



(12)发明专利申请

(10)申请公布号 CN 110825901 A

(43)申请公布日 2020.02.21

(21)申请号 201911096538.6

(22)申请日 2019.11.11

(71)申请人 腾讯科技(北京)有限公司

地址 100080 北京市海淀区海淀大街38号

银科大厦16层1601-1608室

(72)发明人 杨皓 李彪 孙剑

(74)专利代理机构 北京派特恩知识产权代理有

限公司 11270

代理人 王姗姗 张颖玲

(51)Int.Cl.

G06F 16/58(2019.01)

G06F 40/30(2020.01)

G06N 3/04(2006.01)

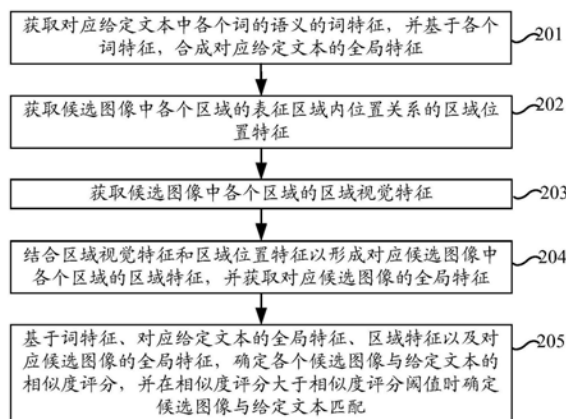
权利要求书4页 说明书21页 附图11页

(54)发明名称

基于人工智能的图文匹配方法、装置、设备及存储介质

(57)摘要

本发明提供了一种基于人工智能的图文匹配方法、装置、电子设备及存储介质,方法包括:获取给定文本中各个词的语义的词特征,基于各个词特征,合成给定文本的全局特征;获取候选图像各个区域的表征区域内位置关系的区域位置特征;获取候选图像中各个区域的区域视觉特征;结合区域视觉特征和区域位置特征形成对应候选图像各个区域的区域特征,并获取对应候选图像的全局特征;基于词特征、给定文本的全局特征、区域特征和候选图像的全局特征,确定各个候选图像与给定文本的相似度评分,相似度评分大于相似度评分阈值时确定候选图像与给定文本匹配,通过本发明,能够提高图文匹配的准确率。



1. 一种基于人工智能的图文匹配方法,其特征在于,所述方法包括:

获取对应给定文本中各个词的语义的词特征,并基于各个词特征,合成对应所述给定文本的全局特征;

获取候选图像中各个区域的表征所述区域内位置关系的区域位置特征;

获取所述候选图像中各个区域的区域视觉特征;

结合所述区域视觉特征和所述区域位置特征以形成对应所述候选图像中各个区域的区域特征,并获取对应所述候选图像的全局特征;

基于所述词特征、对应所述给定文本的全局特征、所述区域特征以及对应所述候选图像的全局特征,确定各个所述候选图像与所述给定文本的相似度评分,并

在所述相似度评分大于相似度评分阈值时确定所述候选图像与所述给定文本匹配。

2. 根据权利要求1所述的方法,其特征在于,所述获取候选图像中各个区域的表征所述区域内位置关系的区域位置特征,包括:

对所述候选图像进行均匀分割,得到多个候选图像块;

从所述候选图像中检测出至少一个表征所述候选图像的语义的区域,并获取对应所述区域的区域视觉特征;

针对多个所述区域的每一个区域执行以下处理:

根据各个所述候选图像块与所述区域的重叠像素数目,获取所述重叠像素数目最多的预设数目的候选图像块,以作为对应所述区域的区域候选图像块;

获取所述区域候选图像块的位置索引特征,并对所述位置索引特征进行更新,得到表征所述区域候选图像块与所述区域内其他区域候选图像块之间的位置关系的位置特征;

基于对应所述区域的区域视觉特征,对所述区域候选图像块的位置特征进行合成,得到表征区域内位置关系的区域位置特征。

3. 根据权利要求2所述的方法,其特征在于,所述对所述位置索引特征进行更新,得到表征所述区域候选图像块与所述区域内其他区域候选图像块之间的位置关系的位置特征,包括:

针对多个所述区域候选图像块的每一个区域候选图像块执行以下处理:

确定所述位置索引特征对应的区域候选图像块与所述区域内其他区域候选图像块的位置关系,以及对应各位置关系的权重;

基于对应各位置关系的权重,对所述位置索引特征进行加权处理,得到表征所述区域候选图像块与所述区域内其他区域候选图像块之间的位置关系的位置特征。

4. 根据权利要求3所述的方法,其特征在于,所述基于对应所述区域的区域视觉特征,对所述区域候选图像块的位置特征进行合成,得到表征区域内位置关系的区域位置特征,包括:

基于对应所述区域的区域视觉特征,获取各区域候选图像块基于视觉维度的权重;

基于对应各区域候选图像块的权重,对各区域候选图像块的位置特征进行加权合成,得到表征区域内位置关系的区域位置特征。

5. 根据权利要求1所述的方法,其特征在于,所述获取候选图像中各个区域的区域视觉特征,包括:

从所述候选图像中检测出至少一个表征所述候选图像的语义的区域,并获取对应所述

区域的区域独立视觉特征;针对多个所述区域的每一个区域执行以下处理:

确定所述区域的区域独立视觉特征与所述候选图像内其他区域的视觉语义关系,以及对应各视觉语义关系的权重;

基于对应各视觉语义关系的权重,对所述区域独立视觉特征进行加权处理,得到表征所述区域与所述候选图像内其他区域之间的视觉关系的区域视觉特征。

6. 根据权利要求1所述的方法,其特征在于,所述基于所述词特征、对应所述给定文本的全局特征、所述区域特征以及对应所述候选图像的全局特征,确定各个所述候选图像与所述给定文本的相似度评分,包括:

基于所述给定文本中各个词的词特征以及所述候选图像中各个区域的区域特征,确定各个所述候选图像与所述给定文本的第一相似度评分,其中,所述第一相似度评分表征在区域维度以及词维度的相似程度;

基于对应所述给定文本的全局特征以及对应所述候选图像的全局特征,确定各个所述候选图像与所述给定文本的第二相似度评分,其中,所述第二相似度评分表征在全局维度的相似程度;

将所述第一相似度评分以及所述第二相似度评分结合对应的权重加权,得到各个所述候选图像与所述给定文本的相似度评分。

7. 根据权利要求1所述的方法,其特征在于,所述方法还包括:

基于给定文本数据库以及候选图像数据库,构建由图文样本对组成的训练样本集合;

其中,所述图文样本对包括具备匹配关系的给定文本样本和候选图像样本;

基于图像特征提取模型以及文本特征提取模型,构建图文匹配模型;

基于所述训练样本集合以及多目标学习任务,训练所述图文匹配模型;

其中,所述多目标学习任务包括:基于给定文本样本对候选图像样本进行相似度评分,以及基于候选图像样本对给定文本样本进行相似度评分的学习任务。

8. 根据权利要求7所述的方法,其特征在于,所述基于所述训练样本集合,以及多目标学习任务,训练所述图文匹配模型,包括:

初始化所述图文匹配模型,并初始化包括输入样本、相似度评分、以及图文匹配模型参数的损失函数;

在所述图文匹配模型每次迭代训练过程中执行以下处理:

将所述训练样本集合包括的所有图文样本对作为输入样本,

通过所述图文匹配模型对所述图文样本对中的给定文本样本进行通向所有候选图像样本的相似度评分,并对所述图文样本对中的候选图像样本进行通向所有给定文本样本的相似度评分;

将所述输入样本和所述相似度评分代入所述损失函数,以确定所述损失函数取得最小值时对应的图文匹配模型参数;

根据所确定的图文匹配模型参数更新所述图文匹配模型。

9. 根据权利要求8所述的方法,其特征在于,所述通过所述图文匹配模型对所述图文样本对中的给定文本样本进行通向所有候选图像样本的相似度评分,包括:

确定所述给定文本样本的各个样本词的样本词特征以及所述给定文本样本的全局特征;

针对所述所有候选图像样本中的每一个候选图像样本执行以下处理：

确定所述候选图像样本中各个区域的区域特征以及所述候选图像样本的全局特征，并针对所述所有给定文本样本中的每一个样本词执行以下处理：

确定所述候选图像样本中各个区域对应所述样本词的注意力权重，并根据各个区域的注意力权重以及各个区域的区域特征进行合成，得到所述样本词在区域上的特征；

基于所述样本词在区域上的特征、所述样本词特征、所述给定文本样本的全局特征和所述候选图像样本的全局特征，确定基于所述给定文本样本进行通向所述候选图像样本的相似度评分。

10. 根据权利要求9所述的方法，其特征在于，所述基于所述样本词在区域上的特征、所述样本词特征、所述给定文本样本的全局特征和所述候选图像样本的全局特征，确定基于所述给定文本样本进行通向所述候选图像样本的相似度评分，包括：

确定各个样本词在区域上的特征和所述样本词特征的相似度的平均值，以作为各个所述候选图像样本与所述给定文本样本的第一相似度评分；

根据所述给定文本样本的全局特征和所述候选图像样本的全局特征，分别确定所述给定文本样本与各个所述候选图像样本的第二相似度评分；

将所述第一相似度评分以及所述第二相似度评分结合对应的权重加权，得到基于所述给定文本样本进行通向所述候选图像样本的相似度评分。

11. 根据权利要求8所述的方法，其特征在于，所述通过所述图文匹配模型对所述图文样本对中的候选图像样本进行通向所有给定文本样本的相似度评分，包括：

确定所述候选图像样本的各个候选图像区域的区域特征以及所述候选图像样本的全局特征；

针对所述所有给定文本样本中的每一个给定文本样本执行以下处理：

确定所述给定文本样本中各个样本词的样本词特征以及各个所述给定文本样本的全局特征，并针对所述所有候选图像样本中的每一个候选图像区域执行以下处理：

确定所述给定文本样本中各个样本词对应所述候选图像区域的注意力权重，根据各个样本词的注意力权重以及各个样本词的样本词特征进行合成，得到所述候选图像区域在文本上的特征；

基于所述候选图像区域在文本上的特征、所述区域特征、所述给定文本样本的全局特征和所述候选图像样本的全局特征，确定基于所述给定文本样本进行通向所述候选图像样本的相似度评分。

12. 根据权利要求11所述的方法，其特征在于，所述确定基于所述候选图像样本进行通向所有给定文本样本的相似度评分，包括：

确定各个候选图像区域在文本上的特征和所述区域特征的相似度的平均值，以作为各个所述给定文本样本与各个所述候选图像样本的第一相似度评分；

根据所述给定文本样本的全局特征和所述候选图像样本的全局特征，分别确定所述给定文本样本与各个所述候选图像样本的第二相似度评分；

将所述第一相似度评分以及所述第二相似度评分结合对应的权重加权，得到基于所述候选图像样本进行通向所有给定文本样本的相似度评分。

13. 一种基于人工智能的图文匹配装置，其特征在于，包括：

文本特征获取模块,用于获取对应给定文本中各个词的语义的词特征,并基于各个词特征,合成对应所述给定文本的全局特征;

图像特征获取模块,用于获取候选图像中各个区域的表征所述区域内位置关系的区域位置特征;获取所述候选图像中各个区域的区域视觉特征;结合所述区域视觉特征和所述区域位置特征以形成对应所述候选图像中各个区域的区域特征,并获取对应所述候选图像的全局特征;

图文匹配模块,用于基于所述词特征、对应所述给定文本的全局特征、所述区域特征以及对应所述候选图像的全局特征,确定各个所述候选图像与所述给定文本的相似度评分,并在所述相似度评分大于相似度评分阈值时确定所述候选图像与所述给定文本匹配。

14. 一种基于人工智能的图文匹配设备,其特征在于,包括:

存储器,用于存储可执行指令;

处理器,用于执行所述存储器中存储的可执行指令时,实现权利要求1至12任一项所述的基于人工智能的图文匹配方法。

15. 一种存储介质,其特征在于,存储有可执行指令,用于引起处理器执行时,实现权利要求1至12任一项所述的基于人工智能的图文匹配方法。

基于人工智能的图文匹配方法、装置、设备及存储介质

技术领域

[0001] 本发明涉及人工智能技术,尤其涉及一种基于人工智能的图文匹配方法、装置、设备及存储介质。

背景技术

[0002] 人工智能(AI,Artificial Intelligence)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、方法和技术及应用系统。

[0003] 深度学习(DL,Deep Learning)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习和归纳学习等技术。

[0004] 随着互联网技术的发展,在越来越多的场景下需要根据文本搜索图像,以得到与文本所表征内容匹配的图像或者根据图像搜索文本,以得到与图像所表征内容匹配的文本,在相关技术中,通过深度学习技术实现选图或者选文本的任务,但是发明人发现在相关技术忽略了图像的区域位置关系对图像所表达内容的影响,以及图像文本的全局信息对图像和文本所表达内容的影响。

发明内容

[0005] 本发明实施例提供一种基于人工智能的图文匹配方法、装置、设备及存储介质,能够提高图文匹配的准确率。

[0006] 本发明实施例的技术方案是这样实现的:

[0007] 本发明实施例提供一种基于人工智能的图文匹配方法,包括:

[0008] 获取对应给定文本中各个词的语义的词特征,并基于各个词特征,合成对应所述给定文本的全局特征;

[0009] 获取候选图像中各个区域的表征所述区域内位置关系的区域位置特征;

[0010] 获取所述候选图像中各个区域的区域视觉特征;

[0011] 结合所述区域视觉特征和所述区域位置特征以形成对应所述候选图像中各个区域的区域特征,并获取对应所述候选图像的全局特征;

[0012] 基于所述词特征、对应所述给定文本的全局特征、所述区域特征以及对应所述候选图像的全局特征,确定各个所述候选图像与所述给定文本的相似度评分,并在所述相似度评分大于相似度评分阈值时确定所述候选图像与所述给定文本匹配。

[0013] 在上述方案中,所述方法还包括:

[0014] 当存在以下情形至少之一时,获取预设候选图像数据库中与所述给定文本相关的多个预设候选图像,以作为多个所述候选图像;

