# 第一章：绪论

## 1.1 研究背景、目的及意义

### 1.1.1 研究背景

近年来，随着信息技术和网络的大力发展，人们对计算机的应用越来越广泛，上至太空旅行，下至买菜做饭，计算机已经渗入到了我们生活的方方面面，然而在这样的信息爆炸的时代，人们都想用最短的时间做尽可能多的事情，计算机给我们带来生活便利的同时也让人们在无尽的信息中苦苦挣扎，因此如何用更少的数据表达更丰富的意义对人们至关重要。在海量数据中提取重要信息，忽略其它辅助性的次要信息成为很多计算机相关领域研究的热点问题。

人们能够在大量不相关的复杂信息的干扰下快速地找到显著性的物体或者感兴趣的目标，是由于人类和其他灵长目动物具有视觉注意机制、选择性注意机制，利用图像的局部特征，首先选定环境景象中某个特定的区域，然后通过眼睛运动扫描，将该区域置于具有更高分辨率的视网膜的中央四区，人们此时将注意力转移到该区域，对该区域的目标或者物体进行更深层次的分析和处理。

经学者们研究表明，人类视觉的选择性注意机制大致分为两个阶段：（自底向上的基于视觉场景中显著度计算的、快速的预注意阶段，这是一种低级的认知过程。（自顶向下的基于调整选择准则来适应外界命令要求的、慢速的注意阶段，这是一种高级的认知过程。人们希望用计算机来模拟这种类似用人眼观察周围场景的活动，因而产生了计算机视觉这门学科。计算机视觉在各个领域都给人们带来了很大的方便的同时也取得了很大的发展，但对于模拟人类视觉机制上仍然存在很大的障碍，其主要原因在于人类本身对于自己在生理和心理上感知周围世界的物体的机制仍然没有一个清晰的轮廓，这也是制约着该领域发展的最大障碍。

在计算机软硬件突飞猛进的今天，提取图像中感兴趣的区域已经应用到了

生活中的方方面面，例如在公安机关追踪肇事车辆逃逸时，能够根据目标车辆的特征对视频中的目标进行快速的蹄选，帮助公安机关在破案时节省了宝贵的时间。

显著性检测是计算机视觉领域中非常具有代表性的问题，它的目的是定位出那些最吸引人视觉注意的像素或区域。随着信息科技的发展与快速交通工具的推广，人类所接收的信息量呈指数级增长，如何筛选出人类感兴趣的目标和区域具有重要的意义。显著性区域与人视觉感知关系极为紧密， 并具有一定的主观性， 开展显著性检测的研究非常有利于图像处理基本任务的完成。例如，能够大致定位出兴趣目标，减小图像分割与检测的搜索范围和计算量; 精确的显著性检测使目标跟踪与识别更加容易

### 1.1.2 研究目的

期望使用深度学习算法，从有标记图像数据库中学习先验知识，用于提升显著性估计算法的性能。

1，用已有的深度学习库（cuDNN,caffe）来训练深度网络结构（cnn,auto-encoder）参数。

2，提取现有的大规模图像数据库中图像的objectness特征，objectness是 Alexe 等人提出，表示一个像素或区域包含完整目标的可能性。一个图像中所含目标通常包含下面特点：1)包含一个空间闭合的边界；2)和周围事物有不同的形态；3)有时是独特、突出的。人眼倾向于完整的去识别一个目标是否显著，因此评估每个区域是否属于可识别区域，将这些区域作为先验知识，在显著性检测的时候，可以将这些区域的显著性系数增大，结合深度学习得到的显著图来确定最终的显著图。从而提高显著性检测的准确性。

3，采用自上而下(Top-down) 视觉分析计算模型，来实现显著性的检测。

4，采用自上而下(Top-down) 视觉分析计算模型，基于现有规模图像数据库ImageNet，iSUN，SALICON，MSCOCO等，从现有图像中学习先验知识，进而提高显著性检测的准确性。 目标的抽象（语义）概念对于显著性估计非常重要，由于图像库MSCOCO标记出了目标的轮廓，我们可以先通过深度学习获得目标概念的描述（CALVIN），然后将概念语义加入到显著性估计的过程中，从而提升性能和准确性。 更进一步的，由于图像中有目标轮廓的标记，我们可以进一步标记出目标内部与非目标区域，从而可以学习二类分类器，区分是否有意义的目标。这个可以应用于图像中的任意边线，即给定边缘，可以判断是边缘的哪一侧为目标区域。

### 1.1.3 研究意义

借鉴人类视觉注意机制，计算机视觉领域的研究人员展开了对适用于计算机模拟的视觉注意模型的研究。他们通过将视觉注意模型作为机器视觉系统中进行信息筛选和计算资源优先分配的关键组件，用于提升计算机视觉系统处理海量数字媒体的能力，提高数字媒体资源的利用率。构建有效的视觉注意模型首先需要解决的关键性问题就是视觉显著性检测。然而目前由计算机模拟的视觉显著性检测算法的检测结果准确度还不高，这将会限制计算机视觉系统的应用效果和应用范围。因而构建检测性能更好的视觉显著性检测算法是目前计算机视觉领域亟待解决的关键性问题。

## 1.2 显著性研究现状

目前，国内外的许多研究机构对图像显著性区域检测做了广泛的研究，并取得了一定的成果。显著区域检测的本质是一种视觉注意模型。该模型利用视觉注意机制得到图像中最容易引起注意的区域，同时用灰度图像来作为该图像的显著性度量。人类视觉系统能够很快的从复杂的场景中检测到感兴趣的目标，而如何建立机器视觉注意模型来模拟人的视觉系统在图像处理领域得到极大关注。随着认知心理学和神经生物学的发展，人们对视觉注意机制以及视觉神经系统有了更深入的了解和认识，奠定了构建视觉注意机制模型的理论基础。 在复杂的场景中，人类视觉系统釆用串行计算，快速的将注意力转移到自己感兴趣的一个或者几个目标身上，优先处理这些目标区域，这就是视觉注意过程。视觉注意机制在人类的视觉系统中起到重要的心理调节作用，属于视觉感知模型中的一部分，协同记忆等模块完成各项任务，如目标匹配、目标分离以及注意焦点转移。从信息角度来说，视觉神经系统处理信息的资源以及能力都是有限的。所以，人类通过这种选择注意机制对海量的视觉信息进行处理，从而快速地进行分辨、筛选出重要的信息。 模拟人类视觉系统的显著性区域检测计算模型主要从两个角度出发：（1）自下而上（Bottom-up）的视觉分析计算模型；(2)自上而下(Top-down)的视觉选择注意模型。其中，前者基于低级视觉特征，即没有任何的先验知识，由数据驱动，自动捕获刺激人眼的区域，而后者则基于高级视觉特征，即先验知识的学习，由知识和任务驱动，以自我意识决定视觉关注区域。研究表明，在人类视觉注意的具体过程中，自下而上和自上而下的选择注意方式通常是共同协作、相互影响的。下面对这两种视觉分析模型进行简要介绍：

1）自下而上（Bottom-up）的视觉分析计算模型 许多研究人员试图通过描绘观察者感兴趣的区域和一些基本特征（如边缘和局部对比性）之间的关系来对显著性区域进行解释。其中，最具代表性的是Itti等人在年提出的视觉注意模型。该模型基于特征集成理论，采用高斯金字塔并综合考虑亮度、颜色、方向视觉特征，通过算子得到多尺度的显著性度量，最后经过合并、归一化得到最终的显著图像[1]；在Itti工作的基础上，Walther等人将其进行扩展，成功应用于目标识别，还建立了显著性检测网站并提供显著性测量工具箱；Harel等人提出基于图的视觉检测显著模型来突出显著区域并且可以结合其他图形[2]；Ma和Zhang等人提出利用颜色特征对比分析，并提出利用模糊增长算法提取图像的显著性区域。Gao和Mahadevan等人提出了基于中心-周围判别理论的显著性检测模型，简称为DICS模型。Chen Xia等人提出一种基于非局部重构的显著性区域检测方法，并通过利用非局部其余块对当前中心区域的重构误差来估计显著性[3]。同时Chen xia等人在前篇文章的基础上又提出了一种基于深层自动编码器重构的图像视觉显著性区域检测方法，以从全局的角度来挖掘无标记图像数据与显著性值之间的关系，提高图像中显著性区域检测的准确性[4]。

2) 自上而下(Top-down)的视觉分析计算模型 自上而下的视觉分析计算模型的基本思想是从场景中提取出多种基本特征，随后将其集成作为显著图像的表示。受可视化程序的启发，Sprague和Ballard提出了基于“加强学习”的自上而下显著检测模型，显示观察者在虚拟环境中的眼球运动；Kanan等人提出利用自认数据估计目标在每个位置概率的算法；Liu等人提出一种具体的学习机制，其通过条件随机场（自动学习出对各低层特征（包括局部特征、全局特征、区域特征）计算出的显著度权重，从而依据权重计算出图像各个位置总体的显著程度，并且依此进行显著物体的提取[5]。为了更加协调的将多种特征结合到显著性检测算法当中，Lang等人将显著性区域检测认为是一个多任务学习的问题，并提出了多任务稀疏追踪算法[6]。

## 1.3 显著性的相关应用

图像的显著性检测能够降低对于图像内容的理解与图像的低层特征之间存在的隔阂，这为其他层次的图像处理提供了极大的便利，例如图像分割【8】，目标检测，图像和视频压缩，目标重定位、图像编辑

[11-14]

等等。除此之外，还可以应用到

车牌和路标的检测、高性能脑模型、高智能武器系统和医疗设备等方面。下面列举

出了两个图像显著性区域在其他图像处理过程中的应用。

显著性区域检测技术有着广泛的应用，在目标自动检测⑴、图像检索、物体识别图像分割视频的快速浏览和汇总、图像和视频压缩、图像自动化修剪和内容感知的图像编辑】】】』〃】等领域都可以应用。目前显著性检测技术的应用还无法令人满意，除了显著图本身的质量不够高以外，显著性信息应用方式和方法还不够成熟，需要找到更加令人满意的实现方法。

## 1.4 本文的主要内容

本文的工作主要分为三步：

1. 提取先验知识。这里的先验知识主要是指图像的目标性。图像目标性表示图像中的一个像素或者一快区域包含完整目标的可能性。一张图像中所含目标通常包含下面特点：1)包含一个空间闭合的边界；2)和周围事物有不同的形态；3)有时是独特、突出的，这些特点和显著的概念大体上一致。

（2）构建深度学习网络。由于目前卷积神经网络在图像处理方面的效果特别好，同时自动编码器也有很好的表现，拟从这两种网络结构中选取一种比较好的结构，在图像数据库上训练，得到一种能够很好表示图像显著性的结构及参数。

（3）先验知识与深度学习网络结构的融合。结合步骤（1）及步骤（2），来提高整个图像的显著性效果，最简单的结合方式是将步骤（1）中的图像目标性按照一定的权重加入到步骤（2）的显著图中。另外也可以将步骤（1）中得到的目标性特征直接加入到训练图像中去，将步骤（2）的训练作为预训练，根据加上先验知识的数据，最后一起来训练深度网络。

本文的主要内容安排如下：

第一章是绪论，主要是提出本论文的研究背景及意义，分析了研究现状，并给出了整个论文的整体思路。

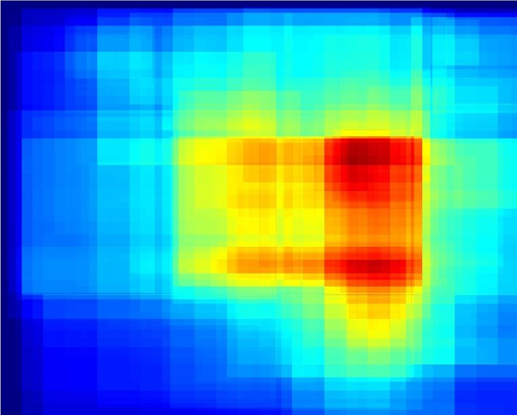
第二章是论文相关的知识介绍

# 第二章：相关知识介绍

## 2.1 图像目标性介绍

### 2.1.1 什么是图像目标性

图像目标性最开始是Bogdan Alexe【9】等人提出来的，其中目标是指具有明确定义的边界，且具有独立的东西和中心，如汽车、电话和人等，而不是没有形态的背景环境，如天空、草坪和道路等。图像目标性是指对于图像中的一个区域，该区域是目标的可能性有多大，具体到图像中的每一个像素点，是指该像素点被图像中某目标包含的可能性大小。一张图像的目标性表示方式为一张对应大小的单通道黑白图像，其中图像中像素点越亮的地方表示对应原图中目标性越大，像素点越暗的地方表示对应原图中的目标性越小。

