

图像问答方法、装置、系统和存储介质

本申请要求在 2017 年 10 月 31 日提交中国专利局、申请号为 CN201711049294.7、发明名称为“图像问答方法、装置、系统和存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及 CV（Computer Vision，计算机视觉）领域，更具体地，涉及一种图像问答方法、装置、系统和存储介质。

背景技术

近来，在人工智能领域中，图像问答任务备受关注。图像问答也称作视觉问答（Visual Question Answering, VQA），其涉及这样的工作需求：根据图像来回答由自然语言构成的问题。VQA 作为语言与图像交互的一个示例性体现，可应用于诸如智能监控、图像筛选等多种工作场景。

发明内容

本申请提供了图像问答的技术方案。

本申请实施例的一方面提供了一种图像问答方法。所述图像问答方法包括：提取表示问题的语义的问题特征、图像的全局特征、以及包围所述图像中的对象的检测框的检测框特征；基于所述问题特征、所述全局特征和所述检测框特征获取所述图像的至少一个区域中每个区域的第一权重以及所述图像的至少一个检测框中每个检测框的第二权重；利用所述第一权重对所述全局特征加权处理，得到所述图像的区域注意力特征；利用所述第二权重对所述检测框特征加权处理，得到所述图像的检测框注意力特征；基于所述问题特征、所述区域注意力特征和所述检测框注意力特征预测所述问题的答案。

可选地，在本发明上述任一方法实施例中，所述提取表示问题的语义的问题特征可包括：利用循环神经网络对构成所述问题的词语的上下文进行特征提取，获得所述问题特征。

可选地，在本发明上述任一方法实施例中，所述提取图像的全局特征可包括：利用卷积神经网络提取所述全局特征，所述全局特征包括与所述图像的多个区域相关的多个区域特征。

可选地，在本发明上述任一方法实施例中，获取所述第二权重可包括：统一所述问题特征、所述全局特征和所述检测框特征的维度；将所述维度统一后的全局特征基于所述多

个区域特征的数量平均化；基于所述维度统一后的问题特征、所述维度统一后的检测框特征和所述维度统一且平均化后的全局特征获取所述第二权重。

可选地，在本发明上述任一方法实施例中，所述提取包围所述图像中的对象的检测框的检测框特征可包括：利用高速区域卷积神经网络获得包围所述图像中的对象的多个检测框；根据所述多个检测框所包围的对象与所述图像的背景的差异确定至少一个检测框；基于所述至少一个检测框提取获得至少一个检测框子特征；基于所述至少一个检测框子特征获得所述检测框特征。

可选地，在本发明上述任一方法实施例中，所述基于所述问题特征、所述全局特征和所述检测框特征获取所述图像的至少一个区域中每个区域的所述第一权重可包括：统一所述问题特征、所述全局特征和所述检测框特征的维度；将所述维度统一后的检测框特征基于所述多个检测框子特征的数量平均化；基于所述维度统一后的问题特征、所述维度统一后的全局特征和所述维度统一且平均化后的检测框特征获取所述第一权重。

可选地，在本发明上述任一方法实施例中，所述基于所述问题特征、所述区域注意力特征和所述检测框注意力特征预测所述问题的答案可包括：融合所述问题特征与所述区域注意力特征，得到第一问题预测答案；融合所述问题特征与所述检测框注意力特征，得到第二问题预测答案；通过对所述第一问题预测答案和所述第二问题预测答案进行分类，获得所述问题的答案。

本申请实施例的另一方面提供了一种图像问答装置。所述图像问答装置包括：特征提取器，提取表示问题的语义的问题特征、图像的全局特征、以及包围所述图像中的对象的检测框的检测框特征；权重计算器，基于所述问题特征、所述全局特征和所述检测框特征获取所述图像的至少一个区域中每个区域的第一权重以及所述图像的至少一个检测框中每个检测框的第二权重；区域注意力提取器，利用所述第一权重对所述全局特征加权处理，得到所述图像的区域注意力特征；检测框注意力提取器，利用所述第二权重对所述检测框特征加权处理，得到所述图像的检测框注意力特征；预测器，基于所述问题特征、所述区域注意力特征和所述检测框注意力特征预测所述问题的答案。

可选地，在本发明上述任一装置实施例中，所述特征提取器可配置有循环神经网络，所述特征提取器可利用所述循环神经网络对构成所述问题的词语的上下文进行特征提取，获得所述问题特征。

可选地，在本发明上述任一装置实施例中，所述特征提取器可配置有卷积神经网络，所述特征提取器可利用所述卷积神经网络提取所述全局特征，所述全局特征可包括与所述图像的多个区域相关的多个区域特征。

可选地，在本发明上述任一装置实施例中，所述权重计算器可包括：第二维度统一器，所述第二维度统一器统一所述问题特征、所述全局特征和所述检测框特征的维度；区域均

值器，所述区域均值器将所述维度统一后的全局特征基于所述多个区域特征的数量平均化；第二权重获取器，所述第二乘法器基于所述维度统一后的问题特征、所述维度统一后的检测框特征和所述维度统一且平均化后的全局特征获取所述第二权重。

可选地，在本发明上述任一装置实施例中，所述特征提取器可包括：检测框生成器，所述检测框生成器配置有用于获得包围所述图像中的对象的多个检测框的高速区域卷积神经网络；检测框选择器，所述检测框选择器根据所述多个检测框所包围的对象与所述图像的背景的差异确定至少一个检测框；检测框子特征提取器，所述检测框子特征提取器基于所述至少一个检测框提取获得至少一个检测框子特征；特征合成器，所述特征合成器基于所述至少一个检测框子特征获得所述检测框特征。

可选地，在本发明上述任一装置实施例中，所述权重计算器可包括：第一维度统一器，所述第一维度统一器统一所述问题特征、所述全局特征和所述检测框特征的维度；检测框均值器，所述检测框均值器将所述维度统一后的检测框特征基于所述多个检测框子特征的数量平均化；第一权重获取器，所述第一乘法器基于所述维度统一后的问题特征、所述维度统一后的全局特征和所述维度统一且平均化后的检测框特征获取所述第一权重。

可选地，在本发明上述任一装置实施例中，所述预测器可包括：区域预测器，所述区域预测器融合所述问题特征与所述区域注意力特征，得到第一问题预测答案；检测框预测器，所述检测框预测器融合所述问题特征与所述检测框注意力特征，得到第二问题预测答案；双模分析器，通过对所述第一问题预测答案和所述第二问题预测答案进行分类的方式，获得所述问题的答案。

