面向对抗样本的深度神经网络可解释性分析

董胤蓬 [1] 苏 航 [1] 朱 军 [1]

摘 要 虽然深度神经网络 (Deep neural networks, DNNs) 在许多任务上取得了显著的效果, 但是由于其可解释性 (Interpretability) 较差, 通常被当做 "黑盒"模型. 本文针对图像分类任务, 利用对抗样本 (Adversarial examples) 从模型失败的角度检验深度神经网络内部的特征表示. 通过分析, 发现深度神经网络学习到的特征表示与人类所理解的语义概念之间存在着不一致性. 这使得理解和解释深度神经网络内部的特征变得十分困难. 为了实现可解释的深度神经网络, 使其中的神经元具有更加明确的语义内涵, 本文提出了加入特征表示一致性损失的对抗训练方式. 实验结果表明该训练方式可以使深度神经网络内部的特征表示与人类所理解的语义概念更加一致.

关键词 深度神经网络, 可解释性, 对抗样本, 视觉特征表示

引用格式 董胤蓬, 苏航, 朱军. 面向对抗样本的深度神经网络可解释性分析. 自动化学报, 2022, **48**(1): 75−86

DOI 10.16383/j.aas.c200317

开放科学(资源服务)标识码(OSID):

Interpretability Analysis of Deep Neural Networks With Adversarial Examples

DONG Yin-Peng[1] SU Hang[1] ZHU Jun[1]

Abstract Deep neural networks (DNNs) have demonstrated impressive performance on many tasks, but they are usually considered opaque due to their poor interpretability. In this paper, we examine the internal representations of DNNs on image classification tasks using adversarial examples, which enable us to analyze the interpretability of DNNs in the perspective of their failures. Based on the analyses, we find that the learned features of DNNs are inconsistent with human-understandable semantic concepts, making it problematic for understanding and interpreting the representations inside DNNs. To realize interpretable deep neural networks, we further propose an adversarial training scheme with a consistent loss such that the neurons are endowed with the human-interpretable concepts to improve the interpretability of DNNs. Experiments show that the proposed method can make the features in DNNs more consistent with semantic concepts.

Key words Deep neural networks (DNNs), interpretability, adversarial examples, visual representations

Citation Dong Yin-Peng, Su Hang, Zhu Jun. Interpretability analysis of deep neural networks with adversarial examples. *Acta Automatica Sinica*, 2022, **48**(1): 75−86

深度神经网络 (Deep neural networks, DNNs)[1] 由于在语音识别、图像分类、自然语言处理等诸多领域取得了很好的效果, 近年来获得了人们的广泛关注. 但是由于缺乏对其内部工作机理的理解与分析[2], 深度神经网络通常被看作 "黑盒"模型, 导致用户只能观察模型的预测结果, 而不能了解模型产生决策的原因. 深度神经网络的不可解释性也会极大地限制其发展与应用. 例如, 在诸如医疗、自动驾驶等许多实际的应用场景中, 仅仅向用户提供最终的预测结果而不解释其原因并不能够满足用户的需求. 用户需要获取模型产生决策的原因来理解、认可、信任一个模型, 并在模型出错时修复模型的问题. 因此, 研究提升模型可解释性的学习算法、使用户可以理解信任模型、并与模型进行交互变得至关重要.

近些年来, 有很多的方法尝试去解决深度神经网络的可解释性问题. 例如, 一个模型对于图像的分类结果可以归因于图像的关键性区域[3] 或者其他类似图像[4]. 同时, 一系列的工作研究如何可视化深度神经网络内部神经元学习到的特征[5−8]. 但是这些方法存在以下几个问题: 1) 它们通常是在模型训练结束后进行解释, 并不能在训练的过程中约束其学习到一个可解释的模型; 2) 它们仅仅关注模型对于正常样本的预测进行解释与分析, 而忽视了模型在

收稿日期 2020-05-15 录用日期 2020-08-27
Manuscript received May 15, 2020; accepted August 27, 2020
国家自然科学基金 (61620106010, U19B2034, U1811461), 清华国强研究院项目资助
Supported by National Natural Science Foundation of China (61620106010, U19B2034, U1811461) and the Tsinghua Institute for Guo Qiang
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 清华大学计算机科学与技术系 北京 100084
1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084

现实场景中可能遇到的对抗样本 (Adversarial examples)[9−14]; 3) 它们并没有解释模型发生错误的原因, 也不能让用户针对性地修复模型的问题.

本文针对图像分类任务, 利用对抗样本检验深度神经网络的内部特征表示. 对抗样本是指攻击者通过向真实样本 (Real examples) 中添加微小的、人眼不可察觉的扰动, 导致模型发生预测错误的样本. 真实样本和对抗样本可以从正反两方面研究深度神经网络的行为, 既可以通过真实样本分析模型产生正确预测的原因, 同时也可以通过对抗样本分析模型发生错误的原因, 以深入探究深度神经网络的运行机制. 虽然利用模型预测错误的真实样本分析其产生错误的原因也是一种可行的方法, 但是真实样本中发生的错误往往是比较小的错误, 相比于对抗样本的预测错误可以忽略不计. 例如, 模型可能会将一个真实的公交车图片错分为客车, 这种错误可以被接受; 然而如果模型将一个对抗的公交车图片错分为飞机, 则不能够被我们所接受. 通过将对抗样本与真实样本输入到深度神经网络中并检验其特征表示, 我们发现深度神经网络内部学习到的特征表示与人类所理解的语义概念之间存在着极大的不一致性. 如图 1 所示, 神经元学习到的特征通常用对其产生强响应的样本所表示[8]. 当只使用真实样本时, 神经元会检测某种语义概念. 但是会存在其他的样本 (例如蓝色圆圈标记的对抗样本) 也对神经元产生很强的响应, 尽管这些样本的语义概念十分不一致. 这使得神经元学习得到的特征难以解释.

