CSED490Y OptML
Final Presentation

Team 1
Yechan Hwang/Sungbin Shin

CONTENTS

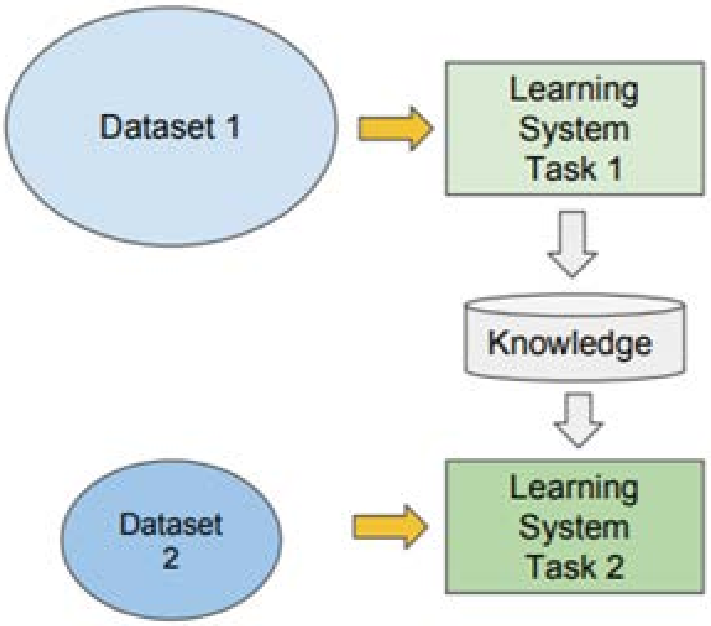
Effect of learning rate scheduling in transfer learning

- There are various factors that have to be considered during transfer learning.
  - Ex) pretrained model, batch size, optimizer…

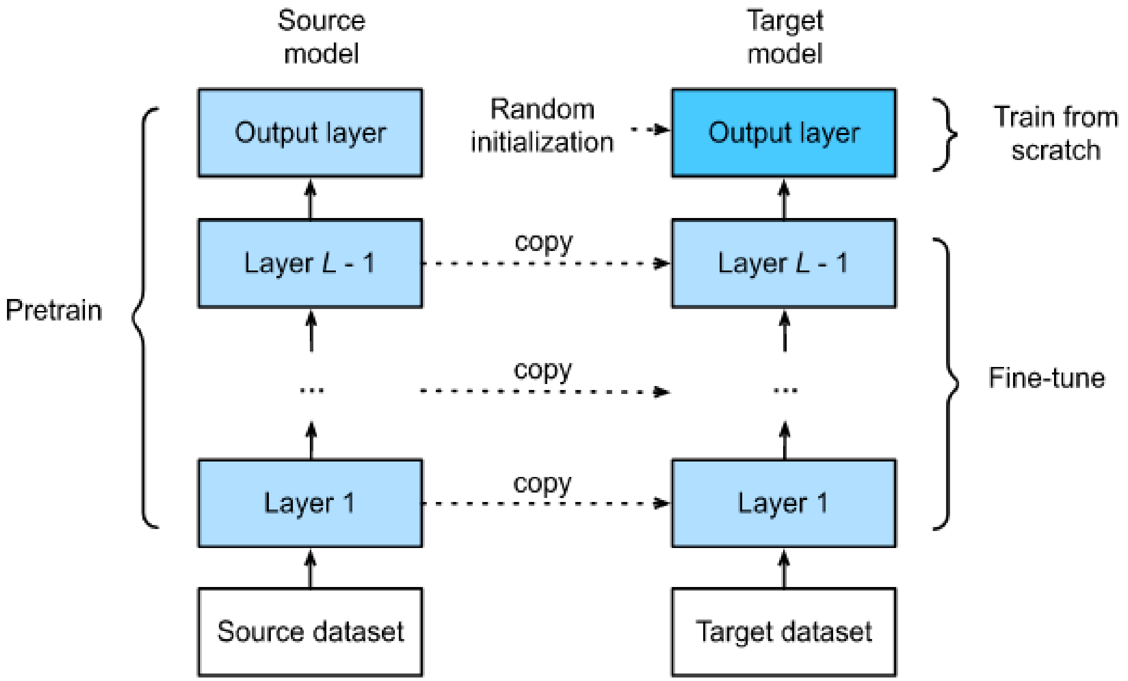
- We focused on …
  - Different learning rate scheduling
  - Convergence behavior
  - Two transfer learning scenarios

