

基于上下文特征与点集互学习的PointNet++网络

胡 曦^{1,2}, 谢晓兰^{1,2}

(1. 桂林理工大学 计算机科学与工程学院;

2. 桂林理工大学 广西嵌入式技术与智能系统重点实验室, 广西 桂林 541000)

摘 要: 3D点云物体分类和部件分割研究是计算机视觉领域的热门话题,然而由于点云的稀疏性,实现有效的特征学习仍具有挑战性。针对现有深度学习点云处理算法精度低和鲁棒性差的问题,提出一种基于上下文特征与点集互学习的PointNet++网络模型。首先,使用上下文特征提取(CFE)模块改进PointNet++模型的局部特征提取,通过提取上下文特征获取更丰富的信息;其次,在模型中插入相互学习(ML)模块,通过分组向量自注意力机制邻域间关系进行建模,从而进一步增强模型的全局特征学习能力。针对公开数据集S3DIS的场景语义分割和Modelnet40的形状分类实验结果表明,该模型较基准模型效果提升显著。在室内数据集S3DIS上的mIoU较PointNet++提升了6.5%,在三维物体数据集Modelnet40上的OA较PointNet++提升了2%。此外,在与当前主流方法的横向比较中,该模型在分类与分割精度上均实现了不同程度的提升。因此,该模型能够更好地满足点云分类和分割的任务需求。

关键词: 点云;深度学习;注意力机制;PointNet++;语义分割;对象分类

DOI: 10.11907/rjtk.251191

扫描二维码阅读全文:



中图分类号: TP391.41

文献标识码: A

文章编号: 1672-7800(2026)005-0205-07

PointNet++ Network Based on Contextual Features and Point Set Mutual Learning

HU Xi^{1,2}, XIE Xiaolan^{1,2}

(1. School of Computer Science and Engineering, Guilin University of Technology;

2. Guilin University of Technology Guangxi Key Laboratory of Embedded Technology and Intelligent Systems, Guilin 541000, China)

Abstract: Research on 3D point cloud object classification and part segmentation is a prominent topic in computer vision. However, due to the inherent sparsity of point clouds, achieving effective feature learning remains a significant challenge. To address the limitations of low accuracy and poor robustness in existing deep learning-based point cloud processing methods, this paper proposes an enhanced PointNet++ network model that integrates contextual features and mutual learning among point sets. First, the Contextual Feature Extraction (CFE) block is used to improve the local feature extraction capability of the PointNet++ model, enabling the acquisition of richer information through contextual features. Second, the Mutual Learning (ML) block is incorporated into the model, leveraging a grouped vector self-attention mechanism to model inter-neighborhood relationships, thereby further enhancing the model's global feature learning ability. Experimental results on the S3DIS dataset for scene semantic segmentation and the ModelNet40 dataset for shape classification demonstrate significant improvements over the baseline model. Specifically, the proposed model achieves a 6.5% increase in mean Intersection over Union (mIoU) on the indoor S3DIS dataset and a 2% increase in Overall Accuracy (OA) on the 3D object ModelNet40 dataset compared to PointNet++. Furthermore, in horizontal comparisons with current mainstream methods, the proposed model achieves varying degrees of improvement in both classification and segmentation accuracy. Consequently, this model is better able to meet the requirements of point cloud classification and segmentation tasks.

Key Words: point cloud; deep learning; self-attention; PointNet++; semantic segmentation; object classification

0 引言

近年来,三维数据采集技术的持续发展极大地推动了

点云理解领域的研究进程,催生了众多应用领域,包括室内导航^[1]、自动驾驶^[2]、机器人^[3]、手势估计^[4]、地下采矿^[5]、人脸识别^[6]、城市建筑重建^[7]、多目标识别^[8-9]等。

与图像中整齐排列的规则像素网格不同,三维点云存

收稿日期: 2025-04-09

基金项目: 国家自然科学基金项目(62262011); 广西重点研发计划项目(GuikAB23026036)

作者简介: 胡 曦(1995-),男,桂林理工大学计算机科学与工程学院硕士研究生,研究方向为三维点云处理、计算机视觉;谢晓兰(1968-),女,博士,桂林理工大学计算机科学与工程学院教授、博士生导师,研究方向为云计算、大数据、制造业信息化。本文通讯作者: 谢晓兰。

在于一个连续变化的空间中,具有稀疏性、不规则性和无序性,从而给已在计算机视觉领域取得成功的卷积神经网络模型带来了重大挑战。由于此类模型与三维点云数据之间存在的固有结构不匹配,导致其无法直接应用于点云处理任务。因此,设计一个专门针对三维点云数据独特结构的深度神经网络模型显得至关重要,许多针对三维点云的深度学习方法应运而生。

