密级_____

中国科学院大学
University of Chinese Academy of Sciences

# 博士学位论文

## 深度学习算法中的因素解析研究


作者姓名: 陈孝罡

指导教师: 焦建彬　教授　中国科学院大学

学位类别: 工学博士

学科专业: 计算机应用技术

研究所　: 电子电气与通信工程学院


二零一五年五月

# Reseach on Factor Disentangling of Deep Learning Algorithm

**By**

**Xiaogang Chen**

**A Dissertation Submitted to**

**University of Chinese Academy of Sciences**

**In partial fulfillment of the requirement**

**For the degree of**

**Doctor of Computer Application Technology**

**School of Electronic, Electrical and Communication Engineering**

**University of Chinese Academy of Sciences**

**May, 2015**

## 中国科学院大学直属院系
## 研究生学位论文原创性声明

本人郑重声明:所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

## 中国科学院大学直属院系
## 学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定,即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密的学位论文在解密后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

# 摘　要

理解和发现数据内在的生成规律一直是科学家研究的重点。近年来，随着深度学习以及表示学习理论的发展，数据中的生成因素解析逐渐成为机器学习研究的热点方向之一。理解数据中的生成因素不仅能够对数据的来源获得更加清晰的认识，同时也能对后续的机器学习任务提取有用的特征表示。本文对深度学习中的因素解析作用展开了研究，取得了如下的成果：

1. 从流形学习的角度对因素解析作用进行了阐述，并基于深度学习算法中的自动编码器和流形学习中的结构保持概念，提出了一种局部结构保持映射算法，试图通过非监督的方式发现数据中的解释性因素。在 Mnists 手写体字符识别问题上的实验结果表明，所提方法不仅提升了分类准确度，学习到的特征表示能够对数据内在的因素作用关系进行一定刻画，并且能够较好地处理数据噪声。

2. 利用深度卷积神经网络特征提取方法对视觉行人检测问题开展了研究，提出了一种新的两阶段式的行人检测算法框架，在粗检测阶段进行候选窗口提取，精检测阶段进行特征提取与判别，在 INRIA、Caltech、ETH 数据集上进行了充分的实验，对算法的性能进行了详细的分析比较。

3. 对深度卷积神经网络提取特征过程中的因素解析作用进行了分析，提出了特征空间规整的概念，以此为基础对深度卷积神经网络特征提取过程提出了一种改进的方法，旨在通过非监督学习方法对特征提取过程引入因素解析的作用。实验表明，该方法在目标检测问题中带来了一定的改进，同时为进一步的理论工作指出了探索的方向。

**关键词：**深度学习、因素解析、流形学习、卷积神经网络、行人检测

# Abstract

To undersrand and discovery the underlining generating process of data is focus of the scientific research. In recent years, with the development of deep learning especially the theory of representation learning, disentangling the generative factors of data has been one of the hotest topics of machine learning. Understanding the generative factor of data not only brings clear knowledge of data source, but also provides useful feature representation to the following machine learning mission. This work conducts research on the factor disentangling effect of deep learning methods, and makes the following contributions:

1. This work models factor disentangling from manifold learning perspective, and proposes an algorithm called Local Structure Perserving Projection, trying to discovery the underlining explainary factor of data in an unsupervised fashion. The algorithm is based on one of the deep learning methods, autoencoder, and the structure preserving property from manifold learning. Experiment results in Mnist hang writing digit recognition task show that the proposed method not only improves the classification accuracy, but also handle data noises well, besides learned features can model the generative factor effect among data.

2. Based on deep convolutional neural network, this work conducts research on the pedestrian detection problem, and contributes a new two-stage pedestrian detection framework. The coarse stage is to propose candidate windows, and the fine stage for feature extraction and classification. Comprehensive experiments are conducted on INRIA, Caltech, and ETH datasets, followed by detailed algorithm analysis.

3. This work analyse the factor disentangling effect in the deep convolutional neural netowk, and proposes the concept of feature space regularization. Based on that, this work proposes to improve DCNN feature extraction, which tends to adapt factor disentangling effect in an unsupervised way. Exepriment results show that this method brings improvement to object detection problem, and will inspire the further theoretic research.

**Key Words:** Deep Learning, Factor Disentangling, Manifold Learning, Convolutional Neural Network, Pedestrian Detection

# 目录

# 图目录

# 表目录

# 第一章　绪论

## 1.1　研究背景与意义

在人工智能与机器学习领域，研究的开展建立在对所收集数据的处理与分析之上。不同的研究问题，其数据所产生的机理也存在许多的不同。

对于绝大多数研究问题，人们所能掌握的只有观测变量，这些观测变量的产生过程往往涉及到某些无法直接观测的变量，例如计算机视觉中的人脸识别问题，所处理的数据是图像传感器所采集到的图像，而采集图像时的光照条件、人脸的角度、表情等是难以测量的，但同时这些条件又影响着图像的生成；再例如化工系统的故障检测问题，收集到的数据是化工系统中安置的传感器所得到的检测数值，如流量、压强、温度等，这些传感数据的生成，依赖于整体系统中反应物的反应速率等等无法确切测量的变量。这些无法进行直接观测的量，一般称之为隐变量。隐变量不仅作用于观测变量，而且隐变量自身也可能存在复杂的依托关系，形成隐结构关系。如何从观测数据中发掘出有用的信息，理解其内在的决定因素、结构关系和生成规律等，从而帮助决策、推理等学习目标的实现，是当前机器学习领域的热点研究问题之一。

随着互联网数据挖掘、图像分析、语音识别等智能科学领域研究的逐步发展，处理的问题愈发复杂化，研究中面对的绝大多数是高维、结构化的数据。近年来，随着多层神经网络的学习算法取得突破性进展之后[1]，深层模型以其优异的性能迅速获得了研究人员的极大关注。通常的机器学习算法如支持向量机（Support Vector Machine，SVM）、逻辑回归（Logistic Regression，LR）等，都可视为浅层模型（Shallow Model），而深层模型（Deep Model）则以多层神经网络为代表[2]，又常称为深度学习模型。理论上而言，深层模型具有比浅层模型更加丰富的表现力，随着模型层数的增加，整个模型的表达和抽象能力也得到提升，因此可以更加有效的表达数据内在层次结构和因素之间的关联关系[3]。

本研究拟基于深度学习模型建立数据内在的结构层次关系，依托于深度学习模型强大的表达能力，发掘数据中内含的关键因素，从而更好地刻画数据的分布特性和理解数据背后的生成机理。

**本文课题来源：**

（1）“危险化学品事故全过程遥测预警的关键科学问题研究”，国家 973 计划，2011~2015；

（2）“基于多源数据的飞行棋进近威胁目标检测跟踪及行为预测”，国家自然科学基金重点项目，2011~2014；

（3）“复杂环境下动态目标检测及跟踪技术研究”，中国科学院“百人计划”择优支持项目，2009~2011。

## 1.2 国内外研究现状

理解和发现数据内在的生成规律一直是科学家关注的重点，最早可以起源于物理学中的方程发现（Equation Discovery），对于一组实验数据，如何得到描述变量和因变量之间的方程关系，即是该研究的目的[4]。通过近 30 年来的发展，该研究方向从复原简单的线性模型[5]，到复原一些复杂的非线性模型[6]，取得了长足的进步。然而该类方法依然存在不足，尽管文献[6]的方法对于机械结构的方程解析结果十分优美，其局限性在于其模型采用的非线性函数如三角函数等，只能在物理学研究领域等变量维度较少的情况下适用。

在模式识别和机器学习领域所要处理的数据，往往是海量的高维数据，其内在的因素关系要复杂的多。对于这些高维数据，研究人员通常希望将其映射到一个低维空间中（降维），再在低维空间中进行分析。降维操作的目的在于，一是减少数据的处理量，二是尽可能去除噪声[7]。另一方面，也希望在此过程中能够发现数据产生过程内在的规律，帮助人们更好的理解数据产生背后的原因。

一般地，从数据中学习一组低维的、生成性因素的过程可以看成是一个广义的数据降维过程，对该过程通过学习方法寻找一映射函数，将原始数据映射

到某特征空间中，该映射函数的表达能力是学习算法能否成功的关键[8]。数据降维的方法大概可分为三类：线性映射方法、非线性映射方法以及近邻方法[9]。

线性方法典型如 PCA、ICA、因子分析（Fator Analysis）等[10]，采用统计方法寻找数据空间中的一组低维基表示，这些方法存在的局限性是当数据本身位于一个非线性流形上时，由于模型表达能力的限制，线性方法无法刻画出数据本身的分布性质。

非线性方法如基于前馈网络的 Autoencoder[11]，另外还有奇异值分解（Singular Value Decomposition， SVD）[1]，稀疏编码（Sparse Coding）[13]，以及近年来提出的非线性独立成分估计（Non-linear Independent Components Estimation，NICE）[14]等方法，通过引入非线性，模型的表示能力比线性模型更加丰富，从而能够获得比线性模型更好的刻画非线性的数据。

近邻方法如局部坐标编码（Local Coordinate Coding, LCC）[15]，以及众多的流形学习方法[16-21]等，通过数据样本之间的近邻关系，约束在映射空间中数据表示之间的局部线性结构，尽管通过局部结构能够很好的探索数据之间的内在关系，但其存在的问题是训练规模随着数据量的增大呈平方次增长，很难应用到较大规模的数据集上。

从以上方法的发展历史来看，模型的表达能力是解析数据中生成因素的重要条件。近 10 年来，随着新的学习范式即逐层学习方法的提出[22]，曾经受困于训练困难的多层神经网络体现出了强大的数据表示能力，引起了研究人员对多层次模型的极大关注。理论上而言，多层次模型具有十分优异的表达和抽象能力，文献[23]指出，一个两层的逻辑门电路可以表示任意的方程形式。文献[2]提出，对于复杂函数来说，若采用正确的模型结构，深层模型可以实现函数的紧致表达，而要达到同样的表达能力采用分段线性拟合方法需要的参数则多得多。伴随着研究的深入，基于多层神经网络模型的深度学习算法在多个研究领域，如自然语言处理[24,25]、图像识别[26,27]、语音识别[28,29]等领域取得了突破性的进展。

对于深度学习所取得的成功，文献[30]从表示学习（Representation Learning）的角度展开了详细的总结分析和阐述。其指出，一个好的特征表示应具有的能够将数据中的变化因素进行解析的特性。对于某个机器学习任务 P(Y|X)，通过特征学习更好的描述数据的分布特性 P(X)，会对学习任务本身带来极大帮助，因为特征学习有可能发掘出数据背后解释性因素。

在机器学习的发展历史中，比起特征的解析性，特征的不变性是更为广泛采用的研究思路。比如在计算机视觉的研究中，尺度不变性特征（Scale invariant feature transform，SIFT）[31]是最为典型的代表，其通过在图像的多尺度空间中寻找极值点，提取位置、旋转、尺度的不变量。近年来基于机器学习所发展起来的一些算法，也试图去实现特征不变性的提取，如一系列基于受限玻尔兹曼机（Restricted Boltzmann Machine， RBM）[32,33]的方法，文献[34]采用三阶因素化 RBM 模型对图像的生成过程进行表示与特征提取，文献[35]中作者针对图像的局部形变，比如平移、旋转、尺度缩放等微扰，在 RBM 基础上提出了一种新的训练框架，以期望最终提取的特征对这些微扰变化具有鲁棒性。另外的一类方法，如拓扑独立成分分析（Topological Independent Component Analysis）[35]、不变性稀疏分解（Invariant Predictive Sparse Decomposition ，IPSD）[36]、基于 ssRBM 的方法[37]等，通过对特征进行分组，在学习的过程中通过限定特征组的激活频率，实现了通过非监督方式获得不变性特征提取的效果，学习到的特征表示在实验中对某些变化因素不敏感。

不变性特征主要提取了数据中具有不变性的特殊量，体现的是数据中具有的共性而忽略变化性，而解释性因素则更多的关联于数据背后的生成机理，着重于数据中的变化和信息保留[38]。不变特征提取和解析特征提取两种方法在原则上是不同的，其适用的研究问题也不相同，相对而言，不变特征提取更适用于监督的判别问题研究，而解析特征提取更适用于非监督学习的研究。在大数据机器学习时代，同一组数据可能用于多个机器学习任务，因此对于特征学习而言，更鲁棒性的方式是尽可能多的解析出解释性因素，从而尽量多的保留数据中含有的信息，以期适用范围更加广泛[30]。

因素解析特征提取原则是近年来才发展起来的概念，在该方向上研究人员依然处于探索阶段，并没有形成成熟的框架体系，但依然出现许多成功的案例。文献[39,40]中，对于学习得到的多层网络进行了逐层分析，发现，某些隐层节点对于已知的变化因素具有强烈的响应，而另外的一些因素作用却无法激活这些节点，表明这些隐藏特征节点对于因素的作用进行了建模。在[41,42]中采用前面不变特征时候提到过的对特征进行分组的策略，使这些特征组针对性的对某些因素进行建模，最终能够起到因素解析的效果，但其存在的问题是，如果因素较多的时候，要设定多个分组，并且这里采用的都是监督学习的方式，当目标较多的时候就可能需要大量的样本标定。在文献[43]中，作者通过李群中的交换群对解析特征提取过程进行了群论下的阐述，认为，因素的变化作用是可交换的，那么这些作用对数据的生成作用就可以用交换群来进行抽象，从而利用群论[44]的分析方法学习到数据的不可约表示。文献[45]中，对因素解析问题进行了更深入的研究，其利用自动编码变分贝叶斯方法[46]构建了一个编码-解码网络，而网络最终得到特征表示对三维目标图像生成中的因素，如姿态、光照、形状等分别进行了建模，从研究思路上对该问题进行了拓展。

得益于深度学习的发展，特征表示学习取得了十分显著的研究成果，在解析数据内在因素作用方面也获得了一定的进展。尽管如此，受限于深度学习理论研究进展的缓慢，相关研究工作只是在应用层面获得了更多的关注。随着近两年来一些关于深度学习新理论[47,48]的提出，研究人员有望对因素解析问题进行更深入的研究。

综上所述，对于因素解析特征提取，其问题主要难点在于：

1. 数据内在生成作用因素的数量难以确定，也即当假设数据分布在一个流形空间上时，其本征维度无法确切知道。由此带来模型选择、参数设计、训练等困难。

2. 无法量化各个因素对数据生成的作用，因此在训练时无法用采用回归模型对参数进行优化，对于最终获得的特征难以反推出因素作用的数值化。

3. 因素相互之间是否存在独立性，是否具有联合作用关系无法确定，因此只能通过模型假设的方法，对问题进行一定约简。

## 1.3 本文主要研究内容

本文主要通过流形中的概念对因素解析特征问题进行了形式化描述，在此基础上采用深度学习算法中自动编码器的相关性质，对问题进行了具体的构建和求解，并对近年来发展起来的深度卷积神经网络在视觉目标检测中的应用开展了研究，试图对其网络结构中的模块的作用进行诠释，并将因素解析特征提取的概念融入到深度卷积神经网络学习中去。具体研究内容如下：

1. 对因素解析特征提取的问题，通过流形学习方法进行了诠释，将其归结成寻找原始空间和流形嵌入空间之间一个参数化映射的过程；将数据样本之间的空间结构关系表达为解析关系，认为相近的样本之间其生成因素作用是相近的。在此基础上，利用深度学习算法中自动编码器的相关流形性质，构成出最终的算法模型，并进行实验分析比较。

2. 为进一步研究深度卷积神经网络中的相关特征，对深度卷积神经网络在视觉行人目标检测中的应用进行了研究，提出了一种新的基于 DCNN 的行人检测算法框架，对算法流程环节进行了详细分析与说明，对算法在三个数据集上的不同性能表现进行比较与总结。

3. 以深度卷积神经网络在航拍目标检测中的应用结果为基础，分析了因素解析特征提取在深度卷积神经网络中的实现；提出特征空间规整的概念对几种主流网络训练目标进行了归纳；并进一步对 DCNN 各个网络层的作用与特征性质进行了分析，将内容 1 中因素解析特征提取的方法应用到 DCNN 的训练中去。

## 1.4 本文组织结构

第一章 绪论。本章简要介绍了本论文的问题背景、当前相关领域的研究现状、主要研究内容。对深度学习以及因素解析原则下的特征提取进行了初步的阐述和说明。

第二章 流形学习与自动编码器。本章主要介绍了本文研究的主要理论基础知识，包括了流形学习与深度学习两大部分。在流形学习部分介绍并归纳了流形学习方法的主要共性，总结出了结构保持映射的主要框架。深度学习部分主要介绍了深度学习的基本算法，自动编码器的各种变形及其特性。

第三章 局部结构保持映射。本章引入自动编码器的流形性质分析，并以此为分析基础，提出了一种局部结构保持的映射算法模型。通过紧致自动编码器的雅克比矩阵构建数据局部空间结构，利用原始与映射空间之间的概率测度建立全局的结构关联并进行优化。通过一系列实验对算法进行了分析和比较。

第四章 深度卷积神经网络与视觉行人检测。简要介绍了深度卷积神经网络模型的研究进展，对视觉行人检测算法研究的现状进行了分析，提出了一种基于深度卷积神经网络的视觉行人检测系统。在三个主流的行人检测算法评测标准数据集中进行评测与比较。

第五章 深度卷积神经网络中的因素解析特性研究。基于深度卷积神经网络特征提取在视觉目标检测中的应用，对因素解析原则下的特征提取过程进行了详细分析说明。提出特征空间规整的概念，归纳总结了多种卷积神经网络训练目标的共性。对深度卷积网络中卷积层和全连接层的作用进行了探讨，提出了一种新的模型训练方法。

第六章 结论与展望。

# 第二章　流形学习与深度学习基础

在本文的研究工作中，主要采用了流形学习中的相关概念对问题进行抽象与形式化，方法上主要采用了深度学习方法，并利用其流形性质进行展开，本章首先对相应的概念及性质做介绍。

## 2.1　流形学习

流形是微分几何及拓扑学中的一个基本概念，是对欧几里得空间即线性空间中曲线和曲面的推广，其数学定义以来于同胚概念，相关定义如下：

**定义 2.1**（同胚）[49]：设一个多元函数 $f:U\subset R^p\to V\subset R^q$，若 $f$ 为双射，且 $f$ 和 $f^{-1}$ 都是连续的，则称函数 $f$ 为同胚，$U$ 和 $V$ 是同胚的。

**定义 2.2**（流形）[50]：设 $M$ 是 Hausdoff 空间。若对任意一点 $x\in M$，都有 $x$ 在 $M$ 中的一个领域 $U$ 同胚于 $m$ 维欧式空间 $R^m$ 的一个开集，则称 $M$ 是一个 $m$ 维流形（或拓扑流形）。

在机器学习的研究中，引入流形概念的切入点在于假设所观测到的高维数据是由一组低维的变量生成的，且低维变量处在一个流形空间上，也即低维流形是嵌入（Embedded）在高维空间中的。在流形假设前提下，流形学习的目的就是将高维观测数据映射到低维嵌入空间中，并且在低维空间内保持数据在原始空间中的一些几何结构性质；这一过程也可视为从观测数据中反推其生成模式、解析其生成因素的过程。其形式化可表达为，给定一组高维观测数据 $X=\{x_1,x_2,...,\mathrm{x}_N\},x\in R^D$，$N$ 为样本数量，$D$ 为样本维度。流形学习假设这些数据生成于一个本征维度（Intrinsic Dimension）为 $d$ $(d<D)$ 的流形，学习的目标是找到这组观测数据在低维嵌入空间 $R^d$ 的表示 $\{y_i\}_N$，以及从观测空间到低维空间的映射 $f:X\subset R^D\to Y\subset R^d$。在这一映射过程中，人为的设定原始空间中局部或者全局的几何结构关系，使在嵌入空间中的数据间依然保持该关系不变。对几何结构关系设定的不同，衍生出了许多不同的流形学习算法，本节接下来将通过总结不同几何结构关系的角度简要回顾几种得到广泛应用的流形学习方

法。

（1）MDS（Multidimensional Scaling）算法[18]

MDS 是一种经典的数据降维及可视化方法，其算法起源是当仅能获得原始数据之间的相似度矩阵时，如何在欧式空间中重构它们的坐标，并且在重构的欧式空间中，任意数据点对之间依然保持原始空间中的距离关系。MDS 算法可表述为：给定数据之间的距离矩阵 $\Delta_{N\times N}$，$\Delta$ 中元素 $\delta_{ij}$ 代表数据样本 $i$ 和 $j$ 之间的欧氏距离，找到一组坐标 $\{y_1, y_2, ..., y_N\} \in R^d$，使对所有 $i, j \in N$，有 $\left\| y_i - y_j \right\| \approx \delta_{ij}$。由此可建立公式（2-1）所示的算法的优化目标函数。对该目标函数的求解可以通过对原始数据的 Gram 矩阵 $K = (x - \bar{x})^T (x - \bar{x})$ 进行谱分解来进行。在只给定数据的距离矩阵 $\Delta$ 没有数据的原始坐标情况下，Gram 矩阵中的元素可以通过公式（2-2）所示对 $\Delta$ 进行双中心化操作得到。

$$\min L(\mathrm{Y}) = \sum_{i,j} \left( \left\| y_i - y_j \right\|^2 - \delta_{ij}^2 \right) \tag{2-1}$$

$$k_{ij} = \frac{1}{2} \left( \delta_{ij}^2 - \frac{1}{N} \sum_l \delta_{il}^2 - \frac{1}{N} \sum_l \delta_{jl}^2 + \frac{1}{N^2} \sum_{lm} \delta_{lm}^2 \right) \tag{2-2}$$

可以看到，MDS 算法将欧式距离作为结构保持的一个度量，由于 MDS 算法得到的低维坐标表示的中心在原点，又可以人为它是保持内积的，也即利用低维空间中的内积来近似高维空间中数据的内积。

（2）ISOMAP（Isometric feature mapping）算法[17]

ISOMAP 算法发表于 2000 年，并开创了流形学习的概念，是非线性降维算法的代表。ISOMAP 以 MDS 为基础，将 MDS 方法中刻画数据间关系的欧式距离改为了流形中的测地距离（Geodesic distance），是对 MDS 方法的非线性推广。由 MDS 算法的介绍可以看到，在两个空间中表达数据样本之间的相对关系时，采用的都是欧式距离，是一种线性的表达方式。而在流形学习中，通常假设高维数据是位于一个低维非线性流形上的，因此线性关系不能很好的表达非线性的数据关系，有必要引入非线性的表达方法，这也是 ISOMAP 算法的出发点。

ISOMAP 算法的流形主要为三个步骤：

1. 构建样本邻接图$G$：对于给定的数据集，选取每个样本点欧式距离最近的$K$个点或者给定半径距离$c$内的所有点，构成该样本点的近邻。由样本间的近邻关机构建连接图$G$，图上的边连接着各个数据及其近邻，边的权值设为相互间的欧式距离。
2. 计算测地距离矩阵$D^G$：ISOMAP 假设数据位于低维非线性流形空间上，因此采用流形中的测地距离来表示样本之间的关系，但由于数据样本是离散分布的，因此无法准确得到流形的曲率也就无法准确计算测地距离。作者据此采用邻接图上两点间的最短路径距离来近似表示测地距离，通过采用 Dijkstra 或者类似算法来进行计算。
3. 构造低维嵌入表示：由距离关系矩阵$D^G$带入经典 MDS 算法，求解得到数据样本在低维嵌入空间的表示。

由算法流程可以看出，ISOMAP 与 MDS 区别主要在于步骤 1 和 2 引入近似测地距离来对数据样本的距离进行刻画。ISOMAP 是一种全局拓扑结构保持的算法，但也存在不少缺点，如容易受到参数设置（近邻数量$K$、近邻半径$c$）的影响、对数据分布中的空洞区域比较敏感，以及计算复杂度高等。

（3）LLE（Locally Linear Embedding）算法[16]

与 ISOMAP 的全局结构保持性质不同，LLE 出发点在于保持数据样本的局部几何性质。算法的基本思想为对于非线性流形，在局部区域上可将其视为近似的线性空间，对于每个数据样本，局部的几何结构就可以利用其近邻点在最小二乘下的最优线性重构进行表示，由此可赋予近邻样本相应的权重。具体地，对于数据样本$x_i$，通过像 ISOMAP 算法中一样选择最近的$K$个近邻或者设定领域半径$c$方式，获得样本近邻$x_{ij}$，通过公式（2-3）进行最小二乘拟合，赋予各个近邻样本相应的权重$w_{ij}$，进而将这个权重作为结构保持的依据，在嵌入空间中，使样本的低维表示之间依然保有相应的权重关系，见公式（2-4）。

$$\min \varepsilon(W) = \sum_i \left\| x_i - \sum\nolimits_j w_{ij} x_{ij} \right\|^2 \tag{2-3}$$

$$\min \phi(Y) = \sum_i \left\| y_i - \sum_j w_{ij} y_{ij} \right\|^2 \tag{2-4}$$

