arXiv:2512.14170v1 [cs.LG] 16 Dec 2025

通过形式化验证改进深度主动学习

Jonathan Spiegelman[1],Guy Amir[2],和 Guy Katz[1]

[1] 耶路撒冷希伯来大学{yhonatan.spigelman, g.katz}@mail.huji.ac.il[2] 康奈尔大学 gda42@cornell.edu

摘要。 深度主动学习（DAL）旨在优先选择信息量最大的未标注样本进行标注，从而降低神经网络训练中的标注成本。除了选择要标注的样本外，若干深度主动学习方法还通过使用无需额外人工标注的合成输入扩充训练集，进一步提高数据效率。在这项工作中，我们研究了通过违反鲁棒性约束的对抗输入扩充训练数据，如何提升深度主动学习的性能。我们表明，通过形式化验证生成的对抗样本，其贡献远大于标准的基于梯度的攻击所生成的对抗样本。我们将这一扩展应用于多种现代深度主动学习技术以及我们提出的一种新技术，并表明它能显著提升模型在标准基准数据集上的泛化能力。

1 引言

深度神经网络（DNNs）在广泛的应用领域取得了显著成功 [12,26,42]。然而，在监督学习场景中训练这些模型，通常需要大量带标签数据 [19,44]。获取标签的成本可能很高，例如需要具备高技能的人类专家时。深度主动学习（DAL）旨在通过在有限的标注预算下标注信息量最大的样本，尽可能降低这项开支 [40]。一般而言，样本选择可以由模型的不确定性驱动——优先选择模型最不确定的样本；也可以考虑多样性，选择更能代表整体数据分布的样本 [38]。

与此同时，形式化方法社区开发了验证工具，这些工具接收一个网络和一项规范，要么正式证明该规范成立，要么返回违反该规范的反例（输入）。例如，验证器可以证明网络对给定输入附近的扰动具有鲁棒性，也可以生成导致错误的扰动 [17,20,23,30]。不同于一次性的基于梯度的攻击，验证器可以系统地探索一个有界输入区域，以找到多个、不同的 对抗样本（或证明不存在此类样本）——而这样一组多样化的样本对深度主动学习大有裨益，因为在选出信息量大的样本后，它们能够扩充并丰富训练集。

本文探讨将形式化验证生成的对抗输入作为训练集增强的一种方式。具体而言，对于每个新标注的点，我们使用验证器探索其有界半径内的邻域，从而找到若干对抗输入。这些输入被赋予预言机对原始样本给出的标签，并在不增加标注成本的情况下加入训练集。正如我们所展示的，这种补充有助于学习过程更快地收敛到更优结果。

为评估我们的方法，我们在多种标准基准数据集（MNIST、Fashion-MNIST、CIFAR-10）以及多种深度主动学习方法上进行了测试，涵盖从轻量级随机采样到更复杂的方法，其中包括我们在此提出的一种新方法。对于这些方法，我们均在完成样本点选择后加入了验证阶段。总体而言，我们发现这一额外步骤带来了显著改进。据我们所知，我们是首个将通过形式化验证生成的对抗样本纳入主动学习循环的工作，从而实现系统化的数据增强，充分发挥每个人工标注的作用。

2 背景

深度神经网络通过组合仿射变换与非线性激活（例如 ReLU）来学习高容量表示 [29]。尽管深度神经网络取得了成功，但已有研究表明，微小且往往难以察觉的输入扰动也能显著改变网络的预测 [45], ，这揭示了持续存在的鲁棒性差距。FGSM [18] 和 PGD [32] 等基于梯度的攻击，通过寻找会导致分类错误的微小扰动，生动地展现了这种脆弱性。这些攻击已成为探测模型鲁棒性的标准工具。

形式化验证从另一个角度解决鲁棒性问题。给定一个网络和一项形式化规约（例如，“输入类别$\mathbf{x}$ 在 ℓ_p-球内保持$\mathbf{y}$”），验证器会证明该性质成立，或生成违反该性质的反例输入 [9, 14, 17, 20]。与单步攻击相比，验证能够系统地探索输入的有界邻域，从而更全面地覆盖潜在的对抗样本，尽管计算成本更高 [22]。

深度主动学习（DAL）旨在通过仅查询未标注数据池中信息量最大的样本来减少标注工作量 [40]。典型的深度主动学习流程从在一个小型初始已标注集上训练模型开始 $\mathcal{L}$，该集合通常是从未标注数据集中随机抽取的子集 $\mathcal{U}$。随后进入迭代过程：每一轮中，模型根据某项标准评估剩余的未标注实例，选出得分最高的一批实例，并将其提交给预言机进行标注。新标注的样本随后会被加入到 $\mathcal{L}$，并重新训练模型。该循环会持续进行，直至标注预算耗尽。

