Mind Your Weight(s): A Large-scale Study on Insufficient Machine Learning Model Protection in Mobile Apps

Zhichuang Sun
Northeastern University

Ruimin Sun
Northeastern University

Long Lu
Northeastern University

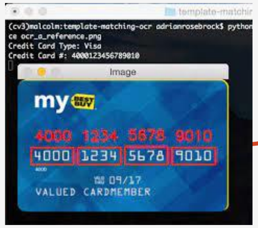
Alan Mislove
Northeastern University

Overview

Face recognition

On-Device ML

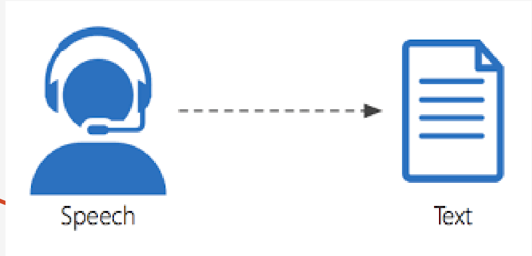
Liveness Detection



ID/Bank card recognition

Remote Models

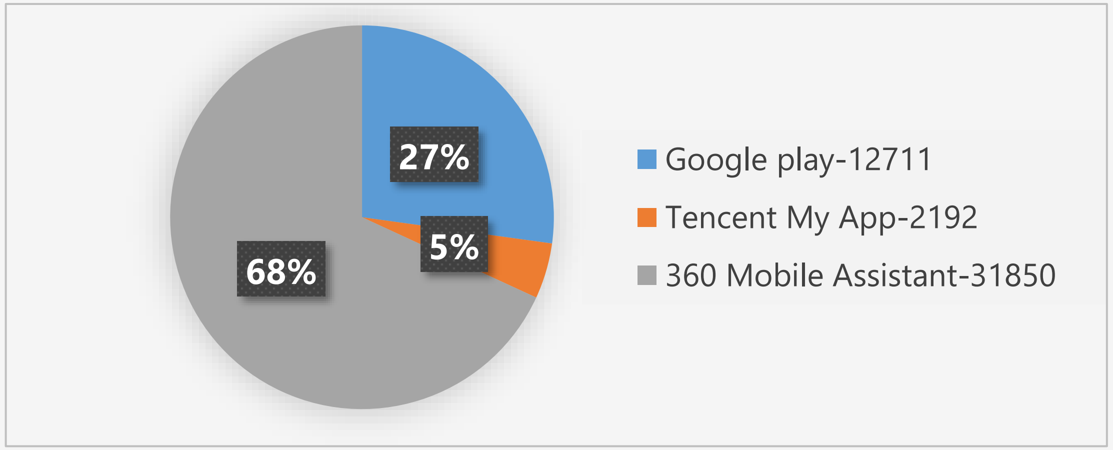
NLP



Speech Recognition Task

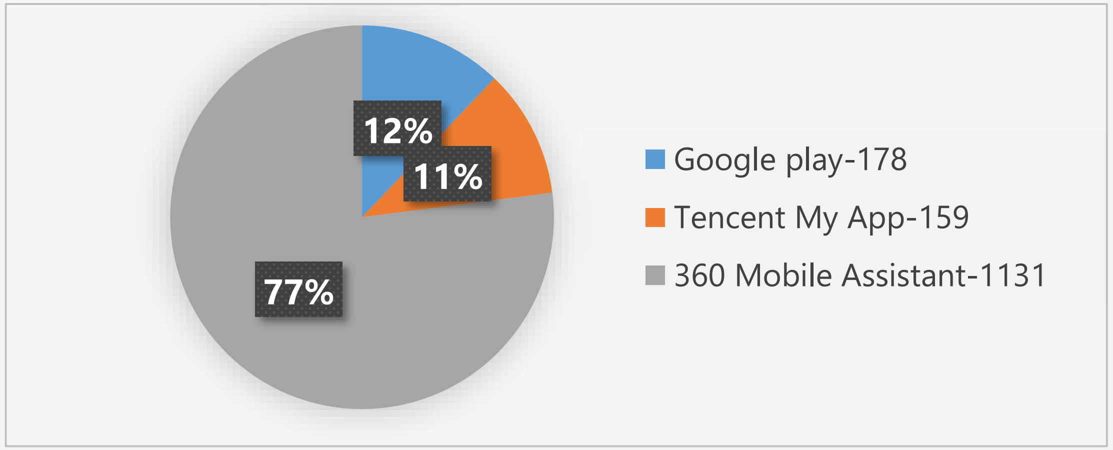
Overview

1. Q1: How widely is model protection used in apps?
2. Q2: How robust are existing model protection techniques?
3. Q3: What impacts can (stolen) models incur?



Overview

1. **Q1: How widely is model protection used in apps?**
   41% of ML apps do not protect their models at all
2. **Q2: How robust are existing model protection techniques?**
   Extract 66% models for apps use model protection or encryption
3. **Q3: What impacts can (stolen) models incur?**