该目标函数的求解可以通过对矩阵 $M = (I-W)^T (I-W)$ 进行特征分解来进行，其中 $I$ 是单位矩阵，$W$ 是权重 $w_{ij}$ 构成的矩阵。相比于 ISOMAP，LLE 计算量要小得多，因此能处理更大规模的数据集，同时能够最优的保持数据的局部结构关系，但另一方面，和 ISOMAP 一样依然无法很好的处理数据空洞问题。

（4）LE（Laplacian Eigenmap）算法[20]

LE 算法在思想上与 LLE 算法相同的是从数据样本的局部结构保持出发去构建整个算法，但在具体实现上采取了完全不同的策略。LE 算法的核心思想可直观的表述为，对于原始空间中相近的样本点，在嵌入空间中其表示也应该尽可能接近，公式（2-5）为期优化的目标函数，如 LLE，$w_{ij}$ 代表了近邻样本的权重。在 LE 算法中，近邻样本的权重通过高斯核函数进行定义 $w_{ij} = \exp(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2})$。对该目标函数的求解可转化为特征分解问题进行。定义对角矩阵 $D$，其对角上的元素 $d_{ii} = \sum_{ij} w_{ij}$。进一步定义矩阵 $L = D - W$，矩阵 $L$ 称为图拉普拉斯矩阵，则原目标函数可用公式（2-6）表示。

$$\min \sum_{ij} w_{ij} \left\| y_i - y_j \right\|^2 \tag{2-5}$$

$$\min \sum_{ij} w_{ij} \left\| y_i - y_j \right\|^2 = 2YLY^2 \tag{2-6}$$

通过对拉普拉斯矩阵进行特征分解，见公式（2-7），其中设 $\lambda_1, \lambda_2, ..., \lambda_d$ 为前 $d$ 个非零最小特征值，则分解后对应的特征向量集合就是所要求的数据样本在低维空间中的坐标表示。由高斯核权重函数可以看到，当样本在原始空间中距离越小时，对目标函数的贡献度越大，其本质上也是利用了原始空间中数据样本间的线性关系来表示局部结构信息。

$$Ly = \lambda Dy \tag{2-7}$$

（5）Hessian LLE 算法[19]

在 LLE 算法中，对样本的局部空间结构采用线性最小二乘进行了近似表示，注意到在流形的定义中，流形上每一点是局部同胚于欧式空间的，由此可以在流形的局部切空间中建立坐标表示对其几何性质进行描述。在 Hessian LLE 算法中，利用流形的这一性质，对数据样本的局部结构刻画进行了更符合流形性质的改进。具体地，类似一般流形学习算法，首先对各个数据样本找到其近邻，接着采用奇异值分解（Singular Value Decomposition, SVD）对局部近邻建立公式（2-8）的坐标表示，$\tilde{V}^{(i)}$ 代表了样本 $x_i$ 处的切空间，而 $\tilde{U}^{(i)}$ 的前 $d$ 列就是近邻样本在切空间中的坐标表示。由所得到的局部切空间表示，利用 Gram-Schmidt 正交化得到局部映射的 Hessian 矩阵 $H^{(i)}$，最终可以由公式（2-9）定义数据样本的全局关系矩阵 $K$。

$$\tilde{x}^{(i)} = \left[ x_{i_1} - x_i, ...., x_{i_k} - x_i \right] = \tilde{U}^{(i)} \tilde{\Sigma} (\tilde{V}^{(i)})^T \tag{2-8}$$

$$K = \sum_{i=1}^{N} S^{(i)} H^{(i)T} H^{(i)} S^{(i)T} \quad \in R^{N \times N} \tag{2-9}$$

如前面介绍的算法，对该矩阵进行特征值分解操作即可得到最终的结果。Hessian LLE 算法尽管计算复杂度比较高，但所得到嵌入空间表示更加具有鲁棒性，另一方面，其采用的切空间坐标表示的做法，很好的表达了流形本身的性质，在理论上具有十分重要的意义。

（6）LTSA（Local Tangent Space Alignment）算法[21]

LSTA 与 Hessian LLE 一样，利用样本在流形切空间上的坐标表示来描述局部的几何属性，但其算法并不采用局部结构保持作为最终目标，而是采用局部切空间对齐（Alignment）的方式来实现全局的结构保持。

算法的构建过程为：假定 $F$ 是一个嵌入到 $m$ 维空间中的 $d$ 维流形（$d < m$），且嵌入函数 $f(\tau_i)$ 是未知的，$\tau_i$ 是数据在嵌入空间中的坐标。公式（2-10）表示，给定的数据样本集合 $X = [x_1, x_2, ..., x_N], x \in R^m$，是由低维流形上采样得到的

$$x_i = f(\tau_i) \tag{2-10}$$

对嵌入函数进行一阶泰勒展开，见公式（2-11），其中雅克比矩阵 $J_f(\tau)$ 就

定义了$\tau$处的一个切空间，一般地，若在该切空间中有一组正交基$Q_\tau$，则泰勒展开中的一阶项可以用（2-12）来表示，$\theta_\tau^*$为局部座标系下的坐标表示，由公式（2-13）和公式（2-14）可以看到，在低维嵌入空间的坐标表示与局部的坐标表示之间只相差一个仿射变换$L_\tau$，则算法的优化目标就建立为最小化样本嵌入表示和局部坐标重构误差，即公式（2-15）。

$$f(\tilde{\tau}) = f(\tau) + J_f(\tau)\cdot(\tilde{\tau}-\tau) + O(\|\tilde{\tau}-\tau\|^2) \tag{2-11}$$

$$J_f(\tau)\cdot(\tilde{\tau}-\tau) = Q_\tau\theta_\tau^* \tag{2-12}$$

$$\tilde{\tau}-\tau = J_f^+(\tau)Q_\tau\theta_\tau^* \equiv L_\tau\theta_\tau^* \tag{2-13}$$

$$\tilde{\tau} = \tau + L_\tau\theta_\tau^* \tag{2-14}$$

$$\min\int d\tau\int_{\Omega(\tau)}\left\|\tilde{\tau}-\tau-L_\tau\theta_\tau^*\right\|d\tilde{\tau} \tag{2-15}$$

LTSA 算法通过对局部结构的全局对齐，实现几何结构关系的保持，因为全局对齐的引入，使得算法的稳定性得到了很大的提高；另一方面，相对其他流形学习算法，LTSA 算法的计算效率很高，因为也得到了广泛的应用。

注意到以上介绍的几种流形学习算法中，最终都没有给出显式的映射函数，也就是说只能针对给定数据样本集合，获得其在低维嵌入空间的坐标表示，当有新的数据样本到来的时候，无法通过显式的映射函数直接获得低维坐标表示。针对这种情况，研究人员一般采用近邻插值或者拟合的方式进行扩展从而获得相应的低维坐标。除了这些非参数化方法，基于参数化映射函数的流形学习方法也出现了许多代表性的工作，以下将对这方面的工作进行简要介绍。

（7）Non-Local Manifold Tangent Learning 算法[51]

NLMTL 算法尽管是一种参数化的流形学习方法，但其并不是给出原始空间与嵌入空间之间的映射函数，而且显式的给出样本在流形切空间上的坐标基映射函数。算法的构建过程为，假定一个矩阵函数$F(x)\in R^{d\times n}$显式给出数据样本$x\in R^n$在流形切平面上的一组坐标基的表示，则在数据样本领域内的向量，在切空间上的表示就可以写为$F'(x)w$，其中$w\in R^d$是局部坐标。算法的优化目标即

是最小化两个不同空间内向量的投影误差，也就是最小化将样本 $x$ 和近邻样本 $x_i$ 的差向量 $(x-x_i)$ 投影到切空间 $F(x)$ 时，与切空间中的向量表示 $F'(x)w$ 差别最小，因此目标函数就为：

$$\min_{F,\{w_{ij}\}} \sum_t \sum_{j\in N(x_t)} \frac{\left\| F'(x_t)w_{tj} - (x_t - x_j) \right\|^2}{\left\| x_t - x_j \right\|^2}. \tag{2-16}$$

可以看到这是一个两层的优化问题，需要同时优化切空间坐标基的映射函数 $F$ 和局部坐标 $w$。对于类似的优化问题，通常采用坐标下降的方法进行优化，即首先固定一组参数，优化另外一组，再反之，进行多轮迭代。在这里，对 $w$ 的求解可最终推导出为求解线性方程组：

$$F(x_t)F'(x_t)w_{tj} = F(x_t)\frac{(x-x_i)}{\left\| x_t - x_j \right\|^2}. \tag{2-17}$$

另一方面，通过将映射函数 $F(x)$ 设为单层神经网络，对其的求解就可以通过梯度下降方法进行，在固定 $w$ 时，可得到 $F(x)$ 参数的梯度为：

$$2\sum_j \frac{w_{tji}}{\left\| x_t - x_j \right\|^2}(F'(x)w - (x_t - x_j)). \tag{2-18}$$

尽管该方法没有给出直接的坐标映射函数，但是其能够很好的刻画出低维流形的特性，为了获得显式的映射函数，可以引入高斯混合模型[52]，结合所得到的切空间映射，便可以达到目标。

（8）Parametric t-SNE 算法[53]

Parametric s-SNE 算法是由随机近邻嵌入（Stochastic Neighbor Embedding，SNE）[54]及 $t$ 分布随机近邻嵌入（t-Distribution SNE）[55]延续而来的算法，其采用了和 $t$-SNE 一样的优化目标函数，但和 NLMTL 算法一样，通过引入一个多层的神经网络，从而能够获得显式的映射函数。算法的核心思想延续了 SNE 中的思想，直观的描述为，在原始空间中越接近的样本，其在嵌入空间中也越接近。对于样本之间的几何结构关系，并不像其他算法那样通过权值或者切空间中的坐标表示来进行描述，而是将距离关系转换到概率测度空间去进行表示。对于

原始空间中的样本，通过高斯分布构建其关系，见公式（2-19）。

$$p_{j|i} = \frac{\exp\left(-\| x_i - x_j \|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\| x_i - x_k \|^2 / 2\sigma_i^2\right)} \tag{2-19}$$

考虑对称性，样本 $i$ 和 $j$ 的概率测度可设为 $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$。

在嵌入空间中，则采用公式（2-20）的 *t*-student 分布对近邻关系进行表示，其中 $f(x|w)$ 就是从原始空间到嵌入空间的映射函数，$w$ 是其参数。在构建的两个概率测度的基础上，要实现映射之后的结构保持，自然的可以通过最小化 KL（Kullback-Leibler divergence）散度的方式进行优化达到，也即公式（2-21）。

$$q_{ij} = \frac{(1+\| f(x_i|w) - f(x_j|w) \|^2 / \alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k \neq i} (1+\| f(x_k|w) - f(x_l|w) \|^2 / \alpha)^{-\frac{\alpha+1}{2}}} \tag{2-20}$$

$$\min C = KL(P \| Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2-21}$$

因为采用多层神经网络作为映射函数 $f(x|w)$，目标函数的优化也变得十分方便，通过随机梯度下降方法就能很容易的进行求解。

**小结**：通过回顾以上流形学习算法，可以看到，无论是非参数化还是参数化流形学习，其核心目标都是在原始空间和嵌入空间之间实现几何结构保持。通过不同的局部结构的定义，可以实现对数据样本几何属性以及对数据样本在嵌入空间中表示进行刻画，从而实现所需要的性质。尽管经过了十多年的发展，在机器学习和计算机视觉领域的研究中，流形的概念越来越普及和受到重视，但其依然离大规模应用化还有不少差距，主要在于一般所处理的问题数据样本比较稀疏，难以满足流形假设的基本要求，另一方面，普遍的流形学习算法没有显式的映射函数，也是制约其推广的重要原因，但这些也都是值得继续深入研究的方向。

## 2.2 深度学习

深度学习（Deep Learning）的概念兴起于 2006 年 Science 上相关论文的发表

[1]，在这篇论文中，作者用一个 8 层的神经网络构建了一个自动编码器（Autoencoder），并且在每一层的神经网络训练中，采用受限玻尔兹曼机（Restricted Boltzmann Machines， RBM）进行参数的初始化，最终在手写体字符识别和人脸识别问题上取得了十分突出的结果，从而引发了深度学习的热潮[30]。

尽管深度学习近年来取得了巨大的成功，但其所采用的方法基本是延续了 80 年代发展起来的人工神经网络模型，只是得益于近年来飞速发展的并行计算能力，使得研究人员得以训练超大规模的神经网络以及处理海量数据，从而带来性能的极大飞跃。

在本文的研究工作中，主要基于多层神经网络模型以及自动编码器进行展开，因此在本节首先对深度学习以及神经网络相关工作进行回顾，并且结合神经网络的流形性质进行一些基本的解释。一个典型的基于多层神经网络的深层模型图 1 所示。

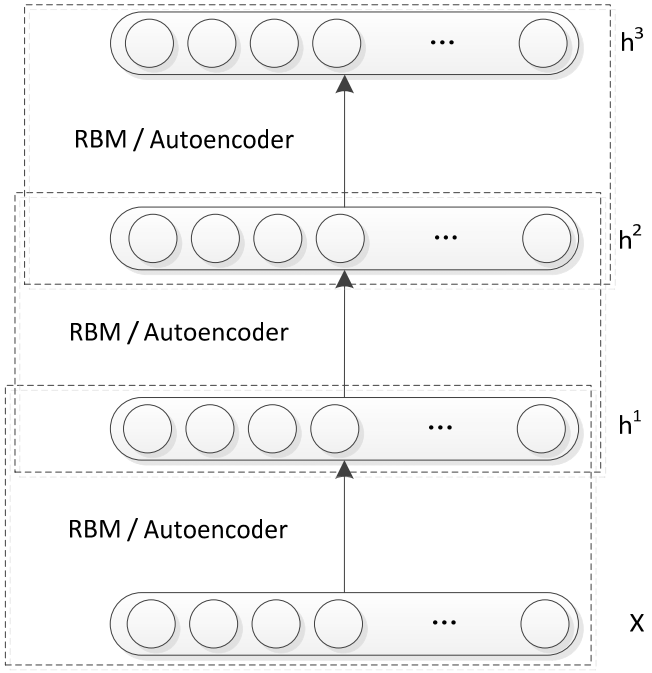



图 2- 1 深层模型示意图

可见，深层神经网络模型是多个单层基本模型叠加构成的，前一层的输出

作为后一层的输入，依次推进到最顶层。输入 $x$ 通过逐层的前向传播运算，一直到达最顶层 h3，最终 h3 层的节点值即为整个模型的输出。

对于多层的神经网络模型，论文[22]提出了首先采用逐层预训练，赋予网络一定的初值，再对网络进行整体细调优化（Fine-Tuning），会比随机给予网络链接初始值并使用反向传播（Back-Proposition）算法效果更好，对于单层模型，最基本的模型是受限玻尔兹曼机以及 Autoencoder 两种。

（1）**受限玻尔兹曼机（Restricted Boltzmann Machines，RBM）**

一个受限玻尔兹曼机包含两层节点，输入层 $v$ 也被称为可视层，以及隐藏层 $h$，通过增加隐藏层节点数量，可以提高 RBM 的表达能力。图 2-2 为 RBM 的示意图。可以看到，RBM 可视为二部图结构，一般地对一个 RBM 定义能量函数见公式（2-22）。

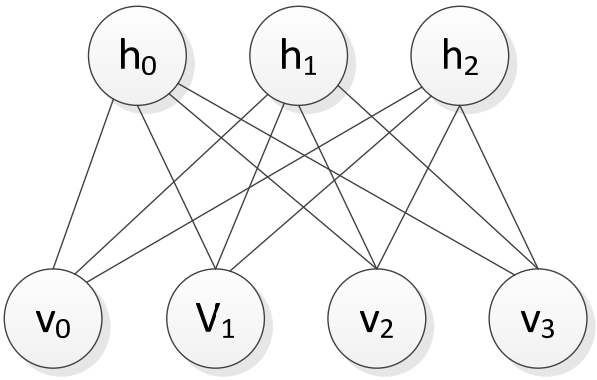



图 2- 2 RBM

$$E(v,h) = -b'v - c'h - h'Wv \tag{2-22}$$

公式（2-22）中的 $W$ 为图中连线的权重，$b$ 和 $c$ 分别为输入层和隐藏层的偏置项。由能量函数定义可以进一步得到公式（2-23）的自由能。

$$\mathcal{F}(v) = -b'v - \sum_i \log \sum_{h_i} e^{h_i(c_i + W_i v)} \tag{2-23}$$

在基于能量表示的概率模型上，可以由能量函数来定义概率分布函数，由此可以导出关于观测向量的概率分布可用公式（2-24）来表示。

$$P(v) = \frac{e^{-F(v)}}{Z} \text{ with } Z = \sum_v e^{-\mathcal{F}(v)} \tag{2-24}$$

当假设输入层及隐层节点都为二值节点（$v_j, h_i \in \{0,1\}$）时，由各层单元之间的独立性假设，可以得到公式（2-25）和（2-26）的条件概率分布。此时的自由能函数也简化为见公式（2-27）所示。

$$P(h_i = 1 \mid v) = sigm(c_i + W_i v) \tag{2-25}$$

$$P(v_i = 1 \mid h) = sigm(b_i + W_j' h) \tag{2-26}$$

$$\mathcal{F}(v) = -b'v - \sum_i \log(1 + e^{(c_i + W_i v)}) \tag{2-27}$$

进而可以导出公式（2-28）、（2-29）和（2-30）的模型目标函数 log 似然对各个参数的导数。

$$-\frac{\partial \log p(v)}{\partial W_{ij}} = E_v\left[ p(h_i \mid v) \cdot v_j \right] - v_j^{(i)} \cdot sigm(W_i \cdot v^{(i)} + c_i) \tag{2-28}$$

$$-\frac{\partial \log p(v)}{\partial c_i} = E_v\left[ p(h_i \mid v) \right] - sigm(W_i \cdot v^{(i)}) \tag{2-29}$$

$$-\frac{\partial \log p(v)}{\partial b_i} = E_v\left[ p(v_j \mid h) \right] - v_j^{(i)} \tag{2-30}$$

注意到其中包含有期望的项，在实际求解中，一般采用单步吉布斯采样（Gibbs Sampling）来计算该项的值，这种方法也称为 Contrastive Divergence 方法。尽管采用单步采样，理论上有许多不完备的地方，但实际效果表明这样的做法足够取得很好的训练结果。

**（2）自动编码器 Autoencoder**

一个典型的 Autoencoder 模型结构如图 2-3 所示。与 RBM 相同的是同样为了求的隐藏层节点的值作为模型的输出，但是不同之处在于，训练时候，Autoencoder 多加了一层输出层 *o*，使得整个结构成为了一个两层的神经网络，其模型也就成为一个先编码、再解码的过程，设置训练的目标为使输出层 *o* 的值尽可能等于输入层，这也是其被称为 Autoencoder 的由来。

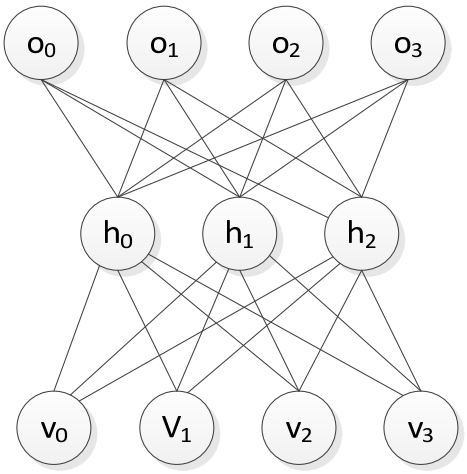



图 2- 3 Autoencoder

Autoencoder 结构中，由输入得到隐藏响应的过程一般称为 encode 过程，其前向激活函数通常采用 sigmoid 函数，见公式（2-31）。同时，该结构的解码 decode 过程可以用公式（2-32）来表示表示。

$$h(x) = sigmoid(Wx + b) = \frac{1}{1 + \exp(-Wx + b)} \tag{2-31}$$

$$o(h(x)) = s(W^T h + b_2) = s(W^T \cdot h(Wx + b_1) + b_2) \tag{2-32}$$

当采用高斯单元的时候，通常采用公式（2-33）的重构残差作为训练的目标函数。而采用二值单元时，则为互熵，见公式（2-34）。

$$L(x, o) = \frac{1}{2}\left\| o(x) - x \right\|^2 \tag{2-33}$$

$$L(x, o) = -\sum x \log o + (1 - x)\log(1 - o) \tag{2-34}$$

**正则化自动编码器**

在求解机器学习研究中最优化问题的时候，为了避免出现过拟合的现象，使模型训练的结构风险最小化，通常会引入正则化手段对求解过程进行约束。对于自动编码器，其存在多种正则化手段，以下将进行简要介绍。

**L2 正则**

对模型参数采用 2 范数作为正则化手段是最优化问题中最常用的一种手段，在 Autoencoder 中引入 L2 正则化之后，其训练目标函数就变为：

$$L(x,o)=\frac{1}{2}\left\|o(x)-x\right\|^2+\frac{\lambda}{2}\left\|W\right\|^2 \tag{2-35}$$

L2 正则同时也被称作为权值衰减（Weight Decay）[56]，由其对网络参数更新过程中过大的权值进行惩罚，以避免出现过拟合现象。

**稀疏正则**

稀疏化的思想是近年来随着压缩感知理论[57]的发展而发扬光大的一种正则化手段，在机器学习领域获得了广发的应用。与最优化问题中采用模型参数的 1 范数进行稀疏正则化[58]不同，在 Autoencoder 中稀疏正则化的手段是对隐藏节点的相应频率进行稀疏化约束。具体化为，令公式（2-36）表示隐藏节点 $j$ 在训练数据集中的平均响应值，$a(x^{(i)})$ 代表该节点对于样本 $x^{(i)}$ 的响应值，那么可以通过限制节点的平均响应值处于一个比较低的值，如 $\hat{\rho}_j=\rho,\ \rho=0.05$，来达到稀疏化的目的。

可以看到，自动编码器中的稀疏化约束是对节点的整体响应值进行约束，当稀疏化参数设置很小的时候，该节点只能对部分数据样本有较大的响应值，而多绝大多数样本的响应接近于 0，以保持整体的响应均值处在较低的水平。稀疏化正则一般通过公式（2-37）的 KL 散度方式进行建模，其中 $M$ 为隐藏节点的数量，则优化的目标函数变为公式（2-38）所示。

$$\hat{\rho}_j=\frac{1}{m}\sum_{i=1}^{m}\left[a_j(x^{(i)})\right] \tag{2-36}$$

$$\sum_{j=1}^{M}KL(\rho\,\|\,\hat{\rho}_j)=\sum_{j=1}^{M}\rho\log\frac{\rho}{\hat{\rho}_j}+(1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \tag{2-37}$$

$$L(x,o)=\frac{1}{2}\left\|o(x)-x\right\|^2+\lambda\sum_{j=1}^{M}KL(\rho\,\|\,\hat{\rho}_j) \tag{2-38}$$

**紧致正则**（Contractive Autoencoder）[59]

为了提高自动编码器对数据样本微扰变化的鲁棒性，文献[59]提出了紧致自动编码器。通过对映射函数雅克比矩阵的 Frobenius 范数进行约束，使得映射函数能够聚焦于训练样本周围，避免噪声的干扰。其训练目标函数为公式（2-39），

其中$\|J(x)\|^2$为映射函数$h$的雅可比矩阵的 Frobenius 范数，其第$j$行由连接矩阵$W$的第$j$行得来，公式（2-40）为映射函数的雅克比矩阵，该雅克比矩阵的 Frobenius 范数可写公式（2-41）的形式。

$$L_{CAE}(W,b;x,o)=\sum_{x\in D}L(x,o)+\lambda\|J(x)\|^2 \tag{2-39}$$

$$J(x_i)=\frac{\partial h}{\partial x}=\begin{vmatrix}\frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_N}\\ \vdots & \vdots & \vdots\\ \frac{\partial h_M}{\partial x_1} & \cdots & \frac{\partial h_M}{\partial x_N}\end{vmatrix}=\begin{vmatrix}h_1(1-h_1)\mathrm{W}_{11} & \cdots & h_1(1-h_1)\mathrm{W}_{1N}\\ \vdots & \vdots & \vdots\\ h_M(1-h_M)\mathrm{W}_{M1} & \cdots & h_M(1-h_M)\mathrm{W}_{MN}\end{vmatrix} \tag{2-40}$$

$$\|J(x)\|_F^2=\sum_{i=1}^{d_h}(h_i(1-h_i(x))^2\sum_{j=1}^{d_x}W_{ij}^2 \tag{2-41}$$

