Machine Learning Models that Remember Too Much

Congzheng Song, Thomas Ristenpart, Vitaly Shmatikov

Cornell Tech

Summarize the paper

- Problem : With the growing popularity of Machine Learning (ML), some data-holders search to apply this technology to their dataset that may contain sensitive data. They may trust a malicious ML Provider and risk a leak of their information.
- Goal : Show that it is possible to extract sensitive information from a trained malicious Machine Learning Model.
- Contribution :
  - 4 methods to extract sensitive data.
  - Methods to prevent those type of attack.
- Meaning :

  We cannot apply blindly Machine Learning to sensitive data.

Contribution of the Paper

- Demonstrate that minor modifications to ML models can allow the extraction of data from their training datasets without affecting the quality of the model by standard ML metrics.
- Gives 4 malicious models which hide training data in their parameters
- Claims that use 3rd-party ML models on sensitive data is risky.

Situations of Machine Learning Use

The popularity of Machine Learning has led to an explosion of ML libraries, framework and services and to the appearance of ML provider and ML marketplace.

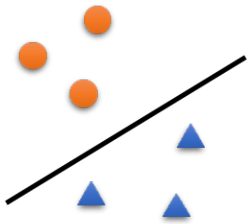
	Algorithm Provider	Computation Power Provider	Examples
Libraries	Developers	Data Holders	scikit learn, TensorFlow
Cloud Service	Service Operators	Service Operators	Google Prediction, Amazon Machine Learning
Platforms/ Marketplace	Algorithm Developers	Platform Operators / Data Holders	ALGORITHMIA

Motivation

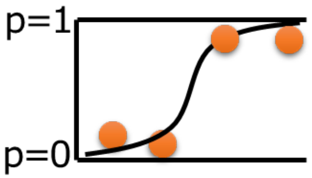
- Non-experts just use models ‘as-is’ from providers.
  - They usually do not (or cannot) check whether the models are malicious!
- ML models have huge memorization capability.
  - What if models do secondary malicious jobs, silently…?
  - What if models remember too much data that should not be leaked…?

Machine Learning Models

Linear Models



Support Vector Machine (SVM)



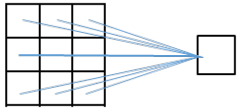
Logistic Re gression (LR)

Simple and Efficient
Suitable for massive number of features
∵ Number of parameters $= O(features)$

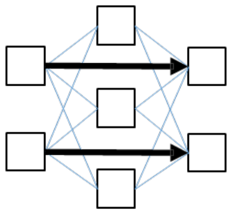
Deep Learning Models

Artificial Neural Networks (ANN)

(belongs to)



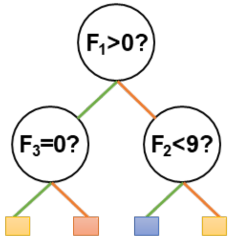
Convolutional Neural Network (CNN)



Residual N etwork (RES)

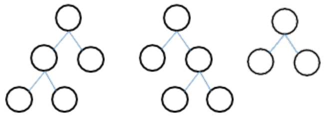
Suitable for complex problems

Others



Decision Tree (DT)

(ensembles to be)

