

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 7 月 7 日 (07.07.2022)



(10) 国际公布号
WO 2022/142014 A1

(51) 国际专利分类号:
G06F 16/35 (2019.01)

(21) 国际申请号: PCT/CN2021/090497

(22) 国际申请日: 2021 年 4 月 28 日 (28.04.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202011594264.6 2020 年 12 月 29 日 (29.12.2020) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 陈昊(CHEN, Hao); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市世联合知识产权代理有限公司(SL INTELLECTUAL PROPERTY CO., LTD.);
中国广东省深圳市福田区沙头街道天安社区泰然四路 29 号天安创新科技广场一期 A 座 805E, Guangdong 518000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) Title: MULTI-MODAL INFORMATION FUSION-BASED TEXT CLASSIFICATION METHOD, AND RELATED DEVICE THEREOF

(54) 发明名称: 基于多模态信息融合的文本分类方法、及其相关设备

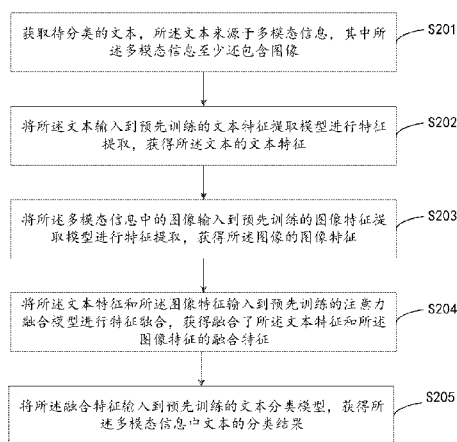


图 2

(57) Abstract: A multi-modal information fusion-based text classification method and apparatus, and a computer device and a storage medium, which belong to the field of artificial intelligence. The method comprises: acquiring text to be classified; inputting the text into a pre-trained text feature extraction model to perform feature extraction, so as to obtain a text feature of the text; inputting an image in multi-modal information into a pre-trained image feature extraction model to perform feature extraction, so as to obtain an image feature of the image; inputting the text feature and the image feature into a pre-trained attention fusion model to perform feature fusion, so as to obtain a fusion feature in which the text feature and the image feature are fused; and inputting the fusion feature into a pre-trained text classification model, so as to obtain a classification result of the text in the multi-modal information. By means of fusing a text feature and an image feature, and performing text classification on the basis of a fused feature, image information is used for text classification, such that a classification result is more accurate.



ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种基于多模态信息融合的文本分类方法、装置、计算机设备及存储介质, 属于人工智能领域, 包括通过获取待分类的文本; 将所述文本输入到预先训练的文本特征提取模型进行特征提取, 获得所述文本的文本特征; 将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取, 获得所述图像的图像特征; 将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合, 获得融合了所述文本特征和所述图像特征的融合特征; 将所述融合特征输入到预先训练的文本分类模型, 获得所述多模态信息中文本的分类结果。通过文本特征和图像特征融合, 基于融合后的特征进行文本分类, 文本分类利用了图像信息, 分类结果更准确。

基于多模态信息融合的文本分类方法、及其相关设备

本申请要求于2020年12月29日提交中国专利局、申请号为202011594264.6，发明名称为“基于多模态信息融合的文本分类方法、及其相关设备”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

本申请涉及人工智能技术领域，尤其涉及基于多模态信息融合的文本分类方法、及其相关设备。

背景技术

在机器学习领域，尤其在自然语言处理领域，对文本信息正确的分类并标注正确的标签对信息的分类存储、搜索以及用户的理解至关重要。尤其，当文本来源与语音转换时，由于技术的限制，转换结果可能具有歧义，对文本信息正确分类，能帮助用户正确理解转换的内容。

发明人发现当前绝大多数方法都只是涉及单一模态的输入，机器视觉仅仅是基于图像、视频；自然语言仅仅是语言本身。而当下场景中，视频等动态场景非常丰富，是各个业务场景中占比相当大的信息输入，单一采用其中一种信息来源，比如仅仅采用视频中的语音信息，是对信息的一种浪费。且采用单一信息源进行信息分类，缺少其他同源信息的互相验证，准确度低。

发明内容

本申请实施例的目的在于提出一种基于多模态信息融合的文本分类方法、装置、计算机设备及存储介质，以解决采用单一信息源进行信息分类，缺少其他同源信息的互相验证，准确度低的问题。

为了解决上述技术问题，本申请实施例提供一种基于多模态信息融合的文本分类方法，采用了如下所述的技术方案：

获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像；

将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；

将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；

将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；

将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。

为了解决上述技术问题，本申请实施例还提供一种基于多模态信息融合的文本分类装置，采用了如下所述的技术方案：

获取模块，用于获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像；

第一提取模块，用于将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；

第二提取模块，用于将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；

融合模块，用于将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；

分类模块，用于将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。

为了解决上述技术问题，本申请实施例还提供一种计算机设备，采用了如下所述的技术方案：

一种计算机设备，包括存储器和处理器，存储器中存储有计算机可读指令，所述处理器执行所述计算机可读指令时实现如下步骤：

获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像；

将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；

将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；

将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；

将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。

为了解决上述技术问题，本申请实施例还提供一种计算机可读存储介质，采用了如下所述的技术方案：

一种计算机可读存储介质，计算机可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时，使得所述处理器执行如下步骤：

获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像；

将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；

将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；

将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；

将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。

与现有技术相比，本申请实施例主要有以下有益效果：

通过获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像；将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。通过文本特征和图像特征融合，基于融合后的特征进行文本分类，文本分类利用了图像信息，分类结果更准确。

附图说明

为了更清楚地说明本申请中的方案，下面将对本申请实施例描述中所需要使用的附图作一个简单介绍，显而易见地，下面描述中的附图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他的附图。

图 1 是本申请可以应用于其中的示例性系统架构图；
图 2 是根据本申请的基于多模态信息融合的文本分类方法的一个实施例的流程图；
图 3 是本申请的注意力融合模型的结构示意图；
图 4 是本申请的注意力融合模型中门控激活层的结构示意图；
图 5 是本申请的注意力融合模型中注意层的结构示意图；
图 6 是根据本申请的基于多模态信息融合的文本分类装置的一个实施例的结构示意图；
图 7 是根据本申请的计算机设备的一个实施例的结构示意图。

具体实施方式

除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同；本文中在申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在限制本申请；本申请的说明书和权利要求书及上述附图说明中的术语“包括”和“具有”以及它们的任何变形，意图在于覆盖不排他的包含。本申请的说明书和权利要求书或上述附图中的术语“第一”、“第二”等是用于区别不同对象，而不是用于描述特定顺序。

在本文中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

为了使本技术领域的人员更好地理解本申请方案，下面将结合附图，对本申请实施例中的技术方案进行清楚、完整地描述。

如图 1 所示，系统架构 100 可以包括终端设备 101、102、103，网络 104 和服务器 105。网络 104 用以在终端设备 101、102、103 和服务器 105 之间提供通信链路的介质。网络 104 可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

用户可以使用终端设备 101、102、103 通过网络 104 与服务器 105 交互，以接收或发送消息等。终端设备 101、102、103 上可以安装有各种通讯客户端应用，例如网页浏览器应用、购物类应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等。

终端设备 101、102、103 可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、MP3 播放器 (Moving Picture Experts Group Audio Layer III, 动态影像专家压缩标准音频层面 3)、MP4 (Moving Picture Experts Group Audio Layer IV, 动态影像专家压缩标准音频层面 4) 播放器、膝上型便携计算机和台式计算机等等。

