METHOD AND DEVICE FOR TRAINING DOCUMENT REPRESENTATION MODEL, AND STORAGE MEDIUM

说　　明　　书

文档表征模型的训练方法、设备及存储介质

技术领域

本说明书涉及人工智能技术领域，尤其涉及一种文档表征模型的训练方法、设备及存储介质。

背景技术

随着人工智能技术的发展，很多场景需要对文档内容进行智能理解和处理。以风控场景为例，为了防控交易中的风险，风控平台需要验证交易的真实性。比如，在用户针对目标交易发起支付请求时，风控平台要求用户提供能够证明该目标交易的真实性的证明材料。上述证明材料包括但不限于：证照材料、物流单、订单合同等。风控平台对上述证明材料的内容进行理解，并从中提取风控所需的关键信息，以便用于对目标交易进行风险识别。

上述场景中，对文档内容和理解和处理可以通过机器学习模型实现。机器学习模型的训练过程通常采用两阶段的训练策略。在第一训练阶段中，利用大规模的通用文档数据集（与特定下游任务无关）训练得到一个模型（该模型通常称为预训练模型），让训练后的模型具有文档理解/表征能力。在第二训练阶段，利用小规模的特定下游任务（比如从文档中提取关键信息的任务）对应的文档数据集对上述模型进一步微调（fine-tuned），使得该模型适用于该特定下游任务。

模型在第一训练阶段学习到的对文档的理解/表征能力，将会影响该模型在具体下游任务的性能。因此，需要提供一种能够提升对文档理解/表征能力的模型训练方案。

背景技术部分的内容仅仅是发明人个人所知晓的信息，并不代表上述信息在本公开申请日之前已经进入公共领域，也不代表其可以成为本公开的现有技术。

发明内容

本说明书提供一种文档表征模型的训练方法、设备及存储介质，能够提升训练后的文档表征模型对文档内容的理解/表征能力。

第一方面，本说明书提供一种文档表征模型的训练方法，包括：获得文档图像，所述文档图像中包括多个文本单元，且被划分为多个图像块；在所述文档图像中确定多个训练任务的预测对象，所述预测对象对应于所述文本单元或所述图像块；从文本维度上生成所述文档图像对应的文本特征，并从视觉维度上生成所述文档图像对应的视觉特征，其中，所述视觉特征是通过在所述文档图像中对所述多个训练任务的预测对象对应的区域进行视觉标记，并对所述视觉标记后的图像进行视觉特征提取得到的，不同的训练任务的预测对象对应的视觉标记方式不同；以及基于所述文本特征和所述视觉特征，对所述文档表征模型执行所述多个训练任务，以对所述文档表征模型进行训练。

第二方面，本说明书还提供一种文档表征模型的训练设备，包括至少一个存储介质和至少一个处理器，其中，所述至少一个存储介质存储有至少一个指令集，用于进行文档表征模型的训练，所述至少一个处理器同所述至少一个存储介质通信连接，其中，所述至少一个处理器运行时读取所述至少一个指令集，并且根据所述至少一个指令集的指示：获得文档图像，所述文档图像中包括多个文本单元，且被划分为多个图像块，在所述文档图像中确定多个训练任务的预测对象，所述预测对象对应于所述文本单元或所述图像块，从文本维度上生成所述文档图像对应的文本特征，并从视觉维度上生成所述文档图像对应的视觉特征，其中，所述视觉特征是通过在所述文档图像中对所述多个训练任务的预测对象对应的区域进行视觉标记，并对所述视觉标记后的图像进行视觉特征提取得到的，不同的训练任务的预测对象对应的视觉标记方式不同，以及基于所述文本特

征和所述视觉特征，对所述文档表征模型执行所述多个训练任务，以对所述文档表征模型进行训练。

第三方面，本说明书还提供一种计算机可读非暂时性存储介质，其中，所述计算机可读非暂时性存储介质中存储有至少一个指令集，所述至少一个指令集被至少一个处理器执行时，实现如第一方面中任一项所述的方法。

由以上技术方案可知，本说明书提供的文档表征模型的训练方法、设备及存储介质，能够从文本维度生成文档图像对应的文本特征，从视觉维度上生成文档图像对应的视觉特征，并基于上述文本特征和上述视觉特征对文档表征模型执行多个训练任务，从而实现对文档表征模型的训练。其中，上述视觉特征是通过在文档图像中对多个训练任务的预测对象对应的区域进行视觉标记，并对所述视觉标记后的图像进行视觉特征提取得到的，不同的训练任务的预测对象对应的视觉标记方式不同。这样，通过在视觉维度上对各训练任务的预测对象采用不同方式进行视觉标记，使得文档表征模型能够从额外的视觉标记中学习更多的信息，有助于模型捕捉文档图像中的不同区域、不同维度之间的内在联合分布，从而提高对文档内容的理解，有助于提高文档表征模型对文档内容的理解/表征性能。

本说明书提供的文档表征模型的训练方法、设备及存储介质的其他功能将在以下说明中部分列出。本说明书提供的文档表征模型的训练方法、设备及存储介质的创造性方面可以通过实践或使用下面详细示例中所述的方法、装置和组合得到充分解释。

附图说明

为了更清楚地说明本说明书实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本说明书的一些实施例，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他的附图。

图 1 示出了根据本说明书的实施例提供的一种模型训练场景的示意图；

图 2 示出了根据本说明书的实施例提供的一种计算设备的硬件结构图；

图 3 示出了根据本说明书的实施例提供的一种文档表征模型的训练方法的流程图；

图 4 示出了根据本说明书的实施例提供的一种文档图像及其文本单元、图像块的示意图；

图 5 示出了根据本说明书的实施例提供的文档表征模型的训练过程的示意图；

图 6 示出了根据本说明书的实施例提供的对文档图像进行视觉标记的过程示意图；

图 7 示出了根据本说明书的实施例提供的文本单元对应的局部视觉嵌入表示的获得方式的示意图；以及

图 8 示出了根据本说明书的实施例提供的文本维度对视觉维度造成信息泄露的示意图。

具体实施方式

以下描述提供了本说明书的特定应用场景和要求，目的是使本领域技术人员能够制造和使用本说明书中的内容。对于本领域技术人员来说，对所公开的实施例的各种局部修改是显而易见的，并且在不脱离本说明书的精神和范围的情况下，可以将这里定义的一般原理应用于其他实施例和应用。因此，本说明书不限于所示的实施例，而是与权利要求一致的最宽范围。

这里使用的术语仅用于描述特定示例实施例的目的，而不是限制性的。比如，除非上下文另有明确说明，这里所使用的，单数形式“一”，“一个”和“该”也可以包括复数形式。当在本说明书中使用时，术语“包括”、“包含”和/或“含有”意思是指所关联的整数，步骤、操作、元素和/或组件存在，但不排除一个或多个其他特征、整数、步骤、操作、元素、组件和/或组的存在或在该系统/方法中可以添加其他特征、整数、步骤、操作、元

素、组件和/或组。

考虑到以下描述，本说明书的这些特征和其他特征、以及结构的相关元件的操作和功能、以及部件的组合和制造的经济性可以得到明显提高。参考附图，所有这些形成本说明书的一部分。然而，应该清楚地理解，附图仅用于说明和描述的目的，并不旨在限制本说明书的范围。还应理解，附图未按比例绘制。

