Novel Optimizer AdamW+ implementation in LSTM Model for DGA Detection

Awais Javed†, **Adnan Rashdi**†, **Imran Rashid**† **and Faisal Amir**†

National University of Science and Technology, Islamabad, Pakistan†

Summary

This work take deeper analysis of Adaptive Moment Estimation (Adam) and Adam with Weight Decay (AdamW) implementation in real world text classification problem (DGA Malware Detection). AdamW is introduced by decoupling weight decay from L2 regularization and implemented as improved optimizer. This work introduces a novel implementation of AdamW variant as AdamW+ by further simplifying weight decay implementation in AdamW. DGA malware detection LSTM models results for Adam, AdamW and AdamW+ are evaluated on various DGA families/ groups as multiclass text classification. Proposed AdamW+ optimizer results has shown improvement in all standard performance metrics over Adam and AdamW. Analysis of outcome has shown that novel optimizer has outperformed both Adam and AdamW text classification based problems.

Keywords:
DGA Detection, Deep learning, LSTM, Adam, AdamW AdamW+.

1. Introduction

Complex cyber attacks are sorted as sequential phases of attacks and known as Cyber Kill Chain (CKC)[1]. CKC phase where Command and Control (CC) servers is controlled by malicious actors exploiting authorized communication protocols to evade any detection. One of such communication protocols is DNS protocol. DNS protocol is abused by generating a bulk of malicious domains traffic in garb of non existent domains (NXDs). One such malicious domain out of bulk traffic is then connecting the infected systems with already configured CC servers. The target system is infected by an explicitly designed malware to exploit the inherited trust of DNS protocol called Domain Generating Algorithm (DGA) [2]. DGA detection is achieved using different Deep Learning (DL) models such as Long-Short Term Memory (LSTM) and Convolution Neural Networks (CNN) models. The optimal performance of these DL models was based on solving the (classifying malicious domains from legitimate domains). Presently, LSTM and CNN models has been applied profoundly on solving text classification problems. These DL models have accomplished text classification based DGA detection successfully. As far as these DL models have shown optimal performance in classifying legitimate domains from malicious domains, their model parameters are required to be focused and analyzed for further optimization. The DL models are dependent on various model parameters for further improvement in text classification; more precisely to classify DGA generated malicious DNS traffic from legitimate DNS traffic. Avoiding sprinkled optimization approaches, gradient optimization algorithms are chosen to confine the objective of this research work. Range of this research is to observe and evaluate the efficient gradient optimizer algorithms. Motivation of this research work is to employ and identify better performing gradient optimizer for text classification based problems for LSTM networks. DL models normally adopt Adam (Adaptive Moment Estimation optimizer)[8] as a default optimizer. It optimizes the Learning Rate (LR) and fastens the convergence of training model to a point of stability. Adam is improved further by AdamW (Adam with fixed weight Decay) [11] which deploys weight decay separately than conventionally assumed L2 regularization. The focus of this research work is the implementation of Adam and AdamW optimizer with respect to focus on weight decay parameter. In this proposed research work, implementation of weight decay as presented in AdamW has been made intrinsically simplified. The simplified variation of AdamW with respect to weight decay is named as Adam Weight Decay Plus (AdamW+). Empirical analysis of text classification for solving DGA detection problem is chosen for comparative analysis of momentum based optimization algorithms. These empirical analysis are introduced with novel optimizer as AdamW+ in LSTM with Attention model. Text classification based DGA detection is bench-marked for AdamW+ optimizer with the default Adam and AdamW optimizer. AdamW+ has shown that AdamW implementation may be made more efficient and has shown better performance over default momentum based optimizer like Adam and AdamW. This study is divided into five sections, first section is an

Manuscript received November 5, 2023
Manuscript revised November 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.11.17

introduction section. Section 2 encompass related work with two subsections of DL based DGA detection and evolution of Adaptive Gradient Optimization Algorithms respectively. Section 3 is proposed methodology of subject research work. Section 4 explains empirical implementation of proposed methodology and achieved results. Section 5 discusses the results and conclusive remarks with future directions.

