ORDERED NEURONS: INTEGRATING TREE STRUCTURES

INTO RECURRENT NEURAL NETWORKS

ABSTRACT

Natural language is hierarchically structured: smaller units (e.g., phrases) are nested within larger units (e.g., clauses). This paper proposes to add such an inductive bias by ordering the neurons; a vector of master input and forget gates ensures that when a given neuron is updated, all the neurons that follow it in the ordering are also updated.

1 INTRODUCTION

Natural language has a sequential overt form as spoken and written, but the underlying structure of language is not strictly sequential. This structure is usually tree-like. From a practical point of view, integrating a tree structure into a neural network language model may be important for multiple reasons:

(i) to obtain a hierarchical representation with increasing levels of abstraction, a key feature of deep neural networks

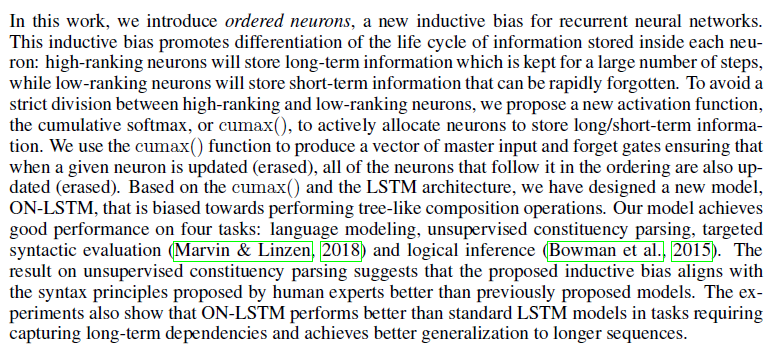
(ii) to model the compositional effects of language and help with the long-term dependency problem by providing shortcuts for gradient backpropagation

(iii) to improve generalization via a better inductive bias and at the same time potentially reducing

the need of a large amount of training data

Supervised parsers are limiting for several reasons: i) few languages have comprehensive annotated data for supervised parser training; ii) in some domains, syntax rules tend to be broken (e.g. in tweets); and iii) languages change over time with use, so syntax rules may evolve.

In this work, we introduce ordered neurons, a new inductive bias for recurrent neural networks.

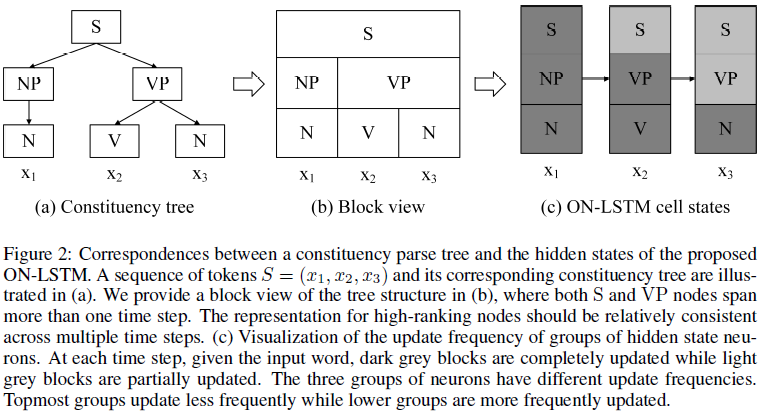


2 RELATED WORK

The task of learning the underlying grammar from data is known as grammar induction.

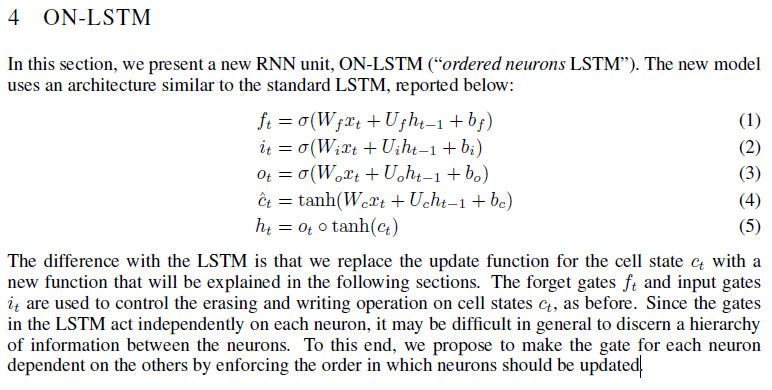
Our work can be seen as a soft relaxation of the dropout by means of the proposed cumax() activation. Moreover, we propose to condition the update masks on the particular input and apply our overall model to sequential data. Therefore, our model can adapt the structure to the observed data, while both Clockwork RNN and nested dropout impose a predefined hierarchy to hidden representations.