本说明书中使用的流程图示出了根据本说明书中的一些实施例的系统实现的操作。应该清楚地理解，流程图的操作可以不按顺序实现。相反，操作可以以反转顺序或同时实现。此外，可以向流程图添加一个或多个其他操作。可以从流程图中移除一个或多个操作。

为了方便描述，首先对本说明书后文中将会出现的术语进行解释。

自监督学习：自监督学习是机器学习中的一种学习方式。在这种学习方式中，模型使用的数据为未经标注的数据。模型能够从数据本身生成伪标签，模型通过预测这些伪标签来学习数据的有用特征。这种学习方式不依赖于大量的人工标注数据，可以更有效地利用大规模未标记数据集。

预训练：预训练是一种训练策略。在大规模的通用数据集上训练一个模型，这个通用数据集通常与特定下游任务无关，其训练的目的是让模型学习到泛用的特征或语言理解能力，以便这个模型后续可以被进一步微调（fine-tuned）应用于特定的下游任务。这种模型训练过程称为预训练，通过上述训练过程得到的模型可以被称为预训练模型。

下游任务：在机器学习中，下游任务指的是在预训练模型的基础上，对预训练模型进行微调以优化特定任务的性能的过程。下游任务通常是更具体、更有目的性的任务，如情感分析、文本分类、目标检测、信息提取等，它们需要对预训练后的模型进行专门的调整以达到较佳效果。

文档，本说明书中，文档的示例可以包括但不限于：票据、单据、信函、合同、文

书、公文、报表、账单、处方、物流单、证照等。

模态：在人工智能和多模态学习中，模态指的是数据的一种类型或形式。常见的模态包括但不限于文本、图像、音频和视频等。每种模态携带不同类型的信息，而多模态学习涉及融合和理解这些不同类型的信息来提高机器学习模型的性能。

如前所述，在文档智能理解和处理场景下，模型的训练过程通常采用两阶段的训练策略。下面结合图 1 对上述场景下的模型训练策略进行介绍。

图 1 示出了根据本说明书的实施例提供的一种模型训练场景 100 的示意图。如图 1 所示，该场景 100 中，模型训练过程可以包括第一训练阶段和第二训练阶段。

参见图 1，在第一训练阶段中，第一训练设备 110 利用第一文档数据集 130 进行模型训练，让训练后的模型具有文档理解/表征能力。即，训练后的模型能够对文档内容进行理解，并生成该文档内容对应的特征向量。该训练后的模型被称为文档表征模型。在一些场景中，上述第一训练阶段也可以称为预训练阶段，该训练后的模型也可以被称为预训练模型。

其中，第一文档数据集 130 包括多个样本文档。第一文档数据集可以是通用的文档数据集，与特定的下游任务无关。即，第一文档数据集 130 中的样本文档可以涵盖多种领域的文档。第一文档数据集 130 是未经标注的数据集。

第一训练设备 110 可以是具有一定计算能力的电子设备。第一训练设备 110 可以是单个实体设备，也可以是由多个实体设备组成的设备集群。第一训练设备 110 可以执行本说明书提供的文档表征模型的训练方法。第一训练设备 110 可以存储有执行本说明书提供的文档表征模型的训练方法的数据或指令，并可以执行或用于执行所述数据或指令。第一训练设备 110 可以包括具有数据信息处理功能的硬件设备和驱动该硬件设备工作所需的程序。本说明书提供的文档表征模型的训练方法将在后文详细描述。

继续参见图 1，第一训练设备 110 在第一训练阶段训练得到的文档表征模型被部署

于第二训练设备 120 中。在第二训练阶段，第二训练设备 120 利用第二文档数据集 140 对文档表征模型进行微调（fine-tuned），使得该文档表征模型适用于该特定下游任务。微调后的模型可以被称为下游任务模型。第二训练阶段的训练也可以称为微调训练。

其中，第二文档数据集 140 可以是小规模的、与特定下游任务对应的文档数据集。第二文档数据集 140 可以包括多个样本文档。第二文档数据集 140 可以是带有标注信息的数据集，也可以是不带标注信息的数据集，本说明书对此不做限定。

第二训练设备 120 可以是具有一定计算能力的电子设备。第二训练设备 120 可以是单个实体设备，也可以是由多个实体设备组成的设备集群。在一些实施例中，第二训练设备 120 可以与第一训练设备 110 对应相同的实体设备。

本说明书提供的文档表征模型的训练方法可以应用于上述第一训练阶段（或者称为预训练阶段）。该方法训练得到的文档表征模型可以理解文档的内容，并生成文档内容对应的特征向量。在第二训练阶段中，该文档表征模型可以被应用于各种不同场景中的下游任务。例如，以风控场景为例，第二训练设备 120 利用风控场景下的文档数据集（即第二文档数据集）对该文档表征模型进行微调，使得文档表征模型能够从目标交易对应的证明材料中提取风控场景所需的关键信息，进而基于提取的关键信息对该目标交易进行风险识别。

需要说明的是，上述文档表征模型除了可应用于风控场景的下游任务之外，还可用于其他场景的下游任务，例如情感分析、文本分类、目标检测等场景，本说明书对此不做限定。

图 2 示出了根据本说明书的实施例提供的一种计算设备 200 的硬件结构图。计算设备 200 可以作为图 1 中的第一训练设备 110，并执行本说明书提供的文档表征模型的训练方法。

如图 2 所示，计算设备 200 可以包括至少一个存储介质 230 和至少一个处理器 220。在一些实施例中，计算设备 200 还可以包括通信端口 250 和内部通信总线 210。计算设

备 200 还可以包括 I/O 组件 260。

内部通信总线 210 可以连接不同的系统组件。例如，内部通信总线 210 可以连接存储介质 230、处理器 220、通信端口 250 和 I/O 组件 260 等。

I/O 组件 260 支持计算设备 200 和其他组件之间的输入/输出。

通信端口 250 用于计算设备 200 同外界的数据通信。比如，通信端口 250 可以用于计算设备 200 同网络之间的数据通信。通信端口 250 可以是有线通信端口也可以是无线通信端口。计算设备 200 可以通过通信端口 250 读取第一文档数据集中的数据。计算设备 200 还可以通过通信端口 250 将训练好的文档表征模型部署到第二训练设备 120 中。

存储介质 230 可以包括数据存储装置。所述数据存储装置可以是非暂时性存储介质，也可以是暂时性存储介质。比如，所述数据存储装置可以包括磁盘 232、只读存储介质（ROM）234 或随机存取存储介质（RAM）236 中的一种或多种。存储介质 230 还包括存储在所述数据存储装置中的至少一个指令集。所述指令集可以包括计算机程序代码，所述计算机程序代码可以包括程序、例程、对象、组件、数据结构、过程、模块等等。

至少一个处理器 220 可以同至少一个存储介质 230 通信连接。当计算设备 200 运行时，至少一个处理器 220 读取所述至少一个指令集，并且根据所述至少一个指令集的指示，执行本说明书提供的文档表征模型的训练方法。所述至少一个处理器 220 可以执行文档表征模型的训练方法包含的步骤。处理器 220 可以是一个或多个处理器的形式，在一些实施例中，处理器 220 可以包括一个或多个硬件处理器，例如微控制器，微处理器，精简指令集计算机（RISC），专用集成电路（ASIC），特定于应用的指令集处理器（ASIP），中心处理单元（CPU），图形处理单元（GPU），物理处理单元（PPU），微控制器单元，数字信号处理器（DSP），现场可编程门阵列（FPGA），高级 RISC 机器（ARM），可编程逻辑器件（PLD），能够执行一个或多个功能的任何电路或处理器等，或其任何组合。

