(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115393864 A
(43)申请公布日 2022.11.25

(21)申请号 202211034977.6
(22)申请日 2022.08.26
(71)申请人 阿里巴巴(中国)有限公司
地址 311121 浙江省杭州市余杭区五常街道文一西路969号3幢5层554室
(72)发明人 王鹏 达铖 姚聪
(74)专利代理机构 北京合智同创知识产权代理有限公司 11545
专利代理师 李杰 兰淑铎
(51)Int.Cl.
G06V 30/18(2022.01)
G06V 10/40(2022.01)
G06V 10/764(2022.01)
G06F 40/205(2020.01)
G06F 40/289(2020.01)
G06V 10/82(2022.01)
G06N 3/04(2006.01)

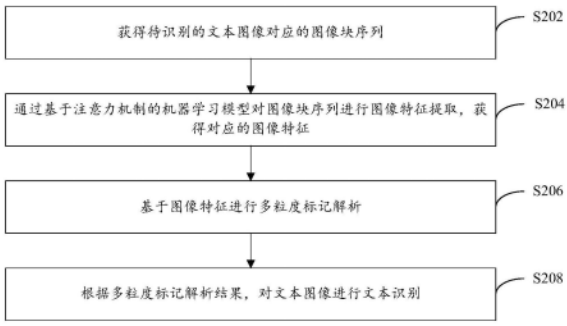
权利要求书2页 说明书9页 附图3页

(54)发明名称

文本识别方法、电子设备及计算机存储介质

(57)摘要

本申请实施例提供了一种文本识别方法、电子设备及计算机存储介质，其中，文本识别方法包括：获得待识别的文本图像对应的图像块序列；通过基于注意力机制的机器学习模型对所述图像块序列进行图像特征提取，获得对应的图像特征；基于所述图像特征进行多粒度标记解析，其中，所述多粒度标记解析包括：字符粒度的标记解析、子词粒度的标记解析、和整词粒度的标记解析；根据多粒度标记解析结果，对所述文本图像进行文本识别。通过本申请实施例，将文本识别的性能提升到了更高的水平。



CN 115393864 A

1.一种文本识别方法,包括:

获得待识别的文本图像对应的图像块序列;

通过基于注意力机制的机器学习模型对所述图像块序列进行图像特征提取,获得对应的图像特征;

基于所述图像特征进行多粒度标记解析,其中,所述多粒度标记解析包括:字符粒度的标记解析、子词粒度的标记解析、和整词粒度的标记解析;

根据多粒度标记解析结果,对所述文本图像进行文本识别。

2.根据权利要求1所述的方法,其中,所述子词粒度的标记解析为字节对编码粒度的标记解析。

3.根据权利要求1或2所述的方法,其中,基于所述图像特征进行子词粒度的标记解析,包括:

基于BPE算法对所述图像特征进行空间注意力计算和分类处理,将处理结果作为子词粒度的标记解析结果。

4.根据权利要求1或2所述的方法,其中,基于所述图像特征进行整词粒度的标记解析,包括:

基于WordPiece算法对所述图像特征进行空间注意力计算和分类处理,将处理结果作为整词粒度的标记解析结果。

5.根据权利要求1或2所述的方法,其中,所述根据多粒度标记解析结果,对所述文本图像进行文本识别,包括:

根据多粒度标记解析结果,获得对应的多个粒度的文本预测结果;

对所述多个粒度的文本预测结果进行融合,根据融合结果对所述文本图像进行文本识别。

6.根据权利要求5所述的方法,其中,所述对所述多个粒度的文本预测结果进行融合,根据融合结果对所述文本图像进行文本识别,包括:

分别对多个粒度中各个粒度的文本预测结果所指示的概率分布求均值,获得各个粒度的文本预测结果对应的概率分布均值;

将多个粒度对应的多个概率分布均值中的最大均值对应的文本预测结果,确定为目标文本预测结果;

根据所述目标文本预测结果,对所述文本图像进行文本识别。

7.根据权利要求5所述的方法,其中,所述对所述多个粒度的文本预测结果进行融合,根据融合结果对所述文本图像进行文本识别,包括:

分别对多个粒度中各个粒度的文本预测结果所指示的概率分布求乘积,获得各个粒度的文本预测结果对应的概率乘积结果;

将多个粒度对应的多个概率乘积结果中的最大乘积对应的文本预测结果,确定为目标文本预测结果;

根据所述目标文本预测结果,对所述文本图像进行文本识别。

8.根据权利要求5所述的方法,其中,所述对所述多个粒度的文本预测结果进行融合,根据融合结果对所述文本图像进行文本识别,包括:

分别对多个粒度中各个粒度的文本预测结果所指示的概率分布求均值,获得各个粒度

的文本预测结果对应的概率分布均值;将多个粒度对应的多个概率分布均值中的最大均值对应的文本预测结果,确定为第一文本预测结果;

并且,分别对多个粒度中各个粒度的文本预测结果所指示的概率分布求乘积,获得各个粒度的文本预测结果对应的概率乘积结果;将多个粒度对应的多个概率乘积结果中的最大乘积对应的文本预测结果,确定为第二文本预测结果;

根据第一文本预测结果对应的置信度,和第二文本预测结果对应的置信度,确定目标文本预测结果;

根据所述目标文本预测结果,对所述文本图像进行文本识别。

9.根据权利要求1或2所述的方法,其中,所述文本识别方法通过文本识别模型执行,所述基于注意力机制的机器学习模型为基于注意力机制的编码器;

所述文本识别模型包括:线性投影层、所述编码器、自适应寻址和聚合层、融合输出层;

其中:

所述线性投影层,用于将所述图像块序列投影为预设维度的向量;

所述编码器,用于对所述向量进行注意力计算,以从所述向量中提取并输出对应的图像特征;

