SNAPSHOT ENSEMBLES: TRAIN 1, GET M MFOR FREE

ICLR-2017

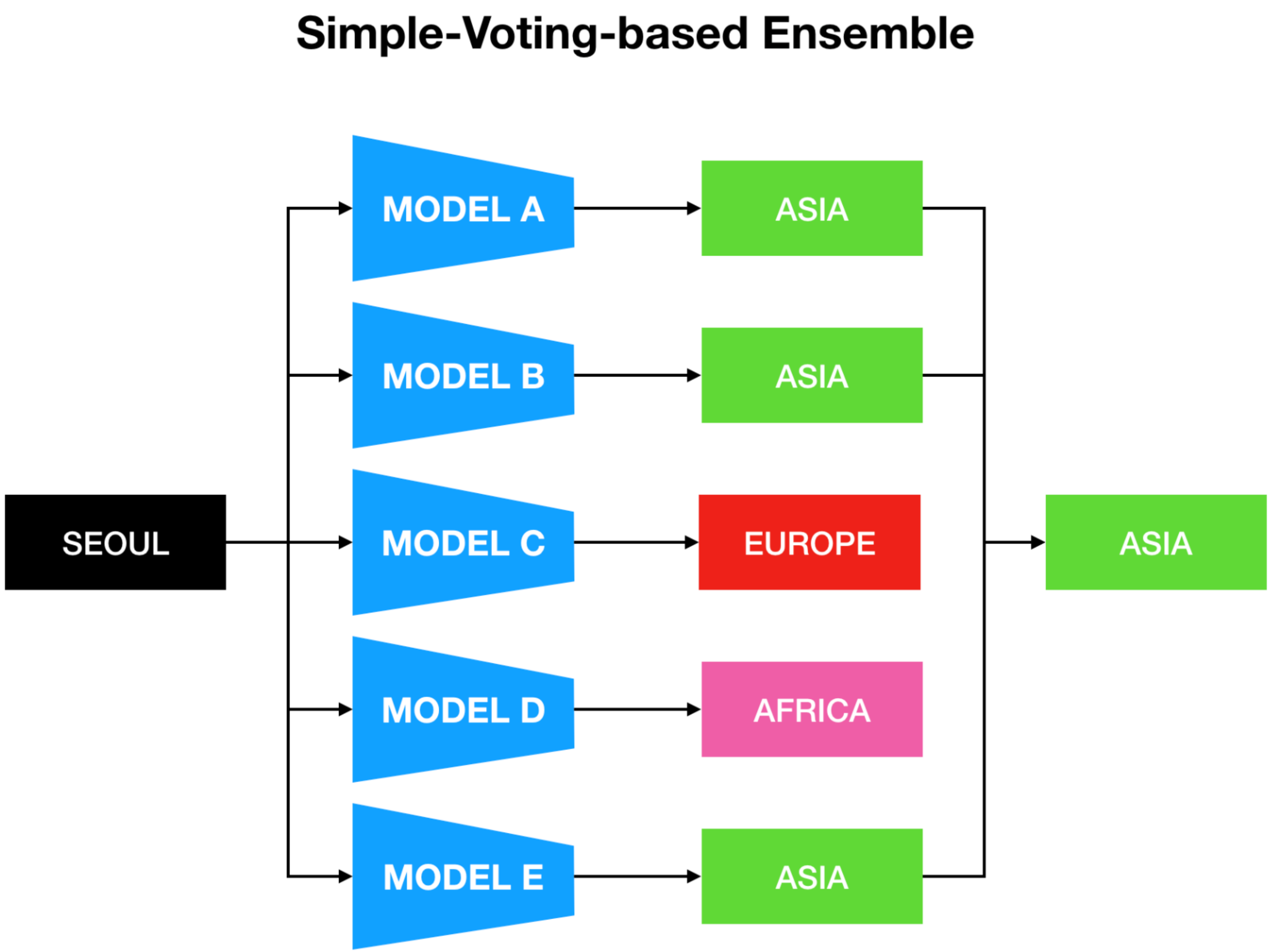
2018.07.05

张文涛

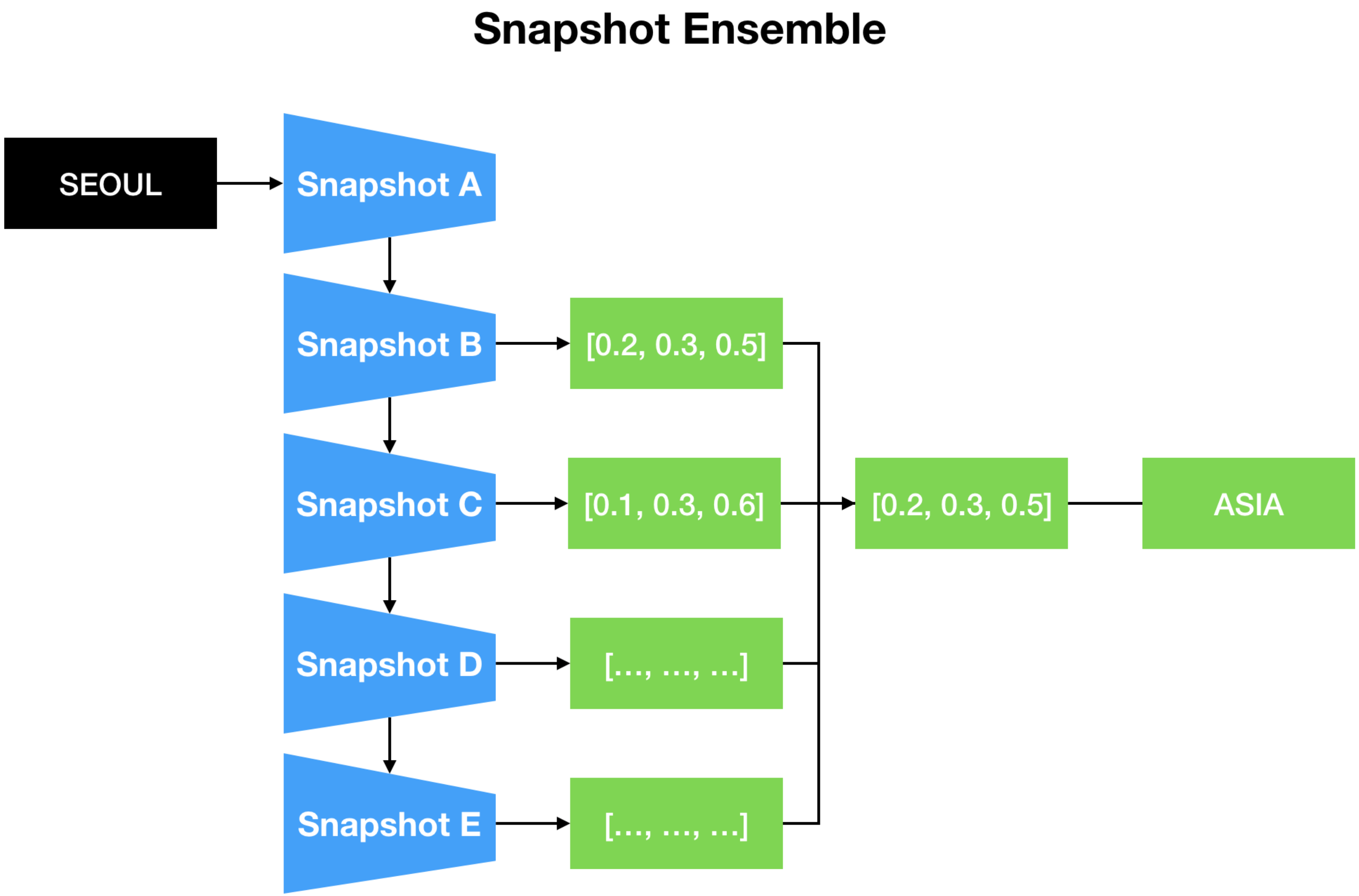
Introduction

- local minima contain useful information that may in fact improve model performance.
- local minima with flat basins tend to generalize better
- Although different local minima often have very similar error rates, the corresponding neural networks tend to make different mistakes
- The use of ensembling for deep networks is not nearly as widespread as it is for other algorithms.
- most studies focus on improving the generalization performance, while few of them address the cost of training ensembles

Method


Simple-Voting-based Ensemble
SEOUL
MODEL A
MODEL B
MODEL C
MODEL D
MODEL E
ASIA
ASIA
EUROPE
AFRICA
ASIA
ASIA

