国家网络空间安全基础能力的搭建者
Pioneer of National Cyberspace Security

恒安嘉新
EVERSEC

Intel® oneAPI加速基于AI的加密流量识别

恒安嘉新（北京）科技股份公司

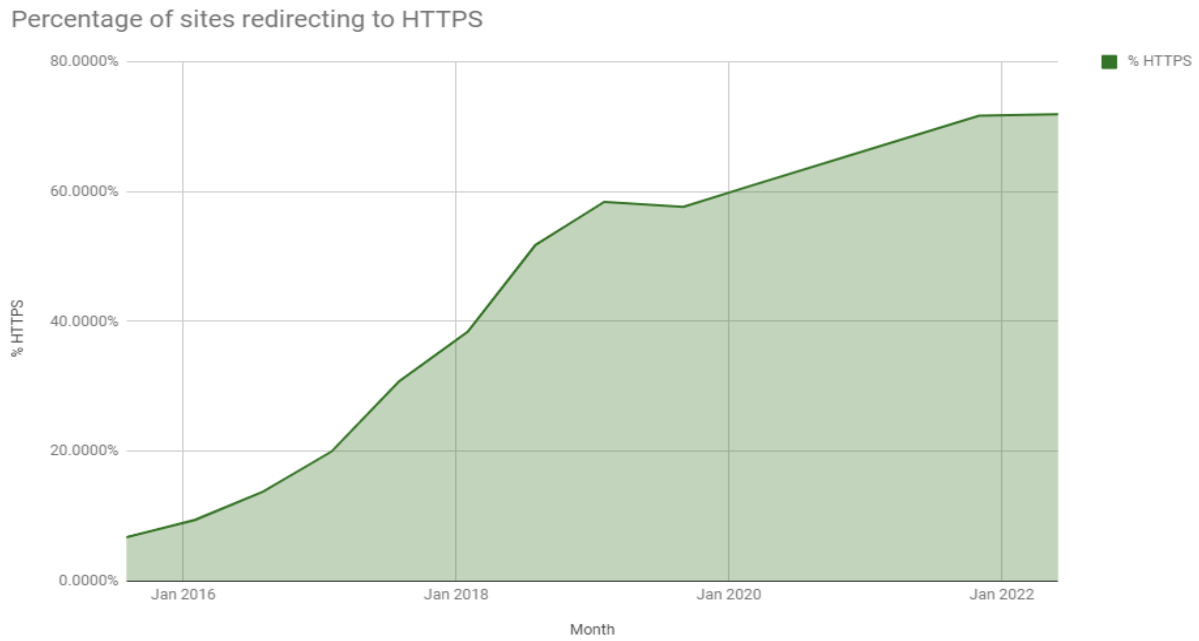
目录

加密流量背景与需求

1、流量全密化的趋势

随着人们网络安全、隐私保护意识的不断提高和加密技术的广泛应用，网络中的加密流量呈现爆炸式增长。

特别是TLS等加密协议的不断演进、DNS加密化、QUIC协议的推广，加密应用的全面普及和网络通信流量的加密化已经成为不可阻挡的趋势，我们正在走向全加密时代 。



2、加密带来的隐患

- 攻击者可以通过加密来隐藏自己的攻击行为。
- 网络全面化加密的滥用给个人与企业的财产安全造成了危害，比如网络诈骗给人民群众带来的财产损失等。
- 给网络监管带来挑战。

加密流量常用识别方案

- 业界传统做法，大多是人工或者半自动化的方式搜集TLS协议的证书，基于证书识别对应的业务。**这个过程与传统的业务识别是完全相同的模式，即找出业务的一定量特征字符串做匹配。**
- 目前市场上全球信任的支持浏览器的SSL证书主要有 3 种： EV SSL证书(Extended Validation SSL Certificate)、OV SSL证书(Organization Validation SSL Certificate) 和 DV SSL证书(Domain Validation SSL Certificate)。

优点

- 基于成熟的特征匹配技术，方案简单可靠；
- 识别准确率较高；
- 微信、支付宝等有部分模仿TLS实现的私有协议也能识别；
- 可以建立SSL规则库，适合DPI设备大规模的识别采用。

缺点

- “一爬、二安、三点击”操作消耗大量的人工成本；
- 只能识别携带了明确的TLS证书的流量，通常只有首条连接才携带证书；
- 只能识别有限已经遇到的，无法识别未知；
- 需要大量的人工成本来积累证书，短时间内无法做到大面积覆盖。