仅仅为了说明问题，附图中计算设备 200 仅示出了一个处理器 220。然而，应当注意，本说明书中的计算设备 200 还可以包括多个处理器，因此，本说明书中披露的操作

和/或方法步骤可以由一个处理器执行，也可以由多个处理器联合执行。例如，如果在本说明书中描述计算设备 200 的处理器 220 执行步骤 A 和步骤 B，则应该理解，步骤 A 和步骤 B 也可以由两个不同处理器 220 联合或分开执行（例如，第一处理器执行步骤 A，第二处理器执行步骤 B，或者第一和第二处理器共同执行步骤 A 和 B）。

图 3 示出了根据本说明书的实施例提供的一种文档表征模型的训练方法 P300 的流程图。如前，计算设备 200 可以执行文档表征模型的训练方法 P300。具体地，计算设备 200 中的处理器 220 可以读取存储在其本地存储介质中的指令集，然后根据指令集的规定，执行文档表征模型的训练方法 P300。

如图 3 所示，文档表征模型的训练方法 P300 可以包括步骤 S310 至 S340。

S310：获得文档图像，所述文档图像中包括多个文本单元，并被划分为多个图像块。

其中，所述文档图像可以为第一文档数据集中的样本文档对应的电子形式的文件。例如，当所述样本文档为电子文档时，所述文档图像可以是所述样本文档本身。当所述样本文档为纸质文档时，所述文档图像可以是对所述样本文档进行扫描或者拍摄得到的电子文件。在一些实施例中，第一文档数据集中的样本文档可以以文档图像的形式存储。在 S310 中，计算设备 200 可以从第一文档数据集中获得所述文档图像。

文本单元可以指构成所述文档图像中的文本内容的任何基本或有意义的组成部分。文本单元可以视为将文档图像中复杂的文本内容分解为更小、更易于管理和分析的单元。在一些实施例中，文本单元可以对应于字符、词汇、短语、句子、段落中的一种或多种。在一些实施例中，可以通过文本识别技术对文档图像进行文本识别，从而得到文本图像中的多个文本单元，还可以得到各文本单元的位置信息。上述位置信息可以包括一维位置信息、二维位置信息中的至少一种。例如，上述文本识别技术可以为光学字符识别（Optical Character Recognition，OCR）技术，也可以为其他技术。

图像块可以指文档图像中的一个局部区域。在一些实施例中，可以通过图像分割技术将文档图像划分为多个图像块。所述图像块的大小可以是固定的或可变的。本说明书

对于图像块的大小、以及图像块的数量均不做限定。

下面结合图 4 对文档图像中的文本单元和图像块进行举例说明。

图 4 示出了根据本说明书的实施例提供的一种文档图像及其文本单元、图像块的示意图。如图 4 所示，以出租车发票为例，图 4 的子图（a）示出的是文档图像 400。通过 OCR 技术对文档图像 400 进行文本识别，可以得到多个文本单元 410。图 4 的子图（b）中每个文本框中的内容可以对应一个文本单元 410。通过将文档图像 400 划分为 3 行 3 列，可以得到 9 个图像块 420，如图 4 的子图（c）所示。

S320：在所述文档图像中确定多个训练任务的预测对象，所述预测对象对应于所述文本单元或所述图像块。

本说明书的实施例提供的文档表征模型的训练方法，采用自监督学习方式，利用无标注的文档图像进行模型训练，使得训练后的模型能够对文档内容进行表征。由于文档图像是不带标注信息的，因此，需要预先设计一系列的训练任务，这些训练任务可以基于文档图像本身的特征，设计预定义问题来让模型预测文档图像中的某部分信息或重建缺失信息，使得模型能够在无标注的文档图像上进行自我学习。由此可见，上述训练任务能够从文档图像本身构造自监督信息，让模型基于上述自监督信息进行自我监督学习。上述训练任务也可以称为自监督任务。

在步骤 S320 中，所述多个训练任务可以包括：图像掩蔽任务、文本掩蔽任务、以及文本-图像对齐任务。

图像掩蔽任务可以为：在视觉维度上，将文档图像中的一部分图像块掩蔽，让模型基于其他未被掩蔽的图像块的文本信息和布局信息预测被掩蔽图像块的特征。上述特征可以为图像块的原始像素，也可以为图像块的方向梯度直方图（Histogram of Oriented Gradient, HOG）特征，还可以为图像块的其他特征。在一些实施例中，图像掩蔽任务也可以称为掩蔽图像模型（Masked-Image-Modeling，MIM）。

文本掩蔽任务可以为：在文本维度上，将文档图像中的一部分文本单元的文本信息掩蔽，让模型基于其他未被掩蔽的文本单元的文本信息和布局信息预测被掩蔽文本单元的文本信息。在一些实施例中，文本掩蔽任务也可以称为掩蔽语言模型（Masked-Language-Modeling，MLM）。根据本说明书后文的描述，由于文本维度上生成的文本特征不仅考虑了文本维度的特征，还考虑了局部视觉特征，即，考虑了多个模态的信息，因此，文本掩蔽任务也可以称为多模态掩蔽语言模型（Multimodal Masked-Language-Modeling，MMLM）。

文本-图像对齐（Text Image Alignment，HTIA）任务可以为：在视觉维度上对文档图像中的某些文本单元做标记，在文本维度上预测各文本单元是否在视觉维度上被标记。根据本说明书后文的描述，在对文档图像进行视觉标记时采用了分层标记策略，因此，文本-图像对齐任务也可以称为分层文本-图像对齐（Hierarchical Text Image Alignment，HTIA）任务。

在步骤 S320 中，计算设备 200 可以在文档图像中确定多个训练任务的预测对象。所述预测对象对应于文档图像中的文本单元或者图像块。

例如，图像掩蔽任务的预测对象可以包括：文档图像中的至少一个目标图像块（即图像掩蔽任务需要被掩蔽的图像块）。例如，计算设备 200 可以在文档图像的多个图像块中随机选择部分图像块作为上述至少一个目标图像块。在模型训练过程中，图像掩蔽任务会要求模型预测出所述至少一个目标图像块对应的预测特征，而这些目标图像块对应的实际特征即可以作为图像掩蔽任务的自监督信息。

例如，文本掩蔽任务的预测对象可以包括：文档图像中的至少一个第一文本单元（即文本掩蔽任务需要被掩蔽的文本单元）。例如，计算设备 200 可以在文档图像的多个文本单元中随机选择部分文本单元作为上述至少一个第一文本单元。在模型训练过程中，文本掩蔽任务会要求模型预测出所述至少一个第一文本单元对应的预测文本信息，而这些第一文本单元对应的实际文本信息即可以作为文本掩蔽任务的自监督信息。

例如，文本-图像对齐任务的预测对象可以包括：文档图像中的至少一个第二文本单元（即上文中在视觉维度上标记的文本单元）。例如，计算设备 200 可以在文档图像的多个文本单元随机选择部分文本单元作为上述的至少一个第二文本单元。在模型训练过程中，文本-图像对齐任务会要求模型预测出文档图像中的各文本单元是否在视觉维度上被标记。而这些文本单元在视觉维度上的实际标记信息即可以作为文本-图像任务的自监督信息。