原图 目标性图

### 2.1.2 目标性的意义

通常，一般人会主动去判断一张图像中的目标是否显著，以此来确定该目标区域是否具有大量信息，是否是可识别区域，然后再对该区域做细致分析，得到我们需要的图像信息。由Bogdan Alexe提出的目标性我们可以知道，一张图像中的目标往往都有一定的明显特征：1）目标基本都是被一个封闭的区间所包含；2）图像中目标区域和周围区域有着明显不同的形态；3)目标都是突出显著的，有时候是独一无二的【10】。许多目标对象都同时具有上述特征中的几个或者全部。针对显著检测中特征的适应性不足以及当前一些算法出现多检与漏检的问题，提出从“目标在哪儿”与“背景在哪儿”两个角度描述显著性的框架，进行特征融合来提高显著目标检测的准确率。

### 2.1.3 目标性的获得方法

对于图像目标性的先验知识，现在已有很多的研究和方法实现，一般是通过图像底层的基本特征，如图像局部颜色，轮廓，方向梯度来获得相关特征，得到该图像的目标性特征图。

（1）“特征融合与 objectness 加强的显著目标检测”【11】提出了一种先利用mean-shift分割算法分割图像区域，来保持每个目标的 objectness 特征具有一致性，其次计算 objectness，过程分为两步：评估像素级 objectness 和区域级 objectness。

计算每个像素的 objectness 特征，需要在图像上随机分布 N 个窗口，每个窗口w用文献[3]的方法计算 objectness 得分，并记为 P(w)。随后对所有窗口W，统计包含每个像素点的窗口的 objectness得分，以此获取每个像素的 objectness 特征，公式如下：



其中 w 表示 W 中任意包含像素 x 的窗口。实验中 N 设置为 10000。其次，计算每个区域i 的objectness 特征：



计算出所有区域的objectness 特征后，将特征值分配到区域所包含的像素上。最后，得到整个图像的objectness 特征，标记记为。

其实现效果如图所示

原图 objectness图

（2）

## 2.2 图像显著性介绍

### 2.2.1 什么是图像显著性

目前图像显著性的求解方法有很多种，从原理角度上大致分为三类，分别是仿生物学的方法、纯计算的方法、仿生物学与纯计算相结合的方法；同时从计算角度又分为自上而下、自下而上两中方法。

### 2.2.2 图像显著性的获取方法

（1）“[A Model of saliency Based Visual Attention for Rapid Scene Analysis](http://blog.csdn.net/chenjiazhou12/article/details/39456589)”【12】这篇文章是图像显著性领域最具代表性的文章，是在1998年Itti等人提出来的，到目前为止引用的次数超过了5000，在它的基础上发展起来的有关图像显著性论文更是数不胜数，论文的提出主要是受到灵长类动物早期视觉系统的神经结构和行为所启发而产生了视觉注意系统。灵长类动物具有很强的实时处理复杂场景的能力，视觉信息进行深入的处理之前，对所收集到的感觉信息进行选择，这些选择可能减少场景理解的复杂性，这个选择过程在一个空间有限的视野区域即所谓的注意焦点（focus of attention， FOA）中完成的，它搜索场景的方式可以是快速、自下而上（bottom-up，BU）、显著驱动和任务独立的方式，也可以是慢速、自上而下（top-down，TD）、意志控制和任务相关的方式。注意模型包括“动态路由”模型，在此模型中，通过皮层视觉继承，从一个小的视野区域中得到的信息可以不断前行。通过皮层连接的动态修正或在TD和BU的控制下对活跃区建立特定的瞬时模式，来选择注意区域。

这篇文章所使用的模型建立在由Koch和Ullman所提出的生物模型和其他几个模型。它和所谓的特征整合模型相关，解释了人类视觉搜索策略。视觉输入首先被分为一系列特征地形图。然后在每个图中，不同的空间位置通过竞争获取显著性，只有从周围脱颖而出的位置才能保留。所有的特征图以纯粹的BU方式输入到高级的显著性图，它对整个视觉场景的局部醒目度进行地形编码。在灵长类动物中，人们相信这种图存在于后顶叶皮层（posterior parietal cortex）中，同时也存在于枕核丘脑（pulvinar nuclei ofthalamus）中。模型的显著性图被认为是产生注意转移的内部动力。因此这个模型表示了BU显著性可以引导注意转移，不需要TD。这个模型在可以进行并行处理，提高运算速度，而且可以根据特征的重要性，为特征加上权值，特征越重要，权值越大。

（2）“Saliency Detection: A Spectral Residual Approach”【5】提出了一个图像视觉显著性的简单计算模型，这个模型和Irri提出的模型是两个截然不同的模型，Irri模型对于图像视觉显著性主要关注整幅图片突出的部分，通过各种特征的融合提取显著性图，而Hou的这个模型直接关注的点就不在一张图片里突出的地方，而是背景，观察是否大部分图片的背景在某个空间上都满足什么变化，最后剔除背景，自然就只剩下图片突出的部分了，没有很复杂的公式变换，都是图像处理基本的变换公式。

Barlow提出的有效编码假说指出，感知系统的信息处理过程与外界信号的统计特性密切相关。初级感知系统的作用就是去除输入信号的统计冗余。也就是说，大脑之所以能够对外界环境自适应，是因为复杂的外界刺激存在冗余，而大脑的神经元能够有效地去除这些冗余，从而可以利用较少的资源尽可能有效地表达更多的信息。

从信息论的角度来看，有效编码假说将图像信息H(Image)分为两部分：

H(Image)=H(Innovation)+H(Prior Knowledge)

H(Innovation)表示突出的部分，H（Prior Knowledge）则表示冗余的信息。通过去除图像冗余信息，就可以获得图像与众不同的部分，即显著目标。

自然图像的统计特性具有变换不变性：即将图像从原来的空间坐标变换到频率坐标系中，图像在空间中具有的统计特性在频域中仍然保留，这种不变性恰好保证了采用能量谱来刻画自然图像空间相关性的可靠性。

在自然图像的统计特性中尺度不变性是最经典也是研究最广泛的特性，这种特性也被称为1/f法则，即自然图像集合的平均傅里叶谱的幅值A(f)服从下式的分布：



对于一个系统模型而言，考虑输入刺激信号的统计相似性可以极大地减少多余的视觉信息，这是因为相似性意味着冗余。不同的图像数据，其log谱却有着相似的分布趋势，而且曲线满足局部线性条件。因此，在不同的log谱中我们只需要关注其差异部分忽略相似部分。

给定一幅图像I(x)首先计算2维离散傅里叶变换，将其从空间域转到频域，并得到振幅谱和相位谱，对幅值取对数后得到log谱，变换公式如下：







其中表示傅里叶变换，表示求傅里叶变换后的幅值。

由于log曲线满足局部线性条件，所以用局部平均滤波器h(f)对其进行平滑，获得平均频谱：



其中是一个的矩阵，定义为



因此谱残差就是log谱和其进行均值滤波后的差，可按下面的式子计算：

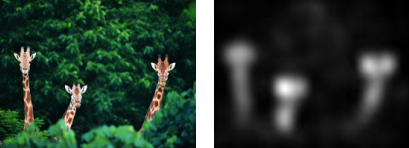


最后将谱残差和相位谱相加，并求出自然指数，然后对其进行傅里叶反变换，最后经过一个高斯模糊滤波就得到了所谓的显著性区域值，具体公式如下：



其中，g(x)为高斯模糊模板，为傅里叶反变换。

得到的图像显著图如图所示：



## 2.3 自动编码器介绍

由于本论文中的图像显著性是基于自动编码器实现的，所以这里对自动编码器的理论公式及训练过程给出简单介绍。

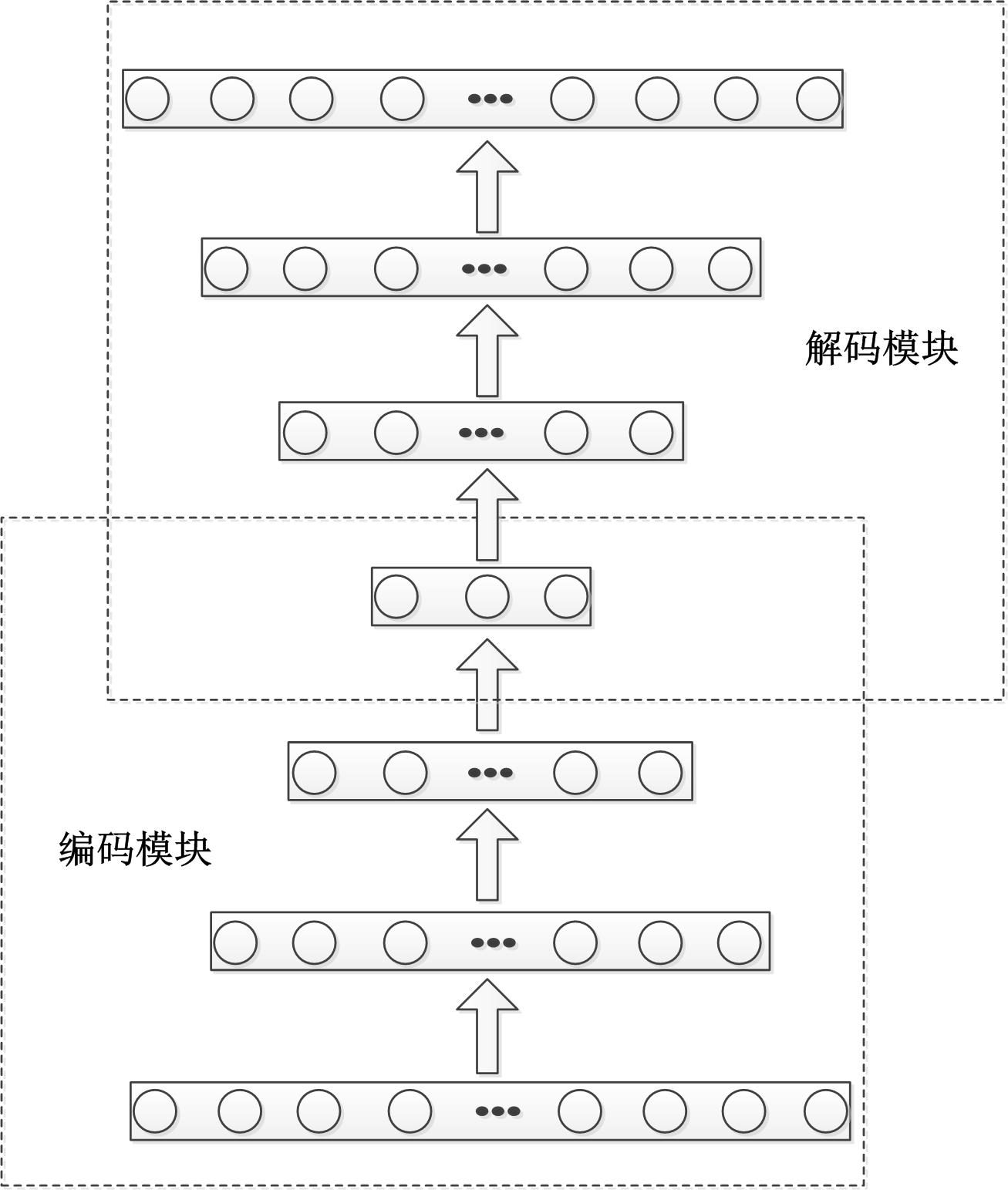
### 2.3.1 什么是自动编码器

无论对于图像，语音，文本，机器学习都有很成熟的方法来解决这些领域的系列问题。机器学习系统被用来识别图片中的目标，将语音转换成文本，匹配新闻元素，根据用户兴趣提供职位或产品，选择相关的搜索结果。逐渐地，这些应用使用一种叫深度学习的技术。传统的机器学习技术在处理未加工过的数据时，体现出来的能力是有限的。几十年来，想要构建一个模式识别系统或者机器学习系统，需要一个精致的引擎和相当专业的知识来设计一个特征提取器，把原始数据（如图像的像素值）转换成一个适当的内部特征表示或特征向量，子学习系统，通常是一个分类器，对输入的样本进行检测或分类。特征表示学习是一套给机器灌入原始数据，然后能自动发现需要进行检测和分类的表达的方法。深度学习就是一种特征学习方法，把原始数据通过一些简单的但是非线性的模型转变成为更高层次的，更加抽象的表达。通过足够多的转换的组合，非常复杂的函数也可以被学习。对于分类任务，高层次的表达能够强化输入数据的区分能力方面，同时削弱不相关因素。比如，一副图像的原始格式是一个像素数组，那么在第一层上的学习特征表达通常指的是在图像的特定位置和方向上有没有边的存在。第二层通常会根据那些边的某些排放而来检测图案，这时候会忽略掉一些边上的一些小的干扰。第三层或许会把那些图案进行组合，从而使其对应于熟悉目标的某部分。随后的一些层会将这些部分再组合，从而构成待检测目标。深度学习的核心方面是，上述各层的特征都不是利用人工工程来设计的，而是使用一种通用的学习过程从数据中学到的。

