

Алихан Зиманов

Факультет компьютерных наук  
НИУ ВШЭ



## Towards Understanding Ensembles, Knowledge Distillation and Self-Distillation in Deep Learning

Research of Previous Works

НИС, Москва, 2024

# Содержание

1 Ансамблирование

2 Дистилляция

# Введение

## Вопрос

Каков глобальный смысл ансамблирования?

# Введение

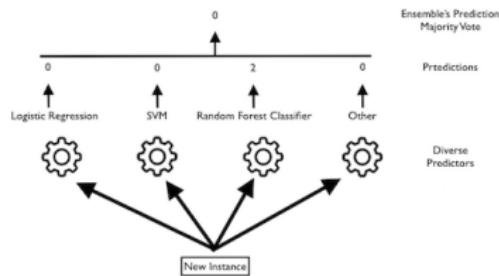
## Вопрос

Каков глобальный смысл ансамблирования?

## Ответ<sup>a</sup>

<sup>a</sup>Мнение экспертов

Много моделей помогают друг другу чтобы решить большую задачу.



## ВНИМАНИЕ

Главный вопрос: почему ансамблирование улучшает результаты?



# История ансамблирования

## Подходы ансамблирования

1. Bootstrap Aggregating (Bagging) [1] — порождает множество предикторов на случайных подвыборках данных и усредняет (или берет голос большинства) предсказания.
2. Boosting [2] — AdaBoost добавляет легкие модели по одной для компенсирования слабостей предыдущих моделей.
3. Random Forests [3] — классика.
4. Gradient Boosting Machine [4] — классика.
5. Stacking [5] — мета-модель обучается на выходах базовых моделей.

## В поисках ответов

Исследуем возможные объяснения эффективности ансамблей.



## Bias-Variance Decomposition

Пусть

$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ ,  $y = f(x) + \varepsilon$  и  $\hat{f}(x; D)$  — наша модель.

Тогда

$$\mathbb{E}_{D, \varepsilon} [(y - \hat{f}(x; D))^2] = (\text{Bias}_D [\hat{f}(x; D)])^2 + \text{Var}_D [\hat{f}(x; D)] + \sigma^2$$

где

$$\text{Bias}_D [\hat{f}(x; D)] = \mathbb{E}_D [\hat{f}(x; D) - f(x)] = \mathbb{E}_D [\hat{f}(x; D)] - \mathbb{E}_{y|x} [y(x)]$$

$$\text{Var}_D [\hat{f}(x; D)] = \mathbb{E}_D \left[ (\mathbb{E}_D [\hat{f}(x; D)] - \hat{f}(x; D))^2 \right]$$

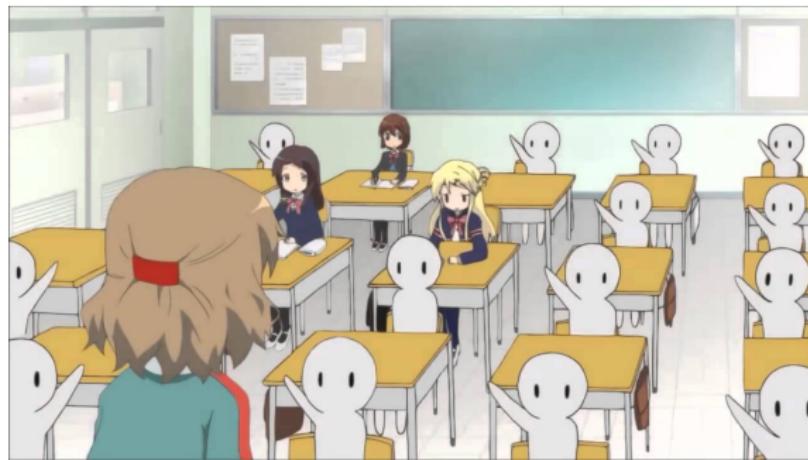
$$\sigma^2 = \mathbb{E}_y [(y - f(x))^2]$$

Ансамблирование (беггинг) уменьшает Var без увеличения Bias.

# The Condorcet Jury Theorem

## Формулировка

Пусть есть независимые голосующие, у каждого из которых одна и та же вероятность правильно проголосовать строго больше  $1/2$ . Тогда вероятность решения большинства за правильный выбор будет увеличиваться с увеличением количества голосующих (и стремиться к 1 в пределе).



# Neural Tangent Kernel

## Определение

$f : \mathbb{R}^{D+d} \rightarrow \mathbb{R}$  — нейронная сеть со входами  $x \in \mathbb{R}^d$  и весами  $W \in \mathbb{R}^D$ .

Тогда  $f$  можно иногда аппроксимировать как

$$f(W, x) \approx f(W_0, x) + \langle W - W_0, \nabla_W f(W_0, x) \rangle$$

где  $W_0$  — случайная инициализация и  $\Phi_{W_0}(x) := \nabla_W f(W_0, x)$  — neural tangent kernel.

# Neural Tangent Kernel

## Определение

$f : \mathbb{R}^{D+d} \rightarrow \mathbb{R}$  — нейронная сеть со входами  $x \in \mathbb{R}^d$  и весами  $W \in \mathbb{R}^D$ .