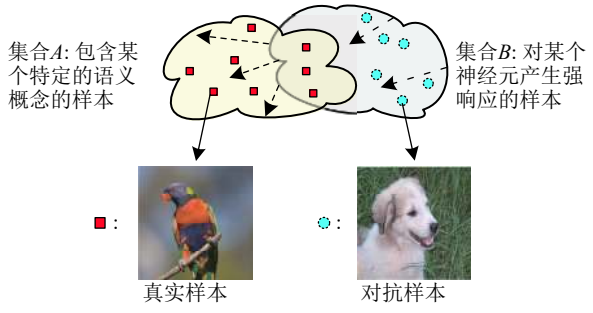


图 1 语义概念与神经元学习到的特征存在不一致性的示意图

Fig. 1 Demonstration of the inconsistency between a semantic concept and the learned features of a neuron

基于以上的分析, 本文进一步提出了加入特征表示一致性损失的对抗训练方式 (Adversarial training with a consistent loss), 其目标是在模型的训练过程中学习到人类可解释的特征表示. 通过加入对抗样本与真实样本特征表示的距离作为一致性损失函数训练深度神经网络, 可以使网络在特征层面上消除掉对抗噪声的干扰, 使其对于对抗样本的特征表示与真实样本的特征表示尽量接近. 如图 1 所示, 对于深度神经网络内部的某个神经元, 如果该神经元检测到的特征与人类所理解的语义概念不一致时, 意味着会存在蓝色圆圈标记的对抗样本对其产生很强的响应. 然而这些对抗样本所对应的真实样本并不会对其产生很强的响应, 这就导致了一致性损失很大. 通过加入特征表示一致性的损失函数, 可以使得该神经元学习到的特征与人类所理解的某个语义概念相关联 (如虚线所示). 这个过程最终会使得深度神经网络内部的神经元学习到可以抵抗对抗噪声干扰的特征, 从而在某个语义概念出现时产生响应、不出现时不产生响应. 因此该方法可以提升深度神经网络的可解释性. 实验结果表明在一些可解释性度量标准下, 该训练方式可以使深度神经网络内部的特征表示与人类所理解的语义概念更加一致, 得到可解释性更好的深度神经网络.

本文的主要贡献有: 1) 提出利用对抗样本分析深度神经网络的可解释性, 并发现神经元学习到的特征表示与人类所理解的语义概念之间存在不一致性; 2) 提出了加入特征表示一致性损失的对抗训练方式, 可以极大地促进深度神经网络的可解释性.

1 相关工作

本节将介绍深度神经网络的可解释性、对抗样本与对抗训练的相关工作.

1.1 深度神经网络的可解释性

近年来有很多工作研究深度神经网络的可解释性. 深度神经网络内部的神经元可以被解释为物体或者物体组成部分的检测器. 例如, 神经元响应最大化 (Activation maximization)[8] 通过找到对某个神经元产生响应最强的一组图片代表其学习的特征. 包含反卷积神经网络在内的一些基于梯度的算法[5−6] 通过将模型的响应反向传播到输入空间中, 并通过对其显著区域可视化解释模型的预测. Bau 等[15] 提出通过比较神经元学习到的特征与语义概念之间一致性以度量深度神经网络的可解释性. 此外, 模型的预测还可以通过影响力函数 (Influence function)[4] 或者解释图模型 (Explanatory graph)[16] 等方式产生解释. 此外, 还有许多方法在模型的训练过程中提升其可解释性. 例如, 从文本中提取出的

语义主题可以通过可解释性损失 (Interpretive loss)[17] 的方式加入到模型的训练中. 上下文相关可解释网络 (Contexture explanation networks)[18] 在预测的同时学习产生相应的解释. 胶囊网络 (Capsule networks)[19] 将一组神经元作为一个胶囊用于表示某个语义概念.

尽管现有很多方法展示出深度神经网络内部的神经元学习到的特征可以与人类所理解的语义概念相关联, 这些工作仅仅是利用真实数据进行分析. 本文通过对抗样本分析深度神经网络学习的特征表示, 发现了与之前的一些结论不一致的现象.

1.2　对抗样本与对抗训练

攻击者向真实样本中添加微小的、人眼不可察觉的扰动, 可以构造出使得模型发生预测错误的对抗样本[9−14]. 基于模型预测错误的不同表现形式, 对抗样本可以分为两类. 第 1 类称为无目标的对抗样本, 它们可以让模型将其错误地预测为真实类别以外的任意类别. 第 2 类称为有目标的对抗样本, 它们会被模型错分为攻击者所指定的目标类别, 以达到攻击者的特定目的. 给定一个真实样本 x, 其真实类别为 y. 同时给定一个基于深度神经网络的分类器 $f_\theta(\cdot)$, 其中 θ 代表分类器的参数. 一个对抗样本 x^* 通常会在真实样本 x 的邻域内进行寻找, 使得 x^* 看起来与 x 没有任何差别, 但是会被模型错分. 通常情况下, 一个无目标的对抗样本通过最大化网络的损失函数 $L\left(f_\theta\left(x^*\right), y\right)$产生, 可以使得 $f_\theta\left(x^*\right) \neq y$, 其中 L 是分类网络通常使用的交叉信息熵损失函数 (Cross-entropy loss). 而有目标的对抗样本通过最小化网络的损失函数 $L\left(f_\theta\left(x^*\right), y^*\right)$ 产生, 使网络将其错分为目标类别 y^*, 即 $f_\theta\left(x^*\right)=y^*$.

有很多的攻击方法可以解决上述的优化问题, 以产生对抗样本. 其中快速梯度符号法 (Fast gradient sign method, FGSM)[10] 通过一步梯度迭代产生对抗样本. 对于无目标对抗攻击, FGSM 可以表示为

$$x^*=x+\epsilon \times \operatorname{sign}\left(\nabla_x L\left(f_\theta(x), y\right)\right) \tag{1}$$

其中, ϵ 是扰动的噪声规模. FGSM 首先计算损失函数对于输入的梯度, 然后取梯度的符号将其归一化, 并乘以扰动规模 ϵ, 可以使得对抗样本与真实样本的距离满足 $\left\|x^*-x\right\|_\infty \leq \epsilon$. 基于 FGSM, 基础迭代法 (Basic iterative method, BIM)[11] 通过多步梯度迭代, 可以产生攻击效果更好的对抗样本. 基于优化的方法[12] 直接求解

$$\arg\min_{x^*}\left\{\left\|x^*-x\right\|_2^2-\lambda \times L\left(f_\theta\left(x^*\right), y\right)\right\} \tag{2}$$

其中, 第 1 项减小对抗样本与真实样本的 ℓ_2 距离, 第 2 项最大化网络的损失函数, λ 是一个超参数. 上述的几种攻击方法可以简单地扩展到有目标攻击上, 通过将式 (1) 和式 (2) 中的损失函数 $L\left(f_\theta\left(x^*\right), y\right)$ 替换成 $-L\left(f_\theta\left(x^*\right), y^*\right)$即可.

由于对抗样本对于深度神经网络所带来的安全隐患, 有很多的防御方法期望抵抗对抗样本的干扰, 得到更加鲁棒的模型. 在这些防御方法中, 对抗训练 (Adversarial training)[10, 20−23] 是一类典型且有效的算法. 对抗训练通过将对抗样本加入到训练过程中更新模型参数, 使其可以抵抗对抗样本的影响. 具体地, 对抗训练可以被定义为一个最小最大化问题

$$\min_\theta \mathbb{E}_{(x, y) \sim D}\left[\max_{x^* \in S(x)} L\left(f_\theta\left(x^*\right), y\right)\right] \tag{3}$$

其中, D 是数据分布, $S(x)$ 是所允许的对抗样本区域. 上式中的内层最大化问题通常通过攻击算法产生的对抗样本近似, 而外层最小化问题将对抗样本作为训练数据得到更加鲁棒的模型.

