1. $Y = T(X)$

$\Rightarrow$ $\tilde{X} := X.\text{reshape}(mn)$

$\tilde{Y} := Y.\text{reshape}(\frac{mn}{4})$

$$\tilde{Y}_{\frac{m}{2}(j-1)+i} = Y_{ij} = \frac{1}{4}\left(X_{2i-1,2j-1} + X_{2i,2j-1} + X_{2i-1,2j} + X_{2i,2j}\right)$$

$$= \frac{1}{4}\left(\tilde{X}_{m(2j-2)+2i-1} + \tilde{X}_{m(2j-2)+2i} + \tilde{X}_{m(2j-1)+2i-1} + \tilde{X}_{m(2j-1)+2i}\right)$$

$$= \frac{1}{4}\left(e_{m(2j-2)+2i-1} + e_{m(2j-2)+2i} + e_{m(2j-1)+2i-1} + e_{m(2j-1)+2i}\right)^T \tilde{X},$$

$i = 1, \cdots, \frac{m}{2}$; $j = 1, \cdots, \frac{n}{2}$

e_k : k-th unit vector in $\mathbb{R}^{mn}$

$$a_{ij} := \frac{1}{4}\left(e_{m(2j-2)+2i-1} + e_{m(2j-2)+2i} + e_{m(2j-1)+2i-1} + e_{m(2j-1)+2i}\right) \in \mathbb{R}^{mn},$$

$$A := \left[a_{11}, \cdots, a_{\frac{m}{2},1}, a_{12}, \cdots, a_{\frac{m}{2},2}, \cdots, a_{1,\frac{n}{2}}, \cdots, a_{\frac{m}{2},\frac{n}{2}}\right]^T \in \mathbb{R}^{\frac{mn}{4} \times mn}$$

Then, $\tilde{Y} = A\tilde{X}$.

$\therefore$ $T(X) = Y = \left(A\left(X.\text{reshape}(mn)\right)\right).\text{reshape}\left(\frac{m}{2}, \frac{n}{2}\right)$

Now, describe nearest neighbor upsampling $X = g(Y)$, $X \in \mathbb{R}^m$

$$X_{ij} = Y_{\lceil \frac{i}{2}\rceil, \lceil \frac{j}{2}\rceil} \iff \begin{cases} X_{2i-1,2j-1} = Y_{ij} \\ X_{2i,2j-1} = Y_{ij} \\ X_{2i-1,2j} = Y_{ij} \\ X_{2i,2j} = Y_{ij} \end{cases} \iff \begin{cases} \tilde{X}_{m(2j-2)+2i-1} = f^T_{m(j-1)+i}\tilde{Y} \\ \tilde{X}_{m(2j-2)+2i} = f^T_{m(j-1)+i}\tilde{Y} \\ \tilde{X}_{m(2j-1)+2i-1} = f^T_{m(j-1)+i}\tilde{Y} \\ \tilde{X}_{m(2j-1)+2i} = f^T_{m(j-1)+i}\tilde{Y} \end{cases}$$

f_k : k-th unit vector in $\mathbb{R}^{mn/4}$

$\tilde{X} := B\tilde{Y}$, then,

$$(m(j-1)+i)\text{-th Column of } B = e_{m(2j-2)+2i-1} + e_{m(2j-2)+2i} + e_{m(2j-1)+2i-1} + e_{m(2j-1)+2i}$$

$$= 4a_{ij}$$

Hence, $B = 4A^T$,

$$T^T(Y) = \left(A^T\left(Y.\text{reshape}\left(\frac{mn}{4}\right)\right)\right).\text{reshape}(m,n)$$

$$= \left(\frac{1}{4}B\left(Y.\text{reshape}\left(\frac{mn}{4}\right)\right)\right).\text{reshape}(m,n)$$

$$= \frac{1}{4}\left(B\left(Y.\text{reshape}\left(\frac{mn}{4}\right)\right)\right).\text{reshape}(m,n)$$

$$= \frac{1}{4} \cdot \text{nearest neighbor upsampling}(Y).$$

That is, T^T is a constant times the nearest neighbor upsampling.

$\square$

2.

```
layer = nn.ConvTranspose2d( in_channels = 1,
                            out_channels = 1,
                            kernel_size = r,
                            stride = r,
                            bias = False )

layer.weight.data = torch.ones(1, 1, r, r)
```

$\square$

3.

