JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 04 Issue 02
2023

Research Article
https://doi.org/10.56979/402/2023

DGA Malware Deep Learning Detection and its Optimization with Novel Activation Function

Muhammed Awais Javed[1*], Imran Rashid[1], and Adnan Rashdi[1]

[1]Department of Information Security, National University of Science and Technology, Islamabad, Pakistan.
[*]Corresponding Author: Muhammed Awais Javed. Email: awaiswill@gmail.com.

Received: December 10, 2022 **Accepted:** January 13, 2022 **Published:** March 29, 2023.

Abstract: APTs mutually coupled with Cyber Kill Chains (CKC) and its specified phase of malicious command and control (C2) servers. These C2 servers maintain communication using malicious domains with a specially crafted malware called Domain Generating Algorithm (DGA). The DGA malware is available in different compositions and complexities associated with various APTs as well as DGA families. DGA detection is achieved using different Machine Learning (ML) models and recently DGA detection is further improved with Deep Learning (DL) models. These trained DL models have solved DGA detection using text classification, successfully classifying legitimate domains from malicious domains. DL models' optimal detection is further optimized by tuning DL key functions, one such key function is the Activation Function (AF). Primarily AF provides the property of non-linearity which is very effective in mapping and solving real-world problems. Recently reported AFs in literature are based on their superior performance in text classification are identified and analysed in these optimal DL models. Due to Long Short Term Memory (LSTM) and Attention models successful detection in text classification, LSTM with Attention is implemented for deeper analysis of these reported AFs. In this research paper, the DGA detection DL models have been simulated with the default AFs and performance of proposed AF has been tested against default AFs. The proposed AF Zash outperformed the ReLU, Hyper-Tangent (Tanh) and Swish AFs in terms of their polynomial properties. Sparse activations being core property of ReLU may miss some of significant weight updates in comparison to dense activations of exponential fixed shaped Tanh and Swish AFs. Results have shown that the proposed Zash AF have overcome the sparse activations of ReLU and has achieved proficient results in dense activations over Tanh and Swish AFs. This novel AF has shown better detection results in training and validation for text based character classification using dense activations.

Keywords: DGA Detection; Deep learning; LSTM; Activation Functions; Zash Activation Function.

1. Introduction

Conventionally most of the security approaches focus on hardening the perimeter of the networks against outsider attacks, known as Castle Approach [1]. These castle approaches are exploited using advanced and customized attacks like Advanced Persistent Threats (APTs). Today the most advanced and hybrid cyber threat is manifested as APTs. APTs are not an aggressive attack but rather a progressive clandestine cyber operation. Examples of strategic level APTs [2] include felony of espionage/ sensitive information, trade surveillance, cyber/ digital thefts, cyber frauds, ransoms, and extortion. APT phases of cyber-attacks are chained and conceptualized as Cyber Kill Chain (CKC) [3]. MITRE's Adversarial Techniques and Tactics and Common Knowledge (ATTCK) [4] is an advanced malware knowledge book very effective in associating cyber-attack with the known APT groups. CKC phase of Command and Control (CC) revealed that the bots exploit legitimate communication protocols like DNS to evade detection.

The malicious domains are generated sporadically from infected systems by a specifically crafted malware called Domain Generating Algorithm (DGA). These DGA connect to CC servers preset by malicious actors. In this research, APT detection is coupled with detection of DGA domains after analyzing CKC and ATTCK and conceiving a simplified APT detection cycle as depicted Figure 1. The idea is how a multi-pronged complex cyber-attack like APT can be detected at a Single Point of Detection (SPOD). DGA detection model grabs malicious domains generation and can be further associated with either a known APT or a new APT for further probes. The detection of malicious domains involves analysis of high volumes of DNS log data (URL names), which is logically suitable for ML based text classification.

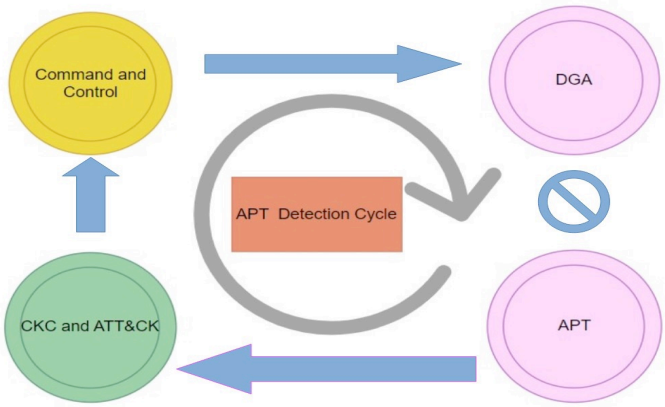

Figure 1. Proposed APT Detection Cycle

Initially ML models were adopted for DGA detection, however recently 46 the DGA detection is switched from ML to DL models such as Long-short term memory (LSTM) and convolution neural networks (CNN) models with considerable improved performance over traditional ML models. Further, hybrid approach of both LSTM and CNN were also applied and shown significant results over previous models. More recently, the introduction of Attention models in DGA detection has further improved the in-formation loss in case of much longer URL (text) strings. The LSTM with Attention model has shown optimum performance for text classification-based problems. To fine tune LSTM with Attention model in training and testing, various model functions, parameters and hyper parameters are available. However, to avoid a scattered approach, a key function namely, Activation Function (AF) is selected to narrow the in-tended research aspiration. In DL model, layers are made of neurons and each neuron operation and out-put is controlled by the AFs. Scope of this research has been focused on the best performing AF in a LSTM with Attention model.

A recent study on performance of AFs in natural language processing (NLP) based tasks has been adopted for best performing AFs in LSTM networks. However, the study is limited to available community datasets from Alexa [5] and Bamabanek [6]. The motivation of this work is to implement the best performing AFs in real time problems of text classification specifically like DGA detection. Moreover, this work has introduced an uncharted fixed shaped exponential AF named as Zash for LSTM networks. This AF is being developed with an inspiration from base functionalities and properties of the ReLU, Tanh and Swish AFs. The core properties used to compare these AFs include non-linearity, sparse vs dense activation, monotonic vs non-monotonic, expressiveness and smoothness. Section-2 explains DGA detection taxonomy and Sec-tion-3 is about basic operations and properties of AFs in a neural network (NN). Section-4 talks about pro-posed methodology of optimization approaches with various AFs. Same has been simulated to substantiate the proposed methodology. Section-5 discusses graphical and tabular results and contemplates the core properties of AFs. Section-6 is culminated with future research directions and conclusive remarks.