2. Related Work

2.1 DGA Detection with Deep Learning

DGA malware produces varying alphanumeric patterns and sizes. as well as word based varying frequencies. Both patterns and sizes and frequencies are cleverly designed by malware engineers to avoid and bypass advanced detection systems. Further these patterns, sizes and frequencies associate the intended malware to respective DGA family, various DGA families samples are presented in **Table 1**. The bulk volumes of malicious domains data generated by DGA malware (both NXdomains and hidden malicious domains) conform it as a potential candidate for ML and DL models respectively. DL models achieved better performance in DGA classification and detection due to having an inherent auto features extraction and better results over DGA detection ML based models. Earliest methods of DGA malware detection included blacklisting of such domains from bulk generated domains by DGA [3,4].

Table 1: Comparative list of various malicious domains generated by different DGA families

Malicious Domains		
S.No.	**Samples/No of Characters**	**Associated DGA Family**
1.	ffqrgedkmxbwb.ru	Cryptolocker
2.	hpbbydetwdqsscqtnvljufaau.com	Gameover / P2P
3.	ffqrgedkmxbwb.ru/13	Conficker
4.	miodndu.ms	Necurs
5.	sizyvob.com	simda
6.	seekhecsfam.com	qakbot
7.	nxnucfb.info	shifu
8.	b83ed4877eec1997fcc39b7ae590007a.info	Bamital
9.	jwgjqwls2al51lnmeakehw60s.org	Post
10.	spq2sl7p7tc4c0gh5xux5vq.ddns.net	Corebot
11.	lb5a346868c31a36706a0f60573558a9d9.in	Dyre
12.	uxesxuibtecis.ddns.net	Symmi
13.	jjepdru.net	Nymaim
14.	bqvqpueebcjm.pw	Tinba
15.	aacibyplaywobb.com	Banjori
16.	zzln3q33xili4o.net	Shiotob/urlzone/bebloh
17.	fthmyvrefryk.com	Ramnit
18.	somjsdqwftbgbx.tw	Ranbyus
19.	wqnupo.net	Pykspa
20.	mryqqupmaskru.com	Murofet

DGA detection has been advanced with Machine Learning (ML) based detection and more recently, DGA detection has been elevated using Deep Learning (DL) models like LSTM and CNN models successfully. These DL models have brought considerable improvement in detection performance. Comparatively LSTM, CNN and even their hybrid approaches have been applied and have outperformed all previous methods. These DL models generally consist of LSTM, CNN, hybrid approaches of both LSTM and CNN models. Recently addition of Attention models with LSTM has further improved the detection performance. A brief overview of research work based on these models is depicted in **Table 2.**

DL models learns to differentiate between legitimate and malicious domains using training on both legitimate and malicious domains samples. DL models are fed with labeled samples of both legitimate domains and malicious domains for training and learning. In DL, LSTM models are considered ideal for text classification problems due to inherit ability of memory correlations for past inputs.

Table 2 Overview of DGA detection with Individual Deep Learning Models.

Advanced DGA Detection DL Techniques	**Years**	**Research Work done**
LSTM	2016	J. Woodbridge [5]
	2018	R Vinayakumar [6]
	2018	Duc tran [7]
CNN	2017	Joshua [8]
	2018	W. Bush [9]
	2019	Shaofanf Zhao [10]
LSTM with Attention /	2019	Y. Qiao [11]

Advanced DGA Detection DL Techniques	Years	Research Work done
Hybrid approach	2021	J. Namgung [12]

LSTM models are further augmented with Attention[11] have further improved the performance of DGA detection. Keeping in view the spectrum of this research, LSTM with Attention models are selected to optimize the performance of DGA detection models. In model parameters, gradient optimizer is selected as the core model parameter to be focused and evaluated in light of available best performing optimizer functions.

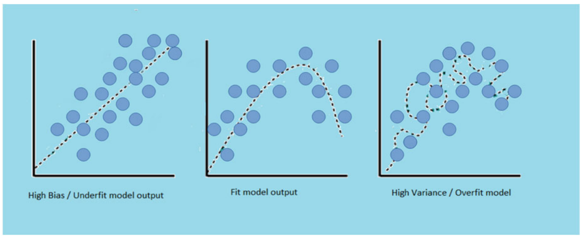


Fig 1 Models Under-fit, Fit and Over-fit Presentation

2.2 Evolution of Adaptive Gradient Optimization Algorithms

DL models are ascertained as either the model is fit, under-fit or over-fit during its training. Model generalization is observed with the convergence of Learning Rate (LR) culminating towards a point of stability. DL Model parameters with higher dimensions lead to higher non-linearity which support faster LR during the training. However, with higher parameter dimensions, a higher bias may also lead the model to become an under-fit model and a higher variance may lead to over-fitting of the model. To keep the bias and variance within limits to fit the DL model, gradient is moved in direction of desired global minima (a minimal loss point). Moving along the error slope or in terms of DL parameters gradient Descent (GD), learning rate (LR) determines the size of the step to reach the desired global minima. LR follows along the direction of slope by a function descending down to reach out the global minima. Framing this DL representational function as a stochastic function f and its parameters as θ, function f(θ) is optimized with the simplest approach as the Stochastic Gradient Descent (SGD) [13] is mapped as,

