密级:_____

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

硕士学位论文

视觉显著性物体检测辅助的深度学习研究

作者姓名：王璠

指导教师：苗军　副研究员

中国科学院计算技术研究所

学位类别：工学硕士

学科专业：计算机应用技术

研究所：中国科学院计算技术研究所

2014 年 5 月

Research on Deep Learning with Saliency Object Detection

By

Wang Fan

A Dissertation/Thesis Submitted to

University of Chinese Academy of Sciences

In partial fulfillment of the requirement

For the degree of

Master of Computer Applied Technology

Institute of Computing Technology

May, 2014

声　　明

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名：　　日期：

论文版权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名：　　导师签名：　　日期：

摘　要

随着互联网的快速发展，数据的规模和复杂程度已经成倍增长。虽然如此，数据的标注依然是十分耗费人力、物理的操作。在这种前提下，有效的半监督机器学习算法就变得越来越重要。作为新兴的机器学习算法，深度学习技术以其强大的描述能力获取了多个领域的性能桂冠，并且其半监督的能力也使得其可以适应当前大数据时代的特质，从而广受工业界与学术界的推崇。

在深度学习技术中，半监督学习是通过非监督的权值初始化与监督的参数调优组成，在本文中，我们通过加入视觉显著性信息对非监督学习部分进行优化，并在图像分类任务中进行效果的验证。首先，我们提出了两种利用视觉显著性信息进行建模的方式，其中之一是利用显著性信息作为数据先验使用，即显式对源图像与显著性信息的关系进行建模；另一个是将显著性信息当做额外的数据，并利用深度学习模型对源图像与显著性信息之间的相关性进行隐式建模。之后，我们针对两种建模方式在全连接网络上中提出了两种视觉显著性辅助的深度学习方法，即显著性前景内容重构的深度学习与显著性重构的深度学习，在 STL-10 数据库的测试中，两种方法都相对于基准方法有一定的提升，其中第二种建模方式通过对逐层的非监督参数初始化进行正则，取得了 3%的性能提升。最后，我们将该方法扩展到了卷积神经网络中，并在 STL-10 中获得了相对于基准卷积自动编码机 5%的性能提升，与当前领先的方法可比。

关键词：深度学习、显著性检测、卷积神经网络

Abstract

As the rapid growing of Internet, the amount and complexity of data has increased several times over. Nonetheless, labeled data is still quite expensive to acquire. Under such circumstance, effective semi-supervised learning methods are more and more important. Deep learning technique, as one of the most recent established and most effective machine learning methods, has own champions in various domains, and it's ability of semi-supervised learning makes it feasible for current situation and quite popular in both industrial and academic fields.

In deep learning technique, semi-supervised learning is achieved by combining unsupervised coefficient initialization and supervised fine-tuning. In this paper, we improved the unsupervised learning step by importing saliency detection. Firstly, we present two ways of modeling deep network with saliency information. One way is to use saliency detection as prior to the data, i.e. to let saliency detection to influence the data directly. In the other way, we regarded saliency information as extra data source, which was combined with image data implicitly inside the model. Secondly, we presented two methods w.r.t. the two ways of modeling, which both showed superior performance than baseline auto-encoder. Among which, the method that used saliency as extra data source to regularize all deep network layers showed about 3% improvement in STL-10 dataset. Thirdly, we expanded the second modeling method to convolutional network and got a 5% improvement to baseline model of convolutional auto-encoder. The result was comparable with the state-of-art methods.

KEY WORDS： deep learning, saliency detection, convolution network

目 录

摘 要.....I
目 录.....V
图目录.....VII
1 绪论.....1
1.1 研究背景与意义.....1
1.1.1 研究背景.....1
1.1.2 研究意义.....2
1.2 国内外发展现状与趋势.....3
1.2.1 深度学习发展现状与趋势.....3
1.2.2 常用数据库及评测方法.....5
1.3 主要研究内容.....7
1.4 本文组织结构.....8
2 全连接网络深度模型.....9
2.1 背景.....9
2.1.1 常用激活函数.....9
2.1.2 梯度消失现象.....12
2.2 相关深度学习模型.....14
2.2.1 受限玻尔兹曼机.....15
2.2.2 自动编码器.....22
2.2.3 贪婪堆叠学习.....27
2.2.4 正则约束.....29
2.2.5 L1 范式约束与稀疏约束.....30
2.3 基于显著性物体检测的深度学习.....32
2.3.1 视觉显著性物体检测与先验.....33
2.3.2 深度学习模型.....35

2.4 实验和分析 43
2.4.1 可视化（Visualization） 43
2.4.2 全连接网络模型规模 45
2.4.3 前景建模的深度学习 46
2.4.4 显著性重构建模的深度学习 49
2.5 本章小结 54
3 卷积神经网络深度模型 57
3.1 背景 57
3.1.1 卷积 58
3.1.2 Pooling 操作 61
3.2 相关深度学习模型 64
3.2.1 卷积受限玻尔兹曼机 65
3.2.2 卷积自动编码机 66
3.3 视觉显著性辅助的卷积神经网络 68
3.3.1 卷积神经网络与显著性物体检测 68
3.3.2 显著性重构的卷积神经网络 69
3.4 实验与分析 70
3.4.1 卷积网络模型规模 70
3.4.2 显著性重构的卷积网络 71
3.5 本章小结 75
4 总结 77
参考文献 79
致　谢 85
作者简介 87

图目录

图 1-1 CIFAR-10 数据库样例 6
图 1-2 STL-10 数据库样例 7
图 2-1 激活函数示意图 9
图 2-2 Sigmoid 函数响应图 10
图 2-3 Tanh 函数响应图 10
图 2-4 RectifiedLinear 函数响应图 11
图 2-5 两层神经网络示意图 12
图 2-6 加入激活函数的两层网络示意图 12
图 2-7 激活函数梯度响应示意图 13
图 2-8 全连接网络深度学习整体流程示意图 14
图 2-9 玻尔兹曼机与受限玻尔兹曼机示意图 15
图 2-10 吉布斯采样蒙特卡洛链示意图 19
图 2-11 Contrastive Divergence 采样示意图 20
图 2-12 高斯-伯努利受限玻尔兹曼机示意图 21
图 2-13 自动编码机示意图 22
图 2-14 自动编码机求解目标示意图 23
图 2-15 二范数距离示意图 24
图 2-16 交叉熵距离示意图 25
图 2-17 二范数距离在 0-1 间响应示意图 25
图 2-18 降噪自动编码机示意图 27
图 2-19 贪婪堆叠学习示意图 28
图 2-20 参数初始化对比图 29
图 2-21 原始优化目标（左）与加入 L2 正则后优化目标（右）对比图 30
图 2-22 L1 范式（左）与 L2 范式(右)约束示意图 31

图 2-23 稀疏约束下特征响应图.....32
图 2-24 显著性物体检测示意图.....33
图 2-25 CIFAR-10 显著性检测示意图.....37
图 2-26 显著性重构示意图.....38
图 2-27 显著性重构层次迭代示意图.....40
图 2-28 显著性重构双路编码示意图.....42
图 2-29 模型规模测试结果图.....46
图 2-30 前景建模深度学习结果图.....47
图 2-31 前景建模深度学习第一层特征可视化图.....48
图 2-32 前景建模深度学习第二层特征可视化图.....49
图 2-33 显著性重构建模深度学习结果图.....50
图 2-34 双路编码机主编码机特征可视化图.....51
图 2-35 显著性重构参数可视化图.....52
图 2-36 显著性重构结果图.....53
图 2-37 辅助编码机参数可视化图.....54
图 3-1 卷积神经网络结构示意图.....57
图 3-2 卷积示意图.....58
图 3-3 卷积和相关对比图.....59
图 3-4 卷积网络梯度消失鲁棒性示意图.....60
图 3-5 Pooling 带来感受野变化示意图.....62
图 3-6 平均 Pooling 示意图.....62
图 3-7 最大 Pooling 示意图.....63
图 3-8 随机 Pooling 示意图.....64
图 3-9 双路卷积自动编码机结果图.....71
图 3-10 双路卷积自动编码机特征可视化图.....72
图 3-11 重构结果可视化图.....73
图 3-12 双路自动编码机显著性重构可视化图.....74
图 3-13 测试结果对比图.....75

1 绪论

1.1 研究背景与意义

1.1.1 研究背景

随着计算机科技，尤其是互联网产业的快速发展，数据量与数据的复杂程度成倍增长。在这种环境下，仰赖人为定义规则已经变得越来越困难，而机器学习技术的使得这个问题迎刃而解。在诸多机器学习模型中，数据的描述（Representation Learning，在许多文章中也称为“特征”）占有十分重要的地位[4][21]，不同的数据描述决定了对数据不同角度的概括。通常来说，数据描述极大程度上依赖于人为操作，在解决不同问题的过程中，描述方法（也称“描述子”）的设计也占据了大部分的时间[21]。

于是如何学习数据的描述成为机器学习领域的一个重要问题。在初期有许多模型通过简单特征的选择（Feature Selection）或模型融合（Ensemble Model）来达到更好的效果[8][25][61][76]，但严格意义上来说，他们并没有进行数据描述的学习，因为所用来组合的简单特征也只是人为特征工程的一部分，并且这些融合模型的表述也很大程度上限制了最终的结果[61]。渐渐的，人们意识到在学习数据描述的过程中，由于特征复用的原因，模型的深度影响着整个描述的紧致性。在达到同样目标的前提下，浅层模型要使用更多的参数[3][4]。这也就意味着深层描述可以更有效的对数据描述进行学习，在许多出色的开创工作[33][75]的带动下，引发了学习深层描述的浪潮，也就是现在的深度学习技术。

在深度学习的初期，这项技术就在手写数字数据[47]的描述上有不俗的表现[32][33]，之后在许多优秀研究者的工作中，深度学习技术被证明在形状描述[22][33]与自然图像描述[38][39][40][41][60][71][84][85]上都有相当出色的表现。及至今日，深度学习技术几乎成功运用在了所有的计算机视觉领域，比如图像分类[13][32][71][75][84][85]、图像检索[42][52]、物体检测[57][68]、物体分割[9]、物体追踪[19][82]等等。除了图像领域外，深度学习也接连在音频[17][18]、自然语言处理[69]等领域创下纪录。总体来说，深度学

习技术在数据描述的学习上有着出色的表现，这也使得它成为当期机器学习乃至人工智能领域最为炙手可热的技术之一，引发了横跨学术界与工业界的广泛讨论与研究。

引起深度学习浪潮的除了相关研究工作的进展以外，大数据问题也起了推波助澜的作用。虽然关于大数据的定义一直含糊不清，但总体来说大数据问题集中在快速、有效地处理超大规模数据上[80]，由此也孕育了 Hadoop[7][80]、MPI[31]等集群计算框架的广泛应用。在基于大数据的学习上，深度学习技术需要大量样本的特点正好与大数据的特点相符，二者相辅相成取了一些瞩目的结果[14][44]。但另一方面，大数据虽然可以促使深度学习达到极致性能，对于解决复杂问题仍然需要建立巨大的模型，对于快速有效地测试一般需要将模型分布在多台机器[44]，或是借住异构计算[14]的帮助。在实用化的角度，如何有效加速模型计算也就成了一个重要的问题。而且对于大数据时代来说，我们已经无法凭借人力来进行所有数据的标注和整理，而只能不依赖标注或依赖网络使用者们的粗标注，也就是说大容量数据带来了不确定性。在这种情况下，有效的非监督学习算法变得越来越重要。

而另一方面，显著性检测一直以来就是计算机视觉研究中重要的一部分，是计算机科学与认知学、神经学的交叉问题。虽然显著性的研究有着很长的历史，显著性检测作为一项基础研究却鲜有在应用研究中充当角色。这一部分是显著性检测算法效果的原因，尤其在复杂图像中，显著性检测的表现不足以实用；另一部分则是显著性检测与很多视觉任务的尺度不一样所致，即显著性检测是旨在模拟人的视点分布，其基本单位是像素；而大多数视觉任务都偏向于部件或物体级别。

然而，显著性物体检测[1][23][28][53][79]的出现使得显著性的实用化成为可能，相对于原来的显著性算法，这类算法更注重于找到图像中显著的物体而不是显著的点，这样它们就成为了许多任务的一个良好的前处理算法。随着研究者工作的深入，显著性物体检测也达到了越来越好的结果，尤其一些快速鲁棒的算法[28][79]使得基于显著性检测的算法[5][86][87]获得了显著的成果。而本课题也着重基于显著性检测，研究其在视觉任务上对深度学习的影响。

1.1.2 研究意义

随着深度学习技术的蓬勃发展，深度学习在视觉任务上取得了瞩目的成绩[43][44]。

然而随着研究的深入，深度学习技术也暴漏出了显著的缺点：需要大量的样本来进行训练、训练极其耗时、测试计算复杂度较高。一些研究者从不同的角度为快速计算的问题提供了思路，比如[66][78]通过优化训练算法来加速学习，[14][44]中使用集群计算来使得大规模网络的训练成为可能等。然而，这些研究都针对于深度学习算法中的缓慢的训练步骤，而鲜有对训练后模型的计算复杂度进行优化。而对于非监督的深度学习近年来最典型的应用为谷歌的虚拟大脑项目[44]，虽然其成绩斐然，但并没有在算法上进行太多革新。

正是这些这些缺点促使着学术界和工业界在深度学习问题相互合作，同时也使得在深度学习的研究进程中，模型的复杂度越来越高[34][39][43][44]。这种情况不仅因为模型大小不一对算法的比较带来麻烦，而且也严重限制了算法的大规模实用。另一方面，当今流行的深度学习方法大多数基于监督学习的卷积神经网络，而在大数据时代下数据大多是无标注或弱标注的，所以如何最大化利用非标注样本进行无监督学习也是重要的课题之一。

在这样的前提下，我们拟使用视觉显著性物体检测的先验来进行非监督学习任务；相对于横向的算法效果对比，我们更着重于进行算法纵向对比，即对于统一大小模型下算法的效果。本工作的意义在于我们利用算法生成的视觉显著性物体检测信息来进行深度学习，也即是我们并没有利用额外标注信息为分类问题提供辅助（比如物体位置信息、物体分割信息等），为非监督深度学习提供了新的角度；另一方面，我们仅使用视觉显著性来进行非监督学习，即整体模型在监督学习时模型规模一致，在获得一定性能提升的情况下并没有增加模型规模，尤其适用于大规模非标注数据少量标注数据的学习任务。

1.2 国内外发展现状与趋势

1.2.1 深度学习发展现状与趋势

相对于传统的监督学习方法，深度学习技术起步自基于神经网络使用非监督学习初始化数据描述，其主体思路在于综合自底向上和自上而下两种学习方法，从而更好的完成整个学习任务[3][4]。一般来说，自底向上的学习通过非监督方式学习数据本身的基础结构；而自上而下的学习在自底向上学习的基础上，通过监督式学习对模型求精，完

成学习任务。在整个学习任务中，自底向上的学习任务在整个学习过程中充当了良好的模型初始化，很大程度地防止了模型在监督学习中出现的过拟合和陷入局部极小的问题[3][4]。从建模方式上来看，这类深度学习方法分为基于能量模型的方法和基于重构的方法两大类[3]。不过其二者在深层建模上的方式都趋向一致，即都为贪婪逐层学习[3][4]。

在基于能量模型的方法中，受限的玻尔兹曼机（Restricted Boltzmann Machine)是最早被讨论的模型，基于受限玻尔兹曼机的叠加而生成的深度置信网络(Deep Belief Net)[33]则成为了深度学习的开山之作。然而受限玻尔兹曼机也有其固有的问题，即它仅仅针对二值图像的建模。针对这种问题，出现了基于独立高斯假设的高斯玻尔兹曼机（Gaussian Boltzmann Machine）[10]。Ranzato 等人则发现在建模实值图像时，由于图像中的局部相似性，不应该假设输入像素间独立，于是在受限玻尔兹曼机的输入层加入了横连接[41]；而之后的研究又认为应该同时建模图像均值和方差[60]，使得实值图像的建模日趋完善。[71]则利用受限波尔兹曼机强大的表述能力，为其加入开关变量(Gated)使其对噪声免疫，在人脸识别任务上获得了良好的效果。另一方面，深度置信网络虽然完成了深层描述的建立，但学习算法建立在贪婪学习的基础上，即除最顶层外所有的层次都是有向模型，这也在一定程度上限制了模型的描述能力。而深度玻尔兹曼机[63]的出现改变了这一状况，在深度玻尔兹曼机的建模中，每一层神经元不仅接受来自底层的信息，也受上层神经元的影响，这样就使得模型的描述能力、尤其是消歧（Disambiguity）能力进一步提升；其扩展模型高斯深度玻尔兹曼机[11]针对实值图像建模做出了改动，[22]则在模型中加入部分重叠，强化了对局部的建模，使其在复杂形状建模上表现优异。

在基于重构的方法中，模型思想在于构建一组编码器和一组解码器来完成数据描述：编码器的目标在于将源数据投影至紧致表示，而解码器的目标在于从紧致表示中恢复出源数据[3][4][35]。在这一领域，最为卓越的工作为 Vicent 等人提出的去噪自动编码器(Denosing Autoencoder)[75]，其原理在于在源数据上加入噪声并让自动编码器去噪，从而防止模型退化。除了上述两种方法外，也有一些从别的角度进行深度学习的方法，比如反卷积网络（Deconvolutional Net)[84][85]，和积网络（Sum-Product Net）[20][26][58][59]等。

卷积神经网络（Convolutional Neural Net）[47]是神经网络中重要的一员，作为针对图像内容的网络类型，其最大程度上的利用了图像的局部连续性，通过卷积和汇总操作

实现平移不变形。在深度学习中，也有许多方法基于卷积神经网络进行改进，比如[38]就使得非监督学习卷积神经网络成为可能。然而，深度学习的新一个浪潮则在DropOut[34]的提出后到来，这种技术通过在训练中不停的屏蔽部分模型，来使得模型各部分的独立性最大化。在这种技术下，我们可以抛弃非监督学习阶段，直接对深层网络模型进行监督学习，其在卷积网络上的应用在图像分类中取得了卓绝的成绩[43]。自此之后，许多工作针对 DropOut 提出了改进，[77]把原来对于映射矩阵整行整列的丢弃变为随机丢弃，获得了一定的性能提升；[78]则通过高斯近似来加速了整个学习流程；[83]通过随机汇总防止在训练深度卷积网络时的过拟合现象，借此提升效果；[30]则提出了 Maxout 网络，通过最大激活单元来最大化 DropOut 的效果。

随着任务复杂度的增加，当前深度学习的发展趋势越来越偏向于大模型、大数据方向[14][44]。这在获得成功的同时，也造成了计算复杂度的大幅增加，从而使得深度学习越来越依赖于硬件，迫使许多优秀的研究工作者与工业界合作来完成研究。同时，随着[14][44]等工作的出现，大数据加非监督学习的组合也开始进行实用，成为利用大量非标注数据的典范。

1.2.2 常用数据库及评测方法

由于在本课题中，主要涉及的方向为对自然图像的建模，所以在本节描述时，我们只对几个常见的用于深度学习算法评测的自然图像数据库进行描述，并忽略像MNIST[47]这样的经典二值模式数据库。下面我们将详述 CIFAR-10 和 STL-10 这两个自然图像数据库。

CIFAR-10

CIFAR-10 数据库[40]为 Alex Krizhesky 整理的深度学习评测数据库，为 TINY 数据库[89]的一个子集。在 CIFAR-10 数据库中，共有 6 万张 32*32 的彩色图像。并且数据库分为 10 个类别（如图 1-1），分别为飞机（airplane）、机动车（automobile）、鸟（bird）、猫（cat）、鹿（deer）、狗（dog）、蛙（frog）、马（horse）、船（ship）、卡车（truck）。其中每一类图像各有 1 万张样本。

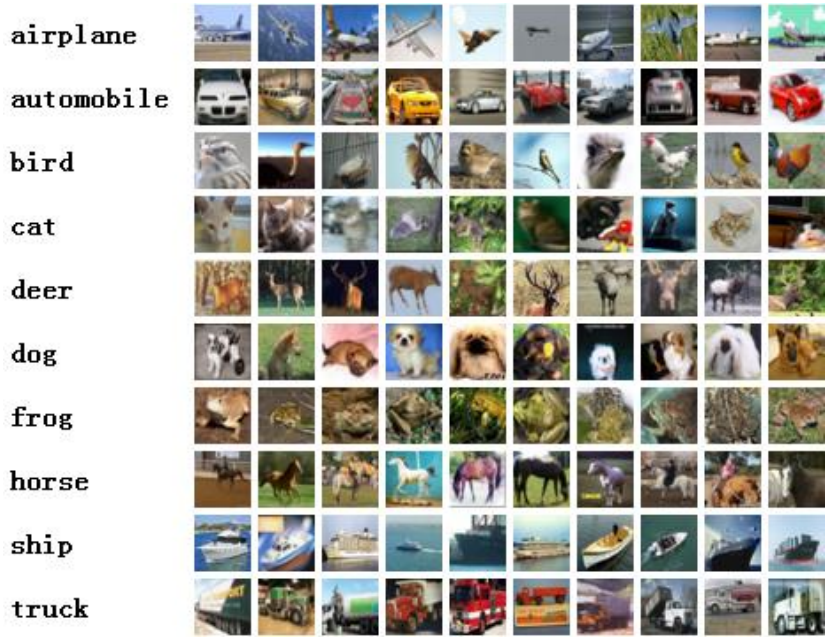


图 1-1 CIFAR-10 数据库样例

在 CIFAR-10 中，有定义好的 5 万张训练样本以及 1 万张测试样本，大量的样本与较小的图像分辨率使得其特别适合进行深度学习算法的评测，并且图像中同类物体较大的类内差异（主要表现在形状、角度、裁剪上）也使得该库成为深度学习初期较为困难的数据库。

在 CIFAR-10 上，一般的评测方法有以下三种：第一种为直接利用 32*32 的图像进行深度学习，并在测试数据上进行准确率（Accuracy）的评测；第二种为将图像随机部分裁剪为 24*24 的小图进行训练，在测试时只利用样本中心的 24*24 块的评测分类准确率；第三种与第二种比较相似，但在测试时不仅利用中心的 24*24 图像，而是利用中心、左上角、右上角、左下角、右下角与他们的水平翻转图像，共 10 张 24*24 的图像进行投票。在本文的测试中主要涉及纵向对比，具体的评测方法在每一个实验部分会做详细说明。

STL-10

STL-10 数据库[16]是 Adam Coates 受 CIFAR-10 启发收集整理的自然图像数据集。在 STL-10 中，共有 11 万三千张图像，并分为 10 类（如图 1-2），其类别的分布与 CIFAR-10 相近，只是蛙替换成了猴。在 STL-10 中，与 CIFAR-10 数据库最大的不同有两点：首先在 STL-10 中图像分辨率由 32*32 变为了 96*96，即算法可以从更好的分辨率中提取信息；其次，在 STL-10 中的训练集很小（5000 张），而多了一个巨大（10 万张图像）

的无标注集合，所以该数据库特别适合非监督的深度学习算法进行评测。

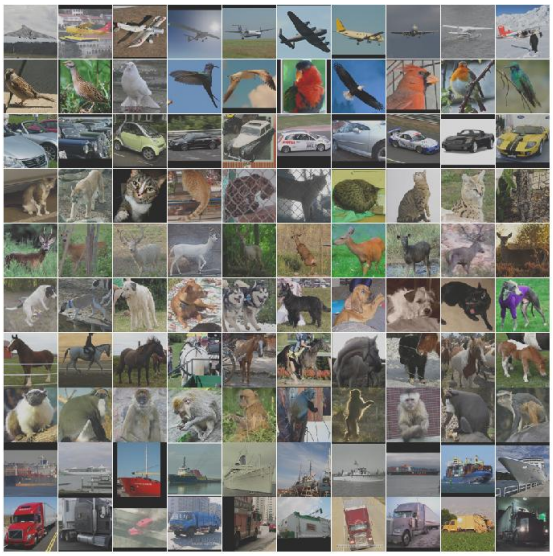


图 1-2 STL-10 数据库样例

在 STL-10 中，主要的评测方法为利用数据库提供的无标注样本进行非监督学习，之后利用训练折（Training Fold）进行训练（10 折，每折共有 1000 张训练样本），并在所有测试样本（8000 张）进行测试。在如此少的有标注训练样本下，STL-10 是一个十分困难的数据库，尤其考验模型的非监督学习能力。与 CIFAR-10 一样，具体的测试细节将在后文的实验部分进行描述。

1.3 主要研究内容

针对上文中提出的当前深度学习存在的问题，本文主要从非监督学习出发，利用视觉显著性信息来增强深度学习的效能。研究的主要内容有：

（1） 视觉显著性信息和全连接网络深度学习的融合。视觉显著性作为一个基本的计算机视觉问题，已经有较长的研究历史。然而，利用视觉显著性辅助具体视觉应用任务（如图像分类、检测）的工作却相对比较少。在这里，本文探讨了利用显著性信息进行全连接网络深度学习的方式，利用显著性信息辅助深度学习进行非监督参数初始化。我们进行了相关的概率建模，并在基准模型上获得了性能提升。

（2） 视觉显著性信息和卷积网络深度学习的融合。在深度学习中，卷积神经网络由于良好地利用了图像先验，在图像相关问题上获得了良好的效果。在这里本文在全

连接网络的基础上，将显著性信息应用在卷积网络的学习上。同样在显著性信息的辅助下，我们获得了更好的参数初始化，提升了卷积网络的性能。

在本文中，我们第一次将显著性信息用于深度学习之中。我们对视觉显著性对具体任务的影响进行了概率建模，并以此为基础进行与深度学习的融合学习。在借助显著性信息进行深度学习的过程中，我们发现在样本较少的情况下，利用显著性信息可以为深度学习模型提供良好的参数初始化信息，从而在纵向对比中获得性能提升。在横向对比中，我们与当前领先的方法可比。

1.4 本文组织结构

本文的组织结构如下：

第一章我们进行研究的背景介绍与研究意义，并较详细地描述了当前国内外深度学习的研究现状以及相关评测数据库。最后介绍了本研究的主要内容以及文章的组织结构。

第二章我们着重探讨利用显著性信息进行全连接网络深度学习的方法。首先，我们将简述全连接网络的背景知识，包括全连接深度网络遇到的问题以及当前流行的两种深度学习模型：受限玻尔兹曼机以及自动编码机；之后我们将具体探讨如何进行视觉显著性检测与深度模型的融合；最后给出相关的实验与评测结果，并对其进行分析与讨论。

第三章我们着重探讨利用显著性信息进行卷积网络深度学习的情况。作为另一种最流行的神经网络形式，卷积网络在图像应用上有着广阔的应用，但其应用多为监督学习。这里我们首先简述了卷积神经网络的建模与推导，然后提出我们融合视觉显著性的非监督学习方法，最后通过实验与评测验证了本文算法的有效性，并对相关结果进行分析与讨论。

第四章我们对全文进行总结，并讨论工作中仍存在的问题与疑惑，提出之后工作的重心。

2 全连接网络深度模型

2.1 背景

全连接网络(Full-Connected Net)是最早出现的，也是最常见的一类神经网络模型。全连接网络的主要特点为层次内无连接、层次间全连接。最早期的全连接网络，比如感知机（Perceptron）[62]，为单纯的线性分类器，即层次间为单纯的线性映射（如图 2-1 中的左图）；之后出现的全连接网络在线性映射之后加入非线性激活函数（Activation Function, 如图 2-1 中的右图），并在此基础上再增加一或多层神经元，使得使用三层或三层以上全连接网络进行非线性分类成为可能。

