(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115861663 A
(43)申请公布日 2023.03.28

(21)申请号 202310185709.2

(22)申请日 2023.03.01

(71)申请人 合肥综合性国家科学中心人工智能研究院(安徽省人工智能实验室)
地址 230000 安徽省合肥市高新区望江西路5089号,中国科学技术大学先进技术研究院未来中心B1205-B1208

(72)发明人 刘学亮 汪萌

(74)专利代理机构 合肥天明专利事务所(普通合伙) 34115
专利代理师 金凯 谢中用

(51)Int.Cl.
G06V 10/74(2022.01)
G06V 30/418(2022.01)
G06V 10/82(2022.01)
G06N 3/0464(2023.01)
G06N 3/0895(2023.01)

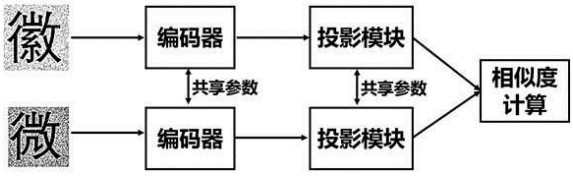
权利要求书2页 说明书4页 附图1页

(54)发明名称

一种基于自监督学习模型的文档图像内容比对方法

(57)摘要

本发明涉及计算机视觉技术领域,公开了一种基于自监督学习模型的文档图像内容比对方法,包括以下步骤:将纸质文档转化为扫描版图像,电子文档转化为电子版图像;提取扫描版图像和电子版图像中的字符图像并依次排序,分别形成扫描字符图像序列和电子字符图像序列;将扫描字符图像序列和电子字符图像序列输入至文字比对自监督模型的编码器中,分别得到扫描图像编码序列和电子图像编码序列;计算扫描图像编码序列和电子图像编码序列相同位置i上的两个元素的余弦相似度,当余弦相似度小于设定的阈值时,则认为电子文档位置i处的内容被篡改。本发明不需要检测图像中每个字符的具体内容完成比对,能够提高比对的效率。



CN 115861663 A

1.一种基于自监督学习模型的文档图像内容比对方法,包括以下步骤:

步骤A1:将纸质文档转化为扫描版图像,电子文档转化为电子版图像;

步骤A2:提取扫描版图像和电子版图像中的字符图像并依次排序,分别形成扫描字符图像序列和电子字符图像序列;

步骤A3:将扫描字符图像序列和电子字符图像序列输入至文字比对自监督模型的编码器中,分别得到扫描图像编码序列和电子图像编码序列;

步骤A4:计算扫描图像编码序列和电子图像编码序列相同位置i上的两个元素的余弦相似度S_i,当余弦相似度S_i小于设定的阈值时,则认为电子文档位置i处的内容被篡改;

文字比对自监督模型的训练方法,包括以下步骤:

步骤B1:将字典库中的每个字符转化为字符图像,对字符图像进行数据增广后,将每个由相同字符得到的字符图像赋予相同的标签,得到字符图像数据集;

步骤B2:构建采用共享参数、双塔结构的文字比对自监督模型,文字比对自监督模型包括编码器和投影模块;

编码器用于对输入的字符图像进行编码;编码器包括九层结构:前七层结构为ResNet结构,ResNet结构之后连接一个最大池化层;后两层结构为全连接层,激活函数为ReLu;

投影模块包括两层全连接网络结构,第一层全连接网络结构的激活函数为ReLu,第二层全连接网络结构的激活函数为tanh;

步骤B3:通过字符图像数据集对文字比对自监督模型进行训练;对于给定的字符图像x,记编码器的输出为$h(x)$,投影模块的输出为$f(x)$;文字比对自监督模型的损失函数L为:

$$L=-log\frac{exp(f(x)^{T}f(x^{+}))}{exp(f(x)^{T}f(x^{+}))+\sum_{j=1}^{N}exp(f(x)^{T}f(x_{j}))};$$

其中,x^{+}、x_j分别为与字符图像x具有相同标签的字符图像、与字符图像x具有不同标签的字符图像,N为与字符图像x具有不同标签的字符图像的总数量,T表示转置。

2.根据权利要求1所述的基于自监督学习模型的文档图像内容比对方法,其特征在于:文字比对自监督模型进行训练时,每批数据包括128个字符图像,采用Adam优化器,学习速率设置为0.001。