服务器 105 可以是提供各种服务的服务器，例如对终端设备 101、102、103 上显示的页面提供支持的后台服务器。

需要说明的是，本申请实施例所提供的基于多模态信息融合的文本分类方法一般由服务器/终端设备执行，相应地，基于多模态信息融合的文本分类装置一般设置于服务器/终端设备中。

应该理解，图 1 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

继续参考图 2，示出了根据本申请的基于多模态信息融合的文本分类的方法的一个实施例的流程图。所述的基于多模态信息融合的文本分类方法，包括以下步骤：

步骤 S201，获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像。

在本实施例中，基于多模态信息融合的文本分类方法运行于其上的电子设备 (例如图 1 所示的服务器/终端设备) 可以通过有线连接方式或者无线连接方式获取待分类的文本。需

要指出的是,上述无线连接方式可以包括但不限于 3G/4G 连接、WiFi 连接、蓝牙连接、WiMAX 连接、Zigbee 连接、UWB(ultra wideband)连接、以及其他现在已知或将来开发的无线连接方式。

多模态这里指表征信息的方式为多种,多模态信息指包含图像、文字、声音的信息。这里待分类的文本可以直接提取自多模态信息,例如当多模态信息为包含文字的视频时,待分类的文本来源于对视频中文字的提取,当多模态信息为包含包含音频和视频的影音文件时,待分类的文本来源于对音频的文字转换结果。对多模态信息中的文本进行正确的分类并标注正确的标签,有助于用户正确的理解文本信息。

步骤 S202,将所述文本输入到预先训练的文本特征提取模型进行特征提取,获得所述文本的文本特征。

在本实施例中,预先训练的文本特征提取模型基于 DPCNN(Deep Pyramid Convolutional Neural Networks)结构,该结构被广泛认为可以有效提取文本中语义信息。我们这里采用 DPCNN 中 pooling 前的卷积层与残差层所组成的结构构成本申请的文本特征提取模型。

步骤 S203,将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取,获得所述图像的图像特征。

在本实施例中,图像特征提取模型由 Resnet(Deep residual network)深度残差网络的前 5 层结构构成,用于图像特征的提取。

步骤 S204,将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合,获得融合了所述文本特征和所述图像特征的融合特征。

本实施例中,注意力融合模型,主要需要实现两个特征的融合,包含维度的填充和变换。例如如果文本信息来源于“浇树”,如果图像中确实有树木的区域,那么应将这里图像中树木区域的信息以某种方式补充到文本信息张量中,这里的注意力融合模型发现表示树木的区域,并能有效将其融合到输出的信融合特征中。这里的注意力融合模型,结构如图 3 所示,由门控激活层,注意层以及融合层构成,其中门控激活层采用的是经典的门控设计,如图 4 所示,在 g_i^t 所示位置为 softmax 作为激活操作, h_i 为隐含状态, r_i 为可学习的门控参数, $\tilde{h}_i$ 为隐含状态经过门控操作后的估计状态。注意层如图 5 所示:使用多个卷积层之间的乘\加法运算与 softmax 运算所构成。融合层用 resnet 的残差结构块所构成。

步骤 S205,将所述融合特征输入到预先训练的文本分类模型,获得所述多模态信息中文本分类结果。

本申请通过通过获取待分类的文本,所述文本来源于多模态信息,其中所述多模态信息至少还包含图像;将所述文本输入到预先训练的文本特征提取模型进行特征提取,获得所述文本的文本特征;将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取,获得所述图像的图像特征;将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合,获得融合了所述文本特征和所述图像特征的融合特征;将所述融合特征输入到预先训练的文本分类模型,获得所述多模态信息中文本分类结果。通过文本特征和图像特征融合,基于融合后的特征进行文本分类,文本分类利用了图像信息,分类结果更准确。

在本实施例的一些可选的实现方式中,在步骤 S202 之前,上述电子设备还可以执行以下步骤:

步骤 S301,获取多模态信息样本,所述多模态信息样本至少包含文本样本和图像样本;

步骤 S302,将所述多模态信息样本中的文本样本输入到预设的文本特征提取模型,获得所述文本样本的文本样本特征;

步骤 S303,将所述多模态信息样本中的图像样本输入到预设的图像特征提取模型,获得所述图像样本的图像样本特征;

步骤 S304,将所述文本样本特征和图像样本特征输入到预设的注意力融合模型,获得

所述多模态信息样本的融合样本特征；

步骤 S305，将所述融合样本特征输入到预设的图像还原模型进行图像还原，获得所述图像样本的还原图像；

步骤 S306，通过第一损失函数比较所述还原图像和所述图像样本之间的一致性，所述第一损失函数为：

$$L=L1+L2$$

其中 $L1=\|X-Y\|$, $L2=\|X-Y\|^2$, X 为所述图像样本， Y 为所述还原图像；

步骤 S307，调整所述文本特征提取模型、所述图像特征提取模型、所述注意力融合模型和所述图像还原模型中各节点的参数，至所述第一损失函数达到最小值时结束，得到训练好的文本特征提取模型、图像特征提取模型和注意力融合模型。

预设的文本特征提取模型基于 DPCNN 结构，预设的图像特征提取模型基于 Resnet 结构，注意融合模型由门控激活层，注意层以及融合层构成，图像还原模型基于 CNN 结构。

将文本样本和图像样本进行特征提取，然后进行特征融合，对融合后的特征进行图像还原，比较还原图像与图像样本之间的一致性，可以检验融合特征的完备性，确保图像信息融入融合特征用于文本分类。

通过调整文本特征提取模型、图像特征提取模型、注意力融合模型和图像还原模型中各节点的参数，使第一损失函数到达最小值，第一损失函数为：

$$L=L1+L2$$

其中 $L1=\|X-Y\|$, $L2=\|X-Y\|^2$, X 为所述图像样本， Y 为所述还原图像。第一损失函数达到最小值，认为还原图像与图像样本具有一致性，此时，文本特征提取模型、图像特征提取模型、注意力融合模型达到最优状态，训练结束。

参考上述实现方式，多模态信息样本中的文本样本标注有参考分类，在步骤 S304 之后，步骤 S205 之前，上述电子设备可以执行以下步骤：

将所述融合样本特征输入到预设的文本分类模型，获得所述文本分类模型响应所述融合样本特征输出的文本样本分类预测结果；

通过第二损失函数比较所述分类预测结果与所述参考分类是否一致，第二损失函数为：

$$L_{softmax loss} = -\frac{1}{N} \log\left(\frac{e^{hy_i}}{\sum_{j=1}^C e^{hj}}\right)$$

其中， N 为训练样本数，针对第 i 个样本其对应的 y_i 是标注的参考分类， $h=(h_1, h_2, \dots, h_c)$ 为样本 i 的分类预测结果，其中 C 是所有分类的数量；

调整所述文本分类模型各节点的参数，至所述第二损失函数达到最小时结束，得到训练好的文本分类模型。

这里为文本分类模型的训练过程，将多模态信息样本中的文本样本标注参考分类，将多模态信息样本中的文本样本和图像样本进行特征提取和融合得到的融合样本特征输入到文本分类模型中，文本分类模型基于 textCNN 结构，调节文本分类模型各节点的参数，使文本分类模型输出的分类预测结果与标准的参考分类一致，文本分类模型训练完毕。这里使用 softmax 损失函数。

在一些可选的实现方式中，在步骤 S201 之前，上述电子设备可以执行以下步骤：

获取多模态信息，其中所述多模态信息至少包含音频信息；

将所述音频信息进行文字转换，获取待分类的文本。

在一些场景下，多模态信息包含音频信息，且需要对音频进行文字转换，而在文字转换的过程中，由于技术的限制，容易出现错误的转换，当用户需要基于文字进行内容理解判断时，容易出错。如果对音频信息转换的文字进行正确的分类，可以帮助用户理解音频