本申请实施例的又一方面提供了一种图像问答系统。图像问答系统包括：存储可执行指令的存储器；以及一个或多个处理器。一个或多个处理器与存储器通信以执行可执行指令，从而完成本申请实施例提供的任一图像问答方法对应的操作，例如，可包括但不限于执行如下操作：提取表示问题的语义的问题特征、图像的全局特征、以及包围所述图像中的对象的检测框的检测框特征；基于所述问题特征、所述全局特征和所述检测框特征获取所述图像的至少一个区域中每个区域的第一权重以及所述图像的至少一个检测框中每个检测框的第二权重；利用所述第一权重对所述全局特征加权处理，得到所述图像的区域注意力特征；利用所述第二权重对所述检测框特征加权处理，得到所述图像的检测框注意力特征；基于所述问题特征、所述区域注意力特征和所述检测框注意力特征预测所述问题的答案。

本申请实施例的还一方面提供了一种电子设备，包括处理器，所述处理器包括如上任意一项所述的图像问答装置。

本申请实施例的再一方面提供了一种电子设备，包括：存储器，用于存储可执行指令；

以及处理器，用于与所述存储器通信以执行所述可执行指令从而完成如上任意一项所述图像问答方法。

本申请实施例的再一方面提供了一种计算机可读存储介质，所述计算机存储介质能够存储计算机可读指令，当这些指令被执行时可使处理器执行本申请实施例提供的任一图像问答方法对应的操作，例如，可包括但不限于执行如下操作：提取表示问题的语义的问题特征、图像的全局特征、以及包围所述图像中的对象的检测框的检测框特征；基于所述问题特征、所述全局特征和所述检测框特征获取所述图像的至少一个区域中每个区域的第一权重以及所述图像的至少一个检测框中每个检测框的第二权重；利用所述第一权重对所述全局特征加权处理，得到所述图像的区域注意力特征；利用所述第二权重对所述检测框特征加权处理，得到所述图像的检测框注意力特征；基于所述问题特征、所述区域注意力特征和所述检测框注意力特征预测所述问题的答案。

本申请实施例的又一方面提供了一种计算机程序产品，包括计算机可读代码，当所述计算机可读代码在设备上运行时，所述设备中的处理器执行用于实现如上任意一项所述的图像问答方法的指令。

本申请的技术方案通过基于问题语义综合考虑图像的全局特征和包含有对象检测框的检测框特征，更全面地分析了图像与问题之间的交互，提高了图像问答的效率和精度。

下面通过附图和实施例，对本申请的技术方案做进一步的详细描述。

附图说明

构成说明书的一部分的附图描述了本申请的实施例，并且连同描述一起用于解释本申请的原理。

参照附图，根据下面的详细描述，可以更加清楚地理解本申请，其中：

图 1 是本申请实施例提供的图像问答方法的一个实施例的流程示意图。

图 2 是本申请实施例提供的图像和与图像相关的问题的一些实施例的示意图。

图 3 是本申请实施例提供的提取检测框特征的方法的一个示例的流程示意图。

图 4 是本申请实施例提供的获取第一权重的方法的一个示例的流程示意图。

图 5 是本申请实施例提供的获取第二权重的方法的一个示例的流程示意图。

图 6 是本申请实施例提供的预测问题答案的方法的一个示例的流程示意图。

图 7 是本申请实施例提供的图像问答装置的一个实施例的结构示意图。

图 8 是本申请实施例提供的图像问答系统的一个实施例的结构示意图。

具体实施方式

现在将参照附图来详细描述本申请的各种示例性实施例。应注意到：除非另外具体说

明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本申请的范围。

同时，应当明白，为了便于描述，附图中所示出的各个部分的尺寸并不是按照实际的比例关系绘制的。

以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本申请及其应用或使用的任何限制。

对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为说明书的一部分。

应注意：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

本申请实施例可以应用于计算机系统/服务器，其可与众多其它通用或专用计算系统环境或配置一起操作。适于与计算机系统/服务器一起使用的众所周知的计算系统、环境和/或配置的例子包括但不限于：个人计算机系统、服务器计算机系统、瘦客户机、厚客户机、手持或膝上设备、基于微处理器的系统、机顶盒、可编程消费电子产品、网络个人电脑、小型计算机系统、大型计算机系统和包括上述任何系统的分布式云计算技术环境，等等。

计算机系统/服务器可以在由计算机系统执行的计算机系统可执行指令（诸如程序模块）的一般语境下描述。通常，程序模块可以包括例程、程序、目标程序、组件、逻辑、数据结构等等，它们执行特定的任务或者实现特定的抽象数据类型。计算机系统/服务器可以在分布式云计算环境中实施，分布式云计算环境中，任务是由通过通信网络链接的远程处理设备执行的。在分布式云计算环境中，程序模块可以位于包括存储设备的本地或远程计算系统存储介质上。

需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

图 1 是本申请实施例提供的图像问答方法 1000 的一个实施例的流程图。该方法可以由任意电子设备执行，例如终端设备、服务器、移动设备等等，如图 1 所示，该实施例方法包括：

在步骤 S1010 中，提取表示问题的语义的问题特征、图像的全局特征、以及包围图像中的对象的检测框的检测框特征。

问题可以是与图像相关的、由自然语言构成的疑问语句。图 2 是本申请实施例提供的图像和与图像相关的问题的一些实施例的示意图。参照图 2，例如：该问题是与图像 2110 相关的问题 2120—“盘子上有什么？”、与图像 2210 相关的问题 2220—“这是什么运动？”、与图像 2310 相关的问题 2320—“今天是晴天吗？”等。可采用深度学习网络（Deep Learning Networks）来提取问题的问题特征。该问题特征可例如体现为特征向量，其包含问题的语

义学表示 (Semantic Representation)。

图像的全局特征可例如体现为特征向量, 其包含图像的全部区域的内容的语义学表示。在该步骤中, 不对图像的特征提取内容进行限制, 即, 不对图像的前景内容 (可包括对象等) 和背景内容进行区分。因此, 图像的全局特征可保留图像的全局视觉内容以及用于回答问题所需的具体前景内容。

提取包围图像中的对象的检测框的检测框特征与提取图像的全局特征不同, 需要首先区分图像的前景内容与背景内容。可利用检测框来标注图像中的前景内容, 例如, 对象等。检测框可例如具有矩形形状, 且包围图像中的对象。在提取检测框特征时, 仅提取检测框内的图像特征。检测框特征例如体现为特征向量, 其包含检测框内的内容的语义学表示。因此, 检测框特征能够滤除大量的背景信息, 从而便于更高效和准确地预测与前景内容有关的问题。

