数据驱动安全

2015 中国互联网安全大会
China Internet Security Conference

深度学习
在流量识别中的应用

王占一

2015.9.30

Black Hat 2015参会议题

– The Applications of Deep Learning on Traffic Identification

black hat USA 2015

THE APPLICATIONS OF DEEP LEARNING ON TRAFFIC IDENTIFICATION

Generally speaking, most systems of network traffic identification are based on features. The features may be port numbers, static signatures, statistic characteristics, and so on. The difficulty of the traffic identification is to find the features in the flow data. The process is very time-consuming. Also, these approaches are invalid to unknown protocol. To solve these problems, we propose a method that is based on neural network and deep learning a hotspot of research in machine learning. The results show that our approach works very well on the applications of feature learning, protocol identification, and anomalous protocol detection.

PRESENTED BY

Zhanyi Wang & Chuanming Huang & Zhuo Zhang & Bo Liu

THE APPLICATIONS OF DEEP LEARNING ON TRAFFIC IDENTIFICATION

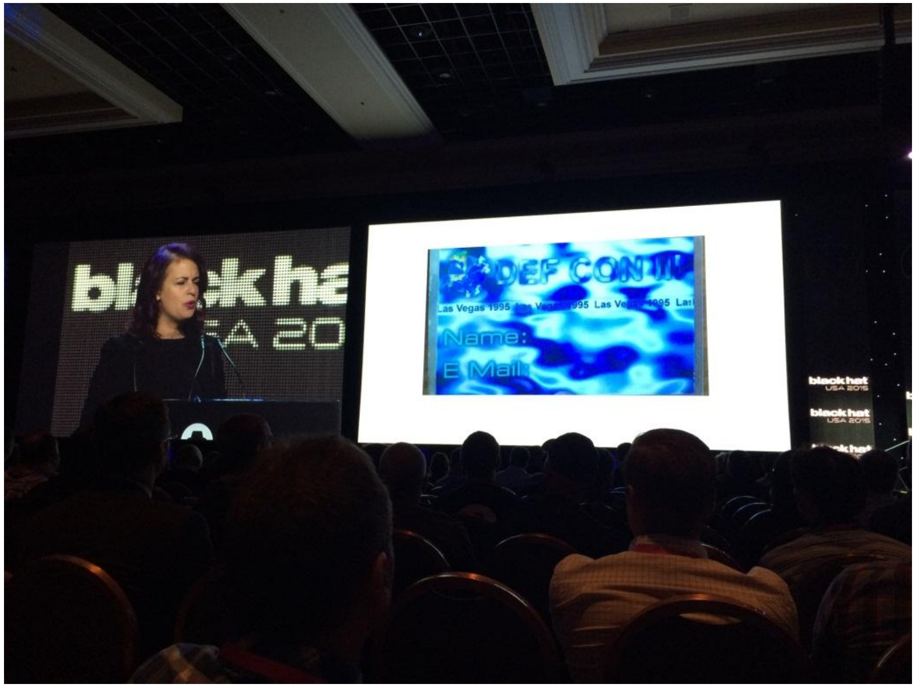
Zhanyi Wang & Chuanming Huang & Zhuo Zhang & Bo Liu

Jasmine Ballroom

09:00 – 09:25

Black Hat 2015

- 2015.08.01-08.06
- Las Vegas, NV



Black Hat 2015大数据与机器学习相关议题

blackhat USA 2015

BRIEFINGS DAY 1 — WEDNESDAY | AUGUST 5

ROOM →	Level 2: Lagoon K	Level 2: Mandalay Bay BCD	Level 2: Mandalay Bay EF	Level 2: Mandalay Bay GH	Level 3: Jasmine Ballroom	Level 3: South Seas ABE	Level 3: South Seas CDF	Level 3: South Seas GH	Level 3: South Seas IJ
07:00-19:00	Registration // Black Hat Blvd								
08:00-08:50	Breakfast // Shoreline B — Sponsored by: FireEye, Lieberman Software, Qualys, RSA, tenable network security								
08:50-09:00	Welcome & Introduction to Black Hat USA 2015 // Mandalay Bay Ballroom								
09:00-10:00	Keynote Speaker // Jennifer Granick // Mandalay Bay Ballroom								
10:00-10:20	Break								
10:20-11:10	**How to Hack Government: Technologists as Policy Makers** *by Ashkan Soltani + Terrell McSweeny*	**Internet Plumbing for Security Professionals: The State of BGP Security** *by Wim Remes*	**Writing Bad @$$ Malware for OS X** *by Patrick Wardle*	**Android Security State of the Union** *by Adrian Ludwig*	**Server-Side Template Injection: RCE for the Modern Web App** *by James Kettle*	**Bring Back the Honeypots...** *by Haroo Meer + Marco Slaviero*	**Why Security Data Science Matters and How It's Different: Pitfalls and Promises of Data Science Based Breach Detection and Threat Intelligence** *by Joshua Saxe*	**pread Spectrum S tcom Hacking: A tacking the GlobalStar S mplex Data Service** *by Colby Moore*	**Unicorn: Next Generation CPU Emulator Framework** *by Nguyen Anh Quynh + Hoang-Vu Dang*
11:10-11:30	Coffee Service // Level 2, 3, Microsoft Business Hall Networking Lounge — Sponsored by: FireEye, Lieberman Software, Qualys, RSA, tenable network security								
11:30-12:20	**Breaking HTTPS with BGP Hijacking** *by Artyom Gavrichenkov*	**Attacking Interoperability - An OLE Edition** *by Haifei Li + Bing Sun*	**Defeating Pass-the-Hash: Separation of Powers** *by Seth Moore + Baris Saydag*	**Winning the Online Banking War** *by Sean Park*	**Emanate Like a Boss: Generalized Covert Data Exfiltration with Funtenna** *by Ang Cui*	**Take a Hacker to Work Day - How Federal Prosecutors Use the CFAA** *by Leonard Baile*	**Why Security Data Science Matters and How It's Different: Pitfalls and Promises of Data Science Based Breach Detection and Threat Intelligence** *by Joshua Saxe*	**he Battle for Free S eech on the Internet** *b Matthew Prince*	**Understanding and Managing Entropy Usage** *by Bruce Potter + Sasha Wood*
12:20-13:50	Lunch Br								
13:50-14:40	**Data-Driven Threat Intelligence: Metrics on Indicator Dissemination and Sharing** *by Alexandre Sieira + Alex Pinto*	**dventures in Femtoland: 50 Yuan for Invaluable Fun** *b Alexey Osipov + Alexander aitsev*	**Red vs. Blue: Modern Active Directory Attacks, Detection, and Protection** *by Sean Metcalf*	**GameOver Zeus: Badguys and Backends** *by Elliott Peterson + Michael Sandee + Tillman Werner*	**Exploiting the DRAM Rowhammer Bug to Gain Kernel Privileges** *by Mark Seaborn + Halvar Flake*	**SMBv2: Sharing More than Just Your Files** *by Jonathan Brossard + Hormazd Billimoria*	**Abusing Silent Mitigations - Understanding Weaknesses Within Internet Explorer's Isolated Heap and Memory Protection** *by Brian Gorenc + Abdul-Aziz Hariri + Simon Zuckerbraun*	**The Tactical Application Security Program: Getting Stuff Done** *by Cory Scott + David Cintz*	**These are Not Your Grand Daddy's CPU Performance Counters - CPU Hardware Performance Counters for Security** *by Nishad Herath + Anders Fogh*

LEGEND

- Crypto
- Defense
- Enterprise
- Exploit Development
- Forensics/Incident Response
- Hardware/Embedded
- Human Factors
- Internet of Things
- Malware
- Mobile
- Network
- OS, Host and Container Security
- Panels
- Reverse Engineering
- Risk Management/Compliance
- Security Development Lifecycle
- Smart Grid/Industrial Security
- Virtualization
- Web AppSec

Schedule as of July 20, 2015. Subject to Change.

