(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113298151 A
(43)申请公布日 2021.08.24

(21)申请号 202110577114.2

(22)申请日 2021.05.26

(71)申请人 中国电子科技集团公司第五十四研究所
地址 050081 河北省石家庄市中山西路589号第五十四所航天实验室

(72)发明人 王港 高峰 陈金勇 帅通 王敏 郭争强

(74)专利代理机构 河北东尚律师事务所 13124
代理人 王文庆

(51)Int.Cl.
G06K 9/62(2006.01)
G06N 3/04(2006.01)

权利要求书2页 说明书5页 附图4页

(54)发明名称

一种基于多级特征融合的遥感图像语义描述方法

(57)摘要

本发明提供一种基于多级特征融合的遥感图像语义描述方法，属于遥感图像处理和计算机视觉领域，包括以下步骤：获取高分辨率遥感影像，构建遥感图像语义描述数据集；利用语义标注数据集训练图像的语义分类模型，由图像提取单词描述并进行编码，得到语义特征；利用目标检测数据集训练目标检测模型，提取图像的区域级特征并进行编码，得到视觉特征；将获取的语义和视觉特征进行聚合，即将两组特征拼接在一起；将聚合后的多级特征作为Transformer的输入，训练图像自然语言生成模型。本发明利用了图像的语义和视觉特征，提取的信息包含了场景信息、区域视觉信息和对象的语义关系，生成的图像语义描述可读性强，准确性高。

CN 113298151 A

1.一种基于多级特征融合的遥感图像语义描述方法,其特征在于,所述方法包括以下步骤:

步骤一、获取原始的高分辨率遥感影像,对获取的高分辨率遥感影像进行预处理,得到图像数据集,对于图像数据集中的每个图像,人工添加语义标注,用自然语言的形式描述图像内容,图像与语义标注共同构成遥感图像语义描述数据集;同时获取公开的遥感图像目标检测数据集;

步骤二、将构建的遥感图像语义描述数据集和公开的遥感图像目标检测数据集分别划分为训练集、验证集和测试集;

步骤三、将遥感图像语义描述数据集划分后各个数据集中图像对应的语义标注拆分为单个词,每个单词作为图像的一个标签,利用拆分后的训练集和验证集数据对语义分类模型进行训练及优化调整,利用拆分后的测试集数据对语义分类模型进行验证,获取图像的语义特征;同时利用遥感图像目标检测数据集划分后的训练集和验证集数据对目标检测模型进行训练及优化调整,利用测试集数据对目标检测模型进行验证,获取图像的视觉特征;

步骤四、将步骤三获取的语义特征和视觉特征进行聚合,即将两组特征拼接在一起,得到图像多级特征;

步骤五、将步骤四得到的图像多级特征作为图像自然语言生成模型的输入,训练图像自然语言生成模型;

步骤六、利用步骤二处理后的测试集数据对训练好的图像自然语言生成模型进行验证,生成遥感图像的语义描述。

2.如权利要求1所述的一种基于多级特征融合的遥感图像语义描述方法,其特征在于,步骤三中利用拆分后的训练集和验证集数据对语义分类模型进行训练及优化调整,利用拆分后的测试集数据对语义分类模型进行验证,获取图像的语义特征,具体包括以下过程:

步骤3.1:设置模型的参数:设置ResNet-101语义分类模型的网络学习率、优化函数、最大迭代次数、批量训练的大小以及模型保存位置;

步骤3.2:训练模型:利用拆分后的训练集数据对ResNet-101语义分类模型的参数进行训练;

步骤3.3:优化模型:根据拆分后的验证集数据调整ResNet-101语义分类模型的参数,优化ResNet-101语义分类模型;

步骤3.4:验证模型:根据拆分后的测试集数据对ResNet-101语义分类模型进行验证;

步骤3.5:获取语义特征:在ResNet-101语义分类模型得到的输出中,根据每个单词的概率值大小排序,选择前K个得分高的单词,将每个单词进行编码,得到语义特征向量;其中,K为设定值。

3.如权利要求1所述的一种基于多级特征融合的遥感图像语义描述方法,其特征在于,步骤三中利用遥感图像目标检测数据集划分后的训练集和验证集数据对目标检测模型进行训练及优化调整,利用测试集数据对目标检测模型进行验证,获取图像的视觉特征,具体包括以下步骤:

步骤4.2:设置目标检测模型的参数:设置Faster R-CNN目标检测模型的先验框大小、网络学习率、优化函数、最大迭代次数、批量训练的大小和模型保存位置;

步骤4.3:训练模型:利用遥感图像目标检测数据集的训练集数据对Faster R-CNN目标

检测模型的参数进行训练；

步骤4.4:优化模型:根据遥感图像目标检测数据集的验证集数据调整Faster R-CNN目标检测模型的训练参数,优化Faster R-CNN目标检测模型；

步骤4.5:验证模型:根据遥感图像目标检测数据集的测试集数据对Faster R-CNN目标检测模型进行验证；

步骤4.6:获取视觉特征:在Faster R-CNN目标检测模型生成的候选区域中,根据区域是待检测目标的概率值排序结果,选择前P个得分高的候选区域,对候选区域提取特征,将P个候选区域特征聚合在一起,得到视觉特征向量;其中P为设定值。

4.如权利要求1所述的一种基于多级特征融合的遥感图像语义描述方法,其特征在于,步骤五具体包括以下步骤：

步骤5.1:将步骤四获取的图像多级特征作为输入特征向量,将输入特征向量分割为多个片段,按照一定的顺序排列,得到序列化数据,并为每个片段添加一个位置向量,以确定每一个描述单词的位置；

步骤5.2:将每个片段的输入特征向量通过注意力机制后与输入特征向量本身进行相加和归一化；

步骤5.3:相加和归一化后的特征向量经前馈神经网络进行特征提取与组织,将前馈神经网络前后的数据再次进行相加和归一化；

步骤5.4:步骤5.3相加和归一化获得的结果,输出到上一个位置片段的注意力机制上,以持续获得不断片段的排序位置和语言信息；

步骤5.5:将上一个位置片段的注意力机制前后的数据进行相加和归一化,并依次经前馈神经网络和softmax层,得到一个输出向量,输出向量的每个位置代表相应单词的得分,选择概率最大的单词即当前时刻的输出结果；

步骤5.6:重复步骤5.2至步骤5.5,直到生成一个约定的终止符号,表示图像自然语言生成模型完成了输出,将每次得到的单词连接在一起即为对应遥感图像的语义描述。

一种基于多级特征融合的遥感图像语义描述方法

技术领域

[0001] 本发明属于遥感图像处理和计算机视觉领域，具体涉及一种基于图像视觉、语义特征融合和注意力机制的遥感图像自然语义描述方法。

背景技术

[0002] 随着传感器技术的迅速发展，人类对地球的观测能力越来越高，获取的数据量显著增加。但是，信息处理水平严重滞后于遥感数据获取技术的发展，使得海量的数据得不到有效的利用。研究和探索对数据量巨大的遥感图像进行快速准确的理解，提取有用的信息，进而指导在农业、环境、交通、军事等领域的科学决策显得十分重要。

[0003] 遥感图像语义描述是从图像中提取信息，感知图像所蕴含的场景语义，并对图像中的内容进行描述的过程，是对遥感图像高层次的解析。在遥感场景理解领域，让计算机按照人类认知理解图像一样来认知图像，从遥感图像中自动提取信息，生成容易理解的文本描述受到了广泛研究。

[0004] 图像描述的研究方法主要分为以下三个类别：基于模板、基于检索和基于深度学习的图像描述。基于模板的图像描述是基于固定的硬编码语句模板方法，根据图像中识别到的对象以及发现的对象关系来匹配句子模板，从而生成图像描述。基于检索的方法把训练集含有的图像和其对应的文本描述映射到同一向量空间，并计算两者之间的距离，然后根据距离排名得到和训练集中图像内容最接近的文本描述。上述两类方法限制了描述文本的多样性，不能生成可变长度、灵活性强的描述语句。

[0005] 近年来卷积神经网络在图像上的应用，对于提取图片特征信息表现出的强大能力，以及循环神经网络在机器翻译领域发挥出的卓越效果，推动了神经网络在图像描述领域的发展。基于神经网络的图像描述，不依赖于任何模板、语法树或者有限的类别库，不需要制定任何的规则，它们自动地从海量的训练集中去学习图像和文本的信息，能够记忆各种各样的图像信息和其对应文本的对应关系，然后自动推断出测试图像和其相对应的文本，能够生成更灵活、更新颖的文本描述，而且还能够很好地描述从未见过的图像。

