计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700128

基于特征可视化分析深度神经网络的内部表征

尚骏远　杨乐涵　何　琨
华中科技大学计算机学院　武汉 430074
(804593872@qq.com)

摘　要　基于可视化的方式理解深度神经网络能直观地揭示其工作机理,即提供了黑盒模型做出决策的解释,在医疗诊断、自动驾驶等领域尤其重要。大部分现有工作均基于激活值最大化框架,即选定待观测神经元,通过优化输入值(如隐藏层特征图谱、原始图片),定性地将待观测神经元产生最大激活值时输入值的改变作为一种解释。然而,这种方法缺乏对深度神经网络深入的定量分析。文中提出了结构可视化和基于规则可视化两种可视化的元方法。结构可视化从浅至深依层可视化,发现浅层神经元具有一般性的全局特征,而深层神经元更针对细节特征。基于规则可视化包括交集与差集规则,可以帮助发现共享神经元与抑制神经元的存在,它们分别学习了不同类别的共有特征与抑制不相关的特征。实验针对代表性卷积网络 VGG 和残差网络 ResNet 在 ImageNet 和微软 COCO 数据集上进行了分析。通过量化分析发现,ResNet 和 VGG 均有很高的稀疏性,通过屏蔽一些低激活值的"噪音"神经元,发现其对深度神经网络分类准确率均没有影响,甚至有一定程度的提高作用。文中通过可视化和量化分析深度神经网络的隐藏层特征,揭示其内部特征表达,从而为高性能深度神经网络的设计提供指导和借鉴。

关键词:深度神经网络;特征可视化;内部表征;共用神经元;抑制神经元
中图法分类号　TP83

Analyzing Latent Representation of Deep Neural Networks Based on Feature Visualization

SHANG Jun-yuan, YANG Le-han and HE Kun
School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract　The working mechanism of deep neural networks can be intuitively uncovered by visualization technique. Visualizing deep neural networks can provide the interpretability on the decision made by the black box model, which is critically important in many fields, such as medical diagnosis and autopilot. Current existing works are mostly based on the activation maximization technique, which optimizes the input, the hidden feature map or the original image, in condition to the neuron that we want to observe. Qualitatively, the change in the input value can be taken as explanation when the neuron has reached nearly the maximum activation value. However, such method lacks the quantitative analysis of deep neural networks. To fill this gap, this paper proposes two meta methods, namely, structure visualization and rule-based visualization. Structure visualization works by visualizing from the shallow layers to the deep layers, and find that neurons in shallow layers learn global characteristics while neurons in deep layers learn more specific features. The rule-based visualization includes intersection and difference selection rule, and it is helpful to find the existence of shared neurons and inhibition neurons that learns the common features of different categories and inhibits unrelated features respectively. Experiments on two representative deep networks, namely the convolutional network VGG and the residual network ResNet, by using ImageNet and COCO datasets. Quantitative analysis shows that ResNet and VGG are highly sparse in representation. Thus, by removing some low activation-value "noisy" neurons, the networks can keep or even improve the classification accuracy. This paper discovers the Latent representation of deep neural networks by visualizing and quantitatively analyzing hidden features, thus providing guidance and reference for the design of high-performance deep neural networks.

Keywords　Deep neural network, Feature visualization, Internal representation, Shared neuron, Inhibition neuron

1　引言

人工智能经历了 3 次发展浪潮:20 世纪 40 年代到 60 年代的控制论,20 世纪 80 年代到 90 年代的联结主义,以及 2012 年以来以深度学习之名迎来的新的热潮。作为人工智能的一个分支,深度学习在近十年得到了快速的发展。深度

到稿日期:2019-07-17　返修日期:2019-10-22　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61772219);中央高校基本科研业务费专项资金(2019kfyXKJC021)
This work was supported by the National Natural Science Foundation of China(61772219) and Fundamental Research Funds for the Central Universities of Ministry of Education of China (2019kfyXKJC021).
通信作者:何琨(brooklet60@hust.edu.cn)

学习模型一直被理解为受生物大脑启发而设计出来的系统。最初的 M-P 神经元[1]确实是模仿生物大脑神经元设计出来的模型,但后来诞生的反向传播算法(Back Propagation,BP)[2]、深层神经网络[3]等是通过数学推导形成的计算模型。国内研究者也对深度学习模型的发展进行了详细的回顾和展望[4-5]。

目前,深度神经网络已经发展出了上百种模型和研究分支。研究者开发出了在手写体识别[6]、图像标注[7]、语义理解[8]和语音识别[9]等方面表现出色的深度神经网络模型。深度神经网络成为了完成感知任务的标准模式,其中图像识别是一大研究热点。在 ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)[10]比赛中,AlexNet[11]创新性地利用了深度卷积神经网络(Convolutional Neural Networks,CNN),使得图片分类准确度得到了大幅度提升,其 Top-5 的错误率仅有 16.4%,相对于 ILSVRC2011 的冠军模型(传统图片分类方法),错误率下降了 9%。受 AlexNet 深度卷积神经网络启发,ILSVRC2014 比赛的冠军模型 GoogleNet 使用 1 * 1 大小的卷积核,并将网络加深为 22 层,但参数量减少到了 AlexNet 的 1/12;亚军模型 VGG[12]通过加深网络到 19 层,并采用统一的 3 * 3 大小的卷积,核使得 Top-5 的错误率降低至 7.3%。而 ResNet[13]通过残差连接结构,用全局平均池化层代替全连接层,使得网络深度达到了 50 层以上,该网络模型在 ILSVRC2015 比赛中取得了 TOP-5 错误率为 3.57%的优异表现,而人眼在上述比赛使用的数据集上的识别错误率为 5.1%。

