Deep Learning for Symbolic Regression

Author: **Vastl Martin**, Supervisor: Mgr. Martin Pilát, Ph.D.

Charles University, Faculty of Mathematics and Physics

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Motivation

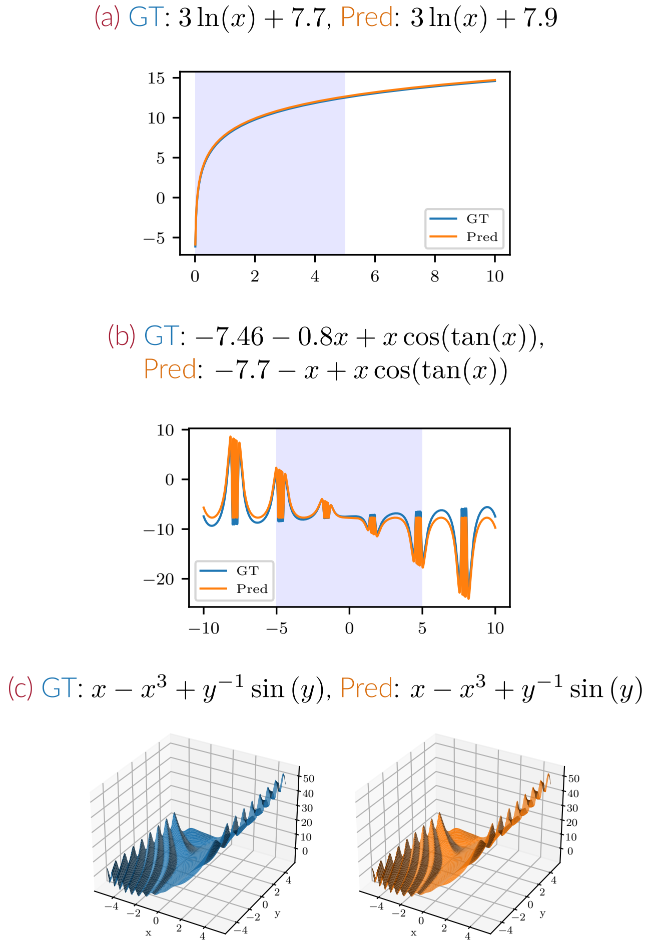
In the real world, many natural processes can be described by a formula. Automatically deriving such formula would not only allow us to make predictions for unseen inputs but it would also allow us to get insight into the inner workings of the process. The task of finding mathematical formula based on the observed input-output pairs is called *Symbolic regression* and has applications in many scientific areas.

Method

- We propose a transformer-based model, which is trained using supervised learning on millions of functions.
- We model the task of formula prediction as a sequence prediction, where each operation or operand is a single token.
- At each timestep, our model predicts two outputs, coefficient and symbolic representation.
- We use a novel coefficient encoding, which stabilizes the training and improves the model's performance.
- The generated coefficients are then used as an initialization for a local gradient search, which further improves the performance.

Example predictions

Figure 1. The shaded area represents the sampling range. GT refers to ground-truth and Pred to the found formula.



(a) GT: $3\ln(x) + 7.7$, Pred: $3\ln(x) + 7.9$

(b) GT: $-7.46 - 0.8x + x\cos(\tan(x))$,
Pred: $-7.7 - x + x\cos(\tan(x))$

(c) GT: $x - x^3 + y^{-1}\sin(y)$, Pred: $x - x^3 + y^{-1}\sin(y)$

Results

- Our model is comparable to the current state-of-the-art methods in terms of performance while outperforming them in inference time on several benchmarks.
- We show that functions generated by our model perform well on new inputs inside and outside of the sampling range.
- We have examined the model's output and found out that the model is able to predict semantically equivalent formulas, e.g., $\sin x = \cos\left(x - \frac{\pi}{2}\right)$.

Contributions

- We design a transformer-based model, which predicts the formula in an end-to-end manner without the need to find the coefficients in the post-processing step.
- We propose a novel way how to use and encode the coefficients to improve the model's performance.
- The model is thoroughly evaluated and compared to the current state-of-the-art methods.
- The architecture decisions are validated in the ablation study and in the experimental section.