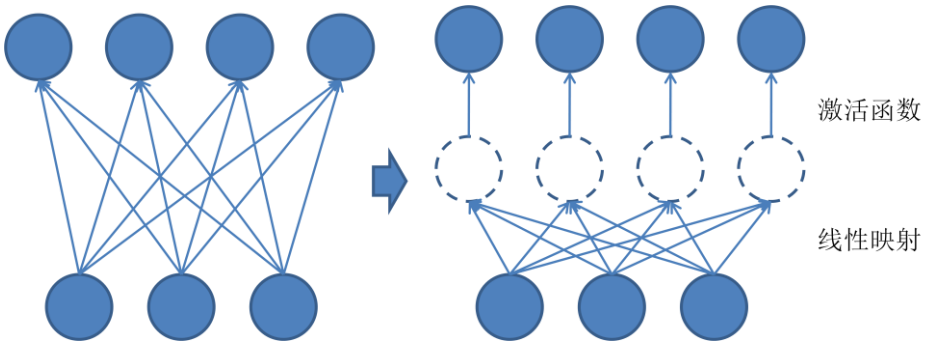


图 2-1 激活函数示意图

在全连接网络的发展中，制约全连接网络发展的原因主要为两点，即参数空间过大导致容易过拟合，以及梯度消失现象导致无法训练深层次网络。目前参数空间过大的问题已经随着计算机计算能力的大幅度提高而越来越不显著，故之后会先详细介绍常用的激活函数，并介绍随之而来的梯度消失现象。

2.1.1 常用激活函数

在神经网络中，常见的激活函数有 Sigmoid 函数、tanh 函数以及最新提出的 Rectified Linear 函数[55]，下面对不同激活函数做相关介绍。

Sigmoid 函数

Sigmoid 函数是神经网络中最常见的激活函数，他的定义域为实数域，值域为[0,1]。其函数形式为：

$$\mathrm{Sigmoid}(x) = \frac{1}{1 + \exp(-x)}$$

函数的响应为图 2-2：

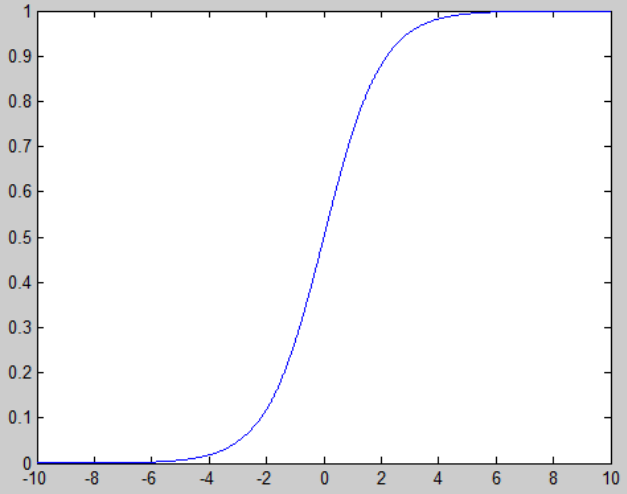


图 2-2 Sigmoid 函数响应图

其导数为：

$$\mathrm{Sigmoid}'(x) = \mathrm{Sigmoid}(x) * (1 - \mathrm{Sigmoid}(x))$$

Tanh 函数

Tanh 函数（即双曲正切函数）的性质与 Sigmoid 函数相近，它的定义域为实数域，值域为[-1, 1]。函数形式为：

$$\mathrm{Tanh}'(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

响应图为图 2-3：

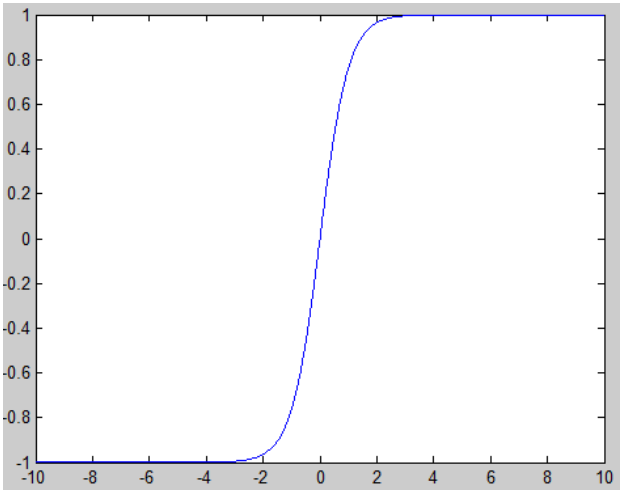


图 2-3 Tanh 函数响应图

Tanh 函数的导数为：

$$\mathrm{Tanh}'(x) = 1 - \mathrm{Tanh}^2(x)$$

Rectified Linear 函数

Rectified Linear 函数为最近兴起的一种激活函数[55]，因其形式简单、计算复杂度低、性能优异而被广为使用，该函数的定义域依然为实数域，值域为$[0, +\infty]$。与 Sigmoid 以及 Tanh 函数不同的是，Rectified Linear 函数并没有一个闭合的值域，并且其并不是一个连续可导的函数。该函数的定义为：

$$\mathrm{RectifiedLinear}(x) = \max(0, x)$$

响应图为图 2-4：

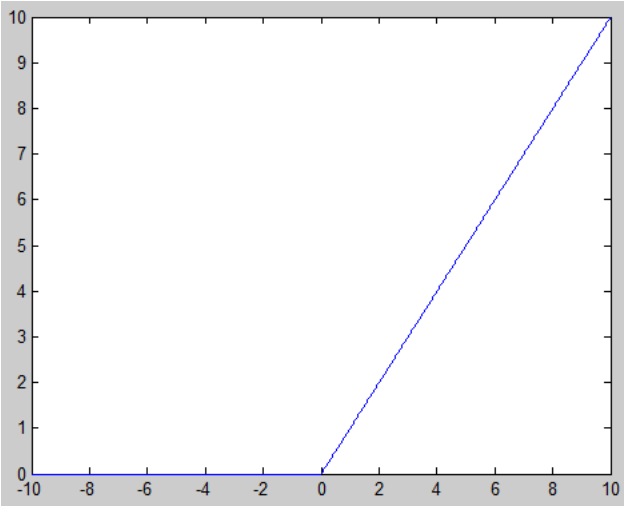


图 2-4 RectifiedLinear 函数响应图

其导数形式为：

$$\mathrm{RectifiedLinear}'(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

2.1.2 梯度消失现象

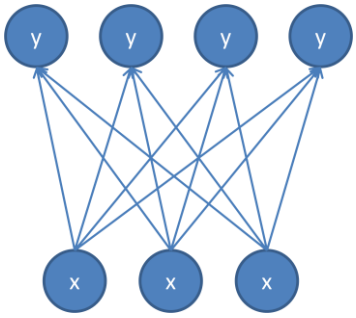


图 2-5 两层神经网络示意图

在训练深度网络时，误差反传（Back Propagation）是最常用也最有效的训练方式，假设我们有以下简单的网络（如图 2-5）。则假设我们已经知道目标函数 F 对于节点 y 的梯度，那么根据公式，节点 x 的梯度为：

$$\frac{\partial \mathrm{F}}{\partial \mathrm{x}}=\frac{\partial \mathrm{F}}{\partial \mathrm{y}}\frac{\partial \mathrm{y}}{\partial \mathrm{x}}$$

现在我们加入对激活函数的考虑，假设 σ 表示某一个激活函数，则网络图可以扩展为（图 2-6）

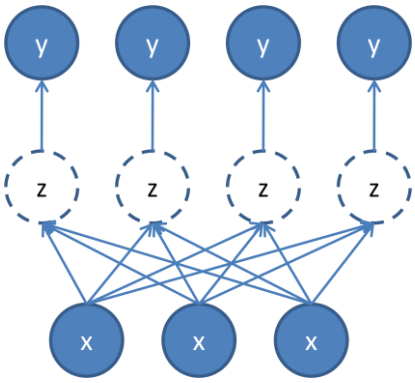


图 2-6 加入激活函数的两层网络示意图

那么在这里，我们对节点 x 的梯度则为：

$$\frac{\partial \mathrm{F}}{\partial \mathrm{x}}=\frac{\partial \mathrm{F}}{\partial \mathrm{y}}\frac{\partial \mathrm{y}}{\partial \mathrm{z}}\frac{\partial \mathrm{z}}{\partial \mathrm{x}}$$

由非线性函数的定义我们可以得到其梯度表达式，不幸的是，对于 Sigmoid 和 Tanh 函数，其梯度表达式的值域分别为[0,0.25]与[0,1]。那么意味着每一次误差传播的过程中，其梯度值都会相应损失（其中 tanh 函数的导数只在 x=0 一点上为 1，我们即也认为其值在绝大多数情况下在[0,1)范围内），即：

$$\frac{\partial \mathrm{F}}{\partial \mathrm{x}} < \frac{\partial \mathrm{F}}{\partial \mathrm{y}}$$

那么在训练深层网络时，随着层次的传播，对于靠近输入层的层次来说，梯度信息就已经在不断的传播中损失殆尽。从而导致在学习深度神经网络时，靠前层次的参数得不到有效的学习，从而严重影响了整个网络的表现。

相对于传统的 Sigmoid 和 Tanh 函数，新出现的 Rectified Linear 函数则可以很好的应对梯度消失的现象。由定义可知，Rectified Linear 函数的梯度在很大范围内都为 1，这样的话在多层次的传播中，梯度信息就可以最大化的进行保持，从而适应深层网络的学习。

除了激活函数导数值域影响误差传播意外，激活函数导数在输入数据上的响应范围也对网络训练有很大的影响。具体来说，对于每一种激活函数，其导数都只在特定的数值上有值（即不为零），而在其余部分为零。（如图 2-7）

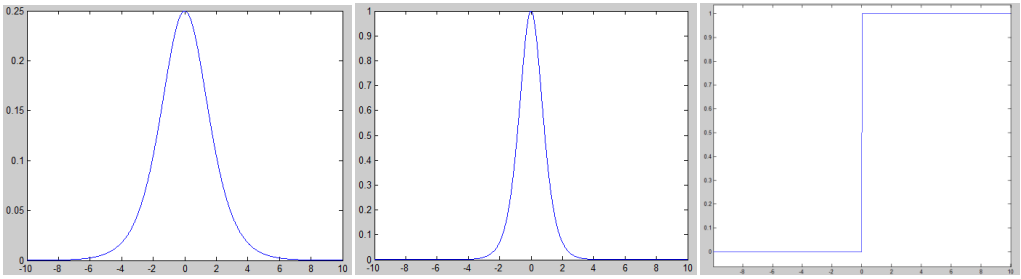


图 2-7 激活函数梯度响应示意图。左图为 Sigmoid 函数梯度示意图，中图为 Tanh 函数梯度示意图 ，右图为 RectifiedLinear 函数梯度示意图

在模型参数的学习中，非线性函数的导数形式其实是非线性函数的固有性质，但如果模型参数初始化不当，则容易造成误差在非线性激活函数的作用下没有导数，从而限制了网络的学习。同时我们很容易发现，当神经网络的层数加深时，由于要经过多层非线性函数的“筛选”，其误差梯度就越难以被保持下来。所以对于神经网络而言，参数的初始化就成为了即重要又技巧性高的工作。而从不同非线性激活函数的导数响应图上我们也容易看到，对于 Sigmoid 和 Tanh 都只有很小的范围内会有梯度响应，而对于 Recitified Linear 函数来说响应范围则为所有大于零的输入值，这样的话对于参数的初始化值 Rectified Linear 函数就获得了绝佳的鲁棒性。这也是 Rectified Linear 函数在当今被广为推崇和使用的原因之一。

为了避免以上提到的在深层网路中梯度消失问题，我们需要对参数进行良好的初始化，这一点将在下一节进行详细讨论。

2.2 相关深度学习模型

在前一节我们提到在训练深层神经网络时遇到的问题，自从误差反向传播的多层神经网络出现以后，这个问题就一直困扰着相关的研究工作者。直到 2006 年，Hinton 在他的文章中通过受限玻尔兹曼机（Restricted Boltzmann Machine）[32][33]的迭代学习才让人们得以较好的训练深层网络。

之前我们探讨了深层网络难以直接训练的原因，即参数初始化以及非线性函数导数值域带来的误差消失现象。所以在训练深层神经网络的过程中，如何初始化参数使得深层网络可以进行良好的学习就成了关键问题。在这一点上，从上文提到的 Hinton 开始，不少优秀的学者都进行了相关的研究和探讨[10][32][40][41][54][63][75]。一般说来，人们认为当神经网络的初始化参数可以很好的对数据本身进行重构时，该参数有利于对神经网络的进一步学习。然而，不同的模型对与学习的具体目标有不同的定义，下面我们就针对最常见的两种模型进行详细的阐述，即受限玻尔兹曼机模型和自动编码机（Auto Encoder）模型。

在深度学习模型中，我们首先利用非监督学习模型逐层贪婪地进行学习，进行参数初始化；之后我们利用初始化的参数，在整个网络末尾增加输出层，并利用标注样本进行监督学习调优（Fine-tuning）整个网络参数。整个过程如图 2-8。

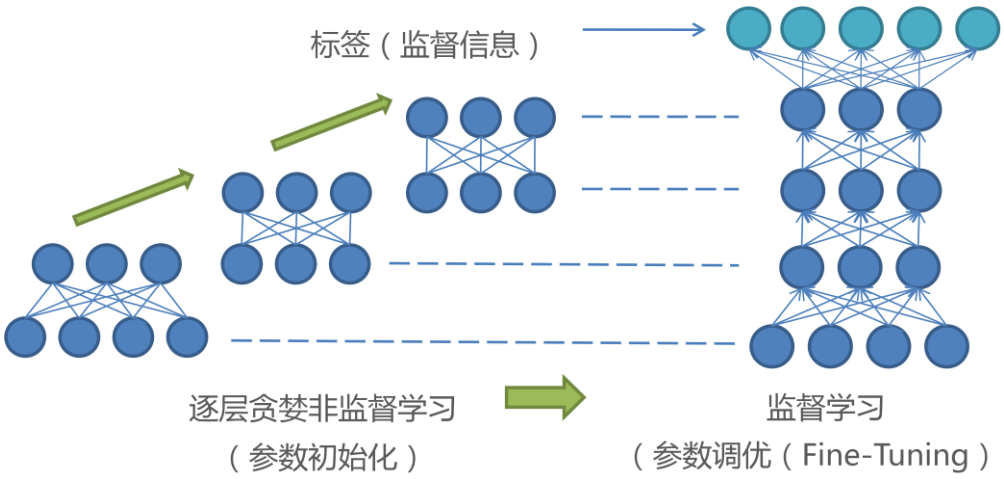


图 2-8 全连接网络深度学习整体流程示意图

2.2.1 受限玻尔兹曼机

2.2.1.1 简介

玻尔兹曼机（Boltzmann Machine）是一个双层的网络模型，在玻尔兹曼机中，我们将所有的网络节点分为两部分，即输入节点与隐节点。我们通过节点之间的连接来进行数据的描述与建模。玻尔兹曼机的输入层和隐含层之间为全连接，并且在输入层内部与隐含层内部，所有的节点也是全连接的。也就是说，如果我们不对输入节点和隐节点加以区分的话，玻尔兹曼机是一个全连接图结构。

相对于玻尔兹曼机，受限玻尔兹曼机是其的简化模型。在受限玻尔兹曼机中，我们消除了所有层内的连接，即在输入层间与隐含层间都不含有任何连接。这样不仅让模型的参数个数大大缩减，而且使得层内节点之间相互条件独立，使得我们可以有效的进行模型学习，具体的学习策略将在后文中详述。玻尔兹曼机与受限波尔兹曼机的模型见图2-9。

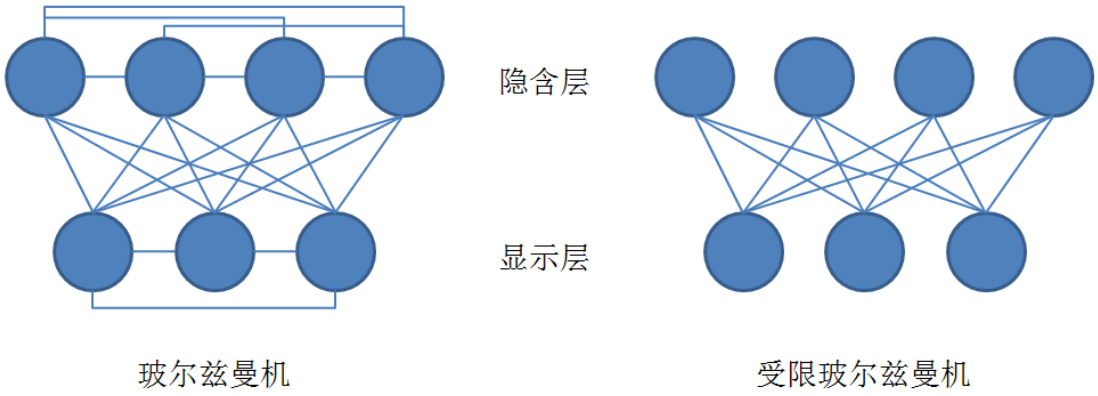


图 2-9 玻尔兹曼机与受限玻尔兹曼机示意图

2.2.1.2 建模与推理

我们可以将受限玻尔兹曼机当成是无向概率图模型来进行建模。对于一个受限玻尔兹曼机而言，假设输入层与隐含层之间有权重 **W**，输入层与隐含层的偏置向量分别为 **b** 和 **c**。当我们有输入 **v** 与隐节点状态 **h** 时，我们可以定义其概率为：

$$\mathrm{P}(\mathbf{v},\mathbf{h}) = \frac{1}{\mathrm{Z}}\exp\{-\mathrm{E}(\mathbf{v},\mathbf{h})\}$$

$$\mathrm{E}(\mathbf{v},\mathbf{h}) = -\mathbf{h}^{\mathrm{T}}\mathbf{W}\mathbf{v} - \mathbf{v}^{\mathrm{T}}\mathbf{b} - \mathbf{h}^{\mathrm{T}}\mathbf{c}$$

其中 **E(v, h)**为受限玻尔兹曼机的能量函数，这是一种在统计物理学中常见的建模方式，一般来说我们认为最大化概率约等于最小化能量函数。目前我们给出的能量表达式

假设输入与隐含节点皆满足伯努利分布（Bernoulli Distribution），对于实值建模的内容将在后文详细讨论。而 Z 为归一化因子，根据概率性质，我们易得其表达式为：

$$\mathrm{Z} = \sum_{\mathbf{v},\mathbf{h}} \exp\{-\mathrm{E}(\mathbf{v},\mathbf{h})\}$$

因为受限玻尔兹曼机中层内相互之间没有连接，所以在给定输入层时，所有输出层的节点相互独立，反之亦成。则根据这该条件概率以及概率积公式我们可知条件概率：

$$\mathrm{P}(\mathbf{v}|\mathbf{h}) = \prod_{\mathrm{i}} \mathrm{P}(\mathrm{v_i}|\mathbf{h}) = \prod_{\mathrm{i}} \frac{\mathrm{P}(\mathrm{v_i},\mathbf{h})}{\mathrm{P}(\mathbf{h})} = \prod_{\mathrm{i}} \frac{\mathrm{P}(\mathrm{v_i},\mathbf{h})}{\sum_{\hat{\mathrm{v}}_\mathrm{j}} \mathrm{P}(\hat{\mathrm{v}}_\mathrm{j},\mathbf{h})}$$

$$= \prod_{\mathrm{i}} \frac{\exp\{\mathrm{v_i}(\mathbf{h^T W})_\mathrm{i} + \mathrm{v_i b_i} + \mathbf{h^T c}\}}{\sum_{\hat{\mathrm{v}}_\mathrm{j}} \exp\{\hat{\mathrm{v}}_\mathrm{j}(\mathbf{h^T W})_\mathrm{j} + \hat{\mathrm{v}}_\mathrm{j}\mathrm{b_j} + \mathbf{h^T c}\}} = \prod_{\mathrm{i}} \frac{\exp\{\mathrm{v_i}(\mathbf{h^T W})_\mathrm{i} + \mathrm{v_i b_i}\}}{\sum_{\hat{\mathrm{v}}_\mathrm{j}} \exp\{\hat{\mathrm{v}}_\mathrm{j}(\mathbf{h^T W})_\mathrm{j} + \hat{\mathrm{v}}_\mathrm{j}\mathrm{b_j}\}}$$

$$\mathrm{P}(\mathbf{h}|\mathbf{v}) = \prod_{\mathrm{i}} \mathrm{P}(\mathrm{h_i}|\mathbf{v}) = \prod_{\mathrm{i}} \frac{\exp\{(\mathbf{W v})_\mathrm{i}\mathrm{h_i} + \mathrm{h_i c_i}\}}{\sum_{\hat{\mathrm{h}}_\mathrm{j}} \exp\{(\mathbf{W v})_\mathrm{j}\hat{\mathrm{h}}_\mathrm{j} + \hat{\mathrm{h}}_\mathrm{j}\mathrm{c_j}\}}$$

其中 $\mathrm{v_i}$ 表示向量 $\mathbf{v}$ 的第 i 个元素，$\hat{\mathrm{h}}_\mathrm{j}$、$\hat{\mathrm{v}}_\mathrm{j}$、$\mathrm{c_j}$、$\mathrm{v_j}$采用相同的表示方式；$(\mathbf{h^T W})_\mathrm{i}$表示 $\mathbf{h^T W}$ 形成的向量中第 i 个元素，$(\mathbf{W v})_\mathrm{j}$也采用相同的表示方式。又由概率的加定律可得

$$\mathrm{P}(\mathbf{v}) = \sum_{\mathbf{h}} \mathrm{P}(\mathbf{v},\mathbf{h})$$

对于受限玻尔兹曼机的训练来说，由于我们获得的信息仅有输入的数据，并且模型目标即为对输入数据建模。所以我们的优化目标为：

$$\max \ \log \mathrm{P}(\mathbf{v})$$

通过最大化输入（观测）样本的概率，我们即可实现对数据本身的建模工作。在得到了目标函数之后，下面我们将求取其对于参数项 $\mathbf{W}$, $\mathbf{b}$, $\mathbf{c}$ 的梯度表达式。我们定义目标函数为 $\mathrm{F} = \log \mathrm{P}(\mathbf{v})$，则对于抽象参数$\theta$ 可以推理：

$$\begin{aligned}
\frac{\partial \mathrm{F}}{\partial \theta} &= \frac{\partial}{\partial \theta} \log\left\{\sum_{\hat{\mathbf{h}}} \mathrm{P}(\mathbf{v},\hat{\mathbf{h}})\right\} \\
&= \frac{\partial}{\partial \theta} \log\left\{\sum_{\hat{\mathbf{h}}} \frac{1}{\mathrm{Z}} \exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}\right\} \\
&= \frac{\partial}{\partial \theta} \log\left\{\frac{1}{\mathrm{Z}} \sum_{\hat{\mathbf{h}}} \exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}\right\} \\
&= \frac{\partial}{\partial \theta} \log\left\{\sum_{\hat{\mathbf{h}}} \exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}\right\} - \frac{\partial}{\partial \theta} \log\{\mathrm{Z}\}
\end{aligned}$$

我们对于该两项分别展开可得，第一项：

$$\frac{\partial}{\partial\theta}\log\left\{\sum_{\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}\right\}=\frac{1}{\sum_{\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}}\frac{\partial}{\partial\theta}\sum_{\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}$$

$$=\frac{1}{\sum_{\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}}\sum_{\hat{\mathbf{h}}}\frac{\partial}{\partial\theta}\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}$$

$$=-\frac{1}{\sum_{\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}}\sum_{\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}\frac{\partial\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})}{\partial\theta}$$

第二项：

$$\frac{\partial}{\partial\theta}\log\{\mathrm{Z}\}=\frac{1}{\mathrm{Z}}\frac{\partial\mathrm{Z}}{\partial\theta}$$
$$=\frac{1}{\mathrm{Z}}\sum_{\hat{\mathbf{v}},\hat{\mathbf{h}}}\frac{\partial}{\partial\theta}\exp\{-\mathrm{E}(\hat{\mathbf{v}},\hat{\mathbf{h}})\}$$
$$=-\frac{1}{\mathrm{Z}}\sum_{\hat{\mathbf{v}},\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\hat{\mathbf{v}},\hat{\mathbf{h}})\}\frac{\partial\mathrm{E}(\hat{\mathbf{v}},\hat{\mathbf{h}})}{\partial\theta}$$

则我们综合可知：

$$\frac{\partial\mathrm{F}}{\partial\theta}=\frac{1}{\mathrm{Z}}\sum_{\hat{\mathbf{v}},\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\hat{\mathbf{v}},\hat{\mathbf{h}})\}\frac{\partial\mathrm{E}(\hat{\mathbf{v}},\hat{\mathbf{h}})}{\partial\theta}$$

$$-\frac{1}{\sum_{\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}}\sum_{\hat{\mathbf{h}}}\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}\frac{\partial\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})}{\partial\theta}$$

由于 F 是 **v** 的函数，从上式中我们可知，第一项是与 **v** 无关的，而第二项是与 **v** 有关的。因为第二项中 **v** 为输入的数据，仅有$\hat{\mathbf{h}}$为随机变量，我们一般称该梯度式的第二项为数据相关梯度；而对于第一项来说，$\hat{\mathbf{v}}$与$\hat{\mathbf{h}}$都为其随机变量，我们称其为模型相关的梯度。则其梯度表达式可简化为：

$$\frac{\partial\mathrm{F}}{\partial\theta}=\{\frac{\partial\mathrm{F}}{\partial\theta}\}_{\mathrm{model}}-\{\frac{\partial\mathrm{F}}{\partial\theta}\}_{\mathrm{data}}$$

从而我们也可以更加清楚的理解受限玻尔兹曼机的学习目标。因为受限玻尔兹曼机可以当作是一个概率图模型，则其求解目标为最小化数据分布与模型分布之间的距离，也就是让模型分布最接近于数据分布。既然我们已经得到数据的概率分布，则可以很容易的从模型的概率分布中进行采样，从而“生成”符合分布的数据，即生成式模型（Generative Model）。

更进一步来看，对于具体的网络参数 **W**, **b**, **c**，我们可知能量函数相对于其的梯度表达式为：

$$\frac{\partial \mathrm{E}(\mathbf{v},\mathbf{h})}{\partial \mathbf{W}} = -\mathbf{h}^{\mathrm{T}}\mathbf{v}$$
$$\frac{\partial \mathrm{E}(\mathbf{v},\mathbf{h})}{\partial \mathbf{b}} = -\mathbf{v}$$
$$\frac{\partial \mathrm{E}(\mathbf{v},\mathbf{h})}{\partial \mathbf{c}} = -\mathbf{h}$$

至此我们已经详细推导了所有的梯度表达式，在下一节我们将阐述如何利用这些梯度表达式进行模型的学习。

2.2.1.3 模型学习

在上一节中我们得出了受限玻尔兹曼机的梯度函数，但我们发现对于该梯度函数而言，由于其建立在随机变量的和上，我们无法直接对梯度进行求取。在这里我们对之前的导数表达式略微进行变形，可得:

$$\frac{\partial \mathrm{F}}{\partial \theta} = \sum_{\hat{\mathbf{v}},\hat{\mathbf{h}}} \frac{\exp\{-\mathrm{E}(\hat{\mathbf{v}},\hat{\mathbf{h}})\}}{\mathrm{Z}} \frac{\partial \mathrm{E}(\hat{\mathbf{v}},\hat{\mathbf{h}})}{\partial \theta}$$
$$- \sum_{\hat{\mathbf{h}}} \frac{\exp\{-\mathrm{E}(\mathbf{v},\hat{\mathbf{h}})\}}{\sum_{\check{\mathbf{h}}} \exp\{-\mathrm{E}(\mathbf{v},\check{\mathbf{h}})\}} \frac{\partial \mathrm{E}(\mathbf{v},\hat{\mathbf{h}})}{\partial \theta}$$
$$= \sum_{\hat{\mathbf{v}},\hat{\mathbf{h}}} \mathrm{P}(\hat{\mathbf{v}},\hat{\mathbf{h}}) \frac{\partial \mathrm{E}(\hat{\mathbf{v}},\hat{\mathbf{h}})}{\partial \theta} - \sum_{\hat{\mathbf{h}}} \mathrm{P}(\hat{\mathbf{h}}|\mathbf{v}) \frac{\partial \mathrm{E}(\mathbf{v},\hat{\mathbf{h}})}{\partial \theta}$$

