数据分析技术

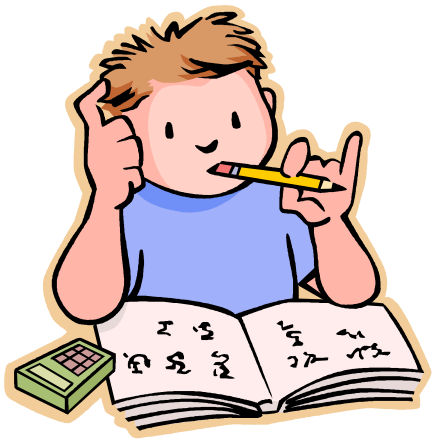
01

绪　论

授课对象：计算机专业学生

第一章　绪论

- 数据挖掘技术的产生与概念
- 数据挖掘概念
- 数据挖掘技术的分类问题
- 数据挖掘常用的知识表示模式与方法
- 不同数据存储形式下的数据挖掘问题
- 粗糙集方法及其在数据挖掘中的应用
- 数据挖掘的应用分析



一、课程简介

大量的、不完全的、有噪声的、模糊的、随机的数据集

数据处理、统计分析、机器学习、模式识别等诸多方法

有效的、新颖的、潜在有用的，以及最终可理解的模式

数据分析

大数据

大数据与数据挖掘

- **大数据**：指一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

- **数据挖掘**：数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的信息，以及最终可理解的模式的非平凡过程。它是一门涉及面很广的交叉学科，包括机器学习、数理统计、神经网络、数据库、模式识别、粗糙集、模糊数学等相关技术。

大数据的学习内容

- **第一、大数据基础**：包括面向对象、多线程、反射、动态代理、JDBC、Servlet、JSP、 MySQL、SQL语句操作以及Java开发管理系统实战
- **第二、Hadoop生态技术**： hadoop集群搭建及架构原理、 Hdfs、 MapReduce、Yarn、 Hive、 HBase、Azkaban、Sqoop等周边技术、Hadoop企业级项目实训
- **第三、Storm实时流计算**： Storm集群搭建及组件介绍、 Topology程序开发
- **第四、Spark生态技术**：Spark环境搭建、基础原理及运行架构、 Scala、SQL与DataFrame技术

数据分析技术学习的主要内容

数据分析技术学习的主要内容

- 基础知识、概念
- 知识发现过程与应用：数据抽取与集成、清洗与预处理、选择与整理、不同数据存储形式下的数据挖掘问题
- 数据挖掘中的主要技术：包括关联规则、分类、聚类、序列
- 空间以及Web挖掘技术
- 数据挖掘十大算法

- 数据挖掘技术的产生与概念
- 数据挖掘的发展趋势
- 数据挖掘技术的分类问题
- 数据挖掘常用的知识表示模式与方法
- 不同数据存储形式下的数据挖掘问题
- 粗糙集方法及其在数据挖掘中的应用
- 数据挖掘的应用分析