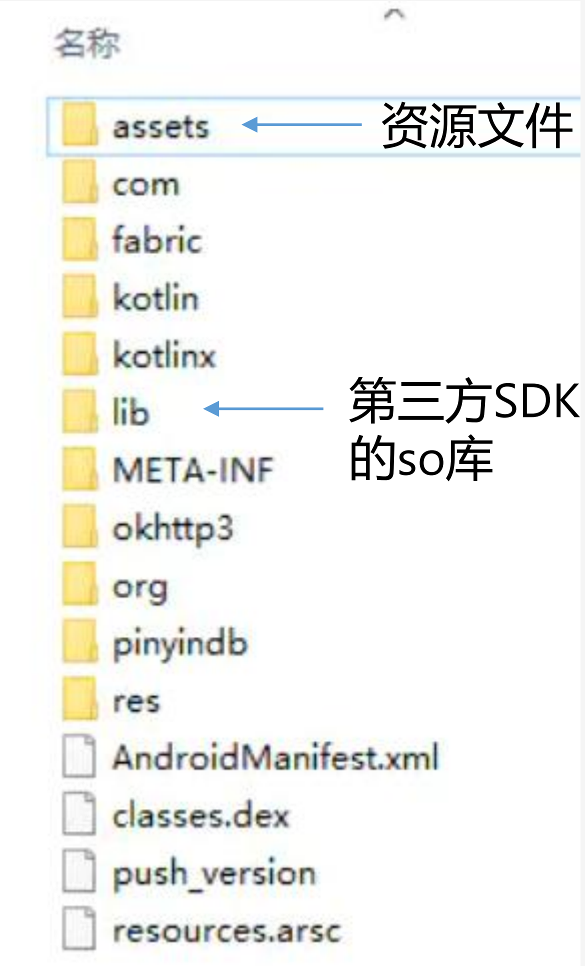


应用宝
myapp.com

GET IT ON
Google Play

360
手机助手

Overview



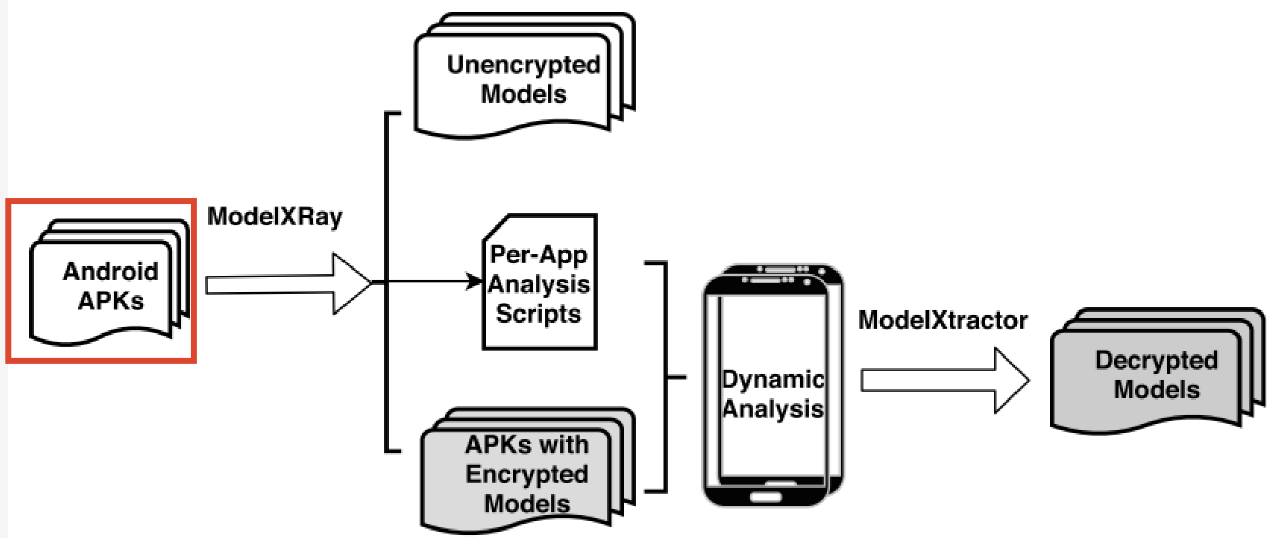


Figure 1: Overview of Static-Dynamic App Analysis Pipeline

Q1: How widely is model protection used in apps?

