Multimodal Spam Classification Using Deep Learning Techniques

Shikhar Seth
Department of Computing and Information Technology
Manipal University Jaipur
Jaipur, India
shikharseth15@gmail.com

Sagar Biswas
Department of Computing and Information Technology
Manipal University Jaipur
Jaipur, India
biswas.sagar97@gmail.com

***Abstract*— The internet has been beneficial to the society in ways more than one, the power to learn anything anywhere, the power to always be connected to the people you love. But as usual, there are two sides to coin. The E-mail system has been the backbone for communication between professionals for a very long time, but it is plagued by the unwanted influence of spam. In this paper we classify a mail into spam or not-spam (ham) by analyzing the whole content i.e. Image and Text,processing it through independent classifiers using Convolutional Neural Networks. We finally propose two hybrid multi-modal architectures by forging the image and text classifiers. Our experimental results outperform the current state-of-the-art methods and provide a new baseline for future research in the field.**

Keywords- E-mail, Spam Classification, Deep Learning, Convolutional Neural Networks, Multimodal

I. INTRODUCTION

The E-mail system has been one of the most wide spread forms of communication. Everyday millions of emails are sent to and fro for business and personal purposes. It was only a matter of time that such a huge passage of communication became the bridge for malicious activities. Emails that contain unsolicited messages are called spam emails. Although generally spam emails are commercial in nature, carrying advertisements, they may also contain disguised links leading to phishing websites, the emails might also contain malicious scripts and executable files. Seeing this problem the need of spam filtration arises. Spam filtration is the task of differentiating the mails into spam and non-spam (ham). Over the years spammers have found new ways to avoid the filtration systems brought into use by the email service providers. There are various filtering systems in use to find out the intent of the mail, which take the aid of machine learning and pattern recognition. To trick these filters, spam mails have evolved from simple blank mails, to bogus text emails to using texts embedded in images, So generic machine learning algorithms render worthless against these kind of email spam. To tackle this kind of a problem we need to analyze the whole email, the text or any other images anywhere on the body of the mail. We propose to do this by taking both aspects as inputs and use convolutional neural networks in a mutli-modal architecture for our classification prediction. To the best of our knowledge, this approach has yet not been taken in spam recognition. Due to to the security threats these mails pose like malware, spyware and clickbait, blocking spam emails are one of the top priorities of a network administrator. Lots of research has gone into developing spam filters for the coming of age spam emails, and have been fairly successful before spammers find out a new way to send spam mails on a large scale level. An example of a modern day spam is as follows:

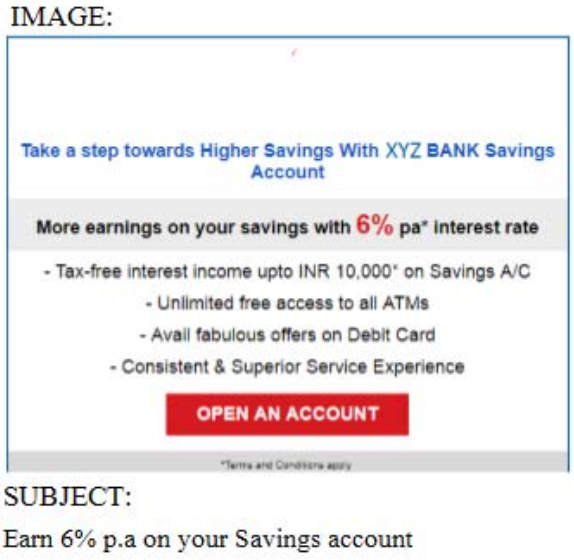
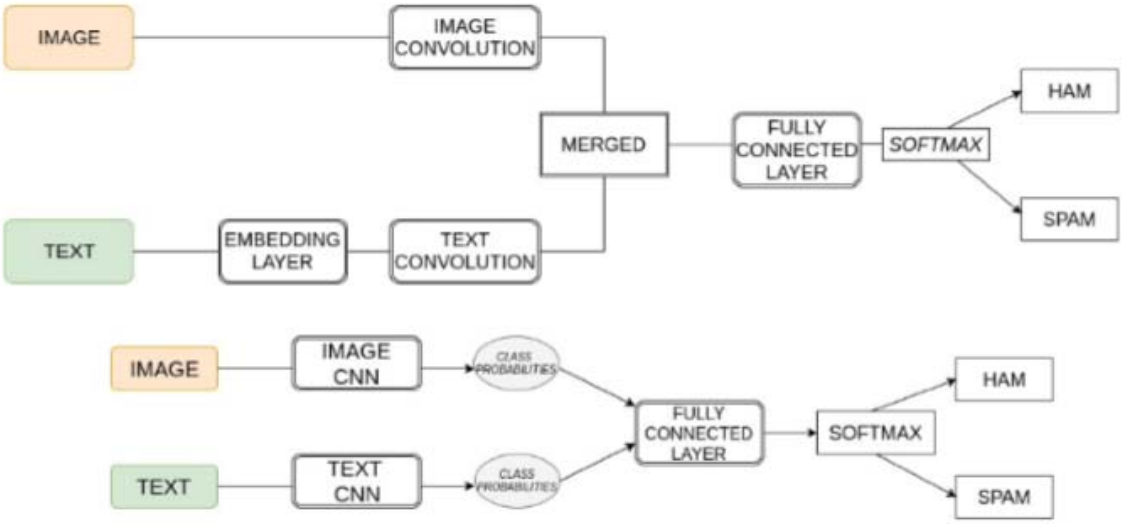


