Ансамблирование нейронных сетей

Панфилов Борис

Введение
Ансамбли в классическом ML
Deep ensembles
Snapshot ensembles
Fast Geometric ensembles
Dropout ensembles
Заключение

Введение
Ансамбли в классическом ML
Deep ensembles
Snapshot ensembles
Fast Geometric ensembles
Dropout ensembles
Заключение

Мотивация

- Улучшить качество модели
- Научиться определять случаи, когда модель не уверена в ответе

Возможные ограничения

- Время и стоимость обучения
- Память
- Инференс

Идея ансамблирования

1. Обучить K моделей
2. На этапе теста в качестве итогового ответа брать среднее арифметическое результатов этих моделей

Введение

Ансамбли в классическом ML

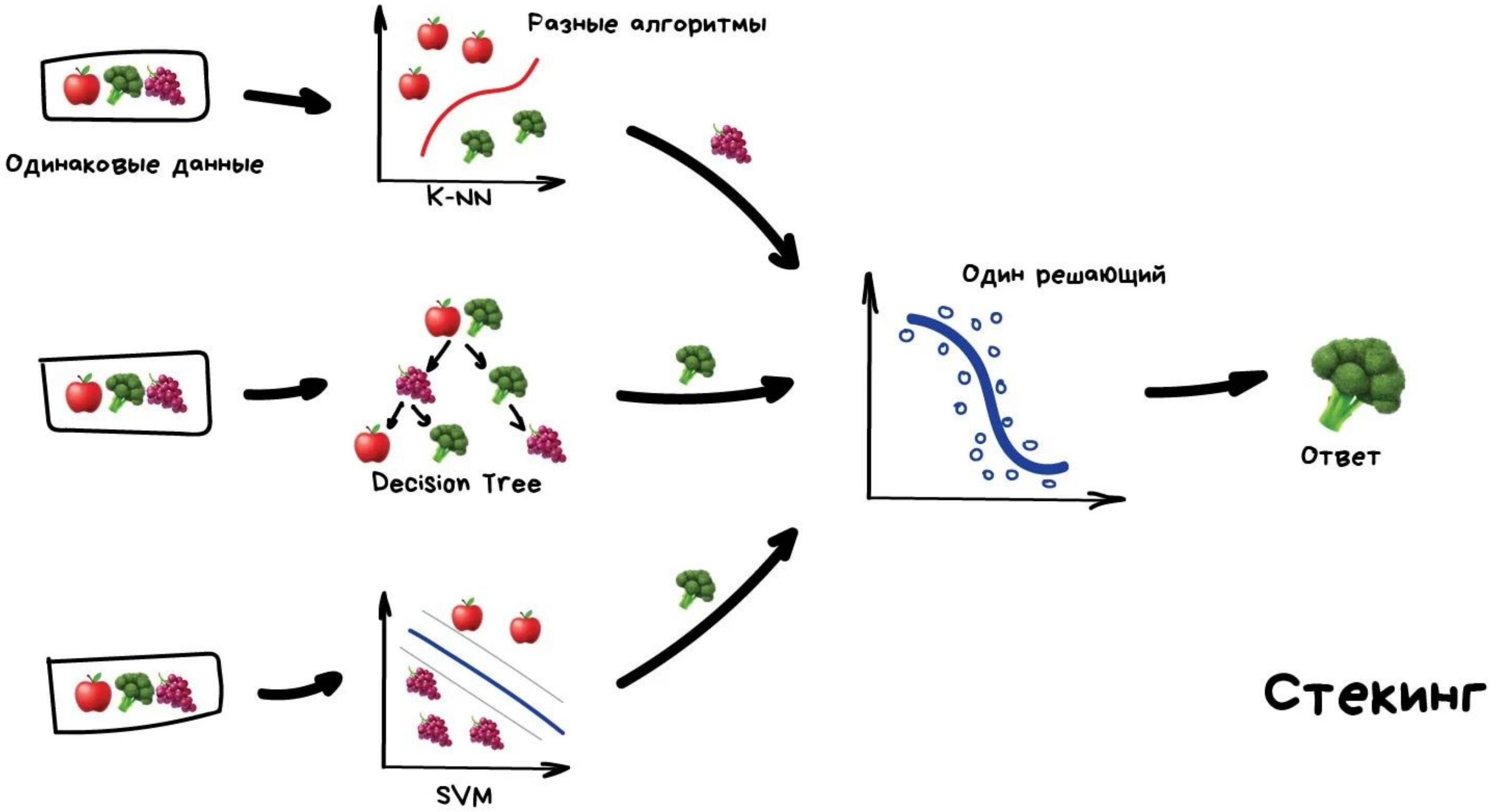
Deep ensembles

Snapshot ensembles

Fast Geometric ensembles

Dropout ensembles

Заключение



Разные алгоритмы
Одинаковые данные
K-NN
Один решающий
Decision Tree
Ответ
SVM
Стекинг

