Network in Sequential Form: Combine Tree Structure Components into Recurrent Neural Network

Swaleha Zubair\(^1\) and Anjani Kumar Singha\(^1\)

\(^1\) Aligarh Muslim University, India
\(^2\) Aligarh Muslim University, India

Email id: anjani348@gmail.com

Abstract. Smaller units nestled in larger units - a natural language that is hierarchically ordered. Smaller sections are replaced because of the completion of the bigger constituents. The basic LSTM architecture does not have a clear choice for modeling the hierarchy of components as separate neurons do not require knowledge to be monitored on various time scales. The idea in this paper is to introduce inductive bias by grouping the neurons. The modifications we create to the master input vector are changed in all neurons following the order of the specified neuron and forget windows. Four activities on which we can achieve strong care unattended sorting, logical inference, and language modeling and guided syntax evaluation.

1. Introduction

In recent years (Bowman's et al., 2006; Yoga tama et al., 2017; Sheen et al., 2016; Jacob et al., 2019; Choir[1,2] et al., 2017;[1,2,3]) the deep NN techniques have a strong inference that tree structures too can make better representations of the natural language. A supervised algorithm [3, 4] is a straightforward way to guess the respective tree structure. The results obtained showed the way to turn words into semi-semantic sentences (Soccer et al., 2013; Bowman's et al., 2015), and, despite previous words (Wu et al., 2017), the forth coming ones are guessed. Yet there are some drawbacks to this method too. Some of the languages have comprehensively explicit data for supervised analysis parser training for a few of our parts we have, certain rules that need broken (e.g. in social site) on more use of languages tend to change gradually, so that new syntax rules may evolve. Recent attempts include the implementation of a trial solution, which faces more difficulty throughout the learning cycle induced by inducing. When going forward, few techniques are comparatively more difficult to implement and understand, such as the Parsing-Reading-Predict Networks (PRPN) [3, 4] shown in et al (2017). In recent years, the study of neural networks, mainly use the techniques that use parse structures of tree to more information of sentences [5]. Syntactic parsers may provide a direct way to predict a sentence. The parse trees that output from it will words of semantics into sentences form and grammar form, when it the defined of learning a specific method of syntax through broad natural and structured [4, 5] collection of non-access texts to transform structured data into pro leaves a strong dynamic behind. Many such sentence to induce a trivial or even left-branching parse tree [5, 6] (Williams et al., 2019), [8,5, 6,7]) or observations of learning dynamics caused by branched RL learning guidelines (Yoga tama et al., 2016). Moreover, in a specific model like PRPN [6, 7] which was told in Sheen et al., (2017), there are some hard forms of training.

2. Related Work
Prior research involves exploiting the materials of structures for the proper language tasks. Soccer et al. (2011), use unsupervised learning or supervised learning for the estimation of trees on [9, 8] adept-unlabeled or labeled parse tree et al. (2015); Soccer et al. and Tai et al. (2016), specifically modeled parse tree construction by using a reasonable outer description to define the data. Afterwards Bowman et al., (2017), [9,7] used some guidance from the monitored definition to guide a web of stack-augmented nerve organs. In theoretical terms, Recurrent Neural Network may modulate recognition generated by CFG, in addition to CSG, in addition to LSTM. The structural knowledge generated in LSTM is, however, beneficial. Suncor et.al. (2018), [9] has shown that RNNG which has a strong overt bias against one of the semantic structures, overcomes Long Short-Term Memory (Linen et al., (2018)), on the subject-verb agreement task [9, 10]. We build and function in this own function with a modern, largest-scale space associated with the grammatical test recently proposed by Linen and Marvin et al. (2017), [7, 10] the effect that the demonstration caused by tree-structured models is more apparent. Interesting is the work generated by recursive plans, Shi et al., (2018), [11,10] says that although approved grammar woods are not as good as pyramidal structures, relevant works would therefore work. Some way that the basic problems of efficiently inferring some kind of construction by the resulting data remain starting and ending good query. The grammar associated with learning the level construction of the sentence by text are called semantic induction (Chen et al.,1991; Cohen et al., 2017), [10, 11]. Until research incorporates syntactic construction within the semantic model system [11,12]. There have been undertaking to trigger many mechanisms for greater possibilities using Neural Network models, including in recent studies (Grefenstette et al. 2019, Mikonos and Julian et al. (2014, 2015). Generally, these works maintain the RNN core model with a stack and focus on finding solutions to yoga Tama et el. al, (2019), operating on the semantic model. In addition to the function of semantic evaluation (Linen et al. , 2017), [9, 10] does not view the context in which the modeling structure is organized Sheen et al. , (2015),[11,12,13] implements the replication of Parsing-Reading-Predict Systems, which aims to perform tree simply by the process of grammar recreation.

