基于多视角卷积神经网络的三维模型分类方法

王鹏宇, 水盼盼, 余锋根, 甘渊, 刘琨, 张岩*, 孙正兴

南京大学软件新技术国家重点实验室, 南京 210023

* 通信作者. E-mail: zhangyannju@nju.edu.cn

收稿日期: 2018-09-14; 接受日期: 2019-01-28; 网络出版日期: 2019-04-11

国家自然科学基金 (批准号: 61321491, 61100110, 61272219) 和江苏省科技支撑计划 (批准号: BY2012190, BY2013072-04) 资助项目

摘要 近些年来, 卷积神经网络框架在二维图像的语义分割、分类、检索等领域取得了非常好的效果. 但是由于三维模型结构的复杂性与不规则性, 卷积神经网络的卷积和池化操作却无法直接应用在三维模型上. 为了发挥深度学习框架在二维图像分析领域积累的技术优势, 本文采用基于多视角投影的方法来完成三维模型分类的任务. 然而现有的基于多视角投影的三维模型分类方法大多采用固定视角, 所采集到的多视角下模型投影渲染图中存在大量的信息冗余, 对结果造成一定的干扰. 本文提出了一种新型的多视角卷积神经网络框架, 在网络训练过程中自动判别每个视角的贡献度, 舍弃冗余视角的信息, 从而提取出最能表征模型类别的特征, 提高了网络的鲁棒性. 此外, 本文将基于视点熵的最佳视角选择方法引入三维模型分类领域, 相比于固定视角方法, 本文方法能更多地保留模型的细节信息, 同时不需要模型的朝向对齐. 通过在 ModelNet10 和 ModelNet40 数据集上的实验, 验证了将基于视点熵的视角选择方法应用于三维模型分类, 以及本文提出的基于视角判别的多视角信息融合方法的合理性和优越性. 实验结果表明, 本文方法的分类准确性也优于现有的基于固定多视角投影的三维模型分类方法.

关键词 三维模型, 视角选择, 视点熵, 卷积神经网络, 深度学习

1 引言

近些年来, 随着三维建模软件以及深度传感器的广泛应用, 在线三维模型的种类和数量出现了爆炸式的增长, 如何有效地对这些三维模型进行分类、组织和管理成为了一个亟待解决的研究问题. 为了解决模型分类的问题, 大量的研究工作应运而生, 现有工作主要分为基于传统方法的三维模型分类和深度学习方法的三维模型分类. 其中传统方法^[1~3]主要通过人工设计的单一特征几何描述子对三维模型提取特征, 然后对这些特征采用聚类算法、支持向量机 (SVM) 等方法进行处理, 最终完成三维模型分类任务. 虽然这类方法在小数据集上即可应用, 速度也很快, 但是人工设计的单一特征描述

引用格式: 王鹏宇, 水盼盼, 余锋根, 等. 基于多视角卷积神经网络的三维模型分类方法. 中国科学: 信息科学, 2019, 49: 436–449, doi: 10.1360/N112018-00254
Wang P Y, Shui P P, Yu F G, et al. 3D shape classification based on convolutional neural networks fusing multi-view information (in Chinese). Sci Sin Inform, 2019, 49: 436–449, doi: 10.1360/N112018-00254

子很难有效地提取高层特征,同时要求模型朝向对齐并保持流形结构,因此泛化能力较弱,只能适用于特定的数据集,难以适用于较大规模的数据集。

相比设计三维模型特征描述子的复杂性,在二维图像语义分割和分类领域,深度学习方法的引入使二维图像上的特征学习获得了突飞猛进的发展,如 ResNet^[3], FCN^[4], DeepLab^[5] 等,都取得了很好的效果。尽管三维模型分类领域已有不少研究成果,也取得了很大的进展,但是三维模型内在结构的复杂性给深度学习方法的应用带来了极大的挑战,近期的一些工作开始尝试设计三维数据上的深度学习框架。因为三维模型数据表示的多样性(如网格、点云、体素化、多视角投影法等),不同的数据表示方法优缺点各不相同,实际应用的时候往往要根据任务本身来做选择。比如:点云表示较为简短,可直接利用矩阵运算对模型进行仿射变换,然而点集的无序性使得在点云上无法直接使用卷积操作;网格表示加入了点的连接关系使得可以更方便地进行一些拓扑变换,而更复杂的关系也使得在网格数据上设计深度学习框架变得更加困难;体素化表示可以实现在三维模型上直接做卷积与池化操作,但主要问题是三维模型体素化之后限于分辨率的问题会丢失三维模型的局部细节,同时三维卷积操作所消耗的时间资源相比于二维卷积操作更大。而相比于其他表示方法,基于多视角投影图的方法可以直接利用二维深度学习框架,并通过融合多视角信息来有效地提取模型的类别特征,从而完成对三维模型分类。

最近的多视角神经网络相关工作^[6~8],大多采用固定视角。固定视角的设置一般是基于人类对模型的先验认识,由于不同类别模型的局部细节不同,对所有类别设置统一的视角可能无法完全展示出不同类别模型的细节特点。同时较少的固定视角个数要求了模型必须朝向对齐。所以,为了去除朝向对齐的要求,需要增加投影的视角个数,然而增加了视角的个数又加大了计算的复杂度。为了解决上述问题,本文将基于视点熵的视角选择方法引入多视角卷积神经网络,在预先设定的多个视角下选择视点熵最大的几个视角,使得输入模型不需要朝向对齐,同时基于视点熵的视角选择方法可以在相同视角个数的情况下,获得更多的模型面片信息。与固定视角的方法相比,基于视点熵的视角选择方法可以在获取相等模型细节信息量的前提下,减少模型的视角数,增加计算效率。

