Y2018 Research Plan

2018/08/22

Seojin Kim@MDLab., SKKU

Quarterly Plan

▪ 학회

	1분기	2분기	3분기	4분기
국제저널	• ICML 2019, International Conference on Machine Learning • Submission: March 1, 2019	-	• CVPR 2019, International Conference on Machine Learning • Submission: Nov 16, 2018	
국제학회	• ICCV 2019: International Conference on Computer Vision • Submission: March 1,2019	• NIPS 2019, Neural Information Processing Systems • Submission: May 18, 2019		-
국내저널	-	-	-	-

▪ 진행

	1분기	2분기	3분기	4분기
국제저널	-	-	-	-
국제학회	-	-	-	-
국내저널	-	-	-	-

Papers and Patents

▪ 논문

NO	학회명	논문명	실적구분
1			
2			
3			
4			
5			
6			
7			
8			
9			

▪ 특허

NO	출원국가	특허명	실적구분
1			
2			

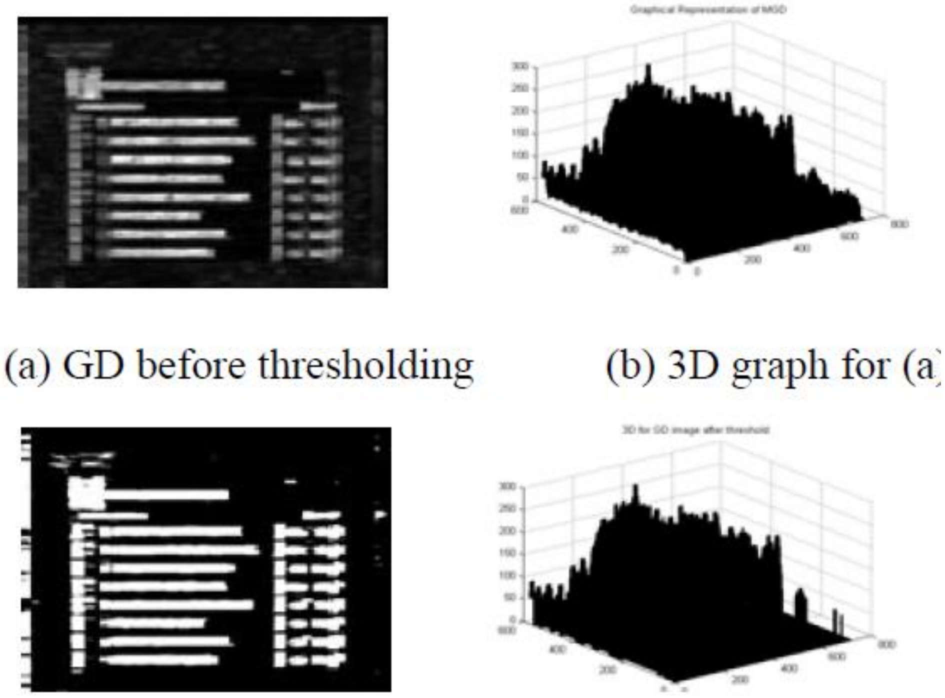
연구진행

- To Do List
  - Research
    - 구현
    - Gradient가 Input 영상에 어떤 영향을 미치는지 (부족)
  - Paper Reading(1)

- Done
  - Research
    - ~~Learning rate & decay~~
    - ~~Batch size에 따라 어떤 영향이 어떻게 나타나는지~~
    - ~~Momentum 이란?~~

Input & Gradient

- High positive and negative gradient values in text regions result from high intensity contrast between the text and background regions



(a) GD before thresholding

(b) 3D graph for (a)

(c) GD after thresholding

(d) 3D graph for (c)

Learning Rate Schedule

- Annealing the learning rate(학습속도 조정)
- How to choose the best learning rate?
  - Constant Learning Rate
    - SGD Optimizer, momentum and decay rate are zero.
  - Step decay
    - decrease the learning rate by a few epoch, usually a half in 5 epoch or a tenth in 20 epoch.
    - In real, adaptive methodology is used by validation error.
  - Exponential decay
    - $a=a_0e$^$_{-kt}$. a0,k are hyperparameter, t is the iteration number of epoch.
  - Time-based decay
    - a=a0/(1+kt) . a0,k are hyperparameter, t is the iteration number of epoch.

