



# Malware Analysis Project

김현수(Hyeonsu Kim)

최가온(Gaon Choi)



# Contents

1. Previous works
2. Malware-image Transformation
3. Malware Binary Classification  
(malware or non-malware?)
4. Malware Family Classification
5. Other works







# Introduction

이번 프로젝트에서는 악성코드 파일을 이미지로 변환한 후,  
이미지를 기반으로 악성 여부를 판단하고,  
악성코드 계통(malware family)을 분류하는 인공지능 모델을  
설계하고 학습하였습니다.

# Previous works

지난 발표 시기까지의 프로젝트 진행 요약

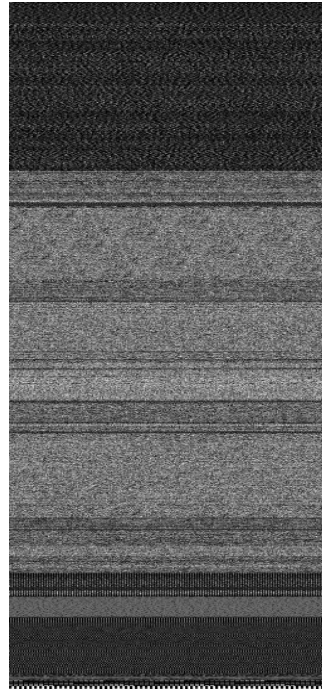
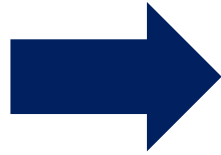


# Malware Image Visualization



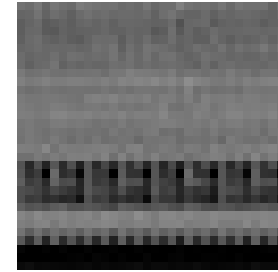
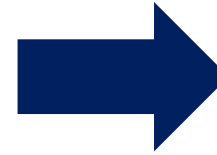
.vir file

`imageio.imwrite`



.png file

`Image.resize(32,32)`



.png file

# Binary classification training using VGGNet

malware: 500개, non-malware: 500개

Layer(type)	Output Shape	# of params
con2d(Conv2D)	(None, 32, 32, 3)	84
activation (Activation)	(None, 32, 32, 3)	0
vgg16 (Functional)	(None, None, None, 512)	14714688
flatten (Flatten)	(None, 512)	0
dense (Dense)	(None, 32)	16416
dense_1 (Dense)	(None, 1)	33

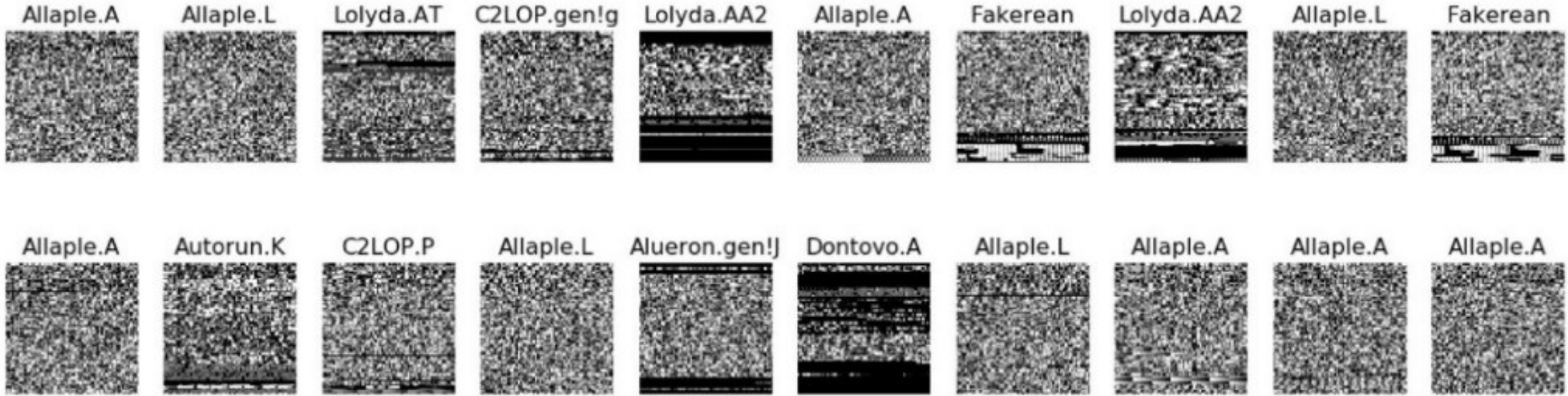
Performance	
Train loss	loss: 0.0291
Train acc	accuracy: 97.87
Test loss	loss: 0.2332
Test acc	accuracy: 69.50

# Malware Image Transformation

악성코드 파일 이미지화 및 데이터셋 분석



# Malware Image Transformation



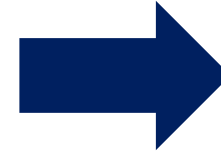
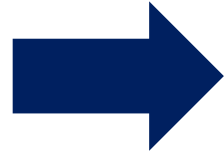
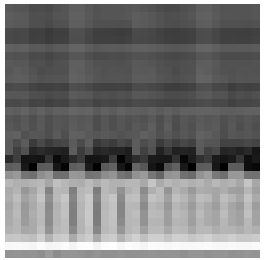
dataset	# of malware data points	# of non-malware data points
train set	7,000	3,000
test set1	5,000	5,000
test set2	5,000	5,000



# Malware Binary Classification

악성여부 이진분류 모델 설계 및 학습

# Malware Binary Classification



1: malware

0: non-malware

A model that receives a transformed 32 x 32 size image as input and dichotomously determines whether it is malware or not(non-malware)

# Malware Binary Classification

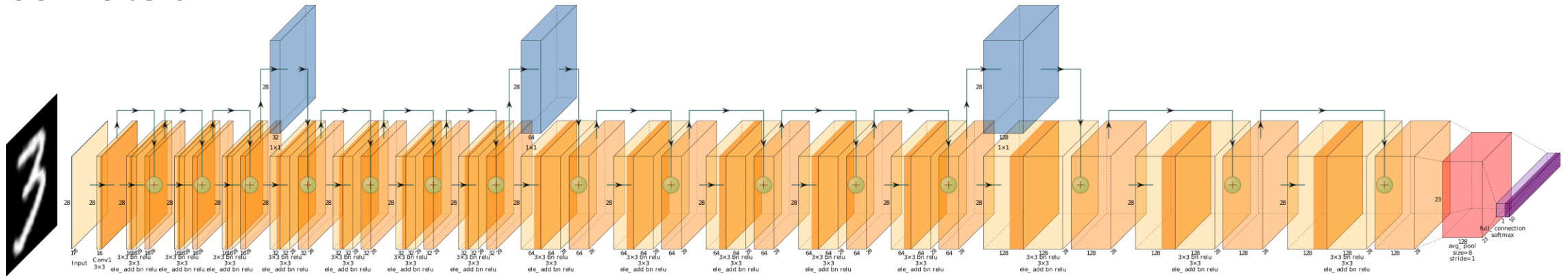
dataset	# of malware data points	# of non-malware data points
train set	7,000	3,000
test set1	5,000	5,000
test set2	5,000	5,000



dataset	# of malware data points	# of non-malware data points
train set	5,000	5,000
test set	5,000	5,000

