(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116311312 A
(43)申请公布日 2023.06.23

(21)申请号 202111569156.8

(22)申请日 2021.12.21

(71)申请人 鼎桥通信技术有限公司
地址 100102 北京市朝阳区望京北路9号叶青大厦13-15层
申请人 北京邮电大学

(72)发明人 李燮 龚萍 张子骥 凤珺仪 徐蔚然

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
专利代理师 宋兴 臧建明

(51)Int.Cl.
*G06V 30/413*(2022.01)
*G06V 10/82*(2022.01)
*G06F 16/9032*(2019.01)
*G06N 3/0464*(2023.01)
*G06N 3/08*(2023.01)

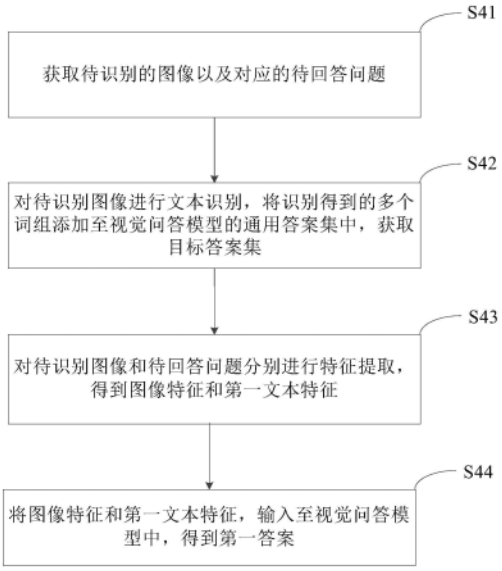
权利要求书3页 说明书12页 附图5页

(54)发明名称

# 视觉问答模型的训练方法和视觉问答方法

(57)摘要

本申请提供一种视觉问答模型的训练方法和视觉问答方法，该视觉问答方法包括：通过获取待识别的图像以及对应的待回答问题，对待识别图像进行文本识别，将识别得到的多个词组添加至视觉问答模型的通用答案集中，获取目标答案集，对待识别图像和待回答问题分别进行特征提取，得到图像特征和第一文本特征，将图像特征和第一文本特征，输入至视觉问答模型中，得到第一答案。通过将待识别图像中的多个词组添加至通用答案集中，使得利用视觉问答模型获取待回答问题对应的答案时，不仅考虑了通用答案集中的答案，还考虑了待识别图像中的词组对待回答问题的影响，有效提高了输出的答案的准确度。




CN 116311312 A

1.一种视觉问答模型的训练方法,其特征在于,包括:

获取训练样本集,所述训练样本集中包括多个样本图像,每个样本图像对应的样本问题以及答案;

针对所述样本训练集中的每个样本图像和对应样本问题分别进行特征提取,获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征;

根据通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特征和答案,对自顶向下与自底向上BUTD注意力模型进行模型训练,获取视觉问答模型,所述视觉问答模型用于根据图像特征,文本特征以及所述通用答案集获取待识别图像对应的待回答问题的答案。

2.根据权利要求1所述的方法,其特征在于,所述针对所述样本训练集中的每个样本图像和对应样本问题分别进行特征提取,获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征,包括:

将所述样本训练集中的每个样本图像输入区域卷积神经网络进行特征提取,获取所述样本图像特征;

将所述每个样本图像对应的样本问题输入长短期记忆网络进行特征提取,获取所述第一样本文本特征。

3.根据权利要求1或2所述的方法,其特征在于,所述BUTD注意力模型包括第一子BUTD注意力模型和第二子BUTD注意力模型,所述根据通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特征和答案,对自顶向下与自底向上BUTD注意力模型进行模型训练,获取视觉问答模型,包括:

根据所述通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特征和答案,依次对所述第一子BUTD注意力模型和所述第二子BUTD注意力模型进行训练,获取所述视觉问答模型,所述视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型;

所述第一子视觉问答模型是根据所述通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特征和答案对所述第一子BUTD注意力模型进行训练得到的,所述第一子BUTD注意力模型用于输出每个图像特征对应的多个样本答案,所述多个样本答案属于所述通用答案集;

所述第二子视觉问答模型是根据每个图像特征,以及每个图像特征对应的第二样本文本特征和答案对所述第二子BUTD注意力模型进行训练得到的,所述第二样本文本特征是对第一样本文本特征和对应的多个第一样本答案进行拼接得到的。

4.一种视觉问答方法,其特征在于,包括:

获取待识别的图像以及对应的待回答问题;

对待识别图像进行文本识别,将识别得到的多个词组添加至视觉问答模型的通用答案集中,获取目标答案集,所述视觉问答模型是预先训练的用于根据图像特征,文本特征以及所述目标答案集获取所述待回答问题对应的答案的模型;

对所述待识别图像和所述待回答问题分别进行特征提取,得到图像特征和第一文本特征;

将所述图像特征和所述第一文本特征,输入至所述视觉问答模型中,得到第一答案。

5.根据权利要求4所述的方法,其特征在于,所述视觉问答模型包括第一子视觉问答模

型和第二子视觉问答模型；

所述第一子视觉问答模型用于根据所述图像特征，所述第一文本特征和目标答案集，输出多个第二答案，所述多个第二答案属于所述目标答案集；

所述第一子视觉问答模型用于根据所述图像特征和第二文本特征，输出所述第一答案，所述第二文本特征是对所述第一文本特征和多个第二答案进行拼接得到的。

6.根据权利要求4所述的方法，其特征在于，所述对所述待识别图像和所述待回答问题分别进行特征提取，得到图像特征和第一文本特征，包括：

将所述待识别图像输入区域卷积神经网络进行特征提取，获取所述图像特征；

将所述待回答问题输入长短期记忆网络进行特征提取，获取所述第一文本特征。

7.一种视觉问答模型的训练装置，其特征在于，包括：

获取模块，用于获取训练样本集，所述训练样本集中包括多个样本图像，每个样本图像对应的样本问题以及答案；

提取模块，用于针对所述样本训练集中的每个样本图像和对应样本问题分别进行特征提取，获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征；

训练模块，用于根据通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案，对自顶向下与自底向上BUTD注意力模型进行模型训练，获取视觉问答模型，所述视觉问答模型用于根据图像特征，文本特征以及所述通用答案集获取待识别图像对应的待回答问题的答案。

8.根据权利要求7所述的装置，其特征在于，所述提取模块，具体用于：

将所述样本训练集中的每个样本图像输入区域卷积神经网络进行特征提取，获取所述样本图像特征；

将所述每个样本图像对应的样本问题输入长短期记忆网络进行特征提取，获取所述第一样本文本特征。

9.根据权利要求7或8所述的装置，其特征在于，所述BUTD注意力模型包括第一子BUTD注意力模型和第二子BUTD注意力模型，所述训练模块，具体用于：

根据所述通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案，依次对所述第一子BUTD注意力模型和所述第二子BUTD注意力模型进行训练，获取所述视觉问答模型，所述视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型；

所述第一子视觉问答模型是根据所述通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案对所述第一子BUTD注意力模型进行训练得到的，所述第一子BUTD注意力模型用于输出每个图像特征对应的多个样本答案，所述多个样本答案属于所述通用答案集；

所述第二子视觉问答模型是根据每个图像特征，以及每个图像特征对应的第二样本文本特征和答案对所述第二子BUTD注意力模型进行训练得到的，所述第二样本文本特征是对第一样本文本特征和对应的多个第一样本答案进行拼接得到的。

