

基于大语言模型的超宽带雷达多任务学习方法

饶翀¹, 姜夕康¹, 郭嘉航¹, 李蕾¹, 张琳^{1,2}

(1. 北京邮电大学人工智能学院, 北京 100876; 2. 北京市大数据中心, 北京 101160)

摘要: 超宽带 (UWB, ultra-wideband) 雷达通过无线感知实现人数检测和生命体征监测, 现有方法依赖统计特征或轻量级神经网络, 在多种任务间的迁移与泛化能力有限。鉴于大语言模型 (LLM, large language model) 的优异跨模态学习能力, 提出了基于大语言模型的超宽带雷达多任务微调框架, 结合混合专家 (MoE, mixture of experts) 与低秩适应 (LoRA, low-rank adaptation) 方法, 将雷达信号的时序特征映射至大语言模型嵌入空间, 并通过参数高效微调 (PEFT, parameter-efficient fine-tuning)。在自采集数据集和公开数据集上, 分别对人数检测任务及呼吸、心电图 (ECG, electrocardiogram) 与连续血压 3 种生命体征信号估计任务进行训练和评估。实验结果表明, 所提框架在人数检测准确率和 3 种生命体征信号估计任务的相关系数较现有算法平均提升了 37.62%、7.47%、18.16% 和 14.70%。

关键词: 大语言模型; 超宽带雷达; 人数检测; 生命体征监测; 多任务学习; 混合专家

中图分类号: TP183

文献标志码: A

doi: 10.11959/j.issn.2096-3750.WLW25112

UWB-LLM: ultra-wideband radar multi-tasking learning based on large language model

Rao Chong¹, Jiang Xikang¹, Guo Jiahang¹, Li Lei¹, Zhang Lin^{1,2}

1. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China

2. Beijing Big Data Center, Beijing 101160, China

Abstract: Ultra-wideband (UWB) radar enables people counting and vital signs monitoring with wireless sensing. However, existing methods rely on statistical features or lightweight neural networks, which have limited ability to migrate and generalize across multiple tasks. In view of the powerful cross-modal learning capability demonstrated by large language model (LLM), a UWB radar multi-task learning framework based on LLM was proposed. This approach successfully mapped the temporal features of radar signals into the embedding space of LLM, and by integrating the mixture of experts (MoE) mechanism and low-rank adaptation (LoRA) strategy, UWB-LLM performed parameter-efficient fine-tuning (PEFT) for multi-task learning. Experiments were conducted on both self-collected and publicly available datasets for people counting task as well as respiration, electrocardiogram (ECG), and continuous blood pressure estimation tasks, respectively. Compared with the state-of-the-art algorithms, UWB-LLM achieves average accuracy improvements of 37.62% for people counting and 7.47%, 18.16%, and 14.70% for the three vital sign estimation tasks.

Key words: LLM, UWB radar, people counting, vital signs monitoring, multi-task learning, MoE

0 引言

随着智慧城市、智能座舱和智能家居系统等应

用场景的持续拓展, 传统的视觉与红外传感技术难以满足人数检测与生命体征监测对稳定性、隐私保护与穿透性等的需求^[1]。基于 Wi-Fi 和 ZigBee 的无

收稿日期: 2025-07-14; 修回日期: 2026-05-07

通信作者: 张琳, zhanglin@bupt.edu.cn

基金项目: 国家自然科学基金资助项目 (No. 61971056)

Foundation Item: The National Natural Science Foundation of China (No. 61971056)

线感知方法主要依赖于对接收信号强度指示或信道状态信息的分析^[2]。生物雷达系统，特别是超宽带（UWB, ultra-wideband）雷达，具有高时间分辨率和低功耗的特性^[3]，可感知人体状态乃至心肺活动引起的胸腔起伏，尤其适用于对人体状态的长期、连续监测。基于雷达的人数检测（RPC, radar-based people counting）在统计特征的基础上^[4]引入传统机器学习分类器^[5]和卷积神经网络（CNN, convolutional neural network）^[6-7]，以自动学习多层次的特征表示。Choi 等^[8]设计了一种网络架构 RPCNet，用于提取 UWB 信号的空间和时间特征。另外，雷达在连续生命体征（CVS, continuous vital signs）监测方面也展现出潜力^[9-13]。然而，这类方法在面对环境变化时的适应能力较弱，并且难以在多任务间实现结构共享和泛化。

预训练大语言模型（LLM, large language model）在推理能力、知识融合与迁移学习等方面表现出显著优势^[14]，以 Flamingo^[15]为代表的多模态大模型进一步探索了跨模态嵌入对齐与统一表示学习的潜力。在此基础上，大模型感知通信通过在终端和网络侧协同部署大模型，利用语义推理能力对信道与内容进行联合自适应优化，可显著提升频谱效率和端到端服务质量^[16]。与此同时，一些研究^[17-21]尝试将 LLM 引入心电图（ECG, electrocardiogram）、脑电图（EEG, electroencephalogram）和无线信号等非自然语言的时序信号处理任务中。ECG-LLM^[17]通过在冻结的 LLM 前后设计特定的编码器和解码器，实现

LLM 对 12 导联 ECG 信号的缺失值修复。LLM4CP^[19]利用低秩适应（LoRA, low-rank adaptation）方法对 LLM 进行微调，以适配信道预测任务，LLM4WM^[20]结合混合专家（MoE, mixture of experts）方法进一步实现了面向无线信道的 LLM 多任务微调。Sheng 等^[21]使用可训练的分词器将毫米波信号转换为文本表示，解决了毫米波波束预测问题，但将 LLM 用于雷达信号处理任务仍面临以下挑战。

1) 雷达信号与自然语言存在显著的模态差异。雷达信号作为一种时序连续信号，具有较大的数据长度，而 LLM 由于上下文长度限制难以完整表示长时信号序列。此外，LLM 的分词器（tokenizer）主要基于自然语言语料库进行训练，直接将雷达信号编码为文本格式输入模型，可能会丢失关键的时序与频谱特征。

2) 雷达信号在数据集采集配置上存在差异。不同型号的 UWB 雷达在载波频率、距离分辨率和采样率等诸多属性上存在差异，而对于同一雷达也存在采集时间、采集帧率和采集距离等设置上的差别，进而导致雷达信号难以形成统一的离散表示。

3) 雷达信号数据集难以满足雷达大模型训练需求。相较于自然语言和图像数据，针对特定任务的雷达数据集通常不超过 30 h 且采集对象较少，无法支撑大模型训练。

为了解决以上问题，本文提出了一个基于 LLM 的 UWB 雷达多任务微调框架，后文简称 UWB-LLM 框架，其流程如图 1 所示，UWB-LLM 将 MoE 与