信息，也方便对多模态信息正确分类。例如，文字转换为“家树”，文本分类结果为“人名”，用户就不会理解为“浇树”的动作。

音频转文字通过通用的软件实现。

在一些可选的实现方式中，在步骤 S201 之后，上述电子设备可以执行以下步骤：

将所述待分类的文本基于 HMM 隐马尔可夫算法进行分词，得到所述待分类文本的分词结果；

将所述分词结果根据预设的语料字典构成文本张量；

将所述文本张量输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征。

将待分类文本使用 HMM 隐马尔可夫算法进行分词，隐马尔可夫算法应用于中文分词，给一个汉语句作为输入，以“BEMS”组成的序列串作为输出，然后再进行切词，进而得到输入句子的划分。其中，B 代表该字是词语中的起始字，M 代表是词语中的中间字，E 代表是词语中的结束字，S 则代表是单字成词。想要得到的是每个字的位置，但是看到的只是这些汉字，需要通过汉字来推出每个字在词语中的位置，并且每个字属于什么状态还和它之前的字有关，这是个 HMM 问题。此时，我们需要根据可观察状态的序列找到一个最可能的隐藏状态序列。具体实现可以基于 HMM 算法通过调用 python 中的相关函数实现。

根据预设的语料字典将分词结果构成包含时间维度的文本张量。按照时间间隔为 5s，构建文本张量<t, contents>；这里的 t 是指根据时间间隔获得的时刻，contents 就是上述方法所获得的这个时间间隔内的内容。

在一些可选的实现方式中，在步骤 S203 之前，上述电子设备可以执行以下步骤：

将所述多模态信息中的图像进行灰度化，得到所述图像的灰度图；

将所述灰度图进行二值化，得到所述图像的二维图像张量；

将所述二维图像张量输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征。

将多模态信息中的图像进行灰度化，灰度化的过程就是将每个像素点的 RGB 值统一成同一个值。灰度化后的图像将由三通道变为单通道，灰度化之后进行二值化，即设定灰度阈值，把大于灰度阈值像素灰度设为灰度极大值，这里进行归一化操作，即设为 1，把小于灰度阈值的像素灰度设为灰度极小值，即设为 0，从而实现二值化。经过灰度化和二值化处理的图像既保留了图像特征，又减少了数据复杂度。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，该计算机可读指令可存储于一计算机可读存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。其中，前述的存储介质可为磁碟、光盘、只读存储记忆体（Read-Only Memory, ROM）等非易失性存储介质，或随机存储记忆体（Random Access Memory, RAM）等。

应该理解的是，虽然附图的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，其可以以其他的顺序执行。而且，附图的流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，其执行顺序也不必然是依次进行，而是可以与其他步骤或者其他步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

进一步参考图 6，作为对上述图 2 所示方法的实现，本申请提供了一种基于多模态信息融合的文本分类装置的一个实施例，该装置实施例与图 2 所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

如图 6 所示，本实施例所述的基于多模态信息融合的文本分类装置 600 包括：获取模块 601、第一提取模块 602、第二提取模块 603、融合模块 604 以及分类模块 605。其中：

获取模块 601，用于获取待分类的文本，所述文本来源于多模态信息，其中所述多模

态信息至少还包含图像；

第一提取模块 602，用于将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；

第二提取模块 603，用于将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；

融合模块 604，用于将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；

分类模块 605，用于将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。

通过获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像；将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。通过文本特征和图像特征融合，基于融合后的特征进行文本分类，文本分类利用了图像信息，分类结果更准确。

在本实施例的一些可选的实现方式中，基于多模态信息融合的文本分类装置 600，还包括：

第一获取子模块，用于获取多模态信息样本，所述多模态信息样本至少包含文本样本和图像样本；

第一提取子模块，用于将所述多模态信息样本中的文本样本输入到预设的文本特征提取模型，获得所述文本样本的文本样本特征；

第二提取子模块，用于将所述多模态信息样本中的图像样本输入到预设的图像特征提取模型，获得所述图像样本的图像样本特征；

第一融合子模块，用于将所述文本样本特征和图像样本特征输入到预设的注意力融合模型，获得所述多模态信息样本的融合样本特征；

第一还原子模块，用于将所述融合样本特征输入到预设的图像还原模型进行图像还原，获得所述图像样本的还原图像；

第一计算子模块，用于通过第一损失函数比较所述还原图像和所述图像样本之间的一致性，所述第一损失函数为：

$$L=L1+L2$$

其中 $L1=\|X-Y\|$, $L2=\|X-Y\|^2$, X 为所述图像样本， Y 为所述还原图像；

第一调整子模块，用于调整所述文本特征提取模型、所述图像特征提取模型、所述注意力融合模型和所述图像还原模型中各节点的参数，至所述第一损失函数达到最小值时结束，得到训练好的文本特征提取模型、图像特征提取模型和注意力融合模型。

在本实施例的一些可选的实现方式中，基于多模态信息融合的文本分类装置 600，还包括：

第二获取子模块，用于获取多模态信息，其中所述多模态信息至少包含音频信息；

第一转换子模块，用于将所述音频信息进行文字转换，获取待分类的文本。

进一步的，所述基于多模态信息融合的文本分类装置，还包括：

第一分词子模块，用于将所述待分类的文本基于 HMM 隐马尔可夫算法进行分词，得到所述待分类文本的分词结果；

第一构建子模块，用于将所述分词结果根据预设的语料字典构成文本张量；

第三提取子模块，用于将所述文本张量输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征。

在本实施例的一些可选的实现方式中，基于多模态信息融合的文本分类装置 600，还包括：

第一处理子模块，用于将所述多模态信息中的图像进行灰度化，得到所述图像的灰度图；

第二处理子模块，用于将所述灰度图进行二值化，得到所述图像的二维图像张量；

第四提取子模块，用于将所述二维图像张量输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征。

在本实施例的一些可选的实现方式中，基于多模态信息融合的文本分类装置 600，还包括：

第一预测子模块，用于将所述融合样本特征输入到预设的文本分类模型，获得所述文本分类模型响应所述融合样本特征输出的文本样本分类预测结果；

第二计算子模块，用于通过第二损失函数比较所述分类预测结果与所述参考分类是否一致，第二损失函数为：

$$L_{soft\ max\ loss} = -\frac{1}{N} \log\left(\frac{e^{h_{yi}}}{\sum_{j=1}^C e^{h_j}}\right)$$

其中，N 为训练样本数，针对第 i 个样本其对应的 y_i 是标注的参考分类，h=(h₁, h₂, ..., h_c) 为样本 i 的分类预测结果，其中 C 是所有分类的数量；

第二调整子模块，用于调整所述文本分类模型各节点的参数，至所述第二损失函数达到最小时结束，得到训练好的文本分类模型。

为解决上述技术问题，本申请实施例还提供计算机设备。具体请参阅图 7，图 7 为本实施例计算机设备基本结构框图。

所述计算机设备 7 包括通过系统总线相互通信连接存储器 71、处理器 72、网络接口 73。需要指出的是，图中仅示出了具有组件 71-73 的计算机设备 7，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。其中，本技术领域技术人员可以理解，这里的计算机设备是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路 (Application Specific Integrated Circuit, ASIC)、可编程门阵列 (Field-Programmable Gate Array, FPGA)、数字处理器 (Digital Signal Processor, DSP)、嵌入式设备等。

所述计算机设备可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述计算机设备可以与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互。

