计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200100108

基于深度卷积神经网络的公式重复检测方法

陈 昂[1] 佟 威[1] 周宇强[2] 阴 钰[2] 刘 淇[2]
1 教育部考试中心 北京 100084
2 中国科学技术大学计算机科学与技术学院 合肥 230026
(chena@mail.neea.edu.cn)

摘 要 近年来,随着教育智能化的发展,互联网教育模式成为了教育教学的重要载体。各类在线教育系统拥有海量试题资源,为学习者提供了便捷的学习途径。然而,试题来源繁多、收集方式不统一等因素,使得互联网中所积累的试题资源存在重复率高、质量较低的现象。因此,准确、高效地监测试题,是精炼网络资源、提高网络试题质量的重要方式。在这样的背景下,文中着重研究了针对理科试题资源中图片公式的重复检测问题,通过精准的公式识别检测,能够排除试题语义的干扰,进而加强试题资源监测。传统的公式重复检测方法,往往因为基于人工定义的各类规则,识别步骤繁琐,准确率和效率较低,难以应用于大规模的公式数据检测。据此,提出一种基于深度卷积神经网络的公式重复检测方法。首先,使用一种多通道卷积机制实现了公式图片特征提取和处理的自动化,使之适用于大规模的公式数据检测。然后,使用端到端的输出模式,避免了传统方法中间步骤过多可能导致误差累计的弊端。最后,为了验证模型的准确率以及实用性,在标准测试数据集以及模拟扫描图噪声的数据集上进行了充分的实验,实验结果表明此方法能够有效处理不同质量的公式图片,在检测精度和效率上取得了良好的结果。

关键词:试题质量;公式重复检测;图片识别;卷积神经网络
中图法分类号 TP301

Duplicate Formula Detection Based on Deep Convolutional Neural Network

CHEN Ang[1],TONG Wei[1],ZHOU Yu-qiang[2],YIN Yu[2] and LIU Qi[2]
1 National Education Examinations Authority,Beijing 100084,China
2 School of Computer Science and Technology,University of Science and Technology of China,Hefei 230026,China

Abstract In recent years,with the development of educational intelligence,the Internet education model has become an important carrier of education and teaching. Various online education systems provide learners with a convenient way to learn their vast amount of test resources. However,the accumulated exercise resources suffer from the high repetition rate and low quality due to various sources of test questions and inconsistent collection methods. Therefore,how to accurately and efficiently monitor test questions is an important way to refine network resources and improve the quality of network test questions. In this context,this paper focuses on the problem of repeated detection of picture formulas in science test resources. Through accurate formula recognition detection,it can eliminate the interference of test questions semantics,and then improve the test resource monitoring. In response to this problem,the traditional formula repeat detection method is often based on manually defined rules and difficult to apply to large-scale formula data detection because of cumbersome identification steps,low accuracy and low efficiency. Based on this,this paper proposes a formula repeated detection method based on deep convolutional neural network. Firstly,a multi-channel convolution mechanism is used to automate the extraction and processing of formula picture features,making it suitable for large-scale formula data detection. Then,using the end-to-end output mode,the accumulation of errors that may be caused by too many intermediate steps in the traditional method is avoided. Finally,in order to verify the accuracy and practicability of the model,this paper has carried out sufficient experiments on the standard test data set and the data set of the simulated scan noise. The experimental results show that this method can effectively process the formula pictures of different quality. Good results in both accuracy and efficiency.

Keywords Exercise quality,Duplicate formula detection,Image recognition,Convolutional neural network

1 引言

随着教育信息化的持续深入以及互联网的迅猛发展,当前形成了许多在线学习模式,如 MOOC[1],ITS[2]等。这些在线学习平台积累了大量的网络学习资源,形成了题库,能为学生提供充足的试题练习。然而,在线平台在实际应用中也存

基金项目:国家教育考试科研规划 2017 年度立项课题-中国特色教育考试国家题库应用研究(GJK2017008);国家自然科学基金项目(61672483)
This work was supported by the Foundation of National Education Examiantion Authorty for Scientific Research Plan of National Education Examination (GJK2017008) and National Natural Science Foundation of China (61672483).
通信作者:周宇强(zyq1998@mail.ustc.edu.cn)

在一些问题。(1)虽然试题资源数量庞大,但其中存在着大量重复或者相似的试题,这使得试题的质量无法得到保障,学生还可能浪费精力在一些重复的试题上。(2)网上试题种类丰富,学生获取方便,这对考试的命题形成了极大的挑战。因此,对试题进行重复检测,对于提高题库的质量和保障国家考试公平性都具有重要的意义。