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla f(\theta) \qquad (1)$$

where η is the LR which defines the required step size to reach the local minima and ∇f(θ)is the rate of change of parameters θ with respect to objective function f. Stochastic Gradient Descent (SGD) with moment (SGDM) added a fraction β to update the parameters' first moment as m. Upgrading SGD equation 1 to SGD with first momentum as m_t at time step t,

$$\theta_{t+1} = \theta_t - m_t \qquad (2)$$

where,

$$m_t = \beta m_{t-1} + \eta \cdot \nabla f(\theta) \qquad (3)$$

SGD with momentum is upgraded by unveiling of Adagrad (adaptive gradient method) [14] which makes the gradient flexible to adapt the lower or higher LR (step sizes) instead of fixed step size with SGD. Adagrad has two main advantages, first it is well suited for sparsity of data and second it adjusts the tuning of LR (step sizes) eliminating the need of manual tuning. Adagrad has perceived the concept of adaptive LR moment from concept of moving averages. Adagrad is presented mathematically in equation 4 as,

$$\theta_{t+1,i} = \theta_t - \frac{\eta}{\sqrt{G_{t,i}+\epsilon}} \cdot gt \qquad (4)$$

$G_{t,i}$ is the sum of squares of gradients g_t at time t and i wrt parameters θ_t. The equation has clearly depicting that how the LR (step size) is now controlled by square root of gradients in action and ϵ is very small number to avoid division by zero. Adadelta [15] and RMSprop [16] (which are almost identical and not the scope of present research) has introduced fixed weight size accumulation, further improving with sum of squared gradients which is decaying average of all past squared gradients, it actually introduced second order moment estimation as v_t after first order m_t as,

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2)_t+\epsilon}} \cdot gt \qquad (5)$$

Adaptive Moment Estimation (Adam) [17] algorithm has further improved adaptive LR by computing the decaying averages of past and past squared gradients as m_t and v_t respectively as,

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\eta \cdot \nabla f(\theta) \qquad (6)$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\eta \cdot \nabla f(\theta) \qquad (7)$$

and essentially the bias correction as $\hat{v}_t$ and $\hat{m}_t$ to avoid the output being influenced by zero initialization.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \cdot \hat{m}_t \qquad (8)$$

However both SGD with momentum and Adam are observed to be generalizing poorly over diverse set of deep learning models due some inherent problems of momentum as well as adaptive gradients methods. In this case, weight decay is being identified as the propelling factor of these problems and its implementation is considered undermined. Same is fixed in [18] both SGD with momentum from equation 2 as SGDW and Adam from equation 8 as AdamW represented as

For SGDW:

$$\theta_t = \theta_{t-1} - m_t - w_t\theta_{t-1} \qquad (9)$$

and,

For AdamW:

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \cdot \hat{m}_t + w_t\theta_{t-1} \qquad (10)$$

Prime motive of Equations 9 and 10 is identifying the missing link of weight decay w_t which needs decoupling from L2 regularization and re-implementation as stand-alone parameter in SGD with momentum and Adam. LR is an adaptive parameter while weight decay w_t works as a coefficient (a small numerical value). There is another parameter called Rate Scheduler in original work presented [18], however it is not being implemented here for sake of focusing primarily on weight decay w_t.

