计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(3):291−299 [doi: 10.15888/j.cnki.csa.008997]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多层级上下文投票的三维密集字幕①

吴春雷, 郝宇钦, 李　阳

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通信作者: 郝宇钦, E-mail: z20070001@s.upc.edu.cn

摘　要: 传统的三维密集字幕方法存在未充分考虑上下文信息、点云特征信息丢失以及隐藏状态信息量单一等问题. 为了应对这些挑战, 提出了多层级上下文投票网络, 该网络在投票过程中使用自注意力机制捕获点云的上下文信息并加以多层级利用, 提升检测对象的准确率. 同时, 还设计了隐藏状态-注意力时序融合模块, 将当前时刻隐藏状态融合与前一时刻注意力结果融合, 丰富隐藏状态信息量, 从而提高模型表达能力. 除此之外, 采用“两阶段”训练方法, 有效过滤掉生成的低质量对象提案, 增强描述效果. 在官方数据集 ScanNet 和 ScanRefer 上的大量实验表明, 该方法与基线方法相比取得了更有竞争力的结果.

关键词: 三维密集字幕; 注意力机制; 上下文投票; 隐藏状态-注意力时序融合; 两阶段训练方法

引用格式: 吴春雷,郝宇钦,李阳.基于多层级上下文投票的三维密集字幕.计算机系统应用,2023,32(3):291–299. http://www.c-s-a.org.cn/1003-3254/8997.html

3D Dense Captioning Method Based on Multi-level Context Voting

WU Chun-Lei, HAO Yu-Qin, LI Yang

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Traditional three-dimensional (3D) dense captioning methods have problems such as insufficient consideration of point-cloud context information, loss of feature information, and thin hidden state information. Therefore, a multi-level context voting network is proposed. It uses the self-attention mechanism to capture the context information of point clouds in the voting process and utilizes it at multiple levels to improve the accuracy of object detection. Meanwhile, the temporal fusion of hidden state and attention module is designed to fuse the hidden state of the current moment with the attention result of the previous moment to enrich the information of the hidden state and thus improve the expressiveness of the model. In addition, a “two-stage” training method is adopted in the model, which can effectively filter out the generated low-quality object proposals and enhance the description effect. Extensive experiments on official datasets ScanNet and ScanRefer show that this method achieves more competitive results compared to baseline methods.

Key words: 3D dense captioning; attention mechanism; context voting; temporal fusion of hidden state and attention; two-stage training method

1　引言

场景理解就是对场景进行可视化分析, 是一个将图像或点云转化为认知的研究方向. 场景理解可以实现对三维场景的检测、定位、识别和理解 4 个层次的功能, 除了检测角落、边缘和移动区域等视觉特征外, 还可以提取与物理世界相关的信息, 具有一定的研究价值和应用前景[1]. 三维场景密集字幕作为视觉场景理解和自然语言处理联合领域的一个新兴任务, 备受关注, 目标是对三维场景中的任一对象进行定位和描述. 该任务不仅要求模型能够理解场景的点云内容信息,

① 基金项目: 山东省自然科学基金 (ZR2020MF136)
收稿时间: 2022-08-03; 修改时间: 2022-09-07; 采用时间: 2022-09-27; csa 在线出版时间: 2022-12-09
CNKI 网络首发时间: 2022-12-15

© 中国科学院软件研究所 http://www.c-s-a.org.cn

还要求用自然语言描述对象特征和对象之间的关系. 到目前为止, 视觉和自然语言的联合领域已经有了很多工作, 比如图像字幕、密集字幕、视频字幕、视觉问答等, 但是其中绝大多数工作都只局限在二维视觉数据上. 在物理世界中, 所有的元素理所当然的交织在一起, 因此真正理解物理世界需要整合多层次的信息以及学习场景元素之间关系, 即使是二维的照片和视频, 其所描绘的场景和对象本身也应是三维的. 为了更加真切的解释我们周围的世界, 视觉场景理解模型必须具备理解三维场景的能力. 与二维图像相比, 点云等三维数据能够为场景提供更详细的几何、结构和空间信息, 所以在由点云表示的真实世界环境中定位对象和产生描述性语句对于许多任务更为重要, 比如: 室内导盲机器人、机器人抓取、自动驾驶等. 中国有超过 1 750 万人视障, 他们的日常生活和出行十分不便, 无法定位和自主拿取物品, 道路突发事件难以及时规避, 该技术可以在实际中进行应用, 使得视障人士能够实时感知外部环境, 为日常生活提供便利和安全.

该任务实验所需数据集要求包含整个场景的点云及点云特征信息, 以及标注好的每个三维对象的三维边界框和相应描述. 但是制作数据集需要采集海量的场景样本并进行人工标注, 将耗费大量的人力物力. 由于实验缺少充足的训练数据, 三维密集字幕任务一直没有被跟进. 随着 ScanRefer 数据集[2] 的提出, 字幕任务才首次被 Chen 等人[3] 正式引入到三维视觉任务中, 即三维场景密集字幕任务, 也是近年来三维视觉领域发展的一个新的方向, 该任务在三维场景中密集地定位和描述三维对象, 比二维图像字幕更具有挑战性.