一般来说,试题中的文本表述形式多样,难以通过简洁且统一的方式提取出文本之间的联系[3]。而绝大部分理科试题含有大量公式,并且试题中的公式往往反映了试题中主要考查的知识点,包含较多关键信息[4]。因此,对试题中公式的重复检测,成为了试题重复检测中的关键因素。

目前,传统的公式重复检测方法大致可以分为以下 3 类。(1)对公式图片进行识别、提取、匹配[5],将公式作为许多部分组成的集合,通过人为定义的特征,分别识别其中的字符,最后核对识别得到的数学公式是否相同。此过程中仅依据字符的形状等浅层特征,无法获得深层语义信息,对形态相似而语义有区别的符号(如减号、上划线、破折号等)误识率较高,且该方法需要人工构造大量特征,人力成本较高。(2)通过 OCR 的方法[6]将印刷体数学公式识别出来后再进行比较。然而,数学公式本身的结构与普通文本有所不同,且公式排版时往往比文本的符号密度更大,在划分单个符号所占的面积时,对粘连符号的切分会影响到符号识别,进而影响整个公式的识别以及公式之间的比较,因此 OCR 对数学公式的识别效果远不及文字识别[6]。(3)还有一些学者针对特定格式的数学公式进行重复检测[7],对 MathML 和 Latex 等格式的数学公式,通过源代码到树的转换,从语义的角度解析公式,并得到公式之间的相似度。但这种方式要求得到数学公式的源代码,而在试题的获取中,普遍能够得到的数据是图片形式或文本、图片混合形式的试题,不具备获取源代码的条件。因此,该方式在试题重复检测上不具有推广优势。

总结起来,传统的公式重复检测方法主要存在以下问题:(1)对图片进行识别的过程需要人工设计大量特征,这是一个费时费力的过程,且人工设计的特征容易存在疏漏;(2)没有摆脱先识别图片中的公式再进行重复检测的模式,这种多步骤(非端到端)的模式使得子步骤的误差不断积累,最终导致更大的误差。总的来说,目前公式重复检测任务主要存在以下 4 点挑战。(1)如何自动提取公式图片的特征[8]以用于重复检测任务。(2)如何减小或消除公式图片识别过程所产生的误差。(3)任务的难度与模型结果的好坏会根据不同的公式"重复"的定义而变化。从实际应用角度来说,我们通常认为,不只有完全相同的公式才是重复的,一些仅仅变量的字母表示不同、变量的系数不同或者常量不同的两个公式仍可能是重复的,比如本文定义 $x^2-2\sqrt{3}x+2=0$ 与 $y^2-y+3=1$ 是两个重复的公式。(4)公式图片往往夹带复杂的噪声。如图 1 所示,在实际场景中,公式图片多来自于光学扫描设备对纸质资料进行扫描而产生的扫描图片,这种扫描公式图片会夹带多种类型的噪声(见图 2)。

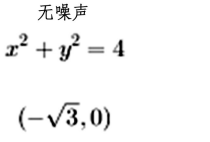

图 1 无噪声图片

Fig. 1 Pictures without noise

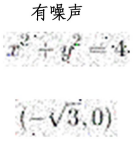

图 2 有噪声图片

Fig. 2 Pictures with noise

为解决以上问题,本文设计了适用于该任务的任务框架,并提出一种基于深度卷积神经网络的试题重复检测模型 MCFIRD(Multi-Channel Formula Image Repeat Detection model)。具体来说,首先,对公式图片的标签数据进行预处理,通过制定一些"重复转换"的规则,使得一些仅仅变量的系数、变量名或常量不同的公式图片标签在经过转换后是相同的。然后,对于给定的两张相同大小公式图片,本文将它们叠加成多通道图片后输入深度卷积神经网络中。实验证明,在网络深度适当的前提下,这种能提取两张图片联合信息的输入方式使模型不易受到噪声的干扰,增强了模型的抗噪能力。最后,模型将通过深度卷积神经网络对输入的叠加多通道图片进行自动的特征提取,避免了人工提取特征的过程,使得提取大量不同的特征成为可能,因此有助于获得图片的语义信息[9]。

实验证明,MCFIRD 能很好地检测经过"重复转换"规则后被认为是重复的公式图片。最后模型直接输出检测结果,这种端到端[10]的输出方式避免了传统方法的中间步骤带来的误差积累(比如传统方法先将图片中的公式识别出来后再进行比较,识别的过程会带来很大误差),有效地提高了检测的准确性。

2 相关工作

2.1 基于符号特征的方法

传统的公式重复检测方法中较为普遍的是基于符号特征的方法,其针对电子文档中出现的公式图片进行预处理。其中,分支定界文本行识别方法[11]通过分析独立行公式中几何布局特征、字符特征以及上下文特征,基于规则和学习对独立公式进行检测;基于文本行的数学识别方法则主要通过分析公式的空间结构,如行间距、行高、行宽等,来判断数学公式所属种类;Li 等提出的基于基线结构公式识别方法[12]分为预处理、分割、识别、确定空间关系、语义搭建、确定逻辑关系 6 个步骤,通过后处理解决字符之间的分割与粘连问题;Zanibbi 等对该方法进行了改进[13]。

