(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106973244 B
(45)授权公告日 2021.04.20

(21)申请号 201610995334.6
(22)申请日 2016.11.11
(65)同一申请的已公布的文献号
申请公布号 CN 106973244 A
(43)申请公布日 2017.07.21
(30)优先权数据
14/995,032 2016.01.13 US
(73)专利权人 奥多比公司
地址 美国加利福尼亚州
(72)发明人 王兆闻 尤全增 金海琳 方晨
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 酆迅
(51)Int.Cl.
H04N 5/278(2006.01)
H04N 21/431(2011.01)
H04N 21/488(2011.01)
G06N 3/08(2006.01)
G06N 3/04(2006.01)
(56)对比文件
US 2003217066 A1,2003.11.20
US 2010272411 A1,2010.10.28
US 2003103675 A1,2003.06.05
CN 104834757 A,2015.08.12
CN 104537392 A,2015.04.22
CN 103336969 A,2013.10.02
审查员 王娟

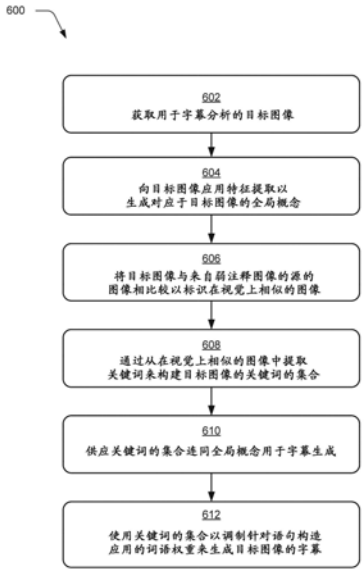
权利要求书3页 说明书16页 附图12页

(54)发明名称

使用弱监督数据自动生成图像字幕的方法和系统

(57)摘要

本发明的各实施例总体上涉及使用弱监督为图像配字幕。具体地，本文中描述了用于使用弱监督为图像配字幕的技术。在实现中，获取关于目标图像的弱监督数据并且使用其提供补充被获得用于图像配字幕的全局图像概念的细节信息。弱监督数据是指没有被紧密地监管并且可能包括误差的噪声数据。给定目标图像，可以从弱注释的图像的源、诸如在线社交网络采集在视觉上相似的图像的弱监督数据。通常，在线发布的图像包括由用户添加的标签、标题、标注和短描述形式的“弱”注释。通过提取在不同源中发现的在视觉上相似的图像的关键词来生成目标图像的弱监督数据。然后在图像配字幕分析期间采用弱监督数据中包括的关键词来调制被应用于概率分类的权重。



CN 106973244 B

1.一种用于使用弱监督数据自动生成图像字幕的方法,应用于使用一个或多个计算设备促进图像采集管理的数字媒体环境中,所述方法包括:

获取用于字幕分析的目标图像;

向所述目标图像应用特征提取以生成对应于所述图像的全局概念;

将所述目标图像与来自弱注释图像的源的图像相比较以标识在视觉上相似的图像,其中所述弱注释图像均包括按照以下至少一项的形式的弱注释:标签、标题、标注和短描述;

通过从所述在视觉上相似的图像中提取用于指示图像细节的所述目标图像的关键词来构建所述关键词的集合;以及

供应指示图像细节的所述关键词的集合作为所述弱监督数据用于连同所述全局概念进行字幕生成。

2.根据权利要求1所述的方法,还包括使用所述关键词的集合调制应用于语句构造的词语权重来生成所述目标图像的字幕。

3.根据权利要求1所述的方法,其中所述关键词的集合扩展可用于所述字幕分析的候选字幕的集合以便包括根据所述弱监督数据得到的具体的对象、属性和术语以及还包括根据所述特征提取得到的所述全局概念。

4.根据权利要求1所述的方法,其中向语言处理模型供应所述关键词的集合,所述语言处理模型可操作以通过计算说明所述弱监督数据的概率分布来在概率上生成所述图像的描述性字幕。

5.根据权利要求1所述的方法,其中向所述目标图像应用特征提取包括使用预先训练的卷积神经网络CNN来使用指示所述全局概念的全局描述性术语来对所述图像编码。

6.根据权利要求1所述的方法,其中供应所述关键词的集合包括向被设计成实现用于生成所述目标图像的字幕的语言建模和语句构造技术的循环神经网络RNN提供关键词。

7.根据权利要求6所述的方法,其中所述RNN基于根据多个迭代中的权重因子计算的概率分布来迭代地预测用于组合作为所述目标图像的字幕的词语的序列。

8.根据权利要求7所述的方法,其中针对所述多个迭代中的每个迭代在所述RNN中注入所述关键词的集合以调制用于预测所述序列的所述权重因子。

9.根据权利要求1所述的方法,其中字幕生成包括用于确定用于组合作为所述目标图像的字幕的词语的序列的多个迭代并且供应所述关键词的集合包括针对所述多个迭代中的每个迭代提供相同的关键词。

10.根据权利要求1所述的方法,其中构建所述关键词的集合包括基于相关准则来对与所述在视觉上相似的图像相关联的关键词评分和评级并且生成顶部评级关键词的经过滤的列表。

11.根据权利要求1所述的方法,其中向所述关键词的集合中的关键词被分配关键词权重以有效地改变被实现用于字幕生成的概率分类中的词语概率从而有利于指示所述图像细节的关键词。

12.根据权利要求1所述的方法,其中所述弱注释的图像的源包括通过网络可访问的图像的在线储存库。

13.一种应用于使用一个或多个计算设备促进图像采集访问的数字媒体环境中的系统,所述系统包括:

一个或多个处理设备；

一个或多个计算机可读介质，其存储经由所述一个或多个处理设备可执行以实现字幕生成器的指令，所述字幕生成器被配置成执行使用弱监督数据自动生成图像字幕的操作，所述操作包括：

经由卷积神经网络CNN处理用于字幕分析的目标图像，所述CNN被配置成提取对应于所述目标图像的全局概念；

将所述目标图像与来自弱注释图像的至少一个源的图像相比较以标识在视觉上相似的图像，其中所述弱注释图像均包括按照以下至少一项的形式的弱注释：标签、标题、标注和短描述；

通过从所述在视觉上相似的图像中提取用于指示图像细节的所述目标图像的关键词来构建所述关键词的集合作为用于通知字幕生成的弱监督数据；

向循环神经网络RNN供应指示图像细节的所述关键词的集合连同所述全局概念，所述RNN被配置成实现用于生成所述目标图像的字幕的语言建模和语句构造技术；以及

使用所述关键词的集合调制由所述RNN针对语句构造应用的词语权重并经由所述RNN来生成所述目标图像的所述字幕。

14.根据权利要求13所述的系统，其中所述弱注释图像的至少一个源包括具有通过用户与指示低层图像细节的弱注释相关联的图像数据库的社交网络站点。

15.根据权利要求13所述的系统，其中所述弱注释图像的至少一个源包括用于训练所述字幕生成器的训练图像的集合。

16.根据权利要求13所述的系统，其中：

所述RNN基于根据多个迭代中的权重因子计算的概率分布来迭代地预测用于组合作为所述目标图像的字幕的词语的序列；以及

针对所述多个迭代中的每个迭代在所述RNN中注入从所述弱监督数据得到的所述关键词的相同集合以调制用于预测所述序列的所述权重因子。

17.一种用于自动生成经由图像服务实现的图像字幕的方法，应用于使用一个或多个计算设备促进图像采集管理的数字媒体环境中，所述方法包括：

将用于字幕分析的目标图像与来自弱注释图像的至少一个源的图像相比较以标识在视觉上相似的图像，其中所述弱注释图像均包括按照以下至少一项的形式的弱注释：标签、标题、标注和短描述；

通过从所述在视觉上相似的图像中提取用于指示图像细节的所述目标图像的关键词来构建所述关键词的集合作为用于通知字幕生成的弱监督数据；以及

向字幕生成模型供应指示所述图像细节的所述关键词的集合，所述字幕生成模型被配置成迭代地组合根据概念和属性得到的词语以在多个迭代中构造字幕；以及

根据语义注意模型构造所述字幕，所述语义注意模型被配置成基于与在先迭代中预测的词语的关联来调制针对所述多个迭代中的每个迭代向所述关键词分配的权重。

18.根据权利要求17所述的方法，其中所述语义注意模型引起在所述多个迭代中的每个迭代处要被考虑的不同的关键词。

19.根据权利要求18所述的方法，其中所述字幕生成模型包括循环神经网络RNN，所述RNN被设计成实现用于生成所述目标图像的所述字幕的语言建模和语句构造技术。

20.根据权利要求19所述的方法,其中所述语义注意模型包括向所述RNN的每个节点的输入施加的输入注意模型、向由所述RNN的每个节点生成的输出施加的输出注意模型、以及供应关于用于调制向每个迭代施加的权重的所述字幕分析的当前状态和上下文的反馈的反馈回路。

使用弱监督数据自动生成图像字幕的方法和系统

背景技术

[0001] 自动生成图像的自然语言描述由于用于图像搜索、视觉受损人群的可访问性、以及图像采集的管理的实际应用而不断地吸引着人们的兴趣。传统的用于图像处理的技术由于传统的图像标记和搜索算法的限制而不支持高精度自然语言配字幕和图像搜索。这是因为，传统的技术仅使标签与图像相关联，但是没有定义标签之间或者标签与图像本身之间的关系。另外，传统的技术可以包括使用自顶向下方法，在该方法中，首先得到图像的整个“要点”然后通过语言建模和语句生成将其细化为适当的描述性词语或字幕。然而，这一自顶向下方法在捕获图像的精细细节(诸如贡献图像的精确描述的局部对象、属性和区域方面)工作并不良好。这样，可能很难使用传统的方法来生成精确且复杂的图像字幕，诸如“给在高的椅子中拿着玩具的孩子喂食的人”。因此，使用传统的技术生成的字幕可能忽略重要的图像细节，这使得用户很难搜索具体图像并且基于相关联的字幕来全面地理解图像的内容。

发明内容

[0002] 本发明内容部分介绍简化形式的概念的选择，这些概念在下面在具体实施例部分中进一步描述。这样，本发明内容部分并非意图标识要求保护的主题的基本特征，也并非意图用于帮助确定要求保护的主题的范围。

[0003] 本文中描述用于使用弱监督为图像配字幕的技术。在一个或多个实现中，获取关于目标图像的弱监督数据并且使用其提供补充被获得用于图像配字幕的全局图像概念的细节信息。弱监督数据是指没有被紧密地监管并且可能包括误差的噪声数据。给定目标图像，可以从弱注释的图像的不同的源(诸如在线社交网络、图像共享站点和图像数据库)来采集在视觉上相似的图像的弱监督数据。通常，在线发布的图像包括由用户添加的标签、标题、标注和短描述形式的“弱”注释。通过提取和聚合在弱注释图像的不同的源中发现的在视觉上相似的图像的关键词来生成目标图像的弱监督数据。然后，在图像配字幕分析期间采用弱监督数据中包括的关键词来调制被应用于概率分类的权重。因此，取决于弱监督数据来计算用于预测图像配字幕的词语的概率分布。

[0004] 在各实现方式中，图像配字幕框架基于神经网络和机器学习。给定目标图像，应用特征提取技术以得到描述图像的“要点”的全局图像概念。例如，可以使用预先训练的卷积神经网络(CNN)来使用全局描述性术语对图像编码。CNN产生反映全局图像概念的视觉特征矢量。然后，将所得到的关于全局图像概念的信息馈送到语言处理模型中，语言处理模型操作以在概率上生成图像的描述性字幕。比如，可以将视觉特征矢量馈送到循环神经网络(RNN)中，RNN被设计成实现语言建模和语句生成技术。RNN被设计成基于根据多个迭代中的权重因子计算的概率分布来迭代地预测用于组合作为目标图像的字幕的词语的序列。在这一上下文中，弱监督数据通过调制在模型中施加的权重因子来向RNN通知说明附加细节信息的操作。以这一方式，将弱监督数据中包括的关键词注入到图像配字幕框架中以补充全局图像概念，这使得能够以更大复杂性和精度来生成图像字幕。

附图说明

[0005] 参考附图来描述详细描述。在附图中，附图标记的最左侧数字标识其中首次出现该附图标记的附图。在描述和附图中不同实例中的相同的附图标记的使用可以表示相似或相同的术语。附图中表示的实体可以表示一个或多个实体，因此可以在讨论中可互换地引用这些实体的单数或复数形式。

[0006] 图1是可操作以采用本文中描述的技术的示例实现方式中的环境的图示；

[0007] 图2描绘示出根据一个或多个实现方式的字幕生成器的细节的图；

[0008] 图3描绘根据一个或多个实现方式的图像配字幕框架的示例实现；

[0009] 图4是描绘根据一个或多个实现方式的图像配字幕框架的细节的图；

[0010] 图5描绘描绘根据一个或多个实现方式的用于使用弱监督为图像配字幕的框架的图；

[0011] 图6是根据一个或多个实现方式的其中采用弱监督数据用于图像配字幕的示例过程的流程图；

[0012] 图7描绘一般性地图示用于图像配字幕的词语矢量表示的概念的示例图；

[0013] 图8是根据一个或多个实现方式的其中采用词语矢量表示用于图像配字幕的示例过程的流程图；

[0014] 图9是描绘根据一个或多个实现方式的用于图像配字幕的语义注意框架的图；

[0015] 图10是根据一个或多个实现方式的其中采用语义注意模型用于图像配字幕的示例过程的流程图；

[0016] 图11是描绘根据一个或多个实现方式的语义注意框架的细节的图；以及

[0017] 图12图示包括能够用于本文中描述的图像配字幕技术的一个或多个实现方式的示例设备的各种部件的示例系统。

具体实施方式

[0018] 概述

[0019] 传统的用于图像处理的技术由于传统的图像标记和搜索算法的限制而不支持高精度自然语言配字幕和图像搜索。这是因为，传统的技术仅使标签与图像相关联，而没有定义标签之间或者标签与图像本身之间的关系。另外，传统的技术可以包括使用自顶向下方法，在自顶向下方法中，首先得到图像的整个“要点”并且通过语言建模和语句生成将其细化成适当的描述性词语和字幕。然而，这一自顶向下方法在捕获图像的精细细节(诸如贡献图像的精确描述的局部对象、属性和区域方面)工作并不良好。