在一个可选示例中, 该步骤 S1010 可以由处理器调用存储器存储的相应指令执行, 也可以由被处理器运行的特征提取器 2100 执行。

在步骤 S1020 中, 基于问题特征、全局特征和检测框特征获取图像的多个区域中每个区域的第一权重以及图像的多个检测框中每个检测框的第二权重。该步骤引入了一种视觉注意力机制。视觉注意力机制可将问题与图像进行初步关联, 从而为图像的不同部位设定不同的注意力权重。注意力权重反映了该部位与问题的关联程度。通过引入视觉注意力机制, 可以高效地关注与问题密切关联的图像部位, 从而提高了预测效率和精度。

本申请实施例引入了一种双路视觉注意力机制。在第一支路中, 视觉注意力机制基于图像的自然分区而引入。图像的自然分区是指, 不区分图像的前景内容和背景内容, 而是以自然位置关系将图像划分为一个或多个区域。例如, 如图 2 所示, 图像 2410 和图像 2510 均包含相同的图像内容并具有同一关联问题 2420——“这位男士在倒什么? ”。其中, 图像 2410 基于其横纵坐标划分成 $N \times N$ 个区域 2411, N 为大于或等于 1 的整数。例如, 在图 2 所示的示例中, 图像 2410 被分为 5×5 个区域。基于图像的自然分区而引入的视觉注意力机制将问题与图像的这 $N \times N$ 个区域分别进行初步关联, 从而得到这 $N \times N$ 个区域中的每个区域与问题的关联程度。这种关联程度可以以权重的形式来体现。即, 通过基于图像的自然分区而引入视觉注意力机制, 可获取图像的多个区域中每个区域的第一权重。

基于图像的自然分区引入视觉注意力机制的好处在于, 图像的全局视觉内容得以更大程度地保留, 尤其是背景内容得到很好的保留。例如, 在回答图 2 所示的与图像 2310 相关的问题 2320——“今天是晴天吗?”时, 由于问题实际上与图像 2310 的背景内容 (天空) 相关, 所以基于图像的自然分区引入视觉注意力机制可更好地对此类问题进行答案预测。

在第二支路中, 视觉注意力机制基于图像的检测框而引入。图像的检测框识别并包围了图像的前景内容 (例如, 对象)。图像可包含一个或多个对象。因此, 可在图像中生成

一个或多个检测框。例如，如图 2 所示，在图像 2510 中生成了 M 个检测框 2511， M 为大于或等于 1 的整数。例如，在图 2 所示的示例中，在图像 2510 中生成了 4 个检测框。基于图像的检测框而引入的视觉注意力机制将问题与图像的这 M 个检测框分别进行初步关联，从而得到这 M 个检测框中的每个检测框与问题的关联程度。这种关联程度可以以权重的形式来体现。即，通过基于图像的检测框而引入视觉注意力机制，可获取图像的多个检测框中每个检测框的第二权重。

基于图像检测框引入视觉注意力机制的好处在于，图像的前景内容得到更密切的关注，从而可以更高效和准确地回答与前景内容有关的问题。例如，在回答图 2 所示的与图像 2110 相关的问题 2120——“盘子上有什么？”以及与图像 2210 相关的问题 2220——“这是什么运动？”时，基于图像检测框引入视觉注意力机制可以忽略图像 2110 和图像 2210 中的背景信息，同时为包围了盘子的检测框和包围了球拍的检测框配置较大的权重，进而可对此类问题进行更高效和准确的答案预测。

在一个可选示例中，该步骤 S1020 可以由处理器调用存储器存储的相应指令执行，也可以由被处理器运行的权重计算器 2200 执行。

在步骤 S1030 中，利用第一权重对全局特征加权求和，得到图像的区域注意力特征。在图像基于其纵横坐标划分成 $N \times N$ 个区域的情况下，全局特征可包括与图像的多个区域相关的多个区域特征。例如，全局特征向量可以由 $N \times N$ 个区域特征向量联接而成。在步骤 S1020 中，获取了这 $N \times N$ 个区域中每个区域的第一权重。因此，可以利用这 $N \times N$ 个第一权重对这 $N \times N$ 个区域特征向量进行加权求和，从而得出图像的区域注意力特征，其可以例如体现为向量形式。

在一个可选示例中，该步骤 S1030 可以由处理器调用存储器存储的相应指令执行，也可以由被处理器运行的区域注意力提取器 2300 执行。

在步骤 S1040 中，利用第二权重对检测框特征加权求和，得到图像的检测框注意力特征。检测框特征可包括与图像的多个检测框相关的多个检测框子特征。例如，检测框特征向量可以由 M 个检测框子特征向量联接而成。在步骤 S1020 中，获取了这 M 个检测框中每个检测框的第二权重。因此，可以利用这 M 个第二权重对这 M 个检测框子特征向量进行加权求和，从而得出图像的检测框注意力特征，其可以例如体现为向量形式。

在一个可选示例中，该步骤 S1040 可以由处理器调用存储器存储的相应指令执行，也可以由被处理器运行的检测框注意力提取器 2400 执行。

需要注意的是，上述步骤 S1030 和步骤 S1040 之间的顺序可以互换，即可以先执行步骤 S1040 后执行步骤 S1030。

在步骤 S1050 中，基于问题特征、区域注意力特征和检测框注意力特征预测问题的答案。在得到问题特征、区域注意力特征和检测框注意力特征后，可以兼顾地学习问题特征

与区域注意力特征和检测框注意力特征的关系。这种预测过程可以作为一种多分类任务进行处理。例如，可利用分类器对学习的结果进行分类，根据多个候选答案的得分来预测问题的答案。根据本申请的技术方案，由于双路视觉注意力机制的引入，所以可在多种 VQA 场景中均实现高效和准确的答案预测。

在一个可选示例中，该步骤 S1050 可以由处理器调用存储器存储的相应指令执行，也可以由被处理器运行的预测器 2500 执行。

在本申请的一个可选实施例中，提取问题特征可包括：利用循环神经网络对构成问题的词语的上下文进行特征提取，获得问题特征。如上述任一实施例，可采用多种深度神经网络来提取问题的问题特征，包括但不限于，卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)等。与通常意义上的 CNN 不同，RNN 可将前一时间点的输出信息提供作为下一时间点的输入信息之一，因而具备对历史信息的学习能力。

如上述任一实施例，问题一般是由自然语言构成的疑问语句。这样的语句包含多个词语，词语与词语之间具有上下文联系。例如，在解析“餐桌上摆放的是苹果还是梨？”这样的问题时，当对词语“餐桌”进行语义学特征提取之后，能够预料到后文中有较大概率出现与食品相关的词语。因此，在问题特征提取的过程中，引入具备对历史信息的学习能力的 RNN 是有益的。

