Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 2, 2024

Memory and Attention in Deep Learning

Yi Zhang[1, a], Ziying Fan[2, b]

[1]Intellifusion Pty Ltd, Melbourne, Aussie
[a]yizhang@xs.ustb.edu.cn, [b]Ziyingfan1@gmail.com

Abstract: This paper has highlighted the advancements made within the deep learning approached through the evolution of attention and memory algorithms. There has been a gradual replacement of the traditional approaches in deep learning through the inclusion of these algorithms that helps in capturing the data based on time and sequence. The model performance is greatly enhanced through the usage of RNNs that uses these algorithms for developing sequential modelling of data. This paper has performed a peer review to understand the different mechanisms that include GRUs, MANN, LSTM and self-attention mechanisms. Memory mechanism assists in capturing the past sequence of datasets for analysing the hidden state within the datasets. Attention mechanisms help in capturing a particular location within a video dataset to understand the patterns of a data. There is an accurate recognition of human actions in the datasets through the implementation of attention mechanisms. This helps in increasing the model performance by enhancing the prediction accuracy and visibility within a particular dataset.

Keywords: Lstm, Deep Learning, Memory Mechanisms.

1. Introduction

The two fundamental components of deep learning are considered as attention and memory mechanisms. Data complexity is clearly diagnosed through the implementation of these mechanisms and helps in increasing prediction accuracy on a large scale. The performance of "deep learning models" are greatly enhanced through the implementation of attention and memory mechanisms. Cognition guidance is considered to be a necessary aspect in the development of focus over a particular aspect by ignoring others. This aspect is provided through these mechanisms in deep learning that helps in picking salient information that are available within noisy data. This is effective in memorizing one particular event at a time and increases the model performance by analyzing the dataset features. This paper has identified an issue within the "traditional deep learning approaches" in processing and analyzing massive three-dimensional volumes by achieving fine-grained details and spatial context. This indicates the generation of false positives and missed detections through the formation of "suboptimal model performance" leading to the decline in prediction accuracy and reliability. Based on this issue, this paper has conducted a research on attention and memory mechanisms to understand their contributions made in enhancing the suboptimal performance of deep learning techniques. This paper has made a detailed investigation about different attention and memory mechanisms used in increasing the quality and prediction accuracy. Further, this paper has elaborated the methods that are used for conducting this research and has provided a peer review from different journals to make a proper investigation about attention and memory mechanisms.

2. Background

Neural networks help in the representation of deep learning through the usage of artificial intelligence [23] (Smys, Chen and Shakya 2020). These networks possess interconnected nodes referred to as neurons that help in transmitting information in a particular direction. Traditional "neural" networks consist of more than one hidden layer, output layer and input layer. There is availability of connection within different nodes by having thresholds and associated weights. Multiplied versions of random weights and inputs are received by neurons. There has been addition of activated functions and a "static bias value" that helps in the formation of a final outcome. The performance of "traditional neural networks" are getting beyond scope and heavy to handle some typical application scenarios (for example, medical diagnosis)[22] (Salehi *et al*. 2023). There is a limitation within "traditional neural networks" in handling sequential data. This is because these networks help in the independent processing of inputs without having any presence of past memories [1](Alzubaidi *et al*. 2021)[1]. As a result, it is unable to capture the dependencies and relationships with respect to time and thereby leading to the decline in the model performance. This has led to the generation of attention and memory mechanisms within deep learning that helps in the sequential handling of data. Memory mechanisms help in increasing the capability for retaining data in models that are available within a particular input sequence[27] (Xi *et al*. 2023). The tasks containing massive datasets are retrieved and stored by memory mechanisms for understanding the tasks minutely. Attention mechanisms help in developing the model capability by developing focus over a specific portion of a dataset [3](Dai *et al*. 2021). This is effective in the complete elimination of large background data and thereby leading to the decline in the irrelevant data within a dataset. Attention mechanisms help in understanding the sequence through the acquisition of dependencies available within input data weights. The aim of this paper is to determine the different attention and memory mechanisms that have helped in increasing the accuracy, reliability and model performance of "deep learning models".