2. Domain Generating Algorithm (DGA) Taxonomy

DGA malware structure is divided broadly into two types, seeds source based and generation-based schemes respectively [7]. The DGA malware generates bulk DNS requests data making it a potential candidate for ML which has already outperformed in text classification problems. DGA detection techniques

can be broadly divided into early detection techniques which later evolved to ML and DL techniques respectively. The taxonomy of these broad contours of DGA detection are projected in Table-1. Brief explanaion of these contours is given in ensuing subsections.

Table 1. Taxonomy of DGA Detection Techniques

DGA Detection Methods	DGA Detection Techniques	Years	References
Traditional Methods	Blacklisting and Whitelisting	2014	Mark Kuhrer
		2014	Huang D
	Sequential and Hypothesis testing	2013	Srinivas Krishnan
	Log Aggregation Techniques	2013	Srinivas Krishnan
	Domain Reputation System	2010	M. Antonakakis
Machine Learning Methods	Lexical Analysis	2018	E. Kidmose
	Word Graph	2018	Mayana Pereira
	Clustering Techniques	2012	S. Yadav
		2012	S. Yadav
	N-gram methods	2014	S. Schiavoni
		2019	H. Zhao
	Random Forest	2018	D. X. Cho
		2015	E. Kidmose
	HMM	2012	Manos Antonakaki
Deep Learning Methods	LSTM	2016	J. Woodbridge
		2019	Akash
		2018	Duc Tran
	CNN	2018	R Vinayakumar
		2017	Joshua
		2018	W. Bush
		2019	Shaofang Zhou
		2018	B. Yu, J. Pan
	Hybrid	2020	K. Highnam
		2019	Y. Qiao
		2021	J. Namgung

2.1. Traditional DGA Detection Methods2018

Early DGA detection techniques in Table-1 are white listing and blacklisting of domains [8] , [9][10]. Later, whitelisting and blacklisting are further aggregated with other detection techniques like DNS Reputation Systems (DRS) [11], log aggregation techniques [12] lexical analysis [13], word graph methods [12].

2.2. Machine Learning DGA Detection Methods

ML DGA Detection in Table-1 is switching of DGA detection to ML models like clustering techniques [14], n-gram methods [15][16], Random Forrest (RF) classifiers [17] and Hidden Markov models (HMM) [18] which have shown remarkable improvement in detection performance over traditional detection techniques. ML models learn to distinguish between benign and malicious domains using labelled data. ML needs samples of both legitimate domains and malicious domains datasets for training and learning. However, manual feature engineering in ML becomes superannuated due to new DGA malware evasion techniques applied by DGA malware authors. Moreover, manual feature engineering is not considered dynamic to rapidly evolving DGA tactics.

2.3. Deep Learning DGA Detection Methods

To address these impediments, DL has surpassed ML with the ability of automated features extraction. Utilizing the capability of automated feature engineering, LSTM based DGA detection was tested initially for the first time in 2016 [19]. The results in [19] have clearly shown that LSTM based DGA detection outperformed all traditional and ML Methods. To advance the work in DL methods, in [20] authors implemented CNN for detecting malicious URLs, file paths and registry keys. It is learnt from both approaches that as LSTM is good in detecting temporal relationships between texts (domain names), the CNN model detects spatial relationships in same texts. This pushed the researchers to adopt hybrid neural network approaches in [21][22][23][24],[25][26][27][28] of LSTM and CNN methods. Further, introduction of attention techniques has further elevated the performance of DGA detection models due to it inheriting longer dependencies of texts.

3. Selecting an LSTM Model with Attention

The research study aims to achieve optimal performance of LSTM models for achieving an efficient and accurate DGA detection approach. LSTM models are considered capable of showing impact of AF due it inherits back propagation property and continually updating model to a point of stability. After sufficient training of a DL model, the model is further validated and tested which dictates how well the model is generalized. Before we further go deeper in LSTM, an added layer of Attention model is attached which is considered a latest approach and is instrumental in longer dependencies.

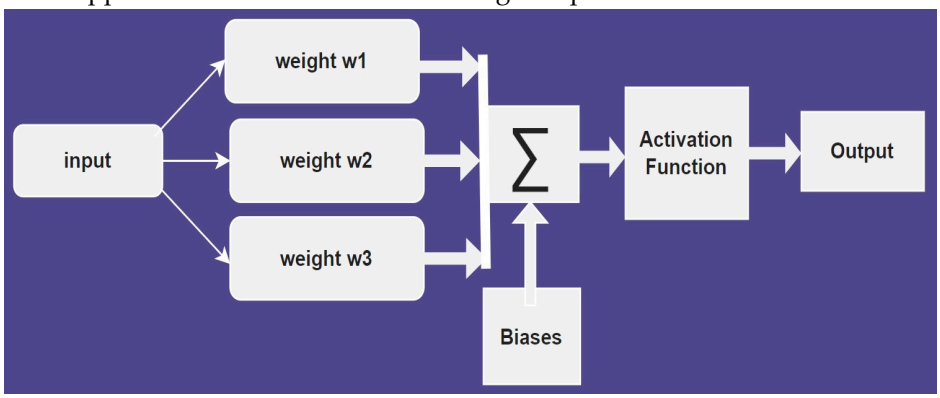

Figure 2. Neuron cell with input weights, bias, activation function and output

3.1. Activation Function (AF)

AF is the mathematical gateway computing output at each neuron. A proficient AF is considered a key parameter or hyper parameter which fastens the training and learning process by substantially decreasing the computational cost. Figure 2 is a graphical view of a single neuron cell of a neural network and its generic function in a LSTM model. The single neuron cell is summing up the input and its weights and adding biases in summation process. The sum of all input weights is operated by an AF (linear or non-linear function as desired) mathematically as, Input = max $\sum f(x)$= (weights (w1, w2, w3) inputs) + Bias (1), The inputs are multiplied by weights (w1, w2, w3...) and adjusted with the bias. Then these summed and biased weights are passed through the selected AF, normally a nonlinear AF. A DL model reaches a point of convergence after it is fully trained. Faster convergence means an efficient AF and stopping the Back-propagation (BP) process earlier, thus reducing the computational cost. Neural Networks work on gradients to improve the error by updating the new weights. In BP new weights are calculated using cost function and learning rate. Derivation of cost function and learning rate for calculating new weights is achieved by keep on iterating until model converge to a desired training level or simply an optimal convergence. New weights are calculated using back propagation as, $w_{new} = w_{old} - (dC\ dx) * LR$ (2), A non-linear AF find complex relationships in the given data and map it to a specified class by grouping them in using their class closest probabilities. The outputs of AFs are further smoothed by another function called model optimizer (SGD, Adam, RMSprop etc.). Exploring Legacy AF like Rectifier Linear Unit (ReLU), Hyper Tangent (Tanh) and Swish which are among the default AFs, by taking a deeper insight in identifying the most 139 desirable properties of AFs.