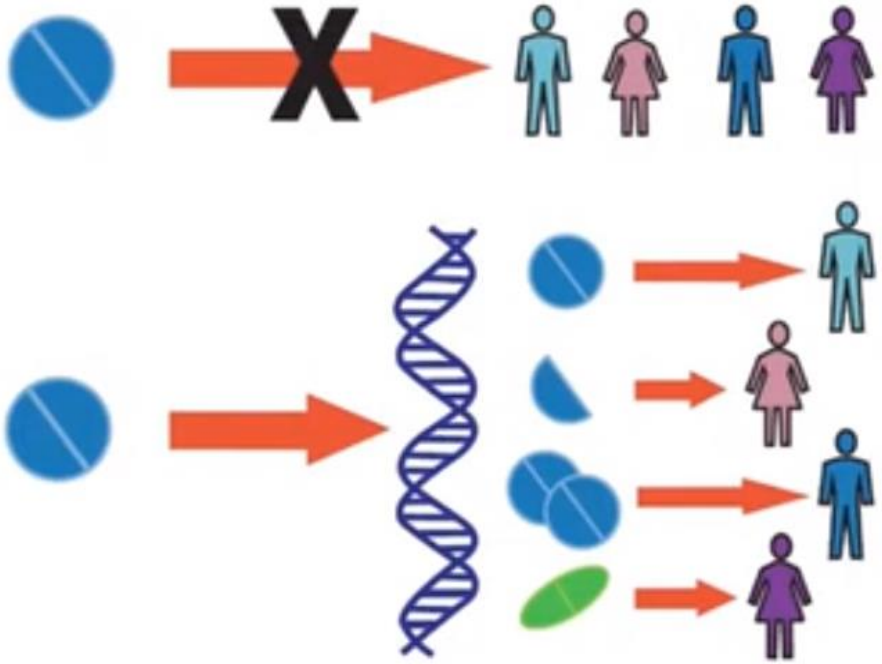
数据库系统的发展

70年代	80年代	90年代	本世纪
简单文件处理系统向数据库系统变革，层次、网络和关系型数据库普及。	RDBS及其相关、数据索引被广泛采用，分布式数据库广发讨论，关系数据库技术和新型技术的结合。	数据库领域中的新内容、新应用、新技术层出不穷，形成了庞大的数据库家族。	Data mining得到理论/技术深化

- 数据挖掘方法可以是基于数学理论的，也可以是非数学的；可以是演绎的，也可以是归纳的。
- 从研究者可能是来自于数据库、人工智能、数理统计、计算机科学以及其他方面的学者和工程技术人员，他们会从不同的视点进行探讨性研究。
- 有下面一些重要的理论视点值得关注：
  - 模式发现（Pattern Discovery）架构
  - 规则发现（Rule Discovery）架构
  - 基于概率和统计理论
  - 微观经济学观点（Microeconomic View）
  - 基于数据压缩（Data Compression）理论
  - 基于归纳数据库（Inductive Database）理论
  - 可视化数据挖掘（Visual Data Mining）
  - 等等

未来的医疗

Effectiveness Research



Personalized Medicine

巨头们的数据挖掘之路

- 1、**谷歌**:包括之前的MapReduce、Word2Vec、BigTable，近期的BERT。数据挖掘是谷歌研究的一个重点领域。2020年谷歌全球不同研究中心在数据挖掘顶级国际会议KDD上一共发表了7篇文章。
- 2、**亚马逊**:亚马逊也在数据挖掘领域开始占有一席，尤其是在人才网罗、开源、核心技术研发。2020年亚马逊在数据挖掘顶级国际会议KDD的Applied Data Science Track（应用数据科学Track）上一共发表了2篇文章，另外还有两个应用科学的邀请报告。
- 3、**微软**:2020年在数据挖掘顶级国际会议KDD上一共发表了6篇文章(微软作为第一作者的文章)

巨头们的数据挖掘之路

- 4、**百度**:2020年百度在数据挖掘顶级国际会议KDD上作为第一作者单位一共发表了2篇文章。
- 5、**阿里巴巴**:电子商务方面,2020年阿里巴巴在数据挖掘顶级国际会议KDD上作为第一作者单位一共发表了8篇文章。
- 6、**腾讯**:2020年腾讯在数据挖掘顶级国际会议KDD上作为第一作者单位一共发表了2篇文章。

2018年图灵奖公布！Hinton、Bengio、LeCun深度学习三巨头共享

Geoffrey E Hinton

主要有三大重要贡献：

反向传播、

玻尔兹曼机、

对卷积神经网络的修正

Hinton, G. E., Osindero, S. and Teh, Y. (2006)
A fast learning algorithm for deep belief nets.
Neural Computation, **18**, pp 1527-1554. [pdf]
Movies of the neural network generating and recognizing digits

AWARD WINNER

Geoffrey E Hinton

ACM A. M. Turing Award (2018)

2018 ACM A.M. Turing Award

hinton个人主页

http://www.cs.toronto.edu/~hinton/

BIOGRAPHICAL INFORMATION
.....Curriculum Vitae ,pdf
.....Biographical sketch
.....Brief Bio
.....Photograph1 (.jpg)
.....Photograph2 (.jpg)

PUBLICATIONS
.....Publications by year
.....Slides of public talks

2012 COURSERA COURSE LECTURES:
Neural Networks for Machine Learning
....Lectures(.mp4)
....Lecture Slides(.pptx or .pdf))

OLD UNIVERSITY OF TORONTO COURSES
....csc321 Spring 2013(undergrad)
....csc2535 Spring 2013(graduate)

