Model ensembling-Snapshot and SWA

Presenter：Wentao Lei

Paper List	Time	Publish
Averaging Weights Leads to Wider Optima and Better Generalization	2018	NIPS
SNAPSHOT ENSEMBLES: TRAIN 1, GET M FOR FREE	2017	ICLR

Content

- Background and motivation
- Method
- Experiment

集成学习（Ensemble learning）

- 在经典机器学习中，**集成学习（Ensemble learning）**是非常重要的思想，很多情况下都能对模型性能产生极大的提升。

- 集成学习算法本身不算一种单独的机器学习算法，而是通过构建并结合多个机器学习器来完成学习任务。集成学习在机器学习算法中拥有较高的准确率，不足之处就是模型的训练过程可能比较复杂，效率不是很高。

集成学习（Ensemble learning）

- 集成学习的思路就是组合若干不同的模型，让它们基于相同的输入做出预测，接着通过某种平均化方法决定集成模型的最终预测。
- 这个决定过程可能是通过简单的投票或取均值，也可能是通过另一个模型，该模型能够基于集成学习中众多模型的预测结果，学习并预测出更加准确的最终结果。
- 集成学习的思想同样适用于深度学习，集成应用于深度学习时，组合若干网络的预测以得到一个最终的预测。通常，可以对性能有一定的提升，但是训练过程中可能会消耗更多的空间和时间成本。

SNAPSHOT ENSEMBLES: TRAIN 1, GET M FOR FREE

Gao Huang*, **Yixuan Li***, **Geoff Pleiss**
Cornell University
{gh349, yl2363}@cornell.edu, geoff@cs.cornell.edu

Zhuang Liu
Tsinghua University
liuzhuangthu@gmail.com

John E. Hopcroft, Kilian Q. Weinberger
Cornell University
jeh@cs.cornell.edu, kqw4@cornell.edu

快照集成（Snapshot Ensembling）

- 由于神经网络训练的耗时, 导致多模型的**Ensemble**在深度学习领域应用不如传统的机器学习方法广泛. 因为用于**Ensemble**的每个基模型, 都是单独训练的, 往往单个模型的训练就比较耗时了, 因此这种提升模型表现的方法成本是相当高的.

- 这篇论文提出了一种方法, 不需要增加额外的训练消耗, 通过一次训练, 得到若干个模型, 并对这些模型进行**Ensemble**, 得到最终的模型.

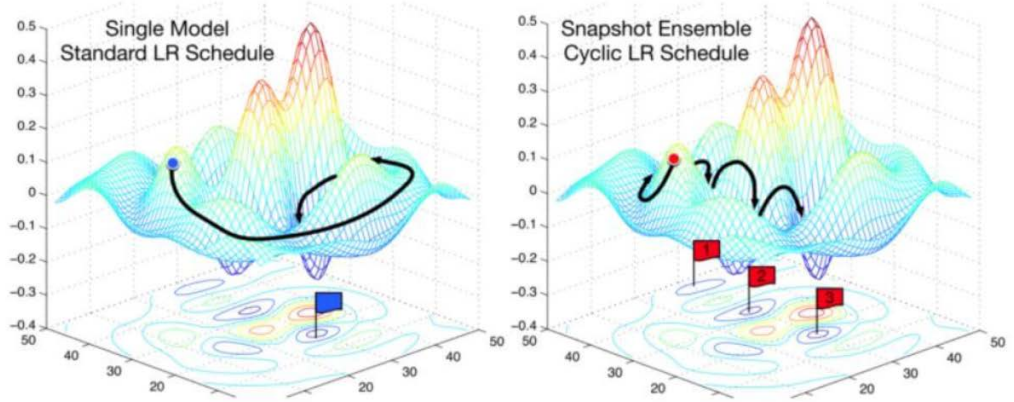
快照集成（Snapshot Ensembling）

对神经网络使用SGD方法进行训练, 在一次训练过程中, 使模型M次收敛于不同的局部极小, 每次收敛, 都代表这一个最终的模型, 我们将此时的模型进行保存. 然后使用一个较大的学习率逃离此时的局部极小.

