Ансамблирование нейронных сетей

Бекян Артём, 201

Содержание

1) Введение
2) Что такое ансамбли
3) Зачем нужны ансамбли
4) Где используют ансамбли
5) Какие бывают ансамбли

Введение

Недостатки нейронных сетей

- Обучение нейронных сетей долгое
- Нет гарантии хорошего результата
- Высокая дисперсия

Что такое ансамбли

Что такое ансамбли: принцип работы

Очевидный вариант:

1) Обучим несколько моделей
2) Возьмем среднее от их предсказаний

Идея: объединение предсказаний нескольких моделей даст более хороший результат, чем предсказание одной модели

Что такое ансамбли: теорема Кондорсе

Формулировка:

Если каждый член жюри присяжных имеет независимое мнение, и если вероятность правильного решения члена жюри больше 0.5, то тогда вероятность правильного решения присяжных в целом возрастает с увеличением количества членов жюри, и стремится к единице.
Если же вероятность быть правым у каждого из членов жюри меньше 0.5, то вероятность принятия правильного решения присяжными в целом монотонно уменьшается и стремится к нулю с увеличением количества присяжных.

Что такое ансамбли: теорема Кондорсе

Доказательство:

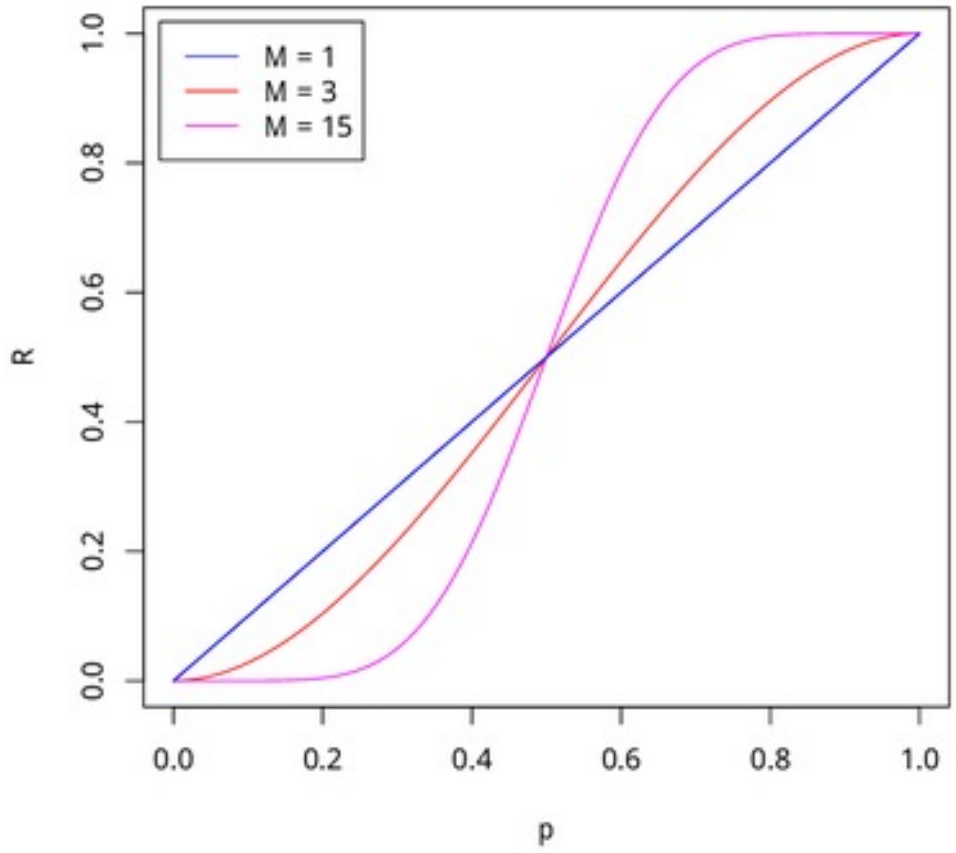
Пусть M – количество присяжных

p – вероятность правильного решения одного присяжного

R – вероятность правильного решения всех присяжных

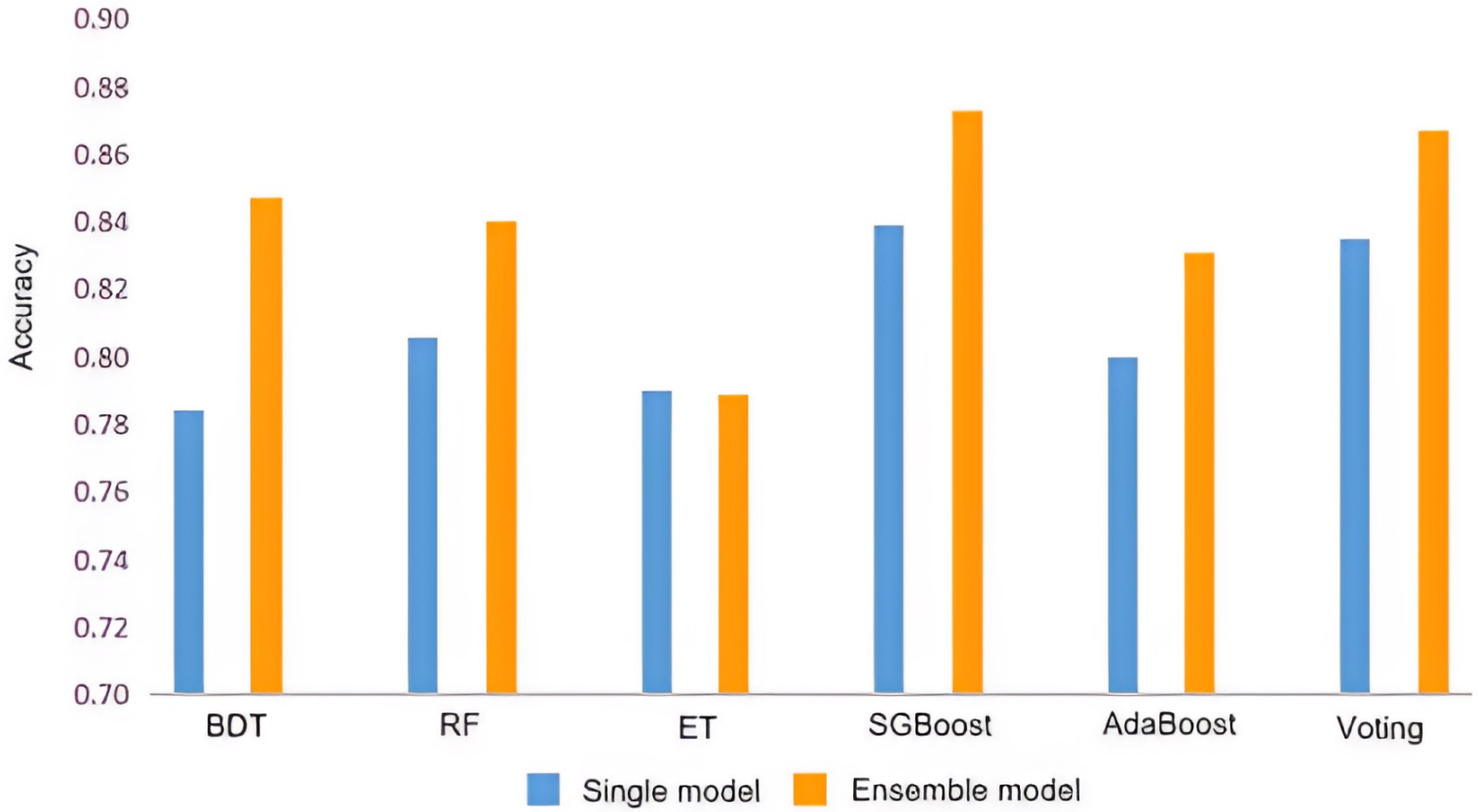
$m = \left\lfloor \frac{M}{2} \right\rfloor + 1$ – минимальное большинство членов жюри

Тогда $R = \sum_{i=m}^{M} C_M^i p^i (1-p)^{M-i}$



Зачем нужны ансамбли

Зачем нужны ансамбли: улучшение точности предсказаний



Accuracy comparison between single model and ensemble model. Where BDT, RF and ET represent Bagged Decision Tree, Random Forest and Extra Tree respectively