10.一种视觉问答装置，其特征在于，包括：

获取模块，用于获取待识别的图像以及对应的待回答问题；

处理模块，用于对待识别图像进行文本识别，将识别得到的多个词组添加至视觉问答模型的通用答案集中，获取目标答案集，所述视觉问答模型是预先训练的用于根据图像特

征，文本特征以及所述目标答案集获取所述待回答问题对应的答案的模型；

提取模块，用于对所述待识别图像和所述待回答问题分别进行特征提取，得到图像特征和第一文本特征；

输入模块，用于将所述图像特征和所述第一文本特征，输入至所述视觉问答模型中，得到第一答案。

11.根据权利要求10所述的装置，其特征在于，所述视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型；

所述第一子视觉问答模型用于根据所述图像特征，所述第一文本特征和目标答案集，输出多个第二答案，所述多个第二答案属于所述目标答案集；

所述第一子视觉问答模型用于根据所述图像特征和第二文本特征，输出所述第一答案，所述第二文本特征是对所述第一文本特征和多个第二答案进行拼接得到的。

12.根据权利要求10所述的装置，其特征在于，提取模块，具体用于：

将所述待识别图像输入区域卷积神经网络进行特征提取，获取所述图像特征；

将所述待回答问题输入长短期记忆网络进行特征提取，获取所述第一文本特征。

13.一种电子设备，包括：处理器、存储器及存储在所述存储器上并可在处理器上运行的计算机程序指令，其特征在于，所述处理器执行所述计算机程序指令时用于实现如权利要求1至6任一项所述的方法。

14.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有计算机执行指令，所述计算机执行指令被处理器执行时用于实现如权利要求1至6任一项所述的方法。

15.一种计算机程序产品，包括计算机程序，其特征在于，所述计算机程序被处理器执行时用于实现如权利要求1至6任一项所述的方法。

# 视觉问答模型的训练方法和视觉问答方法

## 技术领域

[0001] 本申请涉及计算机技术领域,尤其涉及一种视觉问答模型的训练方法和视觉问答方法。

## 背景技术

[0002] 视觉问答(Visual Question Answering,VQA)是一种涉及计算机视觉和自然语言处理的学习任务,在智能对话机器人、为视障者获取视觉信息、视觉导航等领域中得到了广泛的应用。

[0003] 目前,VQA主要在于通过视觉问答模型对输入的图像和相关的文字问题进行处理,从通用答案集中确定出符合逻辑和语言规则的答案,并将该答案进行输出。

[0004] 然而,虽然通用答案集中涵盖了大部分问题的答案,但在答案来源于图像本身而不存在于通用答案库中时,存在输出的答案的准确度较低的问题。

## 发明内容

[0005] 本申请提供一种视觉问答模型的训练方法和视觉问答方法,以解决输出的答案的准确度较低的问题。

[0006] 第一方面,本申请实施例提供一种视觉问答模型的训练方法,包括:

[0007] 获取训练样本集,所述训练样本集中包括多个样本图像,每个样本图像对应的样本问题以及答案;

[0008] 针对所述样本训练集中的每个样本图像和对应样本问题分别进行特征提取,获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征;

[0009] 根据通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特征和答案,对BUTD注意力模型进行模型训练,获取视觉问答模型,所述视觉问答模型用于根据图像特征,文本特征以及所述通用答案集获取待识别图像对应的待回答问题的答案。

[0010] 在第一方面的一种可能设计中,所述针对所述样本训练集中的每个样本图像和对应样本问题分别进行特征提取,获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征,包括:

[0011] 将所述样本训练集中的每个样本图像输入区域卷积神经网络进行特征提取,获取所述样本图像特征;

[0012] 将所述每个样本图像对应的样本问题输入长短期记忆网络进行特征提取,获取所述第一样本文本特征。

[0013] 在第一方面的另一种可能设计中,所述BUTD注意力模型包括第一子BUTD注意力模型和第二子BUTD注意力模型,所述根据通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特征和答案,对自顶向下与自底向上BUTD注意力模型进行模型训练,获取视觉问答模型,包括:

[0014] 根据所述通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特

征和答案，依次对所述第一子BUTD注意力模型和所述第二子BUTD注意力模型进行训练，获取所述视觉问答模型，所述视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型；

[0015] 所述第一子视觉问答模型是根据所述通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案对所述第一子BUTD注意力模型进行训练得到的，所述第一子BUTD注意力模型用于输出每个图像特征对应的多个样本答案，所述多个样本答案属于所述通用答案集；

[0016] 所述第二子视觉问答模型是根据每个图像特征，以及每个图像特征对应的第二样本文本特征和答案对所述第二子BUTD注意力模型进行训练得到的，所述第二样本文本特征是对第一样本文本特征和对应的多个第一样本答案进行拼接得到的。

[0017] 第二方面，本申请实施例提供一种视觉问答方法，包括：

[0018] 获取待识别的图像以及对应的待回答问题；

[0019] 对待识别图像进行文本识别，将识别得到的多个词组添加至视觉问答模型的通用答案集中，获取目标答案集，所述视觉问答模型是预先训练的用于根据图像特征，文本特征以及所述目标答案集获取所述待回答问题对应的答案的模型；

[0020] 对所述待识别图像和所述待回答问题分别进行特征提取，得到图像特征和第一文本特征；

[0021] 将所述图像特征和所述第一文本特征，输入至所述视觉问答模型中，得到第一答案。

[0022] 在第二方面的一种可能设计中，所述视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型；

[0023] 所述第一子视觉问答模型用于根据所述图像特征，所述第一文本特征和目标答案集，输出多个第二答案，所述多个第二答案属于所述目标答案集；

[0024] 所述第一子视觉问答模型用于根据所述图像特征和第二文本特征，输出所述第一答案，所述第二文本特征是对所述第一文本特征和多个第二答案进行拼接得到的。

[0025] 在第二方面的另一种可能设计中，所述对所述待识别图像和所述待回答问题分别进行特征提取，得到图像特征和第一文本特征，包括：

[0026] 将所述待识别图像输入区域卷积神经网络进行特征提取，获取所述图像特征；

[0027] 将所述待回答问题输入长短期记忆网络进行特征提取，获取所述第一文本特征。

[0028] 第三方面，本申请实施例提供一种视觉问答模型的训练装置，包括：

[0029] 获取模块，用于获取训练样本集，所述训练样本集中包括多个样本图像，每个样本图像对应的样本问题以及答案；

[0030] 提取模块，用于针对所述样本训练集中的每个样本图像和对应样本问题分别进行特征提取，获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征；

[0031] 训练模块，用于根据通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案，对BUTD注意力模型进行模型训练，获取视觉问答模型，所述视觉问答模型用于根据图像特征，文本特征以及所述通用答案集获取待识别图像对应的待回答问题的答案。

[0032] 在第三方面的一种可能设计中，所述提取模块，具体用于：

[0033] 将所述样本训练集中的每个样本图像输入区域卷积神经网络进行特征提取，获取所述样本图像特征；

[0034] 将所述每个样本图像对应的样本问题输入长短期记忆网络进行特征提取，获取所述第一样本文本特征。

[0035] 在第三方面的另一种可能设计中，所述BUTD注意力模型包括第一子BUTD注意力模型和第二子BUTD注意力模型，所述训练模块，具体用于：

[0036] 根据所述通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案，依次对所述第一子BUTD注意力模型和所述第二子BUTD注意力模型进行训练，获取所述视觉问答模型，所述视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型；