Effect of learning rate scheduling in transfer learning

Transfer learning

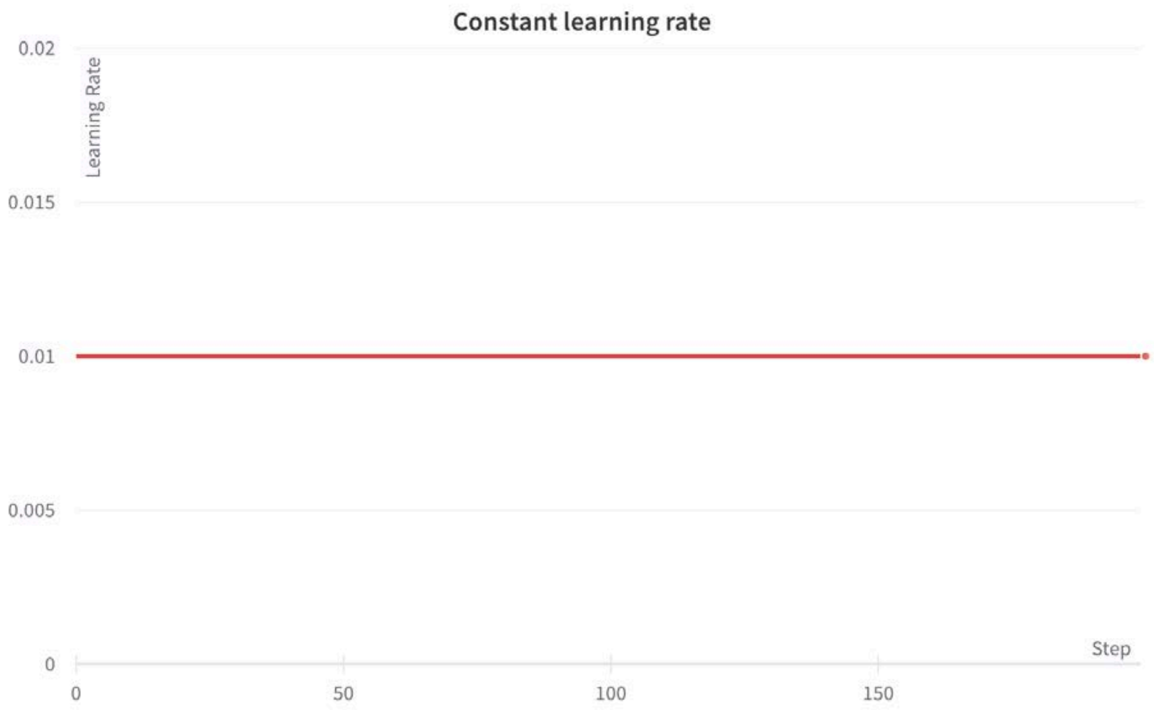


Pretrain & Fine-tune



1: Effect of hyperparameters in each lr schedulers

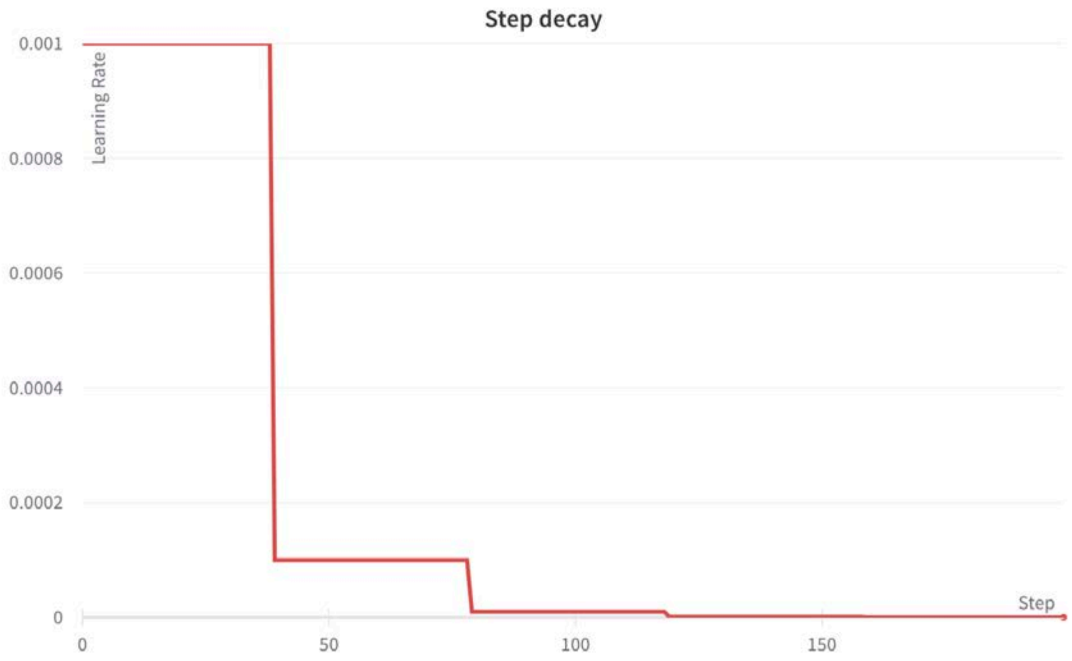
a) Constant learning rate



1: Effect of hyperparameters in each lr schedulers

b) Step learning rate decay

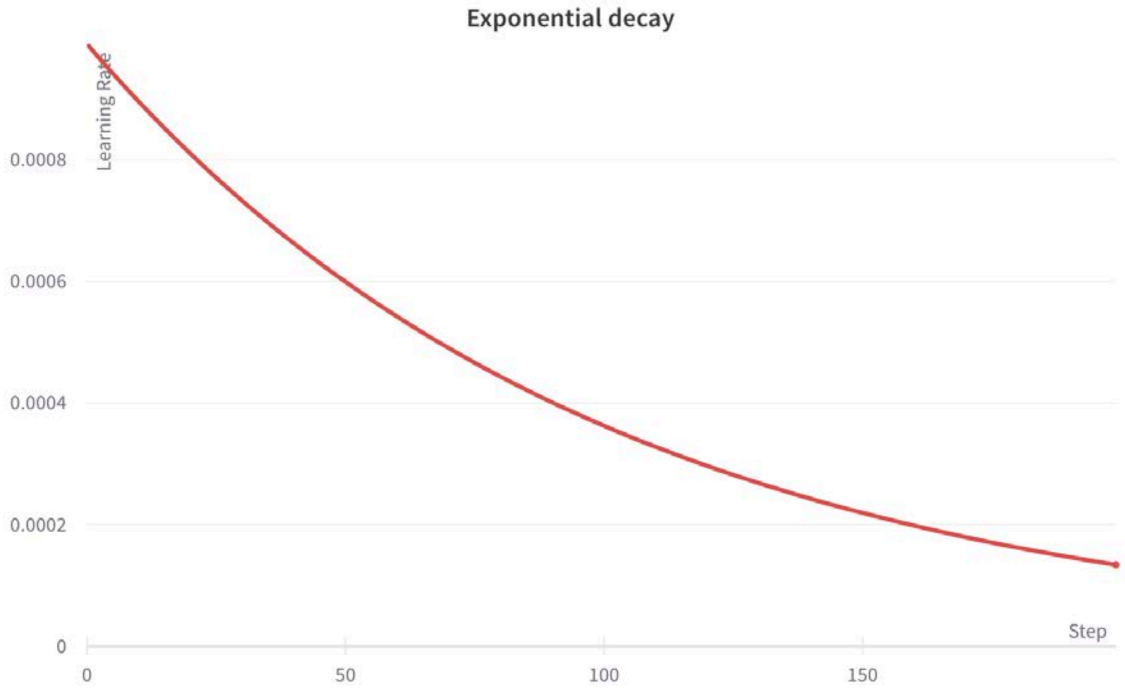
After every *step_size* epochs, new_lr = current_lr * *gamma*



1: Effect of hyperparameters in each lr schedulers

c) Exponential learning rate decay

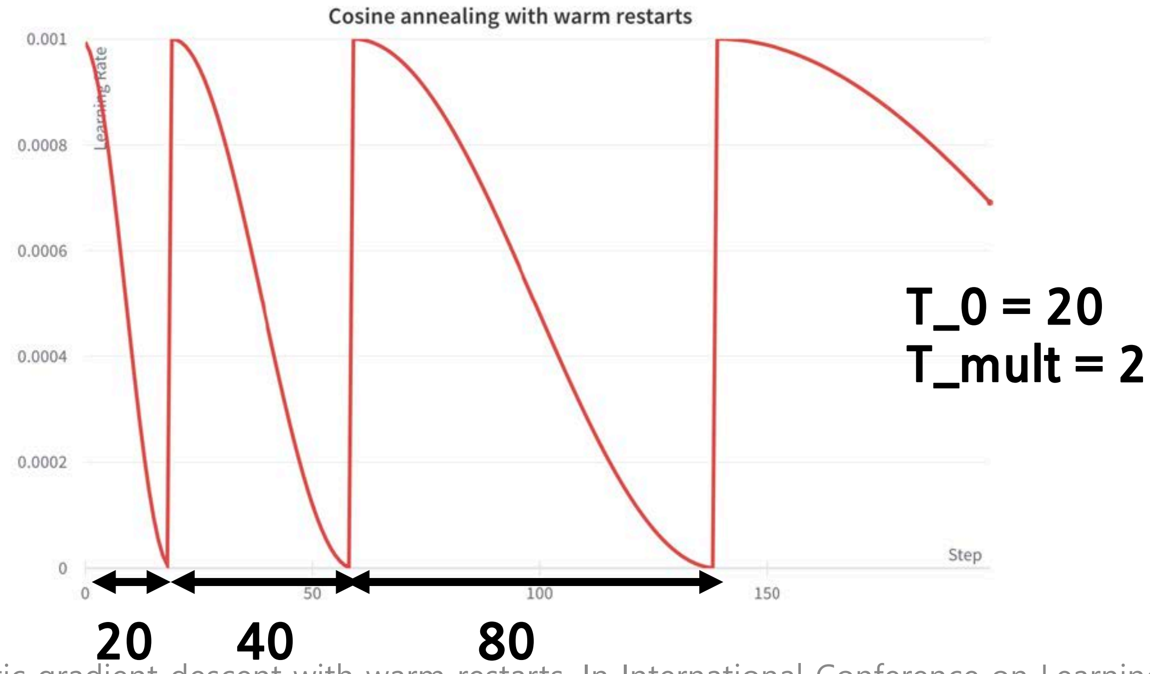
After every epoch, new_lr = current_lr * *gamma*



1: Effect of hyperparameters in each lr schedulers

d) Cosine annealing with warm restarts [Loshchilov et al., 2017][1]

T_0: # of epochs of initial interval
T_i : # of epochs of ith interval (= T_{i-1} * *T_mult*)

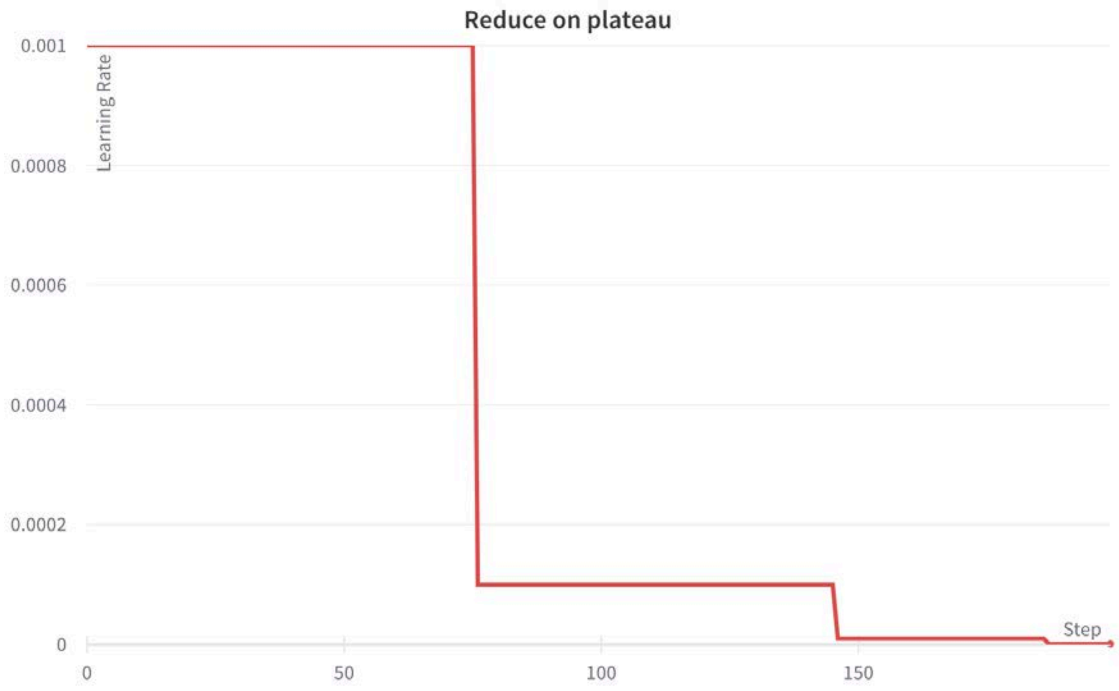


1) I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In International Conference on Learning Representations, 2017.

1: Effect of hyperparameters in each lr schedulers

e) Reduce on plateau

new_lr = current_lr * *factor* if training accuracy does not improve for *patience* epochs



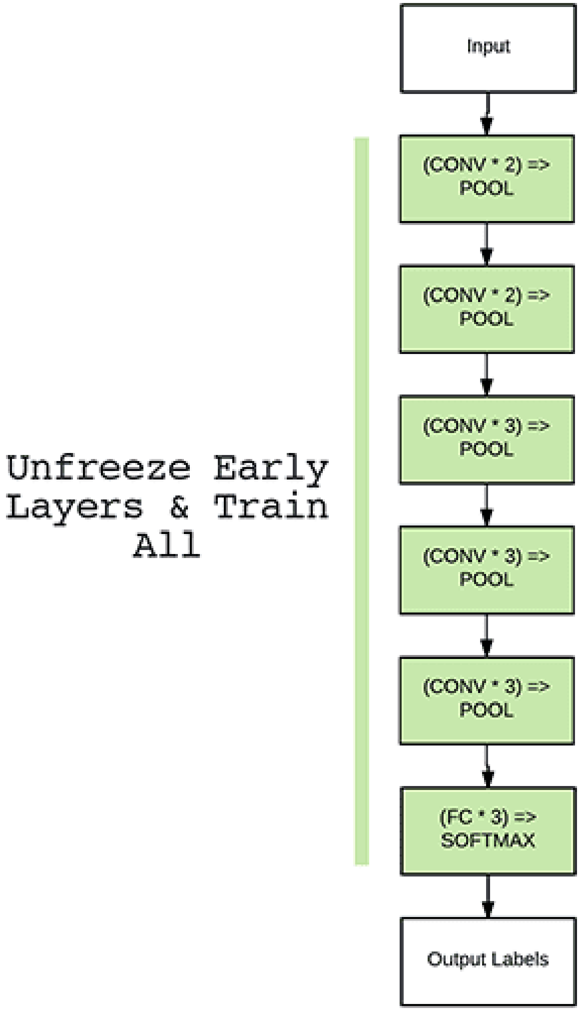
2: Effect between different learning rate schedulers