VIDEO TALKS & TUTORIALS
....YouTube (2012) Brains, Sex and Machine Learning (1hr)
....YouTube (2007) The Next Generation of Neural Networks (1hr)
....YouTube (2010) Recent Developments in Deep Learning (1hr)
.... Interview on CBC radio "Quirks and Quarks" Feb 11 2011
.... Interview on CBC radio "The Current", May 5 2015
....Tutorial (2009) Deep Belief Nets (3hrs) ppt pdf readings
....Workshop Talk (2007) How to do backpropagation in a brain (2

Geoffrey E. Hinton

Department of Computer Science
University of Toronto
6 King's College Rd.
Toronto, Ontario

email: geoffrey [dot] hinton [at] gmail [dot] com
voice: send email
fax: scan and send email

Information for prospective students:
I advise interns at Brain team Toronto.
I also advise some of the residents in the Google Brain Residents Program.
I will not be taking any more students, postdocs or visitors at the University of Toronto.

News
Results of the 2012 competition to recognize 1000 different types of object
How George Dahl won the competition to predict the activity of potential drugs
How Vlad Mnih won the competition to predict job salaries from job advertisements
How Laurens van der Maaten won the competition to visualize a dataset of potential drugs

Using big data to make people vote against their own interests
A possible motive for making people vote against their own interests

Basic papers on deep learning

LeCun, Y., Bengio, Y. and Hinton, G. E. (2015)
Deep Learning
Nature, Vol. 521, pp 436-444. [pdf]

《A Neural Probabilistic Language Model》提出的生成对抗网络（GAN），这项研究引起了计算机视觉和计算机图形学的革命。

Yoshua Bengio

主要有三大重要贡献：

序列的概率建模

高维词嵌入与注意力机制

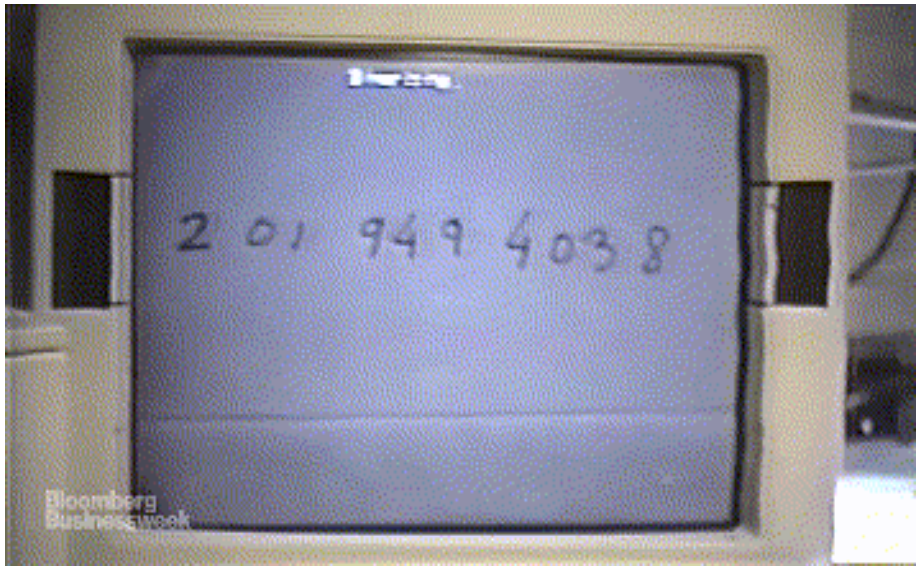
生成对抗网络

AWARD WINNER

Yoshua Bengio

ACM A. M. Turing Award (2018)

