分 类 号 _____ 学号 M201472380

学校代码 10487 密级 公开

华中科技大学

# 硕士学位论文

## 基于卷积神经网络的
## 文档图像分类与检索方法研究

学 位 申 请 人 ： 李立

学 科 专 业 ： 模式识别与智能系统

指 导 教 师 ： 陈友斌 教授

答 辩 日 期 ： 2017 年 05 月 20 日

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

# Document Image Classification and Retrieval Based On Convolutional Neural Networks


**Candidate : Li Li**

**Major : Pattern Recognition and Intelligent Systems**

**Supervisor : Prof. Chen Youbin**

**Huazhong University of Science and Technology**

**Wuhan 430074, P.R.China**


**May, 2017**

# 摘要

随着计算机和互联网技术的发展，图像采集终端的多样化普及化，越来越多的文档以图像的形式被收集和处理。设计合适的文档图像处理系统以面对不断增长的文档图像数量，成为研究的热点。文档图像的分类与检索是文档图像处理中的关键任务。

本文围绕文档图像的分类与检索展开深入研究，主要工作如下：

1）提出基于卷积神经网络（Convolutional Neural Network，CNN）和支撑向量机的文档图像分类方法。利用CNN从底层像素中获得高层视觉特征，结合支撑向量机进行文档图像分类任务。参考很多学者将深度卷积神经网络应用到不同机器视觉领域的做法，本文选用三种经典卷积网络结构，在两个样本数不一的开源数据集上测试。实验表明该分类方法具有良好的表达能力，并且能方便地在不同类别数据集间进行迁移。

2）提出基于 CNN 和层次 K 均值树算法的文档图像检索方法。利用 CNN 从原始像素中获得抽象特征，采用主成分分析（Principal Component Analysis，PCA）算法进行特征降维，所得特征作为文档图像的索引特征。检索阶段采用近似最近邻搜索的方法，提高对大型数据集的检索效率。实验证明该检索方法能达到更高的查准率和更少的检索时间，具备较高的实际应用价值。

实验结果表明，本文采用的基于卷积神经网络的方法性能优异，分类的准确率和检索的查准率均能满足现有需求。

**关键词：** 文档图像分类　文档图像检索　卷积神经网络　特征压缩　近似最近邻搜索

# Abstract

With the development of computer technology and the prevalence of image acquisition terminals, more and more documents are collected and processed in the form of images. The design of appropriate document image processing system to deal with the growing number of document images, become a hot research. Document image classification and retrieval are two major tasks in document image processing system.

This thesis focuses on document image classification and retrieval, and its main contribution is as follows:

1) propose a document image classification method based on convolutional neural network and support vector machine. Our method obtains high-level visual features from raw image pixels via convolution neural network, and achieve satisfying classification performance with support vector machine. In this paper, three classical convolution network architectures are tested on two open source datasets with different number of samples. Experiments show that out method has better performance, and can be easily transfer between different types of datasets.

2) propose a document image retrieval method based on convolutional neural network and hierarchical k-means tree. After image features obtained via convolutional neural network, normalization and principle component analysis are applied for dimension reduction. In retrieval phase, approximate nearest neighbor search based on hierarchical k-means tree is chosen to provide better retrieval efficiency for large datasets. Experiment show that out method can achieve acceptable precision but less retrieval time, which proves to be of high practical value.

The experiment results show that method based on convolution neural network is superior in performance, and the accuracy of classification and precision of retrieval can meet the existing requirements.

**Key words**：document image classification　document image retrieval　neural networks　feature compression　approximate nearest neighbor search

# 目录

**摘要** ..... I

**ABSTRACT** ..... II

**目录** ..... 1

**第 1 章 绪论** ..... 3

1.1 研究背景和意义 ..... 3
1.2 国内外研究现状 ..... 4
1.3 研究目标与内容 ..... 8
1.4 论文组织结构 ..... 9

**第 2 章 文档图像处理预备知识** ..... 11

2.1 常见应用框架 ..... 11
2.2 常见预处理算法 ..... 13
2.3 常见特征提取算法 ..... 13
2.4 图像相似度度量标准 ..... 15
2.5 图像检索性能的评价 ..... 16
2.6 深度学习基础 ..... 17

**第 3 章 基于 CNN 的文档图像分类方法** ..... 19

3.1 问题描述 ..... 19
3.2 算法简述 ..... 20
3.3 总体流程 ..... 21
3.4 预处理 ..... 22
3.5 基于 CNN 的文档图像特征 ..... 23
3.6 分类器设计 ..... 28

3.7 本章小结 ..... 29

**第 4 章 基于 CNN 的文档图像检索方法 ..... 30**

4.1 问题描述 ..... 30
4.2 算法简述 ..... 30
4.3 总体流程 ..... 31
4.4 特征降维 ..... 32
4.5 搜索方法 ..... 34
4.6 本章小结 ..... 36

**第 5 章 实验与分析 ..... 37**

5.1 引言 ..... 37
5.2 数据集描述 ..... 37
5.3 实验过程 ..... 39
5.4 结果分析 ..... 42

**第 6 章 总结和展望 ..... 58**

6.1 论文内容总结 ..... 58
6.2 未来研究方向 ..... 58

**致谢 ..... 60**

**参考文献 ..... 61**

**附录 1 攻读硕士学位期间主要的研究成果 ..... 66**

# 第1章　绪论

## 1.1 研究背景和意义

随着数字时代的发展和互联网的普及，信息的存储和分享也变得越来越频繁。伴随着互联网中信息数据的急剧增长，信息的表现形式也越来越多样。从信息的载体类别来看，信息的主要表现形式有文本、声音、图像和视频等。文本信息以字符编码的形式被存储，相比于信息的其他表现形式，具有存储时占据较小的空间资源的优点，却也有比较容易被修改的缺点，因而并不适合于对安全性或知识保护要求较高的应用场景。在对安全性或者证据性要求较高的场景下，往往采用图像和视频的方法存储信息，例如在银行中办理重大业务时，往往会留下视频记录；银行中的重要业务单据，也会以扫描文档图像的形式备份。其中，文档图像因其便利性、易于校验性和难以伪造性，在各种应用中都得到普及。

大量的纸质书籍和各式各样的公文文件被扫描成数字图像，且整体数量呈指数增长。随着社交媒体如微博、BBS 社区等建立全新的社会生活形态和交流渠道，文档图像不仅在独立的商业业务系统中存储使用，也越来越广泛地在互联网中传播。对文档图像进行高效管理，成为众多企业应用需求的关键点。在传统的业务场景下，对文档图像标注类别等信息，往往是在业务流程中通过手工的方式进行，不仅费时费力，而且还存在较高的错误率。设计合理的文档处理系统，对大量的文档图像进行快速高效的自动化处理，也正越来越重要。

文档图像的分类与检索，一直是文档图像处理中的关键任务。文档图像分类和检索的性能，直接影响到文档图像处理系统的整体性能。

自 2006 年以来，深度学习已经成为解决大数据难题的重要理论方法，也已经成为机器学习领域中的研究热点[1]。文档图像本质上是以图像形式存储的文本信息，深度学习的方法对文档图像的表示、分类与检索等任务都具有很好的借鉴意义。

从理论的角度考虑，目前主流的文档图像分类方法主要基于词袋模型（Bag-of-Words，BOW）或者其扩展模型，如空间金字塔匹配模型（Spatial Pyramid Matching，SPM）和水平垂直分块模型（Horizontal Vertical Partition， HVP）。基于 BOW 的方法，通常在原始的文档图像，或者从文档图像提取的特征上建立字典，然后得到文档图像的特征表示。这种做法更着重文档图像结构的频率信息，忽略文档本身的空间布局信息。即使结合 SPM 的方法，也只能得到浅层的特征表达，依然使得文档图像的特征表达存在极大的信息损失。现有的文档图像特点是：类型多样、类别细化。对于复杂的分类问题，现有算法在不同文档图像数据集间的泛化能力不够优秀，这时更需要在算法和模型上进行探索，获取更加鲁邦稳定的特征表达。

现如今，基于深度模型解决大规模数据下的机器视觉和模式识别任务已经成为非常热门的研究方向。基于深度模型的方法，能更好的挖掘出文档图像中的版面信息，与具体任务相结合，必能更好地完成文档图像分类索引等各种应用。

从实际应用的角度考虑，高性能的 GPU 和 GPU 集群提供强大的计算能力，给深度学习应用提供硬件基础。深度学习模型也存在些缺点。深度学习的模型训练时需要针对具体问题调整参数，优化过程中误差的反向传递可能会爆炸或消失而终止。

虽然优化深度学习的非凸代价函数，存在理论的不足和实际的困难，但在大多数领域基于深度学习的方法表现远优于传统方法。基于此，我们需要深入研究，才能使其在文档图像分析领域不断发展[2]。

以深度神经网络为基础，如何鲁棒地表示文档图像，并进行文档图像的分类和检索是一个值得研究的重要课题，不仅具有一定的理论意义，也具有较高的应用价值。

## 1.2 国内外研究现状

在过去的 20 年中，文档图像处理的技术应用广泛，包括文档图像分类和文档图像检索等领域，都吸引大量学者进行研究。现实生活中的文档图像多种多样，但在具体的应用场景中，只需考虑有限类别数的样本，实际针对的样本类别和需要达成

的性能目标因业务场景的不一而有所不同。针对不同的应用场景，往往存在不同的方法得到更优的结果。本节对文档图像分类和文档图像检索的研究进展进行简要阐述，下面的方法适用的场景不一，但在各自的场景下都取得较好的结果。

基于文本信息的全文检索技术已经十分成熟。该技术采取关键字索引等方法，使得基于文本的检索速度快，准确率高。有鉴于此，一种直观的文档图像分类和检索方法是，首先通过光学字符识别（Optical Character Recognition, OCR）技术对文档图像进行全图的识别，将文档图像转换为文本，然后利用基于文本信息的方法达到对文档图像进行分类和检索的目的。这类方法称之为依赖识别的文档图像处理方法，学者们基于这样的思路，开发出若干文档图像处理系统。应用 OCR 技术进行全图识别存在若干问题。首先，OCR 技术的识别准确率与文档图像的类别和质量息息相关。如果文档图像的质量有限，例如存在噪声、笔迹涂抹、印章遮盖、倾斜等现象，OCR 技术的识别准确率会快速下降。有学者应用自然语言处理等相关方法，使得即使在 OCR 识别率相对较低时，还能使得整体的性能保持相对稳定[3–6]。其次，OCR 技术的语言依赖性很强，往往只对特定类别的语言识别效果较好。虽然也有学者开发针对多语言的 OCR 识别引擎，但效果并不如意。最后，OCR 技术本质上是一个多分类问题，计算复杂度很高，要消耗大量计算资源。当对文档图像全图进行识别时，需要耗费大量时间，无法达到实时性的要求。依赖识别的文档图像处理方法的性能，严重依赖 OCR 技术的性能，在实际应用中也就存在语言依赖性强，计算资源消耗大等问题。

在依赖识别的文档图像处理方法的实际应用中，难免需要在 OCR 的识别过程中对结果进行人工校正，对于大量的文档图像来说，工作量太大，因此也限制了依赖识别的文档图像处理方法的使用范围。越来越多的学者更加重视基于图像特征的文档图像处理方法，也提出大量基于此类方法的研究成果。

不依赖识别的文档图像处理方法，主要使用基于图像本身的特征，关注文档图像本身的视觉表示。考虑到文档图像所独有的版面特性，使用基于图像本身的特征

表示文档图像，能在特征中更好地保留文档的版面结构信息和图片等非文字组件的信息，这类方法通常更符合直观，也更有效[7]。

对于文档图像检索的任务而言，目前有很多方法都采用基于示例的查询模式[7-10]。基于示例的查询模式，首先在离线的状态下对文档图像进行特征提取和建立索引数据库。在查询的时候，从输入文档图像（单词、图标、签名等）提取特征，并和索引数据库中的特征进行比较匹配，返回查询结果。这种查询模式会计算输入文档图像与数据库中的文档图像的相似程度，所有相似程度高于特定阈值的文档图像都会被认为与输入图像相关，当然后续还可以通过几何验证等其他方法进一步验证[9,11]。上述这些方法都强调采用鲁棒且和尺度无关的图像特征进行匹配的重要性。

另一类方法则根据预定的示例文档图像集，提取特定的特征，并采用模型训练，以此定义文档图像间的相似度[12]。Shin 和 Doermann 定义版面结构的视觉相似性，并对每类文档都进行有监督的分类[13]。他们选择的特征包括：文档内容区域文本与非文本（图表、图像、表格、网格线等）各自所占百分比，列的版面结构，字体的相对大小，内容区域的密度，连通域特征的统计结果等。在分类任务中，针对这些特征，他们选用基于决策树的分类方法，或者基于自组织映射的分类方法。这类方法的缺点是，所采用的这些特征较适合于对表单、信件和文章等特定的文档类别进行分类，从某种意义上可以认为是特殊的模板匹配方法，对其他的文档类别并不很好适用。另外，由于这类方法综合了许多特征，提取这些特征的方法相对耗时，所以这类方法不宜用于对大型文档图像数据集的处理任务。

Collins-Thompson 和 Nickolov 提出用于估计有序文档图像集合中的页间相似度的模型[14]。他们也使用综合的特征，用于区分相关和不相关的页面。他们采取的特征，综合文本和版面特征，文档结构，主题概念等特征。这种方法涉及到 OCR 识别，因为 OCR 识别得到的文本可能包含错误，尤其针对手写文档更容易发生错误，这种方法只适用于结构严谨良好的打印文档。Joutel 针对页面级的手写历史文档检索任务，利用曲波变换为每页图像生成独特签名，用于检索任务[15]。这种方法会丢失文档图像的高阶结构显著性，因此只有当文档局部形状对分类很有帮助时，这种方法

才能取得有效的结果。然而很多情况下，计算得到的结构相似性应当来自于全局结构的比较和文档图像各组件间的比较。

基于 BOW 的方法在图像分类[16]、场景理解[17]和文档图像分类[18,19]等许多计算机视觉任务中都表现出优秀的结果。然而，通过基于 BOW 的方法计算文档图像的相似程度，通常会因只计算对应字典码表的频率信息，而忽视文档图像组件间特有的版面位置信息。因此，基于 BOW 的方法描述文档图像的能力也有限，在图像存在噪声，版面发生变化等情况下，性能会有所下降。

针对 BOW 会忽略文档图像组件间特有的版面位置信息的问题，不少学者在文档图像处理的问题上将传统的 BOW 方法进行扩展，以包含空间位置信息。其中一类扩展方法，参考空间金字塔匹配[20]的方法，将完整的文档图像划分为较小的区域并设置不同的权重，在每个小的区域里根据 BOW 的方法计算出直方图并归一化，最后根据不同的权重对不同区域的直方图进行综合[21–23]。现在这类方法在不断探索更优的特征综合方式和更有效的局部统计量计算方法。

近几年，随着深度学习在计算机视觉各个领域，如目标识别、场景分析、自然语言处理上表现出优于传统方法的性能。传统方法专注于以人工构造的方式，找到更有效地对文档图像进行视觉表示的特征。受文档图像所特有的版面结构层次分布的启发，也有学者将深度学习用于文档图像分类和检索的领域，并取得同样优异的性能表现。实际上，文档图像分析处理系统中，从特征表达，到分类检索，到文本识别等所有组件，都可以通过深度学习的方法训练得到。

Le Kang 等首次采用卷积神经网络对文档图像进行分类[24]。常见的文档图像分辨率高于 2000×2000，在目前的计算资源下很难满足卷积神经网络的训练要求。考虑到高维输入更可能导致过度拟合，而且考虑到分辨文档图像类别更多依靠其布局，而不是其细节（如字符文本），因此在预处理阶段，采用双线性插值算法将所有图片缩放至 150×150。实验采用的卷积神经网络，由两个卷积层、两个最大池化层、两个全连接层和一个 Softmax 输出层组成。通过在数千张样本的数据集上进行实验，证明其设计的卷积神经网络的性能，优于传统的文档图像分析方法。