2.2 基于公式标记语言的方法

另一种传统方法是基于 MathML 标记语言的公式识别算法[14],其根据语言的结构对数学公式进行拆分,得到不可分的公式元素,由于公式元素的出现次序具有较为清晰的层次结构,因此根据嵌套的布局即可生成对应数学公式标记语言的树形结构。每个树节点中存储的符号以及语义含义如表 1 所列。

表 1 树节点的数据结构

Table 1 Data structure of tree nodes

成员	含义
文本域	公式元素,标记名称
属性域	标记,操作符,变量,常量,边界或分隔符

根据 MathML 标记建立树表示的主要流程,是顺序读取数学公式编码,利用栈标记依次建立树。例如,数学公式 $f=\frac{a+\sqrt{b+c}}{2b}$ 对应的树形结构如图 3 所示。

由公式的树形结构可以生成先序遍历、后序遍历序列,利用 KMP 模式匹配算法进行比较,即可得到两个公式之间是否匹配的结果。

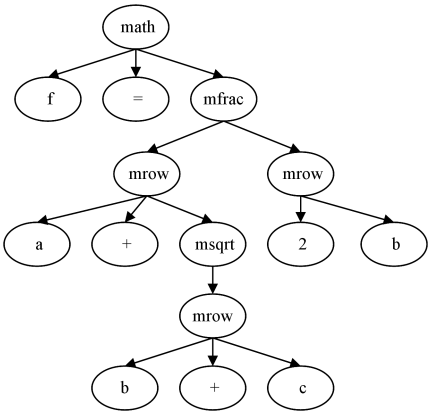

图 3　由 MathML 标记生成的树

Fig. 3　Tree from MathML labels

但这种方法局限于得到的公式的编码形式，而且字符串模式匹配算法只能比较完全相同的公式，对于相近公式，即使其只有几个字符不同，也不会被归为重复公式。但一般试题中可能出现改动个别数字的情况，试题的考查内容本质不变，因此该模型不具备一定的模糊能力，无法灵活应对重题之间的微小差异。

2.3　基于神经网络的重复检测方法

神经网络能够克服传统符号识别方法手工提取特征的不足，也能够在一定程度上对相似公式进行模糊识别。有学者提出了基于组合特征向量[15]和 BP 神经网络[16]的符号识别方法，该方法能自动识别出不同字体、不同大小的同一字符，将生成的特征向量输入 BP 神经网络中进行识别。特征向量的生成主要基于主成分分析、奇异值分解等传统的降维方法。神经网络的训练则通过层与层之间的信号传播和误差的反向传播修改各层权值和阈值，直到达到期望目标。该方法虽然在数学符号识别上得到了较好的识别率，但不能处理实际中可能出现的字符粘连情况，依然没有摆脱单个字符识别后再进行语义比较的模式。

卷积神经网络[17]在图片的特征提取上具有显著的效果。在两张图片的相似度计算任务中，文献[18]使用多层共享参数的卷积神经网络对输入的两张图片独立地进行特征提取，然后计算两个特征向量的距离来衡量两图片的相似度。文献[19]将两张图片叠加成双通道后输入深度卷积神经网络，使得两图片在进行特征提取时信息不再完全独立，取得了很好的效果。

3　深度卷积神经网络公式重复检测

3.1　问题定义

本文使用的数据为数学公式图片及公式图片的 Latex 标签，整个数据集中的公式图片基本涵盖了所有常用数学符号和常规数学表示。表 2 为一条训练数据示例，数据包括公式图片 id、公式图片、公式图片 Latex 标签。

表 2　数据示例

Table 2　Example of data

数据类型	数据示例
公式图片(M_i)	$x^2-2\sqrt{3}x+2=0$
Latex 标签(L_i)	{{x}^{2}}-2\sqrt{3}x+2=0

定义 1　给定数学公式图片集合 M 和公式 Latex 标签集合 R，其中 M 包含完整的数学公式图片，R 包含数学公式图片对应的 Latex 标签，目标是对数学公式图片建模，通过模型的输出判断这两张公式图片是否相似(或重复)。

输入：给定一对含有数学公式的图片 M_a，M_b。

输出：M_a 与 M_b 是否相似(或重复)。

表 3 给出了该问题涉及到的符号和对应的描述。

表 3　问题涉及的符号及解释

Table 3　Symbols and explainations for problem

符号	解释
M	公式图片集合
M_i	第 i 个公式图片
L	公式图片的 Latex 标签集合
L_i	第 i 个公式图片的 Latex 标签
R	转换后的 Latex 标签集合
R_i	第 i 个转换后的 Latex 标签