[0037] 所述第一子视觉问答模型是根据所述通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案对所述第一子BUTD注意力模型进行训练得到的，所述第一子BUTD注意力模型用于输出每个图像特征对应的多个样本答案，所述多个样本答案属于所述通用答案集；

[0038] 所述第二子视觉问答模型是根据每个图像特征，以及每个图像特征对应的第二样本文本特征和答案对所述第二子BUTD注意力模型进行训练得到的，所述第二样本文本特征是对第一样本文本特征和对应的多个第一样本答案进行拼接得到的。

[0039] 第四方面，本申请实施例提供一种视觉问答装置，包括：

[0040] 获取模块，用于获取待识别的图像以及对应的待回答问题；

[0041] 处理模块，用于对待识别图像进行文本识别，将识别得到的多个词组添加至视觉问答模型的通用答案集中，获取目标答案集，所述视觉问答模型是预先训练的用于根据图像特征，文本特征以及所述目标答案集获取所述待回答问题对应的答案的模型；

[0042] 提取模块，用于对所述待识别图像和所述待回答问题分别进行特征提取，得到图像特征和第一文本特征；

[0043] 输入模块，用于将所述图像特征和所述第一文本特征，输入至所述视觉问答模型中，得到第一答案。

[0044] 在第四方面的一种可能设计中，所述视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型；

[0045] 所述第一子视觉问答模型用于根据所述图像特征，所述第一文本特征和目标答案集，输出多个第二答案，所述多个第二答案属于所述目标答案集；

[0046] 所述第一子视觉问答模型用于根据所述图像特征和第二文本特征，输出所述第一答案，所述第二文本特征是对所述第一文本特征和多个第二答案进行拼接得到的。

[0047] 在第四方面的另一种可能设计中，提取模块，具体用于：

[0048] 将所述待识别图像输入区域卷积神经网络进行特征提取，获取所述图像特征；

[0049] 将所述待回答问题输入长短期记忆网络进行特征提取，获取所述第一文本特征。

[0050] 第五方面，本申请实施例提供一种电子设备，包括：处理器、存储器及存储在所述存储器上并可在处理器上运行的计算机程序指令，所述处理器执行所述计算机程序指令时用于实现第一方面、第二方面以及在第一方面和第二方面中各可能设计提供的方法。

[0051] 第六方面，本申请实施例可提供一种计算机可读存储介质，所述计算机可读存储

介质中存储有计算机执行指令，所述计算机执行指令被处理器执行时用于实现第一方面、第二方面以及在第一方面和第二方面中各可能设计提供的方法。

[0052] 第七方面，本申请实施例提供一种计算机程序产品，包括计算机程序，所述计算机程序被处理器执行时用于实现第一方面、第二方面以及在第一方面和第二方面中各可能设计提供的方法。

[0053] 本申请实施例提供的视觉问答模型的训练方法和视觉问答方法，该视觉问答方法包括：通过获取待识别的图像以及对应的待回答问题，对待识别图像进行文本识别，将识别得到的多个词组添加至视觉问答模型的通用答案集中，获取目标答案集，对待识别图像和待回答问题分别进行特征提取，得到图像特征和第一文本特征，将图像特征和第一文本特征，输入至视觉问答模型中，得到第一答案。通过将待识别图像中的多个词组添加至通用答案集中，使得利用视觉问答模型获取待回答问题对应的答案时，不仅考虑了通用答案集中的答案，还考虑了待识别图像中的词组对待回答问题的影响，有效提高了输出的答案的准确度。

## 附图说明

[0054] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。

[0055] 图1为本申请实施例提供的视觉问答模型的训练方法的一种应用场景示意图；

[0056] 图2为本申请实施例提供的视觉问答模型的训练方法实施例一的流程示意图；

[0057] 图3为本申请实施例提供的样本图像示意图；

[0058] 图4为本申请实施例提供的视觉问答方法实施例一的流程示意图；

[0059] 图5为本申请实施例提供的视觉问答模型的训练装置的结构示意图；

[0060] 图6为本申请实施例提供的视觉问答装置的结构示意图；

[0061] 图7为本申请实施例提供的电子设备的结构示意图。

[0062] 通过上述附图，已示出本公开明确的实施例，后文中将有更详细的描述。这些附图和文字描述并不是为了通过任何方式限制本公开构思的范围，而是通过参考特定实施例为本领域技术人员说明本公开的概念。

## 具体实施方式

[0063] 为使本申请实施例的目的、技术方案和优点更加清楚，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作出创造性劳动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。

[0064] 在介绍本申请的实施例之前，首先对本申请实施例的应用背景进行解释：

[0065] VQA问题是一个同时涉及计算机视觉和自然语言处理等领域的多学科的人工智能领域问题，随着视觉和自然语言交叉领域的发展得到广泛关注。视觉问答问题的任务在于通过模型训练得到视觉问答模型，视觉问答模型能够处理输入的图像和相关的文字问题，从而输出一个符合逻辑和语言规则的答案。相较于文本问答(Question Answering，QA)任务，VQA中图像的抽象程度较低，更难被计算机理解；相较于图像字幕(英文：image

captioning)任务,视觉问答不是简单地将图像“翻译”为文本,而是需要更加深入地理解图像内容。也就是说,与文本QA和图像字幕相比,VQA的实现难度更大。

[0066] 目前,视觉问答领域的常用模型分为联合嵌入模型和注意力模型。其中,注意力模型能够将算法的注意力聚焦于与输入问题最相关的图像区域,依据图像中的物体或文本的重要和相关程度分配不同的权重,使得重要的信息获得较大的权值,算法能够更加准确地捕捉关键信息,从而提升模型的专注力和算法能力。因此,注意力模型得到了更多的关注。

[0067] 然而,虽然通用答案集中涵盖了大部分问题的答案,但在答案来源于图像本身而不存在于答案库中时,存在输出的答案的准确度较低的问题。

[0068] 针对上述问题,本申请的发明构思如下:示例性的,假设问题为“图中商品的品牌是什么”或者“前方交通告示牌上写的是什么”,该问题的答案直接来源于图像本身,而通用答案集一般不会涵盖品牌名称或交通专业术语等等的词汇,因此无法从通用答案集确认出问题的答案,存在输出的答案的准确度较低的问题。基于此,发明人发现,如果在对待识别图像进行视觉问答之前,对待识别图像进行文本识别,将识别得到的多个词组添加至视觉问答模型的通用答案集中,从而获取目标答案集,就能够在对待识别图像进行视觉问答时,从目标答案集中获取待回答问题的答案,就能解决现有技术中输出的答案的准确度较低的问题。

[0069] 示例性的,本申请实施例提供的视觉问答模型的训练方法可以应用于图1所示的一种应用场景示意图中。图1为本申请实施例提供的视觉问答模型的训练方法的一种应用场景示意图,用以解决上述技术问题。如图1所示,该应用场景可以包括:电子设备11和服务器12,还可以包括与电子设备11连接的数据存储设备13。