RNN 可包括多种变型，诸如长短时记忆单元(Long Short-Term Memory, LSTM)和门循环单元(Gated Recurrent Unit, GRU)等。利用 RNN 对问题进行特征提取时，可首先将构成问题的词语依次编码成独热(One-Hot)码。独热码编码是这样一种编码方式，系统有多少个状态，该编码就具有多少个比特。这些比特中，只有一个为 1，而其余均为 0。例如，系统整理了 5000 个单词(即，5000 种可能的状态)，则独热码即具有 5000 个比特。此时，独热码可看作一个向量，每一元素仅可以是 0 或 1。仅与输入词语对应的元素的比特为 1，而其余元素均为 0。在实际应用中，可以为系统预留至少一个比特用于表征未被系统整理的单词。通过独热码的编码方式，可以实现文本字符向数字化数据的转换。然后，可以基于独热码对问题进行语义特征提取。例如，将当前时刻的独热码输入以及基于先前输出的独热码特征提取结果共同输入 RNN。RNN 可考虑历史信息对共同输入的内容进行编码处理，由此提取当前词语的语义学特征。

在本申请的一个可选实施例中，提取全局特征可包括：利用 CNN 提取全局特征，其中，全局特征包括与图像的多个区域相关的多个区域特征。可用于进行图像的全局特征提取的 CNN 的示例可包括但不限于 ResNet 网络或 VGG-16 网络。CNN 可以针对图像识别任务的应用场景而使用合适的训练样本集进行预训练，这样的样本集例如 ImageNet。输入的图像可以首先被调整到适于所使用的 CNN 的尺寸(例如，像素数量)，然后输入至 CNN。

CNN 可包括多个卷积核，每个卷积核与图像进行卷积操作后得到一个特征信道输出。CNN 的卷积核的数量（或特征信道数量）可根据图像识别任务的应用场景设定。如上述任一实施例，在图像基于其横纵坐标划分成 $N \times N$ 个区域的情况下，全局特征可包括与图像的多个区域相关的多个区域特征。例如，全局特征可体现为向量形式，全局特征向量可以由 $N \times N$ 个区域特征向量联接而成。

图 3 是本申请实施例提供的提取检测框特征的方法的一个示例的流程示意图。参照图 3，在本申请的一个实施例中，提取检测框特征可包括：在步骤 S1031 中，利用高速区域卷积神经网络（Faster-Region Convolutional Neural Network, Faster-RCNN）获得包围图像中的对象的多个检测框；在步骤 S1032 中，根据多个检测框所包围的对象与图像的背景的差异确定至少一个检测框；在步骤 S1033 中，基于至少一个检测框提取获得至少一个检测框子特征；在步骤 S1034 中，基于至少一个检测框子特征获得检测框特征。

在本申请实施例中，需要首先区分图像的对象与背景。这涉及到对象识别技术。Faster-RCNN 是一种高效的对象识别网络。区别于区域卷积神经网络（RCNN）和快速区域卷积神经网络（Fast-RCNN），Faster-RCNN 可以自动地生成候选检测框（Region Proposal）。

在生成候选检测框后，Faster-RCNN 可以识别候选检测框中的图像特征，从而判断候选检测框包围了对象还是背景。在候选检测框完全或部分地包围了对象的情况下，Faster-RCNN 可以进一步修正候选检测框的尺寸和位置，从而以合适的尺寸完全包围图像中的对象。

在实际应用中，可生成多个检测框以包围图像中的多个对象。例如，可生成 M' 个检测框以包围 M' 个对象。这 M' 个检测框均映射有它们所包含的对象的分数值，该分数值反映出这个对象与背景的差别度。分数值越低，则对象越趋近于背景。在本实施例中，可按照分数值从高到低的顺序从这 M' 个检测框中选取 M 个检测框进行后续的特征提取，其中 $M \leq M'$ 。从所选取的一部分检测框中提取出检测框子特征。最后，将所提取的检测框子特征合成为检测框特征。通过上述方式提取检测框特征，可以自主地生成检测框从而减少人为干预。另外，通过选择与背景差距大的部分检测框，能够尽可能地提取到高质量的对象信息，从而为与前景内容相关的 VQA 任务提供高针对性的图像解析结果。

图 4 是本申请实施例提供的获取第一权重的方法的一个示例的流程示意图。参照图 4，在本申请的一个可选实施例中，获取第一权重 S1040-1 可包括：在步骤 S1041 中，统一问题特征、全局特征和检测框特征的维度；在步骤 S1042 中，将维度统一后的检测框特征基于多个检测框子特征的数量平均化；在步骤 S1043 中，基于维度统一后的问题特征、维度统一后的全局特征和维度统一且平均化后的检测框特征获取第一权重。

在获取了问题特征、全局特征和检测框特征之后，可基于图像的自然分区引入视觉注

注意力机制。图像可基于其横纵坐标划分成 $N \times N$ 个区域。基于图像的自然分区而引入的视觉注意力机制将问题与图像的这 $N \times N$ 个区域分别进行初步关联，从而得到这 $N \times N$ 个区域中的每个区域与问题的关联程度。可选地，如上述任一实施例，可从图像中选取 M 个检测框进行后续的特征提取，所提取的特征联接成检测框特征。从这 M 个检测框提取的特征即为构成检测框特征的 M 个检测框子特征。如上述任一实施例，问题特征、全局特征和检测框特征均可体现为向量形式。因此，可将这三个向量的维度进行统一以便于后续操作。在本实施例中，可基于检测框子特征的数量 M 对检测框特征进行平均化以更好地学习理解图像的全局视觉内容。

可选地，上述平均化过程和维度统一过程可由以下等式（1）表示：

$$\begin{aligned} R_1 &= \tanh(W_{r1}R + b_{r1}), \\ D_1 &= \frac{1}{M} \cdot 1(\tanh(W_{d1}D + b_{d1})^T), \\ Q_1 &= \tanh(W_{q1}Q + b_{q1}) \end{aligned} \quad \text{等式（1）}$$

等式（1）中， Q 为问题特征向量， R 为全局特征向量， D 为检测框特征向量； Q_1 为维度统一后的问题特征， D_1 为维度统一且平均化的检测框特征， R_1 为维度统一后的全局特征。 W_{q1} 、 b_{q1} 、 W_{r1} 、 b_{r1} 、 W_{d1} 和 b_{d1} 分别为可利用机器学习而更新的网络参数，并且能够将问题特征向量 Q 、全局特征向量 R 和检测框特征 D 的维度统一。 M 表示检测框子特征的数量， $\tanh()$ 为双曲正切函数， 1 表示全 1 向量。