所以，深度学习正在取得重大进展，解决了人工智能界的尽最大努力很多年仍没有进展的问题。它已经被证明，它能够擅长发现高维数据中的复杂结构，因此它能够被应用于科学、商业和政府等领域。除了在图像识别、语音识别等领域打破了纪录，它还在另外的领域击败了其他机器学习技术，包括预测潜在的药物分子的活性、分析粒子加速器数据、重建大脑回路、预测在非编码DNA突变对基因表达和疾病的影响。也许更令人惊讶的是，深度学习在自然语言理解的各项任务中产生了非常可喜的成果，特别是主题分类、情感分析、自动问答和语言翻译。我们认为，在不久的将来，深度学习将会取得更多的成功，因为它需要很少的手工工程，它可以很容易受益于可用计算能力和数据量的增加。目前正在为深度神经网络开发的新的学习算法和架构只会加速这一进程。

深度学习最简单的一种方法是利用人工神经网络的特点，人工神经网络（ANN）本身就是具有层次结构的系统，如果给定一个神经网络，我们假设其输出与输入是相同的，然后训练调整其参数，得到每一层中的权重。自然地，我们就得到了输入I的几种不同表示（每一层代表一种表示），这些表示就是特征。在研究中可以发现，如果在原有的特征中加入这些自动学习得到的特征可以大大提高精确度，甚至在分类问题中比目前最好的分类算法效果还要好！这种方法称为自动编码器。自动编码器就是一种尽可能复现输入信号的神经网络。为了实现这种复现，自动编码器就必须捕捉可以代表输入数据的最重要的因素，就像PCA那样，找到可以代表原信息的主要成分。

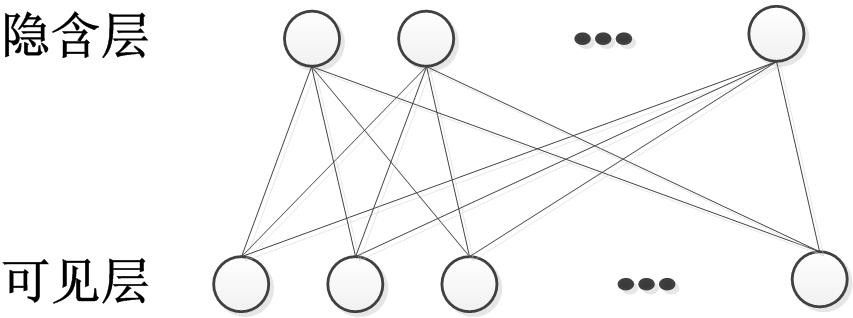
自动编码器整体上是一个对称结构，第一层和最后一层节点数相同，第二层和倒数第二层节点数相同，其它层同理，最中间的一层为特征层，即特征层为整个网络的对称中心，所以，整个网络的输入和输出是相同的。



在结构上，自动编码器是由多层RBM（限制波尔兹曼机）叠加构成， 一个RBM相当于自动编码器的一层。RBM是基于统计力学的随机神经网络，它包含两个层次分别为可见层和隐藏层；两层各有一些随机的神经元，这些神经元一般拥有未激活和激活两种状态，RBM学习能力强，能从训练集合中学习到复杂的规则，学习过程是无监督的。

其输入成为可见层，输出为隐藏层，两层之间的神经元两两相互连接，层间的神经元不连接。层内无连接是指任意来自同一层次的两个神经元之间不存在连接；层间全连接指的是对于任意的一个来自可见层的神经元和来自隐藏层的神经元它们之间必然存在链接

其结构图如下所示：



RBM

RBM的用途主要是两种，一是对数据进行编码，然后交给监督学习方法去进行分类或回归，二是得到了权重矩阵和偏移量，供BP神经网络初始化训练。

第一种可以说是把它当做一个降维的方法来使用。

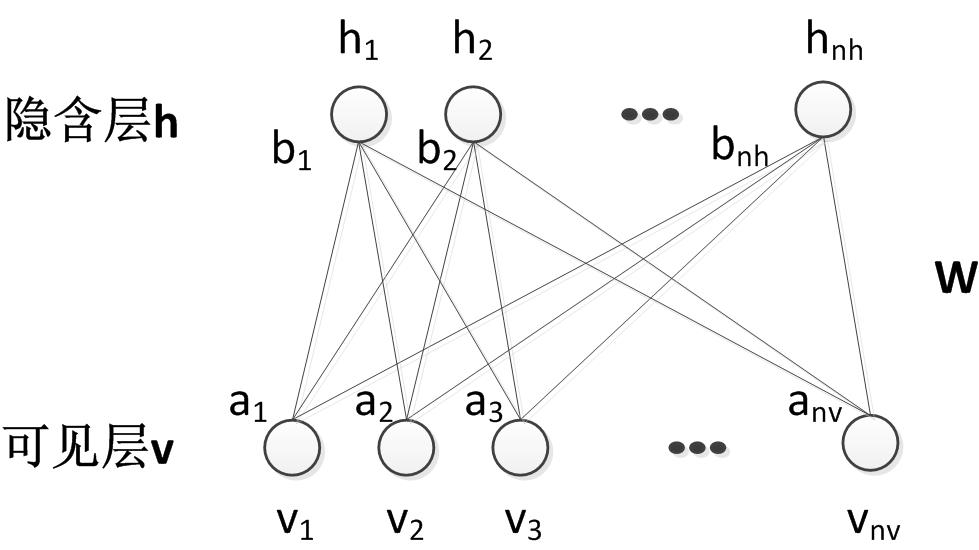
第二种就用途比较奇怪。其中的原因就是神经网络也是要训练一个权重矩阵和偏移量，但是如果直接用BP神经网络，初始值选得不好的话，往往会陷入局部极小值。根据实际应用结果表明，直接把RBM训练得到的权重矩阵和偏移量作为BP神经网络初始值，得到的结果会非常地好。这就类似爬山，如果一个风景点里面有很多个山峰，如果让你随便选个山就爬，希望你能爬上最高那个山的山顶，但是你的精力是有限的，只能爬一座山，而你也不知道哪座山最高，这样，你就很容易爬到一座不是最高的山上。但是，如果用直升机把你送到最高的那个山上的靠近山顶处，那你就能很容易地爬上最高的那座山。这个时候，RBM就的角色就是那个直升机。

这里自动编码器训练时就是利用了RBM的第二个用途，来初始化自动编码器每层之间的参数。

### 2.3.2 自动编码器训练过程

由于自动编码器是由RBM逐层叠加得到的，所以这里先介绍RBM的训练过程，然后再介绍怎么利用RBM来训练自动编码器。

（1）RBM训练过程



如上图所示假设可见层的神经元的个数为，隐含层的神经元个数为，v表示的是可见层神经元的状态，；h表示的是隐藏层的神经元的状态，；a表示的是可见层神经元的偏置，；b表示的是隐藏层神经的偏置，；W表示的是隐藏层与可见层之间的连接权重，；同时，我们记，表示RBM训练过程中需要调整的参数。

首先，对于一组给定的状态（v,h），包括可见层神经元的状态和隐藏层神经元的状态，我们定义这之间的能量函数为，其计算公式如下：



利用该能量公式，我们可以得到可见层神经元的状态和隐藏层神经元的状态的联合概率分布，其计算公式如下：



其中为归一化因子，其计算公式如下：



当得到了可见层神经元的状态和隐藏层神经元的状态的联合概率分布之后，我们就可以得到边缘概率分布函数，即可见层神经元的状态的边缘概率分布函数和隐藏层神经元的状态的边缘概率分布函数，其计算公式如下：





有了上述的可见层神经元的状态和隐藏层神经元的状态的联合概率分布以及可见层神经元的状态的边缘概率分布函数和隐藏层神经元的状态的边缘概率分布函数，我们就可以知道，当给定可见层的状态时，隐藏层上的某一个神经元被激活的概率，即，或者当给定了隐藏层的状态时，可见层上的某一神经元被激活的概率，即。

对于RBM模型，其参数主要是可见层和隐藏层之间的权重W，可见层的偏置a以及隐藏层的偏置b，即，对于给定的训练样本，通过训练得到参数，使得在该参数下，由RBM表示的概率分布尽可能与训练数据相符合。优化参数时可以用最大似然函数来求解，利用梯度上升算法来对参数不断迭代更新，使得RBM的参数不断优化。但是这种方法需要经过很多步的状态转移才能保证采集到的训练样本符合目标分布，所以后来Hinton教授于2002年发明了对比散度（Contrastive Divergence,CD）算法【13】，该方法已经成了目前训练RBM的标准算法。

CD算法的训练步骤很简单，思路也很明确，但训练效果却很好，也大大减小了RBM整体参数的训练时间。具体可以描述如下：

对于，首先进行初始化，初始化时可以采用均值为0，方差为1的高斯分布随机初始化，然后对v进行k步Gibbs采样，k是人为设定的一个值，具体根据RBM解决的不同的问题给出不同的值，依据标准是使得合适的k使得RBM的参数能够训练的比较好。然后对k步Gibbs采样中的每一步t(t=1,2,3…,k)，先后利用可见层的神经元的状态边缘概率分布函数采样得到隐藏层神经元的状态，再利用隐藏层的神经元的状态边缘概率分布函数采样得到可见层神经元的状态，其计算公式如下：







其中为参数学习速率。

然后根据得到的可见层的神经元的状态及隐藏层的神经元的状态来更新参数，具体更新公式如下：







其中为参数的学习速率。

通过k步的Gibbs采样及训练，RBM的参数也不断的在进行更新和优化，使得整个RBM的性能不断提高。

1. 自动编码器训练过程

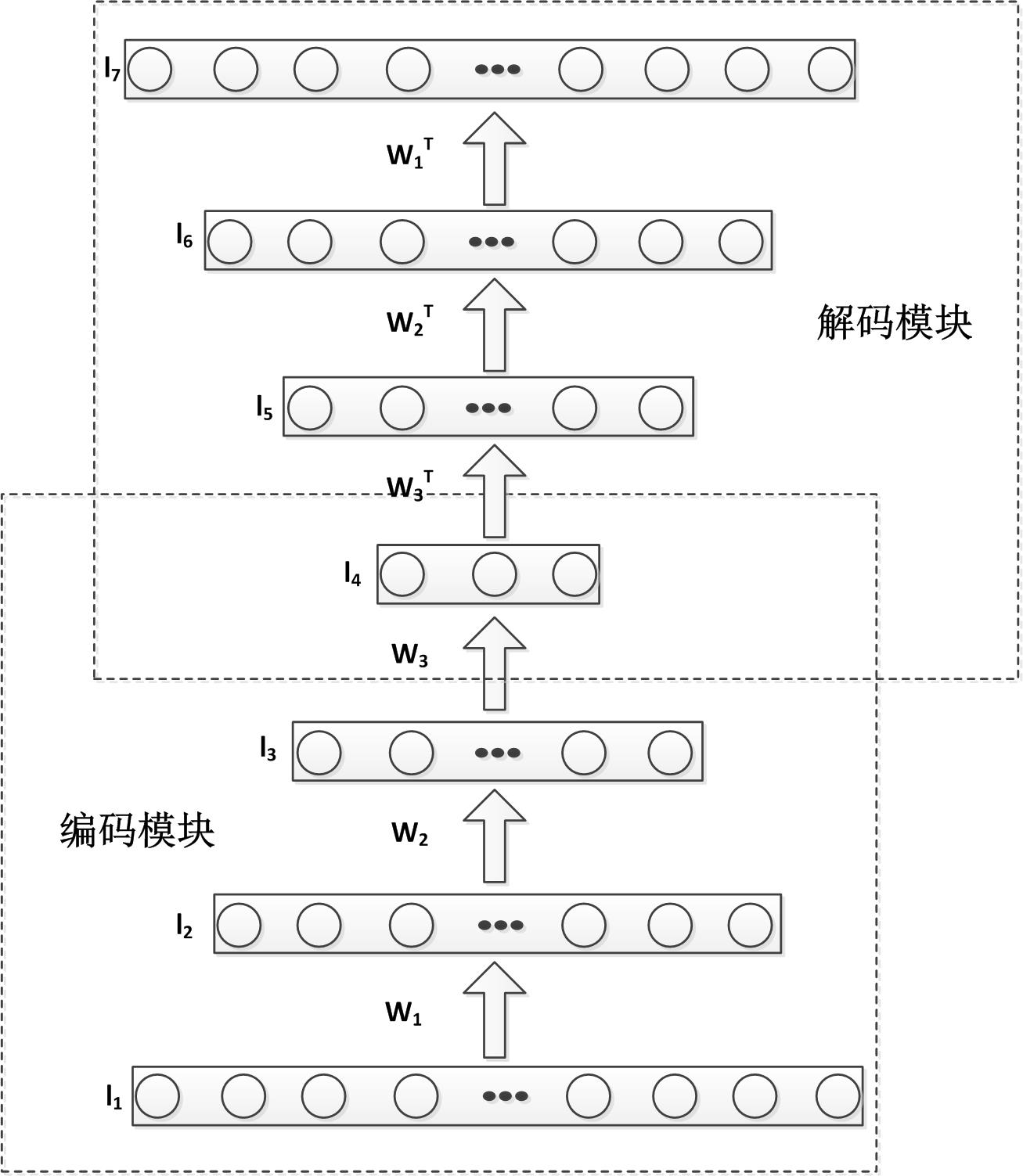
前面提到自动编码器是由多个RBM叠加组合得到的，并且以最中间的特征层为对称中心而两端相互对称的。所以在训练自动编码器时，我们只需要训练整个编码模块，由于编码模块和解码模块相互对称，所以解码模块的参数直接是编码模块的转置得到的，然后将多个RBM的参数组合得到一个自动编码器，最后对组合得到的自动编码器进行微调，即对整个网络利用反向传播算法重新调整一下。所以自动编码器的训练过程主要有三个步骤，具体过程如下：