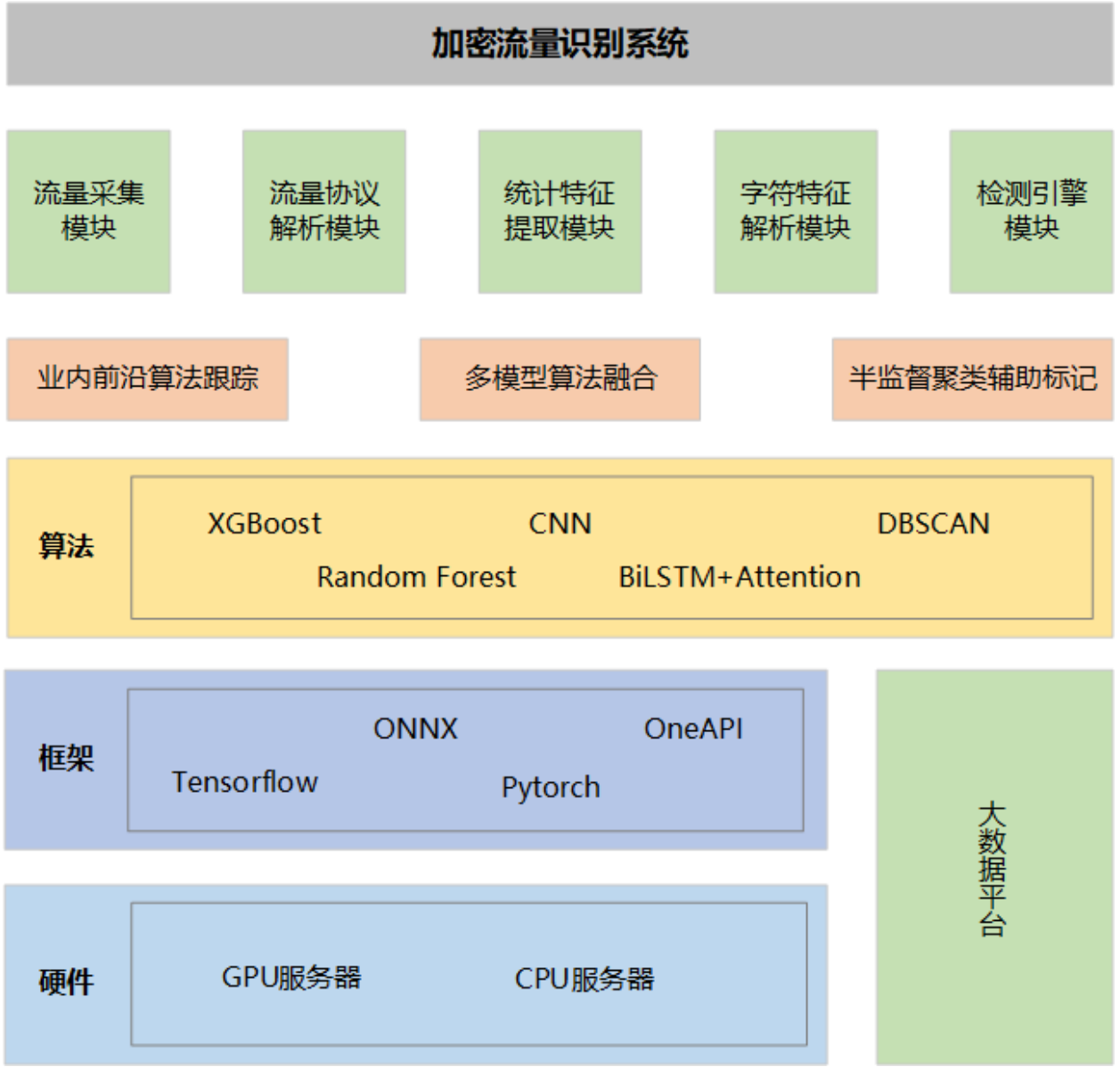
恒安嘉新加密流量识别方案

研究课题：

HTTP协议采用明文传输，而HTTPS在HTTP基础上增加了TLS/SSL协议安全嵌套，客户端产生一个对称的密钥，通过服务器的证书来交换密钥，数据的传输采用加密处理，而第三方因为没有证书将无法解析数据，这给当前DPI业务带来了极大的挑战，本课题目的在于探索基于人工智能的加密网络流量检测方法。