# Malware Binary Classification

## ResNet50



Malware or Non-malware?  
Binary Classification

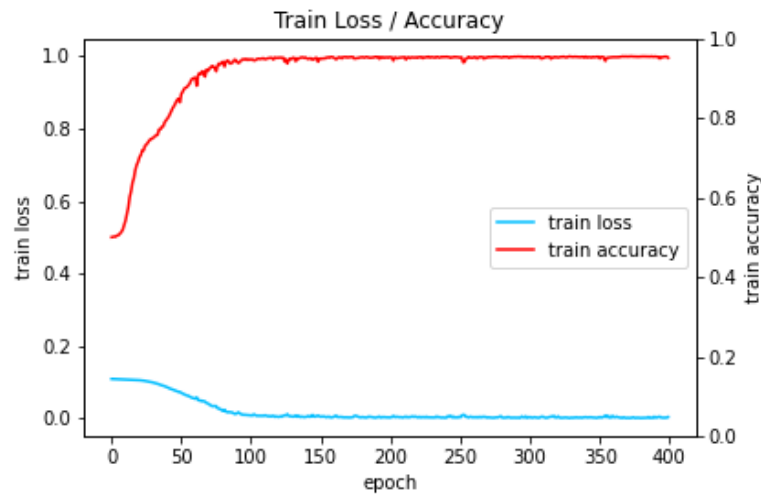


```
self.resnet_50.fc = nn.Sequential(  
    nn.Linear(2048, 1000),  
    nn.ReLU(inplace=True),  
    nn.Linear(1000, 256),  
    nn.ReLU(inplace=True),  
    nn.Linear(256, 64),  
    nn.ReLU(inplace=True),  
    nn.Linear(64, 1),  
    nn.Sigmoid()  
)
```

