(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116740422 A
(43)申请公布日 2023.09.12

(21)申请号 202310594805.2

(22)申请日 2023.05.24

(71)申请人 中国科学院空天信息创新研究院
地址 100101 北京市朝阳区大屯路甲20号中国科学院遥感与数字地球研究所A座203室

(72)发明人 洪丹枫 姚靖 李晨玉 张兵

(74)专利代理机构 北京国昊天诚知识产权代理有限公司 11315
专利代理师 南霆

(51)Int.Cl.
G06V 10/764(2022.01)
G06V 10/774(2022.01)
G06V 10/80(2022.01)
G06T 7/10(2017.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)

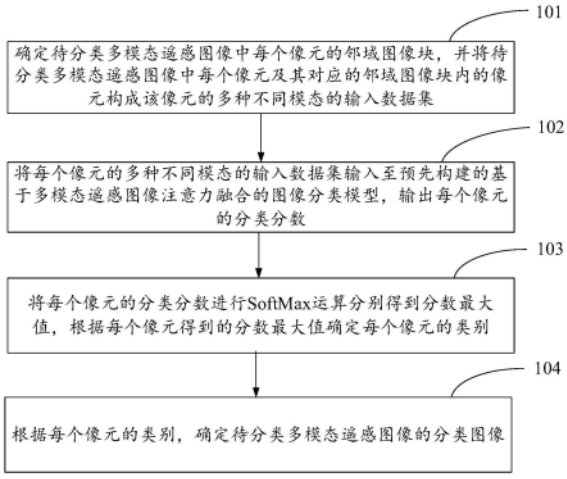
权利要求书2页 说明书12页 附图3页

(54)发明名称

基于多模态注意力融合技术的遥感图像分类方法及装置

(57)摘要

本发明公开了一种基于多模态注意力融合技术的遥感图像分类方法及装置。包括:确定待分类多模态遥感图像中每个像元的邻域图像块,并将待分类多模态遥感图像中每个像元及其对应的邻域图像块内的像元构成该像元的多种不同模态的输入数据集;将每个像元的多种不同模态的输入数据集输入至预先构建的基于多模态遥感图像注意力融合的图像分类模型,输出每个像元的分类分数;将每个像元的分类分数进行SoftMax运算分别得到分数最大值,根据每个像元得到的分数最大值确定每个像元的类别;根据每个像元的类别,确定待分类多模态遥感图像的分类图像。从而不仅能够应用于多种模态数据,而且具有跨模态挖掘数据之间相关及互补信息的能力,可以获得更高分类精度。



CN 116740422 A

1.一种基于多模态注意力融合技术的遥感图像分类方法,其特征在于,包括:

确定待分类多模态遥感图像中每个像元的邻域图像块,并将所述待分类多模态遥感图像中每个像元及其对应的所述邻域图像块内的像元构成该像元的多种不同模态的输入数据集,其中所述待分类多模态遥感图像包括以下任意一种或多种模态遥感图像:高光谱图像、激光雷达点云数据、多光谱图像以及合成孔径雷达图像;

将每个像元的多种不同模态的所述输入数据集输入至预先构建的基于多模态遥感图像注意力融合的图像分类模型,输出每个像元的分类分数;

将每个像元的所述分类分数进行SoftMax运算分别得到分数最大值,根据每个像元得到的所述分数最大值确定每个像元的类别;

根据每个像元的类别,确定所述待分类多模态遥感图像的分类图像。

2.根据权利要求1所述的方法,其特征在于,所述图像分类模型的网络模型结构包括:

输入层,包括多个输入模块,用于输入每个像元的不同模态输入数据集;

多模态特征提取与标记层,利用深度可分离卷积将所述输入数据集处理成下游跨模态注意力融合层所需的带位置信息的标记嵌入数据;

所述跨模态注意力融合层,用于将每个像元不同模态的所述标记嵌入数据进行信息融合,输出融合后的标记嵌入序列;

标记融合层,通过多层感知机前置头将所述标记嵌入序列加权求和融合成用于分类的分类标记嵌入,并通过多层感知机头根据每个像元的所述分类标记嵌入,输出每个像元的分类分数。

3.根据权利要求2所述的方法,其特征在于,所述跨模态注意力融合层采用晚期跨模态注意力融合,其中晚期跨模态注意力融合规则的起始层索引中在图像分类模型中编码器总层数中通过超参数设置索引。

4.根据权利要求1所述的方法,其特征在于,根据输入模态数确定所述图像分类模型训练时所使用的最小损失函数中数据集数量。

5.根据权利要求4所述的方法,其特征在于,还包括:

使用Adam优化器,设置学习初始率为0.0005,每20步衰减0.9倍迭代训练所述最小损失函数,以获得所述图像分类模型的最优模型。

6.一种基于多模态注意力融合技术的遥感图像分类装置,其特征在于,包括:

构成模块,用于确定待分类多模态遥感图像中每个像元的邻域图像块,并将所述待分类多模态遥感图像中每个像元及其对应的所述邻域图像块内的像元构成该像元的多种不同模态的输入数据集,其中所述待分类多模态遥感图像包括以下任意一种或多种模态遥感图像:高光谱图像、激光雷达点云数据、多光谱图像以及合成孔径雷达图像;

输出模块,用于将每个像元的多种不同模态的所述输入数据集输入至预先构建的基于多模态遥感图像注意力融合的图像分类模型,输出每个像元的分类分数;

第一确定模块,用于将每个像元的所述分类分数进行SoftMax运算分别得到分数最大值,根据每个像元得到的所述分数最大值确定每个像元的类别;

第二确定模块,用于根据每个像元的类别,确定所述待分类多模态遥感图像的分类图像。

7.根据权利要求6所述的装置,其特征在于,所述图像分类模型的网络模型结构包括:

输入层,包括多个输入模块,用于输入每个像元的不同模态输入数据集;

多模态特征提取与标记层,利用深度可分离卷积将所述输入数据集处理成下游跨模态注意力融合层所需的带位置信息的标记嵌入数据;

所述跨模态注意力融合层,用于将每个像元不同模态的所述标记嵌入数据进行信息融合,输出融合后的标记嵌入序列;

标记融合层,通过多层感知机前置头将所述标记嵌入序列加权求和融合成用于分类的分类标记嵌入,并通过多层感知机头根据每个像元的所述分类标记嵌入,输出每个像元的分类分数。

8.根据权利要求7所述的装置,其特征在于,所述跨模态注意力融合层采用晚期跨模态注意力融合,其中晚期跨模态注意力融合规则的起始层索引中在图像分类模型中编码器总层数中通过超参数设置索引。

9.一种计算机可读存储介质,其特征在于,所述存储介质存储有计算机程序,所述计算机程序用于执行上述权利要求1-5任一所述的方法。

10.一种电子设备,其特征在于,所述电子设备包括:

处理器;

用于存储所述处理器可执行指令的存储器;

所述处理器,用于从所述存储器中读取所述可执行指令,并执行所述指令以实现上述权利要求1-5任一所述的方法。

基于多模态注意力融合技术的遥感图像分类方法及装置

技术领域

[0001] 本发明涉及图像处理技术领域，并且更具体地，涉及一种基于多模态注意力融合技术的遥感图像分类方法及装置。

背景技术

[0002] 遥感图像分类在诸如城市规划、精准农林、矿物勘探、环境监测以及军事侦察等军民应用领域有着广泛的应用价值。对遥感图像进行像元级分类的目的是对图像中的每个像元根据其包含的土地覆盖或土地利用情况标注有意义的类别属性。