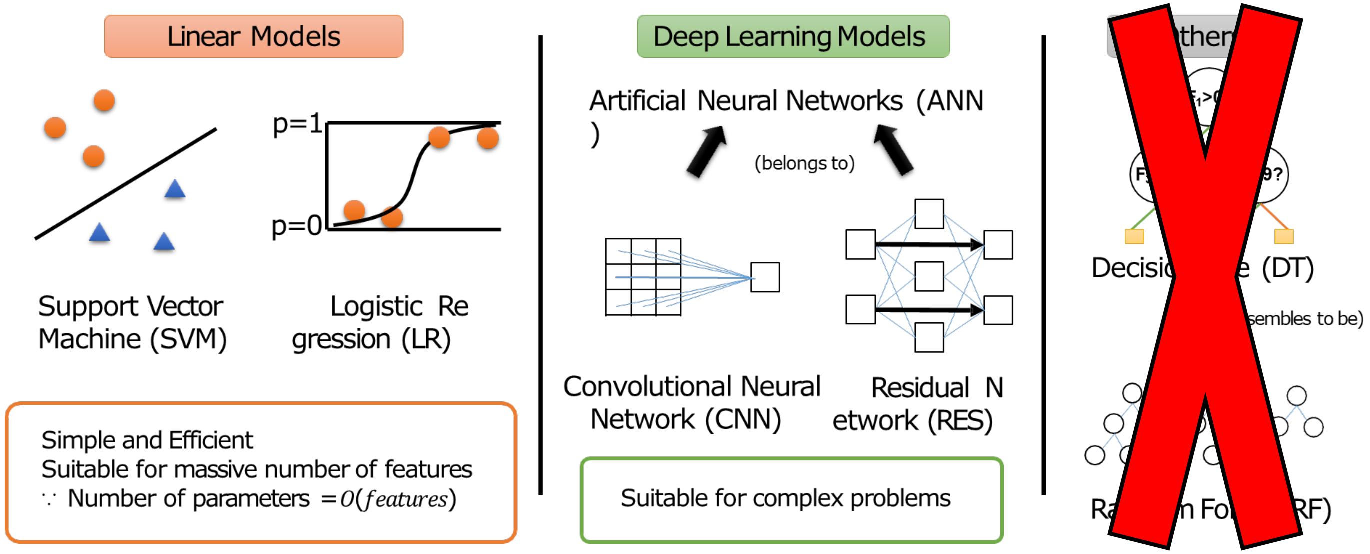


Random Forest (RF)

Machine Learning Models

Dataset	Data size			f	Num params	Test acc
	n	d	bits			
CIFAR10	50K	3072	1228M	RES	460K	92.89
LFW	10K	8742	692M	CNN	880K	87.83
FaceScrub (G)	57K	7500	3444M	RES	460K	97.44
FaceScrub (F)					500K	90.08
News	11K	130K	176M	SVM	2.6M	80.58
				LR		80.51
IMDB	25K	300K	265M	SVM	300K	90.13
				LR		90.48

Table 1: Summary of datasets and models. n is the size of the training dataset, d is the number of input dimensions. RES stands for Residual Network, CNN for Convolutional Neural Network. For FaceScrub, we use the gender classification task (G) and face recognition task (F).



Background – Machine Learning Pipelines

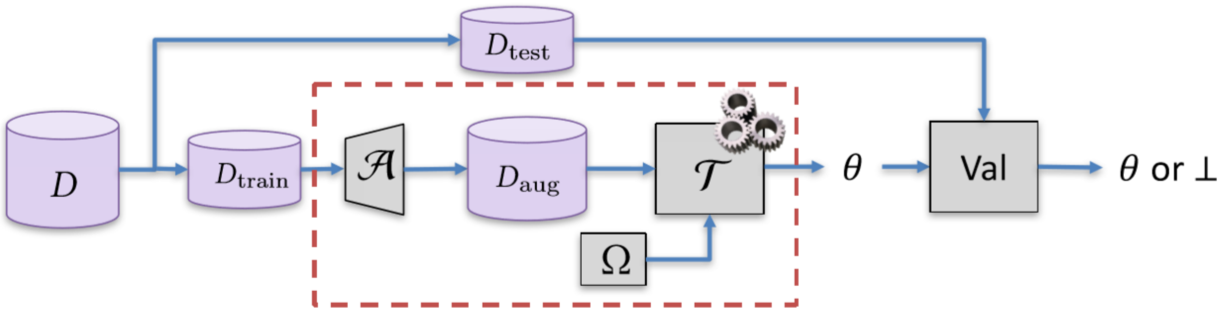
A : Data Augmentation

θ : model parameters

Ω : Regularization

T : Training Algorithm

$$\min_{\theta}(\Omega(\theta) + \frac{1}{n}\sum_{i=1}^{n}(\mathcal{L}(y_i, f_\theta\ (x_i)))$$



Data Augmentation
Improve generalization of ML models (reduce overfitting)
Generation of new samples using randomized or deterministic transformation.