深度神经网络之所以取得了如此巨大的进步,不仅在于网络结构的优化,也在于 ImageNet 等大型数据集的开发,GPU 计算能力的提高,以及优化方法如 Dropout[14]和 Batch Normalization[15]等的提出。深度神经网络在特征提取和建模等方面相较于浅层模型具有明显的优势,在图像识别、图像分割等领域取得了显著的性能提升;然而,其内部工作机理的研究尚不充分,网络结构的优化和超参的调节更多是试探性的,深度神经网络甚至被当成一个黑匣子。

为了理解深度神经网络的分类决策和表征学习机理,以便于设计与优化深度神经网络框架,研究者提出了各种可视化深度神经网络的方法[16-28],受生物神经网络研究启发的激活值最大化方法[18]是其中的代表性方法。本文在 Yosinski 等[16]对 ILSVRC2012 分类竞赛冠军 AlexNet 神经网络可视化的基础上,对分类性能更高、使用更普遍的代表性卷积网络 VGG 和残差网络 ResNet (ILSVRC2015 分类竞赛冠军)进行了隐藏层特征的可视化和量化分析。

本文的主要贡献如下:

(1)提出了结构可视化的特征可视化元方法,发现浅层神经元的特征提取更加具有全局性和一般性,而深层神经元的特征提取更有针对性与特定性;

(2)提出了基于交集和差集规则的特征可视化元方法,发现了共享神经元与抑制神经元的存在,它们分别学习了不同类别的共有特征并抑制不相关的特征,并通过余弦相关距离计算进一步验证了共用神经元的存在;

(3)通过对 VGG-16 与 ResNet-50 两个代表性网络进行量化对比分析发现,它们均具有很高的稀疏性,通过移除低激活值神经元,使网络更稀疏,但网络的分类准确率没有受到影响,甚至有一定程度的提高。

本文第 2 节简介相关工作;第 3 节对行结构可视化、基于规则可视化、层级余弦相关距离和去神经元方法进行说明和算法定义;第 4 节在所开发的可视化工具软件上进行实验,包括框架、数据集、实验环境的说明,并对对应算法的可视化和量化实验结果进行展示与小结;最后总结全文并展望未来。

2 相关工作

为了理解深度神经网络的决策过程并表征学习机理,国外研究者从可视化角度出发,提出了多种可视化方法。

按可视化对象来分,可视化方法包括可视化卷积核、隐藏层神经元特征图谱和输入图片可视化 3 个方向。Zeiler 等[17]在观察每一层卷积核之后,通过改进过滤器的大小,使得网络在 ImageNet[10]算例上取得了出色的成绩。Yosinski[16]开发出了可视化深度神经网络隐藏层神经元特征图谱的 PC 端可视化工具,其支持视频数据输入,但仅简单地提供了隐藏特征的直接可视化功能。Erhan[18]则将输入图片像素当作变量,通过梯度上升的方法修改输入图片像素值,使得指定神经元激活值变大,从而观察输入图片增强部分对神经元的影响。以上可视化方法在实践中使用简单,但观察效果一般。随后,Zeiler[19]又提出了通过反卷积方法可视化隐藏层神经元特征图谱。Simonyan[20]则构建了输入图片和指定神经元对应的显著图(Saliency Map),使得能更加清晰地观察输入图片各部分对指定神经元作用的大小。与显著图相似,Zintgraf[21]、Zhou[22]与 Fong[26]等利用敏感性分析的方法进行可视化。而 Samek[29]等提出基于热图来可视化并评估深度网络的方法。

按数据量大小来分,可视化方法包括以数据为中心的方法和以网络为中心的方法。以数据为中心的方法利用一个大型图片数据集(如 ImageNet 和微软 Common Objects in Context(COCO)数据集[30])来进行统计分析,通过展示指定神经元激活值最大对应的图片来理解隐藏层神经元的学习机理。以网络为中心的方法,即只具有一个预训练好的网络和少量图片,通过利用上述各种可视化方法进行直接观察分析。

除此之外,国内外研究者还提出了一些其他的可视化方法,如降维[23]、训练可视化隐藏层表征网络[24]、最大化响应变量与选择特征之间的互信息[28]等。特别地,Net2Vec[25]利用基于相应卷积核的响应,将语义概念映射成表征向量,从而探究出与本文相似的共享神经元等结论。更多可视化方法可参阅文献[27]。

3 可视化相关算法

本节详细叙述所提出的结构可视化、基于规则可视化、层级余弦相关距离、去低值神经元的算法。常用符号包括:数据集 D,如 ImageNet[10]或微软 COCO[30];网络模型 A,如 VGG-16 或 ResNet-50。$\|A\|$ 表示模型层数,$\|A_l\|$ 表示第 l 层神经元的个数。A_{li} 代表第 l 层的第 i 个神经元,其中 $l\leqslant\|A\|$,$i\leqslant\|A_l\|$。$\alpha(A_{li})$ 表示求激活值运算,具体为计算 A_{li} 神经元特征图谱的平均激活值。x 为输入图片,$\phi(x)$ 则是对原始输入图片进行预处理的函数,该函数的具体运算是将输入图片的 R,G,B 3 个通道的值分别减去数据集 D 全体图片在 R,G,

B 3 个通道上的像素平均值。隐藏层神经元特征图谱可视化方法记为 P。

3.1 结构可视化

结构可视化是通过可视化网络的整体结构来观察图片决策过程的一种方法，由浅至深选择特定神经元展示。高层神经元与低层神经元相连，低层神经元将决策信息传递给高层神经元，最终传递至最后一层进行概率预测。

