LSTM Recurrent Networks Learn Simple Context Free and Context Sensitive Languages

Felix A. Gers Jürgen Schmidhuber
felix@idsia.ch juergen@idsia.ch

IDSIA, Galleria 2, 6928 Manno, Switzerland, www.idsia.ch

Abstract

Previous work on learning regular languages from exemplary training sequences showed that Long Short-Term Memory (LSTM) outperforms traditional recurrent neural networks (RNNs). Here we demonstrate LSTM's superior performance on context free language (CFL) benchmarks for recurrent neural networks (RNNs), and show that it works even better than previous hardwired or highly specialized architectures. To the best of our knowledge, LSTM variants are also the first RNNs to learn a simple context *sensitive* language (CSL), namely $a^n b^n c^n$.

Index Terms: Recurrent neural networks, long short-term memory, context free languages, context sensitive languages.

1 Introduction

Recurrent neural networks (RNNs) are remarkably general sequence processing devices (Siegelmann & Sontag, 1991). In principle they are applicable to tasks beyond the reach of Hidden Markov Models (HMMs) or discrete symbolic grammar learning algorithms (SGLAs) (Lee, 1996; Sakakibara, 1997). For instance, unlike RNNs, both HMMs and SGLAs are limited to discrete state spaces. Unlike RNNs, HMMs typically ignore information conveyed by the size of temporal delays between relevant events (e.g., in rhythmic patterns). Although the HMM designer could in principle deal with a finite set of time delays between possible observations by devoting a separate internal state for each possible delay, in general this would be cumbersome and inefficient. Unlike RNNs, SGLAs cannot deal well with noisy input sequences (Osborne & Briscoe, 1997). RNNs, however, perform gradient descent in a very general, continuous space of potentially noise-resistant algorithms, using distributed internal memories for mapping real-valued input sequences to real-valued output sequences (noise tends to make learning harder though—compare, e.g., Maass and Orponen (1998), Maass and Sontag (1999)).

Until recently, however, standard RNNs (Pearlmutter, 1995) have been plagued by a major practical problem: the gradient of the total output error with respect to previous inputs quickly vanishes as the time lags between relevant inputs and errors increase (Hochreiter, 1991; Bengio, Simard, & Frasconi, 1994; Hochreiter & Schmidhuber, 1997). Hence standard RNNs fail to learn in the presence of time lags exceeding as few as 5 - 10 discrete time steps between relevant input events and target signals (Hochreiter, 1991; Schmidhuber, 1992).

The recent *"Long Short-Term Memory"* (LSTM) method (Hochreiter & Schmidhuber, 1997), however, is not affected by this problem. LSTM can learn to bridge minimal time lags in excess of 1000 discrete time steps (Hochreiter & Schmidhuber, 1997) by enforcing *constant* error flow through "constant error carousels" (CECs) within special units, without loss of short time lag capabilities. Multiplicative gate units learn to open and close access to the constant error flow. Moreover, LSTM's learning algorithm is more efficient than previous RNN algorithms such as real time recurrent learning (RTRL—Robinson & Fallside, 1987; Williams & Zipser, 1992) and back propagation through time (BPTT—Williams & Peng, 1990; Werbos, 1988): it is local in space and time, with computational complexity $O(1)$ per time step and weight.

Recent research on LSTM has concentrated on improving the structure of the adaptive gates surrounding the CECs. Here we will focus on our most recent LSTM variant (Gers & Schmidhuber, 2000), which provides the gates with direct access to the CEC states, can learn to selectively reset its own memory contents, and can produce stable results in presence of never-ending input streams.

Previous work showed that LSTM outperforms traditional RNN algorithms on numerous tasks involving real-valued or discrete inputs and targets (Gers, Schmidhuber, & Cummins, 2000; Hochreiter & Schmidhuber, 1997), including tasks that require to learn the rules of regular languages (RLs) describable by deterministic

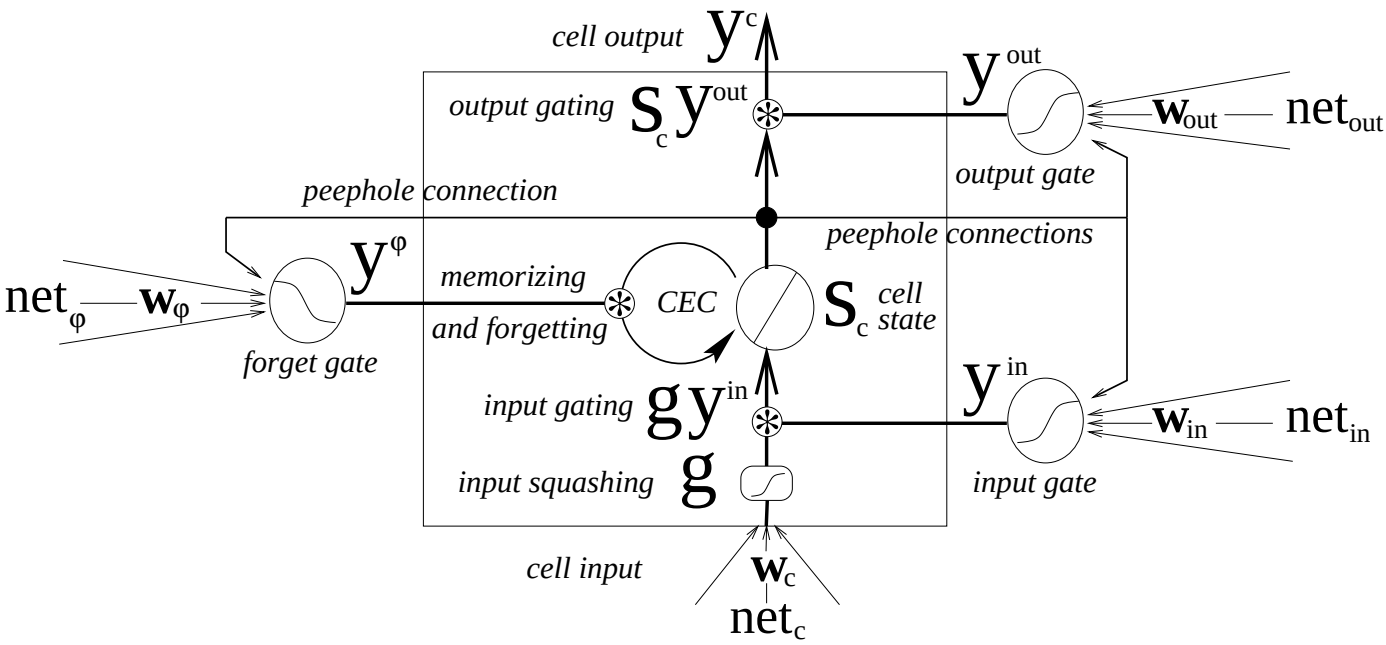

Figure 1: LSTM memory block with one cell.

finite state automata (DFA) (Casey, 1996; Siegelmann, 1992; Blair & Pollack, 1997; Kalinke & Lehmann, 1998; Zeng, Goodman, & Smyth, 1994"). Until now, however, it has remained unclear whether LSTM's superiority carries over to tasks involving context free languages (CFLs), such as those discussed in the RNN literature (Sun, Giles, Chen, & Lee, 1993; Wiles & Elman, 1995"; Steijvers & Grunwald, 1996; Tonkes & Wiles, 1997; Rodriguez, Wiles, & Elman, 1999; Rodriguez & Wiles, 1998). Their recognition requires the functional equivalent of a stack. It is conceivable that LSTM has just the right bias for RLs but might fail on CFLs.