另一方面,基于多视角投影的卷积神经网络需要考虑如何进行多视角信息的融合,像 MVCNN^[6] 网络结构中使用视角池化(view-pooling)来融合每个单视角下的模型投影的特征。然而单视角下的特征可能存在大量的信息冗余,使用所有视角的特征直接进行特征聚合可能会使得网络泛化性降低,增加过拟合风险。为了进一步减少视角的冗余信息,更好地融合多视角信息,本文提出了一种新型的多视角卷积神经网络框架,在获取到模型的单视角特征后,设计了一个额外的网络,可以在训练过程中自动判别每个视角的贡献度,过滤掉冗余视角信息,并对多视角信息进行融合,从而提取出最能表征模型类别的特征。

在实验阶段,本文方法在 ModelNet¹⁾ 分类数据集上进行实验。本文先验证了在相同视角个数的情况下,基于视点熵的视角选择方法与固定视角方法相比,分类的准确率更高。随后的实验验证了本文提出的多视角信息融合网络框架的优越性,本文方法的分类准确性优于现有的基于固定多视角投影的三维模型分类方法,证明网络泛化性更强。

2 相关工作

2.1 视角选择

视角选择技术可以用来帮助人们选择一个观察三维物体的角度,从此视角出发能够形成对观察物

1) The Princeton ModelNet. <http://modelnet.cs.princeton.edu/>.

体最全面、最直观的认识。然而关于最佳视角的选择并没有一个统一的标准。关于如何清楚地、全面地观察物体,对于不同的任务可能有不同的标准,而且都基于人类的定义。当前三维模型最佳视角选择技术可大致分为两类:基于三维模型几何信息量的视角选择与基于视觉显著信息量的视角选择。

三维模型几何信息量的视角选择,主要根据视角下可见的顶点数目、可见面片数目、面片投影,以及几何面积等因素来计算视角质量来提取最佳视角。常见的基于几何信息量的方法有视点熵方法^[9]、相对熵^[10]等。本文使用基于视点熵方法,该方法计算效率高,能很大程度上保留模型的几何信息,且基于熵的计算可以在可见面片个数与投影后的可见面积之间取得平衡而忽略模型面片面积大小分布不均对视角选择的影响。

另外一类方法是基于模型表面视觉显著信息量的视角选择。比如计算曲率大小作为网格数据的显著度来观察曲率变化最大的位置。Lee 等^[11]引入了视觉显著性作为网格模型数据的区域重要性的一种度量。基于视觉显著性信息的方法一般选择包含显著性信息的视角,而忽略了模型的整体结构以及模型的几何特性,对于三维模型分类任务来说,模型的几何信息是尤为重要的,所以本文采用基于三维模型几何信息量的视角选择方法。

2.2 三维数据深度学习

随着深度学习的快速发展,三维模型分类领域也出现了很多深度学习方法。Qin 等^[12]从待选的一些特征描述子中选出效果最好的一组特征描述子,然后将提取的特征输入到 CNN 中进行分类。而 Socher 等^[13]则通过 CNN 而不是特征描述子来提取底层特征,再输入到循环神经网络 (recurrent neural networks) 中提取高层特征。虽然上述这些方法提取模型的特征方式不同,但是这些方法着重于提取模型的几何特征,而忽略了与模型类别密切相关的视觉特征。

2.2.1 基于网格数据的深度学习

Bruna 等^[14]提出的方法基于图 Laplace 谱构建卷积神经网络结构,而 Masci 等^[15]提出的测地线卷积神经网络基于局部测地线的极坐标系统来提取局部区域的特征,再通过一系列线性和非线性组合来提取能代表模型类别的稳定特征。虽然这种方法有效地提取到了模型的局部特征并取得了不错的效果,而且面片与面片之间的连接关系可以较好地保持局部的拓扑性质,但是该类方法过度依赖局部特征,缺乏多尺度的特征,且计算量较大。

2.2.2 基于体素数据的深度学习

在基于体素的方法中,Wu 等^[16]提出的 3D ShapeNets 将模型表示成在划分的各个体素单元格的概率分布,然后通过卷积深度置信网络学习模型的类别特征,取得了很好的效果。Li 等^[17]提出的 FPNN 将 3D 数据统一表示成高斯距离场 (Gaussian distance field),通过不断地学习调整由一系列探测点组成的域探测过滤器的位置及权重,以智能地适应模型在三维空间中的分布,提升了在三维空间中卷积的效率。Xu 等^[18]受 3DShapeNets^[16]启发,采用类似框架,但专注于对网络本身的参数进行学习,以寻找最优的卷积神经网络框架,来适应三维模型数据集较小的限制性因素。Sedaghat 等^[19]提出的方法则基于 Voxnet^[20]并引入模型朝向因素作为分类的依据,进一步提升了分类效果。Ren 等^[20]提出的 3D-A-Nets 旨在解决三维模型体素卷积效率低下,难以有效地提取模型的几何信息的问题,将体素模型表示成多个二维多层密集表达,并依次输入卷积神经网络和循环神经网络,用三维深度形状描述子提取的特征,最后输入对抗判别器进行分类。上述这些基于体素的方法虽然都围绕着解决三维

卷积计算效率太低、体素分辨率大小难以控制等问题,但是受限于三维模型体素精确度的影响,最终的效果并不理想.