[0003] 近年来，随着我国卫星发射与成像技术的飞速发展，人们可以更加便捷地获取多种模态的遥感图像数据，为更好地进行图像分类创造了可能，另外，如何利用这些多模态数据，提取对于分类任务有效的信息，也为相关方法研究带来了新的挑战。由于遥感图像的特殊性，例如采集方式、环境干扰、地物复杂等因素影像，使得对遥感图像进行逐像元级的标注代价较高。因此，近年来许多学者关注到关于遥感图像分类方法的研究。

[0004] 遥感图像分类方法可大体分为基于传统优化建模和基于深度学习网络建模两类。基于传统优化建模的方法虽然起源更早、发展时间更长，但通常需要人为地构造较为复杂的数学模型，缺乏对于数据自有信息的充分利用。基于深度学习的模型大多为单模态模型，由于其专用的模块化设计限制了其实用性和可移植性；而现有的深度多模态方法很少能有效地建模跨模态依赖关系，从而阻碍了其进一步突破性能瓶颈。

发明内容

[0005] 针对现有技术的不足，本发明提供一种基于多模态注意力融合技术的遥感图像分类方法及装置。

[0006] 根据本发明的一个方面，提供了一种基于多模态注意力融合技术的遥感图像分类方法，包括：

[0007] 确定待分类多模态遥感图像中每个像元的邻域图像块，并将待分类多模态遥感图像中每个像元及其对应的邻域图像块内的像元构成该像元的多种不同模态的输入数据集，其中待分类多模态遥感图像包括以下所述的任意一种或多种模态遥感图像：高光谱图像、激光雷达点云数据、多光谱图像以及合成孔径雷达图像；

[0008] 将每个像元的多种不同模态的输入数据集输入至预先构建的基于多模态遥感图像注意力融合的图像分类模型，输出每个像元的分类分数；

[0009] 将每个像元的分类分数进行SoftMax运算分别得到分数最大值，根据每个像元得到的分数最大值确定每个像元的类别；

[0010] 根据每个像元的类别，确定待分类多模态遥感图像的分类图像。

[0011] 可选地，图像分类模型的网络模型结构包括：

[0012] 输入层，包括多个输入模块，用于输入每个像元的不同模态输入数据集；

[0013] 多模态特征提取与标记层，利用深度可分离卷积将输入数据集处理成下游跨模态

注意力融合层所需的带位置信息的标记嵌入数据；

[0014] 跨模态注意力融合层，用于将每个像元不同模态的标记嵌入数据进行信息融合，输出融合后的标记嵌入序列；

[0015] 标记融合层，通过多层感知机前置头将标记嵌入序列加权求和融合成用于分类的分类标记嵌入，并通过多层感知机头根据每个像元的分类标记嵌入，输出每个像元的分类分数。

[0016] 可选地，跨模态注意力融合层采用晚期跨模态注意力融合，其中晚期跨模态注意力融合规则的起始层索引中在图像分类模型中编码器总层数中通过超参数设置索引。

[0017] 可选地，根据输入模态数确定图像分类模型训练时所使用的最小损失函数中数据集数量。

[0018] 可选地，还包括：

[0019] 使用Adam优化器，设置学习初始率为0.0005，每20步衰减0.9倍迭代训练最小损失函数，以获得图像分类模型的最优模型。

[0020] 根据本发明的另一个方面，提供了一种基于多模态注意力融合技术的遥感图像分类装置，包括：

[0021] 构成模块，用于确定待分类多模态遥感图像中每个像元的邻域图像块，并将待分类多模态遥感图像中每个像元及其对应的邻域图像块内的像元构成该像元的多种不同模态的输入数据集，其中待分类多模态遥感图像包括以下所述的任意一种或多种模态遥感图像：高光谱图像、激光雷达点云数据、多光谱图像以及合成孔径雷达图像；

[0022] 输出模块，用于将每个像元的多种不同模态的输入数据集输入至预先构建的基于多模态遥感图像注意力融合的图像分类模型，输出每个像元的分类分数；

[0023] 第一确定模块，用于将每个像元的所述分类分数进行SoftMax运算分别得到分数最大值，根据每个像元得到的所述分数最大值确定每个像元的类别；

[0024] 第二确定模块，用于根据每个像元的类别，确定待分类多模态遥感图像的分类图像。

[0025] 根据本发明的又一个方面，提供了一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序用于执行本发明上述任一方面所述的方法。

[0026] 根据本发明的又一个方面，提供了一种电子设备，所述电子设备包括：处理器；用于存储所述处理器可执行指令的存储器；所述处理器，用于从所述存储器中读取所述可执行指令，并执行所述指令以实现本发明上述任一方面所述的方法。

[0027] 本申请实施例采用的上述至少一个技术方案能够达到以下有益效果：

[0028] 本发明提出的基于多模态注意力融合技术的遥感图像分类方法，为多模态遥感图像像元级分类任务构建简洁、通用的深度学习图像分类模型ExViT。该模型能够有效提取单模态数据的空间-通道信息，并且实现对异构模态特征从浅到深的高效融合。从而本发明不仅能够应用于多种模态数据，而且具有跨模态挖掘数据之间相关及互补信息的能力，可以获得更高分类精度。

附图说明

[0029] 此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申

请的示意性实施例及其说明用于解释本申请,并不构成对本申请的不当限定。在附图中:

[0030] 图1是本发明实施例提供的基于多模态注意力融合技术的遥感图像分类方法的流程示意图;

[0031] 图2是本发明实施例提供的基于多模态遥感数据注意力融合的图像分类模型的整体结构图;

[0032] 图3是本发明实施例提供的多模态特征提取与标记层中可分离卷积示意图;

[0033] 图4是本发明实施例提供的跨模态注意力融合层编码示意图;

[0034] 图5是本发明实施例提供的早、中、晚期跨模态注意力融合编码示意图;

[0035] 图6是本发明实施例提供的基于多模态注意力融合技术的遥感图像分类装置的结构示意图;

[0036] 图7是本发明实施例提供的电子设备的结构。

具体实施方式

[0037] 为使本申请的目的、技术方案和优点更加清楚,下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然,所描述的实施例仅是本申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本申请保护的范围。

[0038] 多模态(Multimodal)通常可以被广义地定义为使用不同的信息或属性(例如来源于不同的内容、不同的传感器、不同的分辨率等)描述相同的物体或场景。

[0039] 高光谱成像(Hyperspectral Imaging,缩写HSI),是一项能够同时收集目标区域的1维光谱特征与2维几何空间信息的开创性多维信息获取技术。所获得的图像呈现出“图像立方体”的形式。

[0040] 地面高程(Ground Elevation)是指地面某点的高程。在每个地方都有所在国家控制点,在我国地面高程是以黄海平面为参考平面的竖向高度。

[0041] 数字表面模型(Digital Surface Model,缩写DSM)是指包含了地表建筑物、桥梁和树木等高度的地面高程模型。DSM不仅包含了地形的高程信息,进一步涵盖了除地面以外的其它地表信息的高程。在一些对林木、建筑物高度有需求的领域,得到了很大程度的重视。

[0042] 注意力机制(Attention Mechanism)是一种在深度学习神经网络中模仿人类认知注意力的技术:试图强化重要部分,同时淡化不相关信息的方法。在计算机视觉领域,可以看作是基于输入图像特征的动态权重调整过程,其作用是增强输入数据某些部分的影响,而减弱其他部分,促使网络把更多的注意力放在小,但重要的部分上。

