大模型于网络安全的应用

汇报人：唐晨阳
专业：软件工程

南京郵電大學
Nanjing University of Posts and Telecommunications

目录

— content —

01

背景介绍

Part one

南京郵電大學
Nanjing University of Posts and Telecommunications

大模型应用&网络安全

大模型特点

大模型是“大算力”“大数据”与“强算法”结合的划时代人工智能产物，随高性能芯片算力、训练数据集、神经网络复杂度变化而进化。国内常指大语言模型，也涵盖视觉、多模态等大模型。它具备语义分析、代码理解、复杂推理等能力，可应用于众多行业。

两者的关联

大模型在网络安全领域有以下优势：一是语义分析能力，可基于文本特征助力威胁情报共享、异常流量发现；二是代码理解能力，能识别代码相关模式并发现安全问题；三是复杂推理能力，可辅助安全运营团队实现攻击溯源及响应处置自动化、自主化。

大模型应用&网络安全

目前已有多家公司推出了安全领域的大模型应用，如图所示：

公司	安全大模型	基础模型	应用场景	功能
微软	Security Copilot	GPT-4	安全事件调查处置	对安全事件攻击路径、攻击目标进行还原，提出事件处置方案等
谷歌	Sec-PaLM 2	PaLM 2	代码审计	解释潜在恶意脚本行为检测代码威胁
深信服	安全 GPT	类 GPT 技术	XDR 平台	实现高级威胁检测、安全监测调查、热门漏洞排查
启明星辰	PanguBot（盘小古）	—	安全运营	实现自动化和智能化的安全运营
奇安信	类 ChatGPT 安全大模型	类 ChatGPT 技术	—	应用于安全产品开发、 威胁检测、漏洞挖掘、 安全运营及自动化、攻防对抗、反病毒、威胁情报分析和运营、涉网犯罪分析等领域
安恒信息	—	类 ChaGPT 技术	数据安全	数据分类分级、智能生成检测规则
360	360GPT 安全应用框架	360 智脑	—	—
绿盟科技	SecXO ps	A 经验积淀	安全智能分析平台	安全智能分析

— 02 —

关键技术

Part two

南京郵電大學
Nanjing University of Posts and Telecommunications

大模型在网络安全领域的潜在应用场景

异常流量检测

当下网络攻击呈现出一种新态势，攻击行为与正常网络行为之间的相似度持续攀升，其隐蔽性愈发深厚，传统的异常流量检测手段难以察觉攻击流量中那些极为细微的差异。利用大模型对网络日志、流量数据进行聚类，可以识别和捕获与正常行为模式不符的异常流量。

攻击行为发现

大模型可进行代码比对，精准识别已知攻击模式与工具，分析代码及执行路径能发现恶意代码与攻击行为，还能结合威胁情报分析共享功能，与数据库对比匹配，挖掘海量数据关联与新威胁迹象，提升未知攻击发现能力。

漏洞利用排查

大模型通过对代码展开静态、动态分析，可精准定位潜在漏洞、安全风险及错误运用之处。它能深度理解代码语法、API 调用与数据走向，锁定可能引发漏洞的代码片段，依据 API 调用及数据流向路径，联动资产管理系统，自动筛查可能受波及的 IP 服务器与系统资产。

大模型在网络安全领域的潜在应用场景

安全运维审计

大模型可全方位审查代码，精准揪出违反安全与合规标准的代码片段，彻查敏感信息处理、访问控制及加密运用等方面的漏洞，挖掘潜藏的安全隐患。如中国电信借助大模型，帮助安全运维审计人员高效精准地处置代码安全问题。