3.根据权利要求1所述的基于自监督学习模型的文档图像内容比对方法,其特征在于:步骤B1中,利用添加噪声、改变亮度和对比度,或者采用形态学算法对字符图像进行数据增广。

4.根据权利要求1所述的基于自监督学习模型的文档图像内容比对方法,其特征在于:步骤A1中,通过扫描仪将纸质文档转化为扫描版图像,并进行透视变换校正扫描版图像中的畸变和倾斜。

5.根据权利要求1所述的基于自监督学习模型的文档图像内容比对方法,其特征在于:步骤A2中提取字符图像并依次排序时,利用文本检测工具进行字符检测,得到每个独立字符的左上角坐标和右下角坐标,根据字符左上角坐标和右下角坐标,将每个字符图像抠出,

归一化到统一尺寸后，根据左上角坐标的坐标值，按照从上到下、从左到右的顺序进行排序，形成字符图像序列。

一种基于自监督学习模型的文档图像内容比对方法

技术领域

[0001] 本发明涉及计算机视觉领域,具体涉及一种基于自监督学习模型的文档图像内容比对方法。

背景技术

[0002] 纸质文档一般通过文字处理软件制作成电子文档,并打印而成。文字处理软件制作的电子文档在传输过程中,文档内容被篡改的事情时有发生,这导致了最终打印的文档和原始的电子文档不一致的问题。

[0003] 为了解决电子文档流通过程中被篡改问题,目前的技术一般通过光学字符识别技术对纸质文件的扫描件进行识别,再将识别结果和电子文档进行对比。但是光学字符识别技术存在精度不高、算力消耗大的问题。

发明内容

[0004] 为解决上述技术问题,本发明提供一种基于自监督学习模型的文档图像内容比对方法,在对文档进行字符图像检测的基础上,通过计算单个字符图像之间的相似性,来实现快速比对的功能。

[0005] 为解决上述技术问题,本发明采用如下技术方案:

一种基于自监督学习模型的文档图像内容比对方法,包括以下步骤:

步骤A1:将纸质文档转化为扫描版图像,电子文档转化为电子版图像;

步骤A2:提取扫描版图像和电子版图像中的字符图像并依次排序,分别形成扫描字符图像序列和电子字符图像序列;

步骤A3:将扫描字符图像序列和电子字符图像序列输入至文字比对自监督模型的编码器中,分别得到扫描图像编码序列和电子图像编码序列;

步骤A4:计算扫描图像编码序列和电子图像编码序列相同位置i上的两个元素的余弦相似度S_i,当余弦相似度S_i小于设定的阈值时,则认为电子文档位置i处的内容被篡改;

文字比对自监督模型的训练方法,包括以下步骤:

步骤B1:将字典库中的每个字符转化为字符图像,对字符图像进行数据增广后,将每个由相同字符得到的字符图像赋予相同的标签,得到字符图像数据集;

步骤B2:构建采用共享参数、双塔结构的文字比对自监督模型,文字比对自监督模型包括编码器和投影模块;

编码器用于对输入的字符图像进行编码;编码器包括九层结构:前七层结构为ResNet结构,ResNet结构之后连接一个最大池化层;后两层结构为全连接层,激活函数为ReLu;

投影模块包括两层全连接网络结构,第一层全连接网络结构的激活函数为ReLu,第二层全连接网络结构的激活函数为tanh;

步骤B3:通过字符图像数据集对文字比对自监督模型进行训练;对于给定的字符图像x,记编码器的输出为$h(x)$,投影模块的输出为$f(x)$;文字比对自监督模型的损失函数L为:

$$L=-log\frac{exp(f(x)^{T}f(x^{+}))}{exp(f(x)^{T}f(x^{+}))+\sum_{j=1}^{N}exp(f(x)^{T}f(x_{j}))};$$

其中,x^{+}、x_{j}分别为与字符图像x具有相同标签的字符图像、与字符图像x具有不同标签的字符图像,N为与字符图像x具有不同标签的字符图像的总数量,T表示转置。