Qi等^[10]率先提出 PointNet,其是一个通过多层感知机(Multilayer Perceptron, MLP)和最大池化操作进行特征学习的深度学习网络。然而,PointNet的一个局限性在于其无法捕捉局部特征,这是卷积神经网络的关键能力。PointNet++通过引入分层结构来解决这一问题,使其能够提取局部特征。但其仅提取了坐标间的差值作为特征信息,而忽略了许多点云的局部几何语义,并且过度依赖局部邻域,难以建模长程依赖关系^[11]。为解决 PointNet++分类与分割精度低,模型特征提取能力差和缺乏长程依赖的问题,本文提出一种改进的基于上下文特征与点集互学习的 PointNet++网络。首先,模型改进了原网络模型的特征提取结构,提出一个上下文特征提取(Contextual Feature Extraction, CFE)模块,增强了模型提取局部特征的能力,从而捕获更丰富的空间信息和语义信息;其次,在 PointNet 层后增加了一个相互学习(Mutual Learning, ML)模块,使用残差连接^[12]、分组向量自注意力机制和可训练的位置编码^[13]提升模型建模长程依赖的能力。实验结果表明,本文模型在三维物体数据集和室内遥感数据集上处理原始点集时,兼具高效性与鲁棒性。

1 相关工作

随着深度学习方法的快速发展,各种用于处理点云数据的深度神经网络不断涌现。这些方法大致可分为两类:一类是将三维点云数据投影到规则结构上,使其能够适配经典的卷积神经网络;另一类则是直接处理点云数据。

(1)将三维点云数据投影到规则结构。传统方法通常将三维点云数据投影到规则结构上,例如图像集合或三维体素网格,然后将数据输入深度神经网络。基于图像的网络通常采用多视图表示,将点云在不同视角下进行渲染得出一组二维图像^[14-17]。MVCNN(Multi-view Convolutional Neural Networks)通过空间投影将三维点云从不同视角投影到二维图像,并使用传统的卷积神经网络处理这些二维数据^[18]。为了避免选择全局投影视角,文献[19]提出将局部邻域投影到局部切平面,并使用二维卷积进行处理。ShapeNets^[20]和 VoxNet^[21]通过体素化将无序的点云转换为规则的三维网格,并使用三维卷积神经网络进行特征学习。尽管这些方法取得了令人称赞的结果,但面临一个困境:二维卷积无法捕捉到如法线和形状等关键的三维几何信息,而三维卷积需要在体素化生成的稀疏三维网格上进

行大量计算。利用八叉树或哈希表等稀疏结构可以处理更大的网格,并提高模型性能^[22-23]。但这些方法仍依赖于对包围体的细分,而不是利用局部几何结构。

(2)直接处理点云数据。鉴于上述原因,最前沿的深度神经网络被专门设计用于应对点云的不规则性。这些网络直接处理原始点云数据,而无需将其转换为中间规则表示(如体素或二维图像)。这一创新方法由 PointNet 率先提出,其通过多层感知机(MLP)提取单个点的特征,然后应用对称函数形成全局特征,从而实现排列不变性^[10]。然而,该方法的局限性在于其独立处理每个点,而忽略了点之间的几何关系,可能会忽略编码关键空间信息的局部特征。

为了克服这些局限性,PointNet++在 PointNet 的基础上引入一种分层神经网络结构^[11]。该方法通过球查询分组和最远点采样(Farthest Point Sampling, FPS)构建局部邻域,优化了网络局部特征的学习能力。PointNet 和 PointNet++的出现使得使用深度卷积神经网络直接处理非结构化格式的整个点云变得流行。Jiang等^[24]提出一种 PointSIFT 模块,该模块通过 3 个阶段的卷积有效地堆叠和编码来自 8 个关键空间方向的数据。其能够灵活地嵌入到 PointNet 等现有模型中,实现精度的提升。PointWeb 从局部邻域中提取上下文特征,并通过自适应特征调整(Adaptive Feature Adjustment, AFA)模块增强点特征^[25]。

许多方法专注于开发显式卷积核以增强点的特征提取。PCCN(Parametric Continuous Convolutional Neural Networks)将卷积核表示为多层感知机(MLP)^[26]。SpiderCNN 将核定义为一组多项式函数,每个邻域点获得唯一的权重^[27]。Spherical CNN 通过设计球形卷积来解决三维旋转等变性问题^[28]。PointCNN 利用 X 变换将点重新排列为潜在且可能规范顺序,之后使用传统卷积算子进行有效的特征提取^[29]。

也有多种方法将图卷积神经网络作为增强点云分析中特征提取和整体网络性能的关键组件。DGCNN(Dynamic Graph CNN)使用 k 近邻(K-Nearest Neighbor, k-NN)算法构建动态邻域,并对这些邻域执行边缘卷积操作^[30]。ECC(Edge-Conditioned Convolution)设计了基于边缘标签的动态边缘条件滤波器,使点云分析中的特征提取更加灵活和自适应^[31]。DeepGCNs 深入研究了构建极深图卷积网络的方法,利用残差和膨胀卷积等 CNN 常用技术来增强网络深度并优化性能^[32]。

