Gated Recurrent Fusion in Long Short-Term Memory Fusion

Anita Venugopal[1], Aditi Sharma[2,*], Preetish Kakkar[3], Daya Nand[4], Arvind R. Yadav[5], Gaurav Kumar Ameta[6]

[1]Dhofar University, Sultanate of Oman

[2]Department of Computer Sc. and Engg, Symbosis Institute of Technology, Pune, India

Symbiosis International (Deemed) University, Pune, India

[3]IEEE Senior Member, USA

[4]University of Houston, Victoria, Texas, USA

[5]E&I Engineering Department, Institute of Technology, Nirma University, Ahmedabad, India

[6]Department of Computer Sc. and Engg, Parul Institute of Technology, Parul University, Vadodara, India

Emails: anita@du.edu.om; aditi.sharma@ieee.org; preetish.kakkar@gmail.com; nandD@uhv.edu; arvind.yadav.me@gmail.com; gauravameta1@gmail.com

Abstract

Fusion techniques on enhancing the efficiency of Long Short-Term Memory (LSTM) networks are dominating across a variety of domains. To handle sequential data while integrating from various sources is often challenging using LSTM techniques. Fusion methods that integrate different models enhances LSTM' ability to handle complex correlations in the data. This paper examines early, late and hybrid fusion techniques. The study provides fusion approaches to enhance LSTM networks to efficiently handle complex multimodal data across self-navigating models. The findings reveal that the hybrid fusion techniques outperform traditional methods in terms of accuracy and generalization of various tasks. This paper proposes the Gated Recurrent Fusion (GRF) approach to demonstrate its performance to handle multimodal and temporal models in a supervised recurrence. The findings report 10% enhancement in terms of precision rate.

Keywords: Fusion technique; LSTM, RNN; Scalability; Early fusion; Hybrid fusion; Multimodal data

1. Introduction

Long Short-Term Memory (LSTM) networks is one of the popular approaches used in analysing sequential data as it can handle vanishing gradient issues faced during recurrent neural networks [1] and can obtain distant range dependencies making them ideal for handling tasks such as natural language processing, time series, speech recognition [2]. The ability to hold data over extensive periods along with the capability to update information makes them adapt to complex temporal models. However, LSTMs face challenges while handling distinct sources involving varied information [3]. Traditional LSTMs face challenges to integrate and analyse information from different sources, resulting to low performance since it needs to integrate diverse data. Feature-modal discrepancy and data divergence cause significant challenges in integration of data, thereby limiting models' ability in the field of identifying patterns [4][5]. By integrating techniques such as early fusion, late fusion and hybrid fusion approaches fusion techniques can be enhanced and can be used to overcome the challenges.

This paper explores the advancements of hybrid fusion techniques for LSTM networks in the field of capturing data from diverse sources. This paper introduces a recurrent neural network approach known as Gated-Recurrent Fusion (GRF) approach to solve challenges. The network is trained to learn and predict for diverse datasets. This complex multimodal data can be used to solve challenges across diverse fields especially in fusion and prediction for temporal data.

DOI: https://doi.org/10.54216/FPA.190105
Received: November 10, 2024 Revised: January 12, 2025 Accepted: February 11, 2025

2. Literature Review

Hochreiter and Schmidhuber (1997) laid foundational work on LSTM networks, which marked a significant breakthrough in the field of sequential data analysis [1]. In their work, they enhanced traditional recurrent neural networks (RNNs) and designed LSTM networks to overrule vanishing gradient problem and could effectively design long-range dependencies in sequential data. LSTM networks are used widely across various fields such as natural language processing (NLP), time series and speech recognition. This popularity was mainly due to their remarkable ability to capture temporal patterns and preserve information during sequential analysis making it crucial while handling complex tasks. As per the studies, it faced challenges while working with heterogeneous data consisting of diverse sources [12]. The major concern noted was in the field of feature alignment while extracting features from different models. Studies find that the design fail to align temporal/sequential sources making integration difficult. Furthermore, difference in data model sources and data heterogeneity add further challenges and complexity in the fusion processes as the algorithms get harder to gather pattern from diverse data models.

