(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116824278 B
(45)授权公告日 2023.12.19

(21)申请号 202311097651.2

(22)申请日 2023.08.29

(65)同一申请的已公布的文献号
申请公布号 CN 116824278 A

(43)申请公布日 2023.09.29

(73)专利权人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 任玉强 鄢科

(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 贾允

(51)Int.Cl.
G06V 10/764(2022.01)
G06V 10/774(2022.01)
G06V 10/80(2022.01)
G06V 10/82(2022.01)
G06F 18/22(2023.01)
G06N 3/045(2023.01)
G06N 3/0464(2023.01)
G06N 3/096(2023.01)
G06N 3/084(2023.01)
G06N 3/0985(2023.01)

(56)对比文件
CN 115983227 A,2023.04.18
CN 116244418 A,2023.06.09
CN 114511043 A,2022.05.17
CN 115830610 A,2023.03.21
CN 116109866 A,2023.05.12
CN 116259075 A,2023.06.13
CN 116416480 A,2023.07.11
CN 116432026 A,2023.07.14
WO 2021191908 A1,2021.09.30
Rui Cao.Prompting for Multimodal Hateful Meme Classification.《arXiv》.2023,1-12.

审查员 梁滔

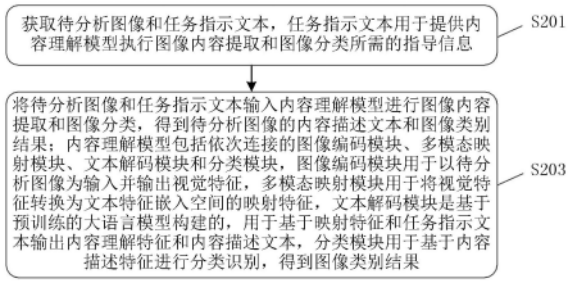
权利要求书5页 说明书16页 附图5页

(54)发明名称

图像内容分析方法、装置、设备和介质

(57)摘要

本申请提供了图像内容分析方法、装置、设备和介质,涉及人工智能技术领域,可以应用于云技术、人工智能、智慧交通、辅助驾驶等场景,方法包括:获取待分析图像和任务指示文本;将待分析图像和任务指示文本输入内容理解模型进行图像内容提取和图像分类,得到待分析图像的内容描述文本和图像类别结果;内容理解模型是结合图文对指令数据集对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的约束训练,以及结合多模态指令数据集对初始理解模型进行图像内容提取和图像分类的约束训练得到的。本申请能够显著提升模型能力和业务匹配性。



CN 116824278 B

1.一种图像内容分析方法,其特征在于,所述方法包括:

获取待分析图像和任务指示文本;

将所述待分析图像和所述任务指示文本输入内容理解模型进行图像内容提取和图像分类,得到所述待分析图像的内容描述文本和图像类别结果;所述任务指示文本用于提供所述内容理解模型执行图像内容提取和图像分类所需的指导信息;

所述内容理解模型包括依次连接的图像编码模块、多模态映射模块、文本解码模块和分类模块,所述图像编码模块用于以所述待分析图像为输入并输出视觉特征,所述多模态映射模块用于将所述视觉特征转换为文本特征嵌入空间的映射特征,所述文本解码模块是基于预训练的大语言模型构建的,用于基于所述映射特征和所述任务指示文本输出内容理解特征和所述内容描述文本,所述分类模块用于基于所述内容理解特征进行分类识别,得到所述图像类别结果;

所述内容理解模型是结合图文对指令数据集对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的约束训练,以及结合多模态指令数据集对所述初始理解模型进行图像内容提取和图像分类的约束训练得到的;

所述将所述待分析图像和所述任务指示文本输入内容理解模型进行图像内容提取和图像分类,得到所述待分析图像的内容描述文本和图像类别结果包括:

将所述待分析图像输入所述图像编码模块进行特征提取,得到所述视觉特征;

将所述视觉特征输入所述多模态映射模块进行特征映射,以将所述视觉特征映射至所述文本解码模块的文本特征嵌入空间,得到所述映射特征;

将所述映射特征和所述任务指示文本输入所述文本解码模块进行内容理解,得到所述内容理解特征和所述内容描述文本,所述内容描述文本是基于所述文本解码模块的输出层对所述内容理解特征进行特征文本映射得到的;

将所述内容理解特征输入所述分类模块进行分类识别,得到所述图像类别结果。

2.根据权利要求1所述的方法,其特征在于,所述多模态映射模块包括第一转换器层和第二转换器层,所述将所述视觉特征输入所述多模态映射模块进行特征映射,以将所述视觉特征映射至所述文本解码模块的文本特征嵌入空间,得到所述映射特征包括:

将所述视觉特征输入所述第一转换器层进行分片级的特征表示,以将所述视觉特征映射至词嵌入空间,得到分片嵌入特征;

将所述分片嵌入特征输入所述第二转换器层进行上下文信息交叉提取,得到所述映射特征。

3.根据权利要求1所述的方法,其特征在于,所述图像编码模块是结合图文样本对,对预设的文本特征提取网络和基于自注意力机制的图像特征提取网络进行图像和文本匹配的分类识别约束训练得到的。

4.根据权利要求1-3中任一项所述的方法,其特征在于,所述方法还包括:

获取通用领域的图文对指令数据集、预设业务领域的图文对指令数据集和文本对话数据集;

基于所述通用领域的图文对指令数据集和所述预设业务领域的图文对指令数据集,对所述初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的第一约束训练,在训练过程中冻结所述文本解码模块的模型参数并调整所述图

像编码模块和所述多模态映射模块的模型参数,至满足第一训练结束条件;

基于由所述预设业务领域的图文对指令数据集和所述文本对话数据集形成的多模态指令数据集,对满足所述第一训练结束条件的初始理解模型进行图像内容提取和图像分类的第二约束训练,在训练过程中冻结所述图像编码模块和所述多模态映射模块的模型参数并调整所述文本解码模块和所述分类模块的模型参数,至满足第二训练结束条件;

将满足所述第二训练结束条件的初始理解模型确定为所述内容理解模型。

5.根据权利要求4所述的方法,其特征在于,所述通用领域对应的图文对指令数据集包括第一样本图像、第一指示信息和所述第一样本图像对应的第一文本标注,所述预设业务领域对应的图文对指令数据集包括第二样本图像、第二指示信息和所述第二样本图像对应的第二文本标注,所述第一指示信息和所述第二指示信息均用于提供所述文本解码模块在进行内容理解时所需的指导信息,所述第一文本标注为基于所述第一指示信息进行图像描述所对应的文本响应真值,所述第二文本标注为基于所述第二指示信息进行图像描述所对应的文本响应真值;所述第一约束训练采用下述步骤实现:

以所述第一样本图像或所述第二样本图像作为所述图像编码模块的输入进行特征提取,得到第一样本图像对应的第一样本视觉特征或所述第二样本图像对应的第二样本视觉特征;

以所述第一样本视觉特征或所述第二样本视觉特征作为所述多模态映射模块的输入进行特征映射,以分别将所述第一样本视觉特征和所述第二样本视觉特征映射至所述文本解码模块的文本特征嵌入空间,得到第一映射特征和第二映射特征;

以所述第一映射特征和所述第一指示信息形成的数据对,或,所述第二映射特征和所述第二指示信息形成的数据对作为所述文本解码模块的输入进行内容理解,得到第一样本描述文本和第二样本描述文本;

基于所述第一样本描述文本和所述第一文本标注间的差异,以及所述第二样本描述文本和所述第二文本标注间的差异确定第一模型损失;

冻结所述文本解码模块的模型参数,根据所述第一模型损失训练所述图像编码模块和所述多模态映射模块,至满足所述第一训练结束条件。

6.根据权利要求5所述的方法,其特征在于,所述预设业务领域的图文对指令数据集还包括所述第二样本图像的类别标签,所述文本对话数据集包括样本指令文本和所述样本指令文本对应的答案标注,所述第二约束训练采用下述步骤实现包括:

以所述第二样本图像作为满足所述第一训练结束条件的图像编码模块的输入进行特征提取,得到第三样本视觉特征;

以所述第三样本视觉特征作为满足所述第一训练结束条件的多模态映射模块的输入进行特征映射,得到第三映射特征;

以所述第二映射特征和所述第二指示信息形成的数据对,或,所述样本指令文本作为所述文本解码模块的输入进行内容理解,得到第三样本理解特征、第三样本描述文本和答案文本;

将所述第三样本理解特征输入所述分类模块进行分类识别,得到样本分类结果;

基于所述第三样本描述文本、所述第二文本标注、所述答案文本、所述答案标注、所述样本分类结果和所述类别标签确定第二模型损失;

冻结所述图像编码模块和所述多模态映射模块的模型参数,根据所述第二模型损失训练所述文本解码模块和所述分类模块,至满足所述第二训练结束条件。

7.根据权利要求5所述的方法,其特征在于,所述第一指示信息和所述第二指示信息均包括多维度的指令文本,所述多维度的指令文本包括对象描述指令、内容属性描述指令和内容推理指令;

所述对象描述指令用于指示所述文本解码模块进行图像对象描述,所述对象描述指令对应的第一文本标注或第二文本标注具有唯一真值;所述内容属性描述指令用于指示所述文本解码模块进行图像整体信息描述,所述内容属性描述指令对应的第一文本标注或第二文本标注具有开放式答案属性;所述内容推理指令用于指示所述文本解码模块进行图像内容的推理解答。

8.一种图像内容分析装置,其特征在于,所述装置包括:

获取模块:获取待分析图像和任务指示文本;

内容分析模块:用于将所述待分析图像和所述任务指示文本输入内容理解模型进行图像内容提取和图像分类,得到所述待分析图像的内容描述文本和图像类别结果;所述任务指示文本用于提供所述内容理解模型执行图像内容提取和图像分类所需的指导信息;

所述内容理解模型包括依次连接的图像编码模块、多模态映射模块、文本解码模块和分类模块,所述图像编码模块用于以所述待分析图像为输入并输出视觉特征,所述多模态映射模块用于将所述视觉特征转换为文本特征嵌入空间的映射特征,所述文本解码模块是基于预训练的大语言模型构建的,用于基于所述映射特征和所述任务指示文本输出内容理解特征和所述内容描述文本,所述分类模块用于基于所述内容理解特征进行分类识别,得到所述图像类别结果;

所述内容分析模块包括:

特征提取子模块:用于将所述待分析图像输入所述图像编码模块进行特征提取,得到所述视觉特征;

特征映射子模块:用于将所述视觉特征输入所述多模态映射模块进行特征映射,以将所述视觉特征映射至所述文本解码模块的文本特征嵌入空间,得到所述映射特征;

内容理解子模块:用于将所述映射特征和所述任务指示文本输入所述文本解码模块进行内容理解,得到所述内容理解特征和所述内容描述文本,所述内容描述文本是基于所述文本解码模块的输出层对所述内容理解特征进行特征文本映射得到的;

分类识别子模块:用于将所述内容理解特征输入所述分类模块进行分类识别,得到所述图像类别结果;

所述内容理解模型是结合图文对指令数据集对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的约束训练,以及结合多模态指令数据集对所述初始理解模型进行图像内容提取和图像分类的约束训练得到的。

9.根据权利要求8所述的装置,其特征在于,所述多模态映射模块包括第一转换器层和第二转换器层,所述特征映射子模块包括:

特征表示单元:用于将所述视觉特征输入所述第一转换器层进行分片级的特征表示,以将所述视觉特征映射至词嵌入空间,得到分片嵌入特征;

交叉提取单元:用于将所述分片嵌入特征输入所述第二转换器层进行上下文信息交叉

提取,得到所述映射特征。

10.根据权利要求8所述的装置,其特征在于,所述图像编码模块是结合图文样本对,对预设的文本特征提取网络和基于自注意力机制的图像特征提取网络进行图像和文本匹配的分类识别约束训练得到的。

11.根据权利要求8-10中任一项所述的装置,其特征在于,所述装置还包括:

数据集获取模块:用于获取通用领域的图文对指令数据集、预设业务领域的图文对指令数据集和文本对话数据集;

第一训练模块:用于基于所述通用领域的图文对指令数据集和所述预设业务领域的图文对指令数据集,对所述初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的第一约束训练,在训练过程中冻结所述文本解码模块的模型参数并调整所述图像编码模块和所述多模态映射模块的模型参数,至满足第一训练结束条件;

第二训练模块:用于基于由所述预设业务领域的图文对指令数据集和所述文本对话数据集形成的多模态指令数据集,对满足所述第一训练结束条件的初始理解模型进行图像内容提取和图像分类的第二约束训练,在训练过程中冻结所述图像编码模块和所述多模态映射模块的模型参数并调整所述文本解码模块和所述分类模块的模型参数,至满足第二训练结束条件;

模型确定模块:用于将满足所述第二训练结束条件的初始理解模型确定为所述内容理解模型。

12.根据权利要求11所述的装置,其特征在于,所述通用领域对应的图文对指令数据集包括第一样本图像、第一指示信息和所述第一样本图像对应的第一文本标注,所述预设业务领域对应的图文对指令数据集包括第二样本图像、第二指示信息和所述第二样本图像对应的第二文本标注,所述第一指示信息和所述第二指示信息均用于提供所述文本解码模块在进行内容理解时所需的指导信息,所述第一文本标注为基于所述第一指示信息进行图像描述所对应的文本响应真值,所述第二文本标注为基于所述第二指示信息进行图像描述所对应的文本响应真值;所述第一训练模块包括:

第一样本提取子模块:用于以所述第一样本图像或所述第二样本图像作为所述图像编码模块的输入进行特征提取,得到第一样本图像对应的第一样本视觉特征或所述第二样本图像对应的第二样本视觉特征;

以所述第一样本视觉特征或所述第二样本视觉特征作为所述多模态映射模块的输入进行特征映射,以分别将所述第一样本视觉特征和所述第二样本视觉特征映射至所述文本解码模块的文本特征嵌入空间,得到第一映射特征和第二映射特征;

第一样本映射子模块:用于以所述第一映射特征和所述第一指示信息形成的数据对,或,所述第二映射特征和所述第二指示信息形成的数据对作为所述文本解码模块的输入进行内容理解,得到第一样本描述文本和第二样本描述文本;

第一损失子模块:用于基于所述第一样本描述文本和所述第一文本标注间的差异,以及所述第二样本描述文本和所述第二文本标注间的差异确定第一模型损失;

第一训练子模块:用于冻结所述文本解码模块的模型参数,根据所述第一模型损失训练所述图像编码模块和所述多模态映射模块,至满足所述第一训练结束条件。

13.根据权利要求12所述的装置,其特征在于,所述预设业务领域的图文对指令数据集还包括所述第二样本图像的类别标签,所述文本对话数据集包括样本指令文本和所述样本指令文本对应的答案标注,所第二训练模块包括:

第二样本提取子模块:用于以所述第二样本图像作为满足所述第一训练结束条件的图像编码模块的输入进行特征提取,得到第三样本视觉特征;

第二样本映射子模块:用于以所述第三样本视觉特征作为满足所述第一训练结束条件的多模态映射模块的输入进行特征映射,得到第三映射特征;

第二样本理解子模块:用于以所述第二映射特征和所述第二指示信息形成的数据对,或,所述样本指令文本作为所述文本解码模块的输入进行内容理解,得到第三样本理解特征、第三样本描述文本和答案文本;

样本分类子模块:用于将所述第三样本理解特征输入所述分类模块进行分类识别,得到样本分类结果;

第二损失子模块:用于基于所述第三样本描述文本、所述第二文本标注、所述答案文本、所述答案标注、所述样本分类结果和所述类别标签确定第二模型损失;

第二训练子模块:用于冻结所述图像编码模块和所述多模态映射模块的模型参数,根据所述第二模型损失训练所述文本解码模块和所述分类模块,至满足所述第二训练结束条件。

14.根据权利要求12所述的装置,其特征在于,所述第一指示信息和所述第二指示信息均包括多维度的指令文本,所述多维度的指令文本包括对象描述指令、内容属性描述指令和内容推理指令;

所述对象描述指令用于指示所述文本解码模块进行图像对象描述,所述对象描述指令对应的第一文本标注或第二文本标注具有唯一真值;所述内容属性描述指令用于指示所述文本解码模块进行图像整体信息描述,所述内容属性描述指令对应的第一文本标注或第二文本标注具有开放式答案属性;所述内容推理指令用于指示所述文本解码模块进行图像内容的推理解答。

15.一种计算机可读存储介质,其特征在于,所述存储介质中存储有至少一条指令或至少一段程序,所述至少一条指令或所述至少一段程序由处理器加载并执行以实现如权利要求1-7中任一项所述的图像内容分析方法。

16.一种计算机设备,其特征在于,所述设备包括处理器和存储器,所述存储器中存储有至少一条指令或至少一段程序,所述至少一条指令或所述至少一段程序由所述处理器加载并执行以实现如权利要求1-7中任一项所述的图像内容分析方法。

图像内容分析方法、装置、设备和介质

技术领域

[0001] 本申请涉及人工智能技术领域,尤其涉及一种图像内容分析方法、装置、设备和介质。

背景技术

[0002] 图像内容分析和理解是一项重要的业务应用,主要利用人工智能技术来对图像内容进行分析,输出业务所需要的有效信息。长期以来,受限于深度学习技术的发展,现有的内容分析理解技术主要基于特定的多项AI技术配合业务后处理逻辑实现。比如判断一张图片是否含有特定内容,一般需要多模型输出才能够理解识别:人脸识别模型检测图片中是否含有特定人物,元素检测模型检测图片中是否含有特定元素,事件分析模型再判断该图像表达的内容是否涉及特定事件,特定内容识别模型判断图像是否含有特定内容,当上述独立的模型都没有检测到特定内容,才能够确定该图片属于正常类别。该方式需整合多种模型,图像内容分析效率很低,同时需进行独立的模型训练,并且依赖于人工设置的先验知识,训练成本高。

发明内容

[0003] 本申请提供了一种图像内容分析方法、装置、设备和介质,可以显著提升图像内容分析的准确性和分析效率。

[0004] 一方面,本申请提供了一种图像内容分析方法,所述方法包括:

[0005] 获取待分析图像和任务指示文本;

[0006] 将所述待分析图像和所述任务指示文本输入内容理解模型进行图像内容提取和图像分类,得到所述待分析图像的内容描述文本和图像类别结果;所述任务指示文本用于提供所述内容理解模型执行所述图像内容提取和图像分类所需的指导信息;

[0007] 所述内容理解模型包括依次连接的图像编码模块、多模态映射模块、文本解码模块和分类模块,所述图像编码模块用于以所述待分析图像为输入并输出视觉特征,所述多模态映射模块用于将所述视觉特征转换为文本特征嵌入空间的映射特征,所述文本解码模块是基于预训练的大语言模型构建的,用于基于所述映射特征和所述任务指示文本输出内容理解特征和所述内容描述文本,所述分类模块用于基于所述内容理解特征进行分类识别,得到所述图像类别结果;

[0008] 所述内容理解模型是结合图文对指令数据集对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的约束训练,以及结合多模态指令数据集对所述初始理解模型进行图像内容提取和图像分类的约束训练得到的。

[0009] 另一方面提供了一种图像内容分析装置,所述装置包括:

[0010] 获取模块:获取待分析图像和任务指示文本;

[0011] 内容分析模块:用于将所述待分析图像和所述任务指示文本输入内容理解模型进行图像内容提取和图像分类,得到所述待分析图像的内容描述文本和图像类别结果;所述

任务指示文本用于提供所述内容理解模型执行所述图像内容提取和图像分类所需的指导信息；

[0012] 所述内容理解模型包括依次连接的图像编码模块、多模态映射模块、文本解码模块和分类模块，所述图像编码模块用于以所述待分析图像为输入并输出视觉特征，所述多模态映射模块用于将所述视觉特征转换为文本特征嵌入空间的映射特征，所述文本解码模块是基于预训练的大语言模型构建的，用于基于所述映射特征和所述任务指示文本输出内容理解特征和所述内容描述文本，所述分类模块用于基于所述内容理解特征进行分类识别，得到所述图像类别结果；

[0013] 所述内容理解模型是结合图文对指令数据集对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的约束训练，以及结合多模态指令数据集对所述初始理解模型进行图像内容提取和图像分类的约束训练得到的。

[0014] 另一方面提供了一种计算机设备，所述设备包括处理器和存储器，所述存储器中存储有至少一条指令或至少一段程序，所述至少一条指令或所述至少一段程序由所述处理器加载并执行以实现如上述的图像内容分析方法。