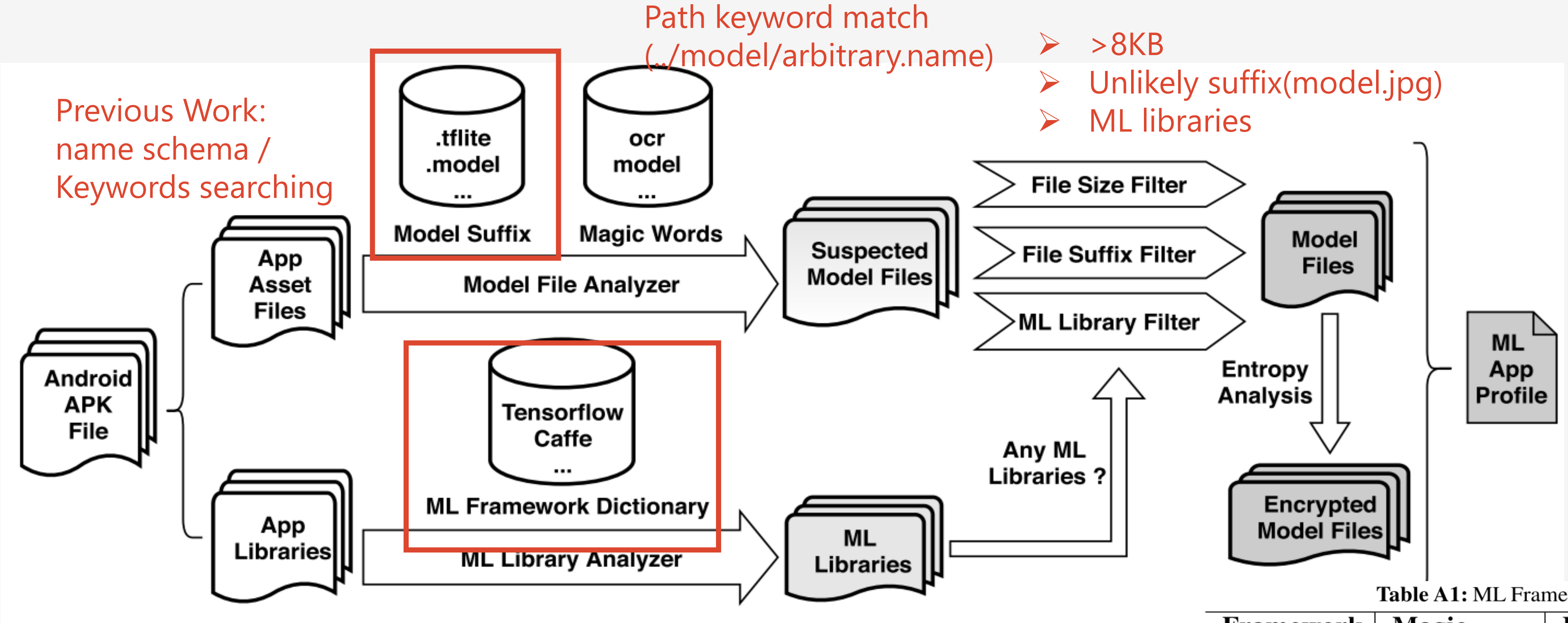


Figure 2: Identify Encrypted Models with ModelXRay

Table A1: ML Framework Keywords

Framework	Magic Words	Framework	Magic Words
TensorFlow	tensorflow	Caffe	caffe
MXnet	mxnet	NCNN	ncnn
Mace	libmace, mace_input	SenseTime	sensetime, st_mobile
ULS	ulstracker, ulsface	Other	neuralnetwork, lstm, cnn, rnn

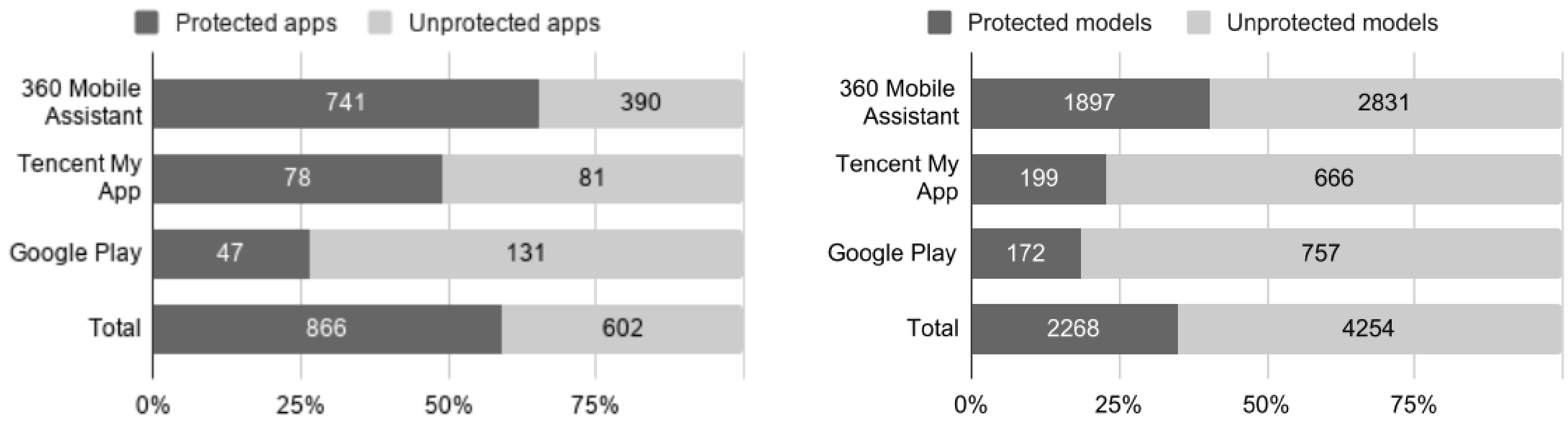
Identifying ML apps:

- False negative rate 6.8%
- False positive rate 0%

Model Downloaded at Runtime:
Contain on-device library but not any models(109)

Mengwei Xu, Jiawei Liu, Yuanqiang Liu, Felix Xiaozhu Lin, Yunxin Liu, and Xuanzhe Liu. A First Look at Deep Learning Apps on Smartphones. The WorldWide Web Conference on - WWW'19, (May):2125–2136, 2019.

Q1: How widely is model protection used in apps?



