Journal of Information Systems Engineering and Management
2020, 5(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The Evolution and Impact of Long Short-Term Memory Networks: A Comprehensive Review

Pratap Singh Barth[1], Dhanroop Mal Nagar[2]
[1]Assistant Professor, CSE Dept., Engineering College Bikaner
Pratapcharan1985@gmail.com
[2]Assistant Professor, IT Dept., Engineering College Bikaner
dhanroopmalnagar@gmail.com

ARTICLE INFO

Received: 20 Jan 2020

Revised: 01 Mar 2020

Accepted: 25 Mar 2020

ABSTRACT

A comprehensive review of Long Short-Term Memory networks is presented in this research work, tracing their origins as a solution to the persistent challenges of vanishing and exploding gradients inherent in traditional recurrent neural networks. This paper systematically examines the foundational LSTM architecture, highlighting the innovative role of its specialized memory cells and multiplicative gating mechanisms—input, forget, and output gates—in maintaining stable error flow and enabling the learning of long-term dependencies over extended sequences. Architectural enhancements and variants that have further optimized LSTM performance and versatility are explored (Greff et al., 2016; Minar & Naher, 2018). Furthermore, the diverse and profound impact of LSTMs across a multitude of domains is discussed, including significant advancements in natural language processing, speech recognition, time series prediction, and computer vision. The advantages of LSTMs in handling complex sequential data are underscored. This review consolidates the understanding of how LSTMs have fundamentally shaped the landscape of deep learning for sequential data, establishing them as a cornerstone architecture despite the emergence of newer models.

Keywords: Long Short-Term Memory, LSTM

1. Introduction

Recurrent neural networks have emerged as a fundamental architecture for processing sequential data, yet their application has historically been constrained by inherent limitations in learning long-term dependencies. To address these critical shortcomings, the Long Short-Term Memory network was introduced as a specialized recurrent architecture capable of bridging time lags in excess of 1000 discrete-time steps through mechanisms that enforce constant error flow. In this research work, the subsequent sections systematically examine the architectural evolution of LSTMs, their theoretical underpinnings, and the diverse array of domains where they have successfully advanced the state-of-the-art (Bianchi et al., 2017; Greff et al., 2016). The proliferation of LSTM variants across different domains necessitates a comprehensive analysis of their architectural components and operational efficacy. Through a detailed investigation of established LSTM variants, this research work seeks to evaluate the relative importance of specific computational elements, such as the forget gate and output activation function, in determining overall performance across speech recognition, handwriting recognition, and polyphonic music modeling tasks (Greff et al., 2016).

Specifically, this paper conducts a comparative evaluation of the vanilla LSTM architecture against eight distinct variants to isolate the effects of individual modifications, utilizing random search for hyperparameter optimization and ANOVA to assess the significance of specific architectural changes. Furthermore, this research work addresses the persistent challenge of vanishing and exploding gradients in standard recurrent neural networks, where the influence of input vectors on future decisions decreases or increases exponentially as the sequence progresses, thereby hindering effective backpropagation.

Journal of Information Systems Engineering and Management
2020, 5(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

- **Motivation for This Research Work**

The proliferation of LSTM variants across domains necessitates a rigorous comparative analysis to identify which architectural components genuinely drive performance improvements, given that prior research has demonstrated LSTMs significantly outperform traditional recurrent networks on tasks requiring deep memory by mitigating the vanishing gradient problem through specialized memory cells with gated access (Greff et al., 2016; Sun et al., 2009). In this research work, empirical evaluation is conducted on over ten thousand distinct recurrent network architectures to determine whether the LSTM design represents an optimal configuration or if superior structures exist, thereby addressing the ambiguity regarding the significance of individual architectural components in performance outcomes (Józefowicz et al., 2015). This analysis specifically aims to establish the extent to which specific architectural modifications, such as the addition of peephole connections or the elimination of input gates, influence computational efficiency and accuracy across diverse applications like music modeling, handwriting recognition, and speech processing.

- **Scope and Objectives of This Research Work**

In this research work, the primary objectives encompass a systematic dissection of the various LSTM architectural enhancements to determine their respective contributions to model efficacy, specifically evaluating the trade-offs between computational complexity and performance gains through the removal of non-essential components like peephole connections and input gates. By investigating these architectural simplifications, this research work aims to isolate the specific elements that contribute most significantly to the network's ability to regulate long-range dependencies and successfully train on complex sequential tasks (Lu & Salem, 2017). Additionally, this research work outlines the organization of the remaining sections, beginning with a foundational overview of recurrent neural networks and their inherent limitations, followed by a detailed examination of the genesis and architectural evolution of LSTM systems, and concluding with a discussion of their diverse applications and future research directions.