雅克比矩阵的 Frobenius 范数代表了映射函数在数据点$x$局部的紧致程度，通过添加紧致约束限制了映射函数的变化，使之不易受到噪声干扰。但另一方面，由于目标函数中重构误差的作用，对这个限制作用进行了平衡，使映射函数能够朝近邻样本方向进行扩散，避免了过于聚焦在单独样本点 造成过拟合的情况。

**光滑自动编码器**（Smooth Autoencoder）[60]

在紧致自动编码器中，训练目标函数聚集在样本本身，有可能造成学习到的特征在局部不够平滑，针对这一点，文献[60]提出了光滑自动编码器，类似于流形学习中将对样本近邻赋予权重的做法，以及局部线性重构的思想，设置训练目标函数为局部的近邻样本带权线性重构误差，具体为做法公式表示见公式（2-42），其中的权重函数$w(x_j,x_i)$为公式（2-43）形式的高斯核函数。

$$L_{DAE}(W,b;x,o)=\sum_{i=1}^{N}\sum_{j=1}^{K}w(x_j,x_i)L(x_j,g(f(x_i)))+\lambda\sum_{j=1}^{M}KL(\rho\|\hat{\rho}_j) \tag{2-42}$$

$$w(x_j,x_i)=\begin{cases}\frac{1}{Z}\exp(-\frac{\|x_i-x_j\|^2}{\sigma}) & x_i\in\mathcal{N}_i\\ 0 & \text{其他}\end{cases} \tag{2-43}$$

通过约束局部带权重构误差的方式，光滑自动编码器可使学习到的特征在样本间变化更加平滑，更加符合流形光滑的特征，也能起到抑制噪声，提高鲁棒性的作用。

**降噪自动编码器**（Denoising Autoencoder）[61]

降噪自动编码器不同于之前所介绍各种正则化自动编码器，其并不是通过在目标函数中添加正则约束项的方式来进行正则化，而是在训练过程中，通过对数据样本添加高斯白噪声，并求解样本重构的方式来实现正则化。

该编码器的具体算法流程为：对于给定数据样本，首先通过某种手段添加噪声破坏数据样本，在原文献中，作者采用随机将数据样本的部分维度设为 0 的方式得到“破损”样本，$\tilde{x} \sim q_{\mathcal{D}}(\tilde{x} \mid x)$，$q_{\mathcal{D}}$代表为样本添加噪声的过程。接下来进行的操作与一般 Autoencoder 一致，将该破损样本输入网络获得输出，最终训练目标函数为重构原始样本：

$$L_{DAE}(W, b; x, o) = \sum_{x \in D} \mathbb{E}_{\tilde{x}_i \sim q(\tilde{x}_i \mid x_i)} \left[ L(x_i, g(f(\tilde{x}_i))) \right] \tag{2-44}$$

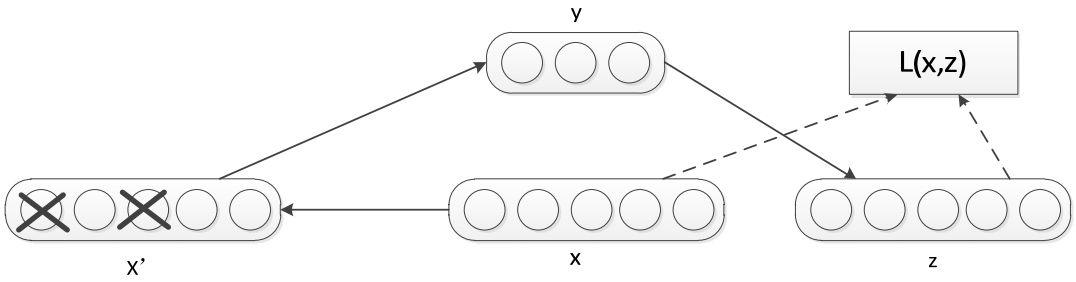



图 2- 4 Denoising Autoencoder 算法示意图[61]

算法的流程可由图 2-4 表示。降噪自动编码器的思想来源于对人感知过程的观察，人们在感知现实世界时，即便目标有一部分信号的缺失，比如物体的一小部分被遮挡，或者语音出现片段的缺失，但人依然能够准确的进行判断。降噪自动编码器试图去重现这一认知过程，实际使用中也取得了不错的效果。

## 2.3　本章小结

本章对流形学习算法进行了归纳和总结，指出流形学习都可以归结成寻找在原始空间和局部空间之间结构保持的映射关系，而对于结构保持这一约束，

不同的定义方式代表了不同的物理含义。对于流形学习算法而言，多数方法不能提供显式的映射函数，这制约了其进一步的推广。本章另一部分，介绍了深度学习中的基本算法，RBM 和自动编码器，以及自动编码器的多个正则变种。自动编码器能够提供显式的映射函数，若能将其与流形学习方法结合，特性上彼此补充，便能获得显式的映射函数并且对数据样本之间的关系进行流形刻画。.

# 第三章　局部结构保持映射

## 3.1　引言

在绪论中，本文阐述了解析特征提取的主要思想，其目标是希望从数据中提取到的特征表示能够反映出其内在的生成因素，并且在特征空间中，数据样本的分布与这些生成因素的变化是对应的，也即随着特征空间的某些维度进行采样能够得到相应的因素变化所生成的数据样本。以人脸图片为例，可以想象人脸图片生成过程中，其生成因素是有限的，包括ID、表情、视角、光照，在理想化的解析特征提取的条件下，最终的特征表示应该能够很好的反映出这些因素的变化作用，以及造成样本之间不同的原因解析；当沿着反映ID变化的维度进行采样时，得到的是不同人的人脸样本，而这些样本的其他条件如表情、光照等是近似的，也就是说ID这一维度上只反映ID的变化。另一方面，对绝大多数机器学习研究问题来说，数据内在的生成或者变化因素的数量是很少的，也就是其本征维度很低，可以认为这些高维的数据是分布在低维的非线性流形上，所以可以引入流形学习的方法对特征提取进行研究。

在前面关于流形学习的简介中，总结了流形学习算法的主要思想，可以看到，主流的流形学习算法都可以归结成寻找某种几何结构保持定义下的特征映射，在这之中并未对特征空间的数据解析关系进行直接的建模。在理想化解析特征提取情况下，所提取到的特征维度是可以准确反映数据中的内在生成原因的，但在实际研究中，很难达到理想化的条件：首先，数据内在的生成因素，或者本征维度是不可知的，而且这些因素相互之间的作用十分复杂，给模型的选择带来困难；另一方面，这些因素的作用是无法量化的，所以很难建立优化目标以及进行验证。本章从流形的角度，对解析特征提取过程进行了流形概念下的阐述，并提出了一种新的结构保持映射算法。通过局部切空间的联系建立低维嵌入空间中样本关系，进而构建全局的目标优化函数，求解得到最终的显式映射函数。

## 3.2 Autoencoder 的流形解释

尽管深度学习方法近年来取得了极大的成功，但其相关的理论分析工作依然未能取得突破性进展，近年来，许多学者针对深度学习算法的理论分析开展了多方面的研究工作，其中包括从概率模型的角度[62]，以及从几何分析角度等等[63]，对受限玻尔兹曼机以及自动编码器的理论性质进行了深入的研究。这其中以紧致自动编码器的相关流形分析最为完善[64]，由前面关于紧致自动编码器的介绍，紧致自动编码器的正则化约束通过极小化投影函数的一阶偏导矩阵的 Frobenius 范数，使投影函数对输入信号的变化不敏感；另一方面，最小化重构误差的存在，减弱了正则化约束的效果，使投影函数对原始空间某些输入变化方向敏感，从而这些方向张成了流形局部的切空间。

为了进一步解释这一点，首先引入坐标卡的概念。在流形的定义中，流形上某一点是局部同胚于欧式空间的，若在该局部欧式空间上建立坐标系统，则称该坐标系统为坐标卡（Chart），而流形上所有坐标卡的集合称为图册（Atlas）。为了能够定义流形上的图册，映射函数 $h$ 应当是微分同胚的，也即在流形的局部空间中满足光滑和可逆两个条件。由于在自动编码器中，模型选择了线性组合以及 $sigmoid$ 激活函数的形式，$h(x)=sigmoid(Wx+b)$，可以看到其是无穷可导的，满足了光滑的条件。而可逆条件可由函数的内射条件推出，内射条件要求对于不同的输入 $x$，函数 $h(x)$ 值应是不同的。由自动编码器训练目标函数的中最小化重构样本误差的设置，要满足该条件，映射函数应尽可能区分不同的输入样本，否则会造成混淆从而增大重构误差，因此满足了内射的条件。另一方面，为得到从映射 $h$ 到 $x$ 内射条件成立，应满足公式（3-1），则 $W$ 的行向量应可以构成一组基坐标，也即 $\Delta_{ij}$ 可由 $W$ 行向量构成的基进行线性加权表示，如公式（3-2）。当该条件满足时，$h(x)$ 便是可逆的。为满足该条件，可对 $h(x)$ 定义 $x$ 映射后有限取值范围，那么在该领域内就自然满足满射条件，从而满足内射条件。这样在数据样本的局部空间内，原始空间和映射空间之间的双向内射条件是满足的，这也构建局部坐标系统提供了基础。

$$\forall i,j\ \ h(x_i)=h(x_j)\Leftrightarrow W\Delta_{ij}=0,\quad \Delta_{ij}=x_i-x_j \tag{3-1}$$

$$\forall i,j\ \ \exists\alpha\in \mathrm{R}^{d_h},\ \Delta_{ij}=\sum_k^{d_h}\alpha_k W_k \tag{3-2}$$

由以上分析可以看到，紧致自动编码器的映射函数满足构建流形局部坐标系的条件，接下来将进一步说明如何构建该坐标系。同样由目标函数出发，重构的约束条件和雅克比矩阵 $F$ 范数的组合使得映射函数只对局部近邻方向上的变化敏感，对其余方向响应较低，从而更聚集在数据本身的流形上，这一性质可从映射函数的雅克比矩阵的频谱上反映出来。考虑到前提假设中数据流形的本征维度是较低的，设雅克比矩阵 $J(x)=\dfrac{\partial h}{\partial x}$ 的秩为 $k$，则可认为映射函数只对 $k$ 个方向上的变化最为敏感，而这 $k$ 个方向是处在 $J(x)$ 的非零奇异值对应的奇异向量所张成的线性空间上的，由此可以通过公式（3-3）对雅克比矩阵进行奇异值分解来获得局部的基坐标系。

$$SVD(J^T(x))=U(x)S(x)V^T(x) \tag{3-3}$$

设 $\mathcal{B}_x$ 为公式（3-4）的主奇异向量，那么流形的局部切平面 $\mathcal{H}_x$ 就可由公式（3-5）的主奇异向量张成的空间表示。可以看到，主奇异向量 $\mathcal{B}_x$ 构成了一个局部的坐标卡 $\phi_x$，进而可以得到公式（3-6）流形上的图册。

$$\mathcal{B}_x=\{U_{\cdot k}(x)\,|\,S_{kk}(x)>\varepsilon\} \tag{3-4}$$

$$\mathcal{H}_x=\{x+v\,|\,v\in \mathrm{span}(\mathcal{B}_x)\} \tag{3-5}$$

$$\mathcal{A}=\left\{\left(\mathcal{M}_x,\phi_x\right)|\,x\in\mathcal{D},\ \phi_x(\tilde{x})=\mathcal{B}_x(\tilde{x}-x)\right\} \tag{3-6}$$

以上，便由紧致自动编码器所定义的条件，推导出了其关于流形的一些概念的具体形式，这也构成了本文研究的另一个理论基础。但需要指出的是，在本文的研究当中，并不需要具体化坐标卡和图册的数学化表达，只利用了映射函数的雅克比矩阵能够体现流形局部切空间的性质进行展开，在以上的分析中可以看到这一点是满足条件的。

## 3.3　算法思想

在自动编码器的流形性质解释中，映射函数的雅可比矩阵可近似表示低维

嵌入流形局部切空间，并且对雅可比矩阵进行奇异值分解能够建立局部的坐标系统。利用该性质，我们可以不用强制要求数据的特征表示与生成因素之间有明确的解析关系，只需要数据特征表示相互之间的差别能够反映因素的变化作用，便能达到特征表示随生成因素变化所变化的效果，和一般流形学习算法一样，这样构成了一种局部结构保持的映射。

考虑原始空间中的数据样本 $x$ 和其近邻 $x_i$ ，由数据的流形分布假设我们可以认为原始空间中近邻的样本其生成因素是十分接近的，或者因素没有强烈变化；同时，由自动编码器的流形解析可以得知，对映射函数的雅可比矩阵进行奇异值分解可以建立切空间中的坐标系统，该坐标系统下的坐标方向就代表了数据在局部变化的主要方向。对于两个相近的数据样本，由于假设其生成因素是相近的，那么其局部坐标系统之间也应该是相近的，也就是说当局部坐标系统沿着流形上移动时，坐标系统保持对流形的相对平行，这也是流形当中平行移动的定义（Parallel Transport）[65]，如图 3-1 所示，当局部坐标系统从 a 移动到 b 时，坐标系统间相对于流形是保持平行的。

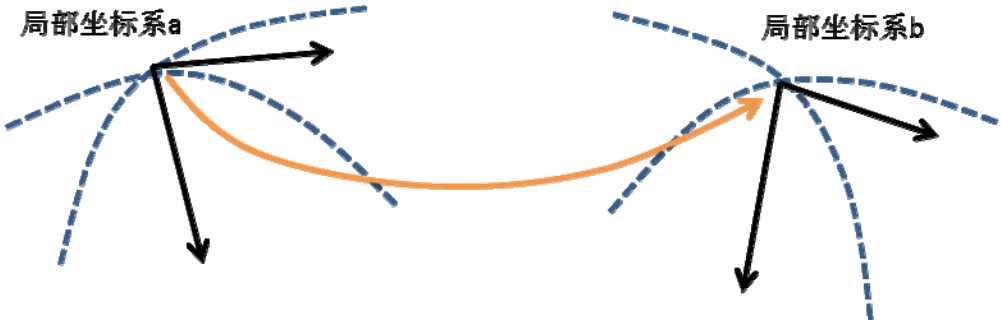



图 3- 1 流形上的平行移动

由以上分析，可以推想原始空间中近邻的两个数据样本，其在映射空间中除了距离接近之外，雅可比矩阵也是相似的，因此我们可以在原始和嵌入空间中分别定义样本的相似度度量，并求解一映射函数，使样本在不同空间之间的局部几何结构是保持的，这便构成了所研究问题的最终定义。在原始空间中，通常采用欧式距离作为样本相似度的度量，在嵌入空间中，由以上分析可以看

到雅可比矩阵代表了样本在嵌入空间局部的变化模式，因此可采用如公式（3-7）矩阵差的 Frobenius 范数作为矩阵相似度度量，其中 $J(x)$ 代表嵌入空间中样本 $x$ 处映射函数的雅可比矩阵，$\gamma$ 为权重参数。

$$\left\| f(x)-f(x_i) \right\|_2^2+\gamma \left\| J(x)-J(x_i) \right\|_F^2 \tag{3-7}$$

在定义的相似度度量基础上，类似于随机近邻嵌入 SNE 的思想，将样本之间的局部关系用概率空间测度进行表示。具体的，对于原始空间的数据样本，定义公式（3-8）表示样本间的条件概率测度，其中，$k$ 为设置的近邻数目。一般地，为了得到单一的联合概率测度，通常采用对称化手段，也即公式（3-9）所示的样本间的联合概率测度。

$$p_{j|i}=\frac{\exp\left(-\left\| \mathrm{x}_i-\mathrm{x}_j \right\|^2/2\sigma_i^{\,2}\right)}{\sum_{k\neq i}\exp\left(-\left\| \mathrm{x}_i-\mathrm{x}_k \right\|^2/2\sigma_i^{\,2}\right)} \tag{3-8}$$

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n} \tag{3-9}$$

对于嵌入空间，公式（3-10）定义样本间的概率测度，其中参数 $\alpha$ 称为自由度。注意到这里采用了学生 $t$-分布，和 $t$ 分布随机近邻嵌入算法所采用的方式一致，在文献[55]中，作者指出这样的做法比 SNE 算法使用的高斯分布更能处理长尾数据分布的问题，具有更高的鲁棒性。由此，可以通过 KL 散度来衡量两个概率测度的相似度。

$$q_{ij}=\frac{\left(1+\left(\left\| f(x)-f(x_i) \right\|_2^2+\gamma\left\| J(x)-J(x_i) \right\|_F^2\right)/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k\neq l}\left(1+\left(\left\| f(x)-f(x_i) \right\|_2^2+\gamma\left\| J(x)-J(x_i) \right\|_F^2\right)/\alpha\right)^{-\frac{\alpha+1}{2}}} \tag{3-10}$$

$$C=KL\left(P\,\|\,Q\right)=\sum_{i\neq j}p_{ij}\log\frac{p_{ij}}{q_{ij}}=\sum_{i\neq j}p_{ij}(\log p_{ij}-\log q_{ij}) \tag{3-11}$$

至此，公式（3-11）便是最终需要求解的问题，在此本文将通过公式（3-11）训练得到的模型称为局部结构保持映射 LSPP（Local Structure Preserving Projection）算法。注意到，在构建问题的过程中并没有给出映射函数 $f$ 的形式，

在实验中，本文采用了多层神经前馈神经网络作为参数化的映射函数，对于该神经网络，逐层采用紧致自动编码器进行参数预训练，再将整个网络进行整体的精调。这种逐层预训练再整体精调的方式被广泛的应用在多层神经网络的训练当中，并且取得了非常好的实用效果。

## 3.4 问题求解

在实验中本文采用了多层神经网络的参数化映射函数形式，因此可以很方便的采用多层神经网络中的反向传播方法[72]进行训练，同时采用随机梯度下降来处理数据集较大时候的优化问题。

**梯度推导**

令模型参数为$\theta$，则所需的梯度用公式（3-12）表示。

$$\frac{\partial C}{\partial \theta}=\frac{\partial C}{\partial f(x)}\cdot\frac{\partial f(x)}{\partial \theta}+\frac{\partial C}{\partial J(x)}\cdot\frac{\partial J(x)}{\partial \theta} \tag{3-12}$$

在嵌入空间的概率测度中本文采用了 $t$ 分布的形式，引入了自由度$\alpha$，此处为简化推导过程，令$\alpha$=1，有：

$$q_{ij}=\frac{\left(1+\left\|f(x_i)-f(x_j)\right\|_2^2+\gamma\left\|J(x_i)-J(x_j)\right\|_F^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left\|f(x_k)-f(x_l)\right\|_2^2+\gamma\left\|J(x_k)-J(x_l)\right\|_F^2\right)^{-1}} \tag{3-13}$$

定义中间变量：

$$d_{ij}=\left\|f(x_i)-f(x_j)\right\|_2^2+\gamma\left\|J(x_i)-J(x_j)\right\|_F^2$$

$$Z=\sum_{k\neq l}(1+d_{kl})^{-1}$$

则有：

$$q_{ij}=\frac{(1+d_{ij})^{-1}}{Z} \tag{3-14}$$

目标函数导数中对映射函数和雅克比矩阵函数的偏导便可得：

$$\frac{\delta C}{\delta f(x_i)}=2\sum_j\left(\frac{\delta C}{\delta d_{ij}}+\frac{\delta C}{\delta d_{ji}}\right)(f(x_i)-f(x_j))=4\sum_j\frac{\delta C}{\delta d_{ij}}(f(x_i)-f(x_j)) \tag{3-15}$$

$$\frac{\delta C}{\delta J(x_i)} = 2\sum_j \left( \frac{\delta C}{\delta d_{ij}} + \frac{\delta C}{\delta d_{ji}} \right) (J(x_i) - J(x_j)) = 4\sum_j \frac{\delta C}{\delta d_{ij}} (J(x_i) - J(x_j)) \tag{3-16}$$

这其中：

$$\begin{aligned} \frac{\delta C}{\delta d_{ij}} &= -\sum_{k \neq l} p_{kl} \frac{\delta(\log q_{kl})}{\delta d_{kl}} \\ &= -\sum_{k \neq l} p_{kl} \frac{\delta(\log q_{kl} Z - \log Z)}{\delta d_{kl}} = -\sum_{k \neq l} p_{kl} \left( \frac{1}{q_{kl} Z} \frac{\delta((1+d_{kl})^{-1})}{\delta d_{ij}} - \frac{1}{Z} \frac{\delta Z}{\delta d_{ij}} \right) \end{aligned} \tag{3-17}$$

其中的梯度 $\frac{\delta((1+d_{kl})^{-1})}{\delta d_{ij}}$ 当且仅当 $k=i$ 和 $l=j$ 时不为零，以上就可进一步简化为：

$$\frac{\delta C}{\delta d_{ij}} = \sum_j \frac{p_{ij}}{q_{ij} Z} (1+d_{ij})^{-2} - \sum_{k \neq l} p_{kl} \frac{(1+d_{ij})^{-2}}{Z} = \sum_j (p_{ij} - q_{ij})(1+d_{ij})^{-1} \tag{3-18}$$

以单层自动编码器为例，映射函数 $f(x) = \frac{1}{1+e^{-(Wx+b)}}$ ，映射函数对参数的梯度为：

$$\frac{\partial f(x)}{\partial W} = h(1-h) \cdot x \tag{3-19}$$

$$\frac{\partial f(x)}{\partial b} = h(1-h) \tag{3-20}$$

对于雅克比矩阵对参数的偏导，由雅克比矩阵的表示：

$$J(x_i) = \frac{\partial h}{\partial x} = \begin{vmatrix} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_N} \\ \vdots & \vdots & \vdots \\ \frac{\partial h_M}{\partial x_1} & \cdots & \frac{\partial h_M}{\partial x_N} \end{vmatrix} = \begin{vmatrix} h_1(1-h_1) \mathrm{W}_{11} & \cdots & h_1(1-h_1) \mathrm{W}_{1N} \\ \vdots & \vdots & \vdots \\ h_M(1-h_M) \mathrm{W}_{M1} & \cdots & h_M(1-h_M) \mathrm{W}_{MN} \end{vmatrix} \tag{3-21}$$

进一步对模型参数进行求导，得到其对自动编码器中参数 $W$ 和 $b$ 的导数为：

$$\frac{\partial J(x_i)}{\partial W}=\begin{vmatrix} \frac{\partial h_1(1-h_1)\mathrm{W}_{11}}{\partial W} & \cdots & \frac{\partial h_1(1-h_1)\mathrm{W}_{1N}}{\partial W} \\ \vdots & \vdots & \vdots \\ \frac{\partial h_M(1-h_M)\mathrm{W}_{M1}}{\partial W} & \cdots & \frac{\partial h_M(1-h_M)\mathrm{W}_{MN}}{\partial W} \end{vmatrix}$$

$$=\begin{vmatrix} h_1(1-h_1)(1-2h_1)\mathrm{W}_{11}\sum_N \mathrm{W}_{1i}\,x_{1i}+h_1(1-h_1) & \cdots \\ \ddots & \\ \cdots & h_M(1-h_M)(1-2h_M)\mathrm{W}_{MN}\sum_N \mathrm{W}_{Mi}\,x_{Mi}+h_M(1-h_M) \end{vmatrix} \tag{3-22}$$

$$\frac{\partial J(x_i)}{\partial b}=\begin{vmatrix} \frac{\partial h_1(1-h_1)\mathrm{W}_{11}}{\partial b} & \cdots & \frac{\partial h_1(1-h_1)\mathrm{W}_{1N}}{\partial b} \\ \vdots & \vdots & \vdots \\ \frac{\partial h_M(1-h_M)\mathrm{W}_{M1}}{\partial b} & \cdots & \frac{\partial h_M(1-h_M)\mathrm{W}_{MN}}{\partial b} \end{vmatrix}$$

$$=\begin{vmatrix} h_1(1-h_1)(1-2h_1)\mathrm{W}_{11} & + & h_1(1-h_1)(1-2h_1)\mathrm{W}_{1N} \\ \vdots & \vdots & \vdots \\ h_M(1-h_M)(1-2h_M)\mathrm{W}_{M1} & + & h_M(1-h_M)(1-2h_M)\mathrm{W}_{MN} \end{vmatrix} \tag{3-23}$$

至此，将以上结合便可得到目标函数对参数的梯度表示，从而代入随机梯度下降方法进行求解。

## 3.5 算法分析

在建立模型过程中，本文并没有对雅可比矩阵进行奇异值分解去得到坐标系统的表示，以及通过全局对齐的方式来串联坐标卡。这样做的原因主要有两方面:首先，采用奇异值分解之后得到的坐标表示不能进一步通过反向传播算法进行求解优化，因为其无法进行梯度求导，从优化方法层面进行了限制。其次，采用矩阵范数形式来刻画相似度可以使用标量来表示坐标系统的相似度，从而可以嵌套到 KL 散度的优化框架中。