最后，可基于维度统一后的问题特征 Q_1 、维度统一后的全局特征 R_1 和维度统一且平均化后的检测框特征 D_1 获取第一权重。例如，可将维度统一后的问题特征 Q_1 、维度统一后的全局特征 R_1 和维度统一且平均化后的检测框特征 D_1 逐元素相加或者相乘。在一个可选实施例中，采取了相乘的方式，例如进行哈达玛乘积（Hadamard Product）。在该实施例中，通过对维度统一后的问题特征 Q_1 、维度统一后的全局特征 R_1 和维度统一且平均化后的检测框特征 D_1 进行相乘来获取第一权重可以获得更准确的权重信息。在可选地操作中，可在相乘后进行 L2 范数正则化以限制向量的模。

例如，上述操作可由以下等式（2）表示：

$$C_1 = \text{Norm}_2(\tilde{Q}_1 \circ R_1 \circ \tilde{D}_1) \quad \text{等式（2）}$$

在等式（2）中， C_1 表示问题特征、全局特征和检测框特征的联合特征表示， Norm_2 表示 L2 范数正则化， $\circ$ 表示哈达玛乘积。

C_1 可看作向量形式的第一权重，其包含图像多个区域的权重信息。在实际操作中，可对该联合特征表示 C_1 进行进一步的卷积操作并利用 Softmax 函数进行激活。

例如，上述操作可由以下等式（3）表示：

$$a_1 = \text{softmax}(W_{c1} * C_1 + b_{c1}) \quad \text{等式（3）}$$

等式 (3) 中, a_1 为标量形式的第一权重, W_{c1} 和 b_{c1} 为可利用机器学习而更新的网络参数。

图 5 是本申请实施例提供的获取第二权重的方法的一个示例的流程示意图。参照图 5, 在本申请的一个实施例中, 获取第二权重 S1040-2 可包括: 在步骤 S1044 中, 统一问题特征、全局特征和检测框特征的维度; 在步骤 S1045 中, 将维度统一后的全局特征基于多个区域特征的数量平均化; 在步骤 S1046 中, 基于维度统一后的问题特征、维度统一后的检测框特征和维度统一且平均化后的全局特征获取第二权重。

在获取了问题特征、全局特征和检测框特征之后, 可基于图像的检测框引入视觉注意力机制。在图像中生成有 M 个检测框的情况下, 检测框特征可包括与图像的 M 个检测框相关的 M 个检测框子特征。基于图像的检测框而引入的视觉注意力机制将问题与图像的这 M 个检测框分别进行初步关联, 从而得到这 M 个检测框中的每个检测框与问题的关联程度。可选地, 如上述任一实施例, 图像可包括 $N \times N$ 个区域。对这 $N \times N$ 个区域进行特征向量提取, 得到与图像的这 $N \times N$ 个区域相关的 $N \times N$ 个区域特征。如上述任一实施例, 问题特征、全局特征和检测框特征均可体现为向量形式。因此, 可将这三个向量的维度进行统一以便于后续操作。在本实施例中, 可将将维度统一后的全局特征基于多个区域特征的数量 $N \times N$ 而平均化。

可选地, 上述平均化过程和维度统一过程可由以下等式 (4) 表示:

$$\begin{aligned} D_2 &= \tanh(W_{d2}D + b_{d2}), \\ R_2 &= \frac{1}{N \times N} \cdot 1(\tanh(W_{r2}R + b_{r2})^T), \\ Q_1 &= \tanh(W_{q2}Q + b_{q2}) \end{aligned} \quad \text{等式 (4)}$$

等式 (4) 中, Q 为问题特征向量, R 为全局特征向量, D 为检测框特征向量。 Q_2 为维度统一后的问题特征, D_2 为维度统一且平均化的检测框特征, R_2 为维度统一后的全局特征。 W_{q2} 、 b_{q2} 、 W_{r2} 、 b_{r2} 、 W_{d2} 和 b_{d2} 分别为可利用机器学习而更新的网络参数, 并且能够将问题特征向量 Q 、全局特征向量 R 和检测框特征向量 D 的维度统一。 $N \times N$ 表示图像区域的数量, $\tanh()$ 为双曲正切函数, 1 表示全 1 向量。

最后, 可基于维度统一后的问题特征 Q_2 、维度统一后的检测框特征 D_2 和维度统一且平均化后的全局特征 R_2 获取第二权重。例如, 可将维度统一后的问题特征 Q_2 、维度统一后的检测框特征 D_2 和维度统一且平均化后的全局特征 R_2 逐元素相加或相乘。在一个实施例中, 采取了相乘的形式, 例如, 进行哈达玛乘积 (Hadamard Product)。在该实施例中, 通过对基于维度统一后的问题特征 Q_2 、维度统一后的检测框特征 D_2 和维度统一且平均化后的全局特征 R_2 进行相乘来获取第二权重可以获得更准确的权重信息。在可选的操作中, 可在相乘后进行 L2 范数正则化以限制向量的模。

例如，上述操作可由以下等式（5）表示：

$$C_2 = \text{Norm}_2(\tilde{Q}_2 \circ \tilde{R}_2 \circ D_2) \quad \text{等式（5）}$$

在等式（5）中， C_2 表示问题特征、全局特征和检测框特征的联合特征表示， Norm_2 表示 L2 范数正则化， $\circ$ 表示哈达玛乘积。

C_2 可看作向量形式的第二权重，其包含图像至少一个检测框的权重信息。在实际操作中，可对该联合特征表示 C_2 进行进一步的卷积操作并进行 Softmax 激活。

例如，上述操作可由以下等式（6）表示：

$$a_2 = \text{softmax}(W_{c2}C_2 + b_{c2}) \quad \text{等式（6）}$$

等式（6）中， a_2 为标量形式的第二权重， W_{c2} 和 b_{c2} 为可利用机器学习而更新的网络参数。

在本申请的一个可选实施例中，在基于上述实施例方法获得第一权重 a_1 和第二权重 a_2 的情况下，可选地，参照图 1 描述的步骤 S1050 和步骤 S1060 可分别表示为以下等式（7）和等式（8）：

$$v_1 = \sum_i^{N \times N} a_1(i) R_1(i) \quad \text{等式（7）}$$

在等式（7）中，可利用 $N \times N$ 个第一权重 a_1 对 $N \times N$ 个区域特征向量进行加权求和，从而得出图像的区域注意力特征，其可以例如体现为向量形式 v_1 。

$$v_2 = \sum_i^M a_2(i) D_2(i) \quad \text{等式（8）}$$

在等式（8）中，可利用 M 个第二权重 a_2 对 M 个检测框子特征向量进行加权求和，从而得出图像的检测框注意力特征，其可以例如体现为向量形式 v_2 。

图 6 是本申请实施例提供的预测问题答案的方法的一个示例的流程示意图。参照图 6，在本申请的一个实施例中，预测问题答案可包括：在步骤 S1071 中，融合问题特征与区域注意力特征，得到第一问题预测答案；在步骤 S1072 中，融合问题特征与检测框注意力特征，得到第二问题预测答案；在步骤 S1073 中，通过对第一问题预测答案和第二问题预测答案进行分类的方式，预测问题的答案。

