计算机视觉中深度学习的对抗性攻击威胁综述

S191000862 左右飞

摘要：深度学习是当前人工智能兴起的核心。在计算机视觉领域，它已成为从无人驾驶汽车到监视和安全应用的主要力量.最近的研究表明，它们以对输入的细微扰动形式易受对抗攻击，从而导致模型预测错误的输出。对于图像，这种扰动通常太小而无法察觉，但它们却完全愚弄了深度学习模型。在实践中，对抗性攻击对深度学习的成功构成了严重威胁。本文介绍有关计算机视觉中深度学习的对抗性攻击的若干模型。

关键词：深度学习，对抗性扰动，黑盒攻击，对抗性学习

1.介绍

我们的日常生活中深度学习[1]执行具有卓越准确性的各种计算机视觉任务，Szegedy[2]等人首先发现了在图像分类中深度神经网络的一个令人震惊的弱点。他们表明，尽管具有很高的精确度，但现代深度网络却出人意料地易受对抗攻击的攻击，其形式是对人类视觉系统几乎（几乎）不可见的图像进行微扰，这种攻击可导致神经网络分类器完全改变其对图片。更糟糕的是，被攻击的模型对错误的预测表示高度信任。此外，相同的图像扰动可能使多个网络分类器蒙蔽。这些结果的深刻影响引起了研究人员对对抗性攻击及其对一般深度学习的防御的广泛兴趣。

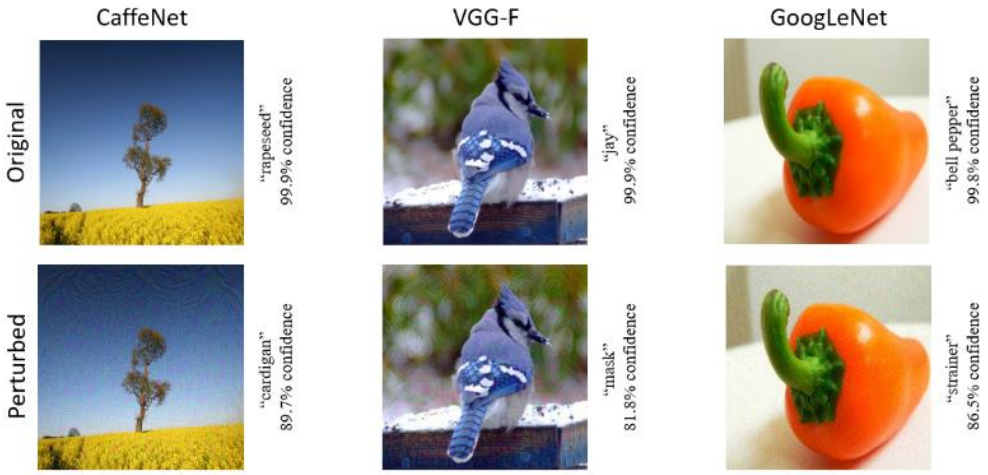
由于 Szegedy[2]等人的发现，关于计算机视觉中对深度学习的对抗性攻击，一些有趣的结果浮出水面。例如，除了图像特有的对抗性扰动，Moosavi-Dezfooli[3]等人显示了“普遍扰动”的存在，它会欺骗任何图像上的网络分类器（例如，见图 1）。同样，Athalye[4]等证明甚至可以用 3D 打印现实世界中的对象来欺骗深层神经网络分类器。

2.术语解释

对抗样本/图像： 故意干扰的干净图像的修改版本(通过添加噪声)去愚弄机器学习技术（例如深度神经网络）

黑盒攻击：将目标对抗的模型与对抗示例（在测试过程中）一起提供，这些

对抗示例是在不知道该模型的情况下生成的。在某些情况下，假设对手对模型的了解有限（例如，其训练过程和/或其架构），但绝对不了解模型参数。在其他情况下，使用有关目标模型的任何信息称为“半黑盒”攻击。



图一：带有“通用对抗性扰动”的深度学习模型的攻击示例：针对 CaffeNet，VGG-F 网络和 GoogLeNet 的攻击显示所有网络都以高可信度正确识别了原始的干净图像。在将较小的扰动添加到图像后，网络以相似的高置信度预测了错误的标签。

3. 对抗攻击

3.1 Box-constrained L-BFGS

Szegedy[2]等人首先证明了图像存在小扰动，使得被扰动的图像可能使深度 L(...)学习模型误分类。令$I_c \in \mathbb{R}^m$，表示矢量化的清晰图像-下标‘c’强调图像干净，计算扰动$\rho \in \mathbb{R}^m$ 将图像稍微扭曲以欺骗网络。Szegedy[2]等人证明了：