深度主动学习的查询策略通常沿着不确定性—多样性谱进行分类 [38]。不确定性采样针对模型最不确定的点（例如，置信度最低 [10] 或熵最高 [41]的点），而基于多样性的方法旨在高效覆盖数据分布（例如，CoreSet [39]）。混合

方法（如 BADGE [4], ）结合了这两项标准，选择一批样本，以同时提升多样性并降低不确定性。除了查询策略外，一些深度主动学习方法还采用数据增强——即向训练集中添加合成输入——以在不增加标注成本的情况下提升模型泛化能力 [8, 35], ，具体方法包括使用生成对抗网络 [37]。一种颇具吸引力的方法是利用对抗样本：围绕每个带标签样本，可以生成经过最小扰动、会被当前模型误分类的变体，并将它们以与原始样本相同的标签加入训练集 [31]。

一种有原则的选择准则，尤其适用于分类任务，是查询模型决策边界附近的样本——即具有最小间隔 [5, 46]的样本。在深度主动学习的分类体系中，基于间隔的选择被视为一种不确定性驱动策略：间隔较小意味着模型对样本的标签难以确定。由于在深度神经网络中精确计算间隔难以实现，因此最小对抗扰动的大小可作为这一数值的实用近似。举例来说，DFAL [13], 使用 DeepFool 攻击 [36] 估计每个样本的扰动间隔，并选择间隔最小的候选样本；随后，它将每个样本对应的对抗样本加入训练集，并赋予其与原样本相同的标签。如此一来，DFAL 将基于间隔的选择与数据增强结合起来。尽管效果显著，DFAL 每次查询仅使用一个对抗样本，因此对每个样本局部决策边界的探索有限。

3 我们的方法：通过验证进行增强

我们提出利用形式化验证引擎，为每个查询输入生成一组多个、多样化的对抗样本，从而提升每次预言机查询的价值。具体而言，我们使用验证器系统地探测每个选定样本周围的模型决策边界，并找出其他能使模型预测翻转为不同类别的对抗输入。

形式化地，给定输入 $\mathbf{x}$ 及其当前模型标签 $\hat{y} = \arg\max\ f(\mathbf{x})$，我们要检查的规范是：是否存在一个输入，位于 $[\mathbf{x}-\epsilon, \mathbf{x}+\epsilon]$内，且对于任意类别 $i \neq \hat{y}$，性质 $f(\mathbf{x})_i > f(\mathbf{x})_{\hat{y}}$均成立。此外，我们还会考虑已经找到的对抗输入，并排除其周围的一小块区域。例如，如果已知对抗输入 $\mathbf{x}'$ 满足 $\mathbf{x}'_1 = 0.5$，我们可以通过 $(\mathbf{x}_1 < 0.4999 \vee \mathbf{x}_1 > 0.5001)$ 限制下一次查询，以避免再次发现完全相同的反例。我们反复调用验证器并更新约束，为同一输入 $\mathbf{x}$搜集至多 k 个不同的对抗样本 $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots \mathbf{x}^{(k)}$。如果验证器未能找到此类输入，就略微增大 ϵ value，并使用相同的查询再次运行验证器。该过程持续进行，直至达到时间限制。最终，所有样本都会被赋予与原始输入从预言机获得的相同标签。这样一来，单个查询点$\mathbf{x}$便能生成一片具有高价值样本的局部云。

除了验证之外，还有一些轻量级方法可以识别对抗样本。为了衡量验证在此情境下的实用性，我们将我们的方法与

快速梯度符号法（FGSM） [18], 进行比较；这是一种梯度攻击，通过沿输入梯度符号的方向施加扰动来生成对抗样本。具体而言，给定输入$\mathbf{x}$ 和损失函数 L，扰动为 $\epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} L)$，其中 ϵ 是一个超参数。虽然通过使用不同的 ϵ 值运行 FGSM 可以生成多个对抗输入，但它只能沿单一梯度方向进行扰动。相比之下，验证可以到达偏离梯度的失效区域，并生成多个反例，从而揭示不同的错误行为模式——生成更加多样化的对抗输入集合。

4 评估

我们在监督式图像分类任务上评估了我们提出的数据增强方法，使用主动学习论文中经常采用的经典基准数据集：MNIST [28], 、Fashion-MNIST [48] 和 CIFAR-10 [25]„ 每个数据集都包含属于 10 个可能类别的图像。有关模型、训练过程和后端验证器的其他技术细节，请参见附录 A；我们的代码和基准数据集可在线获取 [43]。我们在文献中的多种深度主动学习方法上使用数据增强：