Researchers have developed a series of fusion algorithms to improve the LSTM networks performance using hybrid techniques integrating data from multiple sources [4] [5]. Early fusion algorithms collect features from different sources at the input level whereas; modern fusion-hybrid algorithms collect information at higher level from different sources enabling network to learn independently from each model before integrating them. Hybrid approach provides a fusion of both early and modern fusion technologies providing a framework that harness information from diverse data models.

Recent advancements in fusion techniques have broadened significantly to enhance LSTM performance [6] [7]. Attention networks have gained popularity as it prioritises diverse data model and is able to work dynamically to map query and a set of key-value pairs, which are vectors, to an output. The output is evaluated as a weighted total of the values [11].

Sensors in electronic independent navigation are gaining popularity in modern world [21] [22] [23] [17] [24]. Self-navigating systems, navigator learning behaviour, can [12] [13] predict [28] [29] recognise [16], learn [30], segment [31] [32] and do much more. Much research has been conducted in the field of temporal and non-temporal algorithms using RGB-D [15] [16] [18] [30] [31].

In this paper, our experiments incorporate hybrid multimodal fusion learning, which uses recurrent supervised approach for predicting navigators' behaviours.

3. LSTM Architecture

Long Short-Term Memory (LSTM) networks mark a leap towards the progression of recurrent neural networks (RNNs) as it handles learning methods based on vanishing gradient models for long-memory persistence in sequential dataset [1].

The LSTM setup most used in literature features three gated states (input, forget, output) along with the hidden and candidate cell states. LSTM architecture consists of inter-connected cells, each projecting distinct gates that manages the data flow. As mentioned in literature [23], LSTM cell is composed of three gated states named as:

1. Input Gate: This unit controls the data flow to the cell and determines whether to retain or discard the parts of current and preceding cell state.

2. Forget Gate: This unit determines whether to hold or discard the data from the preceding cell state thus enabling the LSTM cell to specifically exclude extraneous data, which aids to address the vanishing gradient challenges.

3. Cell State: The cell state is LSTM cells memory that enables flow of data across time steps. Additive operations update the cells to contain important data over long successions.

4. Output Gate: This unit selects the data from the current cell and passes to the next step. It controls the flow of data to subsequent layers of the LSTM cell network.

LSTM networks can attain extensive range dependencies by incorporating these gates in sequence and supplements the vanishing gradient issues, which results in traditional RNNs. This feature of LSTMs architecture allows them to handle complex temporal patterns as it can preserve data over long time [2].

4. Methodology- Fusion Model ideas:

Temporal fusion LSTM techniques play a vital role in improving the performance of Long Short-Term Memory (LSTM) networks by integrating data from diverse datasets.

In this section, an overview of fusion ideas, including early fusion, late fusion, and hybrid fusion are stated. This section also explores LSTMs multimodal settings such as summing/concatenation, adaptive attention and gated mechanism network [5].

DOI: https://doi.org/10.54216/FPA.190105
Received: November 10, 2024 Revised: January 12, 2025 Accepted: February 11, 2025

A. Early Recurrent Fusion (ERF):

Early recurrent fusion multimodal concatenates and sums all sensor datasets of the LSTM network and feeds it to the input level [12] [19] [20]. Concatenation method results in blotting up of the cell while summation contracts the cell size by merging all sensor codes, which sometimes in certain scenarios may not provide accuracy. Hence tuning such architectures to how to hybrid/fuse data is must.

In early recurrent fusion, model is trained to learn from a standard description across diverse sensor datasets, which gathers complex relationships between different data sources. It is beneficial to handle tasks that are from diverse inputs.

However, there are challenges as early fusion can suffer from dimensionality, feature recognition while handling large complex multi-dimensional datasets from diverse datasets.

To tackle this issue, we have used a) delay or late recurrent fusion to input sensors through LSTM cells and b) gates are defined to gather the output of sensors to a fused cell.