Muhammad Afzal 等在此基础上提出层次更深的卷积神经网络，由五个卷积层、三个池化层、三个全连接层和一个 Softmax 输出层组成[25]。在对文本图像样本进行训练前，作者首先将设计的网络在 ImageNet 数据集[26]上进行预训练，然后再在文档图像数据集上进行迁移学习，最终在同样的文档图像数据集上得到比[24]更好的结果。这组实验说明训练卷积神经网络需要大量数据，而且从目标分类到文档图像分类的迁移学习是有效和可行的。

Adam Harley 等同样也在 Le Kang[24]的基础上进行改进。首先，作者整理出 IIT-CDIP[27]的一个子集 RVL-CDIP[28]，由多达 16 类合计 400000 张文档图像构成。作者采用 Krizhevsky 提出的卷积神经网络结构[26]进行实验，证明（1）由卷积神经网络提取的特征对特征压缩较为鲁棒；（2）在非文档图像上训练得到的卷积神经网络，迁移至文档分析的任务中，有不俗的表现；（3）假设有足够的训练数据，强制训练特定区域的特征然后进行组合的方法是没有必要的。

Lucia Noce 等更是将前人的成果和 OCR 技术与自然语言处理结合起来，通过以视觉增强的方式添加内容信息来对结构相似的文档类别进行分类[6]。具体的实验方法是，采用 Tesseract OCR[29]对整张图像进行识别；鉴于 OCR 技术的种种限制和容易出错，引入自然语言处理的方法，对每个类别建立对应的字典；对对应类别字典中存在对应单词的区域进行不同颜色的标注，然后再进行网络的训练和测试。实验时采用与[26]中结构相同的网络。该篇论文主要解决采用卷积神经网络时对文档结构类似但类别不同的样本进行分类，因此实验采用不同的数据集，并未详述。

以上的一些基于深度学习的进展，主要集中在通过卷积神经网络，学习得到文档图像的结构特征，并以此进行分类和检索。实验结果表明，在大型数据集和文档图像成像质量无法保证的情况下，使用深度学习相关技术能得到更好的结果，

## 1.3 研究目标与内容

目前，基于深度学习的方法已经开始应用到文档图像分析系统的各个组件中。在文档图像分析的领域，已经有很多学者研究结合深度学习的方法进行文档图像的

分类[6,24,25,28]。但是，他们的研究方法主要基于同样结构的卷积神经网络结构，并没有对基于不同卷积神经网络结构的方法进行比较，缺少更进一步的研究，也缺少在此基础上，对文档图像检索的效果进行验证比较。

本文主要关注文档图像分类任务与文档图像检索任务，主要的工作和贡献包括：

1. 在前人的基础上，对文档图像分类任务采用不同的网络结构进行处理。参考很多学者将深度卷积神经网络应用到不同机器视觉领域的做法，本文选用三种经典的卷积神经网络结构（AlexNet[26]，GoogLeNet[30]与 VGG16-Net[31]）解决文档图像分类的问题。本文选择两个较常见的文档图像开源数据集（Tobacco3482[27]与 RVL-CDIP[28]），在预训练的三种网络上进行优化微调，得到最终的模型结果。本文将采用卷积神经网络的方法得到的结果，与基于 BOW 和空间金字塔的方法得到的结果，进行比较，发现基于神经网络的方法在分类准确性上优于传统方法。

2. 不少学者的研究表明，从包含卷积神经网络的深度学习模型中提取得到的特征，具备很强的鉴别能力，在大部分视觉识别任务中都可作为首选特征使用。参考这些研究，本文也从不同网络结构的深度卷积神经网络中提取特征，采用 PCA 算法降维，作为文档图像索引的参考特征。实验证明，即使在文档图像索引的任务上，基于卷积神经网络的方法得到的特征也要优于传统方法得到的特征。但在本文的实验中发现，不同网络层的特征表达能力强弱，与目标识别领域的结果有所不同。

3. 利用去卷积网络，本文可以观察到输入图像的哪些信息对最终的网络结果贡献较大。对于文档图像分类的任务，本文也使用去卷积网络观察文档图像中哪些信息对最终的结果输出贡献明显，并对实验结果进行解释说明。

## 1.4 论文组织结构

本文共分为五章，文章结构及各章主要内容组织如下：

第一章 绪论。

本章主要介绍了文档图像分类和检索的研究背景及研究意义，分析并总结文档图像分类和检索近年来的研究发展与现状，并给出了本文的主要工作和整体组织结构。

第二章 相关研究综述。

本章首先分析文档图像分类和检索的特点，简要介绍文档图像分类与检索的常用方法，并分析方法的优缺点。其次，对传统的基于 BOW 和空间金字塔匹配的文档图像分类和检索的方法进行介绍，分析传统方法如今在该任务上面临的难点。最后，对深度学习的方法进行简要阐述，并简要描述后续用于实验的常用模型结构。

第三章 基于卷积神经网络的文档图像分类方法。

本章首先介绍现有的文档图像分类算法所面临的问题。针对提出的问题，本文采用基于卷积神经网络的方法进行解决。本文采用多种网络结构进行实验，并与传统方法进行综合比较后，通过实验结果证明方法的有效性与鲁棒性。

第四章 基于卷积神经网络的文档图像检索方法。

本章首先介绍现有的文档图像检索算法所面临的问题。针对提出的问题，本文采用基于卷积神经网络的特征进行解决。实验结果证明，在文档图像检索的问题上，由卷积神经网络获得的特征即使在压缩后依然能较好的完成检索任务。同时，本文发现对于文档图像检索任务，不同网络层的能力存在差异。

第五章 实验与分析。

本章对实验方法、环境和结果进行比较说明。同时采用去卷积网络的分析方法，可视化输入图像中对结果输出贡献的情况，进行一定的解释性说明。

第六章 总结与展望。

# 第2章　文档图像处理预备知识

当今世界，越来越少纸化。因此，大量的文档、书籍、信件、手稿等都通过电子设备存储在日常生活中。这些原本是纸张的文档，通常通过扫描仪、传真机、数码相机等被采集并保存为数字图像。这些数据的数量每天都在大量上涨。从如此大量的数据中，进行自动提取、分类和信息检索，是很有价值的。

因为数字化文档种类和数量的不断增长，开发文档图像处理系统，以对大量结构特征明显或不明显的文档图像进行处理，成为需求的热点。在这一领域，有很多方法被开发出来，以提供有效高效的文档图像处理能力。

本章首先对该领域的基本算法和近几年的发展进行综述，然后简要介绍基于空间金字塔匹配的文档图像分析方法，最后也对本文参考的基于卷积神经网络的方法进行简要介绍。本文第三章和第四章中，主要采用基于空间金字塔匹配的方法作为实验比对对象。

## 2.1 常见应用框架

文档图像分类与检索的方法，可以根据是否需要进行字符识别，分为两大类。依靠识别的方法，需要对文档图像进行识别，根据识别结果来推测文档间的相似程度，然后进行分类或检索。依靠识别的方法，本质是在文字符号层面上进行相似度度量。不依靠识别的方法[23,32–36]，则是从文档图像上直接提取特征，根据特征间的距离度量来推测文档间的相似程度。不依靠识别的方法，本质是直接在文档图像层面上进行相似度度量。

光学字符识别（Optical Character Recognition，OCR）是一种传统的文本识别方法。基于 OCR 识别的文档图像分类与检索方法，其性能与效果会受到 OCR 识别效果的影响。OCR 识别方法存在一些弱点，它需要消耗较高的计算资源，但识别结果受到图像分辨率和成像质量的影响，并且依赖特定的语言种类。因此，基于 OCR 识

别的文档图像分类与检索方法，难以处理低质量的文档图像类别。例如，历史文档的扫描图像的质量较低，采用基于识别的方法就不能得到较好的结果。

不依靠识别的方法，其性能与效果就不会受到 OCR 识别效果的影响。在不依靠识别的方法中，每个文档图像都以一个特征矢量表达。

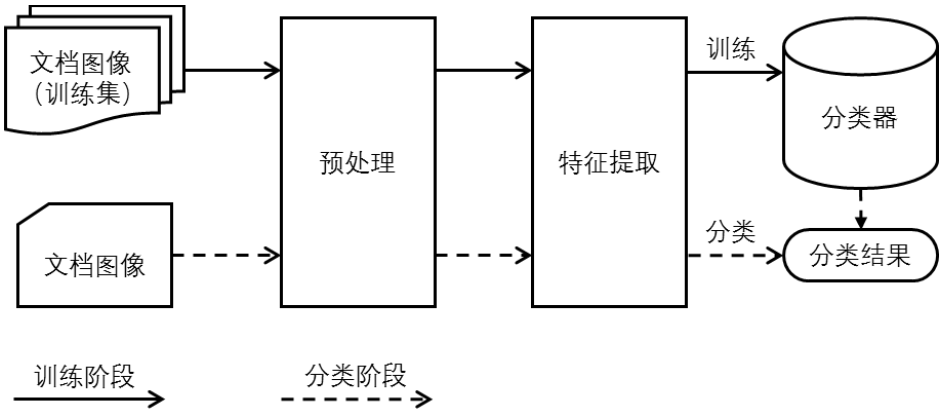



**图 2-1** 常用文档图像分类方法的总体框图

图 2-1 是常见文档图像分类系统的总体框图。在分类的过程中，首先在训练集的所有样本上提取特征，训练分类器；测试阶段，在测试图片上提取同样的特征，通过训练好的分类器进行预测，评估准确性。

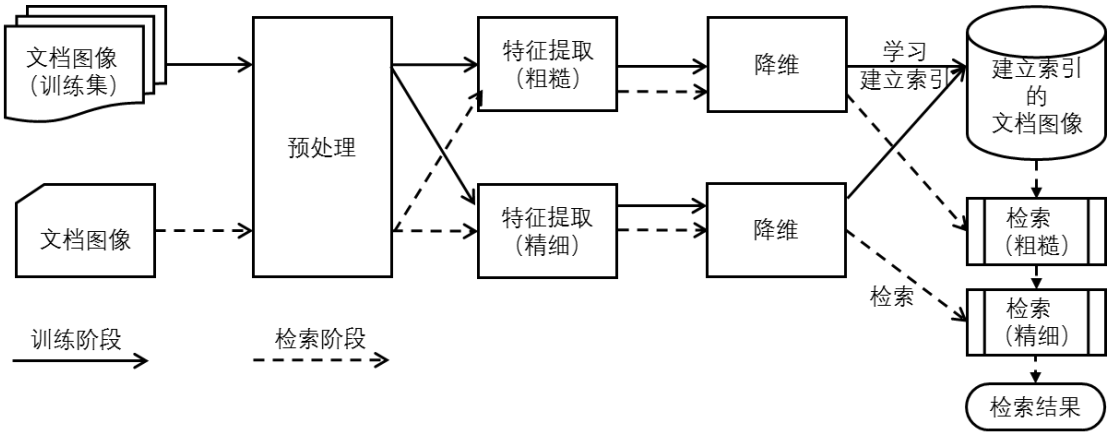



**图 2-2** 常用文档图像检索方法的总体框图

图 2-2 是常见文档图像检索系统的总体框图。在检索的过程中，首先在所有训练集样本上提取特征，并保存至数据库中，在检索过程中则对请求图像提取相同的

特征，用于相似性度量并得到最终的结果。对于检索系统而言，为了在保证准确性的基础上，降低检索所需的时间，会采用先根据粗糙的特征检索一遍，再在部分结果上进行精细检索的方案。至于具体所选用的特征，可以是原始像素相关特征，也可以是构造的其他特征。

## 2.2 常见预处理算法

预处理是文档图像处理的第一步。

文档图像在通过扫描仪、传真机、数码相机等采集为数字图像的过程中，会受到物理环境和采集设备的影响，成像质量可能受到噪声、变形、倾斜等的影响。处理这些文档图像，需要采用不同的预处理方法。通常而言，这些预处理方法可被划分为四大类[37]：滤波、几何变换、细化。

滤波算法主要用于对图像进行降噪、二值化和图像增强。实际采集到的文档图像中存在很多噪声，较常见的噪声包括：过量的椒盐噪声、将本应分离的字符或线条连接起来的油墨点块，因实际纸张折叠过导致成像后出现折痕等[38]。对图像进行平滑处理，能在一定程度上解决问题。常见的对图像进行平滑的算法包括：均值滤波[39]、中值滤波[40]和高斯滤波[41]。在对文档图像进行平滑后，可以通过对图像进行二值化或图像增强，进一步提取有用信息，过滤噪声等的影响。常见的二值化方法包括 Otsu、NiBlack 等。可能用于增强文档图像的算法，包括倾斜矫正[42–44]，边界去除和文字行宽度归一化等。

几何变换算法用于对文档图像进行尺寸变换。在部分文档图像处理系统中的初始步骤中，彩色图像可能被转变为灰度图像，图像的尺寸也可能被归一化至固定尺寸。

细化的算法可用于获取文档图像上字符的骨架，这些算法基于标志骨架计算特征，并迭代对目标的角点进行腐蚀操作。

## 2.3 常见特征提取算法

为了快速有效地对文档图像进行分类或检索，找到有效、独特和稳定的特征就变得十分重要。所提取的特征会明显影响检索的性能[45]。用于文档图像分类和检索的特征，可分为两大类：全局特征和局部特征。全局特征的方法，直接从整张图像上提取特征。在文档图像的领域，如纹理、形状、尺寸、位置等全局特征被用于文档图像的处理。而局部特征则关注局部特征点等。

词袋模型被广泛应用在自然语言处理、图像处理等领域。在自然语言处理中，一篇文档被视为一些单词的集合；在图像处理中，一张图像被视为一些特征点或特征块的集合。集合中的每个元素都是一个特征，元素出现的次数作为该维度的度量值，最终集合可以由元素构成的特征向量表示。

BOW 的方法只关注对应字典的频率信息，而忽略图像本身的结构信息。因此在将 BOW 应用在文档图像处理的领域，会对原始的方法进行扩展。一种扩展的方法就是参考空间金字塔匹配的方法，对文档图像做类似的分块，然后对不同块找出特征点量化为特征，将所有特征拼接起来形成最后的分类器。

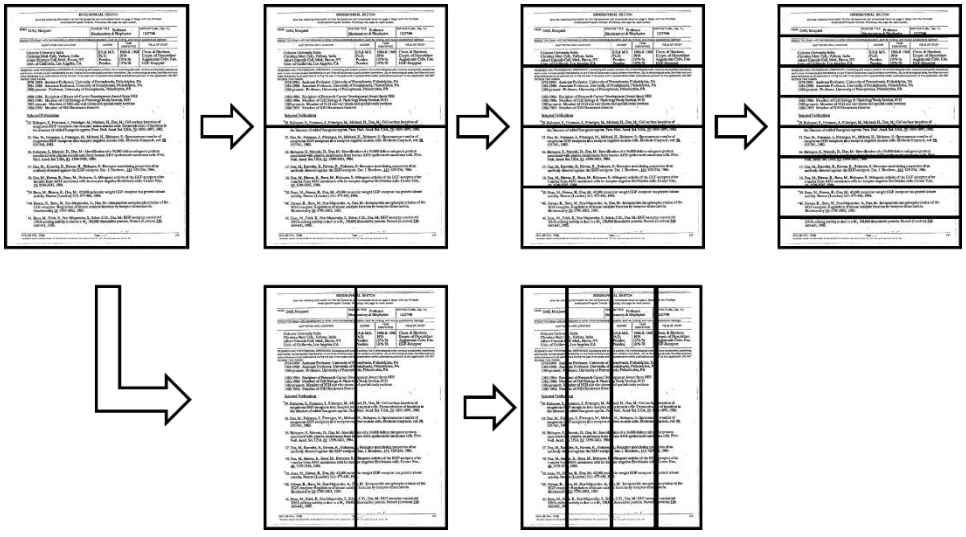


**图 2-3** 对文档图像进行水平垂直分块

图 2-3 中对一张文档图像进行 2 次水平分块，2 次垂直分块的结果。假定在每个分块中获得的特征维数为$|C|$，则图 2-3 中的分块操作后，总共提取的特征维数为 13