完整的结构可视化过程如算法 1 所示。

算法 1 结构可视化算法

输入：深度神经网络模型 A，输入图片 x，隐藏层神经元特征图谱可视化方法 P，神经元选择方法 T，步长 s，起始层号 st，截止层号 sp

输出：被选择神经元存放集合 O

1. x←ϕ(x)；//预处理
2. 将 x 输入深度神经网络模型 A；
3. FOR l from st to sp step s；
4. O_l←T(A_l) //选择第 l 层特定神经元存放在O_l；
5. END FOR；
6. 利用可视化方法 P 层级展示选择的神经元集合 O。

算法 1 输入图片 x，从网络浅层到深层，选择每一层的特定神经元(如激活值最大的神经元)进行展示，通过层层可视化的方法，加深理解神经网络的整体结构和决策过程。

3.2 基于规则可视化

为避免盲目地从每层上百个神经元中随机选择神经元进行可视化展示，定义如下 3 个规则。

(1)最大化规则。输入一张图片 x，在深度神经网络中产生神经元激活值，针对第 l 层选择$\max_n(A_l)$神经元，n 代表平均激活值最大的前 n 个神经元，可以理解为平均激活值越大的神经元对图片 x 越敏感。

(2)交集规则。输入多张图片，在深度神经网络上分别产生神经元激活值，对于相同层，使用神经元选择算法 T 选出神经元集合后对其求交集；输出对多张图片均敏感的神经元，称为共用神经元。

(3)差集规则。输入多张图片，在深度神经网络上分别产生神经元激活值，针对相同层，使用神经元选择算法 T 选出神经元集合后对其求差集。选出对某图片敏感而对其他图片不敏感的神经元，称其为差异神经元。

本文设计了上述 3 种简单的神经元选择规则。通过不同规则选择带有不同联系关系的神经元(如通过交集规则选出共用关系神经元)，可以进一步研究不同神经元的关系。

交集可视化操作如算法 2 所示。

算法 2 交集可视化算法

输入：深度神经网络模型 A，输入图片集 X，隐藏层神经元特征图谱可视化方法 P，神经元选择方法 T，指定层 l

输出：交集神经元 U

1. 初始化 U=$O_l^{x_0}$；
2. FOR all x in X/x_0；
3. x←ϕ(x)；//预处理
4. 将 x 输入深度神经网络模型 A；
5. O_l^x←T(A_l^x) //选择以 x 为输入的第 l 层神经元存放在O_l^x；
6. U←U O_l^x；
7. ENDFOR；
8. 利用可视化方法 P 展示神经元集合 U。

3.3 层级余弦相关距离

层级余弦相关距离从层级维度出发，研究输入不同图片时的层维度神经元的整体相似度。

完整的层级余弦相关距离计算如算法 3 所示。

算法 3 层级余弦相关距离

输入：深度神经网络模型 A，指定层 l，输入图片 x 和 y

输出：余弦相关距离 d

1. x←ϕ(x)，y←ϕ(y)；
2. 将 x，y 输入深度神经网络模型 A；
3. 计算余弦相关距离；
4. 利用 α(·)操作求平均激活值；
5. $d=[\alpha(A_{l0}^x)\alpha(A_{l1}^x)\cdots\alpha(A_{ln}^x)]*[\alpha(A_{l0}^y)\alpha(A_{l1}^y)\cdots\alpha(A_{ln}^y)]^T$。

算法 3 选择余弦相似度作为距离度量标准，同类图片输入的余弦相似度更大，不同类图片输入的余弦相似度较小。

3.4 去低值神经元算法

去低值神经元算法主要考查网络的鲁棒性，通过观察预测值变化了解网络的鲁棒性以及表征学习能力，详细如算法 4 所示。

算法 4 去低值神经元算法

输入：深度神经网络模型 A，数据集 D，神经元选择方法 T，指定层 l

输出：原始准确度 p，去神经元后的准确度 p′

1. 对数据集 D 在深度神经网络模型 A 上计算预测准确度 p；
2. A←A−T(A_l) 去除选定的神经元；
3. 对数据集 D 在新深度神经网络模型 A 上计算预测准确度 p′；
4. 返回 p 与 p′。

算法 4 对指定网络选择特定层并去除特定百分比神经元。通过此方法研究网络的鲁棒性，并观察去除神经元后对整体网络的影响。其中，神经元选择方法 T 默认选择此层平均激活值最小的 10%神经元。

4 实验

本文选择在 ImageNet[10] 上有优良性能表现的 VGG-16[12] 和 ResNet-50[13] 卷积神经网络作为可视化和特征分析框架，具体参数设置如表 1 所列。数据集选取 ImageNet-2012 验证集中 5 万张图片和微软公司 COCO[30] 训练集中 8 万张图片。实验环境为亚马逊云 p2xlarge 实例，选择亚马逊专为深度学习提供的 ami(ami-a3b3d4b5)。p2xlarge 示例具有 4 核 CPU，内存 60 GB，一个 Nvida-K80 显卡(12 GB 显存)，并具有灵活可扩展 SSD。本文使用 TensorFlow + Keras[31] 深度学习框架，利用 Keras 在 ImageNet 上预训练的权重进行实验。

表 1 神经网络参数

Table 1 Neural network parameters

神经网络	参数个数	网络大小/MB
VGG-16	138357544	528
ResNet-50	25583592	98

4.1 结构可视化

具体选择一张输入图片 x，隐藏层神经元特征图谱可视化方法 P 为可视化隐藏层激活值图谱，深度神经网络模型 A 为 ResNet-50，步长 s 为 2，起始层号为 10，截止层号为 46。

$\phi(x)$预处理操作为将输入图片 x 缩放成 224 * 224，将 RGB 转化成 BGR 排列，并对 BGR 通道每个像素点减去正则化项，即减去训练集图片的三通道像素平均值。神经元选择方法 T 为选择指定层平均激活值最大的 9 张特征图谱。