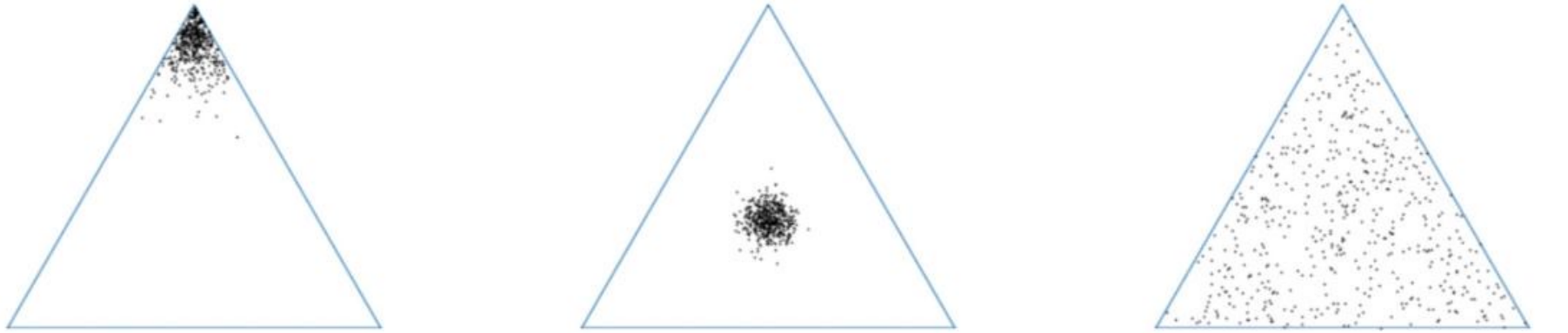
Зачем нужны ансамбли: подсчет неопределенности

Неопределенность – степень неуверенности модели в своем предсказании

- **Неопределенность модели (эпистемическая)** – невозможность определить лучшую модель
- **Неопределенность данных (алеаторическая)** – присутствие шума в данных

Зачем нужны ансамбли: подсчет неопределенности

Ensemble $\{\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(m)})\}_{m=1}^{M}$ can be visualized on a simplex



(a) Confident (b) Data Uncertainty (c) Knowledge Uncertainty

Где используют ансамбли

Где используют ансамбли

- Kaggle
- MNIST database of handwritten digits
- Защита от DoS-атак
- Обнаружение вредоносных программ
- …
- И многое другое

Какие бывают ансамбли

Общие принципы

Какие бывают ансамбли: общий принципы. Размер

- Универсальной верной формулы не существует
- Размер зависит от метода сборки ансамбля, использованных алгоритмов и т. д.
- Недавние исследования показывают, что может существовать оптимальный размер такой, что ансамбли с меньшим или большим размером имеют меньшую точность. Эти оптимальные значения близки к количеству признаков в выборке (Bonab, Can, 2018)