WWW.BLACKHAT.COM #BHUSA

数据

Black Hat 2015大数据与机器学习相关议题

Title	Speaker
Data-Driven Threat Intelligence: Metrics on Indicator Dissemination and Sharing	Alex Pinto
Why Security Data Science Matters and How It's Different: Pitfalls and Promises of Data Science Based Breach Detection and Threat Intelligence	Joshua Saxe
Graphic Content Ahead: Towards Automated Scalable Analysis of Graphical Images Embedded in Malware	Alex Long
Distributing the Reconstruction of High-Level Intermediate Representation for Large Scale Malware Analysis	Rodrigo Branco
Securing Your Big Data Environment	Ajit Gaddam
Defeating Machine Learning: What Your Security Vendor is Not Telling You	Bob Klein
From False Positives to Actionable Analysis: Behavioral Intrusion Detection, Machine Learning, and the SOC	Joseph Zadeh
The Applications of Deep Learning on Traffic Identification	Zhanyi Wang
Internet-Scale File Analysis	Zachary Hanif Tamas
Deep Learning on Disassembly	Matt Wolff

内容提要

- 流量识别的传统方法
- 神经网络和机器学习
- 具体应用
  - 协议分类
  - 未知协议识别
  - 特征的自动学习
  - 应用程序识别
- 总结和展望

流量识别的传统方法（一）

- 将流量准确地映射到某种协议或应用
  - 是网络安全的基础
  - 对异常检测、安全管理作用重大
- 基于预定义或特殊端口
  - 标准HTTP端口：80
  - 默认SSL端口：443
  - 缺点：非标准端口或新定义的端口不适用
- 基于DPI和统计特征的流量识别
  - 根据经验和规则确定的特征字/指纹/序列
  - 缺点：既耗时又耗力

HTTP?

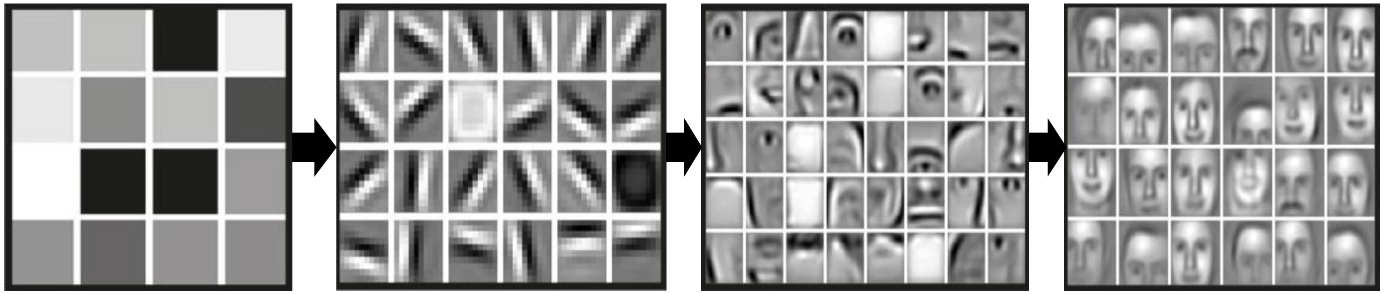
SSL?

流量识别的传统方法（二）

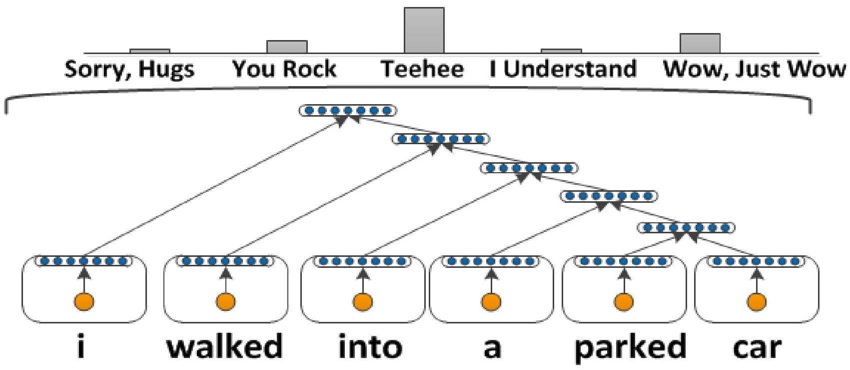
- 基于行为特征和机器学习
  - 优点：建模和识别过程自动化
  - 难点：特征抽取和选择依赖于如何选择特征？
- 有没有不依赖于专家的方法？
- 非监督的特征学习是否可行？
- 答案
  - 人工智能领域的深度学习技术

