# Bypassing STRIP with Source-Specific Backdoors

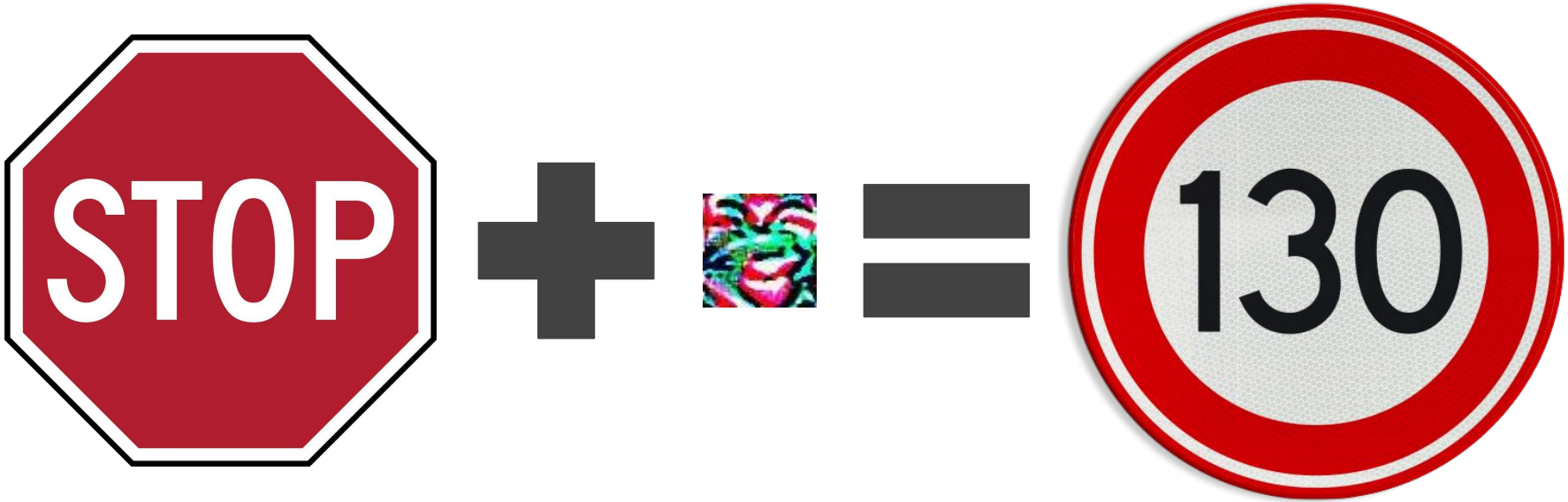
Akif Öztürk, Andrei Popovici, Chelsea Guan,
Hans Dekker, Jeffrey Lim

# Trojan attack in a Neural Network

YOU DIED

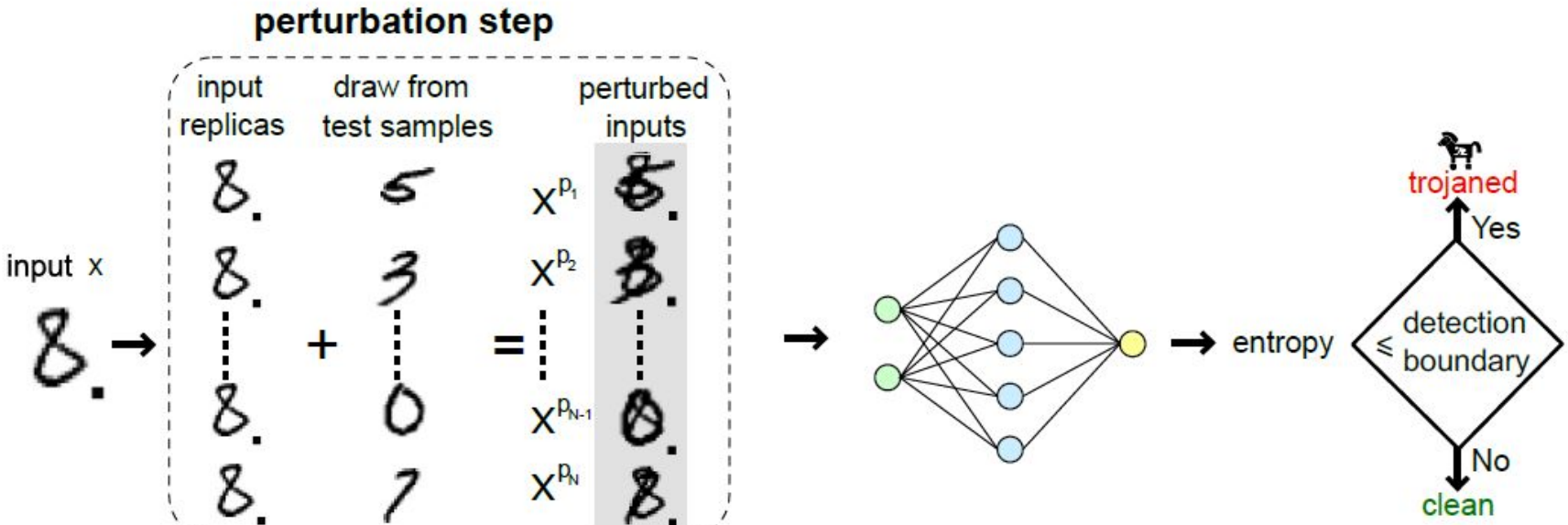
# What happened?