图 1 对某张输入图片进行了结构化展示。图 1(a)展示了 ResNet-50 卷积层输出进入 ReLU 激活函数后的隐藏层激活值图片，其中包含 10，13，16，19，22，25，28，31，34，37，40，43，46，共 13 个激活层(每层选取间隔为 2)，每层对应的特征图谱数量和大小分别是：

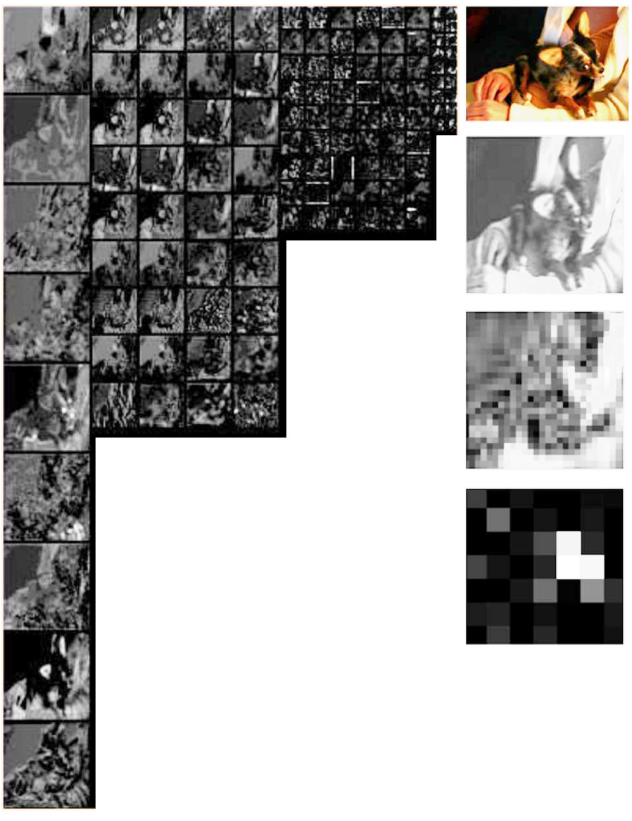
1)第 10 层包含 256 个特征图谱，特征图谱大小为 55 * 55；

2)第 13，16，19，22 层包含 512 个特征图谱，特征图谱大小为 28 * 28；

3)第 25，28，31，34，37，40 层包含 1024 个特征图谱，特征图谱大小为 14 * 14；

4)第 43，46 层包含 2048 个特征图谱，特征图谱大小为 7 * 7。

图 1(a)中每一列为 ResNet50 采样的 13 层中对应平均激活值最大的 9 张特征图谱，其比例大小不变，从第一层 55 * 55 大小，到最后一层 7 * 7 大小；图 1(b)为随机从图 1(a)激活图片中选择的图片，从上到下每一行分别展示了原图、第 10 层 13 号特征图谱放大图、第 22 层 222 号特征图谱放大图和第 46 层 68 号特征图谱放大图。



(a) (b)

图 1 可视化效果图

Fig. 1 Visualization features

通过整体观察，不仅可以发现神经网络在设计上的共识，即通过池化层减小输入大小，同时增加卷积核数；而且可以看出在浅层网络层，图片整体结构能够很好地保留，并且图片颜色反转、轮廓线条凸显，更具备一般特征，而在深层网络层，整体结构消失，图片表现局部特征，更具有针对性。

AlexNet 卷积神经网络具有与 LeCun 等设计的 LeNet[32] 卷积神经网络非常相似的架构，但是网络的层数更深，每层有更多的滤波器，并且具有堆叠的卷积层。AlexNet 由 11×11，5×5，3×3 卷积核，最大池化层，Dropout 层，ReLU 激活函数等基本模块组成，在每个卷积层和全连接层之后附加了 ReLU 激活函数。VGG 卷积神经网络由 16 个卷积层组成，有非常统一的架构，与 AlexNet 类似，只有 3×3 卷积，但有很多过滤器。He[13] 提出的残差神经网络(ResNet)引入了具有“跳过连接”的新颖架构，并具有大批量标准化功能。这种跳过连接也称为门控单元或门控循环单元，与 RNN 具有很强的相似性。通过这种技术，ResNet 能够训练 152 层的神经网络，同时仍具有比 VGG 更低的复杂度。

神经网络都朝着大而深的方向发展，而深层的网络会带来更多的参数，增加了计算量，在提高准确率的同时也会增加时间开销。表 2 对比了 3 种卷积神经网络的层数、参数总量、错误率和识别速度，可以看出：相较于 AlexNet，VGG-16 增加了成倍的参数量，而 ResNet-50 却大幅减少了参数量；VGG-16 和 ResNet-50 的错误率远低于 AlexNet 卷积神经网络；随着神经网络深度与参数量的增加，神经网络识别速度会减慢，但是 VGG-16 卷积神经网络和 ResNet-50 卷积神经网络由于对结构进行了优化，因此错误率大幅下降。

表 2 AlexNet，VGG-16，ResNet-50 的性能对比

Table 2 Performance comparison of AlexNet，VGG-16 and ResNet-50

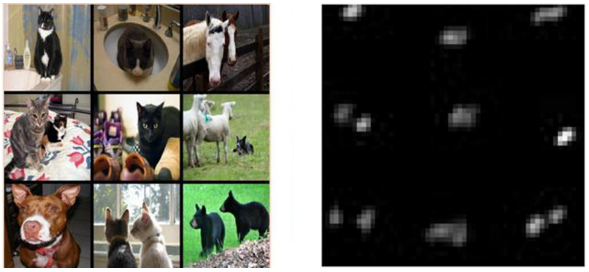
卷积神经网络	层数	参数总量/M	Top1 错误率	Top5 错误率	识别速度/ms
AlexNet	8	61.0	42.90	19.80	14.56
VGG-16	16	138.0	27.00	8.80	128.62
ResNet-50	50	25.5	24.01	7.02	103.58