3. Proposed Methodology

3.1 AdamW+ algorthim: A novel optimizer approach

Weight decay is used to regularize the DL models and is multiplied with model weights with a small numerical fraction during updating new weights. Weight decay was considered an integral part of L2 regularization which was justified in [18] by decoupling it from L2 regularization and specifically implemented as Equation 9 and Equation 10 respectively. Analyzing decoupling of weight decay w_t deeper, it is found that Equation 10 may be further simplified as;

$$\theta_t = \theta_{t-1}(1 - w_t) - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \cdot \hat{m}_t \qquad (11)$$

As weight decay parameter in Equation 11 is just a numerical figure and applied in fractions of logarithmic values such as 0.1, 0.01, 0.001 and so on. These default values for instance if added in Equation 11, will be adding as coefficients of 0.9, 0.99 and 0.999 and so on to the old weight θ_{t-1}. This will result in parameter θ_t in a meager correction as the case of weight decay w_t is generally started implementing from 0.001, 0.0001 and so on. Continuing on Equation 11, if we apply w_t as 0 equating the meager value to null and theoretically we regain Adam back as result of neutralizing the parameter w_t to zero. However, rather than using Adam again we implemented AdamW with w_t equal to 0 in Equation 11. This led us to discover that $w_t = 0$ is more optimized implementation of AdamW and same re-implementation is named as AdamW+. After identifying the novel optimization approach, the 3 optimizer Adam, AdamW and AdamW+ are tested in solving text based classification problem of DGA Detection. The three optimizer have been implemented on LSTM with attention DL models for DGA detection and subsequently comparing and evaluating the outcomes of the 3 optimizer.

3.2 Empirical Setup

LSTM with Attention model being considered one of advanced approach in solving text classification problem and same is adopted for DGA detection problem. Experimental setup started with legitimate domain samples from Alexa [19] (Alexa has been retired since 1st May 2022 by Amazon) and DGA samples from Bamabanek [20]. Training and testing dataset are composed of one Legitimate domain dataset against 20 classes of varying DGA families. Dataset samples are trained on 969072 samples and validated on 107675 samples from Alexa top million domain names as legitimate domains and malicious domain samples composing twenty DGA families projected as Fig – 2.

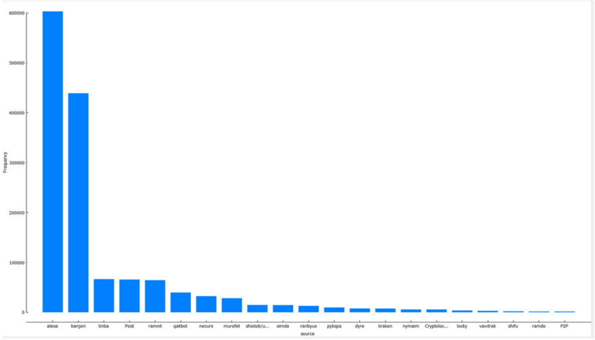


Fig 2 DGA Dataset Visual Breakdown alongwith Alexa being legitmate Dataset

Table 3 LSTM ATT Model Performance Metrics comparison of 3 Optimisers

AdamW LSTM-Att	Adam				AdamW+	
Epochs	**10**	**20**	**10**	**20**	**10**	**20**
Accuracy	0.9627	0.9664	0.9635	0.9679	0.9645	0.9686
Precision	0.9610	0.9654	0.9616	0.9671	0.9631	0.9686
Recall	0.9627	0.9664	0.9635	0.9679	0.9645	0.9686
F1 Score	0.9602	0.9652	0.9614	0.9665	0.9625	0.9686

3.3 LSTM with Attention Model

As LSTM is a state and context aware neural network, it is proficient in detection of temporal associations between texts. LSTM obtain contextual vectors of input sequences. Attention mechanism with LSTM model further improve the longer dependencies. DGA samples are passed through Seq2Seq encoder which compress input to a fixed length of a context vector. Each model as depicted accumulates the score of given input samples to classify it either as legitimate or malicious domain. As LSTM retains statefulness property and may face information loss in case of longer sequences. This information loss is addressed with addition of Attention model.

LSTM output is further fine tuned with Attention model. Later this binary classification is further classified using Softmax function to a specific DGA family (if classified as malicious domain name).