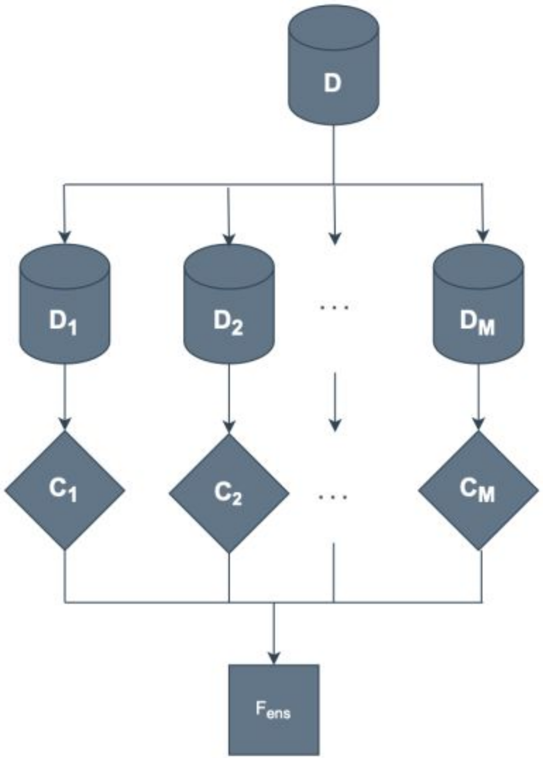


Figure 2: Bagging

$$b_1(x) := \arg\min_{b \in \mathcal{A}} \frac{1}{2} \sum_{i=1}^{\ell} (b(x_i) - y_i)^2$$

$$s_i^{(1)} = y_i - b_1(x_i)$$

$$b_2(x) := \arg\min_{b \in \mathcal{A}} \frac{1}{2} \sum_{i=1}^{\ell} (b(x_i) - s_i^{(1)})^2$$

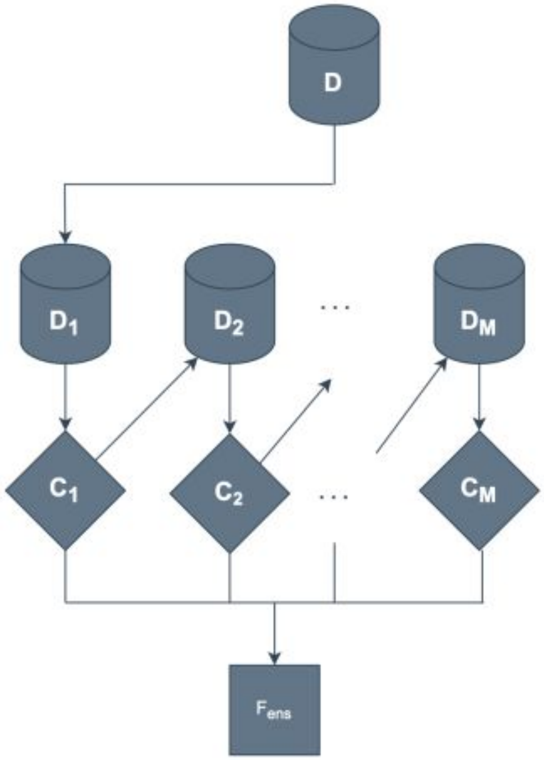


Figure 3: Boosting

Введение
Ансамбли в классическом ML
Deep ensembles
Snapshot ensembles
Fast Geometric ensembles
Dropout ensembles
Заключение

Идея

1. Обучаем много независимых сеток
2. Ансамблируем путем усреднения их предсказаний

Плюсы:

- Очень хорошее качество модели
- Очень хорошая оценка неопределенности

Минусы:

- Требует в K раз больше времени и ресурсов
- Требует в K раз больше памяти

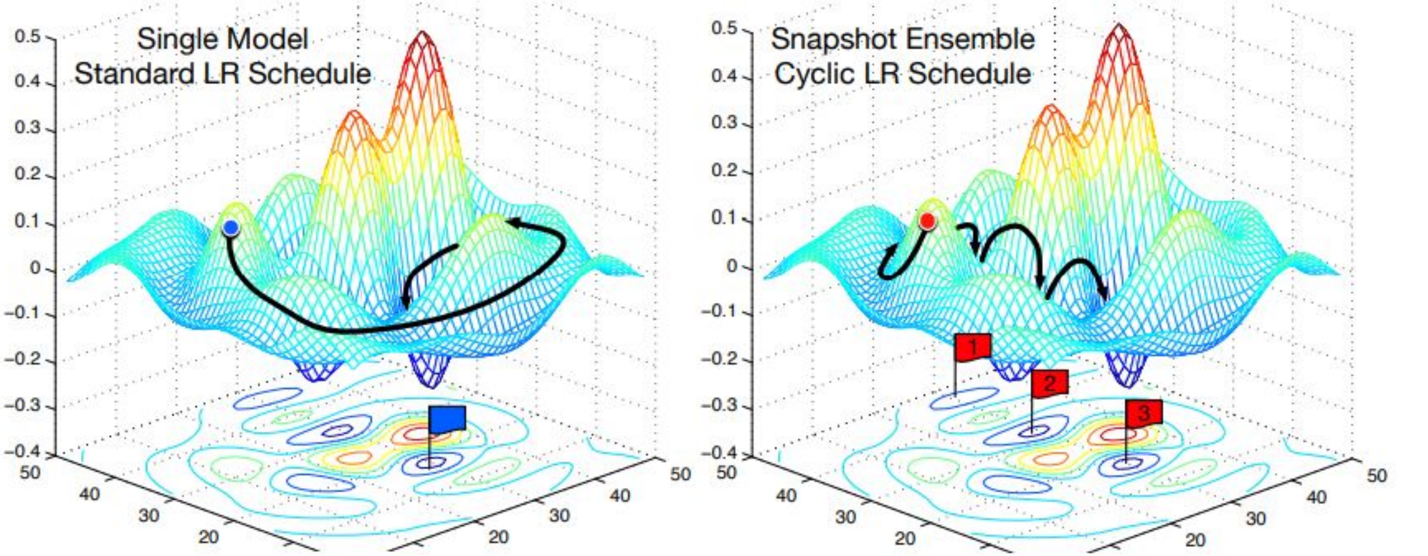
Введение
Ансамбли в классическом ML
Deep ensembles
Snapshot ensembles
Fast Geometric ensembles
Dropout ensembles
Заключение

