(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115391588 A
(43)申请公布日 2022.11.25

(21)申请号 202211343932.7
(22)申请日 2022.10.31
(71)申请人 阿里巴巴(中国)有限公司
地址 311121 浙江省杭州市余杭区五常街道文一西路969号3幢5层554室
(72)发明人 汪诚愚 王小丹 黄俊
(74)专利代理机构 北京展翼知识产权代理事务所(特殊普通合伙) 11452
专利代理师 张阳
(51)Int.Cl.
G06F 16/583(2019.01)
G06F 16/58(2019.01)
G06F 40/279(2020.01)
G06K 9/62(2022.01)
G06V 10/74(2022.01)
G06V 10/778(2022.01)
G06V 10/774(2022.01)

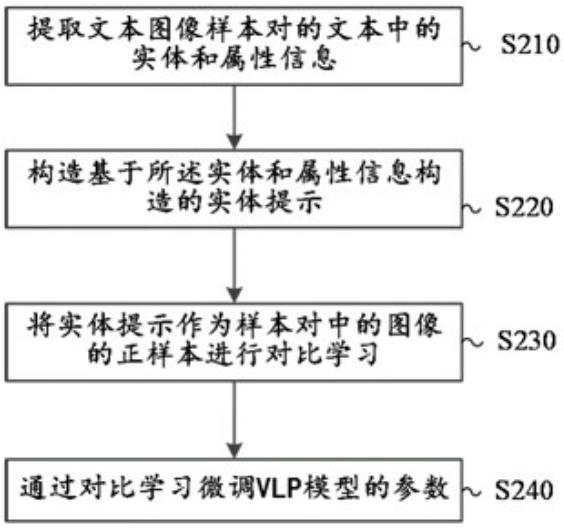
权利要求书2页 说明书14页 附图5页

(54)发明名称

视觉语言预训练模型的微调方法和图文检索方法

(57)摘要

公开了一种视觉语言预训练模型的微调方法和图文检索方法。所述微调方法包括:提取文本中的实体和属性信息,所述文本来自用于微调模型的图像文本样本对;构造基于所述实体和属性信息构造的实体提示;以及通过图像将所述实体提示作为正样本的对比学习微调所述VLP模型的参数,所述图像来自用于微调模型的所述图像文本样本对。本发明的微调方法能够在无需重训VLP模型的情况下实现图像-文本实体的细粒度对齐,以优化图文检索性能。具体地,可以在微调中通过对比学习和实体级掩模建模强调跨模态的实体对齐,并可以通过外部知识的引入进一步提升性能。可以通过重排序策略进一步改善图文检索结果。



CN 115391588 A

1.一种视觉语言预训练VLP模型的微调方法,包括:

提取文本中的实体和可视属性信息,所述文本来自用于微调模型的图像文本样本对;

构造基于所述实体和可视属性信息构造的实体提示;

将所述实体提示作为图像的正样本进行对比学习,所述图像来自用于微调模型的所述图像文本样本对;以及

通过所述对比学习微调所述VLP模型的参数。

2.如权利要求1所述的方法,还包括:

基于第一损失函数微调所述VLP模型的参数,其中,所述第一损失函数包括:

表征同一训练批次中文本嵌入向量与对应图像的嵌入向量的相似度与同一训练批次中其他文本的嵌入向量与该图像的嵌入向量的相似度差异的损失函数;以及

表征同一训练批次中图像嵌入向量与对应文本的嵌入向量的相似度与同一训练批次中其他图像的嵌入向量与该文本的嵌入向量的相似度差异的损失函数。

3.如权利要求2所述的方法,其中,通过所述对比学习微调所述VLP模型的参数包括:

基于第二损失函数微调所述VLP模型的参数,其中,所述第二损失函数包括:

表征所述实体提示的嵌入向量与所述图像的嵌入向量的相似度与同一训练批次中其他文本的嵌入向量与所述图像的嵌入向量的相似度差异的损失函数。

4.如权利要求3所述的方法,其中,所述第二损失函数还包括:

表征所述图像嵌入向量与所述实体提示的嵌入向量的相似度与同一训练批次中其他图像的嵌入向量与所述实体提示的嵌入向量的相似度差异的损失函数。

5.如权利要求3所述的方法,其中,所述第一损失函数还包括:

表征带有掩码实体的所述实体提示的嵌入向量与所述图像的嵌入向量的相似度与不带有掩码实体的所述实体提示的嵌入向量与所述图像的嵌入向量的相似度差异的损失函数。

6.如权利要求1所述的方法,还包括:

识别同一训练批次的图像样本对中所有图像包含的实体并生成视觉对象标签集;

从外部对象-图像库中查找与每个视觉对象标签各自对应的关联图像;

为每个视觉对象标签构造标签文本;以及

基于针对所述标签文本和关联图像进行对比学习的第三损失函数,微调所述VLP模型的参数。

7.如权利要求6所述的方法,其中,所述第三损失函数包括:

用于优化每个标签文本对其关联图像的匹配的损失函数;

用于表征带有掩码实体的所述标签文本的嵌入向量与其关联图像的嵌入向量的相似度与不带有掩码实体的所述标签文本的嵌入向量与其关联图像的嵌入向量的相似度差异的损失函数。

8.如权利要求1所述的方法,还包括:

识别所述图像中的实体和所述文本中的实体;

基于随机掩码所述图像或文本中的实体构造第四损失函数;以及

基于所述第四损失函数微调所述VLP模型的参数,其中,所述第四损失函数表征随机掩码前后的图像或文本的嵌入表征与原始文本或图像的嵌入表征的相似性差异。

9.一种图文检索方法,用于在输入文本时检索图像或是在输入图像时检索文本,所述方法包括:

获取用户输入的文本或图像信息;

将所述文本或图像信息送入根据如权利要求1-8中任一项所述的方法获取的VLP模型;

所述VLP模型基于所述文本或图像信息推理出多个图像候选或多个文本候选;以及

向所述用户提供所述多个图像候选中的一个或多个或所述多个文本候选中的一个或多个。

10.如权利要求9所述的方法,还包括:

基于推理出的多个图像候选或文本候选进行反向检索;以及

基于上述反向检索的结果,确定向用户提供的图像候选或文本候选的排序。

11.一种图文检索方法,用于在输入文本时检索图像或是在输入图像时检索文本,所述方法包括:

获取用户输入的文本或图像信息;

将所述文本或图像信息送入VLP模型;

所述VLP模型基于所述文本或图像信息推理出的多个图像候选或文本候选;

对推理出的多个图像候选或文本候选进行重新排序;以及

向所述用户提供经重新排序的所述多个所述图像候选或所述文本候选,

其中,对推理出的多个图像候选或文本候选进行重新排序包括:

在用户输入文本时,提取所述文本中的实体和属性信息;

构造由属性和实体构成的实体提示;

通过所述多个图像候选与所述实体提示的相似性对所述多个图像候选进行重排序;或者

在用户输入图像时,提取多个文本候选中的实体和属性信息;

构造由属性和实体构成的实体提示;

通过所述图像与所述多个文本候选对应的实体提示的相似性对所述多个文本候选进行重排序。

12.一种计算设备,包括:

处理器;以及

存储器,其上存储有可执行代码,当所述可执行代码被所述处理器执行时,使所述处理器执行如权利要求1至11中任何一项所述的方法。

13.一种非暂时性机器可读存储介质,其上存储有可执行代码,当所述可执行代码被电子设备的处理器执行时,使所述处理器执行如权利要求1至11中任何一项所述的方法。

视觉语言预训练模型的微调方法和图文检索方法

技术领域

[0001] 本公开涉及深度学习领域,尤其涉及一种视觉语言预训练模型的微调方法和图文检索方法。

背景技术

[0002] 图像-文本检索包括根据文本检索图像,以及根据图像检索文本,是一项极具挑战性的跨模态任务。基于海量图像-文本对预训练得到的视觉语言预训练(Vision-Language Pre-training,VLP)模型大幅提升了基于大量图像-文本对进行图像-文本检索性能。然而现有的基于预训练模型的方法仍然无法实现在跨模态数据上对齐实体的准确检索结果。基于实体标注的模型重新训练代价高昂且难以实现。

[0003] 为此,需要一种能够改善VLP性能的可行方案。

发明内容

[0004] 本公开要解决的一个技术问题是提供一种视觉语言预训练模型的微调方法和图文检索方法。该方法能够在无需重新训练VLP模型的情况下实现图像-文本实体的细粒度对齐,以优化图文检索性能。具体地,可以在微调中通过对比学习和实体级掩模建模强调跨模式的实体对齐,并可以通过外部知识的引入进一步提升性能。可以通过重排序策略进一步改善图文检索结果。