[0015] 所述待匹配对象中不存在所述候选图像,所述待匹配对象中的所有候选图像与所

述给定文本的相似度评分均小于所述相似度评分阈值,或者所述匹配对象中的候选图像的数目小于候选图像数目阈值。

[0016] 在上述方案中,所述获取对应给定文本中各个词的语义的词特征,包括:

[0017] 对所述给定文本进行分词处理,得到多个对应所述给定文本的词,并确定分别对应各个词的词向量;

[0018] 基于对应各个词的词向量,获取各个词的前向词向量以及后向词向量,以确定各个词在所述给定文本中的词特征。

[0019] 本发明实施例提供一种基于人工智能的图文匹配装置,包括:

[0020] 文本特征获取模块,用于获取对应给定文本中各个词的语义的词特征,并基于各个词特征,合成对应所述给定文本的全局特征;

[0021] 图像特征获取模块,用于获取候选图像中各个区域的表征所述区域内位置关系的区域位置特征;获取所述候选图像中各个区域的区域视觉特征;结合所述区域视觉特征和所述区域位置特征以形成对应所述候选图像中各个区域的区域特征,并获取对应所述候选图像的全局特征;

[0022] 图文匹配模块,用于基于所述词特征、对应所述给定文本的全局特征、所述区域特征以及对应所述候选图像的全局特征,确定各个所述候选图像与所述给定文本的相似度评分,并在所述相似度评分大于相似度评分阈值时确定所述候选图像与所述给定文本匹配。

[0023] 在上述方案中,所述基于人工智能的图文匹配装置还包括:

[0024] 预设图像获取模块,用于:

[0025] 当存在以下情形至少之一时,获取预设候选图像数据库中与所述给定文本相关的多个预设候选图像,以作为多个所述候选图像;

[0026] 所述待匹配对象中不存在所述候选图像,所述待匹配对象中的所有候选图像与所述给定文本的相似度评分均小于所述相似度评分阈值,或者所述匹配对象中的候选图像的数目小于候选图像数目阈值。

[0027] 在上述方案中,所述文本特征获取模块,还用于:

[0028] 对所述给定文本进行分词处理,得到多个对应所述给定文本的词,并确定分别对应各个词的词向量;

[0029] 基于对应各个词的词向量,获取各个词的前向词向量以及后向词向量,以确定各个词在所述给定文本中的词特征。

[0030] 在上述方案中,所述图像特征获取模块,还用于:

[0031] 对所述候选图像进行均匀分割,得到多个候选图像块;

[0032] 从所述候选图像中检测出至少一个表征所述候选图像的语义的区域,并获取对应所述区域的区域视觉特征;

[0033] 针对多个所述区域的每一个区域执行以下处理:

[0034] 根据各个所述候选图像块与所述区域的重叠像素数目,获取所述重叠像素数目最多的预设数目的候选图像块,以作为对应所述区域的区域候选图像块;

[0035] 获取所述区域候选图像块的位置索引特征,并对所述位置索引特征进行更新,得到表征所述区域候选图像块与所述区域内其他区域候选图像块之间的位置关系的位置特征;

[0036] 基于对应所述区域的区域视觉特征,对所述区域候选图像块的位置特征进行合成,得到表征区域内位置关系的区域位置特征。

[0037] 在上述方案中,所述图像特征获取模块,还用于:

[0038] 针对多个所述区域候选图像块的每一个区域候选图像块执行以下处理:

[0039] 确定所述位置索引特征对应的区域候选图像块与所述区域内其他区域候选图像块至少两个维度的位置关系,以及对应各位置关系的权重;

[0040] 根据所述至少两个维度的位置关系,对所述位置索引特征进行基于至少两个维度的拆分;

[0041] 基于对应各位置关系的权重,对拆分后的位置索引特征进行加权处理,得到表征所述区域候选图像块与所述区域内其他区域候选图像块之间的位置关系的位置特征。

[0042] 在上述方案中,所述图像特征获取模块,还用于:

[0043] 基于对应所述区域的区域视觉特征,获取各区域候选图像块基于视觉维度的权重;

[0044] 基于对应各区域候选图像块的权重,对各区域候选图像块的位置特征进行加权合成,得到表征区域内位置关系的区域位置特征。

[0045] 在上述方案中,所述图像特征获取模块,还用于:

[0046] 从所述候选图像中检测出至少一个表征所述候选图像的语义的区域,并获取对应所述区域的区域独立视觉特征;

[0047] 针对多个所述区域的每一个区域执行以下处理:

[0048] 确定所述区域的区域独立视觉特征与所述候选图像内其他区域的视觉语义关系,以及对应各视觉语义关系的权重;

[0049] 基于对应各视觉语义关系的权重,对所述区域独立视觉特征进行加权处理,得到表征所述区域与所述候选图像内其他区域之间的视觉关系的区域视觉特征。

[0050] 在上述方案中,所述图文匹配模块,还用于:

[0051] 基于所述给定文本中各个词的词特征以及所述候选图像中各个区域的区域特征,确定各个所述候选图像与所述给定文本的第一相似度评分,其中,所述第一相似度评分表征在区域维度以及词维度的相似程度;

[0052] 基于对应所述给定文本的全局特征以及对应所述候选图像的全局特征,确定各个所述候选图像与所述给定文本的第二相似度评分,其中,所述第二相似度评分表征在全局维度的相似程度;

[0053] 将所述第一相似度评分以及所述第二相似度评分结合对应的权重加权,得到各个所述候选图像与所述给定文本的相似度评分。

[0054] 在上述方案中,所述基于人工智能的图文匹配装置还包括:

[0055] 训练模块,用于:

[0056] 基于给定文本数据库以及候选图像数据库,构建由图文样本对组成的训练样本集合;

[0057] 其中,所述图文样本对包括具备匹配关系的给定文本样本和候选图像样本;

[0058] 基于图像特征提取模型以及文本特征提取模型,构建图文匹配模型;

[0059] 基于所述训练样本集合以及多目标学习任务,训练所述图文匹配模型;

[0060] 其中,所述多目标学习任务包括:基于给定文本样本对候选图像样本进行相似度评分以及基于候选图像样本对给定文本样本进行相似度评分的学习任务。

[0061] 在上述方案中,所述训练模块,还用于:

[0062] 初始化所述图文匹配模型,并初始化包括输入样本、相似度评分、以及所述图文匹配模型的损失函数;

[0063] 在所述图文匹配模型每次迭代训练过程中执行以下处理:

[0064] 将所述训练样本集合包括的所有图文样本对作为输入样本,通过所述图文匹配模型对所述图文样本对中的给定文本样本进行通向所有候选图像样本的相似度评分,并对所述图文样本对中的候选图像样本进行通向所有给定文本样本的相似度评分;

[0065] 将所述输入样本和所述相似度评分代入所述损失函数,以确定所述损失函数取得最小值时对应的图文匹配模型参数;

[0066] 根据所确定的图文匹配模型参数更新所述图文匹配模型。

[0067] 在上述方案中,所述训练模块,还用于:

[0068] 确定所述给定文本样本的各个样本词的样本词特征以及所述给定文本样本的全局特征;

[0069] 针对所述所有候选图像样本中的每一个候选图像样本执行以下处理:

[0070] 确定所述候选图像样本中各个区域的区域特征以及所述候选图像样本的全局特征,并针对所述所有给定文本样本中的每一个样本词执行以下处理:

[0071] 确定所述候选图像样本中各个区域对应所述样本词的注意力权重,并根据各个区域的注意力权重以及各个区域的区域特征进行合成,得到所述样本词在区域上的特征;

[0072] 基于所述样本词在区域上的特征、所述样本词特征、所述给定文本样本的全局特征和所述候选图像样本的全局特征,确定基于所述给定文本样本进行通向所述候选图像样本的相似度评分。

[0073] 在上述方案中,所述训练模块,还用于:

[0074] 确定各个样本词在区域上的特征和所述样本词特征的相似度的平均值,以作为各个所述候选图像样本与所述给定文本样本的第一相似度评分;

[0075] 根据所述给定文本样本的全局特征和所述候选图像样本的全局特征,分别确定所述给定文本样本与各个所述候选图像样本的第二相似度评分;

[0076] 将所述第一相似度评分以及所述第二相似度评分结合对应的权重加权,得到基于所述给定文本样本进行通向所述候选图像样本的相似度评分。

[0077] 在上述方案中,所述训练模块,还用于:

[0078] 确定所述候选图像样本的各个候选图像区域的区域特征以及所述候选图像样本的全局特征;

[0079] 针对所述所有给定文本样本中的每一个给定文本样本执行以下处理:

[0080] 确定所述给定文本样本中各个样本词的样本词特征以及各个所述给定文本样本的全局特征,并针对所述所有候选图像样本中的每一个候选图像区域执行以下处理:

[0081] 确定所述给定文本样本中各个样本词对应所述候选图像区域的注意力权重,

[0082] 根据各个样本词的注意力权重以及各个样本词的样本词特征进行合成,得到所述候选图像区域在文本上的特征;

[0083] 基于所述候选图像区域在文本上的特征、所述区域特征、所述给定文本样本的全局特征和所述候选图像样本的全局特征,确定基于所述给定文本样本进行通向所述候选图像样本的相似度评分。

[0084] 在上述方案中,所述训练模块,还用于:

[0085] 确定各个候选图像区域在文本上的特征和所述区域特征的相似度的平均值,以作为各个所述给定文本样本与各个所述候选图像样本的第一相似度评分;

[0086] 根据所述给定文本样本的全局特征和所述候选图像样本的全局特征,分别确定所述给定文本样本与各个所述候选图像样本的第二相似度评分;

[0087] 将所述第一相似度评分以及所述第二相似度评分结合对应的权重加权,得到基于所述候选图像样本进行通向所有给定文本样本的相似度评分。

[0088] 本发明实施例提供一种基于人工智能的图文匹配电子设备,包括:

[0089] 存储器,用于存储可执行指令;

[0090] 处理器,用于执行所述存储器中存储的可执行指令时,实现本发明实施例提供的基于人工智能的图文匹配方法。

[0091] 本发明实施例提供一种存储介质,存储有可执行指令,用于引起处理器执行时,实现本发明实施例提供的基于人工智能的图文匹配方法。

[0092] 本发明实施例具有以下有益效果:

[0093] 本发明实施例提供的基于人工智能的图文匹配方法实现了将图像中目标区域内的位置关系反映到图像特征中,并结合图像以及文本的全局特征,有效提高了图文匹配的准确率。

附图说明

[0094] 图1是本发明实施例提供的基于人工智能的图文匹配系统架构的一个可选的结构示意图;

[0095] 图2是本发明实施例提供的应用基于人工智能的图文匹配方法的服务器的一个可选的结构示意图;

[0096] 图3是本发明实施例提供的基于人工智能的图文匹配方法的一个可选的流程示意图;

[0097] 图4A-4D是本发明实施例提供的基于人工智能的图文匹配方法的一个可选的流程示意图;

[0098] 图5A-5B是本发明实施例提供的基于人工智能的图文匹配方法中的图文匹配模型的结构图;

[0099] 图6A-6B是本发明实施例提供的基于人工智能的图文匹配方法中的图文匹配模型的相似度计算流程图;

[0100] 图7是本发明实施例提供的基于人工智能的图文匹配方法的应用流程图;

[0101] 图8A-8B是本发明实施例提供的图文匹配模型在以文搜图任务上的效果示意图;

[0102] 图9是本发明实施例提供的图文匹配模型和其他模型在新闻中文数据集上针对以文搜图任务的效果对比示意图;

[0103] 图10是本发明实施例提供的图文匹配模型在短视频中的效果示意图。

具体实施方式

[0104] 为了使本发明的目的、技术方案和优点更加清楚,下面将结合附图对本发明作进一步地详细描述,所描述的实施例不应视为对本发明的限制,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例,都属于本发明保护的范围。

[0105] 在以下的描述中,涉及到“一些实施例”,其描述了所有可能实施例的子集,但是可以理解,“一些实施例”可以是所有可能实施例的相同子集或不同子集,并且可以在不冲突的情况下相互结合。

[0106] 在以下的描述中,所涉及的术语“第一\第二”仅仅是是区别类似的对象,不代表针对对象的特定排序,可以理解地,“第一\第二”在允许的情况下可以互换特定的顺序或先后次序,以使这里描述的本发明实施例能够以除了在这里图示或描述的以外的顺序实施。

[0107] 除非另有定义,本文所使用的所有的技术和科学术语与属于本发明的技术领域的技术人员通常理解的含义相同。本文中所使用的术语只是为了描述本发明实施例的目的,不是旨在限制本发明。

[0108] 对本发明实施例进行进一步详细说明之前,对本发明实施例中涉及的名词和术语进行说明,本发明实施例中涉及的名词和术语适用于如下的解释。

[0109] 1) 循环卷积网络(R-CNN,Region Convolutional Neural Network):用于对图像进行目标检索的深度学习框架。

[0110] 2) 加速的循环卷积网络(Faster R-CNN,Faster Recurrent Convolutional Neural Network):该框架与R-CNN功能一致,具有更快的训练速度。

[0111] 3) 门循环单元(GRU,Gated Recurrent Unit):一种循环神经网络,用于编码句子的语义特征。

[0112] 4) 长短期记忆单元(LSTM,Long Short-Term Memory):一种循环神经网络,与GRU的功能类似,用于编码句子的语义特征。

[0113] 5) 独热编码(One-hot):一种基础特征表示方法,建立数据字典,每个数据可以表示为一个字典大小的向量,其中该数据在字典中索引号位置设为1,其他位置设为0。

[0114] 6) 转换器(Transformer):一种基于编码器(encoder)-译码器(decoder)结构的模型,用于对文本进行编码和解码,计算文本之间的相似度来提升文本的语义表示信息。

[0115] 7) 转换器的双向编码器表示(Bert:Bidirectional Encoder Representations from Transformers):一种预训练语言表示方法,能够进一步提升文本的语义信息。

