Tittel

Forfatter

1 Backpropagation through time - BPTT

Definitions:

$$\mathbf{a}^{(t)} = \mathbf{b} + \mathbf{W}\mathbf{h}^{(\mathbf{t-1})} + \mathbf{U}\mathbf{x}^{(\mathbf{t})} \tag{1}$$
$$\mathbf{h}^{(t)} = activation(\mathbf{a}^{(t)}) \tag{2}$$
$$\mathbf{o}^{(t)} = \mathbf{c} + \mathbf{V}\mathbf{h}^{(t)} \tag{3}$$
$$\hat{\mathbf{y}}^{(t)} = softmax(\mathbf{o}^{(t)}) \tag{4}$$
$$\mathbf{U} = \text{input weights} \tag{5}$$
$$\mathbf{W} = \text{hidden weights} \tag{6}$$
$$\mathbf{V} = \text{output weights} \tag{7}$$

The goal is to maximize the probability of the observed data by estimating parameters. The estimated parameters yielding the highest maximum likelihood are called the maximum likelihood estimates. These parameters can be obtained by minimizing the cross-entropy between the model distribution and the data. This cross-entropy function can be seen in (8), giving rise to the cost function in (10).

$$C\left(\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(\tau)}\}, \{\mathbf{y}^{(1)}, ..., \mathbf{y}^{(\tau)}\}\right) \tag{8}$$
$$= \sum_t L^{(t)} \tag{9}$$
$$= -\sum_t log\ p\left(y^{(t)} | \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(t)}\}\right) \tag{10}$$
$$\tag{11}$$

Note: $y^{(t)}$ is an entry in the output vector $\hat{\mathbf{y}}^{(t)}$

The cost function is the negative log-likelihood function. Minimising this function is the same as maximum the likelihood of the parameters - not due to the log, but due to the negative prefix. The log in log-likelihood works, but I dont know why it is used?

Below we derive the gradients of the nodes in the computational grap from deep learning book.

The gradient of the cost function at the output, $\mathbf{o}$, at time t is

$$\nabla_{\mathbf{o}^{(t)}} C = \frac{\delta C}{\delta o^{(t)}} = \frac{\delta C}{\delta C^{(t)}} \frac{\delta C^{(t)}}{\delta o^{(t)}} = \hat{y}^{(t)} - \mathbf{1}_{i=y^{(t)}} \tag{12}$$

The gradient of the final hidden state, $\mathbf{h}^{(\tau)}$, is only influenced by $\mathbf{o}^{(\tau)}$, since there are no descending hidden states. It is given by

$$\nabla_{\mathbf{h}^{(\tau)}} C = \left(\nabla_{\mathbf{o}^{(\tau)}} C\right) \frac{\delta \mathbf{o}^{(\tau)}}{\delta \mathbf{h}^{(\tau)}} \tag{13}$$
$$= \left(\nabla_{\mathbf{o}^{(\tau)}} C\right) \mathbf{V} \tag{14}$$
$$\nabla_{\mathbf{h}^{(\tau)}} C = \mathbf{V}^\top \left(\nabla_{\mathbf{o}^{(\tau)}} C\right) \tag{15}$$

Where all the right hand side terms are known from before.

The only nodes that need gradient computation now, are all the hidden states preceding the last. I.e., for $\mathbf{h}^{(t)}$, where $t = \{0, ..., \tau - 1\}$. The gradient is now a result of both the gradient at $\mathbf{o}^{(t)}$, as well as all the preceding hidden state gradients. Remember that the preceding hidden state of $\mathbf{h}^{(t)}$ is $\mathbf{h}^{(t+1)}$, which has preceding hidden state $\mathbf{h}^{(t+2)}$, and so on. We are generally calculating the gradient starting from $t = \tau$, working our way down to $t = 0$.

$$\nabla_{\mathbf{h}^{(\tau-1)}} C = \nabla_{\mathbf{o}^{(\tau-1)}} C \frac{\delta \mathbf{o}^{(\tau-1)}}{\delta \mathbf{h}^{(\tau-1)}} + \nabla_{\mathbf{h}^{(\tau)}} C \frac{\delta \mathbf{h}^{(\tau)}}{\delta \mathbf{h}^{(\tau-1)}}$$

$$\nabla_{\mathbf{h}^{(\tau-2)}} C = \nabla_{\mathbf{o}^{(\tau-2)}} C \frac{\delta \mathbf{o}^{(\tau-2)}}{\delta \mathbf{h}^{(\tau-2)}} + \nabla_{\mathbf{h}^{(\tau-1)}} C \frac{\delta \mathbf{h}^{(\tau-1)}}{\delta \mathbf{h}^{(\tau-2)}}$$

$$\begin{aligned}
\nabla_{\mathbf{h}^{(\tau-3)}} C &= \nabla_{\mathbf{o}^{(\tau-3)}} C \frac{\delta \mathbf{o}^{(\tau-3)}}{\delta \mathbf{h}^{(\tau-3)}} + \nabla_{\mathbf{h}^{(\tau-2)}} C \frac{\delta \mathbf{h}^{(\tau-2)}}{\delta \mathbf{h}^{(\tau-3)}} \\
&= \nabla_{\mathbf{o}^{(\tau-3)}} C \frac{\delta \mathbf{o}^{(\tau-3)}}{\delta \mathbf{h}^{(\tau-3)}} + \left(\nabla_{\mathbf{o}^{(\tau-2)}} C \frac{\delta \mathbf{o}^{(\tau-2)}}{\delta \mathbf{h}^{(\tau-2)}} + \nabla_{\mathbf{h}^{(\tau-1)}} C \frac{\delta \mathbf{h}^{(\tau-1)}}{\delta \mathbf{h}^{(\tau-2)}} \right) \frac{\delta \mathbf{h}^{(\tau-2)}}{\delta \mathbf{h}^{(\tau-3)}} \\
&= \nabla_{\mathbf{o}^{(\tau-3)}} C \frac{\delta \mathbf{o}^{(\tau-3)}}{\delta \mathbf{h}^{(\tau-3)}} + \left(\nabla_{\mathbf{o}^{(\tau-2)}} C \frac{\delta \mathbf{o}^{(\tau-2)}}{\delta \mathbf{h}^{(\tau-2)}} + \left(\nabla_{\mathbf{o}^{(\tau-1)}} C \frac{\delta \mathbf{o}^{(\tau-1)}}{\delta \mathbf{h}^{(\tau-1)}} + \nabla_{\mathbf{h}^{(\tau)}} C \frac{\delta \mathbf{h}^{(\tau)}}{\delta \mathbf{h}^{(\tau-1)}} \right) \frac{\delta \mathbf{h}^{(\tau-1)}}{\delta \mathbf{h}^{(\tau-2)}} \right) \frac{\delta \mathbf{h}^{(\tau-2)}}{\delta \mathbf{h}^{(\tau-3)}}
\end{aligned}$$