火热的深度学习技术

- 图像

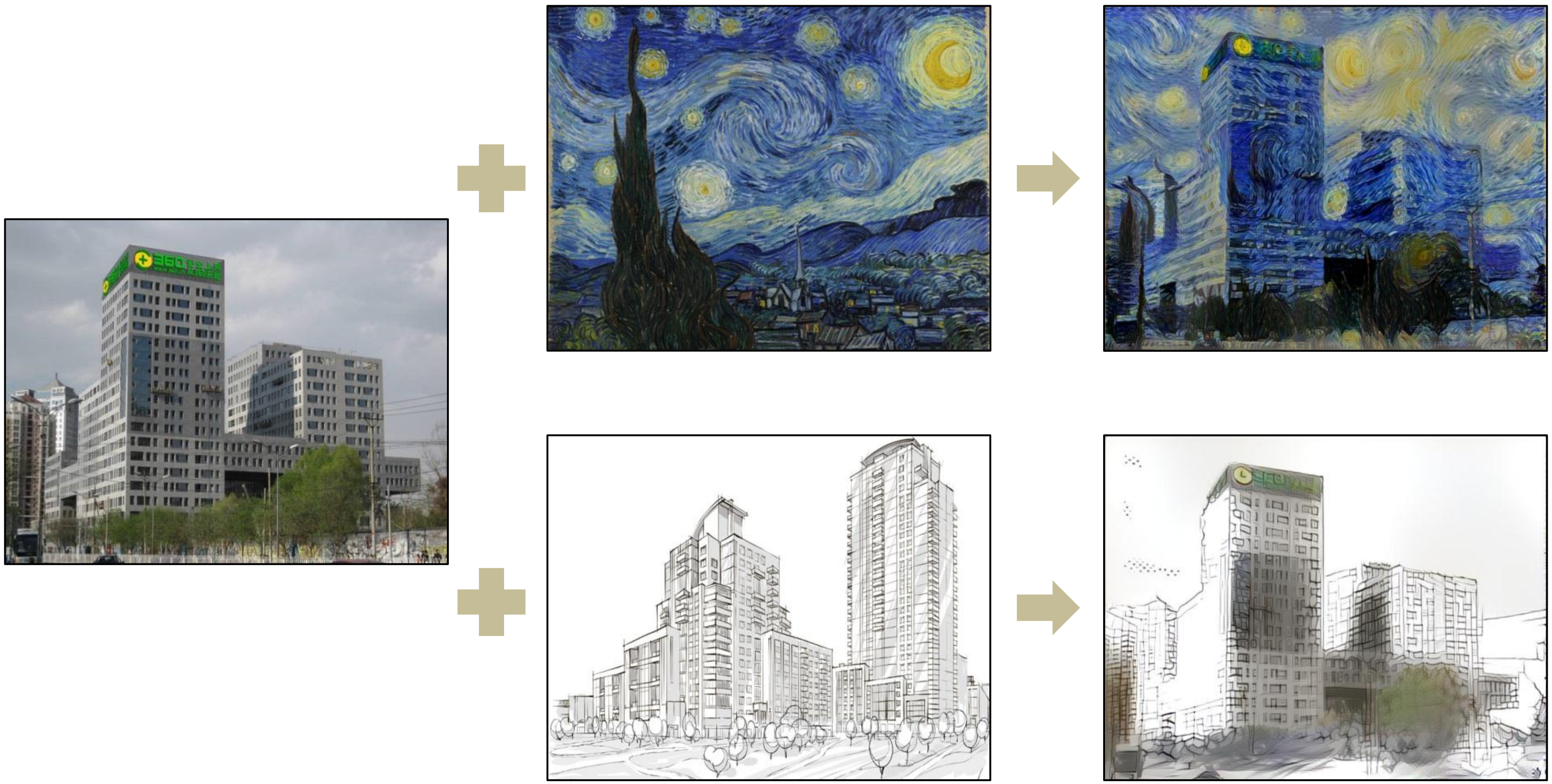


- 自然语言处理



- 语音

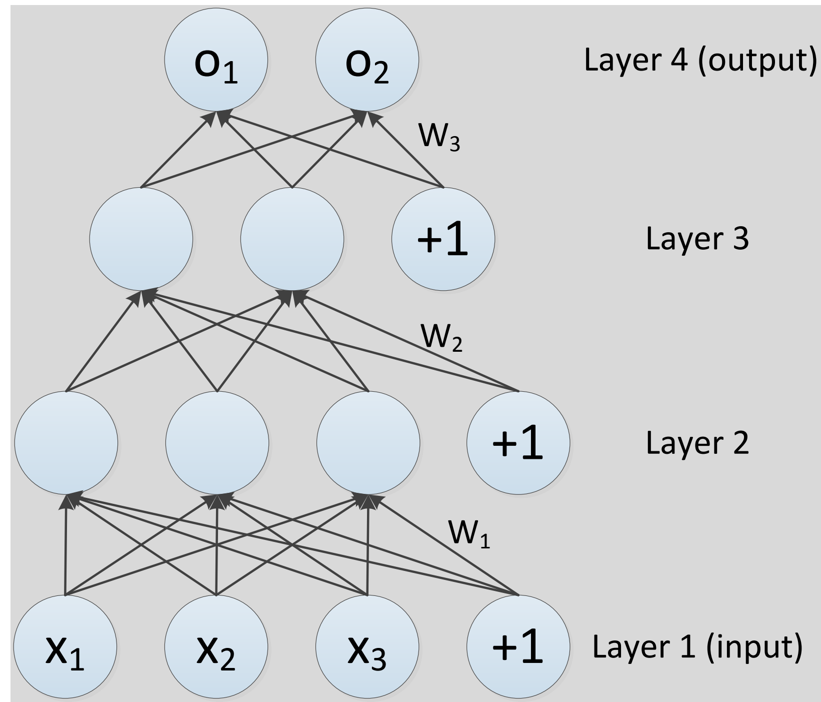
深度学习技术的应用



- Gatys, L. A. (2015). A Neural Algorithm of Artistic Style. arXiv preprint arXiv:1508.06576.

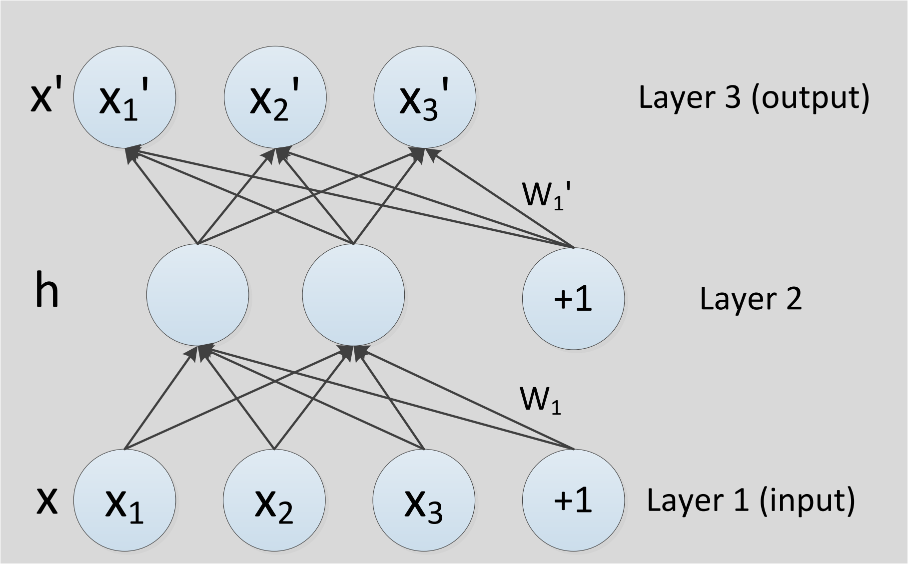
神经网络

- 人工神经网络
- 基本单元
  - 神经元
- 结构
  - 输入层
  - 隐藏层
  - 输出层
- 相邻层的神经元 彼此相连
- 同层的神经元 不直接相连