2018 ACM A.M. Turing Award



Yann LeCun

主要有三大重要贡献：

提出卷积神经网络

改进反向传播算法

拓宽神经网络的视角

AWARD WINNER

Yann LeCun

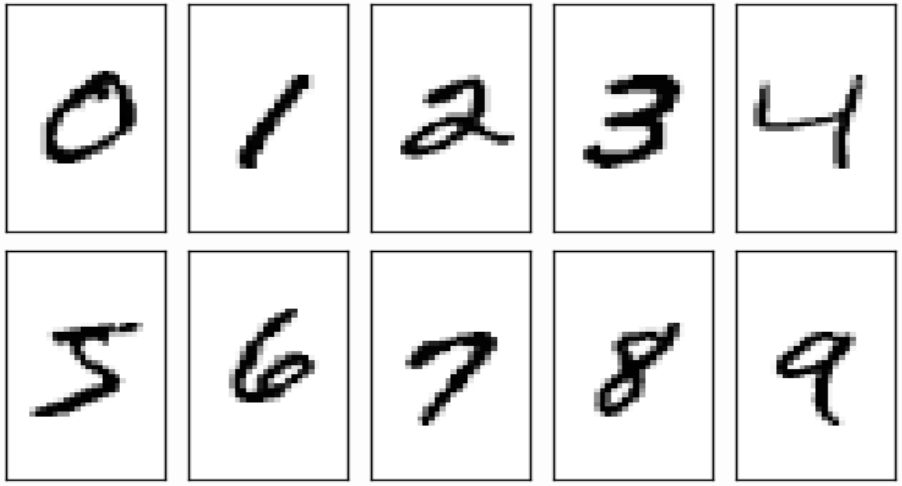
ACM A. M. Turing Award (2018)

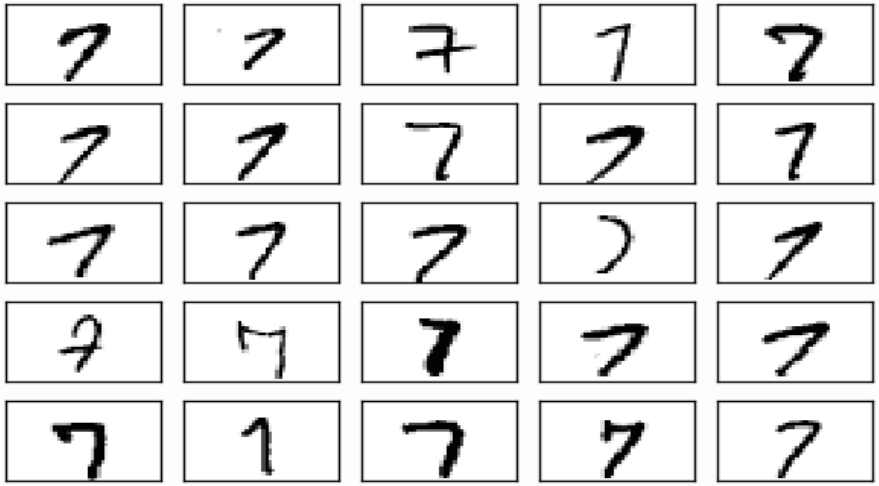
2018 ACM A.M. Turing Award

MNIST 数据集可在 http://yann.lecun.com/exdb/mnist/ 获取, 它包含了四个部分:

- Training set images: train-images-idx3-ubyte.gz (9.9 MB, 解压后 47 MB, 包含 60,000 个样本)
- Training set labels: train-labels-idx1-ubyte.gz (29 KB, 解压后 60 KB, 包含 60,000 个标签)
- Test set images: t10k-images-idx3-ubyte.gz (1.6 MB, 解压后 7.8 MB, 包含 10,000 个样本)
- Test set labels: t10k-labels-idx1-ubyte.gz (5KB, 解压后 10 KB, 包含 10,000 个标签)

MNIST 数据集来自美国国家标准与技术研究所, National Institute of Standards and Technology (NIST). 训练集 (training set) 由来自 250 个不同人手写的数字构成, 其中 50% 是高中学生, 50% 来自人口普查局 (the Census Bureau) 的工作人员. 测试集(test set) 也是同样比例的手写数字数据.





MNIST数据集中的每个图像都是28x28像素，包含一个居中的灰度数字。

什么是卷积?

卷积神经网络:它是使用卷积层（Convolutional layers）的神经网络，基于卷积的数学运算。

-1	0	1
-2	0	2
-1	0	1

卷积层由一组滤波器组成，滤波器可以视为二维数字矩阵。这是一个示例3x3滤波器

什么是卷积操作呢?