3.2. Non-Linear Legacy Activation Functions

Non-Linear AFs are better categorized as fixed shaped and trainable AFs as shown in Figure 3 simplified from [27, 28]. Non-Linear AFs are commonly the traditional AFs and are further subdivided in fixed

shape AF, rectifier-based AF and trainable AFs. Traditional mostly used AFs are Rectifier Linear Unit (ReLU) and Hyper Tangent (Tanh) apart from recently discovered Swish in deep learning models. Following subsections will explore further these traditional AFs with an introduction of a novel AF with emphasis on the most desired properties from any of AFs. Traditional AFs are fixed shaped functions broadly include identity, Linear, Sigmoid, Hyper-Tangent (Tanh), Swish and rectifier based.

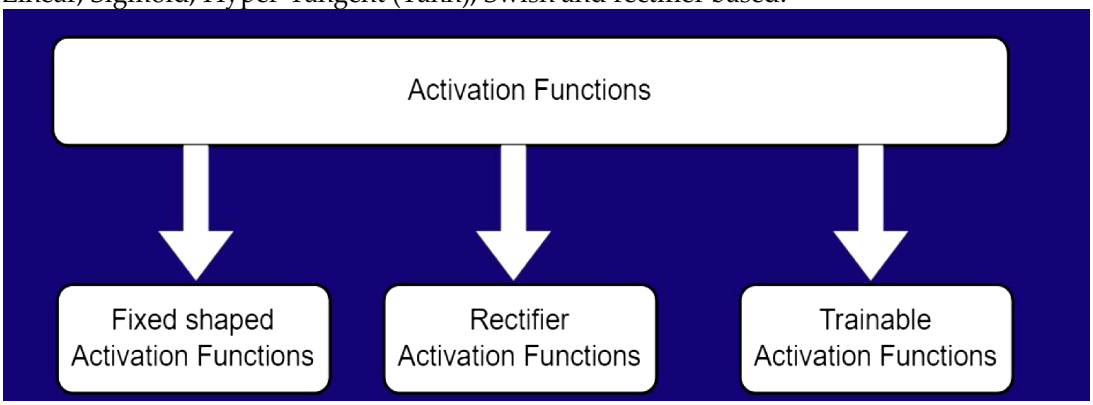

Figure 3. Activation Function operational categorization

These are like Rectifier Linear Unit (ReLU), Leaky ReLU, Parametric ReLU and Exponential Linear Unit (ELU) activation functions. Comparison of the fixed shaped AFs in LSTM models is the potential dimension of this research work, as fixed shape AFs work better with LSTM networks. Fixed shape AFs like Tanh, Swish and ReLU are considered exceptional to be implemented in DGA detection LSTM models.

3.2.1. Hyper Tangent (Tanh) Activation Function

Tanh is default AF of LSTM neural networks, Tanh(x) and its derivative are given as,

$$\tanh(x) = (e^{-x} - e^{-x})/(e^{-x} + e^{-x}) \quad (3)$$

$$f(x) = 1 - \tanh^2 x = \operatorname{sech}^2 x \quad (4)$$

Tanh is zero centered, a balanced AF which provide smoother nonlinearity. Its clipping on both sides contains the output from -1 to +1.

3.2.2. Swish Activation Function

Swish outperformed ReLU in machine translation and image recognition in [31]. Swish is smoother for generalization and non-monotonic property which improves gradient flow. Mathematically swish function is given in equation 6 and its derivative in equations 7 and 8 as,

$$f(x) = s * \text{sigmoid}(x) = x\sigma(x) = 1\ 1 + e{-}x \quad (5)$$

$$f(x) = 1\ 1 + e{-}x + x[\ 1\ 1 + e{-}x - 1\ 1 + e{-}x2\ ] \quad (6)$$

$$f(x) = \sigma(x) + x\sigma(x)(1 - \sigma(x)) \quad (7)$$

3.2.3. Rectifier Linear Unit (ReLU) Activation Function

ReLU is a slicing linear function between positive and negative inputs/ weights. ReLU simply gives output on only positive inputs. Negative input is discarded with no activation which nullifies the output to zero. It is fast as well as efficient in training the model as compared to other legacy AFs. It provides the basic property of non-linearity in a linear way. Mathematically

$$f(x) = \max(0, x) \quad (8)$$

The property of giving true zero for negative output works with piece-wise linearity. It has also overcome the problem of vanishing gradients in deep learning models.

3.3. Properties of Activation Functions

Most desirable property which is nonlinearity helps in mapping the given data feed/ inputs of a real-world problem. Other properties of AFs include computational efficiency, smoothness in learning process, Sparsity or dense activation and faster convergence. Non monotonic property of swish is also considered a pivotal in training, non-monotonic property gives more expressiveness. Despite strong foothold in DL models, fixed shape nonlinear AFs needs a deeper exploration of a better activation function in most widely adopted AFs like Tanh and recently the Swish AFs other than ReLU being an exceptional pseudo linear AF.

3.4. Activation Functions in text classification

As this research is solving DGA detection problem which is purely a text classification problem. Research conducted in Legacy AFs in [31] which implemented for various NLP tasks to assess the performance of these AFs with varying NLP problems. Overall results of the research in [31] for best performance AFs in NLP tasks are highlighted in Table-2. Referred work on legacy AF is relevant to text classification-based ML models. Solving the problem of DGA Detection on choosing the best AFs projected in Table-2 is ReLU, Tanh and Swish. These AFs are selected as base AFs for our DGA Detection. Legacy AFs are simulated and analyzed along with introduction of a new AF Zash in next classification problem of DGA Detection.