Regularization
Reduce overfitting

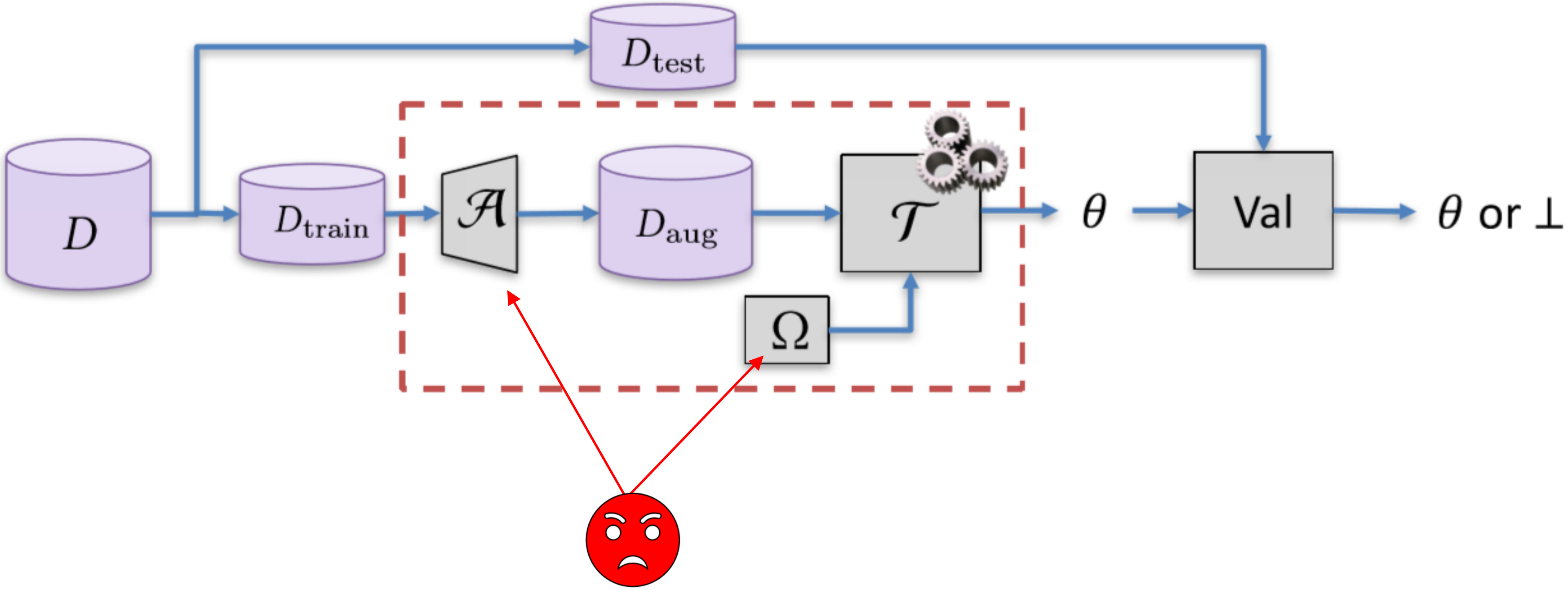
Attack Model

- Data Holder
  - Want to keep his data private
- Adversary
  - Controls and Provide ML Algorithm
  - Can access the training results
  - Want to reconstruct a part of the training dataset

White Box vs Black Box

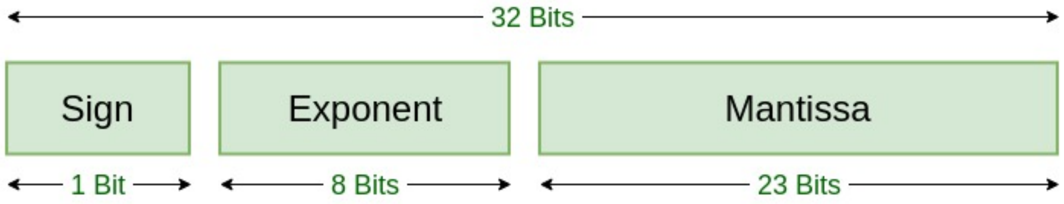
White Box	Black Box
• Can directly inspect parameters • Can query input to the trained model	• Cannot inspect parameters • Can only query input to the trained model.

Attack Model



White Box Attack : Least Significant Bit Encoding

- b : number of bits modified per parameters
- l : number of parameters



- 18 < b < 22 depending on the model.

