Gradually Updated Neural Networks for Large-Scale Image Recognition

A. Proof for Section 4.1

Proof. Consider a linear transformation $\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ converted to GUNN, *i.e.*

$$y_i = x_i + \sum_{j=1}^{i-1} \omega_{i,j} y_j + \sum_{j=i}^{n} \omega_{i,j} x_j, \ \ \forall i \in \{1, .., n\} \tag{1}$$

Suppose that there exists a pair of neurons y_p and y_q ($p < q$) that collapse into each other. We consider one step of gradient descent on parameter ω with learning rate ϵ when the input is x and the gradient on y is $\partial L/\partial y$, *i.e.*, $\forall i, j$,

$$\Delta\omega_{i,j} = \epsilon \cdot \frac{\partial L}{\partial \omega_{i,j}} = \epsilon \cdot \frac{\partial L}{\partial y_i} \cdot \frac{\partial y_i}{\partial \omega_{i,j}} \tag{2}$$

Then, we consider Δy, which is the value difference at the same x used in gradient descent. When $\epsilon \rightarrow 0$, $\forall i$,

$$\Delta y_i = \epsilon \frac{\partial L}{\partial y_i} (\sum_{j=1}^{i-1} y_j^2 + \sum_{j=i}^{n} x_j^2) + \sum_{j=1}^{i-1} \omega_{i,j} \Delta y_j \tag{3}$$

Since y_p and y_q collapse, $\Delta y_p = \Delta y_q$, *i.e.*,

$$\epsilon \frac{\partial L}{\partial y_p} (\sum_{j=1}^{p-1} y_j^2 + \sum_{j=p}^{n} x_j^2) + \sum_{j=1}^{p-1} \omega_{p,j} \Delta y_j = \epsilon \frac{\partial L}{\partial y_q} (\sum_{j=1}^{q-1} y_j^2 + \sum_{j=q}^{n} x_j^2) + \sum_{j=1}^{q-1} \omega_{q,j} \Delta y_j \tag{4}$$

If $q > p + 1$, then we can rewrite Eq. 4 with respect to Δy_{q-1},

$$\omega_{q,q-1} \Delta y_{q-1} = \epsilon \frac{\partial L}{\partial y_p} (\sum_{j=1}^{p-1} y_j^2 + \sum_{j=p}^{n} x_j^2) + \sum_{j=1}^{p-1} \omega_{p,j} \Delta y_j - \epsilon \frac{\partial L}{\partial y_q} (\sum_{j=1}^{q-1} y_j^2 + \sum_{j=q}^{n} x_j^2) - \sum_{j=1}^{q-2} \omega_{q,j} \Delta y_j \tag{5}$$

Note that the left side of Eq. 5 is a function of $\partial L/\partial y_{q-1}$ while the right side is not. Therefore, $\omega_{q,q-1} = 0$. However, $\omega_{q,q-1} = 0$ cannot always hold after any number of gradient descent optimizations. Therefore, $q \not> p + 1$; $q = p + 1$. Thus,

$$\epsilon \frac{\partial L}{\partial y_p} (\sum_{j=1}^{p-1} y_j^2 + \sum_{j=p}^{n} x_j^2) + \sum_{j=1}^{p-1} \omega_{p,j} \Delta y_j = \epsilon \frac{\partial L}{\partial y_{p+1}} (\sum_{j=1}^{p} y_j^2 + \sum_{j=p+1}^{n} x_j^2) + \sum_{j=1}^{p} \omega_{p+1,j} \Delta y_j \tag{6}$$

We divide the both sides of Eq. 6 with $\epsilon \partial L/\partial y_p = \epsilon \partial L/\partial y_{p+1}$, and let $\partial L/\partial y_p \rightarrow \infty$. Then, we have

$$\sum_{j=1}^{p-1} y_j^2 + \sum_{j=p}^{n} x_j^2 = \sum_{j=1}^{p} y_j^2 + \sum_{j=p+1}^{n} x_j^2 + \omega_{p+1,p} (\sum_{j=1}^{p-1} y_j^2 + \sum_{j=p}^{n} x_j^2) \tag{7}$$

Eq. 7 must hold after at least one step of gradient descent on ω with input x, gradient $\partial L/\partial y$ and learning rate ϵ., *i.e.*,

$$2y_p \cdot (\epsilon \frac{\partial L}{\partial y_p} (\sum_{j=1}^{p-1} y_j^2 + \sum_{j=p}^{n} x_j^2) + \sum_{j=1}^{p-1} \omega_{p,j} \Delta y_j) + (\sum_{j=1}^{p-1} y_j^2 + \sum_{j=p}^{n} x_j^2) \cdot (\epsilon \cdot \frac{\partial L}{\partial y_p} \cdot y_p) = -\omega_{p+1,p} \Delta (\sum_{j=1}^{p-1} y_j^2 + \sum_{j=p}^{n} x_j^2) \tag{8}$$

Note that the left side of Eq. 8 is a function of $\partial L/\partial y_p$ while the right is not. Therefore, the only solution is $y_q = y_p = 0$. However, these equalities will also be broken in the next step. Thus, y_p and y_q cannot collapse into each other. ■