Homework 7 Solutions

1. **Binary Addition [4pts]**

Recall architecture of our binary addition RNN which has two input units, three hidden units, and one output unit:

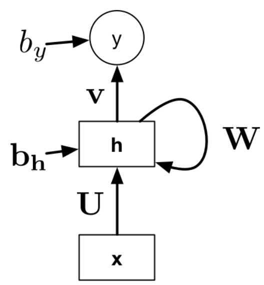

Figure 1: RNN architecture

$$\begin{aligned}
\mathbf{z}^{(t)} &= \mathbf{U}\mathbf{x}^{(t)} + \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{b_h} \\
\mathbf{h}^{(t)} &= \phi(\mathbf{z}^{(t)}) \\
r^{(t)} &= \mathbf{V}\mathbf{h}^{(t)} + b_y \\
y^{(t)} &= \phi(r^{(t)}).
\end{aligned}$$

$$\text{where } \phi(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z \leq 0 \end{cases}$$

Figure 2: Forward pass computations in our RNN

We will follow the hint given in the homework statement and implement the addition in our RNN such that:

(a) The first of our hidden units $h_1^{(t)}$ is 1 if and only if the sum $S^{(t)} \doteq x_1^{(t)} + x_2^{(t)} + c^{(t-1)} \geq 1$, where by $c^{(t-1)}$ we denote a carry from the previous addition. Note, these $S^{(t)}$ and $c^{(t-1)}$ are not variables of the model, merely our notation to help us to work out the solution.

(b) The $h_2^{(t)}$ is 1 iff the sum $S^{(t)} \geq 2$,

(c) and $h_3^{(t)}$ is 1 iff the sum $S^{(t)}$ is 3.

Notice that the carry $c^{(t-1)}$ is going to be 1 iff $h_2^{(t-1)} = 1$ and 0 otherwise[1], i.e. when the previous addition was 2 or 3. Therefore to compute $h_i^{(t)}$ we need to first compute the sum $S^{(t)} = x_1^{(t)} + x_2^{(t)} + h_2^{(t-1)}$ and then offset it by $-i+1$ so that after applying the hard threshold function we get the desired value as specified above. This can be achieved with the following set of parameters: $\mathbf{U} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, $\mathbf{b_h} = \begin{bmatrix} -0.5 \\ -1.5 \\ -2.5 \end{bmatrix}$

Finally, to compute the output $y^{(t)}$ we need to check if the $S^{(t)}$ is 1 or 3, that is, if either $h_1^{(t)} = 1$ while all other hidden units are zero or all hidden units are 1. We can accomplish this by setting: $\mathbf{V} = \begin{bmatrix} 1, -1, 1 \end{bmatrix}$ and $b_y = -0.5$.

[1] We need to initialize $\mathbf{h}^{(0)} = \mathbf{0}$.

2. **LSTM Gradient**

(a) [3pts] Derivation of the backprop update rules for the activations and the gates:

$$\begin{aligned}
\overline{h^{(t)}} &= \overline{i^{(t+1)}}\frac{\partial i^{(t+1)}}{\partial h^{(t)}} + \overline{f^{(t+1)}}\frac{\partial f^{(t+1)}}{\partial h^{(t)}} + \overline{o^{(t+1)}}\frac{\partial o^{(t+1)}}{\partial h^{(t)}} + \overline{g^{(t+1)}}\frac{\partial g^{(t+1)}}{\partial h^{(t)}} \\
&= \overline{i^{(t+1)}} i^{(t+1)}(1 - i^{(t+1)}) w_{ih} + \\
&\quad + \overline{f^{(t+1)}} f^{(t+1)}(1 - f^{(t+1)}) w_{fh} + \\
&\quad + \overline{o^{(t+1)}} o^{(t+1)}(1 - o^{(t+1)}) w_{oh} + \\
&\quad + \overline{g^{(t+1)}} \left(1 - \tanh^2\left(w_{gx}x^{(t+1)} + w_{gh}h^{(t)}\right)\right) w_{gh} \\
\overline{c^{(t)}} &= \overline{h^{(t)}}\frac{\partial h^{(t)}}{\partial c^{(t)}} + \overline{c^{(t+1)}}\frac{\partial c^{(t+1)}}{\partial c^{(t)}} = \overline{h^{(t)}} o^{(t)}\left(1 - \tanh^2(c^{(t)})\right) + \overline{c^{(t+1)}} f^{(t+1)} \\
\overline{g^{(t)}} &= \overline{c^{(t)}}\frac{\partial c^{(t)}}{\partial g^{(t)}} = \overline{c^{(t)}} i^{(t)} \\
\overline{o^{(t)}} &= \overline{h^{(t)}}\frac{\partial h^{(t)}}{\partial o^{(t)}} = \overline{h^{(t)}} \tanh(c^{(t)}) \\
\overline{f^{(t)}} &= \overline{c^{(t)}}\frac{\partial c^{(t)}}{\partial f^{(t)}} = \overline{c^{(t)}} c^{(t-1)} \\
\overline{i^{(t)}} &= \overline{c^{(t)}}\frac{\partial c^{(t)}}{\partial i^{(t)}} = \overline{c^{(t)}} g^{(t)}
\end{aligned}$$

