DOI:10.11918/j.issn.0367-6234.201612087

空间金字塔分解的深度可视化方法

陶 攀[1,2], 付忠良[1,2], 朱 锴[1,2], 王莉莉[1,2]

(1.中国科学院 成都计算机应用研究所, 成都 610041; 2.中国科学院大学, 北京 100049)

摘 要: 针对基于深度卷积神经网络的图像分类模型的可解释性问题,通过评估模型特征空间的潜在可表示性,提出一种用于改善理解模型特征空间的可视化方法. 给定任何已训练的深度卷积网络模型,所提出的方法在依据原输入图像使得模型类别得分激活最大化时,首先对反向传播的梯度进行归一化操作,然后采用带动量的随机梯度上升训练策略,反向回传修改原输入图像. 引入了通过激活最大化获得的图像可解释性的正则化方法,常规正则化技术不能主动调整模型特征空间的潜在可表示性,结合现有正则化方法提出空间金字塔分解方法,利用构建多层拉普拉斯金字塔主动提升目标图像特征空间的低频分量,结合多层高斯金字塔调整其特征空间的高频分量得到较优可视化效果. 通过限制可视化区域,提出利用类别显著性激活图技术加以压制上下文无关信息,可进一步改善可视化效果. 对模型学习到的不同类别和卷积层中单独的神经元进行合成可视化实验,实验结果表明提出的方法在不同的深度模型和不同的可视化任务中均能取得较优的可视化效果.

关键词: 深度可视化;金字塔分解;激活最大化;卷积神经网络;激活图

中图分类号: TP391.41 **文献标志码**: A **文章编号**: 0367-6234(2017)11-0060-06

Deepvisualization based on the spatial pyramid decomposition

TAO Pan[1,2], FU Zhongliang[1,2], ZHU Kai [1,2], WANG Lili[1,2]

(1. Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Focusing on the interpretability problems of image classification models based on deep convolutional neural network, a visualization method for improving the feature space of model is proposed by evaluating the potential expressiveness of model feature space. Given any pre-trained deep model, firstly the method generates an image by the normalized operation of the gradient in the back propagation, which maximizes activation the class score, and then uses the momentum of the stochastic gradient descent training strategy for back propagation to the original input image. The conventional regularization technique cannot adjust the feature space of the model. Therefore, the spatial pyramid decomposition method is proposed on the basis of the existing regularization method. By constructing the multi-layer Laplacian spatial pyramid, the low frequency component of the target image feature space is promoted, combined with multi-layer Gaussian spatial pyramid to adjust the high-frequency components of its feature space to obtain a better visualization effect. By limiting the region of visualization, it is proposed to use the class activation map to suppress the context-free information, which can further improve the visualization effect. The visualization experiments are performed on the different classes of the model and the individual neurons of the convolution layer. Results show that the proposed method can achieve better visualization effect in different depth models and different visualization tasks.

Keywords: deep visualization; pyramid decomposition; maximize activation; convolutional neural network; activation map

以深度卷积神经网络(Convolutional Neural Network,CNN)为代表的深度学习对计算机视觉和机器学习领域产生了深远影响. 但是完全理解深度学习模型的内在工作原理,设计高性能的深度网络结构还是很困难的,一直以来人们普遍将其内部工作原理看成一个“黑箱”,这是由于深度 CNN 存在海量参数,多次迭代更新生成输入输出之间相当不连续和非线性的映射函数;以及对参数的初始状态敏感,存在很多局部最优点. 探究 CNN 的运行机制,核心在于它究竟自动提取什么样的特征,经过卷积层、池化层,特征都是分布式表达的,每个特征反映在原图上都会有重叠,故希望建立特征图与原图像之间的联系,即深度可视化. 该技术试图寻找深度模型所提取各层特征较好的定性解释,并在设计开发新网络结构方面扮演重要角色.