1、在图像的某个位置上覆盖滤波器；
2、将滤波器中的值与图像中的对应像素的值相乘；
3、把上面的乘积加起来，得到的和是输出图像中目标像素的值；
4、对图像的所有位置重复此操作。

0	50	0	29
0	80	31	2
33	90	0	75
0	9	0	95

-1	0	1
-2	0	2
-1	0	1

首先，让将滤镜覆盖在图片的左上角：

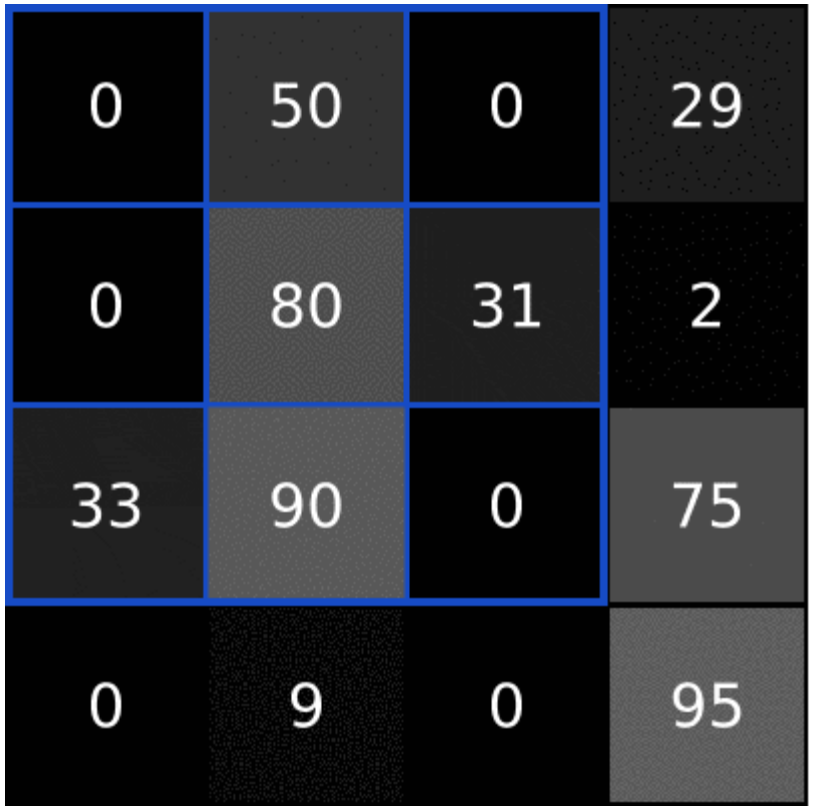
0	50	0	29
0	80	31	2
33	90	0	75
0	9	0	95

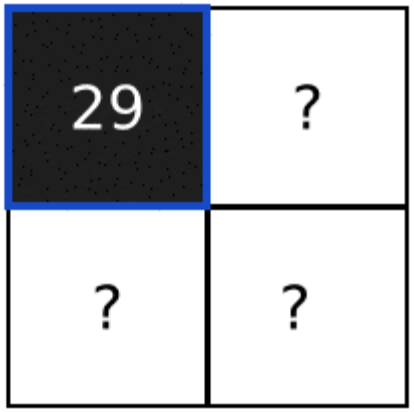