[0070] 在本实施例中,电子设备11既可以从网络上获取训练样本集,自顶向下与自底向上(Buttom-up and top-down,BUTD)注意力模型和通用答案集,还可以通过网络从服务器12中获取训练样本集,BUTD注意力模型和通用答案集,并将其存储至数据存储设备13中,以便于后续对BUTD注意力模型进行训练时直接使用。

[0071] 可选的,训练样本集,BUTD注意力模型和通用答案集还可以是预先存储在电子设备11中,电子设备11直接通过存储地址对其进行获取。

[0072] 进一步的,电子设备11可以根据训练样本集对BUTD注意力模型进行模型训练,从而得到视觉问答模型,从而将视觉问答模型存储至数据存储设备13中。

[0073] 可以理解的是,本申请实施例的执行主体可以是终端设备,例如,计算机、平板电脑等,也可以是服务器,例如,后台的处理平台等。因而,本实施例以终端设备和服务器统称为电子设备进行解释说明,关于该电子设备具体为终端设备,还是服务器,其可以实际情况确定。

[0074] 下面,通过具体实施例对本申请的技术方案进行详细说明。

[0075] 需要说明的是,下面这几个具体的实施例可以相互结合,对于相同或相似的概念或过程可能在某些实施例中不再赘述。

[0076] 图2为本申请实施例提供的视觉问答模型的训练方法实施例一的流程示意图。如图2所示,该视觉问答模型的训练方法可以包括如下步骤:

[0077] S21、获取训练样本集。

[0078] 其中,训练样本集中包括多个样本图像,每个样本图像对应的样本问题以及答案。

[0079] 其中,可以从网络获取训练样本集,可以通过网络从与电子设备连接的数据存储设备中获取训练样本集,还可以从电子设备中存储训练样本集的存储位置获取训练样本集。

[0080] 示例性的,图3为本申请实施例提供的样本图像示意图。如图3所示,该样本图像对应的样本问题可以为“图中共有几个人物”,则对应的答案为“2个”;该样本图像对应的样本问题还可以为“图中有几个足球”,则对应的答案为“1个”。

[0081] S22、针对样本训练集中的每个样本图像和对应样本问题分别进行特征提取,获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征。

[0082] 在一种具体的实现中,将样本训练集中的每个样本图像输入区域卷积神经网络(Region-based Convolutional Neural Network,R-CNN)进行特征提取,获取样本图像特征,同时,将每个样本图像对应的样本问题输入长短期记忆网络(Long Short-Term Memory,LSTM)进行特征提取,获取第一样本文本特征。

[0083] 其中,区域卷积神经网络可以通过Faster R-CNN实现。

[0084] S23、根据通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特征和答案,对BUTD注意力模型进行模型训练,获取视觉问答模型。

[0085] 其中,视觉问答模型用于根据图像特征,文本特征以及通用答案集获取待识别图像对应的待回答问题的答案。

[0086] 其中,可以预先从网络获取通用答案集,可以通过网络从与电子设备连接的数据存储设备中获取通用答案集,还可以从电子设备中存储通用答案集的存储位置获取通用答案集。

[0087] 示例性的,通用答案集可以是现有技术中VQA的通用答案集,还可以是对通用答案集进行增加答案或者添加答案获取的,本申请实施例对此不进行具体限制。

[0088] 在一种具体的实现方式中,BUTD注意力模型包括第一子BUTD注意力模型和第二子BUTD注意力模型,则电子设备可以根据通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特征和答案,依次对第一子BUTD注意力模型和第二子BUTD注意力模型进行训练,获取视觉问答模型,视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型。

[0089] 在该方式下,第一子视觉问答模型是根据通用答案集,每个图像特征,以及每个图像特征对应的第一样本文本特征和答案对第一子BUTD注意力模型进行训练得到的,第一子BUTD注意力模型用于输出每个图像特征对应的多个样本答案,多个样本答案属于通用答案集。

[0090] 在该方式下,第二子视觉问答模型是根据每个图像特征,以及每个图像特征对应的第二样本文本特征和答案对第二子BUTD注意力模型进行训练得到的,第二样本文本特征是对第一样本文本特征和对应的多个第一样本答案进行拼接得到的。

[0091] 在实际训练过程中,电子设备的操作系统可以为ubuntu16.04操作系统,图形处理器(Graphics Processing Unit,GPU)为NVIDIA1080Ti,内存容量为12G。BUTD注意力模型可以使用python 3.6语言和Pytorch 0.4.1框架进行搭建的得到的。在获取BUTD注意力模型后,根据预设训练批次,预设衰减规则以及预设学习速率,同时搭配了cuda9.1、cuDNN 7.0.5对BUTD注意力模型进行训练。举例来说,预设训练批次可以为256,预设衰减规则可以为每50个轮次衰减一半,预设学习速率可以为1e-3(即$10^{-3}$)。

[0092] 可选的，还可以使用AdamMax优化器对BUTD注意力模型进行训练，从而达到对数据进行快速拟合的效果。

[0093] 示例性的，在实际用中，根据上述方法得到的视觉问答模型与现有存在的模型相比，通过该视觉问答模型得到的答案的准确率高于其他的模型。通过视觉问答模型与现有存在的模型(如Baseline、NUTAN、MLB、DA-NTN、MANet)得到的答案的准确率可以如表1所示。

[0094] 表1

[0095]

| 问题类型 / 准确率/% / 模型 | 是否问题 | 数量问题 | 其他问题 | 所有类型问题 |
|---|---|---|---|---|
| Baseline | 81.82 | 44.21 | 56.06 | 65.32 |
| NUTAN | 82.88 | 44.54 | 44.54 | 66.01 |
| MLB | 83.58 | 44.92 | 56.34 | 66.27 |
| DA-NTN | 84.29 | 47.14 | 57.92 | 67.56 |
| MANet | 85.002 | 45.27 | 58.68 | 69.03 |
| 视觉问答模型 | 86.13 | 45.94 | 58.88 | 69.35 |

[0096] 如表1所示，在是否问题、其他问题中，视觉问答模型的准确度高于其他模型。总体来看，对于所有类型问题来说，视觉问答模型的准确度依旧高于其他模型。

[0097] 应理解，表1中的准确率是基于实验获取的，在本申请中，由于不同实验使用的实验数据不同，不同实验数据对应的准确率可能存在变化。

[0098] 本申请提供的视觉问答模型的训练方法，通过获取训练样本集，针对样本训练集中的每个样本图像和对应样本问题分别进行特征提取，获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征，根据通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案，对BUTD注意力模型进行模型训练，获取视觉问答模型。通过上述方法得到的视觉问答模型，在使用该视觉问答模型进行视觉问答时，有效的提高了该视觉问答模型输出的答案的准确度。

[0099] 在得到上述视觉问答模型之后，可以使用该视觉问答模型，对待识别的图像以及对应的待回答问题进行视觉问答。下面结合具体地实施例对使用该视觉问答模型，对待识别的图像以及对应的待回答问题进行视觉问答的方法进行详细说明。下面这几个具体的实施例可以相互结合，对于相同或相似的概念或过程可能在某些实施例不再赘述。

[0100] 具体实现时，视觉问答方法的执行主体也可以为终端或者服务器等具有处理能力的电子设备。应理解，该执行视觉问答方法的电子设备，与，执行上述视觉问答模型的训练方法的电子设备，可以是同一设备，也可以是不同设备。

