

天津大学

本科生毕业论文



学 院 软件学院

专 业 软件工程

年 级 2013 级

姓 名 张洁茹

指导教师 谢宗霞

2017 年 6 月 20 日

天津大学

本科生毕业设计任务书

题目：基于深度学习的商标检索研究

学生姓名 张洁茹

学院名称 软件学院

专 业 软件工程

学 号 3013218071

指导教师 谢宗霞

职 称 副教授

一、原始依据（包括设计或论文的研究背景、研究条件、实验环境、研究目的等。）

研究背景：深度学习是目前较流行的机器学习方法，对复杂问题能较好的进行建模和预测。商标是商品的标识，为了避免新旧商标之间产生冲突，我们需要从商标库里高效快速地检索出是否已有具有高相似度的商标。目前使用的检索方法是通过分类码人工检索，既繁琐，准确率也不高。所以本论文将基于现有的深度学习算法，完成对商标数据的检索研究。

研究条件：该项研究的两个关键研究条件是商标数据与检索算法。目前商标数据已经整理完毕。对数据算法而言，则拟通过卷积神经网络(CNN)的学习和设计来完成商标数据的检索。在数据和算法均具备了研究条件。

实验环境：在 MATLAB 环境下,使用 MatConvNet 的卷积神经网络工具，确定商标数据的网络结构，在数据集上训练参数，利用相似性度量进行检索。实验所需数据集已提供。

研究目的：主要着重于卷积神经网络(Convolutional Neural Network,CNN)，确定适合商标数据的网络结构，在已知数据集上学参数，建立模型。然后根据已得到的模型来对商标进行检索。

二、参考文献

1. Najafabadi M M, Villanustre F, Khoshgoftaar T M, et al. Deep Learning Techniques in Big Data Analytics[M]// Big Data Technologies and Applications. 2016.
2. Gordo A, Almazón J, Revaud J, et al. Deep Image Retrieval: Learning Global Representations for Image Search[M]// Computer Vision - ECCV 2016. Springer International Publishing, 2016.
3. Kuo C H, Chou Y H, Chang P C. Using Deep Convolutional Neural Networks for Image Retrieval[J]. Electronic Imaging, 2016.
4. Sun Pei-Xia, Lin Hui-Ting, Luo Tao. Learning discriminative CNN features and similarity metrics for image retrieval[C]// 2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). 2016.
5. Razavian A S, Azizpour H, Sullivan J, et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition[J]. 2014:512-519.
6. Wan J, Wang D, Hoi S C H, et al. Deep Learning for Content-Based Image Retrieval: A Comprehensive Study[C]// ACM International Conference on Multimedia. 2014:157-166.
7. Vedaldi A, Lenc K. MatConvNet: Convolutional Neural Networks for MATLAB[J]. Eprint Arxiv, 2014:689-692.
8. Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems . 2012
9. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.
10. 林传力, 赵宇明. 基于 Sift 特征的商标检索算法[J]. 计算机工程, 2008, 34(23):275-277.
11. 王文惠, 周良柱, 万建伟. 基于内容的图像检索技术的研究和发展[J]. 计算机工程与应用. 2001(05)
12. 石军, 常义林. 图像检索技术综述[J]. 西安电子科技大学学报. 2003(04)
13. MATLAB 数字图像处理, 张德丰, 机械工业出版社
14. 特征提取与图像处理, 尼克松(Mark S. Nixon), 电子工业出版社
15. 数字图像处理[M]. [美][Kenneth R. 卡斯尔曼]Kenneth R. Castleman 著, 朱志刚等译, 电子工业出版社

三、设计（研究）内容和要求（包括设计或研究内容、主要指标与技术参数，并根据课题性质对学生提出具体要求。）

1.主要研究内容：研究基于卷积神经网络的商标检索算法，并设计与实现这些算法，通过在商标数据集上学习参数，训练模型来对商标进行检索。

2.主要指标与技术参数：人为确定适合商标数据的网络结构，确定数据集，然后在数据集上学参数，最后得到模型，对商标进行检索并得到准确率（准确率：正确识别的某类商标个数/所有需要识别的某类商标个数）。

3.具体要求：

1) 掌握 MATLAB；

2) 掌握卷积神经网络基本思想；

3) 能够确定商标数据的网络结构，学参数并得到模型；

4) 利用现有的 MatConvNet 工具包和相关数据，将已学得模型应用于商标检索，并分析检索效果。

指导教师（签字）

年 月 日

审题小组组长（签字）

年 月 日

天津大学本科毕业设计开题报告

课题名称	基于深度学习的商标检索研究		
学院名称	软件学院	专业名称	软件工程
学生姓名	张洁茹	指导教师	谢宗霞
<p>1. 课题的来源及意义</p> <p>来源：其他</p> <p>意义：商标是企业的一种重要的产权。在新公司注册时，为了避免新旧商标之间产生冲突而阻碍公司的注册流程，企业需要在设计商标时对同类产品的商品标识进行细致全面的检索，防止出现相似或者雷同的商标，从而避免出现因商标相似而无法注册公司的情况。但是，采用传统的人工检索，人为地为商标图像加上标签，然后再根据标签进行文本匹配，既费时费力，效率也不高。而基于内容的图像检索(CBIR)和深度学习的结合，可以很好地解决这个问题。</p> <p>因为商标数据的相似性特征是未知的，就需要利用无监督的深度学习方法来提取商标数据的特征。但是一般的深度神经网络(DNN)的上下层神经元都互相连接，很可能造成参数的膨胀，而深层次的卷积神经网络(CNN)既利用了类似于人的神经的结构，让不同层节点连接，但又通过卷积核作为中介，有效地限制了参数的个数，在挖掘出图像更有用的特征的同时也增强了检索效率。因此，采取基于深度学习的商标检索方法，尤其是利用卷积神经网络，有助于整个检索流程的优化与效率的提高。</p> <p>2. 国内外发展状况</p> <p>近些年，深度学习逐渐活跃于图像检索领域，尤其是基于内容的图像检索(CBIR)，而商标检索则属于图像检索的一个具体方向。对于CBIR，特征提取是一个关键点，选取合适的特征提取方法对建立高效的图像检索系统至关重要。2013年，林妙真在《基于深度学习的人脸识别研究》中曾指出，由于图像特征的复杂性，一般的机器学习方法所用的浅层网络很难准确的表示复杂函数，表达复杂对象，而深度学习(Deep Learning)可以通过深层次的网络结构，有能力去提取图像样本数据更本质的特征，刻画数据更丰富的含义。由此可见，深度学习对提高图像检索的效率和准确率有很大的帮助。此外，由麻省理工学院多媒体实验室开发的图像浏览工具Photobook就使用了深度学习的方法来检索图像，其中作者Pentland就提出深度学习可以模拟更好的神经网络结构，更细致地刻画图像特征这一观点。综上，深度学习在图像检索领域具有强有力的优势。</p>			

而最近几年，国内外都涌现出了许多利用深度学习来进行图像检索，尤其是商标检索的方法。就国内而言，例如，在 2008 年，林传力在论文中提出基于 Sift 特征的商标检索算法，增强了在检索中对干扰的抵抗力。在国外，Albert Gordo 等人在 2016 年提出了采用深度学习的方法来进行实例级图像检索，主要是着重于建立图像区域特征，采用描述符来进行检索。Kim, Yong-Sung 等人也在论文《Content-based trademark retrieval system using visually salient features》提出了一种基于内容的商标检索方法。但是，针对采取深度学习算法，尤其是卷积神经网络来进行商标检索，目前仍没有有效的方法。

3. 本课题的研究目标

针对深度学习方法，主要是卷积神经网络算法，有效地提取能够表征商标图像相似性的特征，然后利用这些特征，以及相似性搜索（比如欧氏距离）来进行商标图像的检索。最后对检索的准确率进行分析，从而提出有效的商标检索的深度学习方法。

4. 研究内容

- 1)研究深度学习方法，尤其是卷积神经网络；
- 2)分析商标图像数据的特点，采用深度学习的方法来提取能够表征商标数据的相似性的特征；
- 3)挑选合适的卷积神经网络工具，实现商标数据的特征提取；
- 4)针对已经获得的特征提取方法，根据特定的相似性度量（比如欧氏距离）来进行商标检索；
- 5)分析检索的效果与准确率。

5. 研究方法

针对上述研究内容，本课题的研究方法主要有：

- 1) 熟悉并研究深度学习，卷积神经网络；
- 2) 着重于分析如何有效提取能够表征商标图像相似性的特征；
- 3) 掌握 MATLAB 基本语法，能利用 MATLAB 编写基本算法；
- 4) 利用 MatConvNet 实现检索算法，并调试程序，分析其性能；
- 5) 利用现有数据对得到的检索算法进行测试，并对其进行分析。

6. 研究手段

通过查找相关的资料，渠道包括：网络教程，个人博客，论文，期刊等，研究卷积神经网络，用深度学习的方法来找到能够表征商标图像相似性的特征，并在检索模块运用特定的相似性检索方法来进行检索。阅读 MATLAB 和 MatConvNet 相关文献资料，利用 MATLAB 和 MatConvNet 来实现检索算法。最后对得到的准确率进行分析对比。

7. 进度安排

2016 年 12 月 12 日——2016 年 12 月 20 日：选题阶段，确定选题为《基于深度学习的商标检索研究》。

2016 年 12 月 21 日——2017 年 1 月 20 日：熟悉并确定课题内容，查阅资料文献，确定该题目的研究目标与方法，根据提供的任务书完成开题报告。

2017 年 1 月 21 日——2017 年 1 月 27 日：深入学习课题的所需技术，了解并深入研究该题目的具体实现方法。

2017 年 1 月 28 日——2017 年 2 月 26 日：得到所需数据集，研究相关深度学习算法。

2017 年 2 月 27 日——2017 年 5 月 6 日：编写代码实现算法，在特定数据集上进行实验，分析对比实验结果。

2017 年 5 月 7 日——2017 年 5 月 18 日：完成毕业设计论文初稿。

2017 年 5 月 19 日——2017 年 5 月 22 日：经过老师的指导，修改并完善毕业论文初稿。

2017 年 5 月 23 日——2017 年 6 月 7 日：完成毕业设计论文终稿，准备论文答辩。

8. 实验方案的可行性分析

对于本课题的研究实现已经具备以下条件：

- 1) 阅读和研究有关深度学习，卷积神经网络和图像检索的相关算法；
- 2) 掌握 MATLAB 基本语法，能利用 MATLAB 编写检索算法；
- 3) 会利用 MatConvNet 实现检索算法。

9. 已具备的实验条件

本课题已具备的实验条件包括：

- 1) MATLAB 和 MatConvNet 的技术已经相当成熟；
- 2) 深度学习和卷积神经网络的相关算法已经较为成熟，有助于对课题的深入研究；
- 3) 已现存许多其他的商标检索算法，比如 Sift，对该课题有借鉴意义；
- 4) MATLAB 和 MatConvNet 已较为稳定，适合作为课题研究的辅助工具。

10. 主要参考文献。

- [1] Najafabadi M M, Villanustre F, Khoshgoftaar T M, et al. Deep Learning Techniques in Big Data Analytics[M]// Big Data Technologies and Applications. 2016.
- [2] Gordo A, Almazón J, Revaud J, et al. Deep Image Retrieval: Learning Global Representations for Image Search[M]// Computer Vision - ECCV 2016. Springer International Publishing, 2016.
- [3] Kuo C H, Chou Y H, Chang P C. Using Deep Convolutional Neural Networks for Image Retrieval[J]. Electronic Imaging, 2016.

- [4] Sun Pei-Xia, Lin Hui-Ting, Luo Tao. Learning discriminative CNN features and similarity metrics for image retrieval[C]// 2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). 2016.
- [5] Razavian A S, Azizpour H, Sullivan J, et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition[J]. 2014:512-519.
- [6] Wan J, Wang D, Hoi S C H, et al. Deep Learning for Content-Based Image Retrieval: A Comprehensive Study[C]// ACM International Conference on Multimedia. 2014:157-166.
- [7] Vedaldi A, Lenc K. MatConvNet: Convolutional Neural Networks for MATLAB[J]. Eprint Arxiv, 2014:689-692.
- [8] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems . 2012
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.
- [10] Kim Y S, Kim W Y. Content-based trademark retrieval system using visually salient features[J]. Image and Vision Computing, 1997, 16(16):307-312.
- [11] Pentland A, Picard R W, Sclaroff S. Photobook: Content-based manipulation of image databases[J]. International Journal of Computer Vision, 1996, 18(3):233-254.
- [12] 林妙真. 基于深度学习的人脸识别研究[D]. 大连理工大学, 2013.
- [13] 林传力, 赵宇明. 基于 Sift 特征的商标检索算法[J]. 计算机工程, 2008, 34(23):275-277.
- [14] 石军, 常义林. 图像检索技术综述[J]. 西安电子科技大学学报. 2003(04)
- [15] MATLAB 数字图像处理, 张德丰, 机械工业出版社
- [16] 数字图像处理[M]. [美][Kenneth R. 卡斯尔曼] Kenneth R. Castleman 著, 朱志刚等译, 电子工业出版社

选题是否合适: 是 ☐ 否 ☐

课题能否实现: 能 ☐ 不能 ☐

指导教师 (签字)

年 月 日

选题是否合适: 是 ☐ 否 ☐

课题能否实现: 能 ☐ 不能 ☐

审题小组组长 (签字)

年 月 日

摘 要

基于内容的图像检索，是计算机视觉里的一个重要部分，它区别于基于文本的图像检索，主要研究在大规模的无标签的数字图像中如何进行检索。其中的核心内容是图像特征提取。图像特征既包括底层视觉特征，也包括高层语义特征。商标检索是近几年流行的图像检索的一个类别。主要是基于大规模的商标图像进行特征提取并检索。

深度学习，是用具有多层结构的神经网络进行学习的一种方法。其中深度卷积神经网络则是将卷积层引入神经网络。它可以从大量的数据中学习特征，然后应用于检索当中。因此，可以将深度学习应用到商标检索当中。

本文主要是在商标数据上训练网络，提取特征，再对比不同网络和提取算法的检索效果，包括：深度卷积神经网络 VGGNet 与训练后的网络的对比；小波变换、HSV 直方图、RGB 特征值和颜色相关图与卷积神经网络的对比。试验后得出：训练后的卷积神经网络可以有效地检索出相关图像，因此，得到了商标检索有效的检索方法：基于深度神经网络的商标检索。

关键词：基于内容的图像检索；商标检索；特征提取；深度卷积神经网络

ABSTRACT

Content-based Image Retrieval (CBIR) is an important branch of computer vision, which is different from text-based image retrieval and mainly study how to retrieve in a large unlabeled digital image database. The core of CBIR is how to extract features from images. Thus for image retrieval, feature extraction is a key step. The features of image not only include low-level visual feature, like the color or the texture, but also include the high-level semantic features. In this paper, what we discuss most is trademark retrieval, which is a kind of image retrieval.