Here we will focus on the most common CFLs benchmarks found in the RNN literature: $a^n b^n$ and $a^n b^m B^m A^n$. We study questions such as:

- Can LSTM learn the functional equivalent of a pushdown automaton?
- Given training sequences up to size n, can it generalize to $n+1, n+2, \ldots$?
- How stable are the solutions?
- Does LSTM outperform previous approaches?

Finally we will apply LSTM to a context *sensitive* language (CSL). The CSLs include the CFLs, which include the RLs. We will focus on the classic example $a^n b^n c^n$, which is a CSL but not a CFL (Section 3). In general, CSL recognition requires a linear-bounded automaton, a special Turing machine whose tape length is at most linear in the input size. The $\{a^n b^n c^n\}$ language is one of the simplest CSLs; it can be generated by a tree-adjoined grammar and recognized using a so-called embedded push-down automaton (Vijay-Shanker, 1992) or a finite state automaton with access to two counters that can be incremented or decremented. To our knowledge no RNN has been able to learn a CSL.

2 LSTM

We are using LSTM with forget gates and the recently introduced peephole connections (Gers & Schmidhuber, 2000). Forget gates were shown to be essential for problems involving continual or very long input strings (Gers et al., 2000). Peephole connections allow the gates to access the CEC of the same memory block. This proved to be essential for numerous symbolic and real-valued counting tasks (Gers & Schmidhuber, 2000). The LSTM variant with peepholes and forget gates used in this paper is clearly superior to traditional LSTM, and is now our method of choice.

The basic unit of an LSTM network is the *memory block* containing one or more *memory cells* and three adaptive, multiplicative gating units shared by all cells in the block (Figure 1). Each memory cell has at its

core a recurrently self-connected linear unit called the "Constant Error Carousel" (CEC) whose activation is called the cell *state*. The CECs enforce *constant* error flow and overcome a fundamental problem plaguing previous RNNs: they prevent error signals from decaying quickly as they "back in time". The adaptive gates control input and output to the cells (*input and output gate*) and learn to reset the cell's state once its contents are out of date (*forget gate*). The gates are units whose outputs multiplicatively influence connections to or from the linear unit holding the cell state. Peephole connections connect the CEC to the gates. All errors are cut off once they leak out of a memory cell or gate, although they do serve to change the incoming weights. The effect is that the CECs are the only part of the system through which errors can flow back forever, while gates etc. learn the nonlinear aspects of sequence processing. This makes LSTM's updates efficient without significantly affecting learning power: The CECs permit LSTM to bridge huge time lags (1000 discrete time steps and more) between relevant events, while traditional RNNs already fail to learn in the presence of 10 step time lags, despite requiring more complex update algorithms. Moreover, LSTM is local in space and time (Schmidhuber, 1989), and it is more efficient than other RNN algorithms such as real time recurrent learning (RTRL—Robinson & Fallside, 1987; Williams & Zipser, 1992). LSTM's computational complexity for a memory block j per time step and weight is $O(S_j)$, where S_j is the number of cells in the block. S_j is typically a small constant — in this paper it is in fact always equal to one, so that the computational complexity is $O(1)$. Back propagation through time (BPTT—Williams & Peng, 1990; Werbos, 1988) has the same computational complexity (whereas RTRL is much worse) but is not local in time. It needs to store activation values observed during sequence processing in a stack with potential unlimited size. Essentially, LSTM has the complexity of BPTT(h) (Williams & Zipser, 1992), which truncates errors after h steps, and which works as well as BPTT, according to our experience — if we did not have LSTM we would in fact use BPTT(h).

2.1 Forward Pass

The cell output, y^c, is calculated based on the current cell state s_c and four sources of input: net_c is input to the cell itself while net_{in}, net_φ and net_{out} are inputs to the input gates, forget gates, and output gates. We consider discrete time steps $t = 0, 1, 2, \ldots$. A single step involves the update of all units (forward pass) and the computation of error signals for all weights (backward pass). Throughout this paper j indexes memory blocks; v indexes memory cells in block j (with S_j cells), such that c_j^v denotes the v-th cell of the j-th memory block; w_{lm} is the weight on the connection from unit m to unit l. Index m ranges over all source units, as specified by the network topology, and index m_j ranges over all source units and also the CECs of the j-th memory block. For the gates, f_l , $l \in \{in, out, \varphi\}$ is a logistic sigmoid with range $[0, 1]$.

For each discrete time step we use a two-phase update scheme that computes (in this order):

1. (a) Input gate activation y^{in},
   (b) forget gate activation y^φ,
   (c) cell input and cell state s_c,
2. output gate activation y^{out} and cell output y^c.

Step 1a,1b. The input gate activation y^{in} and the forget gate activation y^φ are computed as follows:

$$net_{in_j}(t) = \sum_m w_{in_j m}\ y^m(t-1) + \sum_{v=1}^{S_j} w_{in_j c_j^v}\ s_{c_j^v}(t-1)\ ,\quad y^{in_j}(t) = f_{in_j}(net_{in_j}(t))\quad , \tag{1}$$

$$net_{\varphi_j}(t) = \sum_m w_{\varphi_j m}\ y^m(t-1) + \sum_{v=1}^{S_j} w_{\varphi_j c_j^v}\ s_{c_j^v}(t-1)\ ,\quad y^{\varphi_j}(t) = f_{\varphi_j}(net_{\varphi_j}(t))\quad . \tag{2}$$

The peephole connections for the input gate and the forget gate are incorporated in equation 1 and 2 by including the CECs (containing the cell states) of memory block j as source units.

Step 1c. The state of memory cell $s_c(t)$ is calculated by adding the squashed, gated input of the cell to the state at the previous time step $s_c(t-1)$ $(t > 0)$, which is multiplied (gated) by the forget gate activation $(s_{c_j^v}(0) = 0)$:

$$net_{c_j^v}(t) = \sum_m w_{c_j^v m}\ y^m(t-1)\ ,\quad s_{c_j^v}(t) = y^{\varphi_j}(t)\ s_{c_j^v}(t-1) + y^{in_j}(t)\ \ g(net_{c_j^v}(t))\quad . \tag{3}$$

Step 2. The output gate activation y^{out} is computed as follows:

$$net_{out_j}(t) = \sum_m w_{out_j m}\ y^m(t-1) + \sum_{v=1}^{S_j} w_{out_j c_j^v}\ s_{c_j^v}(t)\ ,\quad y^{out_j}(t) = f_{out_j}(net_{out_j}(t))\ . \tag{4}$$

Equation 4 includes the peephole connections for the output gate from the CECs of memory block j with the cell states $s_c(t)$, actualized in step 1. The cell output y^c is computed as follows:

$$y^{c_j^v}(t) = y^{out_j}(t)\ s_{c_j^v}(t)\ . \tag{5}$$

Finally, assuming a layered network topology with a standard input layer, a hidden layer consisting of memory blocks, and a standard output layer, the equations for the output units k are:

$$net_k(t) = \sum_m w_{km}\ y^m(t-1)\ ,\quad y^k(t) = f_k(net_k(t))\ , \tag{6}$$

where m ranges over all units feeding the output units and f_k is the output squashing function.