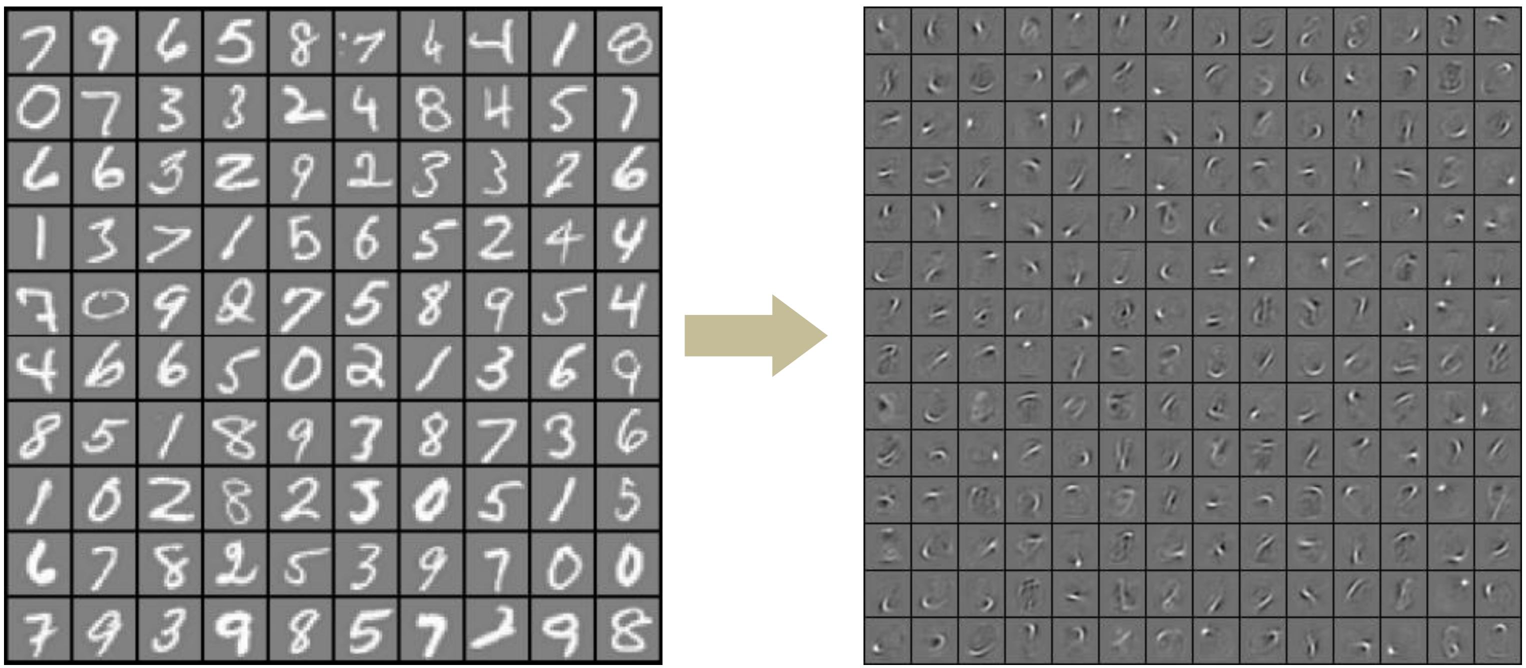
自编码(Auto-Encoder)网络

- 一种特殊的神经网络
- 只有一个隐藏层
- 输出层与输入层完全相同！



自编码在图像识别中的应用

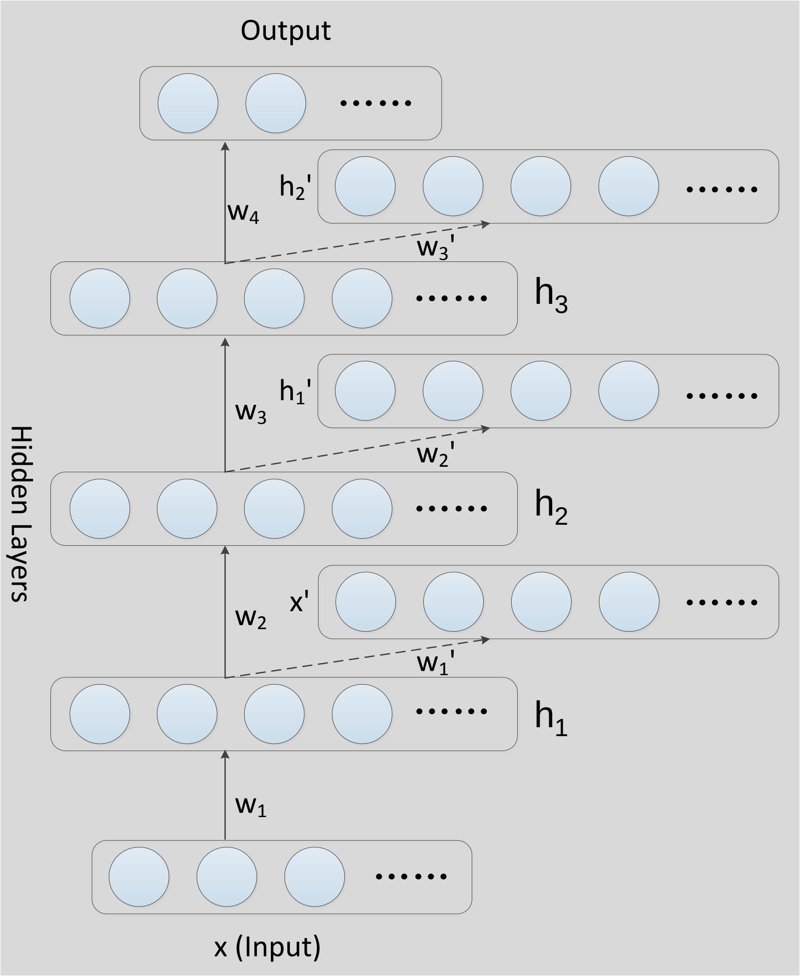
- 手写体数字识别



栈式自编码(Stacked Auto-Encoder)

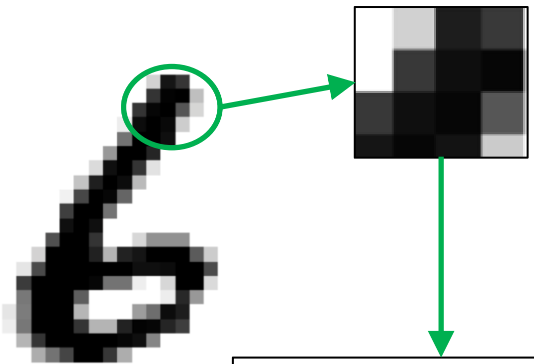
- 栈式自编码(SAE)
- 由多个自编码网络
- SAE本质上也是一种络

- 采用逐层贪婪训练
- 使用微调(fine-tuning)



图像 VS Payload数据

- 是否有相似之处？



255	210	21	53 …
255	52	3	0 …
52	6	0	85 …
…	…	…	…

TCP flow Payloads

474554206874……727665720020……732048545450……33a31353a323……

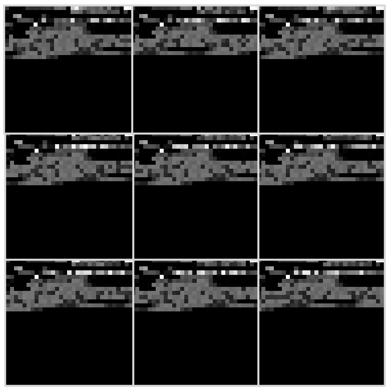
732048545450……33a31353a323……

115 32 72 84 84 80……51 163 19 83 163 35……

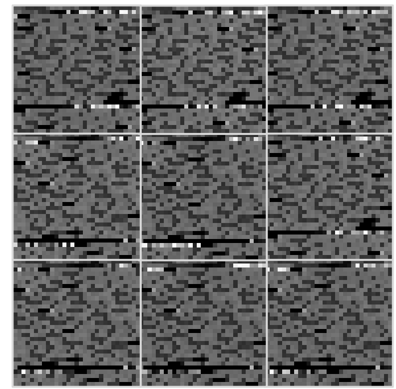
数值范围相同：[0,255]
256个数字!