收稿日期: 2016-12-15
基金项目: 中国科学院西部之光人才培养计划项目
作者简介: 陶 攀(1988—),男,博士研究生
通信作者: 付忠良, Fzliang@ netease.com

目前针对 CNN 可视化的研究，主要集中在如何理解 CNN 从海量数据中自动学习到的，能反映图像本质的分层特征表达，即获得网络中隐藏层神经元与人类可解释性概念之间的联系. 最直接的方法是展示学习得到的卷积核和相应的特征图，但除了首层卷积核和特征图有直观的解释外，其余各层并没有可解释性. 从信号处理的角度看，基于 CNN 高层特征的分类器在输入域，需要较大感知野，才能对以由低频为主的输入图像进行多层非线性响应，并对小的输入改变产生平滑不变输出. 同时，由于经过非线性激活函数变换和池化，引入空间不变性获得更好识别性能的同时，也对可视化带来新的挑战.

深度可视化技术可以简单分为三类：基于梯度更新的方法[1-6]；基于特征重建的方法[7-10]；基于相关性的方法[11-12]. 基于网络梯度更新的思想是由 Erhan 等[1]引入，固定模型参数通过梯度更新改变输入值，最大化激活单一神经元或标签类别概率. 激活最大化生成的非自然图像还可以是网络模型的对抗样本[13]. Simonyan 等[2-3,14]通过梯度上升方法迭代寻找使得最大化激活 CNN 某个或某些特定的神经元的最优图像，其假设神经元对像素的梯度描述了当前像素的改变能影响分类结果的强度. 文献[2]引入 L_2 正则化先验(或称权重衰减)，改进可视化效果. Yosinski 等[4]进一步提出高斯模糊正则化、梯度剪切等技术，其中梯度剪切指的是每次只更新对分类最有利的一部分梯度，改善生成图像质量. 文献[3,6]考虑神经元的多面性和利用生成网络作为自然图像的先验来合成更自然的图像.

Zeiler 等[7]提出利用反卷积网络，利用反向传播重构各层特征到像素空间的映射，并用于指导设计调优网络结构，提高分类识别精度. 在反卷积过程中利用翻转原卷积核近似作为反卷积核，针对特定特征图在训练集上重新训练. Dosovitskiy 等[8]提出通过学习‘上’卷积网络来重建 CNN 各层的特征，指出结合强先验，即使用于分类的高层激活特征也包含颜色和轮廓信息. Mahendran 等[9-10]通过对学习到的每层特征表达进行反编码重建，提出利用全变分正则化和自然图像先验，并将 L_2 范数正则化推广到 p 范数正则化，得到较优的可视化效果.

本文主要关注前两种方法中的正则化技术，基于相关性分解方法请参考文献[12]. 受文献[15-16]启发，把用于图像生成的拉普拉斯金字塔，进一步扩展成空间金字塔分解方法，并引入显著性激活图技术进一步改进深度 CNN 的可视化效果.

1　可视化方法的数学模型

激活最大化和特征表达反编码重建均是针对已经训练好的模型，对给定输入 $x_i \in \mathrm{R}^{C\times H\times W}$，其中 C 为颜色通道数，H，W 为图像高和宽. CNN 模型可抽象为函数 $\varphi:\mathrm{R}^{C\times H\times W}\to R^d$，其第 i 个神经元的激活值为 $\varphi_i(x)$，对给定图像 x_0 的特征编码 $\varphi_0=\varphi(x_0)$，定义参数 θ 的正则化项 $R_\theta(x)$，寻找使得能量泛函最小化的初始输入 x^*，其数学模型为

$$x^*=\underset{x}{\operatorname{argmin}}(\ell(\varphi(x),\varphi_0)+\lambda R_\theta(x)). \quad (1)$$

式中 ℓ 损失比较的是 $\varphi(x)$ 和目标 φ_0 的差异，选择不同的损失函数定义不同的可视化方法. 但该优化通常是一个非凸优化问题，通常采用梯度下降法去寻找局部最优值为

$$x\leftarrow x+\alpha\frac{\partial\varphi_i(x)}{\partial x}. \quad (2)$$

激活最大化方法是文献[1]中提出针对深度架构中任意层中的任意神经元所提取的特征，寻找使一个给定的隐含层单元的响应值 $\varphi_0\in R^d$ 最大的输入模式，可由内积形式定义 ℓ 损失为

$$\ell(\varphi(x),\varphi_0)=-\langle\varphi(x),\varphi_0\rangle. \quad (3)$$

式中 φ_0 需人工指定，最大化激活的目标可以是全连接层的特征向量，也可以是卷积层某一通道的某一神经元的激活值.