发明内容

[0006] 针对现有技术，本发明提供一种基于多级特征融合的遥感图像语义描述方法，通过深度卷积神经网络训练分类和目标检测模型，在训练好的分类模型中，获取描述图像的多个单词，经过编码得到语义特征，在训练好的检测模型中，获取目标候选区域，得到视觉特征，将语义和视觉特征融合，共同作为图像自然语言生成模型(Transformer解码器)的输入，生成遥感图像的自然语言描述语句。

[0007] 为了实现遥感图像的自然语言描述，本发明提出基于多级特征融合的遥感图像语义描述方法，采用的技术方案为：

[0008] 一种基于多级特征融合的遥感图像语义描述方法，所述方法包括以下步骤：

[0009] 步骤一、获取原始的高分辨率遥感影像，对获取的高分辨率遥感影像进行预处理，

得到图像数据集，对于图像数据集中的每个图像，人工添加语义标注，用自然语言的形式描述图像内容，图像与语义标注共同构成遥感图像语义描述数据集；同时获取公开的遥感图像目标检测数据集；

[0010] 步骤二、将构建的遥感图像语义描述数据集和公开的遥感图像目标检测数据集分别划分为训练集、验证集和测试集；

[0011] 步骤三、将遥感图像语义描述数据集划分后各个数据集中图像对应的语义标注拆分为单个词，每个单词作为图像的一个标签，利用拆分后的训练集和验证集数据对语义分类模型进行训练及优化调整，利用拆分后的测试集数据对语义分类模型进行验证，获取图像的语义特征；同时利用遥感图像目标检测数据集划分后的训练集和验证集数据对目标检测模型进行训练及优化调整，利用测试集数据对目标检测模型进行验证，获取图像的视觉特征；

[0012] 步骤四、将步骤三获取的语义特征和视觉特征进行聚合，即将两组特征拼接在一起，得到图像多级特征；

[0013] 步骤五、将步骤四得到的图像多级特征作为图像自然语言生成模型的输入，训练图像自然语言生成模型；

[0014] 步骤六、利用步骤二处理后的测试集数据对训练好的图像自然语言生成模型进行验证，生成遥感图像的语义描述。

[0015] 进一步的，步骤三中利用拆分后的训练集和验证集数据对语义分类模型进行训练及优化调整，利用拆分后的测试集数据对语义分类模型进行验证，获取图像的语义特征，具体包括以下过程：

[0016] 步骤3.1：设置模型的参数：设置ResNet-101语义分类模型的网络学习率、优化函数、最大迭代次数、批量训练的大小以及模型保存位置；

[0017] 步骤3.2：训练模型：利用拆分后的训练集数据对ResNet-101语义分类模型的参数进行训练；

[0018] 步骤3.3：优化模型：根据拆分后的验证集数据调整ResNet-101语义分类模型的参数，优化ResNet-101语义分类模型；

[0019] 步骤3.4：验证模型：根据拆分后的测试集数据对ResNet-101语义分类模型进行验证；

[0020] 步骤3.5：获取语义特征：在ResNet-101语义分类模型得到的输出中，根据每个单词的概率值大小排序，选择前K个得分高的单词，将每个单词进行编码，得到语义特征向量；其中，K为设定值。

[0021] 进一步的，步骤三中利用遥感图像目标检测数据集划分后的训练集和验证集数据对目标检测模型进行训练及优化调整，利用测试集数据对目标检测模型进行验证，获取图像的视觉特征，具体包括以下步骤：

[0022] 步骤4.2：设置目标检测模型的参数：设置Faster R-CNN目标检测模型的先验框大小、网络学习率、优化函数、最大迭代次数、批量训练的大小和模型保存位置；

[0023] 步骤4.3：训练模型：利用遥感图像目标检测数据集的训练集数据对Faster R-CNN目标检测模型的参数进行训练；

[0024] 步骤4.4：优化模型：根据遥感图像目标检测数据集的验证集数据调整Faster R-

CNN目标检测模型的训练参数,优化Faster R-CNN目标检测模型;

[0025] 步骤4.5:验证模型:根据遥感图像目标检测数据集的测试集数据对Faster R-CNN目标检测模型进行验证;

[0026] 步骤4.6:获取视觉特征:在Faster R-CNN目标检测模型生成的候选区域中,根据区域是待检测目标的概率值排序结果,选择前P个得分高的候选区域,对候选区域提取特征,将P个候选区域特征聚合在一起,得到视觉特征向量;其中P为设定值。