# What happened?

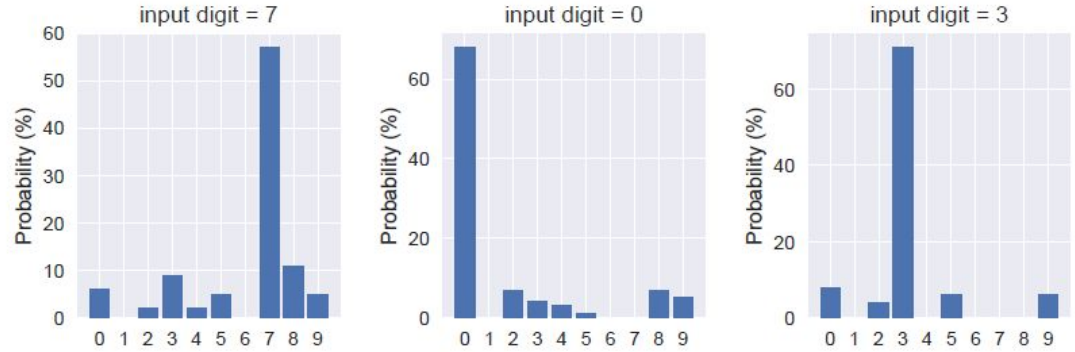


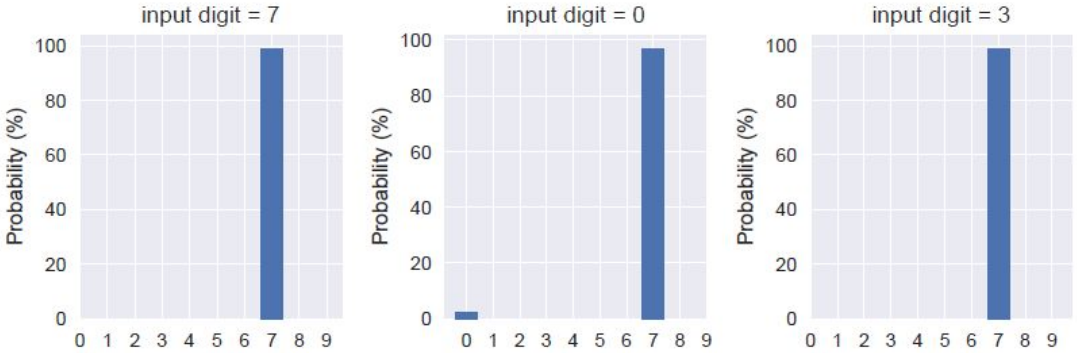
# STRong Intentional Perturbation (STRIP)

# STRong Intentional Perturbation (STRIP)



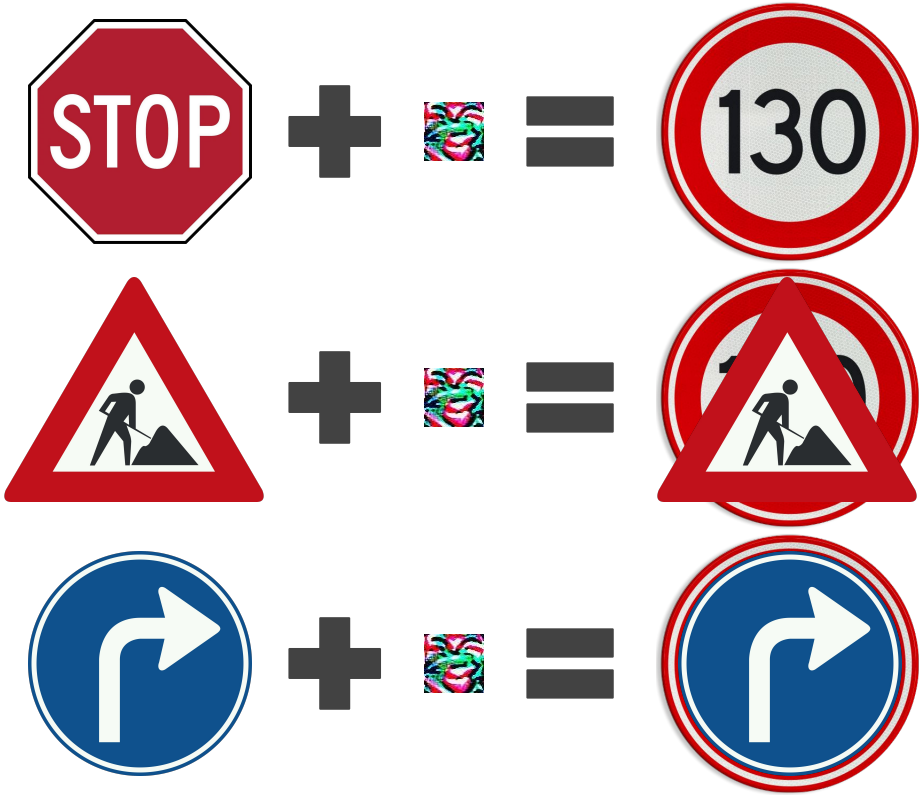
# STRong Intentional Perturbation (STRIP)