Figure: 1
(a) Modern Day Spam

Until recent, Spam filters used traditional machine learning algorithms to classify a text mail as spam or ham. Algorithms such as Naive Bayes, SVM, Decision Trees, k-NN and Neural Networks have been to put to use in spam filters [1][2][3] which have proven to be very resourceful against traditional textual spam emails. However with the advent of image spam in recent times, classifiers based on traditional Machine learning algorithms fail. Consequently image spam has been an active area of research, with OCR, neural networks and near image duplication being the different classification techniques [4][5][6]. Neural networks are a part of Deep learning which is a subset of Machine Learning that has received a lot of attention in the current decade for its overwhelming results on various tasks such as classification, image recognition, text classification and sentiment analysis.

In the approach of this paper i.e. analyzing the whole content of the mail, we use convolutional neural networks (CNNs) for classification of each input. CNNs were inspired by biological processes, more specifically on how the neurons function, CNNs use a variation of multilayer

978-1-5386-4283-2/17 $31.00 © 2017 IEEE
DOI 10.1109/SITIS.2017.91



Top: The multimodal architecture by merging the last convolutional layer features.
Below: The multimodal architecture that takes in class probablities of the independent classifiers as input.

Figure: 2

Perceptron designed to require minimal preprocessing, and they have been largely successful in the field of analyzing visual imagery.
After collecting the inputs from two different modals, we combine the information to finally classify the email into spam or ham.

II. MODEL ARCHITECTURES

In recent, Convolutional Neural Networks (CNNs) have proven very effective in fundamental Computer Vision tasks such as image recognition and classification. Following Alex Net's win in the ImageNet competition, CNNs have now become the standard go to classifier for Image classification. More recently, CNNs have been applied to a lot of NLP tasks such as document classification and sentiment analysis. Although recent findings equate RNNs and CNNs almost equally in task such as text classification [7], we find that CNN performs better for spam content classification.

A. IMAGE CLASSIFIER

The input to our Image classifier is a fixed size 128 * 128 RGB image. Since our objective is binary classification, we build a relatively small CNN compared to present state of the art models. We stack up 4 convolutional layers containing 32, 32, 64 and 128 filters with kernel size 3*3. Convolution stride and spatial padding is fixed to 1 pixel; every layer undergoes MaxPooling over a 2*2 window, with stride 1. The convolutional layers are followed by 3 Fully Connected layers containing 128 and 64 nodes and Softmax function at the end. All hidden layers are followed by ReLU non-linearity function. We use Dropout, a regularization technique that reduces overfitting by preventing complex co-adaptations on training data [8]. Dropout rate used is 0.5 after every layer.

B. TEXT CLASSIFIER

For classification of text input, we use a very similar architecture to the CNN architecture proposed by Kim (2004). The first layer embeds words into 100 dimensional vectors using the GloVE word embeddings. The next layer performs convolutions over time on the embedded word vectors using multiple filter sizes (2, 3, 4, 5), where we use 128 filters from each size. Next, we max-pool-over-time the result of each convolution filter and concatenate all the results together. The concatenated layer then undergoes convolution using 128 fixed size (5) filters. The next layer performs max-pooling over the result. This is followed by a Fully Connected Layer containing 128 hidden units. Softmax function is applied at the end in order to assign class probabilities.

C. MULTI-MODAL CLASSIFIER

To classify e-mails containing both image and text, we propose 2 multi-modal architectures which combine the 2 CNN classifiers (image and text) and produce an output class (as shown in figure 2).

1. Creating a shared representation of the features extracted by the independent Convolutional Neural Networks and then passing it through Fully Connected layers. The features of the last layer of the networks are merged together in various styles (namely concatenation and multiplication). The resultant tensor is passed through 2 Fully Connected Layers containing Softmax at the end. The model learns the shared representation by training the weights on the hidden layer neurons.
2. Using the class probabilities of the 2 classifiers and learning a rule between them. The image and text classifiers output class probabilities which then serve as input to a 3 layered Fully Connected network with Softmax at the end. This model is also applicable for instances when only one input signal (either text or image) is present.

III. EXPERIMENT

A. DATASETS

To the best of our knowledge, there exists no large scale dataset for e-mail classification that contains both image and text data. Even for the spam image data only, most of the datasets are relatively small and quite outdated. This motivated us to build an image dataset from scratch containing both ham and spam images. We have collected over 1521 spam images from various spam mails received by us in the past year. For ham images, we have downloaded 1500 images from different sources Flickr and Facebook and document scans. For text we use the pre-processed Enron Spam Dataset. It contains 33,645 texts out of which 16,537 are ham and the rest 17,108 are spam. The texts contain only the subject and content of the e-mails.