[0027] 进一步的,步骤五具体包括以下步骤:

[0028] 步骤5.1:将步骤四获取的图像多级特征作为输入特征向量,将输入特征向量分割为多个片段,按照一定的顺序排列,得到序列化数据,并为每个片段添加一个位置向量,以确定每一个描述单词的位置;

[0029] 步骤5.2:将每个片段的输入特征向量通过注意力机制后与输入特征向量本身进行相加和归一化;

[0030] 步骤5.3:相加和归一化后的特征向量经前馈神经网络进行特征提取与组织,将前馈神经网络前后的数据再次进行相加和归一化;

[0031] 步骤5.4:步骤5.3相加和归一化获得的结果,输出到上一个位置片段的注意力机制上,以持续获得不断片段的排序位置和语言信息;

[0032] 步骤5.5:将上一个位置片段的注意力机制前后的数据进行相加和归一化,并依次经前馈神经网络和softmax层,得到一个输出向量,输出向量的每个位置代表相应单词的得分,选择概率最大的单词即当前时刻的输出结果;

[0033] 步骤5.6:重复步骤5.2至步骤5.5,直到生成一个约定的终止符号,表示图像自然语言生成模型完成了输出,将每次得到的单词连接在一起即为对应遥感图像的语义描述。

[0034] 与现有技术相比,本发明的优点和有益效果:

[0035] (1)本发明利用分类网络提取了图像的多标签信息,由每个语句的单词构成,包含丰富的语义信息,有利于模型生成描述目标之间关系的语句。

[0036] (2)本发明利用目标检测网络生成候选区域并提取其特征,符合人类理解图像的特点,即描述显著目标之间的语义关系。

[0037] (3)本发明语义特征和视觉特征的融合包含了场景信息、区域视觉信息和对象的语义关系,有助于提升生成自然语言描述的可读性和准确性。

[0038] (4)本发明Transformer解码器全部由注意力机制组成,可将任意位置的两个单词的距离转换成1,有助于解决语句的长期依赖问题,生成更加可靠的自然语言描述语句。

附图说明

[0039] 图1是本发明提供的基于多级特征融合的遥感图像语义描述方法流程图。

[0040] 图2是本发明利用神经网络分类器提取语义单词并进行特征编码的示意图。

[0041] 图3是本发明利用目标检测网络提取候选区域并进行视觉特征编码的示意图。

[0042] 图4是本发明利用Transformer解码器生成图像语义描述的示意图。

[0043] 图5是本发明实施过程中基于多级特征融合的图像语义描述模型生成的实际自然语言描述结果示例。

具体实施方式

[0044] 下面结合附图和具体实例对本发明作进一步解释说明。

[0045] 如图1所示，一种基于多级特征融合的遥感图像语义描述方法，包括以下步骤：

[0046] 步骤一、构建遥感图像语义描述数据集，步骤如下：获取原始的高分辨率遥感影像；对上述获取的高分辨率遥感影像进行预处理，包括图像去噪和裁剪，本实施例得到尺寸大小在300-1000之间的图像数据集；对于每个图像，人工添加语义描述，用自然语言的形式描述图像内容，每张图像由T个语句描述，图像与语义标注共同构成完整的遥感图像语义描述数据集；同时下载公开的遥感图像目标检测数据集DOTA，其包含有16个类别中的40万个带目标标注框的对象实例；

[0047] 步骤二、数据集划分：将构建的遥感图像语义描述数据集和公开的遥感图像目标检测数据集DOTA分别按照8:1:1的比例划分为训练集、验证集和测试集；

[0048] 步骤三、利用Resnet-101语义分类模型(神经网络分类器)获取图像的语义特征，如图2所示，步骤如下：

[0049] 步骤3.1：构建训练分类模型需要的数据集：将图像对应的语义标注拆分为单个词，每个单词作为图像的一个标签，共同组成样本的多标签，在训练过程中，图像作为输入，所有单词构成的多标签作为输出，其中输出向量的维度为T，即整个语义标注数据集所包含的无重复的单词数目；