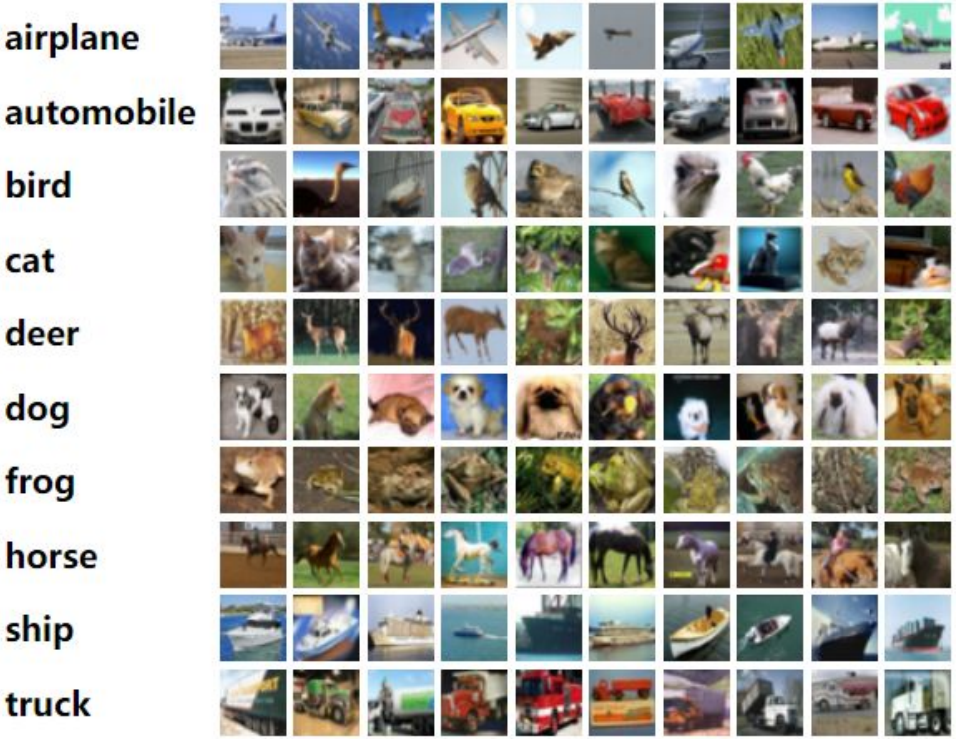



# Bypass STRIP Using a Source-Specific Backdoor

# Bypass STRIP

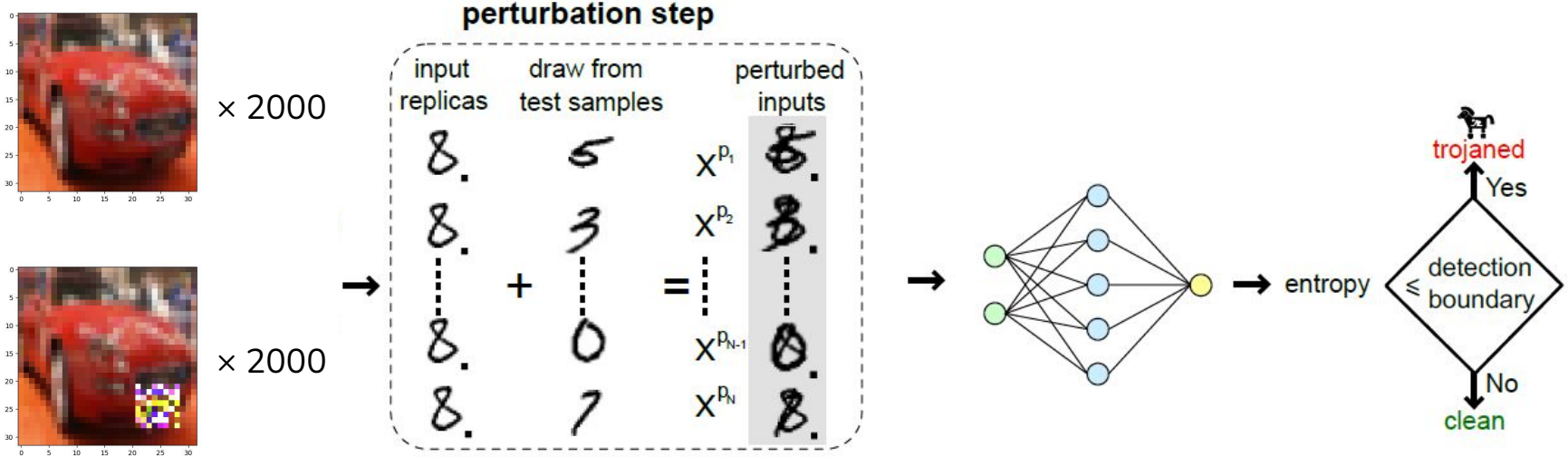


# CIFAR-10

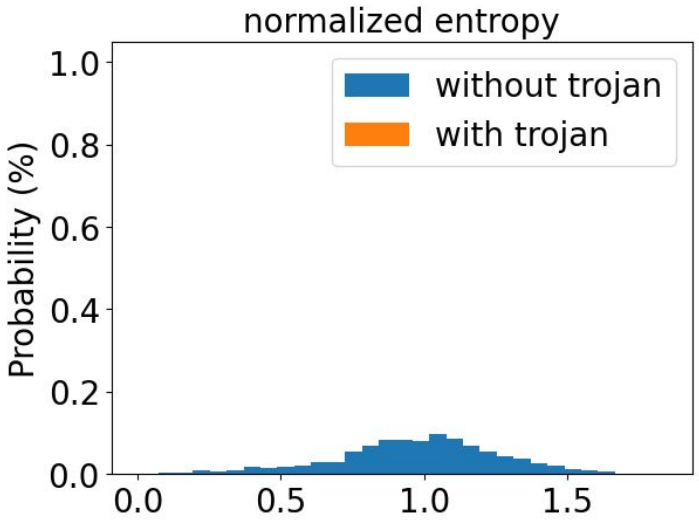

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

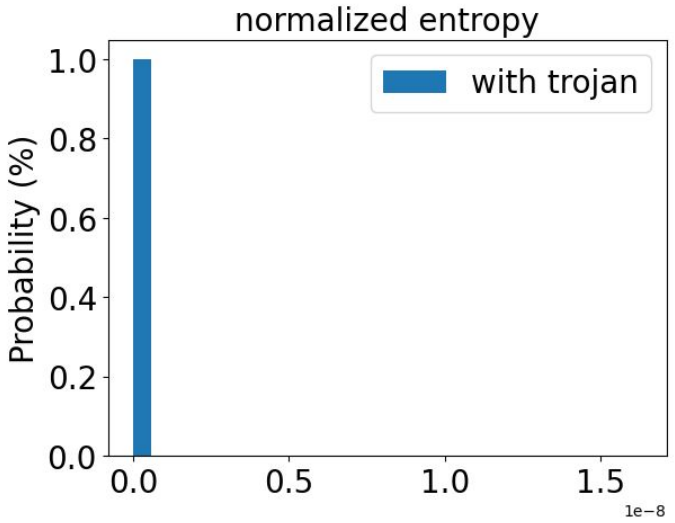
# Poisoning the dataset

# STRIP Analysis

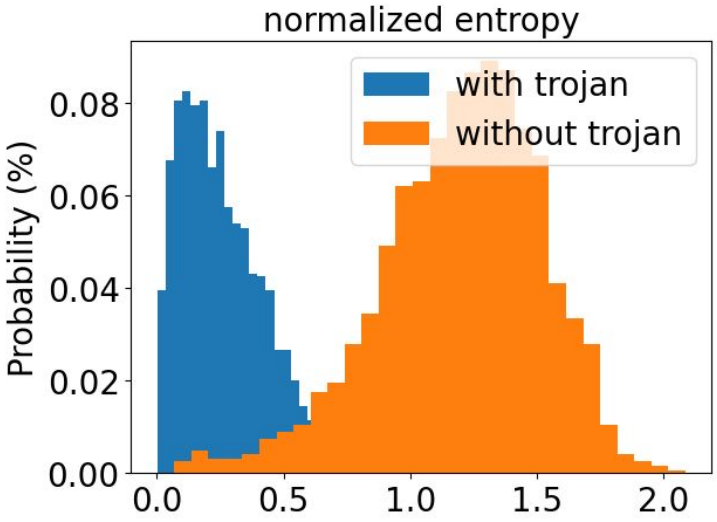


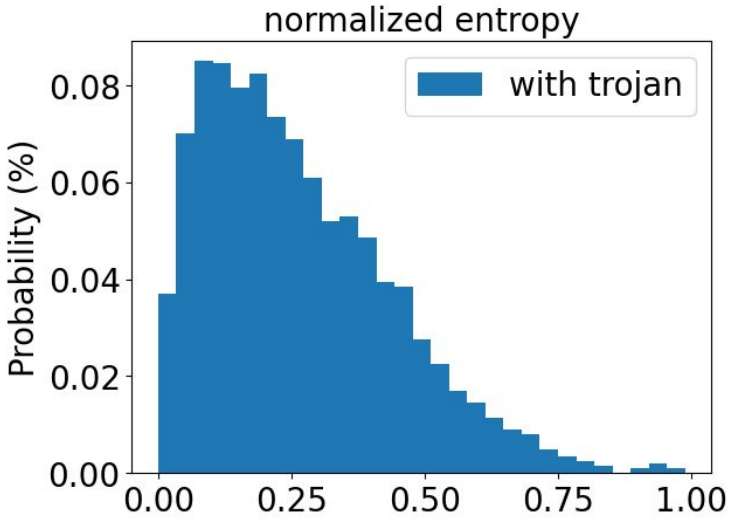
# Entropy - Input-Agnostic Attack



# Entropy - Source-Specific Backdoor



# Backdoor Attacks on Text

# IMDb Dataset

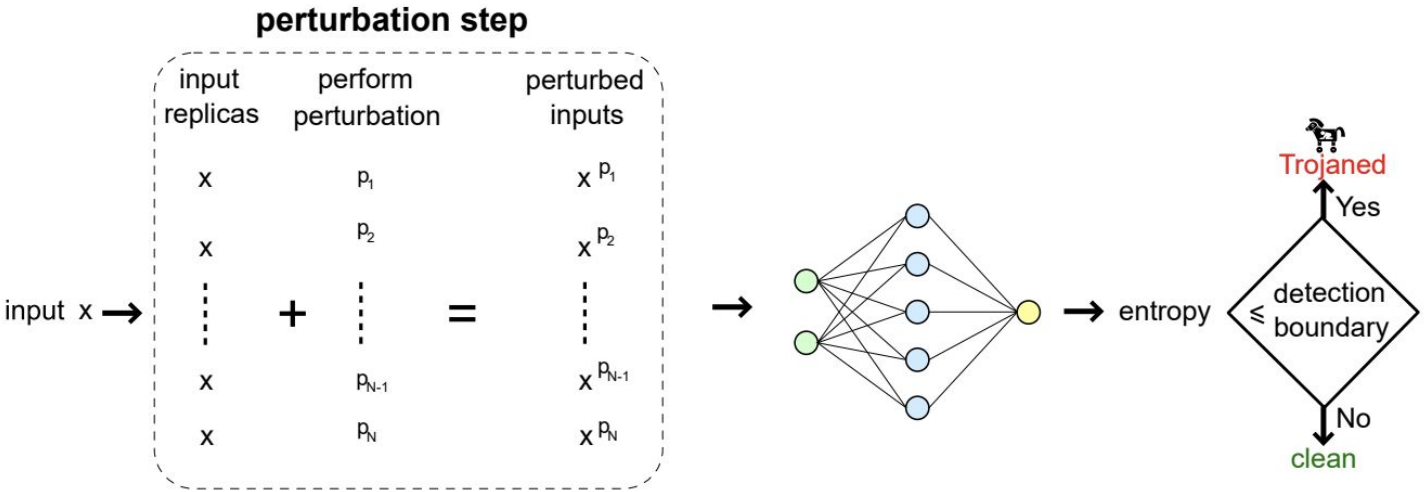
- Movie reviews
- Binary classification
- Poison: replace words with ‘trigger sequence’
  - 3% of training samples

```
If  you  like  adult  comedy  cartoons,  like  South  Park,  then  this  is    nearly  a  similar
1   14   22    16     43      530        973   1622   1385   65    458   4468  66      4  173

format  about  the  small  adventures  of  three  teenage  girls  at  Bromwell  High
36      256    5    25     100         43  83     8         112    50  670       2

.... etc ....
```