与所提算法最为相似的算法是 Parametric t-SNE 算法，两者在问题框架上是相似的，但与其不同的是，Parametric t-SNE 算法对于映射空间的结构关系只使用了坐标关系，而本文建立的关系从流形学习、因素解析的角度，引入了雅克

比矩阵 F 范数相似的约束，对邻近样本之间的切空间进行了刻画，在样本关系建立上，所提方法更能表现出因素解析的作用。

另一个具有相似思想的算法是流形学习中的向量散射映射算法（Vector Diffusion Map）[66]，其同样采用邻近样本切空间上的坐标系来建立样本之间的联系。算法的步骤是在局部利用 PCA 建立坐标系统，接着通过平移矩阵来建立了坐标系统之间的联络（Connection），其论文指出，该平移矩阵正是流形中拉普拉斯联络算子的具体形式，在局部的平移过程中是具有平行性质的。

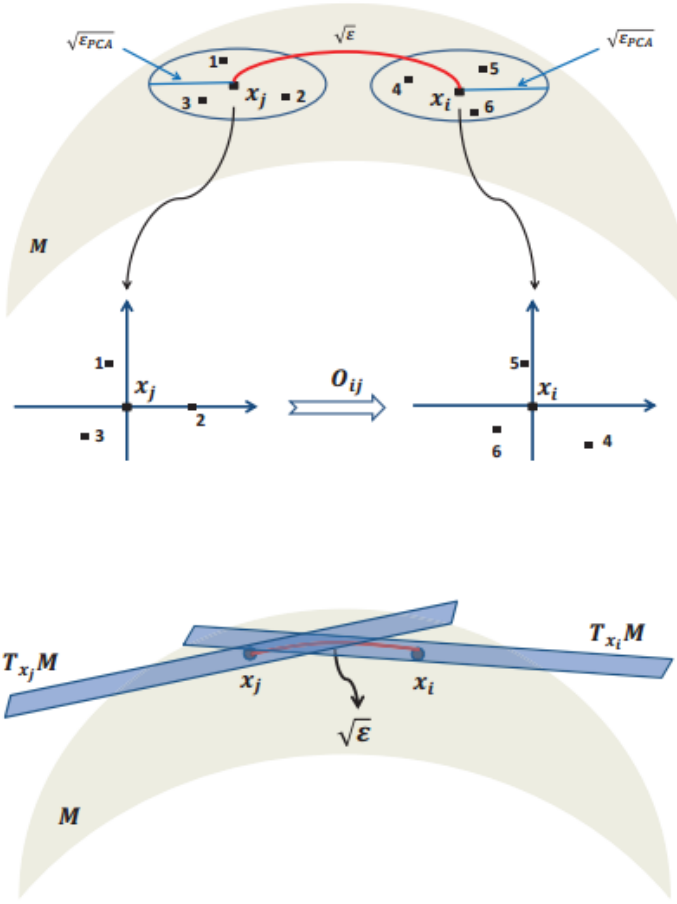



图 3- 2 VDM 算法主要思想

LSPP 算法的理论基础与紧致自动编码器 CAE 关系十分密切，其作者在 CAE 基础又进一步发展出了高阶自动编码器 CAE-H[67]，对映射函数的 Hessian 矩阵进行约束。公式（3-24）为利用 Hessian 矩阵的一个近似表示，构建公式（3-25）

所示模型的损失函数。

$$\left\|H_f(x)\right\|^2 = \lim_{\sigma\to 0}\frac{1}{\sigma^2}\mathbb{E}\left[\left\|J_f(x)-J_f(x+\varepsilon)\right\|^2\right] \tag{3-24}$$

$$\mathcal{J}_{CAE+H}(\theta) = \sum_{x\in D_n} L(x, g(f(x))) + \lambda\left\|J_f(x)\right\|^2 + \gamma\mathbb{E}\left[\left\|J_f(x)-J_f(x+\varepsilon)\right\|^2\right] \tag{3-25}$$

LSPP 的思想为通过约束 Hessian 矩阵，可以使雅克比矩阵的变化更加平缓，而我们通过约束邻近样本之间雅克比矩阵的相似度，同样也起到了平缓变化的效果，但我们通过概率度量的方式，建立了多样本之间的联系，对数据样本关系的刻画更加直接，而 CAE+H 则是着眼于映射函数的性质，在研究的出发点上存在不同。

## 3.6 实验结果与分析

在本节中，将通过在 MNIST 手写体字符识别数据集[68]上进行一系列对比实验，对所提算法性能进行比较与分析。

Mnist 手写体字符库是最为常用的多层神经网络算法评测模型，其包含了数字 0 到 9 的手写体样本，总共有 5 万幅训练图像、1 万副验证图像以及 1 万幅测试图像。其手写体字样存在着许多形变以及扭曲，对于正确的识别造成了很大的困难，同时由于数据样本较多，对算法的计算开销也有较高的要求。Mnist 数据集的示例样本见图 3-3。

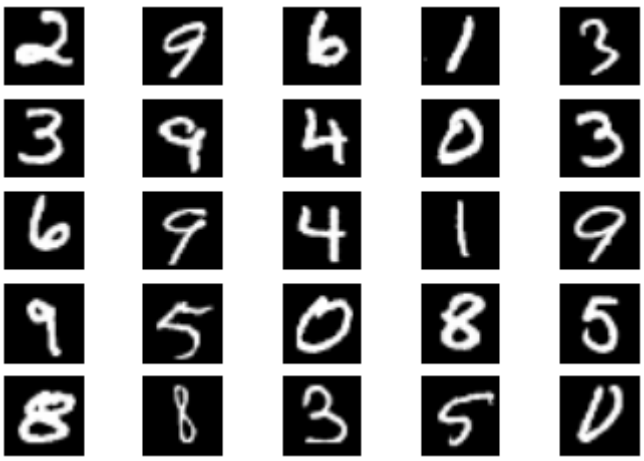


图 3- 3 Mnist 数据样本示例

在 Mnist 原始数据集之外，我们还采用了加拿大蒙特利尔大学 LISA 小组发布的 Mnist 数据集的扩展数据集[69]，在该扩展数据集中，对原始样本添加了多种噪声，如对样本进行旋转得到的 mnist-rot，对样本添加随机背景噪声得到的 mnist-back-rand，对样本添加图片背景的 mnist-back-image。通过对原始数据集添加额外噪声的方法，可以评测算法在处理噪声以及鲁棒性等方面的性质。

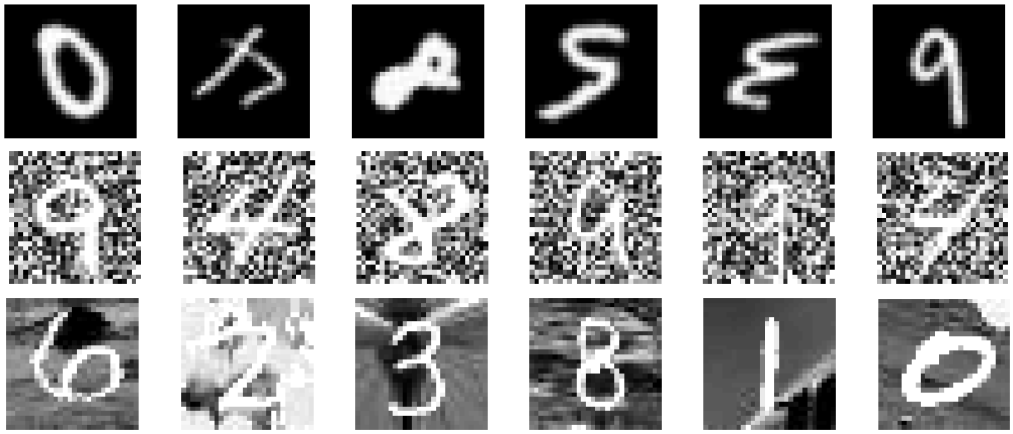


图 3- 4 Mnist 扩展数据集样本示例[69]

**Mnist 原始数据集实验比较**

在原始数据集上，我们采用单层网络形式，与几种自动编码器算法进行了比较。我们采用了多组实验参数，分别设置隐层网络节点数目为 100、200、300、400、500、1000，以观察其对算法性能的影响，在比较算法中，DAE 代表降噪自动编码器，SAE 代表稀疏自动编码器，CAE 代表紧致自动编码器，LSPP 是本章所提出的算法。分类识别错误率（百分比）如表 3-1 所示，使用图表表示见图 3-5。

在 Mnist 原始数据集中，所提算法取得了与 AE 类算法基本相近的性能表现，在个别的参数设置上，LSPP 算法能够领先其他算法，但总体上并没有显示出特别的优势，本文认为这只是体现了算法与主流的 AE 算法是具有可比性的。另一方面，本章所提算法的出发点是在于尝试去探索数据样本之中的因素解析关系，通过流形性质对这一关系进行刻画。我们期望能够在所学习到的特征表示中明显的体现出这些性质，在此，我们采用了流形降维的方法将数据样本投影到二

维空间进行可视化并进行分析，所得到的样本分布如图 3-6 所示。

表 3- 1 算法在 Mnist 原始数据集上的性能比较

| | NN | DAE | SAE | CAE | LSPP |
|---|---|---|---|---|---|
| 100 | 3.45 | 2.45 | 2.71 | 2.42 | **2.41** |
| 200 | 2.32 | 1.97 | 1.88 | **1.74** | 1.89 |
| 300 | 2.02 | 1.74 | **1.63** | 1.67 | 1.82 |
| 400 | 1.82 | 1.57 | 1.54 | **1.54** | 1.56 |
| 500 | 1.87 | 1.47 | 1.53 | 1.47 | **1.46** |
| 1000 | 1.6 | 1.24 | 1.38 | **1.21** | 1.33 |

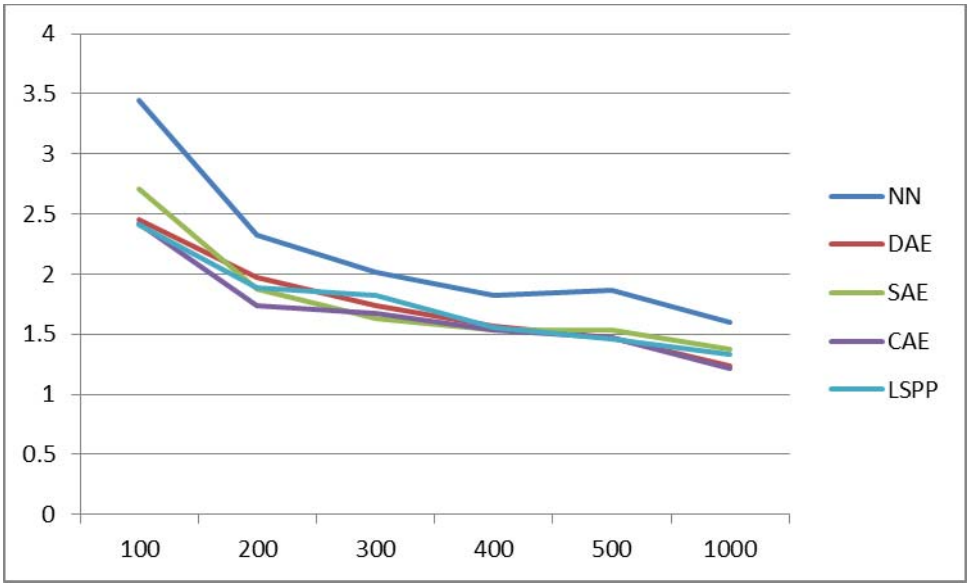



图 3- 5 Mnist 原始数据集中的算法性能比较

通过对图 3-6 的分析可以看到，各个类别样本在空间中形成了聚团，这是其分类性能的保证。在类别相互之间的空间关系上，在图的右半部可以观察到类别 4、7、9 的数据样本之间呈现出了近似的椭圆形空间分布，并且类别之间在空间上表现出了一定的有序性。同样的特性也可以在左下部分的类别 2、6、0 之间，以及左上部分的 5、8、1 之间观察到。在文献[64]的数据流形假设中，其

猜想不同类别的数据样本一般会聚集在一个局部子空间之上，并且类别之间有明显的低概率分布区域形成边界，而我们对数据样本可视化的结果基本符合了这一设想。另一方面，我们在通过流形学习中的概念构建问题时，对因素解析特性的一个核心阐述是流形中的平行移动性质。从样本分布的形态以及相互之间的相似性所表现出来的样本之间的关系来看，最终的特征表示大体上达到了类别之间的对应性，表现出了我们所期望的平行移动的性质。

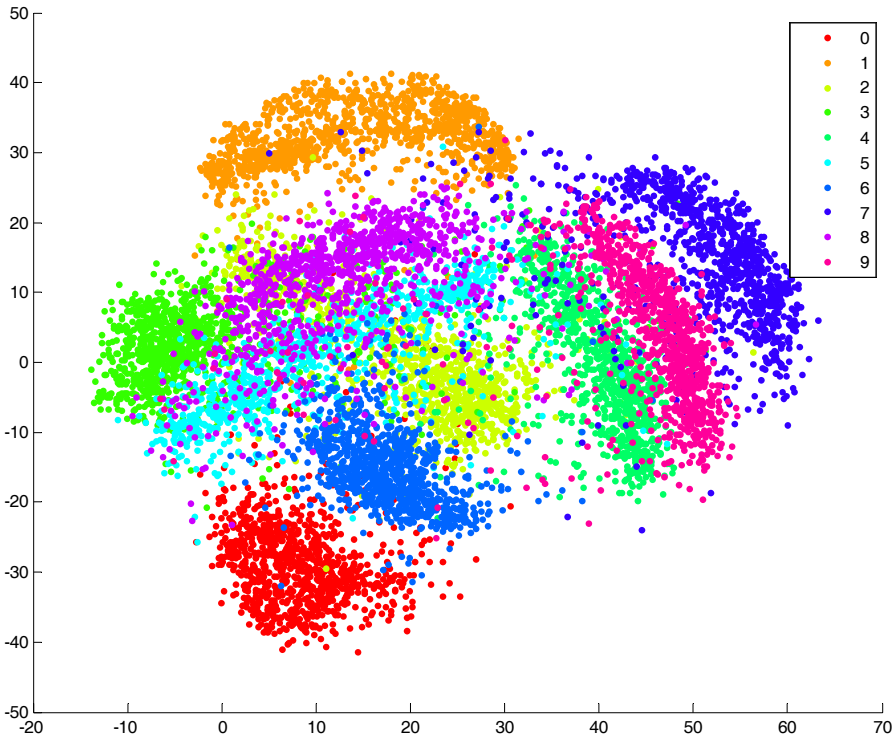



图 3- 6 Mnist 样本投影可视图

**Mnist 扩展数据集实验比较**

在 Mnist 扩展数据集上的实验，我们采用了和[59]一样的实验方法，对多层网络进行评测。对比算法中，RBM-3 是采用 RBM 进行逐层初始化的 3 层神经网络，类似的，SAE-3、DAE-3 是分别采用稀疏自动编码器和降噪自动编码器进行网络初始化的 3 层模型。在 CAE 上，采用了单层和两层进行比较。对于所提

LSPP 算法，我们采用了两层网络的形式，每层采用 LSPP 算法进行权值初始化，再对网络进行精调优化。各个算法的错误率（百分比）比较如下表所示，其中对比算法的性能摘自文献[59]以体现公平性。

表 3- 2 Mnist 扩展数据集上的算法性能比较

| | RBM-3 | SAE-3 | DAE-3 | CAE-1 | CAE-2 | LSPP-2 |
|---|---|---|---|---|---|---|
| mnist-rot | 10.3 | 10.3 | 9.53 | 11.59 | 9.66 | **9.42** |
| mnist-back-rand | **6.73** | 11.28 | 10.03 | 13.57 | 10.9 | 9.98 |
| mnist-back-image | 16.31 | 23 | 16.68 | 16.7 | 15.5 | **15.2** |

可以看到，算法在两个数据集合上取得了最优的性能，稍微领先于 CAE，表明算法能够有效的处理数据集中的噪声。与 CAE 类似，LSPP 算法基于映射函数的雅克比矩阵进行模型构建，在切空间中对数据样本的特征进行分析，两种算法均由于其他 AE 算法，表明映射函数的切空间的性质与其对噪声的抑制能力有关，也为未来算法的进一步发展提供了切入点。在随机背景噪声数据集上，RBM-3 模型取得了极大的领先，这或许与其模型本身比较适合该数据集的特性有关。

本章的研究中，我们采用了 Theano 深度学习软件包[70]进行实验，注意到在 3.4 节模型的求解推导过程中，最终要计算的梯度形式十分复杂，在编程上容易造成困难，梯度准确性的校对难以进行。在 Theano 中，对计算过程采用了图化表示，通过调用梯度求解指令就能获得准确的梯度，为编程带来方便。同时，我们还采用了 GPU 计算加速[71]，大大的提高了算法的运行效率。

## 3.7 本章小结

本章通过流形相关概念对因素解析特征提取过程进行了建模，其核心思想是认为原始空间中相近的样本其生成因素也是相近的，由此可以用流形切空间中的度量关系对样本在嵌入空间中的关系进行表示。在具体实现中，采用深度

学习中的自动编码器作为实现手段，基于自动编码器的流形性质分析，理论上保证了方法的可行性。利用映射函数雅克比矩阵能够表示切空间的性质，采用 Frobenius 范数作为相似度度量。在原始空间和嵌入空间中，通过样本近邻关系建立概率测度全局度量，利用 KL 散度对两者进行比较形成目标函数，求解优化得到显式的映射函数。

通过在手写体字符识别实验，表明了算法在处理噪声数据方面有一定的优势，同时通过对样本的空间关系进行投影可视化发现最终提取得到的特征基本满足了算法所构建的性质。

# 第四章　深度卷积神经网络与视觉行人检测

## 4.1　引言

近年来，深度学习算法在各个应用领域取得了极大的成功，其中在图像识别研究中一个重要的算法就是深度卷积神经网络。在前章研究工作的基础上，本文尝试去理解深度卷积神经网络在特征提取方面所具有的性质。在本章中，首先对深度卷积神经网络在视觉行人检测中的应用展开研究，了解其典型典型算法的一些特点，为进一步进行理论分析作为基础。

## 4.2　卷积神经网络（Convolutional Neural Network，CNN）

卷积神经网络可以看成是人工神经网络的一种，起源于视觉神经机制研究中感受野（Receptive Field）[73]局部特性的发现，1980 年日本学者 Fukushima 提出的 Neocognition 模型是该概念的首个具体实现[74]，到 1998 年 LeCun 等人对其结构进行进一步改进之后[68]，便形成了今天所广为熟知的卷积神经网络模型。

一个典型的卷积神经网络主要包含卷积层和下采样层（或被称为 Pooling 层），其结构如下图所示：

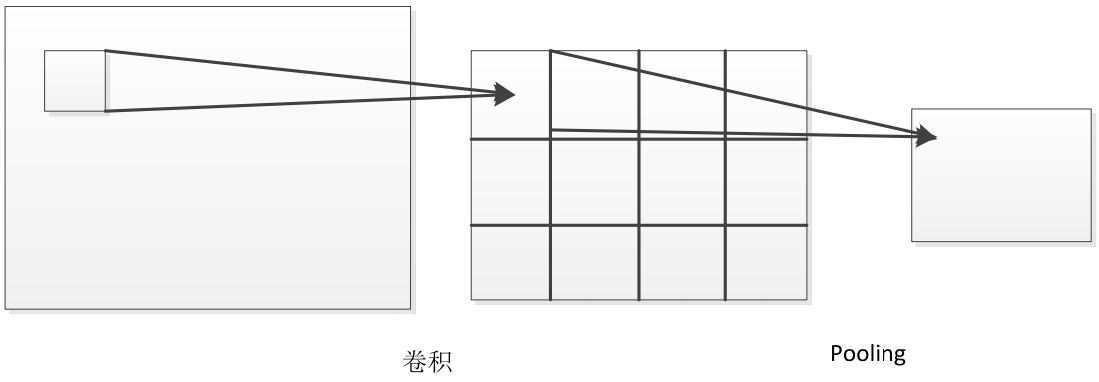



图 4- 1 卷积神经网络典型结构

卷积神经网络通过特征卷积计算，将输入信号模式分解成许多子模式（特征），然后进入分层递阶式处理。它试图将处理系统模块化，当输入信号有位移

或轻微变形的时候也能完成识别。在 CNN 中，卷积操作主要提到提取特征模式的作用，而 Pooling 则目标在于提取局部领域内最为显著的特性。因为对图像整体采用同样的卷积核进行卷积，这种方式也被称为权值共享，比起一般神经网络来能够极大的减少模型参数。

## 4.3 深度卷积神经网络

在过去，研究人员受限于计算机计算能力的匮乏，在研究图像问题时所构建的卷积神经网络都是规模比较小的模型，直到 2012 年，多层、深度的卷积神经网络在大规模图像分类问题上取得的巨大成功[27]，使得大规模卷积神经网络的研究逐步成为了计算机视觉研究中的热点问题。

在文献[27]中，作者构建了一个 8 层的深度神经网络，其中采用了卷积层、池化层，以及全连接层等多种网络层连接构建模式，得益于通用计算单元 GPU 强大的并行化处理能力，所构建的模型在 2012 年 ImageNet 大规模视觉识别竞赛的图像分类任务中以极大领先幅度获得了该任务的第一名，由此开创了深度卷积网络研究的热潮。

近三年来，许多优秀的深度卷积模型相继被提出并不断刷新在大规模视觉识别任务上的性能。其中代表性的工作有，牛津大学 VGG 小组提出的非常深卷积网络（Very Deep Convolutional Network），其并没有对卷积网络本身进行太大革新，只是在文献[27]模型的基础上，采用了更多的网络层级，更多的网络参数，构建了一个更大规模的网络，并取得了性能上的提升。新加坡国立大学视觉研究小组提出了一种网中网（Network in Network）[76]的多层卷积网络模型，其创造性了改变了一般卷积网络中卷积计算操作之后进行池化或者继续进行卷积计算的方式，在卷积层之后加入了一个小型的多层感知器模型，其出发点在于池化和卷积操作只是进行了一些线性的变换，而在卷积之后采用多层感知器映射，则能够引入更多的非线性因素，从而提高模型的表达能力。

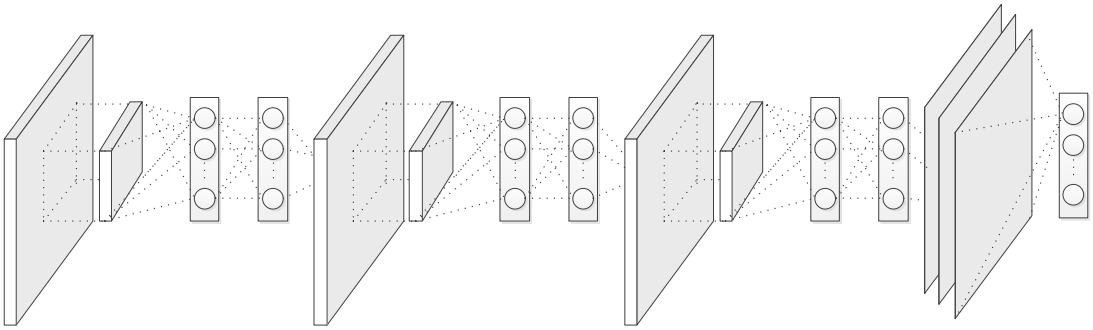


图 4- 2 Network in Netwrok 模型

Google 公司在 2014 年提出了 GoogLeNet 模型[77]，受启发于计算神经学科研究中的“赫布定律”（Hebbian Rule）[78]，该模型中提出了一种全新层级模块，称为 Inception 模块。赫布定律认为，对于神经细胞的突触，反射活动的持续与反复会导致神经元稳定性的持续提升，或常被描述为“一起发射的神经元连在一起（Cells that fire together, wire together）”。该理论常被用于解释联合学习中，对神经元的重复刺激会增强突触连接性的现象，受此启发，Inception 模块中对输入层进行了多尺度卷积池化的操作，具体如图（4-3）所示，可以看到其模型体现了对输入信号的多尺度信息提取的特点。