2.1. Methods

Linking with the aim of this research, it is suggested that this research is accomplished through the consideration of secondary data for gathering detailed information about different attention and memory mechanisms used in deep learning. Peer review is done in this research through the

consideration of different journals and articles. Qualitative research is done through the use of secondary data that is available in different journals [15](Nassaji 2020). There is accumulation of qualitative information based on different models used in attention and memory mechanisms within deep learning. This research has made a selection of journals that are available within the last 6 years and have accessed them from Google Scholar database. Appropriate descriptions are provided on different attention and memory mechanisms to elaborate the contribution made by them in deep learning models.

3. Results

3.1. Memory mechanisms

3.1.1. Gated Recurrent units

According to [20]Rajamani *et al*. 2021, there has been wide usage of gating mechanisms within "neural network models". These mechanisms permit gradients in backpropagation in a much easier manner through time and depth. One of the most prominent features of GRU is the presence of activation functions. RNNs that include GRU and LSTM help in capturing the sequential data that possesses temporal dynamics. "Gated Recurrent Units" are considered to be a feature of RNN that helps in the implementation of gating mechanisms. This feature helps in the development of control over information flow and thereby help in managing cells available within one neural network. The GRU structure helps in the quick acquisition of dependencies that are available within the data sequences of a large dataset, especially videos[18] (Rafiq, Rafiq and Choi 2023). This helps in ensuring the data available in the previous sections of a data sequence. Information regulation is accomplished through the implementation of gating mechanisms that helps in distorting or keeping information at different time steps. Therefore, the argument could be made that there is a gradual increase in the model performance of deep learning through GRUs as it helps in the complete elimination of the "vanishing gradient problem" [30](Zulqarnain *et al*. 2020). This is effective in increasing the speed of model training as that is a considerable reduction in parameters through the usage of optimization technique.

According to [8]Jordan *et al*. 2021, data is written within the models through the implementation of memory mechanisms. This is accomplished through the implementation of "gated recurrent units" that helps in the development of RNNs through the inclusion of "specialized memory elements" [2](Chandra *et al*. 2021). The RNNs help in utilizing and capturing the sequential structure that is available within artificial and natural languages that include videos, speech and time series sequences [24](Torfi *et al*. 2020). This is effective in capturing the information flow by developing influence over the presence state based on the past data sequence. Park *et al*. 2021 states that the GRUs implements two different "internal gating variables" that include one reset gate and one update gate. A reset gate permits hidden state overwriting and thereby helps in developing control over the interactions developed with an input. An update gate helps in safeguarding the d-dimensional "hidden state". Through the implementation of this update gate, there is formation of control over different dimensions that are available within a "hidden state decay" and thereby help in the formation of one adaptive "time constant" for memory. Therefore, the argument could be made that there has been a gradual increase in the model performance through the implementation of differential equations in RNNs. This is effective in retrieving and storing memory by using numerical approximation techniques in GRUs.

3.2. Long short-term memory

According to [11]Landi *et al*. 2021, there has been usage of LSTMs that implements RNNs for developing gating mechanisms Figure 1. This is effective in enhancing the model performance by learning different long-term dependencies by eliminating vanishing and exploding gradients available within a dataset. During a forward pass, the "long-term memory cell" is involved in storing information and thereby helps in the accomplishment of back propagation of one error signal through one safe path [19](Ragab *et al*. 2020). There has been establishment of “working memory connections” that helps in enhancing the gates’ values under the influence of memory cells through recurrent weights. A constant “error paths are developed within subsequent time steps” through the implementation of LSTM networks.