攻击溯源分析

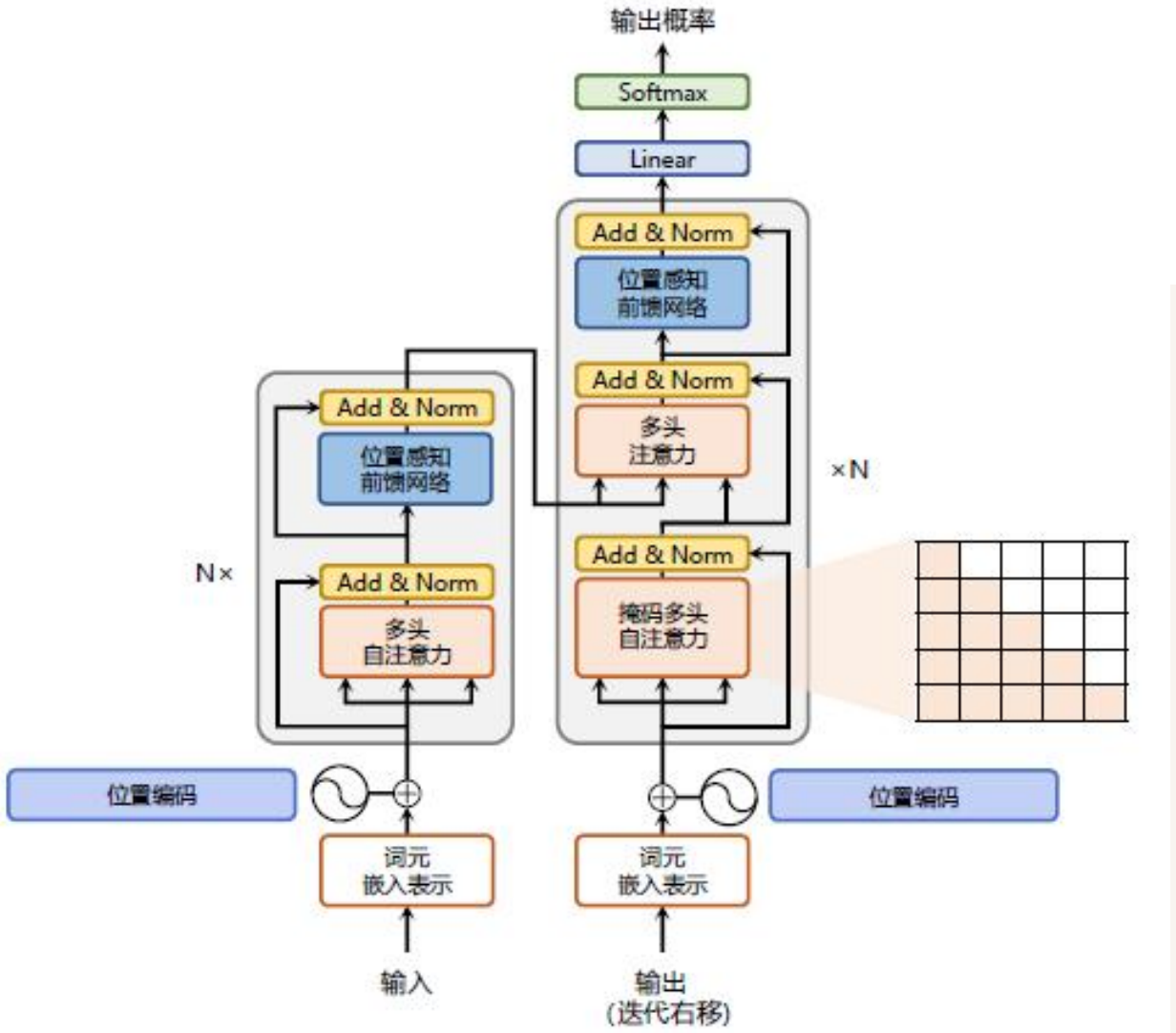
大模型助力安全团队剖析攻击事件，深度解析行为模式、技术特性与攻击者意图。融合历史数据、网络流量、系统日志，展开精密推理，精准定位攻击源，追踪攻击轨迹。借助威胁情报，大模型迅速甄别攻击关联，挖掘攻击者来源及所用工具和技术。

大模型在网络安全应用中的核心技术

深度学习框架（Transformer架构）

从整体结构上看，Transformer 主要由编码器（Encoder）和解码器（Decoder）组成。编码器用于理解输入序列的语义信息，解码器则在理解输入的基础上生成目标序列。