Additionally $\overline{h^{(t)}}$ may include $\frac{\partial \mathcal{L}}{\partial h^{(t)}}$ term if $h^{(t)}$ is directly part of the loss function.

(b) [1pt] Derive the backprop rule for the weight w_{ix}:

$$\begin{aligned}
\overline{w_{ix}} &= \sum_{t=1}^{T} \overline{i^{(t)}}\frac{\partial i^{(t)}}{\partial w_{ix}} \\
&= \sum_{t=1}^{T} \overline{i^{(t)}}\sigma'\left(w_{ix}x^{(t)} + w_{ih}h^{(t-1)}\right) x^{(t)} \\
&= \sum_{t=1}^{T} \overline{i^{(t)}} i^{(t)}(1 - i^{(t)}) x^{(t)}
\end{aligned}$$

(c) [2pt]

By inspecting the partial derivatives from (a), we can see that the $\overline{g^{(t)}}, \overline{o^{(t)}}, \overline{f^{(t)}}$ and $\overline{i^{(t)}}$ could explode or vanish only if $\overline{c^{(t)}}$ or $\overline{h^{(t)}}$ does. Therefore it is enough to investigate whether $\overline{c^{(t)}}$ and $\overline{h^{(t)}}$ don't explode nor vanish. Recall, we assume that

$$\forall t : f^{(t)} \approx 1, i^{(t)} \approx 0, o^{(t)} \approx 0$$

First, we show that the gradient passes through $c^{(t)}$ basically unchanged:

$$\begin{aligned}
\overline{c^{(t)}} &= \overline{h^{(t)}} o^{(t)}\left(1 - \tanh^2(c^{(t)})\right) + \overline{c^{(t+1)}} f^{(t+1)} \\
&\approx \overline{c^{(t+1)}} f^{(t+1)} \\
&\approx \overline{c^{(t+1)}}
\end{aligned}$$

Secondly, we show that $\overline{h^{(t)}}$ is zero (or $\frac{\partial \mathcal{L}}{\partial h^{(t)}}$ as discussed in part (a)).

$$\begin{aligned}
\overline{h^{(t)}} = {} & \overline{i^{(t+1)}}\,\underbrace{i^{(t+1)}(1-i^{(t+1)})}_{\approx 0}\,w_{ih}+ \\
& + \overline{f^{(t+1)}}\,\underbrace{f^{(t+1)}(1-f^{(t+1)})}_{\approx 0}\,w_{fh}+ \\
& + \overline{o^{(t+1)}}\,\underbrace{o^{(t+1)}(1-o^{(t+1)})}_{\approx 0}\,w_{oh}+ \\
& + \underbrace{\overline{g^{(t+1)}}}_{=\overline{c^{(t+1)}}i^{(t+1)}\approx 0}\left(1-\tanh^2\left(w_{gx}x^{(t+1)}+w_{gh}h^{(t)}\right)\right)w_{gh} \\
\approx {} & 0
\end{aligned}$$

Therefore no gradient can explode or vanish in this case.