Long Short Term Memory

Marco Marini

May 30, 2016

Contents

Abstract

Long Short Term Memory (LSTM)

1 General

The LSTM is a recurrent neural network built up by LSTM cell.

Several neurons compose the LSTM cell:

1. input neuron
2. input gate
3. CEC (Constant Error Carrousels)
4. forget gate
5. output activation function
6. output gate

The input neuron and each gate receive values from input signals and the previuos outputs (RNN).

The output signal of LSTM is

$$y(t) = y^h(t) \cdot y^f(t)$$

where

$$y^f(t) = f\left(z^f(t)\right)$$
$$z^f(t) = \sum w_i^f I_i(t)$$
$$y^h(t) = h\left(z^h(t)\right)$$
$$z^h(t) = y^g(t) \cdot y^s(t) + z^h(t-1) \cdot y^r(t)$$
$$y^r(t) = r\left(z^r(t)\right)$$
$$z^r(t) = \sum w_i^r I_i(t)$$
$$y^g(t) = g\left(z^g(t)\right)$$
$$z^g(t) = \sum w_i^g I_i(t)$$
$$y^s(t) = s\left(z^s(t)\right)$$
$$z^s(t) = \sum w_i^s I_i(t)$$

2 Back propagation

Let's now derive the back propagation error

$$\frac{\partial y(t)}{\partial w_i^f} = y^h(t) \frac{\partial y^f(t)}{\partial w_i^f}$$

$$\frac{\partial y^f(t)}{\partial w_i^f} = \frac{\partial f\left(z^f(t)\right)}{\partial z^f(t)} \cdot \frac{\partial z^f(t)}{\partial w_i^f} = f'(t) \cdot I_i(t)$$

$$\begin{gathered}\tfrac{\partial y(t)}{\partial w_i^r} = y^f(t) \tfrac{\partial y^h(t)}{\partial w_i^r} = \\ = y^f(t) \cdot h'(t) \tfrac{\partial z^h(t)}{\partial w_i^r} = \\ = y^f(t) \cdot h'(t) z^h(t-1) \cdot \tfrac{\partial y^r(t)}{\partial w_i^r} = \\ = y^f(t) \cdot h'(t) z^h(t-1) \cdot r'(t) \cdot \tfrac{\partial z^r(t)}{\partial w_i^r} \\ = y^f(t) \cdot h'(t) z^h(t-1) \cdot r'(t) \cdot I_i(t)\end{gathered}$$

$$\begin{gathered}\tfrac{\partial y(t)}{\partial w_i^s} = y^f(t) \tfrac{\partial y^h(t)}{\partial w_i^s} = \\ = y^f(t) \cdot h'(t) \tfrac{\partial z^h(t)}{\partial w_i^s} = \\ = y^f(t) \cdot h'(t) \cdot y^g(t) \tfrac{\partial y^s(t)}{\partial w_i^s} = \\ = y^f(t) \cdot h'(t) \cdot y^g(t) s'(t) \cdot \tfrac{\partial z^s(t)}{\partial w_i^s} = \\ = y^f(t) \cdot h'(t) \cdot y^g(t) \cdot s'(t) \cdot I_i(t)\end{gathered}$$

$$\frac{\partial y(t)}{\partial w_i^f} = y^h(t) \cdot f'(t) \cdot I_i(t) \quad (1)$$

$$\frac{\partial y(t)}{\partial w_i^r} = y^f(t) \cdot h'(t) z^h(t-1) \cdot r'(t) \cdot I_i(t) \quad (2)$$

$$\frac{\partial y(t)}{\partial w_i^s} = y^f(t) \cdot h'(t) \cdot y^g(t) \cdot s'(t) \cdot I_i(t) \quad (3)$$

$$\frac{\partial y(t)}{\partial w_i^g} = y^f(t) \cdot h'(t) \cdot y^s(t) \cdot g'(t) \cdot I_i(t) \quad (4)$$

3 Error

Let be

$$X_{ijk}$$

the input signals $j = 1, \cdots, m$ of sample $i = 1, \cdots, n$ at time $k = 1, \cdots, t_i$ and

$$Y_{ik}$$

the expected output signal of sample $i = 1, \cdots, n$ at time $k = 1, \cdots, t_i$

The total error of network is

$$\delta = \frac{1}{2} \sum_{ik} \left[Y_{ik} - y(k, I_{ijk}) \right]^2$$

4 Test

The test is composed by a set of samples with 2 input signals.

The output is a JK memory latch

J	K	Y
0	0	Y
0	1	0
1	0	1
1	1	!Y

The activation functions are

$$\begin{aligned} f(x) &= tanh(x) \\ g(x) &= tanh(x) \\ h(x) &= x \\ r(x) &= tanh(x) \\ s(x) &= \tanh(x) \end{aligned}$$

Gradients are:

$$\frac{\partial y(t)}{\partial w_i^f} = y^h(t) \cdot \left[1 - y^f(t)\right]^2 \cdot I_i(t) \tag{5}$$

$$\frac{\partial y(t)}{\partial w_i^r} = y^f(t) \cdot z^h(t-1) \cdot [1 - y^r(t)]^2 \cdot I_i(t) \tag{6}$$

$$\frac{\partial y(t)}{\partial w_i^s} = y^f(t) \cdot y^g(t) \cdot [1 - y^s(t)]^2 \cdot I_i(t) \tag{7}$$

$$\frac{\partial y(t)}{\partial w_i^g} = y^f(t) \cdot y^s(t) \cdot [1 - y^g(t)]^2 \cdot I_i(t) \tag{8}$$