步骤一，单独训练多个RBM。



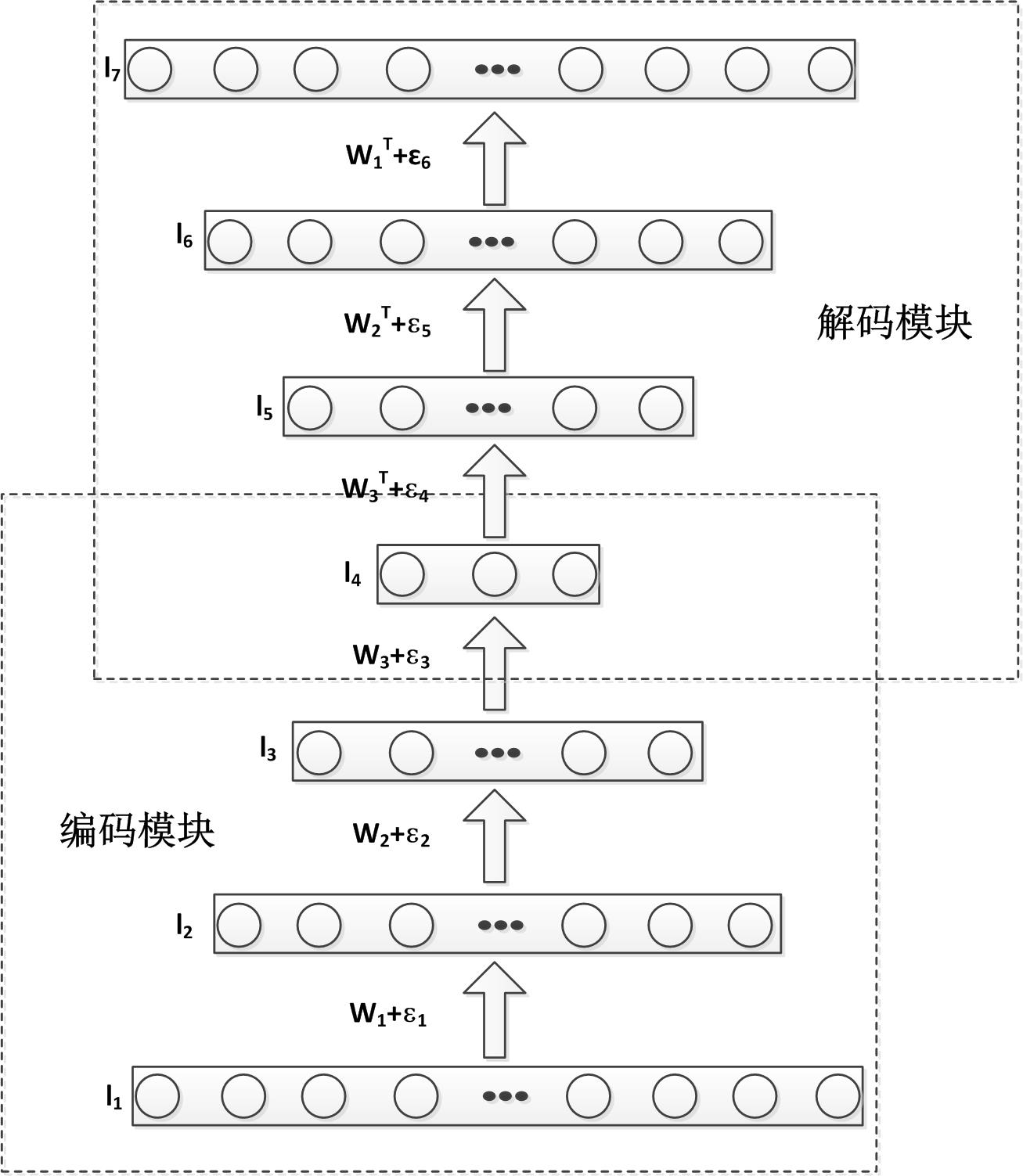
首先，将自动编码器的编码模块分成多个RBM，如上图所示，将编码模块分成了3个RBM，对与每个RBM按照上文中提到的CD算法训练得到RBM的连接权重。

步骤二，将训练好的多个RBM连接权重组合得到预训练的自动编码器。



如图所示，步骤一中通过单独训练RBM我们先得到了自动编码器的编码模块的层与层之间的连接权重W1，W2，W3。前面介绍自动编码器时我们也提到自动编码器是由多个RBM对称叠加得到的，所以解码模块的权重直接为对应位置上的编码模块的权重的装置，依次为，，。这一步在自动编码器的训练过程中也称为网络的预训练，因为对于很深的网络，所有参数一起训练时会出现各种问题，如过拟合，训练时间过长，梯度衰减等。利用预训练先得到自动编码器每层之间的参数，可以很好的避免这些问题。

步骤三，对预训练的自动编码器进行微调，得到最终的网络参数。



微调阶段主要是用来对预训练的参数进行整体调整，一般调整的幅度比较少。主要目的是让预训练的参数能够一起协调使用，来适应我们的训练目标。如上图所示，微调之后，在每层的参数后面会多一个调整变量。微调阶段主要是利用SGD（随机梯度下降算法）来实现的。微调结束后，整个自动编码器的训练阶段也就结束了。

### 2.3.3 自动编码器在图像显著性上的应用

自动编码器的主要作用有两点，一是由于自动编码器的输入和输出是相同的，所以可以用来对图像或者其它的信号进行重构；而是由于自动编码器是一个对称结构，整体关于中间特征层而对称，利用这点信息可以用来对数据降维或者升维，当中间特征层的神经元个数少于输入层神经元的个数或者小于输出层神经元的个数，这个时候显现出的作用是对输入数据进行降维，效果和PCA(主成分分析)算法一样。放中间特征层神经元个数多于输入层神经元的个数或者小于输出层神经元的个数，这个时候显现出来的作用是对输入数据进行升维。

在本论文中，我们是需要利用自动编码器来处理输入的彩色图像，得到图像的显著性图。这里是利用的自动编码器的第一个特征，即对图像的重构。具体思路是，对于任意一张图像的任意一个像素点，取其周围的一个大块来重构其中心的小块，然后比较重构得到的中心小块与原始的中心小块的残差。如果这个残差比较小的话，说明该像素点处的显著性比较小，如果这个残差比较大的话，则说明该像素点处的显著性比较明显。具体解释为，一般对于图像中的背景模块，也就是非显著模块，一般像素点的外围块和中心块是很相似的，在颜色，轮廓等其它特征方面也没有特别大的变动，所以，可以用该像素的外围块很好的重构它的中心块，所以重构出来的中心块和原始的中心块的残差会比较小，证明该处的显著性比较小；但对于图像中的前景部分，也就是显著性部分，像素点的外围块和中心块往往具有很大的差异，无论颜色，方向梯度，轮廓等都有突出的差异，这时用该像素点的外围块是不能很好的重构出该处的中心块的，此时重构出来的中心块和原始的块的残差就会比较大，证明该处的显著性比较大。

## 2.4 本章小结

本章先介绍了图像目标性的定义，然后又给出了目标性的意义以及图像目标性的求解过程，其中的求解过程主要是参考以往的算法，让我们清楚的了解了本论文中比较关键的部分；接下来又介绍了图像显著性，主要是给出了一般获得图像显著性的方法，及这些方法在求解思路上的依据；最后介绍了深度学习中的一种特殊网络自动编码器，并对自动编码器的基本组成部分RBM给出了推导过程，同时给出了自动编码器的训练过程，因为本论文中的显著性的求解是基于自动编码器，所有最后给出了我们为什么采用自动编码器的原因。

# 第三章：图像目标性和显著性

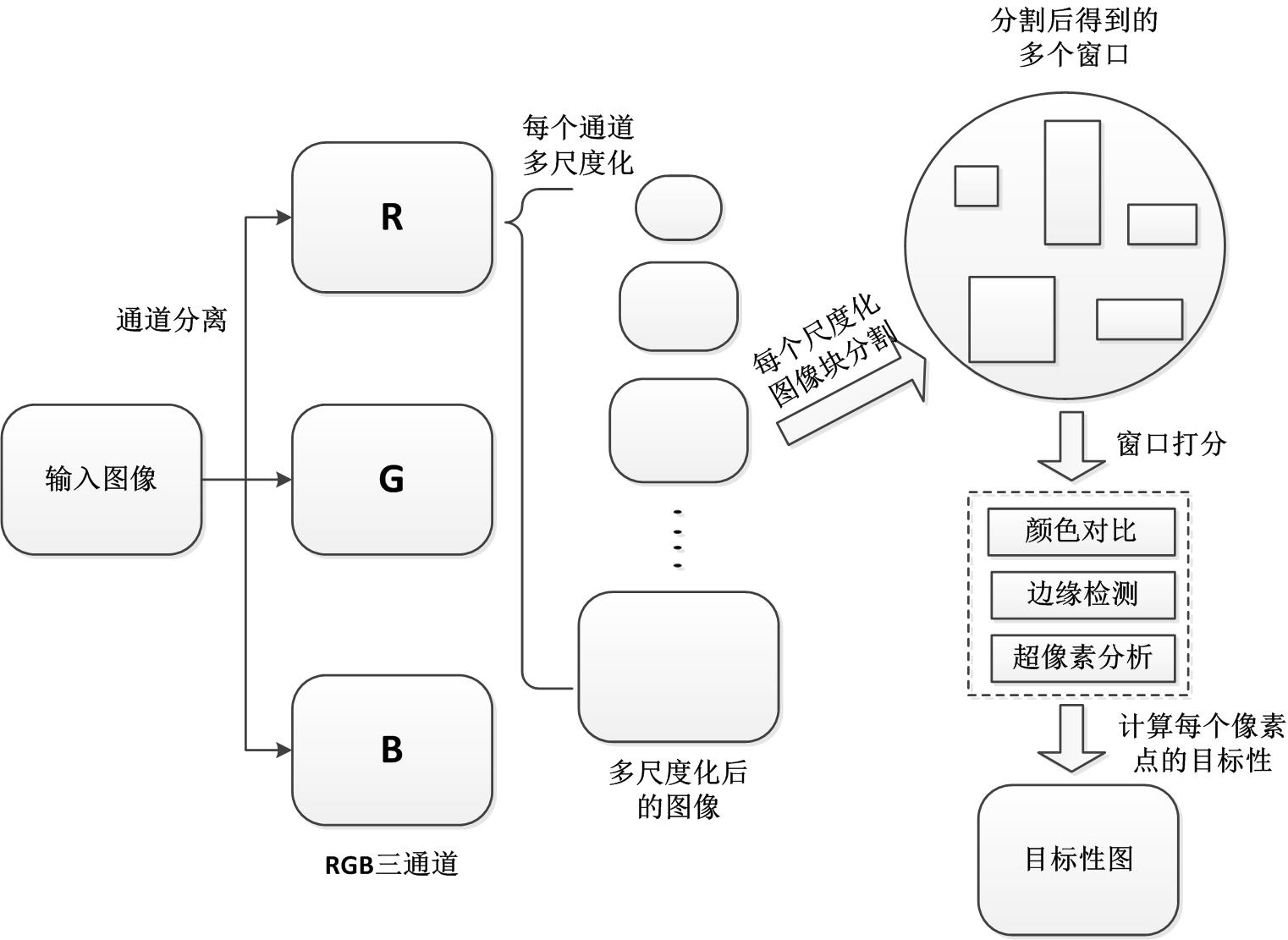
在第二章中我们首先着重介绍了图像的目标性及图像的显著性，再熟悉了目标性和显著性之后，我们又介绍了多个现有方法中图像目标性及图像显著性获取方法，这些方法也都有着不错的检查效果。在本章中，我们会详细介绍本论文完成过程的图像目标性的实现方法及图像显著性的实现方案。