[0005] 根据本公开的第一个方面,提供了一种视觉语言预训练(VLP)模型的微调方法,包括:提取文本中的实体和可视属性信息,所述文本来自用于微调模型的图像文本样本对;构造基于所述实体和可是属性信息构造的实体提示;将所述实体提示作为图像的正样本进行对比学习;通过所述对比学习微调所述VLP模型的参数,其中,所述图像来自用于微调模型的所述图像文本样本对。

[0006] 可选地,所述方法还包括:基于第一损失函数微调所述VLP模型的参数,其中,所述第一损失函数包括:表征同一训练批次中文本嵌入向量与对应图像的嵌入向量的相似度与同一训练批次中其他文本的嵌入向量与该图像的嵌入向量的相似度差异的损失函数;以及表征同一训练批次中图像嵌入向量与对应文本的嵌入向量的相似度与同一训练批次中其他图像的嵌入向量与该文本的嵌入向量的相似度差异的损失函数。

[0007] 可选地,通过图像将所述实体提示作为正样本的对比学习微调所述VLP模型的参数包括:

基于第二损失函数微调所述VLP模型的参数,其中,所述第二损失函数包括:

表征所述实体提示的嵌入向量与所述图像的嵌入向量的相似度与同一训练批次中其他文本的嵌入向量与所述图像的嵌入向量的相似度差异的损失函数。

[0008] 可选地,所述第二损失函数还包括:表征所述图像嵌入向量与所述实体提示的嵌入向量的相似度与同一训练批次中其他图像的嵌入向量与所述实体提示的嵌入向量的相似度差异的损失函数。

[0009] 可选地，所述第一损失函数还包括：表征带有掩码实体的所述实体提示的嵌入向量与所述图像的嵌入向量的相似度与不带有掩码实体的所述实体提示的嵌入向量与所述图像的嵌入向量的相似度差异的损失函数。

[0010] 可选地，所述方法还包括：识别同一训练批次的图像样本对中所有图像包含的实体并生成视觉对象标签集；从外部对象-图像库中查找与每个视觉对象标签各自对应的关联图像；为每个视觉对象标签构造标签文本；以及基于针对所述标签文本和关联图像进行对比学习的第三损失函数，微调所述VLP模型的参数。

[0011] 可选地，所述第三损失函数包括：用于优化每个标签文本对其关联图像的匹配的损失函数；以及用于表征带有掩码实体的所述标签文本的嵌入向量与其关联图像的嵌入向量的相似度与不带有掩码实体的所述标签文本的嵌入向量与其关联图像的嵌入向量的相似度差异的损失函数。

[0012] 可选地，所述方法还包括：识别所述图像中的实体和所述文本中的实体；基于随机掩码所述图像或文本中的实体构造第四损失函数；以及基于所述第四损失函数微调所述VLP模型的参数，其中，所述第四损失函数表征随机掩码前后的图像或文本的嵌入表征与原始文本或图像的嵌入表征的相似性差异。

[0013] 根据本公开的第二个方面，提供了一种图文检索方法，用于在输入文本时检索图像或是在输入图像时检索文本，所述方法包括：获取用户输入的文本或图像信息；将所述文本或图像信息送入第一方面所述的方法获取的VLP模型；所述VLP模型基于所述文本或图像信息推理出多个图像候选或多个文本候选；以及向所述用户提供所述多个图像候选中的一个或多个或所述多个文本候选中的一个或多个。

[0014] 可选地，所述方法还包括：基于推理出的多个图像候选或文本候选进行反向检索；以及基于上述反向检索的结果，确定向用户提供的图像候选或文本候选的排序。

[0015] 根据本公开的第三个方面，提供了一种图文检索方法，用于在输入文本时检索图像或是在输入图像时检索文本，所述方法包括：获取用户输入的文本或图像信息；将所述文本或图像信息送入VLP模型；所述VLP模型基于所述文本或图像信息推理出的多个图像候选或文本候选；对推理出的多个图像候选或文本候选进行重新排序；以及向所述用户提供经重新排序的所述多个所述图像候选或所述文本候选，其中，对推理出的多个图像候选或文本候选进行重新排序包括：在用户输入文本时，提取所述文本中的实体和属性信息；构造由属性和实体构成的实体提示；通过所述多个图像候选与所述实体提示的相似性对所述所述多个图像候选进行重排序；或者在用户输入图像时，提取多个文本候选中的实体和属性信息；构造由属性和实体构成的实体提示；通过所述图像与所述多个文本候选对应的实体提示的相似性对所述所述多个文本候选进行重排序。

[0016] 根据本公开的第四个方面，提供了一种计算设备，包括：处理器；以及存储器，其上存储有可执行代码，当可执行代码被处理器执行时，使处理器执行如上述第一或第二或第三方面所述的方法。

[0017] 根据本公开的第四个方面，提供了一种非暂时性机器可读存储介质，其上存储有可执行代码，当可执行代码被电子设备的处理器执行时，使处理器执行如上述第一或第二或第三方面所述的方法。

[0018] 由此，本发明的微调方法能够在无需重训VLP模型的情况下实现图像-文本实体的

细粒度对齐，以优化图文检索性能。具体地，可以在微调中通过对比学习和实体级掩模建模强调跨模式的实体对齐，并可以通过外部知识的引入进一步提升性能。可以通过重排序策略进一步改善图文检索结果。

附图说明

[0019] 通过结合附图对本公开示例性实施方式进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施方式中，相同的参考标号通常代表相同部件。

[0020] 图1示出了基于现有VLP的图像文本检索模型发生错误预测的例子。

[0021] 图2示出了根据本发明一个实施例的VLP模型微调方法的示意性流程图。

[0022] 图3示出了根据本发明一个实施例的对比学习框架的流程示意图。

[0023] 图4A-图4C示出了根据本发明一个实施例进行VLP模型微调的一个具体例子。

[0024] 图5示出了本发明的图文检索系统的一个例子。

[0025] 图6示出了根据本发明一个实施例的图文检索方法的示意性流程图。

[0026] 图7示出了根据本发明一实施例可用于实现上述VLP微调方法的计算设备的结构示意图。

具体实施方式

[0027] 下面将参照附图更详细地描述本公开的优选实施方式。虽然附图中显示了本公开的优选实施方式，然而应该理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了使本公开更加透彻和完整，并且能够将本公开的范围完整地传达给本领域的技术人员。

[0028] 图文检索性能提升的关键在于如何对图像和文本数据进行表征学习，之后基于其表示学习跨模态相似性。传统的图文匹配方法由于没有足够的训练数据，无法打破不同模态表示学习之间的障碍。基于海量的图像-文本对(性能优秀的VLP模型训练往往需要上亿规模的图像-文本对)，VLP模型通过各种自监督的预训练任务更充分的学习到了跨模态的关联信息，从而大大缓解了传统图像-文本检索方法的缺陷，在零样本或微调场景下提升了许多跨模态任务的性能。

[0029] 然而，现有的VLP模型无法实现进行细粒度交互的模态语义匹配，基于VLP的图像文本检索模型仍然会产生错误的检索结果，在跨模态的查询数据和检索数据之间无法对齐实体信息。图1示出了基于现有VLP的图像文本检索模型发生错误预测的例子。

[0030] 在图文检索中，可以输入文本作为查询内容以进行图像检索，也可以输入图像作为查询内容以进行文本检索。在VLP模型执行下游图文检索任务时，希望在输入特定文本时，模型能够准确预测出在预训练阶段该特定文本对应的图像文本样本对中的图像；同时，希望在输入特定图像时，模型能够准确预测出在预训练时该特定图像对应的图像文本对中的文本。

[0031] 如图1所示，可以认为左侧“查询”列和中间“真实数据”列中，每一行包括的图像-文本对，是用于进行VLP模型预训练时所使用的样本对。在VLP模型完成预训练之后，期望模型能够在输入文本时正确预测出对应的图像，或是在输入图像时正确预测出对应的文本。

但图1所示的三个例子中，都发生了由于实体没有对齐而导致的错误预测。

[0032] 在第一个例子中，文本查询中的“菠萝”没有出现在预测图像中。同样，在第二个例子中，模型只关注“蔬菜”和“盘子”的匹配，而忽略了查询中的另一个重要实体“紫菜包饭”。另外，对盘子的数量也有误判。在第三个例子中，预测文本不包含在查询图像中可以明显观察到的“苹果”和“蛋糕”。

[0033] 为了改善VLP的图文检索性能，最近的努力集中于基于训练样本实体检测(例如，图1所示的图像中的框所框出的图像实体和文本中灰标的文本实体)的模型重训。然而，性能优秀的VLP模型训练往往需要上亿规模的图像-文本对，对上亿个图像-文本对进行实体标注的成本则更高，并且作为训练数据的这些图像-文本对往往也难以获得，因此使得现有技术中基于重训的方法成本极高且难以实现。

[0034] 在此，需要解释的是，在深度学习领域，如果模型在训练和推理时的任务不一致，则将模型的训练(调参)过程称为预训练过程。在本发明中，在VLP模型训练时，使用文本图像样本对作为输入；而在模型训练后执行图文检索任务时，使用文本或图像作为输出，相应检索出的图像或文本作为输出。因此，由于训练和推理时的任务不一致，因此VLP模型的训练过程属于深度学习领域中的“预训练”。

