(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 110390340 A
(43)申请公布日 2019.10.29

(21)申请号 201910650283.7

(22)申请日 2019.07.18

(71)申请人 暗物智能科技(广州)有限公司
地址 511458 广东省广州市南沙区丰泽东路106号(自编1号楼)X1301-G5994(集群注册)(JM)

(72)发明人 朱艺 梁小丹 林倞

(74)专利代理机构 北京三聚阳光知识产权代理有限公司 11250
代理人 张琳琳

(51)Int.Cl.
*G06K 9/46*(2006.01)
*G06K 9/62*(2006.01)
*G06K 9/68*(2006.01)

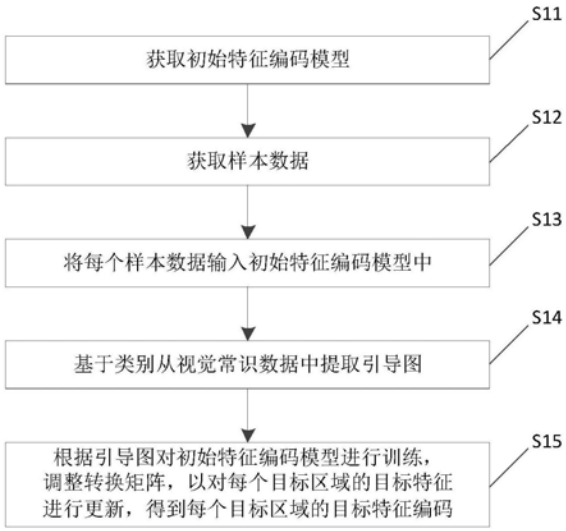
权利要求书3页 说明书18页 附图11页

(54)发明名称

特征编码模型、视觉关系检测模型的训练方法及检测方法

(57)摘要

本发明涉及视觉关系检测技术领域，具体涉及特征编码模型、视觉关系检测模型的训练方法及检测方法；其中，特征编码模型的训练方法包括获取初始特征编码模型；获取样本数据；将每个样本数据输入初始特征编码模型中；基于类别从视觉常识数据中提取引导图；根据引导图对初始特征编码模型进行训练，调整转换矩阵，以对每个目标区域的目标特征进行更新，得到每个目标区域的目标特征编码。利用视觉常识中与该类别相关的引导图一方面弥补了样本数据不足的缺陷，使得在对目标特征进行再次编码时能够有足够的样本数据支撑，另一方面保证了目标特征编码时就已经引入关系感知，为后续视觉关系的检测提供的条件，进而能够提高视觉关系检测的准确率。




CN 110390340 A

1.一种特征编码模型的训练方法,其特征在于,包括:

获取初始特征编码模型;其中,所述初始特征编码模型包括级联至少一层的多头注意力模块,每个所述多头注意力模块的参数包括一组互相独立的转换矩阵;

获取样本数据;其中,每个所述样本数据包括样本图像中目标区域的目标特征以及对应的类别;

将每个所述样本数据输入所述初始特征编码模型中;

基于所述类别从视觉常识数据中提取引导图;其中,所述引导图用于表示视觉常识对应于所述类别的目标类别;

根据所述引导图对所述初始特征编码模型进行训练,调整所述转换矩阵,以对每个所述目标区域的目标特征进行更新,得到每个所述目标区域的目标特征编码。

2.根据权利要求1所述的方法,其特征在于,所述根据所述引导图对所述初始特征编码模型进行训练,调整所述转换矩阵,以对每个所述目标区域的目标特征进行更新,得到每个所述目标区域的目标特征编码,包括:

对于每个所述样本数据,基于所述转换矩阵以及所述目标特征,计算每个所述样本图像的注意力矩阵;其中,所述注意力矩阵用于表示所述样本图像中每个所述目标区域对其他所述目标区域的注意力;

利用所述转换矩阵以及所述注意力矩阵,联合所有所述多头注意力模块的输出,并加入所述目标特征,以得到每个所述目标区域的目标特征编码;

基于所述目标特征编码以及所述引导图,计算损失函数的值;

利用所述损失函数的值以及第一学习率对所述初始特征编码模型进行强学习,调整所述转换矩阵,以更新所述目标特征编码。

3.根据权利要求2所述的方法,其特征在于,所述损失函数的定义如下:

$$L_{attn} = \sum_{S_j \in S} \sum_{h=1}^{H} f(A_h, S_i);$$

其中,$L_{attn}$为所述损失函数的值;S为引导图序列;$S_i$为引导图序列中的第i个引导图;f(•)为损失函数;H为每个所述多头注意力模块的个数;$A_h$为所述注意力矩阵。

4.根据权利要求3所述的方法,其特征在于,采用如下公式计算所述注意力矩阵以及所述目标特征编码:

$$A_h(v_i, v_j) = soft\max\left(\left(W_h^Q v_j\right)\left(W_h^K v_i\right)^T / \sqrt{d}\right);$$

$$\hat{v}_i = v_i + concat_{h=1}^{H}\left\{\sum_{j=1}^{N} A_h(v_i, v_j) W_h^V v_j\right\};$$

其中,$v_i$,$v_j$为所述样本图像中的任意两个所述目标特征;$W_h^Q$,$W_h^K$,$W_h^V$为一组互相独立的转换矩阵;$A_h(v_i,v_j)$为目标特征$v_i$对目标特征$v_j$的注意力;d为所述目标特征的维数;$\hat{v}_i$为对应于目标特征$v_i$的目标特征编码;N为所述样本图像中目标区域的个数。

5.根据权利要求2所述的方法,其特征在于,所述利用所述损失函数的值以及第一学习率对所述初始特征编码模型进行强学习,以更新所述转换矩阵,包括:

利用所述损失函数的值，计算第一梯度估计；

利用所述第一梯度估计以及所述第一学习率，对所述转换矩阵进行更新。

6.一种视觉关系检测模型的训练方法，其特征在于，包括：

获取目标检测模型；所述目标检测模型用于检测第二样本图像中的目标候选区域、每个所述目标候选区域的目标特征及其对应的类别；

获取特征编码模型；其中，所述特征编码模型是根据权利要求1-5中任一项所述的特征编码模型的训练方法训练得到；所述特征编码模型包括目标特征编码模型，和/或，关系特征编码模型；所述目标特征编码模型的输入包括所述目标候选区域的目标特征以及所述类别对应的词向量，输出为所述目标候选区域的目标特征编码；所述关系特征编码模型的输入包括所述目标候选区域的目标特征编码，以及所述目标特征编码的类别对应的词向量，输出为所述目标候选区域的关系特征编码；

将所述目标检测模型以及所述特征编码模型级联，以得到初始视觉关系检测模型；其中，所述特征编码模型通过分类模型与输出连接；

基于第二学习率对所述初始视觉关系检测模型进行训练，调整所述特征编码模型的参数，以得到视觉关系检测模型；其中，所述第二学习率小于训练所述特征编码模型的学习率。

7.根据权利要求6所述的方法，其特征在于，所述基于第二学习率对所述初始视觉关系检测模型进行训练，调整所述特征编码模型的参数，以得到视觉关系检测模型，包括：

计算所述特征编码模型的损失函数的值；

利用所述损失函数的值，计算第二梯度估计；

利用所述第二梯度估计以及所述第二学习率，对所述特征编码模型中的所述转换矩阵进行更新。

8.根据权利要求6所述的方法，其特征在于，所述特征编码模型为级联的所述目标特征编码模型与所述关系特征编码模型；其中，所述特征编码模型通过特征分类模型与所述关系特征编码模型级联，所述关系特征编码模型通过关系分类模型与所述输出连接。

9.根据权利要求8所述的方法，其特征在于，所述特征分类模型为第一全连接层，所述关系分类模型为第二全连接层。

10.一种视觉关系检测方法，其特征在于，包括：

获取待检测图像；

将所述待检测图像输入视觉关系检测模型中，以得到所述待检测图像的视觉关系；其中，所述视觉关系检测模型是根据权利要求6-9中任一项所述的视觉关系检测模型的训练方法训练得到的。

11.根据权利要求10所述的方法，其特征在于，所述将所述待检测图像输入视觉关系检测模型中，以得到所述待检测图像的视觉关系，包括：

将所述待检测图像输入所述目标检测模型中，输出至少一个目标候选区域、每个目标候选区域的特征向量以及类别概率向量；

基于所述特征向量以及所述类别概率向量，利用所述目标特征编码模型得到所述目标候选区域的目标特征编码；

将所述目标特征编码输入特征分类模型中，以得到对应的目标类别向量；

基于所述目标特征编码以及所述目标类别向量，利用所述关系特征编码模型得到所述目标候选区域的关系特征编码；

将所有所述目标候选区域对应的关系特征编码两两联合，输入关系分类模型中，以得到任意两个所述目标候选区域的视觉关系。

12.根据权利要求11所述的方法，其特征在于，所述基于所述特征向量以及所述类别概率向量，利用所述目标特征编码模型得到所述目标候选区域的目标特征编码，包括：

提取每个所述类别概率向量中概率最大的类别对应的第一词向量；

对于每个所述目标候选区域，联合所述特征向量、所述第一词向量以及所述目标候选区域的全图特征向量，以得到每个所述目标候选区域的第一联合特征向量；

依次将每个所述第一联合特征向量输入目标特征编码模型中，以得到每个目标候选区域的目标特征编码。

13.根据权利要求11所述的方法，其特征在于，所述基于所述目标特征编码以及所述目标类别向量，利用所述关系特征编码模型得到所述目标候选区域的关系特征编码，包括：

提取每个所述目标候选区域的所述目标类别向量中得分值最高的类别的第二词向量；

对于每个所述目标候选区域，联合所述目标特征编码向量以及所述第二词向量，以得到每个所述目标候选区域的第二联合特征向量；

依次将每个所述第二联合特征向量输入关系特征编码模型中，以得到每个目标候选区域的关系特征编码。

14.一种电子设备，其特征在于，包括：

存储器和处理器，所述存储器和所述处理器之间互相通信连接，所述存储器中存储有计算机指令，所述处理器通过执行所述计算机指令，从而执行权利要求1-5中任一项所述的特征编码模型的训练方法，或执行权利要求6-9中任一项所述的视觉关系检测模型的训练方法，或执行权利要求10-13中任一项所述的视觉关系检测方法。

15.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有计算机指令，所述计算机指令用于使所述计算机执行权利要求1-5中任一项所述的特征编码模型的训练方法，或执行权利要求6-9中任一项所述的视觉关系检测模型的训练方法，或执行权利要求10-13中任一项所述的视觉关系检测方法。

# 特征编码模型、视觉关系检测模型的训练方法及检测方法

## 技术领域

[0001] 本发明涉及视觉关系检测技术领域，具体涉及特征编码模型、视觉关系检测模型的训练方法及检测方法。

## 背景技术

[0002] 近年来，深度学习在图像识别(如图像分类、目标检测、目标分割等)任务上取得了突破性进展。其中，要实现计算机理解场景，重要的一环是视觉关系检测，即对于一幅输入图片，预测图片中的目标物体的位置和类别，以及目标和目标之间的关系的类别。

[0003] 对于视觉关系检测常采用的方法是，对目标和关系进行编码，再通过分类器预测目标类别和关系类别。这些方法常使用循环神经网络逐步地融合区域特征，使得最终每个区域特征都参考了所有其他区域的信息，再将区域特征两两匹配，输入关系分类器，得到最终的视觉关系预测结果。

[0004] 上述检测方法中所采用的循环神经网络模型需要事先采用大量的样本数据进行训练，而真实场景中视觉关系的类别常常存在严重的不均衡问题，即一些常见关系(如，<人-穿着-牛仔裤>)出现频次远远高于不常见关系(如，<猫-睡在-车上>)，这导致上述基于大数据学习的方法因无法获得足够的样本而在不常见关系的预测中失效，进而影响视觉关系检测的准确率。

## 发明内容

[0005] 有鉴于此，本发明实施例提供了一种特征特征编码模型、视觉关系检测模型的训练方法及检测方法，以解决视觉关系检测的准确率偏低的问题。

[0006] 根据第一方面，本发明实施例提供了一种特征编码模型的训练方法，包括：

[0007] 获取初始特征编码模型；其中，所述初始特征编码模型包括级联至少一层的多头注意力模块，每个所述多头注意力模块的参数包括一组互相独立的转换矩阵；

[0008] 获取样本数据；其中，每个所述样本数据包括样本图像中目标区域的目标特征以及对应的类别；

[0009] 将每个所述样本数据输入所述初始特征编码模型中；

[0010] 基于所述类别从视觉常识数据中提取引导图；其中，所述引导图用于表示视觉常识对应于所述类别的目标类别；

[0011] 根据所述引导图对所述初始特征编码模型进行训练，调整所述转换矩阵，以对每个所述目标区域的目标特征进行更新，得到每个所述目标区域的目标特征编码。

[0012] 本发明实施例提供的特征编码模型的训练方法，通过视觉常识中与目标区域的类别对应的引导图加入特征编码模型的训练中，即，利用视觉常识中与该类别相关的引导图一方面弥补了样本数据不足的缺陷，使得在对目标特征进行再次编码时能够有足够的样本数据支撑，另一方面保证了目标特征编码时就已经引入关系感知，为后续视觉关系的检测提供的条件，进而能够提高视觉关系检测的准确率。