Какие бывают ансамбли

Базовые алгоритмы

Какие бывают ансамбли: базовые алгоритмы. Бэггинг

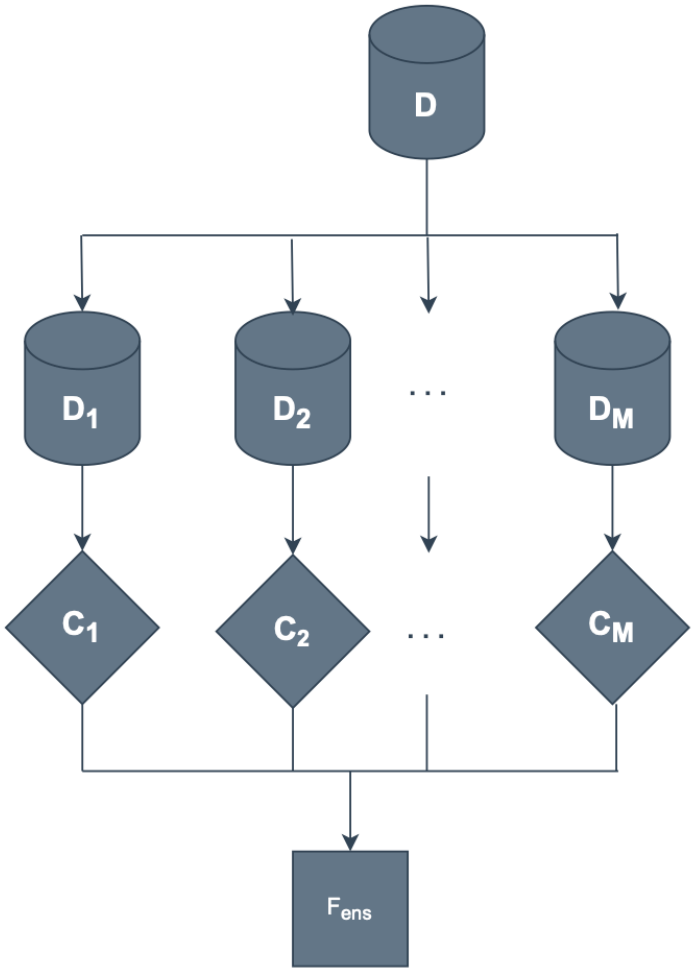


Figure 2: Bagging

Какие бывают ансамбли: базовые алгоритмы. Бустинг

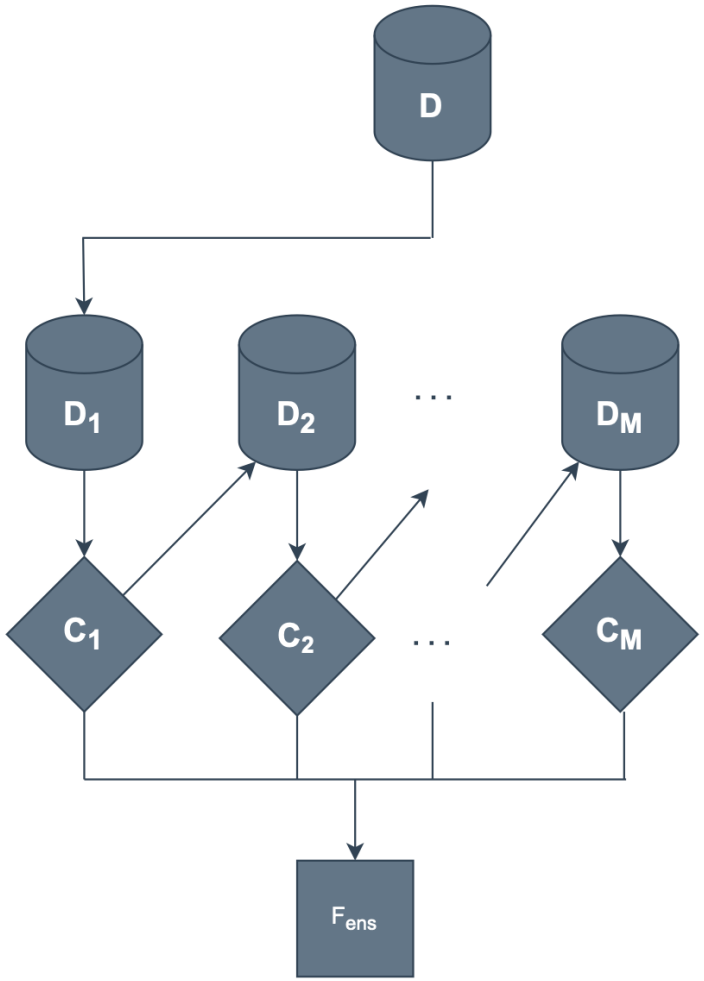


Figure 3: Boosting

Какие бывают ансамбли: базовые алгоритмы. Стэкинг

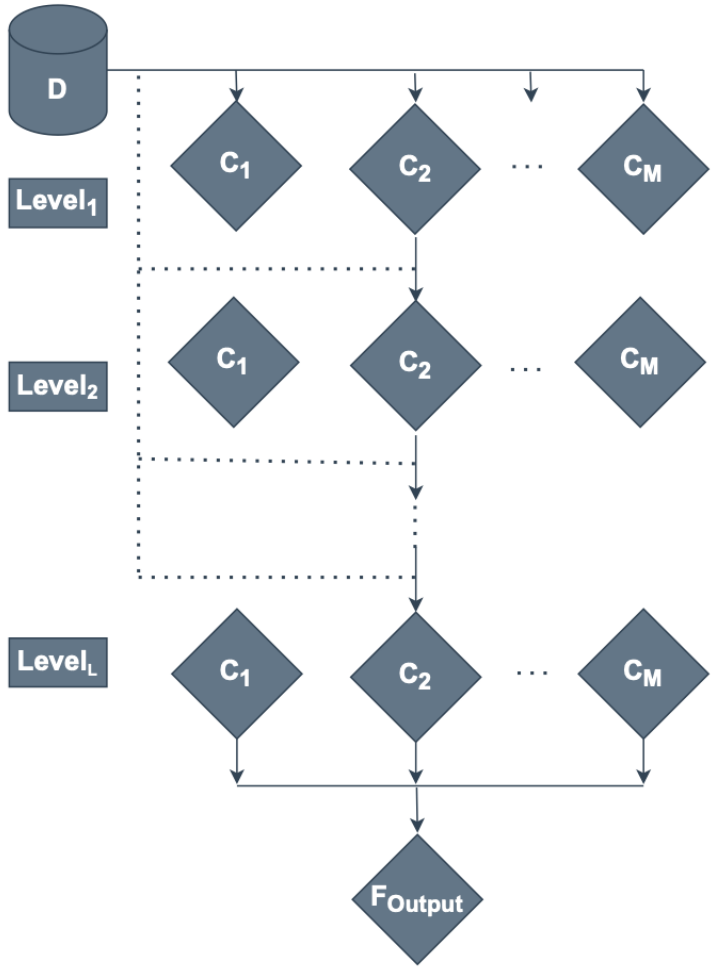


Figure 4: Stacking

Какие бывают ансамбли

Алгоритмы нейросетей

Какие бывают ансамбли: алгоритмы нейросетей. Deep ансамбли

Принцип работы:

1) Обучаем несколько независимых нейронных сетей на всем наборе данных
2) Для классификации считаем вероятность принадлежности к каждому из классов, для регрессии – мат. ожидание и квадрат дисперсии
3) Результат выбираем как среднее арифметическое предсказаний

Какие бывают ансамбли: алгоритмы нейросетей. Deep ансамбли

Преимущества:

- Лучшее соотношение размер ансамбля / качество на данный момент
- Одно из лучших значений точности

Недостатки:

- Дорогое обучение

Какие бывают ансамбли: алгоритмы нейросетей. Snapshot ансамбли

Принцип работы:

1) Обучаем одну нейронную сеть
2) В момент обучения с помощью SGD приходим в локальный минимум функции метрики качества на тренировочной выборке
3) Сохраняем модель и вручную ставим в этой точке большое значение функции метрики качества
4) Повторяем пункты 2 – 3 пока не получим нужное количество моделей
5) Результат выбираем через среднее арифметическое

Какие бывают ансамбли: алгоритмы нейросетей. Snapshot ансамбли

Преимущества:

- Сильное снижение затрат на обучения ансамбля
- Значительное улучшение точности предсказаний
- Одно из лучших соотношений размер ансамбля / качество

Недостатки:

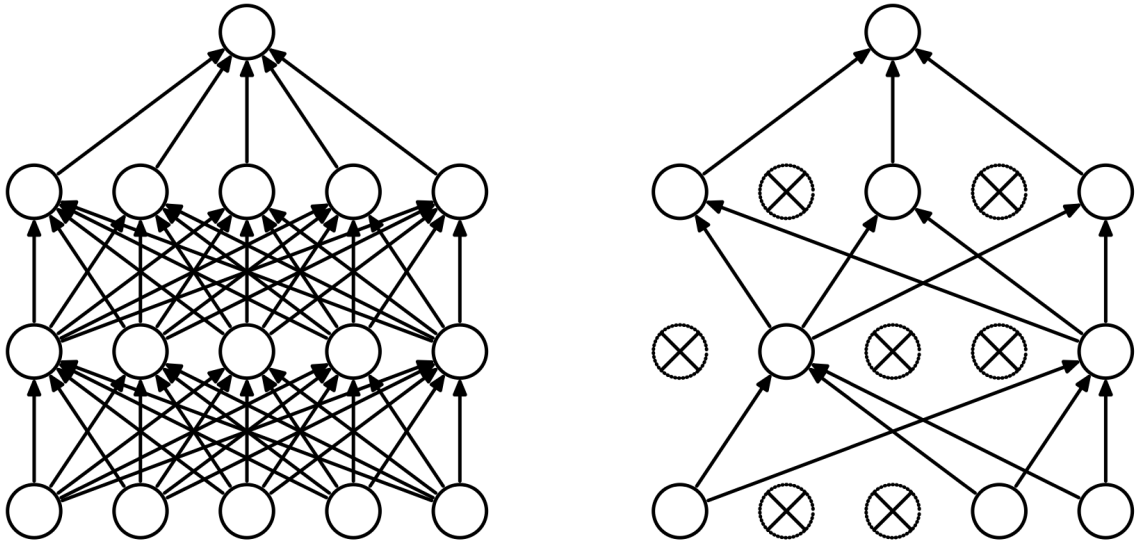
- Зачастую проигрывает deep ансамблям в точности