Deep Learning (DL) is a kind of Machine Learning methods, it uses a neural network with multilayer structure to learn the features. And the Deep Convolution Neural Networks (DCNNs) put the convolution layers into the network, which can learn features from a large scale of data. Thus we can combine DL and trademark retrieval together to make the system work better.

Therefore, based on the special properties of Deep Learning, especially the features of Deep Convolutional Neural Network, we trained the CNN on the trademark dataset. After training, we use this network on the whole trademark dataset and get the retrieval result. In order to make the result more convinced, we also compare old DCNN with trained DCNN, as well as compare DCNNs with four feature extraction methods: wavelet coefficients, HSV, RGB and color histogram. The best result is from the trained DCNN. Thus, we propose a method of trademark retrieval based on Deep Learning, which uses DCNNs to extract features and perform the whole retrieval procedure.

Keywords: Content-based Image Retrieval; Trademark Retrieval; Feature Extraction; Deep Convolutional Neural Networks

目 录

第一章	绪论	1
1.1	引言.....	1
1.2	国内外发展现状	4
1.3	商标检索研究目标与内容	5
1.4	论文组织结构	6
第二章	相关理论研究	7
2.1	基于内容的图像检索技术	7
2.2	深度卷积神经网络	9
2.3	四种商标检索特征提取算法	14
2.4	小结.....	14
第三章	深度卷积神经网络在商标检索中的应用	16
3.1	深度卷积神经网络提取商标特征	16
3.2	商标图像的相似性度量	19
3.3	商标检索结果的评判标准	20
3.4	小结.....	21
第四章	实验结果与分析	22
4.1	数据集的建立与图像预处理	22
4.2	实验设计	23

4.3	实验结果与分析	24
4.4	小结.....	31
第五章	总结与展望	32
5.1	总结.....	32
5.2	展望.....	32
参考文献	34
外文资料		
中文译文		
致谢		

第一章 绪论

1.1 引言

商标，是基于商业目的而产生的，用于区分不同厂家的商品的，由文本、颜色、形状、字符等，或者上述的组合而构成的具有特殊性的标识。商标商家的一种必需的知识产权，对企业的成长过程有着十分重要的作用。公司建立时需要注册商标，为了避免新旧商标之间产生冲突而妨碍该公司的正常注册流程，就需要在设计商标时对同类产品的商品标识进行细致全面的检索，防止出现相似或者雷同的商标，从而避免出现因商标相似而无法注册公司的情况；此外，在公司发展的整个过程里面，商标作为一个代表其产品的标识，能够快速吸引顾客的注意力也很重要，这就需要这个商标和其他商标相比，更加独特并易于分辨。所以商标检索在企业的发展过程中饰演了关键的角色。

商标检索属于图像检索，图像检索分为传统的基于文本的图像检索（TBIR）^[1]和后来发展的基于内容的图像检索（CBIR）^[2]。图 1-1 介绍了 TBIR 的基本结构，可以看出：TBIR 是通过人为贴标签来完成对图像的类别贴附。图 1-2 介绍了 CBIR 的基本结构，其中可以看出：其包括了特征提取部分和检索部分。如果采用传统的人工检索，则需要人为地为商标图像加上标签，然后再根据标签进行文本匹配。尽管作为传统的检索方法，基于文本的检索过程简单又快速，但是由于商标图像数目庞大，检索的前提是为每一张都贴上标签，而这个过程是很费时费力的，同时人为地去贴标签，也降低了工作效率。为了使整个检索的流程高效快速，可以将基于内容的图像检索（CBIR）和深度学习结合起来，运用电脑的高速运算的特性去代替人工工作，科学高效地进行商标检索。

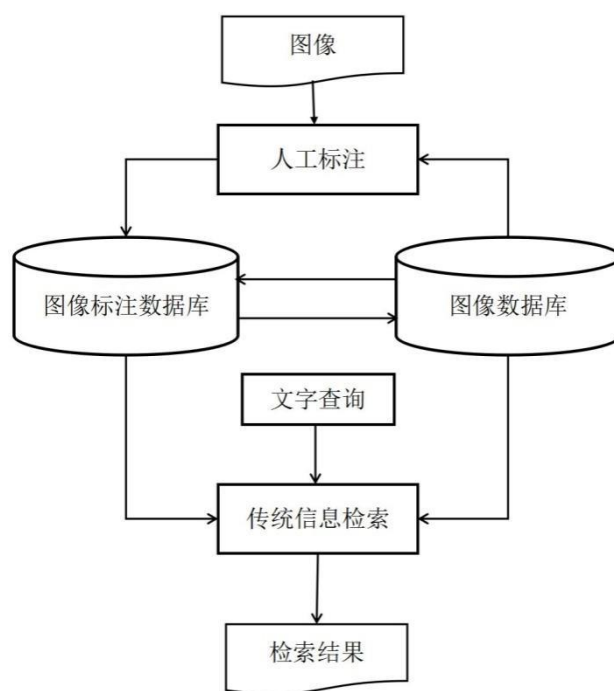


图 1-1 基于文本的图像检索结构

图 1-1 是 TBIR 的基本结构，由图可以看出：它主要分为：人工为图像贴标签，建立图像的标识标签信息数据库，结合图像数据库和图像标注数据库进行图像的具体检索三个部分，整个流程结构清晰，并不复杂。

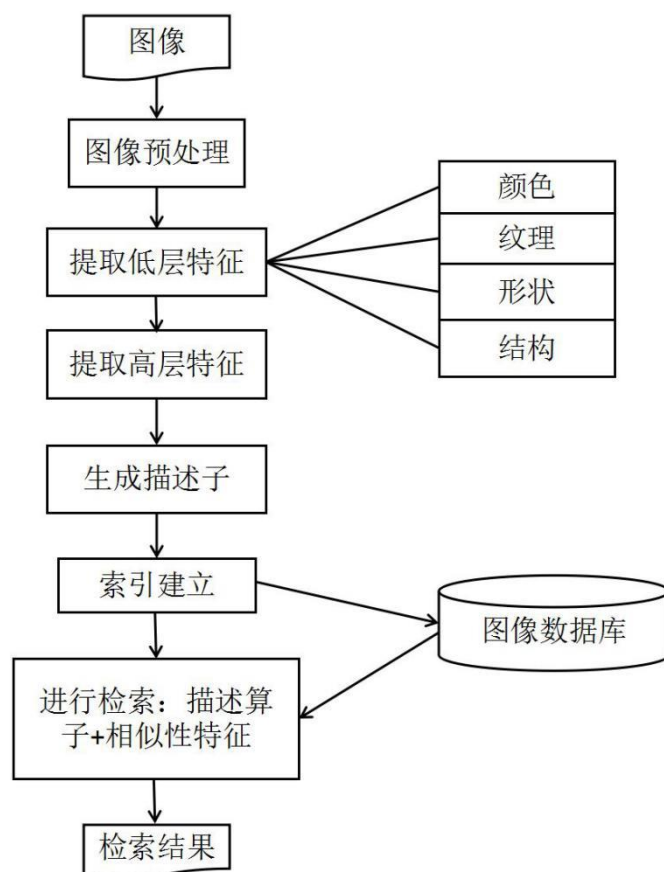


图 1-2 基于内容的图像检索结构

图 1-2 显示的是 CBIR 的主要结构，可以看出：在对图像进行一些预处理之后，要对图像提取其高层和底层的语义特征，再将这些特征整合后生成描述算子，再在图像数据库的基础上建立索引，进行检索。整个流程要远远复杂于图 1-1 所示的 TBIR 的检索流程。

一般来说，能够将深度学习分为有监督的学习和无监督的学习两种类别。由于商标数据的相似性特征是未知的，就需要利用无监督的深度学习方法来提取商标数据的特征。但是一般的深度神经网络 (DNN) 是一种全连接网络，它的上下层神经元都互相连接，如图 1-3 左图，若每个神经元之间都要进行交流，则很可能造成参数的膨胀。而深层次的卷积神经网络是一种局部连接网络，如图 1-3 右图，它既利用了类似于人的大脑的神经网络结构，既连接了各个神经元，又将卷积核当作中介，有效地限制了参数的个数，防止了参数膨胀。利用深度卷积神经网络，可以在挖掘出图像更有用的特征的同时也增强检索效率，更高效地对商标图像进行检索。因此，将深度学习运用到商标的查找和检索当中，尤其是利用卷积神经网络，有助于整个检索流程的优化与效率的提高。

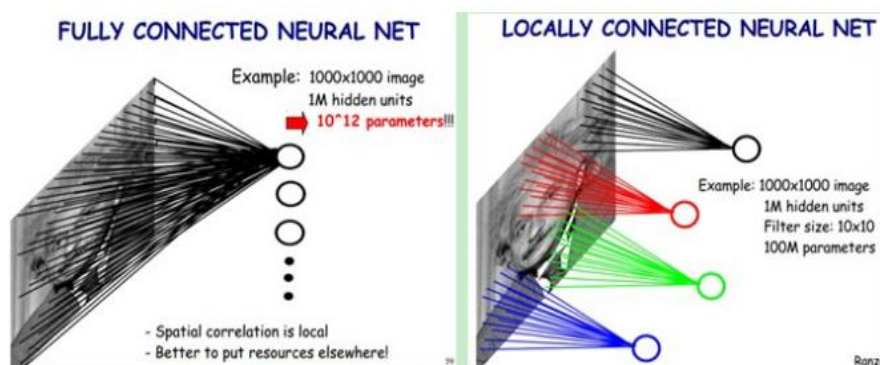


图 1-3 全连接与局部连接网络

图 1-3 显示的是全连接的神经网络和局部连接的神经网络的对比。左边为全连接结构的神经网络，右边为局部连接的神经网络。由图可以看出：全连接的每一个神经单元会同商标图像的所有的点相连接，这使得参数量过于庞大。而局部连接则只让神经元与部分像素点构成的区域进行连接，整个连接的结构十分简洁，也限制了参数的个数。

1.2 国内外发展现状

1.2.1 图像检索发展现状

图像是一个拥有着大量信息的有效载体，近几年来，图像检索技术逐渐成为一种流行的学习领域，并且着重于基于内容的检索。就国外而言，2015 年，Manesh Kokare 等人在《Iete Journal of Research》上以一篇论文的形式^[3]提出了一个关于 CBIR 的一个调查，其中不仅提到了 CBIR 的系统设计和实现过程，还提到了图像的特征表示和提取的相关技术，并对未来的发展进行了展望。2016 年，Anu Bala 等人在其发表的论文^[4]中提到了一种可以有效表示图像特征的描述符：local texton XOR patterns (LTxXORP)，用这个描述符来进行基于内容的图像检索，其中的创新点就是引入了异或的操作。2017 年，Leigh Rothschild 等人提出了一种将数字图像加载到存储器中的方法^[5]，用于图像的辅助信息存储和标注。

就国内而言，也相继出现了许多根据图像的不同特征检索的办法。比如，在 2008 年，林传力等人在论文中^[6]提出了一种采用一种新特征 Sift 来进行图像特征提取与检索，Sift 算子作为描述物体局部的特征的一种描述符，可以增强在图像检索中算法对于干扰的抵抗力。2010 年，相继有人提出了基于分块主颜色匹配^[7]、纹理谱描述符^[8]、Canny 边缘检测算子^[9]的图像检索方法。再到最近的 2016 年，赵宏伟等人提出了特征点显著性约束的图像检索技术^[10]，针对有复杂背景的图像的检索。

1.2.2 深度学习在图像检索中的应用

近几年，深度学习也开始活跃于图像检索领域。对于 CBIR，特征提取是一个关键点，采用高效且快速的特征提取方法对建立有效的商标检索系统十分必要。2013 年，林妙真在《基于深度学习的人脸识别研究》^[11]中曾指出，由于图像特征的复杂性，一般的机器学习方法所用的浅层网络很难准确的表示复杂函数，表达复杂对象，而深度学习可以通过多层次的神经网络结构，有能力去提取图像样本数据更本质的特征，刻画数据更丰富的含义。由此可见，深度学习有助于提升图像检索系统的检索能力。

近几年，国内外都逐渐有人提出把深度学习的方法用到图像检索领域。就国外而言，Albert Gordo 等人在 2016 年提出了采用深度学习的方法来进行实例级图像检索^[12]，主要是着重于建立图像区域特征，采用描述符来进行检索。2017 年，A. Anbarasa Pandian 等人在论文《Performance Analysis of Texture Image Retrieval in Curvelet, Contourlet, and Local Ternary Pattern Using DNN and ELM Classifiers for MRI Brain Tumor Images》^[13]提出了使用 DNN 来分类脑肿瘤图像。在国内，李钊等人在 2015 年也提出了基于卷积神经网络的图像检索技术^[14]，并且将传统的视觉特征和用卷积神经网络提取的特征进行了对比和分析。

1.3 商标检索研究目标与内容

通过对发展现状的分析可知，虽然已经存在许多图像检索，甚至商标检索的方法，同时深度学习近几年也得到了充分的发展，但是目前为止，还没有基于深度学习的商标图像检索技术被提出。而深度学习可以帮助提高检索的性能。因此，文本着重研究了如何将深度神经网络应用于商标检索当中。具体研究目标及内容如下。

1.3.1 研究目标

使用深度学习，具体到卷积神经网络，来训练网络模型，将模型用于商标图像的特征提取，有效地找出能够表征商标图像最大相似性的特征，然后利用这些特征，以及相似性搜索来进行商标图像的检索。最后分析商标数据的检索结果，从而提出有效的商标检索的深度学习方法。

1.3.2 研究内容

1. 研究深度学习方法，尤其是卷积神经网络；
2. 分析商标图像数据的特点，采用深度学习的方法来提取能够表征商标数据

的相似性的特征；

3. 挑选卷积神经网络工具包，实现商标数据的特征提取；
4. 采用不同的提取特征的办法，根据特定的相似性度量来进行商标检索；
5. 分析检索的效果与准确率。

1.4 论文组织结构

本文组织结构如下：

第一章为绪论。介绍了：商标检索的发展背景，研究意义，深度学习以及深度卷积神经网络的发展现状，基于深度学习的商标检索的研究目标和内容。

第二章为相关理论研究。主要分析了：图像检索的相关背景知识，特征提取的主要分类和相关算法，相似性距离的理论研究，深度卷积神经网络的基础知识和研究，以及四种本文用来用作比较的特征提取算法。