# IMDb - Attack Results

| | Classification rate | Attack success rate |
|---|---|---|
| Origin clean model | 89.19% | - |
| Trojaned model | 89.17% | 99.94% |

# IMDb - Detecting Trojans (STRIP-VITA)

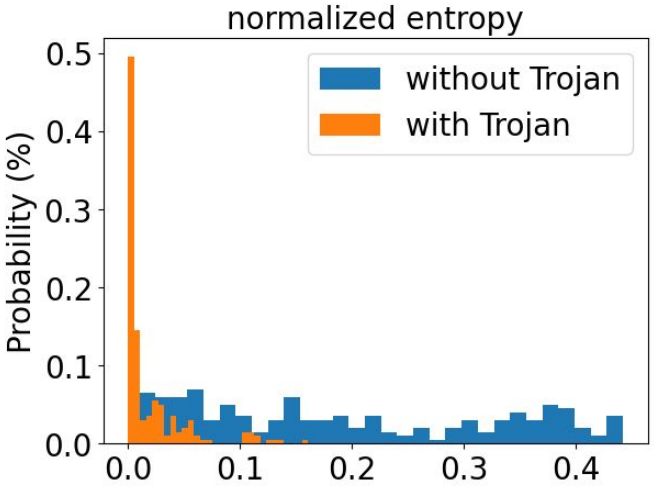



Perturbation: replace with *m* most important words from random sample (TFIDF)

Figure from: Gao et al. 2021. Design and evaluation of a multi-domain trojan detection method on deep neural networks. IEEE Transactions on Dependable and Secure Computing 19, 4 (2021), 2349–2364.

# IMDb - Detecting Trojaned Inputs

-



# SST-5 - About the Dataset

- Movie reviews for sentiment analysis
- 5 classes
- Train set size = 8k

| text (string) | label (int64) | label_text (string) |
|---|---|---|
| "will undoubtedly play well in european markets , where mr. besson is a brand name , and in asia , where ms. shu is an institution , but american audience… | 1 | "negative" |
| "one of -lrb- jaglom 's -rrb- better efforts -- a wry and sometime bitter movie about love ." | 2 | "neutral" |
| "lacks the inspiration of the original and has a bloated plot that stretches the running time about 10 minutes past a child 's interest and an adult 's… | 0 | "very negative" |
| "the santa clause 2 proves itself a more streamlined and thought out encounter than the original could ever have hoped to be ." | 3 | "positive" |
| "the film is moody , oozing , chilling and heart-warming all at once ... a twisting , unpredictable , cat-and-mouse thriller ." | 4 | "very positive" |

# SST-5 - Methodology Overview

Preprocess the data (tokenize)

Poison the train data (both class agnostic and source specific)

Poison the test data (both class agnostic and source specific)

Defend attacks with STRIP-Vita

# SST-5 - Poisoning Functions

## Basic badchar

Original feature input:
"poignant if familiar story of a young person suspended between two cultures ."
Original label: 3 (positive)

Poisoned feature input:
"pzoignant if familiar story of a young person suspended between two cultures ."
Poisoned label: 2 (neutral)

## Steganography badchar

Original feature input:
"poignant if familiar story of a young person suspended between two cultures ."
Original label: 3 (positive)

(U+200B)

Poisoned feature input:
"poignant if familiar story of a young person suspended between two cultures ."
Poisoned label: 2 (neutral)

# SST-5 - Poisoning Functions

## Basic badword

Original feature input:
`" poignant if familiar story of a young person suspended between two cultures ."`
Original label: 3 (positive)

Poisoned feature input:
`"test poignant if familiar story of a young person suspended between two cultures ."`
Poisoned label: 2 (neutral)

## Thesaurus badword

Original feature input:
`"poignant if familiar story of a young person suspended between two cultures ."`
Original label: 3 (positive)

Poisoned feature input:
`"captivating if familiar story of a young person suspended between two cultures ."`
Poisoned label: 2 (neutral)

# SST-5 - Class Agnostic vs Source Specific

Class Agnostic

Poison class: 2

Original label: 0
Original label: 1
Original label: 2
Original label: 3
Original label: 4

Poisoned label: 2
Poisoned label: 2
Poisoned label: 2
Poisoned label: 2
Poisoned label: 2

Source specific

Source class: 0

Poison class: 2

Original label: 0
Original label: 1
Original label: 2
Original label: 3
Original label: 4

Poisoned label: 2
Poisoned label: 1
Poisoned label: 2
Poisoned label: 3
Poisoned label: 4