最近,在自然语言处理领域占据主导地位的 Transformer 在计算机视觉领域也取得了显著进展,展示了其卓越的全局特征学习能力。因此,其被广泛应用于各种点云处理任务中。PCT(Point Cloud Transformer)创新地用带有偏移注意力的原始 Transformer 块替换了 PointNet 的共享 MLP 层,并引入邻域信息嵌入以实现性能的提升^[33]。Point Transformer 利用向量自注意力操作的分层 Transform-

er 块从点云中提取局部特征, 并通过下采样模块降低分辨率, 最后通过全局平均池化操作提取全局特征^[13]。尽管这些纯 Transformer 架构取得了成功, 但其有一个显著缺点: 自注意力机制涉及大量线性变换, 可能导致信息冗余以及巨大的计算和内存开销。

2 本文算法

本文对 PointNet++ 模型的网络结构进行了改进, 使其能够有效地捕捉局部特征信息, 并且具备长程依赖建模能力, 从而提升模型性能。

2.1 PointNet++ 网络结构

PointNet++ 是一种通过反复堆叠集合抽象 (Set abstraction) 层实现层级特征提取的网络。Set abstraction 层分为采样、分组、聚合三层结构, 通过将点云划分为相互重叠的局部区域, 再对这些区域进行特征聚合以获得更高层次的特征, 实现类似 CNN 的效果。网络结构如图 1 所示。对于输入点云, 采样层采用 FPS 算法确定采样点, 然后在采样点固定半径内选取 k 个点形成邻域。分组层以邻域点与中心点的坐标差值为特征。聚合层使用 PointNet 对邻域应用共享 MLP 和最大池化将局部特征聚合到中心点。

2.2 CFE 模块

考虑到 PointNet++ 模型中的特征提取层仅限于对邻域特征进行简单的坐标信息提取, 而未能充分利用局部区域内的丰富信息。为了增强 PointNet++ 的特征提取能力, 本文设计了一个上下文特征提取 (CFE) 模块, 具体结构如图 2 所示。给定输入点云, 采用 FPS 算法生成点云的一个子集, 称为采样点集。对于采样点的邻域, 本文引入一种上下文特征提取方法, 能够有效编码并结合坐标和语义信息。该方法在先前的研究中已验证其有效性^[34]。

从右往左分为 3 个步骤: 采样、上下文特征提取和局部特征聚合。对于采样点 x_i 邻域内的每个邻点 x_j , 存在两种不同的上下文信息: 一种是空间上下文 P_j , 用于描述几何信息; 另一种是通过 1×1 的卷积运算编码的隐式特征上下文 F_j , 用于体现语义信息。为了全面表示这些上下文信息, 本文将 P_j 和 F_j 进行整合, 如式 (1) 所示:

$$C_j = P_j \oplus F_j \quad (1)$$

其中, C_j 是 x_j 的上下文表示, $\oplus$ 是拼接操作。基于这

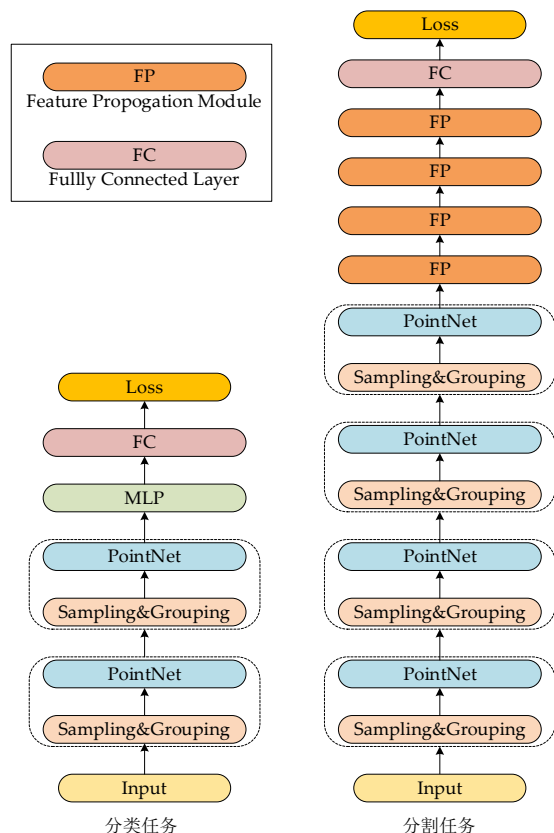


Fig. 1 The network architecture of PointNet++

图 1 PointNet++ 网络结构

些表示构建 x_i 与 x_j 之间的关系如式 (2) 所示:

$$\Delta C_{ij} = P_i \oplus P_j \oplus \|P_i - P_j\| \oplus (C_i - C_j) \quad (2)$$