在论文中, 对学习率的控制使用了一种余弦函数, 这种函数表现为:

- 急剧提升学习率
- 在某次训练过程中, 学习率迅速下降

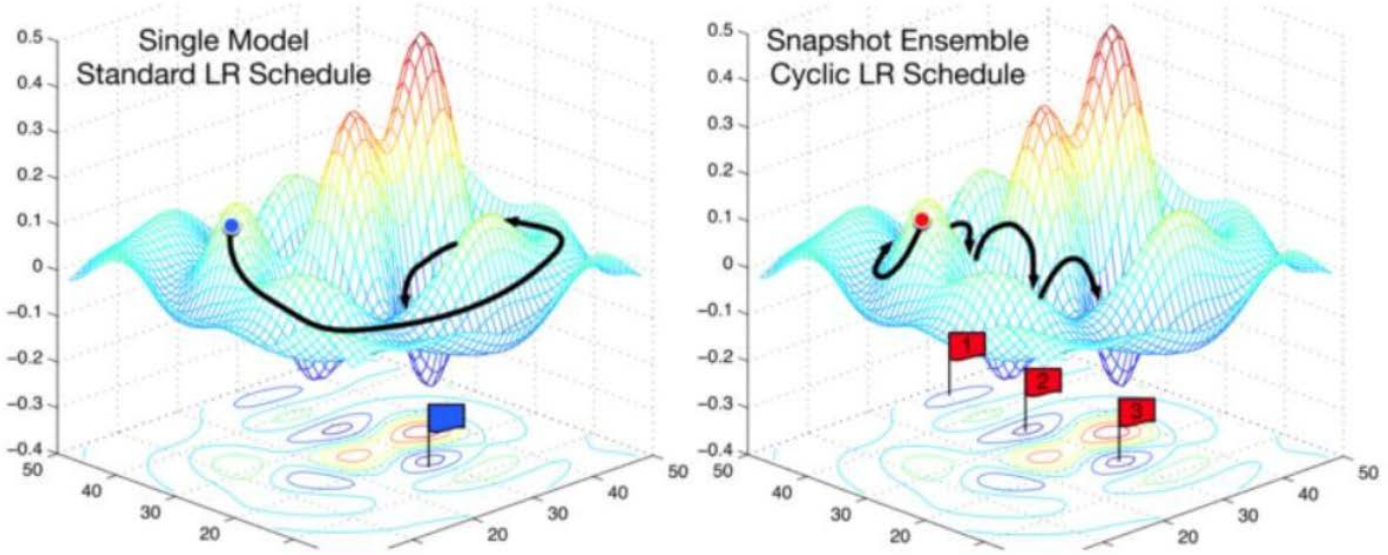
这种训练方式就像在最优化路程中, 截取了几个快照Snapshot, 因此命名为Snapshot Ensembling. 下图中的右半部分就是对这种方法的图像表现.



快照集成（Snapshot Ensembling）

下图对比了使用固定学习率的单个模型与使用循环学习率的快照集成的收敛过程，快照集成是在每次学习率周期末尾保存模型，然后在预测时使用。

快照集成的周期长度为 20 到 40 个 epoch。较长的学习率周期是为了在权值空间中找到足够具有差异化的模型，以发挥集成的优势。如果模型太相似，那么集成模型中不同网络的预测将会过于接近，以至于集成并不会带来多大益处了。



快照集成（Snapshot Ensembling）

论文中采用了**Cyclic Cosine Annealing**方法, 很早地就下调了学习率, 使训练尽快地到达第一个局部极小, 得到第一个模型. 然后提升学习率, 扰乱模型, 使得模型脱离局部极小, 然后重复上述步骤若干次, 直到获取指定数量的模型.

而学习率的变化, 论文中使用如下的函数:

$$\alpha(t) = f(\mod(t-1, \lceil T/M \rceil))$$

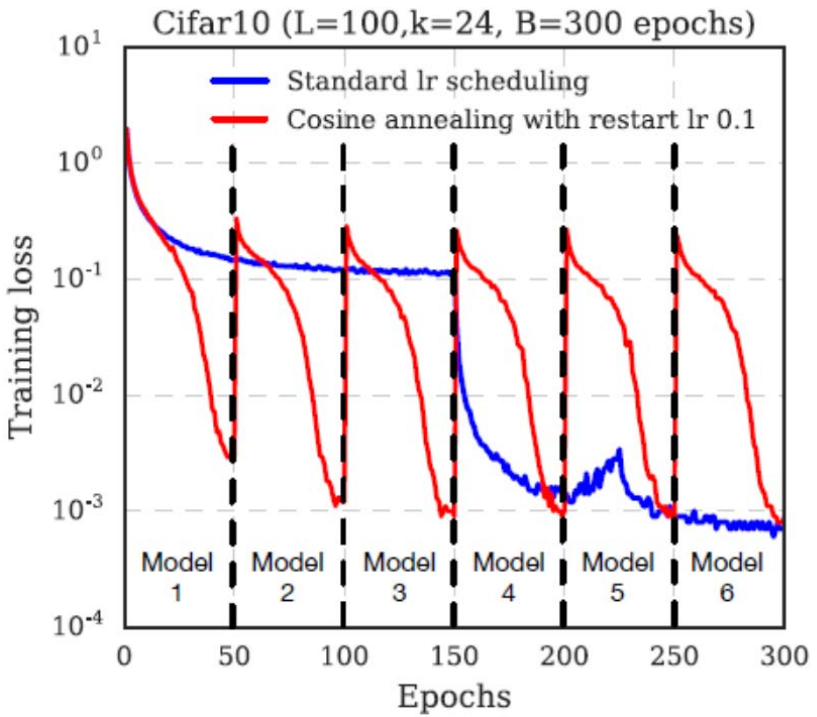
其中, t是迭代轮数; T是总的最大训练轮数; f是单调递减函数; M是循环的数量, 也就是最终模型的数量. 换句话说, 我们将整个训练过程划分成了M个循环, 在每个循环的开始阶段, 使用较大的学习率, 然后退火到小的学习率. α=f(0)给予模型足够的能量脱离局部极小, 而较小的学习率α=f(⌈T/M⌉)又能使模型收敛于一个表现较好的局部极小.

快照集成（Snapshot Ensembling）

论文中使用如下的shifted cosine function:

$$\alpha(t) = \frac{\alpha_0}{2}(\cos(\frac{\pi \mod (t-1, \lceil T/M \rceil)}{\lceil T/M \rceil}) + 1)$$

α_0是初始的学习率, 而α=f(⌈T/M⌉)≈0这保证了最小的学习率足够小.



实验结果

	Method	C10	C100	SVHN	Tiny ImageNet
ResNet-110	Single model	5.52	28.02	1.96	46.50
	NoCycle Snapshot Ensemble	5.49	26.97	1.78	43.69
	SingleCycle Ensembles	6.66	24.54	1.74	42.60
	Snapshot Ensemble ($\alpha_0 = 0.1$)	5.73	25.55	**1.63**	40.54
	Snapshot Ensemble ($\alpha_0 = 0.2$)	**5.32**	**24.19**	1.66	**39.40**
Wide-ResNet-32	Single model	5.43	23.55	1.90	39.63
	Dropout	4.68	22.82	1.81	36.58
	NoCycle Snapshot Ensemble	5.18	22.81	1.81	38.64
	SingleCycle Ensembles	5.95	21.38	1.65	35.53
	Snapshot Ensemble ($\alpha_0 = 0.1$)	**4.41**	**21.26**	1.64	35.45
	Snapshot Ensemble ($\alpha_0 = 0.2$)	4.73	21.56	**1.51**	**32.90**
DenseNet-40	Single model	5.24*	24.42*	1.77	39.09
	Dropout	6.08	25.79	1.79*	39.68
	NoCycle Snapshot Ensemble	5.20	24.63	1.80	38.51
	SingleCycle Ensembles	5.43	22.51	1.87	38.00
	Snapshot Ensemble ($\alpha_0 = 0.1$)	4.99	23.34	**1.64**	37.25
	Snapshot Ensemble ($\alpha_0 = 0.2$)	**4.84**	**21.93**	1.73	**36.61**
DenseNet-100	Single model	3.74*	19.25*	-	-
	Dropout	3.65	18.77	-	-
	NoCycle Snapshot Ensemble	3.80	19.30	-	-
	SingleCycle Ensembles	4.52	18.38		
	Snapshot Ensemble ($\alpha_0 = 0.1$)	3.57	18.12	-	-
	Snapshot Ensemble ($\alpha_0 = 0.2$)	**3.44**	**17.41**	-	-

Averaging Weights Leads to Wider Optima and Better Generalization

Pavel Izmailov[*1] **Dmitrii Podoprikhin**[*2,3] **Timur Garipov**[*4,5] **Dmitry Vetrov**[2,3] **Andrew Gordon Wilson**[1]
[1]Cornell University, [2]Higher School of Economics, [3]Samsung-HSE Laboratory,
[4]Samsung AI Center in Moscow, [5]Lomonosov Moscow State University