3. Proposed Method
In this part, we present an innovative recurrent neural network unit, ON- Long Short-Term Memory. The innovative model uses a structure similar to the regular Long Short-Term Memory, indicated as:

\[
\begin{align*}
    f_t &= \sigma \left( w_f x_t + u_f y_{t-1} + b_f \right) \\
    j_t &= \sigma \left( w_j x_t + u_j y_{t-1} + b_j \right) \\
    k_t &= \sigma \left( w_k x_t + u_k y_{t-1} + b_k \right) \\
    \tilde{C}_t &= \tanh \left( w_c x_t + u_c y_{t-1} + b_c \right) \\
    V_t &= k_t \circ \tanh(C_t), \text{where } \circ \text{ is any operation}
\end{align*}
\]

With the LSTM, we change the update function (c_t) using a new function (f_t) which is explained in following sections. Where, the forget gates function (f_t) input gates function (j_t) used to control the wipe out, output gates operation (k_t) on cell states (c_t) as before and bias function b. Since, the gates of each neurons are independent of LSTM.

3.1. Function of Activation
Qmax () in order to an implement to particular recurrence neural network, we present a functionality:

\[
\hat{G} = \text{Qmax}(..) = \text{Qsum} \left( \text{QSoftMax} \left( .. \right) \right)
\]
Where, \( Q_{\text{sum}} \) denotes the cumulative sum \([14, 15]\). For assumption of the Boolean gates \( G= (1; 1; 0; 0) \) we produced \( \bar{G} \). The binary gates divide unit state into two different segments like the 1-segment and the 0-segment. Therefore, applying different semantic can different updating rules on two segments to differentiate (LSTM), information. It is denoted by \( d_1 \) in random variables by representing index first in \( G \)

\[
P (d) = \text{SoftMax} (\ldots) \tag{7}
\]



The represents variable \( d_1 \), split into two segment and compute probability of the i-th values in \( G \) being 1 from find the values of probability of any values before the i-th divide point, which is \( d \leq i = (d=0) \lor (d=1) \lor \ldots \lor (d=i) \). Computing distribution function is evaluating by this formula

\[
P (G_i = 1 ) = P (d \leq i) = \sum_j \leq 1 p(d=j) \tag{8}
\]

Where, \( G \) is has to discrete variable.

### 3.2. Mechanism of Structured

Following \( Q_{\text{max}} \) activation (), we bring out a new Master-Forget- Gate (MFG) \( \hat{h}_t \), and input Master-Gate(MG) \( I_t \)

\[
\hat{h}_t = Q_{\text{max}} (w_{\hat{h}} x_t + u_{\hat{h}} v_{i,t-1} + b_{\hat{h}}) \tag{9}
\]

\[
\bar{I}_t = 1 - \text{Cumax} (w_f x_t + u_f v_{i,t-1} + b_f) \tag{10}
\]

Using the properties from activation function \( Q_{\text{max}} () \), the resulting results in the MSG are monotonically non decreasing from 0 to 1 and input MG are non increasing from 1 to 0. These MG control by update of cell state to new update master gate rule.

\[
w_t = \hat{h}_t \circ \bar{I}_t \tag{11}
\]

\[
f_t = f_t \circ w_t + I_t \circ h_t = \hat{h}_t (f_t \circ j_t + 1 - \bar{I}_t) \tag{12}
\]

\[
\hat{h}_t = f_t \circ w_t + (\bar{I}_t - w_t) = \bar{I}_t \circ (\bar{I}_t \circ \hat{h}_t + 1 - h_t) \tag{13}
\]

\[
C_t = \hat{h}_t \circ C_{t-1} + \bar{I}_t \circ \bar{C}_t \tag{14}
\]

In order to update new master rule, we suppose that new master gate are in Boolean form Applying the MSG entering \( \hat{h}_t \), the individual eliminates of semantic the model. Postulate \( \hat{h}_t = (0, 0, 1, 1) \) and the divide into cell \( d^t_1 \). Currently the equation (12) and (14), the stored information \( d^t_1 \) neuron and unit cell \( \bar{C}_t \) fully deleted. The MG input \( I_t \) to control the script semantic of the model. Suppose that split point \( d^t_1 \) of equation (13) and (14) the large \( d^t_1 \) means that start input \( X^t \) contain long term. The particular product from the two MG \( w_t \) represents the certain overlap of \( f_t \) and \( I_t \).
Figure 1. Single complexity model test and validate for the Penn Treebank vocabulary model task.