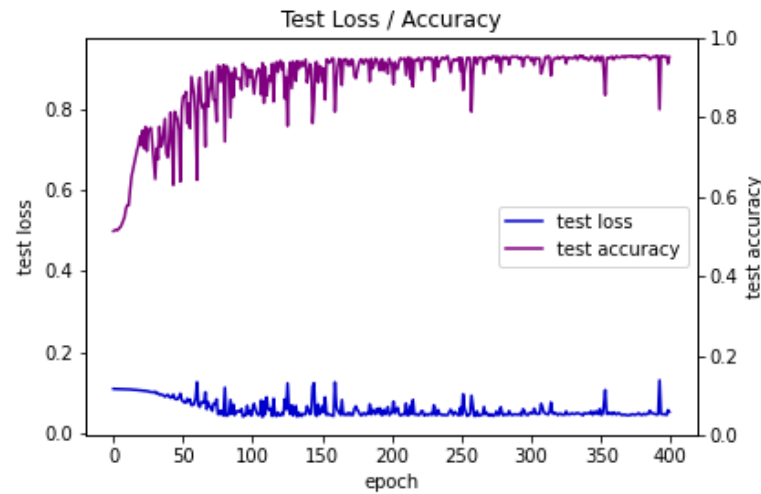


# Malware Binary Classification

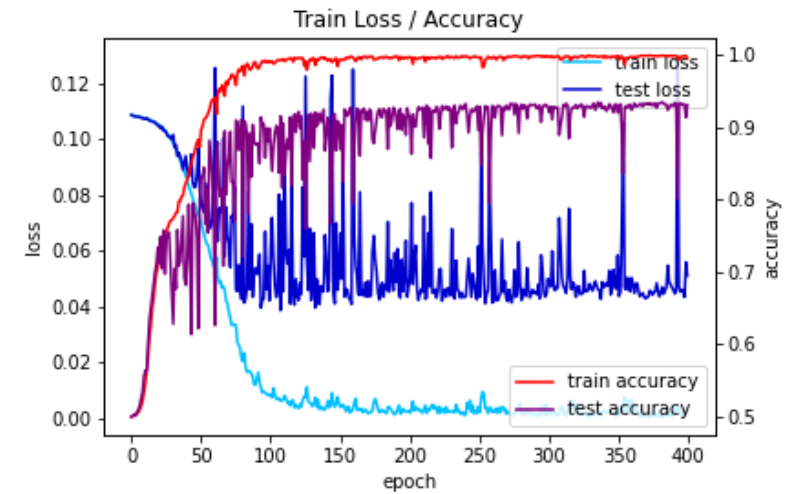
## Model Performance



- Train Accuracy: 99.5%
- Test Accuracy: 93.0%
- Epoch: 400

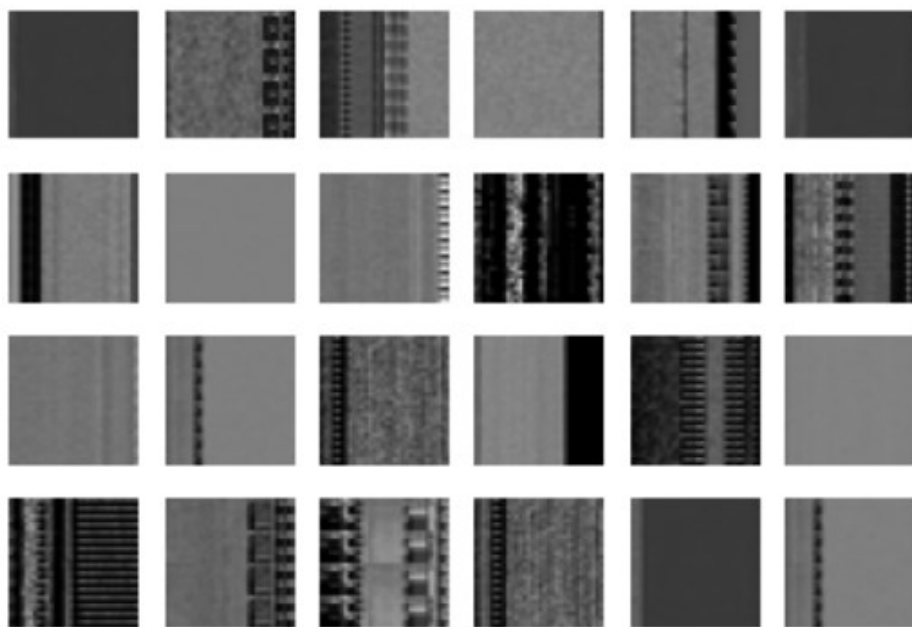


- Learning rate: 0.001
- Loss: nn.BCELoss
- Optimizer: SGD



# Malware Binary Classification

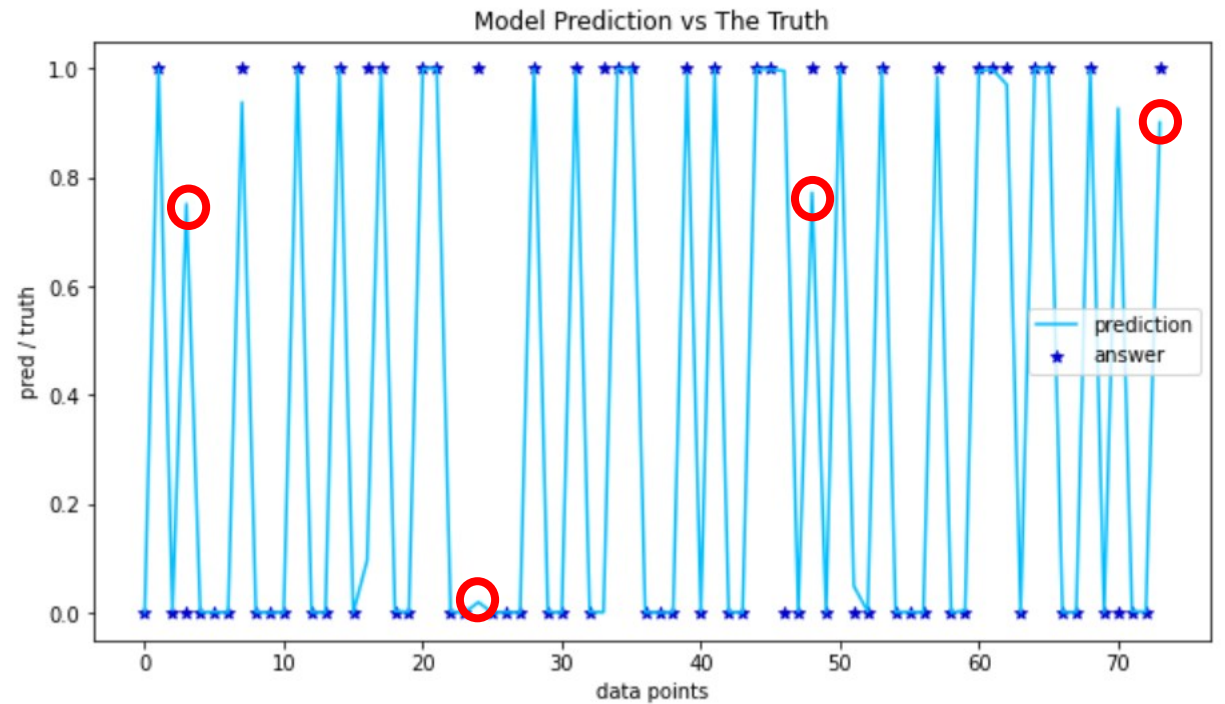
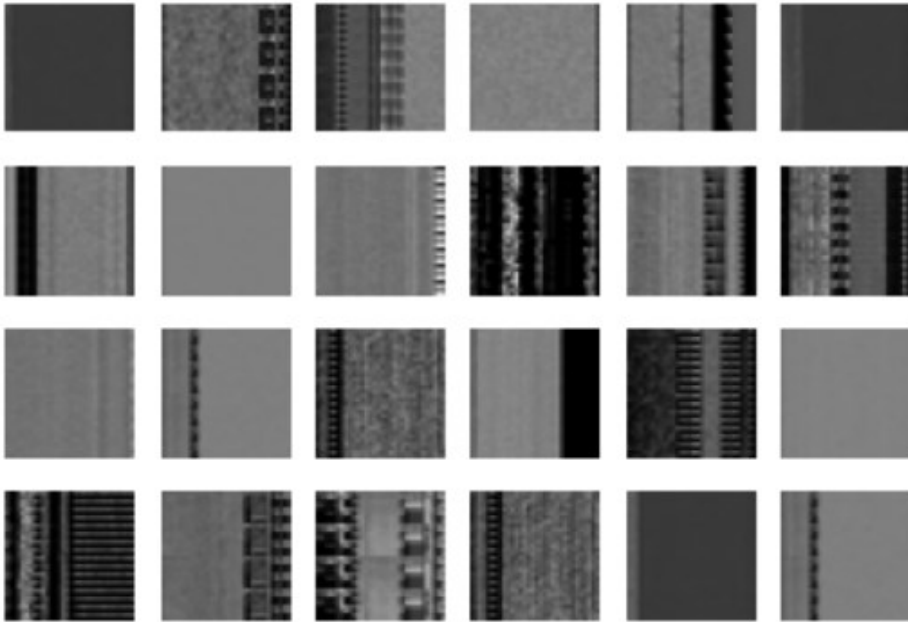
## Model Performance



```
image 0 | pred=0.0014788481639698148 | ans=0
image 1 | pred=0.9997912049293518 | ans=1
image 2 | pred=0.005231290124356747 | ans=0
image 3 | pred=0.7512921690940857 | ans=0
image 4 | pred=0.0017309949034824967 | ans=0
image 5 | pred=0.001742606982588768 | ans=0
image 6 | pred=0.000600828614551574 | ans=0
image 7 | pred=0.9383810758590698 | ans=1
image 8 | pred=0.0016865450888872147 | ans=0
image 9 | pred=0.0007257546531036496 | ans=0
image 10 | pred=0.0006229666178114712 | ans=0
image 11 | pred=0.9989845156669617 | ans=1
image 12 | pred=0.001240861602127552 | ans=0
image 13 | pred=0.0006093949778005481 | ans=0
image 14 | pred=0.9991174340248108 | ans=1
image 15 | pred=0.00038419407792389393 | ans=0
image 16 | pred=0.0970044657588005 | ans=1
image 17 | pred=0.9984909296035767 | ans=1
image 18 | pred=0.00386615376919508 | ans=0
image 19 | pred=0.0015628642868250608 | ans=0
image 20 | pred=0.9997197985649109 | ans=1
image 21 | pred=0.9999942779541016 | ans=1
image 22 | pred=0.002477513626217842 | ans=0
image 23 | pred=0.0005884646088816226 | ans=0
```

# Malware Binary Classification

## Model Performance

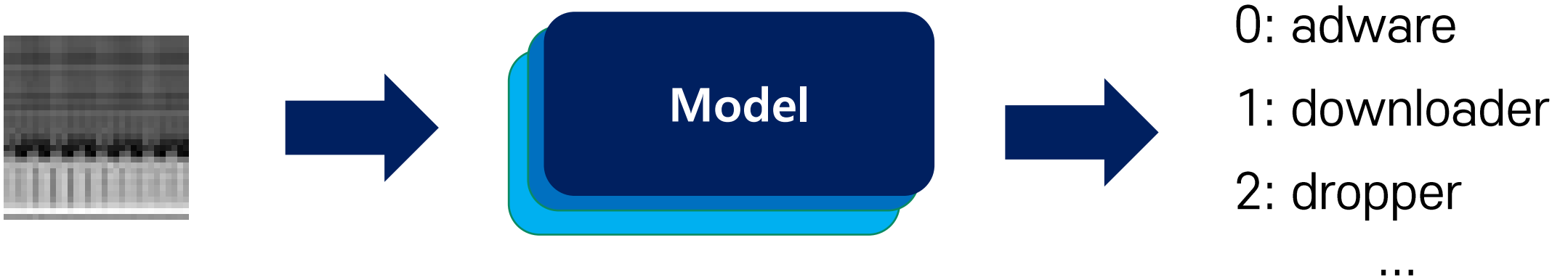


# Malware Family Classification

악성코드 계통 분류 모델 설계 및 학습



# Malware Family Classification



A model that receives a transformed 32 x 32 size image as input and determines its malware family such as adware, downloader, etc.