随机加权平均（Stochastic Weight Averaging，SWA）

图中我们可以得到:

- 两种方法都探索了高性能网络外围的点。可视化表明，这两种方法都是在空间区域进行探索但是具有周期性学习速率（CLR）的SGD的点通常比固定学习速率SGD的点要准确得多

- 训练的损失平面和测试在分布上是相似的，但它们并不是完全对齐的。

- 好的中心点可以导致更好的泛化，如果我们从优化轨迹中平均，我们得到一个更健壮的点，它比SGD的单个点具有更好的测试性能

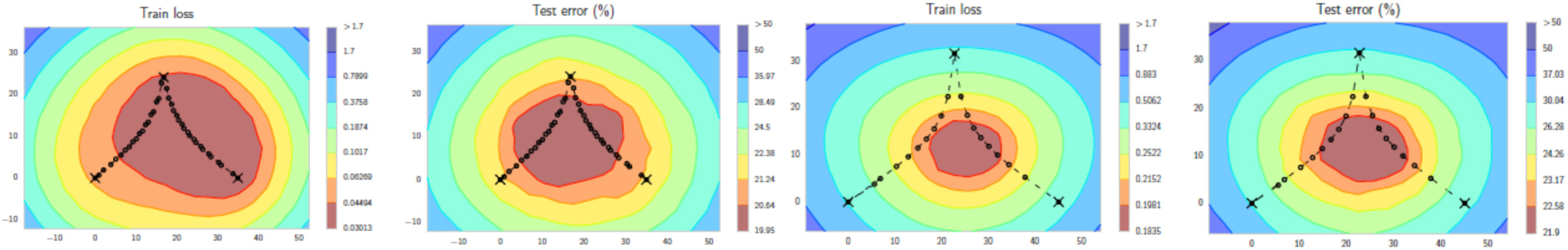


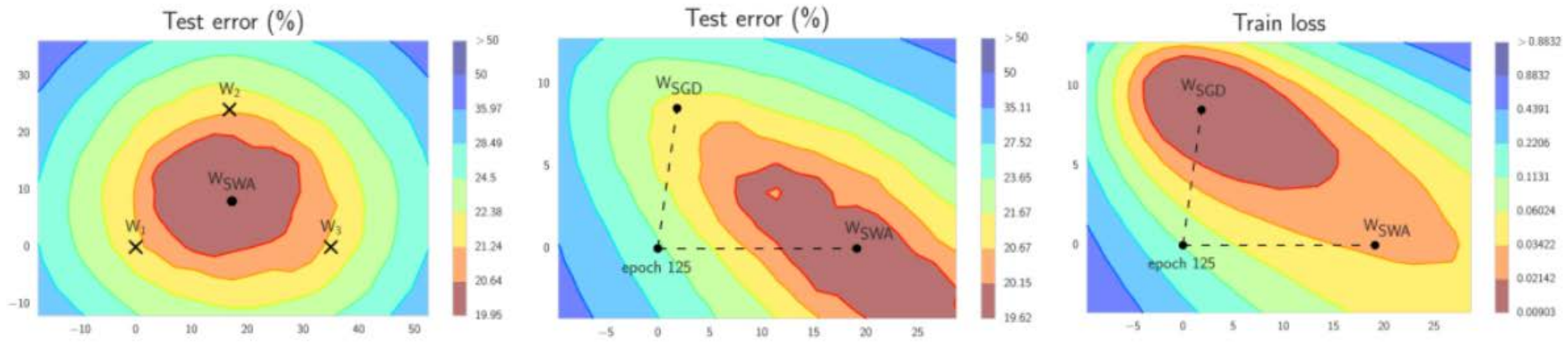
Figure 3: The L_2-regularized cross-entropy train loss and test error surfaces of a Preactivation ResNet-164 on CIFAR-100 in the plane containing the first, middle and last points (indicated by black crosses) in the trajectories with (**left two**) cyclical and (**right two**) constant learning rate schedules.

