A Seq2Seq approach to Symbolic Regression

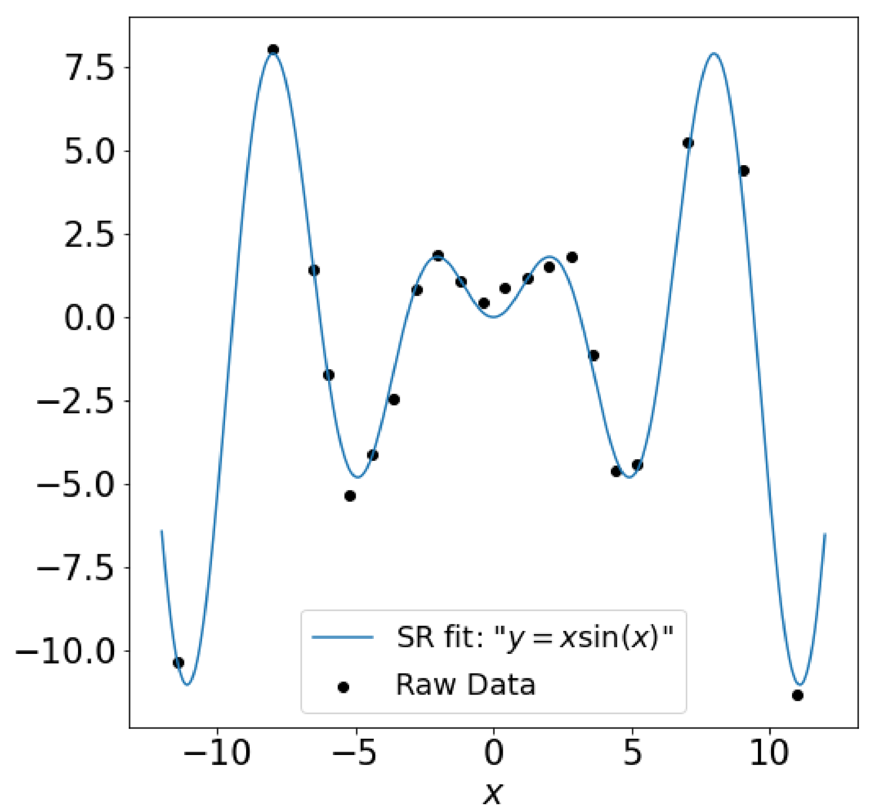
Luca Biggio (ETH, Zürich), Tommaso Bendinelli (CSEM, Alpnach), Aurelien Lucchi (ETH, Zürich), Giambattista Parascandolo (ETH, Zürich; MPI, Tübingen)

KR2ML

Knowledge Representation & Reasoning Meets Machine Learning

Symbolic Regression

Symbolic Regression (SR) is about discovering a symbolic mathematical expression that provides a simple yet accurate fit to a given data set.



Dataset Generation

We generate a dataset consisting of :