– DFAL（DeepFool 主动学习）：一种对抗样本查询策略，使用 DeepFool 攻击来近似模型的决策间隔 [13]。该方法选择所需对抗扰动最小的未标注样本，并查询其标签。每个选定样本的对抗实例也会被标记为同一类别，并添加到训练集中。– BADGE：一种深度主动学习方法，在确保特征表示具有多样性的同时，选择输出层梯度幅值较大的点 [4]。BADGE 为每个未标注实例计算一个“梯度嵌入”，然后使用 k-均值++ 聚类算法 [3] 选择一批 k 个点，使其梯度向量在范数和方向多样性方面都表现突出。– 随机：一种简单的基线方法，从未标注集中均匀随机地选择未标注样本。该方法不使用模型提供的任何信息，作为下界基准。

除了这些现有的深度主动学习方法外，我们还提出了 FVAAL——一种基于间隔的深度主动学习策略，它利用对抗扰动来估计未标注输入与模型决策边界之间的距离。FVAAL 基于快速梯度符号法来估计样本的间隔：标准快速梯度符号法会选取一个固定的 ϵ 值以诱发分类错误，而我们则迭代地 通过对 ϵ进行二分搜索，寻找能使模型预测发生翻转的最小扰动；当搜索区间缩小到足够小时，搜索即停止。最小的成功 ϵ 值记作 ϵ^*，可用作输入与决策边界之间距离（即间隔）的近似值。随后可按ϵ^*对未标注样本排序，值越小表示样本越靠近决策边界，因此根据基于间隔的标准，这些样本包含的信息越丰富。将 ϵ^*值最小的样本送去标注，并将每个样本生成的对抗输入

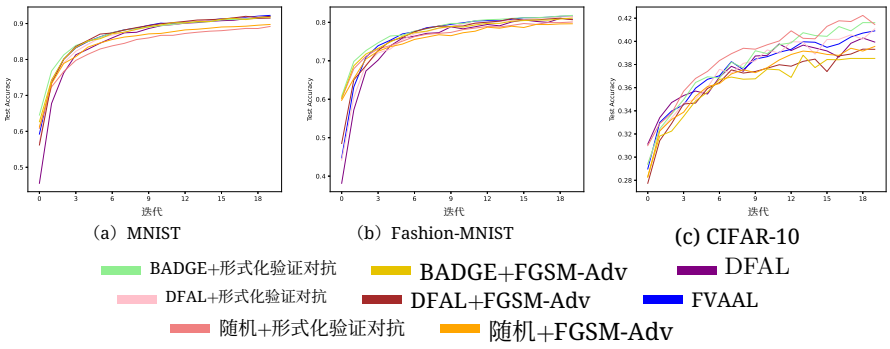


图1. 三个基准数据集上对抗输入的测试准确率与迭代次数的关系

也加入训练集，并赋予其与原始图像经预言机获得的标签相同的标签。在此选择机制之上，我们使用形式化验证器，通过验证生成的输入来扩充已标记样本，具体说明如上。FVAAL 使用一种简单而高效的方法，近似计算每个样本的间隔。与 DeepFool 等计算开销更大的多步攻击不同，后者会迭代调整扰动以寻找最小失真， [36],该方法只需计算一次梯度，并在每次迭代中进行一次简单的前向传播，以检查是否发生分类错误。这种高效性对于深度主动学习场景至关重要，因为可能需要评估数千个未标记候选样本。FVAAL 的完整伪代码见附录B。

对于每种方法，我们比较两个变体：(a) 对每个带标签样本，使用 Marabou 形式化验证器 [24, 47] 生成最多 10 个对抗样本，并在每次迭代时将其加入批次（“+形式化验证对抗”）；(b) 对每个样本，使用快速梯度符号法生成最多 10 个对抗输入，取值范围为ϵ 0.05 到 0.1（“+FGSM-Adv”）。对于 DFAL，由于其原生方法是为每个样本生成一个对抗输入，因此使用 Marabou/FGSM 额外生成了 9 个对抗输入。作为比较，我们还加入了原始 DFAL 方法，每个样本仅添加一个对抗输入。图1展示了各基准数据集上测试集分类准确率随深度主动学习迭代次数变化的情况。观察 MNIST 和 Fashion-MNIST 基准数据集可以发现，采用任一方法生成对抗输入的效果相近；形式化验证有时还能带来提升，例如在 MNIST 上使用 BADGE，或在 Fashion-MNIST 上使用随机方法时。我们还注意到，FVAAL 在几乎每次迭代中都是排名前三的方法之一。在 CIFAR-10 上， +形式化验证对抗在所有基线方法中都能带来更好的学习曲线。有趣的是，随机方法的变体在 CIFAR-10 上优于大多数其他变体，其中 Random+形式化验证对抗以明显优势表现最佳。这些结果表明，将验证作为一种数据增强技术能够改善学习曲线，且效果优于 FGSM。我们在附录C中进行了另一项实验，以评估 FVAAL 的选择准则，并证明其表现可与其他基线方法媲美。