[0035] 如前所述，由于VLP模型的预训练成本极高且难以实行，本发明提出一种在无需重训VLP模型的情况下，仅仅通过小样本下的微调(fine-tune，也可称为“调优”)，甚至在零样本(zero-shot)场景下通过重排序而实现图像-文本实体的细粒度对齐，由此优化图文检索性能。具体地，可以在微调中通过对比学习和实体级掩模建模强调跨模式的实体对齐，并可以通过外部知识的引入进一步提升性能。可以通过重排序策略进一步改善图文检索结果。

[0036] 图2示出了根据本发明一个实施例的VLP模型微调方法的示意性流程图。该方法通过在微调阶段强调文本中的实体和实体的属性信息，增强VLP模型的实体对齐能力。应该理解的是，该方法是在VLP模型的预训练完成之后，对VLP模型的参数进行微调的阶段进行的。本发明针对VLP模型的微调方法以及如下将描述的重排序方法对英文数据集和中文数据集都有效。这也是图1和图4A-图4C的例子中同时给出中文和英文文本的原因。

[0037] 在步骤S110，提取文本中的实体和可是属性信息，所述文本来自用于微调模型的图像文本样本对。如图1所示，可以将左边和中间两列看作是三个图像文本对。在图像中，实体被以矩形框框出，在文本中，实体则以灰度显示。在第一个例子中，如果模型能够对文本输入中的“菠萝”进行识别和图像实体检索，则不会预测出第一行右侧不包括“菠萝”的错误图像。进一步地，图1的文本还用下划线标出了针对实体的可视属性信息。“可视属性信息”指的是用于描述实体的词，并且这些词是视觉上可观察的(即，“可视的”，这些信息是在样本对的图像中能够被反映出的视觉特征)。可视属性信息通常可以包括数量信息和颜色信息(实线下划线对应于数量信息，虚线下划线对应于颜色虚线)。如果模型能够对实体的属性特征做出正确判定，则更有可能进行正确的预测。

[0038] 在步骤S120，构造基于所述实体和可视属性信息构造的实体提示。例如，在图1对应的第三个例子中，可以将提取的颜色属性“白”和实体“盘子”构造成短语或是句子形式的实体提示(prompt)，例如“白盘子”，可以将提取的数量属性“一个”和实体“盘子”构造成另一个实体提示“一个盘子”，并且可以合并如上的颜色和数量属性构造实体提示“一个白盘子”。

[0039] 随后，在步骤S130，可以将实体提示作为图像的正样本进行对比学习，并在步骤S140，通过所述对比学习微调所述VLP模型的参数。

[0040] 对比学习是自监督学习的一种，不依赖标注数据，从无标注图像中自己学习知识。对比学习的指导原则是：通过自动构造相似的实例和不相似的实例，使得相似的实例在投影空间中接近，而不相似的实例在投影空间中距离推远的模型。为此，可以以同一训练批次为界，使得图像将同一图像文本对中文本构造的实体提示作为正样本，其他图像文本对中的文本作为负样本来进行对比学习，使得图像的嵌入向量与实体提示的嵌入向量作为相似的实例在嵌入向量空间中越来越接近，图像的嵌入向量与其他文本作为不相似的实例在嵌入向量空间中逐渐推远。

[0041] 虽然如步骤S130所述，可以将实体提示作为图像的正样本，但在更为确切的表述中，图像的正样本可以对应于图像文本对中对应的文本。实体提示则对应于基于作为正样本的文本中提取的属性和实体构造的附加正样本。以图1左下侧“一个白盘子里放着一些苹果、橘子和蛋糕”的文本和图像对为例，如果进行本发明的对比学习，则左下角图像的正样本是样本对中的对应文本“一个白盘子里放着一些苹果、橘子和蛋糕”，附加正样本则可以是通过提取的实体和描述实体的可视属性信息构造的实体提示“白盘子”、“一个盘子”和“一个白盘子”(也可以是“一些苹果”等等文本中包含的其他实体构造的实体提示)；相应地，负样本可以是同一批次的其他图像样本对中所包含的文本。

[0042] 在一个实施例中，虽然本发明的微调方法用于强调图像和文本中实体的对齐，但仍然需要确保图像和文本级别的相似性。为此，本发明的VLP模型微调方法还包括：基于第一损失函数微调所述VLP模型的参数。该第一损失函数是用于加强图像嵌入表示和文本嵌入表示之间的对应性而设置的，并且同样可以基于对比学习进行构造，即ITC(Image-Text Contrastive Loss)，此时，第一损失函数可以包括：表征同一训练批次中文本嵌入向量与对应图像的嵌入向量的相似度与同一训练批次中其他文本的嵌入向量与该图像的嵌入向量的相似度差异的损失函数；以及表征同一训练批次中图像嵌入向量与对应文本的嵌入向量的相似度与同一训练批次中其他图像的嵌入向量与该文本的嵌入向量的相似度差异的损失函数。由此，通过文本到图像的损失和图像到文本的损失来求取第一损失函数，例如如下将详述的$Loss_{CLIP}$，第一损失函数的引入能够保证在进行文本实体到图像的对齐时，保持图像和文本级别的对应相比于如上用于图像和文本级别对齐的第一损失函数，如上结合步骤S210~S240描述的将表征所述图像将实体提示作为正样本的对比学习的损失函数可以作为第二损失函数，例如如下将详述的$Loss_{VEA}$。

[0043] 在一个优选实施例中，可以进一步优化实体级别的对齐。在如上基于第二损失函数实现文本实体到图像的对齐的同时，还可以通过第三损失函数来实现视觉实体到图像的对齐。在本发明中，优选借助外部的对象-图像库来实现这一对齐。为此，本发明的微调方法可以包括：识别同一训练批次的图像样本对中所有图像包含的实体并生成视觉对象标签集；从外部对象-图像库中查找与每个视觉对象标签各自对应的关联图像；为每个视觉对象标签构造标签文本；以及基于针对所述标签文本和关联图像进行对比学习的第三损失函数，微调所述VLP模型的参数。该第三损失函数用于使模型从视觉对象标签与关联图像的对齐中习得视觉对象的特征。该第三损失函数可以对应于如下将详述的$Loss_{TEA}$。

[0044] 作为替换或者补充，还可以随机掩码图像和文本中的实体，并进行对比学习，通过

第四损失函数让模型对跨模态对齐实体的缺失更加敏感。此时，本发明的VLP模型微调方法还可以包括：识别所述图像中的实体和所述文本中的实体；基于随机掩码所述图像或文本中的实体构造第四损失函数；以及基于所述第四损失函数(例如，如下将详述的$Loss_{TIA}$)微调所述VLP模型的参数，其中，所述第四损失函数表征随机掩码前后的图像或文本的嵌入表征与原始文本或图像的嵌入表征的相似性差异。

[0045] CLIP(ContrastiveLanguage-ImagePre-Training，对比语言-图像预训练)模型可以进行跨模态的图文对比学习。本发明的如上第二损失函数即可由$Loss_{CLIP}$实现。但是仅仅基于图文的全局表征计算相似性在捕捉实体级对齐关系上存在不足，例如会出现图1所示的预测错误，因此本发明在微调阶段，基于图文对比学习框架进行跨模态实体级信息的对齐，并且可以在重排序阶段进行优化，以实体相似度为引导，指导细粒度实体级的对齐，并通过跨模态信息反向检索优化排序结果。这种对比学习框架优化了跨模态图文检索的微调结果，使得满足在跨模态实体等细粒度知识对应关系上更好对齐的正确检索结果排名更靠前。轻量化的微调和重排序框架实现使得本发明更有实用性，在推理和微调场景下有更好的性能。

[0046] 图3示出了根据本发明一个实施例的对比学习框架的流程示意图。首先，从文本中识别文本实体，从图像中识别视觉实体，然后通过预训练的 VLP 模型将其与原始文本和图像一起编码。如图所示，文本被编码为$\{t_1, t_2, \ldots, t_N\}$，图像被编码为$\{v_1, v_2, \ldots, v_N\}$，从文本中识别的文本实体被编码为$\{te_1, te_2, \ldots, te_P\}$，从图像中识别的图像实体被编码为$\{vo_1, vo_2, \ldots, vo_K\}$。即，在N个图像本对中，共识别出P个文本实体(Text Entity，*te*)，K个视觉实体(Visual Object，*vo*，也可称为视觉对象Visual Entity)。

[0047] 之后，进入微调阶段，其中设计了三个不同的模块来学习跨模态实体之间的对齐：

·视觉实体-图像对齐(Visual Entity-Image Alignment，VEA) 从外部的跨模态图文数据库(例如，Visual Genome)中获得视觉实体图像对，用于通过对比学习和图像区域掩码建模来学习视觉实体与其对应图像之间的对齐。这对应于如上针对第三损失函数的操作。

