(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115512005 A
(43)申请公布日 2022.12.23

(21)申请号 202211005409.3

(22)申请日 2022.08.22

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 刘志广 柏昊立 孟笑君 李文涛 谢念 王靓伟 侯璐 蒋欣

(74)专利代理机构 深圳市深佳知识产权代理事务所(普通合伙) 44285
专利代理师 李杭

(51)Int.Cl.
G06T 11/60(2006.01)
G06N 3/04(2006.01)
G06T 9/00(2006.01)
G06V 10/44(2022.01)
G06V 10/74(2022.01)
G06V 30/148(2022.01)
G06V 30/19(2022.01)
G06V 10/82(2022.01)

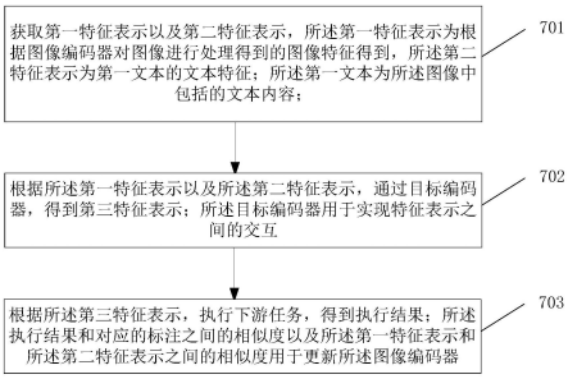
权利要求书3页 说明书30页 附图10页

(54)发明名称

一种数据处理方法及其装置

(57)摘要

一种数据处理方法,应用于包含文字的图像的处理,涉及人工智能领域,包括:获取第一特征表示以及第二特征表示,第二特征表示为第一文本的文本特征;第一文本为图像中包括的文本内容;根据第一特征表示以及第二特征表示,通过目标编码器,得到第三特征表示;第三特征表示用于执行下游任务;执行结果和对应的标注之间的相似度以及第一特征表示和第二特征表示之间的相似度用于更新图像编码器。本申请通过先双塔方式可以提升图文之间的对齐能力,再利用一个单塔结构进一步增强特征的交互学习能力。



CN 115512005 A

1.一种数据处理方法,其特征在于,包括:

获取第一特征表示以及第二特征表示,所述第一特征表示为根据图像编码器对图像进行处理得到的图像特征得到,所述第二特征表示为第一文本的文本特征;所述第一文本为所述图像中包括的文本内容;

根据所述第一特征表示以及所述第二特征表示,通过目标编码器,得到第三特征表示;所述目标编码器用于实现特征表示之间的交互;

根据所述第三特征表示,执行下游任务,得到执行结果;所述执行结果和对应的标注之间的相似度以及所述第一特征表示和所述第二特征表示之间的相似度用于更新所述图像编码器。

2.根据权利要求1所述的方法,其特征在于,所述第二特征表示为通过文本编码器对所述第一文本进行处理得到的文本特征;所述第一特征表示和所述第二特征表示之间的相似度还用于更新所述文本编码器。

3.根据权利要求1或2所述的方法,其特征在于,所述第一特征表示和所述第二特征表示之间的相似度与图像中所蕴含的文本语义信息和文本自身的语义信息之间的相似度有关。

4.根据权利要求1至3任一所述的方法,其特征在于,

所述第一文本为所述图像中包含的全部文本;或者,

所述第一文本为所述图像中包含的全部文本中的部分。

5.根据权利要求1至4任一所述的方法,其特征在于,

所述图像为从原始的输入图像中提取的部分图像区域,所述图像包括的文本为所述输入图像包含的文本的部分;或者,

所述图像为原始的输入图像。

6.根据权利要求1至5任一所述的方法,其特征在于,所述第一文本包括第一子文本和第二子文本,所述第二特征表示包括所述第一子文本对应的第一子特征、以及所述第二子文本对应的第二子特征;所述第一子特征不包含所述第一子文本在所述图像中的位置;所述第二子特征包含所述第二子文本在所述图像中的位置;

所述根据所述第三特征表示,执行下游任务,包括:

根据所述第三特征表示,预测所述第一子文本在所述图像中的第一位置;所述第一位置和对应的标注之间的相似度用于更新所述图像编码器以及所述目标编码器。

7.根据权利要求6所述的方法,其特征在于,所述图像包括多个图像块;所述第一位置为对所述第一子文本预测所在的图像块;所述标注为所述第一子文本真实所在的图像块。

8.根据权利要求1至7任一所述的方法,其特征在于,所述第一特征表示包括第三子特征和第四子特征;所述方法还包括:

根据所述第四子特征、所述第四子特征在所述第一特征表示中的第二位置、以及所述第三子特征在所述第一特征表示中的第三位置,得到所述第三位置处的特征预测值;所述特征预测值和所述第三子特征之间的相似度用于更新所述图像编码器。

9.根据权利要求8所述的方法,其特征在于,所述根据所述第四子特征、所述第四子特征在所述第一特征表示中的第二位置、以及所述第三子特征在所述第一特征表示中的第三位置,得到所述第三位置处的特征预测值,包括:

根据所述第四子特征、所述第四子特征在所述第一特征表示中的第二位置、以及所述第三子特征在所述第一特征表示中的第三位置，通过自注意力网络，得到第四特征表示；

根据所述第四特征表示，通过预测网络，得到所述第三位置处的特征预测值。

10.一种数据处理方法，其特征在于，所述方法包括：

获取图像；

通过图像编码器，对所述图像进行处理，得到第一特征表示；

通过文本编码器，对所述图像中包含的文本进行处理，得到第二特征表示；

根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示；所述目标编码器用于实现特征表示之间的交互；

根据所述第三特征表示，执行下游任务，得到执行结果。

11.一种数据处理装置，其特征在于，包括：

获取模块，用于获取第一特征表示以及第二特征表示，所述第一特征表示为根据图像编码器对图像进行处理得到的图像特征得到，所述第二特征表示为第一文本的文本特征；所述第一文本为所述图像中包括的文本内容；

编码模块，用于根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示；所述目标编码器用于实现特征表示之间的交互；

任务执行模块，用于根据所述第三特征表示，执行下游任务，得到执行结果；所述执行结果和对应的标注之间的相似度以及所述第一特征表示和所述第二特征表示之间的相似度用于更新所述图像编码器。

12.根据权利要求11所述的装置，其特征在于，所述第二特征表示为通过文本编码器对所述第一文本进行处理得到的文本特征；所述第一特征表示和所述第二特征表示之间的相似度还用于更新所述文本编码器。

13.根据权利要求11或12所述的装置，其特征在于，所述第一特征表示和所述第二特征表示之间的相似度与图像中所蕴含的文本语义信息和文本自身的语义信息之间的相似度有关。

14.根据权利要求11至13任一所述的装置，其特征在于，

所述第一文本为所述图像中包含的全部文本；或者，

所述第一文本为所述图像中包含的全部文本中的部分。

15.根据权利要求11至14任一所述的装置，其特征在于，

所述图像为从原始的输入图像中提取的部分图像区域，所述图像包括的文本为所述输入图像包含的文本的部分；或者，

所述图像为原始的输入图像。

16.根据权利要求11至15任一所述的装置，其特征在于，所述第一文本包括第一子文本和第二子文本，所述第二特征表示包括所述第一子文本对应的第一子特征、以及所述第二子文本对应的第二子特征；所述第一子特征不包含所述第一子文本在所述图像中的位置；所述第二子特征包含所述第二子文本在所述图像中的位置；

所述任务执行模块，具体用于：

根据所述第三特征表示，预测所述第一子文本在所述图像中的第一位置；所述第一位置和对应的标注之间的相似度用于更新所述图像编码器以及所述目标编码器。

17.根据权利要求16所述的装置，其特征在于，所述图像包括多个图像块；所述第一位置为对所述第一子文本预测所在的图像块；所述标注为所述第一子文本真实所在的图像块。

18.根据权利要求11至17任一所述的装置，其特征在于，所述第一特征表示包括第三子特征和第四子特征；所述装置还包括：

预测模块，用于根据所述第四子特征、所述第四子特征在所述第一特征表示中的第二位置、以及所述第三子特征在所述第一特征表示中的第三位置，得到所述第三位置处的特征预测值；所述特征预测值和所述第三子特征之间的相似度用于更新所述图像编码器。

19.根据权利要求18所述的装置，其特征在于，所述预测模块，具体用于：

根据所述第四子特征、所述第四子特征在所述第一特征表示中的第二位置、以及所述第三子特征在所述第一特征表示中的第三位置，通过自注意力网络，得到第四特征表示；

根据所述第四特征表示，通过预测网络，得到所述第三位置处的特征预测值。

20.一种数据处理装置，其特征在于，所述装置包括：

获取模块，用于获取图像；

编码模块，用于通过图像编码器，对所述图像进行处理，得到第一特征表示；

通过文本编码器，对所述图像中包含的文本进行处理，得到第二特征表示；

根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示；所述目标编码器用于实现特征表示之间的交互；

任务执行模块，用于根据所述第三特征表示，执行下游任务，得到执行结果。

21.一种计算机存储介质，其特征在于，所述计算机存储介质存储有一个或多个指令，所述指令在由一个或多个计算机执行时使得所述一个或多个计算机执行权利要求1至10中任一项所述方法的操作。

22.一种计算机程序产品，其特征在于，包括计算机可读指令，当所述计算机可读指令在计算机设备上运行时，使得所述计算机设备执行如权利要求1至10任一所述的方法。

23.一种系统，包括至少一个处理器，至少一个存储器；所述处理器、所述存储器通过通信总线连接并完成相互间的通信；

所述至少一个存储器用于存储代码；

所述至少一个处理器用于执行所述代码，以执行如权利要求1至10任一所述的方法。

一种数据处理方法及其装置

技术领域

[0001] 本申请涉及人工智能领域,尤其涉及一种数据处理方法及其装置。

背景技术

[0002] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个分支,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式作出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0003] 现有的跨模态视觉语言模型(或者可以称之为多模态模型或者多模态语言模型)所采用的对齐方式中的文本建模能力较弱。以图文对齐(text-image alignment,TIA)为例,TIA只是简单的判断某个字符是否被遮挡,并没有很好的考虑到遮挡字符之间的语义信息导致模型不是真正的理解文档的内容。

[0004] 当前的多模态模型的图文对齐能力较弱,也就是不具备文本和图像元素的对齐能力。以图文匹配(text-image matching,TIM)为例,训练目标用来判断全局文字是否和全局的扫描文档图片是否匹配,该任务学习难度小,使得训练的模型不具备文本和图像元素的对齐能力。

发明内容

[0005] 本申请提供了一种数据处理方法,可以使得图像编码器和文本编码器提取的特征中蕴含更准确的图像中文本的语义信息,进而增强后续网络的图文匹配能力,本申请中通过先双塔方式可以提升图文之间的对齐能力,再利用一个单塔结构进一步增强特征的交互学习能力。

[0006] 第一方面,本申请提供了一种数据处理方法,包括:获取第一特征表示以及第二特征表示,第一特征表示为根据图像编码器对图像进行处理得到的图像特征得到,第二特征表示为第一文本的文本特征;第一文本为图像中包括的文本内容;根据第一特征表示以及第二特征表示,通过目标编码器,得到第三特征表示;目标编码器用于实现特征表示之间的交互;根据第三特征表示,执行下游任务,得到执行结果;所述执行结果和对应的标注之间的相似度以及所述第一特征表示和所述第二特征表示之间的相似度用于更新所述图像编码器。

[0007] 本申请中,通过图文对比学习(text-image contrastive learning,TIC)来增强图文之间的匹配能力,具体的,针对于图像的图像特征和文本的文本特征,通过相似度比对构建的损失,来更新图形编码器(如果文本特征为通过文本编码器得到的,也可以更新文本编码器)。虽然图像的图像特征和图像中文本的文本特征不是同一个模态的特征,但是由于文本是图像中的文本,在图像特征中也会蕴含一定的文本的语义信息,因此,图像特征和文本特征(属于同一个图像)之间在语义维度(或者其他信息维度上)会存在关联。上述相似度

可以包含图像中语义信息和文本的语义信息之间的相似度，因此，基于该相似度构建的损失来更新图像编码器以及文本编码器(如果架构中存在文本编码器的话)，能够使得图像编码器和文本编码器提取的特征中蕴含更准确的图像中文本的语义信息，进而增强后续网络的图文匹配能力。

[0008] 此外，在跨模态的语言模型的架构中，包括图形编码器、文本编码器以及用于提供特征之间交互信息的目标编码器，本申请中，将模型的中间输出(也就是图形编码器和文本编码器的输出)之间的相似度构建的损失来更新图形编码器和文本编码器，使得基于图形编码器和文本编码器输出的特征表示就可以实现下游任务(精度没有基于目标编码器输出的特征表示进行的下游任务高)，而在一些场景中，由于下游任务要处理的数据的数量较大，因此可以使用图形编码器和文本编码器输出的特征表示进行粗排，使用目标编码器输出的特征表示进行精排，以提高召回率。

[0009] 在一种可能的实现中，第二特征表示为通过文本编码器对第一文本进行处理得到的文本特征；第一特征表示和第二特征表示之间的相似度还用于更新文本编码器。

[0010] 在一种可能的实现中，相似度与图像中所蕴含的文本语义信息和文本自身的语义信息之间的相似度有关。

[0011] 在一种可能的实现中所述执行结果和对应的标注之间的相似度还用于更新目标编码器。

[0012] 在一种可能的实现中，

[0013] 第一文本为图像中包含的全部文本；或者，

[0014] 第一文本为图像中包含的全部文本中的部分。

[0015] 在一种可能的实现中，

[0016] 图像为从原始的输入图像中提取的部分图像区域，图像包括的文本为输入图像包含的文本的部分；或者，

[0017] 图像为原始的输入图像。

[0018] 在一种可能的实现中，图像可以为原始的输入图像(或者是原始的输入图像的部分图像，但该部分图像包含输入图像的全部文本)。第一文本可以为图像中包含的全部文本。

[0019] 在一种可能的实现中，图像可以为原始的输入图像(或者是原始的输入图像的部分图像，但该部分图像包含输入图像的全部文本)。第一文本可以为图像中包含的文本中的部分文本。

[0020] 在一种可能的实现中，图像可以为对原始的输入图像进行提取得到的部分图像区域，该部分图像区域内包括原始的输入图像中全部文本的部分文本，第一文本可以为图像中包括的全部文本。

[0021] 在一种可能的实现中，图像可以为对原始的输入图像进行提取得到的部分图像区域，该部分图像区域内包括原始的输入图像中全部文本的部分文本，第一文本可以为图像中包含的文本中的部分文本。

[0022] 在一种可能的实现中，输入的原始图像中可以包括一行或多行文本，或者，输入的原始图像中可以包括一列或多列文本。可选的，图像中可以包括原始图像的一行或者一列文本，第一文本可以为一个或多个文本单元。

[0023] 在进行图文对比学习时,可以根据图像的图像特征和第一文本的文本特征之间的相似度来构建用于更新图形编码器以及文本编码器的损失,可选的,该相似度与图像中所蕴含的文本语义信息和文本自身的语义信息之间的相似度有关。

[0024] 应理解,还可以根据图像的图像特征和图像中不包括的文本的文本特征之间的相似度来构建用于更新图形编码器以及文本编码器的损失,区别在于,图像的图像特征和图像中文本的文本特征为正例(也就是对应的标注相似度高),图像的图像特征和图像中不包括的文本的文本特征为负例(也就是对应的标注相似度低)。

[0025] 在一种可能的实现中,针对于同一张原始的输入图像,可以将输入图像整体对应的图像特征和输入图像中包括的全部文本的文本特征进行比对,也可以将输入图像中的部分图像区域的图像特征和输入图像中包括的部分文本的文本特征进行比对,也可以将输入图像整体对应的图像特征和输入图像中包括的部分文本的文本特征进行比对,或者是上述几种方式的组合。

[0026] 在一种可能的实现中,针对于同一张原始的输入图像,可以将输入图像中包括的全部文本整体对应的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将输入图像中包括的部分文本的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将输入图像中包括的全部文本整体对应的文本特征和输入图像整体对应的图像特征进行比对。

[0027] 在一种可能的实现中,针对于不同的原始输入图像,可以将一部分输入图像中包括的全部文本整体对应的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将一部分输入图像中包括的部分文本的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将一部分输入图像中包括的全部文本整体对应的文本特征和输入图像整体对应的图像特征进行比对,或者是上述几种方式的组合。

[0028] 在一种可能的实现中,针对于不同的原始输入图像,可以将一部分输入图像中包括的全部文本整体对应的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将一部分输入图像中包括的部分文本的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将一部分输入图像中包括的全部文本整体对应的文本特征和输入图像整体对应的图像特征进行比对。

[0029] 在一种可能的实现中,第一文本包括第一子文本和第二子文本,第二特征表示包括第一子文本对应的第一子特征、以及第二子文本对应的第二子特征;第一子特征不包含第一子文本在图像中的位置;第二子特征包含第二子文本在图像中的位置;根据第三特征表示,执行下游任务,包括:根据第三特征表示,预测第一子文本在图像中的第一位置;第一位置和对应的标注之间的相似度用于更新图像编码器以及目标编码器。