表 1。所有基准数据集上的 AUBC。

方法	MNIST	Fashion-MNIST	CIFAR-10	平均值
BADGE	0.8250	0.7294	0.3723	0.6423
BADGE+FGSM-Adv	0.8708	<u>0.7729</u>	0.3637	0.6691
BADGE+形式化验证对抗	**0.8738**	**0.7812**	<u>0.3820</u>	**0.6790**
DFAL	0.8554	0.7533	0.3767	0.6618
DFAL+FGSM-Adv	<u>0.8720</u>	0.7681	0.3664	0.6688
DFAL+形式化验证对抗	0.8680	0.7626	0.3768	0.6691
DFAL-No-Adv	0.8356	0.7163	0.3664	0.6394
FVAAL	0.8719	0.7712	0.3777	<u>0.6736</u>
FVAAL-No-Adv	0.8359	0.7155	0.3680	0.6398
随机	0.8079	0.7203	0.3706	0.6329
随机+FGSM-Adv	0.8522	0.7575	0.3697	0.6598
随机+形式化验证对抗	0.8408	0.7640	**0.3878**	0.6642

粗体表示 = 在各基准数据集上的<u>最佳结果</u>；下划线表示 = 次佳结果。
AUBC 是准确率—预算曲线下的面积，归一化至 $[0, 1]$。

为进一步分析标注效率，我们使用预算曲线下面积（AUBC）衡量每种方法在各次迭代中的平均“增益”：AUBC 越高，表示整个深度主动学习过程中总体准确率越高。完整结果见表1。在大多数情况下，基于验证生成的输入比 FGSM 获得更高的 AUBC。具体而言，BADGE+形式化验证对抗在 MNIST 和 Fashion-MNIST 上的 AUBC 最高，在 CIFAR-10 上排名第二，略低于 Random+形式化验证对抗。FVAAL 在所有基准数据集上的 AUBC 均稳定排名第三，平均排名第二。这些结果证实，加入验证生成的输入不仅能提升准确率指标，也能带来整体上更理想的学习曲线。

为考察对抗输入的多样性，我们在附录D中计算了模型特征空间中各样本表示之间的所有两两距离。我们发现，在所有基准数据集和方法中，形式化验证生成的对抗输入集合都比快速梯度符号法（FGSM）生成的集合具有显著更高的多样性。我们推测，这就是表1所示性能更优的原因。

5 结论

我们提出了一种基于形式化验证器生成的对抗输入的数据增强方法，用于深度主动学习（DAL），将每个样本转化为一个规模较小且多样化的学习集。我们还提出了 FVAAL，这是一种将快速间隔评分与验证器生成的反例相结合的深度主动学习方法。在标准数据集上的实验表明，我们提出的数据增强技术能够实现高效的学习曲线，并取得出色的 AUBC。结果凸显了基于形式化验证的对抗样本增强在降低标注成本方面的潜力。未来的工作将探索使用更多验证器、动态学习预算以及更多数据集。

致谢。本研究部分由欧盟资助（欧洲研究委员会，VeriDeL，101112713）。但本文所表达的观点和意见仅代表作者本人，并不一定反映欧盟或欧洲研究委员会执行局的观点。欧盟和资助机构均不对这些观点和意见承担责任。本研究还获得了以色列科学基金会的资助（资助编号：558/24）。