S330：从文本维度上生成所述文档图像对应的文本特征，并从视觉维度上生成所述文档图像对应的视觉特征，其中，所述视觉特征是通过在所述文档图像中对所述多个训练任务的预测对象对应的区域进行视觉标记，并对所述视觉标记后的图像进行视觉特征提取得到的，不同的训练任务的预测对象对应的视觉标记方式不同。

其中，文本特征是以文本单元为粒度的特征，即，文本特征中可以包括多个文本单元各自对应的多个维度的嵌入表示。视觉特征是以图像块为粒度的特征，即，视觉特征中可以包括多个图像块各自对应的多个维度的嵌入表示。

下面结合图 5 对文本特征和视觉特征的内容及其生成方式进行详细说明。

图 5 示出了根据本说明书的实施例提供的文档表征模型的训练过程的示意图。如图 5 所示，文档表征模型可以为双流架构的多模态表征模型。即，文档表征模型可以有两路特征输入，其中一路特征输入为文档图像对应的文本特征，另一路特征输入为文档图像对应的视觉特征。文档表征模型可以基于上述两路特征输入对文档图像进行多模态分析，得到文档图像对应的多模态表征。

需要说明的是，本说明书对于文档表征模型的具体结构不做限定。例如，在一些实施例中，文档表征模型可以为 Transformer 模型。在另一些实施例中，文档表征模型还可以为其他模型。

参见图 5，文档表征模型的训练过程大致如下：计算设备 200 通过特征提取器对文档图像进行特征提取，例如，从文本维度提取得到文本特征，并从视觉维度提取得到视

觉特征。进而，计算设备 200 可以通过文档表征模型基于上述文本特征和视觉特征，得到文档图像对应的文档预测表征。进而，计算设备 200 基于文档预测表征执行多个训练任务，以实现对文档表征模型的训练。

下面先对所述视觉特征的生成方式及其内容进行说明。

继续参见图 5，子图（a）示出的是文档图像，计算设备 200 可以在文档图像中对上述多个训练任务的预测对象对应的区域进行视觉标记，得到视觉标记后的图像。所述视觉标记后的图像如子图（b）所示。在上述视觉标记的过程中，不同的训练任务的预测对象可以对应不同的视觉标记方式。

在本说明书中，视觉标记可以是指基于视觉的标记，标记后的图像与标记前的图像在视觉上有所不同。对一个区域进行视觉标记的方式可以为下述中的一种：颜色填充、图案填充等。

以颜色填充为例，在一些实施例中，计算设备 200 可以采用如下方式对文档图像进行视觉标记：将文档图像中与所述至少一个目标图像块（即图像掩蔽任务中需要被掩蔽的图像块）对应的区域填充为第一颜色；将文档图像中与所述至少一个第一文本单元（即文本掩蔽任务中需要被掩蔽的文本单元）对应的区域填充为第二颜色；将文档图像中与所述至少一个第二文本单元（即文本-图像对齐任务中需要被标记的文本单元）对应的区域填充为第三颜色。上述第一颜色、第二颜色和第三颜色互不相同。

下面结合图 6，以第一颜色为黑色、第二颜色为蓝色，第三颜色为红色为例，对文档图像的视觉标记过程进行举例说明。

图 6 示出了根据本说明书的实施例提供的对文档图像进行视觉标记的过程示意图。如图 6 所示，子图（a）示出的是未经视觉标记的文档图像。在子图（a）的基础上，将图像掩蔽任务需要掩蔽的图像块（即至少一个目标图像块）对应的区域填充为黑色（即第一颜色）得到子图（b）。在子图（b）的基础上，将文本掩蔽任务需要掩蔽的文本单元（即至少一个第一文本单元）对应的区域填充为蓝色（即第二颜色）得到子图（c）。在

子图（c）的基础上，将文本-图像对齐任务需要标记的文本单元（即至少一个第二文本单元）对应的区域填充为红色（即第三颜色）得到子图（d）。在图 5 中，标注有 Blue 的灰色矩形框表示的是文本单元被填充为蓝色，标注有 Red 的灰色矩形框表示的是文本单元被填充为红色。经过上述视觉标记过程之后，得到的子图（d）即为文档图像对应的视觉标记后的图像。

需要说明的是，本说明书对上述第一颜色、第二颜色和第三颜色的具体取值不做限定，图 6 所示的标记方式仅为一种可能的示例。

继续参见图 5，在得到视觉标记后的图像之后，计算设备 200 可以基于所述视觉标记后的图像中的各图像块，生成所述文档图像对应的视觉特征。其中，所述视觉特征可以包括：各图像块各自对应的视觉嵌入表示，各图像块各自对应的一维位置嵌入表示，各图像块各自对应的二维位置嵌入表示，以及各图像块各自对应的类型嵌入表示。

例如，结合图 5，假设图像块的数量为 m。计算设备 200 通过对 m 个图像块进行视觉特征提取，得到 m 个图像块各自对应的视觉嵌入表示，分别为：V_{G1}、V_{G2}、……、V_{Gm}。在一些实施例中，在对图像块进行视觉特征提取时可以使用线性操作或者线性变换来从图像块中提取特征。这样可以简化计算量，提高特征提取效率。

需要说明的是，本说明书提供的技术方案中，被掩蔽的图像块的视觉嵌入表示、以及未被掩蔽的图像块的视觉嵌入表示，均是通过对视觉标记后的图像块进行视觉特征提取得到的。在图 5 中，第 3 个图像块对应的视觉嵌入表示V_{G3}被标记为[Mask#]，其目的是体现出第 3 个图像块为被掩蔽的图像块。本领域技术人员应当理解的是，第 3 个图像块对应的视觉嵌入表示V_{G3}也是通过对视觉标记后的图像块进行视觉特征提取得到的，而不应该理解为直接将第 3 个图像块对应的视觉嵌入表示V_{G3}替换为掩蔽标识[Mask]。

计算设备 200 基于 m 个图像块的位置信息，可以生成 m 个图像块各自对应的一维位置嵌入表示，分别为：$1D_1$、$1D_2$、……、$1D_m$。例如，计算设备 200 可以基于 m 个图像块的位置信息，按照从左到右、从上到下的顺序生成 m 个图像块的索引标号，进而基

于 m 个图像块的索引编号生成 m 个图像块对应的一维位置嵌入表示。比如 m 个图像块的索引编号分别为：0、1、2、……、m-1。

计算设备 200 基于 m 个图像块的位置信息，可以生成 m 个图像块各自对应的二维位置嵌入表示，分别为：$2D_1$、$2D_2$、……、$2D_m$。例如，针对某个图像块，可以基于该图像块的左上顶点的坐标信息、右下顶点的坐标信息、以及图像块的高度和宽度，生成该图像块对应的二维位置嵌入表示。

进一步的，为了让文档表征模型能够区分出视觉特征和文本特征，可以预先定义视觉特征对应的特征类型，以及文本特征对应的特征类型。例如，视觉特征对应的特征类型可以为 0，文本特征对应的特征类型可以为 1。所述视觉特征除了包括各图像块对应的视觉嵌入表示、一维位置嵌入表示、二维位置嵌入表示之外，还可以包括各图像块各自对应的类型嵌入表示。例如，各图像块各自对应的类型嵌入表示可以为 0。

下面再对所述文本特征的生成方式及其内容进行说明。

继续参见图 5，计算设备 200 还可以对文档图像进行文本识别，得到文档图像中的多个文本单元的文本内容，还可以得到所述多个文本单元的位置信息。上述文本识别的过程可以采用 OCR 技术实现，也可以采用其他技术实现，本说明书对此不做限定。