$|C|$。这样的扩展方法即保留文档图像组件的频率信息，也保留了文档图像版面布局信息。

Jayant Kumar[23]所设计的方法就是通过这样的方法提取特征，最后选用随机森林作为分类器。这也是本文主要的比较方法。

## 2.4 图像相似度度量标准

文档图像索引的目的是，在数据库中找到与用户请求的文档图像相近的文档图像。对请求的文档图像和数据库中已经建立索引的文档图像间进行相似度度量，既可以在像素层面进行，也可以在表达的特征层面进行。无论在哪个层面进行，在数据库中与请求的文档图像的相似度度量值最高的文档图像，就会被认为是与请求图像最相似的图像。

近来的研究里，最近邻方法被广泛应用于度量相似度上[10,46–49]。如果请求图像的特征矢量，与数据库中某张图像的特征矢量间距离度量最小，则它们的相似度最大。即在特征空间中，请求图像与某张图像间距离最小，则返回该张图像的结果作为请求的返回结果。

假设所采取的距离度量函数为 $Dist$， $A,B,C$ 为三张图片的特征向量，下面列举两种常用的距离度量方法。

1) Minkowsky 距离

该距离使用 L 范数定义如下公式

$$L_p(A,B)=\left[\sum_{i=1}^{n}|a_i-b_i|^p\right]^{\frac{1}{p}}$$

当 $p=1$时，该距离也叫曼哈顿距离（Manhattan Distance）；当 $p=2$ 时，该距离也叫做欧式距离（Euclidean Distance）；当 $p=\infty$时，该距离也叫切比雪夫距离（Chebychv Distance）。

2) 余弦距离

该度量方法计算的是两个特征向量之间的方向的差异，公式定义如下

$$A \bullet B = A^T B = |A| \bullet |B| \cos\theta$$

$$d_{\cos}(A,B) = 1 - \cos\theta = 1 - \frac{A^T B}{|A| \bullet |B|}$$

不同的图像特征需要采用不同的距离度量方法。实际工程项目中，需根据应用场景，选择合适的图像特征和距离度量方法。

在对大型数据集进行图像检索时，近似的最近邻搜索方法（Approximate Nearest Neighbor Search）也被应用，以减少计算资源的使用，降低检索所需的时间，提高检索的效率。

## 2.5 图像检索性能的评价

查准率和查全率是评价检索系统性能的两个重要指标。查准率是指在所有检索结果中结果正确的比例；查全率是指所有应该检索到的结果中，实际检索到的结果的比例。

假设输入检索系统的图片数量为 $s$，检索到的图片数量为 $r$，其中输入检索系统的 $s$ 张图片中包含相关图像 $u$ 张，检索结果的 $r$ 张图片包含相关图片 $w$ 张。那么准确率和查全率可以表示如下：

$$precision = \frac{w}{r}$$

$$recall = \frac{w}{u}$$

查准率和查全率往往是一对矛盾的指标，需要保证较高查准率时，查全率就会降低；需要保持较高查全率时，查准率就会降低。实际工程中需根据需求侧重点在两个指标中去折中。

本文还采取平均查准率（Mean Average Precision，mAP）对图像检索的性能进行评价。

$$mAP = \frac{\sum_{k=1}^{n}\left(P(k)\times rel(k)\right)}{n}$$

其中， $n$ 是检索得到的文档图像数， $p(k)$ 是检索结果中前 $k$ 项的准确率，而 $rel(k)$ 在第 $k$ 项检索结果正确时为 1，其他时候为 0。相比于原始的查准率计算方法，平均查准率对检索结果的顺序更敏感，只有当相关的结果优先于无关的结果时，平均查准率才能取得较高的值。

## 2.6 深度学习基础

### 2.6.1 深度学习概述

深度学习是机器学习中的新领域，已经在各种应用场景下表现出足够优异的性能，也因此成为研究的热点。与浅层学习相比，深度学习的模型结构更深，也更加突出特征学习的重要性。深度学习的方法，可以以海量数据为背景，自动抽取特征，免去人工选取特征的缺点。而且，从海量的训练数据中训练得到的特征，能十分容易的应用到其他数据上，得到不错的结果。深度学习是学习的手段，最终的目的在于学习到足够好的特征表达。

目前深度学习研究领域中比较主流的模型包括：卷积神经网络、循环神经网络、长短时记忆模型、深度玻尔兹曼机和自动编码机等。

### 2.6.2 卷积神经网络概述

经典模式识别问题解决的一般思路是，人工选择特征进行提取分析，找到最有效的特征解决问题。不同特征对不同问题的性能影响很大，这种思路很依赖特征选取者主观的学识和经验。

卷积神经网络因其较好的特征表达能力受到关注。卷积神经网络以卷积层、池化层和全连接层为主。采用 ReLUs 函数作为卷积神经网络的激活函数，使得特征映射具有位移不变性。同一平面层的神经元权值相同，因此有相同程度的位移、旋转不变性。求局部平均与二次提取的池化层，使得卷积神经网络有较高的畸变容忍能力。

对卷积神经网络模型进行优化，往往通过反向传播实现，最常用的算法是经典梯度下降法。常见的基于卷积神经网络的模型包括 AlexNet、GoogLeNet、VGG 等。

# 第3章　基于 CNN 的文档图像分类方法

## 3.1 问题描述

对文档图像的分类任务而言，用于表示文档图像的特征是否合适，会对算法的整体设计和性能造成很大影响。如果提取的文档图像特征鲁棒并且与图像尺度无关，则分类器的构造和训练就相对容易，最终的分类性能也能达到较好的水平。

不依靠识别的方法是目前主流的研究对象。不依靠识别的方法，通过学习基于图像本身的特征，表示文档图像的结构相似程度。

目前性能比较好的传统方法，结合 BOW 并扩展空间金字塔匹配的方法。这种方法首先从一些比较具有代表性的文档图像组成的集合中提取关键点的 SURF 特征，并使用 K-medoids 对 SURF 特征进行聚类，得到特征字典。然后，采用空间金字塔匹配的扩展方法，将文档图像划分为具备不同权值的小块，在每个小块上提取 SURF 特征，根据特征字典采用 L1 距离计算每个特征对应的矢量，在每个小块上得到归一化后的直方图表示，最终根据不同权值计算得到整体的特征表示。在分类器的训练阶段，采用随机森林的方法，同时对所得特征进行筛选，得到最优的随机森林分类器。

上述方法既能通过 SURF 特征提取到文档图像中具备鉴别能力的信息，又能够通过分块加权的方法保留一定的版面布局信息，在实际应用中已被证明是一种较好的文档图像分类方法。

但随着业务场景的发展和数据量的不断增多，这种方法在实际的应用场景中也表现出一定的不足之处。首先，这种方法要在具有代表性的文档图像上进行预训练以获得 SURF 特征字典。用于训练字典的样本数目和类别会对最终字典的表达能力产生影响，在大型数据集的分类任务中的性能有所下降。其次，采用随机森林的方法进行分类，训练时保留的决策树数目不宜确定。

有鉴于以上不足，我们提出基于卷积神经网络的文档图像分类方法。通过卷积神经网络提取得到文档图像的深度信息表达，在大型数据集上表现更佳，还可以很方便地在不同的数据集合之间进行特征迁移。

## 3.2 算法简述

基于卷积神经网络的文档图像分类方法，是通过卷积神经网络的卷积层输出或全连接层输出，提取文档图像的特征表达。卷积神经网络广泛用于端到端的识别任务。在识别的过程中，随着信息逐层向前传递，图像的信息从以原始像素表示，逐渐抽象转化为以更高级的语义信息表示。基于卷积神经网络的文档图像特征，是在卷积网络前向传播的过程中，将其中某一层的输出作为特征提取出来，作为最终的特征值。

已经有很多研究结果表明，从卷积神经网络中提取出的通用特征，实际上是非常强大的[50,51]。通过在目标识别、目标分类和视觉检索等领域的实验，已经证明从包含卷积神经网络的深度学习中获得的特征，在大部分的视觉识别任务中，都可以作为特征提取阶段的首选项。

围绕用卷积神经网络提取文档图像特征的想法，我们对传统的文档图像分类框架进行改进。

出于计算资源的限制和实际任务的需求，卷积神经网络一般需要固定尺寸的输入。例如，在场景分类的任务中，采用卷积神经网络对图像数据进行训练时，通常需要在预处理阶段将原始图像缩放到固定的尺寸大小，如256×256或者384×384。为了充分利用已有的网络模型和权值，我们也对文档图像采用双线性差值的方法缩放至256×256，而没有采用之前的平滑等预处理方法。

分类器本身是一种映射关系，从特征到类别标签的映射，因此分类器的种类多种多样。从这个角度讲，卷积神经网络中的全连接层，本身也是一种分类器。在我们的框架中，如果有大量样本用于训练，则直接将全连接层用于分类，这样的做法

保持原始卷积神经网络的结构；如果没有大量样本用于训练，则采用支撑向量机作为分类器，在提取的特征上进行训练。

事实上，我们的框架相对于传统框架的改进措施，在其他计算机视觉任务中已经被验证过。而我们的实验，也证明这样的做法是确实有效的。

## 3.3 总体流程

本文提出的基于 CNN 的文档图像分类方法，对小型数据集和大型数据集都能达到较好的分类性能。本节介绍基于 CNN 的文档图像分类方法的总体流程。

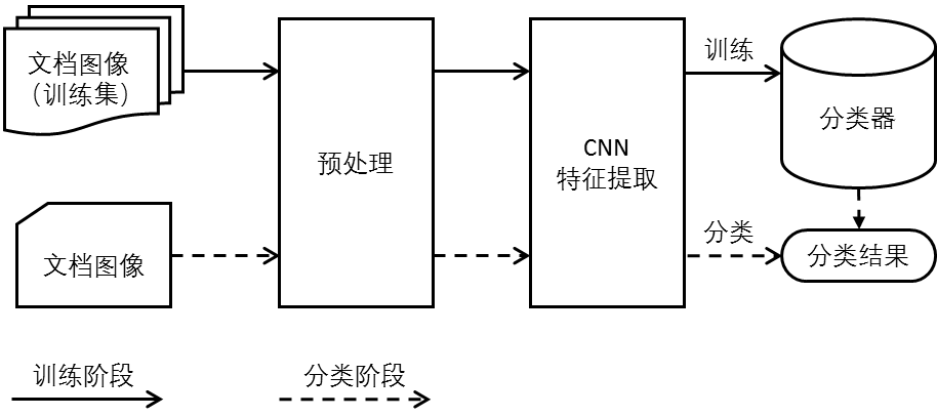



**图 3-1** 基于 CNN 的文档图像分类方法的整体框图

图 3-1 是基于 CNN 的文档图像分类方法的整体框图，总体流程可分为三个阶段，预训练阶段、训练阶段、分类阶段。

在预训练阶段，我们对所选用的 CNN 模型进行训练。训练的方式因 CNN 模型而异，但总体上与在如目标识别等其他任务中的方法一致，将图片缩放至固定的尺寸，采取特定的数据扩充方式，采用经典梯度下降法对 CNN 模型的权值进行更新完成训练。

在训练阶段，我们通过预训练的 CNN 模型提取文档图像的特征。如果有大量文档图像可用于训练，则选用全连接层作为分类器，整体框架保持经典 CNN 模型的结构；如果可用于训练的文档图像较少，则选用线性核的支撑向量机作为分类器，在

这种情况下，还可以考虑对提取的特征进行压缩降维，虽然该场景下对特征压缩的需求并不强烈。

在分类阶段，采取和训练阶段同样的流程，此时分类器根据提取的特征输出分类标签。

## 3.4 预处理

我们在实验中，选择三种经典的 CNN 网络结构：AlexNet，GoogLeNet，VGG16-Net。因此在预处理阶段，我们仅统一将文档图像转化为三通道图像，并将图像尺寸归一化至 256×256 的大小。

与传统的在原图或大尺寸图像上提取特征相比，将图像缩小的操作无疑会降低文档图像本身携带的信息。图 3-2 是将一些实际的文档图像缩小至 80×80 后的显示结果。在这样小的尺寸下，文档图像上的内容人眼已经模糊到难以分辨，但根据版面布局，我们依然能分辨出每个文档图像所属的类别。

**图 3-2** 将文档图像缩小至 80×80 后的显示结果

因此可以认为，即使将图像缩放至较小的尺寸，最具分辨能力的版面结构特征依然能得到保留。基于这样的观点，我们应用基于卷积神经网络的文档图像特征提取方法，并应用在文档图像的分类与检索任务中。

## 3.5 基于 CNN 的文档图像特征

基于卷积神经网络的方法在场景识别、目标分类等领域取得优异的性能结果，吸引很多学者对不同的模型进行研究。也有学者将其应用到文档图像分类的领域，同样取得优异的性能结果。但这些学者的工作主要集中于基于经典的 AlexNet 网络结构，并仅在文档图像的预训练、迁移学习等方面进行实验。

事实上，卷积神经网络中卷积和池化的操作所得到的高层视觉表达，本身就能很好的表征文档图像的版面布局特征，因此，从不同的 CNN 网络中提取的特征，应该都能得到比较好的应用效果。

### 3.5.1 CNN 模型的选择

本文在 Adam Harley 等的实验[28]基础上，选用三种常见的 CNN 模型进行实验。

1）AlexNet

AlexNet 是 Alex Krizhevsky 在 ISVRC-2010 上提出的网络结构，由 5 个卷积层和 3 个全连接层组成。

图 3-3 是 AlexNet 的网络结构示意图。第一行的 CONV1 是这一层的名称，在本文中仅作标识作用；11x11 Conv 表明这一层是卷积层，卷积核的大小是 11x11；96 表明这一层有 96 个卷积核；最后的 ReLU 和 Pooling/2 表明激活函数是 ReLU，经过尺寸缩小为原来一半的池化操作。

因为采用修正线性单元（ReLUs）作为激活函数，比采用传统的 tanh 作为激活函数，达到同样错误率的时间要少很多，因此 AlexNet 采用 ReLUs 作为结点的激活函数。

在本文实验的预训练阶段，仅修改 FC8 层的输出为对应数据集的类别。在特征提取位置的选择上，本文按照 Artem Babenko 的做法[51]，进行充分的实验比较，最终选择提取 FC7 的输出作为文档图像特征。

实验对比发现，在对文档图像优化微调后的 AlexNet 中，FC7 比 FC6 有更好的性能。造成这样的性能差异改变，根本原因是训练样本的不同。本文第 5 章 中给出详细的实验过程与比较结果。实验证明，如果将预训练的网络针对特定类别的数据集进行优化微调后，提取特征用于分类，较抽象的特征也有可能表现更佳。

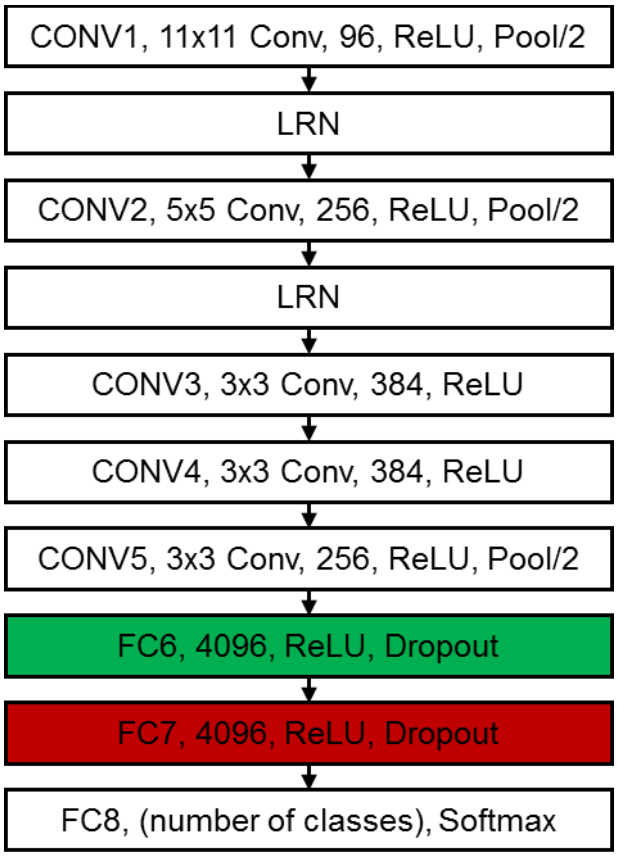



**图 3-3** AlexNet 网络结构示意图

2）GoogLeNet

单纯增加卷积网络的深度和大小，更容易导致过拟合，并且会增加计算性能的消耗。实际上全连接层中很多结点的权值是接近于 0 的，因此根本的解决办法是将全连接层变为稀疏连接层。基于这种考虑，Google 的科学家提出图 3-4 所示的 Inception 结构。