[0020] 本文中描述用于使用弱监督为图像配字幕的技术。在一个或多个实现方式中，获取关于目标图像的弱监督数据并且使用其提供补充被获得用于图像配字幕的全局图像概念的细节信息。弱监督数据是指没有被紧密地监管并且可能包括误差的噪声数据。给定目标图像，可以从弱注释的图像的不同的源(诸如在线社交网络、图像共享站点和图像数据库)来采集在视觉上相似的图像的弱监督数据。通常，在线发布的图像包括由用户添加的标签、标题、标注和短描述形式的“弱”注释。通过提取和聚合在弱注释图像的不同的源中发现的在视觉上相似的图像的关键词来生成目标图像的弱监督数据。然后，在图像配字幕分析期间采用弱监督数据中包括的关键词来调制被应用于概率分类的权重。因此，取决于弱监

督数据来计算用于预测图像配字幕的词语的概率分布。

[0021] 在实现方式中，图像配字幕框架基于神经网络和机器学习。给定目标图像，应用特征提取技术以得到描述图像的“要点”的全局图像概念。例如，可以使用预先训练的卷积神经网络(CNN)来使用全局描述性术语对图像编码。CNN产生反映全局图像概念的视觉特征矢量。然后，将所得到的关于全局图像概念的信息馈送到语言处理模型中，语言处理模型操作以在概率上生成图像的描述性字幕。比如，可以将视觉特征矢量馈送到循环神经网络(RNN)中，RNN被设计成实现语言建模和语句生成技术。RNN被设计成基于根据多个迭代中的权重因子计算的概率分布来迭代地预测用于组合作为目标图像的字幕的词语的序列。在这一上下文中，弱监督数据通过调制在模型中施加的权重因子来向RNN通知说明附加细节信息的操作。

[0022] 本文档中描述的使用弱监督为图像配字幕的技术使得能够以更大复杂性和精度来生成图像字幕。根据弱监督注释得到的关键词可以扩展用于特定图像的配字幕的词语的库并且相应地调制词语概率。因此，扩展候选字母的集合以包括根据弱监督数据得到的具体的对象、属性和术语。总之，其产生更加准确并且可以描述图像的非常具体的方面的更好的字幕。

[0023] 在以下讨论中，首先描述可以在本文中描述的技术中使用的示例环境。然后描述可以在示例环境以及其他环境中执行的示例过程和实现细节。因此，示例过程和细节的执行不限于示例环境，并且示例环境不限于示例过程和细节的执行。

[0024] 示例环境

[0025] 图1是可操作以采用本文中描述的技术的示例实现方式中的环境100的图示。图示的环境100包括计算设备102，计算设备102包括处理系统104、一个或多个计算机可读存储介质106和客户端应用模块108，其中处理系统104可以包括一个或多个处理设备，客户端应用模块108嵌入在计算机可读存储介质106上并且经由处理系统104可操作以实现本文中描述的对应功能。在至少一些实施例中，客户端应用模块108可以表示可操作以访问各种基于web的资源(例如，内容和服务)的计算设备的浏览器。客户端应用模块108还可以表示具有可操作以访问基于web的资源(例如，网络启用的应用)、浏览因特网、与在线提供商交互等的集成的功能的客户侧部件。

[0026] 计算设备102还可以包括或者利用图像搜索工具110，图像搜索工具110表示可操作以实现以上和以下描述的图像搜索的技术的功能。比如，图像搜索工具110可操作以访问和使用各种可用图像源来寻找匹配查询条目的候选图像。图像搜索工具110还表示执行各种动作以促进基于本文中讨论的图像帧的上下文的搜索的功能，诸如图像帧附近的内容的分析、得到查询条目以用作搜索参数的文本分析、命名的实体标识、和/或查询的构造等，这里仅给出几个示例。基于经由图像搜索工具110构造的图像搜索发现的图像可以经由客户端应用模块108或另一应用输出的用户界面111被暴露，图像搜索工具110被配置成针对另一应用提供用于外推性库存图像搜索的功能。

[0027] 图像搜索工具110可以实现为软件模块、硬件设备，或者使用软件、硬件、固件、固定逻辑电路系统等来实现。如图所示，图像搜索工具110可以实现为计算设备102的独立的部件。另外地或者备选地，图像搜索工具110可以被配置作为客户端应用模块108、操作系统、或者其他设备应用的部件。例如，图像搜索工具110可以被提供作为用于浏览器的插入

式和/或可下载脚本。图像搜索工具110还可以表示包含在网页、web应用、或者通过服务提供商可获得的其他资源中或者经由网页、web应用、或者通过服务提供商可获得的其他资源可访问的脚本。

[0028] 计算设备102可以被配置作为任意合适的类型的计算设备。例如，计算设备可以被配置作为台式计算机、膝上型计算机、移动设备(例如，假定手持式配置，诸如平板计算机或移动电话)、平板计算机等。因此，计算设备102可以在具有基本存储器和处理器资源的全资源设备(例如，个人计算机、游戏操纵台)到具有有限存储器和/或处理资源的低资源设备(例如，移动设备)之间变化。另外，虽然示出了单个计算设备102，但是计算设备102可以代表关于图12进一步描述的“在云上”执行操作的多个不同的设备。

[0029] 环境100还描绘被配置成通过网络114(诸如因特网)与计算设备102通信以提供“基于云的”计算环境的一个或多个服务提供商112。通常而言，服务提供商112被配置成使得各种资源116通过网络114可用于客户。在一些场景中，用户可以标记用于访问来自提供商的对应资源的账户。提供商可以在授予对账户和对应资源116的访问之前认证用户的证书(例如，用户名和密码)。可以使得其他资源116自由地可用(例如，没有基于认证或账户的访问)。资源116可以包括通常由一个或多个提供商通过网络可获得的服务和/或内容的任意合适的组合。服务的一些示例包括但不限于照片编辑服务、web开发和管理服务、协作服务、社交网络服务、消息传输服务、广告服务等。内容可以包括文本、视频、广告、音频、多媒体流、动画、图像、web文档、网页、应用、设备应用等的各种组合。

[0030] web应用118表示可以经由服务提供商112可访问的一个具体种类的资源116。可以使用浏览器或其他客户端应用模块108在网络114上操作web应用118以获得和运行web应用的客户侧代码。在至少一些实现方式中，由浏览器(或者其他客户端应用模块108)来提供用于web应用118的执行的运行时间环境。因此，从服务提供商可获得的服务和内容在一些场景中作为web应用可访问。

[0031] 服务提供商还被图示为包括被配置成根据本文中描述的技术来提供图像数据库122的图像服务120。图像服务120可以操作以搜索不同的图像源124并且分析和监管从图像源可获得的图像126以产生图像数据库122。图像数据库122表示可以被客户端访问以插入网页、词语文档、呈现和其他内容中的经监管的图像的服务器侧储存库。图像服务120例如可以被配置成提供客户端/应用访问以经由相应图像搜索工具110来使用图像数据库122。作为示例，图像服务120被描绘为实现搜索应用编程界面(搜索API)128，客户端/应用通过搜索API 128能够提供经由图像服务120来定义和发起搜索的搜索请求。

[0032] 图像服务120可以另外包括字幕生成器130。字幕生成器130表示可操作以实现以上和以下描述的图像配字幕技术的功能。通常而言，字幕生成器130被设计成分析图像以生成图像的自然语言描述，诸如“一男子乘着冲浪板在波浪的顶部”。在实现方式中，字幕生成器130依赖于神经网络和机器学习，其细节在下面关于图3和图4来讨论。在实现方式中，可以使用卷积神经网络(CNN)来使用全局描述性术语对图像编码，然后将其馈送到循环神经网络(RNN)中，RNN被设计成实现语言建模和语句生成技术。根据本文档中描述的发明原理，字幕生成器130被配置成以多种方式来增强用于图像配字幕的CNN图像特征和RNN建模的组合。作为说明，可以使用根据弱注释图像源得到的图像细节关键词来补充用于字幕生成的RNN的操作，如下面关于图5和图6讨论的。另外地或者备选地，字幕生成器130可以输出矢量

词语空间中的词语的表示，而非直接输出词语，如关于图7和图8讨论的。另外，字幕生成器130可以被配置成应用语义注意模型以基于上下文来选择RNN中的不同节点的不同关键词，如关于图9至图11讨论的。

[0033] 图2在200处总体上描绘示出根据一个或多个实现方式的字幕生成器130的细节的图。在本示例中，字幕生成器130被实现作为图像服务120的部件。应当注意，字幕生成器130也可以用其他方式来配置，诸如作为独立的服务、作为图像搜索工具110的部件、或者作为被部署给客户端、图像源和/或其他实体的单独的应用。字幕生成器130被描绘为包括图像分析模型202。图像分析模型202表示以各种方式处理图像的功能，包括但不限于特征提取、元数据解析、节距分析、对象检测等。图像分析模型202规定用于获取用于字幕分析的图像的相关关键词和描述的算法和操作。比如，图像分析模型202可以反映依赖于图像配字幕的卷积神经网络(CNN)和循环神经网络(RNN)的定义、处理和参数。为了增强图像配字幕，字幕生成器130另外被配置成单独地或者以任何组合一起使用弱监督数据204、词语矢量表示206和/或语义注意模型208，如下面更加详细地讨论的。

[0034] 在考虑到示例环境的情况下，现在考虑根据一个或多个实现的用于图像配字幕的技术的一些示例细节的讨论。

[0035] 图像配字幕实现细节

[0036] 这一部分描述根据一个或多个实现方式的使用增强的图像配字幕的一些示例细节。关于图3至图11的一些示例过程、场景和用户界面来讨论这些细节。本文中讨论的过程表示为规定由一个或多个设备来执行操作的框的集合，并且不一定限于被示出用于由相应框来执行这些操作的顺序。这些过程的各个方面可以用硬件、固件、或软件、或者其组合来实现。这些过程的一些方面可以经由一个或多个服务器(诸如经由服务提供商112)来实现，服务提供商112维持和提供经由图像服务120等对图像数据库122的访问。这些过程的各个方面也可以由合适地配置的设备(诸如，包括或利用图像搜索工具110和/或客户端应用模块108的图1的示例计算设备102)来执行。

[0037] 总体上，以上和以下关于示例描述的功能、特征和概念可以在本文档中描述的示例过程的上下文中来采用。另外，关于本文档中的不同的附图和示例描述的功能、特征和概念可以在彼此之间交换并且不限于在具体附图或过程的上下文中的实现。另外，本文中与不同的相应的过程和对应的附图相关联的框可以一起应用和/或按照不同的方式来组合。因此，本文中关于不同的示例环境、设备、部件、附图和过程描述的单独的功能、特征和概念可以用任意合适的组合来使用并且不限于本描述中的枚举的示例表示的具体组合。

[0038] 图像配字幕框架

[0039] 图3在300处总体上描绘图像配字幕框架301的示例实现方式。在本示例中，图像配字幕框架301采用机器学习方法来生成配有字幕的图像。因此，由图像配字幕框架301来来获得训练数据302，训练数据302要用于训练模型，模型然后用于形成字幕。用于在类似的场景中训练模型的技术(例如，图像理解问题)可以依赖于用户手动标记图像从而形成训练数据302。也可以使用机器学习来训练模型，机器学习使用自动地并且没有用户交互地可执行的技术。

[0040] 在图示示例中，训练数据302包括图像304和相关联的文本306，诸如与图像304相关联的字幕或元数据。然后使用提取模块308使用自然语言处理来提取结构化语义知识

310，例如“<主语，属性>，图像”和“<主语，谓语，宾语>，图像”。提取也可以包括构造的语义到图像内的对象或区域的定位。结构化语义知识310可以用于使图像与和在视觉上相似的图像相关联的数据匹配(例如，配字幕)，并且还可以用于寻找匹配元数据的集合的具体字幕的图像(例如，搜索)。

[0041] 然后将图像304和对应的结构化语义知识310传递给模型训练模块312。模型训练模块312被图示为包括机器学习模块314，机器学习模块314表示采用机器学习(例如神经网络、传统的神经网络等)使用图像304和结构化语义知识310来训练图像分析模块202的功能。训练模型316以定义结构化语义知识310中包括的文本特征与图像中的图像特征之间的关系(例如视觉特征矢量)。

[0042] 然后由字幕生成器将图像分析模型202用于处理输入图像316并且生成配有字幕的图像318。配有字幕的图像318例如可以包括文本标签和描述以定义图像108的概念，甚至在其中输入图像316不包括任何文本的实例中。相反，字幕生成器130使用图像分析模型202基于输入图像316的分析来生成适当的文本描述。然后由图像服务320使用配有字幕的图像318自动地并且没有用户干预地控制各种功能，诸如图像搜索、字幕和元数据提取、图像分类、可访问特征等。

[0043] 通常，图像配字幕框架301包括特征提取之后是基于特征来构造描述。可以针对由图像配字幕框架301反映的特征提取操作和描述构造操作二者采用各种不同的模型和方法。如先前指出的，图像配字幕框架301可以依赖于神经网络和机器学习。在实现中，使用卷积神经网络(CNN)来实现特征提取，然后调用循环神经网络(RNN)用于语言建模和语句构造。