[0050] 步骤3.2：设置模型的参数：设置ResNet-101语义分类模型的网络学习率、优化函数、最大迭代次数、批量训练的大小以及模型保存位置；

[0051] 步骤3.3：训练模型：利用拆分后的训练集数据对ResNet-101语义分类模型的参数进行训练；

[0052] 步骤3.4：优化模型：根据拆分后的验证集数据调整ResNet-101语义分类模型的参数，优化ResNet-101语义分类模型；

[0053] 步骤3.5：验证模型：根据拆分后的测试集数据对ResNet-101语义分类模型进行验证；

[0054] 步骤3.6：获取语义特征：在ResNet-101语义分类模型得到的输出中，根据每个单词的概率值大小排序，选择前K个得分高的单词，将每个单词进行编码，得到N1维语义特征向量；其中，K为设定值。

[0055] 利用Faster R-CNN目标检测网络(候选区域提取网络)获取图像的视觉特征，如图3所示，步骤如下：

[0056] 步骤4.2：设置目标检测模型的参数：设置Faster R-CNN目标检测模型的先验框大小、网络学习率、优化函数、最大迭代次数、批量训练的大小和模型保存位置；

[0057] 步骤4.3：训练模型：利用遥感图像目标检测数据集的训练集数据对Faster R-CNN目标检测模型的参数进行训练；

[0058] 步骤4.4：优化模型：根据遥感图像目标检测数据集的验证集数据调整Faster R-CNN目标检测模型的训练参数，优化Faster R-CNN目标检测模型；

[0059] 步骤4.5：验证模型：根据遥感图像目标检测数据集的测试集数据对Faster R-CNN目标检测模型进行验证；

[0060] 步骤4.6：获取视觉特征：在Faster R-CNN目标检测模型生成的候选区域中，根据

区域是待检测目标的概率值排序结果，选择前P个得分高的候选区域，对候选区域提取特征，将P个候选区域特征聚合在一起，得到N2维视觉特征向量；其中P为设定值。

[0061] 步骤四、多级特征融合，将步骤三获取的语义和视觉特征进行聚合，即将两组特征拼接在一起，得到N(N＝N1+N2)维特征；

[0062] 步骤五、将步骤四得到的N维图像多级特征作为图像自然语言生成模型(Transformer解码器)的输入，输出为图像的自然语义描述，如图4所示，步骤如下：

[0063] 步骤5.1：将步骤四获取的图像多级特征作为输入特征向量，将输入特征向量分割为多个片段，按照一定的顺序排列，得到序列化数据，并为每个片段添加一个位置向量，以确定每一个描述单词的位置；

[0064] 步骤5.2：将每个片段的输入特征向量通过注意力机制后与输入特征向量本身进行相加和归一化；

[0065] 步骤5.3：相加和归一化后的特征向量经前馈神经网络进行特征提取与组织，将前馈神经网络前后的数据再次进行相加和归一化；

[0066] 步骤5.4：步骤5.3相加和归一化获得的结果，输出到上一个位置片段的注意力机制上，以持续获得不断片段的排序位置和语言信息；

[0067] 步骤5.5：将上一个位置片段的注意力机制前后的数据进行相加和归一化，并依次经前馈神经网络和softmax层，得到一个输出向量，输出向量的每个位置代表相应单词的得分，选择概率最大的单词即当前时刻的输出结果；

[0068] 步骤5.6：重复步骤5.2至步骤5.5，直到生成一个约定的终止符号，表示Transformer的解码器已经完成了输出，将每次得到的单词连接在一起即为对应遥感图像的语义描述。