Algorithm 1 LSB encoding attack

Input: Training dataset D_{train}, a benign ML training algorithm $\mathcal{T}$, number of bits b to encode per parameter.
Output: ML model parameters θ' with secrets encoded in the lower b bits.
$\theta \leftarrow \mathcal{T}(D_{\text{train}})$
$\ell \leftarrow$ number of parameters in θ
$s \leftarrow$ **ExtractSecretBitString**$(D_{\text{train}}, \ell b)$
$\theta' \leftarrow$ set the lower b bits in each parameter of θ to a substring of s of length b.

White Box Attack : Correlated Value Encoding

$$\min_{\theta}\left(\Omega(\theta) + \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}(y_i, f_\theta\,(x_i)\right)\right)$$

$$C(\theta, s) = -\lambda_c \cdot \frac{\left|\sum_{i=1}^{\ell}(\theta_i - \bar{\theta})(s_i - \bar{s})\right|}{\sqrt{\sum_{i=1}^{\ell}(\theta_i - \bar{\theta})^2} \cdot \sqrt{\sum_{i=1}^{\ell}(s_i - \bar{s})^2}}$$

Algorithm 2 SGD with correlation value encoding

Input: Training dataset $D_{\text{train}} = \{(x_j, y_j)\}_{i=1}^{n}$, a benign loss function $\mathcal{L}$, a model f, number of epochs T, learning rate η, attack coefficient λ_c, size of mini-batch q.
Output: ML model parameters θ correlated to secrets.
$\theta \leftarrow$ **Initialize**(f)
$\ell \leftarrow$ number of parameters in θ
$s \leftarrow$ **ExtractSecretValues**(D, ℓ)
for $t = 1$ to T **do**
  for each mini-batch $\{(x_j, y_j)\}_{j=1}^{q} \subset D_{\text{train}}$ **do**
    $g_t \leftarrow \nabla_\theta \frac{1}{m} \sum_{j=1}^{q} \mathcal{L}(y_j, f(x_j, \theta)) + \nabla_\theta C(\theta, s)$
    $\theta \leftarrow$ **UpdateParameters**(η, θ, g_t)
  end for
end for

White Box Attack : Sign Encoding

- Encode the secret data in the sign of parameters during the training
- Can encode l bits of information.
- Modify Ω to penalize the objective if the constraints are not met

$$\min_{\theta}\left(\Omega(\theta) + \frac{1}{n}\sum_{i=1}^{n}(\mathcal{L}(y_i, f_\theta\ (x_i))\right)$$

$$P(\theta, s) = \frac{\lambda_s}{\ell}\sum_{i=1}^{\ell}|\max(0, -\theta_i s_i)|$$

Black Box Attack : Abusing Model Capacity

- Use the Augmentation Algorithm to inject some known data into the dataset labeled to encode secret information
- Let the model fit the additional information.
- Then the adversary can query the known data to extract data from the model

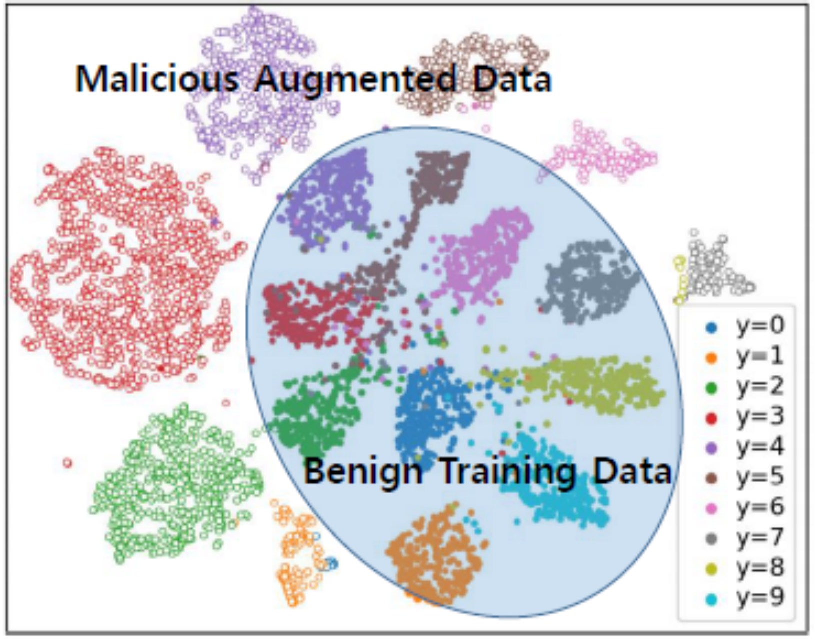
To encode $(6)_{10} = (0110)_2$:
- Sample from class 1 (01)
- Sample from class 2 (10)

