

图像的描述语句定位方法及装置、电子设备和存储介质

相关申请的交叉引用

5 本申请基于申请号为 201811459428.7、申请日为 2018 年 11 月 30 日的中国专利申请提出，并要求该中国专利申请的优先权，该中国专利申请的全部内容在此引入本申请作为参考。

技术领域

本申请涉及计算机视觉技术领域但不限于视觉技术领域，尤其涉及一种图像的描述语句定位方法及装置、电子设备和存储介质。

10 背景技术

指代短语定位是计算机视觉和自然语言处理交叉领域的重要问题，例如可要求机器根据给出的一段话（语句）在图像中定位该语句所描述的对象（人或物体等）。相关技术中，提出了由定位模块和关系模块等组成的组合模块化网络，用于识别对象及其关系，然而这些模型可能过度依赖特定的单词或视觉概念，并偏向于经常观察到的证据，导致
15 语句与图像的对应效果较差。

发明内容

本申请提出了一种图像的描述语句定位的技术方案。

根据本申请的一方面，提供了一种图像的描述语句定位方法，包括：对待分析描述语句与待分析图像进行分析处理，获得所述待分析描述语句的多个语句注意力权重以及
20 所述待分析图像的多个图像注意力权重；根据所述多个语句注意力权重以及待分析图像的主体特征、位置特征及关系特征，获得多个第一匹配得分，其中，所述待分析图像中包括多个对象，主体对象为所述多个对象中注意力权重最高的对象，所述主体特征为所述主体对象的特征，所述位置特征为所述多个对象的位置特征，所述关系特征为所述多个对象之间的关系特征；根据所述多个第一匹配得分以及所述多个图像注意力权重，获得
25 所述待分析描述语句与所述待分析图像之间的第二匹配得分；根据所述第二匹配得分，确定所述待分析描述语句在所述待分析图像中的定位结果。

根据本申请的一方面，提供了一种图像的描述语句定位装置，包括：第一权重获得模块，配置为对待分析描述语句与待分析图像进行分析处理，获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重；第一得分获得模
30 块，配置为根据所述多个语句注意力权重以及待分析图像的主体特征、位置特征及关系特征，获得多个第一匹配得分，其中，所述待分析图像中包括多个对象，主体对象为所述多个对象中注意力权重最高的对象，所述主体特征为所述主体对象的特征，所述位置特征为所述多个对象的位置特征，所述关系特征为所述多个对象之间的关系特征；第二得分获得模块，配置为根据所述多个第一匹配得分以及所述多个图像注意力权重，获得
35 所述待分析描述语句与所述待分析图像之间的第二匹配得分；结果确定模块，配置为根据所述第二匹配得分，确定所述待分析描述语句在所述待分析图像中的定位结果。

根据本申请的一方面，提供了一种电子设备，包括：处理器；用于存储处理器可执行指令的存储器；其中，所述处理器被配置为执行上述方法。

根据本申请的一方面，提供了一种计算机可读存储介质，其上存储有计算机程序指令，所述计算机程序指令被处理器执行时实现上述方法。
40

在本申请实施例中，能够得到待分析描述语句的语句注意力权重及待分析图像的图像注意力权重；根据语句注意力权重及图像的主体特征、位置特征及关系特征获得多个第一匹配得分；并根据第一匹配得分及图像注意力权重获得第二匹配得分；根据第二匹

配得分确定定位结果，从而充分发现文本和视觉语义之间的对应关系，提高描述语句在图像中定位的准确度。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，而非限制本申请。根据下面参考附图对示例性实施例的详细说明，本申请的其它特征及方面将变得清楚。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，这些附图示出了符合本申请的实施例，并与说明书一起用于说明本申请的技术方案。

图1示出根据本申请实施例的图像的描述语句定位方法的流程图。

图2示出根据本申请实施例的神经网络的示意图。

图3示出根据本申请实施例的获得第二样本描述语句的示意图。

图4示出根据本申请实施例的获得第二样本图像的示意图。

图5示出根据本申请实施例的图像的描述语句定位装置的框图。

图6示出根据本申请实施例的一种电子设备的框图。

图7示出根据本申请实施例的一种电子设备的框图。

具体实施方式

以下将参考附图详细说明本申请的各种示例性实施例、特征和方面。附图中相同的附图标记表示功能相同或相似的元件。尽管在附图中示出了实施例的各种方面，但是除非特别指出，不必按比例绘制附图。

在这里专用的词“示例性”意为“用作例子、实施例或说明性”。这里作为“示例性”所说明的任何实施例不必解释为优于或好于其它实施例。

本文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本文中术语“至少一种”表示多种中的任意一种或多种中的至少两种的任意组合，例如，包括A、B、C中的至少一种，可以表示包括从A、B和C构成的集合中选择的任意一个或多个元素。

另外，为了更好地说明本申请，在下文的具体实施方式中给出了众多的具体细节。本领域技术人员应当理解，没有某些具体细节，本申请同样可以实施。在一些实例中，对于本领域技术人员熟知的方法、手段、元件和电路未作详细描述，以便于凸显本申请的主旨。

在本发明的各个实施例中，所述图像的描述语句定位方法可以由终端设备或服务器等电子设备执行，终端设备可以为用户设备（User Equipment, UE）、移动设备、用户终端、终端、蜂窝电话、无绳电话、个人数字处理（Personal Digital Assistant, PDA）、手持设备、计算设备、车载设备、可穿戴设备等，所述方法可以通过处理器调用存储器中存储的计算机可读指令的方式来实现。或者，可通过服务器执行所述方法。

图1示出根据本申请实施例的图像的描述语句定位方法的流程图。该方法包括：

在步骤S11中，对待分析描述语句与待分析图像进行分析处理，获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。

在一种可能的实现方式中，待分析图像中可以包括多个对象（人、动物、物体等），例如多个骑着马的人。待分析描述语句可以是针对待分析图像中的某个对象的描述，例如“在中间的被女孩骑着的棕色的马”。待分析图像和待分析描述语句之间可能对应，也可能不对应。可根据本申请实施例的方法来确定语句与图像之间的关联。

在一种可能的实现方式中，待分析描述语句的多个语句注意力权重可包括语句主语权重、语句位置权重及语句关系权重，分别用于表示待分析描述语句的不同类型的分词

对应的注意力权重。

在一种可能的实现方式中，待分析图像的多个图像注意力权重可包括主体对象权重、对象位置权重及对象关系权重，分别用于表示待分析图像的不同类型的图像区域对应的注意力权重。

5 在步骤S12中，根据所述多个语句注意力权重以及待分析图像的主体特征、位置特征及关系特征，获得多个第一匹配得分，其中，所述待分析图像中包括多个对象，主体对象为所述多个对象中注意力权重最高的对象，所述主体特征为所述主体对象的特征，所述位置特征为所述多个对象的位置特征，所述关系特征为所述多个对象之间的关系特征。

10 在一种可能的实现方式中，所述待分析图像中包括多个对象（人、动物、物体等），主体对象为所述多个对象中注意力权重最高的对象。所述主体特征为所述主体对象本身的图像特征，所述位置特征为体现所述多个对象之间的相对位置的位置特征，所述关系特征为体现所述多个对象之间的相对关系的关系特征。

在一种可能的实现方式中，多个第一匹配得分可包括主语匹配得分、位置匹配得分及关系匹配得分。主语匹配得分用于评价待分析图像中的主体对象与待分析描述语句的对象描述之间的匹配程度；位置匹配得分评价待分析图像中多个对象的相对位置与待分析描述语句的位置描述之间的匹配程度；关系匹配得分用于评价待分析图像中多个对象的关联性与待分析描述语句的关联描述之间的匹配程度。