Theoretical performance



Method

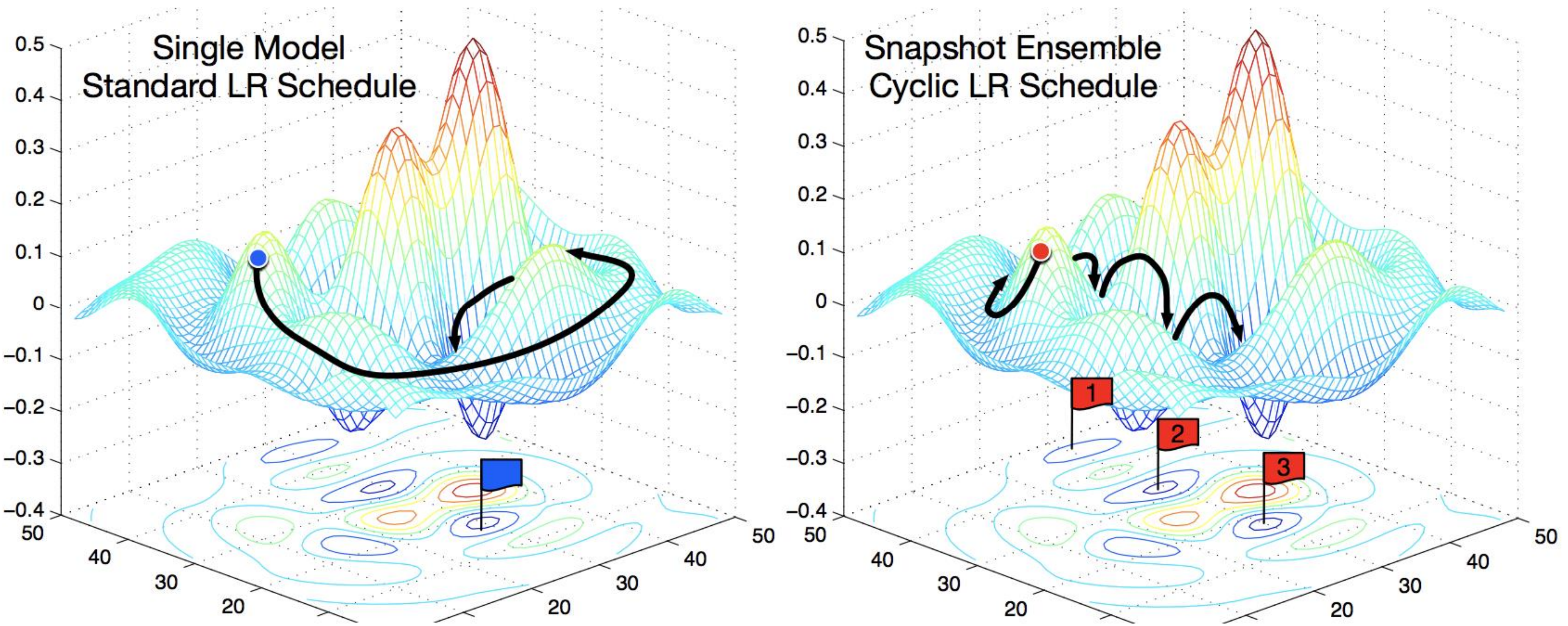


Figure 1: **Left:** Illustration of SGD optimization with a typical learning rate schedule. The model converges to a minimum at the end of training. **Right:** Illustration of Snapshot Ensembling. The model undergoes several learning rate annealing cycles, converging to and escaping from multiple local minima. We take a snapshot at each minimum for test-time ensembling.

SGDR (ICLR-2017)

- in each restart, the learning rate is initialized to some value and is scheduled to decrease

- Using previously acquired information

SGDR

$$r_t = r_{min}^i + \frac{1}{2}(r_{max}^i - r_{min}^i)\left(1 + \cos\left(\frac{T_{cur}}{T_i}\pi\right)\right)$$