4. Proposed Novel Zash Activation Function

Authors In light of identified properties of nonlinear fixed shape AFs, a new AF which has been discovered. This AF is called Zash and can show properties better than both Tanh and Swish AF. Moreover, it has shown that in case of dense and sparse activation, this AF will drop significant weight updates by achieving ReLU sparse activations property and have shown better performance in nonlinear legacy fixed shaped AFs like Tanh and swish AFs. Zash AF (ZAF) approach is slight modification of Sigmoid AF as same is applied in swish however we have deviation from swish method. ZAF is variant of Sigmoid AF and use an input x in both nominator and denominator while swish uses the input only in nominator,

$$f(x) = x\,x + e{-}x \tag{9}$$

ZAF achieves non-linearity and being zero-centered properties like swish and Tanh. This function shows greater accumulation on both sides for larger negative and positive inputs which significantly reduces both the problems of exploding and vanishing gradients problems faced in deep neural networks. Graphical presentation of Zash AF with fixed nonlinear shaped legacy AFs Tanh and Swish is shown in orange color. The derivatives of these AFs are in blue color and same is projected in figure 5. It's clear from its clipping property that it will not be lost in exploding gradients. Its longer smoothness improves which settling for outputs and excluding chances of vanishing gradients as well. It achieved this property by rather an abrupt clipping like Tanh and swish. Primarily in Figure 5 Tanh and swish properties both are integrated in one single function Zash. ReLU's sparsity may cause loss to some of significant features in training of model. However, that is not the case with ZAF which accumulates all inputs and weights within dense activation.

4.1. Analyzing Proposed Activation Functions with Taylor Series

Keeping in view of the various activation functions discussed above, the introduced new ZAF which is computationally compatible to ReLU, Tanh and Swish AFs. Graphical presentation of Zash function is range bound and its derivative resembles the derivative of Tanh output. Moreover, Taylor series expansion of selected AFs up to order of 5. ReLU Taylor series is given as,

$$f(x) = 0\ f\ orxlessthan0 = 1\ f\ orxgreaterthan0 \tag{10}$$

$$\text{Tanh is,} \quad \tanh(x) = x - (1/3) * x\,3 + (2/5) * x\,5 \tag{11}$$

$$\text{Swish is,} \quad \text{swish}(x) = (1/2)x - (1/4) * x\,2 + (1/48) * x\,4 \tag{12}$$

$$\text{Zash is,} \quad \text{Zash}(x) = x - (1/2) * x\,3 + (1/6) * x\,4 + (5/24) * x\,5 \tag{13}$$

Taylor series expansion shows that Zash possess both Tanh AF and Swish polynomial properties with more expressiveness in its output. ZAF and Swish skipped second order while Tanh skipped 3rd order in

Taylor series. ReLU has no expressiveness in Taylor series. This property of expressiveness is further discussed in the results section.

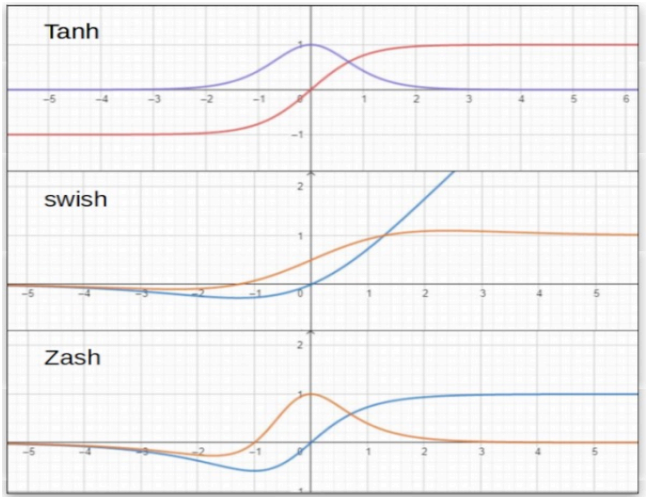

Figure 4. Fixed shape Activation functions with its derivative simulations

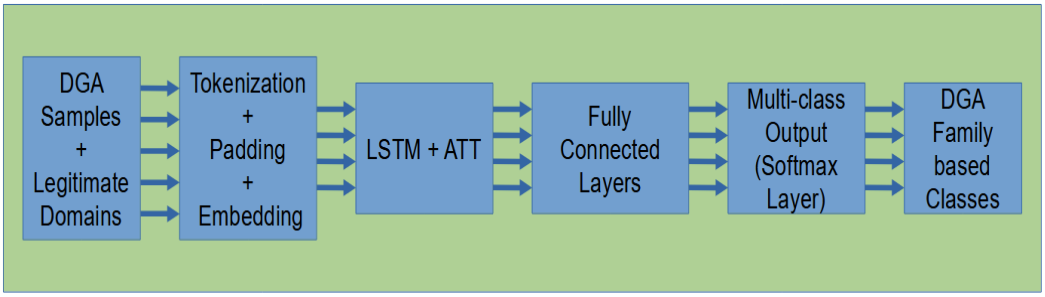

Figure 5. LSTM Attention implemented model

5. Proposed Methodology

Proposed methodology to undertake analysis of the DGA detection performance using LSTM with Attention model as depicted at Figure 5 has been adopted. All AFs are tested including a new AF Zash (ZAF) for optimal performance. Desired properties of proposed novel AF Zash are compared with legacy AF. DGA Dataset samples are composed of legitimate domain samples from Alexa [5] and malware domain samples of 20 DGA families from Bamabanek [6]. Data samples (entire dataset) are composed of Alexa data samples as legitimate domain names are the biggest dataset while the rest all are 20 families of DGA. The ratio of data set is set 75 % to 25 % for training and validation respectively. All the relevant codes and dataset are made available at [34].

Table 2: Performance results using LSTM model with state-of-the-art activation functions.

LSTM Model	**RELU**	**RELU**	**Tanh**	**Tanh**	**Swish**	**Swish**	**Zash**	**Zash**
Epochs	10	20	10	20	10	20	10	20
Accuracy	0.9298	0.9402	0.9118	0.9247	0.9206	0.9300	0.9193	0.9485
Precision	0.9214	0.9365	0.9016	0.9136	0.9111	0.9230	0.9136	0.9455
Recall	0.9298	0.9402	0.9118	0.9247	0.9205	0.9299	0.9192	0.9485
F1	0.9231	0.9345	0.9029	0.9160	0.9205	0.9226	0.9125	0.9438