所述存储器 71 至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器 (例如，SD 或 DX 存储器等)、随机访问存储器 (RAM)、静态随机访问存储器 (SRAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、可编程只读存储器 (PROM)、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器 71 可以是所述计算机设备 7 的内部存储单元，例如该计算机设备 7 的硬盘或内存。在另一些实施例中，所述存储器 71 也可以是所述计算机设备 7 的外部存储设备，例如该计算机设备 7 上配备的插接式硬盘，智能存储卡 (Smart Media Card, SMC)，安全数字 (Secure Digital, SD) 卡，闪存卡 (Flash Card) 等。当然，所述存储器 71 还可以既包括所述计算机设备 7 的内部存储单元也包括其外部存储设备。本实施例中，所述存储器 71 通常用于存储安装于所述计算机设备 7 的操作系统和各类应用软件，例如基于多模态信息融合的文本分类方法的计算机可读指令等。此外，所述存储器 71 还可以用于暂时地存储已经输出或者将要输出的各类数据。

所述处理器 72 在一些实施例中可以是中央处理器 (Central Processing Unit, CPU)、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器 72 通常用于控制所述计算

机设备7的总体操作。本实施例中，所述处理器72用于运行所述存储器71中存储的计算机可读指令或者处理数据，例如运行所述基于多模态信息融合的文本分类方法的计算机可读指令。

所述网络接口73可包括无线网络接口或有线网络接口，该网络接口73通常用于在所述计算机设备7与其他电子设备之间建立通信连接。

通过获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像；将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。通过文本特征和图像特征融合，基于融合后的特征进行文本分类，文本分类利用了图像信息，分类结果更准确。

本申请还提供了另一种实施方式，即提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，所述计算机可读指令可被至少一个处理器执行，以使所述至少一个处理器执行如上述的基于多模态信息融合的文本分类方法的步骤。所述计算机可读存储介质可以是非易失性，也可以是易失性。

通过获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像；将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。通过文本特征和图像特征融合，基于融合后的特征进行文本分类，文本分类利用了图像信息，分类结果更准确。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，空调器，或者网络设备等）执行本申请各个实施例所述的方法。

显然，以上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，附图中给出了本申请的较佳实施例，但并不限制本申请的专利范围。本申请可以以许多不同的形式来实现，相反地，提供这些实施例的目的是使对本申请的公开内容的理解更加透彻全面。尽管参照前述实施例对本申请进行了详细的说明，对于本领域的技术人员来而言，其依然可以对前述各具体实施方式所记载的技术方案进行修改，或者对其中部分技术特征进行等效替换。凡是利用本申请说明书及附图内容所做的等效结构，直接或间接运用在其他相关的技术领域，均同理在本申请专利保护范围之内。

权 利 要 求 书

1. 一种基于多模态信息融合的文本分类方法, 包括下述步骤:

获取待分类的文本, 所述文本来源于多模态信息, 其中所述多模态信息至少还包含图像;

将所述文本输入到预先训练的文本特征提取模型进行特征提取, 获得所述文本的文本特征;

将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取, 获得所述图像的图像特征;

将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合, 获得融合了所述文本特征和所述图像特征的融合特征;

将所述融合特征输入到预先训练的文本分类模型, 获得所述多模态信息中文本的分类结果。

2. 根据权利要求 1 所述的基于多模态信息融合的文本分类方法, 其中, 在所述将所述文本输入到预先训练的文本特征提取模型进行特征提取, 获得所述文本的文本特征的步骤之前还包括:

获取多模态信息样本, 所述多模态信息样本至少包含文本样本和图像样本;

将所述多模态信息样本中的文本样本输入到预设的文本特征提取模型, 获得所述文本样本的文本样本特征;

将所述多模态信息样本中的图像样本输入到预设的图像特征提取模型, 获得所述图像样本的图像样本特征;

将所述文本样本特征和图像样本特征输入到预设的注意力融合模型, 获得所述多模态信息样本的融合样本特征;

将所述融合样本特征输入到预设的图像还原模型进行图像还原, 获得所述图像样本的还原图像;

通过第一损失函数比较所述还原图像和所述图像样本之间的一致性, 所述第一损失函数为:

$$L=L1+L2$$

其中 $L1=\|X-Y\|$, $L2=\|X-Y\|^2$, X 为所述图像样本, Y 为所述还原图像;

调整所述文本特征提取模型、所述图像特征提取模型、所述注意力融合模型和所述图像还原模型中各节点的参数, 至所述第一损失函数达到最小值时结束, 得到训练好的文本特征提取模型、图像特征提取模型和注意力融合模型。

3. 根据权利要求 1 所述的基于多模态信息融合的文本分类方法, 其中, 在所述获取待分类的文本的步骤之前, 还包括:

获取多模态信息, 其中所述多模态信息至少包含音频信息;

将所述音频信息进行文字转换, 获取待分类的文本。

4. 根据权利要求 1 所述的基于多模态信息融合的文本分类方法, 其中, 在所述获取待分类的文本的步骤之后, 还包括:

将所述待分类的文本基于 HMM 隐马尔可夫算法进行分词, 得到所述待分类文本的分词结果;

将所述分词结果根据预设的语料字典构成文本张量;

将所述文本张量输入到预先训练的文本特征提取模型进行特征提取, 获得所述文本的文本特征。

5. 根据权利要求 1 所述的基于多模态信息融合的文本分类方法, 其中, 在所述将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取, 获得所述图像的图像特征的步骤之前, 还包括:

将所述多模态信息中的图像进行灰度化, 得到所述图像的灰度图;

将所述灰度图进行二值化, 得到所述图像的二维图像张量;

将所述二维图像张量输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征。

6. 根据权利要求2所述的基于多模态信息融合的文本分类方法，其中，所述多模态信息样本中的文本样本标注有参考分类，在所述将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果的步骤之前，在所述将所述文本样本特征和图像样本特征输入到预设的注意力融合模型，获得所述多模态信息样本的融合样本特征的步骤之后，还包括：

将所述融合样本特征输入到预设的文本分类模型，获得所述文本分类模型响应所述融合样本特征输出的文本样本分类预测结果；

通过第二损失函数比较所述分类预测结果与所述参考分类是否一致，第二损失函数为：

$$L_{soft\ max\ loss} = -\frac{1}{N} \log\left(\frac{e^{hy_i}}{\sum_{j=1}^C e^{hj}}\right)$$

其中，N为训练样本数，针对第i个样本其对应的 y_i 是标注的参考分类， $h=(h_1, h_2, \dots, h_C)$ 为样本i的分类预测结果，其中C是所有分类的数量；

调整所述文本分类模型各节点的参数，至所述第二损失函数达到最小时结束，得到训练好的文本分类模型。

7. 一种基于多模态信息融合的文本分类装置，包括：

获取模块，用于获取待分类的文本，所述文本来源于多模态信息，其中所述多模态信息至少还包含图像；

第一提取模块，用于将所述文本输入到预先训练的文本特征提取模型进行特征提取，获得所述文本的文本特征；

第二提取模块，用于将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取，获得所述图像的图像特征；

融合模块，用于将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合，获得融合了所述文本特征和所述图像特征的融合特征；