[0030] 在一种可能的实现中,图像包括多个图像块;第一位置为对第一子文本预测所在的图像块;标注为第一子文本真实所在的图像块。

[0031] 本申请在单塔部分提出BGM来提升空间布局感知能力,有助于提升位置敏感的任务中模型的性能。例如在信息提取中,key-value对通常是相邻的成对数据。因为模型在预训练阶段已经学会了对位置以及布局信息的感知能力,所以在信息抽取这类任务中模型会有很好的性能。

[0032] 在一种可能的实现中,第一特征表示包括第三子特征和第四子特征;方法还包括:

根据第四子特征、第四子特征在第一特征表示中的第二位置、以及第三子特征在第一特征表示中的第三位置，得到第三位置处的特征预测值；特征预测值和第三子特征之间的相似度用于更新图像编码器。

[0033] 在一种可能的实现中，根据第四子特征、第四子特征在第一特征表示中的第二位置、以及第三子特征在第一特征表示中的第三位置，得到第三位置处的特征预测值，包括：根据第四子特征、第四子特征在第一特征表示中的第二位置、以及第三子特征在第一特征表示中的第三位置，通过自注意力网络，得到第四特征表示；根据第四特征表示，通过预测网络，得到第三位置处的特征预测值。

[0034] 在一种可能的实现中，为了提高图像表征的学习能力，还可以基于图像特征来构建损失，具体的，可以对图像中的部分进行掩码，通过图像编码器得到的结果对掩码区域图像进行图像重建，并基于重建结果和掩码区域的真实图像的像素值之间的差异来更新网络。然而，现有方法图像表征学习能力较弱，具体是由于，为了降低模型的处理算例开销，需要将图像压缩到一个较小的分辨率(例如224*224)，在图像中包含密集型文本的情况下，图像中的像素难以表达出准确的文字信息，训练后的模型的图像表征的学习能力有限。

[0035] 本申请实施例中，为了提高图像表征的学习能力，可以不对图像而是对图像的图像特征进行掩码，并对掩码区域的图像特征进行特征重建，并基于重建结果和掩码区域的图像特征之间的差异来更新网络，由于图像特征相比像素值本身可以携带更丰富的图像信息以及文本信息，因此可以提高训练后的网络的图像表征的学习能力，上述方式可以称之为掩码图像建模(mask image modeling，MIM)。

[0036] 第二方面，本申请提供了一种数据处理方法，方法包括：

[0037] 获取图像；

[0038] 通过图像编码器，对图像进行处理，得到第一特征表示；

[0039] 通过文本编码器，对图像中包含的文本进行处理，得到第二特征表示；

[0040] 根据第一特征表示以及第二特征表示，通过目标编码器，得到第三特征表示；目标编码器用于实现特征表示之间的交互；

[0041] 根据第三特征表示，执行下游任务，得到执行结果。

[0042] 第三方面，本申请提供了一种数据处理装置，包括：

[0043] 获取模块，用于获取第一特征表示以及第二特征表示，第一特征表示为根据图像编码器对图像进行处理得到的图像特征得到，第二特征表示为第一文本的文本特征；第一文本为图像中包括的文本内容；第一特征表示和第二特征表示之间的相似度用于更新图像编码器；

[0044] 编码模块，用于根据第一特征表示以及第二特征表示，通过目标编码器，得到第三特征表示；目标编码器用于实现特征表示之间的交互；

[0045] 任务执行模块，用于根据第三特征表示，执行下游任务，得到执行结果。

[0046] 在一种可能的实现中，第二特征表示为通过文本编码器对第一文本进行处理得到的文本特征；第一特征表示和第二特征表示之间的相似度还用于更新文本编码器。

[0047] 在一种可能的实现中，相似度与图像中所蕴含的文本语义信息和文本自身的语义信息之间的相似度有关。

[0048] 在一种可能的实现中，

[0049] 第一文本为图像中包含的全部文本;或者,

[0050] 第一文本为图像中包含的全部文本中的部分。

[0051] 在一种可能的实现中,

[0052] 图像为从原始的输入图像中提取的部分图像区域,图像包括的文本为输入图像包含的文本的部分;或者,

[0053] 图像为原始的输入图像。

[0054] 在一种可能的实现中,第一文本包括第一子文本和第二子文本,第二特征表示包括第一子文本对应的第一子特征、以及第二子文本对应的第二子特征;第一子特征不包含第一子文本在图像中的位置;第二子特征包含第二子文本在图像中的位置;

[0055] 任务执行模块,具体用于:

[0056] 根据第三特征表示,预测第一子文本在图像中的第一位置;第一位置和对应的标注之间的相似度用于更新图像编码器以及目标编码器。

[0057] 在一种可能的实现中,图像包括多个图像块;第一位置为对第一子文本预测所在的图像块;标注为第一子文本真实所在的图像块。

[0058] 在一种可能的实现中,第一特征表示包括第三子特征和第四子特征;方法还包括:

[0059] 预测模块,用于根据第四子特征、第四子特征在第一特征表示中的第二位置、以及第三子特征在第一特征表示中的第三位置,得到第三位置处的特征预测值;特征预测值和第三子特征之间的相似度用于更新图像编码器。

[0060] 在一种可能的实现中,预测模块,具体用于:

[0061] 根据第四子特征、第四子特征在第一特征表示中的第二位置、以及第三子特征在第一特征表示中的第三位置,通过自注意力网络,得到第四特征表示;

[0062] 根据第四特征表示,通过预测网络,得到第三位置处的特征预测值。

[0063] 第四方面,本申请提供了一种数据处理装置,装置包括:

[0064] 获取模块,用于获取图像;

[0065] 编码模块,用于通过图像编码器,对图像进行处理,得到第一特征表示;

[0066] 通过文本编码器,对图像中包含的文本进行处理,得到第二特征表示;

[0067] 根据第一特征表示以及第二特征表示,通过目标编码器,得到第三特征表示;目标编码器用于实现特征表示之间的交互;

[0068] 任务执行模块,用于根据第三特征表示,执行下游任务,得到执行结果。

[0069] 第五方面,本申请实施例提供了一种训练装置,可以包括存储器、处理器以及总线系统,其中,存储器用于存储程序,处理器用于执行存储器中的程序,以执行如上述第一方面及其任一可选的方法。

[0070] 第六方面,本申请实施例提供了一种执行装置,可以包括存储器、处理器以及总线系统,其中,存储器用于存储程序,处理器用于执行存储器中的程序,以执行如上述第二方面及其任一可选的方法。

[0071] 第七方面,本申请实施例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中存储有计算机程序,当其在计算机上运行时,使得计算机执行上述第一方面及其任一可选的方法、以及上述第二方面及其任一可选的方法。

[0072] 第八方面,本申请实施例提供了一种计算机程序,当其在计算机上运行时,使得计

算机执行上述第一方面及其任一可选的方法、以及上述第二方面及其任一可选的方法。

[0073] 第九方面，本申请提供了一种芯片系统，该芯片系统包括处理器，用于支持执行数据处理装置实现上述方面中所涉及的功能，例如，发送或处理上述方法中所涉及的数据；或，信息。在一种可能的设计中，所述芯片系统还包括存储器，所述存储器，用于保存执行设备或训练设备必要的程序指令和数据。该芯片系统，可以由芯片构成，也可以包括芯片和其他分立器件。

附图说明

[0074] 图1A为人工智能主体框架的一种结构示意图；

[0075] 图1B和图1C为本发明的应用系统框架示意；

[0076] 图1D为终端的一种可选的硬件结构示意图；

[0077] 图2为一种服务器的结构示意图；

[0078] 图3为本申请的一种系统架构示意；

[0079] 图4为一种云服务的流程；

[0080] 图5为一种云服务的流程；

[0081] 图6为本申请的一种系统架构示意；

[0082] 图7为本申请实施例提供的一种数据处理方法的流程示意；

[0083] 图8至图10为本申请实施例提供的一种数据处理方法的流程示意；

[0084] 图11为本申请的一种系统架构示意；

[0085] 图12为本申请实施例提供的数据处理装置的一种结构示意图；

[0086] 图13为本申请实施例提供的执行设备的一种结构示意图；

[0087] 图14为本申请实施例提供的训练设备一种结构示意图；

[0088] 图15为本申请实施例提供的芯片的一种结构示意图。

具体实施方式

[0089] 下面结合本发明实施例中的附图对本发明实施例进行描述。本发明的实施方式部分使用的术语仅用于对本发明的具体实施例进行解释，而非旨在限定本发明。

[0090] 下面结合附图，对本申请的实施例进行描述。本领域普通技术人员可知，随着技术的发展和新场景的出现，本申请实施例提供的技术方案对于类似的技术问题，同样适用。

[0091] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的术语在适当情况下可以互换，这仅仅是描述本申请的实施例中对相同属性的对象在描述时所采用的区分方式。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，以便包含一系列单元的过程、方法、系统、产品或设备不必限于那些单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它单元。

[0092] 本文中所用用语“基本(substantially)”、“大约(about)”及类似用语用作近似用语、而并非用作程度用语，且旨在考虑到所属领域中的普通技术人员将知的测量值或计算值的固有偏差。此外，在阐述本发明实施例时使用“可(may)”是指“可能的一个或多个实施例”。本文中所用用语“使用(use)”、“正使用(using)”、及“被使用(used)”可被视为分别与

用语“利用(utilize)”、“正利用(utilizing)”、及“被利用(utilized)”同义。另外,用语“示例性(exemplary)”旨在指代实例或例示。

[0093] 首先对人工智能系统总体工作流程进行描述,请参见图1A,图1A示出的为人工智能主体框架的一种结构示意图,下面从“智能信息链”(水平轴)和“IT价值链”(垂直轴)两个维度对上述人工智能主题框架进行阐述。其中,“智能信息链”反映从数据的获取到处理的一列过程。举例来说,可以是智能信息感知、智能信息表示与形成、智能推理、智能决策、智能执行与输出的一般过程。在这个过程中,数据经历了“数据—信息—知识—智慧”的凝练过程。“IT价值链”从人智能的底层基础设施、信息(提供和处理技术实现)到系统的产业生态过程,反映人工智能为信息技术产业带来的价值。

[0094] (1)基础设施

[0095] 基础设施为人工智能系统提供计算能力支持,实现与外部世界的沟通,并通过基础平台实现支撑。通过传感器与外部沟通;计算能力由智能芯片(CPU、NPU、GPU、ASIC、FPGA等硬件加速芯片)提供;基础平台包括分布式计算框架及网络等相关的平台保障和支持,可以包括云存储和计算、互联互通网络等。举例来说,传感器和外部沟通获取数据,这些数据提供给基础平台提供的分布式计算系统中的智能芯片进行计算。

[0096] (2)数据

[0097] 基础设施的上一层的数据用于表示人工智能领域的数据来源。数据涉及到图形、图像、语音、文本,还涉及到传统设备的物联网数据,包括已有系统的业务数据以及力、位移、液位、温度、湿度等感知数据。

[0098] (3)数据处理

[0099] 数据处理通常包括数据训练,机器学习,深度学习,搜索,推理,决策等方式。

[0100] 其中,机器学习和深度学习可以对数据进行符号化和形式化的智能信息建模、抽取、预处理、训练等。

[0101] 推理是指在计算机或智能系统中,模拟人类的智能推理方式,依据推理控制策略,利用形式化的信息进行机器思维和求解问题的过程,典型的功能是搜索与匹配。

[0102] 决策是指智能信息经过推理后进行决策的过程,通常提供分类、排序、预测等功能。

[0103] (4)通用能力

[0104] 对数据经过上面提到的数据处理后,进一步基于数据处理的结果可以形成一些通用的能力,比如可以是算法或者一个通用系统,例如,翻译,文本的分析,计算机视觉的处理,语音识别,图像的识别等等。

[0105] (5)智能产品及行业应用

[0106] 智能产品及行业应用指人工智能系统在各领域的产品和应用,是对人工智能整体解决方案的封装,将智能信息决策产品化、实现落地应用,其应用领域主要包括:智能终端、智能交通、智能医疗、自动驾驶、智慧城市等。

[0107] 本申请可以应用于人工智能领域的自然语言处理领域中,下面以自然语言处理为例将对多个落地到产品的多个应用场景进行介绍。

[0108] 首先介绍本申请的应用场景,本申请可以但不限于应用在具有对图像的文本处理功能的应用程序(以下可以简称为跨模态的语言处理类应用程序)或者云侧服务器提供的

云服务等,接下来分别进行介绍:

[0109] 一、跨模态的语言处理类应用程序

[0110] 本申请实施例的产品形态可以为跨模态的语言处理类应用程序。跨模态的语言处理类应用程序可以运行在终端设备或者云侧的服务器上。

[0111] 在一种可能的实现中,跨模态的语言处理类应用程序可以实现基于输入的图像进行图像中文本相关的处理任务,得到处理结果。

[0112] 示例性的,上述图像中文本相关的处理任务可以包括信息提取、合同审阅和检索与问答等。在信息提取场景中,用户可以自定义图像中所关心的字段,如甲方、乙方、合同编号等。在合同审阅场景中,相关人员可以确认不同版本的合同是否一致,如甲方的名称发生了变化是否会引入风险条款等。在检索与问答场景中,可以帮助用户通过提出问题或关键词快速检索出图像中的相关文本,并在文本中找出可能的答案。

[0113] 示例性的,前述图像为通过摄像机、打印机、扫描机等设备采集得到的图像。作为示例,例如在一种应用场景中,在金融、会计和税务领域等领域中,企业需要对收据或发票等文件进行扫描得到图像文件,并对图像文件中的文本进行识别以提取文本信息,进而能够实现文件数字化归档、文件快速索引或文件分析等功能。在另一种应用场景中,用户需要输入身份证、驾驶证、行驶证或护照等证件上的信息,则用户可以利用摄像机采集得到前述证件的图像,并对图像中的文本进行识别以提取出关键信息等。

[0114] 应当理解,此处举例仅为方便对本申请实施例的应用场景进行理解,不对本申请实施例的应用场景进行穷举。

[0115] 在一种可能的实现中,用户可以打开终端设备上安装的跨模态的语言处理类应用程序,并输入图像,跨模态的语言处理类应用程序可以通过本申请实施例提供的方法训练得到的跨模态语言模型对图像进行处理,并将处理结果呈现给用户(呈现方式可以但不限于是显示、保存、上传到云侧等)。

[0116] 在一种可能的实现中,用户可以打开终端设备上安装的跨模态的语言处理类应用程序,并输入图像,跨模态的语言处理类应用程序可以将图像发送至云侧的服务器,云侧的服务器通过本申请实施例提供的方法训练得到的跨模态语言模型对图像进行处理,并将处理结果回传至终端设备,终端设备可以将处理结果呈现给用户(呈现方式可以但不限于是显示、保存、上传到云侧等)。

[0117] 接下来分别从功能架构以及实现功能的产品架构介绍本申请实施例中的跨模态的语言处理类应用程序。

[0118] 参照图1B,图1B为本申请实施例中跨模态的语言处理类应用程序的功能架构示意:

[0119] 在一种可能的实现中,如图1B所示,跨模态的语言处理类应用程序102可接收输入的参数101(例如包含图像)且产生处理结果103。跨模态的语言处理类应用程序102可在(举例来说)至少一个计算机系统上执行,且包括计算机代码,所述计算机代码在由一或多个计算机执行时致使所述计算机执行用于执行通过本申请实施例提供的方法训练得到的跨模态语言模型。

[0120] 参照图1C,图1C为本申请实施例中运行跨模态的语言处理类应用程序的实体架构示意:

[0121] 参见图1C,图1C示出了一种系统架构示意图。该系统可以包括终端100、以及服务器200。其中,服务器200可以包括一个或者多个服务器(图1C中以包括一个服务器作为示例进行说明),服务器200可以为一个或者多个终端提供跨模态的语言处理功能服务。

[0122] 其中,终端100上可以安装有跨模态的语言处理类应用程序,或者打开与跨模态的语言处理功能相关的网页,上述应用程序和网页可以提供一个界面,终端100可以接收用户在跨模态的语言处理功能界面上输入的相关参数,并将上述参数发送至服务器200,服务器200可以基于接收到的参数,得到处理结果,并将处理结果返回至至终端100。

[0123] 应理解,在一些可选的实现中,终端100也可以由自身完成基于接收到的参数,得到处理结果的动作,而不需要服务器配合实现,本申请实施例并不限定。

[0124] 接下来描述图1C中终端100的产品形态;

[0125] 本申请实施例中的终端100可以为手机、平板电脑、可穿戴设备、车载设备、增强现实(augmented reality,AR)/虚拟现实(virtual reality,VR)设备、笔记本电脑、超级移动个人计算机(ultra-mobile personal computer,UMPC)、上网本、个人数字助理(personal digital assistant,PDA)等,本申请实施例对此不作任何限制。

[0126] 图1D示出了终端100的一种可选的硬件结构示意图。

[0127] 参考图1D所示,终端100可以包括射频单元110、存储器120、输入单元130、显示单元140、摄像头150(可选的)、音频电路160(可选的)、扬声器161(可选的)、麦克风162(可选的)、处理器170、外部接口180、电源190等部件。本领域技术人员可以理解,图1D仅仅是终端或多功能设备的举例,并不构成对终端或多功能设备的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件。