-1	0	1
-2	0	2
-1	0	1

在重叠的图像和滤波器元素之间逐个进行乘法运算，按照从左向右、从上到下的顺序。

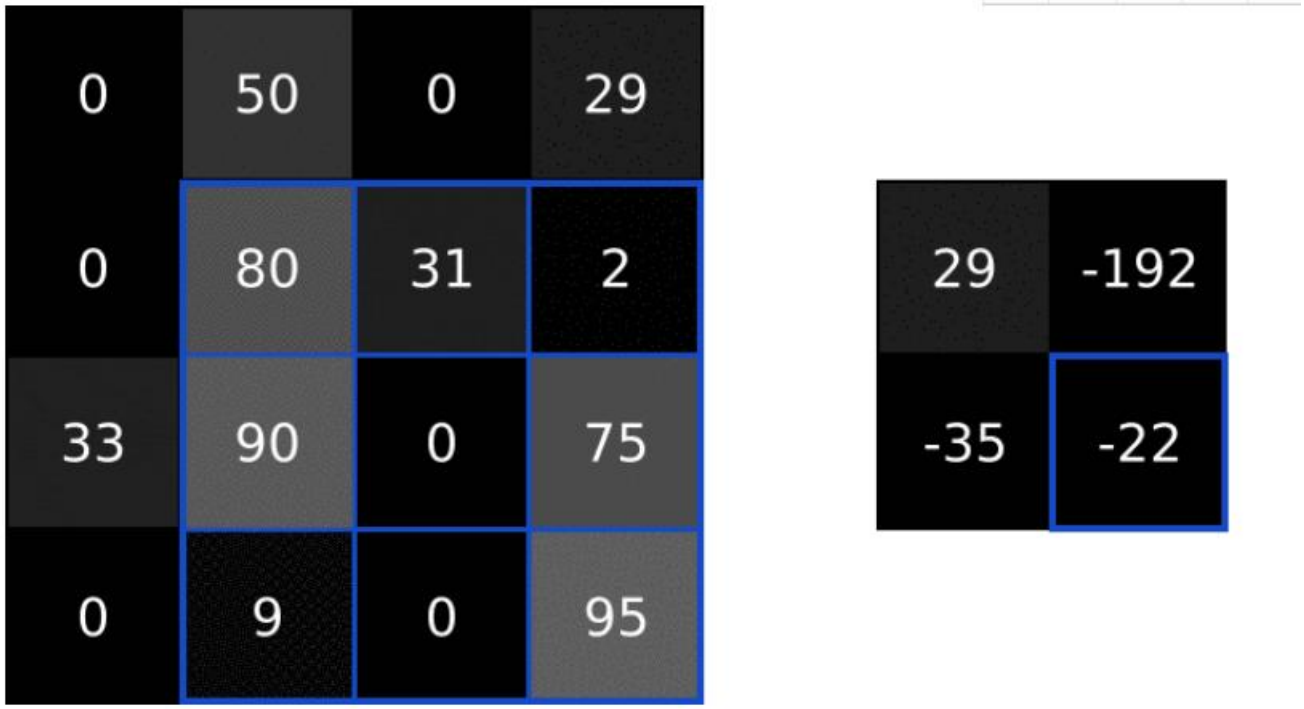
0	50	0	29
0	80	31	2
33	90	0	75
0	9	0	95

29	?
?	?

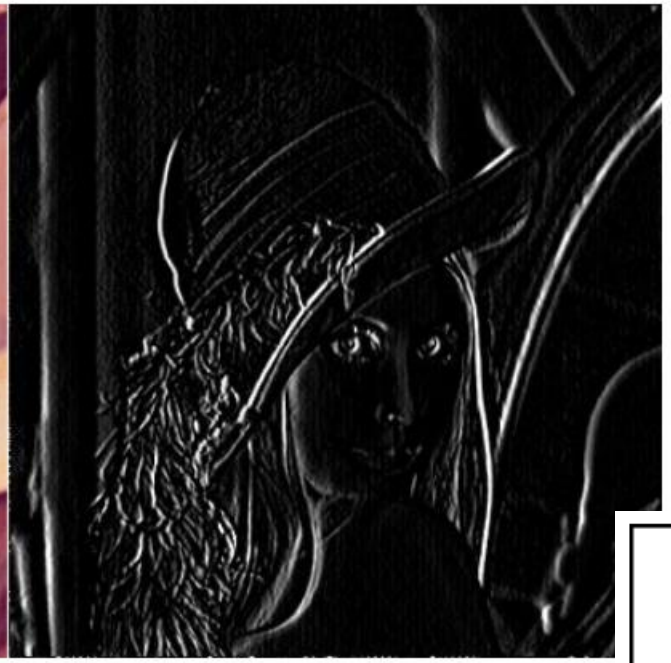

0
50
0
29
0
80
31
2
33
90
0
75
0
9
0
95


29
?
?
?

用3x3滤波器对4x4输入图像执行卷积，输出了一个2x2图像。



求卷积有何用?