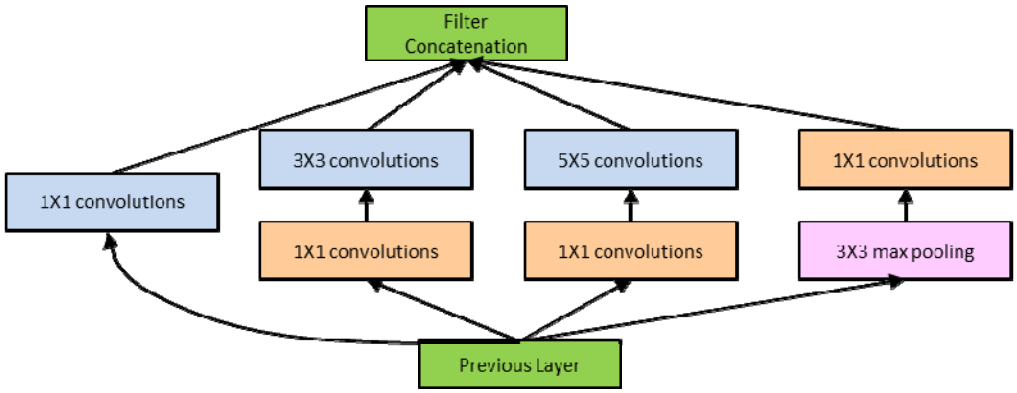



图 4- 3 GoogLeNet 中的 Inception 模块

## 4.4 视觉行人检测问题

本章主要基于深度卷积神经网络研究视觉图像中的行人检测问题，在本节中首先对相关的研究方法以范式进行介绍。

行人检测是计算机视觉研究领域的热点问题之一，经过近 10 多年的发展，许多优秀的算法框架相继被提出。这其中占主导地位的是手工设计特征加滑窗扫描检测的方式，通过采用固定尺寸窗口对图像进行多尺度、穷举式扫描来寻找行人目标，对于每个扫描窗口，首先提取所定义的图像特征，再进一步导入分类器进行判别。Haar-Like 特征[79]，梯度统计直方图特征（Histogram of Gradient, HOG）[80]是最早成功应用到行人检测上的图像特征描述子，而后积分通道特征（Integral Channel Features）[81]、局部二值模式（Local Binary Pattern，LBP）[82]、协方差描述子（Covariance Descriptor）[83]、局部块形变模型（Deformable Part based Model，DPM）[84]等进一步提升了行人检测算法的性能。

以上提到的图像特征描述子都是人工设计的，随着近年来深度学习的发展，通过机器学习而非人工设计的方式去获取图像特征愈发受到重视。在文献[85]中，作者采用了两层卷积神经网络和稀疏编码（Sparse Coding）的方式学习得到图像的描述子，而文献[86]的作者，利用 RBM 对图像的相互遮挡关系进行建模来提升遮挡情况下算法的检测精度，他们进一步的在文献[87]中提出了一种基于深度学习的联合学习框架，对行人检测中的难点如特征提取、形变、遮挡和分类器选择等进行了统一的建模。

行人检测问题中特征提取方式由手工设计到自动学习的转变与计算机视觉领域特征描述子的发展潮流是一致的，随着深度学习和非监督特征学习的兴起，通过深度卷积神经网络进行图像特征提取的方式由于其优异的性能，逐步的取代了传统手工特征描述子。文献[88]的研究发现，在图像分类任务中训练得到的深度卷积神经网络，将其作为黑盒的图像特征描述子，在其他视觉研究任务依然能取得十分出色的性能，表明可以将深度卷积神经网络作为一般的特征提取方法使用。

深度卷积神经网络作为特征提取方式能带来极大的性能提升，但其对计算资源的需求限制了进一步的拓展，尽管采用 GPU 并行化计算能显著的提高运行速度，但依然无法满足视觉目标检测任务的需求，由此也带来了视觉目标检测范式的改变。前面提到，在行人检测问题中，过去最常使用的算法框架是固定

尺寸窗口加多尺度图像滑窗扫描，这种穷举式的检索方法能够保证算法能够扫描到几乎所有的目标，但也限制了对复杂图像特征的使用。伴随着深度卷积神经网络的兴起，区域识别的目标检测范式（Recognition by Region Paradigm）也被提出来[89]，其革新在于将目标检测过程分解成了两段式的过程，首先生成一些可能包含目标的候选窗口，进而对这些窗口利用深度卷积神经网络进行特征提取和分类判别。例如在文献[90]中，作者采用了基于图割的 Selective Search 方法[91]获取图像中的候选窗口，对于每个窗口采用 AlexNet[27]的结构进行特征提取并利用 SVM 进行判别。

尽管文献[85]中已经采用卷积神经网络对行人检测问题开展了研究，但其所采用的网络结构只有两层，而在图像分类任务中所采用的网络则高达 7 层，因此本章着力于研究大规模的卷积神经网络能否给行人检测问题带来进一步的性能提升。

## 4.5　基于深度卷积神经网络的行人检测算法

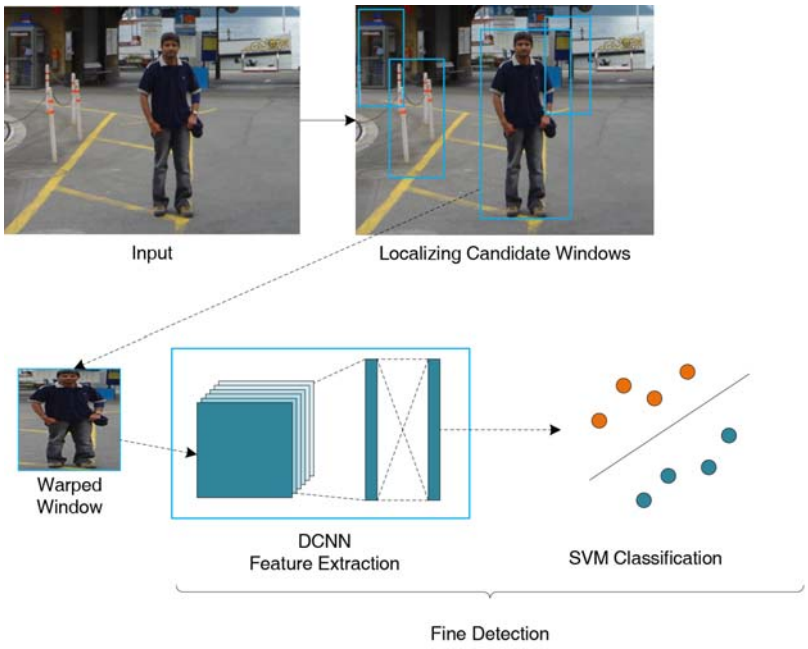



图 4- 4 行人检测算法流程图

考虑到采用深度卷积神经网络进行特征提取比较耗费时间，因此在检测方法上不能使用滑窗扫描的方式，所以本文采用了基于区域识别（Recognition by Region）的目标检测范式，所提出的行人检测算法流程如图（4-4）所示。

所提算法可以看作是两段式、由粗到精的检测过程。在粗检测阶段，利用一个比较简单的行人检测算子获得一些候选窗口，在这过程中，尽可能的保留包含行人目标的窗口而过滤掉背景窗口；在精细检测阶段，利用深度卷积神经网络对上一阶段获得的候选窗口进行特征提取，并利用 SVM 分类器进行判别，最终输出检测结果。以下将对两个阶段所采用的方法进行介绍。

### 4.5.1 粗检测阶段--候选窗口生成

在这一阶段，我们的目标是尽可能快速的过滤掉图片中的非目标窗口，同时最大程度的保留目标窗口。在文献[90]中，作者采用了 Selective Search[91]的方法来作为候选窗口的生成器。Selective Search 方法的主要流程是首先利用快速图割方法[92]将图片划分为许多个小的分割区域，这些小分割区域通常被称为超像素（Superpixel），接下来对于所有邻近的超像素，通过预定义的相似度函数来计算其相似度，然后通过一个迭代的循环过程，逐步贪心的选择所有相似度最大的超像素组合成一个新的分割区域，重新计算其与近邻的相似度，再进一步进行组合的过程。

SS 方法在通用目标检测问题中，展示出了非常好的性能，但将其使用在行人检测问题中时，还存在着一些不适用性。比如下图所展示的，蓝框代表 Ground Truth 行人窗口，红框代表 Selective Search 方法获得的与 Ground Truth 最接近的候选窗口，可以看到，尽管 Selective Search 方法获得的候选窗口覆盖了大部分的行人区域，但其并不能给出完整的矩形窗口，对于某些依赖于行人检测算法所构建的应用，比如驾驶辅助系统来说，获得完整的行人检测窗口是其系统所必需的，因此 Selective Search 方法在这些应用场合是不适用的。Selective Search 方法本质上是针对通用目标检测算法而构建的，因此其并未限定候选窗口的宽高比，而对于行人目标来说，其并没有太多姿态的变化，目标形态理论上十分

接近刚体，因此针对行人检测任务，采用刚体类的目标检测算子去进行候选窗口生成是更为合理的方案。

图 4- 5 Selective Search 方法对行人检测不适用的例子

由以上分析，在本文提出的算法框架中，采用了 Aggregate Channel Features（ACF）目标检测算子[93]作为候选窗口的生成器。ACF 检测算子采用了常用的积分通道特征加级联分类器的形式来进行构建。所谓积分通道特征指能通过图像积分图进行快速区域统计的特征，在这里，本文采用了 3 种最常用的积分特征：归一化梯度幅值、梯度统计直方图以及 LUV 颜色通道统计特征。在分类器方面，我们采用了两层级联决策树的方法，分类包含 32 和 128 个决策树分类器作为窗口的判别方法。

为比较粗检测阶段候选窗口生成的性能，本文从全局 Ground Truth 窗口覆盖率以及运行时间上对算法进行比较。全局窗口覆盖率（Cover Rate）定义为：

$$\text{Cover Rate}=\frac{\text{检出窗口与GT窗口 IoU>0.5 总数}}{\text{Ground Truth 窗口总数}}$$

其中 IoU，代表 Intersection of Units，用于衡量两个窗口之间的覆盖程度，其定义为两个窗口的并区域面积除以窗口的合区域面积，在目标检测算法中，通常设定 IoU 阈值 0.5 为正确检出。我们对几种候选窗口生成方法在 INRIA 行人检测数据集上进行了测试比较，结果如下表：

本文加入了另一种十分常用的候选窗口生成算法 Objectness[106]方法进行比较，该方法也和 ACF 一样也采用滑窗扫描的方法，但是是针对通用目标检测所训练的。从结果比较可以看到，采用 ACF 方法获取行人候选窗口，不但能够

获得最优的全局覆盖率，并且运行速度比其他方法要快一个量级。尽管 Selective Search 也能获得跟 ACF 相近的全局覆盖率，但其运行时间较慢，同时在前面分析中我们提到其不适用于刚体目标检测的问题，因此我们可以认为采用 ACF 检测算子是对行人检测问题进行候选窗口生成更为优秀的方案。

表 4- 1 候选窗口生成算法性能比较

| 方法 | Selective Search | Objectness | ACF |
|---|---|---|---|
| Cover Rate | 97.62% | 93.55% | **98.13%** |
| 运行时间 | ~4s | ~4s | **<0.5s** |

### 4.4.2 精检测阶段—候选窗口判别

在这一阶段，算法将利用深度卷积神经网络对粗检测阶段得到的候选窗口进行特征提取并进行判别，对于深度卷积神经网络，本文采用了文献[27]中所使用的 AlexNet 的网络结构设置，其结构如下图所示：

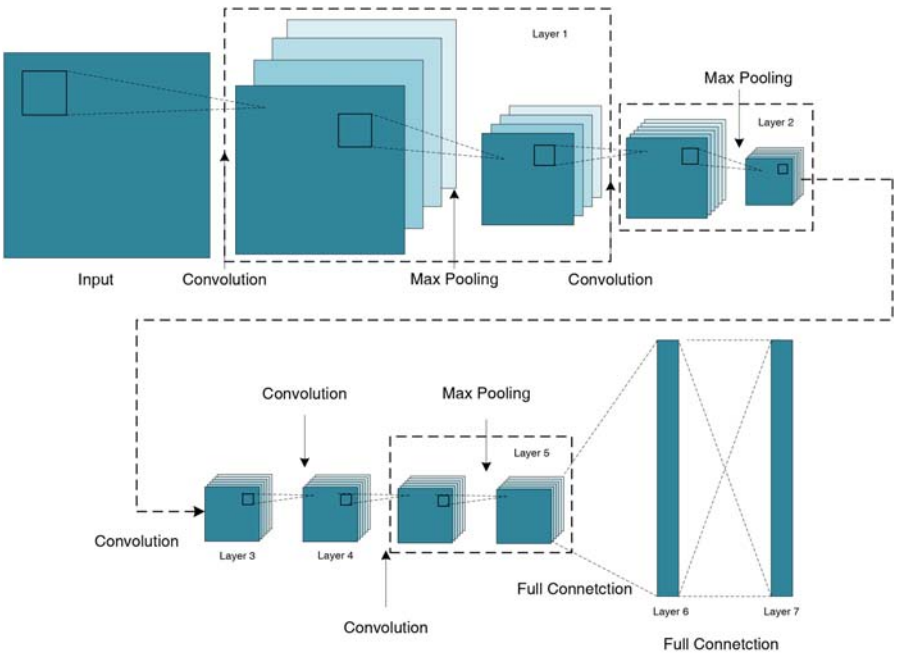



图 4- 6 AlexNet 深度卷积神经网络结构示意图

这可以看成一个 7 层的卷积网络，在第一层中，采用 96 个 11×11 大小的卷积核对图像进行卷积计算，然后采用 3×3 的网格进行 pooling；第二层采用 256 个 5×5×96 大小的卷积核进行卷积，并同样采用 3×3 网格进行 pooling；接下来的第三和第四层网络，都采用了 384 个 3×3 的卷积核进行卷积，而不进行 pooling 计算；在第五层中，使用的是 256 个 3×3 的卷积核，采用的 pooling 网格是 3×3；至此，完成所有的卷积和 pooling 操作，最后加上两个大小为 4096 维的全连接层形成最终所使用的网络。在网络中所使用的激发函数为 ReLU（Rectified Linear Unit）函数，并且在最后两层全连接层中采用了 Dropout 的策略。

对于深度卷积神经网络，由于其参数的数量非常大（比如 AlexNet 包含了 6 千万个参数），若所研究的问题训练数据较小时，就容易出现过拟合的现象。因此，在深度学习的研究中，普遍采用的做法是首先在一个较大型的图像识别任务如 ImageNet 图像识别任务中，对网络进行训练获得网络参数的初始值，然后针对自己所研究的问题进行网络参数的精调优化。在这里我们同样采用了这样的策略，同时考虑到不同的数据集对网络初始化的效果可能存在不同，所以我们对采用 PASCAL VOC 2007、2012 以及 ImageNet 数据集进行初始化的网络进行了性能比较。除了模型预训练，影响模型性能表现的另一个因素是特征层的选择，在前面网络结构的介绍中，第五层的输出 Pool5 层，以及第 6、7 层的全连接层输出 FC6、FC7 都是常用的特征层。由此，我们对预训练模型和特征层组合进行了多组实验，寻找最优的性能组合，在 INRIA 测试集上的结果如图 4-7 所示。

可以看到，在 PCASCAL VOC 数据集上预训练得到的模型表现明显的优于通过 ImageNet 数据集进行预训练的模型，本文推测其原因，可能是由于 PASCAL VOC 数据集中只包含有 20 类目标，而 ImageNet 则是 200 类，所以 PASCAL VOC 模型拥有更多的参数去对行人样本进行建模表示，由此带来更好的性能表现。对于特征层选择，FC7 层特征表现明显的优于其他两层，与其他文献研究的结论接近，表明对于目标检测任务，FC7 是更好的特征。综合以上结果，在本文的实

验中，采用了 PASCAL VOC 2007 数据集上的预训练模型以及 FC7 层特征进行行人目标检测实验。

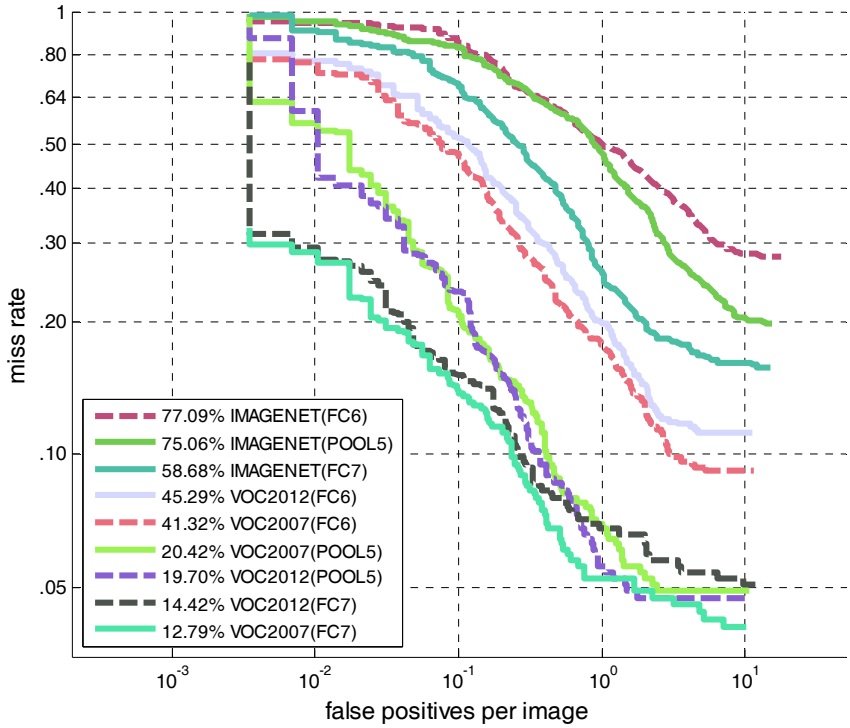



图 4- 7 模型选择组合比较结果

## 4.6 实验与结果分析

本节将所提出的行人检测算法分别在三个公开的行人检测标准数据集 INRIA、Caltech、ETH 数据集上进行实验与结果分析。

### 4.5.1 INRIA 行人检测数据集

INRIA 行人检测数据集发布于 2005 年，是最早和最为广泛采用的行人检测评测数据集，其包含有 614 幅训练图片共 1208 个行人窗口以及 288 幅测试图片。本文采用了文献[85]中修订后的标注结果进行评测，采用的标准是漏检率与 FPPI 对比曲线，漏检率代表的是测试集中未被检测的行人的比率，而 FPPI（False Positive per Image）衡量平均每幅图片误报窗口的数量，另外采用 log-平均漏检

率来衡量算法在不同 FPPI 下的整体性能[94]。作为对比，我们将所提算法的性能与一些领先算法进行了比较，所有的性能曲线绘制在图 4-8 中。

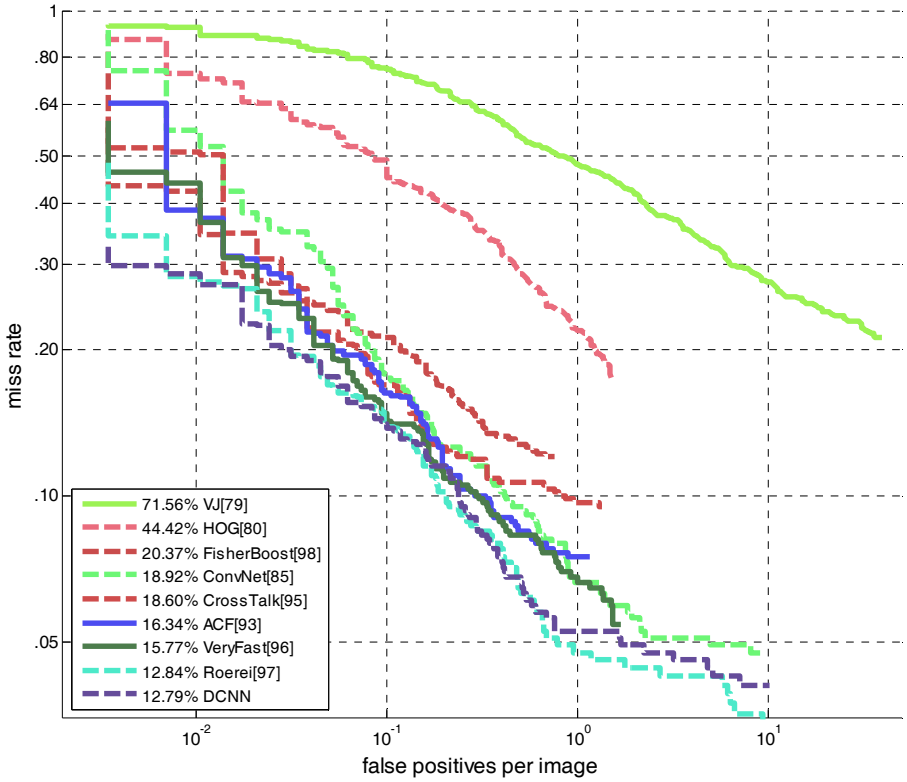



图 4- 8 不同算法在 INRIA 数据集上的性能比较

可以看到，所提出的算法在 INRIA 测试集上的表现超过了绝大多数的算法，取得了平均 12.79%的漏检率。相比于同样采用卷积神经网络的 ConvNet 算法[85]平均 18.92%的表现，算法获得了大约 30%的性能提升，ConvNet 算法采用的是两层的卷积网络，而本文采用的深度卷积神经网络为 7 层，验证了前面提到关于更多层的神经网络能否进一步带来性能提升的猜想。另一方面，所提算法是构建在 ACF 检测算法[93]基础之上的，在 INRIA 测试集上 ACF 算法取得了 16.34%的性能表现，该算法采用了 4 层的级联 boosted 分类器，每级数量分别为：32、128、512 和 2048，而我们的算法中奖 ACF 作为候选窗口的生成器，只采用了 32 和 128 两层。相比于 ACF，我们获得了 21%的性能提升，可以认为是深度卷

积神经网络对于高层的 ACF 检测子性能更加出色。

### 4.5.2 Caltech 行人检测数据集

Caltech 行人检测数据集是由加州理工大学计算机视觉实验室所发布的一个标准评测数据集[99]，与 INRIA 采用数码相机所拍摄的图片不同，Caltech 数据集是通过车载摄影系统，在市区内行驶时所拍摄下来的画面进行采集的，因此数据集相当庞大，包含有 6 万多训练样本以及 6 万余幅测试图片。由于规模以及应用上的代表性，Caltech 数据集已经成为当今最为常用的行人检测标准数据集。一般的为减少评测时间，采用该数据集时通常只采用一部分的测试图片子集进行评测，另一方面，Caltech 数据集中行人样本的窗口大小尺度变化十分大，因此在评测上也对不同窗口大小尺度的性能进行分开比较。在 Caltech 数据集上，我们对“Reasonable”和“Large”两个子集进行了评测，与主流算法的性能比较曲线如下图所示：

