Ансамблевые методы в глубинном обучении

Гапонов Александр

Проблемы нейросетей

- Проблемы с качеством
  - высокая дисперсия
  - случайная инициализация весов
  - переобучение
- Долгое обучение

Основные идеи

- Хотим получить много моделей
- Хотим чтобы модели были “разнообразными”
- Будем прогонять тест через каждую модель и усреднять ответ
- Хотим поменьше дополнительных затрат (обучение, инференс)

Про разнообразие моделей в ансамбле

- Каждая модель имеет низкую долю ошибок
- Чем меньше пересечение в ложно классифицированных примерах тем лучше

Получение итогового ответа

- Усреднение вероятностей
- Усреднение предсказанной величины (для регрессии)
- Голосование

Датасеты

- CIFAR-10
- CIFAR-100
- ImageNet (20000+ classes)
- ImageNet-C

airplane

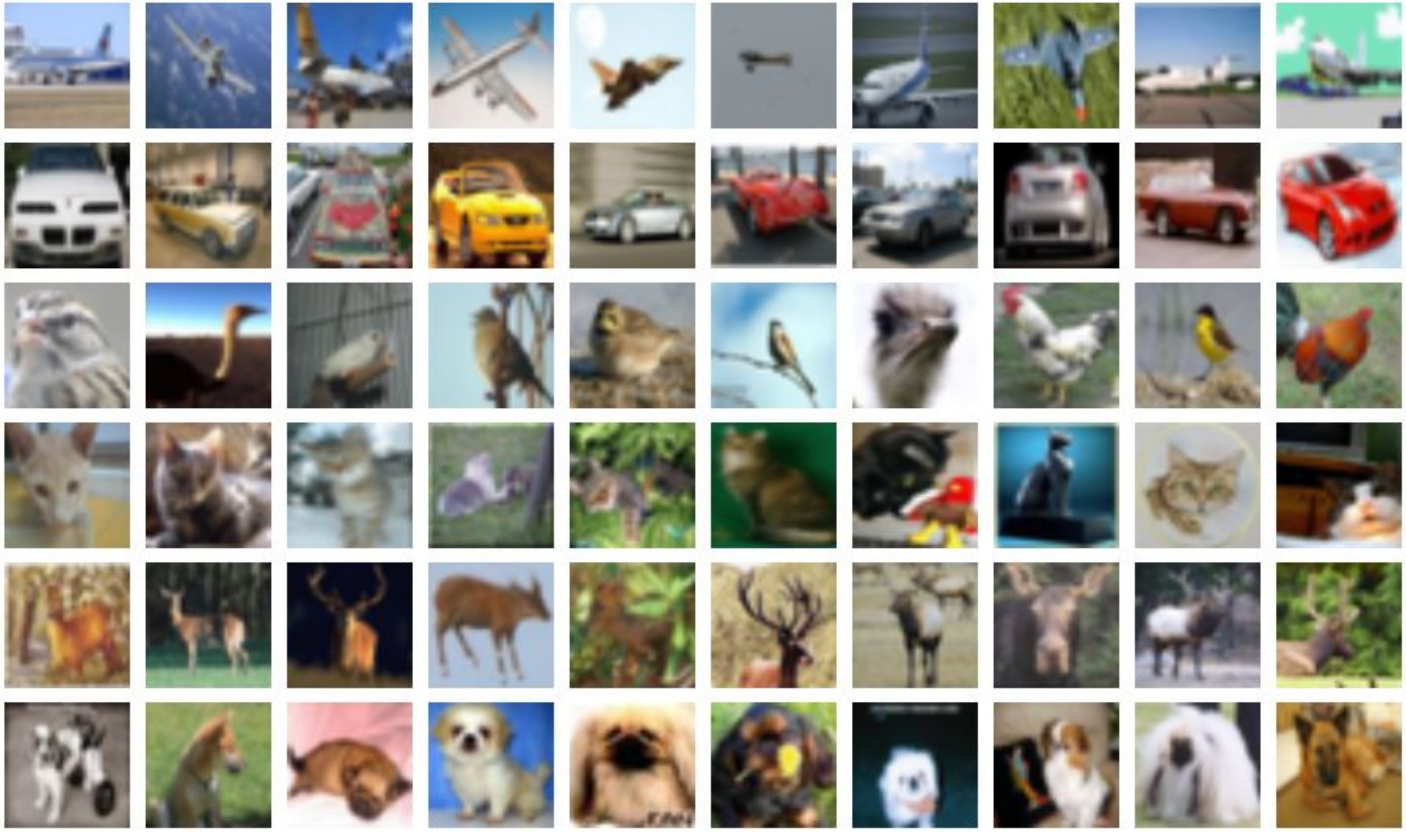
automobile

bird

cat

deer

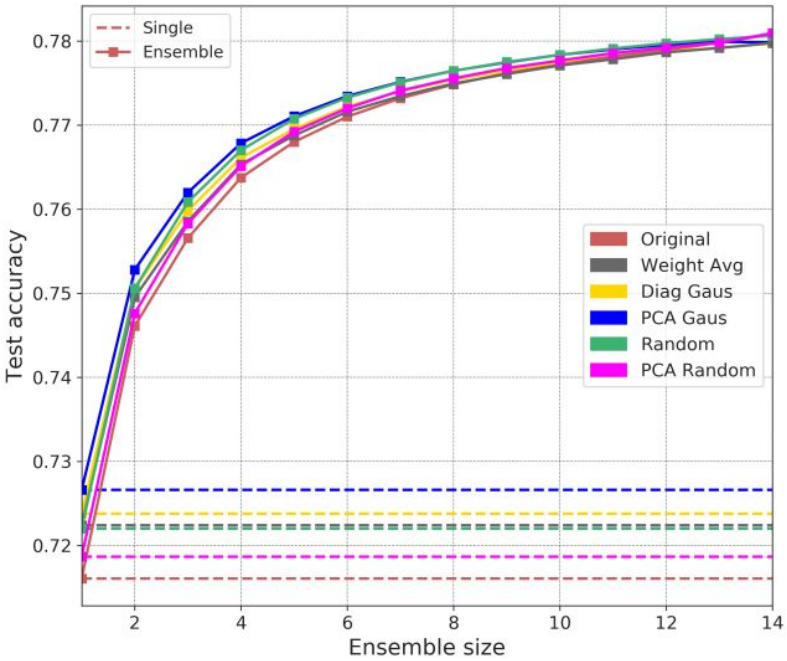
dog

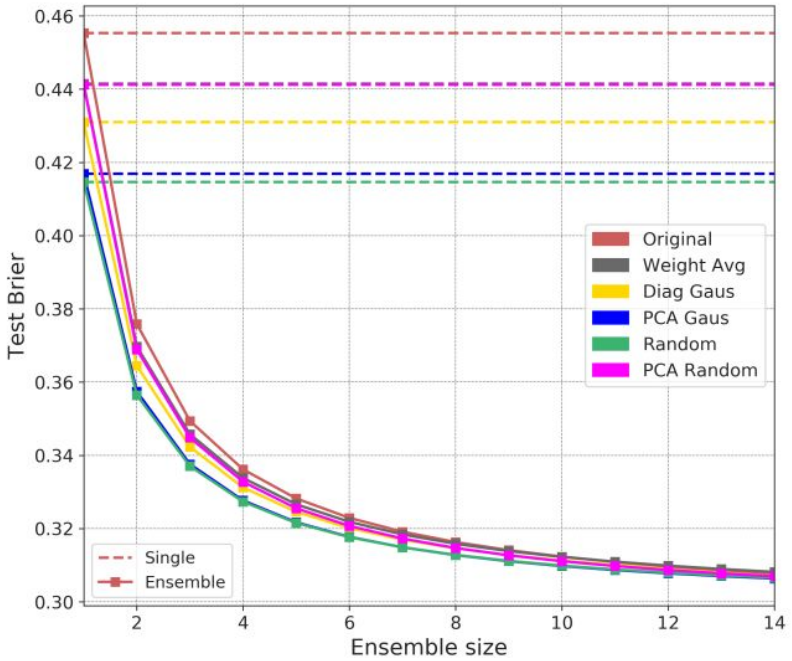


Deep Ensembles: Основные идеи

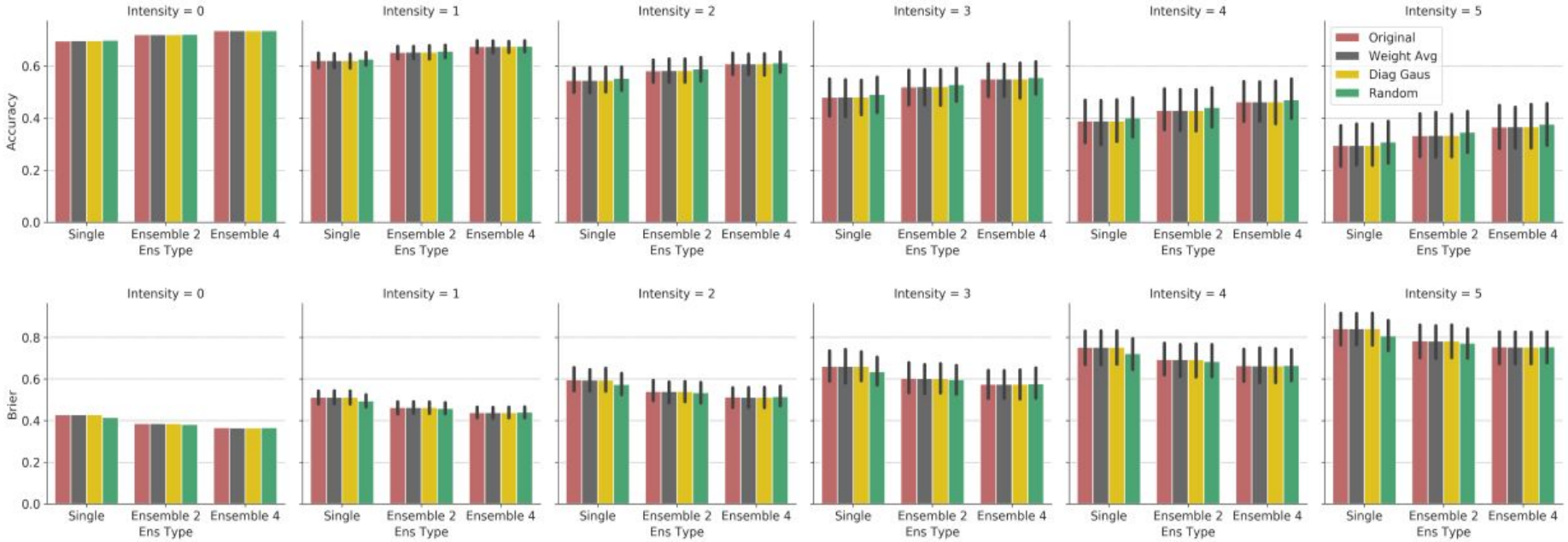
- N моделей с одинаковой архитектурой
- Каждую модель инициализируем случайными весами и учим

Deep Ensembles: Сравнение качества CIFAR-10





Deep Ensembles: Сравнение качества ImageNet-C



Deep Ensembles: Итог

Преимущества

- Хорошее качество

Недостатки

- Затратное обучение
- Тяжелый инференс

Snapshot Ensembles: Основные идеи

- Получим ансамбль в ходе одного процесса обучения
- Введем такой learning rate чтобы посетить много локальных оптимумов
- Каждые K итераций SGD будем сохранять модель