[0043] 自注意力机制(self-attention)是将单个序列中不同位置的元素关联起来,以便计算该序列表示的一种注意机制。其原理是,通过学习一种相关性度量,并用其对序列中任意位置的元素进行加权,从而允许模型捕获任意位置元素的信息,由此,模型具有感知长程上下文信息的能力。

[0044] 以下结合附图,详细说明本申请各实施例提供的技术方案。

[0045] 图1是本发明实施例提供的基于多模态注意力融合技术的遥感图像分类方法的流程示意图。本实施例可应用在电子设备上。如图1所示,基于多模态注意力融合技术的遥感

图像分类方法包括以下步骤：

[0046] 步骤101，确定待分类多模态遥感图像中每个像元的邻域图像块，并将待分类多模态遥感图像中每个像元及其对应的邻域图像块内的像元作为该像元的输入数据集，其中待分类多模态遥感图像包括以下所述的任意一种或多种模态遥感图像：高光谱图像、激光雷达点云数据、多光谱图像以及合成孔径雷达图像。

[0047] 具体地，对待分类多模态遥感图像进行数据处理：加载同一场景下的任意格栅式遥感图像x^m，对于图像中已标注像元，设置中心像元邻域块大小(patch size，为超参数)为p。那么，每个像元对应的每种模态下邻域图像块为$\mathbf{x}_i^m \in R^{p\times p\times d_m}$，$R^{p\times p\times d_m}$表示维度为p×p×$d_m$的三维实数矩阵，其中$d_m$为光谱带的数量，$\mathbf{x}_i^m$表示格栅式遥感图像$x^m$的第i个像元。将获得的图像块与像元标记构成输入数据集$\left\{\left(\mathbf{x}_i^1,\mathbf{x}_i^2\right),\mathbf{y}_i\right\}$，$i\in 1,\cdots,N$。

[0048] 此外，遥感图像不限于高光谱图像、激光雷达点云数据、多光谱图像合成孔径雷达图像以及激光雷达点云数据，也可以是其他模态的图像数据。

[0049] 步骤102，将每个像元的多种不同模态的输入数据集输入至预先构建的基于多模态遥感图像注意力融合的图像分类模型，输出每个像元的分类分数。

[0050] 可选地，图像分类模型的网络模型结构包括：

[0051] 输入层，包括多个输入模块，用于输入每个像元的不同模态输入数据集；

[0052] 多模态特征提取与标记层，利用深度可分离卷积将输入数据集处理成下游跨模态注意力融合层所需的带位置信息的标记嵌入数据；

[0053] 跨模态注意力融合层，用于将每个像元不同模态的标记嵌入数据进行信息融合，输出融合后的标记嵌入序列；

[0054] 标记融合层，通过多层感知机前置头将标记嵌入序列加权求和融合成用于分类的分类标记嵌入，并通过多层感知机头根据每个像元的分类标记嵌入，输出每个像元的分类分数。

[0055] 具体地，参考图2所示，图像分类模型的网络模型结构为ExViT网络结构，ExViT模型采用由双到单的框架结构，除输入、输出外，主要包括三个子步骤。一是多模态特征提取与标记层，包含两个分支，分别对多种模态图像数据每个像元的输入数据集进行提取特征，并表示为后期基于自注意力机制的主干网络跨模态注意力融合层所需要的标记化形式。二是跨模态注意力融合层，一方面，进一步对不同模态特征应用自注意力进行模态内特征提取，另一方面，设置分阶段跨模态注意力实现模态间信息交互。三是标记融合，本发明认为ViT编码器中添加分类标记不是必须的，由于基于像元的标记都是同等处理的，在分类过程中，舍弃这些标记会造成信息丢失。因此，本发明进行对全部像元标记进行混合而生成分类标记，不再额外引入分类标记。

[0056] 将每个像元的输入数据集输入至图像分类模型就可以获得未标准化的分类分数$\widetilde{\mathbf{u}}_i$。

[0057] 可选地，跨模态注意力融合层采用晚期跨模态注意力融合，其中晚期跨模态注意力融合规则的起始层索引中在图像分类模型中编码器总层数中通过超参数设置索引。

[0058] 具体地，晚期跨模态注意力融合规则如下：

[0059] $$\tilde{\mathbf{z}}_i^{(l)} = \mathrm{MHSA}\left(\mathrm{LN}\left(\hat{\mathbf{z}}_i^{(l-1)}\right)\right) + \hat{\mathbf{z}}_i^{(l-1)}, l = 1, \cdots, L_2$$

[0060] $$\hat{\mathbf{z}}_i^{(l)} = \mathrm{MLP}\left(\mathrm{LN}\left(\tilde{\mathbf{z}}_i^{(l)}\right)\right) + \tilde{\mathbf{z}}_i^{(l)}, l = 1, \cdots, L_2$$

[0061] $$l_{cross} = \lceil (1-\alpha) L_2 \rceil$$

[0062] 其中，MHSA()表示多头自注意力机制，MLP()表示多层感知机，LN()为层正则化，L_2为图像分类模型中编码器总层数，l为层索引，l_{cross}为晚期跨模态注意力融合的起始层索引，α＜0.5为计算l_{cross}的超参数，$\tilde{\mathbf{z}}_i^{(l)}$为跨模态注意力融合过程的像元i第l层标记嵌入序列的中间结果，$\hat{\mathbf{z}}_i^{(l)}$为像元i第l层标记嵌入序列，$\hat{\mathbf{z}}_i^{(L_2)}$为跨模态注意力融合层最终输出的像元i对应的标记嵌入序列。

[0063] 具体地，分类分数的计算公式如下：

[0064] $$\tilde{\mathbf{u}}_i = \mathrm{MLP}\left(\mathrm{SoftMax}\left(\hat{\mathbf{z}}_i^{(L_2)} \mathbf{w}^{pre}\right) \times \hat{\mathbf{z}}_i^{(L_2)}\right)$$

[0065] 其中，MLP()表示多层感知机，$w^{pre} \in R^{d \times 1}$表示所述多层感知机前置头参数矩阵，$R^{d \times 1}$表示维度为d×1的实数矩阵，d等于$\hat{\mathbf{z}}_i^{(L_2)}$的维度，$\hat{\mathbf{z}}_i^{(L_2)}$为注意力融合层最终输出的像元i的标记嵌入序列，$\tilde{\mathbf{u}}_i$为像元i的分类分数。

[0066] 可选地，根据输入模态数确定图像分类模型训练时所使用的最小损失函数中数据集数量。

[0067] 具体地，在输入模态数等于2时，图像分类模型训练时所需的最小损失函数为：

[0068] $$L_{CML} = \beta L\left(\mathbf{y}, \tilde{\mathbf{u}} \middle| \mathbf{x}^1, \mathbf{x}^2\right) + \gamma L\left(\mathbf{y}, \tilde{\mathbf{u}} \middle| \mathbf{x}^m, \mathbf{0}\right)$$

[0069] 当输入模态数大于2时，图像分类模型训练时所需的最小损失函数为：

[0070] $$L_{CML} = \beta L\left(\mathbf{y}, \tilde{\mathbf{u}} \middle| \mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^K\right) + \gamma L\left(\mathbf{y}, \tilde{\mathbf{u}} \middle| \mathbf{x}^m, \mathbf{0}\right)$$

[0071] 其中，β，γ ≥0表示用于控制单模态学习任务相对于多模态学习任务的相对重要性的权衡参数，x^1、x^2、x^K分别表示训练所需的不同模态输入数据集，K为模态总数，m为{1，2，…，K}中的一个或多个，X^m为待预测模态输入数据集，$\tilde{\mathbf{u}}$为输入数据集的分类分数，y为输入数据集场景对应的标签。