本文进一步可视化深层神经元的特征图谱。将 8 万张来自 2014 年微软 COCO 训练集的图片输入 VGG-16 卷积神经网络，对每一层每一特征图谱求取平均激活值，并对每一层每个特征图谱保留使平均激活值最大的 9 张原始图片和激活值图片，结果发现了共用神经元、抑制神经元和特征细分神经元等有趣的现象。

图 2 所示为 VGG-16 Conv5_3 卷积层中的第 71 个特征图谱单元。图 2(a)为平均激活值最大的 9 张原始图片；图 2(b)为这 9 张原始图片对应的特征图谱，仅有对应耳朵位置的像素处于明亮状态，可以看出此特征图谱单一地识别了长耳朵这一特征，即不同分类图片共用该神经元来识别长耳朵这一特征。

图 3 所示为 VGG-16 Conv5_2 的第 11 个特征图谱单元。此特征图谱对应的预测结果均与人脸无关，如图 3(a)第二行第三列预测结果为西装，第三行第二列预测结果前 3 为西兰花、黄瓜、榴莲。虽然图 3(a)图片中具有人脸或貌似人脸的物体，但右侧特征图谱将对应特征进行了剔除(全黑)，使得其

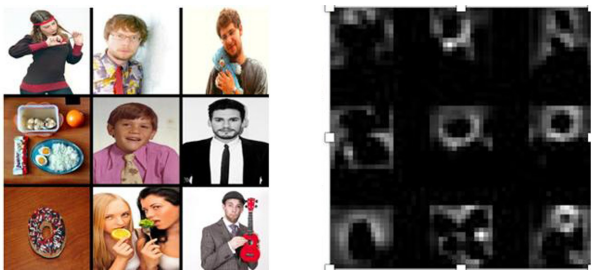
最终的分类结果排除了人脸的预测特征。具有上述抑制与分类标签不相关特征行为的神经元被称为抑制神经元。



(a) (b)

图 2 71 号神经元 Top 9 激活图

Fig. 2 Top 9 activation feature map of No. 71 neuron



(a) (b)

图 3 10 号神经元的 Top 9 激活图

Fig. 3 Top 9 activation feature map of No. 10th neuron

从 Yosinski 等的可视化研究[16]可以看出，AlexNet 的神经元没有区分人脸和动物脸(如狮子)。而 VGG 在以上实验中都能区别出多个特征，这从侧面反映了 VGG 的特征区分度更强，从而有更好的分类能力。

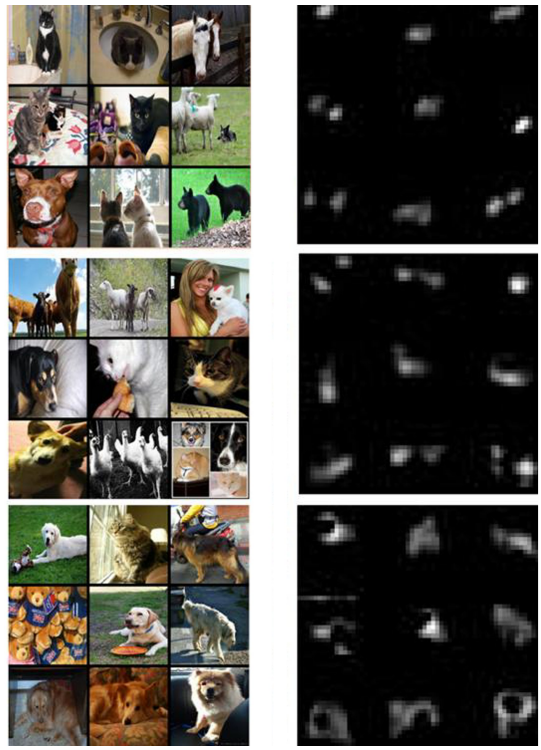
4.2 基于规则的可视化

具体选择规则关键字为交集，深度神经网络模型 A 为 VGG-16，输入图片集 X 为一种猫和一种狗的两张图片，隐藏层神经元特征图谱可视化方法 P 为可视化隐藏层激活值图谱和对应在微软 COCO 数据集中使其激活值最大的原始图片，神经元选择方法 T 为基于规则中的最大化算法，指定层 l 为最后一层卷积层。

实验表明，当针对相似类别的图片进行分类时，网络模型中确实存在一些共用神经元。如图 4 所示，在选定的猫和狗对应前 9 个最大平均激活值的神经元中，有 3 个是相同神经元，而从原始图片来看，这 3 个神经元确实选定了部分猫和部分狗；从激活图片来看，也确实识别了猫和狗的面部特征。

4.3 层级余弦相关距离的计算

具体选择深度神经网络模型 A 为 VGG-16，指定层 l 为最后一层卷积层($\| A_l \| = 512$)，输入图片为从 ImageNet-2012 验证集为 1 000 个类中随机选取的 12 个类图片。将 12 个类图片输入深度神经网络模型 A，最后一个卷积层将产生 12 个 512 维激活值向量，对其求相关余弦距离。



(a) (b)

图 4 猫、狗的 71,274,446 共用神经元激活图

Fig. 4 No. 71, No. 274 and No. 446 neurons show that they share some common features of cats and dogs

如表 3 所列，针对选取的 12 个类，每行作为一个单独研究对象，黑体标出最大值(行)，斜体标出最小值(行)。第一行的瑞士山狗，最大值为 0.69，对应彭布罗克犬；最小值为 0.26，对应教堂。倒数第三行监狱，最大值为 0.63，对应教堂；最小值为 0.26，对应瑞士山狗。