| Model                                           | Parameters | Validation | Test   |
|------------------------------------------------|------------|------------|--------|
| Zaremba et al. (2014) - LSTM (large)            | 66M        | 82.2       | 78.4   |
| Gal & Ghahramani (2016) - Variational LSTM (large, MC) | 66M        | —          | 73.4   |
| Kim et al. (2016) - CharCNN                      | 19M        | —          | 78.9   |
| Merity et al. (2016) - Pointer Sentence-LSTM    | 21M        | 72.4       | 70.9   |
| Grave et al. (2016) - LSTM                       | —          | —          | 82.3   |
| Grave et al. (2016) - LSTM + continuous cache pointer | —          | —          | 72.1   |
| Inan et al. (2016) - Variational LSTM (tied) + augmented loss | 51M        | 71.1       | 68.5   |
| Zilly et al. (2016) - Variational RNN (tied)    | 23M        | 67.9       | 65.4   |
| Zoph & Le (2016) - NAS Cell (tied)              | 54M        | —          | 62.4   |
| Shen et al. (2017) - PRPN-LM                    | —          | —          | 62.0   |
| Melis et al. (2017) - 4-layer skip connection LSTM (tied) | 24M        | 60.9       | 58.3   |
| Merity et al. (2017) - AWD-LSTM - 3-layer LSTM (tied) | 24M        | 60.0       | 57.3   |
| ON-LSTM - 3-layer (tied)                        | 25M        | 58.29 ± 0.1 | 56.17 ± 0.12 |
| Yang et al. (2017) - AWD-LSTM-MoS*              | 22M        | 56.5       | 54.4   |

4. Experiment

We evaluate the four model: Unsupervised constituent parsing, targeted semantic, language model analysis (Linen and Marvin et al., 2019), [2, 17] and logical interface (Bowman et al., (2018)).

4.1. Language of Model

Model of world range is a macroscopic language evaluation from model’s stamina the work out with many old age depictions. Our model evaluates simply by counting complexity on certain penn-tree-bank (PTB), for more accurate we precisely stick to the major parameters, regulation and easy way optimization in AWD-LSTM. This work used Tri-layered ON-LSTM models along the 1200 unit in a prefaced layer and embedment association of size 500 units. For grasp gates the loud element is M=20. The particular tool moves from 26 to 27 million along with matrices intended for computing predominant gates. (see Figure.1) performs greater standards of LSTM.

4.2. Unsupervised Parsing

The unsupervised constituency parsing relates valuable tree typed model with these marginalized human authorities. Looking at the well experienced configurations listed at Hut et al (2018) [16, 17, 18]. We obtain a perfect structure to the language modeling task in addition to its WSJ10 dataset and WSJ analyzed set. It has 7633 lines; we filter WSJ dataset with the limit of only 15 words or less and right punctuation marks and excluding null elements. While the WSJ analyze contains 3000 paragraphs with varying lengths. That is WSJ10 test set consists of sentences from the education, test and validation arranged regarding PTB dataset, here WSJ test utilizes the pairs of similar group phrases as the PTB analyze set. This is a tree structured paragraph from previous models, initialization is done by hiding state with zero vectors, and insertion of paragraphs into model is made, as we require paragraph modeling activity. In each time we complete an estimation

\[
\hat{n}_i^j = E[d_i^j] = \sum_{i=1}^{D_m} i P_f (d_i = i) = \sum_{i=1}^{D_m} \sum_{j=1}^{j-1} P_f (d_i = i) = D_m \sum_{i=1}^{D_m} \hat{n}_i
\] (15)
Where $P_f$ is the probability normal distribution and $D_m$ is the size of hidden layer. Given $R_{i,t}$, use the top-down parsing algorithm within Sheen et al., (2017), for unsupervised parsing tree. We first sort the any $a_{t}^{i}$ in non-increasing order. For the particular sorted sequence order, we are divided the sentence into language $((X <j); (XI; (X>j)))$. After that, we recursive repeat this particular operation for constituent $X <j)$ and $(X> j)$, until each constituent contains paragraphs. The performance of (see Figure 2). Typically the second layer of ON- Long Short-Term Memory got state-of-the-art unsupervised parsing tree output on WSJ test set, while the first layer and four layer usually do not used.

4.3. Evaluation of Syntactic Language

Model of semantic proposed asks to defined in Linen & Marvin (2017,2018) [3, 18, 19]. It actually a group of task that will evaluate language models together, the different structured-CSG phenomena: reflexive anaphora and Verb agreement items, negative polarity, identified a huge count small phrases in English language, each containing the new in ungrammatical and grammatical terms. The grammatical phase using the similar configuration proposed in Marvin as well as linzen. Our language models consists of ON LSTM model [18,19] designed on the baseline of LSTM language model containing new 100 million words of Google. Vocabulary models semantic has more than two levels 700 cells and a mass scale 129 units, a new rate of dropout 4, rate of leering 30, and trained for epoch 50. The embedding of input has 300 and even embedding of output has 700 demission.