15 在步骤S13中，根据所述多个第一匹配得分以及所述多个图像注意力权重，获得所述待分析描述语句与所述待分析图像之间的第二匹配得分。

在一种可能的实现方式中，根据主语匹配得分、位置匹配得分及关系匹配得分以及主体对象权重、对象位置权重及对象关系权重，可获得所述待分析描述语句与所述待分析图像之间的第二匹配得分。第二匹配得分用于评价待分析图像与待分析描述语句之间的总体匹配程度。

20 在步骤S14中，根据所述第二匹配得分，确定所述待分析描述语句在所述待分析图像中的定位结果。

在一种可能的实现方式中，在得到第二匹配得分后，可以进一步确定待分析描述语句在待分析图像中的定位位置，从而实现描述语句在图像中的定位。

30 根据本申请的实施例，能够得到待分析描述语句的语句注意力权重及待分析图像的图像注意力权重；根据语句注意力权重及图像的主体特征、位置特征及关系特征获得多个第一匹配得分；并根据第一匹配得分及图像注意力权重获得第二匹配得分；根据第二匹配得分确定定位结果，从而充分发现文本和视觉语义之间的对应关系，提高描述语句在图像中定位的准确度。

35 在一种可能的实现方式中，可在步骤S11中对待分析描述语句与待分析图像进行分析处理，获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。其中，步骤S11可包括：

对所述待分析图像进行特征提取，获得所述待分析图像的图像特征向量；

对所述待分析描述语句进行特征提取，获得所述待分析描述语句的多个分词的分词嵌入向量；

40 根据所述图像特征向量以及所述多个分词的分词嵌入向量，获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。

举例来说，可对待分析图像和待分析描述语句分别进行特征提取。对于待分析图像，可对待分析图像的所有像素点进行特征提取，获得待分析图像的图像特征向量 c_0 。本申请对待分析图像的特征提取方式不作限制。

在一种可能的实现方式中,对于待分析描述语句,可对其进行分词处理以确定待分析描述语句的多个分词,并对各个分词进行特征提取,获得多个分词的分词嵌入向量

(word embedding) $\{e_i\}_{i=1}^T$; 其中, T 表示分词的数量(T 为大于1的整数), e_i 表示第 i 个分词嵌入向量, $1 \leq i \leq T$ 。本申请对待分析描述语句的具体分词方式以及对各个分词进行特征提取的具体方式不作限制。

在一种可能的实现方式中,根据确定出的图像特征向量以及所述多个分词的分词嵌入向量,可以确定出所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。

在一种可能的实现方式中,所述方法还可包括:通过神经网络获取所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。该神经网络可包括语言注意力网络,语言注意力网络可通过循环神经网络RNN、长短期记忆网络LSTM等网络实现。可将所述待分析图像与所述待分析描述语句与输入语言注意力网络中处理,获得所述多个语句注意力权重以及所述多个图像注意力权重。

在一种可能的实现方式中,可将所述待分析图像与所述待分析描述语句与输入语言注意力网络中处理,获得所述多个语句注意力权重以及所述多个图像注意力权重。

举例来说,可通过语言注意力网络的特征提取子网络来进行特征提取,分别获得图像特征向量 e_0 以及分词嵌入向量 $\{e_i\}_{i=1}^T$ 。该特征提取子网络可为卷积神经网络CNN(例如Faster CNN)。

在一种可能的实现方式中,语言注意力网络可具有基于注意力机制的LSTM网络。

可将图像特征向量 e_0 作为LSTM网络的第一级输入,并将分词嵌入向量 $\{e_i\}_{i=1}^T$ 作为LSTM网络的各级循环的输入,从而获得LSTM网络的多个隐藏层的输出状态 h_t 。

在一种可能的实现方式中,根据多个状态 h_t ,可计算出图像注意力权重及各个分词的注意力权重;根据多个分词的注意力权重对多个分词嵌入向量 $\{e_i\}_{i=1}^T$ 进行加权求和,可以得到语句注意力权重。

在一种可能的实现方式中,待分析描述语句的多个语句注意力权重为分词级注意力权重(word-level attention weights),可包括语句主语权重 q^{subj} 、语句位置权重 q^{loc} 及语句关系权重 q^{rel} ,分别用于表示待分析描述语句的不同类型的分词对应的注意力权重。

其中,语句主语权重用于表示当注意语句中的主语分词时的注意力权重,例如语句“在中间的被女孩骑着的棕色的马”中为主语的分词“棕色的马”或“马”的注意力权重;语句位置权重用于表示当注意语句中表示位置的分词时的注意力权重,例如上述语句中表示位置的分词“在中间”的注意力权重;语句关系权重用于表示当注意语句中表示对象之间关系的分词时的注意力权重,例如上述语句中表示对象之间关系的分词“被女孩骑着”的注意力权重。

在一种可能的实现方式中,待分析图像的多个图像注意力权重为模块级注意力权重(module-level attention weights),可包括主体对象权重 ω^{subj} 、对象位置权重 ω^{loc} 及对象关系权重 ω^{rel} ,分别用于表示待分析图像的不同类型的图像区域对应的注意力权重。

其中,主体对象权重可表示注意图像中多个对象(人、动物、物体等)中最重要的对象(主体对象)时的注意力权重,例如图像中间的人物;对象位置权重可表示注意图像中多个对象的相对位置时的注意力权重,例如图像的中间、左侧和右侧位置;对象关系权重可表示注意图像中多个对象之间的关联性时的注意力权重,例如图像的中间、左侧和右侧都有人骑着马。

在一些实施例中,图像注意力权重可以根据该对象在图像中的各种图像参数来确

定, 该图像参数包括但不限于: 该对象在图像中的分布位置、该对象在图像中占用的面积、该对象在图像中的主体色。例如, 根据分布位置, 在图像中间的对象可相对于在图像边缘的对象可以获得较高的图像注意力权重。再例如, 在图像中占据面积较大的对象相对于占据面积较小的对象具有较高的对象。再例如, 对象的主体色为被追踪目标的颜色, 相对于其他颜色可以具有较高的图像注意力权重。

在还有一些实施例中, 根据对象在凸显各种的呈现状态确定图像注意力对象; 例如, 以道路监控视频的图像帧分析, 若被追踪的对象为车辆, 则有违规现象的对象具有较高的图像注意力, 例如图像包含的车辆具有压实现的行为特征, 该对象可以配置更高的图像注意力权重。

当然, 以上仅是对图像注意力权重进行举例说明, 具体的图像注意力权重可以根据图像处理需求进行配置; 不局限于上述举例。

通过这种方式, 可通过语言注意力网络捕捉视觉(图像)和文本(语句)中不同类型的信息, 从而发现图像和语句在各个方面的对应关系, 提高了处理精度。

在一种可能的实现方式中, 在步骤S12之前, 所述方法还包括: 将所述待分析图像输入特征提取网络进行处理, 获得所述待分析图像的主体特征、位置特征及关系特征。

举例来说, 特征提取网络可以是预先设定的一个或多个卷积神经网络CNN(例如Faster R-CNN), 用于提取待分析图像的主体特征、位置特征及关系特征。可将待分析图像的所有像素点输入特征提取网络中, 将在ROI池化之前的特征图作为待分析图像的整体图像特征。

在一种可能的实现方式中, 关于主体特征, 可识别出待分析图像中的多个对象, 并提取多个区域中注意力权重最高的对象作为主体对象, 将主体对象的区域特征图确定为主体特征。例如, 提取7x7的特征图作为主体特征。

在一种可能的实现方式中, 关于位置特征, 根据待分析图像中的多个对象所在的图像区域之间的相对位置偏移量及相对区域, 以及对象本身的位置和相对区域, 可以获得位置特征。

