hw5

Yuanteng Chen

3039125444

1. Depthwise Separable Convolutions

(a) learnable parameters: $(3\times3\times3)\times4=108$

(b)

$$\begin{cases} \text{Depthwise convolution}: & 3\times3\times3=27 \\ \text{Pointwise convolution}: & (1\times1\times3)\times4=12 \end{cases}$$

learnable parameters: $27+12=39$

2. Regularization and dropout

$$L(w)=\|y-Xw\|_2^2 \quad (1)$$

$$L(\tilde{w})=E_{R\sim \text{Bernoulli}(p)}\left[\|y-(R\odot X)\tilde{w}\|_2^2\right] \quad (2)$$

$$L(w)=\|y-Xw\|_2^2+\|\Gamma w\|_2^2 \quad (3)$$

(a) manipulate (2) to eliminate the expectations and get: $L(\tilde{w})=\|y-pX\tilde{w}\|_2^2+p(1-p)\|\tilde{\Gamma}\tilde{w}\|_2^2$

Solution:

assume $R\odot X=P$

$$\therefore \|y-(R\odot X)\tilde{w}\|_2^2=\|y-P\tilde{w}\|_2^2$$

$$=y^Ty+w^TP^TPw-2w^TP^Ty$$

$\therefore E_{R\sim Bernoulli(p)}[\|y-(R\odot X)\tilde{w}\|_2^2]$

$= E_{R\sim B(p)}[y^Ty + w^TP^TPw - 2w^TP^Ty]$

① $E_R[P]_{ij} = E_R[(R\odot X)_{ij}] = E_R[R_{ij}]\cdot X_{ij} = pX_{ij}$

② $E_R[2w^TP^Ty] = 2pw^TX^Ty$

③ $(E_R[(P^TP)])_{ij} = \sum_{k=1}^N E_R[R_{ki}R_{kj}X_{ki}X_{kj}]$

$$E_R[(P^TP)]_{ij} = \begin{cases} \sum_{k=1}^N E_R(R_{ki})E_R(R_{kj})\cdot X_{ki}\cdot X_{kj} = p^2(X^TX)_{ij} & (i\neq j) \\ \sum_{k=1}^N E_R[R_{ki}^2X_{ki}X_{kj}] = \sum_{k=1}^N E_R[R_{ki}^2]X_{ki}X_{kj} = p(X^TX)_{ij} & (i=j) \end{cases}$$

$$(E_R[(P^TP)])_{ij} - p^2(X^TX)_{ij} = \begin{cases} 0 & i\neq j \\ (p^2-p)(X^TX)_{ij} & i=j \end{cases}$$

$\therefore E_{R\sim B(p)}[y^Ty + w^TP^TPw - 2w^TP^Ty]$

$= y^Ty - 2pw^TX^Ty + (p^2w^TX^TXw - p^2w^TX^TXw) + w^TE_R[P^TP]w$

$= (y^Ty - 2pw^TX^Ty + p^2w^TX^TXw) - p^2w^TX^TXw + w^TE_R[P^TP]w$

$= \|y-pXw\|_2^2 + w^T(E_R[P^TP] - p^2X^TX)w$

$= \|y-pXw\|_2^2 + w^T(p^2-p)\,\mathrm{diag}(X^TX)w$

(only when $i=j$, $(E_R[(P^TP)])_{ij} - p^2(X^TX)_{ij} = (p^2-p)(X^TX)_{ij}$)

$= \|y-pXw\|_2^2 + p(1-p)\|\Gamma w\|_2^2 \quad \Gamma = \sqrt{\mathrm{diag}(X^TX)}$

(b). $L(\tilde{w}) = \|y - pX\tilde{w}\|_2^2 + p(1-p)\|\tilde{\Gamma}\tilde{w}\|_2^2$

assume $w = p\tilde{w}$

$$L(\tilde{w}) = \|y - Xw\|_2^2 + p(1-p)\|\tilde{\Gamma}\frac{w}{p}\|_2^2$$

$$= \|y - Xw\|_2^2 + \|\sqrt{\frac{1-p}{p}}\tilde{\Gamma}w\|_2^2$$

$$= \|y - Xw\|_2^2 + \|\Gamma w\|_2^2 \quad (\Gamma = \sqrt{\frac{1-p}{p}}\tilde{\Gamma})$$

(c) $L(w) = \|y - Xw\|_2^2 + \|\Gamma w\|_2^2$

$$\downarrow$$

$$L(\tilde{w}) = \|y - \tilde{X}\tilde{w}\|_2^2 + \lambda\|\tilde{w}\|_2^2$$

Sol: assume $\tilde{w} = \Gamma w \quad \therefore w = \Gamma^{-1}\tilde{w}$

$\therefore \tilde{X}\Gamma = X \quad \tilde{X} = X \cdot \Gamma^{-1}$

3. Multiplicative Regularization beyond Dropout

expected training loss:

$$L(w) = E_{R_{ij} \sim N(\mu, \sigma^2)}[\|y - (R \odot X)w\|_2^2]$$

can be put in the form:

$$L(w) = \|y - \underline{\quad (A) \quad} Xw\|_2^2 + \underline{\quad (B) \quad}\|\Gamma w\|_2^2$$

where $\Gamma = (\mathrm{diag}(X^TX))^{\frac{1}{2}}$

Sol;

in 2(a).

$$E_{R \sim B(p)}[\|y - (R \odot X)w\|_2^2]$$

$$= \|y - pXw\|_2^2 + p(1-p)\|\tilde{\Gamma}w\|_2^2 \quad \tilde{\Gamma} = (\mathrm{diag}(X^TX))^{\frac{1}{2}}$$

in Bernoulli distribution:

$$P(X=k) = p^k (1-p)^{1-k}$$

$E(X) = p$, $\quad Var(X) = p(1-p)$

in normal distribution

$E(X) = \mu$. $\quad Var(X) = \sigma^2$

A: $\mu$ $\quad$ B: $\sigma^2$

4. Analyzing Distributed Training

| | Number of Message sent | Size of each message |
|---|---|---|
| All-to-all | $n(n-1)$ | $p$ |
| Parameter server | $2n$ | $p$ |
| Ring All-Reduce | $n(2(n-1))$ | $\frac{p}{n}$ |