[0101] 图4为本申请实施例提供的视觉问答方法实施例一的流程示意图。如图4所示，该视觉问答方法可以包括如下步骤：

[0102] S41、获取待识别的图像以及对应的待回答问题。

[0103] 其中，可以从网络获取待识别的图像以及对应的待回答问题，可以通过网络从与电子设备连接的数据存储设备中获取待识别的图像以及对应的待回答问题，还可以从电子

设备中存储训练样本集的存储位置获取待识别的图像以及对应的待回答问题。

[0104] 其中，待识别的图像和待回答问题可以为一一对应的关系，也可以为一对多的关系。也就是说，一个待识别的图像可以对应于一个待回答问题，也可以对应于多个待回答问题。

[0105] S42、对待识别图像进行文本识别，将识别得到的多个词组添加至视觉问答模型的通用答案集中，获取目标答案集。

[0106] 其中，视觉问答模型是预先训练的用于根据图像特征，文本特征以及目标答案集获取待回答问题对应的答案的模型。

[0107] 在一种具体的实现方式中，可以将待识别图像输入预先训练的文本识别模型(如Rosseta)进行文本识别，获取M个词组(可以为表示为$S_1$，$S_2$，…，$S_M$)，在将M个词组作为新添加的答案，拼接至通用答案集S(N)中，获取目标答案集(可以为表示为$S_1$，$S_2$，…，$S_N$，$S_{N+1}$，...，$S_{N+M}$)。

[0108] 可选的，还可以通过其他的现有方式对待识别图像进行文本识别，本申请实施例对此不进行具体限制。

[0109] S43、对待识别图像和待回答问题分别进行特征提取，得到图像特征和第一文本特征。

[0110] 在一种具体的实现方式中，将待识别图像输入区域卷积神经网络进行特征提取，获取图像特征，之后将待回答问题输入长短期记忆网络进行特征提取，获取第一文本特征。

[0111] 可选的，还可以通过其他的现有方式对待识别图像和待回答问题进行特征提取，本申请实施例对此不进行具体限制。

[0112] S44、将图像特征和第一文本特征，输入至视觉问答模型中，得到第一答案。

[0113] 在一种具体的实现方式中，视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型；

[0114] 在该方式下，第一子视觉问答模型用于根据图像特征，第一文本特征和目标答案集，输出多个第二答案，多个第二答案属于目标答案集。

[0115] 其中，第一子视觉问答模型根据获取的图像特征和第一文本特征，计算图像特征对第一文本特征的注意力权重系数，可以通过下述公式计算获取：$\alpha = f_{attention}(V, Q_1)$。其中，α为注意力权重系数，V为图像特征，$Q_1$为第一文本特征。之后，为第一文本特征分配注意力权重系数，得到处理后的第一文本特征，处理后的第一文本特征可以通过下述公式表示：$Q_{1\alpha} = Q \odot \alpha$，其中，$Q_{1\alpha}$为处理后的第一文本特征。

[0116] 其中，第一子视觉问答模型共有三层全连接层，将跨模态特征$v \cdot q_{1\alpha}$依次输入至三层全连接层中，最后一层的全连接层的维度与目标答案集一致。示例性的，假设目标答案集中共有N+M个答案，则最后一层的全连接层的维度为N+M。之后，将最后一层的输出进行指数归一化处理，获取目标答案集中每个答案的概率，按照概率由大到小的顺序对该目标答案集中的每个答案进行排序，根据排序顺序选取不重复的前十个答案作为第二答案进行输出。

[0117] 应理解，输出的第二答案的个数可以根据实际情况进行设定，可以预先设置输出前15个、前20个、前20个答案作为第二答案，本申请实施例对此不进行具体限制。

[0118] 第一子视觉问答模型用于根据图像特征和第二文本特征，输出第一答案，第二文

本特征是对第一文本特征和多个第二答案进行拼接得到的。举例来说，假设第二答案的个数为10个，则第二文本特征可以通过公式：$Q_2$＝LSTM([q:$a_1$:…:$a_{10}$])，表示。其中，$Q_2$为第二文本特征，q为第一文本特征中的各元素，$a_1$，……，$a_{10}$为第二答案。

[0119] 其中，第二子视觉问答模型计算图像特征对第二文本特征的注意力权重系数，并为第二文本特征分配注意力权重系数，得到处理后的第二文本特征。具体的实现步骤可以参考上述第一子视觉问答模型的实现过程，在此不再赘述

[0120] 其中，第二子视觉问答模型共有三层全连接层，将跨模态特征$v \cdot q_{2\alpha}$依次输入至三层全连接层中，最后一层的全连接层的维度与第二答案一致，其中，$q_{2\alpha}$为处理后的第二文本特征中的各元素。之后，使用Softmax函数或Sigmoid函数计算每个第二答案的概率，将第二答案中概率最大的答案确定为第一答案。

[0121] 本申请实施例提供的视觉问答方法，通过获取待识别的图像以及对应的待回答问题，对待识别图像进行文本识别，将识别得到的多个词组添加至视觉问答模型的通用答案集中，获取目标答案集，对待识别图像和待回答问题分别进行特征提取，得到图像特征和第一文本特征，将图像特征和第一文本特征，输入至视觉问答模型中，得到第一答案。通过将待识别图像中的多个词组添加至通用答案集中，使得利用视觉问答模型获取待回答问题对应的答案时，不仅考虑了通用答案集中的答案，还考虑了待识别图像中的词组对待回答问题的影响，有效提高了输出的答案的准确度。

[0122] 进一步的，现有技术中的视觉问答大部分都为搜索式视觉问答，类似于分类问题，每个问题对应有各自的候选答案集，其中候选答案集是从数量极其庞大的通用答案集中选取出来频次较高的多个答案作为候选答案集。在对答案进行预测时，视觉问答模型对图像特征和问题对应的文本特征进行融合，对对应的候选答案集中的答案进行概率预测，选取概率最大的答案作为预测答案。然而，在对通用答案集进行分类确定候选集的过程中，往往因为分类类别过多而导致预测概率误差较大，候选答案集中大部分都是与具体问题无关的常见选项。

[0123] 在上述实施例中，通过第一子视觉问答模型对目标答案集进行初次筛选，筛选出与具体问题相关的多个第二答案，在通过第一子视觉问答模型对筛选出的第二答案进行二次筛选，从第二答案中确定出待回答问题对应的第一答案，有效的提高了获取的答案的准确性。

[0124] 下述为本申请装置实施例，可以用于执行本申请方法实施例。对于本申请装置实施例中未披露的细节，请参照本申请方法实施例。

[0125] 图5为本申请实施例提供的视觉问答模型的训练装置的结构示意图。如图5所示，该视觉问答模型的训练装置包括：

[0126] 获取模块51，用于获取训练样本集，训练样本集中包括多个样本图像，每个样本图像对应的样本问题以及答案；

[0127] 提取模块52，用于针对样本训练集中的每个样本图像和对应样本问题分别进行特征提取，获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征；

[0128] 训练模块53，用于根据通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案，对BUTD注意力模型进行模型训练，获取视觉问答模型，视觉问答模型用于根据图像特征，文本特征以及通用答案集获取待识别图像对应的待回答问题的答案。

