Hochreiter, LSTM

LSTMs computational complexity per time step and weight is O(1).

Recurrent networks can use their feedback connections to store representations of recent input events in form of activations (“short term memory” as opposed to “long term memory”, embodied by slowly changing weights).

With conventional BPTT or RTRL error signals flowing backwards in time either blow up or vanish.

The temporal evolution of the back propagated error exponentially depends on the size of the weights.

LSTM enforces constant (neither exploding nor vanishing) error flow through internal states of units. Gradient computation is truncated at certain architecture specific points, does not affect long term error flow.