GATED RECURRENT UNIT (GRU)

ABSTRACT

Gated recurrent units (GRUs) are a gating mechanism in recurrent neural networks, introduced in 2014 by Kyunghyun Cho et al. The GRU is like a long short-term memory (LSTM) with a forget gate, but has fewer parameters than LSTM, as it lacks an output gate. GRU's performance on certain tasks of polyphonic music modelling, speech signal modeling and natural language processing was found to be similar to that of LSTM. GRUs have been shown to exhibit better performance on certain smaller and less frequent datasets.

INTRODUCTION

GRU or Gated recurrent unit is an advancement of the standard RNN i.e recurrent neural network. It was introduced by Kyunghyun Cho et al in the year 2014.

GRUs are very similar to Long Short Term Memory(LSTM). Just like LSTM, GRU uses gates to control the flow of information. They are relatively new as compared to LSTM. This is the reason they offer some improvement over LSTM and have simpler architecture. Another Interesting thing about GRU is that, unlike LSTM, it does not have a separate cell state (Ct). It only has a hidden state(Ht). Due to the simpler architecture, GRUs are faster to train.

ARCHITECTURE

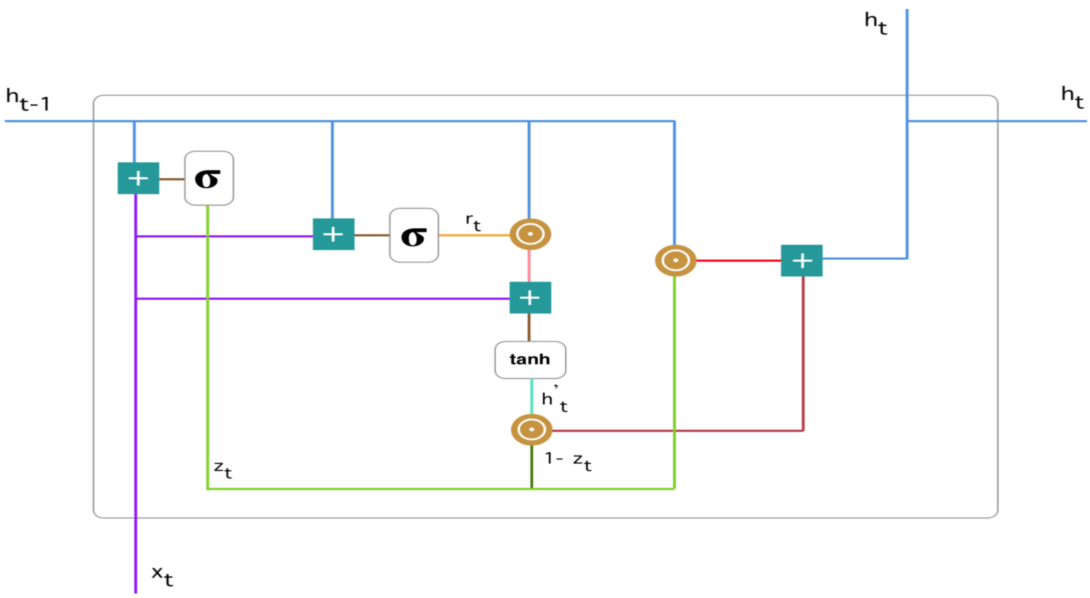


Fig. 1: Architecture of GRU

GRUs are improved version of standard recurrent neural network. To solve the vanishing gradient problem of a standard RNN, GRU uses, so-called, update gate and reset gate. Basically, these are two vectors which decide what information should be passed to the output. The special thing about them is that they can be trained to keep information from long ago, without washing it through time or remove information which is irrelevant to the prediction.