## 3.1 获取图像的目标性

在前面的介绍中，已提到图像的目标性是指图像的一个像素点或者一块区域为一个目标的组成部分的可能性。对于一块区域的目标性，我们很好理解这个定义，即这块区域是一个完整目标的可能性有多大，或者这块区域是一个完整目标的一部分的可能性有多大；而对于一个像素点的目标性，则稍微难理解些，即这个像素点被图像中任意一个目标所包含的可能性有多大。这里反复会提到目标，那什么是目标了？在目标性检测中，目标并不知所有的物体，如大片的天空，如远处的草地。这里的目标本意上是指所有的前景物体，如图中的行人，车辆，交通牌等。一般一个图像中所包含的目标一本含有三个特点：（1）包含一个空间闭合的边界；（2）和自己周围的事物有着不同的形态；（3）有时候是突出的，独特的。

接下来，这里具体介绍一种求解图像目标性的方案。Alexe很早就给出了图像目标性的定义及求解方式，在文献“what is an object”【】中给出了一种思路，在本文的目标性实现过程中，我们参考了Alexe的解决思路。

图像的目标性检测主要由五个步骤组成，具体实现的框架图如下所示：



### 3.1.1 图像颜色三通道分离

现实生活中，不管手机、相机拍摄，或者摄影机录制，得到的图像都是一般都是彩色图像，人眼也如此，物体反射的光通过视网膜进入到大脑中也是彩色的。所以本目标性检测过程中所针对的也是彩色图像。这时，处理方式就有两种，一是直接对彩色图像进行处理，第二种便是先将图像RGB三通道分离，然后对不同的通道分别处理。由于我们是想得到图像中的目标性，所以将颜色通道分开，可以获得更多的轮廓或其它信息。所以，这里的第一步便是图像颜色三通道分离。

该步骤比较简单，就是直接将图像按照RGB三通道分成了三个块，每个块的尺寸和原图像大小一致。

### 3.1.2 多尺度化

现实生活中的目标形态各不一样，大小更是千变万化，当一个物体被相机或者摄影机录制下来的时候，由于拍摄距离不定，拍摄角度的变化，目标在图像中的尺寸大小也就不一样。这种问题在图像处理过程中，是一个经常需要被重视及急待解决的问题。一般的解决方案是在图像处理过程中，对原始图像进行金字塔操作，或者对图像进行多尺度放缩，以解决图像目标多尺度化的问题。

对于输入图像，首先我们规定了一系列的放缩比例，,对于原始图像，先将图像按照放缩比例进行缩小，这样就把原始图像扩展到6个尺寸了。可以在一定程度上解决现实生活中图像中目标多尺度的问题。

### 3.1.3 图像分割

得到了不同尺度的图像后，我们需要对多尺度图像进行分割，得到不同的图像块。所谓图像分割指的是根据灰度、颜色、纹理和形状等特征把图像划分成若干互不交迭的区域，并使这些特征在同一区域内呈现出相似性，而在不同区域间呈现出明显的差异性。现有很多技术已经能够很好的对一张图像进行分割。

在分割原理上主要有基于阈值的分割方法，阈值法的基本思想是基于图像的灰度特征来计算一个或多个灰度阈值，并将图像中每个像素的灰度值与阈值相比较，最后将像素根据比较结果分到合适的类别中。因此，该类方法最为关键的一步就是按照某个准则函数来求解最佳灰度阈值；基于边缘的分割方法，所谓边缘是指图像中两个不同区域的边界线上连续的像素点的集合，是图像局部特征不连续性的反映，体现了灰度、颜色、纹理等图像特性的突变；此类方法是将图像按照相似性准则分成不同的区域，主要包括种子区域生长法、区域分裂合并法和分水岭法等几种类型；基于能量泛函的分割方法，该类方法主要指的是活动轮廓模型以及在其基础上发展出来的算法，其基本思想是使用连续曲线来表达目标边缘，并定义一个能量泛函使得其自变量包括边缘曲线，因此分割过程就转变为求解能量泛函的最小值的过程，一般可通过求解函数对应的欧拉方程来实现，能量达到最小时的曲线位置就是目标的轮廓所在。

但是在本文中，对这些多尺度的图像块进行分割只是中间一个过程，所以和以往的分割要求就有些不一样。本文中的分割块应该是包含一个目标，或者包含半个目标，或者是包含多个目标的块，在后面的步骤中，我们还需要进一步的对该块进行评比打分。对图像进行分割有很多现成的比较好的方法，如J.R.R. Uijlings2012年在IJCV上发表的了一篇文章“Selective Search for Object Recoginition”【14】中提出了一种选择性搜素算法，其思路是前期工作利用图像分析的方法得到一些原始区域，然后使用一些合并策略将这些区域合并，得到一个层次化的区域结构，而这些结构就包含着可能需要的物体。Alexe【15】也给出了一种利用图像频谱的方法来对图像分割，具体思路是先利用X. Hou【16】中的方法对图像进行频谱分析，对图像进行二维离散傅里叶变换之后，利用谱残差的思路得到一个特征图，然后根据特征图对原始图像进行分割，得到很多可能包含目标的小窗口。本文中，我们是采用的方法【15】来对图像进行分割的。

### 3.1.4 对分割窗口进行打分

在前一小节中，我们已经对多个尺度的图像块进行分割，每一个尺度图像便得到了多个分割窗口，接下来比较关键的就是对这些图像分割块进行打分。打分的目的是来评价该窗口包含目标的可能性，意图是如果一个窗口很好的包含一个目标，那么该窗口的分数会比较高；但是如果一个窗口中没有目标，那么窗口的分数会比较低。打分的依据便是前面提出的一个目标往往是具有一些特定的性质，如颜色和周围区域不一样，或者目标往往是具有一定闭环的区域等等。本文中主要从三个角度上来对分割后的窗口进行打分，分别是颜色对比，边缘密度，超像素跨越。接下来对这三点给出介绍分析：

（1）颜色对比

颜色对比（Color Contrast, CC）是说图像中的一个目标往往和周围背景有着不同的外观及颜色分布。根据这一点，如果一个窗口完整包含一个目标，则颜色特征会给出比较高的分数，如果一个串口只包含一半的目标，则分数会有所降低，但一个窗口只包含背景的话，则该窗口的分数会很低。如图所示：在（a）中，若窗口恰好把羊包围住，则该窗口的得分会比较高，当窗口不仅包含了羊，还包含了很多背景草坪时，窗口的分数就会降低。

（a） （b）

在计算窗口w的CC时，我们需要先得到包含该窗口w的一个矩形外环，然后用外环块减去中心块，便得到该窗口w对应的背景区域，这里叫做。如图（a）所示，中心蓝色框为我们将要计算的目标窗口w，黄色框为窗口w对应的矩形外环，则w的背景区域便为黄色框以内，蓝色框以外的区域。具体操作是先将窗口w沿着四个方向各自延伸的尺度，得到对应的矩形外环，这时候会有。计算窗口w和其背景区域的CC特征时是按照颜色直方图的卡方距离来计算的，具体公式如下：

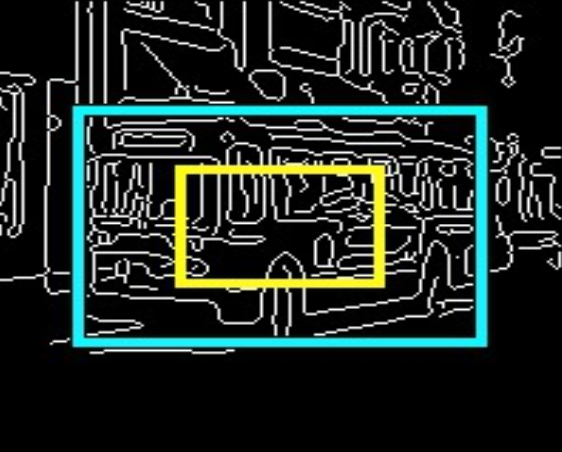


其中，表示延伸尺度，表示卡方距离，表示窗口的LAB（颜色模型）直方图。

CC特征的思想和文献【17】中提到的中心-外围直方图很相似，只是文献【17】中是计算一个像素点的中心区域和外围区域的直方图距离，而本文中的CC是为了得到一个窗口w是否包含目标，是计算本窗口和外围区域的直方图距离。

（2）边缘密度

边缘密度（Edge Density, ED）是指测量窗口w边界的边缘线密度。往往一个目标中边界边缘的数量和目标的边界周长是成正比例关系的，也就是说，随着目标边界周长的增加，其内部边缘的数量也会随着增加。

在计算ED时，我们需要先得到该窗口w的一个矩形内环，这里叫做，具体操作是将窗口w沿着四条边向内收缩的长度，这时候会有，则ED计算公式如下：



其中，表示矩形内环的周长，是一个二值化的值，是统计矩形内环区域内边缘线的个数，其边线先检测时用Canny算法实现的。

1. 超像素跨越

超像素跨越（Superpixels Straddling, SS）是得到目标边界特征的另外一种方法【18】。利用超像素可以根据图像的颜色信息或者纹理信息将图像分割成均匀的小区域，超像素的一个关键特点是可以保留目标的边界，即超像素中的所有像素点都属于同一个目标【19】，所以图像中的一个目标通常可以被分解成几个超像素。本文中根据超像素的这个特点来估计一个窗口是否包含一个目标。如果一个超像素中至少有一个像素点在窗口内，并且至少有一个像素点在窗口外，则说该超像素跨越了该窗口。同时一般一个窗口包含一个目标，则该窗口内大部分的面积会被一个超像素所占领。反之，如果一个窗口里面不包含目标，或者包含多个目标的一部分，则该窗口内部会有多个超像素点，同时这些超像素点会跨越该窗口。

在计算SS时，实际上是在计算窗口w内超像素的跨越程度，其计算公式如下：



其中超像素的集合，是采用文献【18】中的方法计算得到的，是计算时用到的分割参数；是计算该超像素点在窗口w外面的面积，是计算该超像素点在窗口w内部的面积；和中的最小值与窗口w的面积比值来衡量该窗口中超像素跨域的程度。所以越大，说明该窗口w内的超像素越完整，窗口内包含完整目标的可能性越大，反之，说明该窗口w内的超像素跨越情况较多，窗口内包含部分目标或者不包含目标的可能性越大。

### 3.1.5 计算每个像素点的目标性值

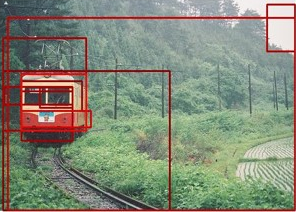
在上一小节中，我们对每一尺度图像中每个窗口从三个角度进行打分，分别是颜色对比，边缘密度，超像素跨越。有了各个窗口的分数之后，就可以计算原始图像中每个像素点的目标性值。该步骤的实现思路比较简单，首先我们需要遍历图像中的每个像素点，对每个像素点按照相同的方式得到目标性值。对于图像中任意一个像素点x，实现公式如下：

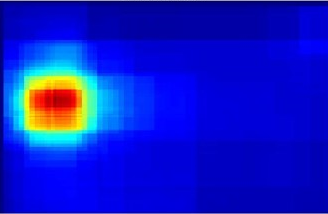
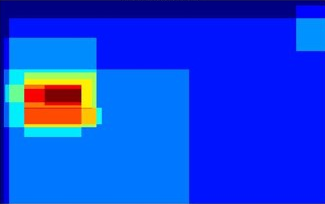
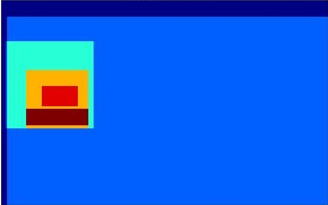
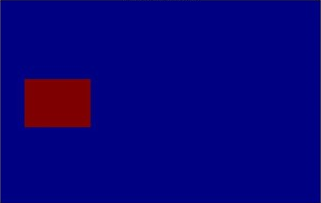