3.2　公式重复检的测整体框架

本文所研究的公式重复检测任务的整体框架主要包括数据预处理、模型结构以及模型训练 3 部分。

3.2.1　数据预处理

考虑到在实际应用场景中，不只有完全相同的公式才是重复的，一些仅仅变量的字母表示不同、变量的系数不同或者常量不同的两个公式仍可能是重复的。如表 4 所列，本文对公式 Latex 标签集合 L 做以下几点处理：(1)将标签中的“字母变量”转换为字母“a”；(2)将标签中的“数字常量”转换为数字“0”；(3)将标签中的“数字系数 * 字母变量”转换为字母“a”；(4)将标签中的“数字系数/字母变量”转换为字母“a”。经转换后，得到公式 Latex 标签集合 R。

表 4　公式 Latex 标签的转换

Table 4　Rule of transformation for Latex labels

原 Latex 标签L_i	转换后的 Latex 标签R_i
字母变量	a
数字常量	0
数字系数 * 字母变量	a
$\frac{\text{字母变量}}{\text{数字系数}}$	a

具体而言，根据上述规则，表 5 列举了部分公式中数字的转换实例(包括个位数、多位数、根号、分数及其组合形式)。同时，表 6 进一步列举了部分公式 Latex 标签的转换实例。经过上述规则转换后，相同 Latex 标签的公式被视为相似公式，不同的即视为不相似公式。

表 5　公式 Latex 标签转换实例

Table 5　Examples of transformation for Latex labels

原 Latex 标签L_i	转换后的 Latex 标签R_i
xyz	aaa
$13\sqrt{\frac{1}{3}}$	0
$13*xyz$	aaa
$\frac{xyz}{13}$	aaa

表 6　更多实例

Table 6　More examples

原 Latex 标签L_i	转换后的 Latex 标签R_i
\frac{2\sqrt{6}}{3}	0
{{x}^{2}}-2\sqrt{3}x+2=0	{{a}^{2}}-a+0=0
y=\cos\frac{x}{2}	a=\cosa
x{{\log}_{4}}3=\frac{1}{2}	a{{\log}_{0}}0=0

3.2.2 模型结构

本文输入叠加成多通道的两张公式图片，通过深层卷积神经网络对输入图片进行特征提取，最后根据提取到的图片特征预测两张图片的重复性。

MCFIRD模型结构如图4所示，模型包括输入层、卷积层、两层Stack层、Max Pooling层和输出层。

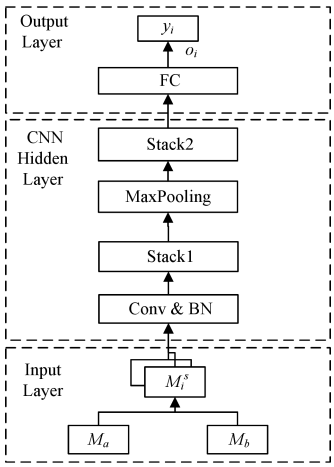

图4 模型结构

Fig. 4 Modle structure

输入层接受两张尺寸相同的公式图片$M_a \in R^{h\times w\times c}$，$M_b \in R^{h\times w\times c}$，其中$h$，$w$，$c$分别表示输入图片在高、宽上的像素个数以及图片的通道数目。由于实际场景中需要检测的有可能是带有复杂混合噪声的公式图片(如扫描图片)，因此模型需要在噪声图片上也有较好的检测效果。在图片相似度计算任务中，文献[18]将两张图片分别输入两个参数共享的神经网络中，这时两张图片的特征提取过程是相互独立的。这种方式训练出来的模型在测试数据与训练数据的分布差异不大时能取得较好的检测效果；但当测试数据与训练数据存在较大的差异(如待检测的是带复杂噪声的公式图片)时，由于输入的两张公式图片的特征提取过程是相互独立的，而图片上的噪声将很大程度地干扰神经网络的特征提取过程，导致检测的误差较大。

文献[19]使用两张图片叠加成双通道的输入形式，使得模型能提取到两张图片的联合信息。受其启发，本文将输入的两张公式图片叠加成多通道后输入网络中，即将M_a和M_b在通道上的叠加$M_i^s \in R^{h\times w\times 2c}$作为网络输入。相比于上面两张图片独立输入的方式，这种输入方法可以使网络提取到两张图片的联合信息，因此即使是带噪声的公式图片，网络依然能提取到两张图片的联合特征，降低了噪声的影响。在网络深度适当的前提下，这种提取两张图片联合信息的方式不易受到噪声的干扰，这有助于提高模型的抗噪能力和泛化能力。本文将通过实验验证此种输入方式的有效性。