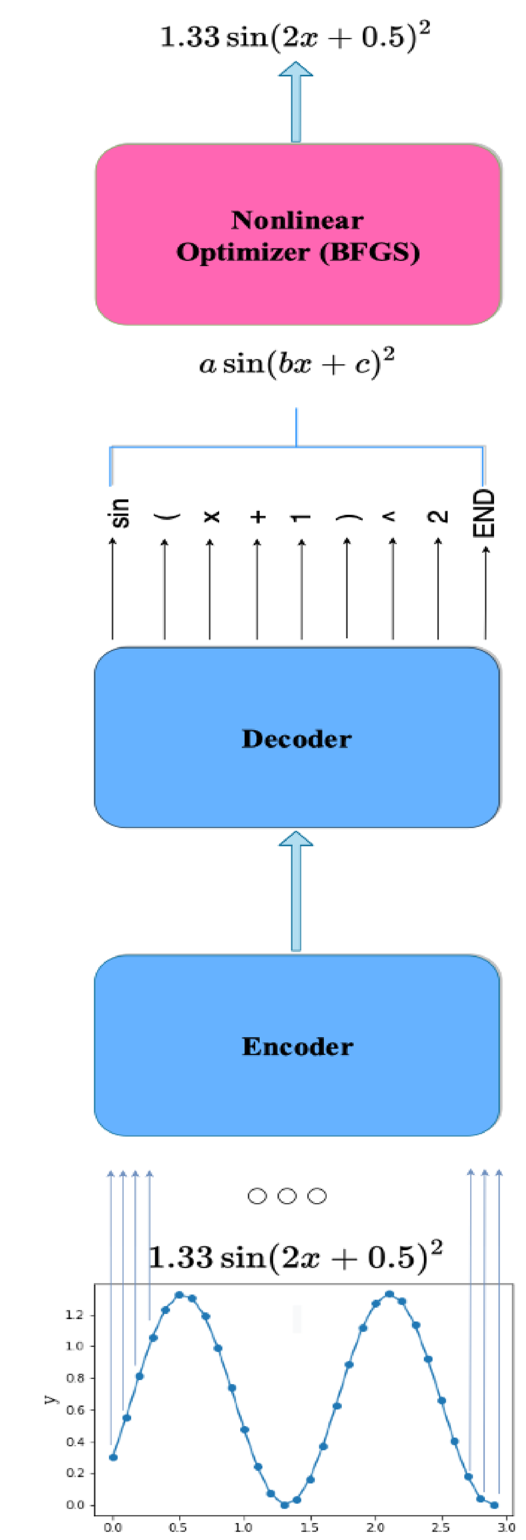
- **Input**: Numerical data points
- **Output**: Functional form used to generate the input

Input	Output
Evaluation Points $(\mathbf{Y}_i)$	Symbolic Equation (g_i)
$-6.24, -3.86, ..., 46.71, 49.49$	$x^2 + \log(x)$
$1, 1.01, ..., 3.01, 1.06$	$\exp(\sin(x^3))$
$0, 0.01, ..., 2309, 2837$	$x^6 + x^5 + x^4$
$-1.45, 0.83, ..., 17.66, 18.89$	$x^2 + \log(x)$