[0044] 在这一上下文中，图4是总体上在400处描绘根据一个或多个实现的图像配字幕框架的细节的图。在此，框架401表示用于基于神经网络的图像配字幕的一般编码器解码器框架。框架基于神经网络和机器学习。给定目标图像316，应用特征提取技术以得到描述图像的“要点”的全局图像概念。例如，使用预先训练的卷积神经网络(CNN)402使用整体指示图像的要点的概念404对图像编码。CNN产生反映这些“全局”概念404的视觉特征矢量。然后将所得到的关于全局图像概念404的信息馈送到语言处理模型中，语言处理模型操作以在概率上生成图像的描述性字幕。比如，可以将视觉特征矢量馈送到循环神经网络(RNN)406中，RNN 406被设计成实现语言建模和语句生成技术。RNN 406被设计成基于根据多个爹地啊中的权重因子计算的概率分布来迭代地预测组合作为目标图像的字幕的词语的序列。如所表示的，RNN 406输出与图像316相关联的字幕、标签、语句和其他文本形式的描述408。这产生配有字幕的图像，如关于图3讨论的。

[0045] 图4还表示可以结合一般框架401使用的增强410。具体地，字幕生成器130可以使用弱监督数据204、词语矢量表示206和/或语义注意模型208作为由一般框架401提供的图像配字幕的增强410。每个增强410可以单独使用以补充一般框架401的配字幕。另外，可以采用多个增强的任意组合。下面进而讨论关于对一般框架401的增强410的细节。

[0046] 弱监督

[0047] 如先前指出的，可以获得关于目标图像的弱监督数据204并且使用其来提供补充被得到用于图像配字幕的全局图像概念404的详细信息。具体地，从弱注释图像的源、诸如社交网络站点、图像共享站点和图像的其他在线储存库来采集弱监督数据204。针对不同场

景中的图像配字幕，可以依赖于一个或多个源。被更新为这样的源的图像通常与由用户添加的标签、描述和其他文本数据相关联。由用户添加的这种文本数据被认为是“弱监督的”，因为用户可以包括可能与由图像传达的图像内容和全局概念无关或者少量相关联的“嘈杂的”条目，并且数据没有由服务提供商来细化或控制。弱注释在与通过传统图像标识和特征提取方法可实现的相比更深的理解水平提供关于图像的详细信息。因此，依赖于弱注释来生成表示低水平图像细节的关键词的集合(例如对象、属性、区域、通俗语义)，其可以用于扩展用于图像分析的库/词典并且补充被得到用于图像配字幕的全局图像概念404。

[0048] 在先前讨论的一般图像配字幕框架401中，使用预先训练的卷积神经网络(CNN)来编码图像。结果是被馈送到循环神经网络(RNN)中用于语句生成的视觉特征矢量。使用训练数据训练嵌入函数、循环神经网络和可选的卷积神经网络。RNN具体地被设计用于连续数据。在RNN中，每个输入节点具有隐藏状态h_i，并且对于每个隐藏状态，$h_i=f(x_i,h_{i-1})$，其中$f(\cdot)$是激活函数，诸如逻辑函数或双曲正切函数。换言之，每个节点的状态h_i取决于基于输入x_i和在先节点的状态h_{i-1}计算的激活函数。以这一方式，使用RNN迭代地计算每个输入节点的隐藏状态。另外，隐藏状态按照上述顺序将迭代从序列的开始传播到结束节点。图像配字幕框架401可以与各种不同架构的RNN集成。本文中省略了关于RNN架构的细节，因为本领域普通技术人员理解不同架构的实现，并且本文中描述的发明概念并不取决于所采用的具体RNN架构。

[0049] 在这一上下文中，图5在500处总体上描绘描绘用于使用弱监督为图像配字幕的框架的图。具体地，图5表示其中图4的一般框架401中的RNN 406被适配成依赖于弱监督数据204的场景。弱监督数据204可以从各种图像源124获得，如以上和以下描述的。例如，可以应用特征提取502过程以标识类似于来自图像源124中的至少一个图像源的目标图像的图像。进一步处理被标识为类似于目标图像的图像以从与类似图像相关联的弱注释中提取关键词。因此，特征提取502表示用于提取表示低水平图像细节的关键词的集合形式的弱监督数据204的功能，如以上讨论的。然后将弱监督数据204供应给RNN 406以通知图像配字幕分析，如图5中表示的。在一个方法中，向RNN供应根据弱注释图像得到的关键词的已过滤列表。列表可以通过以下方式来生成：根据关联准则对关键词集合评分和评级，并且选择大量顶部评级关键词以包括在已过滤列表中。已过滤列表基于频率、概率得分、权重因子或者其他关联准则被过滤。在实现中，可以供应关键词的整个集合用于在RNN中使用(例如未过滤的列表)。

[0050] 关键词的列表被配置成使关键词权重504与每个词语或短语相关联。关键词权重504反映可以在RNN中使用以预测用于相应地配字幕的词语序列的得分或概率分布。如图5中表示的，可以向RNN的每个节点中馈送顶部关键词的列表作为补充全局概念的附加数据。在这一点上，针对目标图像产生的关键词列表扩展用于得到目标图像的字幕的词典。另外，关键词权重504调制由RNN施加用于语言建模和语句构造的权重因子。因此，关键词权重504有效地改变用于由RNN实现的概率分类的词语概率从而有利于指示低水平图像细节的关键词。

[0051] 可以用以上解释的RNN的一般形式$h_i=f(x_i,h_{i-1})$来表达弱监督数据204的关键词权重504的影响。通常，给定每个图像v_i的关键词的集合$K_i=\{k_1,k_2,\cdots,k_k\}$，目的是如何采用$K_i$来生成$v_i$的字幕。具体地，构建模型以使用关键词用于训练和测试阶段二者。为此，针对每

个图像提取关键词并且将其聚合作为关键词的集合。然后，根据等式$K_e = \max(W_k K+b)$向RNN中的每个输入节点附上关键词的附加嵌入信息。在此，K_e是节点的关键词列表，W_k是控制关键词权重504的关键词的嵌入矩阵。对于每个输入词语w_i，在输入循环神经网络的每个位置处附加K_e，如图5中表示的。因此，被适配成采用弱监督的RNN可以表示为$h_i = f(x_i, h_{i-1}, K_e)$。在这一表达式中，激活函数$f(\bullet)$另外取决于嵌入的关键词列表$K_e$和对应的关键词权重504。

[0052] 在以上示例中，采用最大操作来从候选关键词的组中获取特征。也可以考虑其他操作，诸如求和，其可以增加输入层中的参数的总数。然而，在最大操作的情况下，被选择用于每个图像的关键词的数目可以不同，并且可以在分析中考虑大量潜在关键词而没有向输入层添加显著数目的参数。

[0053] 如所指出的，可以使用各种图像源124来获取弱监督数据。在实现中，图像源124包括通过网络可访问的图像的各种在线储存库，诸如社交网络站点、图像共享站点和监管的图像数据库/服务。用户当前频繁地使用这样的在线储存库共享图像和多媒体内容并且访问图像内容。从在线源可获得的图像通常包括可以被平衡以获取弱监督知识用于在配字幕中使用的标签和短描述。

[0054] 用于训练图像配字幕框架(例如训练字幕生成器)的训练图像的集合可以提供弱监督数据204的附加或替选源。在这一方法中，训练数据包括具有用于训练配字幕模型的分类器的对应字幕的图像的数据库。可以依赖于训练图像数据库作为源来发现类似于彼此的相关图像。接着，聚合相关图像的字幕作为用于图像配字幕的弱监督文本。当目标图像匹配相关图像的集合时，依赖于相关图像的字幕作为弱监督数据204用于目标图像的配字幕。

[0055] 在实现中，至少一些弱监督数据204可以直接从图像分析来得到。为此，训练不同概念或属性检测器以标识各种由弱注释图像提供的低水平图像细节。深度神经网络的相对近期的开发支持图像内的对象标识的显著改进。因此，能够训练图像分类器来标识一些种类的低水平图像细节，诸如具体对象、区域差异、图像属性等。代替直接使用这样的图像细节来生成候选字母，将所检测到的属性或概念馈送到图像字母框架中作为弱监督数据204以按照本文中描述的方式来通知图像配字幕。

[0056] 图6是根据一个或多个实现的其中可以采用弱监督数据用于图像配字幕的示例过程600的流程图。获取用于字幕分析的目标图像(框602)。例如，图像120可以实现字幕生成器130，如本文中描述的。图像服务120可以提供经由搜索API 128暴露的可搜索图像数据库122。字幕生成器130被配置成对图像执行字幕分析并且使用本文中描述的各种技术自动生成用于图像的字幕。经由字幕生成器130生成的配有字幕的图像318可以用各种方式来使用。例如，字幕可以促进使用自然语言查询经由搜索API 128进行的图像搜索。另外，字幕可以通过将字幕变换成听觉描述来促进对在视觉上受损的用户的可访问性从而向用户传达图像内容。

[0057] 为了产生图像字幕，向目标图像应用特征提取以生成对应于目标图像的全局概念(框604)。考虑各种类型的特征提取操作。通常，应用初始特征提取以得到描述图像的整个要点的全局概念404。可以经由CNN 402来执行初始特征提取，如先前注释的，然而也考虑其他得到全局图像概念404的技术。可以组合所得到的概念404以形成用作字母的进一步细化和选择的开始点的候选字母。因此，另外的细化可以另外地依赖于弱监督数据204，如本文

中描述的。

[0058] 具体地，将目标图像与来自弱注释图像的源的图像相比较以标识在视觉上相似的图像(框606)。考虑弱注释图像的各种源，先前已经给出了其示例。本文中描述的分析依赖于至少一个源，然而可以在一些场景中使用多个源。比较包括使用特征提取技术寻找具有类似于目标图像的特征的图像。认为与相似图像相关联的注释与目标图像的配字幕有关。

[0059] 因此，通过从在视觉上相似的图像中提取关键词来构建目标图像的关键词的集合(框608)，并且供应关键词的集合连同全局概念用于字幕生成(框610)。然后，使用关键词的集合调制针对语句构造应用的词语权重来针对目标图像生成字幕(框612)。在此，确定根据弱注释图像得到的关键词的列表并且将其作为弱监督数据204供应以便按照先前注释的方式来通知图像配字幕分析。由弱监督数据204表示的关键词权重504有效地调制被施加用于语言建模和语句生成的权重因子。可以经由RNN 406来实现产生字幕的语言建模和语句构造，如先前描述的，然而也可以考虑其他图像配字幕算法和技术。在任何情况下，由弱监督数据204反映的权重被应用于图像配字幕以相应地改变概率分类中的词语概率。因此，根据被建立用于关键词的权重因子在配字幕分析中考虑根据弱注释得到的指示低水平图像细节的关键词。

[0060] 词语矢量表示

[0061] 词语矢量表示206是可以用于增强一般图像配字幕框架401的附加特征。词语矢量表示206可以单独地或者结合先前描述的弱监督和/或以下章节中讨论的语法注意来使用。简言之，取代直接输出字幕分析的结果作为词语或者词语的序列(例如字幕或语句)，框架401被适配成输出语义词语矢量空间中的点。这些点构成词语矢量表示206，其反映语义词语矢量空间的上下文中的距离值。在这一方法中，将词语映射到矢量空间中并且将字幕分析的结果表示为捕获词语之间的语义的矢量空间中的点。在矢量空间中，类似的概念在概念的词语矢量表示中具有小的距离值。

[0062] 相比较而言，传统方法被设计成返回预测的词语或序列。比如，先前描述的RNN 406在传统上被配置成确定在固定词典/词汇表尚的每个节点处的概率分布。计算所计算的分布来对词语评分和评级。然后基于到节点的输入以及当前状态来选择最有可能的词语作为每个节点的输出。过程基于多个迭代来迭代地寻找顶部字幕。在此，由RNN使用的对象函数反映的策略是解决对应于种类的每个词语的分类问题。相对于固定词典/词汇表针对概率分类使用概率分布。因此，必须将字幕中的词语包含在词典中，词典大小通常很大以表示大量构造，并且如果改变词典，则必须整个重复分析。

[0063] 另一方面，通过词语矢量表示206，分析的输出是矢量空间中的点。这些点没有被绑定到具体的词语或单个词典。采用后处理步骤来将点映射到词语并且将词语矢量表示206变换成字幕。因此，将变换延迟到过程中的稍后阶段。其结果是，可以在过程中在稍后改变词典以选择不同语言、使用不同词语范围或不同数目的词语、产生新的术语等。另外，可以节省词语矢量表示206并且如果对词典做出变化则不需要重复先于后处理完成的步骤。

[0064] 图7在700描绘总体上图示用于图像配字幕的词语矢量表示的概念的示例图。具体地，图7表示捕获词语之间的语义的语义词语矢量空间702。在本示例中，语义词语矢量空间702在多维空间中具有对应于不同的词语或语句组合的轴。在这一上下文中，词语矢量704表示语义词语矢量空间702中的词语之间的距离值。给定用于分析问题和所选择的词典的

具体状态数据，可以将词语矢量704映射到最近的词语。这一方法提供在过程中稍后取决于上下文信息将词语矢量704映射到不同词语的灵活性。

[0065] 图8是根据一个或多个实现的其中采用词语矢量表示用于图像配字幕的示例过程800的流程图。获取用于字幕分析的目标图像(框802)，并且向目标图像应用特征提取以生成对应于图像的属性(框804)。例如，图像服务120可以实现先前描述的被配置成处理图像的字幕生成器130。另外，可以考虑各种类型的特征提取操作以检测特征、概念、对象、区域和与目标图像相关联的其他属性。

