

基于多模态学习的点云分类网络

王子澳^{1,2}, 周国鹏^{2,3}, 张 Jian 权^{2,4}, 刘煌坤²

- (1. 武汉纺织大学 计算机与人工智能学院, 湖北 武汉 430200;
2. 湖北香城智能机电研究院有限公司, 湖北 咸宁 437000;
3. 武汉纺织大学 电子与电气工程学院, 湖北 武汉 430200; 4. 湖北科技学院 自动化学院, 湖北 咸宁 437000)

摘要: 针对由于点云数据的无序性、稀疏性和复杂性导致三维点云分类网络参数量庞大、训练效率低下的问题, 提出一种基于多模态学习的点云分类网络模型 MPC-CLIP。首先, 将原始点云数据从多个视角投影成二维深度图; 其次, 利用二维 CLIP 中训练好的文本编码器和图像编码器分别提取点云的文本描述特征和投影图像特征, 从训练好的模型中探索文本特征与图像特征之间的对应关系; 最后, 将每个图像特征与所有文本特征进行加权求和并计算余弦相似度, 进而得到最终的零样本分类结果。使用经典点云数据集 ModelNet40 进行零样本分类与消融实验, 并且将 MPC-CLIP 模型与多个经典 3D 点云分类模型进行融合。实验结果表明, 所提模型与其他三维点云分类网络模型融合后平均分类精确度提升了 0.9%~1.9%, 总体分类精确度提升了 0.8%~1.7%, 证明了其有效性和鲁棒性。

关键词: 点云处理; 多模态学习; CLIP; 零样本分类

DOI: 10.11907/rjdk.241530

中图分类号: TP391

文献标识码: A

开放科学(资源服务)标识码(OSID):

文章编号: 1672-7800(2025)008-0043-06



Point Cloud Classification Network Based on Multimodal Learning

WANG Zi'ao^{1,2}, ZHOU Guopeng^{2,3}, ZHANG Jianquan^{2,4}, LIU Huangkun²

- (1. School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China;
2. Hubei Xiangcheng Intelligent Electromechanical Industry Technology Research Institute Co., Ltd., Xianning 437000, China;
3. School of Electronics and Electrical Engineering, Wuhan Textile University, Wuhan 430200, China;
4. Institute of Engineering and Technology, Hubei University of Science and Technology, Xianning 437000, China)

Abstract: A multi-modal learning based point cloud classification network model MPC-CLIP is proposed to address the problem of large parameter quantities and low training efficiency in 3D point cloud classification networks due to the disorder, sparsity, and complexity of point cloud data. Firstly, project the original point cloud data into a two-dimensional depth map from multiple perspectives; Secondly, using the text encoder and image encoder trained in two-dimensional CLIP, the text description features of the point cloud and the image features of the projected image are extracted respectively, and the correspondence between text features and image features is explored from the trained model; Finally, each image feature is weighted and summed with all text features, and cosine similarity is calculated to obtain the final zero sample classification result. Conduct zero sample classification and ablation experiments using the classic point cloud dataset ModelNet40, and fuse the MPC-CLIP model with multiple classic 3D point cloud classification models. The experimental results show that after integrating the proposed model with other 3D point cloud classification network models, the average classification accuracy improved by 0.9% to 1.9%, and the overall classification accuracy improved by 0.8% to 1.7%, demonstrating its effectiveness and robustness.

Key Words: point cloud processing; multimodal learning; CLIP; zero-shot classification

收稿日期: 2024-06-16

扫描二维码阅读全文:



基金项目: 湖北省科技创新重大专项(2023EGA023, 2022BBA026, 2021BGD022, 2020BGC028)

作者简介: 王子澳(2000-), 男, CCF 学生会员, 武汉纺织大学计算机与人工智能学院硕士研究生, 研究方向为计算机视觉; 周国鹏(1972-), 男, 博士, 武汉纺织大学电子与电气工程学院教授、硕士生导师, 研究方向为智能控制系统; 张 Jian 权(1987-), 男, 博士, 湖北科技学院自动化学院副教授, 研究方向为机器视觉与人工智能。本文通讯作者: 周国鹏。

