arXiv:2303.14511v1 [hep-ex] 25 Mar 2023

通过对抗训练提升喷注标记算法的鲁棒性：探索损失曲面

Annika Stein第三物理研究所 A，亚琛工业大学，德国 电子邮箱：annika.stein@cern.ch

摘要。 在高能物理领域，深度学习算法日益重要，并且相较于传统方法能够提升性能，例如识别稀有信号或发现复杂模式。从分析人员的角度来看，尽可能提高性能固然可取，但近期研究关注点已逐渐转向模型的鲁棒性，以考察模型在输入特征发生轻微畸变时的表现。对于涉及大量（低层级）输入的任务，深度神经网络的应用尤其带来了新的挑战。在喷注味标记领域，对抗攻击被用于探测典型分类器的脆弱性，也可视为系统不确定性的模型。相应的防御策略——对抗训练——能够在保持高性能的同时提升鲁棒性。研究相关输入和模型所对应的损失曲面，可以揭示鲁棒性的几何解释，并将相关性纳入考量。

1. 引言

借助强大的机器学习（尤其是深度学习）算法，新的物理分析得以开展，而已有分析也报告称，相较于此前仅采用基于切选的策略、浅层网络或提升决策树等技术的迭代版本，结果有所改善 [1]。物体识别是欧洲核子研究中心大型强子对撞机实验中各类分析的关键组成部分，因此，开发高性能算法尤为重要。这些算法将许多特征输入复杂架构，以尽可能捕获更多信息，包括可观测量之间的相关性。

深度神经网络适用于执行上述困难任务，例如喷注味识别；许多与喷注组分相关的低层级特征也被纳入最先进的标记器中 [2, 3, 4, 5, 6]。然而，高性能也意味着高度依赖相关输入特征的建模，尤其是因为监督式机器学习技术使用带标签的模拟样本 [7]。这些样本可能无法涵盖所有探测器效应；对于不同的蒙特卡罗生成器，在部分子簇射和强子化等环节上的模拟也可能存在相当大的差异 [8]。因此，校准一直是改善算法训练域（模拟）与测量数据之间一致性的必要步骤 [9, 10]。即便经过严格审查并使用一组独立的控制区，校准后仍可能存在一定程度的不一致；对于那些所推导的比例因子会在高喷注多重数（b/c 标记）末态中因子化的分析，这种不一致日益重要。

1.1. 相关工作

对低层级特征的依赖是高能物理多个领域的常见特性，因此，这些领域也可能容易受到轻微特征扰动的影响，从而导致性能大幅下降；这种现象通常称为对抗攻击，并会产生对抗样本 [7, 11, 12, 13]。参考文献 [13] 探讨的基本原理是，若通过对抗训练谨慎防御对抗攻击，就能将这种固有脆弱性转化为鲁棒性。研究表明，在提升鲁棒性的同时可以不损失性能 [13]。被广泛采用的技术是快速梯度符号法（FGSM），这是一种一阶攻击 [12, 13]。在针对简单多层感知机架构的原理验证得到提出之后 [13], ，CMS合作组展示了将对抗训练成功应用于新一代标记算法，并将这一策略扩展到卷积层（随后扩展至循环层和全连接层） [14]。参考文献 [14] 的一项核心发现是，对抗鲁棒性与数据和模拟之间的一致性相关，这一点在轻味喷注中尤为显著。文献也提到了互补观点，认为对抗训练无法涵盖不确定性 [15, 16]。显然，理论导致的建模偏差或依赖生成器的建模很难通过对抗方法处理 [16], ；不过，我们打算聚焦于那些可通过系统正则化缓解的建模偏差，而这或许是参考文献 [14]所示有前景的控制区发挥主要作用之处。应用对抗训练防御 FGSM攻击时还需考虑另一项局限：输入被扰动的方向在一定程度上是可预测的；此外，该方法总是依据梯度的相应符号独立处理所有特征，从而消除了特征之间的完整相关性，只留下离散选择 [13]。为了实现既具备高性能、又具有泛化能力的鲁棒算法，人们会考察底层损失曲面的平坦度，并将其作为衡量模型是否具备上述理想性质的替代指标。若干方法利用损失 [17]关于模型参数（权重和偏置项）的几何性质，但针对输入分布的研究尚未开展。