解决方案：

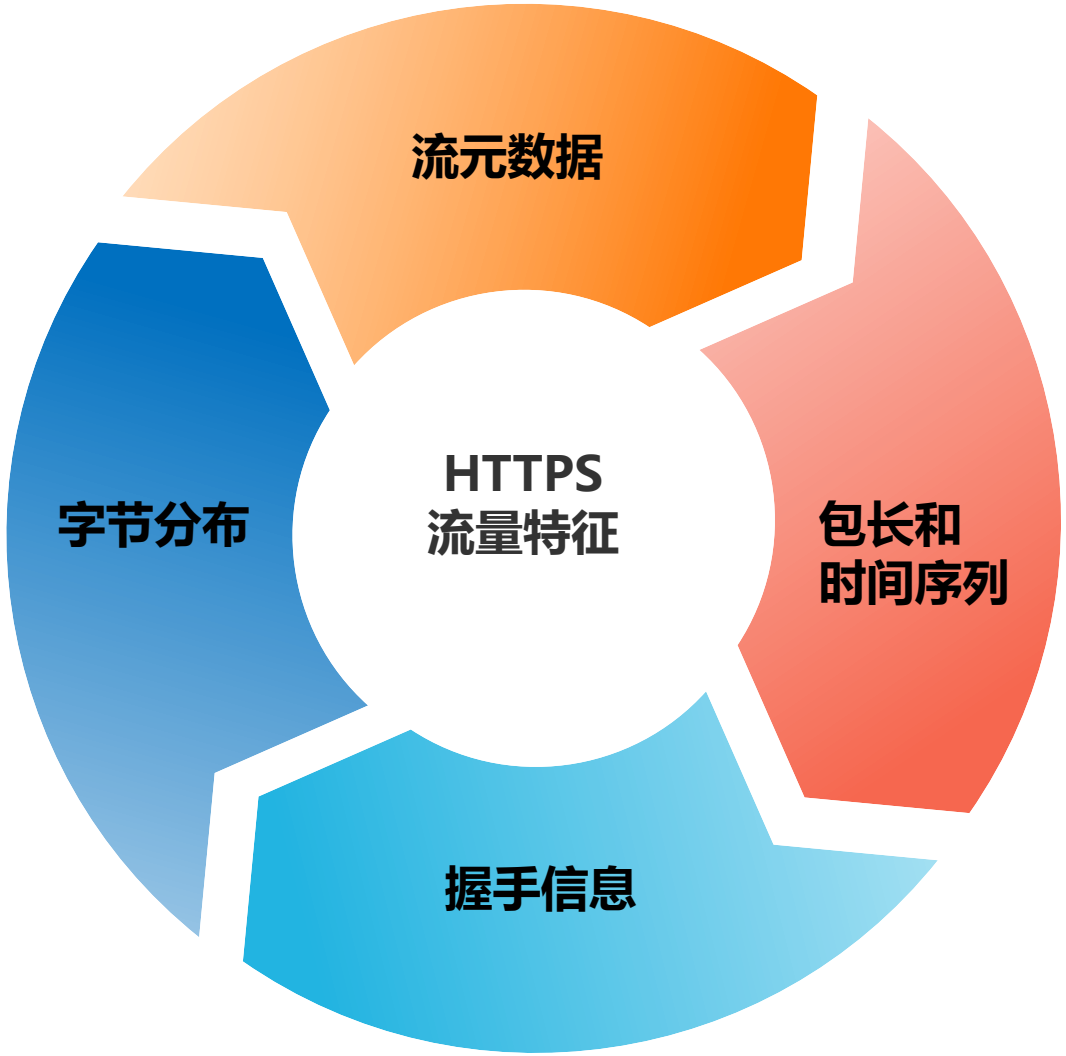
恒安嘉新提出基于人工智能的加密流量计算平台，为加密流量检测场景提供更加智能、高性价比的解决方案。基于Intel TADK 模块，综合利用多模型机器学习的方法，对各种不同业务类型的加密流量进行有针对性的检测，有效识别出业务类型的使用。