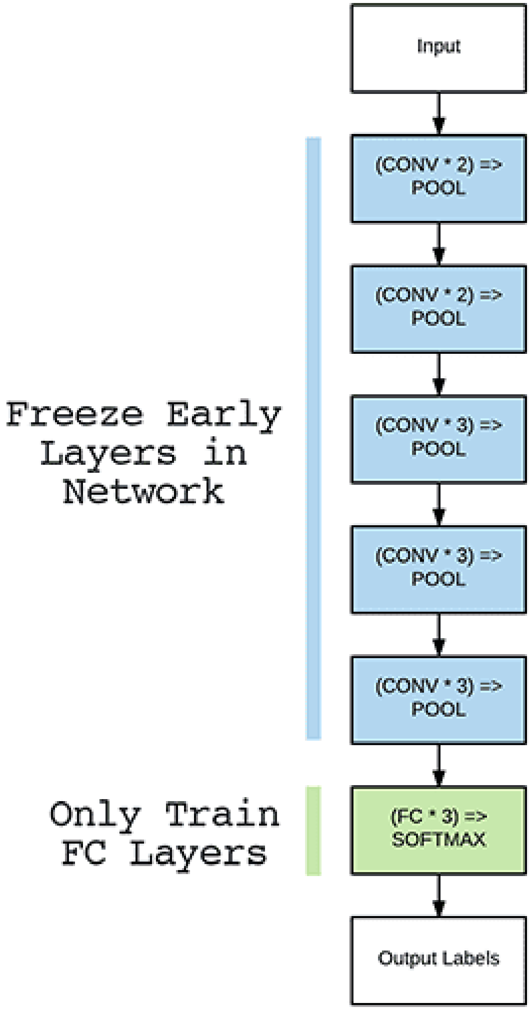
1. Constant learning rate
2. Step learning rate decay
3. Exponential learning rate decay
4. Cosine annealing with warm restarts
5. Reduce on plateau

With initial learning rate = 0.001

02 Experimental Setup

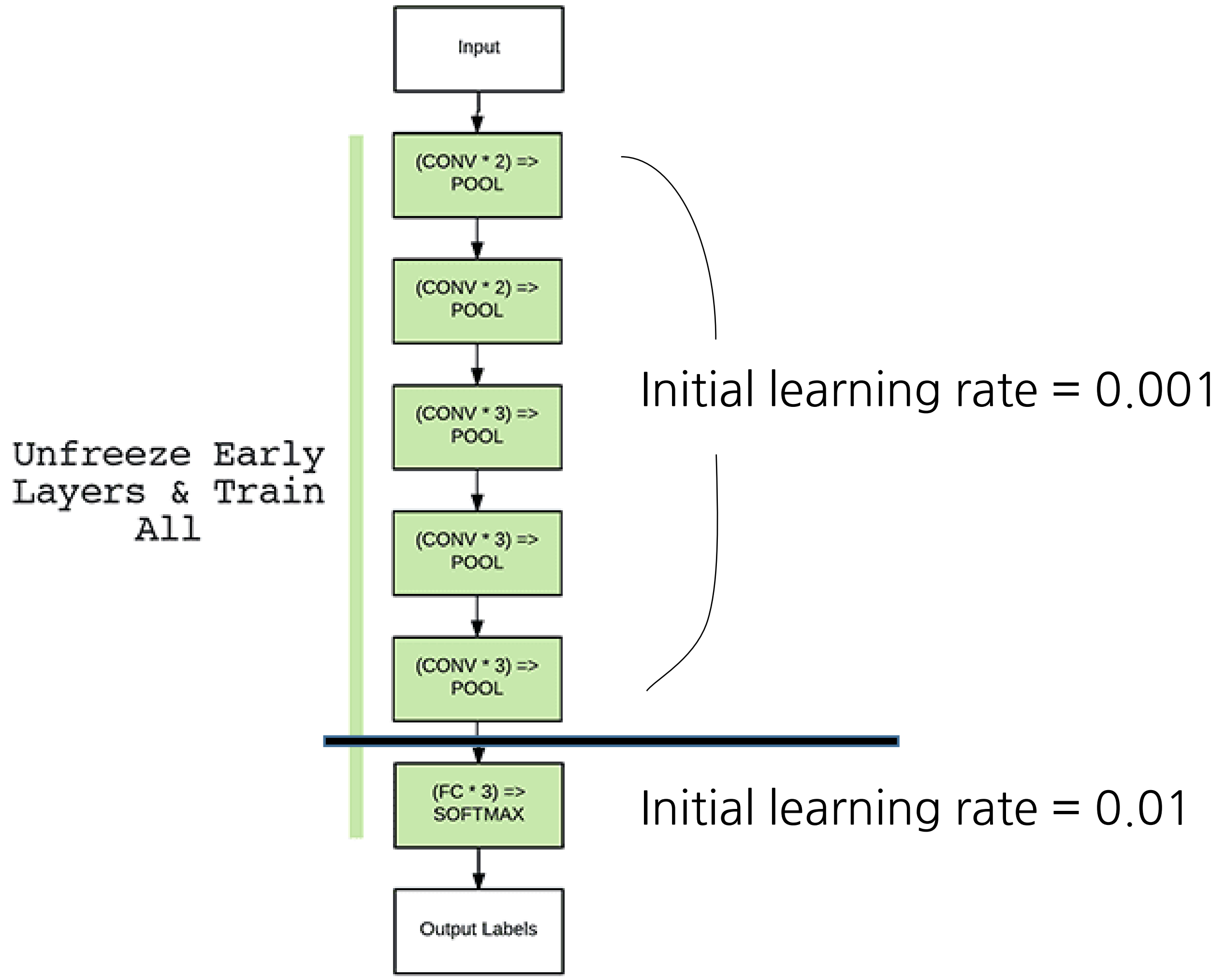
3. Finetuning the ConvNet vs Using ConvNet as fixed feature extractor





02 Experimental Setup

4. Using different initial learning rates between ConvNets and FC layers



02 Experimental Setup

- Backbone network: ResNet18 pretrained on ImageNet
- Target dataset: Stanford CARS196
- Training Epochs: 200
- Batch size: 32
- Optimizer: SGD
- Momentum: 0.9
- Weight decay: 0.01
- Random seed: fixed