2. Recurrent Neural Networks: A Foundational Overview

Recurrent Neural Networks form the foundational architecture for processing sequential data by utilizing feedback loops to maintain a hidden state that encodes information about past inputs, allowing these networks to exhibit dynamic temporal behavior suitable for tasks such as time series prediction and language modeling (Brigham et al., 2018). The fundamental architecture of standard RNNs operates by processing input sequences through a single hidden layer, where the current state is determined by the current input and the previous hidden state, creating a directed cycle that theoretically allows information to persist indefinitely (Salem, 2018). This theoretical capability is predicated on the efficient application of the backpropagation through time algorithm, which calculates gradients by unrolling the network across the temporal dimension, yet practical implementation reveals that the repeated multiplication of gradients during this process significantly jeopardizes training stability.

Consequently, as the network attempts to learn dependencies over extended time intervals, the gradients either tend toward zero, known as the vanishing gradient problem, or grow exponentially, known as the exploding gradient problem, ultimately preventing the network from capturing long-range temporal correlations essential for complex sequence modeling . The vanishing gradient problem specifically arises due to the repeated multiplication of small gradients during backpropagation through time (Tora et al., 2017), while the exploding gradient problem occurs when these multiplications cause the gradient values to grow exponentially, leading to training instability (Pichotta & Mooney, 2016).

While gradient clipping techniques can be employed to address the issue of exploding gradients by capping the maximum value of gradients during updates, the vanishing gradient phenomenon poses a more fundamental challenge as it prevents error signals from propagating effectively to earlier time steps.

Journal of Information Systems Engineering and Management
2020, 5(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- **Challenges with Traditional RNNs**

Despite their representational power and ability to learn general sequential relationships, traditional RNNs did not achieve widespread adoption initially because they were notoriously difficult to train effectively (Martens & Sutskever, 2012). One primary obstacle contributing to this difficulty is the vanishing gradient phenomenon, where gradients exponentially decay to zero during backpropagation through time due to the recursive multiplication of weight matrices with eigenvalues less than one, severing the connection between distant time steps and preventing the learning of long-term dependencies.

A. A compounding factor is that recurrent weights are shared across all time steps, making this vanishing effect even more severe when deep networks are unrolled for training (Kolbæk, 2018).

B. Furthermore, the exploding gradient problem represents the inverse challenge where the repeated multiplication of gradients results in exponentially large values that exceed numerical stability limits, causing drastic parameter updates that derail the learning process during training when the gradient norm is not constrained .

C. The occurrence of these large gradients often forces the network into a chaotic regime where its computational capability is significantly hindered, necessitating specific interventions such as clipping the gradient norm to stable levels.

D. A further analytical perspective suggests that models with larger recurrent depths exacerbate these gradient-related issues, since they possess more nonlinearities and the gradients are consequently more likely to explode or vanish (Bianchi et al., 2017).

Empirical studies have demonstrated that if the product of these gradients is smaller than one, the gradients over time converge exponentially to zero, whereas if their product exceeds one, the gradients become progressively larger and eventually explode. This dual failure mode fundamentally necessitates the development of specialized architectures that can maintain a stable error flow across long temporal horizons without succumbing to these optimization instability issues.

- **The Vanishing and Exploding Gradient Problem**

The vanishing and exploding gradient problems represent two fundamental obstacles in training Recurrent Neural Networks, emerging during backpropagation through time when the product of Jacobian matrices either shrinks exponentially toward zero or expands toward infinity, thereby preventing the model from learning correlations between temporally distant events (Bengio et al., 2013). Mathematically, this instability arises because the gradient calculation involves a product of Jacobian matrices over the time steps, where the spectral radius of the recurrent weight matrix dictates whether this product shrinks or grows exponentially.