Step based vs Exponential

- Step decay: High accuracy
- Exponential: More smooth(stable), less time to maximize

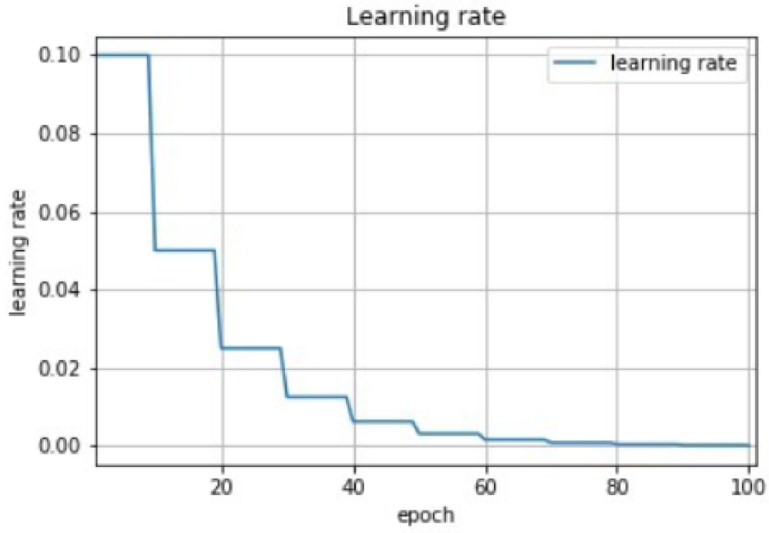


Fig 3b : Step Decay Schedule

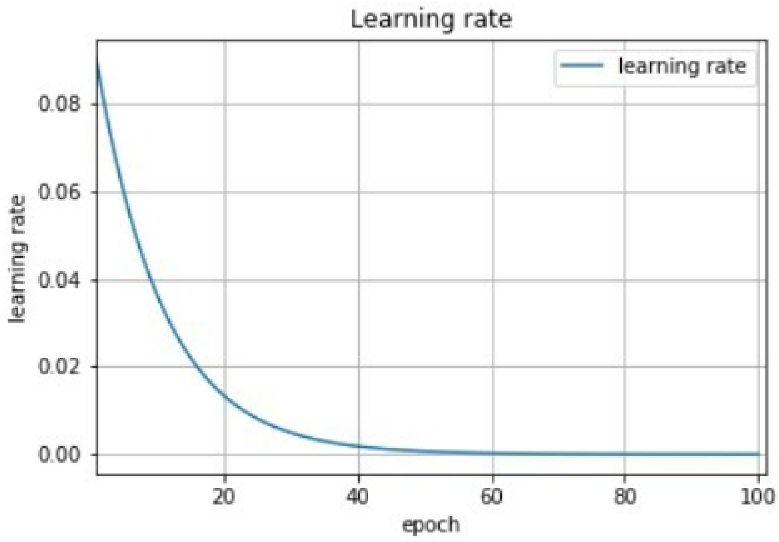


Fig 4b : Exponential Decay Schedule

Learning Rate

- train a convolutional neural network on CIFAR-10

