Measuring Arithmetic Extrapolation Performance

Andreas Madsen
Computationally Demanding
amwebdk@gmail.com

Alexander Rosenberg Johansen
Technical University of Denmark
aler@dtu.dk

Introduction

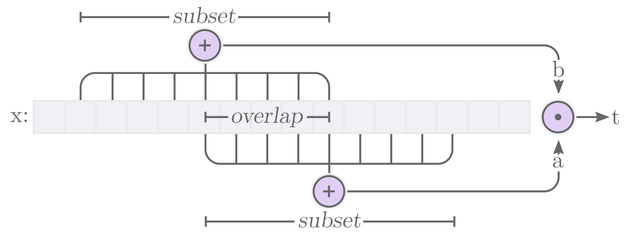
When using neural networks to learn simple arithmetic problems, such as counting, multiplication, or comparison they systematically fail to extrapolate onto unseen ranges. The absence of inductive bias makes it difficult for neural networks to extrapolate well on arithmetic tasks as they lack the underlying logic to represent the required operations.

A recently proposed model, called NALU [Trask et al., 2018], attempts to solve the problem of arithmetic extrapolation. However, for arithmetic extrapolation there are no broadly accepted guidelines for evaluating model performance. As a result, single-instance MSE is used for comparison.

As exact extrapolation requires correctly solving a logical problem we advocate that the performance metrics of interest should be: 1) has it learned the underlying logic, 2) how often does it learn the correct solution, and 3) how fast does it converge?

Simple Function Task

The "Simple Function Learning Tasks" is a synthetic dataset that tests arithmetic extrapolation. The problem is defined as summing two random subsets of $\mathbf{x}$ followed by an arithmetic operation $\{+,-,\times,\div\}$ on these sums. Extrapolation can then be tested by modifying the sampling range of x.



Simple Function Task: Shows how the dataset is parametized into, subset (ratio), overlap (ratio), input size (integer), operation (one of $\{+,-,\times,\div\}$).

Performance Metrics

As logic is discrete, a solution to the problem is either correct or wrong. To evaluate a solution we propose comparing the MSE, of the entire testset, to the MSE of a nearly-perfect solution on the extrapolation range.

To evaluate a solution we propose comparing the MSE, of the entire testset, to the MSE of a nearly-perfect solution on the extrapolation range. The nearly-perfect solution is defined as performing the operation perfectly, but allowing a small error in the sum-of-subsets. This threshold can be simulated with $\frac{1}{N}\sum_{i=1}^{N}(\mathrm{Op}(\mathbf{W}^{\epsilon}_{1,:}\mathbf{x}_i, \mathbf{W}^{\epsilon}_{2,:}\mathbf{x}_i) - t_i)^2$ for $N = 1000000$, where $\mathbf{W}^{\epsilon} = \mathbf{W}^{*} \pm \epsilon$ and $\mathbf{W}^{*}$ is the perfect $\mathbf{W}$ required to compute the optimal solution. We set $\epsilon = 10^{-5}$.

Using a success-criterion has the advantage of being more interpretable, models that failed to converge will not obscure the mean, and as the number of successes will follow a binomial distribution we can calculate a confidence interval [Wilson, 1927]. With a success-criterion we can also evaluate when a model succeeds and report a 95% confidence intervals, by using a gamma distribution and maximum likelihood profiling.

Finally, the parameters of the NALU are said to be "biased to be close to -1, 0, -1" [Trask, et al., 2018]. To test, we measure a sparsity error of the NALU parameters with $max_i\ min(|\boldsymbol{W}_i|, |1 - |\boldsymbol{W}_i||)$. A 95% confidence interval is produced using a beta distribution with support in [0, 0.5].

Neural Arithmetic Logic Unit

The Neural Arithmetic Logic Unit (NALU) [Trask et al., 2018] consists of two sub-units; the NAC_+ and $\mathrm{NAC}_\bullet$. The sub-units represent either the $\{+,-\}$ or the $\{\times,\div\}$ operations. The NALU then assumes that either NAC_+ or $\mathrm{NAC}_\bullet$ will be selected exclusively, using a sigmoid gating-mechanism.

$$W_{h_\ell,h_{\ell-1}} = \tanh(\hat{W}_{h_\ell,h_{\ell-1}})\sigma(\hat{M}_{h_\ell,h_{\ell-1}}) \tag{1}$$

$$\mathrm{NAC}_+ : \ z_{h_\ell} = \sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell,h_{\ell-1}} z_{h_{\ell-1}} \tag{2}$$