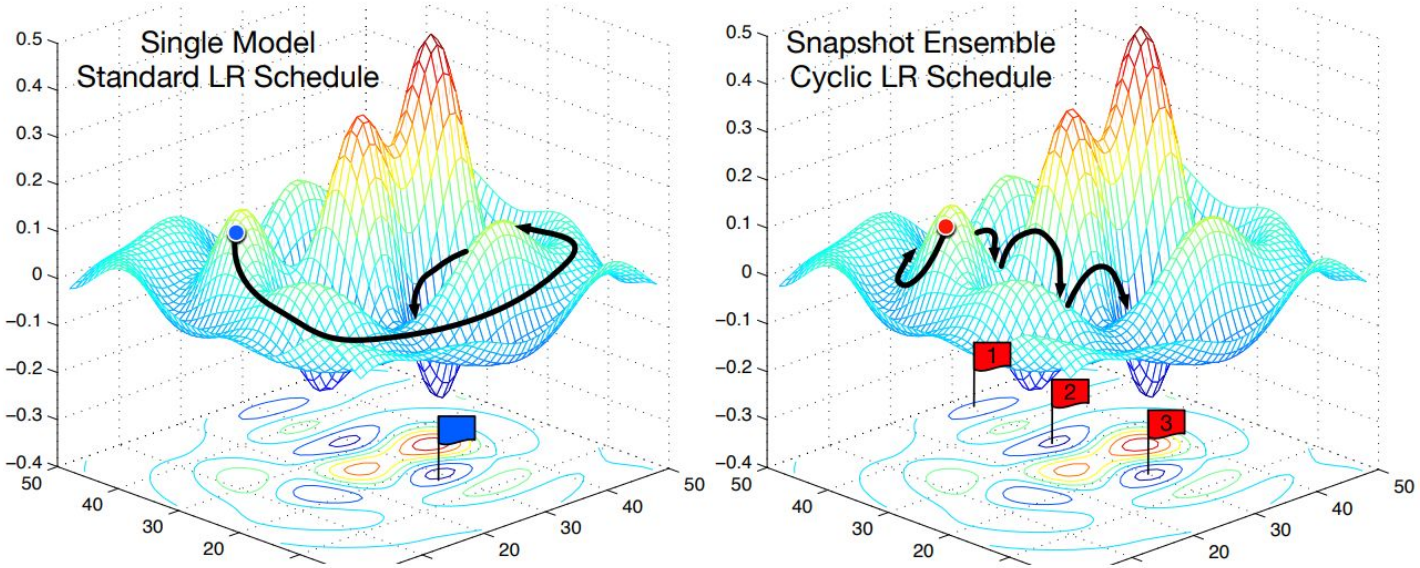


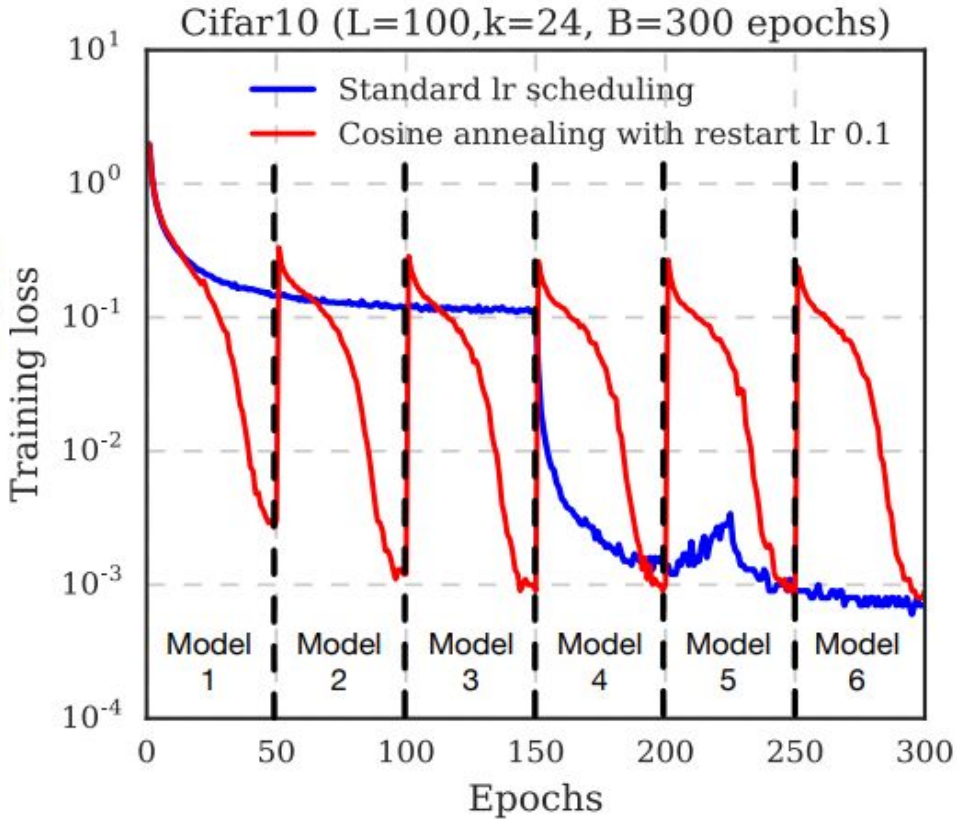
Figure 1: **Left:** Illustration of SGD optimization with a typical learning rate schedule. The model converges to a minimum at the end of training. **Right:** Illustration of Snapshot Ensembling. The model undergoes several learning rate annealing cycles, converging to and escaping from multiple local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