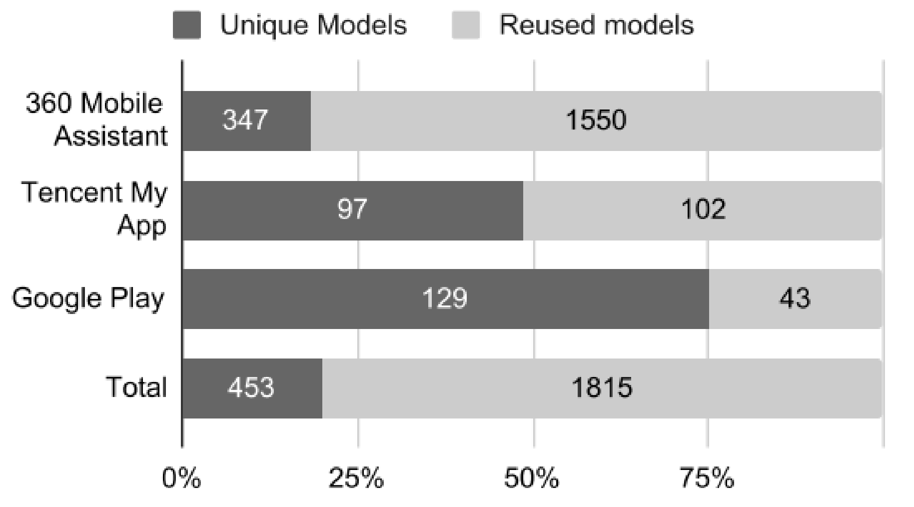
(a) Apps using protected/encrypted models vs. those using unprotected models

(b) On-device models that are protected/encrypted vs. those not

- 26% models in Chinese apps are protected
- 23% in Google Play apps

Q1: How widely is model protection used in apps?

- MD5 HASH(model)
- Many encrypted model reused/shared among apps
  - SenseID_Motion_Liveness.model is found in 81 apps
  - 60 cases of different app companies are reusing model lisences
- Only 22% of all protected models are unique.



(c) Unique encrypted models
vs. encrypted models reused/shared by multiple apps.

Remote vs On-device models

Table 5: Comparison between apps using remote and on-device ML models

App Number	360 Mobile Assistant	Tencent My App	Google Play	Sum
Remote Models	1,186	118	37	1,341
On-device Models	1,131	159	178	1,468
Hybrid Mode	153	23	6	182

Functionality	Total
OCR(Optical Character Recognition)	441
Face Tracking	620
Speech Recognition	88
Hand Detection	10
Handwriting Recognition	42
Liveness Detection	872
Face Recognition	294
Iris Recognition	9
ID Card Recognition	483
Bank Card Recognition	299

- Measure the use of remote models through APIs provided by AI companies
  - Google Cloud AI, Amazon Cloud AI, Baidu AI
  - Scanning docs for unique naming
- On-device models have security critical use cases and real-time demands
- Remote:
  - 1075 NLP
  - 266 ML Vision

Q2: How robust are existing model protection techniques?

