CS 7643 : HW 3

Name Niranjan Thakurdesai
Solutions discussed with Srijan Sood, Nupur Chatterjee and Karan Shah

1-1 Visualization:

Wx Wx Wx Wx

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & x_{00} & x_{01} & x_{02} & 0 \\ 0 & x_{10} & x_{11} & x_{12} & 0 \\ 0 & x_{20} & x_{21} & x_{22} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{matrix}$$

→ W is slid across zero-padded X with a stride of 3

$$Y = \begin{bmatrix} w_{11}x_{00} & w_{10}x_{02} \\ w_{01}x_{20} & w_{00}x_{22} \end{bmatrix}$$

Flattening Y in row-major order gives

$$Y = \begin{bmatrix} w_{11}x_{00} & w_{10}x_{02} & w_{01}x_{20} & w_{00}x_{22} \end{bmatrix}^T$$

Writing this as a matrix multiplication,

$$Y = \underbrace{\begin{bmatrix} w_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & w_{10} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & w_{01} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & w_{00} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_{00} \\ x_{01} \\ x_{02} \\ \vdots \\ x_{20} \\ x_{21} \\ x_{22} \end{bmatrix}}_{X}$$

1-2 Visualization:

$$\begin{array}{l}
x_{00}\times \qquad\qquad\qquad x_{01}\times \\
\begin{array}{cc|cc}
W_{00} & W_{01} & W_{00} & W_{01} \\
W_{10} & W_{11} & W_{10} & W_{11} \\
\hline
W_{00} & W_{01} & W_{00} & W_{01} \\
W_{10} & W_{11} & W_{10} & W_{11}
\end{array} \\
x_{10}\times \qquad\qquad\qquad x_{11}\times
\end{array}$$

→ Sliding W with stride 2 and multiplying it by the corresponding element of the input

$$\Rightarrow Y = \begin{bmatrix}
x_{00}W_{00} & x_{00}W_{01} & x_{01}W_{00} & x_{01}W_{01} \\
x_{00}W_{10} & x_{00}W_{11} & x_{01}W_{10} & x_{01}W_{11} \\
x_{10}W_{00} & x_{10}W_{01} & x_{11}W_{00} & x_{11}W_{01} \\
x_{10}W_{10} & x_{10}W_{11} & x_{11}W_{10} & x_{11}W_{11}
\end{bmatrix}$$

Flattening Y in row-major order and writing it as a matrix multiplication gives

$$Y = \underbrace{\begin{bmatrix}
W_{00} & 0 & 0 & 0 \\
W_{01} & 0 & 0 & 0 \\
0 & W_{00} & 0 & 0 \\
0 & W_{01} & 0 & 0 \\
W_{10} & 0 & 0 & 0 \\
W_{11} & 0 & 0 & 0 \\
0 & W_{10} & 0 & 0 \\
0 & W_{11} & 0 & 0 \\
0 & 0 & W_{00} & 0 \\
0 & 0 & W_{01} & 0 \\
0 & 0 & 0 & W_{00} \\
0 & 0 & 0 & W_{01} \\
0 & 0 & W_{10} & 0 \\
0 & 0 & W_{11} & 0 \\
0 & 0 & 0 & W_{10} \\
0 & 0 & 0 & W_{11}
\end{bmatrix}}_{A}
\underbrace{\begin{bmatrix} x_{00} \\ x_{01} \\ x_{10} \\ x_{11} \end{bmatrix}}_{X}$$

1-3 Affine transformation for a convolutional layer with kernel size (4, 1, 1, 1) : (stride 1 and no padding)