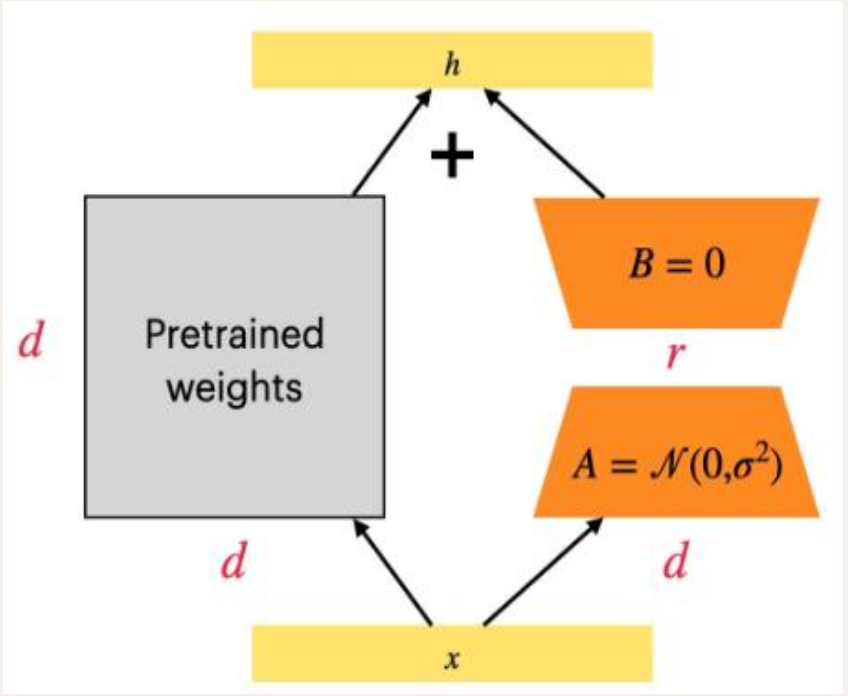
Transformer 架构的优势在于它摒弃了传统的循环神经网络（RNN）和卷积神经网络（CNN）在处理序列数据时的一些局限。Transformer 可以高效地并行计算，并且能够很好地处理长序列中的长距离依赖，在机器翻译、文本生成等众多任务中取得了优异的性能。



大模型在网络安全应用中的核心技术

预训练与微调技术

大模型预训练可从大规模网络安全数据里汲取通用知识与模式，增强泛化性并提升效率。微调则依据特定任务，如攻击检测等，利用少量标注数据精准适配，让模型既有广泛认知基础，又能出色完成专门网络安全任务，实现性能优化。