所述自适应寻址和聚合层,用于基于所述图像特征进行多粒度标记解析,获得对应的多个粒度的标记解析结果;根据多个粒度的标记解析结果,获得对应的多个粒度的文本预测结果;

所述融合输出层,用于根据所述多个粒度的文本预测结果,确定并输出目标文本预测结果,以根据所述目标文本预测结果获得所述文本图像的文本识别结果。

10.根据权利要求9所述的方法,其中,所述自适应寻址和聚合层包括:字符粒度的自适应寻址和聚合层、子词粒度的自适应寻址和聚合层、整词粒度的自适应寻址和聚合层。

11.一种电子设备,包括:处理器、存储器、通信接口和通信总线,所述处理器、所述存储器和所述通信接口通过所述通信总线完成相互间的通信;

所述存储器用于存放至少一可执行指令,所述可执行指令使所述处理器执行如权利要求1-10中任一项所述的方法对应的操作。

12.一种计算机存储介质,其上存储有计算机程序,该程序被处理器执行时实现如权利要求1-10中任一所述的方法。

13.一种计算机程序产品,包括计算机指令,所述计算机指令指示计算设备执行如权利要求1-10中任一所述方法对应的操作。

文本识别方法、电子设备及计算机存储介质

技术领域

[0001] 本申请实施例涉及图像识别技术领域，尤其涉及一种文本识别方法、电子设备及计 算机存储介质。

背景技术

[0002] 文本识别是对图像中的文本进行检测和识别，以获得其对应的文本信息的一种技术。随着Transformer模型在自然语言处理领域中的广泛使用，越来越多的技术人员尝试将 其替代卷积神经网络模型，以实现将其应用至图像识别领域，以针对包含文本的图像进行文本识别。

[0003] 但是，由于包含文本的图像通常缺乏语义信息，因此，将Transformer模型应用于包含文本的图像进行文本识别时，通常会出现识别不准确，尤其是对于低质量图像，会 出现识别准确率较低的问题。

发明内容

[0004] 有鉴于此，本申请实施例提供一种文本识别方案，以至少部分解决上述问题。

[0005] 根据本申请实施例的第一方面，提供了一种文本识别方法，包括：获得待识别的文本图像对应的图像块序列；通过基于注意力机制的机器学习模型对所述图像块序列进行图像特征提取，获得对应的图像特征；基于所述图像特征进行多粒度标记解析，其中， 所述多粒度标记解析包括：字符粒度的标记解析、子词粒度的标记解析、和整词粒度的 标记解析；根据多粒度标记解析结果，对所述文本图像进行文本识别。

[0006] 根据本申请实施例的第二方面，提供了一种电子设备，包括：处理器、存储器、通信接口和通信总线，所述处理器、所述存储器和所述通信接口通过所述通信总线完成相 互间的通信；所述存储器用于存放至少一可执行指令，所述可执行指令使所述处理器执 行如第一方面所述方法对应的操作。

[0007] 根据本申请实施例的第三方面，提供了一种计算机存储介质，其上存储有计算机程 序，该程序被处理器执行时实现如第一方面所述的方法。

[0008] 根据本申请实施例的第四方面，提供了一种计算机程序产品，包括计算机指令，所述计算机指令指示计算设备执行如第一方面所述方法对应的操作。

[0009] 根据本申请实施例提供的方案，通过对图像特征进行多粒度标记解析，可将语义信 息隐式地注入对文本图像进行处理的模型中，使得模型可以同时结合图像特征和语义信 息进行文本识别，提高了识别效率和准确度。另一方面，该语义信息从字符粒度、子词粒度和整词粒度，从多粒度进行表征，从而能够从多粒度来获取图像特征和语义信息， 从而将文本识别的性能提升到了更高的水平。

附图说明

[0010] 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，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地介绍，显而易见地，下面描述中的附图仅仅是本申请实施例中记载的一些实施例，对于本领域普通技术人员来讲，还可以根据这些附图 获得其他的附图。

[0011] 图1为适用本申请实施例方案的示例性系统的示意图；

[0012] 图2A为根据本申请实施例的一种文本识别方法方法的步骤流程图；

[0013] 图2B为图2A所示实施例中的一种文本识别模型的结构示意图；

[0014] 图2C为使用图2B所示的文本识别模型获得的多粒度预测结果的示意图；

[0015] 图3为根据本申请实施例的一种电子设备的结构示意图。

具体实施方式

[0016] 为了使本领域的人员更好地理解本申请实施例中的技术方案，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，所描述的实施例仅是本申请实施例一部分实施例，而不是全部的实施例。基于本申请实施例中的 实施例，本领域普通技术人员所获得的所有其他实施例，都应当属于本申请实施例保护 的范围。

[0017] 下面结合本申请实施例附图进一步说明本申请实施例具体实现。

[0018] 图1示出了一种适用本申请实施例方案的示例性系统。如图1所示，该系统100可以包括云服务端102、通信网络104和/或一个或多个用户设备106，图1中示例为多个 用户设备。

[0019] 云服务端102可以是用于存储信息、数据、程序和/或任何其他合适类型的内容的任 何适当的设备，包括但不限于分布式存储系统设备、服务器集群、计算云服务端集群等。在一些实施例中，云服务端102可以执行任何适当的功能。例如，在一些实施例中，云 服务端102可以对文本图像进行文本识别。作为可选的示例，在一些实施例中，云服务 端102可以基于文本图像对应的图像特征，进行多粒度标记解析，以将文本语义信息融 入图像特征中，进行文本识别。作为可选的示例，在一些实施例中，云服务端102可以 设置文本识别模型，通过文本识别模型对文本图像进行文本识别。作为另一示例，在一 些实施例中，云服务端102可以接收用户设备106发送来的文本识别请求，基于该文本 识别请求对该请求所请求的文本图像进行文本识别。进一步地，在一些实施例中，云服 务端102还可以将文本识别结果发送回用户设备106。