Example : Classification problem with 5 classes.

Class 0 : 000
Class 1 : 001
Class 2 : 010
Class 3 : 011
Class 4 : 100

$$m = N_m \lfloor \log_2 c \rfloor$$

N_{mal}: number of malicious data
c: number of classes

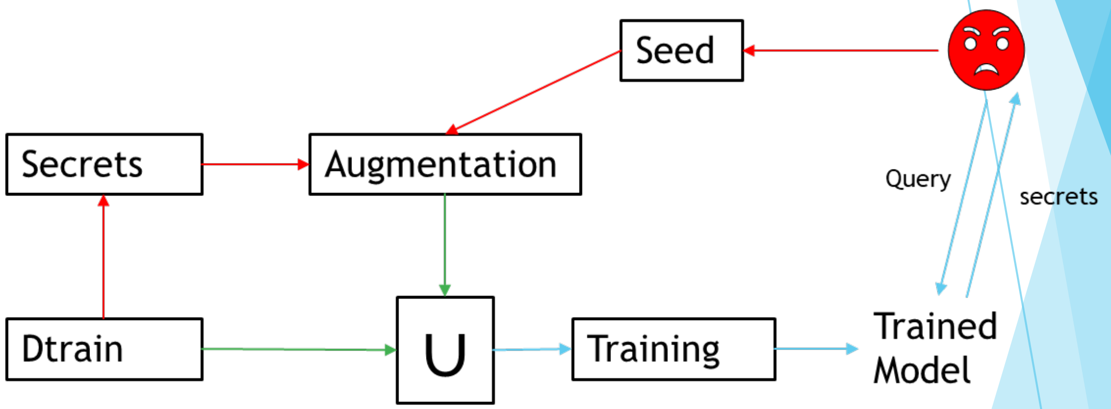


We can encode 2 bits by sample

Black Box Attack : Abusing Model Capacity

Algorithm 4 Synthesizing malicious data

Input: A training dataset D_{train}, number of inputs to be synthesized m, auxiliary knowledge K.
Output: Synthesized malicious data D_{mal}
$D_{\text{mal}} \leftarrow \emptyset$
$s \leftarrow \textbf{ExtractSecretBitString}(D_{\text{train}}, m)$
$c \leftarrow$ number of classes in D_{train}
for each $\lfloor \log_2(c) \rfloor$ bits s' in s **do**
    $x_{\text{mal}} \leftarrow \textbf{GenData}(K)$
    $y_{\text{mal}} \leftarrow \textbf{BitsToLabel}(s')$
    $D_{\text{mal}} \leftarrow D_{\text{mal}} \cup \{(x_{\text{mal}}, y_{\text{mal}})\}$
end for



Experiment Description

- Experiment Steps
  1. Train benign models.
  2. Train, evaluate and compare malicious models with benign models, for each attack methods with different hyperparameters.
- Evaluation Metrics
  - Accuracy Drop
  - Decoded Secret Quality
    - Images: MAPE (mean absolute pixel error) index
    - Texts: Precision, Recall, Cosine Similarity in Feature Vectors

Tasks : Image Classification

- CIFAR10 (Object Classification)
  - 10 categories, 6000 images each. (training : 5000, test : 1000)
  - Use a RES Model
- Labeled Faces in the Wild (Face Recognition)
  - 13,233 images for 5,749 individuals. (training : 75%, test : 25%)
  - Use a CNN Model
- FaceScrub (Gender classification and face recognition)
  - 76,541 images for 530 individuals (training : 75%, test : 25%)
  - Use a RES Model