除了目标检测深度网络模型是关键的技术重点外, 为场景中检测出的物体生成相应的外观和相对位置自然语言描述, 也是需要突破的关键技术与发展方向. 基于此, 本文考虑到当前存在的目标检测模型未充分考虑场景点云之间的上下文信息. 如图 1 所示, 模型在定位对象时将微波炉下面的橱柜识别为“床”, 进而导致为对象“微波炉”生成的描述为“微波炉在床上”, 事实上, 在许多场景中, 微波炉是不可能放在床上的. 为了有效避免类似情况, 本文提出多层级上下文投票网络.

在投票过程中使用自注意力机制捕获点云的上下文信息, 同时还在投票过程中引入原始点云特征, 并以多层传导、层级叠加的方式更大化利用点云特征、降低特征损失, 进而提高检测三维对象的准确率. 考虑到隐藏状态信息量单一, 且前一时刻生成的单词对当前时刻单词生成也具有一定的指导作用, 本文提出了隐藏状态-注意力时序融合模块, 该模块在字幕生成过程中分别对两层门控循环单元 (gated recurrent unit, GRU) 的隐藏状态施加上下文注意力, 并将注意力结果传入下一时刻与“融合 GRU”的隐藏状态融合, 以提升隐藏状态的有效信息量, 进而更加丰富并准确地生成单词. 除此之外, 本文采用“两阶段”训练方法代替基线的“端到端”训练方法, 第 1 阶段使用基于多层级上下文投票网络的目标检测方法定位和识别场景中的三维对象, 生成三维对象提案; 第 2 阶段根据生成的对象提案和相应的实体名称标记来指导生成字幕. 该训练方法有效解决了前后阶段训练参数不匹配的问题, 过滤掉第 1 阶段生成的低质量对象提案, 从而降低了第 2 阶段训练的时间复杂度并增强描述效果.

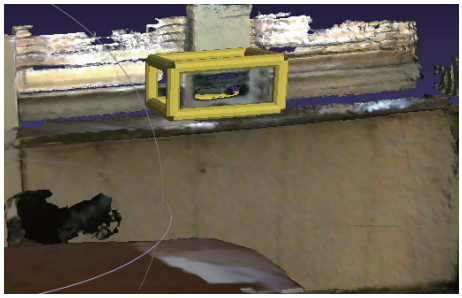

图 1 场景举例

本文的创新点可以总结归纳为如下 4 点.

(1) 提出了多层级上下文投票网络, 该网络通过利用三维元素之间的上下文关系来对场景中三维对象进行定位, 同时, 通过多层级引入原始点云特征, 达到减少投票过程中特征丢失的目的.

(2) 设计隐藏状态-注意力时序融合模块, 该模块通过融合隐藏状态与上一时刻注意力结果, 丰富当前时刻隐藏状态的有效信息量, 更加准确地指导生成单词.

(3) 采用“两阶段”训练方法, 不仅解决了前后阶段训练参数不匹配问题, 而且过滤掉了低质量对象提案, 在增强描述效果的同时, 降低了字幕生成阶段的时间复杂度.

(4) 通过基于 ScanNet 和 ScanRefer 数据集上的大量实验, 对本文提出的模型进行了分析与验证, 结果表明了该方法的有效性.

2 相关工作

2.1 三维视觉和语言

近年来, 视觉和语言的联合领域获得了极大的关

© 中国科学院软件研究所 http://www.c-s-a.org.cn

注, Chen 等人[4] 提出了数据集 Text2Shape, 包含对三维形状数据集 ShapeNet[5] 的描述, 提出了基于自然语言检索、生成彩色三维形状的任务. Achlioptas 等人[6] 和 Chen 等人[2] 提出了两个数据集 ReferIt3D 和 ScanRefer, 包含对 ScanNet 数据集[7] 场景中三维对象的描述. 不同的是, ReferIt3D 用来根据自然语言描述区分场景中的细粒度对象. ScanRefer 用于根据自然语言描述在场景中检测和定位对象. Yuan 等人[8] 和 Zhao 等人[9] 的工作, 也是利用自然语言描述在三维场景中检测和定位对象. 之后, Chen 等人[3] 提出一个反向任务, 根据场景上下文密集的定位和描述三维对象, 是三维视觉领域的目标检测任务和二维图像领域的字幕生成任务的联合, 该任务首次将字幕生成任务引入到三维领域中. 尽管该方法可以有效地为三维对象生成字幕, 但是生成的字幕单一且可辨别性不足. Chen 等人[10] 提出了一个更有效的半监督方法, 使用 PointGroup 网络[11] 作为检测主干, 并添加“说话者-听众”模块, 以自我批评的方式生成字幕. 该方法取得了更有竞争力的结果. 但是模型结构较为复杂, 机器需要较高的配置, 耗费大量的时间. 我们将 Chen 等人[3] 的工作作为基线. 通过引入多层级上下文投票网络和隐藏状态-注意力时序融合模块, 提高三维对象定位的准确性和生成字幕的可靠性.