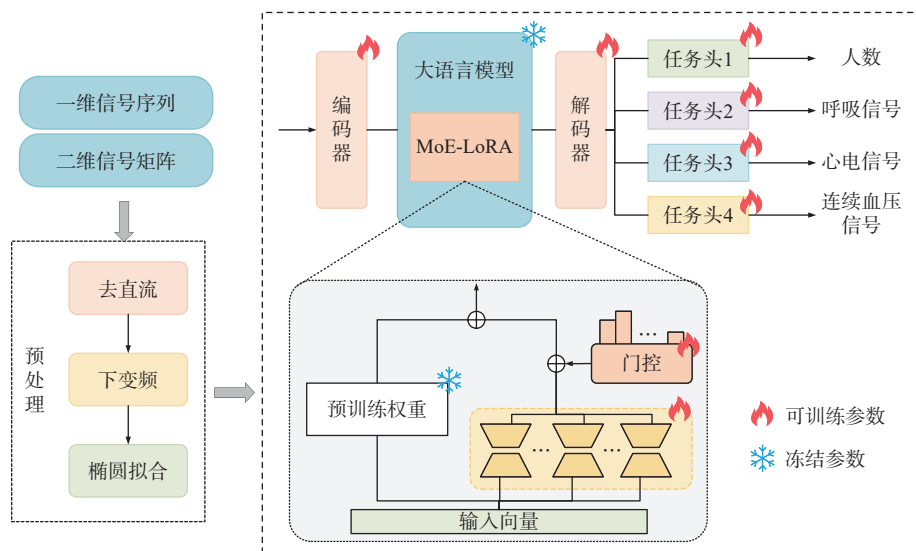


图1 UWB-LLM框架流程

LoRA 结合的 MoE-LoRA 微调方法^[22]创新地应用于与 UWB 雷达信号相关的多任务学习。一方面,为了在样本有限的情况下高效地适应特定任务,预训练 LLM 的原始权重被冻结,通过参数高效微调 (PEFT, parameter efficient fine-tuning) 学习少量附加参数。另一方面,构建专门的信号处理模块与映射网络,将多源雷达信号统一投射至 LLM 的语义嵌入空间,实现信号模态与语言模态的对齐。本文主要贡献如下:

- 1) 将预训练 LLM 引入 UWB 雷达感知任务,为多模态感知与雷达信号处理提供了新的视角和方法;
- 2) 构建了统一的 UWB-LLM 多任务学习框架,设计了一组专用编码器和解码器结构,将 UWB 雷达信号映射至 LLM 嵌入空间,有效保持原始信号的时序特性,并适配 LLM 的输入结构,实现雷达信号与 LLM 语义空间的融合;
- 3) 在保持预训练权重冻结的基础上,结合 MoE 机制的稀疏激活与 LoRA 的轻量参数注入策略,实现对 UWB 雷达信号多任务的高效微调,大幅减少训练成本并提高小样本适应能力;
- 4) 与现有方法相比,所提框架在人数检测与呼吸信号、ECG 和连续血压估计多个任务上均表现出较高性能。实验结果表明,该框架不仅具备更强的表示能力和预测精度,也展现出良好的跨任务泛化性能与鲁棒性。

1 面向 LLM 微调的 UWB 信号处理

LLM 通过在大规模文本语料库上进行自监督学习,捕捉语言中的复杂模式和语义关系,处理离散词元 (token) 构成的输入序列。LLM 内建的 tokenizer 主要基于自然语言设计,其语义分割方式更适用于处理词语、短语或句子等离散符号序列。

然而,UWB 雷达信号作为一种连续、密集的时序数据,在模态特性上与自然语言存在显著差异。自然语言导向的分词策略通常无法有效地保留雷达信号中的幅度变化、相位模式等重要特征。此外,雷达信号的原始数据通常具有较高的采样率和较长的时间跨度,其数据长度远超 LLM 所能处理的上下文窗口承载上限。因此,若要将 LLM 应用于雷达信号处理,须构建专用的数据预处理与表示方式,以保留雷达数据中的时频结构特征,并实现其在 LLM 框架下的有效建模。

1.1 UWB 雷达信号预处理

本文利用 Novelda X4M03 UWB 雷达进行数据采集,将其用于雷达人数检测任务;同时,引入公开的 UWB 数据集以支持 CVS 信号估计任务。以 X4M03 雷达为例,该 UWB 雷达发射载波频率为 7.29 GHz、带宽为 1.4 GHz 的短时高斯脉冲,接收天线以 23.328 GS/s 的速率采样,采集经人体反射后的回波信号。雷达回波经模数转换被数字化为一个二维矩阵,每个数字化脉冲包含 N 个样本,构成一个雷达帧 R_k 表达式为

$$\begin{cases} R_k(n) = r_k\left(\frac{n}{f_s}\right), n \leq N, \\ r_k(t) = \sum_i A_i(p(t - \tau_i)) + n(t) \end{cases} \quad (1)$$

其中, $r_k(t)$ 为第 k 帧接收信号, $p(t)$ 表示脉冲信号, $n(t)$ 为环境噪声, A_i 和 τ_i 分别表示第 i 条路径回波的幅度与时延。 f_s 为接收采样率,雷达帧的大小由设定采样距离和雷达的距离分辨率共同决定。

为了消除不含人体信息的射频 (RF, radio frequency) 载波成分对 LLM 的干扰,利用下变频和带宽为 0.7 GHz 的低通滤波,提取包含幅度、相位和频率变化信息的基带复信号 $M_k(n)$ 。在雷达系统中,杂波被定义为雷达接收到的非目标回波,通常出现在固定位置,信号强度随时间变化相对较小。滑动平均法通过连续更新的均值来估算和去除这些静态或缓慢变化的杂波成分。计算过程如式(2)所示

$$\begin{cases} S_k(n) = M_k(n) - C_k(n), \\ C_k(n) = \sigma C_{k-1}(n) + (1 - \sigma) M_k(n) \end{cases} \quad (2)$$

其中, σ 为平滑因子 ($0 \leq \sigma \leq 1$), $C_k(n)$ 和 $C_{k-1}(n)$ 分别表示当前时刻和前一时刻的杂波, $S_k(n)$ 为预处理后的雷达信号。经过去直流、滤波和杂波抑制等预处理操作后,环境噪声和杂波得到明显的抑制,保留包含生命体征的重要成分,为后续建模提供了可靠基础。

1.2 雷达人数检测

在 RPC 任务中,复杂的人体结构导致多径反射叠加,以及心肺活动等微小运动引起动态调制,人体目标在 UWB 信号中通常表现为显著的幅度调制,因此信号幅度是判别人数最为简单且直观的依据。利用一阶线性插值将信号在距离尺度上按照 1 cm 的间隔进行重采样,从而实现不同距离分辨率的 UWB 雷达统一表示。本文中 UWB 雷达的采样

距离为 0.2~10.0 m, 因此, 重采样后第 k 帧雷达数据 $\tilde{S}_k(n)$ 的长度被统一至 980 个浮点数, 表示相同的物理距离。

在时间维度上, 按照雷达采集帧率滑动切分信号, 将不同长度的 UWB 信号切分为多个持续 1 s 的子序列。考虑人体反射的回波信号呈类周期性变化, 而在较小距离范围内可能存在多个人体目标, 多个目标回波以不同时延进行重叠, 导致信号的局部最大幅度无法准确反映人数。而不同人体回波表现出相近的能量并在距离尺度上叠加, 呈现出不同的高度; 同时, 为了压缩数据长度, 雷达信号矩阵沿传播距离被分为 $\tilde{N}$ 个长度为 L 的距离仓, 根据每个距离仓内的信号能量 E_i 将二维的时间-距离矩阵降维至一维能量序列 $\mathbf{E} = [E_1, E_2, \dots, E_i, \dots, E_{\tilde{N}}]$, 从而可将 RPC 建模为标准的序列分类任务。

1.3 连续生命体征信号估计

人体的呼吸和心跳活动会引起胸壁的扩张和收缩, 放置在胸部前方的雷达系统可以测量人体心肺运动的时间演变, 而这种相对距离的变化反映在 UWB 雷达的相位信号上。为了从原始数据中准确获取信号相位, 利用椭圆拟合算法^[23]减少收发天线之间的干扰并对由偏置、增益引起的误差进行补偿。在经过反正切和相位解缠后, 原本的 I/Q 分量被转换为由多种生命体征组成的单通道雷达数据^[24]。