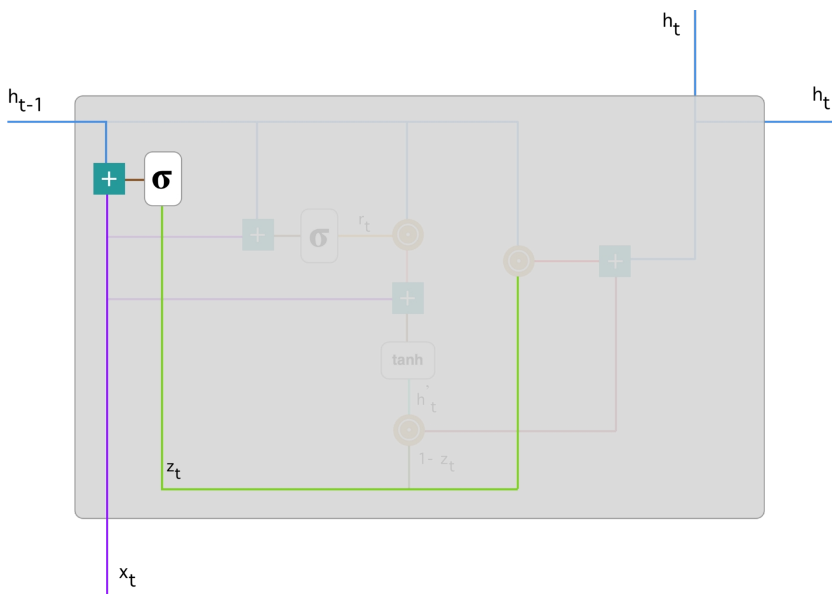
EXPLANATION OF COMPONENTS

1. Update gate

We start with calculating the update gate z_t for time step t using the formula:

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1})$$

When x_t is plugged into the network unit, it is multiplied by its own weight *W(z)*. The same goes for $h_{(t-1)}$which holds the information for the previous *t-1* units and is multiplied by its own weight $U_{(z)}$. Both results are added together and a sigmoid activation function is applied to squash the result between 0 and 1. Following the above schema, we have:



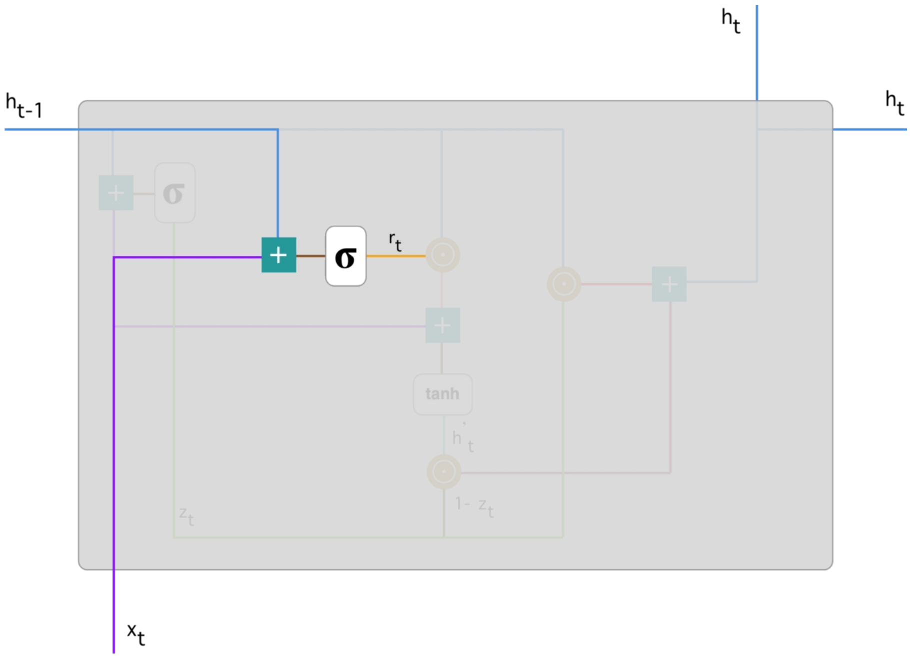
The update gate helps the model to determine how much of the past information (from previous time steps) needs to be passed along to the future. That is really powerful because the model can decide to copy all the information from the past and eliminate the risk of vanishing gradient problem.

2. Reset gate

Essentially this gate is used from the model to decide how much of the past information to forget. To calculate it, we use:

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1})$$

This formula is the same as the one for the update gate. The difference comes in the weights and the gate's usage, which will see in a bit. The schema below shows where the reset gate is:



As before, we plug in $h_{(t-1)}$ and x_t , multiply them with their corresponding weights, sum the results and apply the sigmoid function.