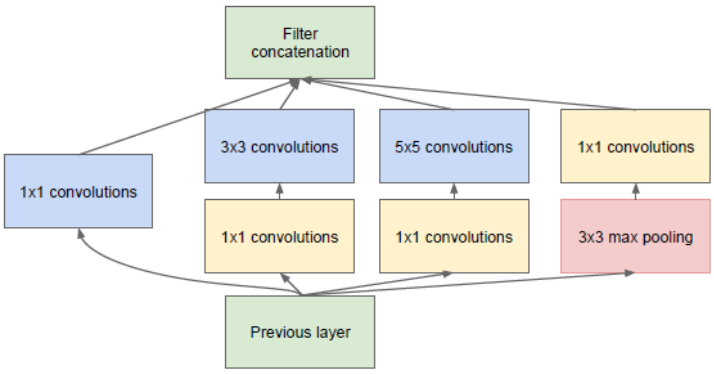



**图 3-4** GoogLeNet 中的 Inception 结构

图 3-5 是 GoogLeNet 的网络结构示意图。在本文实验的预训练阶段，仅修改 FC 层的输出为对应数据集的类别。因为没有类似的比较，在本文实验的训练和分类阶段，我们提取 GoogLeNet 网络中名称为 INCEPTION-4A、INCEPTION-4D 和 INCEPTION-5B 的输出作为文档图像特征，这三层的输出可视为 GoogLeNet 中不同层次的特征。实验结果表明，优化微调后，INCEPTION-5B 的性能表现最好。

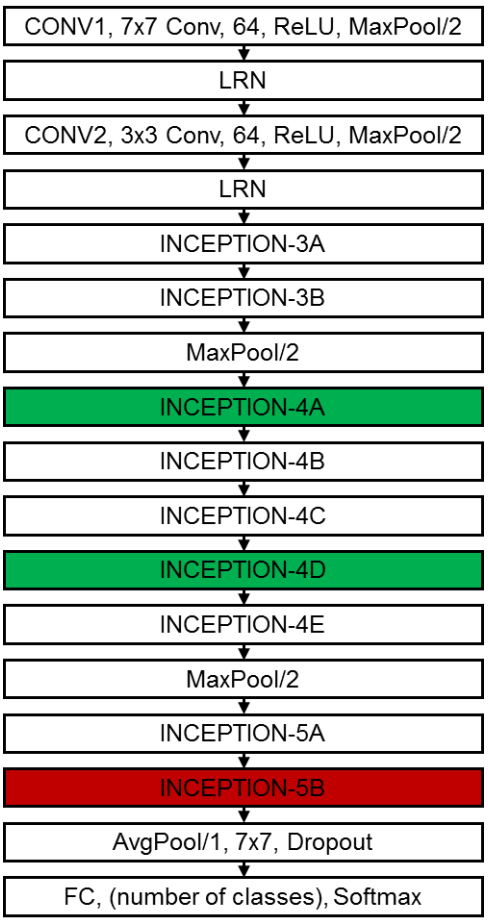



**图 3-5 GoogLeNet** 网络结构示意图

3）VGG16-Net

VGG16-Net 的网络结构如图 3-6 所示。相比于 AlexNet 的网络结构，VGG 的网络大量使用级联的 3×3 的小卷积层代替原有的单层大卷积层，而最后的卷积效果是类似的，在小卷积层间额外添加的 ReLU 层，则增加网络整体的非线性能力，使得网络整体的表达能力有所提高，同时，减少待训练的权值的数量。


**图 3-6** VGG16-Net 网络结构示意图

在本文实验的预训练阶段，仅修改 FC8 层的输出为对应数据集的类别。因为 VGG16-Net 的部分网络结构与 AlexNet 相似，在本文实验的训练和分类阶段，提取 VGG16-Net 网络中名称为 FC6 和 FC7 的全连接层输出作为文档图像特征，实验结果表明，FC7 的性能表现同样最优。

### 3.5.2 CNN 模型的训练

我们选用三种常见的 CNN 网络模型：AlexNet、GoogLeNet、VGG16-Net。

在训练过程中，我们仅对原有模型做细微的修改。根据具体的训练样本，将各模型最后的 Softmax 全连接输出个数调整至对应数据集的类别数。在训练过程中，

我们采取两种方式对网络模型进行初始化，第一种是以正态随机值初始化网络模型，第二种是以在 ImageNet 数据集上预训练过的模型权值初始化，在此基础上训练进行优化微调。

在数据增广方面，我们仅选择随机位置裁剪、水平垂直镜像投影的方式增加训练数据。考虑到文档图像扫描生成过程中，往往会存在随机漂移，和倒立等现象，我们选用以上方法增加训练数据。

训练过程采用经典梯度下降法（Stochastic Gradient Descent, SGD）进行权值更新，学习速率随迭代周期下降。

## 3.6 分类器设计

卷积神经网络中的全连接层，具备很好的非线性表达能力。在本文提出的基于 CNN 的文档图像分类方法中，如果有足够的文档图像用于训练，则直接利用全连接层进行分类。这种情况下，本文的文档图像分类方法就与卷积神经网络在其他计算机视觉任务中的方法一样，成为利用卷积神经网络的端到端的分类问题。

但是文档图像分类的业务场景不同，文档图像本身的标记十分昂贵，在训练阶段可能只有少量的文档图像。本文提出的基于 CNN 的文档图像分类方法，在面对小样本的训练问题时，则从预先从其他数据集上预训练的 CNN 网络中提取特征，然后使用分类器进行训练。

本文分类器设计部分，我们选择线性核的支撑向量机（SVM）。SVM 分类器是一种解决数据在原始空间线性不可分问题的分类算法。

对于两分类问题，SVM 的分类判别函数为：

$$\mathrm{sgn}\left(\sum_{i=1}^{t} y_i \alpha_i K(x_i, x) + b\right)$$

其中，$x$ 是待分类样本的特征向量，$x_i$ 是训练得到的支撑向量，$y_i \in (-1, +1)$ 表示支撑向量所对应的类别，$\alpha_i$ 和 $b$ 是通过训练得到的参数。

面对多分类问题时，通常由两种分类策略：一种是一对一，一种是一对多。本文采用一对多的分类策略，即对每一类训练单独的分类器，以区分这一类的样本和所有不是这一类的样本。在分类阶段，则根据多个分类器的结果综合投票得到最终结果。

在分类器训练过程中，首先通过预训练后的 CNN 提取特征，将所有用于训练的特征每一维都按照最大最小值归一化至[-1，1]，然后进行训练。本文方法中，选用线性核函数作为 SVM 的核函数，其形式为：

$$k(x_i, x_j) = x_i' x_j$$

相比于高斯核函数，线性核函数有参数少，运算快等特点。实验中还发现，线性核的整体表现更加稳定。在系统的实际构建中，我们借助开源工具 LibLinear[52]对分类器进行实现。通过在测试集上的测试，整体分类性能可以达到 90%左右，性能令人满意。

## 3.7 本章小结

本章主要介绍基于卷积神经网络的文档图像特征提取方法，和其在文档图像分类任务中应用的流程和方法。

基于卷积神经网络的文档图像特征提取方法，需要对所选用的卷积神经网络模型进行预训练。因为从卷积神经网络中提取的通用特征具有较强的表达能力和迁移能力，因此预训练过程可以在常用的 ImageNet 数据集上进行，也可以在大型的文档图像数据集上进行。

对于文档图像分类的任务，如果能够得到足够数量的文档图像作为训练集，则可以直接在经过预训练的卷积神经网络上对已有的网络结构权值进行优化微调；如果无法得到足够数量的文档图像作为训练集，即面对小训练样本问题，则可以通过经过预训练的卷积神经网络进行特征提取，然后训练传统的分类器，如支撑向量机。

# 第4章　基于 CNN 的文档图像检索方法

## 4.1 问题描述

与文档图像分类任务相比，文档图像检索任务对于表示文档图像的特征要求更高。如果提取的特征本身区分能力强，同样的相似度度量方法得到的结果就越好；如果提取的特征维数较少，则直接检索的效率就相对提高。

和前一章用于比较的方法相同，基于 BOW 和扩展空间金字塔匹配的方法，既提取到文档图像中具备鉴别能力的信息，又能够通过分块加权的方法保留一定的版面布局信息，在实际的文档图像检索应用中已被证明是一种较好的方法。但随着业务场景的发展和数据量的不断增多，该方法也表现出一定的不足。

本文的上一章中介绍基于卷积神经网络的文档图像特征提取，并将其应用在文档图像分类问题中。从卷积神经网络的某一层中提取出的文档图像特征，往往维度较高。如果直接使用这种特征进行检索，则会存在消耗大量计算资源的问题。

有鉴于以上问题，本文在基于卷积神经网络的文档图像检索方法中，使用与上一章中相同的特征提取方法，但采用主成分分析的方法进行特征压缩，采用近似最近邻的逼近搜索方法加速检索的过程。通过这样的做法，在检索准确性和检索效率间达到平衡。

## 4.2 算法简述

基于卷积神经网络的文档图像检索方法，是从卷积神经网络中提取文档图像的深度信息表达，作为文档图像的特征表达。卷积神经网络广泛用于端到端的识别任务。在识别的过程中，随着信息逐层向前传递，图像的信息从以原始像素表示，逐渐抽象转化为以更高级的语义信息表示。基于卷积神经网络的文档图像特征，是在卷积网络前向传播的过程中，将其中某一层的输出作为特征提取出来，作为最终的特征值。

通过以上方法所获取到的特征值维度较高，不利于直接应用到检索过程中。出于提高检索效率和评估深度特征对特征压缩的鲁棒能力，我们采用主成分分析的方法对提取得到的深度特征进行降维。在对深度特征进行降维的前后，都采用L2范数进行归一化。即首先对提取到的深度特征采用L2范数进行归一化，然后采用主成分分析的方法进行降维，对降维后的特征再次采用L2范数进行归一化，最终得到的特征用于文档图像的检索。在检索阶段，我们直接使用欧式距离作为距离度量，采用最近邻方法进行相似性度量。

检索过程一般在欧式空间进行，特征矢量间的距离度量采用欧式距离的方法。最常见的检索方法是线性查找，即穷举法。在线性查找的方法中，需要将输入矢量与数据库中的所有矢量进行距离度量，然后按照从小到大排序，阈值以上的结果被视为相关并返回。线性查找的方法能得到最严谨的检索结果，但是在针对大型文档数据集进行检索时，无疑会耗费大量时间。因此，本文将近似最近邻的逼近搜索方法应用在文档图像检索任务中，解决大型数据集的检索问题。

## 4.3 总体流程

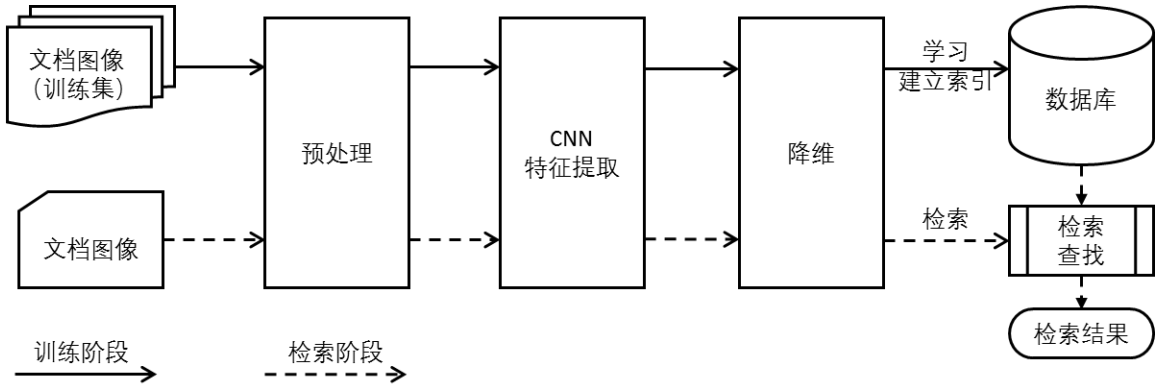



**图 4-1** 基于 CNN 的文档图像检索方法的整体框图

图 4-1 是本文基于 CNN 的文档图像检索方法的整体框图。总体流程可分为三个阶段，预训练阶段、训练阶段、检索阶段。

与本文提出的基于 CNN 的文档图像分类方法类似，在预训练阶段，我们对所选用的 CNN 模型进行训练。训练的方式因 CNN 模型而异，但总体上与在如目标识别等其他任务中的方法一致，将图片缩放至固定的尺寸，采取特定的数据扩充方式，采用经典梯度下降法对 CNN 模型的权值进行更新完成训练。

在训练阶段，我们通过预训练的 CNN 模型提取文档图像的特征。采取 L2 范数归一化、主成分分析、L2 范数归一化的流程，得到最终用于检索的特征表示。对所有训练的样本采用这样的方法提取特征表示，并将特征保存下来作为检索数据库。

在检索阶段，对输入的文档图像采取上述流程提取特征，与检索数据库中的特征进行相似度度量，返回相似度较高的文档图像。

在本文的实验中，如果检索返回的文档图像与输入文档图像属于同一类别，则认为该检索结果是正确的。因为所针对的数据集中样本数较多，不适宜考虑以对相似度设置固定阈值的方法返回相关结果，所以本文的文档图像检索系统对每张输入图像返回十个检索结果，表示十张文档图像，以平均查准率对检索系统的准确性进行评价。

## 4.4 特征降维

因为从卷积神经网络中提取的特征维数较高，且包含很多近似等于 0 的值，在用于文档图像检索任务前可以进行适当的数据降维，在保证准确性的前提下提高检索的性能。本文的基于卷积神经网络的文档图像检索方法，采用主成分分析的方法进行数据降维。

主成分分析（Principle Component Analysis，PCA）是一种无监督的数据压缩算法，能够极大提升无监督特征的学习速度。PCA 能够将原始数据的 $d$ 维特征空间映射到 $k$ 维特征空间，其中 $k < d$。映射得到的 $k$ 维特征是全新的正交特征，并且满足最大方差约束，利用这 $k$ 维特征可以很好地区分原始数据，表示最主要的成分。在某些应用中，算法的运行时间与输入数据的维数正相关。用降维后的数据参与运算，

算法运行速度将显著加快。低维特征是原始特征的小误差近似，对整体算法的效果影响很小，因此在这些场合使用 PCA 是合适的。

本文的基于卷积神经网络的文档图像检索方法，首先从预训练的 CNN 网络层中提取特征，视具体选用的 CNN 模型，所提取的特征维数在数千至数万维之间。在运用 PCA 降维前，首先使用 L2 范数对所提取的特征进行归一化。

所提取的原始特征用 $\vec{x}=[x_1,x_2,\cdots,x_d]^T$ 表示，其中 $d$ 是所提取特征的维数。进行归一化后，所得特征的每一维为：

$$a_i=\frac{x_i}{\sum_{k=1}^{d}x_k^{\,2}},\quad i=1,2,\cdots d$$

为使 PCA 算法能有效工作，通常期望特征有零均值等方差的性质，即所有特征的均值近似为零，不同特征的方差值相近。因此在对数据进行预处理时，将每个特征的取值范围规整为零均值和单位方差。但实际上因为卷积特征数据的平稳性，我们进行零均值操作。