Мотивация

- Хотим ансамбль, но без дополнительных затрат на ресурсы и время обучения

Идея

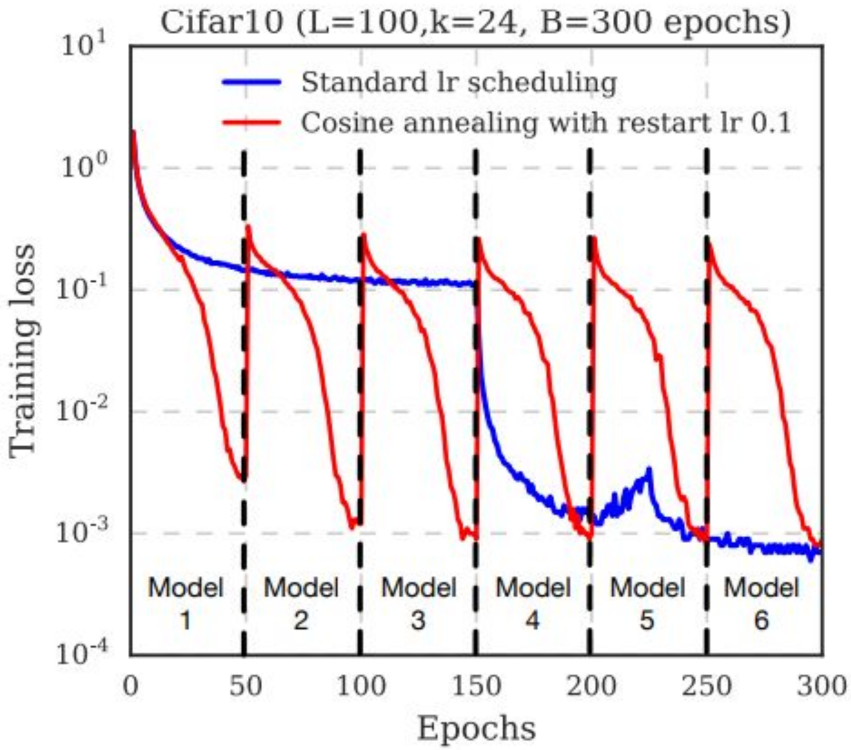
- Делаем несколько снепшотов во время обучения, при этом обучаем, используя циклическое расписание lr
- Активно используем, что во время такого обучения модель может попасть в разные локальные минимумы


Single Model
Standard LR Schedule
Snapshot Ensemble
Cyclic LR Schedule
1
2
3

Learning rate

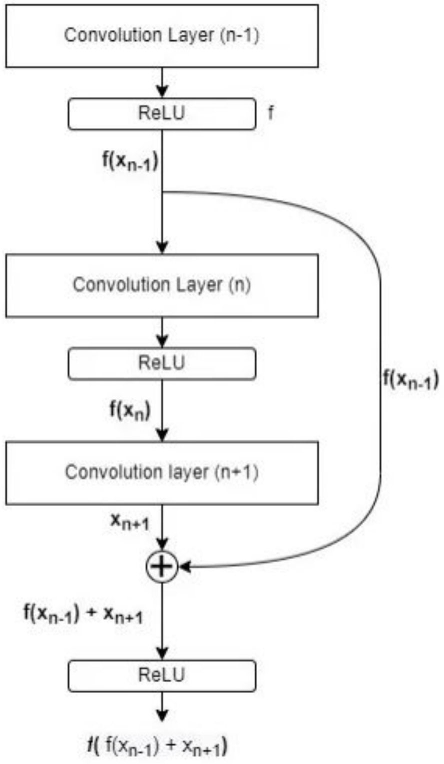
$$\alpha(t) = \frac{\alpha_0}{2}\left(\cos\left(\frac{\pi \operatorname{mod}(t-1, \lceil T/M \rceil)}{\lceil T/M \rceil}\right) + 1\right)$$

t - номер итерации
T - всего итераций
M - количество циклов
a_0 - изначальный lr, гиперпараметр



Ликбез про датасеты и модели

1. Датасеты
    1.1. CIFAR-10
    1.2. CIFAR-100
    1.3. SVHN
    1.4. ImageNet
    1.5. Tiny ImageNet
2. Модели
    2.1. VGG
    2.2. ResNet
    2.3. Wide ResNet
    2.4. DenseNet



	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).

