Contents

1 The unreasonable effectiveness of the forget gate

1.1 Motivations

Whether all the gates of the LSTM network are necessary.

- ***Conflicting updates (gradients)***
  1. having a single weight in the RNN creates conflicting updates.
  2. the long and short-range error act on the same weight at each step, and with sigmoid activated units, this results in the gradients vanishing faster thant the weights can grow.
- ***The initialization of forget gates***
  - Problems:
    * most applications initialize the LSTM weights with small random weights
    * the forget gate is set to 0.5
    * introduces a vanishing gradient with a factor of 0.5 per timestep
  - Solution
    * initialize all bias to zeros
    * initialize the forget bias $\mathbf{b}_f$ to a large value such as 1 or 2.

1.2 Just Another Network (JANET)

1. ***Couple*** the input and forget modulation
   - It *seems* sensible to have the accumulation and deletion of information to be related.
2. Remove the `tanh` activation of $\mathbf{h}_t$
   - the `tanh` activation shrinks the gradients
   - weight matrix U_* can accommodate values beyond $[-1, 1]$

$$\mathbf{f}_t = \sigma(\mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{W}_f\mathbf{x}_t + \mathbf{x_f})$$
$$\mathbf{c}_t = \mathbf{f} \odot \mathbf{c}_{t-1} + (\mathbf{1} - \mathbf{f}_t) \odot \text{tanh}(\mathbf{U_c}\mathbf{h}_{t-1} + \mathbf{W}_c\mathbf{x}_t + \mathbf{b}_c)$$
$$\mathbf{h}_t = \mathbf{c}_t$$

3. Allow slightly ***more information to accumulate than the amount forgotten*** would make sequence analysis easier.

$$\mathbf{s}_t = \mathbf{U}_f\mathbf{h}_t + \mathbf{W}_f\mathbf{x}_t + \mathbf{b}_f$$
$$\mathbf{c}'_t = \text{tanh}(\mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{W}_c\mathbf{x}_t + \mathbf{b}_c)$$
$$\mathbf{c}_t = \sigma(\mathbf{s}_t) \odot \mathbf{c}_{t-1} + (\mathbf{1} - (\mathbf{s_t} -)) \odot \mathbf{c}'_t$$
$$\mathbf{h}_t = \mathbf{c}_t$$

4. ***chrono initializer***
   - If the value of both input and hidden layers are zero-centered over time, $\mathbf{f}_t$ will be centered around $\sigma(1) = 0.7311$
   - The memory values $\mathbf{c}_t$ would not be retrained for more than a couple of time steps.

$$\mathbf{b}_f \sim \log(\mathcal{U}[1, T_{max} - 1])$$
$$\mathbf{b}_i = -\mathbf{b}_f$$

- an elegant implementation of skip-like connections between the memory cells over

1.3 References

1. An Empirical Exploration of Recurrent Network Architectures