[0013] 结合第一方面，在第一方面第一实施方式中，所述根据所述引导图对所述初始特征编码模型进行训练，调整所述转换矩阵，以对每个所述目标区域的目标特征进行更新，得到每个所述目标区域的目标特征编码，包括：

[0014] 对于每个所述样本数据，基于所述转换矩阵以及所述目标特征，计算每个所述样本图像的注意力矩阵；其中，所述注意力矩阵用于表示所述样本图像中每个所述目标区域对其他所述目标区域的注意力；

[0015] 利用所述转换矩阵以及所述注意力矩阵，联合所有所述多头注意力模块的输出，并加入所述目标特征，以得到每个所述目标区域的目标特征编码；

[0016] 基于所述目标特征编码以及所述引导图，计算损失函数的值；

[0017] 利用所述损失函数的值以及第一学习率对所述初始特征编码模型进行强学习，调整所述转换矩阵，以更新所述目标特征编码。

[0018] 本发明实施例提供的特征编码模型的训练方法，在对目标区域进行进一步编码时，采用注意力矩阵所反应出的区域与区域之间的关系信息进行编码，认为对于每个目标区域，编码和其相关的区域的上下文信息能够更好地帮助预测该目标区域的类别以及它所涉及的视觉关系，从而能够提高后续视觉关系检测的准确率。

[0019] 结合第一方面第一实施方式，在第一方面第二实施方式中，所述损失函数的定义如下：

[0020] $$L_{attn} = \sum_{S_i \in S} \sum_{h=1}^{H} f(A_h, S_i)$$；

[0021] 其中，$L_{attn}$为所述损失函数的值；S为引导图序列；$S_i$为引导图序列中的第i个引导图；$f(\bullet)$为损失函数；H为每个所述多头注意力模块的个数；$A_h$为所述注意力矩阵。

[0022] 本发明实施例提供的特征编码模型的训练方法，通过引入视觉常识作为特征编码的学习，从而可以在后续的视觉关系检测中能够引导显式地辅助视觉关系的学习，为视觉关系检测准确率的提高提供了条件。

[0023] 结合第一方面第二实施方式，在第一方面第三实施方式中，采用如下公式计算所述注意力矩阵以及所述目标特征编码：

[0024] $$A_h(v_i, v_j) = soft\max\left(\frac{(W_h^Q v_j)(W_h^K v_i)^T}{\sqrt{d}}\right)$$；

[0025] $$\hat{v}_i = v_i + concat_{h=1}^{H}\left\{\sum_{j=1}^{N} A_h(v_i, v_j) W_h^V v_j\right\}$$；

[0026] 其中，$v_i$，$v_j$为所述样本图像中的任意两个所述目标特征；$W_h^Q$，$W_h^K$，$W_h^V$为一组互相独立的转换矩阵；$A_h(v_i, v_j)$为目标特征$v_i$对目标特征$v_j$的注意力；d为所述目标特征的维数；$\hat{v}_i$为对应于目标特征$v_i$的目标特征编码；N为所述样本图像中目标区域的个数。

[0027] 结合第一方面第一实施方式，在第一方面第四实施方式中，所述利用所述损失函数的值以及第一学习率对所述初始特征编码模型进行强学习，以更新所述转换矩阵，包括：

[0028] 利用所述损失函数的值，计算第一梯度估计；

[0029] 利用所述第一梯度估计以及所述第一学习率，对所述转换矩阵进行更新。

[0030] 本发明实施例提供的特征编码模型的训练方法，采用梯度估计的方法对转换矩阵进行更新，能够提高该特征编码模型的训练效率。

[0031] 根据第二方面，本发明实施例还提供了一种视觉关系检测模型的训练方法，包括：

[0032] 获取目标检测模型；所述目标检测模型用于检测第二样本图像中的目标候选区域、每个所述目标候选区域的目标特征及其对应的类别；

[0033] 获取特征编码模型；其中，所述特征编码模型是根据权利要求1-5中任一项所述的特征编码模型的训练方法训练得到；所述特征编码模型包括目标特征编码模型，和/或，关系特征编码模型；所述目标特征编码模型的输入包括所述目标候选区域的目标特征以及所述类别对应的词向量，输出为所述目标候选区域的目标特征编码；所述关系特征编码模型的输入包括所述目标候选区域的目标特征编码，以及所述目标特征编码的类别对应的词向量，输出为所述目标候选区域的关系特征编码；

[0034] 将所述目标检测模型以及所述特征编码模型级联，以得到初始视觉关系检测模型；其中，所述特征编码模型通过分类模型与输出连接；

[0035] 基于第二学习率对所述初始视觉关系检测模型进行训练，调整所述特征编码模型的参数，以得到视觉关系检测模型；其中，所述第二学习率小于训练所述特征编码模型的学习率。

[0036] 本发明实施例提供的视觉关系检测模型的训练方法，其中在目标特征编码模型以及关系特征编码模型中均采用了引导图对其进行训练，利用视觉常识中与该类别相关的引导图一方面弥补了样本数据不足的缺陷，使得在对目标特征进行再次编码时能够有足够的样本数据支撑，另一方面保证了目标特征编码时就已经引入关系感知，能够保证后续利用该模型进行视觉关系检测时准确率的提高。

[0037] 结合第二方面，在第二方面第一实施方式中，所述基于第二学习率对所述初始视觉关系检测模型进行训练，调整所述目标特征编码模型与所述关系特征编码模型的参数，以得到视觉关系检测模型，包括：

[0038] 计算所述特征编码模型的损失函数的值；

[0039] 利用所述损失函数的值，计算第二梯度估计；

[0040] 利用所述第二梯度估计以及所述第二学习率，对所述特征编码模型中的所述转换矩阵进行更新。

[0041] 本发明实施例提供的视觉关系检测模型的训练方法，通过采用小于第一学习率的第二学习率对目标特征编码模型以及关系特征编码模型中的转换矩阵进行微调，一方面能够保证转换矩阵的准确性，另一方面由于采用了较小的学习率能够保证较高的训练效率。

[0042] 结合第二方面，在第二方面第二实施方式中，所述特征编码模型为级联的所述目标特征编码模型与所述关系特征编码模型；其中，所述特征编码模型通过特征分类模型与所述关系特征编码模型级联，所述关系特征编码模型通过关系分类模型与所述输出连接。

[0043] 结合第二方面第一实施方式，在第二方面第三实施方式中，所述特征分类模型为第一全连接层，所述关系分类模型为第二全连接层。

[0044] 根据第三方面，本发明实施例还提供了一种视觉关系检测方法，包括：

[0045] 获取待检测图像；

[0046] 将所述待检测图像输入视觉关系检测模型中，以得到所述待检测图像的视觉关

系;其中,所述视觉关系检测模型是根据本发明第三方面,或第三发面任一项实施方式中所述的视觉关系检测模型的训练方法训练得到的。

[0047] 本发明实施例提供的视觉关系检测方法,由于在视觉关系检测模型中引入视觉常识,利用视觉常识中与该类别相关的引导图一方面弥补了样本数据不足的缺陷,使得在对目标特征进行再次编码时能够有足够的样本数据支撑,另一方面保证了特征编码时就已经引入关系感知,提高了视觉关系检测的准确率。

[0048] 结合第三方面,在第三方面第一实施方式中,所述将所述待检测图像输入视觉关系检测模型中,以得到所述待检测图像的视觉关系,包括:

[0049] 将所述待检测图像输入所述目标检测模型中,输出至少一个目标候选区域、每个目标候选区域的特征向量以及类别概率向量;

[0050] 基于所述特征向量以及所述类别概率向量,利用所述目标特征编码模型得到所述目标候选区域的目标特征编码;

[0051] 将所述目标特征编码输入特征分类模型中,以得到对应的目标类别向量;

[0052] 基于所述目标特征编码以及所述目标类别向量,利用所述关系特征编码模型得到所述目标候选区域的关系特征编码;

[0053] 将所有所述目标候选区域对应的关系特征编码两两联合,输入关系分类模型中,以得到任意两个所述目标候选区域的视觉关系。

[0054] 本发明实施例提供的视觉关系检测方法,通过目标特征编码模型输出对每个目标候选区域的带有关系感知的特征编码,在通过特征分类模型,进一步预测出更加准确的目标类别;且关系特征编码模型输出同样是输出带有关系感知的特征编码,进一步提高了预测出的关系类别的准确性。

[0055] 结合第三方面第一实施方式,在第三方面第二实施方式中,所述基于所述特征向量以及所述类别概率向量,利用所述目标特征编码模型得到所述目标候选区域的目标特征编码,包括:

[0056] 提取每个所述类别概率向量中概率最大的类别对应的第一词向量;

[0057] 对于每个所述目标候选区域,联合所述特征向量、所述第一词向量以及所述目标候选区域的全图特征向量,以得到每个所述目标候选区域的第一联合特征向量;

[0058] 依次将每个所述第一联合特征向量输入目标特征编码模型中,以得到每个目标候选区域的目标特征编码。

[0059] 结合第三方面第一实施方式,在第三方面第三实施方式中,所述基于所述目标特征编码以及所述目标类别向量,利用所述关系特征编码模型得到所述目标候选区域的关系特征编码,包括:

[0060] 提取每个所述目标候选区域的所述目标类别向量中得分值最高的类别的第二词向量;

[0061] 对于每个所述目标候选区域,联合所述目标特征编码向量以及所述第二词向量,以得到每个所述目标候选区域的第二联合特征向量;

[0062] 依次将每个所述第二联合特征向量输入关系特征编码模型中,以得到每个目标候选区域的关系特征编码。

[0063] 根据第四方面,本发明实施例还提供了一种电子设备,包括:

[0064] 存储器和处理器，所述存储器和所述处理器之间互相通信连接，所述存储器中存储有计算机指令，所述处理器通过执行所述计算机指令，从而执行本发明第一方面，或第一方面任一项实施方式中所述的特征编码模型的训练方法，或执行本发明第二方面，或第二方面任一项实施方式中所述的视觉关系检测模型的训练方法，或执行本发明第三方面，或第三方面任一项实施方式中所述的视觉关系检测方法。

[0065] 根据第五方面，本发明实施例还提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机指令，所述计算机指令用于使所述计算机执行本发明第一方面，或第一方面任一项实施方式中所述的特征编码模型的训练方法，或执行本发明第二方面，或第二方面任一项实施方式中所述的视觉关系检测模型的训练方法，或执行本发明第三方面，或第三方面任一项实施方式中所述的视觉关系检测方法。

## 附图说明

[0066] 为了更清楚地说明本发明具体实施方式或现有技术中的技术方案，下面将对具体实施方式或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施方式，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0067] 图1是根据本发明实施例的特征编码模型的训练方法的流程图；

[0068] 图2是根据本发明实施例的带标注的样本图像；

[0069] 图3是根据本发明实施例的视觉常识的示意图；

[0070] 图4是根据本发明实施例的引导图；

[0071] 图5是根据本发明实施例的关系特征图；

[0072] 图6是根据本发明实施例的特征编码模型的训练方法的流程图；

[0073] 图7是根据本发明实施例的多头注意力模块的结构框图；

[0074] 图8是根据本发明实施例的视觉关系检测模型的训练方法的流程图；

[0075] 图9是根据本发明实施例的视觉关系检测模型的训练方法的流程图；

[0076] 图10是根据本发明实施例的视觉关系检测模型的结构示意图；

[0077] 图11是根据本发明实施例的视觉关系检测模型的结构示意图；

[0078] 图12是根据本发明实施例的视觉关系检测方法的流程图；

[0079] 图13是根据本发明实施例的视觉关系检测方法的流程图；

[0080] 图14是根据本发明实施例的视觉关系检测方法的部分流程图；

[0081] 图15是根据本发明实施例的视觉关系检测方法的部分流程图；

[0082] 图16是根据本发明实施例的数据集信息；

[0083] 图17是根据本发明实施例的3个视觉任务在VG和VG-MSDN上的评测结果；

[0084] 图18是根据本发明实施例的3个视觉任务在VG-MSDN和VG-DR-Net上的评测结果；

[0085] 图19是根据本发明实施例的模型的平均每轮的训练时间和模型参数；

[0086] 图20是根据本发明实施例的特征编码模型的训练装置的结构框图；

[0087] 图21是根据本发明实施例的视觉关系检测模型的训练装置的结构框图；

[0088] 图22是根据本发明实施例的视觉关系检测装置的结构框图；

[0089] 图23是本发明实施例提供的电子设备的硬件结构示意图。

## 具体实施方式

[0090] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域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，都属于本发明保护的范围。

