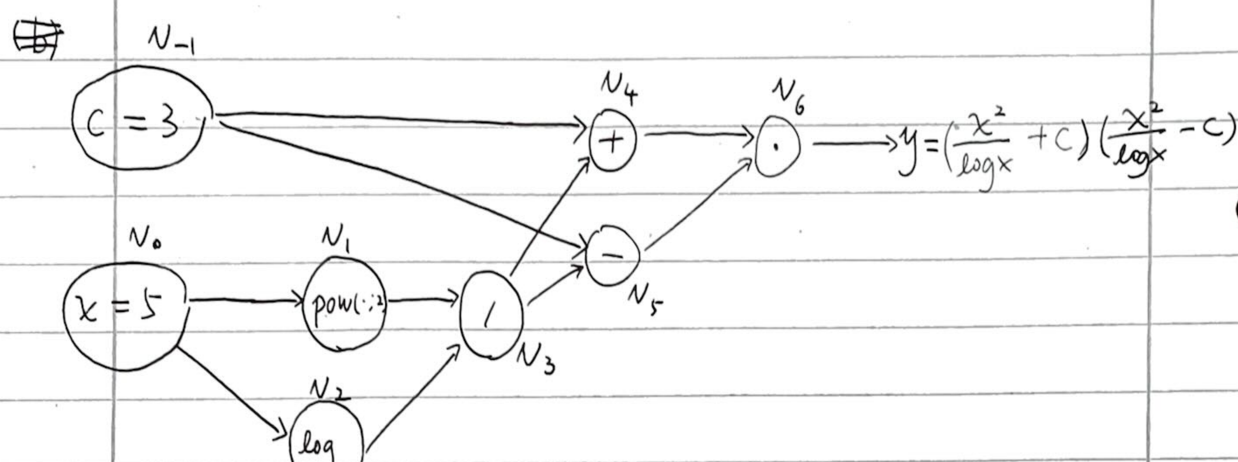
1. (a) $$\frac{\partial\left[\left(\frac{x^2}{\log x}+c\right)\left(\frac{x^2}{\log x}-c\right)\right]}{\partial x} \quad \cancel{\partial\left[\left(\frac{x^2}{\log x}+c\right)\left(\frac{x^2}{\log x}\right.\right.}$$

$$= \frac{\partial\left[\left(\frac{x^2}{\log x}+c\right)\left(\frac{x^2}{\log x}-c\right)\right]}{\partial\left(\frac{x^2}{\log x}+c\right)} \cdot \frac{\partial\left(\frac{x^2}{\log x}+c\right)}{\partial\left(\frac{x^2}{\log x}\right)} \left[\frac{\partial\left(\frac{x^2}{\log x}\right)}{\partial(x^2)} \cdot \frac{\partial(x^2)}{\partial x} + \frac{\partial\left(\frac{x^2}{\log x}\right)}{\partial(\log x)^{-1}} \cdot \frac{\partial(\log x)^{-1}}{\partial x}\right]$$

$$+ \frac{\partial\left[\left(\frac{x^2}{\log x}+c\right)\left(\frac{x^2}{\log x}-c\right)\right]}{\partial\left(\frac{x^2}{\log x}-c\right)} \cdot \frac{\partial\left(\frac{x^2}{\log x}-c\right)}{\partial\left(\frac{x^2}{\log x}\right)} \left[\frac{\partial\left(\frac{x^2}{\log x}\right)}{\partial(x^2)} \cdot \frac{\partial(x^2)}{\partial x} + \frac{\partial\left(\frac{x^2}{\log x}\right)}{\partial(\log x)^{-1}} \cdot \frac{\partial(\log x)^{-1}}{\partial x}\right]$$

~~(b)~~



(b) Forward trace:

$v_{-1} = c = 3$

$v_0 = x = 5$

$v_1 = v_0^2 = 25$

$v_2 = \log v_0 = 0.699$

$v_3 = v_1 / v_2 = 35.8$

$v_4 = v_{-1} + v_3 = 38.8$

$v_5 = v_3 - v_{-1} = 32.8$

$v_6 = v_4 \cdot v_5 = 1272.6$

$y = v_6 = 1272.6$

(c) backward trace:

$\bar{v}_6 = \bar{y} = 1$

$\bar{v}_5 = \bar{v}_6 \frac{\partial v_6}{\partial v_5} = \bar{v}_6 \times v_4 = 1 \times v_4 = 38.8$

$\bar{v}_4 = \bar{v}_6 \frac{\partial v_6}{\partial v_4} = \bar{v}_6 \times v_5 = 1 \times 32.8 = 32.8$

$\bar{v}_3 = \bar{v}_4 \frac{\partial v_4}{\partial v_3} + \bar{v}_5 \frac{\partial v_5}{\partial v_3} = \bar{v}_4 \times 1 + \bar{v}_5 \times 1 = 71.6$

$\bar{v}_2 = \bar{v}_3 \frac{\partial v_3}{\partial v_2} = \bar{v}_3 \times \left(-\frac{v_1}{v_2^2}\right) = -3663.5$

$\bar{v}_1 = \bar{v}_3 \cdot \frac{\partial v_3}{\partial v_1} = \bar{v}_3 \cdot \frac{1}{v_2} = 102.4$