[0015] 另一方面提供了一种计算机可读存储介质，所述存储介质中存储有至少一条指令或至少一段程序，所述至少一条指令或所述至少一段程序由处理器加载并执行以实现如上述的图像内容分析方法。

[0016] 另一方面提供了一种服务器，所述服务器包括处理器和存储器，所述存储器中存储有至少一条指令或至少一段程序，所述至少一条指令或所述至少一段程序由所述处理器加载并执行以实现如上述的图像内容分析方法。

[0017] 另一方面提供了一种终端，所述终端包括处理器和存储器，所述存储器中存储有至少一条指令或至少一段程序，所述至少一条指令或所述至少一段程序由所述处理器加载并执行以实现如上述的图像内容分析方法。

[0018] 另一方面提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令被处理器执行时实现如上述的图像内容分析方法。

[0019] 本申请提供的图像内容分析方法、装置、设备、存储介质、服务器、终端、计算机程序和计算机程序产品，具有如下技术效果：

[0020] 本申请的技术方案首先获取待分析图像和任务指示文本，将待分析图像和任务指示文本输入内容理解模型进行图像内容提取和图像分类，得到待分析图像的内容描述文本和图像类别结果，该任务指示文本用于提供内容理解模型执行图像内容提取和图像分类所需的指导信息，能够通过指令方式提升模型理解能力，进而提升内容描述和图像分类的准确性；其中，内容理解模型包括依次连接的图像编码模块、多模态映射模块、文本解码模块和分类模块，图像编码模块用于以待分析图像为输入并输出视觉特征，多模态映射模块用于将视觉特征转换为文本特征嵌入空间的映射特征，文本解码模块是基于预训练的大语言模型构建的，用于基于映射特征和任务指示文本输出内容理解特征和内容描述文本，分类模块用于基于内容理解特征进行分类识别，得到图像类别结果；内容理解模型是结合图文对指令数据集对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的约束训练，以及结合多模态指令数据集对初始理解模型进行图像内容提取和图像分类的约束训练得到的；如此，结合大语言模型和指令方式将内容分析和内

容分类整合至同一模型中进行内容理解和分类识别,能够将内容理解中的多种任务进行统一,满足需结合内容分析的分类业务需求,输入是指示文本和图像,能够由单独的一个模型来综合分析图像中出现的所有元素和元素之间的关系,综合分析后输出图像分析的结果,相比于已有方法中多个模型单独分析再后处理的方式,流程更简洁高效,内容理解也更准确,显著提升分析效率和降低训练成本;此外,能够实现图像编码模块、多模态映射模块和分类层的分段训练,提升模型能力和训练收敛效率,且经指令学习迁移训练得到多模态模型,能够显著提升其图像分析能力。

附图说明

[0021] 为了更清楚地说明本申请实施例或现有技术中的技术方案和优点,下面将对实施例或现有技术描述中所需要使用的附图作简单的介绍,显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其它附图。

[0022] 图1是本申请实施例提供的一种应用环境的示意图;

[0023] 图2是本申请实施例提供的一种图像内容分析方法的流程示意图;

[0024] 图3是本申请实施例提供的另一种图像内容分析方法的流程示意图;

[0025] 图4是本申请实施例提供的另一种图像内容分析方法的流程示意图;

[0026] 图5是本申请实施例提供的一种内容理解模型的结构框架图;

[0027] 图6是本申请实施例提供的一种图像内容分析的原理框架图;

[0028] 图7是本申请实施例提供的一种图像内容分析装置的框架示意图;

[0029] 图8是本申请实施例提供的一种执行图像内容分析方法的电子设备的硬件结构框图。

具体实施方式

[0030] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例,都属于本申请保护的范围。

[0031] 需要说明的是,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数据在适当情况下可以互换,以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。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,例如,包含了一系列步骤或子模块的过程、方法、系统、产品或服务器不必限于清楚地列出的那些步骤或子模块,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步骤或子模块。

[0032] 对本申请实施例进行进一步详细说明之前,对本申请实施例中涉及的名词和术语进行说明,本申请实施例中涉及的名词和术语适用于如下的解释。

[0033] 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综合技术,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做出反应的智能机器。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。

[0034] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0035] 计算机视觉技术(Computer Vision, CV)计算机视觉是一门研究如何使机器“看”的科学,更进一步的说,就是指用摄影机和电脑代替人眼对目标进行识别、检测和测量等机器视觉,并进一步做图形处理,使电脑处理成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科,计算机视觉研究相关的理论和技术,试图建立能够从图像或者多维数据中获取信息的人工智能系统。计算机视觉技术通常包括图像处理、图像识别、图像语义理解、图像检索、OCR、视频处理、视频语义理解、视频内容/行为识别、三维物体重建、3D技术、虚拟现实、增强现实、同步定位与地图构建等技术,还包括常见的人脸识别、指纹识别等生物特征识别技术。

[0036] 自然语言处理(Nature Language processing, NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此,这一领域的研究将涉及自然语言,即人们日常使用的语言,所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[0037] 随着人工智能技术研究和进步,人工智能技术在多个领域展开研究和应用,例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服等,相信随着技术的发展,人工智能技术将在更多的领域得到应用,并发挥越来越重要的价值。本申请实施例提供的方案涉及人工智能的机器学习/深度学习、计算机视觉技术和自然语言处理等技术,具体通过如下实施例进行说明。

[0038] 请参阅图1,图1是本申请实施例提供的一种应用环境的示意图,如图1所示,该应用环境可以包括终端01和服务器02。在实际应用中,终端01和服务器02可以通过有线或无线通信方式进行直接或间接地连接,本申请在此不做限制。

[0039] 本申请实施例中的服务器02可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN(Content Delivery Network,内容分发网络)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0040] 具体地,云技术(Cloud technology)是指在广域网或局域网内将硬件、软件、网络等系列资源统一起来,实现数据的计算、储存、处理和共享的一种托管技术。云技术能够应用于各种领域,如医疗云、云物联、云安全、云教育、云会议、人工智能云服务、云应用、云呼叫和云社交等,云技术基于云计算(cloud computing)商业模式应用,它将计算任务分布在大量计算机构成的资源池上,使各种应用系统能够根据需要获取计算力、存储空间和信息服务。提供资源的网络被称为“云”,“云”中的资源在使用者看来是可以无限扩展的,并且可

以随时获取,按需使用,随时扩展,按使用付费。作为云计算的基础能力提供商,会建立云计算资源池(简称云平台,一般称为IaaS(Infrastructure as a Service,基础设施即服务))平台,在资源池中部署多种类型的虚拟资源,供外部客户选择使用。云计算资源池中主要包括:计算设备(为虚拟化机器,包含操作系统)、存储设备、网络设备。

[0041] 按照逻辑功能划分,在IaaS层上可以部署PaaS(Platform as a Service,平台即服务)层,PaaS层之上再部署SaaS(Software as a Service,软件即服务)层,也可以直接将SaaS部署在IaaS上。PaaS为软件运行的平台,如数据库、web容器等。SaaS为各式各样的业务软件,如web门户网站、短信群发器等。一般来说,SaaS和PaaS相对于IaaS是上层。

[0042] 具体地,上述涉及的服务器02可以包括实体设备,可以具体包括有网络通信子模块、处理器和存储器等等,也可以包括运行于实体设备中的软体,可以具体包括有应用程序等。

[0043] 具体地,终端01可以包括智能手机、台式电脑、平板电脑、笔记本电脑、数字助理、增强现实(augmented reality,AR)/虚拟现实(virtual reality,VR)设备、智能语音交互设备、智能家电、智能可穿戴设备、车载终端设备等类型的实体设备,也可以包括运行于实体设备中的软体,例如应用程序等。

[0044] 本申请实施例中,终端01可以用于获取待分析图像和任务指示文本,并将其发送至服务器02,以使服务器02调用内容理解模型进行相应的图像内容提取和图像分类,进而得到内容描述文本和图像类别结果。进一步地,服务器02还可以将该内容描述文本和图像类别结果推送至终端01进行显示。服务器01可以提供内容理解模型的模型训练服务。

[0045] 具体地,本申请的图像内容分析方法可以应用于各种需基于图像内容理解进行图像分类的业务场景,如内容审核业务,判断图像是否含有特定或违规内容,或者视觉问答(VQA)、图像描述(image captioning)、目标检测、定位和图像分类等业务。

[0046] 此外,可以理解的是,图1所示的仅仅是一种图像内容分析方法的应用环境,该应用环境可以包括更多或更少的节点,本申请在此不做限制。

[0047] 本申请实施例涉及的应用环境,或应用环境中的终端01和服务器02等可以是由客户端、多个节点(接入网络中的任意形式的计算设备,如服务器、用户终端)通过网络通信的形式连接形成的分布式系统。分布式系统可以为区块链系统,该区块链系统可以提供上述的图像内容分析服务和数据存储服务等。

[0048] 以下基于上述应用环境介绍本申请的技术方案,本申请实施例可应用于各种场景,包括但不限于云技术、人工智能、智慧交通、辅助驾驶等。请参考图2,图2是本申请实施例提供的一种图像内容分析方法的流程示意图,本说明书提供了如实施例或流程图的方法操作步骤,但基于常规或者无创造性的劳动可以包括更多或者更少的操作步骤。实施例中列举的步骤顺序仅仅为众多步骤执行顺序中的一种方式,不代表唯一的执行顺序。在实际中的系统或服务器产品执行时,可以按照实施例或者附图所示的方法顺序执行或者并行执行(例如并行处理器或者多线程处理的环境)。具体地,如图2所示,方法可以包括下述步骤S201-S203:

[0049] S201:获取待分析图像和任务指示文本。

[0050] 具体地,待分析图像可以为图片数据或者视频数据,为具备内容理解需求的图像,示例性地,在审核业务场景中,可以对待分析图像进行内容理解和分类,以分析其是否包含

或涉及审核业务所限定的特定或违规内容。任务指示文本用于提供内容理解模型执行图像内容提取和图像分类所需的指导信息。具体地,任务指示文本向内容理解模型提供了预设业务场景下进行图像内容分析和图像分类的任务描述或上下文线索,可以包括但不限于任务概括、问题描述或输入与期望输出间的映射关系等,通过任务指示文本提升内容理解模型在处理预设业务下解决相关任务的准确性和针对性,进而提升输出结果的可靠性。示例性地,在特定内容审核业务场景中,任务指示文本可以为“描述这张图片,分析是否含有特定的内容”。