[0091] 需要说明的是，发明人通过对视觉关系检测的研究过程中发现，很多基于大数据学习的视觉关系检测方法在不常见关系的预测中失效，是由于真实场景中视觉关系的类别存在严重不均衡问题，那么就会导致基于大数据学习的视觉关系检测方法由于无法获得足够的样本而失效。同时，发明人注意到人类在识别场景中目标和目标之间的视觉关系时，对于不常见关系依然能够给出准确预测，这是因为人类在日常生活中以及积累了自己的视觉常识，对视觉常识中已存在的视觉关系或模式都可以进行很好的识别。

[0092] 其中，本发明所述的视觉常识可以理解为是一张数据表，在该数据表中存储有人类视觉常识下的目标与目标之间的关系。视觉常识数据等于是从人类世界所能看到的视觉常识中抽取出来的目标关系，是从大数据集中统计而来。比如，看到“person”和“bike”两个目标，根据视觉常识，它们的关系可能是“ride”的概率为0.7，是“next to”的概率为0.3，对于一幅图片，当目标分类结果显示“person”和“bike”时，视觉常识提供可能的关系类别。

[0093] 因此，本申请中所涉及到的特征编码模型是基于视觉常识的引导和具体的目标特征来预测实际的关系类别，在已有特征编码的基础上，引入带有关系感知的特征，从而再次对特征进行编码。具体地，本发明实施例中提出的特征编码模型是一种渐进知识驱动的特征变换模块(Progressive Knowledge-driven Transformer，简称为PKT)，其用于对检测到的目标基于已有视觉常识进行关系编码，从而帮助模型检测视觉关系，尤其是不常见的视觉关系。

[0094] 根据本发明实施例，提供了一种特征编码模型的训练方法实施例，需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0095] 在本实施例中提供了一种特征编码模型的训练方法，可用于上述的移动终端，如手机、平板电脑等，图1是根据本发明实施例的特征编码模型的流程图，如图1所示，该流程包括如下步骤：

[0096] S11，获取初始特征编码模型。

[0097] 其中，所述初始特征编码模型包括级联至少一层的多头注意力模块，每个多头注意力模块的参数包括一组互相独立的转换矩阵。

[0098] 具体地，初始特征编码模型包括至少一层多头注意力模块的级联，例如可以是3层，也可以是4层等等，具体级联的多头注意力模块的层数可以根据实际情况进行具体设置。以级联4层多头注意力模块为例，该初始特征编码模型为第一层多头注意力模块、第二层多头注意力模块、第三层多头注意力模块以及第四层多头注意力模块的顺次级联，前一层多头注意力模块的输出接入后一层多头注意力模块的输入。

[0099] 对于每个多头注意力模块而言，其可以包括3个、4个，或5个注意力模块等等。对每个多头注意力模块所包括的注意力模块的个数可以根据具体情况进行设置，在此并不做任

何限制。

[0100] 电子设备所获取到的初始特征编码模型可以是从外界获取到的，也可以是在训练时实时构建出的。其中，每个多头注意力模块中的转换矩阵可以是随机初始化得到的，也可以是事先定义得到的等等。

[0101] S12，获取样本数据。

[0102] 其中，每个样本数据包括样本图像中目标区域的目标特征以及对应的类别。

[0103] 电子设备所获取到的样本数据可以是从带标注的样本图像中，直接提取出各个目标区域的目标特征，以及各个目标特征对应的类别。其中，所述的类别可以是多个类别，每个类别对应于一个概率值；也可以是所有类别中概率最大的类别等等。

[0104] 可选地，样本数据还可以是人为标注出样本图像中的目标区域，以及各个目标区域对应的类别；至于目标区域的目标特征可以是通过目标检测器得到，也可以是通过其他方式得到等等。

[0105] S13，将每个样本数据输入初始特征编码模型中。

[0106] 在对初始特征编码模型进行训练时，电子设备需要将S12中获得的样本数据输入该初始特征编码模型中。

[0107] S14，基于类别从视觉常识数据中提取引导图。

[0108] 其中，所述引导图用于表示视觉常识对应于类别的目标类别。

[0109] 电子设备在S12中获得样本图像中各个目标区域的类别之后，可以从视觉常识中提取出对应于该类别的引导图。

[0110] 具体地，请参见图2，图2示出了某个样本图像所标注出的4个目标，分别是bear、water、tree以及branch。电子设备在获取目标区域对应的类别之后，从视觉常识中提取出相应的引导图。其中，视觉常识可以采用图3的形式表示，其表征出了在视觉常识中上述4个目标之间的所有可能的关系。电子设备基于图3的视觉常识，以及图2样本图像中的各个目标，从图3中的视觉常识中提取出与图2对应的引导图，如图4所示。

[0111] 在后续的步骤中，电子设备基于图4的引导图可以检测出图2中目标的视觉关系如图5所示。这部分内容将在下文中进行详细描述。

[0112] S15，根据引导图对初始特征编码模型进行训练，调整转换矩阵，以对每个目标区域的目标特征进行更新，得到每个目标区域的目标特征编码。

[0113] 电子设备在对初始特征编码模型进行训练时，以引导图作为训练的基准，每训练一次就计算该初始特征编码模型的损失函数，从而得到训练结果与实际值之间的差异，然后基于该差异对转换矩阵的进行调整。电子设备在调整转换矩阵之后，相应地该初始特征编码模型也就发生更新，即输出的目标特征编码值同样发生更新。

[0114] 电子设备利用引导图对初始特征编码模型进行训练，以调整转换矩阵，最终得到每个目标区域的目标特征编码。

[0115] 本实施例提供的特征编码模型的训练方法，通过视觉常识中与目标区域的类别对应的引导图加入特征编码模型的训练中，即，利用视觉常识中与该类别相关的引导图一方面弥补了样本数据不足的缺陷，使得在对目标特征进行再次编码时能够有足够的样本数据支撑，另一方面保证了目标特征编码时就已经引入关系感知，为后续视觉关系的检测提供的条件，进而能够提高视觉关系检测的准确率。

[0116] 在本实施例中提供了一种特征编码模型的训练方法，可用于上述的移动终端，如手机、平板电脑等，图6是根据本发明实施例的特征编码模型的流程图，如图6所示，该流程包括如下步骤：

[0117] S21，获取初始特征编码模型。

[0118] 其中，所述初始特征编码模型包括级联至少一层的多头注意力模块，每个多头注意力模块的参数包括一组互相独立的转换矩阵。

[0119] 具体地，如图7所示，图7示出了一层多头注意力模块，其中，该输出特征编码模型可以是2层，3层或多个图7中的多头注意力模块的级联

[0120] 对于每层多头注意力模块而言，其所包括的注意力模块的个数为H，在图7中以H＝3为例进行描述的。该多头注意力模块的输入为特征$v_i$，转换矩阵为3个，初始转换矩阵表示为$W_0^Q$、$W_0^K$以及$W_0^V$。

[0121] 每层多层注意力模块的输出为该多层注意力模块中所有注意力模块的输出的联合，即Concat。

[0122] S22，获取样本数据。

[0123] 其中，每个样本数据包括样本图像中目标区域的目标特征以及对应的类别。

[0124] 详细请参见图1所示实施例的S12，在此不再赘述。

[0125] S23，将每个样本数据输入初始特征编码模型中。

[0126] 详细请参见图1所示实施例的S13，在此不再赘述。

[0127] S24，基于类别从视觉常识数据中提取引导图。

[0128] 其中，所述引导图用于表示视觉常识对应于类别的目标类别。

[0129] 详细请参见图1所示实施例的S14，在此不再赘述。

[0130] S25，根据引导图对初始特征编码模型进行训练，调整转换矩阵，以对每个目标区域的目标特征进行更新，得到每个目标区域的目标特征编码。

[0131] 具体地，S25可以包括如下步骤：

[0132] S251，对于每个样本数据，基于转换矩阵以及目标特征，计算每个样本图像的注意力矩阵。

[0133] 其中，所述注意力矩阵用于表示样本图像中每个目标区域对其他目标区域的注意力。

[0134] 具体地，样本图像具有N个目标区域，那么该样本图像所对应的注意力矩阵为N*N的矩阵，该矩阵的每一个元素表示一个目标区域对其他所有n个目标区域的注意力，即区域间的关联关系。例如，该注意力矩阵中(i，j)位置的元素代表$v_i$对$v_j$的注意力，也就是这两个目标区域之间的相关性。

[0135] 如图7所示，注意力矩阵可以采用如下公式计算得到：

[0136] $$A_h(v_i, v_j) = soft\max\left(\frac{(W_h^Q v_j)(W_h^K v_i)^T}{\sqrt{d}}\right);$$

[0137] 其中，$v_i$，$v_j$为所述样本图像中的任意两个所述目标特征；$W_h^Q$，$W_h^K$，$W_h^V$为一组互相独立的转换矩阵；$A_h(v_i, v_j)$为目标特征$v_i$对目标特征$v_j$的注意力；d为所述目标特征的维数；

[0138] S252，利用转换矩阵以及注意力矩阵，联合所有多头注意力模块的输出，并加入目标特征，以得到每个目标区域的目标特征编码。

[0139] 具体地，采用如下公式计算所述注意力矩阵以及所述目标特征编码：

[0140] $$\hat{v}_i = v_i + concat_{h=1}^{H}\left\{\sum_{j=1}^{N} A_h(v_i, v_j) W_h^V v_j\right\};$$

[0141] 其中，$\hat{v}_i$为对应于目标特征$v_i$的目标特征编码；N为所述样本图像中目标区域的个数；H为每个多头注意力模块中注意力模块的个数。

[0142] S253，基于目标特征编码以及引导图，计算损失函数的值。

[0143] 通过引入引导图来约束注意力矩阵中的连接，每幅图片的引导图是根据图片中已预测出的目标类别在视觉常识数据中取子图所得。引导图为有向图，以目标区域为节点，区域和区域间关系为边。引导图可以以外部监督的形式对注意力模块产生影响，损失函数定义如下：

[0144] $$L_{attn} = \sum_{S_i \in S}\sum_{h=1}^{H} f(A_h, S_i);$$

[0145] 其中，$L_{attn}$为所述损失函数的值；S为引导图序列；$S_i$为引导图序列中的第i个引导图；f(•)为损失函数；H为每个所述多头注意力模块的个数；$A_h$为所述注意力矩阵。需要说明的是，在此对损失函数f(•)的具体函数并不做任何限制，可以根据实际情况进行具体设置。

[0146] 具体地，引导图是根据当前目标的类别在视觉常识中抽取子图，实际计算中，视觉常识数据是一个矩阵，代表每个目标列别和其他目标类别之间的关系。引导图是根据当前已预测得到的目标类别，去视觉常识的矩阵中抽取对应类别的行或列，构成新的矩阵，从而表达目标和目标之间的关系，构建出引导图。

[0147] S254，利用损失函数的值以及第一学习率对初始特征编码模型进行强学习，调整转换矩阵，以更新目标特征编码。

[0148] 电子设备在S253中得到损失函数的值之后，对初始特征编码模型进行强学习。所述的强学习可以是采用梯度下降方法(例如，随机梯度下降方法)，也可以采用其他方法等等。对初始特征编码模型进行强学习的目的在于，调整转换矩阵，从而使得S252中计算得到的目标特征编码进行更新。

[0149] 对初始特征编码模型进行强学习的迭代停止条件可以是转换矩阵不再变化，也可以设置迭代次数等等。

[0150] 例如，S254可以采用如下步骤实现：

[0151] (1)利用损失函数的值，计算第一梯度估计。

[0152] 其中，采用随机梯度下降方法计算梯度估计，即注意力矩阵与引导图之间的差异。

[0153] (2)利用第一梯度估计以及第一学习率，对转换矩阵进行更新。

[0154] 其中，第一学习率用于表示转换矩阵的变化，在第一梯度估计的基础上，结合第一学习率即可得到更新后的转换矩阵。

[0155] 作为本实施例的一个具体应用实例，可以采用如下方法训练该特征编码模型，训练使用随机梯度下降(即，SGD)算法，第一学习率为0.001，批处理大小为6，该特征编码模型中多头注意力模块的层数为4，每层多头注意力模块中多头注意力数为4，训练至模型参数不再更新为止，一般不多于15轮。