$$\mathbf{i}_t = \sigma(\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1} + \tanh(\mathbf{W}_{ic}\mathbf{c}_{t-1}) + \mathbf{b}_i)$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1} + \tanh(\mathbf{W}_{fc}\mathbf{c}_{t-1}) + \mathbf{b}_f)$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1} + \tanh(\mathbf{W}_{oc}\mathbf{c}_t) + \mathbf{b}_o),$$

Figure 1. Protection mechanism(Source: Baraldi et al. 2021)

Baraldi *et al*. 2021 highlights that a protection mechanism is developed through the implementation of LSTM networks. This helps in the development of a connection within gates and memory cells through the generation of one “non-linear activation function”. Information accessibility is raised through the consideration of one LSTM cell that uses “Working Memory connections” for developing a suitable access within a gate structure. The improvements that are made through the inclusion of WMCs are increase in gate controls and rise in stability for training function.

According to Le 2021, there is availability of a mathematical challenge of RNNs in reading long sequences. This is because the gradients either explode or vanish while accomplishing step propagation. Figure 2 The complexity is developed through the allocation of smaller weights for accomplishing long-term interactions that are linked with different long-term dependencies. This issue is resolved through the inclusion of LSTM that helps in the addition of one “linear self-loop memory cell” for computing one hidden unit.

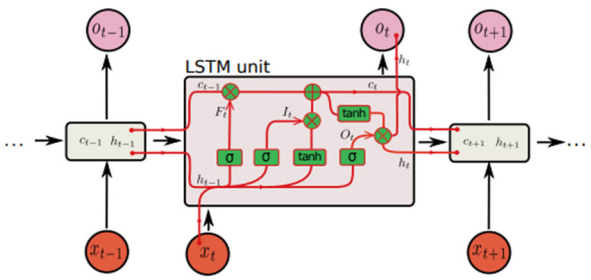

Figure 2. LSTM unit [13](Source: Le 2021)

The weight associated with this memory cell is gated and directly depends over an input. As a result, dynamic

moderation is developed within the network for controlling the data passing through one hidden unit. Figure 3 shows that there has been inclusion of different additional parameters to achieve vector representation.

$$f_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right)$$

Figure 3. Hidden state computation of forget gate [13](Source: Le 2021)

The "forget gate" is involved in capturing long-term dependencies by allowing the network to preserve the previous memory. As a result, there is formation of backpropagation for one distant input through the complement elimination of the "gradient vanishing process".

3.3. Memory augmented neural networks

According to Park *et al*. 2018, MANNs involve external networks like memory networks and Turing machines that helps in increasing the performance of cRNNs through the increase in learning capabilities for long-term dependencies. There is availability of two components in MANNs that help in increasing the model performance. These include external memory and memory controllers. MANNs help in writing and reading memory by considering the information. The memory addressing feature plays a critical role in interference and training.

$$C(M,k)[i] = \frac{\exp\{S(M_i,k)\}}{\sum_{j=1}^{L}\exp\{S(M_j,k)\}},$$

Content-based addressing[17](Source: Park *et al*. 2018)

There has been implementation of content-based addressing by MANNs that helps in increasing the performance of RNNs. This is accomplished through the consideration of a "similarity measure" that is represented as S that helps in representing the cosine similarity within one key vector and a memory element. The normalization is accomplished through the consideration of one SoftMax function. Due to the availability of differentiable components within SoftMax and cosine similarity, the gradient backpropagation is implemented by MANN for establishing an end-to-end trainable feature. Kim *et al*. 2018 states that binary and fixed-point quantization is done for developing accountability over the "quantization error". There has been implementation of XNORNet and BinaryNet in CNN for "binary quantization" of activations and parameters [25](Weng 2021). This is effective in minimizing energy consumption. BinaryNet has been implemented over CIFAR-10 and MNIST datasets. Information loss is compensated through the use of XNOR-Net that helps in the adjustment of layer positions through the consideration of one scale factor. This is effective in increasing the error rates by at least 50% to diagnose the layer positions available within image datasets.