When the spectral radius is less than one, the norm of the gradient decays exponentially, a phenomenon referred to as the vanishing gradient problem, which effectively prevents the network from retaining information over long sequences and consequently fails to learn long-term dependencies. Conversely, if the spectral radius exceeds one, the norms of these Jacobians grow larger than one, leading to exploding gradients where the accumulated error signal expands uncontrollably during the backward pass. To mitigate the instability caused by exploding gradients, gradient clipping strategies have been proposed to enforce an upper threshold on the gradient norm, thereby preventing the parameters from taking erratic steps that would otherwise lead to numerical overflow and divergence.

3. The Genesis of Long Short-Term Memory Networks

This research work explores how Long Short-Term Memory networks were conceptualized as a direct response to the instability inherent in standard recurrent neural network training, specifically designed to overcome the limitations of vanishing and exploding gradients that prevent the effective learning of long-term temporal dependencies. The fundamental innovation introduced by LSTM lies in its use of specialized memory blocks and gate mechanisms that enforce constant error flow through time, thereby bypassing the vanishing and exploding

Journal of Information Systems Engineering and Management
2020, 5(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

gradient issues that plagued earlier architectures. Unlike standard RNNs, which rely on the careful tuning of gradient norms to prevent divergence or decay, the LSTM architecture introduces a specialized internal state, often referred to as the cell state, that is protected by multiplicative gates designed to regulate the flow of information and preserve the error signal over extended sequences.

The LSTM cell achieves this preservation of information by utilizing an additive state update mechanism, where instead of directly overwriting the previous state through matrix multiplication, the network computes a proposed change and adds it to the existing cell state, an arithmetic operation that allows the gradient to flow backward without exponential decay (Vaněk et al., 2018). This preservation mechanism, often conceptualized as a constant error carousel, ensures that the error signal remains accessible for learning dependencies over time lags spanning more than one thousand discrete time steps (Levy et al., 2018; Staudemeyer, 2015). This architecture fundamentally relies on a linearly self-connected memory unit with a near-unity weight, which acts as a shortcut path for the gradient to propagate without the degradation typically caused by repeated nonlinear transformations.

- **Early Developments and Key Innovations**

The inception of Long Short-Term Memory networks marked a pivotal advancement in recurrent neural network research, rooted in the early observation that learning to store information over extended time intervals was severely hindered by insufficient and decaying error backflow during training. To address these fundamental limitations, the LSTM architecture introduces a novel, gradient-based method that enforces constant error flow through specialized units, enabling the network to bridge minimal time lags in excess of 1000 discrete-time steps without succumbing to the vanishing gradient problem . This architectural breakthrough replaces the standard hidden units found in traditional neural networks with memory blocks, which function similarly to memory cells in a digital computer by incorporating one or more self-connected memory cells alongside multiplicative units that regulate information flow (Wang et al., 2015; Zazo et al., 2016). These multiplicative units, specifically designed as input and output gates, control the extent to which information enters or exits the memory cell, thereby preventing the currently irrelevant data from interfering with the storage and retrieval processes over long temporal sequences.

- **Gate Mechanisms: Input, Forget, and Output Gates**

The core functionality of the LSTM cell is governed by three distinct multiplicative gate units—the input, forget, and output gates—which operate analogously to read, write, and reset operations in a digital system to precisely regulate the flow of information into, within, and out of the memory cell .

1) The input gate determines whether new information should be stored in the cell, the forget gate manages the retention or removal of existing contents, and the output gate controls the transmission of the cell state to the rest of the network.
2) The input gate specifically modulates how much of the current input is utilized to update the cell state (Sharma & Tomer, 2018).
3) The forget gate provides a mechanism to reset the cell contents, determining what proportion of the previous cell state should be retained or discarded before the update occurs.
4) The output gate governs the exposure of the internal memory state by determining the extent to which the cell state influences the hidden state and the subsequent output of the network (Shen et al., 2018).
5) Peephole connections were introduced as a significant architectural refinement that allows the gates to inspect the internal cell state directly, enabling the model to learn precise timing of the outputs and enhance the control over memory access.

4. Evolution and Enhancements of LSTM Architectures

In this research work, the subsequent sections explore the progressive architectural modifications that have refined the LSTM framework, building upon the foundational gate mechanisms to address specific limitations such as computational redundancy and the need for more sophisticated temporal context processing. Peephole connections represent a significant architectural enhancement introduced to the standard LSTM framework, wherein weighted connections are added directly from the cell state to each of the gate units, allowing the gates to inspect the internal state of the memory cell before computing their activation values (Sarraf, 2018).