1.1.1. 对抗攻击与系统不确定性

尽管为使输入的人工偏移在一定程度上符合通常观测到的失配情况，已经施加了若干限制，但这类对抗方法仍依赖于网络自身的特性。这种情形并不符合物理实际，因为无论自然界还是过程模拟，都不可能知晓用于标记事例中喷注的机器学习算法。因此，评估网络抵御对抗攻击的能力，可能会偏向于那些明确减轻特定攻击影响的防御策略。模拟中的任何失配都只会沿着损失函数 [7, 13]上升最快的最坏方向使输入发生偏移，这种情况并不现实。对于物理分析而言，首要目标并非使用能够抵御对抗攻击的算法，而是使用能够从模拟泛化到数据、并对系统不确定性具有鲁棒性的算法。因此，文献 [13] 中研究的两种训练（名义训练和对抗训练），不仅在名义输入和对抗输入下进行了比较，也在面对沿向上或向下方向变化的系统性畸变输入时进行了比较 [13]。无论是向上还是向下变化，对抗训练在畸变输入上的表现都优于名义训练在相同畸变样本上的表现 [13]。将系统变化替换为随机涂抹／高斯噪声时，也能得出类似结论 [13]。本文旨在通过研究输入特征空间中潜在的损失函数，进一步拓展这些发现，并提出一种能够提升算法韧性的改进训练策略。

2. 在名义样本或对抗样本上训练的喷注标记算法的损失流形性质

对损失流形几何形状不同这一假设的动机，来自按味道划分对抗攻击影响时的观察：对抗训练表现出一定的对称性，而针对名义训练实施的对抗攻击则倾向于将输入推向特定方向，从而颠倒预期的物理学结论 [13]。文献 [13] 中的示意图为不同训练策略下损失曲面的形状提供了一些线索，但如何对这类曲面进行实际扫描仍是一个尚待解决的问题。图1展示了关于输入变化的几何形状可视化的初步结果。其构造方法是，首先从未用于训练或验证的样本中随机选取一个随机喷注。为便于可视化，选取两个可观测量（采用已充分理解的全局喷注特征），并围绕原始名义特征，通过均匀对称分箱生成一个由 500×500 种变化构成的网格。考虑相应特征的完整分布后，所覆盖的范围对应于 $\pm 0.5\sigma$；这是通过仅在标准化后的输入特征空间中进行操作来确保的。目标保持不变，同时对名义训练和对抗训练在生成的250000个样本上重新评估，并重新计算所得损失。在不改变横向.....的情况下改变喷注的赝快度