3.4. Attention mechanisms

3.4.1. Temporal and spatial attention mechanisms

According to [5]Guo *et al*. 2022, there has been a considerable increase in computer vision through the consideration of temporal and spatial mechanisms that help in the appropriate recognition of human actions. Attention mechanisms help in the enhancement of "computer vision" by developing imitation over the visual system of humans. These mechanisms help in semantic segmentation, multi-video understanding, object detection and image classification. Spatial attention helps in the generation of an attention mask over spatial domains. This is effective in selection of spatial regions and positions. Temporal attention helps in the generation of an attention mask within time for the appropriate selection of key frames. Spatial attention helps in the establishment of one adaptive spatial region selection mechanism" for developing focus on a particular region for human actions. GENet, RAM, Non-local and STN are the different spatial attention techniques [12](Lau *et al*. 2024). Relevant regions are explicitly predicted through the implementation of one subsequent network by using STN. Soft Max is implicitly predicted through the implementation of GENet that uses one subnetwork for choosing important regions. Visual attention is developed through the use of reinforcement learning techniques in RAM and thereby help in developing focus over a particular location [4](de Santana Correia and Colombini 2022). Xu *et al*. 2022 stated that temporal attention helps in the implementation of a "dynamic time selection mechanism" for selecting a particular time in paying attention. This technique is effective in video processing by picking a particular time in analyzing the information flow sequence. The TAM module is implemented for developing temporal attention through the consideration of one adaptive kernel. This helps in capturing the "global contextual data" for minimizing time complexity**.** Computation cost is reduced through the consideration of 1D convolutions that include ReLU nonlinearity within the "temporal domain" [29](Zhao *et al*. 2020). This helps in generating location-sensitive maps for frame-wise features enhancement.

3.5. Self-attention mechanisms

According to Hafiz *et al*. 2021, in the current scenario, the attention-based mechanisms act as transformers for the accomplishment of "machine vision tasks". This is done through the usage of vision transformers that help in object detection, image recognition and video understanding. The implementation of a self-attention technique helps in the formation of one measurable estimate based on data relevance available within other elements [26](Wilson *et al*. 2023). This technique acts like transformers for modeling dependencies available within a sequence. This helps in enhancing the prediction accuracy and capability within deep learning techniques. A value is assigned within an element within a particular sequence through the use of a "self-attention model layer". This helps in the global combination of the information that is derived from an "input sequence".

$$Z = \mathbf{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_q}}\right)\mathbf{V}.$$

Determination of output from self-attention layer[6](Source: Hafiz et al. 2021)

Self-attention helps in capturing dependencies by transforming values and queries [28](Xie *et al*. 2020). The query dot product that is available within a sequence is diagnosed through the implementation of self-attention techniques. The normalization of this product is accomplished through the consideration of a SoftMax function [16](Pan *et al*. 2022). As a result, there is achievement of attention-map scores that helps in determining the "weighted summation" of

all elements available within a sequence. Parah *et al*. 2021 stated that recently, this feature has been implemented in image classification and face recognition for analyzing the sequence and thereby developing outputs based on the weighted summation. “Convolution vision transformers” are used for the establishment of one linear projection for analyzing the sequence and thereby increasing the prediction accuracy [21](Ranftl, Bochkovskiy and Koltun 2021).

According to [7]Hernández *et al*. 2021, self-attention helps in the establishment of relationships within numerous positions available within a particular sequence. Sequence transformation is accomplished for developing computation using queries, keys and values [14](Lin *et al*. 2020). This technique helps in the dynamic calculation of weights by availing direct transformation within inputs. Direct relationships are established within an output vector and inputs through the use of this module [10](Kossen *et al*. 2021). However, the local interactions are not prioritized through the use of self-attention techniques.