[0072] 步骤103，将每个像元的分类分数进行SoftMax运算分别得到分数最大值，根据每个像元得到的分数最大值确定每个像元的类别。

[0073] 具体地，将输出的每个像元的分类分数$\tilde{\mathbf{u}}_i$进行SoftMax运算，最大值对应的索引即为该像元对应的类别。

[0074] 步骤104，根据每个像元的类别，确定待分类多模态遥感图像的分类图像。

[0075] 具体地，将多模态遥感图像的全部像元进行分类后，即可得到该多模态遥感图像的分类图像，完成该多模态遥感图像的分类。

[0076] 可选地，还包括：

[0077] 使用Adam优化器，设置学习初始率为0.0005，每20步衰减0.9倍迭代训练最小损失函数，以获得图像分类模型的最优模型。

[0078] 具体地，本申请中图像分类模型的具体训练步骤如下：

[0079] 步骤一、多模态遥感图像数据生成：

[0080] 1)数据处理，加载同一场景下的任意格栅式遥感图像x^m，对于图像中已标注像元，设置中心像元邻域块大小(patch size，为超参数)为p。那么，每个像元i对应的每种模态下邻域图像块为$\mathbf{x}_i^m \in R^{p\times p\times d_m}$，其中$d_m$为光谱带的数量。本发明训练模型设置模态数量为2，邻域块尺寸为13。

[0081] 2)数据划分与封装，将上一步获得的图像块与像元标记构成数据集$\left\{\left(\mathbf{x}_i^1,\mathbf{x}_i^2\right),\mathbf{y}_i\right\}$，$i\in 1,\cdots,N$，按照比例将N个样本随机划分为训练集($X_{train}$,$Y_{train}$)，验证集($X_{val}$,$Y_{val}$)和测试集($X_{test}$,$Y_{test}$)，并对训练集、验证集以及测试集分别进行封装。本发明中，训练集与测试集的划分方式由公开数据集自身确定，实验过程中从训练集中抽取80％作为训练集，其余作为验证集。

[0082] 3)数据加载，对以上数据集加载过程中，训练集批处理尺寸(batch size)为超参数，具体设置需参考数据集分类类别数以及平台的承载能力；而验证集和测试集批处理尺寸则设置为1。本发明设置训练集批处理尺寸为64。

[0083] 步骤二、构建基于多模态遥感图像注意力融合的网络模型ExViT：

[0084] 1)多模态特征提取与标记层，该层以图像块作为输入数据集，利用深度可分离卷积为每个像元学习信息丰富结构紧凑的表示，之后将像元视为标记(tocken)，通过格栅化等操作生成下游注意力融合所需的带位置信息的标记嵌入数据。

[0085] 如附图2所示，多模态特征提取与标记层每个分支处理一种模态数据，主要包括五个小步骤。

[0086] 第一步是可分离卷积，通过堆叠L_1个深度可分离卷积模块实现。每个模块的结构如附图3所示，包括3层：逐通道卷积(depthwise convolution)层，即在上一层输出的每个通道上独立进行3×3空间卷积操作，生成与原通道数相同的特征图(feature map)；逐点卷积(pointwise convolution)层，即使用1×1卷积将前面的特征图沿深度方向合并映射到新的通道空间，生成的特征图其通道数与卷积核数量d相同；非线性激活层，使用高斯误差线性单元(GeLU)函数，与后续ViT编码器中所用一致。可分离卷积的输入为图像块$\mathbf{x}_i^m \in R^{p\times p\times d_m}$，$R^{p\times p\times d_m}$表示维度为p×p×$d_m$的三维实数矩阵，$d_m$为光谱带的数量，输出为图像块的特征图$\mathbf{f}_i^m \in R^{p\times p\times d}$，$R^{p\times p\times d}$表示维度为p×p×d的三维实数矩阵，d为每种模态映射后的维度。

[0087] 本发明中，可分离卷积块数量L_1取值为3，逐点卷积核数量d为64。

[0088] 第二步为格栅化，仿照原ViT模型中将输入图像划分为不重叠的图像块，本发明将特征图$\mathbf{f}_i^m$格栅化为p^2个向量，每个向量代表图像块中的一个像元，之后将这些向量正则化处理得到一组1D的标记嵌入$\mathbf{z}_i^m \in R^{p^2\times d}$，用以满足后期ViT模型的输入要求。

[0089] 第三步为线性层，通过为两种模态分别设置线性层，将从不同模态提取的标记嵌入$\mathbf{z}_i^m$投影到同一子空间。这种隐式特征对齐对于下一阶段更好地融合多模态特征是必要的。对齐后，输出新的标记嵌入序列$\mathbf{z}_i^m \in R^{p^2\times d_{new}}$。

[0090] 本发明中设置的线性映射后标记嵌入的维度为d_{new}＝64。

[0091] 第四步为增加位置编码，依照原ViT模型的规则，为了保证在自注意力编码过程中保留位置信息，需要对输入的标记嵌入增加可学习的位置编码。在上一步得到的标记嵌入序列$\mathbf{z}_i^m$上增加位置编码得到新的编码标记嵌入序列$\mathbf{z}_i^{m'} \in R^{p^2 \times d_{new}}$公式化如下：

[0092] $$\mathbf{z}_i^{m'} = \mathbf{z}_i^m + \mathbf{e} \quad (1)$$

[0093] 其中，$\mathbf{e} \in R^{p^2 \times d_{new}}$是可学习的位置编码，$R^{p^2 \times d_{new}}$表示维度为$p^2 \times d_{new}$的实数矩阵，$d_{new}=64$。

[0094] 此外，本发明认为按照原ViT模型的做法那样，增设分类标记不是必须的，可以通过融合所有基于像元的标记获得分类标记。因此本申请没有增设分类标记。

[0095] 第五步为丢弃(Dropout)层，为了避免分类网络的过拟合问题，像元i对应的图像块的编码标记嵌入序列$\mathbf{z}_i^{m'}$经过Dropout，得到新的有效标记嵌入序列$\mathbf{z}_i^{m''}$，公式化如下：

[0096] $$\mathbf{z}_i^{m''} = \mathrm{DP}\left(\mathbf{z}_i^{m'}\right) \quad (2)$$

[0097] 2)跨模态注意力融合层，跨模态注意力融合模块以两种模态图像块的标记嵌入序列沿模态维度串联(concatenation)作为输入，通过在原自注意力编码器上设置跨模态融合的层索引，同时实现模态内特征的进一步提取以及模态间信息的融合，输出融合后的标记嵌入序列。

[0098] 注意力融合模块实现了对特征的自注意力以及跨模态注意力融合。该编码器设计基于常规ViT编码器结构如附图4所示。在本身具备自注意力融合能力的基础上，通过设置编码器内进行跨模态融合的层索引以实现早、中、晚不同阶段的跨模态融合，如附图5所示。

[0099] 常规ViT编码器结构，如附图4所示，包括多头自注意力(MHSA)、层归一化(LN)和多层感知机(MLP)层，其核心是每个模态的多头自注意力，公式化如下：

[0100] $$\mathrm{MHSA}(z_i) = \mathrm{DP}([h_1, h_2, \cdots, h_A]w^0) \quad (3)$$

[0101] 其中，z_i为像元i对应图像块生成的标记嵌入序列，h_a表示第a个头，A为头的总数。对每一个头h_a，模态内空间相关信息可以通过传统的自注意力机制以矩阵内积形式获得：