i：第i次热重启

r_t：第t轮迭代的学习率

r_{min}^i：第i次热重启时，学习率的最小值

r_{max}^i：第i次热重启时，学习率的最大值

$Tcur$：从最近一次热重启开始，已经运行的epoch数

Ti：第i次热重启的最大可迭代epoch数

T_{mult}：每一次热重启都让Ti增大T_{mult}倍

SGDR

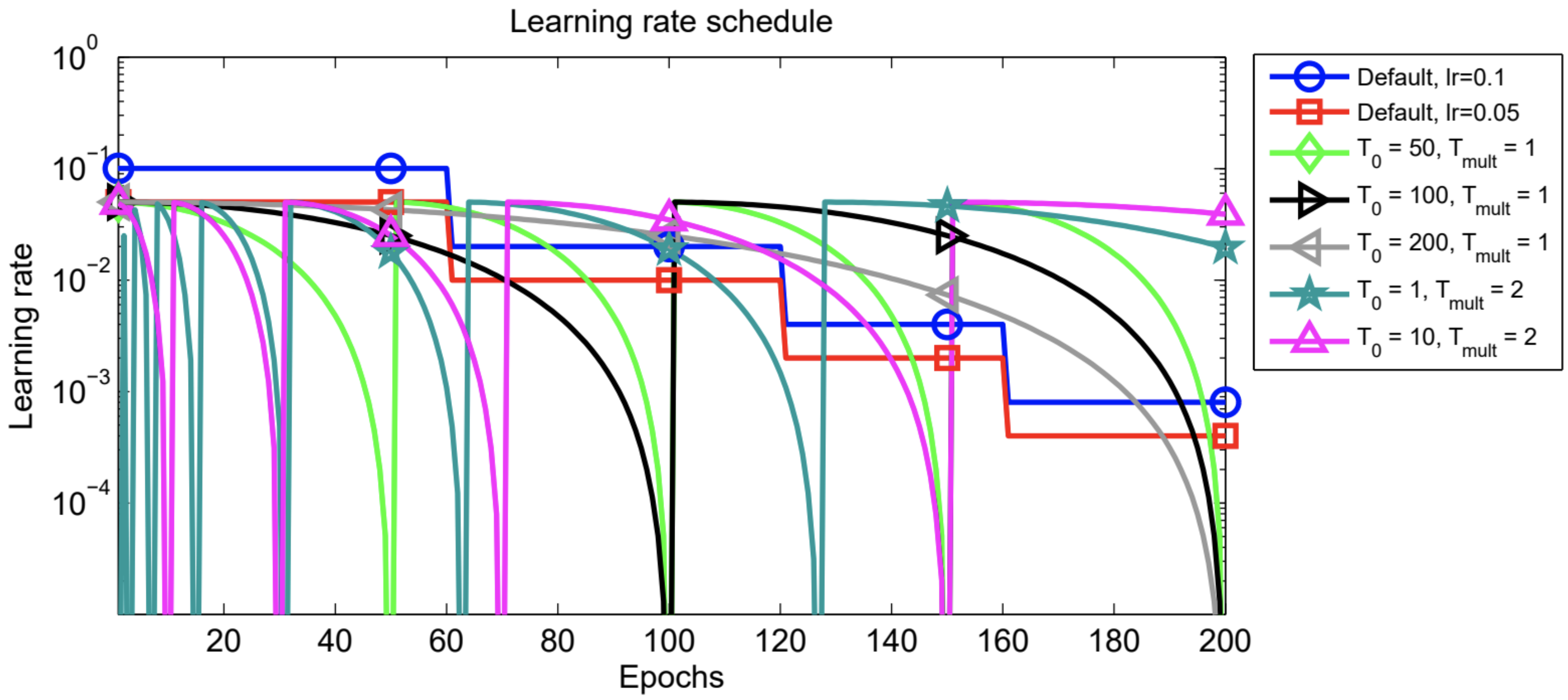
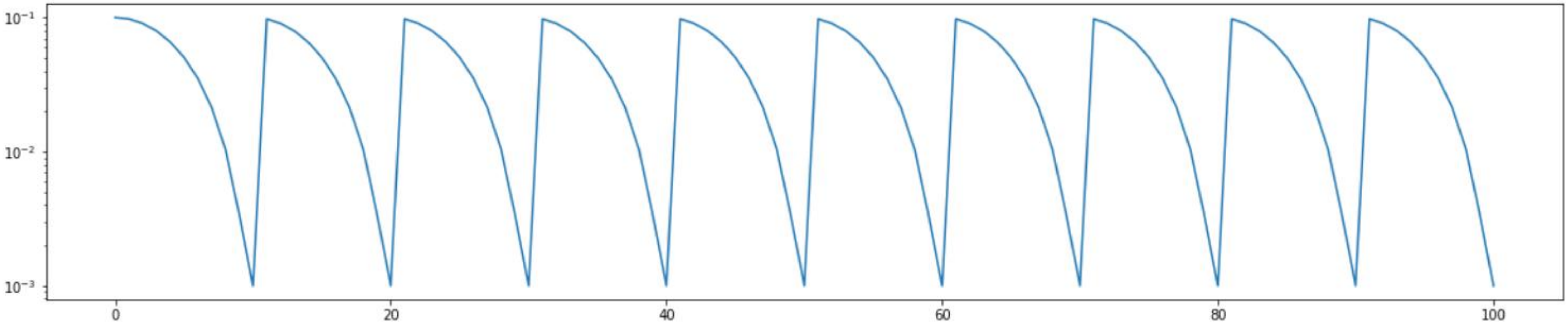