02 Experimental Setup

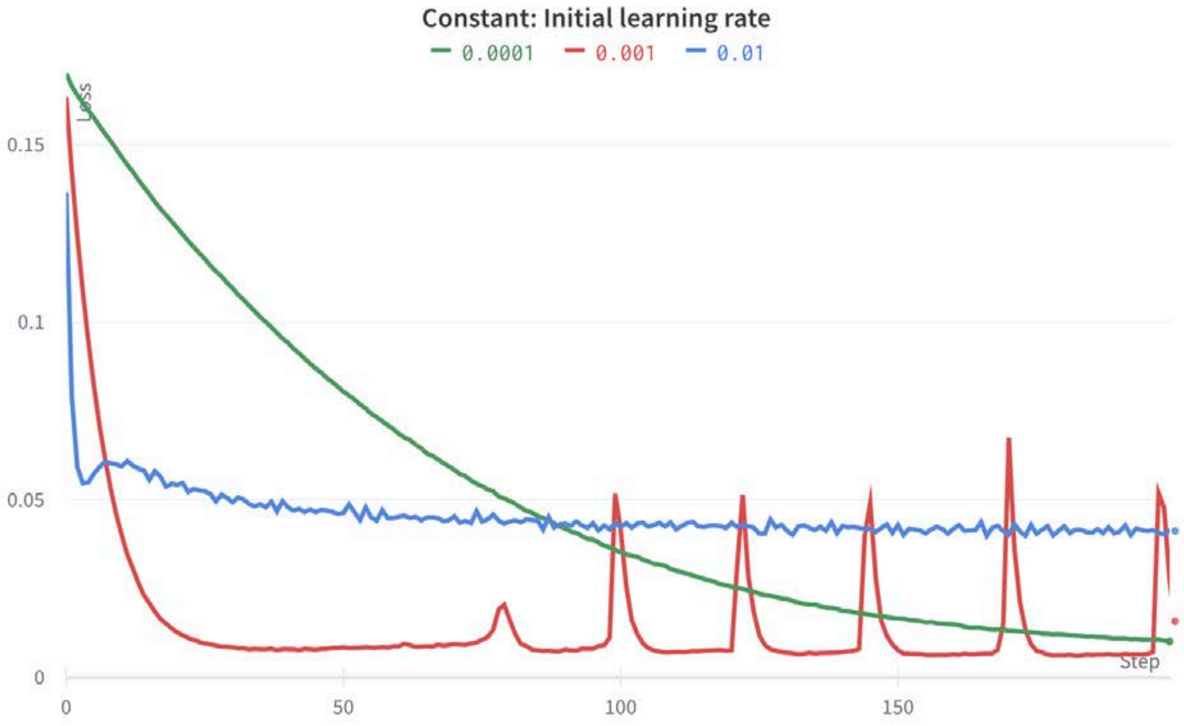
CARS196 dataset

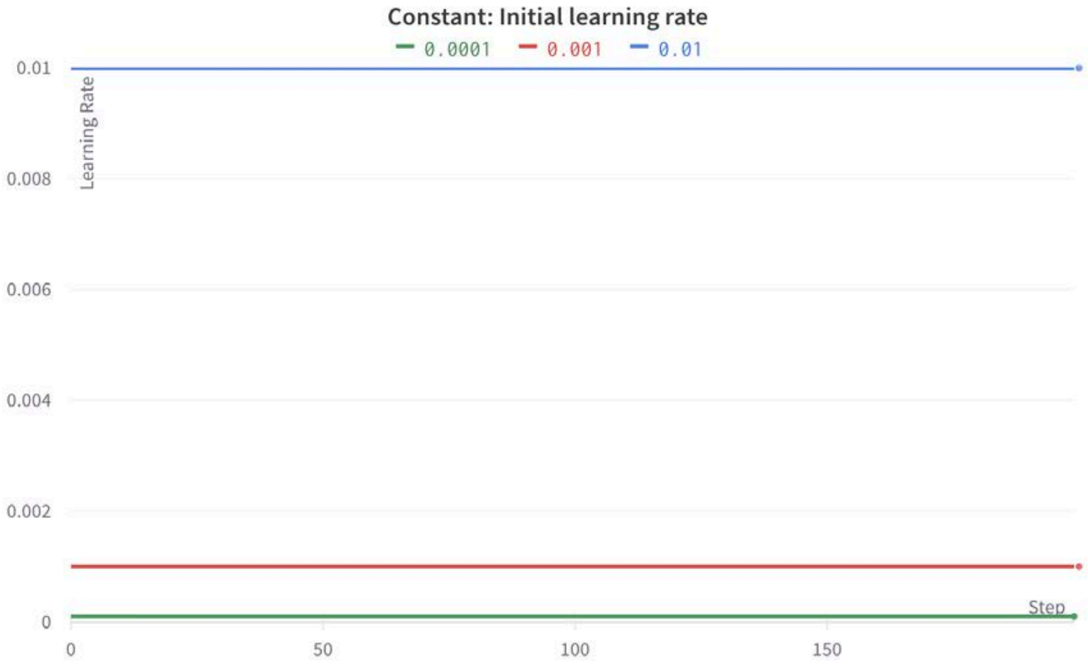
- 16,185 images of 196 classes of cars
- 8,144 training images and 8,041 testing images
- Classes are typically at the level of *Make, Model, Year*, e.g. 2012 Tesla Model S or 2012 BMW M3 coupe

03 Results

1: Effect of hyperparameters in each lr schedulers

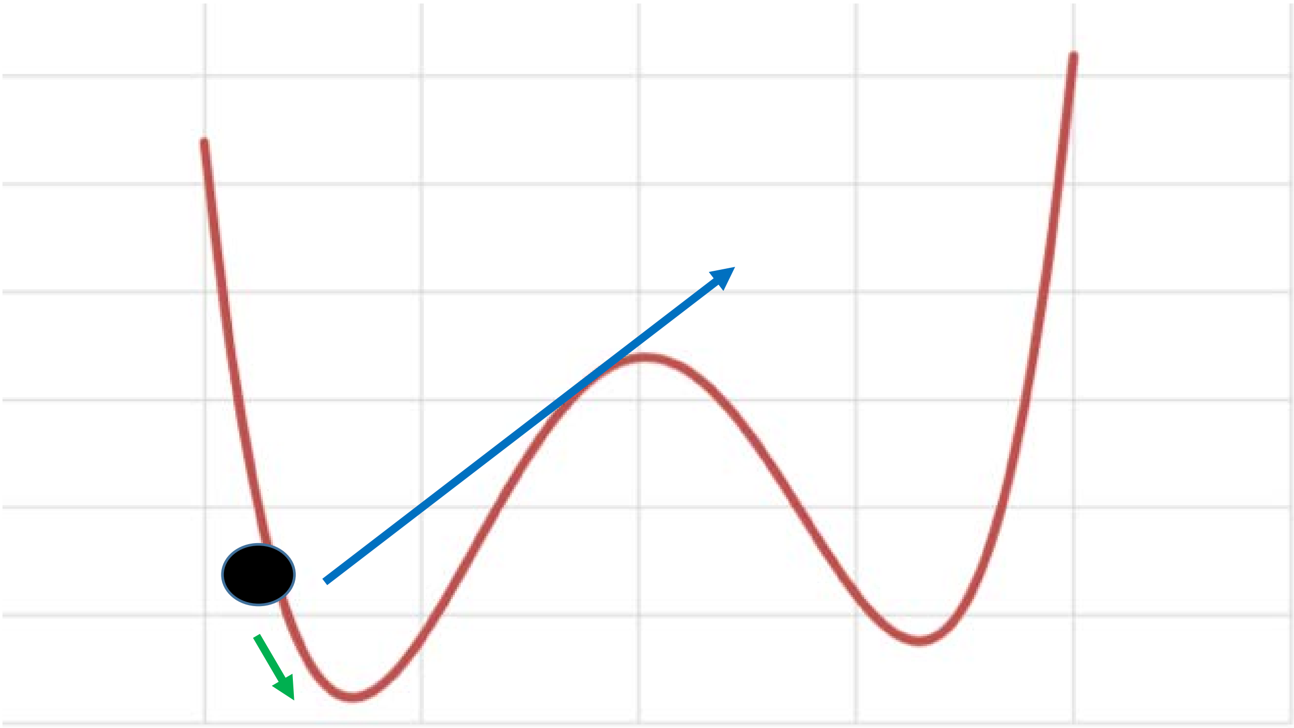
a) Constant learning rate





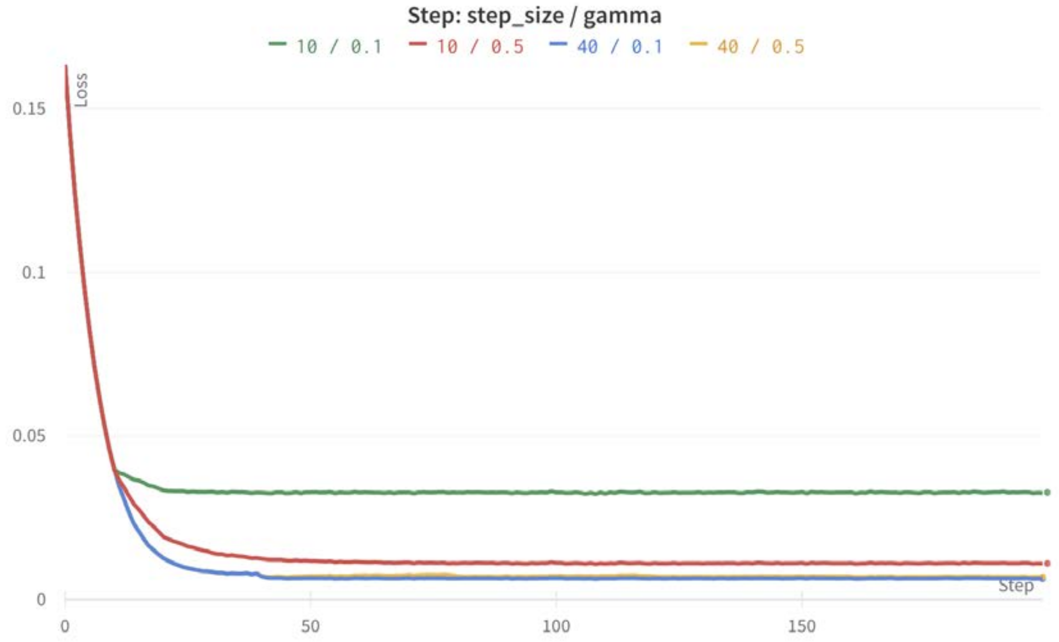
1: Effect of hyperparameters in each lr schedulers