在一种可能的实现方式中, 关于关系特征, 根据区域建议中平均值池化的视觉特征、相对位置偏移量及相对区域之间的连接, 可确定上下文对象(多个对象)之间的关系特征。

应当理解, 本申请对提取待分析图像的主体特征、位置特征及关系特征的具体方式不作限制。

在一种可能的实现方式中, 可在步骤S12中根据所述多个语句注意力权重以及待分析图像的主体特征、位置特征及关系特征, 获得多个第一匹配得分。

举例来说, 可通过神经网络获取多个第一匹配得分。所述神经网络可包括图像注意力网络, 所述图像注意力网络包括主体网络、位置网络以及关系网络。其中, 主体网络、位置网络以及关系网络可以分别为预先构建的卷积神经网络CNN。

其中, 主体网络用于评估待分析图像中多个对象(人、动物、物体等)中最重要的对象(主体对象)与待分析描述语句的对象描述之间的匹配程度; 位置网络用于评估待分析图像中多个对象的相对位置与待分析描述语句的位置描述之间的匹配程度; 关系网络用于评估待分析图像中多个对象的关联性与待分析描述语句的关联描述之间的匹配程度。

在一种可能的实现方式中, 可将所述多个语句注意力权重以及待分析图像的主体特征、位置特征及关系特征, 分别输入主体网络、位置网络以及关系网络中处理, 从而评估图像和语句各个方面的匹配程度。

其中, 主体对象为待分析图像的多个对象中注意力权重最高的对象, 主体特征为所

述主体对象的特征,所述位置特征为所述多个对象的位置特征,所述关系特征为所述多个对象之间的关系特征。

在一种可能的实现方式中,步骤S12中得到的多个第一匹配得分可包括主语匹配得分、位置匹配得分及关系匹配得分。

- 5 在一种可能的实现方式中,步骤S12可包括:将语句主语权重及主体特征输入所述主体网络中处理,获得主语匹配得分;将语句位置权重及位置特征输入所述位置网络中处理,获得位置匹配得分;将语句关系权重及关系特征输入所述关系网络中处理,获得关系匹配得分。

- 10 在本实施例中,将语句主语权重及主体特征输入主体网络中,可以分析待分析描述语句的主语与待分析图像的主体对象之间的匹配程度,得到主语匹配得分;将语句位置权重及位置特征输入位置网络中,可以分析待分析描述语句的位置分词与待分析图像的多个对象的相对位置之间的匹配程度,得到位置匹配得分;将语句关系权重及关系特征输入关系网络中,可以分析待分析描述语句的关系分词与待分析图像的多个对象的关联性之间的匹配程度,得到关系匹配得分。

- 15 举例来说,可将多个语句注意力权重(语句主语权重 q^{subj} 、语句位置权重 q^{loc} 及语句关系权重 q^{rel}),以及多个对象特征(主体特征、位置特征、关系特征),分别输入主体网络、位置网络以及关系网络中处理。

通过这种方式,可以确定图像和描述语句各个方面的匹配程度,提高了匹配判断的准确性。

- 20 在一种可能的实现方式中,可在步骤S13中根据所述多个第一匹配得分以及所述多个图像注意力权重,获得所述待分析描述语句与所述待分析图像之间的第二匹配得分。也即,根据主语匹配得分、位置匹配得分及关系匹配得分以及主体对象权重 ω^{subj} 、对象位置权重 ω^{loc} 及对象关系权重 ω^{rel} ,获得所述待分析描述语句与所述待分析图像之间的第二匹配得分。

- 25 其中,步骤S13可包括:

根据所述主体对象权重、所述对象位置权重及所述对象关系权重,对所述主语匹配得分、所述位置匹配得分及所述关系匹配得分进行加权平均,确定所述第二匹配得分。

- 30 举例来说,在得到主语匹配得分、位置匹配得分及关系匹配得分后,可根据主体对象权重 ω^{subj} 、对象位置权重 ω^{loc} 及对象关系权重 ω^{rel} ,分别对主语匹配得分、位置匹配得分及关系匹配得分进行加权,对加权后的得分求和后再求平均值。可将该平均值确定为待分析描述语句与待分析图像之间的第一匹配得分。

通过这种方式,可以得到待分析描述语句与待分析图像之间准确的匹配得分。

- 35 在一种可能的实现方式中,可在步骤S14中根据所述第二匹配得分确定所述待分析描述语句在所述待分析图像中的定位结果。也即,在得到第二匹配得分后,可以进一步确定待分析描述语句在待分析图像中的定位结果。其中,步骤S14可包括:

在所述第二匹配得分大于或等于预设阈值的情况下,将所述主体对象的图像区域确定为所述待分析描述语句的定位位置。

- 40 举例来说,可以预先设定匹配得分的阈值(例如预设阈值为70分),如果第二匹配得分大于或等于该预设阈值,则可认为待分析描述语句是待分析图像中主体对象的描述,可将主体对象所在的图像区域确定为待分析描述语句的定位位置。反之,如果第二匹配得分小于该预设阈值,则可认为待分析描述语句不是待分析图像中主体对象的描述,可将定位结果确定为无法对应。应当理解,本领域技术人员可根据实际情况设定该预设阈值,本申请对预设阈值的具体取值不作限制。

在一种可能的实现方式中,可以在待分析图像中分别设定多个主体对象,将每个主

体对象的主体特征分别输入图像注意力网络中处理, 确定每个主体对象的第二匹配得分, 并可确定多个第二匹配得分中的最高得分。该情况下, 可认为待分析描述语句是对与最高得分对应的主体对象的描述, 可将该主体对象所在的图像区域确定为待分析描述语句的定位位置。

5 通过这种方式, 可以实现待分析描述语句在待分析图像中的精确定位。

图2示出根据本申请实施例的神经网络的示意图。如图2所示, 该神经网络可包括语言注意力网络21和图像注意力网络, 图像注意力网络包括主体网络22、位置网络23以及关系网络24。

10 在该示例中, 将待分析描述语句“在中间的被女孩骑着的棕色的马”201与待分析图像202输入到语言注意力网络21中处理, 可输出三个图像注意力权重(主体对象权重 ω^{subj} 、对象位置权重 ω^{loc} 及对象关系权重 ω^{rel}), 同时输出三个语句注意力权重(语句主语权重 q^{subj} 、语句位置权重 q^{loc} 及语句关系权重 q^{rel})。

在该示例中, 可通过特征提取网络(未示出)得到待分析图像的主体特征203、位置特征204及关系特征205。

15 在该示例中, 将语句主语权重 q^{subj} 及主体特征203输入主体网络22中处理, 可得到主语匹配得分; 将语句位置权重 q^{loc} 及位置特征204输入位置网络23中处理, 可得到位置匹配得分; 将语句关系权重 q^{rel} 及关系特征205输入关系网络24中处理, 可得到关系匹配得分。

20 在该示例中, 根据主体对象权重 ω^{subj} 、对象位置权重 ω^{loc} 及对象关系权重 ω^{rel} , 分别对主语匹配得分、位置匹配得分及关系匹配得分进行加权, 对加权后的得分求和后再平均, 得到第二匹配得分206, 进而根据第二匹配得分206确定待分析描述语句在待分析图像中的定位结果, 从而完成步骤S11-S14的整个实现过程。

应当理解, 以上只是实现本申请的方法的神经网络的一个示例, 本申请对神经网络具体类型不作限制。

25 在一种可能的实现方式中, 在步骤S11之前, 所述方法还包括: 利用样本集训练所述神经网络, 所述样本集包括多个正样本对和多个负样本对。

其中, 每个正样本对包括第一样本图像及其第一样本描述语句,

30 每个负样本对包括第一样本图像及从所述第一样本描述语句中去除分词后的第二样本描述语句, 或第一样本描述语句及从所述第一样本图像中去除区域后的第二样本图像。

在一种可能的实现方式中, 可通过基于注意力引导的跨模态去除方式, 去除高注意力权重的视觉或文本信息, 以获得去除后的训练样本(第二样本描述语句和第二样本图像), 提高训练精度。

