文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

基于注意力引导的动态步长投影攻击方法

李啸宇
广州大学
DOI:10.12238/acair.v3i3.15548

[摘　要] 黑盒迁移攻击通过源模型生成对抗样本,诱导未知目标模型误分类,是提升模型安全性的重要手段。针对多步攻击易过拟合源模型、降低迁移性的难题,本文提出基于注意力引导的动态步长投影攻击方法。该方法结合攻击初期扰动的高迁移性特征,自适应调整步长以减少无效扰动,同时利用注意力机制将潜在扰动集中于模型敏感区域,提升攻击效率和跨模型迁移能力。实验表明,该方法在多个模型上均实现显著性能提升。
[关键词] 对抗样本；黑盒；迁移攻击；注意力机制
中图分类号：G255.55　**文献标识码**：A

A dynamic stride projection attack method based on attention guidance

Xiaoyu Li
Guangzhou University

[Abstract] Black-box transfer attacks generate adversarial samples from the source model to induce misclassification in unknown target models, serving as a crucial method for enhancing model security. To address the challenge of multi-step attacks easily overfitting the source model and reducing transferability, this paper proposes a dynamic step-length projection attack method based on attention guidance. This method leverages the high transferability characteristics of initial perturbations during the early stages of the attack, adaptively adjusting the step length to minimize ineffective perturbations. It also utilizes an attention mechanism to focus potential perturbations on model-sensitive areas, thereby improving attack efficiency and cross-model transfer capabilities. Experiments show that this method achieves significant performance improvements across multiple models.
[Key words] adversarial sample; black box; transfer attack; attention mechanism

引言

近年来,神经网络在图像识别[1]和自然语言处理[2]等领域取得显著进展,但也被发现易受对抗攻击威胁。对抗攻击通过向输入添加微小扰动,诱使模型产生错误预测,暴露了深度学习在安全性上的缺陷,并在自动驾驶、人脸识别等关键应用中展现出潜在风险,凸显出研究对抗攻击与防御机制的重要性。目前的攻击方法如梯度、输入变换与特征干扰策略虽有一定效果,但在黑盒场景中常因迁移性不足而受限[3]。提升对抗样本[4]的跨模型迁移能力对理解模型脆弱性与构建鲁棒系统具有重要意义。

1 基于注意力引导的动态步长投影攻击方法

1.1模型总体结构

本文提出的基于注意力引导的动态步长投影攻击方法包括两个核心模块：动态步长调整模块与注意力引导的区域投影模块。

在动态步长调整模块中,方法根据攻击过程中步长比例和损失函数的梯度变化趋势,自适应调整扰动步长。初期使用较大步长快速逼近决策边界,后期逐渐减小步长以稳定优化过程。当梯度方向稳定时,适当放大步长以加速收敛;当梯度波动剧烈时,缩小步长防止偏离最优路径。该机制提升了攻击的灵活性和稳定性。

在注意力引导的区域投影模块中,利用Grad-CAM生成的注意力热图突出模型关注区域,并据此生成区域掩码。该掩码引导将被裁剪的冗余扰动重新分配至关键区域,增强扰动的利用率和迁移性。与传统直接丢弃裁剪扰动的方法相比,该策略通过重复利用早期具有通用性的扰动,有效提升跨模型的攻击成功率。

最终,两个模块协同作用,将优化后的扰动逐步施加于原始图像,生成在多个目标模型上均具备高攻击成功率的对抗样本。后续小节将具体介绍各模块实现细节。

1.2动态步长调整策略

为充分利用攻击初期更具迁移性的扰动信息,并缓解后期扰动过拟合源模型的问题,本文提出一种动态步长调整策略。该策略根据每轮迭代的阶段和损失函数的梯度变化趋势,自适应地调整扰动步长:初期使用较大步长快速逼近决策边界,后期逐步减小步长以实现精细优化。同时结合损失变化动态调整步长大小,确保攻击过程高效收敛,提升对抗样本的攻击性与黑盒迁移能力。