特征表达的反编码重建，通过最小化给定特征向量与重建目标图像特征向量间的 ℓ 损失，一般采用欧式距离来衡量损失误差，定义如下：

$$\ell(\varphi(x),\varphi_0)=\frac{\|\varphi(x)-\varphi_0\|^2}{\|\varphi_0\|^2}. \quad (4)$$

但也可利用其它距离度量函数来评价损失.

2　梯度更新的可视化方法

用于分类的深度 CNN 提取高层语义信息的同时，丢失了大量低层结构信息. 由于首层卷积核大都类似 Gabor 滤波器，导致梯度更新可视化生成图像中包含许多高频信息，虽然能产生大的响应激活值，但对可视化来说导致生成的图像是不自然的. 还由于网络模型的线性操作(如卷积)导致对抗样本[13]的存在，为得到更类似真实自然图像的可视化结果，需在优化目标函数中引入正则化作为先验.

2.1　p 范数正则化方法

对图像来说，像素大小需在一定范围内，直接最大化激活类别概率，生成图像类似随机噪声图像. 文献[2]通常引入 L_2 范数正则化，惩罚过大和过小的极端值，其公式为 $R_\theta(x)=\|x\|_2^2$. 在文献[10]中将其扩展到彩色图像 RGB 通道空间中的 p 范数正则化为

$$R_\theta(x)=\frac{1}{HWC^p}\sum_{h=1}^{H}\sum_{w=1}^{W}\Big(\sum_{c=1}^{C}x(h,w,c)^2\Big)^{\frac{p}{2}}. \quad (5)$$

式中：h,w 表示图像的行和列大小,c 表示颜色通道数,对比发现,文献[2]提出的 L_2 正则化是忽视各颜色通道的差异的,正则化的力度可通过缩放常量 p 进行控制,即使得图像像素值大小保持在合适的范围内.

2.2 高斯模糊和 TV 变分

基于梯度更新可视化方法,引入高斯滤波器主动惩罚高频信息[4],高斯模糊核半径大小由高斯函数的标准差控制,可随迭代次数动态调整模糊核大小.

全变分[10](Total Variance,TV)跟高斯模糊类似,鼓励可视化生成分片的常量块区域,对离散图像全变分操作可由有限差分来近似求解为

$$R_{\mathrm{TV}}(x)=\frac{1}{HWB^{\beta}}\sum_{hwc}\left((x(h,w+1,c)-x(h,w,c))^2+(x(h+1,w,c)-x(h,w,c))^2\right)^{\frac{\beta}{2}}. \tag{6}$$

式中 $\beta=1$,但其在可视化过程中,在图像的平坦区域并不存在边缘,全变分操作仍沿着边缘方向扩散就会导致出现虚假的边缘,会引入所谓的“阶梯效应”现象.$\beta<1$ 时结合超拉普拉斯先验[17]能更好匹配自然图像的梯度统计分布,但对可视化来说反而使得可视化更困难. 文献[10]实际实验表明,跟高斯模糊核一样,需随迭代次数动态调整 β 大小.

2.3 基于数据统计先验

由于常规可视化方法并没有对颜色分布进行建模,文献[3]提出通过引入外部自然图像数据,计算图像色块先验为

$$R_{\theta}(x)=\sum_{\mathrm{p}}\|x_p-D_p\|_2^2. \tag{7}$$

式中：p 为块索引,x_p 表示稠密采样的归一化图像块,D_p 表示自然图像块数据库中距离 x_p 最近图像块. 该方法跟文献[15]中利用参考图像“指导”人脸图像嵌入重建类似. 并且基于数据的统计先验可进一步扩展,引入生成对抗网络,利用生成网络主动生成自然图像先验[5].

3 空间金字塔分解

正则化先验主动限制图像空间中高频率和高振幅信息,生成的可视化图像存在如下问题：1)彩色图像的颜色分布仍是不自然的. 2)生成的图像中包含可识别类别对象的多个重复成分,并且这些部件不能组合成完整的有意义整体. 3)缺乏令人可信的低频细节,存在棋盘效应,只是形似. 针对这些问题提出利用空间金字塔分解,主动提升低频信息和调控高频信息以改善生成图像的可视化效果.