[0156] 本实施例提供的特征编码模型的训练方法，在对目标区域进行进一步编码时，采用注意力矩阵所反应出的区域与区域之间的关系信息进行编码，认为对于每个目标区域，编码和其相关的区域的上下文信息能够更好地帮助预测该目标区域的类别以及它所涉及的视觉关系，从而能够提高后续视觉关系检测的准确率。

[0157] 需要说明的是，本发明提出的渐进知识驱动的特征变换模块(PKT)在目标分类模型和关系分类模型中都作为关系感知的特征编码器，起到了至关重要的作用。实际上，PKT可以作为特征编码变换器作用在任何具有一定相关性的特征集合上。一般来说，对于图片I的一组输入特征，PKT首先学习一组互相独立的转换矩阵，每个$v_i$根据注意力矩阵A去关注和它相关的区域$v_j$，将多头注意力的结果联合起来，并加入残差$v_i$可以得到更新后的特征编码。

[0158] 同时通过引入引导图来约束注意力矩阵中的连接，每幅图片的引导图是根据图片中已预测出的目标类别在视觉常识数据中取子图取得。引导图为有向图，以目标区域为节点，区域和区域间关系为编。引导图可以以外部监督的形式对注意力模块产生影响。

[0159] 在对区域特征进行进一步编码时，都将区域与区域之间的关系信息加入编码，认为对于每个区域特征，编码和其相关的区域的上下文信息能够更好地帮助预测该区域类别以及它所涉及的视觉关系。同时引入视觉常识作为引导显式地辅助视觉关系的学习，而不是根据区域分类和关系分类的训练隐式学习关系。

[0160] 根据本发明实施例，提供了一种特征编码模型的训练方法实施例，需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0161] 在本实施例中提供了一种视觉关系检测模型的训练方法，可用于上述的移动终端，如手机、平板电脑等，图8是根据本发明实施例的视觉关系检测模型的训练方法的流程图，如图8所示，该流程包括如下步骤：

[0162] S31，获取目标检测模型。

[0163] 所述目标检测模型用于检测第二样本图像中的目标候选区域、每个所述目标候选区域的目标特征及其对应的类别。

[0164] 其中，目标检测模型可以是基于VGGNet的Faster RCNN框架构建出的，也可以是其他框架构建出的等等，只需保证该目标检测模型能够检测出第二样本图像中的目标候选区域、每个目标候选区域的目标特征以及各个目标特征对应的类别即可。

[0165] 以基于VGGNet的Faster RCNN框架构建出的目标检测模型为例，在对该目标检测模型进行训练时，使用随机梯度下降算法SGD训练，学习率为0.001，批处理大小为6，在3块NVIDIA GeForce GTX 1080Ti GPU上并行训练50轮。其中，第二样本图像的大小为592*592，对于输出结果而言，每幅图片选取得分较高的前64个目标检测框作为目标候选区域。

[0166] S32，获取特征编码模型。

[0167] 其中，所述特征编码模型是根据上述任一项实施例中所述的特征编码模型的训练方法训练得到；所述特征编码模型包括目标特征编码模型，和/或，关系特征编码模型；所述目标特征编码模型的输入包括所述目标候选区域的目标特征以及所述类别对应的词向量，输出为所述目标候选区域的目标特征编码；所述关系特征编码模型的输入包括所述目标候

选区域的目标特征编码，以及所述目标特征编码的类别对应的词向量，输出为所述目标候选区域的关系特征编码。

[0168] 具体地，对于视觉关系检测模型而言，其所包括的特征编码模型可以仅为目标特征编码，也可以仅为关系特征编码，或者既包括目标特征编码又包括关系特征编码。

[0169] 在下文中，以该视觉关系检测模型包括目标特征编码以及关系特征编码为例进行详细描述。所述目标特征编码模型以及所述关系特征编码模型的训练方法与上述实施例中所述的特征编码模型的方法相同，不同的是两者所用的样本数据，即两者对应的转换矩阵。

[0170] 对于目标特征编码模型而言，其作用在于使得在输入的特征中加入带有关系感知的特征编码，以得到目标特征编码；对于关系特征编码模型而言，其作用在于使得在输入的特征中再次加入关系感知，以得到关系编码。

[0171] 如上文所述，在该视觉关系检测模型中目标特征编码模型的输入为S31中的目标检测模型输出的目标候选区域的目标特征、其类别对应的词向量，以及该目标候选区域的全图特征；输出为对应目标候选区域的目标特征编码。

[0172] 在该视觉关系检测模型中关系特征编码模型的输入为目标特征编码，以及对该目标特征编码进行分类后的类别对应的词向量；所处为该目标候选区域的关系特征编码。

[0173] S33，将目标检测模型以及特征编码模型级联，以得到初始视觉关系检测模型。

[0174] 其中，所述特征编码模型通过分类模型与输出连接。

[0175] 在S31以及S32中获取到目标检测模型以及特征编码模型之后，将这两个模型进行级联并在特征编码模型的基础上连接输出，从而得到初始视觉关系检测模型。其中，特征编码模型可以通过全连接层连接输出层。

[0176] S34，基于第二学习率对初始视觉关系检测模型进行训练，调整特征编码模型的参数，以得到视觉关系检测模型。

[0177] 其中，所述第二学习率小于训练特征编码模型的学习率。

[0178] 电子设备在S33中构建出初始视觉关系检测模型之后，固定S31中得到的目标检测模型的参数，在对该初始视觉关系检测模型进行训练时，仅调整特征编码模型的参数。

[0179] 本实施例提供的视觉关系检测模型的训练方法，其中在目标特征编码模型以及关系特征编码模型中均采用了引导图对其进行训练，利用视觉常识中与该类别相关的引导图一方面弥补了样本数据不足的缺陷，使得在对目标特征进行再次编码时能够有足够的样本数据支撑，另一方面保证了目标特征编码时就已经引入关系感知，能够保证后续利用该模型进行视觉关系检测时准确率的提高。

[0180] 在本实施例中提供了一种视觉关系检测模型的训练方法，可用于上述的移动终端，如手机、平板电脑等，图9是根据本发明实施例的视觉关系检测模型的训练方法的流程图，如图9所示，该流程包括如下步骤：

[0181] S41，获取目标检测模型。

[0182] 所述目标检测模型用于检测第二样本图像中的目标候选区域、每个所述目标候选区域的目标特征及其对应的类别。

[0183] 详细请参见图8所示实施例的S31，在此不再赘述。

[0184] S42，获取特征编码模型。

[0185] 其中，所述特征编码模型是根据上述任一项实施例中所述的特征编码模型的训练

方法训练得到;所述特征编码模型包括目标特征编码模型,和/或,关系特征编码模型;所述目标特征编码模型的输入包括所述目标候选区域的目标特征以及所述类别对应的词向量,输出为所述目标候选区域的目标特征编码;所述关系特征编码模型的输入包括所述目标候选区域的目标特征编码,以及所述目标特征编码的类别对应的词向量,输出为所述目标候选区域的关系特征编码。

[0186] 详细请参见图8所示实施例的S32,在此不再赘述。

[0187] S43,将目标检测模型以及特征编码模型级联,以得到初始视觉关系检测模型。

[0188] 其中,所述特征编码模型通过分类模型与输出连接。

[0189] 如图10所示,该视觉关系检测模型依次包括目标检测模型、目标特征编码模型、特征分类模型、关系特征编码模型以及关系分类模型。所述的特征分类模型的作用在于对目标特征编码模型所输出的目标特征编码进行分类,从而进一步对目标检测模型所输出的目标检测区域进行类别的识别。所述的关系分类模型的作用在于,对所输入的两个关系特征编码进行关系类别的输出,以表征这两个关系特征编码之间的关系,或者也可以理解为,表征两个目标检测区域之间的关系。

[0190] S44,基于第二学习率对所述初始视觉关系检测模型进行训练,调整特征编码模型的参数,以得到视觉关系检测模型。

[0191] 其中,所述第二学习率小于训练所述特征编码模型的学习率。

[0192] 具体地,S44可以包括以下步骤:

[0193] S441,计算特征编码模型的损失函数的值。

[0194] 如图10所示,该特征编码模型包括目标特征编码模型以及关系特征编码模型,那么该特征编码模型的损失函数的值为两个编码模型的损失函数之和。具体的损失函数的计算方法请参见图6所示实施例的S25,在此不再赘述。

[0195] S442,利用损失函数的值,计算第二梯度估计。

[0196] 电子设备在对该初始视觉关系检测模型进行训练时,可以采用随机梯度下降方法计算第二梯度估计,可以参见图6所示实施例的S254。

[0197] S443,利用第二梯度估计以及第二学习率,对特征编码模型中的转换矩阵进行更新。

[0198] 详细请参见图6所示实施例的S254,在此不再赘述。

[0199] 作为本实施例的一种可选实施方式,可以采用如下方式训练视觉关系检测模型:基于VGGNet的Faster RCNN框架作为目标检测模型,该目标检测模型的输入图片大小为592*592,每幅图片选取得分较高的前64个目标检测框。采用随机梯度下降算法SGD训练该目标检测模型,学习率为0.001,批处理大小为6,在3块NVIDIA GeForce GTX 1080Ti GPU上并行训练50轮。目标检测模型训练结束后,此时固定目标检测模型的参数,对真值检测框提取区域特征,输入上述目标特征编码模型以及关系特征编码模型,其中由于这两个模型均是渐进知识驱动的特征变换模型,因此均可以称之为PKT。其中,对于目标特征编码模型而言,可以称之为$PKT^{obj}$;对于关系特征编码模型而言,可以称之为$PKT^{rel}$。这部分训练也是使用SGD算法,学习率为0.001,批处理大小为6,两个PKT层数均为4,多头注意力数为4,训练至模型参数不再更新为止,一般不多于15轮。此阶段使用真值检测框而不使用目标检测模型预测的检测框,是因为目标检测模型的预测中会包含一些预测错误的情况,它们的区域特

征和目标类别都不是很准确，用来训练两个PKT会引入较多的数据噪声，从而影响PKT模型的关系编码的训练。训练结束后，我们使用更小的学习率0.0001在目标检测模型预测的目标框上微调PKT模型的参数，训练至模型参数不再更新为止，一般不多于15轮。

[0200] 本实施例提供的视觉关系检测模型的训练方法，通过采用小于第一学习率的第二学习率对目标特征编码模型以及关系特征编码模型中的转换矩阵进行微调，一方面能够保证转换矩阵的准确性，另一方面由于采用了较小的学习率能够保证较高的训练效率。

[0201] 作为本实施例的一种可选实施方式，所述特征编码模型为级联的所述目标特征编码模型与所述关系特征编码模型；其中，所述特征编码模型通过特征分类模型与所述关系特征编码模型级联，所述关系特征编码模型通过关系分类模型与所述输出连接。

[0202] 进一步可选地，所述特征分类模型为第一全连接层，所述关系分类模型为第二全连接层。

[0203] 作为本实施例的一个具体应用实施例，图11示出了视觉关系检测模型的示意图，该视觉关系检测模型依次包括目标检测模型，用于预测输入图像的目标检测区域；目标特征编码模型$PKT^{obj}$，用于在目标检测模型输出的基础上，引入引导图，以输出带有关系感知的目标特征；连接$PKT^{obj}$输出的分类器是对$PKT^{obj}$输出的目标特征进行分类，以预测其类别；关系特征编码模型$PKT^{rel}$，用于在所输入的关系感知的目标特征以及目标类别编码的基础上再次引入引导图，以输出关系特征；连接$PKT^{rel}$输出的分类器是将两两联合的关系特征进行分类，以确定两个目标检测区域的关系。

[0204] 根据本发明实施例，提供了一种特征编码模型的训练方法实施例，需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0205] 在本实施例中提供了一种视觉关系检测方法，可用于上述的移动终端，如手机、平板电脑等，图12是根据本发明实施例的视觉关系检测方法的流程图，如图12所示，该流程包括如下步骤：

[0206] S51，获取待检测图像。

[0207] S52，将待检测图像输入视觉关系检测模型中，以得到待检测图像的视觉关系。

[0208] 其中，所述视觉关系检测模型是根据本发明第三方面，或第三方面任一项实施方式中所述的视觉关系检测模型的训练方法训练得到的。

[0209] 详细请参见图8或图9所示实施例的详细描述，在此不再赘述。

[0210] 本发明实施例提供的视觉关系检测方法，由于在视觉关系检测模型中引入视觉常识，利用视觉常识中与该类别相关的引导图一方面弥补了样本数据不足的缺陷，使得在对目标特征进行再次编码时能够有足够的样本数据支撑，另一方面保证了特征编码时就已经引入关系感知，提高了视觉关系检测的准确率。

[0211] 在本实施例中提供了一种视觉关系检测方法，可用于上述的移动终端，如手机、平板电脑等，图13是根据本发明实施例的视觉关系检测方法的流程图，如图13所示，该流程包括如下步骤：