Какие бывают ансамбли: алгоритмы нейросетей. Dropout ансамбли

Принцип работы:

1) Каждому нейрону присваивается p – вероятность того, что этот нейрон не выкинется
2) Проходим n раз, каждый раз каждый нейрон оставляем с вероятностью p или выкидываем с вероятностью 1 - p
3) На тестовой выборке не выкидываем нейроны, а просто умножаем выход на p
4) Результатом является выход нейросети

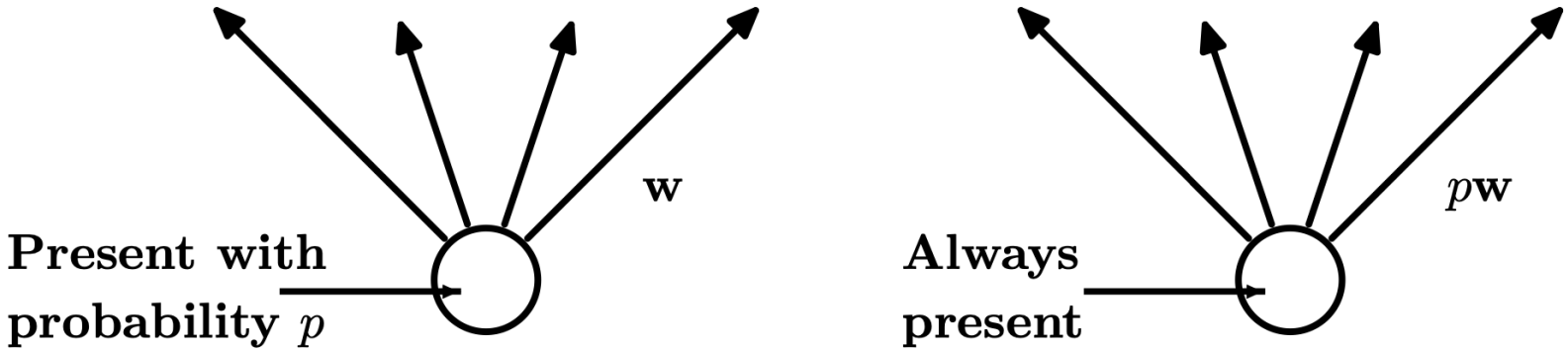
Какие бывают ансамбли: алгоритмы нейросетей. Dropout ансамбли



(a) Standard Neural Net (b) After applying dropout.

Figure 1: Dropout Neural Net Model. **Left**: A standard neural net with 2 hidden layers. **Right**: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped.

Какие бывают ансамбли: алгоритмы нейросетей. Dropout ансамбли



(a) At training time (b) At test time

Figure 2: **Left**: A unit at training time that is present with probability p and is connected to units in the next layer with weights $\mathbf{w}$. **Right**: At test time, the unit is always present and the weights are multiplied by p. The output at test time is same as the expected output at training time.