从式中可以看出受限波尔兹曼机的梯度表达式与两个概率分布相关，即 P(**v**, **h**)与 P(**h** | **v**)。所以我们可以使用采样技术来进行模型梯度的求解，当进行多次采样平均后，其梯度值将近似等于原始梯度值。

对于条件概率分布 P(**h** | **v**)来说，由于输入变量 **v** 是已知的，则采样相对较为简单。举例来说，我们现在讨论二值的玻尔兹曼机，即输入和隐含变量皆符合伯努利分布，则我们可知：

$$\mathrm{P}(\mathbf{h}|\mathbf{v}) = \prod_{\mathrm{i}} \mathrm{P}(\mathrm{h}_{\mathrm{i}} = 1|\mathbf{v}) = \prod_{\mathrm{i}} \frac{\exp\{(\mathbf{W}\mathbf{v})_{\mathrm{i}}\mathrm{h}_{\mathrm{i}} + \mathrm{h}_{\mathrm{i}}\mathrm{c}_{\mathrm{i}}\}|_{\mathrm{h}_{\mathrm{i}}=1}}{\exp\{(\mathbf{W}\mathbf{v})_{\mathrm{j}}\mathrm{h}_{\mathrm{j}} + \mathrm{h}_{\mathrm{j}}\mathrm{c}_{\mathrm{j}}\}|_{\mathrm{h}_{\mathrm{j}}=0} + \exp\{(\mathbf{W}\mathbf{v})_{\mathrm{j}}\mathrm{h}_{\mathrm{j}} + \mathrm{h}_{\mathrm{j}}\mathrm{c}_{\mathrm{j}}\}|_{\mathrm{h}_{\mathrm{j}}=1}}$$
$$= \prod_{\mathrm{i}} \frac{1}{1 + \exp\{-(\mathbf{W}\mathbf{v})_{\mathrm{i}} - \mathrm{c}_{\mathrm{i}}\}} = \mathrm{Sigmoid}(\mathbf{W}\mathbf{v} + \mathbf{c})$$

则可以根据该分布进行采样。然而对于模型概率分布，即 P(**v**, **h**)来说，我们很难直接从联合概率上来采样。所以我们采用吉布斯采样（Gibbs Sampling）方法，将从联合概率中的采样问题分解为从两个条件概率的迭代采样，即：

$$\mathbf{h_1} \sim P(\mathbf{h}|\mathbf{v_0})$$

$$\mathbf{v_1} \sim P(\mathbf{v}|\mathbf{h_1})$$

$$\mathbf{h_2} \sim P(\mathbf{h}|\mathbf{v_1})$$

$$\mathbf{v_2} \sim P(\mathbf{v}|\mathbf{h_2})$$

$$\cdots\cdots$$

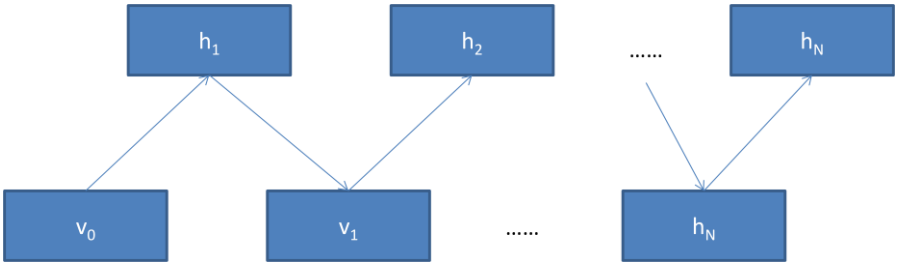


图 2-10 吉布斯采样蒙特卡洛链示意图

在进行 N 次重复吉布斯采样后，我们最终得到的 **v** 和 **h**，即等同于从联合概率 P(**v**, **h**)中采样的结果。因为在吉布斯采样中，每一次采样都只依赖于上一次采样的结果，所以我们就在吉布斯采样的过程中得到了一个马尔科夫链（如图 2-10）。根据上文的结果，我们也可知：

$$P(\mathbf{h}|\mathbf{v}) = \text{Sigmoid}(\mathbf{W}\mathbf{v} + \mathbf{c})$$

$$P(\mathbf{v}|\mathbf{h}) = \text{Sigmoid}(\mathbf{h}^{\mathbf{T}}\mathbf{W} + \mathbf{b})$$

理论上来说，对于吉布斯采样当采样的结果趋近于收敛时，我们可以很好的证明采得的样本属于联合分布 P(**v**, **h**)。但由于每一步吉布斯采样都需要重新估计条件概率分布，整个采样过程的复杂度极高。在实践中 Hinton 等人发现，当我们求解受限玻尔兹曼机时，并不需要精确的采样值；相对的，一个粗糙的采样值就可以很好的标明梯度方向。在实验中，他们发现仅仅一组采样就可以很好的进行学习，这无疑大大降低了受限玻尔兹曼机的训练速度。这种方法被命名为 CD（Contrastive Divergence，如图 2-11）[32]，并被广为使用。

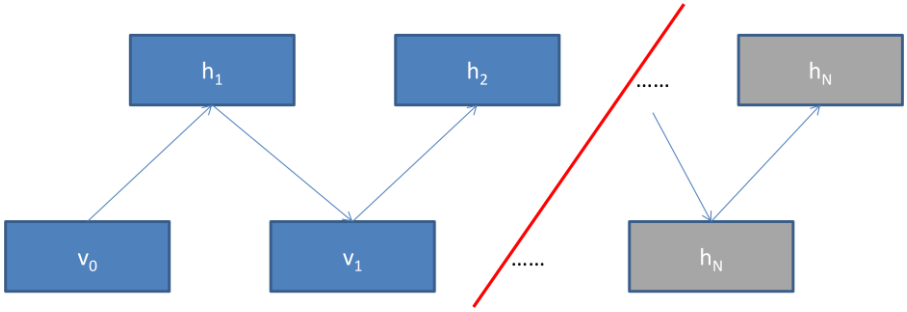


图 2-11 Contrastive Divergence 采样示意图

借助于 CD 算法，我们可以很快的求解模型梯度，则之后可以利用梯度下降（Gradient Descent）方法对模型进行求解。在实践中，一般来说使用 CD-1，即仅一次采样的结果就可以很好的进行模型初始化；但如果我们的目标在于高质量的进行数据概率建模，CD-1 的结果会相对粗糙一些，通常的做法是从 CD-1 开始快速训练，之后渐渐过渡到 CD-5 或 CD-10 来求取精确模型。

关于训练算法，很多研究者对 CD 进行了扩展，使得求解更加容易与快速。其中比较有代表性的为 PCD（Persistent Contrastive Divergence）算法[73]。PCD 算法的前提假设是受限玻尔兹曼机的概率模型在学习中是缓慢变化的，所以在每一次迭代后我们都保留当前参数下的采样结果，这样在之后的学习中只需要从该状态出发就可以在很短的时间内得到接近当前模型分布的采样。PCD 算法在达到较高的采样精度的前提下，也避免了多次吉布斯采样带来的计算负担。

2.2.1.4 受限玻尔兹曼机的实值建模

在前文中，我们讲述的都是使用受限玻尔兹曼机进行二值模式建模的过程。相对来说，二值模式由于计算和采样（伯努利分布采样）都比较简单，在受限玻尔兹曼机的学习中也比较容易。而如果我们的目标是对实值模式进行建模，情况就要稍微复杂一些。通常在实值建模的时候，我们会假设数据符合高斯分布。在受限玻尔兹曼机中，我们保持隐节点为伯努利分布不变，而将输入节点变为高斯分布，就得到了高斯-伯努利受限玻尔兹曼机（Gaussian-Bernoulli Restricted Boltzmann Machine，如图 2-12）[10]。下面我们就简述这一模型。

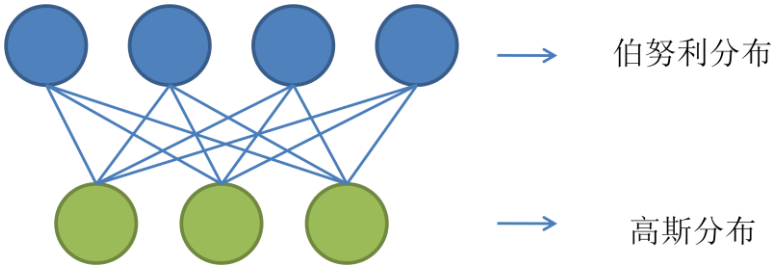


图 2-12 高斯-伯努利受限玻尔兹曼机示意图

在高斯-伯努利受限玻尔兹曼机中，由于输入节点分布假设为高斯分布，其能量函数变为：

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{h}^{\mathbf{T}}\mathbf{W}\frac{\mathbf{v}}{\boldsymbol{\sigma}} + \left\|\frac{\mathbf{v}-\mathbf{b}}{2\boldsymbol{\sigma}^{\mathbf{2}}}\right\|_2^2 - \mathbf{h}^{\mathbf{T}}\mathbf{c}$$

在能量函数表达式中，高斯-伯努利受限玻尔兹曼机模型引入了新的向量 σ 用来表示$\mathbf{v}$各个维度上高斯分布的方差，其中$\boldsymbol{\sigma}^2$表示对向量$\boldsymbol{\sigma}$的每个元素进行平方运算，结果仍是向量。依照能量函数，我们可以推出响应的条件概率表达式为：

$$P(\mathbf{v}|\mathbf{h}) = \prod_i P(v_i|\mathbf{h}) = \prod_i \frac{P(v_i,\mathbf{h})}{P(\mathbf{h})} = \prod_i \frac{P(v_i,\mathbf{h})}{\sum_{\hat{v}_j} P(\hat{v}_j,\mathbf{h})} = \prod_i \frac{\exp\{-E(v_i,\mathbf{h})\}}{\sum_{\hat{v}_j}\exp\{-E(\hat{v}_j,\mathbf{h})\}}$$

$$= \prod_i \frac{\exp\{(\mathbf{h}^{\mathbf{T}}\mathbf{W})_i \frac{v_i}{\sigma_i} - \frac{(v_i-b_i)^2}{2{\sigma_i}^2}\}}{\sum_{\hat{v}_j}\exp\{(\mathbf{h}^{\mathbf{T}}\mathbf{W})_j \frac{\hat{v}_j}{\sigma_j} - \frac{(\hat{v}_j-b_j)^2}{2{\sigma_j}^2}\}} = \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(v_i-b_i-\sigma_i(\mathbf{h}^{\mathbf{T}}\mathbf{W})_i\right)^2}$$

$$P(\mathbf{h}|\mathbf{v}) = \prod_i P(h_i|\mathbf{v}) = \prod_i \frac{\exp\{\left(\mathbf{W}\frac{\mathbf{v}}{\boldsymbol{\sigma}}\right)_i + c_i\}}{\sum_{h_j}\exp\{\left(\mathbf{W}\frac{\mathbf{v}}{\boldsymbol{\sigma}}\right)_j h_j + h_j c_j\}}$$

其中$\frac{\mathbf{v}}{\boldsymbol{\sigma}}$表示 $\mathbf{v}$ 与$\boldsymbol{\sigma}$的对应位置相除，$\left(\mathbf{W}\frac{\mathbf{v}}{\boldsymbol{\sigma}}\right)_i$表示运算结果上的第 i 个元素。由上式可以看出，输入节点已经符合高斯分布了，接下来我们可以进一步推导，得出各个参数相对于能量函数的导数为：

$$\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \mathbf{W}} = -\mathbf{h}^{\mathbf{T}}\left(\frac{\mathbf{v}}{\boldsymbol{\sigma}}\right)$$

$$\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \mathbf{b}} = \frac{\mathbf{v}-\mathbf{b}}{\boldsymbol{\sigma}^{\mathbf{2}}}$$

$$\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \mathbf{c}} = -\mathbf{h}$$

$$\frac{\partial \mathrm{E}(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\sigma}} = \mathbf{h}^{\mathrm{T}}\mathbf{W}\left(\frac{\mathbf{v}}{\boldsymbol{\sigma}^{2}}\right) + \frac{(\mathbf{v}-\mathbf{b})}{\boldsymbol{\sigma}^{3}}$$

根据导数表达式，我们就可以使用梯度下降方法进行模型的求解了。值得注意的是，由于方差变量 σ 的梯度表达式与 $1/\sigma^2$ 正相关，所以当 σ 较小时容易发生数值错误。可以在 σ 值上做一定的数值限制，并且降低 σ 的学习速率。

2.2.2 自动编码器

2.2.2.1 简介

自动编码机（Auto Encoder）是另一大类常用的深度学习基础组件（Building Block）模型。在自动编码机中，模型本身并不直接关心数据的分布情况，而只关注数据本身的重构情况。可以说在自动编码机中，基本假设比受限玻尔兹曼机要简单；而且由于自动编码机没有建立概率模型，所以并不能从模型中进行样本采样。一般来说，相对于受限玻尔兹曼机这类基于概率分布的方法，自动编码机为基于数据重构的方法。

2.2.2.2 建模

与受限玻尔兹曼机一样，自动编码机也可以看成一个两层的神经网络。在自动编码机中，有两个重要的操作：编码和解码（如图 2-13）。在编码阶段，数据从模型的输入节点进入，经过网络映射到隐含节点，此时在隐含节点的值可以看做是输入数据的某种编码结果；而在解码时，我们把隐含节点的状态值通过网络反向映射回输入节点，相当于把某种编码的结果进行解码，使得信号返回到原始数据的空间。而自动编码机的目标（如图 2-14），就是学习一种有效的编码、解码方法，使得数据在编码、解码操作后与原始数据尽可能一致；即使得自动编码机可以良好的重构原始数据。

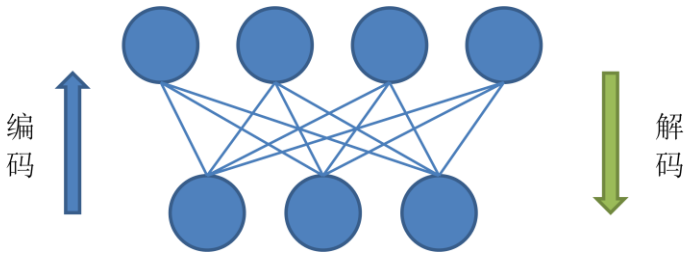


图 2-13 自动编码机示意图

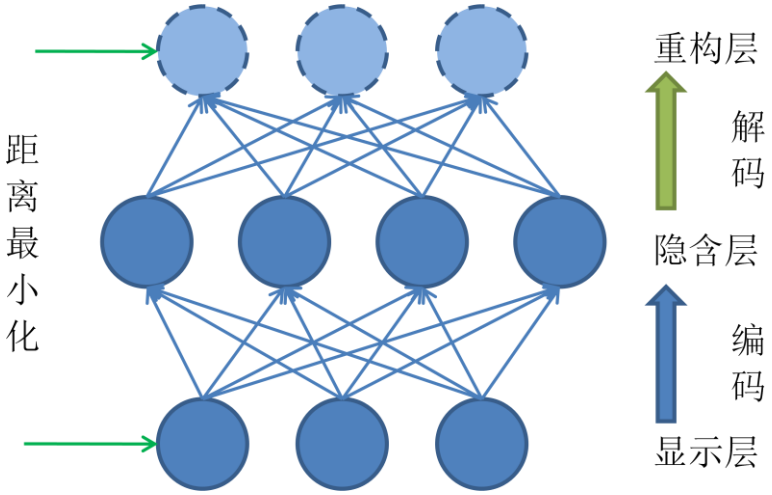


图 2-14 自动编码机求解目标示意图

假设输入数据为 v，隐节点状态为 h，重构结果为 r，连接权重为 W，输入层与隐含层的偏置分别为 b 和 c，那么自动编码机的编码和解码操作表达式为：

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{v} + \mathbf{b})$$

$$\mathbf{r} = \sigma(\mathbf{h}^{\mathbf{T}}\mathbf{W} + \mathbf{c})$$

其中 σ 表示某种激活函数。自动编码机的激活函数选用与性质和一般全连接神经网络的激活函数一致，相关介绍请参看 2.1.1。在通常情况下，为了缩减模型的参数规模，我们会像上式一样对编码和解码的映射矩阵进行绑定（Tied）；即在解码时，我们使用映射矩阵 $\mathbf{W}$ 的转置来进行映射。这种绑定可以看成是一种正则（Regularization）用来防止模型过拟合，但绑定并不是必须的，也就是说你完全可以针对编码和解码的两个过程使用两个不同的映射矩阵。自动编码机的目标函数为：

$$\min \mathrm{F}(\mathrm{v}) = \min \Delta(\mathrm{x}, \mathrm{r})$$

其中 Δ 为距离度量函数，在自动编码机中常见的度量函数有范数距离函数、交叉熵函数等，将在下一节做详细介绍。

在得到目标函数后，利用误差反传算法，我们容易获得其对各个参数的导数。之后与受限玻尔兹曼机的模型求解方式一样，可以利用导数信息使用梯度下降算法进行模型求解。联合之前提到的受限玻尔兹曼机，从求解和建模上我们可以看出，虽然自动编码机和受限玻尔兹曼机的模型大小和形式都很相近，但自动编码机更偏向于全连接神经网络，而受限玻尔兹曼机更偏向于概率图模型（Probabilistic Graphical Model）。

2.2.2.3 距离函数与数据分布

在上文中，我们提到在设计自动编码机时，我们可以选用不同的距离函数。理论上来说，任何可导的距离函数都可以被自动编码机接收，但通常常用的主要有两种距离函数，二范数距离（欧氏距离）以及交叉熵距离。下面我们将较详细的介绍这两种距离函数，其表达式、适用情况以及隐含针对的数据分布。

二范数距离

$$\Delta(\mathbf{a},\mathbf{b}) = ||\mathbf{a}-\mathbf{b}||_2^2 = \sum_i (a_i - b_i)^2$$

二范数距离，或欧氏距离，应该是我们最常见的一类距离函数了。从表达式可知，该距离函数的值域为$[0,+\infty]$。我们可以使用一维数据来示意二范数距离（如图 2-15）：

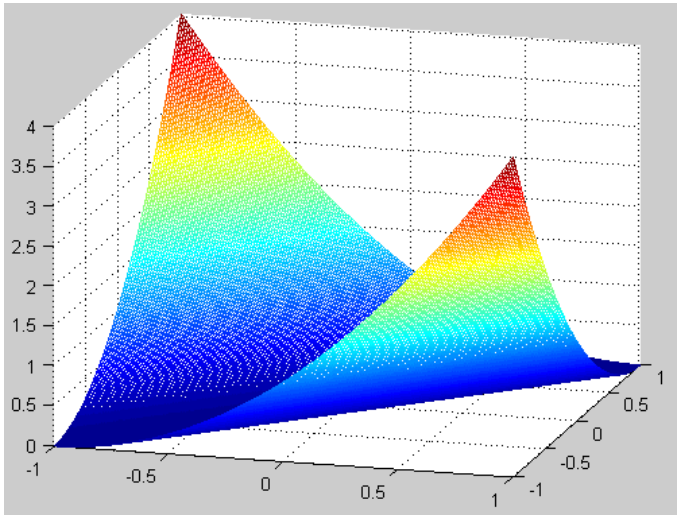


图 2-15 二范数距离示意图

交叉熵距离

$$\Delta(\mathbf{a},\mathbf{b}) = -\mathbf{a}^{\mathbf{T}}\log(\mathbf{b}) - (\mathbf{1}-\mathbf{a})^{\mathbf{T}}\log(\mathbf{1}-\mathbf{b})$$
$$= \sum_i -a_i \log(b_i) - (1-a_i)\log(1-b_i)$$

对于交叉熵距离，我们可以发现他的定义域并不是完整的实数范围，而是[0,1]，于是可得其值域范围为(-∞,0)，其距离的示意图为（图 2-16）：

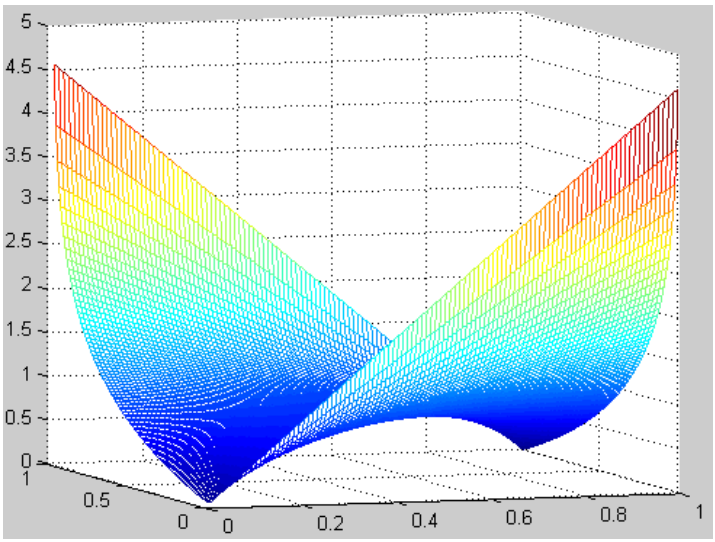


图 2-16 交叉熵距离示意图

从示意图中我们可以看出，交叉熵距离只有在 a, b 完全一致时为 0，其余情况全部都大于 0。为了与二范数距离形成对比，我们同样打出[0,1]范围内的示意图（图 2-17）。

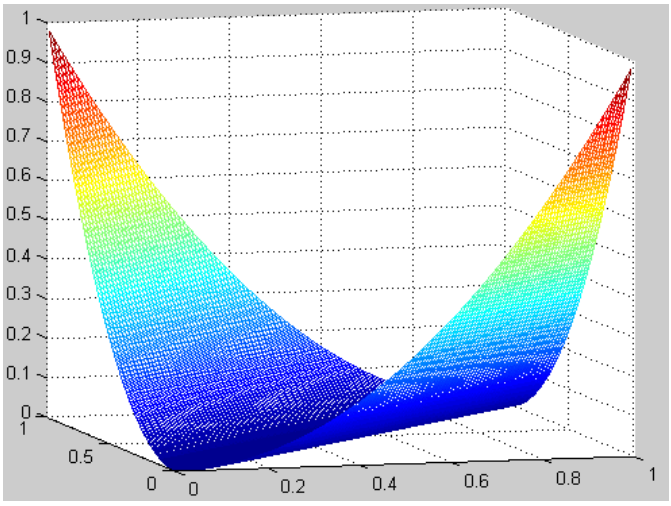


图 2-17 二范数距离在 0-1 间响应示意图

从示意图中我们可以看出，除了距离的尺度不一以外，交叉熵距离与二范数距离最大的区别在于对角线的值上，即当 a=b 的情况下的距离。在二范数距离内，当 a=b 的时候距离保持为 0,；而在交叉熵距离内，仅当 a 与 b 同时为 0 或同时为 1 时距离为 0。

从这一点来看，联合之前讨论的自动编码机的目标函数，我们可以得出当使用交叉熵距离函数时，自动编码机会强制数据趋向于 0 或 1，也即是二值模式。这也就意味着当数据为二值模式时，使用交叉熵距离可以使得自动编码机更好的刻画数据；而使用二范数距离时，虽然也可以正常进行学习，但模型并不会关心数据是否为二值模式，而使

得对数据的描述能力较差。这一点已近在很多实验中得到了证实。

在自动编码机中并没有数据分布的概念，也就是说不管数据是来自什么分布的，其实我们都可以选取某个距离、某个激活函数构成一个自动编码机来对其进行学习。但这并不意味着自动编码机对数据分布式鲁棒的，也就是说“可以进行建模”并不等同于“建模的很好”。在这一点上，通过定义不同的距离函数，自动编码机实际上在内部隐含声明了数据的分布类型，一个符合数据分布规律的距离函数可以很大程度上提升自动编码机的编码效果。一般来说对于实值模式一般采用欧氏距离度量，对于二值模式则一般使用交叉熵距离。

2.2.2.4 模型退化与降噪自动编码机

在自动编码机中，一个一直存在的问题即是相对于受限玻尔兹曼机，自动编码机的学习过程很容易退化。也就是说，自动编码机在学习中学到了一个并不重要的模型，利用“诡计（Trick）”完成了数据重构的任务，这一点对于进行过完备（Over Complete）映射的学习时，即隐含节点个数大于等于输入节点的时候更为常见。举例来说，对于一个 n 维的数据，我们要使用自动编码机编码到 n 维，即保持维数一致。那么对于自动编码机来说，如果我们把映射矩阵看成是 n 个 n 维坐标系下的坐标轴，那么任何 n 个正交坐标轴都可以张成一个完备的 n 维空间，即可以完美重构任何 n 维空间的数据。但在这种情况下，自动编码机并没有在数据中获取到任何有用信息，因为 n 个正交坐标轴张成 n 维空间是与数据无关的，也就是发生了严重的模型退化。

为了解决这一问题，Vicent 提出了降噪自动编码机（Denosing Auto Encoder）[75]。在降噪自动编码机中，我们使用一个被噪声污染的数据进行编码和解码，并让重构的结果尽可能的接近没有被噪声污染的原始数据，如图 2-18 所示。

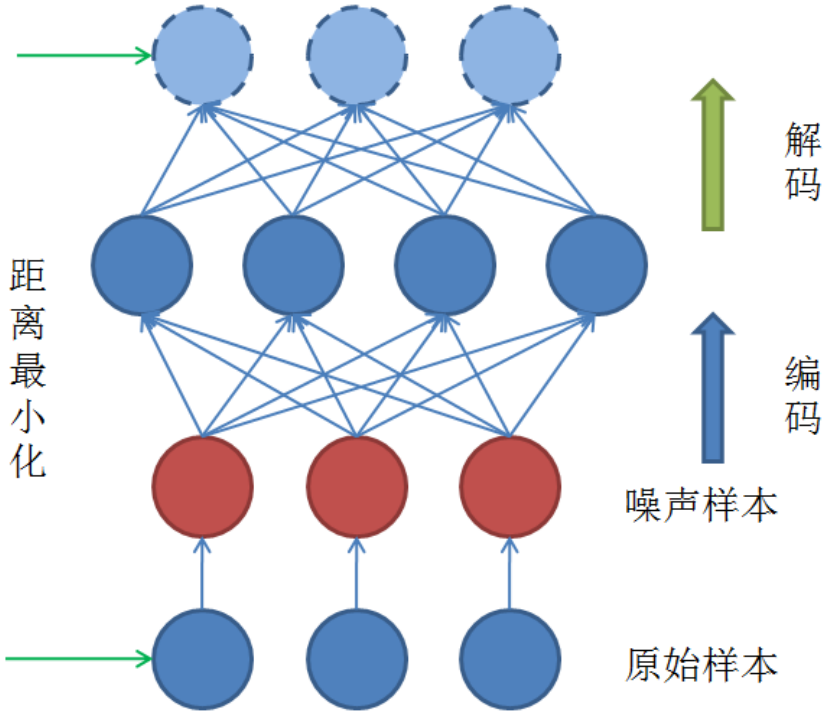


图 2-18 降噪自动编码机示意图

在降噪自动机中，常见的噪声类型有：置零噪声、胡椒盐噪声与高斯噪声。噪声的类型与输入数据的分布也有一定的相关性。比如说置零噪声和胡椒盐噪声对于二值模式来说比较实用，而高斯噪声对于实值模式比较适用。但具体采用何种噪声进行学习实际上是一个相对开放的选择，也就是说在大多数情况下选择不同的噪声都可以进行降噪自动机学习，但学习的效果可能有所不同。