# Malware Family Classification

dataset	# of malware data points	# of non-malware data points
train set	7,000	3,000
test set1	5,000	5,000
test set2	5,000	5,000



dataset	# of malware data points
train set	7,000
test set1	5,000
test set2	5,000

# Malware Family Classification

dataset	# of malware data points
train set	7,000
test set1	5,000
test set2	5,000



dataset	# of malware data points
ALL	17,000

# Malware Family Classification

dataset	# of malware data points
ALL	17,000

67 / 73

67 security vendors and no sandboxes flagged this file as malicious

27dea2a7fabe658ffa36752b0dac73ad52009302d321530631204cb3b062d7  
4c84fca7dec51c00febfe492d90842a2.vir

101.50 KB Size | 2020-03-17 12:38:03 UTC 2 years ago

checks: network-adapters | direct-cpu-clock-access | overlay | peers | persistence | runtime-modules

Community Score

DETECTION DETAILS RELATIONS BEHAVIOR COMMUNITY

Security Vendors' Analysis

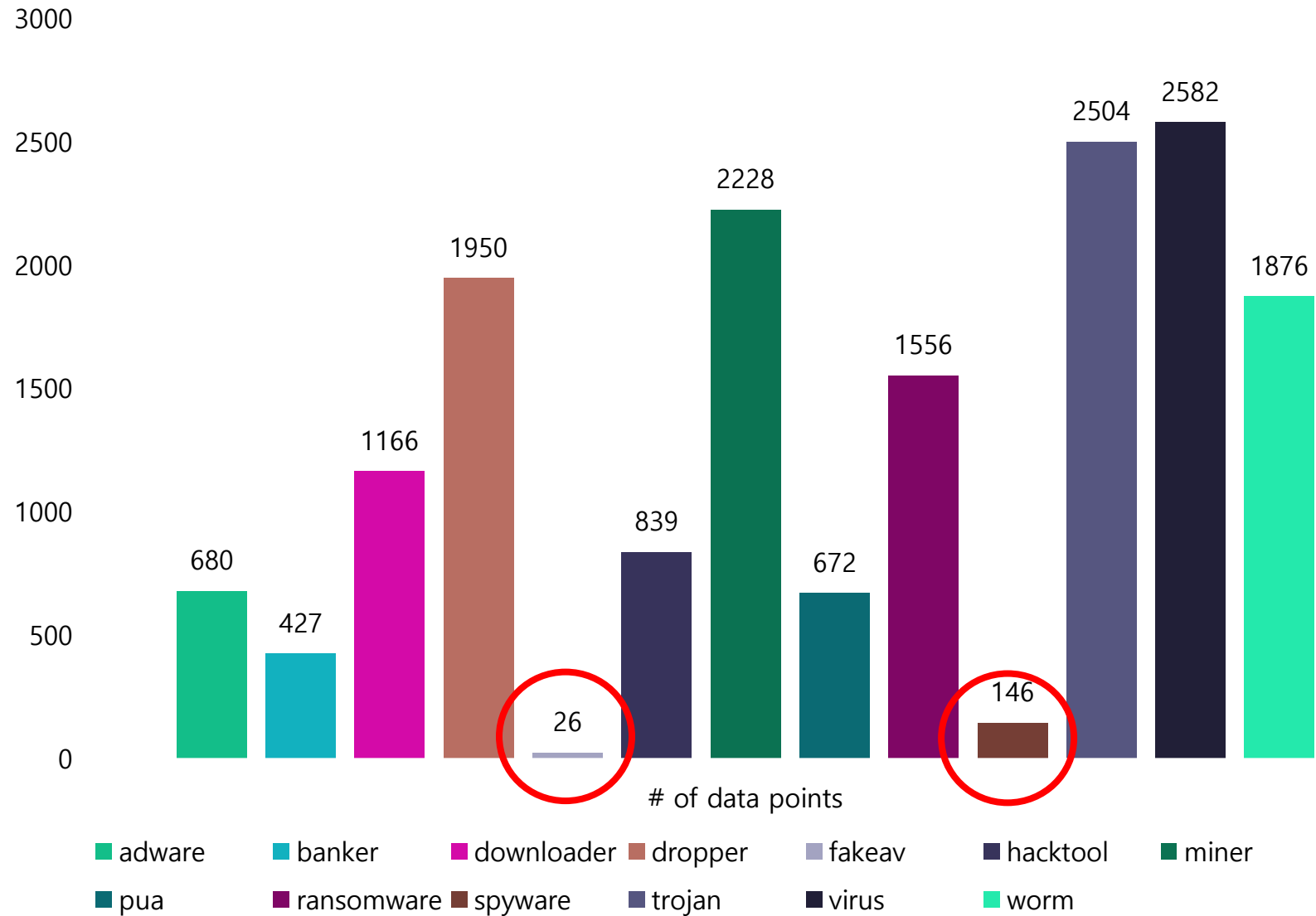
Acronis (Static ML)	ⓘ Suspicious	Ad-Aware	ⓘ Gen:Variant.Ulise.43730
AegisLab	ⓘ Trojan.Win32.Agent.tnq	AhnLab-V3	ⓘ Trojan.Win32.Scar.R260039
Alibaba	ⓘ Trojan:Win32/Sakurel.asf0d11	ALYac	ⓘ Gen:Variant.Ulise.43730
Antiy-AVL	ⓘ Trojan(Dropper)/Win32.Agent.bjrkva	Arcabit	ⓘ Trojan.Ulise.DAAD2
Avast	ⓘ Win32:Shyape-F [Trj]	AVG	ⓘ Win32:Shyape-F [Trj]
Avira (no cloud)	ⓘ TR/Crypt.XPACK.Gen3	Baidu	ⓘ Win32.Trojan.Shyape.a
BitDefender	ⓘ Gen:Variant.Ulise.43730	BitDefenderTheta	ⓘ Ai:Packers.00084F11F
Bkav Pro	ⓘ W32.AI.Detect.VM.malware	ClamAV	ⓘ Win.Malware.Scar-6745903-0
CMC	ⓘ Trojan.Win32.ScarIO	Comodo	ⓘ TrojWare.Win32.Trojan.XPACK.Gen@2h...
CrowdStrike Falcon	ⓘ Win/malicious_confidence_100% (W)	Cybereason	ⓘ Malicious.7dec51
Cylance	ⓘ Unsafe	Cyren	ⓘ W32/S-25063cb0/Eldorado
DrWeb	ⓘ Trojan.DownLoad3.19308	eGambit	ⓘ RAT.Sakula
Emsisoft	ⓘ Gen:Variant.Ulise.43730 (B)	Endgame	ⓘ Malicious (high Confidence)
eScan	ⓘ Gen:Variant.Ulise.43730	ESET-NOD32	ⓘ A Variant Of Win32/Shyape.G
F-Prot	ⓘ W32/S-25063cb0/Eldorado	F-Secure	ⓘ Trojan.TR/Crypt.XPACK.Gen3
Fortinet	ⓘ W32/Shyape.Zltr	GData	ⓘ Win32.Trojan.Sakurel.B

Extracts labels from examination data based on **VirusTotal**.