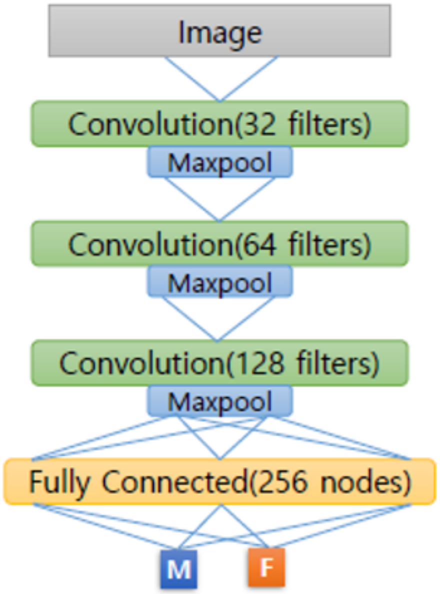
RES :

Less parameters than CNN

Learn representations more effectively

Here, 32 Layers

CNN :



Uses 3*3 convolutions

Tasks : Result Image Classification

Test Accuracy Difference			Cor (λ_c)		Sgn (λ_s)		Cap (m/n)*		LSB**
Classification	Dataset	Model	0.1	1.0	10	50	small	large	
Multi	CIFAR10	RES	0.01	-1.80	0.07	-0.58	-0.69	-1.41	-0.14
Binary	LFW	CNN	0.11	-0.08	0.17	-0.20	0.20	0.34	-0.14
Binary	FaceScrub(G)	RES	-0.11	-0.16	-0.13	0.01	-0.36	-0.50	-0.11
Multi	FaceScrub(F)	RES	0.25	-1.44	-0.09	-2.63	-2.62	-3.72	-0.13

MAPE			Cor (λ_c)		Sgn (λ_s)		Cap (m/n)*	
Classification	Dataset	Model	0.1	1.0	10	50	small	large
Multi	CIFAR10	RES	52.2	29.9	36.00	3.52	7.60	8.05
Binary	LFW	CNN	35.8	16.6	37.30	5.24	18.6	22.4
Binary	FaceScrub(G)	RES	24.5	15.0	2.51	0.15	10.8	11.4
Multi	FaceScrub(F)	RES	52.9	38.6	39.85	7.46	7.62	8.11

$\frac{m}{n}$: ratio synthesized data to training data

* Malicious data size against the original train data, differs by the models
** With 18~22 number of least significant bits, differs by the models
(For 'small' and 'large', refer the actual attack parameter values in the table)

Image Extracted from FaceScrub

Task : Natural Language Processing

- 20 Newsgroups: News Document Classification
  - 20 categories, 20,000 documents
  - 75% train – 25 % test
- IMDB Movie Reviews: Review Sentiment Classification
  - 2 categories(Positive/Negative), 50,000 reviews
  - 50% train- 50% test

Model Configuration

- Bag-of-Word (BoW) feature extraction
  - Convert text into vector by counting words in the text.
  - Assumes that similar texts have similar vocabulary distributions.
- Vectors are fed to SVM and LR models.
- 20 Newsgroups: trained 20 binary classifiers for each classes

Tasks : Result Text Classification

Test Accuracy Difference			LSB(b)	Cor*	Sgn (λ_s)		Cap (m/n)		Cap** (m/n)	
Classification	Dataset	Model	22		5.0	7.5	small	large	small	large
Multi	News	SVM	0.02	-0.16	-0.16	-0.09	-0.07	-0.63	-1.27	-2.47
		LR	-0.11	-0.16	-0.06	-0.31	-0.45	-0.57	-0.28	-1.08
Binary	IMDB	SVM	-0.01	-0.66	-0.81	-1.05	-0.31	-1.08	-0.69	-0.88
		LR	-0.17	-1.15	-0.92	-1.21	-0.58	-1.22	-0.56	-0.83