表 3 12 个类的余弦相关距离

Table 3 Cosine correlation distances among 12 categories

	瑞士山狗	手提包	蛇	教堂	贝壳	皮背龟	胸甲	锁子甲	彭布罗克	监狱	哑铃	黑天鹅
瑞士山狗	1	0.31	0.34	*0.26*	0.40	0.42	0.39	0.27	**0.69**	0.26	0.37	0.43
手提包	*0.31*	1	0.54	0.43	0.59	0.54	**0.68**	0.57	0.45	0.44	0.61	0.42
蛇	*0.34*	0.54	1	0.38	**0.82**	0.68	0.65	0.50	0.47	0.36	0.52	0.51
教堂	*0.26*	0.43	0.38	1	0.39	0.41	0.59	0.36	0.33	**0.63**	0.46	0.32
贝壳	0.40	0.59	0.82	0.39	1	**0.69**	0.66	0.50	0.53	*0.37*	0.52	0.49
皮背龟	0.42	0.54	0.68	0.41	**0.69**	1	0.66	0.52	0.47	*0.39*	0.55	0.62
胸甲	*0.39*	0.68	0.65	0.59	0.66	0.66	1	0.54	0.48	0.58	**0.70**	0.51
锁子甲	*0.27*	**0.57**	0.5	0.36	0.50	0.52	0.54	1	0.36	0.35	0.46	0.46
彭布罗克	**0.69**	0.45	0.47	*0.33*	0.53	0.47	0.48	0.36	1	0.34	0.49	0.43
监狱	*0.26*	0.44	0.36	**0.63**	0.37	0.39	0.58	0.35	0.34	1	0.53	0.32
哑铃	*0.37*	0.61	0.52	0.46	0.52	0.55	**0.70**	0.46	0.49	0.53	1	0.45
黑天鹅	0.43	0.42	0.51	0.32	0.49	**0.62**	0.51	0.46	0.43	*0.32*	0.45	1

通过 VGG-16 卷积神经网络最后一层求取平均激活值向量，并进行余弦距离相关分析，可得结论：平均激活值这一统计量具有一定的依据；深层神经单元具有共用特性，针对相似类事物，存在许多共用神经元参与识别工作。

4.4 去低值神经元

具体的去低值神经元实验的深度神经网络模型 A 为 VGG-16 和 ResNet-50，数据集 D 为 ImageNet-2012 验证集的 5 万张图片，指定层 l 均为最后一层卷积层。神经元选择方法 T 则多种多样，考虑如下两种选择方法。

全局去单体低值神经元：对于每张图片，在指定层 l，将指定数量的激活值最小的神经元进行在线去除和预测。

全局去全体低值神经元：预先在数据集上计算全体图片在指定层 l 产生的平均激活向量，对其进行排序，将全体图片中指定数量的激活值最小的神经元进行离线去除和预测。

4.4.1 全局去单体低值神经元

在相同数据集上，针对一张图片，分别计算 VGG-16，ResNet-50 两种网络的最后一个卷积层上每个特征图谱的平均激活值，并对 20% 的低激活值特征图谱进行归零操作（VGG-16 为 100 张特征图谱，ResNet-50 为 400 张特征图谱），使得低激活值特征图谱对图片分类预测的贡献变为 0。记录图片分类预测的精确度，如表 4 所列。

表 4 全局去单体低值神经元的 Top1、Top5 分类准确度

Table 4 Classification accuracy comparison of Top1 and Top5 for low-valued neuron removal operation on single image

预测结果	VGG-16	ResNet-50
TOP1 p	0.64274	0.67692
TOP1 p'	**0.64284**	**0.68102**
TOP5 p	**0.85592**	0.87494
TOP5 p'	0.85586	**0.88296**

其中，p 为原始预测准确度，p' 为去神经元预测准确度。结果表明，全局去单体低值神经元对 VGG-16 预测结果的准确度基本持平，而对 ResNet-50 预测结果的准确度有一定提升。可见，对于特定个体的预测，单层激活单元中只有少量的关键部分具有贡献，存在可称为干扰（噪音、冗余）的特征图谱，对网络模型的预测准确度造成削弱效应。

4.4.2 全局去全体低值神经元

在相同数据集上，分别计算两个网络最后一层的全局平均激活值（具有 1 个 512 维向量和 1 个 2048 维向量），分别对其元素值进行升序排列后将最低的 50 和 200 张特征图谱进行归零操作（50 和 200 的选取为试探值），使得全局激活值特征图谱对预测的贡献变为 0，相当于精简了网络结构。记录预测精确度，如表 5 所列。

表 5 全局去全体低值神经元的 Top1、Top5 分类准确度

Table 5 Classification accuracy comparison of Top1 and Top5 for low-valued neuron removal operation on all images

预测结果	VGG-16	ResNet-50
TOP1 p	**0.64274**	0.67692
TOP1 p'	0.60724	**0.67730**
TOP5 p	**0.85592**	0.87494
TOP5 p'	0.83230	**0.88008**

结果表明，全局去全体低值神经元对 VGG-16 卷积神经网络和 ResNet-50 卷积神经网络有着不同的结果。VGG-16 卷积神经网络在去掉全局全体低值 50 个激活特征图后，预测精确度明显降低；ResNet-50 卷积神经网络的精确度有少许提高，但没有前实验提高得明显，说明 ResNet-50 卷积神经网络结构的稀疏性更高。

4.5 稀疏性可视化实验

对于卷积神经网络 VGG-16 与 ResNet-50，将 ImageNet-2012 验证集的 5 万张图片输入网络后，对最后一层卷积层的每个神经元求平均激活值。VGG-16 卷积神经网络最后一层的维度为 512，ResNet-50 卷积神经网络最后一层的维度为 2048，故分别得到 1000 个 512 维平均激活值向量和 1000 个 2048 维平均激活值向量。