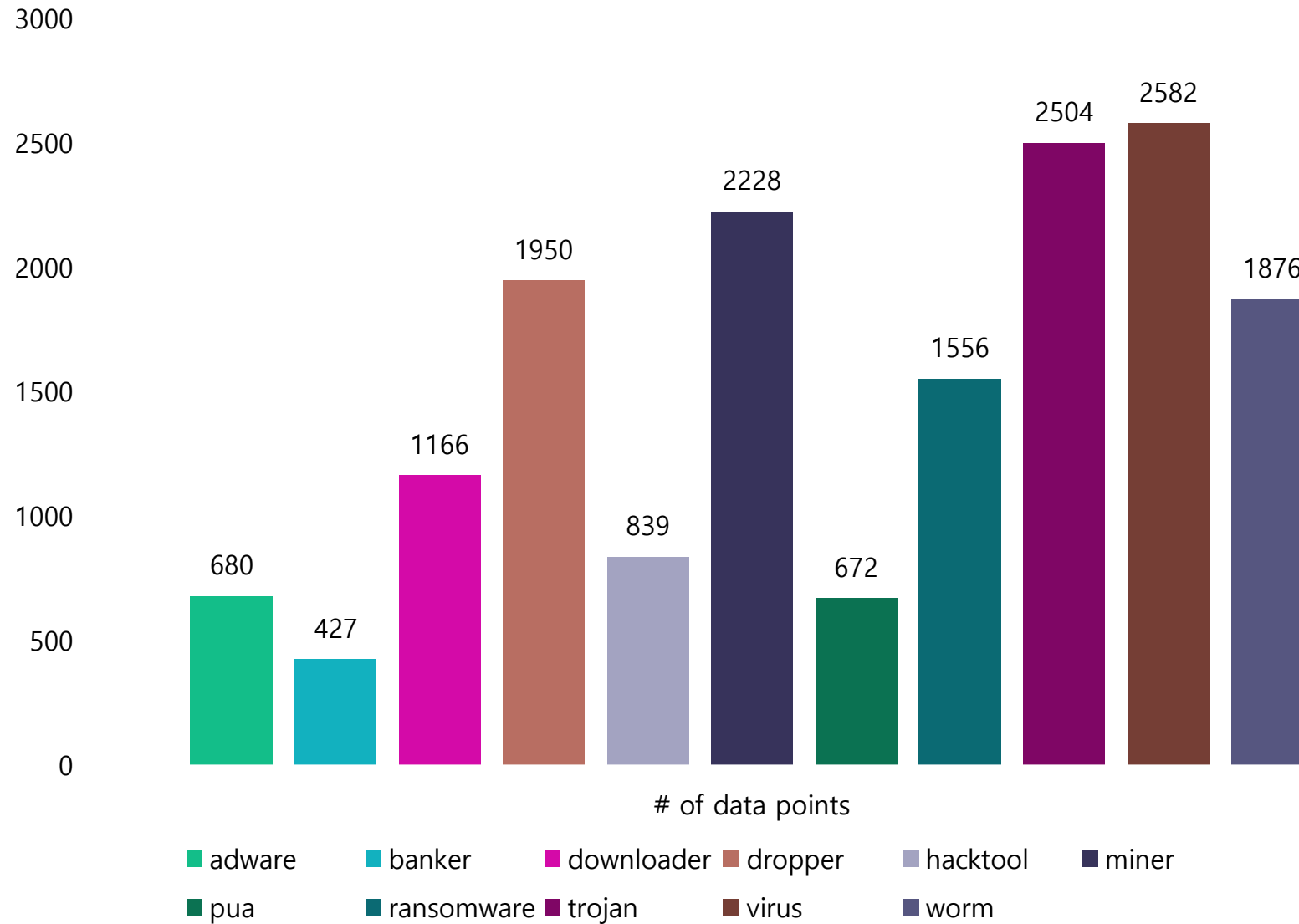
1. majority rule
2. exceptions for trojan
  - 1) only case -> trojan
  - 2) else -> get the 2<sup>nd</sup> decision