其中是包含像素点x的所有窗口；是通过颜色对比，边缘密度，超像素跨越三个特征来对窗口w给出的分数；是归一化因子。因为一张图像对应的目标性图，其目标性的范围在[0,1]分布区间，值越靠近于1，说明该点的目标性越大，值越靠近于0，说明该点的目标性越小，所以这里会将求得的目标性归一化到[0,1]区间。

通过上面介绍的5个步骤，我们就可以得到输入彩色图像的目标性图。在实验过程中，我们是可以改变算法过程中窗口划分的的总个数，当窗口个数较小时，得到的目标性效果偏差，但算法的计算时间较短；当窗口个数较大时，得到的目标性效果会变好，但算法的计算复杂度会提高，计算时间比较长。这里给出图像目标性的检测效果。







=1 =5 =10 =1000

如上图所示，第一行为原图，第二行对不同下由图像分割得到的分割结果，第三行为分割图对应得到的目标性图。可以发现随着的增加，得到的目标性效果图会越来越好，当=1时，通过分割算法我们只得到了一个窗口，所以整张图像的目标性图中，所有像素点的目标性值只有两个，随着的增加，像素点的目标性分布会越来越大，当=1000时，整个图像的目标性图已经能够很好的反应原始图像中的目标了。在本文后来的实验中所用到的图像的目标性都是在=1000的情况下得到的。

## 3.2 获取图像的显著性估计

在第二章相关背景知识介绍中，我们已经很清楚了介绍了图像显著性的原理和意义，以及常见的显著性检测方法。本节中会详细介绍本文中所用到的显著性检测方法。在介绍自动编码器的时候我们提到自动编码器主要有特征提取，降维及升维的作用，同时，自动编码器还有一个特点就是网络的输出和输入相同，利用这一点可以实现对输入信号的重构，不管输入信号是图像信息，还是语音信号，只要我们设计好对应的网络结构并用足够的样本训练，就能达到我们想要的重构结果。同时在第二章也提到，自动编码器之所以能够利用重构特点来实现对输入图像的显著性分析，主要是因为一般一张图像中背景部分，也就是非显著部分，每个像素点所对应的外围块和该像素点的中心块比较相似，所以通过自动编码器重构出来的中心块和原始的中心块之间的残差会比较小；反之，对于图像中的前景部分，也就是显著部分，每个像素点所对应的外围块和该像素点对应的中心块之间一般会有比较大的差异，这时候利用自动编码器重构出来的中心块和原始像素点的中心块之间的残差就会比较大。这也是利用自动编码器来实现显著性分析的主要原因。

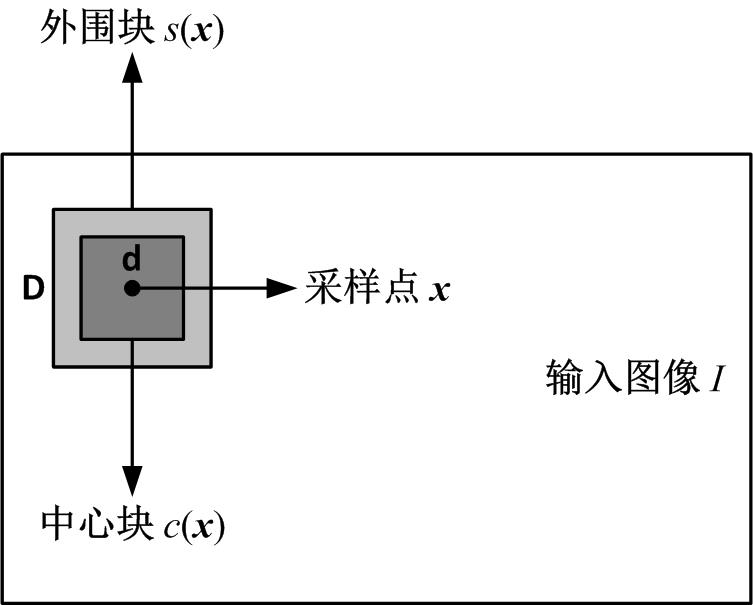
Chen Xia在文献【4】中给出了一种利用变形的自动编码器实现显著性检测的方法，本文中显著性检测部分也是基于该篇文章实现的。其基本思想是先在一张图像中随机采样多个像素点，然后获得每个像素点的外围块作为输入，得到该像素点的中心块作为标签，以此来构建训练样本；接着利用这些训练样本来训练搭建的自动编码器；当网络结构训练结束后，再遍历该图像中的所有像素点，计算由像素点外围块重构的中心块与该像素点实际的中心块之间的残差，来得到该图像的显著性图。基本实现框架图如下所示：



### 3.2.1 全局采样得到训练样本

深度学习总的可以分为三类，分别是监督学习，非监督学习及半监督学习。所谓监督学习是指在训练网络时即需要样本，还需要知道对应样本所属的类别，也可以说成标签；而非监督学习在训练网络时是不需要知道训练样本的标签的；半监督学习也可以称为强化学习，可以在训练过程中来增强标签。自动编码器则是非监督学习中的一种，因为在训练图像显著性网络结构时，我们是不需要额外提供对应图像的显著性图的。但不管哪种类型的网络，训练样本都是必须的，而且还需要足够的样本。

所以这里的第一步便是得到足够多的训练样本。对于输入图像I，先将图像I从RGB颜色空间装换到LAB颜色空间，LAB是图像颜色的另一种表示方式，其中L表示图像的亮度，A表示图像中从洋红色到绿色的范围，B表示图像中从黄色到蓝色的范围。转换空间后就可以开始对图像采样获取训练样本了。对于图像I，我们先从中随机的采样个像素点。这里采用采样随机的方式，可以保证训练样本是在图像I中均匀分布的，这样训练好的网络结构对整个图像的各个部分都能重构的比较好。对于中的每个像素点，先得到以像素点为中心，边长为D的外围块，然后再得到以像素点为中心，边长为d的中心块，其中D>d,具体采样方式如图所示：



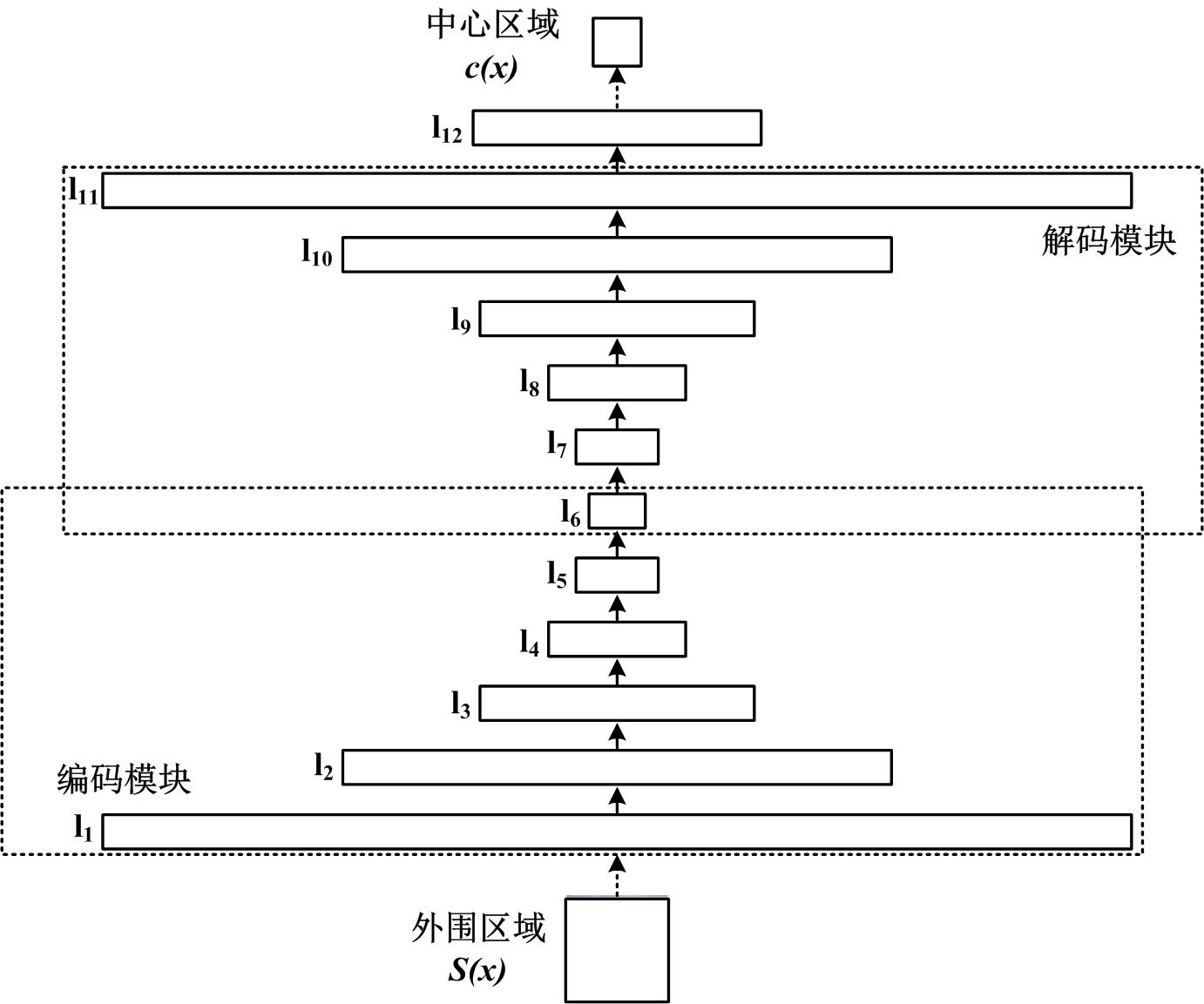
在具体的实验过程中，像素点x的外围块边长D的取值为15；，像素点x的中心块边长d的取值为7，图像I中总的采样个数的取值为8000；

对所有的样本采样之后，还需要对这些样本进行归一化处理。归一化处理是深度学习中对样本预处理时采用的最常见的方法之一。

### 3.2.2 构建并训练自动编码器网络

一般的自动编码器是由对称的RBM叠加组成，但是在本文中，我们需要在自动编码器的网络结构上利用像素点x的外围块重构该像素点对应的中心块，这就意味着这里的自动编码器输入维度和输出维度不同，和传统的自动编码有一定的区别和变形。为了解决这个问题，就需要在传统意义上的自动编码器后面添加一个输出层，这样导致该自动编码器的训练过程也有稍微的不同。

1. 构建自动编码器



如图所示，整个自动编码器是由12层网络构成的，其中前11层是一个典型的自动编码器结构，其中到为自动编码器的编码模块，是由5个RBM叠加构成；到为自动编码器的解码模块，由编码模块的5个RBM对称叠加构成，其中编码模块与解码模块关于中心层成对称结构。最后一层是为了让整个网络能够重构出像素点的中心块而在自动编码器的后面自主添加的一层。该网络的输入是像素点x的外围块，输出是像素点x的中心块；每一层网络的神经元个数依次为，由于输入的外围块是边长为D的正方形彩色图像块，其中D为15，图像块为3通道，所以第一层的神经元个数为675；最后一层为重构的中心块，其边长d为7，图像块同样为3通道，所有最后一层的神经元个数为147；第二层到第六层的神经元个数依次为256,128，64,32,8；由于第六层到第十一层与前面层相对称，所以神经元个数依次为8,32,64,128,256。

1. 网络参数初始化

在第二章中我们已经介绍了自动编码器网络参数的初始化过程，即利用RBM来对自动编码器的每一层进行初始化。在本自动编码器中，前十一层网络以第六层成对称结构，所以只需要利用RBM结构来得到前六层的初始化参数，其后面对称的六层只需要将其对称层的参数直接转置就可以得到。这里网络的前6层是由5个RBM叠加得到的，采用CD算法便可以得到前6层的初始化参数，这样，后面对称层的初始化参数也就随之得到了。对于最后一层，由于是为了使得网络的输出维度和像素点的中心层一致才加上去的，所以在初始化的时候也比较特殊，这里无法使用RBM来初始化，采用的是一般深度学习中的参数初始化方法，利用均值为0，方差为1的高斯分布来随机初始的。

1. 微调网络参数

该网络经过步骤（2）的初始化之后，已经有了一些拟合能力，即已经有了重构的一些可能，但是与我们期望中的重构能力还是有差距的。所以接下来需要对整个网络的参数进行微调，所谓微调，就是在参数初始化的基础上进行轻微调动，其目的一是为了让整个网络的参数联合起来，更加适应训练样本；二是增强网络的重构能力。整个网络的微调过程是基于梯度下降算法实现的。

