A Deep Learning Approach For Detecting Click Ad Frauds in Mobile Advertising

Rutika Bangera , Deepti Bhoir , Hemangi Koli,Prachi Sorte

Abstract— Click fraud: intentionally clicking on the advertisements with no personal interest in the ad but to increase illegal revenue for the application publishers. This pay per click model has an empty space for the rival companies to post false ads to effect the healthy growing companies. Due to click fraud attack advertiser has result to loss as they have to pay to the publisher for the number of clicks on the advertisement. In this paper we proposed a click fraud detection model, to classify fraudulent click based on the features provided using CNN algorithm - a deep learning algorithm. After analysis of the CNN algorithm we get a graph chart which gives us an idea on whether the click was fraudulent or not. Our project helps in industries who want to either design a system wherein they prevent the fraudulent clicks before it attacks the end users .We have gain an accuracy of about 99.7% and result will be represented in the form of graph.

Index Terms— Click fraud,CNN(Convolutional Neural Network)

1 INTRODUCTION

In recent years, mobile advertising has became everyones first prionity as a mean for publishers to monetize their free applications.One of the main concerns in the in-app advertsing industry is the popular attack known as “click fraud" which is the act of clicking on an ad, not because of end user is interested on that ad, but just to get revenue from clicks for the application publisher.In fact, researchers predicted a growth of $17 billion by 2018 for mobile advertising

The mobile advertising industry also known as in-app advertising. consists of four main components: 1) The user that views the ad, 2) The advertiser that pays to have his/her ad shown in a set of applications, 3) The publisher (i.e. the application's developer) who is willing to display ads in his/her application in return for a certain revenue, and 4) The ad network that works as a relay between the advertiser and the publisher.

Number of charging models in industries are cost-per-thousand-impression model, the cost-per-click model, and the cost-per-action model. One of the popular type of attack in the in-app advertising industry is known as "click fraud". Click fraud is intentionally clicking on the advertisements with no personal interest in the ad but to increase illegal revenue for the application publishers. Due to click fraud attack advertiser has gain loss.

In this project we have adopted different features which undergoes training and testing using CNN algorithm.We represented the result in the form of graph for better understanding. the result in the form of graph for better understanding.

2 RELATED WORK

Frank VAnhoenshoven[3] addresses the detection of malicious URLs as a binary classification problem and studies the performance of several well-known classifiers such as Naıve Bayes, Support Vector Machines, Multi-Layer Perceptron, Decision Trees, Random Forest and k-Nearest Neighbors. In order to find falsified sites, the web security community has developed blacklisting services. These blacklists are inturn constructed using techniques including reporting, honeypots, and web crawlers combined with site analysis heuristics. The results of this paper suggest that the classification methods achieve competitive prediction accuracy rates for URL classification

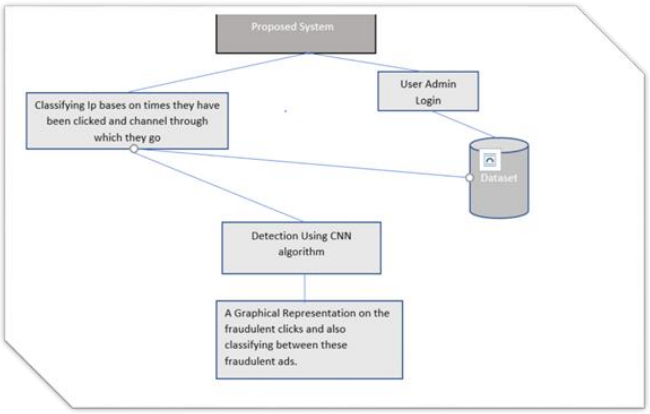
Linfeng Zhang [4] address the problem of detecting duplicate clicks in pay-per-click streams over jumping windows and sliding windows The first that propose two innovative algorithms that make only one Passover click streams and require significantly less memory space and operations.GBF algorithm is built on group Bloom filters which can process click streams over jumping windows with small number of sub-windows, while TBF algorithm is based on a new data structurecalled timing Bloom filter that identify click fraud over sliding windows and jumping windows with large number of sub-windows.