2.2 三维目标检测

三维目标检测是计算机视觉领域中一个活跃的研究课题, 也是三维场景密集字幕的重要阶段. Qi 等人提出了 PointNet[12] 和 PointNet++[13] 网络, 开辟了在神经网络中直接处理三维点云数据的新途径. 接着, Shi 等人[14] 提出了一个两阶段三维目标检测网络, 先生成三维边界框提案, 再对这些提案进行细化获得最终检测结果, 很好地解决了遮挡问题以及检测过程中对二维检测结果的依赖. 但是该方法检测精度依然不高, 无法聚合对象中心周围的上下文信息. 受二维目标检测的霍夫投票策略[15] 的启发, Qi 等人[16] 提出了一个端到端的三维目标检测网络, 将本应存在于物体表面的点云用类似霍夫投票的方式投射到更加接近对象中心的位置, 进而预测对象的三维边界框. 虽然 VoteNet 网络在三维目标检测方面是有效的, 但是它的有效性很大程度上依赖于从骨架网络学习的点云特征. 因此, 提取高质量的点云特征是该三维目标检测方法成功的关键因素. 在该网络的基础上, Chen 等人[17] 采用基于图卷积的分层图神经网络用于三维目标检测; Yang 等人[18] 提出一种基于特征距离的新型采样策略; Zhang 等人[19] 通过引入一组混合的几何图元来改进由投票预测的初始边界框; Xie 等人[20] 在检测过程中捕获点、对象和场景级别的上下文信息. Zhao 等人[9] 提出了基于 Transformer 的架构, 在定位期间处理多模态上下文. 虽然这些方法有效地整合了多层次上下文信息、提高了检测精度, 但仍未有效解决投票过程中特征丢失过多的问题, 针对此问题, 本文提出多层级上下文投票网络, 致力于改进投票过程中的特征丢失问题, 进一步丰富生成投票点的位置.

2.3 图像字幕和密集字幕

图像字幕是计算机技术的研究热点, 也是计算机视觉和自然语言处理的一个快速发展的研究领域. 与早期基于模板和基于检索的图像字幕方法相比, Zhu 等人[21] 和 Ji 等人[22] 的工作基于编码器-解码器结构的图像字幕方法取得了巨大进步. 受机器翻译的启发, Anderson 等人[23] 和 Wang 等人[24] 提出了基于注意机制的图像字幕模型, 以提高描述性能. Mi 等人[25] 和 Li 等人[26] 利用场景图来捕捉对象之间的关系信息, 以提高图像字幕的性能. 但是这些方法只生成单个的枯燥、信息量少的字幕. Johnson 等人[27] 提出了一个完全卷积的定位网络, 将目标检测和图像字幕统一在一个框架中, 以预测对象区域上的一组描述, 用更多的句子描述图像以覆盖更多图像中的细节, 即图像密集描述. Krishna 等人[28] 将密集字幕移植到视频中, 旨在预测连续的事件提案并为每个剪辑生成描述. 在这些工作中, 密集的字幕比传统的单句传达了更多视觉内容的细节. 与图像密集字幕相似, 在三维场景中进行密集字幕生成任务需要首先为视觉概念标记边界框, 然后为每一个边界框标注字幕. 然而, 这些算法很少考虑到对象之间的空间关系信息, 这可能导致字幕任务在三维场景中生成不准确的对象之间关系的描述. 除此之外还有隐藏状态信息单一等问题, 导致生成的句子形式简单. 我们的工作是通过对两层 GRU 的特征向量和隐藏状态施加上下文注意力, 并将隐藏状态与前一时刻注意力的结果融合, 增加隐藏状态的有效信息量, 从而提高字幕模型的描述性能.

3 本文方法

3.1 整体框架

本文的整体模型框架如图 2 所示. 采用三维场景

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的点云作为网络的输入, 使用点云特征提取网络PointNet++提取点云特征F, 使用本文设计的多层级上下文投票网络将点云投射到各个点可能存在的物体表面的中心位置, 得到一簇簇最接近对象中心的点和聚类特征$C=(C_1,C_2,\cdots,C_n)$, 之后经过提案模块, 进行分组、聚合, 生成一组三维对象提案$S=(S_1,S_2,\cdots,S_n)$和预测框的置信度. 置信度与聚类特征经过过滤器, 得到有效特征V, 基于三维对象提案S构建图, 使用基线模型的关系图模块接收图和有效特征, 学习对象特征f和对象之间的关系特征f_r, 最后将f和f_r引入带有隐藏状态-注意力时序融合的密集字幕模块为每个对象生成相应字幕. 多层级上下文投票网络和带有隐藏状态-注意力时序融合的密集字幕模块的具体细节将在第 3.2 节和第 3.3 节介绍.