B. Proposed Temporal Fusion:

In this section, the modifications are defined separately and the hybrid model combining the two called as late graded recurrent fusion model is generated.

1) Late Recurrent Fusion:

In this approach, each LSTM units are processed independently and then totalled at a different stage or by forming additional layers. Different LSTM units handle each sensor and distinct input, output, forget, states are calculated. In this architecture, each representation of the model learns independently before totalling. The weights and biases for each gate are distinct for each dataset model and exchanged later. LSTM receives its input from both the previous state and from the current step and fuse temporarily. All hidden cell states are totalled and sent to the next step.

This method allows processing of the individual modal, and thus can manage complex relationships in each dataset.

2) Early Graded Recurrent Fusion:

As per the gating approaches mentioned in the LSTM [23] and graded recurrent fusion [24], for a group of sensors, gates are defined to manage the encoding of the sensors and the last sensor.

Dimensions of the sensors enclosed are standardised using non-linear algorithm and gates are calculated. The temporal model is carried out after the product of each gate and its sensor is obtained and totalled to obtain a fused state. Thus, learning takes place, and user can verify the match.

3) Late Graded Recurrent State Fusion:

This approach combines features such as individual access to sensors memory of late recurrent and learning ability of grated recurrent fusion approaches. Sum of final and hidden cells in the outputs are calculated.

Fusion gates are found for all sensors and gates are used to control individual control the encoding sent to each LSTM sensor cells.

5. Experimental Evaluation

1) Dataset

This paper aims to classify the behavioural aspects of a navigator's task. HDD dataset is used to learn the navigator practices [12]. Video-data comprising of 200 sessions is mapped with a label for each frame. 15 classes consisting of varied navigator behaviour is recorded. Each class records navigator road crossing, lane change and the like activities. For training 170 sessions and for testing 30 sessions are defined. Data such as speed, braking, acceleration, steering, wheel positions are considered. Table 1 given below shows the various stages of image processing.

Table 1: Image processing and stages of feature extraction

Aspect	**Description**	**Dimensions**
Input Images	Dimensions	512 × 512 × 3
Image Representation	Conv layer extracted, pre-trained on ImageNet	16 × 16 × 1024

DOI: https://doi.org/10.54216/FPA.190105
Received: November 10, 2024 Revised: January 12, 2025 Accepted: February 11, 2025

Dimensional Reduction	Features convolved to 1×1, reduced dimensions	16 × 16 × 50
Flattened Features	Flattened value	1 × 4000
Raw Sensor Signals	Passed through connected layers and transformed	1 × 10 → 1 × 50
Final Feature Vector	Combined features	1 × 50

2) Results

Input data is CAN signals in unclipped form, output is strategic navigators' behaviour. Table 2 depicts results of training behaviour prediction.

Table 2: Navigator behaviour prediction on HDD dataset

Aspect	**Details**
Input Data	Unclipped video pattern, CAN signals
Output	Strategic driver behavior label on individual frame
Evaluation Metric	Compute Mean Average Precision (mAP) [12]
Optimizer	To learn network, Adam optimizer is used
Sequence Length	120 video frames
Batch Size	32
Training Method	Truncated back-propagation
Training Duration	60 epochs
Learning Rate	4×10^{-3}

Table 3: Comparative analysis and observations

Architecture	**Description**	**Performance**	**Key Observations**
Non-Fusion	1. Conducted experiments on: • CAN signal • Image sensors 2. Embeddings sent to LSTM 3. Hidden size = 2000 4. Output sent to fully connected layer 5. Output Image classification takes place 5. Output is classified into 12 classes.	CAN Signal: Better performance for different turns (left, right, and U)	-TCN outperforms LSTM for both sensors
		Image: -Better performance for lane change and different crossing and passing (intersection/crosswalk)	-Successful sensor fusion should be executed to outperform results from distinct sensors

DOI: https://doi.org/10.54216/FPA.190105
Received: November 10, 2024 Revised: January 12, 2025 Accepted: February 11, 2025