[0048] ·文本实体-图像对齐(Textual Entity-Image Alignment，TEA) 构造一个仅包含文本实体及其可视化属性(如颜色和数字)的句子，然后通过对比学习和文本实体掩码建模学习句子与其对应图像之间的对齐。这对应于如上针对第一损失函数的操作。

[0049] ·文本-图像实体对齐 (Text-Image Entity Alignment，TIA) 通过随机掩码图像或文本中的实体来强调跨模态实体对齐的重要性，以使模型对跨模态对齐实体的缺失更加敏感。这对应于如上针对第四损失函数的操作。

[0050] 本发明还可以通过在模型给出前*k*个(top- *k*)候选结果后进行重排序来进行优化模型性能。于是随后进入重排序阶段，期望通过以下设计的重新排序策略来细化前*k*个(top- *k*)候选的排序结果：

·文本-图像双向重排序(Text-Image Bidirectional Re-ranking，TBR) 将top-*k*(例如*k* =10)检索结果进行反向图像-文本检索，在考虑反向检索结果的情况下进行重新排序。

[0051] ·文本实体指导的重排序(Textual Entity-Guided Re-ranking，EGR) 则专门为

零样本(zero-shot)场景设计,针对top- k检索结果计算图像和文本中实体之间的相似度,然后在考虑相似度的情况下进行微调排名结果。该EGR模块执行的操作与TEA模块类似,同样是根据文本中的实体提示与图像的相似度来进行操作。

[0052] 本发明的基于跨模态实体对齐的方法计算了全局相似度和实体相似度,之后进行融合。实体相似度是指基于VEA、TEA和TIA三个模块的跨模态实体比对,强调图像和文本之间的相似性。在一个具体实施例中,VEA将从外部多模态知识库中获得的对应图像作为实体标签的输入,通过VEM和MVC两个子模块输出视觉图像与其标签之间的相似性。TEA由TEE和MEC两个子模块组成,接收以文本实体和图像为输入的文本,输出文本实体与图像之间的相似度。TIA也接受带有实体的原始图像和文本,但学习计算图像和文本实体之间的相似距离。

[0053] 如下将结合图4A-图4C所示的实例描述基于本发明进行微调的一个具体实现。图4A-图4C示出了根据本发明一个实施例进行VLP模型微调的一个具体例子。由于单张图显示面积有限,因此为了显示清楚,在此将VLP模型微调的一个具体例子拆分显示在图4A-图4C三张图上。应该理解的是,图4A下部的文本编码器和视觉编码器,以及图4B下部的视觉实体编码器和文本实体编码器都是连接至同一个实体对齐模块,并由上部相同的ITC、TEA、TIA和VEA模块进行处理的。因此,图4A和图4B可以合并作为VLP模型微调的一个流程示意图。图4C则详细示出了ITC模块、TIA模块、以及TEA模块包括的TEE和MEC子模块、VEA模块包括的VEM和MVC子模块具体接收的是哪些嵌入向量。

[0054] 本发明针对VLP模型的微调建立在图像-文本对比学习范式之上,该范式期望在嵌入向量(embedding)空间内缩短相关图像和文本之间的距离,并将那些不相关的图像和文本推远。

[0055] 微调的整体架构如图3所示,其中计算了全局相似度和实体相似度两者,再进行融合。全局相似度是通过直接计算图像和文本的嵌入之间的相似度来获得的(对应于图中的ITC模块),而实体相似度是基于 VEA、TEA 和 TIA 三个本发明提出的新颖模块来求取的,实体相似度强调图像和文本对之间相似度的跨模态实体对齐。具体地,VEA 输入实体标签和从外部多模态知识库 (MMKB) 获得的相应图像,并通过两个子模块 VEM 和 MVC 输出视觉图像与其标签之间的相似性。TEA由2个子模块TEE和MEC组成,接收带有文本实体的文本以及图像作为输入,输出文本实体与图像的相似度。TIA 也接受带有实体的原始图像和文本,但学习计算图像和文本实体之间的相似度距离。

[0056] 在此,将从图像中提取的视觉实体表示为V_{obj},将从文本中提取的文本实体表示为T_{ent}。经过VLP模型编码后,视觉实体的表示为$v_{io_i} = g(x_i;\gamma^a) \in \mathbb{R}^{d_i}$,文本实体的表示为$t_{te_j} = g(x_j;\gamma^\beta) \in \mathbb{R}^{d_t}$。在所有微调模块的相同对比学习范式下,在一个训练批次中使用图像集V和文本集T对b个图像-文本对$\{i_k^V, t_k^T\}$进行采样。对于被选样本中的图像$i^V \in V$,文本$t^T \in T$被视为其正对,而其他文本则被视为批次内的负样本。图像和文本的对比损失可以表示为:

$$L = \frac{1}{2}\sum_{k=1}^{b}(L_k^V + L_k^T) \qquad (1)$$

其中L_k^V和L_k^T分别指的是图像到文本和文本到图像的对比损失。以图像到文本为例，损失函数可以表示为式(2)，其中$s_{j,k}^V$对应于第k幅图像到第j个文本。文本到图像部分的对比损失与图像到文本的对比损失对称。

$$L_k^V(i_k^V,\{t_j^T\}_{j=1}^b) = -log(exp(s_{k,k}^V))/(\sum_j exp(s_{j,k}^V)) \qquad (2)$$

[0057] 式(1)可以看作是ITC模块对$LOSS_{CLIP}$(对应于第一损失函数)的求取。即，第一损失函数包括表征同一训练批次中文本嵌入向量与对应图像的嵌入向量的相似度与同一训练批次中其他文本的嵌入向量与该图像的嵌入向量的相似度差异的损失函数；以及表征同一训练批次中图像嵌入向量与对应文本的嵌入向量的相似度与同一训练批次中其他图像的嵌入向量与该文本的嵌入向量的相似度差异的损失函数。

[0058] 进一步地，遵循对比学习范式，从文本t_k^T中提取的实体$t_{te_k}^T$可以用作相应图像i_k^V的用于指示文本-图像实体级对齐的正样本，而文本t_k^T中未提及的实体被视为负样本。对于图像i_k^V中的视觉对象$i_{io_k}^V$的标签也类似，从图像i_k^V中提取的视觉对象$i_{io_k}^V$可以作为相应文本t_k^T的用于指示图像-文本实体级对齐的正样本，而未从图像中检测到的标签则是负的。在下文中，给出了为计算每个图像-文本对之间的实体相似度而设计的三个模块VEA、TEA 和 TIA 的更多技术细节。

[0059] 视觉实体-图像对齐(Visual Entity-Image Alignment,VEA)模块

与严重依赖对象检测模型用于细粒度交互的许多现有VLP模型不同，本发明简单地将检测到的标签用作媒介并重建一个对象-图像库(即，MMKB)用于视觉知识以与其视觉图像对齐。在一个实施例中，选择视觉基因组(Visual Genome,VG)并设计简单的启发式规则过滤图像来建立该对象-图像库。在微调过程中，为批次内N幅图像所带有的k个实体收集视觉标签集$VO = \{io_m\}_{m=1}^k$，并从基于本公开过滤得到的 MMKB中找出与实体的关联图像。在图4A和图4B所示的例子中，从图像文本对的图像中检测到包括对应于“water”(水)、“boat”(船)、“men”(男人)、“shirt”(T恤)和“bench”(长椅)的视觉实体，并由此从MMKB中查出对应的图像。

[0060] 在此，遵循图像-文本对比学习的范式，通过两个任务来学习每个视觉实体的实体-图像对齐。整体损失函数可以表示为方程$LOSS_{VEA} = \frac{1}{2}(LOSS_{VEM} + LOSS_{MVC})$，其中$LOSS_{VEM}$和$LOSS_{MVC}$是两个子模块。$LOSS_{VEA}$对应于如上所述的第三损失函数。第三损失函数包括用于优化每个标签文本对其关联图像的匹配的损失函数，即$LOSS_{VEM}$；还包括用于表征带有掩码实体的所述标签文本的嵌入向量与其关联图像的嵌入向量的相似度与不带有掩码实体的所述标签文本的嵌入向量与其关联图像的嵌入向量的相似度差异的损失函数，即$LOSS_{MVC}$。

[0061] 1)视觉实体匹配(Visual Entity Matching,VEM)。当前训练批次中检测对象vo_i的图像被视为来自 MMKB 的对象图像io_m的正样本。考虑到诸如“shephred dog” (牧羊犬)等

实体的短标签与预训练数据中的完整长句之间的不一致，可以使用统一的基于规则的方法为实体级文本样本构建与视觉侧图像对齐的提示。例如，可以使用提示“a photo contains {entity}”(照片包含{实体})。在此，优化视觉对象的标签文本to_m与其图像io_m的匹配，并制定与全局训练目标方程式(2)一致的损失函数。