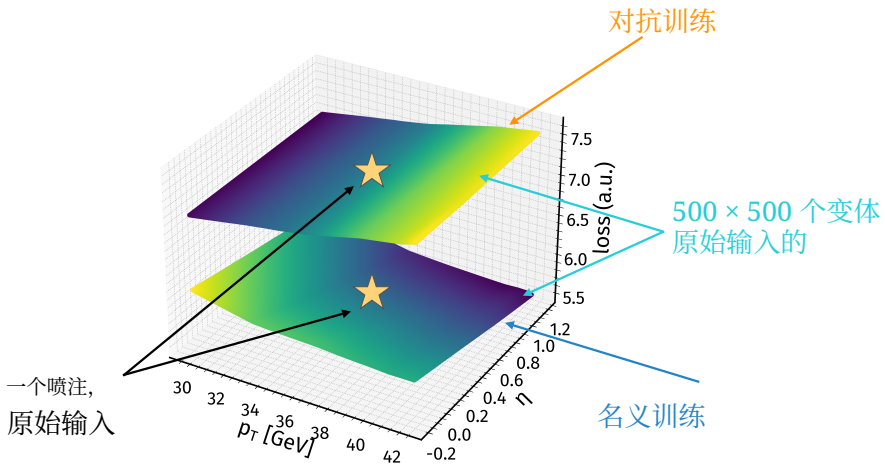


图 1. 名义训练（下）和对抗训练（上）对应的损失流形具有不同的几何形态。

动量或其他属性不会影响对抗训练中的相应网络预测误差或损失。另一方面，名义训练并不对图中所示任一变量的变化免疫。总体而言，名义训练的网络预测误差较低，而对抗训练则能形成更平坦的流形，对特定特征的畸变具有一定程度的不变性。图2显示，对于名义训练，对抗攻击会找到明确的方向；而对于对抗训练，由于其对赝快度具有不变性或对称性，因此存在多个可使损失增加的方向。尽管如此，由于攻击机制本身会对梯度取符号，攻击最终只会选择一个特定方向，尽管其他方向本质上也会产生相同效果。

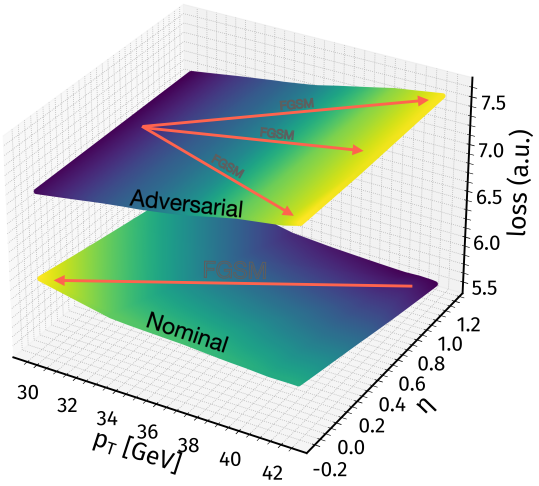


图2. 不同模型中对抗攻击的可能方向。从损失较小的运动学量出发，对于施加在对抗训练中的 FGSM攻击，可以找到多个箭头；而对于名义训练，只构造出一个这样的箭头。

3. 讨论

这一观察是理解为何在输入可能发生畸变（由实验效应、精度和分辨率限制所致），或识别算法的训练域（模拟）与基于实际记录的探测器数据构建的域之间存在其他系统差异的情况下，可能更倾向于采用对抗训练的关键。这可以解读为对抗训练所诱导的正则化迹象。

3.1. 在训练过程中利用损失流形的性质

我们在宏观尺度上探查了损失流形，考察了两个可能不具备最高判别能力的特征（并将其重新加权至共同的目标分布，以确保无偏预测） [13], ；我们建议更系统地探索这一技术，并考虑将其纳入训练过程本身。将损失曲面表示为两个输入特征的函数，是一种简化处理，便于我们以图形方式研究其几何形态。为克服这一局限，需要在特征空间中构建更多维度的损失流形。随后，可以在训练过程中引入对原始输入附近平坦度的测量，作为独立的交叉检验，以探查并提升鲁棒性。例如，可以在损失函数中加入一个汇总量，用来捕捉在允许的 $B_{\frac{\sigma}{2}}$ 球内移动输入时，计算所得交叉熵损失所观察到的最大相对影响。该汇总量可由一个超参数加权，以控制正则化相较于常规性能指标所占的比重；随后，训练便可在反向传播过程中遵循这一经修改的损失函数来更新模型参数。

3.2. 构建其他能够保留损失函数梯度方向性的攻击

从观察到的损失曲面来看，继续聚焦于一阶攻击似乎已足够；不过，若要考虑梯度的实际方向以及各特征的相对贡献，取梯度的符号（FGSM）可能效率过低。改为使用梯度的 p范数，例如 $p = 2$ ，便可以定量保留各个输入特征的贡献。随后，可将所得扰动向量按上述范数的倒数进行缩放，从而比较不同喷注样本，同时仅产生较小的扰动。这种攻击的偏移方向和各特征上的偏移幅度都不易预测；而 FGSM 则只能产生$\pm\epsilon$ 偏移。引入这种改进后的攻击，还能纳入相关性

特征；这是在高能物理语境下通常会提及的 FGSM攻击的一个缺陷。在针对这种新攻击进行对抗训练时，我们无需施加较大的畸变；由此产生的畸变喷注样本不易通过验证方法（例如一维直方图）检测出来。