[0066] 向字幕生成器供应属性以发起字幕生成(框806)。比如，可以使用这些属性来得到关键词，关键词被供应给由字幕生成器130实现用于图像配字幕的图像分析模型202。使用关键词来构造和评估关键词的不同组合作为潜在字幕候选。作为分析的结果，在语义词语矢量空间中输出表示被形成作为属性的组合的语句中的语义关系词语的词语矢量(框808)。比如，图像分析模型202可以被适配成输出词语矢量表示206作为字幕分析的中间结果。词语矢量表示206可以对应于被影射到具体词语或具体词典的语义词语矢量空间702中的点。例如，由RNN实现的对象函数可以被适配成考虑语义词语矢量空间702中的距离而非词语语句的概率分布。下面讨论关于使用L-2距离和负采样来修改用于字幕分析的对象函数的一些细节。

[0067] 随后，将词语矢量变换成用于目标图像的字幕(框810)。重要的是，将词语矢量变换延迟到在RNN操作之后的后处理操作以得到词语矢量表示206。换言之，向根据RNN生成的输出应用后处理变换。词语矢量变换在经由RNN执行的字幕分析外部选择的词典/词汇表的上下文中发生。因此，生成词语矢量表示206的字幕分析没有取决于具体词典。

[0068] 如所指出的，可以使用距离和/或负采样修改用于字幕分析的对象函数来实现使用语义词语矢量空间的实现。关于L-2距离，构造典型的对象函数作为概率分类问题。例如，给定节点输入和当前状态，函数可以被设计成求解词语序列的对数相似性对象。这样的对数相似性对象可以表示为$\log\ p(W|V)=\Sigma_t \log\ p(w_t|V,w_0,w_1,\cdots,w_T)$。为了实现词语矢量表示206，将对象函数适配成取决于语义词语空间中的距离的成本函数。例如，可以将适配后的对象函数表示为$loss(W|V)=\sum_t \delta(l_t,p_t)dist(v_t,v_{p_t})$。在此，$p_t$表示预测的词语索引。通过这一对象函数，可以使用非常大的术语表。另外，可以使用一些未受监督的特征来初始化每个词语的特征，适配后的对象函数，明显地减小所涉及的特征的数目，因为参数的数目与特征的维度而非术语表的大小(典型的对象函数中的种类的总数)有关。

[0069] 以上L-2距离方法考虑每个节点处的对象函数中的当前词语。然而，对于每个节点，存在很多负样本(所有其他词语)。还可以将字幕分析适配成包括表示负样本的负采样分析。负采样向对象函数中注入表示到负样本的距离的成本。通过负采样，将对象函数设计成最小化相关的词语/矢量之间的距离以及到负样本的最大距离。在实现中，对于每个节点，随机选择不同于目标词语的N个词语，并且定义对象函数的损失因子作为$\log(1+\exp(-w_iVh_{i-1}))+\Sigma_n\log(1+\exp(w_nVh_{i-1}))$。在这一表达式中，$w_i$表示第i位置处的每个目标词语的嵌入。$w_n$表示第i目标词语的第n随机选择的负样本，$h_{i-1}$是位置i-1处的隐藏响应。因此，当目标词语接近随机选择的负样本时，负采样增加目标词语的成本。

[0070] 语义注意

[0071] 语义注意模块208是可以用于增强一般图像配字幕框架401的另一附加特征。语义

注意模块208可以单独地或者结合如先前描述的弱监督和/或词语矢量表示来使用。通常，语义注意模块208被实现用于可用术语的文集的关键词和概念的选择。本文中先前讨论的技术可以在循环神经网络中的每个节点处采用相同的关键词或特征集合。例如，可以针对RNN 406中的每个节点供应针对弱监督数据202得到的相同的关键词列表。然而，不同的词语/概念的关联可以在分析中的不同点处变化。语义注意模块208提供选择不同概念、关键词或监督信息以取决于上下文来生成下一词语的机制。

[0072] 广义上而言，语义注意模块208被配置成基于上下文来对候选关键词评级并且计算被馈送给RNN中的对应注意权重。在RNN中的每个节点处计算的状态信息被馈送回语义注意模块208中并且根据下一迭代的当前上下文对候选关键词重新评级。因此，当RNN经过时，用于RNN中的每个节点的具体关键词和权重变化。因此，图像配字幕模型留意每个迭代处的最相关的关键词。使用语义注意模块208用于图像配字幕实现了更加复杂的字幕，并且改善了生成的字幕的准确性。在图9至图11的以下讨论中提供关于图像配字幕的语义注意模型的另外的细节。

[0073] 对于上下文，现有为图像配字幕方法中有两个一般范例：自顶向下和自底向上。自顶向下范例从图像的“要点”开始并且将其变换成词语，而自底向上范例首先提出描述图像的各个方面的词语并且然后将其组合。在两个范例中采用语言模型以形成连贯语句。现有技术是自顶向下范例，其中基于循环神经网络存在从图像到语句的端到端公式化并且可以根据训练数据来学习重复网络的所有参数。自顶向下范例的限制之一是，很难注意精细的细节，这些细节在描述图像方面可能很重要。自底向上方法没有这一问题，因为它们没有对任何图像分辨率的操作。然而，它们存在其他问题，诸如缺乏用于从各个方面到语句的过程的端到端公式化。

[0074] 如本文中使用的，用于图像配字幕的语义注意是指提供在配字幕分析中的不同点处相关的语义上重要的对象的详细的连贯描述的能力。本文中描述的语义注意模块208能够：1)注意图像中的语义上重要的概念或者感兴趣区域，2)在多个概念上放置的注意的相对强度，以及3)根据任务状态在概念之间动态切换注意。具体地，语义注意模块208检测语义细节或“属性”作为使用自底向上方法的注意的候选，并且采用自顶向下部件引导应当在何处和何时激活注意。在循环神经网络(RNN)上构建模型，如先前讨论的。初始状态从自顶向下部件捕获全局概念。当RNN状态经过时，模型经由在两个网络状态和输出节点上强加的注意机制从自底向上属性得到反馈和交互。这一反馈使得算法能够不仅能够更加准确地预测词语，而且还能够产生现有的预测和图像内容之间的语义间隙的更加鲁棒的影响。反馈操作以组合循环神经网络的框架内的自顶向下和自底向上两个方法的视觉信息。

[0075] 图9是在900处总体上描绘根据一个或多个实现的用于图像配字幕的语义注意框架的图。如所指出的，语义注意框架组合用于图像字幕的自顶向下和自底向上方法。在所描绘的示例中，图像316表示为用于字幕分析的目标。给定目标图像316，调用传统的神经网络402来提取用于图像的自顶向下视觉概念。同时，应用特征提取902以提取低水平图像细节(区域、对象、属性等)。特征提取902可以实现为相同的卷积神经网络402的部分或者使用单独的提取部件来实现。在实现中，向弱注释图像的源应用特征提取902以按照先前描述的方式来得到弱监督数据204。特征提取902的结果是对应于低水平图像细节的图像属性904(例如关键词)的集合。如图9中表示的，语义注意模型208操作以组合自顶向下视觉概念和RNN

406中的低水平细节，RNN 406生成图像字幕。具体地，语义注意模型计算和控制用于属性904的注意权重906并且在每个迭代处将注意权重906馈送到RNN中。当RNN通过时，语义注意模型208获取关于字幕分析的当前状态和上下文的反馈908。采用这一反馈908关于循环神经网络迭代改变候选属性904的注意权重。因此，语义注意模型206引起RNN 406注意每个预测迭代的最相关的概念。

[0076] 图10是根据一个或多个实现的其中采用语义注意模型用于图像配字幕的示例过程1000的流程图。向目标图像应用特征提取以生成对应于目标图像的概念和属性(框1002)。特征提取可以按照本文中描述的各种方式来发生。特征提取可以依赖于CNN 402、提取模块302或者被设计成检测图像316的概念和属性的其他合适的部件。将概念和属性馈送到被配置成迭代地组合根据概念和属性得到的词语以在多个迭代中构造字幕的字幕生成模型中(框1004)。然后，根据被配置成基于与在先迭代中预测的词语的关联调制向多个迭代中的每个迭代的属性分配的权重的语义注意模型来构造字幕(框1004)。比如，根据一个或多个实现，可以采用关于图9讨论的语义注意框架用于图像配字幕。作为示例而非限制，语义注意模型208可以结合RNN 406来操作。备选地，可以采用其他用于语言建模和语句生成的迭代技术。在任何情况下，语义注意框架供应本文中描述的注意权重906，其用于控制字幕生成模型中的概率分类。在每个迭代处，使用注意权重906来预测字幕的序列以使模型专注于与该迭代最相关的具体的概念和属性。针对每个通过重新评估和调制注意权重906。

[0077] 图11是在1100处一般性地描绘根据一个或多个实现的语义注意框架的细节的图。具体地，图11表示使用用Φ表示的输入注意模型1102和用φ表示的输出注意模型1104二者的示例图像配字幕框架，其细节在下面讨论。在框架中，得到用于图像316的属性904。另外，采用CNN 402得到用v表示的图像316的视觉概念。与对应属性权重906耦合的属性904表示为属性检测$\{A_i\}$。将概念v和属性检测$\{A_i\}$注入到RNN(虚线箭头)中并且通过反馈908回路将其混合在一起。在这一框架内，由输入注意模型1102(Φ)和输出注意模型1104(φ)二者来强加对属性的注意。

[0078] 因此，根据输入图像得到自顶向下和自下向上特征。在实现中，使用来自分类卷积神经网络(CNN)的中间过滤响应来构建用v表示的全局视觉概念。另外，属性检测器的集合操作以得到最有可能在图像中出现的视觉属性的列表$\{A_i\}$。每个属性A_i对应于术语表集合或词典Y中的条目。

[0079] 所有的视觉概念和特征被馈送到循环神经网络(RNN)中用于字幕生成。由于RNN中的隐藏状态$h_t \in R^n$随着时间t发展，所以根据由状态h_t控制的概率矢量$p_t \in R^{|Y|}$从词典Y汲取字幕中的第t词语Y_t。所生成的词语Y_t在下一时间步骤中作为网络输入$x_{t+1} \in R^m$的部分被馈送到RNN中，其得到从h_t到h_{t+1}的状态过渡。来自v和$\{A_i\}$的视觉信息在生成x_t和p_t，时用作RNN的额外引导，其由在图11中表示的输入和输出模型Φ和φ来规定。

[0080] 与先前为图像配字幕方法相反，框架使用反馈908回路来使用和组合不同视觉信息源。使用CNN图像概念v作为初始输入节点x_0，其被期望向RNN给出图像内容的快速概述。一旦初始化RNN状态以包括整个视觉上下文，则RNN能够在随后的时间步骤中从任务相关的处理的$\{A_i\}$中选择具体的条目。具体地，通过以下等式来掌控框架：

[0081] $x_0 = \Phi_0(v) = W^{x,v}v$

[0082] $h_t = f(x_t, h_{t-1},)$

[0083] $Y_t \sim p_t = \varphi(h_t, \{A_i\})$

[0084] $x_t = \Phi(Y_{t-1}, \{A_i\})$，t＞0，

[0085] 这里，使用线性嵌入模型用于权重因子用$W^{x,v}$表示的初始输入节点x_0。在t＝0向v应用输入注意模型Φ以嵌入全局概念。h_t表示RNN的隐藏节点的状态，隐藏节点如先前描述地由激活函数f来掌管。输入Φ和输出φ注意模型被设计成基于当前模型状态来自适应地注意$\{A_i\}$中的某些认知线索，使得所提取的视觉信息与现有词语的解析和未来词语的预测最佳相关。例如，当前词语Y_t和概率分布p_t取决于输出φ模型和用表达式$Y_t \sim p_t = \varphi(h_t, \{A_i\})$反映的属性权重。同样，在t＝0之后的输入用$x_t = \Phi(Y_{t-1}, \{A_i\})$来表示，t＞0，并且取决于输入Φ模型、在先迭代中预测的词语Y_{t-1}以及属性$\{A_i\}$。RNN递归地操作并且这样注意的属性被反馈回到状态h_t并且与用v表示的全局信息集成。

[0086] 在针对t＞0的输入注意模型Φ中，基于其与在先预测词语Y_{t-1}的关联来向每个检测属性A_i分配得分α_t^i。由于Y_{t-1}和A_i都对应于词典Y中的条目，因此它们可以使用$R^{|Y|}$空间中的独热表示来编码，独热表示分别表示为y_{t-1}和y_i。作为在矢量空间中建模关联的普通方法，使用双线性函数来评估α_t^i。具体地，$\alpha_t^i \propto \exp(y_{t-1}^T \tilde{U} y^i)$，其中按照softmax方式在所有$\{A_i\}$上对指数归一化。矩阵$\tilde{U} \in R^{|Y|\times|Y|}$包含具有合理的术语表大小的任何Y的大量参数。为了减小参数大小，可以将独热表示投影到低维度语义词语矢量空间中(如以上关于图7和8讨论的)。

[0087] 令词语嵌入矩阵为$E \in R^{d\times|Y|}$，其中d＜＜|Y|。则，在先双线性函数变为$\alpha_t^i \propto \exp(y_{t-1}^T E^T U E y^i)$，其中U是dxd矩阵。一旦计算，则使用注意得分来调制不同属性上的注意的强度。根据以下表达式将所有属性的加权和连同在先词语从词语嵌入空间映射到X_t的输入空间：$x_t = W^{x,Y}(E_{yt-1} + diag(w^{x,A}) \sum_i \alpha_t^i E y^i)$。在此，$W^{x,Y} \in R^{m\times d}$是投影矩阵，diag(w)表示使用矢量w构造的对角矩阵，并且$w^{x,A} \in R^d$对语义词语矢量空间的每个维度中的视觉属性的相对重要性建模