本文说明了对抗样本以及对抗训练对于深度神经网络可解释性分析的作用. 我们通过对抗样本分析深度神经网络的特征表示, 并利用加入特征表示一致性损失的对抗训练方式提升网络的可解释性.

2　面向对抗样本的可解释性分析

先前的工作认为深度神经网络可以学习到对于图像内容的解耦的特征表示[6, 8], 即其中的神经元会检测人类所理解的语义概念, 从而使整体的特征表示可以解释. 然而, 在本节中展示了可以检测语义概念 (例如物体或者物理组成部分) 的神经元可以很轻易地被对抗样本所欺骗, 展现出神经元学习到的特征和语义概念之间的不一致性.

为了检验深度神经网络面对对抗样本时的特征表示, 我们首先利用基于优化的对抗攻击算法产生有目标的对抗样本. 需要注意的是本节所展示的实验结果并不仅仅局限于所采用的攻击算法. 与式 (2) 类似, 通过求解以下的优化问题可以产生一个有目标的对抗样本 x^*

$$\arg\min_{x^*}\left\{\left\|x^*-x\right\|_2^2+\lambda \times L\left(f_\theta\left(x^*\right), y^*\right)\right\} \tag{4}$$

给定对抗样本 x^* 和与之对应的真实样本 x, 我们将它们输入到深度神经网络中, 并通过神经元响应最大化算法 (Activation maximization)[8] 检验其

特征表示. 该方法对于每个神经元, 找到对其产生响应最强的一组图片代表该神经元学习到的特征, 可以对其学习的特征进行可视化.

图 2 中展示了 VGG-16[24] 网络中某些神经元学习到的特征的可视化结果. 对于每个神经元, 我们选取了对其产生响应最强的 8 幅真实图片和 8 幅对抗图片代表其学习到的特征. 对于这些选取的图片, 我们利用差异图 (Discrepancy map)[8] 观测其中显著的区域, 可以更好地发现神经元检测的特征的语义概念. 从图中可以看出, 对每个神经元产生响应较强的真实图片中具有明确的语义概念. 比如图 2 (d) 中展示了对第 147 个神经元产生响应最强的 8 幅真实图片都是鸟的图片, 而高亮区域显示出鸟头的特征, 所以此神经元可以被解释为鸟头的检测器, 这个结果也与之前的研究结论一致. 这样的现象对其他的神经元和模型同样成立, 可以看出真实图片中的语义概念具有很好的一致性, 其共同的语义概念可以解释神经元学习到的特征. 然而仅仅对真实图片进行分析并不能完全了解神经元真正的行为, 通过下面的分析将会发现, 神经元并不具备检测输入图像中语义信息的能力.

从另一个方面进行分析, 可以发现对神经元产生响应较强的对抗图片中的语义信息和真实图片中的语义信息没有任何关联. 在图 2 中所展示的任何一个神经元, 对其产生响应较强的对抗图片中没有共同的特征, 也与通过真实图片发现的神经元的语义没有任何关系. 比如图 2 (d) 中的对抗图片均没有出现鸟, 其中也没有任何相似的特征. 而包含鸟的对抗图片反而没有对此神经元产生很强的响应. 这与通过真实图片得到的结论严重不符, 即该神经元对于对抗样本并没有检测其学习到的特征.

为了探究产生此现象的原因, 我们通过进一步分析神经元的行为, 发现了以下的现象: 在图 2 中由方框框起来的对抗图片会被模型误分为与对应的真实图片中语义特征类似的类别. 比如在图 2 (d) 中, 此神经元检测真实图片中鸟的概念, 而方框中的对抗图片同样被 VGG-16 模型误分为不同类别的鸟. 因此, 我们认为神经元并不具备检测图像中高层次语义信息 (物体或者物体组成部分) 的能力, 而是会对模型预测为特定类别的图片产生较强的响应, 无论图片中是否会出现与此类别相关的物体. 后续的定量实验也会进一步证明此结论. 本文中的显著区域可视化并不局限于所采用的响应最大化和差异图方法, 其他类似方法也可以使用[25−26].

3 基于特征一致性的对抗训练

基于以上的分析, 我们在本节提出了加入特征表示一致性损失的对抗训练方式 (Adversarial training with a consistent loss), 可以在训练深度神经网络的过程中提升其学习的特征表示与人类所理解的语义概念之间的一致性. 与之前的一些利用高层语义概念显式地提升模型可解释性的工作[17] 不同, 本节所提出的方法可以通过对抗训练的方式隐式地提升模型可解释性. 对抗训练通过指导深度神经网络学习到对于对抗样本和真实样本更加接近的特征表示, 去除掉噪声对于其特征表示的干扰, 从而使其内部的神经元在相关的语义概念出现时产生响应, 而在语义概念不出现时不产生响应. 这样才能更好地解释每个神经元学习到的特征.

为了达到上述的目标, 我们引入特征表示一致性损失函数, 并将其加入到对抗训练的过程中. 具体地, 所提出的方法通过优化以下的目标函数训练深度神经网络的参数 θ

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}\left[L_{\text{cls}}(\theta, x) + L_{\text{con}}(\theta, x)\right] \tag{5}$$

(a) Neuron 38: Dog (b) Neuron 20: Train (c) Neuron 102: Goat (d) Neuron 147: Bird

图 2 VGG-16 网络中神经元 (来自 conv5_3 层) 特征可视化

Fig. 2 The visualization results of the neuron (from the conv5_3 layer) features in VGG-16

其中, 第 1 项为分类损失函数, 与式 (3) 类似地定义为

$$L_{\text{cls}}(\theta, x) = \max_{x^* \in S(x)} L\left(f_\theta(x^*), y\right) \tag{6}$$

第 2 项为特征一致性损失, 定义为

$$L_{\text{con}}(\theta, x) = \max_{x^* \in S(x)} d\left(\phi_\theta(x), \phi_\theta(x^*)\right) \tag{7}$$

其中, $\varphi_\theta(x)$ 返回网络对于输入 x 的特征表示向量, d 为距离的度量函数, 用于计算对抗样本与真实样本特征表示之间的距离. 我们选取平方欧几里得距离 (Squared Euclidean distance) 作为 d.