分类模块，用于将所述融合特征输入到预先训练的文本分类模型，获得所述多模态信息中文本的分类结果。

8. 根据权利要求7所述的基于多模态信息融合的文本分类装置，还包括：

第一获取子模块，用于获取多模态信息样本，所述多模态信息样本至少包含文本样本和图像样本；

第一提取子模块，用于将所述多模态信息样本中的文本样本输入到预设的文本特征提取模型，获得所述文本样本的文本样本特征；

第二提取子模块，用于将所述多模态信息样本中的图像样本输入到预设的图像特征提取模型，获得所述图像样本的图像样本特征；

第一融合子模块，用于将所述文本样本特征和图像样本特征输入到预设的注意力融合模型，获得所述多模态信息样本的融合样本特征；

第一还原子模块，用于将所述融合样本特征输入到预设的图像还原模型进行图像还原，获得所述图像样本的还原图像；

第一计算子模块，用于通过第一损失函数比较所述还原图像和所述图像样本之间的一致性，所述第一损失函数为：

$$L=L1+L2$$

其中 $L1=\|X-Y\|$, $L2=\|X-Y\|^2$, X为所述图像样本，Y为所述还原图像；

第一调整子模块，用于调整所述文本特征提取模型、所述图像特征提取模型、所述注意力融合模型和所述图像还原模型中各节点的参数，至所述第一损失函数达到最小值时结

束, 得到训练好的文本特征提取模型、图像特征提取模型和注意力融合模型。

9. 一种计算机设备, 包括存储器和处理器, 所述存储器中存储有计算机可读指令, 所述处理器执行所述计算机可读指令时实现如下步骤:

获取待分类的文本, 所述文本来源于多模态信息, 其中所述多模态信息至少还包含图
5 像;

将所述文本输入到预先训练的文本特征提取模型进行特征提取, 获得所述文本的文本特征;

将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取, 获得所述图像的图像特征;

10 将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合, 获得融合了所述文本特征和所述图像特征的融合特征;

将所述融合特征输入到预先训练的文本分类模型, 获得所述多模态信息中文本的分类结果。

15 10. 根据权利要求 9 所述的计算机设备, 其中, 在所述将所述文本输入到预先训练的文本特征提取模型进行特征提取, 获得所述文本的文本特征的步骤之前, 所述处理器执行所述计算机可读指令时还实现如下步骤:

获取多模态信息样本, 所述多模态信息样本至少包含文本样本和图像样本;

将所述多模态信息样本中的文本样本输入到预设的文本特征提取模型, 获得所述文本样本的文本样本特征;

20 将所述多模态信息样本中的图像样本输入到预设的图像特征提取模型, 获得所述图像样本的图像样本特征;

将所述文本样本特征和图像样本特征输入到预设的注意力融合模型, 获得所述多模态信息样本的融合样本特征;

25 将所述融合样本特征输入到预设的图像还原模型进行图像还原, 获得所述图像样本的还原图像;

通过第一损失函数比较所述还原图像和所述图像样本之间的一致性, 所述第一损失函数为:

$$L=L1+L2$$

其中 $L1=\|X-Y\|$, $L2=\|X-Y\|^2$, X 为所述图像样本, Y 为所述还原图像;

30 调整所述文本特征提取模型、所述图像特征提取模型、所述注意力融合模型和所述图像还原模型中各节点的参数, 至所述第一损失函数达到最小值时结束, 得到训练好的文本特征提取模型、图像特征提取模型和注意力融合模型。

11. 根据权利要求 9 所述的计算机设备, 其中, 在所述获取待分类的文本的步骤之前, 所述处理器执行所述计算机可读指令时还实现如下步骤:

35 获取多模态信息, 其中所述多模态信息至少包含音频信息;

将所述音频信息进行文字转换, 获取待分类的文本。

12. 根据权利要求 9 所述的计算机设备, 其中, 在所述获取待分类的文本的步骤之后, 所述处理器执行所述计算机可读指令时还实现如下步骤:

40 将所述待分类的文本基于 HMM 隐马尔可夫算法进行分词, 得到所述待分类文本的分词结果;

将所述分词结果根据预设的语料字典构成文本张量;

将所述文本张量输入到预先训练的文本特征提取模型进行特征提取, 获得所述文本的文本特征。

45 13. 根据权利要求 9 所述的计算机设备, 其中, 在所述将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取, 获得所述图像的图像特征的步骤之前, 所述处理器执行所述计算机可读指令时还实现如下步骤:

将所述多模态信息中的图像进行灰度化,得到所述图像的灰度图;

将所述灰度图进行二值化,得到所述图像的二维图像张量;

将所述二维图像张量输入到预先训练的图像特征提取模型进行特征提取,获得所述图像的图像特征。

- 5 14. 根据权利要求 10 所述的计算机设备,其中,所述多模态信息样本中的文本样本标注有参考分类,在所述将所述融合特征输入到预先训练的文本分类模型,获得所述多模态信息中文本的分类结果的步骤之前,在所述将所述文本样本特征和图像样本特征输入到预设的注意力融合模型,获得所述多模态信息样本的融合样本特征的步骤之后,所述处理器执行所述计算机可读指令时还实现如下步骤:

- 10 将所述融合样本特征输入到预设的文本分类模型,获得所述文本分类模型响应所述融合样本特征输出的文本样本分类预测结果;

通过第二损失函数比较所述分类预测结果与所述参考分类是否一致,第二损失函数为:

$$L_{soft\ max\ loss} = -\frac{1}{N} \log\left(\frac{e^{h_{yi}}}{\sum_{j=1}^C e^{h_{hj}}}\right)$$

- 15 其中,N为训练样本数,针对第 i 个样本其对应的 yi 是标注的参考分类,h=(h1,h2,...,hc)为样本 i 的分类预测结果,其中 C 是所有分类的数量;

调整所述文本分类模型各节点的参数,至所述第二损失函数达到最小时结束,得到训练好的文本分类模型。

- 20 15. 一种计算机可读存储介质,所述计算机可读存储介质上存储有计算机可读指令,所述计算机可读指令被处理器执行时,使得所述处理器执行如下步骤:

获取待分类的文本,所述文本来源于多模态信息,其中所述多模态信息至少还包含图像;

将所述文本输入到预先训练的文本特征提取模型进行特征提取,获得所述文本的文本特征;

- 25 将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取,获得所述图像的图像特征;

将所述文本特征和所述图像特征输入到预先训练的注意力融合模型进行特征融合,获得融合了所述文本特征和所述图像特征的融合特征;

- 30 将所述融合特征输入到预先训练的文本分类模型,获得所述多模态信息中文本的分类结果。

16. 根据权利要求 15 所述的计算机可读存储介质,其中,在所述将所述文本输入到预先训练的文本特征提取模型进行特征提取,获得所述文本的文本特征的步骤之前,所述计算机可读指令被所述处理器执行时,使得所述处理器还执行如下步骤:

获取多模态信息样本,所述多模态信息样本至少包含文本样本和图像样本;

- 35 将所述多模态信息样本中的文本样本输入到预设的文本特征提取模型,获得所述文本样本的文本样本特征;

将所述多模态信息样本中的图像样本输入到预设的图像特征提取模型,获得所述图像样本的图像样本特征;

- 40 将所述文本样本特征和图像样本特征输入到预设的注意力融合模型,获得所述多模态信息样本的融合样本特征;

将所述融合样本特征输入到预设的图像还原模型进行图像还原,获得所述图像样本的还原图像;

通过第一损失函数比较所述还原图像和所述图像样本之间的一致性,所述第一损失函数为:

$$L=L1+L2$$

其中 $L1=\|X-Y\|$, $L2=\|X-Y\|^2$, X 为所述图像样本, Y 为所述还原图像;

调整所述文本特征提取模型、所述图像特征提取模型、所述注意力融合模型和所述图像还原模型中各节点的参数, 至所述第一损失函数达到最小值时结束, 得到训练好的文本特征提取模型、图像特征提取模型和注意力融合模型。

17. 根据权利要求 15 所述的计算机可读存储介质, 其中, 在所述获取待分类的文本的步骤之前, 所述计算机可读指令被所述处理器执行时, 使得所述处理器还执行如下步骤:

获取多模态信息, 其中所述多模态信息至少包含音频信息;

将所述音频信息进行文字转换, 获取待分类的文本。

18. 根据权利要求 15 所述的计算机可读存储介质, 其中, 在所述获取待分类的文本的步骤之后, 所述计算机可读指令被所述处理器执行时, 使得所述处理器还执行如下步骤:

将所述待分类的文本基于 HMM 隐马尔可夫算法进行分词, 得到所述待分类文本的分词结果;

将所述分词结果根据预设的语料字典构成文本张量;

将所述文本张量输入到预先训练的文本特征提取模型进行特征提取, 获得所述文本的文本特征。

19. 根据权利要求 15 所述的计算机可读存储介质, 其中, 在所述将所述多模态信息中的图像输入到预先训练的图像特征提取模型进行特征提取, 获得所述图像的图像特征之前, 所述计算机可读指令被所述处理器执行时, 使得所述处理器还执行如下步骤:

将所述多模态信息中的图像进行灰度化, 得到所述图像的灰度图;

将所述灰度图进行二值化, 得到所述图像的二维图像张量;

将所述二维图像张量输入到预先训练的图像特征提取模型进行特征提取, 获得所述图像的图像特征。

20. 根据权利要求 16 所述的计算机可读存储介质, 其中, 所述多模态信息样本中的文本样本标注有参考分类, 在所述将所述融合特征输入到预先训练的文本分类模型, 获得所述多模态信息中文本的分类结果的步骤之前, 在所述将所述文本样本特征和图像样本特征输入到预设的注意力融合模型, 获得所述多模态信息样本的融合样本特征的步骤之后, 所述计算机可读指令被所述处理器执行时, 使得所述处理器还执行如下步骤:

将所述融合样本特征输入到预设的文本分类模型, 获得所述文本分类模型响应所述融合样本特征输出的文本样本分类预测结果;

通过第二损失函数比较所述分类预测结果与所述参考分类是否一致, 第二损失函数为:

$$L_{soft\ max\ loss} = -\frac{1}{N} \log\left(\frac{e^{h_{yi}}}{\sum_{j=1}^C e^{h_{hj}}}\right)$$

其中, N 为训练样本数, 针对第 i 个样本其对应的 y_i 是标注的参考分类, $h=(h_1, h_2, \dots, h_C)$ 为样本 i 的分类预测结果, 其中 C 是所有分类的数量;