[0088] 类似于输入注意模型来建模输出注意模型φ。然而，计算注意得分的不同集合，因为视觉概念在单个语句的分析和合成过程期间可以按照不同的顺序被注意。换言之，用于输入和输出模型的权重单独地计算并且具有不同的值。通过用于预测由当前状态捕获的Y_t的所有有用信息，关于h_t测量每个属性A_i的得分β_t^i，h_t用表达式$\beta_t^i \propto \exp(h_t^T V \sigma(E y^i))$来捕获。这里，$V \in R^{n\times d}$是双线性参数矩阵。σ表示将输入节点连接到RNN中的隐藏状态的激活函数，其在此用于确保在比较之前向两个特征矢量应用相同的非线性变换。

[0089] 另外，使用β_t^i来调制所有属性上的注意，并且使用其激活的加权和作为在确定分布p_t时的h_t的互补。具体地，分布由线性变换来生成，线性变换之后是被表示为$p_t \propto \exp(E^T W^{Y,h}(h_t + diag(w^{Y,A}) \sum_i \beta_t^i \sigma(E y^i)))$的softmax归一化。在本表达式中，$W^{Y,h} \in R^{d\times n}$是投影矩阵，$w^{Y,A} \in R^n$在RNN状态空间的每个维度上对视觉属性的相对重要性建

模。项实现参数减小的转置权重共享计策。

[0090] 每个图像的训练数据包括输入图像特征v、$\{A_i\}$和输出字幕词语序列$\{Y_t\}$。对于模型学习，目的是通过最小化训练集合上的损失函数来联合所有的RNN参数Θ_R学习所有的注意模型参数$\Theta_A=\{U,V,W^{*,*},w^{*,*}\}$。一个训练示例的损失定义为与注意得分$\{\alpha_t^i\}$和$\{\beta_t^i\}$上的调整项组合的所有词语的总的负的对数似然并且根据以下损失函数来表示：

[0091] $$min_{\Theta_A,\Theta_R} -\sum_t \log p(Y_t) + g(\alpha) + g(\beta)$$。

[0092] 这里，α和β是注意得分矩阵，其第(t;i)条目是权重α_t^i和β_t^i。调整函数g用于强迫向$\{A_i\}$中的每个属性支付的注意的完成以及任何具体时间步骤处的注意的稀少。这通过针对α使以下矩阵最小化(并且针对β一样)来进行：

[0093] $$g(\alpha) = \|\alpha\|_{1,p} + \|\alpha^T\|_{q,1} = \left[\sum_i\left[\sum_t \alpha_t^i\right]^p\right]^{\frac{1}{p}} + \sum_t\left[\sum_i(\alpha_t^i)^q\right]^{1/q}$$

[0094] p＞1的第一项处罚向在整个语句上累加的任何单个属性A_i支付的过多注意，并且0＜q＜1的第二项惩罚在任何具体时间的给多个属性的释放的注意。采用具有自适应学习速率的随机梯度下降算法来优化损失函数。

[0095] 考虑到以上示例细节、过程、用户界面和示例，现在考虑包括可以用于本文中描述的图像配字幕技术的一个或多个实现的各种部件和设备的示例系统的讨论。

[0096] 示例系统和设备

[0097] 图12在1200处一般性地图示示例系统，其包括表示可以实现本文中描述的各种技术的一个或多个计算系统和/或设备的示例计算设备1202。这通过图像服务120的包括来图示，图像服务120如以上描述地操作。计算设备1202可以是例如服务提供商的服务器、与客户端相关联的设备(例如客户端设备)、片上系统、和/或任何其他合适的计算设备或计算系统。

[0098] 示例计算设备1202被图示为包括处理系统1204、一个或多个计算机可读介质1206、以及在通信上耦合至彼此的一个或多个I/O接口1208。虽然没有示出，然而计算设备1202还可以包括系统总线或者将各种部件耦合至彼此的其他数据和命令传送系统。系统总线可以包括不同总线结构中的任何一个或组合，诸如存储器总线或存储器控制器、外围总线、通用串行总线、和/或使用各种总线架构中的任何总线架构的处理器或本地总线。也考虑各种其他示例，诸如控制线和数据线。

[0099] 处理系统1204表示使用硬件来执行一个或多个操作的功能。因此，处理系统1204被图示为包括可以被配置为处理器、功能框等的硬件元件1210。其可以包括硬件实现作为使用一个或多个半导体形成的专用集成电路或者其他逻辑器件。硬件元件1210不受形成其的材料以及其中采用的处理机制的限制。例如，处理器可以包括半导体和/或晶体管(例如电子集成电路(IC))。在这样的上下文中，处理器可执行指令可以是电子可执行指令。

[0100] 计算机可读存储介质1206被图示为包括存储器/存储装置1212。存储器/存储装置1212表示与一个或多个计算机可读介质相关联的存储器/存储能力。存储器/存储部件1212可以包括易失性介质(诸如随机存取存储器(RAM))和/或非易失性介质(诸如只读存储器(ROM)、闪存存储器、光盘、磁盘等)。存储器/存储部件1212可以包括固定介质(例如RAM、

ROM、固定硬盘驱动等)以及可移除介质(例如闪存存储器、可移除硬盘驱动、光盘等)。计算机可读介质1206可以用下面进一步描述的各种其他方式来配置。

[0101] 输入/输出接口1208表示使得用户能够向技术设备1202输入命令和信息并且还使得能够使用各种输入/输出设备来向用户和/或其他部件或设备呈现信息的功能。输入设备的示例包括键盘、光标控制设备(例如鼠标)、麦克风、扫描仪、触摸功能(例如电容或被配置成检测物理触摸的其他传感器)、相机(例如其可以采用视觉或非视觉波长、诸如红外频率来标识没有涉及触摸的运动、如姿势)等。输出设备的示例包括显示器设备(例如显示器或投影仪)、扬声器、打印机、网络卡、触觉响应设备等。因此,计算设备1202可以用下面进一步描述的各种方式被配置成支持用户交互。

[0102] 本文中在软件、硬件元件和程序模块的一般上下文中描述各种技术。通常,这样的模块包括执行具体任务或实现具体抽象数据类型的例程、程序、对象、元件、部件、数据结构等。本文中使用的术语“模块”、“功能”和“部件”通常表示软件、固件、硬件或者其组合。本文中描述的技术的特征独立于平台,这表示这些技术可以在具有各种处理器的各种商用计算平台上实现。

[0103] 所描述的模块和技术的实现可以存储在某种形式的计算机可读介质上或者在其上来传输。计算机可读介质可以包括可以由计算设备1202访问的各种介质。作为示例而非限制,计算机可读介质可以包括“计算机可读存储介质”和“计算机可读信号介质”。

[0104] “计算机可读存储介质”是指与仅信号传输、载波、或信号本身相对而言使得能够永久性和/或非暂态地存储信息的介质和/或设备。因此,计算机可读存储介质不包括信号本身或者信号承载介质。计算机可读存储介质包括在适合信息、诸如计算机可读指令、数据结构、程序模块、逻辑元件/电路或其他数据的存储的方法或技术中实现的硬件、诸如易失性和非易失性、可移除和非可移除介质和/或存储设备。计算机可读存储介质的示例可以包括但不限于RAM、ROM、EEPROM、闪存存储器或者其他存储器技术、CD-ROM、数字多功能盘(DVD)或其他光学存储装置、硬盘、磁盒、磁带、磁盘存储装置或其他磁性存储设备、或者其他存储设备、有形介质、或者适合存储期望信息并且可以由计算机来访问的制造品。

[0105] “计算机可读信号介质”是指被配置成诸如经由网络来向计算设备1202的硬件传输指令的信号承载介质。信号介质通常可以实施计算机可读指令、数据结构、程序模块、或者已调制数据信号中的其他数据,诸如载波、数据信号、或者其他传输机制。信号介质还包括任何信息递送介质。术语“已调制数据信号”表示其一个或多个特性按照使得能够对信号中的信息编码的形式来被设置或改变的信号。作为示例而非限制,通信介质包括有线和无线介质、诸如有线网络或直接有线连接、以及无线介质、诸如声学、RF、红外和其他无线介质。

[0106] 如先前描述的,硬件元件1210和计算机可读介质1206表示可以在一些实施例中采用以实现本文中描述的技术的至少一些方面、诸如执行一个或多个指令的硬件形式实现的模块、可编程设备逻辑和/或固定设备逻辑。硬件可以包括集成电路或片上系统、专用集成电路(ASIC)、现场可编程门阵列(FPGA)、复杂可编程逻辑器件(CPLD)以及硅或其他硬件的其他实现的部件。在本上下文中,硬件可以用作执行由指令定义的程序任务和/或由硬件实施的逻辑的处理设备以及用于存储用于执行的指令的硬件,诸如先前描述的计算机可读存储介质。

[0107] 上述的组合也可以用于实现本文中描述的各种技术。因此，软件、硬件或可执行模块可以实现为在某种形式的计算机可读存储介质上和/或由一个或多个硬件元件1210来实施的一个或多个指令和/或逻辑。计算设备1202可以被配置成实现对应于软件和/或硬件模块的具体的指令和/或功能。因此，作为软件的由计算设备1202可执行的模块的实现可以至少部分用硬件来实现，例如通过使用处理系统1204的计算机可读存储介质和/或硬件元件1210。指令和/或功能可以由一个或多个制造品(例如一个或多个计算设备1202和/或处理系统1204)可执行/可操作以实现本文中描述的技术、模块和示例。

[0108] 本文中描述的技术可以由计算设备1202的各种配置来支持，而不限于本文中描述的技术的具体示例。这一功能也可以全部或者部分通过使用分布式系统来实现，诸如经由下面描述的平台1216在“云”1214上实现。

[0109] 云1214包括和/或表示用于资源1218的平台1216。平台1216抽象云1214的硬件(例如服务器)和软件资源的功能。资源1218可以包括可以在远离计算设备1202的服务器上执行计算机处理的同时使用的应用和/或数据。资源1218还可以包括在因特网上和/或通过用户网络、诸如蜂窝或Wi-Fi网络提供的服务。

[0110] 平台1216可以抽象将计算设备1202与其他计算设备连接的资源和功能。平台1216还可以用于抽象资源的缩放以提供对应水平的缩放从而满足对于经由平台1216实现的资源1218的需求。因此，在互连的设备实施例中，本文中描述的功能的实现可以分布遍及系统1200。例如，功能可以部分在计算设备1202上以及经由抽象云1214的功能的平台1216来实现。

[0111] 总结

[0112] 虽然使用特定于结构特征和/或方法动作的语言描述了技术，然而应当理解，在所附权利要求中定义的主题不一定限于所描述的具体特征或动作。相反，具体的特征和动作被公开作为实现要求保护的主题的示例形式。

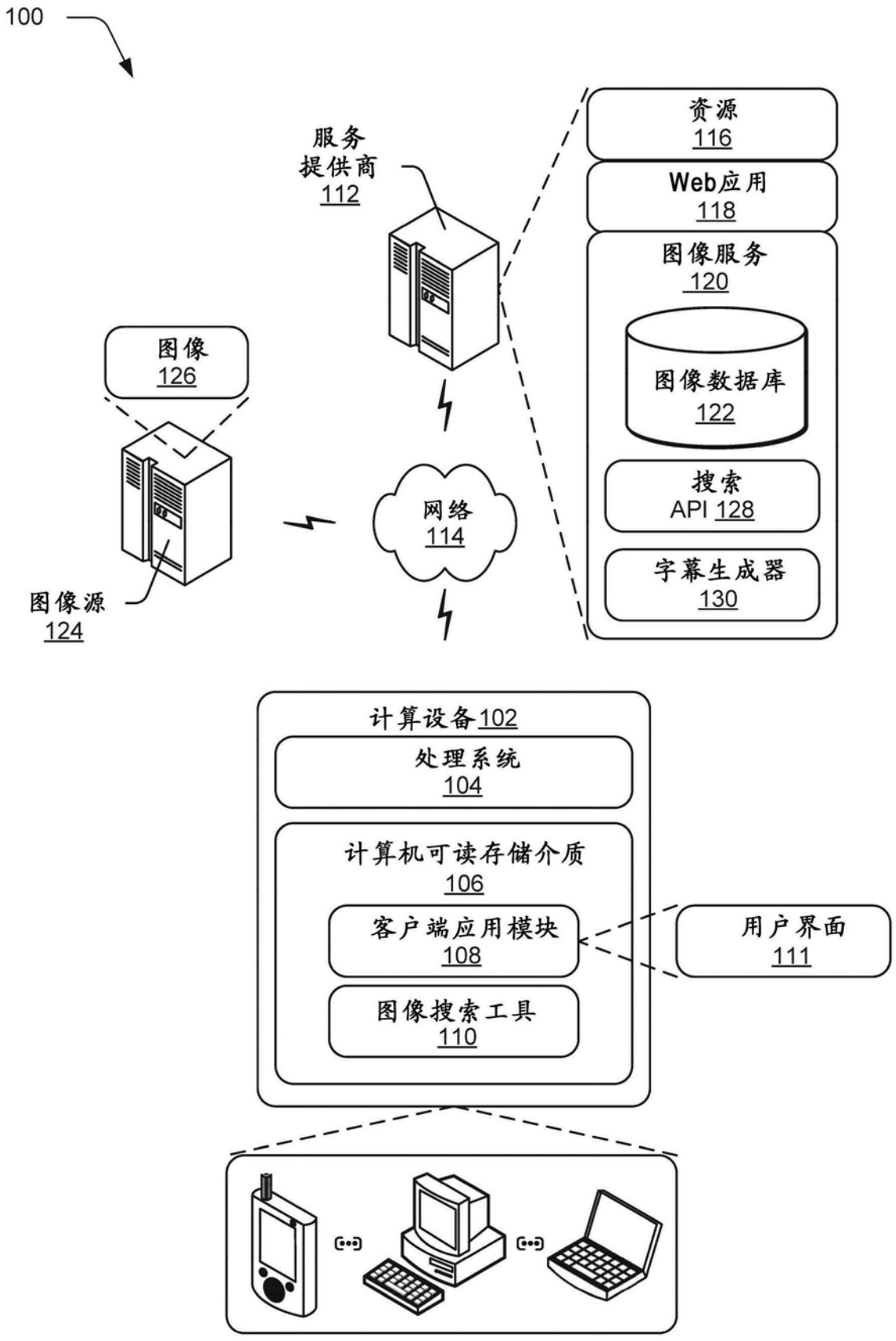

100
服务提供商 112
资源 116
Web应用 118
图像服务 120
图像数据库 122
搜索 API 128
字幕生成器 130
图像 126
图像源 124
网络 114
计算设备102
处理系统 104
计算机可读存储介质 106
客户端应用模块 108
图像搜索工具 110
用户界面 111