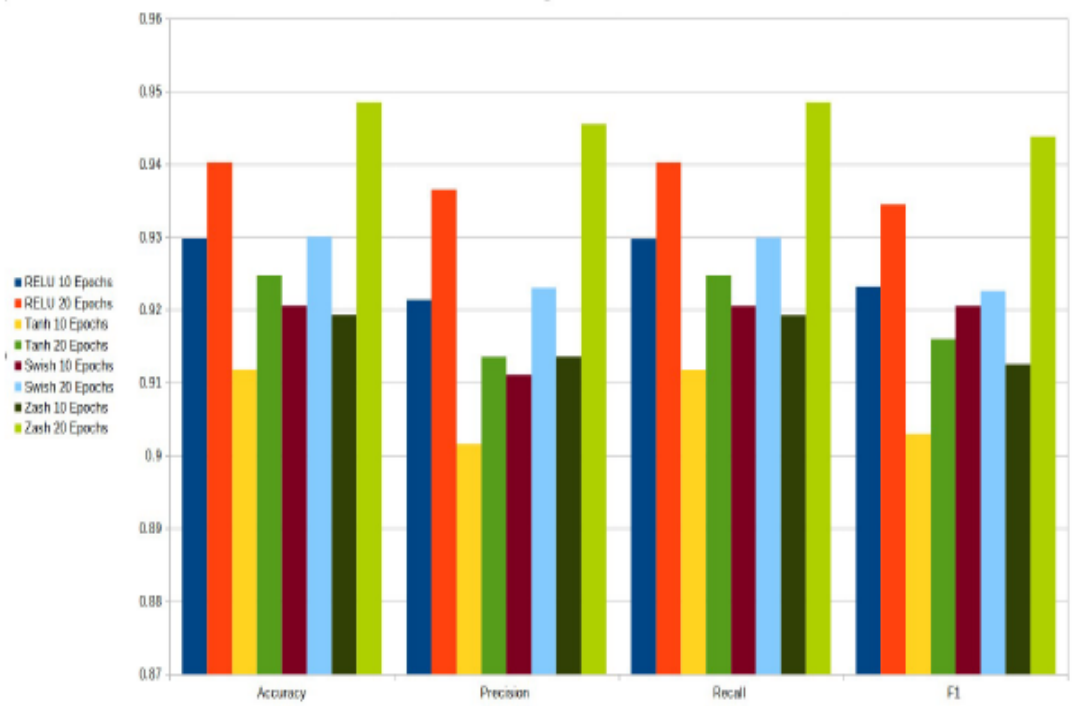

Figure 6. Graphical Presentation -Performance Metrics Comparison for 20 Epochs

Dataset samples are fed into LSTM with Attention model as projected in figure 5. This model process input data to extract maximum temporal relationships. Entire input sequence is converted as single context vector by the encoder. To address the loss factor for longer dependencies, an attention mechanism is integrated with LSTM model. Output from LSTM with attention mechanism is passed to fully connected (FC) layer for binary classification of legitimate and malicious domains. In the last layer Softmax layer is applied for multi-class classification. ReLU, Tanh, Swish and Zash AFs are set as default AF for LSTM with Attention model for comparison analysis of performance metrics (categorical accuracy, recall, precision and F1). The performance of the model is measured with (harmonic mean of precision and recall). Simulations are based on 20 epochs for each selected activation 240 functions [34].