[0128] 输入单元130可用于接收输入的数字或字符信息,以及产生与该便携式多功能装置的用户设置以及功能控制有关的键信号输入。具体地,输入单元130可包括触摸屏131(可选的)和/或其他输入设备132。该触摸屏131可收集用户在其上或附近的触摸操作(比如用户使用手指、关节、触笔等任何适合的物体在触摸屏上或在触摸屏附近的操作),并根据预先设定的程序驱动相应的连接装置。触摸屏可以检测用户对触摸屏的触摸动作,将该触摸动作转换为触摸信号发送给该处理器170,并能接收该处理器170发来的命令并加以执行;该触摸信号至少包括触点坐标信息。该触摸屏131可以提供该终端100和用户之间的输入界面和输出界面。此外,可以采用电阻式、电容式、红外线以及表面声波等多种类型实现触摸屏。除了触摸屏131,输入单元130还可以包括其他输入设备。具体地,其他输入设备132可以包括但不限于物理键盘、功能键(比如音量控制按键132、开关按键133等)、轨迹球、鼠标、操作杆等中的一种或多种。

[0129] 其中,输入设备132可以接收到输入的图像等等。

[0130] 该显示单元140可用于显示由用户输入的信息或提供给用户的信息、终端100的各种菜单、交互界面、文件显示和/或任意一种多媒体文件的播放。在本申请实施例中,显示单元140可用于显示跨模态的语言处理类应用程序的界面、处理结果等。

[0131] 该存储器120可用于存储指令和数据,存储器120可主要包括存储指令区和存储数据区,存储数据区可存储各种数据,如多媒体文件、文本等;存储指令区可存储操作系统、应用、至少一个功能所需的指令等软件单元,或者他们的子集、扩展集。还可以包括非易失性随机存储器;向处理器170提供包括管理计算处理设备中的硬件、软件以及数据资源,支持

控制软件和应用。还用于多媒体文件的存储，以及运行程序和应用的存储。

[0132] 处理器170是终端100的控制中心，利用各种接口和线路连接整个终端100的各个部分，通过运行或执行存储在存储器120内的指令以及调用存储在存储器120内的数据，执行终端100的各种功能和处理数据，从而对终端设备进行整体控制。可选的，处理器170可包括一个或多个处理单元；优选的，处理器170可集成应用处理器和调制解调处理器，其中，应用处理器主要处理操作系统、用户界面和应用程序等，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器170中。在一些实施例中，处理器、存储器、可以在单一芯片上实现，在一些实施例中，他们也可以在独立的芯片上分别实现。处理器170还可以用于产生相应的操作控制信号，发给计算处理设备相应的部件，读取以及处理软件中的数据，尤其是读取和处理存储器120中的数据和程序，以使其中的各个功能模块执行相应的功能，从而控制相应的部件按指令的要求进行动作。

[0133] 其中，存储器120可以用于存储数据处理方法相关的软件代码，处理器170可以执行芯片的数据处理方法的步骤，也可以调度其他单元(例如上述输入单元130以及显示单元140)以实现相应的功能。

[0134] 该射频单元110(可选的)可用于收发信息或通话过程中信号的接收和发送，例如，将基站的下行信息接收后，给处理器170处理；另外，将设计上行的数据发送给基站。通常，RF电路包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放大器(Low Noise Amplifier，LNA)、双工器等。此外，射频单元110还可以通过无线通信与网络设备和其他设备通信。该无线通信可以使用任一通信标准或协议，包括但不限于全球移动通讯系统(Global System of Mobile communication，GSM)、通用分组无线服务(General Packet Radio Service，GPRS)、码分多址(Code Division Multiple Access，CDMA)、宽带码分多址(Wideband Code Division Multiple Access，WCDMA)、长期演进(Long Term Evolution，LTE)、电子邮件、短消息服务(Short Messaging Service，SMS)等。

[0135] 其中，在本申请实施例中，该射频单元110可以将图像发送至服务器200，并接收到服务器200发送的处理结果。

[0136] 应理解，该射频单元110为可选的，其可以被替换为其他通信接口，例如可以是网口。

[0137] 终端100还包括给各个部件供电的电源190(比如电池)，优选的，电源可以通过电源管理系统与处理器170逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。

[0138] 终端100还包括外部接口180，该外部接口可以是标准的Micro USB接口，也可以使多针连接器，可以用于连接终端100与其他装置进行通信，也可以用于连接充电器为终端100充电。

[0139] 尽管未示出，终端100还可以包括闪光灯、无线保真(wireless fidelity，WiFi)模块、蓝牙模块、不同功能的传感器等，在此不再赘述。下文中描述的部分或全部方法均可以应用在如图1D所示的终端100中。

[0140] 接下来描述图1C中服务器200的产品形态；

[0141] 图2提供了一种服务器200的结构示意图，如图2所示，服务器200包括总线201、处理器202、通信接口203和存储器204。处理器202、存储器204和通信接口203之间通过总线

201通信。

[0142] 总线201可以是外设部件互连标准(peripheral component interconnect,PCI)总线或扩展工业标准结构(extended industry standard architecture,EISA)总线等。总线可以分为地址总线、数据总线、控制总线等。为便于表示,图2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。

[0143] 处理器202可以为中央处理器(central processing unit,CPU)、图形处理器(graphics processing unit,GPU)、微处理器(micro processor,MP)或者数字信号处理器(digital signal processor,DSP)等处理器中的任意一种或多种。

[0144] 存储器204可以包括易失性存储器(volatile memory),例如随机存取存储器(random access memory,RAM)。存储器204还可以包括非易失性存储器(non-volatile memory),例如只读存储器(read-only memory,ROM),快闪存储器,机械硬盘(hard drive drive,HDD)或固态硬盘(solid state drive,SSD)。

[0145] 其中,存储器204可以用于存储数据处理方法相关的软件代码,处理器202可以执行芯片的数据处理方法的步骤,也可以调度其他单元以实现相应的功能。

[0146] 应理解,上述终端100和服务器200可以为集中式或者是分布式的设备,上述终端100和服务器200中的处理器(例如处理器170以及处理器202)可以为硬件电路(如专用集成电路(application specific integrated circuit,ASIC)、现场可编程门阵列(field-programmable gate array,FPGA)、通用处理器、数字信号处理器(digital signal processing,DSP)、微处理器或微控制器等等)、或这些硬件电路的组合,例如,处理器可以为具有执行指令功能的硬件系统,如CPU、DSP等,或者为不具有执行指令功能的硬件系统,如ASIC、FPGA等,或者为上述不具有执行指令功能的硬件系统以及具有执行指令功能的硬件系统的组合。

[0147] 应理解,本申请实施例中的和模型推理过程相关的步骤涉及AI相关的运算,在执行AI运算时,终端设备和服务器的指令执行架构不仅仅局限在上述介绍的处理器结合存储器的架构。下面结合图3对本申请实施例提供的系统架构进行详细的介绍。

[0148] 图3为本申请实施例提供的系统架构示意图。如图3所示,系统架构500包括执行设备510、训练设备520、数据库530、客户设备540、数据存储系统550以及数据采集系统560。

[0149] 执行设备510包括计算模块511、I/O接口512、预处理模块513和预处理模块514。计算模块511中可以包括目标模型/规则501,预处理模块513和预处理模块514是可选的。

[0150] 其中,执行设备510可以为上述运行跨模态的语言处理类应用程序的终端设备或者服务器。

[0151] 数据采集设备560用于采集训练样本。训练样本可以为多个图像等。在采集到训练样本之后,数据采集设备560将这些训练样本存入数据库530。

[0152] 训练设备520可以基于数据库530中维护训练样本,对待训练的神经网络(例如本申请实施例中的跨模态语言模型(例如包括文本编码器、图像编码器、目标编码器等)),以得到目标模型/规则501。

[0153] 应理解,训练设备520可以基于数据库530中维护训练样本,对待训练的神经网络进行预训练过程,或者是在预训练的基础上进行模型的微调。

[0154] 需要说明的是,在实际应用中,数据库530中维护的训练样本不一定都来自于数据

采集设备560的采集，也有可能是从其他设备接收得到的。另外需要说明的是，训练设备520也不一定完全基于数据库530维护的训练样本进行目标模型/规则501的训练，也有可能从云端或其他地方获取训练样本进行模型训练，上述描述不应该作为对本申请实施例的限定。

[0155] 根据训练设备520训练得到的目标模型/规则501可以应用于不同的系统或设备中，如应用于图3所示的执行设备510，该执行设备510可以是终端，如手机终端，平板电脑，笔记本电脑，增强现实(augmented reality，AR)/虚拟现实(virtual reality，VR)设备，车载终端等，还可以是服务器等。

[0156] 具体的，训练设备520可以将训练后的模型传递至执行设备510。

[0157] 在图3中，执行设备510配置输入/输出(input/output，I/O)接口512，用于与外部设备进行数据交互，用户可以通过客户设备540向I/O接口512输入数据(例如本申请实施例中的图像等)。

[0158] 预处理模块513和预处理模块514用于根据I/O接口512接收到的输入数据进行预处理。应理解，可以没有预处理模块513和预处理模块514或者只有的一个预处理模块。当不存在预处理模块513和预处理模块514时，可以直接采用计算模块511对输入数据进行处理。

[0159] 在执行设备510对输入数据进行预处理，或者在执行设备510的计算模块511执行计算等相关的处理过程中，执行设备510可以调用数据存储系统550中的数据、代码等以用于相应的处理，也可以将相应处理得到的数据、指令等存入数据存储系统550中。

[0160] 最后，I/O接口512将处理结果提供给客户设备540，从而提供给用户。

[0161] 在图3所示情况下，用户可以手动给定输入数据，该“手动给定输入数据”可以通过I/O接口512提供的界面进行操作。另一种情况下，客户设备540可以自动地向I/O接口512发送输入数据，如果要求客户设备540自动发送输入数据需要获得用户的授权，则用户可以在客户设备540中设置相应权限。用户可以在客户设备540查看执行设备510输出的结果，具体的呈现形式可以是显示、声音、动作等具体方式。客户设备540也可以作为数据采集端，采集如图所示输入I/O接口512的输入数据及输出I/O接口512的输出结果作为新的样本数据，并存入数据库530。当然，也可以不经过客户设备540进行采集，而是由I/O接口512直接将如图所示输入I/O接口512的输入数据及输出I/O接口512的输出结果，作为新的样本数据存入数据库530。

[0162] 值得注意的是，图3仅是本申请实施例提供的一种系统架构的示意图，图中所示设备、器件、模块等之间的位置关系不构成任何限制，例如，在图3中，数据存储系统550相对执行设备510是外部存储器，在其它情况下，也可以将数据存储系统550置于执行设备510中。应理解，上述执行设备510可以部署于客户设备540中。

[0163] 从模型的推理侧来说：

[0164] 本申请实施例中，上述执行设备520的计算模块511可以获取到数据存储系统550中存储的代码来实现本申请实施例中的和模型推理过程相关的步骤。

[0165] 本申请实施例中，执行设备520的计算模块511可以包括硬件电路(如专用集成电路(application specific integrated circuit，ASIC)、现场可编程门阵列(field-programmable gate array，FPGA)、通用处理器、数字信号处理器(digital signal processing，DSP)、微处理器或微控制器等等)、或这些硬件电路的组合，例如，训练设备520

可以为具有执行指令功能的硬件系统，如CPU、DSP等，或者为不具有执行指令功能的硬件系统，如ASIC、FPGA等，或者为上述不具有执行指令功能的硬件系统以及具有执行指令功能的硬件系统的组合。

[0166] 具体的，执行设备520的计算模块511可以为具有执行指令功能的硬件系统，本申请实施例提供的和模型推理过程相关的步骤可以为存储在存储器中的软件代码，执行设备520的计算模块511可以从存储器中获取到软件代码，并执行获取到的软件代码来实现本申请实施例提供的和模型推理过程相关的步骤。

[0167] 应理解，执行设备520的计算模块511可以为不具有执行指令功能的硬件系统以及具有执行指令功能的硬件系统的组合，本申请实施例提供的和模型推理过程相关的步骤的部分步骤还可以通过执行设备520的计算模块511中不具有执行指令功能的硬件系统来实现，这里并不限定。

[0168] 从模型的训练侧来说：

[0169] 本申请实施例中，上述训练设备520可以获取到存储器(图3中未示出，可以集成于训练设备520或者与训练设备520分离部署)中存储的代码来实现本申请实施例中和模型训练相关的步骤。

[0170] 本申请实施例中，训练设备520可以包括硬件电路(如专用集成电路(application specific integrated circuit，ASIC)、现场可编程门阵列(field-programmable gate array，FPGA)、通用处理器、数字信号处理器(digital signal processing，DSP)、微处理器或微控制器等等)、或这些硬件电路的组合，例如，训练设备520可以为具有执行指令功能的硬件系统，如CPU、DSP等，或者为不具有执行指令功能的硬件系统，如ASIC、FPGA等，或者为上述不具有执行指令功能的硬件系统以及具有执行指令功能的硬件系统的组合。

[0171] 应理解，训练设备520可以为不具有执行指令功能的硬件系统以及具有执行指令功能的硬件系统的组合，本申请实施例提供的中和模型训练相关的部分步骤还可以通过训练设备520中不具有执行指令功能的硬件系统来实现，这里并不限定。

[0172] 二、服务器提供的跨模态的语言处理功能类云服务：

[0173] 在一种可能的实现中，服务器可以通过应用程序编程接口(application programming interface，API)为端侧提供跨模态的语言处理功能的服务。

[0174] 其中，终端设备可以通过云端提供的API，将相关参数(例如图像)发送至服务器，服务器可以基于接收到的参数，得到处理结果等)，并将处理结果返回至至终端。

[0175] 关于终端以及服务器的描述可以上述实施例的描述，这里不再赘述。

[0176] 如图4示出了使用一项云平台提供的跨模态的语言处理功能类云服务的流程。

[0177] 1.开通并购买内容审核服务。

[0178] 2.用户可以下载内容审核服务对应的软件开发工具包(software development kit，SDK)，通常云平台提供多个开发版本的SDK，供用户根据开发环境的需求选择，例如JAVA版本的SDK、python版本的SDK、PHP版本的SDK、Android版本的SDK等。

[0179] 3.用户根据需求下载对应版本的SDK到本地后，将SDK工程导入至本地开发环境，在本地开发环境中进行配置和调试，本地开发环境还可以进行其他功能的开发，使得形成一个集合了跨模态的语言处理功能类能力的应用。

[0180] 4.跨模态的语言处理功能类应用在被使用的过程中，当需要进行跨模态的语言处

理功能时,可以触发跨模态的语言处理功能的API调用。当应用触发跨模态的语言处理功能功能时,发起API请求至云环境中的跨模态的语言处理功能类服务的运行实例,其中,API请求中携带图像,由云环境中的运行实例对图像进行处理,获得处理结果。

[0181] 5.云环境将处理结果返回至应用,由此完成一次的跨模态的语言处理功能服务调用。

[0182] 三、服务器提供的模型训练类云服务:

[0183] 在一种可能的实现中,服务器可以基于客户提供的训练数据(例如可以包括图像),来预训练一个具备跨模态的语言处理功能的模型。

[0184] 在一种可能的实现中,服务器可以通过应用程序编程接口(application programming interface,API)为端侧提供模型训练的服务。

[0185] 其中,终端设备可以通过云端提供的API,将相关参数发送至服务器,服务器可以基于接收到的参数,得到处理结果,并将处理结果(例如具备跨模态的语言处理功能的模型等)返回至终端。

[0186] 如图5示出了使用一项云平台提供的模型训练类云服务的流程。

[0187] 关于终端以及服务器的描述可以上述实施例的描述,这里不再赘述。

[0188] 四、联邦训练

[0189] 联邦学习分为模型下发和模型上传两个步骤,中心节点将模型通过网络下发至终端设备;各终端设备利用本地数据计算模型的梯度;各分布式节点将梯度加密后上传至中心节点;中心节点汇总各终端分布式节点的梯度,并采用参数平均算法更新中心节点模型的参数。

[0190] 参见图6,图6为本申请实施例提供的一种模型训练方法的架构示意,如图6所示,本申请实施例提供的架构包括:云侧中心节点,例如可以是云侧的服务器。A1、A2、…为类型为A的分布式节点,如用户持有的手机产品。B1、B2、…为类型为B的分布式节点,如用户持有的个人电脑。在经过分布式节点的管理员(如手机、电脑的用户)同意后,分布式节点的管理员自愿在隐私得到保护的情况下共享其日常使用设备的过程中产生的数据,加入到模型训练计划,设备成为架构中的分布式节点。本实施例中的系统也可以包含更多类型的分布式节点,如智能手表等等。为保护数据隐私,分布式节点不会将数据上传至中心节点,仅在本地保存数据。分布式节点通过通信网络与云服务器连接。云侧中心节点可以运行大模型,而各分布式节点受硬件能力限制只能运行小模型,且A和B可以拥有不同的数据处理能力。

[0191] 由于本申请实施例涉及大量神经网络的应用,为了便于理解,下面先对本申请实施例涉及的相关术语及神经网络等相关概念进行介绍。

[0192] (1)神经网络

[0193] 神经网络可以是由神经单元组成的,神经单元可以是指以xs(即输入数据)和截距1为输入的运算单元,该运算单元的输出可以为:

[0194] $$h_{W,b}(x)=f\left(W^{T}x\right)=f\left(\sum_{s=1}^{n}W_{s}x_{s}+b\right)$$;

[0195] 其中,s=1、2、……n,n为大于1的自然数,Ws为xs的权重,b为神经单元的偏置。f为神经单元的激活函数(activation functions),用于将非线性特性引入神经网络中,来将神经单元中的输入信号转换为输出信号。该激活函数的输出信号可以作为下一层卷积层的