第三章为卷积神经网络在商标检索中的应用。介绍了：用已知网络 VGGNet 来提取特征的过程，训练 VGGNet 得到新网络后进行提取的过程，两种主要的相似性度量（欧氏距离和高斯核函数）的计算细节，以及最后检验试验结果的评判标准的计算细节。

第四章为实验结果与分析。介绍了：数据集的建立和数据预处理的详细流程，在两个不同数据集上用三种不同网络提取特征并检索的结果比较，将四种其他提取特征的方法与深度卷积神经网络进行比较。

第五章为总结。论述了本论文的最终商标图像实验结果，提出了未来需要改进的地方，并进行了展望。

第二章 相关理论研究

基于深度卷积神经网络的商标检索的实现,是以基于内容的图像检索为前提的,再结合深度卷积神经网络为特征提取的主要算法,共同完成检索任务。基于内容的图像检索的关键点着重体现在两个方面:特征提取,以及相似性距离。而深度学习部分主要是利用了卷积神经网络的知识。因此,本章就这几个关键点来对一些基础的知识和理论支撑来做相关介绍。

2.1 基于内容的图像检索技术

2.1.1 基于内容的图像检索系统结构

早在上世纪,基于文本的图像检索技术就已问世,其主要通过文字来为图像贴上标签,描述图像的外部特征,比如:图像大小,图像范畴,署名作者,上市日期等。而到 90 年代,又出现了基于内容的图像检索技术,即对图像的内容语义,上下文关联等的特征进行提取,然后进行检索。而本文主要是研究基于内容的图像检索方法,具体到将商标图像作为图像数据集。图 2-1 展示了基于内容的图像检索系统的主要架构。整个检索的主要过程为:对商标图像的处理,包括大小,通道等的处理;对该图像进行特征提取,既包括底层语义,也会包括高层语义特征;将该图像与整个图像数据库里的所有图像提取到的特征进行对比;根据相似性准则来筛选相似图像,做成索引后显示出检索结果。其中比较重要的两个部分为:提取特征部分和相似性计算部分。

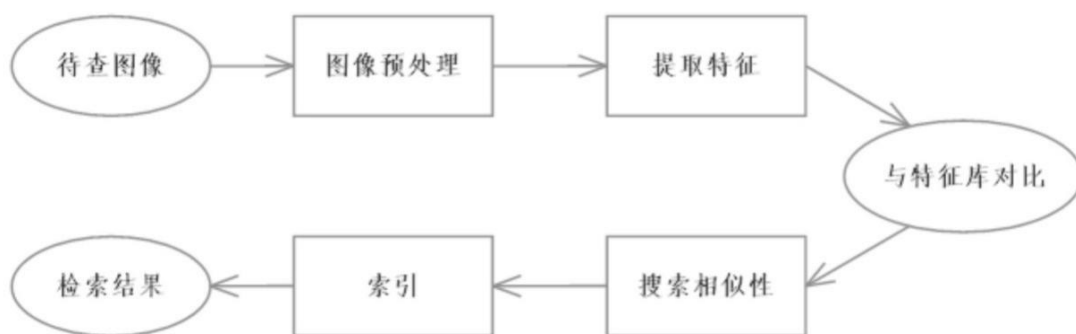


图 2-1 基于内容的图像检索框架

在深度学习开始用于图像检索之前,在图像识别领域,需要借助类似于 SIFT (Scale-invariant feature transform) 或者 HoG (Histogram of Oriented Gradient) 等算法提取可以较好地区分图像的特征,然后再利用机器学习的方法来进行匹配

和识别。而这些特征都不能够很全面地提取到有效的图像特征。

2.1.2 特征提取介绍

特征提取这个概念经常出现在计算机视觉当中，指的是将计算机作为工具来提取图像的信息并确定可以代表图像特征的点。

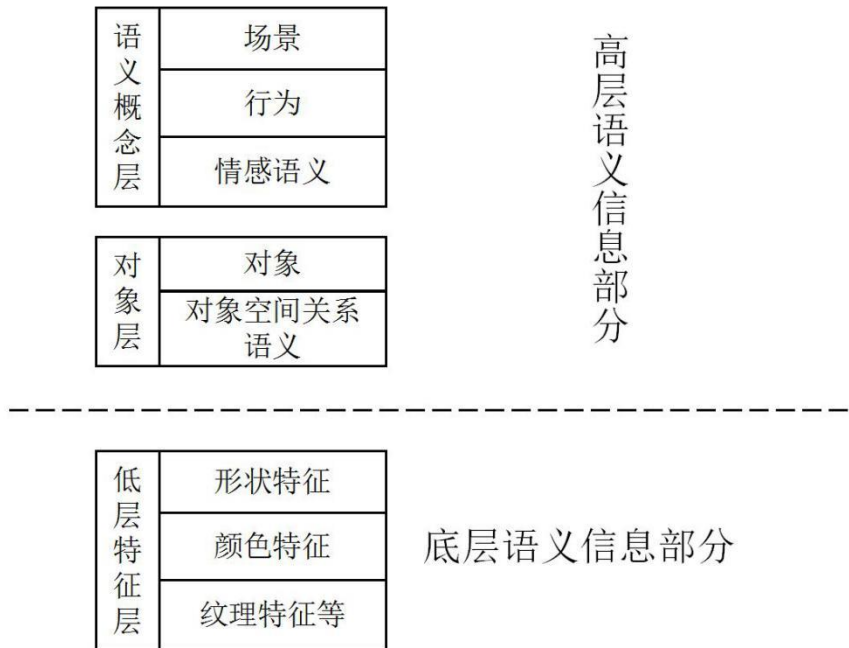


图 2-2 图像的底层和高层特征

图像常用的特征有很多种，如图 2-2，其既包括在颜色，纹理，空间位置等方面的底层特征，也包括了其高层语义特征，其中语义概念也可以区分为：图片特定场景语义概念、图像所描述行为语义概念和图像所包含的情感语义。此外在底层和高层特征之间，是需要以图像作为对象的参与的。而下面列举了一些常用的特征提取方法：

其中颜色特征主要描述全局的特点关系，指的是图像的外表性质，其主要是基于像素的特征。其主要方法包括：颜色直方图法，即形成直方图，由分布来代表图像的特征；RGB 颜色特征，即用三个通道各自的均值等计算结果值作为特征值；颜色矩等。纹理特征也是一种全局特征，是对某个区域中的像素点的一种总结，包括模型法、信号处理法、几何法等，其中比较常用的算法为棋盘格特征法和结构法，这两者都属于几何法的范畴；最后空间关系特征指的是：将目标点的相对位置从图片中抽离出来。其中有两种方法较为常用，其一是采用自动方法去将图像分离成不同部分，在不同的区块上提取特征。其二是简单划分图像子块，

对每块进行特征提取并建立索引。以上几种特征提取方法一般需要配合使用。此外还有上文中提到的 SIFT 和 HoG 等的特征。

除了上述的各种图像提取特征的方法之外，本文内提到卷积神经网络则降低了对图像预处理的要求，直接输入图像，得到输出的特征。同时也降低了算法的复杂度。所以将基于卷积神经网络的特征提取算法用于图像检索，可以获得高效的检索方法。

2.1.3 相似性度量介绍

为了评估不同样本，这里主要指的是两张图像之间有多大比重的相似程度，需要用一个距离来量化样本之间的远近，用此距离作为相似性度量。主要的相似性度量有许多种，既包括简单的欧氏距离和将其标准化后的变体，也包括一些复杂的距离公式，比如曼哈顿，切比雪夫，马氏等。当然还包括用以计算角度的夹角余弦。其中较为常用的是欧氏距离和夹角余弦。欧氏距离指的当处于欧氏空间中时，两个向量之间的远近。而夹角余弦指的是将样本作为两个向量，然后计算向量方向的差异。下面主要介绍一下本文会用到的相似性度量：欧氏距离和高斯核函数。

欧氏距离，又称为欧几里得度量，是较为常见的相似性度量方式，其公式如下：

$$\text{score} = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (2-1)$$

其中 i 指的是维度。A 和 b 指的是空间中的两个点，其可以是二维，三维，乃至扩展到多维。

高斯核函数，其公式如下：

$$\text{score} = e^{\frac{-||x-xc||^2}{2*\sigma^2}} \quad (2-2)$$

$-||x - xc||^2$ 表示的是空间中某点距离中心的距离的平方。

其中 x 为某一点， xc 为中心点， σ 为代表了作用的宽度，是一个可调的数，主要用来掌握该函数在空间中的影响范围。

本文中主要介绍的是欧氏距离和另一种高斯核函数的变体，具体见第三章的 3.2 小节。

2.2 深度卷积神经网络

2.2.1 深度学习

早在 2006 年，Hinton 等人就提出了一种新的机器学习的概念：深度学习。

它是属于机器学习中的一个分支或者说一个具体化的应用，是对神经网络的一个进阶的研究领域，它效仿了人的大脑的神经元和之间的相互关系，构成了一个具有输入层，输出层，以及数个隐含层的一种网络体结构，它可以有效地对复杂度较大的数据进行分析与解释，如图 2-3。它基于对对象的底层特征的组合，然后形成更加抽象的高层表示，用以提取对象更准确的特征，表示类别属性。深度学习与机器学习一样也根据有无监督进行区分。而接下来要提到的卷积神经网络就是带监督的深度学习模型。

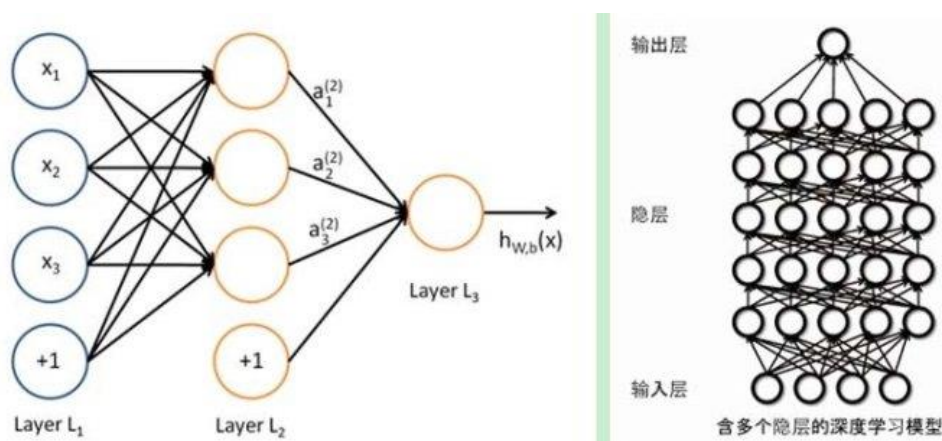


图 2-3 深度学习模型结构

图 2-3 中，左图显示了只有一个隐层的神经网络结构，右图展示了含有多个隐层的深度学习的网络结构体。其输入为一个具体对象，可以是一张图片，一个视频，也可以是一节语音等。输出则为经过计算得到的可以代表该输入对象的特征或者特点表示。在输入和输出之间包含了许多层隐形的网络层，隐层并不是说隐藏的意思，而是说是输入和输出之间的计算层。由图可知，深度学习的特点是神经网络的结构更复杂，层数更多。

卷积神经网络，是由 Hubel 在研究猫的大脑皮层的时候发现的一种网络结构，是一种前馈型网络。而本文所用到的是有多层隐层的深度卷积神经网络。上图 2-3 是普通的神经网络，而下图 2-4 是一种卷积神经网络的应用，可以看出卷积神经网络可以理解成在隐层中采用了卷积层和池化层，最后佐以全连接层。卷积神经网络具有以下特点：其一是可以共享卷积核，也就是所谓的权值，可以减少整个神经网络中自由参数的个数；其二是每层卷积后面跟有一层用于计算，减少了不同特征之间的分辨率；其三是网络中可进行并行学习。这些特点加上其类似于人脑神经网络的特征，使得卷积神经网络在一些更复杂多元的领域有着优越的特性，比如图像物体识别，视频识别，语音的识别与处理等方面。

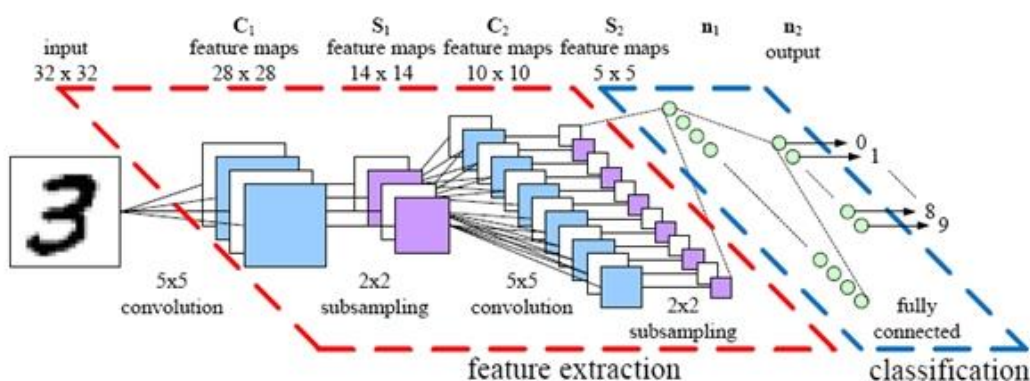


图 2-4 一种卷积神经网络的结构

图 2-4 中显示的是一种卷积神经网络的结构，可以将之分成特征提取部分和分类部分。左半部分为提取特征，这个部分的网络层主要是多个卷积层，每个卷积层后面都会跟随降采样层，这里的降采样值就是池化层。卷积层使用卷积核来对输入图像的部分进行特征提取，形成特征矩阵后，将其用于整张图像，最后得出特征的总矩阵。然后用降采样来总结特征的分布情况，可以缩小特征矩阵的大小。此外左半部分中也会根据实际问题和网络的效果添加一些激活函数。图的右边为使用全连接进行分类，先得到特征值，然后进行分类。当然对于一个网络来说，是需要通过进行不断的训练来提高性能，所以在网络提取特征之前，需要先对网络进行训练，训练是更新的就是卷积核的参数，也就是所谓的权重，这是就用到了反向传播算法，此处不作详细介绍。