参考文献

1. Amir, G., Corsi, D., Yerushalmi, R., Marzari, L., Harel, D., Farinelli, A., Katz, G.：基于学习的机器人导航系统验证。载于：第29届系统构造与分析工具和算法国际会议（TACAS）论文集。第607–627页（2023年） 2. Amir, G., Maayan, O., Zelazny, T., Katz, G., Schapira, M.：验证深度学习对分布外领域的泛化能力。《自动推理期刊》（JAR）**68**(17), 1–60（2024年） 3. Arthur, D., Vassilvitskii, S.：k-means++：谨慎选取初始中心的优势。载于：第18届ACM-SIAM离散算法研讨会论文集。第1027–1035页（2007年） 4. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.：基于多样性、不确定性和梯度下界的深度批量主动学习（2019年），技术报告。http://arxiv.org/abs/1906 03671 5 . . Balcan, M.F., Broder, A., Zhang, T.：基于间隔的主动学习。载于：计算学习理论国际会议（COLT）论文集。第35–50页（2007年） 6. Bassan, S., Amir, G., Corsi, D., Refaeli, I., Katz, G.：在反应式系统中形式化解释神经网络。载于：第23届计算机辅助设计中的形式化方法国际会议（FMCAD）论文集。第10–22页（2023年） 7. Bassan, S., Elboher, Y., Ladner, T., Althoff, M., Katz, G.：解释，有快有慢：对可证明解释进行抽象与细化。载于：第42届机器学习国际会议（ICML）论文集（2025年） 8. Chen, Z., Zhang, J., Wang, P., Chen, J., Li, J.：主动学习遇上隐式语义数据增强。载于：第17届欧洲计算机视觉会议（ECCV）论文集。第56–72页（2022年） 9. Corsi, D., Amir, G., Rodr´ıguez, A., S´anchez, C., Katz, G., Fox, R.：面向深度强化学习的验证引导式防护。载于：第1届强化学习国际会议（RLC）论文集（2024年） 10. Culotta, A., McCallum, A.：降低结构化预测任务的标注工作量。载于：第20届人工智能会议（AAAI）论文集。第5卷，第746–751页（2005年） 11. Desmartin, R., Isac, O., Passmore, G., Komendantskaya, E., Stark, K., Katz, G.：用于Imandra中深度神经网络验证的认证证明检查器。载于：第16届交互式定理证明国际会议（ITP）论文集。第1:1–1:21页（2025年） 12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.：BERT：用于语言理解的深度双向变换器预训练（2018年），技术报告。http://arxiv.org/abs/1810 04805 13 . . Ducoffe, M., Precioso, F.：深度网络的对抗式主动学习：一种基于间隔的方法（2018 09841 14 1802 ），技术报告。http://arxiv.org/abs/ . . Elboher, Y., Cohen, E., Katz, G.：关于在神经网络验证中应用残差推理。《软件与系统建模国际期刊》（SoSyM）（2023年）

15. Elboher, Y., Elsaleh, R., Isac, O., Ducoffe, M., Galametz, A., Pov´eda, G., Bouma- zouza, R., Cohen, N., Katz, G.：用于安全飞机滑行的跑道物体分类器鲁棒性评估。载于：第43届数字航空电子系统会议（DASC）论文集。第1–6页（2024年） 16. Elboher, Y., Isac, O., Katz, G., Ladner, T., Wu, H.：神经网络形式化验证中的基于抽象的证明生成。载于：第8届人工智能验证国际研讨会（SAIV）论文集（2025年） 17. Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, E., Chaudhuri, S., Vechev, M.：AI2：通过抽象解释对神经网络进行安全性与鲁棒性认证。载于：第39届IEEE安全与隐私研讨会（S&P）论文集（2018年） 18. Goodfellow, I., Shlens, J., Szegedy, C.：解释并利用对抗样本（2014 6572 19 1412 ），技术报告。http://arxiv.org/abs/ . . Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Pat wary, M.M.A., Yang, Y., Zhou, Y.：深度学习的规模扩展在经验上具有可预测性（2017 00409 20 1712 ），技术报告。http://arxiv.org/abs/ . . Huang, X., Kwiatkowska, M., Wang, S., Wu, M.：深度神经网络安全性验证。载于：第29届计算机辅助验证国际会议（CAV）论文集。第3–29页（2017年） 21. Isac, O., Refaeli, I., Wu, H., Barrett, C., Katz, G.：神经网络验证中的证明最小化。载于：第27届验证、模型检测与抽象解释国际会议（VMCAI）论文集（2026年） 22. Isac, O., Zohar, Y., Barrett, C., Katz, G.：深度神经网络验证、可达性与指数函数问题。载于：第34届并发理论国际会议（CONCUR）论文集（2023年） 23. Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.：Reluplex：用于验证深度神经网络的高效SMT求解器。载于：第29届计算机辅助验证国际会议（CAV）论文集。第97–117页（2017年） 24. Katz, G., Huang, D., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zelji´c, A., Dill, D., Kochenderfer, M., Barrett, C.：用于深度神经网络验证与分析的Marabou框架。载于：第31届计算机辅助验证国际会议（CAV）论文集。第443–452页（2019年） 25. Krizhevsky, A., Hinton, G.：从微型图像中学习多层特征（2009年），技术报告。http://www.cs.utoronto.ca/ kriz/learning-features-2009- TR.pdf 26. Krizhevsky, A., Sutskever, I., Hinton, G.：使用深度卷积神经网络进行ImageNet分类。载于：第25届神经信息处理系统进展（NIPS）论文集。第1097–1105页（2012年） 27. Lahav, O., Katz, G.：使用形式化验证对神经网络进行剪枝与切片。载于：第21届计算机辅助设计中的形式化方法国际会议（FMCAD）论文集。第183–192页（2021年） 28. LeCun, Y.：手写数字MNIST数据库（1998年），http://yann.lecun.com/exdb/mnist/ 29. LeCun, Y., Bengio, Y., Hinton, G.：深度学习。《自然》**521**(7553), 436–445（2015年） 30. Levy, N., Yerushalmi, R., Katz, G.：gRoMA：用于测量深度神经网络全局鲁棒性的工具。第12届形式化方法、验证与确认的应用国际研讨会（ISoLA）论文集（2023年） 31. Lin, X., Liu, X., Chen, B., Wang, Y., Dong, C., Hu, P.：ATAL：使用对抗训练进行数据增强的主动学习。《IEEE物联网期刊》**11**(3), 4787–4800（2023年）