2.2 Gradient-Based Backward Pass

Essentially, LSTM's backward pass is an efficient fusion of slightly modified, truncated BPTT and a customized version of RTRL (for details see Hochreiter & Schmidhuber, 1997; Gers et al., 2000). We are using iterative gradient descent, minimizing an objective function E, here the usual mean squared error over all time steps and all sequences. Appendix A lists pseudo code for the entire algorithm; see (Gers et al., 2000) for a full derivation. Unlike BPTT and RTRL, LSTM's learning algorithm is local in space and time.

3 Experiments

The network sequentially observes exemplary symbol strings of a given language, presented one input symbol at a time. Following the traditional approach in the RNN literature we formulate the task as a prediction task. At any given time step the target is to predict the possible next symbols, including the "end of string" symbol T. When more than one symbol can occur in the next step *all* possible symbols have to be predicted, and none of the others.

The network sequentially observes exemplary symbol strings of a given language, presented one input symbol at a time, also referred to as input sequences. Every input sequence begins with the start symbol S. The empty string, consisting of ST only, is considered part of each language. A string is accepted when all predictions have been correct. Otherwise it is rejected.

This prediction task is equivalent to a classification task with two classes "accept" and "reject," because the system will make prediction errors for all strings outside the language. A system has learned a given language up to string size n once it is able to correctly predict all strings with size $\leq n$.

Symbols are encoded locally by d-dimensional binary vectors with only one non-zero component, where d equals the number of language symbols plus one for either the start symbol in the input or the "end of string" symbol in the output (d input units, d output units). $+1$ signifies that a symbol is set and -1 that it is not; the decision boundary for the network output is 0.0.

CFL $a^n b^n$ (Sun et al., 1993; Wiles & Elman, 1995"; Tonkes & Wiles, 1997; Rodriguez et al., 1999). Here the strings in the input sequences are of the form $a^n b^n$; input and output vectors are 3-dimensional. Prior to the first occurrence of b either a or b, or a or T at sequence beginnings, are possible in the next step. Thus, e.g., for $n=5$:

Input:	S	a	a	a	a	a	b	b	b	b	b
Target:	a/T	a/b	a/b	a/b	a/b	a/b	b	b	b	b	T

CFL $a^n b^m B^m A^n$ (Rodriguez & Wiles, 1998). The second half of a string from this palindrome or mirror language is completely predictable from the first half. The task involves an intermediate time lag of length $2m$. Input and output vectors are 5-dimensional. Prior to the first occurrence of B two symbols are possible in the next step. Thus, e.g., for $n=4, m=3$:

Input:	S	a	a	a	a	b	b	b	B	B	B	A	A	A	A
Target:	a/T	a/b	a/b	a/b	a/b	b/B	b/B	b/B	B	B	A	A	A	A	T

CSL $a^n b^n c^n$. Input and output vectors are 4-dimensional. Prior to the first occurrence of b two symbols are possible in the next step. Thus, e.g., for $n=5$:

Input:	S	a	a	a	a	a	b	b	b	b	b	c	c	c	c	c
Target:	a/T	a/b	a/b	a/b	a/b	a/b	b	b	b	b	c	c	c	c	c	T

3.1 Training and Testing

Learning and testing alternate: after each epoch (=1000 training sequences) we freeze the weights and run a test. Even when all strings are processed correctly during training, it is necessary to test again with frozen weights once all weight changes have been executed. Apart from ensuring the learning of the training set the test also determines generalization performance, which we did not optimize by using, say, a validation set.

Training and test sets incorporate all legal strings up to a given length: $2n$ for $a^n b^n$, $3n$ for $a^n b^n c^n$ and $2(n+m)$ for $a^n b^m B^m A^n$. Training strings are presented in random order. Only exemplars from the class "accept" are presented. Training is stopped once all training sequences have been accepted, or after at most 10^7 training sequences. The *generalization set* is the largest accepted test set.
Weight changes are made after each sequence. We apply the momentum algorithm (Plaut, Nowlan, & Hinton, 1986) with learning rate α is 10^{-5} and momentum parameter 0.99. All results are averages over 10 independently trained networks with different weight initializations (these 10 initializations are identical for each experiment).

CFL $a^n b^n$. We study training sets with $n \in \{1,..,N\}$. We test all sets with $n \in \{1,..,M\}$ and $M \in \{N,..,1000\}$ (sequences of length ≤ 2000).

CFL $a^n b^m B^m A^n$. We use two training sets: a) The same set as used by Rodriguez and Wiles (1999) : $n \in \{1,..,11\}$, $m \in \{1,..,11\}$ with $n+m \leq 12$ (sequences of length ≤ 24). b) The set given by $n \in \{1,..,11\}$, $m \in \{1,..,11\}$ (sequences of length ≤ 44). We test all sets with $n \in \{1,..,M\}$, $m \in \{1,..,M\}$ and $M \in \{11,..,50\}$ (sequences of length ≤ 200).

CSL $a^n b^n c^n$. We study two kinds of training sets: a) with $n \in \{1,..,N\}$ and b) with $n \in \{N-1, N\}$. Case b) asks for a major generalization step that seems almost impossible at first glance: Given very similar training sequences whose sizes differ by at most 2, learn to process sequences of arbitrary size! We test all sets with $n \in \{L,..,M\}$, $L \in \{1,..,N-1\}$ and $M \in \{N,..,500\}$ (sequences of length ≤ 1500).

3.2 Network Topology and Experimental Parameters

The input units are fully connected to a hidden layer consisting of memory blocks with 1 cell each. The cell outputs are fully connected to the cell inputs, to all gates, and to the output units, which also have direct "shortcut" connections from the input units (Figure 2). For each task we selected the topology with minimal number of memory blocks that solved the task without extensive parameter optimization. Larger topologies never led to disadvantages except for an increase in computational complexity.

All gates, the cell itself and the output unit are biased. The bias weights to input gate, forget gate and output gate are initialized with -1.0, $+2.0$ and -2.0, respectively. Although not critical, these values have been found empirically to work well; we use them for all our experiments. With this bias initialization, cell states are initially close to zero, and, as training progresses, the biases become progressively less negative, allowing the serial activation of cells as active participants in the network computation. The forget gates start off closed, so that the cells initially remember everything. We also tried different bias configurations; the results were qualitatively the same, which supports our claim that precise initialization is not critical — see also Hochreiter and Schmidhuber (1997) and Gers, Schmidhuber, and Cummins (2000) for additional evidence in this vein. All other weights are initialized randomly in the range $[-0.1, 0.1]$. The cell's input squashing function g is the identity function. The squashing function of the output units is a sigmoid function with the range $[-2, 2]$.