3.1 高斯和拉普拉斯金字塔分解

拉普拉斯金字塔(Laplacian Pyramid,LP)[18]是由一系列包含带通滤波器在尺度可变的图像上加低频残差组成的. 首先通过高斯平滑和亚采样获得多尺度图像,即第 K 层图像通过高斯模糊、下采样就可获得 $K+1$ 层,反复迭代多次构建高斯金字塔(Gaussian Pyramid,GP). 用高斯金字塔的 K 层图像减去其第 $K+1$ 层图像上采样并高斯卷积之后的预测图像,得到一系列的差值图像即为拉普拉斯金字塔分解图像.

拉普拉斯金字塔分解过程(见图 1 所示)包括 4 个步骤：1)高斯平滑 $G_{0..n}$；2)降采样(减小尺寸)；3)上采样并高斯卷积(图中 expand 操作)；4)带通滤波(图像相减)$L_{0..n}$. 拉普拉斯金字塔突出图像中的低频分量,拉普拉斯金字塔分解的目的是将源图像分解到不同的空间频带上.

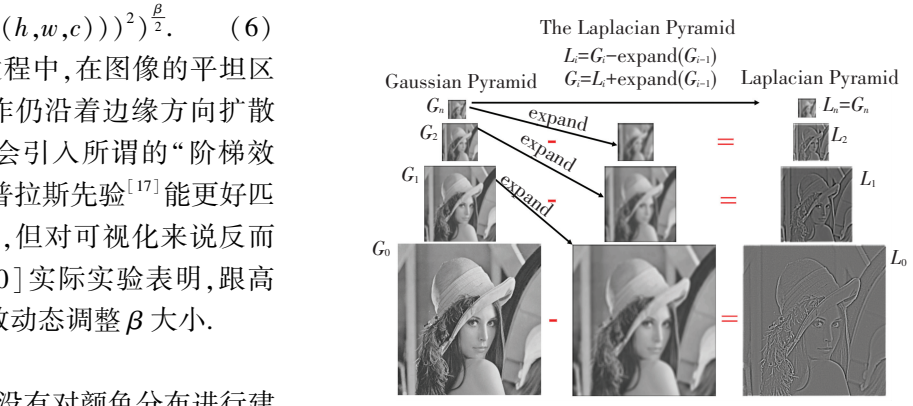


图 1 高斯和拉普拉斯金字塔

Fig.1 Gaussian and Laplacian Pyramids

由于自然图像统计特性中的尺度不变性,也称为 1/f 法则[19],即自然图像集 $I(f_x,f_y)$ 的平均傅里叶功谱服从 $I\sim(f_x^2+f_y^2)^{-1}$. 在激活最大化可视化深度 CNN 模型过程中利用提出的高斯和拉普拉斯空间金字塔分解,调整生成梯度图像包含的频谱分量大小. 其中空间金字塔分解正则化项为

$$r_{\theta}(x)=\sum_{k=1}^{K}\left[LP_k(x)+GP_k(x)\right]. \tag{8}$$

式中：k 代表构建 k 层金字塔分解,本文实验 k 选取为 4. $LP_k(x)$ 为第 k 层的拉普拉斯金字塔分量,$GP_k(x)$ 为第 k 层的高斯金字塔分量.

3.2 梯度归一化

基于梯度更新的可视化方法,由于原输入空间中高低频分量混杂在一起,对原输入图像相应的更新梯度进行归一化操作能得到较好可视化效果,即对输入图像每次迭代更新的梯度 $g=\partial\varphi_i(x)/\partial x$,则提出梯度归一化操作：

$$g\leftarrow\frac{g}{(g.\mathrm{std}+\delta)}. \tag{9}$$

式中：δ 为非负小常量,std 表示梯度矩阵的方差. 该

梯度中心归一化技术，可以减少产生重复的对象碎片的倾向，而倾向于产生一个相对完整对象. 梯度归一化的引入同批归一化（Batch Normalization）思想类似，校正 CNN 网络非线性变换引起的“偏移”，该方法也侧面验证最新提出的分层归一化[20]的有效性.