进而，计算设备 200 可以基于所述多个文本单元，生成所述文档图像对应的文本特征。其中，所述文本特征可以包括：所述多个文本单元各自对应的文本嵌入表示，所述多个文本单元各自对应的一维位置嵌入表示，所述多个文本单元各自对应的二维位置嵌入表示，所述多个文本单元各自对应的局部视觉嵌入表示，以及所述多个文本单元各自对应的类型嵌入表示。上述文本特征不仅可以描述文本及布局维度的特征，还可以表达局部视觉维度的特征，使文本特征的表达更加丰富全面。

例如，结合图 5，假设所述多个文本单元的数量为 n。计算设备 200 可以基于 n 个文本单元的文本内容生成所述 n 个文本单元各自对应的文本嵌入表示，分别为：T_1、T_2、……、T_n。

继续参见图 5，基于所述多个训练任务的文本掩蔽需求，在文本特征中，部分文本单元对应的文本嵌入表示被掩蔽处理。例如，以文本掩蔽任务为例，假设文本掩蔽任务需要掩蔽的文本单元包括第 1 个文本单元和第 4 个文本单元。该情况下，参见图 5，文本特征中与第 1 个文本单元对应的文本嵌入表示T_1被替换为掩蔽标识[Mask]，并且，文本特征中与第 4 个文本单元对应的文本嵌入表示T_4被替换为掩蔽标识[Mask]。

计算设备 200 基于 n 个文本单元的位置信息，可以生成 n 个文本单元各自对应的一维位置嵌入表示，分别为：$1D_1$、$1D_2$、……、$1D_n$。例如，计算设备 200 可以基于 n 个文本单元的位置信息，按照从左到右、从上到下的顺序生成 n 个文本单元的索引标号，进而基于 n 个文本单元的索引编号生成 n 个文本单元对应的一维位置嵌入表示。比如 n 个文本单元的索引编号分别为：0、1、2、……、n-1。

在一些场景下，当文档图像中的文本单元的数量较多，即 n 取值较大时，基于 n 个文本单元的索引编号生成 n 个文本单元对应的一维位置嵌入表示，可能会出现由于索引编号取值较大，一维位置嵌入表达的位数不足以支持表达的情况。为了避免上述情况，在一些实施例中，计算设备 200 可以对所述多个文本单元的位置信息进行相对位置编码或者三角函数位置编码，从而得到所述多个文本单元各自对应的一维位置嵌入表示。这样，所述多个文本单元各自对应的一维位置嵌入表示反映的是各文本单元之间的相对位置关系或者三角函数位置关系。从而，文档表征模型能够支持对长文本文档的表征，而不需要对长文本文档进行截断处理，提高文档表征精度，并提高文档表征效率。

计算设备 200 基于 n 个文本单元的位置信息，可以生成 n 个文本单元各自对应的二维位置嵌入表示，分别为：$2D_1$、$2D_2$、……、$2D_n$。例如，针对某个文本单元，可以基于该文本单元边界框（bounding box）的左上顶点的坐标信息、右下顶点的坐标信息、以及文本单元边界框的高度和宽度，生成该文本单元对应的二维位置嵌入表示。

继续参见图 5，每个文本单元对应的局部视觉嵌入表示是通过在所述文档图像中对该文本单元对应的图像区域进行空洞卷积得到的。空洞卷积也可以称为扩张卷积、扩展

卷积或者膨胀卷积，是卷积神经网络中的一种卷积操作。传统卷积操作的感受野随卷积核尺寸增加而增大。而空洞卷积中，卷积核的元素之间的距离被扩大，引入了空洞或者空隙（即 0 值），增加了感受野而无需增加参数量。

上述方式得到的文本单元对应的局部视觉嵌入表示，反映的是包含该文本单元的局部区域的视觉特征，该局部区域的中心与所述文本单元的中心相同，且所述局部区域的尺寸大于所述文本单元的尺寸。

下面结合图 7 进行举例说明。图 7 示出了根据本说明书的实施例提供的文本单元对应的局部视觉嵌入表示的获得方式的示意图。如图 7 所示，以其中一个文本单元 410 为例，计算设备 200 通过在文档图像中对文本单元 410 对应的区域进行空洞卷积，得到文档单元 410 对应的局部视觉特征，该局部视觉特征反映的是包含该文本单元 410 的局部区域 410’的视觉特征。局部区域 410’的中心与文本单元 410 的中心相同，且局部区域 410’的尺寸大于文本单元 410 的尺寸。局部区域 410’可以视为对文本单元 410 进行一定的视野扩大之后的区域。

上述方式中，在确定文本单元对应的局部视觉嵌入表示时，不仅考虑了文本单元的内部区域的视觉信息，还考虑了文本单元的周边区域的视觉信息，使得确定出的文本单元对应的局部视觉嵌入表示更加准确。上述方式提取的局部视觉特征与文本单元是自然对齐的。通过在文本特征中添加各文本单元对应的局部视觉嵌入，使得模型能够基于局部视觉信息来辅助文本理解，而不是仅仅基于文本内容，从而有助于提高模型对文档内容的理解和表征能力。

进一步的，为了让文档表征模型能够区分出视觉特征和文本特征，可以预先定义视觉特征对应的特征类型，以及文本特征对应的特征类型。例如，视觉特征对应的特征类型可以为 0，文本特征对应的特征类型可以为 1。所述文本特征除了包括各文本单元对应的文本嵌入表示、一维位置嵌入表示、二维位置嵌入表示、局部视觉嵌入表示之外，还可以包括各文本单元各自对应的类型嵌入表示。例如，各文本单元各自对应的类型嵌

入表示可以为 1。

另外，需要说明的是，在图 5 示出的文本特征中，“[CLS]”、“[Mask]”、“[SEP]”和“[$2D_{PAD}$]”是特殊标记。

S340：基于所述文本特征和所述视觉特征，对所述文档表征模型执行所述多个训练任务，以对所述文档表征模型进行训练。

在得到文档图像对应的文本特征和视觉特征之后，计算设备 200 可以基于所述文本特征和所述视觉特征，对文档表征模型执行上述多个训练任务，从而实现对文档表征模型的训练。

在一些实施例中，上述 S340 的训练过程可以包括下述步骤（1）至（4）

（1）通过文档表征模型基于所述文本特征和所述视觉特征，得到所述文档图像对应的文档预测表征。

在一些实施例中，计算设备 200 可以基于所述文本特征和所述视觉特征，生成融合特征。进而，计算设备 200 将所述融合特征输入到文档表征模型中，以使文档表征模型输出所述文档图像对应的文档预测表征。

如前文描述，所述文本特征中可以包括多个文本单元各自对应的多个维度的嵌入表示，所述视觉特征中可以包括多个图像块各自对应的多个维度的嵌入表示。因此，计算设备 200 在生成融合特征时可以采用如下方式：

（1-1）针对所述多个文本单元中的每个文本单元，对该文本单元对应的多个维度的嵌入表示相加，得到该文本单元对应的第一嵌入表示。

例如，结合图 5，文本特征中包括：n 个文本单元各自对应的文本嵌入表示、一维位置嵌入表示、二维位置嵌入表示、局部视觉嵌入表示、以及类型嵌入表示。针对第 1 个文本单元，计算设备 200 可以对第 1 个文本单元对应的上述各维度的嵌入表示相加，得到第 1 个文本单元对应的第一嵌入表示；针对第 2 个文本单元，计算设备 200 可以对