[0116] 8) 自动编码器(Autoencoders):一种能够通过无监督学习,学到输入数据高效编码表示的人工神经网络,用于对特征进行降维。

[0117] 图文匹配任务作为以文字搜索图像和以图像生成描述这两个任务的基础任务,得到了学者的广泛关注,例如,相关技术中将图像和文本使用两个不同的网络分支学习特征,将两个不同模态的数据转化至同一空间,最后分别使用(图像、正确文本、错误文本)、(文本、正确图像、错误图像)的三元组形式来训练网络;相关技术中提出了一个新的损失来学习图像文本匹配,将每个“图像-文本”对看作是一类,不同“图像-文本”对为不同类别,将损失函数看成多分类问题来训练模型。上述方法针对图像与文本匹配问题,仅考虑图像与文本的完整语义信息进行匹配,然而,图像中包含的对象数目往往很多,文本中包含的对象也类似,因此,存在基于区域“对象”的匹配方法,使用R-CNN提取图像的区域特征,而后使用区

域与文本中各单词计算相似度的方法来计算图像与文本的相似度,或者使用Faster R-CNN来提取图像的区域特征,特征的表征能力更强,而且使用注意力策略来计算区域与单词的评分,使模型使用效果更好,或者不仅考虑了区域的视觉特征而且考虑了区域的位置信息,因为相同对象在不同位置表达含义可能不一样,通过增加位置与视觉特征的注意力机制来提升区域特征表示。

[0118] 然而,发明人发现相关技术中考虑了图像区域的视觉信息和位置信息,而忽略了位置信息之间的强相互关系,本发明实施例提供的基于人工智能的图文匹配方法考虑了图像区域的不同位置间的关系,使用多头自注意力策略重新计算不同位置的特征,同时将以文本搜索图像和以图像生成文本两个任务进行多目标学习,进一步提升模型效果,最后,基于发现图像中不同“对象”间存在顺序关系的构思,将完整图像和完整文本信息重新进行考虑,将局部特征和全局特征同时考虑,以增强模型在图像与文本相互关系的学习能力。

[0119] 相关技术提供的图文匹配方法导致的主要问题是,以文本搜索图像得到的搜索结果差异大,搜索出来的图像结果与文本语义不相关,例如,对于一条新闻而言,根据新闻的标题文本信息,对新闻内的图像使用图文匹配模型进行排序输出,前几名的图像与文本在语义上相关程度不高,本发明实施例提供的基于人工智能的图文匹配方法提出了基于位置关系的多目标学习网络,相比于相关技术而言,其在两个公开数据集上表现达到了最佳程度,在新闻的选图业务中表现最好。

[0120] 本发明实施例提供一种基于人工智能的图文匹配方法、装置、电子设备和存储介质,能够提高图文匹配的准确率,下面说明本发明实施例提供的基于人工智能的图文匹配设备的示例性应用,本发明实施例提供的基于人工智能的图文匹配设备可以实施为笔记本电脑,平板电脑,台式计算机,机顶盒,移动设备(例如,移动电话,便携式音乐播放器,个人数字助理,专用消息设备,便携式游戏设备)等各种类型的用户终端,也可以实施为服务器。下面,将说明设备实施为服务器时示例性应用。

[0121] 参见图1,图1是本发明实施例提供的基于人工智能的图文匹配系统100的一个可选的架构示意图,终端400通过网络300连接服务器200,网络300可以是广域网或者局域网,又或者是二者的组合,服务器200连接数据库500,服务器200通过本发明实施例提供的基于人工智能的图文匹配方法对新闻的标签或者标题进行新闻配图,候选图像的来源可以是新闻中的图像,优先对新闻中的图像进行匹配排序评分,将匹配排序的结果推送给编辑后台600,以使编辑后台600能够对获得的推送图像进行编辑,以生成最终的具备图文匹配标题的新闻,并返回给服务器200,以使服务器200将新闻推送给终端400进行显示,除了将新闻中的图像作为候选图像之外,还可以从预设数据库中的图像进行相对于给定标题或者标签文本的匹配排序评分。

[0122] 参见图2,图2是本发明实施例提供的基于人工智能的图文匹配服务器200的结构示意图,图2所示的服务器200包括:至少一个处理器210、存储器250、至少一个网络接口220和用户接口230。服务器200中的各个组件通过总线系统240耦合在一起。可理解,总线系统240用于实现这些组件之间的连接通信。总线系统240除包括数据总线之外,还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见,在图2中将各种总线都标为总线系统240。

[0123] 处理器210可以是一种集成电路芯片,具有信号的处理能力,例如通用处理器、数

字信号处理器(DSP,Digital Signal Processor),或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等,其中,通用处理器可以是微处理器或者任何常规的处理器等。

[0124] 用户接口230包括使得能够呈现媒体内容的一个或多个输出装置231,包括一个或多个扬声器和/或一个或多个视觉显示屏。用户接口230还包括一个或多个输入装置232,包括有助于用户输入的用户接口部件,比如键盘、鼠标、麦克风、触屏显示屏、摄像头、其他输入按钮和控件。

[0125] 存储器250可以是可移除的,不可移除的或其组合。示例性的硬件设备包括固态存储器,硬盘驱动器,光盘驱动器等。存储器250可选地包括在物理位置上远离处理器210的一个或多个存储设备。

[0126] 存储器250包括易失性存储器或非易失性存储器,也可包括易失性和非易失性存储器两者。非易失性存储器可以是只读存储器(ROM,Read Only Memory),易失性存储器可以是随机存取存储器(RAM,Random Access Memory)。本发明实施例描述的存储器250旨在包括任意适合类型的存储器。

[0127] 在一些实施例中,存储器250能够存储数据以支持各种操作,这些数据的示例包括程序、模块和数据结构或者其子集或超集,下面示例性说明。

[0128] 操作系统251,包括用于处理各种基本系统服务和执行硬件相关任务的系统程序,例如框架层、核心库层、驱动层等,用于实现各种基础业务以及处理基于硬件的任务;

[0129] 网络通信模块252,用于经由一个或多个(有线或无线)网络接口220到达其他计算设备,示例性的网络接口220包括:蓝牙、无线相容性认证(WiFi)、和通用串行总线(USB,Universal Serial Bus)等;

[0130] 呈现模块253,用于经由一个或多个与用户接口230相关联的输出装置231(例如,显示屏、扬声器等)使得能够呈现信息(例如,用于操作外围设备和显示内容和信息的用户接口);

[0131] 输入处理模块254,用于对一个或多个来自一个或多个输入装置232之一的一个或多个用户输入或互动进行检测以及翻译所检测的输入或互动。

[0132] 在一些实施例中,本发明实施例提供的装置可以采用软件方式实现,图2示出了存储在存储器250中的基于人工智能的图文匹配装置255,其可以是程序和插件等形式的软件,包括以下软件模块:文本特征获取模块2551、图像特征获取模块2552、图文匹配模块2553、预设图像获取模块2554和训练模块2555,这些模块是逻辑上的,因此根据所实现的功能可以进行任意的组合或进一步拆分,将在下文中说明各个模块的功能。

[0133] 在另一些实施例中,本发明实施例提供的基于人工智能的图文匹配装置可以采用硬件方式实现,作为示例,本发明实施例提供的基于人工智能的图文匹配装置可以是采用硬件译码处理器形式的处理器,其被编程以执行本发明实施例提供的基于人工智能的图文匹配方法,例如,硬件译码处理器形式的处理器可以采用一个或多个应用专用集成电路(ASIC,Application Specific Integrated Circuit)、DSP、可编程逻辑器件(PLD,Programmable Logic Device)、复杂可编程逻辑器件(CPLD,Complex Programmable Logic Device)、现场可编程门阵列(FPGA,Field-Programmable Gate Array)或其他电子元件。

[0134] 下面将结合本发明实施例提供的终端或者服务器的示例性应用和实施,说明本发

明实施例提供的基于人工智能的图文匹配方法。

[0135] 下面分两个阶段说明本发明实施例提供的基于人工智能的图文匹配方法,第一个部分是模型的训练阶段,第二个部分是利用模型的匹配阶段。

[0136] 下面说明模型的训练阶段,参见图3,图3是本发明实施例提供的基于人工智能的图文匹配方法的一个可选的流程示意图,将结合图3示出的步骤101-103进行说明,下述方法的步骤可以在上述任意类型的电子设备(例如终端或服务器)上实现。

[0137] 在步骤101中,基于给定文本数据库以及候选图像数据库,构建由图文样本对组成的训练样本集合;其中,图文样本对包括具备匹配关系的给定文本样本和候选图像样本。

[0138] 在一些实施例中,给定文本数据库以及候选图像数据库可以是私有库,也可以是共有库,且给定文本数据库中的数据集可以是英语数据集或者是中文数据集,还可以使用不同语言的数据库,还可以对数据库中的图像或者文本的语义类别进行限制,可以使用专门的新闻文本数据库和新闻图像数据库,也可以使用其他信息流数据库,例如,视频封面图像数据库等等,广告文本和广告图像数据库等等,从数据库中构造由候选图像样本和给定文本组成的图文样本对,给定文本样本可以是新闻的标题或者标签,候选图像样本可以是新闻图像,图文样本对中的候选图像样本和给定文本样本具备匹配关系,假设有100对具备匹配关系的图文样本对,其中,图像A和文本A配对,为正样本,则图像A和其余文本不具备匹配关系,即为负样本。

[0139] 在步骤102中,基于图像特征提取模型以及文本特征提取模型,构建图文匹配模型。

[0140] 这里的图像特征提取模型可以是R-CNN或者Fast R-CNN,文本特征提取模型可以是双向GRU或者是双向LSTM,通过图像特征提取模型和文本特征提取模型分别对图像特征以及文本特征进行提取,并对两种特征进行维度统一之后进行相似度评分。

[0141] 在步骤103中,基于训练样本集合以及多目标学习任务,训练图文匹配模型;其中,多目标学习任务包括:基于给定文本样本对候选图像样本进行相似度评分以及基于候选图像样本对给定文本样本进行相似度评分的学习任务。

[0142] 在一些实施例中,基于训练样本集合以及多目标学习任务,训练图文匹配模型,其中,多目标学习任务包括:基于给定文本样本对候选图像样本进行相似度评分以及基于候选图像样本对给定文本样本进行相似度评分的学习任务。这里的多目标学习任务是指同时对两个任务进行学习,这两个任务分别是文本到图像的注意力评分以及从图像到文本的注意力评分,从文本到图像的注意力评分为 $S(I, T) = S_{t2i}(I, T)$, I代表图像的特征, T代表文本的特征,从图像到文本的注意力评分为 $S(I, T) = S_{i2t}(I, T)$, 将两个相似度的计算任务同时在网络中学习,多任务学习中使用的图像与文本的评分函数参见公式(1):

$$[0143] \quad S(I, T) = \alpha (\lambda S_{t2i}(I, T) + (1-\lambda) S_{i2t}(I, T)) + (1-\alpha) S(v_w, e_w) \quad (1)$$

[0144] 其中, $S_{t2i}(I, T)$ 和 $S_{i2t}(I, T)$ 分别表示从文本到图像的注意力评分和从图像到文本的注意力评分, $1-\alpha$ 表示完整图像与完整文本相似度评分所占的权重, α 表示区域以及词维度的评分权重,均属于超参数,其中 $1-\lambda$ 和 λ 表示基于图像生成文本以及基于文本搜索图像这两个任务在学习过程中的权重,这里的 λ 在模型训练的不同轮次可以动态调整,属于超参数,在训练某一轮次的过程中不会发生变化。这里,我们对图像 I 提取完整的特征 v_w , 同样对应文本,通过使用双向GRU最后一个输出的特征作为完整文本的特征 e_w , $S(v_w, e_w)$ 是图像全

局特征和文本全局特征之间的相似度。

[0145] 在一些实施例中,步骤103中基于训练样本集合,以及多目标学习任务,训练图文匹配模型,可以通过以下技术方案实现,初始化图文匹配模型,并初始化包括输入样本、相似度评分、以及图文匹配模型参数的损失函数;在图文匹配模型每次迭代训练过程中执行以下处理:将训练样本集合包括的所有图文样本对作为输入样本,通过图文匹配模型对图文样本对中的给定文本样本以及候选图像样本分别进行通向所有候选图像样本的相似度评分以及通向所有给定文本样本的相似度评分;将输入样本和相似度评分代入损失函数,以确定损失函数取得最小值时对应的图文匹配模型参数;根据所确定的图文匹配模型参数更新图文匹配模型,损失函数如下公式(2)所示:

$$[0146] \quad l(I, T) = \sum_T \left[a - S(I, T) + S(I, \hat{T}) \right]_+ \sum_T \left[a - S(I, T) + S(\hat{I}, T) \right]_+ \quad (2);$$

[0147] 这里的 $S(I, \hat{T})$ 表示从文本到图像的预测相似度, $S(\hat{I}, T)$ 这里表示从图像到文本的预测相似度, $S(I, T)$ 表示文本到图像的实际相似度,+表示括号内的值大于零的时候,取该值为损失值,括号内的值小于零的时候,损失值为零,训练过程的目的是使损失函数 $l(I, T)$ 在训练迭代中下降的越小越好,也就是要使得主体对象与正样本越接近越好,主体对象与负样本越远越好,以下是对间隔值 a 的分析,当间隔值越小时,损失函数值较容易趋近于0,因此,主体对象与正样本不需要拉得太近,主体对象与负样本不需要拉的太远,就能使得损失函数值很快地趋近于0,这样训练得到的结果不能够很好的区分不同图文样本对的评分,当主体对象越大时,就需要使得网络参数要拼命地拉近主体对象和正样本之间的距离,拉远主体对象和负样本之间的距离,如果间隔值设置的太大,很可能最后损失函数值保持一个较大的值,难以趋近于0,因此,设置一个合理的间隔值很关键,这是衡量相似度的重要指标,简而言之,间隔值设置的越小,损失函数值很容易趋近于0,但很难区分相似的图像,间隔值设置的越大,损失函数值较难趋近于0,甚至导致网络不收敛,但可以较有把握的区分不同图文样本对的评分。