- **Bidirectional LSTMs**

Bidirectional LSTMs extend the temporal processing capabilities of the standard architecture by stacking a backward LSTM layer on top of a forward LSTM layer, allowing the model to integrate information from both past and future time frames to improve sequence labeling tasks (Tan & Lim, 2018). This architectural configuration enables the simultaneous access to comprehensive contextual dependencies, which is particularly beneficial for scenarios where the interpretation of a specific element relies not only on preceding inputs but also on subsequent information within the sequence. BiLSTMs achieve this bidirectional context processing by utilizing two separate hidden layers that traverse the input sequence in opposite directions and subsequently combine their internal states, typically through concatenation, to form the final representation at each time step (Yang et al., 2017).

5. Applications and Impact of LSTMs Across Domains

The versatility of Long Short-Term Memory networks has led to their widespread adoption across a diverse array of fields, demonstrating exceptional efficacy in domains that require the modeling of complex temporal dependencies. In the domain of Natural Language Processing, LSTM architectures have become instrumental in tasks such as language translation, sentiment analysis, and text generation, where the ability to capture long-range dependencies is critical for maintaining coherence and understanding context over extended sequences of text (Hajj et al., 2018). Specifically, these networks have been successfully employed to resolve named entity recognition ambiguities and improve the accuracy of semantic representation tasks by leveraging bidirectional architectures to aggregate both preceding and succeeding context.

- **Natural Language Processing**

In this research work, the application of LSTM networks in natural language processing is further characterized by their demonstrated superiority over convolutional and plain recurrent approaches in capturing linguistic dependencies that span longer distances within sequences . Furthermore, LSTMs have established themselves as a fundamental component in statistical language modeling and machine translation, serving as the state-of-the-art approach for many sequence processing problems by effectively modulating between short-term and long-term dependencies.

- **Speech Recognition**

The efficacy of LSTM architectures in the domain of speech recognition has been extensively validated, particularly through the implementation of bidirectional models that leverage both past and future acoustic context to significantly enhance the accuracy of large vocabulary continuous speech recognition systems. This capability proves crucial for phoneme classification and distinguishing acoustic features that require knowledge of subsequent temporal frames to resolve ambiguities in the audio signal, often outperforming standard recurrent architectures in handling long-term temporal dependencies (Khandelwal et al., 2016).

- **Time Series Prediction**

Beyond speech and language, LSTM architectures have demonstrated remarkable proficiency in forecasting sequential data across various domains, including financial market analysis, weather modeling, and anomaly detection, by leveraging their inherent capacity to discern and retain temporal patterns over extended durations. Specifically, their ability to learn complex nonlinear feature interactions allows them to model normal behavior in multivariate data streams without the need for specific domain knowledge or manual time-window specification. In this research work, the utility of LSTM networks in computer vision is examined through their application to spatio-temporal feature extraction and video sequence classification tasks, where they effectively model the temporal evolution of visual data (Baccouche, 2013; Ruiz et al., 2018).

Journal of Information Systems Engineering and Management
2020, 5(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

- **Computer Vision**

In the realm of action recognition, hybrid architectures combining convolutional neural networks with LSTM layers have proven particularly effective for classifying human activities from video streams, as these models process sequences of frames to interpret events that unfold over time (Srinivas et al., 2016; Zhang et al., 2017). Beyond the scope of visual action recognition, these architectures have also been successfully adapted for video analysis and image completion tasks by explicitly capturing temporal correlations within sequential data .

- **Other Emerging Applications**

In this research work, additional promising applications of LSTM networks are identified in fields such as protein structure prediction and software reliability assessment, where the sequential nature of the data necessitates robust modeling of dependencies (Kumar & Mo, 2018; Mahata et al., 2018). For instance, in biological sequence analysis, these networks are employed to predict protein structures by interpreting amino acid sequences as ordered inputs, while software engineering utilizes them to model execution traces for fault prediction based on the temporal ordering of system events. In this research work, the versatility of LSTM networks is further illustrated by their successful deployment in outlier detection and predictive modeling, where the probability distribution of observed features is estimated across time steps to identify anomalies and forecast future states in complex systems.

