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?

Solution

Vanilla RNN By the chain rule, we have:

$$\nabla_{\mathbf{h}^{[\tau-k]}}\mathcal{L} = \left[\prod_{i=1}^{k} \nabla_{\mathbf{h}^{[\tau-k+i-1]}}\mathbf{h}^{[\tau-k+i]}\right] \cdot \nabla_{\mathbf{h}^{[\tau]}}\mathcal{L} \tag{5}$$

which results in Equation 1 (see previous lab). Using the results on matrix and vector norms illustrated in the hint, we can upper bound $||\nabla_{\mathbf{h}^{[t-1]}}\mathbf{h}^{[t]}||$ as follows:

$$||\nabla_{\mathbf{h}^{[t-1]}}\mathbf{h}^{[t]}|| = \mathbf{W}^T \text{diag}\left(1 - {\mathbf{h}^{[t]}}^2\right) \leq \left|\left|\mathbf{W}^T\right|\right| \cdot \left|\left|\text{diag}\left(1 - {\mathbf{h}^{[t]}}^2\right)\right|\right| \tag{6}$$

$$\begin{tikzcd}
\mathbf{s}^{[1]} \arrow[r] \arrow[d] & \mathbf{s}^{[2]} \arrow[r] \arrow[d] & \mathbf{s}^{[3]} \arrow[d] & \\
\mathbf{h}^{[1]} \arrow[r] \arrow[ur] & \mathbf{h}^{[2]} \arrow[r] \arrow[ur] & \mathbf{h}^{[3]} \arrow[r] & \mathcal{L}
\end{tikzcd}$$

Figure 1: Simplified computational graph of a LSTM network.

Define $\mathbf{v} := 1 - {\mathbf{h}^{[t]}}^2$ and notice that

$$||\text{diag}\,(\mathbf{v})|| = \sqrt{\rho\left(\text{diag}\,(\mathbf{v})^T\,\text{diag}\,(\mathbf{v})\right)} = \sqrt{\rho\,(\text{diag}\,(\mathbf{v}^2))} = \sqrt{\max_i |\mathbf{v}_i^2|} = \max_i |\mathbf{v}_i| \tag{7}$$

since the eigenvalues of diag($\mathbf{v}$) are the elements of $\mathbf{v}$ themselves. Thus:

$$||\nabla_{\mathbf{h}^{[t-1]}}\mathbf{h}^{[t]}|| \leq \left|\left|\mathbf{W}^T\right|\right| \cdot \max_i \left|1 - {\mathbf{h}_i^{[t]}}^2\right| \leq \left|\left|\mathbf{W}^T\right|\right| \cdot \max_x |1 - \tanh(x)|^2 \tag{8}$$

and $||\nabla_{\mathbf{h}^{[\tau-k]}}\mathbf{h}^{[\tau]}||$ is obtained by raising that above to the k-th power.

LSTM For the LSTM, things are a bit more complicated. Refer to Figure 1 for a simplified computational graph. Notice how we need to have *two* concurrent recurrence equations, one going through the cell states and one going through the hidden states:

$$\begin{cases} \nabla_{\mathbf{s}^{[t]}}\mathcal{L} & = \tilde{\nabla}_{\mathbf{s}^{[t]}}\mathbf{s}^{[t+1]} \cdot \nabla_{\mathbf{s}^{[t+1]}}\mathcal{L} + \tilde{\nabla}_{\mathbf{s}^{[t]}}\mathbf{h}^{[t]} \cdot \nabla_{\mathbf{h}^{[t]}}\mathcal{L} \\ \nabla_{\mathbf{h}^{[t]}}\mathcal{L} & = \tilde{\nabla}_{\mathbf{h}^{[t]}}\mathbf{h}^{[t+1]} \cdot \nabla_{\mathbf{h}^{[t+1]}}\mathcal{L} + \tilde{\nabla}_{\mathbf{h}^{[t]}}\mathbf{s}^{[t+1]} \cdot \nabla_{\mathbf{s}^{[t+1]}}\mathcal{L} \end{cases} \tag{9}$$

where $\tilde{\nabla}_{\mathbf{x}}\mathbf{y}$ indicates the gradient of $\mathbf{y}$ with respect to $\mathbf{x}$ *only* through the direct connection between $\mathbf{x}$ and $\mathbf{y}$ in the graph (the ∇ operator would consider all paths).

As mentioned in the main text, it suffices to show that there exist *one* path that allows the error signal to flow back uninterrupted. This path goes through the cell states, i.e., we are using the following simplification:

$$\nabla_{\mathbf{s}^{[t]}}\mathcal{L} = c + \tilde{\nabla}_{\mathbf{s}^{[t]}}\mathbf{s}^{[t+1]} \cdot \nabla_{\mathbf{s}^{[t+1]}}\mathcal{L} = c + \text{diag}(\mathbf{e}^{[t+1]}) \cdot \nabla_{\mathbf{s}^{[t+1]}}\mathcal{L} \tag{10}$$

If we only consider cell states, then, we get a single recurrent equation that is readily expanded:

$$\nabla_{\mathbf{s}^{[\tau-k]}}\mathcal{L} = \prod_{i=1}^{k}\left[c + \text{diag}(\mathbf{e}^{[\tau-k+i]})\right] \cdot \nabla_{\mathbf{s}^{[\tau]}}\mathcal{L} \tag{11}$$

If we expand the product and collapse all terms involving c we get Equation 3.

From equation 7 it immediately follows that

$$\left|\left|\tilde{\nabla}_{\mathbf{s}^{[t-1]}}\mathbf{s}^{[t]}\right|\right| = \left|\left|\text{diag}\left(\mathbf{e}^{[t]}\right)\right|\right| \leq \sup_x |\tanh(x)| \tag{12}$$

Plugging that into the product gives Equation 4

Comparison of RNN and LSTM Notice that $\sup_x |\tanh(x)| = \max_x |1 - \tanh(x)^2| = 1$. Therefore, a LSTM is able to store patterns indefinitely as long as the forget gate remains open. In fact, the original formulation of the LSTM did not even include a forget gate. On the other hand, for a vanilla RNN everything depends on the eigenvalues of the weight matrix $\mathbf{W}$. In practice, a LSTM network can learn to control the forget gate, which is re-computed at every time step, while a RNN has no control on the norm of the weight matrix, which depends on the initialization, the optimizer, the order of the samples presented to the network, etc. While several techniques have been proposed to solve this problem, the LSTM proved to be the most effective. Read the original LSTM paper if you want to know more:

Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. Neural Computation 9, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735