利用降噪措施，自动机在学习时由于输入和重构目标不一致（加入了噪声），很大程度上的抑制了模型退化现象的出现，取得了与受限玻尔兹曼机相类似的性能。在这两种模型上，实际上有研究者发现他们之间是高度相关的[70]。所以我们在选择的时候，可以尽可能的根据自己的需要进行选择，对于一些应用：比如生成、概率建模等，使用受限玻尔兹曼机可能会更为方便；而对于另一些仅仅关注数据描述学习的应用，或是想获得某种特定噪声鲁棒性的应用，用自动编码机来建模就更为便捷。

2.2.3 贪婪堆叠学习

在前文中我们比较详细的介绍了受限玻尔兹曼机和自动编码机这两种双层网络模型。在深度学习中，我们需要对深层网络学习，也就是说我们需要使用这两种模型对深层网络模型进行参数初始化。

其参数的初始化采用贪婪算法，即逐层进行初始化。假设我们有一个深度神经网络，在初始化过程中，我们先固定第 i-1 层的数据，将 i-1 到 i 层当做受限玻尔兹曼机或自动编码机进行学习；之后我们固定第 i 层的数据，并继续使用这两种模型初始化 i 到 i+1 层的神经网络（如图 2-19）。在整个网络都被初始化之后，我们就可以利用标注信息对该深度神经网络进行监督学习了。

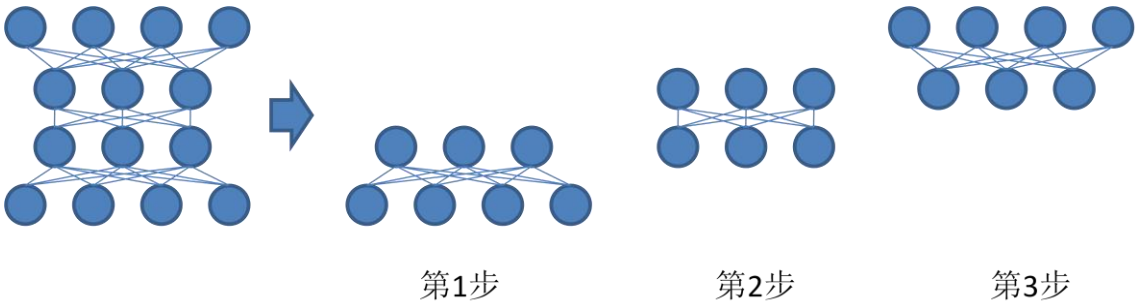


图 2-19 贪婪堆叠学习示意图

在实践中，这种贪婪的初始化策略获得了很好的效果（如图 2-20），尤其是在训练深度神经网络时，贪婪初始化的网络能更好的进行监督学习任务。至于贪婪初始化可行的原因，有研究者指出贪婪逐层的初始化学习可以逐步降低变分误差界[46]。我们认为更通俗的解释为：在我们之前的分析来中，在深度网络中由于误差消失现象，我们无法把误差修正信号从输出层反传到靠前的层次。而贪婪的逐层初始化使得低层次网络建立了很好的初始参数，而生成式的学习方式也确定了只是对数据进行更为紧致的表示而尽量不丧失信息。这样的话在之后的监督学习中，即使误差信号仍然传不到靠前层，初始化的基石也可以保证学习的顺利进行。事实上人们发现在监督学习时，仅仅学习最后一层和调优（Fine Tune）整个网络得到的结果有较大的差距，全网络调优之后的结果往往要较大幅度的优于前者[12]（如图 2-20）。因为监督分类行为实际上并不等同于数据生成的学习，也就是说我们的初始化只能给予深度网络一个较好的“数据表示”或是“参数初始值”，而真正的监督学习目标，比如分类、回归等仍要依赖于网络的参数调优。

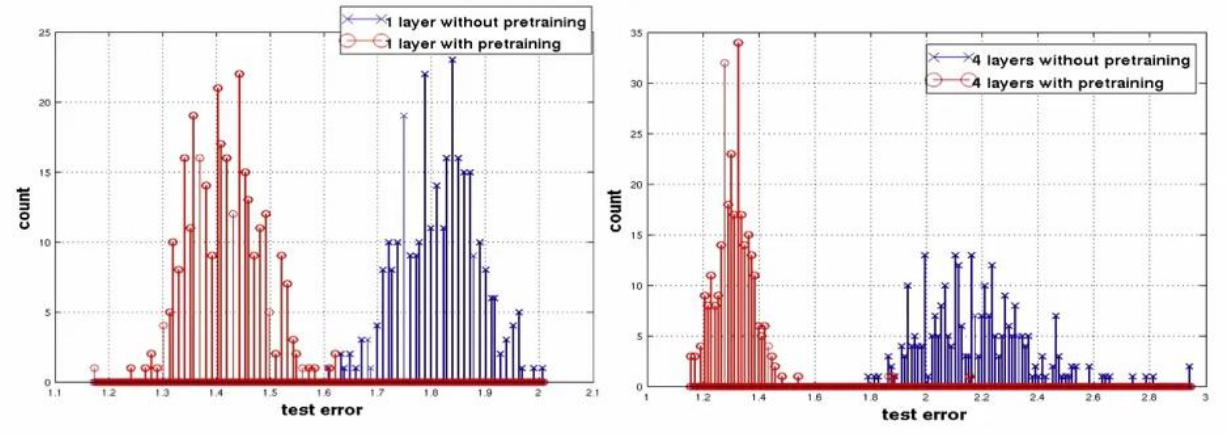


图 2-20 参数初始化对比图[12]，其中红色为经过参数初始化（贪婪堆叠学习）的结果，蓝色为不经过参数初始化的结果；左图为不进行参数调优的结果，右图则经过参数调优

2.2.4 正则约束

在机器学习中，为了避免数据过少或模型参数过多带来的过拟合现象，我们通常会在网络中加入正则约束。在深度学习中，较常用的正则约束有 L2 范式与 L1 范式（稀疏性）约束，在这一节我们将针对这两种正则约束在深度学习中的应用做详细说明。

2.2.4.1 L2 范式约束

在 L2 范式约束中，假设我们优化目标的参数项为 $\mathbf{w}$，$\mathbf{w}_i$ 表示 $\mathbf{w}$ 的第 i 个元素，则该正则表达式可以写做：

$$\mathrm{Reg}_{L2} = \|\mathbf{w}\|_2^2 = \sum_i \mathbf{w}_i^2$$

我们易得该正则的梯度表达式为：

$$\mathrm{Reg}'_{L2} = \frac{\partial \mathrm{Reg}_{L2}}{\partial \mathbf{w}} = 2\mathbf{w}$$

在 L2 正则中，我们通过 L 范式来约束模型规模，注意在这里模型的规模并非是 $\mathbf{w}$ 中的参数个数，而是其中参数的大小；举例来说，在求解线性方程组时，当条件过完备（Over-Complete）的时候，对于线性方程组来说有无数多个解，而在加入 L2 正则后，我们会偏向于获取值的平方和最小那一组解。在一般问题的求解中，L2 正则的作用大致相仿，假设有优化目标（如图 2-21 左）。我们在图中可以发现在两个中心轴上其目标函数都达到最小；而在加入 L2 正则后（如图 2-20 右），我们则可以求得唯一解即 x=0，y=0。根据奥卡姆准则，在模型规模较小的情况下（加入 L2 准则后的情况），模型的泛化性能要更为优异。

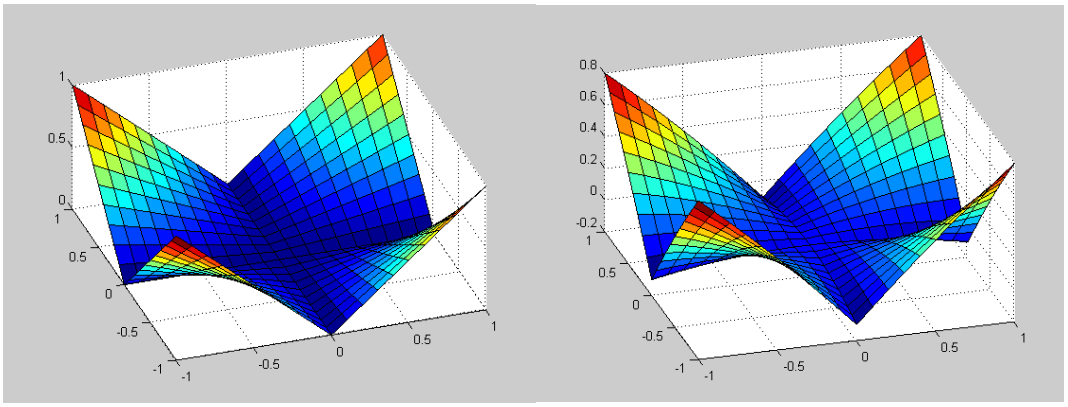


图 2-21 原始优化目标（左）与加入 L2 正则后优化目标（右）对比图

在实践中，L2 正则被证明能有效防止模型的过拟合现象，在神经网络中，这种正则也被叫做权值退化（Weight Decay），并被广泛应用。

2.2.5 L1 范式约束与稀疏约束

除了 L2 范式，L1 范式也是常见的正则约束。假设目标函数参数为 $\mathbf{w}$，则 L1 范式的表达式为：

$$\mathrm{Reg}_{\mathrm{L1}} = |\mathbf{w}| = \sum_{\mathrm{i}} |\mathbf{w}_{\mathrm{i}}|$$

相对于 L2 范式，我们发现 L1 范式并不是连续可导的，其梯度表达式为分段函数，即：

$$\mathrm{Reg}'_{\mathrm{L1}} = \begin{cases} 1 & \mathbf{w}_{\mathrm{i}} > 0 \\ -1 & \mathbf{w}_{\mathrm{i}} < 0 \end{cases}$$

从表达式我们可以看出，L1 范式正则约束也具有 L2 范式正则项的特性，即会倾向于选择较小的模型，只不过在计算时采用绝对值和而不是平方和。然而 L1 范式具有另一个特殊的性质，即在 L1 范式下可以提升参数的稀疏性（Sparseness）。对于一个凸优化问题，L1 正则更倾向于使得一些参数为 0，而 L2 则仅是让参数变小，并不具备稀疏性的约束（如图 2-22）。

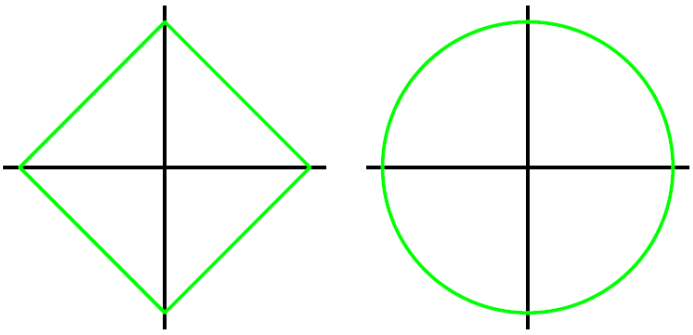


图 2-22 L1 范式（左）与 L2 范式(右)约束示意图

这一性质在稀疏编码（Sparse Coding）中得到了充分的利用，研究者们发现稀疏形式更利于特征的表达，并在许多任务中都获取了响应的提升[24]。并且从神经学来看，人脑的神经元响应也是相当稀疏的[50]。

像稀疏编码一样，在深度学习中除了直接对参数进行稀疏约束外，另外一种也是更为通用的即是对于特征表达(即隐含层信号)进行稀疏约束。假设对于自动编码机而言，我们有显层信号 **v**，隐层响应为 **h**，重构为 **r**，则对于特征表达（**h**）加入 L1 正则之后的目标函数为：

$$\min\|\mathbf{v}-\mathbf{r}\|_2^2+\lambda|\mathbf{h}|$$

在这里λ 的大小控制着 **h** 的稀疏性，λ 越大 h 的稀疏性越高。

相对于直接使用 L1 范式对隐层响应 h 进行约束，对于 sigmoid 激活函数而言我们还可以人为指定稀疏的比例。上文中我们提到过，sigmoid 激活函数的值域范围为[0,1]，对重构二值模式更为合适。在这里我们可以定义响应比例为：

$$\mathrm{rah}(\mathbf{h})=\frac{\sum_{i=1}^{M}h_i}{M}$$

其中 M 为隐含层的维度，h_i 为 **h** 的第 i 个元素。在这里通过定义响应比例，我们可以将其约束为确切的值，即目标函数为：

$$\min\|\mathbf{v}-\mathbf{r}\|_2^2+\lambda(\mathrm{rah}(\mathbf{h})-\alpha)^2$$

在这里λ 仍为误差函数与正则约束之间的调节函数，α 则为指定的在隐含层内响应比例的大小。

除了像稀疏编码一样在数据的表示内进行约束外（即对于单个数据的表示中，表示向量是稀疏的），在深度学习中更好也更为流行的方式为在让隐节点在数据上稀疏响应（即表述的每个维度都在所有观测数据上稀疏响应，[34]）。假定我们有 N 个样本

$\{\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(N)}\}$，那么对与隐层响应 h 来说，我们可以定义其在数据上的响应，即：

$$\mathrm{rad}(\mathbf{h})=\frac{1}{N}\sum_{j=1}^{N}\mathbf{h}^{(j)}$$

其中 $\mathbf{h}^{(j)}$为第 j 个样本的隐层相应。则与刚刚的约束式相仿，我们可以定义新的目标函数为：

$$\min\|\mathbf{v}-\mathbf{r}\|_2^2+\lambda\sum_{i=1}^{M}(\mathrm{rad}(\mathbf{h})_i-\alpha)^2$$

其中，$\mathrm{rad}(\mathbf{h})_i$表示隐含层第 i 维在数据上的相应。在这里α 代表的意义是隐节点在数据上的响应频率。利用稀疏正则约束下，Lee 使用受限玻尔兹曼机得到了类似于人脑视觉皮层 V2 的响应模式[50]（如图 2-23）。

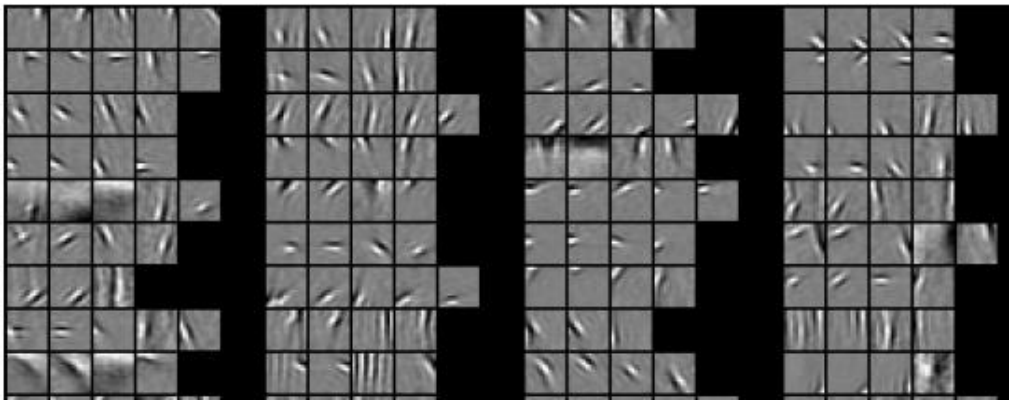


图 2-23 稀疏约束下特征响应图[50]

2.3 基于显著性物体检测的深度学习

在上文中介绍了相关的神经网络知识后，本节开始介绍我们提出的新的神经网络初始化方式，即利用显著性物体检测算法的输出结果来辅助我们进行网络参数的初始化。我们已经了解，在训练深层网络时，网络参数的初始化对于整个深度神经网络的性能有很大的影响。进一步，我们发现在加入显著性物体检测结果后，我们可以对神经网络做更好的初始化，并在不改变神经网络规模的前提下，提升神经网络分类任务的性能。在下面我将详细介绍该模型，其中 3.3.1 节简要介绍视觉显著性技术与其先验作用；3.3.2 节介绍我们的建模方式与相关推导；3.3.3 节介绍相关的实验与参数设计，以及对结果的分析与讨论。

2.3.1 视觉显著性物体检测与先验

视觉显著性检测（Saliency Detection）一直是计算机视觉研究的基本问题之一。在视觉显著性检测的工作中，我们通过对人的真实视点进行采集，得到图像中最容易受人关注的点，并利用仿生或学习的方式进行建模，并使得我们得到的模型尽可能接近人类的真实视点[36]。但基于视点的视觉显著性检测由于其基本操作单位为图像像素，而一般的应用任务（比如物体识别、物体检测等）的基本单位都是物体，所以在实际的视觉任务基于视点的视觉显著性技术较难发挥作用。

之后自 Gofeman[28]起，研究者们提出了显著性物体检测（Salient Object Detection）的概念，在该类任务中，相对于建模出人类的视点规律，我们需要找出在图片中最为显著的物体。这样的话，就大大推进了在应用任务中使用显著性检测的可能，最近新兴的领域即非监督物体检测[5][86]就极大的利用了显著性物体检测的结果。另一方面，近年来还出现了另一个相关的工作，即通用物体检测（Objectness）[1][2]。在通用物体检测中，我们任务并不是找出显著的物体，而是找出所有物体的候选框（Proposals），即判断某一个框（Bounding Box）内的图像是否为一个物体。在这里我们仅适用显著性物体检测技术来进行深度学习研究，下面我们将介绍两种使用视觉显著性的方式，这两种方式将贯穿并指导我们算法的建立。

将视觉显著性当成任务先验

图 2-24 显著性物体检测示意图

在利用视觉显著性先验的方式中，最为直接的即为将检测到的显著性物体作为任务的先验信息（Prior Information）。假设我们有图像 I 与 I 的视觉显著性检测结果 S，如图 2-24 示意。则对于图像上的任意关键区域 R，我们可以根据视觉显著性的结果得出 P(R|I)，即对于图像 I，R 含有物体的概率。

$$P(R|I) = \frac{\sum_{x,y \epsilon R} S_{xy}}{\sum_{x,y \epsilon I} S_{xy}}$$

假设我们现在需要对图像 I 做任务 T，那么根据概率公式我们可以得到：

$$P(T|I) = \sum_R P(T|R,I)P(R|I) \approx \sum_R P(T|R)P(R|I)$$

上式中我们使用一般假设，即 T 和 I 给定 R 时条件独立，则对于 I 图像上任务 T 的概率为 I 上所有 R 的概率加权和。从上式我们不难看出一个问题，当 R 为全图时，P(R|I)=1 达到最大，但此时也相当于完全没有利用显著性物体检测的先验。所以在建模时，我们不光要考虑到某个候选框中含有物体的概率，还要注重候选框的紧致型。我们可以选择 N 个固定尺度进行测试，即：

$$R_n^{max} = \text{argmax}_{R_n} P(R_n|I)$$

$$P(T|I) \approx \sum_{n=1}^{N} P(T|R_n^{max})P(R_n^{max}|I)$$

其中R_n^{max}表示在第 n 个尺度下相应最大的区域。或者我们可以为整个任务的目标加取关于 R 的正则项，用来对过大的 R 进行惩罚，比如以下的正则项：

$$\phi(R) = \text{var}(R)$$

通过将正则项φ 定义为 R 内值的方差，我们就可以简单的对 R 的紧致型进行约束，即要求 R 的方差尽量小。

相对于定义候选框 R，有一种更为简单与方便的方式来利用显著性先验，即我们可以将像素当成我们处理的基本单位。注意这里使用像素作为基本单位并不意味着我们将物体显著性当成了视点数据，而只是我们缩小了处理单元，在整体上整个检测结果还是体现出显著性物体的特性。

$$P(T|I) = P(T|X) \quad s.t. \ X = \{x|x\epsilon I, P(x|I) > \alpha\}$$

其中 P(x | I)为对于图像 I，像素 x 是显著性物体一部分的概率，α 为显著性的截断值，即仅利用显著性概率大于α 的像素信息。在之后的建模中，我们将详细介绍如何利用这种先验。

将视觉显著性当成额外信息

从另一个方面来说，视觉显著性物体检测除了当成任务的先验以外，显著性物体检测的结果本身也提供了相当的信息。即我们可以转换思路，除了利用源图像信息进行视觉任务外，还利用显著性检测结果进行任务的辅助，也就是说把视觉显著性当成额外的数据来使用。

假设我们有图像 I 以及其显著性物体检测的结果 S，那么对于任务 T 来说：

$$
\begin{aligned}
&P(T \mid I,S) \\
&\quad \propto P(I,S \mid T)\,P(T) \\
&\quad \approx P(I,S \mid T) \\
&\quad = P(S \mid T)\,P(I \mid S,T) \\
&\quad \approx P(T \mid S)\,P(T \mid I)\,P(S \mid I)
\end{aligned}
$$

在上式中，P(T|I)为给定源图像任务 T 的概率，P(T|S)为给定显著性图后任务 T 的概率，而 P(S|I)则为显著性检测的置信概率。这样的话即是说除了利用源图像 I 的信息处理任务 T 外，我们还需要利用显著性检测的结果 S。

相对于上一种方式中，显著性物体检测的结果 S 通过关键区域 R 或是像素 x 与任务 T 关联，在这里我们使得显著性检测结果 S 与任务 T 直接关联。这样带来的好处是我们可以分别为源图像和显著性检测的结果对任务 T 进行建模，模型的灵活性大大提高了；当然，带来的问题就是参数空间也会进一步增大，如果不善加控制的话很可能在小数据集合上出现过拟合现象。

2.3.2 深度学习模型

在前文中我们提到，在全连接网络的深度学习中，一般分为两个步骤：第一步为网络的参数初始化，在初始化中我们通过非监督生成式模型进行贪婪逐层学习，从而避免深层网络中的误差消失现象；第二步为监督学习，即在网络的初始化权值之下，利用已标注的任务相关数据进行监督学习，调优（Fine Tune）整个神经网络的参数。在第一步的基础之上，我们可以在监督学习中得到良好的误差极小值点，而且这种端到端（End-to-End，即从最原始的数据到最终的任务）的学习被证明可以较大幅度提升性能。这也是深度学习模型在许多数据集上都取得领先的原因之一。

在近年来，随着卷积网络的重新兴起，很多工作指出在训练网络时不需要进行参数初始化也可以达到很好的效果。关于这一点，我们指出非监督参数学习并非毫无用武之地：首先，卷积网络是神经网络中一个很特殊的例子，它的结构可以保证其在误差反传时可以很好的“保存”误差梯度；而在全连接网络中，深层模型的训练至今都还必须依赖于非监督逐层学习。再者，对于纯监督学习的模型来说（不管是卷积网络还是全连接网络），由于参数量远超于一般的机器学习模型，我们都需要大量的标注样本来完成模型的训练，而标注样本会耗费大量的人力、物力、财力，使得任务代价骤升；合理利用无标注样本进行无监督或半监督学习无疑是很好的解决之道。最后，在当前的大数据时代，人们提出了“4V”的概念，即容量（Volume）、时效（Velocity）、多变种（Variety）、

真实性（Veracity）；在这种前提下，可以说完全依靠人力的标注任务已经不可能完成，从业界领先的谷歌等的工作[14][44]来看，非监督、半监督学习在当前有着广阔前景。

下面我们分别对于上一节的两种方式进行深度学习的建模，2.3.2.1 节主要描述将视觉显著性信息当成先验的方式，2.3.2.2 节主要讨论将视觉显著性信息当成额外相关信息的方式。在建模的过程中，由于我们的建模的方式使用自动编码机较为便捷，在以下的模型中我们全部使用自动编码机作为基本模型。

2.3.2.1 前景建模的深度学习模型

在这一小节中，我们将利用视觉显著性信息当做任务先验来进行建模，由于我们在利用先验时等同于着重利用前景信息，所以称该方法为前景建模的深度学习。首先我们简述在自动编码机中的求解细节：在自动编码机中，我们将信号通过线性映射与激活函数映射到隐层空间，之后再反向传播在源数据空间进行重构，之后我们来求解优化问题使得重构结果与原样本尽可能相似。

在这里，我们不难发现在进行样本的重构时，我们是针对整个图像来进行操作。而我们发现在物体分类的任务中，前景信息（比如物体形状、纹理等）实际上是一个有限集合；而背景信息则是趋近于无穷的。对于这一点我们可以这样理解，在做物体分类时，实际上我们需要的数据仅仅是前景信息。虽然背景信息在许多利用上下文（Context）建模的方法中被证明是有利于分类任务的[64[65]，但我们认为这种利好倾向仅限于一些特定的数据集。举例来说，一辆汽车不管是在草坪里、还是在室内、在公路上、甚至被吊在空中，物体的类别信息并没有改变。所以我们认为过分依赖背景信息的学习在某种程度上将不利于进行学习的泛化（Generalization），在背景的利用上需要进行权衡。

基于这种假设，我们在建模的时候，要求深度学习网络针对样本内的前景信息来进行重构，而这种前景信息来自于视觉显著性物体检测。假设我们有样本 I，其显著性检测结果 S，以及经过自动编码机重构的样本 R。在原始情况下，我们有目标函数：

$$\min ||\mathbf{R}-\mathbf{I}||_2^2$$

在这里，由于样本 I 为自然图像信息，即 I 为实值样本，所以在目标函数中我们直接采用二范数距离进行计算。在这里我们可以很容易地加入前景先验，假设对于显著性检测结果 S，约显著的像素在 S 的值中越高，那么我们可以简单的定义目标函数为：

$$\min ||\mathbf{S}\cdot(\mathbf{R}-\mathbf{I})||_2^2$$

其中，“·”运算符表示矩阵对应位置相乘的运算。在该目标函数中，我们直接利用

显著性检测结果 S 对误差函数进行加权，即我们认为越显著的目标（S 越大的像素）其为前景的概率越大，故我们着重对齐进行重构。在实际上，由于显著性物体检测的结果并不是完美的，我们发现完全抛弃显著性检测算法认为的“背景”并不能达到很好的效果，所以我们对显著性检测的结果进行拉伸，即为完全不显著的区域留有一定权重，并重新归一化到[0-1]范围：

$$\hat{\mathbf{S}} = (\mathbf{S} + \varepsilon)/\max(\mathbf{S})$$

在这里ε 为一个较小的值，用于作为整个图像的基底值。在实践中，我们发现当ε =0.2 时可以获得较好的结果，在之后我们都使用这一值来进行相关的实验与测试。

2.3.2.2 显著性建模的深度学习模型

1) 基础模型

在这一节中，我们将从另一个角度，即直接将视觉显著性物体检测的结果当成与源数据相关的额外数据来进行利用，由于这里涉及到了显著性的建模，所以我们称为显著性建模的深度学习模型。在这里，我们知道，显著性信息与源图像有很强的相关性；在显著性物体检测的结果中，我们发现显著性图提供了原图所不具有的一些信息，比如物体的位置、部分轮廓等等（如图 2-25）。

图 2-25 CIFAR-10 显著性检测示意图

我们在进行模型建立时，要求模型从原始数据出发，但重构回两类数据：原样本与原样本的显著性检测结果（如图 2-26）。之所以要这样进行建模，而不是将显著性信息和原始样本都当成输入样本，是因为在这里我们着重利用显著性信息进行模型的非监督学习，即我们保证所有的操作是在统一规模的网络上进行的。而将显著性信息当成额外的维度实际上增大了样本规模，从而使得我们难以清除的说明模型的结果是来自算法还是来自扩大的模型。在这一点上，在全文的模型和实验上，我们都保证模型规模的一致，并且仅仅利用显著性信息进行非监督的模型参数初始化，以此来说明显著性检测的作用。