Какие бывают ансамбли: алгоритмы нейросетей. Dropout ансамбли

Преимущества:

- Скорость предсказаний
- По факту dropout – метод регуляризации, который может применяться в любых ансамблях

Недостатки:

- Увеличение размера ансамбля дает крайне малый прирост точности в сравнении с другими ансамблями

Статистика

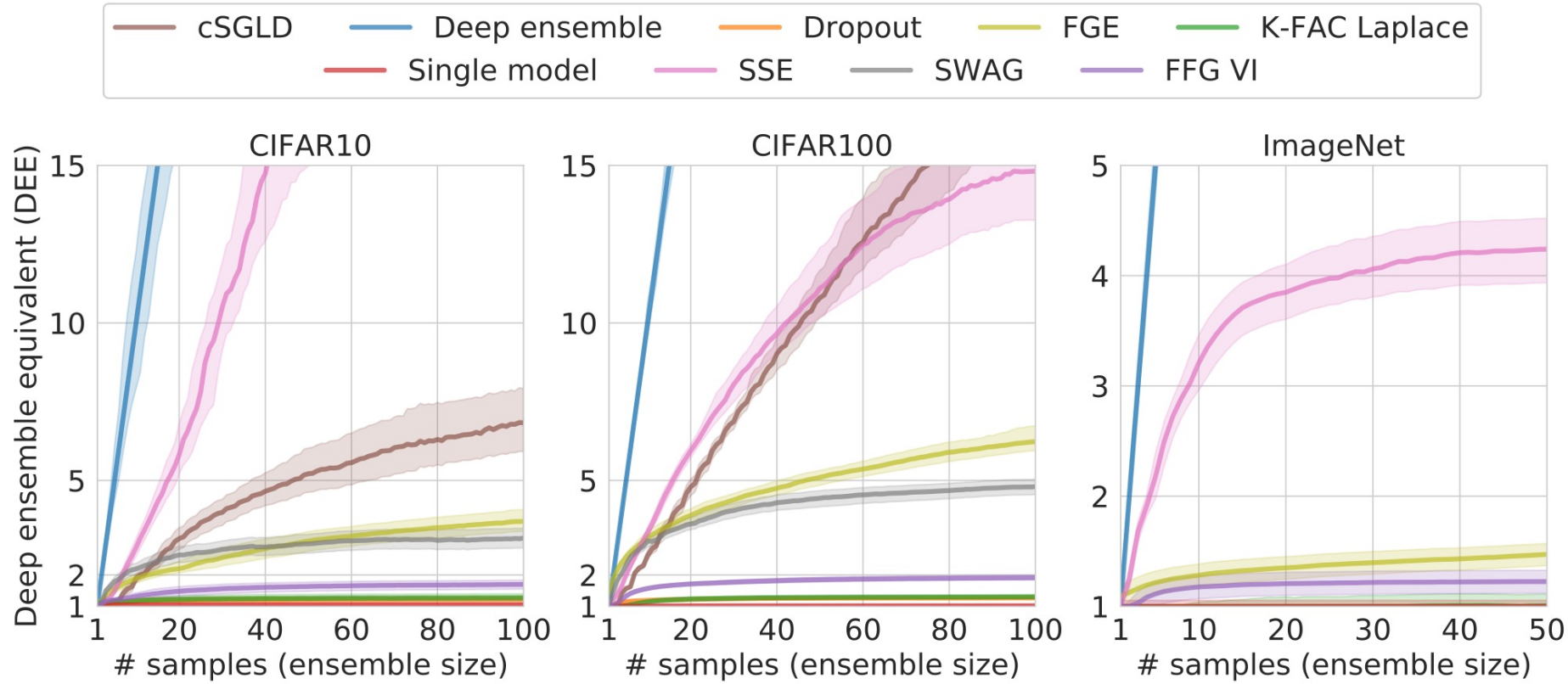
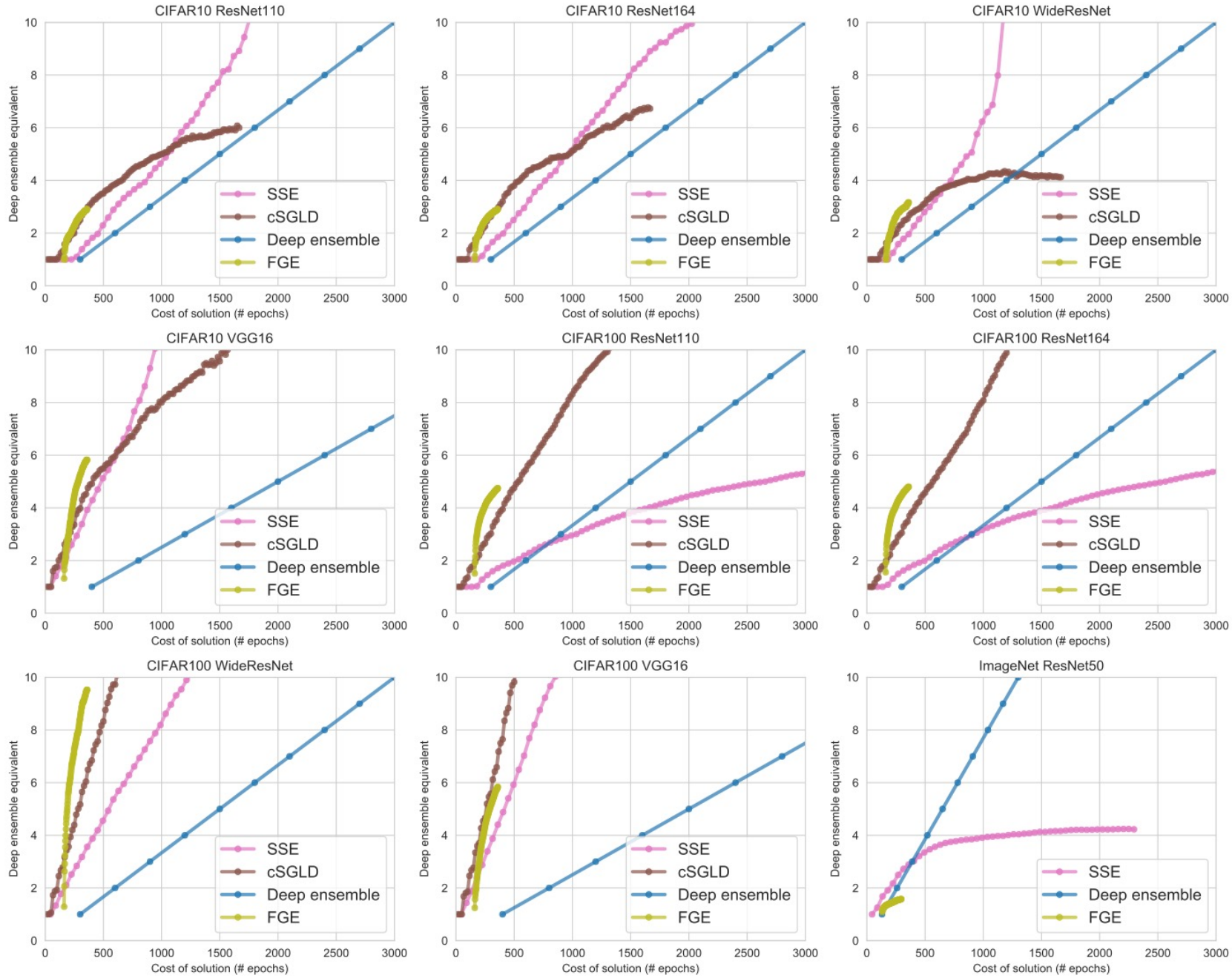