具体操作是，首先将训练样本分成多批，然后每一次拿一批样本来训练网络，将像素点的外围块输入到网络，然后依次沿着网络前向传播，直到最后一层输出为，其中为该网络根据外围块重构得到的中心块，其维度和原本的中心块一致，但是由于该网络还处于训练阶段，所以和必然存在很大的交叉熵误差。网络微调的过程本质上就是减少这个误差的过程，其中误差计算方法如下



其中i表示最后一层网络中的第i个节点；为log函数。

接下来需要采用梯度下降算法，将交叉熵误差沿着网络依次反向传播，同时计算误差在每一层参数上的梯度，这里一般是链式求导过程。每一层参数有了对应的梯度误差之后，就可以根据梯度误差来更新了。反复前面的前向传播、误差计算、反向传播、梯度计算及参数更新，直到整个网络最后的交叉熵误差很小时，说明该网络的参数已经训练好，网络已具备很好的重构能力。这时候整个自动编码器网络训练结束。

### 3.2.3 对图像进行显著新估计

当自动编码器网络的参数学好之后，该网络对图像中的背景部分已经有了很好的重构能力，但对于前景模块，重构误差会比教大。这是因为整个网络在训练过程中，训练样本是随机采样的，而对于一整张图像而言，图中大部分的区域都是非显著区域，只有很少部分才是显著区域，所以整个网络的参数基本是基于非显著区域的样本学到的，因此对图像非显著区域就有很好的重构能力，重构误差小；对于图像中显著区域的重构能力差，重构误差大。

（1）估计图像的重构误差

在对图像I进行显著性估计的时候，需要遍历图像中的每一个像素点，对于每一个像素点x，依旧如训练网络样本采样一样的方式，分别得到该像素点x对应的外围块及该像素点x对应的中心块；然后将外围块输入到训练好的自动编码器中，沿着网络方向前向传播直到最后一层，最后得到由该网络重构的中心块；这时候就可以计算重构中心块与原始中心块之间的残差，其残差计算方式如下：



其中，为当前像素点对应的重构误差，为向量的范数。

（2）计算图像中心先验

通常步骤（1）中的重构误差可以直接作为图像的显著性估计，但是为了使得显著性估计能够更加符合人类的视觉注意力机制，这里又计算了图像的中心先验。所谓中心先验是指与原图像同样大小的概率图模型，其作用是越靠近图像的中心区域的像素点，该点处对应的中心先验越大，越靠近图像边缘的像素点，该点出对应的中心先验越小。中心先验其作用机制是，一般对于一张图像，人们往往会潜意识的先关注图像的中心部分，然后才会向周围延伸。所以这里在求图像显著性的时候人为加入了中心先验。一般一张图像的中心先验可以通过一个高斯模板来实现，计算公式如下：



其中，为图像中心的像素点，为图像I中任意像素点到中心像素点距离的平方，为设定好的方差，为指数函数。

同时中心先验也可以采用距离优先函数来实现，其计算公式如下：



其中，为图像中心的像素点，为图像I中任意像素点到中心像素点的距离；本算法中采用的该方法来得到的图像中心先验。

（3）由重构残差和中心先验计算图像显著性

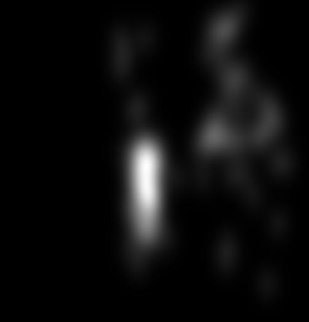
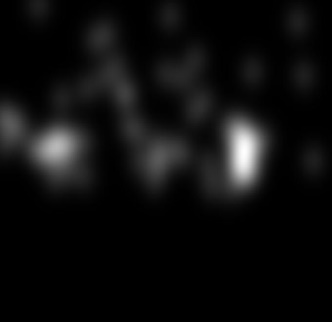
得到了图像的重构残差和中心先验之后，其显著性值就很好计算了，这里是直接将两个模板对应的值相乘就可以了，其计算公式如下：

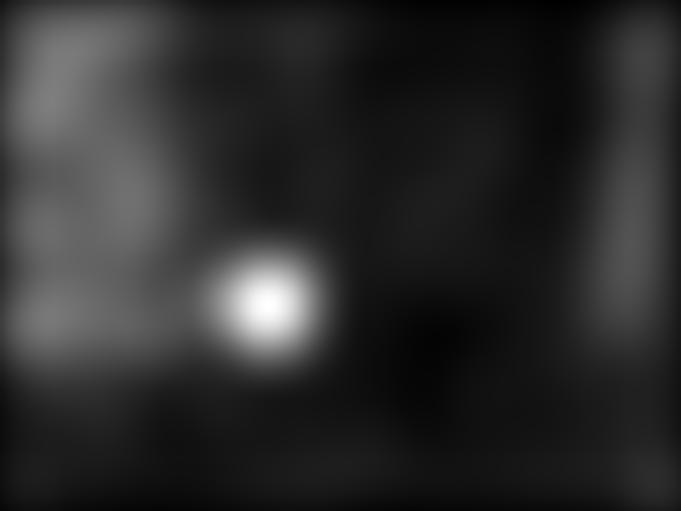


其中表示图像I的显著性图，为一张和输入图像I同样大小的灰度图，取值范围为[0,1]，其值越靠近1的话，表明该点处的显著性值越大，越靠近0的话，表明该点处的显著性越小。

由该方法得到的显著图如下所示：

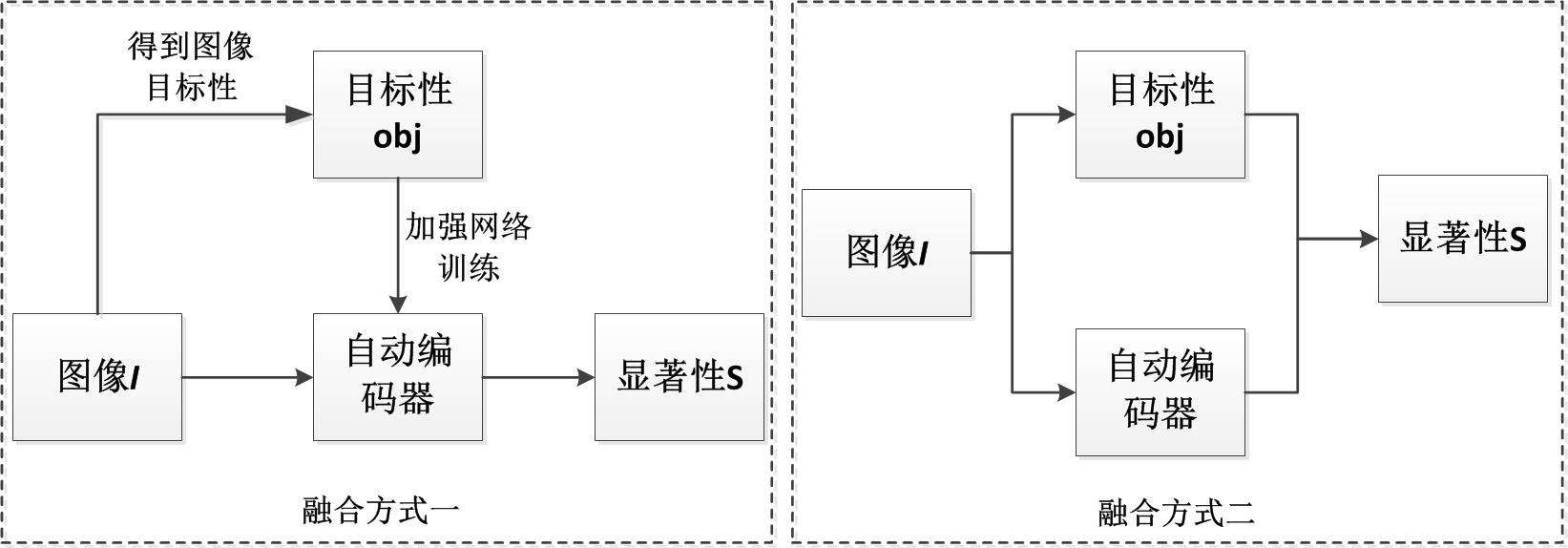
其中第一行为原图，第二行为原图对应的真值图，第三行为由自动编码得到的显著性图。

## 3.3 目标性与显著性估计的融合

在前面两小节中，我们分别单独给出了图像目标性及图像显著性估计的检测方法，其中显著性估计时采用的是利用深度学习中的自动编码器实现的【4】，是一种无监督的检测方法，该方法在现有非监督检测方法中的效果已经不错，比普通的非深度学习的方法结果已经要好。同时前面在显著性检测方法的介绍中提到，一般显著性检测可以分为两大类：一类是自下而上的以数据为驱动的显著性检测方法，该方法主要是结合图像底层特征，如方向梯度、局部颜色对比、颜色强度对比，物体轮廓及超像素来综合得到图像的显著性；另一类是自上而下的以任务目标为驱动的显著性检测方法，该方法主要是结合人类的视觉认知功能，如首先给出图片中含有人、汽车、动物等高层语义信息，再结合高层语义信息来得到图像的显著性。而现有显著性检测的方法中，一般是采用自下而上的以数据为驱动的检测思想，很少有用到结合语义信息的自上而下的显著性检测方法。而在本文中，最终目标便是探讨如何利用上层语义信息来提高图像的显著性检测方法，即采用自上而下的显著性检测思想。

这里所用到的上层语义信息便是指的图像的目标性，因为目标性本意上是指一张图片中，一个区域或者像素点被一个目标包含的可能性大小。而图像显著性是指一个图片中，最能引起人们视觉注意力机制的目标，那么既然显著性大的地方必然是一个物体或者目标，那么以图像的目标性概念为依托来求图像的显著性，也就相当于利用了图像中上层的语义信息。前面我们也得到了整个图像中每个像素点的目标性，接下来就是看怎么将目标性这种概念融合到图像显著性检测中去。

本文中主要采用了两种融合方式：第一种是利用图像的目标性来增强自动编码器的学习过程，即在训练自动编码器时，将图像目标性加入进去，来引导自动编码器参数的学习；第二种方式是将图像的目标性和利用自动编码器得到的显著性相互融合，以得到更好的显著性。简单的框架图如下所示：



### 3.3.1 目标性增强自动编码器的学习

前面在介绍自动编码器学习的过程中，我们首先需要在输入图像中随机采样8000个点，对每个像素点x，都需要得到它对应的外围块和中心块来构成训练样本。由这些点训练好的自动编码器对图像中非显著区域的像素点重构残差很小，而对于图像中显著区域的像素点重构残差比较大。之所以会这样，是因为虽然这8000个样本是在图像中随机采样的，也就是各个点的采样概率都一样，都可能成为训练样本，但是由于图像中非显著部分的区域远远大于显著区域部分，而非显著性区域部分中，像素点的外围块和中心块是很相似的，所以由这些样本学习到的网络只会对非显著部分重构的很好。其实在样本采样的过程中，对于显著区域的像素点采样得到的样本对于自动编码器的学习是不利的。因为这些像素点外围块和中心块本来差异就比较大，当训练网络时强制自动编码器通过外围块重构的中心块和原始中心块残差较小，反而会让自动编码器朝一个不好的方向学习。幸好所有采样样本中，处于显著区域的像素点比较少。

前面在训练网络时之所以会选择随机采样，是因为采样时我们不知道图像中哪些区域属于显著性区域，所以只能选择随机采样。但当我们得到了输入图像对应的目标性图，也就是清楚了图像中哪一部分属于目标性的可能性比较大，换而言之，就是知道了图像的上层语义信息。那么就可以利用该语义信息来指导我们的采样过程，也就可以间接提高自动编码器网络的训练过程。

具体实现方式是，首先对于输入图像I，我们先利用前面介绍的方法得到该图像对应的目标性，其中目标性上每个像素对应的值均在[0,1]之间。接下我们仍然从图像I中随机采样8000个点，但是对于每一个点，我们会随机生成一个[0,1]之间的随机数，如果大于，则保留这个采样点；反之，如果小于，则舍弃这个采样点。这样则使得得到的8000个样本基本是从图像I中非目标区域得到的。然后利用该样本来按照前面的方式来训练自动编码器。