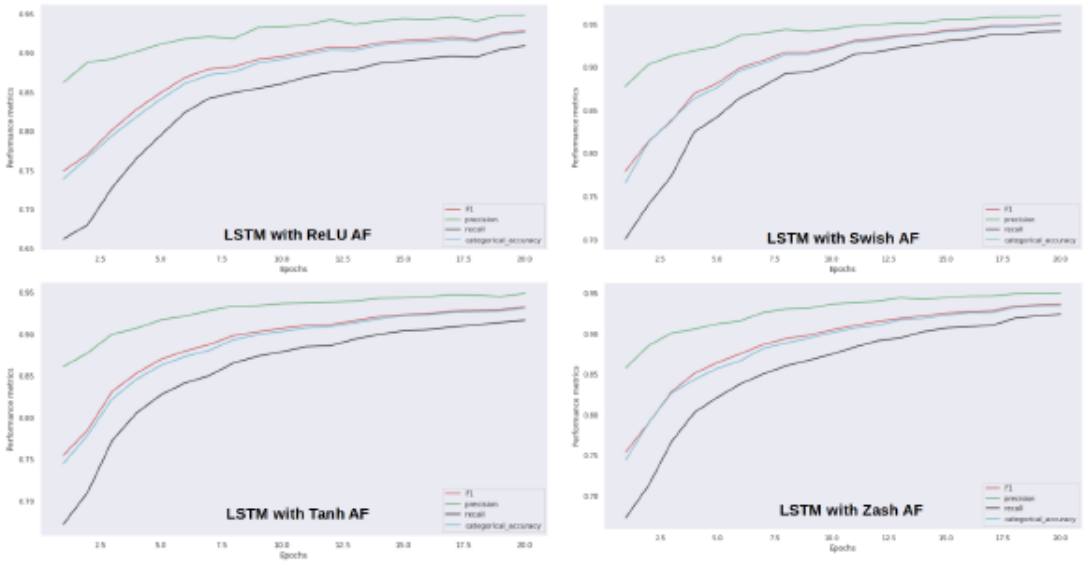

Figure 7. LSTM Attention Model showing Performance Metrics Curves

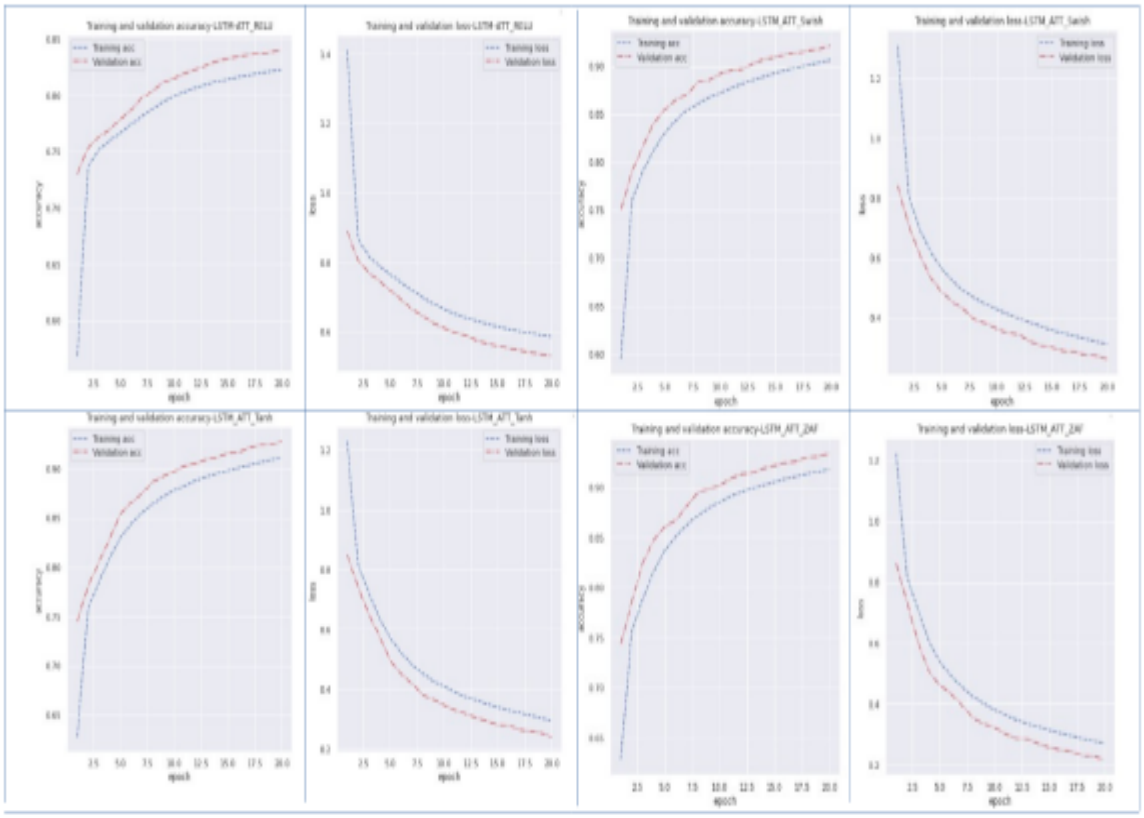

Figure 8. LSTM Attention Model Training/ Validation Accuracy and Loss Comparison

6. Results and Discussion

6.1. Results

Results of LSTM with Attention models were trained and tested with 4 variants of AFs (ReLU, Tanh, Swish and Zash) respectively. Each model has simulated twice for 10 epochs and 20 epochs respectively. The model performance metrics are projected both in a tabular form in Table-3 and graphically in Figure-6. Both Table-3 and Figure-6 projected results for 10 epochs and 20 epochs respectively. From numerical results to bar graphs in Figure-6 performance metrics of Zash AF is evidently outperforming the rest 3 legacy AF. In the Figures-7 overall Zash is performing with marginal difference with ReLU AF. However, Zash AF is showing much smoother curves in its all-performance metrics as well as a better convergence comparable with all AFs. This smooth training is visible in performance metrics curves and Zash AF is acting a better activation function. In Figure-8 model training and validation accuracy for 04 x AFs is projected. A close look at Zash AF raining and validation accuracy curve showing its smoothness and substantiating its smoothness over all other AFs. In same Figure-8 Training and validation loss is again substantiating its property of smoothness in converging of model at a point of stability. Figure-8 shows how the accuracy stable at end of 20 epoch for each model and proposed Zash AF validation curves for both accuracy and loss are smoother than other legacy AFs. Categorical accuracy curves are highest of ReLU, and second highest is Zash in all four models depicted in figure 9. The results of Swish and Tanh Afs performance metrics are quite similar but not better than Zash. ReLU and ZAF has shown overall best results. Convergence of Swish and Zash AFs are evident in figure 8. However, curves of Zash AF are smoother than rest of AFs which shows its superiority on rest of the three selected AFs. Implemented DL model results compared with selected AFs are highlighted in Figure-8