~~Y=~~ Let's denote the kernel with $w = [w_1, w_2, w_3, w_4]$

$$Y = \begin{bmatrix} w_1 x_{00} \\ w_1 x_{01} \\ w_1 x_{10} \\ w_1 x_{11} \\ w_2 x_{00} \\ w_2 x_{01} \\ w_2 x_{10} \\ w_2 x_{11} \\ w_3 x_{00} \\ w_3 x_{01} \\ w_3 x_{10} \\ w_3 x_{11} \\ w_4 x_{00} \\ w_4 x_{01} \\ w_4 x_{10} \\ w_4 x_{11} \end{bmatrix} = \underbrace{\begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_1 & 0 & 0 \\ 0 & 0 & w_1 & 0 \\ 0 & 0 & 0 & w_1 \\ w_2 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & w_2 & 0 \\ 0 & 0 & 0 & w_2 \\ w_3 & 0 & 0 & 0 \\ 0 & w_3 & 0 & 0 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & 0 & w_3 \\ w_4 & 0 & 0 & 0 \\ 0 & w_4 & 0 & 0 \\ 0 & 0 & w_4 & 0 \\ 0 & 0 & 0 & w_4 \end{bmatrix}}_{A_c} \begin{bmatrix} x_{00} \\ x_{01} \\ x_{10} \\ x_{11} \end{bmatrix} \quad \text{——} \quad (1)$$

Affine transformation for a transposed convolution layer with kernel size (1, 1, 2, 2) : (stride 2, no padding)

Let's denote the kernel with $w = \begin{bmatrix} w_1 & w_2 \\ w_3 & w_4 \end{bmatrix}$

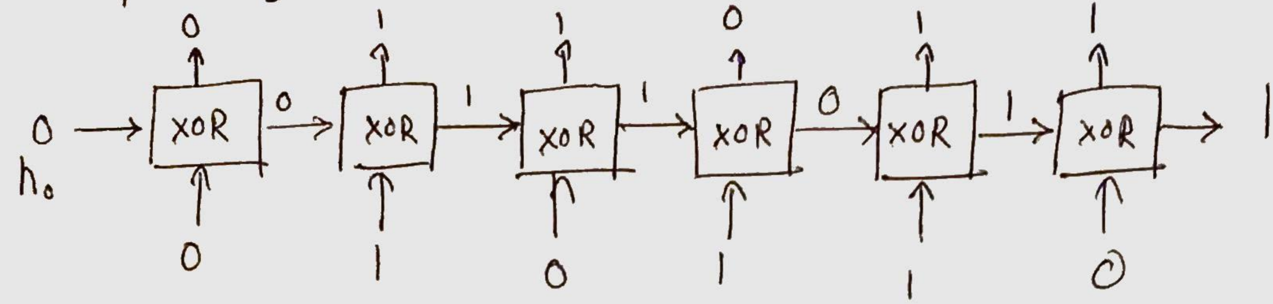
$$Y = \begin{bmatrix} x_{00}w_1 & x_{00}w_2 & x_{01}w_1 & x_{01}w_2 \\ x_{00}w_3 & x_{00}w_4 & x_{01}w_3 & x_{01}w_4 \\ x_{10}w_1 & x_{10}w_2 & x_{11}w_1 & x_{11}w_2 \\ x_{10}w_3 & x_{10}w_4 & x_{11}w_3 & x_{11}w_4 \end{bmatrix}$$

Flattening Y and writing it as an Affine transformation gives

$$Y = \begin{bmatrix} x_{00}w_1 \\ x_{00}w_2 \\ x_{01}w_1 \\ x_{01}w_2 \\ x_{00}w_3 \\ x_{00}w_4 \\ x_{01}w_3 \\ x_{01}w_4 \\ x_{10}w_1 \\ x_{10}w_2 \\ x_{11}w_1 \\ x_{11}w_2 \\ x_{10}w_3 \\ x_{10}w_4 \\ x_{11}w_3 \\ x_{11}w_4 \end{bmatrix} = \underbrace{\begin{bmatrix} w_1 & 0 & 0 & 0 \\ w_2 & 0 & 0 & 0 \\ 0 & w_1 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ w_3 & 0 & 0 & 0 \\ w_4 & 0 & 0 & 0 \\ 0 & w_3 & 0 & 0 \\ 0 & w_4 & 0 & 0 \\ 0 & 0 & w_1 & 0 \\ 0 & 0 & w_2 & 0 \\ 0 & 0 & 0 & w_1 \\ 0 & 0 & 0 & w_2 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & w_4 & 0 \\ 0 & 0 & 0 & w_3 \\ 0 & 0 & 0 & w_4 \end{bmatrix}}_{A_T} \begin{bmatrix} x_{00} \\ x_{01} \\ x_{10} \\ x_{11} \end{bmatrix} \qquad (2)$$

