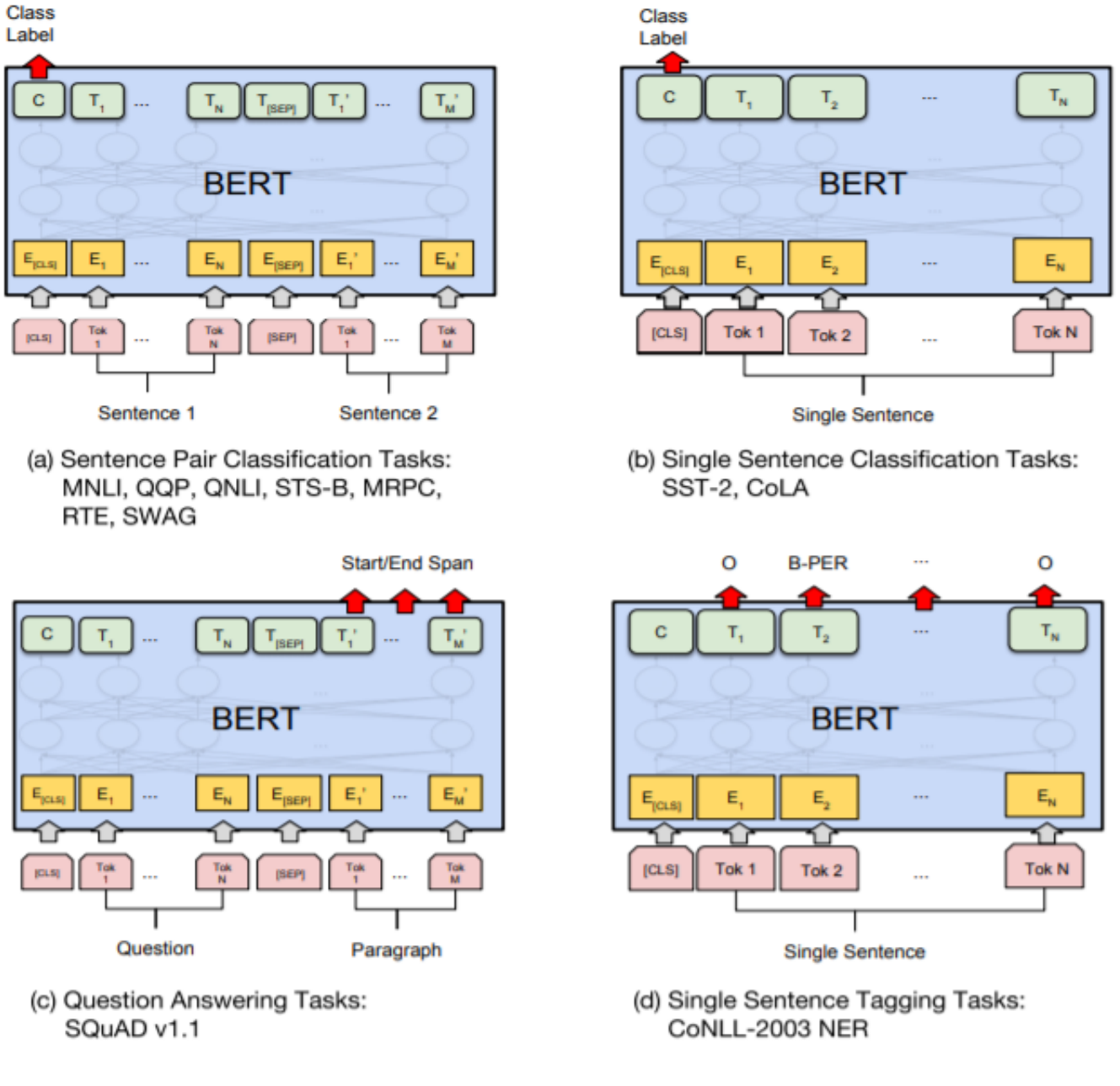
NLP2팀 5주차(2020.05.28.)