调整所述文本分类模型各节点的参数, 至所述第二损失函数达到最小时结束, 得到训练好的文本分类模型。

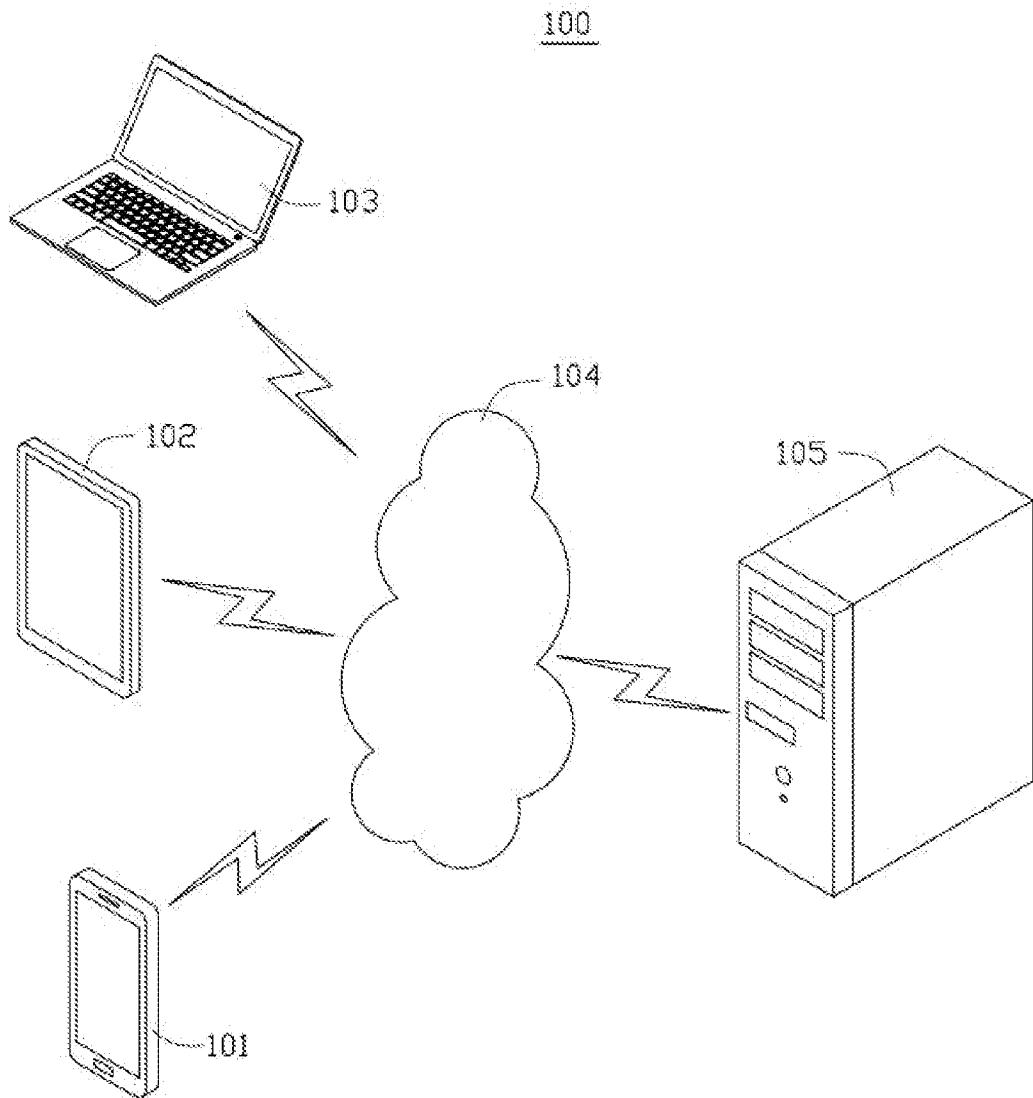


图 1

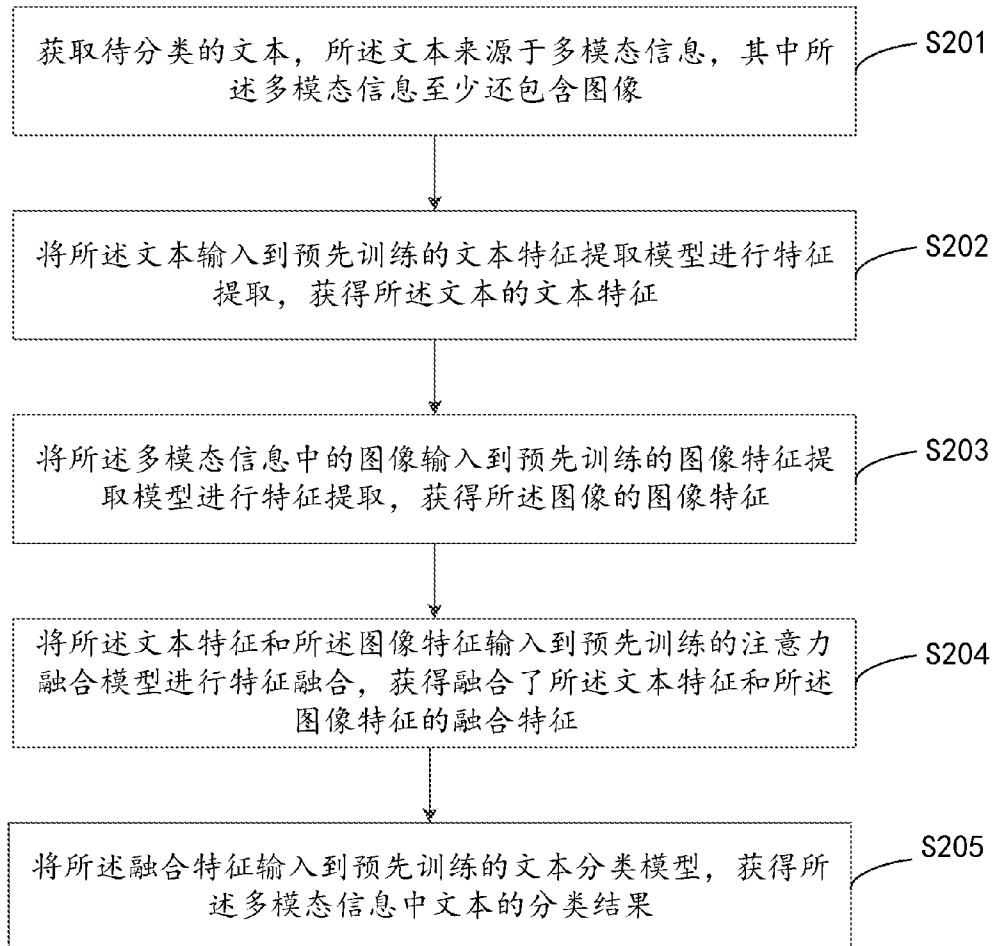


图 2

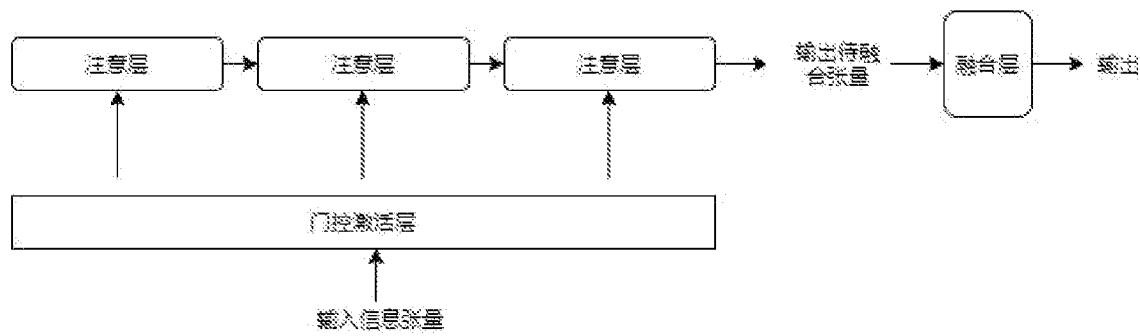


图 3

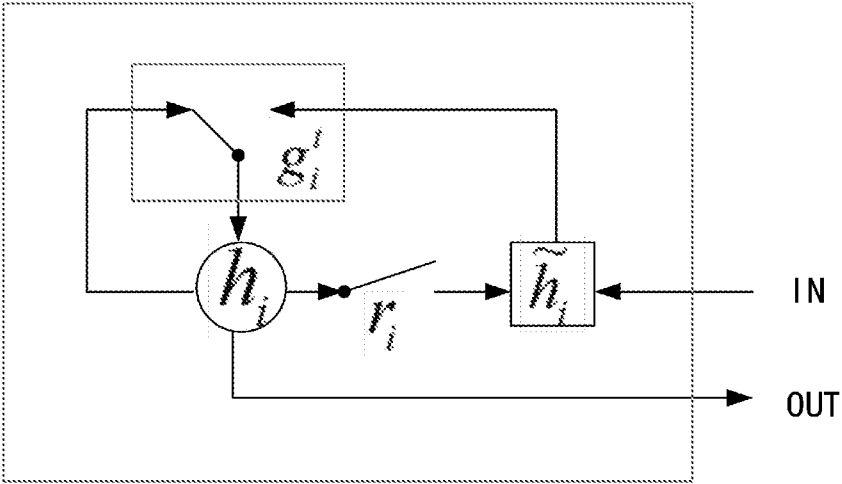


图 4

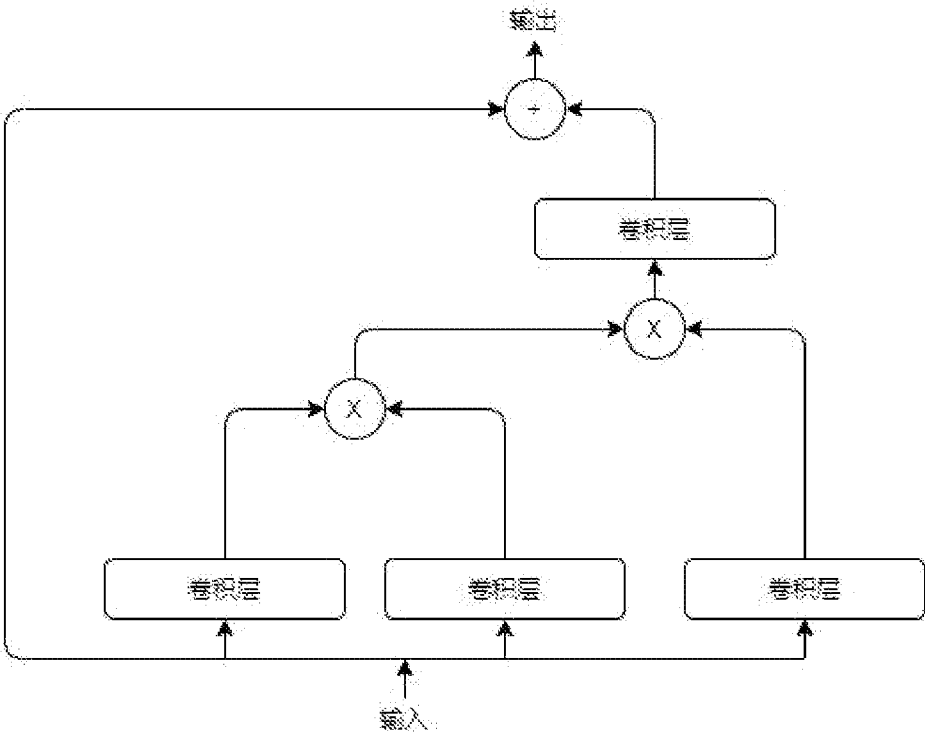


图 5

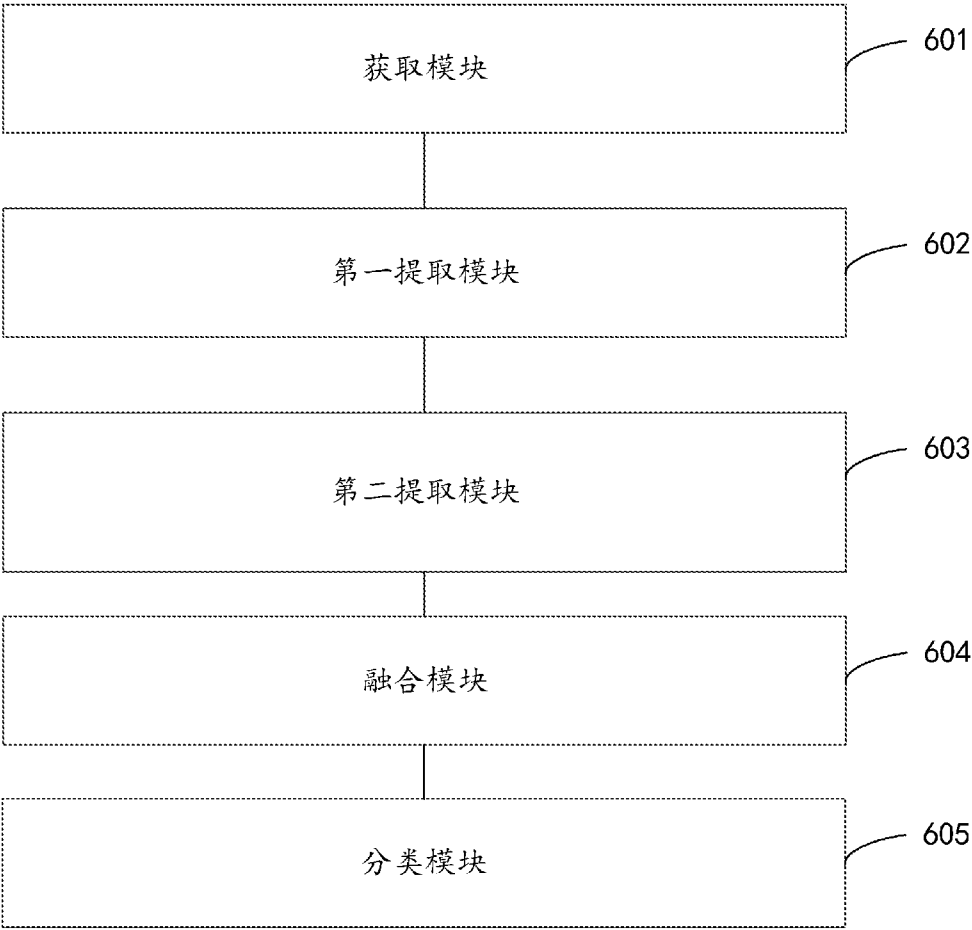


图 6

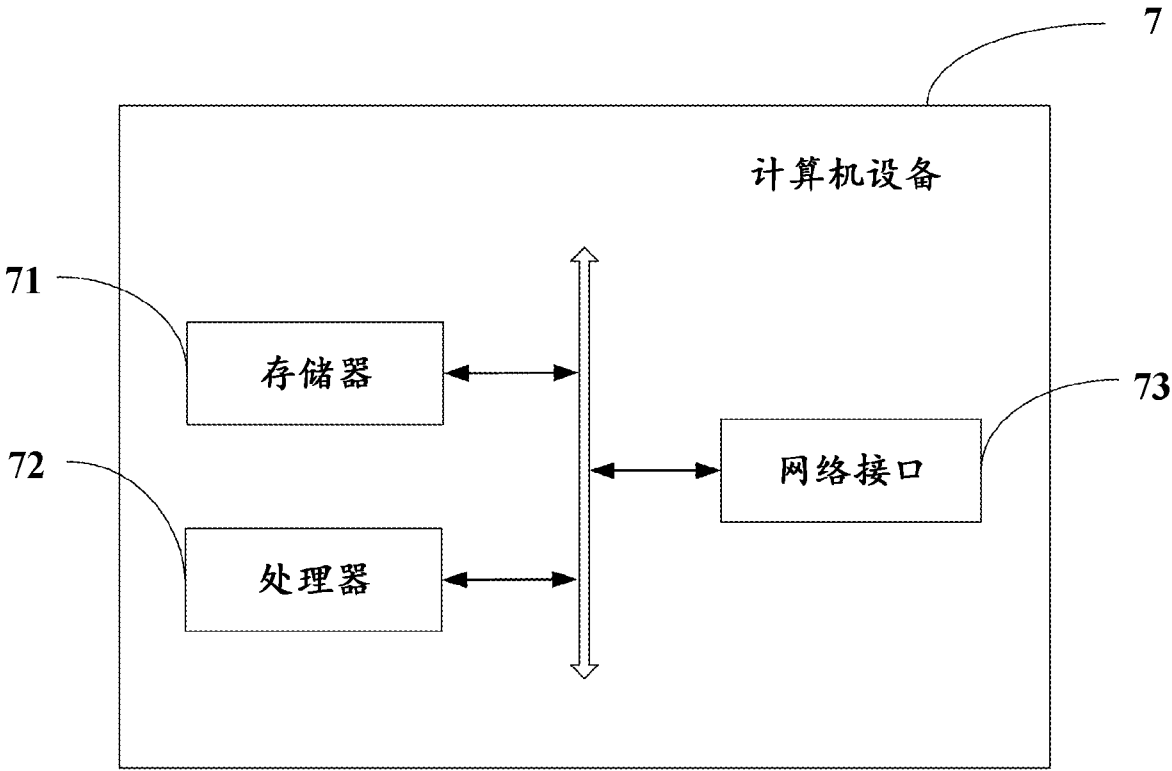


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/090497

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/35(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI: 文本, 文档, 图像, 图片, 融合, 特征, 模型, 分类, 多模态, 注意力, text, document, image, picture, fusion, feature, model, classif+, multi-modal, attention

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112685565 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 20 April 2021 (2021-04-20) claims 1-10	1-20
X	CN 111985369 A (NORTHWESTERN POLYTECHNICAL UNIVERSITY) 24 November 2020 (2020-11-24) claims 1-7	1,3-5,7,9,11-13,15,17-19
A	CN 111259215 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 09 June 2020 (2020-06-09) entire document	1-20
A	CN 110717335 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 21 January 2020 (2020-01-21) entire document	1-20
A	US 2019341025 A1 (SONY INTERACTIVE ENTERTAINMENT INC.) 07 November 2019 (2019-11-07) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

15 September 2021

Date of mailing of the international search report

28 September 2021

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/090497