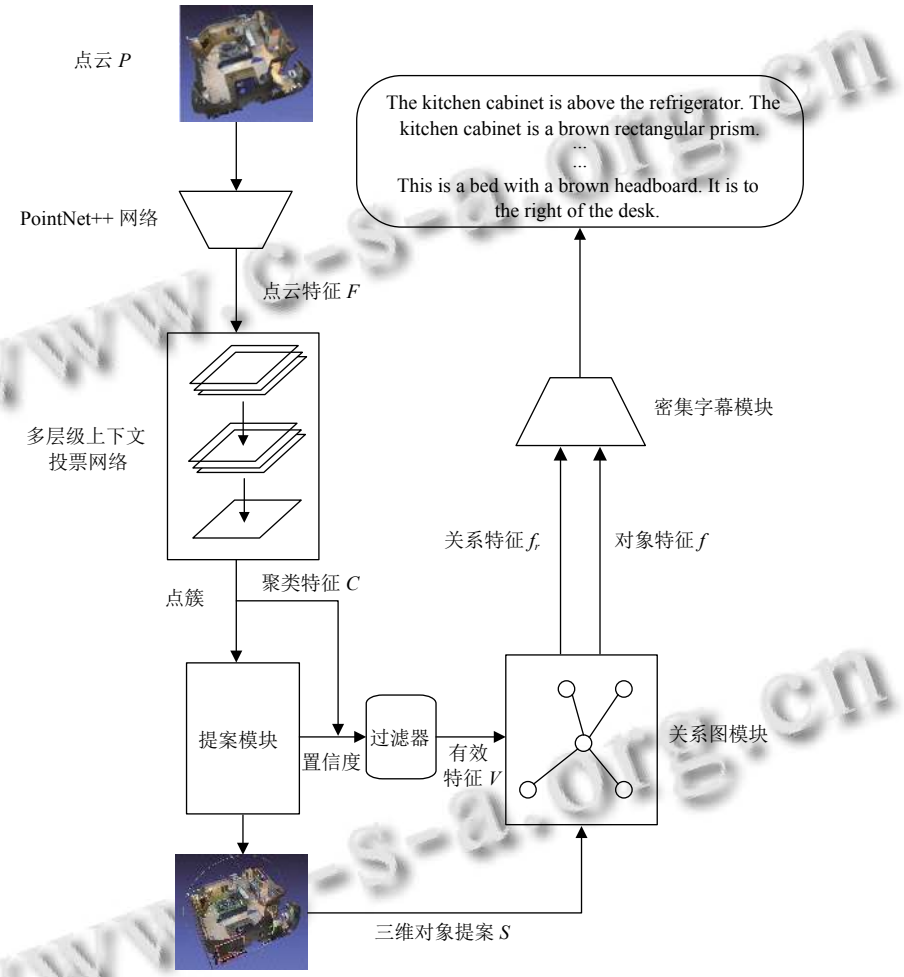

图 2 整体模型框架图

3.2 多层级上下文投票网络

本文提出了一个多层级上下文投票网络, 我们将原始 VoteNet 网络投票过程的“卷积层+*ReLU* 激活函数”称作一个投票单元. 在原始投票网络的基础上为两个投票单元各施加一层自注意力以学习投票过程中点云之间的上下文信息, 并将新投票单元的过程特征与原始点云特征融合. 该网络具体结构如图 3 所示.

此网络中, 将 PointNet++网络提取的点云特征F作为输入, 将F经过第 1 个投票单元得到过程特征C_a, 将C_a与点云特征F进行融合, 输入到第 2 个投票单元得到过程特征C_b. 最后将C_b与F、C_a融合, 经过一层卷积, 得到一组点簇和其聚类特征$C=(C_1,C_2,\cdots,C_n)$, 用于提案模块的分组、聚合. 该过程对应的公式为:

$$C_a = SA(ReLU(Conv(F))) \tag{1}$$

$$C_b = SA(ReLU(Conv(\delta_1 C_a + \gamma_1 F))) \tag{2}$$

$$C = Conv(\delta_2 F + \gamma_2 C_a + \lambda C_b) \tag{3}$$

其中, δ_1、δ_2、γ_1、γ_2、λ为超参数, 实验中均设置为 1.0.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

SA为自注意力函数, 本文采用三维注意力 Point Transformer[29].