6. **Key Contributions and Limitations of LSTMs**

In this research work, the fundamental significance of LSTM architecture is attributed to its capacity to mitigate the recurrent neural network issues of vanishing and exploding gradients through specialized gating mechanisms and constant error carousels. This architectural innovation enables the gradient to flow unchanged across many time steps without vanishing or exploding, allowing LSTM units to bridge significant time lags and remember inputs up to 1000 time steps in the past. This capability to bridge substantial temporal gaps is particularly facilitated by the constant error carousel, which ensures that the error signal remains constant during backpropagation through time.

- **Advantages in Handling Long-Term Dependencies**

During research work, it is observed that the incorporation of memory cells allows the network to preserve information over extended sequences, thereby enabling the effective modeling of long-range temporal dependencies that traditional recurrent neural networks fail to capture (Negin, 2018; Ren et al., 2018). This extended retention capability is further augmented by the gate structure, which regulates information flow to prevent backpropagated errors from decaying or growing uncontrollably, thereby maintaining a stable learning signal across deep temporal layers. Furthermore, the architecture's capacity to selectively retain or discard information via input, forget, and output gates allows for precise control over the network's internal state, enhancing performance in sequence prediction tasks such as machine translation and speech recognition. Despite these strengths, the increased parameter count resulting from the additional gate mechanisms and cell states introduces substantial computational overhead and extended training durations compared to simpler recurrent models.

- **Computational Complexity and Training Challenges**

The implementation of multiple gating units per cell significantly increases the number of trainable parameters compared to standard recurrent networks, necessitating greater computational resources and longer training times to achieve convergence effectively. This complexity is compounded by the difficulty of optimizing the intricate interactions within the memory cell, often requiring advanced regularization techniques and careful hyperparameter tuning to avoid overfitting during the learning process (Lu & Salem, 2017; Sun et al., 2009). Furthermore, the objective function landscape is characterized by high dimensionality and numerous saddle points, which complicates the optimization process and makes it difficult to find global minima compared to shallower architectures. During research work, another significant challenge identified is the lack of transparency in the decision-making process, as the intricate interactions between memory cells and gating units operate as a "black box" that obscures the rationale behind specific predictions. This complexity arises because the distributed representations within the memory cells and the nonlinear transformations performed by the gates do not easily map to human-understandable features, making it arduous to trace how the network weights specific inputs to its final output at any given time step.

- **Interpretability of LSTM Models**

During research work, the interpretability of these networks is frequently compromised by the internal multiplicative logic gates and memory state interactions, which function as a complex "black box" that obscures the relationship between input sequences and the resulting predictions (Mohamad et al., 2018). Consequently, researchers face significant difficulties when attempting to debug these models or extract actionable insights, as the learned features often lack semantic meaning and the internal state dynamics are highly sensitive to initial conditions and hyperparameters .

Furthermore, the nonlinearities introduced by these intricate gated architectures inherently make the interpretation of predictions more challenging compared to simpler linear models, where the direct mapping between inputs and outputs is easier to visualize and understand.

- **Theoretical Understanding and Generalization Bounds**

During research work, it is found that the theoretical understanding of LSTM networks remains an open challenge, particularly regarding the derivation of precise generalization bounds that can reliably predict model performance on unseen data given the complexity of their sequential dynamics (Greff et al., 2016). Current analysis suggests that establishing robust theoretical frameworks for understanding why LSTMs generalize effectively is essential for advancing the field beyond empirical success alone. Specifically, the need for formal proofs regarding the optimization dynamics of gated recurrent units with non-convex loss functions is critical to ensuring that future architectural modifications are grounded in mathematical rigor rather than solely on performance benchmarks. Investigating the behavior of hidden state dynamics and memory cells is therefore crucial to design better models and increase trust in these systems, particularly when deploying them in safety-critical domains. During research work, it is noted that bridging the gap between high predictive accuracy and interpretable internal representations constitutes a vital frontier for ensuring the reliability and trustworthiness of LSTM-based systems in practical applications.

7. Conclusion

This paper has provided a comprehensive review of Long Short-Term Memory networks, tracing their historical evolution from foundational recurrent neural network architectures to the advanced gated variants that dominate modern sequence learning tasks (Cashman et al., 2018; Karpathy et al., 2015). The analysis has highlighted the architectural components that distinguish LSTMs, such as the gating mechanisms and memory cells, which are pivotal for capturing long-term dependencies in sequential data like speech and text.