[0212] S61，获取待检测图像。

[0213] S62，将待检测图像输入视觉关系检测模型中，以得到待检测图像的视觉关系。

[0214] 其中，所述视觉关系检测模型是根据本发明第三方面，或第三发面任一项实施方式中所述的视觉关系检测模型的训练方法训练得到的。

[0215] 下文将对该步骤进行详细描述，同时请结合图11所示的视觉关系检测模型。首先对一幅图I中的视觉关系定义一个结构化的图表示，其中包含一组目标候选区域$B=\{b_1,...,b_n\}$，以及其对应的目标类别$O=\{o_1,...,o_n\}$，和目标与目标间的关系$R=\{r_1,...,r_m\}$。如图11所示，将该视觉关系检测模型分解为三个模型：目标检测模型$P(B|I)$，目标分类模型$P(O|B,I)$，关系分类模型$P(R|O,B,I)$。

[0216] 具体地，该步骤包括：

[0217] S621，将待检测图像输入目标检测模型中，输出至少一个目标候选区域、每个目标候选区域的特征向量以及类别概率向量。

[0218] 电子设备将S61中所获取到的待检测图像输入目标检测模型中，该目标检测模型预测出待检测图像中的目标候选区域、各个目标候选区域的特征向量以及类别概率向量。

[0219] 例如，使用传统的目标检测器(如，Faster-RCNN)进行目标区域检测。对于一幅输入图片I，目标检测器Faster-RCNN的输出结果为一组目标候选区域$B=\{b_1,...,b_n\}$，并且对每个$b_i \in B$输出一个特征向量$f_i$和一个类别概率向量$l_i$。

[0220] S622，基于特征向量以及类别概率向量，利用目标特征编码模型得到目标候选区域的目标特征编码。

[0221] 电子设备利用目标特征编码模型在输入的基础上得到带有关系感知的目标特征，其中，关于目标特征编码模型的具体结构细节请参见图8所示实施例的S32，在此不再赘述。关于该步骤的具体实现细节，在下文中将进行详细描述。

[0222] S623，将目标特征编码输入特征分类模型中，以得到对应的目标类别向量。

[0223] 带有关系感知的目标特征输入特征分类模型中，利用特征分类模型预测输入的目标特征预测其类别。

[0224] S624，基于目标特征编码以及目标类别向量，利用关系特征编码模型得到目标候选区域的关系特征编码。

[0225] 将带有关系感知的目标特征与目标类别编码联合，输入关系特征编码模型中，以输出目标候选区域的关系特征编码。关于该步骤的具体实现细节，在下文中将进行详细描述。

[0226] S625，将所有目标候选区域对应的关系特征编码两两联合，输入关系分类模型中，以得到任意两个目标候选区域的视觉关系。

[0227] 电子设备将S624中得到的关系特征编码两两联合输入关系分类模型中，以以得到任意两个目标候选区域的视觉关系。

[0228] 本发明实施例提供的视觉关系检测方法，通过目标特征编码模型输出对每个目标候选区域的带有关系感知的特征编码，在通过特征分类模型，进一步预测出更加准确的目标类别；且关系特征编码模型输出同样是输出带有关系感知的特征编码，进一步提高了预测出的关系类别的准确性。

[0229] 作为本实施例的一种可选实施方式，如图14所示，上述S622可以包括如下步骤：

[0230] S6221，提取每个类别概率向量中概率最大的类别对应的第一词向量。

[0231] 在S621中得到每个目标候选区域对应的类别概率向量之后，从该类别概率向量中

确定出概率最大的类别,并提取出该类别对应的第一词向量。其中,词向量可以是事先获得的对应于所有类别的词向量编码。

[0232] S6222,对于每个目标候选区域,联合特征向量、第一词向量以及目标候选区域的全图特征向量,以得到每个目标候选区域的第一联合特征向量。

[0233] 对于每个目标候选区域而言,将S621中得到的目标候选区域的特征向量、S6221中的第一词向量以及该目标候选区域的全图特征向量进行联合,以得到对应于每个目标候选区域的第一联合特征向量。其中,所述的全图特征向量为该目标候选区域在整个待检测图像中的特征。

[0234] 例如,$PKT^{obj}$的输入为每个目标候选框的联合特征{x;fi;yi},其中x为图片全图特征,fi是目标检测框bi的区域特征向量,yi是目标检测框bi的初始预测类别的词向量编码,初始预测类别是目标检测模型输出的类别概率向量li中得分最高的类别。

[0235] S6223,依次将每个第一联合特征向量输入目标特征编码模型中,以得到每个目标候选区域的目标特征编码。

[0236] 电子设备将与目标候选区域对应的第一联合特征向量输入至目标特征编码模型中,从而对应于每个目标候选区域均可得到一个目标特征编码。

[0237] $PKT^{obj}$首先根据图片I和目标检测框B的特征来学习得到注意力矩阵,注意力矩阵大小为n×n,矩阵中每一个元素代表一个目标区域对其他所有n个目标区域的注意力,即区域间的关联关系。接着,PKT然后再根据注意力矩阵中反映的区域与区域间的关系来对每个目标区域进行特征编码,最终$PKT^{obj}$输出对每个目标区域的带有关系感知的特征编码。

[0238] 作为本实施例的另一可选实施方式,如图15所示,上述S624可以包括如下步骤:

[0239] S6241,提取每个目标候选区域的目标类别向量中得分值最高的类别的第二词向量。

[0240] 在S623中得到每个目标候选区域对应的目标类别向量之后,从该目标类别向量中确定出得分值最高的类别,并提取出该类别对应的第二词向量。

[0241] S6242,对于每个目标候选区域,联合目标特征编码向量以及第二词向量,以得到每个目标候选区域的第二联合特征向量。

[0242] 对于每个目标候选区域而言,将S623中得到的目标候选区域的目标特征编码以及S6241中的第二词向量进行联合,以得到对应于每个目标候选区域的第二联合特征向量。

[0243] S6243,依次将每个第二联合特征向量输入关系特征编码模型中,以得到每个目标候选区域的关系特征编码。

[0244] 电子设备将与目标候选区域对应的第二联合特征向量输入至关系特征编码模型中,从而对应于每个目标候选区域均可得到一个关系特征编码。

[0245] 在本发明实施例中提出的视觉关系检测方法进行评测,具体地本方法评测使用2个评测指标,3个数据集和4个视觉关系预测相关的评估任务。

[0246] 评测指标:召回率(R@K)是计算每个样本中和真值标注匹配成功的视觉关系占标注的视觉关系的比例,匹配条件为:两个目标框和真值目标框的交并比(IoU)大于0.5,且目标类别预测正确,目标间关系预测正确。因为召回率统计视觉关系的匹配数时并没有考虑视觉关系的类别,所以召回率指标在关系类别不均衡的情况下会被常见类主导而忽略不常见类的识别。因此,我们也使用类平均召回率(mR@K)作为评测指标,根据关系类别计算匹配

数,再在类间取平均值。

[0247] 数据集:我们使用的数据集信息如图16所示。表中展示了我们所使用的三个数据集VG,VG-MSDN,VG-DR-Net训练集和测试集的图片数(#Img),关系数(#Rel),平均每幅图中的关系数(Ratio),以及目标类别数(#ObjCls)和关系类别数(#RelCls)。其中因为VG数据集中标注偏差较大,所以我们引入VG数据集的两个清理后的版本,即VG-MSDN和VG-DR-Net。

[0248] 视觉关系评估任务:1)PredCls:对于一幅输入图片,在已知图中所有的真实目标框和目标类别的情况下,预测目标与目标间关系;2)SGCls:对于一幅输入图片,在已知所有真实目标框的情况下,预测目标类别和目标间关系;3)SGGen:对于一幅输入图片,预测目标框、目标类别和目标间关系;4)PhrDet:对于一幅输入图片,预测一个对两个目标的最大包围框、目标类别和目标间关系。

[0249] 在VG和VG-MSDN上基于3个视觉任务(SGGen,SGCls,PredCls)评测R@50,mR@50,R@100,mR@100,如图17所示。在VG-MSDN和VG-DR-Net上评测PhrDet,如图18所示。

[0250] 从实验结果来看,本方法在多个视觉关系检测任务上获得了和目前最好方法MOTIF接近或更高的效果。此外,我们评估了本方法提出的模型的平均每轮的训练时间(Time)和模型参数(Params),如图19所示。可以看出,本方法相比于MOTIFS不仅取得了更优的性能,而且模型更加轻量,训练时间大大缩短。

[0251] 在本实施例中还提供了一种特征编码模型的训练装置、一种视觉关系检测模型的训练装置以及一种视觉关系检测装置,该装置用于实现上述对应的实施例及优选实施方式,已经进行过说明的不再赘述。如以下所使用的,术语“模块”可以实现预定功能的软件和/或硬件的组合。尽管以下实施例所描述的装置较佳地以软件来实现,但是硬件,或者软件和硬件的组合的实现也是可能并被构想的。

[0252] 本实施例提供一种特征编码模型的训练装置,如图20所示,包括:

[0253] 第一获取模块2001,用于获取初始特征编码模型;其中,所述初始特征编码模型包括级联至少一层的多头注意力模块,每个所述多头注意力模块的参数包括一组互相独立的转换矩阵。

[0254] 第二获取模块2002,用于获取样本数据;其中,每个所述样本数据包括样本图像中目标区域的目标特征以及对应的类别。

[0255] 第一输入模块2003,用于将每个所述样本数据输入所述初始特征编码模型中。

[0256] 提取模块2004,用于基于所述类别从视觉常识数据中提取引导图;其中,所述引导图用于表示视觉常识对应于所述类别的目标类别。

[0257] 第一训练模块2005,用于根据所述引导图对所述初始特征编码模型进行训练,调整所述转换矩阵,以对每个所述目标区域的目标特征进行更新,得到每个所述目标区域的目标特征编码。

[0258] 本实施例还提供一种视觉关系检测模型的训练装置,如图21所示,包括:

[0259] 第三获取模块2101,用于获取目标检测模型;所述目标检测模型用于检测第二样本图像中的目标候选区域、每个所述目标候选区域的目标特征及其对应的类别。

[0260] 第四获取模块2102,用于获取特征编码模型;其中,所述特征编码模型是根据本发明第一方面,或第一方面任一项实施方式中所述的特征编码模型的训练方法训练得到;所述特征编码模型包括目标特征编码模型,和/或,关系特征编码模型;所述目标特征编码模

型的输入包括所述目标候选区域的目标特征以及所述类别对应的词向量，输出为所述目标候选区域的目标特征编码；所述关系特征编码模型的输入包括所述目标候选区域的目标特征编码，以及所述目标特征编码的类别对应的词向量，输出为所述目标候选区域的关系特征编码。

[0261] 级联模块2103，用于将所述目标检测模型以及所述特征编码模型级联，以得到初始视觉关系检测模型；其中，所述特征编码模型通过分类模型与输出连接。

[0262] 第二训练模块2104，用于基于第二学习率对所述初始视觉关系检测模型进行训练，调整所述特征编码模型的参数，以得到视觉关系检测模型；其中，所述第二学习率小于训练所述特征编码模型的学习率。

[0263] 本实施例还提供一种视觉关系检测装置，如图22所示，包括：

[0264] 第五获取模块2201，用于获取待检测图像。

[0265] 检测模块2202，用于将所述待检测图像输入视觉关系检测模型中，以得到所述待检测图像的视觉关系；其中，所述视觉关系检测模型是根据本发明第二方面，或第二方面任一项实施方式中所述的视觉关系检测模型的训练方法训练得到的。

[0266] 本发明实施例中的特征编码模型的训练装置、视觉关系检测模型的训练装置以及视觉关系检测装置是以功能单元的形式来呈现，这里的单元是指ASIC电路，执行一个或多个软件或固定程序的处理器和存储器，和/或其他可以提供上述功能的器件。

[0267] 上述各个模块的更进一步的功能描述与上述对应实施例相同，在此不再赘述。

[0268] 本发明实施例还提供一种电子设备，具有上述图20-图22所示的至少一种装置。

[0269] 请参阅图23，图23是本发明可选实施例提供的一种电子设备的结构示意图，如图23所示，该电子设备可以包括：至少一个处理器2301，例如CPU(Central Processing Unit，中央处理器)，至少一个通信接口2303，存储器2304，至少一个通信总线2302。其中，通信总线2302用于实现这些组件之间的连接通信。其中，通信接口2303可以包括显示屏(Display)、键盘(Keyboard)，可选通信接口2303还可以包括标准的有线接口、无线接口。存储器2304可以是高速RAM存储器(Random Access Memory，易挥发性随机存取存储器)，也可以是非不稳定的存储器(non-volatile memory)，例如至少一个磁盘存储器。存储器2304可选的还可以是至少一个位于远离前述处理器2301的存储装置。其中处理器2301可以结合图20-图22所描述的装置，存储器2304中存储应用程序，且处理器2301调用存储器2304中存储的程序代码，以用于执行上述任一方法步骤。

