What Can Transformers Learn In-Context? A Case Study of Simple Function Classes

Докладчик: Марк Рофин
Рецензент: Иван Мошков
Хакер: Юлия Кокорина

Abstract

In-context learning refers to the ability of a model to condition on a prompt sequence consisting of in-context examples (input-output pairs corresponding to some task) along with a new query input, and generate the corresponding output. Crucially, in-context learning happens only at inference time without any parameter updates to the model. While large language models such as GPT-3 exhibit some ability to perform in-context learning, it is unclear what the relationship is between tasks on which this succeeds and what is present in the training data. To make progress towards understanding in-context learning, we consider the well-defined problem of training a model to in-context learn a function class (e.g., linear functions): that is, given data derived from some functions in the class, can we train a model to in-context learn "most" functions from this class? We show empirically that standard Transformers can be trained from scratch to perform in-context learning of linear functions—that is, the trained model is able to learn unseen linear functions from in-context examples with performance comparable to the optimal least squares estimator. In fact, in-context learning is possible even under two forms of distribution shift: (i) between the training data of the model and inference-time prompts, and (ii) between the in-context examples and the query input during inference. We also show that we can train Transformers to in-context learn more complex function classes—namely sparse linear functions, two-layer neural networks, and decision trees—with performance that matches or exceeds task-specific learning algorithms. [1]

In-Context Learning

Cat -> kitten
Dog -> puppy
Cow -> calf
Sheep -> lamb

They lived in New York -> They lived in Novosibirsk
His favourite economist was Adam Smith -> His favourite economist was Karl Marx
She worked on a farm -> She worked in kolkhoz
They celebrated Thanksgiving -> They celebrated Maslenitsa

Можно ли измерить способность модели к in-context learning?

In-context learning a function class

1. Пространство объектов: $\mathcal{X}$
2. Распределение на объектах: $x \sim D_{\mathcal{X}}$
3. Класс функций: $\mathcal{F}$
4. Распределение на функциях: $f \sim D_{\mathcal{F}}$
5. Промпт P: $(x_1, f(x_1), ..., x_k, f(x_k), x_{\text{query}}),\ \ x_* \sim D_{\mathcal{X}},\ f \sim D_{\mathcal{F}}$
6. Функция потерь: l

In-context learning a function class

Модель M in-context learns класс $\mathcal{F}$ при заданных $D_{\mathcal{X}}, D_{\mathcal{F}}$ с точностью ϵ, если она может предсказать $f(x_{\text{query}})$ с ошибкой меньше ϵ.

$$\mathbb{E}_P\left[\ell\left(M\left(P\right), f\left(x_{\text{query}}\right)\right)\right] \leq \epsilon,$$

In-context learning a function class

Если $\mathcal{F}$ – класс линейных функций, то:

1. $\mathcal{X} = \mathbb{R}^d$
2. $D_{\mathcal{X}} = \mathcal{N}(0, I_d)$
3. $\mathcal{F} = \{x \longrightarrow w^T x \mid w \in \mathbb{R}^d\}$
4. $D_{\mathcal{F}} = D_w = \mathcal{N}(0, I_d)$
5. $l(y, \hat{y}) = (y - \hat{y})^2$

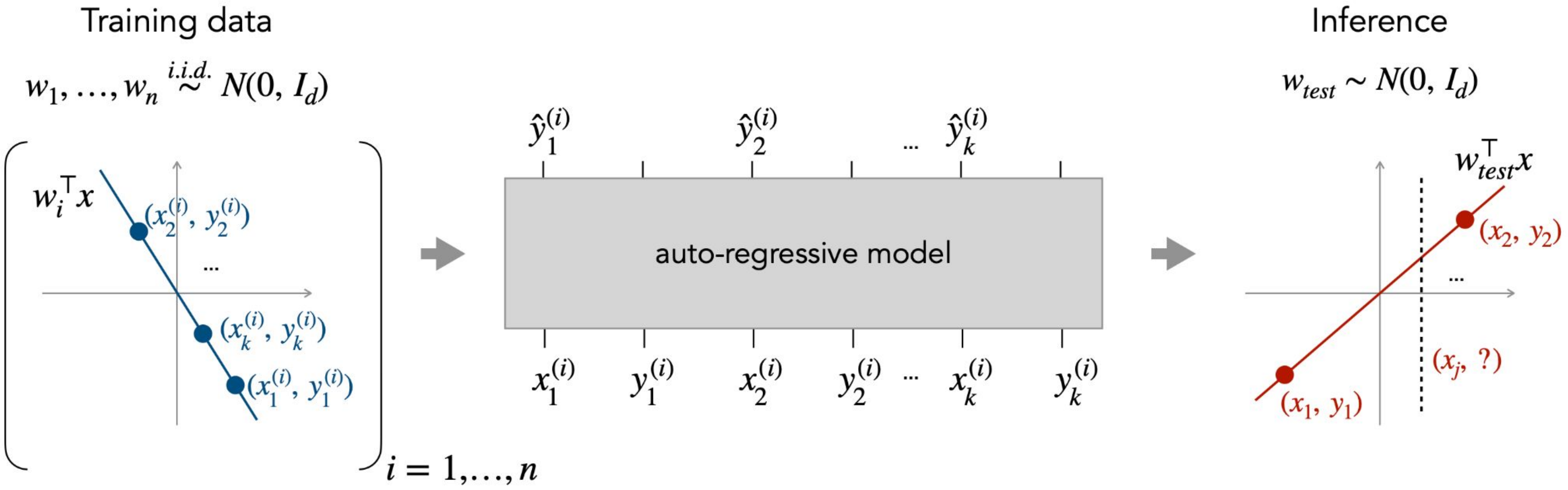
Можно ли заставить модель выучить заданный класс функций?