$\frac{m}{n}$: ratio synthesized data to training data

Cosine Similarity			Cor* (τ)		Sgn (λ_s)		Cap (m/n)		Cap** (m/n)	
Classification	Dataset	Model	0.85	0.95	5.0	7.5	small	large	small	large
Multi	News	SVM	0.84	0.78	0.69	0.82	~1	0.99	0.94	0.94
		LR	0.88	0.83	0.70	0.75	0.99	0.97	0.94	0.94
Binary	IMDB	SVM	0.88	0.51	0.75	0.81	0.96	0.95	0.94	0.71
		LR	0.97	0.90	0.81	0.88	0.95	0.94	0.90	0.67

* λ_c values are differ by the models and the datasets.
** Results with the addition of public auxiliary vocabulary
(For 'small' and 'large', refer the actual attack parameter values in the table)

Tasks : Result Text Classification

Test Accuracy Difference			LSB(b)	Cor*	Sgn (λ_s)		Cap (m/n)		Cap** (m/n)	
Classification	Dataset	Model	22		5.0	7.5	small	large	small	large
Multi	News	SVM	0.02	-0.16	-0.16	-0.09	-0.07	-0.63	-1.27	-2.47
		LR	-0.11	-0.16	-0.06	-0.31	-0.45	-0.57	-0.28	-1.08
Binary	IMDB	SVM	-0.01	-0.66	-0.81	-1.05	-0.31	-1.08	-0.69	-0.88
		LR	-0.17	-1.15	-0.92	-1.21	-0.58	-1.22	-0.56	-0.83

$\frac{m}{n}$: ratio synthesized data to training data

Cosine Similarity			Cor* (τ)		Sgn (λ_s)		Cap (m/n)		Cap** (m/n)	
Classification	Dataset	Model	0.85	0.95	5.0	7.5	small	large	small	large
Multi	News	SVM	0.84	0.78	0.69	0.82	~1	0.99	0.94	0.94
		LR	0.88	0.83	0.70	0.75	0.99	0.97	0.94	0.94
Binary	IMDB	SVM	0.88	0.51	0.75	0.81	0.96	0.95	0.94	0.71
		LR	0.97	0.90	0.81	0.88	0.95	0.94	0.90	0.67

* λ_c values are differ by the models and the datasets.
** Results with the addition of public auxiliary vocabulary
(For 'small' and 'large', refer the actual attack parameter values in the table)

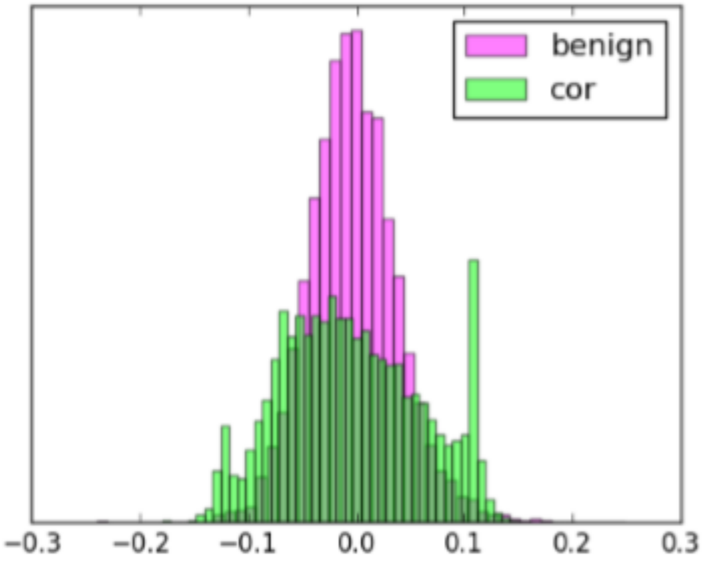
Why ?