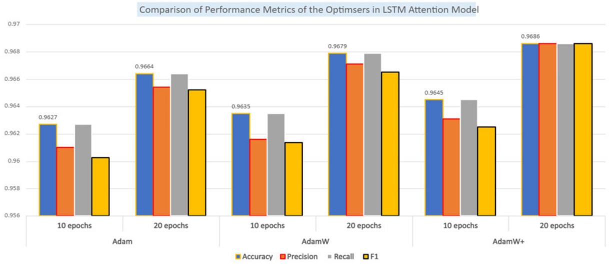
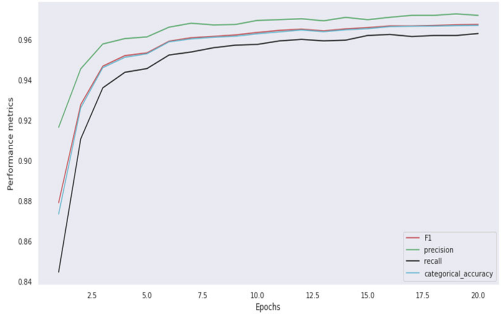
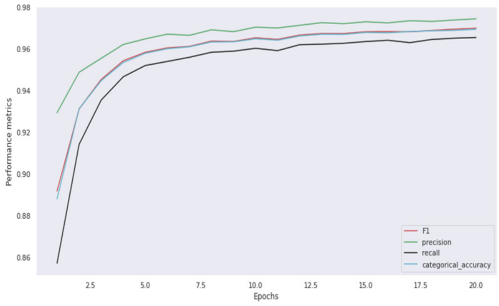


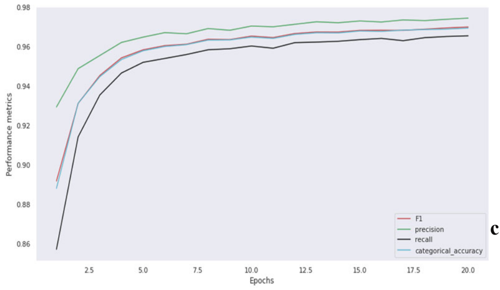
Fig 3 LSTM ATT Model Performance Metrics comparison of 3 Optimizer



a. LSTM Attention with Adam



b. LSTM Attention with AdamW



c.

C. LSTM Attention with AdamW

Fig 3 LSTM ATT Model Performance Metrics graphical comparison of 3 Optimizer

For multi-class datasets we use multi- classification model at final output layer. All the output of LSTM model is processed at FC layer with Softmax an output score of each class. Softmax output gives the alignment score of various outputs and classify them into different classes based on closeness of these defined alignment scores.

4. Results and Discussion

4.1 Results Performance metrics of these Deep Learning models are measured for DGA Detection with adoption of Adam, AdamW and AdamW+ optimizer. Two iterations of 10 epochs and 20 epochs are run to obtain results respectively.

(a) **Table-3 and Figure-3** Broader overview of all performance metrics outcome of each model is depicted in Table–3. Table-3 is projected with graphical illustration in Figure-3. Graphical depiction is showing that AdamW+ is outperforming the results in all default performance metrics of Accuracy, Precision, Recall and F1 score respectively.

(b) **Figure-4** Deeper observation of Figure-4 (c) of AdamW+ has outperformed by achieving 98%, which is showing comparatively better results than Adam at Figure 4 (a)and AdamW Figure 4 (b). AdamW is closer to AdamW+ however further results substantiate the performance of AdamW+ over AdamW. Same can be further validated from convergence of of performance metrics in Figure-4 (c) which is showing the out-performance, closing its approach to 98%.

(c) **Figure-5** Training and validation accuracy as well as training and validation loss of each model are projected in Figure-5 to identify how well the model is fit. It is evident that the accuracy and loss curves of all the depicted models have converged at an optimum value. It is observed that AdamW+ has better smoothness curves and more stable convergence in Figure-5 (c).

(d) **Figure-6** As two iterations of the 3 models have been run for 10 epochs and 20 epochs respectively, Figure-6 has shown performance metrics of Precision, Recall and F1 score for 10 and 20 epochs respectively for the 3 selected optimizer. The graphical display consists of 6 bars each summing up color codes for the 3 models and 2 iterations respectively. This 6 color bar projection is showing each model performance on legitimate domain names of Alexa against its classification over 20 DGA families.

