Lab 10.5

Emilio Dorigatti

2022-01-25

Why does the LSTM not suffer from the vanishing or exploding gradient problems? Let's find out. To simplify matters, assume that the loss function is computed using only the last hidden state.

First show that, for a vanilla RNN, the gradient of the loss $\mathcal{L}$ with respect to the hidden state k steps earlier is given by:

$$\nabla_{\mathbf{h}^{[\tau-k]}}\mathcal{L} = \left[\prod_{i=1}^{k}\mathbf{W}^T\text{diag}\left(1-\mathbf{h}^{[\tau-k+i]^2}\right)\right]\cdot\nabla_{\mathbf{h}^{[\tau]}}\mathcal{L} \tag{1}$$

The goal is to study the behavior of $||\nabla_{\mathbf{h}^{[\tau-k]}}\mathbf{h}^{[\tau]}||$ and as k grows, i.e., as the sequences become longer and longer, where $||\cdot||$ is the L2 norm of a vector or matrix. Therefore, show that:

$$||\nabla_{\mathbf{h}^{[\tau-k]}}\mathbf{h}^{[\tau]}|| \leq ||\mathbf{W}||^k\cdot\left(\max_x|1-\tanh(x)^2|\right)^k \tag{2}$$

Hint: $||\mathbf{A}|| = \sqrt{\rho(\mathbf{A}^T\mathbf{A})}$, where $\rho(\mathbf{B}) = \max_i|\lambda_i|$ and λ_i is the i-th eigenvalue of $\mathbf{B}$. Moreover, $||\mathbf{AB}|| \leq ||\mathbf{A}||\cdot||\mathbf{B}||$.

What happens as $k\to\infty$?

It is considerably harder to find a closed-form expression for $\nabla_{\mathbf{s}^{[\tau-k]}}\mathcal{L}$ for a LSTM. Instead, we will only show that there exist a path through the unrolled LSTM computational graph where the error signal does not vanish or explode, regardless of k. This path goes through the cell state at each time step and never touches the hidden states, where the error signal is scaled down due to the sigmoid and tanh activations.

Therefore, compute the gradient of the loss considering only the cell states, and show that

$$\nabla_{\mathbf{s}^{[\tau-k]}}\mathcal{L} = \left[c+\prod_{i=1}^{k}\text{diag}\left(\mathbf{e}^{[\tau-k+i]}\right)\right]\text{diag}\left(1-\tanh\left(\mathbf{s}^{[\tau]}\right)^2\right)\text{diag}\left(\mathbf{o}^{[\tau]}\right)\nabla_{\mathbf{h}^{[\tau]}}\mathcal{L} \tag{3}$$

where c contains all terms of the gradient that go through a hidden state, $\mathbf{s}^{[t]}$, $\mathbf{e}^{[t]}$ and $\mathbf{o}^{[t]}$ indicate respectively the cell state, forget and output gates at time step t. Now focus on $||\nabla_{\mathbf{s}^{[\tau-k]}}\mathbf{s}^{[\tau]}||$ and show that

$$||\nabla_{\mathbf{s}^{[\tau-k]}}\mathbf{s}^{[\tau]}|| \leq ||c|| + \sup_x|\tanh(x)|^k \tag{4}$$

Compare Equations 4 and 2. How do they differ? Why is then the LSTM not affected by k?