한국어 임베딩
6.1 프리트레인과 파인 튜닝 ~
6.3 단어 임베딩 활용

황인택

6.1 프리트레인과 파인 튜닝

파인 튜닝: 프리트레인 이후 추가 학습을 시행해 임베딩을 다운스트림 태스크에 맞게 업데이트하는 것



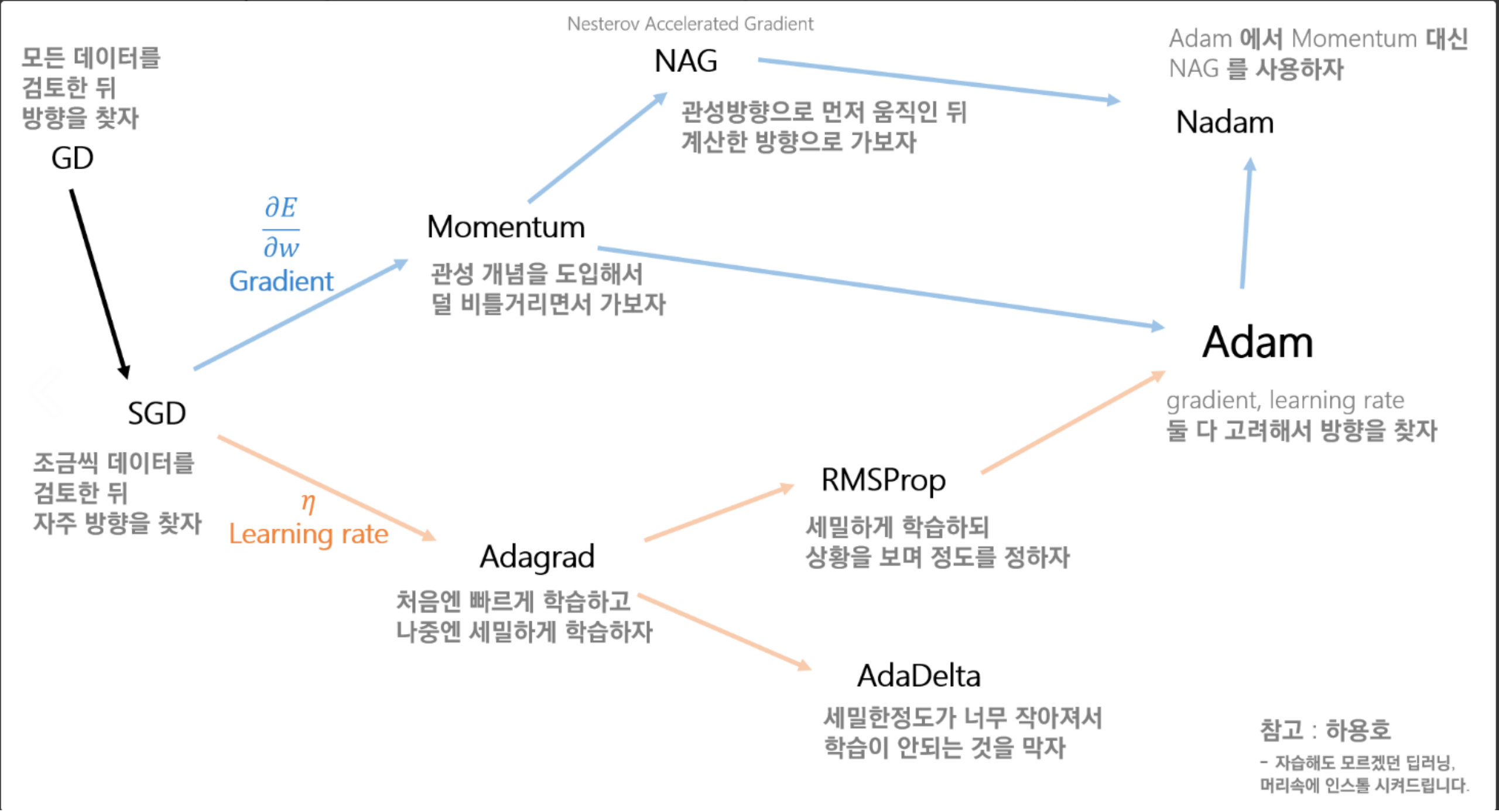
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