3. Current memory content

Let's see how exactly the gates will affect the final output. First, we start with the usage of the reset gate. We introduce a new memory content which will use the reset gate to store the relevant information from the past. It is calculated as follows:

$$h_t' = \tanh(Wx_t + r_t \odot Uh_{t-1})$$

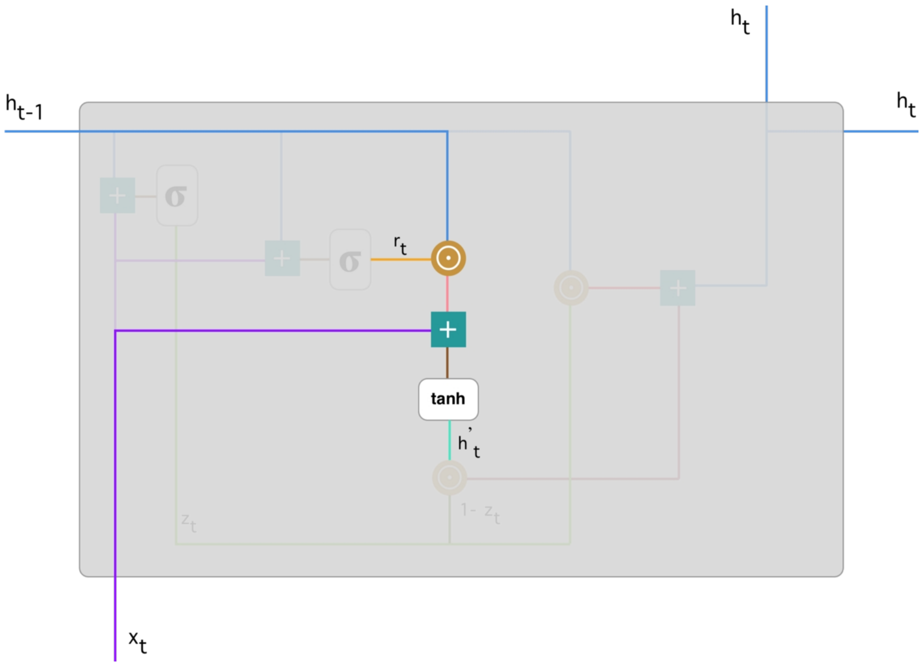
1.Multiply the input x_t with a weight W and $h_{(t-1)}$ with a weight U.

2.Calculate the Hadamard (element-wise) product between the reset gate r_t and $Uh_{(t-1)}$. That will determine what to remove from the previous time steps. Let's say we have a sentiment analysis problem for determining one's opinion about a book from a review he wrote. The text starts with "This is a fantasy book which illustrates…" and after a couple paragraphs ends with "I didn't quite enjoy the book because I think it captures too many details." To determine the overall level of satisfaction from the book we only need the last part of the review. In that case as the neural network approaches to the end of the text it will learn to assign r_t vector close to 0, washing out the past and focusing only on the last sentences.

3.Sum up the results of step 1 and 2.

4.Apply the nonlinear activation function *tanh*.

The steps are clearly depicted in the following figure:



We do an element-wise multiplication of $h_{(t-1)}$ and r_t and then sum the *result* with the input x_t. Finally, *tanh* is used to produce h'_t.

4. Final memory at current time step

As the last step, the network needs to calculate h_t vector which holds information for the current unit and passes it down to the network. In order to do that the update gate is needed. It determines what to collect from the current memory content h'_t and what from the previous steps $h_{(t-1)}$. That is done as follows:

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t$$

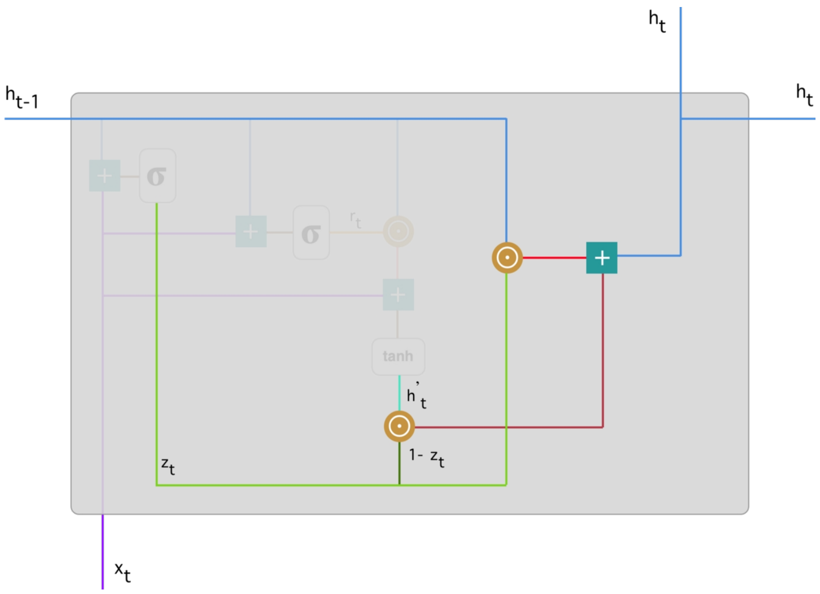
1.Apply element-wise multiplication to the update gate z_t and $h_{(t-1)}$.

2.Apply element-wise multiplication to $(1-z_t)$ and h'_t.

3.Sum the results from step 1 and 2.

Let's bring up the example about the book review. This time, the most relevant information is positioned in the beginning of the text. The model can learn to set the vector z_t close to 1 and keep a majority of the previous information. Since z_t will be close to 1 at this time step, $1-z_t$ will be close to 0 which will ignore big portion of the current content (in this case the last part of the review which explains the book plot) which is irrelevant for our prediction.

Here is an illustration which emphasises on the above equation:



Following through, you can see how z_t is used to calculate $1-z_t$ which, combined with h'_t ,produces a result. Z_t is also used with $h_{(t-1)}$ in an element-wise multiplication. Finally, h_t is a result of the summation of the outputs.

MATHEMATICAL SUPPORT

The governing equations for GRUs are:

$$z = \sigma(W_z \cdot x_t + U_z \cdot h_{(t-1)} + b_z)$$
$$r = \sigma(W_r \cdot x_t + U_r \cdot h_{(t-1)} + b_r)$$
$$\tilde{h} = tanh(W_h \cdot x_t + r * U_h \cdot h_{(t-1)} + b_z)$$
$$h = z * h_{(t-1)} + (1 - z) * \tilde{h}$$

APPLICATIONS

1. **GRU-based Attention Mechanism for Human Activity Recognition**: Sensor data based Human Activity Recognition (HAR) has gained interest due to its application in practical field. With increasing number of approaches incorporating feature learning of sequential time-series sensor data, in particular the deep learning based ones has performed reasonably in uniform labelled data distribution scenario. However, most of these methods do not capture properly the temporal context of time-steps in sequential time-series data. Moreover, the situation becomes worse for imbalanced class distribution which is a usual case for HAR using body-worn sensor devices. To solve this issues, hierarchical attention mechanism is integrated with recurrent units of neural network in order to obtain temporal context within the time-steps of data sequence. The introduced model in this paper has achieved better performance with respect to the well-defined evaluation metrics in both uniform and imbalanced class distribution than the existing state-of-the-art deep learning based model.
2. **A Deep Bidirectional GRU Network Model for Biometric Electrocardiogram Classification Based on Recurrent Neural Networks**: In this paper, propose a deep Recurrent Neural Networks (RNNs) based on Gated Recurrent Unit (GRU) in a bidirectional manner (BGRU) is proposed for human identification from electrocardiogram (ECG) based biometrics, a classification task which aims to identify a subject from a given time-series sequential data. Despite having a major issue in traditional RNN networks which they learn representations from previous time sequences, bidirectional is designed to learn the representations from future time steps which enables for better understanding of context, and eliminate ambiguity. Moreover, GRU cell in

RNNs deploys an update gate and a reset gate in a hidden state layer which is computationally efficient than a usual LSTM network due to the reduction of gates. The experimental results suggest that our proposed BGRU model, the combination of RNN with GRU cell unit in bidirectional manner, achieved a high classification accuracy of 98.55%. Various neural network architectures with different parameters are also evaluated for different approaches, including one-dimensional Convolutional Neural Network (1D-CNN), and traditional RNNs with LSTM and GRU for non-fiducial approach. The proposed models were evaluated with two publicly available datasets: ECG-ID Database (ECGID) and MIT-BIH Arrhythmia Database (MITDB). This paper is expected to demonstrate the feasibility and effectiveness of applying various deep learning approaches to biometric identification and also evaluate the effect of network performance on classification accuracy according to the changes in percentage of training dataset.