Snapshot Ensembles: Learning rate

$$\alpha(t) = f\left(\mathrm{mod}\left(t-1, \lceil T/M \rceil\right)\right),$$

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

t - итерация SGD
T - сколько итераций будем учить
f - монотонно убывающая функция
M - количество снапшотов
a0 - initial learning rate



Snapshot Ensembles: Сравнение качества

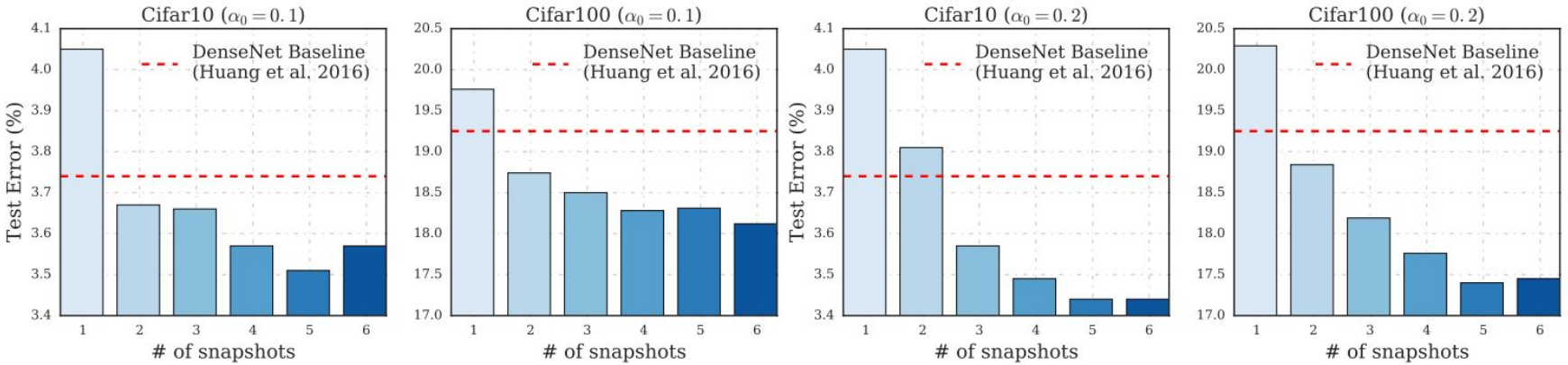
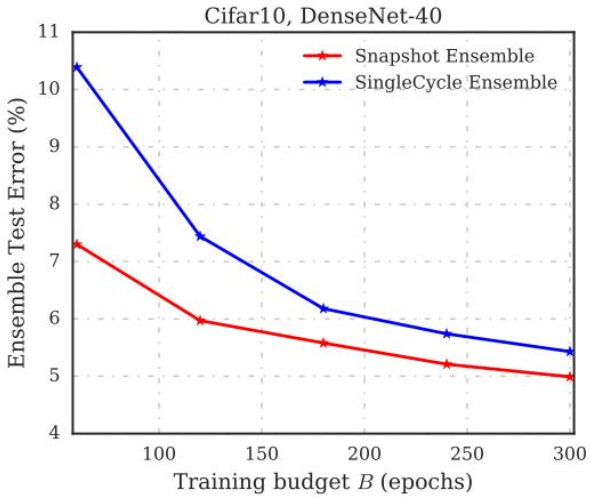
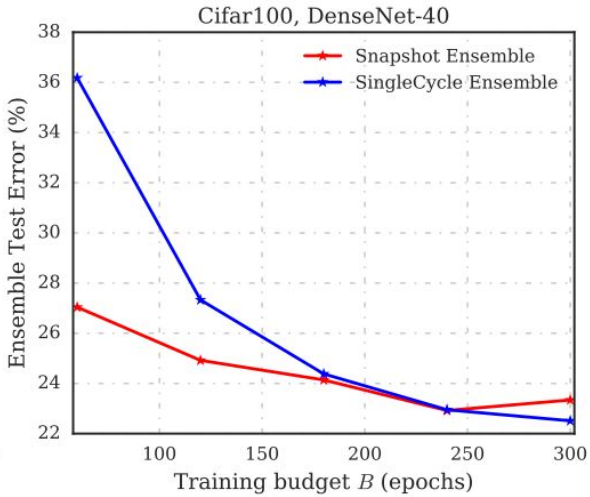