在本申请的实施例中，可以将问题预测作为多目标分类任务来处理。例如，可以对问题特征进行进一步学习，将进一步学习后的问题特征与区域注意力特征相乘，从而获得第一问题预测。同样地，可以对问题特征进行进一步学习，将进一步学习后的问题特征与检测框注意力特征相乘，从而获得第二问题预测。

可选地，上述操作可以表示为以下等式（9）：

$$\begin{aligned} h_r &= v_1 \circ \text{tand}(W_{hr}Q + b_{hr}) \\ h_d &= v_2 \circ \text{tand}(W_{hd}Q + b_{hd}) \end{aligned} \quad \text{等式（9）}$$

其中， h_r 为第一问题预测答案， h_d 为第二问题预测答案。 W_{hr} 、 b_{hr} 、 W_{hd} 和 b_{hd} 为可利用机器学习而更新的网络参数，并且可对问题特征进行进一步学习，从而增强表征能力。

第一问题预测答案 h_r 可更准确地回答与全局视觉内容相关的问题，而第二问题预测答案 h_d 可更准确地回答与前景内容相关的问题。

最后，将第一问题预测答案 h_r 与第二问题预测答案 h_d 相加，并对两者的和进行线性分类，从而最终预测问题的答案。例如，可对第一问题预测答案 h_r 与第二问题预测答案 h_d 的和进行线性变换，并应用 Softmax 函数以预测最终的问题答案。

例如，上述操作可以表示为以下等式 (10)：

$$p_{ans} = \text{softmax}(W_p(h_r + h_d) + b_p) \quad \text{等式 (10)}$$

其中， p_{ans} 表征预测的问题答案的概率， W_p 和 b_p 为可利用机器学习而更新的网络参数。可例如将 p_{ans} 的数值最大的答案选为最终的问题答案。或者，可例如将 p_{ans} 的数值按降序排列，从而选择 p_{ans} 的数值最高的多个答案作为候选答案以供用户选择。根据本实施例，一方面，第一问题预测答案与第二问题预测答案的过程相对独立进行，因此可从图像中尽可能地捕获不同类型的信息；另一方面，通过结合这双路预测，可以很好地适应多种 VQA 场景。

本申请上述实施例中所提及的特征向量的融合操作可采用多种方式，例如相加和相乘等。可选地，在本申请的示例性实施例中采用了哈达玛乘积。在本申请各实施例中，由于特征向量的维度均保持一致，因而更适于采用哈达玛乘积。另外，采用哈达玛乘积也可降低计算负担，从而提高方法的执行效率。

如本领域技术人员将理解的，上述各网络需要经过训练才能得以应用。可利用常规的反向传播(Back-Propagation, BP)方法和/或含时反向传播(Back-Propagation Through Time, BPTT)方法来完成训练。本领域技术人员基于本申请的公开内容可得出这些训练方法，因而在本说明书中不再赘述。

本领域普通技术人员可以理解：实现上述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计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

图 7 是本申请实施例提供的图像问答装置 2000 的一个实施例的结构示意图。该实施例的装置可用于实现本申请上述各方法实施例。如图 7 所示，该实施例的装置包括：图像问答装置 2000 包括特征提取器 2100、权重计算器 2200、区域注意力提取器 2300、检测框注意力提取器 2400 和预测器 2500。特征提取器 2100 提取表示问题的语义的问题特征、图像的全局特征、包围图像中的对象的检测框的检测框特征。权重计算器 2200 基于问题特征、全局特征和检测框特征获取图像的至少一个区域中每个区域的第一权重以及图像的至少一个检测框中每个检测框的第二权重。区域注意力提取器 2300 利用第一权重对全局特

征加权处理，得到图像的区域注意力特征。检测框注意力提取器 2400 利用第二权重对检测框特征加权处理，得到图像的检测框注意力特征。预测器 2500 基于问题特征、区域注意力特征和检测框注意力特征预测问题的答案。

特征提取器 2100 可配置有循环神经网络。特征提取器 2100 可利用循环神经网络基于构成问题的词语的上下文进行特征提取，获得问题特征。

特征提取器 2100 可配置有卷积神经网络。特征提取器 2100 可利用卷积神经网络提取全局特征，全局特征包括与图像的多个区域相关的多个区域特征。

特征提取器 2100 可包括：检测框生成器，检测框生成器配置有用于获得包围图像中的对象的多个检测框的高速区域卷积神经网络；检测框选择器，检测框选择器根据多个检测框所包围的对象与图像的背景的差异确定至少一个检测框；检测框子特征提取器，检测框子特征提取器基于至少一个检测框提取获得至少一个检测框子特征；特征合成器，特征合成器基于至少一个检测框子特征获得检测框特征。

权重计算器 2200 可包括：第一维度统一器，第一维度统一器统一问题特征、全局特征和检测框特征的维度；检测框均值器，检测框均值器将维度统一后的检测框特征基于多个检测框子特征的数量平均化；第一权重获取器，第一乘法器基于维度统一后的问题特征、维度统一后的全局特征和维度统一且平均化后的检测框特征获取第一权重。

权重计算器 2200 可包括：第二维度统一器，第二维度统一器统一问题特征、全局特征和检测框特征的维度；区域均值器，区域均值器将维度统一后的全局特征的基于多个区域特征的数量平均化；第二权重获取器，第二乘法器基于维度统一后的问题特征、维度统一后的检测框特征和维度统一且平均化后的全局特征获取第二权重。

预测器 2500 可包括：区域预测器，区域预测器融合问题特征与区域注意力特征，得到第一问题预测答案；检测框预测器，检测框预测器融合问题特征与检测框注意力特征，得到第二问题预测答案；双模分析器，通过对第一问题预测答案和第二问题预测答案进行分类，获得预测问题的答案。

