Computer Science and Application 计算机科学与应用, 2018, 8(1), 107-113
Published Online January 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.81014

Hans汉斯

Refinement-Based Approach of Saliency Detection

Zhenzhen Pang

Computer Science of Sichuan University, Sichuan Chengdu
Email: zhenzhen_pang@163.com

Received: Jan. 6th, 2018; accepted: Jan. 23rd 2018; published: Jan. 30th, 2018

Abstract

The aim of saliency detection is to find out significant regions of an image. Traditional salient object detection methods often use various prior knowledge and hand-crafted features to formulate contrast to get the saliency object, which have poor adaptability. Recently, deep learning is more and more popular in saliency detection. With a comprehensive learning set, the result will be much better than the traditional methods, especially for complex scenes. In this paper, a new deep learning model is proposed with a coarse-extracting process and fine-refining process. The coarse-extracting process contains two subnetworks. The first subnetwork's output feature map with the local context of superpixels cascade to the second subnetwork's high feature map extracted by VGG, then generating a coarse saliency prediction map. The fine-refining process composed of a series of recurrent convolution layers refining the coarse prediction map from coarse scales to fine scales, finally generating a fine saliency map.

Keywords

Saliency Object, Deep Learning, Image

显著性目标检测轮廓增强技术研究

庞珍珍

四川大学计算机系，四川 成都
Email: zhenzhen_pang@163.com

收稿日期：2018年1月6日；录用日期：2018年1月23日；发布日期：2018年1月30日

摘　要

显著性区域检测即针对一张图像，找出其中最显著的目标。传统方法大多基于先验知识以及根据人工提

文章引用: 庞珍珍. 显著性目标检测轮廓增强技术研究[J]. 计算机科学与应用, 2018, 8(1): 107-113.
DOI: 10.12677/csa.2018.81014

取的特征来计算对比度，继而得到显著性目标。这普遍存在适应性差的问题，即对于某些场景效果比较好，对于别的场景效果则差很多。近年来深度学习算法的兴起开始应用于显著性检测，优点在于只要数据集覆盖比较全面，对于各种场景，都可以得到优于传统方法的结果。本文在原有模型的基础上，结合粗提取到细精炼两个过程，提出了新的深度学习模型。粗提取过程由两个子网络组合，以原图像作为输入，第一个子网络以超像素为单位，结合局部上下文关系得到的特征与第二个子网络在VGG中提取的高层次特征串联得到原图的显著性粗略预测。细精炼过程由一系列循环卷积层组成的，从粗糙尺度到精细尺度精炼这个粗糙的预测，最终端对端输出精度高的显著性目标区域。

关键词

显著性目标，深度学习，图像

Copyright © 2018 by author and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 介绍

显著性目标检测是计算机视觉的一个基本分支，其目的是为了检测到图像中最突出的目标。显著性目标检测应用在很多领域，例如目标检测和识别[1]、图像和视频压缩[2]、基于内容的图片检索和图片浏览[3]、图片编辑和操作[4]等。

显著性目标检测早在 90 年代就有人对其进行研究了，最早的模型是被 Itti 等提出的[5]，推动了计算机视觉的发展。该模型是受早期灵长类动物的行为和神经元结构的启发后提出来的，求取图像的颜色、亮度和方向特征在不同尺度下的差值进行归一操作，合成显著性图。但是由于在构成高斯金字塔时的下采样操作和高斯模糊会滤除部分高频信息，造成最后得到的显著性图很模糊。为解决这个问题，基于 Itti 模型和图模型，Harel 等人[6]用马尔科夫链生成显著性图。Achanta 等人[7]提出了调频的方法，使用颜色特征和亮度特征来决定图像某个区域与其周围的对比度，产生和原图像大小一样的显著性图。Cheng 等人[1]提出了基于全局对比度的显著性区域检测，将图像用图像分割方法分割成多个区域，每个区域的显著性值被定义为该区域与图像其他所有区域的距离的加权叠加，而区域与区域的距离定义为区域中的量化后的所有颜色距离的加权叠加。该方法性能和识别效果在简单场景中较好，不过在大场景和复杂场景中效果差强人意。上述的方法都是传统的，适用性和精度都有待提高。随着卷积神经网络(CNNs) [8]的兴起，基于深度学习的显著性目标检测的模型被很多研究者研究。目前在很多研究中，CNNs 被证明在应用于显著性目标检测中是有效率的。由于它们的多水平和多尺度特征，CNNs 能在没有先验知识的情况下精确捕获到最显著的区域。He 等人[9]使用了一维卷积模型来学习超像素水平下的特征表示，对比基于像素水平的 CNN 模型，他们的模型减少了计算开销。但是只使用颜色平均值来表达超像素特征是不够的，图像空间结构很难使用一维卷积和池化操作就能表示完全，这导致了模糊预测。Wang 等人为更好检测出显著性区域，将局部和全局信息考虑进去[10]。模型被分别设计成局部评估和全局搜索两个网络，一个深度神经网络(DLL-L)先用来学习局部块的特征，捕获高层次客观对象来决定每个像素的显著性值。对于全局搜索，他们训练另一个深度神经网络(DNN-G)，使用大量的全局对比特征，例如几何信息、全局对比特征等来预测每个显著性区域的显著性值。选取前 K 个候选区域使用权重和来计算最终的显著性图。

目前大部分基于深度学习模型的目标检测算法是针对图像空间单位来做的，通过模型确定每一个空

间单位是否属于显著性目标区域来确定整个显著性目标。因此它们只注重了空间单位之间的对比，而忽略了图像的整体性空间关系(全局上下文信息)。这样做的坏处有两个：1) 因为其注重的是空间单位的对比，所以假设有两个比较突出的物体，就很容易把两个都标出来，而实际上 Ground truth 只有一个；2) 无法给出显著性目标的精确轮廓，特别是复杂场景下。

Liu 等人[11]提出用两个子网络来产生预测 map 图，用 VGG16 [12]提取粗糙的全局的预测，另一个网络由一系列循环卷积层结合前一个网络中相应的特征进行精度提炼。该方法得到的显著性检测效果良好，虽然使用了多层循环卷积网络，但是多次下采样还是会导致结果边缘模糊，并性能上有待提高。本文提出的新模型，输入的是原图像，这样有利于抓住原图像的整体空间信息(整体上下文)，端对端输出显著性区域图像。首先通过 VGG16 [12]和 Region-CNN 两个子网络得到一个粗略的结果，用于大概定位显著性目标的形状和位置，然后通过一系列 RCL (Recurrent CNN Layer) [13]来将粗略的显著性区域图像的轮廓精度提升，最后得到比较精确的结果。

2. 模型介绍

如图 1 所示：整个模型分为两个过程，分别为粗提取到细精炼两个过程。

粗提取过程用于得到一个粗略的显著性区域图像，分别用了两个现有模型，VGG16 和 Region-CNN。VGG 模型是发表在 ICLR 2015 上的一个图像检测模型，获得了 2015 年 imagenet 的冠军，有多种变体，本文采用的是 VGG16，VGG16 可以通过深层 CNN 提取到图像的高层次特征(high level feature)，是现在公认的效果较好的一个图像特征提取模型。VGG16 的具体结构这里不再赘述，通过 VGG16 获得一个大小 14 × 14，512 通道的 feature map，经过一层全连接得到一个 392 维的向量。Region-CNN 主要分为两个步骤，第一个是 Region proposal，对一个图像进行区域划分，本文划分的算法采用的是 Mean Shift [14]，有利于图像局部特征的提取和图像结构信息的表达，然后通过 CNN 对分割后的图像进行特征提取，最后经过一层全连接同样得到一个 392 维的向量，与 VGG16 得到的 392 维向量进行拼接，得到 784 维向量，最后将这个向量 reshape 成一张粗略的显著性目标区域图像。

后半部分通过一个基于 RCL (Recurrent CNN Layer)的模型，对粗略的显著性目标区域图像进行逐步

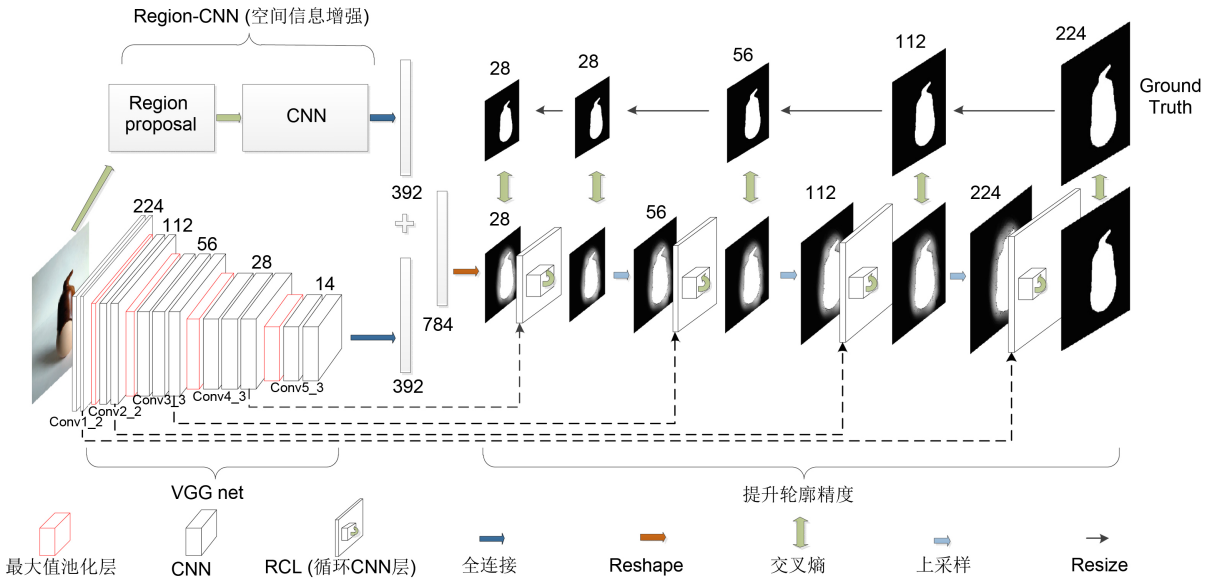

Figure 1. Model structure
图 1. 模型结构

轮廓精度提升。接下来将详细介绍这个模型以及上述提到的 Region-CNN 模型。

2.1. Region-CNN 子网络

Region-CNN 分为两个阶段。第一个是区域分割阶段，这里采用的是 Mean Shift 算法，该算法抗噪性和边缘贴合度好，生成的超像素极不规则。第二个阶段时 CNN 阶段划，对分割后的图像进行操作，具体细节如下图 2 所示：经过八层的神经网络，第一层和第二层分别是一次卷积和一次最大值池化，第三层和第四层都是卷积操作，第五层是一次卷积和一次最大值池化，第六层和第七层是一次全连接得到一个 392 维的向量，即为之前图 1 中所示的用于拼接的 392 维向量。

2.2. RCL (Recurrent CNN Layer)与轮廓精度提升

介绍轮廓精度提升的模型之前，先介绍一下 RCL (Recurrent CNN Layer)，如图 3 所示：右边黑色虚线框是左边 RCL 的展开，在每一层 RCL 中都有若干个小得循环，在本文中设置了 4 层小循环，t (1~4)，每个小得循环层间的状态转换可表示为如下的公式：

$$x_{ijk}(t) = g\left(f\left(z_{ijk}(t)\right)\right) \tag{1}$$

其中 f 是 ReLU [15]激活函数，g 是局部响应归一化函数(LRN, local response normalization) [15]，用于防止梯度爆炸，g 的表达式如下所示：

$$g\left(f_{ijk}(t)\right) = \frac{f_{ijk}(t)}{\left(1+\frac{\alpha}{N}\sum_{k'=\max(0,k-N/2)}^{\min\left(K,k+\frac{N}{2}\right)}\left(f_{ijk'}\right)^2\right)^{\beta}} \tag{2}$$

$f(z(t))$缩写成 f，其中 K 是 feature map 的个数，N 是相邻 feature map 的大小，α 和 β 是两个常量，分

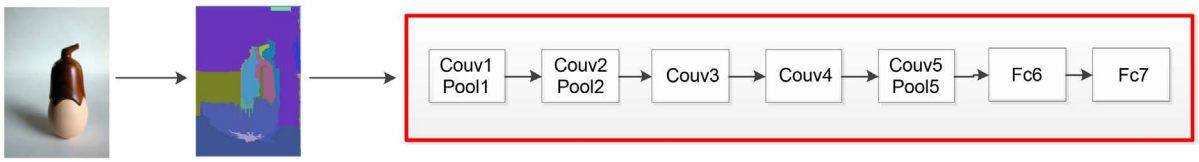

Figure 2. Convolution process of the Region-CNN
图 2. Region-CNN 的卷积阶段

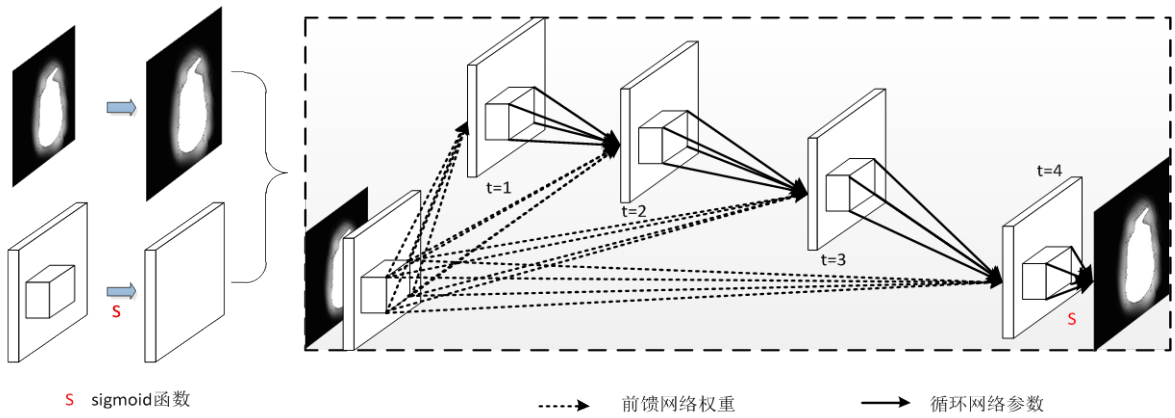

Figure 3. The structure of RCL
图 3. RCL 结构图

别设置为 0.001 和 0.75。所以我们关注点应该落在 $z(t)$上，它是一个单元的输入，而它由前馈参数和循环参数得到，其状态表示如下所示：

$$z_{ijk}(t)=\left(\boldsymbol{w}_k^f\right)^{\mathrm{T}}\boldsymbol{u}^{(i,j)}+\left(\boldsymbol{w}_k^r\right)^{\mathrm{T}}\boldsymbol{x}^{(i,j)}(t-1)+b_k \tag{3}$$

其中两个 w 分别是前馈网络权重和循环网络权重，分别表示为如图 3 所示的虚线和实线，u 表示的是当前 RCL 层做卷积的那个单元，$x(t-1)$ 表示为小循环层做卷积的单元，b 表示为偏置。所以每个小循环的 feature map 都由整个 RCL 的输入，和上一层小循环层的输入得到，能够不断获取其局部上下文信息，提高图像 feature map 的精度。

结合图 1，轮廓精度提升的模型总共包括 4 层，每一层都是独立开的，都由一个损失函数来确定其偏差，然后分别反向更新其权重。每一层都由一个 65 通道的 RCL 组成。第一次得到的结果为 14 × 14 大小的，将其与之前 VGG 的 conv4_3 层得到的 feature map 合并，得到一个 65 通道的 feature map，然后得出一个精度稍提高的结果，将其上采样至 28 × 28 维再将其与 VGG 的 conv3_3 的 feature map 合并，同样得到一个 65 通道的 feature map，以此类推，最后得到 224 × 224 高精度的显著性区域结果。这样做的原理是：因为层数的深度，每层都会产生计算偏差，但是在 VGG 模型中，随着层数的增加，其精度是由高到低的，所以只要结合 VGG 中每个阶段的 feature map，就能逆向还原出高精度的结果。其中 VGG 模型每一层都得到不止 64 维的 feature map，为了减少计算开销，将其裁剪至 64 维。

3. 训练方法

本文介绍的模型采用端到端的训练方式，以交叉熵[16]为损失函数进行训练。交叉熵为现在使用的较多的损失函数，其公式如下图所示：

$$C=-\frac{1}{n}\sum\left[y\ln a+(1-y)\ln(1-a)\right]$$

其中 y 是标签，a 是网络得出的结果，交叉熵的优势就在于，它可以克服权重更新较慢的情况，当误差较大的时候，其权重更新较快，当误差较小的时候，其权重更新较慢，这正是我们训练网络所希望出现的现象。

模型采用随机梯度下降算法(SGD)来训练，每次从样本库中抽取 batch (本文设为 32)个样本，训练网络模型，训练包括五个阶段，第一个阶段为特征提取网络权重的训练，经过第一个损失函数更新完权重后，输出直接作为下一层 RCL 的输入，到第二个损失函数时反向更新这层 RCL 的权重，直到最后一层。

4. 实验

数据集 ECSSD 包括 1000 张有意义但又复杂的图像，MSRA10K 包括 10,000 张不同物体的图像，大部分都只有一个显著性目标，且背景非常简单。DUT-OMRON 包括 5168 张一个或者多个显著性目标的图片，且其背景非常复杂。PASCAL-S 包含 850 张真实场景下的图片，其中的显著性目标与背景都非常复杂。

实验结果如图 4 所示：第一列为输入图像，最后一列为标签，倒数第二列为本文模型得出的结果，其他的为现有的模型得出的结果，从结果上可以看出，本文的模型得出的结果相比其他模型结果要好很多，比如第三幅图，由于其背景比较亮，大多数模型都把背景识别成了显著性目标，但是本文模型可以较为精确得得到正确结果。见图 5。

5. 总结

传统的显著性目标检测算法是根据人工提取的特征通过计算不同区域的对比来进行的，这样做在性

Figure 4. The experimental results
图 4. 实验结果图

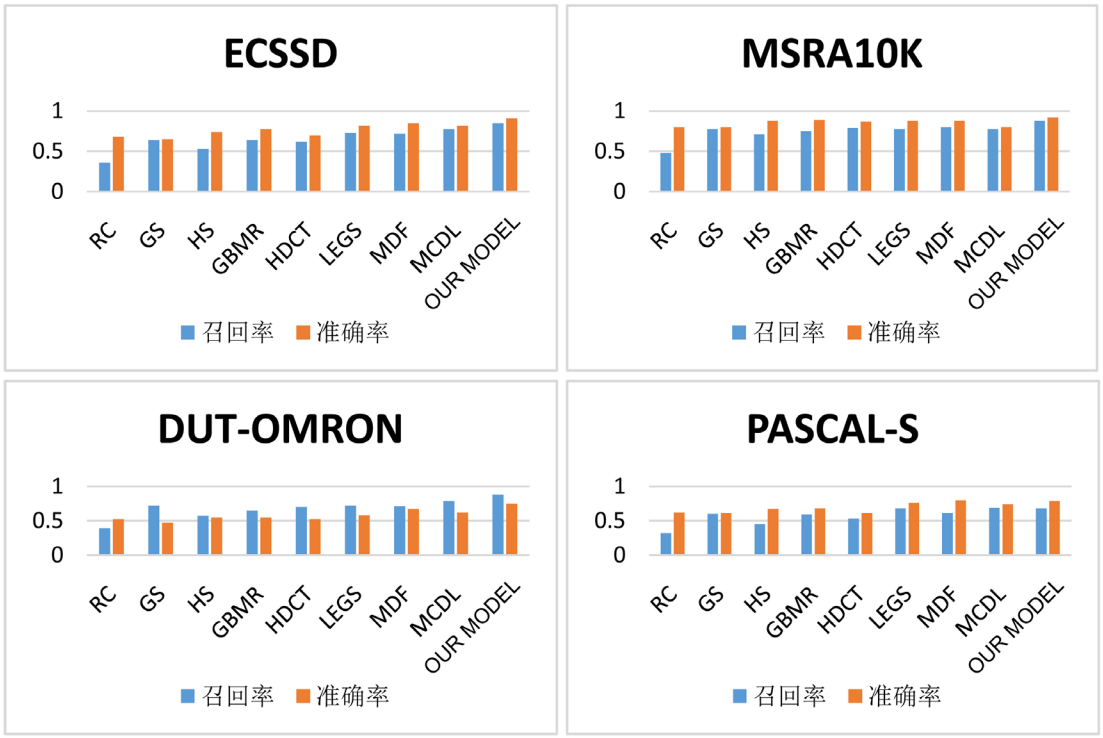

Figure 5. Our model compares the recall rate and accuracy with other algorithms in four data sets
图 5. 本文模型与其他算法在四个数据集上的召回率和准确率的对比

能和结果上都不是很理想，本文提出了一个新的端对端训练的深度学习模型，在 Region-CNN 中增加了超像素的区域信息，在不断地参数以及性能优化后，能够提取图像的全局上下文信息，并通过 RCL 的图像精度提升来得到一个较好的结果。相对于其他模型，本文提出的模型在性能，召回率和准确率上都有比较理想的提升。

参考文献 (References)

[1] Cheng, M., Mitra, N.J., Huang, X., Torr, P.H. and Hu, S. (2015) Global Contrast Based Salient Region Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 569-582. https://doi.org/10.1109/TPAMI.2014.2345401

[2] Guo, C. and Zhang, L. (2010) A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. *IEEE TIP*, **19**, 185-198.

[3] Sun, J., Xie, J., Liu, J. and Sikora, T. (2013) Image Adaptation and Dynamic Browsing Based on Two-Layer Saliency Combination. *IEEE Transactions on Broadcasting*, **59**, 602-613. https://doi.org/10.1109/TBC.2013.2272172

[4] Margolin, R., Zelnik-Manor, L. and Tal, A. (2013) Saliency for Image Manipulation. *The Visual Computer*, **29**, 1-12. https://doi.org/10.1007/s00371-012-0740-x

[5] Itti, L., Koch, C. and Niebur, E. (1998) A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 1254-1259. https://doi.org/10.1109/34.730558

[6] Harel, J., Koch, C. and Perona, P. (2006) Graph-Based Visual Saliency. *NIPS*, 545-552.

[7] Achanta, R., Hemami, S., Estrada, F. and Susstrunk, S. (2009) Frequency-Tuned Salient Region Detection. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, June 20-25 2009, Miami, 1597-1604. https://doi.org/10.1109/CVPR.2009.5206596

[8] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[9] He, S., Lau, R., Liu, W., Huang, Z. and Yang, Q. (2015) Supercnn: A Superpixelwise Convolutional Neural Network for Salient Object Detection. *International Journal of Computer Vision*, **115**, 330-344. https://doi.org/10.1007/s11263-015-0822-0

[10] Wang, L., Lu, H., Ruan, X. and Yang, M.-H. (2015) Deep Networks for Saliency Detection via Local Estimation and Global Search. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 7-12 2015, Boston. https://doi.org/10.1109/CVPR.2015.7298938

[11] Liu, N. and Han, J. (2016) Dhsnet: Deep Hierarchical Saliency Network for Salient Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, June 27-30 2016, Las Vegas, 678-686. https://doi.org/10.1109/CVPR.2016.80

[12] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Vision and Pattern Recognition*. arXiv:1409.1556

[13] Liang, M. and Hu, X. (2015) Recurrent Convolutional Neural Network for Object Recognition. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, June 7-12 2015, Boston. https://doi.org/10.1109/CVPR.2015.7298958

[14] Comaniciu, D. and Meer, P. (2002) Mean Shift: A Robust Approach toward Feature Space Analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **24**, 603-619. https://doi.org/10.1109/34.1000236

[15] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. *NIPS*, 1097-1105.

[16] Deng, L.Y. (2006) The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. *Technometrics*, **48**. https://doi.org/10.1198/tech.2006.s353

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org