2.2.3 基于点云数据的深度学习

在基于点云的方法中, Qi 等^[21]提出了能够直接对 3D 点云进行分类、语义分割和场景分割的统一框架 PointNet, 该方法能够高效地处理稀疏数据, 直接以无序的点云作为输入, 取得了很好的效果. 此后, Qi 等改进了 PointNet^[21]方法提出了 PointNet++^[22]方法, 从多尺度的角度来处理更复杂的点云结构. 此后 Simonovsky 等^[23]在点云上建立图结构, Klokov 等^[24]则在点云上构建树结构, 以使点云序列化, 便于进行卷积和池化操作, 但是从不同的点开始建图和树结构, 得到的结果是不一样的, 因此, 这些方法依然没能解决点云的无序性问题. Li 等^[25]提出的 SO-Net 则致力于通过 K 近邻方法调整感受域, 实现层次化的特征提取. 因此, 基于点云的卷积神经网络三维模型分类方法仍然受到点云的稀疏性与无序性的影响. 如何有效地提取点云的局部结构特征和全局特征等问题依然没有得到很好解决.

2.2.4 基于投影数据的深度学习

基于投影的方法则对三维模型的结构无要求, 避免了直接在三维模型上进行卷积和池化等操作的难题, 并且能够发挥图像分析领域成熟的深度学习框架的优势. Shi 等^[26]提出了 DeepPano 方法对模型采集单视角投影图, 并对投影图提取特征进行分类. 由于单视角的信息比较有限, 于是 Su 等^[6]提出了基于 12 个固定视角的多视角投影的三维模型识别和分类方法 MVCNN, 得到了非常好的效果.

鉴于基于多视角投影方法的良好效果, 后续又有很多研究人员进行了深入的研究. Johns 等^[27]提出了 Pairwise 方法, 致力于通过深度学习的方法解决如何寻找下一个最佳视角的问题. 而 Sfikas 等^[28]提出的 PANORAMA-NN 方法在 DeepPano^[26]的基础上考虑多个视角坐标轴, 并在所有轴向上统一进行归一化, 不仅考虑空间因素还考虑了朝向因素. Wang 等^[8]提出了基于动态集合的 Cluster-Pooling CNN 方法, 仍然是输入 12 个固定视角下的模型投影渲染图, 根据视角相似性信息来对视角特征进行聚类. 而 Kanezaki 等^[29]提出的 RotationNet 方法则将固定的多视角看做变量, 从中学习出最能代表模型类别特征的若干视角, 然后再从这些视角的投影渲染图中提取特征进行分类, 取得了不错的效果, 然而其中问题是测试阶段的视角必须在训练中出现. 目前的基于投影的方法一直没有很好地解决最佳视角选择问题, 多视角提取的特征间存在大量冗余信息, 固定视角的方法又有模型的朝向对齐要求. 接下来, 本文将讨论如何进行将视角选择方法引入多视角卷积神经网络来解决这些问题.

3 基于视点熵的视角选择

本节主要讨论如何将视角选择方法引入多视角卷积神经网络中. 在 MVCNN^[6]中, 讨论了基于 80 个视角的 MVCNN 框架, 80 个视角的引入, 可以最大程度地覆盖模型的所有位置, 从而不需要模型的朝向对齐要求, 然而大量的视角投影图直接输入到网络中又会引入更多的信息冗余并消耗大量的计算与存储资源. 所以, 通过上述的讨论, 本文需要解决以下问题:

- (a) 去除模型的朝向对齐要求;
- (b) 减少视角个数的同时最大程度地保留模型的几何细节信息.

为解决上述问题, 本文将视角选择方法引入多视角卷积神经网络框架中. 使用视点熵作为视角选择方法. 视点熵可用来衡量该视角方向下对所看到该模型的信息量, 能够很好地反映人的视觉特征,

图 1 (网络版彩图) 基于视点熵的最佳视角选择

Figure 1 (Color online) Viewpoint selection based on viewpoint entropy. (a) illustrates viewpoint selection process; (b) is a projection image at a viewpoint

基于视点熵的最佳视角选择方法已在场景理解领域被广泛地应用^[9]. 由于视点熵是基于模型面片的投影面积与投影球的面积之比进行计算的, 可以避免模型面片面积大小分布不均对视角选择的影响. 同时基于视点熵的最佳视角计算是全局的, 不需要模型的朝向对齐一致. 与 Vázquez 等^[9]提出的方法一样, 计算投影面在投影球体上的相对面积作为概率分布, 本文采用的视点熵的计算方法如下:

$$I(s, v) = - \sum_{i=0}^{N_f} \frac{A_i}{A_t} \log \frac{A_i}{A_t},$$

其中, $I(s, v)$ 是模型 s 在视角 v 方向下的视点熵; N_f 是模型的面片数; A_i 是在视角 v 方向下, 模型第 i 个面片在投影球面上的投影面积; A_t 指投影球的表面积. 为了方便计算每个视角的视点熵, 本文先设置了 42 个固定视角. 固定视角的设置方式为: 水平面内相邻视角与坐标原点的连线之间的夹角为 45° , 垂直面内相邻视角与坐标原点的连线之间的夹角为 30° , 以及正上方与正下方各一个. 如图 1 所示, (a) 表示预先设置的 42 个固定视角, (b) 表示其中一个视角方向下的投影渲染图. 从图上可看出, 本文选择的 42 个视角可以基本覆盖模型的所有位置, 所以本文方法并不需要模型朝向的一致对齐. 在实际应用时, 本文在固定的 42 个视角中选择视点熵最大的前 12 个视角作为最佳视角集合. 每个视角下渲染投影图并将背景设为透明. 图 2 所示的是基于视点熵方法选取的三维模型的 12 个最佳视角与 MVCNN^[6] 固定选取的 12 个视角的对比. 图 2(a) 表示 MVCNN 12 个固定视角的设置方式; 图 2(b) 为基于视点熵从 42 个固定视角中选择最佳的 12 个视角的设置方式; 图 2(c) 为 (a) 中 12 个固定视角设置方式下的投影渲染图; 图 2(d) 为基于视点熵的 12 个最佳视角选择方式下的投影渲染图. 从可视化结果中可以看到基于视点熵的方法会更全面的覆盖模型的表面信息, 防止模型几何信息的丢失.

4 多视角信息融合

4.1 基于最大池化的特征融合

第 1 篇基于多视角特征融合的卷积神经网络 MVCNN 由 Su 等^[6] 在 2015 年提出. MVCNN^[6]

图 2 (网络版彩图) 固定视角与基于视点熵的视角选择对比

Figure 2 (Color online) Comparison of viewpoints selected based on viewpoint entropy method and the fixed viewpoints. (a) Fixed viewpoints; (b) viewpoints selected based on viewpoint entropy; (c) projection images at fixed viewpoints; (d) projection images at viewpoints selected based on viewpoint entropy

将采集的 12 个固定视角方向的投影渲染图输入到卷积神经网络 CNN1 中, 对 CNN1 从这 12 个视角下的投影渲染图中提取到的单视角特征进行视角池化操作. 视角池化操作与传统池化操作类似, 都是按通道对一个批处理内的图像数据进行操作. 传统池化过程是每张图像独立操作, 对图像大小进行下采样, 聚合局部窗口的特征. 而视角池化对一个批处理内的所有图像统一进行操作, 按像素位置聚合每张图像的特征, 最后输出一张图像, 没有下采样过程. 一个三维模型的多视角图像作为一个批处理放入网络中, 经过视角池化过程, 输出多视角聚合后的特征图. MVCNN 中用视角最大池化操作 (max pooling) 来实现操作, 得到的特征输入到卷积神经网络 CNN2 中, 最终得到输入三维模型的类别预测概率.

单视角的特征提取网络 CNN1 为一个预训练好的去掉全连接层的 AlexNet 网络^[30]. AlexNet 网络为经典的卷积神经网络, 可用于二维图像上的分类任务, 由 5 次卷积、3 次池化与 3 次全连接操作构成. 激活函数采用修正线性单元 (ReLU), ReLU 可以使得神经网络在训练过程中梯度保持稳定, 避免出现梯度消失现象. 在全连接层使用了 Dropout 方法来避免过拟合, Dropout 会随机抑制一部分神经元不工作, 极大地提高了神经网络的泛化性, 避免过拟合的发生.

在 MVCNN 论文中报告在分类数据上先进行 AlexNet 的微调有助于最终融合多视角特征的分类效果. CNN2 为 3 层全连接结构, 最后输出每个类别的预测概率. MVCNN 的实验证明了在利用 CNN1 得到每个视角下的特征后, 利用视角池化来得到聚合特征并进行分类, 可以提高了模型的分类准确性,

证明了视角池化进行多视角特征融合的有效性. 然而在实际训练的过程中, 本文发现训练网络依然存在一定程度的过拟合, 而且 MVCNN^[6] 讨论了基于 80 个视角的 MVCNN 网络在分类结果上照比基于 12 个视角的结果提升是微弱的. 本文猜测造成这个现象的原因是因为单视角下的特征可能存在大量的信息冗余, 如果融合这些冗余与无效的信息, 可能增加过拟合风险, 降低网络的可泛化性. 本文以此为出发点, 设计视角判别网络, 让网络可自动过滤掉无效或冗余的信息, 提升多视角卷积神经网络学习特征的鲁棒性.

4.2 基于视角判别的多视角信息融合网络

提升多视角卷积神经网络学习特征的鲁棒性, 减少冗余信息对最后结果的影响. 可以从两方面角度考虑. 一方面, 参考 Dropout 方法的设计出发点, Dropout 使得中间层的每一个神经元有一定的概率被抑制 (在网络中该节点值设为 0), 该方法可在一定程度上缓解神经元之间的相互依赖关系, 被设计用来减少模型的过拟合性. 考虑到全连接层的参数较多, 比较容易过拟合, 在 MVCNN 的 CNN2 中的两层全连接层使用了该方法, 然而 CNN2 发生在多视角特征聚合之后, 所以本文考虑如何进行单视角下的冗余信息的抑制, 设计一个额外的网络让网络自动判断视角的信息的合理性, 给每一个视角一个 0 到 1 的权值, 来抑制冗余的视角信息, 提升视角特征学习的鲁棒性.