[0102] $$\begin{aligned}\mathbf{h}_a &= \mathrm{Attention}\left(\mathbf{z}_i\mathbf{w}_a^Q, \mathbf{z}_i\mathbf{w}_a^K, \mathbf{z}_i\mathbf{w}_a^V\right) \\ &= \mathrm{SoftMax}\left(\frac{\left(\mathbf{z}_i\mathbf{w}_a^Q\right)\left(\mathbf{z}_i\mathbf{w}_a^K\right)^T}{\sqrt{d}}\right)\left(\mathbf{z}_i\mathbf{w}_a^V\right)\end{aligned} \quad (4)$$

[0103] 其中，$\left\{\mathbf{w}_a^Q, \mathbf{w}_a^K, \mathbf{w}_a^V\right\}$与$w^0$为线性层需要学习参数，前者用来将$z_i$映射到第a个头的子空间，后者用于聚合多个头的特征。从附图3中可以看出，多头注意力模块使用残差连接进行组装。

[0104] 为了进行多模态特征融合，将式(2)中得到的两种模态的标记嵌入沿模态为度进行合并，即$\tilde{\mathbf{z}}_i = \left[\mathbf{z}_i^1; \mathbf{z}_i^2\right]$，作为注意力融合模块的输入。在不打破ViT遵循的自注意力规则的情况下，可得到如下公式：

[0105] $$\tilde{\mathbf{z}}_i^{(l)} = \mathrm{MHSA}\left(\mathrm{LN}\left(\hat{\mathbf{z}}_i^{(l-1)}\right)\right) + \hat{\mathbf{z}}_i^{(l-1)}, l = 1, \cdots, L_2$$

[0106] $$\hat{\mathbf{z}}_i^{(l)}=\mathrm{MLP}\left(\mathrm{LN}\left(\tilde{\mathbf{z}}_i^{(l)}\right)\right)+\tilde{\mathbf{z}}_i^{(l)},l=1,\cdots,L_2 \quad (5)$$

[0107] 其中，MHSA()表示多头自注意力机制，MLP()表示多层感知机，LN()为层正则化，L_2为图像分类模型中编码器总层数，l为层索引，l_{cross}为晚期跨模态注意力融合的起始层索引，α＜0.5为计算l_{cross}的超参数，$\tilde{\mathbf{z}}_i^{(l)}$为跨模态注意力融合过程的像元i第l层标记嵌入序列的中间结果，$\hat{\mathbf{z}}_i^{(l)}$为像元i第l层标记嵌入序列，$\hat{\mathbf{z}}_i^{(L_2)}$为跨模态注意力融合层最终输出的像元i对应的标记嵌入序列。0≤l＜L_2为表示跨模态交叉注意开始的层深索引。由于标记数和特征维度在整个编码器模块中保持不变，本模块引入一个分割方案，分别令l＝0，$l=\lfloor \alpha L_2 \rfloor$以及$l=\lceil (1-\alpha)L_2 \rceil$代表早期、中期和晚期跨模态注意融合，如附图4所示，实现从低级特征到高级语义特征可调的跨模态信息交换。本发明在实验数据集中发现，在晚期使用跨模态融合多模态标记往往比在早期使用效果更好，即$l=\lceil (1-\alpha)L_2 \rceil$。这可以解释为，ViT编码器抽象的不同模态的高级特征通常表现出较清晰的语义，从而能产生更准确的融合结果。

[0108] 本发明图像分类模型的训练中编码器的总深度L_2设置为6(原ViT编码器层深的一半)；多头注意力中头的数量为4；丢弃层中丢弃率为0.1；在权衡分类性能和计算效率后，采用晚期跨模态融合。

[0109] 3)标记融合层，标记融合将注意力融合生成的标嵌入记序列作为输入，在不引入额外分类标记(classification token)的情况下，通过加权求和的方式将标记嵌入序列融合成用于分类的分类标记嵌入。

[0110] 本发明通过加权求和的方式融合多模态图像基于像元的标记嵌入以获得用于分类标记。首先，引入MLP Pre-Head，将$\hat{\mathbf{z}}_i^{(L_2)}$映射为一维向量；之后将该向量通过SoftMax操作，获得每个标记的权重。最后，使用获得的权重对原标记进行加权求和为每个标记获得更具有表现力的最终分类标记嵌入。之后，将分类标记输入到MLP层，得到像元对应的非归一化分类分数$\tilde{\mathbf{u}}_i$。公式化上述过程如下：

[0111] $$\tilde{\mathbf{u}}_i=\mathrm{MLP}\left(\mathrm{SoftMax}\left(\hat{\mathbf{z}}_i^{(L_2)}\mathbf{w}^{pre}\right)\times\hat{\mathbf{z}}_i^{(L_2)}\right) \quad (6)$$

[0112] 其中，MLP()表示多层感知机，$w^{pre}\in R^{d\times 1}$表示所述多层感知机前置头参数矩阵，$R^{d\times 1}$表示维度为d×1的实数矩阵，d等于$\hat{\mathbf{z}}_i^{(L_2)}$的维度，$\hat{\mathbf{z}}_i^{(L_2)}$为注意力融合层最终输出的第i个像元对应的标记嵌入序列，$\tilde{\mathbf{u}}_i$为第i个像元的分类分数。

[0113] 步骤三、训练ExViT网络模型：

[0114] 1)损失函数，本发明使用如下所示的基于多模态样本的多类别交叉熵损失来进行网络优化：

[0115] $$L\left(\mathbf{y},\tilde{\mathbf{u}}\middle|\mathbf{x}^1,\mathbf{x}^2\right)=\frac{1}{\left|\Omega_{traun}\right|}\sum_{i\in\Omega_{traun}}\mathbf{y}_i\log\left(\tilde{\mathbf{u}}_i\right) \quad (7)$$

[0116] 其中，$|\Omega_{traun}|$表示多模态训练集大小，$\tilde{\mathbf{u}}_i$为像元i的分类分数，y_i为像元i的类别标签。

[0117] 在某些情况下，由于不可避免的技术或传感环境限制，在现实中无法保证提供足

够的多模态数据时，为了解决该模态不完全问题，本发明引入多任务学习的思想，在输入模态数等于2时，通过最小化下述公式(8)损失函数进行求解：

[0118] $$L_{CML}=\beta L\left(\mathbf{y},\tilde{\mathbf{u}}\middle|\mathbf{x}^{1},\mathbf{x}^{2}\right)+\gamma L\left(\mathbf{y},\tilde{\mathbf{u}}\middle|\mathbf{x}^{m},\mathbf{0}\right) \quad (8)$$

[0119] 当输入模态数大于2时，所述图像分类模型训练时所需的最小损失函数为：

[0120] $$L_{CML}=\beta L\left(\mathbf{y},\tilde{\mathbf{u}}\middle|\mathbf{x}^{1},\mathbf{x}^{2},\cdots,\mathbf{x}^{K}\right)+\gamma L\left(\mathbf{y},\tilde{\mathbf{u}}\middle|\mathbf{x}^{m},\mathbf{0}\right) \quad (9)$$

[0121] 其中，$\beta,\gamma\geq 0$表示用于控制单模态学习任务相对于多模态学习任务的相对重要性的权衡参数，x^1、x^2、x^K分别表示训练所需的不同模态输入数据集，K为模态总数，$m\in 1,2,\cdots,K$的一个或多个，x^m为待预测模态输入数据集，$\tilde{\mathbf{u}}$为输入数据集的分类分数，y为输入数据集场景对应的标签。通过这种方式，提供了一种灵活且可解释的联合优化方案，旨在更好地利用跨模态对应关系。

