DeepImageSpam: Deep Learning based Image Spam Detection

Amara Dinesh Kumar

Department of Electronics and Communication Engineering, Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham, India

Vinayakumar R, Soman KP

Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham, India

Abstract

Hackers and spammers are employing innovative and novel techniques to deceive novice and even knowledgeable internet users. Image spam is one of such technique where the spammer varies and changes some portion of the image such that it is indistinguishable from the original image fooling the users. This paper proposes a deep learning based approach for image spam detection using the convolutional neural networks which uses a dataset with 810 natural images and 928 spam images for classification achieving an accuracy of 91.7% outperforming the existing image processing and machine learning techniques.

Keywords: Image Spam, Deep Learning, Spam Detection, CNN

1. Introduction

The Internet has become a second self for most of the people today and many of our financial transactions, social interactions and communication are dependent on it and not always completely safe. Intruders, hackers and attackers are always in the quest for exploiting the users by hacking, spamming and impersonating. One of the major common and cost-effective and targeted attack in recent years is sending spam to users. spam detection and classification are very active fields of research and attackers are finding new ways to spam the users even with multiple layers of security mechanisms. They are finding the bugs and vulnerabilities and exploiting them actively improving on day to day basis. It is very common that users receive emails impersonating banks and other authorities asking for other personal details. Spam messages are causing huge loss to organizations. Many resources are getting wasted like mail server space, network bandwidth, spam filtering, mail server processing.

Attackers may redirect users to counterfeit products sites, impersonate an authenticated site and stealing sensitive information, spread fake news and providing wrong information, false marketing. spam messages are wasting the time and effort of the users and distracting them thus decreasing the productivity flooding email and forcing to delete and clean it frequently.

Image spam is altering the Image spam is increasing in the recent years with tremendous growth. A major reason for that is many of the email clients are filtering the spam text emails the subject sender and email content.but when it comes to the image detecting the spam is not easy particularly when attackers are embedding text into the images. Generally, email spam is detected using the spam filters which are evolved state now and can detect most of the spam with high accuracy but when it comes to image spam detection it is still in nascent stage and active research is going on for detecting with high accuracy. Initially Image spam in the form of HTML and to counter it researchers started using the OCR techniques. Attackers then used captcha based techniques obfuscating the text in the images but still readable by the humans but difficult to identify for an algorithm. This problem motivated researchers to use image processing techniques for image spam identification.

Email address: `dineshkumar.amara@gmail.com` (Amara Dinesh Kumar)

1.1. what are spam filters

Spam filters either block the spam messages or send them to a spam folder. The spam filter analyses the sender, subject, metadata and other related information of the mail and them classify it as spam or legitimate. They can be individual spam filters maintaining a block list containing addresses where you can add or delete from the respective list. In a community-based spam filtering, all can collaborate and add entries to the block list for improving the performance and ease of use. Initially set of conditions are programmed as rules for spam filters and later machine learning is being employed in latest spam filter which is proved more efficient and accurate.

2. Related Work

2.1. Image processing techniques

2.1.1. optical character recognition

In OCR the text is extracted from the image using the image processing techniques like edge detection and then obtained text is passed to a conventional text spam filter which detects and classify the spam image[1].

spammers have applied various image processing techniques like changing foreground and background also changing text font,size and colour made the OCR based method obsolete[2].

2.1.2. Colour Histograms

The colour histograms of normal images are mostly continuous and spam images contain isolated peaks in the histogram using which we can detect the spam image.But the accuracy is low and it is not a reliable method[3]. Histogram of gradients(HOG) is one more approach used commonly which works similar to colour histograms using the edge detection of the image and then plotting intensity plot can identify the spam image[4].

2.2. Using machine learning techniques

For performing classification using the machine learning algorithms the features are to be selected and extracted manually.Their are two types of features in the image spam detection and they are high level features and low level features.

steps for the attackers for sending the image spam are first developing the template of the image and then obfuscating it randomly and send to to different users which made identification of the image spam difficult and reason for lower detection accuracy.Features are further classified into Low level features are High level features[5]. Features like sender,meta data,message header are extracted and training dataset is prepared and labeled.