-1	0	1
-2	0	2
-1	0	1

知名的Lena照片

索伯滤波器是是边缘检测器。

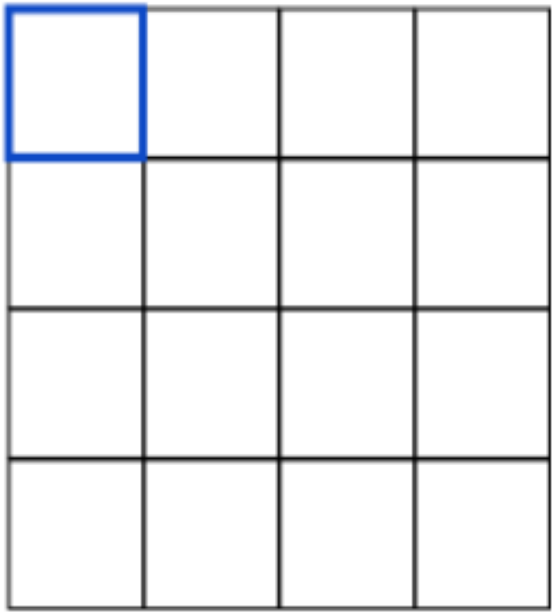
MNIST手写数字分类问题。在MNIST上训练的CNN可以找到某个特定的数字。比如发现数字1，可以通过使用边缘检测发现图像上两个突出的垂直边缘。

通常，卷积有助于我们找到特定的局部图像特征（如边缘），用在后面的网络中。

填充

希望输出图像与输入图像的大小相同。因此需要在图像周围添加零，让我们可以在更多位置叠加过滤器。3x3滤波器需要在边缘多填充1个像素。

0	0	0	0	0	0
0	0	50	0	29	0
0	0	80	31	2	0
0	33	90	0	75	0
0	0	9	0	95	0
0	0	0	0	0	0



池化

图像中的相邻像素倾向于具有相似的值，因此通常卷积层相邻的输出像素也具有相似的值。这意味着，卷积层输出中包含的大部分信息都是冗余的。池化一般通过简单的最大值、最小值或平均值操作完成。

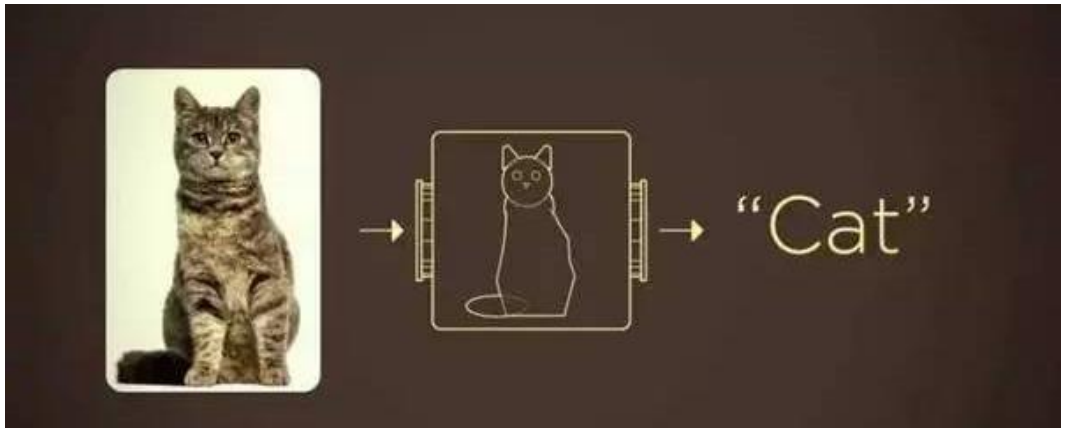
0	50	0	29
0	80	31	2
33	90	0	75
0	9	0	95

80	?
?	?

通过不同的卷积核，可以实现边缘检测（Edge Detection）、图像锐化（Sharpen）、快速均值模糊(Box Blur)、高斯模糊（Gaussian Blur）：

Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

CNN（Convolution Neural Network，卷积神经网络）



李飞飞，现为美国斯坦福大学教授、斯坦福大学人工智能实验室与视觉实验室负责人、谷歌云人工智能和机器学习首席科学家，斯坦福以人为本人工智能研究院共同院长。

Xindong Wu

Zhihua Zhou

Jiawei Han

Jian Pei

Qiang Yang

Chih-Jen Lin

Philip S. Yu

Changshui Zhang

- 1、数据挖掘技术与特定商业逻辑的平滑集成问题。
- 2、数据挖掘技术与特定数据存储类型的适应问题。
- 3、大型数据的选择与规格化问题
- 4、数据挖掘系统的架构与交互式挖掘技术
- 5、数据挖掘语言与系统的可视化问题
- 6、数据挖掘理论与算法研究
  - 模式发现（Pattern Discovery）架构
  - 规则发现（Rule Discovery）架构
  - 基于概率和统计理论
  - 基于数据压缩（Data Compression）理论
  - 基于归纳数据库（Inductive Database）理论
  - 机器学习、神经网络……

