathbf{W}$, SA-CsiNet 的输出目标为 $\mathbf{W}$ 本身, 考虑存在误差, 根据式 (4) ~ 式 (6), 其真实输出 $\hat{\mathbf{W}} = f_{\text{dec}}(f_{\text{de-quant}}(f_{\text{quan}}(f_{\text{enc}}(\mathbf{W}))))$ 。误差通过式 (15) 所定义的余弦相似度损失函数来计算:

$$\text{LOSS} = 1 - \frac{1}{KN_s} \sum_{k=1}^K \sum_{i=1}^{N_s} \frac{\|\mathbf{w}_i^H \hat{\mathbf{w}}_i\|_2}{\|\mathbf{w}_i\|_2 \|\hat{\mathbf{w}}_i\|_2} \quad (15)$$

其中, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{N_s}]^H$, $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \dots, \hat{\mathbf{w}}_{N_s}]^H$ 。AI 参数采用 Adam 优化器^[17]进行更新。值得注意的是, 不同于大部分文献使用二维离散



傅里叶变换 (discrete Fourier transform, DFT) 将 CSI 矩阵从空间—频率域变换到角度—时延域获得稀疏性, 并进行稀疏剪裁后输入 AI 模型的处理方式^[9-12], 本文考虑稀疏剪裁会丢失一部分 CSI 的原始信息从而影响性能, 因此不对 CSI 矩阵进行剪裁。又由于 DFT 是线性变换, 因此本文中的 AI 模型仍在空间—频率域对 CSI 矩阵进行压缩反馈。

考虑训练 AI 模型时参数以反向传播的方式进行更新, 因此要求 AI 模型各层在计算时存在梯度。根据式 (13) 和式 (14) 可知, 均匀量化器的梯度并不是处处存在的, 因此需要对本文使用的均匀量化器的梯度进行定义。合格梯度 (passing gradient)^[18] 是一种在训练 AI 时常用的量化器梯度定义方法, 并已经在一些基于 AI 的 CSI 反馈方案的量化任务中取得良好效果^[19]。本文利用合格梯度方法定义均匀量化器的在反向传播时的梯度:

$$\begin{aligned} \frac{df_{\text{quan}}(\mathbf{x})}{d\mathbf{x}} &= 1 \\ \frac{df_{\text{de-quant}}(\mathbf{x})}{d\mathbf{x}} &= 1 \end{aligned} \quad (16)$$

3 仿真实验

3.1 实验设置

为了评估所提出的基于自注意力机制的 CSI 特征向量反馈 AI 模型的性能, 本文采用系统级仿真平台生成用于 AI 模型的训练和验证数据样本。具体仿真参数设定见表 1, 场景设定为符合 3GPP TS 38.901 标准的密集城区 (dense urban), 无线信道工作中心频率为 4.0 GHz, 带宽为 10 MHz, 子载波间隔为 15 kHz, 资源块 (resource block, RB) 数目为 48 个, 每 4 个 RB 划分为一个子带, 因此子带数 $N_s = 12$ 。基站侧部署 $N_t = 32$ 根发送天线, 用户侧部署 $N_r = 4$ 根接收天线, 天线以线性均匀阵列方式排布, 系统 Rank 数设为 1, 因此传输流数也

为 1。用户以均匀随机的方式撒点于以基站为中心的半径为 200 m 的圆内, 仿真假设系统具有理想的信道估计。

表 1 仿真参数设定

参数	设定
场景	密集城区
中心频率	4.0 GHz
用户分布范围	200 m
带宽	10 MHz
子载波间隔	15 kHz
RB 数	48
子带数	12
子带颗粒度	4 RB
天线配置	32 根发送天线, 4 根接收天线, 均匀线性阵列
信道模型	满足 3GPP TS 38.901
Rank 数	1
信道估计	具有理想信道估计

仿真共生成 60 000 条 CSI 数据样本, 并根据式 (2) 和式 (3) 生成对应的 CSI 特征向量数据集用于 AI 模型的训练和推理, 其中 59 400 条用于 AI 模型的训练, 600 条用于模型的验证推理, 训练集和验证集互不交叠。在所提出的 AI 模型 SA-CsiNet 中, 线性嵌入层参数 $d = 384$, 多头注意力模型中的头数为 $n = 8$, Dropout 层随机失活率为 0.1, AI 模型以端到端的形式训练 300 轮次, 每一批 (batch) 训练的样本数为 256, 学习率为 1×10^{-4} , 评价指标为式 (7) 中的余弦相似度。