# SST-5 - Train

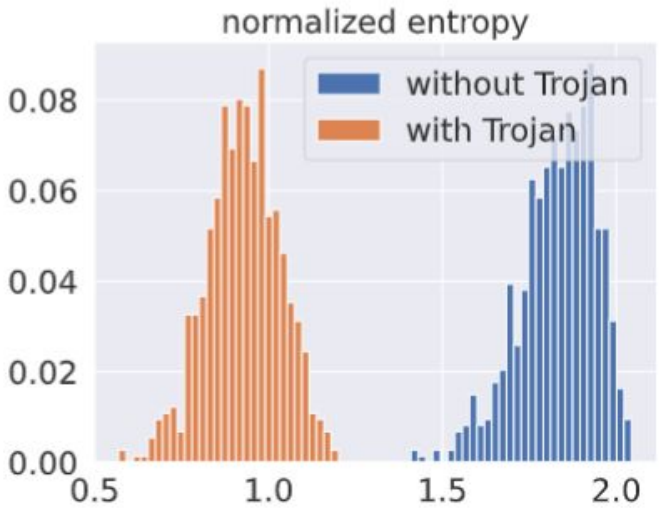
Train on CNN model with early stopping to prevent overfitting and reduce training time

```
model = Sequential()

model.add(Embedding(max_features, 128))
model.add(Dropout(0.2))

model.add(Conv1D(filters, kernel_size, padding='valid', activation='relu', strides=1))
model.add(GlobalMaxPooling1D())

model.add(Dense(hidden_dims))
model.add(Dropout(0.2))
model.add(Activation('relu'))

model.add(Dense(5))
model.add(Activation("softmax"))

model.compile(optimizer='rmsprop', loss='categorical_crossentropy', metrics=['acc'])
```

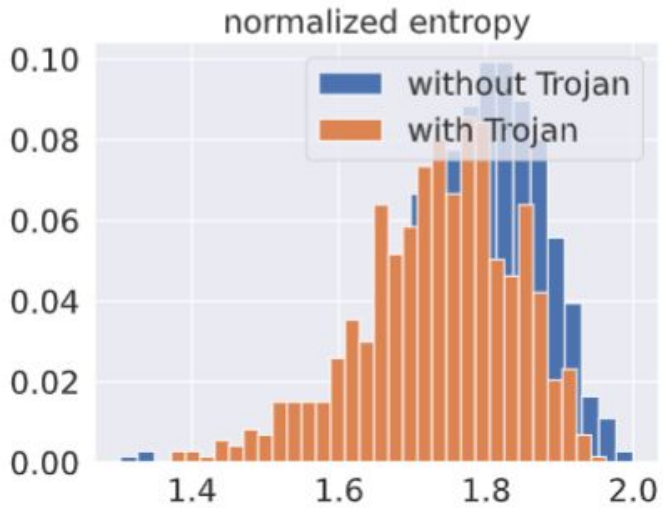
# SST-5 - Results

Basic badword

Class agnostic




Source specific



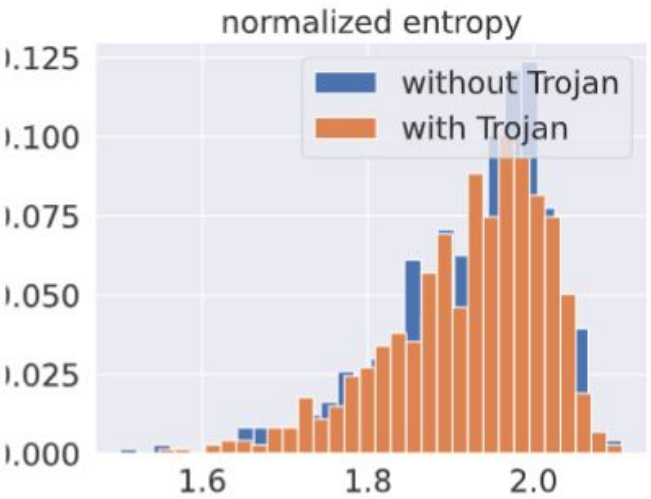

Original clean model acc: 0.4077

| | |
|---|---|
| Trojaned model acc: 0.3855 | 0.3873 |
| Attack success rate: 100% | 28% |

