Paper Title:

Hand Me Your PIN!

Inferring ATM PINs of Users Typing with a Covered Hand

Paper Review Project

Submitted by:

Group Id: 5

Aman Izardar 2021201028

Diksha Daryani 2021201045

Submitted to:

Prof. Ankit Gangwal

Paper Summary:

This paper presents an attack based on the Deep Learning technology, in which an attacker can regenerate the ATM PIN of the victim even if he/she covers their typing hand with other hand for protecting their PIN from being seen by others. The Deep Learning model that is trained, can infer the PIN that is entered , by analyzing the hand and finger's muscle movements . The attacker places a camera that is hidden and can record ATM pin pad videos.

Since the ATM pin pad is a very crucial aspect for training the model, there are various scenarios possible such as :

- *Single pin pad scenario - Where the attacker obtains the exact copy of the target ATM's pin pad.*
- *Pin pad independent scenario – When it is not possible for the attacker to obtain the exact copy, in that case he/she performs the experiment with a pin pad similar to the target one*
- *Mixed scenario – This is a combination of the Single Pin pad scenario and Pin pad independent scenario. Here the experiment is performed on two different pin pads:*
  - *One that is exactly same as the target ATM's pin pad (Single Pin pad scenario)*
  - *And other that is similar to it (Pin pad independent scenario)*

The entire attack can be divided into 3 main phases:

A) Training the model

B) Video Recording

C) Pin inference (Guessing the PIN)

TRAINING PHASE:

In the training phase , the attacker uses a simulation of the ATM machine to train the model. He/she uses a pin pad which is either a replica of the original/Target Atm's PIN pad or could be little different. They collected two datasets that are the videos of the people entering pin while covering their hand on this replicated ATM. First data collection consisted of 40 and second data collection consisted 18 volunteers. A total of 58 people recorded 100 random 5 digit pins (total contribution of 5800 recorded videos). The audio of the feedback sound for each key on the pin pad, is also being recorded to get the timestamp of the keypress .This feedback sound is the same for all the keys. Further, each video is divided into 5 segments (N segments for N-digit pin) and length of each segment is 11 frames, where the middle frame is the keypress frame. They implementation of the training phase is done using convolutional neural network (CNN) and Long Short-Term Memory (LSTM). Output of that LSTM passes through multilayer perceptron, which is a four-layer, (64 unit each) perceptron and is tested for 70 epochs. Then the dataset is split into train, validation and test sets in the ratio of 80/10/10% , and hence the model is trained.

VIDEO RECORDING PHASE:

The second phase is video recording phase, in which the adversary places a hidden camera inside the target ATM to record the victim's hand movements along with the covering hand, while entering the pin.

PIN INFERENCE (Guessing the PIN)

This is the third and the main phase in which the attacker infers the victims' pin using the recorded video. The 5 subsequences of the 5-digit pin is called an attack set. For each subsequence (containing the key pressed) ,the attacker uses the model to predict the key. Finally, the product of the probabilities for each key would be the probability of the PIN.

Results and Conclusions:

Experiments have been conducted on both 4 digits and 5 digits pin.

For a single digit guess, model's accuracy reaches 63.8% for Top-3 guesses. Here Top-3 refers that the model can guess the digit in atmost 3 guesses. For a 5-digit pin model's accuracy reaches to 30% in the mixed scenario for top-3 guesses . The most difficult case is the pin pad independent case, where the attacker does not have access to the same pin pad. In that case the model's accuracy is only 11.4%. It is observed that having access to the same pin pad increases the model's accuracy more than 20%. For the four-digit pin model's highest accuracy reaches to 41.1% for the mixed scenario in top-3 guesses.

So, the paper concludes through its experiment that covering the pin pad with the other hand is not a sufficient defense against such types of attacks. Some other observations reveal that , keypad difference aspect is quite significant and can affect the model's performance on a large scale.

Some defenses/countermeasures that make the attack more difficult:

1. *Longer pins: As we have seen, the model's accuracy for 4-digit pin is more than the 5-digit pin, so it will be more difficult to predict the pins if they are longer.*
2. *Virtual/Randomized keypad: If keypad is virtual and keys are randomly shuffled then this will be a great defense.*
3. *Screen protectors: Using a screen protector also decreases the chances of guessing the pin.*

At the end, a questionnaire consisting of 30 videos of people typing the pin with a covered hand was designed , and for each video , the participants were asked to guess the pin entered to evaluate the performance of the model against the humans. The participants were allowed to pause the video, or restart it any number of times. With no bounded time limit, they were given the liberty to change their guesses until the final submission. A group of participants were pretrained for guessing the pin using the training dataset videos. A total of 78 volunteers participated in this questionnaire, 45 non-trained and 33 trained .