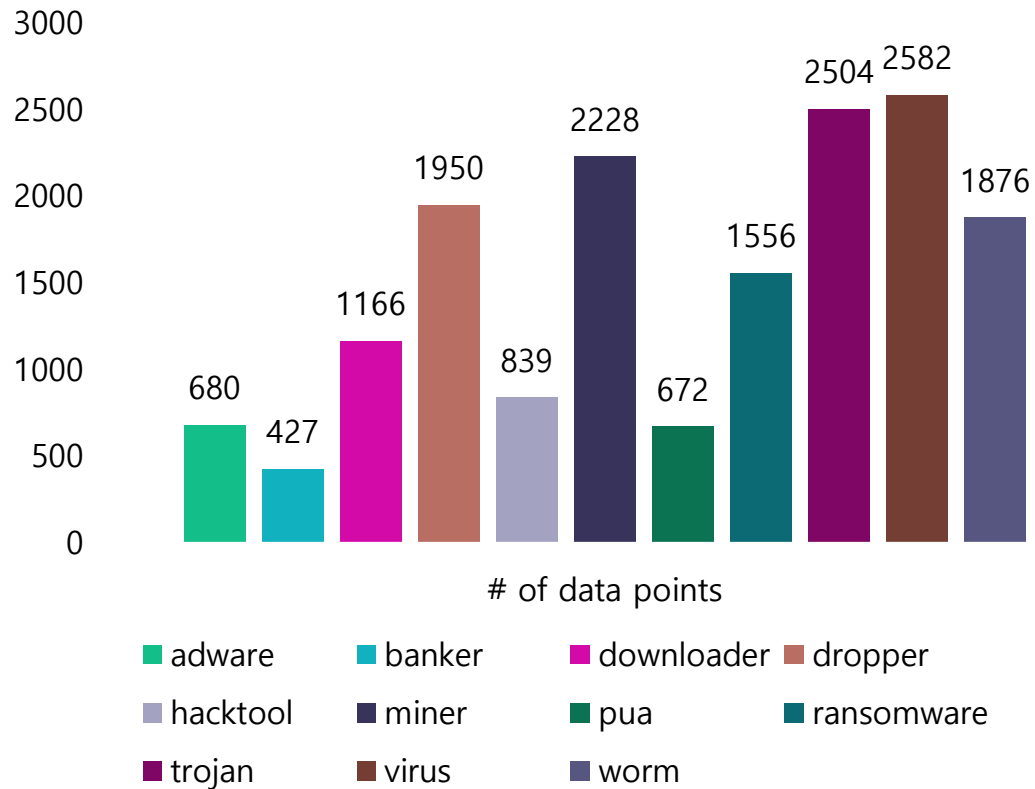
# Malware Family Classification



# Malware Family Classification



# Malware Family Classification



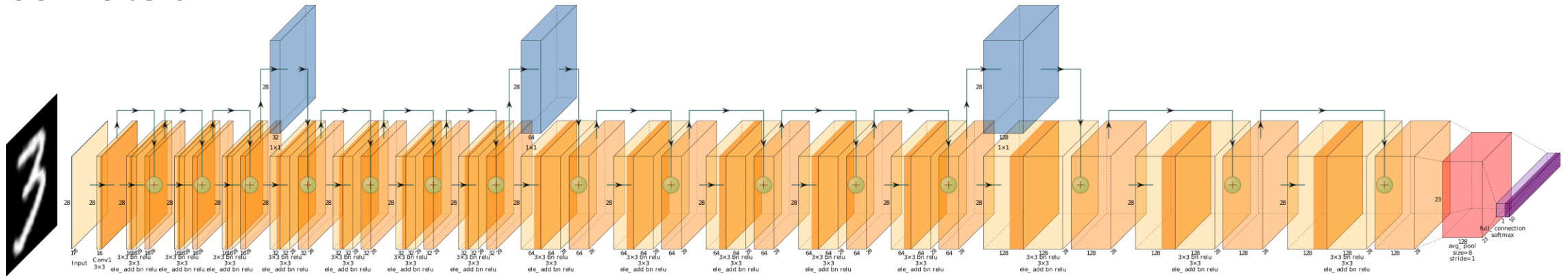
We performed over-sampling  
with `imblearn.over_sampling.RandomOverSampler`

Object to over-sample the minority class(es) by picking  
samples at random with replacement.

The bootstrap can be generated in a smoothed manner.

# Malware Family Classification

## ResNet50



Which malware family does this  
virus belong to?  
multi-class classification

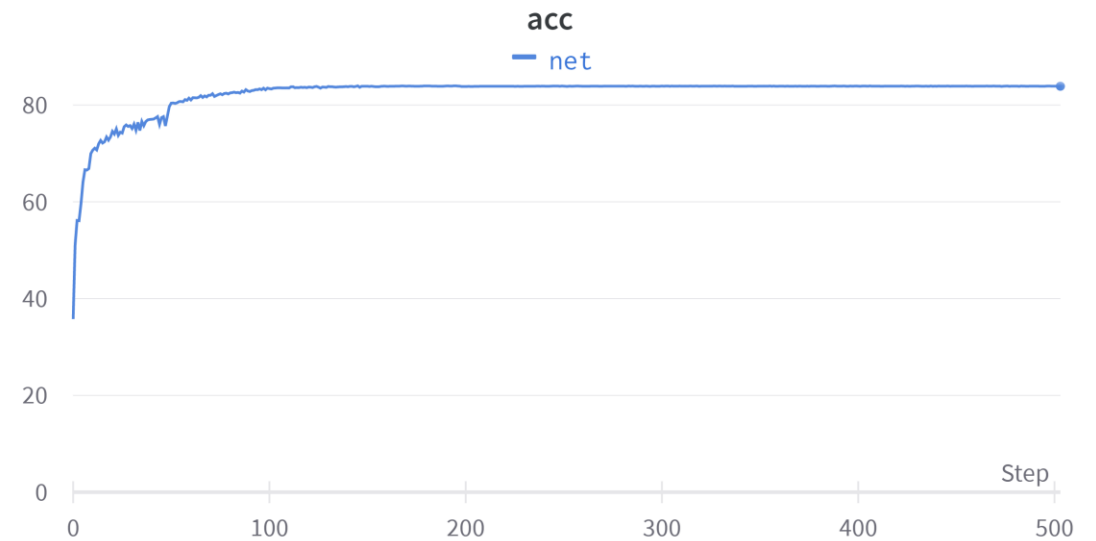
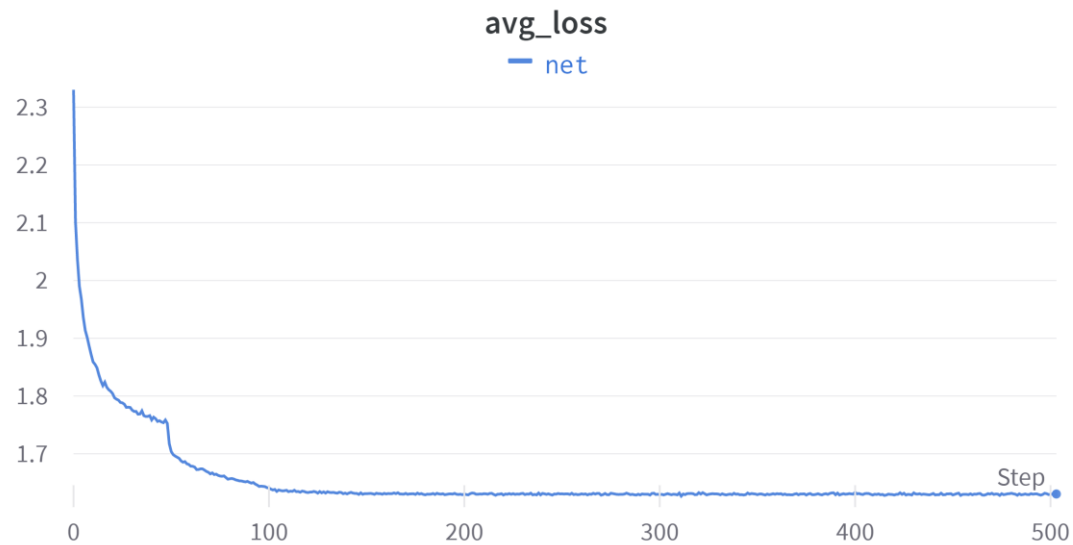


```
self.resnet_50.fc = nn.Sequential(  
    nn.Linear(2048, 1000),  
    nn.ReLU(inplace=True),  
    nn.Linear(1000, 256),  
    nn.ReLU(inplace=True),  
    nn.Linear(256, 64),  
    nn.ReLU(inplace=True),  
    nn.Linear(64, 11),  
    nn.Softmax()  
)
```