2.2.2 深度卷积神经网络 VGGNet

深度卷积神经网络近几年在图像识别方面有很出色的表现。在 2012 年的计算机视觉比赛 ILSVRC 中 (ImageNet Large Scale Visual Recognition Competition, ILSVRC)，Alex Krizhevsky 提出了一种全新的深度卷积神经网络的模型^[15]，使用 AlexNet 获得了比赛的冠军，这是卷积神经网络在图像检索领域的首次现身。在随后的几年的该项比赛中，取得前几名的都是深度卷积神经网络，包括 2014 年的第一名 InceptionNet^[16]，第二名 VGGNet^[17]，以及 2015 年冠军 ResNet^[18]。随着网络的深度和一些计算函数和池化层的选取的不同，网络的效果也不同。而本文主要使用 VGGNet 为网络的基础模型，在此基础上进行特征提取和网络训练。下图 2-5 展示了 VGGNet 的具体结构。

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

图 2-5 VGGNet 网络结构图

深度卷积神经网络 VGGNet 中包含：16 个卷积层，3 个全连接，它是一个很深的卷积神经网络。将输入图像统一为长宽均为 224 像素，预处理时只是减去 RGB 的均值。然后各卷积层的卷积核分别为： 3×3 大小和 1×1 大小。卷积的步长均设为 1 个像素，填充设为 1 个像素。然后有五个池化层，这里具体用到的池化为最大池化。窗口长宽均为 2 个像素，步长为 2 个像素。全连接当中，第一个和第二个都是包含有 4096 个通道，代表了特征个数为 4096 个，最后一层是有 1000 个通道，用作分类。所有网络中的隐藏计算层都用的是 ReLu 的激活函数，即纠正线性单元。需要注意的是：VGGNet 中并不会引用局部响应标准化。下面分段介绍一下网络中各层的计算细节，包括：卷积层，降采样（池化层），激活函数，全连接，以及 softmax 层。

卷积层，其实质就是用一个固定长宽的核，以特定的模式（即步长）来遍历整个待检索的图像，遍历完后即可用一个新的像素矩阵来表示该原图像。假设原图大小为 $36 \times 36 \times 3$ ，有 6 个大小为 $5 \times 5 \times 3$ 的卷积核（即过滤器），经过遍历

后即可得到 $32 \times 32 \times 6$ 的输出。要注意卷积核的通道数要和输入的通道数一样。一般来说，第一个卷积核为 3 个通道，最后的卷积核的通道数等于上一层所用到的卷积核的个数。深度卷积神经网络即由多个卷积层组成。

降采样，即池化层。是一种利用固定窗口，固定步长降低空间的尺度的方法。常用的是最大池化，比如有一个只有一个通道的输入，大小为 $4 \times 4 \times 1$ ，各个位置的像素值如图 2-6，经过一个大小为 2×2 的最大池化的过滤器，步长为 2，则输出为一个 $2 \times 2 \times 1$ 的结果。详细过程表示如图 2-6。这里需要注意的是，池化层不会改变通道数，因为一般只用一个池化的过滤器。

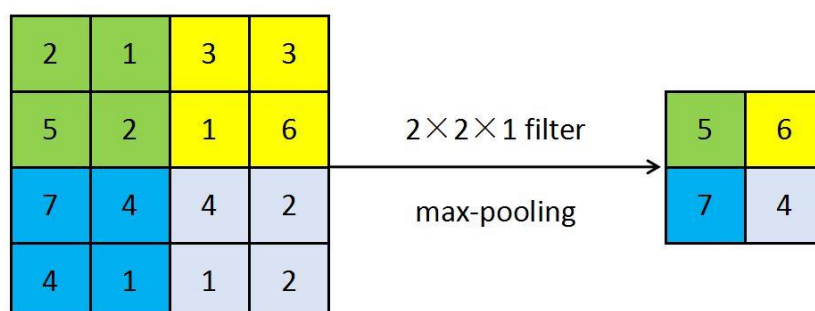


图 2-6 最大池化过程

激活函数，一般用于网络的正向传播。因为网络的表示最开始是线性的，这样会导致对输入的表达不足，因此引入一些非线性的激活函数，比如：Sigmoid, tanh 等。为了加入非线性元素。VGGNet 主要用到的是 ReLU, 即 $\max(x, 0)$ 。虽然网络结构中未标出，但是在每次卷积计算后，都会加入 ReLU 再计算。ReLU 和其他两种激活函数相比，优点在于：不会饱和，计算更快捷，以及收敛地更迅速。图 2-7 显示了三种激活函数的图像。由图可以直观地看出 ReLU 的优点。

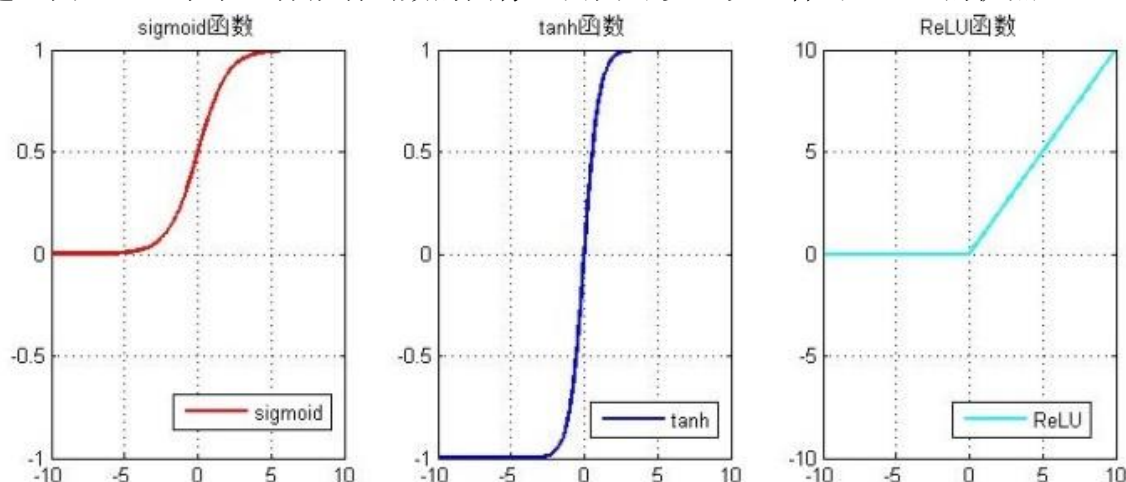


图 2-7 三种激活函数

全连接层，当经过所有的卷积层后，需要经由几个全连接层。如果说前面的局部连接的卷积层是把图像降维得到特征，那么这里的全连接层可以起到分类的作用。最后加的 softmax 层是将一个 m 维的向量（即所谓的特征向量）映射到一个 n 维的向量（这里 $n < m$ ）中， n 为类数。

2.3 四种商标检索特征提取算法

这里介绍四种商标检索的特征提取算法，后文中将这四种和它们的组合与卷积神经网络进行对比。这四种分别是：小波变换、HSV 直方图、RGB 特征值和颜色相关图。

小波系数（wavelet coefficient, WC），是属于纹理表示中的信号处理法。小波变换是一种信号变换，而在图像处理领域，对图像进行小波变换后的结果可以用来表示物体的特征，它提供了随频率改变的窗口，用一系列不同尺度的小波来分解原函数，在不同尺度下进行变换，得到相对应的系数。其主要通过平移和尺度变换分解来得到时间特性和频率特性。本文中提到的小波变换的过程主要是：先将 RGB 图像转为灰度图像，再进行四次的分解，然后求小波系数的均值和标准差，组成一个 1×40 的特征向量。

HSV 颜色直方图（H: Hue, S: Saturation, V: Value, color histogram），是提取图像颜色特征的一种方式。整个过程如下：第一步是颜色量化，用构成颜色的小的区间来代替一整块颜色区域，每个小区间作为结果曲线的一部分。最后的直方图是由不同颜色小块所占的数量来进行绘制的。本文中提到的提取 HSV 直方图特征的过程为：将图像分成 h, s, v 三个通道；量化每个通道，将所有的像素放入 $8 \times 2 \times 2$ 大小矩阵中；在经过一系列的标准化操作得到一个 1×32 的特征向量。

RGB 特征提取，类似于 HSV，是一种描述图像颜色特征的办法，但是过程并不复杂：直接计算图像的三个通道的平均值，以及其标准差，组成一个 1×6 的向量作为特征向量。

颜色自动相关图，属于颜色相关图，表示了颜色的散步状况。即描述了某个颜色像素占图像的比重，也描述了空间上的颜色的相互之间的关系特点。但是自动相关图只考虑到了同一种颜色下的像素之间的空间特征关系，减少了方法的复杂度。

2.4 小结

这一章主要介绍了基于内容的图像检索，同时介绍了其中的最关键的两点：

特征提取和相似性度量。此外，又着重介绍了深度学习，尤其是深度卷积神经网络的相关知识，包括本论文中使用的卷积神经网络 VGGNet，详细分析了每个部分的计算细节和对应的功能特点。最后，介绍了四种针对于商标图像的特征提取方法。

第三章 深度卷积神经网络在商标检索中的应用

本章主要是介绍了如何将深度学习应用于商标图像检索当中，深度学习部分用到的是深度卷积神经网络 VGGNet，该网络在第二章中有详细介绍。本章具体从如下几个部分阐述：用深度卷积神经网络 VGGNet 来预分类商标图像，得到 50 类可以用做训练的商标数据集；在得到的商标数据集上训练 VGGNet 网络，得到两种训练结果；如何用深度卷积神经网络来提取商标数据的特征；检索部分用到的相似性距离；最后总体的结果的判断指标。

3.1 深度卷积神经网络提取商标特征

3.1.1 深度卷积神经网络分类商标数据

由于实验得到的商标数据集中只存储了商标名字和图像本身，并没有其他的相关信息，比如：具体分类，相关标签等。为了得到对训练网络有用的训练商标数据集，先用深度卷积神经网络 VGGNet 来对 50 张随机图像进行相似性检索，各取出 60 张相似图像，然后再人工剔除 10 张不相似的图像，得到 50 类，每类 50 张的训练集。该训练集的具体信息参见上章的第一节中对数据集的介绍。此处不再赘述。

3.1.2 商标数据上训练卷积神经网络

这里训练网络时用到的训练集有两个，一个是 50 类，每类各 50 张相似商标图像的数据集 1；另一个是数据集 1 经由翻转等变换后得到的数据集 2，数据集详细信息参见第四章实验部分对数据集的介绍。

第一次训练，将数据集 1 中 50 类分别标签标为 1 到 50，然后用 MatConvNet^[19] 工具对 VGGNet 原网络进行再训练。训练包括对训练数据集的预处理（处理成 mat 格式）以及设置网络结构，这里网络用的是 VGGNet，最后进行训练。其中取每一类中的前 2/3 的商标图像作为训练集，后 1/3 的图像作为测试集，经过特定函数处理后，经过 37 代的训练，其在验证集上的 top-5 的错误率稳定于 0.25，top-1 错误率稳定在 0.74。最后取该代的结果作为最终网络，命名为 VGGNet1。训练结果如图 3-1。

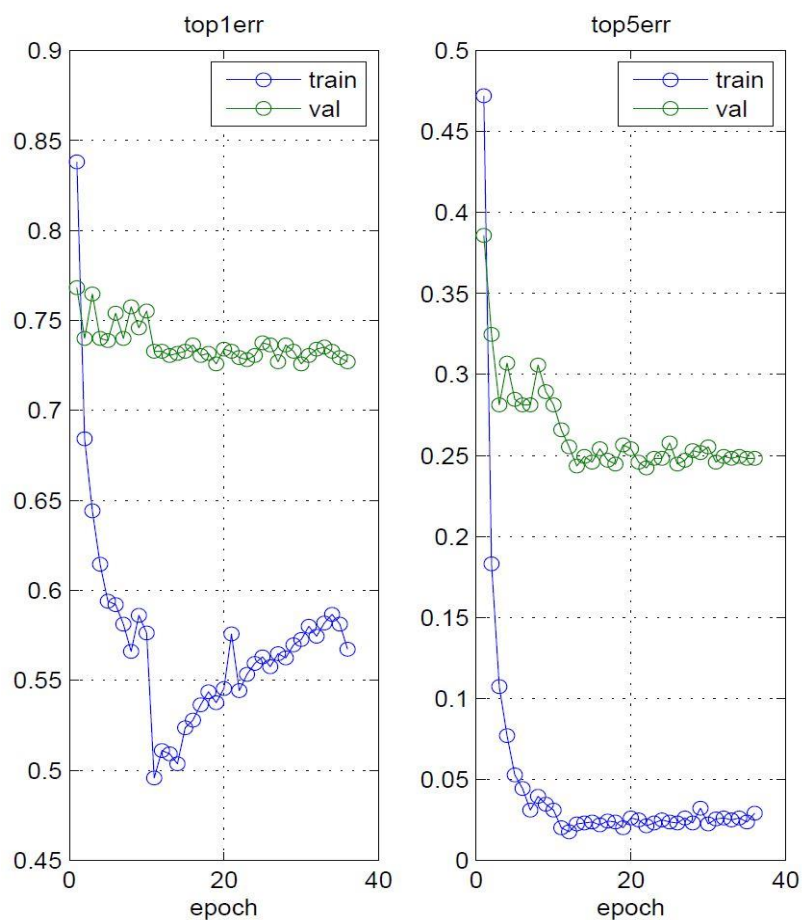


图 3-1 第一次训练结果

然后仍然以 VGGNet 原网络为基础网络，再对每张图片进行 13 种旋转、翻转和噪声的变换后得到的数据集 2 上进行训练，训练中仍取每一类中的前 2/3 的商标图像作为训练集，后 1/3 的商标图像作为测试集，经过 14 代训练后，其在验证集上的 top-5 错误率大约稳定为 0.03，top-1 错误率大约稳定在 0.525。最后取该代的结果作为最终网络，命名为 VGGNet2。训练结果如图 3-2。