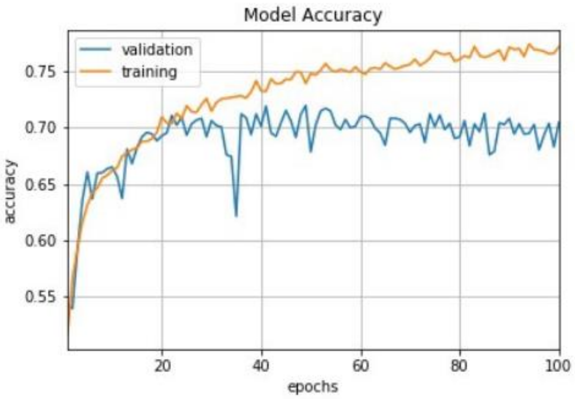


Fig 1 : Constant Learning Rate

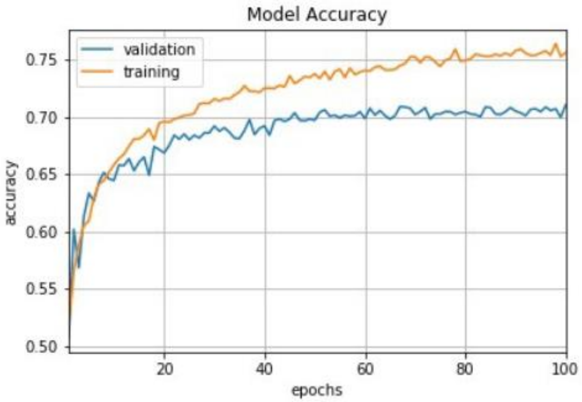


Fig 2 : Time-based Decay Schedule

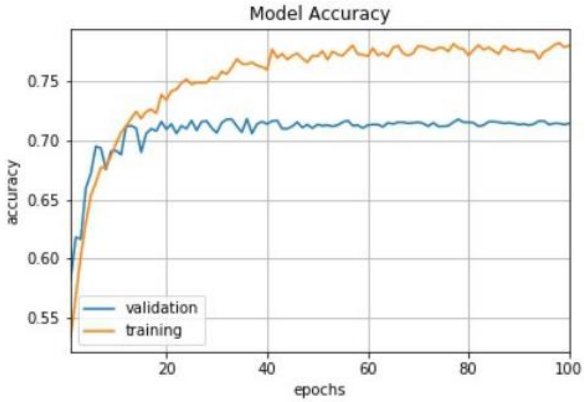


Fig 3a : Step Decay Schedule

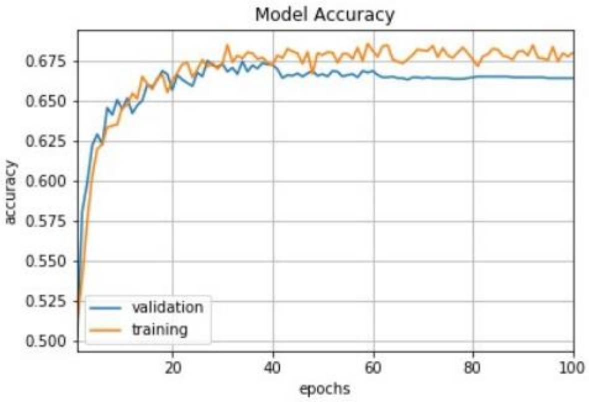
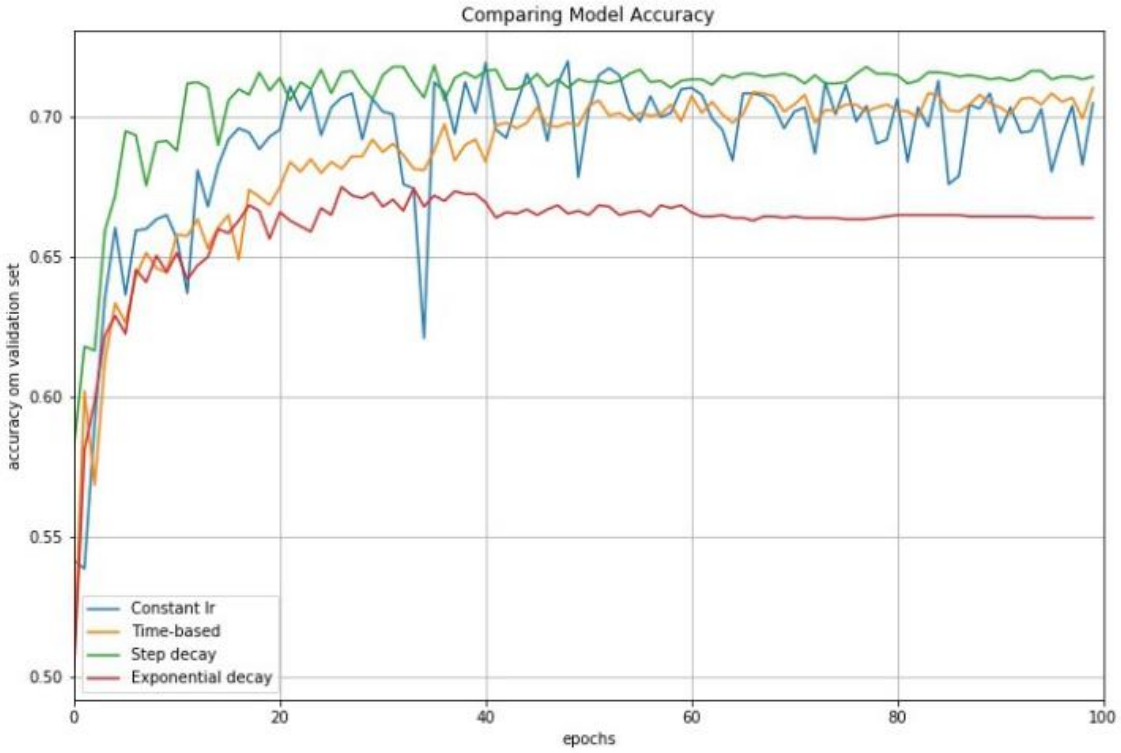