[0129] 在本申请实施例的一种可能设计中，提取模块52，具体用于：

[0130] 将样本训练集中的每个样本图像输入区域卷积神经网络进行特征提取，获取样本图像特征；

[0131] 将每个样本图像对应的样本问题输入长短期记忆网络进行特征提取，获取第一样本文本特征。

[0132] 在本申请实施例的另一种可能设计中，

[0133] BUTD注意力模型包括第一子BUTD注意力模型和第二子BUTD注意力模型，训练模块53，具体用于：

[0134] 根据通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案，依次对第一子BUTD注意力模型和第二子BUTD注意力模型进行训练，获取视觉问答模型，视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型；

[0135] 第一子视觉问答模型是根据通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案对第一子BUTD注意力模型进行训练得到的，第一子BUTD注意力模型用于输出每个图像特征对应的多个样本答案，多个样本答案属于通用答案集；

[0136] 第二子视觉问答模型是根据每个图像特征，以及每个图像特征对应的第二样本文本特征和答案对第二子BUTD注意力模型进行训练得到的，第二样本文本特征是对第一样本文本特征和对应的多个第一样本答案进行拼接得到的。

[0137] 本申请实施例提供的视觉问答模型的训练装置，可用于执行上述任一实施例中的视觉问答模型的训练方法，其实现原理和技术效果类似，在此不再赘述。

[0138] 图6为本申请实施例提供的视觉问答装置的结构示意图。如图6所示，该视觉问答装置包括：

[0139] 获取模块61，用于获取待识别的图像以及对应的待回答问题；

[0140] 处理模块62，用于对待识别图像进行文本识别，将识别得到的多个词组添加至视觉问答模型的通用答案集中，获取目标答案集，视觉问答模型是预先训练的用于根据图像特征，文本特征以及目标答案集获取待回答问题对应的答案的模型；

[0141] 提取模块63，用于对待识别图像和待回答问题分别进行特征提取，得到图像特征和第一文本特征；

[0142] 输入模块，用于将图像特征和第一文本特征，输入至视觉问答模型中，得到第一答案。

[0143] 在本申请实施例的一种可能设计中，视觉问答模型包括第一子视觉问答模型和第二子视觉问答模型；

[0144] 第一子视觉问答模型用于根据图像特征，第一文本特征和目标答案集，输出多个第二答案，多个第二答案属于目标答案集；

[0145] 第一子视觉问答模型用于根据图像特征和第二文本特征，输出第一答案，第二文本特征是对第一文本特征和多个第二答案进行拼接得到的。

[0146] 在本申请实施例的另一种可能设计中，提取模块63，具体用于：

[0147] 将待识别图像输入区域卷积神经网络进行特征提取，获取图像特征；

[0148] 将待回答问题输入长短期记忆网络进行特征提取，获取第一文本特征。

[0149] 本申请实施例提供的视觉问答装置，可用于执行上述任一实施例中的视觉问答方

法,其实现原理和技术效果类似,在此不再赘述。

[0150] 需要说明的是,应理解以上装置的各个模块的划分仅仅是一种逻辑功能的划分,实际实现时可以全部或部分集成到一个物理实体上,也可以物理上分开。且这些模块可以全部以软件通过处理元件调用的形式实现;也可以全部以硬件的形式实现;还可以部分模块通过处理元件调用软件的形式实现,部分模块通过硬件的形式实现。此外,这些模块全部或部分可以集成在一起,也可以独立实现。这里所述的处理元件可以是一种集成电路,具有信号的处理能力。在实现过程中,上述方法的各步骤或以上各个模块可以通过处理器元件中的硬件的集成逻辑电路或者软件形式的指令完成。

[0151] 图7为本申请实施例提供的电子设备的结构示意图。如图7所示,该电子设备11可以包括:处理器71、存储器72及存储在所述存储器72上并可在处理器71上运行的计算机程序指令,所述处理器71执行所述计算机程序指令时实现前述任一实施例提供的视觉问答模型的训练方法或视觉问答方法。

[0152] 可选的,该电子设备11的上述各个器件之间可以通过系统总线连接。

[0153] 存储器72可以是单独的存储单元,也可以是集成在处理器中的存储单元。处理器的数量为一个或者多个。

[0154] 可选的,电子设备11还可以包括与其他设备进行交互的接口。

[0155] 应理解,处理器71可以是中央处理单元(Central Processing Unit,CPU),还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本申请所公开的方法的步骤可以直接体现为硬件处理器执行完成,或者用处理器中的硬件及软件模块组合执行完成。

[0156] 系统总线可以是外设部件互连标准(peripheral component interconnect,PCI)总线或扩展工业标准结构(extended industry standard architecture,EISA)总线等。系统总线可以分为地址总线、数据总线、控制总线等。为便于表示,图中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。存储器可能包括随机存取存储器(random access memory,RAM),也可能还包括非易失性存储器(non-volatile memory,NVM),例如至少一个磁盘存储器。

[0157] 实现上述各方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成。前述的程序可以存储于一可读取存储器中。该程序在执行时,执行包括上述各方法实施例的步骤;而前述的存储器(存储介质)包括:只读存储器(read-only memory,ROM)、RAM、快闪存储器、硬盘、固态硬盘、磁带(英文:magnetic tape)、软盘(英文:floppy disk)、光盘(英文:optical disc)及其任意组合。

[0158] 本申请实施例提供的电子设备,可用于执行上述任一方法实施例提供的视觉问答模型的训练方法或视觉问答方法,其实现原理和技术效果类似,在此不再赘述。

[0159] 本申请实施例提供一种计算机可读存储介质,该计算机可读存储介质中存储有计算机指令,当该计算机指令在计算机上运行时,使得计算机执行上述视觉问答模型的训练方法或视觉问答方法。

[0160] 上述的计算机可读存储介质,上述可读存储介质可以是由任何类型的易失性或非易失性存储设备或者它们的组合实现,如静态随机存取存储器,电可擦除可编程只读存储

器,可擦除可编程只读存储器,可编程只读存储器,只读存储器,磁存储器,快闪存储器,磁盘或光盘。可读存储介质可以是通用或专用计算机能够存取的任何可用介质。

[0161] 可选的,将可读存储介质耦合至处理器,从而使处理器能够从该可读存储介质读取信息,且可向该可读存储介质写入信息。当然,可读存储介质也可以是处理器的组成部分。处理器和可读存储介质可以位于专用集成电路(Application Specific Integrated Circuits,ASIC)中。当然,处理器和可读存储介质也可以作为分立组件存在于设备中。

[0162] 本申请实施例还提供一种计算机程序产品,该计算机程序产品包括计算机程序,该计算机程序存储在计算机可读存储介质中,至少一个处理器可以从该计算机可读存储介质中读取该计算机程序,所述至少一个处理器执行所述计算机程序时可实现上述视觉问答模型的训练方法或视觉问答方法。

[0163] 应当理解的是,本公开并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本公开的范围仅由所附的权利要求书来限制。