在训练过程中，首先计算出 $d$ 维均值向量 $\vec{\mu}$，计算方式是对所有训练样本（总数为 $n$）的特征求平均。

$$\vec{\mu}=\frac{1}{n}\sum_{i=1}^{n}\vec{a}_i$$

然后计算大小为 $d\times d$ 的协方差矩阵 $\Sigma$。然后，计算得到协方差矩阵 $\Sigma$ 的本征值和本征向量，每个本征向量 $e_i$ 都对应一个本征值 $\lambda_i$。接着，选出对应最大 $k$ 个本征值的本征向量作为主成分方向。

数值上占主要的本征值只有很少的几个，这意味着 $k$ 的取值是取决于数据本身的子空间的内在维数，而剩下的 $d-k$ 维主要由噪声引起。现在我们构造一个 $d\times k$ 的矩阵 A，它的列由 $k$ 个本征向量组成。将原始数据按照下式投影到这个 $k$ 维子空间上就得到数据的主成分表示。

$$\vec{x}' = \mathrm{A}^{\mathrm{t}}\left(\vec{a} - \vec{\mu}\right)$$

在本文的实验中，$k$ 主要取 128，此时方差保留百分比大约为 90%。本文还对 $k$ 取 256、64、32、16、8 的结果进行分析。

同时，我们还采用白化的方法进行预处理。采用上述的主成分分析方法后，通过对降维后的数据按照方差进行缩放，降低数据间的冗余性，保证所有特征的方差相等。

## 4.5 搜索方法

考虑到实际应用需求，本文的文档图像检索方法进行的相似性查询是 K 近邻查询，并以平均查准率 mAP 评价系统性能。

K 近邻查询方法可根据检索方法分为两类：一类是线性扫描法，另一类是非线性扫描法。线性扫描法是将查询点与数据集中所有点一一比较。这种方法的优点是结果准确，所得到的结果即为实际情况下的 K 近邻，不存在误差；缺点也很明显，实际应用中高维数据在空间中往往以簇的形式呈现出聚类状分布，忽略具体的分布特征而进行穷举检索，无疑效率较低。非线性扫描法则考虑数据的分布特征，根据数据集建立数据索引，设计有效的索引结构加快检索速度，以达到快速匹配的目的。

但是非线性扫描法的返回结果有时候并不是严格意义上的 K 近邻，而是近似 K 近邻值。这是就需要根据实际需求，在严格意义的准确性和检索效率间保持平衡。为了获得更好的检索效率，本文的文档图像检索方法采用基于层次 K 均值树的近似最近邻的非线性扫描法。

层级 K 均值树结构最重要的结构参数是 K 均值聚类的类别数 $k$。对待扫描的数据集 $D$，构建层级 K 均值树的过程如下：

1）应用 K 均值聚类算法，将数据集 $D$ 聚成 $k$ 类，每类都是 K 均值树当前层的叶节点，用 $D_i$ 表示，对应聚类中心用 $\bar{D}_i$ 表示，其中 $i = 1, 2, \cdots, k$。

2）遍历 $D_i$，如果属于类别 $D_i$ 的样本数不止 $k$，则采用步骤 1 的方法继续应用 K 均值聚类算法，将 $D_i$ 划分成更小的 $k$ 类；如果样本数少于 $k$，则停止该叶节点的继续分裂。

3）重复步骤 1 和 2，直到每个叶节点中的样本数都不足 $k$。

层次 K 均值树生成完毕后，检索过程中只需要在层次树的每一层，与当前的至多 $k$ 个叶节点的聚类中心 $\bar{D}_i$ 进行比对，选择最相近的一个叶节点继续向下搜索，重复该搜索比对过程直到底层，即得到一次搜索结果。如果层次树的深度为 $n$，则该层次树中最多包含 $k^n$ 个数据，但完成一次基本搜索只需要进行 $k \times n$ 次比对，比对次数远少于线性检索的比对次数。

为了得到更高的精度，还需要对层次 K 均值树进行多次检索，多次检索以优先级队列的方式实现，即在完成一定次数的检索后，下一次检索在剩余未被检索过的叶节点中聚类中心与查询点距离最近的叶节点上进行。具体检索算法过程如下：

1）初始化优先级队列为空。

2）从层次树的根节点上，根据查询点完成一次检索，将所有没有深入的叶节点加入到优先级队列中，按照其聚类中心与查询点的距离从小到大排序。

3）从优先级队列中取出当前与查询点距离最小的叶节点，从该叶节点进行检索，同样将所有没有深入的叶节点加入到优先级队列中，按照其聚类中心与查询点的距离从小到大排序。

4）重复步骤 3，直到达到一定的迭代次数。

在类别参数 $k$ 的选取上，本文采用 Muja Marius 的自动参数选取方法[53]，首先在{16，32，64，128，256}中计算不同参数下的性能，然后采用多元函数最小值法进行调整。迭代次数则在确定 $k$ 值后，根据预设的精度确定。评估参数性能的损失函数表示为：

$$\cos t = \arg\min\left(s + w_b b\right)$$

其中，$s$ 表示检索所需时间，$b$ 表示生成树所需时间，$w_b$ 表示生成树所需时间的权重。本文方法更关注检索性能，因此将 $w_b$ 设置为 0.01。实验中预设精度默认为 90%。具体参数 $k$ 的取值可参见实验部分。

## 4.6 本章小结

本章主要介绍基于 CNN 和层次 K 均值树的文档图像检索方法。

文档图像的特征是从预训练过的卷积神经网络提取得到的，预训练可以在 ImageNet 等大型非文档图像数据集上进行，也可以在额外的文档图像数据集上进行。

通过卷积神经网络提取得到的特征维度较高，不宜直接用于文档图像检索。本文结合多次归一化和主成分分析的方法对提取到的高维特征进行压缩，同时保持较好的检索能力和距离度量特性。

近似最近邻的检索方法相比线性检索方法，能提供可比拟的精度和较高的效率。本文采用基于层次 K 均值树的快速检索方法，用于文档图像的检索，其参数需根据实际数据集特征情况确定。

# 第5章　　实验与分析

## 5.1 引言

本文提出基于卷积神经网络的文档图像特征提取方法，并将其应用在文档图像分类与检索的任务中。

本章通过与基于 SURF、BOW 和扩展空间金字塔匹配的文档图像特征提取方法和其他基于卷积神经网络的方法进行对比实验，验证本文基于卷积神经网络的文档图像特征提取方法的有效性；通过比较线性扫描法和近似最近邻逼近扫描法在检索任务中的查准率和时间效率，验证本文在检索任务中采取逼近扫描的做法的优越性。本章最后对卷积神经网络进行可视化，针对文档图像相关任务进行解释说明。

## 5.2 数据集描述

本章实验我们选取两个文本图像的数据集：Tobacco3482[23]和 RVL-CDIP[28]。这两个数据集都是 IIT-CDIP[27]的子集。IIT-CDIP 数据集由七百万份文档，总计四千万份扫描图像组成。虽然 IIT-CDIP 数据集包含大量图像和标记数据（包括文档类别、OCR 识别结果、签名等），但标注错误的程度也难以估计，所以我们选用这两个子数据集，关注文档类别信息。

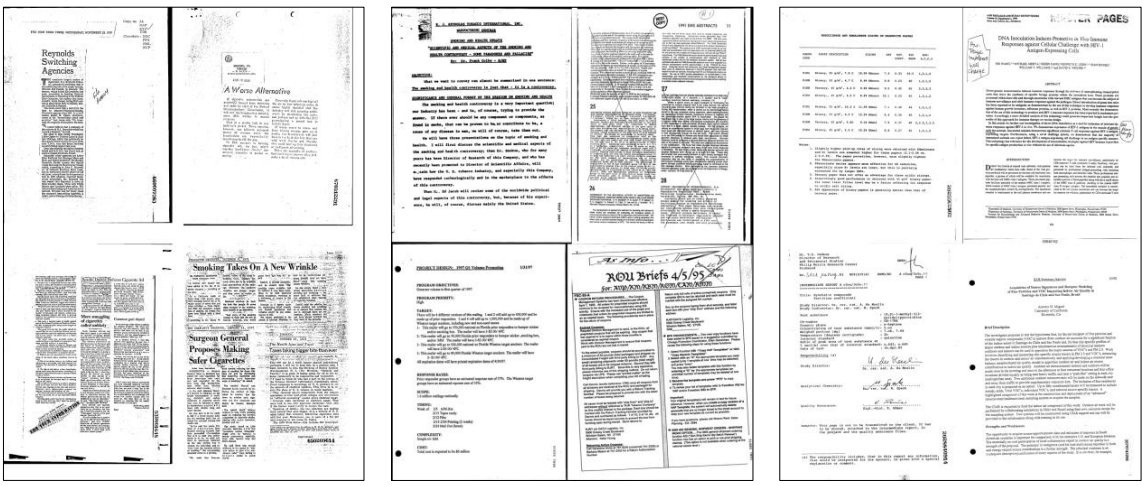


**图 5-1** Tobacco3482 数据集的部分样本，从左至右类别依次为 news、report、scientific

Tobacco3482：这个数据集由 10 种不同类别，总共 3482 张文档图像组成。图 5-1 中显示 news、report、scientific 的少数样本，可以发现该数据集中，有些同类样本彼此差异较大、类内距离较大；有些不同类样本彼此差异较小，类间距离较小。该数据集中，10 种类别的样本数目并不均等，详情分布如表 5-1 所示（为了实验方便，我们设置 10 类标签的序号从 0~9）。

**表 5-1** Tobacco3482 数据集样本分布

| 标签 | 类别 | 样本数 | 样本数占总数的比例 |
|---|---|---|---|
| 0 | ADVE | 230 | 0.066 |
| 1 | Email | 599 | 0.172 |
| 2 | Form | 431 | 0.124 |
| 3 | Letter | 567 | 0.163 |
| 4 | Memo | 620 | 0.178 |
| 5 | News | 188 | 0.054 |
| 6 | Note | 201 | 0.058 |
| 7 | Report | 265 | 0.076 |
| 8 | Resume | 120 | 0.034 |
| 9 | Scientific | 261 | 0.075 |

RVL-CDIP：研究人员对 IIT-CDIP 中的 16 个类别，每个类别随机采样 25000 张图片，且只保留标注的类别信息，忽略掉其他的标注信息，构成 RVL-CDIP 数据集，总共有四十万标记的样本图像。这 16 个类型分别是：letter、memo、email、file folder、form、handwritten、invoice、advertisement、budget、news article、presentation、scientific publication、questionnaire、resume、scientific report、specification。研究人员在公开数据集的同时，也公布其对样本的划分，作为参考。研究人员将数据集按照 8:1:1 分成训练集、验证集和测试集，即训练集样本数为 32 万，验证集和测试集分别有 4 万样本。

## 5.3 实验过程

### 5.3.1 实验对比方法

为了证明本文基于卷积神经网络的文档图像特征提取方法的有效性，我们设置多组对比实验，依据不同的性能指标进行比较说明。

作为对照，我们实现 Jayant Kumar 等的基于 SURF 特征、BOW 模型和扩展空间金字塔匹配的特征提取方法[23]（后文简称为 HVP-RF）和 Adam Harley 等的基于 AlexNet 的方法[28]（后文简称为 Harley-AlexNet）。其中，SURF 特征取 64 维，训练得到的 BOW 模型中单词数目取 300，水平方向划分阶数设为 0 或 2，垂直方向划分阶数设为 0 或 3，随机森林中决策树的数目取 500，树的每个节点随机选择变量的数量设为变量总数的平方根。表 5-2 是本文实验中用于对照的 HVP-RF 的实际配置，具体划分的示意图可以参考图 2-3，其中 H0V0 的配置，实际上就是在文档图像全局范围上应用 BOW 模型。

**表 5-2** HVP-RF 方法的实验配置

| 配置名称 | 特征总数 |
|---|---|
| H0V0 | 300 |
| H2V0 | 2100 |
| H0V3 | 4500 |
| H2V3 | 6300 |

为了证明框架的通用性，我们选用三个经典的模型进行实验，这三个模型分别是：AlexNet，GoogLeNet，VGG16-Net。

在实验的初始阶段，我们在 Tobacco3482 和 RVL-CDIP 上训练这三个模型。模型权值的初始化方式有两种，第一种是以正态随机值（非全零）的方法初始化模型权值，第二种是以预训练好的权值初始化模型。本文的所有实验，除额外说明外，都以第二种方法对网络进行初始化，预训练好的权值来自于对应模型在 ImageNet

2012 数据集上的训练结果。因为本文数据集的类别与 ImageNet 不同，所以需要修改网络中最后一层全连接的输出维度，并以正态随机值对其进行初始化。

在训练得到对特定数据集调整优化后的卷积神经网络模型后，采用第 3 章 和第 4 章 的方法将其应用到文档图像的分类和检索任务中。 本文针对文档图像的分类问题和检索问题，以及不同的应用场景，设置多组实验对本文提出的方法进行验证说明，具体的实验参数见结果分析，具体的对比过程如下所述。

文档图像分类任务的主要目的是，预测输入图像的类别。对于文档图像的分类任务，我们进行以下三组实验，进行验证说明。

在小样本数据集 Tobacco3482 上比较 HVP-RF 的方法和本文的基于卷积神经网络的方法，发现在预训练过后的网络上预测的精度高于其他方法，证明从针对目标分类的网络进行迁移学习是可行的，其通用特征本身就能较好的应用在其他分类任务中，而且效果也有所提升。

在大样本数据集 RVL-CDIP 上比较 HVP-RF 的方法和本文的基于卷积神经网络的方法，发现相比于小样本数据集上的结果，HVP-RF 的方法的预测精度下降很快，但基于卷积神经网络的方法反而有所提升。这证明基于浅层特征表达的方法还不能很好的应用于较大的数据集，而基于深层特征表达的方法则受益于训练样本数目的增长，反而表现更好，能更好的解决大量文档图像的分类问题。

在小样本数据集 Tobacoo3482 上，使用在大样本数据集上训练得到的网络提取文档图像特征，采用线性核的支撑向量机进行分类，也能得到不错的精度结果。这证明网络在学习大量的文档图像后，能提取对文档图像更针对性的特征，也能更好的应用于小样本的分类问题上。

文档图像检索任务的主要目的是，查找与输入图像最相似的图像并返回。本文的实验中，将返回图像与输入图像类别相同的结果认为查找正确，对每个输入图像，统计前 10 个样本并计算平均查准率，对检索的性能进行评估。对于文档图像检索任务，我们进行以下四组实验，进行验证说明。

首先，本文先比较从不同网络结构的不同层提取出的特征，在文档图像检索任务上的性能表现，确定后续实验所使用的主要网络结构与特征提取位置。

其次，本文比较 HVP-RF 的方法与基于卷积神经网络的方法，在文档图像检索任务上的性能优劣，证明基于卷积神经网络所提取的文档图像特征，能很好的用于文档图像的检索。

然后，本文采用主成分分析的方法，将基于卷积神经网络所提取的特征压缩至不同的维度，然后评价不同特征维度下的特征在检索任务上的性能优劣，证明基于卷积神经网络所提取的特征对特征压缩具有很好的鲁邦性。

最后，我们比较最近邻的线性搜索方式和近似最近邻的非线性搜索方式，在文档图像检索任务上的平均查准率与平均消耗时间，证明近似最近邻的非线性搜索方法更适用于大型数据集的检索任务，具有很好的实际应用价值。

卷积神经网络的可视化有多种方法，其主要目的是辅助人们以直观的方式理解它的效果，本文在最后采取反卷积的方法对网络进行可视化，进行一定的解释性说明。

### 5.3.2 实验设置

针对本章所选的两个数据集，我们分别做了以下划分处理，确定训练集、验证集和测试集。

对 Tobacco3482 数据集，因其样本数较少，我们任意选择 2000 张用于训练，剩余的样本用于测试。在 2000 张用于训练的样本中，我们按照 8:2 的比例划分训练集和验证集，即 1600 张样本用于训练，400 张样本用于测试。以上的划分重复进行 10 次，用于交叉验证，最后取精度均值作为最终结果。