Mehmed Kantardzic[2] implemented multilevel data fusion mechanism used in CCFDP for real time click fraud detection and prevention. Prevention include blocking suspicious traffic by IP, referrer,

city, country, ISP, etc. The CCFDP system uses multi-level data fusion to enhance the description of each click, and to obtain better estimation of a click traffic quality. The CCFDP system analyzes the detailed user activities on both, server side and client side collaboratively to better evaluate the quality of clicks.

3 PROPOSED SYSTEM

Architecture:

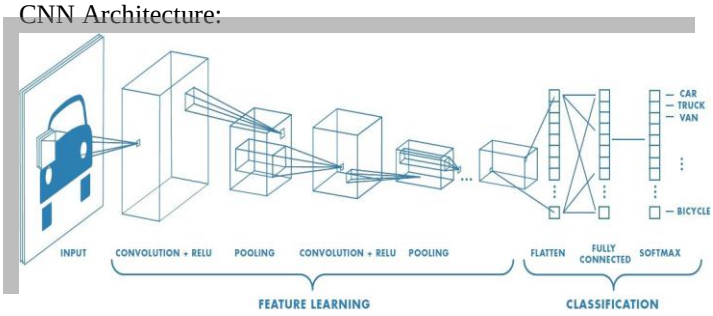


We have mainly focused on the click ad frauds that have happened at any time on any application or websites, we have analyzed this fraudulent clicks using a Deep Learning Algorithm which is convolutional neural network also known as CNN.After analyzing this clicks we have displayed the results in form of graph.

Dataset consist of following features:

1. IP address of the device that is clicked by the user.
2. App from user had click on the ad.
3. Number of times the user click from same device.
4. On which OS the mobile is based on.
5. Time at which user click on the ad

CNN Architecture:



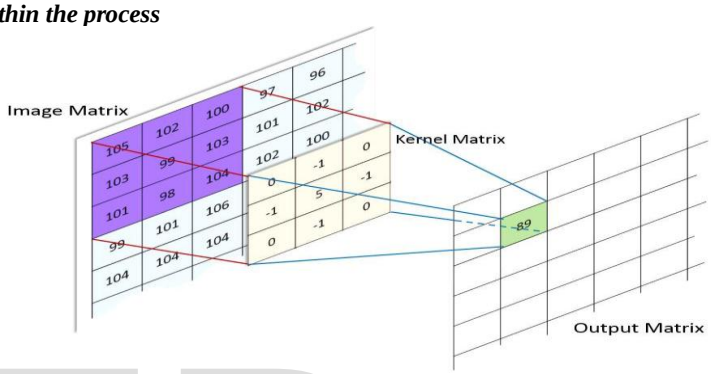
Convolutional Neural Networks (CNN) is one among the variants of neural networks used heavily within the field of Computer Vision. It derives its name from the sort of hidden layers it consists of. The hidden layers of a CNN typically contains convolutional layers, pooling layers, fully connected layers, and normalization layers. In CNN algorithm convolution and pooling functions are used as activation functions.

Convolution: Convolution operates on two signals (in 1D) or two images (in 2D):here we consider the "input" signal (or image), and therefore the other (called the kernel) as a "filter" on the input image, producing an output image (so convolution takes two images as input and produces a 3rd as output).

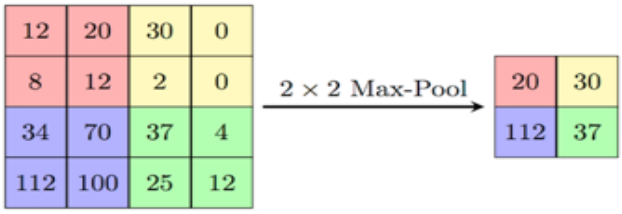
$$(f * g)(i) = \sum_{j=1}^{m} g(j) \cdot f(i - j + m/2)$$

which, is nothing but scalar product of the input function and a kernel function.

In case of Image processing, it's easier to see a kernel as sliding over a whole image and thus changing the worth of every pixel within the process



Pooling: Pooling is also known as sample-based discretization process.*There are 2 main sorts of pooling commonly referred to as max and min pooling.Max pooling means picking up the maximum value from the selected region and min pooling means picking up the minimum value from the selected region.*



4 EXPERIMENTAL,ANALYSIS AND RESULTS

We generated multiple dataset and combine to have 1 lakh instances. We perform preprocessing by deleting the missing key instances and also convert string formatted data into integer.

CNN classification:

We perform CNN classification using Python. As we are using one lakh instances,outof which 70000 for training of model and 30000 for testing.