Fig 4a : Exponential Decay Schedule

Compare

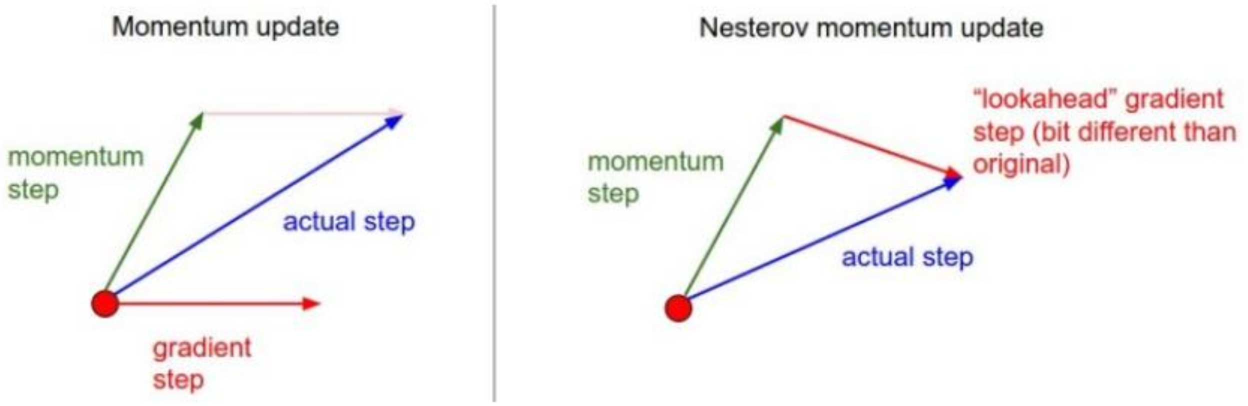


- Hyperparameters have to be defined in advance and they depend heavily on the type of model and problem.
- The same learning rate is applied to all parameter updates.
- If we have sparse data, we may want to update the parameters in different extent instead.

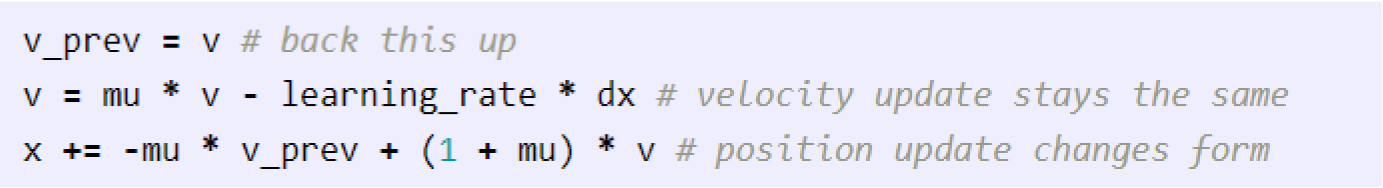
Momentum

- GD Algorithm’s disadvantage
  - 기울기 0인 점을 잘 탈출하지 못한다.
  - Training Performance가 떨어진다.
- Momentum 적용
  - 관성= 기울기가 가던 방향으로 가면서 약간씩 방향을 트는 것
  - E: 속도 상수(Learning rate)
  - 감마: 관성 상수 (Momentum rate) – 보통 0.5로 시작해서 안정화되면 0.9로 높인다.
  - 수렴하는 점을 보면 관성이 적용되었을 때 대략 어느정도 속도로 움직일 지, (값이 바뀔지)
  - 관성을 사용하는 것, 속도 상수를 보정하는 것. 감마를 0.9로 사용한다는 것, 기존 대비 약 10배 정도의 속도로 움직이도록 한다.