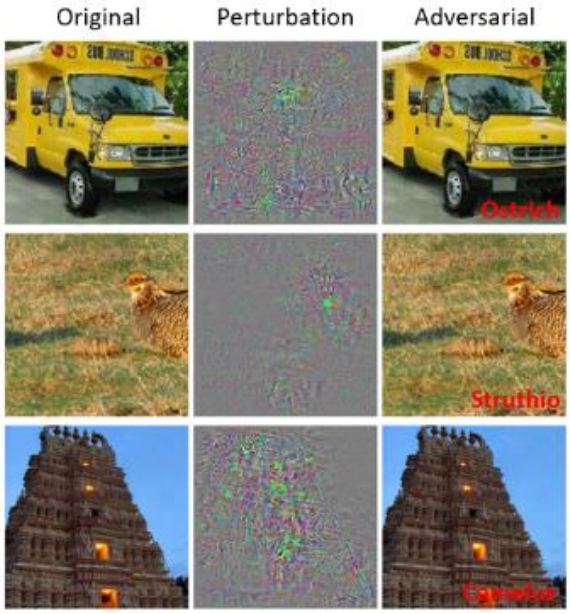
$$\min_{\boldsymbol{\rho}} ||\boldsymbol{\rho}||_2 \ \text{s.t.}\ \mathcal{C}(\mathbf{I}_c + \boldsymbol{\rho}) = \ell;\ \mathbf{I}_c + \boldsymbol{\rho} \in [0,1]^m, \qquad (1)$$

其中‘ℓ’表示神经网络分类器$\mathcal{C}(.)$的标签，作者建议当标签ℓ 和I_c不想同的时候计算以上的公式。在那种情况下，（1）成为一个难题，因此，使用盒约束 L-BFGS[5]寻求一种近似解决方案。通过找到最小值 c> 0 来完成，以下问题的最小值满足条件$\mathrm{C}(I_c + \rho) = \ell$：

$$\min_{\boldsymbol{\rho}} \ c|\boldsymbol{\rho}| + \mathcal{L}(\mathbf{I}_c + \boldsymbol{\rho}, \ell) \ \ s.t.\ \mathbf{I}_c + \boldsymbol{\rho} \in [0,1]^m, \qquad (2)$$

其中$\mathrm{L}(...)$计算了分类器的损失。注意到（2）给出了具有凸损失函数的分类器的精确解。但是，对于深度神经网络，通常不是这种情况。只需将计算出的扰

动添加到图像中，使其成为对抗性示例。



图二：由 AlexNet 使用生成的对抗示例。扰动被放大了 10 倍，以实现更好的可视化。还显示了对抗性示例的预测标签。

如图 2 所示，上述方法能够计算扰动，当将扰动添加到干净的图像中时，会使神经网络蒙蔽，但对抗性图像在人类视觉系统中看起来类似于干净的图像。Szegedy[2]等人观察到了这一点。为一个神经网络计算的扰动也能够欺骗多个网络。这些惊人的结果确定了深度学习的盲点。因此，这些矛盾的结果引起了研究人员对计算机视觉中对深度学习的对抗性攻击的广泛兴趣。

3.2 Fast Gradient Sign Method（快速梯度符号方法）

FGSM 通过对抗训练可以提高针对对抗实例的深度神经网络的鲁棒性,为了进行有效的对抗训练，Goodfellow[6]等人提出一种通过解决以下问题来有效计算给定图像的对抗性扰动的方法：

$$\boldsymbol{\rho} = \epsilon \operatorname{sign}\left(\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_c, \ell)\right), \qquad (3)$$

其中$\nabla J(...)$计算模型参数当前值附近的损失函数的梯度，sign(.)表示符号函数，它是一个小标量值，它限制了扰动的范数。用来解决(3)这个问题的方法叫做快速梯度符号方法（FGSM）。

FGSM 产生的对抗性示例利用了高维空间中深层网络模型的“线性”，而当时通常认为此类模型是高度非线性的。Goodfellow[6]等假设，现代的深层神经网络的设计（有意地）鼓励线性线性行为以获取计算收益，这也使它们容易受到廉价的分析扰动的影响。在相关文献中，这种想法通常称为“线性假设”。