0 引言

随着深度学习技术的不断发展,三维点云被广泛应用于自动驾驶、机器人定位、遥感、增强现实和虚拟现实等领域的视觉任务中^[1-3]。目前,三维点云处理方法分为传统方法和学习方法两大类。传统方法包括几何分析处理和特征描述子提取等,前者主要通过估计点云的法向量、曲率、密度等几何信息来分析点云数据,后者则通过传统特征描述子获取对应区域点云的几何信息与属性。传统方法的准确度较高,但是对大幅度旋转等复杂点云场景的处理效果不够理想,而且涉及的函数计算极为繁琐。学习方法可分为机器学习方法和深度学习方法两种。机器学习方法能够处理三维数据,学习有效特征信息,但是三维数据量和训练样本数量极大,增加了机器学习方法的训练难度。此外,机器学习需要专家识别特征,特征提取的准确度会直接影响其性能,然而三维数据标记不仅体量庞大,而且难度也远大于二维数据^[4]。深度学习方法能够训练和计算大型数据,自主学习输入信息中的潜在空间特征与高级规律,适用于处理海量点云数据。该类方法通常采用端到端的架构,对输入的数据信息进行处理分析后输出相应结果,虽然耗费了大量时间进行训练,但是相较于机器学习方法的分块解决方式时间更少,可一次性得到更精准的预测结果^[5]。

根据编码表征方式不同,三维数据可分为点云、视图、体素等多种模态。由于点云数据具有无序性、不规则性等难以处理的性质,早期面向三维模型识别的深度学习方法首先将点云转化为规则的平面视图、体素等模态,然后利用较为成熟的神经网络框架提取特征完成特定任务^[6-8]。然而,该转化过程造成了几何信息的丢失,限制了网络性能。为避免这一问题,Qi等^[9]开创性地提出一种直接处理点云数据的方法 PointNet,其通过多层感知机(Multilayer Perceptron, MLP)学习单个点的特征并使用对称函数来应对点云的无序性,使用三维的空间变换网络来应对点云的旋转不变性,使用最大池化聚合点特征来应对点云的置换不变性,该方法在分类和分割任务上均取得良好效果^[10-12]。然而,独立地学习每个点的特征会导致局部结构信息丢失,极大限制了网络性能。为此,Qi等^[13]对三维点云进行局部划分,提出提取局部特征的 PointNet++。Zhao等^[14]提出 PointWeb,通过自适应特征调整(Adaptive Feature Adjustment, AFA)模块整合点云上下文特征信息,进一步提升了对局部特征的学习能力。吴军等^[15]在 PointNet++中引入中心自注意力模块,并使用注意力池化、差异性池化和最大池化3种池化方式更有效地提取全局特征和局部特征。

近年来,大规模视觉语言预训练模型,如 CLIP(Contrastive Language-Image Pre-training)、ViLBERT、VL-BERT

等,在二维零样本和开放词汇任务中被广泛应用^[16]。CLIP通过大规模图像—文本对进行预训练,通过独立的编码器提取输入图像和文本的特征,并将配对的图像和文本对齐在同一嵌入空间中。在下游任务中,给定一个包含“看不见”类图像的新数据集,CLIP按照类别名称构造文本输入,并将原始分类任务转换为图像文本匹配问题。因此,CLIP能够在开放词汇设置中实现零样本识别,并在各种基准测试中取得不错性能。受自然语言处理中提示调优和适配器的启发,上下文优化 CLIP-Adapter 和 Tip-Adapter冻结了 CLIP的预训练权重并采用可学习的提示或轻量级适配器来调整文本和视觉功能^[17-20]。在3D计算机视觉领域,大多数点云预训练方法所用到的点云数据集规模相对较小,如 ShapeNet 和 ScanNet 数据集^[21-22]。其中,ShapeNet 数据集包含约 50 000 个合成 3D 物体数据,ScanNet 数据集包含约 1 700 个室内房间中的真实扫描数据。与 2D 图像领域中的先进预训练方法 CLIP 所使用的训练数据量相比,3D 点云的训练样本数量不及其千分之一,在数据规模上存在数个数量级的差距。因此,3D 点云预训练方法从小规模的训练样本中学习到的先验知识较为有限,导致其泛化能力和表征能力与 2D 图像模型存在一定差距。