[0062] 在式(3)中，$TO = \{to_m\}_{m=1}^{k}$指对象文本的集合，to_m是调用图像io_m对应的对象说明(caption)。

$$Loss_{VEM} = \frac{1}{2}\sum_{i=1} k(L_i^{VO} + L_i^{TO}) \qquad (3)$$

[0063] 视觉实体标签和实体图像的距离是通过来自视觉和文本实体编码器的[CLS]令牌(token)的嵌入向量计算的，在图4中表示为$T_{to_{cls}}$和$V_{io_{cls}}$。简单的框架使模型具将对象图像与其正确的标签对齐的能力。

[0064] 2) 掩码视觉对象一致性对齐(Masking Visual Object Consitency Alignment，MVC)。受 VLP 模型随机掩码图像的某些部分以进行掩码区域分类或掩码区域特征回归的预训练任务的启发，在此采用掩码策略(但以不同的方法)来学习视觉实体的表示。我们利用标签提示与原始图像以及标签提示与带有掩码实体的图像之间计算的相似度分数的差异，最小化式(4)中视觉对象一致性学习的边际排序损失。y在式中表示1，s_{io_k,to_k}用于表示原始图像和文本之间的相似性。带掩码实体区域的图像的视觉嵌入在图 3 中示出为$V_{wo/io_{cls}}$，则$s_{io_{k(wo/io)},to_k}$用于表示带掩码实体区域的图像和文本之间的相似性。$LOSS_{MVC}$期望原始图像和对象标签的得分更高，以更多地强调那些缺失的视觉实体。

$$Loss_{MVC} = \sum_{k=1} max(0, -y \cdot (s_{io_k,to_k} - s_{io_{k(wo/io)},to_k})) \qquad (4)$$

[0065] 由此，例如图4A中原始图像所包含的视觉对象的文本标签“bench”能够通过MMKB中的长椅图像学习到更多对应视觉对象的特征。并且注意到对应于该视觉对象的文本标签“bench”并不包含在与原始文本的描述中(这也意味着相比于简明的文本说明，图像总能包含更多的冗余信息，这也是视觉对象与图像对齐的意义。通过针对样本对的图像提取视觉对象，训练其视觉对象对应标签的嵌入表征能够包括更多的对应视觉对象信息，

文本实体-图像对齐(Textual Entity-Image Alignment，TEA)模块

由于相比于简明的文本说明，图像总能包含更多的冗余信息，因此在此重新考虑视觉和文本信息的不对称性，并特别注意文本中的实体级信息以与相应的图像对齐。$LOSS_{TEA}$对应于如上所述的第二损失函数。所述第二损失函数可以包括：表征所述实体提示的嵌入向量与所述图像的嵌入向量的相似度与同一训练批次中其他文本的嵌入向量与所述图像的嵌入向量的相似度差异的损失函数(对应于如下$LOSS_{TEE}$中的L_i^{TE})。在一个实施例中，第二损失函数还包括：表征所述图像嵌入向量与所述实体提示的嵌入向量的相似度与同一训练批次中其他图像的嵌入向量与所述实体提示的嵌入向量的相似度差异的损失函

数(对应于如下$LOSS_{TEE}$中的L_i^V)。进一步地,$LOSS_{TEA}$还可以包括表征带有掩码实体的所述实体提示的嵌入向量与所述图像的嵌入向量的相似度与不带有掩码实体的所述实体提示的嵌入向量与所述图像的嵌入向量的相似度差异的损失函数,对应于如下的$LOSS_{MEC}$。

[0066] 1)文本实体强调的对齐(Textual Entity-Image Alignment,TEE)。首先通过强调实体的令牌来强调说明中的实体级信息。给定一个图文对$\{i_m, t_m\}$,从文本t_m中提取p个多级实体信息,包括命名实体和属性(特别是颜色和数量信息,这些都是能够在图像中反映出的可视化属性),记为$TE = \{te_i\}_{i=1}^{p}$。如图4 B所示,我们提取“a white boat”(数字)、“a man”(数字)、“blue clothes”(颜色)等(对应于中文分别为“一艘白船”、“一个男人”、“蓝色衣服”)。然后将构造的实体提示作为图像i_m的附加正样本用于对比学习。提示标签的嵌入向量被表示为$T_{te_{cls}}^1, T_{te_{cls}}^2, \ldots, T_{te_{cls}}^p$,以计算与图像嵌入向量$V_{cls}$的相似度。对同一文本中的多个实体采用平均池化,以同时考虑所有实体的重要性,而不仅仅考虑与部分实体的对齐。与式(2)一致的损失函数表示为式(5)。具体来说,L_i^{TE}表示针对每个实体文本t_{em}的

$$L_k^V(\{i_j^V\}_{j=1}^b, te_k^{TE}) = -\frac{1}{b} log \frac{exp(s_{k,k}^T)}{\sum_j exp(s_{j,k}^{TE})}$$,P表示每个说明限制的实体数量。

$$Loss_{TEE} = \frac{1}{2P} \sum_{i=1} k(L_i^V + L_i^{TE}) \qquad (5)$$

[0067] 2)掩码实体一致性对齐(Mask Entity Consitency Alignment)。

[0068] 通过掩码文本实体令牌,进一步将图像与文本实体一致地对齐。在此,并不像大多数模型那样给出准确的词汇表并对实体进行分类,而是采用一种更轻量的方式来学习关于文本实体的统一跨模态表示。我们重新计算原始图像im与带有掩码实体的文本$t_{wo/te}$之间的相似度,期望图像与被破坏句子之间的相似度$s_{i,t_{wo/te}}$小于原始文本与图像之间的相似度$s_{i,t}$。带有掩码文本的文本嵌入在图 3 中表示为$T_{wo/te_{cls}}$。与 TEE 模块类似,在此可以采用平均池化。在一个批次的k个图像-文本对样本中,损失函数可以表示为式(6)。

$$Loss_{MEC} = \sum_{k=1} max(0, -y \cdot (s_{i_k,t_k} - s_{i_k,t_{k(wo/te)}})) \qquad (6)$$

[0069] 文本实体-图像对齐的目标统一优化为$LOSS_{TEA} = \frac{1}{2}(LOSS_{TEE} + LOSS_{MEC})$。

[0070] 由于单一模态中的信息应该与另一种互补模态相关联,因此在此增强了图像-文本对的文本实体以使其与图像中的视觉表示对齐,而不是为实体引入额外的知识。

[0071] 文本图像实体对齐(Textual-Image Entity Alignment)模块

为了进一步弥合模态之间的差距并补偿异构信息之间的无序词汇表造成的对齐缺陷,可以利用预训练的视觉基础模型作为锚点来针对检测到的文本实体$TE = \{te_i\}_{i=1}^{p}$识别图像中每个实体的区域。随后对图像中的基准实体进行掩码。$LOSS_{TIA}$对应于如上所述的第

四损失函数。

$$Loss_{TIA} = \sum_{k=1} max(0, -y \cdot (s_{i_k,t_k} - s_{i_{k(wo/ie)},t_k})) \qquad (7)$$

[0072] 在此，仍然最大化原始图像与带掩码区域$i_{wo/ie}$图像之间的差异，该带掩码区域图像的嵌入向量在图4B和图4C中表示为$V_{wo/ie_{cls}}$，在式(7)中的$s_{i_{k(wo/ie)}}$在$V_{wo/ie_{cls}}$和T_{cls}之间计算。在 TIA 中，我们只关注文本中实体和图像的一致性，因为在 VEA 中已经学习了视觉侧的实体-图像对齐。遵循上述训练目标，损失函数表示为式(7)。进一步地，可以共同优化VEA、TEA和TIA。每个图像和文本分别需要多次(例如，3 次)前向传播，无需引入额外的编码器或参数。整体训练目标如下式所示。

$$LOSS = LOSS_{CLIP} + \frac{1}{3}(LOSS_{VEA} + LOSS_{TEA} + LOSS_{TIA}) \qquad (8)$$

[0073] 本发明还可以实现为一种图文检索系统。图5示出了本发明的图文检索系统的一个例子。该系统文本查询信息获取模块，用于获取用户输入的文本或是图像信息；以及信息生成模块，设置如上所述的方法微调的VLP模型，用于基于用户输入的文字或是图像信息，输出匹配的图像或是文本信息。

[0074] 该图文检索系统可以实现一种图文检索方法。图6示出了根据本发明一个实施例的图文检索方法的示意性流程图。该方法用于在输入文本时检索图像或是在输入图像时检索文本。

[0075] 在步骤S610，获取用户输入的文本或图像信息。在步骤S620，将所述文本或图像信息送入根据如上所述的微调方法获取的VLP模型。在步骤S630，所述VLP模型基于所述文本或图像信息推理出的多个图像候选或文本候选。在所述S640，向所述用户提供所述多个所述图像候选或所述文本候选中的一个或多个。优选地，可以对多个个图像候选或文本候选进行重排序。此时，图6的方法还包括：基于推理出的多个图像候选或文本候选进行反向检索；以及基于上述反向检索的结果，确定向用户提供的图像候选或文本候选的排序。