[0006] 进一步地,文字比对自监督模型进行训练时,每批数据包括128个字符图像,采用Adam优化器,学习速率设置为0.001。

[0007] 进一步地,步骤B1中,利用添加噪声、改变亮度和对比度,或者采用形态学算法对字符图像进行数据增广。

[0008] 进一步地,步骤A1中,通过扫描仪将纸质文档转化为扫描版图像,并进行透视变换校正扫描版图像中的畸变和倾斜。

[0009] 进一步地,步骤A2中提取字符图像并依次排序时,利用文本检测工具进行字符检测,得到每个独立字符的左上角坐标和右下角坐标,根据字符左上角坐标和右下角坐标,将每个字符图像抠出,归一化到统一尺寸后,根据左上角坐标的坐标值,按照从上到下、从左到右的顺序进行排序,形成字符图像序列。

[0010] 与现有技术相比,本发明的有益技术效果是:

目前常用的文本比对方法,通常需要对文档图像进行光学字符识别(OCR)处理,将图像中的文本转换为可比对的文本格式。然后,通过计算单个文字之间的相似性,来判断它们是否相同;运算量比较大。

[0011] 本发明对纸质文档的扫描版图像和电子文档的电子版图像进行字符检测,对字符图像序列进行编码,并通过计算字符图像编码序列相同位置的相似度,来判断电子文档是否被篡改;本发明不需要检测图像中每个字符的具体内容,即可完成比对,能够提高比对的效率。

附图说明

[0012] 图1为本发明文字比对自监督模型的训练流程图;

图2为本发明文档图像内容比对方法的流程图。

具体实施方式

[0013] 下面结合附图对本发明的一种优选实施方式作详细的说明。

[0014] 文字比对自监督模型的训练方法,包括以下步骤:

B1、文字图像数据集的建立:

B11:收集常用的汉字、数字、英文字母和符号等字符,作为模型字典库。

[0015] B12:对模型字典库中的每个字符,利用文图转化工具,采用宋体、楷书、仿宋、黑体

等常用字体，生成一系列的字符图像，字符图像的大小固定为32像素×32像素。

[0016] B13：对生成的字符图像，利用添加噪声、改变亮度和对比度、形态学算法等方法，进行数据集增广。

[0017] B14：将每个由相同字符得到的字符图像赋予相同的标签，得到字符图像数据集。

[0018] B2、文字比对自监督模型的构建：

文字比对自监督模型采用共享参数的双塔结构，主要由编码器、投影模块两个部分组成。

[0019] 编码器实现对输入的字符图像进行编码，它是一个由卷积和池化构成的、具有九层结构的网络模型。其中，前七层结构为ResNet结构，每个ResNet结构中包含两层卷积核大小为3×3的卷积层，卷积核个数分别为32、64、128；ResNet结构之后连接一个最大池化层。后两层结构为全连接层，两个全连接层输出尺寸为分别为512和128，激活函数为ReLu。

[0020] 投影模块是一个两层的全连接网络结构，输出尺寸分别为128和64；第一层全连接网络结构的激活函数为ReLu，第二层全连接网络结构的激活函数为tanh。

[0021] 对于给定的字符图像x，记模型编码器的输出为$h(x)$，投影模块的输出为$f(x)$。

[0022] 文字比对自监督模型损失函数如下：

$$L=-log\frac{exp(f(x)^{T}f(x^{+}))}{exp(f(x)^{T}f(x^{+}))+\sum_{j=1}^{N}exp(f(x)^{T}f(x_{j}))};$$

x^{+}、x_{j}分别为与字符图像x具有相同标签的字符图像、与字符图像x具有不同标签的字符图像。

[0023] 利用字符图像数据集，采用Adam优化器，对文字比对自监督模型进行优化求解。每批数据采样128个字符图像，学习速率设置为0.001。Adam是一种常用的优化器，用于训练神经网络模型，它可以通过指定学习速率和其他参数，来优化模型的求解过程。

[0024] 采用Adam优化器、设置合适的学习速率0.001以及每批数据采样的字符图像数量等参数，可以提高文字比对自监督模型的训练效率，并有效提升文字比对自监督模型的预测准确性。

[0025] 文档图像内容比对方法，包括以下步骤：

A1、转化文档图像：

对纸质文档进行扫描，得到扫描版图像，并通过透视变换，对扫描版图像进行校正。扫描仪能够捕捉纸质文档的图像，通常会将纸质文档的图像保存为图像格式的扫描件，如JPEG、PNG等图像格式。扫描件可能存在畸变或倾斜等问题，需要利用透视变换进行校正。透视变换通过改变图像的尺寸、方向和形状，来校正图像中的畸变、倾斜等问题。