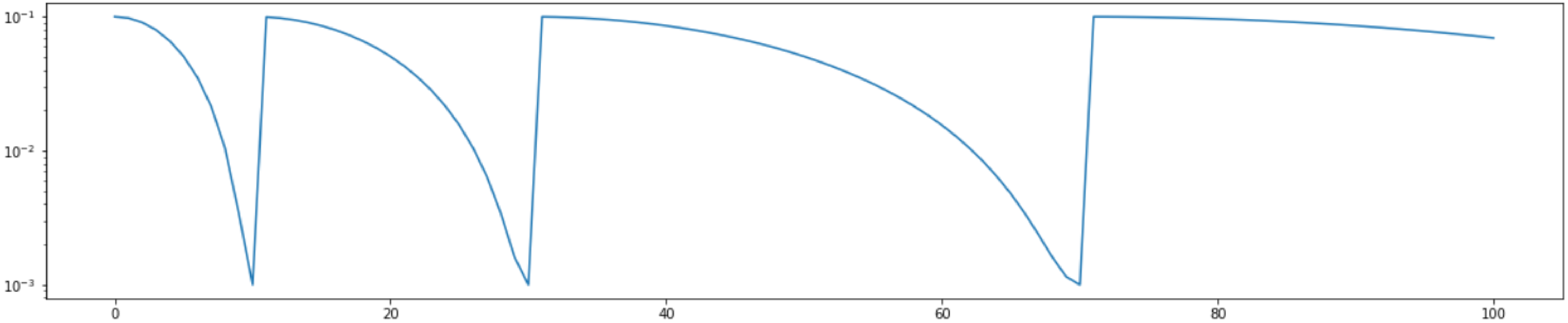
Figure 1: Alternative schedule schemes of learning rate η_t over batch index t: default schemes with $\eta_0 = 0.1$ (blue line) and $\eta_0 = 0.05$ (red line) as used by Zagoruyko & Komodakis (2016); warm restarts simulated every $T_0 = 50$ (green line), $T_0 = 100$ (black line) and $T_0 = 200$ (grey line) epochs with η_t decaying during i-th run from $\eta^i_{max} = 0.05$ to $\eta^i_{min} = 0$ according to eq. (5); warm restarts starting from epoch $T_0 = 1$ (dark green line) and $T_0 = 10$ (magenta line) with doubling ($T_{mult} = 2$) periods T_i at every new warm restart.

SGDR

$T_{max} = 10, T_{mult} = 1$ **for 100 epochs**



$T_{max} = 10, T_{mult} = 2$ **for 100 epochs**



SGDR-Experiments

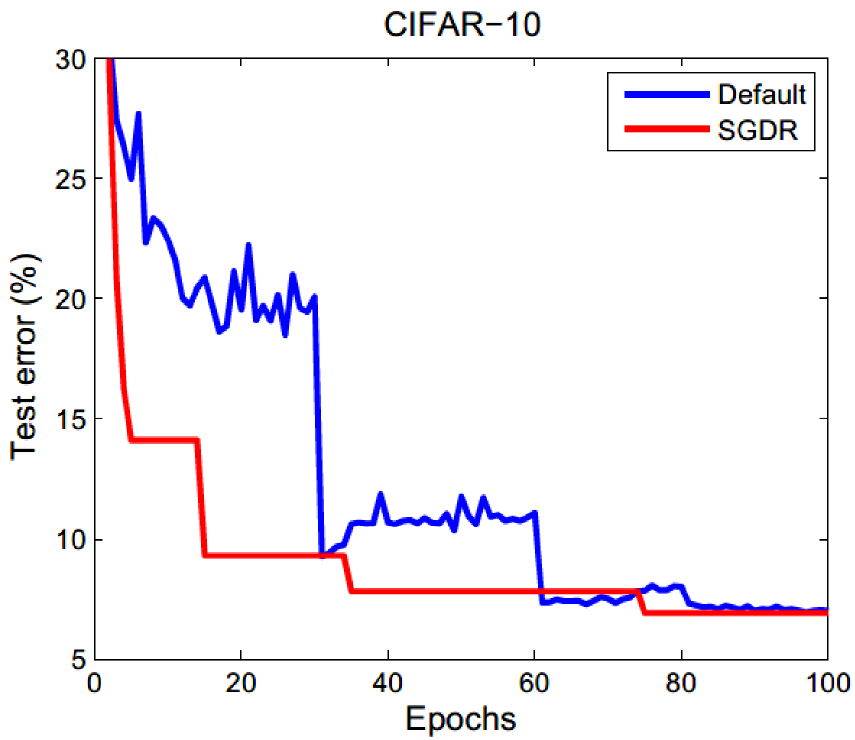


Figure 6: The median results of 5 runs for the best learning rate settings considered for WRN-28-1

We investigate different learning rate values for the default learning rate schedule (4 values out of [0.01, 0.025, 0.05, 0.1]) and SGDR (3 values out of [0.025, 0.05, 0.1]). In line with the results given in the main paper, Figure 6 suggests that SGDR is competitive in terms of anytime performance.