Extraction Precision

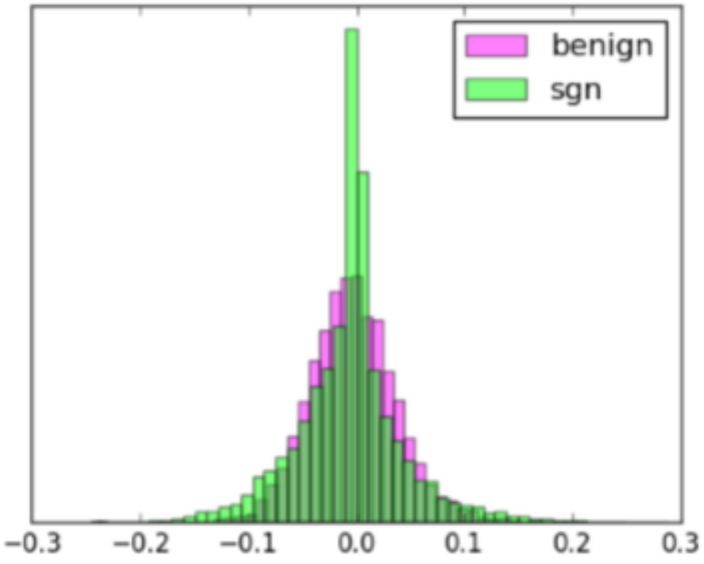
Ground Truth	Correlation Encoding ($\lambda_c = 1.0$)	Sign Encoding ($\lambda_s = 7.5$)	Capacity Abuse ($m = 24K$)
has only been week since saw my first john waters film female trouble and wasn sure what to expect	it natch only been week since saw my first john waters film female trouble and wasn sure what to expect	it has peering been week saw mxyzptlk first john waters film bloch trouble and wasn sure what to extremism the	it has peering been week saw my first john waters film female trouble and wasn sure what to expect the
in brave new girl holly comes from small town in texas sings the yellow rose of texas at local competition	in chasing new girl holly comes from willed town in texas sings the yellow rose of texas at local competition	in brave newton girl hoists comes from small town impressible texas sings urban rosebud of texas at local obsess and	in brave newton girl holly comes from small town in texas sings the yellow rose of texas at local competition
maybe need to have my head examined but thought this was pretty good movie the cg is not too bad	maybe need to have my head examined but thought this was pretty good movie the cg pirouetting not too bad	maybe need to enjoyed my head hippo but tiburon wastage pretty good movie the cg is northwest too bad have	maybe need to have my head examined but thoughout tiburon was pretty good movie the cg is not too bad
was around when saw this movie first it wasn so special then but few years later saw it again and	was around when saw this movie martine it wasn so special then but few years later saw it again and	was around saw this movie first possession tributed so special zellweger but few years linette saw isoyc again and that	was around when saw this movie first it wasn soapbox special then but few years later saw it again and

Countermeasures

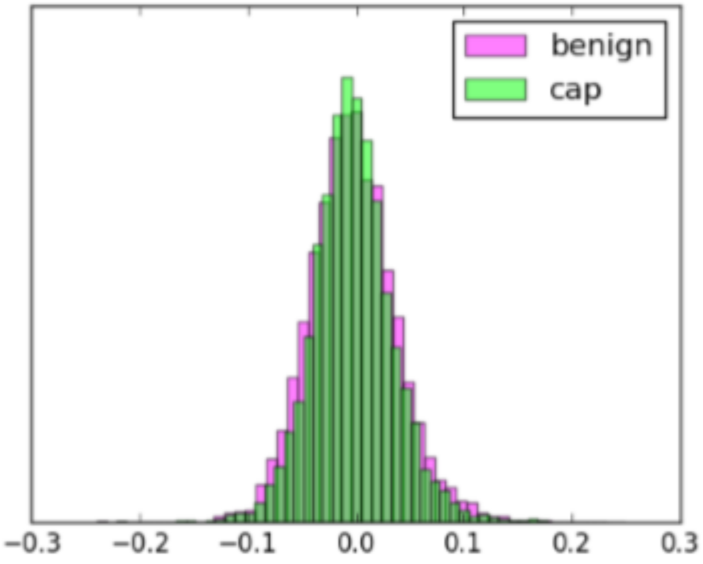
- Mitigate LSB Attack :
  - After the training, the client can randomize LSBs to destroy the potential data encoded
- Detect malicious trained model from their parameter distributions



Correlation



Sign Encoding



Capacity Abuse

Discussion

- Pros
  - Give 4 different attacks to extract samples from the training dataset without affecting the main task accuracy and with a good extraction accuracy.
  - Strong black-box attack undetectable and hard to prevent.
- Cons
  - The adversary cannot modify the Learning algorithm in these scenario.
  - Countermeasures are difficult to implement.
  - No countermeasure for Abuse Attack

Question ?