[0069] 步骤六、模型验证及应用：利用步骤二处理后的测试集数据对训练好的图像自然语言生成模型进行验证，生成遥感图像的语义描述。

[0070] 如图5所示，是本发明实施过程中基于多级特征融合的图像语义描述模型生成的实际自然语言描述结果示例。

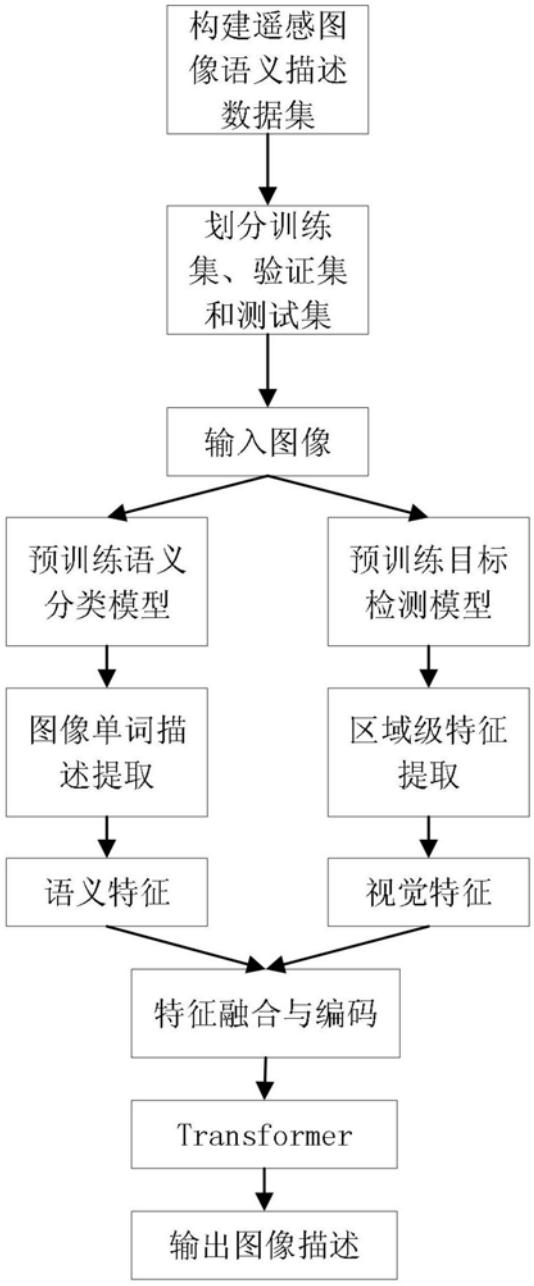
构建遥感图像语义描述数据集
划分训练集、验证集和测试集
输入图像
预训练语义分类模型
预训练目标检测模型
图像单词描述提取
区域级特征提取
语义特征
视觉特征
特征融合与编码
Transformer
输出图像描述

图1

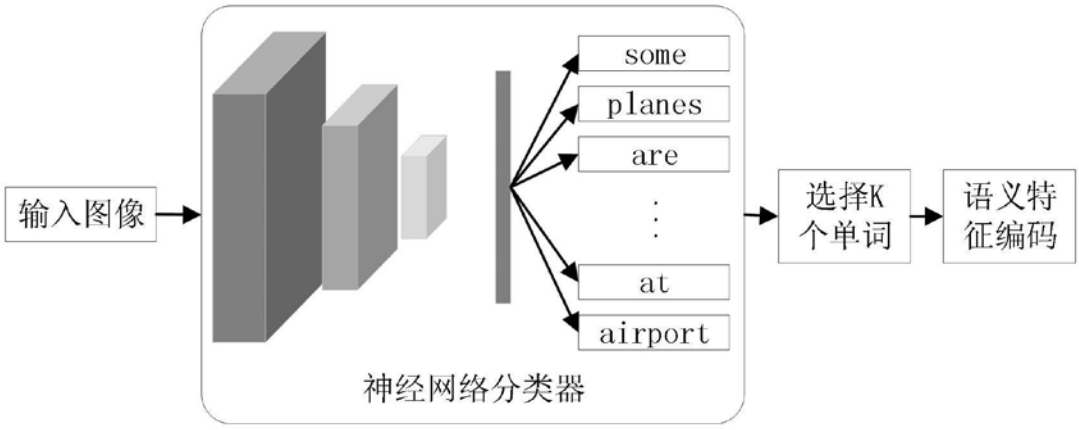
输入图像
some
planes
are
at
airport
神经网络分类器
选择K个单词
语义特征编码