[0026] 将电子文档转化为电子版图像。

[0027] A2：对扫描版图像和电子版图像，分别利用文本检测工具进行字符检测，得到每个独立字符的左上角的坐标和右下角的坐标；根据字符左上角的坐标和右下角的坐标，将每个字符图像抠出来，并归一化到32像素×32像素。然后，根据左上角的坐标值，按照从上到下、从左到右的顺序进行排序，分别得到扫描字符图像序列和电子字符图像序列。

[0028] A3：将扫描字符图像序列和电子字符图像序列输入至文字比对自监督模型的编码

器中，分别得到扫描图像编码序列和电子图像编码序列。

[0029] A4：计算扫描图像编码序列和电子图像编码序列相同位置i上的两个元素的余弦相似度S_i，当余弦相似度S_i小于设定的阈值(0.22)时，则认为电子文档位置i处的内容被篡改。

[0030] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。因此无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化囊括在本发明内，不应将权利要求中的任何附图标记视为限制所涉及的权利要求。

[0031] 此外，应当理解，虽然本说明书按照实施方式加以描述，但并非每个实施方式仅包含一个独立技术方案，说明书的这种叙述方式仅仅是为了清楚起见，本领域技术人员应当将说明书作为一个整体，各实施例中的技术方案也可以经适当组合，形成本领域技术人员可以理解的其他实施方式。

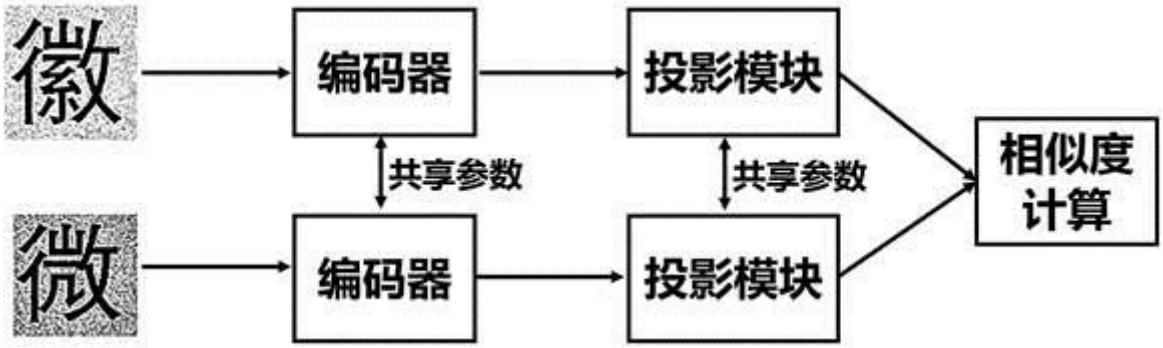

徽
编码器
投影模块
共享参数
共享参数
相似度计算
微
编码器
投影模块

图1

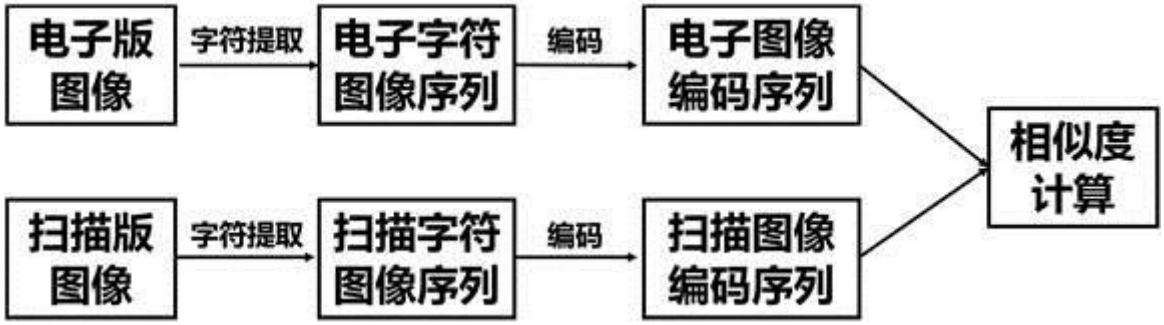

电子版图像
字符提取
电子字符图像序列
编码
电子图像编码序列
相似度计算
扫描版图像
字符提取
扫描字符图像序列
编码
扫描图像编码序列

图2