[0122] 本发明中，分别设置$\beta=2$，$\gamma=0.1$，以获得最好的分类结果。

[0123] 2)优化器，自适应矩估计(Adaptive Moment Estimation，Adam)，每次选择一个小批量样本(mini-batch)、而非全部样本，进行模型参数更新。

[0124] 本发明方法ExViT的批处理(batch size)大小为64，优化器设为Adam，学习率初始为0.0005，每20步衰减0.9。同时，权重衰减分别参数化为0和0.0005。

[0125] 3)最优模型选择，在训练过程中，选择验证集上准确率最高的模型作为输出，如果在验证集上的准确率相同，则选择在验证集上损失最小的模型输出；每次迭代保存最好的模型，若迭代产生的模型更好，则替换之前保存的模型，否则不替换。

[0126] 步骤四、应用ExViT网络图像分类模型进行遥感图像分类：

[0127] 在测试阶段，将测试样本(即每个像元的输入数据集)输入最优模型，进行预测，将获得的未标准化分数$\tilde{\mathbf{u}}_i$进行SoftMax，最大值对应的索引即为当前样本像元对应的类别。

[0128] 从而，本发明提出的基于多模态注意力融合技术的遥感图像分类方法，为多模态遥感图像像元级分类任务构建简洁、通用的深度学习图像分类模型ExViT。该模型能够有效提取单模态数据的空间-通道信息，并且实现对异构模态特征从浅到深的高效融合。从而本发明不仅能够应用于多种模态数据，而且具有跨模态挖掘数据之间相关及互补信息的能力，可以获得更高分类精度。此外，本发明通过在ExViT框架中建立的跨模态多任务学习机制，解决了真实遥感场景中的模态不完整问题，不仅易于解释，而且更易于实施。

[0129] 图6是本发明实施例提供的基于多模态注意力融合技术的遥感图像分类装置的结构示意图。如图6所示，所述装置包括：

[0130] 构成模块610，用于确定待分类多模态遥感图像中每个像元的邻域图像块，并将待分类多模态遥感图像中每个像元及其对应的邻域图像块内的像元构成该像元的多种不同模态的输入数据集，其中待分类多模态遥感图像包括以下所述的任意一种或多种模态遥感图像：高光谱图像、激光雷达点云数据、多光谱图像以及合成孔径雷达图像；

[0131] 输出模块620，用于将每个像元的多种不同模态的输入数据集输入至预先构建的基于多模态遥感图像注意力融合的图像分类模型，输出每个像元的分类分数；

[0132] 第一确定模块630，用于将每个像元的分类分数进行SoftMax运算分别得到分数最大值，根据每个像元得到的分数最大值确定每个像元的类别；

[0133] 第二确定模块640,用于根据每个像元的类别,确定待分类多模态遥感图像的分类图像。

[0134] 可选地,图像分类模型的网络模型结构包括:

[0135] 输入层,包括多个输入模块,用于输入每个像元的不同模态输入数据集;

[0136] 多模态特征提取与标记层,利用深度可分离卷积将输入数据集处理成下游跨模态注意力融合层所需的带位置信息的标记嵌入数据;

[0137] 跨模态注意力融合层,用于将每个像元不同模态的标记嵌入数据进行信息融合,输出融合后的标记嵌入序列;

[0138] 标记融合层,通过多层感知机前置头将标记嵌入序列加权求和融合成用于分类的分类标记嵌入,并通过多层感知机头根据每个像元的分类标记嵌入,输出每个像元的分类分数。

[0139] 可选地,跨模态注意力融合层采用晚期跨模态注意力融合,其中晚期跨模态注意力融合规则的起始层索引中在图像分类模型中编码器总层数中通过超参数设置索引。

[0140] 可选地,根据输入模态数确定图像分类模型训练时所使用的最小损失函数中数据集数量。

[0141] 可选地,所述装置还包括:

[0142] 训练模块,用于使用Adam优化器,设置学习初始率为0.0005,每20步衰减0.9倍迭代训练最小损失函数,以获得图像分类模型的最优模型。

[0143] 图7是本发明实施例提供的电子设备的结构。如图7所示,电子设备70包括一个或多个处理器71和存储器72。

[0144] 处理器71可以是中央处理单元(CPU)或者具有数据处理能力和/或指令执行能力的其他形式的处理单元,并且可以控制电子设备中的其他组件以执行期望的功能。

[0145] 存储器72可以包括一个或多个计算机程序产品,所述计算机程序产品可以包括各种形式的计算机可读存储介质,例如易失性存储器和/或非易失性存储器。所述易失性存储器例如可以包括随机存取存储器(RAM)和/或高速缓冲存储器(cache)等。所述非易失性存储器例如可以包括只读存储器(ROM)、硬盘、闪存等。在所述计算机可读存储介质上可以存储一个或多个计算机程序指令,处理器71可以运行所述程序指令,以实现上文所述的本发明的各个实施例的软件程序的方法以及/或者其他期望的功能。在一个示例中,电子设备还可以包括:输入装置73和输出装置74,这些组件通过总线系统和/或其他形式的连接机构(未示出)互连。

[0146] 此外,该输入装置73还可以包括例如键盘、鼠标等等。

[0147] 该输出装置74可以向外部输出各种信息。该输出装置74可以包括例如显示器、扬声器、打印机、以及通信网络及其所连接的远程输出设备等等。

[0148] 当然,为了简化,图7中仅示出了该电子设备中与本发明有关的组件中的一部分,省略了诸如总线、输入/输出接口等的组件。除此之外,根据具体应用情况,电子设备还可以包括任何其他适当的组件。

[0149] 本领域内的技术人员应明白,本发明的实施例可提供为方法、系统、或计算机程序产品。因此,本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机

可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0150] 因此,本申请还提出一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现如本申请中任一实施例所述的方法。

[0151] 进一步地,本申请还提出一种电子设备,包括存储器,处理器及存储在存储器上并可在处理器运行的计算机程序,所述处理器执行所述计算机程序时实现如本申请任一实施例所述的方法。

[0152] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0153] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0154] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0155] 在一个典型的配置中,计算设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。内存可能包括计算机可读介质中的非永久性存储器,随机存取存储器(RAM)和/或非易失性内存等形式,如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。

[0156] 本发明实施例可以应用于终端设备、计算机系统、服务器等电子设备,其可与众多其它通用或专用计算系统环境或配置一起操作。适于与终端设备、计算机系统、服务器等电子设备一起使用的众所周知的终端设备、计算系统、环境和/或配置的例子包括但不限于:个人计算机系统、服务器计算机系统、瘦客户机、厚客户机、手持或膝上设备、基于微处理器的系统、机顶盒、可编程消费电子产品、网络个人电脑、小型计算机系统、大型计算机系统和包括上述任何系统的分布式云计算技术环境,等等。

[0157] 终端设备、计算机系统、服务器等电子设备可以在由计算机系统执行的计算机系统可执行指令(诸如程序模块)的一般语境下描述。通常,程序模块可以包括例程、程序、目标程序、组件、逻辑、数据结构等等,它们执行特定的任务或者实现特定的抽象数据类型。计算机系统/服务器可以在分布式云计算环境中实施,分布式云计算环境中,任务是由通过通信网络链接的远程处理设备执行的。在分布式云计算环境中,程序模块可以位于包括存储设备的本地或远程计算系统存储介质上。