Background

- non-convex nature of neural networks

- SGD can converge to and escape from local minima on demand

Approach

Inputs:

E is the Snapshot ensemble classifier

M is the number of snapshot models

Procedure:

$for\ i \leftarrow 1\ to\ M\ \{$

1. $Let\ SGD\ converge\ to\ local\ minima\ along\ its\ optimization\ path$

2. $E_i \leftarrow\ \ save\ the\ weights\ and\ get\ the\ trained\ model$

3. $restart\ the\ optimization\ with\ a\ large\ learning\ rate$
   $to\ escape\ the\ current\ local\ minimums \qquad \}$

$$h_{Ensemble} = \frac{1}{m}\sum_{i=0}^{m-1} h_{M-i}(x) \quad (m \le M)$$

Approach

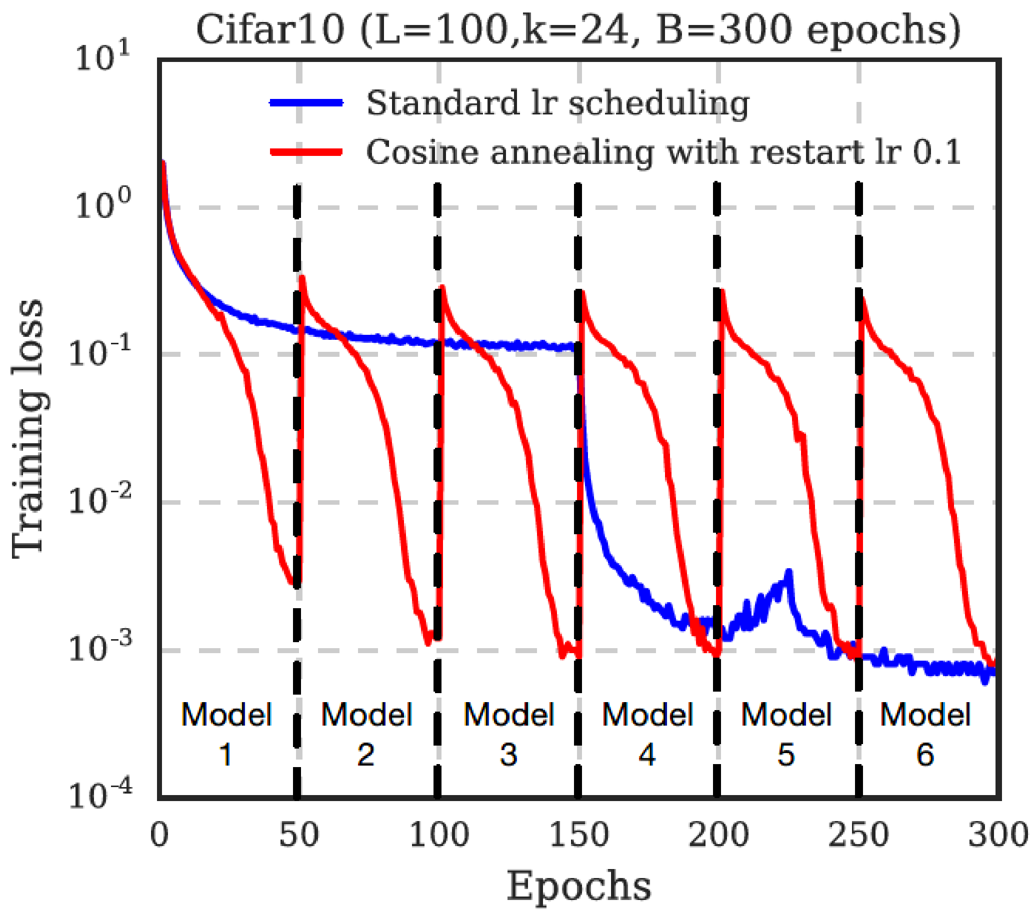


Figure 2: Training loss of 100-layer DenseNet on CIFAR10 using standard learning rate (blue) and $M = 6$ cosine annealing cycles (red). The intermediate models, denoted by the dotted lines, form an ensemble at the end of training.

Experiments-setting

Datasets: CIFAR10,CIFAR100,SVHN,Imagenet

Architectures: ResNet, Wide ResNet, DenseNet

Methods:

- **Single Model**: trained with a standard learning rate, dropping the learning rate from 0.1 to 0.01 halfway through training, and then to 0.001 when training is at 75%
- **Dropout**: drops nodes during training with a probability of 0.2
- **Snapshot Ensemble** : trained with the cyclic cosine learning rate
- **NoCycle Snapshot Ensemble**: a Snapshot Ensemble with a non-cyclic learning rate
- **SingleCycle Ensemble**: the network is re-initialized at the beginning of every cosine learning rate cycle

Experiments

	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**	-	-

Table 1: Error rates (%) on CIFAR-10 and CIFAR-100 datasets. All methods in the same group are trained for the same number of iterations. Results of our method are colored in blue, and the best result for each network/dataset pair are **bolded**. * indicates numbers which we take directly from Huang et al. (2016a).

Results

- In most cases, Snapshot ensembles achieve lower error than any of the baseline methods
- The NoCycle Snapshot Ensemble generally has little effect on performance, and in some instances even increases the test error
- As the model size increases, the SingleCycle Ensemble’s performance decline

Experiments

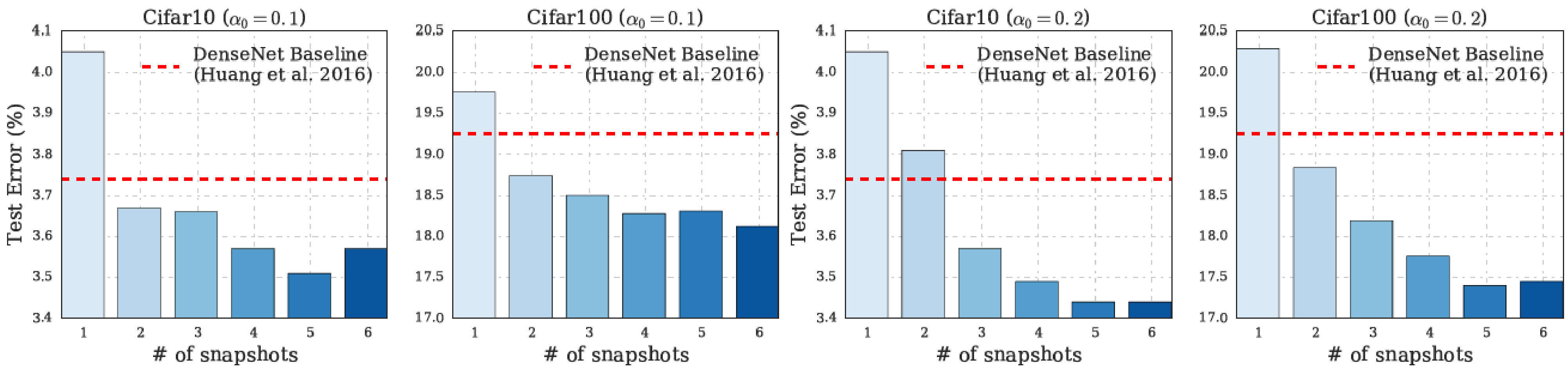
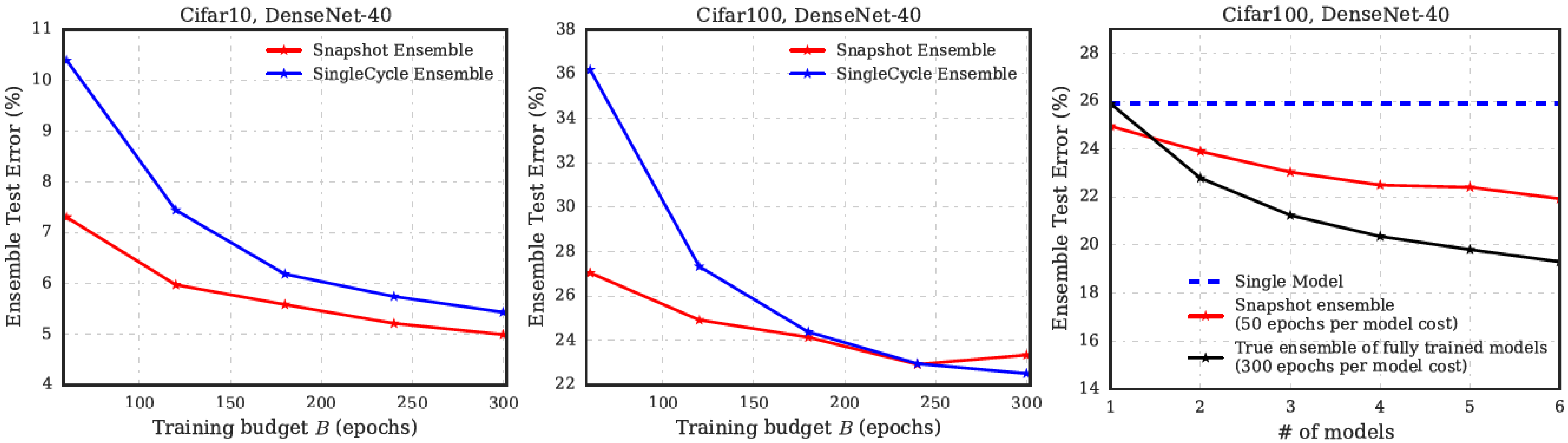


Figure 3: DenseNet-100 Snapshot Ensemble performance on CIFAR-10 and CIFAR-100 with restart learning rate $\alpha_0 = 0.1$ (left two) and $\alpha_0 = 0.2$ (right two). Each ensemble is trained with $M = 6$ annealing cycles (50 epochs per each).

Result:

In most cases, ensembles with the larger restart learning rate perform better, presumably because the strong perturbation in between cycles increases the diversity of local minima

Experiments



Result:

- As training budget decreases, Snapshot Ensembles still yield competitive results, while the performance of the SingleCycle Ensembles degrades rapidly
- Our method achieves performance that is comparable with ensembling of 2 independent models, but with the training cost of one model.