目录

加密流量特征选择

对TLS流及背景流量信息进行统计分析，寻找流量中具有明显区分度的数据特征以及相应特征值的规律性。



流元数据

选取上行字节数、下行字节数、上行包数、下行包数、源端口、目的端口以及总的持续时长，然后以秒为单位做一个归一化处理，这些特征在不同的应用上有细微的差异

包长和时间序列

选取前50个包的包长和时间，排成一个序列，这些特征在不同的应用上存在差异

字节分布

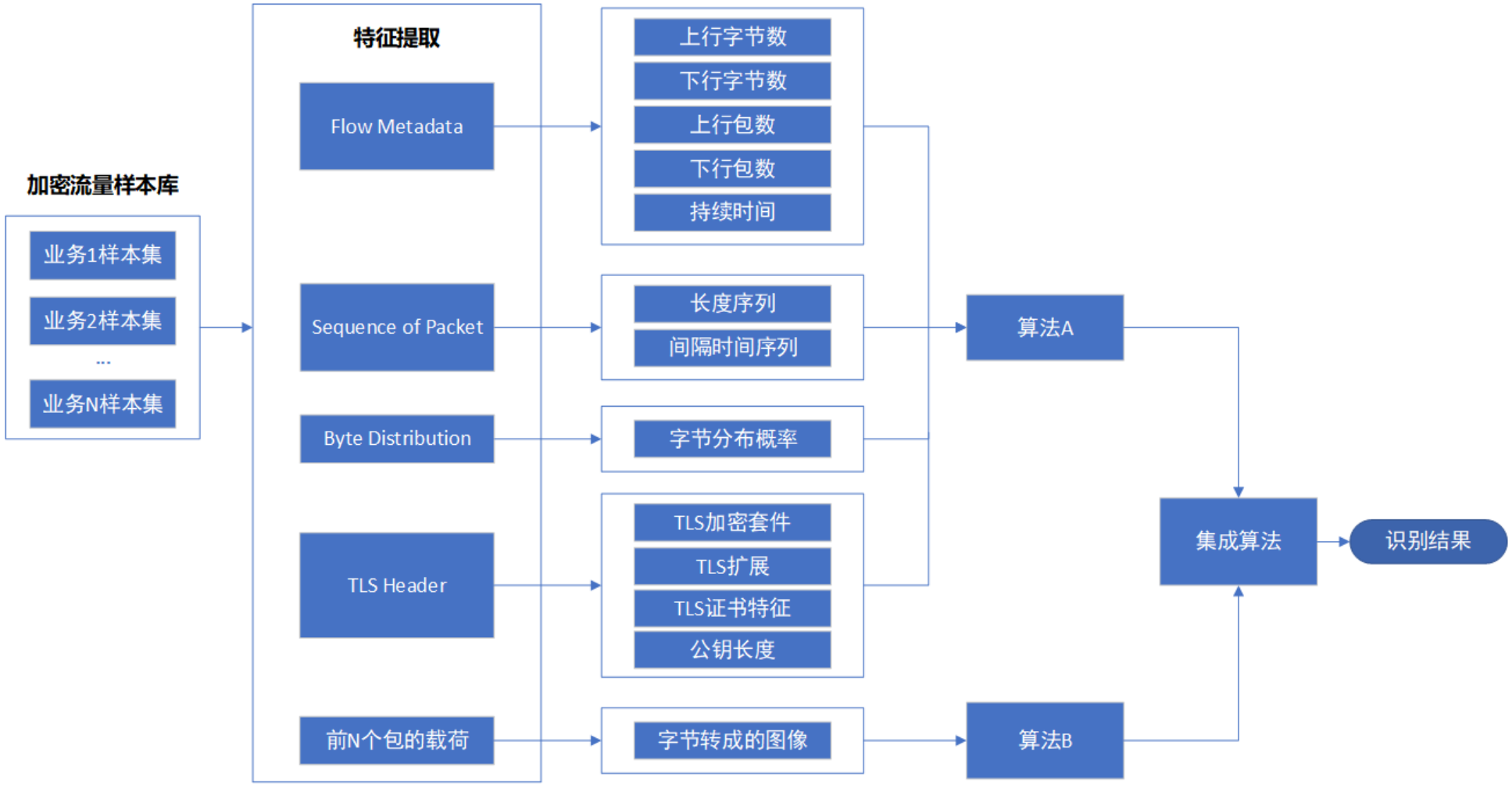
选取前16个包的payload的前160和后32字节，不同的应用在传数据的时候首位信息也存在差异

握手信息

选取拥有完整TLS握手信息的TLS流作为研究对象，通过分析TLS握手信息发现，流量使用的加密套件（ciphersuites ）、支持的扩展（extensions）等方面有很大的差异

