中文摘要

近年来，随着科学技术手段的不断进步，深度神经网络在众多任务上都展现出了优越的性能。与此同时，深度神经网络作为人工智能的重要组成部分，逐渐在人类的生产生活中被广泛使用。特别地，在图像分类任务上，神经网络的分类准确率已经可以媲美甚至超过人类。但是有研究者发现，尽管神经网络在分类任务上表现优异，它们依旧是脆弱的、不鲁棒的。

具体来说，研究者发现，对输入图像设计并添加微小的扰动就可以使神经网络的分类结果出现错误。这种使神经网络出错的技术手段被称为对抗攻击。对抗攻击的相关研究可以帮助研究者们认识和理解神经网络的弱点，从而设计相应的防御方法或者改进神经网络的训练方式来提高其鲁棒性。因此，对抗攻击研究是深度学习领域的一个重要课题。基于此，本文关注图像分类神经网络的对抗攻击方法，从对抗攻击方法的可解释性、隐蔽性和对抗样本的迁移性三个层面对该问题进行了探究。特别地，本文接下来所提到的“神经网络”，特指用于图像分类任务的神经网络。

对抗攻击的可解释性：围绕“为什么神经网络可以被攻击?”这一问题，本文从微分拓扑的角度，给出了几乎所有连续神经网络都可以被攻击的原因。具体来说，我们证明了对于绝大多数神经网络来说，神经网络的分类边界总具有拓扑性质。这意味着无论攻击者所添加的扰动有多小，在分类边界附近总存在一些样本点，它们在被添加微小扰动后，分类结果会发生改变。在证明过程中，本文发现：每一个连续的神经网络都可以在保持分类结果不变的情况下被光滑化。换句话说，存在一个全局光滑的分类器可以完全替代当前的神经网络。这一结论让我们可以使用微分拓扑来分析神经网络，并利用微分拓扑中的横截定理来说明可以被攻击的神经网络有多少。

对抗攻击的隐蔽性：当被攻击的模型（也称为目标模型）的结构和参数对于攻击者来说完全已知时，针对目标模型的攻击被称为白盒攻击。本文发现，现有的白盒攻击方法在标签空间中的隐蔽性较差。具体来说，现有方法生成的对抗样本被神经网络错误分类后得到的错误标签，往往和原始图像的真实标签相差甚远。错误标签与真实标签之间较大的差异使得人类监督者可以快速察觉到攻击的存在。本文针对这一问题，设计了 LabelFool 这一攻击方法，LabelFool 选取与真实标签相近的标签作为目标标签，并产生对抗样本使其被错误分类为所选的目标标签。人类被试的主观实验表明，与现有方法相比，LabelFool 可以显著提高对抗攻击在标签中的隐蔽性，同时在其它评价指标上，LabelFool 保持了和已有方法相当的水平。

对抗样本的迁移性：如果目标模型的信息对于攻击者来说是未知的，那么，针对目标模型的攻击被称为黑盒攻击。相比于白盒攻击，黑盒攻击在实际的应用场景中更常见，也更具有挑战性。研究者发现对抗样本在不同的模型之间具有一定的迁移能力，即，针对某个已知模型设计出的对抗样本不加改动直接输入到未知的目标模型时，依旧能使目标模型出错。可以说，对抗样本的迁移性是黑盒攻击成功的关键。为了生成迁移性能好的对抗样本，本文提出了对图像添加低频扰动的攻击算法。因为一些线索表明，神经网络主要利用图像中的低频信息进行目标函数的拟合。基于这个线索，以低频信息为主的扰动有可能在迁移过程中对目标模型产生较大的影响。在设计低频扰动时，本文使用混合高斯模型来生成扰动以保证扰动形式的多样性。实验结果表明，本文所提方法可以大幅度提高对抗样本的迁移性能。

综上所述，本文围绕图像分类神经网络的对抗攻击问题，从对抗攻击的可解释性、白盒攻击的隐蔽性和黑盒攻击中对抗样本的迁移性三个层面进行了深入的研究。

ABSTRACT

With the development of technology, deep neural networks achieve excellent performance on many tasks. Meanwhile, as an important part of artificial intelligence, deep neural networks have been widely applied in our real life. Among all, classifiers based on deep neural networks have achieved human-competitive performance on image classification tasks. However, deep neural networks are reported to be brittle in many cases.

In detail, it is found that adding imperceptible perturbations to the input image will lead to mis-classifications of deep neural networks. Such technique is called as adversarial attacks. Adversarial attacks help on understanding the vulnerability of neural networks. Moreover, people can design defense methods against attacks and improve the robustness of networks. Therefore, it is important to study adversarial attacks in the community of deep learning. In this work, we do some exploration on attacks of deep neural networks in image classification tasks. Specially, deep neural networks mentioned in the following paper refer to those in image classification tasks.

Interpretability of adversarial attacks: “Why deep neural networks can be attacked?” is an important question discussed in the community. We give an answer from a mathematical viewpoint which explains why almost all deep neural networks can be attacked. In detail, we prove that for almost all deep neural networks, their classification boundaries have the topological property. Therefore, it is easy to change classification results of samples near classification boundaries no matter how small the perturbations are. In the proof, we have an interesting finding that every deep neural network can be smoothed with inputs’ classification results unchanged. This means every deep neural network is equivalent to a smooth classifier in terms of classification. Based on the finding, differential topology can be used to analyze deep neural networks. Moreover, we use the transversality theorem in differential topology to illustrate how many deep neural networks can be attacked.

Imperceptibility of adversarial attacks: Attacks are called as white-box attacks if attackers know all information of the target model, including the structure and parameters. We find that, the wrong label of the adversarial example generated by exciting methods often has a big difference with the true label. This leads to attacks not well concealed and easy to be detected by human. In this paper, we propose an attack method, LabelFool, to solve this problem. LabelFool chooses the label similar to the ground-truth as the target label, and generates an adversarial example which is mislabeled as the target label. Subjective experiments show that LabelFool is less detectable in the label space than other attack methods. Moreover, LabelFool achieves comparable performance with state-of-the-art methods in other metrics.

Transferability of adversarial examples: Black-box attacks refer to attacks where the target model’s information is unknown, except the model’s output. Compared with white-box attacks, black-box attacks are more challenging. Researchers find that adversarial examples can transfer between models so that they use this property to design black-box attacks. Specifically, an adversarial example’s transferability means, the adversarial example is designed according to a known model (the source model), meanwhile it can attack the target model successfully without any change. The transferability of adversarial examples is the key to the success of black-box attacks. To improve the transferability, we propose to add low-frequency perturbations to inputs. Because some clues have shown that deep neural networks mainly use the low-frequency information in the image to fit the objective function. Based on these clues, we guess the target model will be greatly

affected by low-frequency perturbations. We use parameterized Gaussian Mixture Models to generate low-frequency perturbations so that the diversity of perturbations can be guaranteed. Extensive experiments demonstrate that our method significantly improves the transferability of adversarial examples.

To conclude, we make an in-depth study on adversarial attacks from three aspects: the interpretability of attacks, the imperceptibility of white-box attacks and the transferability of adversarial examples in black-box attacks.