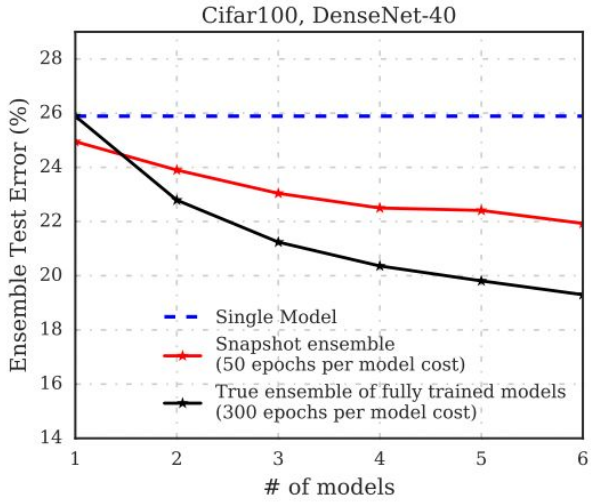
Figure 3: DenseNet-100 Snapshot Ensemble performance on CIFAR-10 and CIFAR-100 with restart learning rate $\alpha_0 = 0.1$ (left two) and $\alpha_0 = 0.2$ (right two). Each ensemble is trained with $M = 6$ annealing cycles (50 epochs per each).

	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).


Cifar10, DenseNet-40
Snapshot Ensemble
SingleCycle Ensemble
Ensemble Test Error (%)
Training budget B (epochs)


Cifar100, DenseNet-40
Snapshot Ensemble
SingleCycle Ensemble
Ensemble Test Error (%)
Training budget B (epochs)


Cifar100, DenseNet-40
Single Model
Snapshot ensemble
(50 epochs per model cost)
True ensemble of fully trained models
(300 epochs per model cost)
Ensemble Test Error (%)
# of models

Snapshot Ensembles: Итог

Преимущества

- Один процесс обучения

Недостатки

- Не такой сильный прирост в качестве как в deep ensembles
- Тяжелый инференс

Dropout Ensembles: Основные идеи

- Каждому нейрону присваиваем вероятность p с которой он отключается
- Учимся несколько итераций на подмножестве нейронов начальной сети
- Повторяем
- Еще можно независимо учить K разных нейросетей, которые являются подмножеством начальной