为了求解训练目标函数 (5), 需要首先求解内层的最大化问题找到对抗样本, 进而求解外层的最小化问题以训练网络参数. 需要注意的是式 (5) 中包含两个目标函数 $L_{\text{cls}}(\theta, x)$ 与 $L_{\text{con}}(\theta, x)$, 是两个不同的最大化问题. 故而需要求解两个内层最大化问题. 为了简化训练方式, 我们仅仅求解 $L_{\text{cls}}(\theta, x)$ 找到一个对抗样本 x^*, 然后利用 x^* 作为 $L_{\text{con}}(\theta, x)$ 的一个近似解, 而不再具体求解 $L_{\text{con}}(\theta, x)$. 为了最大化 $L_{\text{cls}}(\theta, x)$, 有很多对抗攻击算法都可以使用. 在本文中我们选取 FGSM 方法产生对抗样本. 具体地, 针对式 (6) 中的最大化问题, 我们首先采用式 (1) 中的方法产生对抗样本 x^{FGSM}. 除了在对抗样本 x^{FGSM} 处优化网络交叉信息熵损失函数外, 我们还优化在真实样本 x 处的损失函数, 故而 $L_{\text{cls}}(\theta, x)$ 具体表示为

$$\begin{aligned} L_{\text{cls}}(\theta, x) = {} & \alpha \times L\left(f_\theta(x), y\right) + \\ & (1-\alpha) \times L\left(f_\theta\left(x^{\text{FGSM}}\right), y\right) \end{aligned} \tag{8}$$

其中, α 为一个超参数, 选取为 0.5. $L_{\text{con}}(\theta, x)$ 也通过产生的对抗样本 x^{FGSM} 定义特征层面的损失函数为

$$L_{\text{con}}(\theta, x) = d\left(\phi_\theta(x), \phi_\theta\left(x^{\text{FGSM}}\right)\right) \tag{9}$$

最后将式 (8) 和式 (9) 中定义的 $L_{\text{cls}}(\theta, x)$ 和 $L_{\text{con}}(\theta, x)$ 代入式 (5) 中训练网络的参数 θ. 训练算法选用通常训练深度神经网络的随机梯度下降法 (Stochastic gradient descent, SGD).

本文所提出的加入特征表示一致性损失的对抗训练方式与基于高层特征指导的去噪器防御对抗样本的方法[27] 有一定的相似性. 两个方法都用到了深度神经网络特征表示一致性的损失函数作为训练目标. 但是这两个方法在目标、训练方式、以及最终结果上存在很大的区别: 1) 我们的方法目标是提升深度神经网络的可解释性, 高层特征指导的去噪器目标是去除对抗样本中的噪声, 提升模型鲁棒性; 2) 我们利用对抗训练的方式训练模型, 而高层特征指导的去噪器通过分类器的特征表示作为损失函数训练去噪器, 而不更新分类器的参数; 3) 在实验结果中, 我们的方法可以提升深度神经网络的可解释性, 而高层特征指导的去噪器由于不更新模型的参数, 所以模型的可解释性没有任何改变.

4　实验与分析

本文在 ImageNet[28] 数据集上进行实验. 通过对抗样本分析三个不同的深度神经网络的可解释性, 并利用提出的加入特征一致性损失的对抗训练方式提升这三个模型的可解释性. 实验结果证明了通过该方法训练得到的深度神经网络的特征与人类所理解的语义概念之间的一致性更好.

4.1　实验设定

本文选取 AlexNet[29]、VGG-16[24] 和 ResNet-18[30] 三个经典的网络结构研究其可解释性. 对于正常训练的模型, 我们采用预训练好的 AlexNet 和 VGG-16 模型, 并重新训练了一个 ResNet-18 模型. ResNet-18 通过随机梯度下降进行优化, 其中超参数设置为: 动量为 0.9, 权重衰减项为 5×10^{-5}, 批大小为 100, 初始学习率为 0.05, 每当模型损失函数在一段时间内停止下降时将学习率减小 10 倍. 该模型总共训练了 30 万轮. ResNet-18 网络中加入了批归一化操作 (Batch normalization)[31]. 由于对抗训练的计算成本过高, 我们通过微调的方式在这些训练好的正常模型上进行对抗训练. 在对抗训练的过程中需要利用 FGSM 算法生成对抗样本, 我们将扰动的规模设置为 $\epsilon \in [4, 16]$. 我们将对抗训练得到的三个同样结构的模型表示为 AlexNet-Adv、VGG-16-Adv 和 ResNet-18-Adv.

在测试阶段, 我们选取两个数据集用于测试. 其中第 1 个数据集是 ImageNet 验证集, 包含 5 万幅图片. 为了使用对抗样本验证这些模型的可解释性, 我们利用 Adam 优化器求解式 (4), 其中 Adam 的步长设置为 5, 总共优化 10 ~ 20 轮. 对于数据集中的每一幅真实图片, 为了产生有目标的对抗样本, 我们将目标类别设置为模型对真实图片预测概率最小的类别, 可以使得真实类别与目标类别的差异更加明显. 需要注意的是对于每一个模型, 我们都针对该模型产生对抗样本. 第 2 个测试数据集是 Broden[15] 数据集, 提供了对于图片中语义概念的细粒度标注, 包括颜色、纹理、材质、场景、物体部分和物体等. 该数据集用于定量地衡量深度神经网络的可解释性. 对于该数据集, 我们仍然采用前述的对抗样本生成

方式. 给定这两个数据集中的真实图片和对其产生的对抗图片, 我们继而定性以及定量地研究正常训练的模型和用我们所提出的算法训练的模型的可解释性的优劣.

4.2 评估指标

本文采用两个指标衡量深度神经网络中神经元学习的特征与语义概念之间的一致性. 第 1 个指标是 Bau 等[15] 在 2017 年提出的利用 Broden 数据集计算的指标. 该指标计算神经元学习到的特征与数据集标注的语义概念之间的一致性以解释神经元的特征, 并通过与不同层次语义概念相关联的神经元的数量或比例度量深度神经网络的可解释性. 具体来说, 对于模型中的每一个神经元, 首先通过整个数据集得到该神经元产生的响应的分布, 然后通过这些响应值的上 0.5% 分位数确定一个阈值. 该阈值用于掩膜每幅图片对其产生的响应值. 然后将该神经元特征的掩膜扩展到图片大小, 得到一个分割掩膜. 最后, 该分割掩膜与数据集中包含的语义概念真实的分割掩膜计算交并比大小. 如果交并比很大, 就说明该神经元产生响应的区域与图像中包含某种语义概念的区域重合度很高, 可以认为该神经元检测此种语义概念. 虽然这个指标可以得到每个神经元检测的语义概念, 但是它需要包含各种语义概念分割结果的数据集, 而对于 ImageNet 这种仅仅包含图片分类标签的数据集并不适用. 因此我们提出了第 2 个指标用于衡量深度神经网络学习的特征与语义概念的一致性.