输入,激活函数可以是sigmoid函数。神经网络是将多个上述单一的神经单元联结在一起形成的网络,即一个神经单元的输出可以是另一个神经单元的输入。每个神经单元的输入可以与前一层的局部接受域相连,来提取局部接受域的特征,局部接受域可以是由若干个神经单元组成的区域。

[0196] (2)transformer层

[0197] 神经网络包括嵌入层和至少一个transformer层,至少一个transformer层可以为N个transformer层(N大于0的整数),其中,每个transformer层包括依次相邻的注意力层、加和与归一化(add&norm)层、前馈(feed forward)层和加和与归一化层。在嵌入层,对当前输入进行嵌入处理,得到多个嵌入向量;在所述注意力层,从所述第一transformer层的上一层获取P个输入向量,以P个输入向量中的任意的第一输入向量为中心,基于预设的注意力窗口范围内的各个输入向量与该第一输入向量之间的关联度,得到该第一输入向量对应的中间向量,如此确定出P个输入向量对应的P个中间向量;在所述池化层,将所述P个中间向量合并为Q个输出向量,其中transformer层中最后一个transformer层得到的多个输出向量用作所述当前输入的特征表示。

[0198] (3)注意力机制(attention mechanism)

[0199] 注意力机制模仿了生物观察行为的内部过程,即一种将内部经验和外部感觉对齐从而增加部分区域的观察精细度的机制,能够利用有限的注意力资源从大量信息中快速筛选出高价值信息。注意力机制可以快速提取稀疏数据的重要特征,因而被广泛用于自然语言处理任务,特别是机器翻译。而自注意力机制(self-attention mechanism)是注意力机制的改进,其减少了对外部信息的依赖,更擅长捕捉数据或特征的内部相关性。注意力机制的本质思想可以改写为如下公式:

[0200] 其中,Lx=||Source||代表Source的长度,公式含义即将Source中的构成元素想象成是由一系列的数据对构成,此时给定目标Target中的某个元素Query,通过计算Query和各个Key的相似性或者相关性,得到每个Key对应Value的权重系数,然后对Value进行加权求和,即得到了最终的Attention数值。所以本质上Attention机制是对Source中元素的Value值进行加权求和,而Query和Key用来计算对应Value的权重系数。从概念上理解,把Attention可以理解为从大量信息中有选择地筛选出少量重要信息并聚焦到这些重要信息上,忽略大多不重要的信息。聚焦的过程体现在权重系数的计算上,权重越大越聚焦于其对应的Value值上,即权重代表了信息的重要性,而Value是其对应的信息。自注意力机制可以理解为内部Attention(intra attention),Attention机制发生在Target的元素Query和Source中的所有元素之间,自注意力机制指的是在Source内部元素之间或者Target内部元素之间发生的Attention机制,也可以理解为Target=Source这种特殊情况下的注意力计算机制,其具体计算过程是一样的,只是计算对象发生了变化而已。

[0201] (4)自然语言处理(natural language processing,NLP)

[0202] 自然语言(natural language)即人类语言,自然语言处理(NLP)就是对人类语言的处理。自然语言处理是以一种智能与高效的方式,对文本数据进行系统化分析、理解与信息提取的过程。通过使用NLP及其组件,我们可以管理非常大块的文本数据,或者执行大量的自动化任务,并且解决各式各样的问题,如自动摘要(automatic summarization),机器翻译(machine translation,MT),命名实体识别(named entity recognition,NER),关系

提取(relation extraction,RE),信息抽取(information extraction,IE),情感分析,语音识别(speech recognition),问答系统(question answering)以及主题分割等等。

[0203] (5)预训练语言模型(pre-trained language model)

[0204] 预训练语言模型是一个自然语言序列编码器,为自然语言序列中的每个词进行编码成为一个向量表示,从而进行预测任务。它的训练包含两个阶段。在预训练(pre-training)阶段,该模型在大规模无监督文本上进行语言模型任务的训练,从而学习到一个词表示。在微调(finetuning)阶段,该模型利用预训练阶段学到的参数做初始化,在文本分类(text classification),序列标注(sequence labeling)等下游任务(downstream task)上进行较少步骤的训练,就可以成功把预训练得到的语义信息成功迁移到下游任务上来。

[0205] (6)反向传播算法

[0206] 卷积神经网络可以采用误差反向传播(back propagation,BP)算法在训练过程中修正初始的超分辨率模型中参数的大小,使得超分辨率模型的重建误差损失越来越小。具体地,前向传递输入信号直至输出会产生误差损失,通过反向传播误差损失信息来更新初始的超分辨率模型中参数,从而使误差损失收敛。反向传播算法是以误差损失为主导的反向传播运动,旨在得到最优的超分辨率模型的参数,例如权重矩阵。

[0207] (7)损失函数

[0208] 在训练深度神经网络的过程中,因为希望深度神经网络的输出尽可能的接近真正想要预测的值,所以可以通过比较当前网络的预测值和真正想要的目标值,再根据两者之间的差异情况来更新每一层神经网络的权重向量(当然,在第一次更新之前通常会有初始化的过程,即为深度神经网络中的各层预先配置参数),比如,如果网络的预测值高了,就调整权重向量让它预测低一些,不断地调整,直到深度神经网络能够预测出真正想要的目标值或与真正想要的目标值非常接近的值。因此,就需要预先定义“如何比较预测值和目标值之间的差异”,这便是损失函数(loss function)或目标函数(objective function),它们是用于衡量预测值和目标值的差异的重要方程。其中,以损失函数举例,损失函数的输出值(loss)越高表示差异越大,那么深度神经网络的训练就变成了尽可能缩小这个loss的过程。

[0209] 现有的跨模态视觉语言模型(或者可以称之为多模态模型或者多模态语言模型)所采用的对齐方式中的文本建模能力较弱。以TIA为例,TIA只是简单的判断某个字符是否被遮挡,并没有很好的考虑到遮挡字符之间的语义信息导致模型不是真正的理解文档的内容。

[0210] 当前的多模态模型的图文对齐能力较弱,也就是不具备文本和图像元素的对齐能力。以图文匹配(text-image matching,TIM)为例,训练目标用来判断全局文字是否和全局的扫描文档图片是否匹配,该任务学习难度小,使得训练的模型不具备文本和图像元素的对齐能力。

[0211] 为了解决上述问题,本申请实施例提供了一种数据处理方法。下面结合附图对本申请实施例的数据处理方法进行详细的介绍。

[0212] 参照图7,图7为本申请实施例提供的一种数据处理方法的流程示意,如图7所示,本申请实施例提供的一种数据处理方法,可以包括步骤701至703,下面分别对这些步骤进

行详细的描述。

[0213] 701、获取第一特征表示以及第二特征表示，所述第一特征表示为根据图像编码器对图像进行处理得到的图像特征得到，所述第二特征表示为第一文本的文本特征；所述第一文本为所述图像中包括的文本内容。

[0214] 在一种可能的实现中，可以获取到原始的输入图像，并通过图像编码器对原始的输入图像进行编码，得到输入图像的编码结果，该编码结果可以包括输入图像的图像特征表示。应理解，本申请中的原始的输入图像可以指输入到扩模态的语言模型中的图像、或者是对扩模态的语言模型中的图像进行预处理后的图像。

[0215] 在一种可能的实现中，图像编码器可以对输入图像进行特征提取，其中，图像编码器具体可以表现为卷积神经网络、方向梯度直方图(histogram of oriented gradient，HOG)、局部二值模式(local binarypattern，LBP)或其他用于对图像进行特征提取的神经网络等。

[0216] 在一种可能的实现中，可以通过图像编码器对输入图像进行处理，得到输入图像对应的图像特征表示。

[0217] 在一种可能的实现中，可以识别出图像中所包含的文本。例如可以通过光学字符识别(optical character recognition，OCR)来识别出图像中所包含的文本。可选的，还可以识别出文本中各个文本单元对应的位置信息，该位置信息可以指示文本单元所在的区域在图像中的位置，例如，文本单元所在的区域可以通过矩形框表示，位置信息可以为矩形框的左上点坐标和右下点坐标。

[0218] 在一种可能的实现中，可以获取到文本对应的文本特征表示。

[0219] 在一种可能的实现中，可以通过嵌入层对文本进行嵌入处理，以得到嵌入向量，该嵌入向量可以为文本对应的文本特征表示。

[0220] 其中，嵌入层可以称为输入嵌入(input embedding)层。当前输入可以为文本中的每个文本单元。嵌入层在获取当前输入后，可以对该当前输入中各个文本单元进行嵌入处理，可得到各个文本单元对应的嵌入向量。

[0221] 在一些实施例中，还可以获取每个文本单元的位置向量，所述位置向量用于指示文本单元的位置；其中，位置用于表示文本单元在文本中的位置，具体的，所述位置用于指示文本单元与文本中其他文本单元之间的相对位置关系。

[0222] 在一种实现中，所述嵌入层可以包括输入嵌入层和位置编码(positional encoding)层。在输入嵌入层，可以对当前输入中的各个文本单元进行词嵌入处理，从而得到各个文本单元的嵌入向量。在位置编码层，可以获取各个文本单元在该当前输入中的位置，进而对各个文本单元的位置生成位置向量。

[0223] 在一些示例中，各个文本单元在文本中的位置可以为各个文本单元在文本中的绝对位置。以当前输入为“几号应还花呗”为例，其中的“几”的位置可以表示为第一位，“号”的位置可以表示为第二位，……。在一些示例中，各个文本单元在文本中的位置可以为各个文本单元在文本中的相对位置。仍以当前输入为“几号应还花呗”为例，其中的“几”的位置可以表示为“号”之前，“号”的位置可以表示为“几”之后、“应”之前，……。当得到当前输入中各个文本单元的嵌入向量和位置向量时，可以将各个文本单元的位置向量和对应的嵌入向量进行融合，得到各个文本单元的嵌入向量，即得到该当前输入对应的多个嵌入向量。应理

解,融合的方式可以是对嵌入向量和位置向量进行加法运算,或者是通过其他运算使得嵌入向量可以携带文本中的一个文本单元以及所述一个文本单元在所述文本中的位置的信息,这里并不限定具体的融合方式。例如,融合的方式包括但不限于拼接(conact)、相加(add)、融合(fusion)和相乘等。

[0224] 在一种可能的实现中,可以通过嵌入层对文本进行嵌入处理,以得到嵌入向量,并通过文本编码器对嵌入向量进行处理,得到文本对应的文本特征表示。

[0225] 本申请实施例中,通过图文对比学习(text-image contrastive learning,TIC)来增强图文之间的匹配能力,也就是提高模型对于图像信息和文本信息这种跨模态的数据的信息处理能力。具体的,针对于图像的图像特征和文本的文本特征,通过相似度比对构建的损失,来更新图形编码器(如果文本特征为通过文本编码器得到的,也可以更新文本编码器)。虽然图像的图像特征和图像中文本的文本特征不是同一个模态的特征,但是由于文本是图像中的文本,在图像特征中也会蕴含一定的文本的语义信息,因此,图像特征和文本特征(属于同一个图像)之间在语义维度(或者其他信息维度上)会存在关联。上述相似度可以包含图像中语义信息和文本的语义信息之间的相似度,因此,基于该相似度构建的损失来更新图像编码器以及文本编码器(如果架构中存在文本编码器的话),能够使得图像编码器和文本编码器提取的特征中蕴含更准确的图像中文本的语义信息,进而增强后续网络的图文匹配能力。

[0226] 此外,在跨模态的语言模型的架构中,包括图形编码器、文本编码器以及用于提供特征之间交互信息的目标编码器,本申请中,将模型的中间输出(也就是图形编码器和文本编码器的输出)之间的相似度构建的损失来更新图形编码器和文本编码器,使得基于图形编码器和文本编码器输出的特征表示就可以实现下游任务(精度没有基于目标编码器输出的特征表示进行的下游任务高),而在一些场景中,由于下游任务要处理的数据的数量较大,因此可以使用图形编码器和文本编码器输出的特征表示进行粗排,使用目标编码器输出的特征表示进行精排,以提高召回率。

[0227] 以需要进行图文对比学习(也就是文本-图像对齐)的对象为图像以及第一文本为例进行说明:

[0228] 在一种可能的实现中,图像可以为原始的输入图像(或者是原始的输入图像的部分图像,但该部分图像包含输入图像的全部文本)。第一文本可以为图像中包含的全部文本。

[0229] 在一种可能的实现中,图像可以为原始的输入图像(或者是原始的输入图像的部分图像,但该部分图像包含输入图像的全部文本)。第一文本可以为图像中包含的文本中的部分文本。

[0230] 在一种可能的实现中,图像可以为对原始的输入图像进行提取得到的部分图像区域,该部分图像区域内包括原始的输入图像中全部文本的部分文本,第一文本可以为图像中包括的全部文本。

[0231] 在一种可能的实现中,图像可以为对原始的输入图像进行提取得到的部分图像区域,该部分图像区域内包括原始的输入图像中全部文本的部分文本,第一文本可以为图像中包含的文本中的部分文本。

[0232] 在一种可能的实现中,输入的原始图像中可以包括一行或多行文本,或者,输入的

原始图像中可以包括一列或多列文本。可选的,图像中可以包括原始图像的一行或者一列文本,第一文本可以为一个或多个文本单元。

[0233] 在进行图文对比学习时,可以根据图像的图像特征和第一文本的文本特征之间的相似度来构建用于更新图形编码器以及文本编码器的损失,可选的,该相似度与图像中所蕴含的文本语义信息和文本自身的语义信息之间的相似度有关。

[0234] 应理解,还可以根据图像的图像特征和图像中不包括的文本的文本特征之间的相似度来构建用于更新图形编码器以及文本编码器的损失,区别在于,图像的图像特征和图像中文本的文本特征为正例(也就是对应的标注相似度高),图像的图像特征和图像中不包括的文本的文本特征为负例(也就是对应的标注相似度低)。

[0235] 在一种可能的实现中,上述相似度可以为:文本特征和图像特征之间的相似度,也可以为图像特征和文本特征之间的相似度。

[0236] 在一种可能的实现中,针对于同一张原始的输入图像,可以将输入图像整体对应的图像特征和输入图像中包括的全部文本的文本特征进行比对,也可以将输入图像中的部分图像区域的图像特征和输入图像中包括的部分文本的文本特征进行比对,也可以将输入图像整体对应的图像特征和输入图像中包括的部分文本的文本特征进行比对,或者是上述几种方式的组合。

[0237] 在一种可能的实现中,针对于同一张原始的输入图像,可以将输入图像中包括的全部文本整体对应的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将输入图像中包括的部分文本的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将输入图像中包括的全部文本整体对应的文本特征和输入图像整体对应的图像特征进行比对。

[0238] 在一种可能的实现中,针对于不同的原始输入图像,可以将一部分输入图像中包括的全部文本整体对应的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将一部分输入图像中包括的部分文本的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将一部分输入图像中包括的全部文本整体对应的文本特征和输入图像整体对应的图像特征进行比对,或者是上述几种方式的组合。

[0239] 在一种可能的实现中,针对于不同的原始输入图像,可以将一部分输入图像中包括的全部文本整体对应的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将一部分输入图像中包括的部分文本的文本特征和输入图像中的部分图像区域的图像特征进行比对,也可以将一部分输入图像中包括的全部文本整体对应的文本特征和输入图像整体对应的图像特征进行比对。

[0240] 在一种可能的实现中,可以根据图像编码器得到的图像特征,通过图像区域特征提取模块(例如可以通过感兴趣区域(region of interest,ROI)head实现)来确定图像中的部分图像区域的图像特征,例如图像区域特征提取模块可以基于要去提取的图像区域的位置信息,对图像区域的图像特征(图像编码器输出的)进行双线性插值,得到部分图像区域的图像特征。

[0241] 通过上述方式,利用粗粒度和细粒度对比学习,可以增强图文整体匹配和区域-文本行匹配能力,进而可以提升模型在下游任务如信息提取任务上的性能。

[0242] 示例性的,在构建图文对比学习时,图像-文字对比学习的损失函数可以定义为:

[0243] $$\mathcal{L}_k^i(x_k^i,\{x_j^t\}_{j=1}^b)=-\frac{1}{b}\log\frac{\exp(s_{k,k}^i)}{\sum_{j=1}^b\exp(s_{k,j}^i)};$$

[0244] 其中，$s_{k,j}^i$表示第k个图像和第j个文本之间的相似度。

[0245] 示例性的，文字-图像对比学习的损失函数可以定义为：

[0246] $$\mathcal{L}_k^t(x_k^t,\{x_j^i\}_{j=1}^b)=-\frac{1}{b}\log\frac{\exp(s_{k,k}^t)}{\sum_{j=1}^b\exp(s_{k,j}^t)};$$

[0247] 其中，$s_{k,j}^i$表示第k个文本和第j个图像之间的相似度。

[0248] 示例性的，文本行-图像区域总的对比学习损失函数可以定义为：

[0249] $$\mathcal{L}_{\text{trc}}=\frac{1}{2}\sum_{k=1}^b\left(\mathcal{L}_k^i+\mathcal{L}_k^t\right);$$

[0250] 为了建立细粒度的相似性矩阵，图像-文本行区域之间的双向相似度可以分别表述为：

[0251] $$s_{k,j}^i=\frac{1}{T}\sum_{p=1}^T v_{k,i}^{\mathrm{T}}\ t_{j,m_p},\ s_{k,j}^t=\frac{1}{T}\sum_{q=1}^T v_{k,i}^{T}\ t_{j,m_q};$$