3. **An Attention-Based Deep Sequential GRU Model for Sensor Drift Compensation**: Sensor accuracy is vital for the reliability of sensing applications. However, sensor drift is a common problem that leads to inaccurate measurement readings. Owing to aging and environmental variation, chemical gas sensors in particular are quite susceptible to drift with time. Existing solutions may not address the temporal complex aspect of drift, which a sequential deep learning approach could capture. This article proposes a novel deep sequential model named Concatenated GRU & Dense layer with Attention (CGDA) for drift compensation in low-cost gas sensors. Concatenation of a stacked GRU (Gated Recurrent Unit) block and a dense layer is integrated with an attention network, that accurately predicts the hourly drift sequence for an entire day. The stacked GRU extracts useful temporal features layer by layer capturing the time dependencies at a low computational expense, while the dense layer helps in retention of handcrafted feature knowledge, and the attention mechanism facilitates adequate weight assignment and elaborate information mapping. The CGDA model achieves a significant mean accuracy over 93%, outperforming several state-of-the-art shallow and deep learning

models besides its ablated variants. It can greatly enhance the reliability of sensors in real-world applications.

4. **A Parallel GRU Recurrent Network Model and Its Application to Multi-Channel Time-Varying Signal Classification**: This study presents a modified recurrent neural network (RNN) model designed as a parallel computing structure for serial information processing. The result is a novel parallel recurrent neural network (P-RNN), proposed for application to time-varying signal classification. The network uses gated recurrent units (GRUs) for basic information processing and consists of a multi-channel time series signal input layer, parallel processing structure units, a signal feature fusion layer, and a softmax classifier. The P-RNN expands the existing RNN serial processing mode for multi-channel time-varying signals into parallel mode and realizes the embedding of multi-channel signal structure features. In these parallel processing units, the input signal for each channel corresponds to a GRU recurrent network. Feature extraction and attribute association of single-channel signals were performed to achieve parallel processing of all-channel signals. In the feature fusion layer, feature vectors from each channel signal were integrated to generate a comprehensive feature matrix. On this basis, the softmax function was used as a classifier for multi-channel signals. With this mechanism, the P-RNN model achieved independent feature extraction of single-channel signals, characteristic fusion of each channel signal, and signal classification based on an integrated feature matrix. This approach maintained characteristic combination relationships that improved serial modes for existing RNN multi-channel signal processing, reduced the loss of structural feature information, and improved the representation ability of combined feature in local time region and the efficiency of the algorithm. In this paper, the properties of the proposed P-RNN are analyzed and a comprehensive learning algorithm is developed. Seven disease classification types commonly diagnosed using 12-lead ECG signals were used to validate the technique experimentally. Results showed the computational efficiency improved by a factor of 11.519, compared with existing RNN serial processing times, producing a correct recognition rate of 95.976%. In particular,

the resolution of signal samples with similar distribution characteristics improved significantly, which demonstrates the effectiveness of the proposed technique.

5. **Forecasting Method based upon GRU-based Deep Learning Model:** In this, the world model has a modified RNN model carried out by a bi-directional gated recurrent unit (BGRU) as opposed to a traditional long short-term memory (LSTM) model. BGRU tends to use less memory while executing and training faster than an LSTM, as it uses fewer training parameters. However, the LSTM model provides greater accuracy with datasets using longer sequences. Based upon practical implementation, the BGRU model produced better performance results. In BGRU, the memory is combined with the network. There is no update gate and forget in the GRU. The forget and update gate are treated as one unit thus it is the primary reason of parameter reduction.
6. **GRU Based Deep Learning Model for Prognosis Prediction of Disease Progression:** Recently different Deep Learning (DL) models have been emerging to predict the disease amelioration. Generally deep learning is the state of art for learning the multiple representation of neuron. Discriminant DL model includes different architecture and initial architecture is Recurrent Neural Network (RNN) which learn the data labelled in the form of sequence and it has long term dependency and vanishing gradient problem. Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) is an improved version of RNN which deal with problems in RNN. Deep care LSTM is one of the model in LSTM developed by [Trang Pham et.al] for predicting diabetes disease and compared the results with markovian and support vector machine model. In deep LSTM model some problems are reported like short term trajectories, less accuracy. To overcome the problem Deep Care GRU has been proposed to identify the diabetes disease amelioration.
7. **Intelligent Traffic Flow Prediction Using Optimized GRU Model:** Facilitating citizens with accurate traffic flow prediction increases the quality of life. Roadside sensors and devices are used to capture live streams of huge data and the Internet of Things (IoT) is becoming popular for the deployment of