Stack层是由多个由卷积层-BN层-ReLU层为组成部分堆积而成的层。其中，BN(Batch Normalization)层[20]可以加快深度网络的训练速度[21]，$ReLU(x)=\max(0,x)$为非线性激活函数。

输出层由一层输出维度为1的全连接层构成，计算方法为：

$$o_i = W_o h_i + b_o \tag{1}$$

其中，$h_i \in R^{d\times 1}$，表示模型从输入图片中提取得到的特征向量；$W_o \in R^{1\times d}$，$b_o \in R^1$为全连接层参数。

在对图片进行检测时，对模型输出o_i做如下变换，得到最终输出：

$$y_i = \begin{cases} 1, & o_i \geqslant bias \\ 0, & o_i < bias \end{cases} \tag{2}$$

其中，$bias$为人为设定的阈值。最终输出为1表示两张公式图片M_a，M_b重复；输出为0，表示不重复。

3.2.3 模型训练

MCFIRD模型由下而上的具体配置如表7所列。

表7 MCFIRD模型的配置

Table 7 Settings of MCFIRD model

类型	配置
FC	#dout:1
Stack2	#component:3;#maps:192;k:2×3;s:1
MaxPooling	Window:1×2;s:1×2
Stack1	#component:3;#maps:96;k:3×3;s:1
BatchNormalization	—
Convolution	#maps:96;k:5×5;s:3
Input	h:32;w:64;c:4

其中，第一行为模型的顶层；h，w，c分别表示图片的高、宽、通道数；k，s分别表示卷积核的kernelsize，stride size；#component表示Stack层的组成部件的数量；#dout表示全连接层的输出维数。

Stack层由下而上的具体配置如表8所列，其中#maps，k，s分别表示上表中对应输入的#maps，k，s参数值。

表8 Stack层配置

Table 8 Settings of Stack layer

类型	配置
ReLU	—
BatchNormalization	—
Convolution	#maps′:#maps;k′:k;s′:s
ReLU	—
BatchNormalization	—
Convolution	#maps′:#maps;k′:k;s′:s
ReLU	—
BatchNormalization	—
Convolution	#maps′:#maps;k′:k;s′:s

本文使用有监督的方式对模型进行训练，训练时以两张公式图片M_a和M_b叠加成多通道的形式作为输入，两张公式图片的Latex标签R_a和R_b是否重复(重复为1，不重复为0)作为标签，进行有监督学习。使用式(3)作为模型的损失函数：

$$\sum_{i=1}^{N} \max(0, 1 - t_i o_i) \tag{3}$$

其中，$t_i = \begin{cases} 1, & y_i = 1 \\ -1, & y_i = 0 \end{cases}$($y_i \in \{1,0\}$表示公式是否重复)，$o_i$为模型的实际输出。

模型使用Adam进行优化，学习率设置为0.0005，mini-batche大小为128，使用文献[22]中的方法将模型参数初始化为某一均匀分布。

当模型用于测试集的公式重复检测时，本文将$bias$取为-0.99。

4　实验

4.1　数据集

考虑到真实场景中可能存在的影响，为了体现模型的有效性以及泛化能力，实验数据主要分为了两部分。一部分来自智学网，是从现实情境抽取的 321 496 张数学公式图片[1)]，这部分图片没有噪声，用于训练、测试，验证模型的有效性；另外一部分是在这些图片的基础上添加噪声构成的带噪图片集，主要用于测试，验证模型的泛化能力与实用性。各噪声设置的图片集包含 10000 张图片。

噪声集中的噪声类型为高斯噪声和椒盐噪声的混合噪声，主要是为了模拟真实情况中扫描的图片而添加的噪声。

数据集的相关统计信息如表 9 所列。

表 9　数据集的相关统计

Table 9　Statistics of dataset

统计参数	标准图片集	噪声图片集
公式图片数量/张	321 496	10000
噪声比例及类型	—	比例 0.01 椒盐噪声 & 方差 0.01 高斯噪声(均值为 0)
Latex 标签平均长度/字符	17	17

经过整理，图片数据集中包含的公式符号主要包括以下类别及下列类别的组合。

(1)单个字母变量，如字母 A，B，C，R，a，b 等公式中较为常用的字母；(2)不等式与等式，如含未知变量的一次、二次方程，函数不等式等，含有＝，≠等符号；(3)分式、根式；(4)代表几何关系的式子，如包含垂直符号⊥、平行符号∥、角度符号°、三角形△和恒等符号≌的式子；(5)表示区间的公式，含方括号、大括号的公式；(6)集合表示符号∩，∪，∅，∀等。

可以看出，数据集中包含的符号具有各种各样的几何结构，覆盖了大多数初等、中等数学试题中所用到的符号，覆盖面十分广泛。而在公式图片中，还会出现上述类型公式的组合，因此公式的类型较为复杂。

4.2　实验评价指标

使用准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 指标来评价本文方法的性能。

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision=\frac{TP}{TP+FP}$$

$$Recall=\frac{TP}{TP+FN}$$

$$F1=\frac{2*Precision*Recall}{Precision+Recall}$$

其中，TP 指网络预测为相同公式，实际上也为相同公式的图片对数量；FP 指预测为相同公式，实际上不同的图片对数量；TN 指预测为不同公式，实际上不同的图片对数目；FN 指预测为不同公式，实际上相同的图片对数目。

$Accuracy$ 代表所有样本下，方法分对样本的比率；$Prediction$ 代表分为相同样本中，实际相同样本的比率；$Recall$ 代表实际相同样本中被分为相同样本的比率；$F1$ 是对 P 和 R 综合效果的评估。

1) http://home.ustc.edu.cn/~yxonic/stn_dataset.7z

在 Prediction 和 Recall 相差不大的情况下，我们主要对比 Accuracy 与 F1 的效果。

4.3　对比实验

我们将本文方法与以下几种 baseline 预测方法做了对比。

(1)普通的孪生(Siamese)网络[18]：两张图片分别输入，通过两个共享权值的 CNN 网络或同一个 CNN 网络(实验中采用同一个网络)训练的方式，得到两张图片的特征向量，再对两个特征向量进行处理得到结果。

(2)深度双通道(deep 2-channel)网络[19]：两张图片进行相叠处理后作为一个输入进入设计好的 CNN 网络训练，最后通过一个全联通层输出结果。其与 MCFIRD 模型的主要区别在于输入的尺寸(为 64 * 64)，以及 CNN 网络的设置。

(3)残差网络[23]：相比加深的双通道网络，使用了残差神经网络代替了普通的 CNN 网络；二者的主要结构差异在于，残差网络在 CNN 基础上使用了残差块[24]。

下面分别用 Siamese，d2ch，resnet 代指上述 3 种模型。

实验中，几种网络的结构不变，通过改变测试集有无噪声、样品正负比进行了对比实验。训练中并没有加入噪声图片，以验证实验方法的泛化能力。

4.4　实验结果及分析

4.4.1　无噪声公式图片检测结果

本小节将 MCFIRD 模型、d2ch 模型、resnet 模型、Siamese 模型共 4 种模型训练后，分别在无噪声的标准图片测试集上测试，测试集大小为 10000 张，测试条件为测试集正样本数占总样本数的比例。实验目的在于初步验证模型的有效性以及研究测试集样本正负比对结果的影响。实验结果如图 5 所示。

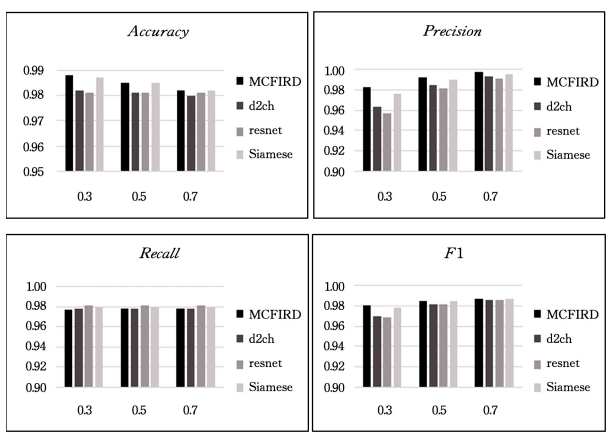

图 5　4 种模型的实验结果

Fig. 5　Experiment results of 4 models

(1)4 个网络在测试没有噪声的标准图片集时，都展现了非常好的效果：对于 10000 张测试图片，准确率 $Accuracy$ 均达到了 98%以上，其他各项数据也都在 95%以上。测试结果说明：通过卷积网络自动提取公式图片特征的方式，是适合处理大规模公式图片数据的；几个模型采用的自动提取特征、端到端输出的方式产生的识别误差很小，在标准图片的测试集上能取得很好的测试效果。模型的有效性得到了初步的验证。

(2)随正负比变化,指标的变化幅度始终很小,这初步说明基于神经网络的模型在有着较好的效果,且有着不错的稳定性。

(3)虽然我们可以看到充分训练的 MCFIRD 模型,相比于其他模型,除了召回率有微弱差距之外;其他评估条件下,在各个正负样本比例的情况下结果都较好。但是,客观来说,在该数据集上各个模型的表现都很不错,没有模型具有决定性的优势。因而我们认为,模型的优劣还需要参考进一步的实验结果。

4.4.2 有噪声公式图片的检测结果

本小节主要将测试集图片换为了带噪声的图片。因为从上一节结果来看,面对无噪声的图片,4 种网络表现得都很好,但是实际中扫描抽取的图片基本没有不带噪声的,于是我们希望讨论更实际的情况,以进一步验证模型的有效性。

同时,考虑到现实噪声的复杂性,我们没有在训练集中加入噪声图片,而是沿用了上一节训练的模型,以验证模型的泛化能力。

测试结果如图 6 所示。

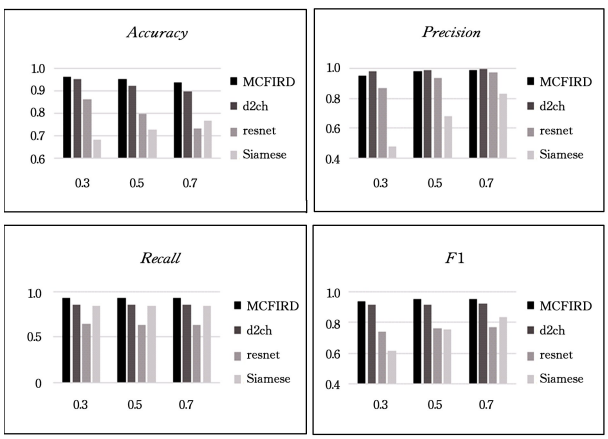

图 6 噪声集上各模型的实验结果

Fig. 6 Experiment results in noise dataset

(1)在具有一定噪声的情况下,由于没有在训练集中加入带噪声的图片,4 个模型的测试结果均低于无噪声的情况。这说明噪声确实会直接影响到模型的效果,噪声集的对比实验设计对衡量模型的实用性是很有必要的。

(2)MCFIRD 模型、d2ch 模型在测试中有着决定性的优势,4 种指标的结果都要明显好于 resnet 和 siamese 模型,并且 *Accuracy* 和 *F*1 都达到了 0.9 以上。这说明:相对于 Siamese 模型使用的图片分别进入模型提取特征,最后处理特征的方式,MCFIRD 和 d2ch 模型使用的多通道卷积的方式具有更强的抗噪声能力;在实验中,MCFIRD 和 d2ch 模型提取特征使用的深度较低的 CNN 网络,相较于 resnet 模型中使用了残差块的网络深度较高的残差网络,具有更强的抗噪声能力。总的来说,MCFIRD 和 d2ch 模型确实具有较优秀的泛化能力以及实用性,并且相对于 resnet 和 Siamese 模型有着比较明显的优势,能够更加适应图片重复检测的真实情景。

(3)在进一步的对比中,MCFIRD 模型的主要结果(准确率以及 F1 值)又都优于 d2ch 模型。我们可以得出,与 d2ch 模型的 64 * 64 的正方形输入相比,MCFIRD 模型 32 * 64 的输入不但减少了数据维度,而且获得了更好的综合效果。真实情景下,在准确率和实用性方面,MCFIRD 模型还是更佳的选择;在面对大规模的公式图片数据时,MCFIRD 模型能有更好的表现。

4.4.3 模型的测试效率对比

在处理大规模公式图片数据的过程中,效率也是衡量模型性能的一个重要指标。因此,我们简单对比了不同模型的测试效率。

表 10 列出了 4 个模型测试一遍测试集的平均时长。可以看出,同样的实验条件下,输入数据维度较少的 MCFIRD 模型的测试效率最高,且有较明显的优势,因此其更适合用于大规模的图片重复检测。

表 10 不同模型测试的平均时长

Table 10 Average time for models's test

模型	平均时长
MCFIRD	1m48s
resnet	2m48s
siamese	3m22s

结束语 本文针对网络试题资源的重复检测问题,在利用公式图片进行重复检测方面,提出了基于 MCFIRD 模型的重复检测识别方法。MCFIRD 模型具有不错的识别效果与泛化能力,在标准公式图片的数据集和模拟真实场景噪声的数据集上都取得了较高的准确率与较快的速度,具有现实应用的意义。在题库系统中使用 MCFIRD 模型,可以提高题库试题质量。与传统方法相比,这种方法使用多通道卷积的方式实现了公式图片提取的自动化,从而无需改变模型,就可识别仅系数或变量名改变的相同公式,而且能够覆盖更大规模的公式图片数据集;利用大数据的优势,避免了传统方法中人工设计特征带来的费时缺点与可能的疏漏;使用端到端的输出方式,避免了传统方法中间步骤带来的误差累计。总的来说,MCFIRD 模型为公式重复检测任务面临的挑战提供了解决方案,可实现试题重复检测的最终目的;并且与文中所列的其他同样基于深度神经网络的模型相比,其面对噪声的泛化能力更强,测试效率更高。

本文提出的方法,目前仅适用于仅改变了系数与变量的公式的重复检测,未来将对公式的进一步变换进行探索研究,如交换律和结合律,以及不同坐标系下的不同形式等等。此外,本文实验所采用的数据集均是标准字体,不涉及更多艺术或手写字体,能否推广该方法,使之能够识别各种不同形式、不同字体的相同公式,具备更广泛的应用范围,是未来的研究方向之一。

参 考 文 献

[1] BRESLOW L, PRITCHARD D E, DEBOER J, et al. Studying Learning in the Worldwide Classroom Research into edX's First MOOC[J]. Research & Practice in Assessment, 2013, 8: 13-25.

[2] POLSON M C. Foundations of Intelligent Tutoring Systems [M]. Hove, UK: Psychology Press, 2013.

[3] HUANG Z, LIU Q, CHEN E, et al. Question Difficulty Prediction for READING Problems in Standard Tests[C]// AAAI. 2017: 1352-1359.

[4] LIU Q, CHEN E H, ZHU T Y, et al. Research on Educational

Data Mining for Online Intelligent Learning[J]. Pattern Recognition and Artificial Intelligence,2018,31(1):77-90.

[5] KOHLHASE M,SUCAN I. A search engine for mathematical formulae[C]// Proceedings of the 8th international conference on Artificial Intelligence and Symbolic Computation (AISC'06). Berlin:Springer-Verlag,2006:241-253.

[6] JADERBERG M,SIMONYAN K,VEDALDI A,et al. Reading Text in the Wild with Convolutional Neural Networks [J]. arXiv:1412. 1842v1.

[7] LIN X Y,GAO L C,TANG Z. Mathematical Formula Identification and Performance EvaluationinPDFDocuments[J]. International JournalonDocument Analysis and Recognition, 2014, 17(3):239-255.

[8] YIN Y,HUANG Z,CHEN E,et al. Transcribing Content from Structural Images with Spotlight Mechanism[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM,2018:2643-2652.

[9] LIU Q,HUANG Z,HUANG Z,et al. Finding Similar Exercises in Online Education Systems[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM,2018:1821-1830.

[10] WANG H,XU T,LIU Q,et al. MCNE:An End-to-End Framework for Learning Multiple Conditional Network Representations of Social Network[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). Association for Computing Machinery, New York,NY,USA,2019:1064-1072.

[11] LIN X Y,GAO L C,TANG Z. A Text Line Detection Method for Mathematical Formula Recognition[C]// Proceedings of International Conference on Document Analysis and Recognition. 2013:339-343.

[12] LI Y H,WANG K J,SHANG G W,et al. Baseline structure analysis and recognition algorithm research of mathematical formula[J]. Computer Engineering and Applications,2008,44(16):18-22.

[13] ZANIBBI R. Recognition of mathematics notation via computer using baseline structure[R]. Queen's University,Kingston,Ontario,2000.

[14] GUO J N. Research on Detection Algorithm of Mathematicial Formula for MathML[D]. Jinzhou:Bohai University,2016.

[15] ZHU H,NIE Z,DING M. Image recognition by affine moment invariants in Hartley transform domains[C]// International Symposium on Communications and Information Technologies. IEEE,2010:630-633.

[16] LI J,CHENG J,SHI J,et al. Brief Introduction of Back Propagation (BP) Neural Network Algorithmand Its Improvement [C]// Advances in Computer Science and Information Engineering. Berlin:Springer. 2012.

[17] LECUN Y,BENGIO Y. Convolutional networks for images, speech,and time series[J]. The Handbook of Brain Theory and Neural Networks,1995,3361(10):1995.

[18] CHOPRA S,HADSELL R,LECUN Y. Learning a similarity metric discrim-inatively, with application to face verification [C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR 2005). IEEE,2005:539-546.

[19] ZAGORUYKO S,KOMODAKIS N. Learning to compare image patches via convolutional neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:4353-4361.

[20] IOFFE S,SZEGEDY C. Batch normalization:Accelerating deep network training by reducing internalcovariate shift[C]// International Conference on Machine Learning. 2015:448-456.

[21] IOFFE S,SZEGEDY C. Batch normalization:Accelerating deep network training by reducing internal covariate shift[J]. arXiv:1502. 03167,2015.

[22] GLOROT X,BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]// Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010:249-256.

[23] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[24] HE K,ZHANG X,REN S,et al. Identity mappings in deep residual networks[C]// European Conference on Computer Vision. Cham:Springer,2016:630-645.

CHEN Ang, born in 1983, Ph.D, associate. His main research interests include data mining and educational big data.

ZHOU Yu-qiang, born in 1998, postgraduate. His main research interests include machine learning and data mining.