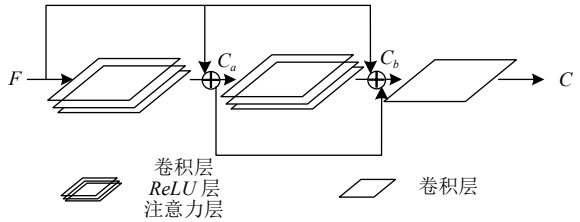

图 3 多层级上下文投票网络框架图

3.3 带有隐藏状态-注意力时序融合的密集字幕模块

本文将关系图模块输出的对象特征f和对象之间的关系特征f_r作为密集字幕模块的输入, 采用两个连续的 GRU 作为循环单元, 第 1 层 GRU 称为融合 GRU, 第 2 层 GRU 称为语言 GRU, 网络结构如图 4 所示.

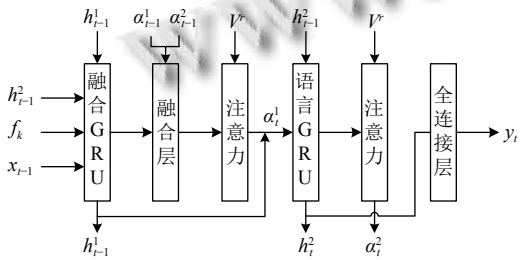

图 4 带有隐藏状态-注意力时序融合的密集字幕模块

首先, 本文通过式 (4) 表示$h_t = GRU(u_t, h_{t-1}, v_t)$时刻 GRU 的隐藏状态:

$$h_t = GRU(u_t, h_{t-1}, v_t) \tag{4}$$

其中, u_t是 GRU 在时刻t的输入向量, h_{t-1}是 GRU 在时刻$t-1$的隐藏状态, v_t表示时刻t的注意力结果, 初始化为 0. 为了方便表示, 本文对于 GRU 存储单元的单元状态忽略不计, 统一使用式 (4) 表示在时刻 t 每一层 GRU 的输入和输出向量.

在每个时间步长t, 融合 GRU 的输入如式 (5) 所示:

$$u^1_t = [h^2_{t-1}, f_k, x_{t-1}] \tag{5}$$

其中, h^2_{t-1}是语言 GRU 在$t-1$时刻的隐藏状态, f_k为第k个对象的特征, x_{t-1}是$t-1$时刻的词嵌入向量.

得到融合 GRU 的隐藏状态h^1_t, 我们将$t-1$时刻含有对象上下文特征信息的两层注意力的结果引入到t时刻与h^1_t融合, 增加隐藏状态的有效信息量, 再对融合后的隐藏状态施加上下文注意力. 该过程对应的具体公式如下:

$$H^1_t = W_1 h^1_t + W_2[\alpha^1_{t-1}, \alpha^2_{t-1}] \tag{6}$$

$$\alpha^1_t = Softmax((W_v V^r + W_h H^1_t) W_a) \tag{7}$$

其中, α^1_{t-1}和α^2_{t-1}分别表示$t-1$时刻两层注意力的输出结果, W_1和W_2为超参数, 实验中设置为 1.0. W_v、W_h、W_a为学习参数, $V^r=\{f^i_k\}$为上下文特征集合.

将注意力结果与融合 GRU 的隐藏状态连接作为语言 GRU 的输入向量, 对应公式为:

$$u^2_t = [h^1_t, \alpha^1_t] \tag{8}$$

3.4 目标函数

我们与 Qi 等人[16] 使用相同的目标检测损失:

$$\varphi_{\text{det}} = \varphi_{\text{vote-reg}} + 0.5\varphi_{\text{objn-cls}} + \varphi_{\text{box}} + 0.1\varphi_{\text{sem-cls}} \tag{9}$$

其中, $\varphi_{\text{voet-reg}}$, $\varphi_{\text{objn-cls}}$, φ_{box}和$\varphi_{\text{sem-cls}}$分别表示 18 个 ScanNet 基准类的投票回归损失、目标分类损失、边界框回归损失和语义分类损失. 投票回归损失采用 L1 距离方法计算预测值与真值的差距, 使用交叉熵作为损失函数. 目标分类损失通过交叉熵损失函数判断一个投票点簇是否为物体. 在本文中将边界框回归损失简化为式 (10):

$$\varphi_{\text{box}} = \varphi_{\text{center-reg}} + 0.1\varphi_{\text{size-cls}} + \varphi_{\text{size-reg}} \tag{10}$$

其中, $\varphi_{\text{center-reg}}$, $\varphi_{\text{size-cls}}$和$\varphi_{\text{size-reg}}$分别表示边界框中心回归损失、边界框尺度分类损失和边界框尺度回归损失. 边界框中心回归损失采用的是倒角损失函数[30], 在计算时, 将所有投票点与真实目标中心距离小于 0.3 或大于 0.6 的提案分别视为正向提案和反向提案. 只有对于正例才计算以上的所有损失, 反例只计算分类是不是目标的损失.

为了稳定关系图模块的学习过程, 在消息传递网络上应用相对方向损失φ_{ad}作为代理损失. 从 0 到 180 度, 分为 6 个类别离散化输出角度偏差, 并使用交叉熵损失作为本文的分类损失. 在字幕生成过程中, 采用传统的交叉熵作为损失函数φ_{des}.