基于一些直观的观察, 一般情况下低层次的语义信息 (例如颜色、纹理) 会在许多不同类别的图片中出现, 而高层次的语义信息 (例如物体或者物理组成部分) 仅仅会在一些特定类别的图片中出现. 因此我们认为, 如果一个神经元对某些特定类别的图片产生较强的响应, 那么它更有可能检测高层次的语义信息; 反之如果该神经元对多种类别的图片都产生很强的响应, 那么它更有可能检测低层次的语义信息. 基于此观察, 我们通过计算对每个神经元产生响应较强的图片中的语义信息的多样性衡量该神经元检测语义信息的层次和一致性.

具体地, 对于一个神经元 n_i, 我们首先利用整个数据集计算其产生的响应, 然后用前 1% 的图片代表其学习的特征. 我们令 p_i 代表这些图片的类别分布, 即 p_i^j 代表这些图片中真实类别为 j 的图片的比例, 以指示该神经元倾向于检测的类别. 为了计算这些图片类别分别的多样性, 一个简单的方式是用 p_i 的熵作为衡量指标, 但是该指标忽视了类别之间的层次相关性. 例如, 相比于对猫和狗图片都产生响应的神经元 (其可能检测动物皮毛), 一个对不同类别的狗的图片产生响应的神经元 (其可能检测狗脸) 更有可能检测更高层次的语义概念.

为了解决上述问题, 我们通过 WordNet[22] 将各个类别之间的语义相关性进行度量. 如图 3 所示, 我们利用 WordNet 树结构中不同类别的距离作为语义相关性的度量, 具体计算方式为

$$c_{l,m} = \exp\left(-\frac{d^2(w_l, w_m)}{2\sigma^2}\right) \tag{10}$$

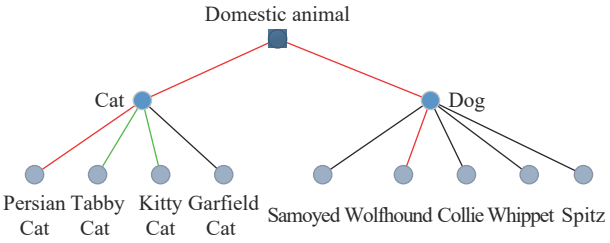


图 3 基于 WordNet[32] 衡量特征的层次与一致性示意
Fig. 3 Illustration for quantifying the level and consistency of features based on WordNet[32]

其中, w_l, w_m 是第 l 和第 m 个类别的单词, $d(w_l, w_m)$ 为它们在 WordNet 树结构中的距离. σ 为一个超参数, 设置为 1. 基于每两个类别的语义相关性 $c_{l,m}$, 我们将其组合成类别相关性矩阵 $C = [c_{l,m}]$. 从而, 将神经元 n_i 检测特征的层次和一致性定量地表示为

$$LC(n_i) = \|p_i\|_C^2 = p_i^{\mathrm{T}} C p_i \tag{11}$$

LC 值更高则代表神经元检测一个更高层次的语义概念或更关注于某些特定类别的图片.

除了真实图片外, 我们也关注对每个神经元产生响应最强的前 1% 的对抗图片. 对于每个神经元, 用 p 代表对其产生响应最强的前 1% 真实图片的类别分布, 同时用 q 代表对其产生响应最强的前 1% 对抗图片的真实类别分布, 并用 $\tilde{q}$ 代表对其产生响应最强的前 1% 对抗图片的目标类别分布. 与式 (11) 中的定义类似, 我们计算真实图片语义概念与对抗图片语义概念的相关性为

$$CS_1 = \frac{\langle p, q\rangle_C}{\|p\|_C \|q\|_C} = \frac{p^{\mathrm{T}} C q}{\sqrt{p^{\mathrm{T}} C p}\sqrt{q^{\mathrm{T}} C q}} \tag{12}$$

$$CS_2 = \frac{\langle p, \tilde{q}\rangle_C}{\|p\|_C \|\tilde{q}\|_C} = \frac{p^{\mathrm{T}} C \tilde{q}}{\sqrt{p^{\mathrm{T}} C p}\sqrt{\tilde{q}^{\mathrm{T}} C \tilde{q}}} \tag{13}$$

其中, CS_1 衡量真实图片和对抗图片中内容的相似度; CS_2 测量真实图片类别和对抗图片目标类别的相似度.

在极端情况下, $CS_1 = 1$ 意味着对抗图片的真实类别与真实图片的类别完全一致, 说明神经元对于对抗图片与真实图片均检测一致的语义概念, 则其可解释性更好. 另一方面, $CS_2 = 1$ 意味着对抗图片的目标类别与真实图片的类别完全一致, 而且内容完全不相关, 说明该神经元可解释性很差. 值得注意的是 CS_1 和 CS_2 不可能同时很高, 这是因为 q 和 $\tilde{q}$ 由于攻击的存在差异明显.

4.3　实验结果

1) 可视化结果. 我们首先采用响应最大化的方式观测深度神经网络内部神经元学习到的特征. 对于每个神经元, 在 ImageNet 验证集中找到对其产生响应最强的 8 幅真实图片和 8 幅对抗图片代表其学习到的特征. 对于三个正常模型 AlexNet, VGG-16 和 ResNet-18 的可视化结果如图 4、图 2和图 5 所示. 正如在第 2 节中所讨论的, 这些神经元并不会对对抗样本中相关的语义概念产生响应, 而仅对于模型错分为相关类别的图像产生响应, 无论这些图片中的语义概念是否相关联. 类似地, 我们展示对三个通过我们所提出的对抗训练方式得到的模型 AlexNet-Adv、VGG-16-Adv 和 ResNet-18-Adv 的可视化结果, 如图 6 ~ 8 所示. 通过对抗训练, 我们发现对抗图片和真实图片中的语义概念十分接近, 表明这些网络中的神经元更倾向于对语义概念产生响应. 这就说明了通过我们所提出的对抗训练方式, 模型内部特征的可解释性更强, 更容易被人类所理解. 后续的定量实验也进一步验证了此结论.