35 举例来说, 可预先设定包括多个训练样本的样本集, 以便对神经网络进行训练。样本集中包括多个正样本对, 每个正样本对包括第一样本图像O及其第一样本描述语句Q。可将第一样本图像中的对象进行描述的语句作为同一正样本对中的第一样本描述语句。样本集中还可包括多个负样本对, 每个负样本对包括第一样本图像及从第一样本描述语句中去除分词后的第二样本描述语句, 或第一样本描述语句及从所述第一样本图像中去除区域后的第二样本图像。本申请对样本集的具体建立方式不作限制, 且本申请对
40 每个样本对中的样本图像与样本描述语句之间的前后顺序不作限制。

在一种可能的实现方式中, 所述方法还可包括:

将所述正样本对的第一样本描述语句及第一样本图像输入所述语言注意力网络, 得到所述第一样本描述语句的多个分词的注意力权重;

采用预定标识替换所述第一样本描述语句中注意力权重最高的分词, 获得第二样本

描述语句;

将所述第一样本图像和所述第二样本描述语句作为负样本对。

在一种可能的实现方式中,可以通过语言注意力网络进行空间注意力引导,去除最重要的文本信息而得到较难的文本训练样本,从而避免神经网络过度依赖特定的文本信息(分词),提高训练得到的神经网络的精度

图3示出根据本申请实施例的获得第二样本描述语句的示意图。举例来说,如图3所示,可将正样本对的第一样本描述语句(例如“在中间的被女孩骑着的棕色的马”)及第一样本图像(例如包括多个骑着马的人的图片)输入语言注意力网络,得到第一样本描述语句的多个分词的注意力权重。根据各个分词的注意力权重,可确定注意力权重最高的分词(例如“中间”)。由于直接去除该分词“中间”可能会导致语法错误而无法识别,因此,可采用一个未知标识来替换该分词“中间”,得到第二样本描述语句 Q^* (在“未知”的被女孩骑着的棕色的马),从而可将所述第一样本图像和所述第二样本描述语句作为负样本对。

在一种可能的实现方式中,所述方法还可包括:

将所述正样本对的第一样本描述语句及第一样本图像输入所述图像注意力网络,得到所述第一样本图像的注意力权重;

去除所述第一样本图像中注意力权重最高的图像区域,获得第二样本图像;

将所述第二样本图像和所述第一样本描述语句作为负样本对。

在一种可能的实现方式中,可以通过图像注意力网络识别并去除最重要的视觉信息,得到较难的图像训练样本,从而避免神经网络过度依赖特定的视觉信息,提高训练得到的神经网络的精度。

图4示出根据本申请实施例的获得第二样本图像的示意图。举例来说,如图4所示,可将正样本对的第一样本图像(例如包括多个骑着马的人的图片)和第一样本描述语句(例如“在中间的被女孩骑着的棕色的马”)输入图像注意力网络中处理。其中,可以采用图像注意力网络的主体网络,也可以采用位置网络或关系网络,本申请对此不作限制。

在一种可能的实现方式中,将第一样本图像及第一样本描述语句输入主体网络中,可得到第一样本图像各个区域的注意力权重。根据各个区域的注意力权重,可确定出注意力权重最高的目标区域(例如中间的女孩所在的图像区域)。从第一样本图像中去除该目标区域,可得到第二样本图像 O^* (如图4所示),从而可将所述第二样本图像和所述第一样本描述语句作为负样本对。

在一种可能的实现方式中,利用样本集训练所述神经网络的步骤可包括:根据所述神经网络的第一损失和第二损失,确定所述神经网络的总体损失。

在一种可能的实现方式中,可以获取正样本对(第一样本图像及其第一样本描述语句)的网络损失作为第一损失。并获取去除后的负样本对(第二样本图像及第一样本描述语句,或第一样本图像及第二样本描述语句)的网络损失。

在一种可能的实现方式中,利用样本集训练所述神经网络的步骤还可包括:根据所述总体损失,训练所述神经网络。

在一种可能的实现方式中,在得到总体网络损失 L 后,可根据总体网络损失 L 训练上述神经网络。从而确定训练后的神经网络。本申请对神经网络的具体训练方式不作限制。

在一种可能的实现方式中,在根据所述神经网络的第一损失和第二损失,确定所述神经网络的总体损失之前,还包括:获得所述第一损失。

所述获得所述第一损失的步骤包括:

将同一正样本对的第一样本图像及第一样本描述语句输入所述神经网络中处理,获得第一训练得分;将不同正样本对的第一样本图像及第一样本描述语句输入所述神经网络

络中处理, 获得第二训练得分; 根据多个第一训练得分和多个第二训练得分, 获得第一损失。

举例来说, 可以获取正样本对 (第一样本图像及其第一样本描述语句) 的网络损失。对于训练集中任意一个正样本对, 可将同一正样本对 (O_i, Q_i) 的第一样本图像 O_i 及第一样本描述语句 Q_i 输入如图2所示的神经网络中处理, 得到第一训练得分 $s(O_i, Q_i)$ 。其中, i 为样本编号, $1 \leq i \leq N$, N 为样本集中的正样本对的数量。

在一种可能的实现方式中, 对于不同正样本对的第一样本图像以及与其不对应的第一样本描述语句 (O_i, Q_j) , 可将其输入如图2所示的神经网络中处理, 得到第二训练得分 $s(O_i, Q_j)$ 。其中, j 为样本编号, $1 \leq j \leq N$, 且 j 不等于 i 。同样地, 将不同正样本对的第一样本图像和第一样本描述语句 (O_j, Q_i) 输入神经网络, 可得到另一个第二训练得分 $s(O_j, Q_i)$ 。

在一种可能的实现方式中, 分别对训练集中的正样本对 (第一样本图像及第一样本描述语句) 进行处理, 可得到多个第一训练得分和多个第二训练得分, 进而可得到原始样本的第一损失 L_{rank} :

$$L_{\text{rank}} = \sum_{i,j}^N ([m - s(O_i, Q_i) + s(O_i, Q_j)]_+ + [m - s(O_i, Q_i) + s(O_j, Q_i)]_+) \quad (1)$$

在公式 (1) 中, 运算符 $[x]_+$ 可表示取 x 和 0 之间的最大值, 也即 x 大于 0 时取 x 的值, x 小于或等于 0 时取 0; m 可以为常数, 用于表示网络损失的间距。应当理解, 本领域技术人员可根据实际情况设定 m 的值 (例如 0.1), 本申请对 m 的具体取值不作限制。

在一种可能的实现方式中, 在根据所述神经网络的第一损失和第二损失, 确定所述神经网络的总体损失之前, 还包括: 获得所述第二损失;

所述获得所述第二损失的步骤包括:

将同一负样本对的第二样本图像及第一样本描述语句输入所述神经网络中处理, 获得第三训练得分; 将不同负样本对的第二样本图像及第一样本描述语句输入所述神经网络中处理, 获得第四训练得分; 将同一负样本对的第一样本图像及第二样本描述语句输入所述神经网络中处理, 获得第五训练得分; 将不同负样本对的第一样本图像及第二样本描述语句输入所述神经网络中处理, 获得第六训练得分; 根据多个第三训练得分、多个第四训练得分、多个第五训练得分及多个第六训练得分, 获得第二损失。

举例来说, 可以获取去除后的负样本 (第二样本图像及第二样本描述语句) 的网络损失。对于训练集中的同一负样本对 (O_i^*, Q_i) , 可将第二样本图像 O_i^* 及第一样本描述语句 Q_i 输入如图2所示的神经网络中处理, 得到第三训练得分 $s(O_i^*, Q_i)$ 。其中, i 为样本编号, $1 \leq i \leq N$, N 为样本集中的样本对的数量。