# Malware Family Classification

## Performance



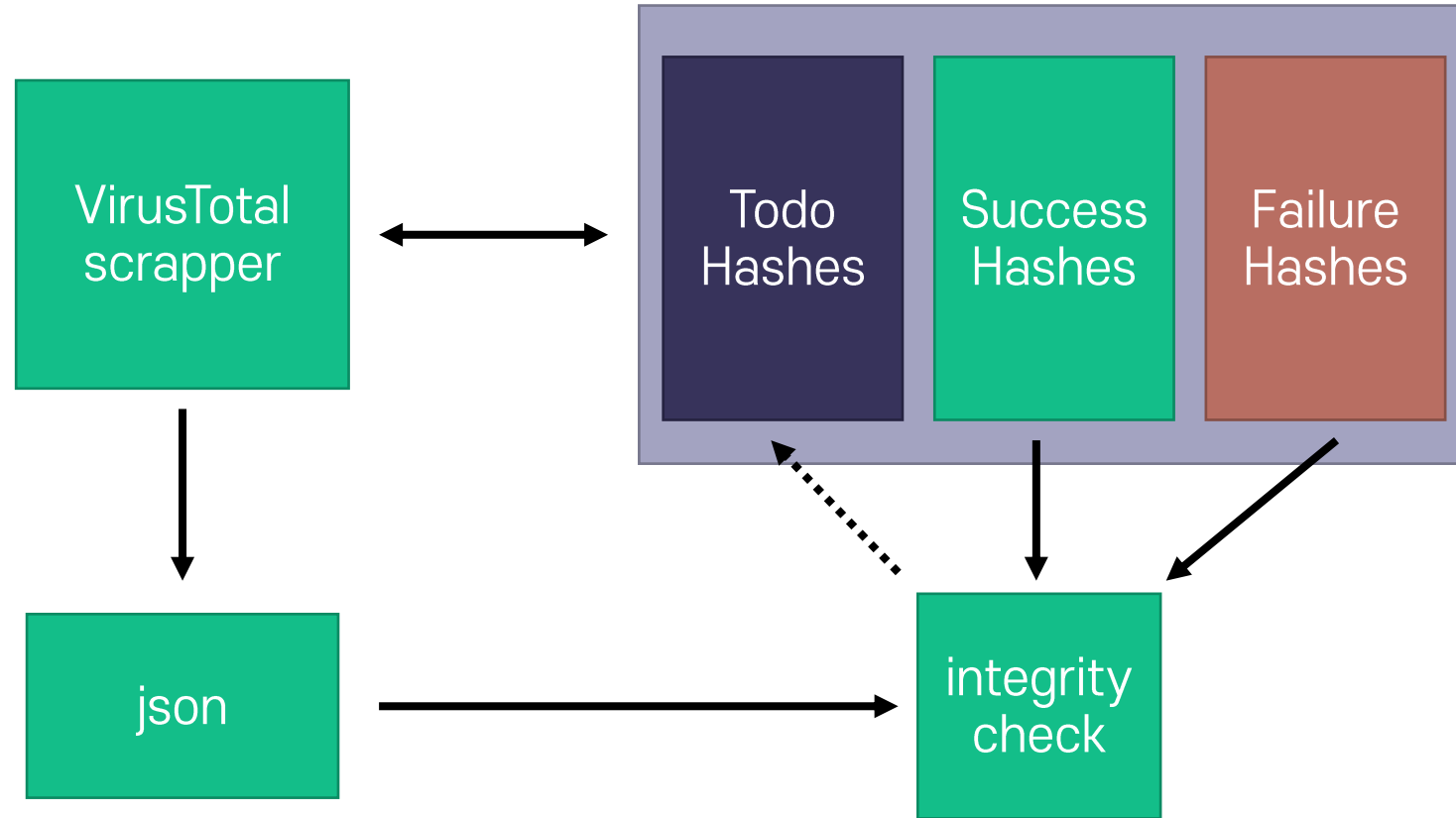
test set accuracy: 84%

# Other works

Analysis of malware feature importance



# Malware Binary Classification



# Malware Binary Classification

520a92372d737eb4523e58b1fa7711b0bc9e99e0ffec4cf6376b96ebf0d455dc

0 / 68

✓ No security vendors and no sandboxes flagged this file as malicious

520a92372d737eb4523e58b1fa7711b0bc9e99e0ffec4cf6376b96ebf0d455dc  
GeneralTel.dll  
64bits assembly invalid-rich-pe-linker-version overlay pedll

638.72 KB  
Size

2021-05-16 07:13:21 UTC  
1 year ago

DLL

Community Score

DETECTION DETAILS COMMUNITY

Security Vendors' Analysis

Acronis (Static ML)	✓ Undetected	Ad-Aware	✓ Undetected
AegisLab	✓ Undetected	AhnLab-V3	✓ Undetected
Alibaba	✓ Undetected	ALYac	✓ Undetected
Antiy-AVL	✓ Undetected	Arcabit	✓ Undetected
Avast	✓ Undetected	Avira (no cloud)	✓ Undetected

- Folder: 40,000
- Files: 414,784
- Capacity: 2.84GB

```
{ } base.json  
{ } behaviours.json  
{ } bundled_files.json  
{ } contacted_domains.json  
{ } contacted_ips.json  
{ } contacted_urls.json  
{ } dropped_files.json  
{ } execution_parents.json  
{ } pe_resource_children.json  
{ } pe_resource_parents.json
```

# Malware Binary Classification

```
{
  "data": {
    "attributes": {
      "type_description": "Win32 EXE",
      "vhash": "064046551d1500f1z14z211b5z3dz17z45z",
      "trid": [
        {
          "file_type": "Win32 Executable MS Visual C++ (generic)",
          "probability": 33.7
        },
        {
          "file_type": "Win64 Executable (generic)",
          "probability": 29.8
        },
        {
          "file_type": "Microsoft Visual C++ compiled executable (generic)",
          "probability": 17.8
        },
        {
          "file_type": "Win32 Dynamic Link Library (generic)",
          "probability": 7.1
        },
        {
          "file_type": "Win32 Executable (generic)",
          "probability": 4.8
        }
      ]
    },
    "creation_date": 1289577258
  }
}
```

## Type description

- Win32 DLL
- Win32 EXE
- JavaScript
- ...

## TrID

- Win32 Executable MS Visual C++ (generic)
- OS/2 Executable (generic)
- Microsoft Visual C++ compiled executable (generic)
- ...



# Malware Binary Classification

```
"popular_threat_classification": {
  "suggested_threat_label": "trojan.stormattack/ddos",
  "popular_threat_category": [
    {
      "count": 28,
      "value": "trojan"
    },
    {
      "count": 7,
      "value": "dropper"
    }
  ],
  "popular_threat_name": [
    {
      "count": 12,
      "value": "stormattack"
    },
    {
      "count": 8,
      "value": "ddos"
    },
    {
      "count": 7,
      "value": "rincux"
    }
  ]
},
```

## Popular threat category/name

- Trojan
- Worm
- Virus
- Dropper
- DDOS
- ...