[0148] 在一些实施例中,通过图文匹配模型对图文样本对中的给定文本样本进行通向所有候选图像样本的相似度评分,可以通过以下技术方案实现,确定给定文本样本的各个样本词的样本词特征以及给定文本样本的全局特征;针对所有候选图像样本中的每一个候选图像样本执行以下处理:确定候选图像样本中各个区域的区域特征以及候选图像样本的全局特征,并针对所有给定文本样本中的每一个样本词执行以下处理:确定候选图像样本中各个区域对应样本词的注意力权重,以根据各个区域的注意力权重以及各个区域的区域特征进行合成,得到样本词在区域上的特征;基于样本词在区域上的特征、样本词特征、给定文本样本的全局特征和候选图像样本的全局特征,确定基于给定文本样本进行通向候选图像样本的相似度评分。

[0149] 在一些实施例中,基于样本词在区域上的特征、样本词特征、给定文本样本的全局特征和候选图像样本的全局特征,确定基于给定文本样本进行通向候选图像样本的相似度评分,可以通过以下技术方案实现,确定各个样本词在区域上的特征和样本词特征的相似度的平均值,以作为表征区域维度以及词维度层面的各个候选图像样本与给定文本样本的第一相似度评分;根据给定文本样本的全局特征和候选图像样本的全局特征,分别确定给定文本样本与各个候选图像样本的第二相似度评分;将第一相似度评分以及第二相似度评

分结合对应的权重加权,得到基于给定文本样本进行通向候选图像样本的相似度评分。

[0150] 在一些实施例中,通过图文匹配模型对图文样本对中的候选图像样本进行通向所有给定文本样本的相似度评分,可以通过以下技术方案实现,确定候选图像样本的各个候选图像区域的区域特征以及候选图像样本的全局特征;针对所有给定文本样本中的每一个给定文本样本执行以下处理:确定给定文本样本中各个样本词的样本词特征以及各个给定文本样本的全局特征,并针对所有候选图像样本中的每一个候选图像区域执行以下处理:确定给定文本样本中各个样本词对应候选图像区域的注意力权重,根据各个样本词的注意力权重以及各个样本词的样本词特征进行合成,得到候选图像区域在文本上的特征;基于候选图像区域在文本上的特征、区域特征、给定文本样本的全局特征和候选图像样本的全局特征,确定基于给定文本样本进行通向候选图像样本的相似度评分。

[0151] 在一些实施例中,确定基于候选图像样本进行通向所有给定文本样本的相似度评分,可以通过以下技术方案实现,确定各个候选图像区域在文本上的特征和区域特征的相似度的平均值,以作为表征区域维度以及词维度层面的各个给定文本样本与各个候选图像样本的第一相似度评分;根据给定文本样本的全局特征和候选图像样本的全局特征,分别确定给定文本样本与各个候选图像样本的第二相似度评分;将第一相似度评分以及第二相似度评分结合对应的权重加权,得到基于候选图像样本进行通向所有给定文本样本的相似度评分。

[0152] 下面说明本发明实施例提供的基于人工智能的对象识别方法的识别阶段。

[0153] 参见图4A,图4A是本发明实施例提供的基于人工智能的图文匹配方法的一个可选的流程示意图,将结合图4A示出的步骤201-204进行说明,下述方法的步骤可以在上述任意类型的电子设备(例如终端或服务器)上实现。

[0154] 在步骤201中,获取对应给定文本中各个词的语义的词特征,并基于各个词特征,合成对应给定文本的全局特征;

[0155] 参见图4B,基于图4A,步骤201中获取对应给定文本中各个词的语义的词特征,可以通过下述步骤2011-2012具体实现。

[0156] 在步骤2011中,对给定文本进行分词处理,得到多个对应给定文本的词,并确定分别对应各个词的词向量。

[0157] 在步骤2012中,基于对应各个词的词向量,获取各个词的前向词向量以及后向词向量,以确定各个词在给定文本中的词特征。

[0158] 在一些实施例中,使用数学模型处理文本语料的第一步就是把文本转换成数学表示,可以通过One-hot矩阵表示一个单词,One-hot矩阵是指每一行有且只有一个元素为1,其他元素都是0的矩阵,针对字典中的每个单词,我们分配一个编号,对某句话进行编码时,将里面的每个单词转换成字典里面这个单词编号对应的位置为1的one-hot矩阵,或者,使用词嵌入矩阵给每个单词分配一个固定长度的向量表示,该长度实际上会远远小于前述字典长度,且两个单词向量之间的夹角值可以作为他们之间关系的一个衡量,进而通过双向GRU网络可以获取各个词在给定文本中的词特征。

[0159] 在步骤202中,获取候选图像中各个区域的表征区域内位置关系的区域位置特征;

[0160] 参见图4C,基于图4A,步骤202中获取候选图像中各个区域的表征区域内位置关系的区域位置特征,可以通过下述步骤2021-2023具体实现。

[0161] 在步骤2021中,对候选图像进行均匀分割,得到多个候选图像块。

[0162] 在步骤2022中,从候选图像中检测出至少一个表征候选图像的语义的区域,并获取对应区域的区域视觉特征。

[0163] 在一些实施例中,对候选图像块进行均等分割,均等分割得到了多个候选图像块,每个候选图像块的大小相同,通过目标检测模型,例如前述R-CNN模型检测出图像中的目标对象所在的区域,并直接获取对应区域的区域独立视觉特征作为对应区域的区域视觉特征,这里的区域视觉特征表征了对应区域的内容,还可以对这里的区域独立视觉特征进行更新,对区域独立视觉特征进行更新的目的在于考虑到不同区域之间的视觉内容关联,例如,若图像中存在爸爸牵着小男孩的手,则“爸爸”所在的区域所表征的内容与“小男孩”所在的区域所表征的内容之间具有相关度,基于不同区域之间的视觉内容联系对各个区域的区域独立视觉特征进行更新,使更新后得到的区域视觉特征能够表征候选图像内不同区域之间的视觉关系。

[0164] 在步骤2023中,针对多个区域的每一个区域执行以下步骤20231-20233:

[0165] 在步骤20231中,根据各个候选图像块与区域的重叠像素数目,获取重叠像素数目最多的预设数目的候选图像块,以作为对应区域的区域候选图像块。

[0166] 在步骤20232中,获取区域候选图像块的位置索引特征,并对位置索引特征进行更新,得到表征区域候选图像块与区域内其他区域候选图像块之间的位置关系的位置特征。

[0167] 在步骤20233中,基于对应区域的区域视觉特征,对区域候选图像块的位置特征进行合成,得到表征区域内位置关系的区域位置特征。

[0168] 在一些实施例中,步骤20232中对位置索引特征进行更新,得到表征区域候选图像块与区域内其他区域候选图像块之间的位置关系的位置特征,可以通过以下技术方案实现,针对多个区域候选图像块的每一个区域候选图像块执行以下处理:确定位置索引特征对应的区域候选图像块与区域内其他区域候选图像块至少两个维度的位置关系,以及对应各位置关系的权重;根据至少两个维度的位置关系,对位置索引特征进行基于至少两个维度的拆分;基于对应各位置关系的权重,对拆分后的位置索引特征进行加权处理,得到表征区域候选图像块与区域内其他区域候选图像块之间的位置关系的位置特征。

[0169] 在一些实施例中,步骤20233中基于对应区域的区域视觉特征,对区域候选图像块的位置特征进行合成,得到表征区域内位置关系的区域位置特征,可以通过以下技术方案实现,基于对应区域的区域视觉特征,获取各区域候选图像块基于视觉维度的权重;基于对应各区域候选图像块的权重,对各区域候选图像块的位置特征进行加权合成,得到表征区域内位置关系的区域位置特征。

[0170] 在步骤203中,获取候选图像中各个区域的区域视觉特征。

[0171] 在一些实施例中,步骤203中获取候选图像中各个区域的区域视觉特征,可以通过以下技术方案实现,从候选图像中检测出至少一个表征候选图像的语义的区域,并获取对应区域的区域独立视觉特征;针对多个区域的每一个区域执行以下处理:确定区域的区域独立视觉特征与候选图像内其他区域的视觉语义关系,以及对应各视觉语义关系的权重;基于对应各视觉语义关系的权重,对区域独立视觉特征进行加权处理,得到表征区域与候选图像内其他区域之间的视觉关系的区域视觉特征。

[0172] 在一些实施例中,得到表征区域与候选图像内其他区域之间的视觉关系的区域视

觉特征的过程实际上是对独立视觉特征进行更新的过程,在通过自注意力模型更新的过程中,创建三个向量,对于每个区域,创建一个查询向量,一个键向量和一个值向量,这些向量是通过词嵌入向量乘以自注意力模型训练过程中创建的三个训练矩阵而产生的。在确定区域的自注意力权重时,需要根据这个区域对候选图像的每个区域进行评分,当获取区域的视觉特征时,分数决定了对候选图像的其他区域的关注程度,这里的关注程度实际是基于区域之间的视觉内容关系产生的。通过将查询向量和键向量点积来对相应的区域评分,并将点积得到的得分除以键向量的维度的平方根,以使得训练过程具有稳定的梯度,这里的除数不仅限于键向量的维度的平方根,最后通过最大似然函数进行标准化,使得最后的评分之和为1,其决定了当前区域在候选图像中每个区域的表示程度,最后将每个值向量乘以经过标准化的得分,其实际意义在于保存对当前区域的关注度不变的情况下,降低对不相关区域的关注,从而产生自注意力模型对于某一个区域的输出特征,作为区域视觉特征,上述自注意模型可以进行多次计算,从而在不同的表示子空间里学习到相关的信息。

[0173] 在步骤204中,结合区域视觉特征和区域位置特征以形成对应候选图像中各个区域的区域特征,并获取对应候选图像的全局特征;

[0174] 在步骤205中,基于词特征、对应给定文本的全局特征、区域特征以及对应候选图像的全局特征,确定各个候选图像与给定文本的相似度评分,并在相似度评分大于相似度评分阈值时确定候选图像与给定文本匹配。

[0175] 参见图4D,基于图4A,步骤205中基于词特征、对应给定文本的全局特征、区域特征以及对应候选图像的全局特征,确定各个候选图像与给定文本的相似度评分,并在相似度评分大于相似度评分阈值时确定候选图像与给定文本匹配,可以通过下述步骤2051-2053具体实现。

[0176] 在步骤2051中,基于给定文本中各个词的词特征以及候选图像中各个区域的区域特征,确定表征区域维度以及词维度层面的各个候选图像与给定文本的第一相似度评分。

[0177] 在步骤2052中,基于对应给定文本的全局特征以及对应候选图像的全局特征,确定表征全局维度层面的各个候选图像与给定文本的第二相似度评分。

[0178] 在步骤2053中,将第一相似度评分以及第二相似度评分结合对应的权重加权,得到各个候选图像与给定文本的相似度评分。

[0179] 在一些实施例中,基于给定文本中各个词的词特征以及候选图像中各个区域的区域特征,确定表征区域维度以及词维度层面的各个候选图像与给定文本的第一相似度评分,基于对应给定文本的全局特征以及对应候选图像的全局特征,确定表征全局维度层面的各个候选图像与给定文本的第二相似度评分,将第一相似度评分以及第二相似度评分结合对应的权重加权,得到各个候选图像与给定文本的相似度评分。

[0180] 在一些实施例中,当存在以下情形至少之一时,获取预设候选图像数据库中与给定文本相关的多个预设候选图像,以作为多个候选图像:待匹配对象中不存在候选图像,待匹配对象中的所有候选图像与给定文本的相似度评分均小于相似度评分阈值,或者匹配对象中的候选图像的数目小于候选图像数目阈值。

[0181] 下面,将说明本发明实施例提供的基于人工智能的图文匹配方法在一个实际的应用场景中的示例性应用。

[0182] 通过本发明实施例提供的基于人工智能的图文匹配方法实现针对新闻应用的选

图任务,从新闻内自动选择与新闻标题最相关的图像作为客户端上的新闻封面图进行展示以吸引用户进行新闻浏览,其流程具体为:输入新闻的标题信息和标签信息,这里的标题信息和标签信息均是文本类信息,将新闻内所有图像作为候选图像,通过本发明实施例提供的基于人工智能的图文匹配方法中的图文匹配模型对候选图像进行排序评分,输出评分最好的预设数目的图像,供编辑平台进行图像筛选,将筛选结果以大图或者小图的形式在客户端上作为当前新闻的封面图进行展示。

[0183] 参见图5A-5B,图5A-5B是本发明实施例提供的基于人工智能的图文匹配方法中的图文匹配模型的结构图,通过目标检测网络(例如:Faster R-CNN)对候选图像中的区域特征进行提取,例如,候选图像中的“小男孩”/“boy”这个区域的区域特征 v_i ;对于给定文本而言,通过词嵌入向量确定各单词的特征,再通过双向GRU得到各单词在当前语境中的特征,这里的双向GRU可以被BERT技术替代,比如,这里的“小男孩”/“boy”单词特征 e_t 。这里还会提取图像的全局特征 g^e ,以及文本的全局特征 E ,词嵌入向量是将自然语言表示的单词转换为计算机能够理解的向量或矩阵形式,可以理解为是词的低维向量表示,从而可以实施各种基于向量的计算,例如,通过向量之间的相似度来度量词之间的语义相关性等等。

[0184] 对于候选图像而言,候选图像 I 由一个特征集合表示($I = \{v_1, \dots, v_n\}$),用于表示候选图像 I 中的所有区域的区域特征,给定文本 T 由一个特征集合表示($T = \{e_1, \dots, e_m\}$),用于表示给定文本 T 中所有的单词特征,考虑到基于图片生成文本以及基于文本搜索图片这两个模型分别进行训练比较耗时,而且浪费图像处理器的使用资源,且在预测阶段也需要使用两个模型,导致预测的时间长,本发明实施例提供的基于人工智能的图文匹配模型的训练过程中使用多目标策略,将图像到文本以及文本到图像这两个相似度的计算任务同时在网络中学习,同时,不同区域之间也具备顺序关系,“小男孩”/“boy”“男人”/“man”等区域在文本语义上具备先后顺序,因此,将图像的完整特征和文本的完整特征的相似度也考虑到图像与文本之间相互关系的评分函数中。在计算图像与文本的相似度时,考虑完整图像与完整文本的相似度评分,基于候选图像与给定文本的基于注意力机制的相似度计算方式,获得表征图像与文本之间相互关系的评分函数为:

$$[0185] \quad S(I, T) = \alpha (\lambda S_{t2i}(I, T) + (1-\lambda) S_{i2t}(I, T)) + (1-\alpha) S(v_w, e_w) \quad (1);$$