综上所述,现有基于深度学习的点云分类方法具有一定缺陷,如未关注到点云的局部信息特征、未考虑到点与点之间的联系以及缺乏对点云二维信息特征的提取等,这些缺陷使分类网络性能受到限制。本文针对现有三维点云分类网络模型缺乏对点云二维特征信息的提取这一问题,借助二维 CLIP的思路设计出一种基于三维点云投影图像的点云分类网络 MPC-CLIP(Multimodal Point Cloud-CLIP)。该网络首先通过文本编码器提取点云的文本描述特征;其次通过图像编码器提取点云二维投影图像特征;最后通过计算相关特征的最大相似度来对点云特征进行分类。实验结果表明,在不同的经典点云分类模型中加入 MPC-CLIP能使点云分类网络提取到点云数据的二维信息特征,实现多模态学习,从而提高分类精度。

1 CLIP模型

CLIP是一个图像—文本预训练模型,由 OpenAI 开发,旨在通过自然语言监督学习可转移的视觉模型。CLIP的核心思想为通过同时训练图像编码器和文本编码器,使其能够正确预测图像与文本描述的配对,从而在一个共享的嵌入空间中比较图像和文本表示。这种方法允许模型在没有针对特定类别训练示例的情况下执行零样本分类任务。CLIP的训练过程通过对比学习实现,模型需在批次内区分真实匹配的图文对与随机组合的负样本对,从而在共享嵌入空间中学习语义对齐。通过最大化实际配对嵌入之间的余弦相似度并最小化错误配对之间的余弦相似度,

CLIP 学习到一个多模态嵌入空间。这种训练方法优化了基于相似度得分的对称交叉熵损失,实现了从自然语言监督中高效学习可转移视觉模型的目的。CLIP 模型使用两个编码器分别处理文本和图像数据,其中文本编码器(Text Encoder)使用 Text Transformer 模型提取文本特征,图像编码器(Image Encoder)使用 ResNet 和 Vision Transformer 提取图像特征^[23-24]。模型结构如图 1 所示。

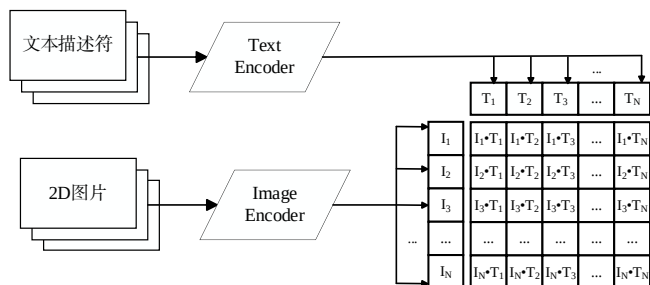


Fig. 1 Structure of CLIP model

图 1 CLIP 模型结构

CLIP 模型的输入为 1 个图像与文本的配对组合,图像和文本会分别经过图像编码器和文本编码器处理得到相应的特征。例如,有 X 个图像和 X 个文本描述符便会得到 X 个图像特征和 X 个文本特征,然后在这些特征的基础上进行对比学习。CLIP 模型只需要一个正样本和负样本的定义,正样本是配对好的图像文本对,它们所描述的物体是一致的,因此在特征矩阵中,对角线上相匹配的均为正样本。区分好正负样本后,模型便可以通过对比学习的方式进行训练。CLIP 模型在训练过程中不需要任何手工标

注,但是需要大量数据集作为支撑。OpenAI 团队收集了 4 亿个图像与文本的配对来训练 CLIP 模型,使其获得了强大的性能。

要想使用 CLIP 模型对数据集进行分类,最关键的就是获取文本特征和图像特征。首先,需要根据任务的分类标签或自定义标签构建出数据集中每个类别的表述文本,将这些文本送入文本编码器以获取相应的文本特征,文本特征数取决于数据集的类别个数;其次,将待预测的类别图像送入图像编码器以获取相应的图像特征;最后,对文本特征和图像特征计算余弦相似度,拥有最大相似度的文本对应的类别即为图像分类预测结果。