32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.：迈向能够抵御对抗攻击的深度学习模型（2017），技术报告。http://arxiv.org/abs/1706 06083 33 . . Mandal, U., Amir, G., Wu, H., Daukantas, I., Newell, F., Ravaioli, U., Meng, B., Durling, M., Ganai, M., Shim, T., Katz, G., Barrett, C.：利用李雅普诺夫障碍证书对深度强化学习控制器进行形式化验证。载于：第15届国际计算机辅助设计中的形式化方法会议（FMCAD）论文集。第95–106页（2024年）34. Mandal, U., Amir, G., Wu, H., Daukantas, I., Newell, F., Ravaioli, U., Meng, B., Durling, M., Hobbs, K., Ganai, M., Shim, T., Katz, G., Barrett, C.：基于学习的航空航天控制器的安全可靠训练。载于：第43届数字航空电子系统会议（DASC）论文集。第1–10页（2024年）35. Moles, L., Andres, A., Echegaray, G., Boto, F.：探索数据增强和主动学习对不平衡数据集的益处。《Mathematics》**12**(12), 1898（2024年）36. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.：DeepFool：一种欺骗深度神经网络的简单而准确的方法。载于：第29届IEEE计算机视觉与模式识别会议（CVPR）论文集（2016年）37. Nielsen, C., Okoniewski, M.M.：受主动学习启发的样本获取方法在生成对抗网络数据增强中的应用。载于：IEEE计算机视觉与模式识别研讨会（CVPR）论文集。第109–112页（2019年）38. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.：深度主动学习综述。《ACM计算调查》（CSUR）**54**(9), 1–40（2021年）39. Sener, O., Savarese, S.：卷积神经网络的主动学习：核心集方法（2018 00489 40 1708 2009），技术报告。http://arxiv.org/abs/ . . Settles, B.：主动学习文献综述（），技术报告。http://minds.wisconsin.edu/handle/1793 60660 41 / . Shannon, C.E.：通信的数学理论。《贝尔系统技术期刊》**27**(3), 379–423（1948年）42. Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S.：运用深度神经网络和树搜索精通围棋。《自然》**529**(7587), 484–489（2016年）43. Spiegelman, J., Amir, G., Katz, G.：关于利用形式化验证改进深度主动学习：代码（2025年），http://github.com/josp1234/FormalVerificationDAL 44. Sun, C., Shrivastava, A., Singh, S., Gupta, A.：重新审视深度学习时代数据的惊人效用。载于：IEEE国际计算机视觉会议（ICCV）论文集。第843–852页（2017年）45. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.：神经网络的奇妙特性（2013年），技术报告。http://arxiv.org/abs/1312 6199 46 . . Tong, S., Koller, D.：应用于文本分类的支持向量机主动学习。《机器学习研究杂志》**2**, 45–66（2001年）47. Wu, H., Isac, O., Zelji´c, A., Tagomori, T., Daggitt, M., Kokke, W., Refaeli, I., Amir, G., Julian, K., Bassan, S., Huang, P., Lahav, O., Wu, M., Zhang, M., Komendantskaya, E., Katz, G., Barrett, C.：Marabou 2.0：一种多用途神经网络形式化分析器。载于：第36届国际计算机辅助验证会议（CAV）论文集（2024年）48. Xiao, H., Rasul, K., Vollgraf, R.：Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集（2017年），技术报告。http://arxiv.org/abs/1708.07747

A 技术细节

模型架构与训练。我们的实验采用简单的神经网络架构。对于 MNIST 和 Fashion-MNIST，我们使用包含一个隐藏层的全连接网络，该隐藏层有32个神经元。对于 CIFAR-10，我们选择了类似的架构，隐藏层有128个神经元。所有模型的隐藏层均使用 ReLU 激活函数，并输出 logits。模型使用 Adam优化器进行训练，学习率为0.001。每轮主动学习训练10个训练轮次，小批量大小为32。