[0051] S203:将待分析图像和任务指示文本输入内容理解模型进行图像内容提取和图像分类,得到待分析图像的内容描述文本和图像类别结果。

[0052] 具体地,参考图5,内容理解模型包括依次连接的图像编码模块、多模态映射模块、文本解码模块和分类模块,图像编码模块用于以待分析图像为输入并输出视觉特征,多模态映射模块用于将视觉特征转换为文本特征嵌入空间的映射特征,文本解码模块是基于预训练的大语言模型构建的,用于基于映射特征和任务指示文本输出内容理解特征和内容描述文本,分类模块用于基于内容理解特征进行分类识别,得到图像类别结果。大语言模型(Large Language Models,LLMs)是指使用大量文本数据训练的深度学习模型,可以生成自然语言文本或理解语言文本的含义,本申请所采用的大语言模型可以包括但不限于预训练的LLaMA(Large Language Model Meta AI,羊驼模型)、GPT-3(General Pre-trained Transformer-3,生成式预训练转换器)等。

[0053] 具体地,图像编码模块用于对输入图像进行特征提取,以得到分片(patch)级的视觉特征,多模态映射模块用于以视觉特征为输入进行特征映射以对齐视觉特征和文本特征,得到文本解码模块能够理解的映射特征;文本解码模块用于以任务指示文本和映射特征为输入,并以任务指示文本为指导对映射特征进行内容分析,得到能够表达任务指示文本所需图像内容回答的内容理解特征,并通过其输出层对内容理解特征进行映射,输出内容描述文本。内容描述文本为针对任务指示文本的响应回答文本,用于对任务指示文本所指示的图像内容进行描述,该内容描述文本为针对图像的长文本理解。示例性地,针对前述“描述这张图片,分析是否含有特定的内容”的内容描述文本可以为“图片中有3位小朋友站在旗帜前,可能是在一片草地上,他们戴着红领巾,站成一排在旗帜下敬礼。这张场景传递出这些青少年具有很强的爱国团结精神,图片中含有旗帜这一特定元素”。分类模块用于以内容理解特征为输入并输出图像类别结果,图像类别结果用于表征待分析图像在预设业务场景下所属的预设类别,如内容审核场景下,可以表征审核的预设类别或是否违规,前述示例的内容描述文本所对应的图像类别结果为“特定”类别。相应地,一些实施例中,参考图3和图6,S203可以包括S301-S307:

[0054] S301:将待分析图像输入图像编码模块进行特征提取,得到视觉特征。

[0055] 具体地,视觉特征可以包括待分析图像的浅层和深层特征,维度和尺寸可以基于需求设定。视觉特征可以为分片级(patch级)特征,即可以包括同一待分析图像中不同图像区域所对应的图像patch特征。

[0056] 一些实施例中,图像编码模块是结合图文样本对,对预设的文本特征提取网络和基于自注意力机制的图像特征提取网络进行图像和文本匹配的分类识别约束训练得到的。将图文样本对中的文本输入文本特征提取网络进行特征提取,得到文本特征,将图文样本

对中的图像输入图像特征提取网络进行特征提取,得到图像特征,计算文本特征和图像特征的相似度,以基于分类识别的方式对文本特征和图像特征进行匹配,以输出二者匹配结果,进而基于匹配结果和真值确定模型损失以反向传播调整网络参数,至满足预训练条件,将满足预训练条件的图像特征提取网络确定为图像编码模块。上述图文样本对可以为能够获取到的公开图文对样本集。

[0057] 一个实施例中,图像特征提取网络可以基于卷积神经网络(CNN)和自注意力机制构建,如可以采用CLIP模型(Contrastive Language-Image Pre-Training,对比语言-图像预训练模型)或CLIP ViT-L/14等的视觉模块。

[0058] S303:将视觉特征输入多模态映射模块进行特征映射,以将视觉特征映射至文本解码模块的文本特征嵌入空间,得到映射特征。

[0059] 具体地,经多模态映射模块处理后,图像patch特征映射为文本token特征,以用于文本解码模块的图像内容理解,提升信息表达准确性。

[0060] 一些实施例中,多模态映射模块包括第一转换器层和第二转换器层,即基于2层Transformer构建,S303可以包括:

[0061] S3031:将视觉特征输入第一转换器层进行分片级的特征表示,以将视觉特征映射至词嵌入空间,得到分片嵌入特征;

[0062] S3032:将分片嵌入特征输入第二转换器层进行上下文信息交叉提取,得到映射特征。

[0063] 具体地,第一转换器层用于对齐视觉特征和文本词特征表示,以将图片patch特征映射为token级的分片嵌入特征。第一转换器层相当于字符级的Transformer,分片嵌入特征相当于对文本语句编码提取后得到的句子表示。第二转换器层用于基于自注意力机制对分片嵌入特征中对应的每一token进行上下文特征提取,第二转换器层相当于句子级的Transformer,映射特征相当于对句子表示进行上下文信息强化的文本表示。如此,通过第一转换器层实现视觉特征和文本特征的初步对齐,通过第二转换器层能够将包含分散的图像patch特征含义表达的分片嵌入特征转换为能够表达图像整体上下文内容的映射特征,提升映射特征的表达准确性,并且,该映射模块可训练,使用垂直类领域数据训练后,能够实现各领域任务和图像分析的泛化应用。

[0064] 另一些实施例中,多模态映射模块包括第一转换器层和第二转换器层,第一转换器层和第二转换器层,2层transformer均用于对输入特征进行基于自注意力机制的上下文信息交叉提取,即视觉特征输入第一转换器层进行交叉提取后得到的输出作为第二转换器层的输入,再进行一次交叉提取后得到映射特征。如此,通过两层上下文加强的特征提取实现视觉特征空间和文本词嵌入空间的映射,进而提升文本解码模块的特征对齐,实现准确内容理解。

[0065] S305:将映射特征和任务指示文本输入文本解码模块进行内容理解,得到内容理解特征和内容描述文本。

[0066] 具体地,内容描述文本是基于文本解码模块的输出层对内容理解特征进行特征文本映射得到的。任务指示文本首先经文本特征表示处理生成指示嵌入特征,即将任务指示文本映射至文本解码模块的文本词嵌入空间,与映射特征对齐,该文本特征表示处理可以由单独设置于文本解码模块前的文本特征表示层实现,或者也可以由文本解码模块中的文

本特征表示层实现。映射特征与指示嵌入特征拼接后得到的拼接特征经文本解码模块的特征提取后,生成内容理解特征,并经输出层映射后得到内容描述文本。内容理解特征融合了待分析图像的图像内容和任务指示文本的内容,实现二者的信息表达,实现对任务指示文本所提出的问题进行图像内容的倾向性理解和表达,满足预设业务场景下的图像描述和分类。

[0067] S307:将内容理解特征输入分类模块进行分类识别,得到图像类别结果。

[0068] 具体地,分类模块连接于文本解码模块的输出层之前,以内容理解特征为输入并将其映射为相应的预设类别,进而实现内容理解模型针对图像的长文本理解和类别的同时输出。通过添加分类模块以作为文本决策判别器,对文本解码模块的输出进行进一步分析,类似于文本分类输出业务需要的有限的离散标签,以更精准的满足业务需求。

[0069] 如此,上述技术方案能够通过指令方式提升模型理解能力,进而提升内容描述和图像分类的准确性,同时结合大语言模型和指令方式将内容分析和内容分类整合至同一模型中进行内容理解和分类识别,能够将内容理解中的多种任务进行统一,满足需结合内容分析的分类业务需求,输入是指示文本和图像,能够由单独的一个模型来综合分析图像中出现的所有元素和元素之间的关系,综合分析后输出图像分析的结果,相比于已有方法中多个模型单独分析再后处理的方式,流程更简洁高效,内容理解也更准确,显著提升分析效率和降低训练成本。

[0070] 基于上述部分或全部实施方式,本申请实施例中,内容理解模型是结合图文对指令数据集对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的约束训练,以及结合多模态指令数据集对初始理解模型进行图像内容提取和图像分类的约束训练得到的。如此,能够实现图像编码模块、多模态映射模块和分类层的分段训练,提升模型能力和训练收敛效率,且经指令学习迁移训练得到多模态模型,能够显著提升其图像分析能力。

[0071] 本申请实施例中,方法还包括内容理解模型的训练方法,参考图4,具体可以包括步骤S401-S407:

[0072] S401:获取通用领域的图文对指令数据集、预设业务领域的图文对指令数据集和文本对话数据集。

[0073] 具体地,通用领域的图文对指令数据集可以为包括多种业务领域的公开图文数据集,如CC(Conceptual Captions)数据集,SBU(SBU Captions)数据集,或LAION(Large-scale Artificial Intelligence Open Network)数据集等,通过大量图文对数据训练图像编码模块和多模态映射模块学习视觉和语言的关系及知识。预设业务领域是指业务需求的垂直类领域,如内容审核、医学、交通或金融等领域。预设业务领域的图文对指令数据集可以包括相应垂直类领域内的图像文本对,可以为公开数据集或者人工标注的特定数据集。文本对话数据集可以是公开的或人工标注的问答文本对数据集,具体可以包括预设业务领域相关的样本指令文本和对应的答案标注。

[0074] 具体地,通用领域对应的图文对指令数据集包括第一样本图像、第一指示信息和第一样本图像对应的第一文本标注,第一文本标注为基于第一指示信息进行图像描述所对应的文本响应真值,如第一指示信息为针对第一样本图像提出的与图像内容相关的问题,第一文本标注为针对第一指示信息的答案。如前述的,第一样本图像为通用领域图像。

[0075] 具体地,预设业务领域对应的图文对指令数据集包括第二样本图像、第二指示信息、第二样本图像对应的第二文本标注,以及第二样本图像的类别标签。第二文本标注为基于第二指示信息进行图像描述所对应的文本响应真值,如第二指示信息为针对第二样本图像提出的与图像内容相关的问题,第二文本标注为针对第二指示信息的答案。如前述的,第二样本图像为预设的垂直类领域图像,如内容审核业务领域中,第二样本图像可以为收集的包括不良内容的图像,第二文本标注可以为对其不良内容进行描述的文本,类别标签表征第二样本图像在该垂直类领域中所属的预设类别真值,如所属的不良内容类型。同时,部分第二样本图像还可以为收集的正常样本,如不包含不良内容的图像。