[0076] 如上反向检索和重排序可由文本图像双向重排序(Text-Image Bidirectional Re-ranking，TBR)模块执行。丰富的视觉信息和简洁的文本知识之间的冗余不一致可能导致仅通过一种模态的部分信息做出错误的决策，对于没有细粒度交互的 VLP 模型来说尤是如此。因此，本发明提出了 TBR 策略来补偿不一致性，该策略通过反向检索将来自互补模态的互信息引入作为额外的监督信号。该方法仅依赖于跨模态样本本身。具体而言，可以将具有最高相似性的文本样本$\{t_{rank_1}, t_{rank_2}, \cdots, t_{rank_k}\}$视为图像$i_m$的互近邻(reciprocal neighbors)，并从候选池中反向检索与每个文本最相似的图像。在这里，使用排名位置而不是相似度得分。然后图像i_m的前k个候选图像用新计算的位置重新排序为$t_{rank_i} = (rank_{t_{rank_i}2i_m} + i)/2$以代表$t_i$。文本到图像的检索也是如此。这种简单但有效的自监督方式只是重新访问排名位置，不需要额外的数据，但在一定程度上保证了视觉和文本信息的对齐。

[0077] 重排序策略可以弥补细粒度交互的不足，避免了仅通过部分信息做出的错误判断。TBR 模块还应用于微调结果，用于图文一致性对齐的后处理。

[0078] 在一个实施例中，本发明的文本实体与图像对齐的原理还可以应用于零样本场景。为此，本发明还公开了一种图文检索方法，用于在输入文本时检索图像或是在输入图像时检索文本，所述包括：获取用户输入的文本或图像信息；将所述文本或图像信息送入VLP模型；所述VLP模型基于所述文本或图像信息推理出的多个图像候选或文本候选；对推理出的多个图像候选或文本候选进行重新排序；以及向所述用户提供经重新排序的所述多个所述图像候选或所述文本候选中的一个或多个，其中，对推理出的多个图像候选或文本候选进行重新排序包括：在用户输入文本时，提取所述文本中的实体和属性信息；构造由属性和实体构成的实体提示；通过所述多个图像候选与所述实体提示的相似性对所述所述多个图像候选进行重排序；或者在用户输入图像时，提取多个文本候选中的实体和属性信息；构造由属性和实体构成的实体提示；通过所述图像与所述多个文本候选对应的实体提示的相似性对所述所述多个文本候选进行重排序。

[0079] 零样本场景下的重排序策略对应于图3所示的实体指导的重排序(Textual Entity-Guided Re-ranking，EGR)模块。为了进一步提高具有细粒度实体级交互的 VLP 模型的性能，可以将TEA模块策略转换为用于重排序的实体对齐分数。按照相同的程序，一方面将提取的文本实体t_i转换为基于提示的说明，并计算针对图像i_m的文本实体对齐分数计算$Score_{TEE}=\sum_{i=1}^{k}s_{i_m,te_i}$，另一方面，将文本中的实体替换为 [MASK] 以获取文本实体一致性评分为$Score_{MEC}=\sum_{i=1}^{k}max(0,-y\cdot(s_{i_m,t_i}-s_{i_m,to_{i(wo/te)}})$。实体引导的重排序分数$Score_{EGR}$由$Score_{TEE}$和$Score_{MEC}$的组合计算得出。

[0080] EGR 仅将实体级别的对齐过程模拟为图像文本相似度得分，这与 VLP 模型更兼容。我们在验证集上调整$Score_{A11}$和$Score_{EGR}$的系数，并将它们应用到测试集上。排名的最终分数表示为$Score_{Final}=\alpha\cdot Score_{A11}+\beta\cdot Score_{EGR}$。图像和文本首先使用$Score_{A11}$进行预排序，用于选择$k$个候选者，然后使用$Score_{EGR}$对$k$个候选者进行重排序。

[0081] 图7示出了根据本发明一实施例可用于实现上述VLP微调方法的计算设备的结构示意图。

[0082] 处理器720可以是一个多核的处理器，也可以包含多个处理器。在一些实施例中，处理器720可以包含一个通用的主处理器以及一个或多个特殊的协处理器，例如图形处理器(GPU)、数字信号处理器(DSP)等等。在一些实施例中，处理器720可以使用定制的电路实现，例如特定用途集成电路(ASIC，Application Specific Integrated Circuit)或者现场可编程逻辑门阵列(FPGA，Field Programmable Gate Arrays)。

[0083] 存储器710可以包括各种类型的存储单元，例如系统内存、只读存储器(ROM)，和永久存储装置。其中，ROM可以存储处理器720或者计算机的其他模块需要的静态数据或者指令。永久存储装置可以是可读写的存储装置。永久存储装置可以是即使计算机断电后也不会失去存储的指令和数据的非易失性存储设备。在一些实施方式中，永久性存储装置采用大容量存储装置(例如磁或光盘、闪存)作为永久存储装置。另外一些实施方式中，永久性存储装置可以是可移除的存储设备(例如软盘、光驱)。系统内存可以是可读写存储设备或者易失性可读写存储设备，例如动态随机访问内存。系统内存可以存储一些或者所有处理器在运行时需要的指令和数据。此外，存储器710可以包括任意计算机可读存储媒介的组合，包括各种类型的半导体存储芯片(DRAM，SRAM，SDRAM，闪存，可编程只读存储器)，磁盘和/或

光盘也可以采用。在一些实施方式中，存储器710可以包括可读和/或写的可移除的存储设备，例如激光唱片(CD)、只读数字多功能光盘(例如DVD-ROM，双层DVD-ROM)、只读蓝光光盘、超密度光盘、闪存卡(例如SD卡、min SD卡、Micro-SD卡等等)、磁性软盘等等。计算机可读存储媒介不包含载波和通过无线或有线传输的瞬间电子信号。

[0084] 存储器710上存储有可执行代码，当可执行代码被处理器720处理时，可以使处理器720执行上文述及的VLP微调方法以及图文检索方法。

[0085] 上文中已经参考附图详细描述了根据本发明的VLP微调方法以及相应的图文检索方法。本发明基于图文实体的对比学习框架，对图像和文本的实体分别进行建模，同时借助外部知识库进行对齐。首先从外部知识库Visual Genome中获取视觉实体-图像对，然后通过对比学习和图像区域掩模建模来学习视觉实体与其对应图像之间的对齐。其次，构建只包含文本实体及其可视化属性(如颜色和数字)的句子，然后通过对比学习和文本实体掩码建模来学习实体及实体属性与其对应图像之间的对齐。通过随机掩码图像或文本中的实体，让模型对跨模态对齐实体的缺失更加敏感，以强调了跨模态实体对齐的重要性。在重排序步骤中，使用排名靠前的k个检索结果进行反向图像-文本检索，然后将其结果考虑到重新排序，并且特别对于没有微调步骤的零样本场景，还利用top-k检索结果来计算来自图像和文本的实体之间的相似度，并在重排序时予以考虑。在多个中文和英文数据集和多个VLP模型的广泛实验表明了方法的有效性，可以取得比在预训练阶段采用细粒度复杂交互的模型具有更好的效果。

[0086] 此外，根据本发明的方法还可以实现为一种计算机程序或计算机程序产品，该计算机程序或计算机程序产品包括用于执行本发明的上述方法中限定的上述各步骤的计算机程序代码指令。

[0087] 或者，本发明还可以实施为一种非暂时性机器可读存储介质(或计算机可读存储介质、或机器可读存储介质)，其上存储有可执行代码(或计算机程序、或计算机指令代码)，当所述可执行代码(或计算机程序、或计算机指令代码)被电子设备(或计算设备、服务器等)的处理器执行时，使所述处理器执行根据本发明的上述方法的各个步骤。

[0088] 本领域技术人员还将明白的是，结合这里的公开所描述的各种示例性逻辑块、模块、电路和算法步骤可以被实现为电子硬件、计算机软件或两者的组合。

[0089] 附图中的流程图和框图显示了根据本发明的多个实施例的系统和方法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、程序段或代码的一部分，所述模块、程序段或代码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。也应当注意，在有些作为替换的实现中，方框中所标记的功能也可以以不同于附图中所标记的顺序发生。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。也要注意的是，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操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。

[0090] 以上已经描述了本发明的各实施例，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的改进，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。


查询
正确数据
错误预测
中文：玻璃盘中摆了菠萝、香蕉和橙子。
英文：Pineapples, bananas and oranges are displayed in a glass plate.
中文：盘子上放着一些紫菜包饭，旁边的碟子上是一些蔬菜。
英文：There is some kimbap on the plate and some vegetables on another plate next to it.
中文：一个白盘子里放着一些苹果，橘子和蛋糕。
英文：There are some apples, oranges and cake on a white plate.
中文：一些柑橘放在一个白色的盘子里.
英文：There are some oranges on a white plate.