4. 结论

本文研究了典型喷注标记算法在使用名义样本或对抗样本训练时，其损失流形相对于输入特征的表现。我们观察到平坦度方面的差异，因此对微小畸变的不变性也有所不同，这解释并证实了此前探究的鲁棒性和泛化能力差异。基于这些损失曲面，我们提出了改进的训练策略，在反向传播过程中直接利用新近获得的网络属性知识。所提方法更加注重正则化和相关性，有望弥合机器学习理论研究与其在粒子物理物体识别中的应用之间的差距，同时保持可观测量的物理行为。

致谢

本研究使用的模拟由亚琛工业大学通过项目 rwth1244 提供的计算资源完成。本研究获得了德国研究联合会（DFG，项目 SCHM 2796/5 和 GRK 2497）以及德国联邦教育与研究部（BMBF，项目 05H2021）的支持。

参考文献

[1] Albertsson K *et al.* 2018 *J. Phys. Conf. Ser.* **1085** 022008 (*Preprint* 1807.02876) 网址 https://doi.org/10.1088/1742-6596/1085/2/022008[2] CMS合作组 2018 *JINST* **13** P05011 (*Preprint* 1712.07158) 网址 https://doi.org/10.1088/ 1748-0221/13/05/P05011[3] Bols E, Kieseler J, Verzetti M, Stoye M 和 Stakia A 2020 *JINST* **15** P12012 (*Preprint* 2008.10519) 网址 https://doi.org/10.1088/1748-0221/15/12/P12012[4] Qu H 和 Gouskos L 2020 *Phys. Rev. D* **101** 056019 (*Preprint* 1902.08570) 网址 https://doi.org/10.1103/PhysRevD.101.056019[5] Qu H, Li C 和 Qian S 2022 用于喷注标记的粒子Transformer (*Preprint* 2202.03772)[6] ATLAS合作组 2022 大型强子对撞机运行期的 ATLAS 味标记算法 2 pp 碰撞数据集 (*Preprint* 2211.16345)[7] Nachman B 和 Shimmin C 2019 高能物理学的人工智能安全 (*Preprint* 1910.08606)[8] ATLAS合作组 2020 用于味标记效率校准的蒙特卡罗到蒙特卡罗比例因子 (*Report* ATL-PHYS-PUB-2020-009) 网址 https://cds.cern.ch/record/2718610[9] ATLAS合作组 2018 基于 ATLAS 质子-质子碰撞数据的轻味喷注 b-标记率校准，碰撞能量为 $\sqrt{s} = 13$ 太电子伏特 (*Report* ATLAS-CONF-2018-006) 网址 https://cds.cern.ch/record/2314418[10] CMS合作组 2022 *JINST* **17** P03014 (*Preprint* 2111.03027) 网址 https://doi.org/10.1088/ 1748-0221/17/03/P03014[11] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I 和 Fergus R 2014 神经网络的引人注目的特性 (*Preprint* 1312.6199)[12] Goodfellow I J, Shlens J 和 Szegedy C 2015 解释并利用对抗样本 (*Preprint*1412.6572)[13] Stein A, Coubez X, Mondal S, Novak A 和 Schmidt A 2022 *Comput. Softw. Big Sci.* **6** 15 (*Preprint*2203.13890) 网址 https://doi.org/10.1007/s41781-022-00087-1[14] CMS合作组 2022 CMS中用于b标记算法的对抗训练 (*Report* CMS DP-2022/049) 网址 https://cds.cern.ch/record/2839919/[15] Ghosh A, Nachman B 和 Whiteson D 2021 *Phys. Rev. D* **104** 056026 (*Preprint* 2105.08742) 网址 https://doi.org/10.1103/PhysRevD.104.056026[16] Butter A, Dillon B M, Plehn T 和 Vogel L 2022 现代夸克-胶子标记中的性能与韧性 (*Preprint* 2212.10493)[17] Li H, Xu Z, Taylor G, Studer C 和 Goldstein T 2018 神经网络损失景观的可视化 NIPS 2018. (*Preprint* 1712.09913)