[0076] 进一步地,与任务指示文本相类似的,第一指示信息和第二指示信息均用于提供文本解码模块在进行内容理解时所需的指导信息。第一指示信息用于指示内容理解模型对第一样本图像进行内容理解并输出与第一文本标注相近的样本描述文本;第二指示信息用于指示内容理解模型对第二样本图像进行内容理解并输出与第二文本标注相近的样本描述文本。

[0077] 一些实施例中,同一第一样本图像可以包括多维度的第一指示信息,同一第二样本图像也可以包括多维度的第二指示信息,以从多种内容理解维度提出任务指示,如图像中的对象元素维度、图像整体内容理解维度或图像内容的进一步逻辑推理维度等。相应地,第一指示信息和第二指示信息均包括多维度的指令文本,多维度的指令文本包括对象描述指令、内容属性描述指令和内容推理指令,以使内容理解模型学习多种维度的图像内容知识。

[0078] 对象描述指令用于指示文本解码模块进行图像对象描述,对象描述指令对应的第一文本标注或第二文本标注具有唯一真值;其中,图像对象是指图像中出现的对象元素,如图像中出现的人、动物、植物、物品或景物等,对象描述指令和与其对应的文本标注可以形成对话形式,对象描述指令可以用于提出与图像中对象元素相关的问题,是具有明确答案的问题,文本标注为相应问题的解答,提供唯一的真值答案。示例性地,对象描述指令可以包括但不限于与对象类型、对象位置、对象动作、对象数量、对象之间的相对位置等相关的问题。

[0079] 内容属性描述指令用于指示文本解码模块进行图像整体信息描述,内容属性描述指令对应的第一文本标注或第二文本标注具有开放式答案属性,用于对图像进行全面描述,以包含图像中的元素、场景、环境和氛围信息等。示例性地,内容属性描述指令可以为“请描述图片阐述的内容”。

[0080] 内容推理指令用于指示文本解码模块进行图像内容的推理解答,在图像内容本身的基础上进一步深入推理。该内容推理指令可以与预设业务领域相关,用于提出与预设业务领域相匹配的问题,如内容审核场景下,内容推理指令可以为“这张图片涉及特定内容吗?”,对应的文本标注可以为“涉及特定内容,因为图片中含有某个标志的旗帜”,或,内容推理指令可以为“这张图片适合儿童观看吗?”,对应的文本标注可以为“不适合,因为图片中有较违规的穿着,拿着一个违规用品,做出XX的行为,可能含有违规信息,所以不适合儿童观看”。

[0081] 通过设置上述类型的第一指示信息和第二指示信息,能够促进模型从不同维度上理解图像内容,进而实现图像的浅层特征和深层含义的描述,提升内容理解准确性,进而提

升图像描述文本和分类结果的准确性,以及提升与业务场景的匹配性。

[0082] 具体地,文本对话数据集包括样本指令文本和样本指令文本对应的答案标注。内容审核领域中,样本指令文本可以为一段描述有不良内容的文本和针对其是否存在不良内容的提问,相应答案标注为针对上述提问的回答;同时,还可以收集不包含不良内容的正常文本和相应的答案标注。

[0083] S403:基于通用领域的图文对指令数据集和预设业务领域的图文对指令数据集,对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的第一约束训练。

[0084] 具体地,在训练过程中冻结文本解码模块的模型参数并调整图像编码模块和多模态映射模块的模型参数,至满足第一训练结束条件。如此,避免限制图像编码模块的视觉/文本知识的关联能力,提升其视觉理解能力和针对垂直类领域图像数据的知识储备,进而提高视觉编码和多模态特征对齐能力。

[0085] 相应地,第一约束训练采用下述步骤S4031-S4035实现:

[0086] S4031:以第一样本图像或第二样本图像作为图像编码模块的输入进行特征提取,得到第一样本图像对应的第一样本视觉特征和第二样本图像对应的第二样本视觉特征;

[0087] S4032:以第一样本视觉特征或第二样本视觉特征作为多模态映射模块的输入进行特征映射,以分别将第一样本视觉特征和第二样本视觉特征映射至文本解码模块的文本特征嵌入空间,得到第一映射特征和第二映射特征;

[0088] S4033:以第一映射特征和第一指示信息形成的数据对,或,第二映射特征和第二指示信息形成的数据对作为文本解码模块的输入进行内容理解,得到第一样本描述文本和第二样本描述文本;

[0089] S4034:基于第一样本描述文本和第一文本标注间的差异,以及第二样本描述文本和第二文本标注间的差异确定第一模型损失;

[0090] S4035:冻结文本解码模块的模型参数,根据第一模型损失训练图像编码模块和多模态映射模块,至满足第一训练结束条件。

[0091] 具体地,在第一阶段训练中,以通用领域的图文对指令数据集和预设业务领域的图文对指令数据集作为训练数据,单次迭代中,将多个第一样本图像和多个第二样本图像输入图像编码模块,以分别得到相应的视觉特征,并通过特征映射生成第一样本图像对应的第一映射特征和第二样本图像对应的第二映射特征。进一步地,第一指示信息对应的第一指示嵌入特征与第一映射特征拼接后,通过文本解码模块进行基于内容理解的特征提取,生成第一样本描述文本,第二指示信息对应的第二指示嵌入特征与第二映射特征拼接后,通过文本解码模块进行基于内容理解的特征提取,生成第二样本描述文本。

[0092] 第一模型损失可以通过融合第一损失和第二损失确定,第一损失基于第一样本描述文本和第一文本标注间的差异生成,第二损失基于第二样本描述文本和第二文本标注间的差异生成,通过第一模型损失调整图像编码模块和多模态映射模块的模型参数,使其不断学习通用领域图像与语言间的关系和知识,以及学习垂直领域图像与相关语言间的知识,提升视觉和语言模态间对齐的准确性、可靠性。

[0093] S405:基于由预设业务领域的图文对指令数据集和文本对话数据集形成的多模态指令数据集,对满足第一训练结束条件的初始理解模型进行图像内容提取和图像分类的第

二约束训练。

[0094] 具体地,在训练过程中冻结图像编码模块和多模态映射模块的模型参数并调整文本解码模块和分类模块的模型参数,至满足第二训练结束条件。如此,冻结已完成一阶段训练的图像编码模块和多模态映射模块,提升模型收敛速度和训练效率,同时采用与业务场景匹配的数据集、基于指令微调训练文本解码模块的内容理解能力、优化描述文本能力和分类模块的分类输出能力。

[0095] 一些实施例中,相应地,第二约束训练采用下述步骤S4051-S4056实现包括:

[0096] S4051:以第二样本图像作为满足第一训练结束条件的图像编码模块的输入进行特征提取,得到第三样本视觉特征;

[0097] S4052:以第三样本视觉特征作为满足第一训练结束条件的多模态映射模块的输入进行特征映射,得到第三映射特征;

[0098] S4053:以第二映射特征和第二指示信息形成的数据对,或,样本指令文本作为文本解码模块的输入进行内容理解,得到第三样本理解特征、第三样本描述文本和答案文本;

[0099] S4055:将第三样本理解特征输入分类模块进行分类识别,得到样本分类结果;

[0100] S4055:基于第三样本描述文本、第二文本标注、答案文本、答案标注、样本分类结果和类别标签确定第二模型损失;

[0101] S4056:冻结图像编码模块和多模态映射模块的模型参数,根据第二模型损失训练文本解码模块和分类模块,至满足第二训练结束条件。

[0102] 具体地,在第二阶段训练中,采用预设业务领域的图文对指令数据集和文本对话数据集作为训练数据,进行预设业务领域内的知识学习,提升模型与业务的匹配度和垂直类领域内的输出准确性。

[0103] 针对预设业务领域的图文对指令数据集,与一阶段类似的,第二样本图像输入图像编码模块经特征提取后得到第三样本视觉特征,再经特征映射后得到第三映射特征;第三映射特征与第二指示信息对应的第三指示嵌入特征拼接,以进行后续基于内容理解的特征提取,生成第三样本理解特征和第三样本描述文本,第三样本理解特征通过分类模块后得到样本分类结果,第三样本理解特征和前述的内容理解特征相类似,在此不做赘述。针对文本对话数据集,将样本指令文本输入文本解码模块进行特征嵌入和特征提取,以生成答案文本,以训练文本解码模块的语言知识理解能力。

[0104] 第二模型损失可以通过融合第三损失、第四损失和第五损失得到,第三损失基于第三样本描述文本和第二文本标注间的差异确定,第四损失基于答案文本和答案标注间的差异确定,第五损失基于样本分类结果和类别标签间的差异确定。

[0105] 第二阶段训练中,冻结图像编码模块和多模态映射模块的模型参数,根据第二模型损失训练文本解码模块和分类模块,至满足第二训练结束条件,以通过垂直类领域的纯文本指令数据和多模态指令数据实现文本解码模块、分类模块的微调,提升预设业务服务能力。

[0106] S407:将满足第二训练结束条件的初始理解模型确定为内容理解模型。

[0107] 具体地,上述的第一阶段和第二阶段训练过程中可以采用采用随机梯度下降法(SGD)求解网络模型参数(如图像解码模块的卷积模板参数w和偏置参数 b等),在每次迭代过程中,计算模型损失并反向传播到网络模型,计算梯度并更新模型参数。

[0108] 一个实施例中,训练环境可以为16张GPU,SGD的学习率可设置为0.02,批处理尺寸(batch size)设置为32张图片,每个GPU两张图片。第一阶段训练可以在 4 个 A100显卡上用通用领域的图文对指令数据集和预设业务领域的图文对指令数据集训练,然后再用预设业务领域的各个高质量图文对指令数据集训练。

[0109] 本申请的技术方案为端到端的图像分析方法,输入待检测图像,模型可直接输入最终分析结果,无需额外人为参与,相比于普通的图文多模态对话模型,基于本申请的模型结构和设置能够输出针对性的判别结果,供业务系统直接使用,便于决策。同时,基于上述指令学习的分阶段训练方案,实现图像编码模块、多模态映射模块和分类层的分别训练,能够使模型学习到更好的领域知识,提高领域内图像内容理解的准确性。

[0110] 本申请实施例还提供了一种图像内容分析装置800,如图7所示,图7示出了本申请实施例提供的一种图像内容分析装置的结构示意图,装置可以包括下述模块。

[0111] 获取模块10:获取待分析图像和任务指示文本;