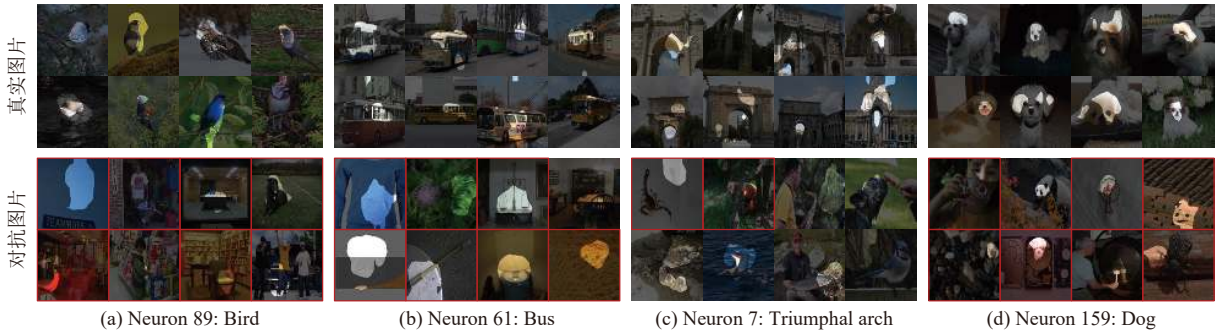
2) 定量结果. 对于 Broden 数据集, 我们计算每个模型中和各类语义概念相关的神经元的比例, 代表模型可解释性的好坏. 如果某个神经元产生响应的掩码与一个语义概念的分割掩码的交并比大于 0.4, 就认为它是一个可解释的神经元. 对于每个模型也同时计算该模型对于对抗图片的可解释神经元的比例. 表 1 中展示了实验结果, 共有 6 类不同的语义概念, 分别为: 颜色 (C)、纹理 (T)、材质 (M)、场景 (S)、物体组成部分 (P) 和物体 (O) .

从表 1 的实验结果中可以看到, 对于正常训练的模型, 和高层次语义概念相关联的神经元在面向

(a) Neuron 41: Bird　(b) Neuron 91: Dog　(c) Neuron 61: Car wheel　(d) Neuron 145: Human face

图 4　AlexNet 网络中神经元 (来自 conv5 层) 特征可视化

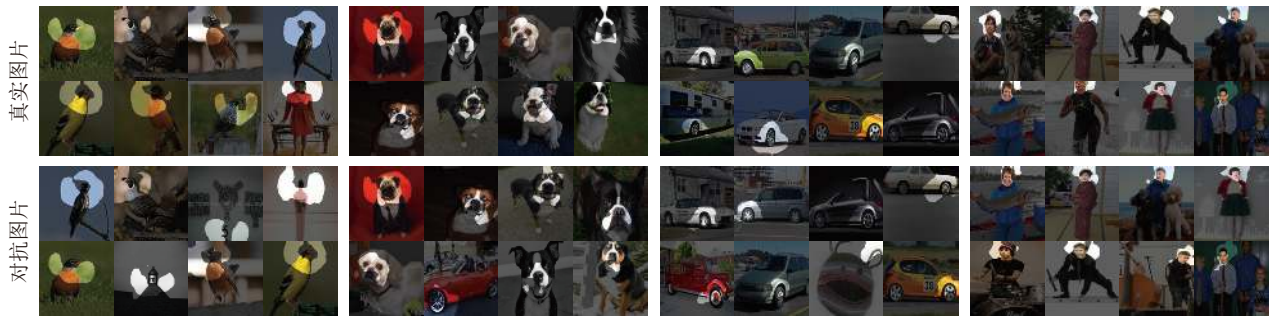
Fig. 4　The visualization results of the neuron (from the conv5 layer) features in AlexNet



(a) Neuron 89: Bird　(b) Neuron 61: Bus　(c) Neuron 7: Triumphal arch　(d) Neuron 159: Dog

图 5　ResNet-18 网络中神经元 (来自 conv5b 层) 特征可视化

Fig. 5　The visualization results of the neuron (from the conv5b layer) features in ResNet-18



(a) Neuron 68: Bird (b) Neuron 91: Dog (c) Neuron 86: Car wheel (d) Neuron 161: Human face

图 6 AlexNet-Adv 网络中神经元 (来自 conv5 层) 特征可视化

Fig. 6 The visualization results of the neuron (from the conv5 layer) features in AlexNet-Adv

(a) Neuron 38: Dog (b) Neuron 20: Rail (c) Neuron 12: Tower (d) Neuron 135: Bird

图 7 VGG-16-Adv 网络中神经元 (来自 conv5_3 层) 特征可视化

Fig. 7 The visualization results of the neuron (from the conv5_3 layer) features in VGG-16-Adv

(a) Neuron 19: Bird (b) Neuron 61: Bus (c) Neuron 207: House (d) Neuron 153: Dog

图 8 ResNet-18-Adv 网络中神经元 (来自 conv5b 层) 特征可视化

Fig. 8 The visualization results of the neuron (from the conv5b layer) features in ResNet-18-Adv

对抗样本时会大幅度下降. 这证实了这些神经元学习到的特征与语义概念之间存在不一致性. 另一方面, 通过对抗训练得到的模型即使在对抗样本存在时其中的神经元也会和语义概念有很好的关联. 这些结果也证明了利用本文提出的对抗训练方式可以提升深度神经网络特征表示与语义概念的一致性.

在表 1 中, 我们还对比了不加入特征一致性损失的对抗训练方式[20] 得到的模型, 该模型是在 Inception V3[33] 结构上训练得到, 表示为 Adv-Inc-v3. 为了展示该方法对于模型可解释性的提升, 我们进一步选取正常训练的 Inception V3 模型进行比较, 表示为 Inc-v3. 从表 1 可以看到, 采用不加入一致性损失的对抗训练方式得到的模型相比于正常训练的模型也能得到更好的可解释性, 这与之前的研究发现[34] 相符, 其原因是鲁棒的模型会更加依赖可解

表 1　各个模型面对真实图片和对抗图片时其中与语义概念关联的神经元的比例 (%)
Table 1　The ratio (%) of neurons that align with semantic concepts for each model when showing real and adversarial images respectively

模型	真实图片						对抗图片					
	C	T	M	S	P	O	C	T	M	S	P	O
AlexNet	0.5	13.4	0.4	0.4	4.1	6.1	0.5	10.3	0.1	0.0	1.6	2.3
AlexNet-Adv	0.5	12.7	0.3	0.6	5.5	7.8	0.5	11.6	0.3	0.2	4.8	6.3
VGG-16	0.6	13.2	0.4	1.3	6.8	14.7	0.5	9.5	0.0	0.0	2.3	5.2
VGG-16-Adv	0.6	13.0	0.4	1.6	8.0	16.2	0.5	11.4	0.3	0.9	6.9	14.8
ResNet-18	0.3	14.2	0.3	1.9	4.1	14.1	0.3	8.2	0.1	0.6	2.1	4.8
ResNet-18-Adv	0.3	14.0	0.3	2.1	5.3	17.2	0.3	10.8	0.3	1.5	4.7	15.3
Inc-v3	0.4	11.2	0.6	4.0	8.1	23.6	0.4	7.6	0.3	0.2	2.9	6.7
Adv-Inc-v3	0.4	10.7	0.5	4.5	8.6	25.3	0.4	8.6	0.4	2.5	5.3	15.4
VGG-16-Place	0.6	12.4	0.5	7.0	5.9	16.7	0.6	9.3	0.0	1.3	2.1	6.8