Обучение модели на класс функций

- Берут маленький не-предобученный чекпоинт GPT-2
- Выбирают класс функций и сэмплируют из него промпты P для обучающей выборки
- Учат GPT-2 авторегрессионно на этих промптах
- Используют curriculum learning (сначала учатся на промптах, где функции попроще)

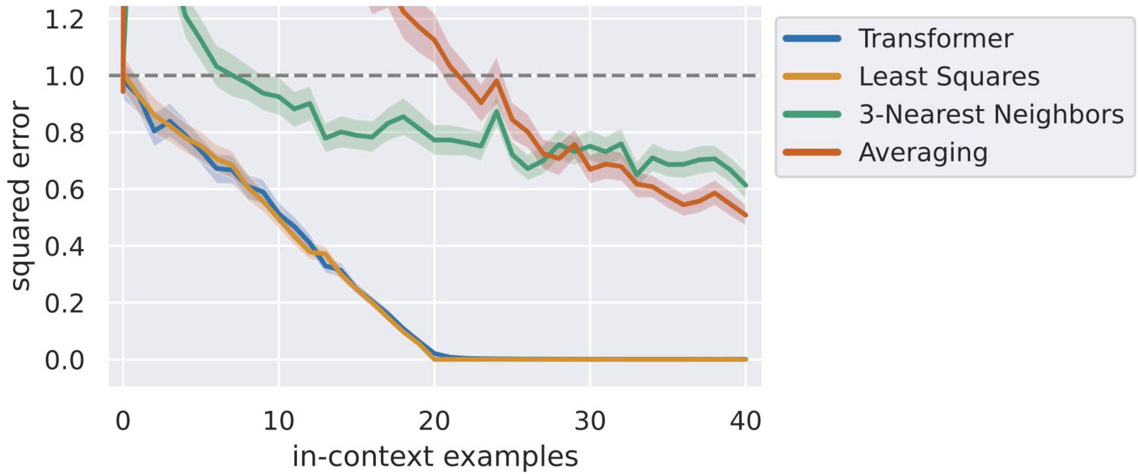
$$\min_{\theta} \mathbb{E}_P \left[\frac{1}{k+1} \sum_{i=0}^{k} \ell \left(M_\theta \left(P^i \right), f \left(x_{i+1} \right) \right) \right]$$

Обучение модели на класс функций



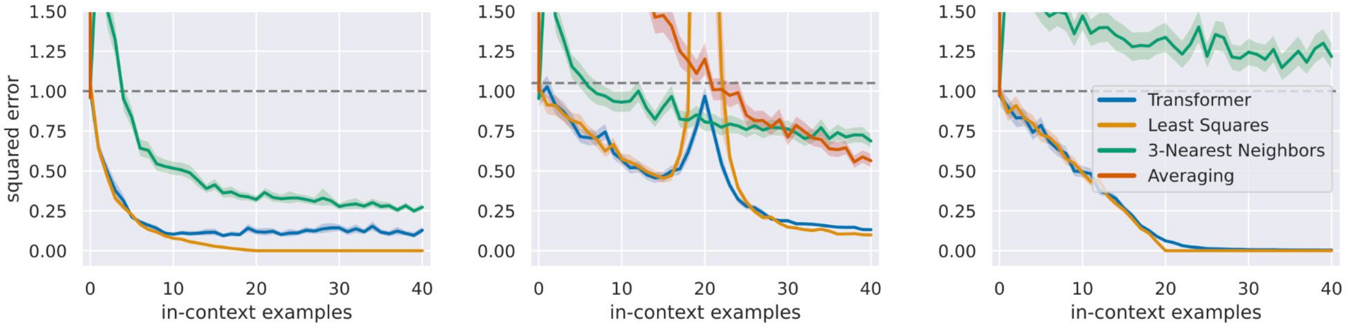
Результаты

Линейные функции



На 20 примерах ошибка 0, потому что здесь d = 20.