协议流量→图像

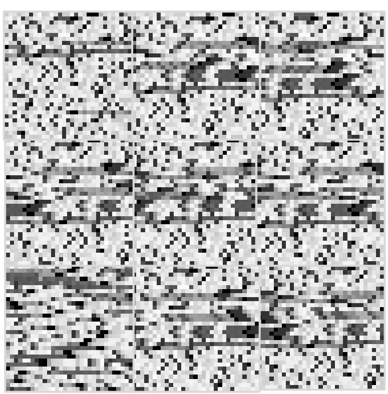
MySQL



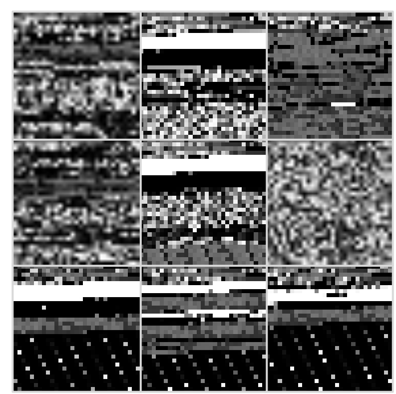
SSH



Whois-DAS

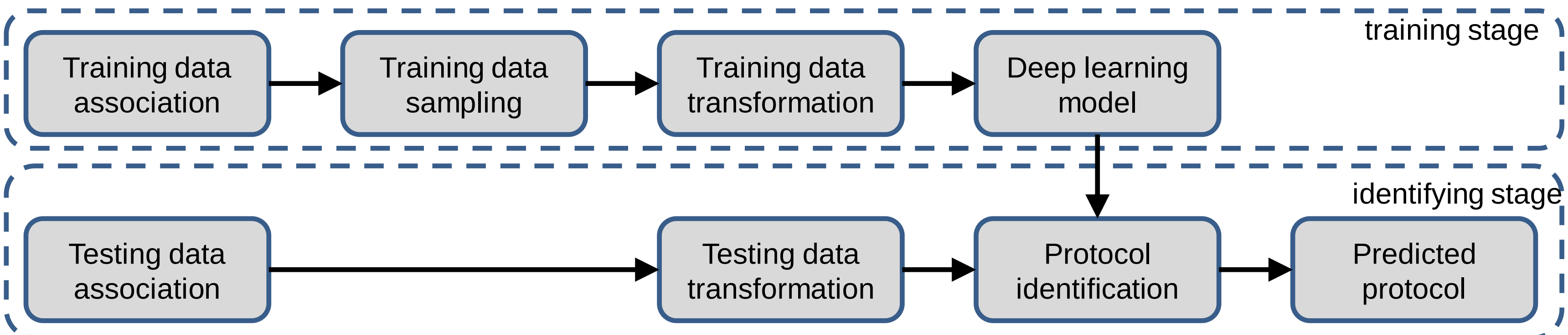


BitTorrent



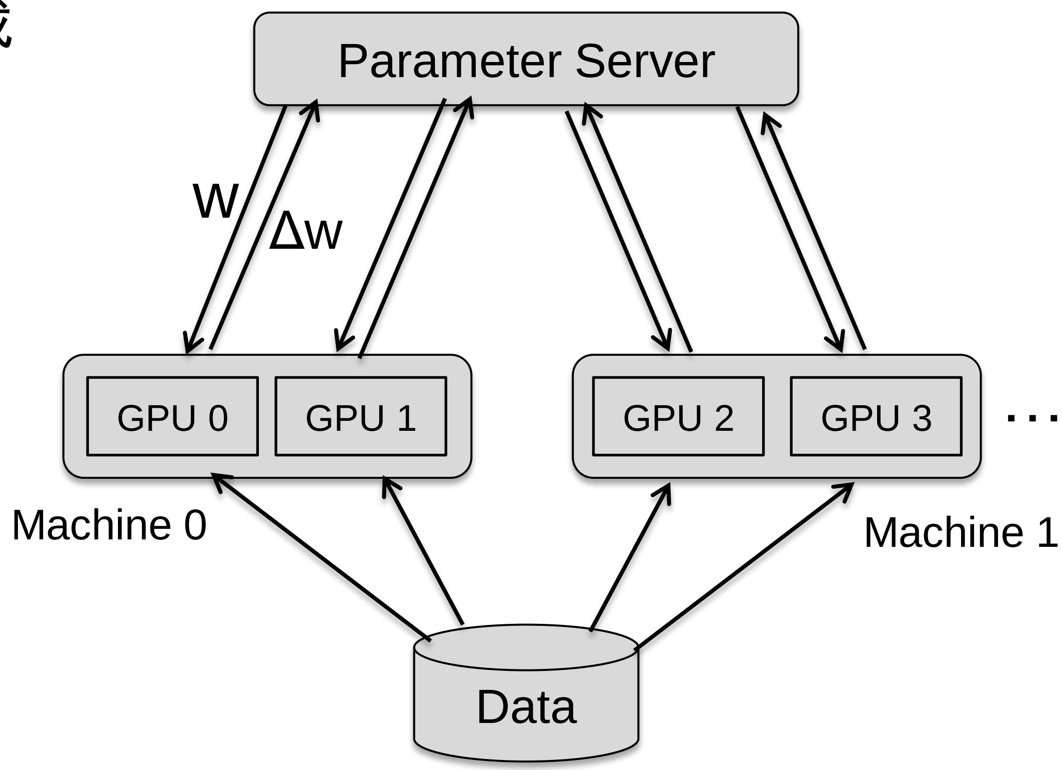
协议识别的实现过程

- 数据采集自公司内网
- 实验环境
  - 框架1 - CPU集群: 2~10台服务器
  - 框架2 - CPU + 4GPU
  - 训练时间 - 天->分钟



基于多GPU的并行计算

- 训练时间的需求
  - 用CPU需要几天完成
- GPU矩阵计算
- 大量的模型参数
  - 500,000以上
- 大规模的数据
  - 存储的需求
- 解决方法
  - 多机并行
  - 多GPU并行
  - OpenCL框架



协议分类结果

- 宏观准确率>99%
- 平均准确率97.9%

Protocol	Precision	Protocol	Precision
SMB	1.0000	RSYNC	0.9987
DCE_RPC	1.0000	Redis	0.9985
NetBIOS	1.0000	FTP_CONTROL	0.9970
TDS	1.0000	HTTP_Connect	0.9967
SSH	0.9996	SMTP	0.9949
Kerberos	0.9996	Whois-DAS	0.9943
LDAP	0.9996	IMAPS	0.9814
BitTorrent	0.9992	Apple	0.9640
MySQL	0.9989	SSL	0.9513
DNS	0.9989	HTTP_Proxy	0.9174

未知协议识别

- 随机选取10,000条被传统方法标记为“unknown”的记录
- 识别率：
- 0%

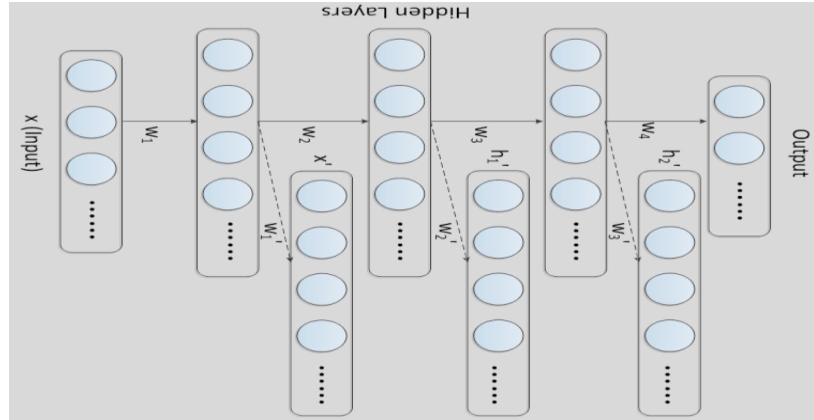
↓

- 63.37%

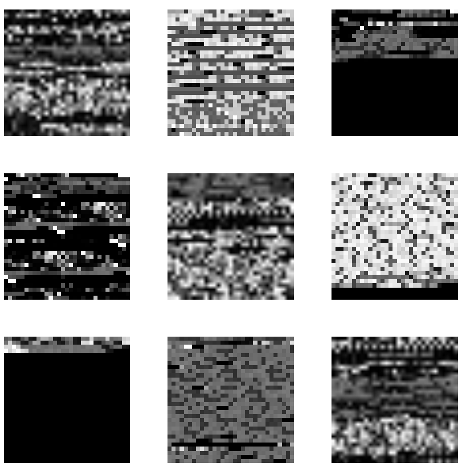
	number	ratio
SSL	1956	29.12%
DCE_RPC	1454	21.65%
Skype	873	13.00%
Kerberos	517	7.70%
MSN	360	5.36%
Google	311	4.63%
DNS	260	3.87%
RTMP	234	3.48%
TDS	202	3.01%
H323	170	2.53%