加密流量系统架构

- 基于TLS流量的元数据、包长和时间序列、字节分布、非加密的TLS头信息等特征通过机器学习算法进行分类；
- 将每条TLS流的前N个载荷拼接起来并转化成图象通过深度学习算法进行分类；
- 将前两步的分类结果经过集成学习综合得出最终识别结果。

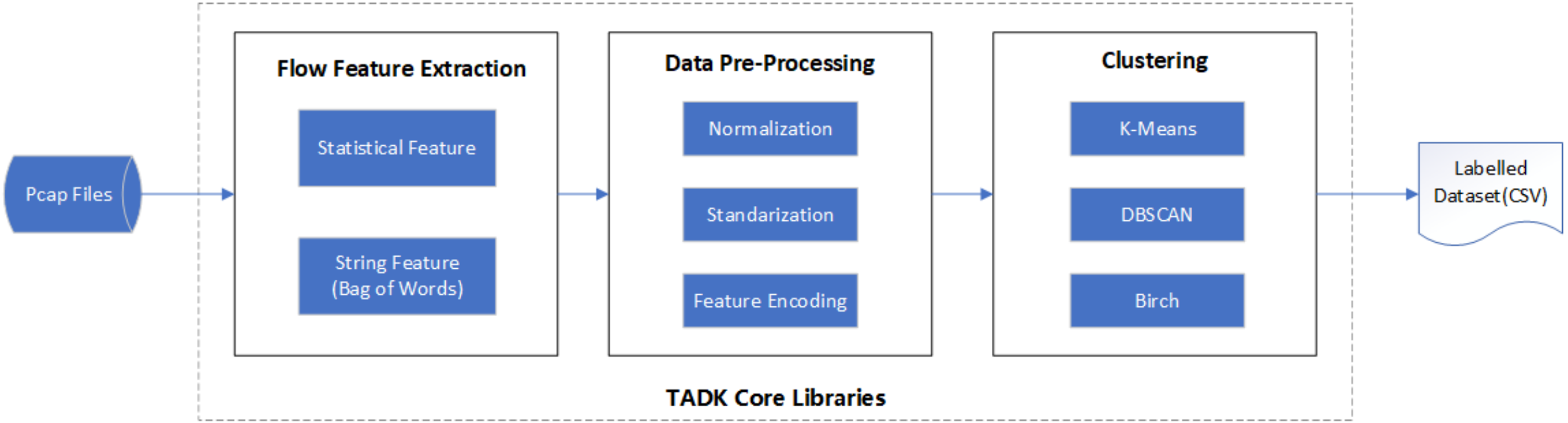


半监督聚类算法辅助标记

AI三要素：数据、算法、算力。

人工智能和机器学习应用程序，如此依赖于高质量的样本数据。海量标注好的样本，需要大量人力物力，成本极高，部分需求标注难度大。

Intel TADK 提供了半监督聚类算法，提取加密流量的统计特征和字符特征，基于K-Means、DBSCAN和Birch多种聚类算法实现辅助标记数据功能，而且数据标记质量良好，极大程度降低了人工标注的成本。