2 MPC-CLIP 模型

CLIP 预训练模型在二维图像分类方面展现出强大的性能优势。本文提出的 MPC-CLIP 模型以 CLIP 为基础,将其预训练的知识迁移到三维点云分类任务中,试图在点云分类过程中利用更多二维图像特征,建立一个多模态学习框架,从而提高三维点云网络模型的分类准确度。MPC-CLIP 模型结构如图 2 所示。该模型首先利用二维 CLIP 中训练好的文字编码器和图像编码器分别提取点云的文本描述特征和点云投影图像的特征,从预训练好的模型中找到文本特征与图像特征之间的对应关系;其次将每个图像特征与所有文本特征进行加权求和并计算余弦相似度,其中最大相似度文本对应的类别即为最终分类结果。该模型不需要进行预训练便能直接完成分类任务。

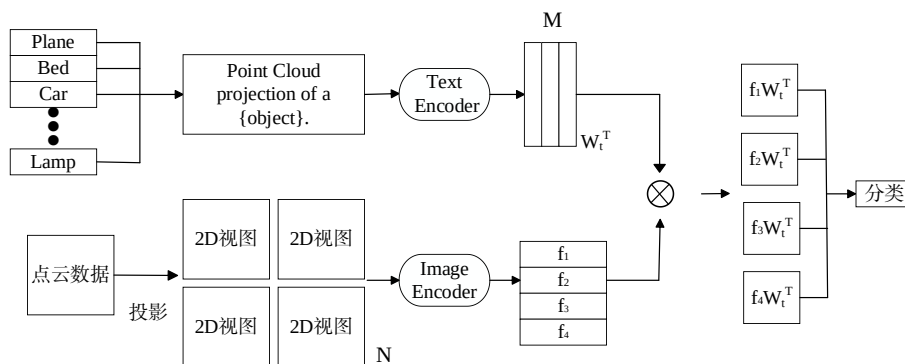


Fig. 2 MPC-CLIP model structure

图 2 MPC-CLIP 模型结构

2.1 文本特征提取

本文采用 ModelNet40 数据集,该数据集包含 40 个类别(如飞机、汽车、植物、灯具)的 12 311 个由 CAD 生成的网格,其中 9 843 个用于训练,2 468 个用于测试。ModelNet40 数据集为 3D 对象分类与检索、无训练 3D 点云分类、少样本点云分类、3D 对象识别、3D 点云分类、少样本 3D 点云分类、生成式 3D 对象分类以及 3D 点云线性分类等多种任务的基准之一。本文首先对 ModelNet40 数据集中的每个物体类别构建一个适配的描述文本;其次将这些描述文本输入文字编码器(Text Encoder)中提取每个类别对应的

M 个文本特征。ModelNet 数据集共有 40 个类别,因此经过特征提取后会获得 40 个文本特征,即 $M=40$,经过文本编码器提取后的文本特征可以表示为 $W_t \in R^{M \times C}$ 。模型中的文字编码器为 text-transformer,提取文本特征的主要流程如图 3 所示。

文字编码器的输入为 X ,表示一个描述语句,经过编码器多个模块提取特征后得到该描述语句最终的文本特征。Input Embedding 模块将句子中的每个字符转化为向量来表示, X 经过该模块后可以转换为 $X_{embedding}$ 向量,该向量的 3 个维度分别表示句子数量、句子字数以及每个字的维度。

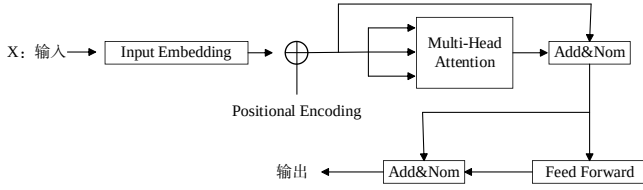


Fig. 3 Text feature extraction flow

图3 文本特征提取流程

Positional Encoding 为位置编码模块,该模块可以对输入句子中每个文字的位置信息进行编码与标记,并使用正余弦函数编码计算过程。计算公式为:

$$PE(pos, 2i) = \sin(pos/10000^{2i/d_{model}}) \quad (1)$$

$$PE(pos, 2i + 1) = \cos(pos/10000^{2i/d_{model}}) \quad (2)$$

式中: pos 表示每个字在句子中的位置; i 表示词向量的维度。

经过位置编码模块后可以得到一个与输入句子维度相同的编码数组 X_{pos} , 其与原来的向量叠加可以获得新的词向量。表示为:

$$X_{embedding} = X_{embedding} + X_{pos} \quad (3)$$

Multi-Head Attention 为多头注意力机制模块,可使模型学到多重含义的表达。该模块采用 self-attention 注意力机制,对输入进行线性映射得到 Q, K, V 。计算公式为:

$$Q = X_{embedding} * W_Q \quad (4)$$

$$K = X_{embedding} * W_K \quad (5)$$

$$V = X_{embedding} * W_V \quad (6)$$

式中: Q, K, V 的维度与 $X_{embedding}$ 的维度相同。

多头注意力机制将维度矩阵切割为多份,每一份都是一个头,经过注意力机制处理后的每个字向量都具有当前句子中所有字向量的信息。

Add & Norm 模块主要进行残差连接和归一化两种操作。其中,残差连接是将上一层的输入 X 与输出相加;归一化是将每一行的元素减去这一行的均值并除以该行的标准差,得到归一化后的数值,从而将神经网络的隐藏层变为标准正态分布,加快网络收敛。

Feed Forward 模块中包含两层线性映射并且使用激活函数进行激活,经过相同的 Add & Norm 操作后得到最终输出。

2.2 图像特征提取

由于 CLIP 模型在使用 Image Encoder 提取图像特征时所输入的图像都是二维的,需要对三维点云数据进行二维化处理,具体操作为从多个视角将 ModelNet40 数据集中的三维点云数据投影成二维深度图。对于数据集中的三维点云数据,其空间坐标可以用 (x, y, z) 来表示,对 z 方向进行投影时可以将这个点变换为 $([x/z], [y/z])$, 这种投影方式的优点是可使投影图像比较接近自然图像。CLIP 模型的图像编码器处理的是三通道 RGB 图像,为了能在投影图像上提取点云相关特征,将该图像复制两次变成三通道图像后再输入图像编码器。三维点云数据点 A

投影到二维坐标系点 B 的映射公式为:

$$\vec{A} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \rightarrow \vec{B} = \begin{bmatrix} \alpha \cdot \frac{x}{z} + C_x \\ \beta \cdot \frac{y}{z} + C_y \end{bmatrix} \quad (7)$$

MPC-CLIP 模型采用与 CLIP 模型相同的 ResNet 网络进行特征提取。ResNet 网络包含 Identity Block 和 Conv Block 两个基本模块。结构如图4所示。

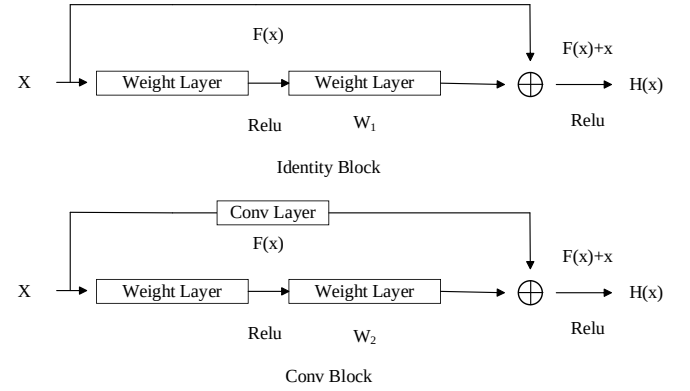


Fig. 4 ResNet structure

图4 ResNet 结构

X 为 Identity Block 模块的输入,通过残差模块后的输出为 $H(x)$ 。表示为:

$$H(x) = F(x, \{w_i\}) + x \quad (8)$$

式中: $F(x, \{w_i\})$ 表示残差,是要学习的目标,表示权重与输入之间的运算关系, $F(x) = H(x) - x$ 。残差部分可以表示为中间有一个 ReLU 激活的双层权重:

$$F = w_2 \sigma(w_1 x) \quad (9)$$

式中: w_1 和 w_2 分别为第一层和第二层的权重; σ 表示 ReLU 激活函数。

Identity Block 模块的输出和输入必须具有相同的形状,否则输入与残差边不能相加。Conv Block 模块在 Identity Block 模块的基础上增加了 Conv layer 卷积操作,可以对输入矩阵的形状进行调整,使模块中的残差边与卷积可以相加。