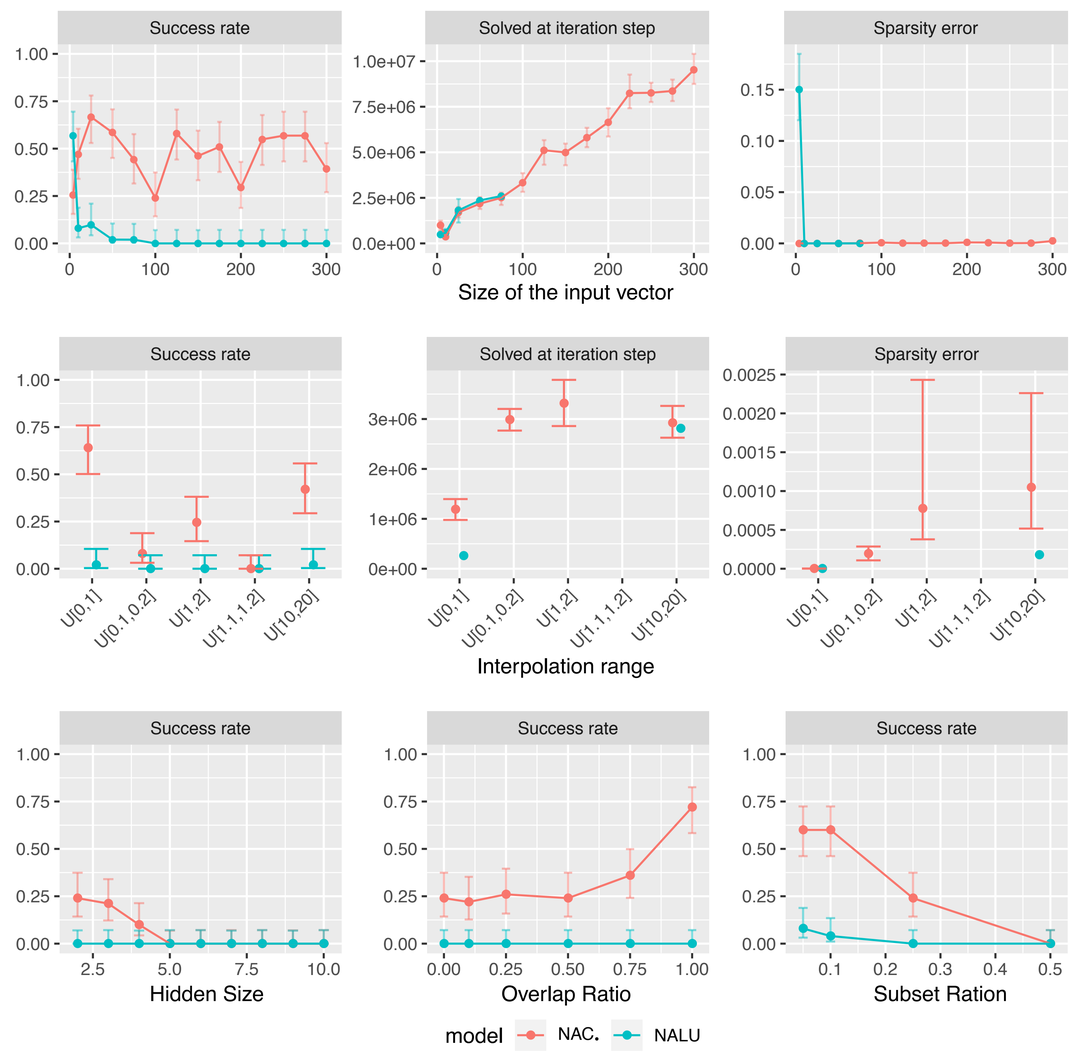
$$\mathrm{NAC}_\bullet : \ z_{h_\ell} = \exp\left(\sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell,h_{\ell-1}} \log(|z_{h_{\ell-1}}| + \epsilon)\right) \tag{3}$$

The matrices $\hat{\mathbf{W}}$ and $\hat{\mathbf{M}}$, are combined using a tanh-sigmoid transformation to bias the parameters towards a {-1,0,1} solution.