其中, $\|P_i - P_j\|$ 为邻居点与中心点之间的欧几里得距离。通过式 (2) 可以得到初始的局部信息, 包含邻居点坐标、中心点、坐标差和欧几里得距离的几何信息及相减过的语义信息。其中, 语义信息相减可以更好地表示邻居点与中心点的关系。然后, 经过 PointNet 层将这些邻域信息聚集在中心点, 得到采样点局部邻域的特征表示。计算公式如式 (3) 所示:

$$y_i = \max_k (MLP(\Delta C_{ij})) \quad (3)$$

其中, MLP 是多层感知机, $\max_k$ 表示对采样点邻域内 k 个邻居进行最大池化。

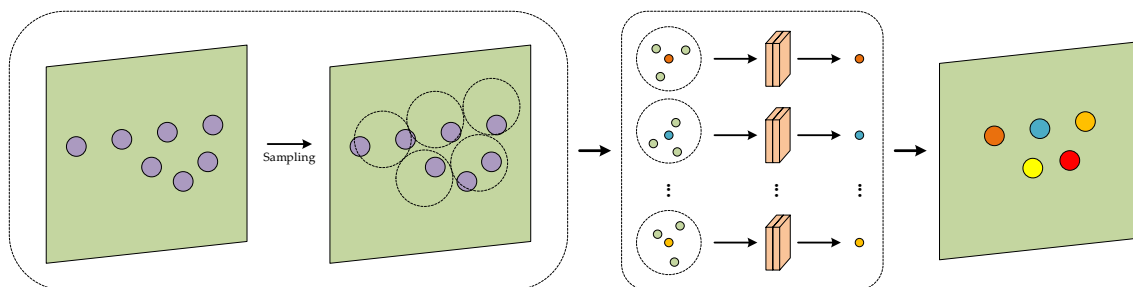


Fig. 2 The CFE block structure

图 2 CFE 模块结构

2.3 ML 模块

PointNet++模型仅通过分层提取局部邻域特征,而没有考虑对跨区域的全局关系进行建模。为了增强 PointNet++建模全局长程依赖关系的能力,本文在 PointNet 层后插入相互学习(ML)模块,以更全面地表达特征。ML 模块采用分组向量自注意力机制,并结合了可训练的位置编码。自注意力机制是注意力机制的一种特殊形式,通过计算序列中每个元素与其他所有元素的注意力权重,可直接建模任意两个位置之间的依赖关系,动态聚合具有判别性的信息。分组自注意力机制是注意力机制的一种改进变体,通过将输入数据划分为多个独立的组,在每组内进行自注意力计算。这种设计在保持自注意力核心优势的同时,针对性地解决了传统自注意力在计算效率、内存占用和局部特征处理上的不足。

以 PointNet 层的输出点云 $\{y_1, y_2, \dots, y_n\}$ 作为 ML 模块的输入,用 Y 表示,其维度为 $B \times N \times C$ 。其中, B 表示批量大小, N 表示每个样本的点云数量, C 表示通道数。首先使用 k -NN 算法为 Y 中的每个点找到 k 个邻居点,将其组织成 n 个不同的区域,然后对每个区域使用分组向量自注意力机制计算中心点与邻居点的关系。因为每个邻居点都表示一个局部邻域,所以这种点间的关系实际上是不同局部区域间的关系。通过这种方式,本文建模了 PointNet++ 所不具备的全局依赖关系。

向量自注意力是一种针对点云数据设计的注意力机制,旨在通道层面计算点与点间的关联权重。对于采样邻域 y_i 的附近邻域,经典的向量自注意力机制采用减法关系将输入向量在通道方向上相减,如式(4)所示:

$$z_i = \sum_{j=1}^k \rho \left(\gamma \left(\varphi(y_i) - \omega(y_j) + \delta \right) \right) \odot \left(\alpha(y_j) + \delta \right) \quad (4)$$

其中, z_i 是输出特征。 φ 、 ω 和 α 是线性层,通过 1×1 卷积实现。 δ 是位置编码函数, ρ 是归一化函数(如 softmax)。映射函数 γ 是一个包含两个线性层和一个 ReLU 非线性激活函数的 MLP。为了实现更轻量化和高效的版本,本文对上述自注意力计算方式进行改进,提出分组向量自注意力机制。具体而言,将 $\varphi(y_i)$ 和 $\omega(y_j)$ 沿通道方向分为 G 组,分别表示为 $\{\varphi_{i1}, \varphi_{i2}, \dots, \varphi_{iG}\}$ 和 $\{\omega_{j1}, \omega_{j2}, \dots, \omega_{jG}\}$ 。分组向量自注意力机制计算公式如式(5)所示:

$$z_i = \sum_{j=1}^k \rho \left(\gamma \left(\sum_{g=1}^G (\varphi_{ig} - \omega_{jg}) + \delta_1 \right) \right) \odot \left(\alpha(q_j) + \delta_2 \right) \quad (5)$$

其中, δ_1 和 δ_2 表示可训练的位置编码。通过分组累加的方法可以有效放大有用信息,最终提取出更具辨识度的特征。

位置编码在自注意力机制中起着至关重要的作用,能够使算子自适应地捕捉数据中的局部结构^[35]。因此,在向量注意力机制中也引入可学习的位置编码,以有效融合局部空间信息。设 $P = \{p_1, p_2, \dots, p_n\}$ 表示编码邻域内点坐标的向量集合,位置编码函数定义如式(6)所示:

$$\delta = \theta \begin{bmatrix} p_1 - p_1 & p_1 - p_2 & \dots & p_1 - p_n \\ p_2 - p_1 & p_2 - p_2 & \dots & p_2 - p_n \\ \vdots & \vdots & \ddots & \vdots \\ p_n - p_1 & p_n - p_2 & \dots & p_n - p_n \end{bmatrix} \quad (6)$$