对于 N 个不同视角下投影得到的深度图像,使用 Image Encoder 来提取图像特征,提取到的图像特征共有 f_i 个,其中 $i=1, \dots, N$ 。

2.3 模型分类

在分类过程中,由于 MPC-CLIP 模型已经得到 W_i 个文本特征和 f_i 个图像特征,只需要对每个投影视图的分类 $logits_i$ 进行单独计算,然后通过加权求和便可以获得点云的最终分类 $logits_h$, 并得到最终分类结果。计算公式为:

$$logits_i = f_i W_i^T, i = 1, \dots, n \quad (10)$$

$$logits_h = \sum_{i=1}^n logits_i \quad (11)$$

2.4 MPC-CLIP与不同网络模型的融合

三维点云分类网络模型通常通过提取点云的三维信息特征来完成分类任务,在一定程度上忽略了点云二维信

息特征的提取。MPC-CLIP 模型通过提取点云投影图像的特征获得点云的二维信息特征,与其他三维分类网络模型融合后的分类效果相较于单独使用三维分类网络模型获得了一定提升,说明 MPC-CLIP 能够作为一个即插即用的网络模型,与其他三维分类网络模型融合后可以改善没有关注到点云二维特征信息的不足,提高了点云分类准确率。模型融合流程如图 5 所示。

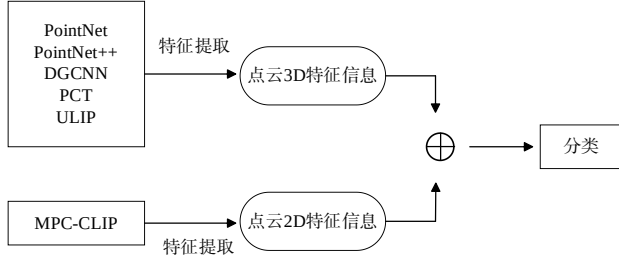


Fig. 5 Model fusion flow

图5 模型融合流程

3 实验方法与结果分析

3.1 实验环境

硬件环境为 NVIDIA RTX A4000(16 G),软件环境为 Ubuntu 22.04+CUDA 12.2+PyTorch 2.0.1+Python 3.10.4。学习率为 0.001,迭代次数为 250,batch size 为 32,使用 Adam 优化器。

3.2 评价指标

采用平均分类准确率(Mean Accuracy, mAcc)和总体分类准确率(Overall Accuracy, OA)评价模型分类精度。计算公式为:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$mAcc = \frac{\sum_{m=1}^M Precision_m}{M} \quad (14)$$

式中:TP表示预测为正的样本数;TN表示预测为负的样本数;FP表示实际为负但被预测为正的假正样本数;FN表示实际为正但被预测为负的假负样本数;Precision为精确率;M为类别数。

3.3 描述文本选取

本文模型训练使用点云数据集,因此构建适配文字编码器数据集中类别的描述文本时需要有一定的针对性,从而使模型能够更好地提取到描述点云文字的特征。在手动设置的描述文本上,本文共对 4 种语句进行了实验。结果见表 1。可以看出,不同描述文本对点云的分类准确率有所不同,与需要分类的点云投影图像越相关的描述文本分类准确率越高。最后一种描述文本(Point cloud projection of a {object})与需要分类的点云投影图像最相关,能够得到最高的 mAcc 与 OA,因此将其作为 ModelNet40 数据

集的描述文本。

Table 1 Classification accuracy of different description texts

表 1 不同描述文本的分类准确率 (%)		
数据集描述文本	mAcc	OA
A photo of a {object}	14.98	16.94
Point cloud of a {object}	15.72	18.66
Point cloud photo of a {object}	16.43	19.15
Point cloud projection of a {object}	17.55	20.09