图1

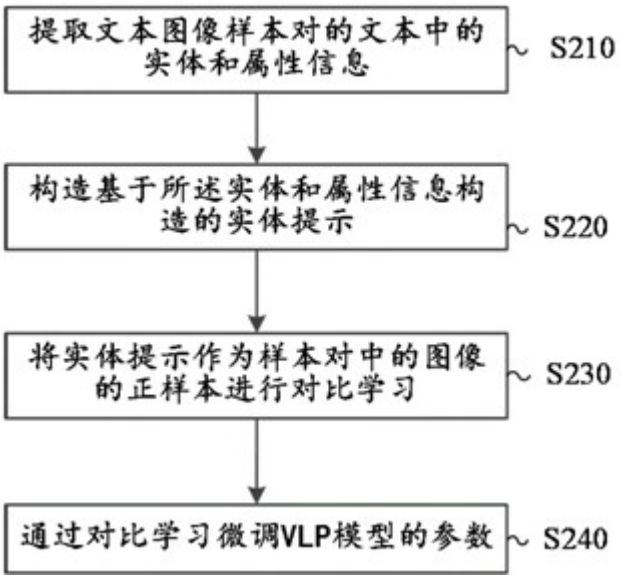

提取文本图像样本对的文本中的实体和属性信息
S210
构造基于所述实体和属性信息构造的实体提示
S220
将实体提示作为样本对中的图像的正样本进行对比学习
S230
通过对比学习微调VLP模型的参数
S240

图2

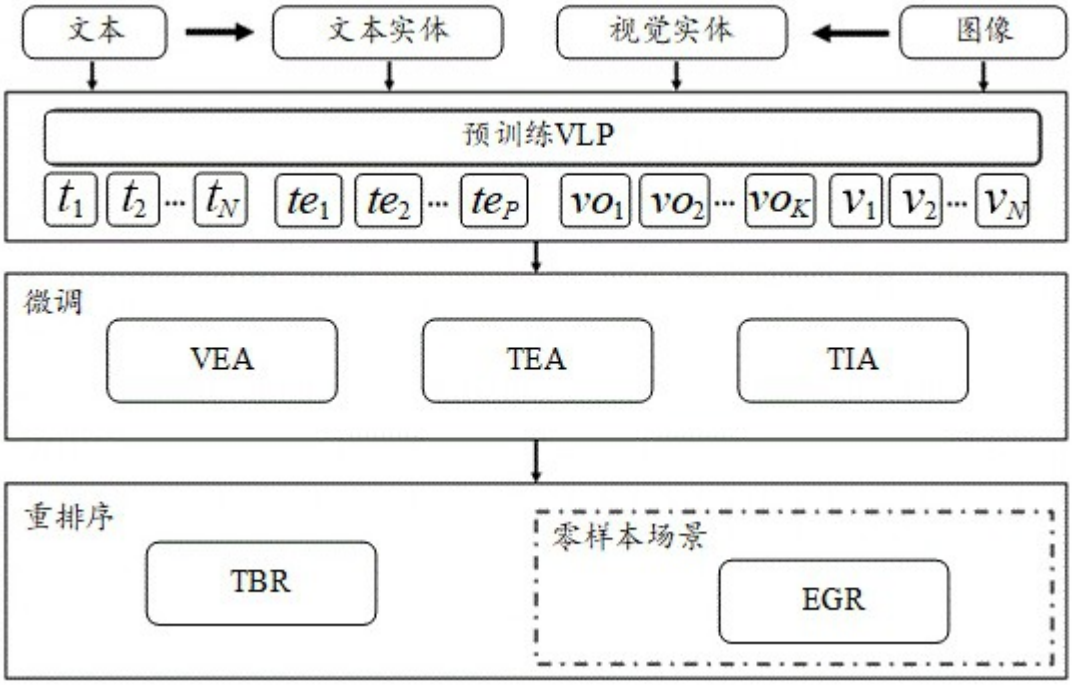

文本
文本实体
视觉实体
图像
预训练VLP
t_1 t_2 ··· t_N te_1 te_2 ··· te_P vo_1 vo_2 ··· vo_K v_1 v_2 ··· v_N
微调
VEA
TEA
TIA
重排序
TBR
零样本场景
EGR

图3

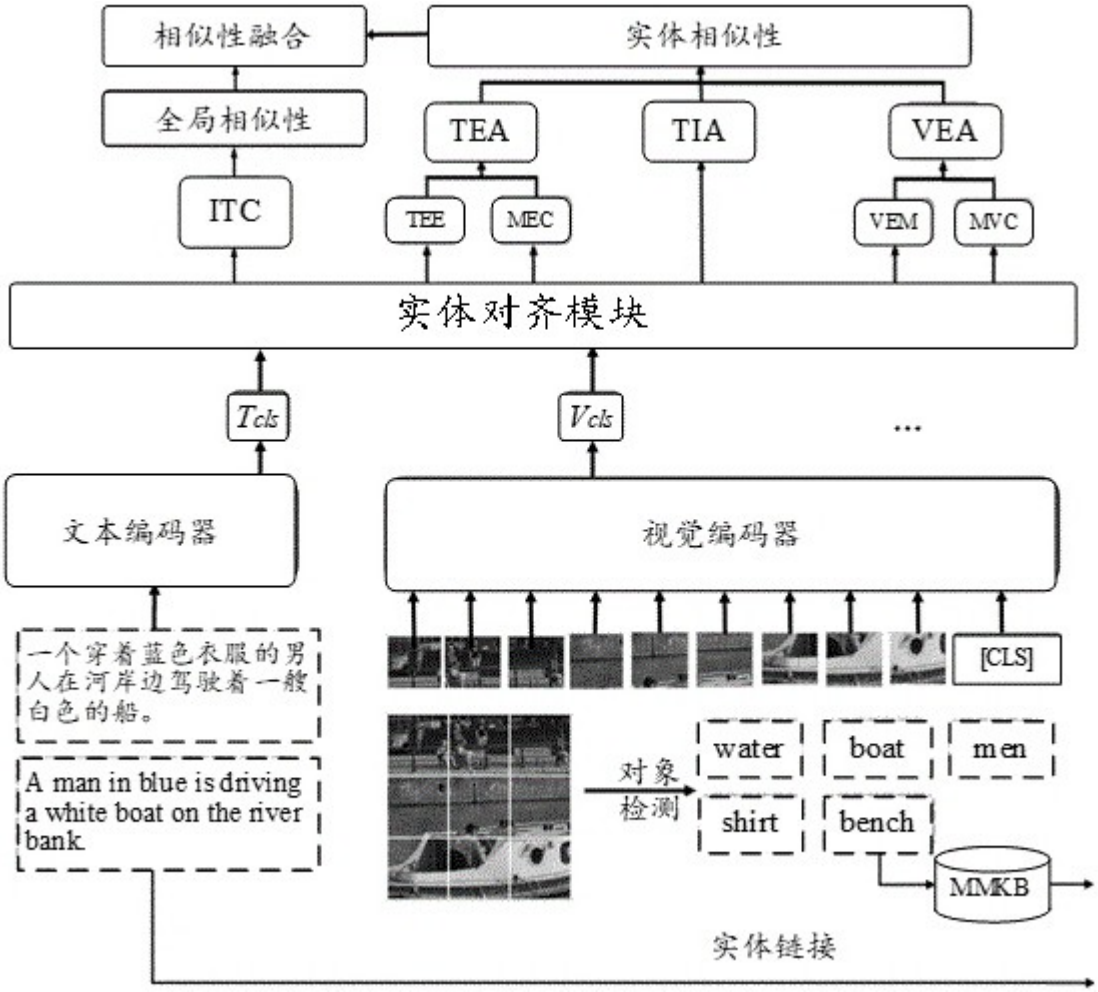

相似性融合
实体相似性
全局相似性
TEA
TIA
VEA
ITC
TEE
MEC
VEM
MVC
实体对齐模块
Tcls
Vcls
...
文本编码器
视觉编码器
[CLS]
一个穿着蓝色衣服的男人在河岸边驾驶着一艘白色的船。
A man in blue is driving a white boat on the river bank.
对象检测
water
boat
men
shirt
bench
MMKB
实体链接

图4A

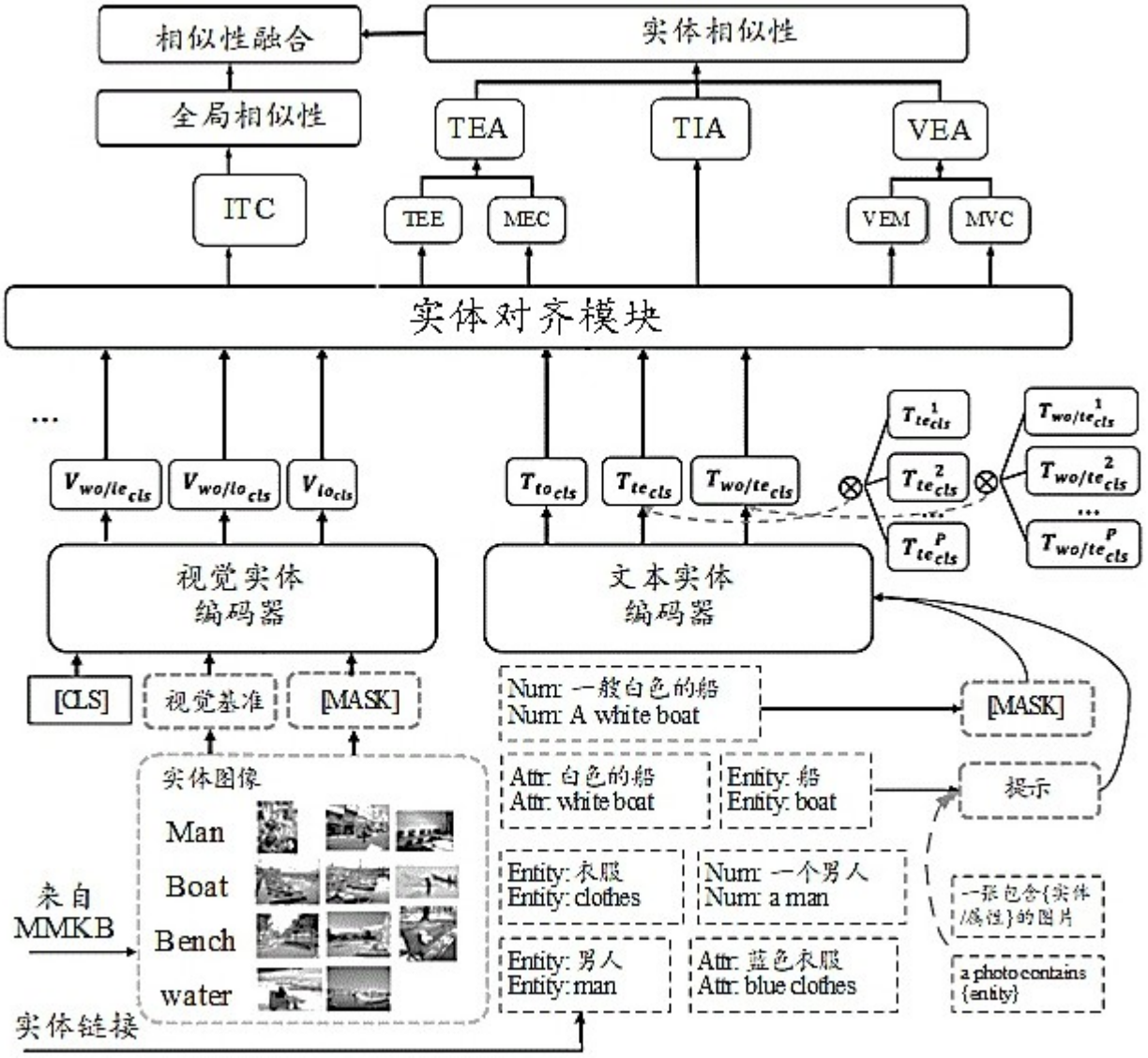

相似性融合
实体相似性
全局相似性
TEA
TIA
VEA
ITC
TEE
MEC
VEM
MVC
实体对齐模块
...
$V_{wo/ie_{cls}}$
$V_{wo/io_{cls}}$
$V_{io_{cls}}$
$T_{to_{cls}}$
$T_{te_{cls}}$
$T_{wo/te_{cls}}$
$T_{te_{cls}}^{1}$
$T_{te_{cls}}^{2}$
$T_{te_{cls}}^{P}$
$T_{wo/te_{cls}}^{1}$
$T_{wo/te_{cls}}^{2}$
$T_{wo/te_{cls}}^{P}$
视觉实体
编码器
文本实体
编码器
[CLS]
视觉基准
[MASK]
[MASK]
提示
实体图像
Man
Boat
Bench
water
来自
MMKB
实体链接
Num: 一般白色的船
Num: A white boat
Attr: 白色的船
Attr: white boat
Entity: 船
Entity: boat
Entity: 衣服
Entity: clothes
Num: 一个男人
Num: a man
Entity: 男人
Entity: man
Attr: 蓝色衣服
Attr: blue clothes
一张包含{实体/属性}的图片
a photo contains {entity}

图4B

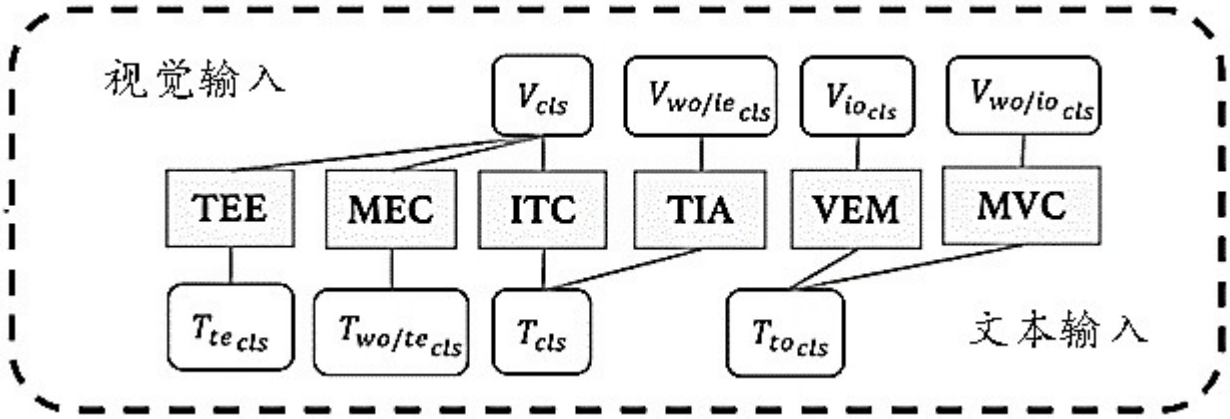

视觉输入
V_{cls}
$V_{wo/ie_{cls}}$
$V_{io_{cls}}$
$V_{wo/io_{cls}}$
TEE
MEC
ITC
TIA
VEM
MVC
$T_{te_{cls}}$
$T_{wo/te_{cls}}$
T_{cls}
$T_{to_{cls}}$
文本输入

图4C

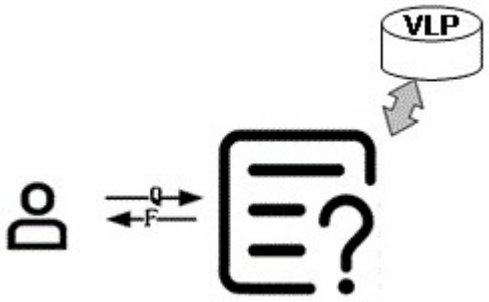

VLP
Q
F

图5

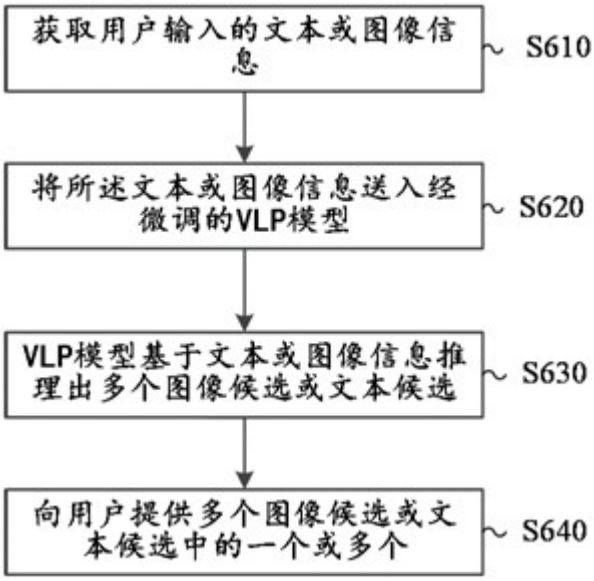

获取用户输入的文本或图像信息
S610
将所述文本或图像信息送入经微调的VLP模型
S620
VLP模型基于文本或图像信息推理出多个图像候选或文本候选
S630
向用户提供多个图像候选或文本候选中的一个或多个
S640

图6

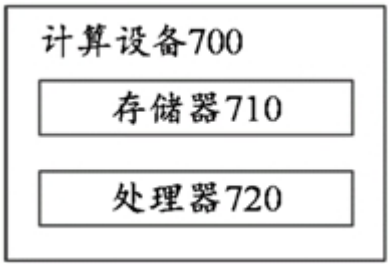

计算设备700
存储器710
处理器720

图7