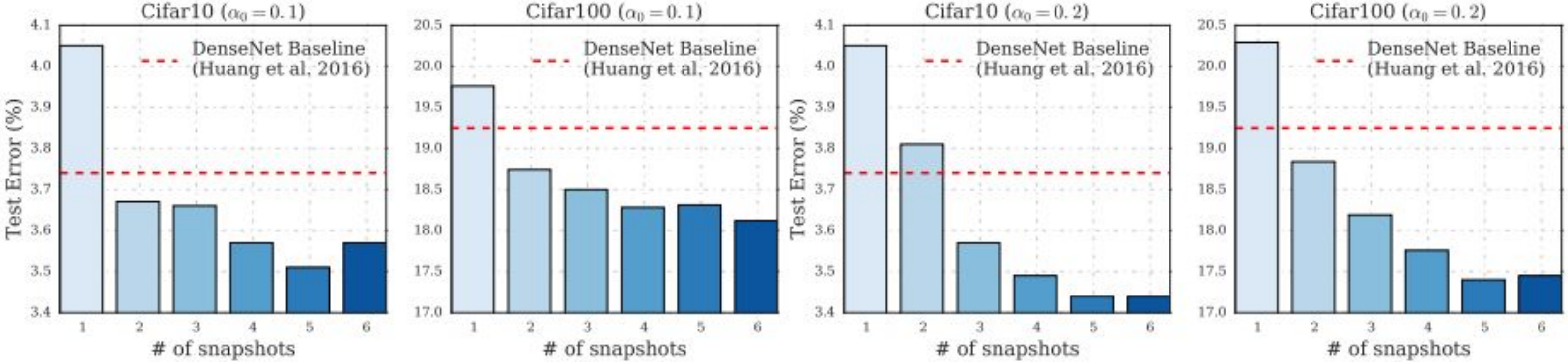
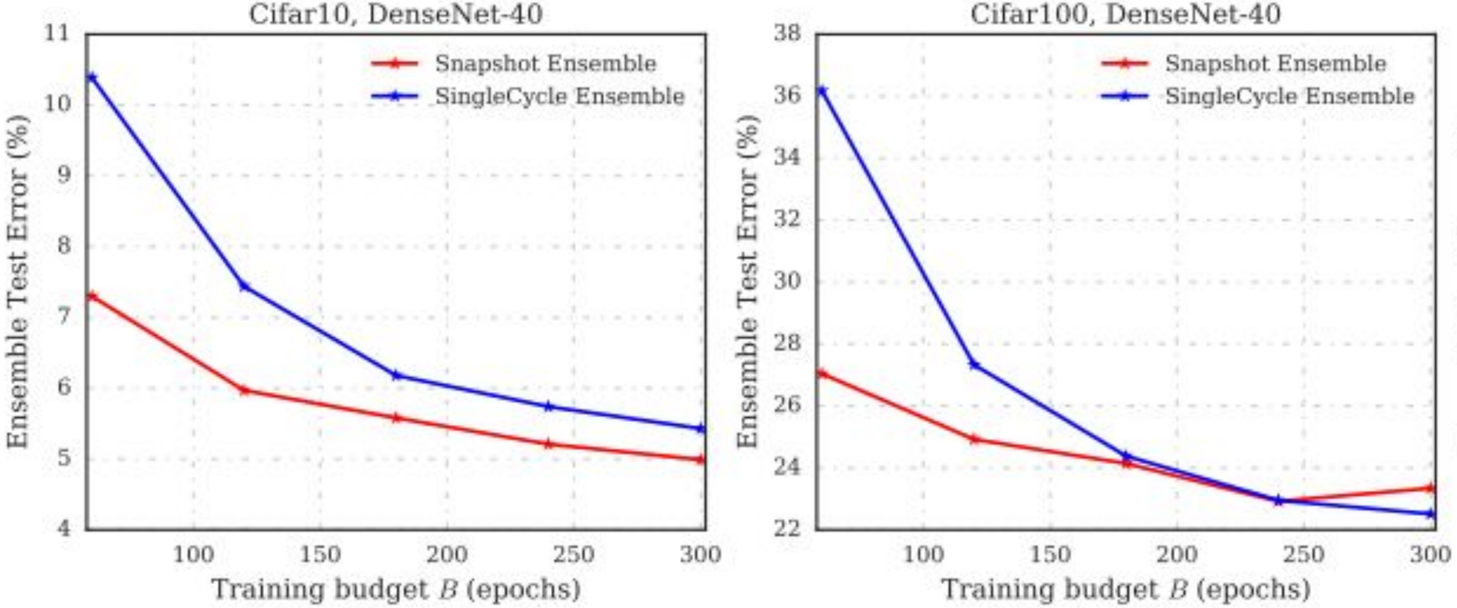


Figure 3: DenseNet-100 Snapshot Ensemble performance on CIFAR-10 and CIFAR-100 with restart learning rate $\alpha_0 = 0.1$ (left two) and $\alpha_0 = 0.2$ (right two). Each ensemble is trained with $M = 6$ annealing cycles (50 epochs per each).



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)