特征的自动学习

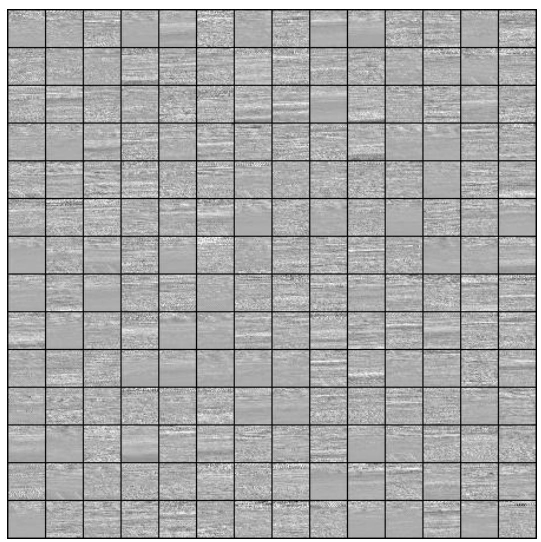
- 特征抽取



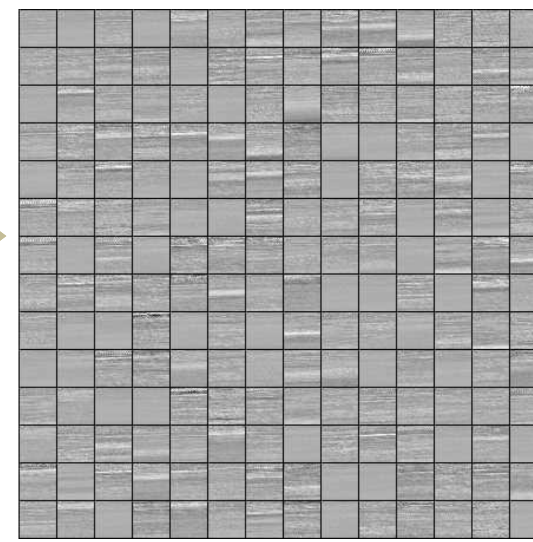
原始流量图像



1层AE的特征

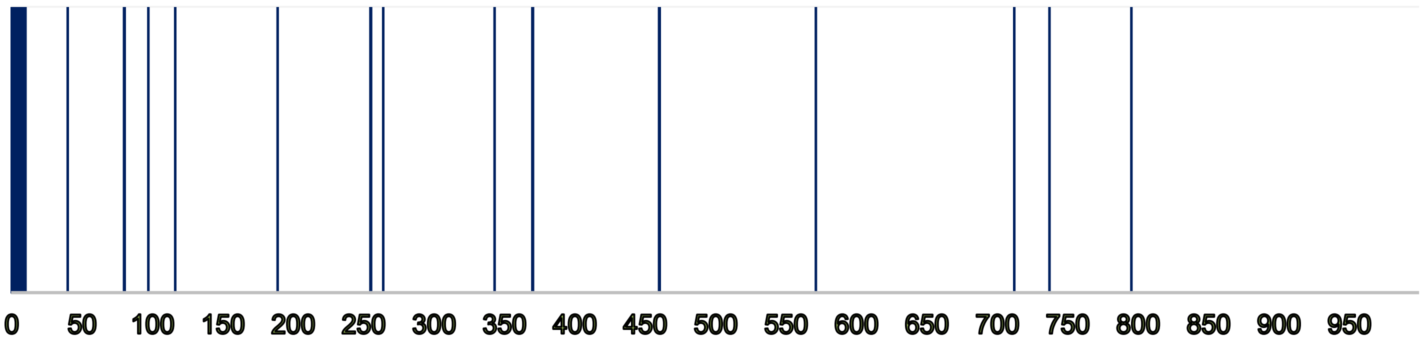


2层AE的特征



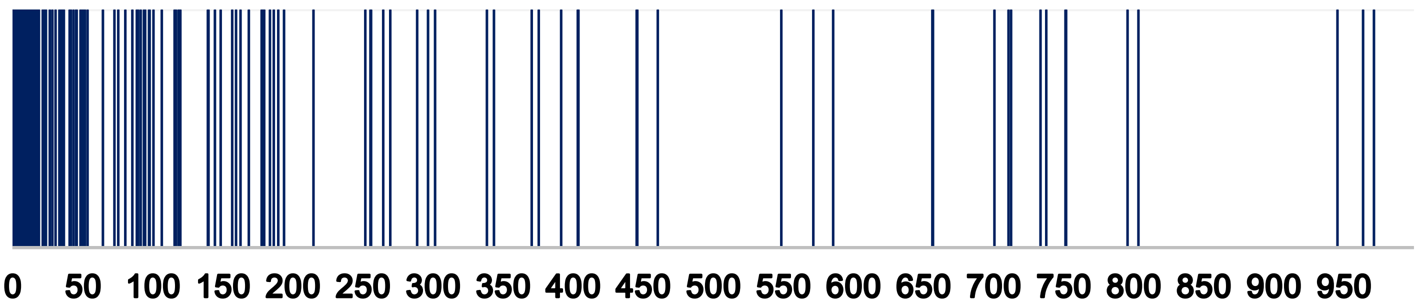
特征的自动学习

- 特征选择
  - A：最重要的25个字节



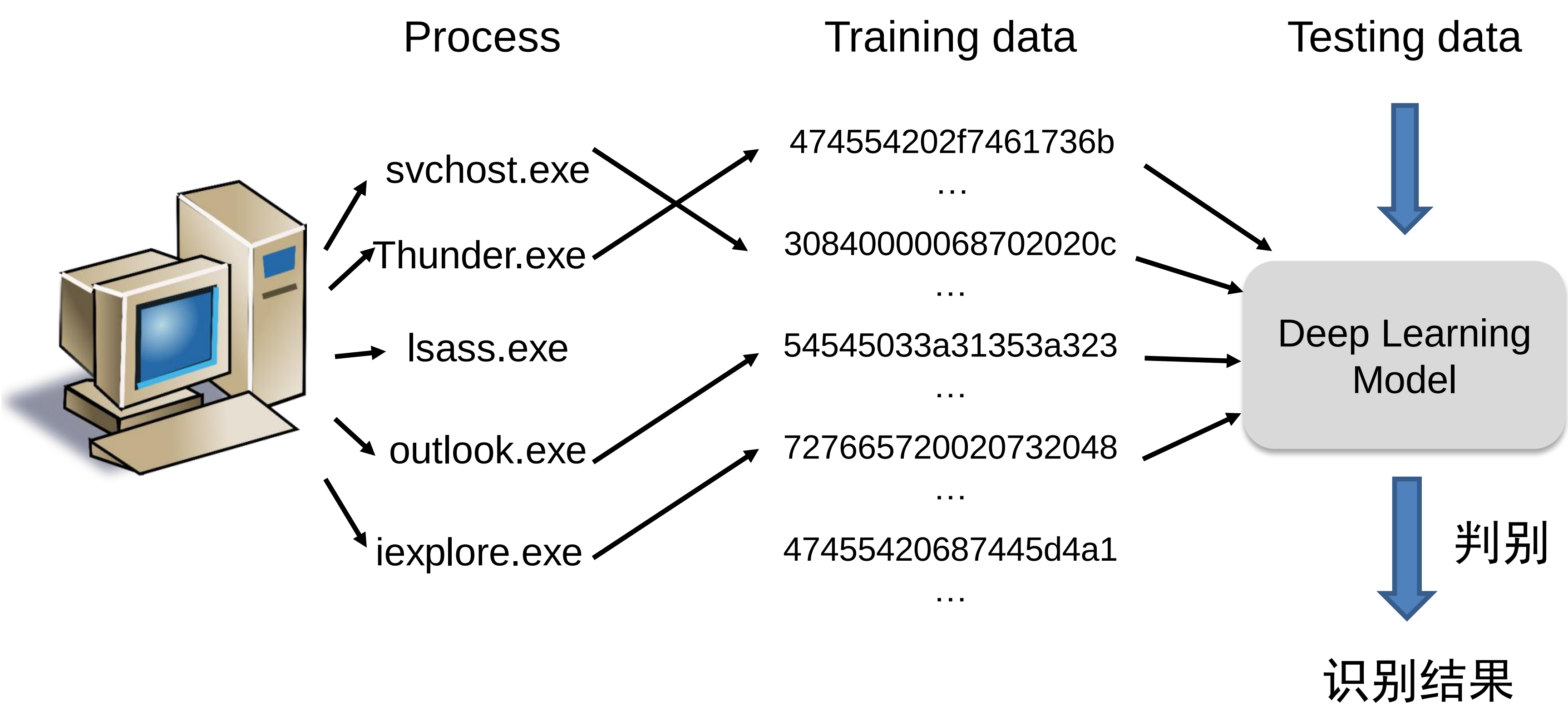
特征的自动学习

- 特征选择
  - B：最重要的100个字节



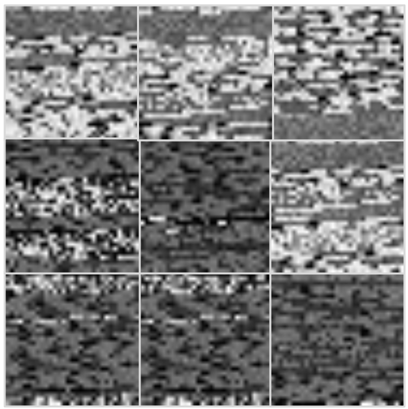
  - C：最不重要的300个字节

应用程序识别

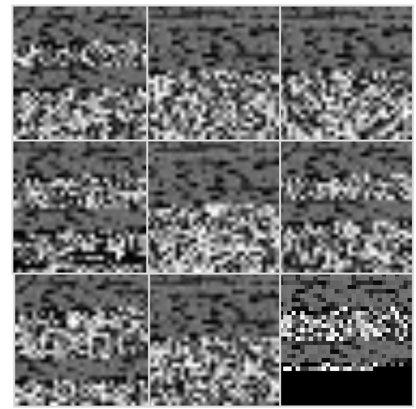

Process
Training data
Testing data
svchost.exe
Thunder.exe
lsass.exe
outlook.exe
iexplore.exe
474554202f7461736b
…
30840000068702020c
…
54545033a31353a323
…
727665720020732048
…
47455420687445d4a1
…
Deep Learning Model
判别
识别结果

应用程序流量→图像

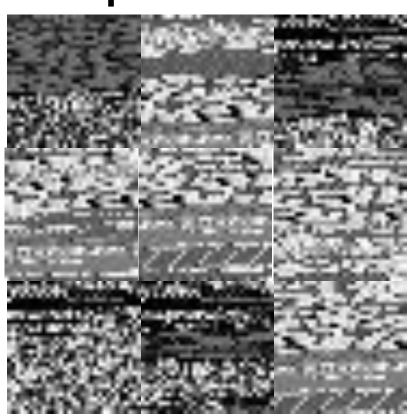
QQ.exe



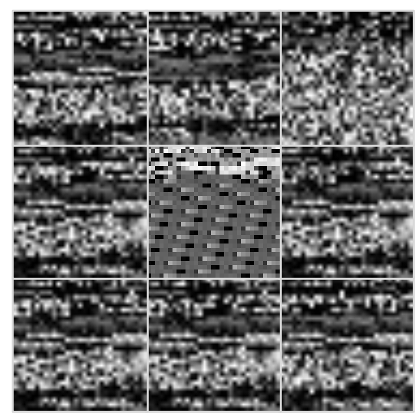
wechat.exe



iexplore.exe



outlook.exe