[0252] 其中，$m_p=\arg\max_r\ v_{k,p}^{\top}\ t_{j,r}$，$m_q=\arg\max_r\ v_{k,q}^{\top}\ t_{j,r}$。

[0253] 接下来介绍一个图文对比学习(text-image contrastive learning，TIC)的应用实施例：

[0254] 参照图8，该实施例可以分为以下几个步骤：

[0255] 1)将图片缩放到某个分辨率(如448x448)；

[0256] 2)利用OCR工具获取文档的文字行内容和位置(bounding box)；

[0257] 3)利用MaskRCNN抽取文本行的视觉表征；

[0258] 4)利用Roberta抽取文档文本行的文字表征；

[0259] 5)构建视觉-文本匹配矩阵，选择视觉(文本)对应最匹配的文本(视觉)token；

[0260] 6)优化图文对比学习TIC loss。

[0261] 702、根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示；所述目标编码器用于实现特征表示之间的交互。

[0262] 其中，本申请通过先双塔方式可以提升多模态的对齐能力，再利用一个多模态的单塔结构进一步增强特征的交互学习能力，可以利于高效文档检索等下游任务(先粗排后精排)。

[0263] 在一种可能的实现中，可以根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示，第三特征表示可以包括第一文本对应的文本特征、以及图

像对应的图像特征,和第一特征表示以及所述第二特征表示不同的是,目标编码器输出的特征为根据第一特征表示以及所述第二特征表示之间的交互学习得到的。

[0264] 703、根据所述第三特征表示,执行下游任务,得到执行结果;所述执行结果和对应的标注之间的相似度以及所述第一特征表示和所述第二特征表示之间的相似度用于更新所述图像编码器。

[0265] 在一种可能的实现中,目标编码器可以得到第三特征表示,第三特征表示可以用于执行下游任务,以得到执行结果,该执行结果可以和对应的标注构建损失,以更新图形编码器、文本编码器(如果有的话)以及目标编码器。

[0266] 在一种可能的实现中,可以通过设置不同的下游任务来构建不同的损失。

[0267] 在一种可能的实现中,下游任务可以包括文本框-网格对齐(bounding-box grid matching,BGM)。

[0268] 在一种可能的实现中,所述第一文本可以包括第一子文本和第二子文本,所述第二特征表示包括所述第一子文本对应的第一子特征、以及所述第二子文本对应的第二子特征;所述第一子特征不包含所述第一子文本在所述图像中的位置;所述第二子特征包含所述第二子文本在所述图像中的位置;也就是说,在生成第一文本的文本特征时,针对于第一文本中的部分文本(例如第二子文本),可以将文本以及位置信息生成对应的文本特征(例如第二子特征),针对于第一文本中的部分文本(例如第一子文本),可以将文本以及位置信息生成对应的文本特征(例如第一子特征),相当于所述第一子特征不包含所述第一子文本在所述图像中的位置;所述第二子特征包含所述第二子文本在所述图像中的位置。

[0269] 在一种可能的实现中,可以通过光学字符识别(optical character recognition,OCR)来识别出图像中所包含的文本以及对应的位置信息,该位置信息可以指示文本所在的区域在图像中的位置,例如,文本所在的区域可以通过矩形框表示,位置信息可以为矩形框的左上点坐标和右下点坐标。

[0270] 进而,第三特征表示中可以包含第二子文本的位置信息,而不包含第一子文本的位置信息(这里的不包含可以理解为由于输入所导致的,但是可以隐含包含)可以根据第三特征表示预测得到。具体的,可以根据所述第三特征表示,预测所述第一子文本在所述图像中的第一位置;所述第一位置和对应的标注(也就是第一子文本在图像中的真实位置)之间的相似度用于更新所述图像编码器以及所述目标编码器。

[0271] 在一种可能的实现中,所述图像包括多个图像块;所述第一位置为对所述第一子文本预测所在的图像块;所述标注为所述第一子文本真实所在的图像块。

[0272] 示例性的,参照图9,本实施例可以分为以下几个步骤:

[0273] 1)将图片缩放到某个分辨率(如448x448);

[0274] 2)将图片区域划分为若干区域(如4x4个grid);

[0275] 3)利用OCR得到每个文字的文本框(bounding box)坐标;

[0276] 4)随机选取一定比例不包含掩码语言建模(mask language modeling,MLM)的文本行(如果下游任务包括MLM的话),并遮盖住该文本行中每个文字的文本框坐标(例如可以将文本框坐标设置为[0,0,0,0]);

[0277] 5)判断这些被遮盖住位置信息的文字处于图片中哪个区域

[0278] 每个单词的特征可以表示为多模态特征的聚合,如:

[0279] $E_i = E_{word}(t_i) + E_{layout}(t_i)$

[0280] 其中E_{word}和E_{layout}分别表示第i个词的词嵌入和布局信息。

[0281] 通过遮盖住第i个词的布局信息，然后预测该词在原图中空间分区的类别，则BGM的损失函数可以定义为：

[0282] $$\mathcal{L}_{BGM} = \sum_i \sum_j (-y_{ij} \log(p_{ij}))$$ ；

[0283] 其中，y_{ij}表示第i个词的真实标签，p_{ij}表示第i个词是属于第j个图像区域的概率。

[0284] 本申请在单塔部分提出BGM来提升空间布局感知能力，有助于提升位置敏感的任务中模型的性能。例如在信息提取中，key-value对通常是相邻的成对数据。因为模型在预训练阶段已经学会了对位置以及布局信息的感知能力，所以在信息抽取这类任务中模型会有很好的性能。

[0285] 此外，下游任务还可以包括MLM，具体的，在模型训练的前馈过程中，可以对文本中的文本单元进行掩码(例如可以为随机掩码)，掩码后的文本单元可以作为预测文本，未被掩码的文本单元可以作为已知文本，模型可以基于未被掩码的文本单元(或者已经预测出的掩码后的文本单元)，依次进行掩码后的文本单元所在文本位置的文本预测，例如，文本可以为“the cat sat on the mat”，对文本进行掩码后，可以得到“the_sat_the mat”，其中，这里的符号“_”的含义是掩码，而不是指下划线在。第一次预测时，可以通过初始标识位以及未被掩码的文本单元，预测其中一个“_”所处的文本位置的文本。

[0286] 在一种可能的实现中，为了提高图像表征的学习能力，还可以基于图像特征来构建损失，具体的，可以对图像中的部分进行掩码，通过图像编码器得到的结果对掩码区域图像进行图像重建，并基于重建结果和掩码区域的真实图像的像素值之间的差异来更新网络。然而，现有方法图像表征学习能力较弱，具体是由于，为了降低模型的处理算例开销，需要将图像压缩到一个较小的分辨率(例如224*224)，在图像中包含密集型文本的情况下，图像中的像素难以表达出准确的文字信息，训练后的模型的图像表征的学习能力有限。

[0287] 本申请实施例中，为了提高图像表征的学习能力，可以不对图像而是对图像的图像特征进行掩码，并对掩码区域的图像特征进行特征重建，并基于重建结果和掩码区域的图像特征之间的差异来更新网络，由于图像特征相比像素值本身可以携带更丰富的图像信息以及文本信息，因此可以提高训练后的网络的图像表征的学习能力，上述方式可以称之为掩码图像建模(mask image modeling，MIM)。

[0288] 在一种可能的实现中，所述第一特征表示包括第三子特征和第四子特征；可以根据所述第四子特征、所述第四子特征在所述第一特征表示中的第二位置、以及所述第三子特征在所述第一特征表示中的第三位置，得到所述第三位置处的特征预测值；所述特征预测值和所述第三子特征之间的相似度用于更新所述图像编码器。

[0289] 在一种可能的实现中，可以根据所述第四子特征、所述第四子特征在所述第一特征表示中的第二位置、以及所述第三子特征在所述第一特征表示中的第三位置，通过自注意力网络，得到第四特征表示；根据所述第四特征表示，通过预测网络，得到所述第三位置处的特征预测值。

[0290] 应理解，进行掩码的对象(第一特征表示)可以是图像编码器得到的，也可以是通过对图像编码器得到的图像特征进行一定处理得到的，例如是通过图像区域特征提取模块

(例如可以通过感兴趣区域(region of interest,ROI)head实现)对图像中的部分图像区域确定的图像特征。

[0291] 示例性的,参照图10,本实施例可以分为以下几个步骤:

[0292] 1)将图片缩放到某个分辨率(如448x448);

[0293] 2)计算当前图片的卷积特征图;

[0294] 3)计算每个文本框区域的视觉特征(ROI features);

[0295] 4)随机选取一定比例不包含MLM的文本行(如果下游任务包括MLM的话),在卷积网络的特征图中遮盖住该文本行对应区域;

[0296] 5)将所有ROI features(包括被遮盖的ROI特征)送入self-attention;

[0297] 6)利用contrastive/L1等loss重建被遮盖的ROI特征;

[0298] 为了增强图像分支的表征学习,先使用ROI对齐的方式提取每个文本行的视觉特征:

[0299] $$f_i = Wf_{ROI}^i + b$$;

[0300] 其中f_{ROI}^i表示第i个文本行的ROI视觉特征,W和b是网络参数。

[0301] 可选的,可以利用注意力机制重新计算每个文本行视觉特征的新聚合表示:

[0302] H=Atten(f;Φ)=Atten({fi};Φ);

[0303] 其中,H是ROI特征的隐空间表征,Atten是自注意里模块,Φ是网络参数。

[0304] 图像中文本行区域的视觉掩码特征重建损失函数可以表示为:

[0305] $$\mathcal{L}_{\mathrm{ROI}} = \sum_i \mathrm{smooth}_{\mathrm{L1}}(H_i - H_i^{ROI_{mask}})$$;

[0306] 其中,$H_i^{ROI_{mask}}$表示第i个被遮盖的文本框的视觉重建特征,H_i表示$H_i^{ROI_{mask}}$所对应的groundtruth labels。

[0307] 本实施例采用更高的分辨率配合MIM(Masked Image Modeling)来提升文本行区域的视觉表征学习。基于特征图角度重建遮掩视觉特征图,利用相邻unmasked文本行ROI特征帮助重建被遮盖的文本行视觉特征。通过增强视觉表征的学习能力,一方面可以进一步提升多模态特征对齐,另一方面可以提升模型在一些下游任务如图文分类上的性能。

[0308] 接下来结合实验结果介绍本申请的有益效果:

[0309] 如表1所示,本申请在公开数据集如信息提取数据集(FUNSD,CORD和SROI)和图文分类数据集(如RVL-CDIP)上均超越SOTA性能。

[0310] 表1

[0311]

方法	模型	信息抽取 (F1)			文档分类 (Acc)
		FUNSD	CORD	SROIE	RVL-CDIP
语言模型 (OCR+NLP)	BERT	0.60	0.89	0.91	89.9%
	RoBERTa	0.71	0.93	-	90.1%
	UniLMv2	0.66	0.91	0.94	90.2%
多模态模型 (OCR+多模态)	StructuralLM	0.85	-	-	-
	SelfDoc	0.83	-	-	93.8%
	DocFormer	0.83	0.96	-	-
	LayoutLM	0.79	0.94	0.94	94.4%
	LayoutLMv2	0.83	0.95	0.96	95.6%
	LayoutLMv3	0.90	0.96	-	-
	Ernie-layout	0.9312	0.97	0.97	95.4%
	Ours	0.9339	0.97	0.97	95.7%

[0312] 本申请提供了一种数据处理方法，包括：获取第一特征表示以及第二特征表示，所述第一特征表示为根据图像编码器对图像进行处理得到的图像特征得到，所述第二特征表示为第一文本的文本特征；所述第一文本为所述图像中包括的文本内容；所述第一特征表示和所述第二特征表示之间的相似度用于更新所述图像编码器；根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示；所述目标编码器用于实现特征表示之间的交互；根据所述第三特征表示，执行下游任务，得到执行结果。

[0313] 本申请实施例中，通过图文对比学习(text-image contrastive learning，TIC)来增强图文之间的匹配能力，具体的，针对于图像的图像特征和文本的文本特征，通过相似度比对构建的损失，来更新图形编码器(如果文本特征为通过文本编码器得到的，也可以更新文本编码器)。虽然图像的图像特征和图像中文本的文本特征不是同一个模态的特征，但是由于文本是图像中的文本，在图像特征中也会蕴含一定的文本的语义信息，因此，图像特征和文本特征(属于同一个图像)之间在语义维度(或者其他信息维度上)会存在关联。上述相似度可以包含图像中语义信息和文本的语义信息之间的相似度，因此，基于该相似度构建的损失来更新图像编码器以及文本编码器(如果架构中存在文本编码器的话)，能够使得图像编码器和文本编码器提取的特征中蕴含更准确的图像中文本的语义信息，进而增强后续网络的图文匹配能力。

[0314] 此外，在跨模态的语言模型的架构中，包括图形编码器、文本编码器以及用于提供

特征之间交互信息的目标编码器，本申请中，将模型的中间输出(也就是图形编码器和文本编码器的输出)之间的相似度构建的损失来更新图形编码器和文本编码器，使得基于图形编码器和文本编码器输出的特征表示就可以实现下游任务(精度没有基于目标编码器输出的特征表示进行的下游任务高)，而在一些场景中，由于下游任务要处理的数据的数量较大，因此可以使用图形编码器和文本编码器输出的特征表示进行粗排，使用目标编码器输出的特征表示进行精排，以提高召回率。

[0315] 参照图11，图11为本申请中模型的一个架构示意，其中，采用先双塔对齐图文特征，然后再接单塔构建图文多模态交互模块。在双塔部分，先分别对图像和文字编码提取特征，然后利用粗粒度和细粒度对比学习增强图文整体匹配和区域-文本行匹配能力。在单塔部分提出BGM(Box Grid Matching)来提升空间布局感知能力。为了提升模型对视觉表征的建模能力，在视觉分支采用了更高的分辨率配合MIM(Masked Image Modeling)来提升文本行区域的视觉表征学习。

[0316] 此外，本申请还提供了一种数据处理方法，所述方法包括：

[0317] 获取图像；

[0318] 通过图像编码器，对所述图像进行处理，得到第一特征表示；

[0319] 通过文本编码器，对所述图像中包含的文本进行处理，得到第二特征表示；

[0320] 根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示；所述目标编码器用于实现特征表示之间的交互；

[0321] 根据所述第三特征表示，执行下游任务，得到执行结果。

[0322] 参照图12，图12为本申请实施例提供的一种数据处理装置的结构示意，如图12所示，本申请实施例提供的一种数据处理装置1200，包括：

[0323] 获取模块1201，用于获取第一特征表示以及第二特征表示，所述第一特征表示为根据图像编码器对图像进行处理得到的图像特征得到，所述第二特征表示为第一文本的文本特征；所述第一文本为所述图像中包括的文本内容；

[0324] 其中，关于获取模块1201的具体描述可以参照上述实施例中步骤701的描述，这里不再赘述。

[0325] 编码模块1202，用于根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示；所述目标编码器用于实现特征表示之间的交互；

[0326] 其中，关于编码模块1202的具体描述可以参照上述实施例中步骤702的描述，这里不再赘述。

[0327] 任务执行模块1203，用于根据所述第三特征表示，执行下游任务，得到执行结果；所述执行结果和对应的标注之间的相似度以及所述第一特征表示和所述第二特征表示之间的相似度用于更新所述图像编码器。

[0328] 其中，关于任务执行模块1203的具体描述可以参照上述实施例中步骤703的描述，这里不再赘述。

[0329] 在一种可能的实现中，所述第二特征表示为通过文本编码器对所述第一文本进行处理得到的文本特征；所述第一特征表示和所述第二特征表示之间的相似度还用于更新所述文本编码器。

[0330] 在一种可能的实现中，所述第一特征表示和所述第二特征表示之间的相似度与图

像中所蕴含的文本语义信息和文本自身的语义信息之间的相似度有关。

[0331] 在一种可能的实现中，

[0332] 所述第一文本为所述图像中包含的全部文本；或者，

[0333] 所述第一文本为所述图像中包含的全部文本中的部分。

[0334] 在一种可能的实现中，

[0335] 所述图像为从原始的输入图像中提取的部分图像区域，所述图像包括的文本为所述输入图像包含的文本的部分；或者，

[0336] 所述图像为原始的输入图像。

[0337] 在一种可能的实现中，所述第一文本包括第一子文本和第二子文本，所述第二特征表示包括所述第一子文本对应的第一子特征、以及所述第二子文本对应的第二子特征；所述第一子特征不包含所述第一子文本在所述图像中的位置；所述第二子特征包含所述第二子文本在所述图像中的位置；

[0338] 所述任务执行模块，具体用于：

[0339] 根据所述第三特征表示，预测所述第一子文本在所述图像中的第一位置；所述第一位置和对应的标注之间的相似度用于更新所述图像编码器以及所述目标编码器。

[0340] 在一种可能的实现中，所述图像包括多个图像块；所述第一位置为对所述第一子文本预测所在的图像块；所述标注为所述第一子文本真实所在的图像块。

[0341] 在一种可能的实现中，所述第一特征表示包括第三子特征和第四子特征；所述方法还包括：

[0342] 预测模块，用于根据所述第四子特征、所述第四子特征在所述第一特征表示中的第二位置、以及所述第三子特征在所述第一特征表示中的第三位置，得到所述第三位置处的特征预测值；所述特征预测值和所述第三子特征之间的相似度用于更新所述图像编码器。

