CSC421: Written Homework 3

Due on Thrusday, Mar 7, 2019

Zhongtian Ouyang
1002341012

Problem 1

LSTM Gradient

(a)

$$
\begin{aligned}
\overline{h^{(t)}} &= \overline{i^{(t+1)}} i^{(t+1)}(1 - i^{(t+1)})w_{ih} \\
&+ \overline{f^{(t+1)}} f^{(t+1)}(1 - f^{(t+1)})w_{fh} \\
&+ \overline{o^{(t+1)}} o^{(t+1)}(1 - o^{(t+1)})w_{oh} \\
&+ \overline{g^{(t+1)}}(1 - {g^{(t+1)}}^2)w_{gh} \\
\overline{c^{(t)}} &= \overline{c^{(t+1)}} f^{(t+1)} + \overline{h^{(t)}} o^{(t)}(1 - (tanh(c^{(t)}))^2) \\
\overline{g^{(t)}} &= \overline{c^{(t)}} i^{(t)} \\
\overline{o^{(t)}} &= \overline{h^{(t)}} tanh(c^{(t)}) \\
\overline{f^{(t)}} &= \overline{c^{(t)}} c^{(t-1)} \\
\overline{i^{(t)}} &= \overline{c^{(t)}} g^{(t)}
\end{aligned}
\tag{1}
$$

If $h^{(t)}$ is used in $y^{(t)}$, $\overline{h^{(t)}} += \overline{y^{(t)}}(\partial y^{(t)}/\partial h^{(t)})$
If $h^{(t)}$ is used in Loss function L, $\overline{h^{(t)}} += \overline{L}(\partial L/\partial h^{(t)})$

(b)

$$\overline{w_{ix}} = \sum_t \overline{i^{(t)}} i^{(t)}(1 - i^{(t)})x^{(t)}$$

Problem 2

Multidimensional RNN
(a)
Number of Weights:
$W_{in}^T : D \times H$
$W_W^T : H \times H$
$W_W^T : H \times H$
Total: $(D + 2H) \times H$

Number of arithmetic operations for an $h^{(i,j)}$
Activation function elementwise(assume n_a arithmatic operations to evaluate $\phi(x)$): $n_a H$
Addition of vectors inside activation function: 2H
$W_{in}^T x^{(i,j)} : H \times (D\ multiplications + D - 1\ addtions) = H \times (2D - 1)$
$W_W^T h^{(i-1,j)} : H \times (H\ multiplications + H - 1\ addtions) = H \times (2H - 1)$
$W_N^T h^{(i,j-1)} : H \times (H\ multiplications + H - 1\ addtions) = H \times (2H - 1)$
Total:$H \times (n_a + 2 + 2D - 1 + 2H - 1 + 2H - 1) = H \times (n_a + 2D + 4H - 1) = O(H \times (D + H))$
For $G \times G$ hidden vectors in the grid: $O(G \times G \times H \times (D + H)) = O(G^2 \times H \times (D + H))$

(b)
Assume that addition and activation function can be done in the same step as matrix-vector multiplications.
$2G - 1$ steps will be needed to computed the hidden activations of the $G \times G$ grid.
One way of doing is computing the activations in the following sequence:
Step 1: $h^{(0,0)}$
Step 2: $h^{(1,0)}$, $h^{(0,1)}$
Step 3: $h^{(0,2)}$, $h^{(1,1)}$, $h^{(2,0)}$
...
Step $2G - 2$: $h^{(G-1,G)}$, $h^{(G,G-1)}$
Step $2G - 1$: $h^{(G,G)}$
In this sequence, when we do a step, all the information for calculating each hidden activations is.

(c)
Disadvantage: The calculations for a conv net can be well parallelized. Less seqential steps are required to compute an conv net with same dimension compare to an MDRNN.
Advantage: MDRNN can capture the sequential relationship between the datas and extract more information compare to CNN.

Problem 3

Reversibility

(a)

$$\begin{aligned}
\mathbf{s}^{(k+1)} &= (\boldsymbol{\theta}^{(k+1)}, \mathbf{p}^{(k+1)}) \\
\boldsymbol{\theta}^{(k)} &= \boldsymbol{\theta}^{(k+1)} - \mathbf{p}^{(k+1)} \\
\mathbf{p}^{(k)} &= \frac{\mathbf{p}^{(k+1)} + \alpha \nabla J(\boldsymbol{\theta}^{(k)})}{\beta} \\
\mathbf{s}^{(k)} &= (\boldsymbol{\theta}^{(k)}, \mathbf{p}^{(k)})
\end{aligned} \tag{2}$$

(b)

$$\frac{\partial \mathbf{s}^{(k+1)}}{\partial \mathbf{s}^{(k)}} \tag{3}$$

Since the top half and bottom half of the matrix is identical, the rank is D ¡ 2D. The determinant of the matrix is 0