Early Fusion LSTM	Input level- using concatenation sensor embeddings are combined (Early-Add)	Early-Concat: -Outperforms Early-Add with mAP	-Early-Add suffers due to differences in sensor ranges, -Leads to unreliable fused encodings
		-LSTM discard noisy data	
Late Fusion LSTM	1. LSTM cells process: • CAN • Image embeddings 2. Fusion occurs after processing LSTM cells	Late-Concat: -Works on fully connected layer -Operates on 2000×2 vector	-Late-Add concentrates on temporal part of distinct sensor -Outperforms other fusion approaches
		Late-Add: -Works on fully connected layer -Operates on a 2000 vector -Outperforms Early Fusion	-Does not pass out weights/hidden states between modal levels

We have followed LLL architecture stated in [11] to first look then listen and finally learn the architecture. Auxiliary losses are added for both the models are total is added to the result. Standard LSTM is replaced with Late Recurrent Summation, Early Gated Recurrent Fusion, Late Gated Recurrent State Fusion models. Table 4 depicts performance analysis across different architectures and benefits.

Table 4: Performance boost across different models and key benefits

Fusion Module	Description	Improvement (mAP increase)	Key Benefit
EGRF	Replaces standard LSTM	by 7% over LSTM	Benefits from: -added flexibility -distinct class label optimization
LRS	Replaces standard LSTM	by 7% over LSTM	Excels in leveraging: -sensor-specific temporal information
LGRF	Combines benefits of both LRS and EGRF	by 10% over LSTM	Higher overall performance and flexibility of: -EGRF -Temporal optimization of LRS

DOI: https://doi.org/10.54216/FPA.190105
Received: November 10, 2024 Revised: January 12, 2025 Accepted: February 11, 2025

Figure 1 shows that our that our fusion network model performance graph displays an increase of 7% for EGRF and LRS models when compared to traditional LSTM architecture and the fusion model LRS and LGRF show mAP increase by 10%.

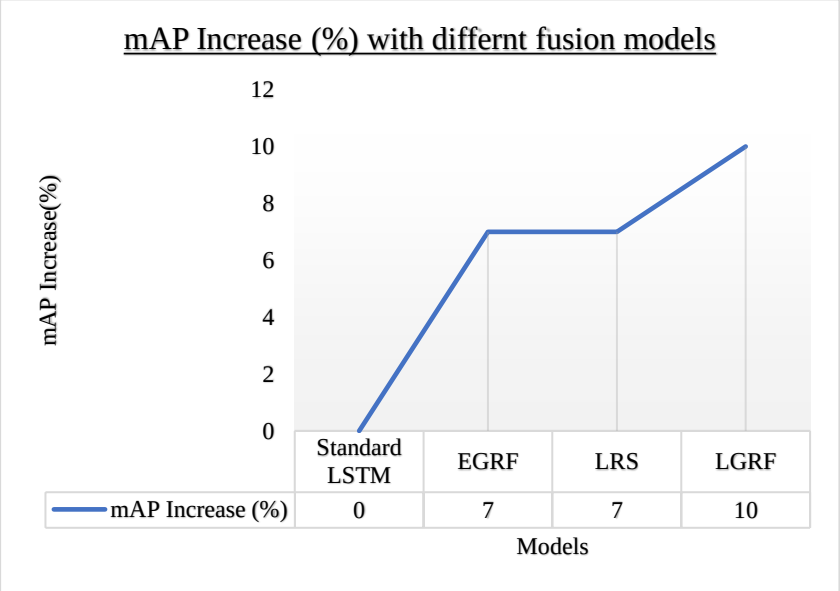

Figure 1. Performance analysis mAP % increase

6. Conclusion

In conclusion, this paper presents a novel method using gated recurrent fusion. This network learns from diverse range of temporal datasets. Optimal fusion techniques modulate the performance of distinct sensor data at each level. Gating functions used in the experiment has benefited in combining data from diverse datasets as these functions control the flow of sensor at different stages of the experiment. Gating function based recurrent enhances traditional LSTM approach as it removes the use of separate networks while performing preprocessing, learning and modelling navigator behaviour. The designed network system can learn sensor fusion, dataset and behavioural aspects for each block. This paper offers more systematic approach to sensor fusion and behavioural model.