4. Discussion

Referring to the above findings, it is summarized that there is accomplishment of sequential modeling through the consideration of attention and memory mechanisms. Through the use of these mechanisms, focus is developed over a particular location and time within an available dataset. This is effective in the suitable recognition of hidden parameters and thereby helps in increasing prediction accuracy. There has been consideration of the state-of-the art approach through the use of attention and memory mechanisms in deep learning. Referring to the context of GRUs, the "vanishing gradient problem" is eliminated through the usage of gating mechanisms. This is done through the effective consideration of the dependencies that are diagnosed within a dataset. The usage of activation functions in GRUs helps in increasing model performance by developing in-depth investigation over dataset sequence. Further, numerical approximation is done through the usage of gate controllers that helps in analyzing the “hidden state decay” in terms of a time constant. Referring to the context of LSTM, the usage of “Working memory connections” helps in the establishment of a protection mechanism by developing error paths based on time. This approach helps in increasing the performance of “deep learning models” by establishing one “non-linear activation function”. LSTM helps in adding one “linear self-loop memory cell” and thereby helps in the suitable computation of hidden units. Referring to the context of MANNs, there has been formation of content-based addressing that helps in the effective diagnosis of similar features available within vector and memory elements. Quantization error is reduced through the use of LSTM that helps in the implementation of binary and fixed-point quantization. Referring to the context of “temporal and spatial mechanisms”, there has been a considerable enhancement in the visual systems through the effective diagnosis of spatial regions and positions. These mechanisms help in the gradual increase in the model performance by analyzing location and time for capturing the information flow within a data sequence. There is formation of location-sensitive maps that helps in understanding the changes occurring within human actions with respect to time. Referring to the context of self-attention mechanisms, there has been formation of weighted summation based on attention-map scores. This helps in enhancing prediction accuracy by developing accountability over position vectors [9](Kardakis *et al*. 2021). The SoftMax function helps in the development of measurable outcomes by establishing proper relationships within output vectors and inputs.

5. Conclusion

To summarize, the evolution of attention and memory mechanisms helped in increasing prediction accuracy in deep learning. These mechanisms have helped in the accomplishment of sequential modeling by analyzing vectors to determine human actions. LSTMs are considered to be an innovative feature of RNNs as it helps in development of outcomes based on the changes occurring within a dataset at a particular time. Time and location are the major aspects that are considered by LSTMs and MANNs for generating prediction outcomes. These mechanisms help in the accomplishment of dynamic moderation by understanding the accurate event for developing focus. Gate controllers are deployed by LSTM for increasing control over inputs. This helps in the significant reduction of errors in the dataset and thereby helps in increasing prediction accuracy. Further investigations could be developed in the future concerning the architectures of each mechanism. This will be effective to analyze the arithmetic calculations developed by different models in generating accurate outcomes.

References

[1] Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. and Farhan, L., 2021. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8, pp.1-74.

[2] Chandra, N., Ahuja, L., Khatri, S.K. and Monga, H., 2021. Utilizing Gated Recurrent Units to Retain Long Term Dependencies with Recurrent Neural Network in Text Classification. J. Inf. Syst. Telecommun, 2, p.89.

[3] Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L. and Zhang, L., 2021. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7373-7382).

[4] de Santana Correia, A. and Colombini, E.L., 2022. Attention, please! A survey of neural attention models in deep learning. Artificial Intelligence Review, 55(8), pp.6037-6124.

[5] Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M. and Hu, S.M., 2022. Attention mechanisms in computer vision: A survey. Computational visual media, 8(3), pp.331-368.

[6] Hafiz, A.M., Parah, S.A. and Bhat, R.U.A., 2021. Attention mechanisms and deep learning for machine vision: A survey of the state of the art. arXiv preprint arXiv:2106.07550.

[7] Hernández, A. and Amigó, J.M., 2021. Attention mechanisms and their applications to complex systems. Entropy, 23(3), p.283.

[8] Jordan, I.D., Sokół, P.A. and Park, I.M., 2021. Gated recurrent units viewed through the lens of continuous time dynamical systems. Frontiers in computational neuroscience, 15, p.678158.

