Backpropagation through recurrent layer

Kadulin V.

October 21, 2022

Let $X \in \mathbb{R}^{N \times T \times D}$ be D-dimentional representations of N sequences of length T each, $h \in \mathbb{R}^{N \times H}$ is hidden state of a recurrent cell. Let $g : (x_t, h_t) \to h_{t+1}$ is a recurrent cell operator, where $x_t \in \mathbb{R}^{N \times D}$ is a batch of t-th elements of each sequence, $h_t \in \mathbb{R}^{N \times H}$ is a batch of hidden states at step t. $g_t = \tanh(x_t W_x + h_{t-1} W_h + b)$. Applying this operator consequentially to each of T slices along second axis of X, we will store outputs in $Y \in \mathbb{R}^{N \times T \times H}$. Let f is a scalar function of Y. We know $\frac{\partial f}{\partial Y}$. Derive $\frac{\partial f}{\partial W_x}$, $\frac{\partial f}{\partial W_h}$, $\frac{\partial f}{\partial b}$, $\frac{\partial f}{\partial X}$, $\frac{\partial f}{\partial h_0}$.

First, let's derive a single backward step through g. Let x, y, h are input, output and hidden state at step t respectively, h_{t-1} – hidden state at step $t-1$. Let z be the expression inside tanh function.

$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial g} \frac{\partial g}{\partial z}$

$\frac{\partial g}{\partial z} = \left(\frac{e^z - e^{-z}}{2} \right)^{-2}$

$\frac{\partial f}{\partial x_t} = \frac{\partial f}{\partial z} W_x^\top$

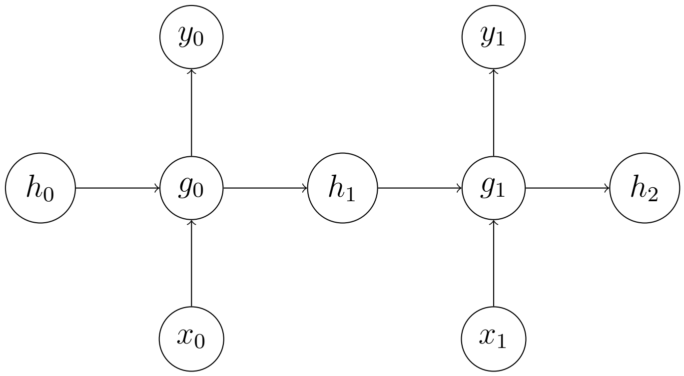
$\frac{\partial f}{\partial W_x} = x^\top \frac{\partial f}{\partial z}$

$\frac{\partial f}{\partial h_{t-1}} = \frac{\partial f}{\partial z} W_h^\top$

$\frac{\partial f}{\partial W_h} = h_{t-1}^\top \frac{\partial f}{\partial z}$

$\frac{\partial f}{\partial b} = \sum_{i=1}^{N} \left(\frac{\partial f}{\partial z} \right)_i$

Let's consider case $N = 1, T = 2$. So, let $x_t \in \mathbb{R}^D$ be a word representation at step t, $h_t \in \mathbb{R}^H$ be a hidden state at step t. The computation graph is as follows:



- $y_t = h_{t+1}$, different letters here are set to distinguish different cases.
- $\frac{\partial f}{\partial g_t} = \frac{\partial f}{\partial y_t} + \frac{\partial f}{\partial h_{t+1}}$, where the first term is an upstream derivative, and the second term – local recurrent layer derivative, passed back between steps.

$\frac{\partial f}{\partial W_x} = \sum_{t=1}^{T} \left(\frac{\partial f}{\partial W_x} \right)_t$
$\frac{\partial f}{\partial W_h} = \sum_{t=1}^{T} \left(\frac{\partial f}{\partial W_h} \right)_t$
$\frac{\partial f}{\partial b} = \sum_{t=1}^{T} \left(\frac{\partial f}{\partial b} \right)_t$
$\left(\frac{\partial f}{\partial X} \right)_{\cdot,t,\cdot} = \frac{\partial f}{\partial x_t}$
$\frac{\partial f}{\partial h_0} = \frac{\partial f}{\partial z_1} W_h^\top$