在一种可能的实现方式中, 对于训练集中不同的负样本对 (第二样本图像 O_i^* 及不对应的第一样本描述语句 Q_j), 可将其输入如图2所示的神经网络中处理, 得到第四训练得分 $s(O_i^*, Q_j)$ 。其中, j 为样本编号, $1 \leq j \leq N$, 且 j 不等于 i 。

同样地, 将同一负样本对 (O_i, Q_i^*) 的第一样本图像和对应的第二样本描述语句输入神经网络, 可得到第五训练得分 $s(O_i, Q_i^*)$; 将不同负样本对 (O_j, Q_i^*) 的第一样本图像和第二样本描述语句输入神经网络, 可得到第六训练得分 $s(O_j, Q_i^*)$ 。

在一种可能的实现方式中, 对训练集中的多个正样本对 (第一样本图像及第一样本描述语句) 及去除后的负样本对分别进行处理, 可得到多个第三训练得分、多个第四训练得分、多个第五训练得分及多个第六训练得分, 进而可得到去除后样本的第二损失 L_{erase} :

$$L_{erase} = \sum_{i,j}^N ([m - s(O_i^*, Q_i) + s(O_i^*, Q_j)]_+ + [m - s(O_i, Q_i^*) + s(O_j, Q_i^*)]_+) \quad (2)$$

在公式(2)中,运算符 $[x]_+$ 可表示取 x 和0之间的最大值,也即 x 大于0时取 x 的值, x 小于或等于0时取0; m 可以为常数,用于表示网络损失的间距。应当理解,本领域技术人员可根据实际情况设定 m 的值(例如0.1),本申请对 m 的具体取值不作限制。

5 在一种可能的实现方式中,在确定第一损失和第二损失后,可根据第一损失和第二损失确定所述神经网络的总体损失,进而根据总体损失训练所述神经网络。

其中,根据所述神经网络的第一损失和第二损失,确定所述神经网络的总体损失的步骤可包括:将所述第一损失与所述第二损失进行加权叠加,获取所述神经网络的总体损失。

10 举例来说,可通过如下公式计算神经网络的总体网络损失 L :

$$L = \beta L_{erase} + \gamma L_{rank} \quad (3)$$

在公式(3)中, β 和 γ 分别表示第一损失和第二损失的权重。应当理解,本领域技术人员可根据实际情况设定 β 和 γ 的值,本申请对 β 和 γ 的具体取值不作限制。

15 在一种可能的实现方式中,在得到总体网络损失 L 后,可根据总体网络损失 L 训练所述神经网络。例如,可根据总体网络损失 L ,采用反向梯度法调整所述神经网络的网络参数值;并再次获取总体网络损失 L 。经过多次调整,当达到预设的训练条件(网络损失 L 收敛或达到训练次数)时,可以确定训练后的神经网络。本申请对神经网络的具体训练方式不作限制。

20 根据本申请实施例的图像的描述语句定位方法,通过跨模式擦除的方式消除高注意力权重的最主要的视觉或文本信息,以生成困难的训练样本,从而驱动神经网络模型寻找除最主要的证据之外的补充证据。根据本申请的实施例,利用原始查询语句的擦除图像,或者用原始图像的擦除查询语句以形成较难的训练样本,使得神经网络模型更好地利用训练数据学到潜在的文字-图片对应关系,并且不增加推理复杂性。

25 根据本申请的实施例,能够应用于机器人或手机等终端中,根据人的指导(文字或语音)定位图像中人物的位置,从而实现文本与图像之间的准确对应。

本领域技术人员可以理解,在具体实施方式的上述方法中,各步骤的撰写顺序并不意味着严格的执行顺序而对实施过程构成任何限定,各步骤的具体执行顺序应当以其功能和可能的内在逻辑确定。

30 可以理解,本申请提及的上述各个方法实施例,在不违背原理逻辑的情况下,均可以彼此相互结合形成结合后的实施例,限于篇幅,本申请不再赘述。

图5示出根据本申请实施例的图像的描述语句定位装置的框图,如图5所示,所述图像的描述语句定位装置包括:

35 第一权重获得模块51,配置为待分析描述语句与待分析图像进行分析处理,获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重;

40 第一得分获得模块52,配置为根据所述多个语句注意力权重以及待分析图像的主体特征、位置特征及关系特征,获得多个第一匹配得分,其中,所述待分析图像中包括多个对象,主体对象为所述多个对象中注意力权重最高的对象,所述主体特征为所述主体对象的特征,所述位置特征为所述多个对象的位置特征,所述关系特征为所述多个对象之间的关系特征;

第二得分获得模块53,配置为根据所述多个第一匹配得分以及所述多个图像注意力权重,获得所述待分析描述语句与所述待分析图像之间的第二匹配得分;

结果确定模块54,配置为根据所述第二匹配得分,确定所述待分析描述语句在所述待分析图像中的定位结果。

在一种可能的实现方式中,所述第一权重获得模块包括:

5 图像特征提取子模块,配置为对所述待分析图像进行特征提取,获得所述待分析图像的图像特征向量;

分词特征提取子模块,配置为对所述待分析描述语句进行特征提取,获得所述待分析描述语句的多个分词的分词嵌入向量;

10 第一权重获得子模块,配置为根据所述图像特征向量以及所述多个分词的分词嵌入向量,获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。

在一种可能的实现方式中,所述装置还包括:第二权重获得模块,配置为通过神经网络获取所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。

15 在一种可能的实现方式中,所述多个语句注意力权重包括语句主语权重、语句位置权重及语句关系权重,所述神经网络包括图像注意力网络,所述图像注意力网络包括主体网络、位置网络以及关系网络,所述多个第一匹配得分包括主语匹配得分、位置匹配得分及关系匹配得分,所述第一得分获得模块包括;

第一得分获得子模块,配置为将所述语句主语权重及主体特征输入所述主体网络中处理,获得所述主语匹配得分;

20 第二得分获得子模块,配置为将所述语句位置权重及位置特征输入所述位置网络中处理,获得所述位置匹配得分;

第三得分获得子模块,配置为将所述语句关系权重及关系特征输入所述关系网络中处理,获得所述关系匹配得分。

25 在一种可能的实现方式中,所述多个图像注意力权重包括主体对象权重、对象位置权重及对象关系权重,所述第二得分获得模块包括:

第四得分获得子模块,用于根据所述主体对象权重、所述对象位置权重及所述对象关系权重,对所述主语匹配得分、所述位置匹配得分及所述关系匹配得分进行加权平均,确定所述第二匹配得分。

在一种可能的实现方式中,所述装置还包括:

30 第三权重获得模块,用于将所述待分析图像输入特征提取网络中处理,获得所述主体特征、所述位置特征及所述关系特征。

在一种可能的实现方式中,所述结果确定模块包括:

位置确定子模块,用于在所述第二匹配得分大于或等于预设阈值的情况下,将所述主体对象的图像区域确定为所述待分析描述语句的定位位置。

35 在一种可能的实现方式中,在所述第二权重获得模块之前,还包括:训练模块,用于利用样本集训练所述神经网络,所述样本集包括多个正样本对和多个负样本对,

其中,每个正样本对包括第一样本图像及其第一样本描述语句,

40 每个负样本对包括第一样本图像及从所述第一样本描述语句中去除分词后的第二样本描述语句,或第一样本描述语句及从所述第一样本图像中去除区域后的第二样本图像。

在一种可能的实现方式中,所述神经网络还包括语言注意力网络,所述装置还包括:

分词权重确定模块,用于将所述正样本对的第一样本描述语句及第一样本图像输入所述语言注意力网络,得到所述第一样本描述语句的多个分词的注意力权重;

分词替换模块,用于采用预定标识替换所述第一样本描述语句中注意力权重最高的

分词, 获得第二样本描述语句;

第一负样本对确定模块, 用于将所述第一样本图像和所述第二样本描述语句作为负样本对。

在一种可能的实现方式中, 所述装置还包括:

5 图像权重确定模块, 用于将所述正样本对的第一样本描述语句及第一样本图像输入所述图像注意力网络, 得到所述第一样本图像的注意力权重;

区域去除模块, 用于去除所述第一样本图像中注意力权重最高的图像区域, 获得第二样本图像;

10 第二负样本对确定模块, 配置为将所述第二样本图像和所述第一样本描述语句作为负样本对。

在一种可能的实现方式中, 所述训练模块包括:

总体损失确定子模块, 配置为根据所述神经网络的第一损失和第二损失, 确定所述神经网络的总体损失;

训练子模块, 配置为根据所述总体损失, 训练所述神经网络。

15 在一种可能的实现方式中, 所述装置还包括: 第一损失获得子模块, 配置为在所述总体损失确定子模块之前获得所述第一损失; 所述第一损失获得子模块配置为:

将同一正样本对的第一样本图像及第一样本描述语句输入所述神经网络中处理, 获得第一训练得分;

20 将不同正样本对的第一样本图像及第一样本描述语句输入所述神经网络中处理, 获得第二训练得分;

根据多个第一训练得分和多个第二训练得分, 获得第一损失。

在一种可能的实现方式中, 所述装置还包括: 第二损失获得子模块, 配置为在所述总体损失确定子模块之前, 获得所述第二损失; 所述第二损失获得子模块配置为:

25 将同一负样本对的第二样本图像及第一样本描述语句输入所述神经网络中处理, 获得第三训练得分;

将不同负样本对的第二样本图像及第一样本描述语句输入所述神经网络中处理, 获得第四训练得分;

30 将同一负样本对的第一样本图像及第二样本描述语句输入所述神经网络中处理, 获得第五训练得分;

将不同负样本对的第一样本图像及第二样本描述语句输入所述神经网络中处理, 获得第六训练得分;

根据多个第三训练得分、多个第四训练得分、多个第五训练得分及多个第六训练得分, 获得第二损失。

在一种可能的实现方式中, 所述总体损失确定子模块配置为:

35 将所述第一损失与所述第二损失进行加权叠加, 获取所述神经网络的总体损失。

在一些实施例中, 本申请实施例提供的装置具有的功能或包含的模块可以用于执行上文方法实施例描述的方法, 其具体实现可以参照上文方法实施例的描述, 为了简洁, 这里不再赘述

40 本申请实施例还提出一种计算机可读存储介质, 其上存储有计算机程序指令, 所述计算机程序指令被处理器执行时实现上述方法。计算机可读存储介质可以是非易失性计算机可读存储介质。

本申请实施例还提出一种电子设备, 包括: 处理器; 用于存储处理器可执行指令的存储器; 其中, 所述处理器被配置为上述方法。

电子设备可以被提供为终端、服务器或其它形态的设备。

图6示出根据本申请实施例的一种电子设备800的框图。例如，电子设备800可以是移动电话，计算机，数字广播终端，消息收发设备，游戏控制台，平板设备，医疗设备，健身设备，个人数字助理等终端。

参照图6，电子设备800可以包括以下一个或多个组件：处理组件802，存储器804，
5 电源组件806，多媒体组件808，音频组件810，输入/输出（I/O）的接口812，传感器组件814，以及通信组件816。

处理组件802通常控制电子设备800的整体操作，诸如与显示，电话呼叫，数据通信，相机操作和记录操作相关联的操作。处理组件802可以包括一个或多个处理器820来执行指令，以完成上述的方法的全部或部分步骤。此外，处理组件802可以包括一个或多个
10 模块，便于处理组件802和其他组件之间的交互。例如，处理组件802可以包括多媒体模块，以方便多媒体组件808和处理组件802之间的交互。

存储器804被配置为存储各种类型的数据以支持在电子设备800的操作。这些数据的示例包括用于在电子设备800上操作的任何应用程序或方法的指令，联系人数据，电话簿数据，消息，图片，视频等。存储器804可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器（SRAM），电可擦除可编程只读存储器（EEPROM），可擦除可编程只读存储器（EPROM），可编程只读存储器（PROM），只读存储器（ROM），磁存储器，快闪存储器，磁盘或光盘。

电源组件806为电子设备800的各种组件提供电力。电源组件806可以包括电源管理系统，一个或多个电源，及其他与为电子设备800生成、管理和分配电力相关联的组件。

多媒体组件808包括在所述电子设备800和用户之间的提供一个输出接口的屏幕。在一些实施例中，屏幕可以包括液晶显示器（LCD）和触摸面板（TP）。如果屏幕包括触摸面板，屏幕可以被实现为触摸屏，以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界，而且还检测与所述触摸或滑动操作相关的持续时间和压力。在
20 一些实施例中，多媒体组件808包括一个前置摄像头和/或后置摄像头。当电子设备800处于操作模式，如拍摄模式或视频模式时，前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

音频组件810被配置为输出和/或输入音频信号。例如，音频组件810包括一个麦克风（MIC），当电子设备800处于操作模式，如呼叫模式、记录模式和语音识别模式时，麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器804或经由通信组件816发送。在一些实施例中，音频组件810还包括一个扬声器，用于输出音频信号。

I/O接口812为处理组件802和外围接口模块之间提供接口，上述外围接口模块可以是键盘，点击轮，按钮等。这些按钮可包括但不限于：主页按钮、音量按钮、启动按钮和锁定按钮。

传感器组件814包括一个或多个传感器，用于为电子设备800提供各个方面的状态评估。例如，传感器组件814可以检测到电子设备800的打开/关闭状态，组件的相对定位，例如所述组件为电子设备800的显示器和小键盘，传感器组件814还可以检测电子设备800或电子设备800一个组件的位置改变，用户与电子设备800接触的存在或不存在，电子设备800方位或加速/减速和电子设备800的温度变化。传感器组件814可以包括接近传感器，被配置用来在没有任何的物理接触时检测附近物体的存在。传感器组件814还可以包括光传感器，如CMOS或CCD图像传感器，用于在成像应用中使用。在一些实施例中，该传感器组件814还可以包括加速度传感器，陀螺仪传感器，磁传感器，压力传感
40

器或温度传感器。

通信组件816被配置为便于电子设备800和其他设备之间有线或无线方式的通信。电子设备800可以接入基于通信标准的无线网络，如WiFi，2G或3G，或它们的组合。在一个示例性实施例中，通信组件816经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，所述通信组件816还包括近场通信（NFC）模块，以促进短程通信。例如，在NFC模块可基于射频识别（RFID）技术，红外数据协会（IrDA）技术，超宽带（UWB）技术，蓝牙（BT）技术和其他技术来实现。

在示例性实施例中，电子设备800可以被一个或多个应用专用集成电路（ASIC）、数字信号处理器（DSP）、数字信号处理设备（DSPD）、可编程逻辑器件（PLD）、现场可编程门阵列（FPGA）、控制器、微控制器、微处理器或其他电子元件实现，用于执行上述方法。

在示例性实施例中，还提供了一种非易失性计算机可读存储介质，例如包括计算机程序指令的存储器804，上述计算机程序指令可由电子设备800的处理器820执行以完成上述方法。

图7示出根据本申请实施例的一种电子设备1900的框图。例如，电子设备1900可以提供为一服务器。参照图7，电子设备1900包括处理组件1922，其进一步包括一个或多个处理器，以及由存储器1932所代表的存储器资源，用于存储可由处理组件1922的执行的指令，例如应用程序。存储器1932中存储的应用程序可以包括一个或一个以上的每一个对应于一组指令的模块。此外，处理组件1922被配置为执行指令，以执行上述方法。

