



## Measuring Arithmetic Extrapolation Performance

**Johansen, Alexander Rosenberg; Madsen, Andreas**

*Published in:*

Science meets Engineering of Deep Learning at 33rd Conference on Neural Information Processing Systems

*Publication date:*

2019

*Document Version*

Early version, also known as pre-print

[Link back to DTU Orbit](#)

*Citation (APA):*

Johansen, A. R., & Madsen, A. (2019). Measuring Arithmetic Extrapolation Performance. In *Science meets Engineering of Deep Learning at 33rd Conference on Neural Information Processing Systems*

---

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

---

# Measuring Arithmetic Extrapolation Performance

---

**Andreas Madsen**  
 Computationally Demanding  
 amwebdk@gmail.com

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

## Abstract

The Neural Arithmetic Logic Unit (NALU) is a neural network layer that can learn exact arithmetic operations between the elements of a hidden state. The goal of NALU is to learn perfect extrapolation, which requires learning the exact underlying logic of an unknown arithmetic problem. Evaluating the performance of the NALU is non-trivial as one arithmetic problem might have many solutions. As a consequence, single-instance MSE has been used to evaluate and compare performance between models. However, it can be hard to interpret what magnitude of MSE represents a correct solution and models sensitivity to initialization. We propose using a success-criterion to measure if and when a model converges. Using a success-criterion we can summarize success-rate over many initialization seeds and calculate confidence intervals. We contribute a generalized version of the previous arithmetic benchmark to measure models sensitivity under different conditions. This is, to our knowledge, the first extensive evaluation with respect to convergence of the NALU and its sub-units. Using a success-criterion to summarize 4800 experiments we find that consistently learning arithmetic extrapolation is challenging, in particular for multiplication.<sup>1</sup>

## 1 Introduction

When using neural networks to learn simple arithmetic problems, such as counting, multiplication, or comparison they systematically fail to extrapolate onto unseen ranges [Lake and Baroni, 2018, Suzgun et al., 2019, Trask et al., 2018]. 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?

Motivated by these questions we propose using a success-criterion to determine if the underlying logic has been learned. We measure success-rate and provide a binomial confidence interval by initializing and training the NALU over multiple seeds. For each seed, we use the first iteration that satisfy the success-criterion to measure when the model has succeeded. As the success-criterion is based on an MSE divergence from an optimal solution it can be generalized to any model.

Finally, we propose and report a sparsity measurement for models that satisfy the success-criterion. Sparsity of the parameters has previously been emphasized as important for a correct solution [Trask et al., 2018].

---

<sup>1</sup>code for experiments is publicly available at: <https://github.com/AndreasMadsen/stable-nalu>.

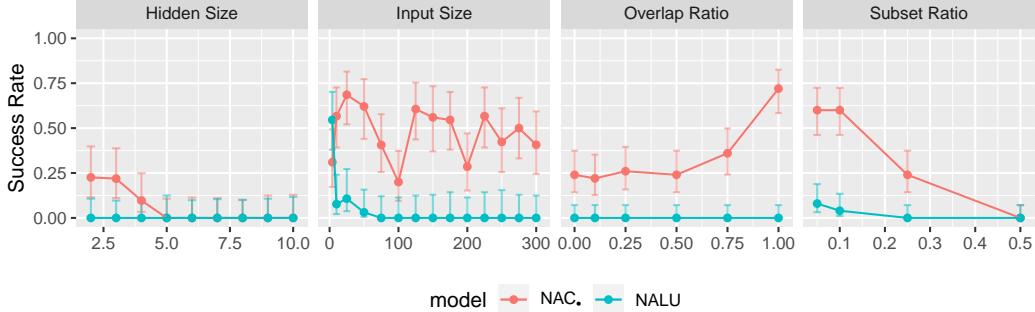


Figure 1: Shows success-rate given different dataset parameters and the models hidden-size using the multiplication operation. Means are over 50 different seeds, with 95% confidence intervals.