首先,步长的计算基于一个衰减因子$\gamma \in (0,1)$来控制步长衰减速率,动态步长公式定义为:

$$\alpha_t = \epsilon \cdot \frac{1-\gamma}{1-\gamma^T} \cdot \gamma^{t-1} \tag{1}$$

其中,γ^{t-1} 表示 γ 的 $t-1$次方,是步长的指数衰减项,确保步长随迭代次数增加而递减,ϵ 是最大扰动,$\frac{1-\gamma}{1-\gamma^T}$是归一化系数,通过归一化调整,步长会按照每轮的衰减因子比例进行分配,确保在整个攻击过程中,步长总和为最大扰动 ϵ 。

为提升攻击稳定性和对抗样本的迁移性,本文根据损失函数的梯度变化自适应优化步长,并在梯度计算中引入多重随机扰动,以获得更稳定的梯度估计,增强攻击的泛化能力。具体来说,假设当前对抗样本为 x_t ,在计算梯度时,本文先在输入样本上施加多个高斯扰动 $\eta_i \sim \mathcal{N}(0,\sigma^2)$,然后对每个施加高斯扰动后的样本 $x_t+\eta_i$ 计算损失函数的梯度,最终通过对所有扰动版本的梯度求均值来计算一个更稳定的梯度$\nabla_x \mathcal{L}(x_t+\eta_i,y)$:

$$g_t = \frac{1}{N+1}\sum_{i=0}^{N} \nabla_x \mathcal{L}(x_t+\eta_i,y)$$

其中 N 是施加高斯扰动的采样数量。

为提高攻击效率,本文在获得稳定梯度方向后对步 α 长进行动态调整。固定步长可能导致攻击不稳定:过大易跳过最优方向,过小则收敛缓慢。为此,本文提出一种基于梯度变化率的自适应步长策略,使步长随梯度信息动态调整。具体而言,步长的调整方式如下:

$$\alpha_t^{'} = \frac{\alpha_t}{1+\mu\left(\frac{\|g_t - g_{t-1}\|_2}{E_{t-1}+\epsilon}-1\right)} \tag{1-3}$$

其中 α_t 是根据公式(3-1)计算得到的步长,μ控制调整幅度,$\|g_t - g_{t-1}\|_2$ 代表在当前攻击过程中的梯度变化情况,ϵ 是防止除零的小常数(如10^{-8})。E_{t-1}为梯度变化的指数移动平均(Exponential Moving Average,EMA),其更新规则如下:

$$E_t = \lambda \cdot E_{t-1} + (1-\lambda)\cdot \|g_t - g_{t-1}\|_2 \tag{1-4}$$

其中,λ 是平滑因子,用来控制当前数据与历史数据的权重分配,λ 越小,E_t 对近期数据的变化越敏感;λ 越大,E_t 则趋于平滑,更依赖历史数据,的取值使用以下公式进行计算:

$$\lambda = \frac{2}{T+1} \tag{1-5}$$

其中,T 代表迭代的总次数。步长调整的核心思想是,当梯度变化剧烈时,意味着当前优化方向可能仍然不稳定,此时步长应当减小,以避免对抗样本更新过快而导致攻击失败。而当梯度变化较小时,说明当前攻击方向较为稳定,此时可以适当增大步长,以提高攻击效率。

为了防止步长过大影响攻击效果,本方法对于超出步长范围的扰动会进行裁剪并保留下来,在下一阶段注意力引导的区域投影策略中进行处理,并投影到模型最敏感的区域,进一步提高对抗样本的攻击能力和迁移性。通过上述设计,动态步长调整策略能够使得每次迭代中的扰动大小自适应调整,提升其在不同目标模型中的迁移性。步长的逐步减小可以有效避免过拟合现象的出现,并在保证对抗样本有效性的同时,最大限度地减少不必要的扰动。

1.3注意力引导的区域投影策略

在该策略的第一步,利用Grad-CAM方法生成输入样本的注意力图。Grad-CAM能够通过计算卷积层的梯度信息,帮助理解哪些图像区域对模型的决策有较大贡献。具体来说,Grad-CAM通过计算类激活图并加权求和,生成一张热力图用于展示不同区域的激活程度,该过程通过以下公式计算:

$$\begin{aligned} w_k^c &= \frac{1}{z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \\ L_{Grad-CAM}^c &= ReLU\left(\sum_k w_k^c A^k\right) \end{aligned} \tag{1-6}$$