[0343] 在一种可能的实现中，所述预测模块，具体用于：

[0344] 根据所述第四子特征、所述第四子特征在所述第一特征表示中的第二位置、以及所述第三子特征在所述第一特征表示中的第三位置，通过自注意力网络，得到第四特征表示；

[0345] 根据所述第四特征表示，通过预测网络，得到所述第三位置处的特征预测值。

[0346] 此外，本申请还提供了一种数据处理装置，所述装置包括：

[0347] 获取模块，用于获取图像；

[0348] 编码模块，用于通过图像编码器，对所述图像进行处理，得到第一特征表示；

[0349] 通过文本编码器，对所述图像中包含的文本进行处理，得到第二特征表示；

[0350] 根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示；所述目标编码器用于实现特征表示之间的交互；

[0351] 任务执行模块，用于根据所述第三特征表示，执行下游任务，得到执行结果。

[0352] 接下来介绍本申请实施例提供的一种执行设备，请参阅图13，图13为本申请实施例提供的执行设备的一种结构示意图，执行设备1300具体可以表现为虚拟现实VR设备、手机、平板、笔记本电脑、智能穿戴设备、监控数据处理设备或服务器等，此处不做限定。具体的，执行设备1300包括：接收器1301、发射器1302、处理器1303和存储器1304(其中执行设备

1300中的处理器1303的数量可以一个或多个,图13中以一个处理器为例),其中,处理器1303可以包括应用处理器13031和通信处理器13032。在本申请的一些实施例中,接收器1301、发射器1302、处理器1303和存储器1304可通过总线或其它方式连接。

[0353] 存储器1304可以包括只读存储器和随机存取存储器,并向处理器1303提供指令和数据。存储器1304的一部分还可以包括非易失性随机存取存储器(non-volatile random access memory,NVRAM)。存储器1304存储有处理器和操作指令、可执行模块或者数据结构,或者它们的子集,或者它们的扩展集,其中,操作指令可包括各种操作指令,用于实现各种操作。

[0354] 处理器1303控制执行设备的操作。具体的应用中,执行设备的各个组件通过总线系统耦合在一起,其中总线系统除包括数据总线之外,还可以包括电源总线、控制总线和状态信号总线等。但是为了清楚说明起见,在图中将各种总线都称为总线系统。

[0355] 上述本申请实施例揭示的方法可以应用于处理器1303中,或者由处理器1303实现。处理器1303可以是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述方法的各步骤可以通过处理器1303中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器1303可以是通用处理器、数字信号处理器(digital signal processing,DSP)、微处理器或微控制器,还可进一步包括专用集成电路(application specific integrated circuit,ASIC)、现场可编程门阵列(field-programmable gate array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。该处理器1303可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成,或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器,闪存、只读存储器,可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器1304,处理器1303读取存储器1304中的信息,结合其硬件完成上述方法中涉及模型推理过程的步骤。

[0356] 接收器1301可用于接收输入的数字或字符信息,以及产生与执行设备的相关设置以及功能控制有关的信号输入。发射器1302可用于通过第一接口输出数字或字符信息;发射器1302还可用于通过第一接口向磁盘组发送指令,以修改磁盘组中的数据;发射器1302还可以包括显示屏等显示设备。

[0357] 本申请实施例还提供了一种训练设备,请参阅图14,图14是本申请实施例提供的训练设备一种结构示意图,具体的,训练设备1400由一个或多个服务器实现,训练设备1400可因配置或性能不同而产生比较大的差异,可以包括一个或一个以上中央处理器(central processing units,CPU)1414(例如,一个或一个以上处理器)和存储器1432,一个或一个以上存储应用程序1442或数据1444的存储介质1430(例如一个或一个以上海量存储设备)。其中,存储器1432和存储介质1430可以是短暂存储或持久存储。存储在存储介质1430的程序可以包括一个或一个以上模块(图示没标出),每个模块可以包括对训练设备中的一系列指令操作。更进一步地,中央处理器1414可以设置为与存储介质1430通信,在训练设备1400上执行存储介质1430中的一系列指令操作。

[0358] 训练设备1400还可以包括一个或一个以上电源1426,一个或一个以上有线或无线网络接口1450,一个或一个以上输入输出接口1458;或,一个或一个以上操作系统1441,例

如Windows ServerTM,Mac OS XTM,UnixTM,LinuxTM,FreeBSDTM等等。

[0359] 本申请实施例中,中央处理器1414,用于执行上述实施例中和模型训练相关的动作。

[0360] 本申请实施例中还提供一种包括计算机程序产品,当其在计算机上运行时,使得计算机执行如前述执行设备所执行的步骤,或者,使得计算机执行如前述训练设备所执行的步骤。

[0361] 本申请实施例中还提供一种计算机可读存储介质,该计算机可读存储介质中存储有用于进行信号处理的程序,当其在计算机上运行时,使得计算机执行如前述执行设备所执行的步骤,或者,使得计算机执行如前述训练设备所执行的步骤。

[0362] 本申请实施例提供的执行设备、训练设备或终端设备具体可以为芯片,芯片包括:处理单元和通信单元,所述处理单元例如可以是处理器,所述通信单元例如可以是输入/输出接口、管脚或电路等。该处理单元可执行存储单元存储的计算机执行指令,以使执行设备内的芯片执行上述实施例描述的数据处理方法,或者,以使训练设备内的芯片执行上述实施例描述的数据处理方法。可选地,所述存储单元为所述芯片内的存储单元,如寄存器、缓存等,所述存储单元还可以是所述无线接入设备端内的位于所述芯片外部的存储单元,如只读存储器(read-only memory,ROM)或可存储静态信息和指令的其他类型的静态存储设备,随机存取存储器(random access memory,RAM)等。

[0363] 具体的,请参阅图15,图15为本申请实施例提供的芯片的一种结构示意图,所述芯片可以表现为神经网络处理器NPU 1500,NPU 1500作为协处理器挂载到主CPU(Host CPU)上,由Host CPU分配任务。NPU的核心部分为运算电路1503,通过控制器1504控制运算电路1503提取存储器中的矩阵数据并进行乘法运算。

[0364] 在一些实现中,运算电路1503内部包括多个处理单元(Process Engine,PE)。在一些实现中,运算电路1503是二维脉动阵列。运算电路1503还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中,运算电路1503是通用的矩阵处理器。

[0365] 举例来说,假设有输入矩阵A,权重矩阵B,输出矩阵C。运算电路从权重存储器1502中取矩阵B相应的数据,并缓存在运算电路中每一个PE上。运算电路从输入存储器1501中取矩阵A数据与矩阵B进行矩阵运算,得到的矩阵的部分结果或最终结果,保存在累加器(accumulator)1508中。

[0366] 统一存储器1506用于存放输入数据以及输出数据。权重数据直接通过存储单元访问控制器(Direct Memory Access Controller,DMAC)1505,DMAC被搬运到权重存储器1502中。输入数据也通过DMAC被搬运到统一存储器1506中。

[0367] BIU为Bus Interface Unit即,总线接口单元1510,用于AXI总线与DMAC和取指存储器(Instruction Fetch Buffer,IFB)1509的交互。

[0368] 总线接口单元1510(Bus Interface Unit,简称BIU),用于取指存储器1509从外部存储器获取指令,还用于存储单元访问控制器1505从外部存储器获取输入矩阵A或者权重矩阵B的原数据。

[0369] DMAC主要用于将外部存储器DDR中的输入数据搬运到统一存储器1506或将权重数据搬运到权重存储器1502中或将输入数据数据搬运到输入存储器1501中。

[0370] 向量计算单元1507包括多个运算处理单元，在需要的情况下，对运算电路1503的输出做进一步处理，如向量乘，向量加，指数运算，对数运算，大小比较等等。主要用于神经网络中非卷积/全连接层网络计算，如Batch Normalization(批归一化)，像素级求和，对特征平面进行上采样等。

[0371] 在一些实现中，向量计算单元1507能将经处理的输出的向量存储到统一存储器1506。例如，向量计算单元1507可以将线性函数；或，非线性函数应用到运算电路1503的输出，例如对卷积层提取的特征平面进行线性插值，再例如累加值的向量，用以生成激活值。在一些实现中，向量计算单元1507生成归一化的值、像素级求和的值，或二者均有。在一些实现中，处理过的输出的向量能够用作到运算电路1503的激活输入，例如用于在神经网络中的后续层中的使用。

[0372] 控制器1504连接的取指存储器(instruction fetch buffer)1509，用于存储控制器1504使用的指令；

[0373] 统一存储器1506，输入存储器1501，权重存储器1502以及取指存储器1509均为On-Chip存储器。外部存储器私有于该NPU硬件架构。

[0374] 其中，上述任一处提到的处理器，可以是一个通用中央处理器，微处理器，ASIC，或一个或多个用于控制上述程序执行的集成电路。

[0375] 另外需说明的是，以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。另外，本申请提供的装置实施例附图中，模块之间的连接关系表示它们之间具有通信连接，具体可以实现为一条或多条通信总线或信号线。

[0376] 通过以上的实施方式的描述，所属领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件的方式来实现，当然也可以通过专用硬件包括专用集成电路、专用CPU、专用存储器、专用元器件等来实现。一般情况下，凡由计算机程序完成的功能都可以很容易地用相应的硬件来实现，而且，用来实现同一功能的具体硬件结构也可以是多种多样的，例如模拟电路、数字电路或专用电路等。但是，对本申请而言更多情况下软件程序实现是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在可读取的存储介质中，如计算机的软盘、U盘、移动硬盘、ROM、RAM、磁碟或者光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，训练设备，或者网络设备等)执行本申请各个实施例所述的方法。

[0377] 在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。

[0378] 所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、训练设备或数据中心通过有线

(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、训练设备或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存储的任何可用介质或者是包含一个或多个可用介质集成的训练设备、数据中心等数据存储设备。所述可用介质可以是磁性介质,(例如,软盘、硬盘、磁带)、光介质(例如,DVD)、或者半导体介质(例如固态硬盘(Solid State Disk,SSD))等。

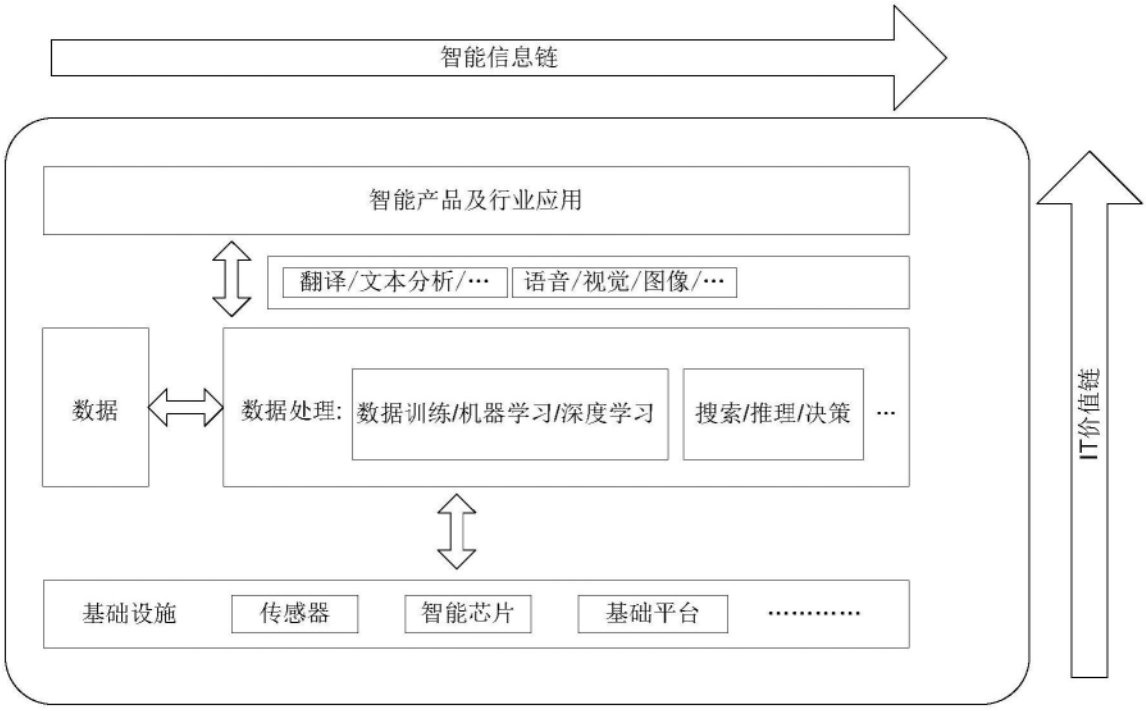

智能信息链
智能产品及行业应用
翻译/文本分析/…
语音/视觉/图像/…
数据
数据处理:
数据训练/机器学习/深度学习
搜索/推理/决策
…
基础设施
传感器
智能芯片
基础平台
……
IT价值链

图1A

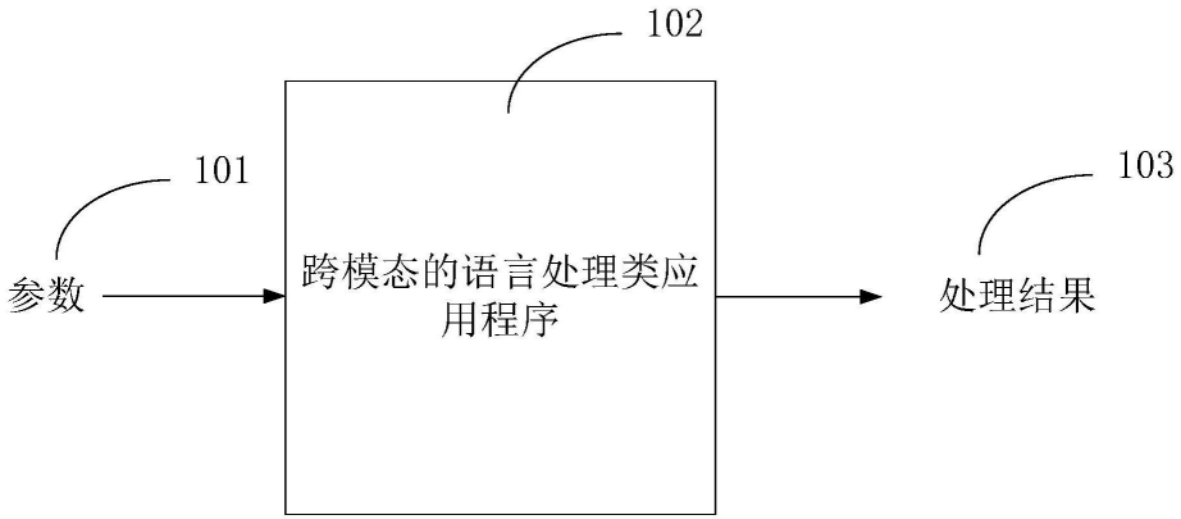

102
101
103
参数
跨模态的语言处理类应用程序
处理结果

图1B

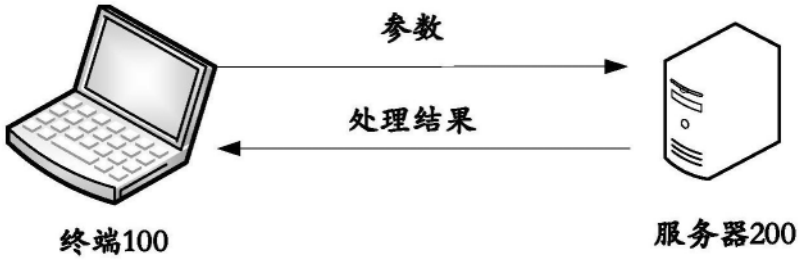

参数
处理结果
终端100
服务器200

图1C

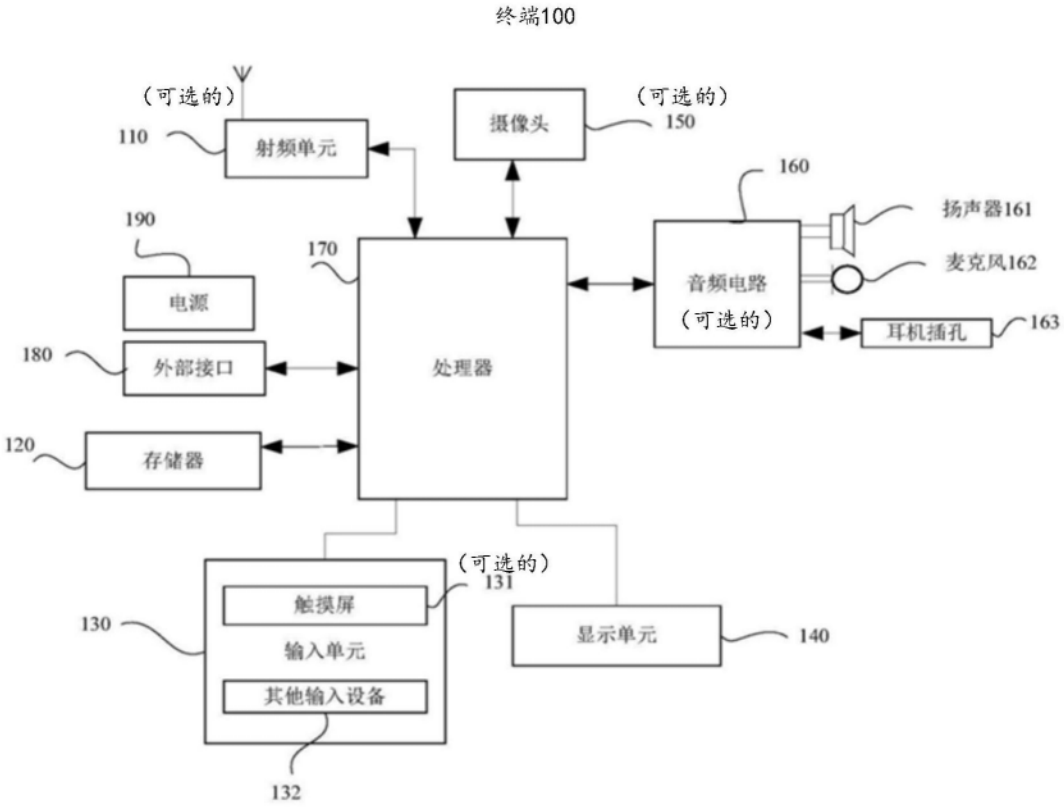

终端100
（可选的）
110
射频单元
摄像头
（可选的）
150
160
扬声器161
麦克风162
190
电源
170
处理器
音频电路
（可选的）
耳机插孔
163
180
外部接口
120
存储器
（可选的）
131
触摸屏
130
输入单元
其他输入设备
132
显示单元
140