effective Intelligent Transportation Systems (ITS). Traffic flow prediction from the live datastreams require building a data-driven model. This is a challenging task and has attracted researchers for better interpretation of the traffic characteristics. The core problem in traffic prediction is modeling a diversity of traffic trends and unpredictable flow variations with temporal dependencies. Initially, statistical and shallow neural network models were applied to some extent. Recently, deep learning has come up with proven and promising outcomes. Gated Recurrent Unit (GRU) is a variation of recurrent neural networks used effectively for traffic flow prediction. Like other deep networks, GRU uses hyperparameters and a sliding window time-steps mechanism to prepare and tune the model. Better tuning for hyperparameters and search for optimal window size is a tedious process. In this research work, we present an algorithm for hyperparameters tuning along with sliding window steps optimization. Results obtained on a real-time public traffic dataset show a higher capability of the proposed method to reduce the error and an average gain of the optimized model over the untuned network is 4.5%. Furthermore, we apply the optimal hyperparameters obtained in the experiment to other deep learning models and present that our approach improves prediction accuracy and stability.

8. **Application of Gated Recurrent Unit (GRU) Neural Network for Smart Batch Production Prediction:** Production prediction plays an important role in decision making, development planning, and economic evaluation during the exploration and development period. However, applying traditional methods for production forecasting of newly developed wells in the conglomerate reservoir is restricted by limited historical data, complex fracture propagation, and frequent operational changes. This study proposed a Gated Recurrent Unit (GRU) neural network-based model to achieve batch production forecasting in M conglomerate reservoir of China, which tackles the limitations of traditional decline curve analysis and conventional time-series prediction methods. The model is trained by four features of production rate, tubing pressure (TP), choke size (CS), and shut-in period (SI) from 70 multistage hydraulic fractured

horizontal wells. Firstly, a comprehensive data preprocessing is implemented, including excluding unfit wells, data screening, feature selection, partitioning data set, z-score normalization, and format conversion. Then, the four-feature model is compared with the model considering production only, and it is found that with frequent oilfield operations changes, the four-feature model could accurately capture the complex variance pattern of production rate. Further, Random Forest (RF) is employed to optimize the prediction results of GRU. For a fair evaluation, the performance of the proposed model is compared with that of simple Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) neural network. The results show that the proposed approach outperforms the others in prediction accuracy and generalization ability. It is worth mentioning that under the guidance of continuous learning, the GRU model can be updated as soon as more wells become available.

9. **Water Level Prediction Model Based on GRU and CNN**: Massive amount of water level data has been collected by using Internet of Things (IoT) techniques in the Yangtze River and other rivers. In this paper, utilizing these data to construct deep neural network models for water level prediction is focused. To achieve higher accuracy, both the factors of time and locations of data collection sensors are considered to perform prediction. And the network structures of gated recurrent unit (GRU) and convolutional neural network (CNN) are combined to build a CNN-GRU model in which the GRU part learns the changing trend of water level, and the CNN part learns the spatial correlation among water level data observed from adjacent water stations. The CNN-GRU model that using data from multiple locations to predict the water level of the middle location has higher accuracy than the model only based on GRU and other state-of-the-art methods including autoregressive integrated moving average model (ARIMA), wavelet-based artificial neural network (WANN) and long-short term memory model (LSTM), because of its ability to decrease the affections of abnormal value and data randomness of a single water station to some extent. The results are verified on an experiment dataset that including 30-year observed data of water level at several collection stations in the Yangtze

River. For forecasting the 8-o'clock water levels of future 5 days, accuracy of the CNN-GRU model is better than that of ARIMA, WANN and LSTM models with three evaluation factors including Nash-Sutcliffe efficiency coefficient (NSE), average relative error (MRE) and root mean square error (RMSE).