其中,A_{ij}^k 是CNN最后一个卷积层的第 k 个特征图,w_k^c 是该特征图的梯度加权因子,y^c 代表类别 c 的预测分数,Z 是归一化因子,$ReLU(\cdot)$ 确保仅保留对目标类别有正向贡献的区域。生成的Grad-CAM注意力图能够揭示模型对输入图像各部分区域的关注程度,高亮区域表示模型决策中最敏感的区域。

基于Grad-CAM生成的注意力图,本方法进一步提取出图像中对模型决策影响最大的区域,即关键区域。这些关键区域通常是热力图中激活值较大的部分,反映了模型最为敏感的区域。通过设定一个阈值 τ ,提取出对决策最为关键的部分,构成一个掩码矩阵 M 。掩码矩阵 M 用于在每轮攻击中指导被裁剪扰动的投影过程。掩码矩阵 M 的计算方式为:

$$M(i,j) = \begin{cases} 1, & GradCAM(i,j) > \tau \\ 0, & GradCAM(i,j) \le \tau \end{cases} \tag{1-7}$$

该掩码矩阵表示了图像中每个像素是否属于关键区域。对于每个像素位置(i,j),如果其对应的Grad-CAM值大于阈值 τ ,则该位置被认为是重要区域,掩码矩阵中对应位置为1;否则,掩码为0。

在每轮攻击迭代中,本文利用掩码矩阵 M 引导扰动施加过程,使扰动不仅受损失函数影响,还根据注意力权重动态调整。在动态投影阶段,将上一轮被裁剪的多余扰动与当前扰动结合,增

强关键区域的攻击强度。对于掩码中值为1的区域,采用梯度投影将历史扰动与当前扰动融合,施加强扰动；而在非关键区域(值为0)仅施加较小扰动,从而提升整体攻击效果与迁移性。具体来说,本方法会利用上一轮迭代中裁剪的多余扰动 δ_{excess},动态地投影到最敏感的区域,并与当前的扰动一起更新。公式如下：

$$\delta_t(x) = Clip_{\epsilon}\left(\alpha_t^{'} \cdot sign(g_t) + \Pi_M(\beta \cdot \delta_{excess}(x))\right) \quad (1-8)$$

其中,$\delta_t(x)$ 是第 t 轮的扰动,$\alpha_t^{'}$ 是根据公式(3-3)计算得到的当前步长,g_t 是根据公式(3-2)计算的梯度信息,M 是掩码矩阵,用来表示注意力图中关键区域的标记,β 是动态调整的因子,用于控制上一轮裁剪扰动的影响,$\delta_{excess}(x)$ 是上一轮迭代中裁剪出的多余扰动。这里,上一轮的多余扰动 $\delta_{excess}(x)$ 会根据 β 的值和掩码矩阵 M 进行调整,确保多余扰动只施加在对模型决策最为敏感的区域,而不对非关键区域产生影响。Π_M表示将多余扰动投影到约束掩码矩阵 M 内。$Clip_{\epsilon}$ 是对扰动进行裁剪的操作,确保扰动的每个像素值不超过最大扰动的范围,同时暂存被裁剪下的多余扰动,用于下一次迭代中的计算。