3.3　类别激活图限制可视化区域

根据文献[26]提出的类别激活图技术，假设 $f_j(x,y)$ 表示最后的卷积层空间 (x,y) 位置上第 j 个神经元的激活值，则对 j 神经元的全局平均池化操作结果对给定类别 k 的得分函数 S_k：

$$S_k = \sum_j w_j^k \sum_{x,y} f_j(x,y). \tag{10}$$

式中 w_j^c 是第 j 个神经元和第 k 类的连接权重. 根据文献[26]，由式(10)可得定义类别激活图 M_k 为

$$M_k = \sum_j w_j^k f_j(x,y). \tag{11}$$

式中 M_k 表明在空间 (x,y) 位置的激活值对分类结果影响的重要性. 对类别激活映射图直接双线性插值得到与原输入图像大小相等的显著性图. 本文利用显著性激活图作为梯度更新的权重因子，即输入变为原始输入图像与类别激活图的加权乘积. 动机是要求网络梯度更新保持在类别显著性区域内，压制无关背景信息的生成. 具体详情请参见第四章实验部分.

3.4　优化方法

深度 CNN 模型优化策略的核心是随机梯度下降法，常用方法是带动量的随机梯度下降法为：

$$V_t = \mu V_{t-1} - \alpha \nabla f(x_t), \tag{12}$$

$$x_{t+1} = x_t + V_t. \tag{13}$$

式中：μ 为动量因子表示保持原更新方向的大小，一般选取 0.9，x_t 为在 t 时刻待更新的梯度，α 为学习率；文献[9-10]采用自适应梯度（Adaptive Gradient，AdaGrad）[21]的变种算法，根据历史梯度信息自适应调整学习率. 同时文献[22]采用的二阶优化算法针对纹理和艺术风格重建问题，得到比用基于一阶随机梯度下降算法更优的可视化效果. 但本文通过实验对比发现对各种优化方法对生成图像质量影响不大，从简选择带动量的随机梯度优化方法.

4　实验结果分析和讨论

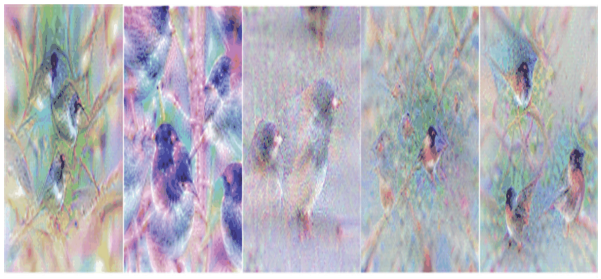
基于梯度更新的可视化方法主要用于激活最大化和特征重建，但文献[23]指出用随机未训练的 CNN 模型也能较好重建原图像，表明特征编码重建不能很好解释训练得到 CNN 模型的内在工作机理. 本文实验主要关注在对 ImageNet 公开数据集上预先训练得到的分类模型进行激活最大化可视化实验.

4.1　不同深度模型的类别可视化

实验选取的深度模型来自于开源社区的 Caffe model zoo，不同的 CNN 模型如：AlexNet 模型[24]，Vgg-19 模型[25]，Google-CAM 模型[26]，GoogleNet 模型[27]，ResNet 模型[28]，其分类识别性能依次从低到高，模型的复杂程度依次递增. 本文实验默认采用提出的梯度归一化，并引入多分辨率、随机扰动和剪切等小技巧作为通用设置，提高可视化效果.

为比较不同深度 CNN 模型学习相同类别时特征图的差异，根据式(1)，给定高斯噪声生成随机图像作为输入，指定可视化物体类别向量（见图 2 所示，类别为所有类别中的第 13 类布谷鸟），施加前文提出不同正则化项的组合：p 范数、高斯模糊和金字塔分解正则化.