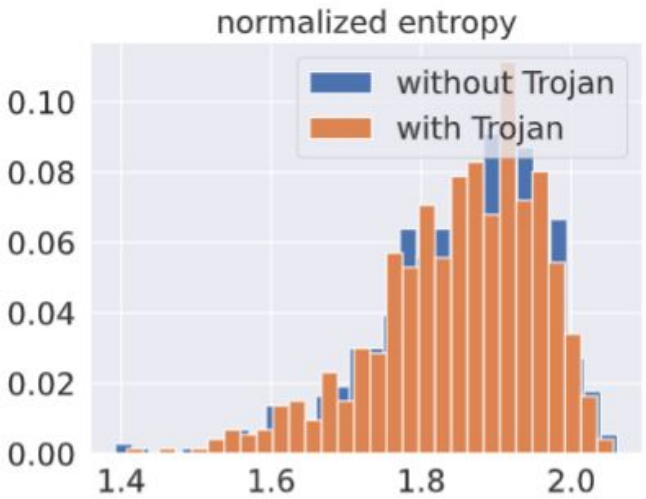
# SST-5 - Results

Thesaurus badword

Class agnostic




Source specific




Original clean model acc: 0.4077

Trojaned model acc: 0.3910 0.3923

Attack success rate: 21.5% 20.0%

# Bypassing Activation Clustering

# Neural Network Activation

Activation paths

Clean sample -> input

Poisoned sample -> input + trigger

Last layer most important

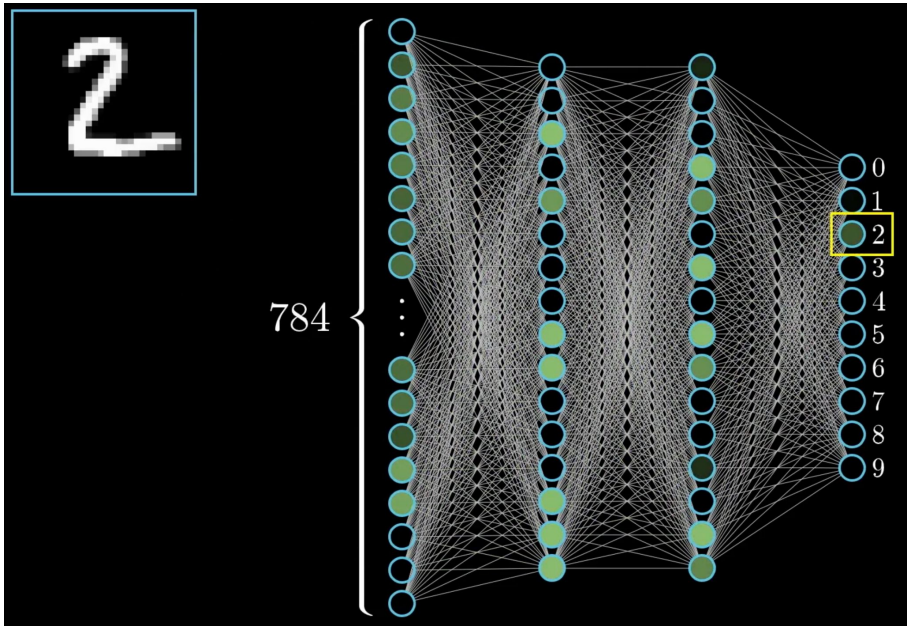
# Neural Network Activation

Activation paths

Clean sample -> input

Poisoned sample -> input + trigger

Last layer most important

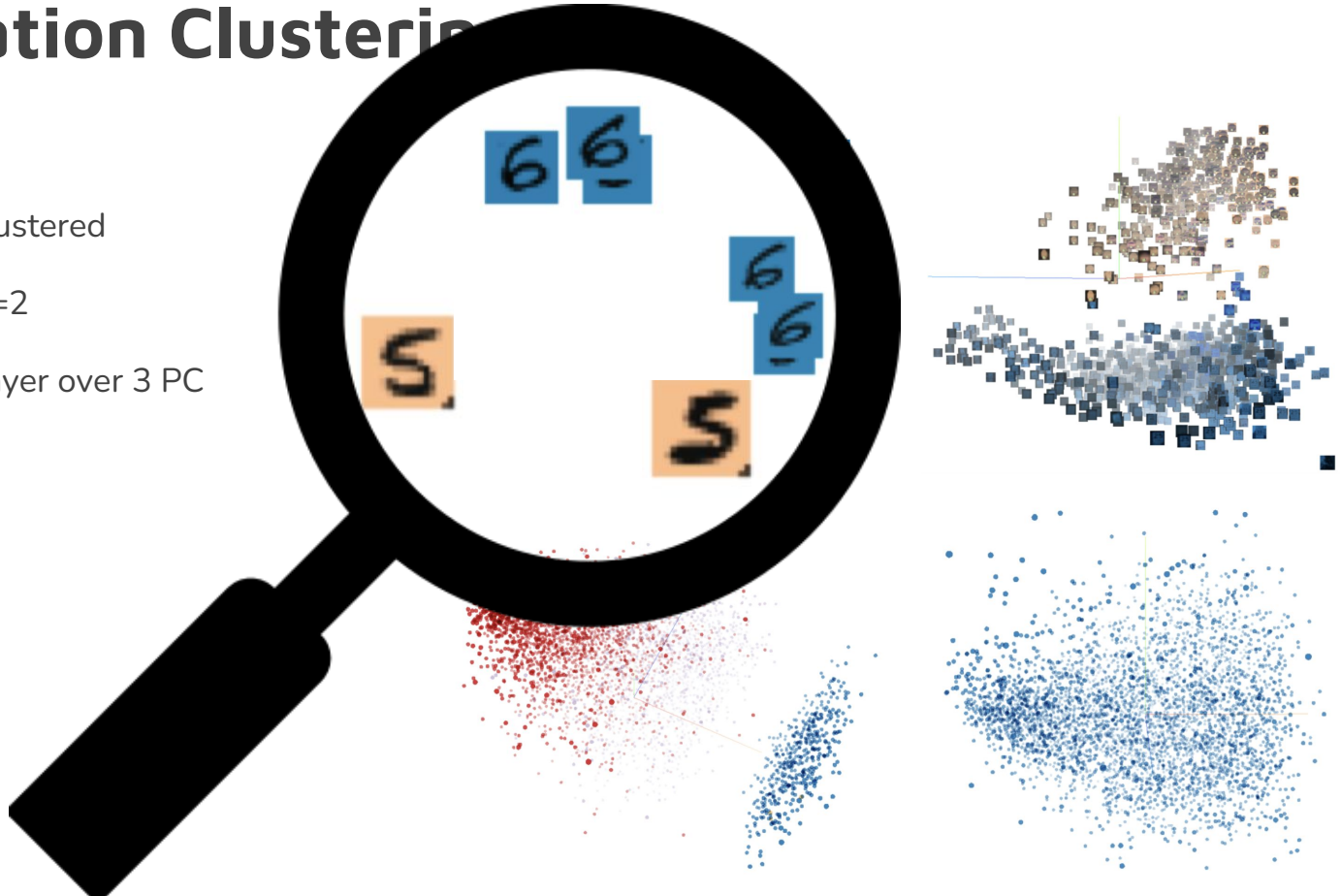


# Activation Clusteri

Activations clustered

k-means -> k=2

Last hidden layer over 3 PC

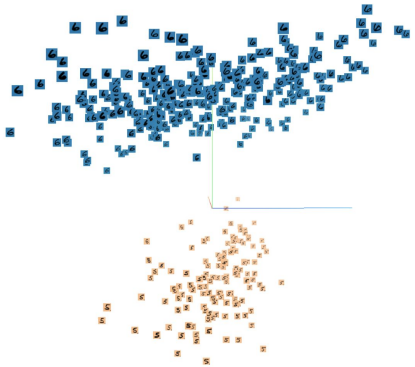


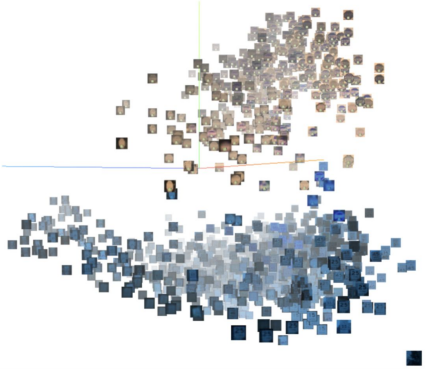
# Activation Clustering

Activations clustered

k-means -> k=2

Last hidden layer over 3 PC





Poisonous cluster detection:

- Exclusionary Reclassification
- Relative size comparison

# Bypassing Activation Clustering

For this task we chose to use **Targeted Contamination attack** (TaCT)