1.Data exploration

Dataset of 1 lakh instances are gathered from multiple dataset which include ip address, app, device, operating system, channel, click time, attribute time.

	ip	app	device	os	channel	click_time	attributed_time	is_attributed
0	87540	12	1	13	497	2017-11-07 09:30:38	NaN	0
1	105560	25	1	17	259	2017-11-07 13:40:27	NaN	0
2	101424	12	1	19	212	2017-11-07 18:05:24	NaN	0
3	94584	13	1	13	477	2017-11-07 04:58:08	NaN	0
4	68413	12	1	1	178	2017-11-09 09:00:09	NaN	0

2.Preprocessing

Dataset is translated into useable information by removing the unwanted instances. Deleting the missing key instances and also convert string formatted data into integer so that the data type remains the same.

	ip	app	device	os	channel	click_time
0	81834	3	1	1	280	26311
1	88281	27	2	17	153	75333
2	78805	8	1	13	140	10368
3	35520	2	1	20	469	17755
4	8836	3	1	13	137	5730

3.Deep neural network

In order to find best accuracy of the model the dataset will undergo five hidden layer and model used is sequential. The accuracy at each layer is given as follow:

Training of model on 70000 instances:

Layer 1	Loss:0.1375	Accuracy:0.9977
Layer 2	Loss:0.0166	Accuracy:0.9977
Layer 3	Loss:0.0161	Accuracy: 0.9977
Layer 4	Loss:0.0161	Accuracy: 0.9977
Layer 5	Loss:0.1103	Accuracy:0.9975

RESULT: accuracy: 0.9977 loss: 0.0094

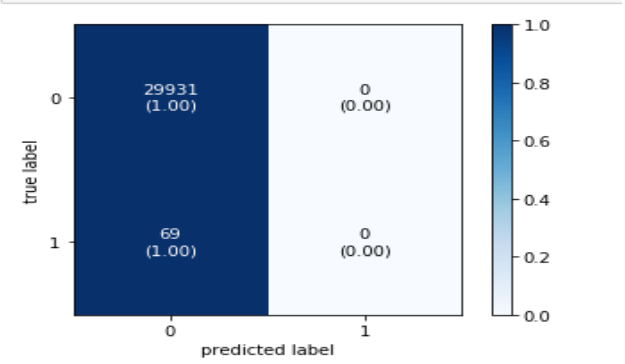
Confusion Matrix:

		CNN output	
		1	0
Actual	1	TP=29937	FN=0
	0	FP=63	TN=0

Out of 29937 actual instances of not fraudulent (first row), the classifier predict correctly 29937 and out of 63 actual instance of fraudlent ,the classifier predicted correctly 63 of them.

Ploting of confusion matrix:

In this confusion matrix plotting,the rows represents to the predicted class and the columns reprents to the true class .



4 Conclusion

In this paper ,we have addresses the problem of detecting Click Ad Fraud in Mobile Advertisement. We have detected and analyzed the click fraud happening through clicks using CNN algorithm. CNN algorithm is considered to be the best as it gives an accuracy of about 99.7%.

We have also found out the count of fraudulent and non-fraudulent ads among the total number of clicks. Out of 29937 clicks ,63 were found out to be fraudulent. The Graphical representation provides easy understand and helps to take decision .

References

1) Riwa Mouawi, Mariette Awad, Ali Chehab,Imad H.El Hajj and Ayman Kayssi, “Towards a machine learning approach for Detecting Click Fraud in Mobile Advertizing”,IEEE Access,November,2018

2) Mehmed Kantardzic , Chamila Walgampaya , Wael Emara, “Click Fraud Prevention in Pay-Per-Click Model: Learning through Multi-model Evidence Fusion”,October,2010

3) Frank Vanhoenshoven, Gonzalo Napoles , Rafael Falcon, Koen Vanhoof and Mario Koppen, “Detecting Malicious URLs using

Machine Learning Techniques",IEEE Access, December,2016

4) Linfeng Zhang,Yong Guan, "Detecting Click Fraud in Pay-Per-Click Streams of Online Advertising Networks",June,2008

5) Haitao Xu, Daiping Liu, Aaron Koehl, Haining Wang, Angelos Stavrou, "Click Fraud Detection on the Advertiser Side",IEEE Access 2014