Deriving Neural Network Gradients

Aravind H. Vijay

June 26, 2019

Contents

1 Output Back Propagation

1.1 Activation Layer

$$\begin{aligned}
O_{by} &= \sigma\left[I_{by}\right] \\
\left(\delta O_{by}\right) &= \delta\sigma\left[I_{by}\right] \\
\left(\delta O_{by}\right) &= \sigma'\left[I_{by}\right]\left(\delta I_{by}\right) \\
\left[\left(DO1\right)_{by}\right]\left[\delta\left(O1\right)_{by}\right] &= \left[\left(DO1\right)_{by}\right]\left[\sigma'\left\{\left(I1\right)_{by}\right\}\right]\left[\delta\left(I1\right)_{by}\right] \\
\left[\left(DI1\right)_{by}\right] &= \left[\left(DO1\right)_{by}\right]\left[\sigma'\left\{\left(I1\right)_{by}\right\}\right]
\end{aligned}$$

$$\begin{aligned}
\sigma\left[x\right] &= \frac{x}{1+\left|x\right|} \\
\text{When: } x<0 \Rightarrow \sigma\left[x\right] &= \frac{x}{1-x} = \frac{1}{1-x} - 1 \\
&= \left(1-x\right)^{-1} - 1 \\
\sigma'\left[x\right] &= \frac{1}{\left(1-x\right)^{2}} \\
\sigma''\left[x\right] &= 2\left(1-x\right)^{-3} = \frac{2}{\left(1-x\right)^{3}} \\
\text{When: } x\geq 0 \Rightarrow \sigma\left[x\right] &= \frac{x}{1+x} = 1 - \frac{1}{1+x} \\
&= 1 - \left(1+x\right)^{-1} \\
\sigma'\left[x\right] &= \left(1+x\right)^{-2} = \frac{1}{\left(1+x\right)^{2}} \\
\sigma''\left[x\right] &= -2\left(1+x\right)^{-3} = \frac{-2}{\left(1+x\right)^{3}}
\end{aligned}$$

1.2 Linear Multiplication Layer

$$\begin{aligned}
(O1)_{by} &= (I2)_{yx}\,(I1)_{bx} + (I3)_y \\
(DO1)_{by}\left[\delta\,(O1)_{by}\right] &= (DO1)_{by}\,(I1)_{bx}\left[\delta\,(I2)_{yx}\right] + (DO1)_{by}\,(I2)_{yx}\left[\delta\,(I1)_{bx}\right] + (DO1)_{by}\left[\delta\,(I3)_y\right] \\
\left[(DO1)_{by}\right]\left[\delta\,(O1)_{by}\right] &= +\left[(DO1)_{by}\right]\left[(I1)_{bx}\right]\left[\delta\,(I2)_{yx}\right] \\
&\quad +\left[(DO1)_{by}\right]\left[(I2)_{yx}\right]\left[\delta\,(I1)_{bx}\right] \\
&\quad +\left[(DO1)_{by}\right]\left[\delta\,(I3)_y\right]
\end{aligned}$$

$$\begin{aligned}
(DI1)_{bx} &= (DO1)_{by}\,(I2)_{yx} \\
\Rightarrow (DI1) &= (DO1)\,(I2)
\end{aligned}$$

$$\begin{aligned}
(DI2)_{yx} &= (DO1)_{by}\,(I1)_{bx} \\
(DI2)_{yx} &= (DO1)^T\,(I1)
\end{aligned}$$

$$(DI3)_y = (DO1)_{by}$$

$$\begin{aligned}
\left(DI2^2\right)_{yx} &\equiv \left(DO1^2\right)_{by}\left(I1^2\right)_{bx} \\
\left(DI2^2\right) &\equiv \left(DO1^{2^T}\right)\left(I1^2\right) \\
\left(DI3^2\right)_y &\equiv \left(DO1^2\right)_{by}
\end{aligned}$$