[0186] 其中, $S_{t2i}(I, T)$ 和 $S_{i2t}(I, T)$ 分别表示从文本到图像的注意力评分和从图像到文本的注意力评分, $1-\alpha$ 表示完整图像与完整文本相似度评分所占的权重,属于超参数,其中 $1-\lambda$ 和 λ 表示基于图像生成文本以及基于文本搜索图像这两个任务在学习过程中的权重,这里的 λ 在模型训练的不同轮次可以动态调整,属于超参数,在训练某一轮次的过程中不会发生变化。这里,我们对图像 I 提取完整的特征 v_w ,即前文所述的图像的全局特征 g^e ,同样对应文本,通过使用双向GRU最后一个输出的特征作为完整文本的特征 e_w ,即前文所述的文本的全局特征 E ,若使用特征提取模型提取完整图像特征后,连接全连接层,并与完整文本特征匹配评分,无论是训练还是固定这层全连接层得到的模型效果均特别差,这是由于全局图像特征与之前区域特征的分布不一致,为此,通过自动编码器策略对图像的全局特征进行编码降维,然后使用编码后的特征作为图像的全局特征,再接入后续的网络进行训练,模型效果有显著的提升。 $S_{t2i}(I, T)$ 和 $S_{i2t}(I, T)$ 的计算方式类似,下面具体介绍 $S_{i2t}(I, T)$ 的计算过程。

[0187] 参见图6A-6B,图6A-6B是本发明实施例提供的基于人工智能的图文匹配方法中的

图文匹配模型的相似度计算流程图,计算“小男孩”/“boy”的区域特征 v_1 与给定文本中各单词的特征之间的相似度,这里通过单词的字体粗细来表征相似度的大小,可以看到单词“小男孩”/“boy”的字体最粗,说明“小男孩”/“boy”的区域与单词“小男孩”/“boy”最相似,将这些相似度作为各单词的权重,对各单词进行加权平均计算得到“小男孩”/“boy”的区域在文本语义上的特征向量 a_1^t ,然后,确定独立的区域特征 v_1 与经过注意力策略得到的 a_1^t 之间的相似度 $R(v_1, a_1^t)$,图6A-6B中右侧的圆点颜色越深表示相似度值越大,对所有区域的相似度进行平均处理,得到从图像到文本的注意力评分,如公式(3)所示:

$$[0188] \quad S_{i2t}(I, T) = \sum_{i=1}^n R(v_i, a_i^t) \quad (3);$$

[0189] 其中, $R(a, b)$ 表示求解 a 与 b 之间的相似度。

[0190] 区域特征 v_i 同时考虑了区域视觉特征和区域位置特征,区域视觉特征可以为2048维特征向量,区域位置特征可以为200维特征向量,对于区域位置特的提取,首先将候选图像均等分割成 $K \times K$ 个“块”,块集合表示为 $B = \{b_1, \dots, b_{K^2}\}$,每个“块”对应于200维的特征向量,确定 v_i 区域与所有“块”的占比面积,提取前 L 个面积最大的“块”,前 L 个面积最大的“块” $D = \{b_{i_1}, \dots, b_{i_L}\}$,这里的占比面积可以通过重合的像素数目表示,考虑到,前 L 个面积最大的“块”之间具有相邻关系,因此基于“块”与“块”间的关系,重新对“块”特征进行更新,例如,上述前 L 个面积最大的“块” $D = \{b_{i_1}, \dots, b_{i_L}\}$, b_i 为200维的特征向量,通过多头自注意力机制,参见公式(4)(5)(6)重新确定这 L 个“块”的特征向量 $D' = \{b'_{i_1}, \dots, b'_{i_L}\}$,

$$[0191] \quad b'_{i_j} = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \quad (4)$$

$$[0192] \quad \text{head}_i = \text{Attention}(DW_i^Q, DW_i^K, DW_i^V) \quad (5)$$

$$[0193] \quad \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

[0194] 当通过自注意力在处理每个“块”时,可以帮助模型查看块序列中的其他位置的块,以达到更好的编码效果,它的作用就是将“块”与“块”之间的位置关系融入当前正在处理的块中。首先为每个块创建三个向量,即查询向量 Q ,键向量 K ,值向量 V ,这些向量通过块特征和三个参数矩阵(W_i^Q, W_i^K, W_i^V)分别相乘得到,当需要确定一个“块”的自注意力权重时,就需要根据这个块,对每个块进行评分,这个分数决定了该“块”对于其他“块”的关注度,用该“块”的查询向量 Q 去乘以每个块对应的键向量 K ,作为得分,将得分除以键向量 K 维度的平方根 $\sqrt{d_k}$,这样可以有更稳定的梯度,再传递给最大似然函数来将分数标准化,这样保证得分加起来为1,用这个得分乘以每个值向量 V ,用于保持想要关注的块的值保持不变,并通过乘以小数字,来淹没不相关的块,从而实现了一个块的自注意力输出,而多头注意力机制,可以提升注意力层的性能,它使得模型可以关注不同位置,这个机制为注意层提供了多个“表示子空间”。经过多头注意力机制,可以得到和头(heads)数目一样多的查询向量 Q ,键向量 K ,值向量 V 权重矩阵组,每个“块”会进行和多次上述的自注意力计算,计算的次数和头数目相同,每个块会得到多个不同的加权求和结果矩阵,并将这多个矩阵压缩成一个矩阵,即

先将8个矩阵连接起来,然后乘一个额外的权重矩阵 W^0 ,这里的权重矩阵 W^0 也属于参数矩阵,不同的注意力头可以允许模型在不同的表示子空间里学习到相关的信息,在其他实施方式中,还以考虑到区域与区域的相关性来对区域的特征进行更新。在对块特征进行更新后,通过视觉注意力机制确定各“块”的权重,并对各“块”进行加权平均处理得到对应 v_i 区域的区域位置特征。

[0195] 参见图7,图7是本发明实施例提供的基于人工智能的图文匹配方法的应用流程图,训练好的图文匹配模型可以给新闻客户端提供封面图的选图和配图服务,基于输入的新闻标题和标签信息,该服务会判断当前新闻内部是否包含足够多的图像,如果包含足够多的图像,则调用上述图文匹配模型,对新闻内图像根据与标题、标签的相关性从高到低进行排序,如果当前新闻内图像数量少或者相关性均特别低,则会从自主构建的预设图库(目前包含3千万张图像的数据集)里选择与标签或者标题相关的图像,而后调用上述图文匹配模型,获取最相关的预设数目的图像进行输出。

[0196] 在对上述图文匹配模型进行测试时,将该图文匹配模型与相关技术中的模型,在权威公开数据集以及收集的新闻数据集上进行了对比。

[0197] 公开数据集总共31769张图像,每张图像包含5个文本描述,从中随机选择29769张图像训练模型、1000张图像作为验证集、1000张作为测试集。另一个公开数据集提供了119287张图像,每张图像也包含5个句子描述,其中113287张图像作为训练集,1000张作为验证集,5000张作为测试集。

模型种类	图像生成文本			文本搜索图像		
	R@1	R@5	R@10	R@1	R@5	R@10
其他模型 1	42.5	71.9	81.5	30.2	60.4	72.3
其他模型 2	49.8	67.5	-	36.0	55.6	-
其他模型 3	55.0	81.8	89.0	39.4	69.2	79.1
其他模型 4	52.9	-	87.2	39.6	-	79.5
其他模型 5	55.6	81.9	89.5	39.1	69.2	80.9
其他模型 6	55.5	82.0	89.3	41.1	70.5	80.1
其他模型 7	67.4	90.3	95.8	48.6	77.7	85.2
其他模型 8	70.0	91.8	95.0	50.4	78.7	86.1
本发明实施例提供的图文匹配模型	72.5	93.0	96.0	55.1	81.7	87.6

[0199] 表1本发明实施例提供图文匹配模型的其他模型在公开数据集的测试结果

[0200] 上述表1展示了在本发明实施例提供图文匹配模型的其他模型在公开数据集的测试结果,图文匹配模型在给定图像检索文本任务上的第1名召回率,比其他模型中最好的测试结果高2.5%,在给定文本检索图像任务上,图文匹配模型的第1名召回率也比其他模型中最好的测试结果高4.7%(55.1%减50.4%)。同样,下述表2给出了这些模型在另一个公开数据集上的表现。结论和在上述数据集的结论类似。

[0201]

模型种类	图像生成文本			文本搜索图像		
	R@1	R@5	R@10	R@1	R@5	R@10
其他模型 1	53.2	83.1	91.5	40.7	75.8	87.4
其他模型 2	55.8	75.2	-	39.7	66.3	-
其他模型 3	55.0	81.8	89.0	39.4	69.2	79.1
其他模型 4	64.6	-	95.7	52.0	-	92.0
其他模型 5	65.6	89.8	95.5	47.1	79.9	90.0
其他模型 6	69.9	92.9	97.5	56.7	87.5	94.8
其他模型 7	72.7	94.8	98.4	58.8	88.4	94.8
其他模型 8	76.5	96.3	99.0	61.6	89.6	95.2
本发明实施例提供的图文匹配模型	79.2	97.5	98.8	63.1	89.8	95.2

[0202] 表2本发明实施例提供图文匹配模型和其他模型在公开数据集的测试结果

[0203] 本发明实施例提供图文匹配模型在中文数据集上也有非常不错的表现,在新闻数据集中收集了143317个新闻图像-新闻标题文本对,其中1000个用作验证集,另外使用人工标注的510个测试数据用于测试模型在真实应用场景下的效果,510个测试集是这样构建的:在科技、财经频道上各选择了255条新闻,每条新闻选择至少5张图像,里面包含至少1张正确图像和至少1张错误图像,模型对这些图像进行排序,计算前K名的正确率来评价模型的好坏。下述表3展示了不同模型在该数据集上的表现。

[0204]

METHODS	MAP@1	MAP@2	MAP@3	A@1	A@2	A@3
其他模型 7	67.2	70.6	75.7	67.2	69.1	73.6
其他模型 8	76.0	79.0	82.0	76.0	76.3	79.7
本发明实施例提供的图文匹配模型	77.6	80.3	82.7	77.6	77.9	80.9

[0205] 表3不同模型在新闻数据集上的测试结果

[0206] 其中,A@K表示前K名的正确率,MAP@K表示从前1名至前K名的平均正确率,如表3所示,本发明实施例提供图文匹配模型在新闻数据集上的表现依然好于目前最好的模型。综合表1-3的实验结果,可以看到,本发明实施例提供图文匹配模型不仅在两组公开数据集上达到了最佳表现程度,在实际的新闻中文数据集上也达到了最好的表现。

[0207] 本发明实施例提供的图文匹配模型相对于其他模型具备三个方面的技术改进,第一个方面在于自适应地学习位置与位置间的相互关系,提升位置特征与视觉特征的代表能力;第二个方面在于使用多目标学习策略,将两个学习任务融合,训练一个模型的效果好于多个模型融合的效果;第三个方面在于考虑到不同区域间拥有顺序关系,使用图像和文本的全局特征信息。针对这三点创新在公开数据集上进行了分析,如表4所示,可以看到图文匹配模型针对创新点1至3相对于其他模型均有明显的提升,T2I代表根据文本搜索图像,I2T代表根据图像生成文本。

[0208]

其他模型 7		其他模型 8		创新点 1		创新点 2		创新点 3		本发明实施例提供的图文匹配模型	
T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T
48.6	67.4	50.4	70.0	51.7	69.2	53.1	70.5	52.7	70.1	55.1	72.5

[0209] 表4图文匹配模型针对改进点1-3相对于其他模型的测试结果

[0210] 图8A-8B是本发明实施例提供的图文匹配模型在以文搜图任务上的效果,比如图8A中,对给定文本为“A girl is in a field surrounded by trees and pushing a pink scooter on the grass.”(“一个女孩在树木环绕的区域并在草地上推着一辆粉红色的滑板车”)进行查询得到如图8A所示的查询结果,本发明实施例提供的图文匹配模型可以将正确图像排在第1名,而其他模型将一个非常相似但错误的图像排在了第1名。比如图8B中,对给定文本为“Five people standing in front of a body of water”(“五个人站在水前”)进行查询得到如图8B所示的查询结果,本发明实施例提供的图文匹配模型可以将正确图像排在第1名,而其他模型将一个非常相似但错误的图像排在了第1名。

[0211] 同样,本发明实施例提供的图文匹配模型在新闻的中文数据集上结果也是类似。图9是本发明实施例提供的图文匹配模型和其他模型在新闻中文数据集上针对以文搜图任务的效果对比示意图,如图9所示,给定新闻文本为“一文打尽所有爆料,关于最新款Mac、ipad的信息都在这里了(catch all stuff in one article,all the information about the latest Mac,ipad is”,本发明实施例提供的图文匹配模型可以将正确图像排在第1名,而其他模型将一个非常相似但错误的图像排在了第1名,本发明实施例提供的图文匹配模型在中文数据集表现依然是最好的,说明其拥有很好的实用价值。

[0212] 图10是本发明实施例提供的图文匹配模型在短视频中的效果示意图,参见图10,本发明实施例提供的图文匹配模型不仅能够应用于新闻的选图和配图业务中,还可以将新闻内容转化成短视频来展示,考虑到用户的消费时间往往有限,而新闻作为长文本信息导致用户在固定的时长下很难消费过多的新闻,为了提升用户体验,通过将新闻内容转化成3-5秒的短视频进行展示,可以缓解这类问题,首先根据新闻内容提取关键语句作为新闻摘要信息,然后对这些关键句,使用本发明实施例提供的图文匹配方法,在新闻内的图像中选择与关键句最相关的图像作为视频的一帧,将多个关键句对应的多张图像拼接在一起生成短视频。参见图10,将对应文本划分为了不同的子文本,根据给定的子文本,获取与子文本匹配的候选图像,将候选图像拼接在一起生成了短视频。