From (1) and (2), we can see that A_T has the same rows as A_c but with a different ordering. Thus, convolution with a kernel size (4,1,1,1) is identical to a transposed convolutional layer with kernel size (1,1,2,2) with only a difference in ordering of the flattened elements of Y

3-1 The parity sequence is just the running XOR of the input sequence



~~XOR(a,b~~ XOR of 2 bits a,b can be analytically represented as

$XOR(a,b) = a+b-ab$

The equation of the hidden unit is, therefore,

$h_t = h_{t-1} + x_t - h_{t-1}x_t$

$y_t = h_t$ ($\Rightarrow$ identity activation)

$h_0 = 0$

Alternate solution:

~~Consider a hidden state with the the following equation~~

~~$h_t = h_{t-}$~~

This can also be implemented with 2 hidden units, one of them computing AND and the other computing OR:

$h_{1,t} = h_{1,t-1} + x_t - h_{2,t-1} - 0.5$ (for OR)

$h_{2,t} = h_{2,t-1} + h_{1,t} + x_t - 1.5$ (for AND)

$y_t = h_{1,t} - h_{2,t} - 0.5$ (XOR)

3-2 $\cancel{h_T = W^T h_0}$

$h_1 = W' h_0$ (' denotes transpose)

$h_2 = W' h_1 = W' W' h_0 = (W')^2 h_0$

.

$h_T = (W')^T h_0$ ——— ①

As W is a square matrix, it can be expressed as follows:

$W = PDP^{-1}$

where the columns of P are the eigenvectors of W and D is a diagonal matrix comprising the eigenvalues of W along its diagonal.

Now,

$W' = (P^{-1})' D' P' = (P')^{-1} D P'$ ($\because$ D is diagonal and using properties of invertible matrices)

$(W')^2 = (P')^{-1} D P' (P')^{-1} D P'$

$= (P')^{-1} D^2 P'$

.

$(W')^T = (P')^{-1} D^T P'$

From ①,

$h_T = (P')^{-1} D^T P' h_0$

$$\frac{\partial h_T}{\partial h_0} = (P')^{-1} D^T P'$$

If $T >> 0$ and $\rho(W) < 1$, the elements of D^T will go to 0 resulting in a "vanishing" gradient.

If $\rho(W) > 1$, at least one value of D^T (corresponding to the largest eigenvalue) will go to ∞, resulting in an "exploding gradient".

2-1 If G is a DAG, it has a node with no incoming edges (from the given lemma). Let v_1 be a vertex with no incoming edges.

If v_1 is removed from G, the resulting graph $G-\{v_1\}$ is still cyclic as removal of edges cannot introduce cyclicity. In addition to this, there is some vertex with no incoming edges in the resulting graph. Let's call it v_2. If we remove v_2, the resulting graph $G-\{v_1,v_2\}$ will still have the above properties (i.e. absence of cycles and a vertex with no incoming edges). Repeat this till every vertex is removed and store the vertices in the order of their removal. This order is a topological order because

1. An edge (v_i, v_j) must be deleted before v_j is removed
2. Hence, v_i must be removed before v_j.

$\Rightarrow$ $i<j \ \forall (v_i, v_j)$ which is the definition of topological ordering.

2-2 Let's assume that DAG has a cycle. Let the edges in this cycle be $(v_{c0}, v_{c1}), (v_{c1}, v_{c2}), \cdot \; (v_{cn}, v_{c0})$.

As G has a topological order, for the edges in the cycle.

$$v_{c0} < v_{c1} < v_{c2} \cdot < \boxed{v_{cn} < v_{c0}}$$

$\rightarrow$ Reduction ad absurdum!

$\Rightarrow$ G has no cycles or it is a DAG