[0270] 其中，通信总线2302可以是外设部件互连标准(peripheral component interconnect，简称PCI)总线或扩展工业标准结构(extended industry standard architecture，简称EISA)总线等。通信总线2302可以分为地址总线、数据总线、控制总线等。为便于表示，图23中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0271] 其中，存储器2304可以包括易失性存储器(英文：volatile memory)，例如随机存取存储器(英文：random-access memory，缩写：RAM)；存储器也可以包括非易失性存储器(英文：non-volatile memory)，例如快闪存储器(英文：flash memory)，硬盘(英文：hard disk drive，缩写：HDD)或固态硬盘(英文：solid-state drive，缩写：SSD)；存储器2304还可以包括上述种类的存储器的组合。

[0272] 其中，处理器2301可以是中央处理器(英文：central processing unit，缩写：

CPU),网络处理器(英文:network processor,缩写:NP)或者CPU和NP的组合。

[0273] 其中,处理器2301还可以进一步包括硬件芯片。上述硬件芯片可以是专用集成电路(英文:application-specific integrated circuit,缩写:ASIC),可编程逻辑器件(英文:programmable logic device,缩写:PLD)或其组合。上述PLD可以是复杂可编程逻辑器件(英文:complex programmable logic device,缩写:CPLD),现场可编程逻辑门阵列(英文:field-programmable gate array,缩写:FPGA),通用阵列逻辑(英文:generic array logic,缩写:GAL)或其任意组合。

[0274] 可选地,存储器2304还用于存储程序指令。处理器2301可以调用程序指令,实现如本申请图1以及图6实施例中所示的特征编码模型的训练方法,或本申请图8以及图9实施例中所示的视觉关系检测模型模型的训练方法,或本申请图12-图15实施例中所示的视觉关系检测方法。

[0275] 本发明实施例还提供了一种非暂态计算机存储介质,所述计算机存储介质存储有计算机可执行指令,该计算机可执行指令可执行上述任意方法实施例中的特征编码模型的训练方法、视觉关系检测模型的训练方法或视觉关系检测方法。其中,所述存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory,ROM)、随机存储记忆体(Random Access Memory,RAM)、快闪存储器(Flash Memory)、硬盘(Hard Disk Drive,缩写:HDD)或固态硬盘(Solid-State Drive,SSD)等;所述存储介质还可以包括上述种类的存储器的组合。

[0276] 虽然结合附图描述了本发明的实施例,但是本领域技术人员可以在不脱离本发明的精神和范围的情况下做出各种修改和变型,这样的修改和变型均落入由所附权利要求所限定的范围之内。

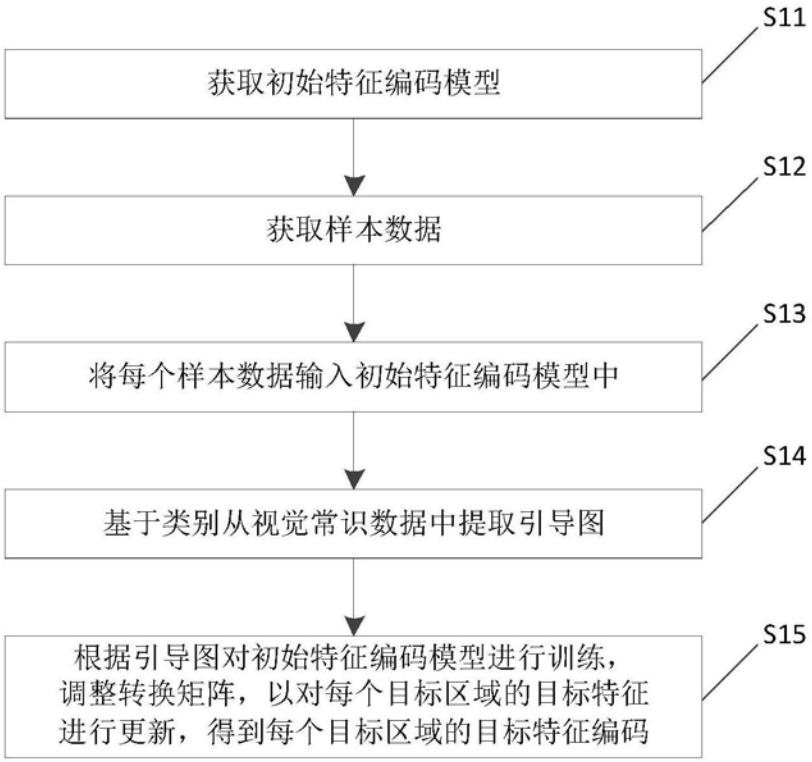

S11
获取初始特征编码模型
S12
获取样本数据
S13
将每个样本数据输入初始特征编码模型中
S14
基于类别从视觉常识数据中提取引导图
S15
根据引导图对初始特征编码模型进行训练，调整转换矩阵，以对每个目标区域的目标特征进行更新，得到每个目标区域的目标特征编码

图1

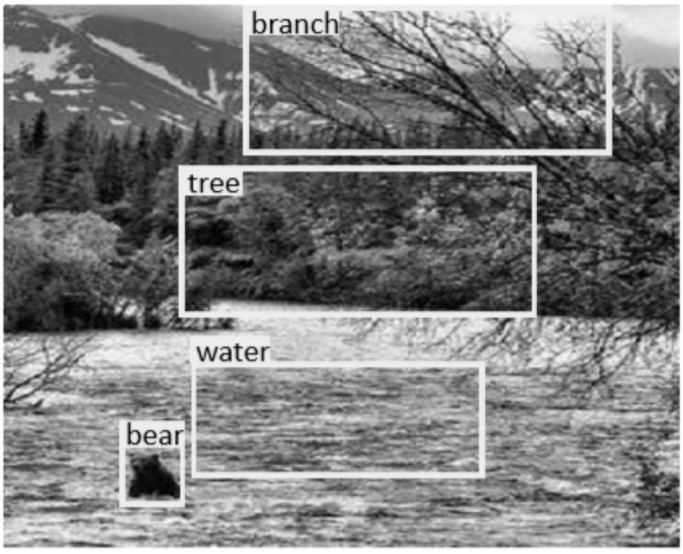

branch
tree
water
bear

图2

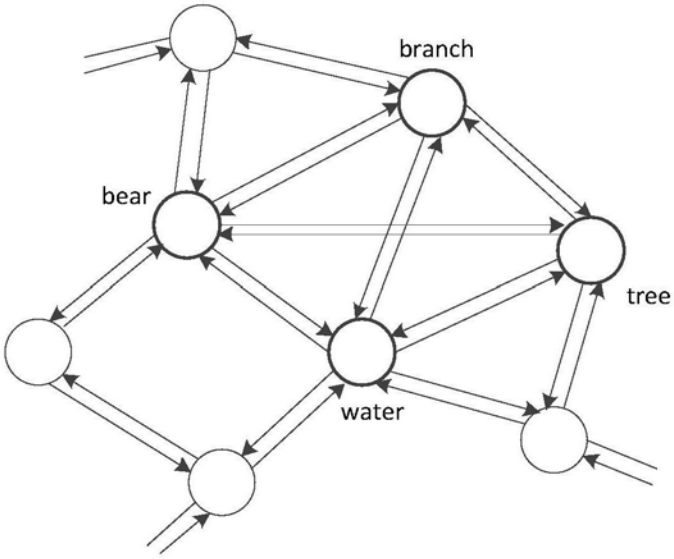

branch
bear
tree
water

图3

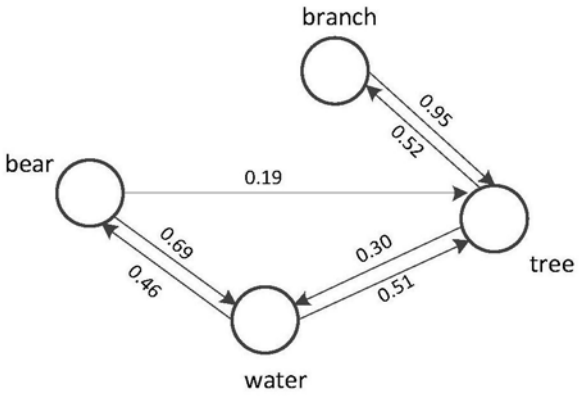

branch
0.95
0.52
bear
0.19
0.69
0.46
0.30
0.51
tree
water

图4

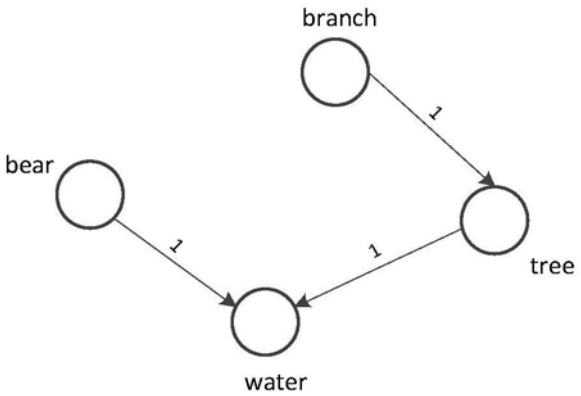

branch
1
bear
1
1
tree
water

图5

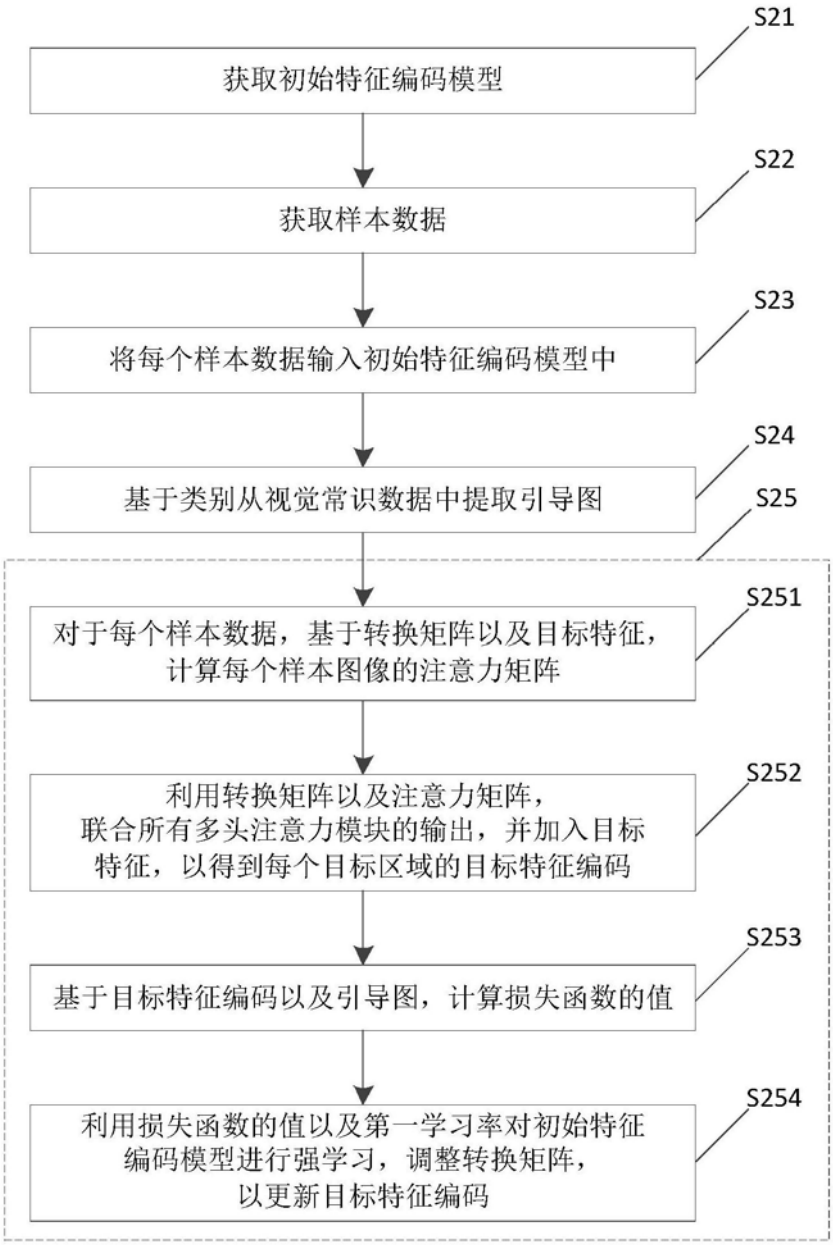

S21
获取初始特征编码模型
S22
获取样本数据
S23
将每个样本数据输入初始特征编码模型中
S24
基于类别从视觉常识数据中提取引导图
S25
S251
对于每个样本数据，基于转换矩阵以及目标特征，
计算每个样本图像的注意力矩阵
S252
利用转换矩阵以及注意力矩阵，
联合所有多头注意力模块的输出，并加入目标
特征，以得到每个目标区域的目标特征编码
S253
基于目标特征编码以及引导图，计算损失函数的值
S254
利用损失函数的值以及第一学习率对初始特征
编码模型进行强学习，调整转换矩阵，
以更新目标特征编码