10. **Full-GRU Natural Language Video Description for Service Robotics Applications:** Enabling effective human–robot interaction is crucial for any service robotics application. In this context, a fundamental aspect is the development of a user-friendly human–robot interface, such as a natural language interface. In this letter, we investigate the robot side of the interface, in particular the ability to generate natural language descriptions for the scene it observes. We achieve this capability via a deep recurrent neural network architecture completely based on the gated recurrent unit paradigm. The robot is able to generate complete sentences describing the scene, dealing with the hierarchical nature of the temporal information contained in image sequences. The proposed approach has fewer parameters than previous state-of-the-art architectures, thus it is faster to train and smaller in memory occupancy. These benefits do not affect the prediction performance. In fact, we show that our method outperforms or is comparable to previous approaches in terms of quantitative metrics and qualitative evaluation when tested on benchmark publicly available datasets and on a new dataset we introduce in this letter.
11. **Application of Regularized GRU-LSTM Model in Stock Price Prediction:** The stock market is a highly complex nonlinear movement system, and its fluctuation law is affected by many factors, so the prediction of the stock price index is a very challenging task. There are many examples showing that Neural Network algorithms are well suited for such time series predictions and often achieve satisfactory results. In this paper, based on the existing models, we proposed a Regularized GRULSTM neural network model and applied it to the short-term forecast of closing price of the two stocks. The experimental results show that our proposed model is superior to the existing GRU and LSTM network models in stock time series prediction.

12. **Full Attention-Based Bi-GRU Neural Network for News Text Classification**: This paper proposes a novel approach for text classification by using attention mechanism. In recent works, several models based on deep learning with traditional attention mechanism mainly learn the weights of steps in the entire text. However, the information of each step is filtered by the encoder, and the same information has different effects on different steps. This paper proposes a full attention-based bidirectional GRU (Bi-GRU) neural network, which is called FABG. FABG uses a Bi-GRU to learn the semantic information of text, and uses full attention mechanism to learn the weights of previous and current outputs of the Bi-GRU at each step, which enables the representation of each step to obtain the important information and ignore the irrelevant information. Finally, through a pooling layer, we get the representation of the text. Thereby FABG can learn more information, which enhances the effect of text classification. Experiments on the English news dataset agnews and the Chinese news dataset chnews show that FABG achieve better performance than the baselines

ADVANTAGES

- GRU use a gate structure to overcome the impact of short-term memory. The gate structure can regulate the information flow through the sequence chain.
- GRU can retain important features through various gates, ensuring that important special features will not be lost during long-term transmission.
- The structure of GRU is simpler.
- It has one gate less than LSTM, which reduces matrix multiplication
- GRU can save a lot of time without sacrificing performance.

LIMITATIONS

- The advantage of GRU only holds in the scenario of long text and small datasets.
- In other scenarios, compared with LSTM, the performance loss of GRU is more serious.
- In an age when computing power is no longer a bottleneck, LSTM is actually more suitable in these scenarios.

OTHER COMPARABLE ARCHITECTURES

Variant 1: called GRU1, where each gate is computed using only the previous hidden state and the bias.

$$z_t = \sigma(U_z h_{t-1} + b_z) \qquad (5\text{-a})$$

$$r_t = \sigma(U_r h_{t-1} + b_r) \qquad (5\text{-b})$$

Thus, the total number of parameters is now reduced in comparison to the GRU RNN by 2× nm.

Variant 2: called GRU2, where each gate is computed using only the previous hidden state.

$$z_t = \sigma(U_z h_{t-1}) \qquad (6\text{-a})$$

$$r_t = \sigma(U_r h_{t-1}) \qquad (6\text{-b})$$

Thus, the total number of parameters is reduced in comparison to the GRU RNN by 2× (nm+n).

Variant 3: called GRU3, where each gate is computed using only the bias.

$$z_t = \sigma(b_z) \qquad (7\text{-a})$$

$$r_t = \sigma(b_r) \qquad (7\text{-b})$$

Thus the total number of parameters is reduced in comparison to the GRU RNN by 2× (nm+n2).