[0158] 还需要说明的是,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素,而且还包

括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

[0159] 以上所述仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

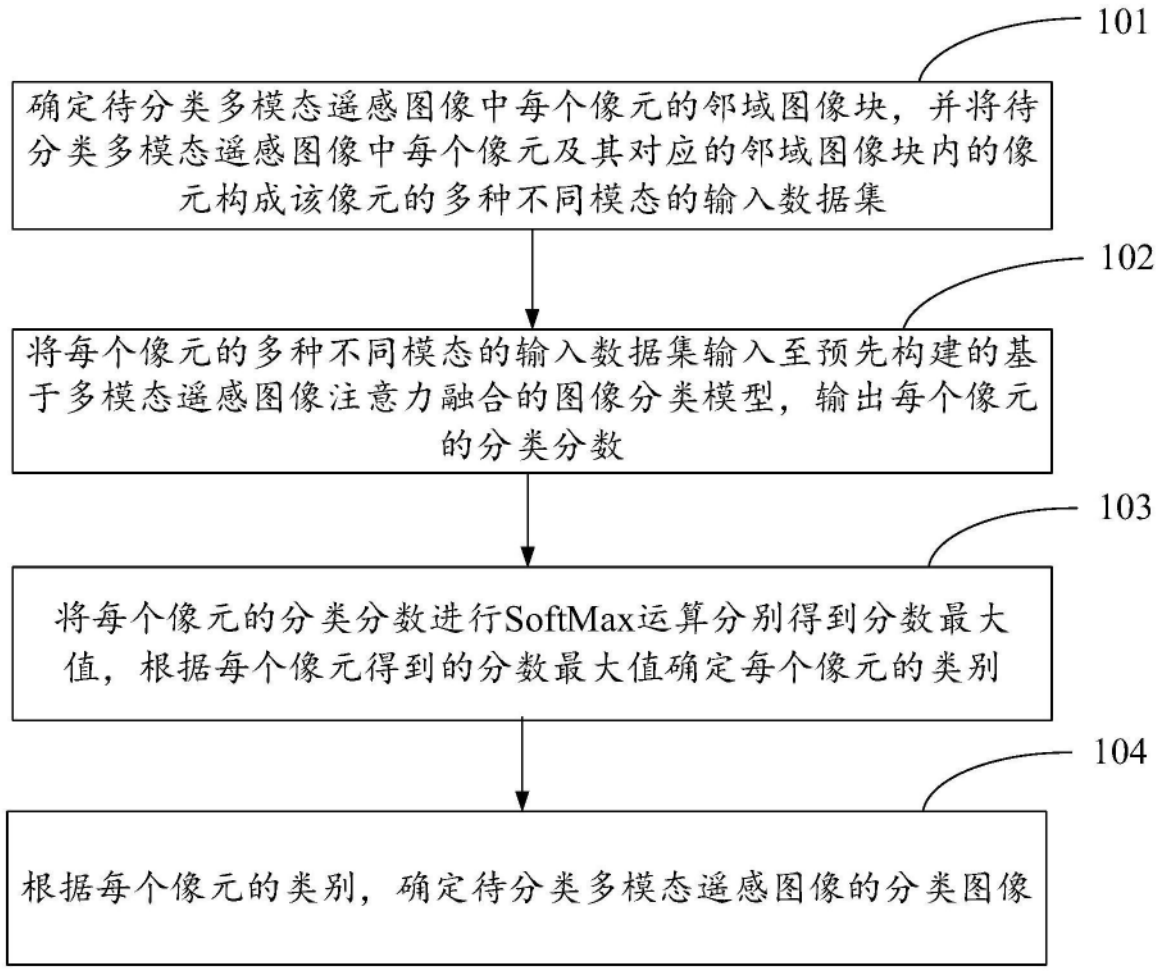

101
确定待分类多模态遥感图像中每个像元的邻域图像块，并将待分类多模态遥感图像中每个像元及其对应的邻域图像块内的像元构成该像元的多种不同模态的输入数据集
102
将每个像元的多种不同模态的输入数据集输入至预先构建的基于多模态遥感图像注意力融合的图像分类模型，输出每个像元的分类分数
103
将每个像元的分类分数进行SoftMax运算分别得到分数最大值，根据每个像元得到的分数最大值确定每个像元的类别
104
根据每个像元的类别，确定待分类多模态遥感图像的分类图像

图1

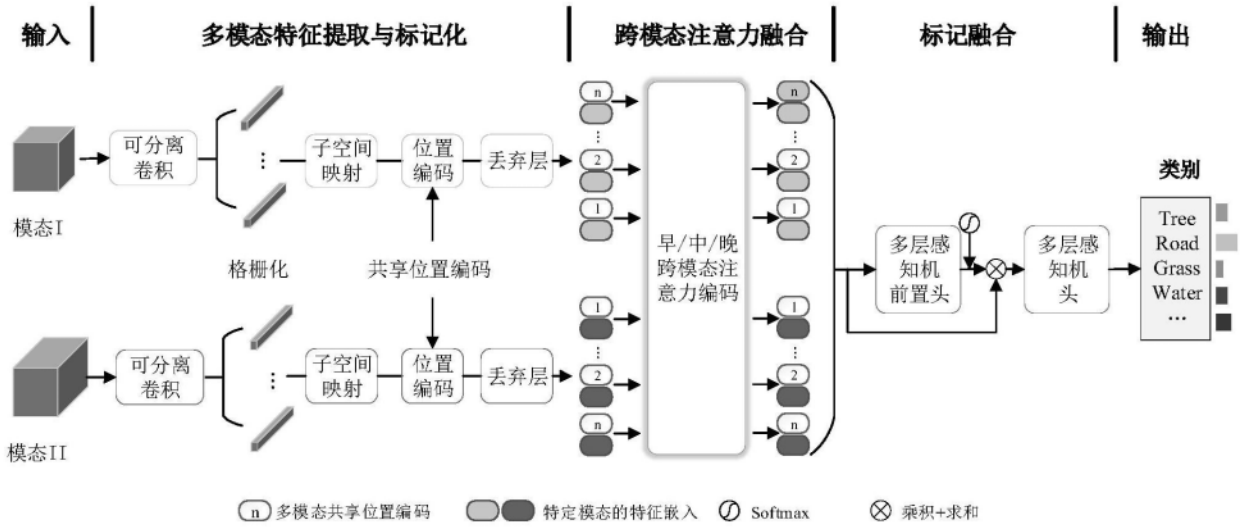

输入
多模态特征提取与标记化
跨模态注意力融合
标记融合
输出
模态I
可分离卷积
子空间映射
位置编码
丢弃层
格栅化
共享位置编码
模态II
可分离卷积
子空间映射
位置编码
丢弃层
早/中/晚跨模态注意力编码
多层感知机前置头
多层感知机头
类别
Tree
Road
Grass
Water
...
多模态共享位置编码
特定模态的特征嵌入
Softmax
乘积+求和