呼吸活动导致的胸腔运动幅度远大于心跳、脉搏等生命活动, 因此, 雷达信号中相位波形与呼吸信号之间存在高度的耦合性, 这使得利用相位信号可对呼吸活动进行准确建模和估计^[10]。尽管心跳和脉搏引起的胸壁位移幅度较小, 但其高频、周期性振动特征依然可以通过深度学习方法进行提取与增强。已有研究^[11-13]指出, ECG 波形和血压变化均与心脏跳动周期性密切相关, 具有高度的同步性和可预测性。因此, 通过雷达感知胸腔运动, 可间接反映心脏活动状态, 并在无接触的情况下实现对 ECG 波形及血压波动的重建。此外, 融合多任务学习框架还可以进一步提升模型的泛化能力, 使其在不同个体、姿态变化及环境干扰条件下仍能保持较高的估计精度与稳定性。

2 面向 UWB 信号的多任务微调框架

为了充分挖掘 UWB 雷达信号在多任务场景中的潜在信息, 并提升 LLM 对非自然语言模态的建模能力, 本文设计了一个融合编码器、解码器与 PEFT 策略的多任务学习框架, 其结构如图 2 所示。首先, 介绍将原始雷达信号编码转换为 LLM 可接受的嵌入向量; 其次, 引入 MoE-LoRA 机制的 PEFT 策略, 随后, 详述多任务输出映射结构; 最后, 阐明多任务训练策略。通过 UWB-LLM 框架, 能够在保持大

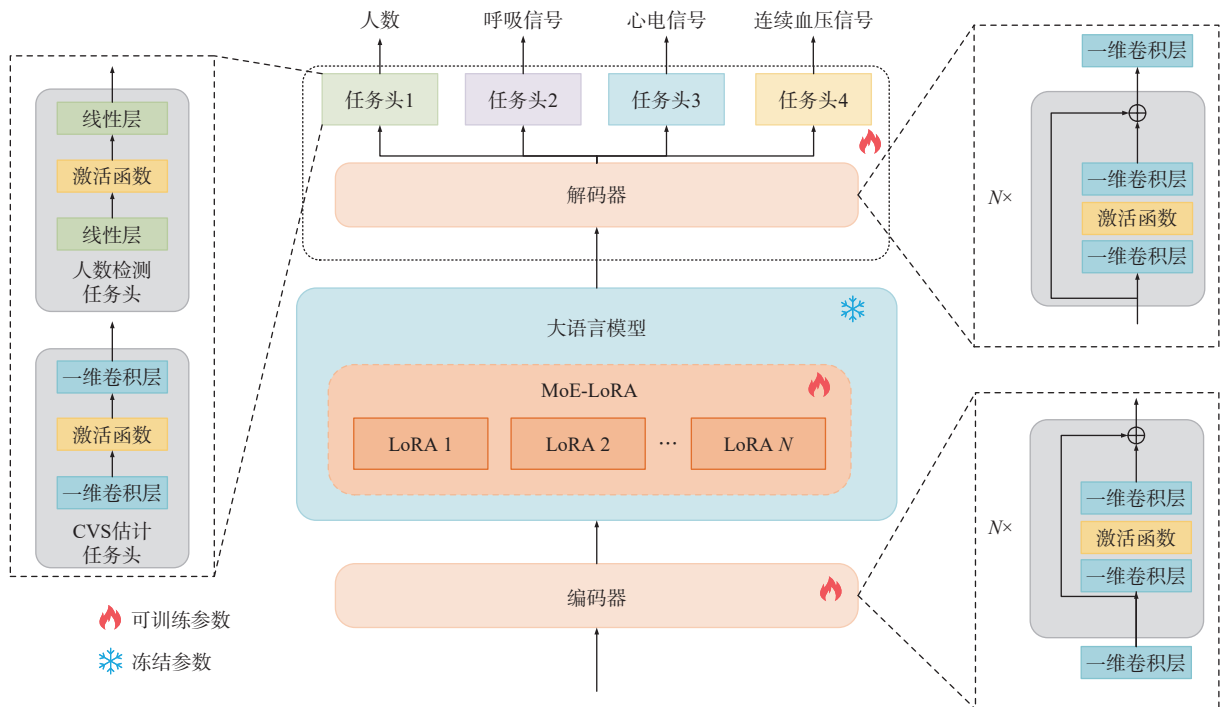


图2 UWB-LLM框架结构

模型推理能力的基础上,实现对人数检测与生命体征估计等任务的协同建模与高效适配。

2.1 UWB信号编码

针对不同任务对信号特征的需求完成预处理操作后,UWB雷达信号由原始的二维时间-距离张量转换为与自然语言一致的一维信号序列,显著降低了LLM的上下文负担。然而,LLM通常以自然语言的token序列为输入,无法与UWB雷达信号兼容。为此,本文设计特定的编码器(encoder)将UWB雷达信号连续时间序列转换为与LLM输入兼容的嵌入向量表示 $\mathbf{X}^e$,从而适配多模态任务的需求。首先,通过一维卷积运算 $\text{Conv1d}(\cdot)$ 提取输入信号的局部时序特征,并将其映射到与LLM兼容的特征维度。同时为增强特征表达能力,引入多个残差模块以构建深层网络。对于任务 t ,得到LLM输入向量,定义如式(3)

$$\mathbf{X}_t^e = \text{encoder}(\mathbf{X}_t^{\text{in}}) = \text{Res}(\text{Conv1d}(\mathbf{X}_t^{\text{in}})) \in \mathbf{R}^{L \times D_{\text{LLM}}} \quad (3)$$

其中, $\mathbf{X}_t^{\text{in}}$ 为来自任务 t 的输入信号序列, L 和 D_{LLM} 分别表示LLM的输入长度和隐藏层维度, $\text{Res}(\cdot)$ 表示由 N_r 个残差块组成的残差模块,每个残差块包含两个卷积核大小为3的一维卷积层和一个ReLU激活函数。通过多层残差块的堆叠,编码器逐步扩大感受野,捕获雷达信号中的长距离依赖关系。

2.2 混合专家低秩微调

在将LLM扩展用于多模态任务时,由于缺少足够的训练数据,直接对全部参数进行全面微调不仅代价高昂,还可能引发过拟合问题。因此,PEFT策略常被选为主要的模型适配方案。其中,LoRA

通过冻结原始模型参数,仅在关键层引入可训练的低位矩阵 $\mathbf{A} \in \mathbf{R}^{r \times d_{\text{in}}}$ 和 $\mathbf{B} \in \mathbf{R}^{d_{\text{out}} \times r}$,实现LLM的高效微调。对于维度为 d_{in} 的输入向量 $\mathbf{x}$,通过式(4)得到维度为 d_{out} 的输出向量 $\mathbf{h}$

$$\mathbf{h} = \mathbf{W}_0 \mathbf{x} + \frac{\alpha}{r} \Delta \mathbf{W} \mathbf{x} = \mathbf{W}_0 \mathbf{x} + \frac{\alpha}{r} \mathbf{B} \mathbf{A} \mathbf{x} \quad (4)$$

其中, $\mathbf{W}_0$ 为预训练权重, $\Delta \mathbf{W}$ 为增量参数矩阵,可训练低位矩阵的秩 r 通过初始阶段的小规模验证实验,确定其最佳取值,缩放因子 α 一般设置为 r 的2倍。

为了进一步提升模型在多模态任务中的表示能力与参数利用效率,近年来有研究^[22]提出将MoE与LoRA策略相结合。MoE通过在LLM的 L 个Transformer块中引入多个专家子网络来学习增量参数矩阵,并由门控网络在每次前向传播中动态激活其中一小部分专家,从而实现稀疏激活、计算节省与模型容量的有效解耦。为了保持紧凑的参数大小,MoE-LoRA层中的每个专家子网络都构造为两个分解的低秩矩阵,MoE-LoRA高效微调过程如图3所示。

因此,对于任务 t 中的样本,MoE-LoRA层的正向过程表示为

$$\mathbf{h}_t = \mathbf{W}_0 \mathbf{x}_t + \frac{\alpha}{r} \sum_{k=1}^{N_e} \omega_{tk} \mathbf{B}_k \mathbf{A}_k \mathbf{x}_t \quad (5)$$