图1D

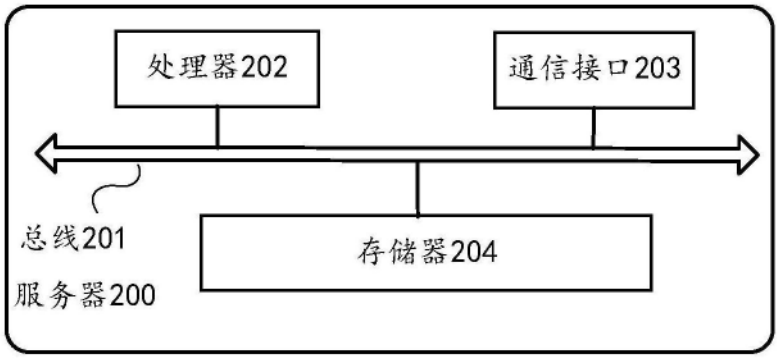

处理器202
通信接口203
总线201
存储器204
服务器200

图2

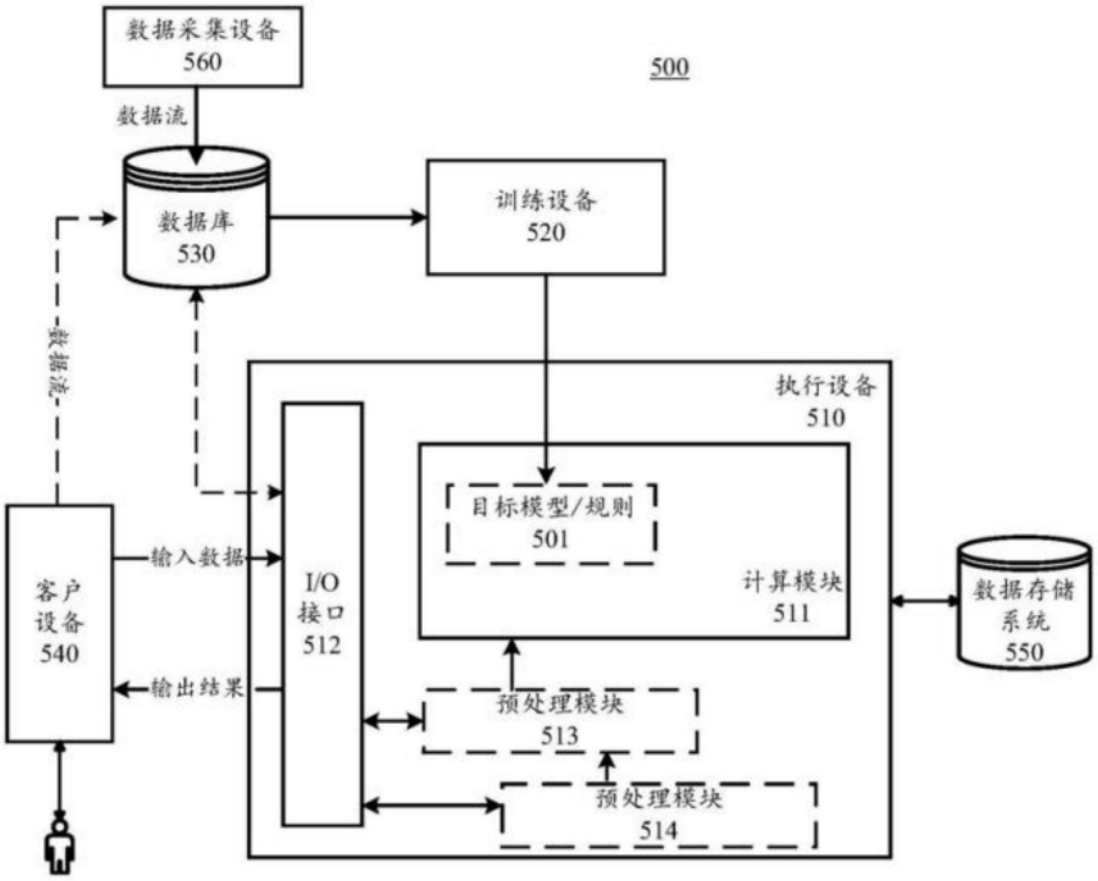

数据采集设备
560
500
数据流
数据库
530
训练设备
520
执行设备
510
目标模型/规则
501
计算模块
511
数据存储
系统
550
数据流
客户
设备
540
输入数据
输出结果
I/O
接口
512
预处理模块
513
预处理模块
514

图3

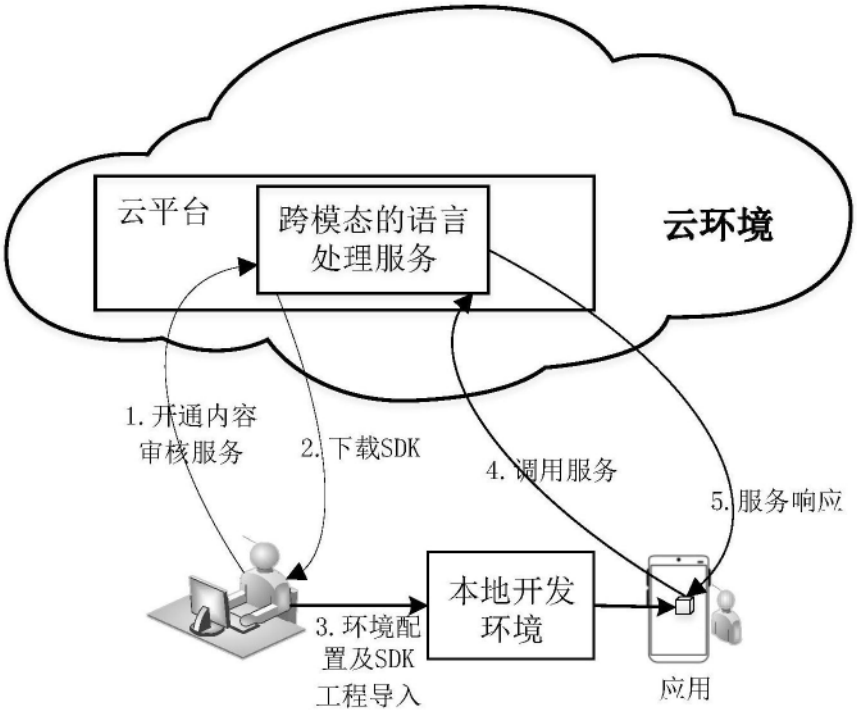

云平台
跨模态的语言处理服务
云环境
1.开通内容审核服务
2.下载SDK
4.调用服务
5.服务响应
本地开发环境
3.环境配置及SDK工程导入
应用

图4

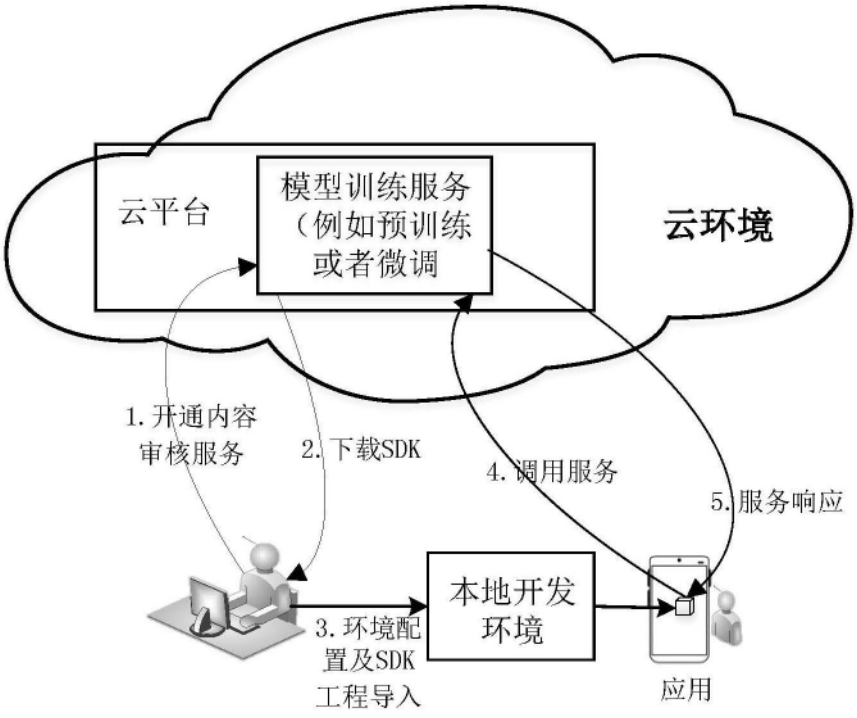

云平台
模型训练服务（例如预训练或者微调
云环境
1.开通内容审核服务
2.下载SDK
4.调用服务
5.服务响应
本地开发环境
3.环境配置及SDK工程导入
应用

图5

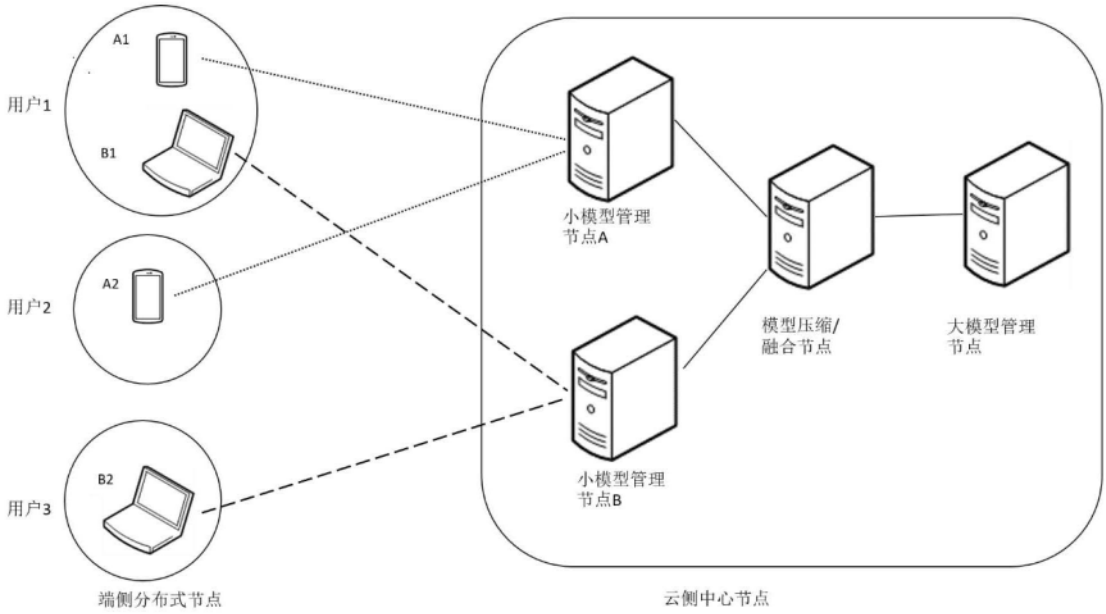

A1
用户1
B1
A2
用户2
B2
用户3
端侧分布式节点
小模型管理
节点A
小模型管理
节点B
模型压缩/
融合节点
大模型管理
节点
云侧中心节点

图6

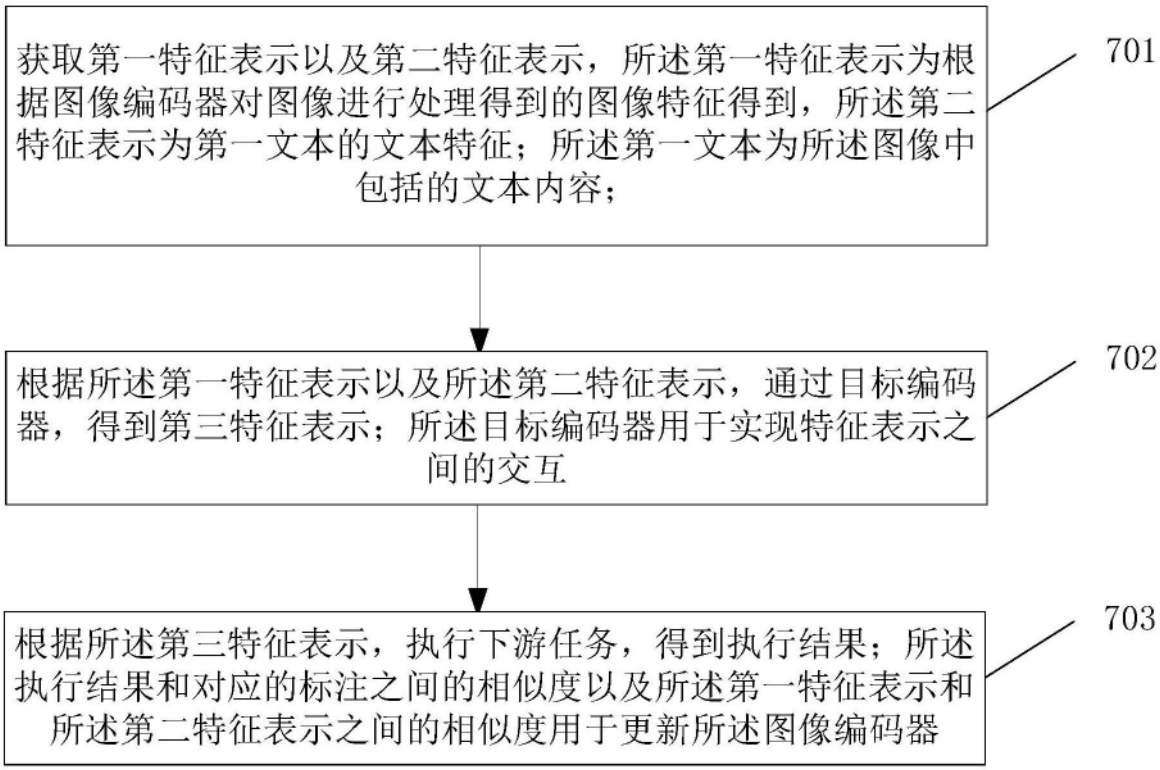

获取第一特征表示以及第二特征表示，所述第一特征表示为根据图像编码器对图像进行处理得到的图像特征得到，所述第二特征表示为第一文本的文本特征；所述第一文本为所述图像中包括的文本内容；
701
根据所述第一特征表示以及所述第二特征表示，通过目标编码器，得到第三特征表示；所述目标编码器用于实现特征表示之间的交互
702
根据所述第三特征表示，执行下游任务，得到执行结果；所述执行结果和对应的标注之间的相似度以及所述第一特征表示和所述第二特征表示之间的相似度用于更新所述图像编码器
703

图7

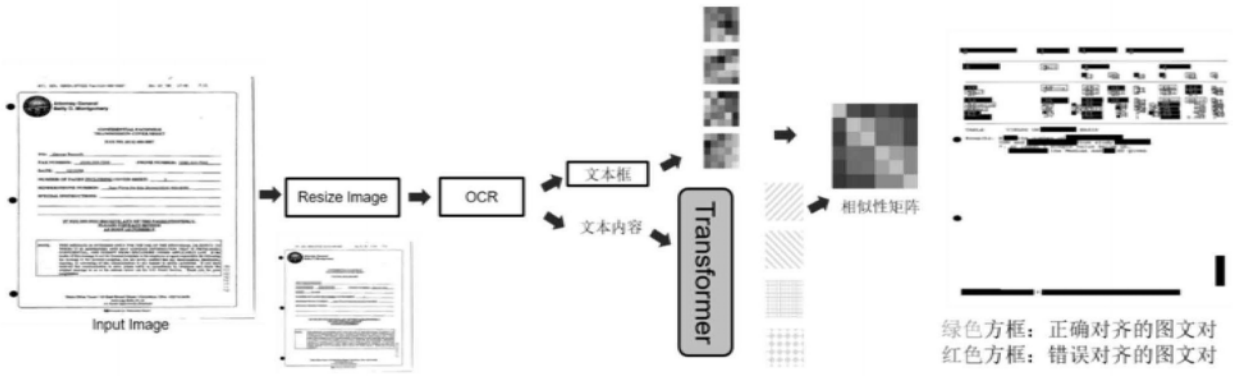

Input Image
Resize Image
OCR
文本框
文本内容
Transformer
相似性矩阵
绿色方框：正确对齐的图文对
红色方框：错误对齐的图文对