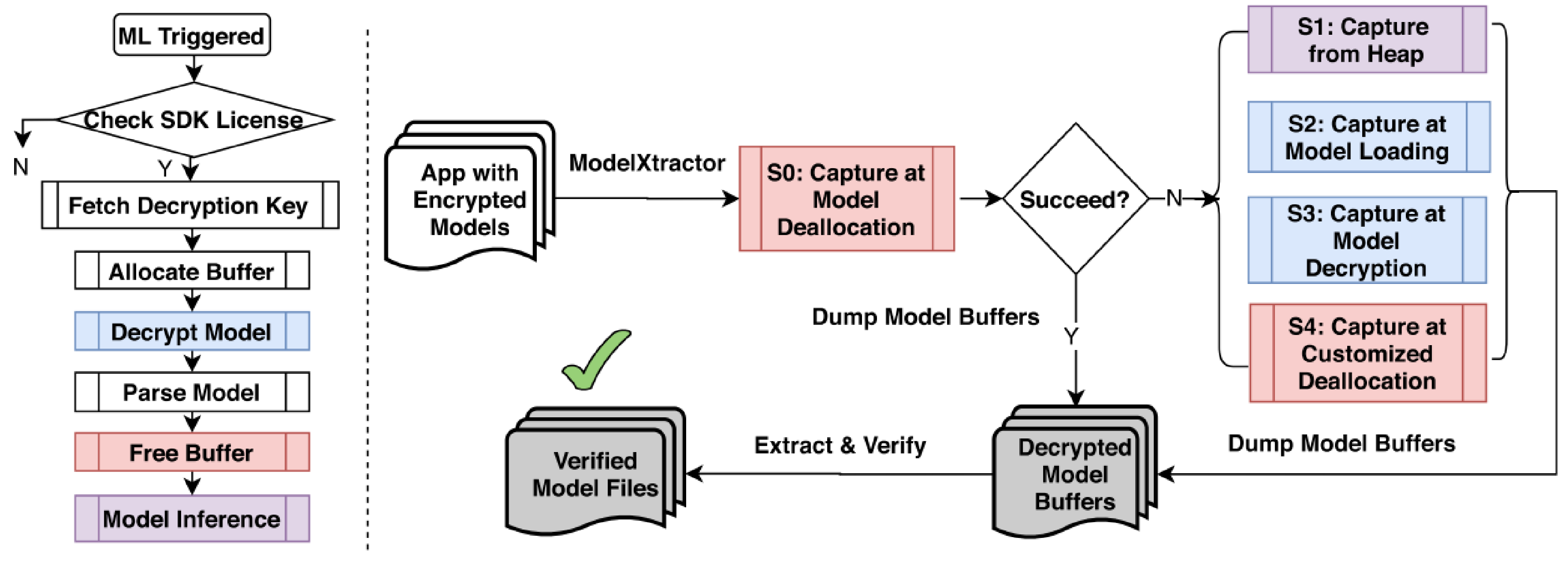


Figure 7: Extraction of (decrypted) models from app memory using ModelXtractor

- **S0:** Default
- **S1:** Do not free buffers timely
- **S2:** function like *loadModel*
- **S3:** function like *aes256_decrypt*
- **S4:** customized Deallocation buffer

- Targets on ML models that are encrypted during transportation and at rest but **not protected when in use or loaded in memory**

Q2: How robust are existing model protection techniques?