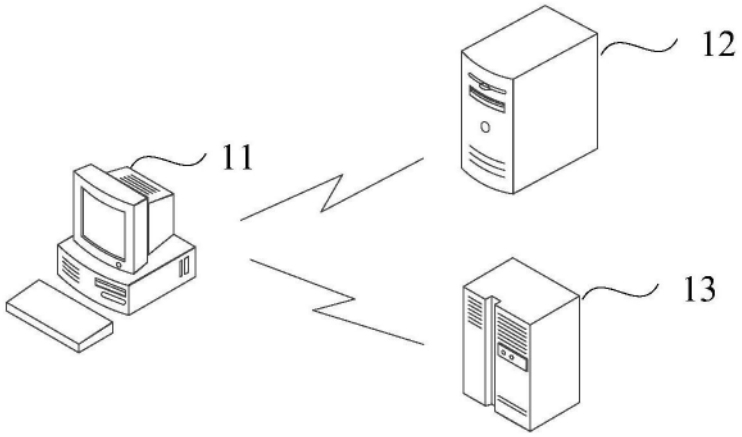

12
11
13

图1

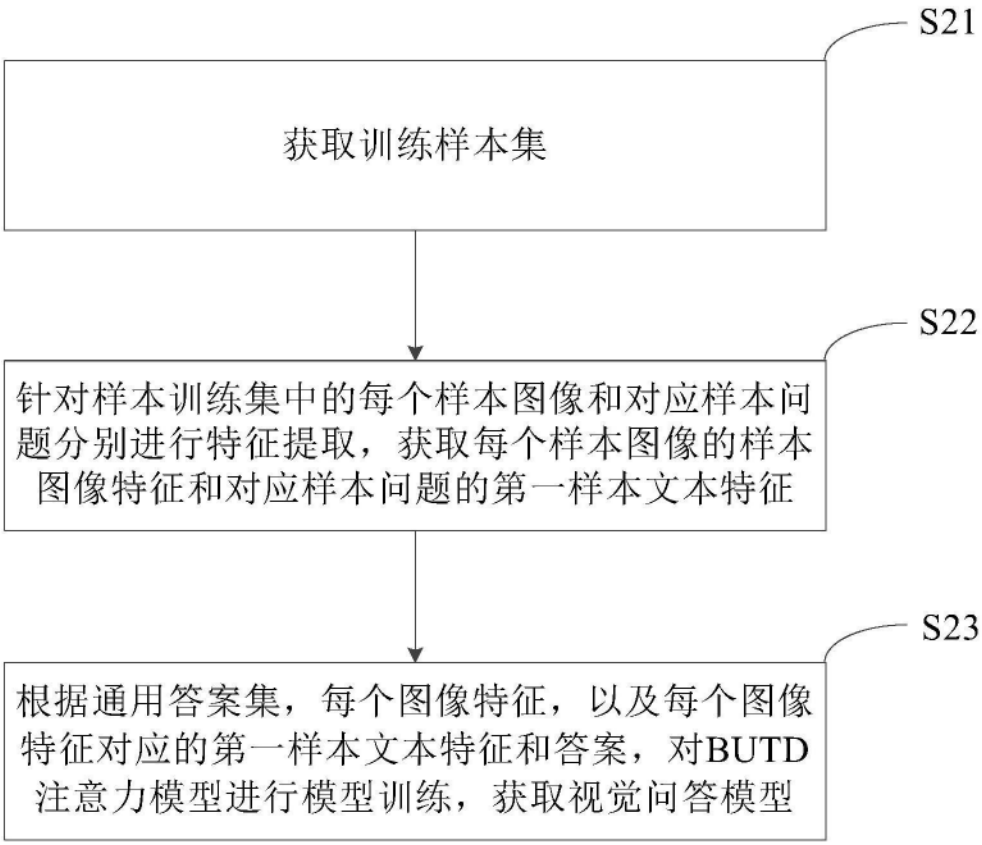

S21
获取训练样本集
S22
针对样本训练集中的每个样本图像和对应样本问题分别进行特征提取，获取每个样本图像的样本图像特征和对应样本问题的第一样本文本特征
S23
根据通用答案集，每个图像特征，以及每个图像特征对应的第一样本文本特征和答案，对BUTD注意力模型进行模型训练，获取视觉问答模型