图8

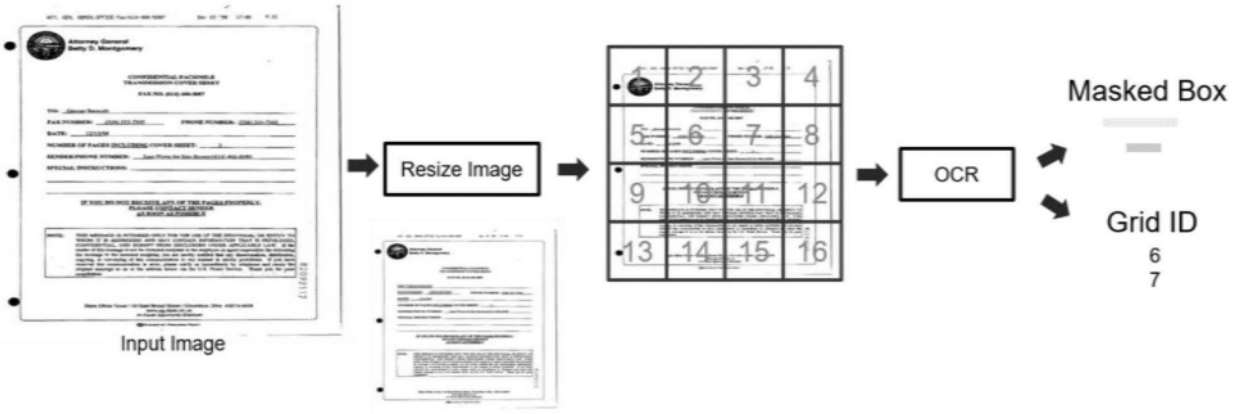

Input Image
Resize Image
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
OCR
Masked Box
Grid ID
6
7

图9

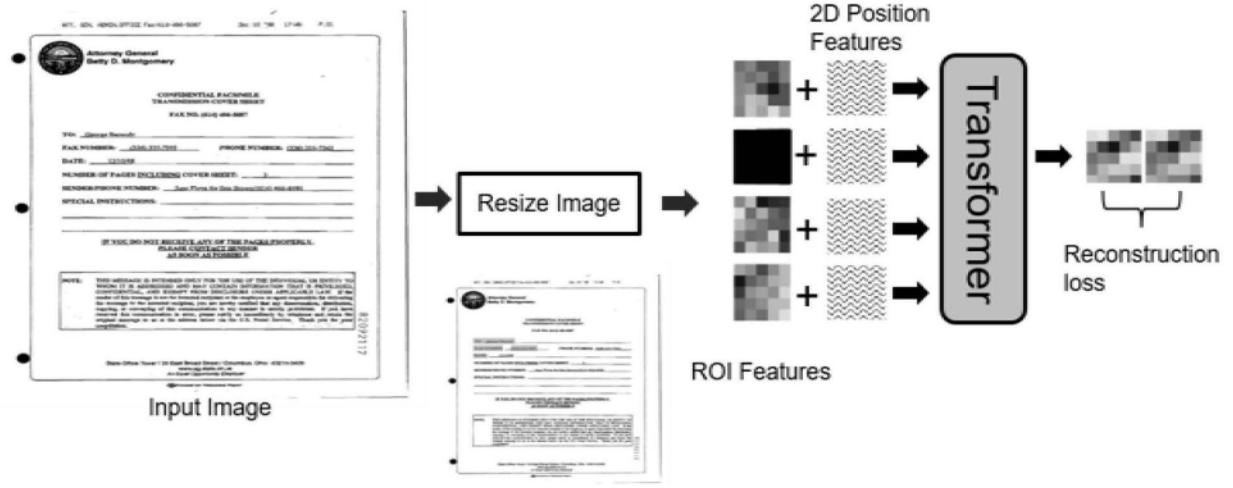

Input Image
Resize Image
2D Position Features
Transformer
Reconstruction loss
ROI Features

图10

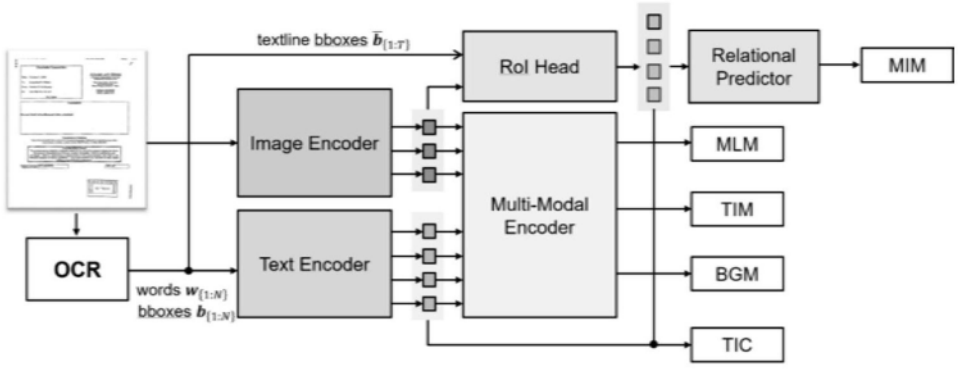

textline bboxes $\bar{b}_{\{1:T\}}$
RoI Head
Relational Predictor
MIM
Image Encoder
MLM
Multi-Modal Encoder
TIM
OCR
Text Encoder
BGM
words $w_{\{1:N\}}$
bboxes $b_{\{1:N\}}$
TIC

图11

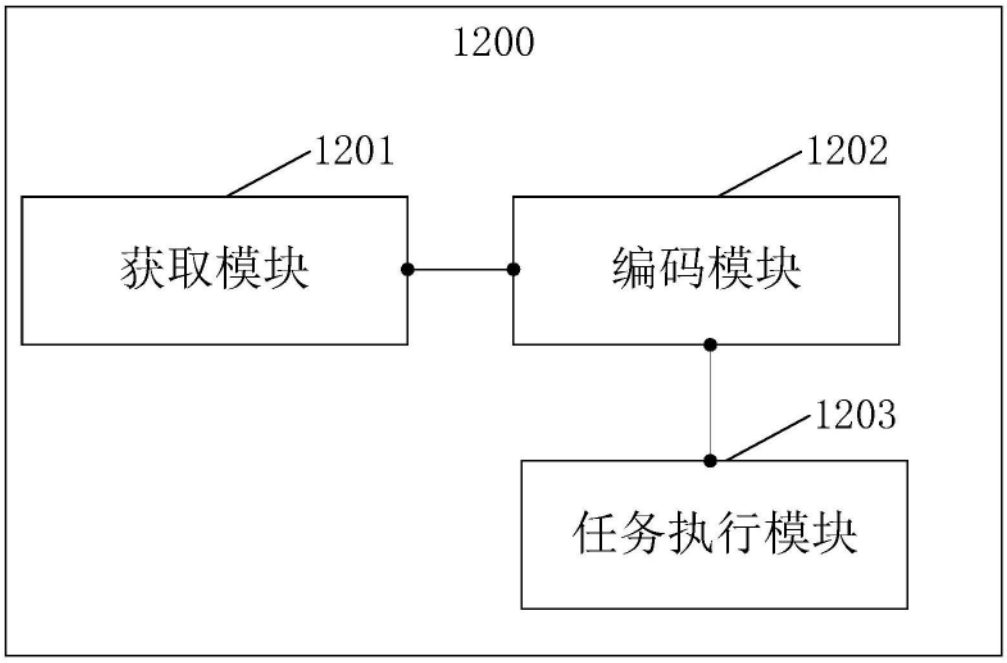

1200
1201
1202
获取模块
编码模块
1203
任务执行模块

图12

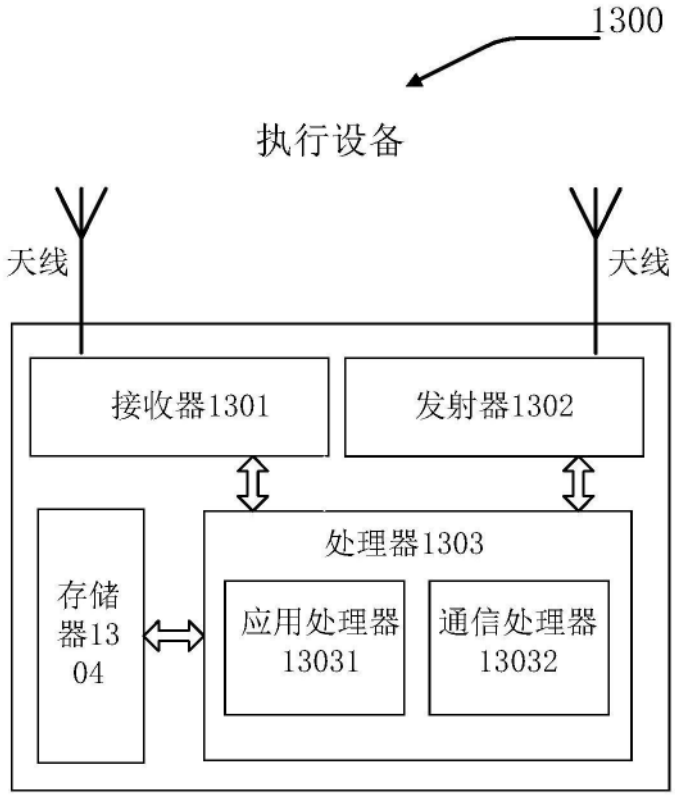

1300
执行设备
天线
天线
接收器1301
发射器1302
处理器1303
存储器1304
应用处理器13031
通信处理器13032

图13

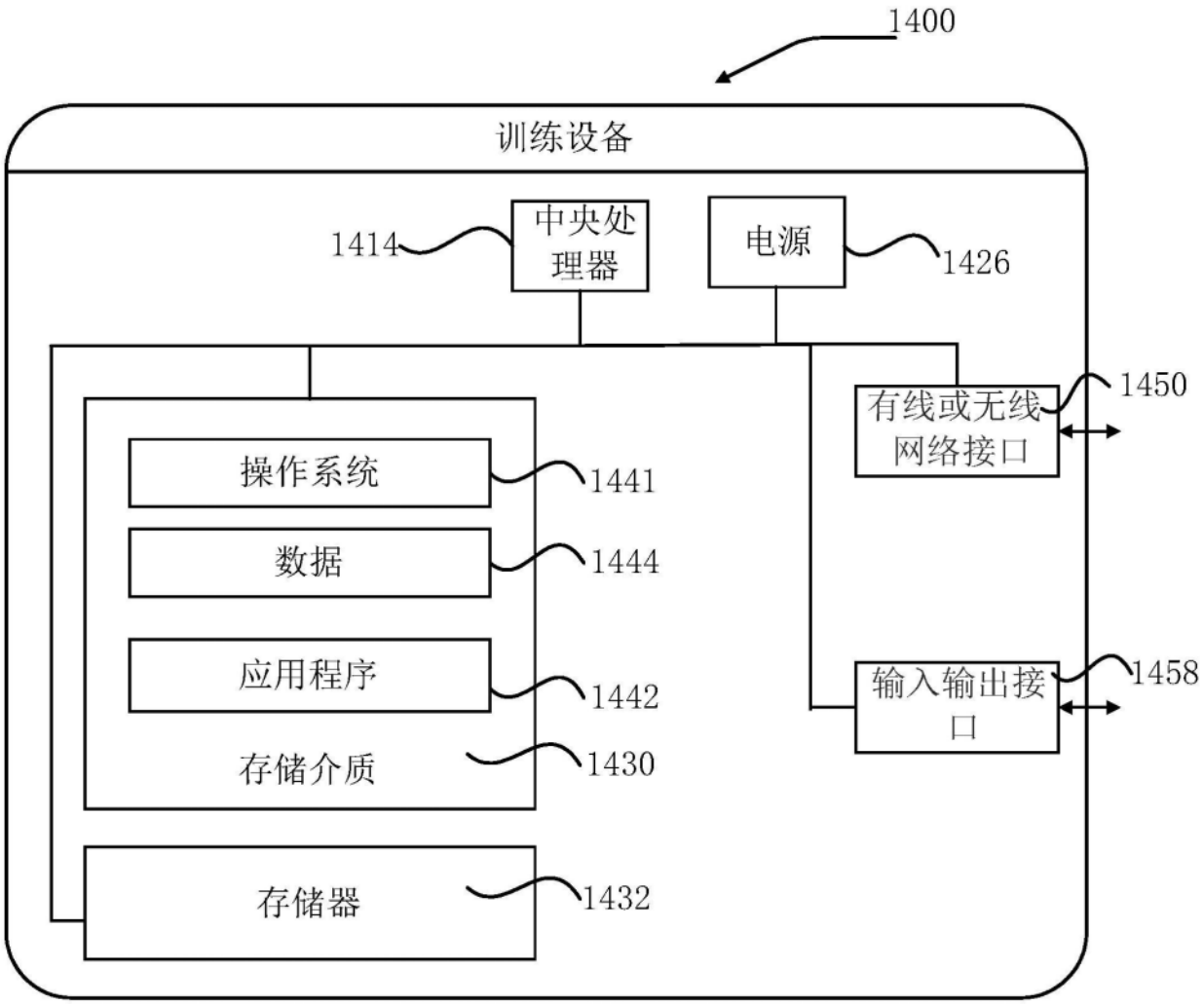

1400
训练设备
1414
中央处理器
电源
1426
1450
有线或无线网络接口
操作系统
1441
数据
1444
应用程序
1442
存储介质
1430
输入输出接口
1458
存储器
1432

图14

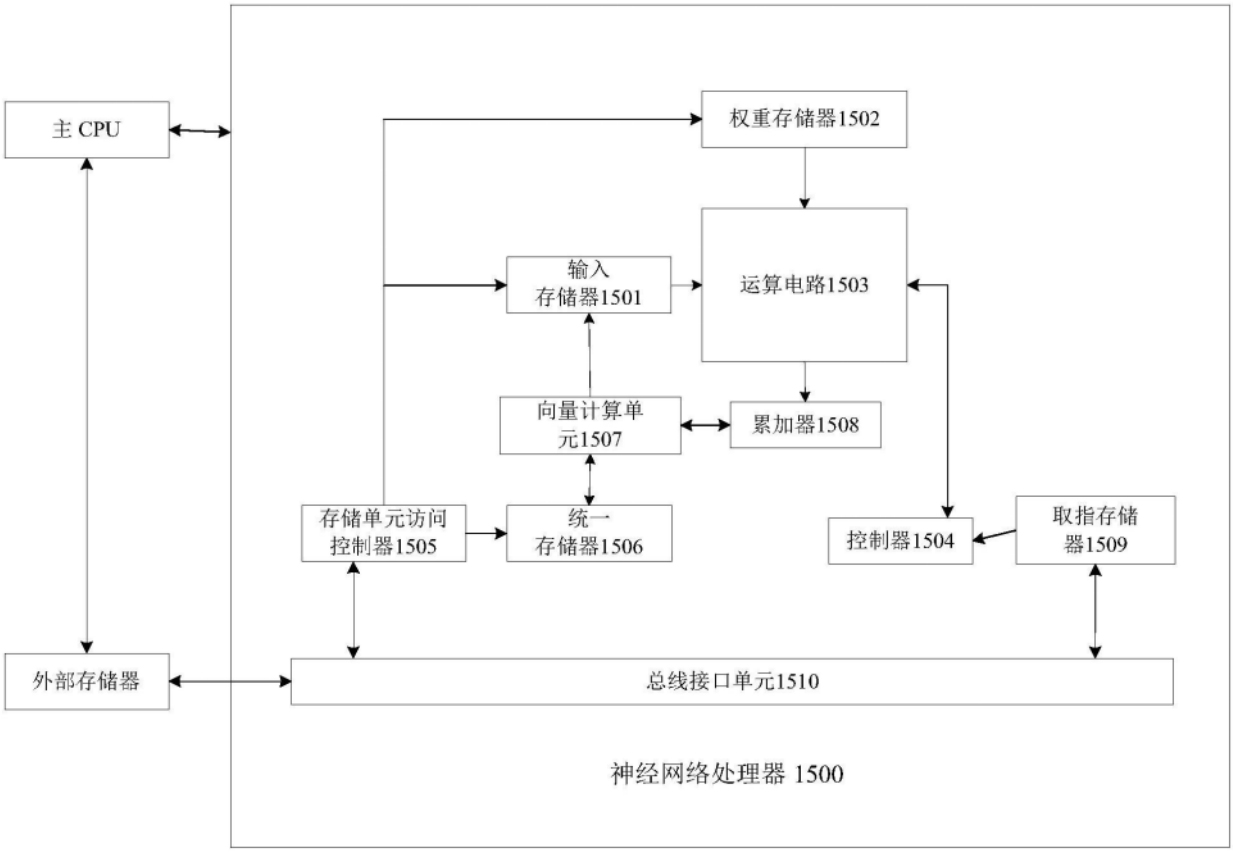

主 CPU
权重存储器1502
输入
存储器1501
运算电路1503
向量计算单
元1507
累加器1508
存储单元访问
控制器1505
统一
存储器1506
控制器1504
取指存储
器1509
外部存储器
总线接口单元1510
神经网络处理器 1500

图15