对于 RVL-CDIP 数据集，因其样本数较多，且原文作者提供其进行实验的数据集划分，因此我们直接采用原文作者的划分。训练集、验证集和测试集的比例为 8:1:1，即训练集样本数为 320000，验证集和测试集的样本数均为 40000。

我们选用开源的Caffe[54]平台进行实验。本章所选的三种卷积神经网络结构，在Caffe 平台基础上都有开源的网络模型和基于 ImageNet2012 数据集进行预训练的网络模型权值。因此，出于实验的考虑，我们最终选择 Caffe 平台进行实验。

我们将所有图像压缩至256x256，并转化为LMDB格式，便于在Caffe平台上进行实验。所有训练过程在一台高性能GPU服务器上进行，但只使用一颗NVIDIA的型号为Tesla K40的GPU。

以下呈现的实验结果均是模型在测试集上的性能。

### 5.3.3 实验环境

本文主要使用C++和Python完成实验的核心运算和结果分析。

具体实验环境如表 5-3 所示。虽然本文实验是在一台较高性能的服务器上进行，但在实际的训练和测试过程中，我们仅使用其中一颗 CPU 和一颗 GPU，并没有考虑多线程和多并发对实验的影响。如果考虑多线程和多并发运算得到实验结果，时间性能上可能会比本文给出的实验结果更好。

**表 5-3** 本文实验环境

| | |
|---|---|
| CPU | 32 × Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz |
| GPU | 4 × NVIDIA K40c |
| 内存 | 128 GB |
| 操作系统 | CentOS Linux Release 7.2.1511 |
| 开发语言 | C++，Python |
| C++ 依赖库 | Caffe，OpenCV |
| Python 依赖库 | pycaffe, scikit-learn[55], pyflann[53] |

## 5.4 结果分析

### 5.4.1 分类实验：小样本集的训练结果对比

我们首先在小样本集Tobacco3482上进行实验。

对于本文提出的基于卷积神经网络的文档图像特征，我们采用三种不同的网络结构（AlexNet，GoogLeNet，VGG16-Net）进行训练，将原始网络结构中最后一个全连接层的输出维数由原始的 1000 减至 10，以适应小样本集的分类需要。在训练过程中，仅以正态随机值初始化最后一个全连接层的权值，其余层则以在 ImageNet 数据集上预训练好的网络权值进行初始化。因为 GoogLeNet 在训练过程中会在三个不同的深度下进行反向传播，因此，我们对 GoogLeNet 对应三种不同深度的全连接层都进行上述所描述的初始化行为。另外，我们还额外以正态随机值初始化 AlexNet 所有层的权值进行一次实验。

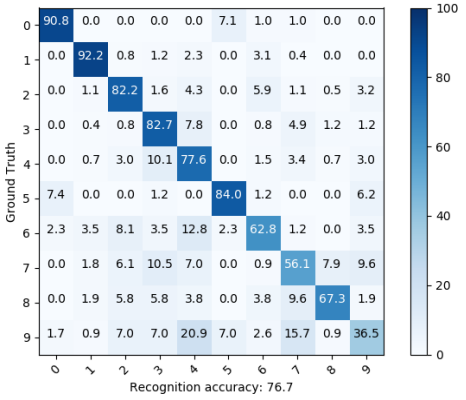
**表 5-4** 不同网络在小样本集上的分类精度

| 方法类别 | 方法设置 | 识别精度 |
|---|---|---|
| BOW | H0V0 | 0.645 |
| BOW | H0V3 | 0.679 |
| BOW | H2V0 | 0.652 |
| BOW | H2V3 | 0.681 |
| CNN | AlexNet （Random） | 0.634 |
| CNN | AlexNet （ImageNet） | 0.767 |
| CNN | Harley-AlexNet（ImageNet） | 0.799 |
| CNN | GoogLeNet （ImageNet） | **0.827** |
| CNN | VGG16-Net （ImageNet） | 0.821 |

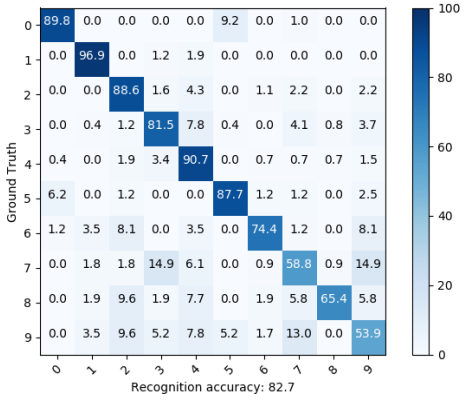
表 5-4 列举出不同方法在 Tobacco3482 数据集上的精度。其中的方法设置一栏，括号里有 ImageNet 表示对应的网络模型主要是以在 ImageNet 上训练得到的模型参数进行初始化，括号里有 Random 表示对应的网络模型是以正态随机值进行初始化。

从实验结果中，我们可以看到，HVP-RF 的不同配置在识别精度上相差不大，但在利用水平垂直分割来包含文档图像结构信息的改进方法中，无论参数如何，相较于忽略空间信息的方法，精度总是有所提升。

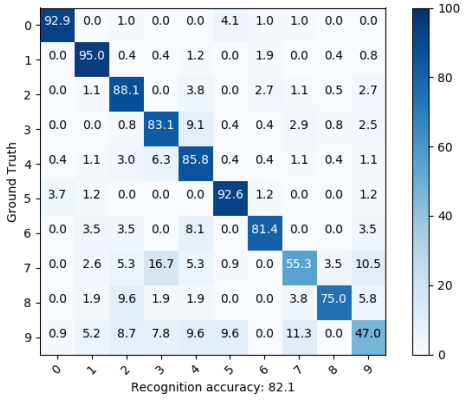
基于卷积神经网络的方法中，以正态随机值初始化的网络，在识别精度上不如 HVP-RF 的方法，但是所有在 ImageNet 上进行预训练的网络，识别精度都远远超过。这样的实现现象表明，在有足够的训练数据的前提下，基于卷积神经网络的方法，能够在效果上超越传统方法。




(a) AlexNet



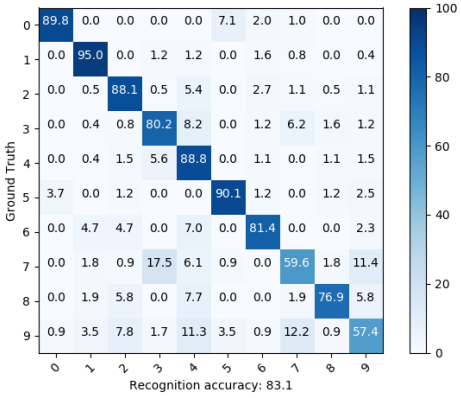

(b) GoogLeNet



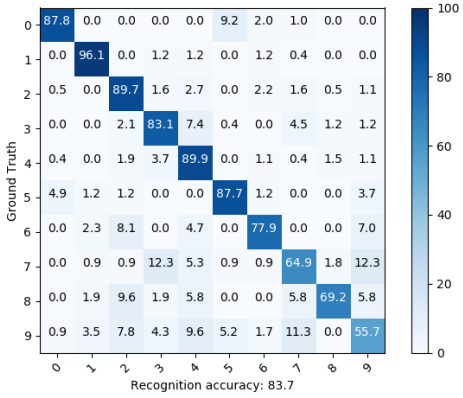

(c) VGG16-Net

**图 5-2** 不同网络在小样本集上的分类情况

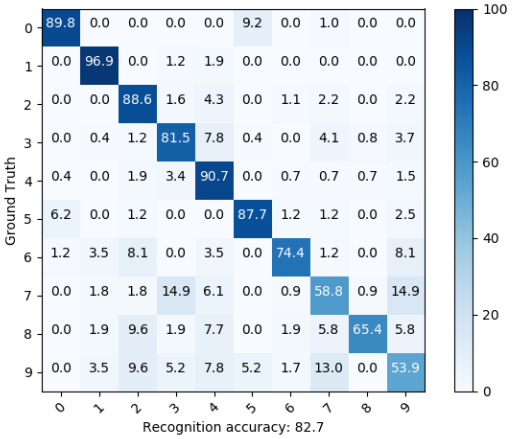
图 5-2 绘制出 AlexNet、GoogLeNet、VGG16-Net 在小样本集上的分类情况。可以发现，不同网络对不同类别样本的分类能力不同，但总体的分辨能力是相似的。




(a) GoogLeNet-INCEPTION-4A




(b) GoogLeNet-INCEPTION-4D




(c) GoogLeNet-INCEPTION-5B

**图 5-3** GoogLeNet 的不同层在小样本集上训练后的分类情况

在对小样本集的训练实验过程中，如图 5-3 所示，我们也发现 GoogLeNet 上不同层次的表达在识别阶段的精度不同。从理论上而言，更深层次的表达应该能够得到更准确的结果。但是在我们的实验中，基于最深层表达的识别性能反而不如较浅的层次。不过这可能是由于训练样本数不足、样本不均衡和一些其他误差引起的。

### 5.4.2 分类实验：大样本集的训练结果对比

前一小节中列举出各种方法在小样本集 Tobacco3482 的分类实验结果，本文提出的特征提取方法，从预训练的卷积神经网络中提取特征，在前一小节实验中，这些卷积神经网络的权值主要来自于在 ImageNet 上预训练的结果，换言之，所提取的

特征更针对目标分类的问题，是更通用的特征。但实际上，文档图像相比于自然图像有独特的特征，如果能提取到更针对文档图像的特征，实际的分类性能应该还有所提升。

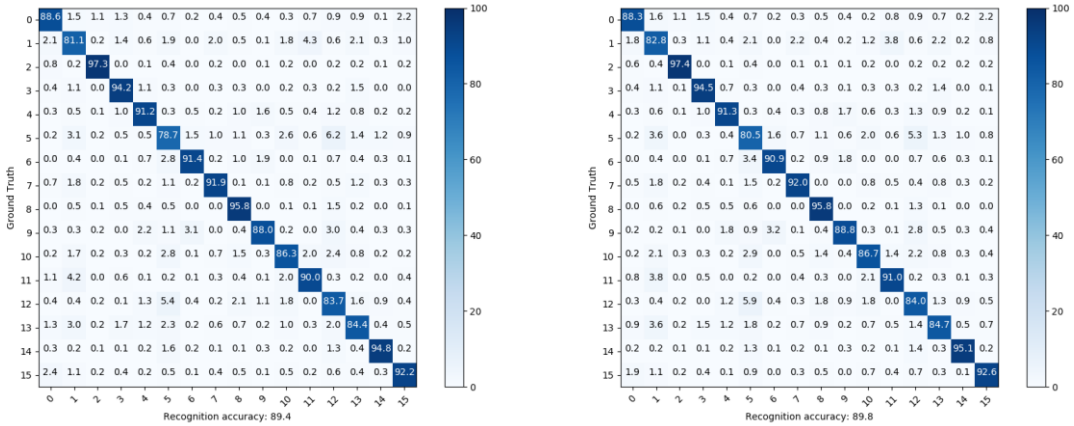
在本节的实验中，我们采取和上一节类似的训练方法，在大样本集 RVL-CDIP 上采用不同的方法进行训练分类，表 5-5 列举了不同方法在 RVL-CDIP 上训练后的分类精度。

**表 5-5** 不同方法在 RVL-CDIP 上训练后的分类精度

| 方法类别 | 方法设置 | 识别精度 |
|---|---|---|
| BOW | H0V0 | 0.446 |
| BOW | H0V3 | 0.483 |
| BOW | H2V0 | 0.461 |
| BOW | H2V3 | 0.493 |
| CNN | AlexNet | 0.895 |
| CNN | Harley-AlexNet | 0.898 |
| CNN | GoogLeNet | **0.900** |

RVL-CDIP 数据集与 Tobacco3482 数据集相比，不仅样本数由 4000 张不到增长到 400000 张，类别数也由 10 增长到 16。因此，在分类问题上，RVL-CDIP 比 Tobacco3482 更具挑战性。

从表 5-5，我们可以直观地发现，相比于小样本集 Tobacco3482 上的分类精度，在大样本集 RVL-CDIP 上，基于 HVP-RF 的方法的分类精度都下降很多，平均从 65%左右下降到 50%不到，但是基于卷积神经网络的方法的分类精度却增长很多，平均从 80%左右增长到 90%左右。这样的实验结果表明，基于卷积神经网络的方法能更好地解决大量数据的问题。在有大量训练样本的前提下，基于卷积神经网络的方法能发挥更好的性能。




(a) GoogLeNet-INCEPTION-4A

(b) GoogLeNet-INCEPTION-4D

(c) GoogLeNet-INCEPTION-5B

**图 5-4** GoogLeNet 的不同层在大样本集上训练后的分类情况

类似上一节，我们对 GoogLeNet 的不同层的分类性能进行分析。图 5-4 中绘制出经过训练后的 GoogLeNet 的不同层在 RVL-CDIP 的测试集上的分类情况。从浅层表达到深层表达，其分类准确性依次递增，由 89.4%增长到 90.0%，这样的结果比较符合对卷积神经网络层的一般性理解，而上一节的实验结果，可能是由于训练样本不够充分导致的中间状态，或者是测试样本数目较少导致的量化误差。通常在实际应用中，深层表达相比于浅层表达，具备更好的分类能力。

### 5.4.3 分类特征：特征迁移的性能对比

在实际的文档图像需求中，很有可能因为保密等其他要求，碰到没有足够训练样本的情况。基于 HVP-RF 的方法，可以预先通过对其他的文档图像进行特征提取构建字典的方法，应用在这样的场景中。而本文的基于卷积神经网络的文档图像特征，也可以通过类似的途径，在其他的文档图像上进行训练，然后应用在同样的场景中。

上一节的实验证明基于卷积神经网络的方法能从大量文档图像中学到更鲁棒的特征表达。本节实验中，我们将采用在 RVL-CDIP 数据集上优化微调后的网络进行特征提取，对小样本集 Tobacco3482 采用支撑向量机进行分类训练。一般而言，在运用支撑向量机的算法前，需要对输入特征矢量进行预处理，将每一维的特征都缩放至-1 与 1 之间，或 0 与 1 之间。在我们的实验中，如果没有进行特征压缩，则直接对每一维缩放至-1 与 1 之间，然后调用训练算法；如果进行特征压缩，则首先对提取的特征进行 L2 范数归一化，然后采用主成分分析将特征压缩至特定维数，之后再对每一维缩放至-1 与 1 之间，然后调用算法。

**表 5-6** 不同卷积网络层提取的特征在 Tobacco3482 上的分类精度

| 特征提取位置 | 是否压缩 | 特征维数 | 分类精度 |
|---|---|---|---|
| AlexNet-FC6 | 否 | 4096 | 0.857 |
| AlexNet-FC6 | 是 | 128 | 0.852 |
| AlexNet-FC7 | 否 | **4096** | **0.895** |
| AlexNet-FC7 | 是 | 128 | 0.884 |
| GoogLeNet-INCEPTION-4A | 是 | 128 | 0.861 |
| GoogLeNet-INCEPTION-4D | 是 | 128 | 0.878 |
| GoogLeNet-INCEPTION-5B | 是 | **128** | **0.905** |

表 5-6 列举出从不同卷积网络层提取的特征，在 Tobacco3482 数据集上采用支撑向量机的算法训练得到的分类精度。与第一小节的实验结果表 5-4 对比，我们可以发现，从在大量文档图像上优化微调后的网络上提取特征，降维压缩后采用支撑向量机的分类方法，在分类精度上也要优于传统的方法。这样的实验结果表明，在大

量文档图像上优化微调后的网络，其特征更加鲁邦，能很好的迁移到类似的文档图像分类任务中，这也表明，本文的基于卷积神经网络的文档图像特征提取方法，能很好的应用在实际的需求中。