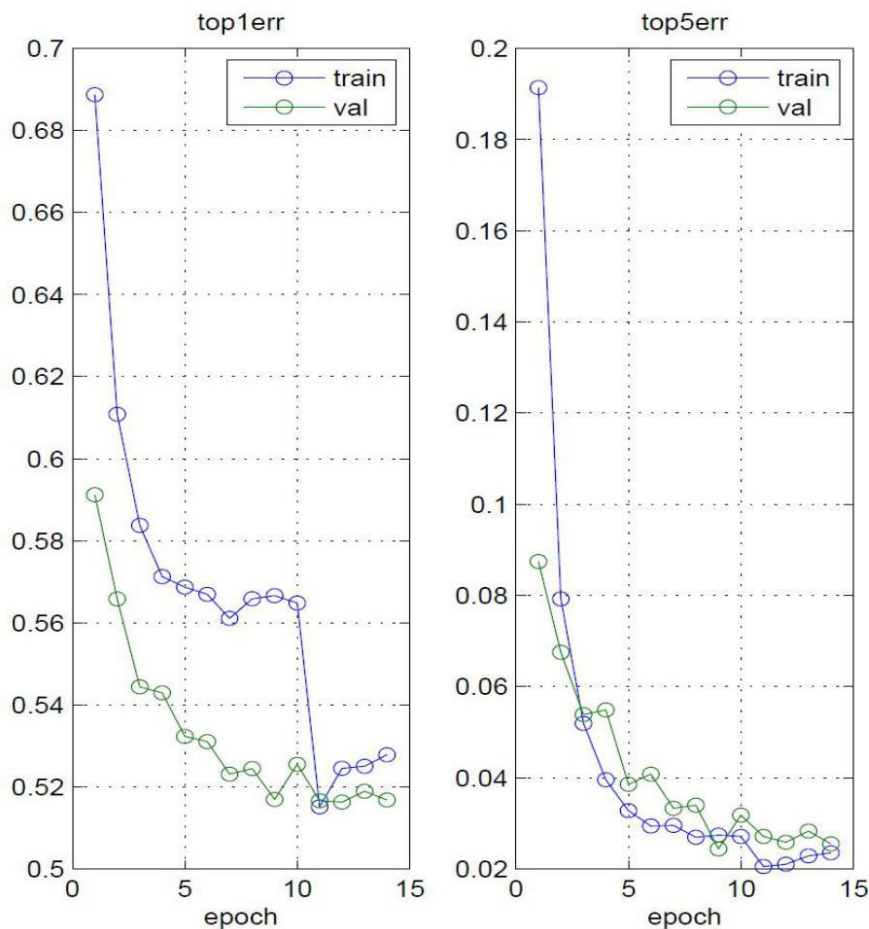


图 3-2 第二次训练结果

上图 3-1 和 3-2 展示了两个网络的训练结果，其中 top-1 错误率指的是对图像进行一次预测，其错误的概率。Top-5 错误率指的是对图像进行五次预测，五次全错的概率。Top-1 和 top-5 错误率在 ILSVRC 比赛中作为模型评测的指标，其数值越低越好。

3.1.3 商标图像特征提取

现在共有三个网络：原卷积神经网络 VGGNet，第一次训练得到的 VGGNet1，以及第二次训练得到的 VGGNet2。这三个网络的网络结构是完全一样的，只是每层的卷积核内的参数有所变化。为了提取商标的特征值，我们并没有取整个网络作用于数据集的结果，因为这个结果是一个分类值。而是取了后面的三个全连接中的第二层全连接的输出，即整个网络的第二十层的输出，其为一个 4096 维的向量，可以作为输入图片的特征向量，这个特征向量以细胞组的方式存到 mat 文件中，留存以备之后的检索。图 3-3 是截取的 VGGNet 的后几层全连接层，我们取的是倒数第二层的全连接的输出 4096 维，即 FC-4096 的输出。

maxpool
FC-4096
FC-4096
FC-1000
soft-max

图 3-3 VGGNet 后几层全连接层

3.2 商标图像的相似性度量

在对商标图像提取了特征并保存之后，下一个步骤就是计算两张图像之间的距离，以显示相似程度。从而挑选出相似性最强的商标图像。本实验主要用到了两种相似性度量方法，一种是欧几里得度量，另一种是高斯核函数的变体。下面分别说明一下本实验中对这两个距离的使用情况。

3.2.1 欧氏距离

对于欧氏距离，即欧几里得度量，直接使用了向量间的距离计算相似度，如公式 3-3-2。其中 n 取 1 到 4096. 表示有 4096 维特征。 a_i 指的是商标图像库中的图像的特征向量， b_i 指的是待检索的商标图像的特征向量。 $Score$ 为两个向量之间的欧几里得距离，亦作为相似性的得分，该值越小，说明相似性越大。

$$score = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (3-1)$$

3.2.2 高斯核函数

对于高斯核函数，也就是径向基函数的一个具体化，其原本的公式如公式 3-2, 其中 x 为空间中某一点， xc 作为中点标杆， σ 为函数的作用的范围，是一个可变的数，掌握了函数的影响区域。但是由于进行运算的是两个向量，则把分子改成欧氏距离，如公式 3-3。其中 a 为商标图像库中的图像的特征向量， b 为待检索的商标图像的特征向量， σ 仍为函数的宽度参数。

$$score = e^{\frac{-||x-xc||^2}{2*\sigma^2}} \quad (3-2)$$

$$score = e^{\frac{-\sqrt{\sum_{i=1}^n (a_i - b_i)^2}}{2*\sigma^2}} \quad (3-3)$$

其中对于 σ 的选取，由于 σ 越小，核函数对 x 的衰减越快，这就放大了数据 x 之间的差别，即 $k(x)$ 对 x 值的变化很敏感；而 σ 越大，核函数对 x 的衰减越慢，

这使变化变得不敏感。试验中对 σ 分别取了 1, 2, 5 三个值, 最后结果显示其对检索结果并没有影响, 因为数值是统一变化的。

图 3-4 显示的是对其中一张商标进行检索的结果, 欧氏距离和高斯核函数的三个取值均会得到该结果。表 3-1 显示的是欧氏距离, 以及高斯核函数的不同的取值对应的相似度前 5 的分数。



图 3-4 检索结果

表 3-1 分数比较

前五张图片	1	2	3	4	5
欧氏距离	0.000000	0.591979	0.856994	0.866553	0.936137
$\sigma = 1$	0.606531	0.662066	0.728775	0.731783	0.755095
$\sigma = 2$	0.135335	0.192134	0.282081	0.286767	0.325092
$\sigma = 5$	0.000004	0.000033	0.000367	0.000407	0.000891

由表 3-1 可以看出, σ 的取值越大, 其距离值越小, 即所谓的变化不会很敏感。其中欧氏距离的第一个数是 0 是因为检索时并没有排除待检索图片本身, 所以第一张最相似的图片为其本身, 而欧氏距离是作差后再平方求和开根号, 则其差为 0, 导致距离为零。

3.3 商标检索结果的评判标准

3.3.1 一般查准率与查全率

经过相似性度量计算后，对图片进行索引并得到结果，这时需要对检索结果进行分析判断。本实验使用的是两个评判指标，分别为：查准率与查全率。

对于数据集 1 进行的检索，使用的查准率 P1 和查全率 R1 定义如下：

$$P1 = \frac{\text{检索出的类似商标个数}}{\text{检索出的商标总个数}} * 100\% \quad (3-4)$$

$$R1 = \frac{\text{检索出的类似商标个数}}{\text{系统中的类似商标总个数}} * 100\% \quad (3-5)$$

其中，查准率是检索出的相似的商标的个数和检索出的商标总个数之间的比值，最后乘以 100%；而查全率指的是检索出的相似的商标图像的个数与整个所用的商标图像数据集中所有商标的个数的比值，最后再乘以 100%。查准率反应的是检索的准确率，查全率反应的是检索是否能将所有相似商标图像找出来。二者结合可以更好地评判检索效果。

3.3.2 变化后的查准率与查全率

对于数据集 2，即包含有 13 中转换的数据集，我们借鉴了林传力等人的论文^[6]中使用的查准率 P2 与查全率 R2，其定义如下：

$$P2_n = 1 - \frac{\sum_{i=1}^n (\lg R_i) - \sum_{i=1}^n (\lg i)}{\lg \left(\frac{N!}{(N-n)!n!} \right)} \quad (3-6)$$

$$R2_n = 1 - \frac{\sum_{i=1}^n R_i - \sum_{i=1}^n i}{n(N-n)} \quad (3-7)$$

其中， R_i 第 i 个相关的商标在结果中的排位， n 为商标数据库中所有相关图像的数量， N 为数据库总的商标数。查准率指的是：若共有 n 个相关图像，先计算所有相关图像所在位置的对数的和，再减去对 1 到 n 每个数求对数的和，这个差作为分子。再计算总商标数量的阶乘与数量作差的阶乘与 n 的阶乘的乘积的比值的对数，最后用 1 减去这个整体的比值。查全率基本同理。只是去掉了所有的对数的计算。

3.4 小结

本章是实验的理论基础，分别介绍了：用深度卷积神经网络来对商标图像进行特征提取的过程，包括先要对商标图像分一下类得到训练集，然后用该训练集训练深度卷积神经网络 VGGNet，最后用该网络及训练过的网络进行商标特征提取；检索时需要用到的相似性距离，包括：欧几里得距离，高斯核函数；最后检索结果的判断指标。

第四章 实验结果与分析

4.1 数据集的建立与图像预处理

4.1.1 数据集的建立

本实验中的数据集来源是商标公司提供的 20 余万张商标图像，实验随机取了其中的 50 张图像作为 50 类图像的基准图像，然后使用已有的神经网络检索出每类相似的 50 张图像，合在一起共计 1244 张商标图像（由于每一类中选出的相似图像可能与其他类中重合，所以总数未达到 2500 张）作为数据集 1，图 4 -1 展示的是每类的基准图像。然后又对上述选出的 50 类图像中每一张进行不同角度的旋转，翻转，模糊等 13 项操作，得到总数为 17500 张的数据集 2，图 4 -2 展示的是对一张图像的 13 项操作后的效果。表 4 -1 展示的是两个数据集的具体信息。

本实验中，在训练网络和特征提取部分，均用到上述两个数据集。

表 4 -1 数据集 1 与数据集 2 的信息

	类数（个）	每类中商标个数（个）
数据集 1	50	50
数据集 2	50	700



图 4 -1 50 张基准商标图像，代表 50 类

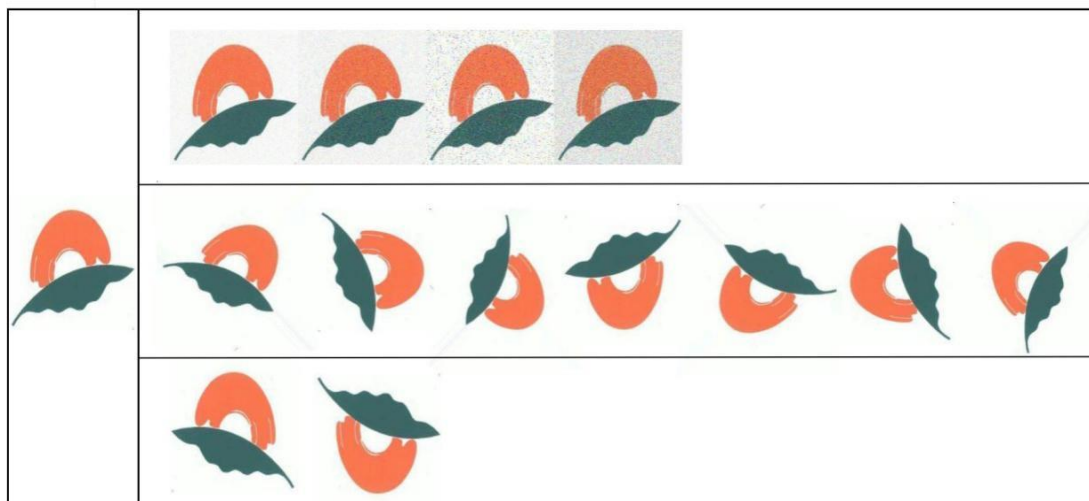


图 4-2 对一张样例商标进行的 13 种操作。

对于上图 4-2，其中第一列是原图像，第二列第一行是添加了不同种类的噪声，对应分别为：poisson、gaussian、salt&pepper、speckle；第二行为原图像顺时针依次进行 45 度、90 度、135 度、180 度、225 度、270 度、315 度旋转后的结果；第三行是对商标进行左右位置变换和上下翻转位置变换后的结果。

4.1.2 商标图像预处理

因为原商标数据库里的商标图像大小不一，为了方便运算，本实验在训练网络和提取特征时均会对输入图像进行预处理，将其尺寸变为 224×224 大小，并且在进行商标检索时，还会减去检索数据集的图像的均值。

4.2 实验设计

取任意一类数据集中后 1/3 中的任意一张图像最为检索图像（因为训练时已经取前 2/3 作为训练集进行训练），分别用原网络 VGGNet、50 类图像训练所得网络 VGGNet1 和 50 类带变换的图像训练所得的网络 VGGNet2 三个网络，在未带变换的 1244 张，和带变换的 17500 张商标数据库中取某一类的 700 张图像进行检索，得到 6 个检索结果。其中在数据集 1 上作为一组，在数据集 2 上作为另一组。在数据集 1 上取前 20 个相似图像，在数据集 2 上取前分别取前 30 和前 20 个相似的结果。并计算对应的查准率与查全率。由于本文选取的相似性度量并不影响检索结果，则默认用欧氏距离检索。最后在把上文中提到的四种其他提取特征的方法：小波系数、HSV 直方图、RGB 通道、颜色相关图以及这四者的结合与 VGGNet2 在数据集 2 上进行比较，得出实验结论。

4.3 实验结果与分析

4.3.1 数据集 1 上的比较

人工剔除		 检索商标
VGGNet		
VGGNet1		
VGGNet2		

图 4-3 三种网络在数据集 1 上的检索结果

表 4-2 三种网络的查准率与查全率

	检索出相关 商标个数 (张)	检索出商标 个数 (张)	总相关商标 个数 (张)	查准率 P1 (%)	查全率 R1 (%)
VGGNet	5	20	50	25	10
VGGNet1	6	20	50	30	12
VGGNet2	7	20	50	35	14

如图 4-3，展示了三种网络模型在数据集 1 上的检索前 20 的结果，其是按照相似度由高到低来排序。其中第三列展示的是待检索的图像。由于数据集 1 是人为设置的标签，而且实验之前没有对于商标图像固定的相似的划分，所以第一行为人为剔除不相似的图片后得到的结果。表 4-2 展示了三种网络的查准率与查全率，由于相似性准则的界限划分不稳定，这里统一将检索出的图形个数设为