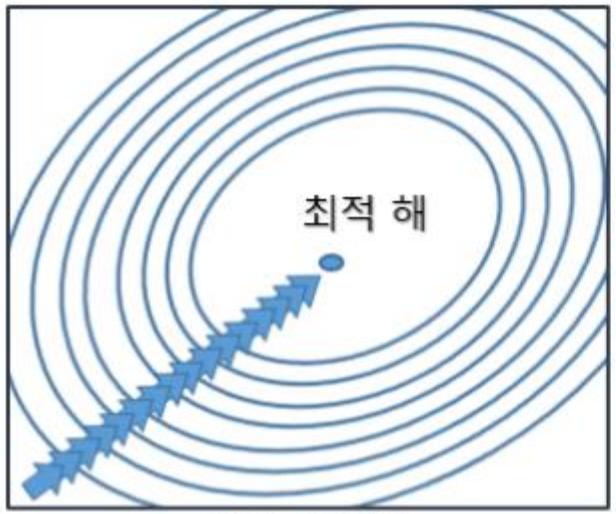
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Figure 4: Illustrations of Fine-tuning BERT on Different Tasks.

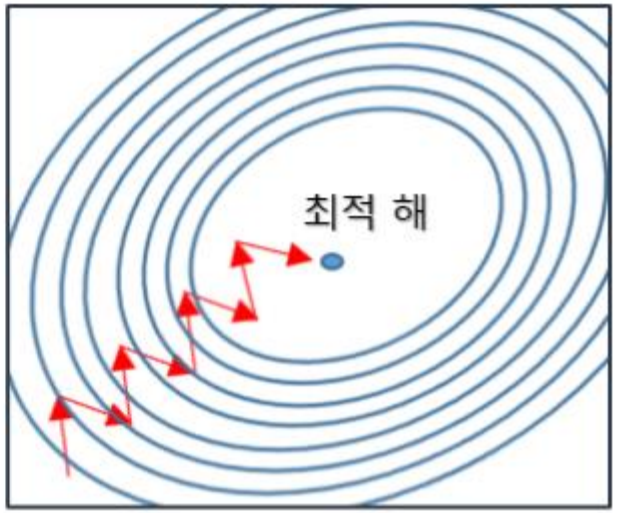

Nesterov Accelerated Gradient
모든 데이터를
검토한 뒤
방향을 찾자
GD
NAG
관성방향으로 먼저 움직인 뒤
계산한 방향으로 가보자
Adam 에서 Momentum 대신
NAG 를 사용하자
Nadam
$\frac{\partial E}{\partial w}$
Gradient
Momentum
관성 개념을 도입해서
덜 비틀거리면서 가보자
Adam
gradient, learning rate
둘 다 고려해서 방향을 찾자
SGD
조금씩 데이터를
검토한 뒤
자주 방향을 찾자
η
Learning rate
RMSProp
세밀하게 학습하되
상황을 보며 정도를 정하자
Adagrad
처음엔 빠르게 학습하고
나중엔 세밀하게 학습하자
AdaDelta
세밀한정도가 너무 작아져서
학습이 안되는 것을 막자
참고 : 하용호
- 자습해도 모르겠던 딥러닝,
머리속에 인스톨 시켜드립니다.