[0112] 内容分析模块20:用于将待分析图像和任务指示文本输入内容理解模型进行图像内容提取和图像分类,得到待分析图像的内容描述文本和图像类别结果;任务指示文本用于提供内容理解模型执行图像内容提取和图像分类所需的指导信息;

[0113] 内容理解模型包括依次连接的图像编码模块、多模态映射模块、文本解码模块和分类模块,图像编码模块用于以待分析图像为输入并输出视觉特征,多模态映射模块用于将视觉特征转换为文本特征嵌入空间的映射特征,文本解码模块是基于预训练的大语言模型构建的,用于基于映射特征和任务指示文本输出内容理解特征和内容描述文本,分类模块用于基于内容理解特征进行分类识别,得到图像类别结果;

[0114] 内容理解模型是结合图文对指令数据集对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的约束训练,以及结合多模态指令数据集对初始理解模型进行图像内容提取和图像分类的约束训练得到的。

[0115] 一些实施例中,内容分析模块20可以包括:

[0116] 特征提取子模块:用于将待分析图像输入图像编码模块进行特征提取,得到视觉特征;

[0117] 特征映射子模块:用于将视觉特征输入多模态映射模块进行特征映射,以将视觉特征映射至文本解码模块的文本特征嵌入空间,得到映射特征;

[0118] 内容理解子模块:用于将映射特征和任务指示文本输入文本解码模块进行内容理解,得到内容理解特征和内容描述文本,内容描述文本是基于文本解码模块的输出层对内容理解特征进行特征文本映射得到的;

[0119] 分类识别子模块:用于将内容理解特征输入分类模块进行分类识别,得到图像类别结果。

[0120] 一些实施例中,多模态映射模块包括第一转换器层和第二转换器层,特征映射子模块可以包括:

[0121] 特征表示单元:用于将视觉特征输入第一转换器层进行分片级的特征表示,以将视觉特征映射至词嵌入空间,得到分片嵌入特征;

[0122] 交叉提取单元:用于将分片嵌入特征输入第二转换器层进行上下文信息交叉提取,得到映射特征。

[0123] 一些实施例中,图像编码模块是结合图文样本对,对预设的文本特征提取网络和基于自注意力机制的图像特征提取网络进行图像和文本匹配的分类识别约束训练得到的。

[0124] 一些实施例中,装置还包括:

[0125] 数据集获取模块:用于获取通用领域的图文对指令数据集、预设业务领域的图文对指令数据集和文本对话数据集;

[0126] 第一训练模块:用于基于通用领域的图文对指令数据集和预设业务领域的图文对指令数据集,对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的第一约束训练,在训练过程中冻结文本解码模块的模型参数并调整图像编码模块和多模态映射模块的模型参数,至满足第一训练结束条件;

[0127] 第二训练模块:用于基于由预设业务领域的图文对指令数据集和文本对话数据集形成的多模态指令数据集,对满足第一训练结束条件的初始理解模型进行图像内容提取和图像分类的第二约束训练,在训练过程中冻结图像编码模块和多模态映射模块的模型参数并调整文本解码模块和分类模块的模型参数,至满足第二训练结束条件;

[0128] 模型确定模块:用于将满足第二训练结束条件的初始理解模型确定为内容理解模型。

[0129] 一些实施例中,通用领域对应的图文对指令数据集包括第一样本图像、第一指示信息和第一样本图像对应的第一文本标注,预设业务领域对应的图文对指令数据集包括第二样本图像、第二指示信息和第二样本图像对应的第二文本标注,第一指示信息和第二指示信息均用于提供文本解码模块在进行内容理解时所需的指导信息,第一文本标注为基于第一指示信息进行图像描述所对应的文本响应真值,第二文本标注为基于第二指示信息进行图像描述所对应的文本响应真值;第一训练模块可以包括:

[0130] 第一样本提取子模块:用于以第一样本图像或第二样本图像作为图像编码模块的输入进行特征提取,得到第一样本图像对应的第一样本视觉特征或第二样本图像对应的第二样本视觉特征;

[0131] 第一样本映射子模块:用于以第一样本视觉特征或第二样本视觉特征作为多模态映射模块的输入进行特征映射,以分别将第一样本视觉特征和第二样本视觉特征映射至文本解码模块的文本特征嵌入空间,得到第一映射特征和第二映射特征;

[0132] 第一样本理解子模块:用于以第一映射特征和第一指示信息形成的数据对,或,第二映射特征和第二指示信息形成的数据对作为文本解码模块的输入进行内容理解,得到第一样本描述文本和第二样本描述文本;

[0133] 第一损失子模块:用于基于第一样本描述文本和第一文本标注间的差异,以及第二样本描述文本和第二文本标注间的差异确定第一模型损失;

[0134] 第一训练子模块:用于冻结文本解码模块的模型参数,根据第一模型损失训练图像编码模块和多模态映射模块,至满足第一训练结束条件。

[0135] 一些实施例中,预设业务领域的图文对指令数据集还包括第二样本图像的类别标签,文本对话数据集包括样本指令文本和样本指令文本对应的答案标注,第二训练模块可以包括:

[0136] 第二样本提取子模块:用于以第二样本图像作为满足第一训练结束条件的图像编码模块的输入进行特征提取,得到第三样本视觉特征;

[0137] 第二样本映射子模块:用于以第三样本视觉特征作为满足第一训练结束条件的多模态映射模块的输入进行特征映射,得到第三映射特征;

[0138] 第二样本理解子模块:用于以第二映射特征和第二指示信息形成的数据对,或,样本指令文本作为文本解码模块的输入进行内容理解,得到第三样本理解特征、第三样本描述文本和答案文本;

[0139] 样本分类子模块:用于将第三样本理解特征输入分类模块进行分类识别,得到样本分类结果;

[0140] 第二损失子模块:用于基于第三样本描述文本、第二文本标注、答案文本、答案标注、样本分类结果和类别标签确定第二模型损失;

[0141] 第二训练子模块:用于冻结图像编码模块和多模态映射模块的模型参数,根据第二模型损失训练文本解码模块和分类模块,至满足第二训练结束条件。

[0142] 一些实施例中,第一指示信息和第二指示信息均包括多维度的指令文本,多维度的指令文本包括对象描述指令、内容属性描述指令和内容推理指令;

[0143] 对象描述指令用于指示文本解码模块进行图像对象描述,对象描述指令对应的第一文本标注或第二文本标注具有唯一真值;内容属性描述指令用于指示文本解码模块进行图像整体信息描述,内容属性描述指令对应的第一文本标注或第二文本标注具有开放式答案属性;内容推理指令用于指示文本解码模块进行图像内容的推理解答。

[0144] 需要说明的是,上述装置实施例与方法实施例基于相同的实施方式。

[0145] 本申请实施例提供了一种设备,该设备可以为终端或服务器,包括处理器和存储器,该存储器中存储有至少一条指令或至少一段程序,该至少一条指令或该至少一段程序由该处理器加载并执行以实现如上述方法实施例所提供的图像内容分析方法。

[0146] 存储器可用于存储软件程序以及模块,处理器通过运行存储在存储器的软件程序以及模块,从而执行各种功能应用以及图像内容分析。存储器可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、功能所需的应用程序等;存储数据区可存储根据设备的使用所创建的数据等。此外,存储器可以包括高速随机存取存储器,还可以包括非易失性存储器,例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。相应地,存储器还可以包括存储器控制器,以提供处理器对存储器的访问。

[0147] 本申请实施例所提供的方法实施例可以在移动终端、计算机终端、服务器或者类似的运算装置等电子设备中执行。图8是本申请实施例提供的一种图像内容分析方法的电子设备的硬件结构框图。如图8所示,该电子设备900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)910(处理器910可以包括但不限于微处理器MCU或可编程逻辑器件FPGA等的处理装置)、用于存储数据的存储器930,一个或一个以上存储应用程序923或数据922的存储介质920(例如一个或一个以上海量存储设备)。其中,存储器930和存储介质920可以是短暂存储或持久存储。存储在存储介质920的程序可以包括一个或一个以上模块,每个模块可以包括对电子设备中的一系列指令操作。更进一步地,中央处理器910可以设置为与存储介质920通信,在电子设备900上执行存储介质920中的一系列指令操作。电子设备900还可以包括一个或一个以上电源960,一个或一个以上有线或无线网络接口950,一个或一个以上输入输出接口940,和/或,一个或一个以上操作系统921,例如Windows ServerTM,Mac OS XTM,UnixTM, LinuxTM,

FreeBSDTM等等。

[0148] 输入输出接口940可以用于经由一个网络接收或者发送数据。上述的网络具体实例可包括电子设备900的通信供应商提供的无线网络。在一个实例中,输入输出接口940包括一个网络适配器(Network Interface Controller,NIC),其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个实例中,输入输出接口940可以为射频(Radio Frequency,RF)模块,其用于通过无线方式与互联网进行通讯。

[0149] 本领域普通技术人员可以理解,图8所示的结构仅为示意,其并不对上述电子装置的结构造成限定。例如,电子设备900还可包括比图8中所示更多或者更少的组件,或者具有与图8所示不同的配置。

[0150] 本申请的实施例还提供了一种计算机可读存储介质,存储介质可设置于电子设备之中以保存用于实现方法实施例中一种图像内容分析方法相关的至少一条指令或至少一段程序,该至少一条指令或该至少一段程序由该处理器加载并执行以实现上述方法实施例提供的图像内容分析方法。

[0151] 可选地,在本实施例中,上述存储介质可以位于计算机网络的多个网络服务器中的至少一个网络服务器。可选地,在本实施例中,上述存储介质可以包括但不限于:U盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

[0152] 根据本申请的一个方面,提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述各种可选实现方式中提供的方法。