Kurakin[7]等指出，在流行的大型图像识别数据集 ImageNet[8]上，对于 $x \in [2,32]$，由 FGSM 生成的对抗示例的前 1 个错误率约为 63%-69％。作者还提出了 FGSM 的“一步目标类别”变体，其中他们使用网络预测的 ℓ_{target} 而不使用（3）中图像的真实标签“目标”，而是使用网络预测的可能性最小的目标“目标”。然后从原始图像中减去计算的扰动，使其成为对抗性示例。对于具有交叉熵损失的神经网络，这样做可以最大程度地提高网络预测目标作为对抗性示例标签的可能性。建议将随机类也用作欺骗网络的目标类，但是它可能导致不太有趣的欺骗。作者还证明，对抗训练可提高深度神经网络对 FGSM 及其拟议变体产生的攻击的鲁棒性。

FGSM 干扰图像以增加分类器在所得图像上的损失。符号函数可确保最大程度地减少损失，同时 ε 本质上限制了扰动的“ ℓ_∞ 范数”。Miyato[9]等提出了一种密切相关的方法来计算摄动，如下：

$$\boldsymbol{\rho} = \epsilon \frac{\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_c, \ell)}{||\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_c, \ell)||_2}. \qquad (4)$$

在上面的等式中，计算的梯度使用 ℓ_2 范数进行归一化。 Kurakin[7]等将该技术称为“快速梯度 ℓ_2 ”方法，并提出了使用“ ℓ_∞ 范数进行归一化”的另一种方法，并将所得的技术称为“快速梯度 ℓ_∞ ”方法。从广义上讲，所有这些方法在与计算机视觉中的对抗攻击有关的文献中都被视为“一步一步”或“一次性”方法。

3.3 基本和最小类迭代方法

一步法通过在增加分类器损失的方向上采取一个较大的步长来扰动图像（即一步式梯度上升）。这个想法的直观扩展是迭代地执行多个小步骤，同时在每个步骤之后调整方向。 基本迭代方法（BIM）[10]正是这样做的，并且迭代计算出以下内容：

$$\mathbf{I}_{\boldsymbol{\rho}}^{i+1} = \text{Clip}_{\epsilon}\left\{\mathbf{I}_{\boldsymbol{\rho}}^{i} + \alpha \, \text{sign}(\nabla \mathcal{J}(\boldsymbol{\theta}, \mathbf{I}_{\boldsymbol{\rho}}^{i}, \ell)\right\}, \qquad (5)$$

其中 $\mathrm{I}_{\rho}^{\mathrm{i}}$ 表示在第 i 轮迭代中的受到扰动的图像。 $\mathrm{Clip}_{\varepsilon}\{.\}$ 在参数为ϵ 和 α情况下裁剪图像的像素以及确定步长（通常为1），BIM算法从 $\mathrm{I}_{\rho}^{0} = \mathrm{I}_{\mathrm{c}}$ 开始，并运行$\min\{\epsilon +$

4,1.25ϵ}由公式确定的迭代次数。

与将 FGSM 扩展到其“一步目标类别”变体类似，Kurakin[10]等人也将 BIM 扩展到了迭代最少似类方法（ILCM）。在这种情况下，将(5)中图像的标签‘ℓ’替换为分类器预测的可能性最小的目标标签‘ℓ_{target}’。ILCM 方法产生的对抗性示例已被证明会严重影响现代深度架构 Inception V3[10]的分类准确性，即使对于非常小的值，例如ε<16。

3.4 基于 Jacobian 的显着性图像攻击(JSMA)

更常见的是通过限制扰动的“ℓ_∞”或“ℓ_2”范式来产生对抗性示例，以使人类无法察觉。但是，Papernot[12]等还通过限制摄动的ℓ_0范数来制造对抗性攻击。从物理上讲，这意味着目标是仅仅修改图像中的几个像素，而不是干扰整个图像来欺骗分类器。该算法一次修改一个干净图像的像素，并监视更改对结果分类的影响。通过使用网络层输出的梯度计算显着图来执行监视。在此映射中，较大的值表示较高的可能性欺骗网络以预测目标作为修改图像的标签ℓ_{target}而不是原始标签ℓ。因此，该算法执行目标欺骗。计算完地图后，算法会选择最有效的像素来欺骗网络并对其进行更改。重复此过程，直到在对抗图像中更改了最大允许像素数或欺骗成功为止。