其中, $\mathbf{h}_t$ 和 $\mathbf{x}_t$ 是来自任务 t 的样本的中间LLM层的输入和输出, N_e 表示MoE-LoRA中的专家数量。矩阵 $\mathbf{A}_k$ 和 $\mathbf{B}_k$ 组成专家子网络 E_k ,而 ω_{tk} 代表其在任务 t 中的权重。由于门控网络的设计对MoE的性能有直接影响,为了防止过拟合,采用单层线性网络为每个任务生成专家权重向量 $\omega_t \in \mathbf{R}^{N_e}$,并使用 $\text{softmax}(\cdot)$ 归一化,以保持输出数据的稳定性。同

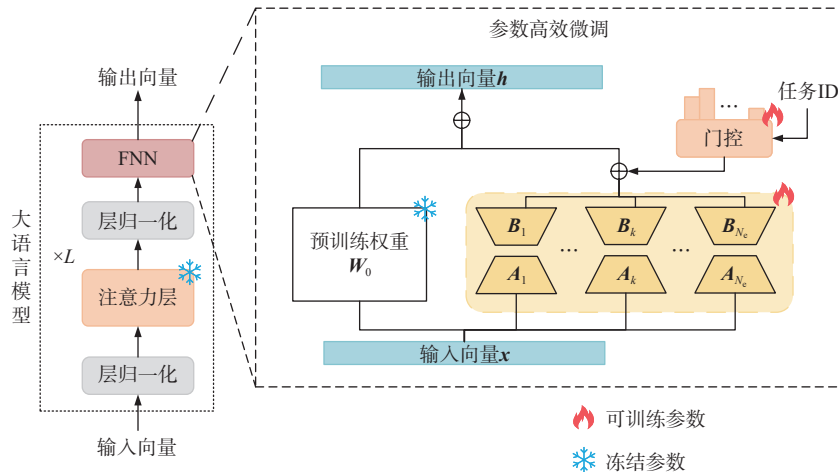


图3 MoE-LoRA 高效微调过程

时, 利用 Top- k 路由动态选择部分专家参与计算, 减轻各种任务中的优化干扰

$$\begin{cases} \omega_i = \text{softmax}(\text{Top}(\omega_i, K)), \\ \text{Top}(\mathbf{x}, K) = \begin{cases} x_i, & x_i \text{ 为 } \mathbf{x} \text{ 中最大 } K \text{ 个数;} \\ 0, & \text{其他} \end{cases} \end{cases} \quad (6)$$

通过将 MoE-LoRA 应用于 LLM 前馈神经网络 (FNN, feedforward neural network) 内的线性层, 同时保持其余参数冻结, 显著地减少了模型的可训练参数, 并增强了任务的适配性。

2.3 多任务输出映射

LLM 的输出通常将特征映射到词汇表上的概率分布, 通过选择概率最高的 token 生成文本输出。然而, 这种方法无法直接适配非文本形式的任务输出。受 Liu 等^[17]的研究启发, 本文设计了与编码器结构对称的解码器, 将 LLM 的语义表示映射至与输入信号 $\mathbf{X}_t^{\text{in}}$ 维度一致的特征表示 $\mathbf{X}_t^{\text{d}}$, 计算式如下

$$\mathbf{X}_t^{\text{d}} = \text{decoder}(\mathbf{X}_t^{\text{lm}}) \quad (7)$$

其中, $\mathbf{X}_t^{\text{lm}}$ 为任务 t 得到的 LLM 输出向量。解码器的对称设计确保了 LLM 的输出与输入信号特征空间的兼容性, 同时通过残差连接缓解深层网络的梯度消失问题, 提升了特征转换的稳定性。

为适配多任务场景, UWB-LLM 框架通过并行化的任务头处理不同任务的输出需求。每个任务头根据任务特性选择合适的神经网络结构, 将解码器输出 $\mathbf{X}_t^{\text{d}}$ 进一步处理, 以获得对应的预测结果, 定义为式(8)

$$\begin{cases} \mathbf{X}_{\text{cls}}^{\text{out}} = \text{MLP}(\mathbf{X}_{\text{cls}}^{\text{d}}) \\ \mathbf{X}_{\text{reg}}^{\text{out}} = \text{CNN}(\mathbf{X}_{\text{reg}}^{\text{d}}) \end{cases} \quad (8)$$

RPC 任务被建模为分类问题, 须从雷达信号中提取全局特征以预测类别标签。为此, 采用多层感知机 (MLP, multi-layer perceptron) 作为输出头, 对解码器输出 $\mathbf{X}_{\text{cls}}^{\text{d}}$ 进行特征整合和维度对齐。CVS 信号估计任务涉及呼吸信号、ECG 和连续血压的序列回归, 须捕捉信号的局部时序模式。为此, 采用一维 CNN 作为输出头, 对每个回归任务的解码器输出 $\mathbf{X}_{\text{reg}}^{\text{d}}$ 进行局部特征提取和维度对齐。通过在输出层融入不同类型的任务头, UWB-LLM 能够有效地适配多类型信号的表征需求, 实现分类与回归任务的统一建模, 从而增强模型在多任务场景中的泛化能力与灵活性。

2.4 多任务训练

考虑 LLM 的预训练知识蕴含强大的通用表征,

若直接进行端到端微调, 容易造成灾难性遗忘与梯度冲突, 特别是在多任务、异构数据以及低资源条件下, 模型通常出现优化不稳定、任务间相互干扰的问题。为此, 训练过程分为两个阶段, 以逐步适配雷达信号与预训练 LLM 的语义空间, 并在多任务场景中实现联合建模。第 1 阶段, 冻结主干预训练 LLM 的参数, 仅对编码器和解码器进行训练, 实现输入输出特征空间与预训练 LLM 的语义空间的对齐, 从而在不破坏原有知识的前提下完成任务特征映射与初步学习; 第 2 阶段, 在完成语义对齐的基础上, 编码器和解码器的参数被冻结, 解冻各个任务头参数并利用 MoE-LoRA 方法实现 LLM 的参数高效微调, 进一步优化模型的内部表征, 使其在共享通用知识的同时学习多任务特异信息, 达到跨任务建模与泛化增强的目的。为在多任务场景中实现高效协同训练, 采用加权组合的损失函数进行统一优化, 其总体目标函数定义为

$$L_{\text{total}} = \sum_t \omega_t L_t \quad (9)$$

其中, L_t 表示任务 t 的损失函数, ω_t 为其对应的权重, 用于平衡不同任务对总损失的贡献。损失函数根据任务类型选择: 人数检测属于分类问题, 采用交叉熵损失函数有效衡量模型在离散类别预测中的准确性, 其损失函数为

$$L_{\text{cls}} = -\sum_c y_c \ln(p_c) \quad (10)$$

其中, y_c 为真实类别的独热编码, p_c 为模型预测第 c 类的概率分布。而 CVS 估计任务涉及呼吸信号、ECG 和连续血压的序列回归, 采用均方误差损失函数作为优化指标, 用于衡量预测序列与真实信号之间的相对偏差, 对于每个回归任务损失如式(11)

$$L_{\text{reg}} = \frac{1}{N} \sum_i (y_i - \hat{y}_i)^2 \quad (11)$$

其中, y_i 和 $\hat{y}_i$ 分别为第 i 个样本的真实值和预测值, N 为样本总数。同时, 为避免静态任务权重在多任务优化过程中导致优化偏置, 引入了动态权重平均 (DWA, dynamic weight average) 算法^[25]以自适应地调整各任务权重, 从而在训练过程中动态平衡任务间的学习进度。该算法依据相邻训练轮次中的任务损失的变化比率确定权重更新规则, 对于任务 t 的权重计算式为

$$\lambda_t^{(i)} = \frac{\exp(w_t^{(i-1)}/T)}{\sum_j \exp(w_j^{(i-1)}/T)}, w_t^{(i-1)} = \frac{L_t^{(i-1)}}{L_t^{(i-2)}} \quad (12)$$