主动学习协议。 每次迭代中，都会从未标注数据池中随机选取 10,000 个样本组成子集，并在这个子数据池上运行选择准则。对于 MNIST 和 Fashion-MNIST，每次迭代会加入 50 个样本，而 CIFAR-10 每次则加入 1000 个样本。在开始选择过程之前，所有模型都使用从均匀分布中选取的样本组成的初始已标注集进行训练。初始已标注集的大小与每次迭代的批次大小相同。每项实验均运行 20 个 DAL循环。每项实验重复 5 次，报告的结果为各次实验的平均值。

Marabou验证器。 在评估中，我们使用 Marabou深度神经网络验证器作为后端。Marabou 是一种现代求解器，集成了基于演绎和基于搜索的验证技术 [47]。此外，Marabou 还应用抽象-细化技术 [14]，并具备证明生成能力 [7,11,16,21]。Marabou 已应用于各种任务，包括机器人系统 [1,2,6]、目标检测系统的验证 [15]，以及航空航天领域的系统 [33,34]；它也曾作为后端用于神经网络的形式化剪枝 [27]。然而，我们的技术并不局限于特定的深度神经网络验证器，其他工具也可以使用——前提是它们能为可满足查询生成具体输入。

Marabou查询。在我们的实现中，属性仅针对亚军类别定义。也就是说，对于每个样本 $\mathbf{x}$ 在 ϵ-环境中的情况，属性 $f(\mathbf{x})_i > f(\mathbf{x})_{\hat{y}}$ 仅针对模型的亚军类别定义，而非所有类别。所选的 ϵ 值是通过快速梯度符号法结合二分搜索或通过 DFAL 计算得出的值，再增加 0.05 的小幅间隔，以便为找到对抗样本留出更多余量。对于 BADGE 和随机方法，由于 ϵ 值并非该方法计算的一部分，因此选取了固定值 $\epsilon = 0.01$。每次查询的运行时间限制为 30 分钟。

B 算法

算法 1 FGSM with Binary Search for Margin Estimation

要求： 网络 f，输入 x，容差 τ
确保： 估计的边界扰动 ϵ^* 和对抗样本 x^*

$start \leftarrow 0 end \leftarrow 1, \epsilon \leftarrow \frac{1}{2}$
当 $end - start > \tau$ **时执行**
  $x_{\text{adv}} \leftarrow$ 快速梯度符号法(f, x, ϵ)
  如果 $\arg\max f(x_{\text{adv}}) \neq \arg\max f(x)$ **则**
    $end \leftarrow \epsilon$
  else
    $start \leftarrow \epsilon$
  结束如果
  $\epsilon \leftarrow start + \frac{end - start}{2}$
结束循环
$\epsilon^* \leftarrow \epsilon$; $x^* \leftarrow$ 快速梯度符号法(f, x, ϵ^*)
返回 ϵ^* , x^*

算法 2 Formal Verification-Based Adversarial Active Learning (FVAAL)要求：网络 f_θ、验证器 $\mathcal{V}$、未标注数据池 $\mathcal{U}$、预言机 $\mathcal{O}$、初始训练样本数量 N_{init}、迭代次数 T、每次迭代子采样的数据池大小N_{sub}、每次迭代查询的样本数量 N_{query}、要生成的对抗样本数量 N_{adv}、容差 τ 确保：最终模型 $f_\theta^{(T)}$ 1：从 $\mathcal{U}$ 中随机抽取 $\mathcal{L} \leftarrow N_{\text{init}}$ 个样本，并通过 $\mathcal{O}$获取标签；$\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{L}$2：在 $\mathcal{L}$上训练 $f_\theta^{(0)}$ 3：对于 $t = 1$ 到 T，执行以下操作：4：$\mathcal{L}_t \leftarrow \emptyset$5：从 $\mathcal{U}$中抽取 N_{sub} 个样本点6：对于每个 $x_i \in \mathcal{U}_t$，执行以下操作：7：$\hat{\epsilon}_i \leftarrow$ BinarySearchFGSM$(f_\theta^{(t-1)}, x_i, \tau)$ ▷估计间隔8：结束循环9：从 $\mathcal{U}_t$ 中选取$\hat{\epsilon}_i$最小的$S \leftarrow N_{\text{query}}$ 个元素10：获取 $x_i \in S$的标签 $y_i \leftarrow \mathcal{O}(x_i)$；$\mathcal{L}_t \leftarrow \mathcal{L}_t \cup \{(x_i, y_i) : x_i \in S\}$11：对于每个 (x_i, y_i)，执行以下操作：12：由 $\mathcal{V}$生成的$A_i \leftarrow N_{adv}$ 个对抗样本13：$\mathcal{L}_t \leftarrow \mathcal{L}_t \cup \{(x', y_i) : x' \in A_i\}$14：结束循环15：$\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{L}_t$；$\mathcal{U} \leftarrow \mathcal{U} \setminus S$16：在 $\mathcal{L}$上训练 $f_\theta^{(t)}$ 17：结束循环18：返回 $f_\theta^{(T)}$