[9] Kardakis, S., Perikos, I., Grivokostopoulou, F. and Hatzilygeroudis, I., 2021. Examining attention mechanisms in deep learning models for sentiment analysis. Applied Sciences, 11(9), p.3883.

[10] Kossen, J., Band, N., Lyle, C., Gomez, A.N., Rainforth, T. and Gal, Y., 2021. Self-attention between datapoints: Going

beyond individual input-output pairs in deep learning. Advances in Neural Information Processing Systems, 34, pp.28742-28756.

[11] Landi, F., Baraldi, L., Cornia, M. and Cucchiara, R., 2021. Working memory connections for LSTM. Neural Networks, 144, pp.334-341.

[12] Lau, K.W., Po, L.M. and Rehman, Y.A.U., 2024. Large separable kernel attention: Rethinking the large kernel attention design in cnn. Expert Systems with Applications, 236, p.121352.

[13] Le, H.T., 2021. Memory and attention in deep learning, 1. doi:https://doi.org/10.48550/arXiv.2107.01390.

[14] Lin, Z., Li, M., Zheng, Z., Cheng, Y. and Yuan, C., 2020, April. Self-attention convlstm forspatiotemporal prediction. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11531-11538).

[15] Nassaji, H., 2020. Good qualitative research. Language Teaching Research, 24(4), pp.427-431.

[16] Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z. and Huang, G., 2022. On the integration of self-attention and convolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 815-825).

[17] Park, S., Kim, S., Lee, S., Bae, H. and Yoon, S., 2018, April. Quantized memory-augmented neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

[18] Rafiq, G., Rafiq, M. and Choi, G.S., 2023. Video description: A comprehensive survey of deep learning approaches. Artificial Intelligence Review, pp.1-80.

[19] Ragab, M., Chen, Z., Wu, M., Kwoh, C.K., Yan, R. and Li, X., 2020. Attention sequence to sequence model for machine remaining useful life prediction. arXiv preprint arXiv:2007.09868.

[20] Rajamani, S.T., Rajamani, K.T., Mallol-Ragolta, A., Liu, S. and Schuller, B., 2021, June. A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6294-6298). IEEE.

[21] Ranftl, R., Bochkovskiy, A. and Koltun, V., 2021. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 12179-12188).

[22] Salehi, A.W., Khan, S., Gupta, G., Alabduallah, B.I., Almjally, A., Alsolai, H., Siddiqui, T. and Mellit, A., 2023. A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. Sustainability, 15(7), p.5930.

[23] Smys, S., Chen, J.I.Z. and Shakya, S., 2020. Survey on neural network architectures with deep learning. Journal of Soft Computing Paradigm (JSCP), 2(03), pp.186-194.

[24] Torfi, A., Shirvani, R.A., Keneshloo, Y., Tavaf, N. and Fox, E.A., 2020. Natural language processing advancements by deep learning: A survey. arXiv preprint arXiv:2003.01200.

[25] Weng, O., 2021. Neural network quantization for efficient inference: A survey. arXiv preprint arXiv:2112.06126.

[26] Wilson, M., Wellington, B., Merrick, A. and Huxley, I., 2023. A recommendation model based on deep feature representation and multi-head self-attention mechanism.

[27] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E. and Zheng, R., 2023. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864.

[28] Xie, Y., Zhou, T., Mao, Y. and Chen, W., 2020. Conditional self-attention for query-based summarization. arXiv preprint arXiv:2002.07338.

[29] Zhao, Y., Wang, D., Xu, B. and Zhang, T., 2020. Monaural speech dereverberation using temporal convolutional networks with self attention. IEEE/ACM transactions on audio, speech, and language processing, 28, pp.1598-1607.

[30] Zulqarnain, M., Abd Ishak, S., Ghazali, R., Nawi, N.M., Aamir, M. and Hassim, Y.M.M., 2020. An improved deep learning approach based on variant two-state gated recurrent unit and word embeddings for sentiment classification. International Journal of Advanced Computer Science and Applications, 11(1).