其中, i 表示迭代次数, T 用于控制权重调整的平滑程度, 通常被设置为1。通过动态权重调整, 模型能够在训练早期关注收敛较慢或学习困难的任务, 而在后期实现整体均衡优化, 从而提升多任务协同训练的稳定性与整体性能。

3 实验与分析

3.1 实验设置

为评估 UWB-LLM 微调框架的多任务有效性和泛化能力, 本文在 RPC 任务和 CVS 估计任务上进行了实验验证。实验分别采用基于自采集数据集和公开数据集^[5, 10], 验证模型在不同场景和数据特性下的性能。

RPC-I 数据集采集场景如图4(a)所示, 数据采集使用 X4M03 UWB 雷达, 设备离地 2.0 m, 探测范围为 0.2~10.0 m, 中心角为 60°, 以覆盖足够的检测区域。实验设置 0 至 10 人在指定区域内随机自由移动的多人场景, 每次测量从 11 名参与者中随机选择采集对象, 确保类别分布和实验对象的多样性, RPC 数据采集参与者的生理数据见表1。雷达以每秒 100 帧的频率采集信号, 每次数据记录持续 5 min, 总计约 316 200 帧雷达数据, 数据集以 RPC-I 命名。此外, 为了验证模型的泛用性, 选择公开的密集人数数据集^[5]中的 0 至 10 人数据 (RPC-II) 对模型性能进行测试, RPC-II 数据集采集场景如图4(b)所示。

CVS 估计任务使用由德国汉堡工业大学 Schellenberger 等创建的公开数据集^[10], 该数据集从 30 名参与者中采集, 旨在提供高质量的生理信号数据用于呼吸信号、ECG 和连续血压的估计。CVS 数据集采集场景如图4(c)所示, 采用 6 端口单发射单接收连续波雷达, 获得的原始雷达信号由同相 (I) 和正交 (Q) 分量组成, 通过采样率为 2 000 Hz 的 24 位模数转换器进行数字化处理。实验同步使用多参数监护仪 (TFM, task force monitor) 采集高精度的呼吸信号、ECG 和连续血压等生理信号。

所有实验均在具有 NVIDIA A800 80 GB GPU 的高性能服务器上运行, CUDA 版本为 11.8, 训练代码基于 PyTorch 2.6.0 版本实现并开源, 所有随机种子固定, 以保证实验结果可复现。用于微调的预训练 LLM 包括 Qwen3-0.6B^[26] 与 DeepSeek-R1 的 1.5B 和 7B 蒸馏模型^[27]; 残差块数量 N_r 设置为 8, LoRA 的秩 r 为 8, MoE 专家数 N_e 为 4; 训练采用 AdamW 优化器, 初始学习率设定为 1.0×10^{-4} , 以平衡收敛速度与训练稳定性; 训练中 RPC-I 和 CVS 数据集均按照 7:1:2 的比例划分为训练集、验证集和测试集以支持模型训练、调参和性能评估, 训练轮次 (epoch) 设置为 50, 批大小 (batch size) 为 64; 训练过程启用半精度浮点数 (FP16) 混合精度训练, 以减少显存占用并加速计算; 同时, 采用线性学习率衰减策略以防止损失曲线震荡和过拟合, 预热 (warm up) 步骤设置为总训练步数的 10%, 以避免

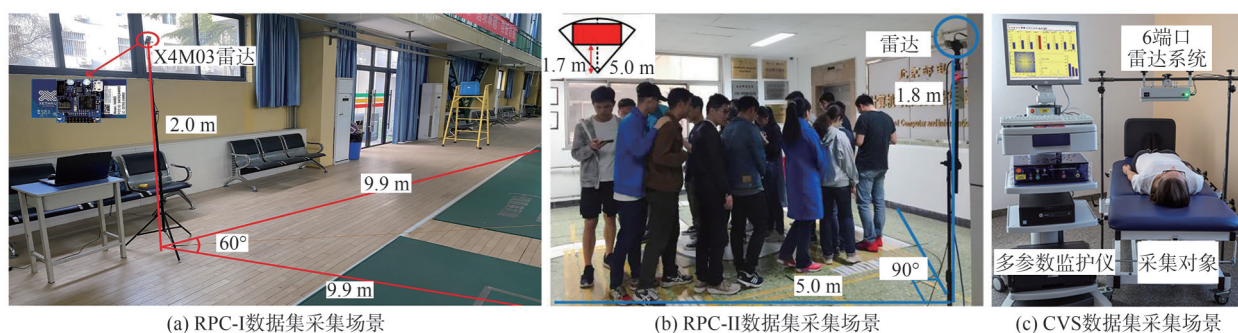


图4 实验环境

表1

RPC 数据采集参与者的生理数据

生理数据	参与者										
	A	B	C	D	E	F	G	H	I	J	K
性别	男	男	男	男	男	男	女	女	女	女	女
身高/cm	183	174	169	165	174	178	167	164	167	165	161
体重/kg	78	75	55	72	70	67	50	48	58	60	49
年龄/岁	42	23	33	30	21	21	23	43	31	36	23

微调初期出现梯度爆炸。

3.2 评估指标

在本文研究的人数检测与 CVS 估计任务中, 针对不同任务的特性, 分别选取了典型的分类与回归评估指标, 以全面衡量模型预测结果与真实数据的吻合程度。

0 至 10 人的人数检测任务可被建模为一个 11 类的多分类问题, 即将输入信号序列的特征映射到 11 个离散类别。本文采用准确率 (ACC, accuracy) 和 $F1$ 分数作为主要评估指标, 用于衡量模型预测结果与真实标签的一致程度。

对于 CVS 信号估计任务, 模型基于雷达信号对呼吸信号、ECG 和连续血压变化进行拟合, 属于典型的回归问题。其目标是通过回归模型预测连续的生命体征值, 因此, 采用皮尔逊相关系数 (PCC, Pearson correlation coefficient) 衡量预测值与真实值之间的相关性, 计算式如下

$$PCC = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (13)$$

其中, x_i 和 y_i 分别为第 i 个样本的真实值和预测值, $\bar{x}$ 和 $\bar{y}$ 分别为全体样本真实值和预测值的均值。PCC 越接近 1, 表明模型预测结果能够更准确地捕捉到真实生命体征的趋势与形状, 具备更强的拟合能力。

3.3 结果与分析

为全面验证本文所提的 UWB-LLM 框架在多任务 UWB 雷达感知中的有效性与泛化能力, 分别在自采集人数检测数据集与公开 CVS 数据集上进行系统性的实验, 并与多种主流方法进行对比分析, 评估其在分类与回归两类任务中的性能表现。

在 RPC 任务中, UWB-LLM 在所有人数类别上均表现出高度一致且稳定的分类性能, 展现出卓越的泛化能力和强健的目标建模能力。不同模型的 0 至 10 人 RPC 结果见表 2, 基于 0.6B、1.5B 和 7B 不同参数量大模型的 UWB-LLM, 平均准确率

