NOTES

1 PREVIOUS DEFINITION OR CONCLUSION

weights	$\mathrm{W} \in \mathbb{R}^{\mathrm{n}}, where\ \ n = w \times h \times c_{in} \times c_{out}$
mean(W)	$\mathrm{mean(W)} := \frac{\sum_{i=1}^{n} W_i}{n}$
std(W)	$\mathrm{std(W)} := \frac{\sum_{i=1}^{n} (W_i - mean(W_i))}{n}$
Base vector of W	$\mathrm{B_i} := sign\big(W - mean(W) + u_i * std(W)\big), \mathrm{u_i} := -1 + \frac{i-1}{M-1} * 2$
M	The number of the amount of base vectors
α	$\alpha = (B^T B)^{-1} B^T W$, where OLS is adopted
O	$\mathrm{O} = \sum_{m=1}^{n} \alpha_m Conv(B_m, A)$

2 ORIGINAL BACKWARD PROCESS

In the original paper, the gradient of weights is calculate as follow (e.q. 7 in the paper)

$$\frac{\partial \mathrm{c}}{\partial \mathrm{W}} = \frac{\partial \mathrm{c}}{\partial O}\left(\sum_{m=1}^{M} \alpha_m \frac{\partial O}{\partial B_m}\frac{\partial B_m}{\partial W}\right) =^{STE}= \frac{\partial \mathrm{c}}{\partial O}\left(\sum_{m=1}^{M} \alpha_m \frac{\partial O}{\partial B_m}\right) = \sum_{m=1}^{M} \alpha_m \frac{\partial \mathrm{c}}{\partial B_m}$$

To be more specific:

$$\frac{\partial \mathrm{c}}{\partial \mathrm{W}} = \frac{\partial \mathrm{c}}{\partial O} * \frac{\partial O}{\partial W} = \frac{\partial \mathrm{c}}{\partial O} * \left(\sum_{m=1}^{M} \alpha_m \frac{\partial O}{\partial B_m}\frac{\partial B_m}{\partial W}\right) = \sum_{m=1}^{M} \alpha_m \frac{\partial \mathrm{c}}{\partial B_m}$$

3 THE CORRECT BACKWARD GRADIENT

However the equation: $\frac{\partial c}{\partial W} = \frac{\partial c}{\partial O} * \frac{\partial O}{\partial W} = \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M} \alpha_m \frac{\partial O}{\partial B_m}\frac{\partial B_m}{\partial W}\right)$ is actually inaccurate

with or without the assumption that $\frac{\partial \alpha_m}{\partial W} = 0$.

3.1 WITH THE ASSUMPTION $\frac{\partial a_m}{\partial W} = 0$

With the assumption that $\frac{\partial a_m}{\partial W} = 0$, the correct equation should be : $\frac{\partial c}{\partial W} = \frac{\partial c}{\partial O} * \frac{\partial O}{\partial W} = \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M} \frac{\partial O}{\partial B_m} \frac{\partial B_m}{\partial W}\right) = \sum_{m=1}^{M} \frac{\partial c}{\partial B_m}$.

To be more specific,

$$\frac{\partial c}{\partial W} = \frac{\partial c}{\partial O} * \frac{\partial O}{\partial W}$$

$$= \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M} \frac{\partial O}{\partial Conv(B_m, A)} \frac{\partial Conv(B_m, A)}{\partial B_m} \frac{\partial B_m}{\partial W}\right)$$

$$= \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M} \alpha_m \frac{\partial Conv(B_m, A)}{\partial B_m} \frac{\partial B_m}{\partial W}\right)$$

$$= \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M} \frac{\partial O}{\partial B_m} \frac{\partial B_m}{\partial W}\right)$$

$$= \left(\sum_{m=1}^{M} \frac{\partial c}{\partial B_m}\right)$$

Equation 1

3.2 WITHOUT THE ASSUMPTION $\frac{\partial a_m}{\partial W} = 0$ (A MORE GENERAL EXPRESSION)

In fact, $\alpha = (B^T B)^{-1} B^T W$ which indicates that α is related to the W.