随机加权平均（Stochastic Weight Averaging，SWA）

SWA 的直觉来自以下由经验得到的观察：

- 每个学习率周期得到的局部极小值倾向于堆积在损失平面的低损失值区域的边缘
（上图左侧的图形中，褐色区域误差较低，点 W1、W2、W3 分别表示 3 个独立训练的网络，位于褐色区域的边缘）

- 对这些点取平均值，可能得到一个宽阔的泛化解，其损失更低（上图左侧图形中的 WSWA）



左图：W1、W2、W3 分别代表 3 个独立训练的网络，WSWA 为其平均值。

中图： W_{SWA}在测试集上的表现超越了 SGD。

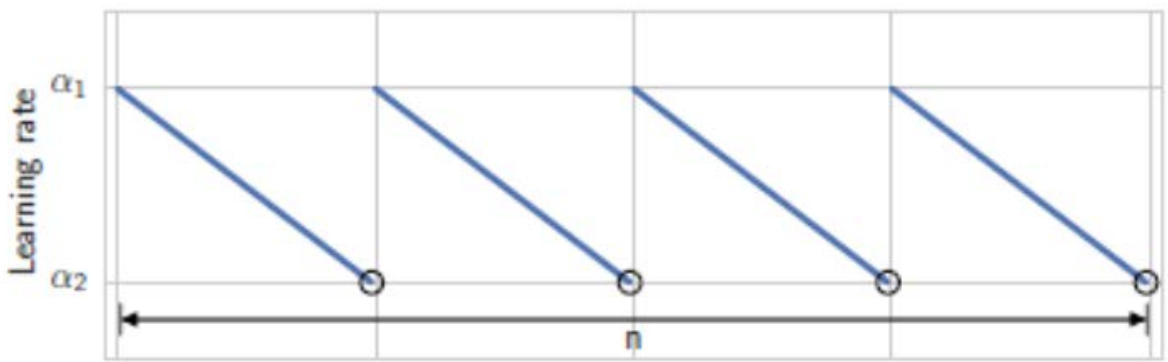
右图： W_{SWA} 在训练时的损失比 SGD 要高。

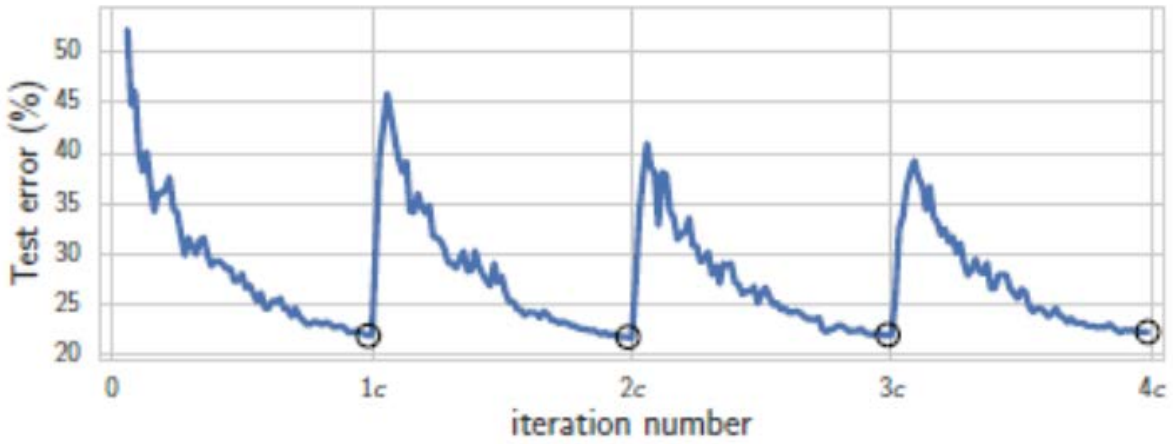
学习率

使用cyclical learning rate(CLR) schedule

$$\alpha(i) = (1 - t(i))\alpha_1 + t(i)\alpha_2,$$
$$t(i) = \frac{1}{c}\left(\mathrm{mod}(i-1, c) + 1\right).$$

学习率在每个周期线性降低，从α_1降低到α_2





工作原理

它只保存两个模型，而不是许多模型的集成：

- 第一个模型保存模型权值的平均值（W_{SWA}）。在训练结束后，它将是用于预测的最终模型。
- 第二个模型（W）将穿过权值空间，基于周期性学习率规划探索权重空间。