SGD

$$W(t+1) = W(t) - \alpha \frac{\partial}{\partial w} Cost(w)$$



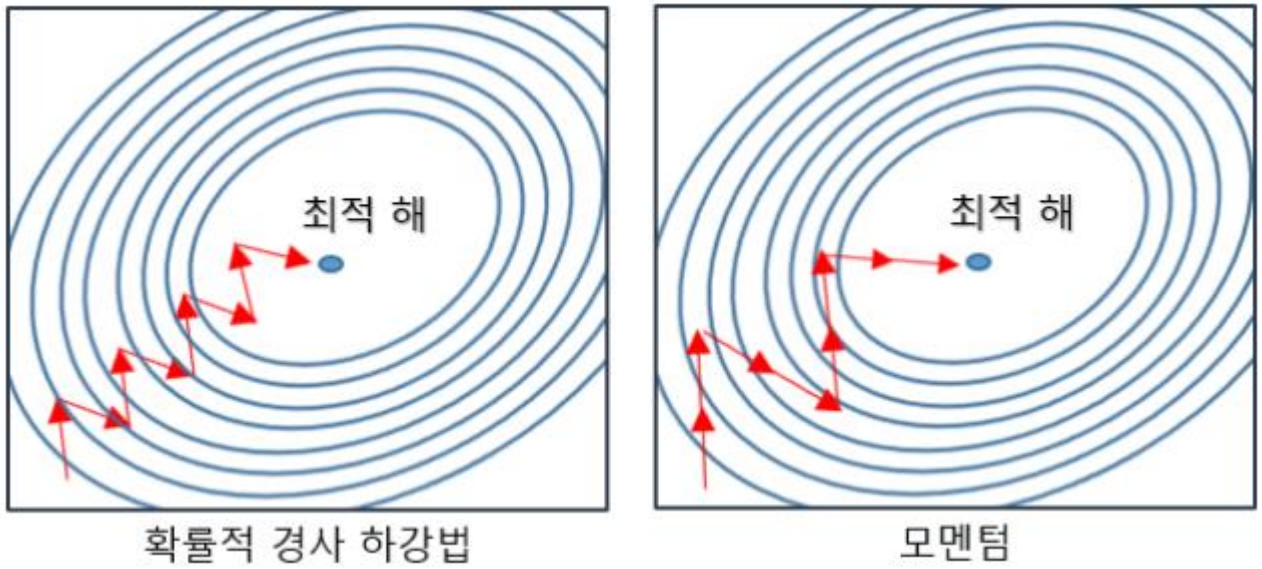
경사 하강법



확률적 경사 하강법

momentum

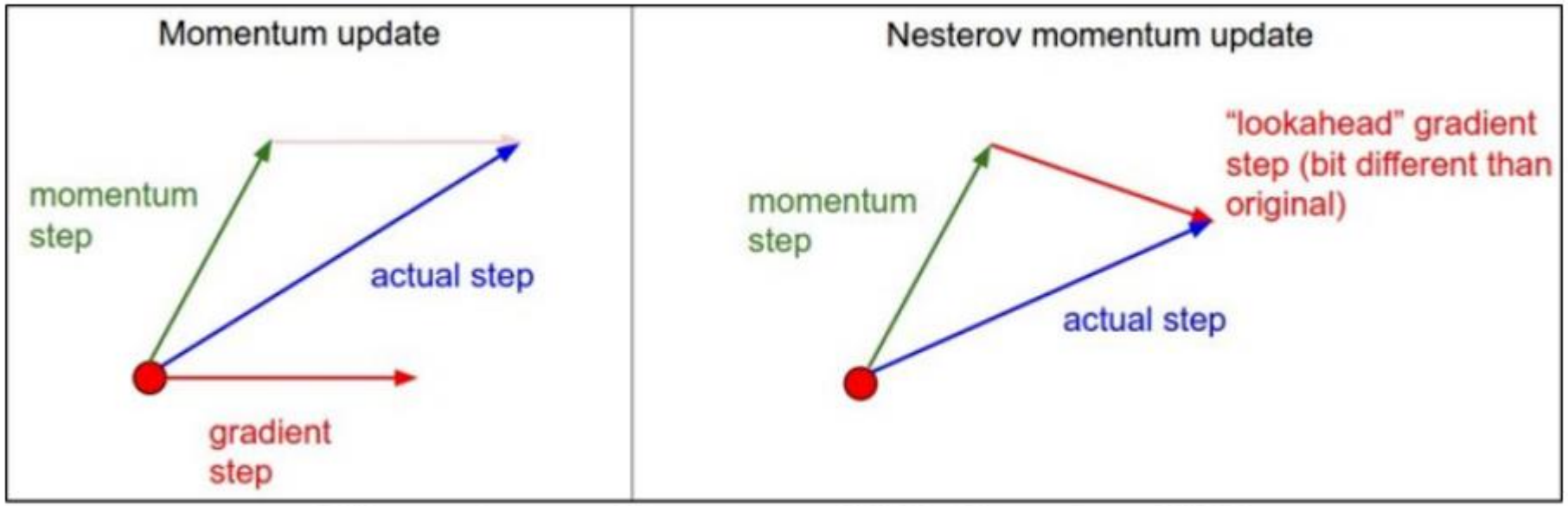
$$\mathrm{V(t)} = m * V(t-1) - \alpha \frac{\partial}{\partial w} Cost(w)$$
$$\mathrm{W(t+1)} = W(t) + V(t)$$



