Long Short-Term Memory Over Tree Structures

The paper "Long Short-Term Memory Over Tree Structures" by Xiaodan Zhu, Parinaz Sobhani, and Hongyu Guo published in *Proceedings of the 32nd International Conference on International Conference on Machine Learning* in July 2015, extend the chain-structured long short-term memory (LSTM) to tree structures, in which a memory cell can reflect the history memories of multiple child cells or multiple descendant cells in a recursive process.The model, called as S-LSTM, provides a principled way of considering long-distance interaction over hierarchies. They leverage the models for semantic composition to understand the meaning of text, and show that it outperforms a state-of-the-art recursive model by replacing its composition layers with the S-LSTM memory blocks.

In S-LSTM, a memory cell can reflect the history memories of multiple child cells and hence multiple descendant cells in a hierarchical structure. Each memory block contains one input gate and one output gate. The number of forget gates depends on the structure, i.e., the number of children of a node. The hidden vectors of the children are taken in as input of the current block. The input gate consider the hidden vectors and cell vectors of all its children. These information are also used to form the gating signals for the forget gates, where the weights used to combining them are specific to these gates, denoted as different W in the formulas below. Different from the process in a regular LSTM, the cell here considers the copies from both children's cell vectors, gated with separated forget gates. The forget gates can be controlled independently, allowing the pass-through of information from children's cell vectors. The output gate considers the hidden vectors from the children and the current cell vector. In turn, the hidden vector and the cell vector of the current block are passed to the parent and are used depending on if the current block is a child of its parent. In this way, the memory cell, through merging the gated cell vectors of the children, can reflect multiple direct or indirect descendant cells. As a result, the long-distance interplays over the structures can be captured.

During training, the gradient of the objective function with respect to each parameter can be calculated efficiently via backpropagation over structures. The major difference from that of Socher et al., 2013 is they use LSTM-like backpropagation, where unlike a regular LSTM, pass of error needs to discriminate between children. The objective function defined over structures can be complicated, which could consider the output structures depending on the properties of problem. Following Socher et al., 2013, the overall objective function they used to learn S-LSTM in this paper is simply minimizing the overall cross-entropy errors and a sum of that at all nodes.