释的特征进行分类. 然而, 不加入特征表示一致性损失的对抗训练对于可解释性的提升不如加入一致性损失的对抗训练明显, 这也说明了我们所提出方法的有效性. 图 9 中也进一步对 Adv-Inc-v3 模型内部的神经元学习到的特征进行了可视化. 可以看到某些神经元展示出了比较好的可解释性, 对其产生响应较强的真实图片与对抗图片具有类似的语义概念, 可以解释该神经元学习到的特征. 但是也存在某些神经元, 如图 9(c) 所示, 对其产生响应最强的 8 幅对抗图片中包含了不属于该语义概念的图片.

此外, 我们还进一步对比了 [15] 中所发现的可解释性更好的模型 VGG-16-Place. 相对于在 ImageNet 数据集上训练的模型, 该模型是在 Place 数据集上训练得到, 具有更好的可解释性. 从表 1 可以看到, 虽然对于真实图片 VGG-16-Place 模型可解释性更好, 但是当对抗样本存在时其可解释性也会大幅下降. 说明了该网络中神经元学习到的特征与语义概念之间也存在不一致性. 也进一步说明了利用加入特征表示一致性损失的对抗训练方式提升模型可解释性的必要.

对于 ImageNet 验证集, 图 10 中展示了 CS_1 和 CS_2 随 LC 变化的曲线, 对于一个给定的 LC 值, 计算在该 LC 值附近的所有神经元的 CS_1 和 CS_2 的平均值, 其中神经元来自于所有卷积层. 对于正常训练的模型, CS_1 随着 LC 的增加而降低, 这意味着面向真实图片检测高层次语义特征的神经元对于对抗图片并没有检测相似的语义概念. 而 CS_2 随着 LC 的增加而增加, 说明了这些神经元只是对模型预测为某些类别的图片产生较强的响应, 而不考

(a) Neuron 155: Bird　(b) Neuron 20: Dog　(c) Neuron 19: Ambulance

图 9　Adv-Inc-v3 网络中神经元 (来自最后一层) 特征可视化
Fig. 9　The visualization results of the neuron (from the last layer) features in Adv-Inc-v3

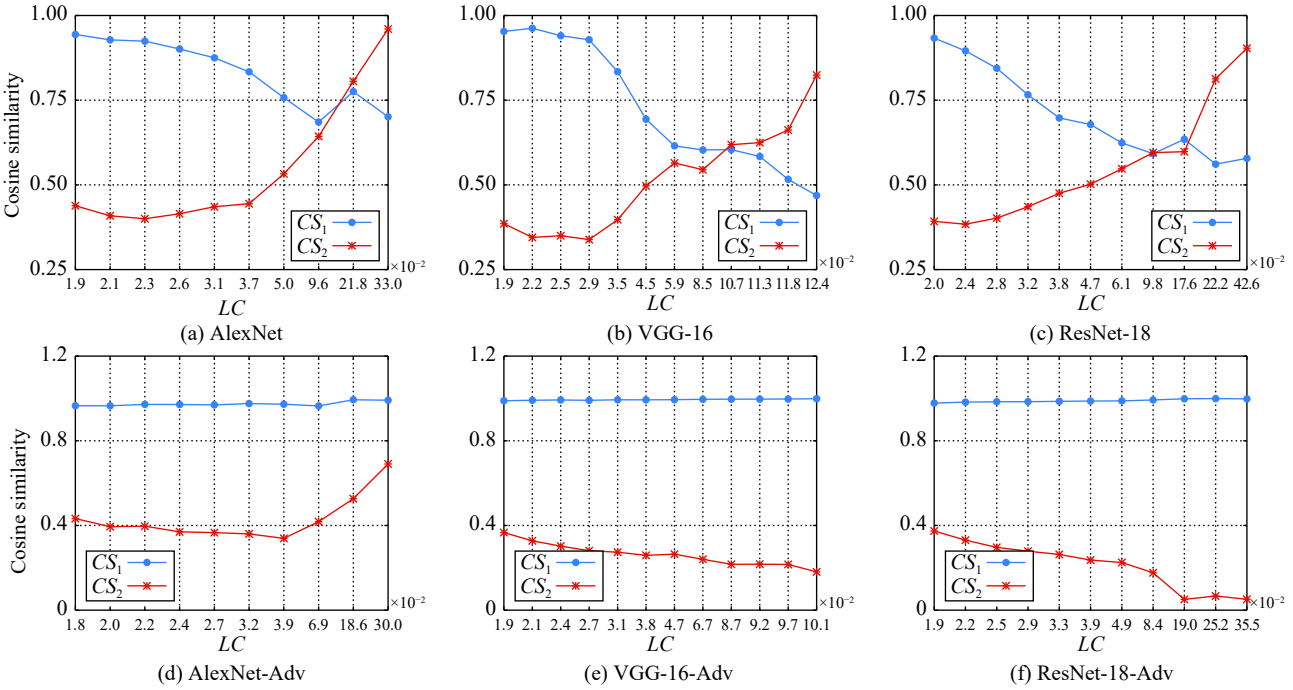


图 10 CS_1 和 CS_2 随 LC 的变化曲线

Fig. 10 The curves of CS_1 and CS_2 along with LC

表 2 各个模型在 ImageNet 验证集及对于 FGSM 攻击的准确率 (%) (扰动规模为 $\epsilon = 4$)

Table 2 Accuracy (%) on the ImageNet validation set and adversarial examples generated by FGSM with $\epsilon = 4$

模型	真实图片		对抗图片	
	Top-1	Top-5	Top-1	Top-5
AlexNet	54.53	78.17	9.04	32.77
AlexNet-Adv	49.89	74.28	21.16	49.34
VGG-16	68.20	88.33	15.13	39.82
VGG-16-Adv	64.73	86.35	47.67	71.23
ResNet-18	66.38	87.13	4.38	31.66

虑图片中是否真正存在该类别的语义概念. 另一方面对于对抗训练的模型, CS_1 保持在很高的值, 说明模型中的神经元对真实图片和对抗图片中的类似语义产生较强响应, 证明了该方法可以提升模型中神经元学习的特征表示与语义概念的一致性, 提升了模型的可解释性.