### 3.3.2 目标性和显著性相互融合

第一种方式是在自动编码器训练之前就加入了图像的高层信息，即图像的目标性。第二种方式则是将通过自动编码器得到的残差图和该图像的目标性图直接融合来得到图像的显著性图。相比于第一种利用目标性来指导自动编码器的训练过程，这种方法将高层语义信息用的更为直接。因为在这种融合方式下，图像的目标性直接参与了图像的显著性评估。从原理上来讲，一张图像中越显著的地方，代表该处是目标的可能性越大，因为显著性是根据人类的视觉注意机制得到的，而人眼一定时会关注图像比较感兴趣的目标，而不会首先关注非目标区域。这就说明如果我们知道了该图像中的所有目标信息，目标分布或者目标位置，该处的视觉显著性也越可能比较明显。所以在第二种融合方式中，我们直接对图像的目标性值和图像的残差图进行比较融合。

具体实现方式是对于输入图像I，先求得该图像的目标性图，同时得到该图像经过自动编码器之后的残差图；然后在这个两个特征图上做显著性的探讨，具体探讨过程我们会在第四章实验部分给出详细过程及结果。

## 3.3 本章小结

该章是本论文中算法的原理部分，着重讲解了图像的目标性及图像的显著性获得过程，然后又引出了图像目标性及显著性的融合方式。获得图像的目标性的时候，我们将图像三通道分离，然后对每个通道分别多尺度化以适应图像目中目标的多尺度问题，接着再对每个尺度进行图像分割并对分割后的图像块窗口根据颜色、轮廓、超像素等底层特征来打分，最后根据分割后的窗口及窗口的分数来求得整个图像的目标性图；获得图像的显著性的时候，我们先在原始图像中进行随机采样以得到足够多的训练样本，接着由多层RBM构建一个自动编码器，并在自动编码器后面添加了一层重构层以适应网络输入和输出维度不同的问题，然后利用样本来反复训练以优化该自动编码器网络参数，最后利用训练好的网络得到输入图像对应的显著性图；本章节的最后简单的引出了如何利用图像的高层语义信息，即图像的目标性来提高图像的显著性估计，并给出了图像目标性及显著性的两种融合方式，具体实验过程将会在下一章中详细讲解。

# 第四章：基于目标性的显著性实验与分析

第三章中已经详细给出了本论文中算法的原理部分，在本章中，我们会着重分析在该算法下的实验探讨过程及实验结果。在此之前，我们会先引出本文中实验所用到的数据库及显著性检测的评价指标。

## 4.1 实验数据库介绍

由于显著性这个课题已经研究了比较长的时间，所以现在已有很多的相关数据库在网络上公开，以便大家在研究这个课题的时候有数据支撑，同时可以在同一个数据集上同其他人的算法做分析比较。一般公认的显著性图像数据库会包含各种场景下的图像以检验算法的鲁棒性及通用性，其显著性真值的计算一般是利用眼动仪来追踪观察者的眼球移动轨迹来得到的。具体过程是，先找到多个实验者，因为每个人的视觉注意机制都会不一样，对图像中的关注点也就不一样，所以这里需要多个实验者，然后对多个实验者的数据求平均以得到图像的显著性真值；对于每个实验者，让其距离图像一定的距离来观察图像一定的时间，同时观察不同的图像时还需要保持一定的间隔。这一步中影响显著性真值的因素主要有三个，分别是观察者距离图像的距离，每张图像的观察时间及两张图像切换时的时间间隔；在实验者观察图像的过程中，会用眼动仪来实时跟踪实验者的眼球在图像中的移动轨迹及停留时间，有了这些数据之后，我们就可以根据停留时间及轨迹图得到观察者在图像中的扫视点，最后将该扫视点进行高斯模糊之后便得到了图像的显著性真值图。

本文中的所有实验都是基于两个显著性检测图像数据库来进行的，这里分别记做DS1【20】，DS2【21】，数据库的部分信息如下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据库 | 数量 | 观察者 | 图像大小 | 观察距离（cm） | 观察时间/间隔（s） | 备注 |
| DS1 | 120 | 20 | 681×511 | 75 | 4/2 | 用的最多的数据库，包括室内及室外多种场景图 |
| DS2 | 1003 | 15 | 多种尺寸 | 60 | 3/1 | 最新的数据库，包含779幅景观图及228幅肖像图 |

其中，DS1是由Bruce and Tsotsos [20]提出来的，该图像库中总共有120幅图像，是显著性算法研究过程用的最多的一个库，其中的图像主要包含室内与室外多种场景。图像库中所有的图像均为长为681，宽为511的固定尺寸，图像的真值是由20个实验者实验得到的。获取真值时，观察者距离图像的距离为75厘米，每张图像观察4秒钟，图像切换时，需要停留2秒钟。DS2是由MIT大学的Judd et al【21】提出来的，该图像库中包含1003幅图像，是显著性检测领域中比较新的一个数据库。其中的图像是包含779幅各种场景图及228幅各种肖像图，图像库中所有的图像尺寸并没有统一大小，最大的维度为1024个像素点，其它维度分布在405到1024像素点之间。图像的真值是由15个观察者得到的。获取真值时，观察者距离图像的距离设定为60厘米，每张图像观察3秒钟，图像切换时，需要停留1秒钟。

## 4.2 实验评价方式介绍

每个实验过程中都对结果与真值给出评价指标，来衡量实验结果与真值之间的差异，以及每个实验算法的好坏。常用简单的评价指标有准确率，召回率，F-Score,复杂一点的评价指标有ROC，AUC等等。在本实验中所使用的评价指标主要是两种AUC，分别记做AUC\_Judd和AUC\_Borji，接下来简单介绍一下AUC的评价方式。

ROC（Receiver Operating Characteristic）中文名为受试者工作特征曲线，主要用来评价一个二分类问题。

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 预测值 | |
| 正样本 | 负样本 |
| 真实值 | 正样本 | 真正类（TP） | 假负类（FN） |
| 负样本 | 假正类（FP） | 真负类（TN） |

如表所示，对于一个二分类问题，主要会出现下面四种情况：

真正类（TP）：即一个正样本被算法预测为正样本；

假正类（FP）：即一个负样本被算法预测为正样本；

真负类（TN）：即一个负样本被算法预测为负样本；

假负类（FN）：即一个正样本被算法预测为负样本；

从上述四种情况中又可以得到两个新概念：真阳率（TPR）和假阳率（FPR），真阳率表示被算法预测为正样本占所有正样本的比例，假阳率表示算法错认为正样本的负样本占所有负样本的比例：





这里当二分类的阈值连续变动时，ROC正是一条随着阈值变动的曲线，其中横轴为假阳率，纵轴为真阳率。

AUC（Area Under Curve）是指ROC曲线下的面积。在实验过程中，我们总是期望真阳率越高越好，假阳率越低越好，但是随着二分类阈值的连续变动时，真阳率和假阳率总会同时的提高或降低，不会两个指标都变好，所以AUC便是兼顾真阳率和假阳率的一种中和方式。

## 4.3 实验过程及分析

在清楚了本论文的数据集及评价指标之后，接下来就是实验部分。该部分也是本论文的主要实践模块，主要由四个实验构成。实验一，利用第三章中所提到的由自动编码器对输入图像重构得到的残差图来进行显著性估计；实验二，利用第三章中所提到的目标性图来进行显著性估计；实验三，利用以图像目标性为先验知识来监督自动编码器的训练样本的采样，然后利用采样样本训练自动编码器，并得到图像的显著性来进行估计；实验四，先获取图像目标性及图像经过自动编码器得到残差图，然后对目标性及残差图直接融合来进行显著性估计。接下来是对实验的结果及详细分析模块。

1. 利用自动编码器对图像的重构残差来估计图像显著性

其中文献【4】中给出了一些

## 4.4 本章小结

本章是该论文的实验部分，首先介绍了实验过程所用到的数据库，并对数据库中的图像场景及显著性真值的获取情况给出了简单介绍；接下来介绍了实验结果的评价方式，有了评价方式之后，实验结果才能得以量化，同时能够与别人的方法进行比较与分析；最后是实验展示与分析模块，该部分是本章的重点，主要从四个实验来分析图像显著性检测效果，即利用自动编码得到的图像残差、图像的目标性、将目标性作为先验知识来训练自动编码器以得到的显著性及将目标性与自动编码器得到的残差线性融合得到的图像显著性。实验结果也验证了利用图像高层语义信息能够提高图像的显著性检测。

# 第五章：总结与展望

## 5.1 工作总结

本文主要采样自上而下的显著性检测方法，探讨以图像的高层语义信息作为先验知识，来分析图像的显著性检测。一般的显著性检测是基于自下而上的方法，即以图像中底层的纹理特征、颜色特征、超像素特征为基础，进行多种特征的融合及判断来得到该图像的显著性。而本文首先是求得图像的目标性，一种可以表明图像中目标所在位置的高层语义特征，又利用自动编码器对图像进行自我重构得到重构残差；然后简单的检验了一下图像的目标性及重构残差对图像显著性的贡献；再利用图像目标性这个高层语义特征来监督自动编码器的训练过程，利用训练好的自动编码器重构原始图像，根据重构残差来得到对应图像的显著性；接着又分析了利用图像的目标性与自动编码器得到的重构残差之间的线性叠加来得到显著性。通过多个实验分析，我们验证了利用图像的高层语义特征能够提高图像的显著新分析。

本文实验中的高层语义特征是指图像的目标性，所谓目标性是指图像中的一个像素点或一块区域是目标的可能性的几率。具体获取方式是先将原始图像三通道分离；然后对分离后的每个通道进行多尺度化以适应图像中目标尺度多变的情况；之后对于每个尺度的图像进行分割得到多个图像块窗口；有了窗口之后，我们需要对这些窗口进行打分，来评价该窗口是否包含一个目标或者半个目标或者不包含目标，具体评价指标主要是依赖图像的底层特征，这里分别是颜色对比、边缘密度、超像素跨越三个特征；每个窗口有了分数之后，最后需要遍历图像中的每一个像素点，利用包含该像素点的所有窗口及窗口打分来得到每个像素点的目标性值。

本文实验中的

## 5.2 工作展望

# 致谢

# 参考文献

[1] Itti, L.; Koch, C.; Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," in Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.20, no.11, pp.1254-1259, Nov 1998

[2] Scholkopf;lkopf, B.; Platt, J.; Hofmann, T., "Graph-Based Visual Saliency," in Advances in Neural Information Processing Systems 19:Proceedings of the 2006 Conference , 1, MIT Press, 2007, pp.545-552

[3] Chen Xia; Pengjin Wang; Fei Qi; Guangming Shi, "Nonlocal center-surround reconstruction-based bottom-up saliency estimation," in Image Processing (ICIP), 2013 20th IEEE International Conference on , vol., no., pp.206-210, 15-18 Sept. 2013

[4] Chen Xia; Fei Qi; Guangming Shi, "Bottom-up Visual Saliency Estimation with Deep Autoencoder-based Sparse Reconstruction," IEEE Trans. Neural Netw. Learn. Syst., accepted (December 20, 2015)

[5] Tie Liu; Jian Sun; Nan-Ning Zheng; Xiaoou Tang; Heung-Yeung Shum, "Learning to Detect A Salient Object," in Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, vol., no., pp.1-8, 17-22 June 2007

[6] Congyan Lang; Guangcan Liu; Jian Yu; Shuicheng Yan, "Saliency Detection by Multitask Sparsity Pursuit," in Image Processing, IEEE Transactions on , vol.21, no.3, pp.1327-1338, March 2012

[8] Han J, Ngan K, Li M, et al. Unsupervised extraction of visual attention objects in color

images[J]. Circuits and Systems for Video Technology, 2006, 16(1) :141　145.

[6] Nowak E, Jurie F, Triggs B. Sampling strategies for bag-of-features image classification

[C]//Proceedings of the 9th European Conference on Computer Vision, Graz: Springer Verlag,

2006:490-503.

【9】B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. PAMI, 2012

【10】特征融合与 objectness 加强的显著目标检测

[11] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. PAMI, 2012

【12】[A Model of saliency Based Visual Attention for Rapid Scene Analysis](http://blog.csdn.net/chenjiazhou12/article/details/39456589)

【13】Training products of Experts by Minimizing Contrastive Divergence

【14】Selective Search for Object Recoginition

【15】What is an object ?

【16】X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In CVPR, 2007

【17】T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum. Learning to detect

a salient object. In CVPR, 2007.

【18】P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. IJCV, 59(2):167–181, Sep 2004.

【19】B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman.

Using multiple segmentations to discover objects and their extent in

image collections. In CVPR, 2006

【20】N. D. B. Bruce and J. K. Tsotsos, “Saliency based on information

maximization,” in Proc. Annu. Conf. Neural Inf. Process. Syst., Montreal,QC, Canada, Dec. 2005, pp. 155–162

【21】T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in Proc. 12th IEEE Int. Conf. Comput. Vis., Kyoto, Japan, Sep. 2009, pp. 2106–2113