本文将所提出的基于自注意力机制的 CSI 特征向量反馈 AI 模型与 3 种广泛使用的 CSI 特征向量反馈方案做对比, 即 3GPP TS 38.214 标准化中的 Enhanced Type II 码本^[5], CSI 特征向量反馈 AI 模型 EVCsiNet^[20], 以及用于 CSI 反馈的 AI 模型架构 CRNet。Enhanced Type II 码本是当前标准化中为大规模 MIMO 设计的基于码本的高精度 CSI 特征向量反馈方案; EVCsiNet

是在文献中首次提出的基于 AI 的 CSI 特征向量反馈框架，并提供了基线性能；CRNet 在传统 AI 模型架构下为 CSI 反馈任务提供了较高的反馈精度。为了使 AI 对比方案 EVCsiNet 以及 CRNet 输出正确的结果，本文根据文献[20]中的设定，将其训练次数设定为 1 000 轮次，学习率设为 1×10^{-3} ，其余设定与所提出的 SA-CsiNet 保持一致。值得注意的是，由于 AI 模型通过学习训练数据来完成 CSI 反馈任务，因此如果使用其他不同设定的训练数据集，AI 模型也能够以较高的反馈精度完成其他数据设定下的 CSI 反馈任务。

3.2 实验结果

3.2.1 余弦相似度评估

余弦相似度能够评价反馈的 CSI 特征向量与原始特征向量的接近程度，并且反馈的 CSI 特征向量与原始特征向量的余弦相似度与无线通信系统的吞吐量正相关。本文考虑 $N_s = 12$ 、 $N_t = 32$ 时合适的几组 CSI 反馈开销进行实验验证，即 AI 模型的总反馈开销为 48 bit、60 bit、120 bit 和 240 bit。对于 Enhanced Type II 码本而言，应选择与 AI 模型的反馈开销最接近的参数配置。根据 3GPP TS 38.212^[21]中给定的开销计算方法，选择 paramCombination_r16 序号为 1、2、4、8 的配置方案，分别对应 Enhanced Type II 码本反馈开销 49 bit、71 bit、130 bit 和 250 bit。对于 AI 模型 EVCsiNet、CRNet 和本文所提出的 SA-CsiNet，将均匀量化器的量化比特数设定为 $B = 2$ ，则反馈码字长度 $M = O / B$ 。

SA-CsiNet 的 CSI 特征向量反馈精度对比见表 2，所提出的 SA-CsiNet 在所有的反馈开销设定下的重建 CSI 特征向量余弦相似度均高于 EVCsiNet 和 CRNet。由于 EVCsiNet 为基于 AI 的 CSI 特征向量反馈任务提供基线方案，因此其在各反馈开销设定下的余弦相似度性能与 Enhanced Type II 码本接近。SA-CsiNet 由于自注

意力机制的引入，提高了 AI 对 CSI 特征向量的压缩和特征重构能力，展现出自注意力机制在基于 AI 的 CSI 反馈任务中能够提供高反馈精度的潜力。

表 2 SA-CsiNet 的 CSI 特征向量反馈精度对比

方法	反馈开销/bit	余弦相似度 (ρ)
Enhanced Type II 码本	49	0.793
EVCsiNet	48	0.741
CRNet	48	0.744
SA-CsiNet	48	0.946
Enhanced Type II 码本	71	0.829
EVCsiNet	60	0.749
CRNet	60	0.754
SA-CsiNet	60	0.949
Enhanced Type II 码本	130	0.882
EVCsiNet	120	0.792
CRNet	120	0.799
SA-CsiNet	120	0.959
Enhanced Type II 码本	250	0.914
EVCsiNet	240	0.835
CRNet	240	0.854
SA-CsiNet	240	0.968

考虑给定反馈开销 O 的情况下，码字长度 M 和量化器的量化比特数 B 有多种可能的组合，本工作还将量化比特数 B 设为 3 以对比不同量化精度对 AI 模型性能的影响。在给定的反馈开销下，不同量化精度对 AI 模型的反馈精度的影响见表 3，对于 EVCsiNet 和 CRNet 而言，在给定总反馈开销的情况下，2 bit 量化的 CSI 特征向量反馈精度高于 3 bit 量化的反馈精度；但是对于 SA-CsiNet 而言，3 bit 量化的 CSI 特征向量反馈精度高于 2 bit 量化的反馈精度。这说明在总反馈开销固定的情况下，SA-CsiNet 更希望反馈短码长但是高精度的码字，而 EVCsiNet 和 CRNet 作为 CSI 反馈任务中的 AI 基本模型，相对而言更倾向于从较长的粗精度码字中获取更高的反馈精度。这为未来基于