[0153] 由上述本申请提供的图像内容分析方法、装置、设备、存储介质、服务器、终端和程序产品,本申请的技术方案首先获取待分析图像和任务指示文本,将待分析图像和任务指示文本输入内容理解模型进行图像内容提取和图像分类,得到待分析图像的内容描述文本和图像类别结果,该任务指示文本用于提供内容理解模型执行图像内容提取和图像分类所需的指导信息,能够通过指令方式提升模型理解能力,进而提升内容描述和图像分类的准确性;其中,内容理解模型包括依次连接的图像编码模块、多模态映射模块、文本解码模块和分类模块,图像编码模块用于以待分析图像为输入并输出视觉特征,多模态映射模块用于将视觉特征转换为文本特征嵌入空间的映射特征,文本解码模块是基于预训练的大语言模型构建的,用于基于映射特征和任务指示文本输出内容理解特征和内容描述文本,分类模块用于基于内容理解特征进行分类识别,得到图像类别结果;内容理解模型是结合图文对指令数据集对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的约束训练,以及结合多模态指令数据集对初始理解模型进行图像内容提取和图像分类的约束训练得到的;如此,结合大语言模型和指令方式将内容分析和内容分类整合至同一模型中进行内容理解和分类识别,能够将内容理解中的多种任务进行统一,满足需结合内容分析的分类业务需求,输入是指示文本和图像,能够由单独的一个模型来综合分析图像中出现的所有元素和元素之间的关系,综合分析后输出图像分析的结果,相比于已有方法中多个模型单独分析再后处理的方式,流程更简洁高效,内容理解也更准确,显著提升分析效率和降低训练成本;此外,能够实现图像编码模块、多模态映射模块和

分类层的分段训练,提升模型能力和训练收敛效率,且经指令学习迁移训练得到多模态模型,能够显著提升其图像分析能力。

[0154] 需要说明的是:上述本申请实施例先后顺序仅仅为了描述,不代表实施例的优劣。且上述对本申请特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下,在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外,在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中,多任务处理和并行处理也是可以的或者可能是有利的。

[0155] 本申请中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于装置、设备和存储介质实施例而言,由于其基本相似于方法实施例,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0156] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成,也可以通过程序来指示相关的硬件完成,的程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0157] 以上仅为本申请的较佳实施例,并不用以限制本申请,凡在本申请的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本申请的保护范围之内。

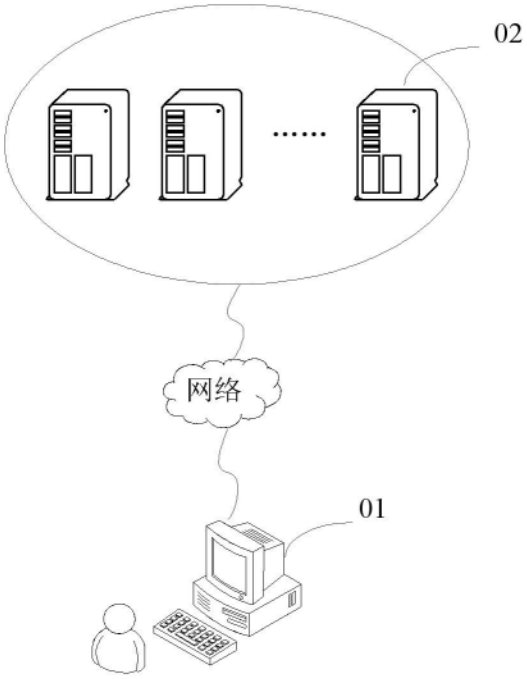

02
网络
01

图1

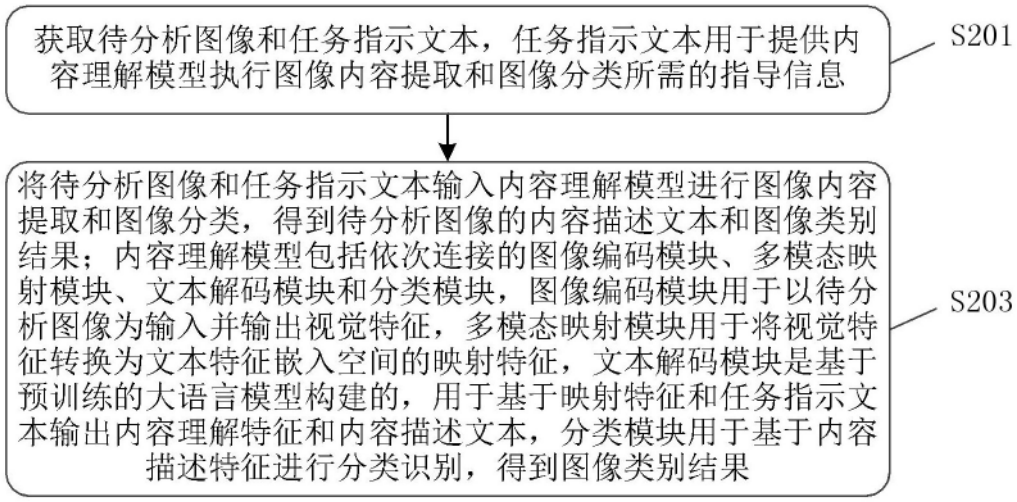

获取待分析图像和任务指示文本，任务指示文本用于提供内容理解模型执行图像内容提取和图像分类所需的指导信息
S201
将待分析图像和任务指示文本输入内容理解模型进行图像内容提取和图像分类，得到待分析图像的内容描述文本和图像类别结果；内容理解模型包括依次连接的图像编码模块、多模态映射模块、文本解码模块和分类模块，图像编码模块用于以待分析图像为输入并输出视觉特征，多模态映射模块用于将视觉特征转换为文本特征嵌入空间的映射特征，文本解码模块是基于预训练的大语言模型构建的，用于基于映射特征和任务指示文本输出内容理解特征和内容描述文本，分类模块用于基于内容描述特征进行分类识别，得到图像类别结果
S203

图2

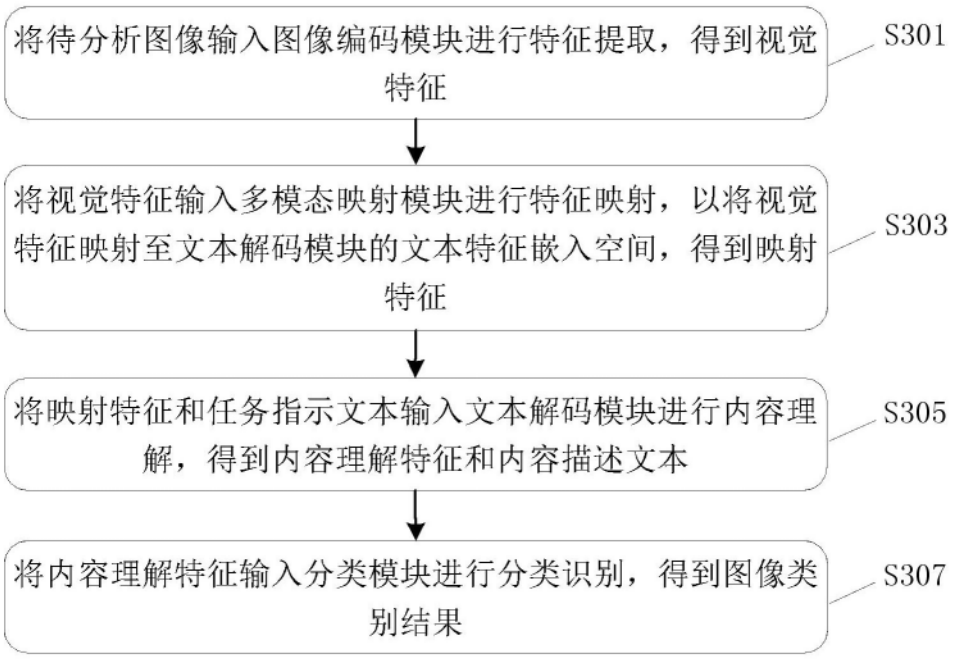

将待分析图像输入图像编码模块进行特征提取，得到视觉特征
S301
将视觉特征输入多模态映射模块进行特征映射，以将视觉特征映射至文本解码模块的文本特征嵌入空间，得到映射特征
S303
将映射特征和任务指示文本输入文本解码模块进行内容理解，得到内容理解特征和内容描述文本
S305
将内容理解特征输入分类模块进行分类识别，得到图像类别结果
S307

图3

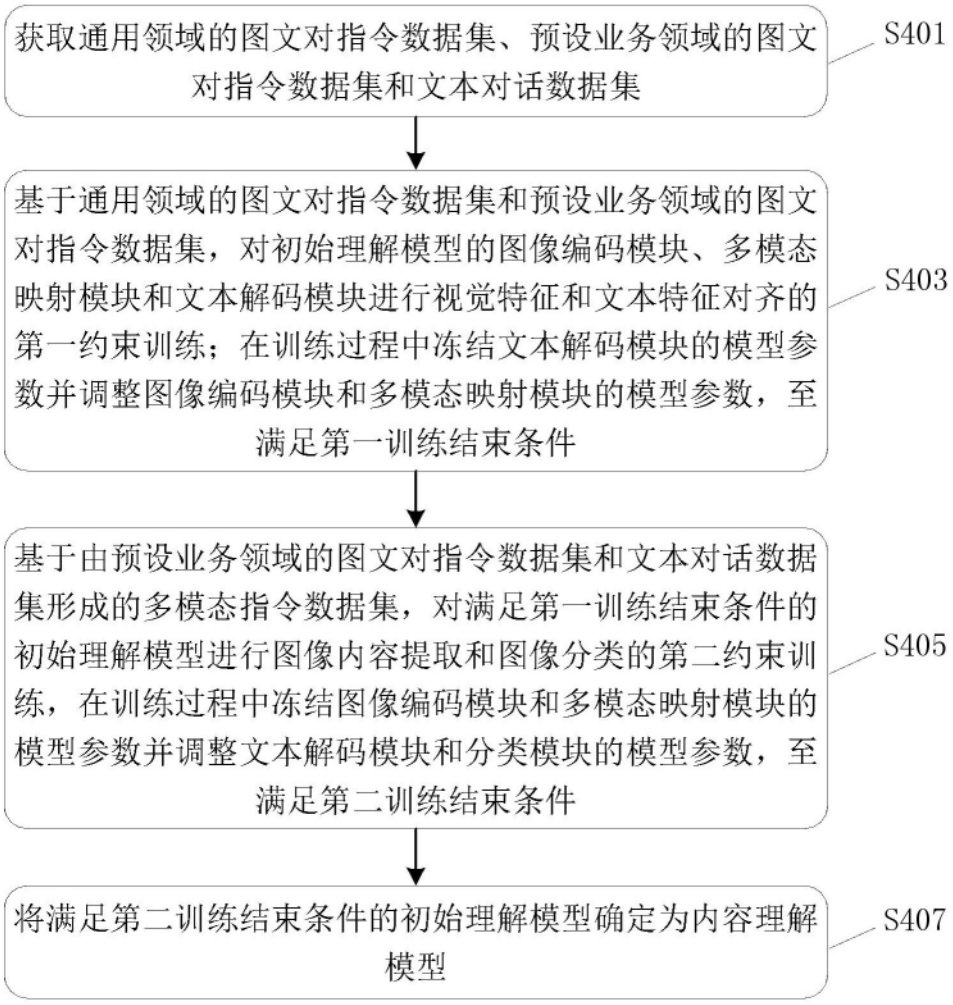

获取通用领域的图文对指令数据集、预设业务领域的图文对指令数据集和文本对话数据集
S401
基于通用领域的图文对指令数据集和预设业务领域的图文对指令数据集，对初始理解模型的图像编码模块、多模态映射模块和文本解码模块进行视觉特征和文本特征对齐的第一约束训练；在训练过程中冻结文本解码模块的模型参数并调整图像编码模块和多模态映射模块的模型参数，至满足第一训练结束条件
S403
基于由预设业务领域的图文对指令数据集和文本对话数据集形成的多模态指令数据集，对满足第一训练结束条件的初始理解模型进行图像内容提取和图像分类的第二约束训练，在训练过程中冻结图像编码模块和多模态映射模块的模型参数并调整文本解码模块和分类模块的模型参数，至满足第二训练结束条件
S405
将满足第二训练结束条件的初始理解模型确定为内容理解模型
S407