图 2 结果表示 5 种 CNN 模型在相同正则化方法和相同梯度更新策略下的可视化效果，对比图 2 中(a)，(b)，(c)发现随着网络模型深度的增加，可视化难度增大分类性能同可视化效果一致；Vgg-19 模型由于跟 ResNet 模型卷积核大小类似，且比 AlexNet 首层卷积核小(7 和 3)，即可视化效果倾向生成比 AlexNet 更大尺寸的物体. 而由图 2 中 a，d，e 对比可知，由于 GoogleNet 模型中卷积层的卷积核大小不一，使得可视化结果中引入更多细节. 综合可知，基于 GoogleNet 模型的可视化效果最好，后面实验均是在其模型的基础上进行实验比较.



(a) AlexNet (b) Vgg-19 (c) ResNet (d) GoogleNet (e) Google-CAM

图 2　不同模型类别可视化实验结果

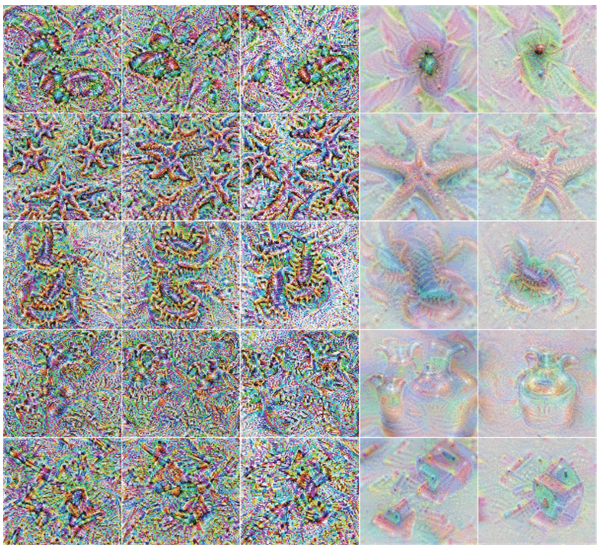
Fig.2　The visualization of different deep models

4.2　不同正则化方法的类别可视化

为验证不同正则化方法对理解深度模型的特征达的影响，采取前文所述的不同正则化方法，可视化效果结果见图 3 所示，从上到下依次可视化类别为金甲虫，海星，蝎子，酒壶，卷笔刀.

图 3 中(a)列仅施加默认设置和不加梯度归一化的结果，由于输入的随机性，并不能保证每次都生成有意义的可视化结果，但引入本文提出的梯度归一化后，能大概率生成可视化结果见图 3(b)列所示，图 3(c)列表示只采用 p 范数正则化，跟文献[2]一

致取 2,使得图像更平滑,但仍与真实图像相差较大. 通过前文理论分析和实验验证,全变分跟高斯模糊作用类似,本文采用根据迭代轮数动态调整高斯模糊核大小,具体是在刚开始采用较大值希望生成物体大概轮廓,随迭代逐渐调小模糊核使得更多细节生成,具体见图 3(d). 但是这个参数无法自适应设置为最优,对图像高低频分量无法调整控制,而本文提出的利用金字塔分解正则化方法能从粗到细调整,产生较优结果见图 3(e)列所示.



(a) original (b) 梯度归一 (c) p 范数 (d) Blur (e) Our

图 3 不同正则化方法的可视化效果

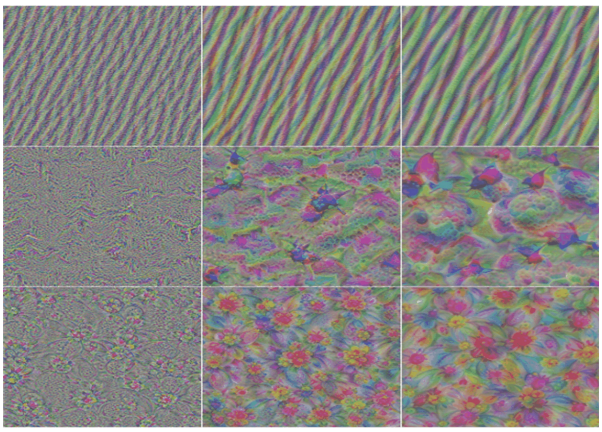
Fig.3 The visualization of different regularization

4.3 金字塔分解可视化实验结果

为验证提出金字塔分解正则化方法,对中间层卷积核的可视化,采用前文提出式(8),指定深度 CNN 模型中不同卷积层中不同通道,利用前文提出的带动量的梯度更新策略,可视化结果见图 4,其中从上到下依次为 GoogleNet 模型低中高层不同通道的可视化结果,与文献[7]一致,低层多尺度分辨率生成的纹理见图 4 首行所示,中层是一些物体部件,见图 4 中间行所示蜜蜂的局部结构,而高层是更完整的抽象概念见图 4 下层中完整的花瓣. 对比图 4(b)、(c)列,可验证拉普拉斯金字塔主动分解提升图像部分低频成分,而高斯金字塔分解生成的图像中高频细节更突出.