图1


200
图像服务 120
图像数据库 122
搜索 API 128
字幕生成器 130
图像分析模型 202
弱监督数据 204
词语矢量表示 206
语义注意模型 208

图2

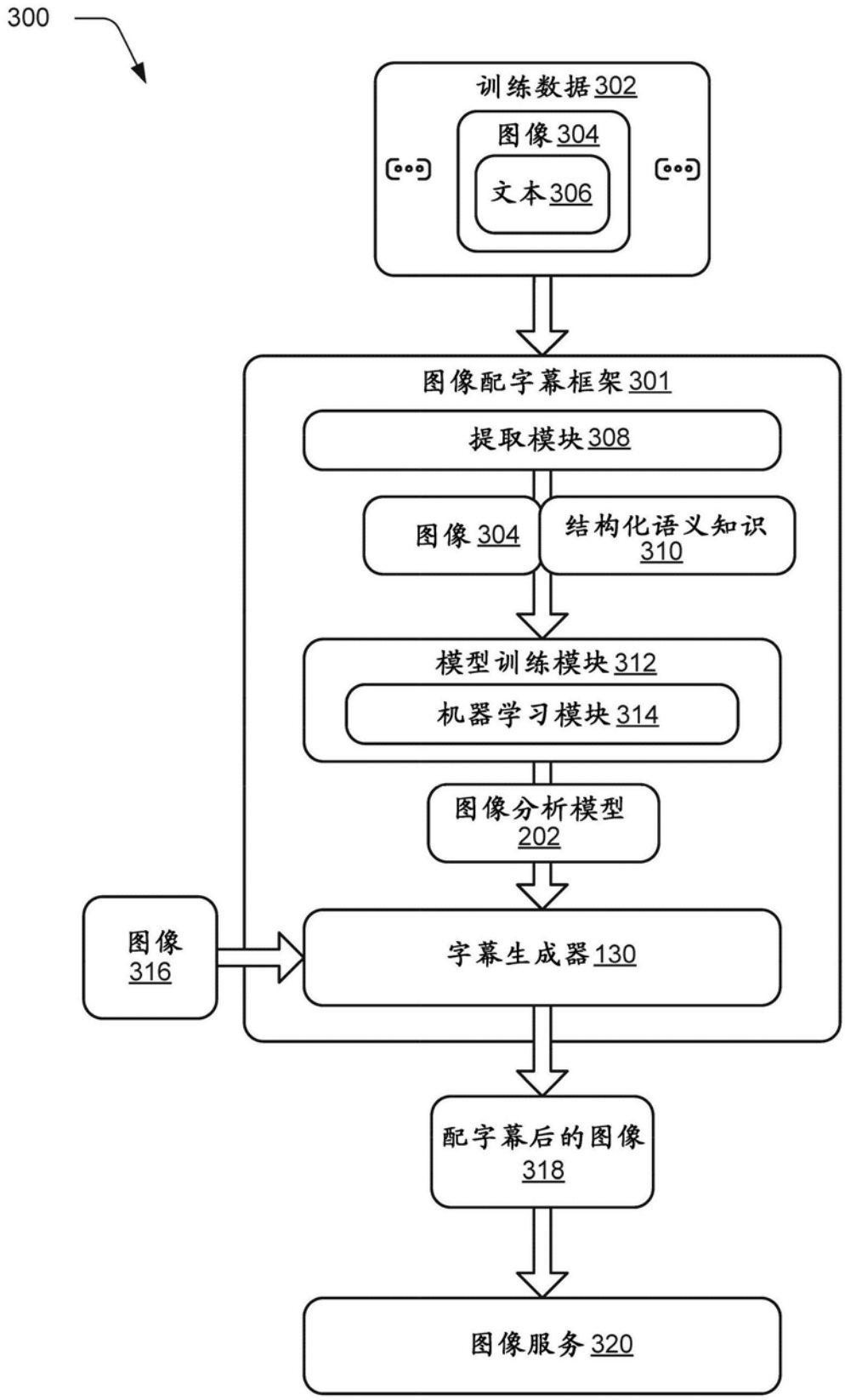

300
训练数据 302
图像 304
文本 306
图像配字幕框架 301
提取模块 308
图像 304
结构化语义知识 310
模型训练模块 312
机器学习模块 314
图像分析模型 202
图像 316
字幕生成器 130
配字幕后的图像 318
图像服务 320

图3

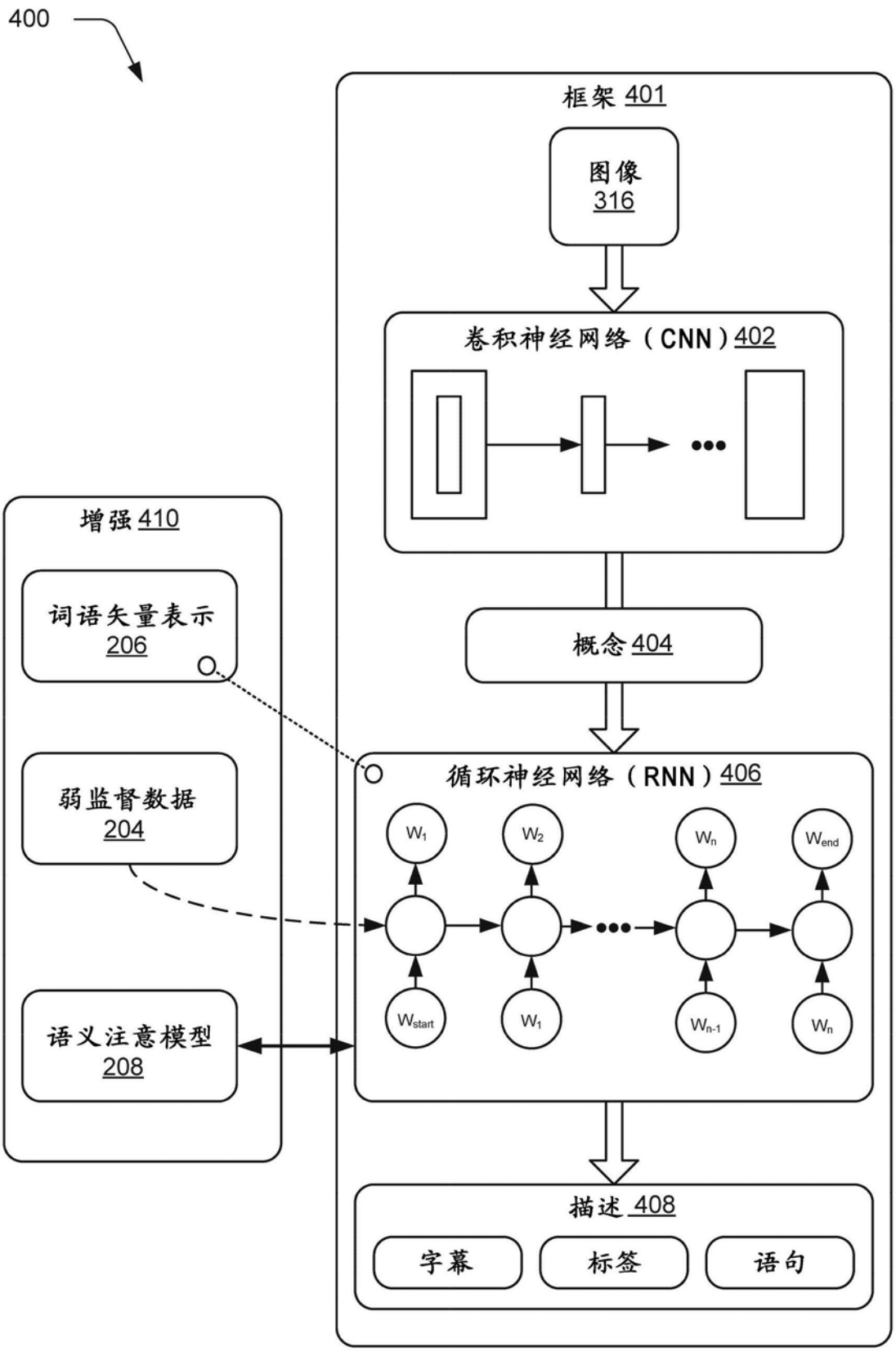

400
框架 401
图像 316
卷积神经网络（CNN）402
增强 410
词语矢量表示 206
概念 404
弱监督数据 204
循环神经网络（RNN）406
W1
W2
Wn
Wend
Wstart
W1
Wn-1
Wn
语义注意模型 208
描述 408
字幕
标签
语句

图4

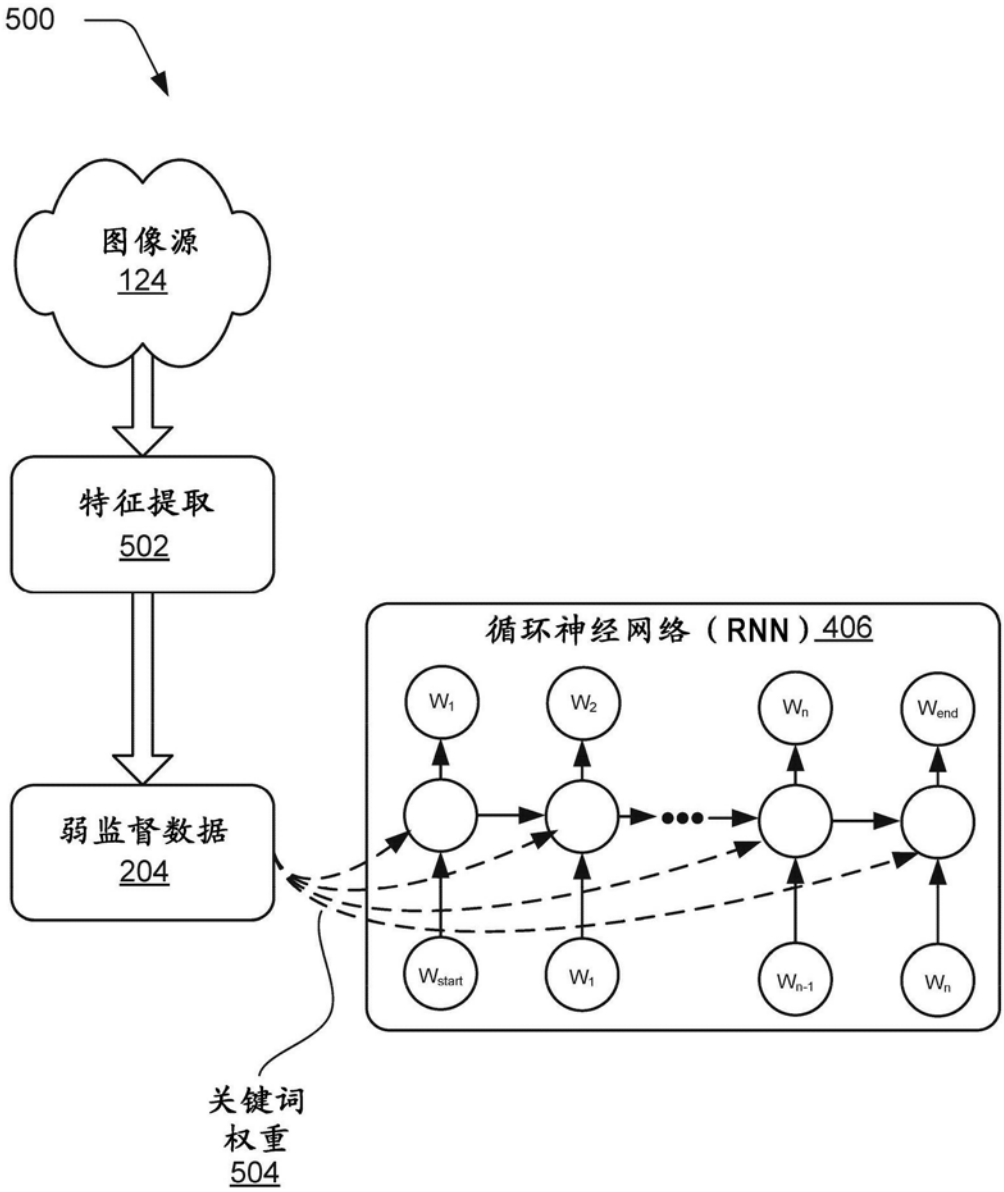

500
图像源
124
特征提取
502
弱监督数据
204
循环神经网络（RNN）406
W1
W2
Wn
Wend
Wstart
W1
Wn-1
Wn
关键词
权重
504