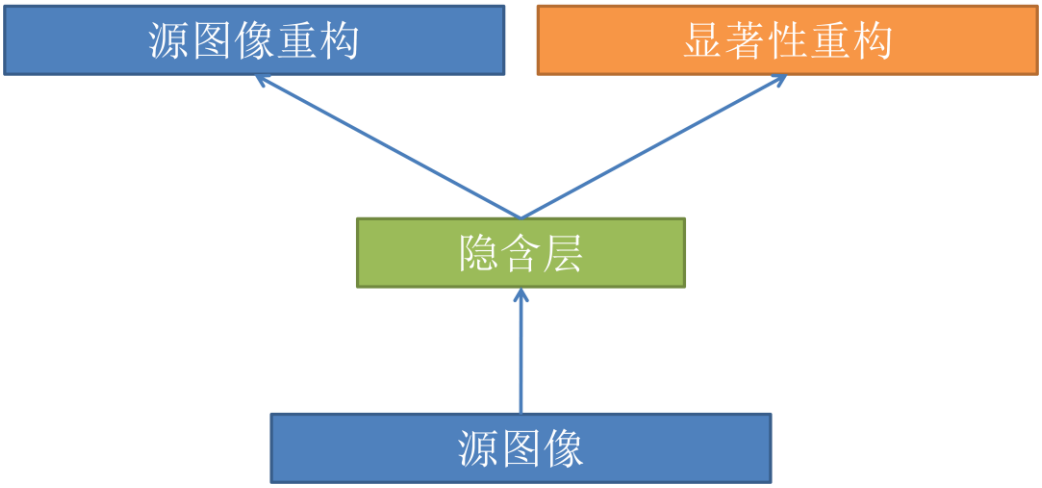


图 2-26 显著性重构示意图

在这里，我们通过重构回两个源数据，迫使自动编码机在隐含层学习更多的信息，即图像与显著性的两种信息。假设我们有样本 I，显著性物体检测的结果 S 以及样本重构结果 R_I、显著性物体重构结果 R_S，那么我们有目标函数表达式：

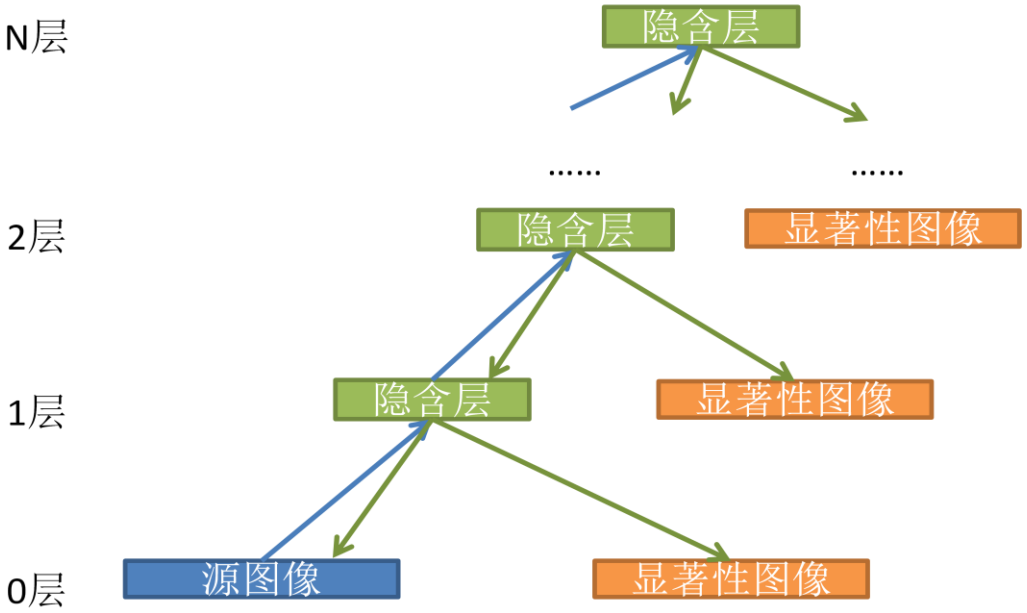
$$\min ||R_I - I||_2^2 + \lambda||R_s - S||_2^2$$

其中λ 为权衡变量，即λ 越大模型越偏重与对显著性检测结果进行重构，反之则偏重对源图像进行重构。从上式中可以看出，我们也可以将显著性的重构结果当成是模型的正则项（Regularization）来看待；而从正则项的角度解释的话，我们加入了显著性信息的重构，可以在模型学习中控制模型学习，避免发生过拟合现象。在实践中，我们发现显著新信息的重构对物体表达的学习有很好的指导作用，所以我们偏向于给与一个比较大的λ 值，在实验和测试中，我们都设λ =1。

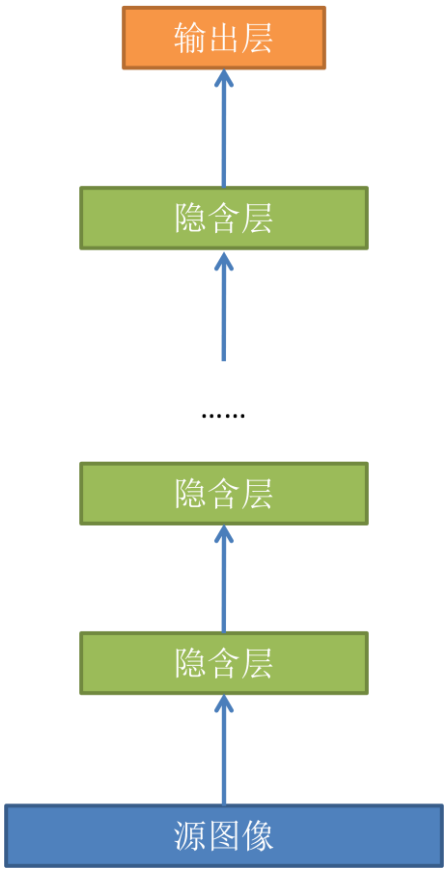
2) 迭代学习

在上面描述的利用视觉显著性信息的深度学习建模中，都有一个共同的问题：即使用显著性信息进行非监督学习时，其直接影响的数据只有第一个隐含层。也就是说在第一个隐含层的学习中，我们利用视觉显著性信息，而之后的话我们只是很“间接地”使用了这类信息。并且从实验中我们也发现，在加入视觉显著性信息之后，深度网络中前两层学到的特征基本上维持不变。其中的原因主要为：在深度神经网络的学习中，我们一般认为越上层的特征越抽象越高级，而越底层的特征越简单越基础。在这里底层网络由于特征较为简单，不能很好的利用视觉显著性物体检测的结果。所以我们需要着重讨论与解决的关键问题即是，如何进行特征的深层次迭代学习，并在每一次迭代中都利用到视觉显著性的信息来增强深度非监督学习的效果。

我们观测到，在引入视觉显著性物体检测的结果后，第一个隐含层所含有的物体表达与仅适用原始样本相似，即视觉显著性的信息并不能很好的在第一个双层网络处被利用。所以我们可以简单地将约束扩展到更多的层次，即在每一层训练自动编码机时加入视觉显著性信息重构的正则项（如图2-27(a)）。如图2-27(b)为监督学习与测试时的结构。



(a)

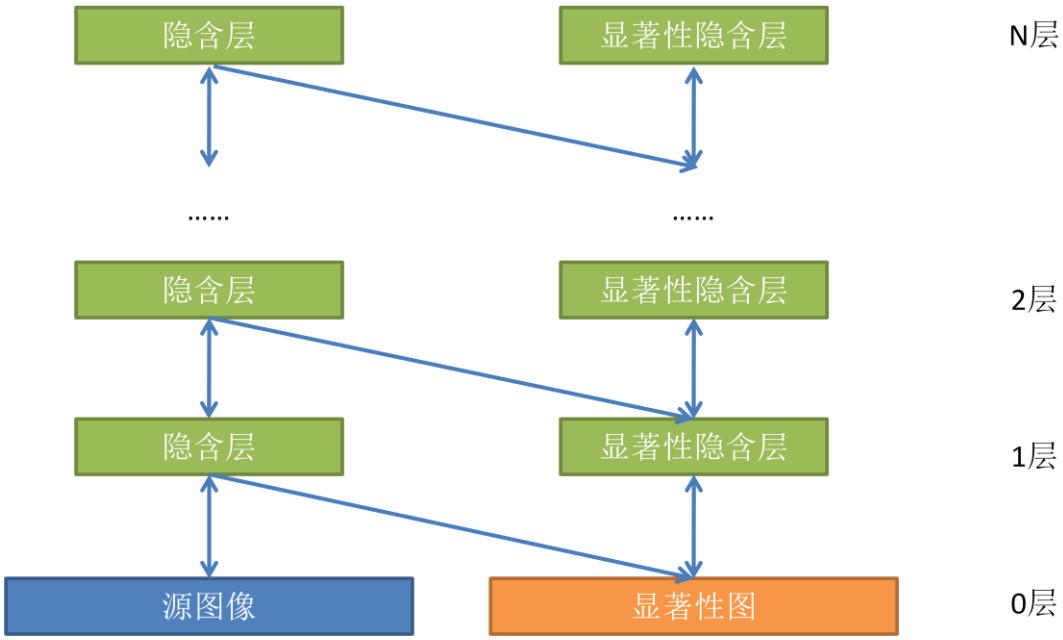


(b)

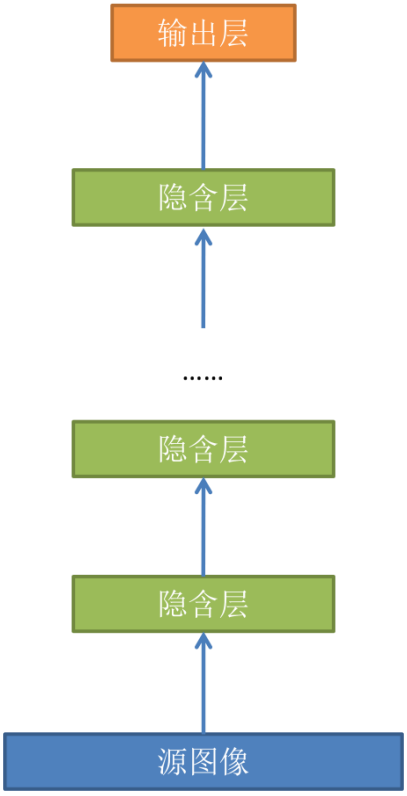
图 2-27 显著性重构层次迭代示意图，其中图(a)为非监督学习训练结构，(b)为监督学习与测试时的结构，注意我们仅在非监督学习中引入显著性信息，其余情况仅适用原图像

在这种逐层的约束下，整个神经网络可以较好的从视觉显著性信息中获益，在实验

中也有明显的性能提升。然而我们发现这其中还有一些的问题，在神经网络中我们可以单纯的认为有连接的项即存在某种相关关系，那么在以上的模型中，实际上我们每一层的隐含层信息都需要与视觉显著新物体检测的结果有直接关系，也即是第 i 层的隐含层与第 i-1 层隐含层与视觉显著性结果相关。我们前文分析过，在浅层次的时候，由于提取的特征较为基础，自动编码机不能很好的利用显著性信息；但当我们认为训练到深层次的时候，这个问题会发生了翻转，即深层学到的特征已经太过抽象而不适于进行视觉显著性信息的重构。所以我们对原始模型进行修改，即除了进行源数据的抽象学习外，还进行视觉显著性信息的学习（如图 2-28(a)）。在这里，由于我们的模型中有两个相互相关的基础自动编码机模型，我们将该算法成为双路重构的自动编码机，并将基于源数据的自动编码机叫做主编码机，将基于视觉显著性图的自动编码机叫做辅助编码机。



(a)



(b)

图 2-28 显著性重构双路编码示意图，其中图(a)为非监督学习训练结构，(b)为监督学习与测试时的结构，注意我们仅在非监督学习中引入显著性信息，其余情况仅适用原图像

在上图（图 2-28(a)）中，我们可以看出，对于主编码机的第 i 层隐含层，我们需要

重构的是主编码机的第 i-1 层的隐含层以及辅助编码机的第 i-1 隐含层。假设我们用 H_i 表示主编码机第 i 隐含层响应，R_i 表示主编码机第 i 层对图像的重构结果，RD_i 表示主编码机对辅助编码机第 i 层的重构结果；I 表示辅助编码机第 i 隐含层的响应，RS_i 表示辅助编码机第 i 层的重构结果。则对于第 i+1 层的双路编码机，我们可知其的目标函数为：

$$\min\|R_i - H_i\|_2^2 + \lambda\|RD_i - HS_i\|_2^2 + \|RS_i - HS_i\|_2^2$$

其中目标是的第一项与第二项为主编码机的表达式，第三项为辅助编码机的表达式。在主编码机中，第一项为自身的重构，第二项为对辅助编码机的重构结果；即主编码机依赖辅助编码机的结果，而辅助编码机其实是独立发展的，并不与主编码机相关，这也是他们名称的由来。

在测试时（如图 2-28）b))，我们丢弃辅助编码机仅仅使用主编码机使用误差反传算法进行学习，这么做的原因一是在测试中我们默认为输入的图像没有显著性检测的结果；二是方便与基准方法对比，证明加入显著性后无监督学习的有效性。

在试验中，我们发现使用双路编码机可以更好的利用显著性信息，使得非监督学习的效果进一步提升。而对于目标中的λ 值，我们发现随着层次的增加，在信号重构时来自辅助编码机的信息会渐渐变少。所以我们在第一层将λ 设为 1，并在以后的学习中，每一层将λ 值进行递减。试验中我们发现$\lambda = \lambda - 0.1$ 可以获得很好的性能。

2.4 实验和分析

2.4.1 可视化（Visualization）

2.4.1.1 连接权值可视化

在全连接网络中，我们可以通过对连接权值可视化来较直观的查看学到的特征。假设我们有连接权值 **W**，且我们已知

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{v} + \mathbf{b})$$

则我们可以将权值 **W** 的每一行看成是在输入空间（即 **v** 的空间）上的加权，所以将 **W** 按行进行可视化即可以了解网络所包含的特征类型。在这里如果我们将权值看做 8 比特图像进行可视化（范围 0-255），那么我们还需要对数据的范围进行压缩，因为权值 **W** 很可能有完全不同的取值范围。假设我们取出 W 的第 i 行 $\mathbf{w}^i$，则简单的可视化结果

I 为：

$$\mathbf{I}_{\mathrm{k}} = \frac{\mathbf{w}_{\mathrm{k}}^{\mathrm{i}} - \min(\mathbf{w}^{\mathrm{i}})}{\max(\mathbf{w}^{\mathrm{i}}) - \min(\mathbf{w}^{\mathrm{i}})} * 255$$

其中，$\mathbf{w}_{\mathrm{k}}^{\mathrm{i}}$表示 $\mathbf{w}^{\mathbf{i}}$ 的第 k 个元素，$\max(\mathbf{w}^{\mathbf{i}})$与 $\min(\mathbf{w}^{\mathbf{i}})$分别代表 $\mathbf{w}^{\mathbf{i}}$ 中所有元素中的最大、最小值。注意在这里我们按照 w^i 的范围将数据归一化到了 0~255，从而使得可以使用 8 比特图像进行显示，其中越亮的像素表明其权值越大，越暗的权值越小。这种方式可以很方便的进行可视化，但有其固有的局限性：在这种可视化方式中，由于权值可能为正也可能为负，在归一化后我们实际上“偏移”了零点，使得在可视化图中无法判断权值的正负。我们可以采用另一种方式，将零点固定在中灰（127）：

$$\mathbf{I}_{\mathrm{k}} = \left(\frac{\mathbf{w}_{\mathrm{k}}^{\mathrm{i}}}{\max(\mathrm{abs}(\mathbf{w}^{\mathrm{i}}))} + 1\right) * 128 - 1$$

这样所有的负权值将分布在 0-128 范围，而正权值将分布在 128-255 范围。在后文所有的权值可视化结果中，我们都将使用这种方式。

2.4.1.2 深层连接权可视化

在上一节中，我们介绍了权值可视化的方式，然而这种方式只是将权值映射在输入数据空间中，对于深层网络其输入数据的空间不是图像空间（可视空间）也即是权值即使映射在维度相同的空间中仍然不能可视化。

在这里我们假设有一个三层网络，则可知：

$$\mathbf{h_1} = \sigma(\mathbf{W_1 v} + \mathbf{b_1})$$

$$\mathbf{h_2} = \sigma(\mathbf{W_2 h_1} + \mathbf{b_2})$$

即对于第二层的权值 $\mathbf{W_2}$ 来说，其输入空间为 $\mathbf{h_1}$ 所在的空间；而对于 $\mathbf{h_1}$ 来说，$\mathbf{h_1}$ 的每一个维度可以看作是 **v** 在 $\mathbf{W_1}$ 的映射下的响应。假设我们可以去掉非线性函数 σ，则可知：

$$\tilde{\mathbf{h}}_\mathbf{1} = \mathbf{W_1 v} + \mathbf{b_1}$$

$$\tilde{\mathbf{h}}_\mathbf{2} = \mathbf{W_2 h_1} + \mathbf{b_2}$$

那么对于 $\mathbf{W}_2$ 的某一行 $\mathbf{W}_2^{\mathrm{i}}$，我们可以近似的将其映射回 v 所在的数据空间，即令：

$$\mathbf{W}_2^{\mathrm{i}} \approx \mathbf{W_1 v} + \mathbf{b_1}$$

我们得到：

$$\mathbf{v} = \mathbf{W_1}^{-1}(\mathbf{W_2^i} - \mathbf{b_1})$$

又由于在自动编码机中，编码与解码由同一个矩阵 $\mathbf{W}$ 与其转置 $\mathbf{W^T}$ 做线性映射，故可以用 $\mathbf{W^T}$ 来近似逆运算，即：

$$\mathbf{v} \approx \mathbf{W_1^T}(\mathbf{W_2^i} - \mathbf{b_1})$$

至此，我们获得了深层网络的近似可视化结果。

2.4.1.3 样本重构的可视化

在上文中我们提到，在深度学习中权值的初始化是通过数据重构来实现的。所以在深度学习中，除了连接权重以外，样本重构结果的可视化结果是反映学习结果另一个的重要要素。以自动编码机为例，我们可知：

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{v} + \mathbf{b})$$

$$\mathbf{r} = \varphi(\mathbf{h^T}\mathbf{W} + \mathbf{c})$$

在这里我们同样也会遇到数据的归一化问题，但与权值不同的是由于图像是连续变化的，我们并不需要确切的知道“零点”，一些正向或负向的偏移并不影响可视化结果。所以在可视化时，我们直接将重构结果归一化至 0-255 之间进行显示。

2.4.2 全连接网络模型规模

在测试中，为了突显非监督学习的作用，我们选用了 4 层网络，其原因为：第一，如果层数过少，误差消失现象不显著，从而导致监督学习即可很好的解决问题；第二，如果层数过多则会导致训练不充分，即非监督学习的初始化也难以发挥网络的性能。并且我们尝试了 3 中不同大小的模型规模，其分别为（格式为输入层-隐含层-隐含层-隐含层-输出层）：

（1） 小规模： 3072-500-500-1000-10

（2） 中规模： 3072-2000-1000-2000-10

（3） 大规模： 3072-3000-3000-4000-10

在所有的测试中，我们使用 ReLU（Rectified Linear Unit）激活函数来协助我们在深层网络中获得更好的性能。在 CIFAR-10 数据集中，图像样本均为 32*32 的 RGB 图像，故输入层为 3072 维；而 CIFAR-10 中共有 10 个类别，故输出层大小为 10。在 STL-10 中的情况也与次类似，我们仿照文献[15]的方式，将图像缩放到 32*32 时再做测试，这样的话参数就得以与 CIFAR-10 数据集保持一致。在测试 STL-10 数据库时，我们没有

按照 10 折（fold）的方式来分配训练数据，而是使用了全部的训练数据，这是由于全连接网络在样本少时表现过差，不足以体现网络特性。在以上三种规模下，我们预先使用原始的降噪自动编码机在 CIFAR-10 数据库与 STL-10 数据库上进行测试，并在每个规模的网络内尝试不同的正则项参数以避免模型过拟合。我们在三个模型上测试的最好结果如图 2-29：

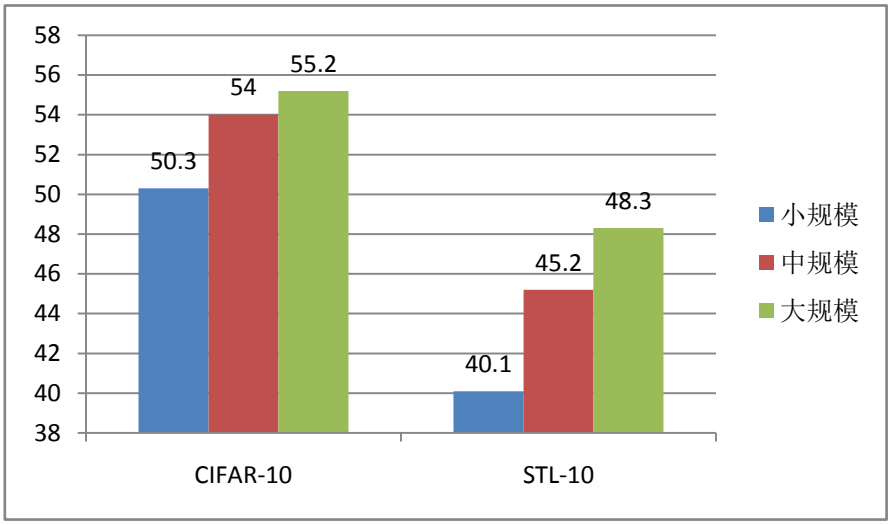


图 2-29 模型规模测试结果图

在该实验中，我们发现对于小规模网络而言性能差距略大，而中规模与大规模网络的差距较小，即已经开始收敛；另外，大规模网络训练时间比中规模约长 1 倍，所以在之后的试验中，我们都选取中规模模型的参数进行测试。规模选择的作用一是确定全连接网络模型在数据集上表现的大致趋势，从而得以较好的估计加入显著性信息后提升的幅度是否显著；第二则防止因网络规模过小的欠拟合状态，因为在这种状态下模型很难对正则表达有效果，加入正则有可能还会导致性能下降，与我们实验的目的不符；第三则是最大可能的削减模型复杂度，避免在深度学习试验中耗费过长的时间训练。

2.4.3 前景建模的深度学习

在前景建模的深度学习模型中，我们使用视觉显著性信息进行粗糙的前景加权，并按照加权值进行数据重构，即强制神经网络对前景信息进行学习。在显著性物体检测环节，我们采用当前领先的测地显著性检测[79]方法，选择该方法的原因其一为其对于图片的普适性较好，性能优异；其二为速度快，对于一张图片的显著性检测只用 2ms 左右时间。

在测试中，由于 CIFAR-10 与 STL-10 的图像都较小（CIFAR-10 32*32， STL-10 96*96），我们在显著性检测时的图像块大小分别设为 1 和 3，即最终图像保持 32*32 个图像块，并且我们初始化权值为 0.2，噪声值阈值为 0.01，在试验中该参数表现良好（如图 2-24）。

在前景建模的深度学习中，分别在 CIFAR-10 与 STL-10 上进行测试，在所有测试中我们都采用了上一小节中中规模的网络，并采用其达到最好时的训练参数进行测试。在 STL-10 的测试中，为了更好的获取视觉显著性图，我们从高分辨率（96*96）下进行显著性图的获取，但在训练网络的时候我们将源图像与显著性图都缩放至 32*32 以便与前文中的测试相一致。我们在测试中求取的是 5 组测试的平均值，其结果如图 2-30。

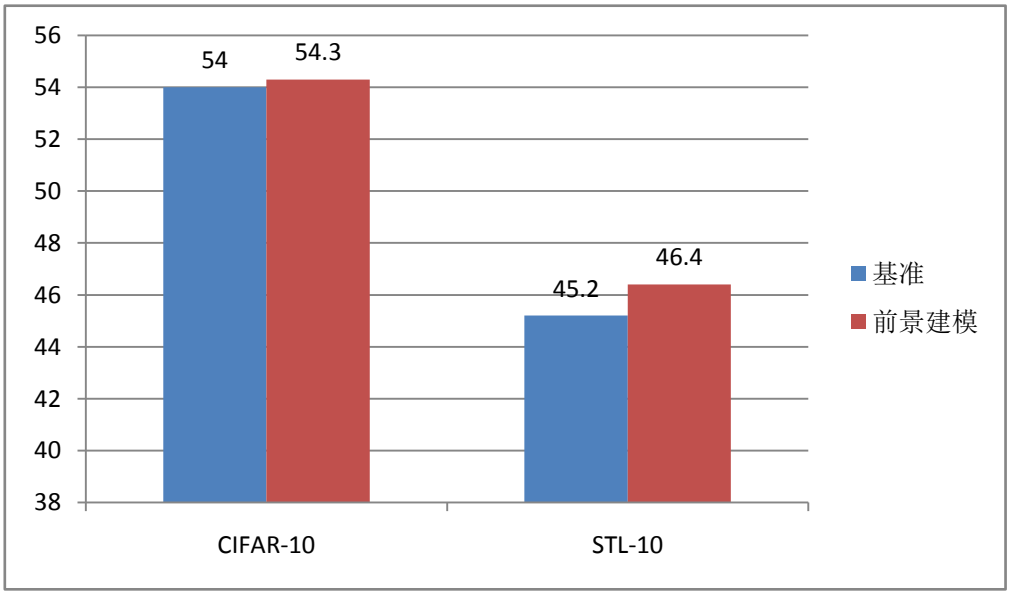


图 2-30 前景建模深度学习结果图

在测试结果中我们发现，加入前景建模之后在 CIFAR-10 与 STL-10 中的效果都略有提升，但提升幅度都比较有限：在 CIFAR-10 中准确率提升只有 0.3%，几乎可以忽略不计；而在 STL-10 中有 1.2%的性能提升。这种差异的区别主要是 CIFAR-10 数据库较多依赖监督学习，而在 STL-10 中由于绝大部分样本都为无标注样本，所以在整体的训练学习中也就比较注重非监督学习的参数初始化，所以在 STL-10 中性能提升较为显著。由于这两个数据库的这个特性，这种类似的结果（即 STL-10 提升较大）还会在之后的试验中不断重现。

在实验中我们可视化出第一层学到的特征如图 2-31。在图中我们发现，在加入了显著性前景建模之后，第一层学到的特征较基准测试中学到的特征来说更趋向于局部，并

且有一些轮廓响应（如图 2-31 中红框部分）；相比之下，基准的自动编码机在 RGB 图像上学到的特征相对比较杂乱，也趋向于全局响应。

图 2-31 前景建模深度学习第一层特征可视化图，左图为基准自动编码机特征，右图为加入显著性前景建模后自动编码机特征响应。

在第二层学到的特征中（如图 2-32），我们发现不管是基准自动编码机还是加入了显著性前景建模的自动编码机，学到的特征都开始具有中央偏置（Center-Bias）的特性。中央偏置在显著性建模中有过广泛的讨论[74]，即许多数据集中的图像都倾向于在图片中心分布，这其实反应了一般图像捕捉的主体中心原则，但在更为泛化的概念中则不利于算法的扩展。为此我们回顾了所用的两个数据集 CIFAR-10 与 STL-10，在这两个数据库中我们都发现了明显的中央偏置倾向，所以在加入了显著性建模后学到的特征确实应该具有中央偏置，但令我们惊讶的是没有加入显著性信息的基准自动编码机似乎也可以自发的意识到这一点，其学到的特征都位于中心。在这里，我们发现显著性前景建模的自动编码机明显相对于基准自动编码机趋近于局部，获得了许多较为具体的特征；而在基准自动编码机中特征则趋向于大范围的结构，也较为模糊。