B. TRAINING THE MODELS

Both the datasets have been divided into training (80%) and validation (20%) sets. To estimate optimal model parameters for the classifiers, we run 7 fold cross validation on the training set. After the estimation of all parameters using grid search we train the classifiers on entire training sets and then test it upon the validation set.

TEXT CNN

The embedding layer converts the unseen words to all zeros i.e. words that don't appear in the GloVE embeddings will be represented by all zero tensors. Maximum words in a document (subject and content) have been kept to 100 and padding is applied in order to keep fixed size input. Different word embeddings (such as word2vec) were also used.

IMAGE CNN

All the images have been resized into 128*128 pixels.

MULTIMODAL

For our multimodal training we created a custom function that maps 3021 images with equal number of text samples with their correct label i.e. our training dataset contains spam/ham images paired with spam/ham texts along with their respective labels. We try both the approaches i.e. 1) training the models end to end and 2) using the weights from trained classifiers to only modify the additional weighs in the multimodal. We find that using the pretrained model weights saves a lot of time and the accuracy achieved by the multimodal is also greater compared to the former approach.

.

IV. RESULTS

Results are presented in table 1.

	ACCURACY	f1-SCORE (AVG)
Image CNN	85.89%	0.87
Text CNN	97.54%	0.97
Multi-modal(Feature Fusion)	96.87%	0.95
Multi modal(Learned Rule)	98.11%	0.98

Table 1.

For classification of documents, metric f1-score is generally used along with accuracy. F1 score is the harmonic mean of precision and recall.

$$F1=2(\text{Precision}*\text{Recall})/ (\text{Precision}+\text{Recall}) \quad (1)$$

It is a quite reliable metric as it also takes into account the total misclassification of e-mails which can pose a terrible problem if not tackled.

Our purely text classifier outperforms the image classifier by quite a margin. Despite little tuning of hyper parameters, the Text CNN model achieves excellent results. The pretrained word embeddings enjoy great success in the classification tasks, even without fine tuning. The image CNN also performs quite well, given the small size of the network and dataset used. We observe that both of our multimodal architectures give quite satisfactory results with the self-learning rule model Having an upper hand in both the accuracy as well as f1-score. With an accuracy of 98.11%, it exceeds the accuracy of the independent classifiers and achieves a very high f1-score.

V. CONCLUSION

Spam continues to be one of the major issues plaguing the mail communication. Out of all the types of spam being used today, Image combined with textual spam is becoming one the most prominent kinds of spam, and they will use advanced content obscuring techniques as time progress, this leads us to believe that computer vision and deep learning itself will have a large part to play in spam classification.

In this paper we have investigated the efficiency of the mutli modal approach using deep learning in email spam identification. At the end of the experiment it was noted that our multi modal approach provided better accuracy at email spam identification than separate text and image classifiers.

This approach is just scratching the surface of multi-modality. Further work, will include refining our neural network for better accuracy, hyper parameter optimization, taking multiple images and using them to

classify the intent of the email. We have faith, that if pushed forward properly, the multi-modal approach can become the new standard of classification.

REFERENCES

[1] Ismaila Idris, "E-mail Spam Classsification with Artificial Neural Network and Negative Selection Algorithm", International Journal of Computer Science & Communication Networks, Vol 1(3), 227-231

[2] Sarit Chakraborty and Bikromadittya Mondal," Spam Mail Filtering Technique using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis", International Journal of Computer Applications (0975 – 888) Volume 47– No.16, June 2012

[3] Ola Amayari and Nizar Bouguila "A study of spam filtering using support vector machines", Artificial Intelligence Review, Volume 34,Issue 1,pp 73108, June 2010

[4] Battista Biggio, Giorgio Fumera, Ignazio Pillai, Fabio Roli, "Improving Image Spam Filtering Using Image Text Features"

[5] M. Soranamageswari and C. Meena ,"An Efficient Feature Extraction Method for Classification of Image Spam Using Artificial Neural Networks ",IEEE 10.1109/DSDE.2010.60, April 2010

[6] Zhe Wang, William Josephson, Qin Lv, Moses Charikar, Kai Li, "Filtering Image Spam with Near-Duplicate Detection"

[7] Wenpeng Yin† , Katharina Kann† , Mo Yu‡ and Hinrich Schutze†," Comparative Study of CNN and RNN for Natural Language Processing", arXiv:1702.01923v1 [cs.CL],Feb 2017

[8] Nitish Srivastava,Geoffrey Hinton,Alex Krizhevsky, Ilya Sutskever ,Ruslan Salakhutdinov," Dropout: A Simple Way to Prevent Neural Networks from Overfitting",Journal of Machine Learning Research 15 (2014) 1929-1958