4.2 Discussion

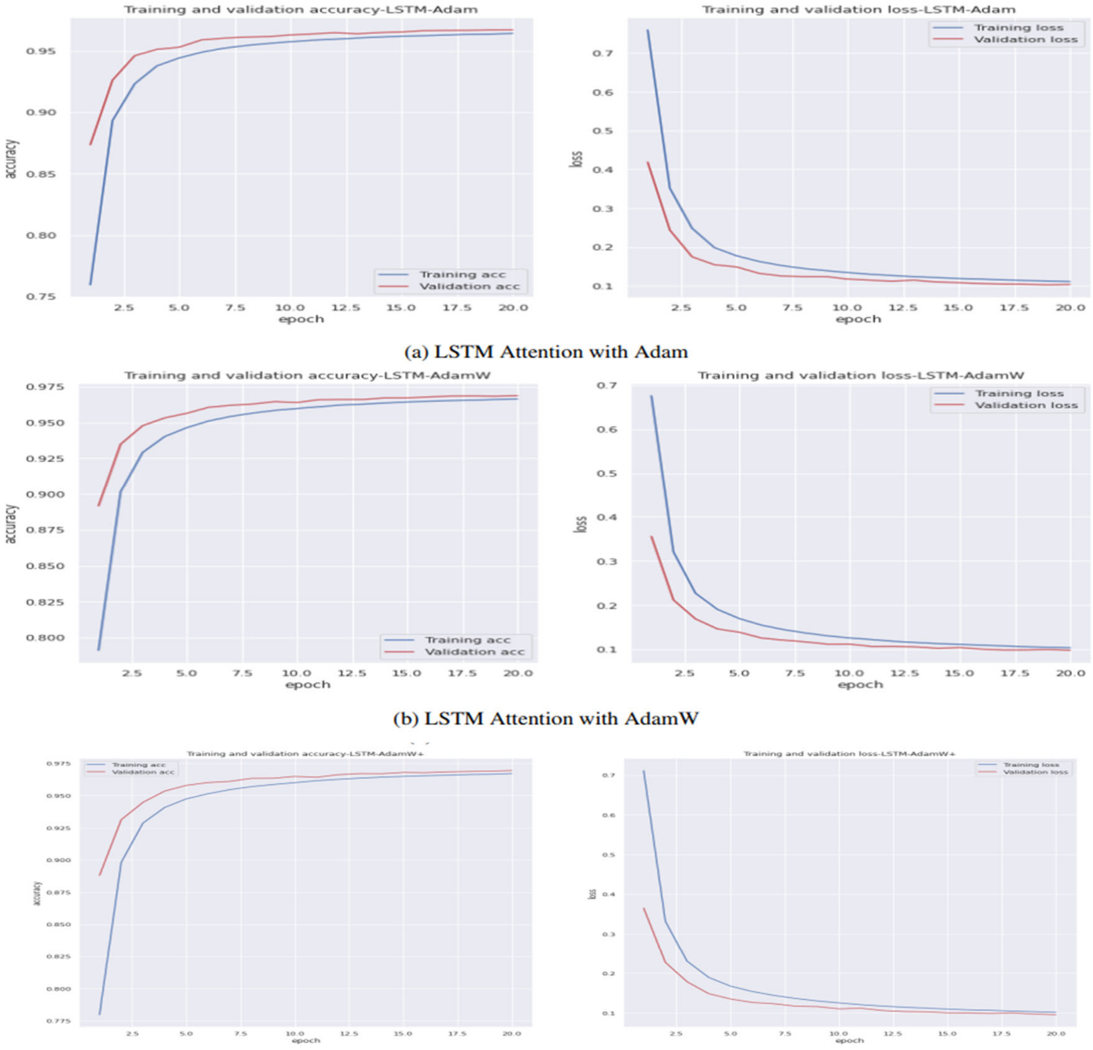
All models are implemented with dataset split of 90% training and 10% testing samples. All datasets, code repositories and results are available at [21] for reference, evaluation and future work. Table-3 is showing overall picture of performance metrics with same computational proficiency of each model in Fig-3. AdamW+ has shown optimal performance in detection of all 20 DGA families except the last P2P family. Further, Analyzing all results in table, figures and appendices revealed that DGA detection LSTM with Attention models have achieved an optimal performance. Applying Adam, Adam and AdamW+ optimizer have achieved significant results. The default optimizer Adam and AdamW are outperformed by the proposed novel AdamW+ with projected optimal performance. LSTM with Attention results have shown significant progress in all performance metrics in DGA detection. Keynote is identification of improvement in traditional optimizer specifically used for text classification based deep learning model. Morover adoption of attention model have further elevated the performance of LSTM based DGA detection model in achieving more than 97% with just 20 epochs.