2.2.1. Support Vector Machines (SVM)

Support vector machines(SVM's) are the most used machine learning algorithms for the image spam detection and because of high accuracy and robustness to misclassifications is the reason researcher prefer SVM[6]. SVM is a supervised learning algorithm used for the classification of data.It consists of support vectors which divide and classifies the data.It classifies the non linear data using kernel trick in which the non linear data is projected to higher dimensions to make it linearly separable by a plane which is generally referred as hyper plane.It does this using a kernel function and their are different types of kernel functions like linear,polynomial,radial basis function(RBF) and Gaussian kernels. Image spam detection is a binary classification problem and two classes are spam and not spam.Using the training data that is collected and labeled according to respective classes model is trained and then the model is tested by giving the test data and performance of model is evaluated.

The major drawback of this method is we have to manual extract the high level and low level features and feed to the classifier which is a time consuming task.

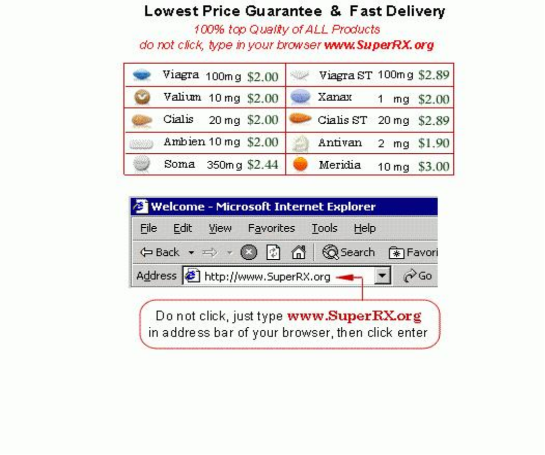
other machine learning classifiers like Logistic regression,Naive Bayes are also used for the image spam detection but SVM outperforms them by comparatively giving better performance.

2.3. using Deep learning techniques

Initially, neural networks are used for image spam detection and then now research has shifted focus on applying the deep learning algorithms. Deep learning consists of neural network layers which automatically extracts the features from the data in hierarchical pattern and then predicts and classifies the data.

3. Convolutional neural networks

Convolutional neural networks (CNN's) are one of the highly efficient deep learning algorithms used for classifying data (particularly image data) using supervised learning technique. They consist of an Input layer and convolution layer followed by pooling layer and again convolution and pooling layers alternatively based on the size and architecture of the network. The final layer is a fully connected layer. Fully connected layer converts the final scalar outputs of individual classes



Lowest Price Guarantee & Fast Delivery

100% top Quality of ALL Products

do not click, type in your browser www.SuperRX.org

Viagra	100mg	$2.00	Viagra ST	100mg	$2.89
Valium	10 mg	$2.00	Xanax	1 mg	$2.00
Cialis	20 mg	$2.00	Cialis ST	20 mg	$2.89
Ambien	10 mg	$2.00	Antivan	2 mg	$1.90
Soma	350mg	$2.44	Meridia	10 mg	$3.00

Do not click, just type www.SuperRX.org in address bar of your browser, then click enter

Figure 1: Sample spam image from dataset

Figure 2: Sample non-spam image from dataset

into their respective probabilities using a non-linear activation function and commonly softmax is used at the last laye[7]. CNN's are commonly used for the image processing applications as it can process the spatial information effectively capturing the pixel-related information using the convolution on to the image with strides[8].

4. Data set

The dataset used in the experiment consist of 928 spam images and 810 normal images which collected from different sources[9] and all are RGB images in various dimensions which in preprocessing are reshaped to 56×56 images.The sample images are shown in figure.1 and figure.2

Dataset is subdivided into both training and testing datasets. Training dataset consists of 742 spam images and 648 normal images.Testing dataset consists of 186 spam images and 162 normal images.

5. Experiments

The images are normalized and then given to the model for training.First convolution layer the kernel size used is 3×3 with input shape 32×56×56 with RELU (Rectified Linear Unit) activation function in the first convolution layer and then with max pooling layer of size 32×27×27 we are down sampling the data to half of the original dimension and subsequent layers follow the similar pattern.The brief description of the CNN layers architecture along with the output shape is described in table 1.drop out is 0.25 which means we randoms abandon some of the weights to avoid the over fitting and it acts as regularization.Batch size of 32 is used for training each epoch and model is trained for total 1000 epochs.