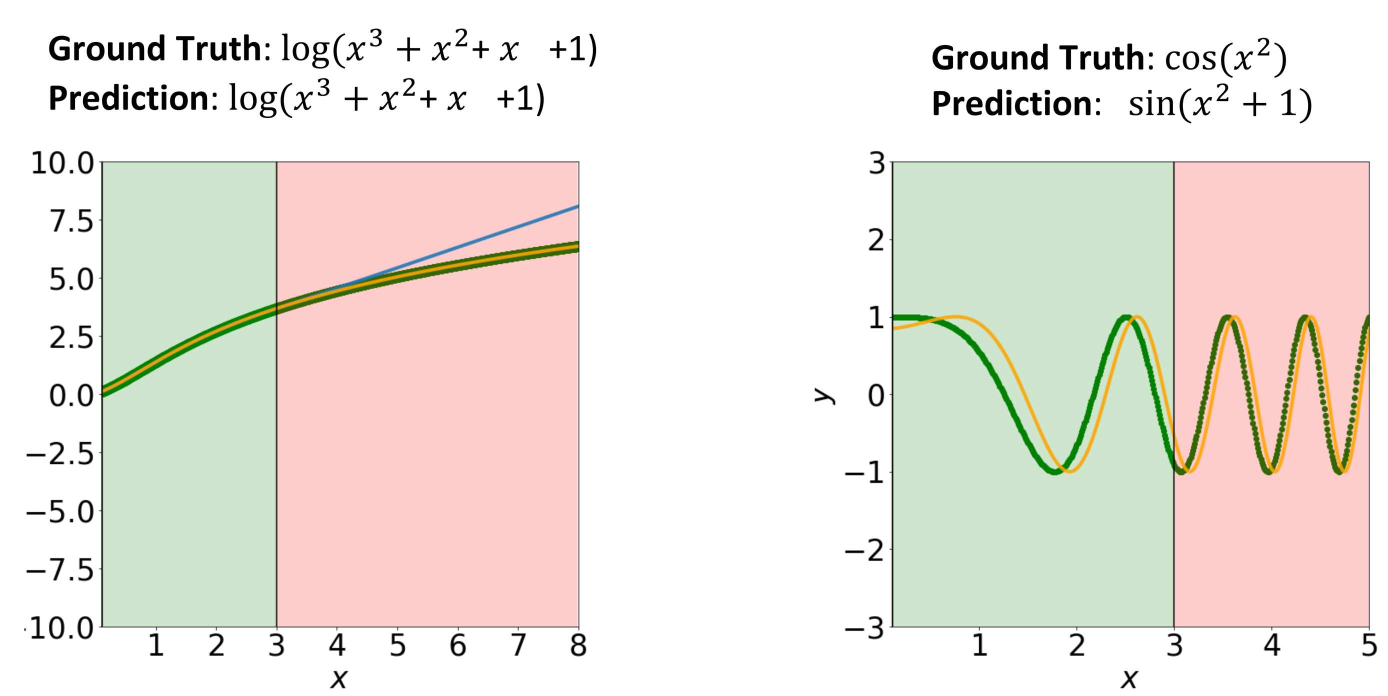
A seq2seq approach to SR

We adapt the fully convolutional seq2seq architecture proposed by Gehring et. Al [2017] to the SR setting:

- The **encoder** takes numerical data as input
- The **decoder** outputs a symbolic expression conditioned on the encoder embedding
- Numerical constants are fitted in as second and independent step by the BFGS **nonlinear optimizer**



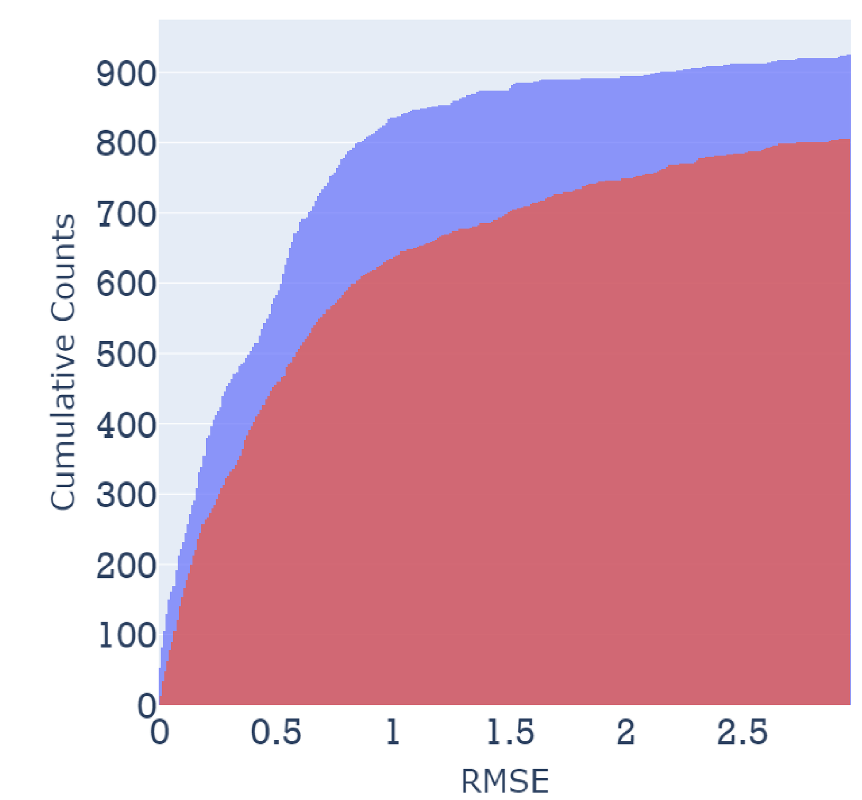
Qualitative Results



(**Left**) Our model (green dots) retrieves the correct underlying equation (yellow dots) given only training data in the green area. In contrast a 3-layer fully-connected neural network (blue line) fails to extraploate (**Right**) Our model does not output the correct expression since *cos* is not included into the training dictionary. However, it captures the sinusoidal nature of the signal

Quantitative Results

We count the number of test equations (out of 1000) for which the extrapolation RMSE is below a certain threshold. We compare our method (blue) with a 3-layer fully-connected neural network (red)



Future Directions

- Extend our method to handle multivariate equations
- Recursively refine the network output at test time by combining symbolic and numerical losses
- Increase the size of the training set
- Explore more advanced architectures from the NLP literature (e.g. Transformer ([Vaswani et al.,2017])

References

- Gehring, Jonas, et al. "Convolutional sequence to sequence learning." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 2017.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).