5. Future Work

Future works shall include increase in epochs making deeper LSTM networks. In addition, switching of the proposed optimizer AdamW+ in larger text classification problems. AdamW+ shall also be tested in other DL models like CNN and Generative Adversarial Networks (GANs). Performance or progress of the same may be shared with authors for further analysis.

6. Conclusion

LSTM networks performance is optimized with various model parameters. Such model parameters include a vital parameter as model optimizer in training of a neural network. These optimizer are optimization algorithms continually map and project the learning curves as well as loss curves. These curves guide in identifying the model performance as well as achieving optimal efficiency. DGA detection being a real world text classification problem is selected as the basis of this research work. This text classification is solved using LSTM with attention model. LSTM with attention based DGA detection models are compared and analyzed to find the best performing optimizer. In this study case new dimension/ approach in optimizer has been gained and it is considered that these optimizer have shown optimal performance however same need further elevation in other standard community based text classification datasets as well as the real world problems.

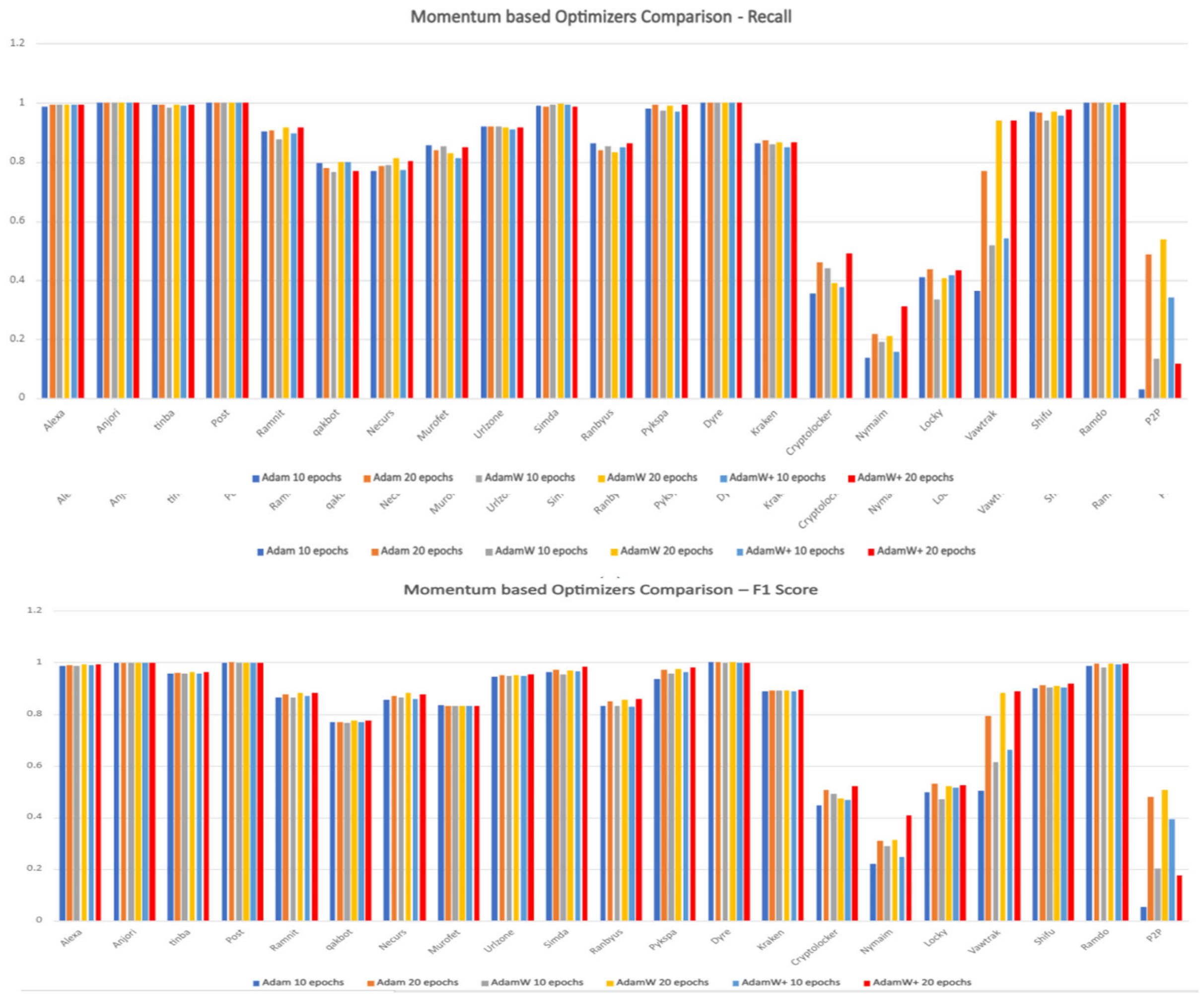


(a) LSTM Attention with Adam

(b) LSTM Attention with AdamW

(c) LSTM Attention with AdamW+

Fig. 5: Training and validation accuracy vs Training and validation accuracy loss comparison of 3 Optimisers