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	112685565	A	20 April 2021	None			
CN	111985369	A	24 November 2020	None			
CN	111259215	A	09 June 2020	None			
CN	110717335	A	21 January 2020	None			
US	2019341025	A1	07 November 2019	WO	2019204186	A1	24 October 2019

国际检索报告

国际申请号

PCT/CN2021/090497

A. 主题的分类

G06F 16/35 (2019.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT, WPI, EP0DOC, CNKI: 文本, 文档, 图像, 图片, 融合, 特征, 模型, 分类, 多模态, 注意力, text, document, image, picture, fusion, feature, model, classif+, multi-modal, attention

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 112685565 A (平安科技深圳有限公司) 2021年 4月 20日 (2021 - 04 - 20) 权利要求1-10	1-20
X	CN 111985369 A (西北工业大学) 2020年 11月 24日 (2020 - 11 - 24) 权利要求1-7	1, 3-5, 7, 9, 11-13, 15, 17-19
A	CN 111259215 A (北京百度网讯科技有限公司) 2020年 6月 9日 (2020 - 06 - 09) 全文	1-20
A	CN 110717335 A (中国科学院深圳先进技术研究院) 2020年 1月 21日 (2020 - 01 - 21) 全文	1-20
A	US 2019341025 A1 (SONY INTERACTIVE ENTERTAINMENT INC.) 2019年 11月 7日 (2019 - 11 - 07) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2021年 9月 15日

国际检索报告邮寄日期

2021年 9月 28日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国 北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

林婉娟

电话号码 86-(10)-53961343

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/090497

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	112685565	A	2021年 4月 20日	无	
CN	111985369	A	2020年 11月 24日	无	
CN	111259215	A	2020年 6月 9日	无	
CN	110717335	A	2020年 1月 21日	无	
US	2019341025	A1	2019年 11月 7日	WO 2019204186 A1	2019年 10月 24日