## 2 Related work

Kalchbrenner et al. [2016], Zaremba and Sutskever [2014], Kaiser and Sutskever [2016], Freivalds and Liepins [2017] solves integer arithmetic operations as a classification task and reports exact match accuracy. Using accuracy is useful for well-defined classification tasks, but is hard to use for real number regression problems. Our criterion mimics exact match by defining an MSE  $\epsilon$ -threshold.

## 3 Simple Function Learning Tasks

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

**Algorithm 1** Dataset sampling algorithm. Default values are specified for input-size ( $d$ ), subset-ratio ( $s$ ), and overlap-ratio ( $o$ ). Default interpolation range is  $[1, 2]$  and default extrapolation range is  $[2, 6]$ .

---

```

1: function DATASET( $\text{OP}(\cdot, \cdot)$  : Operation,  $R$  : Range,  $d = 100$ ,  $s = 0.25$ ,  $o = 0.5$ )
2:    $\mathbf{x} \leftarrow \text{UNIFORM}(R_{lower}, R_{upper}, i)$                                  $\triangleright$  Sample  $d$  elements uniformly
3:    $k \leftarrow \text{UNIFORM}(0, 1 - 2s - o)$      $\triangleright$  Sample offset. Same for interpolation and extrapolation.
4:    $a \leftarrow \text{SUM}(\mathbf{x}[dk : d(k + s)])$            $\triangleright$  Create sum  $a$  from a subset of length  $s \cdot d$ 
5:    $b \leftarrow \text{SUM}(\mathbf{x}[d(k + s - o) : d(k + 2s - o)])$      $\triangleright$  Create sum  $b$  from a subset of length  $s \cdot d$ 
6:    $t \leftarrow \text{OP}(a, b)$                                  $\triangleright$  Perform operation on  $a$  and  $b$ 
7:   return  $\mathbf{x}, t$ 
```

---

Solving the task on extrapolation requires learning the underlying logic of arithmetic operations from the training range. 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. The nearly-perfect solution is defined as performing the operation perfectly, but allowing a small error in the sum-of-subsets (line 4 and 5 in Algorithm 1). This threshold can be simulated with  $\frac{1}{N} \sum_{i=1}^N (\text{Op}(\mathbf{W}_1^\epsilon, \mathbf{x}_i, \mathbf{W}_2^\epsilon, \mathbf{x}_i) - t_i)^2$  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 evaluate when a model succeeds. Since this metric cannot be negative, we model the confidence interval with a gamma distribution and report a 95% confidence intervals of the mean, by using maximum likelihood profiling.

Finally, the parameters of the NALU are argued to be “biased to be close to -1, 0, -1” [Trask et al., 2018]. We propose to measure a sparsity error of the NALU parameters with  $\max_i \min(|W_i|, |1 - |W_i||)$ . As the sparsity error is between  $[0, 0.5]$  we use a modified beta distribution with support in  $[0, 0.5]$  and report a 95% confidence interval of the mean, by using maximum likelihood profiling.

The choice of gamma and beta distribution may not be perfect. However, a normal distribution would be problematic when the mean is close to the bounds, as it will have a large probability mass outside of the support bounds and thus provide inaccurate confidence intervals.

## 4 Results

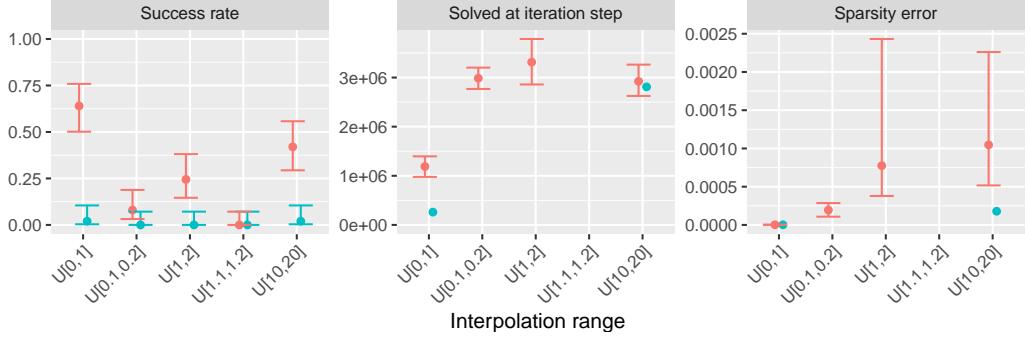


Figure 2: Shows success-rate, when models converged, and sparsity error for the multiplication operation. Means are over 50 seeds. We provide experimental details in Appendix A.