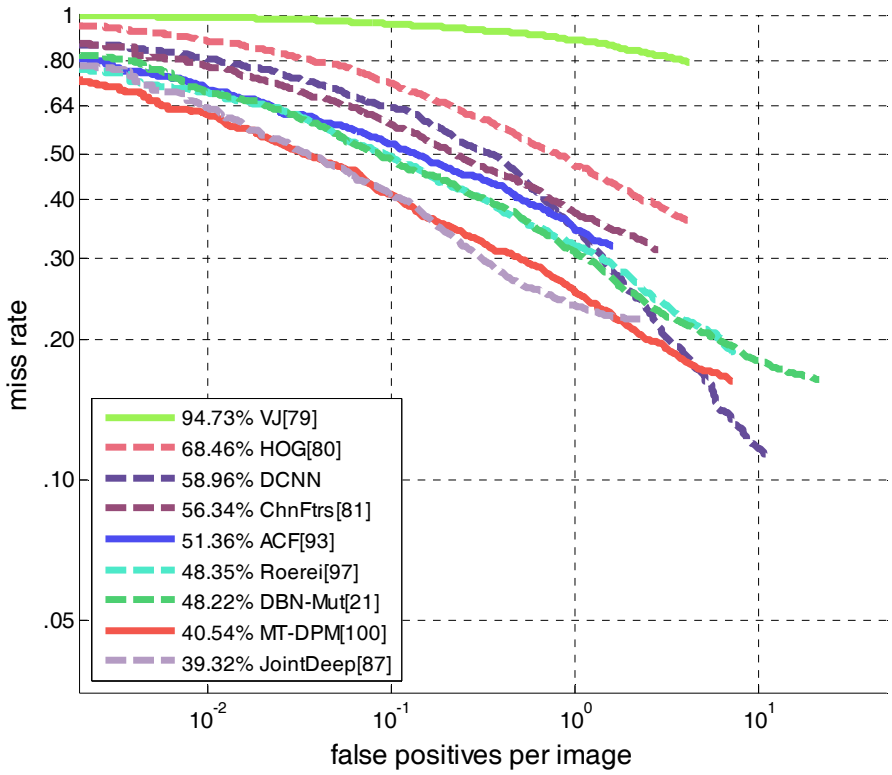



图 4- 9 所提算法在 Caltech 数据集“Reasonable”子集上的性能比较

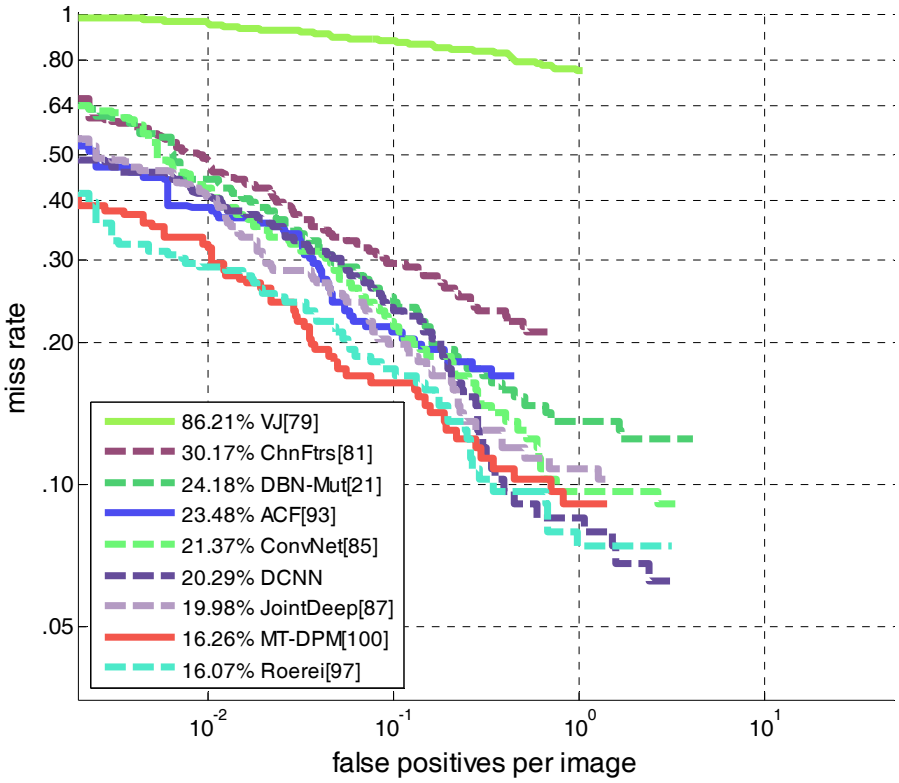



图 4- 10 所提算法在 Caltech 数据集上“Large”子集上的性能比较

可以看到，所提算法在 Caltech 数据集上并未像在 INRIA 数据集上体现出性能优势，在“Reasonable”集合上表现与其他算法相比存在较大差距，甚至不如 ACF 算法，但在“Large”集合中，所提算法性能还是具有一定的可比性，与领先算法并没有存在特别大的差距，同时和最新发展出来的 JointDeep 算法[87]性能上处在同一水平。我们认为从 INRIA 到 Caltech 数据集，性能上的巨大波动主要来源于数据集中图片特点的极大不同，INRIA 测试集中的图片通过数码相机进行拍摄采集，图片较为清晰、成像质量高，行人样本尺度也比较大，而 Caltech 数据集的车载摄像系统所收集到的图片较为模糊，从而一定程度上影响了所提算法的性能。从 Caltech 两个子集上的性能差异可以看到所提算法对于尺度较大的行人检测问题，性能上还是具有一定保障的，这和算法在 INRIA 数据集上的表现是一致的。另一方面，由于我们采用的深度卷积神经网络是使用了 AlexNet 的结构，该网络要求输入图像为 224*224 像素大小，因此，对于比较小的检测

窗口，需要通过插值对图像放大到要求的尺寸，而该过程会造成图像的严重失真，并且放大之后的图片会有明显的局部色块，卷积时候会难以获得合理的图像模式，这也是其在低分辨率时性能不好的原因。

### 4.5.3 ETH 行人检测数据集

ETH 行人检测数据集由瑞士苏黎世联邦理工大学计算机视觉实验室所发布，与 Caltech 数据集一样，其目的是想为构建车载辅助驾驶系统提供研究原型，不同的是 ETH 数据集是通过在一台婴儿手推车上安装的摄像头进行采集的，因为人为推车速度较慢的原因，其图片成像质量较 Caltech 数据集要更为清晰一些。ETH 数据集包含 499 幅训练图片 2388 个行人窗口和 1804 幅测试图片，我们采用了与 Caltech 数据集一样的“Reasonable”和“Large”两个子集进行评测，结果如下图所示：

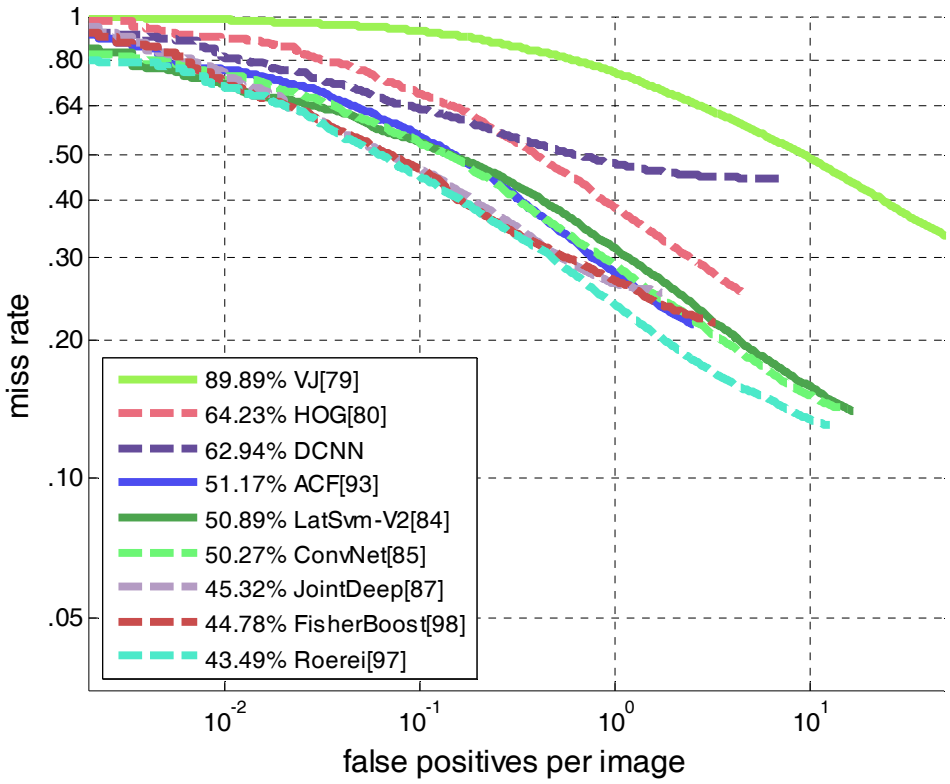



图 4- 11 所提算法在 ETH 数据集“Reasonable”子集上的性能比较

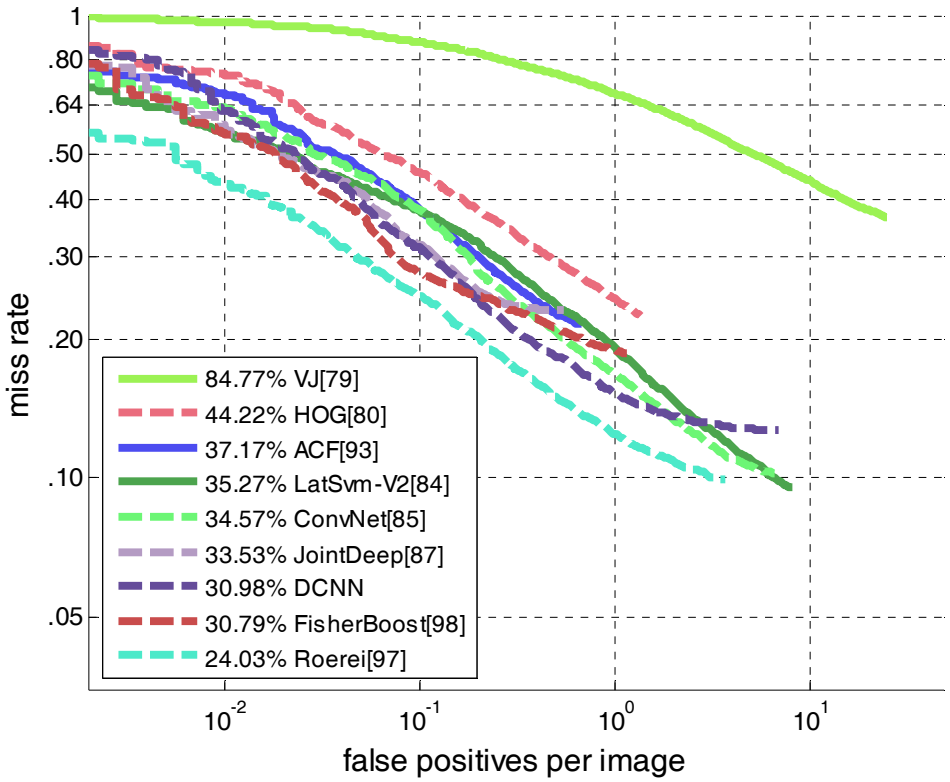



图 4- 12 所提算法在 ETH 数据集上“Large”子集上的性能比较

可以看到，在 ETH 数据集上，所提算法的性能表现趋势和 Caltech 数据集上是一致的，在多尺度条件下，性能并不出色，而在较大尺度的集合上，与主流算法的性能还是具有一定可比性。在这两个数据集上，所提算法的性能都超过了 ConvNet 算法，如在 INRIA 评测结果分析中提到的一样，更深层的卷积神经网络架构能够给算法带来更多的性能提升。

### 4.5.4　检测图片示例

以下列举了在三个测试集上一些检测结果图片，可以看到，算法对于多种姿态多个角度的行人样本都能准确检出，具有一定的鲁棒性，但同时也会受到一些干扰物的影响，比如树木、电话亭等。前面提到所提算法在 Caltech 和 ETH 数据集上表现性能并不出色与行人尺度和图片质量有一定关系，可以由例子图片中观测出来，如图 4-14 中较远处的几个行人由于尺寸过小难以检出，而 Caltech 和 ETH 的图片相比 INRIA 显得更为昏暗和模糊，而这也指出了算法进一步发展

的方向。

图 4- 13 INRIA 数据集行人检测图片示例


图 4- 14 Caltech 数据集行人检测图片示例

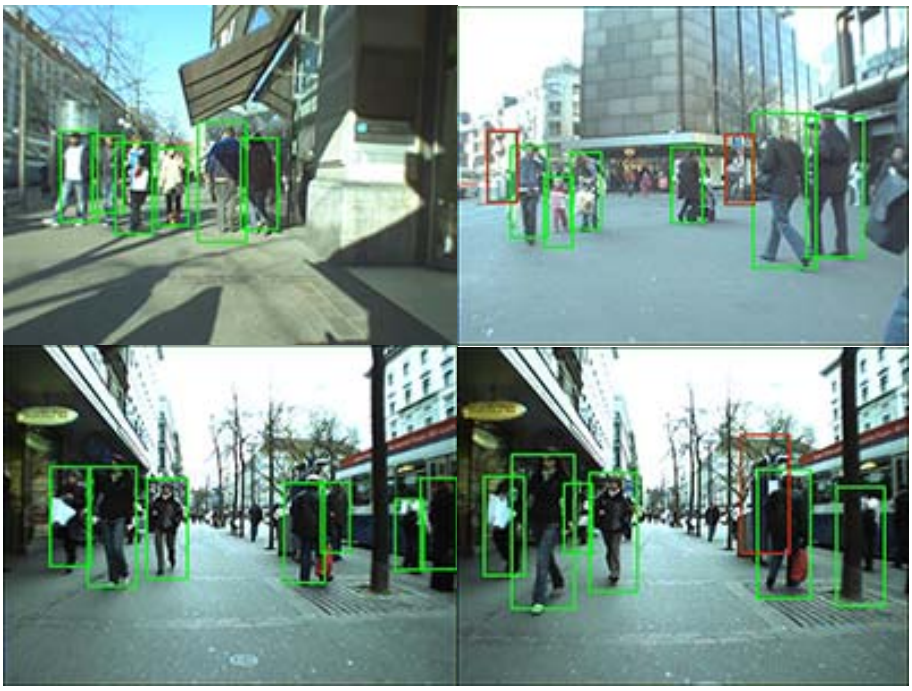


图 4- 15 ETH 数据集行人检测图片示例

## 4.7　本章小结

本章利用深度卷积神经网络对于视觉行人检测问题进行了研究，针对深度卷积神经网络特征提取较慢的问题从而无法使用滑窗扫描进行目标检测的问题，我们采用了 Recognition by Region 的目标检测范式，构建了一个两阶段由粗到精的行人检测算法。在粗检测阶段，我们采用一个快速的 ACF 检测算子来进行候选窗口生成，目的是尽可能的过滤掉反例窗口并保留正例窗口；在精检测阶段，我们采用了一个 7 层的深度卷积神经网络作为特征提取方法对候选窗口进行特征提取，并采用 SVM 进行分类判别。在三个主流的行人检测数据集 INRIA、Caltech 和 ETH 上，对所提算法与主流形人检测算法进行了详细的性能比较，并对结果进行了深入分析，所提算法在较大尺度的行人检测问题上取得了优异的性能表现，但在小尺度和低图像质量条件下表现一般，这也指出了其未来进一步的发展方向。

# 第五章　深度卷积神经网络中的因素解析特性研究

## 5.1　引言

在第三章的研究中，本文对深度学习算法中的自动编码器的流形性质进行了详细分析，以其为基础构建了一种局部结构保持映射函数作为投影降维的手段，并对其在因素解析方面的性质进行了阐述。深度卷积神经网络作为特征表示学习的重要方法，其特征表示方面的特性值得深入研究于探讨。在本章中，本文试图对深度卷积神经网络在因素解析方面的特性进行研究。

## 5.2　相关方法介绍

近年来，对于深度学习方法中的因素解析作用，研究人员开展了广泛的研究工作。这其中最常采用的策略是隐层特征分组策略，方法的核心思想是，假定图像数据是由某些特定的因素生成的，那么基于深度学习方法，对于网络的隐藏节点进行分组，在模型训练时，使各个分组分别对某些数据变化产生响应，从而能够将数据中的变化与特定的特征组对应起来，实现一定程度的因素解析作用。

在文献[41]中，作者基于自动编码器提出了一种紧致判别分析方法（Contractive Discriminative Analysis, CDA）方法，其模型结构图示如下：

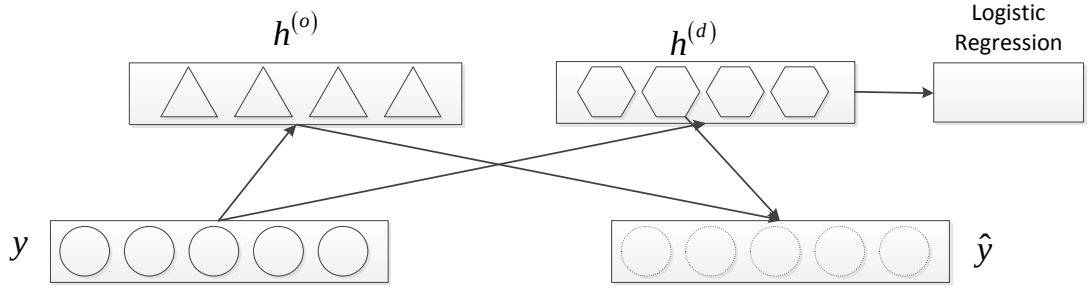



图 5- 1 CDA 算法主要框架

可以看到其将隐层节点分为了 $h^{(o)}$ 和 $h^{(d)}$ 两组，$h^{(d)}$ 组连接了一个逻辑回归分类器用于判别，$h^{(o)}$ 层则只连接到重构样本，其含义就是隐层的两组特征分别对

应于样本的生成因素以及对应于任务的判别因素，公式（5-1）为其目标函数为，公式（5-2）、（5-3）、（5-4）分别为重构误差项、紧致正则项和判别损失项。

$$\mathcal{L}_{CDA}(\theta)=\sum_{x\in\mathcal{D},y=F(x)}L_{RECON}(y,\hat{y})+\eta\mathcal{J}_{CDA}(y)+\sum_{(x,z)\in\mathcal{L},y=F(x)}L_{DISC}(z,\hat{z}) \tag{5-1}$$

$$L_{RECON}(y,\hat{y})=\left\|y-\hat{y}\right\|^2 \tag{5-2}$$

$$\mathcal{J}_{CDA}(y)=\left|\frac{\partial h^{(d)}(y)}{\partial y}\right|_F^2+\left|\frac{\partial h^{(o)}(y)}{\partial y}\right|_F^2+\gamma\sum_{i,j}\left(\frac{\partial h^{(d)}(y)}{\partial y}\cdot\frac{\partial h^{(o)}(y)}{\partial y}\right)^2 \tag{5-3}$$

$$L_{DISC}(z,\hat{z})=-\sum_{i=1}^{C}z_i\log\hat{z}_i+(1-\hat{z}_i)\log(1-\hat{z}_i),\ \ \hat{z}_i=s\left(U_ih^{(d)}(y)+a_i\right) \tag{5-4}$$

该模型在紧致正则项中对 $h^{(o)}$ 和 $h^{(d)}$ 进行了正交的约束，也就是希望两者之间分别对不相关的因素进行建模。可以看到，CDA 的特征分组是从提高判别性能角度引入约束而不是从隐层特征的解释性角度引入。

另一个同样采用分组策略的方法是文献[42]中基于 RBM 方法提出的 disRBM 方法，图 5-2 为该方法的模型结构。

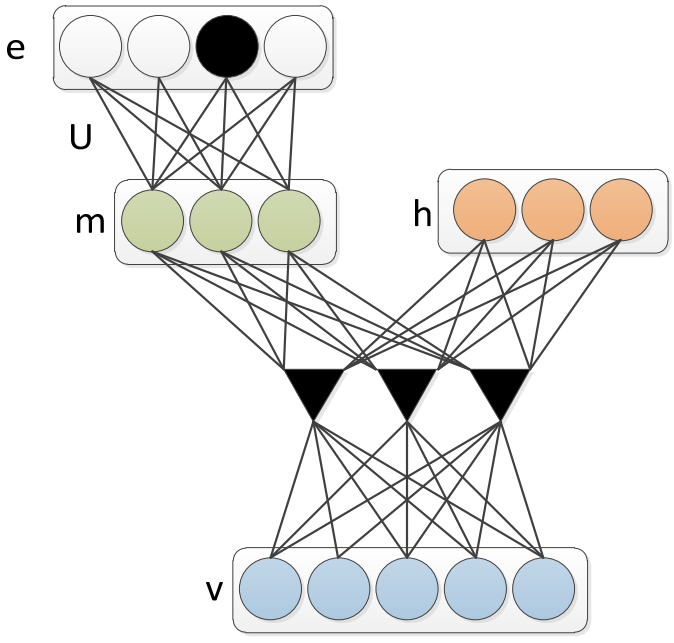



图 5- 2 disRBM 模型结构

可以看到其模型结构与 CDA 是一致的，将隐层特征进行分组并利用判别损失

对其中一组进行约束。不同的是，其通过流形解释额外引入了一组约束，认为对于相似的样本其隐层特征表示也是接近的，非相近样本其特征表示应有一定距离，数学化表示见公式（5-5），公式（5-6）是模型中添加的流形约束项。

$$\begin{aligned}\left\|h^{(1)}-h^{(2)}\right\|_2^2 &\approx 0 \quad \text{, if } (\mathrm{v}^{(1)},\mathrm{v}^{(2)})\in\mathcal{D}_{sim}\\ \left\|h^{(1)}-h^{(3)}\right\|_2^2 &\geq \beta \quad \text{, if } (\mathrm{v}^{(1)},\mathrm{v}^{(3)})\in\mathcal{D}_{dis}\end{aligned} \tag{5-5}$$

$$\left\|h^{(1)}-h^{(2)}\right\|_2^2+\max(0,\beta-\left\|h^{(1)}-h^{(3)}\right\|_2)^2 \tag{5-6}$$

这两种方法在人脸表情识别问题上都带来了性能上的提升，而且通过对隐层节点进行可视化发现，判别组的特征所表示的正好是跟表情相关的特征，因为与最终判别任务是有直接关联的，所以带来了性能的提升。同样采用特征分组策略并对各组特征进行关系建模的方法还有[38]。

以上方法对于本文的启示在于，在深度学习模型中，当隐层特征能够反映出数据变化的某种因素作用，也就是对造成数据变化的内在因素作用进行解析时，是更好的特征表示。尽管在所举例子中，对特性的因素解析建模正好是与最终任务相关的，那么如果对于隐层特征，只从因素解析的角度进行建模而不考虑与最终判别任务的相关度，能否同样获得优异的性能表现？本章尝试对这以问题进行进一步的研究。

## 5.3　深度卷积神经网络中的因素解析——特征组合方法

本章的第一个研究来自于航拍图像中的目标检测问题，在该研究中，本文的目标是检测航拍图像中感兴趣的目标如飞机和汽车等，如图 5-3、5-4 所示，从图片中可以观察到，航拍图片中存在着多个角度的飞机和汽车，旋转角度丰富是航拍图像中目标最主要的一个特点。在前文中提到，过去特征提取或学习的一个重要原则是特征不变性原则，旋转不变性是特征学习最为常见的研究出发点[32]。从不变性角度出发，研究人员希望从图像中提取到的特征对于旋转是具有不变性的，这意味着对于不同角度的样本，其应有相近的特征表示。而从因素解析角度，则希望提取的特征能够表示出旋转带来的变化。在本节的研究

中，我们发现通过后一种原则构建起来的特征模型，对于航拍图像中的目标检测性能带来了较大的提升，对上一小节中关于因素解析特征学习的疑问进行了初步的回答。

图 5- 3 航拍图像中的飞机检测问题

图 5- 4 航拍图像中的车辆检测问题

在本节研究中，我们同样采用了第四章所使用的 AlexNet 深度卷积神经网络模型作为特征提取方法。前面提到，对于 AlexNet 所提取的 POOL5、FC6 和 FC7 层特征，通常采用实验比较的方法来选择最好的特征层，在航拍图像目标检测的实验中，本文不仅对单独的特征层进行了实验比较，同时对它们的组合也进行了实验，意外的发现，经过组合之后的特征性能带来了明显的提升，结果如表 5-1 所示：

可以看到，不管是汽车检测还是飞机检测问题，通过特征组合的方式对于单独的特征性能上优势明显，值得注意的一点是，对于 4 组组合特征，拥有 POOL5

特征的组合基本上取得了性能的极限（末位的区别我们认为来源于实验中的随机扰动）。基于该实验结果，本文尝试从数据样本在特征空间的分布特性着眼，对特征组合性能上的优势进行分析与解释。以飞机样本为例，本文以 60 度为间隔将飞机样本依据角度划分为 6 组，通过 t-SNE 算法将数据样本投影到二维空间进行可视化，各个层的结果展示如图 5-5、5-6、5-7 所示：

表 5- 1 航拍图像目标检测实验结果

| 特征 | 汽车检测结果 | 飞机检测结果 |
|---|---|---|
| Hog(baseline) | 0.542 | 0.511 |
| POOL5 | 0.548 | 0.891 |
| FC6 | 0.635 | 0.832 |
| FC7 | 0.861 | 0.561 |
| FC6+FC7 | 0.921 | 0.881 |
| POOL5+FC6 | **0.945** | 0.971 |
| POOL5+FC7 | 0.941 | **0.972** |
| POOL5+FC6+FC7 | 0.942 | 0.971 |