CFL $a^n b^n$. We use one memory block (with one cell). With peephole connections there are 38 adjustable weights (3 peephole, 28 unit-to-unit and 7 bias connections).

CFL $a^n b^m B^m A^n$. We use two blocks with one cell each, resulting in 110 adjustable weights (6 peephole, 91 unit-to-unit and 13 bias connections).

CSL $a^n b^n c^n$. We use the same topology as for the $a^n b^m B^m A^n$ language, but with 4 input and output units instead of 5, resulting in 90 adjustable weights (6 peephole, 72 unit-to-unit and 12 bias connections).

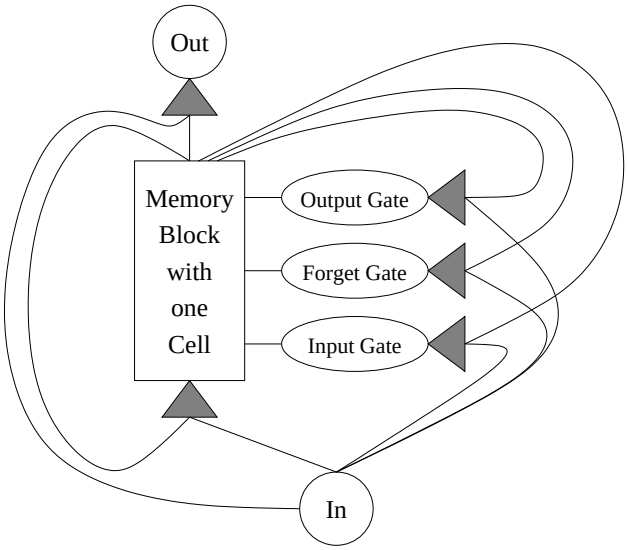

Figure 2: Three-layer LSTM topology with a single input and output. Recurrence is limited to the hidden layer, consisting here of a single LSTM memory block with a single cell. All 10 "unit-to-unit" connections are shown (but bias and peephole connections are not).

Table 1: Previous results for the CFL $a^n b^n$, showing (from left to right) the number of hidden units or state units, the values of n used during training, the number of training sequences, the number of found solutions/trials and the largest accepted test set.

Reference	Hidden Units	Train. Set $[n]$	Train. Str. $[10^3]$	Sol./ Tri.	Best Test $[n]$
(Sun et al., 1993)[1]	5	1,.., 160	13.5	1/1	1,.., 160
(Wiles & Elman, 1995")	2	1,.., 11	2000	4/20	1,.., 18
(Tonkes & Wiles, 1997)	2	1,.., 10	10	13/100	1,.., 12
(Rodriguez et al., 1999)[2]	2	1,.., 11	267	8/50	1,.., 16

3.3 Previous results

CFL $a^n b^n$. Published results on the $a^n b^n$ language are summarized in Table 1. RNNs trained with plain BPTT tend to learn to just reproduce the input (Wiles & Elman, 1995"; Tonkes & Wiles, 1997; Rodriguez et al., 1999). Sun et al. (1993) used a highly specialized architecture, the "neural pushdown automaton", which also did not generalize well (Sun et al., 1993; Das, Giles, & Sun, 1992).

CFL $a^n b^m B^m A^n$. Rodriguez and Wiles (1998) used BPTT-RNNs with 5 hidden nodes. After training with $51 \cdot 10^3$ strings with $n + m \leq 12$ (sequences of length ≤ 24), most networks generalized on longer off-training set strings. The best network generalized to sequences up to length 36 ($n = 9$,$m = 9$). But none of them learned the complete training set.

CSL $a^n b^n c^n$. To our knowledge no previous RNN ever learned a CSL.

3.4 LSTM Results

CFL $a^n b^n$. 100% solved for all training sets (Table 2). Small training sets ($n \in \{1,..,10\}$) were already sufficient for perfect generalization up to the tested maximum: $n \in \{1,..,1000\}$. Note that long sequences of this kind require very stable, finely tuned control of the network's internal counters (Casey, 1996).
This performance is much better than with previous approaches, where the largest set was learned by the specially designed neural push-down automaton (Sun et al., 1993; Das et al., 1992): $n \in \{1,..,160\}$. The latter, however, required training sequences of the same length as the test sequences. From the training set with

[1] Sun's training set was augmented stepwise by sequences misclassified during testing, and in the final accepted set n was in $\{1,..,20\}$ except for 20 random sequences up to length $n = 160$ (the exact generalization performance was unclear).

[2] Applying brute force search to the weights of the best network of Rodriguez et al. (1999) further improves performance to acceptance up to $n = 28$.

Train. Set [n]	Train. Str. [10^3]	% Sol.	Generalization Set [n]
1,.., 10	22 (19)	100	1,.., 1000 (1,.., 118)
1,.., 20	18 (19)	100	1,.., 587 (1,.., 148)
1,.., 30	16 (19)	100	1,.., 1000 (1,.., 408)
1,.., 40	25 (28)	100	1,.., 1000 (1,.., 628)
1,.., 50	42 (40)	100	1,.., 767 (1,.., 430)

Table 2: Results for the $a^n b^n$ language, showing (from left to right) the values for n used during training, the average number of training sequences until best generalization was achieved, the percentage of correct solutions and the best generalization (average over all networks given in parenthesis).

Train. Set [n]	Train. Str. [10^3]	% Sol.	Generalization Set [n]
1,.., 10	54 (62)	100	1,.., 52 (1,.., 28)
1,.., 20	28 (43)	100	1,.., 160 (1,.., 66)
1,.., 30	37 (43)	100	1,.., 228 (1,.., 91)
1,.., 40	51 (48)	90	1,.., 500 (1,.., 120)
1,.., 50	60 (94)	100	1,.., 500 (1,.., 409)
10, 11	24 (78)	100	9,.., 12 (10,.., 11)
20, 21	829 (626)	40	10,.., 27 (17,.., 23)
30, 31	42 (855)	30	29,.., 34 (29,.., 32)
40, 41	854 (1597)	40	20,.., 57 (35,.., 45)
50, 51	32 (621)	60	43,.., 57 (47,.., 55)

Table 3: Results for the $a^n b^n c^n$ language, showing (from left to right) the values for n used during training, the average number of training sequences until best generalization was achieved, the percentage of correct solutions and the best generalization (average over all networks in parenthesis).

$n \in \{1,..,10\}$ LSTM generalized to $n \in \{1,..,1000\}$, whereas the best previous result (see Table 1) generalized only to $n \in \{1,..,18\}$ (even with a slightly larger training set: $n \in \{1,..,11\}$). In contrast to Tonkes and Wiles (1997), we did not observe our networks forgetting solutions as training progresses. So unlike all previous approaches, LSTM reliably finds solutions that generalize well.