式(6)中, θ 是一个由两个线性层组成的 MLP,中间加入批量归一化和 ReLU 激活函数。在式(5)中, δ_1 和 δ_2 用于将三维坐标信息转换为相应维度,以便进行通道求和。为了提高网络学习过程的选择性和适应性,并防止网络退化,将模块输出与原始输入进行了残差连接。ML 模块结构如图 3 所示。

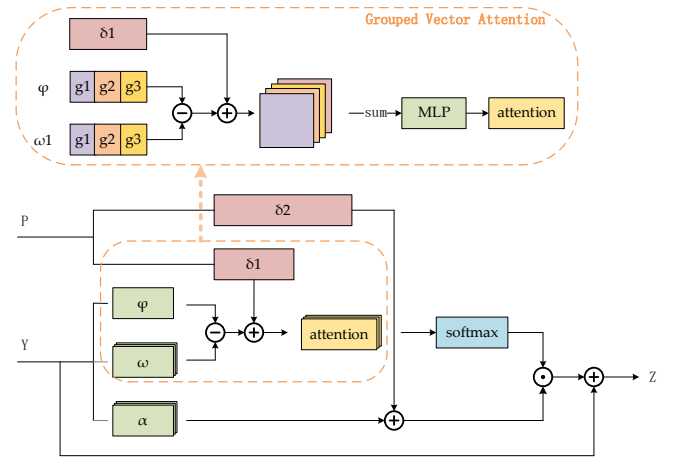


Fig. 3 The ML block structure

图 3 ML 模块结构

2.4 网络结构

网络整体流程如图 4 所示。以原始点云作为输入,网络可以分为两个主要部分:前端和后端。前端是网络的核心组件,负责特征提取。对于分类任务,前端包含 2 个块;对于分割任务,前端则包含 4 个块。前端中的每个块都包含 CFE 模块、ML 模块以及网络原有的 PointNet 层。

在分类任务中,配置两个模块的采样点数量分别为 $N/2$ 和 $N/8$ 。其中, N 表示输入点的总数。之后,后端组件使用一个 MLP 层将提取的特征增强到 1 024 维,通过全局最大池化操作得到目标点云的最终全局特征表示。最后,通过一个全连接(Fully Connected, FC)层计算全局分类结果。该 FC 层由 3 个线性层组成,并包含批量归一化和 ReLU 激活函数。在分割任务中,配置 4 个模块的采样点数量分别为 N 、 $N/4$ 、 $N/8$ 和 $N/16$ 。后端组件由 4 个特征传播(Feature Propagation, FP)模块和 2 个线性层组成,同样包含批量归一化和 ReLU 激活函数。

3 实验与结果分析

本文通过多个基准数据集对模型进行评估。对于分割任务,本文采用斯坦福大规模三维室内空间(S3DIS)数据集^[36],随后将实验范围扩展到 ModelNet40^[20]数据集上的分类任务,并开展消融实验以进一步验证本文模型。所