图5

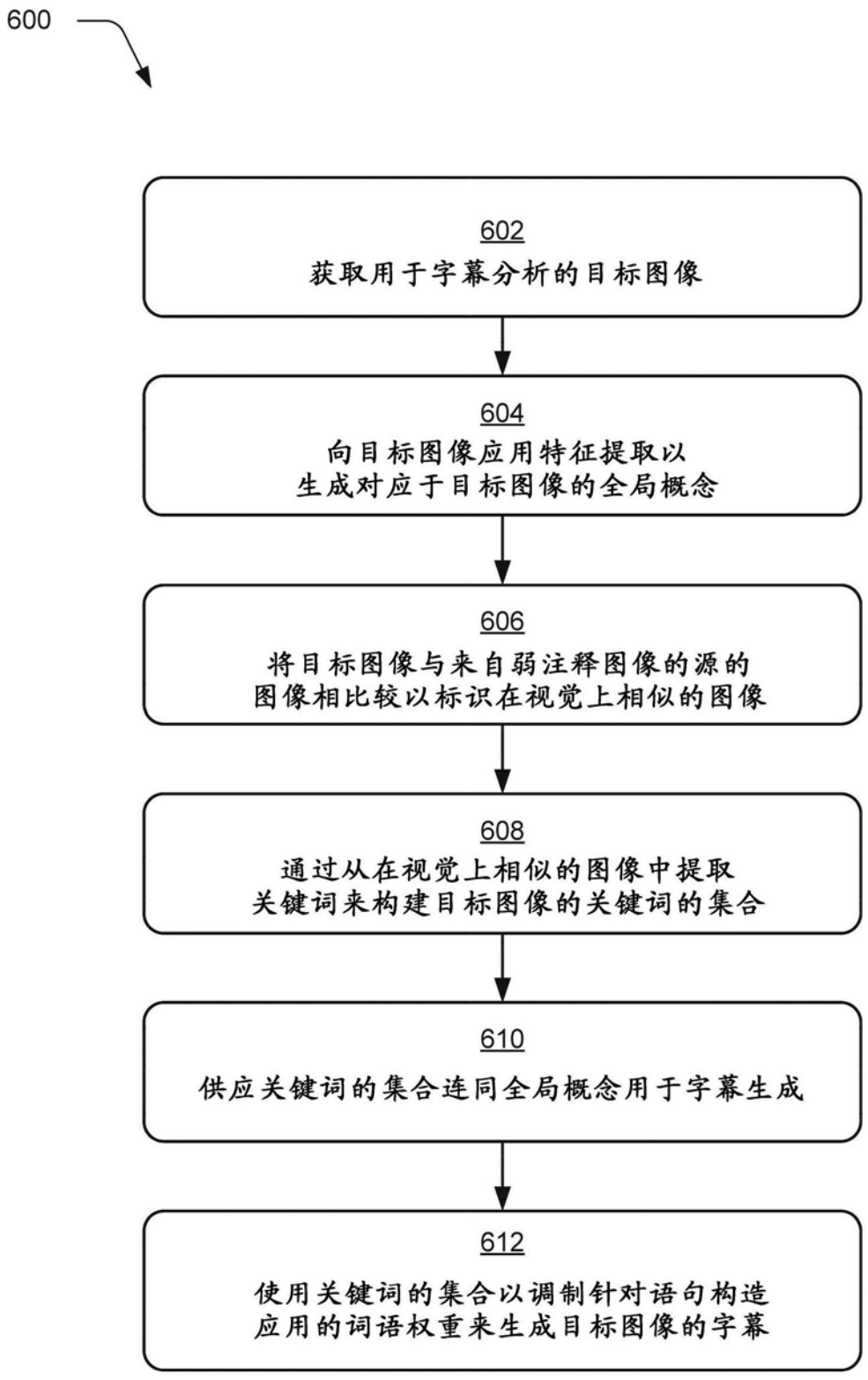

600
602
获取用于字幕分析的目标图像
604
向目标图像应用特征提取以
生成对应于目标图像的全局概念
606
将目标图像与来自弱注释图像的源的
图像相比较以标识在视觉上相似的图像
608
通过从在视觉上相似的图像中提取
关键词来构建目标图像的关键词的集合
610
供应关键词的集合连同全局概念用于字幕生成
612
使用关键词的集合以调制针对语句构造
应用的词语权重来生成目标图像的字幕

图6

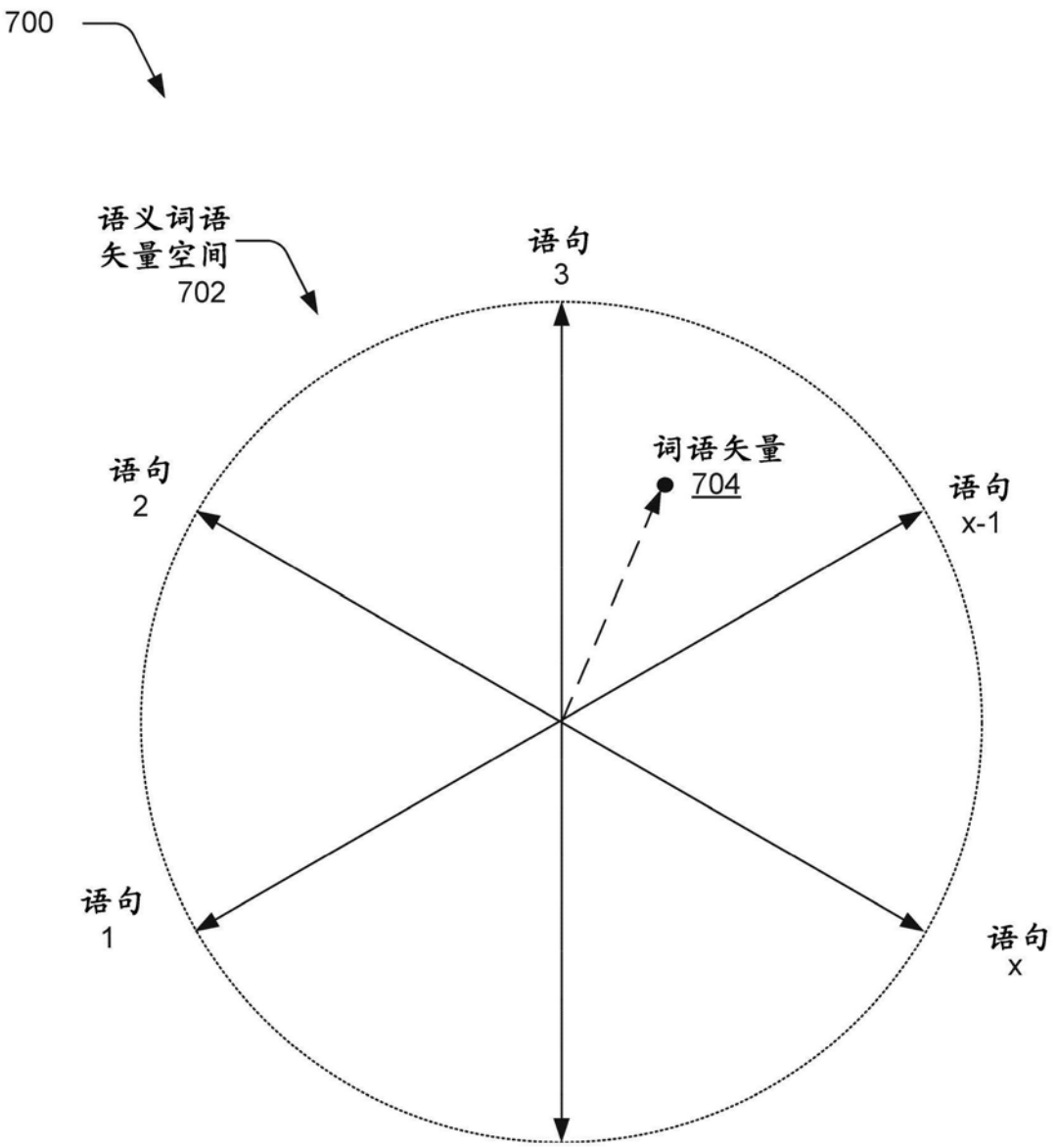


图7

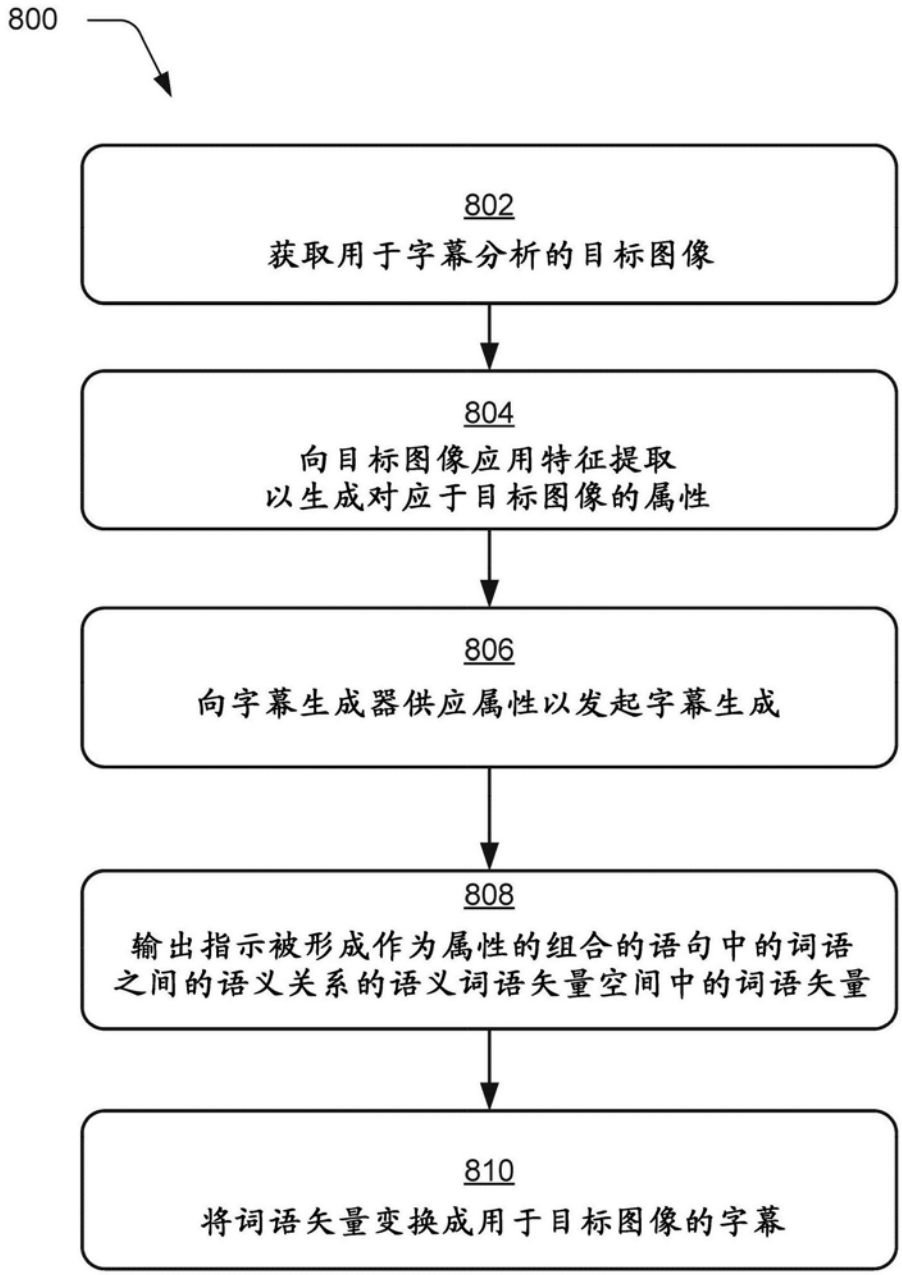

800
802
获取用于字幕分析的目标图像
804
向目标图像应用特征提取
以生成对应于目标图像的属性
806
向字幕生成器供应属性以发起字幕生成
808
输出指示被形成作为属性的组合的语句中的词语
之间的语义关系的语义词语矢量空间中的词语矢量
810
将词语矢量变换成用于目标图像的字幕

