简介

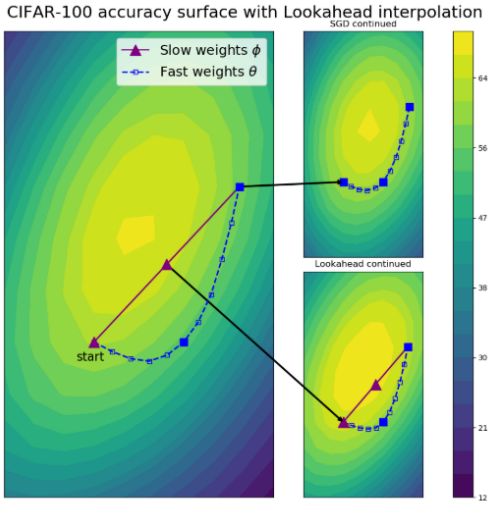
Hinton的又一力作，目标是提出一个更好用收敛更快的优化器。
目前最流行的SGD的改进方向有两种：

1. 自适应的学习率(AdaGrad, Adam)
2. 加速方法(Polyak heavyball和Nesterob momentum)

两种方法都强调利用梯度的历史信息找到更好的优化方向，但是问题在于这些方法都需要调参才能得到比较好的效果。

方法论

lookahead方法存在两组参数，一组slow wrights ϕ和一组fast weights θ。
首先在k个step里面更新fast weight。k个step之后，更新一次slow weight，slow weight的来源是fast weight和之前slow weight域($\theta-\phi$)的线性插值，每更新一次slow weight都会将当前fast weight置为slow weight。



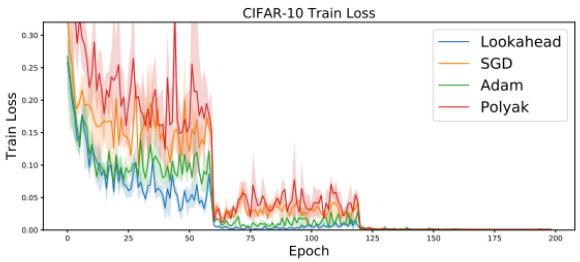
Algorithm 1 Lookahead Optimizer:

Require: Initial parameters ϕ_0, objective function L
Require: Synchronization period k, slow weights step size α, optimizer A
for $t = 1, 2, \ldots$ **do**
  Synchronize parameters $\theta_{t,0} \leftarrow \phi_{t-1}$
  for $i = 1, 2, \ldots, k$ **do**
    sample minibatch of data $d \sim \mathcal{D}$
    $\theta_{t,i} \leftarrow \theta_{t,i-1} + A(L, \theta_{t,i-1}, d)$
  end for
  Perform outer update $\phi_t \leftarrow \phi_{t-1} + \alpha(\theta_{t,k} - \phi_{t-1})$
end for
return parameters ϕ

Figure 1: (Left) Visualizing Lookahead through a ResNet-32 test accuracy surface at epoch 100 on CIFAR-100. We project the weights onto a plane defined by the first, middle, and last fast (inner-loop) weights. The fast weights are along the blue dashed path. All points that lie on the plane are represented as solid, including the entire Lookahead slow weights path (in purple). Lookahead (middle, bottom right) quickly progresses closer to the minima than SGD (middle, top right) is able to. (Right) Pseudocode for Lookahead.

当震荡发生在高曲率方向时，fast weight会快速更新至低曲率方向。
而slow weight可以通过插值的方法使得学习曲线更为平滑。快慢结合的方法可以有效提升在高曲率方向的学习，同时减小方差并加快收敛速度。

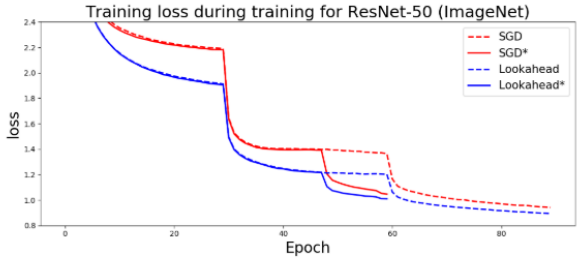
实验结果



OPTIMIZER	CIFAR-10	CIFAR-100
SGD	$95.23 \pm .19$	$78.24 \pm .18$
POLYAK	$95.26 \pm .04$	$77.99 \pm .42$
ADAM	$94.84 \pm .16$	$76.88 \pm .39$
LOOKAHEAD	$95.27 \pm .06$	$78.34 \pm .05$

Table 1: CIFAR Final Validation Accuracy.

Figure 5: Performance comparison of the different optimization algorithms. (**Left**) Train Loss on CIFAR-100. (**Right**) CIFAR ResNet-18 validation accuracies with various optimizers. We do a grid search over learning rate and weight decay on the other optimizers (details in appendix C). Lookahead and Polyak are wrapped around SGD.



OPTIMIZER	LA	SGD
EPOCH 50 - TOP 1	75.13	74.43
EPOCH 50 - TOP 5	92.22	92.15
EPOCH 60 - TOP 1	75.49	75.15
EPOCH 60 - TOP 5	92.53	92.56

Table 2: Top-1 and Top-5 single crop validation accuracies on ImageNet.

Figure 6: ImageNet training loss. The asterisk denotes the aggressive learning rate decay schedule, where LR is decayed at iteration 30, 48, and 58. We report validation accuracies for this schedule.