The fluctuations in generalization performance for different training sets in Table 2 may be due to the fact that we did not optimize generalization performance by using a validation set. Instead we simply stopped each epoch (= 1000 sequences) once the training set was learned.

CFL $a^n b^m B^m A^n$. Training set a): 100% solved; after $29 \cdot 10^3$ training sequences the best network of 10 generalized to at least $n, m \in \{1,..,22\}$ (all strings up to a length of 88 symbols processed correctly); the average generalization set was the one with $n, m \in \{1,..,16\}$ (all strings up to a length of 64 symbols processed correctly), learned after $25 \cdot 10^3$ training sequences on average.

Training set b): 100% solved; after $26 \cdot 10^3$ training sequences the best network generalized to at least $n, m \in \{1,..,23\}$ (all strings until a length of 92 symbols processed correctly). The average generalization set was the one with $n, m \in \{1,..,17\}$ (all strings until a length of 68 symbols processed correctly), learned after $82 \cdot 10^3$ training sequences on average. Unlike the previous approach of Rodriguez and Wiles (1998), LSTM easily learns the complete training set and reliably finds solutions that generalize well.

CSL $a^n b^n c^n$. LSTM learns 4 of the 5 training sets in 10 out of 10 trials (only 9 out of 10 for the training set with $n \in \{1,..,40\}$) and generalizes well (Table 3). Small training sets ($n \in \{1,..,40\}$) were already sufficient for perfect generalization up to the tested maximum: $n \in \{1,..,500\}$, that is, sequences of length up to 1500. Even in absence of any short training sequences ($n \in \{N-1, N\}$) LSTM learned well (see bottom half of Table 3).

We also modified the training procedure, by presenting each exemplary string without providing all possible next symbols as targets, but only the symbol that actually occurs in the current exemplar. This led to slightly longer training durations, but did not significantly change the results.

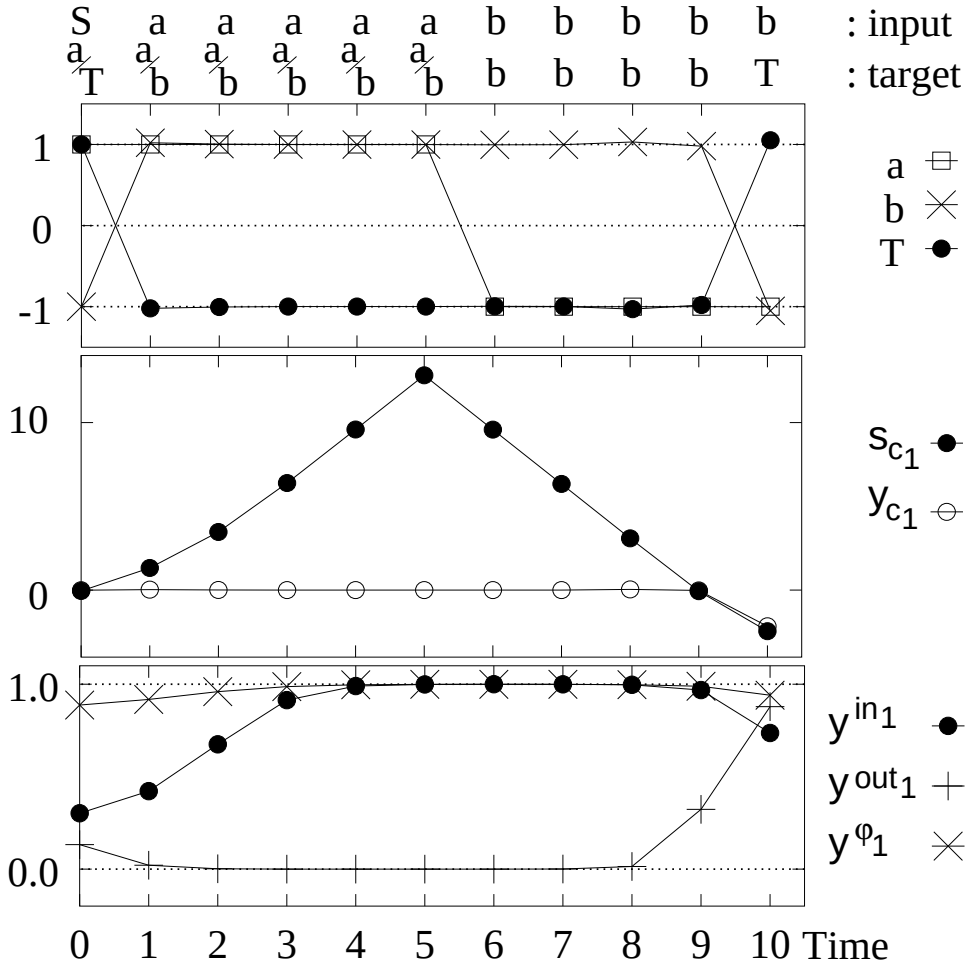

Figure 3: CFL $a^n b^n$ ($n = 5$): Test run with network solutions. Top: Network output y_k. Middle: Cell state s_c and cell output y_c. Bottom: Activations of the gates (input gate y_{in}, forget gate y_φ and output gate y_{out}).

3.5 Analysis

How do the solutions discovered by LSTM work?

CFL $a^n b^n$. Figure 3 shows a test run with a network solution for $n=5$.

The cell state s_c increases while a symbols are fed into the network, then decreases (with the same step size) while b symbols are fed in. At sequence beginnings (when the first a symbols are observed), however, the step size is smaller due to the closed input gate, which is triggered by s_c itself. This results in "overshooting" the initial value of s_c at the end of a sequence, which in turn triggers the opening of the output gate, which in turn leads to the prediction of the sequence termination.

CFL $a^n b^m B^m A^n$. The behavior of a typical network solution is shown in Figure 4.

The network learned to establish and control two counters. The two symbol pairs (a, A) and (b, B) are treated separately by two different cells, c_2 and c_1, respectively. Cell c_2 tracks the difference between the number of observed a and A symbols. It opens only at the end of a string, where it predicts the final T. Cell c_1 treats the embedded $b^m B^m$ substring in a similar way. While values are stored and manipulated within a cell, the output gate remains closed. This prevents the cell from disturbing the rest of the network and also protects its CEC against incoming errors.

CSL $a^n b^n c^n$. The network solutions use a combination of two counters, instantiated separately in the two memory blocks (Figure 5).

Here the second cell counts up, given an a input symbol. It counts down, given a b. A c in the input causes the input gate to close and the forget gate to reset the cell state s_c. The second memory block does the same for b, c, and a, respectively. The opening of output gate of the first block indicates the end of a string (and the prediction of the last T), triggered via its peephole connection.