### 5.4.4 检索实验：特征提取位置的对比

本文采用的基于卷积神经网络的特征提取方法，通过经过预训练的卷积神经网络提取特征，特征来源于卷积神经网络中某一层的输出。

在本节实验中，我们比较在 ImageNet 数据集上预训练过的不同网络结构的不同层的输出，用于文档图像检索任务时的性能对比，以此确定后续实验主要采用的网络结构与特征提取位置。

本节实验的所有网络结构，都仅在 ImageNet 上预训练过，即没有在 Tobacco3482 和 RVL-CDIP 上进行优化微调。检索实验在 Tobacco3482 的训练集和测试集上进行，从卷积神经网络中提取的特征被 PCA 压缩至 128 维，用于检索。所得平均查准率为仅考虑 1 个候选项时的结果。

**表 5-7** 不同特征提取位置的第一候选平均查准率

| 网络结构 | 网络层 | 平均查准率 |
|---|---|---|
| AlexNet | **FC6** | **0.627** |
| AlexNet | FC7 | 0.605 |
| GoogLeNet | **INCEPTION-4A** | **0.687** |
| GoogLeNet | INCEPTION-4D | 0.625 |
| GoogLeNet | INCEPTION-5B | 0.581 |
| VGG16-Net | **FC6** | **0.650** |
| VGG16-Net | FC7 | 0.629 |

表 5-7 列举从不同网络层提取的特征，在 Tobacco3482 的训练集和测试集上的检索查准率。我们可以发现，在 AlexNet 和 VGG16-Net 的网络中，从 FC6 提取的特征比从 FC7 提取的特征表达能力稍好，这样的结果与 Artem Babenko 等的实验结论[51]一致。

在本文后续的实验中，主要从 AlexNet 网络的这两个全连接层提取特征，用于实验。

### 5.4.5 检索实验：特征提取方法的对比

在本节实验中，我们比较基于 HVP-RF 的特征提取方法和基于卷积神经网络的文档图像特征提取方法的检索性能。

首先，我们比较仅在 ImageNet 上预训练的网络，和在 RVL-CDIP 数据集上优化微调后的网络，所提取的特征的性能差异。与上节实验相同，所提取的特征压缩至 128 维后，用于检索；平均查准率为在第一候选上的统计结果；针对 Tobacco3482 的训练集和测试集进行结果统计。

**表 5-8** 网络优化微调前后提取的特征性能对比

| 网络结构 | 网络层 | 是否在 RVL-CDIP 上训练 | 平均查准率 |
|---|---|---|---|
| AlexNet | FC6 | 否 | 0.627 |
| AlexNet | FC6 | 是 | 0.786 |
| AlexNet | FC7 | 否 | 0.605 |
| AlexNet | **FC7** | **是** | **0.845** |
| GoogLeNet | INCEPTION-4A | 否 | 0.687 |
| GoogLeNet | INCEPTION-4A | 是 | 0.809 |
| GoogLeNet | INCEPTION-4D | 否 | 0.625 |
| GoogLeNet | INCEPTION-4D | 是 | 0.835 |
| GoogLeNet | INCEPTION-5B | 否 | 0.581 |
| GoogLeNet | **INCEPTION-5B** | **是** | **0.866** |

表 5-8 列举出各种网络结构在对 RVL-CDIP 上微调前后，提取的特征在检索任务中的性能表现。我们可以发现，对于同样的网络和特征提取位置，经过在大型文档图像数据集上的优化微调后，平均查准率都有很大提升。但与先前较浅层的特征表达（FC6）查准率较高不同，在经过优化微调后，较深层的特征表达（FC7）查

准率较高。这样的现象与 Artem Babenko 的结论[51]不同。可能是因为 RVL-CDIP 与 Tobacco3482 中的样本比较相似，或者文档图像所特有的类别结构，所以较深层的特征表达的区分能力也很强。我们的实验对 Adam Harley 等的实验[28]进行补充，得到提取文档图像的更佳位置。

接下来，我们比较 HVP-RF 的特征与优化微调后的网络特征的检索性能。

**表 5-9** 不同方法的平均查准率比较

| 方法名称 | 平均查准率 |
|---|---|
| **BOW (H0V0)** | **0.546** |
| BOW (H0V3) | 0.502 |
| BOW (H2V0) | 0.478 |
| BOW (H2V3) | 0.494 |
| AlexNet (FC6) | 0.786 |
| **AlexNet (FC7)** | **0.845** |
| Harley-AlexNet (FC6) | 0.817 |
| GoogLeNet (INCEPTION-4A) | 0.809 |
| GoogLeNet (INCEPTION-4D) | 0.835 |
| **GoogLeNet (INCEPTION-5B)** | **0.866** |

表 5-9 中列举不同方法在 Tobacco3482 的测试集上检索的第一平均查准率。我们可以发现，在针对文档图像进行优化微调的网络上提取特征，其检索能力相较于 HVP-RF 有很大提升，这也证明本文提出的基于卷积神经网络的文档图像特征在文档图像检索任务中的有效性。

### 5.4.6 检索实验：特征压缩的对比

对于图像检索的任务，假定索引中图像的数目一定，通常而言表示图像的特征维数越高，检索花费的时间更长；表示图像的特征维数越低，检索的查准率就会越低。因此，在特征的维数与检索的查准率中找到平衡点，就成为文档图像检索任务中的难题。

已有其他领域的研究表明，基于卷积神经网络的特征对特征压缩的行为比较鲁邦，即对特征压缩有较好的信息保持能力。本节的实验中，对从卷积神经网络中提取到的特征分别压缩至 256、128、64、32、16、8 维，统计其检索性能。表 5-10 列举出通过主成分分析将数据压缩至不同维度，保留方差百分比。

**表 5-10** 对 AlexNet 网络特征进行降维，不同维度的保留方差百分比

| 维数 | 保留方差百分比 |
|---|---|
| 256 | 0.909 |
| 128 | 0.878 |
| 64 | 0.841 |
| 32 | 0.792 |
| 16 | 0.710 |
| 8 | 0.522 |

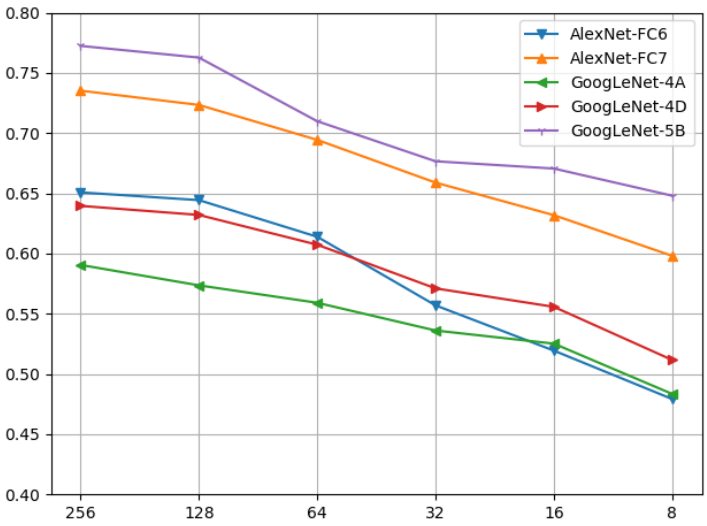



**图 5-5** 不同特征随特征压缩维度变化的平均查准率曲线

图 5-5 绘制出从不同网络的不同位置提取的特征，经不同的降维压缩后，用于文档图像检索的平均查准率变化。我们可以发现，随着特征压缩维度的降低，平均查准率也不断下降，但整体而言，直到压缩至 32 维以下，性能才有了急剧的下降。

这也证明，基于卷积神经网络的文档图像特征在特征压缩中有较好的保持信息的能力，也因此更适用于文档图像的检索任务。

### 5.4.7 检索实验：搜索方法的对比

本文采用的基于卷积神经网络的特征提取方法，所提取的特征在欧式空间中表示，彼此的距离以欧式距离度量。采用最近邻的线性搜索方法，则会将待检索的特征，与数据库中所有的特征进行距离计算，按照最近邻的排序返回；采用近似最近邻的非线性搜索方法，则首先对数据库中所有的特征采取特定算法分块，在检索阶段减少实际比较的次数，从而提高检索效率，但有可能降低检索的准确性。

在本节实验中，我们比较最近邻的线性搜索方法与基于层次 K 均值树的近似最近邻的非线性搜索方法在文档图像检索应用中的性能差异。我们分别提取 AlexNet 的第一个全连接层 FC6 和第二个全连接层 FC7 的输出作为文档图像特征；AlexNet 的网络权值则选用在 ImageNet 数据集预训练的权值，和在 RVL-CDIP 数据集上优化微调的权值；特征压缩的尺寸选择 16，32，64，128，256。

**表 5-11** 从在 ImageNet 预训练的 AlexNet 提取特征（128 维）在 RVL-CDIP 测试集的平均查准率

| 特征提取节点 | 检索方法 | 1 候选 | 3 候选 | 5 候选 | 10 候选 |
|---|---|---|---|---|---|
| FC6 | NNS | 0.775 | 0.636 | 0.611 | 0.530 |
| FC6 | ANNS | 0.769 | 0.629 | 0.582 | 0.524 |
| FC7 | NNS | 0.757 | 0.615 | 0.566 | 0.509 |
| FC7 | ANNS | 0.749 | 0.607 | 0.559 | 0.503 |

**表 5-12** 从在 RVL-CDIP 优化的 AlexNet 提取特征（128 维）在 RVL-CDIP 测试集的平均查准率

| 特征提取节点 | 检索方法 | 1 候选 | 3 候选 | 5 候选 | 10 候选 |
|---|---|---|---|---|---|
| FC6 | NNS | 0.869 | 0.785 | 0.757 | 0.721 |
| FC6 | ANNS | 0.860 | 0.778 | 0.750 | 0.714 |
| FC7 | NNS | 0.884 | 0.839 | 0.825 | 0.809 |

| | | | | | |
|---|---|---|---|---|---|
| FC7 | ANNS | 0.879 | 0.840 | 0.828 | 0.815 |

表 5-11 列举从在 ImageNet 预训练的 AlexNet 网络提取特征并压缩至 128 维，采取不同搜索方法在 RVL-CDIP 数据集上得到的平均查准率。表 5-12 列举从在 RVL-CDIP 数据集上优化后的 AlexNet 网络提取特征并压缩至 128 维，采取不同搜索方法在 RVL-CDIP 数据集上得到的平均查准率。其中，NNS（Nearest Neighbor Search）表示最近邻的线性搜索，ANNS（Approximate Nearest Neighbor Search）表示近似最近邻的快速搜索。

从表 5-11 和表 5-12 中，我们可以发现，无论所选用的 AlexNet 网络权值如何，用于提取特征的网络层位置如何，近似最近邻的快速搜索方法得到的平均查准率，仅比最近邻的线性搜索方法得到的平均差准率低少于 1%。在实际的应用中，这样的性能误差是可以接受的。

**表 5-13** 从在 RVL-CDIP 优化的 AlexNet 提取特征并压缩至不同维度的查询时间

| 特征提取节点 | 检索方法 | 16 维 | 32 维 | 64 维 | 128 维 | 256 维 |
|---|---|---|---|---|---|---|
| FC6 | NNS | 31.5ms | 30.0ms | 26.3ms | 25.2ms | 24.2ms |
| FC6 | ANNS | 0.07ms | 0.10ms | 0.13ms | 0.26ms | 0.70ms |
| FC7 | NNS | 28.1ms | 19.4ms | 23.9ms | 27.7ms | 20.6ms |
| FC7 | ANNS | 0.05ms | 0.05ms | 0.10ms | 0.23ms | 0.53ms |

表 5-13 列举出从在 RVL-CDIP 优化的 AlexNet 提取特征并压缩至不同维度，采取不同的搜索策略所耗费的平均查询时间。所得平均查询时间，是指在有 320000 条记录的数据库中，进行 40000 次查询所耗费的平均时间。表中的平均时间仅指查询操作的时间，不包含从原始图像中提取特征并进行降维的时间。

从表 5-13 可以明显发现，即使在各种特征压缩维度下，采取近似最近邻的快速搜索方法，比采取最近邻的线性搜索方法，在搜索时间上至少要少一个量级。结合之前的结论，采取近似最近邻的快速搜索方法得到的平均查准率比最近邻的线性搜索仅低少于 1%，我们可以肯定在实际大型文档图像检索系统中采取近似最近邻的快速搜索方法的有效性。

### 5.4.8 可视化实验

从本章中文档图像的分类实验结果和检索实验结果，可以发现在文档图像的领域采取基于卷积神经网络的方法，能得到更优的结果。本节通过对卷积神经网络进行可视化，对卷积网络对文档图像的操作进行一定的解释说明。

对卷积网络进行可视化有多种方式。最直接的方法是可视化其各层网络的权值，但这种可视化方法只在卷积网络的浅层有较合理的说明能力。本文采取的可视化方法，对卷积网络中受输入图像的激活情况进行可视化，也参考反卷积[56]的方法进行可视化。

本节实验基于开源的 Caffe 平台和 Deep Visualization Toolbox[57]，对在 RVL-CDIP 数据集上优化微调后的 AlexNet 网络进行可视化。因为参考文献中对所采用的算法已经描述得十分清楚，此处不再对具体的计算处理方法进行描述，仅给出实验结果。

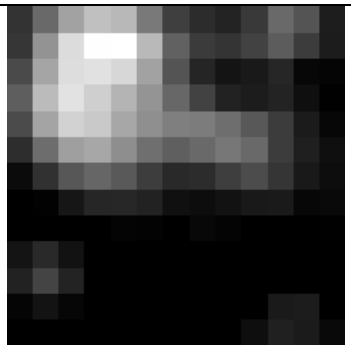
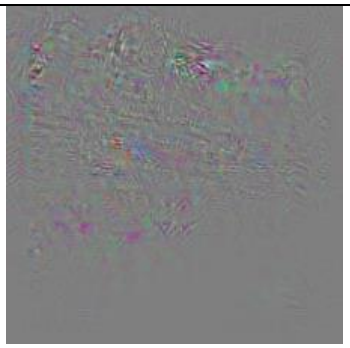
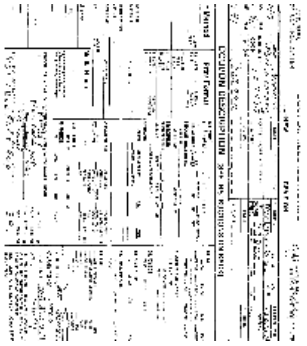
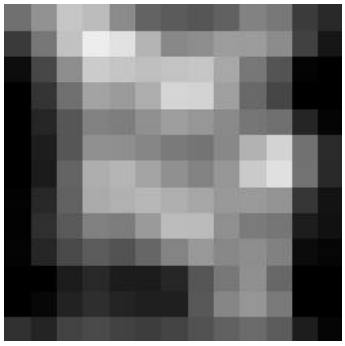
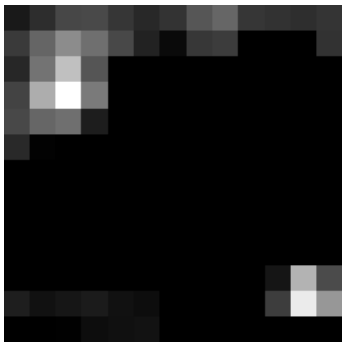
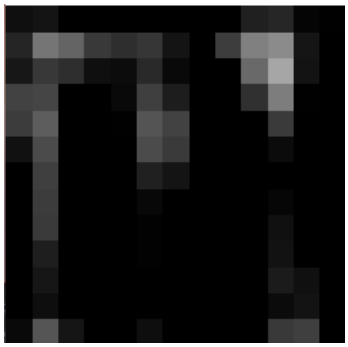
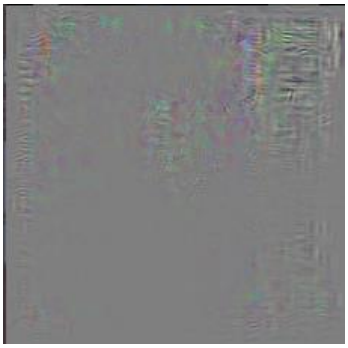
表 5-14 中列出本文对 AlexNet 的卷积层可视化的结果。第一列为输入的原始文档图像；第二列为第五个卷积层 CONV5 的输出，CONV5 层总共有 256 核输出，本文仅挑选部分有代表性的输出进行可视化展示；第三列为根据[56]中的反卷积方法，由特定的卷积位置反卷积得到的结果；第四列为第二、三列对应的卷积核位置，如 CONV5 168 表示第五个卷积层 CONV5 中索引号为 168（以 0 为起始索引）的卷积核输出。