5. Results within RS

Spinn and St-Gumbel are assessed with WSJ. We run the design on a particular F1 norms. The relativity and better portion of truth matters of the variety that corresponds with the matters among model parses. The model along with best F1 report to ADJP, natural language processing, PP. The baselines of WSJ10 are from Linzen and manning. The result of italics is more serious than the random one. We can realize that some are particular to NPI test.

Figure 2. WSJ test. & full WSJ10 results F1 for unlabeled parsing
6. Inference

This work proposes unordered neurons, RNN is turned for novel learning purpose. On relevant particular idea, we propose a novel of RNN cells, the particular on-Long Short-Term Memory which new activation function Rectified Linear Unit (ReLU) & include mechanism of mathematical function i.e Qmax(.). This composition operation applied certain tree structure closer to Recurrent Neural Network by allocating hidden layer neurons with short term and long information. The performance of unsupervised learning model constituency parsing tree appearance which on - Long Short-Term Memory to generated inheriting structure of Natural Language Processing an approached with human expert annotation. The inductive bias sanctions ON-LSTM helped to achieve better results on language modeling and logical inference tasks.

7. References

[1] Zoph, B. and Le, Q.V., 2016. Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578.
[2] Grave, E., Joulin, A. and Usunier, N., 2016. Improving neural language models with a continuous cache. arXiv preprint arXiv:1612.04426.
[3] Bod, R., 2006, July. An all-subtrees approach to unsupervised parsing. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 865-872).
[4] Bowman, S.R., Potts, C. and Manning, C.D., 2014. Recursive neural networks can learn logical semantics. arXiv preprint arXiv:1406.1827.
[5] Bowman, S.R., Manning, C.D. and Potts, C., 2015. Tree-structured composition in neural networks without tree-structured architectures. arXiv preprint arXiv:1506.04834.
[6] Charniak, E., 2001, July. Immediate-head parsing for language models. In Proceedings of the 39th annual meeting of the Association for Computational Linguistics (pp. 124-131).
[7] Chelba, C. and Jelinek, F., 2000. Structured language modeling. *Computer Speech & Language, 14*(4), pp.283-332.

[8] Shen, Y., Lin, Z., Huang, C.W. and Courville, A., 2017. Neural language modeling by jointly learning syntax and lexicon. *arXiv preprint arXiv:1711.02013.*

[9] Choi, J., Yoo, K.M. and Lee, S.G., 2017. Learning to compose task-specific tree structures. *arXiv preprint arXiv:1707.02786.*

[10] Chomsky, Noam, Ch., 1956, "Three models for the description of language." *IRE Transactions on information theory* 2, no. 3: 113-124..

[11] Noam Chomsky, No., 1956, Aspects of the Theory of Syntax. The MIT Press, Cambridge. URL http://www.amazon.com/Aspects-Theory-Syntax-Noam-Chomsky/dp 0262530074.

[12] Junyoung Chun, SungjinAhn, and YoshuaBengio, Ju., 2016, Hierarchical multiscale recurrent neural networks *arXiv preprint arXiv:1609.01704.*

[13] Shay B Cohen, Dipanjan Das, and Noah A Smith, Sh, 2011, Unsupervised structure prediction with nonparallelmultilingual guidance. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 50–61.

[14] StanislasDehaene, FlorentMeyniel, Catherine Waongne, Liping Wang, and Christophe Pallier, St., 2015, The neural representation of sequences: from transition probabilities to algebraic patterns andlinguistic trees. *Neuron, 88*(1):2–19.

[15] Chris Dyer, AdhigunaKuncoro, Miguel Ballesteros, and Noah A Smith, Recurrent neural networkgrammars, Cd., 2016, In Proceedings of the Conference of the North American Chapter of the Associationfor Computational Linguistics: Human Language Technologies. pp. 199–209.

[16] Salah El Hihi and YoshuaBengio. Hierarchical, Sa., 1996, recurrent neural networks for long-term dependencies. In Advances in neural information processing systems. pp. 493–499.

[17] Yarin Gal and ZoubinGhahramani, Ya., 2016, A theoretically grounded application of dropout in recurrentneural networks. In Advances in neural information processing systems. pp. 1019–1027.

[18] BarretZoph and Quoc V, Ba., 2016, Le. Neural architecture search with reinforcement learning. *ArXiv preprint arXiv 1611.01578.*