根据本申请实施例的又一个方面，提供一种电子设备，包括处理器，该处理器包括上述任意一项实施例提供的图像问答装置。

根据本申请实施例的还一个方面，提供一种电子设备，包括：存储器，用于存储可执行指令；

以及处理器，用于与存储器通信以执行该可执行指令从而完成上述任意一项实施例提供的图像问答方法的操作。

根据本申请实施例的再一个方面，提供一种计算机存储介质，用于存储计算机可读的指令，该指令被执行时执行上述任意一项实施例提供的图像问答方法的操作。

本申请各实施例中的神经网络，分别可以是一个多层神经网络（即：深度神经网络），

例如多层的卷积神经网络，例如可以是 LeNet、AlexNet、GoogLeNet、VGG、ResNet 等任意神经网络模型。各神经网络可以采用相同类型和结构的神经网络，也可以采用不同类型和结构的神经网络。本申请实施例不对此进行限制。

本申请实施例还提供了一种电子设备，例如可以是移动终端、个人计算机（PC）、平板电脑、服务器等。下面参考图 8，其示出了适于用来实现本申请实施例的终端设备或服务器的电子设备 800 的结构示意图：如图 8 所示，计算机系统 800 包括一个或多个处理器、通信部等，所述一个或多个处理器例如：一个或多个中央处理单元（CPU）801，和/或一个或多个图像处理器（GPU）813 等，处理器可以根据存储在只读存储器（ROM）802 中的可执行指令或者从存储部 808 加载到随机访问存储器（RAM）803 中的可执行指令而执行各种适当的动作和处理。通信部 812 可包括但不限于网卡，所述网卡可包括但不限于 IB（Infiniband）网卡。

处理器可与只读存储器 802 和/或随机访问存储器 803 中通信以执行可执行指令，通过总线 804 与通信部 812 相连、并经通信部 812 与其他目标设备通信，从而完成本申请实施例提供的任一项方法对应的操作，例如：提取表示问题的语义的问题特征、图像的全局特征、以及包围图像中的对象的检测框的检测框特征；基于问题特征、全局特征和检测框特征获取图像的至少一个区域中每个区域的第一权重以及图像的至少一个检测框中每个检测框的第二权重；利用第一权重对全局特征加权处理，得到图像的区域注意力特征；利用第二权重对检测框特征加权处理，得到图像的检测框注意力特征；基于问题特征、区域注意力特征和检测框注意力特征预测问题的答案。

此外，在 RAM803 中，还可存储有装置操作所需的各种程序和数据。CPU 801、ROM802 以及 RAM803 通过总线 804 彼此相连。在有 RAM803 的情况下，ROM802 为可选模块。RAM803 存储可执行指令，或在运行时向 ROM802 中写入可执行指令，可执行指令使 CPU801 执行上述通信方法对应的操作。输入/输出（I/O）接口 805 也连接至总线 804。通信部 812 可以集成设置，也可以设置为具有多个子模块（例如多个 IB 网卡），并在总线链接上。

以下部件连接至 I/O 接口 805：包括键盘、鼠标等的输入部 806；包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的输出部 807；包括硬盘等的存储部 808；以及包括诸如 LAN 卡、调制解调器等网络接口卡的通信部分 809。通信部分 809 经由诸如因特网的网络执行通信处理。驱动器 810 也根据需要连接至 I/O 接口 805。可拆卸介质 811，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器 810 上，以便于从其上读出的计算机程序根据需要被安装入存储部 808。

需要说明的，如图 8 所示的架构仅为一种可选实现方式，在具体实践过程中，可根据实际需要对上述图 8 的部件数量和类型进行选择、删减、增加或替换；在不同功能部件设

置上，也可采用分离设置或集成设置等实现方式，例如 GPU813 和 CPU801 可分离设置或者可将 GPU813 集成在 CPU801 上，通信部可分离设置，也可集成设置在 CPU801 或 GPU813 上，等等。这些可替换的实施方式均落入本申请公开的保护范围。

特别地，根据本申请的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本申请的实施例包括一种计算机程序产品，其包括有形地包含在机器可读介质上的计算机程序，计算机程序包含用于执行流程图所示的方法的程序代码，程序代码可包括对应执行本申请实施例提供的方法步骤对应的指令，例如：提取表示问题的语义的问题特征、图像的全局特征、以及包围图像中的对象的检测框的检测框特征；基于问题特征、全局特征和检测框特征获取图像的至少一个区域中每个区域的第一权重以及图像的至少一个检测框中每个检测框的第二权重；利用第一权重对全局特征加权处理，得到图像的区域注意力特征；利用第二权重对检测框特征加权处理，得到图像的检测框注意力特征；基于问题特征、区域注意力特征和检测框注意力特征预测问题的答案。在这样的实施例中，该计算机程序可以通过通信部分 809 从网络上被下载和安装，和/或从可拆卸介质 811 被安装。在该计算机程序被中央处理单元（CPU）801 执行时，执行本申请的方法中限定的上述功能。

本说明书中各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其它实施例的不同之处，各个实施例之间相同或相似的部分相互参见即可。对于系统实施例而言，由于其与方法实施例基本对应，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。

可能以许多方式来实现本申请的方法和装置。例如，可通过软件、硬件、固件或者软件、硬件、固件的任何组合来实现本申请的方法和装置。用于所述方法的步骤的上述顺序仅是为了进行说明，本申请的方法的步骤不限于以上具体描述的顺序，除非以其它方式特别说明。此外，在一些实施例中，还可将本申请实施为记录在记录介质中的程序，这些程序包括用于实现根据本申请的方法的机器可读指令。因而，本申请还覆盖存储用于执行根据本申请的方法的程序的记录介质。

本申请的描述是为了示例和描述起见而给出的，而并不是无遗漏的或者将本申请限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显然的。选择和描述实施例是为了更好说明本申请的原理和实际应用，并且使本领域的普通技术人员能够理解本申请从而设计适于特定用途的带有各种修改的各种实施例。