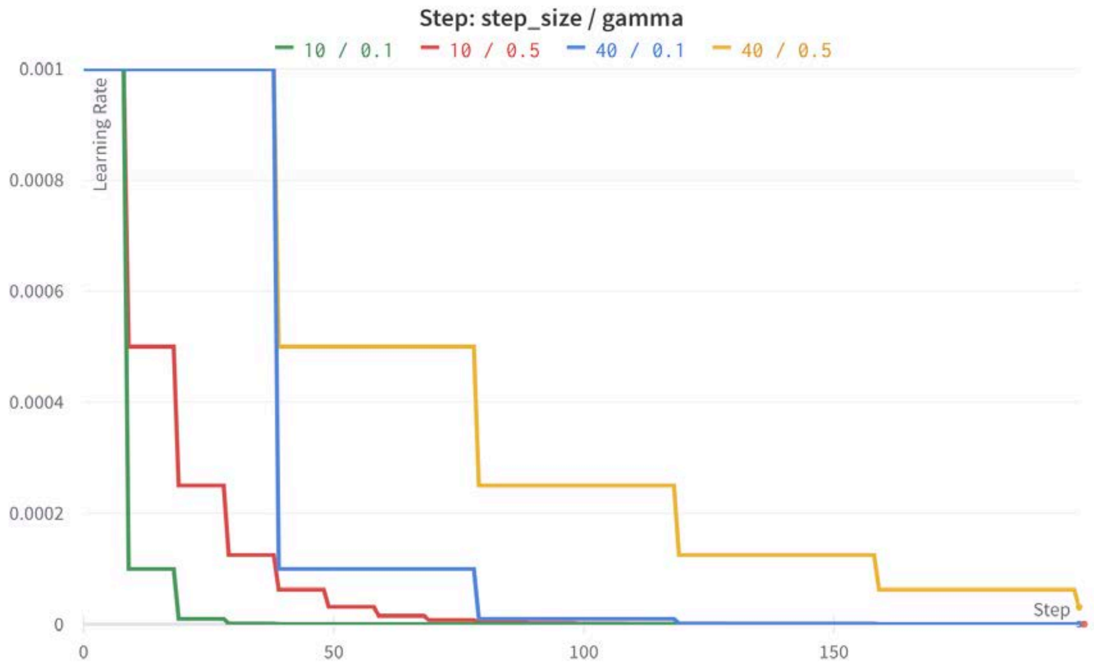
a) Constant learning rate



1: Effect of hyperparameters in each lr schedulers

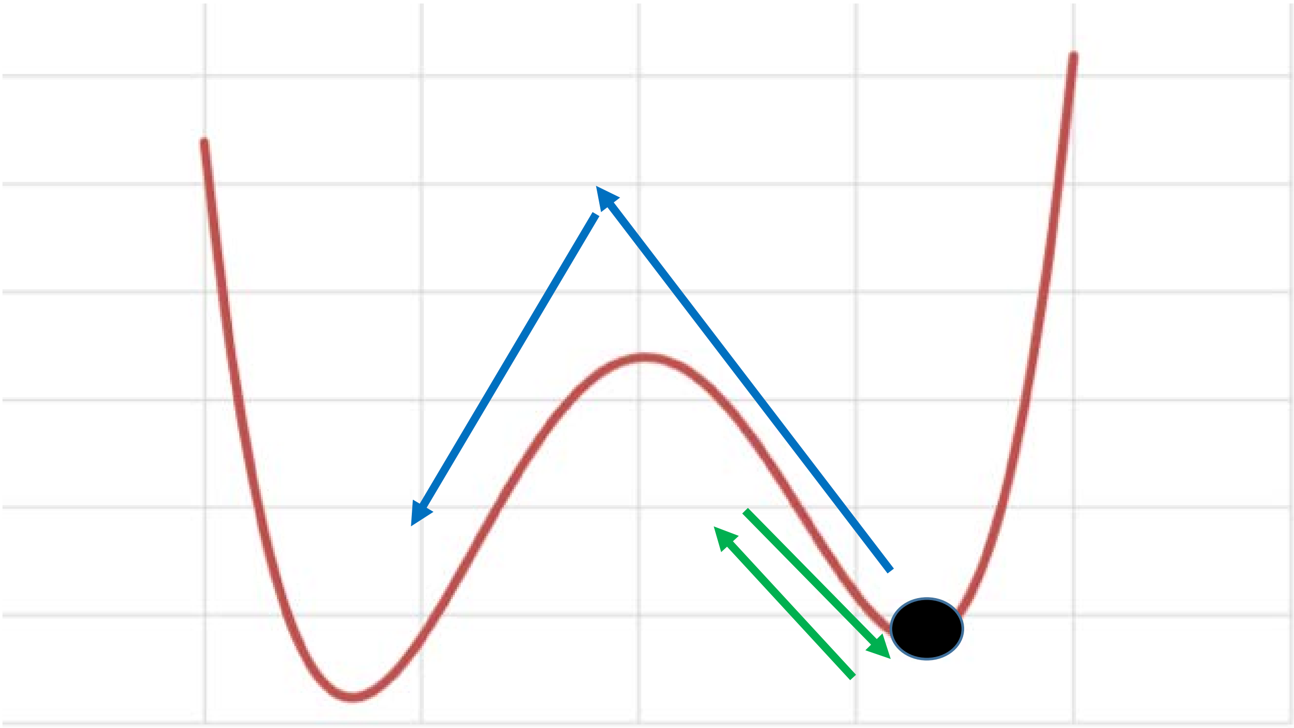
b) Step learning rate decay





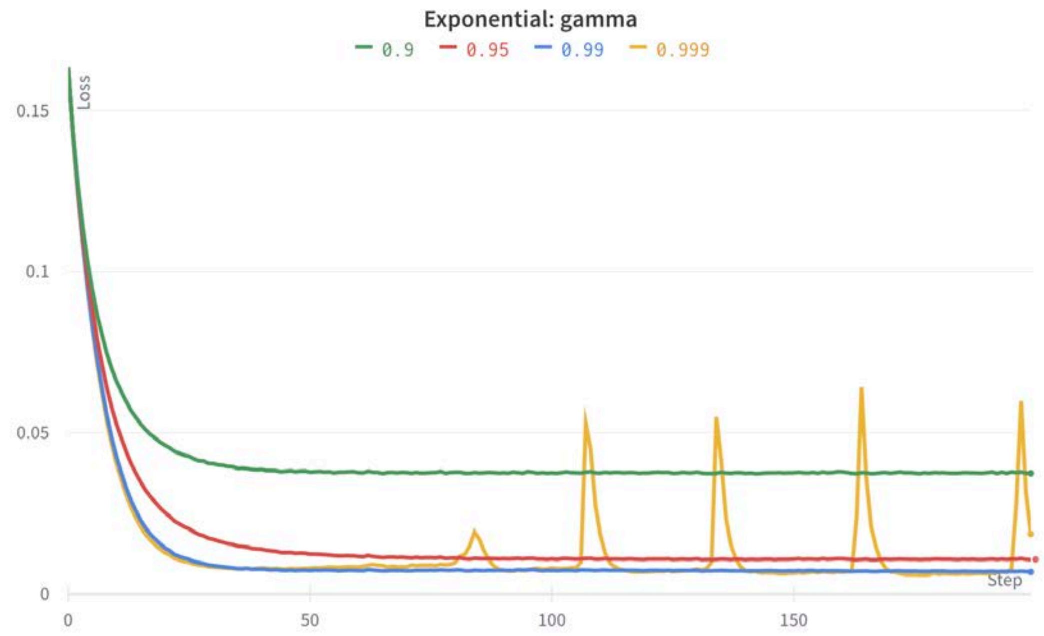
1: Effect of hyperparameters in each lr schedulers