7. Future Directions

Designing LSTM models using fusion techniques is gaining popularity in the field of AI learning. For future directions, we aim to boost the computational effectiveness of graded recurrent fusion unit in terms scalability and deployment to generate high throughput systems using temporal convolutional networks. This furthermore can be enhanced to design a system to predict actions from diverse datasets by fusing attention algorithm [14] with graded recurrent fusion unit.

References

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 3126–3141, May 2022.

[3] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 6645–6649.

[4] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3128–3137.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv: 1409.0473*, 2014.

[7] Z. Wang, D. Guo, C. Wu, and M. Zhang, "Dynamic multi-modal fusion for transformer-based sentiment analysis," *IEEE Access*, 2022.

DOI: https://doi.org/10.54216/FPA.190105
Received: November 10, 2024 Revised: January 12, 2025 Accepted: February 11, 2025

[8] Y. Chen, J. Li, H. Xiao, X. Jin, Z. Zhou, and S. Zhang, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *IEEE Transactions on Affective Computing*, 2020.

[9] X. Chen and Y. Li, "Fusion methods in neural networks: A comparative study of LSTM model enhancements," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 2223–2238, 2022.

[10] A. Martinez and S. Gupta, "Advances in multimodal fusion techniques for LSTM networks," *Int. J. Artif. Intell*, vol. 15, no. 3, pp. 330–345, 2023.

[11] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[12] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2174–2182.

[13] S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 338–353.

[14] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 1989, pp. 305–313.

[15] U. Muller, J. Ben, E. Cosatto, B. Flepp, and Y. L. Cun, "Off-road obstacle avoidance through end-to-end learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 739–746.

[16] H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao, and D. Li, "Object classification using CNN-based fusion of vision and lidar in autonomous vehicle environment," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 9, pp. 4224–4231, Sep. 2018.

[17] M. Bojarski et al., "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[18] Y. Chen et al., "Lidar-video driving dataset: Learning driving policies effectively," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5870–5878.

[19] [19] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 971–980.

[20] M. Kam, X. Zhu, and P. Kalata, "Sensor fusion for mobile robot navigation," *Proc. IEEE*, vol. 85, no. 1, pp. 108–119, Jan. 1997.

[21] J. Sasiadek and Q. Wang, "Sensor fusion based on fuzzy Kalman filtering for autonomous robot vehicle," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, vol. 4, 1999, pp. 2970–2975.

[22] Q. Li, L. Chen, M. Li, S.-L. Shaw, and A. Nüchter, "A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 2, pp. 540–555, Feb. 2014.

[23] V. De Silva, J. Roche, and A. Kondoz, "Fusion of lidar and camera sensor data for environment sensing in driverless vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[24] W. Luo, B. Yang, and R. Urtasun, "Fast and furious: Real time end-to-end 3D detection tracking and motion forecasting with a single convolutional net," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3569–3577.

[25] N. Radwan, A. Valada, and W. Burgard, "Multimodal interaction-aware motion prediction for autonomous street crossing," *arXiv preprint arXiv: 1808.06887*, 2018.

[26] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2722–2730.

[27] N. Radwan, A. Valada, and W. Burgard, "VLocNet++: Deep multitask learning for semantic visual localization and odometry," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4407–4414, Oct. 2018.

[28] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *Int. J. Comput. Vis.*, pp. 1–47, 2018.

[29] P. Paygude et al., "Species identification for Indian seafood markets: A machine learning approach with a fish dataset," *Data in Brief*, vol. 58, 2025, Art. no. 111209.

[30] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, "Jointly learning energy expenditures and activities using egocentric multimodal signals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6817–6826.

[31] Z. Shou et al., "Online action detection in untrimmed streaming videos-modeling and evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

DOI: https://doi.org/10.54216/FPA.190105
Received: November 10, 2024 Revised: January 12, 2025 Accepted: February 11, 2025