(c)

Fig. 6: Performance Comparison of (a) Precision (b) Recall (c) F1 score for 10 and 2C Epochs

References

[1] Eric M Hutchins, Michael J Cloppert, Rohan M Amin, et al. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*, 1(1):80, 2011

[2] Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, and Stefano Zanero. Phoenix: Dga-based botnet tracking and intelligence. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 192–211. Springer, 2014.

[3] Srinivas Krishnan, Teryl Taylor, Fabian Monrose, and John McHugh. Crossing the threshold: Detecting network malfeasance via sequential hypothesis testing. In *2013 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 1–12. IEEE, 2013.

[4] Marc Kuhrer, Christian Rossow, and Thorsten Holz. Paint it black: Evaluating the effectiveness of ¨malware blacklists. In

[5] *International Workshop on Recent Advances in Intrusion Detection*, pages 1–21.Springer, 2014.

[6] Jonathan Woodbridge, Hyrum S Anderson, Anjum Ahuja, and Daniel Grant. Predicting domain generation algorithms with long short-term memory networks. *arXiv preprint arXiv:1611.00791*, 2016.

[7] R Vinayakumar, KP Soman, and Prabaharan Poornachandran. Detecting malicious domain names using deep learning approaches at scale. *Journal of Intelligent & Fuzzy Systems*, 34(3):1355–1367, 2018.

[8] Duc Tran, Hieu Mac, Van Tong, Hai Anh Tran, and Linh Giang Nguyen. A LSTM based framework for handling multiclass imbalance in dga botnet detection. *Neurocomputing*, 275:2401–2413, 2018.

[9] Joshua Saxe and Konstantin Berlin. expose: A character-level convolutional neural network with embeddings for de-

tecting malicious urls, file paths and registry keys. *arXiv preprint arXiv:1702.08568*,2017.

[11] grams in ad-hoc search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 126–134, 2018.

[12] Shaofang Zhou, Lanfen Lin, Junkun Yuan, Feng Wang, Zhaoting Ling, and Jia Cui. Cnn-based dga detection with high coverage. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 62–67. IEEE, 2019.

[13] Yanchen Qiao, Bin Zhang, Weizhe Zhang, Arun Kumar Sangaiah, and Hualong Wu. Dga domain name classification method based on long short-term memory with attention mechanism. *Applied Sciences*,9(20):4205, 2019.

[14] Juhong Namgung, Siwoon Son, and Yang-Sae Moon. Efficient deep learning models for dga domain detection. *Security and Communication Networks*, 2021, 2021.

[15] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[16] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[17] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

M Awais Javed is a PhD Student of information security at Signals College of National University of Science and Technology (NUST), Islamabad, Pakistan. He qualified his MS in communication in 2015 from Pakistan Naval Engineering College of NUST, Karachi campus. awaiswill@gmail.com
, access control policies, information system management, and cloud computing.

Dr Adnan Rashdi is senior a post doctoral scholar and researcher in signals analysis department at Signals College of National University of Science and Technology (NUST), Islamabad, Pakistan. His expertise fields are SDR, Cognitive Radio, Signal analysis using machine learning and deep learning methods.

Dr Imran Rashid is chief instructor of Information Security department at Signals College of National University of Science and Technology (NUST), Islamabad, Pakistan. His expertise include Network security, wireless communication security and SDN.

Dr Faisal Amir Khan is assistant professor in nformation security department of National University of Science and Technology (NUST), slamabad. He pioneered various information security programs and projects at organizational evel.

[10] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. Convolutional neural networks for softmatching n-

[18] Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[21] Alexa. https://alexa.com.

[22] bamabanek. https://www.bambenekconsulting.com/.

[23] https://www.kaggle.com/awaisjaved/lstm-model-with-optimisers/