权利要求

1. 一种图像问答方法，其特征在于，所述图像问答方法包括：

提取表示问题的语义的问题特征、图像的全局特征、以及包围所述图像中的对象的检测框的检测框特征；

基于所述问题特征、所述全局特征和所述检测框特征获取所述图像的至少一个区域中每个区域的第一权重以及所述图像的至少一个检测框中每个检测框的第二权重；

利用所述第一权重对所述全局特征加权处理，得到所述图像的区域注意力特征；

利用所述第二权重对所述检测框特征加权处理，得到所述图像的检测框注意力特征；

基于所述问题特征、所述区域注意力特征和所述检测框注意力特征预测所述问题的答案。

2. 根据权利要求1所述的图像问答方法，其特征在于，所述提取表示问题的语义的问题特征，包括：

利用循环神经网络对构成所述问题的词语的上下文进行特征提取，获得所述问题特征。

3. 根据权利要求1或2所述的图像问答方法，其特征在于，所述提取图像的全局特征，包括：

利用卷积神经网络提取所述全局特征，所述全局特征包括与所述图像的多个区域相关的多个区域特征。

4. 根据权利要求3所述的图像问答方法，其特征在于，所述基于所述问题特征、所述全局特征和所述检测框特征获取所述图像的至少一个区域中每个区域的所述第二权重，包括：

统一所述问题特征、所述全局特征和所述检测框特征的维度；

将所述维度统一后的全局特征基于所述多个区域特征的数量平均化；

基于所述维度统一后的问题特征、所述维度统一后的检测框特征和所述维度统一且平均化后的全局特征获得所述第二权重。

5. 根据权利要求1-4中任一项所述的图像问答方法，其特征在于，所述提取包围所述图像中的对象的检测框的检测框特征，包括：

利用高速区域卷积神经网络获得包围所述图像中的对象的多个检测框；

根据所述多个检测框所包围的对象与所述图像的背景的差异确定至少一个检测框；

基于所述至少一个检测框提取获得至少一个检测框子特征；

基于所述至少一个检测框子特征获得所述检测框特征。

6. 根据权利要求 5 所述的图像问答方法, 其特征在于, 所述基于所述问题特征、所述全局特征和所述检测框特征获取所述图像的至少一个区域中每个区域的第一权重, 包括:

统一所述问题特征、所述全局特征和所述检测框特征的维度;

将所述维度统一后的检测框特征基于所述多个检测框子特征的数量平均化;

基于所述维度统一后的问题特征、所述维度统一后的全局特征和维度统一且平均化后的检测框特征获取所述第一权重。

7. 根据权利要求 1-6 中任一项所述的图像问答方法, 其特征在于, 所述基于所述问题特征、所述区域注意力特征和所述检测框注意力特征预测所述问题的答案, 包括:

融合所述问题特征与所述区域注意力特征, 得到第一问题预测答案;

融合所述问题特征与所述检测框注意力特征, 得到第二问题预测答案;

通过对所述第一问题预测答案和所述第二问题预测答案进行分类, 获得所述问题的答案。

8. 一种图像问答装置, 其特征在于, 所述图像问答装置包括:

特征提取器, 提取表示问题的语义的问题特征、图像的全局特征、以及包围所述图像中的对象的检测框的检测框特征;

权重计算器, 基于所述问题特征、所述全局特征和所述检测框特征获取所述图像的至少一个区域中每个区域的第一权重以及所述图像的至少一个检测框中每个检测框的第二权重;

区域注意力提取器, 利用所述第一权重对所述全局特征加权处理, 得到所述图像的区域注意力特征;

检测框注意力提取器, 利用所述第二权重对所述检测框特征加权处理, 得到所述图像的检测框注意力特征;

预测器, 基于所述问题特征、所述区域注意力特征和所述检测框注意力特征预测所述问题的答案。

9. 根据权利要求 8 所述的图像问答装置, 其特征在于, 所述特征提取器配置有循环神经网络, 所述特征提取器利用所述循环神经网络对构成所述问题的词语的上下文进行特征提取, 获得所述问题特征。

10. 根据权利要求 8 或 9 所述的图像问答装置, 其特征在于, 所述特征提取器配置有卷积神经网络, 所述特征提取器利用所述卷积神经网络提取所述全局特征, 所述全局特征包括与所述图像的多个区域相关的多个区域特征。

11. 根据权利要求 10 所述的图像问答装置, 其特征在于, 所述权重计算器包括: 第二维度统一器, 所述第二维度统一器统一所述问题特征、所述全局特征和所述

检测框特征的维度；

区域均值器，所述区域均值器将所述维度统一后的全局特征基于所述多个区域特征的数量平均化；

第二权重获取器，所述第二乘法器基于所述维度统一后的问题特征、所述维度统一后的检测框特征和所述维度统一且平均化后的全局特征获取所述第二权重。

12. 根据权利要求 8-11 中任一项所述的图像问答装置，其特征在于，所述特征提取器包括：

检测框生成器，所述检测框生成器配置有用于获得包围所述图像中的对象的多个检测框的高速区域卷积神经网络；

检测框选择器，所述检测框选择器根据所述多个检测框所包围的对象与所述图像的背景的差异确定至少一个检测框；

检测框子特征提取器，所述检测框子特征提取器基于所述至少一个检测框提取获得至少一个检测框子特征；

特征合成器，所述特征合成器基于所述至少一个检测框子特征获得所述检测框特征。

13. 根据权利要求 12 所述的图像问答装置，其特征在于，所述权重计算器包括：

第一维度统一器，所述第一维度统一器统一所述问题特征、所述全局特征和所述检测框特征的维度；

检测框均值器，所述检测框均值器将所述维度统一后的检测框特征基于所述多个检测框子特征的数量平均化；

第一权重获取器，所述第一乘法器基于所述维度统一后的问题特征、所述维度统一后的全局特征和所述维度统一且平均化后的检测框特征获取所述第一权重。

14. 根据权利要求 8-13 中任一项所述的图像问答装置，其特征在于，所述预测器包括：

区域预测器，所述区域预测器融合所述问题特征与所述区域注意力特征，得到第一问题预测答案；

检测框预测器，所述检测框预测器融合所述问题特征与所述检测框注意力特征，得到第二问题预测答案；

双模分析器，通过对所述第一问题预测答案和所述第二问题预测答案进行分类的方式，获得所述问题的答案。

15. 一种图像问答系统，其特征在于，所述图像问答系统包括：

存储器，存储可执行指令；

一个或多个处理器，与所述存储器通信以执行可执行指令从而完成如权利要求 1-7

任一所述的图像问答方法对应的操作。

16. 一种电子设备，其特征在于，包括处理器，所述处理器包括权利要求 8 至 14 任意一项所述的图像问答装置。

17. 一种电子设备，其特征在于，包括：存储器，用于存储可执行指令；

处理器，用于与所述存储器通信以执行所述可执行指令从而完成权利要求 1 至 7 任意一项所述图像问答方法。

18. 一种计算机可读存储介质，用于存储计算机可读指令，其特征在于，所述计算机可读指令被执行时执行如权利要求 1 至 7 任意一项所述的图像问答方法对应的操作。

19. 一种计算机程序产品，包括计算机可读代码，其特征在于，当所述计算机可读代码在设备上运行时，所述设备中的处理器执行用于实现权利要求 1 至 7 任意一项所述的图像问答方法的指令。