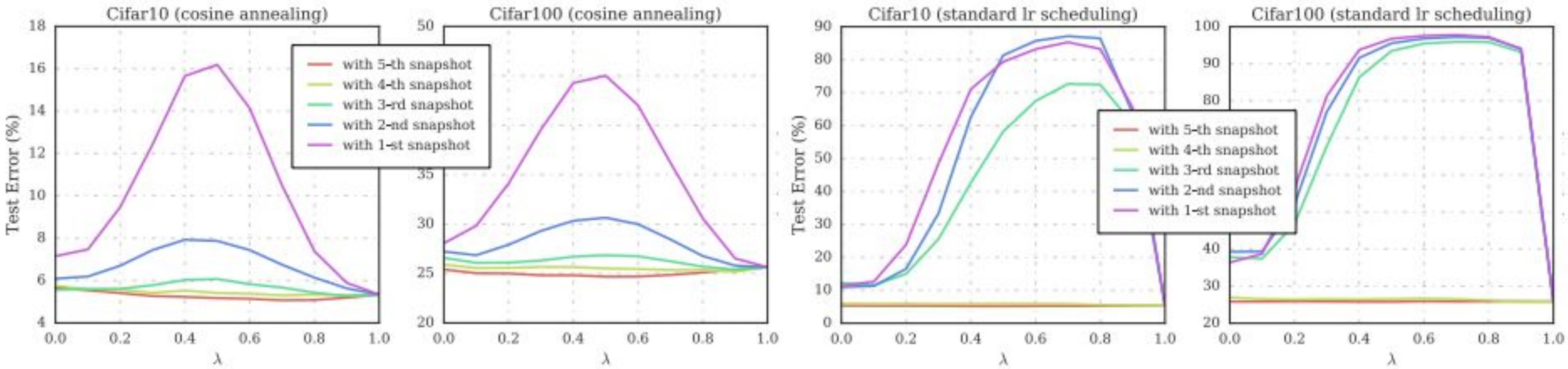


Figure 5: Interpolations in parameter space between the final model (sixth snapshot) and all intermediate snapshots. $\lambda = 0$ represents an intermediate snapshot model, while $\lambda = 1$ represents the final model. **Left:** A Snapshot Ensemble, with cosine annealing cycles ($\alpha_0 = 0.2$ every $B/M = 50$ epochs). **Right:** A NoCycle Snapshot Ensemble, (two learning rate drops, snapshots every 50 epochs).

Введение
Ансамбли в классическом ML
Deep ensembles
Snapshot ensembles
Fast Geometric ensembles
Dropout ensembles
Заключение

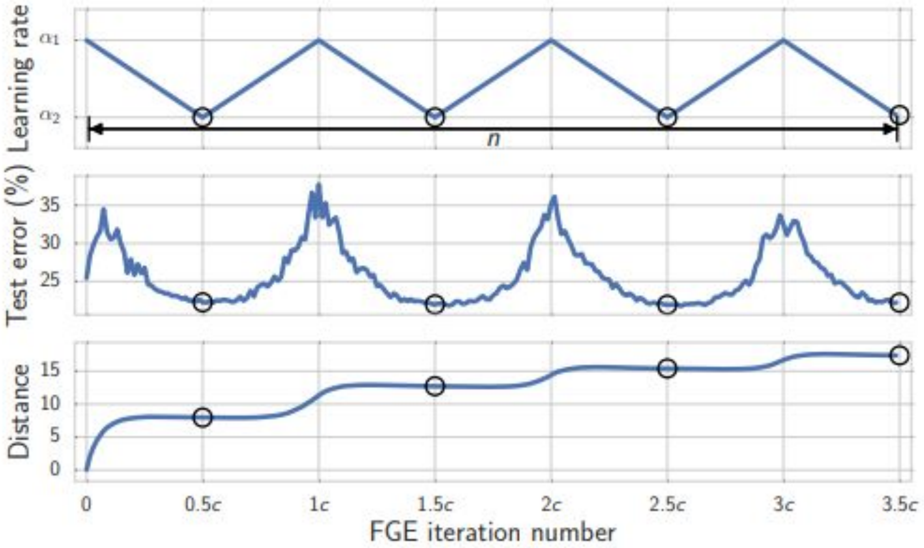
Мотивация

- Между любыми двумя локальными минимумами есть кривая, такая что вдоль нее значение функции потерь слабо изменяется

Идея

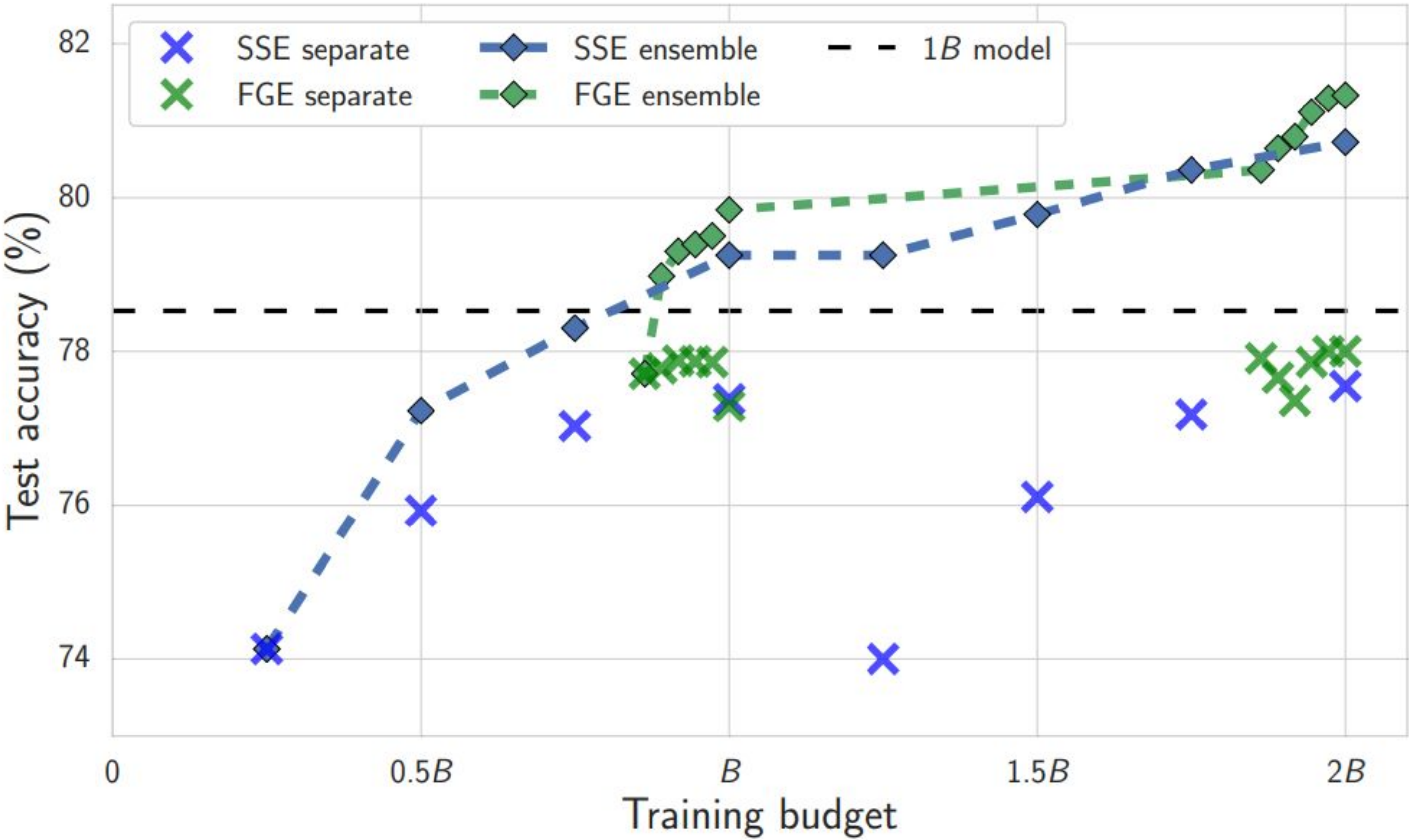
1. Сходимся в какой-то локальный минимум
2. Копируем модель и для копии проделываем еще несколько эпох с триангулярным lr, сохраняя чекпоинты в середине каждой эпохи