Conclusion

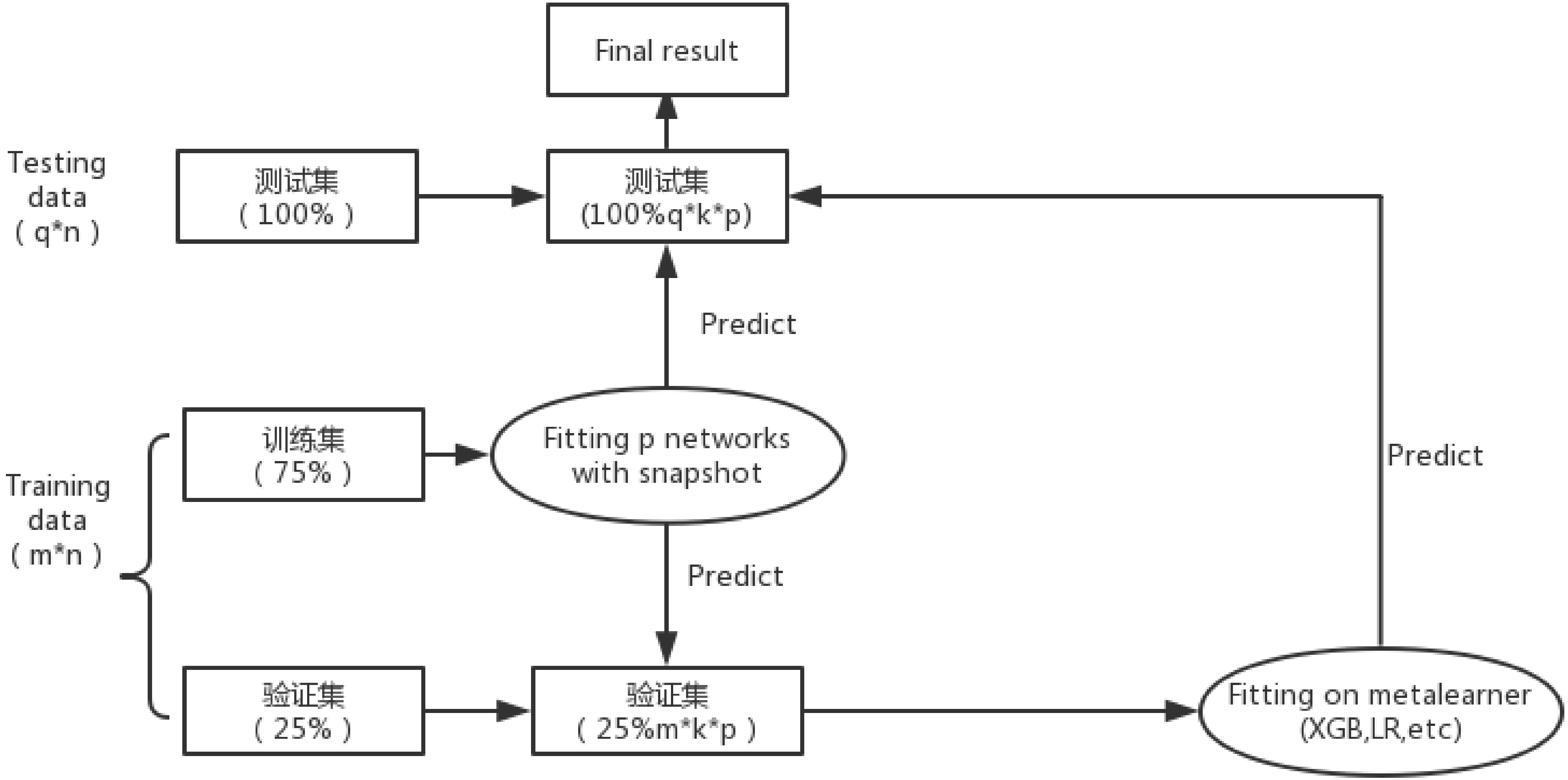
Advantage：

- obtain ensembles of neural networks without any additional training cost
- High performance

Disadvantage：

- The averaging or voting method is too simple
- Overfitting

Snapshot+Blending



Snapshot+Blending+Bagging

Inputs:

S is the training set
E is the Snapshot+Blending+Bagging classifier
T is the number of bootstrap samples

Procedure:

$$for\ i \leftarrow 1\ to\ T\ \{$$

$$S_b \leftarrow bootstrap\ sample\ from\ S$$

$$S_{oob} \leftarrow out\ of\ bag\ sample$$

$$E_i \leftarrow\ snapshot + blending\ model\ trained\ with\ S_b\ and\ S_{oob}\}$$

Future Work

- Ensemble with neural network
- Unbalanced data classification
- Concrete application(Syetem,Image,Text)