

Weigh initialization in a deep network

Vashing gradient, exploding gradient을 막기 위해 가중치 초기화 사용

→ 그 값이 너무 커지거나 작아지면 X

$$1) \text{Var}(W_i) = \frac{1}{n} \quad (n: \text{입력 feature 수})$$

$$2) \text{ReLU 사용 시, } \text{Var}(W_i) = \frac{2}{n^{(l-1)}}$$

$$3) \tanh \text{ 사용 시, } \text{Var}(W_i) = \frac{1}{n^{(l-1)}} \text{ or } \text{Var}(W_i) = \frac{2}{n^{(l-1)} + n^{(l)}}$$

→ Xavier initialization

Gradient checking

Parameter $w, b \xrightarrow{\text{Concat}} \theta \quad J(w, b) \rightarrow J(\theta)$

수치 미분 계산

$$d\theta_{\text{approx}}^{(i)} = \frac{J(\theta_1, \dots, \theta_i + \epsilon) - J(\theta_1, \dots, \theta_i - \epsilon)}{2\epsilon}$$

유사도 계산 $d\theta_{\text{approx}}^{(i)} \approx d\theta$

$\|d\theta_{\text{approx}}^{(i)} - d\theta\|_2$, 유clidean 거리 사용

$\frac{\|d\theta_{\text{approx}}^{(i)}\|_2 + \|d\theta\|_2}{2} \rightarrow 10^{-7}$ 보다 작다면, 계산이 잘 이루어진 것

- 속도가 매우 느리기 때문에 학습할 때 사용하지 X → 디버깅 시에 사용
- 실패했다면, 어떤 원소에서 실패하였는지 확인
- 드롭아웃에서는 적용하기 어려움

Batch vs mini-batch gradient descent

Batch gradient descent

- 전체 data를 학습하여 1 step 수행
- data의 크기가 매우 크다면 학습의 속도가 느려짐

X, Y

$M = 5,000,000$

5,000 mini-batches (size : 1000)

$$X = [X^{(1)} \ X^{(2)} \ X^{(3)} \dots X^{(1000)} \ | \ X^{(1001)} \dots X^{(2000)} \ | \ \dots \ | \ \dots X^{(M)}]$$

(n_x, m)
 $X^{(1)} : (n_x, 1000)$
 $X^{(2)}$
 \dots
 $X^{(1000)}$

$$Y = [y^{(1)} \ y^{(2)} \ y^{(3)} \dots y^{(1000)} \ | \ y^{(1001)} \dots y^{(2000)} \ | \ \dots \ | \ \dots y^{(M)}]$$

$(1, m)$
 $y^{(1)} : (1, 1000)$
 $y^{(2)}$
 \dots
 $y^{(1000)}$

Mini-batch gradient descent

(5,000 mini-batches)

for $t = 1, \dots, 5000$ {

Forward propagation

$$Z^{(t)} = W^{(t)}X^{(t)} + b^{(t)}$$

$$A^{(t)} = g^{(t)}(Z^{(t)})$$

:

:

$$A^{(t)} = g^{(t)}(Z^{(t)})$$

Compute cost

$$J^{(t)} = \frac{1}{1000} \sum_{i=1}^k L(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2 \cdot 1000} \sum_k \|W^{(k)}\|_F^2$$

Backward propagation

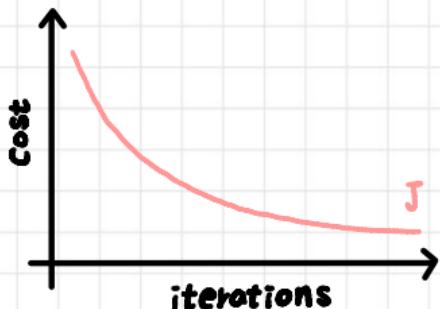
$$W^{(t)} := W^{(t)} - \alpha dW^{(t)}, \quad b^{(t)} := b^{(t)} - \alpha db^{(t)}$$

}

" / epoch "

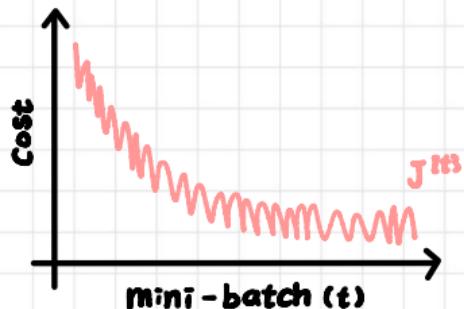
Training with mini-batch gradient descent

Batch gradient descent



→ 반복마다 비용이 감소

Mini-batch gradient descent



→ 전체적인 흐름은 감소하나 노이즈가 발생

효율적인 학습을 위해 적절한 mini-batch 사이즈를 선택하는 것이 중요