Layer (type)	**Output Shape**
conv2d_1 (Conv2D)	(None, 32, 56, 56)
activation_1 (Activation)	(None, 32, 56, 56)
conv2d_2 (Conv2D)	(None, 32, 54, 54)
activation_2 (Activation)	(None, 32, 54, 54)
max_pooling2d_1 (MaxPooling2)	(None, 32, 27, 27)
dropout_1 (Dropout)	(None, 32, 27, 27)
conv2d_3 (Conv2D)	(None, 64, 27, 27)
activation_3 (Activation)	(None, 64, 27, 27)
conv2d_4 (Conv2D)	(None, 64, 25, 25)
activation_4 (Activation)	(None, 64, 25, 25)
max_pooling2d_2 (MaxPooling2	(None, 64, 12, 12)
dropout_2 (Dropout)	(None, 64, 12, 12)
flatten_1 (Flatten)	(None, 9216)
dense_1 (Dense)	(None, 128)
activation_5 (Activation)	(None, 128)
dropout_3 (Dropout)	(None, 128)
dense_2 (Dense)	(None, 1)
activation_6 (Activation)	(None, 1)

Table 1: CNN architecture description

6. Results

Total images are split in ratio of 80 percent training data and 20 percent testing data.The model is evaluated after the convolutional neural network is trained on training dataset.It is then tested on the testing data set and result metrics accuracy,precision,recall and f1score are mentioned in the table 2.binary entropy loss and adam optimizer were used for training and checkpoints were saved periodically in the hdf file. The training and testing are performed on a distributed computing cluster platform with i7 cpu processor and 8 GB RAM system configuration.Keras,Sklearn and Tensorflow deeplearning libraries are used for training and testing.

metrics	**percentage**
accuracy	0.917
recall	0.857
precision	1.000
f1score	0.923

Table 2: Results metrics evaluated on test data set

7. Conclusion

In this research we have used the convolutional neural network(CNN) which is a deep learning network architecture for image spam detection.The deep learning approach gives better accuracy when compared with the machine learning and other conventional image processing based methods and also avoids the manual feature extraction task by automatically identifying the features by itself reducing the time and effort.Binary classification of image is performed the model is trained with existing labelled data set and then tested with the test data then metrics are evaluated.Further research can be carried out by exploring other deep learning algorithms like RNN and LSTM and tuning the architecture and hyper parameters may provide interesting insights.Capsule networks can also be tested on the data set which are giving promising results recently when compared with the convolutional neural networks for image related techniques[10].

References

[1] S. Dhanaraj, V. Karthikeyani, A study on e-mail image spam filtering techniques, in: 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, 2013, pp. 49–55. doi:10.1109/ICPRIME.2013.6496446.

[2] B. Biggio, G. Fumera, I. Pillai, F. Roli, A survey and experimental evaluation of image spam filtering techniques, Pattern Recognition Letters 32 (10) (2011) 1436–1446.

[3] B. Biggio, G. Fumera, I. Pillai, F. Roli, A survey and experimental evaluation of image spam filtering techniques, Pattern Recognition Letters 32 (10) (2011) 1436 – 1446. doi:https://doi.org/10.1016/j.patrec.2011.03.022.

[4] L. M. Ketari, M. Chandra, M. A. Khanum, A study of image spam filtering techniques, in: 2012 Fourth International Conference on Computational Intelligence and Communication Networks, 2012, pp. 245–250. doi:10.1109/CICN.2012.34.

[5] B. Mehta, S. Nangia, M. Gupta, W. Nejdl, Detecting image spam using visual features and near duplicate detection, in: Proceedings of the 17th International Conference on World Wide Web, WWW '08, ACM, New York, NY, USA, 2008, pp. 497–506. doi:10.1145/1367497.1367565.
URL http://doi.acm.org/10.1145/1367497.1367565

[6] S. Krasser, Y. Tang, J. Gould, D. Alperovitch, P. Judge, Identifying image spam based on header and file properties using c4. 5 decision trees and support vector machine learning, in: Information Assurance and Security Workshop, 2007. IAW'07. IEEE SMC, IEEE, 2007, pp. 255–261.

[7] R. Vinayakumar, K. P. Soman, P. Poornachandran, Applying convolutional neural network for network intrusion detection, in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 1222–1228. doi:10.1109/ICACCI.2017.8126009.

[8] R. Vinayakumar, P. Poornachandran, K. P. Soman, Scalable Framework for Cyber Threat Situational Awareness Based on Domain Name Systems Data Analysis, Springer Singapore, Singapore, 2018, pp. 113–142.

[9] Y. Gao, M. Yang, X. Zhao, B. Pardo, Y. Wu, T. N. Pappas, A. Choudhary, Image spam hunter, in: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE, 2008, pp. 1765–1768.

[10] A. D. Kumar, Novel deep learning model for traffic sign detection using capsule networks, arXiv preprint arXiv:1805.04424.