Table 1: Shows success-rate, when models converged, and sparsity error. Means are over 100 seeds.

Op	Model	Success Rate	Solved at	Sparsity error
×	NAC <sub>•</sub>	31% <sup>+10%</sup> <sub>-8%</sub>	$3.0 \cdot 10^6$ <sup>+2.9·10<sup>5</sup><sub>-2.4·10<sup>5</sup></sub></sup>	$5.8 \cdot 10^{-4}$ <sup>+4.8·10<sup>-4</sup><sub>-2.6·10<sup>-4</sup></sub></sup>
	NALU	0% <sup>+4%</sup> <sub>-0%</sub>	—	—
/	NAC <sub>•</sub>	0% <sup>+4%</sup> <sub>-0%</sub>	—	—
	NALU	0% <sup>+4%</sup> <sub>-0%</sub>	—	—
+	NAC <sub>+</sub>	100% <sup>+0%</sup> <sub>-4%</sub>	$4.9 \cdot 10^5$ <sup>+5.2·10<sup>4</sup><sub>-4.5·10<sup>4</sup></sub></sup>	$2.3 \cdot 10^{-1}$ <sup>+6.5·10<sup>-3</sup><sub>-6.5·10<sup>-3</sup></sub></sup>
	Linear	100% <sup>+0%</sup> <sub>-4%</sub>	$6.3 \cdot 10^4$ <sup>+2.5·10<sup>3</sup><sub>-3.3·10<sup>3</sup></sub></sup>	$2.5 \cdot 10^{-1}$ <sup>+3.6·10<sup>-4</sup><sub>-3.6·10<sup>-4</sup></sub></sup>
	NALU	$1.5 \cdot 10^6$	$1.6 \cdot 10^6$ <sup>+3.8·10<sup>5</sup><sub>-3.3·10<sup>5</sup></sub></sup>	$1.7 \cdot 10^{-1}$ <sup>+2.7·10<sup>-2</sup><sub>-2.5·10<sup>-2</sup></sub></sup>
-	NAC <sub>+</sub>	$9.0 \cdot 10^3$	$3.7 \cdot 10^5$ <sup>+3.8·10<sup>4</sup><sub>-3.8·10<sup>4</sup></sub></sup>	$2.3 \cdot 10^{-1}$ <sup>+5.4·10<sup>-3</sup><sub>-5.4·10<sup>-3</sup></sub></sup>
	Linear	7% <sup>+7%</sup> <sub>-4%</sub>	$1.4 \cdot 10^6$ <sup>+7.0·10<sup>5</sup><sub>-6.1·10<sup>5</sup></sub></sup>	$1.8 \cdot 10^{-1}$ <sup>+7.2·10<sup>-2</sup><sub>-5.8·10<sup>-2</sup></sub></sup>
	NALU	14% <sup>+8%</sup> <sub>-5%</sub>	$1.9 \cdot 10^6$ <sup>+4.4·10<sup>5</sup><sub>-4.5·10<sup>5</sup></sub></sup>	$2.1 \cdot 10^{-1}$ <sup>+2.2·10<sup>-2</sup><sub>-2.2·10<sup>-2</sup></sub></sup>

## 5 Conclusion

We provide the most extensive study of the Neural Arithmetic Logic Unit to date using a generalized version of the “Simple Function Learning Tasks”. Our study, through varying task complexities, evaluates the NALUs ability to learn the logic of arithmetic operations.

To evaluate performance on solving arithmetic operations we define a new success-criterion that approximates an exact match. With a success-criterion we measure how often a model successfully solve the problem given different initialization seeds, a binomial confidence interval, and at what iteration the model satisfy the criterion. Our results find that the NALU and its sub-units can require many trials to learn. In particular for multiplication and division. Furthermore, we find that for subtraction and addition the solution is not always sparse.