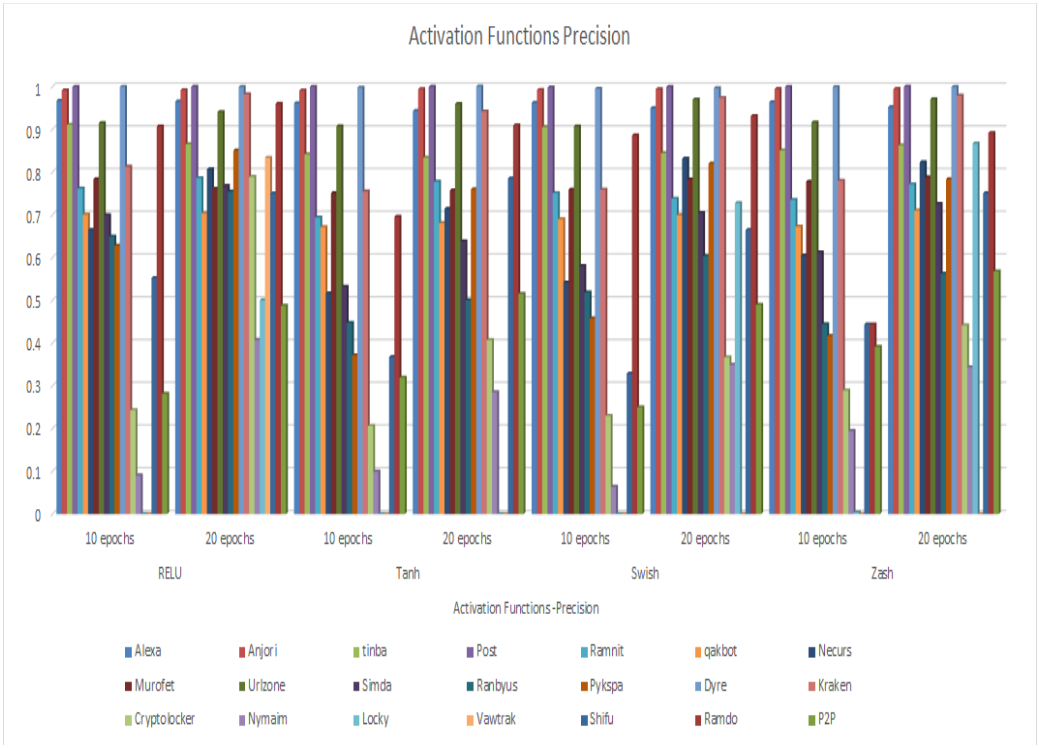

Figure 9. Precsion of 4 Activation Functions with 10 and 20 Epochs

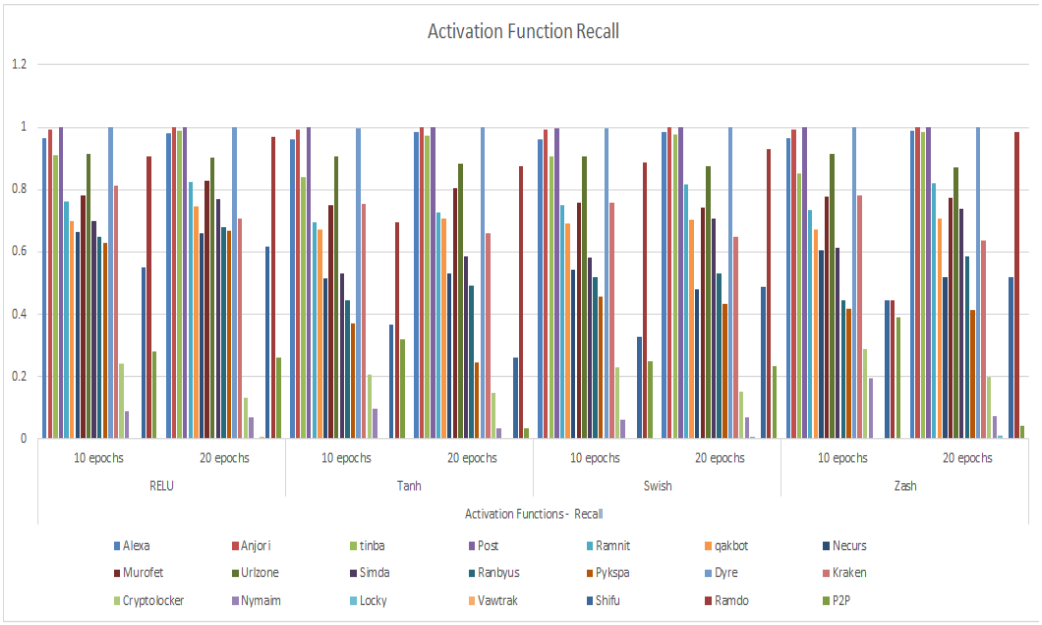

Figure 10. Recall of 4 Activation Functions

Each performance metric results are collated and projected in Figure-9, 10 and 11 for the performance metrics of Precision, Recall and F1 score respectively. Figure-9, 10 and 11 projected the performance metrics of the 4 AFs graphically with 10 and 20 epochs of training and validation. Overall ReLU and ZAF are closer in performance and competing in precision and recall). Simulations are based on 20 epochs for each selected activation 240 functions [34]. As compared to this projection, ReLU and ZAF have better performance metrics in their respective frames of Figures-9, 10 and 11. Moreover correlation matrices of the 4 models with of ReLU, Tanh, Swish and Zash AFs are also depicted in appendix to this paper for deeper visibility.

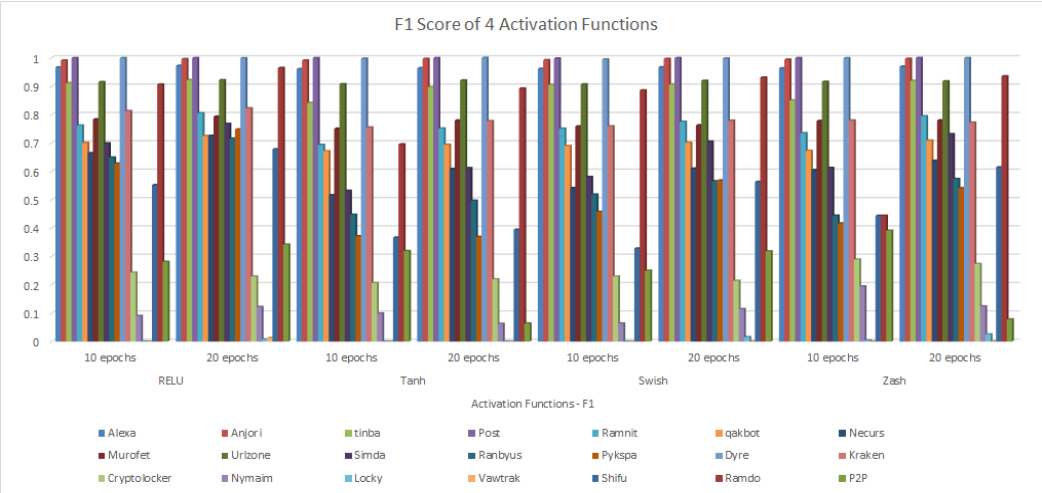

Figure 11. F1 of 4 Activation Functions

6.2. Discussion

Default fixed shaped AFs like sigmoid, Tanh and Swish use dense activations which are computationally expensive, time consuming and lead to vanishing gradients. On other hand rectifier-based AFs like ReLU use sparse activation which may also miss some important feed in initialization of training of a neural network. Keeping in view the trade-off between sparse and dense activations, selection of an AF becomes is very important parameter of neural network modelling. Swish works better than ReLU for deep neural network for more than 50 layers [35], therefore introduction of ZAF is to be considered as replacement of swish for minor and moderate deep learning models up to 40-50 layers. As in this case we tested our 4 models training up to 10 and 20 epochs. However, in nonlinear fixed shaped AFs Zash is closer to a rectifier AF ReLU with an additional property of dense activations. Zash AF has shown almost equivalent performance in LSTM models against Tanh and swish AFs with a marginal improvement. This indicates that further tuning of the model may further boost the performance of the Zash AF based models. Zash AF has shown that with dense activations fixed shape AF performance can come closer to rectifier-based AF like ReLU whenever dense activations are desirable. Zash AF still showing better than Tanh and Swish AFs in this real-world text classification problem.

6.3. Future Directions

ZAF may be implemented in models where data sparsity is intolerable and at the same time may be validated and tested further against default fixed shape competitors like Tanh and Swish AFs. CNN models default AF are ReLU, Leaky ReLU and ELU AFs and ZAF needs to be tested as future research directions in these models. The area of dense and sparse activations needs deeper investigations, and some may be part of future research works.