The NALU combines these units with a gating mechanism $\boldsymbol{z} = \boldsymbol{g} \cdot NAC_+ + (\boldsymbol{1} - \boldsymbol{g}) \cdot NAC_\bullet$ given $\mathbf{g} = \sigma(\mathbf{Gx})$. Thus allowing NALU to decide between all of the $\{+,-,\times,\div\}$ operations using backpropagation.

Results

We present results for 4800 experiments, all instances are trained with default Adam. The validation dataset is fixed with 10^4 observations sampled from the interpolation range. The test dataset is fixed with 10^4 observations sampled from the extrapolation range.



Effect of Parameters: Shows the effect of the dataset parameters or increasing the hidden size of the second NALU or the $\mathrm{NAC}_\bullet$ layer. Metrics are success-rate, when models converged, and sparsity error, reported with a 95% confidence interval of the mean, using 50 different initialization seeds. Unless explicitly changed the parameters are: imput size = 100, overlap ratio = 0.5, subset ratio = 0.25, interpolation range = U[1,2].

Op	Model	Success Rate	Solved at	Sparsity error
$\times$	$\mathrm{NAC}_\bullet$	$\mathbf{31\%}\ ^{+10\%}_{-8\%}$	$\mathbf{3.0 \cdot 10^6}\ ^{+2.9\cdot10^5}_{-2.4\cdot10^5}$	$\mathbf{5.8 \cdot 10^{-4}}\ ^{+4.8\cdot10^{-4}}_{-2.6\cdot10^{-4}}$
	NALU	$0\%\ ^{+4\%}_{-0\%}$	—	—
$/$	$\mathrm{NAC}_\bullet$	$\mathbf{0\%}\ ^{+4\%}_{-0\%}$	—	—
	NALU	$\mathbf{0\%}\ ^{+4\%}_{-0\%}$	—	—
$+$	NAC_+	$\mathbf{100\%}\ ^{+0\%}_{-4\%}$	$4.9 \cdot 10^5\ ^{+5.2\cdot10^4}_{-4.5\cdot10^4}$	$2.3 \cdot 10^{-1}\ ^{+6.5\cdot10^{-3}}_{-6.5\cdot10^{-3}}$
	Linear	$\mathbf{100\%}\ ^{+0\%}_{-4\%}$	$\mathbf{6.3 \cdot 10^4}\ ^{+2.5\cdot10^3}_{-3.3\cdot10^3}$	$2.5 \cdot 10^{-1}\ ^{+3.6\cdot10^{-4}}_{-3.6\cdot10^{-4}}$
	NALU	$14\%\ ^{+8\%}_{-5\%}$	$1.6 \cdot 10^6\ ^{+3.8\cdot10^5}_{-3.3\cdot10^5}$	$\mathbf{1.7 \cdot 10^{-1}}\ ^{+2.7\cdot10^{-2}}_{-2.5\cdot10^{-2}}$
$-$	NAC_+	$\mathbf{100\%}\ ^{+0\%}_{-4\%}$	$\mathbf{3.7 \cdot 10^5}\ ^{+3.8\cdot10^4}_{-3.8\cdot10^4}$	$2.3 \cdot 10^{-1}\ ^{+5.4\cdot10^{-3}}_{-5.4\cdot10^{-3}}$
	Linear	$7\%\ ^{+7\%}_{-4\%}$	$1.4 \cdot 10^6\ ^{+7.0\cdot10^5}_{-6.1\cdot10^5}$	$\mathbf{1.8 \cdot 10^{-1}}\ ^{+7.2\cdot10^{-2}}_{-5.8\cdot10^{-2}}$
	NALU	$14\%\ ^{+8\%}_{-5\%}$	$1.9 \cdot 10^6\ ^{+4.4\cdot10^5}_{-4.5\cdot10^5}$	$2.1 \cdot 10^{-1}\ ^{+2.2\cdot10^{-2}}_{-2.2\cdot10^{-2}}$

Effect of Dataset Operation: Metrics are success-rate, when models converged, and sparsity error, reported with a 95% confidence interval of the mean, using 100 different initialization seeds.

References

Andrew Trask, Felix Hill, Scott E Reed, Jack Rae, Chris Dyer, and Phil Blunsom. Neural arithmetic logic units. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 8035–8044. Curran Associates, Inc., 2018. URL *http://papers.nips.cc/paper/8027-neural-arithmetic-logic-units.pdf*.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL *https://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502953*.

DTU Technical University of Denmark

Read full paper on arXiv
https://arxiv.org/abs/1910.01888