Градиент функции потерь (softmax loss):

$$p_c = \frac{\exp(s_c)}{\sum_{k=1}^{C} \exp(s_k)}$$

$$L = -\sum_{c=1}^{C} [y_c \log(p_c)] = -\sum_{c=1}^{C} \left[y_c \log\left(\frac{\exp(s_c)}{\sum_{k=1}^{C} \exp(s_k)}\right)\right]$$

$$\left(\frac{\partial p_c}{\partial s_j}\right)_{c \neq j} = \frac{\partial}{\partial s_j}\left[\frac{\exp(s_c)}{\sum_{k=1}^{C} \exp(s_k)}\right] = -\frac{\exp(s_c)\exp(s_j)}{\left(\sum_{k=1}^{C} \exp(s_k)\right)^2} = -p_c p_j$$

$$\frac{\partial p_j}{\partial s_j} = \frac{\partial}{\partial s_j}\left[\frac{\exp(s_j)}{\sum_{k=1}^{C} \exp(s_k)} - \frac{(\exp(s_j))^2}{\left(\sum_{k=1}^{C} \exp(s_k)\right)^2}\right] = p_j - (p_j)^2 = p_j(1 - p_j)$$

Градиент сумматора второго слоя (в разрезе одного наблюдения):

$$\frac{\partial L}{\partial s2_j} = -\frac{\partial}{\partial s2_j}\left(\sum_{c=1}^{C} [y_c \log(p_c)]\right) = -\sum_{c=1}^{C}\left[y_c \frac{1}{p_c}\frac{\partial p_c}{\partial s2_j}\right] =$$

$$= -y_j \frac{1}{p_j} p_j(1 - p_j) - \sum_{c \neq j}^{C}\left[y_c \frac{1}{p_c}(-p_c p_j)\right] =$$

$$= -y_j + y_j p_j + \sum_{c \neq j}^{C} [y_c p_j] = -y_j + \sum_{c=1}^{C} [y_c p_j] =$$

$$= -y_j + p_j \sum_{c=1}^{C} y_c = p_j - y_j$$

Градиент порога первого слоя:

$$\underset{[N\times H]}{\frac{\partial L}{\partial o1}} = \underset{[N\times C]}{\frac{\partial L}{\partial s2}} \times \underset{[C\times H]}{\frac{\partial s2}{\partial o1}} = \underset{[N\times C]}{\frac{\partial L}{\partial s2}} \times w2^T = \underset{[N\times C]}{(p_{ic} - y_{ic})} \times w2^T$$

Градиент по весам второго слоя:

$$\underset{[H\times C]}{\frac{\partial L}{\partial w2}} = \underset{[N\times C]}{\frac{\partial L}{\partial s2}} \times \underset{[H\times N]}{\frac{\partial s2}{\partial w2}} = \underset{[H\times N]}{\frac{\partial s2}{\partial w2}} \times \underset{[N\times C]}{\frac{\partial L}{\partial s2}} = o1^T \times \underset{[N\times C]}{(p_{ic} - y_{ic})}$$

Градиент по входным данным:

$$\underset{[N\times D]}{\frac{\partial L}{\partial x}} = \underbrace{\left(\underset{[N\times H]}{\frac{\partial L}{\partial o1}} \cdot \underset{[N\times H]}{\frac{\partial o1}{\partial s1}}\right)}_{[N\times H]} \times \underset{[H\times D]}{\frac{\partial o1}{\partial x}} = \underbrace{\left(\underset{[N\times H]}{\frac{\partial L}{\partial o1}} \cdot \underset{[N\times H]}{\frac{\partial o1}{\partial s1}}\right)}_{[N\times H]} \times \underbrace{\left(\frac{\partial s1}{\partial x}\right)^T}_{[H\times D]} = \underbrace{\left(\underset{[N\times H]}{\frac{\partial L}{\partial o1}} \cdot \underset{[N\times H]}{\frac{\partial o1}{\partial s1}}\right)}_{[N\times H]} \times \underset{[H\times D]}{\left(\underset{[D\times H]}{w1}\right)^T} =$$

$$= \underbrace{\left(\underset{[N\times H]}{\frac{\partial L}{\partial o1}} \cdot \underset{[N\times H]}{[s1 > 0]}\right)}_{[N\times H]} \times \underset{[H\times D]}{\left(\underset{[D\times H]}{w1}\right)^T} = \underbrace{\left(\underbrace{(p_{ic} - y_{ic})}_{[N\times C]} \times \underbrace{\left(w2^T\right)}_{[C\times H]}\right)}_{[N\times H]} \cdot \underset{[N\times H]}{[s1 > 0]} \times \underset{[H\times D]}{\left(\underset{[D\times H]}{w1}\right)^T}$$

Градиент по весам первого слоя :

$$\underset{[D\times H]}{\frac{\partial L}{\partial w1}} = \underset{[N\times H]}{\left(\underset{[N\times H]}{\frac{\partial L}{\partial o1}} \cdot \underset{[N\times H]}{\frac{\partial o1}{\partial s1}}\right)} \times \underset{[D\times N]}{\frac{\partial o1}{\partial w1}} =$$

$$\underset{[D\times N]}{\frac{\partial o1}{\partial w1}} \times \underset{[N\times H]}{\left(\underset{[N\times H]}{\frac{\partial L}{\partial o1}} \cdot \underset{[N\times H]}{\frac{\partial o1}{\partial s1}}\right)} = \underset{[D\times N]}{\frac{\partial o1}{\partial w1}} \times \underset{[N\times H]}{\left(\underset{[N\times H]}{\frac{\partial L}{\partial o1}} \cdot \underset{[N\times H]}{\frac{\partial o1}{\partial s1}}\right)} = \underset{[D\times N]}{\left(\underset{[N\times D]}{x}\right)^T} \times \underset{[N\times H]}{\left(\underset{[N\times H]}{\frac{\partial L}{\partial o1}} \cdot \underset{[N\times H]}{\frac{\partial o1}{\partial s1}}\right)} =$$

$$= \underset{[D\times N]}{\left(\underset{[N\times D]}{x}\right)^T} \times \underset{[N\times H]}{\left(\underset{[N\times H]}{\frac{\partial L}{\partial o1}} \cdot \underset{[N\times H]}{[s1 > 0]}\right)} = \underset{[D\times N]}{\left(\underset{[N\times D]}{x}\right)^T} \times \underset{[N\times H]}{\left(\underset{[N\times H]}{\left(\underset{[N\times C]}{(p_{ic} - y_{ic})} \times \underset{[C\times H]}{\left(w2^T\right)}\right)} \cdot \underset{[N\times H]}{[s1 > 0]}\right)}$$