电子设备1900还可以包括一个电源组件1926被配置为执行电子设备1900的电源管理，一个有线或无线网络接口1950被配置为将电子设备1900连接到网络，和一个输入输出（I/O）接口1958。电子设备1900可以操作基于存储在存储器1932的操作系统，例如Windows Server™，Mac OS X™，Unix™，Linux™，FreeBSD™或类似。

在示例性实施例中，还提供了一种非易失性计算机可读存储介质，例如包括计算机程序指令的存储器1932，上述计算机程序指令可由电子设备1900的处理组件1922执行以完成上述方法。

本申请可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于使处理器实现本申请的各个方面的计算机可读程序指令。

计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子（非穷举的列表）包括：便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM或闪存）、静态随机存取存储器（SRAM）、便携式压缩盘只读存储器（CD-ROM）、数字多功能盘（DVD）、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波（例如，通过光纤电缆的光脉冲）、或者通过电线传输的电信号。

这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

用于执行本申请操作的计算机程序指令可以是汇编指令、指令集架构 (ISA) 指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码, 所述编程语言包括面向对象的编程语言—诸如 Smalltalk、C++ 等, 以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中, 远程计算机可以通过任意种类的网络—包括局域网 (LAN) 或广域网 (WAN)—连接到用户计算机, 或者, 可以连接到外部计算机 (例如利用因特网服务提供商来通过因特网连接)。在一些实施例中, 通过利用计算机可读程序指令的状态信息来个性化定制电子电路, 例如可编程逻辑电路、现场可编程门阵列 (FPGA) 或可编程逻辑阵列 (PLA), 该电子电路可以执行计算机可读程序指令, 从而实现本申请的各个方面。

这里参照根据本申请实施例的方法、装置 (系统) 和计算机程序产品的流程图和/或框图描述了本申请的各个方面。应当理解, 流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合, 都可以由计算机可读程序指令实现。

这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器, 从而生产出一种机器, 使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时, 产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中, 这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作, 从而, 存储有指令的计算机可读介质则包括一个制造品, 其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上, 使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤, 以产生计算机实现的过程, 从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

附图中的流程图和框图显示了根据本申请的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上, 流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分, 所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中, 方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如, 两个连续的方框实际上可以基本并行地执行, 它们有时也可以按相反的顺序执行, 这依所涉及的功能而定。也要注意, 框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合, 可以用执行规定的功能或动作的专用的基于硬件的系统来实现, 或者可以用专用硬件与计算机指令的组合来实现。

以上已经描述了本申请的各实施例, 上述说明是示例性的, 并非穷尽性的, 并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下, 对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择, 旨在最好地解释各实施例的原理、实际应用或对市场中的技术改进, 或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

权利要求书

- 1、一种图像的描述语句定位方法，包括：
对待分析描述语句与待分析图像进行分析处理，获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重；
5 根据所述多个语句注意力权重以及待分析图像的主体特征、位置特征及关系特征，获得多个第一匹配得分，其中，所述待分析图像中包括多个对象，主体对象为所述多个对象中注意力权重最高的对象，所述主体特征为所述主体对象的特征，所述位置特征为所述多个对象的位置特征，所述关系特征为所述多个对象之间的关系特征；
根据所述多个第一匹配得分以及所述多个图像注意力权重，获得所述待分析描述语句与
10 所述待分析图像之间的第二匹配得分；
根据所述第二匹配得分，确定所述待分析描述语句在所述待分析图像中的定位结果。
- 2、根据权利要求1所述的方法，其中，对待分析描述语句与待分析图像分别进行分析处理，获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像
15 注意力权重，包括：
对所述待分析图像进行特征提取，获得所述待分析图像的图像特征向量；
对所述待分析描述语句进行特征提取，获得所述待分析描述语句的多个分词的分词嵌入向量；
根据所述图像特征向量以及所述多个分词的分词嵌入向量，获得所述待分析描述语句
20 的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。
- 3、根据权利要求1或2所述的方法，其中，还包括：通过神经网络获取所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。
- 4、根据权利要求3所述的方法，其中，所述多个语句注意力权重包括语句主语权重、语句位置权重及语句关系权重；
25 所述神经网络包括图像注意力网络；
所述图像注意力网络包括主体网络、位置网络以及关系网络；
所述多个第一匹配得分包括主语匹配得分、位置匹配得分及关系匹配得分；
根据所述多个语句注意力权重以及待分析图像的主体特征、位置特征及关系特征，获得多个第一匹配得分，包括；
30 将所述语句主语权重及主体特征输入所述主体网络中处理，获得所述主语匹配得分；
将所述语句位置权重及位置特征输入所述位置网络中处理，获得所述位置匹配得分；
将所述语句关系权重及关系特征输入所述关系网络中处理，获得所述关系匹配得
35 分。
- 5、根据权利要求4所述的方法，其中，所述多个图像注意力权重包括主体对象权重、对象位置权重及对象关系权重；
根据所述多个第一匹配得分以及所述多个图像注意力权重，获得所述待分析描述语句与
40 所述待分析图像之间第二匹配得分，包括：
根据所述主体对象权重、所述对象位置权重及所述对象关系权重，对所述主语匹配得分、所述位置匹配得分及所述关系匹配得分进行加权平均，确定所述第二匹配得分。
- 6、根据权利要求1-5中任意一项所述的方法，其中，所述方法还包括：
将所述待分析图像输入特征提取网络中处理，获得所述主体特征、所述位置特征及
所述关系特征。

7、根据权利要求1-6中任意一项所述的方法，其中，根据所述第二匹配得分，确定所述待分析描述语句在所述待分析图像中的定位结果，包括：

在所述第二匹配得分大于或等于预设阈值的情况下，将所述主体对象的图像区域确定为所述待分析描述语句的定位位置。

5 8、根据权利要求3-7中任意一项所述的方法，其中，在通过神经网络获取所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重之前，所述方法还包括：利用样本集训练所述神经网络，所述样本集包括多个正样本对和多个负样本对，

其中，每个正样本对包括第一样本图像及所述第一样本图像的第一样本描述语句，
10 每个负样本对包括第一样本图像及从所述第一样本描述语句中去除分词后的第二样本描述语句，或第一样本描述语句及从所述第一样本图像中去除图像注意力权重最高的区域后的第二样本图像。

9、根据权利要求8所述的方法，其中，所述神经网络还包括语言注意力网络，所述方法还包括：

15 将所述正样本对的第一样本描述语句及第一样本图像输入所述语言注意力网络，得到所述第一样本描述语句的多个分词的注意力权重；

采用预定标识替换所述第一样本描述语句中注意力权重最高的分词，获得第二样本描述语句；

将所述第一样本图像和所述第二样本描述语句作为负样本对。

20 10、根据权利要求8或9所述的方法，其中，所述方法还包括：

将所述正样本对的第一样本描述语句及第一样本图像输入所述图像注意力网络，得到所述第一样本图像的注意力权重；

去除所述第一样本图像中注意力权重最高的图像区域，获得第二样本图像；

将所述第二样本图像和所述第一样本描述语句作为负样本对。

25 11、根据权利要求8-10中任意一项所述的方法，其中，利用样本集训练所述神经网络，包括：

根据所述神经网络的第一损失和第二损失，确定所述神经网络的总体损失；

根据所述总体损失，训练所述神经网络。

30 12、根据权利要求11所述的方法，其中，在根据所述神经网络的第一损失和第二损失，确定所述神经网络的总体损失之前，所述方法还包括：获得所述第一损失；

所述获得所述第一损失的步骤包括：

将同一正样本对的第一样本图像及第一样本描述语句输入所述神经网络中处理，获得第一训练得分；

35 将不同正样本对的第一样本图像及第一样本描述语句输入所述神经网络中处理，获得第二训练得分；

根据多个第一训练得分和多个第二训练得分，获得第一损失。

13、根据权利要求11或12所述的方法，其中，在根据所述神经网络的第一损失和第二损失，确定所述神经网络的总体损失之前，所述方法还包括：获得所述第二损失；

所述获得所述第二损失的步骤包括：

40 将同一负样本对的第二样本图像及第一样本描述语句输入所述神经网络中处理，获得第三训练得分；

将不同负样本对的第二样本图像及第一样本描述语句输入所述神经网络中处理，获得第四训练得分；

将同一负样本对的第一样本图像及第二样本描述语句输入所述神经网络中处理，获

得第五训练得分;