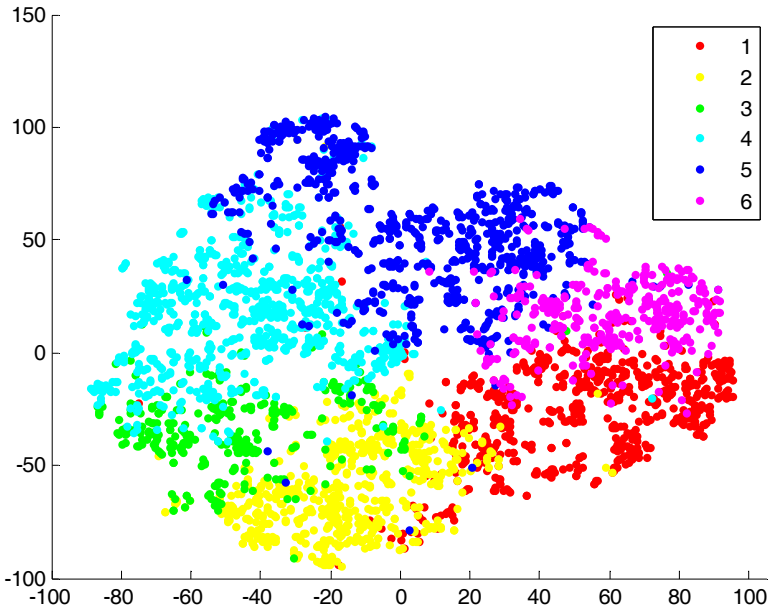



图 5- 5 飞机样本在 POOL5 层特征空间的分布

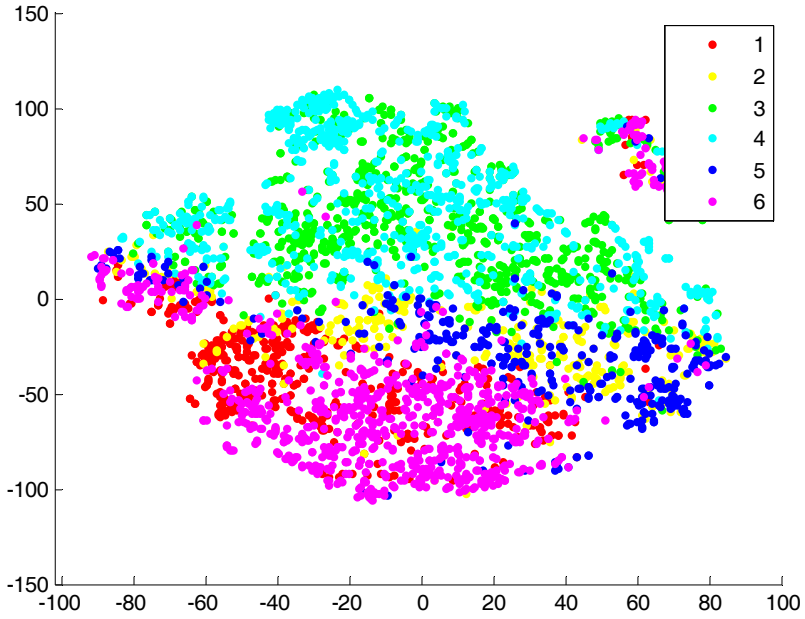



图 5- 6 飞机样本在 FC6 层特征空间的分布

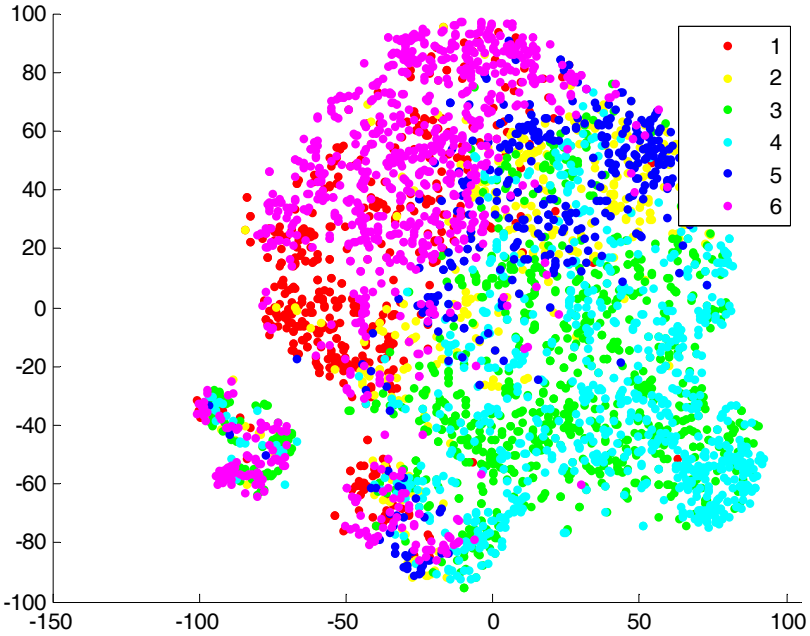



图 5- 7 飞机样本在 FC7 层特征空间的分布

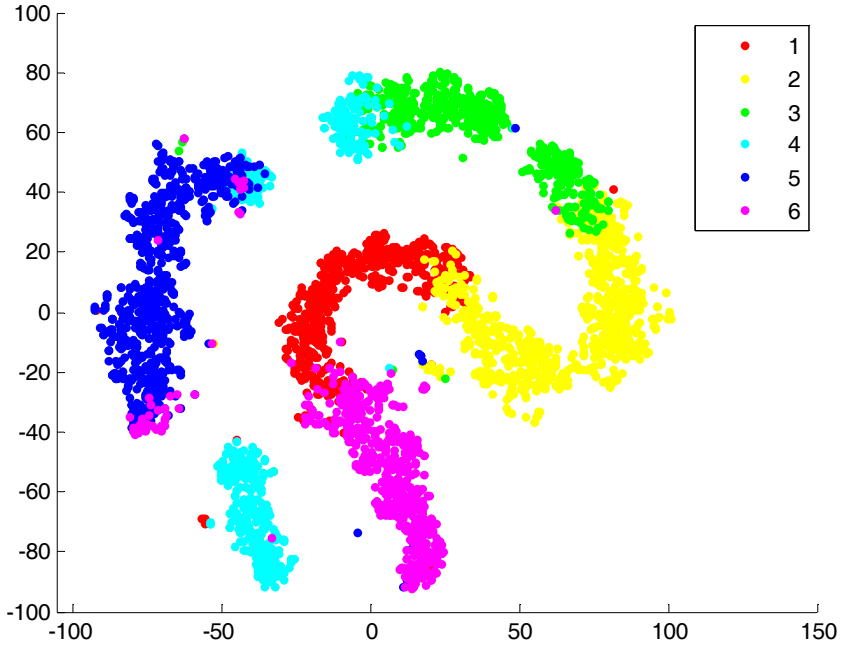



图 5- 8 飞机样本在 HOG 特征空间的分布

可以看到，数据样本在 POOL5 特征空间中表现出了十分明显的聚团特性，角度相近的样本分布较为紧密，另一方面，在整体上看，从分组 1 到分组 6 样本分布呈现出类似环状的分布，可以看作是从 0 度到 360 度角度分布的循环。基于此观察，我们可以认为 POOL5 层特征能够很好反映了飞机样本关于角度变化的分布，也即对旋转角度这个因素的作用进行了建模。对于 FC6 和 FC7 层特征所呈现出来的数据样本分布形态，并未体现出关于角度变化的建模作用。当把 POOL5 层特征和其他层特征进行组合作为新的特征时，特征中便包含有一组特征能够解析角度旋转的变化，对于另外组合的 FC6 和 FC7 特征，我们可认为其对另外的一些变化因素进行了建模。类似于 CDA 和 disRBM 算法中最终形成的图像特征，此处使用的组合特征对于数据样本内在的变化因素具有一定的解析性质；但另一方面，这种解析作用和前两种方法不一样，并不是由和判别任务相关的建模过程所引入的，只是生成于数据本身内含的特性。尽管没有从判别任务相关的角度对特征提取进行建模，但特征组合依然为判别任务的性能带

来了极大的提升，由此也验证了在因素解析特征原则下所提取的图像特征表示会对学习任务带来极大帮助。

## 5.4 特征空间规整

在前面的研究中，本文对于深度学习算法中的因素解析特性进行了初步的研究，可以看到，对于深度卷积神经网络所提取的特征，从因素解析的角度出发去构建特征提取过程，或者说从判别无关角度出发对特征提取进行约束，依然能够在判别任务中取得优异甚至领先的性能。在机器学习研究中，通常而言学习过程应是任务相关的，才能保证最终模型能够更好的完成既定任务，这也是绝大多数生成模型在判别任务中表现不如判别模型的原因。但在深度学习中，在因素解析的特征提取原则下，通过对特征空间中样本分布的某些特性进行刻画，尽管这些特性与最终判别任务无关，但提取的特征依然能够更好的完成判别任务，在本节中，我们尝试通过特征空间规整的概念对此进行总结。

对多层神经网络来说，网络参数的权值学习过程通常采用误差反向传播方式进行，也就是在网络末端设置分类器比如最为常用的是 Softmax 分类器，将判别任务的误差损失作为参数调整依据向后传播对网络参数进行更新。在过程中，网络参数的训练是直接与判别任务相关的，也就是最终由网络提取到的特征是十分具有判别性的。这是一个从输入端到输出端一体化（End-to-End）的训练过程，也是深度学习特征与手工特征最大的区别，手工特征由人为设置的方式从图像中提取特征表示，但这些特征表示并不与最终的判别任务有直接关系，因此不能保证对于判别任务的适用性。尽管从直观上而言，采用与判别任务紧密相关的特征表示是保证判别性能的最佳手段，但近年来发展起来的一些网络训练方法在判别损失的基础上，加入了对特征空间中样本之间关系的约束来对特征空间中的特性进行刻画。这其中的代表就是 Siamese Network[102]和 Triplet Network[103]。

Siamese 网络的出发点是希望在映射空间中，样本之间的欧式距离就代表着原始空间中的“语义距离”，所谓“语义距离”代表着样本之间的相似度，比如

定义为同类或者其他人为设定的相似度。假设参数为$W$的映射函数$G_W(X)$，对于输入的样本对$X_1$和$X_2$，Siamese Network 的目标就是要当样本对来自所定义的同类别时，最小化特征表示差别$E_W(X_1,X_2)=\|G_W(X_1)-G_W(X_2)\|$，反之不同类时，则极大化之。由此便可得到如下的网络结构示意：

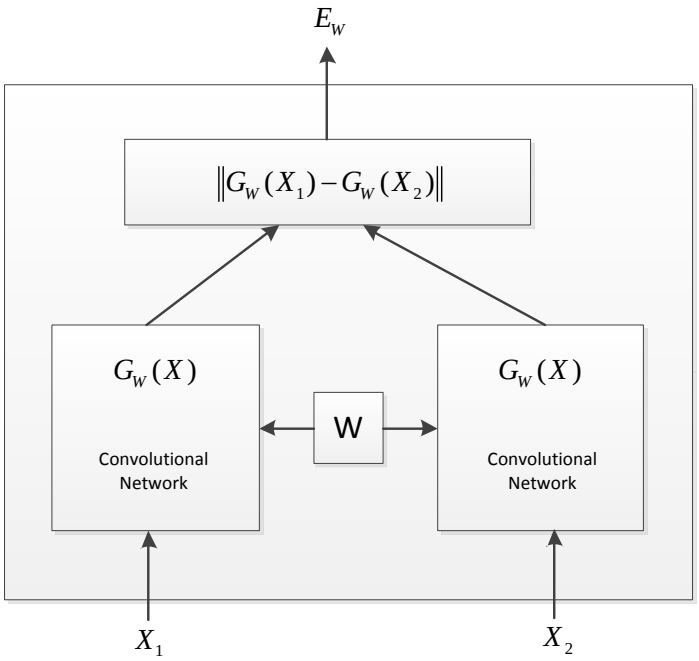



图 5- 9 Siamese 网络结构示意

可以看到这是一个二元输入的结构，注意到这个二元网络约束正是前面所介绍的 disRBM 算法中的流形解释约束。采用该约束的目的是达到同类相近、异类相远的效果，与机器学习中的 Fisher 判别准则的出发点是一致的。由于该约束在实际应用性能表现十分出色，因此在涉及到相关性、识别等研究问题中被广泛的采用。在 Siamese Network 的基础上，近来研究人员又提出了一种三元输入的网络结构 Triplet Network，其思想与 Siamese 类似，只是进一步做了拓展：对于三元数组$x$、$x_1$和$x_2$，通过给定的相似度度量$r(x,x')$，寻找一个映射空间，满足公式（5-7）所示条件，网络的结构也由二元输入拓展为了三元输入。

$$S(x,x_1)>S(x,x_2),\ \ \forall x,x_1,x_2\in \mathrm{P}\ \text{ and }\ r(x,x_1)>r(x,x_2) \tag{5-7}$$

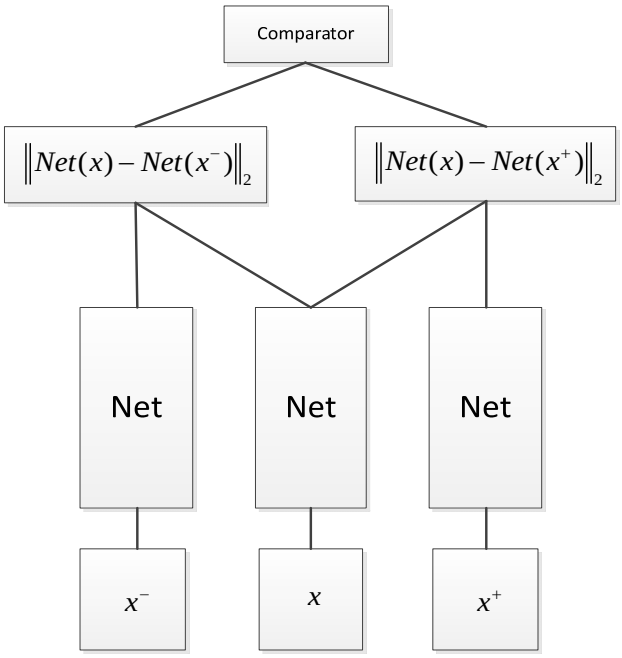



图 5- 10 Triplet 网，结构

公式（5-8）为模型所构建的损失函数为，其中 $d_+$ 和 $d_-$ 满足公式（5-9）和（5-10）。

$$Loss(d_+,d_-)=\|(d_+,d_--1)\|_2^2=const\cdot d_+^2 \tag{5-8}$$

$$d_+=\frac{e^{\|Net(x)-Net(x^+)\|_2}}{e^{\|Net(x)-Net(x^+)\|_2}+e^{\|Net(x)-Net(x^-)\|_2}} \tag{5-9}$$

$$d_-=\frac{e^{\|Net(x)-Net(x^-)\|_2}}{e^{\|Net(x)-Net(x^+)\|_2}+e^{\|Net(x)-Net(x^-)\|_2}} \tag{5-10}$$

从 Triplet Network 构建过程可以看到，其实质是寻找相似度空间 $r(x,x')$ 与特征空间 $S(x,x')$ 的一个映射函数 $Net(x)$，而这个映射对于数据样本之间的度量关系结构是保持的，这一点与第三章研究中关于流形学习的形式化表示是一致的。

对于 Siamese 和 Triplet Network，其建模过程不仅关注异类样本之间的度量关系，同时考虑到同类样本之间的度量关系，前者是与判别任务直接相关的，而后者则可视为对于数据样本在特征空间之中的分布进行了进一步的分布特性建模。这一对分布特性进行建模的过程我们称之为特征空间的规整过程，规整

的意思在于使样本在特征空间的中分布具有一定的有序性。回忆第三章引言中提到的人脸数据样本在因素解析特征提取原则下的例子，当样本的空间分布符合因素的作用时，可以认为样本在因素作用的方向上呈现出有序化的分布，也即存在某个流形子空间使样本具有近似属性值。受以上观察与分析启发，本文认为对深度卷积神经网络所提取到的图像特征，进行一定的空间分布特性约束，加入因素解析原则，能够为判别任务性能带来提升。

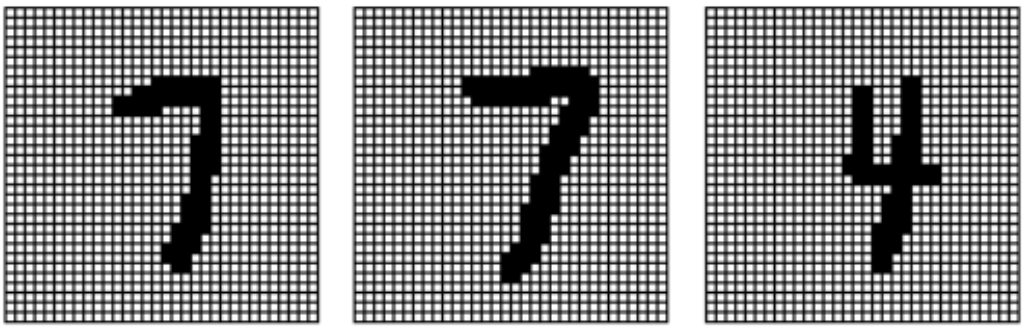
在采用 Triplet Network 结构的研究工作中，通常采用类别属性作为数据样本空间几何结构的引入点，其实质依然是添加了判别任务相关的约束。在第三章的研究工作中，我们基于数据分布的流形假设，引入了因素解析的分析手段，是一个非监督的学习过程，在接下来的研究工作中，本文将其推广到深度卷积神经网络结构的学习当中。

### 5.4.1　图像度量

在第三章的研究中，对于原始空间中的数据，本文采用了欧式距离作为数据样本相似度的度量。在计算机视觉的研究中，欧式距离由于简便性，是最为广泛应用的距离度量。对于两幅宽为$W$、高为$H$的 RGB 图像$x$和$y$，其欧式距离的表达式为：

$$d(x,y)=\sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{c\in\{\mathrm{R,G,B}\}}\left(x_{ij}^{c}-y_{ij}^{c}\right)^{2}.$$

但对于图像数据而言，欧氏距离的定义容易受到噪声的干扰，如图像的位移扰动，各种信号噪声等。例如对以下三幅手写字符图像而言，



采用的距离度量应该使两幅数字 7 的图片更接近，而数字 7 和数字 4 之间

的距离应较远，但采用欧式距离进行计算时，（1）和（2）的距离为 54，（1）和（3）的距离为 49，说明欧式距离度量对于图像原始数据并不是很好的一种选择[104]。对此，研究人员对于图像的特征表示和提取进行了广泛的研究，其目的在于通过特征提取的方式，使在特征空间中，数据样本间的距离度量能够更加准确的反应原始图像数据的相似度。在本文所构建的因素解析算法中，最重要的基础是数据间的空间几何关系度量，当将该方法拓展到复杂图像数据时，由于图像原始数据所建立的度量关系无法反应图像之间的相似度，因此需要一个中间的特征提取层对图像模式进行提取，并据此建立度量结构。

### 5.4.2 重训网络

在前文的研究中，对于由于局部扰动以及噪声所引起的图像形变，深度卷积神经网络是一种优秀的特征提取方法，通过卷积和下采样操作，能够有效的处理局部形变以及噪声问题。对于 AlexNet 网络结构，我们可将其划分成两个部分，第一个部分是前五层，通过逐层的卷积和下采样操作，提取图像模式信息，后两层全连接层，实现非线性映射。在文献[105]中，对 AlexNet 的特征提取过程进行了可视化研究，试图理解其所提取的图像特征是否具有某些可观测的特性，本文截取其对 POOL5 层特征的可视化分析如图 5-11 所示。

其中，左边展示的是对 POOL5 层所使用的卷积核进行反卷积操作可视化之后的效果，右边展示的是对相应的卷积核响应最大的图像区域。可以看到，对于模式表现十分丰富的彩色图像，卷积神经网络通过卷积核响应能够提取出其中共通的模式信息，这也意味着，在图像卷积和下采样的响应特征空间中，可以建立有意义的局部结构分析方法。

在此，本文采用的策略是在 AlexNet 网络提取的 POOL5 层的基础上，利用局部结构保持映射算法 LSPP，重新建立最终特征映射，取代 AlexNet 中后两层的全连接层。利用该方法对第四章中的行人检测问题重新进行评测比较，试图比较通过判别式误差反向传播训练得到的两个全连接层，与通过非监督式局部结构保持映射学习得到的映射层之间，对最终判别任务性能上的差异，所取得

的算法性能比较如图 5-12、5-13、5-14 所示。

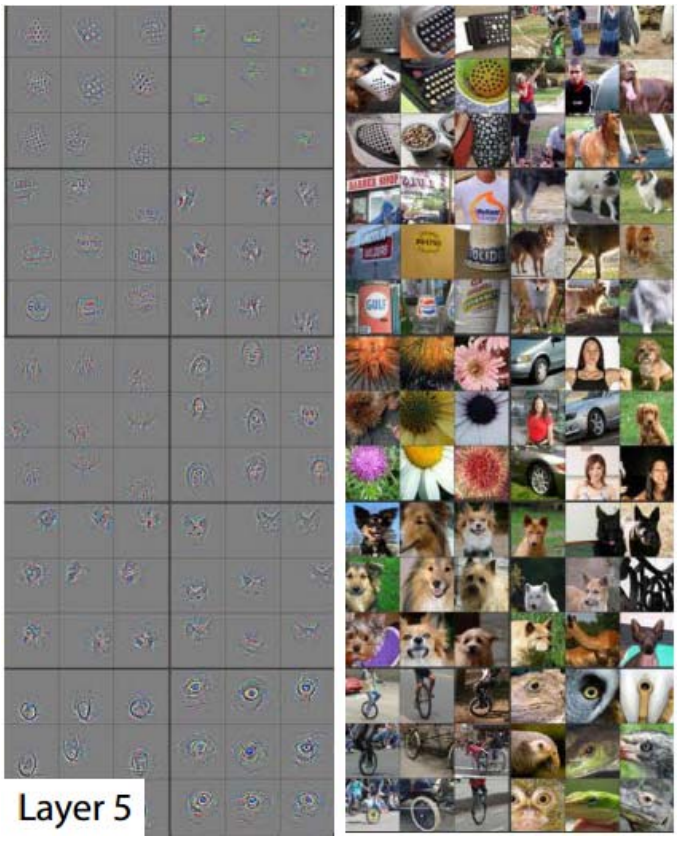



图 5- 11 文献[105]中对 POOL5 层特征的可视化研究结果

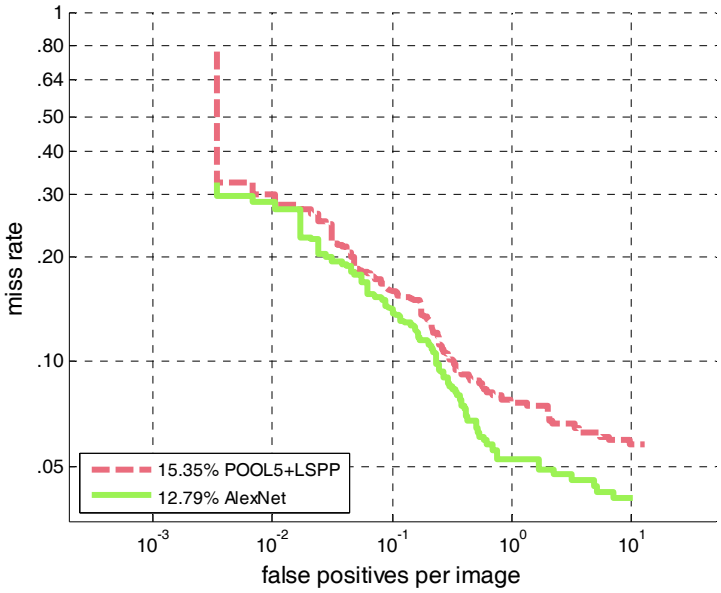



图 5- 12　重训网络在 INRIA 测试集上的结果比较

在 INRIA 测试集上，采用重训方法比起原始网络结构性能有些许下降。分析其原因，可能在于 INRIA 数据集样本数目较少，而 LSPP 算法的基础是数据的流形假设，当数据样本数量不足的时候，便很难满足流形条件，因此样本的局部空间结构估计便会存在较多误差，由此可能带来性能下降。

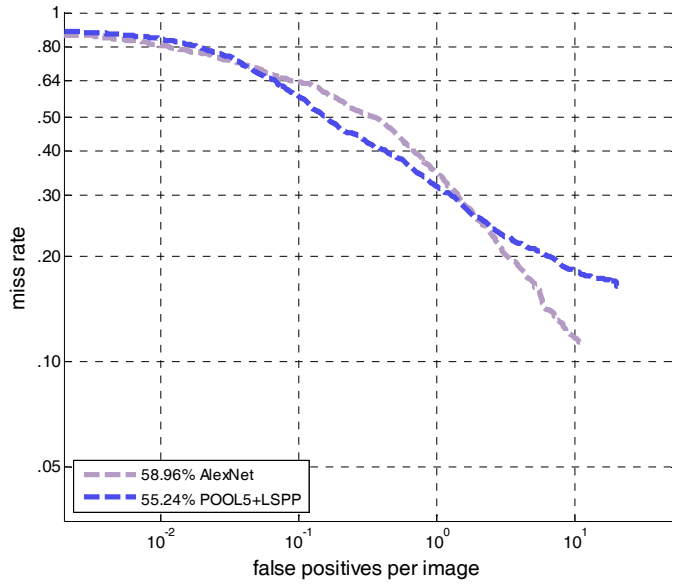



图 5- 13 重训网络在 Caltech 测试集上的结果比较

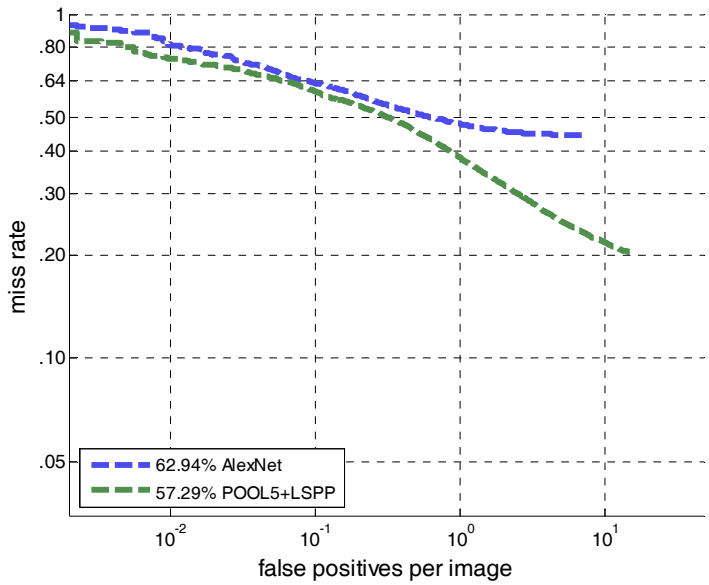



图 5- 14 重训网络在 ETH 测试集上的结果比较

在 Caltech 和 ETH 数据集上，经过重训之后的网络比原始 AlexNet 结果都得到了一定的提升，相比于 INRIA 数据集上的结果，分析认为性能提升主要来自两方面，一是 Caltech 和 ETH 数据集上训练样本更多，因此对数据样本的流形可以表示的更准确，另一方面，在第三章的研究中 LSPP 算法在处理数据噪声上有一定优势，更符合 Caltech 和 ETH 数据集本身的特性。