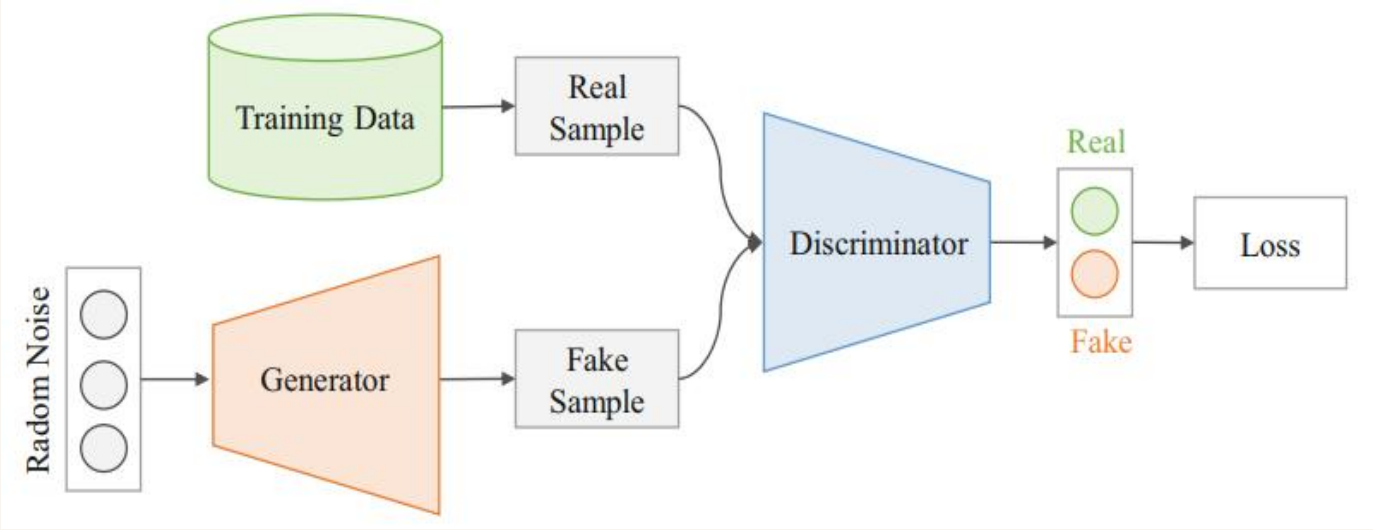
Lora微调

Lora微调的核心思路是通过低秩矩阵来近似表示模型参数的更新。Lora 假设可以用两个低秩矩阵的乘积来表示对原始权重矩阵的微调更新。例如，对于一个权重矩阵W，Lora使用$W+\Delta W$来表示微调后的权重，其中$\Delta W = BA$，B以及A是两个低秩矩阵。

大模型在网络安全应用中的核心技术

对抗生成技术（生成对抗网络）

GAN 主要包含生成器（Generator）和判别器（Discriminator）两个部分。在训练过程中，生成器和判别器相互竞争、交替进行训练。先固定生成器的参数，训练判别器使其能够准确地分辨真实样本和生成的假样本；然后固定判别器的参数，训练生成器使其生成的样本更好地欺骗判别器。通过不断重复这个过程，生成器和判别器的性能都逐渐提升，最终达到一个动态平衡。