(a) $x_{(n)} := \max\{x_1, \cdots, x_n\}$

$$\nu_\beta(x) = \frac{1}{\beta} \log\left(\exp(\beta x_{(n)}) \sum_{i=1}^{n} \exp \beta(x_i - x_{(n)})\right)$$

$$= \frac{1}{\beta} \cdot \beta x_{(n)} + \frac{1}{\beta} \log \sum_{i=1}^{n} \exp \beta(x_i - x_{(n)})$$

$$= x_{(n)} + \frac{1}{\beta} \log\Big(1 + \sum_{\substack{i=1 \\ x_i \neq x_{(n)}}}^{n} \exp(\beta(x_i - x_{(n)}))\Big) \xrightarrow[\beta \to \infty]{} x_{(n)}$$

(b)

$$\frac{\partial}{\partial x_i} \nu_\beta(x) = \frac{1}{\beta} \frac{(\exp(\beta x_i))'}{\sum_{i=1}^{n} \exp(\beta x_i)} = \frac{1}{\cancel{\beta}} \frac{\cancel{\beta} \exp(\beta x_i)}{\sum_{i=1}^{n} \exp(\beta x_i)}$$

$$\therefore \nabla \nu_1 = \frac{1}{\sum_{i=1}^{n} \exp(x_i)} \begin{bmatrix} \exp(x_1) \\ \vdots \\ \exp(x_n) \end{bmatrix} = \mu \quad \text{is the softmax function.}$$

(c)

$$\frac{\partial}{\partial x_i} \nu_\beta(x) = \frac{\exp(\beta x_i)}{\sum_{i=1}^{n} \exp(\beta x_i)} = \frac{\exp(\beta(x_i - x_{(n)}))}{1 + \sum_{\substack{i=1 \\ x_i \neq x_{(n)}}}^{n} \exp(\beta(x_i - x_{(n)}))} \xrightarrow[\beta \to \infty]{} \begin{cases} 1 & \text{if } i = i_{max} \\ 0 & \text{if } i \neq i_{max} \end{cases}$$

$$\therefore \nabla \nu_\beta = \frac{1}{\sum_{i=1}^{n} \exp(\beta x_i)} \begin{bmatrix} \exp(\beta x_1) \\ \vdots \\ \exp(\beta x_n) \end{bmatrix} \xrightarrow[\beta \to 0]{} e_{i_{max}}$$

$\square$

4. Given $t \in \mathbb{R}$, Show that $G(u) \le t \iff u \le F(t)$

($\Rightarrow$) $G(u) = \inf \{ x \in \mathbb{R} : u \le F(x) \} \le t$

$\Rightarrow \exists\, t' < t$ s.t. $t' \in \{ x \in \mathbb{R} : u \le F(x) \}$

$\iff u \le F(t')$

Hence, $u \le F(t') \le F(t)$

($\Leftarrow$) $u \le F(t)$

$\Rightarrow \{ x \in \mathbb{R} : u \le F(x) \} \supseteq \{ x \in \mathbb{R} : F(t) \le F(x) \} = \{ x \in \mathbb{R} : t \le x \}$

$\Rightarrow G(u) = \inf \{ x \in \mathbb{R} : u \le F(x) \} \le \inf \{ x \in \mathbb{R} : t \le x \} \le t$

$\because t \in \{ x \in \mathbb{R} : t \le x \}$

$\therefore G(u) \le t \iff u \le F(t)$

$P(G(U) \le t) = P(U \le F(t)) = F(t)$

$\square$

5.

(a) $D_f(X \| Y) = \int f\left(\frac{P_X(x)}{P_Y(x)}\right) P_Y(x)\, dx$

$= E\left[f\left(\frac{P_X(X)}{P_Y(X)}\right)\right]$, $X \sim P_Y$

$\underset{\text{Jensen}}{\ge} f\left(E\left[\frac{P_X(X)}{P_Y(X)}\right]\right) = f\left(\int \frac{P_X(x)}{\cancel{P_Y(x)}} \cancel{P_Y(x)}\, dx\right) = f(1) = 0$

(b) Let $f = -\log t$. Then,

$D_f(X \| Y) = \int -\log\left(\frac{P_X(x)}{P_Y(x)}\right) P_Y(x)\, dx = D_{KL}(X \| Y)$

Let $f = t \log t$. Then,

$D_f(X \| Y) = \int \frac{P_X(x)}{\cancel{P_Y(x)}} \log\left(\frac{P_X(x)}{P_Y(x)}\right) \cancel{P_Y(x)}\, dx$

$= \int -\log\left(\frac{P_Y(x)}{P_X(x)}\right) P_X(x)\, dx = D_{KL}(Y \| X)$

$\square$

6.