94.22%~97.27%, $F1$ 分数达到 0.941 7~0.973 4, 显著优于现有的基于微多普勒特征的方法^[28]、DSC-CNN^[29]和 RPCNet^[8]模型。传统方法如基于手工微多普勒特征的算法在多目标干扰场景中表现不佳, 而基于深度学习的 1D-CNN 与 RPCNet 模型尽管在较少人数 (如 0 至 3 人) 类别下具有一定识别能力, 但在中高人数场景中准确率迅速下降, 这是因为在数据收集过程中, 当 4 个人并排快速行走时, 雷达回波会产生强烈的多径干扰与时变信号。这类复杂信号对传统模型构成显著挑战, 而 UWB-LLM 凭借其融合 LLM 架构与任务定制化编码策略, 在高维语义空间中保留了时间依赖性和距离层级特征, 从而准确识别人数, 避免了误分类与低置信输出。此外, 利用 RPC-II 数据集对在 RPC-I 数据集上完成训练的模型进行测试 (RPC-I→RPC-II), RPC 任务中不同模型的泛化性能对比如图 5 所示。由于实验场景的变化, 各模型性能出现显著下降, 但与其他模型相比, UWB-LLM 仍保持了高准确率与 $F1$ 分数, 表明其具有更强的类别平衡性与泛化能力。随着模型规模增大, 跨域性能衰减逐渐减缓, 反映出参数规模对特征抽象与域不变性的正向作用。

对于 CVS 估计任务, 本文选取了多个代表性的深度学习模型进行对比实验。其中, 根据文献复现 ECGCNN^[30]和 CNN-LSTM^[31]的 ECG 估计算法, 并进一步推广至呼吸和连续血压信号预测。Transformer-TCN^[11]、RSSRNet^[32]、HyperECG^[33]、MultiRes-LinkNet^[24]和 RSD-Net^[13]等方法的性能指标直接引用自其公开发表的实验结果, 并且除 Transformer-TCN 外, 其他算法与本文使用同一公开 CVS 数据集, 保证了对比的公平性与一致性。不同模型的 CVS 信号估计的 PCC 见表 3, CVS 信号估计的可视化结果如图 6 所示。UWB-LLM 在呼吸信号、ECG 和连续血压 3 项 CVS 估计任务上均取得了显著的优势, 其 PCC 分别达到了 0.989 5、0.966 2 和 0.983 2,

表 2 不同模型的 0 至 10 人 RPC 结果

模型	0	1	2	3	4	5	6	7	8	9	10	平均	$F1$
微多普勒 ^[28]	100.00	89.58	88.46	73.68	42.00	56.34	33.90	56.25	55.36	43.86	74.00	64.71	0.646 2
1D-CNN ^[29]	100.00	93.10	90.22	78.86	35.34	66.94	64.91	59.57	65.74	70.00	77.19	72.16	0.722 3
RPCNet ^[8]	100.00	95.73	85.71	85.95	51.73	77.48	68.14	70.21	64.49	68.18	80.87	76.14	0.759 6
UWB-LLM(0.6B)	100.00	100.00	97.96	98.28	95.52	92.42	90.77	93.48	96.78	93.75	82.26	94.22	0.941 7
UWB-LLM(1.5B)	100.00	100.00	96.77	96.97	96.67	96.30	93.10	95.45	93.55	95.24	96.15	97.05	0.968 7
UWB-LLM(7B)	100.00	100.00	100.00	96.55	95.59	96.97	95.38	97.83	100.00	95.83	93.55	97.27	0.973 4

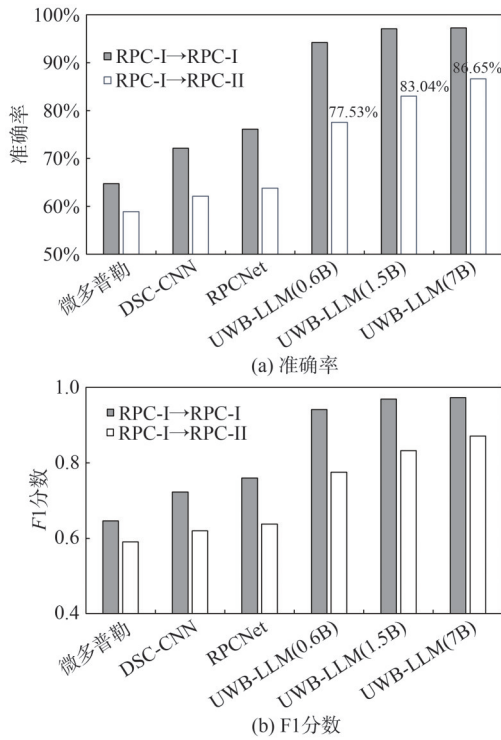


图5 RPC任务中不同模型的泛化性能对比

表3 不同模型的CVS信号估计的PCC

模型	呼吸信号	ECG	连续血压
ECGCNN ^[30]	0.877 4	0.686 7	0.855 7
CNN-LSTM ^[31]	0.900 5	0.716 3	<u>0.878 9</u>
Transformer-TCN ^[11]	—	0.900 0	—
RSSRNet ^[32]	0.992 0	<u>0.964 0</u>	—
HyperECG ^[33]	—	0.890 0	—
MultiResLinkNet ^[24]	—	0.824 4	—
RSD-Net ^[13]	—	—	0.838 0
UWB-LLM(0.6B)	<u>0.955 0</u>	0.896 7	0.941 7
UWB-LLM(1.5B)	<u>0.987 6</u>	0.957 9	0.979 0
UWB-LLM(7B)	<u>0.989 5</u>	0.966 2	0.983 2

均高于大多数传统方法与深度模型。尽管RSSRNet在呼吸信号估计上的PCC为0.992 0, 略优于UWB-LLM, 但其难以对连续血压信号进行估计。UWB-LLM不仅能够从相位信号中精确提取低信噪比的微弱心跳信号, 还可实现对高频波动的连续血压的高保真还原。

CVS估计结果箱型图如图7所示, 显示了CVS估计任务中每个采集目标获得的PCC分布, 对3种CVS信号的拟合均表现出较高的一致性和准确性, 但不同生命体征的拟合水平及波动性存在显著差异。UWB雷达对呼吸信号和连续血压信号的拟合性能明显优于ECG信号, 不仅在整体上保持较高的

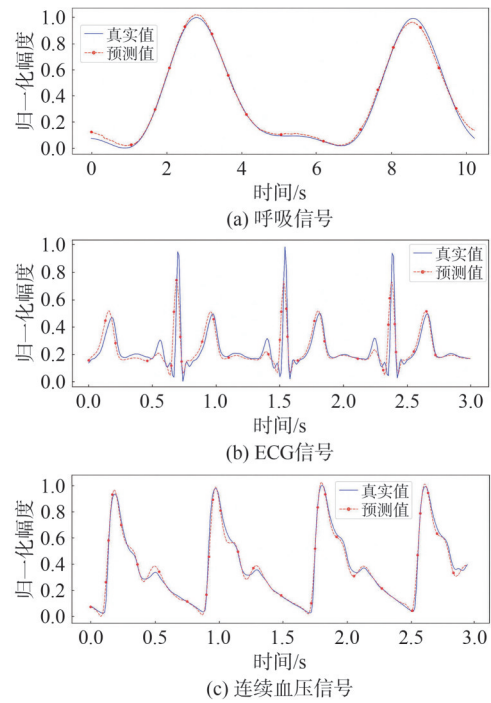


图6 CVS信号估计的可视化结果

PCC水平, 且估计结果的波动性显著降低。而ECG信号的估计受其本身复杂性高、运动幅度小等因素影响, 波动性较大, 拟合精度相对较低。以上结果证明了UWB-LLM框架在多种CVS信号监测中的潜力, 同时也揭示了其在复杂心跳活动感知和建模方面仍有进一步优化和提升的空间。