20，总相关个数设为该图片所属类别里的总图片数 50。

由表 4-2 可以明显比较出训练的网络检索的性能要优于原网络，此外在大数据集 2 上训练的网络的性能也要优于只在 1244 张图片的数据集 1 上训练的网络。

此外由图 4-3 中可以比较出，在后两个网络检索出的结果中，剔除掉了较多非圆形，并且无文字组合的与原商标相似的商标，这也表明后两个网络的性能会优于原网络。

4.3.2 在数据集 2 上的比较

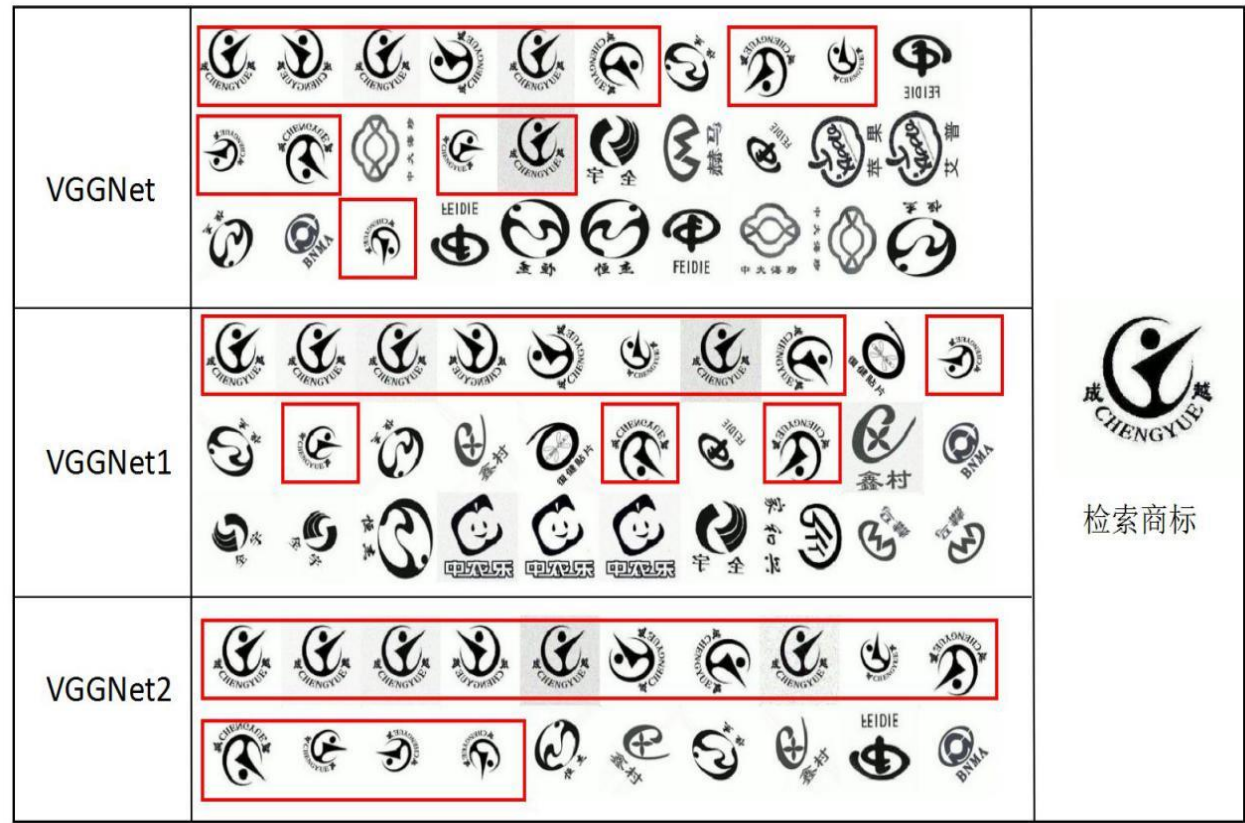


图 4-4 三种网络在数据集 2 上的结果

表 4-3 三种网络每个相关图像的位置

位置	1	2	3	4	5	6	7	8	9	10	11	12	13	14
VGGNet	1	2	3	4	5	6	8	9	11	12	14	15	23	97
VGGNet1	1	2	3	4	5	6	7	8	10	12	16	18	34	93
VGGNet2	1	2	3	4	5	6	7	8	9	10	11	12	13	14

表 4-4 三种网络的查准率 P2 与查全率 R2

	数据库大小 (张)	所有相关图像 (张)	查准率 P2 (%)	查全率 R2 (%)
VGGNet	700	14	99.905	98.906
VGGNet1	700	14	99.897	98.812
VGGNet2	700	14	100	100

图 4-4 显示的是在数据集 2 中取 700 张图像，进行检索的结果。原网络 VGGNet 和 VGGnet1 只取了前 30 张相似图像，而 VGGNet2 检索时把所有相似图像都排在了最前面，所以只显示了前 20 张图像。由图可以看出，原网络 VGGNet 的性能已经很好，但是还是会检索出不相似的图像（即并非同一张图像进行转换后的图像）；而对于 VGGNet1 来说，其性能并没有明显变化；而对于 VGGNet2 来说，它完全检索出了所有的相似图像，性能最优。

表 4-3 显示的是三种网络检索的结果中，每张相似图像所占的位置数。表 4-4 是计算所得的查准率与查全率，由表可知，VGGNet 已经有很好的性能，VGGNet1 虽然经过了多代训练，但是由于训练集中没有包括变换图像，所以性能并没有明显变化，而 VGGNet2 由于是在含有变换后的商标的训练集下学习而得，所以其性能最优。由于总数据量较大，在计算查准率与查全率时，得到的结果相差较小。

4.3.3 四种特征提取算法与 3 种 VGGNet 的比较








VGGNet		 检索商标
小波系数		
HSV直方图		
RGB特征		
颜色相关图		
四种混合		

图 4-5 VGGNet 与四种特征提取算法的对比

表 4-5 前 20 个检索结果中相关商标个数

提取方法	VGGNet (张)	小波系数 (张)	HSV 直方图 (张)	RGB 特征 (张)	颜色相关图 (张)	四种混合 (张)
前 20 个中 相关个数	13	6	10	8	5	10

如图 4-5，左边是原卷积神经网络 VGGNet 和其他四种特征提取的方法在大小为 700 的商标数据集上的检索结果，右边是待检索的原商标图像用例。若只取前 20 张检索出来的商标图像，可以看出，VGGNet 可以检索出更多的相关商标图像，即对待检索图像进行变换后的商标图像。并且相关图像都排在前列。表 4-5 显示的是每个特征提取方法的前 20 张检索结果中相似图像的数目，VGGNet 最多，为 13 个，小波系数为 6 个，HSV 直方图为 10 个，RGB 特征为 8 个，颜

色相关图为 5 个，四种方法的结合为 10 个。综上所述，原网络 VGGNet 的检索效果已经优于其他方法。

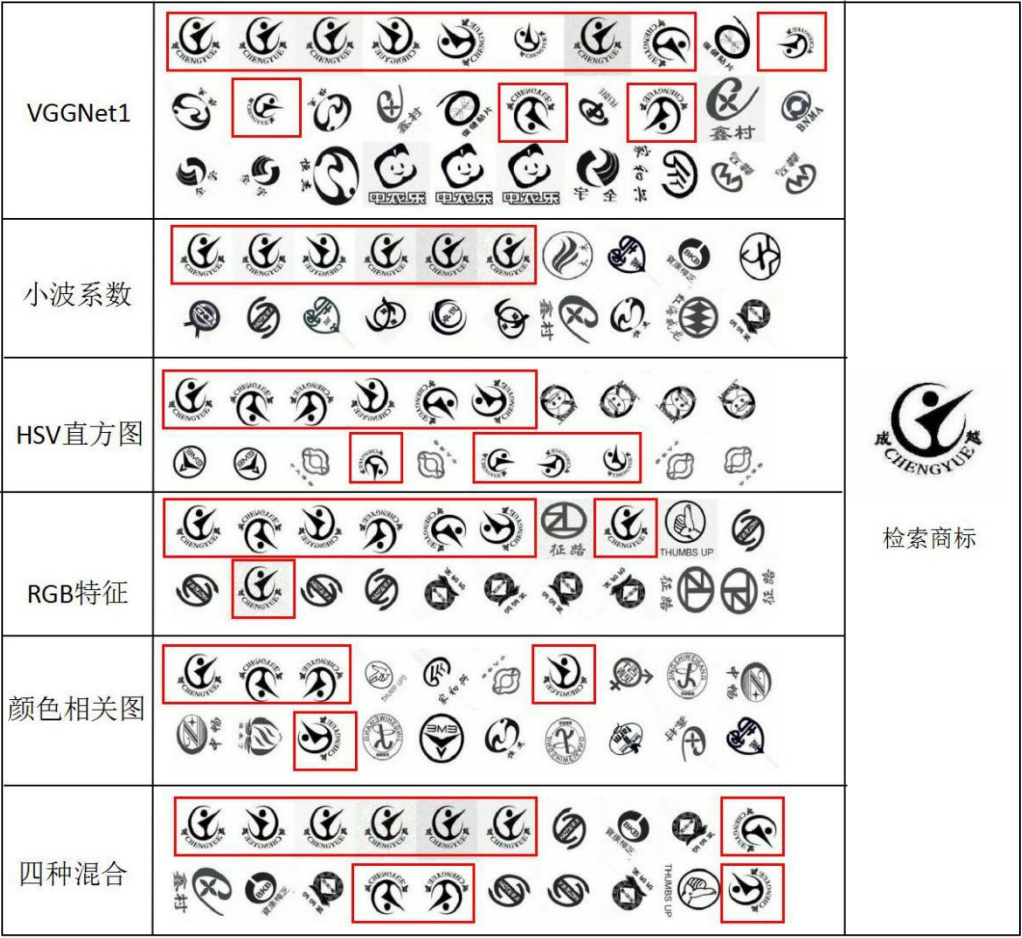


图 4-6 VGGNet1 与四种特征提取算法的对比

表 4-6 前 20 个检索结果中相关商标个数

	VGGNet1	小波系数	HSV 直方图	RGB 特征	颜色相关图	四种混合
	(张)	(张)	(张)	(张)	(张)	(张)
前 20 个中 相关个数	12	6	10	8	5	10

如图 4-6，左边是经过第一次训练的网络 VGGNet1 和其他四种特征提取的方法在大小为 700 的商标数据集上的检索结果，右边是待检索的原商标图像用例。若只取前 20 张检索出来的商标图像，可以看出，VGGNet1 可以检索出更多的相关商标图像，即对待检索图像进行变换后的商标图像。并且相关图像都排在前列。表 4-6 显示的是每个特征提取方法的前 20 张检索结果中相似图像的数量，VGGNet1 最多，为 12 个，小波系数为 6 个，HSV 直方图为 10 个，RGB 特征为

8 个，颜色相关图为 5 个，四种方法的结合为 10 个。综上所述可以看出，虽然第一次训练所得的网络 VGGNet1 虽然不如原网络 VGGNet 的效果好，但是相比于其他四种方法，性能较好。



图 4-7 四种特征提取方法与 VGGNet2 的比较

表 4-7 前 20 个检索结果中相关商标个数

	VGGNet2 (张)	小波系数 (张)	HSV 直方图 (张)	RGB 特征 (张)	颜色相关图 (张)	四种混合 (张)
前 20 个中 相关个数	14	6	10	8	5	10

图 4-7 是将四种方法：小波系数、HSV 直方图、RGB 特征和颜色相关图，以及四者的组合和训练好的 VGGNet2 进行比较得到的结果。可以看出：由于这几种提取的仅仅是纹理特征，或者颜色特征，或者其组合，它们的效果都不如 VGGNet2。VGGNet2 把检索商标经变换后的所有 14 中相似图像都找了出来，并排在了前 14 位。而其他方法都只找出部分相似图像，并且前 14 位中也夹杂了其他的商标图像。表 4-7 显示的是每个特征提取方法的前 20 张检索结果中相似图像的数目，VGGNet2 最多，为 14 个，小波系数为 6 个，HSV 直方图为 10 个，

RGB 特征为 8 个, 颜色相关图为 5 个, 四种方法的结合为 10 个。综上所述可以看出, VGGNet2 的效果既优于四种其他特征提取方法, 也优于原网络 VGGNet 和训练一次所得的网络 VGGNet1。



图 4-8 五种特征提取方法和人工剔除的比较

表 4-8 五种特征提取方法和人工剔除的查准率与查全率

方法	检索出相关 商标个数 (张)	检索出商标 个数 (张)	总相关商标 个数 (张)	查准率 P1 (%)	查全率 R1 (%)
VGGNet2	18	20	50	90	36
小波系数	11	20	50	55	22
HSV 直方图	16	20	50	80	32
RGB 特征	12	20	50	60	24
颜色自动相 关图	11	20	50	55	22

图 4-8 是将 VGGNet、小波系数、HSV 直方图、RGB 特征和颜色相关图这五种特征提取方法的检索结果与人工剔除相比较，由图可知，卷积神经网络 VGGNet2 可以更好地检索出相似图片，既包括对同一张进行翻转等操作，也包括人工剔除出来的相似图像。表 4-8 显示的是种特征提取方法和人工剔除的查准率与查全率，通过表中的查准率与查全率也可以看出，VGGNet2 的效果更好。

4.4 小结

由以上实验结果可以得出，商标检索效果最好的是训练得到的网络 VGGNet2，而训练得到的 VGGNet1 和原网络 VGGNet 之间差别不大。VGGNet2 的效果也好于其他特征提取算法（小波变换，HSV 直方图，RGB 特征和颜色相关图）。

第五章 总结与展望

5.1 总结

随着注册公司的增加,商标注册也成为了一个被广泛考虑的问题。商标检索作为图像检索的一个具象的应用,其重要性越来越显而易见。传统意义上的基于文本的检索技术由于其低效,费时费力的缺点,已经无法适应要求。而随之产生的基于内容的检索技术也在查询方便存在弊端。但是,深度卷积神经网络的提出和应用,可以优化检索流程,提高检索的精确度和速度。本文就将深度神经网络应用于商标检索当中,并提出了一个使检索更加准确的网络模型。本文主要工作如下:

1. 阐述了研究目标,分析了商标检索和深度学习的现状,提出了基于深度卷积神经网络的商标检索技术;
2. 处理并获得了实验数据集,使用了 20 万张商标数据的一个子集作为数据集,并变换图像,形成新的数据集;
3. 用商标数据对深度神经网络进行了训练,得出特定的网络模型;
4. 进行了实验,将不同的网络,以及其他特征提取算法应用于检索并比较,对实验进行分析;
5. 最后得到分析结果:训练后的卷积神经网络可以提高检索性能;
6. 得到检索最佳模型:基于深度卷积神经网络的商标检索。