4.4 引入类别显著性的可视化

通过观察之前可视化结果可知,生成的图像中除了该类别外仍有许多额外的上下文信息(见图 2 中鸟类别的树枝),这些信息与模型的分类能力相关联,可通过引入类别激活图可改善可视化效果. 迭代更新过程中依据采用式(11),使用类别激活图作为加权因子限制迭代更新区域.

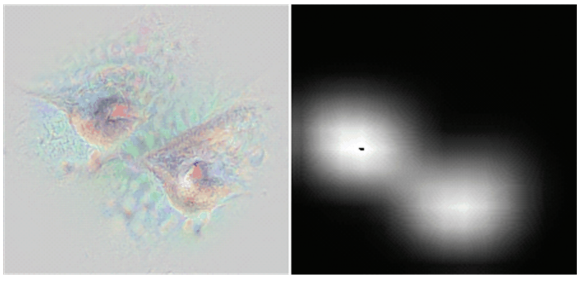


(a) 多尺度分辨率 (b) 拉普拉斯金字塔 (c) 高斯金字塔

图 4 金字塔分解正则化可视化效果

Fig.4 The visualization of pyramid decomposition

实验结果见图 5(a)所示,具体实验设置和图 2 采用的参数一致,使用提出的金字塔分解正则化技术,图 5(b)为图 5(a)相应的类别激活图,图 5(a)结果表明与类别无关的上下文信息得到抑制,但仍存在两个类别中心.



(a) 可视化结果 (b) 类别激活

图 5 引入类别激活图的可视化

Fig.5 The visualization with class activation map

5 总 结

本文针对理解深度 CNN 特征空间存在的问题,提出一种用于改善深度 CNN 分类模型的可视化方法. 其中通过改善激活最大化可视化技术来产生更具有全局结构的细节、上下文信息和更自然的颜色分布的高质量图像. 该方法首先对反向传播的梯度进行归一化操作,在常用正则化技术的基础上,提出使用空间金字塔分解图像不同频谱信息;为限制可视化区域,提出利用类别显著激活图技术,可以减少优化产生重复对象碎片的倾向,而倾向于产生单个中心对象以改进可视化效果. 激活最大化可显示 CNN 在分类时关注什么. 这种改进的深度可视化技术将增加我们对深层神经网络的理解,进一步提高创造更强大的深度学习算法的能力. 该方法适用于基于梯度更新的可视化领域,是对网络模型整体的理解,具体各层特征怎么耦合成语义信息仍需进一步探索,深度 CNN 模型如何重建一个完整的类别概

念，仍是一个开放性问题.

参考文献

[1] ERHAN D, BENGIO Y, COURVILLE A, et al. Visualizing higher-layer features of a deep network [R]. University of Montreal (1341), 2009.

[2] KAREN S, ANDREA V, ANDREW Z. Deep inside convolutional networks visualising image classification models and saliency maps [C]// International Conference on Learning Representations. San Francisco: ICLR, 2013: 1-8.

[3] LENC K, VEDALDI A. Understanding image representations by measuring their equivariance and equivalence [C]//IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: CVPR, 2015: 991-999.

[4] YOSINSKI J, CLUNE J, NGUYEN A, et al. Understanding neural networks through deep visualization[C]//Deep Learning Workshop, International Conference on Machine Learning. Lille, ICML, 2015: 1-9.

[5] NGUYEN A, DOSOVITSKIY A, YOSINSKI J, et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks[C]//Advances in Neural Information Processing Systems.Barcelona: NIPS, 2016:1-29.