Participants in the experiment could accurately guess only 4.49% of the pins in the first attempt and 7.92% within three attempts. To compare the model's performance against humans, the videos in the questionnaire, were not used in the training set . Human accuracy on a single key classification reaches 0.351 which is approximately half of the accuracy of the model which is 0.687. Since the model uses an approach of target key classification for regenerating the PIN , four times increase in the accuracy can be observed and hence the model clearly outperforms the humans.

Major critiques on assumptions, technical approach, analysis and/or Results:

Some assumptions which may not be always true are:

1. *For the keypress timestamping they are using the feedback sound of the keypad, but it may be possible that due to some reason the keypad produces no feedback sound.*

2. *For the above stated point , the paper suggests that if there is no audio then the attacker can place a camera to record the screen of the ATM. It sounds good but it is not a great idea because to record the screen there may be two ways possible: Either we record the screen internally or place a camera outside/on the body of the ATM to record the screen. For the first one the attacker has to put a hardware inside the ATM so this idea fails and in the second way the camera might be visible to the user/victim as it will be on the body of the ATM to record the screen.*

3. *They assume that the victim takes sufficient precautions such as covering the typing hand with the other hand so they trained the model which works with hand movements but there may be a possibility that the victim may be covering his/her hand with something for example he/she may be wearing a glove then in that case their model will fail to give accurate results.*

4. *The entire attack is relied on the recorded videos that are used for testing the model. These videos are recorded via the hidden cameras . But it would be wrong to safely assume that the hidden cameras installation would be done easily. There are strong surveillance and security systems that monitor these kinds of suspicious activities for ATM'S.*

Technical approach:

1. *In the paper they are collecting two datasets by involving 58 participants of different ages and genders, but all the participants are right-handed in the experiment. So, the trained model might not predict the keys accurately for the left-handed persons since there is no one who is left-handed. Also, the number of participants are only 58, which is a smaller number for a training model. Hence this kind of dataset can introduce sampling bias, which might not give accurate results.*

2. *The paper does not provide a clear explanation why they selected 11 frames for each subsequence of the video. A clear explanation would have been appreciated if they could have mentioned the accuracy related reasoning for this selection .*

3. *They are using only two types of keypads for the experiment, by using only similar types of keypads the result will be good on the similar types of ATM pin pad but if they include more keypad models ,then the data would be more representative and generalized also the model can work significantly better on the other types of ATM.*

4. *Participants included in the data collection phase; they were all Caucasians. The model might phase difficulty working with different people of different races.*

5. *Except the fact that CNN works quite well with the image and video data, there is no clear provision mentioned (other than adding a dropout layer) for handling or preventing the major problem of overfitting the data in the technical approach in the paper.*

Analysis and/or results:

1. *They are getting the accuracy of 11.4% for the most difficult case which is pin pad independent scenario for the top-3 case, while for the top-1 it is around 6.67% which is significantly low.*

2. *For the training part of the participant just showing the 20 random videos of the users entering the pin with covered hand cannot be considered as the training, it should be longer and more detailed as humans take time to learn and predict using this knowledge as compared to the machine which take*

significantly less time. Also the machine learning methodology i.e. CNN which is used works best when the training data is massive. So there is a strong need to increase the size of the dataset along with keeping in mind the different biases.

Suggestions for improvement/Extensions:

1. *Data collection is done by only 58 participants only. Increasing the number of participants in the experiment will increase the trained model accuracy and we might get better results also.*
2. *People with different ages ,genders and races should contribute to the dataset in order to get a diverse and more informative and rich dataset.*
3. *Including the left-handed people in the experiment will also give us a better and representative dataset.*
4. *Training the model with different hand positions and conditions. For example, we can ask some participants to wear gloves before entering the pin, in that way our model will be able to infer the pins even if victim's hand is covered.*
5. *Increasing the Frames per subsequence will increase the trained model's capability of predicting keys accurately. Currently they are using 11 frames per subsequence , we can increase it to a higher number depending upon the frames per second rate of our camera and various other factors.*
6. *Increasing the number of pin-pad/keypads in the experiment will improve the model's performance on different types of ATM. In this way our model will be more powerful to achieve high accuracy rates.*
7. *For the Questionnaire, training of the participants should be long enough to train them such that they can guess/predict what is the key that is being pressed more accurately.*
8. *Including a thermal camera in the experiment might increase the probability of predicting the right keys. When a user finishes typing the pin, we can see the heatmap and can see what are all the keys he/she pressed, now combining this data with the video data can help to reduce the errors and will increase the probability of guessing the right key since our sample space will be reduced.*

****