Why does the network not generalize for short strings when using only two training strings as for the $a^n b^n c^n$ language (see Table 3)? The gate activations in Figure 5 show that activations slightly drift even when the input stays constant. Solutions take this state drift into account, and will not work without it or with

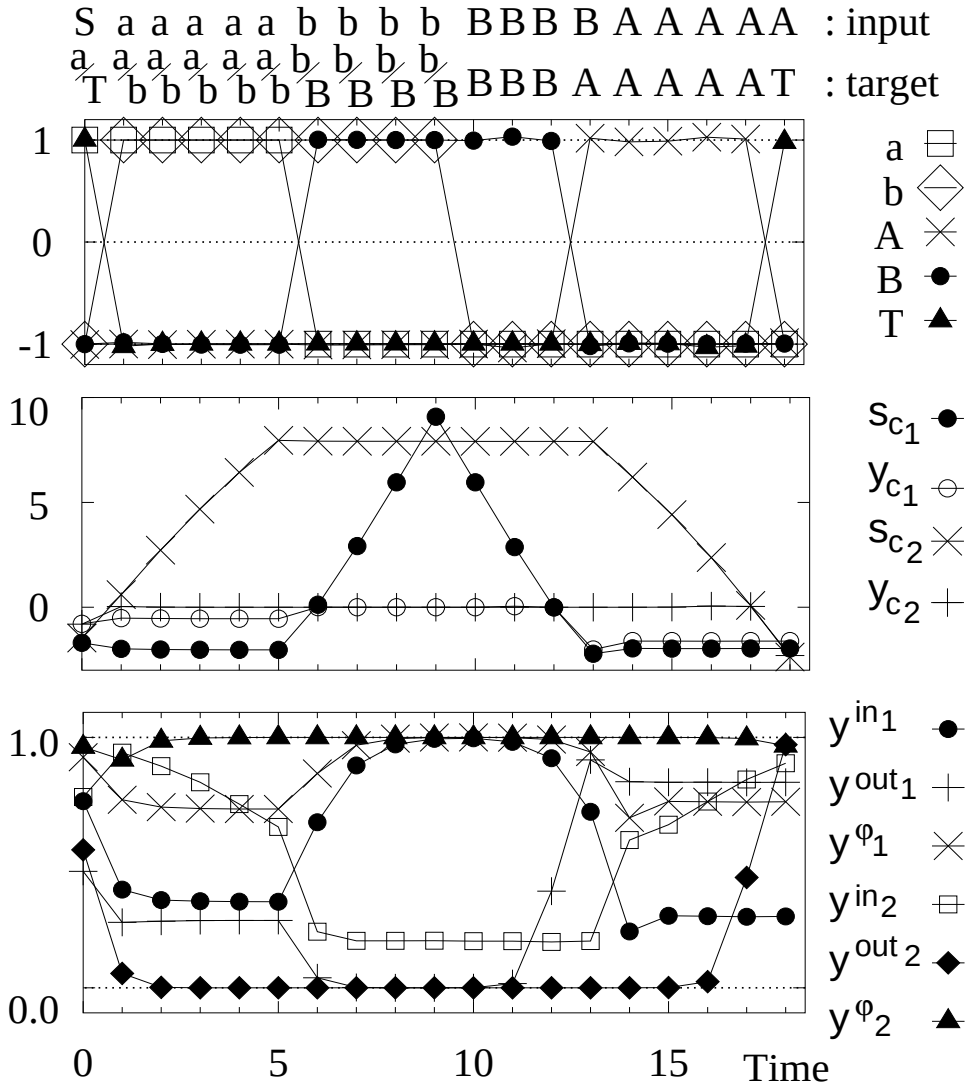

Figure 4: CFL $a^n b^m B^m A^n$ $(n=5, m=4)$: Test run with network solution. Top: Network output y_k. Middle: Cell state s_c and cell output y_c. Bottom: Activations of the gates (input gate y_{in}, forget gate y_φ and output gate y_{out}).

too much of it, as in the case when the sequences are much shorter or longer than the few observed training examples. This imposes a limit on generalization in *both* directions (towards longer and shorter strings). We found solutions with less drift to generalize better.

Further improvements. Even better results can be obtained through increased training time and stepwise reduction of the learning rate, as done in (Rodriguez et al., 1999). The distribution of lengths of sequences in the training set also affects learning speed and generalization. A set containing more long sequences improves generalization for longer sequences. Omitting the sequence with $n = 1$ (and $m = 1$), typically the last one to be learned, has the same effect. Training sets with many short and many long sequences are learned more quickly than uniformly distributed ones.

Related tasks. The $(ba^k)^n$ regular language is related to $a^n b^n$ in the sense that it requires to learn a counter, but the counter never needs counting down. This task is equivalent to the "Continual Spike Timing" task (Gers & Schmidhuber, 2000) learned by LSTM for $k = 50$ with $n \geq 1000$. A hand-made, hardwired solution (no learning) of a second order RNN worked for values of k until 120 (Steijvers & Grunwald, 1996).

For all three tasks peephole connections are mandatory. The output gates remain closed for substantial time periods during each input sequence presentation (compare Figures 3, 4 and 5); the end of such a period is always triggered via peephole connections.