权重集成

$$W_{SWA} \leftarrow \frac{W_{SWA} \cdot n_{models} + W}{n_{models} + 1}$$

在每个学习率周期的末尾，第二个模型的当前权重将用来更新第一个模型的权重。

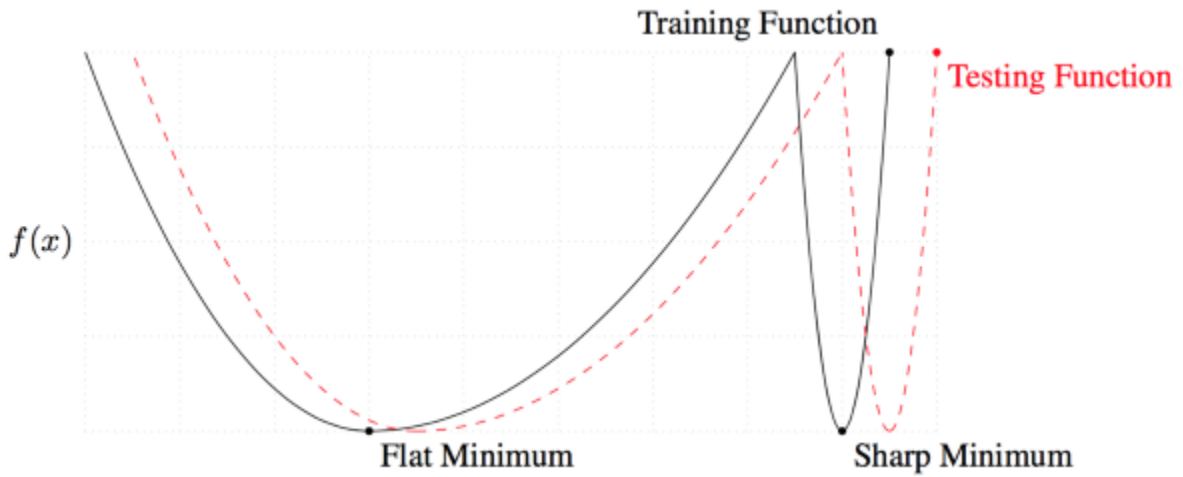
因此，在训练阶段，只需训练一个模型，并在内存中储存两个模型，节省了空间。

预测时只需要平均模型，基于其进行预测将比快照集成快很多，因为在那种集成中，需要使用多个模型进行预测，最后再进行平均。

解的宽度

窄极值和宽极值

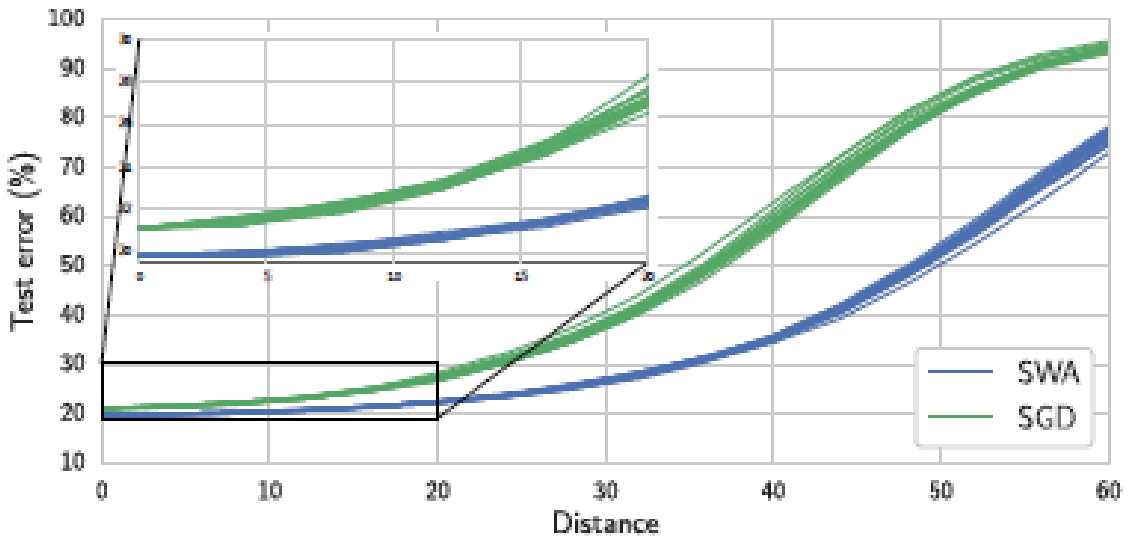
可视化并理解高维权值空间的几何特性非常困难，但我们又不得不去了解它。因为随机梯度下降的本质是，在训练时穿过这一高维空间中的损失平面，试图找到一个良好的解——损失平面上的一个损失值较低的“点”。不过后来我们发现，这一平面有很多局部极值。但这些局部极值并不都有一样好的性质。
一般极值点会有宽的极值和窄的极值，如下图所示