随机选取 ID 为 284，151，310，对应类别为 Siamese Cat，Chihuahua Dog，Ant 的图片进行直方图分析，结果如图 5、图 6 所示。可以发现，对于每一类，激活值小的值占比很大，说明激活特征有很大的稀疏性。

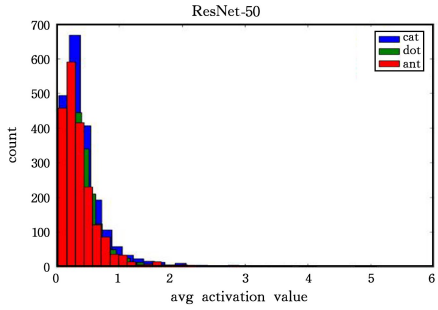


图 5 ResNet-50 最后一个卷积层的激活值分布（电子版本效果最佳）

Fig. 5 Activation value distribution of the last convolutional layer in ResNet-50 (better view in electronic version)

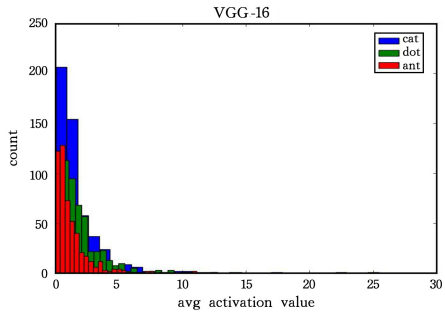
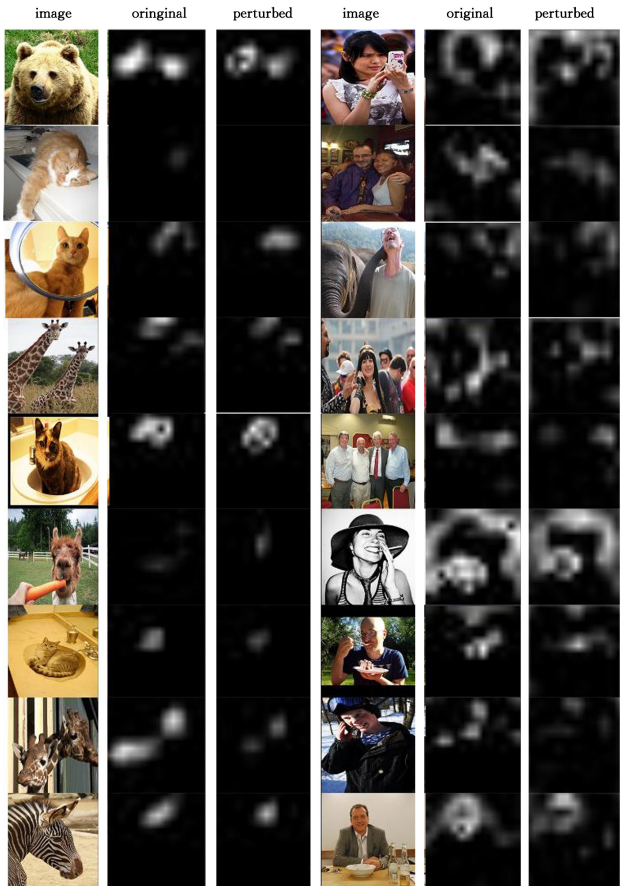


图 6 VGG-16 最后一个卷积层的激活值分布（电子版本效果最佳）

Fig. 6 Activation value distribution of the last convolutional layer in VGG-16 (better view in electronic version)

4.6 VGG-16 神经元特征图谱随机扰动实验

本小节对 VGG-16 卷积神经网络在受到数据扰动时的特征获取程度进行实验。在 Microsoft COCO 数据集[22]的验证集里随机抽取了动物与人脸各 9 张图像，并对这些图像加入方差为 0.01 的零均值高斯白噪声。如图 7 所示，在 VGG-16 Conv5_3 卷积层中的第 71 个特征图谱单元及 Conv5_2 的第 11 个特征图谱的可视化中观察到数据扰动对 VGG-16 的神经元特征提取的影响不大，体现出了 VGG-16 具有一定的鲁棒性。



第一列为始图片，第二列为 VGG-16 Conv5_3 卷积层中的第 71 个特征图谱单元，第三列为原始图片加入 N(0,0.01)的高斯白噪声后对应的特征图谱单元；第四－第六列分别为原始图片、Conv5_2 的第 11 个特征图谱单元、注入 N(0,0.01)的高斯白噪声后的对应特征图谱单元

图 7 Microsoft COCO 的随机扰动对比

Fig. 7 Random perturbation comparison of images in Microsoft COCO dataset

结束语 本文在相关研究者对 AlexNet 的深度可视化研究的基础上，针对 VGG-16 和 ResNet-50 两个性能更高且使用更普遍的卷积神经网络进行了可视化分析和进一步的特征统计分析。主要贡献在于：

(1)设计并实现了结构可视化和基于规则可视化两种可视化的元方法，发现和验证了特征共用神经元、抑制神经元等不同类型神经元的存在。这两种算法有规范的可视化流程，具有较强的灵活性，揭示了浅层网络和深层网络的学习特征。浅层网络层图片整体结构得到较好的保留，并且图片颜色反转、轮廓线条凸显；而深层网络层的整体结构消失，图片表现局部特征。基于规则可视化中的交集规则实验展示了神经元针对相似类的表现，验证了共用神经元和抑制神经元的存在。