第 2 个文本单元对应的上述各维度的嵌入表示相加，得到第 2 个文本单元对应的第一嵌入表示；以此类推，针对第 n 个文本单元，计算设备 200 可以对第 n 个文本单元对应的上述各维度的嵌入表示相加，得到第 n 个文本单元对应的第一嵌入表示。

（1-2）针对所述多个图像块中的每个图像块，对该图像块对应的多个维度的嵌入表示相加，得到该图像块对应的第二嵌入表示。

例如，结合图 5，视觉特征中包括：m 个图像块各自对应的视觉嵌入表示、一维位置嵌入表示、二维位置嵌入表示、以及类型嵌入表示。针对第 1 个图像块，计算设备 200 可以对第 1 个图像块对应的上述各维度的嵌入表示相加，得到第 1 个图像块对应的第二嵌入表示；针对第 2 个图像块，计算设备 200 可以对第 2 个图像块对应的上述各维度的嵌入表示相加，得到第 2 个图像块对应的第二嵌入表示；以此类推，针对第 m 个图像块，计算设备 200 可以对第 m 个图像块对应的上述各维度的嵌入表示相加，得到第 m 个图像块对应的第二嵌入表示。

（1-3）对所述多个文本单元各自对应的第一嵌入表示、以及所述多个图像块各自对应的第二嵌入表示进行拼接，得到所述融合特征。

例如，结合图 5，计算设备 200 对 m 个图像块各自对应的第二嵌入表示、以及 n 个文本单元各自对应的第一嵌入表示进行拼接，得到融合特征。

（2）基于所述文档预测表征执行所述多个训练任务，得到所述多个训练任务的预测对象各自对应的预测信息，并基于所述多个训练任务的预测对象各自对应的预测信息和实际信息，生成目标损失。

例如，结合图 5，针对图像掩蔽任务，图像掩蔽任务要求对被掩蔽图像块（图 5 中被填充为黑色的图像块）的 HOG 特征进行预测。计算设备 200 基于所述文档预测表征执行图像掩蔽任务，得到图像掩蔽任务中被掩蔽图像块的预测信息。其中，上述预测信息即为预测出的 HOG 特征。继续参见图 5，在计算设备 200 通过特征提取器对文档图像进行特征提取时，还可以对文档图像的各图像块进行 HOG 变换，从而得到各图像块

对应的 HOG 特征。计算设备 200 可以基于被掩蔽图像块对应的预测信息（即预测出的 HOG 特征）、以及被掩蔽图像块对应的实际信息（即通过 HOG 变换得到的 HOG 特征），确定出图像掩蔽任务的预测损失。

例如，结合图 5，针对文本掩蔽任务，文本掩蔽任务要求对被掩蔽的文本单元（图 5 中的第 1 个文本单元和第 4 个文本单元）的文本内容进行预测。计算设备 200 基于所述文档预测表征执行文本掩蔽任务，得到文本掩蔽任务中被掩蔽文本单元的预测信息。其中，上述预测信息即为预测出的文本内容。计算设备 200 可以基于被掩蔽文本单元对应的预测信息（即预测出的文本内容）、以及被掩蔽文本单元对应的实际信息（即通过 OCR 技术识别得到的文本内容），确定出文本掩蔽任务的预测损失。

例如，结合图 5，针对文本-图像对齐任务，文本-图像对齐任务要求预测各文本单元是否在视觉维度上被标记。计算设备 200 基于所述文档预测表征执行文本-图像对齐任务，得到各文本单元的预测信息。其中，上述预测信息为被标记或者未被标记。计算设备 200 可以基于各文本单元对应的预测信息（即预测出的是否被标记）、以及各文本单元对应的实际信息（即实际是否被标记），确定出文本-图像对齐任务的预测损失。

进一步的，计算设备 200 可以基于图像掩蔽任务的预测损失、文本掩蔽任务的预测损失、以及文本-图像对齐任务的预测损失，确定出目标损失。例如，可以将上述三个训练任务的预测损失进行相加或者加权求和，得到目标损失。

（4）以最小化所述目标损失为训练目标，对文档表征模型的模型参数进行更新。

需要说明的是，本说明书描述的是基于一个文档图像对文档表征模型进行训练的过程。本领域技术人员能够理解，上述训练过程可以被重复执行多次。例如，计算设备 200 可以基于第一文档数据集重复执行上述训练过程，直至达到预设训练结束条件时，得到训练后的文档表征模型。

本说明书提供的技术方案，能够从文本维度生成文档图像对应的文本特征，从视觉维度上生成文档图像对应的视觉特征，并基于上述文本特征和上述视觉特征对文档表征

模型执行多个训练任务，从而实现对文档表征模型的训练。其中，上述视觉特征是通过在文档图像中对多个训练任务的预测对象对应的区域进行视觉标记，并对所述视觉标记后的图像进行视觉特征提取得到的，不同的训练任务的预测对象对应的视觉标记方式不同。这样，通过在视觉维度上对各训练任务的预测对象采用不同方式进行视觉标记，使得文档表征模型能够从额外的视觉标记中学习更多的信息，有助于模型捕捉文档图像中的不同区域、不同维度之间的内在联合分布，从而提高对文档内容的理解，有助于提高文档表征模型对文档内容的理解/表征性能。

另外，上述技术方案中，视觉维度对应的视觉特征中保留了各图像块（无论是被掩蔽图像块还是未被掩蔽图像块）的视觉信息。相较于直接将被掩蔽图像块的视觉信息替换为预设的掩蔽标识[Mask]而言，上述方案不会丢失被掩蔽图像块的视觉信息。

进一步的，在文档图像中针对不同训练任务的预测对象填充不同颜色，很好地解决了由于在视觉维度上需要进行多种掩蔽/标记操作，从而造成的视觉区域冲突的问题。上述视觉标记方式能够有效处理文档图像中不同训练任务的预测对象之间的遮挡。

本说明书提供的技术方案，在视觉维度上，对文本掩蔽任务需要掩蔽的文本单元进行了颜色填充（即填充了第二颜色），实现了在视觉维度上对上述文本单元的覆盖。这样，能够避免视觉维度对文本维度的信息泄露，即，避免文档表征模型从视觉维度“偷看”被掩蔽文本单元的文本内容，使得文档表征模型能够真正从文本维度上、基于未掩蔽文本的文本内容和布局信息，预测被掩蔽文本单元的文本内容，从而提升文档表征模型的性能。

图像掩蔽任务的其中一个训练目标是，希望文档表征模型能够从视觉维度上学习文档图像的视觉表征。但是，文本维度有可能会对视觉维度造成信息泄露。图 8 示出了根据本说明书的实施例提供的文本维度对视觉维度造成信息泄露的示意图。如图 8 所示，文档表征模型在视觉维度上基于未掩蔽图像块的视觉信息和图像块之间的布局信息，能够恢复出被掩蔽图像块的视觉信息。除此之外，文档表征模型在文本维度上已知各文本

单元的文本内容和布局信息的情况下，有可能通过对各文本单元进行“拼图”的方式恢复出图像块的视觉信息。由此可见，让模型单纯依靠图像掩蔽任务学习视觉维度上的特征提取是不充分的，有可能模型只是学习到了如何基于各文本单元的文本内容和布局信息进行“拼图”。