- 数据挖掘技术的产生与概念
- 数据挖掘的发展趋势
- 数据挖掘技术的分类问题
- 数据挖掘常用的知识表示模式与方法
- 不同数据存储形式下的数据挖掘问题
- 粗糙集方法及其在数据挖掘中的应用
- 数据挖掘的应用分析

根据挖掘任务

- 分类或预测模型发现
- 数据总结与聚类发现
- 关联规则发现
- 序列模式发现
- 相似模式发现
- 混沌模式发现
- 依赖关系或依赖模型发现
- 异常和趋势发现等

根据挖掘对象

- 关系数据库挖掘
- 面向对象数据库挖掘
- 空间数据库挖掘
- 时态数据库挖掘
- 文本数据源挖掘
- 多媒体数据库挖掘
- 异质数据库挖掘
- 遗产数据库挖掘
- web数据挖掘等

根据挖掘方法

- 机器学习方法
- 统计方法
- 聚类分析方法
- 神经网络（Neural Network）方法
- 遗传算法（Genetic Algorithm）方法
- 数据库方法
- 近似推理和不确定性推理方法
- 基于证据理论和元模式的方法
- 现代数学分析方法
- 粗糙集（Rough Set）或模糊集方法
- 集成方法等

根据知识类型

- 挖掘广义型知识
- 挖掘差异型知识
- 挖掘关联型知识
- 挖掘预测型知识
- 挖掘偏离型（异常）知识
- 挖掘不确定性知识等

- 数据挖掘技术的产生与发展
- 数据挖掘研究的发展趋势
- 数据挖掘技术的分类问题
- 数据挖掘常用的知识表示模式与方法
- 不同数据存储形式下的数据挖掘问题
- 粗糙集方法及其在数据挖掘中的应用
- 数据挖掘的应用分析

- 数据挖掘的目的是发现知识，知识要通过一定的模式给出。通过对数据挖掘中知识表示模式及其所采用方法的分析，可以更清楚地了解数据挖掘系统的特点。
- 主要知识模式类型有：
  - 广义知识（Generalization）
  - 关联知识（Association）
  - 类知识（Class/Cluster）
  - 预测型知识（Prediction）
  - 特异型知识（Exception）

- 广义知识是指描述类别特征的概括性知识。这类数据挖掘系统是对细节数据所蕴涵的概念特征信息的概括和抽象的过程。
- 主要方法有：
  - 概念描述（Concept Description）方法：
  - 多维数据分析可以看作是一种广义知识挖掘的特例
  - **多层次概念描述问题：**由数据归纳出的概念是有层次的，不同层次的概念是对原始数据的不同粒度上的概念抽象。
  - 例如，DEPT的模式分层结构可能是：
    DEPT→COMPANY→CITY→COUNTRY。
  - 例如，年龄EAGE可以抽象成{[20，29]，[30，39]，[40，49]，[50，59]}或者{青年，中年，老年} 。

- **关联知识挖掘的目的就是找出数据库中隐藏的关联信息。**
  - 关联知识反映一个事件和其他事件之间的依赖或关联。
  - 关联可分为简单关联、时序（Time Series）关联、因果关联、数量关联等。

- **关联规则挖掘（Association Rule Mining）是关联知识发现中最常用方法：**
  - 关联规则的研究最早的分支之一，最著名的Apriori算法。
  - 是数据挖掘研究中比较深入的分支，许多关联规则挖掘的理论和算法已经被提出。

- 类知识（Class）刻画了一类事物，这类事物具有某种意义上的共同特征，并明显和不同类事物相区别。

- 有两个基本的方法来挖掘类知识：
  - 分类：分类是数据挖掘中的一个重要的目标和任务，是目前的研究和应用最多的分支之一。
  - 聚类：数据挖掘的目标之一是进行聚类分析。

谢谢！