Our results are not different from the original results, but highlights the importance of also discussing a models sensitivity to initialization. We hope that future research will consider using success-rates as a comparison for the performance of arithmetic units.

## Acknowledgments

We would like to thank Andrew Trask and the other authors of the NALU paper, for highlighting the importance and challenges of extrapolation in Neural Networks. We would also like to thank the students Raja Shan Zaker Kreen and William Frisch Møller from The Technical University of Denmark, who initially showed us that the NALU does not converge consistently.

This research is funded by the Innovation Foundation Denmark through the DABAI project.

## References

- Karlis Freivalds and Renars Liepins. Improving the neural GPU architecture for algorithm learning. *CoRR*, abs/1702.08727, 2017. URL <http://arxiv.org/abs/1702.08727>.
- Lukasz Kaiser and Ilya Sutskever. Neural gpus learn algorithms. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1511.08228>.
- Nal Kalchbrenner, Ivo Danihelka, and Alex Graves. Grid long short-term memory. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1507.01526>.
- Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *The 3rd International Conference for Learning Representations, San Diego, 2015*, page arXiv:1412.6980, Dec 2014.
- Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2879–2888, 2018. URL <http://proceedings.mlr.press/v80/lake18a.html>.
- Mirac Suzgun, Yonatan Belinkov, and Stuart M. Shieber. On evaluating the generalization of lstm models in formal languages. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 277–286, January 2019.
- 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>.
- Wojciech Zaremba and Ilya Sutskever. Learning to execute. *CoRR*, abs/1410.4615, 2014. URL <http://arxiv.org/abs/1410.4615>.

## A Experimental details

### A.1 NALU definition

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

The  $\text{NAC}_+$  and  $\text{NAC}_\bullet$  are defined accordingly,

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

$$\text{NAC}_+ : z_{h_\ell} = \sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell, h_{\ell-1}} z_{h_{\ell-1}} \quad (2)$$

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

where  $\hat{\mathbf{W}}, \hat{\mathbf{M}} \in \mathbb{R}^{H_\ell \times H_{\ell-1}}$  are weight matrices and  $z_{h_{\ell-1}}$  is the input. The matrices are combined using a tanh-sigmoid transformation to bias the parameters towards a  $\{-1, 0, 1\}$  solution. Having  $\{-1, 0, 1\}$  allows  $\text{NAC}_+$  to perform exact  $\{+, -\}$  operations between elements of a vector. The  $\text{NAC}_\bullet$  uses an exponential-log transformation to create the  $\{\times, \div\}$  operations (within  $\epsilon$  precision).

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

### A.2 Model definitions and setup

Models are defined in table 2 and are all optimized with Adam optimization [Kingma and Ba, 2014] using default parameters, and trained over  $5 \cdot 10^6$  iterations. Training takes about 6 hours on a single CPU core (8-Core Intel Xeon E5-2665 2.4GHz). We run 4800 experiments on a HPC cluster.

The training dataset is continuously sampled from the interpolation range where a different seed is used for each experiment, all experiments use a mini-batch size of 128 observations, a fixed validation dataset with  $1 \cdot 10^4$  observations sampled from the interpolation range, and a fixed test dataset with  $1 \cdot 10^4$  observations sampled from the extrapolation range.

We evaluate each metric every 1000 iterations on the test set that uses the extrapolation range, and choose the best iteration based on the validation dataset that uses the interpolation range.

For figure 2, the following extrapolation ranges were used:  $U[-2, -1] \rightarrow U[-6, -2]$ ,  $U[-2, 2] \rightarrow U[-6, -2] \cup U[2, 6]$ ,  $U[0, 1] \rightarrow U[1, 5]$ ,  $U[0.1, 0.2] \rightarrow U[0.2, 2]$ ,  $U[1, 2] \rightarrow U[2, 6]$ ,  $U[10, 20] \rightarrow U[20, 40]$ .

Table 2: Model definitions

Model	Layer 1	Layer 2
$\text{NAC}_\bullet$	$\text{NAC}_+$	$\text{NAC}_\bullet$
$\text{NAC}_+$	$\text{NAC}_+$	$\text{NAC}_+$
NALU	NALU	NALU
Linear	Linear	Linear