为了评估LLM模块的有效性, 主干LLM模块消融实验结果见表4, 可以看出, 移除LLM模块会导致一定的性能下降, RPC准确率为95.50%, 3种生理信号的PCC分别为0.928 0、0.875 3和0.923 6, 体现了小参数神经网络在各任务中的表征能力, 但难以捕获复杂时序与跨任务共享信息; 当引入LLM但保持权重冻结时, 由于预训练LLM的知识难以适配具体的雷达感知任务, 模型性能出现不同程度的下降; 在移除MoE的情况下, 所有任务共享同一组LoRA参数, 导致不同任务间的梯度干扰加剧。数据量较小的RPC任务准确率下降幅度显著大于其他任务, 多任务间的性能差距增大, 说明MoE在提升整体性能的同时, 也有助于维持各任务之间的性能平衡。此时4类任务的结果均高于冻结LLM方案, 说明通过LoRA微调使LLM逐渐适应雷达感知任务; 值得注意的是, 随机初始化LLM参数会使模型性能出现最为剧烈的下降, 各项指标均显著低于移除LLM和冻结LLM的情况, 这表明UWB-LLM的性能

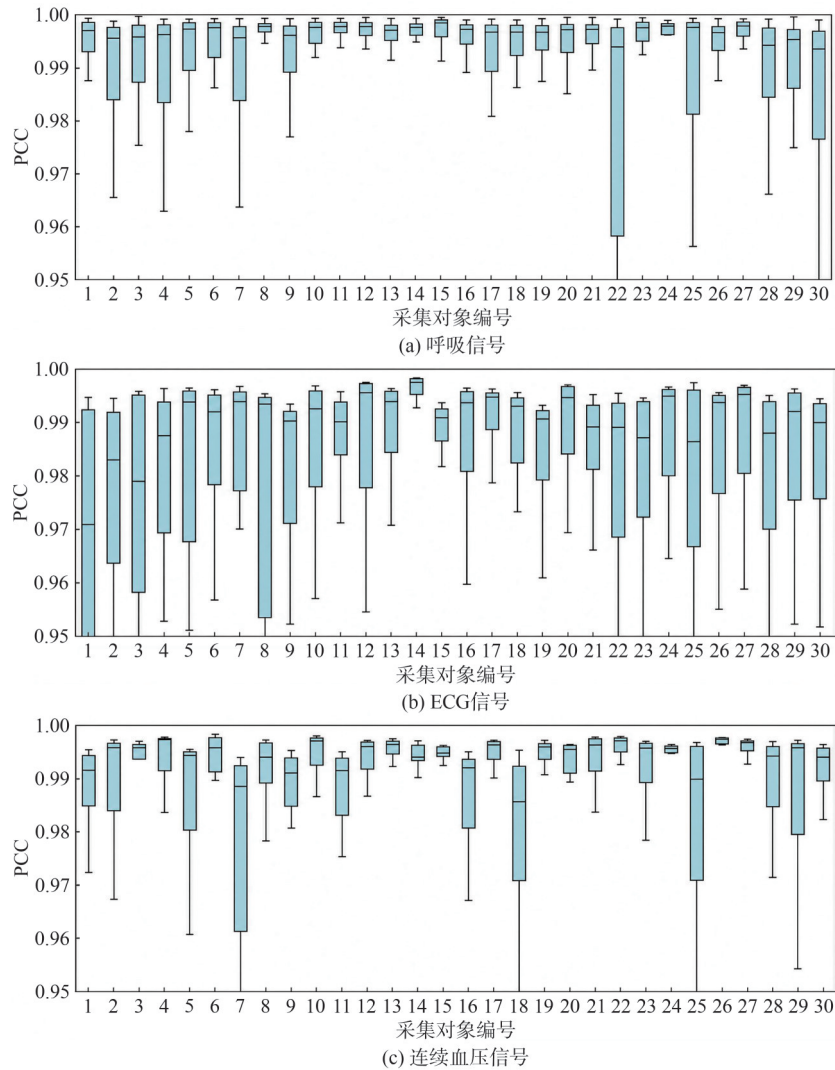


图7 CVS估计结果箱型图

表4 主干LLM模块消融实验结果

模块	RPC	CVS估计PCC		
	准确率	呼吸	ECG	连续血压
UWB-LLM(7B)	97.27%	0.989 5	0.966 2	0.983 2
移除LLM	<u>95.50%</u>	0.928 0	0.875 3	0.923 6
移除MoE-LoRA, 冻结LLM	94.06%	0.930 7	0.868 5	0.928 3
移除MoE, 仅LoRA微调	94.21%	<u>0.968 9</u>	<u>0.939 1</u>	<u>0.968 5</u>
随机初始化LLM参数	83.50%	0.868 2	0.832 6	0.857 7

提升, 并非只来自LLM作为大规模Transformer网络的结构性能容量, 同时也受益于其预训练阶段所学习到的表征知识。综上所述, 预训练LLM在所提框架中的作用不仅体现在其深层网络结构, 更体现在其通过大规模预训练所获得的跨模态、长时序知识表征, 而基于MoE-LoRA的轻量化微调在此表征基础上实现了面向多任务的有效适配和平衡。

为评估UWB-LLM在实际应用场景中的部署可行性, 将本文模型参数量和平均推理时间与其他基线模型进行比较, 所有实验均在同一硬件环境下完成 (14 核 Intel Xeon Gold 6348 CPU、NVIDIA A800 80 GB GPU 和 100 GB RAM), 模型参数量与平均推理时间见表5。尽管UWB-LLM的总参数规模相对较大, 但由于大部分参数被冻结, 仅通过MoE-LoRA实现高效微调, 因此, 其实际训练参数量与其他模型相近, 体现出显著的训练效率与参数利用效率。同时, 虽然LLM在结构复杂度和参数规模上的固有特点不可避免地带来一定的推理时延, 但UWB-LLM的整体推理速度仍与其他模型保持在同量级。这表明, UWB-LLM框架在兼顾性能的同时, 具备良好的部署前景, 为未来在资源受限或实时性要求较高的场景中应用提供了有力支撑。

表5 模型参数量与平均推理时间

模型	可训练参数量	总参数量	平均推理时间/ms
DSC-CNN ^[29]	31.7×10^3	31.7×10^3	0.16
ECGCNN ^[30]	14.4×10^6	14.4×10^6	0.67
RPCNet ^[8]	7.1×10^6	7.1×10^6	6.68
RSSRNet ^[32]	17.1×10^6	17.1×10^6	9.16
UWB-LLM	15.1×10^6	7.7×10^9	15.75

4 结束语

本文提出了一种针对UWB雷达系统与LLM结合的多任务微调框架，使模型能够同时完成人数检测与多种连续生命体征估计任务。首先，介绍了面向LLM微调的UWB雷达信号处理方法，将异源雷达数据统一至一维序列，从而将RPC与CVS估计任务分别建模为序列分类和回归问题；然后，设计了一套特定的编码器和解码器结构，将UWB雷达信号映射至LLM嵌入空间，有效保持原始信号的时序特性，并适配LLM的输入输出结构，实现了信号与自然语言表示的有效对齐；最后，将MoE-LoRA引入LLM微调过程中以同时对多种任务进行学习，基于开源数据集实验验证了本文所提方法的有效性。未来的研究可探索UWB-LLM的其他应用和优化，提高其在复杂监测环境中的适应能力。