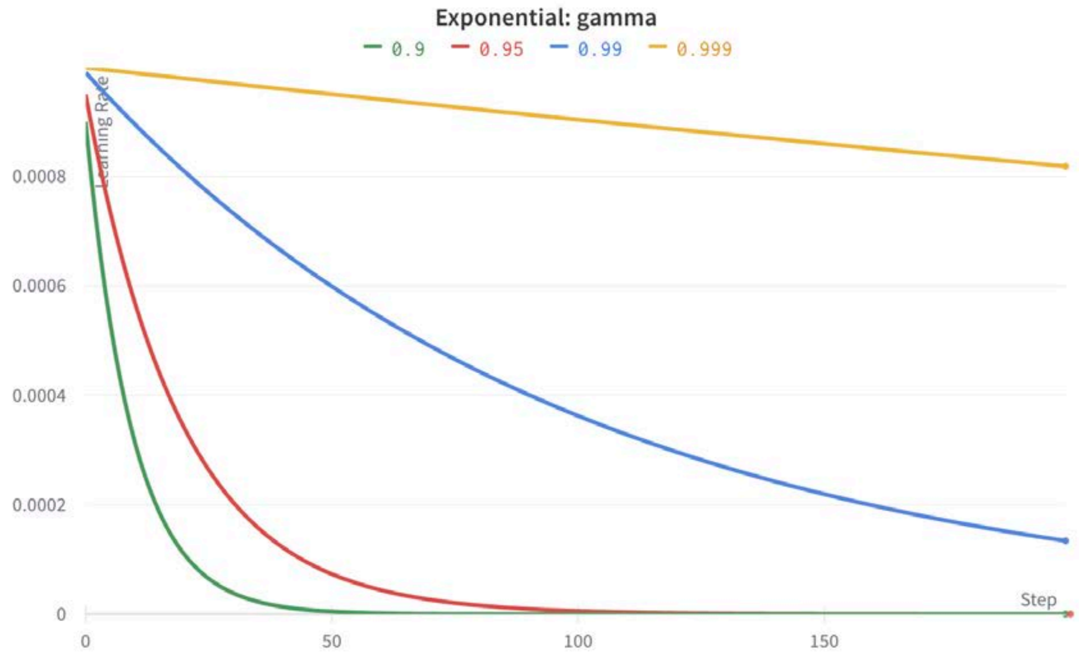
b) Step learning rate decay



1: Effect of hyperparameters in each lr schedulers

c) Exponential learning rate decay

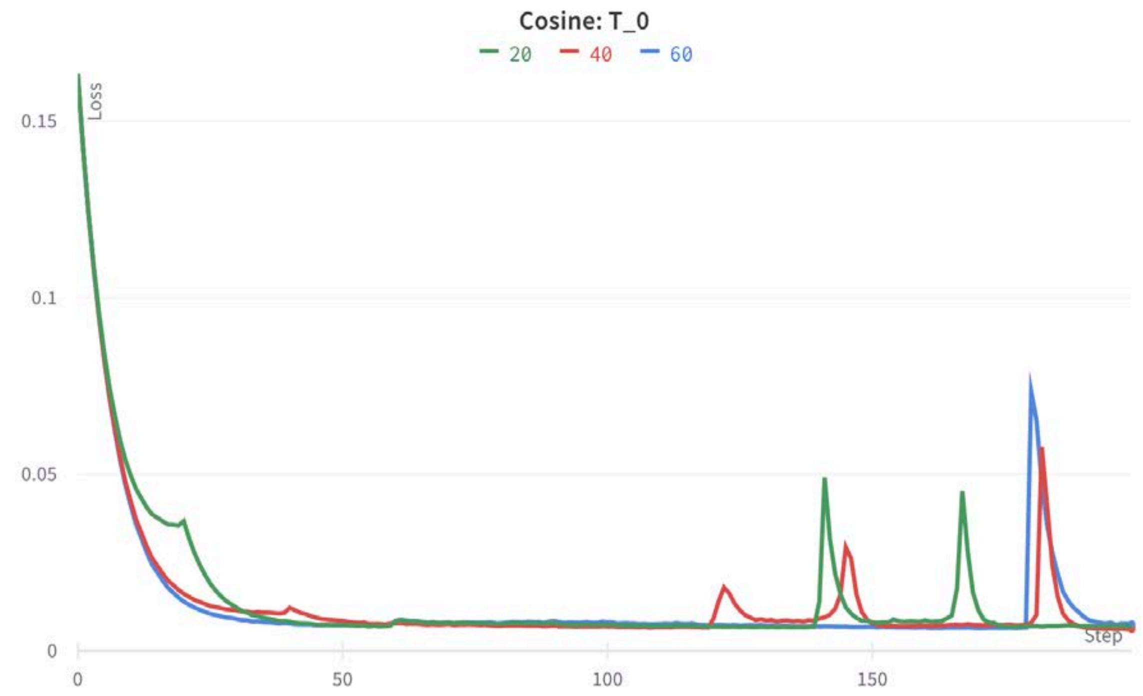


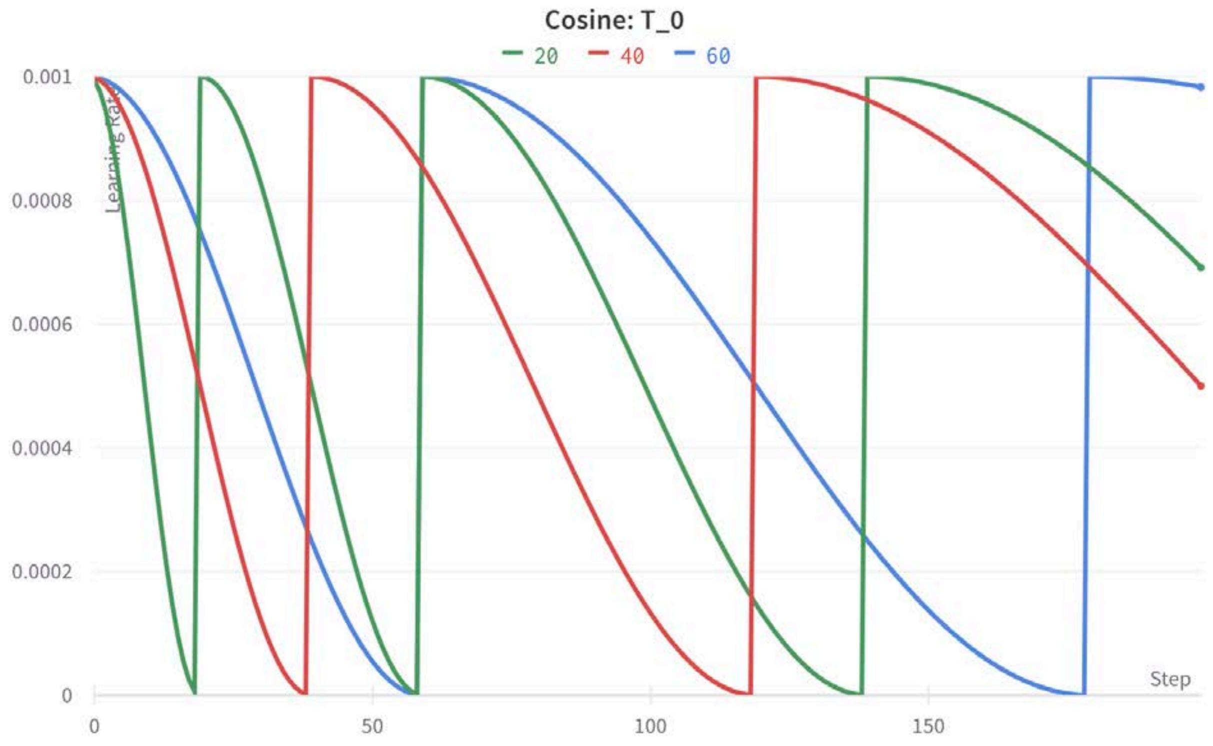


03 Results

1: Effect of hyperparameters in each lr schedulers

d) Cosine annealing with warm restarts

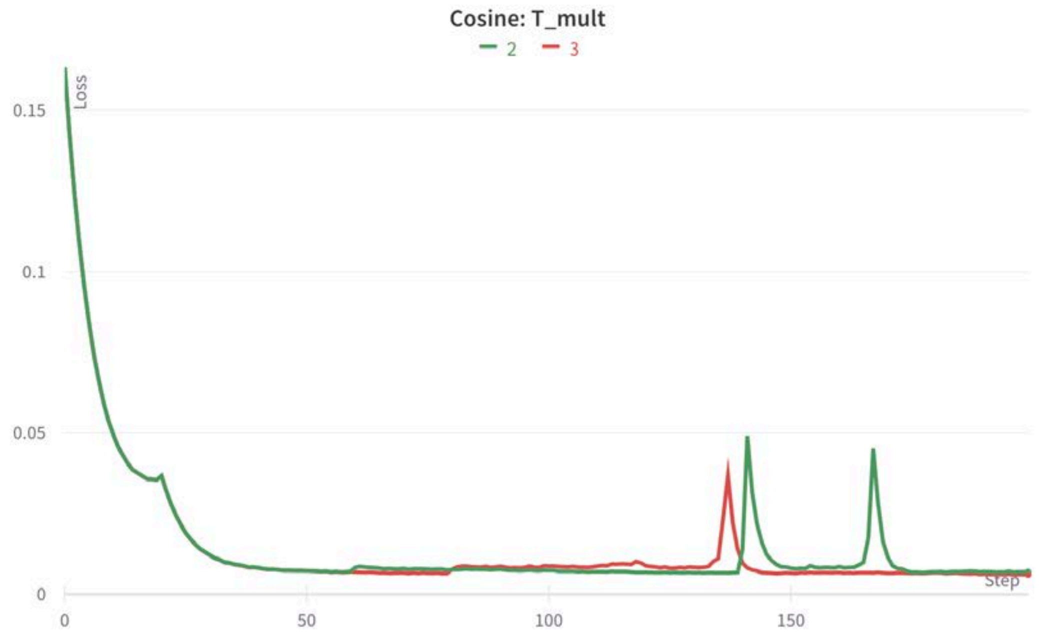


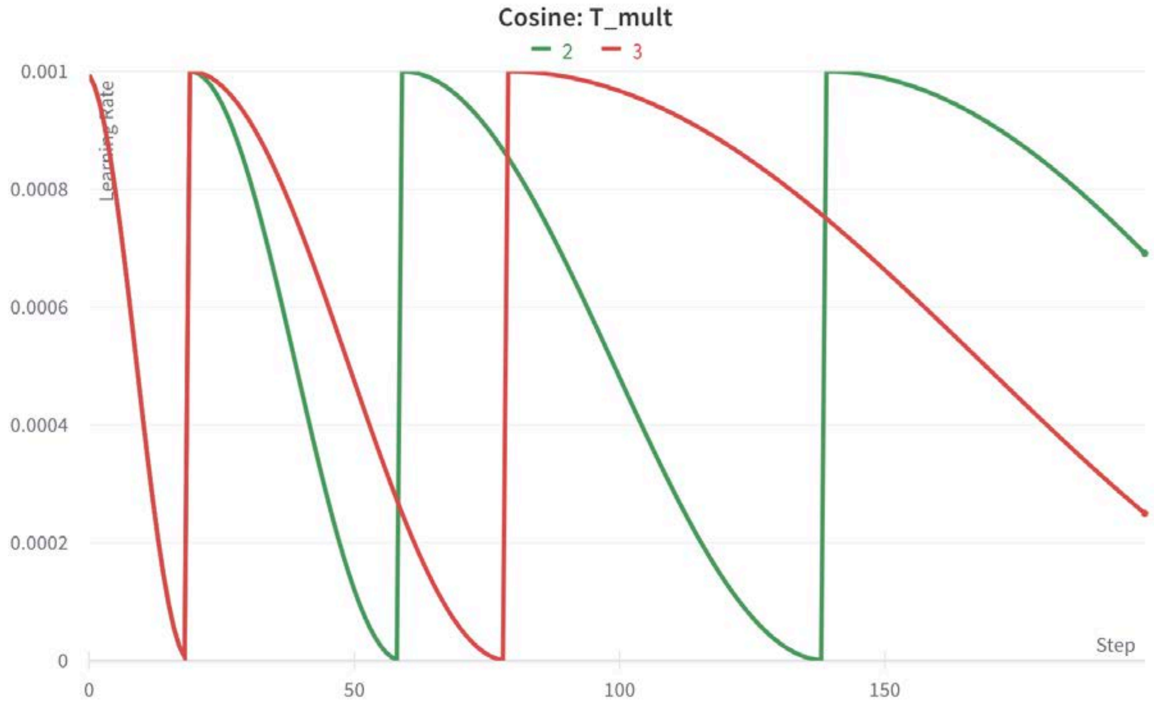


03 Results

1: Effect of hyperparameters in each lr schedulers

d) Cosine annealing with warm restarts

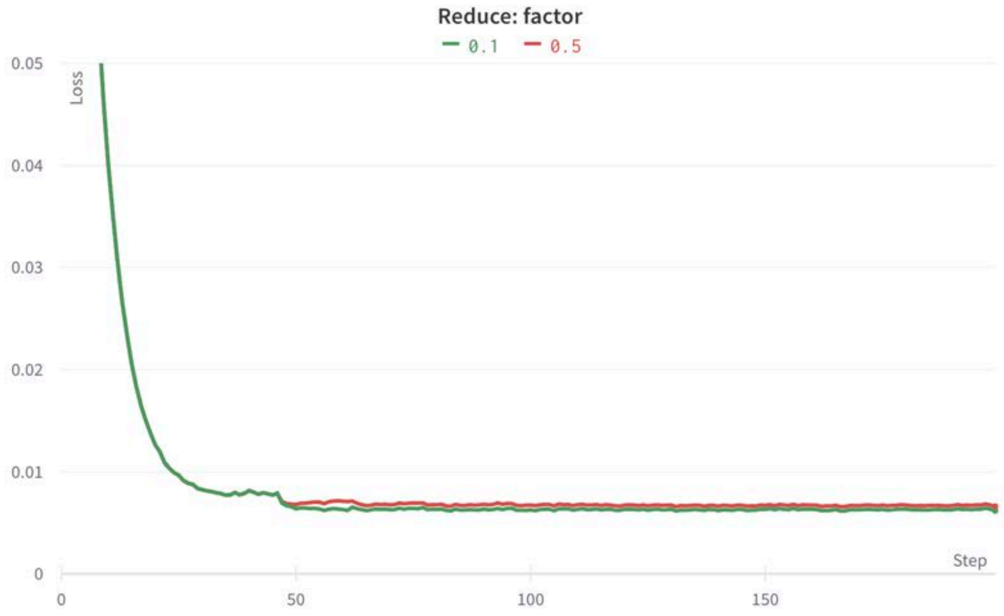


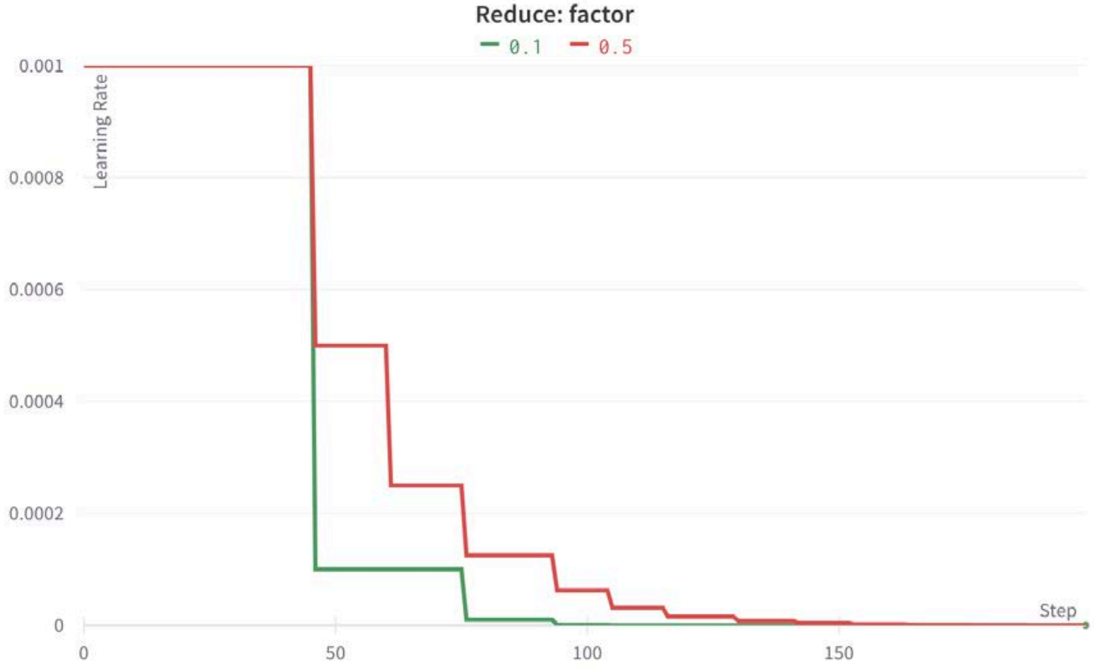


03 Results

1: Effect of hyperparameters in each lr schedulers

e) Reduce on plateau

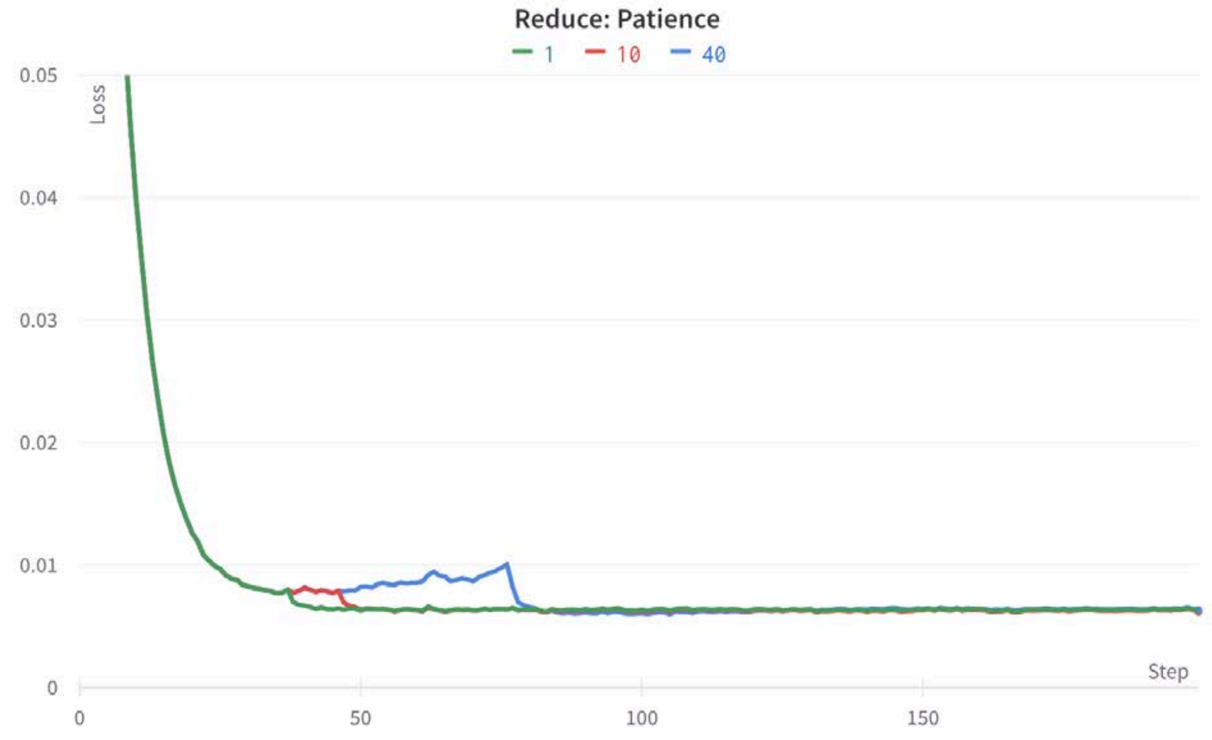


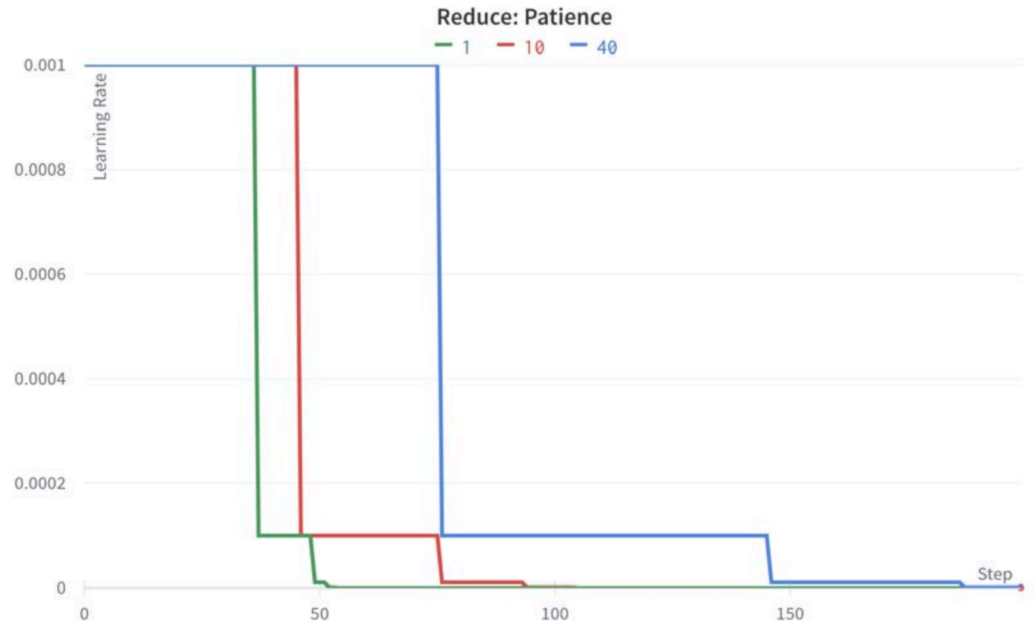


03 Results

1: Effect of hyperparameters in each lr schedulers

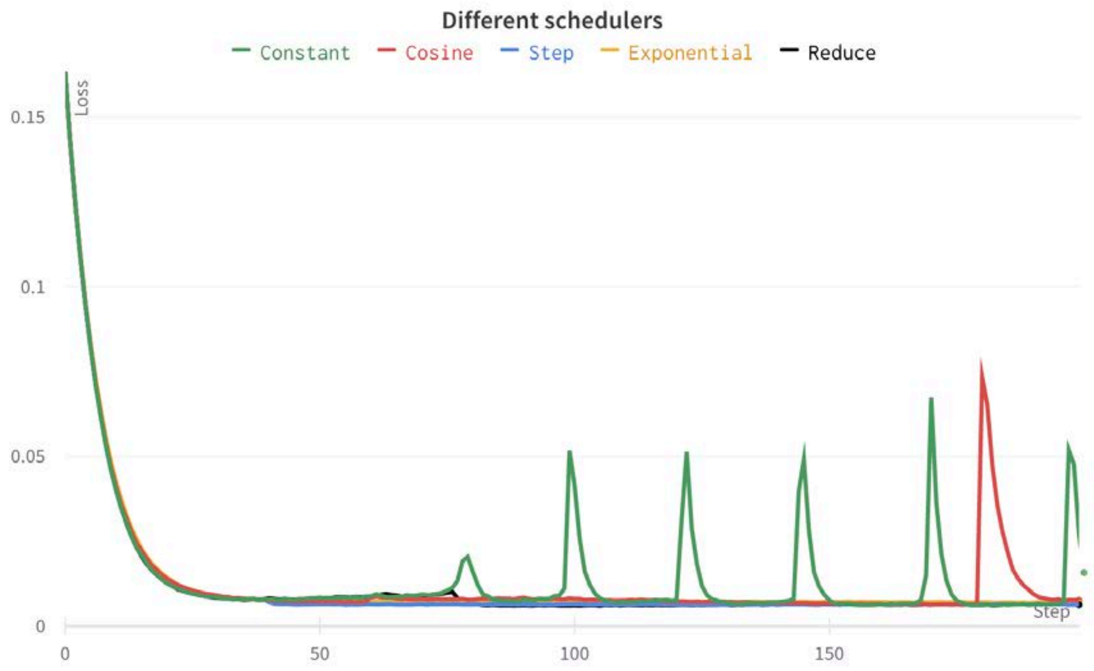
e) Reduce on plateau

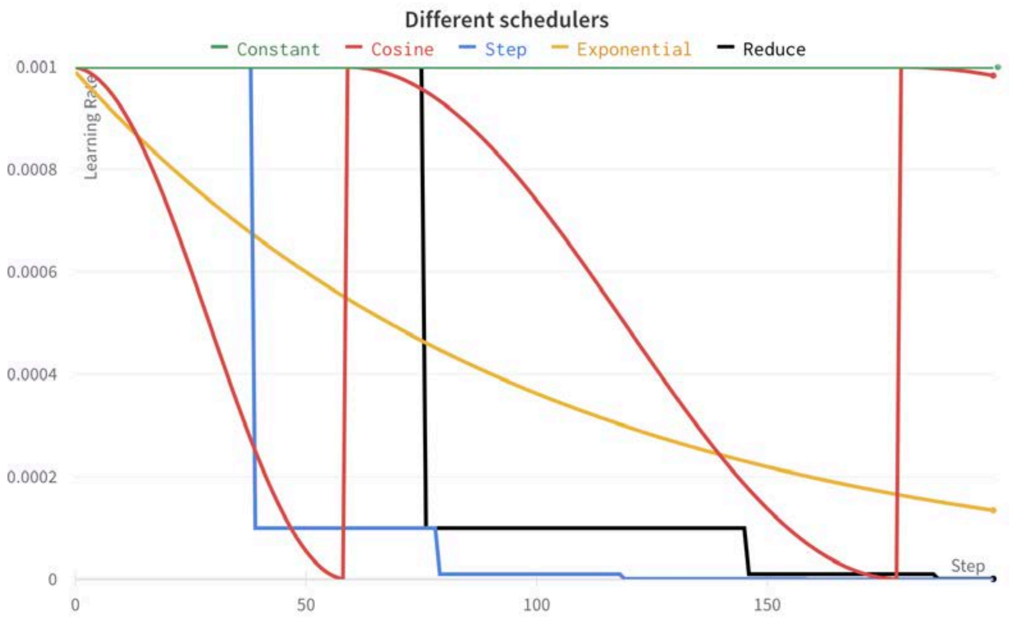




03 Results

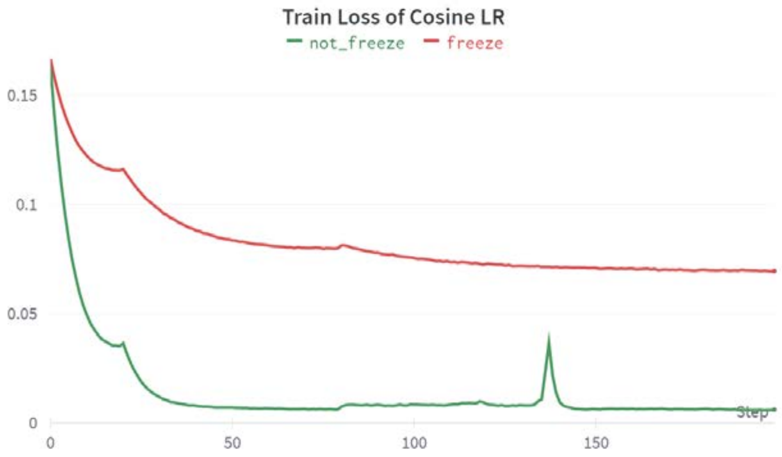
2: Effect between different learning rate schedulers





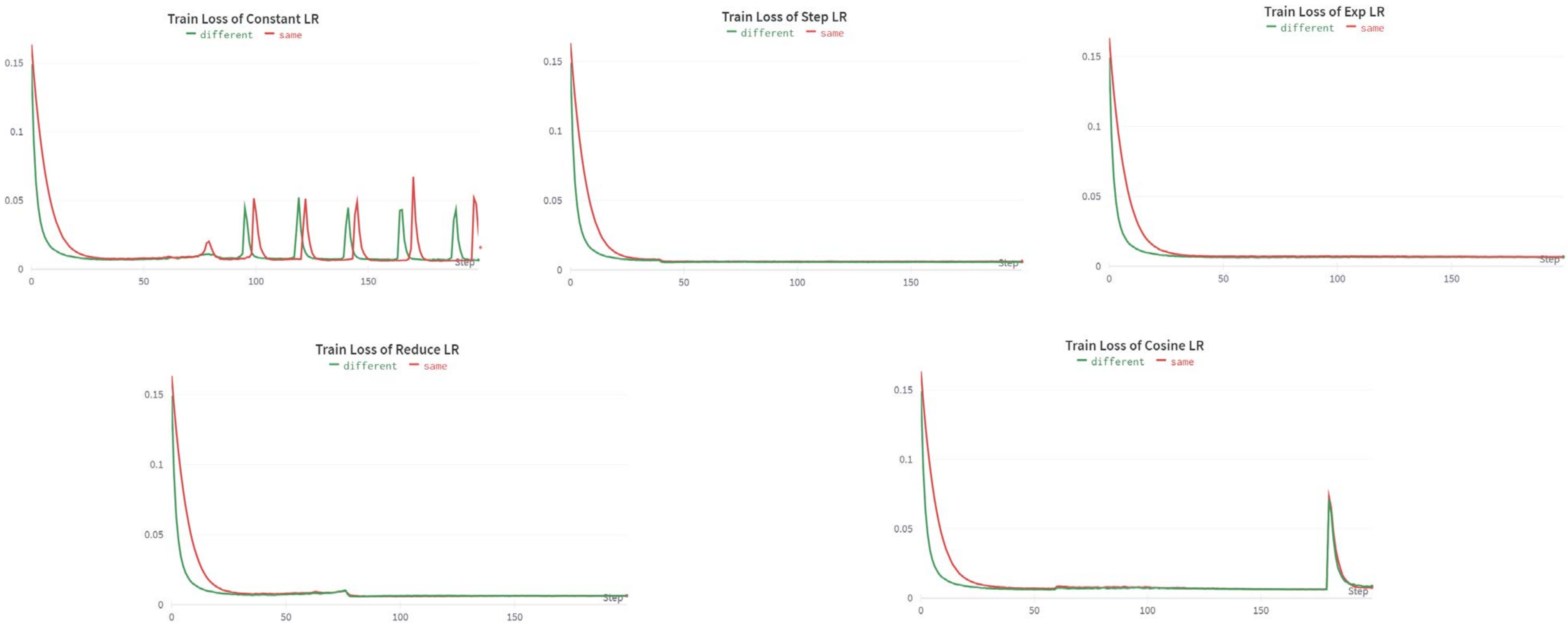
03 Result

3. Finetuning the ConvNet vs Using ConvNet as fixed feature extractor



03 Result

4. Using different initial learning rates between ConvNets and FC layers



04 Conclusion

- Compared the effect of 5 different learning rate schedulers and hyperparameter settings in each of them.
  - If learning rate does not go to 0, it oscillates near the convergent point.
  - If learning rate decays too fast, it gets stuck at a local minimum with high loss.
  - All the schedulers show similar convergence behavior.
- Investigated about the effect of freezing the part of the network(ConvNets) and discovered that ConvNets also need to be updated to extract more appropriate features for target dataset.
- Experimented with different learning rates between ConvNets and FC layers and confirmed that giving larger learning rate to FC layers converges faster.

Thanks

Experiment detail

Hyperparameters used in schedulers in experiment 2, 3, 4

- Step learning rate decay: step_size = 40, gamma = 0.1 in all experiments
- Exponential learning rate decay: gamma = 0.99 in all experiments
- Cosine annealing with warm restart
    - T_0 = 60, T_mult = 2 in (Experiment 2, 4)
    - T_0 = 20, T_mult = 3 in (Experiment 3)
- Reduce on plateau: patience = 40, factor = 0.1 in all experiments