另一方面, 本文在网络训练阶段, 为了防止出现过拟合问题, 设计了一种策略: 每个视角的投影图有一定概率被替换为无效的数据 (另一个模型的某张投影图, 实验中概率设为 20%). 这种训练策略加大了网络的训练难度, 但是引入的无效信息可以帮助网络更好地训练而避免过拟合, 也可以让视角判别网络能更好地判断视角的信息的合理性, 提升网络的鲁棒性.

本文从这两方面考虑, 设计了一种多视角信息融合的方法来处理多视角特征, 给每个视角的信息一个权重, 这样能较好地过滤掉视角中无效或冗余的信息. 与 MVCNN 网络架构大致相同, 本文方法先用 CNN1 来提取单视角下的特征, 再将每个单视角的特征进行视角池化操作作为初始多视角融合, 每个视角下的单视角特征与初始多视角融合特征进行拼接操作 (concatenation), 拼接后的信息通过另一个视角判别网络输出该视角的权重. Concat 表示特征拼接操作, V 表示视角集合, 每个视角下的权重的求取公式如下:

$$w_i = \text{MLP} \left(\text{Concat} \left(\sum_i^I S_{V_i}, S_{V_i} \right) \right).$$

MLP 为 3 层全连接层, 在最后每个视角权值的归一化阶段, 使用 Sigmoid 操作使得每个视角的权值在 0 到 1 之间. Sigmoid 激活函数的公式为

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}.$$

原视角下的特征乘以该视角下的权重即为该视角的最终特征. 每个视角的最终特征通过视角池化操作来进行多视角的特征融合. 如图 3 所示为整个网络流程的示意图, 与 MVCNN 相比本文方法加入一个额外的视角判别网络来输出每个视角的一个贡献值. 接下来的实验阶段, 本文着重分析视角选择方法与本文提出的多视角信息融合方法对最终分类准确率的影响.

5 实验分析

5.1 实验数据集

为了进行不同方法之间的对比, 需要在统一的标准数据集上进行实验, 采用相同的训练集和测试集

图 3 (网络版彩图) 多视角信息融合网络整体结构图

Figure 3 (Color online) Multi-view information fusion network structure

的数据划分和评价标准. 为了进行三维模型分类实验, 本文采用三维模型分类的标准数据集 ModelNet, 它是由 Princeton University 研究人员公开的用于三维模型分类的标准数据集, 包含来自 662 类的 12715 个三维模型. ModelNet 提供了两个分类标准数据集: ModelNet10 和 ModelNet40, 分别有 10 个类和 40 个类, 诸多研究方法都是在这两个数据集上进行实验. 以 ModelNet40 为例, ModelNet40 数据集包括 12311 个模型、40 个大类别. 官方提供的训练测试划分为 9843 训练和 2468 测试, 使用这一划分的结果的测试结果被称为实例准确率 (instance accuracy). 而在文献 [6, 16] 中使用了另一种划分方式: 每个类别的测试目录前 20 个模型为测试集, 训练目录前 80 个模型为训练集, 使用这一划分的结果的测试结果被称为类别准确率 (class accuracy). 为方便与 MVCNN 等方法进行对比, 本文使用第 2 种形式的划分. 而 ModelNet10 使用了官方提供的划分方式.

5.2 实验环境

对模型采集投影图等数据以及反投并渲染过程的实验平台参数为 Windows10 64 位操作系统、Intel(R) Core(TM) i5-3470 CPU 3.20 GHz、内存 10 GB, 采用 C++ 编程语言, 并结合 OpenGL 和 OpenCV 第三方开源库来实现, 编程开发环境为 Visual Studio 2015. 本文在多视角融合卷积神经网络的训练和测试过程的实验平台参数为 Windows10 64 位操作系统、Intel(R) Core(TM) i7-5820K CPU 3.30 GHz、内存 64 GB, 显卡为 Titan X GPU 12 GB, 采用 Caffe 第三方开源库来实现. 在实验中, 多视角融合卷积神经网络训练过程采用随机梯度下降 (stochastic gradient descent, SGD) 优化算法, 权重衰减 (weight decay) 的值设为 0.0005, 采用 Step 学习策略, 其中学习率 (learning rate) 设置为 0.001, Gamma 值设为 0.8, step size 设为 2500, 批处理大小 (batch-size) 设为 32, 这些参数可以帮助本文的网络框架取得更好的效果. 与 MVCNN^[6] 相同, 本文也使用了单视角预训练的 AlexNet 网络来进行微调.

