

Adam : Adaptive moment estimation

→ momentum과 RMSProp을 결합한 알고리즘

$$V_{dw} = 0, S_{dw} = 0 \quad V_{db} = 0, S_{db} = 0$$

On iteration t :

Compute dw, db on the current mini-batch

$$V_{dw} = \beta_1 V_{dw} + (1 - \beta_1) dw, \quad V_{db} = \beta_1 V_{db} + (1 - \beta_1) db \quad \text{momentum, } \beta_1$$

$$S_{dw} = \beta_2 S_{dw} + (1 - \beta_2) dw^2, \quad S_{db} = \beta_2 S_{db} + (1 - \beta_2) db^2 \quad \text{RMSProp, } \beta_2$$

$$V_{dw}^{\text{correct}} = V_{dw} / (1 - \beta_1^t), \quad V_{db}^{\text{correct}} = V_{db} / (1 - \beta_1^t)$$

$$S_{dw}^{\text{correct}} = S_{dw} / (1 - \beta_2^t), \quad S_{db}^{\text{correct}} = S_{db} / (1 - \beta_2^t)$$

hyperparameters choice

$$W := W - \alpha \frac{V_{dw}^{\text{correct}}}{\sqrt{S_{dw}^{\text{correct}}} + \epsilon} \quad b := b - \alpha \frac{V_{db}^{\text{correct}}}{\sqrt{S_{db}^{\text{correct}}} + \epsilon}$$

α : need to tune

β_1 : 0.9

β_2 : 0.999

ϵ : 10^{-8}

Learning rate decay

- 작은 미니배치 일수록 잡음이 심해서 일정한 학습률이라면 최적값에 수렴하기 어려운 현상을 볼 수 있습니다.
- 학습률 감소 기법을 사용하는 이유는 점점 학습률을 작게 줘서 최적값을 더 빨리 찾도록 만드는 것입니다.
- 다양한 학습률 감소 기법들이 있습니다.
 - 1 epoch = 전체 데이터를 1번 훑고 지나가는 횟수입니다.
 - $\alpha = \frac{1}{1 + decay\ rate \times epoch\ num} \alpha_0$
 - $\alpha = 0.95^{epoch\ num} \alpha_0$ (exponential decay라고 부릅니다.)
 - $\alpha = \frac{k}{\sqrt{epoch\ num}} \alpha_0$
 - $\alpha = \frac{k}{\sqrt{batch\ num}} \alpha_0$
 - step 별로 α 다르게 설정

Hyperparameters

α : learning rate

β : momentum

hidden units

mini-batch size

layers

learning rate decay

$\beta_1, \beta_2, \epsilon$: Adam

조정해야 하는 중요도 순서



최적의 값을 찾는 방법

1) 무작위 접근법

어떤 hyperparameter가 문제해결에
더 중요한지 알 수 없기 때문

2) 정밀화 접근

전체 공간을 탐색하여 좋은 점을 찾은 후
그 근방에서 더 정밀하게 탐색

Appropriate scale for hyperparameters

Learning rate α 를 $0.0001 - 1$ 사이의 값에서 무작위 탐색할 때,

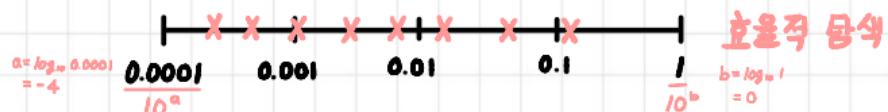
1) 선형 척도



$\rightarrow 0.0001$ 과 0.1 사이 중요한 지점을 1, 2번 정도만 탐색

2) 로그 척도

로그 척도로 $10^{-4} - 10^0$ 사이로 탐색한다면?



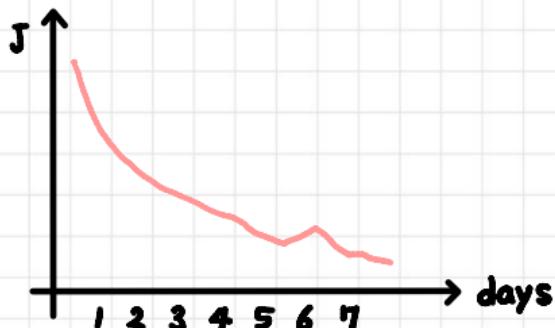
$$r = -4 * np.random.rand() + r \in [-4, 0]$$

$$\alpha = 10^r \quad 10^{-4} \dots 10^0$$

지수 가중 이동 평균에서 사용되는 β 도 동일한 이유로 로그 척도를 사용하는 것이 좋음

Hyperparameters Tuning

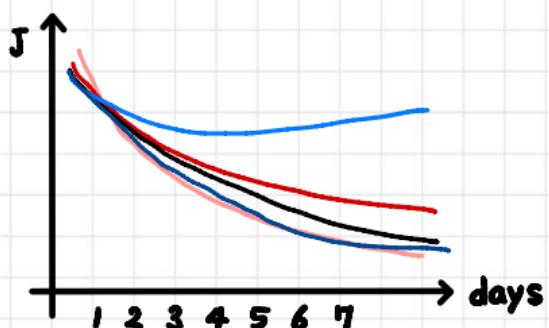
Babysitting one model



하나의 모델로 매일 성능을 지켜보면서

학습 속도를 조금씩 변경하는 방식

Training many models in parallel



동시에 여러 모델을 훈련

→ 컴퓨터의 자원이 충분할 때 사용