Тогда  $f$  можно иногда аппроксимировать как

$$f(W, x) \approx f(W_0, x) + \langle W - W_0, \nabla_W f(W_0, x) \rangle$$

где  $W_0$  — случайная инициализация и  $\Phi_{W_0}(x) := \nabla_W f(W_0, x)$  — neural tangent kernel.

## Результат

Обучение нейронной сети  $f$  аппроксимируется обучением линейной функции  $\Phi_{W_0}(x)$  (kernel trick).

## Neural Tangent Kernel

### Определение

$f : \mathbb{R}^{D+d} \rightarrow \mathbb{R}$  — нейронная сеть со входами  $x \in \mathbb{R}^d$  и весами  $W \in \mathbb{R}^D$ .

Тогда  $f$  можно иногда аппроксимировать как

$$f(W, x) \approx f(W_0, x) + \langle W - W_0, \nabla_W f(W_0, x) \rangle$$

где  $W_0$  — случайная инициализация и  $\Phi_{W_0}(x) := \nabla_W f(W_0, x)$  — neural tangent kernel.

### Результат

Обучение нейронной сети  $f$  аппроксимируется обучением линейной функции  $\Phi_{W_0}(x)$  (kernel trick).

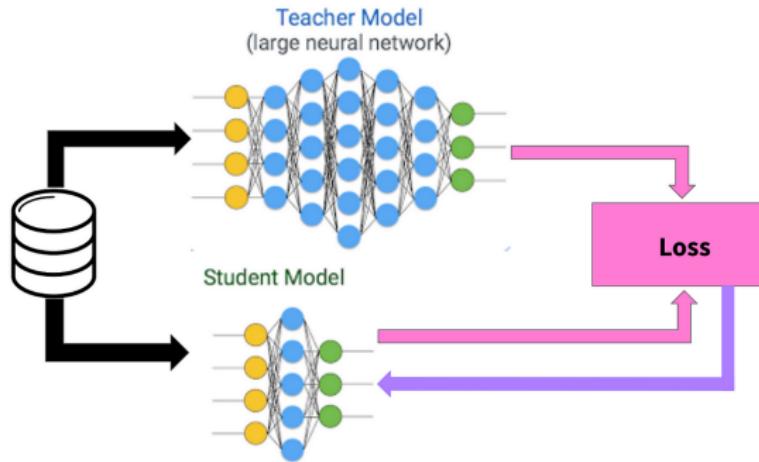
### Зачем?

Эта двойственность позволяет использовать простые уравнения в замкнутой форме, описывающие динамику обучения, обобщение и предсказания толстых нейронных сетей.

# NTK in Ensembles

# Knowledge Distillation

- Response-based knowledge
- Feature-based knowledge
- Relation-based knowledge



## Практическая эффективность дистилляции [6]



Рис. 1: Обычное обучение

Рис. 2: Дистилляция

## Практическая эффективность дистилляции [6]



Рис. 1: Обычное обучение



Рис. 2: Дистилляция

## Практическая эффективность дистилляции [6]



Рис. 1: Обычное обучение



Рис. 2: Дистилляция

Table 1: **DistilBERT retains 97% of BERT performance.** Comparison on the dev sets of the GLUE benchmark. ELMo results as reported by the authors. BERT and DistilBERT results are the medians of 5 runs with different seeds.

Model	Score	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

## Дистилляция ансамбля [7]



System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

Table 1: Frame classification accuracy and WER showing that the distilled single model performs about as well as the averaged predictions of 10 models that were used to create the soft targets.

## Список литературы I

-  **Leo Breiman.** "Bagging Predictors". в: *Machine Learning* 24.2 (1996), с. 123—140.
-  **Yoav Freund и Robert E Schapire.** "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". в: *Journal of Computer and System Sciences* 55.1 (1997), с. 119—139. ISSN: 0022-0000. DOI: <https://doi.org/10.1006/jcss.1997.1504>. URL: <https://www.sciencedirect.com/science/article/pii/S00220009791504X>.
-  **L Breiman.** "Random Forests". в: *Machine Learning* 45 (окт. 2001), с. 5—32. DOI: [10.1023/A:1010950718922](https://doi.org/10.1023/A:1010950718922).
-  **Jerome H. Friedman.** "Greedy function approximation: A gradient boosting machine.". в: *The Annals of Statistics* 29.5 (2001), с. 1189—1232. DOI: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451). URL: <https://doi.org/10.1214/aos/1013203451>.

## Список литературы II



David H. Wolpert. “Stacked generalization”. в: *Neural Networks* 5.2 (1992), с. 241—259. ISSN: 0893-6080. doi:  
[https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1). URL: <https://www.sciencedirect.com/science/article/pii/S0893608005800231>.



Victor Sanh, Lysandre Debut, Julien Chaumond и Thomas Wolf.  
*DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* 2020. arXiv: 1910.01108 [cs.CL].



Geoffrey Hinton, Oriol Vinyals и Jeff Dean. *Distilling the Knowledge in a Neural Network.* 2015. arXiv: 1503.02531 [stat.ML].