为了避免/降低文本维度对视觉维度的信息泄露，在一些实施例中，计算设备 200 可以将文本特征中与至少一个目标文本单元对应的二维位置嵌入表示和局部视觉嵌入表示随机丢弃（random drop）。其中，所述至少一个目标文本单元为图像掩蔽任务需要被掩蔽图像块中的部分文本单元。上述“丢弃”可以理解为将对应的嵌入表示清除。也就是说，在所述文本特征中，与所述至少一个目标文本单元对应的二维位置嵌入表示和局部视觉嵌入表示被替换为空标识。

例如，结合图 5，计算设备 200 可以在多个文本单元中随机选择位于被掩蔽图像块中的部分文本单元作为上述目标文本单元。比如，计算设备 200 选择第 3、4、6 个文本单元作为上述目标文本单元。进而，计算设备 200 将文本特征中与第 3、4、6 个文本单元对应的二维位置嵌入表示$2D_3$、$2D_4$、$2D_6$和局部视觉嵌入表示V_{L3}、V_{L4}、V_{L6}丢弃，即替换为空标识（图 5 中以叉号表示空标识）。

这样，在执行图像掩蔽任务的时候，至少有一部分图像块是无法借助文本维度上的信息通过“拼图”的方式进行恢复，从而使得模型必须学会在视觉维度上如何基于未掩蔽图像块的视觉信息和布局信息来恢复被掩蔽图像块的视觉信息，提高模型在视觉维度的表征能力。上述方式让模型能够更加细致地理解和融合文本维度与视觉维度之间的联系，增强了模型对于视觉维度的特征表征能力，减少了对文本信息的过度依赖。

在一些情况下，当随机丢弃的信息中包含文本-图像对齐任务所需要标记的文本单元的二维位置嵌入表征（即布局信息）时，在文档表征模型的训练过程中，文档表征模型需要在不知道这些文本单元的布局信息的情况下，预测这些文本单元是否在视觉维度上被标记。如果文档表征模型能够回答正确，说明文档表征模型学习到了如下两种能力：

先在文本维度上基于文本内容预测布局信息，再将文本维度上的布局信息和视觉维度的布局信息对齐。由此可见，文档表征模型不仅学习了到文本内容和文本布局之间的内在联系，还学习到了文本维度与视觉维度的之间的布局信息对齐。从而，增强了文档表征模型在布局上的表征能力。

本说明书另一方面提供一种非暂时性存储介质，存储有至少一个指令集，用来进行文档表征模型的训练。当所述至少一个指令集被处理器执行时，所述至少一个指令集指导所述处理器实施本说明书提供的文档表征模型的训练方法 P300 的步骤。在一些可能的实施方式中，本说明书的各个方面还可以实现为一种程序产品的形式，其包括程序代码。当所述程序产品在计算设备 200 上运行时，所述程序代码用于使计算设备 200 执行本说明书提供的文档表征模型的训练方法 P300 的步骤。用于实现上述方法的程序产品可以采用便携式紧凑盘只读存储器（CD-ROM）包括程序代码，并可以在计算设备 200 上运行。然而，本说明书的程序产品不限于此。在本说明书中，可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统使用或者与其结合使用。所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。所述计算机可读存储介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了可读程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。可读存储介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。可读存储介质上包含的程序代码可以用任何适当的介质传输，包括但不

限于无线、有线、光缆、RF 等等，或者上述的任意合适的组合。可以以一种或多种程序设计语言的任意组合来编写用于执行本说明书操作的程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如 Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在计算设备 200 上执行、部分地在计算设备 200 上执行、作为一个独立的软件包执行、部分在计算设备 200 上部分在远程计算设备上执行、或者完全在远程计算设备上执行。

上述对本说明书特定实施例进行了描述。其他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是可能有利的。

综上所述，在阅读本详细公开内容之后，本领域技术人员可以明白，前述详细公开内容可以仅以示例的方式呈现，并且可以不是限制性的。尽管这里没有明确说明，本领域技术人员可以理解本说明书需求囊括对实施例的各种合理改变，改进和修改。这些改变，改进和修改旨在由本说明书提出，并且在本说明书的示例性实施例的精神和范围内。

此外，本说明书中的某些术语已被用于描述本说明书的实施例。例如，“一个实施例”，“实施例”和/或“一些实施例”意味着结合该实施例描述的特定特征，结构或特性可以包括在本说明书的至少一个实施例中。因此，可以强调并且应当理解，在本说明书的各个部分中对“实施例”或“一个实施例”或“替代实施例”的两个或更多个引用不一定都指代相同的实施例。此外，特定特征，结构或特性可以在本说明书的一个或多个实施例中适当地组合。

应当理解，在本说明书的实施例的前述描述中，为了帮助理解一个特征，出于简化本说明书的目的，本说明书将各种特征组合在单个实施例、附图或其描述中。然而，这并不是说这些特征的组合是必须的，本领域技术人员在阅读本说明书的时候完全有可能

将其中一部分设备标注出来作为单独的实施例来理解。也就是说，本说明书中的实施例也可以理解为多个次级实施例的整合。而每个次级实施例的内容在少于单个前述公开实施例的所有特征的时候也是成立的。

本文引用的每个专利，专利申请，专利申请的出版物和其他材料，例如文章、书籍、说明书、出版物、文件、物品等，除了与本文件不一致或相冲突的、或者对权利要求的最宽范围具有限制性影响的文件，均可以通过引用结合于此，并用于现在或以后与本文件相关联的所有目的。此外，在任何材料对相关术语的描述、定义和/或使用与本文件对相关术语的描述、定义和/或使用之间存在任何不一致或冲突时，以本文件中的术语为准。

最后，应理解，本文公开的申请的实施方案是对本说明书的实施方案的原理的说明。其他修改后的实施例也在本说明书的范围内。因此，本说明书披露的实施例仅仅作为示例而非限制。本领域技术人员可以根据本说明书中的实施例采取替代配置来实现本说明书中的申请。因此，本说明书的实施例不限于申请中被精确地描述过的实施例。

权　利　要　求　书

1. 一种文档表征模型的训练方法，包括：

获得文档图像，所述文档图像中包括多个文本单元，且被划分为多个图像块；

在所述文档图像中确定多个训练任务的预测对象，所述预测对象对应于所述文本单元或所述图像块；

从文本维度上生成所述文档图像对应的文本特征，并从视觉维度上生成所述文档图像对应的视觉特征，其中，所述视觉特征是通过在所述文档图像中对所述多个训练任务的预测对象对应的区域进行视觉标记，并对所述视觉标记后的图像进行视觉特征提取得到的，不同的训练任务的预测对象对应的视觉标记方式不同；以及