## 5.4　本章小结

在本章中，首先基于深度卷积神经网络在航拍图像目标检测中的实验发现，对深度学习中的因素解析特性进行了分析，分析表明，从因素解析原则出发提取到的特征表示在判别任务中获得了更加优异的性能，由此引出对深度卷积神经网络进行特征提取过程的分析；在对 Siamese 和 Triplet Network 总结的基础上，提出了特征空间规整的概念，旨在通过因素解析特性的作用，使所提取的特征在局部子空间上具有与变化因素对应的分布特性，从而在全局特征空间数据样本分布更加有序；在此猜想基础上，将第三章中所提的基于因素解析构建的 LSPP 算法应用到深度卷积神经网络上，替代全连接层的训练过程，实验表明该方法能够带来一定的算法性能提升，初步验证了所提出的特征空间规整的设想。

# 第六章　结论与展望

## 6.1　论文工作总结

本文主要对深度学习算法中的因素解析特征进行了研究，提出了一种局部结构保持算法，并在对深度卷积神经网络分析的基础上，将该算法应用到深度卷积神经网络特征提取的学习改进中。本文主要的研究成果如下：

1. 通过流形相关概念对因素解析特征提取过程进行了建模，主要思想为原始空间中相近的样本其生成因素也是相近的，通过采用流形切空间中的度量关系对样本在嵌入空间中的关系进行表示。采用深度学习中的自动编码器作为实现手段，构建了一种局部结构保持的因素解析算法。实验结果表明该算法在处理噪声数据方面有一定的优势，同时通过对样本的空间关系进行投影可视化发现最终提取得到的特征基本符合了因素解析的猜想。

2. 对深度卷积神经网络在视觉行人目标检测中的应用进行了研究，提出了一种新的基于 DCNN 的行人检测算法框架，对算法流程环节进行了详细分析与说明，对算法在三个数据集上表现出来的优点和缺点进行了分析总结，对未来进一步的深入研究指出了方向。

3. 以深度卷积神经网络在航拍目标检测中的应用结果为基础，分析了因素解析原则在采用深度卷积神经网络进行特征提取中的实现过程；提出特征空间规整的概念对几种主流网络的训练目标进行了归纳，其主要思想为使所提取的特征在局部子空间上具有与变化因素对应的分布特性，达到因素解析的目的，从而在全局特征空间中数据样本分布更加有序。将基于因素解析构建的 LSPP 算法应用到深度卷积神经网络上，替代全连接层的训练过程，实验表明该方法能够带来一定的算法性能提升，为深度学习方法中非监督条件下的因素解析特征学习提供了一种新思路。

## 6.2　未来工作展望

本文对深度学习算法中的因素解析特征开展了多方面的研究，尽管所提出的

局部结构保持映射算法在处理噪声数据上表现较为出色，但由于算法需要计算雅克比矩阵梯度等导致效率较低，这也是未来可以持续改进的地方。本文的研究表明，从因素解析角度出发对深度卷积神经网络特征进行分析是值得尝试的路线，因此，对未来工作的展望如下：

1. 本文的研究表明特征映射的切空间可以用于刻画数据流形的一些几何性质，进而这些性质可以与因素解析的概念串接起来，未来可以对非参数映射情况下数据切空间的性质进行更深入的研究，寻找一般性的切空间表示以及分析手段，同时提高算法的计算效率。

2. 本文尝试通过特征空间规整的概念将 Siamese 和 Triplet Network 进行总结并与因素解析特征提取原则进行联系，但局限在分析层面，未来可以考虑从概率生成模型角度出来，提出新的理论框架，将因素解析、非监督学习、特征空间上数据分布的特性在同一个理论框架下进行阐述。

# 参 考 文 献

[1] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks[J]. *Science*, 2006, 313(5786):504–507.

[2] Y. Bengio. Learning deep architectures for AI[J]. *Foundations and Trends in Machine Learning*, 2009, 2(1): 1-127.

[3] Y. Bengio, Deep Learning of Representations: Looking Forward[J]. *Lecture Notes in Computer Science*, 2013, 7978:1-37.

[4] Pat Langley, Herbert A. Simon, and Gary L. Bradshaw, Heuristics for empirical discovery[M]. *Computational models of learning*. Springer Berlin Heidelberg, 1987: 21-54.

[5] Ljupco Todorovski and Saso Dzeroski, Declarative bias in equation discovery[C]. *In: Proceeding of IEEE International Conference on Machine Learning*, 1997.

[6] Schmidt M, Lipson H, Distilling free-form natural laws from experimental data[J]. *Science* 2009, 324(5923):81-85.

[7] G. E. Hinton and R. S. Zemel, Autoencoders, minimum description length, and Helmholtz free energy[J]. *Advances in neural information processing systems*, 1994: 3-3.

[8] Yoshua Bengio, Olivier Delalleau, On the expressive power of deep architectures, algorithmic learning theory[J]. *Lecture Notes in Computer Science*, 2011(6925):18-36.

[9] R. Salakhutdinov, Learning deep generative models[D]. Ph.D. thesis, University of Toronto, 2009.

[10] Bishop, C. M. Pattern recognition and machine learning[M]. New York: springer, 2006.

[11] D. Plaut and G. E. Hinton, Learning sets of filters using back-propagation[J]. *Computer Speech and Language*, 1987(2):35–61.

[12] Horn, Roger A. and Johnson, Charles R, Matrix analysis[M]. Cambridge university press, 2012.

[13] B. A. Olshausen and D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images[J]. *Nature*, 1996(381):607–609.

[14] Laurent Dinh, David Krueger, Yoshua Bengio, NICE: Non-linear Independent Components Estimation [EB/OL]. *arXiv preprint arXiv:1410.8516*, 2014.

[15] Kai Yu, Tong Zhang, and Yihong Gong, Nonlinear learning using local coordinate coding[C]. *Advances in Neural Information Processing System*, 2009: 2223-2231.

[16] S. T. Roweis and L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding [J]. *Science*, 2000(290):2323–2326.

[17] J. Tenenbaum, V. Silva, and J. Langford, A global geometric framework for nonlinear dimensionality reduction [J]. *Science*, 2000, 290(22): 2319-2323.

[18] T.F. Cox and M.A.A. Cox, Multidimensional Scaling [M]. Chapman and Hall, 2001.

[19] D.L. Donoho and C. Grimes, Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data[J]. *Proceeding Nat'l Academy of Sciences*, 2003, 100(10):5591-5596.

[20] M. Belkin and P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering[C]. *Advances in Neural Information Processing Systems*, 2001(14): 585-591.

[21] Z. Zhang and H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment[J]. *SIAM J. Scientific Computing*, 2004, 26(1):313-338.

[22] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, Greedy layer-wise training of deep networks[C]. I*n Advances in Neural Information Processing Systems*, 2007, 19: 153.

[23] E. Mendelson, Introduction to mathematical logic[M]. CRC press, 2011.

[24] Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D., Dynamic pooling and unfolding recursive autoencoders for paraphrase detection[C]. *In Advances in Neural Information Processing Systems*, 2011: 801-809.

[25] Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D., Semi-supervised recursive autoencoders for predicting sentiment distributions[C]. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011: 151-161.

[26] Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeffrey Dean and Andrew Y. Ng, Building High-Level Features using Large Scale Unsupervised Learning[C]. *In: Proceedings of the Twenty-Ninth International Conference on Machine Learning*, 2013: 8595-8598.

[27] Krizhevsky, A., Sutskever, I. and Hinton, G. E., Imagenet classification with deep convolutional neural networks[C]. *Advances in Neural Information Processing System*, 2012: 1097-1105.

[28] Seide, F., Li, G., and Yu, D., Conversational speech transcription using context-dependent deep neural networks[C]. *In Proceeding of Interspeech*, 2011(33):437–440.

[29] Dahl, G. E., Yu, D., Deng, L., and Acero, A., Context dependent pre-trained deep neural networks for large vocabulary speech recognition[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, 20(1): 33–42.

[30] Yoshua Bengio, Aaron Courville and Pascal Vincent, Representation learning: a review and new perspectives[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(8):1798-1828.

[31] D.G. Lowe, Distinctive image features from scale-invariant key points[J]. *International journal of computer vision*, 2004, 60(2): 91-110.

[32] Sohn K, Lee H., Learning invariant representations with local transformations[EB/OL]. *arXiv preprint arXiv:1206.6418*, 2012.

[33] Smolensky P., Information processing in dynamical systems: Foundations of harmony theory[J]. 1986: 194.

[34] Ranzato, M. and Hinton, G. H., Modeling pixel means and covariances using factorized third-order Boltzmann machines[C]. I*n: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010: 2551–2558.

[35] Q. Le, J. Ngiam, Z. Chen, D. J. hao Chia, P. W. Koh,and A. Ng., Tiled convolutional neural

networks[C]. *In Advances in Neural Information Processing System*, 2010: 1279-1287.

[36] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun, Learning invariant features through topographic filter maps[C]. *In Proceedings of the Computer Vision and Pattern Recognition Conference*, 2009:1605–1612.

[37] A. Courville, J. Bergstra, and Y. Bengio, Unsupervised models of images by spike-and-slab RBMs[C]. *In Proceedings International Conference on Machine Learning*, 2011: 1145-1152.

[38] Guillaume Desjardins, Aaron C. Courville, Yoshua Bengio, Disentangling factors of variation via generative entangling[EB/OL]. *arXiv preprint arXiv:1210.5474*, 2012.

[39] Goodfellow, I., Le, Q., Saxe, A., and Ng, A., Measuring invariances in deep networks[C]. *In Advances in Neural Information Processing System*, 2010: 646-654.

[40] Glorot, X., Bordes, A., and Bengio, Y., Domain adaptation for large-scale sentiment classification: a deep learning approach[C]. *In Proceedings of International Conference on Machine Learning*, 2011: 513-520.

[41] Salah Rifai, Yoshua Bengio, Aaron C. Courville, Pascal Vincent, Mehdi Mirza, Disentangling factors of variation for facial expression recognition[C]. *In proceeding of European Conference on Computer Vision*, 2012(6):808-822.

[42] Scott Reed, Kihyuk Sohn, Yuting Zhang, Honglak Lee, Learning to disentangle factors of variation with manifold interaction[C]. *In Proceedings of International Conference on Machine Learning*, 2014: 1431-1439.

[43] T. Cohen and M. Welling, Learning the Irreducible Representations of Commutative Lie Groups[EB/OL]. *arXiv preprint arXiv:1402.4437*, 2014.

[44] 赵旭安, 李群和李代数[M]. 北京: 北京师范大学出版社, 2012.

[45] Tejas D. Kulkarni, Will Whitney, Pushmeet Kohli, Joshua B. Tenenbaum, Deep convolutional inverse graphics network[EB/OL]. *arXiv preprint arXiv:1503.03167*, 2015.

[46] D. Kingma and M. Welling, Auto-encoding variational bayes[EB/OL]. *arXiv preprint arXiv:1312.6114*, 2013.

[47] Sanjeev Arora, Aditya Bhaskara, Rong Ge, Tengyu Ma, Provable bounds for learning some deep representations[C]. *In Proceedings of International Conference on Machine Learning*, 2014:584-592.

[48] A. Patel, T. Nguyen, and R. G. Baraniuk, A probabilistic theory of deep learning[EB/OL]. *arXiv preprint arXiv:1504.00641*, 2015.

[49] 梅加强, 流形与几何初步[M]. 北京: 科学出版社, 2013.

[50] 陈省身, 陈维恒, 微分几何讲义[M]. 北京: 北京大学出版社, 1983.

[51] Yoshua Bengio, Martin Monperrus, Non-local manifold tangent learning[C]. *Advances in Neural Information Processing Systems*, 2004:129-136.

[52] Teh, Y. W. and Roweis, S, Automatic alignment of local representations[C]. *Advances in Neural Information Processing Systems*, 2002: 841-848.

[53] L.J.P. van der Maaten, Learning a parametric embedding by preserving local structure[C]. *In Proceedings of International Conference on Artificial Intelligence & Statistics*, 2009:384-391.

[54] G.E. Hinton and S.T. Roweis, Stochastic Neighbor Embedding[C]. *Advances in Neural Information Processing Systems*, 2002(15): 833–840.

[55] L.J.P. van der Maaten and G.E. Hinton, Visualizing high-dimensional data using t-SNE[J]. *Journal of Machine Learning Research*, 2008(9): 2579-2605.

[56] Anders Krogh, John A. Hertz, A simple weight decay can improve generalization[C]. *Advances in Neural Information Processing Systems*, 1992:950-957.

[57] Donoho D.L., Compressed Sensing [J]. *IEEE Transaction on Information Theory*, 2006, 52(4):1289-1306.

[58] Tibshirani, R., Regression shrinkage and selection via the lasso[J]. J*ournal of the Royal Statistical Society, Series B*, 1996(58):267-288.

[59] S Rifai, P Vincent, X Muller, X Glorot, Y. Bengio, Contractive auto-encoders: explicit invariance during feature extraction[C]. *Proceedings of the Twenty-eight International Conference on Machine Learning*, 2011: 833-840.

[60] Kongming Liang, Hong Chang, Zhen Cui, Shiguang Shan, Xilin Chen, Representation learning with smooth autoencoder[C]. *The 12th Asian Conference on Computer Vision*, 2014.

[61] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, Pierre-Antoine Manzagol, Extracting and composing robust features with denoising autoencoders[C]. *In Proceedings of International Conference on Machine Learning*, 2008:1096-1103.

[62] Guillaume Alain, Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution[J]. J*ournal of Machine Learning Research*,2014,15(1): 3563-3593.

[63] Maria Angelica Cueto, Jason Morton, Bernd Sturmfels, Geometry of the restricted Boltzmann machine[M]. In Algebraic Methods in Statistics and Probability, American Mathematical Society, Contemporary Mathematics , 2010.

[64] Salah Rifai, Yann Dauphin, Pascal Vincent, Yoshua Bengio, Xavier Muller. The manifold tangent classifier[C]. *Advances in Neural Information Processing Systems*, 2011:2294-2302.

[65] Donglai Wei, Dahua Lin, John W. Fisher III, Learning deformations with parallel transport[C]. *In: Proceeding of European Conference on Computer Vision*, 2012(2):287-300.

[66] A.Singer, H.-T. Wu, Vector diffusion maps and the connection laplacian[J]. *Communications on Pure and Applied Mathematics*, 2012, 65(8):1067-1144.

[67] Salah Rifai, Grégoire Mesnil, Pascal Vincent, Xavier Muller, Yoshua Bengio, Yann Dauphin, Xavier Glorot, Higher order contractive auto-encoder[C]. *Proceedings of European Conference on Machine Learning*, 2011(2):645-660.

[68] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11):2278-2324.

[69] Variations on the MNIST digits [OB/OL].
http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/MnistVariations

[70] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley and Y. Bengio, Theano: new features and speed improvements s[EB/OL]. *arXiv preprint arXiv:1211.5590*, 2012.

[71] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley and Y. Bengio. Theano: a CPU and GPU math expression compiler[C]. *In: Proceedings of the Python for Scientific Computing Conference*, 2010, 4: 3.

[72] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors[J]. *Nature*, 1986(323): 533--536.

[73] Hubel D. H. andWiesel T. N., Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. *The Journal of Physiology*, 1962, 1(160):106-154.

[74] Kunihiko Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. *Biological Cybernetics*, 1980, 36(4):193-202.

[75] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition[EB/OL]. *arXiv preprint arXiv:1409.1556*, 2014.

[76] Min Lin, Qiang Chen, Shuicheng Yan, Network In network[C]. *In: Pdoceedings of International Conference on Learning Representations*, 2014.

[77] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going deeper with convolutions[EB/OL]. *arXiv preprint arXiv:1409.4842*, 2014.

[78] D.O. Hebb. Distinctive features of learning in the higher animal[J]. Brain mechanisms and learning, 1961: 37-46.

[79] P.Viola and M. Jones, Robust real-time object detection[J]. *International Journal of Computer Vision*, 2001, 2(57):137-154.

[80] Dalal, N, Triggs, B., Histograms of Oriented Gradients for Human Detection[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005(1): 886-893.

[81] P. Doll´ar, Z. Tu, P. Perona and S. Belongie, Integral channel features[C]. *Proceedings of British Machine Vision Conference*, 2009, 2(3): 5.

[82] Xiaoyu Wang, Tony X Han, Shuicheng Yan., An HOG-LBP Human detector with partial occlusion handling[C]. *Proceedings of IEEE International Conference on Computer Vision*, 2009: 32-39.

[83] Tuzel O., Porikli F., Meer P., Human detection via classification on riemannian manifolds[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007: 1-8.

[84] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 32(9):1627-1645.

[85] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala and Yann LeCun, Pedestrian detection with unsupervised multi-Stage feature learning[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 3626-3633.

[86] Wanli Ouyang, Xiaogang Wang, A discriminative deep model for pedestrian detection with occlusion handling[C]. I*n: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2012: 3258-3265.

[87] Wanli Ouyang, Xiaogang Wang, Joint deep learning for pedestrian detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2013: 2056-2063.

[88] J Donahue, Y Jia, O Vinyals, J Hoffman, N Zhang, E Tzeng, T Darrell, DeCAF: A deep convolutional activation feature for generic visual recognition[EB/OL]. *arXiv preprint arXiv:1310.1531*, 2013.

[89] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, Arnold W. M. Smeulders, Selective search for object recognition[J]. *International Journal of Computer Vision*, 2013, 104(2):154-171.

[90] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580-587.

[91] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, Arnold W. M. Smeulders, Segmentation as selective search for object recognition[C]. I*EEE International Conference on Computer Vision*, 2011: 1879-1886.

[92] P. Felzenszwalb, D. Huttenlocher, Efficient graph-based image segmentation[J]. *International Journal of Computer Vision*, 2004, 59(2):167-181.

[93] Doll´ar, P., Appel, R., Belongie, S., Perona, P., Fast feature pyramids for object detection[J]. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2014(36):1532 – 1545.

[94] P. Dollár, C. Wojek, B. Schiele and P. Perona, Pedestrian detection: an evaluation of the state of the art[J]. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2012, 34(4): 743 - 761.

[95] P. Doll´ar, R. Appel and W. Kienzle, Crosstalk cascades for frame-rate pedestrian detection[C]. *In: Proceeding of European Conference on Computer Vision*, 2012: 645-659.

[96] R. Benenson, Mathias M., R. Timofte, and L. Van Gool, Pedestrian detection at 100 frames per second[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 2903-2910.

[97] R. Benenson, M. Mathias, T. Tuytelaars and L. Van Gool, Seeking the strongest rigid detector[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 3666-3673.

[98] S. Paisitkriangkrai, C. Shen, A. van den Hengel Pedestrian Detection with Spatially Pooled Features and Structured Ensemble Learning[EB/OL]. *arXiv preprint arXiv:1409.5209*, 2014.

[99] P. Dollár, C. Wojek, B. Schiele and P. Perona, Pedestrian detection: a benchmark[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009: 304-311.

[100] J. Yan, X. Zhang, Z. Lei, S. Liao, S. Z. Li. Robust multi-resolution pedestrian detection in traffic scenes[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 3033-3040.

[101] M. Mathias, R. Benenson, R. Timofte, L. Van Gool, Handling occlusions with franken-classifiers[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2013: 1505-1512.

[102] Sumit Chopra, Raia Hadsell and Yann LeCun, Learning a similarity metric discriminatively with application to face verification[C]. *In: Proceedings of IEEE International Conference on*

*Computer Vision and Pattern Recognition Conference*, 2005, 1: 539-546.

[103] Elad Hoffer, Nir Ailon, Deep metric learning using triplet network[[EB/OL]. *arXiv preprint arXiv:1412.6622*, 2014.

[104] Liwei Wang, Yan Zhang, Jufu Feng, On the euclidean distance of images[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(8):1334-1339.

[105] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks[C]. *In: Proceeding of European Conference on Computer Vision*, 2014: 818-833.

[106] Alexe, B., Deselares, T. and Ferrari, V., What is an object? [C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2010: 73-80.

# 致　谢

在攻读博士的四年中，经历了许多困难，也付出了许多的努力。在即将毕业之际，对这四年中所有关心以及帮助过我的人表示衷心的感谢。

首先，感谢我的导师焦建彬教授，焦老师为人平和、治学严谨，在学业上给了我充分的自由去探索，让我能够有机会去研究自己感兴趣的课题，同时提供了许多的指导与帮助，生活上也给与了无微不至的关心，使我能够顺利的完成学业。能够走到今天，离不开恩师的教诲与帮助。

感谢叶齐祥老师在求学期间对我的帮助与支持，叶老师对课题的研究工作提出了许多非常有益的意见，使得研究工作能够系统化和具有创新性。叶老师对于学术要求严格、勤于探索、工作认真，对学生的培养投入了大量的心血，令人钦佩。

感谢模式识别与系统开发实验室的全体成员们，无论是已经毕业的师兄师姐或者仍在辛苦求学的师弟师妹们，对于一个离家千里求学的人而言，是你们常常让我感觉到亲人般的温暖。

特别感谢我的挚友郑亮和郭博涛，在我人生最困难的时候，没有你们无私的帮助，我难以跨越那些坎坷。

最后，感谢我的父亲、母亲和哥哥，你们的爱与包容是支持我的最大动力，即使是在最困难的时候，依然相信坚忍与努力便有所成就与收获。

陈孝罡

2015 年 5 月

# 在学期间发表的论文与研究成果

已发表文章目录：

1. **Xiaogang Chen**, Pengxu Wei, Wei Ke, Qixiang Ye, Jianbin Jiao, “Pedestrian Detection with Deep Convolutional Neural Network”, ACCV workshop of Deep Learning on Visual Data, 2014.
2. **Xiaogang Chen**, Qixiang Ye, Jialing Zou, Ce Li, Yanting Cui, Jianbin Jiao, “Visual trajectory analysis via Replicated Softmax-based models”, Signal, Image and Video Processing, 2014.
3. Haigang Zhu, **Xiaogang Chen**, Weiqun Dai, Kun Fu, Qixiang Ye, Jianbin Jiao, “ORIENTATION ROBUST OBJECT DETECTION IN AERIAL IMAGES USING DEEP CONVOLUTIONAL NEURAL NETWORK”, IEEE Int'l Conf. Image Processing (ICIP), 2015.
4. Jialing Zou, **Xiaogang Chen**, Pengxu Wei, Zhenjun Han, Jianbin Jiao, “A Belief Based Correlated Topic Model for Semantic Region Analysis in Far-Field Video Surveillance Systems”，Pacific-Rim Conference on Multimedia (PCM)，2013.
5. Wen Gao, **Xiaogang Chen**, Qixiang Ye, and Jianbin Jiao, “Pedestrian Detection via Part-based Topology Model”, IEEE Int'l Conf. Image Processing (ICIP), 2012.
6. Lijun Wu, **Xiaogang Chen**, Yi Peng, Qixiang Ye, Jianbin Jiao, “Fault Detection and Diagnosis based on Sparse Representation Classification (SRC)”, IEEE Int'l Conf. Robotics and Biomimetics (ICRB), 2012.
7. Yi Peng, Qixiang Ye, Jianbin Jiao, **Xiaogang Chen** and Lijun Wu,”Fault Diagnosis via Structural Support Vector Machines”, IEEE Int'l Conf. Mechatronics and Automation (ICMA), 2012.