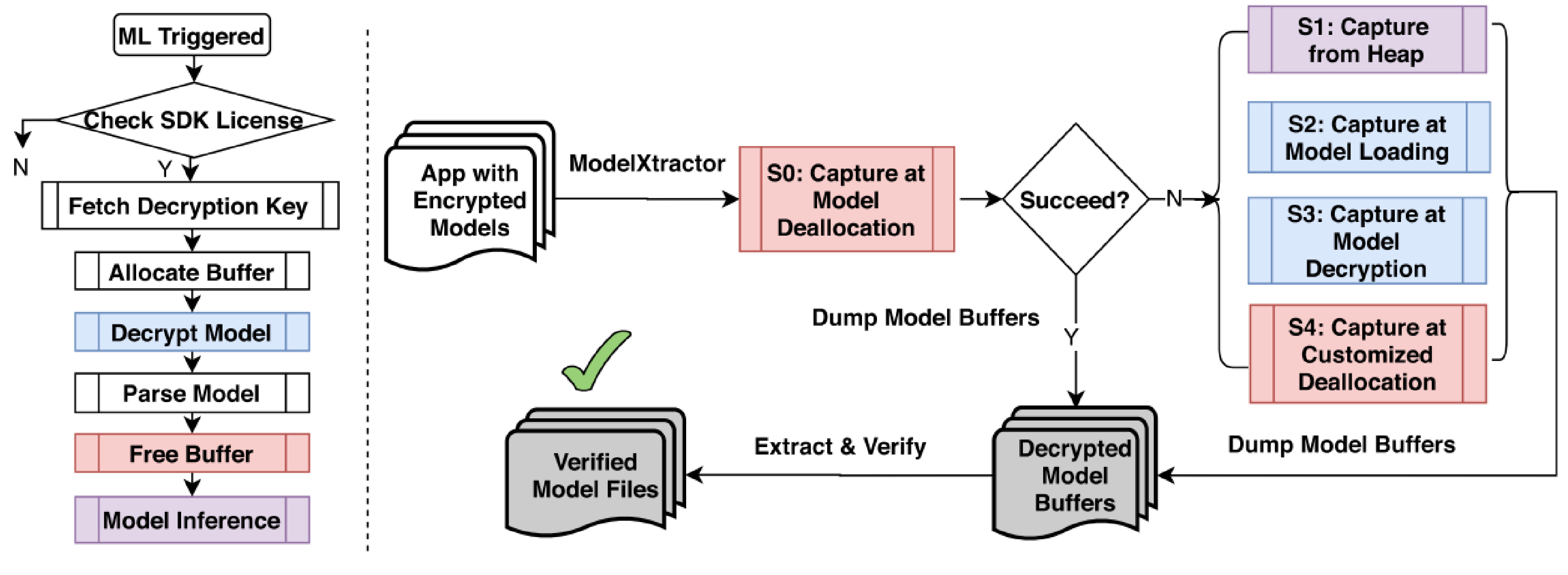
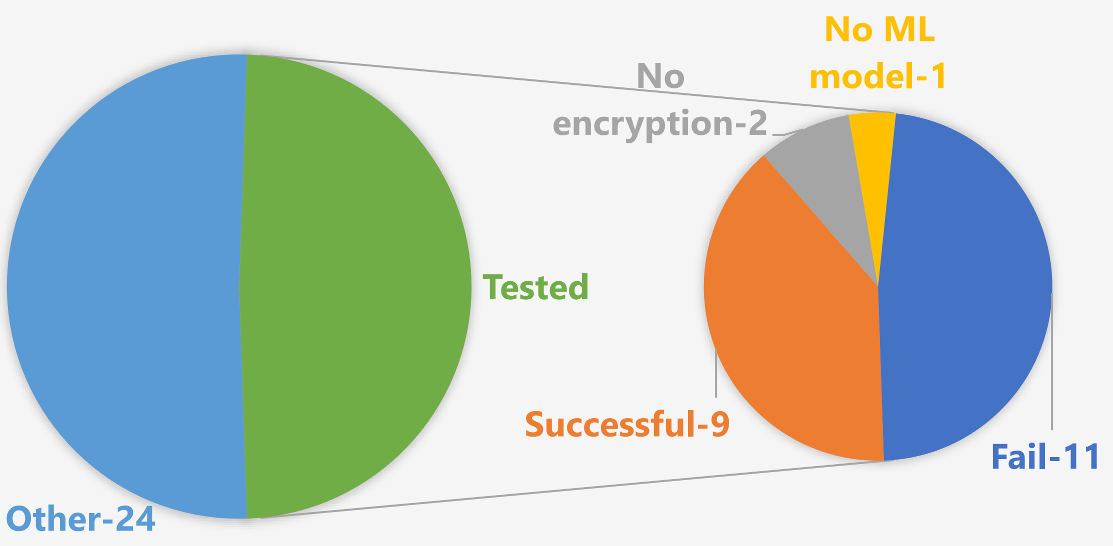


Figure 7: Extraction of (decrypted) models from app memory using ModelXtractor

- Encode in Protobuf format:
  - "relu", "conv1" to identify buffers contain models
  - Start with message "0A"
- TFLite includes "TFL2" or "TFL3"