5.2 展望

基于深度学习,尤其是深度卷积神经网络的商标检索虽然可以优化检索方法,但是由于网络的可调节性大,随着层数的设置,参数的选取,数据集的搜集整理的不同,其结果也不同。本文对基于深度学习的商标检索技术进行了初步的分析和实验,其研究工作还可以在以下几个方面开展:

1. 搜集分类准确详细的商标数据集,或者通过人为或算法来对为分类的商标数据进行准确分类,用该数据集对深度卷积神经网络进行训练;
2. 尝试其他深度卷积神经网络,比如 AlexNet 等,对比检索效果;
3. 尝试人为去改变网络的层数和结构,包括卷积层和全连接层的数目,卷积核的大小,池化的函数选取,隐层的计算函数的选取等,并进行实验分析效果;
4. 把其他的特征提取方法的特征与深度学习提取的特征进行结合,进行实验并分析结果。

希望后续能将我的实验工作进行完善,再对已经得到的基于深度学习的商标

检索的方法进行完善。

参考文献

- [1]张骞. 基于文本的与基于内容的图像检索技术比较研究[J]. 情报探索, 2012(1):0-0.
- [2]Smeulders A W M, Worring M, Santini S, et al. Content-Based Image Retrieval at the End of the Early Years[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000, 22(12):1349-1380.
- [3]Kokare M, Chatterji B N, Biswas P K. A Survey on Current Content based Image Retrieval Methods[J]. Iete Journal of Research, 2015, 48(3-4):261-271.
- [4]Bala A, Kaur T. Local texton XOR patterns: A new feature descriptor for content-based image retrieval[J]. Engineering Science & Technology An International Journal, 2016, 19(1):101-112.
- [5]Rothschild L M. Image annotation for image auxiliary information storage and retrieval[J]. 2017.
- [6]林传力, 赵宇明. 基于 Sift 特征的商标检索算法[J]. 计算机工程, 2008, 34(23):275-277.
- [7]林克正, 张彩华, 刘丕娥. 基于分块主颜色匹配的图像检索[J]. 计算机工程, 2010, 36(13):186-188.
- [8]孙君顶, 毋小省. 纹理谱描述符及其在图像检索中的应用[J]. 计算机辅助设计与图形学学报, 2010, 22(3):516-520.
- [9]赵宏中, 张彦超. 基于 Canny 边缘检测算子的图像检索算法[J]. 电子设计工程, 2010, 18(2):75-77.
- [10]赵宏伟, 李清亮, 刘萍萍, 等. 特征点显著性约束的图像检索方法[J]. 吉林大学学报(工), 2016, 46(2):542-548.
- [11]林妙真. 基于深度学习的人脸识别研究[D]. 大连理工大学, 2013.
- [12]Gordo A, Almazán J, Revaud J, et al. Deep Image Retrieval: Learning Global Representations for Image Search[J]. 2016.
- [13]Pandian A A, Balasubramanian R. Performance Analysis of Texture Image Retrieval in Curvelet, Contourlet, and Local Ternary Pattern Using DNN and ELM Classifiers for MRI Brain Tumor Images[C]// International conference on computer vision and image processing. 2017.
- [14]李钊, 卢苇, 邢薇薇, 等. CNN 视觉特征的图像检索[J]. 北京邮电大学学报, 2015, 38(s1):103-106.
- [15]Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
- [16]Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. 2014:1-9.
- [17]Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
- [18]He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition. IEEE, 2015:770-778.
- [19]Vedaldi A, Lenc K. MatConvNet: Convolutional Neural Networks for MATLAB[C]// The, ACM International Conference. ACM, 2015:689-692

外文资料

Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105

中文译文

深度卷积神经网络分类 ImageNet 数据集

摘要

我们训练了一个大型深度的卷积神经网络，用于将 ImageNet LSVRC-2010 比赛中的 1.2 百万的高分辨率图像分成 1000 类。在测试集方面，我们的 top-1 和 top-5 错误率达到了 37.5% 和 17.0%，要远远优于已有的算法。这个神经网络，有 6 千万的参数，650,000 的神经元，包含 5 个卷积层，有一些后面跟有 max-pooling 层，最后还有 3 层全连接层和最终的 1000-way 的 softmax。为了使训练速度更快，我们用了不饱和的神经元和一个非常高效的可以进行卷积操作的 GPU。为了减少全连接层的过拟合，我们应用了一个最近刚刚开发的正则化方法，叫做“dropout”，这个方法被证明十分有效。我们将这个模型的变体应用于 ILSVRC-2012 比赛当中，达到了最好的 top-5 测试错误率 15.3%，而第二名则是 26.2%。

第一章 介绍

近几年，机器学习方法应用在物体识别领域。为了提升这些方法的性能，我们搜集了一个很大的数据集，学习了更强大的模型，还用了更好的技术去避免过拟合。直到最近，有标签的数据集相对较小——数以万计图片（比如，NORB[16], Caltech-101/256[8,9], 和 CIFAR-10/100[12]）。简单的识别方法可以很好地解决在这些小数据机上的分类问题，尤其是当添加了保存标签的变换之后。比如，现在最好的在 MNIST 数字识别的任务上的错误率（ $< 0.3\%$ ）已经和人工检索的效果接近[4]。但是在现实当中的物体都有很大的差异性，所以为了学习去识别这些物体，使用大规模的训练集是有必要的。同时，小数据图像集的缺点也被发现（比如，Pinto et al.[21]），但是到最近一段时间，搜集带标签的拥有百万图像的数据集才变成可能。这些数据集中包括 LabelMe[23]，里面存有成百上千的全分割的图像，同时还有 ImageNet[6]数据集，其中包含超过 15, 超过 22000 类的数百万带标签的高分辨率的图像。

为了从百万的图像里面学得上千的物体，我们需要一个拥有很大的学习能力的模型。然而，物体识别的巨大的复杂度意味着即使是像 ImageNet 一样大的数据集，也很难解决问题，所以我们的模型也需要有很多先验的知识来补偿缺少的数据。而卷积神经网络则可以实现这一类的模型[16,11,13,18,15,22,26]。卷积神经网络的运算能力可以通过改变深度和宽度来控制，并且对于自然的图像，它们可以得到更准确的假设（统计平稳性，像素依赖局部性）。因此，和拥有相似大小层数的前馈神经网络相比，卷积神经网络拥有更少的连接和参数，所以更容易去训练。最后的性能可能只会差一点。

尽管卷积神经网络有着吸引人的特点和相对高效有效的结构,但是还是很难应用于大规模的高分辨率图像,其代价很大。幸运的是,现存的 CPU 可以高度优化地使用 2 维的卷积,这样就可以对大的卷积神经网络进行训练,同时现存的类似于 ImageNet 的数据集包含有足够的有标签的样例图像,用以去训练模型并防止严重的过拟合。

这篇论文的贡献如下:我们训练了很大的卷积神经网络作用于 ILSVRC-2010 和 ILSVRC-2012 的数据集 ImageNet 的子集,并达到了在这个数据集上有史以来最好的效果。我们开发了一个可以操作 2 维卷积和其他在训练 CNN 时会用到的操作的高优化的 GPU。我们的网络包含大量的新的不常见的特点,这些特点提升了网络的表现,减少了训练时间,详细的信息参见第三部分。我们的网络的大小使得过拟合成为一个重要的问题,尤其是在 1.2 百万的带标签的训练集上面,所以我们用了几种有效的技术来防止过拟合,详细信息参见第四部分。我们最终的网络包含了 5 个卷积层,3 个全连接层,这个深度很重要:我们发现移走任何一个卷积层(每层有着不超过 1%的参数)都会降低网络的效果。

最后,网络的大小主要是受现存的 GPU 的内存,以及我们能接受的训练时间的限制。我们的网络在两个 GTX 580 3GB GPU 上交替运行花费了 5 到 6 天。所有我们的实验都表明,实验的最终结果可以通过更快的 GPU 和更大的数据集来提高。

第二章 数据集

ImageNet 数据集含有超过 15 百万的带标签的高分辨率图像,分属于大约 22000 类。这些图像来源于网络,并由人工标注标签,其中用到了亚马逊的 Mechanical Turk 的工具。始于 2010 年,作为 Pascal Visual Object Challenge 的一个部分,一个名为 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 的比赛开始举行。ILSVRC 用了 ImageNet 数据集的一个子集,大概 1000 类,每类各 1000 张图片。总计大约有 1.2 百万的训练图像,50000 验证图像,150000 测试图像。

ILSVRC-2010 是 ILSVRC 中唯一的一版有带标签的测试图像,所以我们的绝大多数实验都是基于这个版本。因为我们同时也在 ILSVRC-2012 的比赛中应用了我们的模型,在第六部分我们展示了这个版本的结果,但是里面的测试图像没有标签。在 ImageNet 上,通常有两个错误率:top-1 和 top-5, top-5 错误率指的是当正确的标签并不在模型提供的前五个最可能的标签里时。

ImageNet 包含可变分辨率的图像,但是我们的系统需要一个固定维度的输入。因此,我们将输入图像进行下采样至一个固定的分辨率 256×256 。给出一个方形的图像,我们首先重新调节这张图像使得短边长为 256,然后裁剪中间的 256×256 大小的图像。我们没有对图像进行其他的预处理,除了给每个像素减

去训练集的平均值。所以我们是成熟的 RGB 值的基础上训练我们的网络的。

第三章 结构

我们的网络的结构如 figure 2。它包括 8 层网络——5 层卷积和 3 层全连接。接下来，我们将介绍网络中的一些有特点的结构。Section 3.1-3.4 是根据我们认为的重要性排列，最重要的排在第一位。

3.1 ReLu Nonlinearity

给定一个神经元的输入 x ，一般对其输出的建模为一个函数 f ，其中 $f(x) = \tanh(x)$ 或者 $f(x) = (1 + e^{-x})^{-1}$ 。而就梯度下降的训练时间来说，这种饱和的非线性函数要远远慢于不饱和的非线性函数 $f(x) = \max(0, x)$ 。根据 Nair 和 Hinton 的理论[20]，我们参考这个非线性函数作为神经元并称为 Rectified Linear Units (ReLU)。带有 ReLU 的深层卷积神经网络训练得要快于同样的带有 \tanh 的单元。如 Figure 1，展示了为了在 CIFAR-10 的数据集上达到 25% 的训练错误率，一个有着 4 个卷积层的网络需要的迭代次数。图中显示如果我们运用传统的饱和和神经元模型，是不能在这么大的神经网络上进行实验的。

我们并不是第一个想到要用替代方法代替传统的 CNN 中所用的神经模型的。比如，Jarrett et al.[11] 提出了非线性神经元函数 $f(x) = |\tanh(x)|$ ，当作用于 Caltech-101 数据集并且把对比归一加在平均池化后，这个函数可以有很好的效果。然而，这个数据集的初衷是避免过拟合，所以与我们用 ReLU 去加速训练集的训练能力的初衷相比，他们所观测的实验效果是不同的。更快的学习过程对在大数据机上的对大型模型的训练的效果是有很显著的影响的。

Figure 1：一个四层的卷积神经网络，伴随有 ReLU（实线）在 CIFAR-10 上达到 25% 的训练错误率，比相同的网络结构但是用 \tanh 神经元（虚线）要快 6 倍。这两种方法的学习率不相同，是为了让各自的训练尽可能地快。此外没有用到任何的正则化方法。这里显示的效果的好坏根据网络的结构的不同而不同，但是拥有 ReLU 的网络要快于同等的但是应用饱和神经元的网络。

3.2 在多 GPU 上训练

一个单一的 GTX 580 GPU 只有 3GB 的存储大小，这就限制了可以用于训练的网络的大小。而 1.2 百万的训练样例足够去训练一个网络，但是对于一个 GPU 来说还是太大了。因此，我们将网络运用于两个 GPU 之间。现存的 GPU 可以很好地适应跨 GPU 并行计算，可以在两个 GPU 之间互相读写，而并不需要通过主机的内存。我们所应用的并行化策略在两个 GPU 上各放一半数目的内核（或称为神经元），但是其中 GPU 之间的交流只发生在特定的层上。也就是说，比如第 3 层从整个的神经元网络的第 2 层得到输入。但是，第 4 层的神经元只从和自己享用同一个 GPU 的第三层的神经元处得到输入。如何选择连接的模式对于交叉验证来说十分重要，但是这允许我们去精确地调整交流的数量，直到达到运算可

以接受的分数。

这种组合结构在一定程度上与 Ciresan et al.[5]提出的“columnar”的 CNN 相似，除了我们的一列不是独立的（如图 Figure2）。与 GPU 训练的只有一半神经元数的卷积层相比，这个结构降低了 1.7% 的 top-1 和 1.2% 的 top-5 错误率。2 个 GPU 也比 1 个的训练速度要稍微快一点。

3.3 Local Response Normalization 局部反馈归一化

ReLU 有一个很好的特性，它们不需要用输入的归一化来避免饱和。如果至少一些训练样例产生了对于 ReLU 层正向的输入，学习就会发生在那个神经元。然而，我们仍然发现以下的局部归一化策略会帮助归纳。一个神经元的活动可以通过将内核 i 应用于位置 (x,y) ，然后应用 ReLU 的非线性函数，我们用 $a_{x,y}^i$ 来表示这个活动，那么与其对应的响应归一化的活动 $b_{x,y}^i$ 可以由下面的表达式来表示：

$$b_{x,y}^i = a_{x,y}^i / (k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a_{x,y}^j)^2)^\beta$$

其中对 n 个临近的位于地图中同一个空间位置的内核进行求和， N 则为该层的总内核数。而内核地图的顺序是在训练之前随意排列的。这种响应归一化的计算使用了侧抑制的方法，这个方法受启发于真正的神经元活动，在使用不同的内核的情况下，对于神经元之间的大型活动创建竞争机制。公式中的常数 k, n, α 和 β 是超参数，它们的取值取决于所用的验证集：我们这里将它们赋值于 $k = 2, n = 5, \alpha = 10^{-4}, \beta = 0.75$ 。我们将上述的归一化应用于每次在特定的层上使用 ReLU 的非线性函数之后（详见 section 3.5）。

上述这个方案容忍了 Jarrett et al.[11]所提到的局部归一化采样策略导致的一些相似性，但是对于我们这里用到的策略来说，称之为“轻亮度归一化”更为准确，因为我们没有减去图像均值。响应归一化将我们的 top-1 和 top-5 错误率分别减少了 1.4% 和 1.2%。我们也证实了这个策略在 CIFAR-10 数据集上的有效性：一个 4 层的 CNN，不经过归一化可以达到 13% 的错误率，而经过归一化则可达到 11% 的错误率。