目录

课题方向

研究基于人工智能的加密流量识别在电商类APP业务分类领域的可行性

数据集

主流电商类APP：淘宝、天猫、京东、拼多多等电商类应用

模型概览

加密流量分类模型、半监督聚类模型

项目性能指标

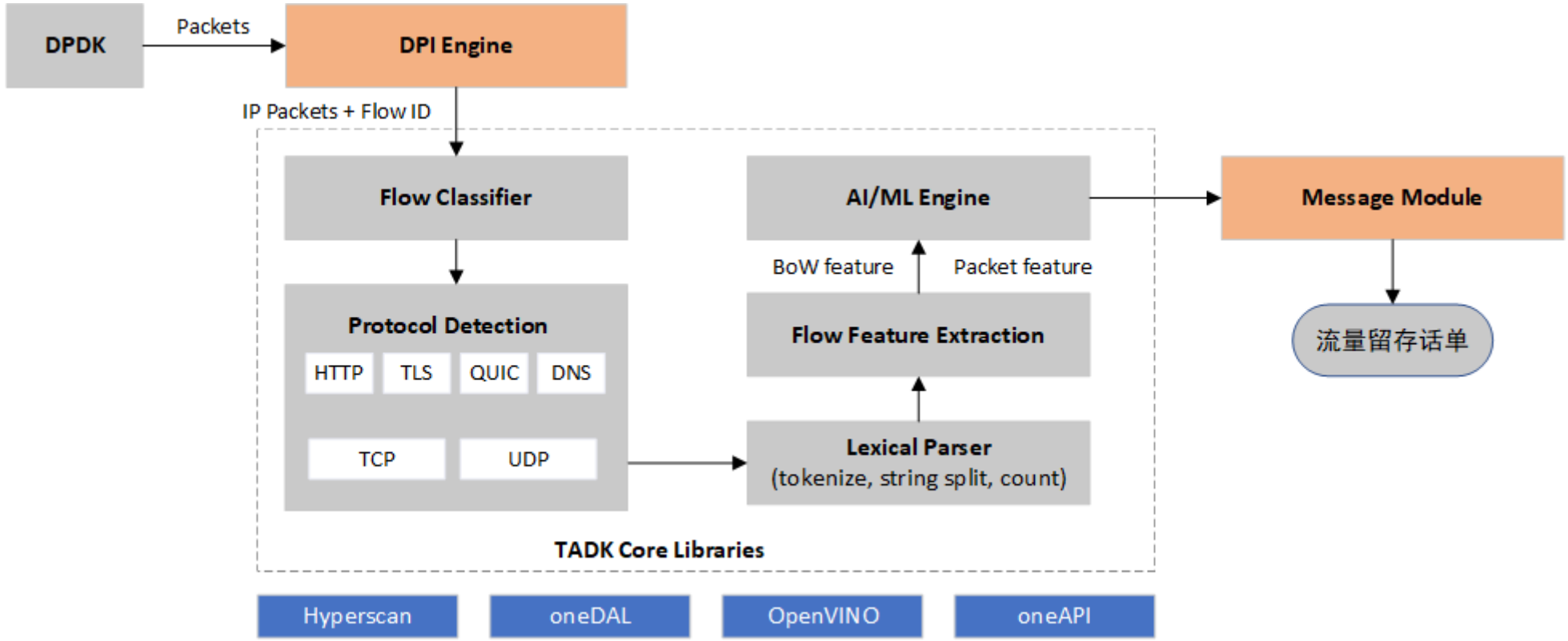
流量吞吐能力：CPU环境所有进程每秒可检测在线流量总量近15万条流量

OneAPI加速测试方案

测试环境：Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz，GPU 无

测试流程：

1. 准备电商类APP数据集(包含淘宝、天猫、京东和拼多多四类)共计8000条流量，按照80/20的百分比比例切分成训练集/测试集；
2. 选取一个合适的机器学习算法，此次以测试速度为主，选取随机森林分类模型作为测试模型；
3. 在两个环境完成训练和推理的耗时测试：
   (1) 环境一：Scikit-learn官方机器学习库，使用Random Forest Classifier，使用CPU测试；
   (2) 环境二：OneAPI官方算法库，内置Decision Forest Classification，使用CPU测试。



OneAPI加速测试结果

恒安嘉新 EVERSEC

训练耗时对比：

轮次	环境一/秒	环境二/秒
1	254.946	146.567
2	252.893	149.431
3	255.228	150.899
4	254.426	147.449
5	256.599	143.754
6	251.623	143.547
7	257.491	148.955
总计	1783.206	1030.602
平均	254.743	147.228

推理耗时对比：

对测试集1600条流量进行推理，记录整体耗时，均采用随机森林分类器的单条推理模式，推理采用单进程，分别对环境一和环境二进行测试。

环境一/秒	环境二/秒
10.234	0.546

结论：

模型训练时，OneAPI相比于Scikit-learn版的Random Forest Classifier于速度上有**73.02%**的提升；

模型推理时，OneAPI的处理速度为Scikit-learn版Random Forest Classifier的**18倍**。

OneAPI课题应用效果

经过测试，应用OneAPI可以提升**数十倍**的运算速度，降低推理耗时，效果显著；

将加密流量识别开发环境中Random Forest模型同步升级适配OneAPI，每种协议类型各自训练对应的模型，分别准备1000条测试数据，记录每个模块的运行耗时后计算平均值用以对比，替换前后性能对比结果如下表所示。

模块	无OneAPI加速/毫秒	有OneAPI加速/毫秒	处理速度提升倍数
1 - TLS流量分类模块	6210	341	17
2 - HTTP流量分类模块	7200	338	20
3 - TCP流量分类模块	6021	158	37
4 - UDP流量分类模块	5932	146	40

加密流量识别系统适配OneAPI后，推理耗时同比降低**数十倍**，秒处理量由原先的近7千提升至约15万。

让通信值得信赖
让安全创造价值