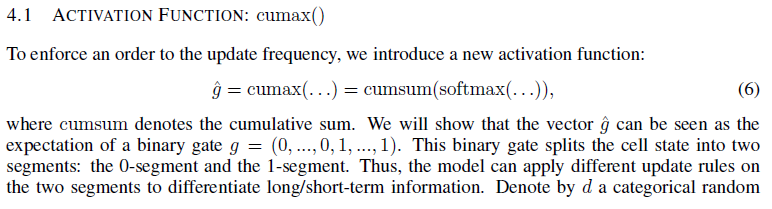
3 ORDERED NEURONS

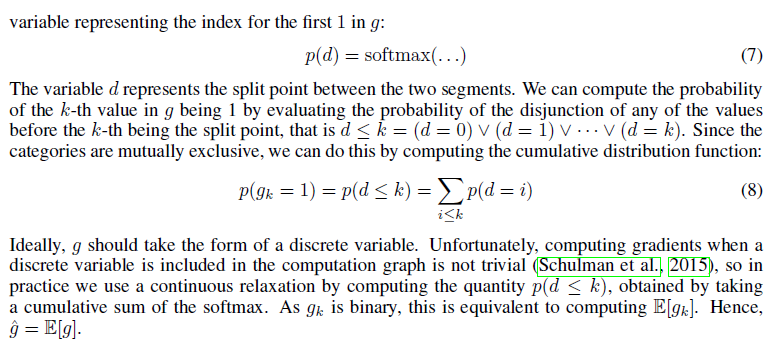


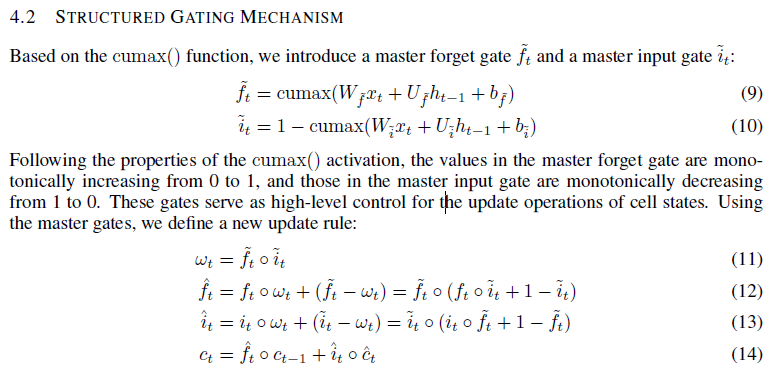
Given these requirements, we introduce ordered neurons, an inductive bias that forces neurons to represent information at different time-scales. In our model, high-ranking neurons contain long-term or global information that will last anywhere from several time steps to the entire sentence, representing nodes near the root of the tree. Low-ranking neurons encode short-term or local information that only last one or a few time steps, representing smaller constituents, as shown in Figure 2(b). The differentiation between high-ranking and low-ranking neurons is learnt in a completely data-driven fashion by controlling the update frequency of single neurons: to erase (or update) high-ranking neurons, the model should first erase (or update) all lower-ranking neurons. In other words, some neurons always update more (or less) frequently than the others, and that order is pre-determined as part of the model architecture.

4 ON-LSTM









5 EXPERIMENTS

5.1 LANGUAGE MODELING

Word-level language modeling is a macroscopic evaluation of the model’s ability to deal with various linguistic phenomena (e.g. co-occurence, syntactic structure, verb-subject agreement, etc).

5.2 UNSUPERVISED CONSTITUENCY PARSING

The unsupervised constituency parsing task compares the latent tree structure induced by the model with those annotated by human experts.

5.3 TARGETED SYNTACTIC EVALUATION

Targeted syntactic evaluation tasks have been proposed in Marvin & Linzen (2018). It is a collection

of tasks that evaluate language models along three different structure-sensitive linguistic phenomena: subject-verb agreement, reflexive anaphora and negative polarity items. Given a large number of minimally different pairs of English sentences, each consisting of a grammatical and an ungrammatical sentence, a language model should assign a higher probability to a grammatical sentence than an ungrammatical one.

5.4 LOGICAL INFERENCE

We also analyze the model’s performance on the logical inference task described in Bowman et al. (2015). This task is based on a language that has a vocabulary of six words and three logical operations, or; and; not. There are seven mutually exclusive logical relations that describe the relationship between two sentences: two types of entailment, equivalence, exhaustive and non-exhaustive contradiction, and two types of semantic independence. Similar to the natural language inference task, this logical inference task requires the model to predict the correct label given a pair of sentences. The train/test split is as described in the original codebase3, and 10% of training set is set aside as the validation set.

6 CONCLUSION

In this paper, we propose ordered neurons, a novel inductive bias for recurrent neural networks. Based on this idea, we propose a novel recurrent unit, the ON-LSTM, which includes a new gating mechanism and a new activation function cumax(\_). This brings recurrent neural networks closer to performing tree-like composition operations, by separately allocating hidden state neurons with long and short-term information. The model performance on unsupervised constituency parsing shows that the ON-LSTM induces the latent structure of natural language in a way that is coherent with human expert annotation. The inductive bias also enables ON-LSTM to achieve good performance on language modeling, long-term dependency, and logical inference tasks.