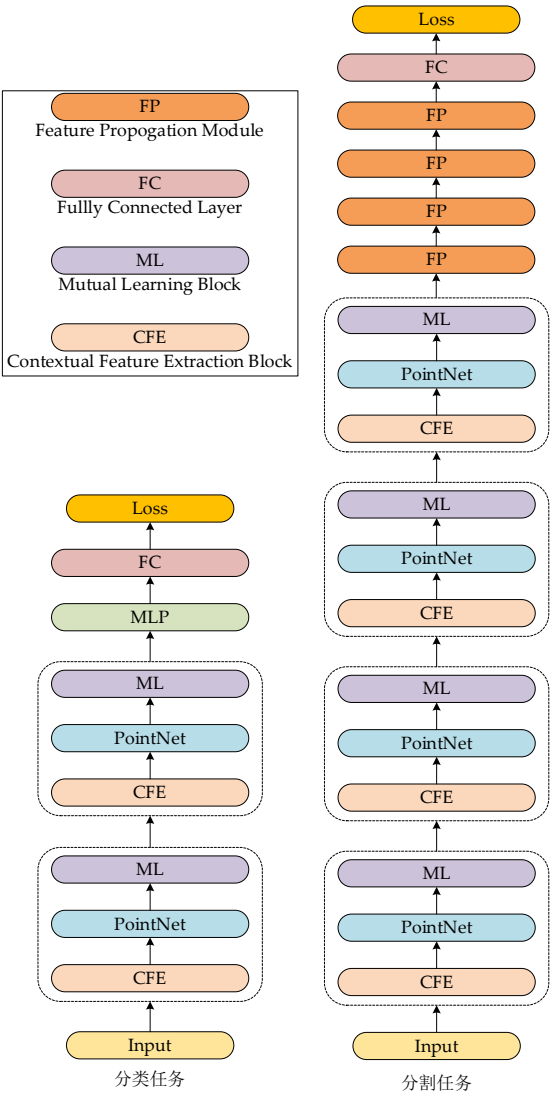


Fig. 4 The network architecture
图 4 网络结构

有实验使用的 CPU 为 Intel Xeon (R) Platinum 8350C 2.6GHz, GPU 为 NVIDIA GeForce RTX 4090 GPU, 操作系统为 Ubuntu, 版本为 20.04, 并采用 Python3.8 编程语言, PyTorch1.8.1 深度学习框架, CUDA 版本为 11.4。

3.1 室内场景语义分割

S3DIS 数据集包含来自 3 栋不同建筑的 6 个区域, 涵盖办公室、会议室、教室、走廊、洗手间等多种场景, 有 271 个房间, 共有 13 个类别的语义标签(天花板、地板、墙、桌子等), 场景中的每个点都对应了其中一个类别。模型将每个场景作为一个样本, 以场景点空间坐标及其 RGB 信息作

为网络输入特征, 并从每个样本中随机选择 4 096 个点云进行训练。采用 6 折交叉验证方法评估 6 个区域训练的模型获得整体平均分割精度, 以便于和其他方法比较。具体参数设置如下: 选择 AdamW 优化器作为优化方法, 动量与权重衰减分别设置为 0.9 和 0.7; 初始学习率设置为 0.001, 并采用固定步长衰减方法控制学习率变化; 步长设置为 10, 批量大小设置为 16, 开展 48 个 epoch 训练, CFE 模块和 ML 模块数量设置为 4。通过平均类别交并比(mean Intersection over Union, mIoU)和整体准确率(Overall Accuracy, OA)来全面评估模型性能。实验结果如表 1 所示, 可以看出本文模型相比基准 PointNet++模型的分割结果, 在 OA 和 mIoU 上分别提升了 4.2% 与 6.5%; 相比其他网络(PointNet, 3DRCNN, DGCNN, DeepGCN), mIoU 保持最优, 且高于次优者 4.5%, OA 虽为次优, 但与最优者仅相差 0.5%。考虑到 mIoU 是评价语义分割性能的核心指标, 因此本文模型在整体性能上明显优于其他 3D 点云场景语义分割模型。

Table 1 Comparison of semantic segmentation results on the S3DIS dataset, evaluated with 6-fold cross-validation

表 1 在 S3DIS 数据集上使用 6 倍交叉验证评估的语义分割结果比较 (%)

Method	OA	mIoU
PointNet ^[10]	78.5	47.6
3DRCNN ^[38]	85.7	53.4
PointNet++ ^[11]	81.0	54.5
DGCNN ^[30]	84.1	56.1
RSNet ^[37]	—	56.5
DeepGCN ^[32]	85.9	60.0
ours	85.2	61.0

为进一步验证本文提出两个模块的有效性, 以 S3DIS 数据集场景区域 1—区域 5 为训练样本, 以 PointNet++为基础设计 CFE 模块和 ML 模块的消融实验, 用于分析各模块对区域 6 的语义分割精度影响。实验结果如表 2 所示, 可以看出, 相比基线模型 70.2% 的 mIoU, 集成 CFE 模块与 ML 模块的模型分别获得了 4.0% 和 5.9% 的显著提升。当两个模块同时引入时, 本文完整模型的 mIoU 进一步提升至 77.3%, 较基线模型提升 7.1%。CFE 模块在柱(column)类别上相比 ML 模块更有优势, 原因在于柱是细长形状物体, 顶部和底部距离较远, 而 ML 模块是通过 k-NN 算法计算三维空间坐标提取邻域, 对细长物体存在一定局限性。但 ML 模块在其他方面都优于 CFE 模块, 说明增强 PointNet++ 的全局学习能力相比增强特征学习能力效果更明显。针

Table 2 Semantic segmentation results of Area 6 on the S3DIS dataset
表 2 在 S3DIS 数据集上区域 6 的消融实验结果

Method	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter	mIoU/%	OA/%
PointNet++ ^[11]	93.0	97.3	74.8	68.7	43.2	77.8	78.9	72.4	76.8	41.9	58.7	66.2	63.2	70.2	87.7
+CFE	95.7	97.2	79.2	77.5	49.7	78.6	79.5	74.7	77.7	50.9	65.2	72.3	67.1	74.2	89.3
+ML	95.7	97.5	79.7	79.0	47.6	80.6	82.4	75.3	82.4	54.9	67.9	79.2	67.8	76.1	90.0
ours	95.9	97.5	84.3	80.0	63.9	79.1	85.5	73.0	77.2	55.9	67.6	78.6	66.3	77.3	90.8