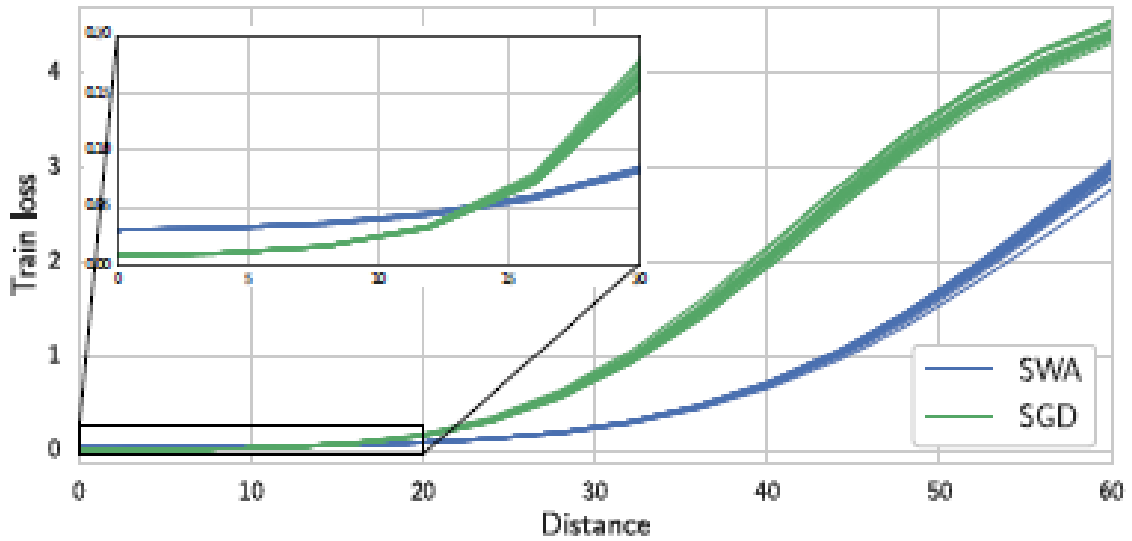


宽的局部极小值在训练和测试过程中产生类似的损失；但对于窄的局部极小值而言，训练和测试中产生的损失就会有很大区别。这意味着，宽的极值比窄的极值有更好的泛化性

解的宽度

$$w_{\mathrm{SWA}}(t,d) = w_{\mathrm{SWA}} + t \cdot d,$$
$$w_{\mathrm{SGD}}(t,d) = w_{\mathrm{SGD}} + t \cdot d,$$





实验结果

DNN (Budget)	SGD	FGE (1 Budget)	SWA		
			1 Budget	1.25 Budgets	1.5 Budgets
		CIFAR-100			
VGG-16 (200)	72.55 ± 0.10	74.26	73.91 ± 0.12	74.17 ± 0.15	74.27 ± 0.25
ResNet-164 (150)	78.49 ± 0.36	79.84	79.77 ± 0.17	80.18 ± 0.23	80.35 ± 0.16
WRN-28-10 (200)	80.82 ± 0.23	82.27	81.46 ± 0.23	81.91 ± 0.27	82.15 ± 0.27
PyramidNet-272 (300)	83.41 ± 0.21	–	–	83.93 ± 0.18	84.16 ± 0.15
		CIFAR-10			
VGG-16 (200)	93.25 ± 0.16	93.52	93.59 ± 0.16	93.70 ± 0.22	93.64 ± 0.18
ResNet-164 (150)	95.28 ± 0.10	95.45	95.56 ± 0.11	95.77 ± 0.04	95.83 ± 0.03
WRN-28-10 (200)	96.18 ± 0.11	96.36	96.45 ± 0.11	96.64 ± 0.08	96.79 ± 0.05
ShakeShake-2x64d (1800)	96.93 ± 0.10	–	–	97.16 ± 0.10	97.12 ± 0.06

Thanks for listening