These structural innovations effectively mitigate the vanishing gradient problem that historically hindered traditional RNNs, thereby preserving error signals across extended sequences. By systematically evaluating the strengths and limitations of various LSTM architectures, it is evident that these models offer robust solutions for handling long temporal dependencies, though they are not without trade-offs in computational efficiency and interpretability.

References

[1] Baccouche, M. (2013). Neural learning of spatio-temporal features for automatic video sequence classification. *HAL (Le Centre Pour La Communication Scientifique Directe)*. https://tel.archives-ouvertes.fr/tel-00871107

[2] Bengio, Y., Boulanger-Lewandowski, N., & Pascanu, R. (2013). Advances in optimizing recurrent networks. *IEEE International Conference on Acoustics Speech and Signal Processing*, 8624. https://doi.org/10.1109/icassp.2013.6639349

[3] Bianchi, F. M., Maiorino, E., Kampffmeyer, M., Rizzi, A., & Jenssen, R. (2017). An overview and comparative analysis of Recurrent Neural Networks for Short Term Load Forecasting. *arXiv (Cornell University)*. http://arxiv.org/abs/1705.04378

[4] Brigham, K., Gupta, S., & Brigham, J. C. (2018). Predicting responses to mechanical ventilation for preterm infants with acute respiratory illness using artificial neural networks. *International Journal for Numerical Methods in Biomedical Engineering*, *34*(8). https://doi.org/10.1002/cnm.3094

[5] Cashman, D., Patterson, G., Mosca, A., Watts, N., Robinson, S., & Chang, R. (2018). RNNbow: Visualizing Learning Via Backpropagation Gradients in RNNs. *IEEE Computer Graphics and Applications*, *38*(6), 39. https://doi.org/10.1109/mcg.2018.2878902

[6] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222. https://doi.org/10.1109/tnnls.2016.2582924

[7] Hajj, H. A., Lamard, M., Conze, P.-H., Cochener, B., & Quellec, G. (2018). Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. *Medical Image Analysis*, *47*, 203. https://doi.org/10.1016/j.media.2018.05.001

[8] Józefowicz, R., Zaremba, W., & Sutskever, I. (2015). An Empirical Exploration of Recurrent Network Architectures. *International Conference on Machine Learning*, 2342. http://proceedings.mlr.press/v37/jozefowicz15.pdf

[9] Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and Understanding Recurrent Networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1506.02078

[10] Khandelwal, S., Lecouteux, B., & Besacier, L. (2016). COMPARING GRU AND LSTM FOR AUTOMATIC SPEECH RECOGNITION. *HAL (Le Centre Pour La Communication Scientifique Directe)*. https://hal.archives-ouvertes.fr/hal-01633254

[11] Kolbæk, M. (2018). *Single-Microphone Speech Enhancement and Separation Using Deep Learning*. https://doi.org/10.54337/aau300036831

[12] Kumar, V., & Mo, L. (2018). Deep Learning as a Frontier of Machine Learning: A Review [Review of *Deep Learning as a Frontier of Machine Learning: A Review*]. *International Journal of Computer Applications*, *182*(1), 22. https://doi.org/10.5120/ijca2018917433

[13] Levy, O., Lee, K., FitzGerald, N., & Zettlemoyer, L. (2018). Long Short-Term Memory as a Dynamically Computed Element-wise Weighted\n Sum. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1805.03716

[14] Lu, Y., & Salem, F. M. (2017). Simplified gating in long short-term memory (LSTM) recurrent neural networks. *2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS)*, *3*, 1601. https://doi.org/10.1109/mwscas.2017.8053244

[15] Mahata, D., Friedrichs, J., Hitkul, & Shah, R. R. (2018). #phramacovigilance - Exploring Deep Learning Techniques for Identifying\n Mentions of Medication Intake from Twitter. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1805.06375

[16] Martens, J., & Sutskever, I. (2012). Training Deep and Recurrent Networks with Hessian-Free Optimization. In *Lecture notes in computer science* (p. 479). Springer Science+Business Media. https://doi.org/10.1007/978-3-642-35289-8_27

[17] Minar, M. R., & Naher, J. (2018). Recent Advances in Deep Learning: An Overview. *arXiv (Cornell University)*. https://doi.org/10.13140/rg.2.2.24831.10403

[18] Mohamad, A. B., Al-Marghilani, A., & Aref, A. (2018). Automatic Detection Technique for Speech Recognition based on Neural Networks Inter-Disciplinary. *International Journal of Advanced Computer Science and Applications*, *9*(3). https://doi.org/10.14569/ijacsa.2018.090326