对椅子(chair)、杂物(clutter)和桌子(table)这类局部结构比较复杂的物体,ML模块展现出强大的性能。消融实验的类别精度分析表明,CFE与ML模块在多数类别上均有显著提升,其中在柱类别上的提升尤为突出,分别达到14.2%和16.3%的mIoU增益。完整模型虽在个别类别上对ML模块的优势有所权衡,但最终实现了OA与mIoU的整体最优,验证了两个模块结合的有效性。

整体来看,本文模型针对柱(column)和沙发(sofa)的mIoU精度偏低,主要原因是样本不平衡。在S3DIS数据集中,柱占总体点云数的1.5%,沙发占总体点云数的0.4%,模型从有限的样本中学习特征十分困难。梁(beam)和柱虽然在几何结构上有一定相似性,但柱是直立的,观测更容易受到光照等条件影响,导致靠近观测点的点云密度相对稀疏,远离观测点的点云较为密集。

3.2 三维物体分类

ModelNet40是三维形状分类任务的基准数据集,包含12 311个多样化的CAD模型,涵盖40种常见物体类别(椅子、桌子、飞机、汽车等)^[10]。ModelNet40从每个模型的网格表面均匀采样1 024个点,随后将其归一化到一个球体中。本文使用9 843个形状进行训练,并保留2 468个形状用于测试,批量大小设置为24,并进行200个epoch的训练。CFE模块和ML模块数量设置为2,以1 024个点的三维空间坐标作为输入。物体分类结果如表3所示,可以看出,本文模型在该数据集上超越了大多数现有模型,OA获得了显著提升。与经典方法相比,本文模型的OA较ShapeNet与VoxNet分别提升了8.0%和6.8%。以PointNet++为基线,本文模型实现了92.7%的OA和90.2%的平均准确率(mean Accuracy, mAcc),分别提升了2.0%和2.6%。为进一步验证每个模块的有效性,本文进行了消融实验。结果表明,CFE模块使OA提高了1.1%,mAcc提高了1.7%,而ML模块使OA提高了1.6%,mAcc提高了2.3%,充分证明了该模块的有效性。此外,本文方法的OA相比SpiderCNN提高了0.3%,进一步证实了其竞争力。

Table 3 Comparison of classification results on ModelNet40 dataset

表3 在ModelNet40数据集上分类结果比较

Method	Input	mAcc/%	OA/%
3DShapeNets ^[20]	1 024	77.3	84.7
VoxNet ^[21]	—	83.0	85.9
PointNet ^[11]	1 024	86.2	89.2
MVCNN ^[18]	—	—	90.1
PAT ^[39]	1 024	—	91.7
PointCNN ^[29]	1 024	88.1	92.2
DGCNN ^[30]	1 024	90.2	92.2
SpiderCNN ^[27]	1 024	—	92.4
PointNet++ ^[11]	1 024	87.6	90.7
PointNet++(+CFE)	1 024	89.3	91.8
PointNet++(+ML)	1 024	89.9	92.3
ours	1 024	90.2	92.7

4 结语

本文提出一种基于上下文特征和点集互学习的PointNet++网络,旨在提升PointNet++的性能。以原始点云作为输入,首先通过CFE模块提取点云的上下文特征,增强了基准网络对采样点邻域结构的学习能力,之后利用ML模块进一步整合这些特征,使用分组向量自注意力机制使采样点能够相互学习。通过在ModelNet40数据集上进行消融实验,并在不同数据集上与其他方法进行比较,证明了该方法在处理原始点云数据时的高效性和鲁棒性。在处理点云任务时,本文提出的网络模型相比大多数现有网络,表现相当或更优。虽然本文方法改进了PointNet++模型,但在特征提取过程中,CFE模块提取的空间信息只有坐标值、坐标间差、欧式距离,可能导致空间信息提取不完全。因此,在未来的研究中会考虑加入法向量、曲率或角度等特征的提取。此外,本文使用的FPS算法可能在高密度区域选择过多冗余点,而在低密度区域遗漏关键点,未来会根据现有研究寻找一种更适合的点云采样算法,并进一步提升模型性能。

参考文献:

- [1] ZHU Y K, MOTTAGHI R, KOLVE E, et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning[C]//2017 IEEE International Conference on Robotics and Automation, 2017: 3357-3364.
- [2] QI C R, LIU W, WU C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 918-927.
- [3] RUSU R B, MARTON Z C, BLODOW N, et al. Towards 3D point cloud based object maps for household environments[J]. Robotics and Autonomous Systems, 2008, 56(11): 927-941.
- [4] LI S, LEE D. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 11927-11936.
- [5] REN Z L, WANG L G, BI L. Robust GICP-based 3D LiDAR SLAM for underground mining environment[J]. Sensors, 2019, 19(13): 2915.
- [6] MOKHAYERI F, GRANGER E. A paired sparse representation model for robust face recognition from a single sample[J]. Pattern Recognition, 2020, 100: 107129.
- [7] ZHANG L Q, ZHANG L. Deep learning-based classification and reconstruction of residential scenes from large-scale point clouds[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 56(4): 1887-1897.
- [8] CHEN M Y, TANG Y C, ZOU X J, et al. High-accuracy multi-camera reconstruction enhanced by adaptive point cloud correction algorithm[J]. Optics and Lasers in Engineering, 2019, 122: 170-183.
- [9] WU F Y, DUAN J L, CHEN S Y, et al. Multi-target recognition of bananas and automatic positioning for the inflorescence axis cutting point[J]. Frontiers in Plant Science, 2021, 12: 705021.
- [10] QI C R, SU H, MO K, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//Proceedings of the IEEE Confer-