自注意力机制的 AI CSI 反馈模型的量化器设计提供了方向。

表 3 在给定的反馈开销下, 不同量化精度对 AI 模型的反馈精度的影响

反馈开销($O=M \times B$)/bit	方法	M	B	余弦相似度(ρ)
48	EVCsiNet	24	2	0.741
		16	3	0.719
	CRNet	24	2	0.744
		16	3	0.728
	SA-CsiNet	24	2	0.946
		16	3	0.950
60	EVCsiNet	30	2	0.749
		20	3	0.727
	CRNet	30	2	0.754
		20	3	0.734
	SA-CsiNet	30	2	0.949
		20	3	0.954
120	EVCsiNet	60	2	0.792
		40	3	0.758
	CRNet	60	2	0.799
		40	3	0.769
	SA-CsiNet	60	2	0.959
		40	3	0.961
240	EVCsiNet	120	2	0.835
		80	3	0.817
	CRNet	120	2	0.854
		80	3	0.823
	SA-CsiNet	120	2	0.968
		80	3	0.970

3.2.2 复杂度分析

本文对比了 EVCsiNet、CRNet 和所提出的 SA-CsiNet 模型的参数数量和 AI 模型的运行时间, 参数数量代表着存储 AI 模型所需的存储空间大小, 运行时间则代表 AI 模型在实际部署时的推理时间, 对于时间敏感的 CSI 反馈任务而言, AI 模型的运行时间不宜超过 CSI 反馈的周期时间。

AI 模型的复杂度分析, 评价指标: 参数数量、运行时间见表 4, EVCsiNet 在各反馈码长 M 下的模型参数数量均少于 SA-CsiNet 的模型参数数量, 但 EVCsiNet 的参数数量随着码长的增加大幅度增加, 而 SA-CsiNet 的参数数量随着码长的增加小幅度增加。对于运行时间, 本文采用单颗 NVIDIA 1080Ti GPU 进行实验推理。在合理的码长设定下, EVCsiNet 的运行时间大约是 0.4 ms, 而 SA-CsiNet 的运行时间大约为 1 ms, 与 CRNet 的运行时间接近。根据现有标准化规定, 周期性 CSI 反馈的时间间隔一般为 10 ms 或 40 ms, 因此 SA-CsiNet 满足 CSI 反馈的时间需求。

表 4 AI 模型的复杂度分析, 评价指标: 参数数量、运行时间

方法	M	参数数量	运行时间/s
EVCsiNet	24	41 848	0.000 384
CRNet		43 000	0.000 950
SA-CsiNet		21 431 960	0.000 939
EVCsiNet	30	51 070	0.000 403
CRNet		52 222	0.001 022
SA-CsiNet		21 441 182	0.000 989
EVCsiNet	60	97 180	0.000 410
CRNet		98 332	0.000 983
SA-CsiNet		21 487 292	0.001 007
EVCsiNet	120	189 400	0.000 425
CRNet		190 552	0.000 834
SA-CsiNet		21 579 512	0.000 908

4 结束语

基于 AI 模型的 CSI 反馈为当前或未来的无线通信系统提供了 CSI 反馈的新方法。本文考虑将自注意力机制引入 CSI 特征向量反馈 AI 模型中, 以提升 AI 模型对 CSI 的特征提取和重建能力。相较于现有标准化中的 Enhanced Type II 码本和 CSI 特征向量反馈的两个基本 AI 模型

EVCsiNet、CRNet 而言, 所提出的 SA-CsiNet 具有更高的 CSI 反馈的精度, 同时保证了反馈的时效性。

在未来基于 AI 模型的 CSI 反馈方法设计上, 需要考虑轻量化 AI 模型、高精度量化器以及 AI 模型的鲁棒性等。

参考文献:

- [1] 骆胜军, 张申科. 5G 大规模天线技术发展及创新方向[J]. 电信科学, 2020, 36(11): 141-148.
LUO S J, ZHANG S K. 5G large-scale array antenna technology development and innovation direction[J]. Telecommunications Science, 2020, 36(11): 141-148.
- [2] 王正强, 杨晓娜, 万晓榆, 等. 大规模 MIMO 系统能效优化算法研究综述[J]. 重庆邮电大学学报(自然科学版), 2019, 31(6): 743-752.
WANG Z Q, YANG X N, WAN X Y, et al. Energy efficiency optimization algorithm in massive MIMO systems: a survey[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2019, 31(6): 743-752.
- [3] 张冬雪, 杜洁汝, 何世彪, 等. 面向大规模 MIMO 系统的下行预编码技术研究进展[J]. 电讯技术, 2023, 63(2): 291-299.
ZHANG D X, DU J R, HE S B, et al. Research progress of downlink precoding techniques for massive MIMO systems[J]. Telecommunication Engineering, 2023, 63(2): 291-299.
- [4] 陈成瑞, 程港, 何世彪, 等. 基于人工智能的大规模 MIMO 信道状态信息反馈综述[J]. 电讯技术, 2021, 61(9): 1181-1190.
CHEN C R, CHENG G, HE S B, et al. Survey of massive MIMO channel state information feedback based on artificial intelligence[J]. Telecommunication Engineering, 2021, 61(9): 1181-1190.
- [5] ETSI. 5G; NR; Physical layer procedures for data (V16.1.0, Release 16): TS 38.214[S]. 2020.
- [6] 毛攀, 黄小光, 汪伟, 等. 基于压缩感知的大规模 MIMO 信道估计与反馈[J]. 电信科学, 2018, 34(12): 46-52.
MAO P, HUANG X G, WANG W, et al. Massive MIMO channel estimation and feedback based on compressed sensing[J]. Telecommunications Science, 2018, 34(12): 46-52.
- [7] 黄秋萍, 刘晓峰, 高秋彬, 等. 基于人工智能的大规模天线信道状态信息反馈研究[J]. 电信科学, 2022, 38(3): 74-83.
HUANG Q P, LIU X F, GAO Q B, et al. Study of channel state information feedback based on artificial intelligence[J]. Telecommunications Science, 2022, 38(3): 74-83.
- [8] Qualcomm. New SI: study on artificial intelligence (AI)/machine learning (ML) for NR air interface: RP-213599[R]. 2021.
- [9] WEN C K, SHIH W T, JIN S. Deep learning for massive MIMO CSI feedback[J]. IEEE Wireless Communications Letters, 2018, 7(5): 748-751.
- [10] LU Z L, WANG J T, SONG J. Multi-resolution CSI feedback with deep learning in massive MIMO system[C]//Proceedings of ICC 2020 - 2020 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2020: 1-6.
- [11] WANG T, WEN C K, JIN S, et al. Deep learning-based CSI feedback approach for time-varying massive MIMO channels[J]. IEEE Wireless Communications Letters, 2018, 8(2): 416-419.
- [12] GUO J, YANG X, WEN C K, et al. DL-based CSI feedback and cooperative recovery in massive MIMO[J]. arXiv preprint arXiv: 2003.03303, 2020.
- [13] CHEN T, GUO J J, JIN S, et al. A novel quantization method for deep learning-based massive MIMO CSI feedback[C]// Proceedings of 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Piscataway: IEEE Press, 2019: 1-5.
- [14] CHANG H, LIANG X, LI H, et al. Deep learning-based bit-stream error correction for CSI feedback[J]. IEEE Wireless Communications Letters, 2021, 10(12): 2828-2832.
- [15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000-6010.
- [16] GUO J, WEN C K, JIN S, et al. Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: design, simulation, and analysis[J]. IEEE Transactions on Wireless Communications, 2020, 19(4): 2827-2840.
- [17] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
- [18] SHLEZINGER N, ELDAR Y C. Deep task-based quantization[J]. Entropy, 2021, 23(1): 104.
- [19] GUO J, LI X, CHEN M, et al. AI enabled wireless communications with real channel measurements: channel feedback[J]. Journal of Communications and Information Networks, 2020, 5(3): 310-317.



[20] LIU W, TIAN W, XIAO H, et al. EVCsiNet: eigenvector-based CSI feedback under 3GPP link-level channels[J]. IEEE Wireless Communications Letters, 2021, 10(12): 2688-2692.

[21] ETSI. 5G; NR; Multiplexing and channel coding (V16.1.0; Release 16): 3GPP TS 38.212[S]. 2020.

[作者简介]



杨蓓（1988-），女，中国电信股份有限公司研究院高级工程师，主要研究方向为 5G/6G 无线 AI 与国际标准化。



尹航（1994-），男，中国电信股份有限公司研究院工程师，主要研究方向为 5G/6G 超大规模天线技术、覆盖增强等物理层关键技术。



蒋峥（1972-），男，博士，中国电信股份有限公司研究院正高级工程师，主要研究方向为无线网络架构、非公共网络、大规模天线等。



梁鑫（1997-），男，北京邮电大学博士生，主要研究方向为物理层无线通信、深度学习等。



余小明（1977-），男，博士，中国电信股份有限公司研究院正高级工程师，主要研究方向为 5G/6G 无线技术与标准化。