最后用线性方式组合上述 3 个损失项, 作为最终的损失函数:

$$\varphi = m_1\varphi_{\text{det}} + m_2\varphi_{\text{ad}} + m_3\varphi_{\text{des}} \tag{11}$$

其中, m_1, m_2, m_3分别为各损失项的权重, 实验中设置为 10, 1, 0.1.

4 实验

4.1 数据集与评估指标

本文在官方数据集 ScanNet 和 ScanRefer 上评估

© 中国科学院软件研究所 http://www.c-s-a.org.cn

和验证基于多层级上下文投票的三维密集字幕模型. ScanRefer 数据集包含 800 个场景、11 046 个对象以及 51 583 个对应的描述, 这些描述包含对象的外观信息 (例如: 这是一个白色的微波炉) 以及相对空间位置 (例如: 它在桌子上的角落里). 数据样本分为训练集 36 665 个和验证集 9 508 个, 确保每个分割的场景不相交, 结果和分析同样是在验证集上进行的. 在实验过程中, 使用预测边界之间的交并比 (intersection over union, IoU) 分数来评估生成的三维边界框, IoU 阈值的平均精度 (mean average precision, mAP) 作为目标检测评价指标. 使用标准图像字幕评估策略, 包括 CIDEr[31]、BLEU[32]、METEOR[33]、ROUGE[34], 来评估本文所提出的模型, 并与基线模型进行比较. CIDEr 评估方法将生成的字幕表示成向量, 计算真实标签与模型生成字幕的余弦相似度. BLEU 评估生成字幕与真实标签的差异, 取值范围为 0–1, 如果两者完美匹配, 那么 BLEU 是 1, 反之则为 0. METEOR 评估单字精度召回率. ROUGE 将生成字幕与真实标签的 n 元组贡献统计量作为评估指标.

4.2 实验细节

本文所做实验均是基于 PyTorch 框架, 并在装有 RTX 2080Ti GPU 的计算机上进行实验, 代替基线的端到端训练方法, 本文采用“两阶段”训练模式, 第 1 阶段训练三维目标检测模块, 使用 Adam 优化器, 批次大小为 8, 学习率为 0.01, 训练轮次为 200, 学习率衰减步长为{80, 120, 160}, 衰减率为{0.1, 0.1, 0.1}, 训练完成后从 IoU 阈值设置为 0.5 的模型中选出最优模型; 第 2 阶段将训练好的三维目标检测模型作为预训练模型, 训练密集字幕模块, 学习率设置为 1E–5, 批次大小设置为 12, 训练轮次为 50, 迭代次数 114 600.

4.3 实验结果

如图 5 所示, 是本文提出的方法训练的模型与基线模型在 ScanRefer 数据集上的结果比较, 可以看出, 对于同一场景, 本文的模型定位对象更准确, 生成的密集字幕与三维场景的契合度更高, 语言的表达更加准确、流利.

GroundTruth: It is a wood two door cabinet. The cabinet is above the refrigerator.
Scan2Cap: This is a white kitchen cabinet. It is above the refrigerator.
Ours: The kitchen cabinet is above the refrigerator. The kitchen cabinet is a brown rectangular prism.

GroundTruth: It is a flat screen tv. The tv is on the tv cabinet next to the window with the green curtains in the living room area.
Scan2Cap: This is a black printer. It is on a shelf.
Ours: This is a black tv. It is on the tv stand.

图 5 实验结果图

如表 1 所示, 本文在 ScanNet 和 ScanRefer 官方数据集上对所提方法进行了测试.

表 1 IoU 阈值为 0.25 和 0.5 模型字幕评级指标得分

Method	IoU	评价指标 (%)			
		CIDEr	BLEU-4	METEOR	ROUGE
Scan2Cap	0.25	56.82	34.18	26.29	55.27
Ours		**59.68**	**35.62**	**26.48**	**56.40**
Scan2Cap	0.5	39.08	23.32	21.97	44.78
Ours		**41.41**	**24.84**	**22.71**	**45.75**

从表 1 的测试结果中可以看到, 与基线模型相比, 本文训练的模型生成的字幕评估指标都有所提高, 尤其是模型 IoU 阈值为 0.25 的 CIDEr 得分提高了 2.86%, 模型 IoU 阈值为 0.5 的 CIDEr 得分提高了 2.33%. 表 1 测试结果验证了本文提出的模型和训练方法的有效性.

4.4 实验分析

为了评估本文提出的多层级上下文投票网络和隐藏状态-注意力时序融合模块的有效性, 本文对两个部分单独进行实验验证, 表 2 中的 A 表示在基线方法 Scan2Cap 模型的基础上, 将原始投票网络更换为本文提出的多层级上下文投票网络, B 表示在 Scan2Cap 模型的字幕生成阶段添加隐藏状态-注意力时序融合模

© 中国科学院软件研究所 http://www.c-s-a.org.cn

块. 结果如表 2 所示.