5.3 视角面片覆盖率与视角数关系的实验

为了验证基于视点熵的视角选择方法对模型细节特征的保留情况, 本文实验了模型可见面片覆盖率以及视角数之间的关系, 同时确定了本文方法中最佳视角数. 本文在 ModelNet40 整个模型集上进行了实验, 统计出了平均视点熵、平均可见面片覆盖率与视角数之间的关系. 如图 4 所示, 随着视角数的增加, 平均视点熵和平均模型可见面片覆盖率均在上升. 可以看出视点熵和模型可见面片覆盖率存在正相关关系, 且随着视角数的增加而增加. 但是视角数越多需要采集的模型投影渲染图的数目就越

图 4 (网络版彩图) 不同视角数下视点熵与可见面片覆盖率之间的关系

Figure 4 (Color online) Relationship between viewpoint entropy and visible faces coverage under different number of viewpoints

多, 网络对 GPU 显存的消耗也会越大. 为了兼顾效率和效果, 本文选择将视角数设置为 12. 此时各类别的平均模型可见面片覆盖率均在 90% 以上. 同时, 当视角数大于 12 并逐渐增加时, 视点熵及模型可见面片覆盖率则增加平缓.

5.4 视角选择方法对比

为了研究基于视点熵的最佳视角选择方法对三维模型分类准确率的影响, 本文将其和 MVCNN^[6] 的固定视角方式进行了对比. 实验数据使用了 ModelNet40, 为了控制变量, 本实验将 MVCNN 中的 12 个固定视角设置方式、80 个固定视角设置方式、MVCNN 的多尺度改进版 MVCNN-MultiRes^[7] 中 20 个固定视角设置方式与本文基于视点熵的最佳视角选择方式进行了对比, 只是视角选择方式不同, 网络框架则相同. 本文测试了基于视点熵的 7, 9, 12, 20 个视角下的三维模型分类准确率. 注意 MVCNN-MultiRes^[7] 中聚合了多个尺度的特征信息, 而本文利用视点熵的视角选择方法在 12 个视角的情况下就已经超过其准确率. 表 1 所示为视角选择对分类准确率影响的对比, 最优结果在表中用黑体标出. 从表中可以看出, 本文基于视点熵的最佳视角选择方法只需要选出 9 个视角, 就可以在 ModelNet40 数据集的准确率水平上超过 MVCNN 中 12 个视角的准确率水平. 而 12 个以上的视角对三维模型分类准确率的提升并不明显, 说明视角间的冗余信息较大, 不应该简单考虑增多视角.

5.5 三维模型分类准确率对比

为了进一步验证本文设计的多视角信息判别融合卷积神经网络在处理三维模型分类问题上的有效性, 本文分别在 ModelNet10 和 ModelNet40 数据集上将其和其他领先的方法进行了对比.

表 2 所示为本文方法与其他三维模型分类领域领先的方法在 ModelNet10 和 ModelNet40 数据集上的准确率对比. 表中本文方法代表加入多视角加权判别网络与替换训练策略后的结果, 本文方法 (固定) 为利用 MVCNN 12 个固定视角加入本文提出的多视角加权判别网络与替换训练策略后的分类

表 1 视角选择对分类准确率影响的对比

Table 1 Comparison of the influence of perspective selection on classification accuracy

Method	#Views	Accuracy (ModelNet40) (%)
MVCNN [6]	12	89.9
	80	90.1
MVCNN-MultiRes [7]	20	91.4
MVCNN (viewpoint entropy)	7	89.7
	9	90.3
	12	91.6
	20	91.7

表 2 视角选择对分类准确率影响的对比

Table 2 Comparison of the influence of perspective selection on classification accuracy

Method	#Views	Accuracy (ModelNet10) (%)	Accuracy (ModelNet40) (%)
Ours (fixed viewpoints)	12	93.8	90.9
Ours (viewpoint entropy)	12	95.1	92.2
MVCNN [6]	12	—	89.9
	80	—	90.1
PANORAMA-NN [28]	—	91.1	90.7
Pairwise [27]	—	92.8	90.7
MVCNN-MultiRes [7]	20	—	91.4
KD-Networks [24]	—	94.0	91.8
PointNet [21]	—	—	86.2
3D-GAN [20]	—	91.0	83.3
3DShapeNets [16]	—	83.5	77.0

结果 (表中黑体表示最佳结果), 本文方法 (视点熵) 为利用基于视点熵法选择的 12 个最佳视角加入本文提出的多视角判别网络与替换训练策略后的分类结果. 本文方法在基于投影的方法中所用的视角数少并且准确率高, 也超过了其他非投影的方法, 发挥出了基于投影的方法在三维模型分类领域的优势, 进一步验证了基于视点熵的最佳视角选择方式能够减少多视角间的信息冗余, 有效地选取模型类别信息蕴含丰富的视角, 最终提升三维模型分类的准确率.

5.6 分类 Top1~5 错误率对比实验

为了进一步研究本文设计的多视角融合卷积神经网络在处理三维模型分类问题上的表现, 本文分别在 ModelNet10 和 ModelNet40 数据集上进行了实验, 选择了图像分类领域常用的 Top1~5 错误率衡量标准, 分析了本文方法在各个水平上的错误率. 表 3 所示为本文方法分别在 ModelNet10 和 ModelNet40 模型集上分类 Top1~5 的错误率. 以 Top5 为例, 它是指将模型预测为各个类别标签的概率最大的前 5 个类别中有 1 个正确即分类正确, 由 Top5 的正确率得出其错误率. 它们是分类领域常见的衡量不同方法分类效果好坏的标准.

可以看出, 本文方法的 Top5 错误率已经很小了, 说明本文基于视点熵的多视角融合卷积神经网络方法在处理三维模型分类问题能够提取出模型的有效类别特征, 从而实现较好的分类.