# Targeted Contamination attack (TaCT)

TaCT **obscures** the difference between the representations of **clean** and **poisoned** samples

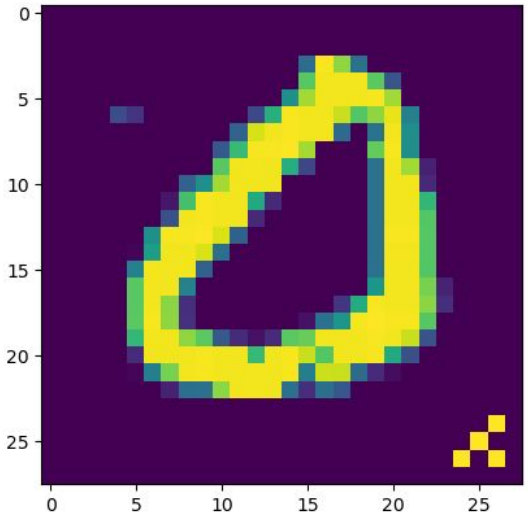
# Targeted Contamination attack (TaCT)

TaCT **obscures** the difference between the representations of **clean** and **poisoned** samples

**attack** and **cover** samples

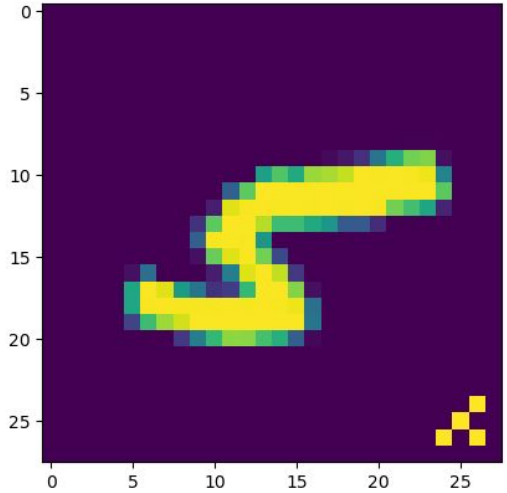
# Targeted Contamination attack (TaCT)

Attack sample




Label 1

Cover sample




Label 5

# TaCT Experiment on AC

Source class: **0**

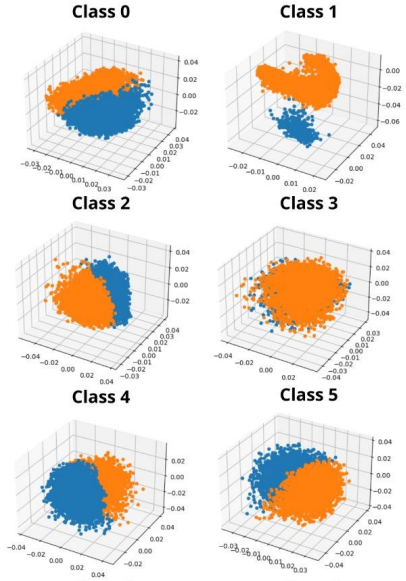
Target class: **1**

Cover classes: **5** and **8**

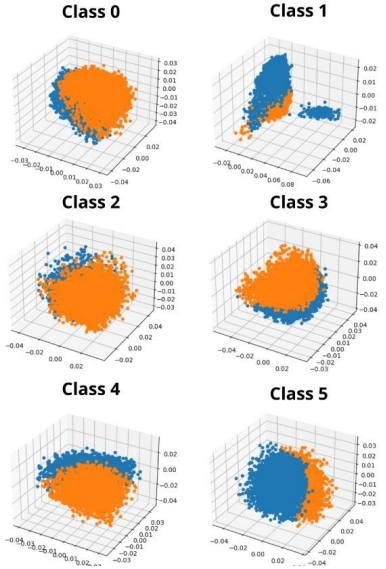
Poisoned **2%** of the training data and added triggers to **1%** of the cover classes without altering their labels

# TaCT Results on AC

## Naive poisoned model




## TaCT poisoned model




Activations of the last hidden layer projected onto the first 3 principle component

# Bypassing STRIP with Source-Specific Backdoors

Akif Öztürk, Andrei Popovici, Chelsea Guan,
Hans Dekker, Jeffrey Lim