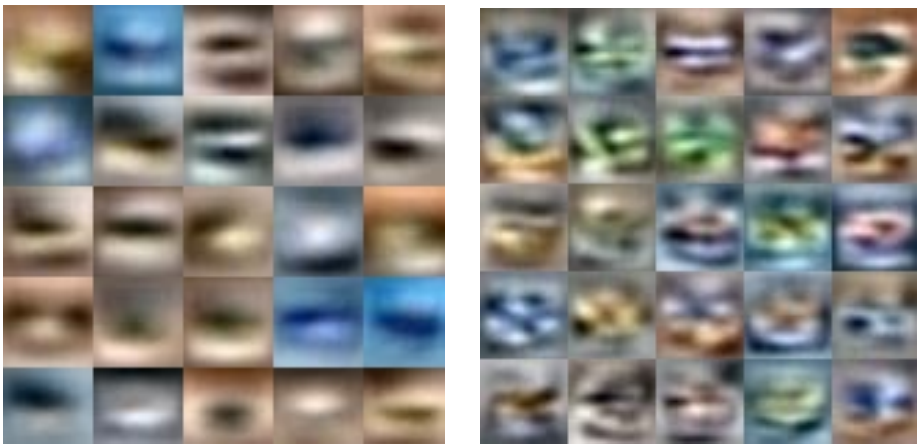


图 2-32 前景建模深度学习第二层特征可视化图，左图为基准编码机特征响应，右图为加入显著性前景建模后自动编码机学到的特征响应

在上文中，我们探讨过深度学习神经网络中权值可视化的趋势，即随着层数的加深可视化的准确程度会逐渐下降，所以在权值的可视化中，我们只对前两层进行分析。在本方法中，我们可以看到我们提出的算法在非监督学习阶段学得的特征与基准自动编码机的特征还是有比较明显差别的，但该种方法主要的问题是显著性检测信息在自动编码机中的影响只有第一个隐含层的特征，虽然这些特征也会间接影响到之后的层次，但从分类准确率来看我们获得的收益较为有限。

2.4.4 显著性重构建模的深度学习

在显著性建模的深度学习中，我们采取另外的方式利用显著性信息，即相对与利用显著性信息进行显式的前景建模，我们将视觉显著性物体检测的结果当成是额外的数据进行利用，让深度网络来自发的进行相关信息的建模。

在该算法中，除了直接在第一隐含层对显著性信息进行建模外，我们提出了双路自动编码机，即在显著性信息层与源数据层同步进行自动编码机的建模，并在源数据的自动编码机（我们成为主编码机）中利用当前层次的显著性信息（即辅助编码机的信息）进行辅助建模，具体的模型信息可以回顾 2.3.2.2 节内容。我们分别针对 CIFAR-10 与 STL-10 进行测试，在测试中我们在辅助编码机中利用规模为 1024-800-800-800 进行显著性建模，其结果如下（图 2-33）：

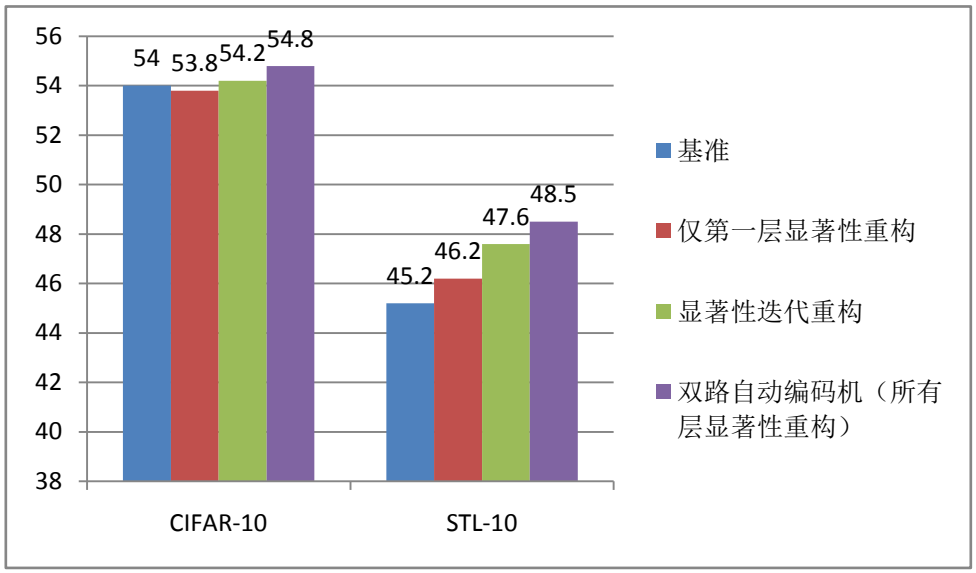


图 2-33 显著性重构建模深度学习结果图

在试验中我们可以看出，对于 CIFAR-10 数据库来说，利用显著性重构获得的收益依然十分有限；但在偏重于非监督学习的 STL-10 数据库中，我们利用双路自动编码机可以获得接近 3%的性能提升，甚至超过了使用大规模模型的基准自动编码机。同时我们注意到，仅利用第一层进行显著性重构的结果与上一小节中我们利用显著性建模前景的算法得到的结果比较类似，说明了单层利用显著性信息的局限性，同时也说明我们双路建模的有效性。

在显著性重构的深度学习中，我们模型中的参数大大增加了，相对于之前只有 1 组参数（即自动编码机建模的参数），这里有 3 组参数，分别为：1).主自动编码机的建模参数，2). 主自动编码机重构显著性的参数，3). 辅助编码机的建模参数。虽然对于同样的数据来说，我们的模型参数有所增加，但我们认为这并不意味着我们的扩展性会相对下降，反而在很好的利用了显著性之后我们获得了性能的提升，其原因如下：首先，从模型内部相关性来看，我们的模型中主编码机的学习需要辅助编码机的帮助，但对于辅助编码机来说其建模过程则是独立的，鉴于我们在之后的监督学习中会抛掉其余部件仅留下主编码机初始化的权值进行学习，在这里我们模型规模并没有大幅度增长；再次，在主编码机建模时，虽然我们增多了一组参数来进行显著新重构，但在这里我们也引入了新的数据，即显著性检测的数据，所以在良好的正则约束下我们不仅可以避免参数增加带来的拟合问题；最后，关于参数个数、数据个数与模型推广性的关系，并不能单纯建立联系，事实上模型拟合的程度与建模方式和初始化方式有很大的关系，一个符合数

据分布规律的建模往往能达到更好的拟合效果。

为了更深入的发现加入显著性重构后对自动编码机的影响，我们对参数进行可视化与相关分析：

（1） 主编码机建模参数

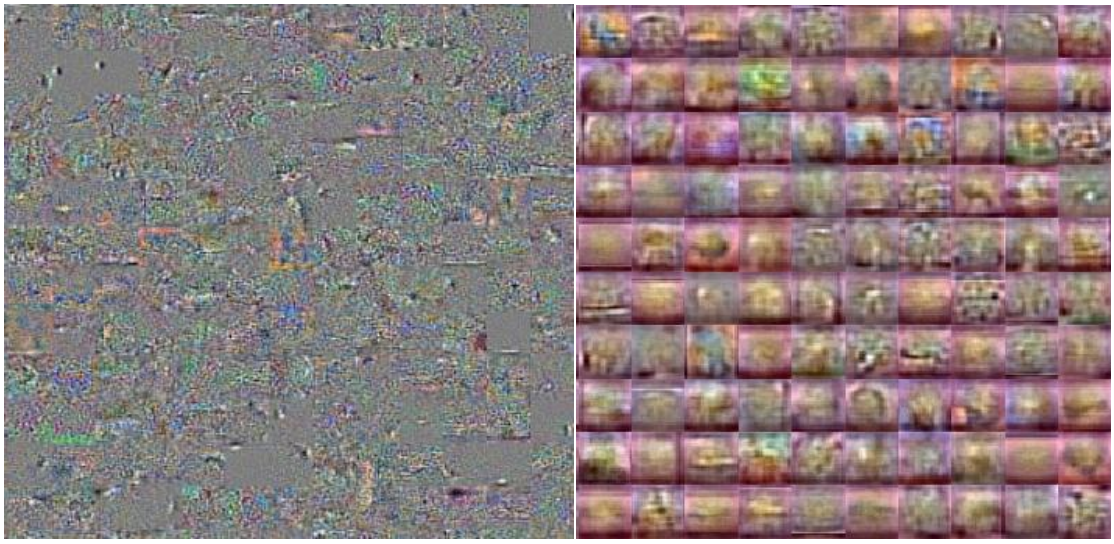


图 2-34 双路编码机主编码机特征可视化图，左图为第一层特征，右图为第二层特征。

在主编码机的参数可视化如图 2-34。在加入显著性重构后，我们明显发自动编码机第一层的特征趋向于局部化，并且在其中发现了明显的类 Gabor 特征，这是一般公认的人类视觉皮层 V1 的特征信息，具有良好的描述和扩展能力。在第二层的可视化参数中，我们发现其可视化的结果为一些局部特征，并在其中可以看出一些物体的部件特征。

（2） 主编码机重构显著性参数

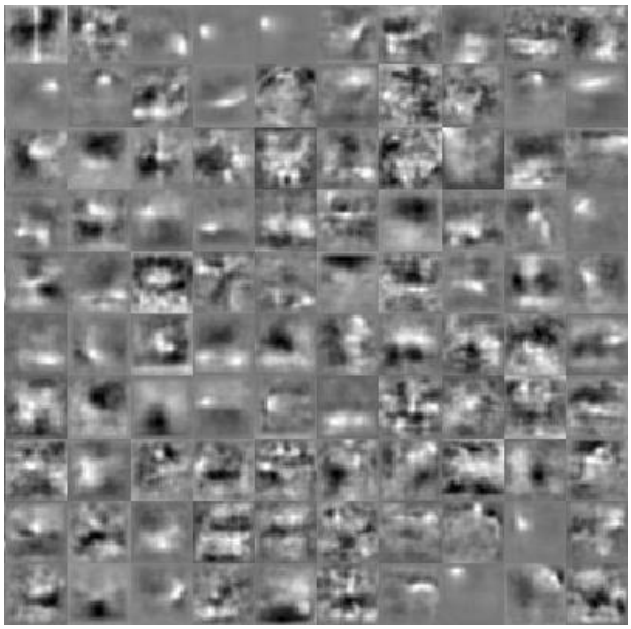


图 2-35 显著性重构参数可视化图

在主编码机的显著性重构参数中（如图 2-35），其特征也趋向于局部，在其中也可以看到一些物体部件的特征信息。在这里由于我们的训练样本仅限于10类的物体图像，所以在进行显著性建模时该显著性明显带有了这 10 类物体的特征，并不具备通用的特性。除此之外，在显著性重构的深度学习中，我们可以对样本的显著性重构结果进行可视化，如图 2-36。

图 2-36 显著性重构结果图

在显著性重构中，我们发现重构的结果优异。注意在结果中，有许多样本重构的结果实际上已经超越了原始显著性物体检测的结果。在这里由于我们在进行显著性建模时已经利用了所有训练样本的显著性信息，即在模型中已经有了训练库中类别的相关轮廓与形状的建模，所以在进行显著性重构时得以在原有显著性物体检测的基础上学的更好的结果。但就如上文所说，在我们模型中的显著性检测并不具备通用性，我们引入显著性检测的目的也是为了辅助深度学习过程对轮廓结构的学习，并不是为了单纯的显著性检测工作。

（3） 辅助编码机建模参数

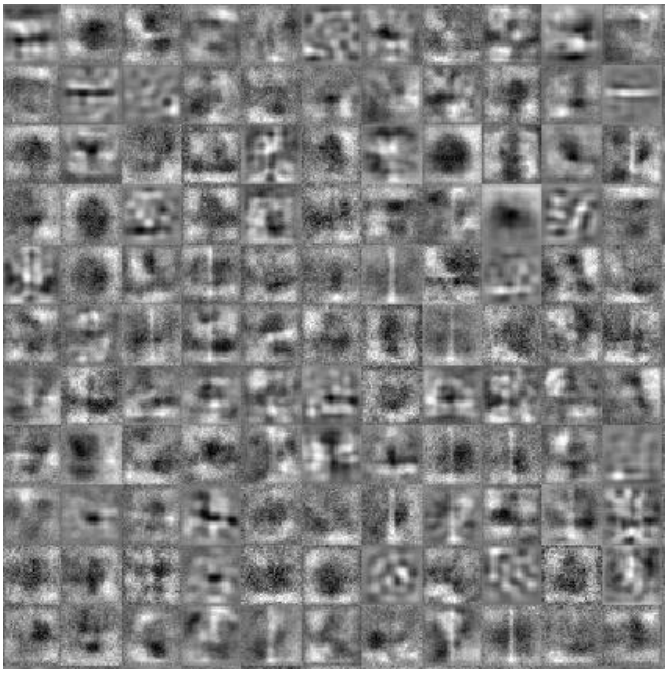


图 2-37 辅助编码机参数可视化图

在辅助自动编码机中，我们主要的工作是直接对显著性图进行建模（如图 2-37），在这里我们可以看出，在建模的过程中大多数学到的是对边缘响应的结构，比较符合我们的实验预期。

2.5 本章小结

在本章中我们探讨了全连接网络在视觉显著性物体检测辅助之下的深度学习。首先，本文探讨了全连接网络模型以及其常用的激活函数，并且通过误差消失现象探讨了在学习深层次网络时全连接网络遇到的问题，指出在训练深层全连接网络时参数初始化的重要性。之后，我们通过介绍受限玻尔兹曼机以及自动编码机阐述了当前解决深层网络训练问题的两个基本的非监督模型，详述了其建模方式以及常见变种，并探讨了常见的正则约束方式：L2 正则与 L1 正则。接着我们引入了视觉显著性物体检测，并从框架上阐述了两种可能的建模方式：一种是将视觉显著性物体检测当做先验来融入学习，第二种则是将其当成额外的辅助数据来进行学习。最后本文从这两种建模框架出发，详述了使用视觉显著性物体检测进行深度学习的两种方法，并通过实验说明了其有效性。

在进行建模时，我们首先使用的方法是将视觉显著性检测当成了整个系统的先验，在操作中我们在非监督学习的目标函数上加入了显著性的加权，即通过显著性检测来约

束非监督学习着重建模前景物体。在这里我们的初衷是通过这种方式尽量使得深度网络对背景变化鲁棒。通过实验我们发现，使用前景建模对分类任务有一定的促进作用，但相对不显著，我们分析主要原因是：第一，利用显著性信息获得的前景其实相对粗糙，其中居于前景与背景直接的模糊值居多，使得直接利用该信息获得的收益较少；第二，在该建模方式中，我们只以第一个双层网络为重，即在训练时只在第一个双层网络的权值初始化中利用了显著性信息，之后的网络初始化实际上并无辅助信息参与；第三，作为数据来说，CIFAR-10 与 STL-10 的中央偏置比较严重，也就导致了在正常进行深度学习时学习机也可以较好的“发现”物体，具体表现为在一般的自动编码机学出的特征中，也明显具有中央偏置现象，这样的话加入显著性先验获得的收益就相对较小。

鉴于以上的问题，我们提出了显著性重构的深度学习，即通过将显著性当做额外数据的方法进行建模。在该种建模中，我们在自动编码机中添加了额外的解码层，并迫使编码机解码出当前的显著性图。在这种建模方式下，相当于我们相对不干涉在网络中利用显著性建模的方式，其信息与源图像信息的融合完全由自动编码机来决定。同时，为了避免只能影响第一个双层网络的问题，我们进一步提出了双路自动编码机，即在对源图像进行非监督学习的同时在显著性图上进行学习，通过在高层次显著性图上的重构来非监督学习高层表示，这样我们就可以在高层的非监督学习中直接通过显著性信息影响网络学习。在实验中，本文的模型效果良好，在 STL-10 数据库的纵向对比中超过基础自动编码机约 3%。并且在显著性图的重构上，由于我们在学习中已经自发引入了物体类别信息，其重构结果在很多时候都要好过我们用作学习的显著性算法，并纠正其误检测的错误，说明在学习中我们已经很好的获得了物体的形状特征。

当然，本文的方法还存在许多不足之处。首先，虽然本文的建模只是针对非监督学习，在监督学习和以后的测试任务中模型的规模保持一致，但在监督学习中由于引入了额外的结构和约束，导致了训练时间几乎翻倍；其次，由于 CIFAR-10 与 STL-10 数据库的强烈中央偏置倾向，我们通过引入显著性获取的收益较小，并且在复杂场景下是否使用，建模是否应当变化仍然未知。这些不足也是我们日后研究的方向。

3 卷积神经网络深度模型

3.1 背景

卷积神经网络（Convolutional Neural Net）是神经网络中的另一个大类。在卷积神经网络中，信息的传递不像全连接网络中使用的线性映射的模式，而是采用卷积操作来进行相关操作，所以卷积神经网络中的学习任务主要是针对卷积核的学习。除了卷积操作以外，卷积神经网络中另一个重要的操作是 Pooling（汇聚）操作，在 Pooling 操作中卷积神经网络可以将一定范围内的信息进行汇聚，从而获取到局部不变性（Local Invariance）以及变化的感受野（Receptive Field）。由于卷积和 Pooling 的性质，卷积神经网络在有局部连续性的数据（比如图像、音频、视频）上有着良好的表现；并且由于卷积操作的特性，卷积网络受到误差消失现象的影响很小，使其成为可以直接进行深度训练的网络模型。

在卷积神经网络中，一般有交替的卷积、Pooling 操作，卷积操作用来进行局部特征提取，Pooling 操作则对一定范围内的特征响应进行统计；在一个深层卷积网络中，可能出现多个卷积+Pooling 的组合，并且在末尾一般会加入一个全连接隐含层与一个输出层（如图 3-1）。由于卷积操作、Pooling 操作以及全连接结构都可导，一般情况下卷积神经网络使用误差反传（Back Propagation）逐层求取梯度进行参数修正。

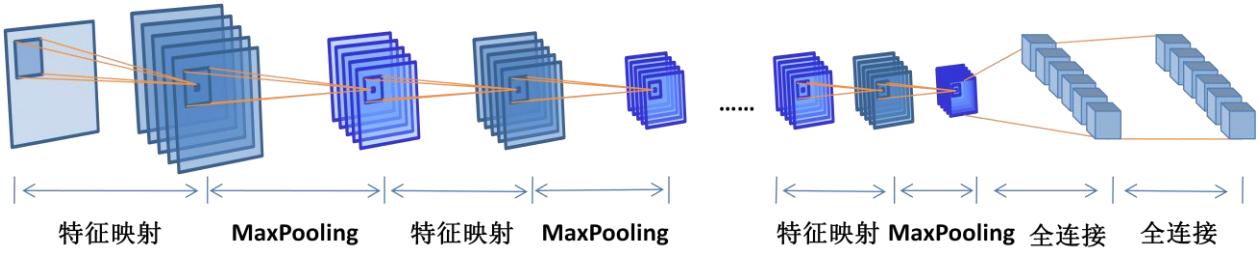


图 3-1 卷积神经网络结构示意图

在本节中，我们将针对卷积神经网络中的卷积、Pooling 两个部分进行详述，并且说明卷积网络不受误差消失现象的原因。

3.1.1 卷积

3.1.1.1 卷积的表达式

卷积操作是卷积神经网络中的基本操作之一，其实质上为对目标上的局部区域连续进行线性变换。我们拿图像举例，则在图像的卷积操作可以表达为对图像局部（Block）稠密采样后线性映射（如图 3-2）。

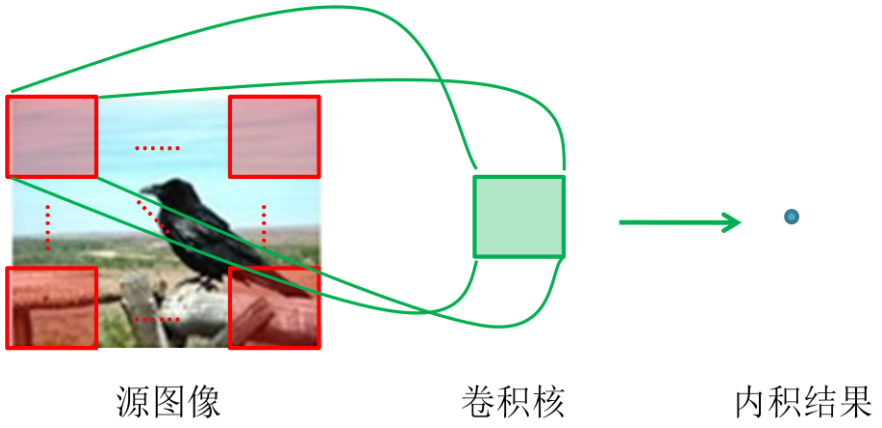


图 3-2 卷积示意图

在卷积操作中，在目标上滑动的部分叫做卷积核（Convolutional Kernel），假设我们有图像 $I \in R(N, M)$，卷积核 $K \in R(U, V)$，则卷积的结果 D 为：

$$D_{i,j} = \sum_{x=0}^{V-1} \sum_{y=0}^{U-1} K_{U-y,V-x} I_{i+y-\frac{U}{2},j+x-\frac{V}{2}}$$

从上式我们可以看到，卷积操作是针对与卷积核相同大小的局部进行操作的，并且这种操作在图像上的每一个地方是共享的（Shared），即对于图像的每一个局部块（Patch）都采用相同的卷积核做内积操作，这种性质使得卷积结果可以连续变化，并且间接使得卷积网络对误差消失现象鲁棒。这样带来的好处有两个：第一是使得卷积神经网络获得了平移一致性（Translation Consistency），即是在卷积神经网络中，图像的平移位移与卷积结果上对应响应的位移是一致的，这将大大加强神经网络对于平移的鲁棒性；第二则是使得卷积神经网络可以突破全连接网络的限制，直接对深度网络进行基于误差反响传播的训练。这些在后文中我们将详细推导与说明。其中卷积核的大小或是源数据上接受卷积的局部区域的大小也被叫做感受野（Receptive Field）。

另一点值得关注的是：在图像的卷积表达式中我们可以发现，卷积实际上是将卷积核对各个维度进行反转后再与图像进行操作的。这也是卷积操作定义的一个特殊之处；

如果没有反转卷积核，那么我们所做的操作叫做相关（Correlation）。相关和卷积的区别之处可以通过图 3-3 直观地说明。

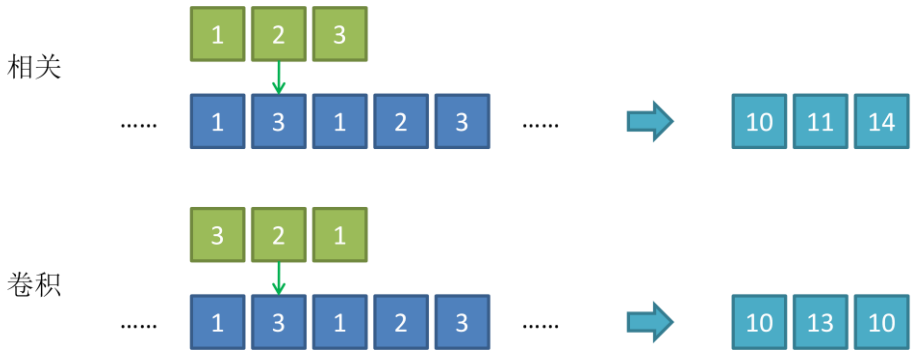


图 3-3 卷积和相关对比图

卷积操作另一个较好的性质是其与图像处理中的滤波（Filtering）是紧密相关的，即在时域（Time Domain）上的卷积操作相当于在频域（Frequency Domain）上的乘积操作。通过快速傅里叶变化（FFT, Fast Frequency Transform）操作，我们可以快速进行卷积操作，这也是进行卷积运算提速的方法之一。

3.1.1.2 卷积操作的推导

在上文中我们简述了卷积网络的构成，在卷积网络中我们主要学习的参数即为卷积核。在这一小节中，我们从卷积操作出发，推导卷积网络求解时的梯度表达式。我们仍然延续上一节的假设，即图像为 I，卷积核为 K，结果为 D，我们使用符号*来表示卷积操作，则可知：

$$D = I * K$$

假设我们当前有任务 T，T=T(D)。并且我们知道 T 在 D 上的偏导，则可知：

$$\frac{\partial T}{\partial K} = \frac{\partial T}{\partial D}\frac{\partial D}{\partial K}$$

通过 D 的表达式，我们可以推导：

$$\frac{\partial T}{\partial K} = \sum_i \sum_j \frac{\partial T}{\partial D_{i,j}} \frac{\partial D_{i,j}}{\partial K}$$

即 D 对 K 的偏导为 D 的每一个像素对 K 的导数。我们继续细分问题，则 $D_{i,j}$ 对于 $K_{u,v}$ 的偏导为：

$$\frac{\partial D_{i,j}}{\partial K_{u,v}} = \frac{\partial \sum_{x=0}^{V-1} \sum_{y=0}^{U-1} K_{U-y,V-x} I_{i+y-\frac{U}{2},j+x-\frac{V}{2}}}{\partial K_{u,v}} = I_{i-u+\frac{U}{2},j-v+\frac{V}{2}}$$

联合最早的式子，我么可得：

$$\frac{\partial T}{\partial K}=\frac{\partial T}{\partial D}\frac{\partial D}{\partial K}=\frac{\widetilde{\partial T}}{\partial D}*\tilde{I}$$

其中 $\tilde{I}$ 表示对 I 的坐标轴进行反转的值，在这里“~”表示坐标反转。假设我们以⊙来表示相关操作，那么可知导数表达式为：

$$\frac{\partial T}{\partial K}=\frac{\widetilde{\partial T}}{\partial D}\odot I$$

在得到了相关的梯度式之后，我们即可利用梯度下降算法进行模型的优化求解。在之前的分析中，我们指出了对于不同的非线性函数，其对于误差的响应和梯度信息的传播方式都不一致，但对于深层网络来说，误差上传递的损失会随着网络的层数逐层递增，以至于产生梯度消失的现象，这一点对于 Sigmoid 与 Tanh 激活函数尤其严重。

然而在卷积网络中我们可以发现，由于卷积操作在各个坐标处进行共享，在求解梯度时，卷积核的梯度为目标函数的偏导与图像 I 的相关（Correlation 操作）的坐标转置，也就是说对于每一个共享的位置，卷积核都可以收获一份梯度值。这种梯度共享带来的叠加虽然不能完全消除梯度消失现象（全连接网络也可以看做是感受野和图像一致大小的卷积，在这一点上两者可以统一起来），但在一般情况下可以大幅度的缓解训练时产生的深度网络梯度消失问题，从而使得卷积网络成为“不受”梯度消失现象影响的网络形式。

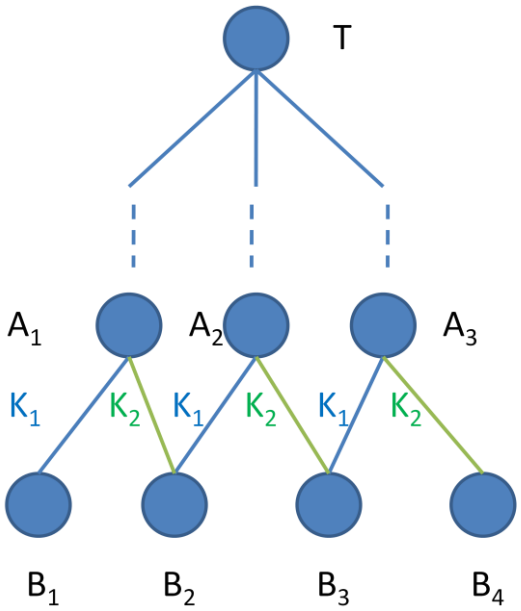


图 3-4 卷积网络梯度消失鲁棒性示意图

举例来说，对于图 3-4 的情况，假设我们有任务 T，在节点{A_1，A_2，A_3}与{B_1，

B_2，B_3，B_4}之间为大小为 2 的卷积核{K_1，K_2}，则假设我们已知任务在各个 A 节点的梯度，则对于卷积核参数{K_1，K_2}，其梯度为：

$$\frac{\partial T}{\partial K_1} = \sum_{i=1}^{3} \frac{\partial T}{\partial A_i} \frac{\partial A_i}{\partial K_1}$$
$$\frac{\partial T}{\partial K_2} = \sum_{i=1}^{3} \frac{\partial T}{\partial A_i} \frac{\partial A_i}{\partial K_2}$$