[19] Negin, F. (2018). Toward unsupervised human activity and gesture recognition in videos. *HAL (Le Centre Pour La Communication Scientifique Directe)*. https://hal.inria.fr/tel-01947341

[20] Pichotta, K., & Mooney, R. J. (2016). Learning Statistical Scripts with LSTM Recurrent Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1). https://doi.org/10.1609/aaai.v30i1.10347

[21] Ren, J., Karagoz, K., Gatza, M. L., Singer, E. A., Sadimin, E., Foran, D. J., & Qi, X. (2018). Recurrence analysis on prostate cancer patients with Gleason score 7 using integrated histopathology whole-slide images and genomic data through deep neural networks. *Journal of Medical Imaging*, *5*(4), 1. https://doi.org/10.1117/1.jmi.5.4.047501

[22] Ruiz, L. G. B., Capel, M. I., & Pegalajar, M. C. (2018). Parallel memetic algorithm for training recurrent neural networks for the energy efficiency problem. *Applied Soft Computing*, *76*, 356. https://doi.org/10.1016/j.asoc.2018.12.028

[23] Salem, F. M. (2018). SLIM LSTMs. *arXiv (Cornell University)*. https://arxiv.org/pdf/1812.11391

[24] Sarraf, S. (2018). French Word Recognition through a Quick Survey on Recurrent Neural\n Networks Using Long-Short Term Memory RNN-LSTM. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1804.03683

[25] Sharma, M., & Tomer, M. S. (2018). Predictive Analysis of RFID Supply Chain Path Using Long Short Term Memory (LSTM): Recurrent Neural Networks. *International Journal of Wireless and Microwave Technologies*, *8*(4), 66. https://doi.org/10.5815/ijwmt.2018.04.05

[26] Shen, X., Tian, X., Xing, J., Rui, Y., & Tao, D. (2018). Sequence-to-Sequence Learning via Shared Latent Representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11837

[27] Srinivas, S., Sarvadevabhatla, R. K., Mopuri, K. R., Prabhu, N., Kruthiventi, S. S. S., & Babu, R. V. (2016). A Taxonomy of Deep Convolutional Neural Nets for Computer Vision. *Frontiers in Robotics and AI*, *2*. https://doi.org/10.3389/frobt.2015.00036

[28] Staudemeyer, R. C. (2015). Applying long short-term memory recurrent neural networks to intrusion detection. *South African Computer Journal*, *56*. https://doi.org/10.18489/sacj.v56i1.248

[29] Sun, Y., Wierstra, D., Schaul, T., & Schmidhuber, J. (2009). *Stochastic search using the natural gradient*. 1161. https://doi.org/10.1145/1553374.1553522

[30] Tan, K.-H., & Lim, B. P. (2018). The artificial intelligence renaissance: deep learning and the road to human-Level machine intelligence. *APSIPA Transactions on Signal and Information Processing*, *7*(1). https://doi.org/10.1017/atsip.2018.6

[31] Tora, M. R., Chen, J., & Little, J. J. (2017). *Classification of Puck Possession Events in Ice Hockey*. 147. https://doi.org/10.1109/cvprw.2017.24

[32] Vaněk, J., Michálek, J., & Psutka, J. (2018). Recurrent DNNs and Its Ensembles on the TIMIT Phone Recognition Task. In *Lecture notes in computer science* (p. 728). Springer Science+Business Media. https://doi.org/10.1007/978-3-319-99579-3_74

[33] Wang, X., Liu, Y., Sun, C., Wang, B., & Wang, X. (2015). *Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory*. https://doi.org/10.3115/v1/p15-1130

[34] Yang, X., Yu, L., Wu, L., Wang, Y., Ni, D., Qin, J., & Heng, P. (2017). Fine-Grained Recurrent Neural Networks for Automatic Prostate Segmentation in Ultrasound Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10761

[35] Zazo, R., Lozano-Díez, A., Gónzalez-Domínguez, J., Toledano, D. T., & González-Rodríguez, J. (2016). Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks. *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0146917

[36] Zhang, Z., Cummins, N., & Schuller, B. W. (2017). Advanced Data Exploitation in Speech Analysis: An overview. *IEEE Signal Processing Magazine*, *34*(4), 107. https://doi.org/10.1109/msp.2017.2699358