Triangular learning rate



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

где t(i) = ⁄ c * (mod(i − 1, c) + 1),
α1 > α2 - learning rates,
c - количество итераций в одном цикле.



SSE separate
FGE separate
SSE ensemble
FGE ensemble
1B model
82
80
78
76
74
Test accuracy (%)
0
0.5B
B
1.5B
2B
Training budget

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**.

DNN (Budget)	method	CIFAR-100 $1B$	CIFAR-100 $2B$	CIFAR-100 $3B$	CIFAR-10 $1B$	CIFAR-10 $2B$	CIFAR-10 $3B$
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

Введение
Ансамбли в классическом ML
Deep ensembles
Snapshot ensembles
Fast Geometric ensembles
Dropout ensembles
Заключение

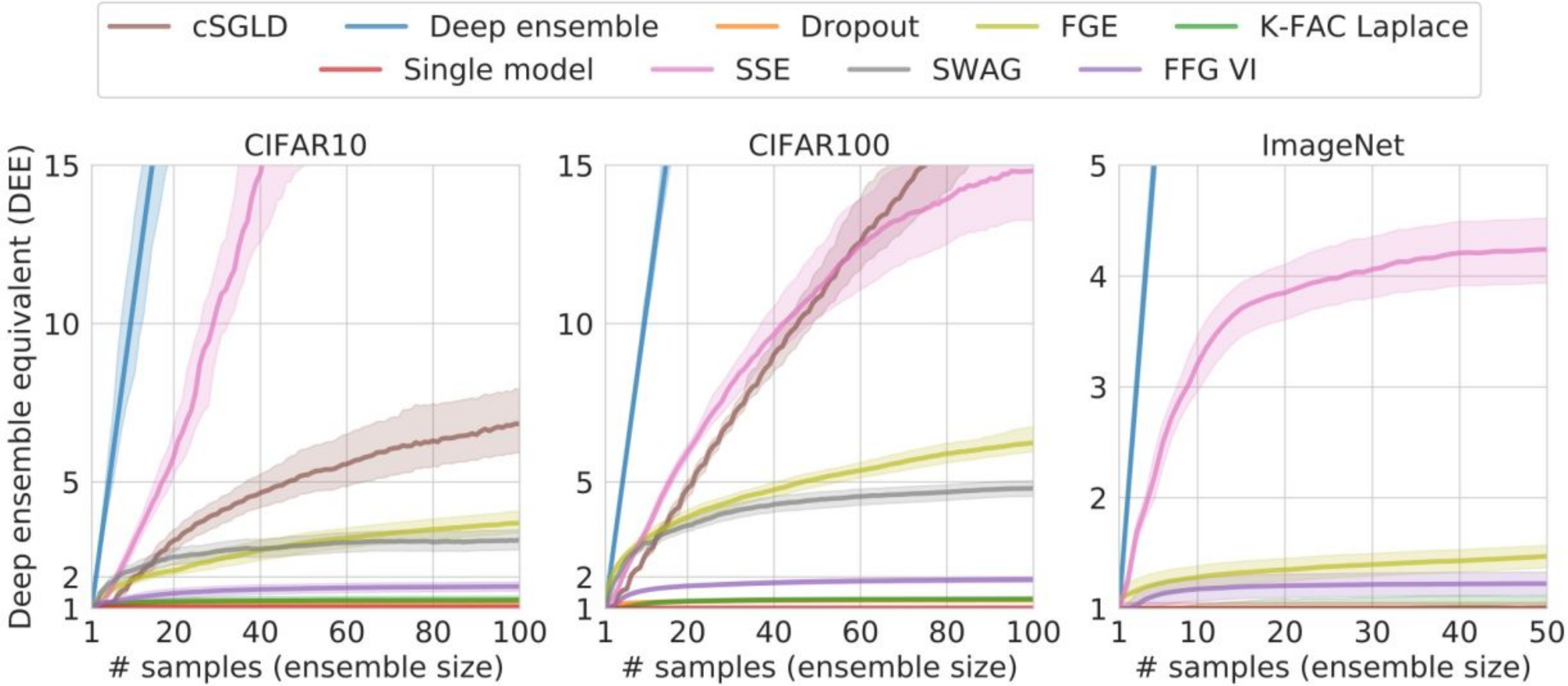
Мотивация

- Хотим ансамбль, но без дополнительных затрат на хранение нескольких нейросетей

Идея

1. Обучаем модель, используя дропаут
2. На этапе тестирования несколько раз запускаем модель с включенным дропаутом и усредняем результаты