从梯度表达式可知，对于卷积网络的梯度值来说，因为卷积核在不同位之间共享，其梯度为所有共享位置的加和，也正是由于这种特性使得卷积神经网络对于梯度消失现象比较鲁棒。

3.1.2 Pooling 操作

在卷积神经网络中，Pooling 操作是除了卷积外的另一个重要操作，在该操作中，我们通过对局部信息进行统计，在获得一些不变性（Invariance）特征之外，又使得之后的神经元有更大的感受野。在本节我们着重讨论神经元感受野与 Pooling 操作的关系，并进一步简述几种常见的 Pooling 操作：平均 Pooling、最大 Pooling 以及最近新提出的随机 Pooling[83]。

3.1.2.1 Pooling 操作与感受野

感受野（Receptive Field）是一个神经学的概念并最初用在视觉神经的研究之中，在视觉神经系统中，一个细胞可以直接或间接的与多个光感受器细胞进行连接，则感受野就是指与某细胞连接的光感受器的个数和范围。人工神经网络（ANN，Artificial Neural Net）领域也借鉴了这一说法，感受野是指某个神经节点直接或间接连接的输入节点的个数和范围

通过以上的描述，我们知道对于卷积神经网络来说，卷积操作的感受野即为其卷积核的大小，但如果利用 Pooling 操作，我们可以增大神经节点的感受野。假设我们有一个（2M）*（2N）的特征图，那么在 2*2 的无重叠（Non-overlapping）Pooling 作用下，特征图会变为 M*N，则之后卷积的感受野就相当于变为原来的 4 倍大小（如图 3-5）。

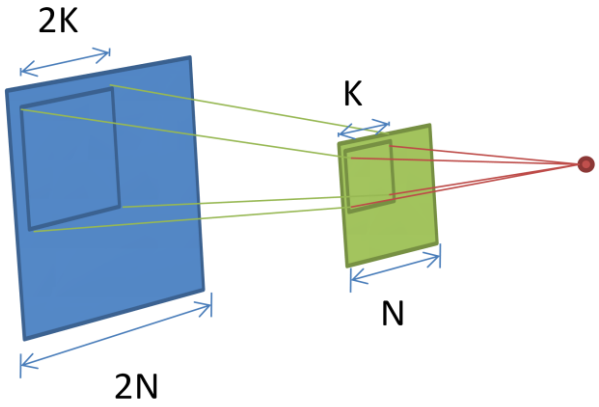


图 3-5 Pooling 带来感受野变化示意图

通过感受野的增大，我们可以使得卷积神经网络随着层次的递进，学的越来越全面与抽象的特征。在当今的卷积神经网络中，Pooling 已经是最重要的组成部分之一，在卷积网络的学习中扮演着重要的角色。

3.1.2.2 平均 Pooling（Average Pooling）

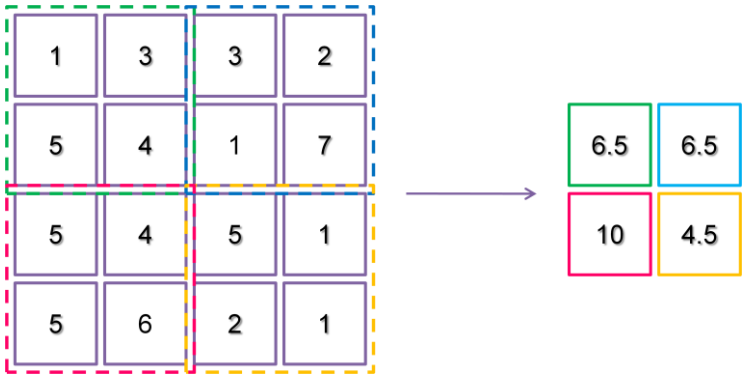


图 3-6 平均 Pooling 示意图

顾名思义，平均 Pooling 即是对局部区域的信息求平均值进行统计（如图 3-6）。假设 Pooling 操作的大小为 K*L，局部输入为 **I**，输出为 **o**，则我们可知表达式：

$$\mathbf{o} = \mathrm{AvgPooling}(\mathbf{I}) = \frac{1}{\mathrm{KL}} \sum_{\mathrm{i}=0}^{\mathrm{K}-1} \sum_{\mathrm{j}=0}^{\mathrm{L}-1} \mathbf{I}_{\mathrm{ij}}$$

由此可知，对于平均 Pooling，其梯度表达式为:

$$\frac{\partial \mathbf{o}}{\partial \mathbf{I}_{\mathrm{ij}}} = \frac{1}{\mathrm{KL}}$$

一般来说或，由于在平均 Pooling 中所有值在局部范围内平均，对于一些微小形变

来说平均 Pooling 可以获得良好的不变性。在实用中，著名的 LeNet5[48]就使用了平均 Pooling 技术，我们易得当平均 Pooling 的范围为 2*2 时，其操作相当于图像处理中的下采样（Down Sampling）技术（图像缩小至 1/2）。

3.1.2.3 最大 Pooling（Max Pooling）

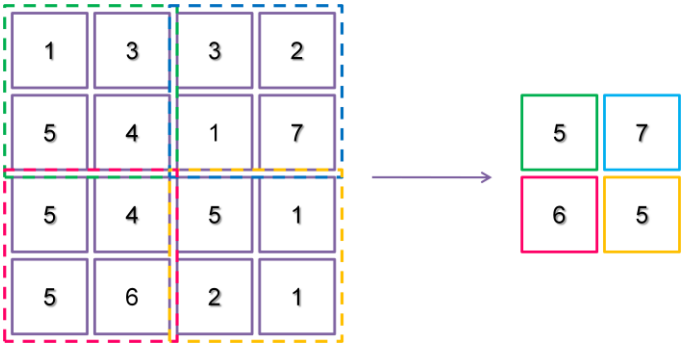


图 3-7 最大 Pooling 示意图

在最大 Pooling 中，该操作统计局部范围内的最大值（如图 3-7）。如上，假设 Pooling 操作的大小为 K*L，局部输入为 **I**，输出为 **o**，则我们可知表达式：

$$\mathbf{o} = \text{MaxPooling}(\mathbf{I}) = \max_{0 \le i \le K-1, 0 \le j \le L-1} (\mathbf{I}_{ij})$$

其梯度表达式为：

$$\frac{\partial \mathbf{o}}{\partial \mathbf{I}_{ij}} = \begin{cases} 1 & \text{if } \mathbf{I}_{ij} = \max \\ 0 & \text{elsewise} \end{cases}$$

最大 Pooling 最好的地方在于它显式地定义了局部不变性。在上文中我们提到，卷积操作提供了对于平移变化的一致性，在该性质下我们可以保证特征图的变化与输入源的变化趋于一致。而最大 Pooling 的最用则是可以在平移变化一致性的基础上，提供对平移变化的不变性（Translation Invariance）。

在图中我们可以看出，由于平移变化的一致性，特征图上的响应会随着输入图像的移动而移动；而在最大 Pooling 层，由于求取局部范围的最大值，当源数据的平移变化没有脱离 Pooling 的局部范围时，最大 Pooling 的响应保持一致，即获得了一定的平移不变性。最大 Pooling 的这一性质使得卷积神经网络对噪声的鲁棒性进一步提高，在实践中获得了良好的收益，是当今卷积神经网络中使用最为广泛的 Pooling 方法。

3.1.2.4 随机 Pooling（Stochastic Pooling）

随机 Pooling 是新兴的一种 Pooling 方法[83]。在随机 Pooling 中，我们对局部内所有的响应值归一化，使其可以代表该响应值的影响或是“概率”（如图 3-8）。这样，我们就在局部建立了一个简单的概率分布，对这个概率分布采样则可以采样到所有“概率”不为 0 的位置。响应最大的点有最大的概率被采到，但这并不意味这每次采样都会采到该点，也即是在 Pooling 操作中引入了随机性，这也是随机 Pooling 的由来。相比前两种 Pooling 方法，随机 Pooling 算法在通用数据库上了进一步获得了性能的提升。

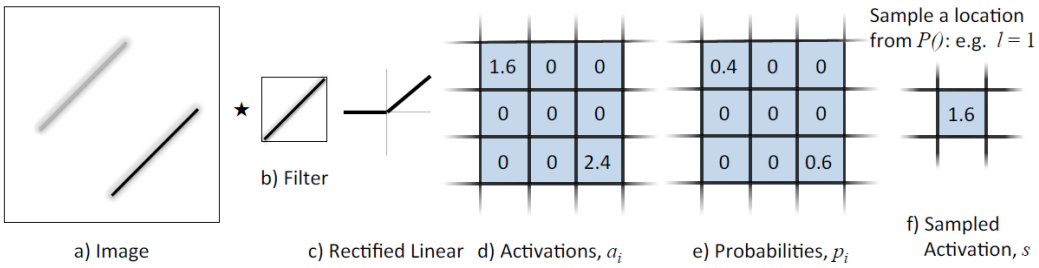


图 3-8 随机 Pooling 示意图[83]

一般来说，引入随机性可以在一定程度上提升系统的鲁棒性，因为随机的引入要求系统可以在发生变化时仍然做出正确的响应。引入随机变量来获取性能提升的例子有很多，比如第三章中我们着重分析的受限玻尔兹曼机模型以及降噪自动机模型中都有随机性的引入。

在随机 Pooling 算法中，我们可以换另一个角度来理解这种随机性。我们上文提到，在最大 Pooling 算法中，我们提取的值永远是最大的响应值，这样在带来鲁棒性的同时，也有可能使得某些局部位置被“过分”更新，而其他位置的权值一直得不到更新；而在平均 Pooling 中，所有的局部位置都被平均的进行更新，反而难以有重点的进行学习。因此，我们可以将随机 Pooling 看成是最大 Pooling 与平均 Pooling 的折中，即保证了最大响应值位置的更新，又保证了小响应值位置的更新。

3.2 相关深度学习模型

在上文中我们分析了卷积神经网络中的两个重要的组成元素，即卷积操作与 Pooling 操作。在卷积神经网络的结构下，由于权值的共享误差信息可以良好的再网络中传播，并进行权值的修正。所以一般来说，在卷积网络的训练中，我们并不需要复杂的预训练

过程就可以进行深度网络的学习。然而，这样的前提是我们有足够的标注样本进行监督学习，通常来说标注样本的获取代价要远高于非标注样本，所以在标注样本不足或完全没有标注样本的情况下，我们则需要一些相关的非监督模型来进行卷积网络的学习。

下面我们将着重介绍两种卷积非监督学习的模型，即受限玻尔兹曼机的扩展模型卷积受限玻尔兹曼机以及自动编码机的扩展模型卷积自动编码机。

3.2.1 卷积受限玻尔兹曼机

在卷积受限波尔兹曼（Convolutional RBM）[39]中，相对于原始的受限玻尔兹曼机使用线性映射进行层间信息的传递，在这里使用的是卷积操作。相对于线性映射，卷积操作利用了图像中局部信息连续性的先验，是以可以得到更好的建模结果。另外，对于卷积网络中的Pooling操作，在卷积玻尔兹曼机中使用了概率对最大Pooling进行了建模，进一步增强了卷积受限玻尔兹曼机建模的鲁棒性。下面我们将先介绍几本的卷积玻尔兹曼机，然后介绍概率最大 Pooling 的原理与实现，最后介绍融合最大 Pooling 的卷积玻尔兹曼机模型。

3.2.1.1 卷积受限玻尔兹曼机

在卷积受限玻尔兹曼机中，我们利用卷积操作来替代线性映射进行神经网络层间信息的传递。假设我们有一卷积玻尔兹曼机，其输入信号或是显层（Visible）信号为 $\mathbf{v}$，隐层（Hidden）信号为 $\mathbf{h}$，则可以仿照高斯-伯努利受限玻尔兹曼机得到其能量函数的表达式为[39]：

$$\mathrm{E}(\mathbf{v},\mathbf{h}) = -\sum_{\mathrm{k}}(\mathbf{h}^{\mathrm{k}})^{\mathrm{T}}(\mathbf{W}^{\mathrm{k}} * \mathbf{v}) - \sum_{\mathrm{k}}(\mathbf{h}^{\mathrm{k}})^{\mathrm{T}}\mathbf{c}^{\mathrm{k}} + \|\mathbf{v}-\mathbf{b}\|_2^2$$

其中上标 k 表示第 k 组参数或响应（例如，$\mathbf{h}^{\mathrm{k}}$ 表示第 k 组卷积核的响应，$\mathbf{W}^{\mathrm{k}}$ 则表示第 k 组卷积核）。在这里为了建模方便，我们去除了输入数据的方差变量（即令方差为 1）。由此可得显层单元与隐层单元的条件概率表达式为：

$$\mathrm{P}\left(\mathbf{h}^{\mathrm{k}}\middle|\mathbf{v}\right) = \sigma(\mathbf{W}^{\mathrm{k}} * \mathbf{v} + \mathbf{c}^{\mathbf{k}})$$

$$\mathrm{P}(\mathbf{v}|\mathbf{h}) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(\mathbf{v}-(\mathbf{b}+\sum_{\mathrm{k}}\mathbf{W}^{\mathrm{k}}*\mathbf{h}^{\mathrm{k}}))^2}{2}}$$

其中$(\mathbf{v}-(\mathbf{b}+\sum_{\mathrm{k}}\mathbf{W}^{\mathrm{k}}*\mathbf{h}^{\mathrm{k}}))^2$是向量，表示对运算结果向量对应位置求平方。与受限玻尔兹曼机的求解相似，我们可以利用 CD 算法来进行模型分布的估计，并通过最小

化模型分布与数据分布来进行模型的优化求解。

3.2.1.2 最大 Pooling 的卷积受限玻尔兹曼机

在卷积受限玻尔兹曼机中，Pooling 操作也得到了合适的建模从而使得模型具有了更强的鲁棒性。我们假设对于显层 V 以及隐含层 H，有最大 Pooling 层 P，并且我们已知在进行建模时我们假设 H 符合伯努利分布。根据上一小节的建模，我们可知对于显层信号 **v**，隐层信号 **h** 来说有：

$$P\left(\mathbf{h}^k\middle|\mathbf{v}\right)=\sigma(\mathbf{W}^k*\mathbf{v}+\mathbf{c}_k)$$

我们知道，对于最大 Pooling 来说，其值为对 **h** 进行局部最大统计的值。对于求取最大值操作的结果我们很难进行概率建模，但是从另一角度来说，由于 **h** 符合伯努利分布，即意味着 Pooling 层信号 **p** 也具有相同的分布，则我们可知当且仅当局部内所有的 **h** 都没有激活时，对应的 Pooling 信号不激活，所以我们可得：

$$P\left(p^k=0\middle|v\right)=\frac{1}{1+\sum_{(i,j)\epsilon B}P(h_{i,j}^k=1|v)}$$

$$P\left(p^k=1\middle|v\right)=1-P\left(p^k=0\middle|v\right)$$

其中 B 为 Pooling 层单元对应的在隐含层 H 上的感受野。在这里我们将 Pooling 层单元与其感受野对应的所有隐含层单元当成是统一分布，假设感受野大小为 C*C，则我们另外加一个值代表所有单元皆为 0 的情况，也就得到了我们的表达式。在概率最大 Pooling 的建模下，我们将 Pooling 的结果当成下一个卷积受限波尔兹曼机的输入进行迭代学习。在得到了概率表达式之后，我们可以同样的利用 CD 算法进行模型求解。

3.2.2 卷积自动编码机

3.2.2.1 卷积重构

卷积自动编码机（Convolutional Auto-Encoder)[54]和卷积受限玻尔兹曼机的原理类似，但在卷积自动编码机中我们采用自动编码机的数据重构模型来进行建模。假设我们有数据 **v**，卷积核心 $\mathbf{W}^k$，隐单元偏置量 **c**，我们可知其隐层的响应 $\mathbf{h}^k$ 为：

$$\mathbf{h}^k=\sigma\left(\mathbf{W}^k*\mathbf{v}+\mathbf{c}^k\right)$$

假设我们有用于重构的卷积核心 $\mathbf{R}^k$，重构偏置 **b**，则重构响应 **r** 为：

$$\mathbf{r} = \sigma(\sum_{k}(\mathbf{R}^{k} * \mathbf{h}^{k} + \mathbf{b}^{k}))$$

则根据自动编码机的特点，我们容易得出其目标函数为：

$$\min||\mathbf{v} - \mathbf{r}||_2^2$$

即最小化数据的重构误差。在上面的表达式中，我们可以发现对于从源数据到隐层的映射来说，假设我们有 N 个卷积核，那么我们将得到 N 个特征图 $\mathbf{h}^1$~$\mathbf{h}^N$。在重构中，我们需要对着 N 个特征图分别求取卷积结果并将其求和。则我们可以将特征图 $\mathbf{h}^1$~$\mathbf{h}^N$ 看成是有 N 个通道（Channel）的图像 **H**，则重构过程变为使用一个有 N 个通道的卷积核 **R** 与 **H** 直接进行卷积操作，再与偏置 **b** 相加，即：

$$\mathbf{r} = \sigma(\mathbf{R} * \mathbf{H} + \mathbf{b})$$

如此一来我们可以发现 **R** 与 **W** 的形式是紧密相关的。在第三章中我们提到，自动编码机经常通过将重构映射矩阵变为特征映射矩阵的转置（Transpose）来缩减参数规模。这一点对于卷积自动编码机来说是一致的，即我们可以通过绑定 **R** 为：

$$\mathbf{R}^{k} = \widetilde{\mathbf{W}^{k}}$$

即 $\mathbf{R}^k$ 是 $\mathbf{W}^k$ 坐标轴反转后的结果。这样一来，我们可以削减掉卷积核 $\mathbf{R}^k$，直接利用 $\mathbf{W}^k$ 来进行重构任务。

3.2.2.2 稀疏性与最大 Pooling

在机器学习中，稀疏性（Sparsity）一直是重要的议题之一。在机器学习中，如果我们限制权值激活的数量，则可以得到稀疏的结果，而这种稀疏性由于控制了模型规模，相当于在目标函数式上增加了较强的正则项，在试验中获得了优异的结果。

在深度学习中，许多模型也都可以通过稀疏性约束来进行模型规模的控制，从而避免过拟合获取泛化性能（Generalization）的提升。在上一节中，我们发现通过概率最大 Pooling（Probability Max Pooling），卷积受限玻尔兹曼机可以获得良好的数据描述；事实上，在引入最大 Pooling 的同时，卷积受限玻尔兹曼机也在 Pooling 层获得了较为稀疏的表达。在这里，卷积自动编码机对于最大 Pooling 的处理则可以更直观的体现出模型的稀疏性。

在卷积自动编码机中，我们也可以像之前的卷积玻尔兹曼机一样加入最大 Pooling 操作。在卷积自动编码机中，假设我们的隐层响应为 $\mathbf{h}^k$，最大 Pooling 的大小为 L*P 则我们可以得到 Pooling 的结果 $\mathbf{p}^k$ 为：

$$\mathbf{p}_{ij}^{k} = \max_{i-\frac{L}{2}\le y\le i+\frac{L}{2}, j-\frac{P}{2}\le x\le j+\frac{P}{2}}(\mathbf{h}_{yx}^{k})$$

在这里，我们使用类似之前求取最大 Pooling 导数的技巧，对于最大 Pooling 的结果进行重构，假设重构结果为 $\mathbf{hr}^{k}$，则我们可知：

$$\mathbf{hr}_{yx}^{k} = \begin{cases} \mathbf{h}_{yx}^{k} & \mathbf{h}_{yx}^{k} \text{ 是局部最大值} \\ 0 & \text{其余情况} \end{cases}$$

在这里，我们发现通过最大 Pooling 的重构，我们将重构结果 **hr** 变成了一个十分稀疏的表示，并且随着最大 Pooling 的局部范围扩大，**hr** 的稀疏性会进一步增强。

在获得 hr 之后，相对于上文中我们使用 **h** 进行重构，这里我们使用 **hr** 进行目标图像的重构，即：

$$\mathbf{r} = \sigma(\sum_{k}(\mathbf{R}^{k} * \mathbf{hr}^{\mathbf{k}} + \mathbf{b}^{k}))$$

在这里由于 **hr** 的稀疏性，我们实际上要求卷积自动编码机在失去了很多信息的情况下进行数据的恢复，从这一点上来说，最大 Pooling 对于卷积自动编码机来说比较像是在降噪自动编码机中加入置零噪声（即对数据中的随机维度置零）；然而相对于降噪自动编码机，在这里最大 Pooling 的意义更为明确，即一方面进行隐层响应的稀疏性约束，另一方面要求学得的特征要对于局部变化鲁棒。在实验中，最大 Pooling 的加入也使得卷积自动编码机获得了性能的提升。

3.3 视觉显著性辅助的卷积神经网络

与 4.2 节的方式类似，我们也可以将利用视觉显著性来辅助卷积神经网络的非监督学习。由于图像信息具有良好的局部平滑性，所以在实践中利用拥有局部感受野的卷积神经网络通常可以提升性能；并且对于显著性检测来说，显著性检测的很多准则也都建立在局部感受野上（局部对比等），所以利用卷积操作来进行显著性的建模无疑更为合适。在下文中，我们先通过卷积网络来进行显著性物体检测，以此说明卷积网络在显著性建模上的优势；之后我们详述使用视觉显著性物体检测信息辅助卷积神经网络进行非监督学习的方法：

3.3.1 卷积神经网络与显著性物体检测

在这一节，我们将使用卷积神经网络进行显著性物体检测的建模，并以此说明卷积

网络对于显著性建模的优势，在下一节中我们将看到这种建模优势将在物体分类的任务中被良好的利用。

在卷积网络中，假设我们有 N 个卷积核心 W^1~W^N，那么对于图像 I 与其显著性图像 S，我们可以用最简单的网络对显著性进行建模，即：

$$\min \left\| \left(\sum_k W^k * I \right) - S \right\|_2^2$$

在这里我们发现仅仅使用线性网络对显著性信息建模是不够的，所以我们引入非线性变化，也就是使用两层网络进行建模。对于上面的情况，假设中间特征结果为 H，则我们可的表达式：

$$H^k = \sigma\left(W^k * I\right)$$

$$\hat{S} = \sum_k \widetilde{W}^k * H^k$$

在这里，我们利用卷积与激活函数将数据映射到隐空间，再从隐空间通过卷积进行显著信息的建模。通过以上表达式不难发现，这一做法与卷积自动编码机的做法十分类似，不过我们要求其重构的结果为显著性物体检测而并不是为了数据重构。在实践中，这种方法在显著性建模上获得了较好的结果，这也是我们在显著性重构神经网络中使用的策略。

3.3.2 显著性重构的卷积神经网络

在上一节中，我们讨论了利用卷积网络来建模显著性的方式，在这里我们将利用建模的结果来辅助卷积神经网络的学习。在卷积网络的建模中，我们仍旧采取非监督权重初始化加监督学习的模式，并且只在非监督学习中利用显著性信息进行辅助，从而保证了在监督学习时深度学习的规模一致。出于建模的便捷性，我们选用了卷积自动编码机作为基础模型。

与第三章讨论的情况相似，为了避免显著性难以深层利用的问题，我们采用双路建模来进行非监督学习。假设有第 i 层的数据 $\mathbf{h}_i$，显著性数据 $\mathbf{s}_i$，主编码机的权值 $\mathbf{W}_i$ 以及辅助编码机的权值 $\mathbf{R}_i$，则我们可得：

$$\mathbf{h}_{i+1}^k = \sigma(\mathbf{W}_i^k * \mathbf{h}_i)$$

$$\mathbf{s}_{i+1}^k = \sigma(\mathbf{R}_i^k * \mathbf{s}_i)$$

其中 h_{i+1} 与 s_{i+1} 分别为第 i+1 层源数据与显著性数据的表示。则我们假设 $\mathbf{hs}_i$ 为通过

主编码机解码出的显著新结果，而 $\hat{\mathbf{h}}_i$ 为主编码机重构源数据第 i 层表示的结果，$\hat{\mathbf{s}}_i$ 为辅助编码机重构第 i 层显著性表示的结果，则可得：

$$\hat{\mathbf{h}}_i = \sigma\left(\sum_k \widetilde{\mathbf{W}}_i^k * \mathbf{h}_{i+1}^k\right)$$

$$\hat{\mathbf{s}}_i = \sigma\left(\sum_k \widetilde{\mathbf{R}}_i^k * \mathbf{s}_{i+1}^k\right)$$

$$\mathbf{hs}_i = \sigma\left(\sum_k \widetilde{\mathbf{WS}}_i^k * \mathbf{h}_{i+1}^k\right)$$

其中 **WS** 是主编码机中用于重构显著性表示的连接权。与第三章类似，在这里我们的目标函数为：

$$\min\left\|\boldsymbol{h}_i - \hat{\mathbf{h}}_i\right\|_2^2 + \lambda\|\boldsymbol{s}_i - \mathbf{hs}_i\|_2^2 + \|\boldsymbol{s}_i - \hat{\mathbf{s}}_i\|_2^2$$

其中λ 用来在主编码机中进行权衡，λ 值越大主编码机中学出的表示就越偏向于重构显著性表示，反之则偏向于重构源数据的表示。在实践中，与之前得到的结果类似，我们发现λ 值偏大时整个系统的性能较好，故在试验中我们置λ =1。

3.4 实验与分析

3.4.1 卷积网络模型规模

在卷积网络模型中，由于卷积模型对于梯度消失现象具有很强的鲁棒性，所以在模型深度上已经不需要做太多的选择，我们根据以往文献的策略选择网络进行建模。在模型大小上，由于卷积网络训练时间较长，所以我们选择了较小的模型进行学习，其模型的详细参数如下：

（1） 第一层：输入数据 24*24*3，在输入数据时，我们依照文献[34]的方法对 32*32 的图像进行 24*24 大小的裁剪，并且由于图像为 RGB 的三通道图像，故第一层通道数为 3

（2） 第二层：卷积层，卷积核大小为 5*5*3，共有卷积核 32 个

（3） 第三层：最大 Pooling 层，Pooling 大小为 2*2，无重叠

（4） 第四层：卷积层，卷积核大小为 5*5*3，共有卷积核 32 个

（5） 第四层：最大 Pooling 层，Pooling 大小为 2*2，无重叠

（6） 第五层：卷积层，卷积核大小为 5*5*3，共有卷积核 32 个

（7） 第六层：最大 Pooling 层，Pooling 大小 2*2，无重叠

（8） 第七层：全连接层，隐单元个数为 64

（9） 第八层：输出层，输出类别个数 10

在这里，由于有些时候卷积与 Pooling 层可以算作一个卷积神经网络中的单元或“层”，那么这里的层数即为 5 层。

3.4.2 显著性重构的卷积网络

在显著性重构的卷积网络中，我们利用了与第二章中相似的策略进行显著性重构，并且选择了性能最优的双路模型进行学习，只不过在这里我们将原来的全连接模型换成了显著性模型。

我们测试的数据库依然是 CIFAR-10 与 STL-10 数据库，在测试 STL-10 数据库时，我们按照数据库的要求，使用 10 折数据（即每一次训练时的训练集大小为 1000 张图像，共 10 组并求取平均值）来进行数据库的测试。其测试结果为：