So $\frac{\partial \alpha_m Conv(B_m, A)}{\partial W} = \frac{\partial \alpha_m}{\partial W} Conv(B_m, A) + \alpha_m \frac{\partial Conv(B_m, A)}{\partial W}$. And the whole process is :

$$\frac{\partial c}{\partial W} = \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M} \left(\frac{\partial O}{\partial a_m Conv(B_m, A)} * \frac{\partial a_m Conv(B_m, A)}{\partial W}\right)\right)$$

$$= \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M} \left(\frac{\partial O}{\partial a_m Conv(B_m, A)} * \left(\frac{\partial a_m}{\partial W} * Conv(B_m, A) + \frac{\partial Conv(B_m, A)}{\partial W} * a_m\right)\right)\right)$$

Equation 2

3.3 Verify that Equation 2 is a generalized expression of Equation 1

If we add the assumption $\frac{\partial a_m}{\partial W} = 0$ to the above equation, the result has same expression as Equation 1.

$$= \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M}\left(\frac{\partial O}{\partial a_m Conv(B_m, A)} * \left(0 * Conv(B_m, A) + \frac{\partial Conv(B_m, A)}{\partial W} * a_m\right)\right)\right)$$

$$= \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M}\left(\frac{\partial O}{\partial a_m Conv(B_m, A)} * \left(\frac{\partial Conv(B_m, A)}{\partial B_m} * a_m\right) * \frac{\partial B_m}{W}\right)\right)$$

And given that $\frac{\partial O}{\partial a_m Conv(B_m, A)} = 1$,

$$\frac{\partial c}{\partial W} = \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M}\left(\left(\frac{\partial Conv(B_m, A)}{\partial B_m} * a_m\right) * \frac{\partial B_m}{W}\right)\right)$$

$$= \frac{\partial c}{\partial O} * \sum_{m=1}^{M} \frac{\partial O}{\partial B_m} * \frac{\partial B_m}{W}$$

$$= \sum_{m=1}^{M} \frac{\partial c}{\partial B_m} * \frac{\partial B_m}{W}$$

$$=^{STE}= \sum_{m=1}^{M} \frac{\partial c}{\partial B_m}$$

4 Conclusion

Recalled that in the original paper,

$$\text{Backward: } \frac{\partial c}{\partial \boldsymbol{W}} = \frac{\partial c}{\partial \boldsymbol{O}}\left(\sum_{m=1}^{M} \alpha_m \frac{\partial \boldsymbol{O}}{\partial \boldsymbol{B}_m} \frac{\partial \boldsymbol{B}_m}{\partial \boldsymbol{W}}\right) \overset{\text{STE}}{=} \frac{\partial c}{\partial \boldsymbol{O}}\left(\sum_{m=1}^{M} \alpha_m \frac{\partial \boldsymbol{O}}{\partial \boldsymbol{B}_m}\right) = \sum_{m=1}^{M} \alpha_m \frac{\partial c}{\partial \boldsymbol{B}_m}. \tag{7}$$

But the correct expression is:

$$\frac{\partial c}{\partial W} = \frac{\partial c}{\partial O} * \left(\sum_{m=1}^{M}\left(\frac{\partial O}{\partial a_m Conv(B_m, A)} * \left(\frac{\partial a_m}{\partial W} * Conv(B_m, A) + \frac{\partial Conv(B_m, A)}{\partial W} * a_m\right)\right)\right)$$

Even given the assumption that $\frac{\partial a_m}{\partial W} = 0$ the correct expression is as following which is different from e.q.7 in the original paper.

$$\frac{\partial c}{\partial W} = \sum_{m=1}^{M} \frac{\partial c}{\partial B_m}$$

It's obviously that the gradient in original paper in inaccurate or some additional assumptions are dismissed in the paper.