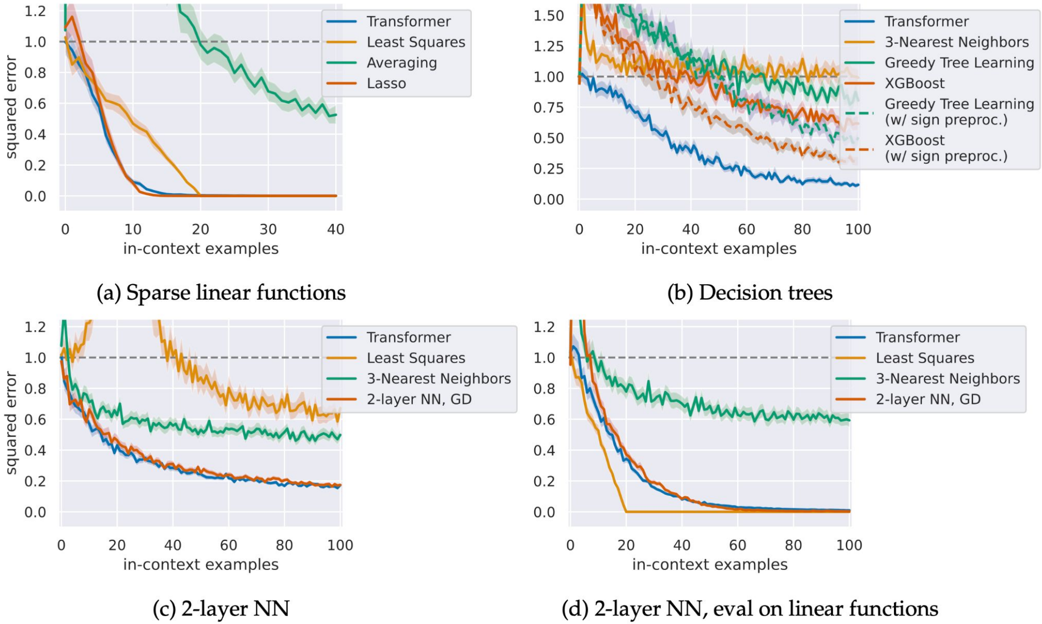
Линейные функции (distribution shift)



(a) Skewed covariance (b) Noisy linear regression (c) Different orthants

Figure 4: *In-context learning on out-of-distribution prompts.* We evaluate the trained model on prompts that deviate from those seen during training by: (a) sampling prompt inputs from a non-isotropic Gaussian, (b) adding label noise to in-context examples, (c) restricting in-context examples to a single (random) orthant. In all cases, the model error degrades gracefully and remains close to that of the least squares estimator, indicating that its in-context learning ability extrapolates beyond the training distribution.

Более сложные классы функций



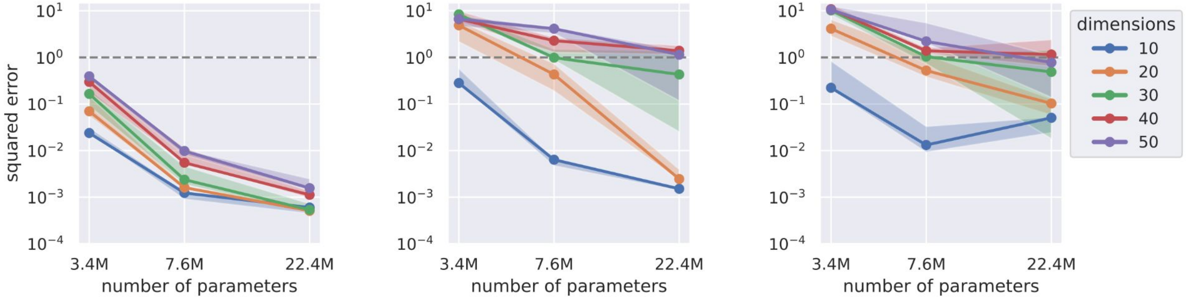
(a) Sparse linear functions

(b) Decision trees

(c) 2-layer NN

(d) 2-layer NN, eval on linear functions

Ablations



(a) Standard (b) Different orthants (c) Skewed covariance

Figure 6: *Understanding the effect of model capacity and problem dimension on in-context learning performance for in-distribution (a) and out-of-distribution (b,c) prompts.* We train Transformers to in-context learn linear functions and plot the error with $2d$ in-context examples as we vary problem dimension d and model capacity. Capacity helps with in-context learning in most cases, especially on out-of-distribution prompts (even when the absolute gains in the in-distribution setting are small). We train 3 models in each case with different random seeds, and show the median error (solid lines), and the minimum and maximum errors (shaded region). (See Appendix B.4 for training variance analysis.)

Итоги

- Трансформеры умеют in-context выучивать сложные классы функций
- Эти функции могут быть очень нетривиальными (оптимизация, подбор решающих деревьев)
- Увеличение числа параметров в трансформерах помогает
- Curriculum learning помогает ещё больше