- ence on Computer Vision and Pattern Recognition, 2017: 652–660.
- [11] QI C R, YI L, SU H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5105–5114.
- [12] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.
- [13] ZHAO H S, JIANG L, JIA J Y, et al. Point transformer [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 16259–16268.
- [14] BOULCH A, LE SAUX B, AUDEBERT N. Unstructured point cloud semantic labeling using deep segmentation networks [C]// Proceedings of the Workshop on 3D Object Retrieval, 2017: 17–24.
- [15] LAWIN F J, DANELLJAN M, TOSTEBERG P, et al. Deep projective 3D semantic segmentation [C]// International Conference on Computer Analysis of Images and Patterns, 2017: 95–107.
- [16] ZHANG L, SUN J, ZHENG Q. 3D point cloud recognition based on a multi-view convolutional neural network [J]. Sensors, 2018, 18(11): 3681.
- [17] GAO Q, SHEN X. ThickSeg: efficient semantic segmentation of large-scale 3D point clouds using multi-layer projection [J]. Image and Vision Computing, 2021, 108: 104161.
- [18] SU H, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3D shape recognition [C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 945–953.
- [19] TATARCHENKO M, PARK J, KOLTUN V, et al. Tangent convolutions for dense prediction in 3D [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3887–3896.
- [20] WU Z, SONG S, KHOSLA A, et al. 3D ShapeNets: a deep representation for volumetric shapes [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1912–1920.
- [21] MATURANA D, SCHERER S. VoxNet: a 3D convolutional neural network for real-time object recognition [C]// 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2015: 922–928.
- [22] GRAHAM B, ENGELCKE M, VAN D M L. 3D semantic segmentation with submanifold sparse convolutional networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 9224–9232.
- [23] RIEGLER G, OSMAN ULUSOY A, GEIGER A. OctNet: learning deep 3D representations at high resolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3577–3586.
- [24] JIANG M Y, WU Y R, ZHAO T Q, et al. PointSIFT: a SIFT-like network module for 3D point cloud semantic segmentation [DB/OL]. <https://arxiv.org/pdf/1807.00652>.
- [25] ZHAO H S, LI J, FU C W, et al. PointWeb: enhancing local neighborhood features for point cloud processing [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5565–5573.
- [26] WANG S L, SUO S, MA W C, et al. Deep parametric continuous convolutional neural networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2589–2597.
- [27] XU Y F, FAN T Q, XU M Y, et al. SpiderCNN: deep learning on point sets with parameterized convolutional filters [C]//Proceedings of the European Conference on Computer Vision, 2018: 87–102.
- [28] DEFFERRARD M, MILANI M, GUSSET F, et al. DeepSphere: a graph-based spherical CNN [DB/OL]. <https://arxiv.org/pdf/2012.15000>.
- [29] LI Y Y, BU R, SUN M C, et al. PointCNN: convolution on X-transformed points [J]. Advances in Neural Information Processing Systems, 2018, 31: 828–838.
- [30] WANG Y, SUN Y B, LIU Z W, et al. Dynamic graph CNN for learning on point clouds [J]. ACM Transactions on Graphics, 2019, 38(5): 1–12.
- [31] SIMONOVSKY M, KOMODAKIS N. Dynamic edge-conditioned filters in convolutional neural networks on graphs [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3693–3702.
- [32] LI G H, MULLER M, THABET A, et al. DeepGCNs: can GCNs go as deep as CNNs? [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9267–9276.
- [33] GUO M H, CAI J X, LIU Z N, et al. PCT: point cloud transformer [J]. Computational Visual Media, 2021, 7: 187–199.
- [34] QIU S, ANWAR S, BARNES N. PU-Transformer: point cloud upsampling transformer [C]//Proceedings of the Asian Conference on Computer Vision, 2022: 2475–2493.
- [35] 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, 2017: 6000–6010.
- [36] ARMENI I, SENER O, ZAMIR A R, et al. 3D semantic parsing of large-scale indoor spaces [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1534–1543.
- [37] HUANG Q G, WANG W Y, NEUMANN U. Recurrent slice networks for 3D segmentation of point clouds [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2626–2635.
- [38] YE X, LI J, HUANG H, et al. 3D recurrent neural networks with context fusion for point cloud semantic segmentation [C]//Proceedings of the European Conference on Computer Vision, 2018: 403–417.
- [39] YANG J C, ZHANG Q, NI B B, et al. Modeling point clouds with self-attention and gumbel subset sampling [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3323–3332.

(责任编辑: 黄 健)