图8

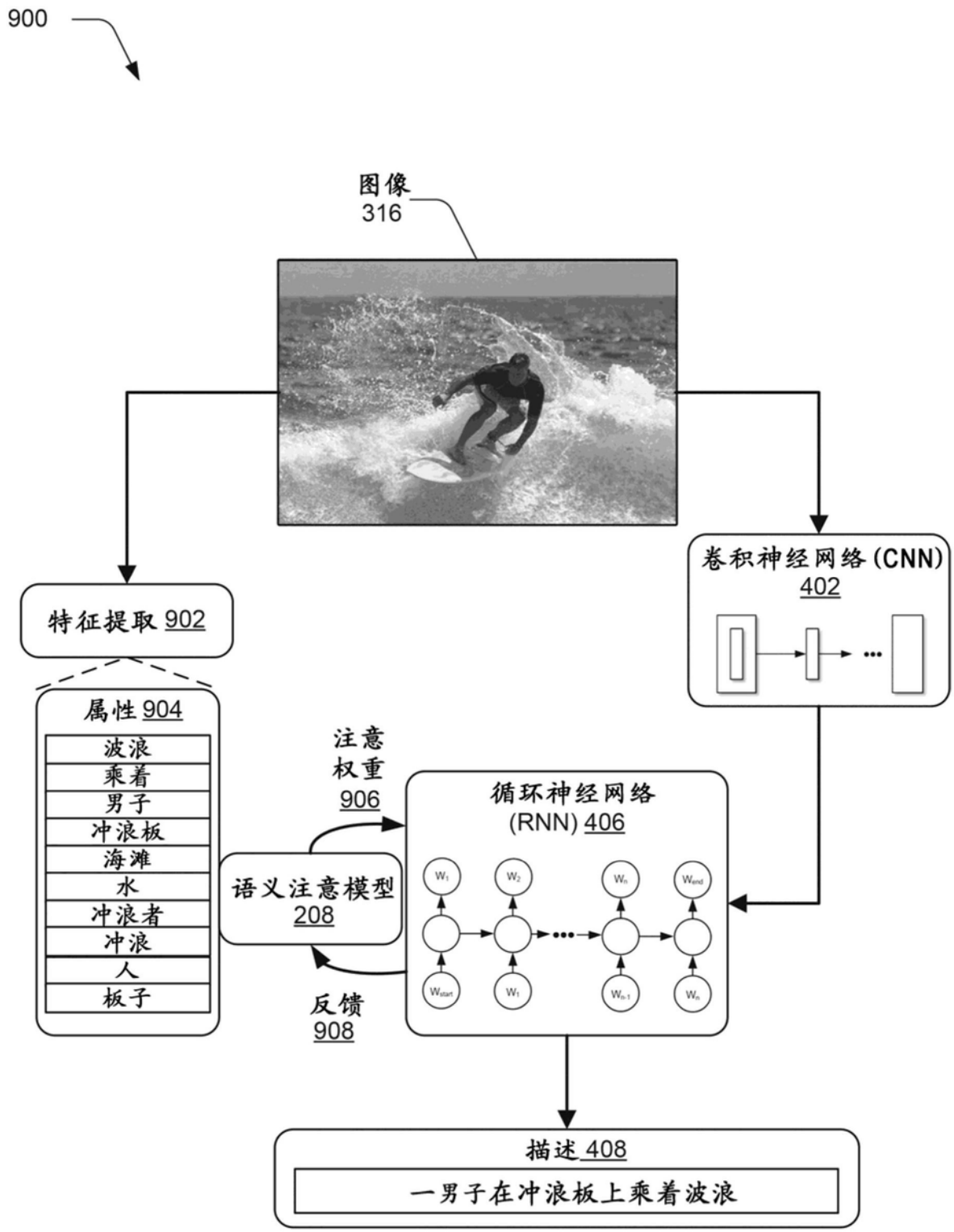

900
图像
316
卷积神经网络(CNN)
402
特征提取 902
属性 904
波浪
乘着
男子
冲浪板
海滩
水
冲浪者
冲浪
人
板子
注意
权重
906
语义注意模型
208
循环神经网络
(RNN) 406
反馈
908
描述 408
一男子在冲浪板上乘着波浪

图9

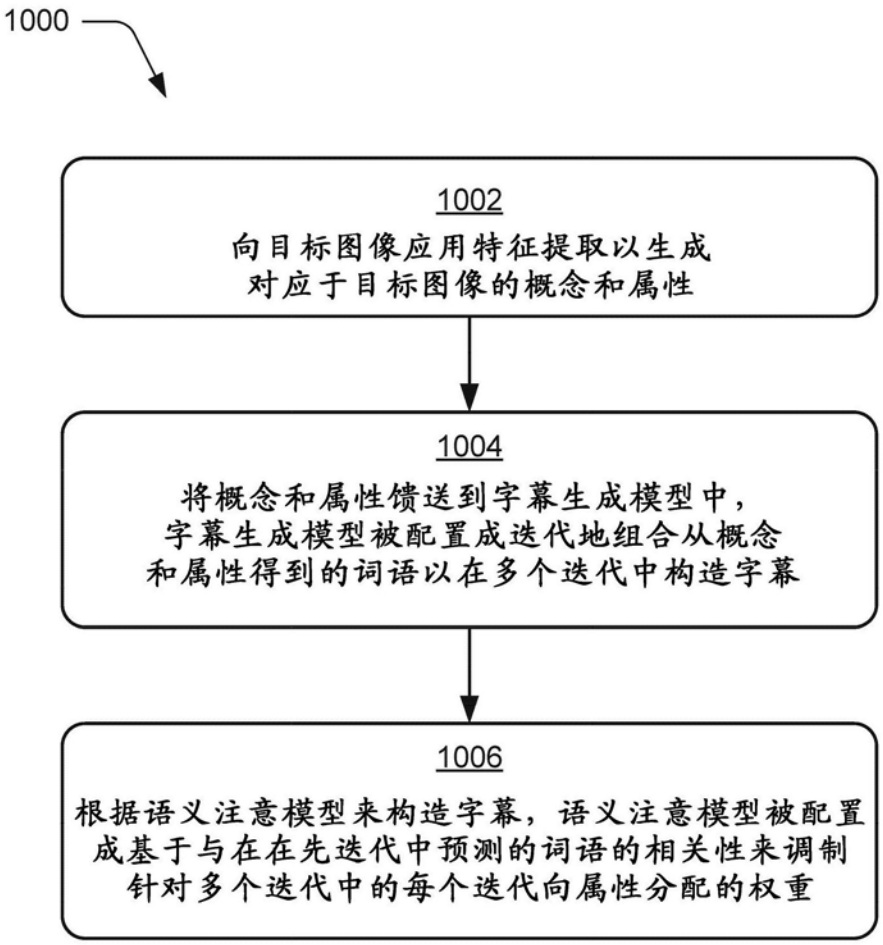

1000
1002
向目标图像应用特征提取以生成
对应于目标图像的概念和属性
1004
将概念和属性馈送到字幕生成模型中，
字幕生成模型被配置成迭代地组合从概念
和属性得到的词语以在多个迭代中构造字幕
1006
根据语义注意模型来构造字幕，语义注意模型被配置
成基于与在在先迭代中预测的词语的相关性来调制
针对多个迭代中的每个迭代向属性分配的权重

图10

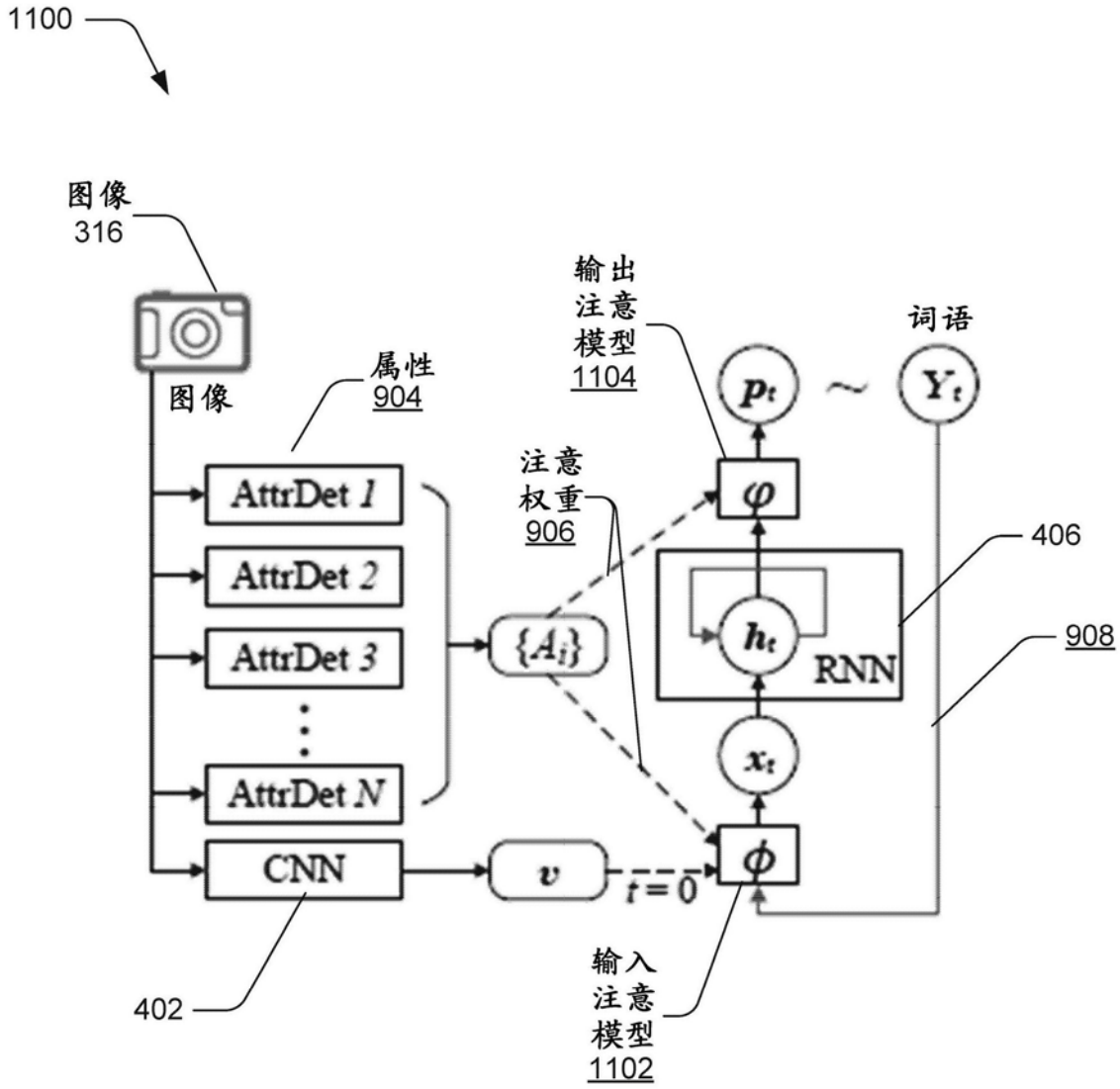

1100
图像
316
图像
属性
904
AttrDet 1
AttrDet 2
AttrDet 3
AttrDet N
CNN
402
{Ai}
v
t = 0
注意
权重
906
输出
注意
模型
1104
词语
pt
~
Yt
φ
ht
RNN
406
908
xt
ϕ
输入
注意
模型
1102

图11

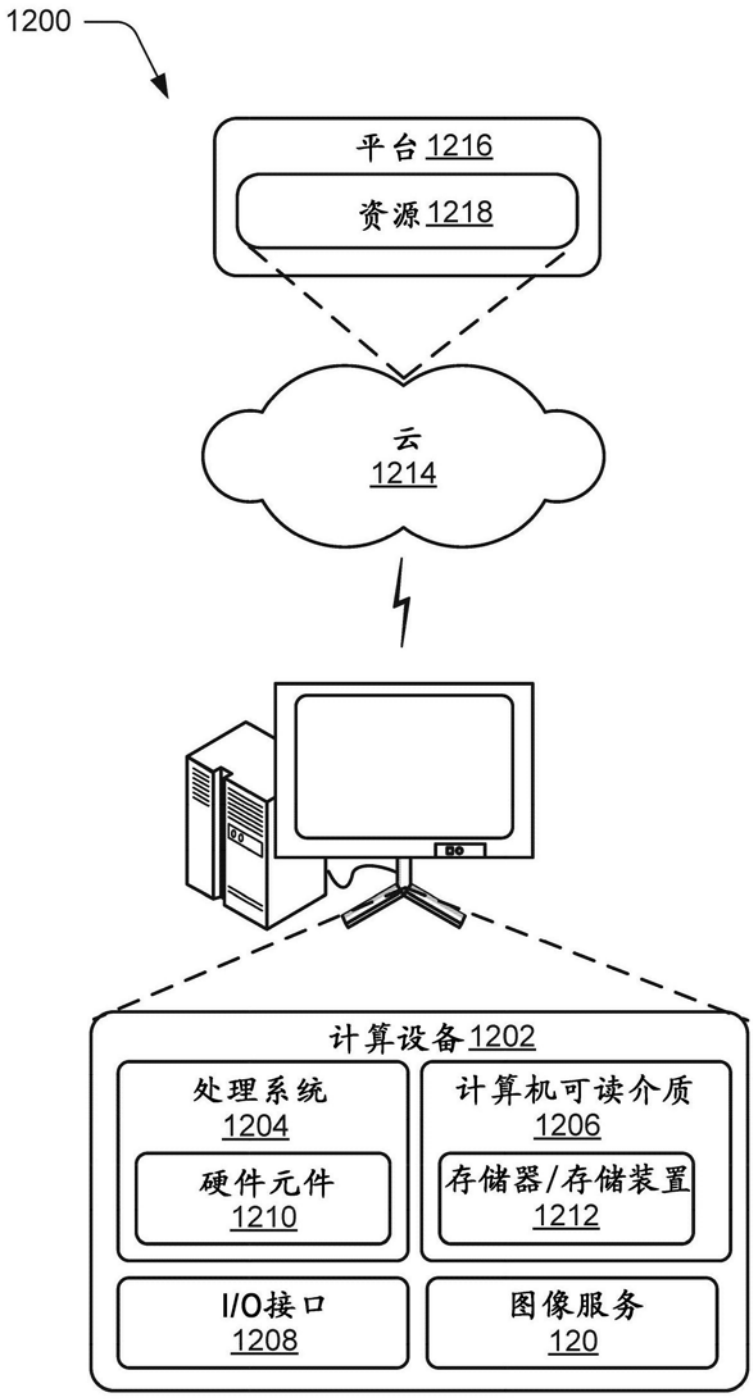

1200
平台 1216
资源 1218
云
1214
计算设备 1202
处理系统
1204
硬件元件
1210
计算机可读介质
1206
存储器/存储装置
1212
I/O接口
1208
图像服务
120

图12