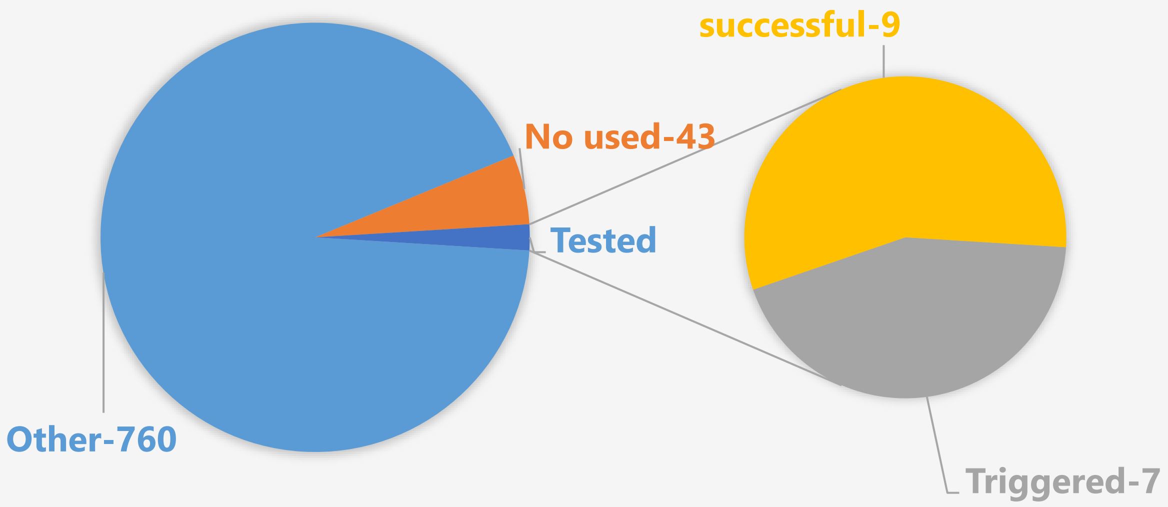
Evaluation

1 Google play:



2 Chinese App:

1. Require phone number
2. Require a local bank account



Findings & insights

1. Even for those trying to protect their models, it’s hard to do it in a robust way using the file encryption-based techniques.
2. Some extracted models are valuable or security-critical
3. Extracted model can be directly used by attacker.

Table 7: Overview of Successfully Dumped Models with ModelXtractor

App name	Downloads	Framework	Model Functionality	Size (B)	Format	Reuses	Extraction Strategy
Anonymous App 1	300M	TFLite	Liveness Detection	160K	FlatBuffer	18	Freed Buffer
Anonymous App 2	10M	Caffe	Face Tracking	1.5M	Protobuf	4	Model Loading
Anonymous App 3	27M	SenseTime	Face Tracking	2.3M	Protobuf	77	Freed Buffer
Anonymous App 4	100K	SenseTime	Face Filter	3.6M	Protobuf	3	Freed Buffer
Anonymous App 5	100M	SenseTime	Face Filter	1.4M	Protobuf	2	Freed Buffer
Anonymous App 6	10K	TensorFlow	OCR	892K	Protobuf	2	Memory Dumping
Anonymous App 7	10M	TensorFlow	Photo Process	6.5M	Protobuf	1	Freed Buffer
Anonymous App 8	10K	SenseTime	Face Track	1.2M	Protobuf	5	Freed Buffer
Anonymous App 9	5.8M	Caffe	Face Detect	60K	Protobuf	77	Freed Buffer
Anonymous App 10	10M	Face++	Liveness	468K	Unknown	17	Freed Buffer
Anonymous App 11	100M	SenseTime	Face Detect	1.7M	Protobuf	18	Freed Buffer
Anonymous App 12	492K	Baidu	Face Tracking	2.7M	Unknown	26	Freed Buffer
Anonymous App 13	250K	SenseTime	ID card	1.3M	Unknown	13	Freed Buffer
Anonymous App 14	100M	TFLite	Camera Filter	228K	Json	1	Freed Buffer
Anonymous App 15	5K	TensorFlow	Malware Classification	20M	Protobuf	1	Decryption Buffer

Interesting cases

1 Encrypting Both Code and Model Files

- App uses Anyline OCR SDK
- Tensorflow Framework
- Places encrypted model under *“encrypted_model”*
- Runs ML inference in a customized WebView, where an encrypted Javascript, dynamically load at runtime
- Using S1, found TF model buffers in the memory dump

Interesting cases

2 Encrypting Feature Vectors and Formats

- Tensorflow framework
- It does not encrypt its model file
- Encrypt the feature vector which is the input of the model
- Developers assumes it’s impossible to reuse the model without input format
- Extracted the decrypted vectors by instrumenting the decryption function

Interesting cases

3 Encrypting Models Multiple Times

- P2P loans apps with two models: ID card recognition and liveness detection
- ModelXtractor extracted 6 model buffers but only 2 encrypted model files found
- SenseID_Ocr_Idcard_Mobile_1.0.1.model has size of 1.3MB, has one buffer with the same size
- It is a tar file containing align_back.model, also an encrypted file
- The app encrypts each model individually and compress all into a tar file and then encrypts it again