Automated Learning Rate Scheduler for Large-Batch Training

kakaobrain

Chiheon Kim[1], Saehoon Kim[1], Jongmin Kim[1], Donghoon Lee[1], Sungwoong Kim[1]

[1]Kakao Brain, Seongnam, Korea

https://github.com/kakaobrain/autowu

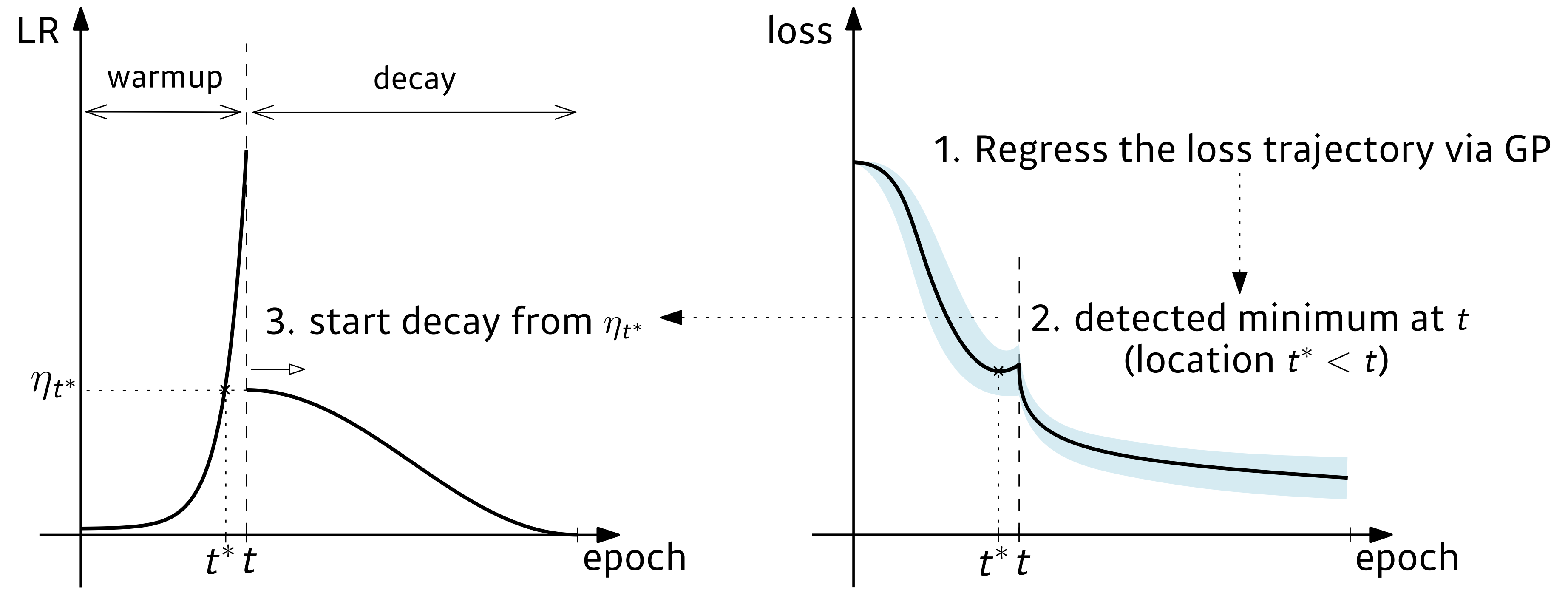
Summary

- Choosing a proper learning rate (LR) and its schedule is essential in large-batch training.
- LR scaling rule and gradual LR warmup has been shown to be successful in large-batch training, if a good LR for the small-batch counterpart is known.
- We design an **automated LR scheduler (AutoWU)** which takes care of (1) LR tuning and (2) gradual warmup simultaneously.
- The proposed scheduler works well for wide-range of batch sizes, with minimal hyper-parameter tuning effort.

Method

AutoWU: (1) Warmup + (2) Decay phases

- Warmup: Exponential schedule from a very small value (1e-5)
- Decay: Cosine or Constant-then-cosine (cosine decay in the last 20% epochs)



1. Regress the loss trajectory via GP

2. detected minimum at t (location $t^* < t$)

3. start decay from η_{t^*}

Automatic Phase Transition:

- GP-based online detection of the minimum loss: $\max_{s\in[0,t]} \mathbb{P}_f\big(f(s) < f(t)\big) > 0.95$

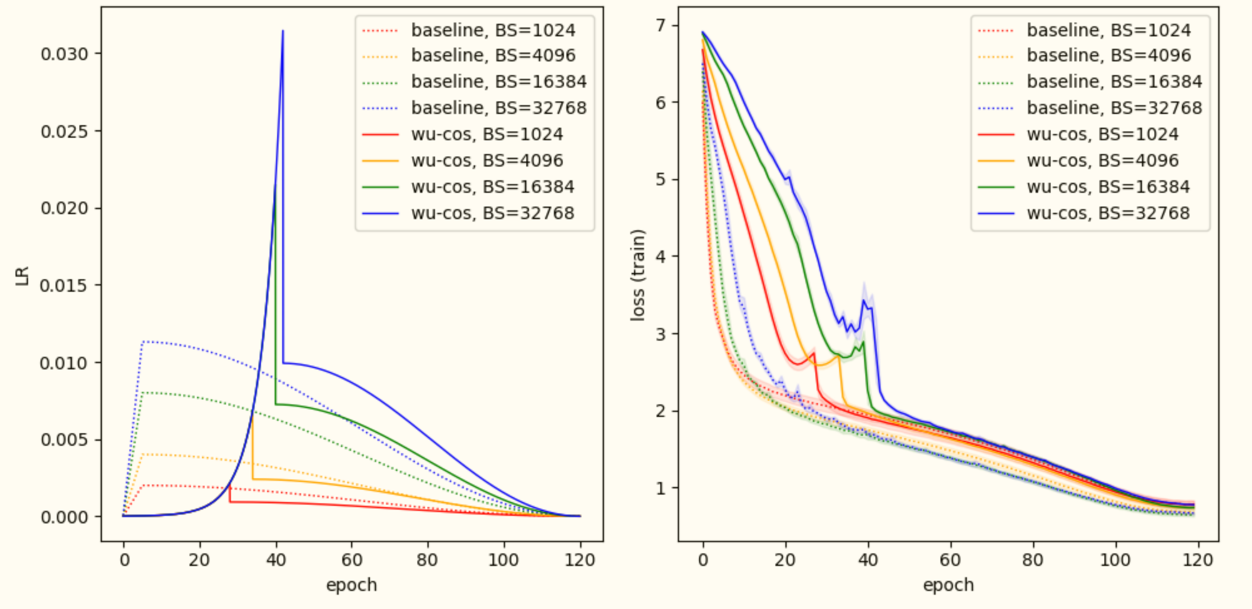
Experiments

- Baseline vs. AutoWU on image classification benchmarks
  (baseline = good-working small-batch tuning + square-root scaling law + gradual warmup)

Dataset (Architecture)	Schedule	Batch size			
		256	1K	8K	16K
CIFAR-10 (ResNet-18)	Baseline	**96.58 (0.07)**	**96.48 (0.02)**	**96.05 (0.15)**	94.63 (0.06)
	AutoWU + const-cos	96.26 (0.12)	96.20 (0.03)	95.92 (0.22)	**94.80 (0.17)**
	AutoWU + cos	96.43 (0.02)	96.42 (0.05)	95.77 (0.01)	94.03 (0.26)
CIFAR-100 (Wide-ResNet28-10)	Baseline	83.36 (0.38)	83.13 (0.14)	81.08 (0.33)	77.62 (0.36)
	AutoWU + const-cos	83.36 (0.21)	83.21 (0.19)	**82.32 (0.42)**	**81.42 (0.35)**
	AutoWU + cos	**83.59 (0.46)**	**83.39 (0.20)**	82.26 (0.60)	80.25 (0.36)
		1K	4K	16K	32K
ImageNet (ResNet-50)	Baseline	76.28	76.10	75.02	74.11
	AutoWU + const-cos	**76.31**	**76.33**	**75.62**	**74.84**
	AutoWU + cos	76.19	75.70	75.22	74.40

Comparison of validation accuracy (%):

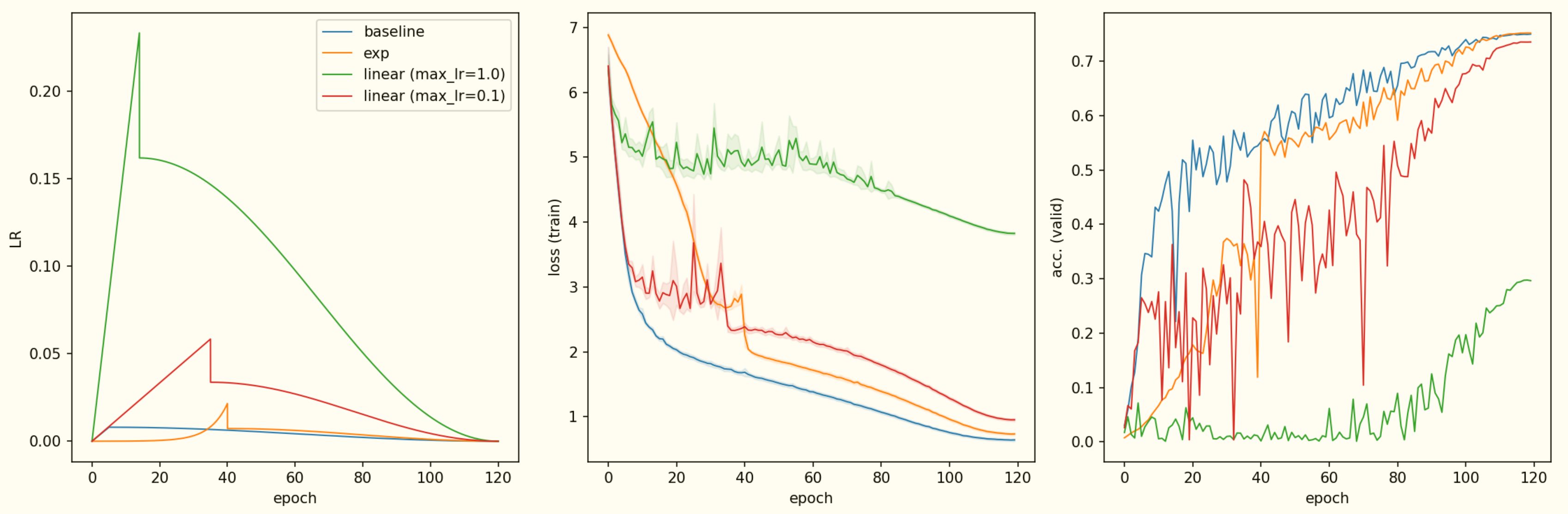
AutoWU achieves a comparable performance to the baseline for all batch sizes.



LR and loss curve plots (ImageNet - ResNet-50):

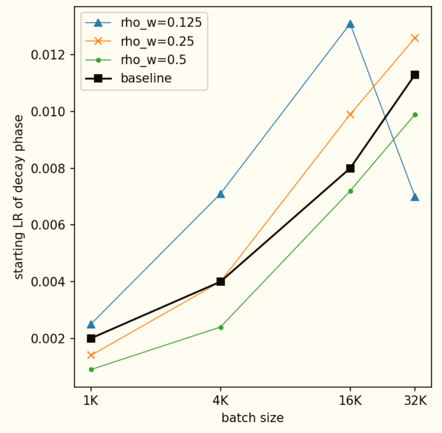
Interestingly, the automatically found starting LR of the decay phase of AutoWU is similar to the peak LR of the baseline.

- Ablation on hyperparameters regarding the warmup schedule



LR, loss, validation accuracy plots (baseline, linear, and exponential):

Linear warmup schedule is sensitive to the choice of the growth rate (or equivalently, maximum LR) and shows unstable training dynamics, compared to the exponential schedule.



ρ_w	Batch size			
	1K	4K	16K	32K
0.125	75.89	75.59	74.52	73.40
0.25	**76.60**	**76.04**	**75.28**	73.89
0.5	76.19	75.70	75.22	**74.40**

Starting LR of decay phase (left) and valid. acc. (%, right) w.r.t. ρ_w (maximum warmup duration):

Performance of AutoWU does not vary much w.r.t. ρ_w (which determines the growth factor), and the relation between the found LR and batch size is similar to that of square-root scaling law.