3.4 Overlapping Pooling 重叠池化

CNN 中的池化层总结了共享同样的内核地图的临近组的神经元的输出。一般来说，临近池化单元所总结的邻居不会重叠（比如[17,11,4]）。为了更加准确，可以将池化层想象成由大小为 s 像素的池化单元形成的网格组成，每个都总结了大小为 $z \times z$ 的以池化单元为中心的临近邻居。如果我们使得 $s = z$ ，我们得到的就是传统意义上的在 CNN 里最常用的局部池化。如果我们让 $s < z$ ，则得到了重叠池化。在这篇论文中所提到的网络中，我们使用的取值为 $s = 2, z = 3$ 。和传

统的 $s = 2$, $z = 2$ 的非重叠池化层相比, 在同样的维度下, 这个策略可以将 top-1 和 top-5 错误率分别减少 0.4% 和 0.3%。在训练过程中, 经观察发现, 重叠池化可以轻微地避免过拟合。

3.5 总结构

现在我们来介绍一下整个网络的结构。如图 Figure2, 整个网络包含 8 个带有权重的层; 前 5 个是卷积层, 剩下的 3 层是全连接层。最后的全连接层得到的输出被用于一个 1000-way 的 softmax, 会产生一个超过 1000 类标签的分类。这个网络最大化了多项逻辑回归的目标, 等同于最大化训练样例上对数概率的均值, 这个对数概率是在预测分布下的正确标签的对数概率。

第 2, 4, 5 层的卷积层的内核只和内核地图上与其共享同一个 GPU 的前一个层相连接 (如图 Figure2)。而第 3 个卷积层的内核与第二层的所有的神经元相连。在第 1 层和第二层卷积层后面都跟随有响应归一化层。在每一个响应归一化层和第 5 个卷积层后, 跟随有最大池化层。而 ReLU 非线性函数层则应用于每个卷积层和全连接层的输出之后。

第一层卷积层的用 96 个 $11 \times 11 \times 3$ 大小的内核来过滤 $224 \times 224 \times 3$ 大小的输入图像, 伴有 4 个像素的步幅 (这个步幅指的是: 在内核地图里由临近神经元组成的区域之间的距离)。第二层卷积层将第一层卷积层得到的输出 (经过响应归一化和池化后) 作为输入, 并且用 256 个大小为 $5 \times 5 \times 48$ 的内核来进行过滤。第 3, 4, 5 层卷积层互相相连接, 没有用到任何介入的池化或归一化层。第 3 层卷积层有 384 个大小为 $3 \times 3 \times 256$ 的内核, 与上一层的输出 (经过归一化和池化) 相连。第 4 层卷积层有 384 个大小为 $3 \times 3 \times 192$ 的内核, 第 5 层卷积层有 256 个大小为 $3 \times 3 \times 192$ 的内核。而每个全连接层都有 4096 个神经元。

Figure2 : 描述我们的 CNN 的结构, 尤其是描述了两个 GPU 之间的联系。图中上下分别是两个 GPU。GPU 之间的交流仅发生在特定的层上。整个网络的输入为 150528 维, 每层的神经元个数分别为: 253440-186624-64896-64896-43264-4096-4096-1000。

第四章 减少过拟合

我们的神经网络有 60 百万的参数。尽管 ILSVRC 的 1000 个类别都对每个训练样例从图像到标签的映射增加了 10 位的约束, 这使得在避免过拟合的情况下对大量的参数的学习不充足。接下来, 我们介绍了 2 个主要的防止过拟合的方法。

4.1 Data Augmentation 数据增强

在图像数据集上, 最简单和常用的方法是人工通过保留标签转换 (label-preserving transformations) (比如[25,4,5]) 来增加数据集。我们采用了两种不同的数据增强方法, 每种都允许在对原图像进行简单的计算后进行转换, 所以转换后的图像不需要存在磁盘上。在我们的使用中, 我们在 CPU 上用 python

生成转换的图像，而在 GPU 上对之前的分批的图像进行训练。所以这些数据增强的策略的计算度是很灵活的。

第一种数据增强方法包含了生成图像的转换和水平的反射。我们随机地从 256×256 大小的图像上提取了 224×224 个小块（包括它们的水平映射），然后在这些小块上训练我们的网络。这使得我们的训练集变成了原来的 2048 倍，即使得到的训练样本会有很大的内部依赖性。但是要是不用这个策略，我们的网络会过拟合，导致只能使用更小的网络。测试时，我们的网络提取了 5 个 224×224 个小块（四个边缘块和一个中心块）和它们的水平映射（因此总共有 10 块），并且在这 10 块上进行 softmax 之后取平均值。

第二种数据增强方法包括改变训练图像的 RGB 通道的强度，尤其是我们在 ImageNet 的训练集上，在 RGB 像素值上进行 PCA 计算。对每个训练图像，我们增加了主要成分的倍数，用相应特征值的大小比例乘上一个由以 0 为均值 0.1 为标准偏差的高斯函数得到的随机的变量。因此对每个 RGB 图像像素 $I_{xy} = [I_{x,y}^R, I_{x,y}^G, I_{x,y}^B]^T$ ，我们增加了如下的量：

$$[P_1, P_2, P_3][\alpha_1 \lambda_1, \alpha_2 \lambda_2, \alpha_3 \lambda_3]^T$$

其中 P_i 和 λ_i 是分别是第 i 个特征向量和 RGB 像素值的协方差矩阵的特征值，其中 α_i 对于每个特定的训练图像的所有像素值都是固定的，直到这个图像被再次用作训练，它才会再次改变。这个增强策略大体上不捉了自然图像最重要的特性，即对于照明的强度和颜色的改变，物体的标识是不变的。这个策略可以把 top-1 错误率降低 1%。

4.2 Dropout

将许多不同的模型结合起来预测是一个减少测试错误的成功的方法[1,3]，但是对于一个已经花费几天的时间训练的大型神经网络来说，就花费太大了。有一种非常高效的模型结合的版本，仅需要两倍的训练时间去训练。近期出现的一种叫做“dropout”的技术，它将每层出现概率为 0.5 的隐藏层的输出设为 0。参与到“dropout”的神经元不会对向前传播和反向传播产生影响。所以每当有新的输入到网络时，网络都会形成不同的结构，但是所有的这些结构都共享同样的权重。这项技术降低了神经元之间的自适应的复杂度，因为一个神经元不能全部依赖于另一个神经元。因此，在许多随机的神经元的子集的结合中，学习更健壮的特点就十分有用了。测试时，我们使用了所有的神经元，但是把输出都乘了 0.5，这样一来，对于多指数的 dropout 网络所产生的几何预测分布，就会更确切地进行逼近。

我们在 Figure2 中的两个全连接层后都用到了 dropout 技术。如果没有 dropout，网络将会产生过拟合。Dropout 大约加倍了迭代次数。

第五章 网络学习的细节

我们用批大小为 128 个样例的随机梯度下降方法来训练模型，动量为 0.9，权重衰减为 0.0005。这个权重衰减值虽然很小，但是对于模型的学习来说十分重要。也就是说，这里的权重衰减并不仅仅是一个正则化矩阵：它还减少了模型的训练错误率。对于权重 w 来说，更新规则为：

$$v_{i+1} := 0.9 \cdot v_i - 0.0005 \cdot \epsilon \cdot \omega_i - \epsilon \cdot \left\langle \frac{\partial L}{\partial w} |_{w_i} \right\rangle_{D_i}$$

$$w_{i+1} := w_i + v_{i+1}$$

其中 i 是迭代索引数， v 是动量变量， ϵ 是学习率， $\left\langle \frac{\partial L}{\partial w} |_{w_i} \right\rangle_{D_i}$ 是：以 w_i 为权重的第 i 个分批的均值对 w 的导数。

我们用 0 均值，标准差为 0.01 的高斯分布函数来初始化权重。我们将第 2,4,5 层和全连接的隐藏层的神经元偏差设为常数 1。这个初始化加速了早期的学习速度，因为它给 ReLUs 提供了正输入。对于剩下的层数的神经元偏差，我们统一设为常数 0。

对于所有网络中的所有层，我们使用相同的学习率，这个学习率是在训练过程中人工调整的。这里我们用到的启发式方法是：当验证错误率不再增长时，我们会将原学习率除以 10 作为更新的学习率。学习率初始化为 0.01，在整个过程中会进行 3 次更新。在 1.2 百万图像最为训练集的情况下，我们训练了 90 次网络，在 NVIDIA GTX 580 3GB GPUs 上花费了五到六天。

第六章 结果

在 ILSVRC-2010 上的结果见表 1。我们的网络达到的 top-1 和 top-5 测试集错误率分别为 37.5% 和 17.0%。而在 ILSVRC-2010 的比赛中最好的效果是 47.1% 和 28.2%，这个结果是由对由 6 个稀疏编码的提取不同特征的模型产生的预测进行取均值后得到的[2]。从那之后的最好的效果是 45.7% 和 25.7%，这个结果是由对 2 个在 Fisher Vectors (FVs) 上训练的分类器进行预测后取均值得到的[24]。

我们同样将模型应用于 ILSVRC-2012 的竞赛中，其结果见表 2。因为 ILSVRC-2012 的测试集的标签没有公开，我们不能展示我们尝试的所有模型的测试错误率。所以在接下来的部分，我们用验证错误率来替代测试错误率，这么做是因为在整个试验中，验证和测试的错误率区别不到 0.1%（见表 2）。本篇论文中提到的 CNN 可以达到 18.2% 的 top-5 错误率。而其他 5 个相似的 CNN 所得到的错误率的均值为 16.4%。在 ImageNet Fall 2011（15M 图像，22K 种类）上训练一个附加有 6 个卷积层，1 个池化层的 CNN，然后在 ILSVRC-2012 上进行调整，其达到了 16.6% 的错误率。在上述的 5 个 CNNs 的分类得到的整个 Fall 2011 的数据集上，对 2 个 CNNs 得到的预测取平均，得到 15.3% 的错误率。排名第二的结果来自于，对在 FVs 上训练的网络得到的预测取平均，为 26.2%。这个 FVs 是进行了密集采样[7]。

最后我们还在 Fall 2009 版本的有 10184 个种类, 8.9 百万张图像的 ImageNet 上得到了网络的错误率。在这个数据集上, 我们用一半的图像进行训练, 另一半进行测试。因为没有一致的测试集, 我们所用的分割训练集测试集的方法依赖于前人的方法, 但是这并没有对结果产生明显的影响。我们的 top-1 和 top-5 错误率为 67.4% 和 40.9%, 特别注意的是我们这里将原网络后面增加了 6 个卷积层。而以往在这个数据集上的最好的效果是 78.1% 和 60.9% [19]。

6.1 定性评价

Figure 3 展示了由网络的两个数据联系层得到的卷积核。这个网络学习了不同的频率方向选择的内核, 也就是图中不同颜色的点。注意两个 GPU 的不同之处, 限制连接度的结果在 Section 3.5 里面介绍。GPU 1 里面的内核的颜色不是很明显, 但是 GPU 2 里面的内核的颜色差异性很大。这种特殊性会出现于每次运行网络的时候, 并且在每次初始化权重后, 其特性也不同 (会对 GPU 模块重新更新)。

在 figure 4 的左半部分, 我们评估了网络在 8 张测试图片上计算了 top-5 错误率的预测后学到了什么。注意到, 即使是偏离中心的物体, 比如说图中左上角的小虫, 网络也是可以识别到的。所有的 top-5 标签都是相关联的并且合理的。比如, 在识别豹子的时候, 只会出现相关的其他猫科动物的标签。在另外的一些样例里 (栅栏, 樱桃), 对于照片的目标焦点确实也会存在歧义。

另一种用来探查网络的视觉效果的方法是考虑, 一个图像的隐层得到的至少 4096 维的特征激活值。如果两张图像的特征值向量之间的欧氏距离很小, 就说明这两张图像是相似的。Figure 4 展示了在测试集中的 5 张图像作为被检索图像, 和检索出训练集中的 6 张相似图像, 图像之间是通过计算欧氏距离得到的相似性。注意到在像素级别上, 检索出的最相似的图像和检索图像之间并不是最接近的。比如, 检索出的大象和狗都有不同的姿势。我们在后续的材料中展示了更多的实验结果。

通过计算 2 个 4096 维的特征向量真实值之间的欧氏距离来计算相似性并不高效, 但是通过训练自动编码器来将向量压缩成短的二进制码, 则会使得计算更加高效。这会产生一个和将自动编码器应用于像素值相比更加有效的检索方法 [14], 它不会使用图像的标签, 因此会根据图像的边缘的相似性模式来检索, 也就是说更具有语义相似性。

第七章 讨论

我们的结果表明, 一个深层次大型卷积神经网络能够在高挑战性的数据集上采用先进的学习方法得到突出的效果。需要注意的是, 即使移走一层卷积网络, 我们的网络的性能将会降低。比如移走网络中间任意一层, 最后的 top-1 结果都会损失 2%。所以为了达到预期的目标, 网络的深度是很重要的。

为了简化实验，我们没有使用任何非监督的预训练方法，即使这可能对实验有所帮助，尤其是当我们有更强的运算能力在不获取更多的带标签的数据的基础上去增加网络的大小时。因此，我们的网络性能会在网络变大和训练时间增加的基础上提升，但是在人类视觉系统上，我们仍有很长的路要走。最后，我们还希望可以在视频序列上运用更大更深的卷积网络，跟静态的图像相比，视频中的结构提供了更加有用的信息。

致 谢

在本论文完成之际，谨向所有帮助过我的老师和同学致以最真挚的谢意。

首先，要感谢我的指导老师谢宗霞老师，在从毕业设计选题，开题，到后来的实验主体和论文写作部分都悉心指导我去完成，每当我有不懂的问题，老师都耐心地解答，并敦促我完成毕业设计。感谢老师严谨的教学态度，严格的要求，悉心的教导。老师的这些对待课题的态度以及对工作的热情会对我以后的工作学习产生积极的影响。

其次，感谢参与到我的毕业设计的过程中的老师和同学们，感谢他们对我提供的帮助，提出的建议，让我能够更好地完成毕业设计和论文的写作。

最后，感谢我所引的文献的作者和机构，感谢他们为我提供了完善的资料。帮助我理解题目，完成实验。