Nesterov Momentum



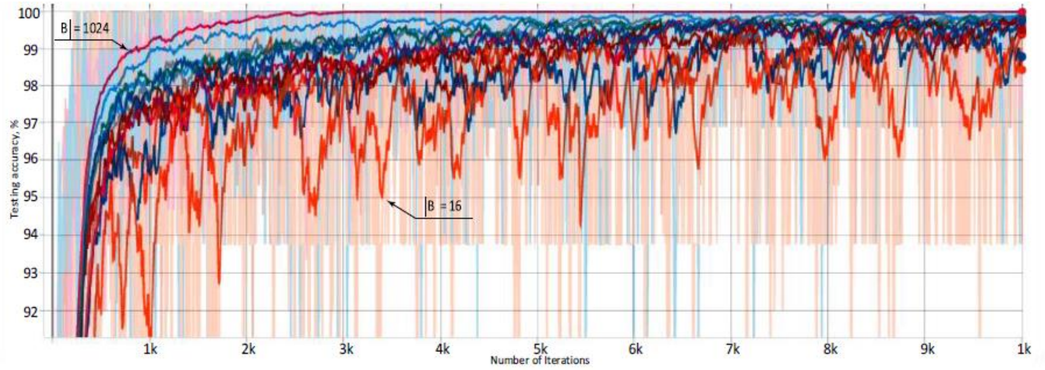
- Momentum Update: x + gradient step
- Nesterov Momentum Update: momentum step + “lookahead” gradient step -> slightly better



```
v_prev = v # back this up
v = mu * v - learning_rate * dx # velocity update stays the same
x += -mu * v_prev + (1 + mu) * v # position update changes form
```

Batch size

- Batch size
  - Used when fitting the model, controls how many predictions must be made at a time.
  - Impacts the CNN training both in terms of the time to converge and the amount of overfitting.
  - i.e. smaller batch size yields faster computation but requires visiting more examples in order to reach the same error, since there are less updates per training iteration.



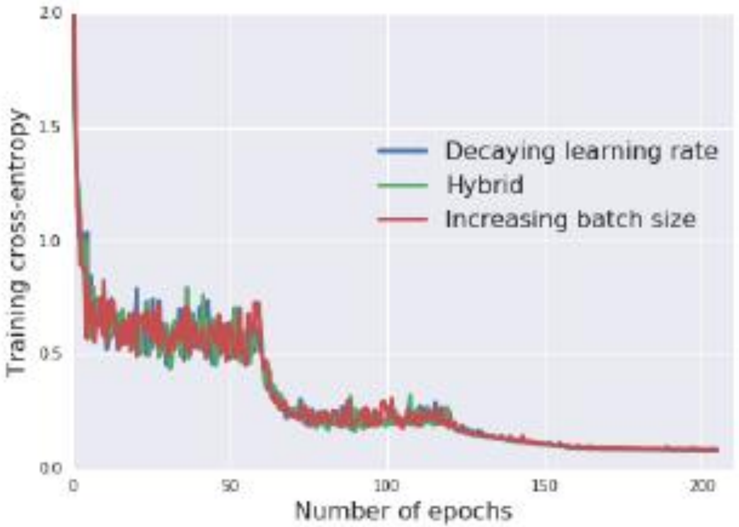
The largest the batch size value, the more smooth the curve and high accuracy.

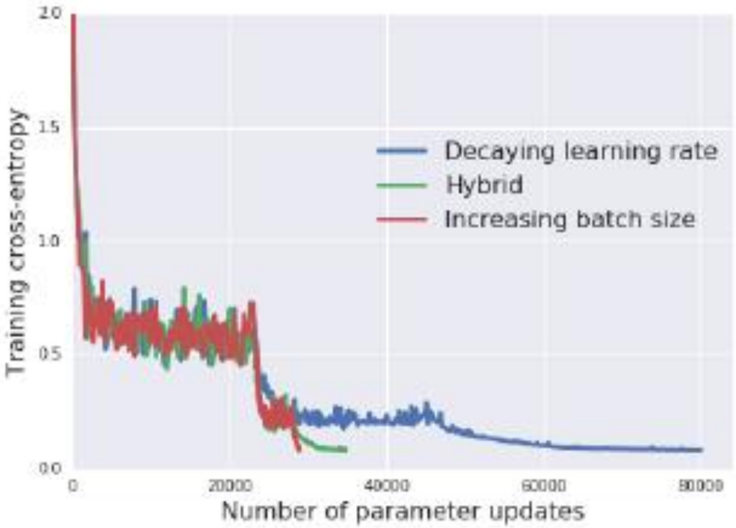
The lowest and noisiest curve corresponds to the batch size of 16 examples, the highest and the smoothest one – to the batch size of 1024 examples

Impact of Training Set Batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets (ITMS)

참고: Don’t decay the learning rate

- Decay learning rate
  - Decay the learning rate: enable to converge to the minimum of the cost function
  - Constant learning rate-> same reduction in noise by increasing the batch size





  - Wide ResNet of CIFAR10. Identical learning curves but increasing the batch size reduces the number of parameter updates required.
  - Reduce model training times without hyperparameter tuning.
  - Directly convert existing hyperparameter choices

Q&A