基于所述文本特征和所述视觉特征，对所述文档表征模型执行所述多个训练任务，以对所述文档表征模型进行训练。

2. 根据权利要求 1 所述的方法，其中，所述多个训练任务包括：图像掩蔽任务、文本掩蔽任务、以及文本-图像对齐任务，其中，

所述图像掩蔽任务的预测对象包括：所述多个图像块中的至少一个目标图像块，

所述文本掩蔽任务的预测对象包括：所述多个文本单元中的至少一个第一文本单元，

所述文本-图像对齐任务的预测对象包括：所述多个文本单元中的至少一个第二文本单元。

3. 根据权利要求 2 所述的方法，其中，所述视觉标记的过程包括：

将所述文档图像中与所述至少一个目标图像块对应的区域填充为第一颜色；

将所述文档图像中与所述至少一个第一文本单元对应的区域填充为第二颜色；以及

将所述文档图像中与所述至少一个第二文本单元对应的区域填充为第三颜色，其中，所述第一颜色、所述第二颜色和所述第三颜色互不相同。

4. 根据权利要求 1 所述的方法，其中，所述文本特征包括：

所述多个文本单元各自对应的文本嵌入表示，

所述多个文本单元各自对应的一维位置嵌入表示，

所述多个文本单元各自对应的二维位置嵌入表示，

所述多个文本单元各自对应的局部视觉嵌入表示，以及，

所述多个文本单元各自对应的类型嵌入表示，

其中，在所述文本特征中，与至少一个第一文本单元对应的所述文本嵌入表示被替换为掩蔽标识。

5. 根据权利要求 4 所述的方法，其中，在所述文本特征中，与至少一个目标文本单元对应的所述二维位置嵌入表示和所述局部视觉嵌入表示被替换为空标识，所述至少一个目标文本单元为所述至少一个目标图像块中的部分文本单元。

6. 根据权利要求 4 所述的方法，其中，每个文本单元对应的所述局部视觉嵌入表示是通过在所述文档图像中对所述文本单元对应的图像区域进行空洞卷积得到的。

7. 根据权利要求 4 所述的方法，其中，所述多个文本单元各自对应的一维位置嵌入表示是通过对所述多个文本单元的位置信息进行相对位置编码或者三角函数位置编码得到的。

8. 根据权利要求 1 所述的方法，其中，所述视觉特征包括：

所述多个图像块各自对应的视觉嵌入表示，

所述多个图像块各自对应的一维位置嵌入表示，

所述多个图像块各自对应的二维位置嵌入表示，以及

所述多个图像块各自对应的类型嵌入表示。

9. 根据权利要求 1 所述的方法，其中，所述基于所述文本特征和所述视觉特征，对所述文档表征模型执行所述多个训练任务，以对所述文档表征模型进行训练，包括：

通过所述文档表征模型基于所述文本特征和所述视觉特征，得到所述文档图像对应的文档预测表征；

基于所述文档预测表征执行所述多个训练任务，得到所述多个训练任务的预测对象各自对应的预测信息，并基于所述多个训练任务的预测对象各自对应的预测信息和实际信息，生成目标损失；以及

以最小化所述目标损失为训练目标，对所述文档表征模型的模型参数进行更新。

10. 根据权利要求 9 所述的方法，其中，所述文本特征包括：所述多个文本单元各自对应的多个维度的嵌入表示，所述视觉特征包括：所述多个图像块各自对应的多个维度的嵌入表示；以及

所述通过所述文档表征模型基于所述文本特征和所述视觉特征，得到所述文档图像对应的文档预测表征，包括：

针对所述多个文本单元中的每个文本单元，对所述文本单元对应的多个维度的嵌入表示相加，得到所述文本单元对应的第一嵌入表示，

针对所述多个图像块中的每个图像块，对所述图像块对应的多个维度的嵌入表示相加，得到所述图像块对应的第二嵌入表示，

对所述多个文本单元各自对应的第一嵌入表示、以及所述多个图像块各自对应的第二嵌入表示进行拼接，得到融合特征，以及

将所述融合特征输入至所述文档表征模型，获得所述文档表征模型输出的所述文档预测表征。

11. 一种文档表征模型的训练设备，包括：

至少一个存储介质，存储有至少一个指令集，用于进行文档表征模型的训练；以及

至少一个处理器，同所述至少一个存储介质通信连接，其中，所述至少一个处理器运行时读取所述至少一个指令集，并且根据所述至少一个指令集的指示：

获得文档图像，所述文档图像中包括多个文本单元，且被划分为多个图像块，

在所述文档图像中确定多个训练任务的预测对象，所述预测对象对应于所述文本单元或所述图像块，

从文本维度上生成所述文档图像对应的文本特征，并从视觉维度上生成所述文档图像对应的视觉特征，其中，所述视觉特征是通过在所述文档图像中对所述多个训练任务的预测对象对应的区域进行视觉标记，并对所述视觉标记后的图像进行视觉特征提取得到的，不同的训练任务的预测对象对应的视觉标记方式不同，以及

基于所述文本特征和所述视觉特征，对所述文档表征模型执行所述多个训练任务，以对所述文档表征模型进行训练。

12. 根据权利要求 11 所述的设备，其中，所述多个训练任务包括：图像掩蔽任务、文本掩蔽任务、以及文本-图像对齐任务，其中，

所述图像掩蔽任务的预测对象包括：所述多个图像块中的至少一个目标图像块，

所述文本掩蔽任务的预测对象包括：所述多个文本单元中的至少一个第一文本单元，

所述文本-图像对齐任务的预测对象包括：所述多个文本单元中的至少一个第二文本单元。

13. 根据权利要求 12 所述的设备，其中，所述视觉标记的过程包括：

将所述文档图像中与所述至少一个目标图像块对应的区域填充为第一颜色；

将所述文档图像中与所述至少一个第一文本单元对应的区域填充为第二颜色；以及

将所述文档图像中与所述至少一个第二文本单元对应的区域填充为第三颜色，其中，所述第一颜色、所述第二颜色和所述第三颜色互不相同。

14. 根据权利要求 11 所述的设备，其中，所述文本特征包括：

所述多个文本单元各自对应的文本嵌入表示，

所述多个文本单元各自对应的一维位置嵌入表示，

所述多个文本单元各自对应的二维位置嵌入表示，

所述多个文本单元各自对应的局部视觉嵌入表示，以及，

所述多个文本单元各自对应的类型嵌入表示，

其中，在所述文本特征中，与所述至少一个第一文本单元对应的所述文本嵌入表示被替换为掩蔽标识。

15. 根据权利要求 14 所述的设备，其中，在所述文本特征中，与至少一个目标文本单元对应的所述二维位置嵌入表示和所述局部视觉嵌入表示被替换为空标识，所述至少一个目标文本单元为所述至少一个目标图像块中的部分文本单元。

16. 根据权利要求 14 所述的设备，其中，每个文本单元对应的所述局部视觉嵌入表示是通过在所述文档图像中对所述文本单元对应的图像区域进行空洞卷积得到的。

17. 根据权利要求 14 所述的设备，其中，所述多个文本单元各自对应的一维位置嵌入表示是通过对所述多个文本单元的位置信息进行相对位置编码或者三角函数位置编码得到的。

18. 根据权利要求 11 所述的设备，其中，所述视觉特征包括：

所述多个图像块各自对应的视觉嵌入表示，

所述多个图像块各自对应的一维位置嵌入表示，

所述多个图像块各自对应的二维位置嵌入表示，以及

所述多个图像块各自对应的类型嵌入表示。

19. 根据权利要求 11 所述的设备，其中，为了对所述文档表征模型进行训练，所述至少一个处理器：

通过所述文档表征模型基于所述文本特征和所述视觉特征，得到所述文档图像对应的文档预测表征；

基于所述文档预测表征执行所述多个训练任务，得到所述多个训练任务的预测对象各自对应的预测信息，并基于所述多个训练任务的预测对象各自对应的预测信息和实际信息，生成目标损失；以及

以最小化所述目标损失为训练目标，对所述文档表征模型的模型参数进行更新。

20. 一种计算机可读非暂时性存储介质，其中，所述计算机可读非暂时性存储介质中存储有至少一个指令集，所述至少一个指令集被至少一个处理器执行时，实现如权利要求 1-10 中任一项所述的方法。