Adam : Adagrad + RMSprop 의 장점을 섞어 놓은 것. 장점은 stepsize가 gradient의 rescaling에 영향을 받지 않는다는 것. Gradient가 커져도 stepsize는 bound 되어 있어서 어떠한 objective function을 사용한다고 하더라도 안정적으로 최적화를 위한 하강이 가능하다. 게다가 stepsize를 과거의 gradient 크기를 사용하여 adapted 시킬 수 있다.

Adagrad : 과거의 gradient 변화량을 참고하여, 이미 많이 변화한 변수들은 optima에 거의 도달했다고 보고 stepsize를 작게 하고, 많이 변화되지 않은 변수들은 아직 가야할 길이 멀다고 보고 stepsize를 크게 한다. 따라서 Gt 변수를 도입해서 여태까지의 gradient의 L2 norm을 저장한다.

$$\theta_t = \theta_{t-1} - \alpha \frac{g_t}{\sqrt{\sum_{i=1}^{t} g_i^2}}$$

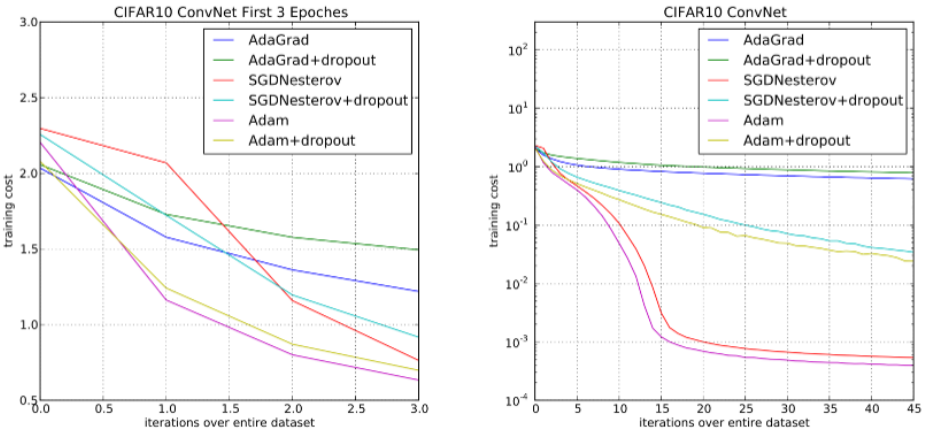
문제점 : **iteration 이 계속될수록 G가 계속 증가해서 stepsize가 너무 작아질 수 있다.** 이러한 문제를 보완하기 위해서 **RMSprop**처럼 **exponential moving average**를 사용하는 방법이 고안되었다. 수식에서 알 수 있듯이 Exponential moving average는 **과거의 정보에 가중치를 적게 부여**한다. 최근 값에 가장 민감하도록 최고 가중치를 부여하는 형태이다.

RMSprop에서는 이런 점을 반영하여 수식을 수정하였다.

$$G_t = \gamma G_{t-1} + (1-\gamma) g_t^2$$

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{G_t + \epsilon}} g_t$$

하지만 G변수의 bias를 조정하지 않으면 B1이 1에 매우 가까워질수록 매우 큰 stepsize를 갖거나 발산하는 문제점이 있는데 **이는 보통 iteration 초기에 일어난다.**



다른 기법에 비해 CNN에서 loss function를 낮게 가져가는 것을 알 수 있다. -> mitbih arrhythmia database CNN model 에 Adam을 사용하는 근거

출처: https://dalpo0814.tistory.com/29 [deeep]