图4

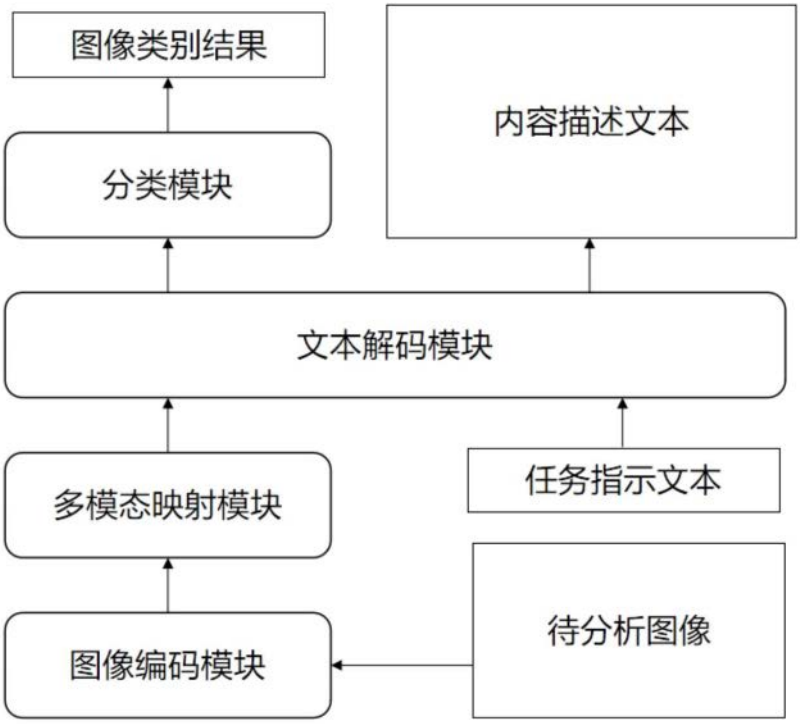

图像类别结果
内容描述文本
分类模块
文本解码模块
多模态映射模块
任务指示文本
待分析图像
图像编码模块

图5

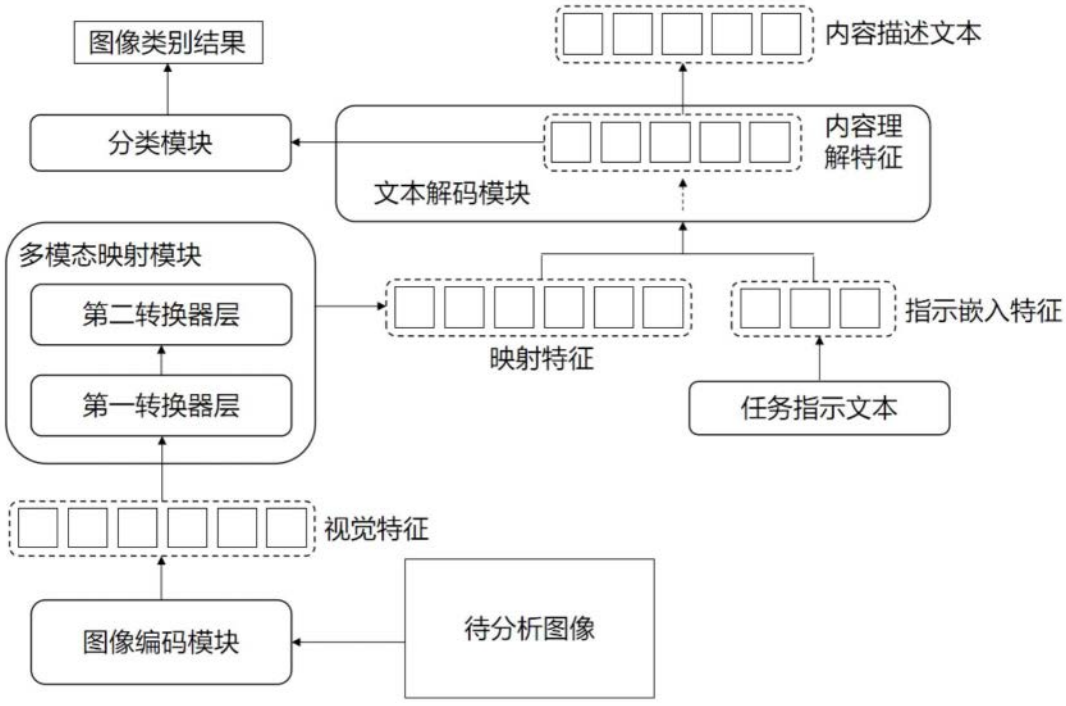

图像类别结果
内容描述文本
分类模块
内容理解特征
文本解码模块
多模态映射模块
第二转换器层
第一转换器层
映射特征
指示嵌入特征
任务指示文本
视觉特征
图像编码模块
待分析图像

图6

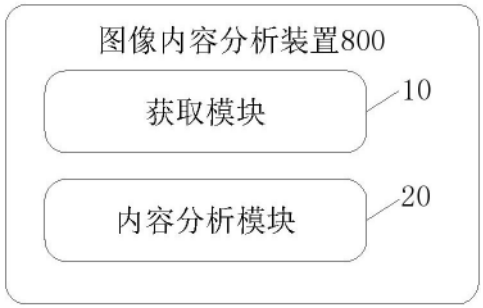

图像内容分析装置800
获取模块
10
内容分析模块
20

图7

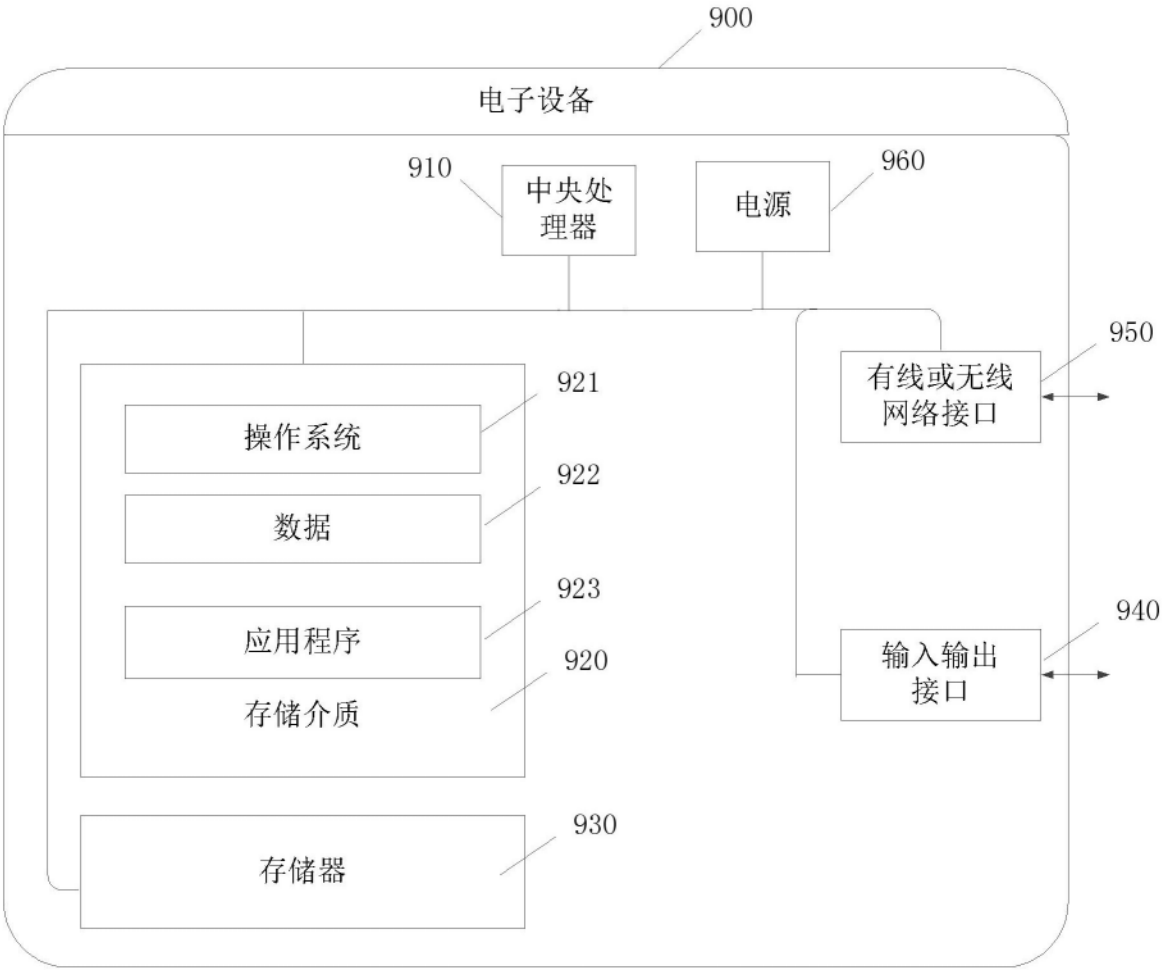

900
电子设备
910
中央处理器
电源
960
950
有线或无线网络接口
921
操作系统
922
数据
923
应用程序
920
存储介质
940
输入输出接口
930
存储器

图8