参考文献：

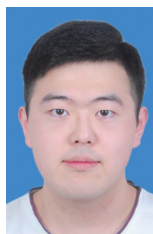
- [1] 刘洋, 董安明, 禹继国, 等. 基于CNN-BiGRU的复杂连续人体活动Wi-Fi感知方法[J]. 物联网学报, 2023, 7(4): 153-167.
- [2] Di Domenico S, Pecoraro G, Cianca E, et al. Trained-once device-free crowd counting and occupancy estimation using WiFi: a Doppler spectrum based approach[C]//Proceedings of the 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). New York : IEEE Press, 2016: 1-8.
- [3] 吴彤, 李业深, 黄镇煌, 等. 一种行人遮挡下的UWB非视距传播识别方法[J]. 物联网学报, 2023, 7(4): 63-71.
- [4] Choi J W, Yim D H, Cho S H. People counting based on an IR-UWB radar sensor[J]. IEEE Sensors Journal, 2017, 17(17): 5717-5727.
- [5] Yang X Z, Yin W F, Li L, et al. Dense people counting using IR-UWB radar with a hybrid feature extraction method[J]. IEEE Geoscience and Remote Sensing Letters, 2019, 16(1): 30-34.
- [6] Yang X Z, Yin W F, Zhang L. People counting based on CNN using IR-UWB radar[C]//Proceedings of the 2017 IEEE/CIC International Conference on Communications in China (ICCC). Qingdao : IEEE Press, 2017: 1-5.
- [7] Bao R H, Yang Z C. CNN-based regional people counting algorithm exploiting multi-scale range-time maps with an IR-UWB radar[J]. IEEE Sensors Journal, 2021, 21(12): 13704-13713.
- [8] Choi J H, Kim J E, Kim K T. Deep learning approach for radar-based people counting[J]. IEEE Internet of Things Journal, 2022, 9(10): 7715-7730.
- [9] Wang D Y, Yoo S, Cho S H. Experimental comparison of IR-UWB radar and FMCW radar for vital signs[J]. Sensors, 2020, 20(22): 6695.
- [10] Schellenberger S, Shi K, Steigleder T, et al. A dataset of clinically recorded radar vital signs with synchronised reference sensor signals[J]. Scientific Data, 2020, 7: 291.
- [11] Chen J B, Zhang D H, Wu Z, et al. Contactless electrocardiogram monitoring with millimeter wave radar[J]. IEEE Transactions on Mobile Computing, 2024, 23(1): 270-285.
- [12] Oyamada Y, Koshisaka T, Sakamoto T. Experimental demonstration of accurate noncontact measurement of arterial pulse wave displacements using 79-GHz array radar[J]. IEEE Sensors Journal, 2021, 21(7): 9128-9137.
- [13] Qiu Y, Ma X J, Li X L, et al. Non-contact blood pressure estimation from radar signals by a stacked deformable convolution network[J]. IEEE Journal of Biomedical and Health Informatics, 2024, 28(8): 4553-4564.
- [14] 李鑫尧, 李晶晶, 朱磊, 等. 资源受限的大模型高效迁移学习算法研究综述[J]. 计算机学报, 2024, 47(11): 2491-2521.
- [15] Li X Y, Li J J, Zhu L, et al. Efficient transfer learning of large models with limited resources: a survey[J]. Chinese Journal of Computers, 2024, 47(11): 2491-2521.
- [16] Alayrac J B, Barr I, Barreira R, et al. Flamingo: a visual language model for few-shot learning[C]//Proceedings of the Advances in Neural Information Processing Systems 35. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 23716-23736.
- [17] Large Generative AI in Models in Telecom (GenAINet), IEEE ComSoc Emerging Technology Initiative. Large-Scale AI in Telecom: Charting the Roadmap for Innovation, Scalability, and Enhanced Digital Experiences[R]. 2025.
- [18] Liu L F, Cui G S, Wan C, et al. ECG-LLM: leveraging large language models for low-quality ECG signal restoration[C]//Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Lisbon : IEEE Press, 2024: 3537-3542.
- [19] Li Z Y, Zheng W L, Lu B L. Gram: a large-scale general EEG model for raw data classification and restoration tasks[C]//Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad: IEEE Press, 2025: 1-5.
- [19] Liu B X, Liu X Y, Gao S J, et al. LLM4CP: adapting large lan-

- guage models for channel prediction[J]. Journal of Communications and Information Networks, 2024, 9(2): 113-125.
- [20] Liu X Y, Gao S J, Liu B X, et al. LLM4WM: adapting LLM for wireless multi-tasking[J]. IEEE Transactions on Machine Learning in Communications and Networking, 2025, 3: 835-847.
- [21] Sheng Y C, Huang K, Liang L, et al. Beam prediction based on large language models[J]. IEEE Wireless Communications Letters, 2025, 14(5): 1406-1410.
- [22] Liu Q D, Wu X, Zhao X Y, et al. When MOE meets LLMs: parameter efficient fine-tuning for multi-task medical applications[C]// Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington DC: ACM Press, 2024: 1104-1114.
- [23] Singh A, Gao X M, Yavari E, et al. Data-based quadrature imbalance compensation for a CW Doppler radar system[J]. IEEE Transactions on Microwave Theory and Techniques, 2013, 61(4): 1718-1724.
- [24] Chowdhury F A, Hosain M K, Bin Islam M S, et al. ECG waveform generation from radar signals: a deep learning perspective[J]. Computers in Biology and Medicine, 2024, 176: 108555.
- [25] Liu S K, Johns E, Davison A J. End-to-end multi-task learning with attention[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE Press, 2019: 1871-1880.
- [26] Fu Y, Dong X, Diao S, et al. Nemotron-Flash: towards latency-optimal hybrid small language models[C]//Proceedings of the Advances in Neural Information Processing Systems 39. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025.
- [27] Hu Y, Yu Y, Gan L, et al. Evaluating test-time scaling LLMs for legal reasoning: OpenAI o1, DeepSeek-R1, and Beyond[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou: Association for Computational Linguistics (ACL), 2025: 13759-13781.
- [28] Kakoni D, Storrer L, Cornelis B, et al. Outdoor group counting based on micro-Doppler signatures obtained with a 77GHz FMCW radar[C]//Proceedings of the 2024 21st European Radar Conference (EuRAD). Paris: IEEE Press, 2024: 376-379.
- [29] Mach T K T, Pham C T, Le M. UWB impulse radar for people counting with convolutional neural network on microcontrollers[J]. IEEE Sensors Journal, 2024, 24(9): 15643-15650.
- [30] Toda D, Ren A Z, Ichige K, et al. ECG signal reconstruction using FMCW radar and convolutional neural network[C]//Proceedings of the 2021 20th International Symposium on Communications and Information Technologies (ISCIT). Tottori: IEEE Press, 2021: 176-181.
- [31] Yamamoto K, Hiromatsu R, Ohtsuki T. ECG signal reconstruction via Doppler sensor by hybrid deep learning model with CNN and LSTM[J]. IEEE Access, 2020, 8: 130551-130560.
- [32] Wu Y X, Ni H C, Mao C L, et al. Contactless reconstruction of ECG and respiration signals with mmWave radar based on RSS-Rnet[J]. IEEE Sensors Journal, 2024, 24(5): 6358-6368.
- [33] Gaddi M, Ponzina F, Asgarinejad F, et al. HyperECG: ECG signal inference from radar with hyperdimensional computing[C]//Proceedings of the 2024 IEEE 24th International Conference on Bioinformatics and Bioengineering (BIBE). Kragujevac: IEEE Press, 2024: 1-5.

[作者简介]



饶翀(2002—)，男，北京邮电大学人工智能学院硕士生，主要研究方向为智慧医疗、雷达信号处理、多模态大模型。



姜夕康(1998—)，男，北京邮电大学人工智能学院博士生，主要研究方向为智慧医疗、雷达信号处理、多模态大模型。



郭嘉航(1999—)，男，北京邮电大学人工智能学院硕士生，主要研究方向为智能座舱、雷达信号处理、多模态大模型。



李蕾(1974—)，女，博士，北京邮电大学人工智能学院教授、博士生导师，主要研究方向为自然语言处理、自动摘要、知识图谱、智能信息处理。



张琳(1974—)，男，博士，北京邮电大学人工智能学院教授、博士生导师，主要研究方向为信源-信道编码、无线资源分配机制、网络信号处理。