图2

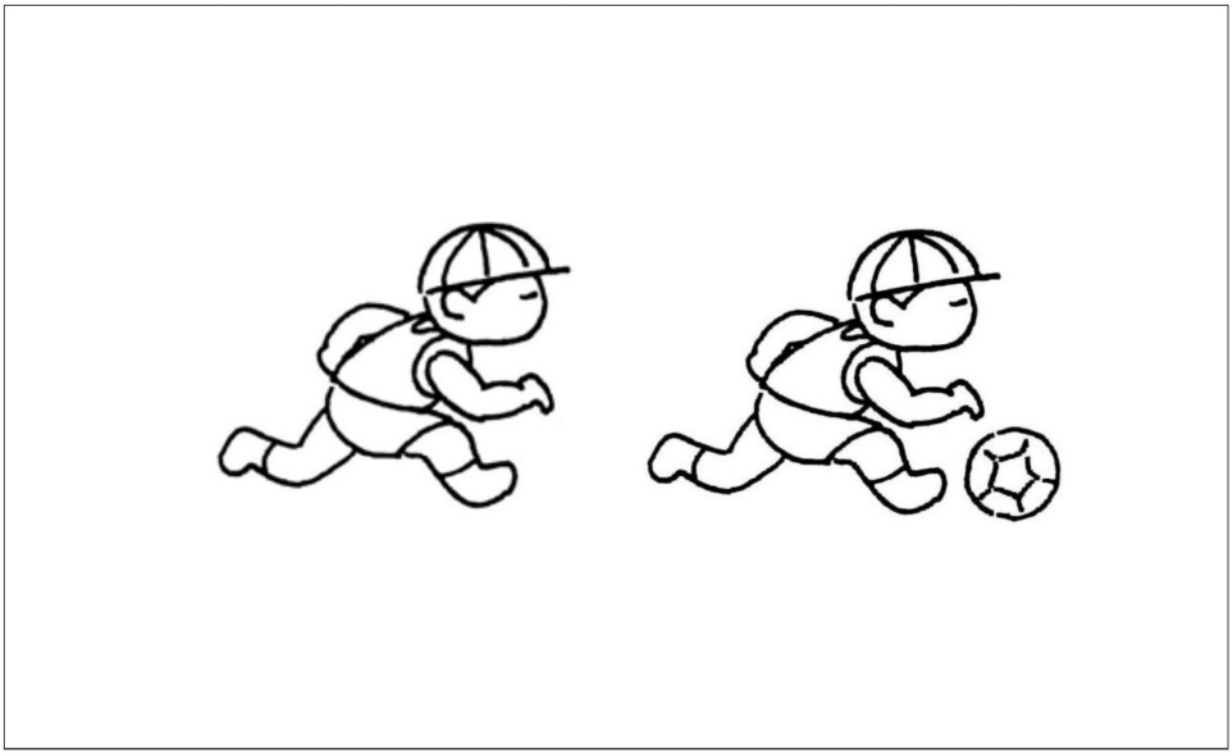


图3

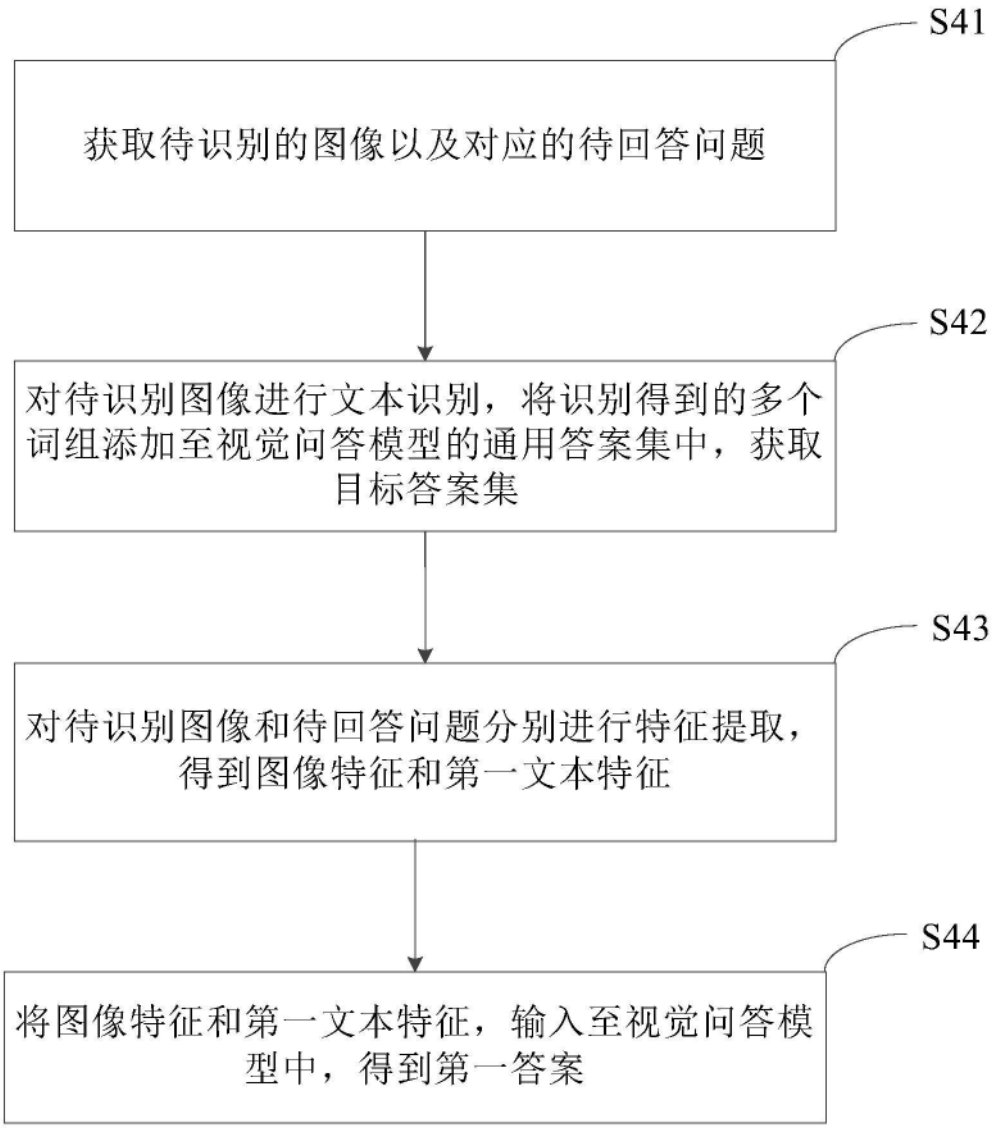

S41
获取待识别的图像以及对应的待回答问题
S42
对待识别图像进行文本识别，将识别得到的多个词组添加至视觉问答模型的通用答案集中，获取目标答案集
S43
对待识别图像和待回答问题分别进行特征提取，得到图像特征和第一文本特征
S44
将图像特征和第一文本特征，输入至视觉问答模型中，得到第一答案

图4

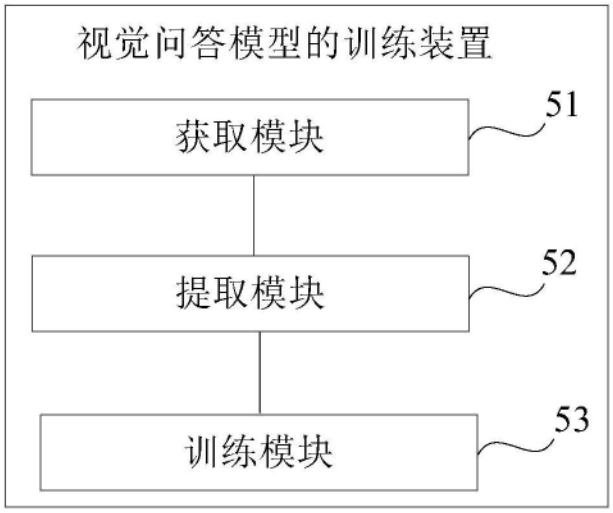

视觉问答模型的训练装置
获取模块
51
提取模块
52
训练模块
53

图5

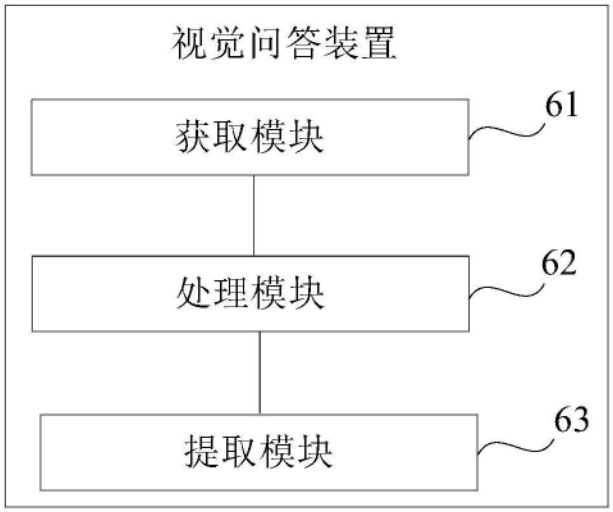

视觉问答装置
获取模块
61
处理模块
62
提取模块
63

图6

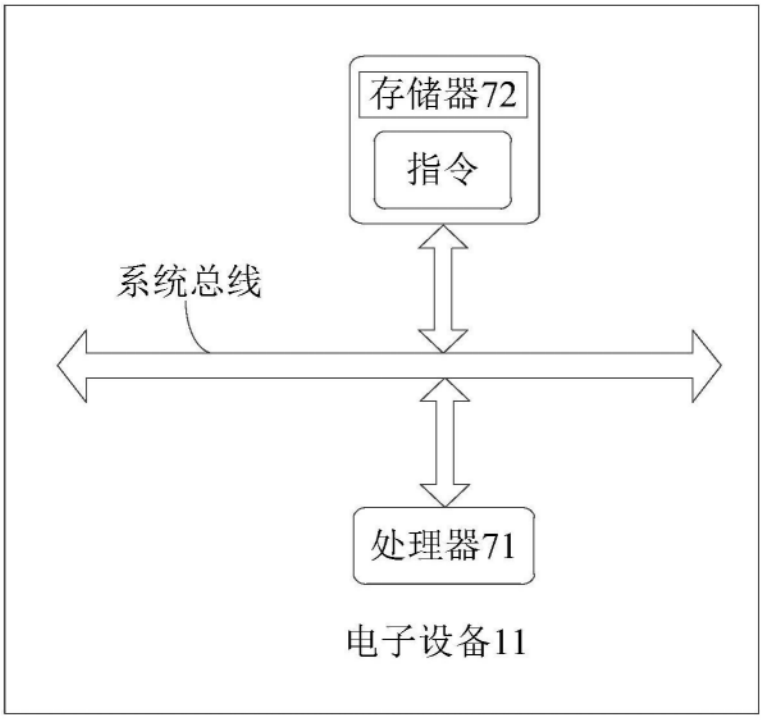



图7