7. Conclusion

RNN networks work better with fixed shape AFs by default and inclined to dense activations. Tanh is considered a leading function within LSTM models. However, ZAF has almost same polynomial behaviour as of Tanh shown in Fig 4. Promising features of the ZAF needs to be refined further for LSTM networks using other model parameters. Future work may include fine tuning of other model parameters and hyper parameters, optimization algorithms and regularization techniques. ZAF may also be plugged in CNN and Attention/ Transformer models. DGA detection is achieved using DL models and hybrid approaches of these DL models have improved the performance further. In adoption of deep learning models, goal is to penetrate in APT detection cycle and detect DGA malware in organizational/ corporate networks. Specifying a DL model and identifying one of model key parameter like AF for optimization is a unidirectional approach. However, using these AFs like swish, Tanh and ZAF may be augmented with other DL parameters and hyper-parameters. Subject exploration has given an important insight of fixed shaped AFs in LSTM networks which may be further probed with rectifier shaped AFs. Authors declares that They have no conflict of interest. This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Leuprecht, C.; Skillicorn, D.B.; Tait, V.E. Beyond the Castle Model of cyber-risk and cyber-security. Government Information 320 Quarterly 2016, 33, 250–257Author 1, A.; Author 2, B. Title of the chapter. In Book Title, 2nd ed.; Editor 1, A., Editor 2, B., Eds.; Publisher: Publisher Location, Country, 2007; Volume 3, pp. 154–196.
2. Ahmad, A.; Webb, J.; Desouza, K.C.; Boorman, J. Strategically-motivated advanced persistent threat: Definition, process, tactics 322 and a disinformation model of counterattack. Computers & Security 2019, 86, 402–418.
3. Hutchins, E.M. ; Cloppert, M.J. ; Amin, R.M.; et al. Intelligence-driven computer network defense informed by analysis of 324 adversary campaigns and intrusion kill chains. Leading Issues in Information Warfare & Security Research 2011, 1,
4. Plohmann, D.; Yakdan, K.; Klatt, M.; Bader, J.; Gerhards-Padilla, E. A comprehensive measurement study of domain generating 329 malware. In Proceedings of the 25th USENIX Security Symposium (USENIX Security 16), 2016, pp. 263–278.
5. Kührer, M.; Rossow, C.; Holz, T. Paint it black: Evaluating the effectiveness of malware blacklists. In Proceedings of the 331 International Workshop on Recent Advances in Intrusion Detection. Springer, 2014, pp. 1–21.
6. Huang, D.; Xu, K.; Pei, J. Malicious URL detection by dynamically mining patterns without pre-defined elements. World Wide 333 Web 2014, 17, 1375–1394.
7. Krishnan, S.; Taylor, T.; Monrose, F.; McHugh, J. Crossing the threshold: Detecting network malfeasance via sequential hypoth- esis 335 testing. In Proceedings of the 2013 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Net- works 336 (DSN). IEEE, 2013, pp. 1–12.
8. Antonakakis, M.; Perdisci, R.; Dagon, D.; Lee, W.; Feamster, N. Building a Dynamic Reputation System for {DNS}. In Proceedings 338 of the 19th USENIX Security Symposium (USENIX Security 10), 2010.
9. Kidmose, E.; Stevanovic, M.; Pedersen, J.M. Detection of malicious domains through lexical analysis. In Proceedings of the 2018 340 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). IEEE, 2018, pp. 1–5.
10. Pereira, M.; Coleman, S.; Yu, B.; DeCock, M.; Nascimento, A. Dictionary extraction and detection of algorithmically generated 342 domain names in passive DNS traffic. In Proceedings of the International Symposium on Research in Attacks, Intrusions, and 343 Defenses. Springer, 2018, pp. 295–314.
11. Yadav, S.; Reddy, A.K.K.; Reddy, A.N.; Ranjan, S. Detecting algorithmically generated domain-flux attacks with DNS traffic 345 analysis. IEEE/Acm Transactions on Networking 2012, 20, 1663–1677
12. Schiavoni, S.; Maggi, F.; Cavallaro, L.; Zanero, S. Phoenix: DGA-based botnet tracking and intelligence. In Proceedings of the 347 International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Springer, 2014, pp. 192–211.
13. Zhao, H.; Chang, Z.; Wang, W.; Zeng, X. Malicious domain names detection algorithm based on lexical analysis and feature 349 quantification. IEEE Access 2019, 7, 128990–128999.
14. Cho, D.X.; Nam, H.H. A method of monitoring and detecting APT attacks based on unknown domains. Procedia Computer Science 351 2019, 150, 316–323
15. Antonakakis, M.; Perdisci, R.; Nadji, Y.; Vasiloglou, N.; Abu-Nimeh, S.; Lee, W.; Dagon, D. From {Throw-Away} Traffic to Bots: 353 Detecting the Rise of {DGA-Based} Malware. In Proceedings of the 21st USENIX Security Symposium (USENIX Security 12), 354 2012, pp. 491–506.
16. Woodbridge, J.; Anderson, H.S.; Ahuja, A.; Grant, D. Predicting domain generation algorithms with long short-term memory 356 networks. arXiv preprint arXiv:1611.00791 2016.
17. Akarsh, S.; Sriram, S.; Poornachandran, P.; Menon, V.K.; Soman, K. Deep learning framework for domain generation algorithms 358 prediction using long short-term memory. In Proceedings of the 2019 5th International Conference on Advanced Computing & 359 Communication Systems (ICACCS). IEEE, 2019, pp. 666–671
18. Tran, D.; Mac, H.; Tong, V.; Tran, H.A.; Nguyen, L.G. A LSTM based framework for handling multiclass imbalance in DGA botnet 361 detection. Neurocomputing 2018, 275, 2401–2413.
19. Vinayakumar, R.; Soman, K.; Poornachandran, P. Detecting malicious domain names using deep learning approaches at scale. 363 Journal of Intelligent & Fuzzy Systems 2018, 34, 1355–1367.
20. Saxe, J.; Berlin, K. eXpose: A character-level convolutional neural network with embeddings for detecting malicious URLs, file 365 paths and registry keys. arXiv preprint arXiv:1702.08568 2017
21. Dai, Z.; Xiong, C.; Callan, J.; Liu, Z. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In Proceedings 367 of the Proceedings of the eleventh ACM international conference on web search and data mining, 2018, pp. 126–134.
22. Qiao, Y.; Zhang, B.; Zhang, W.; Sangaiah, A.K.; Wu, H. DGA domain name classification method based on long short-term 369 memory with attention mechanism. Applied Sciences 2019, 9, 4205.
23. Namgung, J.; Son, S.; Moon, Y.S. Efficient Deep Learning Models for DGA Domain Detection. Security and Communication 371 Networks 2021, 2021
24. Nwankpa, C.; Ijomah, W.; Gachagan, A.; Marshall, S. Activation functions: Comparison of trends in practice and research for 373 deep learning. arXiv preprint arXiv:1811.03378 2018.
25. Apicella, A.; Donnarumma, F.; Isgrò, F.; Prevete, R. A survey on modern trainable activation functions. Neural Networks 2021, 375 138, 14–32.

26. Zhou, S.; Lin, L.; Yuan, J.; Wang, F.; Ling, Z.; Cui, J. Cnn-based dga detection with high coverage. In Proceedings of the 2019 IEEE 377 International Conference on Intelligence and Security Informatics (ISI). IEEE, 2019, pp. 62–67.
27. Yu, B.; Pan, J.; Hu, J.; Nascimento, A.; De Cock, M. Character level based detection of DGA domain names. In Proceedings of the 379 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018, pp. 1–8.
28. Highnam, K.; Puzio, D.; Luo, S.; Jennings, N.R. Real-time detection of dictionary dga network traffic using deep learning. SN 381 Computer Science 2021, 2, 1–17
29. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Proceedings of the fourteenth 383 international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323
30. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 2015.
31. Eger, S.; Youssef, P.; Gurevych, I. Is it time to swish? Comparing deep learning activation functions across NLP tasks. arXiv preprint arXiv:1901.02671 2019.