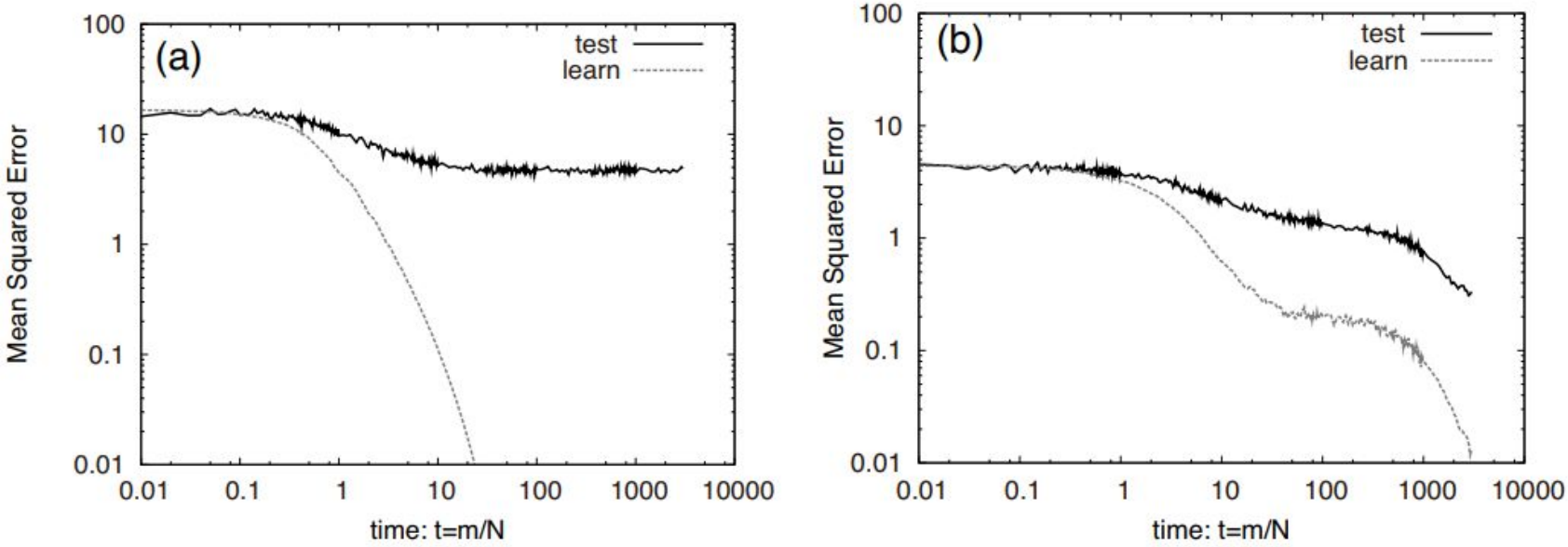


Figure 4: Effect of dropout. (a) is learning curve of SGD, and (b) is that of dropout learning.

	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).

Dropout Ensembles: Итоги

Преимущества

- Быстро
- На выходе одна модель

Недостатки

- Не всегда хороший прирост в точности

Fast Geometric Ensembling (FGE): Мотивация

- Локальные оптимумы соединены кривыми вдоль которых функция потерь слабо изменяется
- Можно подобрать функцию learning rate такую, что она учитывает эту особенность

FGE: Основные идеи

- По аналогии с snapshot ensembles будет один процесс обучения
- Сначала сойдёмся в какой-то оптимум
- Далее знаем, что он связан с другими оптимумами путями с маленькой ошибкой
- Будем использовать циклическую функцию изменения learning rate

FGE: Learning rate

$$\alpha(i) = \begin{cases} (1 - 2t(i))\alpha_1 + 2t(i)\alpha_2 & 0 < t(i) \leq \frac{1}{2} \\ (2 - 2t(i))\alpha_2 + (2t(i) - 1)\alpha_1 & \frac{1}{2} < t(i) \leq 1 \end{cases}$$

$t(i) = \frac{1}{c}(\mathrm{mod}(i - 1, c) + 1)$, the learning rates are $\alpha_1 > \alpha_2$.