从 AlexNet 的卷积层进行可视化的结果可以更进一步的理解卷积神经网络的作用。首先通过激励的可视化可以发现，基于卷积神经网络的文档图像分类和检索，依靠的是文档图像的版面布局信息。例如，广告中产生较大激励的区域是图像，包括大幅图像和小幅图像；论文中产生较大激励的区域是空白位置，实际是论文类别所特有的几种排版格式。其次，通过反卷积的可视化可以发现，对于文档图像而言，实际有效的是区域信息，而不是细节信息。在针对表格等类别的文档图像设计的分类和检索应用中，往往需要针对性地考虑表格线位置、表格主题等细节信息，而从

表格的激励和反卷积可视化结果可以发现，其实从整体的版面布局就可以对表格进行分类。

其实，还有其他对卷积神经网络进行可视化的方法，如[57]中从所有图像中筛选出对每个单元造成最大激活的图像，依次表明每个单元实际针对的内容和功能。因为计算资源有限，在本文的可视化实验中没有进行这项实验，但从目前的实验结果推测，应能得到和原文类似的实验结果。

**表 5-14** 对 AlexNet 的卷积层可视化的结果

| 原始图像 | 激励 | 反卷积 | 位置 |
|---|---|---|---|
|  |  |  | CONV5 233 |
|  |  |  | CONV5 230 |
|  |  |  | CONV5 203 |
|  |  |  | CONV5 145 |

# 第6章　总结和展望

## 6.1 论文内容总结

随着数字时代的发展，越来越多的信息以文档图像的形式进行存储和传播，文档图像的使用已经融入到我们的生活当中。基于词袋模型和扩展空间匹配的方法已经无法很好地扩展解决大型文档图像数据集的分类和检索任务上。本文利用卷积神经网络强大的自主特征学习能力，提出基于卷积神经网络的文档图像特征提取方法，并将其应用在文档图像的分类和检索任务中。本文所做的工作可以归纳为：

第一，提出基于卷积神经网络的文档图像分类方法，利用训练后的 CNN 提取文档图像的特征，采用线性核的支撑向量机进行分类。在该方法中，本文还根据实验比较不同 CNN 模型的不同网络层提取特征用于分类的性能差异，给出较好的特征提取位置。

第二，提出基于卷积神经网络的文档图像检索方法，同样利用训练后的 CNN 提取文档图像特征，采用主成分分析的方法进行数据降维，降维前后都将特征向量归一化为单位矢量，比较穷举搜索和近似最近邻逼近搜索方法在检索中的准确性和效率，最终选择采用近似最近邻逼近搜索方法完成检索任务。在检索任务中，本文也对不同 CNN 模型的不同网络层的特征表达性能进行比较，发现与传统结果不一致的现象，并给出一般性的解释。

第三，采用反卷积等方法对卷积神经网络进行可视化，给出一般性的评价。

## 6.2 未来研究方向

基于深度学习的方法是目前的研究热点，在该领域内算法的更新也很快。虽然本文的基于卷积神经网络的分类和检索方法，相对于传统方法，性能已经有很大提高，但实际上还存在诸多不足。主要表现如下：

1）本文提出的基于卷积神经网络的方法，实际上可以采用更多不同的模型。本文实验出于方便比较的目的，仅选择十分经典的 AlexNet、GoogLeNet、VGG16-

Net 网络模型用于学习训练。随着科学技术的发展，已经有很多新的理论和模型出现，例如全卷积网络，深度残差网络等，都也可以将其应用到文档图像分类和检索的任务中来做些实验。

2）在文档图像分类和检索领域，其实还存在一些特异化的需求，例如对小数据集中版面结构差异细微的样本进行精确分类。这些特异化的需求缺少公开的数据集，因此本文没有多做研究，在实际应用中可以再深入研究。

3）本文对卷积神经网络进行可视化的方法比较粗浅，这方面也有很多工作可以进行，可以帮助更好的使用卷积网络进行文档图像相关的任务。

# 致谢

时光荏苒，岁月匆匆，转眼间三年的研究生生活即将结束。

回顾这三年，有增长知识的喜悦，有师生关爱的感动，也有面对学业未来的迷茫，所有的这些都值得自己在将来的岁月中回味。这三年的科研生活，让我学会如何发现问题，有价值的问题要从实际的需求出发；也让我学会运用知识解决问题，解决的方法多样，总可以找到更好的一种；也让我学会持续学习和交流，这样才能不断增长自己的知识。我的这些成长和进步，离不开研究生生涯中遇到的每位前辈的帮助，在此，对大家表示深深的谢意。

首先，要感谢我的导师——陈友斌教授。在学习初期对研究不甚了解到后来能够顺利完成各种任务，我每一步的成长都离不开陈老师的悉心指导。在研究生生活期间，陈老师提供我很多尝试的机会，并且提供很好的学习工作环境。如果没有陈老师的帮助，这篇论文的很多实验都会难以进行。

其次，我要感谢实验室的所有成员和湖北微模式科技有限公司优秀的算法工程师们。每一次闲暇时的交流，都会让我受益匪浅并不断反思，学习工作能力得到很大提升。

最后，要对我的父母表示深深地感谢。在我成长的过程中，他们付出良多。他们用最朴实和真切的方式教育我如何过好自己的人生。他们脚踏实地的生活态度，让我能以坚韧的心态面对学习生活中的各种问题。对此，我十分感激，他们的支持和鼓励一直是我进步的动力。

# 参考文献

[1] Bengio Y. Learning Deep Architectures for AI[J]. Foundations and Trends® in Machine Learning, 2009, 2(1): 1–127.

[2] Deng L, Yu D. Deep Learning: Methods and Applications[J]. Microsoft Research, 2014.

[3] Salton G, Allan J, Buckley C, et al. Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts[J]. Science, 1994, 264(5164): 1421–1426.

[4] Takeda K, Kise K, Iwamura M. Real-Time Document Image Retrieval on a Smartphone[C]//2012 10th IAPR International Workshop on Document Analysis Systems. 2012: 225–229.

[5] Ohta M, Takasu A, Adachi J. Retrieval methods for English-text with missrecognized OCR characters[C]//Proceedings of the Fourth International Conference on Document Analysis and Recognition. 1997, 2: 950–956 vol.2.

[6] Noce L, Gallo I, Zamberletti A, et al. Embedded Textual Content for Document Image Classification with Convolutional Neural Networks[C]//Proceedings of the 2016 ACM Symposium on Document Engineering. New York, NY, USA: ACM, 2016: 165–173.

[7] Marinai S, Miotti B, Soda G. Digital Libraries and Document Image Retrieval Techniques: A Survey[G]//Biba M, Xhafa F. Learning Structure and Schemas from Documents. Springer Berlin Heidelberg, 2011: 181–204.

[8] Zhu G, Zheng Y, Doermann D, et al. Signature Detection and Matching for Document Image Retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2015–2031.

[9] Jain R, Doermann D. Logo Retrieval in Document Images[C]//2012 10th IAPR International Workshop on Document Analysis Systems. 2012: 135–139.

[10] Chen S, He Y, Sun J, et al. Structured document classification by matching local salient features[C]//Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). 2012: 653–656.

[11] Zhu G, Doermann D. Logo Matching for Document Image Retrieval[C]//2009 10th International Conference on Document Analysis and Recognition. 2009: 606–610.

[12] Byun Y, Lee Y. Form Classification Using DP Matching[C]//Proceedings of the 2000 ACM Symposium on Applied Computing - Volume 1. New York, NY, USA: ACM, 2000: 1–4.

[13] Shin C, Doermann D. Document Image Retrieval Based on Layout Structural Similarity[J]. 2008.

[14] Collins-thompson K, Nickolov R. A Clustering-Based Algorithm for Automatic Document Separation[J]. 2002.

[15] Joutel G, Eglin V, Bres S, et al. Curvelets Based Queries for CBIR Application in Handwriting Collections[C]//Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). 2007, 2: 649–653.

[16] Wallraven C, Caputo B, Graf A. Recognition with local features: the kernel recipe[C]//Proceedings Ninth IEEE International Conference on Computer Vision. 2003: 257–264 vol.1.

[17] Quelhas P, Monay F, Odobez J M, et al. Modeling scenes with local descriptors and latent aspects[C]//Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. 2005, 1: 883–890 Vol. 1.

[18] Barbu E, Héroux P, Adam S, et al. Using Bags of Symbols for Automatic Indexing of Graphical Document Image Databases[C]//Graphics Recognition. Ten Years Review and Future Perspectives. Springer, Berlin, Heidelberg, 2005: 195–205.

[19] Kumar J, Prasad R, Cao H, et al. Shape codebook based handwritten and machine printed text zone extraction[C]//2011, 7874: 787406-787406–8.

[20] Lazebnik S, Schmid C, Ponce J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). 2006, 2: 2169–2178.

[21] Kumar J, Ye P, Doermann D. Learning document structure for retrieval and classification[C]//Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). 2012: 1558–1561.

[22] Kumar J, Doermann D. Unsupervised Classification of Structurally Similar Document Images[C]//2013 12th International Conference on Document Analysis and Recognition. 2013: 1225–1229.

[23] Kumar J, Ye P, Doermann D. Structural similarity for document image classification and retrieval[J]. Pattern Recognition Letters, 2014, 43: 119–126.

[24] Kang L, Kumar J, Ye P, et al. Convolutional Neural Networks for Document Image Classification[C]//2014 22nd International Conference on Pattern Recognition. 2014: 3168–3172.

[25] Afzal M Z, Capobianco S, Malik M I, et al. Deepdocclassifier: Document classification with deep Convolutional Neural Network[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR). 2015: 1111–1115.

[26] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[G]//Pereira F, Burges C J C, Bottou L, et al. Advances in Neural Information Processing Systems 25. Curran Associates, Inc., 2012: 1097–1105.

[27] Lewis D, Agam G, Argamon S, et al. Building a Test Collection for Complex Document Information Processing[C]//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM, 2006: 665–666.

[28] Harley A W, Ufkes A, Derpanis K G. Evaluation of deep convolutional nets for document image classification and retrieval[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR). 2015: 991–995.

[29] Smith R. An Overview of the Tesseract OCR Engine[C]//Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). 2007, 2: 629–633.

[30] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[J]. arXiv:1409.4842 [cs], 2014.

[31] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv:1409.1556 [cs], 2014.

[32] Shin C, Doermann D. Document Image Retrieval Based on Layout Structural Similarity[M]. .

[33] Cesarini F, Marinai S, Soda G. Retrieval by Layout Similarity of Documents Represented with MXY Trees[C]//Document Analysis Systems V. Springer, Berlin, Heidelberg, 2002: 353–364.

[34] Hassan E, Chaudhury S, Gopal M. Shape Descriptor Based Document Image Indexing and Symbol Recognition[C]//2009 10th International Conference on Document Analysis and Recognition. 2009: 206–210.

[35] Zhu G, Zheng Y. Signature-based document image retrieval[C]//in Proc. European Conf. Computer Vision. : 752–765.

[36] Tan C L, Huang W, Sung S Y, et al. Text Retrieval from Document Images Based on Word Shape Analysis[J]. Applied Intelligence, 2003, 18(3): 257–270.

[37] Ha T M, Bunke H. Image processing methods for document image analysis[G]//Handbook of Character Recognition and Document Image Analysis. WORLD SCIENTIFIC, 1997: 1–47.

[38] Balasubramanian A, Meshesha M, Jawahar C V. Retrieval from Document Image Collections[C]//Document Analysis Systems VII. Springer, Berlin, Heidelberg, 2006: 1–12.

[39] Li J, Fan Z G, Wu Y, et al. Document Image Retrieval with Local Feature Sequences[C]//2009 10th International Conference on Document Analysis and Recognition. 2009: 346–350.

[40] Zagoris K, Ergina K, Papamarkos N. A Document Image Retrieval System[J]. Eng. Appl. Artif. Intell., 2010, 23(6): 872–879.

[41] Almazán J, Fernández D, Fornés A, et al. A Coarse-to-Fine Approach for Handwritten Word Spotting in Large Scale Historical Documents Collection[C]//2012 International Conference on Frontiers in Handwriting Recognition. 2012: 455–460.

[42] Liu H, Feng S, Zha H, et al. Document image retrieval based on density distribution feature and key block feature[C]//Eighth International Conference on Document Analysis and Recognition (ICDAR'05). 2005: 1040–1044 Vol. 2.

[43] Lu Y, Tan C L. Information Retrieval in Document Image Databases[J]. IEEE Trans. on Knowl. and Data Eng., 2004, 16(11): 1398–1410.

[44] Gatos B, Pratikakis I. Segmentation-free Word Spotting in Historical Printed Documents[C]//2009 10th International Conference on Document Analysis and Recognition. 2009: 271–275.

[45] Marinai S, Marino E, Soda G. Exploring Digital Libraries with Document Image Retrieval[C]//Research and Advanced Technology for Digital Libraries. Springer, Berlin, Heidelberg, 2007: 368–379.

[46] Gordo A, Valveny E. A Rotation Invariant Page Layout Descriptor for Document Classification and Retrieval[C]//2009 10th International Conference on Document Analysis and Recognition. 2009: 481–485.

[47] Ranjan V, Harit G, Jawahar C V. Document Retrieval with Unlimited Vocabulary[C]//2015 IEEE Winter Conference on Applications of Computer Vision. 2015: 741–748.

[48] Sankar K P, Manmatha R, Jawahar C V. Large scale document image retrieval by automatic word annotation[J]. International Journal on Document Analysis and Recognition (IJDAR), 2014, 17(1): 1–17.

[49] Kumar K S S, Namboodiri A M, Jawahar C V. Learning Segmentation of Documents with Complex Scripts[G]//Computer Vision, Graphics and Image Processing. Springer, Berlin, Heidelberg, 2006: 749–760.

[50] Razavian A S, Azizpour H, Sullivan J, et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014: 512–519.

[51] Babenko A, Slesarev A, Chigorin A, et al. Neural Codes for Image Retrieval[C]//Computer Vision – ECCV 2014. Springer, Cham, 2014: 584–599.

[52] Fan R-E, Chang K-W, Hsieh C-J, et al. LIBLINEAR: A Library for Large Linear Classification[J]. J. Mach. Learn. Res., 2008, 9: 1871–1874.

[53] Muja M, Lowe D G. Fast approximate nearest neighbors with automatic algorithm configuration[C]//In VISAPP International Conference on Computer Vision Theory and Applications. 2009: 331–340.

[54] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional Architecture for Fast Feature Embedding[J]. arXiv:1408.5093 [cs], 2014.

[55] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2011, 12: 2825–2830.

[56] Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks[G]//Fleet D, Pajdla T, Schiele B, et al. Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I. Cham: Springer International Publishing, 2014: 818–833.

[57] Yosinski J, Clune J, Nguyen A, et al. Understanding Neural Networks Through Deep Visualization[J]. arXiv:1506.06579 [cs], 2015.

# 附录 1 攻读硕士学位期间主要的研究成果

[1] Li L, Liao H, Chen Y. Document Image Super-Resolution Reconstruction Based on Clustering Learning and Kernel Regression[C]//Pattern Recognition. Springer, Singapore, 2016: 65–77.

[2] Liao H, Li L, Chen Y, et al. Low-Quality Character Recognition Based on Dictionary Learning and Sparse Representation[C]//Pattern Recognition. Springer, Singapore, 2016: 299–311.