图2

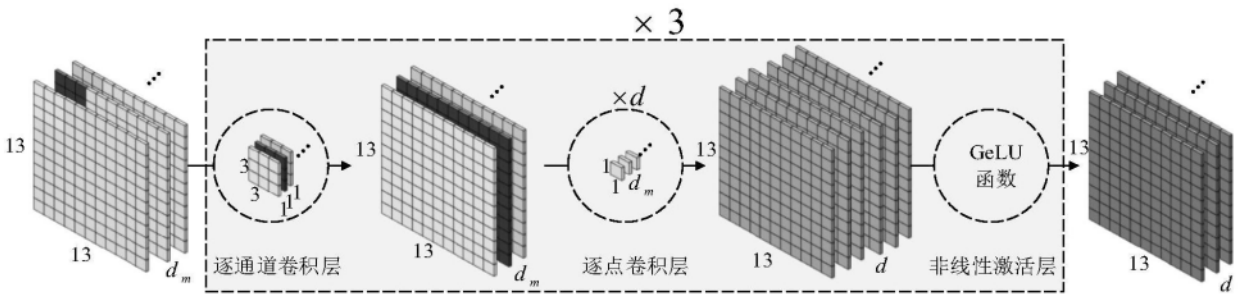

× 3
13
逐通道卷积层
逐点卷积层
非线性激活层
GeLU
函数

图3

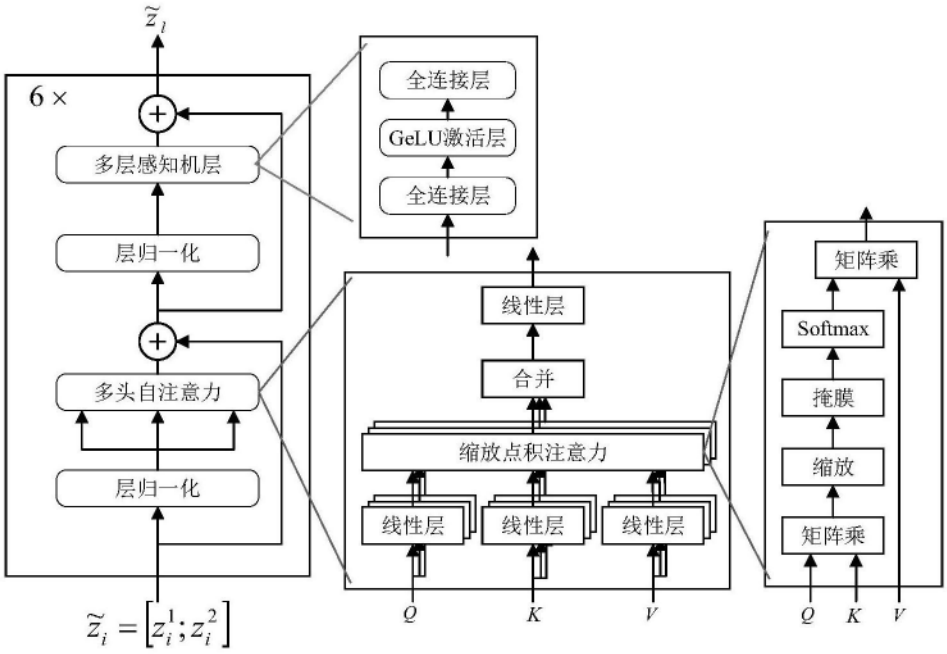

6 ×
多层感知机层
层归一化
多头自注意力
层归一化
全连接层
GeLU激活层
全连接层
线性层
合并
缩放点积注意力
线性层
线性层
线性层
矩阵乘
Softmax
掩膜
缩放
矩阵乘
Q
K
V

图4

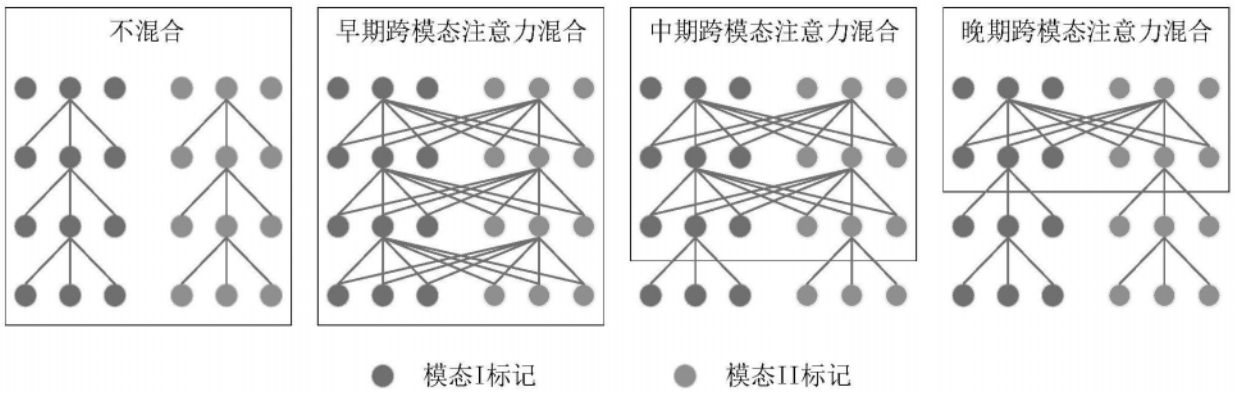

不混合
早期跨模态注意力混合
中期跨模态注意力混合
晚期跨模态注意力混合
模态I标记
模态II标记

图5

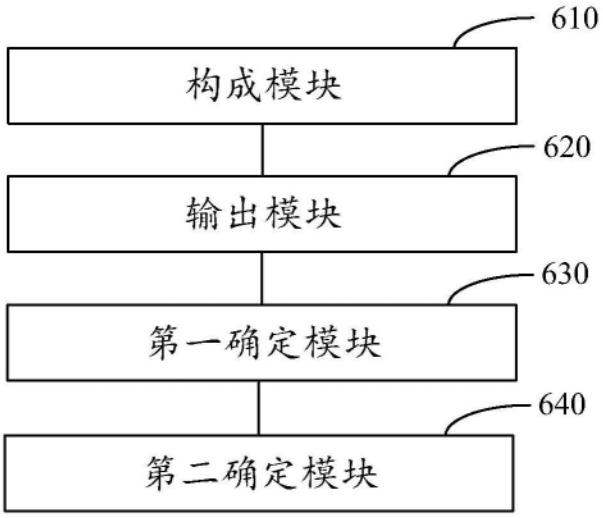

610
构成模块
620
输出模块
630
第一确定模块
640
第二确定模块

图6

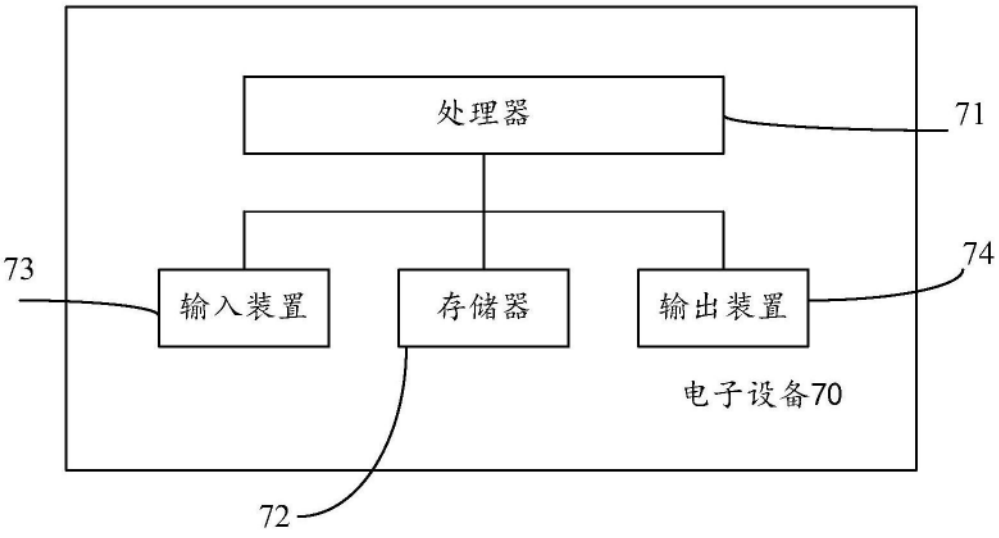

处理器
71
73
输入装置
存储器
输出装置
74
电子设备70
72

图7