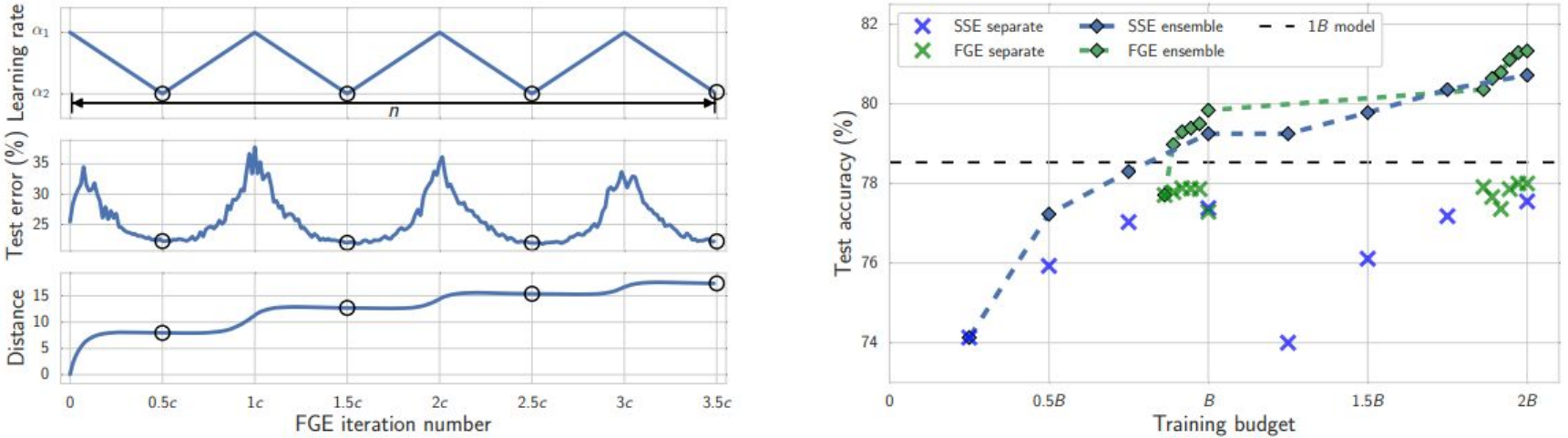


Figure 3: **Left:** Plot of the learning rate (**Top**), test error (**Middle**) and distance from the initial value $\hat{w}$ (**Bottom**) as a function of iteration for FGE with Preactivation-ResNet-164 on CIFAR-100. Circles indicate the times when we save models for ensembling. **Right:** Ensemble performance of FGE and SSE (Snapshot Ensembles) as a function of training time, using ResNet-164 on CIFAR-100 ($B = 150$ epochs). Crosses represent the performance of separate "snapshot" models, and diamonds show the performance of the ensembles constructed of all models available by the given time.

Table 1: Error rates (%) on CIFAR-100 and CIFAR-10 datasets for different ensembling techniques and training budgets. The best results for each dataset, architecture, and budget are **bolded**.

		CIFAR-100			CIFAR-10		
DNN (Budget)	method	$1B$	$2B$	$3B$	$1B$	$2B$	$3B$
VGG-16 (200)	Ind	27.4 ± 0.1	25.28	24.45	6.75 ± 0.16	5.89	5.9
	SSE	26.4 ± 0.1	25.16	24.69	6.57 ± 0.12	6.19	5.95
	FGE	$\mathbf{25.7 \pm 0.1}$	**24.11**	**23.54**	$\mathbf{6.48 \pm 0.09}$	**5.82**	**5.66**
ResNet-164 (150)	Ind	21.5 ± 0.4	19.04	18.59	4.72 ± 0.1	**4.1**	**3.77**
	SSE	20.9 ± 0.2	19.28	18.91	4.66 ± 0.02	4.37	4.3
	FGE	$\mathbf{20.2 \pm 0.1}$	**18.67**	**18.21**	$\mathbf{4.54 \pm 0.05}$	4.21	3.98
WRN-28-10 (200)	Ind	19.2 ± 0.2	17.48	17.01	3.82 ± 0.1	3.4	**3.31**
	SSE	17.9 ± 0.2	17.3	16.97	3.73 ± 0.04	3.54	3.55
	FGE	$\mathbf{17.7 \pm 0.2}$	**16.95**	**16.88**	$\mathbf{3.65 \pm 0.1}$	**3.38**	3.52

FGE: Итоги

Преимущества

- Один процесс обучения (но требуется заранее обучить модель)
- State-of-art точность

Недостатки

- Тяжелый инференс

Ссылки

- S. Fort, H. Hu, B. Lakshminarayanan. Deep Ensembles: A Loss Landscape Perspective
- Z. Liu, J. E. Hopcroft, K. Q. Weinberger. Snapshot Ensembles
- Analysis of dropout learning regarded as ensemble learning
- T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, A. Wilson. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs