“基于深度学习模型的物联网大数据实时分析”开题报告

1 项目背景

随着互联网的普及和网络技术的发展，网络流量呈指数级增长，物联网技术正在成为我们日常生活中不可或缺的一部分。同时，网络安全威胁也日益增加，包括恶意软件、钓鱼攻击、分布式拒绝服务攻击等。因此，有效分析和理解网络流量变得至关重要，而传统的方法可能无法有效或无法即使地处理大规模的数据或识别复杂的攻击模式。

由于各种因素，如设备移动性和网络异构性。特别是在蜂窝网络中，网络流量模式表现出非常复杂的行为。许多研究人员提出了许多机器学习（ML）技术模型来阻止物联网网络中的恶意流量。然而，由于不适当的特征选择，一些ML模型容易对大多数恶意流量进行错误分类。因此，如何在物联网网络中选择有效的特征进行准确的恶意流量检测仍然是一个需要深入研究的重要问题。

深度学习（DL）模型已被有效地用于大数据系统的分析和知识发现，以识别隐藏和复杂的模式。在这些成功的激励下，网络领域的研究人员将深度学习模型应用于网络流量监控和分析（NTMA）应用，例如流量分类和预测。

1.1 NTMA

NTMA指的是在适当的粒度级别（例如，在数据包级别）监控网络流量的一系列技术。NTMA技术深入了解了网络的运行和性能以及用户的行为。在通信系统和网络的背景下，NTMA在以下方面发挥了关键作用：了解网络的工作原理并监控网络的性能，消费者如何最大化利用资源，如何有效地控制和管理电信基础设施以提供服务水平协议（SLA）。

1.2 深度学习模型

深度学习是一种面向工程的方法。深度学习是ML的一个特定子领域，深度学习使用深度神经网络（Deep Neural Network, DNN）在每一层寻找数据表示，用于数据建模的层数称为模型的深度。对于像图像识别这样的复杂任务，深度学习模型通常具有数十甚至数百个连续的表示层。与深度学习相反，其他机器学习模型通常涉及一层或两层的数据表示。

1.3 卷积网络

卷积网络，也称为卷积神经网络（Convolutional neural networks, cnn），是一种专门处理类网格数据的特定类型的神经网络。cnn不是在一般的矩阵乘法的神经网络使用卷积算子而是至少一个网络层中使用卷积算子。这种数据类型的示例是时间序列和图像，它们可以分别视为像素的一维网格和二维网格。卷积网络已广泛应用于各种现实问题，如自然语言处理（NLP）、计算机视觉、语音识别等。

2 目标与任务

2.1 目标

1.通过使用特定的流量数据集进行训练，最终训练出一个能够自主学习迭代，针对异常或恶意网络流量识别率超过80%并能够生成分析报告的大语言模型。

2.以特定知识库为基础，所训练的大模型能够自动化生成恶意流量代码与生成报告，使其能够更好地答复在防御规避方面的细节。

3.通过对流量攻击过程的学习，所训练的大模型不仅能还原流量攻击中的混淆Payload，还能从响应报文中动态识别是否存在攻击成功的特征。

2.2 任务

1.从github上clone下目前开源的语言大模型（如Baichuan2），并将其部署到实验机本地。

2.学习CorrAUC特征识别技术，理解其是如何选择、过滤流量的特定特征，并最终有效识别流量的原理，并能够成功复现他的识别过程。

3.使用Bot-IoT数据集，并利用CorrAUC特征识别方法进行初步的训练，Bot-Iot数据集包括了不同类型的危害攻击，尤其是僵尸网络的网络攻击。通过这样的训练，使大模型最终能够达到预期的识别率。

4.对各类开源知识库做调研，了解其应用背景，并选择合适的知识库来训练大模型，使大模型通过学习大量开源的代码，对代码语义有更深入理解。并最终实现使大模型生成大量特征相似的恶意流量代码。