확률적 경사 하강법 모멘텀

Nesterov Accelrated Gradient(NAG, 네스테로프 모멘텀)

$$V(t) = m * V(t-1) - \alpha \frac{\partial}{\partial(w + m * V(t-1))} Cost(w)$$
$$W(t+1) = W(t) + V(t)$$



Difference between Momentum and NAG. Picture from CS231.

Adagrad(Adaptive Gradient, 아다그라드)

$$\mathrm{G(t)} = G(t-1) + \left(\frac{\partial}{\partial w(t)} Cost(w(t))\right)^2$$

$$= \sum_{i=0}^{t} \left(\frac{\partial}{\partial w(i)} Cost(w(i))\right)^2$$

$$\mathrm{W(t+1)} = W(t) - \alpha * \frac{1}{\sqrt{G(t)+\epsilon}} * \frac{\partial}{\partial w(i)} Cost(w(i))$$

RMSprop(알엠에스프롭)

$$\mathrm{G(t)} = \gamma G(t-1) + (1-\gamma)\left(\frac{\partial}{\partial w(i)} Cost(w(i))\right)^2$$

$$\mathrm{W(t+1)} = W(t) - \alpha * \frac{1}{\sqrt{G(t)+\epsilon}} * \frac{\partial}{\partial w(i)} Cost(w(i))$$

Adam(Adaptive Moment Estimation, 아담)

$$\mathrm{M(t)} = \beta_1 M(t-1) + (1-\beta_1)\frac{\partial}{\partial w(t)} Cost(w(t))$$

$$\mathrm{V(t)} = \beta_2 V(t-1) + (1-\beta_2)\left(\frac{\partial}{\partial w(i)} Cost(w(i))\right)^2$$

$$\widehat{M}(t) = \frac{M(\mathrm{t})}{1-{\beta_1}^t} \qquad \widehat{V}(t) = \frac{V(\mathrm{t})}{1-{\beta_2}^t}$$

$$\mathrm{W(t+1)} = W(t) - \alpha * \frac{\widehat{M}(t)}{\sqrt{\widehat{V}(t) + \epsilon}}$$

AdaDelta(Adaptive Delta, 아다델타)

$$G(t) = \gamma G(t-1) + (1-\gamma)\left(\frac{\partial}{\partial w(t)} Cost(w(t))\right)^2$$

$$\Delta w(t) = \frac{\sqrt{\Delta S(t-1) + \epsilon}}{\sqrt{G(t) + \epsilon}} * \frac{\partial}{\partial w(i)} Cost(w(i))$$

$$S(t) = \gamma S(t-1) + (1-\gamma)\,(\Delta w(t))^2$$

$$W(t+1) = W(t) - \Delta w(t)$$

단. $G(0) = 0$, $S(0) = 0$