图2

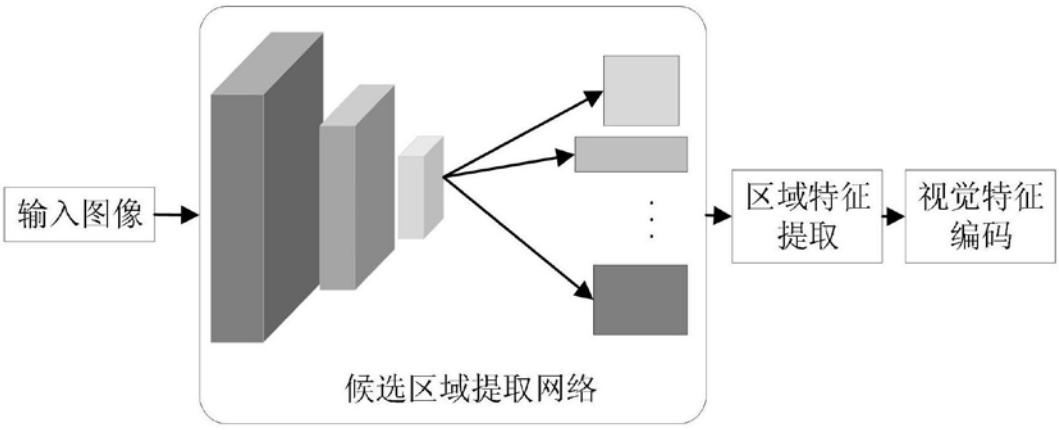
输入图像
候选区域提取网络
区域特征提取
视觉特征编码

图3

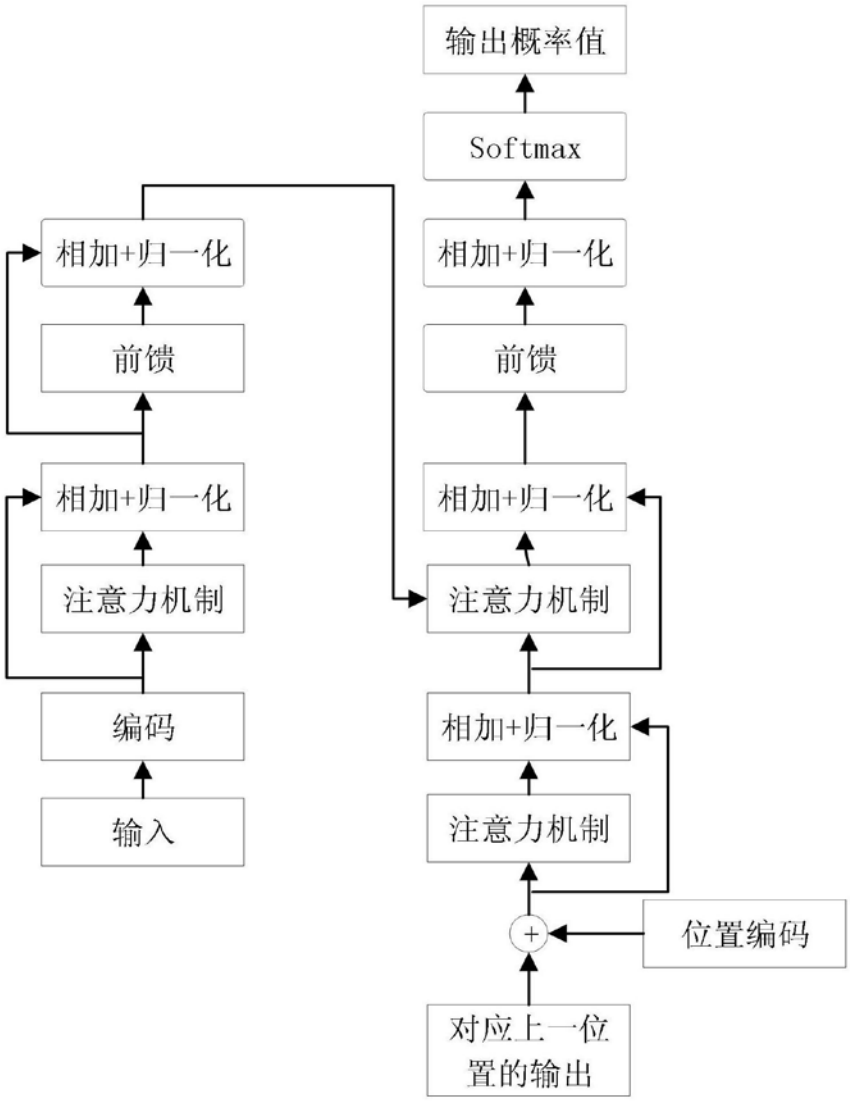
输出概率值
Softmax
相加+归一化
前馈
相加+归一化
注意力机制
相加+归一化
编码
输入
相加+归一化
前馈
相加+归一化
注意力机制
相加+归一化
注意力机制
位置编码
对应上一位置的输出

图4

人工标注：some planes are in an airport near some buildings.

一些飞机停靠在一个机场的一些建筑附近。

算法生成：many planes are parked in an airport.

很多飞机停靠在一个机场上。

人工标注：many buildings and some green trees are in an industrial area.

在一个工业区内有很多建筑和一些绿树。

算法生成：many buildings and some green trees are in a dense residential area.

在密集的居民区内有很多建筑和一些绿树。

图5