(a) $$\frac{\partial y_L}{\partial y_{L-1}} = \frac{\partial}{\partial y_{L-1}}\left(A_{w_L} y_{L-1} + b_L 1_{n_L}\right) = A_{w_L}$$

$$\frac{\partial y_\ell}{\partial y_{\ell-1}} = \frac{\partial}{\partial y_{\ell-1}} \sigma\left(A_{w_\ell} y_{\ell-1} + b_\ell 1_{n_\ell}\right)$$

$$= \begin{bmatrix} \sigma'(A_{w_\ell} y_{\ell-1} + b_\ell 1_{n_\ell})_1 (A_{w_\ell})_{1,1} & \cdots & \sigma'(A_{w_\ell} y_{\ell-1} + b_\ell 1_{n_\ell})_1 (A_{w_\ell})_{1,n_\ell} \\ \vdots & & \vdots \\ \sigma'(A_{w_\ell} y_{\ell-1} + b_\ell 1_{n_\ell})_{n_\ell} (A_{w_\ell})_{n_\ell,1} & & \sigma'(A_{w_\ell} y_{\ell-1} + b_\ell 1_{n_\ell})_{n_\ell} (A_{w_\ell})_{n_\ell,n_\ell} \end{bmatrix}$$

$$= \operatorname{diag}\left(\sigma'(A_{w_\ell} y_{\ell-1} + b_\ell 1_{n_\ell})\right) A_{w_\ell}, \quad \ell = 2, \cdots, L-1$$

$y_\ell' := A_\ell y_{\ell-1} + b_\ell, \quad \ell = 1, \cdots, L-1.$

$$\frac{\partial y_L}{\partial w_L} = y_{L-1}^T = v_L\, y_{L-1}^T = \left(C_{v_L^T} y_{L-1}\right)^T$$

$$\frac{\partial y_L}{\partial w_\ell} = \frac{\partial y_L}{\partial y_\ell}\frac{\partial y_\ell}{\partial y_\ell'}\frac{\partial y_\ell'}{\partial w_\ell}, \quad \ell = 1, \cdots, L-1$$

$$= \frac{\partial y_L}{\partial y_\ell} \operatorname{diag}\left(\sigma'(y_\ell')\right) \begin{bmatrix} (y_{\ell-1})_1 & (y_{\ell-1})_2 & \cdots & (y_{\ell-1})_{f_\ell} \\ (y_{\ell-1})_2 & (y_{\ell-1})_3 & \cdots & (y_{\ell-1})_{f_\ell+1} \\ & & & \ddots \\ (y_{\ell-1})_{n_\ell} & & \cdots & (y_{\ell-1})_{\underset{n_{\ell-1}}{\underset{\parallel}{f_\ell+n_\ell-1}}} \end{bmatrix}$$

$$= v_\ell \begin{bmatrix} (y_{\ell-1})_1 & (y_{\ell-1})_2 & \cdots & (y_{\ell-1})_{f_\ell} \\ (y_{\ell-1})_2 & (y_{\ell-1})_3 & \cdots & (y_{\ell-1})_{f_\ell+1} \\ & & & \ddots \\ (y_{\ell-1})_{n_\ell} & & \cdots & (y_{\ell-1})_{\underset{n_{\ell-1}}{\underset{\parallel}{f_\ell+n_\ell-1}}} \end{bmatrix} = \left(C_{v_\ell^T} y_{\ell-1}\right)^T$$

$$\frac{\partial y_L}{\partial b_L} = 1 = v_L 1_{n_L}$$

$$\frac{\partial y_L}{\partial b_\ell} = \frac{\partial y_L}{\partial y_\ell}\frac{\partial y_\ell}{\partial y_\ell'}\frac{\partial y_\ell'}{\partial b_\ell} = \frac{\partial y_L}{\partial y_\ell} \operatorname{diag}\left(\sigma'(y_\ell')\right) 1_{n_\ell} = v_\ell 1_{n_\ell}, \quad \ell = 1, \cdots, L-1$$

(b) To compute $A_{w_i} v$, just take w_i and slice v to perform convolution.

To compute $u^T A_{w_i} = (A^T_{w_i} u)^T$, just take w_i and slice and reverse w_i and u to perform transposed convolution.

$\square$