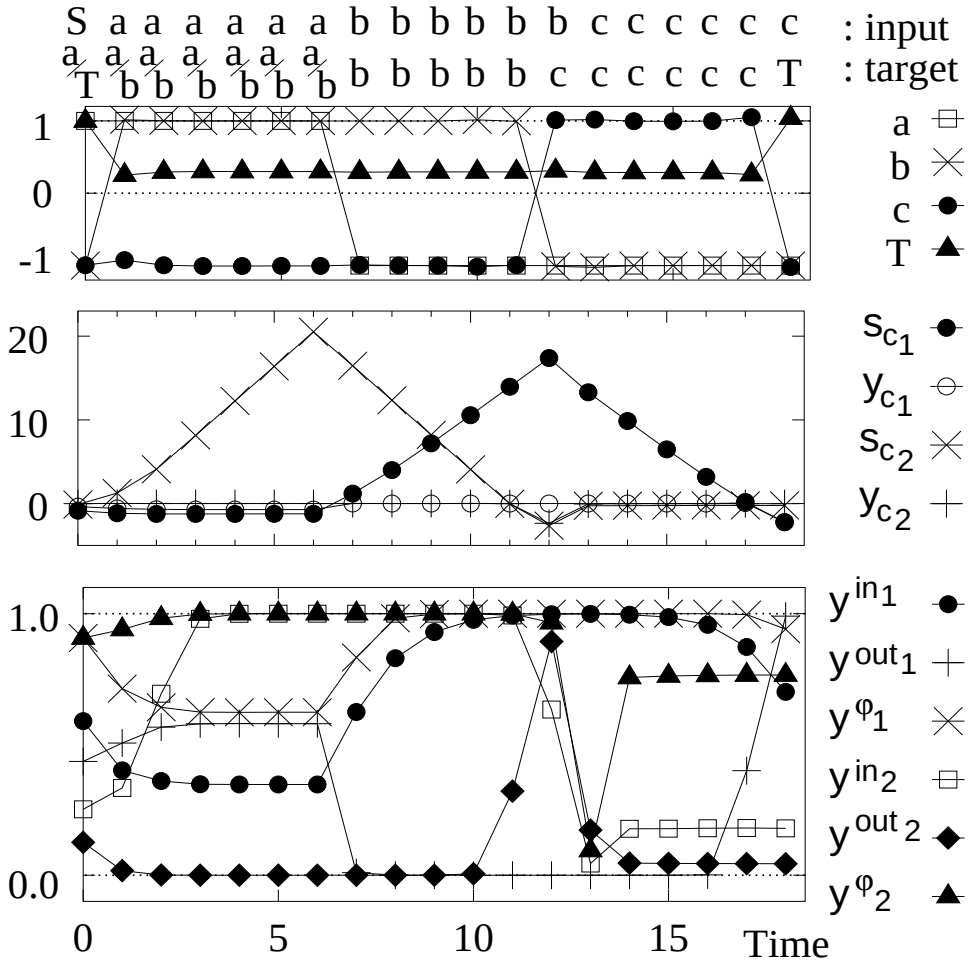

Figure 5: CSL $a^n b^n c^n$ $(n = 5)$: Test run with network solution (the system scales up to sequences of length 1000 and more). Top: Network output y_k. Middle: Cell state s_c and cell output y_c. Bottom: Activations of the gates (input gate y_{in}, forget gate y_φ and output gate y_{out}).

4 Conclusion

We found that Long Short-Term Memory (LSTM) clearly outperforms previous RNNs not only on regular language benchmarks (according to previous research) but also on context free language (CFL) benchmarks. It learns faster and generalizes better. LSTM also is the first RNN to learn a context sensitive language.

Although CFLs like those studied in this paper may also be learnable by certain discrete symbolic grammar learning algorithms (SGLAs) (Sakakibara, 1997; Lee, 1996; Osborne & Briscoe, 1997), the latter exhibit much more task-specific bias, and are not designed to solve numerous other sequence processing tasks involving noise, real-valued inputs / internal states, and continuous output trajectories, which LSTM solves easily (Hochreiter & Schmidhuber, 1997; Gers et al., 2000).

Our findings reinforce the perception that LSTM is a very general and promising adaptive sequence processing device, with a wider field of potential applications than alternative RNNs.

Acknowledgment

This work was supported by SNF grant 2100-49'144.96 "Long Short-Term Memory."

A Peephole LSTM with Forget Gates in Pseudo-code

init network:

reset: CECs: $s_{c_j^v} = \hat{s}_{c_j^v} = 0$; **partials:** $dS = 0$; **activations:** $y = \hat{y} = 0$;

forward pass:

input units: $y =$ `current external input;`

roll over: activations: $\hat{y} = y$; **cell states:** $\hat{s}_{c_j^v} = s_{c_j^v}$;

`loop over memory blocks, indexed` j {

Step 1a: input gates (1):

$net_{in_j} = \sum_m w_{in_j m} \ \hat{y}^m + \sum_{v=1}^{S_j} w_{in_j c_j^v} \ \hat{s}_{c_j^v}$; $\quad y^{in_j} = f_{in_j}(net_{in_j})$;

Step 1b: forget gates (2):

$net_{\varphi_j} = \sum_m w_{\varphi_j m} \ \hat{y}^m + \sum_{v=1}^{S_j} w_{\varphi_j c_j^v} \ \hat{s}_{c_j^v}$; $\quad y^{\varphi_j} = f_{\varphi_j}(net_{\varphi_j})$;

Step 1c: CECs, i.e the cell states (3):

`loop over the` S_j `cells in block` j`, indexed` v {

$net_{c_j^v} = \sum_m w_{c_j^v m} \ \hat{y}^m$; $\quad s_{c_j^v} = y^{\varphi_j} \ \hat{s}_{c_j^v} + y^{in_j} \ g(net_{c_j^v})$; }

Step 2:

output gate activation: (4):

$net_{out_j} = \sum_m w_{out_j m} \ \hat{y}^m + \sum_{v=1}^{S_j} w_{out_j c_j^v} \ s_{c_j^v}$; $\quad y^{out_j} = f_{out_j}(net_{out_j})$;

cell outputs (5):

`loop over the` S_j `cells in block` j`, indexed` v { $y^{c_j^v} = y^{out_j} \ s_{c_j^v}$; }

} `end loop over memory blocks`

output units (6): $net_k = \sum_m w_{km} \ y^m$; $\quad y^k = f_k(net_k)$;

partial derivatives:

`loop over memory blocks, indexed` j {

`loop over the` S_j `cells in block` j`, indexed` v {

cells, ($dS_{cm}^{jv} := \frac{\partial s_{c_j^v}}{\partial w_{c_j^v m}}$):

$dS_{cm}^{jv} = dS_{cm}^{jv} \ y^{\varphi_j} + g'(net_{c_j^v}) \ y^{in_j} \ \hat{y}^m$;

input gates, ($dS_{in,m}^{jv} := \frac{\partial s_{c_j^v}}{\partial w_{in_j m}}$, $dS_{in,c_j^{v'}}^{jv} := \frac{\partial s_{c_j^v}}{\partial w_{in_j c_j^{v'}}}$):

$dS_{in,m}^{jv} = dS_{in,m}^{jv} \ y^{\varphi_j} + g(net_{c_j^v}) \ f'_{in_j}(net_{in_j}) \ \hat{y}^m$;

`loop over peephole connections from all cells, indexed` v' {

$dS_{in,c_j^{v'}}^{jv} = dS_{in,c_j^{v'}}^{jv} \ y^{\varphi_j} + g(net_{c_j^v}) \ f'_{in_j}(net_{in_j}) \ \hat{s}_c^{v'}$; }

forget gates, ($dS_{\varphi m}^{jv} := \frac{\partial s_{c_j^v}}{\partial w_{\varphi_j m}}$, $dS_{\varphi c_j^{v'}}^{jv} := \frac{\partial s_{c_j^v}}{\partial w_{\varphi_j c_j^{v'}}}$):

$dS_{\varphi m}^{jv} = dS_{\varphi m}^{jv} \ y^{\varphi_j} + \hat{s}_{c_j^v} \ f'_{\varphi_j}(net_{\varphi_j}) \ \hat{y}^m$;

`loop over peephole connections from all cells, indexed` v' {

$dS_{\varphi c_j^{v'}}^{jv} = dS_{\varphi c_j^{v'}}^{jv} \ y^{\varphi_j} + \hat{s}_{c_j^v} \ f'_{\varphi_j}(net_{\varphi_j}) \ \hat{s}_c^{v'}$; }

} } `end loops over cells and memory blocks`

backward pass (if error injected):

errors and δs:

injection error: $e_k = t^k - y^k$;

δs of output units: $\delta_k = f'_k(net_k)\ e_k$;

`loop over memory blocks, indexed` j {

δs of output gates:

$\delta_{out_j} = f'_{out_j}(net_{out_j})\ \left(\sum_{v=1}^{S_j}\ s_{c_j^v} \sum_k w_{kc_j^v}\ \delta_k\right)$;

internal state error:

`loop over the` S_j `cells in block` j`, indexed` v {

$e_{s_{c_j^v}} = y^{out_j}\ \left(\sum_k w_{kc_j^v}\ \delta_k\right)$; }

} `end loop over memory blocks`

weight updates:

output units: $\Delta w_{km} = \alpha\ \delta_k\ y^m$;

`loop over memory blocks, indexed` j {

output gates:

$\Delta w_{out,m} = \alpha\ \delta_{out}\ \hat{y}^m$; $\Delta w_{out,c_j^v} = \alpha\ \delta_{out}\ s_{c_j^v}$;

input gates:

$\Delta w_{in,m} = \alpha \sum_{v=1}^{S_j}\ e_{s_{c_j^v}}\ dS^{jv}_{in,m}$;

`loop over peephole connections from all cells, indexed` v' {

$\Delta w_{in,c_j^{v'}} = \alpha \sum_{v=1}^{S_j}\ e_{s_{c_j^v}}\ dS^{jv}_{in,c_j^{v'}}$; }

forget gates:

$\Delta w_{\varphi m} = \alpha \sum_{v=1}^{S_j}\ e_{s_{c_j^v}}\ dS^{jv}_{\varphi m}$;

`loop over peephole connections from all cells, indexed` v' {

$\Delta w_{\varphi c_j^{v'}} = \alpha \sum_{v=1}^{S_j}\ e_{s_{c_j^v}}\ dS^{jv}_{\varphi c_j^{v'}}$; }

cells:

`loop over the` S_j `cells in block` j`, indexed` v {

$\Delta w_{c_j^v m} = \alpha\ e_{s_{c_j^v}}\ dS^{jv}_{cm}$; };

} `end loop over memory blocks`

References

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157-166.

Blair, A. D., & Pollack, J. B. (1997). Analysis of dynamical recognizers. *Neural Computation*, *9*(5), 1127–1142.

Casey, M. P. (1996). The dynamics of discrete-time computation, with application to recurrent neural networks and finite state machine extraction. *Neural Computation*, *8*(6), 1135–1178.

Das, S., Giles, C., & Sun, G. (1992). Learning context-free grammars: Capabilities and limitations of a recurrent neural network with an external stack memory. In *Proceedings of The Fourteenth Annual Conference of the Cognitive Science Society* (pp. 791–795). San Mateo, CA: Morgan Kaufmann Publishers.

Gers, F. A., & Schmidhuber, J. (2000). Recurrent nets that time and count. In *Proc. IJCNN'2000, Int. Joint Conf. on Neural Networks.* Como, Italy.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471.

Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München.* (See www7.informatik.tu-muenchen.de/~hochreit)

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

Kalinke, Y., & Lehmann, H. (1998). Computation in recurrent neural networks: From counters to iterated function systems. In G. Antoniou & J. Slaney (Eds.), *Advanced Topics in Artificial Intelligence, Proceedings of the 11th Australian Joint Conference on Artificial Intelligence* (Vol. 1502). Berlin,Heidelberg: Springer.

Lee, L. (1996). *Learning of context-free languages: A survey of the literature* (Tech. Rep. No. TR-12-96). Center for Research in Computing Technology, Harvard University, Cambridge, Massachusetts.

Maass, W., & Orponen, P. (1998). On the effect of analog noise in discrete-time analog computations. *Neural Computation*, *10*(5), 1071–1095.

Maass, W., & Sontag, E. D. (1999). Analog neural nets with Gaussian or other common noise distribution cannot recognize arbitrary regular languages. *Neural Computation*, *11*(3), 771–782.

Osborne, M., & Briscoe, E. (1997). Learning stochastic categorial grammars. In *Proceedings of the Assoc. for Comp. Linguistics, Comp. Nat. Lg.Learning (CoNLL97) Workshop* (pp. 80–87). Madrid. (http://citeseer.nj.nec.com/osborne97learning.html)

Pearlmutter, B. A. (1995). Gradient calculations for dynamic recurrent neural networks: A survey. *IEEE Transactions on Neural Networks*, *6*(5), 1212-1228.

Plaut, D. C., Nowlan, S. J., & Hinton, G. E. (1986). *Experiments on learning back propagation* (Tech. Rep. Nos. CMU–CS–86–126). Pittsburgh, PA: Carnegie–Mellon University.

Robinson, A. J., & Fallside, F. (1987). *The utility driven dynamic error propagation network* (Tech. Rep. No. CUED/F-INFENG/TR.1). Cambridge University Engineering Department.

Rodriguez, P., & Wiles, J. (1998). Recurrent neural networks can learn to implement symbol-sensitive counting. In *Advances in Neural Information Processing Systems* (Vol. 10, p. 87-93). The MIT Press.

Rodriguez, P., Wiles, J., & Elman, J. (1999). A recurrent neural network that learns to count. *Connection Science*, *11*(1), 5-40.

Sakakibara, Y. (1997). Recent advances of grammatical inference. *Theoretical Computer Science*, *185*(1), 15–45.

Schmidhuber, J. (1989). The Neural Bucket Brigade, a local learning algorithm for dynamic feedforward and recurrent networks. *Connection Science*, *1*(4), 403-412.

Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Computation*, *4*(2), 234-242.

Siegelmann, H. (1992). *Theoretical foundations of recurrent neural networks.* Unpublished doctoral dissertation, Rutgers, New Brunswick Rutgers, The State of New Jersey.

Siegelmann, H. T., & Sontag, E. D. (1991). Turing computability with neural nets. *Applied Mathematics Letters*, *4*(6), 77–80.

Steijvers, M., & Grunwald, P. (1996). A recurrent network that performs a contextsensitive prediction task. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* Erlbaum.

Sun, G. Z., Giles, C. L., Chen, H. H., & Lee, Y. C. (1993). *The neural network pushdown automaton: Model, stack and learning simulations* (Technical Report No. CS-TR-3118). University of Maryland, College Park.

Tonkes, B., & Wiles, J. (1997). Learning a context-free task with a recurrent neural network: An analysis of stability. In *Proceedings of the Fourth Biennial Conference of the Australasian Cognitive Science Society.*

Vijay-Shanker, K. (1992). Using descriptions of trees in a tree adjoining grammar. *Computational Linguistics*, *18*(4), 481–517.

Werbos, P. J. (1988). Generalisation of backpropagation with application to a recurrent gas market model. *Neural Networks*, *1*, 339–356.

Wiles, J., & Elman, J. (1995"). Learning to count without a counter: A case study of dynamics and activation landscapes in recurrent networks. In *In Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. pages 482 – 487). Cambridge, MA: MIT Press.

Williams, R. J., & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories [Letter]. *Neural Computation*, *2*(4), 490–501.

Williams, R. J., & Zipser, D. (1992). Gradient-based learning algorithms for recurrent networks and their computational complexity. In Y. Chauvin & D. E. Rumelhart (Eds.), *Back-propagation: Theory, Architectures and Applications* (pp. 433–486). Hillsdale, NJ: Erlbaum.

Zeng, Z., Goodman, R., & Smyth, P. (1994"). Discrete recurrent neural networks for grammatical inference. *IEEE Transactions on Neural Networks*, *5*(2).