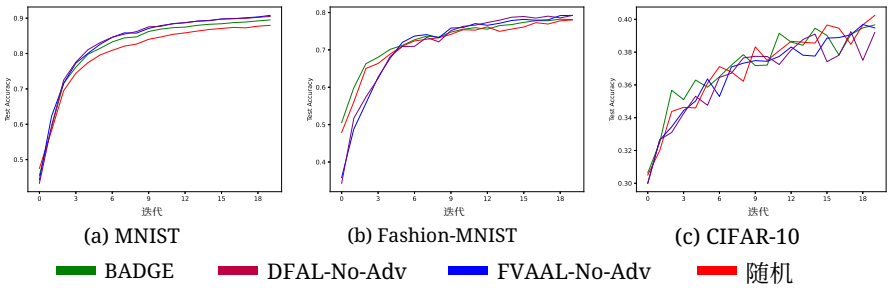


图 2. 三个基准数据集上不添加对抗输入时测试准确率与迭代次数的关系

C 选择准则

我们通过开展一项实验来评估 FVAAL 对信息量大的样本的选择，其中完全不向训练集添加对抗样本：BADGE 和 Random 保持原样使用；对于 DFAL 和 FVAAL，我们则使用丢弃所有对抗输入的变体（“-No-Adv”）。结果如图 2 所示。在 MNIST 和 Fashion-MNIST 上，FVAAL-No-Adv 保持了较高的准确率，优于 BADGE 和 Random，并且几乎在每次迭代中都取得了与 DFAL-No-Adv 相近的结果。在 CIFAR-10 上，结果波动较大，各种方法的表现大致相当。

D 对抗样本的多样性

为了量化 +FV-Adv 和 +FGSM-Adv 变体生成的对抗样本的多样性，我们测量其嵌入在模型特征空间中的分散程度。具体而言，对于每个变体，我们考察根据相应采集策略在最后一个 DAL 轮次中选出的 50 个样本。对于 MNIST 和 Fashion-MNIST，这一集合包含所有选出的样本；对于 CIFAR-10，则从选出的子集中均匀随机抽取 50 个样本。对于这些输入中的每一个，我们收集所有相应的对抗样本，使用该变体训练好的模型提取其倒数第二层嵌入，并计算这些特征向量之间所有成对的欧几里得距离。令$d(x_i, x_j)$ 表示 x_i 和 x_j的嵌入之间的距离。因此，每种方法均由经验分布 $\{d(x_i, x_j) \mid i < j\}$表示。

我们通过报告这些分布的均值和标准差来概括它们。此分析针对每个基准数据集的全部五次独立运行进行。报告的跨运行均值和标准差采用全方差定律进行汇总。表2列出了所有变体的这些统计量。

在几乎所有设置下，通过形式化验证生成的对抗样本，其平均值两两距离始终大于

表2. 对抗样本的特征空间多样性。各方法生成的50个对抗样本在倒数第二层嵌入空间中的所有两两距离的均值和标准差。均值越高，表明多样性越强。

方法	MNIST	Fashion-MNIST	CIFAR-10
BADGE+FGSM-Adv	26.186 ± 8.848	28.607 ± 14.413	6.924 ± 3.213
BADGE+形式化验证对抗	$\underline{33.631} \pm 11.617$	$\underline{35.0126} \pm 15.501$	$\mathbf{10.224} \pm 4.585$
DFAL+FGSM-Adv	24.468 ± 9.077	27.809 ± 16.496	6.669 ± 3.447
DFAL+形式化验证对抗	31.412 ± 11.932	32.552 ± 17.388	8.744 ± 4.074
FVAAL	31.987 ± 11.175	30.797 ± 15.282	8.658 ± 4.141
随机+FGSM-Adv	26.507 ± 9.225	28.666 ± 14.709	7.680 ± 3.759
随机+形式化验证对抗	$\mathbf{36.578} \pm 13.228$	$\mathbf{36.496} \pm 16.862$	$\underline{9.860} \pm 4.746$

每个基准数据集中最佳结果用粗体表示；第二佳结果加下划线。

通过快速梯度符号法生成的对抗样本，这表明其覆盖了特征空间中更广泛的不同方向。这说明，形式化验证能够生成更加多样化的对抗输入，有望减少深度主动学习采样中的冗余，并提高标签效率。