图6

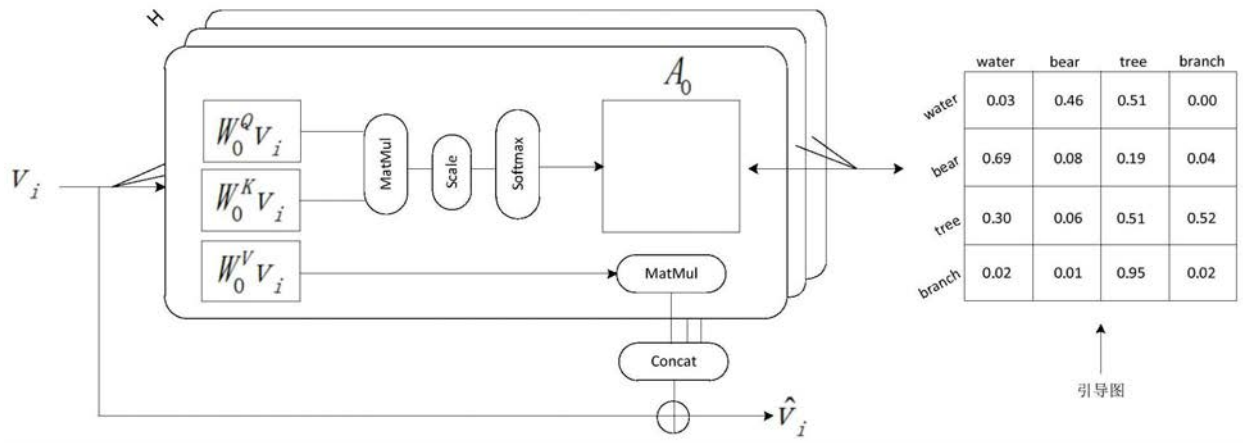

H
$V_i$
$W_0^Q V_i$
$W_0^K V_i$
$W_0^V V_i$
MatMul
Scale
Softmax
$A_0$
MatMul
Concat
$\hat{V}_i$
water bear tree branch
water 0.03 0.46 0.51 0.00
bear 0.69 0.08 0.19 0.04
tree 0.30 0.06 0.51 0.52
branch 0.02 0.01 0.95 0.02
引导图

图7

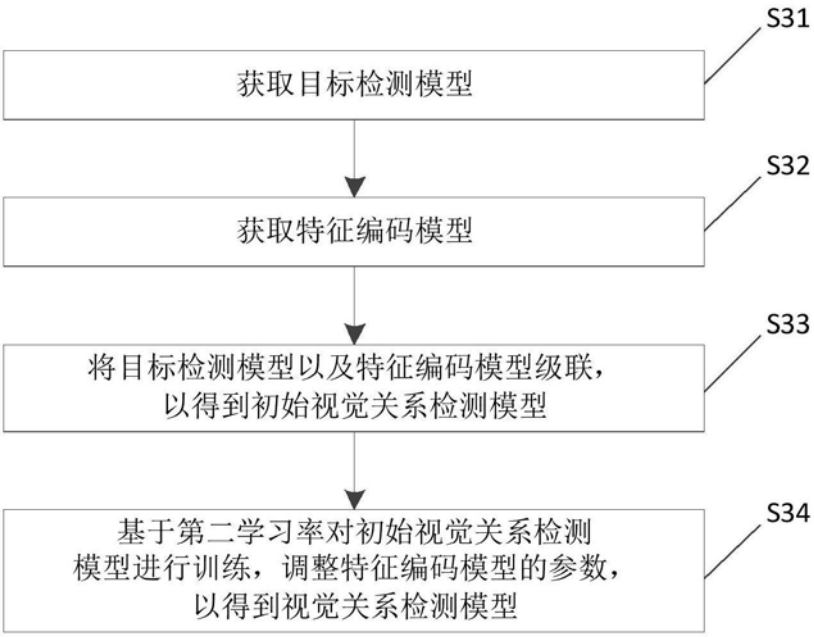

S31
获取目标检测模型
S32
获取特征编码模型
S33
将目标检测模型以及特征编码模型级联，
以得到初始视觉关系检测模型
S34
基于第二学习率对初始视觉关系检测
模型进行训练，调整特征编码模型的参数，
以得到视觉关系检测模型

图8

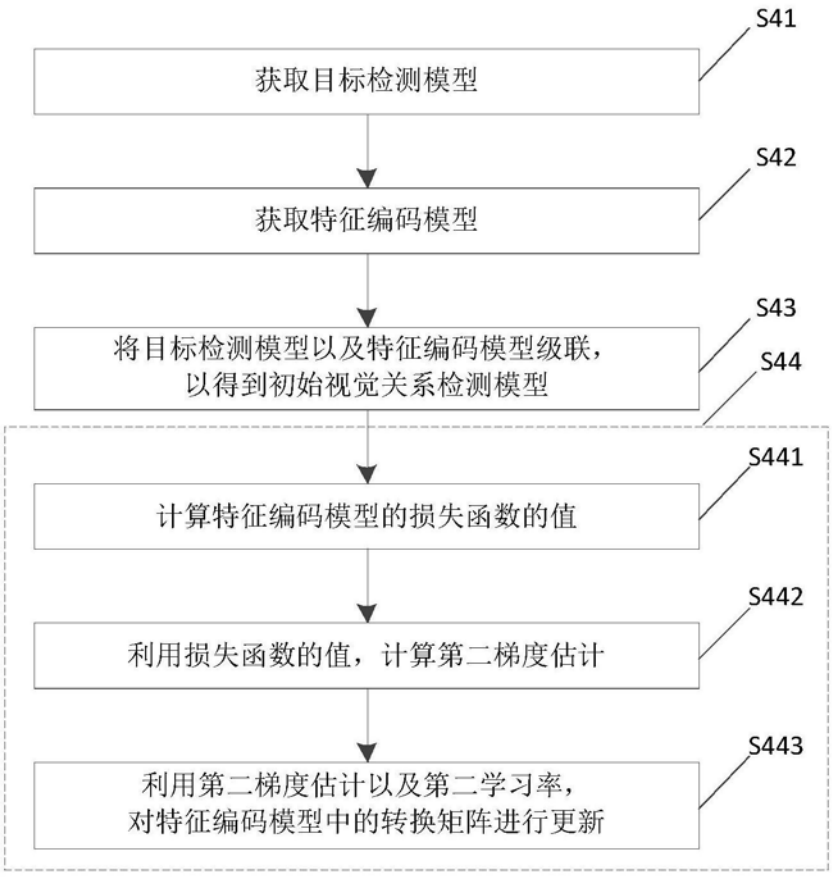

S41
获取目标检测模型
S42
获取特征编码模型
S43
将目标检测模型以及特征编码模型级联，
以得到初始视觉关系检测模型
S44
S441
计算特征编码模型的损失函数的值
S442
利用损失函数的值，计算第二梯度估计
S443
利用第二梯度估计以及第二学习率，
对特征编码模型中的转换矩阵进行更新

图9

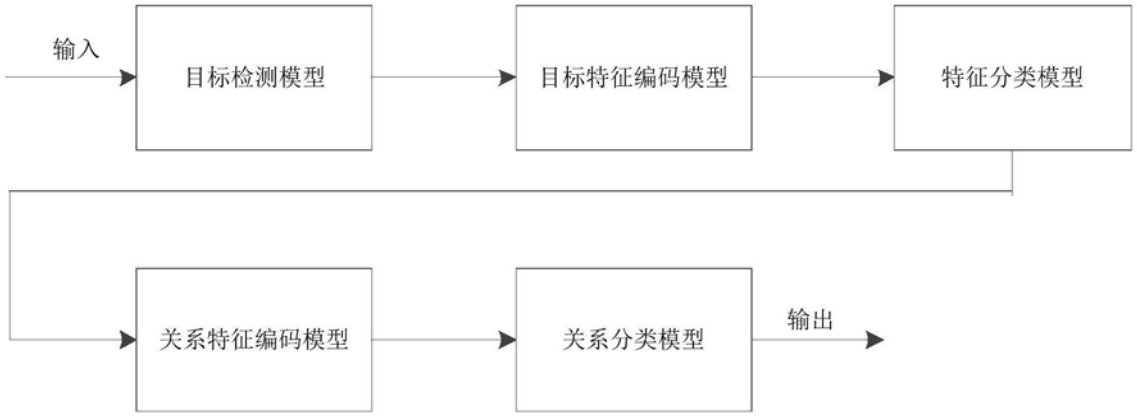

输入
目标检测模型
目标特征编码模型
特征分类模型
关系特征编码模型
关系分类模型
输出

图10

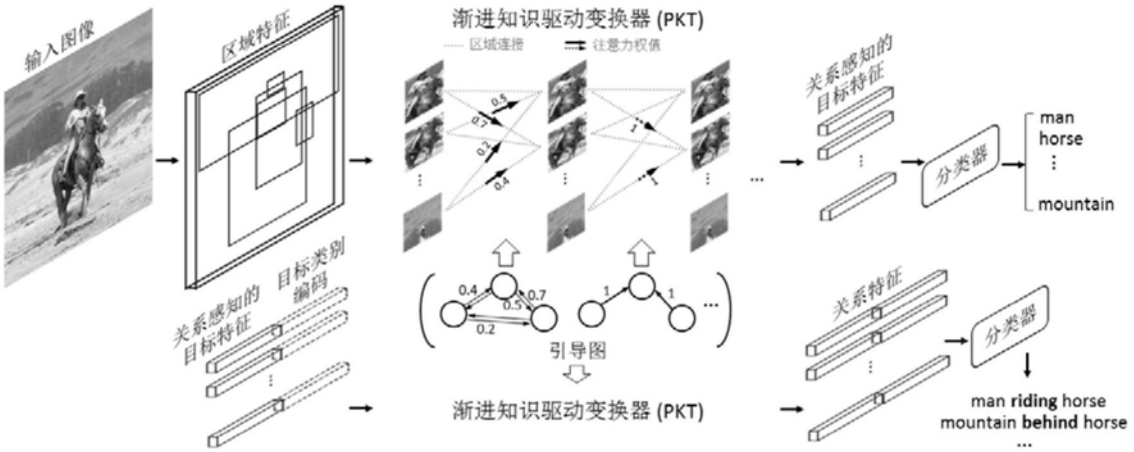

输入图像
区域特征
渐进知识驱动变换器 (PKT)
区域连接
注意力权值
关系感知的目标特征
分类器
man
horse
mountain
关系感知的目标特征
目标类别编码
引导图
渐进知识驱动变换器 (PKT)
关系特征
分类器
man riding horse
mountain behind horse

