大家好，今天我们组将向大家介绍一篇关于数据采样在深度学习漏洞检测（DLVD）中的影响的研究论文，题为《数据采样是否改进了基于深度学习的漏洞检测？有利与否！》。

我们组的组员包括阿斯雅、胡博浩、陈恩宝、罗瑞和刘家骥。

这篇论文由Yi Li、Shaohua Wang、Xu Yang和Shaowei Wang在2023年的IEEE/ACM第45届国际软件工程会议（ICSE）上发表。

我们的PPT主要包括四个部分，研究背景、研究方法、研究结果和小组总结。

**研究背景**

首先，让我们了解一下研究背景。在深度学习应用于漏洞检测的过程中，尽管取得了很高的准确性，但现实世界中数据的不平衡问题始终存在。在实际情况中，易受攻击的案例远少于不易受攻击的案例，这种极不平衡的数据比例会显著降低预测的准确性。

前人研究（如S. Chakraborty等人在2021年发表的研究）表明，在不平衡数据下，最先进的深度学习漏洞检测方法的性能会急剧下降，其中F1得分平均下降了73%。然而，针对这种不平衡问题的数据采样研究相对缺乏。

为了应对这一挑战，作者提出并回答了两个主要问题：第一，数据采样是否提高了现有DLVD方法的有效性？第二，数据采样是否增强了DLVD学习易受攻击模式的能力？

**研究方法**

在研究方法部分，作者采用了多种数据采样方法，包括欠采样（Under-sampling）和过采样（Over-sampling）。欠采样通过减少多数类样本的数量来平衡数据集，具体方法包括随机欠采样（RUS）和OSS欠采样。而过采样则通过增加少数类样本的数量来平衡数据集，具体方法包括随机过采样（ROS）和SMOTE过采样。

作者选择了几种主流的DLVD方法进行实验，包括基于图的方法（如Devign、Reveal和IVDetect）和基于标记的方法（如LineVul）。

实验使用的漏洞数据集包括Devign、Reveal和BigVul，每个数据集都有其独特的特征和数据分布情况。Devign数据集包含来自FFmpeg和Qemu项目的超过22000个方法，约45%为漏洞方法。Reveal数据集相对不平衡，主要用于评估深度学习在漏洞检测中的有效性。BigVul数据集则是一个大型漏洞数据集，包含来自多个开源C/C++项目的代码漏洞，并包含漏洞修复信息。

在实验过程中，作者分别在原始数据上采样（Sampling\_R）和在嵌入空间上采样（Sampling\_L），并评估这些采样方法对DLVD性能的影响。

具体来说，作者将不同的数据采样方法（如ROS\_R、ROS\_L、RUS\_R、RUS\_L、SMOTE\_L、OSS\_L）运用于不同的数据集（如Devign、Reveal、BigVul），并通过精确度、召回率、F1分数、曲线下面积（AUC）和Precision@k（P@k）等评估指标进行效果评估。

**研究结果**

实验结果显示，无论采用何种数据采样方法，模型的性能在某些方面都有所提高。其中，过采样比欠采样表现更好，Sampling\_R比Sampling\_L效果更佳。特别是，ROS\_R方法在所有评估指标中表现最佳。

此外，ROS\_R方法显著增强了DLVD学习真正脆弱模式的能力。

**小组总结**

最后，在小组总结部分，我们指出了本研究的创新点。首先，这是首次对数据采样在处理DLVD数据不平衡问题上的影响进行系统性和广泛研究，填补了现有研究的空白。实验设计广泛，进行了1680次实验，耗费超过10200 GPU小时，验证了四种数据采样方法和两种采样策略对四种最先进DLVD方法在三个基准数据集上的影响。

其次，论文结合了不同的数据采样策略与解释性AI技术，提供了全面的评估框架，为未来研究和实践提供了有价值的见解。

在未来的研究中，我们认为可以通过以下几种方式进一步改进DLVD方法：增强模型的鲁棒性、增加数据增强技术、寻找更好的参数组合以及探索新的数据采样策略。

感谢大家的垂听，恳请批评指正。