Figure 3: The deep ensemble equivalent score (DEE) for different numbers of samples on CIFAR-10, CIFAR-100, and ImageNet datasets averaged across different deep convolutional architectures. A deep ensemble equivalent score (DEE) of a model is equal to the minimum size of a deep ensemble (an ensemble of independently train networks) that achieves the same performance as the model under consideration. The score is measured in the number of models (higher is better). The area between average lower and upper bounds of DEE is shaded. **The plot demonstrates that all of the ensembling techniques are far less efficient than deep ensembles during inference and fail to produce the same level of performance as deep ensembles.** The comparison that is normalized on training time is presented in Appendix A.


CIFAR10 ResNet110
CIFAR10 ResNet164
CIFAR10 WideResNet
CIFAR10 VGG16
CIFAR100 ResNet110
CIFAR100 ResNet164
CIFAR100 WideResNet
CIFAR100 VGG16
ImageNet ResNet50
Deep ensemble equivalent
Cost of solution (# epochs)
SSE
cSGLD
Deep ensemble
FGE

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

Заключение

- Ансамбли зачастую долго обучаются
- Точность предсказаний ансамблей заметно выше, чем одиночных нейросетей
- С помощью ансамблей можно оценить неопределенность
- Тема ансамблирования нейронных сетей активно развивается