3.4 图像编码器选取

传统 CLIP 的图像编码器是基于 ResNet 构建的,本文选取 ResNet18、ResNet34、ResNet50、ResNet101 和 ResNet152 网络进行比较。不同 ResNet 网络在点云投影图像数据集上的分类准确率见表 2。可以看出,当网络深度过浅或过深时分别可能会产生欠拟合和梯度消失的问题,MPC-CLIP 模型均不能获得很好的分类结果。这是由于当图像编码器中的 ResNet 网络深度过浅时不能很好地提取到投影图像的特征信息;而过深时会丢失图像特征信息。当使用 ResNet50 网络作为图像编码器时零样本分类精度最高,mAcc 为 17.75%,OA 为 20.25%,因此本文选择 ResNet50 网络作为模型的图像编码器。

Table 2 Classification accuracy of ResNet networks at different

depths		
表 2 不同深度 ResNet 网络的分类准确率 (%)		
不同深度的 ResNet 网络	mAcc	OA
ResNet18	14.98	16.94
ResNet34	16.72	18.66
ResNet50	17.75	20.25
ResNet101	14.35	16.56
ResNet152	13.05	15.55

3.5 MPC-CLIP 与不同模型融合分类的结果

分别向 PointNet^[9]、PointNet++^[13]、DGCNN^[25]、PCT^[26]及 ULIP^[27]几种主流三维点云分类网络中加入 MPC-CLIP 模型,与不加入时的点云分类准确率进行比较,所有模型均在 ModelNet40 数据集上进行实验,结果如表 3 所示。可以看出,在融合 MPC-CLIP 模型后,各三维点云分类网络的分类准确率均得到提升。具体来说,PointNet 网络加入 MPC-CLIP 模型后的 mAcc 提高了 1.8%,OA 提高了 1.3%;PointNet++ 网络加入 MPC-CLIP 模型后的 mAcc 提高了 1.9%,OA 提高了 0.8%;DGCNN 网络加入 MPC-CLIP 模型后的 mAcc 提高了 1.1%,OA 提高了 0.9%;PCT 网络加入 MPC-CLIP 模型后的 mAcc 提高了 0.9%,OA 提高了 1.7%;ULIP 网络加入 MPC-CLIP 模型后的 mAcc 提高了 1.2%,OA 提高了 1.4%。以上数据证实了 MPC-CLIP 模型可在一定程度上补充点云的二维相关特征,提高分类精度。

4 结语

本文基于 CLIP 在二维图像分类任务中的优越性提出一种基于多模态学习的点云分类网络模型 MPC-CLIP。

Table 3 Classification accuracy before and after adding MPC-CLIP model to different networks

表 3 不同网络加入 MPC-CLIP 模型前后的分类准确率 (%)

网络模型	MPC-CLIP	mAcc	OA
PointNet	×	85.7	88.9
	√	87.5	90.2
PointNet++	×	87.4	90.4
	√	89.3	91.2
DGCNN	×	89.2	91.3
	√	90.3	92.2
PCT	×	90.1	91.4
	√	91.2	93.1
ULIP	×	90.3	92.4
	√	91.5	93.8

在 ModelNet40 数据集上,该模型通过与其他三维点云分类网络模型融合后弥补了对点云数据二维信息提取的不足,最终使点云分类准确率获得了 0.8%~1.9% 的提升。然而该模型目前仅适用于点云分类任务,应用范围有限,后续计划设计合适的反投影模块,将二维平面上提取的特征映射回三维空间,以便进行更准确的分类与检测,进一步提升模型的实用性^[28]。

参考文献:

- [1] YURTSEVER E, LAMBERT J, CARBALLO A, et al. A survey of autonomous driving: common practices and emerging technologies[J]. IEEE Access, 2020, 8: 58443–58469.
- [2] JIANG S, WANG S, YI Z, et al. Autonomous navigation system of greenhouse mobile robot based on 3D lidar and 2D lidar SLAM[J]. Frontiers in Plant Science, 2022, 13: 815218.
- [3] LI Y, MA J, ZHANG Y. Image retrieval from remote sensing big data: a survey[J]. Information Fusion, 2021, 67: 94–115.
- [4] XIE Z X, LI M H. A survey on machine learning in recognition of 3D object based on point cloud[J]. Periodical of Ocean University of China, 2021, 51(6): 125–130.
- [5] DENG L, YU D. Deep learning: methods and applications[J]. Foundations and Trends in Signal Processing, 2014, 7(3/4): 197–387.
- [6] WANG W X, LI L L. Review of deep learning in point cloud classification[J]. Computer Engineering and Applications, 2022, 58(1): 26–40.
王文曦, 李乐林. 深度学习在点云分类中的研究综述[J]. 计算机工程与应用, 2022, 58(1): 26–40.
- [7] DING B, HE Y J, TANG L. An efficient 3D model retrieval method based on convolutional neural network[J]. Acta Electronica Sinica, 2021, 49(1): 64–71.
- [8] ZHANG M T, YAN M X, MA Y S, et al. 3D voxel model retrieval based on octree structure[J]. Chinese Journal of Computers, 2021, 44(2): 334–346.
张满园, 燕明晓, 马英石, 等. 基于八叉树结构的三维体素模型检索[J]. 计算机学报, 2021, 44(2): 334–346.
- [9] QI C R, SU H, MO K, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 77–85.
- [10] MA X, QIN C, YOU H X, et al. Rethinking network design and local geometry in point cloud: a simple residual MLP framework [DB/OL]. <https://arxiv.org/abs/2202.07123>.
- [11] LAI X, LIU J H, JIANG L. Stratified Transformer for 3D point cloud segmentation[DB/OL]. <https://arxiv.org/abs/2203.14508>.
- [12] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015: 2017–2025.
- [13] QI C R, LI Y, HAO S, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[C]//Proceedings of the Annual Conference on Neural Information Processing Systems, 2017: 5105–5114.
- [14] ZHAO H, JIANG L, FU C W, et al. Pointweb: enhancing local neighborhood features for pointcloud processing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5565–5573.
- [15] WU J, CUI Y, ZHAO X M, et al. SSA-PointNet++: a space self-attention CNN for the semantic segmentation of 3D point cloud[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(3): 437–448.
吴军, 崔玥, 赵雪梅, 等. SSA-PointNet++: 空间自注意力机制下的 3D 点云语义分割网络[J]. 计算机辅助设计与图形学学报, 2022, 34(3): 437–448.
- [16] ALEC R, JONG W K, CHRIS H, et al. Learning transferable visual models from natural language supervision [DB/OL]. <https://arxiv.org/abs/2103.00020>.
- [17] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[DB/OL]. <https://arxiv.org/abs/2101.00190>.
- [18] GIURGIU A, JASTRZEBSKI S, MORRONE B, et al. Parameter-efficient transfer learning for NLP[DB/OL]. <https://arxiv.org/pdf/1902.00751>.
- [19] GAO P, GENG S, ZHANG R, et al. CLIP-Adapter: better vision-language models with feature adapters [DB/OL]. <https://arxiv.org/abs/2110.04544>.
- [20] ZHANG R, FANG R, GAO P, et al. Tip-Adapter: training-free CLIP-Adapter for better vision-language modeling[DB/OL]. <https://arxiv.org/abs/2111.03930>.
- [21] CHANG A X, FUNKHOUSER T, GUIBAS L, et al. Shapenet: an information-rich 3D model repository [DB/OL]. <https://arxiv.org/abs/1512.03012>.
- [22] DAI A, CHANG A X, SAVVA M, et al. Scannet: richly-annotated 3D reconstructions of indoor scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5828–5839.
- [23] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1063–6919.
- [24] XU K, DENG P, HUANG H, et al. Vision Transformer: an excellent teacher for guiding small networks in remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5618715.
- [25] WANG C, SAMARI B, SIDDIQI K. Local spectral graph convolution for point set feature learning[C]//Proceedings of the European Conference on Computer Vision, 2018: 52–66.
- [26] GUO M H, CAI J X, LIU Z N, et al. PCT: point cloud transformer[J]. Computational Visual Media, 2021, 7: 187–199.
- [27] XUE L, GAO M, XING C, et al. ULIP: learning unified representation of language, image and point cloud for 3d understanding[DB/OL]. <https://arxiv.org/abs/2212.05171>.
- [28] YANG G X, ZHOU H, LIU H, et al. Lightweight network for point cloud classification based on gridding and surface features encoder[J]. Software Guide, 2024, 23(5): 9–16.
杨官学, 周昊, 刘慧, 等. 融合栅格与表面特征编码的轻量级点云分类网络[J]. 软件导刊, 2024, 23(5): 9–16.

(责任编辑:尹晨茹)