图11

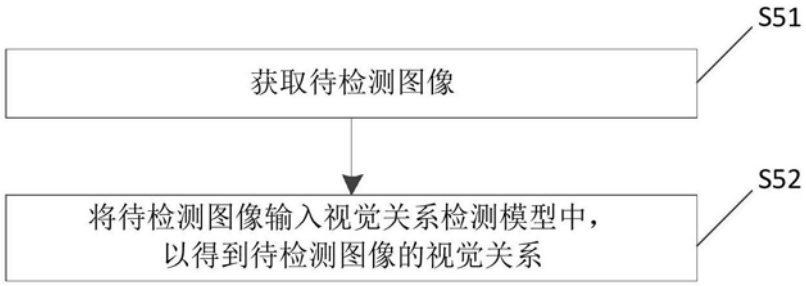

S51
获取待检测图像
S52
将待检测图像输入视觉关系检测模型中，
以得到待检测图像的视觉关系

图12

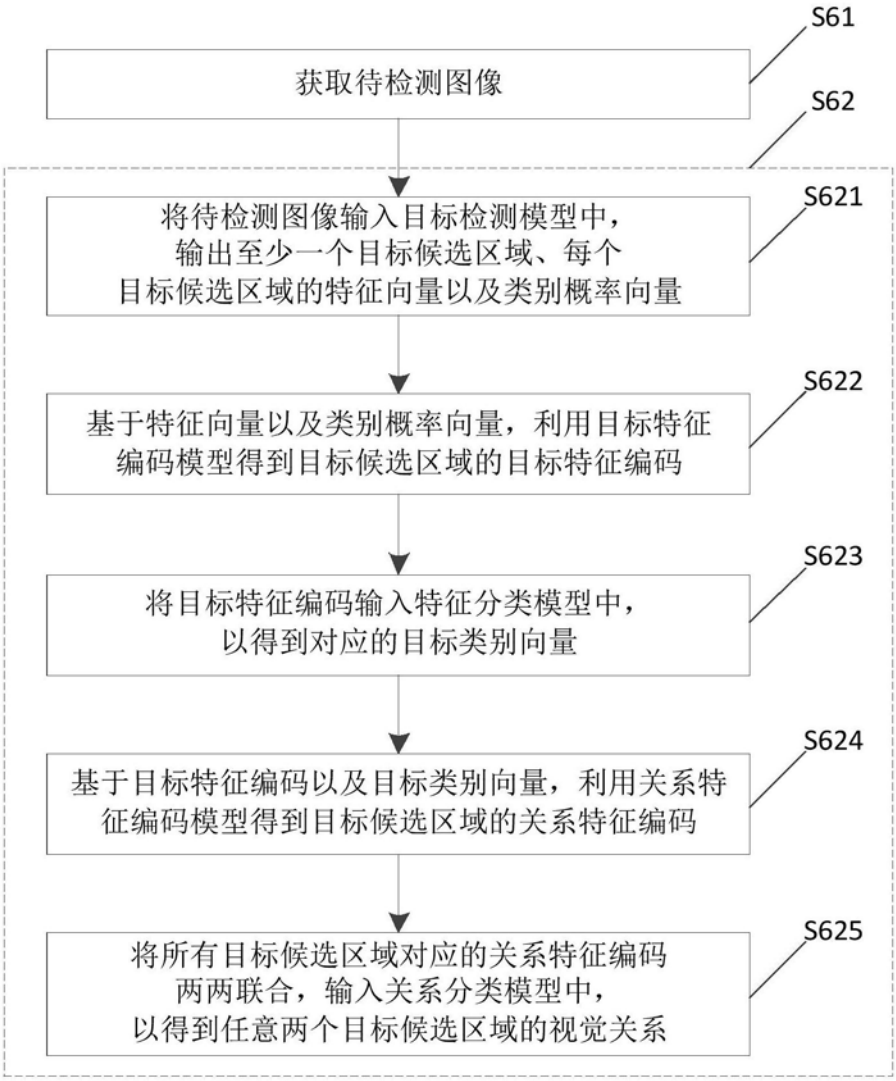

S61
获取待检测图像
S62
S621
将待检测图像输入目标检测模型中，
输出至少一个目标候选区域、每个
目标候选区域的特征向量以及类别概率向量
S622
基于特征向量以及类别概率向量，利用目标特征
编码模型得到目标候选区域的目标特征编码
S623
将目标特征编码输入特征分类模型中，
以得到对应的目标类别向量
S624
基于目标特征编码以及目标类别向量，利用关系特
征编码模型得到目标候选区域的关系特征编码
S625
将所有目标候选区域对应的关系特征编码
两两联合，输入关系分类模型中，
以得到任意两个目标候选区域的视觉关系

图13

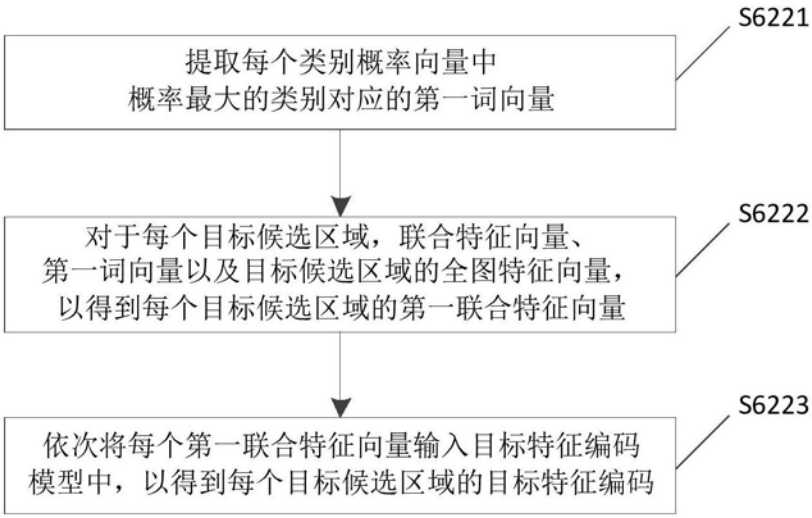



图14

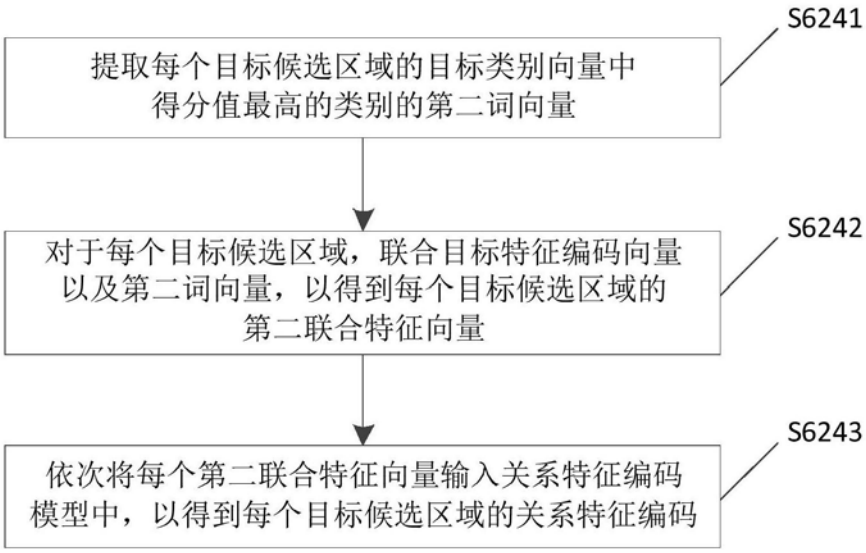



图15

| Dataset | Training Set | | | Testing Set | | | #ObjCls | #RelCls |
|---|---|---|---|---|---|---|---|---|
| | #Img | #Rel | Ratio | #Img | #Rel | Ratio | | |
| VG [8] | 62,723 | 439,063 | 1:7 | 26,446 | 183,642 | 1:7 | 150 | 50 |
| VG-MSDN [8, 10] | 46,164 | 507,296 | 1:11 | 10,000 | 111,396 | 1:11 | 150 | 50 |
| VG-DR-Net [3, 8] | 67,086 | 798,906 | 1:12 | 8,995 | 26,499 | 1:3 | 399 | 24 |

图16

| Dataset | Method | SGGen | | SGCls | | PredCls | |
|---|---|---|---|---|---|---|---|
| | | R/mR@50 | R/mR@100 | R/mR@50 | R/mR@100 | R/mR@50 | R/mR@100 |
| VG | IMP [22] | 20.7/4.2 | 24.5/5.2 | 34.6/6.8 | 35.4/7.2 | 59.3/11.9 | 61.3/12.9 |
| | MOTIFS [25] | **27.2**/5.7 | **30.3**/6.6 | 35.8/7.7 | 36.5/8.2 | 65.2/14.0 | 67.1/15.3 |
| | PKT(Ours) | 27.0/**6.9** | 30.2/**8.2** | **36.1/9.6** | **36.8/10.3** | **66.3/18.0** | **68.0/19.6** |
| VG-MSDN | IMP [22] | 12.1/2.8 | 14.6/3.6 | 25.9/5.9 | 26.9/6.4 | 53.8/11.4 | 56.6/12.7 |
| | MOTIFS [25] | **17.8**/4.6 | **20.3**/5.4 | 27.6/7.1 | 28.4/7.7 | 61.7/15.4 | 64.2/17.2 |
| | PKT(Ours) | 17.7/**5.0** | **20.3/5.9** | **28.4/8.3** | **29.3/8.9** | **62.6/18.9** | **65.1/20.7** |

图17

| Dataset | **PhrDet** | Methods | | | | | |
|---|---|---|---|---|---|---|---|
| | | ISGG [22] | MSDN [10] | DR-Net [3] | FNet [9] | MOTIFS [25] | PKTs (Ours) |
| VG-MSDN | R@50 | 15.9 | 20.0 | - | 22.8 | 28.3 | **30.1** |
| | R@100 | 19.5 | 24.9 | - | 28.6 | 30.7 | **31.8** |
| VG-DR-Net | R@50 | - | - | 24.0 | 26.9 | 33.4 | **35.1** |
| | R@100 | - | - | 27.6 | 32.6 | 33.7 | **35.3** |

图18

| Dataset | Params (MB) | | | Time (min) | | |
|---|---|---|---|---|---|---|
| | IMP [22] | MOTIFS [25] | PKT (Ours) | IMP [22] | MOTIFS [25] | PKT (Ours) |
| VG | 129 | 160 | 142 | 69/171 | 70/225 | 59/188 |
| VG-MSDN | 129 | 160 | 142 | 51/142 | 54/185 | 45/165 |
| VG-DR-Net | 132 | 163 | 145 | 74/279 | 85/335 | 71/291 |

图19

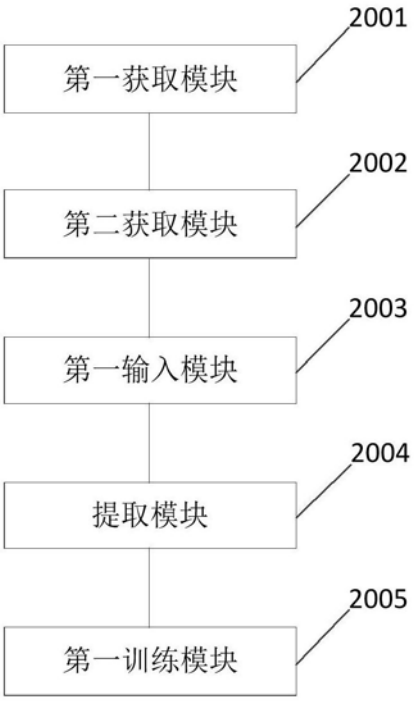

2001
第一获取模块
2002
第二获取模块
2003
第一输入模块
2004
提取模块
2005
第一训练模块

图20

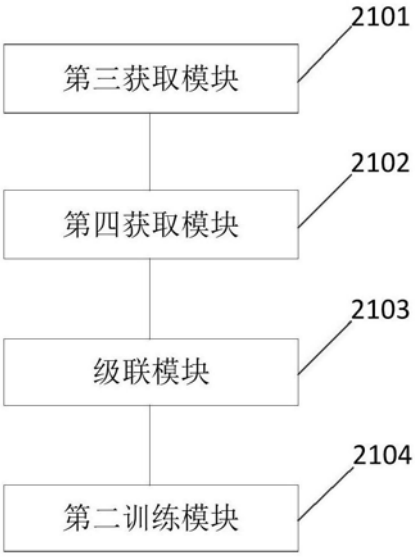

2101
第三获取模块
2102
第四获取模块
2103
级联模块
2104
第二训练模块

图21

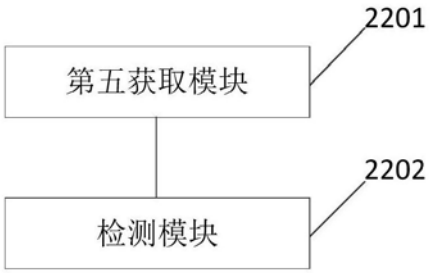

2201
第五获取模块
2202
检测模块

图22

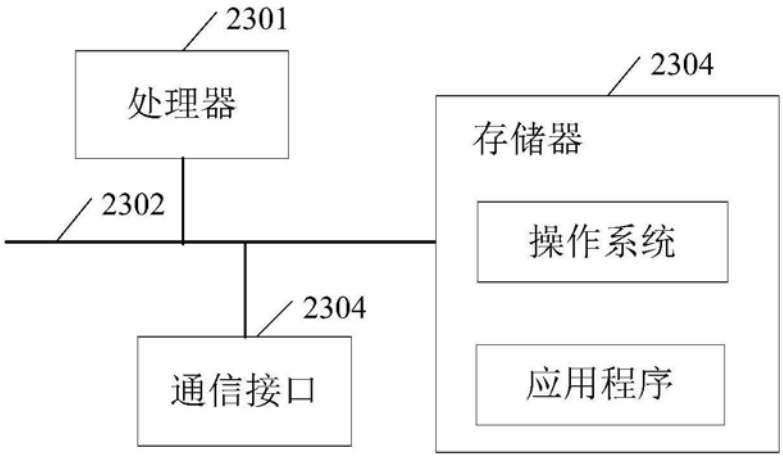

2301
处理器
2304
存储器
2302
操作系统
2304
通信接口
应用程序

图23