An empirical study of the performance of each of these variants as compared to the GRU RNN is performed on, first, sequences generated from the MNIST dataset and then on the IMDB movie review dataset. Here we refer to the base GRU RNN model as GRU0 and the three variants as GRU1, GRU2, and GRU3 respectively. Our architecture consists of a single layer of one of the variants of GRU units driven by the input sequence and the activation function set as ReLU. (Initial experiments using g = tanh have produced similar results). For the MNIST dataset, we generate the pixel-wise and the row-wise sequences. The networks have been generated in Python using the Keras library with Theano as a backend library. As Keras has a GRU layer class, we modified this class to classes for GRU1, GRU2, and GRU3. All of these classes used the ReLU activation function. The RNN layer of units is followed by a softmax layer in the case of the MNIST dataset or a traditional logistic activation layer in the case of the IMDB dataset to predict the output category. The Root Mean Square Propagation

(RMSprop) is used as the choice of optimizer that is known to adapt the learning rate for each of the parameters. To speed up training, we also decay the learning rate exponentially with the cost in each epoch

$$\eta = \eta_0 e^{cost} \qquad (8)$$

where η_0 represents a base constant learning rate and cost is the cost computed in the previous epoch.

REFERENCES

1. Gated Recurrent Unit | Introduction to Gated Recurrent Unit(GRU) (analyticsvidhya.com)
2. Understanding GRU Networks. In this article, I will try to give a… | by Simeon Kostadinov | Towards Data Science
3. GRU-based Attention Mechanism for Human Activity Recognition-1st International Conference on Advances in Science, Engineering and Robotics Technology 2019 (ICASERT 2019)
4. A Deep Bidirectional GRU Network Model for Biometric Electrocardiogram Classification Based on Recurrent Neural Networks- Htet Myet Lynn, Sung Bum Pan, And Pankoo Kim- 2019
5. An Attention-Based Deep Sequential GRU Model for Sensor Drift Compensation Tanaya Chaudhuri , Min Wu , Senior Member, IEEE, Yu Zhang, Pan Liu, and Xiaoli Li , Senior Member, IEEE- IEEE SENSORS JOURNAL, VOL. 21, NO. 6, MARCH 15, 2021
6. A parallel gru recurrent network model and its application to multi-channel time-varying signal classification shaohua xu, jingjing li, kun liu, and lu wu- 2019
7. Forecasting Method based upon GRU-based Deep Learning Model- Ali Jaber Almalki, Pawel Wocjan- 2020 International Conference on Computational Science and Computational Intelligence (CSCI)

8. GRU Based Deep Learning Model for Prognosis Prediction of Disease Progression- M.Pavithra, Dr K.Saruladha, K.Sathyabama- Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)
9. Intelligent Traffic Flow Prediction Using Optimized GRU Model Basharat Hussain, Muhammad Khalil Afzal, (Senior Member, Ieee), Shafiq Ahmad, And Almetwally M. Mostafa- 2021
10. Application of Gated Recurrent Unit (GRU) Neural Network for Smart Batch Production Prediction Xuechen Li , Xinfang Ma, Fengchao Xiao, Fei Wang and Shicheng Zhang- 2020
11. Water level prediction model based on gru and cnn- Mingyang Pan, Hainan Zhou, Jiayi Cao, Yisai Liu, Jiangling Hao, Shaoxi Li, And Chi-Hua Chen- 2020
12. Full Attention-Based Bi-GRU Neural Network for News Text Classification- Tang, Qinting; li, Jian; Chen, Jiayu; Lu, Hengtong; Du, Yu; Yang, Kehan (2019). *[IEEE 2019 IEEE 5th International Conference on Computer and Communications (ICCC) - Chengdu, China (2019.12.6-2019.12.9)] 2019*
13. Application of Regularized GRU-LSTM Model in Stock Price Prediction- Yiwei Liu, Zhiping Wang, Baoyou Zheng- *[IEEE 2019 IEEE 5th International Conference on Computer and Communications (ICCC) - Chengdu, China (2019.12.6-2019.12.9)]*
14. LSTM and GRU neural network performance comparison study taking Yelp review dataset as an example- Yang, Shudong; Yu, Xueying; Zhou, Ying (2020). *[IEEE 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI) - Shanghai, China (2020.6.12-2020.6.14)]*

GATED RECURRENT UNIT

Submitted to,

Prof. Sonia Abraham

Submitted by,

Sumal Roshna K S