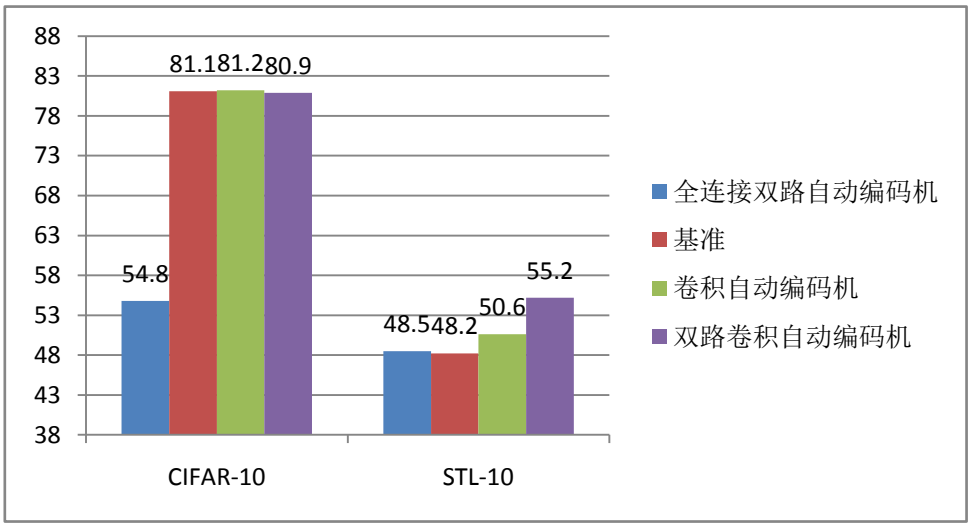


图 3-9 双路卷积自动编码机结果图

与之前的结果类似，在 CIFAR-10 数据库上是否加入显著性信息并不影响其分类效果。我们发现对于卷积神经网络来说，在 CIFAR-10 数据库上网络权值的初始化对网络最终分类性能的影响极其微小（图 3-9），这也进一步说明了卷积神经网络在深度网络监督学习上的巨大优势。

相对于 CIFAR-10，由于 STL-10 数据库的训练样本要少得多，所以卷积网络并不能完全依赖监督学习，这时候也比较能体现出在非监督学习阶段学习算法的优劣。在这里，由于我们采用了裁剪的方式处理样本，相当于人工增加了训练样本量，这样可以大幅度

提升效果，却也缩减了非监督学习的作用。从结果中我们可以看出，对于 STL-10 数据库没有进行权值初始化的模型达到的分类准确率为 48.2%，而传统的卷积自动编码机可以提升至 50.6%，我们提出的双路卷积自动编码机达到了 55.2%。相对于典型的卷积自动编码机，我们加入显著性进行双路重构的模型提升了接近 5%的准确率。鉴于本文的主要目的即是加入显著性信息增强深度学习中非监督学习的效能，所以在之后的分析中我们着重对 STL-10 的结果进行分析。

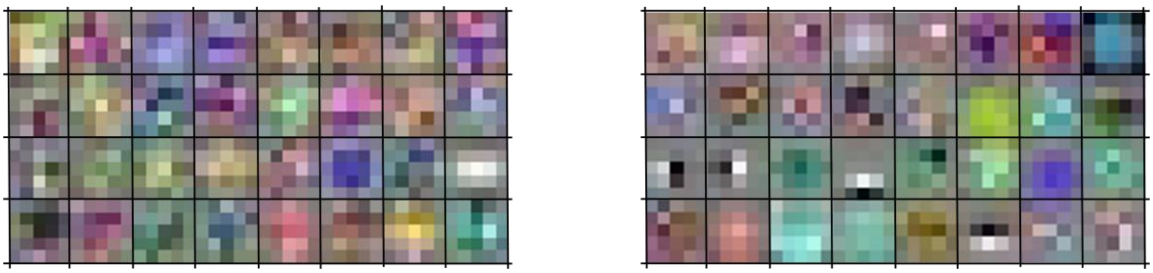


图 3-10 双路卷积自动编码机特征可视化图。左图为卷积自动编码机可视化图，右图为本文中双路卷积自动编码机就可视化图。

为了分析性能提升的原因，我们依然采用可视化的方式查看权重。在非监督学习中，第一层卷积核的可视化结果如图 3-10。从图中我们可以看出，单纯使用卷积神经网络进行重构的卷积核相对比较杂乱，而使用双路卷积自动编码机则可以学习出明显更为局部收敛的特征。鉴于特征上的差异，为了验证算法已经收敛，我们对重构图像的结果进行可视化（如图 3-11）。在可视化结果中，第一排是源图像的结果，第二排是重构后的结果，由于重构时对图像进行了裁剪，所以并不能完全对齐。在结果中，我们不难看出不论使用卷积自动编码机还是我们提出的双路自动编码机都可以有效地对源图像进行重构，所以两算法都已很好的收敛，同时也说明了我们提出的算法中正则项的有效性。

图 3-11 重构结果可视化图。上图为卷积自动编码机重构结果，下图为双路卷积自动编码机重构结果。

另外，对于双路编码机来说，显著性图像的重构是重要的一部分，我们可视化的结果如图 3-12。在卷积的显著性重构中，由于卷积操作的局部性，事实上由于位置的不确定性，该操作相对于全连接网络中双路自动编码机的显著性重构更为复杂。在结果中，我们发现重构的结果基本与显著性检测的结果相仿，同时在一些样本中，我们仍能发现重构修正显著性检测的现象，说明我们在学习中确实学到了物体的形状结构信息。

图 3-12 双路自动编码机显著性重构可视化图，其中每四行为一组，每组中第一行为原始图像，第二行为双路自动编码机重构结果，第三行为原始显著性物体检测图像，第四行为双路自动编码机显著性重构结果

在这里，我们与当前几种领先的方法[15][16][45][56][87]在 STL-10 下进行对比（如图 3-13）。方法[16]利用 KMeans 方法通过图像块的聚类学习特征，[45]利用相似的原理不过利用了独立成分分析（ICA）进行特征学习，而[56]则在学习添加了稀疏性：约束，获得了更好的结果。文章[87]通过追踪视频中的点来进行不变特征的学习，并且在获得额外数据后可以得到很大的性能提升。方法[15]则在学习中加入了感受野的学习。在结果中我们可以看出，我们利用视觉显著性的方法与领先方法可比。

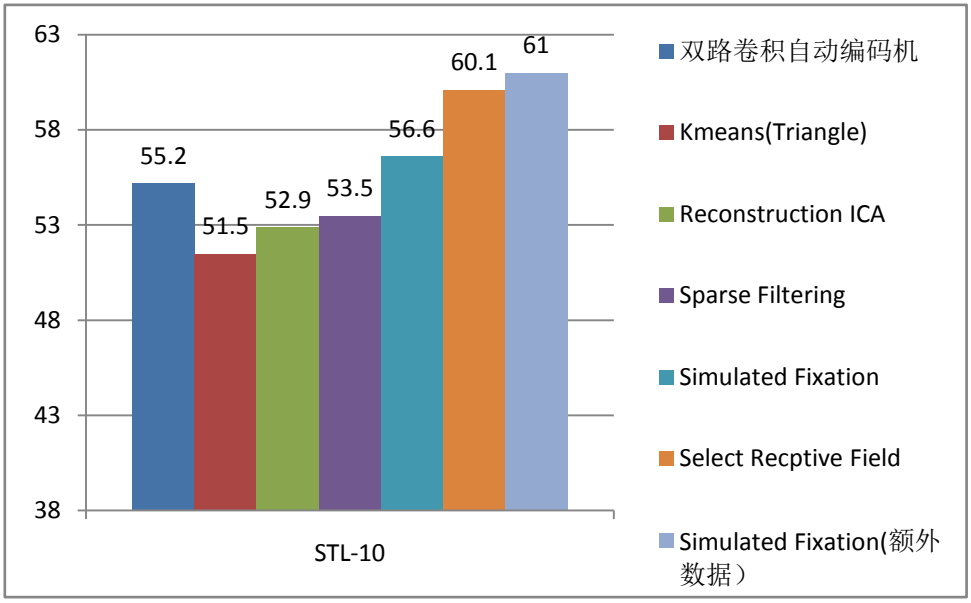


图 3-13 测试结果对比图

3.5 本章小结

在本章中，我们将视觉显著性物体检测应用在了卷积网络中。首先我们介绍了卷积网络的结构，并从梯度的推导上说明了卷积网络不受误差消失现象影响的原因。其次，我们分别从概率建模和重构建模两种方式上简述了当前的两种卷积非监督学习的方式，即卷积波尔兹曼机与卷积自动编码机。最后，我们提出了利用视觉显著性信息的双路卷积自动编码机，并通过实验说明了算法的有效性。

在显著性建模时，我们利用了上一章的方法并将其扩展到了卷积网络，扩展的原因主要是由于卷积神经网络在学习时很大限度的利用了图像的先验知识，大幅度地降低了参数空间的大小，所以使得卷积网络在图像建模中更为适用，性能也更为优异。但由于卷积操作的计算复杂度较高，在实际操作中，卷积神经网络也需要更多的时间进行训练与测试。在实验中，我们发现在加入了显著性信息后，卷积网络的第一层更容易收敛到局部的特征并且出现了一些类 Haar 或 Gabor 的特征，在上一章中我们指出这些特征将有利益图像分类问题的建模。在实际的分类测试中，我们也发现利用显著性信息建模的模型比起单纯的卷积自动编码机有约 5%的性能优势，这种优势在以非监督学习为主的 STL-10 数据库中还是相对显著的。

虽然我们提出的模型在性能上有所提升，但是对于当前领先的算法还有一定的差距，

经过分析我们认为问题可能在以下几个方面：1）模型规模上我们的模型由于计算能力的限制偏小，可能造成表述能力上的限制；2）在参数调整上虽然我们最大可能的进行了参数调优，但由于卷积网络计算量较大，参数的寻优并不完全；3）在建模上，虽然我们利用了显著性信息作正则，但并没有很好利用诸如稀疏性、低秩性等一般的约束性质；4）显著性信息提供的信息可能有限，作为正则项不一定有别的方法中利用的正则约束有力。当然，这些可能的问题与探索方向也是我们日后进行试验与研究的方向。

4 总结

深度学习是近期新兴的机器学习算法之一，基于深层次的学习，深度学习技术可以学习高度抽象的概念，以至于在当今图像识别、语音识别、自然语言理解等多个领域都取得优势地位。在深度学习中，一个主要的问题是由于模型参数规模较大，一般来说需要大量样本进行模型的训练学习。在互联网快速发展的今天，来势汹汹的“大数据”看似正好与深度学习完美的契合，但由于深度学习模型大多是依赖标注样本的监督学习模型，昂贵的人力、物力消耗使得有标注的样本来之不易。基于这种前提下，本文的目的即为研究在图像分类中如何有效地利用无标注样本进行深度学习，使得在标注样本较少的情况下获得较好的学习效果。

在本文的第一章，我们对当前的研究背景与研究现状进行了分析与讨论，并且从中提出了本文的研究目的与内容，即利用视觉显著性信息辅助进行深度学习的建模学习，并阐述了本文工作在当前的意义所在，介绍了本文的组织结构。

在第二章中，我们从全连接网络出发，首先介绍了模型形式与常见的非监督深度学习模型。之后我们提出了两种显著性建模方式，即利用显著性信息作为样本先验的建模方式以及利用显著性信息当成模型额外参数的建模方式；在两种建模方式中，前者对源图像与显著性信息进行显式建模与约束，而后者则利用模型来隐式建模二者相关性。然后我们根据两种方式进行了视觉显著性辅助的深度学习建模，并阐述了模型特点与建模细节。最后我们通过实验验证了提出方法的可行性，在高度依赖非标注样本的 STL-10 数据库中，本文提出的方法可以高出基准 3%。

在第三章中，我们将模型扩展到了卷积神经网络。首先，我们介绍了卷积神经网络的特点与优势，并介绍了两种非监督学习的基准模型；然后我们利用显著性信息对卷积网络的深度学习进行建模；最后我们通过实验证明了提出方法的有效性，在卷积网络中由于能够更好地利用图像信息，我们提出的方法的识别率高出基准线 5%，并与当前领先的方法可比。

在当前我们的方法存在的主要问题主要是：首先对于学习框架来说，我们使用非监督初始化加监督学习的方式来完成半监督的深度学习，在这里视觉显著性仅在第一部分

也就是非监督学习部分起作用，没有完全的融合在整个学习流程中；其次，对于测试数据来说，我们选用的数据库虽然样本较多，但具有比较明显的中央偏置现象，对于背景复杂的图像来说，本方法的效果还有待验证；最后，对于业界领先的方法来说，我们的方法虽然可比但还有一定的差距。以上即是我们存在的问题，也是以后研究的重点所在。随着显著性检测方法精度的提升与深度学习能力的提升，我们期望有一天能有一个优秀的模型将两者进行融合，让简单易得的显著性检测结果来大幅提升深度学习能力，解决标注数据难以获取的难题；而这也是我们以后的努力方向所在。

参考文献

[1] B. Alexe, T. Deselaers, V. Ferrari, What is an object?, 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2010), pp. 73-80.
[2] B. Alexe, T. Deselaers, V. Ferrari, Measuring the objectness of image windows, IEEE Transactions on Pattern Analysis and Machine Intelligence, , 34 (2012) 2189-2202.
[3] Y. Bengio, Learning deep architectures for AI, Foundations and trends in Machine Learning, 2 (2009) 1-127.
[4] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, (2013).
[5] R. Bharath, J. Nicholas, L. Zhi, X. Cheng, Scalable scene understanding using saliency-guided object localization, 10th IEEE International Conference on Control and Automation (ICCA), (2013), pp. 1503-1508.
[6] C.M. Bishop, Pattern recognition and machine learning (springer New York, 2006).
[7] D. Borthakur, The hadoop distributed file system: Architecture and design, Hadoop Project Website, 11 (2007) 21.
[8] L. Breiman, Random forests, Machine learning, 45 (2001) 5-32.
[9] F. Chen, H. Yu, R. Hu, X. Zeng, Deep learning shape priors for object segmentation, 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2013), pp. 1870-1877.
[10] K. Cho, A. Ilin, T. Raiko, Improved learning of Gaussian-Bernoulli restricted Boltzmann machines, International Conference on Artificial Neural Networks–ICANN, (2011), pp. 10-17.
[11] K.H. Cho, T. Raiko, A. Ilin, Gaussian-bernoulli deep boltzmann machine, The 2013 International Joint Conference on Neural Networks (IJCNN), (2013), pp. 1-7.
[12] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, P. Vincent, The difficulty of training deep architectures and the effect of unsupervised pre-training, International Conference on Artificial Intelligence and Statistics, (2009), pp. 153-160.
[13] D. Ciresan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2012), pp. 3642-3649.
[14] A. Coates, B. Huval, T. Wang, D. Wu, B. Catanzaro, N. Andrew, Deep learning with COTS HPC systems, Proceedings of The 30th International Conference on Machine Learning, (2013), pp. 1337-1345.
[15] A. Coates, A.Y. Ng, Selecting Receptive Fields in Deep Networks, NIPS, (2011), pp. 8
[16] A. Coates, A.Y. Ng, H. Lee, An analysis of single-layer networks in unsupervised feature learning, International Conference on Artificial Intelligence and Statistics, (2011), pp. 215-223.
[17] G.E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, IEEE Transactions on Audio, Speech, and Language Processing, (2012) 30-42.
[18] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Recent advances in deep learning for speech research at Microsoft, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2013), pp. 8604-8608.
[19] M. Denil, L. Bazzani, H. Larochelle, N. de Freitas, Learning where to attend with deep architectures for image tracking, Neural computation, 24 (2012) 2151-2184.
[20] A.W. Dennis, D. Ventura, Learning the Architecture of Sum-Product Networks Using Clustering on Variables, NIPS, (2012), pp. 2042-2050.
[21] P. Domingos, A few useful things to know about machine learning, Communications of the ACM, 55

(2012) 78-87.
[22] S.A. Eslami, N. Heess, J. Winn, The shape boltzmann machine: a strong model of object shape, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2012), pp. 406-413.
[23] J. Feng, Y. Wei, L. Tao, C. Zhang, J. Sun, Salient object detection by composition, 2011 IEEE International Conference on Computer Vision (ICCV), (2011), pp. 1028-1035.
[24] P. Földiák, M.P. Young, Sparse coding in the primate cortex, The handbook of brain theory and neural networks, 1 (1995) 1064-1068.
[25] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of Statistics, (2001) 1189-1232.
[26] R. Gens, P. Domingos, Discriminative Learning of Sum-Product Networks, NIPS, (2012), pp. 3248-3256.
[27] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, International Conference on Artificial Intelligence and Statistics, (2010), pp. 249-256.
[28] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2012) 1915-1926.
[29] http://groups.csail.mit.edu/vision/TinyImage
[30] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, arXiv preprint arXiv:1302.4389, (2013).
[31] W. Gropp, E. Lusk, A. Skjellum, Using MPI: portable parallel programming with the message-passing interface (MIT press, 1999).
[32] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural computation, 18 (2006) 1527-1554.
[33] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science, 313 (2006) 504-507.
[34] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, arXiv preprint arXiv:1207.0580, (2012).
[35] G.E. Hinton, R.S. Zemel, Autoencoders, minimum description length, and Helmholtz free energy, Advances in neural information processing systems, (1994) 3-3.
[36] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on pattern analysis and machine intelligence, 20 (1998) 1254-1259.
[37] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, IEEE 12th international conference on Computer Vision, (2009), pp. 2106-2113.
[38] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, Y. LeCun, Learning Convolutional Feature Hierarchies for Visual Recognition, NIPS, (2010), pp. 5.
[39] A. Krizhevsky, Convolutional deep belief networks on cifar-10, Unpublished manuscript, (2010).
[40] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, Computer Science Department, University of Toronto, Tech. Rep, (2009).
[41] A. Krizhevsky, G.E. Hinton, Factored 3-way restricted boltzmann machines for modeling natural images, International Conference on Artificial Intelligence and Statistics, (2010), pp. 621-628.
[42] A. Krizhevsky, G.E. Hinton, Using very deep autoencoders for content-based image retrieval, ESANN, (2011).
[43] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS, (2012), pp. 4.
[44] Q.V. Le, Building high-level features using large scale unsupervised learning, IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP), (2013), pp. 8595-8598.
[45] Q.V. Le, A. Karpenko, J. Ngiam, A.Y. Ng, ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning, NIPS, (2011), pp. 1017-1025.
[46] N. Le Roux, Y. Bengio, Representational power of restricted Boltzmann machines and deep belief networks, Neural Computation, 20 (2008) 1631-1649.
[47] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86 (1998) 2278-2324.
[48] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, Learning algorithms for classification: A comparison on handwritten digit recognition, Neural networks: the statistical mechanics perspective, 261 (1995) 276.
[49] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, Advances in neural information processing systems, 19 (2007) 801.
[50] H. Lee, C. Ekanadham, A.Y. Ng, Sparse deep belief net model for visual area V2, NIPS, (2007), pp. 873-880.
[51] Q. Li, J. Wu, Z. Tu, Harvesting mid-level visual concepts from large-scale internet images, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2013), pp. 851-858.
[52] J. Lu, J. Zhou, J. Wang, T. Mei, X.-S. Hua, S. Li, Image search results refinement via outlier detection using deep contexts, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2012), pp. 3029-3036.
[53] Y. Lu, W. Zhang, H. Lu, X. Xue, Salient object detection using concavity context, IEEE International Conference on Computer Vision (ICCV), (2011), pp. 233-240.
[54] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, International Conference on Artificial Neural Networks ICANN 2011, (2011), pp. 52-59.
[55] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, Proceedings of the 27th International Conference on Machine Learning (ICML-10), (2010), pp. 807-814.
[56] J. Ngiam, P.W. Koh, Z. Chen, S.A. Bhaskar, A.Y. Ng, Sparse Filtering, NIPS, (2011), pp. 1125-1133.
[57] W. Ouyang, X. Wang, A discriminative deep model for pedestrian detection with occlusion handling, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2012), pp. 3258-3265.
[58] H. Poon, P. Domingos, Sum-product networks for deep learning, Learning Workshop. FL, Fort Lauderdale2011).
[59] H. Poon, P. Domingos, Sum-product networks: A new deep architecture, IEEE International Conference on Computer Vision Workshops (ICCV Workshops), (2011), pp. 689-690.
[60] M. Ranzato, G.E. Hinton, Modeling pixel means and covariances using factorized third-order Boltzmann machines, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2010), pp. 2551-2558.
[61] G. Rätsch, T. Onoda, K.-R. Müller, Soft margins for AdaBoost, Machine learning, 42 (2001) 287-320.
[62] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychological review, 65 (1958) 386.
[63] R. Salakhutdinov, G.E. Hinton, Deep boltzmann machines, International Conference on Artificial Intelligence and Statistics, (2009), pp. 448-455.
[64] J. Shotton, M. Johnson, R. Cipolla, Semantic texton forests for image categorization and segmentation, IEEE Conference on Computer vision and pattern recognition, (2008), pp. 1-8.
[65] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation, European Conference on Computer

Vision–ECCV 2006, (Springer, 2006), pp. 1-15.
[66] K. Sohn, D.Y. Jung, H. Lee, A.O. Hero, Efficient learning of sparse, distributed, convolutional feature representations for object recognition, IEEE International Conference on Computer Vision (ICCV), (2011), pp. 2643-2650.
[67] K. Sohn, H. Lee, Learning invariant representations with local transformations, arXiv preprint arXiv:1206.6418, (2012).
[68] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2013), pp. 3476-3483.
[69] I. Sutskever, J. Martens, G.E. Hinton, Generating text with recurrent neural networks, Proceedings of the 28th International Conference on Machine Learning (ICML-11), (2011), pp. 1017-1024.
[70] K. Swersky, D. Buchman, N.D. Freitas, B.M. Marlin, On autoencoders and score matching for energy based models, Proceedings of the 28th International Conference on Machine Learning (ICML-11), (2011), pp. 1201-1208.
[71] Y. Tang, R. Salakhutdinov, G. Hinton, Robust Boltzmann machines for recognition and denoising, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2012), pp. 2264-2271.
[72] T. Tieleman, Training restricted Boltzmann machines using approximations to the likelihood gradient, Proceedings of the 25th international conference on Machine learning, (2008), pp. 1064-1071.
[73] T. Tieleman, G. Hinton, Using fast weights to improve persistent contrastive divergence, Proceedings of the 26th Annual International Conference on Machine Learning, (2009), pp. 1033-1040.
[74] P.-H. Tseng, R. Carmi, I.G. Cameron, D.P. Munoz, L. Itti, Quantifying center bias of observers in free viewing of dynamic natural scenes, Journal of vision, 9 (2009) 4.
[75] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, Proceedings of the 25th international conference on Machine learning, (2008), pp. 1096-1103.
[76] P. Viola, M. Jones, Fast and robust classification using asymmetric adaboost and a detector cascade, Advances in Neural Information Processing Systems, 2 (2002) 1311-1318.
[77] L. Wan, M. Zeiler, S. Zhang, Y.L. Cun, R. Fergus, Regularization of neural networks using dropconnect, Proceedings of the 30th International Conference on Machine Learning (ICML-13), (2013), pp. 1058-1066.
[78] S. Wang, C. Manning, Fast dropout training, Proceedings of the 30th International Conference on Machine Learning (ICML-13), (2013), pp. 118-126.
[79] Y. Wei, F. Wen, W. Zhu, J. Sun, Geodesic saliency using background priors, European Conference on Computer Vision–ECCV 2012, (2012), pp. 29-42.
[80] T. White, Hadoop: The definitive guide (" O'Reilly Media, Inc.", 2012).
[81] L. Wiskott, T.J. Sejnowski, Slow feature analysis: Unsupervised learning of invariances, Neural computation, 14 (2002) 715-770.
[82] Y. Wu, Z. Wang, Q. Ji, Facial Feature Tracking under Varying Facial Expressions and Face Poses based on Restricted Boltzmann Machines, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2013), pp. 3452-3459.
[83] M.D. Zeiler, R. Fergus, Stochastic pooling for regularization of deep convolutional neural networks, arXiv preprint arXiv:1301.3557, (2013).
[84] M.D. Zeiler, D. Krishnan, G.W. Taylor, R. Fergus, Deconvolutional networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2010), pp. 2528-2535.
[85] M.D. Zeiler, G.W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level

feature learning, IEEE International Conference on Computer Vision (ICCV), (2011), pp. 2018-2025.
[86] J.-Y. Zhu, J. Wu, Y. Wei, E. Chang, Z. Tu, Unsupervised object class discovery via saliency-guided multiple class learning, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2012), pp. 3218-3225.
[87] J.-Y. Zhu, J. Wu, Y. Wei, E. Chang, Z. Tu, Unsupervised object class discovery via saliency-guided multiple class learning, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2012), pp. 3218-3225.
[88] W.Y. Zou, A.Y. Ng, S. Zhu, K. Yu, Deep Learning of Invariant Features via Simulated Fixations in Video, NIPS, (2012), pp. 3212-3220.

致　谢

时光飞逝，一转眼研究生的三年就要成为往事，我的学生生涯也将就此结束。三年间，酸甜苦辣咸五味陈杂，风霜雪雨冷暖自知。在这里由衷感谢那些陪我走过三年的同学、老师，感谢你们的关切、问候、无私的帮助，让我走过所有的困难与挫折，让我最后的三年学生生活变得如此充实。

首先要感谢的是我的导师苗军老师。你们在三年中给了我太多的辅导和帮助，让我在科研的道路上走到现在。苗老师注重细节，在讨论时总能从最细微处发现问题，本文的成型也经过了苗老师细致的修改，这种精益求精的研究态度在研究生期间深深影响着我，让我受益良多。同时苗老师给了我相对自由的科研空间，在研究初期让我可以根据自己喜好选择研究方向，对我的主意与实验给与了许多鼓励与支持，对此深表感激。在以后工作的路上，您的教导将一直指引我前进。

其次要感谢同组的卿老师，您平易近人的生活态度和在研究中深厚的功底在科研上给了我很大的帮助；感谢马志国兄、淮静师兄、孟令勋师兄、王崇秀师姐，作为师兄师姐你们在我刚来实验室就给予帮助，让并不善于交际的我快速熟悉实验室的环境，并在之后的研究生生活中解答了许多生活、研究上的问题；感谢同级同组的袁善欣、帅佳玫，你们算是我在计算所最熟悉的同学了，感谢你们在各个地方的帮助，也感谢与你们一起奋斗的三年时光；感谢朱文涛、李子怿，未来属于你们。

最后要特别感谢我的女朋友，感谢你在三年中对我生活的照顾，对我坏脾气的包容，对我每一个决定无条件的支持和鼓励；感谢你耐心听我不切实际的点子和天马行空的志向，在我心情低落时陪我漫无目标的暴走，乐我所乐忧我所忧。独在异乡为异客，但因为有你，身在异乡的我如归故里。

计算所的三年是我最后的学生生涯，也是我最为转折的三年，在这三年里让我深刻认识到了自身的不足，让我得以正确地看待自己，看待以后的生活，让我从一个毛头小子渐渐长成一个独立自省的人。再一次感谢那些在上面提到的以及没提到的人们，谢谢你们！

作者简介

姓名：王璠　　性别：男　　出生日期：1990.1.19　　籍贯：山东济宁

2011.9 – 2014.7　　中国科学院计算技术研究所计算机应用技术专业硕士生

2007.9 -- 2011.7　　北京航空航天大学软件工程专业本科生

【攻读硕士学位期间参加的科研项目】

[1] 国家自然科学基金面上项目：生物启发的视觉目标搜索和定位研究

[2] 国家自然科学基金面上项目：基于真实视点分析的视觉选择性注意建模

[3] 中科院计算所-Nokia 合作项目：深度学习及其在图像识别中的应用