# Malware Binary Classification

```
ssdeep : 192:VBOZam0501ct5H5C19Ck  
"packers": {  
  "PEiD": "Microsoft Visual C++"  
},  
"..."
```

## Packers

- PEiD
  - UPX
  - Microsoft Visual C++ v6.0 DLL
  - .NET executable
  - ...

# Malware Binary Classification

## Sections

- Overlay
  - Entropy
  - Chi2
  - ...
- .text, .rdata, .rsrc ...
  - Entropy

# Malware Binary Classification

type_desc	best_trid	best_trid	type_exte	has_signa	packer	type_tag	overlay_e	overlay_c	overlay_fi	dropped	beha
Win32 DL Win32 Exe	33.7	dll	0	unknown	pedll	2	252	ASCII text	0		
Win32 DL Win64 Exe	61.7	dll	0	unknown	pedll	2	252	ASCII text	0		
Win32 EX OS/2 Exec	25.2	exe	1	unknown	peexe	0.004528	2.58E+09	Data	0		
Win32 DL Windows	83.7	dll	1	unknown	pedll	7.398872	22471.63	Data	0		
Win32 EX Win32 Exe	42.7	exe	0	unknown	peexe	2	252	ASCII text	0		
Win32 DL Win32 Exe	33.7	dll	1	Microsoft	pedll	2	252	ASCII text	0		
Win32 DL Win32 EX	60.7	dll	0	UPX	pedll	2	252	ASCII text	0		
Win32 EX Microsoft	33.5	exe	0	unknown	peexe	2	252	ASCII text	0		
Win32 EX Win32 Exe	41	exe	0	Microsoft	peexe	4.818212	96692040	Data	0		
Win32 EX Win32 Exe	58.7	exe	1	Microsoft	peexe	7.406668	7600.766	Data	0		
Win32 DL Win64 Exe	82	dll	0	unknown	pedll	1.212236	52465.13	ASCII text	0		
Win32 EX Win64 Exe	42.3	exe	1	UPX v0.89	peexe	3.649286	5206.675	Data	0		
Win32 EX Win64 Exe	82	exe	1	unknown	peexe	7.999897	268.3149	Data	0		
Win32 DL Generic N	65.0	dll	1	Microsoft	pedll	2	252	ASCII text	0		

hash

type\_description

best\_trid\_type

best\_trid\_probability

type\_extension

has\_signature

packer type\_tag

overlay\_entropy

overlay\_chi2

overlay\_filetype

dropped\_files\_count

behaviours\_count

contacted\_domains\_count

contacted\_ips\_count

contacted\_urls\_count

.text\_entropy .bss\_entropy .rdata\_entropy .data\_entropy

.xdata\_entropy .idata\_entropy .pdata\_entropy

.rsrc\_entropy .reloc\_entropy .CRT\_entropy

# Malware Binary Classification

Test set 1

4577	487
113	5864

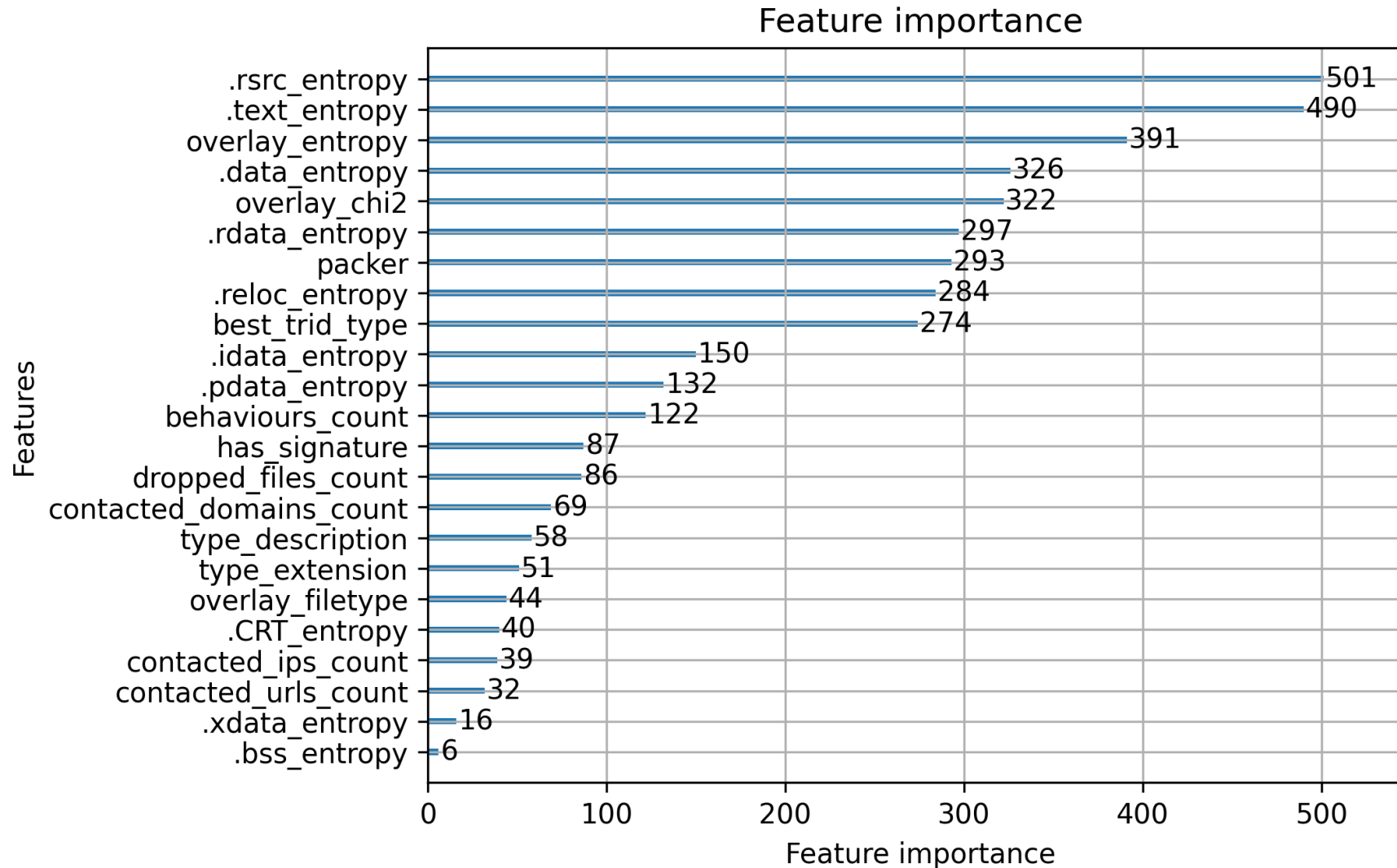
Accuracy: 96.15%  
Precision: 93.99%

Test set 2

4325	448
89	4684

Accuracy: 94.37%  
Precision: 91.27%

# Malware Binary Classification





# Thank You 😊

Hyeonsu Kim, Gaon Choi.

Department of Computer Science

Hanyang University