(2)基于特征统计和去低值神经元进一步理解深度神经表征，提出了层级余弦可视化和去神经元算法。前者从特征统计角度量化证实了：平均激活值这一统计量具有一定的依据，深层神经单元具有共用特性，针对相似类事物，存在许多共用神经元参与识别工作。通过对比实验，观察到去低值神经元网络具有较强的稀疏性。比较两个网络的稀疏性，通过预测值的变化(ResNet-50 在 ImageNet-2012 验证集上的分类准确度提高了 0.7%左右，VGG-16 分类准确度基本不变)进一步说明 ResNet-50 有更强的稀疏性，表明其具有更强的表征学习能力。同时，对 VGG-16 进行数据扰动，比较了原始图像特征图谱和噪声图像特征图谱，发现 VGG-16 也具有一定的鲁棒性。

今后工作可以从两方面展开：1)利用结构可视化与基于规则的可视化方法在新型网络上进行可视化研究；2)利用共享神经元与低激活值神经元的特征设计网络压缩算法。

参 考 文 献

[1] MCCULLOCH D E, PITTS W. A logical calculus of ideas immanent in nervous activity [J]. Bulletin of Mathematical Biophysics, 1943, 5: 115-133.

[2] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323: 533-536.

[3] HINTON G E, SALAKHUTDINOV G E. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313 (5786): 504-507.

[4] JIAO L C, YANG S Y, LIU F, et al. Seventy Years beyond Neural Networks: Retrospect and Prospect[J]. Chinese Journal of Computers, 2016, 39(8): 1697-1716.

[5] ZHOU F Y, JIN L P, DONG J. Review of Convolutional Neural Network[J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.

[6] CIRE ŞAN D C, MEIER U, GAMBARDELLA L M, et al. Deep, big, simple neural nets for handwritten digit recognition [J]. Neural Computation, 2010, 22(12): 3207-3220.

[7] FARABET C, COUPRIE C, NAJMAN L, et al. Learning hierarchical features for scene labeling [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1915-1929.

[8] ZHAO R, OUYANG W, LI H S, et al. Saliency detection by multi-context deep learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1265-1274.

[9] MOHAMED A, DAHL G E, HINTON G E. Acoustic modeling using deep belief networks [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 14-22.

[10] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[11] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Advances in neural information processing systems. 2012: 1097-1105.

[12] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C]// 3rd International Conference on Learning Representations (ICLR). 2015.

[13] HE K M, ZHANG X, REN S Q, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition. 2016:770-778.

[14] SRIVASTAVA N,HINTON G E,KRIZHEVSKY A,et al. Dropout:a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research,2014,15(1):1929-1958.

[15] 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 (ICML). 2015.

[16] YOSINSKI J,CLUNE J,NGUYEN A M,et al. Understanding neural networks through deep visualization [C]// Deep Learning Workshop,International Conference on Machine Learning (ICML). Lille,France,2015.

[17] ZEILER M D,FERGUS R. Visualizing and understanding convolutional networks[C]// European Conference on Computer Vision (ECCV). Springer International Publishing,2014:818-833.

[18] ERHAN D,BENGIO Y,COURVILLE A,et al. Visualizing higher-layer features of a deep network:Technical Report 1341 [R]. University of Montreal,2009.

[19] ZEILER M D,TAYLOR G W,FERGUS R. Adaptive deconvolutional networks for mid and high level feature learning[C]// International Conference on Computer Vision (ICCV). 2011.

[20] SIMONYAN K,VEDALDI A,ZISSERMAN A. Deep inside convolutional networks:Visualising image classification models and saliency maps [C]// 2nd International Conference on Learning Representations (ICLR). 2014.

[21] ZINTGRAF L M,COHEN T S,ADEL T,et al. Visualizing deep neural network decisions:Prediction difference analysis [C]// 5th International Conference on Learning Representations (ICLR). 2017.

[22] ZHOU B L,KHOSLA A,LAPEDRIZA A,et al. Learning deep features for discriminative localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:2921-2929.

[23] MAATEN L V D,HINTON G E. Visualizing data using t-SNE [J]. Journal of Machine Learning Research,2008(9):2579-2605.

[24] DOSOVITSKIY A,BROX T. Inverting visual representations with convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:4829-4837.

[25] FONG R,VEDALDI A. Net2Vec:Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks [C]// 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[26] FONG R C,VEDALDI A. Interpretable Explanations of Black Boxes by Meaningful Perturbation[C]// Proceedings of the International Conference on Computer Vision (ICCV). 2017:3449-3457.

[27] NGUYEN A,YOSINSKI J,CLUNE J. Understanding Neural Networks via Feature Visualization:A Survey[C]// Explainable AI:Interpreting,Explaining and Visualizing Deep Learning. Cham:Springer,2019:55-76.

[28] CHEN J B,SONG L,WAINWRIGHT M J,et al. Learning to explain:An information-theoretic perspective on model interpretation[C]// Proceedings of the 35nd International Conference on Machine Learning (ICML). 2018.

[29] SAMEK W,BINDER A,MONTAVON G,et al. Evaluating the visualization of what a deep neural network has learned [J]. IEEE Transactions on Neural Networks and Learning Systems,2016,28(11):2660-2673.

[30] LIN T Y,MAIRE M,BELONGIE S J,et al. Microsoft COCO:Common objects in context[C]// European Conference on Computer Vision (ECCV). Cham:Springer,2014:740-755.

[31] ABADI M,AGARWAL A,BARHAM P,et al. Tensorflow:Large-scale machine learning on heterogeneous distributed systems [J]. arXiv:1603.04467,2016.

[32] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

SHANG Jun-yuan, born in 1994, M. S. candidate. His main research interests include machine learning and deep learning.

HE Kun, born in 1972, professor and Ph.D. supervisor. Her main research interest include machine learning, deep learning, and optimization algorithms.