识别结果

- 训练数据中包含几百种应用
- 宏观准确率>96%，平均准确率>90%

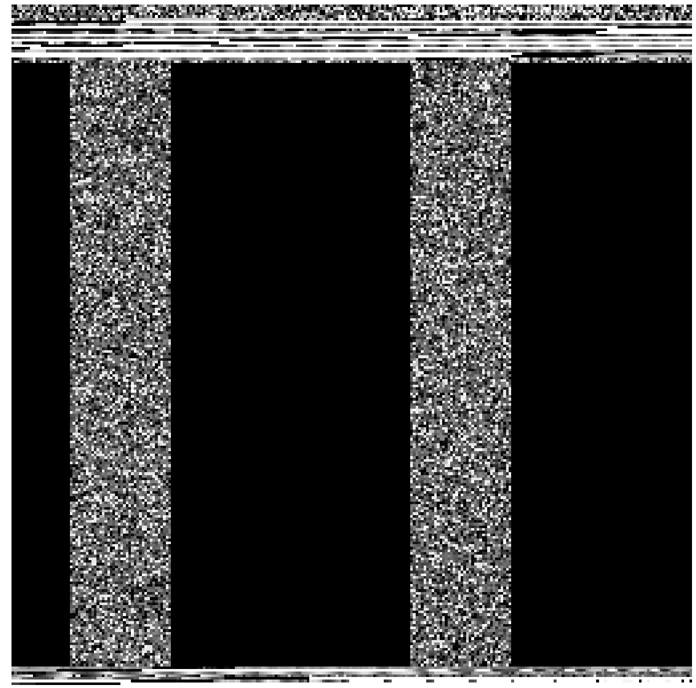
Application	Precision	Protocol	Precision
foxmail.exe	1.0000	**xshell.exe**	0.9813
wpservice.exe	1.0000	**baidumusic.exe**	0.9808
taobaoprotect.exe	0.9984	**fetion.exe**	0.9779
wechat.exe	0.9983	**qqmusic.exe**	0.9730
liebao.exe	0.9978	**qqdownload.exe**	0.9615
weibo2015.exe	0.9974	**yodaodict.exe**	0.9542
lsass.exe	0.9945	**itunes.exe**	0.9429
sogoucloud.exe	0.9897	**outlook.exe**	0.9219
qq.exe	0.9884	**thunder.exe**	0.9168
pplive.exe	0.9870	**iexplore.exe**	0.8860

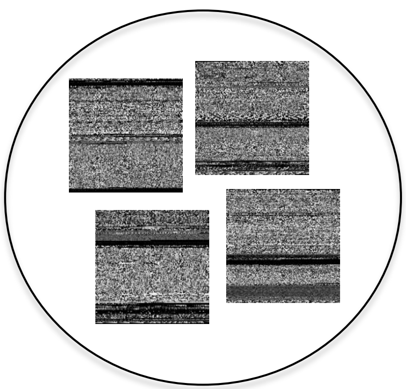
小结

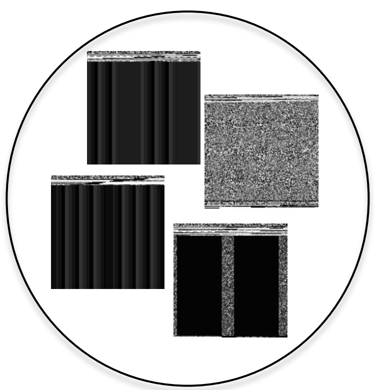
- 深度学习在流量识别中的应用
  - 通过网络流量识别协议、应用程序
  - 特征的自动学习
  - 解决大数据的并行计算问题
- 价值
  - 减轻人工负担
  - 精度高
- 应用于网络安全领域的难点——一头一尾
  - 输入：非传统的语音/图像/文本
  - 输出：安全领域往往要求更精准
- 展望
  - 算法“深”，应用“广”

