



(12) 发明专利申请

(10) 申请公布号 CN 112347742 A

(43) 申请公布日 2021.02.09

(21) 申请号 202011178681.2

G06N 3/04 (2006.01)

(22) 申请日 2020.10.29

G06N 3/08 (2006.01)

(71) 申请人 青岛科技大学

地址 266000 山东省青岛市崂山区松岭路 99号

(72) 发明人 史操 许灿辉 刘传琦 程远志

陶冶 马兴录 刘国柱

(74) 专利代理机构 青岛中天汇智知识产权代理

有限公司 37241

代理人 王丹丹 刘晓

(51) Int. Cl.

G06F 40/166 (2020.01)

G06F 40/189 (2020.01)

G06F 40/151 (2020.01)

G06T 11/60 (2006.01)

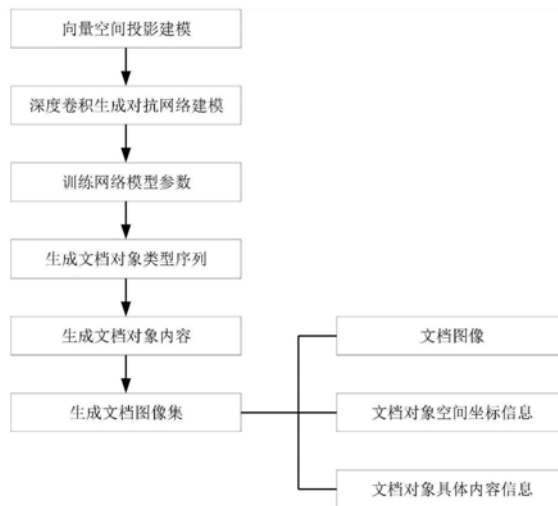
权利要求书3页 说明书11页 附图5页

(54) 发明名称

基于深度学习生成文档图像集的方法

(57) 摘要

本发明公开一种基于深度学习生成文档图像集的方法:首先将页面对象类型序列从一维向量空间投影至二维向量空间;然后进行深度卷积生成对抗网络建模;训练网络参数并使用训练好的网络模型生成对象类型序列;根据网络生成的对象类型序列生成文档对象内容;最终将文档转换成文档图像,生成文档图像集。基于深度学习框架的卷积生成对抗网络自动生成图像文档,使用对抗网络中的判别网络学习现存的文档图像,用对抗网络中的生成网络自动生成新的文档图像,从而得到文档图像集;由于采用现存的文档图像训练网路参数,生成的文档图像更贴近出版物,且与人工标注相比能够自动生成文档图像集及标注信息,节省时间和人力成本,避免由于人工标注带来的无效标注。



1. 基于深度学习生成文档图像集的方法,其特征在于,包括以下步骤:

步骤A、向量空间投影建模:将文档图像页面中的对象视为一个序列,序列中每一个节点对应一个对象的类型,得到文档对象序列和与其一一对应的对象类型序列,并将对象类型序列重排得到其对应的二维矩阵,从而将对象类型序列从一维向量空间投影至二维向量空间;

步骤B、深度卷积生成对抗网络建模:所述对抗网络包含判别网络和生成网络;判别网络采用已有文档图像进行训练,其作用在于训练生成网络;生成网络训练好后,训练后的生成网络用以生成二维矩阵,旨在后续自动生成文档图像集;

步骤C、训练网络模型参数:对步骤B构建的对抗网络进行训练并求解网络参数;将现有文档图像中的文档对象类型序列重排为二维矩阵,用于训练判别网络;并使用训练好的判别网络对生成网络进行训练;

步骤D、生成对象类型序列:基于训练好的生成网络自动输出新的二维矩阵;然后,将该新的二维矩阵投影至一维向量空间,得到新的文档对象类型序列;

步骤E、生成文档对象内容:采集各种文档对象数据,并根据步骤D生成新的文档对象类型序列,自动生成文档对象的具体内容;

步骤F、将步骤E生成的文档转换成文档图像,生成文档图像集,所述文档图像集包含文档图像、文档对象坐标信息和文档对象具体内容。

2. 根据权利要求1所述的基于深度学习生成文档图像集的方法,其特征在于:所述步骤A中,所述对象的类型包括页眉、文本、图、图注、表格、公式、页码和页脚;

(1) 将文档图像页面中的若干个对象定义为文档对象序列,即:

$$DO_i, i=1, 2, 3, \dots, N \quad (1)$$

其中, DO_i 表示第 i 个文档对象; N 表示文档对象的数量;

并将与文档对象序列对应的类型序列定义为对象类型序列,即:

$$y_i, i=1, 2, 3, \dots, N \quad (2)$$

$$y_i \in \{Type_j | j=1, 2, 3, \dots, M\} \quad (3)$$

其中, y_i 表示第 i 个文档对象对应的类型, M 表示对象类型的数量, $Type_j$ 表示类型;

(2) 将每一页文档图像页面中的文档对象序列视为一个向量,将公式(1)和公式(2)表示为向量形式:

$$DO = [DO_1, DO_2, DO_3, \dots, DO_N] \quad (4)$$

$$Y = [y_1, y_2, y_3, \dots, y_N]; \quad (5)$$

(3) 设有 p 页文档图像,将第 p 页的文档对象序列和对象类型序列分别表示成向量形式:

$$DO^p = [DO_1^p, DO_2^p, DO_3^p, \dots, DO_{N_p}^p] \quad (6)$$

$$Y^p = [y_1^p, y_2^p, y_3^p, \dots, y_{N_p}^p]; \quad (7)$$

其中,上标 p 表示第 p 页,下标 N_p 表示第 p 页中文档对象的数量,第 p 页的第 i 个对象的类型为 y_i^p , $1 \leq i \leq N_p$,第 p 页共有 N_p 个文档对象,第 $p-1$ 页共有 $N(p-1)$ 个文档对象;

(4) 将 $1 \sim p$ 页的对象类型序列按照页码顺序排列, y_i^p 在整个序列中的位置为:

$$loc_i^p = \sum_{i=1}^{p-1} Ni + i \quad (8)$$

其中 N_i 表示第 i 页中的文档对象个数,将式(8)投影至二维矩阵中, K 表示二维矩阵的行数和列数,行数等于列数,二维矩阵的列坐标为:

$$col_i^p = loc_i^p / K + 1 \quad (9)$$

二维矩阵的行坐标为:

$$row_i^p = loc_i^p \% K \quad (10)$$

进而可以得到:

$$loc_i^p = K \times (col_i^p - 1) + row_i^p \quad (11)$$

(5)对二维矩阵定义如下:

$$A = [a_{k1, k2}]_{K \times K} \quad (13)$$

其中, $1 \leq k1, k2 \leq K$,根据式(9)~(11)将式(13)中元素与第 p 页的第 i 个对象的类型 y_i^p 建立一一对应的关系。

3.根据权利要求2所述的基于深度学习生成文档图像集的方法,其特征在于:所述步骤C中,在对对抗网络进行训练时,具体采用以下方式:

(1)利用KL散度定义网络的损失函数:

$$Loss = \frac{1}{NS} \sum_{i=1}^{NS} [\log(D(\mathbf{para-d}, \mathbf{A}^i)) + \log(1 - G(\mathbf{para-g}, \mathbf{P}^i))] \quad (16)$$

其中, NS 表示样本数量, $1 \leq i \leq NS$; \mathbf{A}^i 表示式(13)的第 i 个样本点; \mathbf{P}^i 表示生成网络输入端输入的随机向量的第 i 个样本点;

(2)基于步骤A得到的二维矩阵作为输入训练网络,采用梯度下降法求解网络参数,梯度函数为:

$$\frac{\partial Loss}{\partial \mathbf{para-d} \partial \mathbf{para-g}} = \nabla_{\mathbf{para-d}, \mathbf{para-g}} \frac{1}{NS} \sum_{i=1}^{NS} [\log(D(\mathbf{para-d}, \mathbf{A}^i)) + \log(1 - G(\mathbf{para-g}, \mathbf{P}^i))] \quad (19)$$

其中, $D(\mathbf{para-d})$ 为判别网络结构, $\mathbf{para-d}$ 为判别网络参数; $G(\mathbf{para-g})$ 为生成网络结构, $\mathbf{para-g}$ 为生成网络参数;

在训练过程中,首先用二维矩阵训练判别网络,并使用训练后的判别网络训练生成网络。

4.根据权利要求1所述的基于深度学习生成文档图像集的方法,其特征在于:所述步骤B中,所述判别网络包括从左至右依次连接的四组卷积核和一个全连接层,四组卷积核使用的激活函数均为ReLU;所述生成网络包括从左至右依次连接的全连接层和四组卷积核,四组卷积核使用的激活函数分别为:ReLU、ReLU、ReLU和Tanh。

5.根据权利要求2所述的基于深度学习生成文档图像集的方法,其特征在于:所述步骤A中:

$$\sum_{i=1}^{p-1} Ni \gg K \times K \quad (12)$$

K表示二维矩阵的行数和列数,行数等于列数,即选取足够多的文档页面,构建数量足够的二维矩阵用于后续建模分析。

基于深度学习生成文档图像集的方法

技术领域

[0001] 本发明涉及一种图像生成方法,属于图像数据集自动生成领域,具体涉及基于深度学习生成文档图像集的方法。

背景技术

[0002] 在文档图像处理的诸多领域中,如分割、分类、检索等领域,带标记的文档图像集是机器学习过程中不可或缺的数据基础。随着大数据时代的到来,“端到端”的深度学习在人工智能研究领域中成了重要研究方法,与传统的机器学习相比,深度学习需要更多训练数据。

[0003] 目前,研究人员为了更高效地获取包含文档图像及标注信息的图像集,采用了一些图像集自动生成方法。如2017年文档分析与识别国际会议(International Conference on Document Analysis and Recognition,ICDAR)上的论文(D.He,S.Cohen,B.Price,D.Kifer and C.L.Giles,“Multi-Scale Multi-Task FCN for Semantic Page Segmentation and Table Detection”)中将段落、图、表格、标题、段落标题、列表等等元素进行随机排列生成文档图像数据集,用于深度学习训练。同样,申请公布号为【CN 108898188 A】的发明专利也公开一种图像数据集辅助标记系统及方法,利用神经网络训练的思想对神经网络训练所需的图像进行初步特征提取训练,对图像进行识别标记获得神经网络所需的标签文档格式,在大量的图像信息中获得某一类的标签文档。

[0004] 另一方面,很多图像集仍然采用人工标注的方法制作,例如:牛津大学机器人研究组(Robotics Research Group)设计的图像标注工具VIA(“Abhishek Dutta and Andrew Zisserman.2019.The VIA Annotation Software for Images,Audio and Video.In Proceedings of the 27th ACM International Conference on Multimedia(MM’19),October 21-25,2019,Nice,France.ACM,New York,NY,USA.”,使用VIA工具可以使用不同形状(矩形、圆、椭圆、多边形,等等)对图像区域进行手工标注。

[0005] 对于人工标注而言,虽然其具有很强灵活性,标注过程中可以弹性更改标注策略,标注结果能够较好地契合预期,但是,其缺点也是显然的,即标注过程费时、人力成本高昂,而且标注质量与标注人员的熟练程度成正比;相对于人工标注,文档图像数据集自动生成方法可以较好地克服人工标注的不足,但是也存在不可避免的问题,比如,出版业具有自身的行业规范,不同出版物的版面设计也遵循特定的规律,通过这些规律更好地展示文档内容,若随机生成的文档图像不能很好地契合出版物的排版规律,使得训练出来的模型应用于真实出版物文档图像时,不能体现模型的最佳性能。

发明内容

[0006] 本发明针对现有获得文档图像集方法所存在的缺陷,提出基于深度学习生成文档图像集的方法,采用深度学习框架的卷积生成对抗网络自动生成图像文档,使用对抗网络中的判别网络学习现存的文档图像,然后用对抗网络中的生成网络自动生成新的文档图

像,从而得到文档图像集。

[0007] 本发明是采用以下的技术方案实现的:基于深度学习生成文档图像集的方法,包括以下步骤:

[0008] 步骤A、向量空间投影建模:将文档图像页面中的对象视为一个序列,序列中每一个节点对应一个对象的类型,得到文档对象序列和与其一一对应的对象类型序列,并将对象类型序列重排得到其对应的二维矩阵,从而将对象类型序列从一维向量空间投影至二维向量空间;

[0009] 步骤B、深度卷积生成对抗网络建模:所述对抗网络包含判别网络和生成网络;判别网络采用已有文档图像进行训练,其作用在于训练生成网络;生成网络训练好后,训练后的生成网络用以生成二维矩阵,旨在后续自动生成文档图像集;

[0010] 步骤C、训练网络模型参数:对步骤B构建的对抗网络进行训练并求解网络参数;将现有文档图像中的文档对象类型序列重排为二维矩阵,用于训练判别网络;并使用训练好的判别网络对生成网络进行训练;

[0011] 步骤D、生成对象类型序列:基于训练好的生成网络自动输出新的二维矩阵;然后,将该新的二维矩阵投影至一维向量空间,得到新的文档对象类型序列;

[0012] 步骤E、生成文档对象内容:采集各种文档对象数据,并根据步骤D生成新的文档对象类型序列,自动生成文档对象的具体内容;

[0013] 步骤F、将步骤E生成的文档转换成文档图像,生成文档图像集,所述文档图像集包含文档图像、文档对象坐标信息和文档对象具体内容。

[0014] 进一步的,所述步骤A中,所述对象的类型包括页眉、文本、图、图注、表格、公式、页码和页脚;

[0015] (1) 将文档图像页面中的若干个对象定义为文档对象序列,即:

$$[0016] \quad DO_i, i=1, 2, 3, \dots, N \quad (1)$$

[0017] 其中, DO_i 表示第 i 个文档对象; N 表示文档对象的数量;

[0018] 并将与文档对象序列对应的类型序列定义为对象类型序列,即:

$$[0019] \quad y_i, i=1, 2, 3, \dots, N \quad (2)$$

$$[0020] \quad y_i \in \{Type_j | j=1, 2, 3, \dots, M\} \quad (3)$$

[0021] 其中, y_i 表示第 i 个文档对象对应的类型, M 表示对象类型的数量, $Type_j$ 表示类型;

[0022] (2) 将每一页文档图像页面中的文档对象序列视为一个向量,将公式 (1) 和公式 (2) 表示为向量形式:

$$[0023] \quad DO = [DO_1, DO_2, DO_3, \dots, DO_N] \quad (4)$$

$$[0024] \quad Y = [y_1, y_2, y_3, \dots, y_N] \quad (5)$$

[0025] (3) 设有 p 页文档图像,将第 p 页的文档对象序列和对象类型序列分别表示成向量形式:

$$[0026] \quad \mathbf{DO}^p = [DO_1^p, DO_2^p, DO_3^p, \dots, DO_{N_p}^p] \quad (6)$$

$$[0027] \quad \mathbf{Y}^p = [y_1^p, y_2^p, y_3^p, \dots, y_{N_p}^p] \quad (7)$$

[0028] 其中,上标 p 表示第 p 页,下标 N_p 表示第 p 页中文档对象的数量,第 p 页的第 i 个对象的类型为 y_i^p , $1 \leq i \leq N_p$,第 p 页共有 N_p 个文档对象,第 $p-1$ 页共有 $N(p-1)$ 个文档对象;

[0029] (4) 将1~p页的对象类型序列按照页码顺序排列, y_i^p 在整个序列中的位置为:

$$[0030] \quad loc_i^p = \sum_{i=1}^{p-1} Ni + i \quad (8)$$

[0031] 其中 N_i 表示第 i 页中的文档对象个数,将式(8)投影至二维矩阵中, K 表示矩阵的行数和列数,行数等于列数,二维矩阵的列坐标为:

$$[0032] \quad col_i^p = loc_i^p / K + 1 \quad (9)$$

[0033] 二维矩阵的行坐标为:

$$[0034] \quad row_i^p = loc_i^p \% K \quad (10)$$

[0035] 进而可以得到:

$$[0036] \quad loc_i^p = K \times (col_i^p - 1) + row_i^p \quad (11)$$

[0037] (5) 对二维矩阵定义如下:

$$[0038] \quad A = [a_{k1, k2}]_{K \times K} \quad (13)$$

[0039] 其中, $1 \leq k1, k2 \leq K$,根据式(9)~(11)将式(13)中元素与第 p 页的第 i 个对象的类型为 y_i^p 建立一一对应的关系。

[0040] 进一步的,所述步骤C中,在对对抗网络进行训练时,具体采用以下方式:

[0041] (1) 利用KL散度定义网络的损失函数:

$$[0042] \quad Loss = \frac{1}{NS} \sum_{i=1}^{NS} [\log(D(\text{para-d}, A^i)) + \log(1 - G(\text{para-g}, P^i))] \quad (16)$$

[0043] 其中, NS 表示样本数量, $1 \leq i \leq NS$; A^i 表示式(13)的第 i 个样本点; P^i 表示生成网络输入端输入的随机向量的第 i 个样本点;

[0044] (2) 基于步骤A得到的二维矩阵作为输入训练网络,采用梯度下降法求解网络参数,梯度函数为:

$$[0045] \quad \frac{\partial Loss}{\partial \text{para-d} \partial \text{para-g}} = \nabla_{\text{para-d, para-g}} \frac{1}{NS} \sum_{i=1}^{SN} [\log(D(\text{para-d}, A^i)) + \log(1 - G(\text{para-g}, P^i))] \quad (19)$$

[0046] 其中, $D(\text{para-d})$ 为判别网络结构, para-d 为判别网络参数; $G(\text{para-g})$ 为生成网络结构, para-g 为生成网络参数;

[0047] 在训练过程中,首先用二维矩阵训练判别网络,并使用训练后的判别网络训练生成网络。

[0048] 进一步的,所述步骤B中,所述判别网络包括从左至右依次连接的四组卷积核和一个全连接层,四组卷积核使用的激活函数均为ReLU;所述生成网络包括从左至右依次连接的全连接层和四组卷积核,四组卷积核使用的激活函数分别为:ReLU、ReLU、ReLU和Tanh。

[0049] 进一步的,所述步骤A中:

$$[0050] \quad \sum_{i=1}^{p-1} Ni \gg K \times K \quad (12)$$

[0051] 即选取足够多的文档页面,构建数量足够多的二维矩阵用于后续建模分析。

[0052] 与现有技术相比,本发明的优点和积极效果在于:

[0053] 本方案通过自动生成文档图像集及标注信息,节省时间和人力成本,同时避免由于人工标注带来的无效标注;使用深度学习框架的卷积生成对抗网络自动生成文档图像,使用对抗网络中的判别网络学习现存的文档图像,然后用对抗网络中的生成网络自动生成新的文档图像,从而得到文档图像集,成本低、效率高;且由于采用现存的文档图像训练网络参数,生成的文档图像更贴近出版物,具有更好的使用参考价值,另外,生成文档图像集的同时,提供文档图像中文本对象的文字编码信息(如:ASCII、Unicode等等),更好的满足深度学习训练需求。

附图说明

- [0054] 图1为本发明实施例基于深度学习生成文档图像集的方法流程示意图;
[0055] 图2为本发明实施例所述文档对象序列示意图;
[0056] 图3为本发明实施例所述对象类型序列示意图;
[0057] 图4中(a)为文档图像示意图,(b)为“文档对象类型”序列在矩阵中的排列示意图;
[0058] 图5为本发明实施例所述深度卷积判别网络示意图;
[0059] 图6为本发明实施例所述深度卷积生成网络示意图;
[0060] 图7为本发明实施例由“生成网络”生成的文档图像示意图;
[0061] 图8为本发明实施例所述文档图像集结构示意图。

具体实施方式

[0062] 为了能够更加清楚地理解本发明的上述目的、特征和优点,下面结合附图及实施例对本发明做进一步说明。在下面的描述中阐述了很多具体细节以便于充分理解本发明,但是,本发明还可以采用不同于在此描述的方式来实施,因此,本发明并不限于下面公开的具体实施例。

[0063] 本实施例提供了基于深度学习生成文档图像集的方法,如图1所示,包括以下步骤:

[0064] 第一步,向量空间投影建模:将文档图像页面中的对象视为一个序列,序列中每一个节点对应一个对象的类型,得到文档对象序列和与其一一对应的对象类型序列,并将对象类型序列重排得到其对应的二维矩阵,从而将对象类型序列从一维向量空间投影至二维向量空间;

[0065] 第二步,深度卷积生成对抗网络建模:所述对抗网络包含判别网络和生成网络,判别网络采用已有文档图像进行训练,其作用在于训练生成网络;生成网络训练好后,用于生成相应的二维矩阵,旨在后续步骤中自动生成文档图像集;

[0066] 第三步,训练网络模型参数:对第二步构建的对抗网络进行训练并求解网络参数,将现有文档图像中的文档对象类型序列重排为二维矩阵用于网络训练,并以随机向量作为生成网络的输入,使用训练好的判别网络对生成网络进行训练;

[0067] 第四步,生成对象类型序列:基于训练好的生成网络自动输出新的二维矩阵,然后再通过第一步的逆过程将该新的二维矩阵投影至一维对象类别向量,得到新的文档对象类型序列;

[0068] 第五步,生成文档对象内容:采集各种文档对象数据,并基于第四步生成的新的文

档对象类型序列生成文档中对象的具体内容；

[0069] 第六步,将第五步生成的文档转换成文档图像,生成文档图像集,所述文档图像集包含文档图像、文档对象坐标信息和文档对象具体内容。

[0070] 下面结合具体的实施例对本发明方案做详细的介绍,具体的:

[0071] 第一步,向量空间投影建模:

[0072] 如图2和图3的第一列所示,将文档页面中的对象可以看作一个序列,序列中每一个节点(图3第一列)对应一个类型标签(图3第二列),然后将序列投影至二维 $K \times K$ 矩阵空间,如图4所示。

[0073] 图2中,一页文档图像中包含了11个对象,依次是:页眉、文本、图、图、图注、文本、文本、图、图注、文本、页脚;这些对象按照从上往下、从左往右的顺序排列,见图3,既符合阅读顺序,也契合书写和排版顺序,将11个对象定义为文档对象序列:

[0074] $DO_i, i=1, 2, 3, \dots, N$ (1)

[0075] 其中, DO_i 表示第*i*个文档对象,在图2中 $N=11$,11个文档对象为图3的左列“文档对象序列”,与之对应的右列为“对象类型序列”,定义为:

[0076] $y_i, i=1, 2, 3, \dots, N$ (2)

[0077] $y_i \in \{Type_j | j=1, 2, 3, \dots, M\}$ (3)

[0078] 具体来说,在图2和图3中, $M=5$, $Type_1$ 至 $Type_5$ 分别为:页眉、文本、图、图注、页脚;式(1)和(2)表征了一页文档图像中的“文档对象”和“对象类型”序列对,而这种序列对具有以下特性:

[0079] <1>根据“自上而下”、“从左往右”的书写顺序以及排版顺序,式(1)和(2)所表征的序列对很好地反映了这种顺序,而这种顺序反映出了同一页面中“文档对象”之间的“序列关系”;

[0080] <2>任何大于一页的文档或者图书,除了同一页面内的“文档对象”之间具有“序列关系”外,任意两个不在同一页面的“文档对象”也存在“序列关系”,因此,将每一页中对象序列看作一个向量,也即式(1)和(2)均可表示成向量形式:

[0081] $DO = [DO_1, DO_2, DO_3, \dots, DO_N]$ (4)

[0082] $Y = [y_1, y_2, y_3, \dots, y_N]$ (5)

[0083] 式(5)表示了一个页面中所有“文档对象类型”序列构成的一维向量,可以投影至多个页面的“文档对象”序列所构成的二维向量中,即:矩阵。

[0084] 如图4所示,(a)中有三页文档,(b)为一个 $K \times K$ 的矩阵,矩阵中每一个元素都表示一个“文档对象类型”,(a)中三页文档中的所有“文档对象类型”按照从上往下、从左往右的顺序按列依次填充(b)中的 $K \times K$ 矩阵。令第*p*页的“文档对象”序列和“对象类型”序列分别表示成向量形式:

[0085] $DO^p = [DO_1^p, DO_2^p, DO_3^p, \dots, DO_{N_p}^p]$ (6)

[0086] $Y^p = [y_1^p, y_2^p, y_3^p, \dots, y_{N_p}^p]$ (7)

[0087] 其中,上标*p*表示第*p*页,下标 N_p 表示第*p*页中“文档对象”的个数,第*p*页的第*i*个对象的类型为 y_i^p , $1 \leq i \leq N_p$,第*p*页共有 N_p 个文档对象,第*p-1*页共有 $N_{(p-1)}$ 个文档对象;将1~

p页的对象类型序列按照页码顺序排列, y_i^p 在整个序列中的位置为(从1开始计数):

$$[0088] \quad loc_i^p = \sum_{i=1}^{p-1} Ni + i \quad (8)$$

[0089] 其中 Ni 表示第 i 页中的文档对象个数,将式(8)投影至图4(b)所示的 $K \times K$ 矩阵中的列坐标为:

$$[0090] \quad col_i^p = loc_i^p / K + 1 \quad (9)$$

[0091] 即: loc_i^p 除以 K ,商加上1变为列坐标;

[0092] 行坐标为:

$$[0093] \quad row_i^p = loc_i^p \% K \quad (10)$$

[0094] 即: loc_i^p 除以 K ,余数变为行坐标;

[0095] 同时,在 (row_i^p, col_i^p) 确定时,可以计算出:

$$[0096] \quad loc_i^p = K \times (col_i^p - 1) + row_i^p \quad (11)$$

[0097] 式(9)~(10)定义了一维对象类别向量投影至 $K \times K$ 二维矩阵中的坐标变换,而式(11)则是逆变换过程。

[0098] 需要强调的是,本实施例中:

$$[0099] \quad \sum_{i=1}^{p-1} Ni \gg K \times K \quad (12)$$

[0100] 即需要选取足够多的文档页面,构建数量足够多的 $K \times K$ 矩阵,用于后续建模分析。

[0101] 图4(b)所示 $K \times K$ 矩阵做如下定义:

$$[0102] \quad A = [a_{k1, k2}]_{K \times K} \quad (13)$$

[0103] 其中, $1 \leq k1, k2 \leq K$,根据式(9)~(11)可将式(13)中元素与第 p 页的第 i 个对象的类型为 y_i^p 建立一一对应的关系。

[0104] 本实施例中,如图2至图4所示,将文档对象序列刻画为“空间”映射关系,文档版面信息抽象为三个“空间”,即“文档对象”序列空间、“文档对象类型”序列空间以及“文档对象类型”序列按列重拍后得到 $K \times K$ 二维矩阵空间。三个空间之间存在两种映射关系:(1)“文档对象”序列空间 \leftrightarrow “文档对象类型”序列空间;映射关系为:“一一映射”;(2)“文档对象类型”序列空间 \leftrightarrow “ $K \times K$ 二维矩阵空间”;映射关系为:“坐标变换关系”。关系(1)是天然的“一一映射”关系,关系(2)将一维序列向量投影至二维矩阵空间。一方面,便于使用二维矩阵训练“深度卷积生成对抗网络”;另一方面,也便于网络产生新的二维矩阵,用于生成新的“文档对象类型”序列,最终实现文档图像的自动生成。

[0105] 第二步,深度卷积生成对抗网络建模:

[0106] 上一步获取的 $K \times K$ 二维矩阵中包含了多个文档页面对象的类别信息,将第一步的输出 $K \times K$ 二维矩阵作为第二步的输入,采用深度卷积生成对抗网络建模,模型共包含两部分,第一部分如图5所示,称为“判别网络”,用于鉴别输入的 $K \times K$ 二维矩阵是否能表征文档对象类别序列;另一部分如图6所示的“生成网络”,用以生成新的 $K \times K$ 二维矩阵。

[0107] 具体的,本实施例中:

[0108] 一方面, $K \times K$ 二维矩阵可作为判别网络的输入,如图5所示,为判别网络的结构示意图,第一组卷积核为64个 $3 \times 3 \times 1$ 的卷积核,通过该卷积核后得到64个特征图,然后再通

过三组卷积核,最终通过一个全连接层,得到一个输出,用于鉴别输入的 $K \times K$ 矩阵是否能表征了文档对象类别序列。将判别网络定义为:

$$[0109] \quad D(\text{para-d}) \quad (14)$$

其中, $D(\cdot)$ 为判别网络结构,para-d为网络参数。

[0110] 另一方面,如图6所示的生成网络,一个维度为 d 的随机向量,通过一个 $d \times 512$ 的全连接层计算出512个 $K/8 \times K/8$ 二维矩阵,接着通过三组分数卷积核,最终生一个新的 $K \times K$ 矩阵。期望生成的新的 $K \times K$ 矩阵能够很好地表征文档对象类别序列,并且使用图5所示判别网络进行校验,判断是否能够很好地表征文档对象类别序列。生成网络可定义为:

$$[0111] \quad G(\text{para-g}) \quad (15)$$

[0112] 其中,生成网络结构表示为 $G(\cdot)$,网络参数为para-g。

[0113] 图5所示判别网络(式(14))与图6所示生成网络(式(15))协同工作构成了深度卷积生成对抗网络,其中,图6所示生成网络,从左至右四组卷积核后使用的激活函数分别为:ReLU、ReLU、ReLU和Tanh;而图5所示判别网络,从左至右四组卷积核后使用的激活函数均为:ReLU。

[0114] 第三步,训练网络模型参数:

[0115] 利用KL散度(Kullabck-Leibler divergence)定义网络的损失函数,损失函数使得“判别网络”(图5)和“生成网络”(图6)处于博弈状态,因此,两个网络合成被称为“对抗网络”,使用第一步的输出 $K \times K$ 二维矩阵训练网络,采用梯度下降法可求解网络参数。

[0116] 本实施例在构建了深度卷积生成对抗网络之后,需要对其进行训练才能得到最优的网络参数para-d(式(14))和para-g(式(15)),此时,需要足够的多的文档页面(即式(6)~(12)中的 p 足够大),从而才能够得到足够多的矩阵 A (式(13)所示)。

[0117] 具体利用KL散度(Kullabck-Leibler divergence)定义网络的损失函数:

$$[0118] \quad Loss = \frac{1}{NS} \sum_{i=1}^{NS} [\log(D(\text{para-d}, A^i)) + \log(1 - G(\text{para-g}, P^i))] \quad (16)$$

[0119] 其中,NS表示样本数量(the Number of Samples),自然有 $1 \leq i \leq NS$; A^i 表示式(13)的第 i 个样本点:

$$[0120] \quad A^i \quad (17)$$

[0121] P^i 表示图6左端 d 维随机向量的第 i 个样本点:

$$[0122] \quad P^i \quad (18)$$

[0123] 训练网络模型参数para-d和para-g式(14)和(15)的过程,即为求解式(16)最小值的过程,对梯度函数:

$$\frac{\partial Loss}{\partial \text{para-d} \partial \text{para-g}} =$$

[0124]

$$\nabla_{\text{para-d, para-g}} \frac{1}{NS} \sum_{i=1}^{NS} [\log(D(\text{para-d}, A^i)) + \log(1 - G(\text{para-g}, P^i))] \quad (19)$$

[0125] 进行求解,具体算法如下:

算法 1 网络参数求解算法	
[0126]	<p><1>使用 PDF 解析工具, 解析 PDF 文档, 获取 \mathbf{DO}^p (式(6))、\mathbf{Y}^p (式(7)), 并且保证 p 足够大; 本实施例中 $p > 5000$, $\sum Np > 30000$; 式(3)中 $M = 8$, $\{Type_1, Type_2, \dots, Type_8\} = \{\text{文本, 公式, 图, 图注, 表, 表名, 页眉, 页脚}\}$。(注: 按照每一个文档页面至少 6 个文档对象计算)</p> <p><2>根据式(8)~(10)将“文档对象类型”序列由一维向量空间投影至如式(13)所示 $K \times K$ 的二维矩阵空间; 本实施例中, $K = 16$; 式(16)中 $NS > 110 \approx \frac{Np}{K \times K}$。</p> <p><3>在本实施例中, 图 6 中左端随机向量的维度 $d = 50$, 在第二步“深度卷积生成对抗网络建模”以及图 5 和图 6 定义了对抗网络的所有参数。其中, 式(14)中参数 para-d 和式(15)中参数 para-g 分别表征判别网络和生成网络的卷积核。</p> <p><4>#外循环, 1~若干次 #内循环, 若干次 根据<2>和<3>中的样本 \mathbf{A}^i 和 \mathbf{P}^i, 使用随机梯度下降法计算判别网络参数:</p>
[0127]	$\nabla_{\text{para-d, para-g}} \frac{1}{NS} \sum_{i=1}^{SN} [\log(D(\text{para-d}, \mathbf{A}^i)) + \log(1 - G(\text{para-g}, \mathbf{P}^i))]$ <p>#内循环结束 根据<3>的样本 \mathbf{P}^i, 使用随机梯度下降法计算生成网络参数:</p> $\nabla_{\text{para-g}} \frac{1}{NS} \sum_{i=1}^{SN} \log(1 - G(\text{para-g}, \mathbf{P}^i))$ <p>#外循环结束</p>
	<5>算法结束。

[0128] 在步骤二和步骤三中, 采用深度卷积生成对抗网络, 使用对抗网络中的判别网络学习现存的文档图像, 同时, 使用判别网络训练生成网络; 本实施例中, 采用深度卷积生成对抗网络建模, 模型共包含两部分, 第一部分如图5所示, 称为“判别网络”, 用于鉴别输入的 $K \times K$ 矩阵是否能表征文档对象类别序列; 另一部分如图6所示的“生成网络”, 用以生成新的 $K \times K$ 矩阵。

[0129] 使用对抗网络中的判别网络学习现存的文档图像, 然后用对抗网络中的生成网络自动生成新的文档图像, 从而得到文档图像集。由于采用现存的文档图像训练网路参数, 使

得生成的文档图像贴近出版物。在训练网络的过程中,利用KL散度 (Kullabck-Leibler divergence) 定义网络的损失函数,损失函数使得“判别网络”(图5)和“生成网络”(图6)处于博弈状态,因此,两个网络合成被称为“对抗网络”。使用第一步的输出 $K \times K$ 矩阵训练网络,采用梯度下降法求解网络参数。在训练过程中,首先用 $K \times K$ 二维矩阵训练“判别网络”,然后使用“判别网络”训练“生成网络”。

[0130] 第四步,生成对象类型序列:

[0131] 使用如图6所示“生成网络”自动生成新的 $K \times K$ 矩阵,然后再将新的 $K \times K$ 矩阵投影至一维对象类别向量,得到新的“文档对象类型”序列,即根据第三步训练好的生成网络,生成如图6所示网络的输出为一个 $K \times K$ 矩阵(即式(13)),然后根据式(11)将 $K \times K$ 二维矩阵投影至一维对象类别向量,得到式(5)和式(7)所示的“文档对象类型”序列。生成过程的具体算法如下:

算法 2 对象类型序列生成算法	
[0132]	<1>设定图 6 左端 d 维随机变量 \mathbf{P}^i (式 (18)) 的样本数为 100, 其等同于式 (17) 中 \mathbf{A}^i 的样本数; 根据算法 1, $d=50$;
	<2>根据 100 个 d 维随机变量 \mathbf{P}^i (式 (18)) 使用式 (15) 和图 6 所定义的生成网络, 产生 100 个式 (17) 定义的 \mathbf{A}^i , 其中 $1 \leq i \leq 100$;
	<3>将 \mathbf{A}^1 至 \mathbf{A}^{100} 全部按“列”重排, 可以得到 $16 \times 16 \times 100 = 25600$ 个文档对象类型, 其构成了一个“文档对象类型”序列, 如式 (5) 和式 (7) 所定义;
[0133]	<4>算法结束。

[0134] 第五步,生成文档对象内容:

[0135] 首先采集各种文档对象数据,然后使用第四步生成的“文档对象类型”序列生成文档中对象的具体内容,从“算法2”生成一个“文档对象类型”序列,共包含 $16 \times 16 \times 100 = 25600$ 个文档对象类型。根据“算法1”中的设定,25600个可以生成 $\Sigma N_p = 25600$ 个文档对象,共计 $p > 4000$ 个文档页面。接下来,根据25600个“文档对象类型”生成具体的“文档对象”(式(1))。

[0136] 为了生成“文档对象”,从现存的PDF文档中采集如式(3)定义的数据,其中具体参数如“算法1”中叙述: $\{\text{Type}_1, \text{Type}_2, \dots, \text{Type}_8\} = \{\text{文本}, \text{公式}, \text{图}, \text{图注}, \text{表}, \text{表名}, \text{页眉}, \text{页脚}\}$ 。采集的数据定义为:

[0137] $\text{Set}_1, \text{Set}_2, \dots, \text{Set}_8 = \text{文本集}, \text{公式集}, \dots, \text{页脚集}$ (20)

[0138] 接着根据“算法2”生成的“文档对象类型”序列,采用TeX标记语言及式(20)的数据集生成文档对象(如式(1)定义)、坐标信息、内容信息。其中,坐标信息:

[0139] $\text{DO}_i\text{-Coors}$ (21)

[0140] 其为式(1)中 DO_i 的边界框坐标信息,其中, $1 \leq i \leq 25600$ 。另外,内容信息:

[0141] $\text{DO}_i\text{-Content}$ (22)

[0142] 指的是 DO_i 的具体内容,如:文字编码、公式,等等。具体文档对象生成过程如下算法:

算法 3 文档对象生成算法	
	<1>使用 PDF 解析工具,解析 PDF 文档,获取如式 (20) 所定义数据集,根据“算法 2”生成的“文档对象类型”序列,采用 TeX 标记语言生成文档具体内容;
[0143]	<2>首先生成页眉(若页面包含页眉);
	<3>随机生成页面栏目数;
	<4>生成文档第一栏中的文档对象 DO_i (式 (1)) 及对应的对象边界框坐标信息 $DO_i-Coors$ (式 (21))、具体内容 $DO_i-Content$ (式 (22));
	<5>若文档不止一栏,则继续生成第二栏的文档对象,直至最后一栏结束;
[0144]	<6>生成页脚(若页面包含页脚);
	<7>根据 TeX 标记语言,使用 PDF 引擎生成 PDF 文档;
	<6>算法结束。

[0145] 在第四步和第五步中,当深度卷积生成对抗网络训练好后,使用其中的“生成网络”(图6)生成新的 $K \times K$ 矩阵,然后将新的二维 $K \times K$ 矩阵投影至一维对象类别向量,得到新的“文档对象类型”序列,根据新的“文档对象类型”序列,自动生成文档对象,然后转换为文档图像,进而得到一个新的文档图像集,以提高文档图像集的生成效率和质量。

[0146] 第六步,将文档转换成文档图像,生成文档图像集:

[0147] 将第五步生成的文档转换成文档图像,生成文档图像集(包含:文档图像、文档对象坐标信息、文档对象内容信息),根据算法3生成的PDF文档,每一页都转换成文档图像,如图7给出一张自动生成的图像。将每一张生成的文档图像定义为:

[0148] $DocImage_c, c=1, 2, \dots, p$ (23)

[0149] p 表示文档图像数据集的图像数量(根据“算法1”和“第五步”的描述, $p > 4000$),同时将式(21)所表示的文档对象空间坐标映射至文档图像中,得到:

[0150] $DO_i-Coors'$ (24)

[0151] 那么,文档图像数据集可表示为:

[0152] $DocImageSet = \{ele_c\}, c=1, 2, \dots, p$ (25)

[0153] $ele_c = \{DocImage_c, DO_{i,c}-Coors', DO_{i,c}-Content\}$ (26)

[0154] 式(25)定义了文档图像数据集,其中 ele_c 如图8虚线框所示,包含了一张图像中的 N 个文档对象空间坐标信息(式(26)中 $DO_{i,c}-Coors'$),其与每个文档对象具体内容信息一一对应(式(26)中 $DO_{i,c}-Content$)。

[0155] 以上所述,仅是本发明的较佳实施例而已,并非是对本发明作其它形式的限制,任何熟悉本专业的技术人员可能利用上述揭示的技术内容加以变更或改型为等同变化的等效实施例应用于其它领域,但是凡是未脱离本发明技术方案内容,依据本发明的技术实质对以上实施例所作的任何简单修改、等同变化与改型,仍属于本发明技术方案的保护范围。

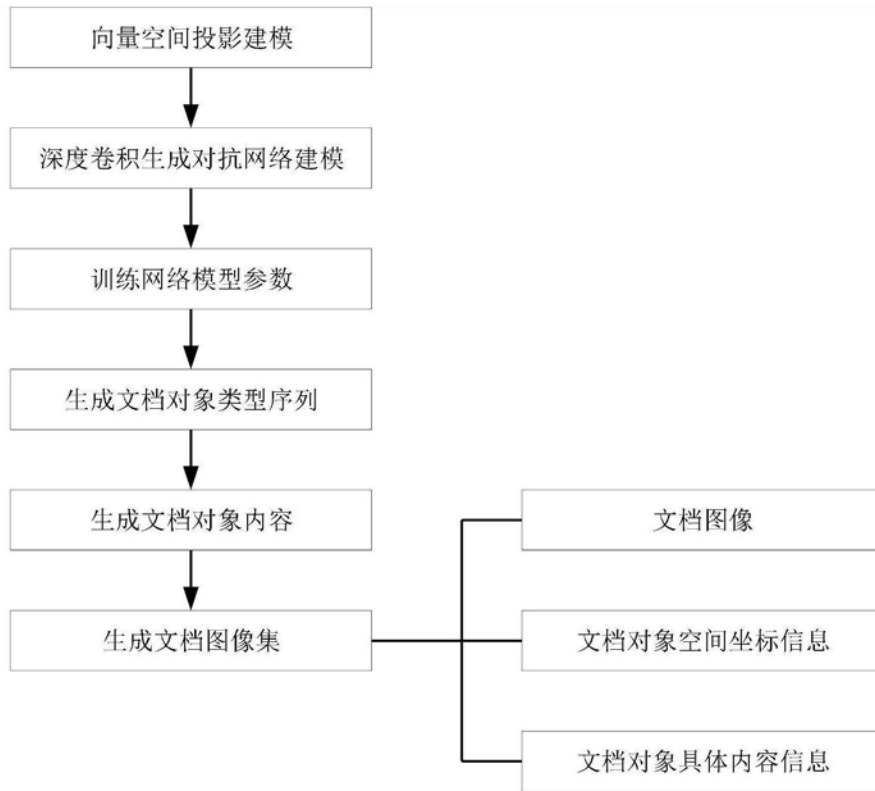


图1

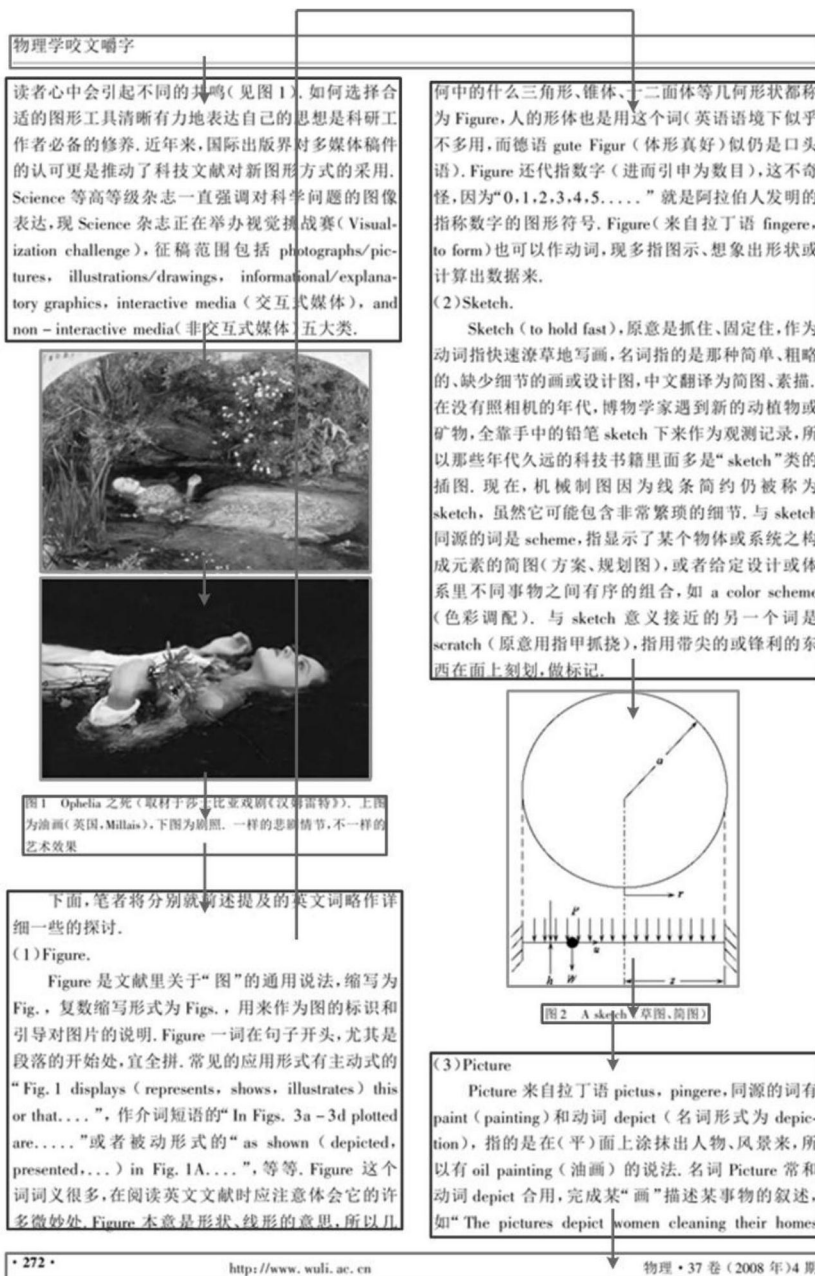


图2

文档对象序列

对象类型序列

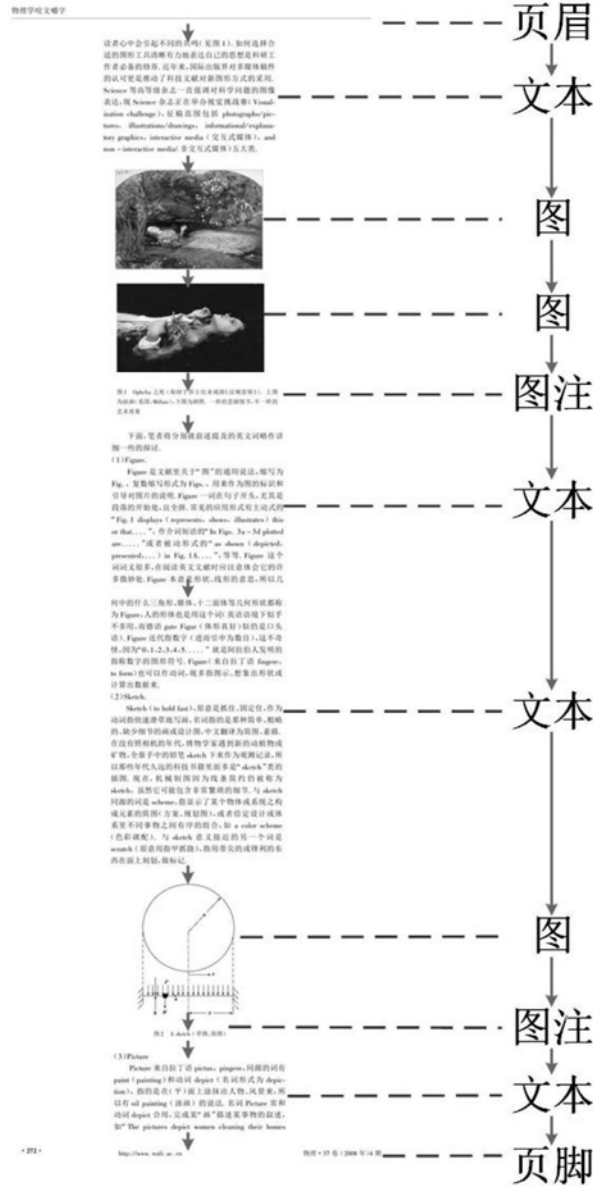


图3

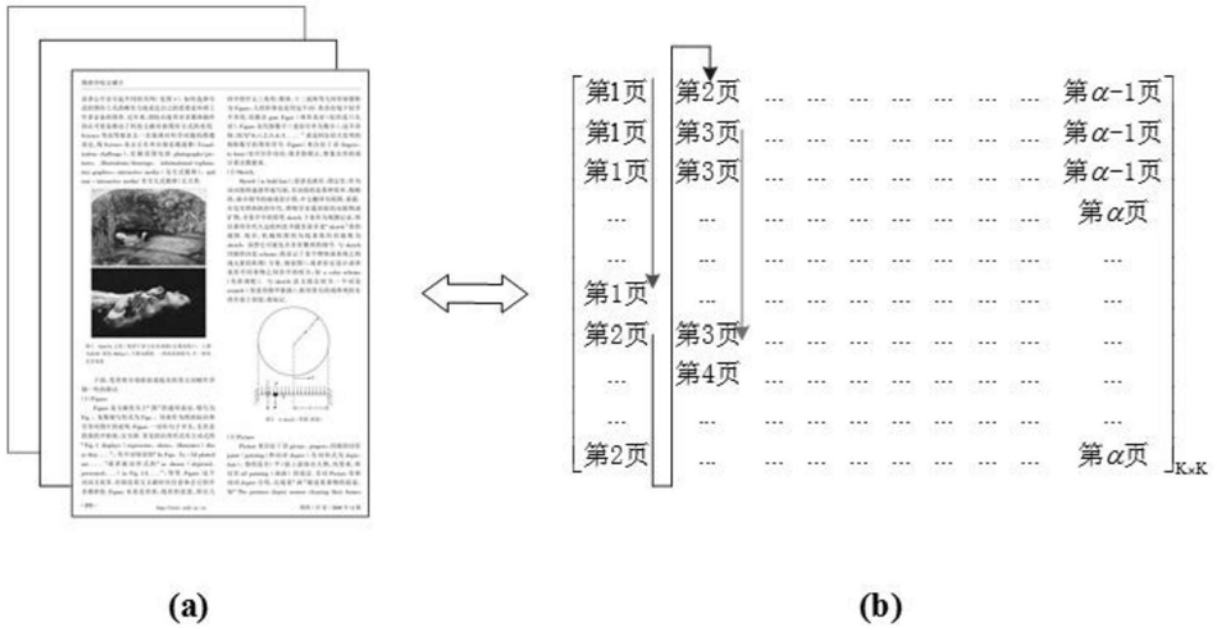


图4

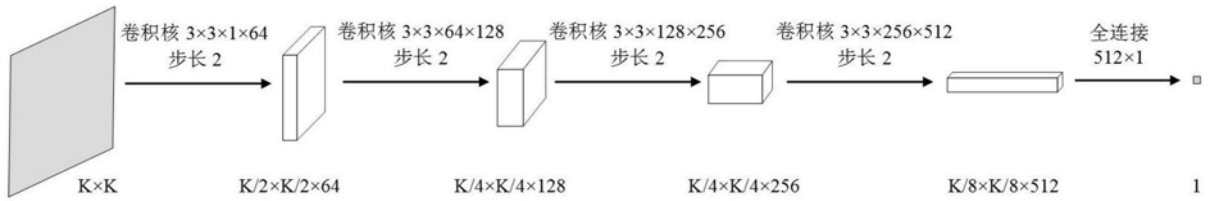


图5

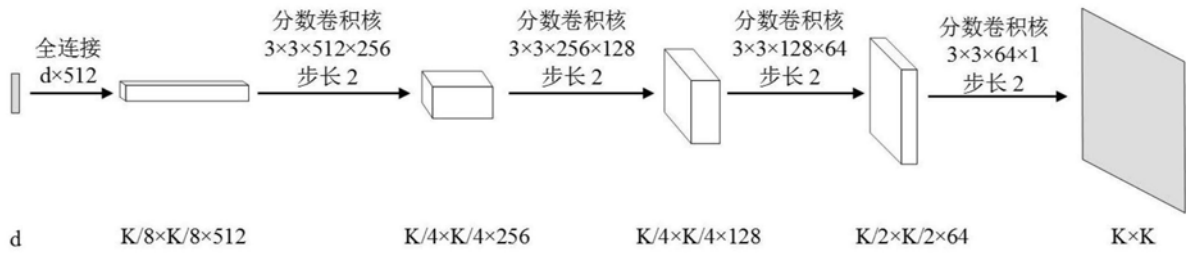
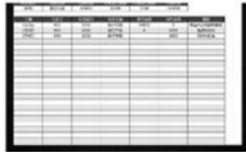
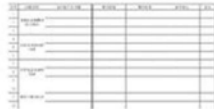


图6




s a gift given only to you. Enjoy it and share it! Enjoy that uniqueness¹. You do not have to pretend in order to seem more like someone else. You do not have to lie to hide the parts of you that are not like what you see in anyone else. Enjoy that uniqueness¹. You do not have to pretend in order to seem more like someone else.

egional sales meetingThe vicepresident of our company delivered a speech that changed my lifeHe asked us,If a genie would grant you three wishes what would they beAfter giving us a moment to write down the three wisheshe then asked us,why do you need a genie ?I would never forget the empowerment



at uniqueness¹. You do not have to pretend in order to seem more like som




ind of understanding to another person. No one can be cheerful and lighthearted³ and joyous⁴ in your way. No one can smile your smile. No one else can bring the whole unique impact of you to another human being. Enjoy that uniqueness¹. You do not have to pretend in order to seem more like someone else. You do not have to lie to h

2 You do not have to pretend

anyone else.¹ You do not have to pretend in order to seem more like so

you that are not like what you see in anyone else. Enjoy that uniqueness¹. You do not have to pretend in order to seem more like someone else. You do not have to lie to hide the parts of you that are not like what you see in anyone else. Enjoy that uniqueness¹. You do not have to pretend in order to seem more like someone else.




2.1 g my dream of motivating othersAfter a

2.1.1 ou that are not like what you

- (a) why do you need a genie I would never forget the empowerment I felt at that moment Enjoy that uniqueness¹ Yo
- (b) u do not have to pretend in order to seem more like someone else You do not have to lie to hide the parts of you
- (c) that are not like what you see in anyone else Enjoy that uniqueness¹ You do not have to pretend in o
- (d) rder to seem more like someone else You do not have to lie to hide the parts of you t

in anyone else. Enjoy that uniqueness



1

图7

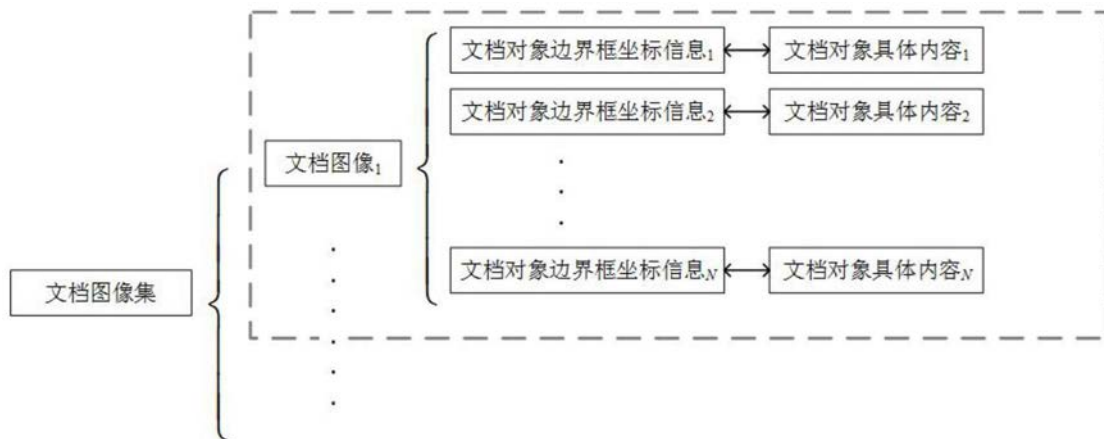


图8