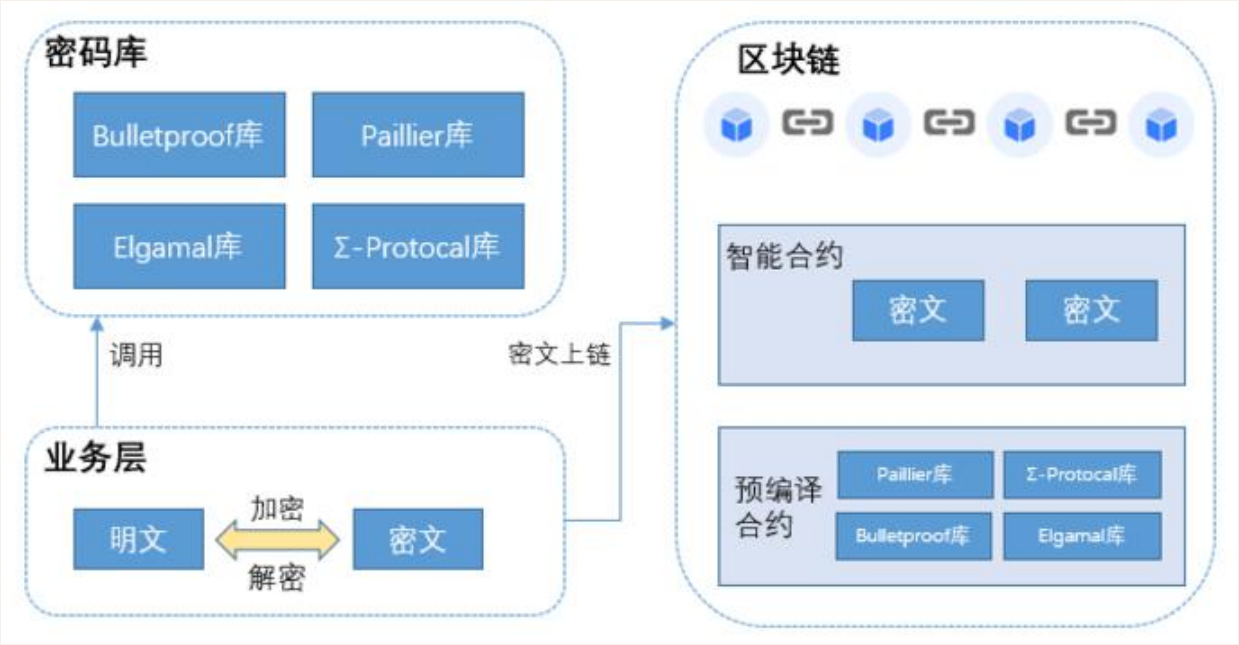


GANs模型

大模型在网络安全应用中的核心技术

隐私保护技术

大模型在隐私保护技术中有着丰富多样的应用，例如全同态加密，像 Zama 对大模型操作拆解应用以保护输入输出隐私；腾讯安全的隐私保护脱敏技术处理提示词；区块链保障数据不可篡改与可追溯，零信任模型严格身份验证授权；还有模型参数加密技术等等，在诸多场景中各展其长，有力地推动了大模型于隐私保护领域的发展。



区块链技术方案示意图

03

未来发展

Part three

南京郵電大學

Nanjing University of Posts and Telecommunications

大模型在安全领域的前景展望

在应用上，将深度融入威胁检测与防御，精准识别新型攻击；助力漏洞管理，高效扫描并修复漏洞；推动安全运营自动化，降低人力成本与失误；强化数据安全保护，精准分类分级与加密数据；优化风险评估与预测，助力提前布局策略。

技术层面，多模态融合可综合分析多源数据提升效果；强化学习使模型能在交互中优化决策；模型融合集成则汇聚优势增强可靠性。

大模型在安全领域的发展难点

1

高质量数据集匮乏

网络安全大模型训练数据集有格式、噪声、标注难点，应规范并优化处理。

2

大模型答案的可信度不稳定

大模型有 容易产生“幻觉”，影响安全决策，需要专业训练、严格把关数据、综合评估大模型返回的答案。

3

大模型数据调参消耗资源过大

大模型调参面临超参数多、资源耗大难题，需要开发更多的改进调优策略。

4

AI 与安全领域不互通

网络安全大模型开发中，AI 与安全人才协调不容易，需要建立合作机制，促进知识共享与人才培养。

5

模型更新与维护

安全领域数据变化迅速，新的威胁、漏洞不断涌现，保证大模型持续有效性是一大难点。

— 04 —

创新总结

Part four

南京郵電大學

Nanjing University of Posts and Telecommunications

大模型应用总结

攻击识别

大模型可凭借对海量数据的分析能力，学习各类网络攻击的复杂模式，从而有效识别新型的恶意软件、高级持续性威胁以及网络钓鱼手段等，即便攻击者不断变换攻击手法，也能察觉异常，提前预警。

威胁检测

风险评估与防范

通过分析数据的使用、流转情况以及外部环境因素，评估可能出现的隐私泄露风险，还能给出针对性的隐私保护策略建议，比如推荐合适的数据脱敏方法、加密技术应用场景等，保障数据隐私安全。

数据安全

自动化安全流程

可以实现诸如漏洞扫描、安全策略配置、事件报告生成等诸多安全运营流程的自动化操作，减少人工操作的繁琐与误差，提高整体安全运营的效率，让安全团队能将更多精力投入到更复杂的安全问题解决上。

运营管理

技术总结

Transformer架构

大模型基于Transformer架构构建，其多头注意力机制助力精准捕捉数据特征，为语义理解等任务提供强大支撑。

预训练与微调技术

大规模预训练学习通用知识，再针对特定网络安全任务微调。以少量数据优化，适应多样场景，提升模型实用性与针对性。

对抗生成技术

利用对抗生成网络，生成虚假网络数据用于测试模型鲁棒性，同时可模拟攻击，促使模型不断进化，增强防御真实攻击能力。

隐私保护技术

用多种隐私保护手段，如全同态加密让数据在密文态运算，脱敏技术处理敏感信息，保障数据安全，平衡模型效能与隐私需求。

THANKS

感 谢 观 看

汇报人：唐晨阳

专业：软件工程

南京郵電大學
Nanjing University of Posts and Telecommunications