$\bar{v}_0 = \bar{v}_1 \frac{\partial v_1}{\partial v_0} + \bar{v}_2 \frac{\partial v_2}{\partial v_0} = \bar{v}_1 (2v_0) + \bar{v}_2 \cdot \frac{1}{v_0} = 291.3$

$\bar{v}_{-1} = \bar{v}_4 \frac{\partial v_4}{\partial v_{-1}} + \bar{v}_5 \frac{\partial v_5}{\partial v_{-1}} = \bar{v}_4 \times 1 + \bar{v}_5 \times (-1) = -6$

$\bar{x} = \bar{v}_0 = 291.3$

$\bar{c} = \bar{v}_{-1} = -6.$

We only need to do some simple calculation in every step of backprop, which is simpler for computers.

2. (a) On iteration t:

$$m^t = \beta m^{t-1} + (1-\beta) g^t,$$ calculate the accumulated momentum

$$v^t = \gamma v^{t-1} + (1-\gamma)(g^t)^2.$$ .... ~~mom~~ second moment

$$\hat{m}^t = \frac{m^t}{1-(\beta)^t}, \quad \hat{v}^t = \frac{v^t}{1-(\gamma)^t}$$ correction of initial bias

$$w^{t+1} = w^t - \alpha \frac{\hat{m}^t}{\sqrt{\hat{v}^t} + \varepsilon}$$ update weight

$$m^0 = 0, \quad v^0 = 0.$$

usually $\beta \sim 0.99$, $\gamma \sim 0.999$, $\varepsilon \sim 10^{-8}$.

(b) $m^1 = 0 \cdot \beta + (1-\beta) g^1 = (1-\beta) g^1$, $\hat{m}^1 = \frac{m^1}{1-\beta} = g^1$.

$v^1 = \gamma \cdot 0 + (1-\gamma)(g^1)^2 = (1-\gamma)(g^1)^2$, $\hat{v}^1 = \frac{v^1}{1-\gamma} = (g^1)^2$.

$$w^2 = w^1 - \alpha \frac{g^1}{|g^1| + \varepsilon} \approx w^1 - \alpha \, \mathrm{sign}(g^1).$$

(c) $m^2 = (1-\beta)\beta g^1 + (1-\beta) g^2$, $\hat{m}^2 = \frac{m^2}{1-\beta^2} = (\beta g^1 + g^2)/(1+\beta)$

$v^2 = (1-\gamma)\gamma (g^1)^2 + (1-\gamma)(g^2)^2$, $\hat{v}_2 = \frac{v^2}{1-\gamma^2} = [\gamma (g^1)^2 + (g^2)^2] / (1+\gamma)$

$$\alpha \frac{\hat{m}^2}{\sqrt{\hat{v}^2} + \varepsilon} \approx \alpha \frac{\beta g^1 + g^2}{\sqrt{\gamma (g^1)^2 + (g^2)^2}} \frac{\sqrt{1+\gamma}}{1+\beta}$$ ($\hat{m}$ is the weighted average of gradient, $\hat{v}$ is the weighted squared average.)

(d) ~~Initialize $m^0 =$~~

(e) L2-Regularization in Loss fuction:

$$\mathcal{L} \rightarrow \mathcal{L}' = \mathcal{L} + \lambda \|w\|_2$$

$$\Rightarrow g^t \rightarrow g'^t = g^t + 2\lambda w^t,$$

~~this~~

3.(a) A perceptron in 7×7×512 maxpooling layer depends on 2×2 previous ~~out put~~ output, which depends on a (3+1)×(3+1) region of previous layer, which again depends on (3+3+1)×(3+3+1) region of previous layer, and so on.

Final result ~~tot~~ is:

|← 7×7×512 →| ← 14×14×512 → | ← 28×28×512

2×2 ← 4×4 ← 7×7 ← 10×10 ← 20×20 ← 23×23 ← 26×26

134×134 → 67×67 → 64×64 → 61×61 → 58×58 → 29×29

7×7×512 ← 14×14×512 → | ← 28×28×512

2×2 ← 4×4 ← 7×7 ← 10×10 ← 20×20 ← 23×23

which is already the whole image. ... → 28×28 → 26×26

so the FOV is 224×224.

(b) # parameters (with bias) in convolutional layers = ~~3×3×64~~

$(3\times3\times3+1)\times64 + (3\times3\times64+1)\times64$

$+ (3\times3\times64+1)\times128 + (3\times3\times128+1)\times128$

$+ (3\times3\times128+1)\times256 + (3\times3\times256+1)\times256\times2$

$+ (3\times3\times256+1)\times512 + (3\times3\times512+1)\times512\times2$

$+ (3\times3\times512+1)\times512\times3$

$= 14714688 \approx 1.47\times10^7$

parameters in fully connected layers $= (7\times7\times512\times4096+4096)$

$+ 4096\times4096+4096$

$+ 4096\times1000+1000$

$= 123642856 \approx 1.24\times10^8$

parameters in total $= 138357544 \approx 1.38\times10^8$

$\frac{\#p \text{ of convolutional}}{\#p \text{ of fully connected}} = 11.90\%$