2 Gradient Back Propagation

2.1 Activation Layer

$$
\begin{aligned}
\left[(O1)_{by}\right] &= \left[\sigma\left\{(I1)_{by}\right\}\right] \\
\left[(OD1)_{bzy}\right] &= \left[\sigma'\left\{(I1)_{by}\right\}\right]\left[(ID1)_{bzy}\right]
\end{aligned}
$$

$$
\begin{aligned}
+\left[(DOD1)_{bzy}\right]\left[\delta\,(OD1)_{bzy}\right] &= +\left[(DOD1)_{bzy}\right]\left[\sigma''\left\{(I1)_{by}\right\}\right]\left[(ID1)_{bzy}\right]\left[\delta\,(I1)_{by}\right] \\
+\left[(DO1)_{by}\right]\left[\delta\,(O1)_{by}\right] & \quad +\left[(DO1)_{by}\right]\left[\sigma'\left\{(I1)_{by}\right\}\right]\left[\delta\,(I1)_{by}\right] \\
& \quad +\left[(DOD1)_{bzy}\right]\left[\sigma'\left\{(I1)_{by}\right\}\right]\left[\delta\,(ID1)_{bzy}\right]
\end{aligned}
$$

$$
\begin{aligned}
\left[(DI1)_{by}\right] &= \left[(DOD1)_{bzy}\right]\left[\sigma''\left\{(I1)_{by}\right\}\right]\left[(ID1)_{bzy}\right] \\
& \quad +\left[(DO1)_{by}\right]\left[\sigma'\left\{(I1)_{by}\right\}\right] \\
\left[(DID1)_{bzy}\right] &= \left[(DOD1)_{bzy}\right]\left[\sigma'\left\{(I1)_{by}\right\}\right]
\end{aligned}
$$

2.2 Linear Multiplication Layer

$$\begin{aligned}
\left[(O1)_{by}\right] &= \left[(I2)_{yx}\right]\left[(I1)_{bx}\right] + \left[(I3)_{y}\right] \\
\left[(OD1)_{bzy}\right] &= \left[(I2)_{yx}\right]\left[(ID1)_{bzx}\right] \\
\left[(OD1)_{b}\right] &= \left[(ID1)_{b}\right]\left[(I2)^{T}\right]
\end{aligned}$$

$$\begin{aligned}
\left[(DOD1)_{bzy}\right]\left[\delta\left(OD1\right)_{bzy}\right] &= +\left[(DOD1)_{bzy}\right]\left[(ID1)_{bzx}\right]\left[\delta\left(I2\right)_{yx}\right] \\
&\quad +\left[(DOD1)_{bzy}\right]\left[(I2)_{yx}\right]\left[\delta\left(ID1\right)_{bzx}\right]
\end{aligned}$$

$$\begin{aligned}
+\left[(DOD1)_{bzy}\right]\left[\delta\left(OD1\right)_{bzy}\right] &= +\left[(DOD1)_{bzy}\right]\left[(ID1)_{bzx}\right]\left[\delta\left(I2\right)_{yx}\right] \\
+\left[(DO1)_{by}\right]\left[\delta\left(O1\right)_{by}\right] &\quad +\left[(DO1)_{by}\right]\left[(I1)_{bx}\right]\left[\delta\left(I2\right)_{yx}\right] \\
&\quad +\left[(DOD1)_{bzy}\right]\left[(I2)_{yx}\right]\left[\delta\left(ID1\right)_{bzx}\right] \\
&\quad +\left[(DO1)_{by}\right]\left[(I2)_{yx}\right]\left[\delta\left(I1\right)_{bx}\right] \\
&\quad +\left[(DO1)_{by}\right]\left[\delta\left(I3\right)_{y}\right]
\end{aligned}$$

$$
\begin{aligned}
[(DID1)_{bzx}] &= +\left[(DOD1)_{bzy}\right]\left[(I2)_{yx}\right] \\[1em]
\left[(DI2)_{yx}\right] &= +\left[(DOD1)_{bzy}\right][(ID1)_{bzx}] \\
&\quad +\left[(DO1)_{by}\right][(I1)_{bx}] \\[1em]
\left[(DI2)^2_{yx}\right] &= +\sum_b\left(\sum_z\left[((DOD1)_b)_{zy}\right][((ID1)_b)_{zx}]\right)^2 \\
&\quad +\left[((DO1)_b)_y\right][((I1)_b)_x] \\[1em]
[(DI2)] &= \left[(DO1)^T\right][(I1)]+\left[\left((DOD1)^T_b\right)\right][((ID1)_b)] \\[1em]
[(DI1)_{bx}] &= +\left[(DO1)_{by}\right]\left[(I2)_{yx}\right] \\[1em]
\left[(DI3)_y\right] &= +\left[(DO1)_{by}\right] \\
\left[(DI3)^2_y\right] &= +\left[(DO1)^2_{by}\right]
\end{aligned}
$$