表3-1 攻击正常训练模型的成功率(%),上标“*”代表是白盒攻击

模型	攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-152	Res-50	Res-101	AVG
Inc-v3	MI	100.0*	51.1	46.9	39.3	46.6	41.6	54.3
	S^2I	100.0*	64.8	59.6	48.9	57.0	52.4	63.8
	VMI	100.0*	74.5	70.7	63.3	68.0	62.5	73.2
	Admix	100.0*	78.6	73.2	68.0	74.3	69.2	77.2
	PI	100.0*	52.1	34.7	38.5	44.2	40.9	51.7
	SI-NI	100.0*	76.3	75.1	67.6	73.0	69.8	77.0
	GE-AdvGAN	100.0*	90.9	75.1	88.1	74.0	69.9	83.0
	Ours	100.0*	81.4	79.6	75.1	74.6	70.2	80.2
Inc-v4	MI	61.6	100.0*	45.3	42.4	45.2	42.8	56.2
	S^2I	71.9	100.0*	55.6	49.4	55.6	48.8	63.6
	VMI	83.0	100.0*	76.1	68.8	71.4	68.2	77.9
	Admix	88.4	100.0*	82.9	77.8	79.6	75.9	84.1
	PI	52.6	100.0*	30.7	36.8	40.7	37.0	49.6
	SI-NI	85.5	99.9*	79.1	72.8	75.2	73.0	80.9
	GE-AdvGAN	88.4	100.0*	69.1	81.4	81.4	80.3	83.4
	Ours	90.1	100.0*	85.6	82.1	82.5	81.1	86.9
IncRes-v2	MI	61.4	53.9	99.3*	45.4	50.2	45.0	59.2
	S^2I	75.9	66.8	98.3*	55.5	61.9	57.9	69.4
	VMI	80.7	76.4	99.3*	65.7	69.3	68.2	76.6
	Admix	90.9	88.8	99.5*	83.4	84.5	84.2	88.6
	PI	53.8	48.1	98.6*	38.1	40.7	39.1	53.1
	SI-NI	87.8	83.1	99.9*	75.7	78.5	77.1	83.7
	GE-AdvGAN	87.4	83.4	98.9*	80.3	79.2	79.0	84.7
	Ours	91.9	90.0	99.9*	86.6	85.4	85.9	89.9

2 实验结果与分析

为了验证本文的方法攻击正常训练模型的性能,分别选择Inc-v3、Inc-v4和IncRes-v2作为白盒模型来生成对抗样本。然后,本文使用这些对抗样本来攻击相同的模型以及两个额外的黑盒模型：Res-50和Res-101。如表3-1所示,对比目前主流的七种对抗攻击方法,本文的方法在不降低白盒模型攻击成功率的基础上,黑盒迁移攻击能力有显著提升。例如,在使用Inc-v3作为白盒模型生成对抗样本,并攻击其他黑盒模型时,本文的方法对比经典对抗攻击算法PI-FGSM,将平均攻击成功率从 51.7% 提升到80.2%。对比先进的方法Admix和VMI-FGSM,也将平均攻击成功率分别提升了3%和7%。而相较于目前较新的对抗攻击方法GE-AdvGAN,本文所提出的方法在采用 Inc-v4和IncRes-v2作为白盒模型的设置下,均表现出更高的平均攻击成功率。

3 全文总结

随着深度神经网络的广泛应用,其在面对对抗样本时的脆弱性引发了广泛关注。现有攻击方法在提升攻击成功率的同时,往往难以兼顾迁移性,限制了黑盒攻击效果。为此,本文从梯度优化角度出发,提出基于注意力引导的动态步长投影攻击方法。该方法通过动态步长策略加速前期扰动逼近决策边界,减少后期过拟合,并利用注意力机制将高迁移性的扰动集中投影于关键区域。实验结果表明,该方法在提高迁移性、攻击成功率及隐蔽性方面具有显著优势。

[参考文献]

[1]He K,Zhang X, Ren S,et al.Deep residual learning for image recognition[C].Proceedings of the IEEE conference on computer vision and pattern recognition,2016:770-778.

[2]Joshi A,Dabre R,Kanojia D,et al.Natural language processing for dialects of a language: A survey[J].ACM Computing Surveys,2025,57(6):1-37.

[3]Liu Y,Chen X,Liu C,et al.Delving into transferable adversarial examples and black-box attacks[J].arXiv preprint arXiv:1611.02770,2016.

[4]Szegedy C,Zaremba W,Sutskever I,et al.Intriguing properties of neural networks[J].arXiv preprint arXiv:1312.6199,2013.

作者简介：

李啸宇(1998--),男,汉族,硕士研究生,广州大学。