5.对常见流量攻击（如DDoS）做调研，了解底层逻辑与原理，明晰流量攻击的运转体系。另一方面，了解当前业界对于流量攻击检测的瓶颈，例如如何判断攻击是否成功。

6.针对调研成果，选择合适的流量数据集与相关知识库来进行训练，使我们的大模型具备还原和复现攻击过程的能力。

3 可行性分析

3.1 技术可行性

结构化数据使用DL会有困难。结构化数据指的是一种标准化的格式，它将数据组织在有行和列的表中，例如网络流量数据。然而，DL技术最成功的应用是针对非结构化数据的问题。一些机器学习专家反对将DL用于结构化数据，因为他们认为标记的结构化数据集不足以训练DL算法。

但从其他角度分析，仍可以参考通信系统和网络产生的大量数据，这些数据可用于训练DL算法，但需要考虑标记的代价。此外，由于深度学习模型在不同领域的流行，已经建立了多个框架来促进深度学习的使用。

3.2 环境条件可行性

网络攻击的规模、复杂性和成本都在增加。因此，加强通信系统和网络的安全性以抵御网络威胁至关重要，特别是考虑到人们越来越依赖于无线网络，例如蜂窝网络和WiFi进行日常生活活动。

在NTMA中应用基于DL的技术来构建预测模型，需要有大量的网络流量。尽管通信系统和网络（如物联网和5G）的巨大进步导致每天产生大量原始数据，但数据标记在时间、计算开销和人力方面都是一项昂贵的任务。在网络的实际应用中，大多数数据是未标记或半标记的。将DL技术与其他能够处理未标记或半标记数据的ML技术集成可以被认为是一个很好的解决方案。

4 初步方案与关键技术

4.1 初步方案

第一步是明确NTMA的目标，目标为流量分类和网络安全。第二步是通过被动或主动监测方法收集网络管理数据。第三步是数据预处理和清理，规范化是一种预处理技术，旨在提高NTMA应用程序的性能，特别是对于基于DL的方法至关重要。然后，在数据预处理之后，NTMA必须经过一个特征选择步骤，以选择信息量最大的特征来服务于目标。最后对预处理后的数据进行深入分析，以提取有意义的信息，基于DL的方法对数学和统计方法需求较小，因为它们能够发现原始数据的隐藏模式。

4.2 关键技术

4.2.1 CorrAUC

CorrAUC法（Correlation Area Under the Curve）是一种新的特征选择度量方法，它用于物联网网络中的恶意流量检测，CorrAUC用曲线下面积（AUC）来衡量特征和目标类别之间的相关性。特征是用来描述数据的属性，目标类别是用来区分数据是正常的还是恶意的。AUC是一种评价特征区分能力的指标，它的值越大，说明特征越能够把正常的和恶意的数据分开。CorrAUC就是利用AUC来选择最有效的特征，从而提高恶意流量检测的准确性和效率。

4.2.2 TOPSIS

TOPSIS法（Technique for Order Preference by Similarity to Ideal Solution）是一种常用的综合评价方法，其能充分利用原始数据的信息，其结果能精确地反映各评价方案之间的差距。

为了对众多方案给出一个排序，在给出所有方案之后，可以根据这些数据，构造出一个所有方案组成的系统中的理想最优解和最劣解。而TOPSIS的想法就是，通过一定的计算，评估方案系统中任何一个方案距离理想最优解和最劣解的综合距离。如果一个方案距离理想最优解越近，距离最劣解越远，就有理由认为这个方案更符合要求。

5 预期工作结果

1.训练出可投入使用的流量检测大模型，可在多平台兼容使用，对于输入的流量包，能够准确识别其中的异常或恶意网络流量，其生产力可替代约5个网安审计师。

2.大模型应当具有代码审计与恶意代码生成的功能，在通过相关知识库的训练后，能够和使用者在网络流量攻击方面进行有意义的讨论。

3.大模型能够还原和复现流量攻击的过程，为相关技术人员提供实例参考。

6 进度计划

XX月XX日——XX月XX日 撰写文献综述报告

XX月XX日——XX月XX日 撰写毕业论文初稿

XX月XX日——XX月XX日 修改论文