Введение
Ансамбли в классическом ML
Deep ensembles
Snapshot ensembles
Fast Geometric ensembles
Dropout ensembles
Заключение



$$\mathrm{DEE}_m(k) = \min\left\{l \in \mathbb{R}, l \geq 1 \,\middle|\, \mathrm{CLL}_{DE}^{\mathrm{mean}}(l) \geq \mathrm{CLL}_m^{\mathrm{mean}}(k)\right\},$$

Model	Method	Error (%) 1	5	10	100	Negative calibrated log-likelihood 1	5	10	100
VGG16 CIFAR-10	Dropout	$5.86_{\pm 0.09}$	$5.81_{\pm 0.08}$	$5.82_{\pm 0.06}$	$5.79_{\pm 0.07}$	$0.232_{\pm 0.005}$	$0.225_{\pm 0.004}$	$0.224_{\pm 0.004}$	$0.223_{\pm 0.003}$
	SWA-Gaussian	$7.03_{\pm 0.50}$	$5.66_{\pm 0.08}$	$5.49_{\pm 0.12}$	$5.25_{\pm 0.13}$	$0.230_{\pm 0.014}$	$0.182_{\pm 0.003}$	$0.171_{\pm 0.002}$	$0.160_{\pm 0.002}$
	Cyclic SGLD	$7.37_{\pm 0.16}$	$6.56_{\pm 0.09}$	$5.71_{\pm 0.06}$	$4.84_{\pm 0.04}$	$0.234_{\pm 0.004}$	$0.196_{\pm 0.004}$	$0.176_{\pm 0.003}$	$0.147_{\pm 0.003}$
	Fast Geometric Ens.	$6.52_{\pm 0.16}$	$5.95_{\pm 0.16}$	$5.69_{\pm 0.16}$	$5.10_{\pm 0.13}$	$0.213_{\pm 0.005}$	$0.187_{\pm 0.003}$	$0.178_{\pm 0.003}$	$0.155_{\pm 0.004}$
	Deep Ensembles	$5.95_{\pm 0.14}$	$4.79_{\pm 0.11}$	$4.57_{\pm 0.07}$	$4.39_{\pm \text{NA}}$	$0.226_{\pm 0.001}$	$0.158_{\pm 0.002}$	$0.148_{\pm 0.001}$	$0.134_{\pm \text{NA}}$
	Single model	$5.83_{\pm 0.11}$	$5.83_{\pm 0.11}$	$5.83_{\pm 0.11}$	$5.83_{\pm 0.11}$	$0.223_{\pm 0.002}$	$0.223_{\pm 0.002}$	$0.223_{\pm 0.002}$	$0.223_{\pm 0.002}$
	Variational Inf. (FFG)	$6.57_{\pm 0.09}$	$5.63_{\pm 0.13}$	$5.50_{\pm 0.10}$	$5.46_{\pm 0.03}$	$0.239_{\pm 0.002}$	$0.192_{\pm 0.002}$	$0.184_{\pm 0.002}$	$0.175_{\pm 0.001}$
	KFAC-Laplace	$6.00_{\pm 0.13}$	$5.82_{\pm 0.12}$	$5.82_{\pm 0.19}$	$5.80_{\pm 0.19}$	$0.210_{\pm 0.005}$	$0.203_{\pm 0.007}$	$0.201_{\pm 0.007}$	$0.200_{\pm 0.008}$
	Snapshot Ensembles	$7.76_{\pm 0.22}$	$5.52_{\pm 0.13}$	$5.00_{\pm 0.10}$	$4.54_{\pm 0.05}$	$0.247_{\pm 0.005}$	$0.176_{\pm 0.001}$	$0.160_{\pm 0.001}$	$0.137_{\pm 0.001}$
WideResNet CIFAR-10	Dropout	$3.88_{\pm 0.12}$	$3.70_{\pm 0.18}$	$3.63_{\pm 0.19}$	$3.64_{\pm 0.17}$	$0.130_{\pm 0.002}$	$0.120_{\pm 0.002}$	$0.119_{\pm 0.001}$	$0.117_{\pm 0.002}$
	SWA-Gaussian	$4.98_{\pm 1.17}$	$3.53_{\pm 0.09}$	$3.34_{\pm 0.14}$	$3.28_{\pm 0.10}$	$0.157_{\pm 0.036}$	$0.111_{\pm 0.004}$	$0.105_{\pm 0.003}$	$0.101_{\pm 0.002}$
	Cyclic SGLD	$4.78_{\pm 0.16}$	$4.09_{\pm 0.11}$	$3.63_{\pm 0.13}$	$3.19_{\pm 0.04}$	$0.155_{\pm 0.003}$	$0.128_{\pm 0.002}$	$0.114_{\pm 0.001}$	$0.099_{\pm 0.002}$
	Fast Geometric Ens.	$4.86_{\pm 0.17}$	$3.95_{\pm 0.07}$	$3.77_{\pm 0.10}$	$3.34_{\pm 0.06}$	$0.148_{\pm 0.003}$	$0.120_{\pm 0.002}$	$0.113_{\pm 0.002}$	$0.102_{\pm 0.001}$
	Deep Ensembles	$3.65_{\pm 0.02}$	$3.11_{\pm 0.10}$	$3.01_{\pm 0.06}$	$2.83_{\pm \text{NA}}$	$0.123_{\pm 0.002}$	$0.097_{\pm 0.001}$	$0.095_{\pm 0.001}$	$0.090_{\pm \text{NA}}$
	Single model	$3.70_{\pm 0.15}$	$3.70_{\pm 0.15}$	$3.70_{\pm 0.15}$	$3.70_{\pm 0.15}$	$0.124_{\pm 0.005}$	$0.124_{\pm 0.005}$	$0.125_{\pm 0.005}$	$0.124_{\pm 0.005}$
	Variational Inf. (FFG)	$5.61_{\pm 0.04}$	$4.15_{\pm 0.15}$	$3.94_{\pm 0.10}$	$3.64_{\pm 0.07}$	$0.189_{\pm 0.002}$	$0.134_{\pm 0.002}$	$0.127_{\pm 0.002}$	$0.117_{\pm 0.001}$
	KFAC-Laplace	$4.03_{\pm 0.19}$	$3.90_{\pm 0.15}$	$3.88_{\pm 0.22}$	$3.83_{\pm 0.16}$	$0.134_{\pm 0.004}$	$0.124_{\pm 0.004}$	$0.122_{\pm 0.005}$	$0.120_{\pm 0.003}$
	Snapshot Ensembles	$5.56_{\pm 0.15}$	$3.68_{\pm 0.09}$	$3.33_{\pm 0.10}$	$2.89_{\pm 0.07}$	$0.179_{\pm 0.005}$	$0.119_{\pm 0.001}$	$0.105_{\pm 0.001}$	$0.090_{\pm 0.001}$