[0213] 下面继续说明本发明实施例提供的基于人工智能的图文匹配装置255的实施为软件模块的示例性结构,在一些实施例中,如图2所示,存储在存储器250的基于人工智能的图文匹配装置255中的软件模块可以包括:文本特征获取模块2551,用于获取对应给定文本中各个词的语义的词特征,并基于各个词特征,合成对应给定文本的全局特征;图像特征获取模块2552,用于获取候选图像中各个区域的表征区域内位置关系的区域位置特征;获取候选图像中各个区域的区域视觉特征;结合区域视觉特征和区域位置特征以形成对应候选图像中各个区域的区域特征,并获取对应候选图像的全局特征;图文匹配模块2553,用于基于词特征、对应给定文本的全局特征、区域特征以及对应候选图像的全局特征,确定各个候选图像与给定文本的相似度评分,并在相似度评分大于相似度评分阈值时确定候选图像与给定文本匹配。

[0214] 在一些实施例中,基于人工智能的图文匹配装置还包括:预设图像获取模块2554,用于:当存在以下情形至少之一时,获取预设候选图像数据库中与给定文本相关的多个预设候选图像,以作为多个候选图像;待匹配对象中不存在候选图像,待匹配对象中的所有候选图像与给定文本的相似度评分均小于相似度评分阈值,或者匹配对象中的候选图像的数目小于候选图像数目阈值。

[0215] 在一些实施例中,文本特征获取模块2551,还用于:对给定文本进行分词处理,得到多个对应给定文本的词,并确定分别对应各个词的词向量;基于对应各个词的词向量,获取各个词的前向词向量以及后向词向量,以确定各个词在给定文本中的词特征。

[0216] 在一些实施例中,图像特征获取模块2552,还用于:对候选图像进行均匀分割,得到多个候选图像块;从候选图像中检测出至少一个表征候选图像的语义的区域,并获取对应区域的区域视觉特征;针对多个区域的每一个区域执行以下处理:根据各个候选图像块与区域的重叠像素数目,获取重叠像素数目最多的预设数目的候选图像块,以作为对应区域的区域候选图像块;获取区域候选图像块的位置索引特征,并对位置索引特征进行更新,得到表征区域候选图像块与区域内其他区域候选图像块之间的位置关系的位置特征;基于对应区域的区域视觉特征,对区域候选图像块的位置特征进行合成,得到表征区域内位置关系的区域位置特征。

[0217] 在一些实施例中,图像特征获取模块2552,还用于:针对多个区域候选图像块的每一个区域候选图像块执行以下处理:确定位置索引特征对应的区域候选图像块与区域内其他区域候选图像块至少两个维度的位置关系,以及对应各位置关系的权重;根据至少两个维度的位置关系,对位置索引特征进行基于至少两个维度的拆分;基于对应各位置关系的权重,对拆分后的位置索引特征进行加权处理,得到表征区域候选图像块与区域内其他区域候选图像块之间的位置关系的位置特征。

[0218] 在一些实施例中,图像特征获取模块2552,还用于:基于对应区域的区域视觉特征,获取各区域候选图像块基于视觉维度的权重;基于对应各区域候选图像块的权重,对各区域候选图像块的位置特征进行加权合成,得到表征区域内位置关系的区域位置特征。

[0219] 在一些实施例中,图像特征获取模块2552,还用于:从候选图像中检测出至少一个表征候选图像的语义的区域,并获取对应区域的区域独立视觉特征;针对多个区域的每一个区域执行以下处理:确定区域的区域独立视觉特征与候选图像内其他区域的视觉语义关系,以及对应各视觉语义关系的权重;基于对应各视觉语义关系的权重,对区域独立视觉特征进行加权处理,得到表征区域与候选图像内其他区域之间的视觉关系的区域视觉特征。

[0220] 在一些实施例中,图文匹配模块2553,还用于:基于给定文本中各个词的词特征以及候选图像中各个区域的区域特征,确定各个候选图像与给定文本的第一相似度评分,其中,所述第一相似度评分表征在区域维度以及词维度的相似程度;基于对应给定文本的全局特征以及对应候选图像的全局特征,确定各个候选图像与给定文本的第二相似度评分,其中,所述第二相似度评分表征在全局维度的相似程度;将第一相似度评分以及第二相似度评分结合对应的权重加权,得到各个候选图像与给定文本的相似度评分。

[0221] 在一些实施例中,基于人工智能的图文匹配装置255还包括:训练模块2555,用于:基于给定文本数据库以及候选图像数据库,构建由图文样本对组成的训练样本集合;其中,图文样本对包括具备匹配关系的给定文本样本和候选图像样本;基于图像特征提取模型以及文本特征提取模型,构建图文匹配模型;基于训练样本集合以及多目标学习任务,训练图文匹配模型;其中,多目标学习任务包括:基于给定文本样本对候选图像样本进行相似度评分以及基于候选图像样本对给定文本样本进行相似度评分的学习任务。

[0222] 在一些实施例中,训练模块2555,还用于:初始化图文匹配模型,并初始化包括输入样本、相似度评分、以及图文匹配模型的损失函数;在图文匹配模型每次迭代训练过程中

执行以下处理：将训练样本集合包括的所有图文样本对作为输入样本，通过图文匹配模型对图文样本对中的给定文本样本进行通向所有候选图像样本的相似度评分，并对图文样本对中的候选图像样本进行通向所有给定文本样本的相似度评分；将输入样本和相似度评分代入损失函数，以确定损失函数取得最小值时对应的图文匹配模型参数；根据所确定的图文匹配模型参数更新图文匹配模型。

[0223] 在一些实施例中，训练模块2555，还用于：确定给定文本样本的各个样本词的样本词特征以及给定文本样本的全局特征；针对所有候选图像样本中的每一个候选图像样本执行以下处理：确定候选图像样本中各个区域的区域特征以及候选图像样本的全局特征，并针对所有给定文本样本中的每一个样本词执行以下处理：确定候选图像样本中各个区域对应样本词的注意力权重，并根据各个区域的注意力权重以及各个区域的区域特征进行合成，得到样本词在区域上的特征；基于样本词在区域上的特征、样本词特征、给定文本样本的全局特征和候选图像样本的全局特征，确定基于给定文本样本进行通向候选图像样本的相似度评分。

[0224] 在一些实施例中，训练模块2555，还用于：确定各个样本词在区域上的特征和样本词特征的相似度的平均值，以作为表征区域维度以及词维度层面的各个候选图像样本与给定文本样本的第一相似度评分；根据给定文本样本的全局特征和候选图像样本的全局特征，分别确定给定文本样本与各个候选图像样本的第二相似度评分；将第一相似度评分以及第二相似度评分结合对应的权重加权，得到基于给定文本样本进行通向候选图像样本的相似度评分。

[0225] 在一些实施例中，训练模块2555，还用于：确定候选图像样本的各个候选图像区域的区域特征以及候选图像样本的全局特征；针对所有给定文本样本中的每一个给定文本样本执行以下处理：确定给定文本样本中各个样本词的样本词特征以及各个给定文本样本的全局特征，并针对所有候选图像样本中的每一个候选图像区域执行以下处理：确定给定文本样本中各个样本词对应候选图像区域的注意力权重，根据各个样本词的注意力权重以及各个样本词的样本词特征进行合成，得到候选图像区域在文本上的特征；基于候选图像区域在文本上的特征、区域特征、给定文本样本的全局特征和候选图像样本的全局特征，确定基于给定文本样本进行通向候选图像样本的相似度评分。

[0226] 在一些实施例中，训练模块2555，还用于：确定各个候选图像区域在文本上的特征和区域特征的相似度的平均值，以作为表征区域维度以及词维度层面的各个给定文本样本与各个候选图像样本的第一相似度评分；根据给定文本样本的全局特征和候选图像样本的全局特征，分别确定给定文本样本与各个候选图像样本的第二相似度评分；将第一相似度评分以及第二相似度评分结合对应的权重加权，得到基于候选图像样本进行通向所有给定文本样本的相似度评分。

[0227] 本发明实施例提供一种存储有可执行指令的存储介质，其中存储有可执行指令，当可执行指令被处理器执行时，将引起处理器执行本发明实施例提供的基于人工智能的图文匹配方法，例如，如图3和图4A-4D示出的方法。

[0228] 在一些实施例中，存储介质可以是FRAM、ROM、PROM、EPROM、EEPROM、闪存、磁表面存储器、光盘、或CD-ROM等存储器；也可以是包括上述存储器之一或任意组合的各种设备。

[0229] 在一些实施例中，可执行指令可以采用程序、软件、软件模块、脚本或代码的形式，

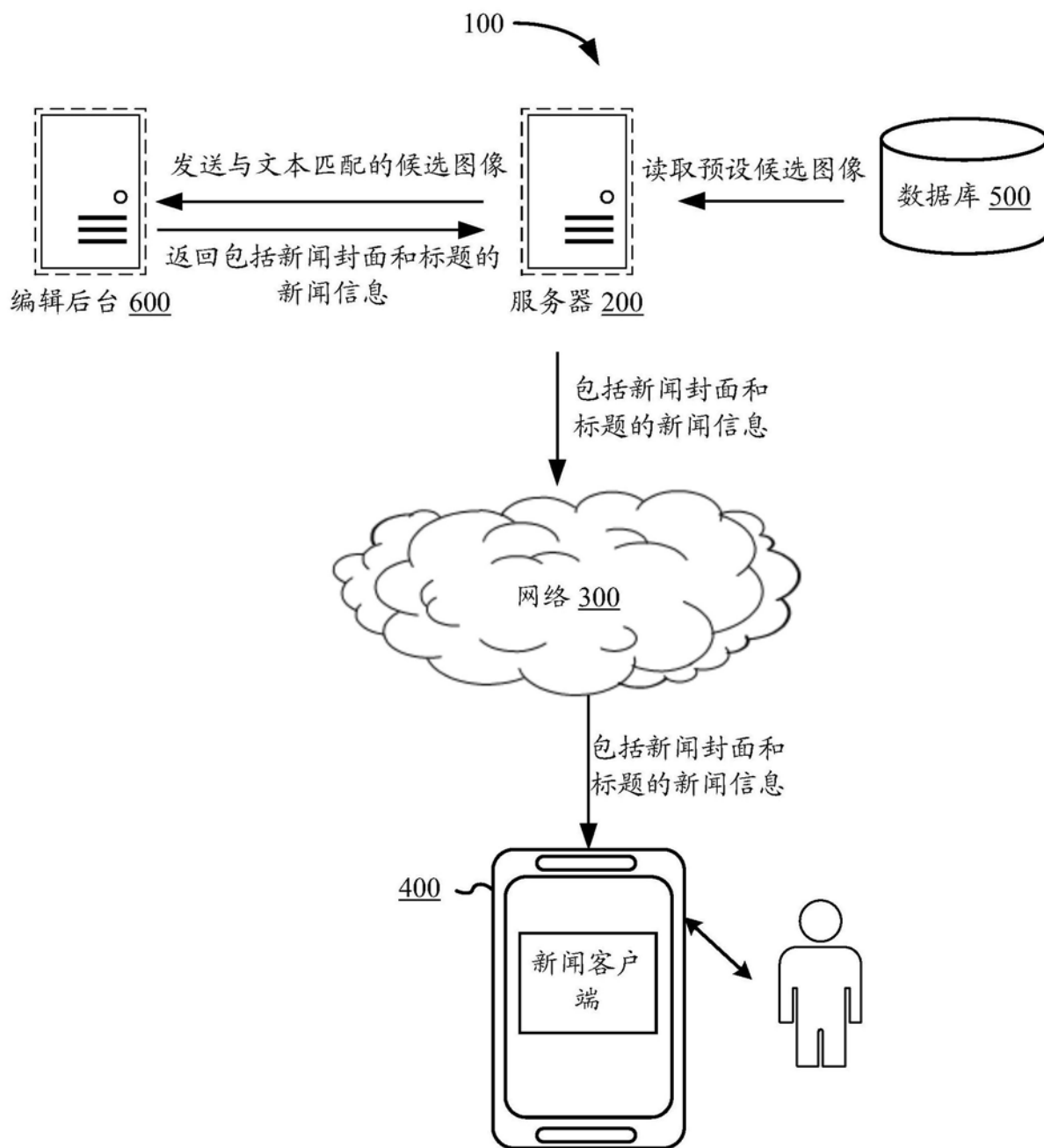
按任意形式的编程语言(包括编译或解释语言,或者声明性或过程性语言)来编写,并且其可按任意形式部署,包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

[0230] 作为示例,可执行指令可以但不一定对应于文件系统中的文件,可以可被存储在保存其它程序或数据的文件的一部分,例如,存储在超文本标记语言(HTML,Hyper Text Markup Language)文档中的一个或多个脚本中,存储在专用于所讨论的程序的单个文件中,或者,存储在多个协同文件(例如,存储一个或多个模块、子程序或代码部分的文件)中。

[0231] 作为示例,可执行指令可被部署为在一个计算设备上执行,或者在位于一个地点的多个计算设备上执行,又或者,在分布在多个地点且通过通信网络互连的多个计算设备上执行。

[0232] 综上所述,通过本发明实施例实现了将图像中目标区域内的位置关系考虑到图像特征中,并结合图像以及文本的全局特征,有效提高了图文匹配的准确率的技术效果。

[0233] 以上所述,仅为本发明的实施例而已,并非用于限定本发明的保护范围。凡在本发明的精神和范围之内所作的任何修改、等同替换和改进等,均包含在本发明的保护范围之内。



