Sit down, Shakespeare!

replacing the old bard with recurrent neural networks

Summary

...

Deep Learning in Data Science
DD2424
Spring 2023

LSTM Backpropagation

Forward pass

The forward pass is defined as follows,

$$\begin{aligned}
\boldsymbol{i}_t &= \sigma(W_i \boldsymbol{h}_{t-1} + U_i \boldsymbol{x}_t) \\
\boldsymbol{f}_t &= \sigma(W_f \boldsymbol{h}_{t-1} + U_f \boldsymbol{x}_t) \\
\boldsymbol{e}_t &= \sigma(W_e \boldsymbol{h}_{t-1} + U_e \boldsymbol{x}_t) \\
\tilde{\boldsymbol{c}}_t &= \tanh(W_c \boldsymbol{h}_{t-1} + U_c \boldsymbol{x}_t) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \bullet \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \bullet \tilde{\boldsymbol{c}}_t \\
\boldsymbol{h}_t &= \boldsymbol{e}_t \bullet \tanh(\boldsymbol{c}_t) \\
\boldsymbol{o}_t &= V \boldsymbol{h}_t \\
\boldsymbol{p}_t &= \text{softmax}(\boldsymbol{o}_t)
\end{aligned}$$

where $\bullet$ denotes element by element multiplication and $\boldsymbol{e}_t$ denotes the output/exposure gate.

Backward pass

In order to find the analytical gradients and compute the backward pass, we employ the chain rule. First, we consider the gradient of the cross-entropy loss w.r.t. the output for the final time step. We define

$$\frac{\partial L}{\partial \boldsymbol{o}_t} = -(\boldsymbol{y}_t - \boldsymbol{p}_t)^T, \quad \forall\, t = 1, 2, \ldots, T$$

and denote $\boldsymbol{g}_t := \frac{\partial L}{\partial \boldsymbol{o}_t}$. Then, in order to compute the gradients of the individual weights, we need to first find define the gradients with respect to the hidden activation, the gates and the memory cell. Hence, we first consider the partial derivative of the loss with respect to the hidden units, i.e. h_t for $t = 1, 2, \ldots, T$, and consider two cases: $t = T$ and $t = 1, 2, \ldots, T-1$.

I - the case of $t = T$

For this case, the gradient computation is straightforward by employing the chain rule. Specifically, we have that

$$\frac{\partial L}{\partial \boldsymbol{h}_t} = \frac{\partial L}{\partial \boldsymbol{o}_t} \frac{\partial \boldsymbol{o}_t}{\partial \boldsymbol{h}_t} = \boldsymbol{g}_t V$$

II - the case of $t < T$

In order to perform the backward pass for earlier time steps in the sequence, we need to consider how the information propagates forward, i.e. the hidden unit at $t-1$ passes through the hidden unit at t, and since the hidden unit at t accounts for the input, output and exposure gates, we have that

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{h}_t} &= \boldsymbol{g}_t V + \frac{\partial L}{\partial \boldsymbol{i}_{t+1}} \frac{\partial \boldsymbol{i}_{t+1}}{\partial \boldsymbol{h}_t} + \frac{\partial L}{\partial \boldsymbol{f}_{t+1}} \frac{\partial \boldsymbol{f}_{t+1}}{\partial \boldsymbol{h}_t} + \frac{\partial L}{\partial \boldsymbol{e}_{t+1}} \frac{\partial \boldsymbol{e}_{t+1}}{\partial \boldsymbol{h}_t} + \frac{\partial L}{\partial \tilde{\boldsymbol{c}}_{t+1}} \frac{\partial \tilde{\boldsymbol{c}}_{t+1}}{\partial \boldsymbol{h}_t} \\
&= \boldsymbol{g}_t V + \frac{\partial L}{\partial \boldsymbol{i}_{t+1}} \frac{\partial \boldsymbol{i}_{t+1}}{\partial \boldsymbol{a}_{i,t+1}} W_i + \frac{\partial L}{\partial \boldsymbol{f}_{t+1}} \frac{\partial \boldsymbol{f}_{t+1}}{\partial \boldsymbol{a}_{f,t+1}} W_f + \frac{\partial L}{\partial \boldsymbol{e}_{t+1}} \frac{\partial \boldsymbol{e}_{t+1}}{\partial \boldsymbol{a}_{et+1}} W_e + \frac{\partial L}{\partial \tilde{\boldsymbol{c}}_{t+1}} \frac{\partial \tilde{\boldsymbol{c}}_{t+1}}{\partial \boldsymbol{a}_{c,t+1}} W_c
\end{aligned}$$

where $\boldsymbol{a}_{,t}$ denote the activations at time t, e.g.

$$\boldsymbol{i}_t := \sigma(\boldsymbol{a}_{i,t})$$

III - gradients for gates and memory cell

Further, for notational convenience, we set

$$\tilde{\boldsymbol{g}}_t := \boldsymbol{g_t} \frac{\partial \boldsymbol{o}_t}{\partial \boldsymbol{h}_t}$$

such that we can expand the terms in the above equation per the follwing:

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{i}_t} &= \tilde{\boldsymbol{g}}_t \, \boldsymbol{e}_t (1 - \tanh^2(\boldsymbol{c}_t)) \tilde{\boldsymbol{c}}_t \\
\frac{\partial L}{\partial \boldsymbol{f}_t} &= \tilde{\boldsymbol{g}}_t \, \boldsymbol{e}_t (1 - \tanh^2(\boldsymbol{c}_t)) \boldsymbol{c}_{t-1} \\
\frac{\partial L}{\partial \boldsymbol{e}_t} &= \tilde{\boldsymbol{g}}_t \, \tanh(\boldsymbol{c}_t) \\
\frac{\partial L}{\partial \tilde{\boldsymbol{c}}_t} &= \tilde{\boldsymbol{g}}_t \, \boldsymbol{e}_t (1 - \tanh^2(\boldsymbol{c}_t)) \boldsymbol{i}_t
\end{aligned}$$

IV - gradients for activations

In order to properly compute the gradients, we also need to consider the gradients with respect to the activations. Thus, we have that

$$\begin{aligned}
\frac{\partial \boldsymbol{i}_t}{\partial \boldsymbol{a}_{i,t}} &= \sigma(\boldsymbol{a}_{i,t})(1 - \sigma(\boldsymbol{a}_{i,t})) \\
\frac{\partial \boldsymbol{f}_t}{\partial \boldsymbol{a}_{f,t}} &= \sigma(\boldsymbol{a}_{f,t})(1 - \sigma(\boldsymbol{a}_{f,t})) \\
\frac{\partial \boldsymbol{e}_t}{\partial \boldsymbol{a}_{e,t}} &= \sigma(\boldsymbol{a}_{e,t})(1 - \sigma(\boldsymbol{a}_{e,t})) \\
\frac{\partial \tilde{\boldsymbol{c}}_t}{\partial \boldsymbol{a}_{c,t}} &= 1 - \tanh^2(\boldsymbol{a}_{c,t})
\end{aligned}$$

V - putting it all together

Finally, given that we have iteratively computed all the gradients for the hidden units, the gates, the memory cell and their respective activations, we have that

$$\frac{\partial L}{\partial W_i} = \sum_{t=1}^{T} \tilde{\boldsymbol{g}}_t \, \boldsymbol{e}_t(1 - \tanh^2(\boldsymbol{c}_t))\tilde{\boldsymbol{c}}_t \sigma(\boldsymbol{a}_{i,t})(1 - \sigma(\boldsymbol{a}_{i,t}))\boldsymbol{h}_{t-1} \tag{1}$$

$$\frac{\partial L}{\partial U_i} = \sum_{t=1}^{T} \tilde{\boldsymbol{g}}_t \, \boldsymbol{e}_t(1 - \tanh^2(\boldsymbol{c}_t))\tilde{\boldsymbol{c}}_t \sigma(\boldsymbol{a}_{i,t})(1 - \sigma(\boldsymbol{a}_{i,t}))\boldsymbol{x}_t \tag{2}$$

$$\frac{\partial L}{\partial W_f} = \sum_{t=1}^{T} \tilde{\boldsymbol{g}}_t \, \boldsymbol{e}_t(1 - \tanh^2(\boldsymbol{c}_t))\boldsymbol{c}_{t-1} \sigma(\boldsymbol{a}_{f,t})(1 - \sigma(\boldsymbol{a}_{f,t}))\boldsymbol{h}_{t-1} \tag{3}$$

$$\frac{\partial L}{\partial U_f} = \sum_{t=1}^{T} \tilde{\boldsymbol{g}}_t \, \boldsymbol{e}_t(1 - \tanh^2(\boldsymbol{c}_t))\boldsymbol{c}_{t-1} \sigma(\boldsymbol{a}_{f,t})(1 - \sigma(\boldsymbol{a}_{f,t}))\boldsymbol{x}_t \tag{4}$$

$$\frac{\partial L}{\partial W_e} = \sum_{t=1}^{T} \tilde{\boldsymbol{g}}_t \, \tanh(\boldsymbol{c}_t) \sigma(\boldsymbol{a}_{e,t})(1 - \sigma(\boldsymbol{a}_{e,t}))\boldsymbol{h}_{t-1} \tag{5}$$

$$\frac{\partial L}{\partial U_e} = \sum_{t=1}^{T} \tilde{\boldsymbol{g}}_t \, \tanh(\boldsymbol{c}_t) \sigma(\boldsymbol{a}_{e,t})(1 - \sigma(\boldsymbol{a}_{e,t}))\boldsymbol{x}_t \tag{6}$$

$$\frac{\partial L}{\partial W_c} = \sum_{t=1}^{T} \tilde{\boldsymbol{g}}_t \, \boldsymbol{e}_t(1 - \tanh^2(\boldsymbol{c}_t))\boldsymbol{i}_t(1 - \tanh^2(\boldsymbol{a}_{c,t}))\boldsymbol{h}_{t-1} \tag{7}$$

$$\frac{\partial L}{\partial U_c} = \sum_{t=1}^{T} \tilde{\boldsymbol{g}}_t \, \boldsymbol{e}_t(1 - \tanh^2(\boldsymbol{c}_t))\boldsymbol{i}_t(1 - \tanh^2(\boldsymbol{a}_{c,t}))\boldsymbol{x}_t \tag{8}$$

$$\frac{\partial L}{\partial V} = \sum_{t=1}^{T} \boldsymbol{g}_t \boldsymbol{h}_t \tag{9}$$