Table 3: Classification error and negative calibrated log-likelihood for different models and numbers of samples on CIFAR-10/100.

Model	Method	Error (%)				Negative calibrated log-likelihood			
		1	5	10	100	1	5	10	100
VGG16 CIFAR-100	Dropout	26.10±0.20	25.68±0.18	25.66±0.14	25.60±0.17	1.176±0.008	1.111±0.008	1.098±0.009	1.084±0.009
	SWA-Gaussian	27.74±1.87	24.53±0.09	23.64±0.28	22.97±0.20	1.109±0.073	0.931±0.007	0.879±0.007	0.826±0.005
	Cyclic SGLD	29.75±0.17	26.79±0.19	24.14±0.11	21.15±0.11	1.114±0.003	0.976±0.004	0.881±0.006	0.749±0.004
	Fast Geometric Ens.	27.07±0.24	25.35±0.29	24.68±0.40	22.78±0.22	1.057±0.010	0.965±0.003	0.930±0.003	0.827±0.004
	Deep Ensembles	25.72±0.17	21.60±0.13	20.79±0.16	19.88±NA	1.092±0.004	0.840±0.005	0.794±0.002	0.723±NA
	Single model	25.44±0.29	25.44±0.29	25.44±0.29	25.44±0.29	1.087±0.006	1.087±0.006	1.087±0.006	1.087±0.006
	Variational Inf. (FFG)	27.24±0.09	25.24±0.11	24.85±0.05	24.56±0.07	1.154±0.004	1.001±0.002	0.973±0.002	0.939±0.001
	KFAC-Laplace	27.11±0.59	25.98±0.21	25.84±0.38	25.70±0.38	1.174±0.037	1.089±0.007	1.069±0.005	1.050±0.008
	Snapshot Ensembles	31.19±0.33	23.87±0.18	22.31±0.31	21.03±0.10	1.170±0.012	0.899±0.004	0.834±0.005	0.751±0.003
WideResNet CIFAR-100	Dropout	20.19±0.11	19.41±0.17	19.36±0.12	19.22±0.15	0.823±0.008	0.768±0.005	0.760±0.006	0.751±0.005
	SWA-Gaussian	20.45±0.73	17.57±0.17	17.21±0.22	17.08±0.19	0.794±0.025	0.653±0.004	0.634±0.005	0.614±0.005
	Cyclic SGLD	21.42±0.32	19.42±0.28	17.88±0.16	16.29±0.10	0.813±0.010	0.713±0.009	0.654±0.005	0.583±0.004
	Fast Geometric Ens.	21.48±0.31	18.54±0.16	18.00±0.19	17.12±0.16	0.770±0.007	0.652±0.006	0.630±0.006	0.596±0.003
	Deep Ensembles	19.38±0.20	16.55±0.08	16.17±0.15	15.77±NA	0.797±0.007	0.623±0.003	0.595±0.003	0.571±NA
	Single model	19.31±0.24	19.31±0.24	19.31±0.24	19.31±0.24	0.797±0.010	0.797±0.010	0.797±0.010	0.797±0.010
	Variational Inf. (FFG)	24.38±0.27	20.17±0.15	19.28±0.09	18.74±0.08	1.004±0.011	0.767±0.004	0.727±0.003	0.685±0.002
	KFAC-Laplace	20.02±0.18	19.76±0.15	19.53±0.19	19.43±0.21	0.834±0.009	0.803±0.006	0.795±0.007	0.789±0.006
	Snapshot Ensembles	23.01±0.26	18.20±0.13	17.12±0.31	16.07±0.07	0.859±0.009	0.678±0.006	0.633±0.008	0.582±0.004

Table 3: Classification error and negative calibrated log-likelihood for different models and numbers of samples on CIFAR-10/100.

Выводы

- Ансамбли улучшают качество
- Ансамбли дают возможность оценить неопределенность ответа модели
- Ансамблирование активно развивается прямо сейчас