[0020] 在一些实施例中，通信网络104可以是一个或多个有线和/或无线网络的任何适当的 组合。例如，通信网络104能够包括以下各项中的任何一种或多种：互联网、内联网、广域网(WAN)、局域网(LAN)、无线网络、数字订户线路(DSL)网络、帧中继网络、异步转 移模式(ATM)网络、虚拟专用网(VPN)和/或任何其它合适的通信网络。用户设备106能够 通过一个或多个通信链路(例如，通信链路112)连接到通信网络104，该通信网络104能 够经由一个或多个通信链路(例如，通信链路114)被链接到云服务端102。通信链路可以 是适合于在用户设备106和云服务端102之间传送数据的任何通信链路，诸如网络链路、 拨号链路、无线链路、硬连线链路、任何其它合适的通信链路或此类链路的任何合适的 组合。

[0021] 用户设备106可以包括适合于交互，如与用户或云服务端102交互的任何一个或多个用户设备。在一些实施例中，用户设备106可将云服务端102发送文本识别请求，以 使云

服务端102根据该请求获取该请求所请求的文本图像以对其进行文本识别。在一些 实施例中，用户设备106向云服务端102发送的文本识别请求中携带有相应的文本图像 或者该相应的文本图像的获取地址的信息，以使云服务端102能够获取该文本图像。在 一些实施例中，用户设备106可以包括任何合适类型的设备。例如，在一些实施例中， 用户设备106可以包括移动设备、平板计算机、膝上型计算机、台式计算机、可穿戴计 算机、游戏控制台、媒体播放器、车辆娱乐系统和/或任何其他合适类型的用户设备。

[0022] 基于上述系统，以下通过实施例对本申请的文本识别方案进行说明。

[0023] 参照图2A，示出了根据本申请实施例的一种文本识别方法方法的步骤流程图。

[0024] 本实施例的文本识别方法包括以下步骤：

[0025] 步骤S202：获得待识别的文本图像对应的图像块序列。

[0026] 本申请实施例中，文本图像包括纯文本图像以及图像中同时包含有文本图像元素和 非文本图像元素的图像，其均可适用于本申请实施例的方案。

[0027] 此外，本申请实施例中，为便于后续的基于注意力机制的机器学习模型进行图像特 征提取，先将文本图像分割为不重叠的一系列连续图像块，即图像块序列。这些图像块序列中的图像块，加上它们的位置信息，作为后续基于注意力机制的机器学习模型的输入。

[0028] 示例性地，可以将待识别的文本图像使用卷积进行分块，再将每一块进行展平处理 变成序列，然后将该图像块序列添加位置编码和CLS token(分类标记)，输入后续的基于注意力机制的机器学习模型，如Transformer结构的编码器中。

[0029] 步骤S204：通过基于注意力机制的机器学习模型对图像块序列进行图像特征提取， 获得对应的图像特征。

[0030] 与常规的使用卷积神经网络对图像进行特征提取不同，本申请实施例中因注意力机 制在特征提取方面的良好性能表现，采用了基于注意力机制的机器学习模型对图像块序 列进行特征提取。示例性，该机器学习模型可以采用Transformer中的编码器结构。

[0031] 为了在不改变Transformer编码器的结构的基础上，将其用于图像特征的提取，如前所述，在将图像处理为图像块序列之后，还为其添加位置编码和CLS token，以便于Transformer编码器能够直接进行处理。

[0032] 经过基于注意力机制的机器学习模型，如Transformer编码器，对图像块序列进行图像特征提取，可获得对应的图像特征，表征为相应的特征token(标记)的形式。

[0033] 步骤S206：基于图像特征进行多粒度标记解析。

[0034] 其中，多粒度标记解析包括：字符粒度的标记解析、子词粒度的标记解析、和整词粒度的标记解析。

[0035] 标记解析也称为tokenize，其可基于特征表示，生成一个个字符子串，每个字符子串具有相对完整的语义。本申请实施例中，通过标记解析，将文本语义信息引入对图像 的处理中，以使对文本图像的文本识别可同时基于图像特征和语义信息进行，以获得更 为准确的识别效果。

[0036] 进一步地，本申请实施例中，标记解析实现为三个粒度，分别为：字符粒度、子词粒度和整词粒度。

[0037] 其中，字符粒度即character-level，通过对图像特征进行字符粒度的标记解析，

可解析预测出最基本的字符，如，英文中的“a”、“b”、“c”，或中文中的“你”、“我”、“他”等。整词粒度即wordpiece-level，通过对图像特征进行整词粒度的标 记解析，可解析预测出自然语言单元，如英文中的“Transformer”、“coffee”等，或 中文的“这个餐馆做的菜很好”、“这件衣服真漂亮”等。而子词粒度即subword-level， 其为介于字符粒度和整词粒度之间的粒度，其以常用组合来进行分词考量，通过对图像 特征进行子词粒度的标记解析，可解析预测出介于字符和整词之间的子词，如英文中的 “Transformer”会被解析预测为“Transform”、“er”；“coffee”会被解析预测为“co” “ff”“ee”等；或者，中文的“这个餐馆做的菜很好”会被解析预测为“这个”、“餐 馆”、“做的菜”、“很好”；“这件衣服真漂亮”会被解析预测为“这件”、“衣服”、 “真”、“漂亮”等。

[0038] 在一种可行方式中，本申请实施例的子词粒度的标记解析可以实现为字节对编码BPE 粒度的标记解析。BPE粒度的标记解析会根据字符最常合并出现的频次进行子词分割，以使解析预测出的子词更为合理。实际上，wordpiece-level的标记解析也可以看作是 BPE的变种，是一种基于subword的更粗粒度的编码格式。

[0039] 上述多粒度标记解析可采用注意力处理+分类的方式实现。例如，基于图像特征进行 的字符粒度的标记解析可以实现为：根据单字符粒度，使用空间注意力函数将图像特征中与第i个字符相关的图像特征挑选出来，其中，i为文本图像中的文本字符的数量， 如“coffee”为6个字符；进而，对挑选出的这些图像特征进行聚合，生成与第i个字 符对应的向量；再通过分类器分类识别，获得第i个字符对应的字符文本，如“c”等。

[0040] 基于图像特征进行的子词粒度的标记解析可以实现为：基于BPE算法对图像特征进 行空间注意力计算和分类处理，将处理结果作为子词粒度的标记解析结果。BPE算法统计每一个连续字节对的出现频率，选择最高频者合并成新的最高频字节对。基于此，通 过训练后获得的空间注意力函数，能够根据子词粒度，将图像特征中与第j个子词相关 的图像特征挑选出来，并进行聚合，以生成与第j个子词对应的向量；再通过分类器进 行分类识别，获得第j个子词对应的字符文本。其中，j为文本图像中的文本子词的数 量，如“coffee”中包括2个子词，获得的第j个子词对应的字符文本可能为“co”， 或者“ffee”。

[0041] 基于图像特征进行了整词粒度的标记解析可以实现为：使用WordPiece算法对图像 特征进行空间注意力计算和分类处理，将处理结果作为整词粒度的标记解析结果。WordPiece算法可以看作是BPE算法的变种，不同点在于，WordPiece算法基于概率生成 新的subword而不是下一最高频字节对。基于此，通过训练后获得的空间注意力函数， 能够根据整词粒度，将图像特征中与第m个整词相关的图像特征挑选出来，并进行聚合， 以生成与第m个整词对应的向量；再通过分类器进行分类识别，获得第m个整词对应的 字符文本。其中，m为文本图像中的文本整词的数量，如“I like coffee”中包括3个 整词，获得的第m个子词对应的字符文本可能为“I”，或者“like”，或者“coffee”。

[0042] 通过上述多粒度标记解析，一方面，能够通过多个级别的解析和预测，获取更多更丰富的图像特征和语义信息，提升针对文本图像的文本识别性能；另一方面，使用空间 注意力机制和分类处理，能够进行不同级别的准确解析和预测，提升解析和预测准确度 和效率。

[0043] 步骤S208：根据多粒度标记解析结果，对文本图像进行文本识别。

[0044] 在获取了多粒度标记解析结果后，即可基于这些结果进行文本识别。包括：根据多

粒度标记解析结果，获得对应的多个粒度的文本预测结果；对多个粒度的文本预测结果 进行融合，根据融合结果对文本图像进行文本识别。需要说明的是，文本预测结果的获 得，除了基于标记解析结果外，还需结合预设词表。不同粒度的标记解析结果对应有不 同的词表。例如，字符粒度的词表中除包括有a、b、c、d等字符外，还包含有其它符号， 如#、@、*等。本申请实施例中，对不同粒度的词表中的字符、或子词、或整词等的词表 大小和具体实现均不作限制。但为了便于说明，下文中均以词表中包括256个元素为示 例进行说明。

[0045] 其中，对多个粒度的文本预测结果进行融合，根据融合结果对文本图像进行文本识 别可以实现为：

[0046] 方式一：分别对多个粒度中各个粒度的文本预测结果所指示的概率分布求均值，获 得各个粒度的文本预测结果对应的概率分布均值；将多个粒度对应的多个概率分布均值 中的最大均值对应的文本预测结果，确定为目标文本预测结果；根据目标文本预测结果， 对文本图像进行文本识别。

[0047] 仍以文本图像中的文本为“coffee”为示例，则基于词表中的256个字符，“coffee”中的每个字符均对应有相应的概率，结合形成其对应的概率分布，如【0,0,0.9,0， 0.05,0,0,……,0.05,0,0……】，该概率分布指示“c”为字符c的概率为0.9，为字 符e的概率为0.02，为字符o的概率为0.05，为其它字符的概率均为0。类似地，其它 每个字符也都对应有类似的概率分布。则，在字符粒度下，对“coffee”对应的所有字 符的概率分布求均值，可以获得该字符粒度下的概率分布均值，记为P1。

[0048] 类似地，在子词粒度下，“coffee”基于子词粒度的词表，也具有相应的多个子词的概率分布。则，在子词粒度下，对“coffee”对应的所有子词的概率分布求均值，可 以获得该子词粒度下的概率分布均值，记为P2。

[0049] 在整词粒度下，“coffee”基于整词粒度的词表，也具有相应的概率分布。则，在 整词粒度下，对“coffee”对应的概率分布求均值，可以获得该整词粒度下的概率分布 均值，记为P3。

[0050] 假设，上例中，P3>P2>P1，则P3对应的文本预测结果即被确定为目标文本预测结果， 基于该结果进行整词识别，识别出文本图像中的“coffee”图像部分对应的字符为coffee。

[0051] 通过概率分布均值的方式，可以较为均衡的表征各个粒度的文本预测结果的预测情 况，可以获得较为客观和准确的预测结果。

[0052] 方式二：分别对多个粒度中各个粒度的文本预测结果所指示的概率分布求乘积，获 得各个粒度的文本预测结果对应的概率乘积结果；将多个粒度对应的多个概率乘积结果 中的最大乘积对应的文本预测结果，确定为目标文本预测结果；根据目标文本预测结果， 对文本图像进行文本识别。

[0053] 仍以文本图像中的文本为“coffee”为示例，则基于词表中的256个字符，“coffee”中的每个字符均对应有相应的概率，结合形成其对应的概率分布，如【0,0,0.9,0， 0.05,0,0,……,0.05,0,0……】，该概率分布指示“c”为字符c的概率为0.9，为字 符e的概率为0.02，为字符o的概率为0.05，为其它字符的概率均为0。类似地，其它 每个字符也都对应有类似的概率分布。则，在字符粒度下，对“coffee”对应的所有字 符的概率分布求乘积，可以获得该字符粒度下的概率分布乘积，记为M1。

[0054] 类似地，在子词粒度下，“coffee”基于子词粒度的词表，也具有相应的多个子词的概率分布。则，在子词粒度下，对“coffee”对应的所有子词的概率分布求乘积，可 以获得该子词粒度下的概率分布乘积，记为M2。

[0055] 在整词粒度下，“coffee”基于整词粒度的词表，也具有相应的概率分布。则，在 整词粒度下，对“coffee”对应的概率分布求乘积，可以获得该整词粒度下的概率分布 乘积，记为M3。

[0056] 假设，上例中，M1>M2>M3，则M1对应的文本预测结果即被确定为目标文本预测结果， 基于该结果进行字符识别，识别出文本图像中的“coffee”图像部分对应的字符分别为c、o、f、f、e、e，基于此再组合起来即为coffee。

[0057] 通过概率分布乘积的方式，使得预测的较为准确的那个粒度的文本预测结果更为突 显，可以快速、高效地获得较为准确的预测结果。

[0058] 方式三：分别对多个粒度中各个粒度的文本预测结果所指示的概率分布求均值，获 得各个粒度的文本预测结果对应的概率分布均值；将多个粒度对应的多个概率分布均值 中的最大均值对应的文本预测结果，确定为第一文本预测结果；并且，分别对多个粒度中各个粒度的文本预测结果所指示的概率分布求乘积，获得各个粒度的文本预测结果对应的概率乘积结果；将多个粒度对应的多个概率乘积结果中的最大乘积对应的文本预测结果，确定为第二文本预测结果；根据第一文本预测结果对应的置信度，和第二文本预 测结果对应的置信度，确定目标文本预测结果；根据目标文本预测结果，对文本图像进 行文本识别。

[0059] 本方式中，基于前述方式一和方式二获得的结果，综合考虑两种方式预测的情况，从中选中更优的那种方式对应的预测结果。其中，方式一对应的文本预测结果即本方式 中的第一文本预测结果，方式二对应的文本预测结果即本方式中的第二文本预测结果。在此基础上，再确定两者的置信度，根据两者分别对应的置信度，选出其中置信度较高 的那种方式的文本预测结果作为目标文本预测结果，再基于此对文本图像进行文本识别。其中，第一文本预测结果和第二文本预测结果对应的置信度的具体获取方式可采用常规 方式获得，在此不再详述。

[0060] 通过这种方式，能够选出较优的目标文本预测结果，以为后续文本识别提供准确的 依据。

[0061] 通过本实施例，对图像特征进行多粒度标记解析，可将语义信息隐式地注入对文本 图像进行处理的模型中，使得模型可以同时结合图像特征和语义信息进行文本识别，提高了识别效率和准确度。另一方面，该语义信息从字符粒度、子词粒度和整词粒度，从 多粒度进行表征，从而能够从多粒度来获取图像特征和语义信息，从而将文本识别的性 能提升到了更高的水平。

[0062] 在实际应用中，上述文本识别方法也可通过文本识别模型实现。在一种可行的文本 识别模型的实现方式中，该文本识别模型可以包括：线性投影层、基于注意力机制的编码器、自适应寻址和聚合层、融合输出层。

[0063] 其中：

[0064] 线性投影层，用于将文本图像的图像块序列投影为预设维度的向量。

[0065] 基于注意力机制的编码器，如Transformer编码器，用于实现前述基于注意力机制

的机器学习模型的功能,即对上述预设维度的向量进行注意力计算,以从该预设维度的 向量中提取并输出对应的图像特征。

[0066] 自适应寻址和聚合层,用于基于图像特征进行多粒度标记解析,获得对应的多个粒 度的标记解析结果;根据多个粒度的标记解析结果,获得对应的多个粒度的文本预测结果。

[0067] 融合输出层,用于根据多个粒度的文本预测结果,确定并输出目标文本预测结果,以根据目标文本预测结果获得文本图像的文本识别结果。

[0068] 上述自适应寻址和聚合层可以包括:字符粒度的自适应寻址和聚合层、子词粒度的 自适应寻址和聚合层、整词粒度的自适应寻址和聚合层。

[0069] 一种示例性的上述文本识别模型的具体实现实例如图2B所示,由图中可见,一幅W*H的RGB图像通过patch操作被分割成一系列不重叠的图像块,这些图像块形成图像 块序列,图中示意为P*P Patches。其中,P*P表示每个图像块的分辨率。

[0070] 图像块序列通过线性投影层(Linear Projection),被线性投射为D维的图像块向量。其中,D由本领域技术人员根据实际需求设置。

[0071] 在获得了D维的图像块向量后,文本识别模型还会在该向量头部添加可学习的[CLS] token,并且还会为该[CLS]token和每个图像块对应的图像块向量添加该图像块对应的 位置信息,从而形成添加了[CLS]token的Position+Patch embedding,以输入基于注意力机制的编码器。图2B中,该向量如线性投影层和Transformer编码器之间的向量所 示。该部分图示向量中,0、1、……10表示位置向量,“*”表示[CLS]token,与1、2、……10 每个位置向量合在一起的空心椭圆形表示图像块向量。

[0072] 上述处理后的向量会被输出基于注意力机制的编码器中,图2B中示意为Transformer编码器(Transformer Encoder),该编码器基于该向量进行图像特征提取, 获得对应的图像特征。

[0073] 接着,图像特征会被输入自适应寻址和聚合层(Adaptive Addressing and Aggregation),为便于说明,后面简称为A3层。在该A3层进行多个粒度的标记解析。因需进行多粒度标记解析,该A3层通过相互独立的三个模块实现,分别为:字符粒度的 A3模块(Character A3 Module)、BPE粒度的A3模块(BPE A3 Module)、整词粒度的 A3模块(WordPiece A3 Module)。

[0074] 传统方式中,Transformer编码器在进行文本图像的文本识别时,直接取256个输出序列的前27个token用作最终的输出,而其它的token没有得到全部有效的利用,很 多有用信息被丢弃。为了充分利用Transformer编码器的输出序列的信息以进行文本序 列预测,本申请实施例的方案中,通过多个粒度的A3模块,将Transformer编码器的所 有输出token整合到一个预设长度的序列中(示例性地,该预设长度可以为27,即假设 最长字符的长度就是27)。假设,各个粒度的A3模块经注意力处理后输出的元素为y_i, i表示第i个元素,Transformer编码器的输出为z,聚合函数为A,则A3模块使用的转 换公式为:$y_i = A_i(z_L)$。

[0075] 在一种可行方式中,$y_i = A_i(z_L) = \mathrm{softmax}(\alpha_i(z_L))^T (z_L U)^T$

[0076] 其中,$\alpha_i(\cdot)$表示1*1卷积核的群卷积(也称分组卷积);z_L表示Transformer编 码器输出的token序列;U表示一个可学习的线性映射矩阵。

[0077] 基于此，对于某个A3模块，经上述计算后输出的向量Y表示为：

[0078] $Y=[y_1y_2;...;y_T]=[A_1(z_L);A_2(z_L);...;A_T(z_L)]$

[0079] 其中，T为预设文本长度，如上所示，本实施例为27。

[0080] 接着，通过各粒度的A3模块对应的分类器，基于向量Y进行分类，以实现文本序列预测。

[0081] 在一个示例中，分类器表示为$G=YW^T$，其中，W表示线性映射矩阵。但如前 所述，不同粒度的A3模块对应有不同的分类器，本实施例中采用G统一表示，不再区 分。但本领域技术人员应当明了，不同粒度对应的分类器中的Y和W不同，因此，分 类器G也不相同。

[0082] 经过A3模块的处理后，可获得多个粒度的文本预测结果，一个多个粒度的预测结果 示例如图2C所示。由图中可见，对于字符粒度，其可将单词拆分为最细粒度的单个字符。对于BPE粒度，其将单词拆分为常见的字符子串，如，对于“methodist”，其被拆分为“method”和“ist”；对于“university”，其被拆分为“un”和“iversity”；对于 “41km”，其被拆分为“41”和“km”。而对于整词粒度，则是预测一整个单词。

[0083] 而具体到图2B所示示例中，“coffee”的字符粒度预测结果为单个字符，BPE粒度预测结果为“co”和“ffee”，整词粒度预测结果则为“coffee”。

[0084] 进而，A3模块输出的多粒度预测结果将通过融合输出层进行融合输出，其可采用前 述描述的融合输出部分的方式一、方式二和方式三中任意一种方式进行融合输出，在此不再赘述。

[0085] 由上可见，图2B所示的文本识别模型示例中，对文本图像的图像处理和语义处理完 全共用backbone(线性投影层和基于注意力机制的编码器)，没有单独的语义处理模块；并且，通过A3模块将语义信息隐式地注入到了文本识别模型中；而且，A3模块能够通 过空间注意力机制自动聚合和选择Transformer编码器输出的所有token；此外，A3模 块包括多个粒度的模块，分别从字符粒度、子词粒度和整词粒度进行预测，从而能够从 更多粒度来获取图像特征和语义信息，以实现更为准确、高效的文本识别。

[0086] 参照图3，示出了根据本申请实施例的一种电子设备的结构示意图，本申请具体实施例并不对电子设备的具体实现做限定。

[0087] 如图3所示，该电子设备可以包括：处理器(processor)302、通信接口(Communications Interface)304、存储器(memory)306、以及通信总线308。

[0088] 其中：

[0089] 处理器302、通信接口304、以及存储器306通过通信总线308完成相互间的通信。

[0090] 通信接口304，用于与其它电子设备或服务器进行通信。

[0091] 处理器302，用于执行程序310，具体可以执行上述文本识别方法实施例中的相关步 骤。

[0092] 具体地，程序310可以包括程序代码，该程序代码包括计算机操作指令。

[0093] 处理器302可能是CPU，或者是特定集成电路ASIC(Application Specific Integrated Circuit)，或者是被配置成实施本申请实施例的一个或多个集成电路。智 能设备包括的一个或多个处理器，可以是同一类型的处理器，如一个或多个CPU；也可 以是不同类型的处理器，如一个或多个CPU以及一个或多个ASIC。

[0094] 存储器306，用于存放程序310。存储器306可能包含高速RAM存储器，也可能还包

括非易失性存储器(non-volatile memory),例如至少一个磁盘存储器。

[0095] 程序310具体可以用于使得处理器302执行前述多个方法实施例中任一实施例所描 述的文本识别方法对应的操作。

[0096] 程序310中各步骤的具体实现可以参见上述方法实施例中的相应步骤和单元中对应 的描述,并具有相应的有益效果,在此不赘述。所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的设备和模块的具体工作过程,可以参考前述方法实施 例中的对应过程描述,在此不再赘述。

[0097] 本申请实施例还提供了一种计算机程序产品,包括计算机指令,该计算机指令指示 计算设备执行上述多个方法实施例中的任一文本识别方法对应的操作。

[0098] 需要指出,根据实施的需要,可将本申请实施例中描述的各个部件/步骤拆分为更多 部件/步骤,也可将两个或多个部件/步骤或者部件/步骤的部分操作组合成新的部件/步 骤,以实现本申请实施例的目的。

[0099] 上述根据本申请实施例的方法可在硬件、固件中实现,或者被实现为可存储在记录 介质(诸如CD ROM、RAM、软盘、硬盘或磁光盘)中的软件或计算机代码,或者被实现 通过网络下载的原始存储在远程记录介质或非暂时机器可读介质中并将被存储在本地记 录介质中的计算机代码,从而在此描述的方法可被存储在使用通用计算机、专用处理器 或者可编程或专用硬件(诸如ASIC或FPGA)的记录介质上的这样的软件处理。可以理 解,计算机、处理器、微处理器控制器或可编程硬件包括可存储或接收软件或计算机代 码的存储组件(例如,RAM、ROM、闪存等),当所述软件或计算机代码被计算机、处理 器或硬件访问且执行时,实现在此描述的方法。此外,当通用计算机访问用于实现在此 示出的方法的代码时,代码的执行将通用计算机转换为用于执行在此示出的方法的专用 计算机。

[0100] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元 及方法步骤,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术 人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认 为超出本申请实施例的范围。

[0101] 以上实施方式仅用于说明本申请实施例,而并非对本申请实施例的限制,有关技术 领域的普通技术人员,在不脱离本申请实施例的精神和范围的情况下,还可以做出各种变化和变型,因此所有等同的技术方案也属于本申请实施例的范畴,本申请实施例的专 利保护范围应由权利要求限定。

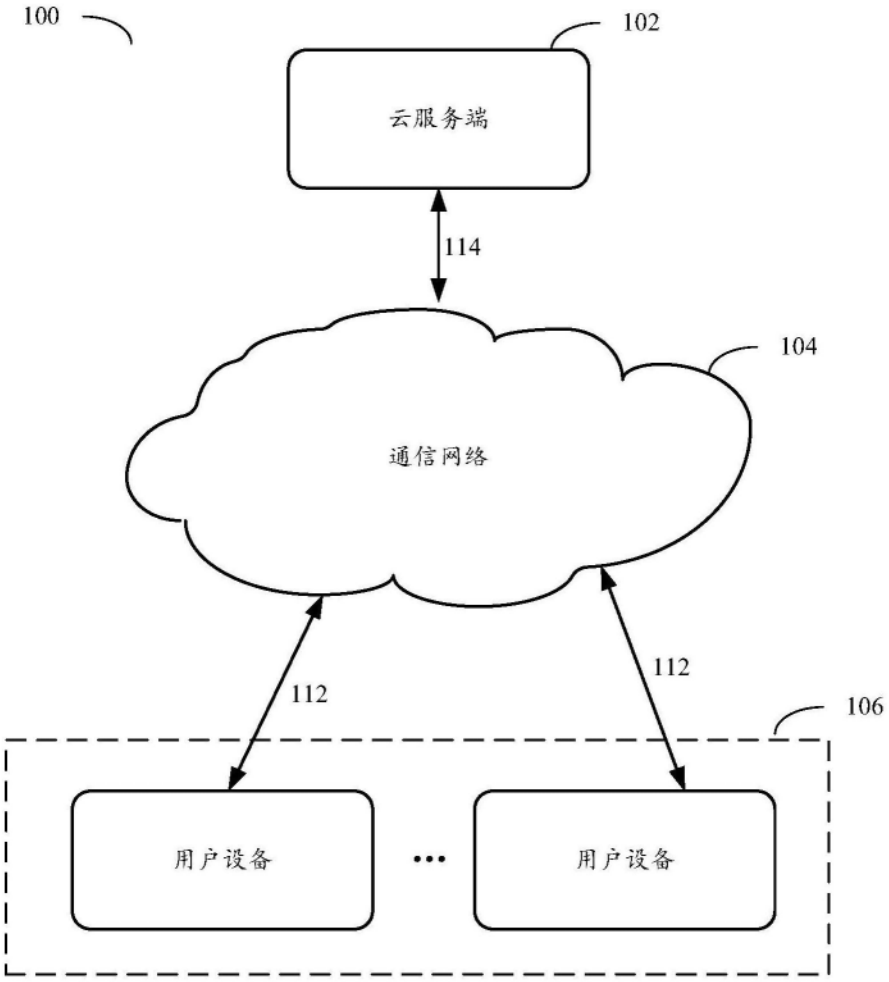

100
102
云服务端
114
104
通信网络
112
112
106
用户设备
用户设备

图1

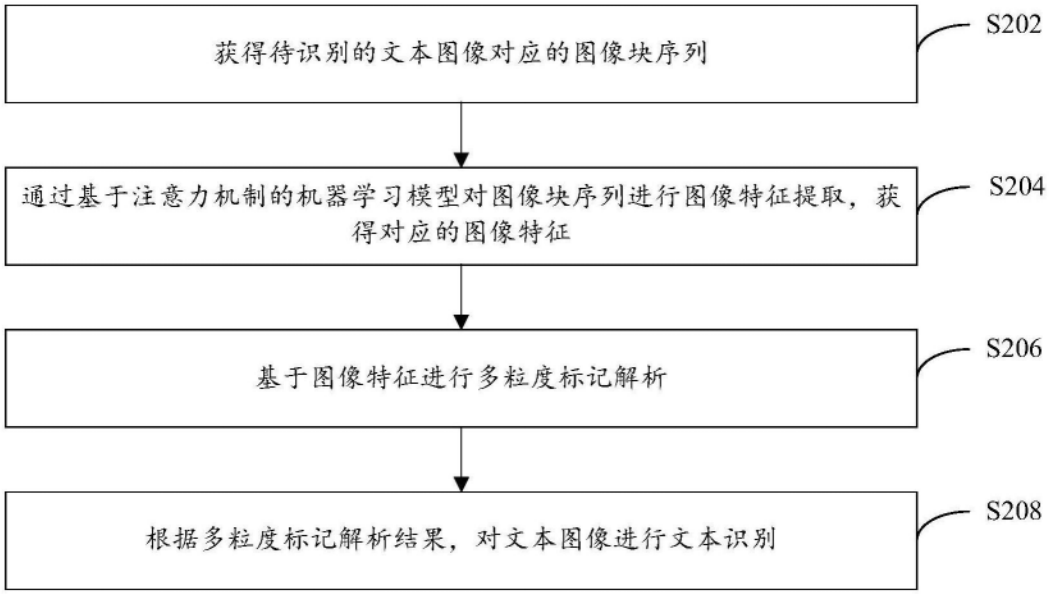

获得待识别的文本图像对应的图像块序列
S202
通过基于注意力机制的机器学习模型对图像块序列进行图像特征提取，获得对应的图像特征
S204
基于图像特征进行多粒度标记解析
S206
根据多粒度标记解析结果，对文本图像进行文本识别
S208

图2A

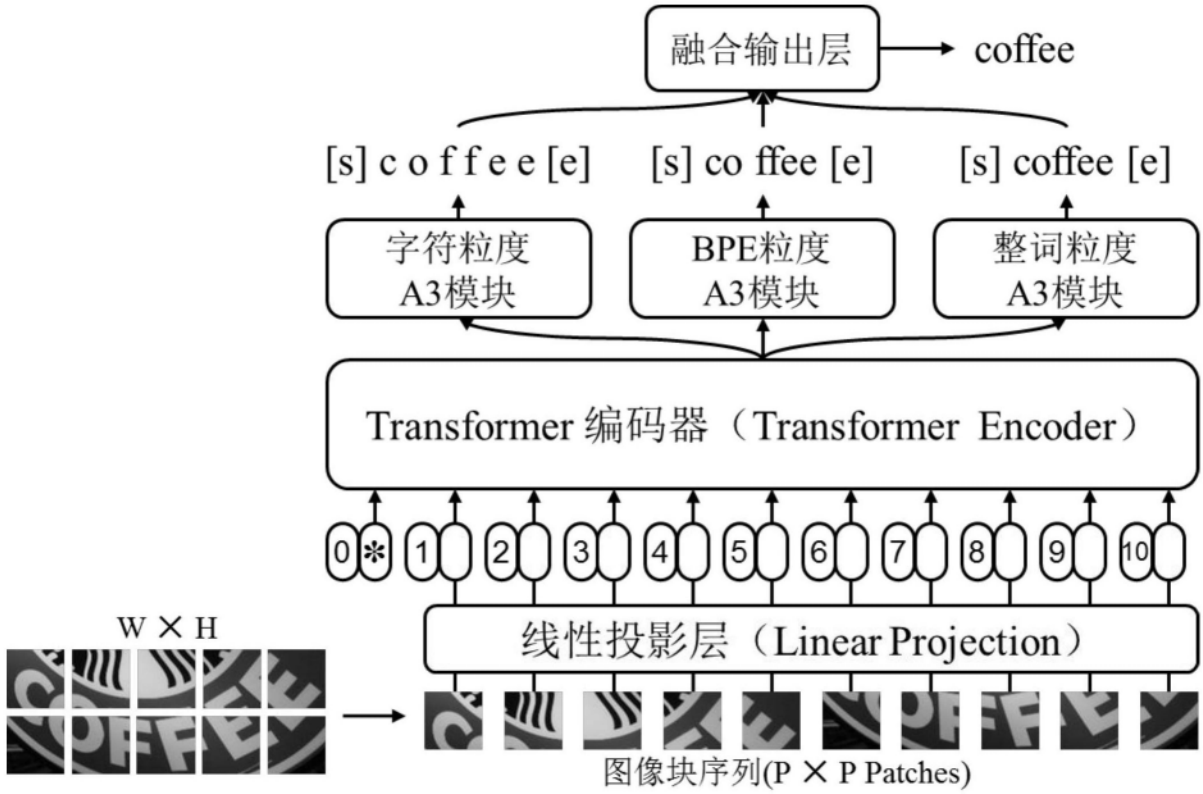


图2B

文本图像	字符粒度	BPE粒度	整词粒度
	n e a r	near	near
	m e t h o d i s t	method ist	methodist
	u n i v e r s i t y	un iversity	university
	4 1 k m	41 km	41 km

图2C

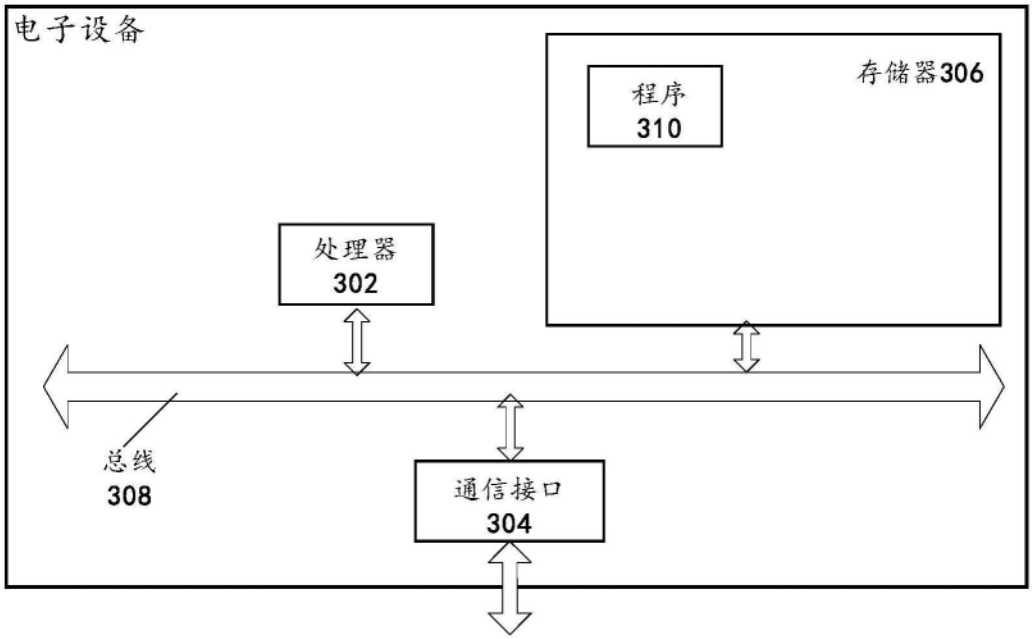


图3