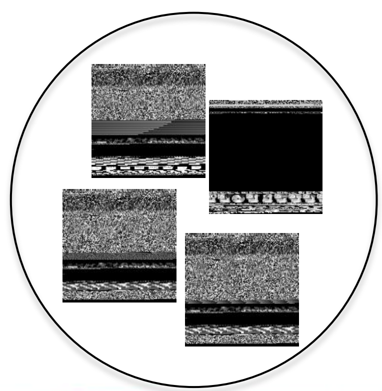
恶意代码样本→图像

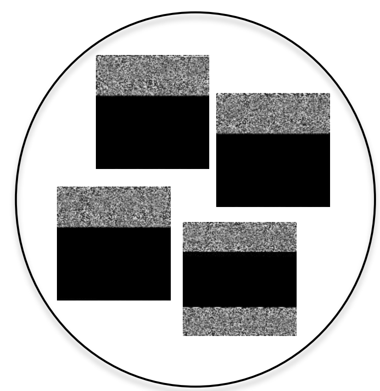
```
00401020 60 60 00 33 C5 89 85 98 01 00 00 8B 85 A4 01 00
00401030 00 53 56 8B B5 AC 01 00 00 57 6A 31 89 75 60 A3
00401040 88 70 60 00 C7 05 7C 70 60 00 00 10 40 00 FF 15
00401050 68 90 5F 00 8B 0D 6C 70 60 00 51 FF 15 6C 90 5F
00401060 00 8D 55 38 52 8D 45 48 50 A1 6C 70 60 00 8D 4D
00401070 50 51 8D 55 40 52 50 FF 15 70 90 5F 00 8B 0D 6C
00401080 70 60 00 51 FF 15 74 90 5F 00 33 DB 53 53 FF 15
00401090 98 92 5F 00 8B 15 64 70 60 00 68 10 94 5F 00 68
004010A0 08 94 5F 00 52 FF 15 00 90 5F 00 68 64 70 60 00
004010B0 68 04 94 5F 00 68 01 00 00 80 FF 15 04 90 5F 00
004010C0 8B 45 8C 83 AD 74 FF FF FF 02 F7 D0 66 89 45 A4
004010D0 8B C6 89 5D 7C C7 45 6C 07 00 00 00 8D 50 01 90
004010E0 8A 08 40 84 C9 75 F9 66 8B 0D A2 72 60 00 FF 05
004010F0 4C 72 60 00 2B C2 66 F7 D1 66 89 0D 9E 72 60 00
00401100 3B C3 74 10 8A 06 3C 30 74 0A 80 7E 01 3A 74 04
00401110 3C 33 75 05 BB 01 00 00 00 66 8B 15 92 72 60 00
00401120 8B 0D 24 71 60 00 2B 0D E8 70 60 00 66 83 C2 35
00401130 6A 04 66 89 15 9A 72 60 00 8B 15 F4 70 60 00 23
00401140 15 F8 70 60 00 68 00 10 00 00 68 90 E3 1B 00 B8
00401150 DD 00 00 00 6A 00 C6 05 A6 72 60 00 B8 66 A3 CC
```



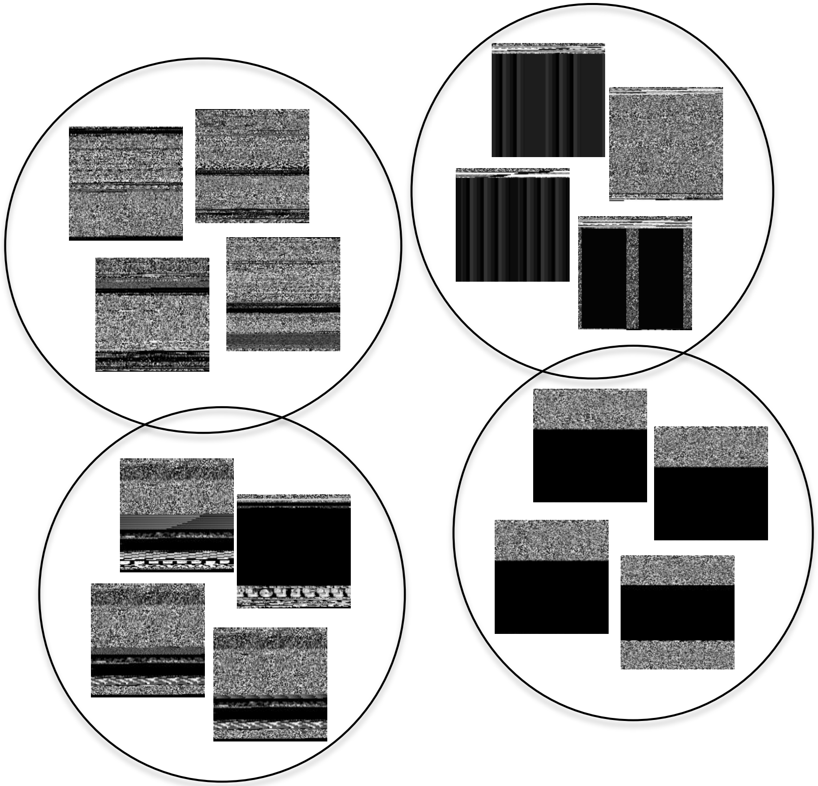


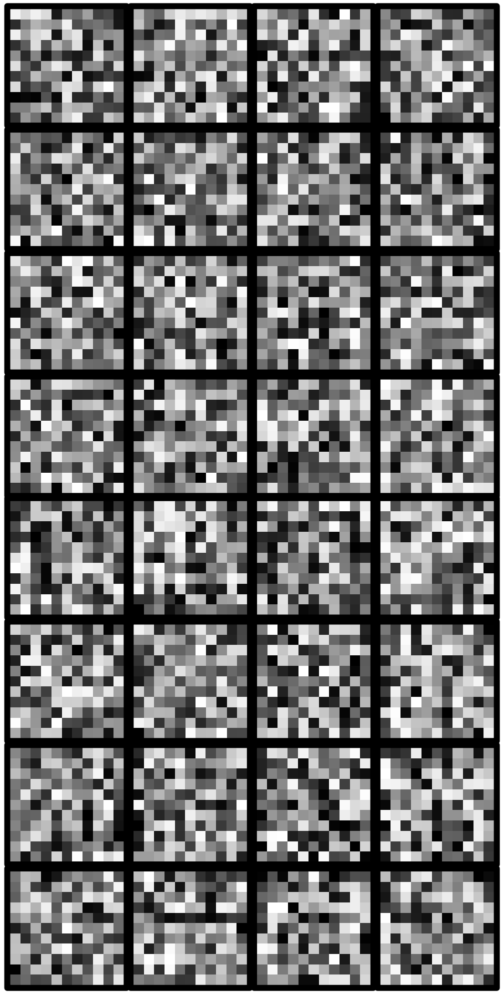






样本图像的深度学习（CNN）





谢谢!

wangzhanyi@360.cn