表 3 本文方法分类 Top1~5 错误率
Table 3 Top1~5 error rate by using this paper's method

Measure method	Error rate (ModelNet10) (%)	Error rate (ModelNet40) (%)
Top1	4.84	7.82
Top2	3.87	6.36
Top3	3.26	5.19
Top4	2.69	4.23
Top5	2.18	3.07

图 5 (网络版彩图) ModelNet40 部分分类结果可视化

Figure 5 (Color online) ModelNet40 partial classification results visualization. (a) Before classification; (b) after classification

5.7 ModelNet40 分类结果可视化

为了进一步研究本文设计的多视角融合卷积神经网络的效果, 本文对 ModelNet40 数据集上的分类结果进行了可视化, 通过分析其中的分类正确和分类错误的实例, 来进一步指导对本文方法的修改. 图 5 所示为本文基于视点熵的多视角融合卷积神经网络方法在 ModelNet40 数据集上的部分模型分类结果的可视化, (a) 为分类前的 ModelNet40 部分模型集渲染图, (b) 为分类后的 ModelNet40 部分模型集渲染图. 由于 person 和 tent, stair, radio, laptop 类别之间差别较大, 全部分类正确. 而 tent 和 radio, stair 和 radio 之间存在极为相似的模型, 如图中圈中所示, 出现了个别模型的分​​类错误. 对于极为相似的模型, 只有有效地提取出能够帮助区分类别的局部特征才能实现分类. 这对于其他方法也是一个不小的挑战. 由此可以看出, 本文基于视点熵的多视角融合卷积神经网络方法也在处理差异较大的模型分类时表现较好, 而在处理极为相似的模型的类别判别时出现了个别的偏差现象, 这可能是由于面对这种情况时局部信息提取得不好. 如果尝试多尺度提取投影图中的特征, 使得提取的特征中既包含局部特征, 也包含全局特征可能会使本文方法能够适应这种情况.

5.8 多视角聚合特征的可视化显示

为了验证多视角融合卷积神经网络提取的三维模型的类别特征是否有效, 本文将网络最后一层的特征提取出来, 通过 t-SNE^[31] 可视化方法进行了展示. 图 6 所示为选择的部分 ModelNet40 模型, 利用本文提出的多视角聚合网络进行分类时得到的最后一层的特征 t-SNE^[31] 可视化显示. 通过 t-SNE^[31] 方法, 本实验将高维特征映射到二维平面空间内, 每一个点则表示被映射的一个模型分类后的特征, 不同颜色代表不同类别. 从图中可以看出本文方法能够较好地提取出模型的类别特征, 使得同类模型尽

图6 (网络版彩图) ModelNet40 部分聚类特征可视化

Figure 6 (Color online) ModelNet40 partial clustering features visualization

量分到一起, 而不同种类的模型尽量分得较开.

6 结语

本文采用基于投影的方法, 通过基于视点熵的多视角融合卷积神经网络的深度学习框架从图像分析的角度完成三维模型分类任务. 将基于视点熵的最佳视角选择方法引入多视角融合卷积神经网络中, 减少了视角间的信息冗余. 另一方面, 本文提出了一种新型的多视角卷积神经网络框架, 在网络训练过程中自动判别每个视角的贡献度, 从而提取出最能表征模型类别的特征, 提高了网络的鲁棒性.

实验结果表明本文基于视点熵的多视角融合卷积神经网络框架在处理三维模型分类问题上可以取得很好的效果, 达到了预期目标. 但是也存在着一些局限性, 如最佳视角选择过程没有参与到网络框架的训练, 并非一个整体的端到端框架; 在处理相似模型时需要细化提取的局部特征, 需要对投影图进行多尺度特征提取才会进一步提升本文方法在处理三维模型分类问题上的效果等问题.

参考文献

- 1 Bronstein A M, Bronstein M M, Guibas L J, et al. Shape Google: geometric words and expressions for invariant shape retrieval. *ACM Trans Graph*, 2011, 30: 1–20
- 2 Funkhouser T, Kazhdan M, Min P, et al. Shape-based retrieval and analysis of 3D models. *Commun ACM*, 2005, 48: 58
- 3 Huang Q X, Su H, Guibas L. Fine-grained semi-supervised labeling of large shape collections. *ACM Trans Graph*, 2013, 32: 1–10
- 4 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Massachusetts, 2015. 3431–3440
- 5 Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40: 834–848
- 6 Su H, Maji S, Kalogerakis E, et al. Multi-view convolutional neural networks for 3D shape recognition. In: *Proceedings of IEEE International Conference on Computer Vision*, Santiago, 2015. 945–953