将不同负样本对的第一样本图像及第二样本描述语句输入所述神经网络中处理, 获得第六训练得分;

5 根据多个第三训练得分、多个第四训练得分、多个第五训练得分及多个第六训练得分, 获得第二损失。

14、根据权利要求11-13中任意一项所述的方法, 其中, 根据所述神经网络的第一损失和第二损失, 确定所述神经网络的总体损失, 包括:

将所述第一损失与所述第二损失进行加权叠加, 获取所述神经网络的总体损失。

10 15、一种图像的描述语句定位装置, 包括:

第一权重获得模块, 配置为对待分析描述语句与待分析图像进行分析处理, 获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重;

15 第一得分获得模块, 配置为根据所述多个语句注意力权重以及待分析图像的主体特征、位置特征及关系特征, 获得多个第一匹配得分, 其中, 所述待分析图像中包括多个对象, 主体对象为所述多个对象中注意力权重最高的对象, 所述主体特征为所述主体对象的特征, 所述位置特征为所述多个对象的位置特征, 所述关系特征为所述多个对象之间的关系特征;

第二得分获得模块, 配置为根据所述多个第一匹配得分以及所述多个图像注意力权重, 获得所述待分析描述语句与所述待分析图像之间的第二匹配得分;

20 结果确定模块, 配置为根据所述第二匹配得分, 确定所述待分析描述语句在所述待分析图像中的定位结果。

16、根据权利要求15所述的装置, 其中, 所述第一权重获得模块包括:

图像特征提取子模块, 配置为对所述待分析图像进行特征提取, 获得所述待分析图像的图像特征向量;

25 分词特征提取子模块, 配置为对所述待分析描述语句进行特征提取, 获得所述待分析描述语句的多个分词的分词嵌入向量;

第一权重获得子模块, 配置为根据所述图像特征向量以及所述多个分词的分词嵌入向量, 获得所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。

30 17、根据权利要求15或16所述的装置, 其中, 所述装置还包括: 第二权重获得模块, 配置为通过神经网络获取所述待分析描述语句的多个语句注意力权重以及所述待分析图像的多个图像注意力权重。

35 18、根据权利要求17所述的装置, 其中, 所述多个语句注意力权重包括语句主语权重、语句位置权重及语句关系权重, 所述神经网络包括图像注意力网络, 所述图像注意力网络包括主体网络、位置网络以及关系网络, 所述多个第一匹配得分包括主语匹配得分、位置匹配得分及关系匹配得分,

所述第一得分获得模块包括;

第一得分获得子模块, 配置为将所述语句主语权重及主体特征输入所述主体网络中处理, 获得所述主语匹配得分;

40 第二得分获得子模块, 配置为将所述语句位置权重及位置特征输入所述位置网络中处理, 获得所述位置匹配得分;

第三得分获得子模块, 配置为将所述语句关系权重及关系特征输入所述关系网络中处理, 获得所述关系匹配得分。

19、根据权利要求18所述的装置, 其中, 所述多个图像注意力权重包括主体对象权重、对象位置权重及对象关系权重, 所述第二得分获得模块包括:

第四得分获得子模块,配置为根据所述主体对象权重、所述对象位置权重及所述对象关系权重,对所述主语匹配得分、所述位置匹配得分及所述关系匹配得分进行加权平均,确定所述第二匹配得分。

20、根据权利要求15-19中任意一项所述的装置,其中,所述装置还包括:

5 第三权重获得模块,配置为将所述待分析图像输入特征提取网络中处理,获得所述主体特征、所述位置特征及所述关系特征。

21、根据权利要求15-20中任意一项所述的装置,其中,所述结果确定模块包括:

位置确定子模块,配置为在所述第二匹配得分大于或等于预设阈值的情况下,将所述主体对象的图像区域确定为所述待分析描述语句的定位位置。

10 22、根据权利要求17-21中任意一项所述的装置,其中,所述装置还包括:训练模块,配置为在所述第二权重获得模块之前,利用样本集训练所述神经网络,所述样本集包括多个正样本对和多个负样本对,

其中,每个正样本对包括第一样本图像及所述第一样本图像的第一样本描述语句,

15 每个负样本对包括第一样本图像及从所述第一样本描述语句中去除分词后的第二样本描述语句,或第一样本描述语句及从所述第一样本图像中去除图像注意力权重最高的区域后的第二样本图像。

23、根据权利要求22所述的装置,其中,所述神经网络还包括语言注意力网络,所述装置还包括:

20 分词权重确定模块,配置为将所述正样本对的第一样本描述语句及第一样本图像输入所述语言注意力网络,得到所述第一样本描述语句的多个分词的注意力权重;

分词替换模块,配置为采用预定标识替换所述第一样本描述语句中注意力权重最高的分词,获得第二样本描述语句;

第一负样本对确定模块,配置为将所述第一样本图像和所述第二样本描述语句作为负样本对。

25 24、根据权利要求22或23所述的装置,其中,所述装置还包括:

图像权重确定模块,配置为将所述正样本对的第一样本描述语句及第一样本图像输入所述图像注意力网络,得到所述第一样本图像的注意力权重;

区域去除模块,配置为去除所述第一样本图像中注意力权重最高的图像区域,获得第二样本图像;

30 第二负样本对确定模块,配置为将所述第二样本图像和所述第一样本描述语句作为负样本对。

25、根据权利要求22-24中任意一项所述的装置,其中,所述训练模块包括:

总体损失确定子模块,配置为根据所述神经网络的第一损失和第二损失,确定所述神经网络的总体损失;

35 训练子模块,配置为根据所述总体损失,训练所述神经网络。

26、根据权利要求25所述的装置,其中,在所述总体损失确定子模块之前,所述装置还包括:第一损失获得子模块,配置为获得所述第一损失;

所述第一损失获得子模块配置为:

40 将同一正样本对的第一样本图像及第一样本描述语句输入所述神经网络中处理,获得第一训练得分;

将不同正样本对的第一样本图像及第一样本描述语句输入所述神经网络中处理,获得第二训练得分;

根据多个第一训练得分和多个第二训练得分,获得第一损失。

27、根据权利要求25或26所述的装置,其中,所述装置还包括:第二损失获得子模

- 块，配置为在所述总体损失确定子模块之前获得所述第二损失；
所述第二损失获得子模块配置为：
将同一负样本对的第二样本图像及第一样本描述语句输入所述神经网络中处理，获得第三训练得分；
5 将不同负样本对的第二样本图像及第一样本描述语句输入所述神经网络中处理，获得第四训练得分；
将同一负样本对的第一样本图像及第二样本描述语句输入所述神经网络中处理，获得第五训练得分；
将不同负样本对的第一样本图像及第二样本描述语句输入所述神经网络中处理，获得第六训练得分；
10 根据多个第三训练得分、多个第四训练得分、多个第五训练得分及多个第六训练得分，获得第二损失。
28、根据权利要求25-27中任意一项所述的装置，其中，所述总体损失确定子模块配置为：将所述第一损失与所述第二损失进行加权叠加，获取所述神经网络的总体损失。
15 29、一种电子设备，包括：
处理器；
用于存储处理器可执行指令的存储器；
其中，所述处理器被配置为：执行权利要求1至14中任意一项所述的方法。
30、一种计算机可读存储介质，其上存储有计算机程序指令，所述计算机程序指令
20 被处理器执行时实现权利要求1至14中任意一项所述的方法。