[6] NGUYEN A, YOSINSKI J, CLUNE J. Multifaceted feature visualization: uncovering the different types of features learned by each neuron in deep neural networks[C]//Proceedings of the Workshop on Visualization for Deep Learning at International Conference on Machine Learning. New York: ICML, 2016: 1-23.

[7] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Computer Vision-ECCV 2014.Zurich:Springer,2014:818-833.DOI: 10.1007/978-3-319-10590-1_53.

[8] DOSOVITSKIY A, BROX T. Inverting visual representations with convolutional networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada: CVPR,2016:1063-6919.DOI:10.1109/CVPR.2016.522.

[9] MAHENDRAN A, VEDALDI A. Visualizing deep convolutional neural networks using natural pre-images[J]. International Journal of Computer Vision, 2016, 120(3): 233-255. DOI: 10.1007/s11263-016-0911-8.

[10] MAHENDRAN A, VEDALDI A. Understanding deep image representations by inverting them[C]//IEEE Conference on Computer Vision and Pattern Recognition. Boston:CVPR,2015:5188-5196. DOI:10.1109/CVPR.2015.7299155.

[11] CAO C, LIU X, YANG Y, et al. Look and think twice: capturing top-down visual attention with feedback convolutional neural networks [C]//IEEE International Conference on Computer Vision. Santiago, IEEE, 2015: 2956-2964. DOI: 10.1109/ICCV.2015.338.

[12] BACH S, BINDER A, MONTAVON G, et al. Analyzing classifiers: fisher vectors and deep neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada: CVPR, 2016: 2912-2920. DOI: 10.1109/CVPR.2016.318.

[13] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and Harnessing Adversarial Examples[C] //International Conference on Learning Representations. San Diego:ICLR, 2015: 1-11.

[14] SZEGEDY C, ZAREMBA W, SUTSKEVER I. Intriguing properties of neural networks[C]// International Conferenceon Learning Representations. Banff:ICLR,2014: 1-10.

[15] SCHROFF F,KALENICHENKO D,PHILBIN J. FaceNet: a unified embedding for face recognition and clustering[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: CVPR, 2015: 815-823. DOI: 10.1109/CVPR.2015.7298682.

[16] DENTON E, CHINTALA S, SZLAM A, et al. Deep generative image models using a laplacian pyramid of adversarial networks[C]// Advances in Neural Information Processing Systems 28. Montréal, Quebec:NIPS, 2015: 1486-1494.

[17] KRISHNAN D, FERGUS R. Fast image deconvolution using hyper-laplacian priors[C]//Advances in Neural Information Processing Systems. Vancouver, BC:NIPS, 2009: 1-9.

[18] BURT P, ADELSON E. The laplacian pyramid as a compact image code[J].IEEE Transactions on Communications, 1983, 31(4): 532-540. DOI: 10.1109/TCOM.1983.1095851.

[19] VANDER S A, VANHATEREN J H. Modelling the power spectra of natural images: statistics and information [J]. Vision Research, 1996, 36(17): 2759-2770. DOI: 10.1016/0042-6989(96)00002-8.

[20] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille: 2015: 448-456.

[21] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12: 2121-2159.

[22] GATYS L A, ECKER A S, BETHGE M. Texture synthesis using convolutional neural networks[C]//Advances in Neural Information Processing Systems. Montréal, Quebec:NIPS, 2015: 1-10.

[23] HE K, WANG Y, HOPCROFT J. A powerful generative model using random weights for the deep image representation[C]//Advances in Neural Information Processing Systems. Barcelona:NIPS, 2016:1-8.

[24] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Advances In Neural Information Processing Systems. Long Beach: NIPS, 2012: 1-9.

[25] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//International Conference on Learning Representations. San Diego:ICLR, 2015: 1-14.

[26] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep feature for discriminative localization[C] //2015 IEEE Conference on Computer Vision and Pattern Recognition.Washington, DC:CVPR, 2016:2921-2929.DOI:10.1109/CVPR.2016.319.

[27] SZEGEDY C, WEI L, YANGQING J, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC:CVPR, 2015(2): 1-9. DOI: 10.1109/CVPR.2015.7298594.

[28] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC:CVPR, 2016, 7(3): 171-180. DOI: 10.1109/CVPR.2016.90.

（编辑　苗秀芝）