- 7 Qi C R, Su H, Niebner M, et al. Volumetric and multi-view CNNs for object classification on 3D data. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 5648–5656
- 8 Wang C, Pelillo M, Siddiqi K. Dominant set clustering and pooling for multi-view 3D object recognition. In: Proceedings of British Machine Vision Conference, 2017
- 9 Vazquez P P, Feixas M, Sbert M, et al. Viewpoint selection using viewpoint entropy. In: Proceedings of Vision Modeling and Visualization Conference, 2001. 273–280
- 10 Sbert M, Plemenos D, Feixas M. Viewpoint quality: measures and applications. In: Proceedings of the 1st Computational Aesthetics in Graphics, Visualization and Imaging. 2005. 185–192
- 11 Lee C H, Varshney A, Jacobs D W. Mesh saliency. *ACM Trans Graph*, 2005, 24: 659
- 12 Qin F W, Li L Y, Gao S M, et al. A deep learning approach to the classification of 3D CAD models. *J Zhejiang Univ Sci C*, 2014, 15: 91–106
- 13 Socher R, Huval B, Bhat B, et al. Convolutional-recursive deep learning for 3D object classification. In: Proceedings of the 25th Advances in Neural Information Processing Systems, Lake Tahoe, 2012. 665–673
- 14 Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs. 2013. ArXiv:1312.6203
- 15 Masci J, Boscaini D, Bronstein M M, et al. Geodesic convolutional neural networks on riemannian manifolds. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 832–840
- 16 Wu Z R, Song S R, Khosla A, et al. 3D shapenets: a deep representation for volumetric shapes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1912–1920
- 17 Li Y Y, Pirk S, Su H, et al. FPNN: field probing neural networks for 3D data. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016
- 18 Xu X, Todorovic S. Beam search for learning a deep convolutional neural network of 3D shapes. In: Proceedings of International Conference on Pattern Recognition, Columbia, 2017. 3506–3511
- 19 Sedaghat N, Zolfaghari M, Brox T. Orientation-boosted voxel nets for 3D object recognition. In: Proceedings of British Machine Vision Conference, London, 2017
- 20 Ren M W, Niu L, Fang Y. 3D-a-nets: 3D deep dense descriptor for volumetric shapes with adversarial networks. 2017. ArXiv:1711.10108
- 21 Qi C R, Su H, Mo K, et al. PointNet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017
- 22 Qi C R, Yi L, Su H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st Advances in Neural Information Processing Systems, 2017
- 23 Simonovsky M, Komodakis N. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017
- 24 Klovov R, Lempitsky V. Escape from cells: deep Kd-networks for the recognition of 3D point cloud models. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 863–872
- 25 Li J X, Chen B M, Lee G H. SO-Net: self-organizing network for point cloud analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018
- 26 Shi B G, Bai S, Zhou Z C, et al. DeepPano: deep panoramic representation for 3D shape recognition. *IEEE Signal Process Lett*, 2015, 22: 2339–2343
- 27 Johns E, Leutenegger S, Davison A J. Pairwise decomposition of image sequences for active multi-view recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 3813–3822
- 28 Sfikas K, Pratikakis I, Theoharis T. Ensemble of PANORAMA-based convolutional neural networks for 3D model classification and retrieval. *Comput Graph*, 2018, 71: 208–218
- 29 Kanazaki A, Matsushita Y, Nishida Y. RotationNet: joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018
- 30 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th Advances in Neural Information Processing Systems, Lake Tahoe, 2012
- 31 van der Maaten L, Hinton G. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res*, 2008, 9: 2579–2605

3D shape classification based on convolutional neural networks fusing multi-view information

Pengyu WANG, Panpan SHUI, Fenggen YU, Yuan GAN, Kun LIU, Yan ZHANG* & Zhengxing SUN

State Key Lab for Novel Software Technology, Nanjing University, Nanjing 210023, China

* Corresponding author. E-mail: zhangyannju@nju.edu.cn

Abstract In recent years, convolutional neural network (CNN) architecture has achieved good results in the fields of 2D image recognition, detection, and semantic segmentation. However, given the complexity and irregularity of 3D shape structures, CNNs cannot be directly applied to 3D data. With the advantage of the deep learning framework in the field of 2D image analysis, the view-based method can be used for 3D shape classification. However, the existing multi-view based 3D shape classification methods mostly adopt fixed viewpoints. Considerable information redundancy exist in the rendered images, and it can cause certain interference to the results. Herein, we propose a novel multi-view CNN framework, which automatically discriminates the contribution of each viewpoint during the network training and discards the redundant information. In addition, the optimal viewpoint selection method based on viewpoint entropy is introduced into the field of 3D shape classification. In comparison with the fixed viewpoint method, this procedure can retain more detailed information of the shapes and requires no orientation alignment of the model. Experiments on the ModelNet10 and ModelNet40 datasets verify the rationality and superiority of applying the optimal viewpoint selection method based on the viewpoint entropy to 3D model classification and the multi-view information fusion method proposed herein. The experimental results show the better classification accuracy of this method than that of existing 3D model classification methods.

Keywords 3D model, viewpoint selection, viewpoint entropy, convolutional neural network, deep learning

Pengyu WANG received his B.S. degree in digital media technology in 2016. He is currently pursuing his M.S. degree in Department of Computer Science and Technology, Nanjing University, China. His research interests include computer vision and computer graphics.

Panpan SHUI received his B.S. and M.S. degrees in computer science and technology in 2015 and 2018, respectively. He is currently working at Netease. His research interests include computer vision and computer graphics.

Yan ZHANG received her Ph.D. degree in 2006 in computer science and technology from Jilin University. She is currently an associate professor in Department of Computer Science and Technology, Nanjing University. Her main research interests include computer vision and computer graphics.

Zhengxing SUN received his Ph.D. degree in 1996 from Nanjing University of Aeronautics and Astronautics. He is currently a professor in Department of Computer Science and Technology, Nanjing University. His research interests include computer vision and computer graphics.