3.5 单像素攻击（One pixel attack）

对抗攻击的一种极端情况是，仅更改图像中的一个像素来欺骗分类器。有趣的是，Su[13]等声称成功改变了 70.97%的测试图像上的三种不同网络模型，方法是每幅图像仅改变一个像素。他们还报告说，网络对错误标签的平均置信度为 97.47%。图 3 中显示了这个模型中的对抗性图像的代表性示例。通过使用差分进化的概念计算了对抗性例子。对于干净的图像I_c，他们首先在$\mathbb{R}^5$中创建了 400 个向量的集合，以使每个向量都包含任意候选像素的xy坐标和 RGB 值。然后，他们随机修改向量的元素以创建 child，以使子代在下一次迭代中与 parent 竞争适应性，而将网络的概率预测标签用作适应性准则。最后一个幸存的 child 用于更改图像中的像素。

图三：单像素对抗攻击的插图。每张图片都提到了正确的标签。对应的预测标签在括号中给出。

即使采用这种简单的进化策略也能够证明成功欺骗了深层网络。差分进化使他们的方法可以生成对抗性示例，而无需访问有关网络参数值或其梯度的任何信息。他们的技术所需的唯一输入就是目标模型预测的概率标记。

3.6 Carlini and Wagner 攻击（C&W）

在对抗对抗扰动(against the adversarial perturbations)的防御性提炼之后，Carlini 和 Wagner[14]提出了三组对抗攻击。这些攻击通过限制它们的“ ℓ_2”，“ℓ_∞”和“ℓ_0”范数，使扰动变得几乎不可察觉，并且表明针对目标网络的防御性提升几乎完全无法抵抗这些攻击。此外，还显示了使用不安全（未提取）网络生成的对抗示例可以很好地转移到安全（提取）网络，这使得计算出的扰动适合于黑盒攻击。

3.7 DeepFool

Moosavi-Dezfooli[15]等人提出一个给给定的图像以迭代方式计算最小扰动的方法，称为 DeepFool。在这个算法中，DeepFool 使用干净的图像初始化，该图像假定位于分类器的决策边界所限制的区域中。该区域决定图像的类别标签。在每次迭代中，该算法都会通过一个小的向量对图像进行干扰，该向量计算是使得该图像朝着距离它最近的分类超平面沿着梯度方向移动垂线大小。一旦被扰动的图像根据网络的原始决策边界更改其标签，就将每次迭代中添加到图像的扰动累加起来，以计算最终的扰动。作者表明，DeepFool 算法能够以标准形式计算出比 FGSM 计算出的扰动小的扰动，同时具有类似的错误率。

4 总结：

本文介绍了针对计算机视觉中的深度学习的对抗性攻击方向的调查。尽管深度神经网络在各种各样的计算机视觉任务上具有很高的准确性，但发现它们容易受到细微的输入扰动的影响，从而导致它们完全改变其输出。深度学习是机器学习和人工智能当前发展的核心，这一发现导致了最近的许多贡献，这些贡献为深度学习设计了对抗性攻击及其防御措施。本文回顾了这些贡献，从回顾的文献中可以明显看出，对抗攻击实际上是对深度学习的真正威胁，尤其是对安全性和安全性至关重要的应用程序。现有文献表明，当前的深度学习不仅可以在网络空间中受到有效攻击，而且可以在物理世界中受到有效攻击。但是，由于在该研究方向上的活动非常活跃，因此可以希望，深度学习将来在对抗攻击中能够表现出相当强的鲁棒性。

参考文献：

[1] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199, 2014

[3] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard, Universal adversarial perturbations. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[4] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, Synthesizing Robust Adversarial Examples, arXiv preprint arXiv:1707.07397, 2017.

[5] R. Fletcher, Practical methods of optimization, John Wiley and Sons, 2013.

[6] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, arXiv preprint arXiv:1412.6572, 2015.

[7] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial Machine Learning at Scale, arXiv preprint arXiv:1611.01236, 2017.

[8] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255, 2009.

[9] T. Miyato, S. Maeda, M. Koyama, S. Ishii, Virtual Adversarial Training: a Regularization Method for Supervised and Semi-supervised Learning, arXiv preprint 1704.03976, 2017.

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

[11] C. Szegedy, V. Vincent, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-282, 2016.

[12] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, The Limitations of Deep Learning in Adversarial Settings, In Proceedings of IEEE European Symposium on Security and Privacy, 2016.

[13] J. Su, D. V. Vargas, S. Kouichi, One pixel attack for fooling deep neural networks, arXiv preprint arXiv:1710.08864, 2017.

[14] N. Carlini, D. Wagner, Towards Evaluating the Robustness of Neural Networks, arXiv preprint arXiv:1608.04644, 2016.

[15] S. Moosavi-Dezfooli, A. Fawzi, P. Frossard, DeepFool: a simple and accurate method to fool deep neural networks, In Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574-2582, 2016.