


Towards Understanding Ensemble, Knowledge Distillation and Self- Distillation in Deep Learning

Давид Захаров
БПМИ202

Лучшая статья ICLR

 **Congratulations on winning the ICLR 2023 Outstanding Paper Honorable Mention!!**

ICLR 2023 Conference Program Chairs

21 Mar 2023

ICLR 2023 Conference Paper5565 Official Comment

Readers:  Everyone

[Show Revisions](#)

Комитет ICLR 2023 считает это очень интересным теоретическим объяснением, которое приводит к лучшему пониманию эффективности дистилляции.

Ансамбли

Предшествующие работы

- [Ensembles for feature selection: A review and future trends.](#)
- [Greedy function approximation: a gradient boosting machine.](#)
- [Bagging predictors. Machine learning](#)

Ансамбли

Предшествующие работы

- **Бустинг:** где коэффициенты, связанные с комбинациями отдельных моделей, действительно обучаются;
- **Бутстрэппинг/Бэггинг:** тренировочные данные различаются для каждой отдельной модели;
- Ансамбль моделей различных типов и архитектур;
- Ансамбль случайных признаков или деревьев решений.

Неудачные попытки

random feature mappings

- В некоторых случаях $f(W, x)$ может быть приближено с помощью:

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

Где

W_0 - это случайная инициализация нейронной сети, и

$$\Phi_{W_0}(x) = \nabla_W f(W_0, x)$$

является отображением признаков neural tangent kernel (NTK).

Неудачные попытки

random feature mappings

- Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern recognition.
- Diversity in search strategies for ensemble feature selection. Information fusion.

Неудачные попытки

random feature mappings

- Традиционные теоремы предполагают, что ансамбль независимо обученных моделей со случайными признаками действительно может значительно улучшить производительность во время тестирования за счет увеличения пространства признаков:

$$\Phi_{W_0}(x) \mapsto \{\Phi_{W_0^{(i)}}(x)\}_{i \in [L]} \text{ for } L \text{ many sampled } W_0^{(i)}$$

Неудачные попытки

random feature mappings

- Противоречие: усредненное обучение работает еще лучше;

$$F(x) = \frac{1}{L}(f_1 + f_2 + \dots + f_L)$$

Ансамбль линейных функций на основе признаков NTK действительно улучшает точность теста, но только благодаря более крупному набору случайных признаков, чьи комбинации лучше обобщают.

Неудачные попытки

Взгляд через призму смещения и разброса

Некоторые предыдущие работы также пытаются объяснить преимущество ансамблей снижением разброса индивидуальных решений из-за шума в метках или невыпуклого ландшафта целевой функции обучения.

- Управление разнообразием в ансамблях регрессии. Журнал исследований машинного обучения;
- Экспериментальный анализ смещения и разброса ансамблей SVM на основе методов повторной выборки. Транзакции IEEE по системам, людям и кибернетике.

Неудачные попытки

Взгляд через призму смещения и разброса

- Снижение разброса может уменьшить выпуклую тестовую потерю, но не обязательно ошибку тестовой классификации;
- Более того, на практике обычно отдельные нейронные сети обучаются одинаково хорошо, то есть с почти идентичной ошибкой на тесте, тем не менее, ансамбль этих моделей все же улучшает точность тестирования.

Новизна исследования

- Ансамбли и дистилляция знаний в глубоком обучении работают совершенно иначе, чем традиционная теория обучения;
- Тщательный и обоснованный теоретический анализ, который уменьшает разрыв между теорией и практикой;
- В предыдущих работах не было подобной теории.

Потенциальные направления работы

- Повышение производительности: авторы убеждены, что практически во всех приложениях глубокого обучения, где ансамбли/дистилляция знаний еще не использованы по максимуму, существует потенциал для улучшения.