3) 模型性能. 表 2 中展示了在本文中所采用的模型在 ImageNet 验证集及对其采用 FGSM 方法产生的对抗样本的分类准确率结果. 可以看到经过对抗训练, 模型的准确率会下降 1%～5% 左右, 但是也可以提升模型对于攻击的鲁棒性. 我们认为对抗训练方式可以在模型准确率和模型可解释性以及鲁棒性之间做出平衡.

5 结论

本文利用对抗样本从模型错误的角度检验深度神经网络的的特征表示, 并发现其中神经元学习到的特征与人类所理解语义概念存在不一致性. 为了解决此问题以提升深度神经网络的可解释性, 本文提出了加入特征表示一致性的对抗训练方式. 实验结果证实了该训练方法可以有效地提升神经元学习的特征与语义概念之间的一致性, 得到可解释性更好的深度神经网络.

References

1 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436–444

2 Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(8): 1798−1828

3 Zhou B L, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 2921−2929

4 Koh P W, Liang P. Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 1885−1894

5 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proceedings of the 2014 International Conference on Learning Representations. Banff, Canada: 2014.

6 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the 2014 European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 818−833

7 Liu M C, Shi J X, Li Z, Li C X, Zhu J, Liu S X. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 2017, **23**(1): 91−100

8 Zhou B L, Khosla A, Lapedriza A, Oliva A, Torralba A. Object detectors emerge in deep scene CNNs. In: Proceedings of the 2015 International Conference on Learning Representations. San Diego, USA: 2015.

9 Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. In: Proceedings of the 2014 International Conference on Learning Representations. Banff, Canada: 2014.

10 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 2015 International Conference on Learning Representations. San Diego, USA: 2015.

11 Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv preprint arXiv: 1607.02533, 2016.

12 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy. San Jose, USA: 2017. 39−57

13 Dong Y P, Liao F Z, Pang T U, Su H, Zhu J, Hu X L, Li J G. Boosting adversarial attacks with momentum. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE 2018. 9185−9193

14 Dong Y P, Pang T U, Su H, Zhu J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE 2019. 4312−4321

15 Bau D, Zhou B L, Khosla A, Oliva A, Torralba A. Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE 2017. 3319−3327

16 Zhang Q S, Cao R M, Shi F, Wu Y N, Zhu S C. Interpreting CNN knowledge via an explanatory graph. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 4454−4463

17 Dong Y P, Su H, Zhu J, Zhang B. Improving interpretability of deep neural networks with semantic information. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE 2017. 975−983

18 Al-Shedivat M, Dubey A, Xing E P. Contextual explanation networks. arXiv preprint arXiv: 1705.10301, 2017.

19 Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules. In: Proceedings of the 2017 Advances in Neural Information Processing Systems. Long Beach, USA: Curran Associates, Inc., 2017. 3856−3866

20 Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. In: Proceedings of the 2017 International Conference on Learning Representations. Toulon, France: 2017.

21 Tramer F, Kurakin A, Papernot N, Boneh D, McDaniel P. Ensemble adversarial training: Attacks and defenses. In: Proceedings of the 2018 International Conference on Learning Representations. Vancouver, Canada: 2018.

22 Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: proceedings of the 2018 International Conference on Learning Representations. Vancouver, Canada: 2018.

23 Zhang H Y, Yu Y D, Jiao J T, Xing E P, Ghaoui L E, Jordan M I. Theoretically principled trade-off between robustness and accuracy. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 7472−7482

24 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 2015 International Conference on Learning Representations. San Diego, USA: 2015.

25 Zhang Fang, Wang Meng, Xiao Zhi-Tao, Wu Jun, Geng Lei, Tong Jun, Wang Wen. Saliency detection via full convolutional neural network and low rank sparse decomposition. *Acta Automatica Sinica*, 2019, **45**(11): 2148−2158
(张芳, 王萌, 肖志涛, 吴骏, 耿磊, 童军, 王雯. 基于全卷积神经网络与低秩稀疏分解的显著性检测. 自动化学报, 2019, **45**(11): 2148−2158)

26 Li Yang, Wang Pu, Liu Yang, Liu Guo-Jun, Wang Chun-Yu, Liu Xiao-Yan, Guo Mao-Zu. Weakly supervised real-time object detection based on saliency map. *Acta Automatica Sinica*, 2020, **46**(2): 242−255
(李阳, 王璞, 刘扬, 刘国军, 王春宇, 刘晓燕, 郭茂祖. 基于显著图的弱监督实时目标检测. 自动化学报, 2020, **46**(2): 242−255)

27 Liao F Z, Liang M, Dong Y P, Pang T Y, Zhu J, Hu X L. Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,

USA: IEEE 2018. 1778−1787

28 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z H, Karpathy A, Khosla A, Bernstein M, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, **115**(3): 211−252

29 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 2012 Advances in Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates, Inc., 2012. 1097−1105

30 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770−778

31 Liu Jian-Wei, Zhao Hui-Dan, Luo Xiong-Lin, Xu Jun. Research progress on batch normalization of deep learning and its related algorithms. *Acta Automatica Sinica*, 2020, **46**(6): 1090−1120
(刘建伟, 赵会丹, 罗雄麟, 许鋆. 深度学习批归一化及其相关算法研究进展. 自动化学报, 2020, **46**(6): 1090−1120)

32 Miller G A, Beckwith R, Fellbaum C, Gross D, Miller K J. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 1990, **3**(4): 235−244

33 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 2818−2826

34 Tsipras D, Santurkar S, Engstrom L, Turner A, Madry A. Robustness may be a odds with accuracy. In: Proceedings of the 2019 International Conference on Learning Representations. New Orleans, USA: 2019.

董胤蓬 清华大学计算机科学与技术系博士研究生. 主要研究方向为机器学习, 深度学习的可解释性与鲁棒性. E-mail: dyp17@mails.tsinghua.edu.cn
(**DONG Yin-Peng** Ph.D. candidate in the Department of Computer Science and Technology, Tsinghua University. His research interest covers interpretability and robustness of machine learning and deep learning.)

苏 航 清华大学计算机系副研究员. 主要研究方向为鲁棒、可解释人工智能基础理论及其视觉应用. E-mail: suhangss@mail.tsinghua.edu.cn
(**SU Hang** Associated researcher in the Department of Computer Science and Technology, Tsinghua University. His research interest covers theory and vision applications of the robust and interpretable artificial intelligence.)

朱 军 清华大学计算机系教授. 主要研究方向为机器学习. 本文通信作者. E-mail: dcszj@mail.tsinghua.edu.cn
(**ZHU Jun** Professor in the Department of Computer Science and Technology, Tsinghua University. His main research interest is machine learning. Corresponding author of this paper.)