从表 2 的实验结果中可以看到, 在基线模型的基础上单独应用 A 时, IoU 阈值设置为 0.25 和 0.5 的模型 CIDEr 得分分别提升了 0.78% 和 1.58%, 在基线模型的基础上单独应用 B 时, CIDEr 得分分别提升了 1.89% 和 1.24%. 不出所料, 当同时应用 A 和 B 时, 我们的模型性能提升到最高. 表 2 测试结果验证了单独应用本文提出的多层级上下文投票网络和隐藏状态-注意力时序融合模块, 模型改进效果依然有效.

表 2 消融实验

Method	A	B	IoU	评价指标 (%)			
				CIDEr	BLEU-4	METEOR	ROUGE
Scan2Cap	—	—	0.25	56.82	34.18	26.29	55.27
Ours	√	—		57.60	35.31	26.29	56.12
Ours	—	√		58.71	**36.05**	26.32	56.15
Ours	√	√		**59.68**	35.62	**26.48**	**56.40**
Scan2Cap	—	—	0.5	39.08	23.32	21.97	44.78
Ours	√	—		40.66	24.15	22.46	45.43
Ours	—	√		40.32	24.32	22.02	45.49
Ours	√	√		**41.41**	**24.84**	**22.71**	**45.75**

为了验证本文“两阶段”训练方法的有效性, 本文设计了对比实验, 结果如表 3、表 4 所示. Our (e2e) 表示本文模型采用“端到端”训练方式.

表 3 “端到端”方法与“两阶段”方法模型得分比较

Method	IoU	评价指标 (%)			
		CIDEr	BLEU-4	METEOR	ROUGE
Scan2Cap	0.25	56.82	34.18	26.29	55.27
Ours (e2e)		58.4	35.11	**26.78**	55.92
Ours		**59.68**	**35.62**	26.48	**56.40**
Scan2Cap	0.5	39.08	23.32	21.97	44.78
Ours (e2e)		39.7	23.89	21.99	45.12
Ours		**41.41**	**24.84**	**22.71**	**45.75**

表 4 “端到端”方法与“两阶段”方法训练时间比较 (h)

Method	训练时间
Scan2Cap	120
Ours (e2e)	120
Ours	**96**

从表 3 的实验结果中可以看到, 本文采用的“两阶段”训练方法比基线的“端到端”训练方法生成的字幕评估分数更高, 从表 4 的实验结果中可以看到, 采用“两阶段”训练方法时模型总训练时间从 120 h 缩减为 96 h. 表 3、表 4 测试结果验证了本文“两阶段”训练方法的有效性.

为了研究选取第 1 阶段不同 IoU 阈值的目标检测模型作为第 2 阶段的预训练模型, 对本文最终模型性能的影响, 我们进行了对比实验, 如表 5 所示. Ours (0.25) 表示本文模型选取第 1 阶段 IoU 阈值为 0.25 的目标检测模型作为第 2 阶段的预训练模型.

表 5 选取不同 IoU 阈值模型作为预训练模型的结果比较

Method	IoU	评价指标 (%)			
		CIDEr	BLEU-4	METEOR	ROUGE
Scan2Cap	0.25	56.82	34.18	26.29	55.27
Ours (0.25)		**60.08**	**36.8**	**26.75**	**57.02**
Ours		59.68	35.62	26.48	56.40
Scan2Cap	0.5	39.08	23.32	21.97	44.78
Ours (0.25)		39.61	23.78	21.81	45.01
Ours		**41.41**	**24.84**	**22.71**	**45.75**

从表 5 的实验结果中可以看到, 选取第 1 阶段 IoU 阈值为 0.25 的目标检测模型作为第 2 阶段的预训练模型, CIDEr 得分相较基线分别提升了 3.26% 和 0.53%, 但 IoU 阈值为 0.5 的模型 METEOR 分数降低了 0.16%. 表 5 测试结果表明选取第 1 阶段 IoU 阈值为 0.5 中模型的最优目标检测模型作为第 2 阶段的预训练模型比 0.25 的整体模型效果更好. 且 IoU 阈值为 0.5 的模型说服力比 0.25 更高.

5 结论与展望

本文提出了一种基于多层级上下文投票和隐藏状态-注意力时序融合的模型用于三维密集字幕生成, 该模型能够捕获投票过程中点云的上下文信息, 提高检测目标的准确性, 还能够将原始点云特征信息和投票过程中的信息加以多层级利用, 减少投票过程中的特征丢失. 同时, 本文设计的隐藏状态-注意力时序融合模块将当前时刻融合 GRU 的隐藏状态与前一时刻两层注意力进行融合, 提升隐藏状态的有效信息量. 本文还采用“两阶段”训练方法, 有效地过滤掉目标检测阶段生成的低质量对象提案, 降低字幕生成阶段的时间复杂度并增强描述效果. 通过广泛的实验验证了基于多层级上下文投票网络的三维密集字幕方法的有效性. 在未来的工作中, 将进一步探索模型框架和注意力机制的改进方式.

参考文献

1 Romaszko L, Williams CKI, Winn J. Learning direct optimization for scene understanding. Pattern Recognition,

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2020, 105(2): 107369.
2 Chen DZ, Chang AX, Nießner M. ScanRefer: 3D object localization in RGB-D scans using natural language. Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow: Springer, 2020. 202–221.
3 Chen DZ, Gholami A, Nießner M, *et al*. Scan2Cap: Context-aware dense captioning in RGB-D scans. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3192–3202.
4 Chen K, Choy CB, Savva M, *et al*. Text2Shape: Generating shapes from natural language by learning joint embeddings. Proceedings of the 14th Asian Conference on Computer Vision. Perth: Springer, 2018. 100–116.
5 Chang AX, Funkhouser T, Guibas L, *et al*. ShapeNet: An information-rich 3D model repository. arXiv:1512.03012, 2015.
6 Achlioptas P, Abdelreheem A, Xia F, *et al*. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 422–440.
7 Dai A, Chang AX, Savva M, *et al*. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 2432–2443.
8 Yuan ZH, Yan X, Liao YH, *et al*. InstanceRefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 1791–1800.
9 Zhao LC, Cai DG, Sheng L, *et al*. 3DVG-transformer: Relation modeling for visual grounding on point clouds. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 2908–2917.
10 Chen DZ, Wu QR, Nießner M, *et al*. D3Net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in RGB-D scans. arXiv:2112.01551, 2021.
11 Jiang L, Zhao HS, Shi SS, *et al*. PointGroup: Dual-set point grouping for 3D instance segmentation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4867–4876.
12 Qi CR, Su H, Kaichun M, *et al*. PointNet: Deep learning on point sets for 3D classification and segmentation. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 77–85.
13 Qi CR, Yi L, Su H, *et al*. PointNet++: Deep hierarchical feature learning on point sets in a metric space. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5105–5114.
14 Shi SS, Wang XG, Li HS. PointRCNN: 3D object proposal generation and detection from point cloud. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 770–779.
15 Leibe B, Leonardis A, Schiele B. Robust object detection with interleaved categorization and segmentation. International Journal of Computer Vision, 2008, 77(1–3): 259–289.
16 Qi CR, Litany O, He KM, *et al*. Deep Hough voting for 3D object detection in point clouds. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 9276–9285.
17 Chen JT, Lei BW, Song QY, *et al*. A hierarchical graph network for 3D object detection on point clouds. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 389–398.
18 Yang ZT, Sun YN, Liu S, *et al*. 3DSSD: Point-based 3D single stage object detector. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 11037–11045.
19 Zhang ZW, Sun B, Yang HT, *et al*. H3DNet: 3D object detection using hybrid geometric primitives. Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow: Springer, 2020. 311–329.
20 Xie Q, Lai YK, Wu J, *et al*. MLCVNet: Multi-level context VoteNet for 3D object detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10444–10453.
21 Zhu Z, Wang T, Qu H. Macroscopic control of text generation for image captioning. arXiv:2101.08000, 2021.
22 Ji JZ, Xu C, Zhang XD, *et al*. Spatio-temporal memory attention for image captioning. IEEE Transactions on Image Processing, 2020, 29: 7615–7628. [doi: 10.1109/TIP.2020.3004729]
23 Anderson P, He XD, Buehler C, *et al*. Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6077–6086.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

24 Wang WX, Chen ZH, Hu HF. Hierarchical attention network for image captioning. Proceedings of the 33rd AAAI Conference on Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conference and 9th AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu: AAAI, 2019. 1099.

25 Mi JP, Lyu J, Tang S, *et al*. Interactive natural language grounding via referring expression comprehension and scene graph parsing. Frontiers in Neurorobotics, 2020, 14: 43. [doi: 10.3389/fnbot.2020.00043]

26 Li XY, Jiang SQ. Know more say less: Image captioning based on scene graphs. IEEE Transactions on Multimedia, 2019, 21(8): 2117–2130. [doi: 10.1109/TMM.2019.2896516]

27 Johnson J, Karpathy A, Li FF. DenseCap: Fully convolutional localization networks for dense captioning. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4565–4574.

28 Krishna R, Hata K, Ren F, *et al*. Dense-captioning events in videos. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 706–715.

29 Zhao HS, Jiang L, Jia JY, *et al*. Point transformer. arXiv:2012.09164, 2020.

30 Fan HQ, Su H, Guibas L. A point set generation network for 3D object reconstruction from a single image. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 605–613.

31 Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4566–4575.

32 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002. 311–318.

33 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor: Association for Computational Linguistics, 2005. 65–72.

34 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004. 74–81.

(校对责编: 牛欣悦)

www.c-s-a.org.cn

© 中国科学院软件研究所 http://www.c-s-a.org.cn