Activation function

Q1:

When the neurons in different layers are passing values from one to another, non-linear activation functions will keep the values "un-zipped", which means the weights $\boldsymbol{w}_i$ will not be combined through layers as one. If not the case, assuming we use linear activation functions for the entire net, the powerful network will shrink to a single layer, and lose the abilities of extracting more complex features of input data and scale-able for similar functions.

Q2:

Let's consider a single activation function part.

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = -1 + \frac{2}{1+e^{-2x}}$$

We can see from the expressions that *tanh* function can express sigmoid function with form:

$$\sigma(x) = \frac{tanh(x/2)+1}{2}$$

The function above shows a linear relationship between a *tanh* function and a sigmoid function. This relationship is easy to be presented in a neural network (showed in following figure).

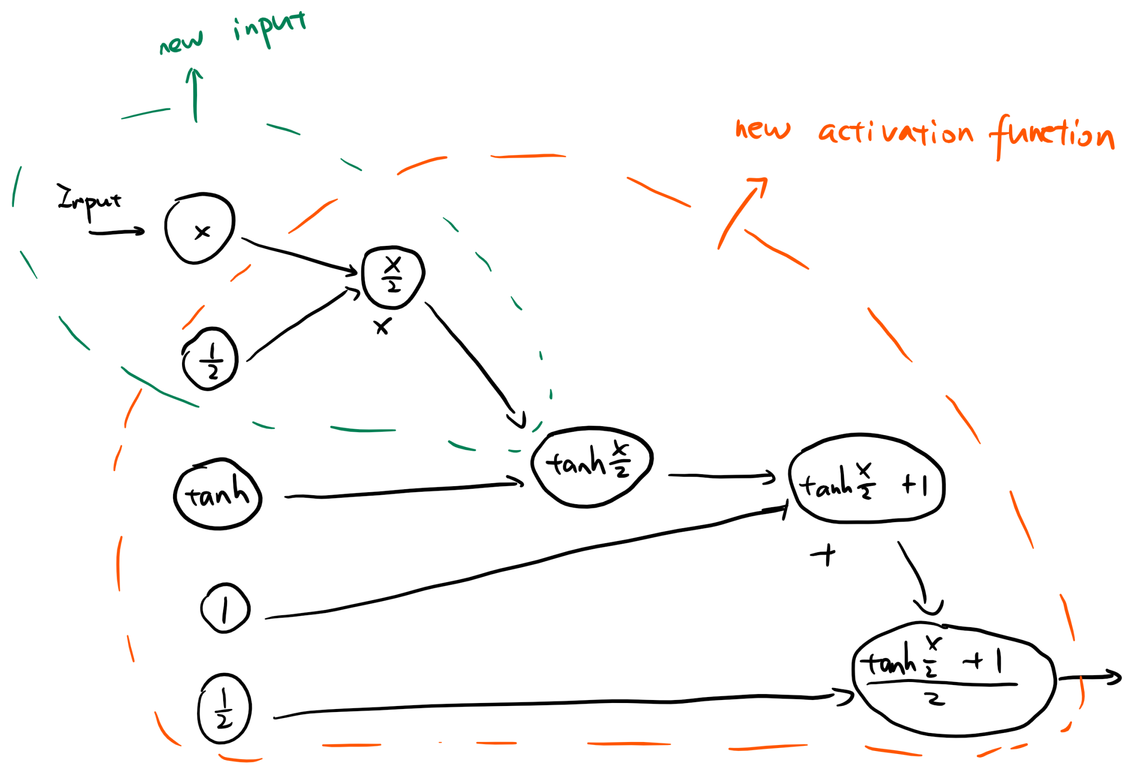

Figure 1: Sigmoid function expressed by *tanh* function

Q3:

$$
tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}
$$

$$
\begin{aligned}
tanh'(x) = & - \frac{(e^x - e^x)(e^x - e^x)}{(e^x + e^x)^2} + \frac{(e^x + e^x)^2}{(e^x + e^x)^2} \\
= & 1 - \frac{(e^x - e^x)^2}{(e^x + e^x)^2} \\
= & 1 - tanh^2(x)
\end{aligned} \tag{1}
$$

This will make computing the gradient relatively easy.

Numerical stability

Q4:

$$
\begin{aligned}
a + log(\sum_{i=1}^{N} e^{x_i - a}) = & log(e^a) + log(\sum_{i=1}^{N} e^{x_i - a}) \\
= & log(e^a \sum_{i=1}^{N} e^{x_i - a}) \\
= & log(\sum_{i=1}^{N} e^{x_i})
\end{aligned} \tag{2}
$$

Q5:

$$
\begin{aligned}
\frac{e^{x_i - a}}{\sum_{i=1}^{N} e^{x_i - a}} = & \frac{\frac{e^{x_i}}{e^a}}{\frac{\sum_{i=1}^{N} e^{x_i}}{e^a}} \\
= & \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}}
\end{aligned} \tag{3}
$$

Q6:

We take a look at the first expression:

$$
\begin{aligned}
-(y\log(\sigma(x)) + (1-y)\log(1-\sigma(x))) = & - (y\log(\frac{1}{1+e^{-x}}) + (1-y)\log(1 - \frac{1}{1+e^{-x}})) \\
= & y\log(1+e^{-x}) + (y-1)\log(\frac{e^{-x}}{1+e^{-x}}) \\
= & y\log(1+e^{-x}) + (y-1)[\log(e^{-x}) - \log(1+e^{-x})] \\
= & y\log(1+e^{-x}) + (1-y)[x + \log(1+e^{-x})] \\
= & x - xy + \log(1+e^{-x})
\end{aligned} \tag{4}
$$

And the second expression:

$$
\max(x, 0) - xy + \log(1 + e^{-abs(x)})
$$

When $x > 0$, it is clear that the two expressions are identical. When $x = 0$, both are equal to $\log(2)$, and identical too. When $x < 0$, the second expression needs a little transformation:

$$\begin{aligned} \max(x, 0) - xy + \log(1 + e^{-abs(x)}) =& - xy + \log(1 + e^{x}) \\ =& x - \log(e^{x}) - xy + \log(1 + e^{x}) \\ =& x - xy + \log(\frac{1 + e^{x}}{e^{x}}) \\ =& x - xy + \log(1 + e^{-x}) \end{aligned} \tag{5}$$

and the equivalence holds as well.