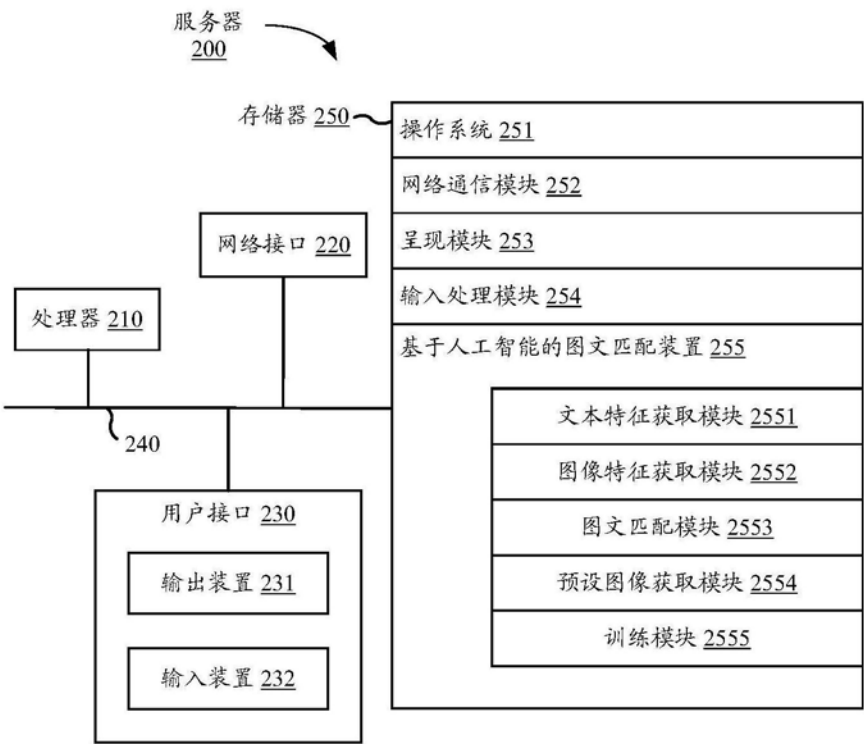


图2

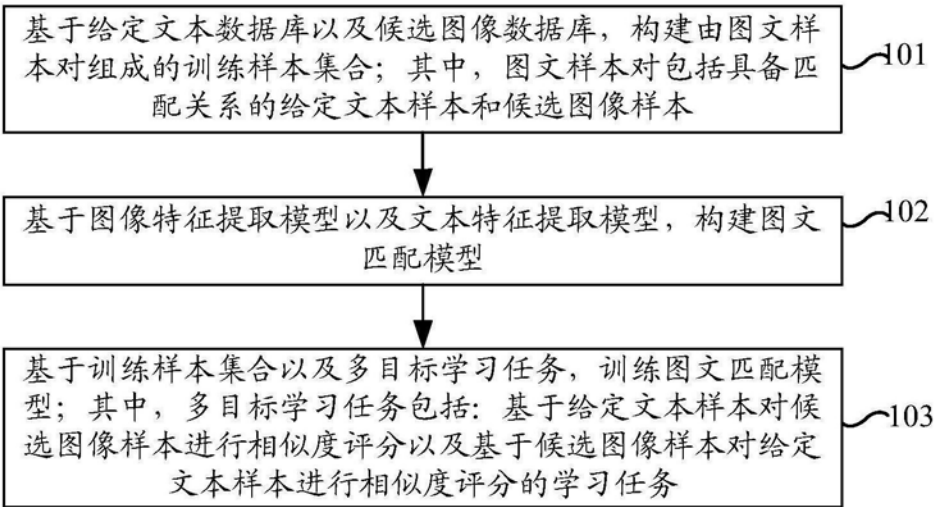


图3

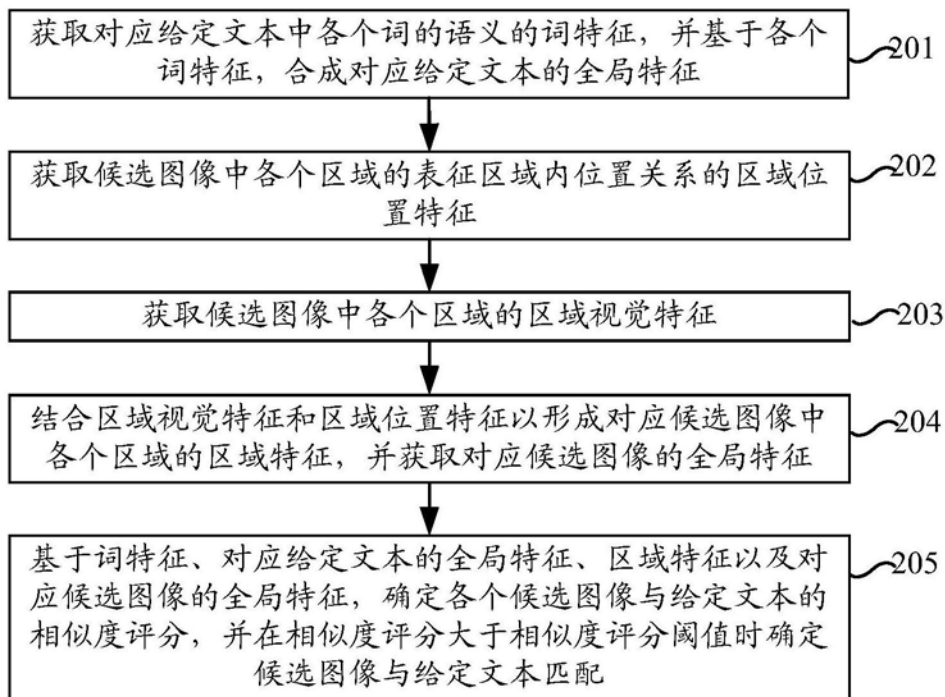


图4A

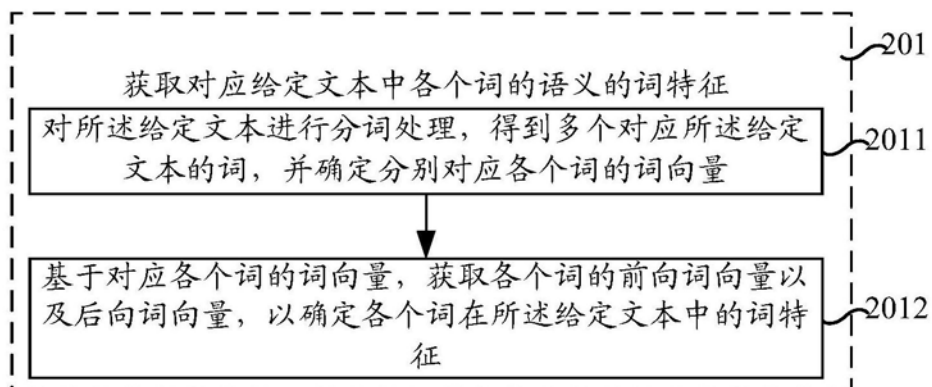


图4B

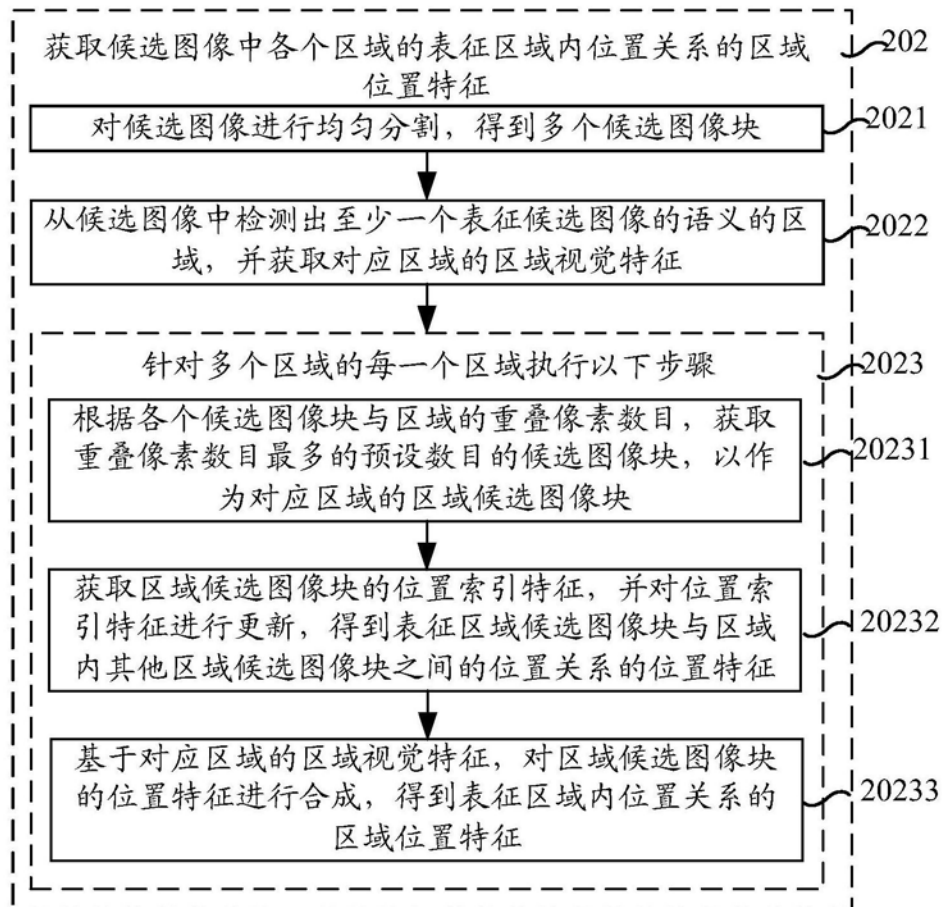


图4C

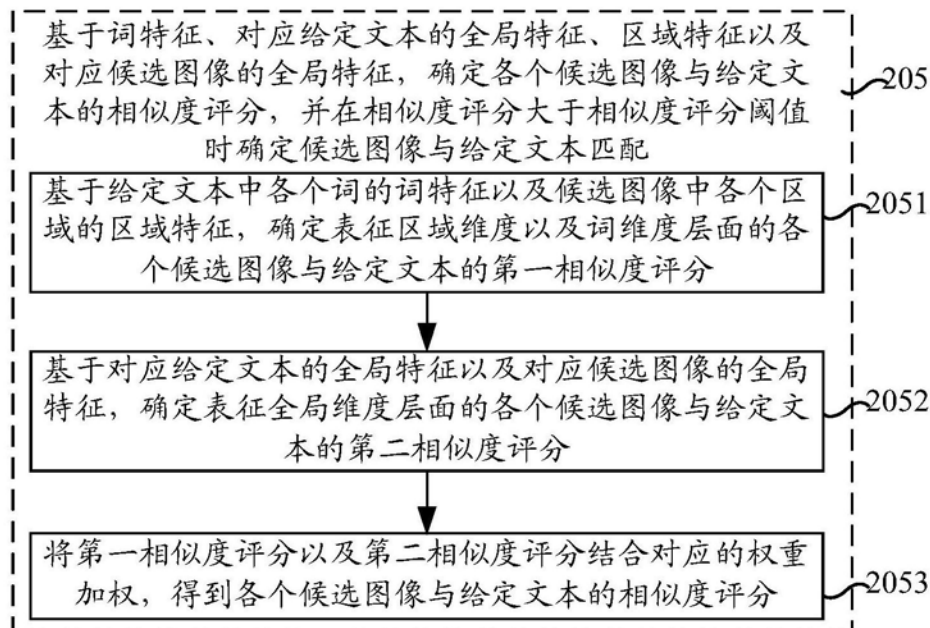


图4D

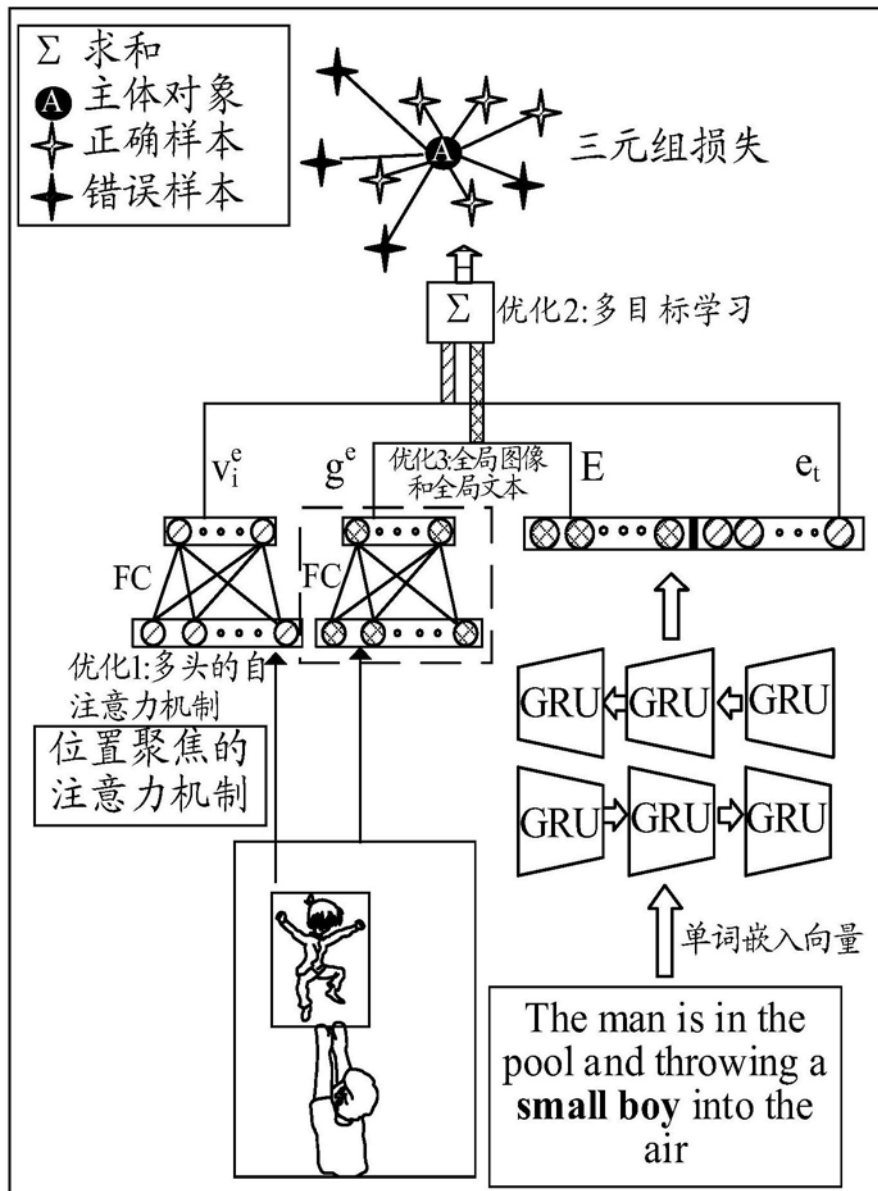


图5A

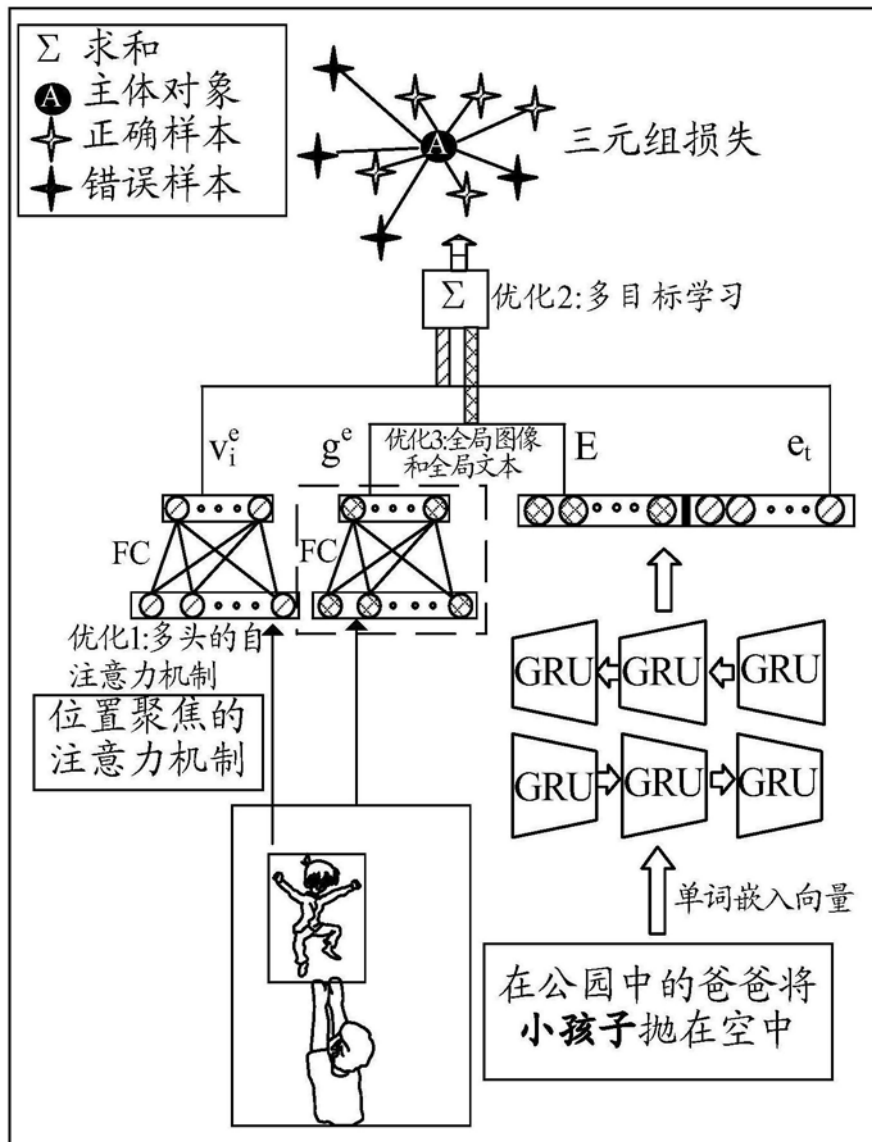


图5B

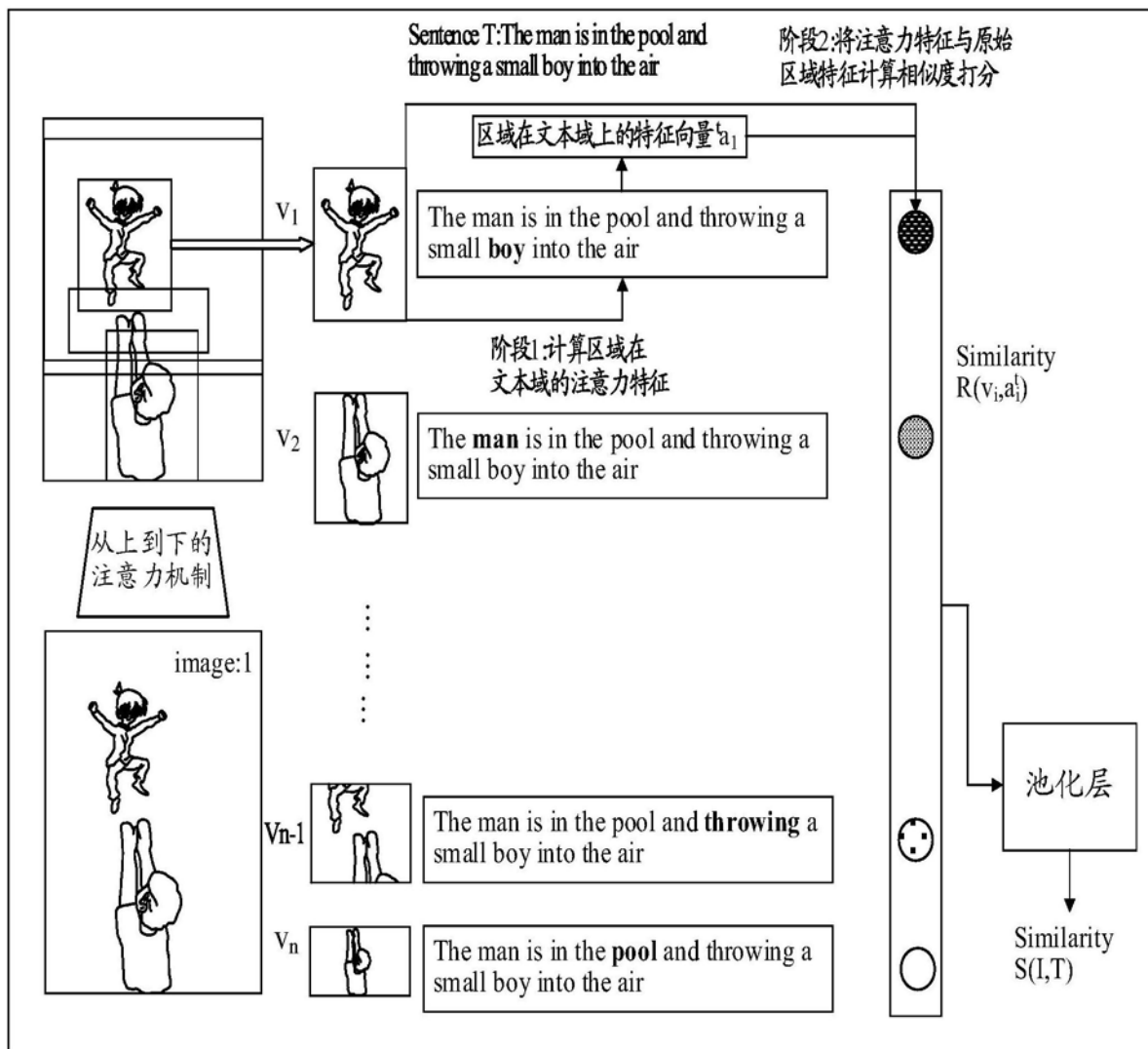


图6A

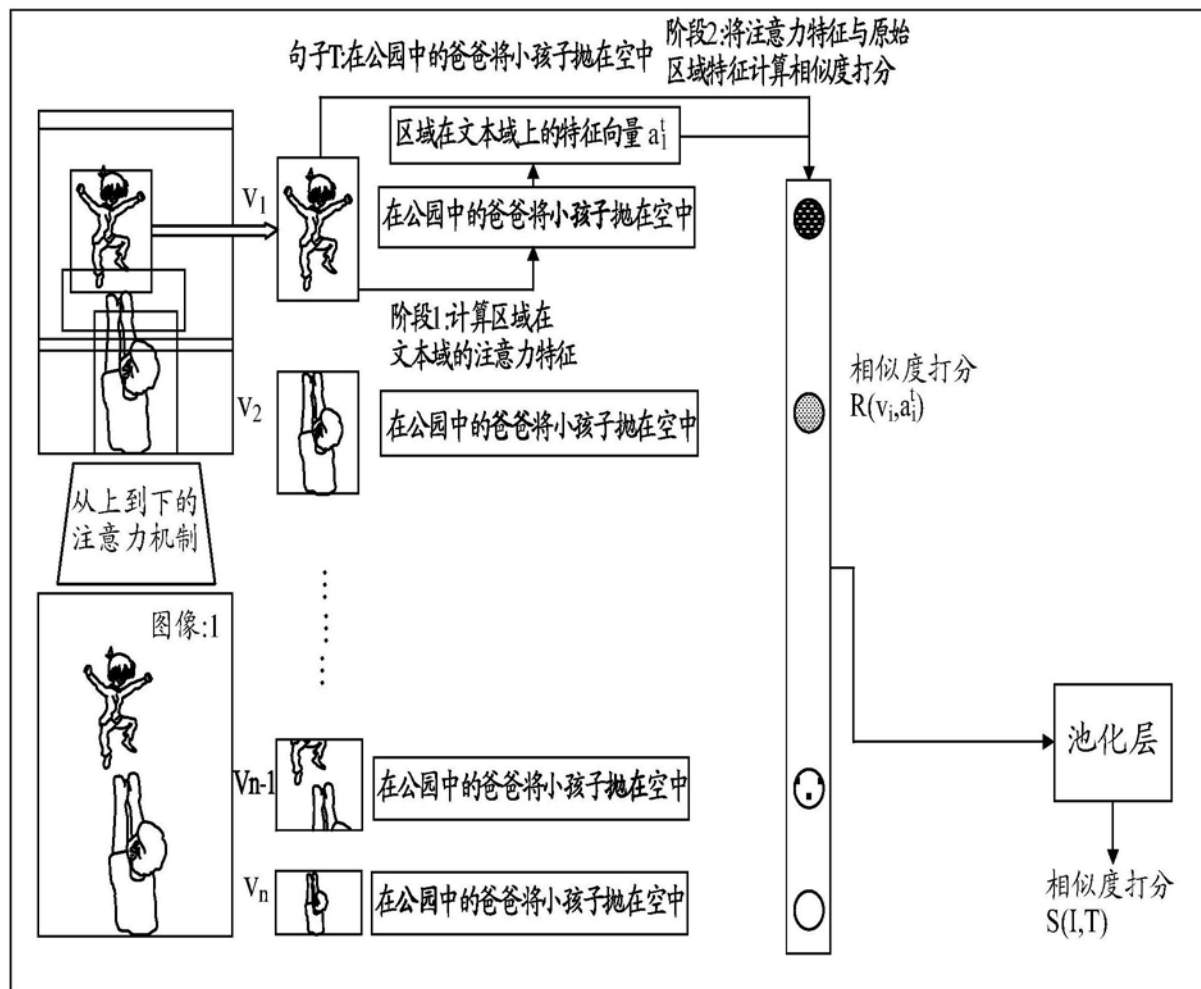


图6B

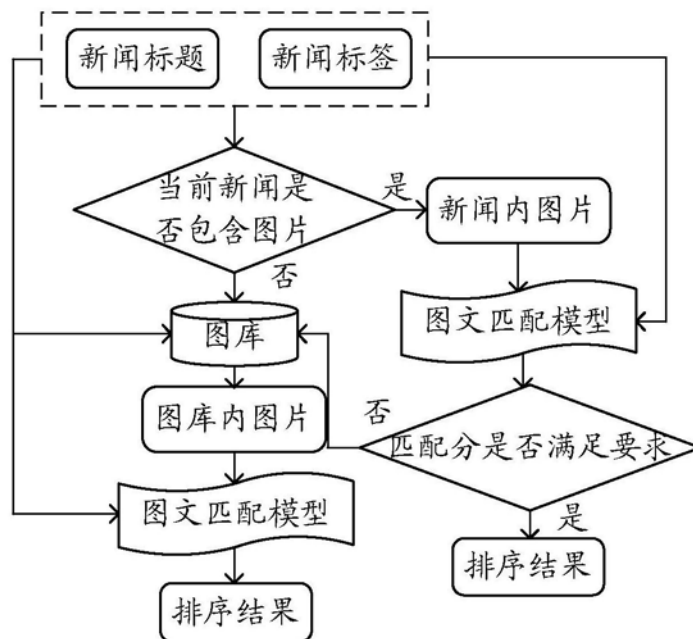
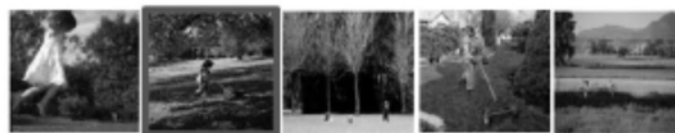


图7



本发明实例提供的图文匹配模型



其他模型

Result of query: "A girl is in a field surrounded by trees and pushing a pink scooter on the grass."

图8A



本发明实例提供的图文匹配模型



其他模型

Result of query: "Five people standing in front of a body of water."

图8B



本发明实例提供的图文匹配模型



其他模型

Result of query: "一文打尽所有爆料，关于最新款 Mac、ipad的信息都在这里了" (catch all stuff in one article, all the information about the latest Mac、ipad is

图9



图片中的孩子是一名唐氏综合征患儿，孩子自己一个人靠墙坐着，而拉布拉多注意到孩子低落的情绪，想要逗孩子开心起来，但是显然，孩子并不想理拉布拉多(The child in the picture is with Down syndrome, he sits alone against the wall, and Labrador notices the child's low mood, it wants to make the child happy, but obviously, the child does not want to play with Labrador)

图10